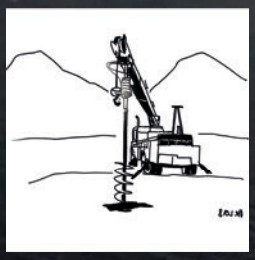
ETH zürich

Infozine S·2

Special Issue 2 | November 15, 2018

Surfing versus Drilling for Knowledge in Science:

When should you use your computer?
When should you use your brain?



Editors:
Philippe Hünenberger
Oliver Renn

Infozine is published by the
Chemistry | Biology | Pharmacy Information Center, ETH Zurich

infozentrum.ethz.ch
Chemie | Biologie | Pharmazie

Imprint

Infozine Special Issue S2
Surfing versus Drilling for Knowledge in Science:
When should you use your computer? When should you use your brain?

Editors:
Prof. Dr. Philippe Hünenberger
ETH Zurich, Laboratory of Physical Chemistry
Vladimir-Prelog Weg 2
8093 Zurich, Switzerland
Phone +41 44 632 55 03, phil@igc.phys.chem.ethz.ch

Dr. Oliver Renn
ETH Zurich, Chemistry | Biology | Pharmacy Information Center
Vladimir-Prelog Weg 10
8093 Zurich, Switzerland
Phone +41 44 632 29 64, renn@chem.ethz.ch

Copy-Editors: Oliver Renn, Philippe Hünenberger
Layout: Oliver Renn
Cover illustration: Oliver Renn, iPad Pro drawing using Procreate. Cover photo courtesy of Antonia Renn.

Please use the hashtag #infozineS2
and address @infozentrum (for Instagram, Facebook) or @icbpeth (for Twitter)
when mentioning articles in social media.

Infozine and its Special Issues are published by the Chemistry | Biology | Pharmacy Information Center,
a function within the Department of Chemistry and Applied Biosciences and the
Department of Biology at the ETH Zurich.
Infozine is published as an English and German edition, Special Issues in English only.
ISSN (English Edition) 2504-1851
ISSN (German Edition) 2504-1843
www.infozentrum.ethz.ch

Contents

Philippe Hünenberger and Oliver Renn
ETH Zurich

Editorial

Surfing versus Drilling for Knowledge in Science:

When should you use your computer? When should you use your brain?

Around 1660, the French scientist and philosopher Blaise Pascal wrote a beautiful text (reproduced as an addendum to this editorial) about the position of humans, stuck somewhere between two infinities: the infinitely large and the infinitely small. For mathematics lovers, the same double-infinities situation occurs in the ensemble of real numbers: however small the interval between two distinct numbers, it still contains an infinity of other ones; and however large a finite number, it is still surpassed in size by an infinity of larger ones. Somewhat similarly, the Gödel theorem shows that beyond a certain complexity, it is impossible to enumerate systematically all the theorems of an axiomatic system from within this system; i.e. the knowledge of any complex system cannot become complete unless you transcend its own limitations.

Although Pascal invented the first "arithmetic machines", ancestors of computers, he was probably far from envisioning the "digital world" we live in. The 21st century technologies, and in particular the internet, bring both infinities directly into our life – even into our pockets – as a permanent invite to search, process, and spread new information. But are we equipped mentally and socially to deal with the double-infinity in such an immediate and permanent proximity?

There is only so much earth you can shovel in a day. If you dig deep, you dig narrow. If you dig wide, you dig shallow. Diving into the infinitely small, i.e. "drilling", makes sense when the nuggets are hidden deep below the ground. It is a time-consuming and often lonely activity, that brings depth, quality, and insight. Dissolving into the infinitely large, i.e. "surfing", makes sense when the nuggets are widely spread just under the grass. It is a comparatively faster and easier activity, that brings overview, throughput and interactivity, but also a risk of information overflow.

The drilling vs. surfing duality can be interpreted in a variety of ways, e.g. researching a topic thoroughly vs. exploratively, thinking analytically vs. synthetically, relying on analog vs. digital technologies, or using your brain vs. your computer. These perspectives are not necessarily entirely equivalent. For example, one may argue that thorough vs. explorative work and analytical vs. synthetic thinking apply all the same to both mental (e.g. solving a scientific problem) and computational (e.g. searching the internet for information) activities.

A key ability of good students, teachers, academic or industry researchers, and software developers has always been to find the right balance between drilling and surfing. Undoubtedly, the digital-liberal orientation of our modern society is strongly influencing our ability to switch between the two lenses of these bifocal glasses. Even the way we read has dramatically changed. The printed book, permanent and finite, is (was?) definitely an invitation to drill. The web, fluctuating and open, is certainly more of an invitation to surf. And the growing pressure on scientists and research institutions to justify their "usefulness" on a continuous basis is yet another incentive to short-term surfing for immediate exposure, as opposed to long-term drilling for deeper achievements.

On the one hand, the digital-liberal era provides scientists with an unprecedented power in terms of sources (databases and search engines), processing (machine learning and artificial intelligence), analytics (visualization and aggregation), reach (electronic publishing and media), and interactions (social networks and collaborative software). On the other hand, these amazing new extensions to the human brain may create handicaps for which they work as prostheses (substitution of brain-learning by bookmarking, reduced exposure to challenging/contradicting information, overweighting of quantity over quality, enhancement of short-termism and superficiality). As was the case for all technology leaps in the past, the balance between risk and gain will depend on how wisely we use these technologies, i.e. to which extent we – as humans and as scientists – actively manage to remain their masters, or passively drift to become their victims.

For this second Special Issue of Infozine, we have invited students, teachers, researchers, and software developers to share their opinions about one or the other aspect of this broad topic: how to balance drilling (for depth) vs. surfing (for breadth) in scientific learning, teaching, research, and software design – and how the modern digital-liberal system affects our ability to strike this balance. This special issue is meant to provide a wide and unbiased spectrum of possible viewpoints on the topic, helping readers to define lucidly their own position and information use behavior.

Citation: Hünenberger P, Renn O: Surfing versus Drilling for knowledge in science: When should you use your computer? When should you use your brain? Infozine **2018**, Special Issue 2, 2–3
DOI 10.3929/ethz-b-000298711

Published: November 15, 2018

Addendum: Blaise Pascal: Les deux infinis – The two infinities

Let man then contemplate the whole of nature in her full and lofty majesty, let him turn his gaze away from the lowly objects around him; let him see the dazzling light set like an eternal lamp to light up the universe, let him see the earth as a mere speck compared to the vast orbit described by this star, and let him marvel at finding this vast orbit itself to be no more than the tiniest point compared to that described by the stars revolving in the firmament. But if our eyes stop there, let our imagination proceed further; it will grow weary of conceiving things before nature tires of producing them. The whole visible world is only an imperceptible dot in nature's ample bosom. No idea comes near it; it is no good inflating our conceptions beyond imaginable space, we only bring forth atoms compared to the reality of things. Nature is an infinite sphere whose center is everywhere and the circumference nowhere. In the end, the greatest palpable sign of the omnipotence of God is that our imagination loses itself in thinking about it.

Let man, returning to himself, consider what he is in comparison to what exists; let him regard himself lost, and from his little dungeon, in which he finds himself lodged, I mean in the universe, let him take the earth, its realms, its cities, its houses and himself at their proper value. What is man in the infinite?

But, to offer him another prodigy equally astounding, let him look into the tiniest thing he knows. Let a mite show him in its minute body incomparably more minute parts, legs with joints, veins in its legs, blood in the veins, humors in the blood, drops in the humors, vapors in the drops: let him divide these things still further until he has exhausted his powers of imagination, and let the last thing he comes down to now be the subject of our discourse. He will perhaps think that this is the ultimate of minuteness in nature.

I want to show him a new abyss. I want to depict to him not only the visible universe, but all the conceivable immensity of nature enclosed in this miniature atom. Let him see there an infinity of universes, each with its firmament, its planets, its earth, in the same proportion to the visible world, and on that earth, animals, and finally mites, in which he will find again the same results as in the first; and finding the same thing yet again in the others without end or respite, he will be lost in such wonders, as astounding in their minuteness as the others in their amplitude. For who will not marvel that our body, a moment ago imperceptible in a universe, itself imperceptible in the bosom of the whole, should now be a colossus, a world, or rather a whole, compared with the nothingness beyond our reach.

Anyone who considers himself in this way will be terrified at himself, and, seeing his mass as given by nature, supporting him between these two abysses of infinity and nothingness, will tremble at these marvels. I believe that with his curiosity changing into wonder he will be more disposed to contemplate them in silence than investigate them with presumption.

For, after all, what is man in nature? A nothing compared to the infinite, a whole compared to the nothing, a middle point between all and nothing, infinitely remote from an understanding of the extremes; the end of things and their principles are unattainably hidden from him in impenetrable secrecy. He is equally incapable of seeing the nothingness out of which he was drawn and the infinite in which he is engulfed.

Les deux infinis (French version)

Que l'homme contemple donc la nature entière dans sa haute et pleine majesté, qu'il éloigne sa vue des objets bas qui l'environnent. Qu'il regarde cette éclatante lumière, mise comme une lampe éternelle pour éclairer l'univers, que la terre lui paraisse comme un point au prix du vaste tour que cet astre décrit et qu'il s'étonne de ce que ce vaste tour lui-même n'est qu'une pointe très délicate à l'égard de celui que les astres qui roulent dans le firmament embrassent. Mais si notre vue s'arrête là, que l'imagination passe outre; elle se lassera plutôt de concevoir, que la nature de fournir. Tout ce monde visible n'est qu'un trait imperceptible dans l'ample sein de la nature. Nulle idée n'en approche. Nous avons beau enfler nos conceptions au-delà des espaces imaginables, nous n'enfantons que des atomes, au prix de la réalité des choses. C'est une sphère dont le centre est partout, la circonférence nulle part. Enfin, c'est le plus grand caractère sensible de la toute puissance de Dieu, que notre imagination se perde dans cette pensée.

Que l'homme, étant revenu à soi, considère ce qu'il est au prix de ce qui est; qu'il se regarde comme égaré dans ce canton détourné de la nature; et que de ce petit cachot où il se trouve logé, j'entends l'univers, il apprenne à estimer la terre, les royaumes, les villes et soi-même son juste prix. Qu'est-ce qu'un homme dans l'infini ?

Mais pour lui présenter un autre prodige aussi étonnant, qu'il recherche dans ce qu'il connaît les choses les plus délicates. Qu'un ciron lui offre dans la petitesse de son corps des parties incomparablement plus petites, des jambes avec des jointures, des veines dans ces jambes, du sang dans ces veines, des humeurs dans ce sang, des gouttes dans ces humeurs, des vapeurs dans ces gouttes; que, divisant encore ces dernières choses, il épuise ses forces en ces conceptions, et que le dernier objet où il peut arriver soit maintenant celui de notre discours; il pensera peut-être que c'est là l'extrême petitesse de la nature.

Je veux lui faire voir là dedans un abîme nouveau. Je lui veux peindre non seulement l'univers visible, mais l'immensité qu'on peut concevoir de la nature, dans l'enceinte de ce raccourci d'atome. Qu'il y voie une infinité d'univers, dont chacun a son firmament, ses planètes, sa terre, en la même proportion que le monde visible; dans cette terre, des animaux, et enfin des cirons, dans lesquels il retrouvera ce que les premiers ont donné; et trouvant encore dans les autres la même chose sans fin et sans repos, qu'il se perde dans ses merveilles, aussi étonnantes dans leur petitesse que les autres par leur étendue; car qui n'admirera que notre corps, qui tantôt n'était pas perceptible dans l'univers, imperceptible lui-même dans le sein du tout, soit à présent un colosse, un monde, ou plutôt un tout, à l'égard du néant où l'on ne peut arriver ?

Qui se considérera de la sorte s'effrayera de soi-même, et, se considérant soutenu dans la masse que la nature lui a donnée, entre ces deux abîmes de l'infini et du néant, il tremblera dans la vue de ces merveilles; et je crois que sa curiosité, se changeant en admiration, il sera plus disposé à les contempler en silence qu'à les rechercher avec présomption.

Car enfin qu'est-ce que l'homme dans la nature ? Un néant à l'égard de l'infini, un tout à l'égard du néant, un milieu entre rien et tout. Infiniment éloigné de comprendre les extrêmes, la fin des choses et leur principe sont pour lui invinciblement cachés dans un secret impénétrable, également incapable de voir le néant d'où il est tiré, et l'infini où il est englouti.

The "Two infinities" has been published with the collection of Pensées (Thoughts)

Blaise Pascal (1623–1662) was a French mathematician, physicist, inventor, writer and Catholic theologian.

Antonio Loprieno
University of Basel

"Surfing" vs. "drilling" in the modern scientific world

When should you use your computer and when your brain? Ideally, always both at the same time! "Surfing" and "drilling" are two concomitant sides of scientific or scholarly activity; to distinguish them on the basis of allocated timeslots is a somewhat arbitrary, if not spurious endeavor. But although "surfing" and "drilling" are two equally important aspects of science practice, they do require rather different investigative qualities: while "surfing" – whether with a computer on the web, with a catalogue in the library or with a surfboard on the ocean's waves – requires juggling competences, experienced balancing and synthetic judgment, "drilling" – whether physically with a drill or metaphorically with your brain – demands meticulous delving into the meanders of analytic investigation, which are more often than not very muddy. So, here is a first generalization: surfing is synthetic, drilling is analytical.

But be careful: neither does this generalization imply that surfing is a more superficial activity than drilling, nor does this imply that drilling into a scientific issue is tantamount to displaying only visionless resolve. Neither of the two procedures is easier or more profane than the other: thus, drawing a distinction between "using the computer" and "using the brain" is very debatable, since we may tend to assume that "using the brain" is a more accomplished enterprise than "using the computer", and thus should be preferred to it. This is not the case: we always need both the synthetic and the analytic mode, although some of us are stronger in one, some in the other.

Some researchers are better at writing a handbook, at offering an overview class, or at developing state-of-the-art presentations, which are all prototypical "surfing" enterprises. Other scientists are better at setting a new research agenda, at discovering a hitherto neglected piece of evidence, or at recognizing the weak points in the colleague's latest paper, all of which are "drilling" endeavors. We may individually prefer one of them, but we always need both.

What should guide us in finding the right dosage between synthetic and analytic procedures is what is now called "critical thinking" – which is, very appropriately, one of the strategic goals of the ETH Zurich. Ideally, we should surf with the drill in our mind, ready to use it relentlessly whenever we are not satisfied with the received information or whenever we judge that there is room for improvement in the accepted state of the art. But at the same time, while holding the drill in our hands, we should also keep the surfboard tight under our feet and float on the recurrent waves, which may assume different shapes: expectations from our students, political or societal regulations, scientists working on the same topic, etc.

This constant search for equilibrium between surfing and drilling is a common trait to all sciences and scholarly activities. So, whenever you are told that sciences and humanities are separated by a different culture (a misconception shared by many natural and social scientists alike), don't believe it. The real difference is not between the alleged "culture" of physicists vs. historians or life scientists vs. cultural scholars, but the one between good and bad science altogether. We are all in need of a constant interaction of "surfing", by trying to look at the general picture, and "drilling", by trying to get deeper into the mystery surrounding any scientific enterprise. Rather than a choice of principle between "surfing" and "drilling", or between "sciences" and "humanities", the challenge we are confronted with is the choice of the most appropriate methodological tools in order to deal with a specific research object. In mathematics, one privileges the "axiomatic" approach, in which cogent scientific proof is provided by the internal, self-referential coherence of the argument. In most natural and physical sciences, on the contrary, evidence is treated in an "empirical" way, i.e., it is derived on the basis of validation provided though experiments. On the other hand, in most humanities, where for various reasons (historical, conceptual, individual, etc.), scientific evidence is not experimentally verifiable, the privileged approach is "hermeneutic", i.e., it is based on the need to make intellectual sense out of frequently inconsistent bits of information. Thus, inevitably, because of the very nature of the evidence, humanists will tend to cherish "surfing", for example by referring more frequently to other scholars' ideas, whereas scientists will privilege "drilling", e.g. by concentrating on ever more minute fragments of information. There is no binary opposition involved here, but a smooth continuum steered by the individual – or institutional – research agenda.

Another aspect of the continuum linking surfing to drilling that deserves attention here is the relationship between research as production and research as transmission of knowledge. In this context, teaching and research are not two radically opposite endeavors, but rather two aspects of the same scientific enterprise. It is impossible to produce new knowledge without intense "drilling" into the most

detailed facets of a problem. But it is equally impossible to transmit scientifically derived knowledge (whether to a colleague, to a post-doc or to a student) without "surfing" it, by framing the issue at stake into the bigger picture. In fact, it is precisely this subtle capacity to understand the context of scientific results that makes out what we refer to as "critical thinking", which is the quality that allows us to discriminate between real science and pseudo-science, between plausible and implausible assumptions. Especially in our current post-factual climate, in which science seems to be challenged by aggressive populist fabrications that jeopardize its primate in society, a critical contextualization of scientific results may turn out to be the most crucial contribution that we as scientists can offer to the society in which we are actively embedded.

Prof. Dr. Dr. hc.
Antonio Loprieno
History of Institutions
Faculty of Business and Economics
University of Basel
Peter Merian-Weg 6
4052 Basel, Switzerland
Phone +41 61 207 33 31/+41 61 207 29 89
a.loprieno@unibas.ch
President, Swiss Academies of Arts and Sciences, Laupenstrasse 7
3001 Bern, Switzerland
Phone: +41 31 306 92 30
antonio.loprieno@akademien-schweiz.ch
ORCID 0000-0001-5152-8815

Citation: Loprieno A: "Surfing" vs. "drilling" in the modern scientific world. Infozine **2018**, Special Issue 2, 4–5
DOI 10.3929/ethz-b-000297316

Published: November 15, 2018

Philippe Hünenberger
ETH Zurich

Of millimeter paper and machine learning

My father used to be a gymnasium teacher. When he was not giving class, he loved to work in the cafés of our home-town, preparing his courses or correcting his copies. Sometimes, I was allowed to join him and then, I would ask him for "a function". In return, I would get a sheet of millimeter paper, his HP35 pocket calculator (with the fascinating "reverse Polish notation"), and … a mathematical expression returning a value of y for a given value of x. And off to work I was, calculating and painstakingly reporting dots on the paper to make a graph. This took hours, gratefully saved by my father for working quietly on his copies. And it also afforded me some surprises. For example, the function $y=\sqrt{(100-x^2)}$ only gave the upper half of a circle, not the lower one, and the HP calculator stubbornly refused to return a y for any x above 10 or below -10. Of course, my father would ultimately clarify these issues with me, but not before I had had time (literally hours!) to brood over them on my own ...

This was clearly drilling – deep and narrow in scope over long open-ended time stretches: a lonely, tedious and time-consuming task relying on manual exploration, intellectual processing, critical questioning and iterative experimenting. Besides a solid reputation of nerd-kid in the cafés, this approach rewarded me with a truly emotional respect for mathematical functions (they became life-long "friends") and an understanding that drilling can underpin thrilling subsequent discoveries (e.g., for the circle, complex numbers).

A few years later, I received my first computer, an Apple IIe. Now, given ten minutes or so of programming, I could plot any function on the screen, i.e. achieve almost instantaneously what had taken me a full morning a few years before. I was ready to explore a mysterious new world, that of mathematical functions: cardioids, epicycloids, astroids, Lissajous curves, Cornu spirals, … I could find new ideas in books, ask my father or other teachers, or simply try at random – I could even "play" the program as a game with friends.

I was now engaged in surfing – broad and shallow in scope over multiple short segments of time: the fast-paced and playful consideration of possibilities, relying on interactive exploration and comparatively superficial observations. This new approach rewarded me with a feeling for the richness of mathematical functions (they became a new "universe" to discover) and also with more visible achievements (even "marketable" ones, as I could earn a high-school prize for a program rotating the five Platonic solids on screen). But the mere surfing left me a bit frustrated, at least as long as I did not complement it by subsequent re-drilling on one function or the other. Just like a tour of 10 capitals of Europe in 10 days leaves you wishing you could spend afterwards 10 days in each of them separately.

When I started my PhD in Theoretical Chemistry in 1992, the research job matched my expectations: a back-and-forth oscillation (zoom in, zoom out) in Pascal's "double infinity", alternating surfing for breadth and drilling for depth, with a largely self-determined alternation schedule. Computers were useful tools, data was at the service of science, and the e-mail and internet were convenient devices for targeted and asynchronous communication. Since then, and especially over the last decade, things have changed a lot.

Undoubtedly, modern digital tools represent a fantastic extension of the human brain in terms of data access, processing throughput and communication reach. But in addition to that, they have also become overly invasive companions. Data seems to no longer be at the service of science, rather the opposite. The e-mail and internet, reinforced by an army of surveys, newsletters, mailing lists, evaluation tools and social networks, has evolved into an overflowing stream of information and an inexhaustible source of interrupts. In this noisy digital world, short-term surfing activities seem to take most of the space, while the quiet drilling activities have become a luxury. This may just be an exaggerated swing of the pendulum, triggered by the relative novelty of digital technologies. And the pendulum could certainly return to a more comfortable position provided that individual researchers and research managers both take the challenge seriously – and use their (human!) brains to control the present evolution. To this purpose, I have listed below three propositions.

Causality is stronger than correlation

Proposition one (epistemological): Causality-based models (from drilling) should be credited with a higher intrinsic value than correlation-based models (from surfing), irrespective of their relative current predictive powers for specific applications.

This judgment of value is not obvious to defend in times where big-data correlations relying on machine learning (ML) and artificial intelligence (AI) are becoming increasingly predictive, often more predictive nowadays than causal models based on elementary physical principles and numerical computations. Imagine a village where the "ancient" would predict the yield of the upcoming crop with 80% confidence based on rational thinking and a deep knowledge of the climate, plants and insects, but the "idiot" would do the same with 95% confidence by

correlating intuitively a large number of more or less relevant observations in an entirely unknown fashion. Wouldn't it make sense to call the method of the "idiot" a "new paradigm" – and then maybe just kill the "ancient" to save food? Bad idea in my opinion … for at least three reasons.

First, comparing current predictive powers for specific applications is short sighted. As long as their physical Ansätze are correct, causal models have the potential of becoming fully predictive over all applications. The bottleneck is in their numerical evaluation, bounded by the current computing power. In contrast, correlation models are intrinsically limited in scope by the selection of a training set and in accuracy by the selection of input observables, irrespective of the available computing power.

Second, causal models have explicit Ansätze which are systematically improvable (e.g. Newtonian to quantum/relativistic ones), and their processing is amenable to human understanding and supervision. In contrast, the "Ansätze" of a correlation model are non-transparent, buried in the selection of training set and input observables. These may suffer from many biases, including proxy effects (if B is similar to A, then B must behave like A; punishment of the exception), assumed causality (if A correlates with B, then changing A will change B; action on a symptom), and design bias (voluntary or involuntary tuning of the training set and observables to get results matching prior expectations). In addition, the data throughput of correlation-based models is typically so high that they are beyond human supervision, and the coupling of the output of such a model to its input (e.g. funding of scientists made in proportion to their publication metrics) may create pernicious feedback loops. As a result, the uncritical use of large-scale correlation models tends to discourage serendipitous discoveries and promote self-fulfilling prophecies.

Third, only causal models represent what one should call knowledge in a humanistic perspective. In this sense, the terms ML and AI are misleading. Computers don't "learn" and they are not "intelligent". These are human characteristics, implying far more than correlation-picking (e.g. critical and orthogonal thinking, creativity, ethical accountability, emotional and social intelligence, …). Following Plato, I am convinced that knowledge is about finding what causes the shadows on the wall of the cavern, not merely about predicting patterns of motions in these shadows.

Surfing should be viewed as an extension to drilling

Proposition two (scaling-up): Surfing should be viewed as an extension to drilling, i.e. procedural understanding should precede automated application; this holds not only for scientific research, but also for learning, teaching, and education in general.

You don't give toddlers a Porsche to explore the city traffic. First, over many years, they learn how to crawl, then walk, then bike, and then drive (and then they can start saving for the Porsche!). Along the way, they progressively refine their procedural competences and feeling for danger, in parallel to scaling-up in terms of locomotion reach and speed. They also learn to distinguish between what does not need human thinking (and can thus be automated) and what definitely does (and therefore implies full brain awareness). Why should we do it differently with computers? In the context of teaching, this is what I wished to illustrate with my explorations of mathematical functions: first millimeter paper, then hacking a curve-drawing program, then surfing the function space.

You will often hear that computer-assisted techniques should be introduced as early as possible in teaching (this hype clearly extends up to the university level and beyond). The usual arguments sound like: (1) surfing is playful and interactive, thus likely to promote curiosity; (2) digital supports can be adjusted to individual learning curves; (3) this will prepare the child/student for a world where digital tools play a central role. None of the above arguments convinces me, because: (1) the type of "curiosity" induced by playful surfing is superficial and short-lived (nothing like the deep and long-lasting thirst one calls scientific curiosity); (2) creating an artificial world that adjusts miraculously to a person's needs actually impairs the development of adaptation skills (very unfortunate considering that neuroplasticity will be a key asset in the upcoming job market); (3) computer-surfing skills are relatively easy to learn if you have brain-drilling skills (but the opposite is definitely much harder).

A key pedagogical element in teaching is to trigger a curiosity-based itching for the next level of abstraction or throughput, thereby motivating the usefulness/necessity of this next level. Just as one should teach Chemistry starting from experimental observations and promoting an itching for the theoretical model explaining them, one should teach computer skills starting from step-by-step procedures and inducing a similar itching for the automation of the repetitive steps. Importantly, this scaling-up ensures that the assumptions and shortcomings of the modeling/automation procedure are evidenced explicitly, so that critical thinking is fully preserved in a subsequent faster-paced surfing phase.

Besides this scaling-up pedagogy, an open and respectful teaching atmosphere in the classroom, the promotion of critical and creative thinking, a thorough preparation, and an exemplary role of the teacher – which was already in essence the good old teaching recipe of my father – I am not sure there is so much to gain by introducing too many "innovations" in teaching, and especially not digital ones.

Similar considerations apply to research. Clearly, the modern scientific world is too complex and multi-faceted for anyone to know every technique in entire depth at any time. This is not even desirable. We all rely on a number of black boxes, i.e. procedural components (theories, models, algorithms, equations, software, data, …) for which we know the input and output, but not the inner workings down to the last details. The key question is rather about the extent of ignorance we are willing to tolerate when using a black box, without running the risk of being "fooled" by it. This limit is crossed as soon as we are no longer able to assess based on our own knowledge and thinking whether the black box is working correctly or

not. An education that has involved an explicit scaling-up in the construction of a number of "standard" black boxes is definitely an asset for performing these types of assessments. The more we lazily skip to the surfing without spending effort onto the preliminary drilling (both in education and in research), the more our society will consider computers as wizards or oracles rather than tools.

Finding the balance between drilling and surfing is a major challenge nowadays

Proposition three (management of resources): Striking the appropriate balance and schedule between drilling and surfing in terms of allocated time, means and rewards is a major challenge nowadays; wise choices in this regard are of extreme importance for the long-term success of the scientific endeavor.

Individual researchers (in particular group leaders) could easily fill their agendas with surfing activities, leaving little room for drilling ones. Digital tools are not the direct cause for this, but an aggravating factor. This is because they allow a massive flow of information and requests to reach us on a quasi-instantaneous basis from all over the world, and because they represent a permanent invitation to inefficient reactive processing (ping-pong) and multitasking habits, themselves again contributing to increasing the digital flow. Yet, I am convinced that most researchers possess in principle the necessary skills and wisdom to strike the balance on their own, with "protection" tricks including: ignoring or declining most requests, delegating tasks, batching on-line periods, agreeing on communication policies, practicing temporary unreachability, and … being "sloppy" when something does not matter.

However, it is not clear how much they still have the freedom to do so in practice, considering the raise of two phenomena at the research-management level: the wish to increase the apparent productivity and immediate visibility of research, and the wish to reinforce its top-down steering. Both result in an increasing pressure on researchers to enhance what is considered to be their efficiency (the "ratio of research output to taxpayer franc", a nice expression I read recently in the NZZ) and quantifiable impact (university rankings, publication numbers and related metrics), and to work in directions that are imposed from the top based on immediate societal relevance and fashion trends (strategic goals, dedicated funding). Surfing activities tend to be more extravert, interactive, drivable, fast-paced and visible. Thus, they are more easily steered, quantified, recognized, financed and rewarded. Drilling activities, on the other hand, are typically introvert, slow, quiet and self-driven, and their effect on research quality is only visible in the long term. As a result, with a top-down management towards productivity and visibility, drilling becomes associated with a negative connotation of unproductive off time. This leads to an unhealthy tendency to minimize these activities or shift them into recovery time, as if they were no longer part of the job.

Fundamental research in an academic environment should be in first priority rigorous and creative, and only in second priority productive and visible. Historically, the production of the most efficient things (fundamental discoveries) has often been a rather inefficient process (trial-and-error, persistent work, well interpreted failures and … a bit of luck). In a society obsessed by efficiency, one should thus think carefully whether one wishes the research process to look efficient, or the research outcome to be efficient. If the latter is desired, the current management trends should be opposed, i.e. one should reinforce the trust in individual researchers.

The three above propositions are only invitations to your own thinking, a few personal suggestions for putting a new value on drilling in a world that is a bit too crazy about surfing. Maybe this thinking can help to avoid a possible future where data is the new currency and algorithms are the new priests, and in which technocrats drive the world based on curves from machine-learning, without ever having themselves put a single dot on a sheet of millimeter paper.

Prof. Dr. Philippe Hünenberger
ETH Zürich
Laboratory of Physical Chemistry
HCI G233
8093 Zurich
Switzerland
Phone +41 44 632 5503
phil@igc.phys.chem.ethz.ch
ORCID 0000-0002-9420-7998
http://www.csms.ethz.ch

Citation: Hünenberger P: Of millimeter paper and machine learning. Infozine **2018**, Special Issue 2, 6–8.
DOI 10.3929/ethz-b-000294364

Published: November 15, 2018

Janne Soetbeer
ETH Zurich

From one to many, from breadth to depth – industrializing research

The objective of science, including research, remained unchanged during the last centuries: as much as Newton was driven to comprehend his surrounding, my own research is driven by the same curiosity. What has evidently changed over time, is the setting in which we carry out research.

Newton, Mendel, Einstein, and more recently Higgs, all generated groundbreaking science, single-handedly. Nowadays, collaborations and interdisciplinary efforts shape the scientific activities of the 21th century. This change is due to two factors. First, our scientific knowledge has become increasingly more complex and detailed. Second, we rely on ever more advanced and costly instrumentation. Both factors promote specialists over generalists, and in this respect we live in an age of drilling. The real deep drilling, i.e. fundamental research, remains a privilege of academia. Despite increasing pressure toward being useful, academia should defend its right for curiosity- and not solely need-driven research. After all, fundamental research provides an unpredictable source for potentially disruptive innovation beyond what the general public may identify as a need.

If the way of mining these innovations has changed, what is the impact on the individual miner and the field at large? The scientific genius is already said to be extinct [1], because no new disciplines are founded and current ones are not revolutionized by an individual, the genius, anymore. Instead, hybrids of existing disciplines emerge and collaborative teams tend to produce the cutting-edge research of our time [1]. To put it bluntly, we have industrialized research through division of labor. The web facilitated this development as the essential infrastructure for efficient information exchange. However, all this comes at a cost. Specialized scientists require specialized training. Thus, as more time is invested into education, the individual can contribute to the body of knowledge substantially later in life than a century ago [2]. The presented parallels to the industrial revolution, should not allude to the scientist as an assembly-line worker, but instead point out risks associated with specialization. In its essence, an expert sacrifices his or her cognitive diversity, to boost the overall productivity, i.e. scientific advancement. While the individual may lose the broader perspective, the field can be more diversely mined. Therefore, the balance between digging and surfing has to be assessed at least at two levels.

As the industrial revolution profoundly changed the rhythm of life in the 19th century, the web has strongly accelerated the publishing process. As such, the web is the steam engine of today's science. Notably, the web's dynamic format supports the nodal structure of knowledge more naturally than the static and linear book does. However, as with any technological progress, critical voices warn about the deteriorating effect on humankind. This being, skimming instead of reading and short attention spans of up to 280 characters, making us shallow individuals. On the other hand, these practices are perhaps merely a necessity to cope with the large amount of information available. Skimming is superior to reading, when looking up factual information. Research instead steps beyond the current body of knowledge, and hence inherently requires depth. As a researcher, I find myself exercising both, best illustrated during literature research. I skim to identify key pieces, but read the relevant papers thoroughly. At the interface of breadth and depth, I feel the two opposing forces. And even stronger, when communicating my research to a layman.

Research in natural sciences has evolved from an individual's effort to a collective's effort of specialists. The complexity of science required, the web infrastructure enabled this trend. As specialists, we should avoid to be blinkered, and instead maintain a generalist's perspective. Sampling ideas outside of the own expert sphere fosters creativity and innovation. This level of breadth should not be limited to one as a receiver, but also as a sender.

References

[1] Simonton DK: Nature **2013**, 493, 602.
[2] Jones BF: Rev Econ Stat **2010**, 92, 1–14.

Janne Soetbeer
Doctoral Student
ETH Zurich
Laboratory of Physical Chemistry
HCI F 238
8093 Zurich
Switzerland
janne.soetbeer@phys.chem.ethz.ch
ORCID 0000-0003-0008-3494

Citation: Soetbeer J: From one to many, from breadth to depth – industrializing research. Infozine 2018, Special Issue 2, 9
DOI 10.3929/ethz-b-000294365

Published: November 15, 2018

Gerd Folkers[1] and Laura Folkers[2]
[1]ETH Zurich, [2]Lunds Universitet

"Deep drilling" requires "surfing"

Deep drilling is a very attractive term. Apart from geological discussions, we heard it for the first time in a conversation with Gottfried Boehm, who coined the term "*iconic turn*". The debate was about the interpretation of a photographic image that depicted the situation in an operating theatre. This was quick and easy to interpret superficially (surfing). But the question was: Did the photographer want to present a very critical situation or moment? Or was the photograph just a random snapshot? How can you decide or know if the photograph has a particular message? Does the expression of the surgeon, the assistant or the anesthetist reveal something? Is there a special type or constellation of medical equipment depicted? Can one read and conclude anything from the displays of the monitoring instruments that are visible in the photograph?

In this example, the surface of the image does no longer provide any obvious information, probably except to the experts, the surgeon and his team. Finding the answers, the possibly additional information, requires deep drilling at specific locations on the image's surface. This immediately raises a new question. Where to drill? Some points are evident, such as the surgeon. If the drilling provides information about his specialty, we can at least make a guess. A liver specialist will intervene less frequently in an emergency situation, as maybe will the urologist. The more detailed analyses of the monitoring instruments or the apparatuses provides less information, because often the same parameters are measured, the same constellation of technical aids are used in an operating theater. Thus, the number of questions increases: Why have which parameters been measured? Who determines the parameters by which a normal situation is distinguished from a critical situation? The decision-making processes associated with these analyses are anything but trivial. The frequent outcome of going deeper is the accumulation of more questions.

Of course, the patient himself is the most promising person for a deep drilling. However, the information about him is securely protected, so that the chisel ends up on photograph's negative.

Not visible of course, but present, is the photographer. This is where a deep well is most likely to be found. What was to be shown, what was the motivation for taking the photograph? Public relations for the clinic, or the description of a working environment, or showing a prominent surgeon or a prominent patient?

The deeper the driller's hole goes, the more information is disclosed. Not all information is equally useful in answering the question: "What does this image want to tell us?". The assistant's mean blood pressure is probably irrelevant, as is the anesthetist's cat's name. The photographer's shoe size is probably irrelevant, but the chemical composition of the photographic paper is not.

Therefore, an important perspective has to be added, that of the (presumed) relevance. A discussion of relevance relations prevents a too deep drilling, which is too often seen, probably due to the widespread conviction that the "truth" can always be found at the lowest level. This unfortunately applies neither to images nor to oil.

Subatomic states can nowadays be measured, simulated and constructed, and there is undoubtedly a connection between them and the photograph. However, knowledge of these subatomic states is not necessary for an interpretation of the image. Hence, deep drilling should explore the level of granularity, needed for interpretation of what can be seen on the surface. The aforementioned is explored by surfing. Only a certain distance, surfing above the surface will grant you with the detection of "hot spots", pictorial elements with putative importance for understanding the whole, hinting at promising drilling points. Like in aerial archeology.

The very same question about the relevance applies to teaching.

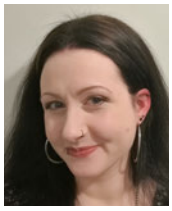
It is only natural that the further one advances with one's education, the fewer and the more involved the topics become. That would be drilling deep. This is good and useful, yet those responsible for Higher Education keep constantly trying to prevent their students from becoming what in German may be called a "*Fachidiot*".

Therefore, curricula are getting crammed with more and more topics – and most of them are still fitting into the general field of the particular study. However, more and more additional topics like ethics appear on the agenda. Ethics is important beyond doubt, but is it to the extent of cutting courses shorter which actually belong to the study field and enable students to drill themselves? Those measures foster surfing in two aspects. One is the fact that the time available to become excellent in either field of specialization is reduced, the other is the – wrong – belief that ethics can be taught in a couple of hours. By that, ethics itself becomes subject to surfing, while it should be an attitude, not an examination topic. Thus, make ethics an integral part of a scientist by serving always as a role model and don't award credits points for ethics.

In a similar fashion, one could wonder about the benefits of keeping the curriculum topics broad – even in the master year of the study program. In case of the chemistry programs at the ETH Zurich, there are three major fields

– organic, inorganic and physical chemistry –, which are all continued until the end. On one hand, this is brilliant as it gives all graduates an overview over all those topics that reaches further than named reactions, batteries and some selection rules. On the other side, the proper "granularity" provides the understanding of the whole image. Subsequent surfing donates the pleasure to identify "hot spots" for individual deep drilling. It is the challenge (time and space) to get the course granularity fine enough for advanced students to identify their special field where they want to drill deep with all their enthusiasm.

Prof. (em.) Dr. Gerd Folkers
ETH Zurich
Department of Humanities, Social and Political Sciences (D-GESS)
8092 Zurich, Switzerland
Phone +41 44 633 87 07
gerd.folkers@gess.ethz.ch
ORCID 0000-0002-3620-705X



Laura Folkers
MSc Chemistry (ETH)
PhD Student Inorganic Chemistry
Lunds Universitet
Sweden
laura.folkers@chem.lu.se
ORCID 0000-0002-3424-1932

Citation: Folkers G, Folkers L: "Deep drilling" requires "surfing". Infozine **2018**, Special Issue 2, 10–11
DOI 10.3929/ethz-b-000296242

Published: November 15, 2018

Alžbeta Kubincová
ETH Zurich

Surfing vs. drilling in science: A delicate balance

Imagine there was once a group of people assembled to build a pyramid. After locating a good spot, time came to decide on how to proceed. Soon, the construction workers grouped themselves into two opposing teams: the first one pleaded for starting with the construction right away in order to finish the work as quickly as possible, and the other one proposed to first carefully examine the foundations to make sure they are sufficiently stable to hold the construct. However, an agreement was not in sight and eventually, they started working at the same time. And so, they became a source of uneasiness to each other, for every dig may cause the collapse of the whole pyramid, as well as every stone laid on its top will increase the damage once this happens. This caused the two teams to grow more and more apart.

We have arrived in the 21st century. The pyramid has reached a far greater height than any single person could build in a lifetime. Indeed, science has made the transition from the work of a few scholars in isolation to a *collective* activity. It is no longer sufficient to set up experiments in a systematic way and establish sound theories to explain their outcome. Now, one has to build on the mountain of work that has been performed before – Newton, Maxwell and many others had to prepare the floor before Einstein could make an entrance and revolutionize physics. A good overview of the state of the art in the respective field is therefore of crucial importance. It is clear that a mix of both *breadth* and *depth* is necessary to score high in research these days, but where is the right balance? And how far away from it are we now?

A personal approach

First, let us build up the pyramid bottom-up: although research is collective in its nature, its elementary units are still the scientists involved. And due to their differing characters, most of them will display a preference for the one or the other direction, which can be used as a criterion to split them into two groups – let's call them the *surfers* and the *drillers*.

The *surfers* are usually busy people, who are very skilled at skimming dozens of papers in the shortest time and filtering out those which are not meaningful or sufficiently relevant. In addition, scientific podcasts and conferences pave them a way to the top of the pyramid. Being able to quickly internalize new concepts and approaches, they gain a great overview over the state-of-the-art methods, which have the potential to be assembled into even greater projects and workflows. Having a good overview over a certain subject also facilitates the communication between scientists in neighboring fields, as well as the propagation of their research over the boundary of their sole team in an easily digestible way. For the same reasons, they also tend to be good teachers, who excel at embedding new concepts into the framework of the students' previous knowledge.

The *drillers,* on the other hand, are driven by a desire to learn and gain insight, rather than focusing on the immediate applicability. Reading papers, they would first have a look at the theory and method sections, striving to understand the approach in detail and what it builds upon. They tend to be devoted and persistent workers, and their intense focus on one particular field allows them to be more critical towards established and widely used methods. Applying the same critical attitude towards their own results adds to their reliability. Most of the scientific breakthroughs can be attributed to this type of people, who had the courage and determination to undermine existing foundations and replace them by more robust ones, allowing for the construction of a much higher building than ever before.

Of course, those are only the two extremes of a continuous spectrum, though many scientists seem to display a preference for the one or the other side. The problem is that the different working habits cause the two groups to phase-separate: Information which is easily accessible to the one type will be poorly digestible for the other. Therefore, these differences are prone to be misinterpreted as shortcomings – the *drillers* hence tend to perceive the *surfers* as superficial and over-selling, while they are judged by the latter as narrow-minded and impractical.

Research landscapes

Apart from the bias on a personal level, it is also important to consider the research field of interest. Chemistry, for instance, has an interface to virtually every natural science – even the two research groups concerned with molecular dynamics simulations at D-CHAB (Department of Chemistry and Applied Biosciences), one which I belong to, while being both on the theoretical side of the spectrum, have collaborations reaching into organic chemistry, NMR spectroscopy and nanoscience.

Similarly, when I first informed myself on the web about the chemistry study program at ETH Zurich, it was stated that students should bring in a broad range of interests – a

statement which clearly encourages *surfers* to join the field.

A counter-pole to chemistry, attracting a high percentage of *drillers*, would be mathematics, which is the foundation of everything, and yet is built on a small number of axioms. Working out proofs is a very lonely and lengthy activity, and their communication to the greater community may be very challenging. A rather extreme example of this principle is the Japanese mathematician Shinichi Mochizuki, who claimed to have proven the so-called *abc conjecture* in 2012, which is a statement about the distinct prime factors of a sum of integers. He did this by publishing a 500-pages document on his website after 10 years of work. To this day, the community has not come to an agreement regarding the correctness of the proof, as only few of the experts in the field claim to understand it.

Minorities in danger of extinction

Combining these two influences, one may now spot the characteristics of a self-reinforcing system: An excessive need for the one or the other type of work will cause the targeting of this specific group, which will hire further workforce of the same kind, as similar qualities and attitudes make the collaboration much easier. However, doing so only brings us closer to an academic monoculture, where the other group ends up being underrepresented. In this context, the high demand for interdisciplinarity and applicability might have scared many *drillers* away from chemistry (and from most applied sciences). For instance, I don't dispute the fact that the use of big data and machine-learning models may prove useful to obtain good predictions of the properties and reactivity of molecules. But suppose we had had these tools (and the required computing power) hundred years ago, and we had invested the same percentage of funds in their development at that time as we do now: how far would disciplines such as quantum chemistry and spectroscopy have progressed to this day? There is still much ground to uncover beyond what can be extrapolated from the current data – but to get forward in this direction, we would need to increase the diversity of our academic staff, to make sure that everyone can contribute in a way which exploits their potential to the fullest.

Alžbeta Kubincová
Doctoral Student
ETH Zürich
Laboratory of Physical Chemistry
HCI G227
8093 Zurich
Switzerland
Phone +41 44 633 45 93
alzbeta.kubincova@phys.chem.ethz.ch
http://www.csms.ethz.ch

Citation: Kubincová A: Surfing vs. drilling in science: A delicate balance. Infozine **2018**, Special Issue 2, 12–13
DOI 10.3929/ethz-b-000296253

Published: November 15, 2018

Leif-Thore Deck
ETH Zurich

Digital trends in academia – for the sake of critical thinking or comfort?

Despite the fast pace at which the world is changing nowadays, people tend to overestimate the influence of new trends. Since its birth, academia has undergone steady changes, and the impact of digital technologies will be a revolution neither in teaching nor in research, but only one single step in a long-term evolution. Let's go back a few centuries to the origins …

In the Renaissance, extraordinary people like Leonardo da Vinci or Galileo Galilei shaped the ideal of brilliant geniuses with universal knowledge in all existing fields of science and humanities. This idea – that being brilliant means to be an expert in nearly all fields at the same time – was followed for a few centuries; accordingly, philosophy was the most studied field at that time. A later example of such a universal genius is the 18th century German poet and writer Johann Wolfgang von Goethe. Although he contributed more to German literature than most of his contemporaries, he also actively carried out research in diverse fields, and thought that his development of a theory of colors (later to be falsified!) was his greatest achievement.

When mankind's accumulated knowledge became too complex for one person to master, specializations started to develop towards the main disciplines as we know them today; the first university programs for Chemistry were created in the 1850s, for Chemical Engineering in the 1890s. Since then, countless other fields arose and, in 2018, there are about 19,000 distinct degree programs at universities only in Germany.

The novel digital technologies that have emerged in the last few decades enable both students and researchers to gather relevant data from many fields, to read published articles online, and to communicate with peers all over the world. Along with further specialization, a new concept of generalism has also appeared, as people are also needed to connect the specialists of all these distinct fields. With the development of the internet, of advanced online encyclopedias and of powerful searching tools, knowledge has become ubiquitous and hence, has lost some of its value. Searching for information about chemistry with modern search engines is faster and more comfortable than pursuing studies for several years. As a result, in academic teaching, a shift from naïve memorizing to deep understanding and critical thinking is inevitable. Otherwise, digital assistants like Alexa will replace all ETH Zurich graduates in the near future ...

But as large as the changes in teaching have been over the last decades, as misguided and arbitrary they were as well. Whereas books and physical attendance of students at lectures were the fundament of teaching for centuries, many of today's students focus on some kind of online material and watch streamed lectures or YouTube videos on the web. The age of books as a tool for studying appears to be past; many students do not even borrow a single one during their entire Bachelor studies. Academia as a whole evolves towards a teaching philosophy that does not involve books anymore.

Still, most lecturers provide scripts paraphrasing the lecture's content and containing all the "must-knows" for the exam, albeit with an immense variance regarding scope and quality. Exercises and their solutions are usually uploaded on a magnificent potpourri of webpages, and the lecture itself is often streamed, so that it can later be watched online – the latter at least for the larger courses in the Bachelor programs. One might argue that these developments render the attendance of the lectures in person unnecessary and, to a certain degree, this seems correct.

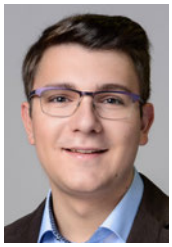
All these trends are caused by digital technologies; but instead of focusing the student (and lecturer) on critical thinking, they are ultimately rather beneficial to their personal comfort and to the economy of resources. Let's have a close look at one characteristic example: the former first-year Biology course for Chemists and Chemical Engineers. All lectures were streamed. For most topics, there were clearly stated learning goals along with voluntary multiple-choice tests and links to digital sources providing extra information; there were also old examinations on the VCS webpage (VCS = Federation of Chemistry Students). All this material – as useful and exemplary as it may be – mainly led to one thing: The lowest average attendance of all first-year lectures, with sometimes fewer than 20 students (of 150!). With this overflow of digital tools, many students did not feel there was any benefit in attending the lecture. It was more comfortable to sleep longer – the lecture started early in the morning – and to study the content later, instead of discussing questions with colleagues or with the lecturer.

This clearly shows that digital technologies have to be used with care. They can definitely support both lecturers and students in learning and teaching, but may also have negative side effects. To take this into account, we have to think about the added value of a lecture for the students who attend it. This is a highly subjective question, but, in

my opinion, there is one main component: The direct communication between the students and the lecturer, connected with the possibility for the students to formulate and discuss questions, and with the need for the lecturer to react to these questions; and thereby simultaneously receive feedback for improving the lecture continuously.

Ideally, digital technologies should support the lecturer in transmitting her/his knowledge and skills to the students while improving or, at least, not mitigating the communication between them. A highly elaborated script containing all the relevant content of the class might actually work against this goal. On the other hand, a link to a YouTube video of another professor explaining the same topic with a different approach might be more beneficial – as it can broaden the students' horizon and encourage them to critically evaluate and compare both approaches. As a guideline, novel technologies should not be included in teaching only because of their availability, but based on a detailed pro-con analysis – and with concrete benefits in mind (other than merely sleeping longer in the morning!).

All in all, the digital revolution and the associated novel technologies will definitely shape academia in a new way. But, from my point of view, it is not mainly a question of surfing versus drilling in the available "big data", but rather of critical thinking versus comfort. The biggest mistake we can make right now is not to miss some fancy recent developments; the biggest mistake would be to adopt all these emerging trends without carefully assessing their relevance, benefits and side-effects.



Leif-Thore Deck
Chemical and Bioengineering Master Student
ETH Zurich
Peter-Debye-Weg 13
8049 Zurich
Switzerland
deckl@ethz.ch

Citation: Deck LT: Digital trends in academia – for the sake of critical thinking or comfort? Infozine **2018**, Special Issue 2, 14–15
DOI 10.3929/ethz-b-000296256

Published: November 15, 2018

Yi Zheng
ETH Zurich

I diagnose, therefore I am a Doctor?

Will drilling computer software replace human doctors in the future?

In August 2016, IBM Watson, an artificial intelligence (AI)-based software system of the company IBM, corrected a misdiagnosis made by numerous dermatology experts of the Medical Institute of the University of Tokyo. Watson – drilling information rather than surfing – compared the genome of the patient with millions of genome data and recognized correctly that she was suffering from an extremely rare kind of leukemia. Thanks to Watson, the doctors were able to treat the patient appropriately and saved her life.

Ever since, more and more applications using AI to solve medical problems have appeared. Famous examples of this spectacularly rapid progress are programs which can interpret MR images or the AI-based network convolutional neural network which can diagnose several forms of skin cancer even better than experienced oncologists could. Due to the recent developments, the question arises if future software will not only be able to support doctors but even diagnose and propose therapies – thus substituting human doctors. As a consequence, some people are asking why one should still educate and train human doctors when computers would be faster, more precise and also cheaper.

In this essay, I am going to discuss this hot topic and illustrate why human doctors will still be needed. Furthermore, I am going to suggest how the technical progress will change requirements for the medical education and training in the future.

The development of medical AI application is advancing daily. It seems unavoidable that, sooner or later, computers will be superior to human doctors in every single aspect. However, this is a deception as there are several aspects in which computers cannot replace human doctors.

Firstly, even though computers are able to process many data easily and precisely, this is limited to data that can be quantified and digitalized, such as imaging, laboratory measurements or genomics. For a diagnosis and the choice of therapy, however, it is as important to consider aspects like the patient's appearance, behavior, individual history etc. At least so far, we are not expecting to have software that can gather and process all the quantitative and qualitative information in the foreseeable future. For this, human brains will still be needed.

Secondly, AI-based software systems have to be trained with prepared annotated data. These data are created by human specialists and an AI software is only as good as its training data set. So, if the training sets contain incorrect information or information with unexpected internal correlation, the results obtained from the AI software will display errors as well and will be misleading. Therefore, qualified human doctors are indispensable for both the development and clinical application of AI programs.

Finally, we must consider that doctors are not just human machines that are able to diagnose and treat diseases. As we are talking about human life and humans' wellbeing, there will always be the social aspect! A computer does not show sincere empathy to a patient in a difficult family situation nor can a computer adjust to different family settings; a computer cannot communicate a bad prognosis humanly and carefully; a computer would choose the best therapy according to the diagnosis yet not consider the social and family context of the patient. Human doctors are not only drillers like computers, but also surfers, i.e. they can also use and integrate scientific information in their decision-making and recommending process which is not immediately related or obvious.

Thus, the interaction between a computer and the patient will never be the same as the interaction between two humans, i.e. between a doctor and a patient. For many people who get the diagnosis of an incurable disease, the doctor with his expertise is the person they can trust and who can give situational and individual advices. It would be a dystopian idea to see a computer collecting your data, processing them and then telling you "I am sorry Sir, but most probably you have cancer." In my opinion, this is the main reason why computers will never replace human doctors: Medicine is dealing with humans. Thus, human interactions and emotions will always be part of it. A computer cannot create trust or empathy based on the acquired literature, and cannot feel, so it will not replace human doctors.

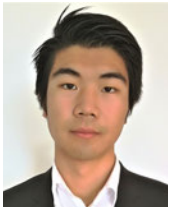
Although computer programs are not able to replace human doctors, it is beyond doubt that computer-based decision procedures can change the medical praxis immensely. As they are able to analyze data such as imaging or lab reports efficiently and precisely, they will facilitate and speed up those tasks and leave the doctors more time to focus on other aspects like the social component of the work, and provide a more personalized treatment with higher quality for the patient. With that, it is important to integrate these new possibilities in the curriculum of

medical education so future doctors will be able to use these powerful tools. Simultaneously, future doctors should be prepared for questions occurring with this development: Where and how will medical software potentially make mistakes? What should physicians do, when the human decision and experience and the output of a software do not agree? To find answers for those questions, it is necessary to not only know how to use these technologies but also to understand how and why they are working.

So metaphorically, the future doctor will be a surfer with broad knowledge who is able to use software as an excellent drilling tool.

To finish, I think the development of medical applications for computer programs can be compared to the invention of modern medical devices likes blood pressure meter or imaging facilities: Before those inventions, the doctor depended on his ability to observe and palpate a patient to make a diagnosis; with those devices quantifying the status of such a patient more precisely, the diagnosis became more reliable and efficient. However, this will not replace human doctors at all. In my opinion, the ability of software to process and diagnose specific scientific data and information will be a big support for doctors to make better decisions on diagnosis and treatment.

In summary, the advance in computer programs in medical application will have big impact on clinical routine and support doctors to be more efficient and precise, so their use and functioning should be integrated into medical education. However, they will not replace the human doctors in medicine.



Yi Zheng
Student Medicine BSc
Department of Health Sciences and Technology
ETH Zurich
8093 Zurich
Switzerland
zhengyi@ethz.ch
ORCID 0000-0001-7564-1775

Citation: Zheng Y: I diagnose, therefore I am a Doctor? Will drilling computer software replace human doctors in the future? Infozine **2018**, Special Issue 2, 16–17
DOI 10.3929/ethz-b-000296250

Published: November 15, 2018

Wilfred F. van Gunsteren
ETH Zurich

Surfing versus drilling in fundamental research

In 1633, as Galileo left the courtroom where his scientific opinions, formed by drilling deep into astronomical data, had been investigated by the Inquisition, he mumbled softly: "Eppur si muove" or "And yet she is moving", i.e. the earth. He had been coerced to confirm the "alternative" fact or theory that the earth does not move and that the sun orbits the earth, this to avoid death at the stake. The current surge in the emergence of alternative facts or theories, spread through surfing on the internet, yet being at odds with the truth, is nothing new. It only has less deadly consequences these days.

In research, questions or hypotheses are being formulated that are to be answered or confirmed/rejected based on facts and logic. Confronted with data obtained through research or gathered in the internet, one has to determine whether the answers and proofs found are reliable. Here four aspects are of relevance:

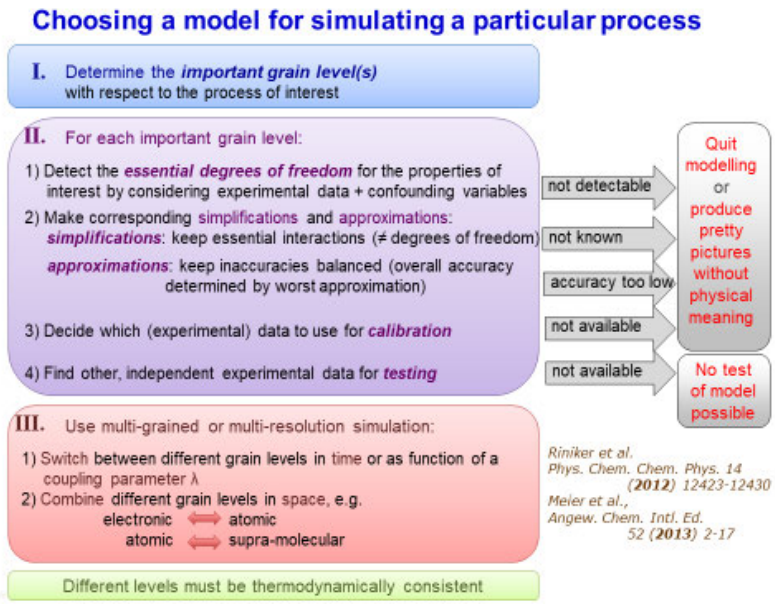
1. Are there any unproven assumptions underlying the research, data or theory?
How will the *approximations* used in the research, or inherent to a model or theory, influence the results obtained?
2. Is there sufficient statistics,
i.e. a sufficient number of independent observations, to draw conclusions?
Are the results unreliable due to *undersampling*, i.e. non-representative sampling? How large is the *uncertainty* (error) due to assumptions, approximations and poor statistics or sampling?
3. Are there any hidden or confounding variables,
i.e. factors that were not considered in the research, model or theory, but may influence/determine an observed correlation?
4. Causality?
Is there an explaining mechanism for an observed process or correlation ?
When searching for information on the internet, the data found are often insufficiently documented to answer these four questions, with the consequence that one cannot determine their reliability. In other words, the internet has a limited value when one wishes to drill deep.

In my research, the major use of the internet is to find literature regarding a specific topic, and to look up basic physical and mathematical data and formulae. Even translational dictionaries available on the internet, which I sometimes consult to find Dutch/English or Dutch/German translations of sayings or expressions, still perform poorly compared to a high-quality dictionary: Many complex language phrases and expressions are not properly translated.

Fundamental research is impossible without drilling deep. Yet surfing has its value when hitting upon analogous data patterns, models or theories in other fields of science. These may induce ideas for solutions to problems in one's own field of interest. This means that surfing should be done, but only rather limited in time, at least compared to drilling.

It is a popular thought that big data in itself can generate new scientific insight. Yet this is doubtful because the availability of lots of data does not guarantee any correlation or underlying mechanism to be abstracted from it. *Figure 1* illustrates this by describing the various steps taken when formulating a model at the atomic or molecular level to simulate bio-molecular behaviour on a computer. Although the bio-molecular data available in the internet is huge, one is often limited to only producing pretty pictures without real sci-

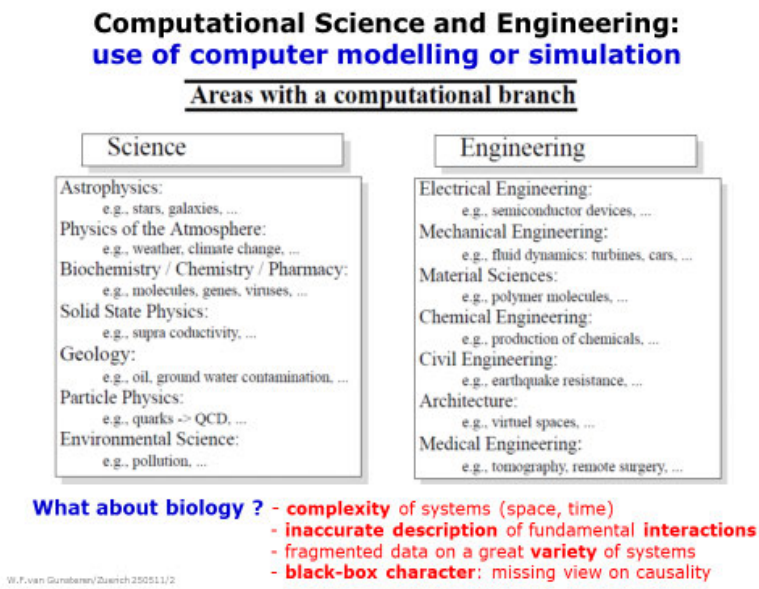
Figure 1



entific content due to the considerations of step II in the figure. This state of affairs is also responsible for the limited predictive power of computational models in biology, at least compared to their application in some other fields of science, see *Figure 2*.

Although pretty pictures without scientific content may be appealing, the tendency to value an image of research or of a research organisation higher than their contents or the truth constitutes a threat to drilling research. This tendency is detrimental for a healthy development of academic institutions, but no cause of real worry as long as the number of deep-drilling researchers stays much larger than the number of surfing ones.

Figure 2



Prof. Dr. Wilfred van Gunsteren
ETH Zurich
Laboratory of Physical Chemistry
HCI G237
8093 Zurich
Switzerland
Phone +41 44 632 5501
wfvgn@igc.phys.chem.ethz.ch
ORCID 0000-0002-9583-7019
http://www.igc.ethz.ch

Citation: van Gunsteren WF: Surfing versus drilling in fundamental research. Infozine **2018**, Special Issue 2,18–19.
DOI 10.3929/ethz-b-000294373

Published: November 15, 2018

Arieh Warshel
University of Southern California, Nobel Prize in Chemistry 2013

Using brain vs. brute force in computational studies of biological systems

Growing up with the emergence of computers, I have been exposed to the gradual increase in their power, starting just at the end of the paper tape period and the beginning of the punch card era. In a way, it was fortunate to start with a very limited computing power. For a long time, we had a rather slow turn-around frequency, getting results at most twice a day, and having to deal with trivial errors in the punch cards as well as major restrictions on the memory size. This meant that we had to think extremely carefully about how the program was written and how the underlying modeling approach was formulated. Beyond this limitation, early experience also taught me that computers can solve problems whose conventional (and ineffective) solution would require reading through enormous amount of written material (e.g. books about normal mode analysis). I also learned that computers provide an exquisite guide and a powerful check for new formulations and concepts.

My first experience in this respect was when I introduced Cartesian normal modes and found out that all the needed treatment ended up with one trivial formula once implemented in a computer program [1]. My selection of computational approaches was also influenced by watching in the very early 70's people trying to conduct conformational analysis of ring molecules by quantum-mechanical approaches. In contrast to this almost pathetic conformational search (without any minimization), our Cartesian second-derivative minimization using a force field found very rapidly the absolute minimum up to ten significant figures. Of course, I appreciated that quantum approaches would become reliable one day, but this was clearly not yet the time to use such seemingly reliable approaches for problems that were handled much better using force fields.

My direction was also determined by the fact that I did not have the option of running a program for a very long time and for using very large explicit-solvent systems. Such expensive setups would not have allowed us to check the results and, more importantly, would not have let us test the underlying assumptions by changing the running conditions. In fact, this constraint on the available computing power led me to formulate a rule that "any modeling of a biological system that has a run length of more than a day is probably incorrect or irrelevant". The reason for this postulate was the need for a time to analyze the results and to explore their stabilities. While my view on long runs has been slightly modified in recent years, with the need for expensive quantum-mechanical results and for the examination of convergence times, I am still convinced about the need for a very frequent examination of the simulation results, and of their sensitivity to the starting conditions and model features.

The realization that with the appropriate approach, computations can be used to study almost any problem (unless one insists on doing everything exactly and explicitly) started to get a hold on me in the middle of the 1970's, and became a clear direction in my early simulations of enzymatic reactions with Mike Levitt [2]. Since we dealt with enzyme catalysis and I did not want to make a fool of myself in entering an area that was so much studied experimentally, I felt that it was essential to include all the possible effects in the emerging model. This included the development of the QM/MM approach, that has already been reviewed many times [3]. Equally important was the fact that the development of the model cracked down the problem of electrostatic energy in proteins and solution. That is, after reading many classical books and talking to experts in electrostatic theories, I realized that the continuum description does not provide any practical help. Thus, I decided to move to an explicit (yet simplified) representation of the environment. This started with moving to a polarizable force field for the protein and representing water molecules by a grid of Langevin dipoles [2]. The fact that water was described with a simple dipole-based model led to never ending criticism: How can one use a dipole when everyone knows that water is not a dipole? (this argument was missing, of course, the fact that the dipoles were "effective", calibrated based on solvation free energies). In fact, for about ten years, we were the only group with a clear physical understanding of solvation and electrostatics in proteins, because we insisted on using a simple but complete model (see discussion in [4]. While we were studying the complete solvation of proteins, the rest of the community was already happy with calculations considering a single ion plus a single HCl molecule as a guide for studying solvation effects.

Another advance of the same period was the invention of a simplified folding model [5]. This model, and the subsequent Go model, was to offer the only way to simulate folding in a reliable fashion for the next three decades (permitting the exploration of the folding process, at a time when it was hopeless to study folding using all-atom simulations).

Another instructive example has been our study of the dynamics of the primary event in the vision process

[6]. This earliest simulation of the dynamics of a biological process predicted remarkably well the experimental reaction time (100 fs), which was unknown at that point. This was possible because the focus was on an extremely fast process, that could easily be simulated using the computers of the mid-70's. However, this was not the case for the attempt to simulate the dynamics of BPTI [7|, that would have required, for meaningful results, a computational power that was not to be available for a long time.

Our next round for more reliable simulations of electrostatic effects in proteins, and the first move to free-energy perturbation (FEP) simulations of a solvated protein [8], was the first time were we needed more resources than were available on the VAX computer. Fortunately (or miraculously), the electrostatic calculations converged in a few days for 20 ps, in part due to the use of powerful spherical boundary conditions. A similar move to the range where computer time might be important was the next generation of studies of enzymatic reactions, where we changed from the protein-dipoles Langevin-dipoles model to an all atom-model for both the protein and the solvent. Here also, the use of the empirical valence bond (EVB) model along with spherical boundary conditions proved highly efficient [9], allowing for the convergence of the activation free energies long before any alternative model could achieve the same result. Our progress was based on the realization that the available computer time was insufficient for any "reliable" calculations with molecular orbital QM/MM calculations.

Another example has been our simulations of the action of voltage-activated ion channels. The studies focused in this case on the use of a coarse grained (CG) model that allowed us to explore, with a limited computer time, the molecular origin of the ion selectivity [10], and establish that this selectivity was caused by having a different effective dielectric response for potassium-potassium and sodium-sodium interactions. Similarly, we succeeded to build a CG model that included explicitly the electrodes, the electrolytes and the membrane-protein system [11]. This provided a very powerful way to understand the nature of voltage activation [12]. Of course, recent years have witnessed the advance of brute force explicit-solvent computer simulations of ion channels, that could characterize the nature of a few conductance events [13]. However, while many will become addictive to such brute force atomistic simulations, it is still very challenging to obtain the free-energy differences relevant for the action of ion channels. Furthermore, even if we can simulate the long-time process realistically, it is crucial to digest the relevant results carefully and to see what parameters control the overall trends. This requires to look at the results frequently, and to consider the influence of different factors. In this respect, CG models still provide a major advantage.

The remarkable insight provided by CG modeling, without waiting for achieving converged results from all-atom simulations, has been illustrated by studies of molecular motors [14] and other complex systems like the ribosome-translocon system [15]. In all of these cases, it has been demonstrated that the insight provided without a major computational effort, but with a physically consistent molecular model, allows one to understand the action of complex biological systems and to know what questions to ask next. It also provided the crucial insight required for guiding subsequent attempts relying on massive computing power. Obviously, there are problems where it is hard to avoid the use of very significant computing power, and a case in point is ab initio calculations of free-energy profiles for enzymatic reactions. However, even in such cases, one can save a lot of computer time by using such approaches as our paradynamics [16].

Maybe one day, the power of computers will be so extensive that it will be possible to trust their brute force predictions in situations where the involved approaches have been carefully validated. An example of this type is provided by ab initio calculations on small molecules in the gas phase. In such cases, the use of the sole human brain may represent a limitation in terms of the problems that can be addressed and in terms of quality control. It is also possible that drug and protein design will once be carried out using dedicated computers, that are optimized based on carful studies, which gradually assess the conditions for obtaining reliable results. However, reaching this stage will require major human thinking. Another related issue is the enormous advance in popularity and impact of artificial intelligence (AI) approaches. Such strategies show a great potential in identifying similarities between related systems, and are expected to represent a major tool when analyzing biological systems. For example, AI can and will greatly help in drug design. But it should not be expected to be predictive in cases where one cannot have a way for direct extrapolation. The problem is that AI by itself is not likely to figure out what is the way to make the approach more effective. This will require the human wisdom gained from multiscale and other simulation approaches.

References

[1] Lifson S, Warshel A. J Chem Phys **1968**, 49(11), 5116–5129.
[2] Warshel A, Levitt M. J Mol Biol **1976**, 103(2), 227–249.
[3] Warshel A. Angew. Chem Int Ed Engl **2014**, 53(38), 10020–10031.
[4] Kamerlin SC, Vicatos S, Dryga A, Warshel A. Annu Rev Phys Chem **2011**, 62, 41–64.
[5] Levitt M, Warshel A. Nature **1975**, 253(5494), 694–698.
[6] Warshel A. Nature **1976**, 260 (5553), 679–683.
[7] Mccammon JA, Gelin BR, Karplus M. Nature **1977**, 267(5612), 585–590.
[8] Warshel A, Sussman F, King G. Biochemistry **1986**, 25(26), 8368–8372.
[9] Warshel A, Sussman F, Hwang J-K. J Mol Biol **1988**, 201(1):139–159.
[10] Burykin A, Kato M, Warshel A. Proteins Struct Funct Bioinf **2003**, 52, 412–426.
[11] Vicatos S, Rychkova A, Mukherjee S, Warshel A. Proteins Struct Funct Bioinf **2014**, 82(7), 1168–1185.
[12] Kim I, Warshel A. Proc Natl Acad Sci USA **2014**, 111(6), 2128–2133.
[13] Jensen MO, et al. Science **2012**, 336(6078), 229–233.
[14] Mukherjee S, Bora RP, Warshel A (2015) Q Rev Biophys **2015**, 48(4), 395–403.
[15] Rychkova A, Mukherjee S, Bora RP, Warshel A. Proc Natl Acad Sci USA **2013**, 110(25), 10195–10200.
[16] Plotnikov NV, Warshel A. J Phys Chem B **2012**, 116(34), 10342–10356.

Prof. Arieh Warshel
Department of Chemistry
University of Southern California
3620 McClintock Avenue
Los Angeles,
CA 90089-1062, USA
Phone +01 213 740 4114
warshel@usc.edu
ORCID: 0000-0001-7971-5401

Citation: Warshel A: Using brain vs. brute force in computational studies of biological systems. Infozine **2018**, Special Issue 2, 20–22
DOI 10.3929/ethz-b-000294355

Published: November 15, 2018

Jeffrey Bode
ETH Zurich

Laboratory literature boards in the digital age

Many, maybe even most, research labs once had a tradition of posting and discussing a "paper of the week", often near the group coffee machine or other high traffic area. Usually, this was some high-profile paper in the field of the lab, or occasionally something truly new in a neighboring field. With the move to electronic publishing and the deluge of journals and articles, this tradition seems to be disappearing. Even the effort to print a paper, pin it up, and encourage discussion among colleagues seems nowadays like too much work.

How, then, can one maintain the important tradition of sharing and discussing the latest research? This may be work directly related to ongoing projects in the lab or simply exciting things going on in science. Given the constant flow of information, and the ever-diminishing signal-to-noise ratio of the current literature, finding opportunities for alerting colleagues and lab mates about new developments is more important than ever.

Our group at ETH Zurich has long fulfilled this important part of science education and research through internal electronic messaging systems. In the early days, we used a product called Yammer, which provided a private messaging system that would automatically collect abstracts and table of contents graphics from a pasted link, and allowed direct online discussion of the paper among group members. After Yammer was bought by another company and largely removed from the small business market, we switched to Slack, which has proven to be a suitable alternative with several important advantages. Slack provides a closed network that can be limited to group members; project students and newcomers can be added or removed easily by one of several group administrators. We can set up many channels, both public and private, for discussions – the most important one being the literature channel. Within the literature channel, group members can post links to papers of interest, allowing for everyone to read and comment them. As a research advisor, this provides me with an excellent forum to post papers or news items that I think are important for students or postdocs to know about, and sometimes to give my personal thoughts about why this work is important (or, occasionally, not as important as it might look at first sight). Although group members do not post or participate as much as I would like, I am always happy to read their posts and see the papers they flag as being sufficiently interesting that everyone should know about them.

Although we initially set up this system with the sole purpose of establishing a platform for literature posting and discussion, Slack has taken on many other important roles and helps to build a group community. Every subgroup has its own public channel, where literature or discussions relevant to specific projects can be housed. Private boards can also be set up, allowing for closed discussions and for sharing files among a few members, which is particularly useful when writing papers or grant proposals. Slack works across many platforms, including smartphones, desktop applications, and web applications. It has largely replaced e-mails as the group communication system.

Finally, Slack allows for direct private messaging between group members. This provides an excellent informal tool for discussions between myself and group members, and is used extensively for intra-group communications. Based on our usage statistics, direct messages contribute more than 95% of all communications on the group system. Our group, like so many other research groups at major research universities, is increasingly delocalized. Team members may be working with collaborators at a remote location. Our group also has a satellite lab in Nagoya, typically staffed not only by local postdocs but often also by group members from Zürich. Our electronic system allows everyone to stay engaged and be part of the lab conversations.

As one improvement over the old paper-based system, Slack allows us to archive all the papers, posts, files, etc. I often recall that I posted some paper to Slack – and a search usually quickly retrieves the appropriate link. There are more advanced features, such as direct integration with Dropbox or Endnote, but for these more technical aspects one must ask my younger group members.

Modern publishing houses promote ever more tools to encourage "sharing" of literature results, with Twitter and other social media quickly becoming one of the primary venues by which we interact with the literature. But it is usually the people closest to us – our students and group members – with whom we want to discuss new papers in a semi-private fashion. While the days of the printed papers posted to the lab pin board may be gone, there are many modern ways to maintain this important part of research and education.

Prof. Dr. Jeffrey Bode
ETH Zurich
Laboratory of Organic Chemistry
HCI F315
8093 Zurich, Switzerland
Phone +41 44 633 2103
bode@org.chem.ethz.ch
http://www.bode.ethz.ch
ORCID 0000-0001-8394-8910

Citation: Bode J: Laboratory literature boards in the digital age. Infozine **2018**, Special Issue 2, 23
DOI 10.3929/ethz-b-000294374

Published: November 15, 2018

Sereina Riniker
ETH Zurich

Research strategies in computational chemistry

Scientists have evolved different strategies to successfully conduct research, depending on their personality, learning style and scientific field. In computational chemistry, we are typically interested in answering fundamental chemical questions while having the skills to automate tasks through computer scripts and programs. I would like to illustrate two extreme research strategies that emerge from these possibilities with an example from computer science.

In graph theory, data can be represented by a so-called *tree* structure, which consists of *nodes* containing the data and *vertices* connecting the nodes [1]. A tree typically starts with a *root* node from which branches emerge, ending in *leaf* nodes (Figure 1). There are two basic algorithms commonly used to search such a tree: depth-first search and breadth-first search. In the depth-first algorithm, starting from the root, a branch in the tree is followed until a leaf is reached and only then are other branches explored. In the breadth-first algorithm, each level of a tree is searched completely before proceeding to the next deeper level. Which of the two algorithms is faster in finding a target node depends on the given data and tree structure. Similarly, the success of *drilling* versus *surfing* when pursuing research depends on the problem at hand, and a mixture of both approaches often turns out to be the most efficient strategy. Thus, as usual in life, the trick is to find the right balance.

As an example, let us have a look at how one can determine computationally the octanol/water partition coefficient of a chemical substance. This coefficient is an important quantity in pharmaceutical research, where it is used as a surrogate for oral bioavailability [2]. It is therefore desirable to have an accurate estimate of this property by computational means before synthesizing a new compound to be tested as a potential drug candidate.

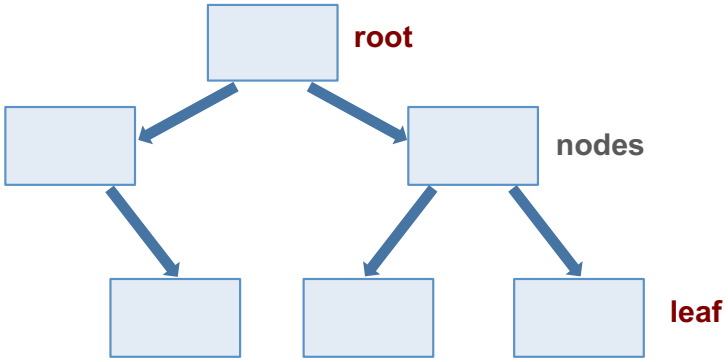


Figure 1. Schematic illustration of a tree data structure with nodes connected by vertices.

On the one hand, a partition coefficient can be calculated with rigorous free-energy methods using molecular dynamics (MD) simulations (see e.g. [3]). On the other hand, a machine-learning (ML) model can be fit to a training set, which consists of structural descriptors for thousands of existing molecules along with their experimentally determined partition coefficients (see e.g. [4]). If the compound to be synthesized is sufficiently similar to molecules in the training set, the model will perform well in predicting its partition coefficient. The first – physics-based – approach requires a considerable amount of computation, but it is more flexible and general as it does not depend on a training set.

The second – statistics-based – approach is usually much faster, but with the drawback that a sufficiently large, diverse and accurate training set must be available and that appropriate descriptors have to be chosen. Which of the two approaches works best to estimate the partition coefficient of a new compound depends on the nature of the molecule and on practical constraints regarding computational cost and accuracy.

In my opinion, success in science often relies on combining the two strategies when conducting research, i.e. to use *surfing* to identify areas for *drilling*. For example, we need large data sets to obtain sufficient statistics for constructing ML models with high predictive powers. However, a large data set means that it is no longer possible (or, at least, desirable) for a human to check each data point individually for its accuracy (e.g. when taking experimental data from the literature). Although the more sophisticated ML methods are typically rather robust against noise, too much noise still impacts the performance achievable by a ML model. It is, therefore, crucial to pay

attention to outliers in the predictions. Outliers can hint towards inconsistencies in the input data or weaknesses in the underlying hypotheses. For example, we may assume a linear relationship between input and output quantities when, in reality, the relationship is non-linear. In the presence of outliers, it is important to *drill*, i.e. to be persistent until you have convinced yourself that the reason for each outlier is understood. To be satisfied before reaching this point is dangerous, as the problem will typically resurface later on (often resulting in an erratum).

With the increasing amount of data that becomes available in chemistry and biology, exciting research opportunities emerge. But these come along with new challenges and with the requirement to curate a vast number of data points. Combining the best of the *drilling* and *surfing* research strategies is the best way to take advantage of these developments and to increase our understanding of nature.

References

[1] Golumbic MC: Algorithmic Graph Theory and Perfect Graphs. Academic Press, New York (1980).

[2] A. Leo A, Hansch C, Elkins D: Partition Coefficients and Their Uses. Chem Rev **1971**, 71, 525–616.

[3] Bannan CC, Calabró G, Kyu DY, Mobley DL: Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. J Chem Theory Comp. **2016**, 12, 4015–4024.

[4] Mannhold R, Poda GI, C. Ostermann C, Tetko IV: Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds. J Pharm Sci **2009**, 98, 861–893.

Prof. Dr. Sereina Riniker
ETH Zurich
Laboratory of Physical Chemistry
HCI G233
8093 Zurich
Switzerland
Phone +41 44 633 4239
sriniker@ethz.ch
ORCID 0000-0003-1893-4031
http://www.riniker.ethz.ch

Citation: Riniker S: Research strategies in computational chemistry. Infozine **2018**, Special Issue 2, 24–25
DOI 10.3929/ethz-b-000294372

Published: November 15, 2018

Nadine Schneider and Nikolaus Stiefl
Novartis Institutes for Biomedical Research

Surfing on the hype waves or drilling deep for knowledge? A perspective from industry

Today, it is easier than ever to get scientific information from all over the world. Scientific publications, conference proceedings and talks, blogs, preprints, podcasts, recorded lectures, contributions on social media platforms, and many more provide access to knowledge and information. In the past, most of the scientific results were discussed by small groups of scientists in labs, at conferences or in access-restricted journals. Today, science is debated more publicly in blogs, forums, or published in open-access journals or pre-print servers. This great evolution allows scientists worldwide to better connect and also enables less well-funded universities to better participate in the scientific community. At the same time, this brave new world creates new challenges: quality vs. quantity of scientific publications due to missing peer-review, or just overlooking important contributions due to the sheer amount of new publications. In fields such as data sciences and informatics, this can even reach a point where ease of access to information becomes the key to its relevance for the community: *If it is not available it does not exist.*

Today, it sometimes seems antiquated to drill deep into one special research topic rather than to surf the next hype wave. Is there a good strategy to find the right balance between the two approaches? And if so, how can this strategy be adapted so that scientists from both worlds, industry and academia, profit most?

The way scientists in industry tackle research questions is certainly different from the way scientists in academia work. In early pharmaceutical research, the major part of the work is performed in project teams, usually a consortium of experts in different fields. In medicinal chemistry projects, for example, when trying to find a new medicine for a certain disease, many scientists have to work together: analytical and organic chemists, *in-vitro*, *in-vivo* and structural biologists, and today, also data scientists. For the individual scientist, being able to speak and understand the language of the various disciplines is essential. Without a high-level understanding, they are not able to adjust their own experiments to the information available. In other words, they need to be able to surf each discipline in order to drill deep into their own specialty. Whereas in academia, drilling often means developing a novel tool, a new assay or a sophisticated synthetic pathway, in medicinal chemistry projects, drilling means combining all the different data in order to gain an in-depth understanding of the underlying problem. This allows the respective experts to adapt and plan their next steps, such as the best synthesis, the assay that provides most insight or the optimal read-outs to guide the next experiment.

Focusing on solving the next problem is often driven and constrained by the allocated time frame and financial budget of a project. Often, decisions need to be taken pragmatically and, more often than not, solving the problem is just good enough. Unfortunately, sometimes, the curiosity of scientists might remain partly unsatisfied when time or budget constraints do not allow for a really deep dive into a scientific question. To fill this gap, industry relies on academia to help with basic research. Recently, various programs have been initiated to open the framework and enable closer partnerships between industry and academia:

- Various pharmaceutical companies offer on-site post-doc programs. Here, junior scientists profit from the experience they gain in an industrial environment. At the same time, industrial researchers quickly reach a deeper insight into novel techniques or special areas of research they do not usually have time to dig into.

- If there is a true need for experience in a certain field, companies collaborate or consult with academic experts to bring the knowledge in-house. This type of drilling helps the industrial partners to immediately tap into an in-depth expert knowledge, and get an opinion on specific scientific questions. To the academic collaborators, apart from potential funding, this provides an opportunity for surfing the "industrial problem wave" – getting insights into real world applications related to their field of expertise, which can in turn lead to novel and applied research questions.

- For research areas that need to be approached using diverse scientific disciplines, consortia between academic and industrial partners are formed. Here, similar to industrial projects, a mix between drilling and surfing is often necessary to cover the different topics. Frequently, these consortia work on cutting-edge methods, where academia is enabling industrial experiments with basic scientific research. The results in turn allow the academic partners to improve their methods much faster.

In essence, there is no preferred option between surfing and drilling. Even though it might sound preferable to be an expert by drilling deep into a research topic, being able to also understand and have a high-level knowledge of multiple disciplines is key in modern science. Disciplines inspire each other and engender hybrid approaches spanning multiple fields (e.g. DNA-encoded chemical libraries, artificial intelligence (AI)-driven chemical synthesis platforms). The more biology, chemistry, physics, data science, and other fields grow together, the more scientists need to be able to combine both surfing and drilling. This is also reflected in the various interdisciplinary studies appearing around the globe. The principal challenge for future researchers will be to figure out when to best apply which strategy. Certainly, the balance between surfing and drilling changes during a scientific career. Initially, building up an in-depth expert knowledge that should be expanded to a broader generalist skillset is a great foundation to be able to surf while drilling deep. We believe that only this phenotype of scientist will succeed in the long run – be it in academia or in industry.

Dr. Nadine Schneider
Investigator II
Novartis Pharma AG
Novartis Institutes for Biomedical Research (NIBR)
Global Discovery Chemistry/CADD
4002 Basel, Switzerland
nadine-1.schneider@novartis.com
ORCID 0000-0001-5824-2764

Dr. Nikolaus Stiefl
Senior Investigator I
Novartis Pharma AG
Novartis Institutes for Biomedical Research (NIBR)
Global Discovery Chemistry/CADD
4002 Basel, Switzerland
nikolaus.stiefl@novartis.com
ORCID 0000-0003-2562-7080

Citation: Schneider N, Stiefl N: Surfing on the hype waves or drilling deep for knowledge? A perspective from industry. Infozine **2018**, Special Issue 2, 26–27
DOI 10.3929/ethz-b-000294375

Published: November 15, 2018

Philip Mark Lund
University of Copenhagen

The use and purpose of articles and scientists

Nature is of an interesting size. Humans try to categorize nature by applying science so that we are able to understand and evolve. By enabling this evolution we have altered the way that knowledge is transferred between generations. From a direct transfer via observations, which is seen in the animal kingdom, to indirect transfer via e.g. the written language. In today's modern society most of the knowledge transfer occurs via indirect methods such as articles, books, and audio/visual media. This, of course, leads to a larger accumulation of information, which has its advantages and disadvantages.

The most obvious advantage is the application of the great amount of knowledge created for further advancements, i.e., creating and developing inventions. These advances can be straightforward, like the development of OLED technology for screens or they can be a byproduct of another research, as is the Post-It note, where the original research aimed for a super strong adhesive.

However, with the huge amount of information gathered it is impossible not to have partially or completely incorrect information. Thus, it is crucial to do iterative testing and have critical discussions, and to validate knowledge. Moreover, categorizing this huge amount of information is a nightmare. Organizations like IUPAC and RCSB Protein Data Bank, which categorize certain types of information, are therefore indispensable.

With the complexity of nature, the way we describe nature has to be complex and it will increase in complexity. It is like trying to fit a circular object into a box and stating that the box describes the circular object. It does so to some extent. By adding additional sides to the box, making it a pentagon, hexagon etc., we increase the precision of the description – but we also increase the complexity of the structure. This decision on the appropriate number of sides used for describing the structure is a frequent task in research. It is a difficult balance between a comprehensive description of the transferred knowledge and the efficiency of the transfer to a target audience. If knowledge is processed into a too complex description to be understood by the audience the entire rendering will be useless.

It is speculated that the time we are living in is the period with the most rapidly expanding knowledge so far, though this is not surprising given the increasing world population. Moore's law states that the number of transistors doubles every second year. This means that computation power increases steadily and – combined with modern techniques like machine learning and artificial intelligence (AI) – automatizes the collection and generation of knowledge. An organization named *Association for the Advancement of Artificial Intelligence* has even had conferences discussing future learning methods when AI gradually takes over research and day-to-day jobs. This opens up a discussion on our future and purpose when AI takes over our jobs.

All these thoughts lead to the question: Is the way we are writing articles in coherence with the way we do research today and how will future articles be written? New formats have been introduced to improve communication, such as letters, article or reviews, and additionally clarifying titles for each section and subsection in the literature. For the Journal *Nature*, the term *research article* was first used in 1933, while *letters* were introduced in the very first issue in 1869. But is the expansion of information progressing too rapid such that the presentation format of knowledge is no longer adequate?

As a greenhorn scientist, I was first truly acquainted with research articles in the second year of my bachelor studies. A half-day introductory course was given on how to read articles and search for them on *Web of Knowledge* and *Scopus*. This has been an imperative course, since it relayed key insights on how to get through an article without having to have had three cups of coffee (or five if it is a complexly written one).

At university, we are given the basic set of tools to understand different topics, as e.g. understanding what makes molecules react or how density function theory works. From here, we then explore the world using these tools, gradually improving our skill set through experience and discussions. The same prioritization should be given for reading articles, which is the bread and butter for development in science. It takes practice to read articles, and to learn how to read for which purpose. Without that introductory course my learning curve from getting the basics of an article to fully understanding the contents would have been much less steep.

Getting educated at a university causes a steep learning curve, in which multiple new skills are learned. One is structuring one's time between individual courses and the additional leisure time there may be. To my knowledge, there are very few students who achieve a graduate degree, who have read and understood every text assigned to them throughout their study program. Thus, it is a constant decision-making process, between reading a text in depth, reading it lightly or not even reading it at all, asking yourself "How important is this text for the lecture/exam?" or "Does it spark my interest?". I believe that these decisions gradually and naturally progress from the teaching material of books to the reading of an article for a project or thesis. There will always be a motivation to

read a text, whether it is for the sake of the carrot or the stick. Apart from motivation, ambition and goals also play a big role in deciding whether to surf over or drill into an article and to which degree. Nobel laureates and recognized scientists probably go far beyond the surfing through texts, and are likely to have a good sense of this decision-making and good reading skills.

It is interesting to view how the field of research has evolved from the early 1900's to modern days, where the sheer quantity of articles published might at times come at the expense of their quality. To be clear, this is not a criticism of the number of articles produced, since each research field has a certain number of articles published a year and adding correct, trustworthy information is crucial. The more reliable information, the merrier. It is criticizing poorly produced articles.

Obtaining a Ph.D. degree does not necessarily mean that one is able to communicate the knowledge one has acquired and found. Thus, it should be obvious that when acquiring a Ph.D. degree communication skills must also be acquired – as knowledge causes responsibility. The responsibility to communicate, clarify and visualize what is otherwise only visible to the few.

The purpose of science is not only to obtain new knowledge but also to make it known to the world, and to develop it further. In order to do so, science must move beyond the boundaries of narrow research groups. This means that scientists must cooperate and share knowledge via indirect communication methods. Simply because the task of the scientist is not only to discover and accumulate knowledge but to communicate it through articles that can enlighten and enrich the whole of our scientific spectrum.

Philip Mark Lund
Master Student
University of Copenhagen
Niels Bohr Institute
Blegdamsvej 17
2100 København
Denmark
nhs827@alumni.ku.dk
ORCID 0000-0001-5935-9753

Citation: Lund PM: The use and purpose of articles and scientists. Infozine **2018**, Special Issue 2, 28–29
DOI 10.3929/ethz-b-000296243

Published: November 15, 2018

Oliver Renn
ETH Zurich

Can you look at papers like artwork?

Are you a driller or a surfer when dealing with information and, in particular, with scientific information? Is one of the two ways smarter and, if yes, which one? Let's keep these questions unanswered for now, and look first into a totally different field: arts.

If you enjoy visiting art museums, there are at least two ways of delving into the artworks. You may start at the entrance and walk yourself through all the rooms, allocating equal time to each piece of art until you reach the exit. This is feasible for a small museum, but hardly applicable to places like the Louvre in Paris or the Eremitage in St. Petersburg! At the Louvre, you might limit yourself to Mona Lisa – or other known knowns. But if you do not know what to expect you need another strategy. My personal strategy is different. I walk quickly through the entire museum, leaving people behind who certainly think I am a philistine, paying no attention to art. However, this way I am gathering an overview of the artworks I want to focus on later. This, I do in a next round, now sitting in front of the pieces I like and studying them very carefully. At the end, I will not be able to recall all the artworks of the museum, but I will have discovered and studied a few of them very well, those which will have an impact on me – as an artist. I doubt that people going for the exhaustive tour – except maybe those with an eidetic memory – will recall much more the following day.

But what if the museum's walls are so packed that you cannot focus on one piece of art? If there is a "Petersburg hanging", and every square inch of the wall is used? And what if you want to see dewy new art that had no chance to be curated by a professional and exhibited in a well-known museum or an art gallery? How can you discover those new talents?

Those questions and strategies can also be applied to information seeking and processing.

Over a million research papers are published every year. And old papers are not leaving the market – unlike artworks that are (at least in part) being sold and hidden in private rooms. Old scientific papers still exist – some will keep their relevance for ever – and add to the overwhelming flood of information.

In the seventies, it may have been possible to cope with all scientific literature. But with up to 90 million papers in literature databases, things are different now – because something else has not changed: The day still has 24 hours and even if you have no other tasks you can hardly allocate more than 14 hours of it to reading scientific literature. So, how to deal with this information overload? Burying your head in the sand and ignoring new research altogether is clearly not an option. So, should you stop digging and limit yourself to surfing, i.e. browsing around and only reading the headlines? Like with artwork, I personally think that one should both dig and surf in science.

Tools for surfing and digging

In order to find relevant information you need to search – which typically starts with surfing – and then, you may end up into digging. **Google** lends itself very much to surfing, but its selection is highly biased: The hit lists are generated by an algorithm you do not know – and most likely, you will not go beyond the results of the first page. **Google Scholar** is better, but searching within the results is limited. The system is not article-centered, but based on links to similar contents. Abstract & Indexing databases such as **Web of Science** or **Scopus** allow for more sophisticated analyses of the hits. Although you cannot drill into the 78,500,000 "Alzheimer" hits of Google, or even the 2,250,000 hits of Google Scholar, you possibly can drill down the 190,000 hits of Scopus or the 967,000 hits of **Dimensions** – a new player in the field. However, all systems are limited to keyword-based searches only. With Scopus, you can continue searching

your results sets for additional words, separated by a to-be-defined number of any other words – even in a particular order.

If you work in the life sciences, you may also use **PubMed** and benefit from the indexers' work, who assign all articles to the sophisticated system of **MeSH** (Medical Subject Headings). However, you need to know the MeSH system if you want to fully benefit from this indexing. Also, the indexing is limited to a number of predefined medical subject headings and subheadings – and manual indexing leaves room for human mistakes.

Tools like **Qinsight** can help you, as they "translate" your search, allow you to search in natural language and look not only for keywords but for relations, dependencies – and synonyms you may not know. Thus, if you want to know if there is a particular relation between A and B, you can use the technical backbone of Qinsight, i.e. artificial intelligence, machine learning, ontologies, text mining etc. If you have no clue if A has any relations to whatsoever, the tool may also help you, as it shows you what A is related to. This may be B, but also K, M, or Y you may have never heard about.

However, to be absolutely sure not do not miss any relevant paper you need to use text mining technologies to analyze contents. This is a time-consuming and expensive approach, but may be worth the effort – especially in a corporate setting.

Learning about the unexpected

But – in the artwork analogy – how do you learn about new artists? In the last century you would have gone to the library shelves, flipped through paper journals and accidentally discovered something that helped you making unexpected connections, thereby generating new knowledge. Today, print journals are gone. And, honestly, do you frequently visit your favorite journals' websites? Maybe you have subscribed to **e-ToCs**? If so, are you reading those regularly? If e-ToCs pile up unread in your e-mail folder, you can instead use the app **Browzine**. This library app notifies you about new issues, and you can read articles as PDF, also offline – share, annotate, store and manage them. This is about the closest you can get back to the print age in the digital era. Another useful source are **alerts**, e.g. Google Alerts or alerts from literature databases, like Scopus. This way, you can receive notifications when your papers are cited, when friends or competitors publish something new, when the literature you like or cite is also cited by others, or if papers matching your search queries are published. You can do this alerting in Web of Science, Scopus (which I prefer), Google, Google Scholar (which I additionally use), Dimensions, PubMed, Qinsight or even in social scientific networks like **ResearchGate** or **Academia**. Another powerful tool is **Twitter**. You do not need to be an active user, i.e. post own contributions. Just follow accounts that look interesting – you can unfollow them anytime. This is particularly useful to learn about new developments. Once you are following a certain number of accounts, Twitter's algorithms will additionally present you matching tweets, and you can even further improve by following the creators of these tweets.

It is probably wise, however, to limit yourself to a small selection of tools. The reason is related to the next important issue: Time for reading.

Getting time for reading

If you subscribe to the e-ToCs of all journals that are important to you, this may result in 50 e-mails a day. More e-mails may come from citation alerts, author alerts, saved keyword searches in too many databases, plus the alerts that come through social networks, like LinkedIn, ResearchGate, or Twitter. Browzine will update you in real-time on the most likely 3-digit number of unread papers, adding to the hundreds of notifications you get through messenger apps or collaboration tools like Yammer or Slack on all your devices. Are you going to read this?

It is not surprising that so many books on digital exhaustion and digital detox have been published recently, as well as a growing number of apps that try to help you to manage your time. Today, focused work is as underrated as multitasking is worshipped. Similarly, open office space is often regarded as the ultimate solution for fostering collaboration and innovation. Serious work is then probably only possible at home offices – as I doubt that one can do "deep work" in such a distracting open-space setting. Deep work is a term coined by Calvin Newport who in 2017 began advocating "digital minimalism".

I am strongly advocating the use of digital tools that allow us to work more effectively and efficiently with scientific information. But I am also recommending not to use more of them than you can digest – like a farmer should not plant more than can be harvested. A wise use of these tools ensures that there remains sufficient time for reading alerts and notifications. Productive knowledge workers try to avoid meetings and digital distraction. As Peter Drucker said, you can either go to meetings or do real work, but never both. Meetings are easily scheduled, especially by those who only practice management but no longer do creative work. For me, decision-making is a useful task but certainly not creative work. And in research, you should be a creator! Many good ideas – that you often get out of your office, e.g. when walking in the woods – need to be developed further in a focused and quiet way. During this process you will need again scientific information and data. It would be a big mistake not to use today's technology tools. What to use and when is a very personal decision, and a learning process, until you find out what suits you best.

Dr. Oliver Renn
ETH Zurich
Lecturer, Head Chemistry | Biology | Pharmacy
Information Center, Head Science Communication
D-CHAB, HCI J57.5
8093 Zurich,Switzerland
Phone +41-44-632 29 64
renn@chem.ethz.ch
ORCID 0000-0002-6966-7757
https://www.infozentrum.ethz.ch

Citation: Renn O: Can you look at papers like artwork? Infozine **2018**, Special Issue 2, 30–31
DOI 10.3929/ethz-b-000294380

Published: November 15, 2018

Frank Perabo
Maia Biotechnology, Inc.

Dynamite fishing in the data swamp

The ability and desire to interpret scientific data contained in an increasingly large public domain and proprietary databases has gained a lot of momentum, noticeably in biology, chemistry, pharmacology and medicine, using sophisticated analytics, or 'artificial intelligence' (AI). Next to the desire to analyze those databases, the mere availability and flood of data can overwhelm researchers and doctors with respect to their day to day activities and decision-making. The solution seems to be in computing and IT advances that promise to make sense of it all and sort the relevant from the irrelevant. And not only that, but these tools claim to enable scientists to find hidden gems of information they were unable to see themselves, identify patterns they thought didn't exist, or help manage the tsunami of data coming at them.

There's no question that the advantage seems obvious. Here are a couple of examples: Epidemiologists and researchers could compare data from patients' genes, their lifestyles and medical records with data from massive databases gathered over the years by companies, healthcare providers and payers. Doctors see the potential to draw insights from tremendous volumes of real-world data and apply it to support clinical decision making with the goal of providing safer and more efficient services. Instead of relying on verbal evidence from patients, that frequently is considered unreliable, prone to variability or may not provide enough information for good decision-making; gathering real-time, patient data via wearable devices for example, could help to produce consistent, objective evidence of actual disease states and the impact of drug efficacy on symptoms, when a range of biometric signals such as heart rate, blood pressure, sleep and activity can be measured 24/7 (goodbye privacy). Oncologists could define small subsets of cancer patients that can benefit from a specific treatment, analyze data to predict the prognosis of the patient and guide different options, including truly personalized medicine. Software could assist pathologists and radiologists in the detection of cancer, which could reduce significantly time-consuming, and possibly biased work.

Last but not least, drug developers see potential to draw insights from tremendous volumes of real-world data and apply it to the design of clinical trials, which could improve data quality, predict patients' responses to therapies using relevant biomarkers, increase patient compliance and retention, more efficiently predict treatment efficacy, and significantly reduce cost.

However, where are we with this? Is this real? What can we expect?

There are several fundamental, I would call them pragmatic 'common sense', approaches to data. One is the data itself. Then the tools we approach data with, and the user working with those tools and data.

The data

Many data in the medical field originate from largely unstructured electronic health records, with data coming from multiple sources which were collected and structured for different purposes. Most routine databases do not have sufficient quality to be used by AI algorithms to achieve the quality standard required for profound analysis. The key difference is frequently data originating from a controlled (chemical libraries) vs. uncontrolled (patient data) environment. Controlled data sets typically have high quality and are coherent, vs uncontrolled data are in many cases a vast data collection with little structure and less quality. Many of the large electronic health records suffer from missing data, or 'dirty data', that lacks quality control. Frequently, the data needed for a particular analysis, may actually just not be there. They are not included in a particular data set. Beyond data sets with missing data values, or a dataset with dirty data, simply data are not present because they were not captured. Critical data may just not have found their way in a dataset, because they were thought to be irrelevant or not related to the data collection purpose. Now, one would assume that collecting 'all' or as many as possible data would minimize that risk, or integrating various databases, or linking to other supplemental sources of information may overcome this issue. At this stage the realm of data integration becomes relevant. It has been incredibly challenging to find standardized ways to integrate complex datasets, given various formats, data fields, and different data structure.

The tools

Many algorithms in the past were only as good as the person who programmed them. Simply put, if the hypothesis was wrong, the results of the programmed code could be misleading. Now, researchers are unleashing AI, often in the form of artificial neural networks, on the data torrents. Unlike earlier attempts at AI, such "deep learning" systems don't need to be programmed with a human expert's knowledge. Instead, they learn on their own, often from large training data sets, until they can see patterns and spot anomalies in data sets that are far larger and messier than human beings can cope with. Those algorithms are suggested to learn, predict and advise based on vast amounts of data. However, unlike a graduate student or a postdoc, neural networks can't explain their thinking: The computations that lead to an outcome are hidden, or inexplicable.

The user

This may be the most fundamental challenge to data analysis. This is the very reason to implement AI, to exclude the human error, or bias. To separate out the bias coming out of so-called paradigms. The change between known or expected patterns. But with AI eliminating the human confounding may be a short-sighted belief in the abilities of AI.

Here's one hypothetical example, admittedly simplified. The paradigm would be 'people in the certain Caucasus regions live very long because they eat a lot of yogurt'. Both may be true, they may live long, and they may eat a lot of yoghurt, but 'age' and 'yoghurt' may have no correlation whatsoever, because the data collected didn't capture other relevant epidemiological, environmental, geographic, social, health or other relevant data. It might be that current population people in this Caucasus region is older in average than other population in a geographic vicinity around it, because the younger population left for bigger cities to find jobs, as a consequence of shifting social and economic patterns. Or as inconceivable as it might seem, maybe there's an unknown regional microbiologic environment such as a particular moss (not yoghurt) which supports the body's immune system to fight off infectious diseases and cancer. To assess whether certain people in the Caucasus live to an old age, and why will require an incredibly complex and massive database – and yet the results may still suffer from the unknown, data not captured. i.e. the prevalence of spores of the moss in drinking water.

Here is where one critical factor comes in beautifully. Too much reliance on AI, could almost completely eliminate the discovery by accident. The lives of millions have been transformed and saved by treatments that scientists were not even looking for. During early disappointing cardiovascular clinical trials of sildenafil, now better known by its trade name Viagra, male volunteers (modern myth has it those were Welsh mine workers), taking the pills reported erections. After further investigation, it turned out that Viagra, designed for a completely different purpose, inhibits an enzyme that is key to erections by relaxing smooth muscle cells in the penis.

In the 1980s, two Australian doctors were ridiculed for suggesting that stomach ulcers were caused not by business lunches and stress, but by infection with a bacterium. Those two researchers noticed that stomach biopsies taken from their ulcer patients all contained the same spiral-shaped bacteria, called *Helicobacter pylori*. And another modern myth has it that one researcher deliberately downed a pint of helicobacter broth that he'd grown in his lab after isolating it from the stomach of one of his patients. Within a week, he had rampant stomach inflammation – which was then completely reversed by taking antibiotics. Their chance discovery has also meant the virtual eradication of a type of stomach cancer caused by helicobacter infection.

Those entertaining examples do not suggest that the future of research should lie in random discoveries, but it highlights that next to data quality, data structure, and data completeness, together with sophisticated analytic tools, the human curiosity, intuition and desire for invention will continue to be the most critical driver of next generation research.

Frank Perabo, MD, PhD
President
Maia Biotechnology, Inc.
1700 Post Oak Boulevard
2 BLVD Place, Suite 600
Houston, TX 77056, USA
Phone +01 713 963-3670
docperabo@yahoo.com
ORCID 0000-0002-2066-7115

Citation: Perabo F: Dynamite fishing in the data swamp. Infozine **2018**, Special Issue 2, 32–33
DOI 10.3929/ethz-b-000294359

Published: November 15, 2018

Jeffrey D. Saffer and Vicki L. Burnett
Quertle

Streetlights, augmented intelligence, and information discovery

Finding and understanding previously published information is the foundation for advancement in any field. As important as this is, it may be surprising to learn that search engines – by the very nature of how they work – generally intensify human cognitive biases, often limiting our ability to discover the most impactful information. Recognizing this is critical for designing smarter search engines that surf information better, drill down to details better, and combine these two aspects into a powerhouse approach that gives tremendously better insights.

Observational biases in search engines

Cognitive biases [1] can be introduced, or made worse, by search engines. Here we address two biases that specifically decrease search effectiveness.

1. Streetlight effect

This bias [2] is explained by a story of a man seen on his hands and knees under a streetlight one night. Asked what he was doing, he said he was looking for his keys, which he lost near the tavern. "Why look here when the tavern is down the street?" "Because this is where the light is!"

This funny situation is not so humorous when looking for critical documents. There are often so many irrelevant results that we are forced to narrow down the search, thus creating our own streetlight in the process. We have been "trained" to drill directly for what we need, bypassing a broader and potentially critical view.

2. Availability cascade

The availability cascade [3] occurs when information seen more frequently is viewed as more important. A form of this bias arises in search engines such as Google [4, 5] where discoverability is not only based on the content, but reflected by other users interest. This may be a great problem, since even the same person doing the query a second time may now be interested in a different perspective. Biasing future exploration using past history is not a valid way to evaluate scientific information.

Modern approach to search engines

The biases above are introduced by traditional search engines because of:

(a) poor search precision, making it impossible to search for a higher-level perspective
(b) ineffective exploration of results, making it impossible to drill down effectively

Artificial intelligence (AI) is one approach to solve these problems. Keep in mind, though, that AI is a very broad category of methods. Just like the choice of which hand tool to use is critical, the choice of AI method – or even how a given method is applied – determines success or failure. However, appropriate AI methods can help remove observational biases if these issues are considered in the very foundation of the implementation.

1. Getting the big picture (surfing with a purpose)

Enabling the big picture is one of the most critical aspects for elevating "search" to "discovery". Suppose you want to know something about the genes that contribute to melanoma. A search for "genes and melanoma" will give over 15K papers. It is not only impractical to understand these results, but most relevant articles will be missing because traditional engines look for articles using the generic term "gene". Questions like this can, however, be effectively addressed by implementing the right AI methods to understand what the user is looking for and, in this example, to search simultaneously for every gene and its relationship to melanoma. This investigation at the systems biology level – finding biological processes in context of the whole – enables greater understanding.

2. Getting down to details (effective drilling)

We have found that effective drilling requires AI-based conceptual searching; that is, finding content based on meaning. For example, "induction" and "activation" are similar conceptually and a searcher would want to find both aspects. In addition, search terms should have meaningful connections in the document, either explicit or implied. Just because a document [6] mentions "banana" and "elbow" doesn't mean bananas affect elbows.

3. Combining surfing and drilling

The need for a big picture as well as details are not mutually exclusive. With combined technologies the real fun – and productive discovery – can begin.

AI can seamlessly combine surfing and drilling. One approach is to use AI methods to automatically identify the concepts of interest to the user. These concepts need not be limited to categories implicit in the user's query and can be extended so that there are no preconceived bounds on what is relevant. In this way, serendipitous discoveries are possible.

The bridge between surfing and drilling can be further enhanced using predictive visual analytics. Here a picture really is worth a thousand words, as Figure 1, an example from a search for "genes and melanoma" shows.

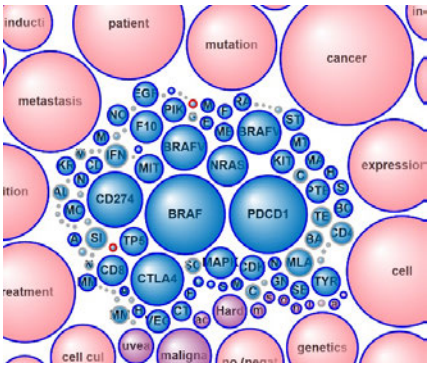


Figure 1

This image helps the user see that BRAF and PDCD1 are more important in the literature (bigger bubbles). The red border on two of the smallest gene (blue) bubbles is an indication that those genes will become more important over time. And by being interactive, the visual analysis allows the user to easily drill down to information that would be missed by traditional methods. Furthermore, when the results are updated based on the user interaction, the information is shown in context so that the user can quickly evaluate the importance.

Summary

From the discussion above, it should be clear that neither surfing alone nor drilling down alone can provide the integrated knowledge, broad context, and opportunity for serendipity that combining the approaches provides. The combination provides augmented intelligence, leading to deeper knowledge with much greater efficiency.

References

[1] List of cognitive biases. Wikipedia http://bit.ly/biasgraph
[2] Streetlight effect. Wikipedia https://en.wikipedia.org/wiki/Streetlight_effect
[3] Kuran T, Sunstein CR: Stanford Law Review 1999, 51, 683-768. DOI: 10.2307/1229439
[4] Schwartz B: Google Posts That Local Results Are Influenced By Clicks, Then Deletes That https://searchengineland.com/google-posts-that-local-results-are-influenced-by-clicks-then-deletes-that-237579
[5] Fishkin R: Queries & Clicks May Influence Google's Results More Directly Than Previously Suspected https://sparktoro.com/blog/queries-clicks-influence-googles-results
[6] Attarian S et al: Orphanet Journal of Rare Diseases 2014, 9, 199. DOI: 10.1186/s13023-014-0199-0

Dr. Jeffrey D. Saffer
Co-founder and CEO
Quertle
2505 Anthem Village Drive, Suite E170, Henderson, Nevada 89052, USA
saffer@quertle.com
https://quertle.com
ORCID 0000-0002-9066-3224

Dr. Vicki L. Burnett
Co-founder and COO
Quertle
2505 Anthem Village Drive, Suite E170, Henderson, Nevada 89052, USA
burnett@quertle.com
https://quertle.com
ORCID 0000-0002-3489-305X

Citation: Saffer JD, Burnett VL: Streetlights, augmented intelligence, and information discovery. Infozine **2018**, Special Issue 2, 34–35
DOI 10.3929/ethz-b-000294368

Published: November 15, 2018

Michiel Kolman and Sjors de Heuvel
International Publishers Association / Elsevier

"Yes Dave. Happy to do that for you."

Why AI, machine learning, and blockchain will lead to deeper "drilling"

Once the cornerstone of scholarly communication across disciplines, the monograph book, in which an author could delve into one topic in great detail, has been proclaimed dead – at least in the exact sciences. Now that it is gone, what will replace it? While the "digital-liberal era" – as the editors of *Infozine* have dubbed it – provides us with access to more and more sources (databases and other software), improved reach (through a variety of media), and contact with peers on a continuous basis (social networks), these tools do not leave scientists with extra time to read up on recently published material. At the same time, the scientific community in general is worried about the integrity and reproducibility of its research. Though still in its infant phase, we believe that modern technologies such as machine learning and artificial intelligence (AI) will allow us to "surf" with unprecedented speed, while balancing the "drilling" depth, quality, and insight attributed to the monograph – but without its physical confinements.

When talking about AI, many of us think of HAL 9000, the antagonistic robot from *2001: A Space Odyssey* who ultimately turns against its human space ship crew (including Dave, for those who did not get the reference in the title). However, according to the Artificial Intelligence Index "today's AI systems have far less common-sense reasoning than that of a five-year-old child. [1]. Nonetheless, the technology is expected to develop rapidly in the coming years and is already being applied by STM (Science, Technology, Medicine) publishers in a variety of ways. As research is becoming increasingly multi-disciplinary, scientists need to quickly come up to speed on topics outside their core discipline. Elsevier's ScienceDirect database now features an extra layer titled "Topics" which draws from citeable book sources to concisely define unfamiliar terms in articles and connect content across disciplines – all using automated approaches for information extraction. In a way, "Topics" is the 21st century answer to the encyclopaedia, serving as an introduction to new subjects. It is powered by AI, but based on reliable peer-reviewed information.

Similar innovations are taking place across the industry – especially where it concerns reliability. In 2017 Digital Science launched "Dimensions," a research information database that makes use of AI to automatically link publications with grants, policy, data and metrics. This year the company joined hands with Springer Nature and Amsterdam-based start-up Katalysis to test how "blockchain" technology might support the peer review process. A strong candidate for "Buzzword of 2018," blockchain has so far powered the Bitcoin system. It is essentially a technology for decentralized, self-regulating data [2]. You could see it as a ledger to which the members of a peer-to-peer network can add "blocks" of information, but single users are unable to modify what is already there. Scientists and publishers alike believe that blockchain might make the exchange of scholarly information more secure and transparent – and peer review less biased, but more visible and trustworthy. Though AI, machine learning, and blockchain should never be regarded as a one-size-fits-all solution to academia's problems, these technologies have the potential to leave us more opportunity for "drilling."

Elsevier's Topics and Digital Science's blockchain experiments are small steps towards an entirely new paradigm in scholarly communication. Already today, we see that on born-online platforms (such as "PLOS") the line between journals is disappearing. In the digital space, there is no need for volumes, issues or even page numbers. And as most research has moved online, articles are increasingly cross-referenced through direct (DOI) links. What was once a print journal – or a printed monograph – has moved through the database age, and is now gradually integrated into an online scientific community in which the exchange of information is continually supported by a social network of peers – and validated by the publisher. As AI and machine learning technologies further develop in the coming years, computers will be able to automatically arrange and instantly present information that is relevant to a specific user. The sharing of datasets through social networks might even give way to an environment in which an experiment can be simulated digitally, leading to research becoming more efficient and focused. Finally, blockchain has the potential to strengthen and modernize peer review, reinforcing the crucial importance of securing trustworthy information in STM publishing.

As happened during the early years of the Internet, many of the predictions we make today will not hold up in the long run, or they will take on a different form than we initially expected. Whatever the future may hold, the innovations publishers are introducing today point towards a near future that not only allows us to "surf" efficiently, but offers equal opportunity to "drill" deeper. Instead of being constrained by the physical shape or length of a book, through innovations like Topics we will be able to delve into a topic drawing from an endless list of sources that are tailored, arranged and

presented to meet the needs of any researcher. Most importantly, the publisher has ensured that information is curated, validated, and reliable through blockchain peer-review.

Whether someone would like to get a quick grasp of an emerging field or take in only the most minute discoveries within their own – it will all be just a click away, saving scientists precious time and effort they would rather spend on what they do best: Research.

References

[1] *Artificial Intelligence Index 2017 Annual Report*, p. 41. Available online via: https://aiindex.org/2017-report.pdf (retrieved 26.9.2018)

[2] van Rossum J,:*Blockchain for Research. Perspectives on a New Paradigm for Scholarly Communication* (London: Digital Science, 2017), p. 2. Available online via https://figshare.com/articles/_/5607778 (retrieved of 26.9.2018)

Dr. Michiel Kolman
Senior Vice President,
Elsevier
Radarweg 29, 1043 NX
Amsterdam,
The Netherlands
Phone +31 204 853 046
M.Kolman@elsevier.com
www.elsevier.com

President
International Publishers Association
23, Avenue de France
1202 Geneva, Switzerland
Phone +41 22 704 18 20
president@internationalpublishers.org
https://www.internationalpublishers.org
ORCID 0000-0001-7394-6110

Sjors de Heuvel
Publisher Relations Manager, Elsevier
Radarweg 29, 1043 NX
Amsterdam, The Netherlands
Phone +31 6 8269 5176
www.elsevier.com
s.heuvel@elsevier.com

Chief of Staff to the President, International Publishers Association, 23 Avenue de France
1202 Geneva, Switzerland
https://www.internationalpublishers.org
ORCID 0000-0003-2174-3420

Citation: Kolman M, de Heuvel S:"Yes Dave. Happy to do that for you.". Infozine **2018**, Special Issue 2, 36–37
DOI 10.3929/ethz-b-000297321

Published: November 15, 2018

Stefan Geißler
Expert System Deutschland GmbH

Trends in scientific document search

One of the ironic observations regarding search today is that in many large corporate environments, search capabilities used by professionals in their daily work are lightyears behind what teenagers today take for granted when using popular platforms on the internet (Google, Amazon, Spotify and the like): Typo-tolerant search, semantic abstraction across synonyms or subterm-superterm relations are still largely absent in many places. As a lot of corporate scientific search requires confidentiality (the company may not want the topics of their searches to become public, let alone the company-internal documents on which it is performed) public platforms dedicated to scientific search such as Google Scholar, Semantic Scholar (www.semanticscholar.org/) or PubMed are not always the complete answer.

This short essay discusses some ingredients to (scientific) document search that should be assessed and considered when planning to update a search environment.

Semantic abstraction, allowing to bridge the gap between the terms used in the user's query and the terms in the relevant documents is perhaps the most beneficial extension beyond simple string matching that search should offer: A user looking for information on "laptops" expects to find matches also when they talk about "notebooks"; what is called "lesion" in some documents might be called "injury" in others. Accounting for these term relations by means of adding a thesaurus is a well-established practice in many domains: A search for a given term is extended automatically to this term's synonyms or subterms. Especially in the medical field, resources like the Medical Subject Headings Thesaurus (NLM's MeSH, www.nlm.nih.gov/mesh) are used extensively to facilitate search. Defining and maintaining a large thesaurus however is a complex project and many smaller domains lack a thesaurus like MeSH to this date.

A relatively new approach to allow search environments to handle term mismatches are the so-called word embeddings [1]. Word embeddings allow to assess term term similarities by comparing the typical contexts of terms as observed in large document collections. Since this does not require manual annotation, just raw text, and since efficient and free implementations of the respective algorithms exist (e.g [2]), word embeddings have become very popular and have established themselves as a kind of de facto standard processing steps for many NLP-related tasks. Pre-computed resources (i.e. the vectors that represent the word embeddings) trained on huge public corpora are freely available (e.g. [3]) and they can with moderate computing power be extended with (or calculated from scratch on) one's own document collections. For a more in-depth introduction to the concept of word x embeddings see [4].

While word embeddings exhibit some striking properties [5], the NLP community sometimes jokingly declares that it is almost illegal to talk about word embeddings without mentioning the famous "king – man + woman = queen" example (that shows that using vector representations, lexical semantics can to some extent be expressed as vector algebra) it is also important to be aware of the limitations of the approach: Word embeddings are good at detecting word similarities but they often have a hard time distinguishing different kinds of relatedness: a term like "diabetes" has a paradigmatic relation to terms like "obesity" or "Crohn's disease" in that all these are medical conditions whereas it has syntagmatic relations to terms like "insulin" or "sugar" in that these terms tend to cooccur with "diabetes". A user searching for "diabetes" however, might be confused to see her query extended in the background with "Crohn's disease". Word embeddings in search environments must be used with care in order to account for these effects [6].

Word embeddings are an impressive approach to semantic word relations but the requirement to enhance the reach and accuracy of scientific search doesn't stop on the word/term level: Many relevant questions that keep users busy, are concerned with specific relations between concepts and entities. An information demand such as "Show me evidence where the administration of estradiol to women of age 50 and beyond lead to decreased bone mineral density!" involves a host of analysis requirements that are way beyond term matching approaches: Properly handling this query would need to account for find a relation between the "administration" and the administered substance "estradiol" as well as between the administration and the observed effect (reduced bone mineral density). Search requirements like that can be interpreted as textual entailment tasks (find documents where the content entails the relations expressed in the query. It doesn't come as a surprise that also on this type of tasks, deep-learning inspired approaches have led to impressive progress recently: The best reported results on the SNLI corpus [7] with ~90% accuracy have been obtained by a sophisticated neural network [8].

These results are highly impressive, given that they address a complex task (deciding whether or not a sentence is semantically entailed in another or not) without prior and manually coded world knowledge. Yet the results are possible largely thanks to the huge SNLI corpus of more than half a million of hand annotated training samples.

Preparing training corpora in commercial projects on new tasks, however, often requires considerable resources in time and money and therefore makes the application many machine learning approaches challenging. There is reason to assume that task-specific approaches to complex search requirements will continue to benefit from NLP-inspired methods. An example of this NLP-driven search environment is the work done at Semiring [9] where legal documents are analyzed, collecting relations between the involved concepts and entities and the resulting collection of facts is fed into a graph database for later search and inference. Ontological knowledge (the CEO of a company has to be of type human) can be added and used to flag conflicting assertions and resolve ambiguities.

Regarding the title of this issue of this publication "Surfing and drilling in the modern scientific world" we can conclude that often both is necessary: Surfing where a user is taken from an initial concept to related topics he or she may not initially have had in mind, as well as drilling, where with the help of both quantitative as well as symbolic methods, searches can be made more complete and more focused at the same time. One exciting aspect of today's landscape around these topics is the immense wealth of established methods, algorithms, libraries and resources that are available to jump start specific search projects: State of the art deep learning libraries (Keras, Torch, TensorFlow), powerful NLP platforms (SpaCy) as well as precomputed models allow implementers to enter a search project "one level up", benefitting from a technology stack which a few years ago would have been unthinkable.

References

[1] Bengio Y et al: Neural Probabilistic Language Models. In: Holmes D.E., Jain L.C. (eds.) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, 2006, p. 194. Springer, Berlin, Heidelberg DOI: 10.1007/3-540-33486-6_6

[2] https://radimrehurek.com/gensim, retrieved 30.9.2018

[3] https://github.com/Hironsan/awesome-embedding-models, retrieved 30.9.2018

[4] http://ruder.io/word-embeddings-1, retrieved 30.9.2018

[5] Shperber G: A gentle introduction to Doc2Vec, https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e, retrieved 30.9.2018

[6] https://luminoso.com, retrieved 30.9.2018

[7] https://nlp.stanford.edu/projects/snli, retrieved 30.9.2018

[8] Kim S, Hong, J-H, Kang I, Kwak N: Semantic sentence matching with densely-connected recurrent and co-attentive Information https://arxiv.org/pdf/1805.11360.pdf

[9] Cavar D, Herring J, Meyer A: Case Law Analysis using Deep NLP and Knowledge Graphs, Proceedings of the LREC 2018, PDF http://lrec-conf.org/workshops/lrec2018/W22/pdf/7_W22.pdf, retrieved 17.10.2018

Stefan Geißler
Expert System
Deutschland GmbH
Blumenstr 15
69115 Heidelberg
Germany
Mobile: +49 174 6595713
skf.geissler@googlemail.com
http://www.linkedin.com/in/stefangeissler

Citation: Geißler S: Trends in scientific document search. Infozine **2018**, Special Issue 2, 38–39
DOI 10.3929/ethz-b-000297324

Published: November 15, 2018

Jane Reed
Linguamatics

Power tools for text mining

Handling the unstructured data overload with augmented intelligence

Ask someone a question they don't know the answer to, say, how far is it to Tipperary, and chances are they'll suggest you should "google it" – certainly that's what the teenagers in my household would say. In daily life, search seems to be sorted, whether that's a particular answer to a specific question, or general information of suggestions about a topic, ranging from appendicitis to holiday resorts.

So why do we struggle so much with search in our professional lives? Finding the right information for decision support in research, whether in academia or industry, is a bigger challenge. In fact, the market research firm IDC estimates that knowledge workers spend almost nine hours each week searching for information needed for their work [1]. This includes scientists searching for knowledge to advance the development of new drugs, or help identify life-preserving gene therapies, or ensure compliance with the latest regulatory rules.

There are several issues at the heart of this problem. First, of course, is the amazing increase in the volume, variety, and velocity of data out there to search (the classic 3 Vs of "big data" [2]). And of course, much of this is unstructured data (text, video, images [3]) which makes it hard search and analyse using traditional manual methods. But let's not lose sight of another issue – and that's the quality of the answers we are trying to achieve. Again, here, professional scientific search differs. To answer the question, "what genes are involved in breast cancer", you want to get a comprehensive list, not just find the first 10 or 15 documents that contain some useful information.

Exactly what tools are best to use to find what you are looking is another issue. Do you need to drill deep into the literature, to find deeply-buried infrequent snippets of information? Or, plough across a broad landscape of text, to gather and amalgamate nuggets scattered far and wide? What power tools do you need?

This is where there is a need for new technologies to assist the human in their efforts. Artificial intelligence (AI) is one of the current buzz words, and Natural Language Processing (NLP) is an AI technology, that enables written text to be interpreted, and rapidly transforms the key content in text documents into quantitative, actionable insights.

Linguamatics provides NLP solutions (I2E [4], iScite [5]) that researchers and clinicians can use in their search across unstructured text. In many ways these can be considered augmented intelligence [6] – as the user can decide whether to build search rules that drill deep into literature for hidden nuggets, or search broadly, ploughing the landscape for the desired information. There's no black box here, the user is in complete control of their search.

To give you a couple of examples, I'll share two use cases. The first one is using NLP to drill deep, to search out and extract the nuggets of information around a particular rare disease, needed for precision medicine.

Shire's use of NLP to uncover genotype-phenotype associations in rare disease

Shire Pharmaceuticals have developed an enzyme replacement therapy for a rare disease, Hunter Syndrome, but this needs to be delivered to the cerebral spinal fluid via an implant device. This is a potentially lifechanging intervention, yet invasive and unpleasant for young patients. Shire wanted to find a reliable way to identify patients who had the greatest potential to benefit.

A text mining project used I2E to extract all the associations reported in full text literature between the relevant gene (IDS, iduronate-2-sulfatase), any mutation or variant, and phenotype descriptions for neurocognitive impairment. As with many rare diseases, the information is very sparse. The researchers designed a set of NLP queries, and used I2E's powerful mutation ontology, to ensure that however authors described the gene, the allelic variation, or the phenotypic outcome, these nuggets of information could be captured. The result was a set of prognostic genetic markers that enabled clinicians to make informed decisions on which infants would benefit from enzyme replacement therapy.

> *"Text mining was remarkably successful. Results were significantly better than any genetic database of reported genotypes available."*
> *Madhu Natarajan, Director, Systems Pharmacology, Shire Pharmaceuticals*

The second use case is using NLP to plough across a broad field of patent literature, extracting key snippets of information about kinase assay technology.

BMS's use of NLP for patent landscape of kinase assay technology

Bristol Myers Squibb needed to strengthen their internal kinase screening technology [7], with the first step being to analyse industry trends and benchmark BMS' capabilities against other pharmaceutical companies, with key questions including: What are the

kinase assay technology trends? for different therapeutic areas? used by the big pharmaceutical companies?

The BMS team used I2E to create effective search queries to extract key information for 500 kinases, 5 screening technologies, 5 therapeutic areas, and across 14 pharmaceutical companies. Use of I2E allowed queries to be designed using domain specific vocabularies, for example using over 10,000 synonyms for the kinases, hugely improving the breadth and recall of the patent searches.

Using this approach, the patent analysis team extracted information from over 7100 full text patents. To put this into perspective, it takes ~1h to manually read one patent for data extraction and a scope this large would require around 175 person-weeks (over 3 years) to accomplish. Using NLP to get this broad landscape of information took 2 patent analysts 3 months (i.e. about 25 weeks) which is a 7-fold saving in FTE time.

NLP augments human intelligence

In these examples, NLP augments the human brain. The human brain designs the search and extraction strategy, as deep or as broad as is needed for the business case. NLP cannot replace the thought processes required for decision making; but it can hugely accelerate the gathering of the information, and transformation of that material into actionable insights.

References

[1] Feldman S, Sherman C: The High Cost of Not Finding Information. An IDC White Paper (link to PDF, retrieved 29.9.2018)

[2] Shafer T: The 42 V's of Big Data and Data Science (link to website, retrieved 29.9.2018)

[3] Walker M: Structured vs. Unstructured Data: The Rise of Data Anarchy (website)

[4] https://www.linguamatics.com/products-services/about-i2e

[5] https://www.linguamatics.com/iscite

[6] Williams S, Wigley C: The Real AI: Augmented Intelligence (link to website, retrieved 29.9.2018)

[7] Yang YY, Klose T, Lippy J, Barcelon-Yang CS, Zhang L: World Patent Information **2014**, 39, 24–34 DOI: 10.1016/j.wpi.2014.09.002

Dr. Jane Z. Reed
Head of Life Science,
Linguamatics
Cambridge Science Park,
Milton Road,
Cambridge
CB4 0WG, UK
Phone +44 1223 651 910
jane.reed@linguamatics.com
ORCID 0000-0002-6773-6237
http://www.linguamatics.com

Citation: Reed J: Power tools for text mining. Infozine **2018**, Special Issue 2, 40–41 DOI 10.3929/ethz-b-000294371

Published: November 15, 2018

Paul Peters
American Chemical Society

Publishing and patenting: Navigating the differences to ensure search success

With more of academic research being funded by third parties and a trend by larger companies to outsource research work to academia, scientists at universities are often going for a dual approach to publishing – once in the highest impact journal that would align best with their research work and once with the patent office(s) that provide the necessary Intellectual Property (IP) rights linked to the countries where the invention could potentially be brought to market. Obviously, scientists need to plan these two publications carefully since the journal publication could invalidate the patent if it is published before the patent is filed at the first patent office.

In terms of content and searchability, journal articles and patents could not be more diverse. Clearly the journal article is written to have an attractive title, a meaningful abstract both inviting the reader to get the full paper and cite it as a relevant document in future, related publications. The patent however, is commonly written to conceal as much as possible for the exact invention, its conditions and commercial potential of this invention. Numerous patents have a title, like *new process* or *new compound*. Even though the patent offices have clear rule about informative titles and abstracts, the skillful patent attorney will draft the document to provide the broadest IP protection against the least amount of disclosure.

This has great consequences in searching for scientific information. Patents are normally the first publications where new compounds are disclosed and where new concepts are described. But the searcher needs to be skillful to find this in a text that could be in many languages, written by attorneys. The meaningful information that journal articles provide in the titles and abstracts must be looked for in the claims and in the text of the examples of the patent. Let's examine these two publications written by scientists from ETH Zurich found in SciFinder™: (*Figure 1*)

1. **Transglutaminase mediated high molecular weight hyaluronan hydrogels**
Quick View PATENTPAK
By Broguiere, Nicolas; Cavalli, Emma; Zenobi-Wong, Marcy
From PCT Int. Appl. (2017), WO 2017191276 A1 20171109. | Language: English, Database: CAPLUS

2. **Factor XIII Cross-Linked Hyaluronan Hydrogels for Cartilage Tissue Engineering**
Quick View Other Sources
By Broguiere, Nicolas; Cavalli, Emma; Salzmann, Gian M.; Applegate, Lee Ann; Zenobi-Wong, Marcy
From ACS Biomaterials Science & Engineering (2016), 2(12), 2176-2184. | Language: English, Database: CAPLUS

Figure 1

(57) Abstract: The invention relates to a process for forming a hyaluronan hydrogel, comprising the steps of a. providing a hyaluronan donor peptide conjugate and a hyaluronan acceptor peptide conjugate each represented by a general formula (I), wherein 10% of R^1 moieties are represented by a general formula (II), wherein L is a 2 to 6 atom linker moiety and Pep is a transglutaminase donor or acceptor peptide, and the rest of R^1 moieties are represented by –COOH. b. adding a factor XIII polypeptide and a thrombin polypeptide, or a factor XIIIa polypeptide. The invention further relates to compositions and hydrogels characterized by the depicted chemistry.

Figure 2a

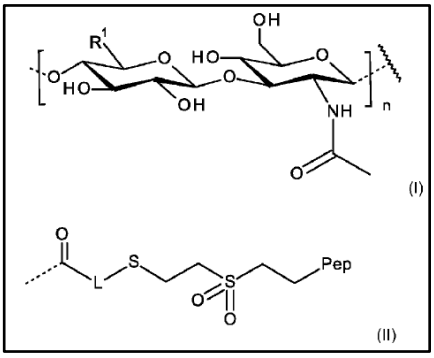


Figure 2b

The research on these hyaluronan hydrogels seems to be done by the same authors (a few more listed for the American Chemical Society (ACS) journal), but the title is very different. How do the two original abstracts compare (first the patent with its Markush-like structure (*Figure 2 and b*) and second the journal article) (*Figure 3*):

ABSTRACT: In this study, transglutaminase-cross-linked hyaluronan (HA-TG) hydrogels are investigated for their potential to treat cartilage lesions. We show the hydrogels fulfill key requirements: they are simultaneously injectable, fast-gelling, biocompatible with encapsulated cells, mitogenic, chondroinductive, and form a stable and strongly adhesive bond to native cartilage. Human chondroprogenitors encapsulated in HA-TG gels simultaneously show good growth and chondrogenesis. Strikingly, within soft gels (~1 kPa), chondroprogenitors proliferate and deposit extracellular matrix to the extent that the hydrogels reach a modulus (~0.3 MPa) approaching that of native cartilage (~1 MPa) within 3 weeks. The combination of such an off-the-shelf human chondroprogenitor cell source with HA-TG hydrogels lays the foundation for a cell-based treatment for cartilage lesions which is based on a minimally invasive one-step procedure, with improved reproducibility due to the defined cells and with improved integration with the surrounding tissue due to the new hydrogel chemistry.
KEYWORDS: *cartilage, hyaluronan, chondroprogenitors, transglutaminase, bioadhesive*

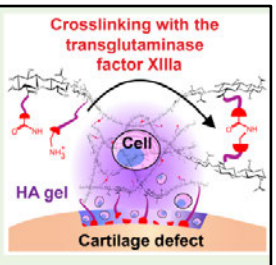


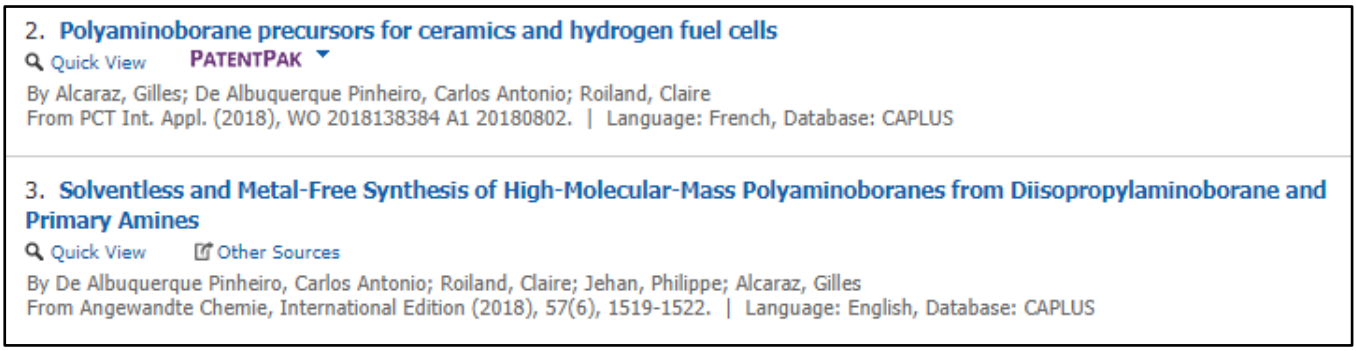
Figure 3

There is a huge difference in these two abstracts, dealing with very similar research work. When we would search on the combination of transglutaminase and hyaluronan hydrogels in Google


1. 2H-1,2,3-Triazole-Based Dipeptidyl Nitriles: Potent, Selective, and Trypanocidal Rhodesain Inhibitors by Structure-Based Design
Quick View Other Sources
By Giroud, Maude; Kuhn, Bernd; Saint-Auret, Sarah; Kuratli, Christoph; Martin, Rainer E.; Schuler, Franz; Diederich, Francois; Kaiser, Marcel; Brun, Reto; Schirmeister, Tanja; et al
From Journal of Medicinal Chemistry (2018), 61(8), 3370-3388. | Language: English, Database: CAPLUS

2. N-1-Cyanocyclopropyl 2-aminoalkanamides as trypanosome inhibitors and their preparation
Quick View PATENTPAK
By Giroud, Maude; Haap, Wolfgang; Kuhn, Bernd; Martin, Rainer E.
From PCT Int. Appl. (2017), WO 2017089389 A1 20170601. | Language: English, Database: CAPLUS

Figure 4: Collaboration between ETH Zurich and Hoffmann-La Roche


2. Polyaminoborane precursors for ceramics and hydrogen fuel cells
Quick View PATENTPAK
By Alcaraz, Gilles; De Albuquerque Pinheiro, Carlos Antonio; Roiland, Claire
From PCT Int. Appl. (2018), WO 2018138384 A1 20180802. | Language: French, Database: CAPLUS

3. Solventless and Metal-Free Synthesis of High-Molecular-Mass Polyaminoboranes from Diisopropylaminoborane and Primary Amines
Quick View Other Sources
By De Albuquerque Pinheiro, Carlos Antonio; Roiland, Claire; Jehan, Philippe; Alcaraz, Gilles
From Angewandte Chemie, International Edition (2018), 57(6), 1519-1522. | Language: English, Database: CAPLUS

Figure 5: Research work from University de Rennes and CNRS (National Center for Scientific Research, France): (patent published in French)


1. Hetero Diels - Alder crosslinker and their use in reversible cross-linking polymer systems
Quick View PATENTPAK
By Schmidt, Friedrich Georg; Paulmann, Uwe; Richter, Christian; Inhestern, Marcel; Meier, Christian; Barner-Kowollik, Christopher; Pahnke, Kai; Sanz, Miguel Angel; Umbreen, Sumaira; Janssen, Christian Ewald
From PCT Int. Appl. (2017), WO 2017129483 A1 20170803. | Language: German, Database: CAPLUS

2. A mild, efficient and catalyst-free thermoreversible ligation system based on dithiooxalates
Quick View Other Sources
By Pahnke, Kai; Haworth, Naomi L.; Brandt, Josef; Paulmann, Uwe; Richter, Christian; Schmidt, Friedrich G.; Lederer, Albena; Coote, Michelle L.; Barner-Kowollik, Christopher
From Polymer Chemistry (2016), 7(19), 3244-3250. | Language: English, Database: CAPLUS

Figure 6: Collaboration between Evonik and the Karlsruhe Institute of Technology (patent published in German)

Scholar, we get 3750 answers (including patents but not citations). The ACS paper is found as the third answer on page one, but the patent doesn't show up in the first 10 pages of results.

Conducting this search in SciFinder from Chemical Abstract Services (CAS) produces 6 results including the ACS paper and the patent from the same ETH Zurich research group. In SciFinder, journals and patents are treated very similarly as a publication of scientific information. For patents, 95% of the titles and abstracts are rewritten to provide more useful information than the original titles and abstracts. Relevant concepts from the patent claims and examples as well as the CAS Registry Numbers™ for the disclosed compounds, materials and biosequences are added to the record that is searchable in one interface.

For these two publications, "Chondrocyte", "Compressive modulus", "Hydrogels" and "Shear modulus" were added as systematic keywords. The patent was also indexed with the keyword "Cartilage" and with "Mesenchymal stem cell", while for the

journal article "Cartilage formation" and "Stem cell" were added. In total, 52 unique CAS Registry Numbers for peptides, small molecules and other compounds were added. There were only 5 that were assigned to both the journal article and the patent. The patent had 43 additional unique compounds indexed and the journal article with 5 additional unique compounds. The difference in indexing comes from a different description of the experiments. Patents try to expand the scope of the invention to maximize their application.

When ETH Zurich scientists opt to patent their work, they usually go directly for a WO patent application that can then be extended into various national and regional patent offices. If we look at other universities in Germany or France, we see that they tend to first apply for a national patent application, followed by a European or WO application. Since EP (European Patent Office) or WO (World Patent Office) patent applications can also be published in German or French, most these patents are not published in English, making them even harder to find. Additionally, WO patent applications may also be published in Chinese, Japanese, Korean, Russian, Spanish, Portuguese, Italian and Arabic. This would have even larger consequences for finding them.

Here are some more examples of scientific work that was sent to the patent office before it was sent to the ACS or other major publisher (*Figure 4–6*).

In the last two examples, it appears that the journal articles were published *before* the patents, but for the patent it is the priority application date that needs to be before the web publication date of the journal article. Patents are usually filed 18 months before their application is published.

If we look at the articles published in ACS journals in the last 18 months (*Figure 7*) that describe the synthesis of compounds, we see that only 35% of the first author affiliations represent a country where English is the native language (USA, UK, Canada, Ireland, Australia, etc). That leaves 65% of the primary authors in ACS journals who are likely to be publishing their patent application in a language other than English. The top countries here are China, Japan, Germany, Korea, France, Spain. Fortunately, in tools like SciFinder, all patents documents from any of the 63 patent offices that CAS covers get an English title and abstract. This important work done by CAS scientists allows researchers at ETH Zurich to keep a close eye on the discoveries published and patented from China and other Asian countries for which we have seen a large increase in volume and a higher level of quality.

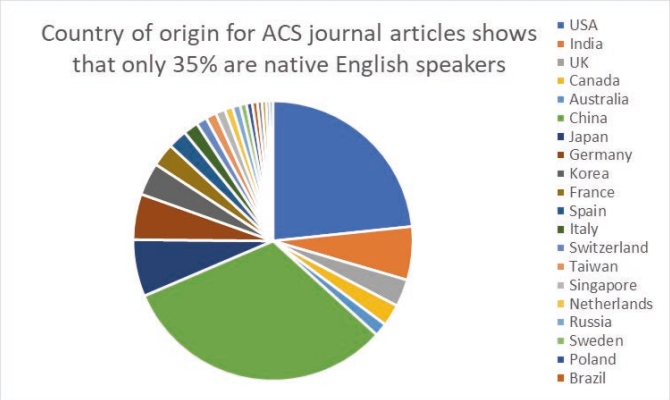


Figure 7

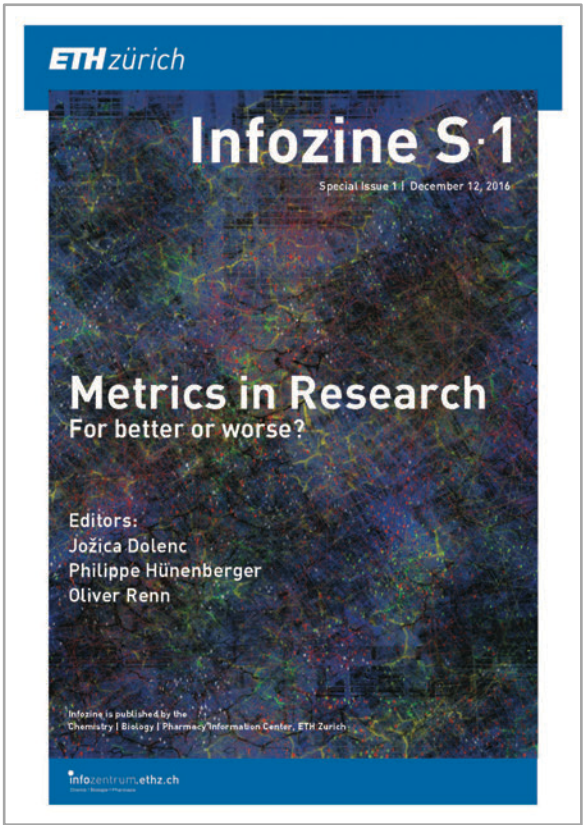
In conclusion, scientists need to be aware that publications in peer-reviewed journals would not automatically be free from any IP rights. The difference between journal articles and patents are extensive, which has a large impact on how they need to be searched. Tools that provide a unified approach to journals and patents in terms of language, titles, abstracts and indexing of keywords and chemical compounds provide a big advantage over other tools.

Paul Peters
ACS International
Alte Döhrener Strasse 50a
30173 Hannover
Germany
ppeters@acs-i.org
www.cas.org
ORCID 0000-0002-7964-9958

Citation: Peters P: Publishing and patenting: Navigating the differences to ensure search success. Infozine **2018**, Special Issue 2, 42–44
DOI 10.3929/ethz-b-000294376

Published: November 15, 2018

Infozine Special Issues Series



Infozine Special Issue 1

Metrics in Research: For better or worse?

has been published December 12, 2016
and is still relevant:

If you want to read many additional interesting comments and editorials on metrics in science, the recent Infozine from ETH Zurich is a fun read.
Jonathan V. Sweedler,
Editor of the ACS Journal Analytical Chemistry

Jonathan V. Sweedler: Metrics to Evaluate Journals, Scientists, and Science: We Are Not There Yet. Anal. Chem. 2017, 89, 5653–5653, DOI 10.1021/acs.analchem.7b01872

Download the free PDF at www.infozentrum.ethz.ch