Communication-Theoretic Design Insights for NextG Infrastructure
From mmWave Systems to Robust Deep Learning

Upamanyu Madhow ("Madhav")

ECE Department

University of California, Santa Barbara

UCLA, February 2024

Funded by NSF, DARPA/SRC JUMP and JUMP 2.0

Key Elements of NextG Infrastructure

- MultiGigabit/s communication
  - With pervasive availability
- Robust, high-resolution sensing
  - At home, on the road

→ Millimeter Wave/THz Systems

Interplay of comm theory, signal proc algorithms, hardware constraints

- Pervasive AI
  - Cloud to edge
  - Invisible plumbing to Chat GPT

→ AI Fundamentals
Robustness & Interpretability

Comm-theoretic inspiration for shaping deep neural networks for robustness

NextG Comm & Sensing (aka mmWave/THz)

The mmWave →THz frontier

- "Unlimited" bandwidth
  - mmWave: Licensed (28 GHz), unlicensed (60 GHz)
  - Towards THz (100+ GHz), regulation TBD
- Tiny wavelengths → miniaturized antenna arrays
- Unique propagation characteristics
- Silicon RFICs, low-cost packaging

mmWave at UCSB (2005-2017): a sampling

NSF
QCOM, Samsung, Nokia
FB, Google

Directional Networking

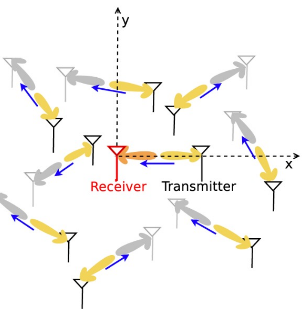
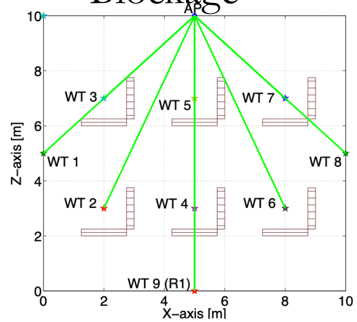


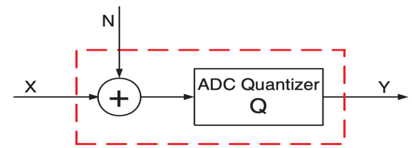
Fig. 2. Network model for interference analysis.

Blockage



(b) Office

ADC Fundamentals



mmWave Picocells

Modeling, protocols, capacity

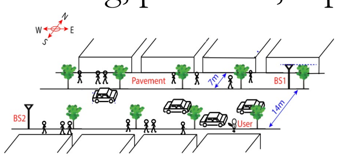
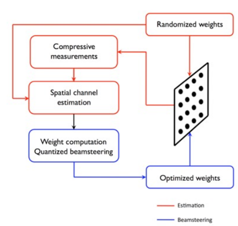


Fig. 1: Picocellular network deployed along an urban canyon

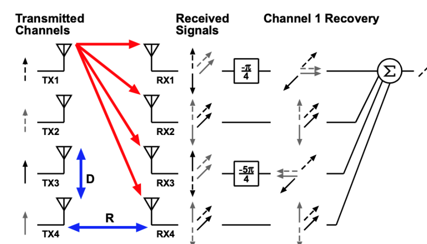
Compressive Estimation

Fundamentals, Algorithms, Demo



LoS MIMO

Fundamentals & Demo



mmWave Mesh Backhaul

Routing & Resource Alloc

Short-range mmWave Imaging

New models, Proof of Concept

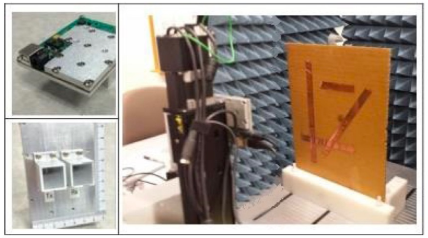


Fig. 5. Experimental data collection using 60 GHz quasi-monostatic radar system.

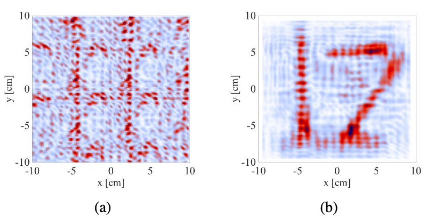
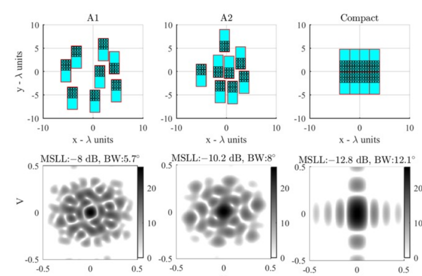


Fig. 8. Sparse array III (a) Point MF (b) Patch MF (1cm × 1cm).

mmWave Sensing

Sensor Geometries, Algorithms



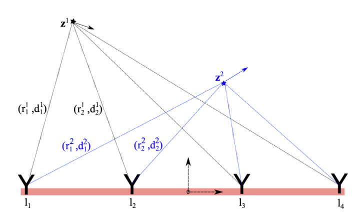


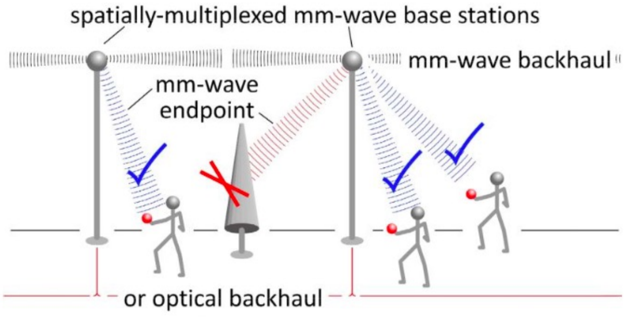
Fig. 1. 2D System model with a linear array of radar sensors placed on x-coordinates, $[l_1, l_2, l_3, l_4]$. The kinematic states z_1, z_2 of two targets are to be estimated using the unordered range and doppler observations from the sensors.

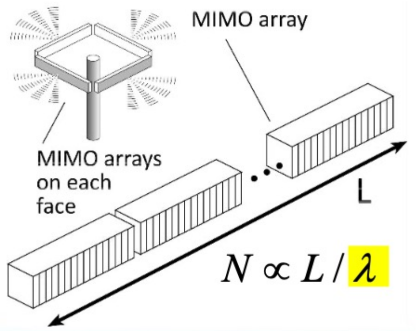
What we knew ~7 years ago

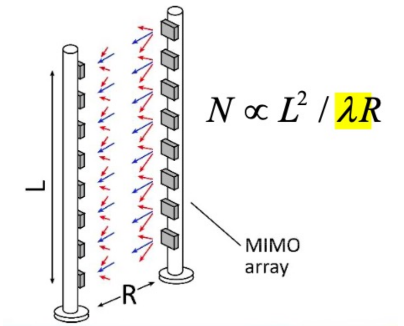
- Sweet spot is at short ranges
  - In-room indoors, ~100 meters outdoors
- Simple models for sparse channels are effective
- Blockage is not a killer: simulations and experiments
- Compressive estimation for efficient channel estimation & tracking
  - New super-resolution algorithms, experimental demonstrations
- LoS MIMO has huge potential: theory and prototyping
- Short-range sensing needs new models and algorithms
  - Patch models for extended objects (theory and experiments)
  - Exploiting geometric constraints

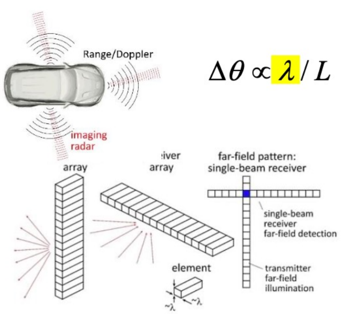
Industry was talking about 5G. What next for academia?

JUMP program: ComSenTer (2018-2022) Communications & Sensing @ Terahertz

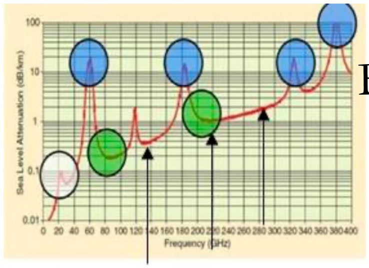








$\Delta\theta \propto \lambda / L$



Band choices avoiding oxygen absorption peaks
(140, 210, 280 GHz)

Can mmWave hardware be scaled to these bands?

Massive increase in #RF chains

Low-cost packaging

Silicon whenever possible, augmented by III/V

How can system designs enable/exploit hardware scale?

Hardware-signal processing co-design

What can we do with lots of antennas?

- Massive MU-MIMO: the most obvious way to push boundaries
  - All-digital ➔ #users scales with #antennas
  - Bottlenecks: RF impairments (nonlinearities, phase noise), ADC precision
  - Secret weapon: channel sparsity, large #antennas
- LoS MIMO
  - Opportunistic deployment, all-digital processing at 10s of GHz bandwidth
- Massive MIMO radar
  - Large arrays to sidestep range versus field of view tradeoffs

2023 onwards: JUMP 2.0 and NSF

- Center for Ubiquitous Connectivity (CUbiC): Optical and wireless
- NSF Rings and 4D100 projects
- Wireless frontiers
  - RF hardware: continue pushing boundaries (higher freqs, #antennas)
  - Signal processing/VLSI: low-power, modularity, scalability
  - Networking: cost-effective dense deployment
  - Sensing: bridging the resolution gap in RF sensing

Who did the work?

Mohammed Abdelghany

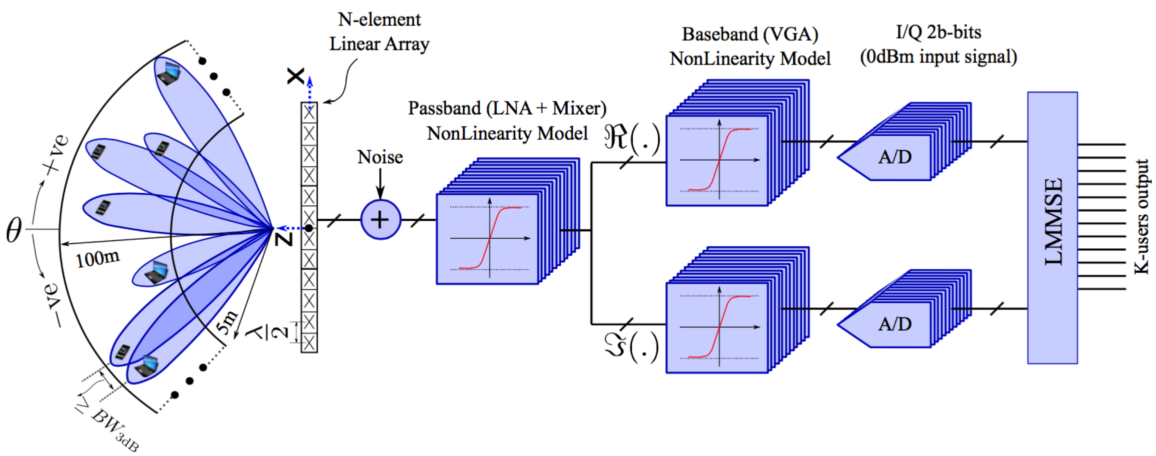
Maryam E. Rasekh

Ahmet Sezer

Lalitha Giridhar

Canan Cebeci

Concept System: Tbps Massive MIMO @140GHz



140 GHz
Picocellular
Uplink
(10 Gbps/user,
100 simultaneous
users)

Key bottlenecks for all-digital architectures

- Need one RF chain for each antenna. Can we relax the specs enough that CMOS works?
- Phase noise is high at millimeter wave and THz. Don't things get worse as we scale to a large number of antennas?
- ADC cost, power consumption and availability is limited as we scale up bandwidth
- Multiuser detection is needed, but classic architectures do not scale

Hardware/signal processing co-design is crucial

Scaling #antennas in all-digital MIMO

- Generic MIMO-OFDM does not scale
- For a large number of antennas and/or large bandwidth, and *generic* space-time block fading channel model, channel estimation is the bottleneck

ISIT 1997, Ulm, Germany, June 29 – July 4

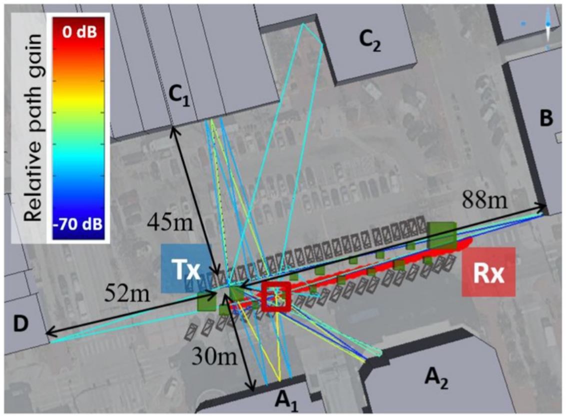
Bandwidth Scaling for Fading Channels

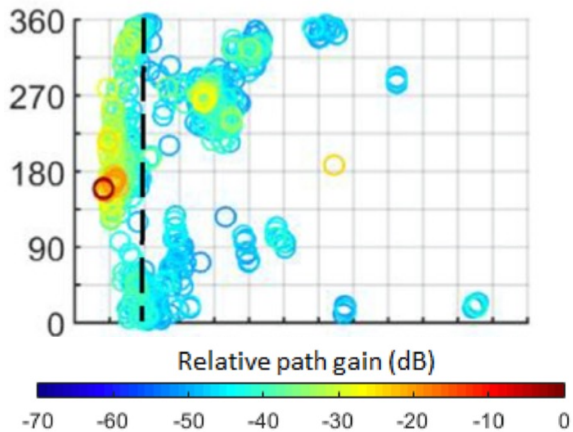
R. Gallager
MIT LIDS
Room 35-207, 77 Massachusetts Ave.
Cambridge, MA 02139
gallager@lids.mit.edu

M. Médard
MIT Lincoln Laboratory
Room C-277, 244 Wood St.
Lexington, MA 02173
medard@ll.mit.edu

- Luckily, the mmWave channel is *not* generic

The mmWave channel is sparse

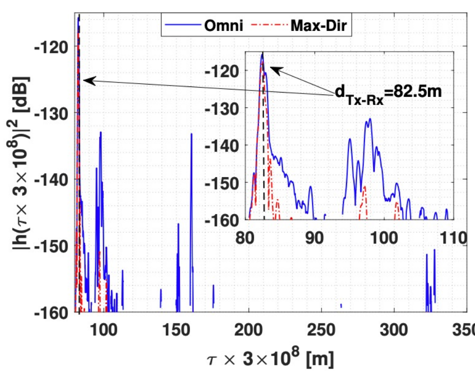


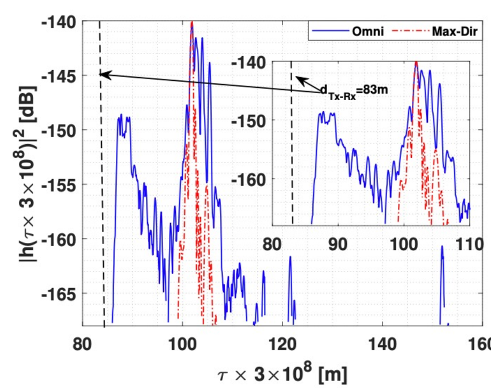


Even at 28 GHz!

(NIST measurements in Boulder, CO. Charbonnier et al, TVT 2020)

Certainly at 140 GHz!





(Molisch group measurements on USC campus, Abbasi et al, TWC, to appear)

Scaling via beamspace

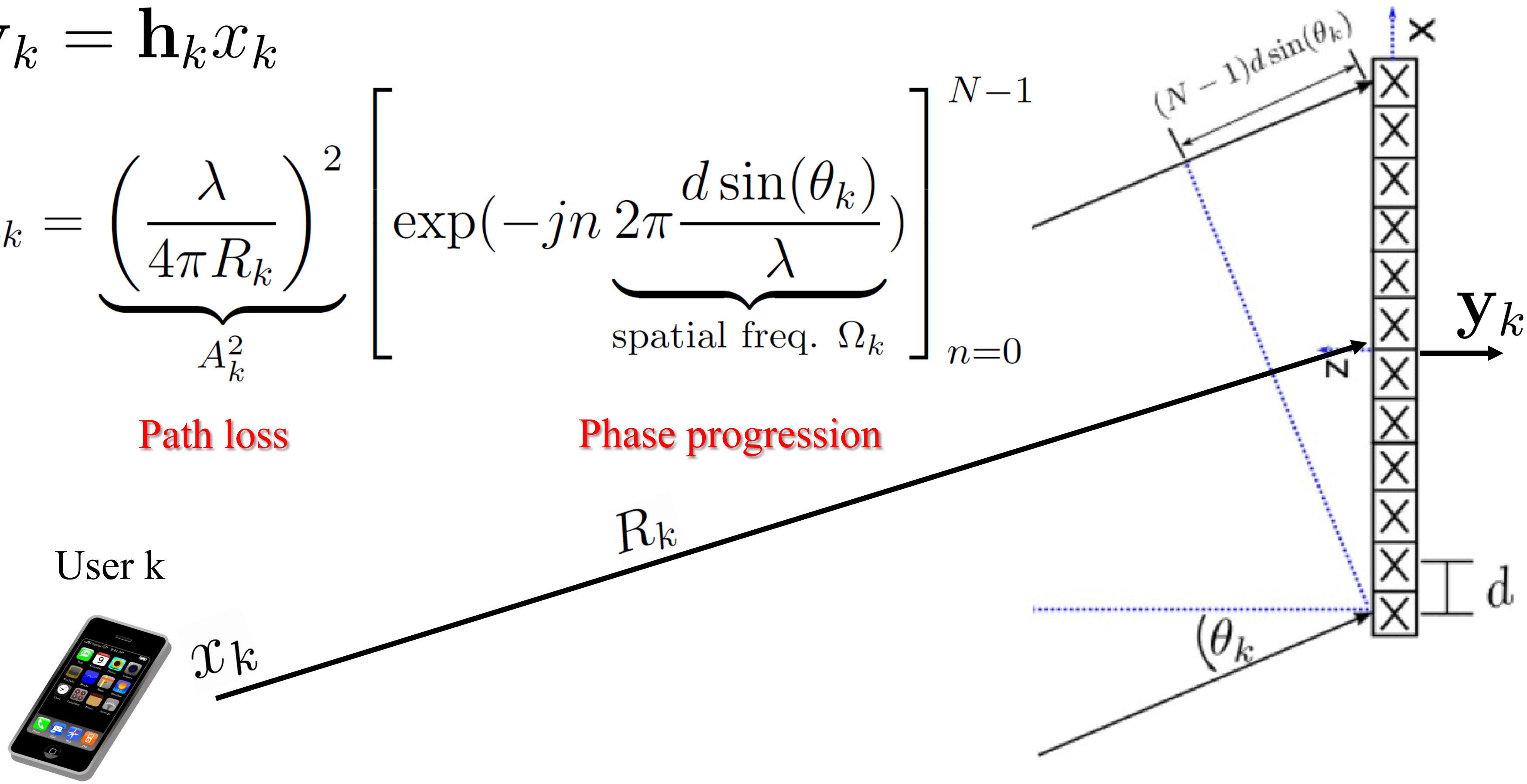
Typical channel: one path is dominant

$$\mathbf{y}_k = \mathbf{h}_k x_k$$

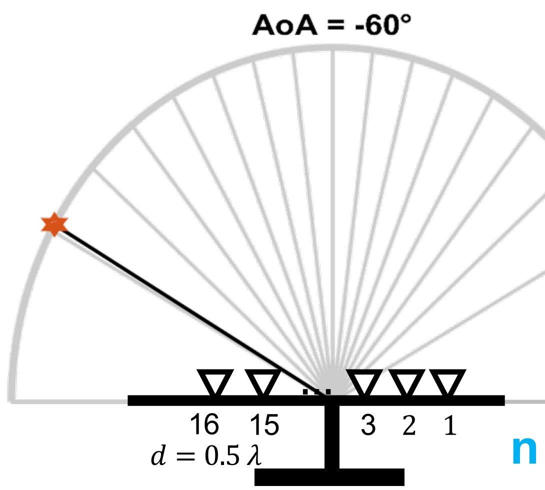
$$\mathbf{h}_k = \underbrace{\left(\frac{\lambda}{4\pi R_k}\right)^2}_{A_k^2} \left[\exp\left(-jn\, \underbrace{2\pi \frac{d \sin(\theta_k)}{\lambda}}_{\text{spatial freq. } \Omega_k}\right) \right]_{n=0}^{N-1}$$

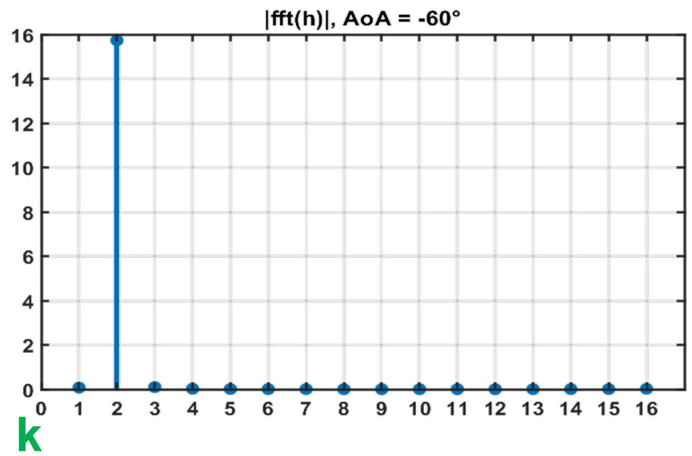
Path loss

Phase progression



Beamspace Representation via spatial FFT

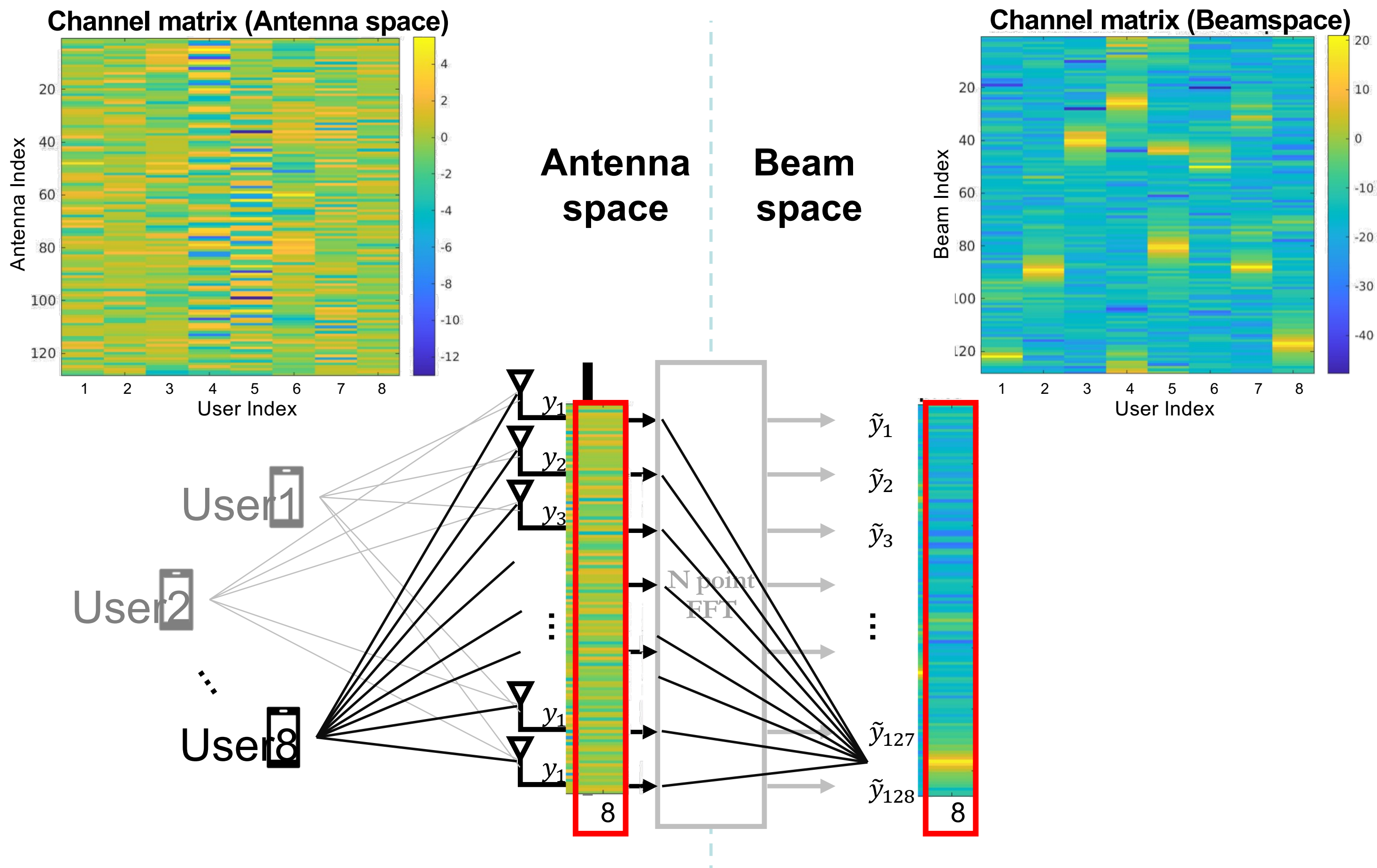




Upfront cost of approximate spatial matched filtering: O(N log N) per sample

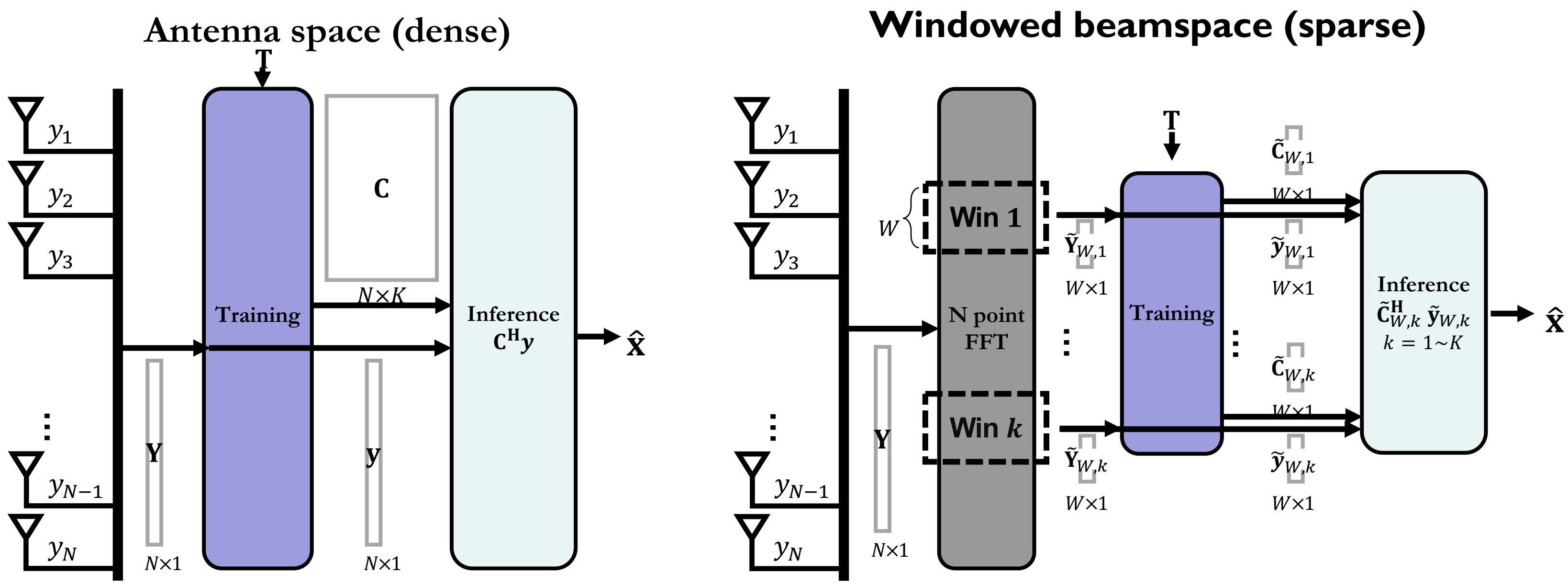
Payoff: Vastly simplified multiuser detection

Channel Sparsity in Beamspace



Each user's energy is concentrated in small beamspace window
➔Reduced-complexity, parallelized demodulation

Antenna space vs Beamspace MU-MIMO



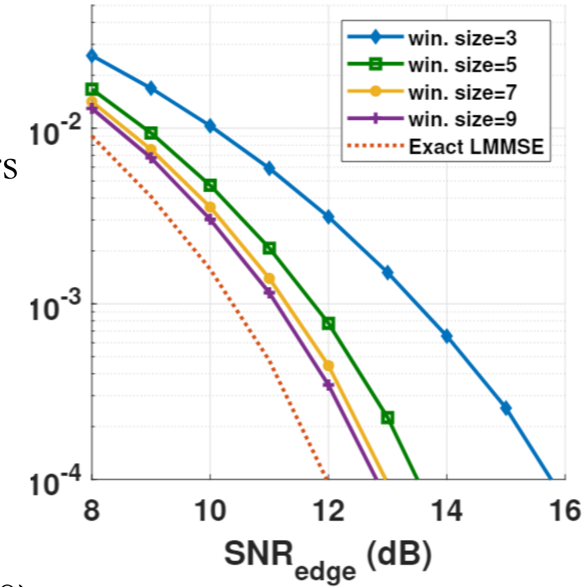
Key observations

- Size of beamspace window does not scale with #antennas
- Window size depends on load factor (#users/#antennas)

➔ Reduced training overhead, reduced complexity

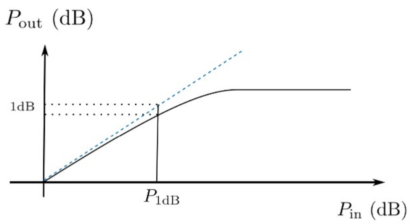
Rethinking multiuser detection in beamspace

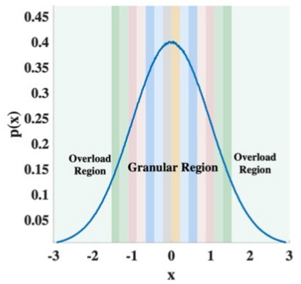
- Local LMMSE detection (SPAWC 2019)
  - Significantly reduces complexity at low load factors

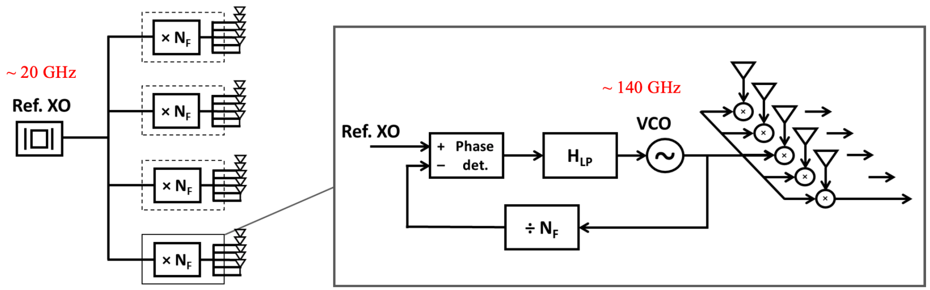


- Nonlinear interference cancellation (SPAWC 2020)
  - SIC on top of LMMSE helps push load factors higher
- Wideband space-time interference suppression (Globecom 2019)
  - Space-time FFT instead of true time delay
- Downlink precoding (Asilomar 2019)
  - Applying uplink-downlink duality

Additional insights on scaling







Scale can be attained with tiling
(phase noise specs can be relaxed)

Severe nonlinearities can be tolerated with scale
(hardware specs can be relaxed)

How far can we go in hardware simplification?
Beamspace with 1-bit ADCs on remote radio heads?

Beamspace with 1-bit ADC

1-bit ADC well matched to low SNR and large # degrees of freedom (classical result: 1.96 dB penalty)

But what if the SNR is too high?

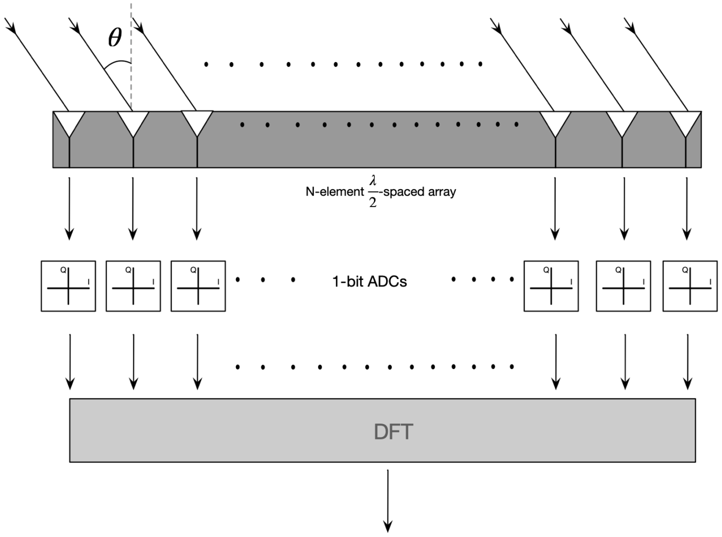
An example worst-case scenario

- Low-cost remote radio heads deployed densely
  - Digital arrays with 1-bit ADC per element
  - Beamspace processing
- Rapidly moving users with limited or no power control
- SNR per element high, #users low

➔ Not enough dithering, input does not look Gaussian

- Need to go back to signals and systems basics

Severely quantized beamspace processing



- Sparse mmWave channel (single path) on linear array

➔ complex exponential spatial response

- All-digital: 1-bit ADC on I and Q
- Input to antennas don't look Gaussian
- ➔ Bussgang decomposition is not accurate

What does the output of the spatial FFT look like?
Can we accurately estimate the channel?
Can we recover the modulated information?

Fourier analysis of "hardlimiting" in space

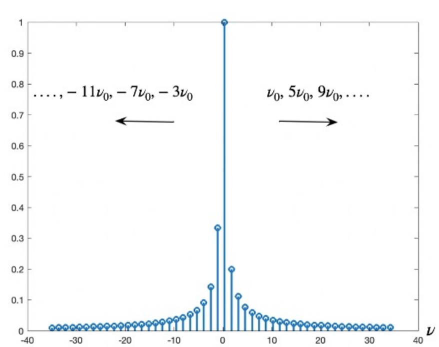
① $[e^{j0}e^{jk}e^{j2k}\ldots\ldots e^{j(N-1)k}] \longrightarrow$ 1-bit Quantizer $\longrightarrow$

① can be written as:

② $s(x) = e^{jkx} \longrightarrow$ Sample at L = 1 $\xrightarrow{s[n]}$ Window of length N $\xrightarrow{s[n]W_N[n]}$ 1-bit Quantizer $\xrightarrow{Q(s[n])W_N[n]}$

We can rearrange the order of the above operations. ② is equivalent to ③ :

③ $e^{jkx} \longrightarrow$ 1-bit Quantizer $\xrightarrow{Q(s(x))}$ Sample at L = 1 $\xrightarrow{Q(s[n])}$ Window of length N $\xrightarrow{Q(s[n])W_N[n]}$



Déjà vu anyone?

34 IEEE TRANSACTIONS ON INFORMATION THEORY January

Hard-Limiting of Two Signals in Random Noise*

J. J. JONES†, MEMBER, IRE

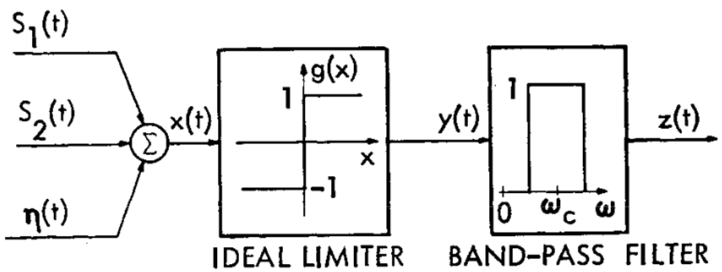


Fig. 1—An ideal band-pass limiter showing an input composed of two signals in random noise.

2 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-15, NO. 1, JANUARY 1969

Hard Limiting of Three and Four Sinusoidal Signals

WILLIAM SOLLFREY

Fourier analysis of "hardlimiting" in space

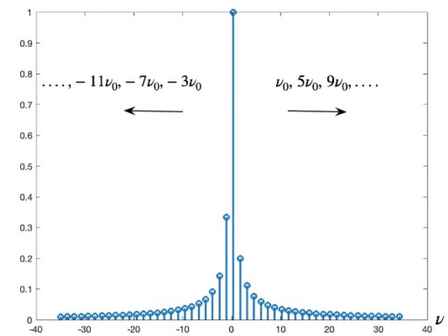
① $[e^{j0}e^{jk}e^{j2k}\ldots\ldots e^{j(N-1)k}] \longrightarrow$ 1-bit Quantizer $\longrightarrow$

① can be written as:

② $s(x) = e^{jkx} \longrightarrow$ Sample at L = 1 $\xrightarrow{s[n]}$ Window of length N $\xrightarrow{s[n]W_N[n]}$ 1-bit Quantizer $\xrightarrow{Q(s[n])W_N[n]}$

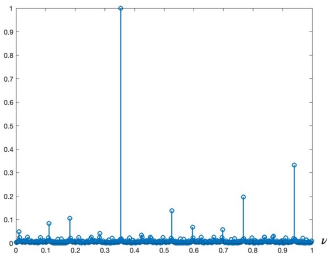
We can rearrange the order of the above operations. ② is equivalent to ③ :

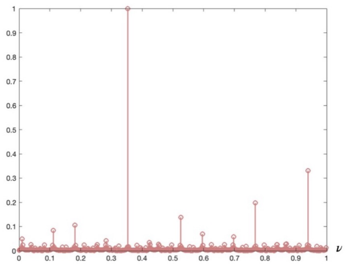
③ $e^{jkx} \longrightarrow$ 1-bit Quantizer $\xrightarrow{Q(s(x))}$ Sample at L = 1 $\xrightarrow{Q(s[n])}$ Window of length N $\xrightarrow{Q(s[n])W_N[n]}$



- Analogous to passband hardlimiter literature from the 1950s (time replaced by space), BUT
- Complex exponentials instead of real-valued sinusoids
- Spatial sampling ➔ aliasing of spatial frequencies
- Spatial windowing ➔ spreading of spatial frequencies
- No passband filter ➔ need some other means of rejecting undesired spatial frequencies

Example conclusions

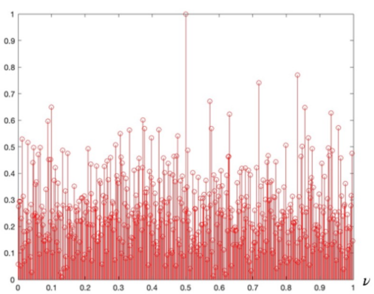


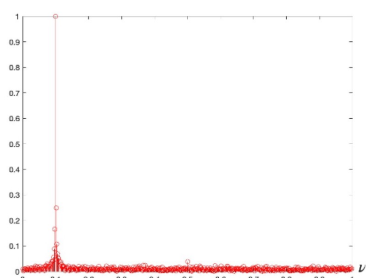


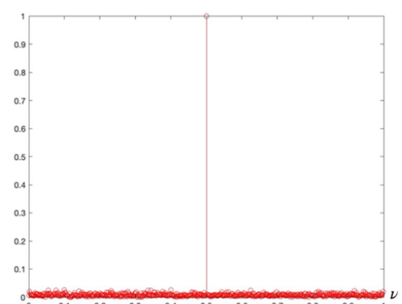
Simulation Analysis

Beamspace outputs for a single user

- Analysis accurately predicts DFT outputs
- ➔ can design based on it
- Design take-away: training sequences with phase ramp to suppress harmonics during channel estimation
- Correlation against QPSK training sequence cannot distinguish between fundamental & harmonics







Raw DFT (2 users) Post-correlation (user 1)

(5th harmonic of user 1 equals fundamental of user 2)

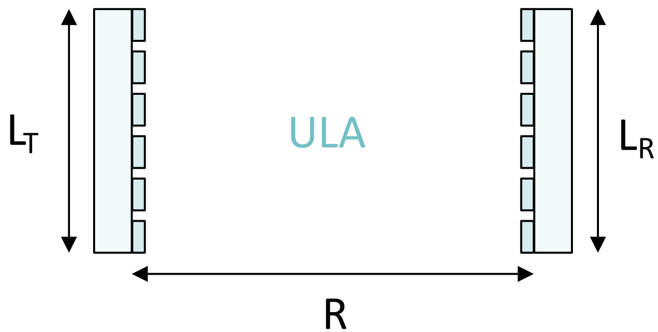
Post-correlation (user 2)

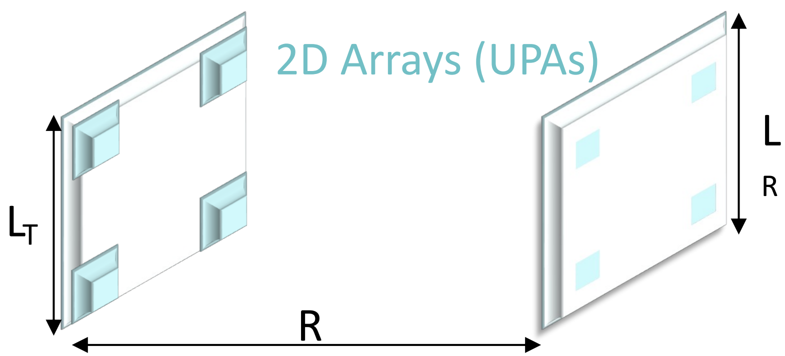
All-digital mmWave MU-MIMO Summary and Status

- Promising first steps for multiuser MIMO for massive scale
- Scale simplifies design of individual hardware components
- Rich space for continued research on scaling antennas & bandwidth
  - Computational complexity
  - Precision constraints (ADC and DAC) and analog nonlinearities
  - Interactions with mobility and power control

LoS MIMO everywhere

LoS MIMO is a natural concept for mmWave and THz





Number of spatial degrees of freedom
(based on information-theoretic considerations):

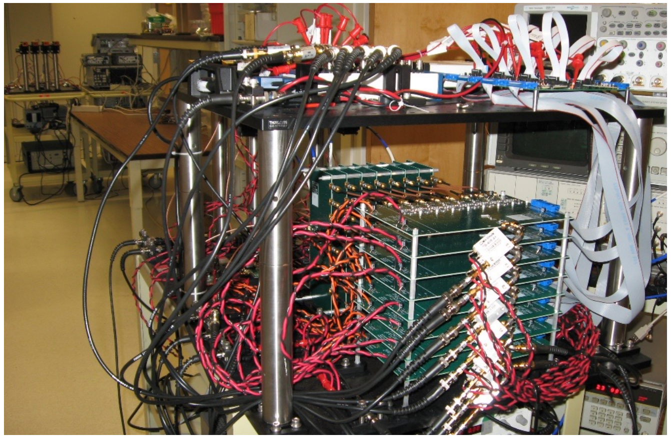
$$N \approx \frac{L_T L_R}{R\lambda} + 1$$

$$N \approx \left(\frac{L_T L_R}{R\lambda}\right)^2 + 1$$

Bandwidth also scales with carrier frequency
($fc \propto 1/\lambda$) where λ : wavelength

Data Rate $\propto f_c^3$

Significant progress in past decade



2 orders of magnitude in range & data rate

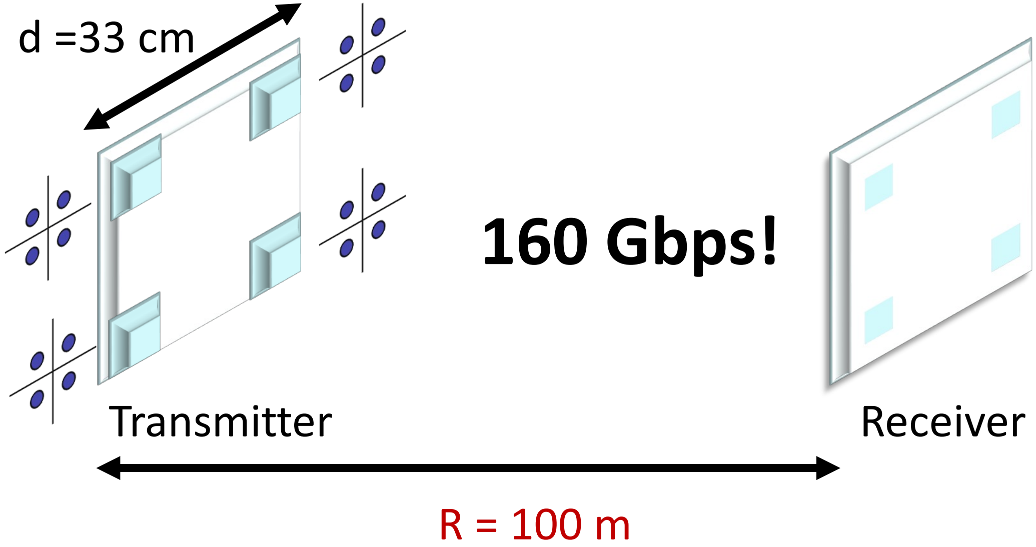
UCSB lab demo @ 60 GHz (2010)
4-fold spatial multiplexing
2.4 Gbps aggregate data rate
Range 10-40 meters

Ericsson prototype link in E-band (2019)
8-fold multiplexing: 4 spatial, 2X polarization
100 Gbps aggregate data rate
Range 1500 meters

Widespread deployment of LoS MIMO requires less bulky and expensive equipment

Short-range backhaul more interesting?

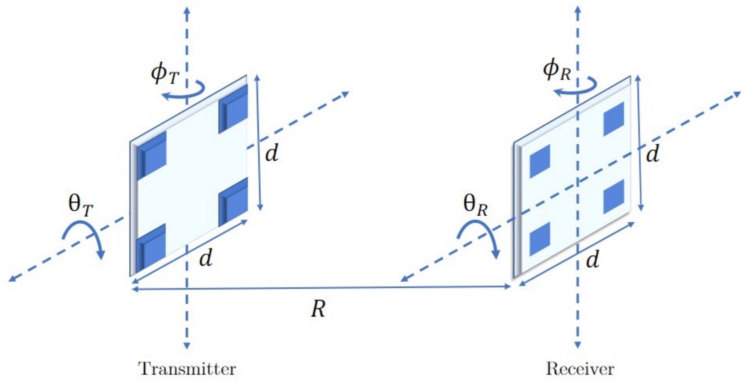
4 x 4 MIMO
130/140 GHz carrier frequency
40 Gbps per stream
Antenna spacing 33cm
(*lamppost-compatible*)

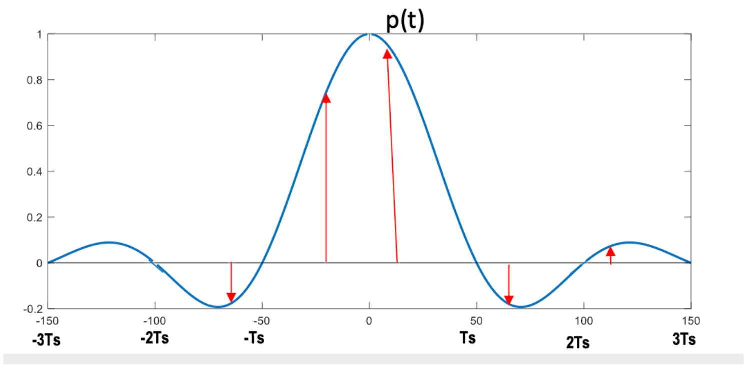


Reasonable form factor, but how about cost & power?

- CMOS or SiGe RFICs with required power are within reach
- DSP is key to economies of scale in baseband processing
- ADC is a bottleneck at 10-20 GHz bandwidths
- Geometric misalignments result in channel dispersion

Endemically misaligned LoS MIMO

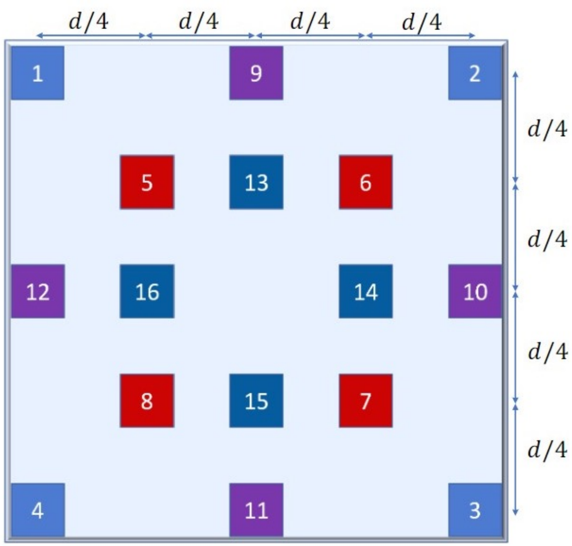


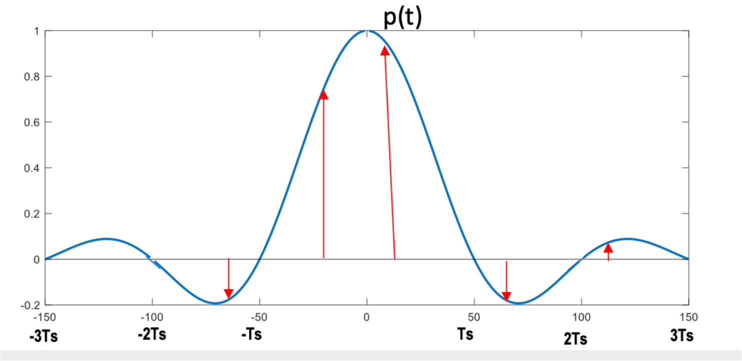


- Misalignment will be routine in mesh networks with LoS MIMO links
  ➔ Inter- and intra-stream interference
- Can we invert this space-time channel?
- Time domain oversampling not on the cards at 20 GHz bandwidth
- The answer: spatial oversampling

Spatial oversampling for robust LoS MIMO

Aperture is 100s of wavelengths ➔ room to fit additional elements

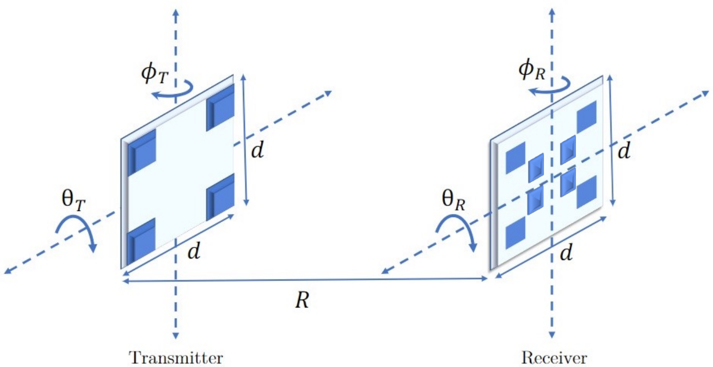




Different versions of sampled response at different elements

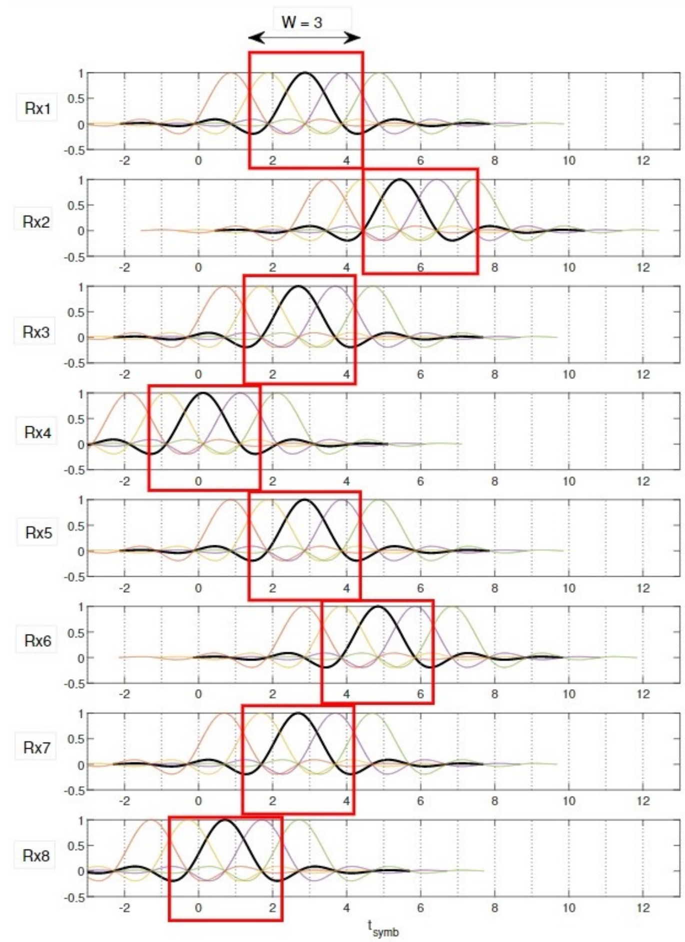
- QPSK modulation
- BW = 20 GHz for f_c = 130 GHz
- Symbol duration T = 50 ps and λ = 2.3 mm
- Transmit pulse - RC waveform with β = 0.25
- Symbol rate sampling: T_S = T

Adaptive windowing with spatial oversampling

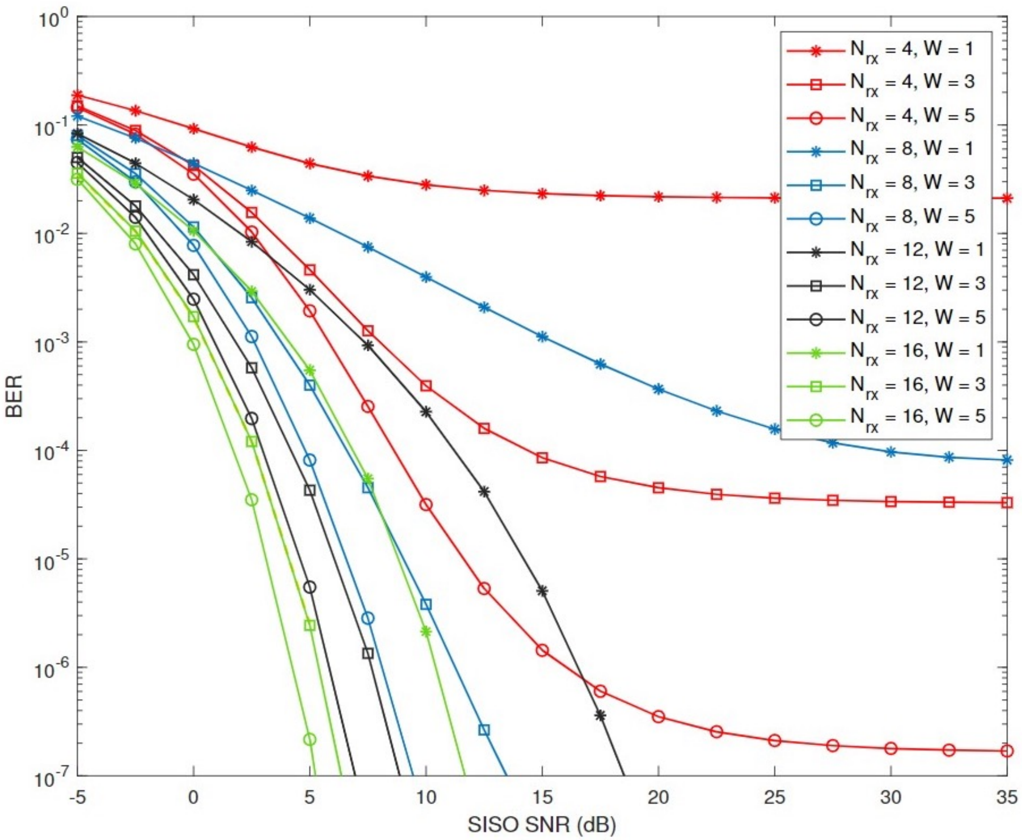


Misalignment example:

$$\theta_T = 3.67^\circ$$
$$\varphi_T = -4.30^\circ$$
$$\theta_R = 6.36^\circ$$
$$\varphi_R = 7.19^\circ$$



BER curves and dimension counting



W	Interference Vectors	Dimension of Signal Space				
		$N_{\mathrm{RX}}=4$	$N_{\mathrm{RX}}=6$	$N_{\mathrm{RX}}=8$	$N_{\mathrm{RX}}=12$	$N_{\mathrm{RX}}=16$
1	19	4	6	8	12	16
3	27	12	18	24	36	48
5	35	20	30	40	60	80

Error floors avoided when signal space dimension is bigger than # strong interference vectors

Opportunistic LoS MIMO Summary and Status

Spatial oversampling increases resilience to impairments (misalignment, precision constraints)

Many hardware, architecture and algorithm issues remain

But in principle, flexible, low-cost "wireless fiber" is feasible

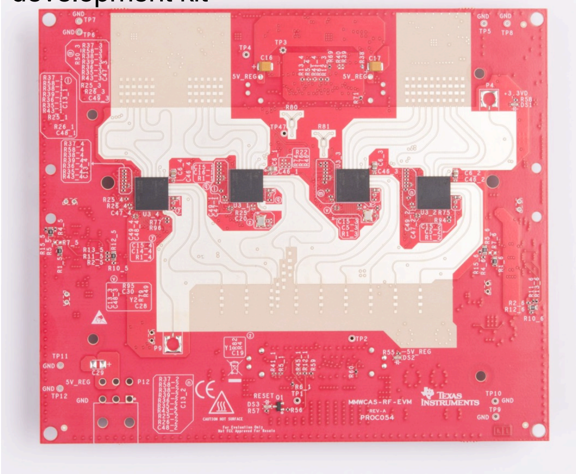
Massive MIMO Radar

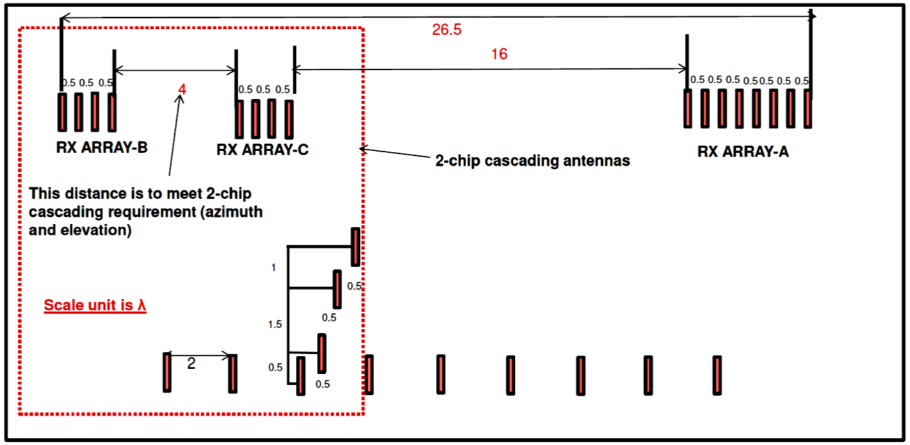
A Compressive Approach

Current state of MIMO radar

- Small number of TX antennas
- Larger, $\lambda/2$-spaced receive antenna array
- Virtual transmit array: TXs take turns

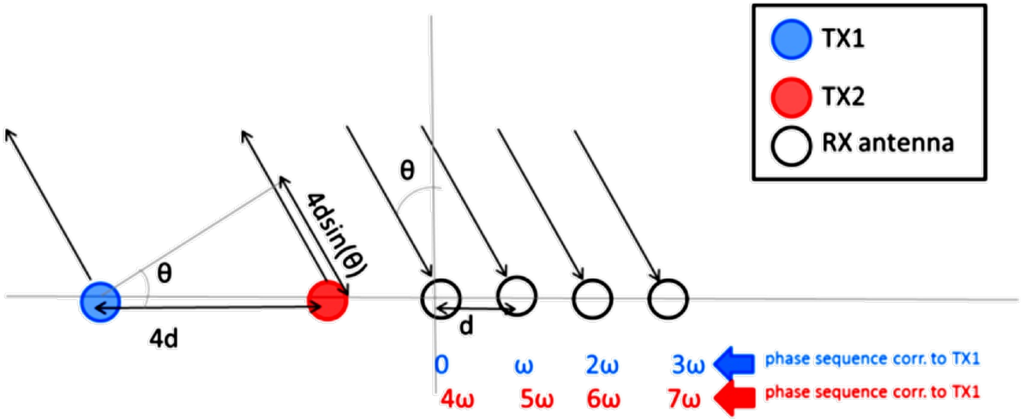
TI's AWR2243 Cascade Radar RF development kit

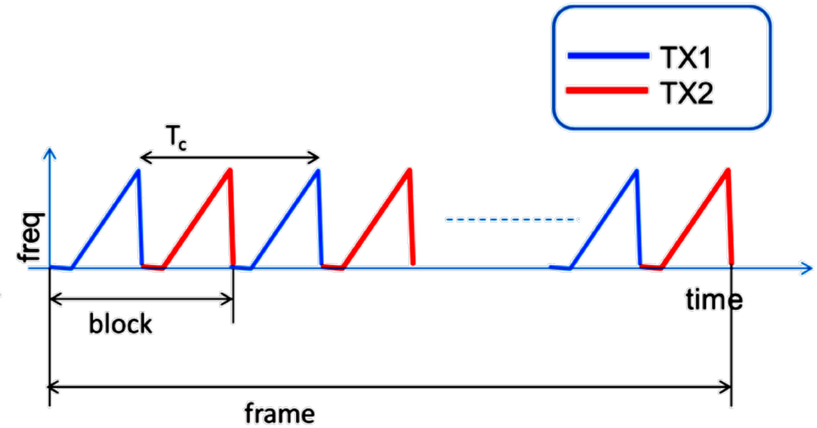




Figures courtesy
Texas Instruments

Current state of MIMO radar





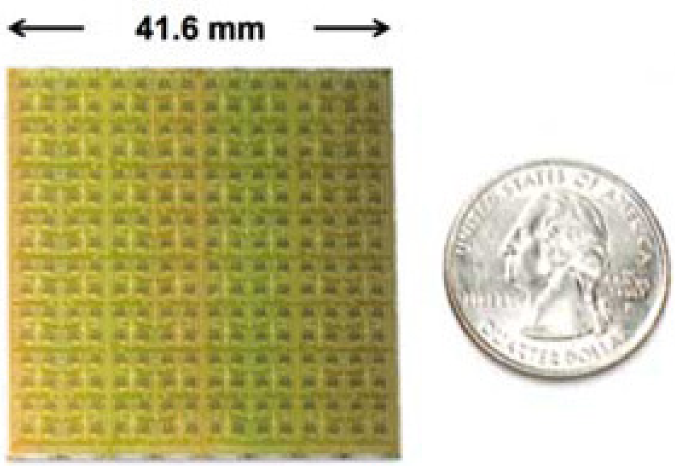
- Drawbacks
- × No transmit beamforming gain – FOV/range tradeoff limited by element directivity
- × Scalability – frame grows large as # TX antennas increases

Figures courtesy of Texas Instruments

Can we leverage developments in comm hardware?

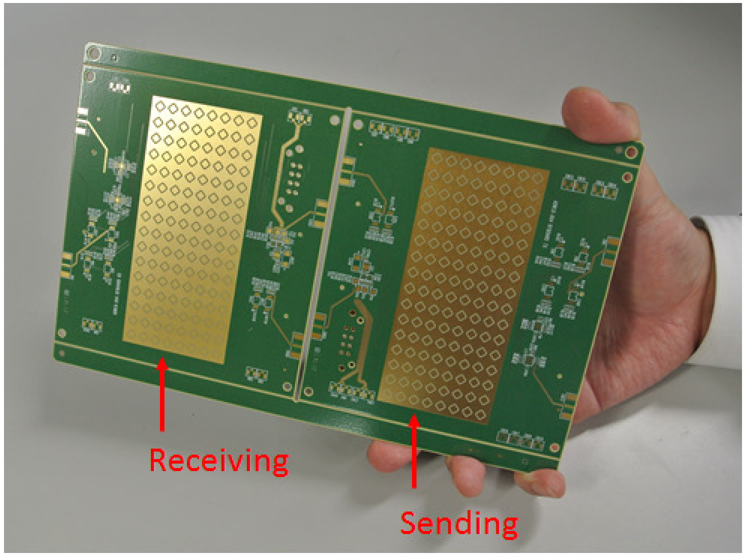
- Large transmit arrays with RF (analog) beamforming
- Large field of view for each element, sharp beams: range/FOV tradeoff eliminated by increasing # TX elements

TowerJazz and UCSD:
256-element (16 x 16), 60GHz phased-array transmitter
56-65GHz frequency range



256-Element Array Chip

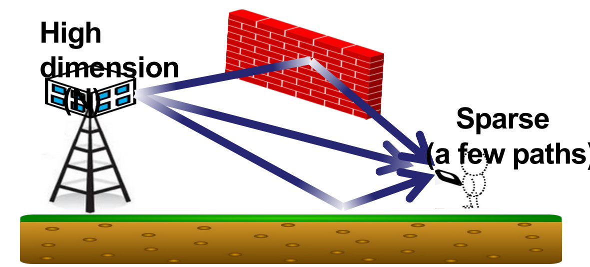
Fujitsu: 28 GHz, 128-element phased array antenna
16 x 8 patch antenna elements (per direction)
Four independent beams, steerable horizontally and vertically.

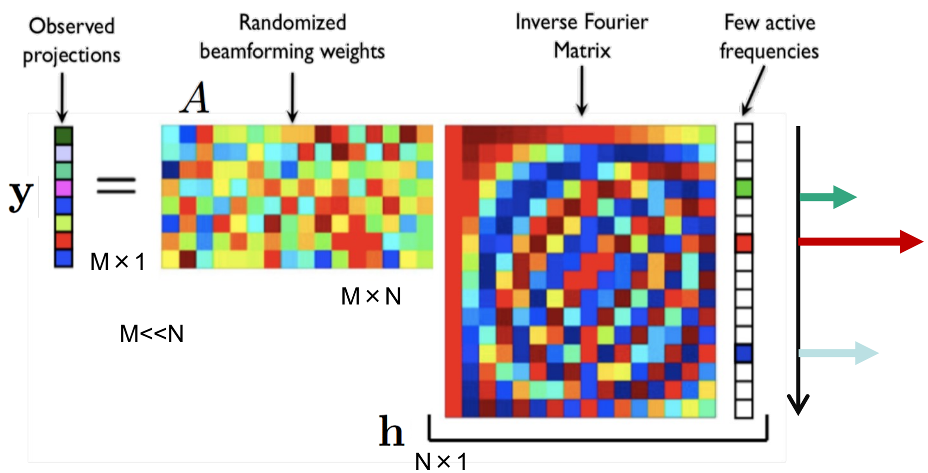


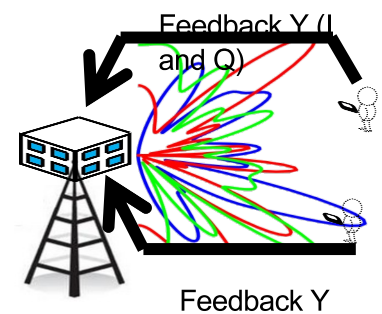
Which concepts from MIMO comm can we bring into MIMO radar?

Concept 1: Compressive estimation of sparse channels

- Large transmit arrays with RF (analog) beamforming
- Locating users in FOV? Sparse channel estimation problem
- (Off-grid) Compressive channel estimation







Ramasamy, Dinesh, Sriram Venkateswaran, and Upamanyu Madhow. “Compressive tracking with 1000-element arrays: A framework for multi-Gbps mm wave cellular downlinks.” *Allerton*, 2012.
Marzi, Zhinus, Dinesh Ramasamy, and Upamanyu Madhow. “Compressive channel estimation and tracking for large arrays in mm-wave picocells.” *IEEE Journal of Selected Topics in Signal Processing*, 2016

Concept 2: Efficient off-grid estimation with Newtonized OMP (NOMP)

- **Super-resolution**

  Estimating over the continuum

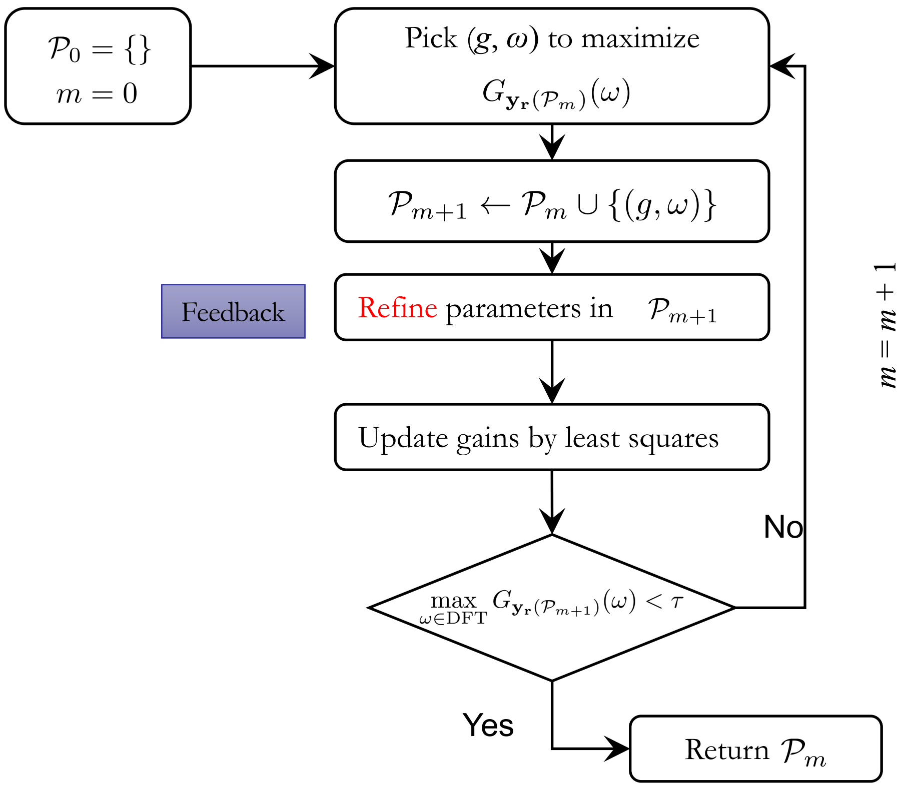
GLRT cost function

$$G_{\mathbf{y}}(\omega) = \frac{|\mathbf{x}(\omega)^H \mathbf{y}|^2}{||\mathbf{x}(\omega)||^2}$$

Residual response

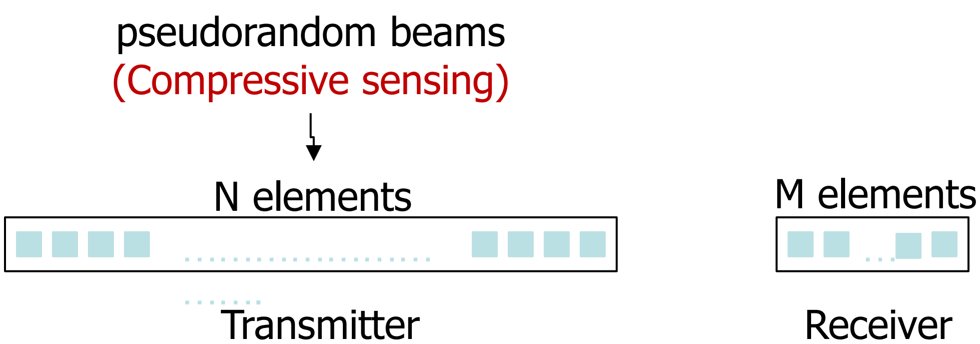
$$\mathcal{P} = \{(g_l, \omega_l), l = 1, \ldots, k\}$$

$$\mathbf{y_r}(\mathcal{P}) = \mathbf{y} - \sum_{l=1}^{k} g_l \mathbf{x}(\omega_l)$$



B. Mamandipoor, D. Ramasamy, U. Madhow, "Newtonized Orthogonal Matching Pursuit: Frequency Estimation over the Continuum," in IEEE Transactions on Signal Processing, 2016.

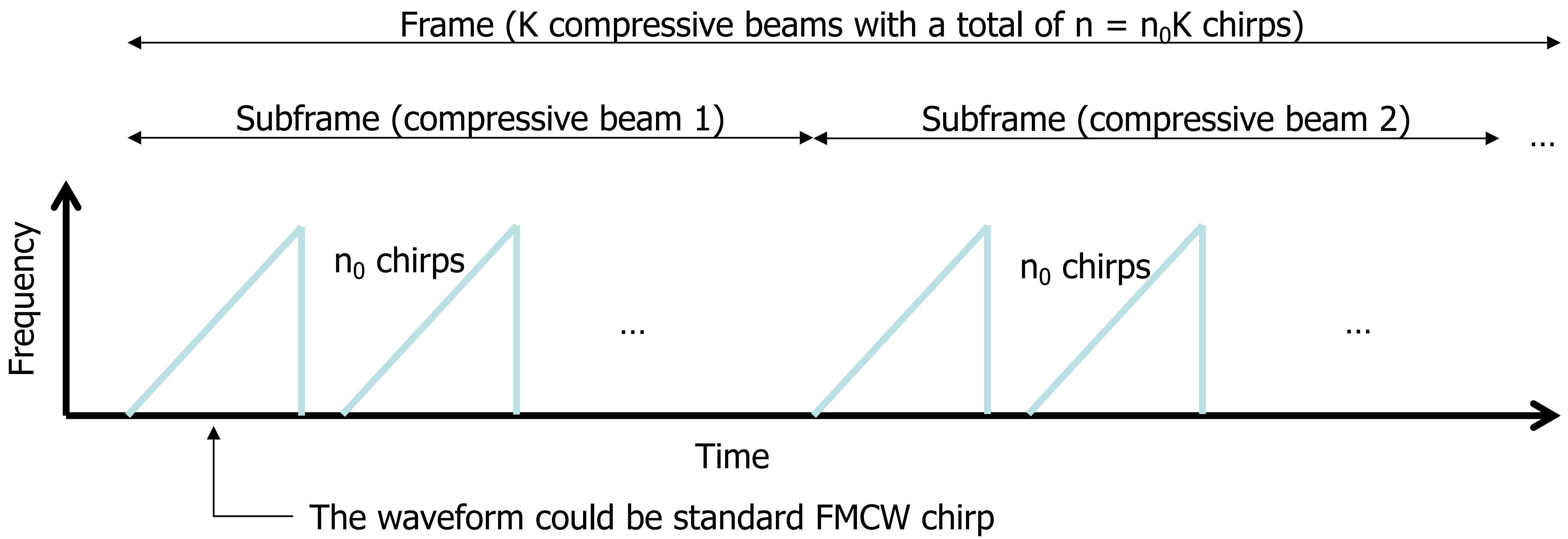
From sparse channel estimation to target detection

- Can we apply the same principle to CWFM radar?
- RF beamformed transmitter (phased array)
- Digital receiver(s)
- 

A virtual array of MN elements can be created

Overlay compressive angular scanning on range/Doppler estimation

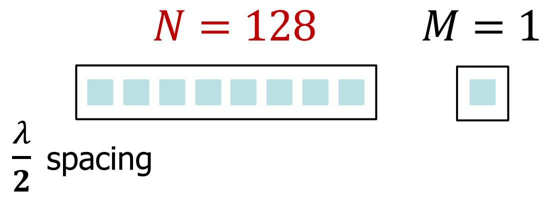
Compressive scanning

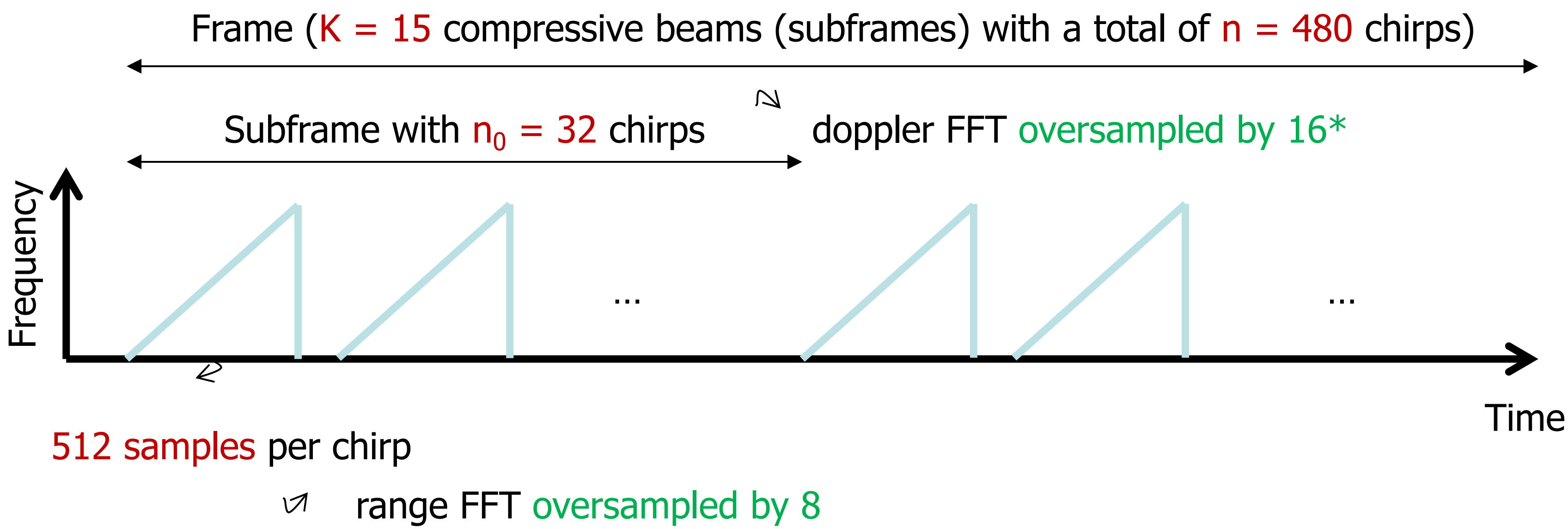


- Range processing unchanged
- Need to modify Doppler processing to handle angle-dependent phase shifts across subframes.

Preliminary results - setup

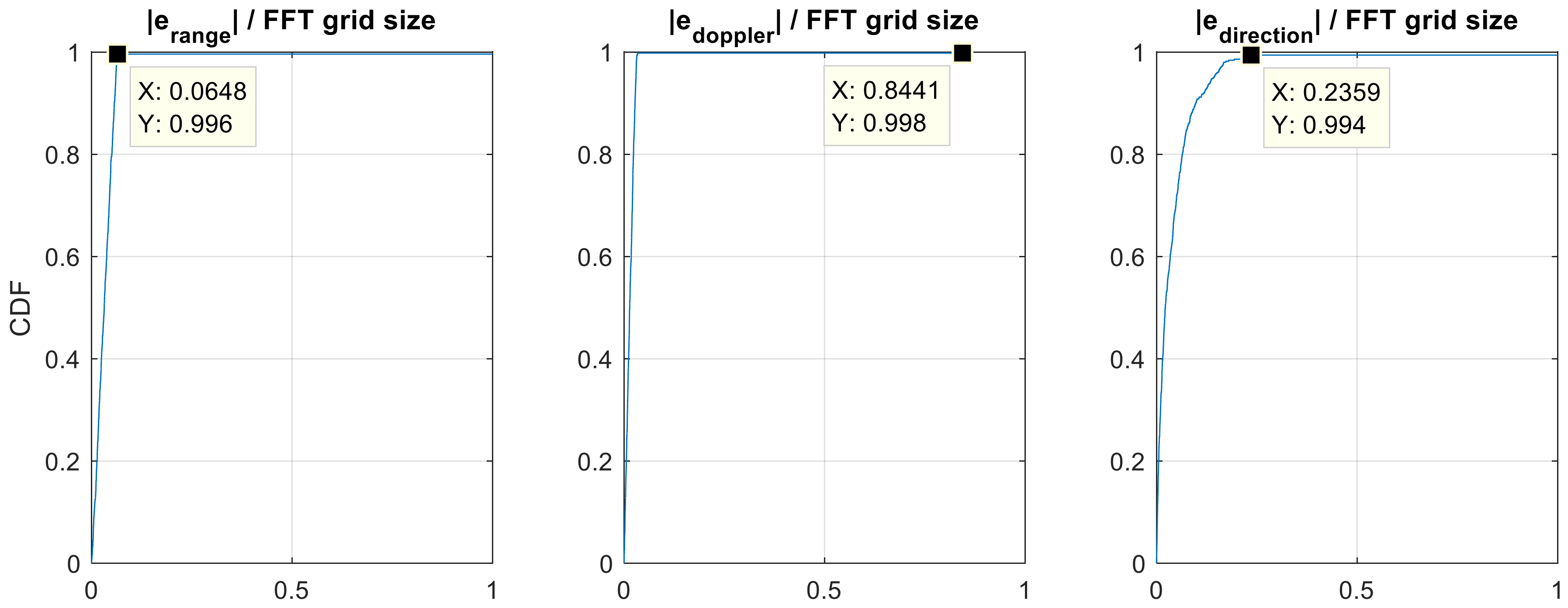
- Number of targets $\sim U(\{1,\ldots,9\})$, dynamic range ≤ 18 dB
- SNR per element = 5 dB





Preliminary results

- Super-resolution even at relatively low SNR



Scaling MIMO radar Summary and Status

Massive MIMO radar enabled by compressive scanning

Sidesteps range/FOV tradeoffs in existing MIMO radar

Preliminary results show promise; more detailed eval needed

Ultimate Goal: bridge the resolution gap between RF & optical sensing

Significant opportunities in joint communication & sensing

Signal Processing for mmWave/THz

- Pushing to higher carrier frequencies keeps opening up new intellectual challenges via hardware/signal processing entanglement
  - Hardware bottlenecks force system innovations
  - Hardware advances open up new system possibilities
  - Key ideas: antenna scaling, bandwidth scaling, sparsity, geometry
- Ambitious system specs today become industry focus ~10 yrs from now
  - The only legitimate barriers are physics and information theory fundamentals
- (sub)-THz sensing is the new frontier
  - Unprecedented spatiotemporal resolution
  - Privacy-preserving, more robust than optical
  - We have only scratched the surface of massive and distributed MIMO for sensing

For further exploration

- Wireless Communication and Sensornets Lab (WCSL) home page: https://wcsl.ece.ucsb.edu
- NSF Giganets project (2015-2020): https://wcsl.ece.ucsb.edu/giganets
  - Papers from interdisciplinary collaborations involving hardware, signal processing and systems
  - Tutorial material (IISc summer school 2016, ACM SigComm 2017)
- ComSenTer (2018-2022): https://comsenter.engr.ucsb.edu/
  - UCSB-led center funded by DARPA and SRC
  - Pushing the limits of mm-wave and THz comm and sensing: both hardware and signal processing
- **CUbiC (2023-27):** https://cubic.engineering.columbia.edu/
  - Columbia-led center on NextG connectivity
  - Technology for low-cost, ubiquitous deployment

CUbiC

And now for something completely different….

Communication theory for robust deep learning

Who's involved

Actually doing the work!

Bhagyashree Puranik
(PhD student)

In an advisory role…

Ahmad Beirami
(Google)

Yao Qin
(UCSB)

U. Madhow
(UCSB)

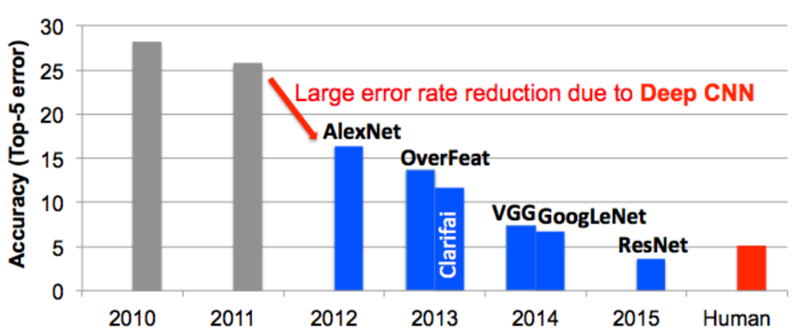
Builds on prior work by…

Metehan Cekic
(now at AWS AI)

Can Bakiskan
(now at Intel)

The big picture

- AI = Deep Neural Networks (for all practical purposes)
- We all know that DNNs are brittle black boxes
  - That does not stop us from using them *everywhere*
- Soft failures are OK in many applications
- But lack of robustness and interpretability blocks many others



AlexNet

~12 years of furious activity →

"cats dancing tango" by Microsoft Image Creator

What's behind the DNN revolution?

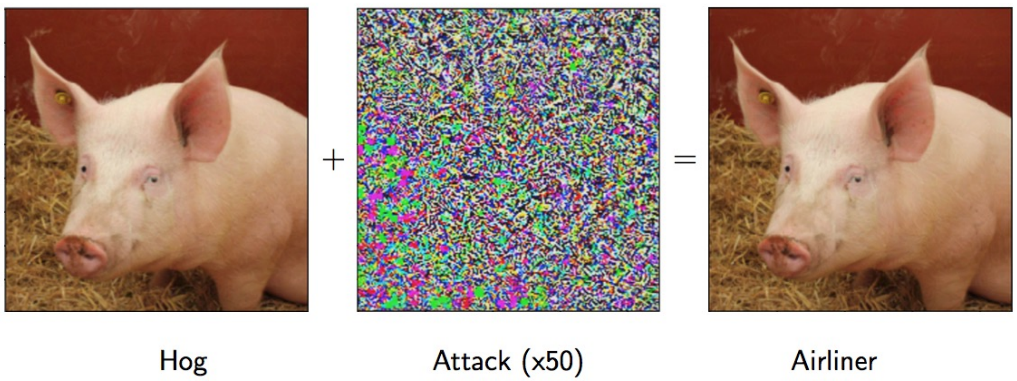
- Backprop works
  - Big data, big compute
  - Sigmoid→ ReLU so gradients propagate down
  - Deeper is better, overparametrization is better
- Design approach is very open to experimentation
  - Define cost function (depends on learning modality)
  - Play with architectures and hyperparameters
- Empirical results blow away the state of the art in most applications

→ We try to work around concerns such as (lack of) robustness, interpretability, fairness in applications…

Can we do better?

Early warning signal: adversarial examples

Szegedy et al, *Intriguing properties of neural networks,* 2013-14.
Goodfellow et al, *Explaining and harnessing adversarial examples,* 2014-15.



Pig → Plane

Stop sign → 45 mph

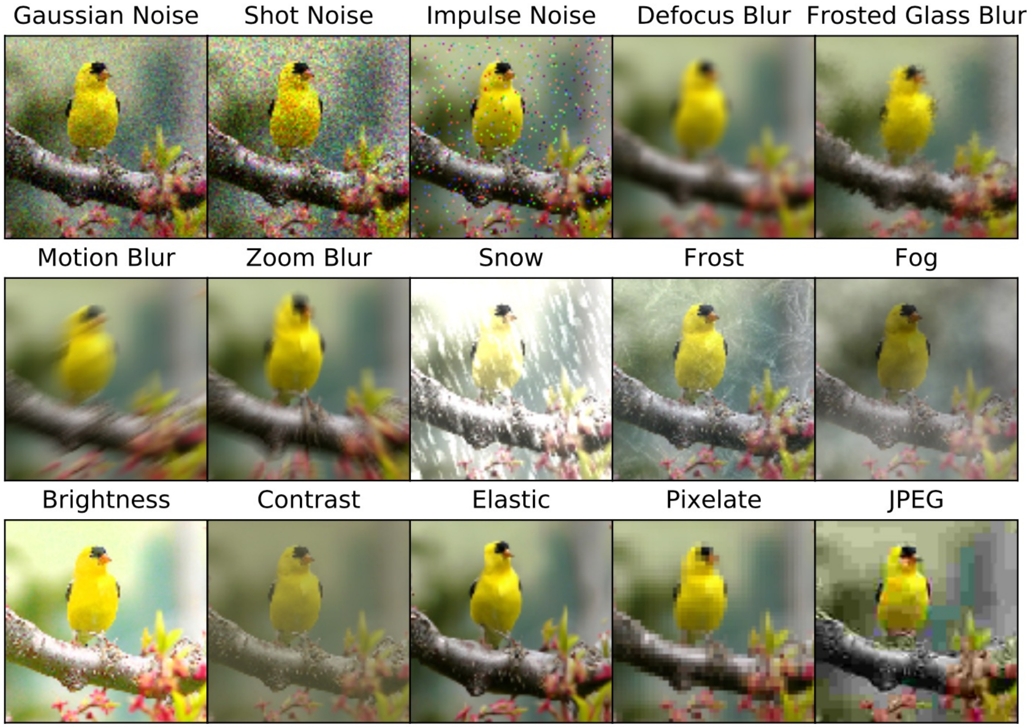
Key insights from a decade ago

- DNNs are too linear
- Small perturbations can add up to large numbers in high dimensions

have not led to concrete design guidelines for robustness

A more pressing (?) concern: OOD robustness

Hard to define "out of distribution" precisely
But you know it when you see it



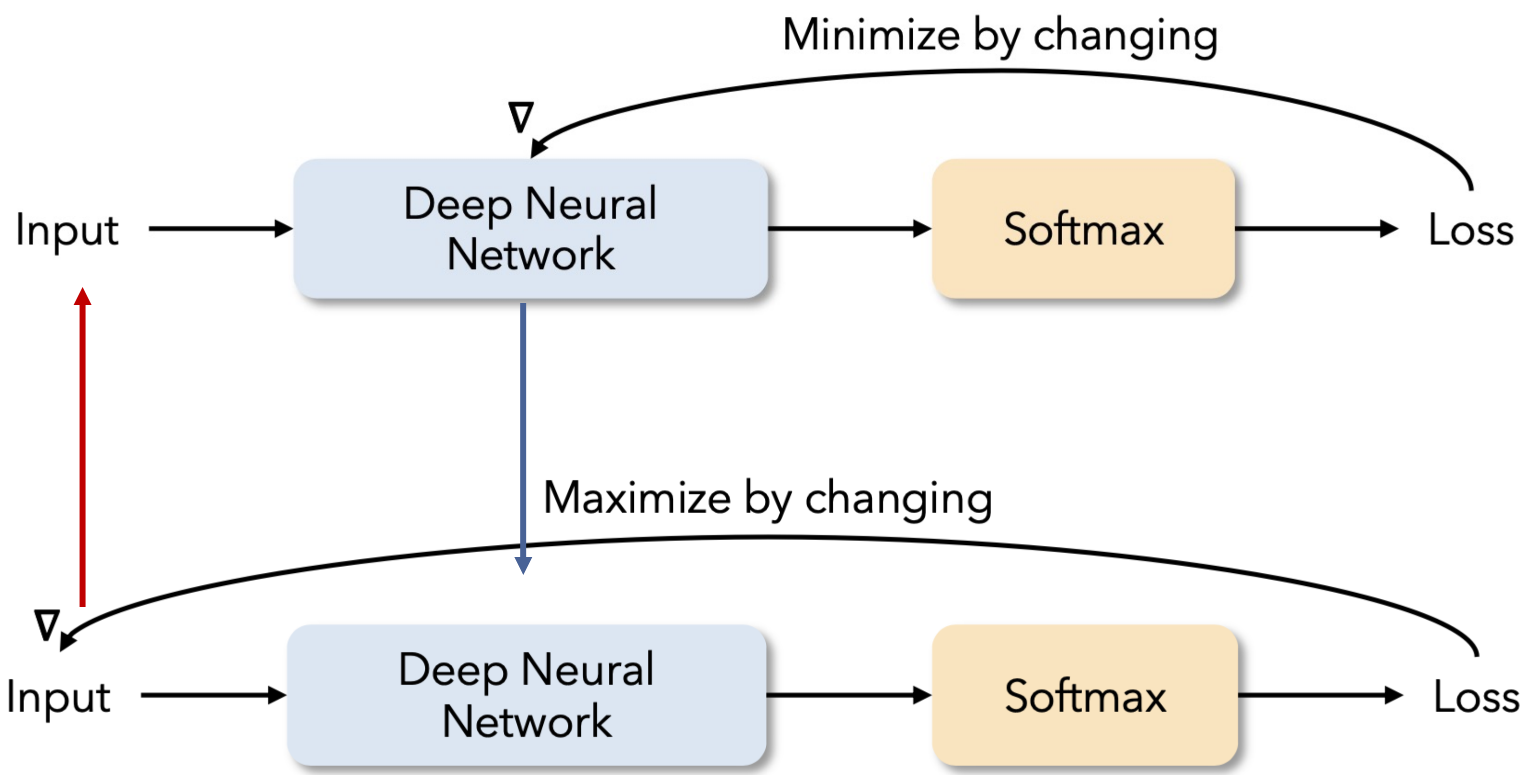
We expect DNNs to be robust to "common corruptions"

SOTA approach for “robust” DNNs is data augmentation

SOTA defense against adversarial perturbations

Adversarial training

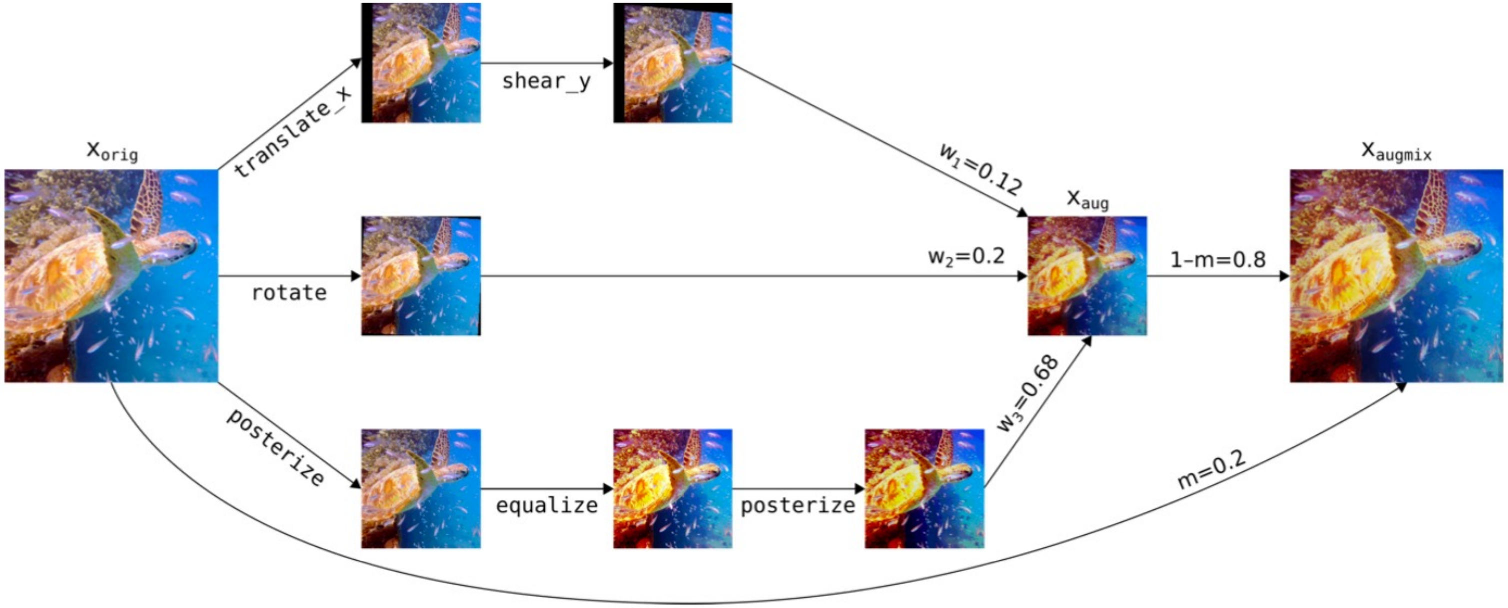
augment with adversarial examples generated while training



SOTA for OOD robustness

is also data augmentation!

AugMix, RandAugment, AutoAugment,…



Hendrycks *et al*, Augmix: A simple data processing method to improve robustness and uncertainty, *ICLR 2020*.

Our thesis: lack of robustness is a *symptom*

End-to-end training (with or without augmentation) ➔ Black box

The disease: we do not control the features DNNs are extracting

➔ We cannot design in robustness *guarantees*

➔ We cannot *interpret* what DNNs are doing

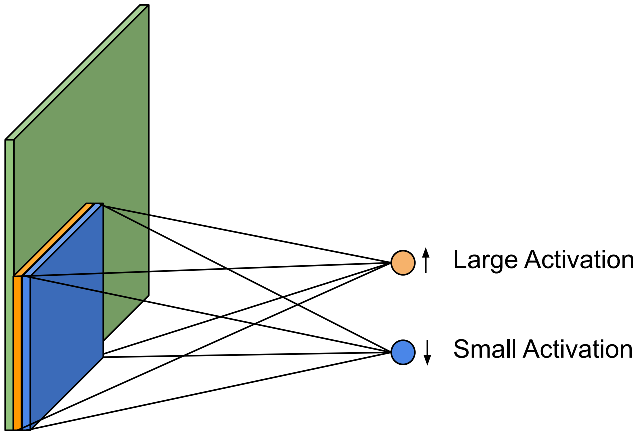
Our approach: shaping DNNs for robustness

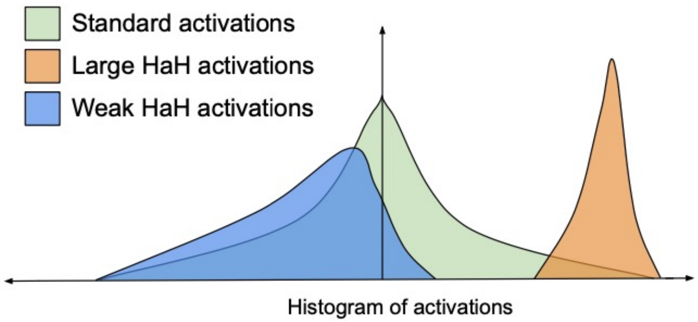
- **Axiom:** Learning can only work if data has low-dimensional structure
- Can we "match" our DNN layers to these low-dimensional manifolds?
- **Idea:** shape each layer of DNNs to produce sparse, strong activations
  - Small fraction of strong activations are harder to perturb
  - Large fraction of weak activations can be attenuated/removed
  - Increased resilience, potentially better generalization and interpretability
- How does communication theory come in?
  - Learn matched filters at each layer (using layerwise objectives)
  - Output posterior probabilities at each layer
  - Codifies initial insights from neuroscience used in our prior work

Our prior work inspired by neuroscience

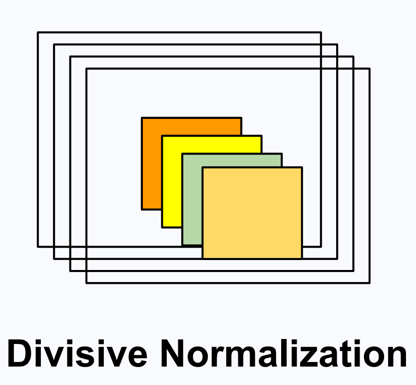
Neuronal competition during training

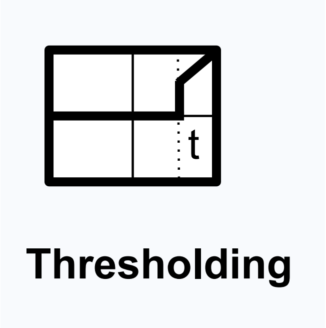
via layerwise Hebbian/anti-Hebbian (HaH) learning





Neuronal competition during inference





(Cekic *et al*, ICIP 2022, 2022 ICML Workshop Adv ML)

Communication theory principles yield better (and neuro-plausible) designs

Learning neuronal "matched filters"

A communication-theoretic formulation ➔ tilted exponentials

M-ary hypothesis testing in Gaussian noise

$$H_i : \ \mathbf{x} = \mathbf{s}_i + \mathbf{n} \ , \ i = 1, ..., M$$

Wish to *learn* the signal templates $\theta = \{\mathbf{s}_i \ , \ i = 1, ..., M\}$

Likelihood function conditioned on hypothesis

$$L_\theta(\mathbf{x}|H_i) = \exp\left(\frac{1}{\sigma^2}\left(\langle \mathbf{x}, \mathbf{s}_i \rangle - ||\mathbf{s}_i||^2/2\right)\right)$$

"Data noise" σ^2

Likelihood function averaged across hypotheses

$$L_\theta(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_i \rangle\right)$$

Log likelihood (for adding across data samples) ➔ tilted exponential

$$T_\theta(\mathbf{x}) = \log\left(\frac{1}{M}\sum_{i=1}^{M} \exp(t \ a_i)\right)$$

$t = 1/\sigma^2$ Tilt (hyperparameter)

$a_i = \langle \mathbf{x}, \mathbf{s}_i \rangle$ Activation of *i*th neuron

TEXP learning is Hebbian

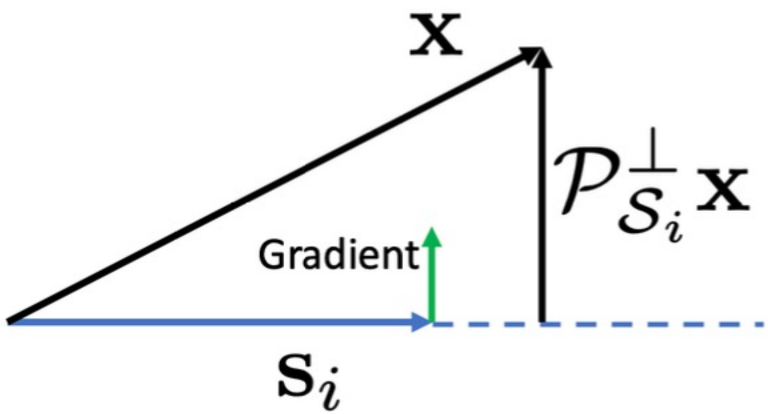
- Implicitly normalize activations for fair competition

$$a_i = \langle \mathbf{x}, \frac{\mathbf{s}_i}{||\mathbf{s}_i||} \rangle$$

Recall TEXP objective to be maximized

$$T_\theta(\mathbf{x}) = \log \left(\frac{1}{M} \sum_{i=1}^{M} \exp(t\ a_i) \right)$$

- Geometry of TEXP gradient update



Rotate template towards input

$$\nabla_{\mathbf{s}_i} T_\theta = \sigma_i(t\mathbf{a}) \frac{\mathcal{P}^{\perp}_{\mathcal{S}_i} \mathbf{x}}{||\mathbf{s}_i||_2}$$

Relatively stronger activations weighted more heavily

TEXP inference: soft decisions

M-ary hypothesis testing in Gaussian noise

$$H_i : \mathbf{x} = \mathbf{s}_i + \mathbf{n} \ , \ i = 1, \ldots, M$$

Soft decisions (posterior probabilities)

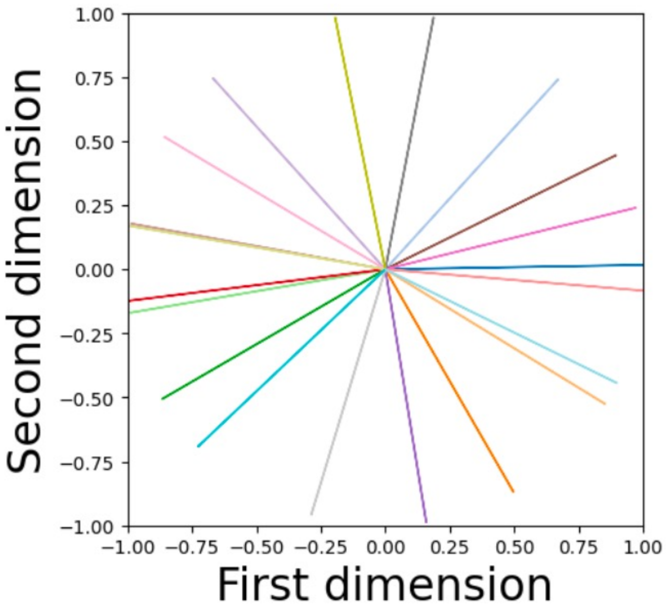
$$P(H_i|\mathbf{x}) = \frac{L_\theta(\mathbf{x}|H_i)P(H_i)}{\sum_{j=1}^{M} L_\theta(\mathbf{x}|H_j)P(H_j)} = \frac{\exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_i \rangle\right)}{\sum_{j=1}^{M} \exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{s}_j \rangle\right)} = \mathrm{Softmax}(t \ a_i)$$

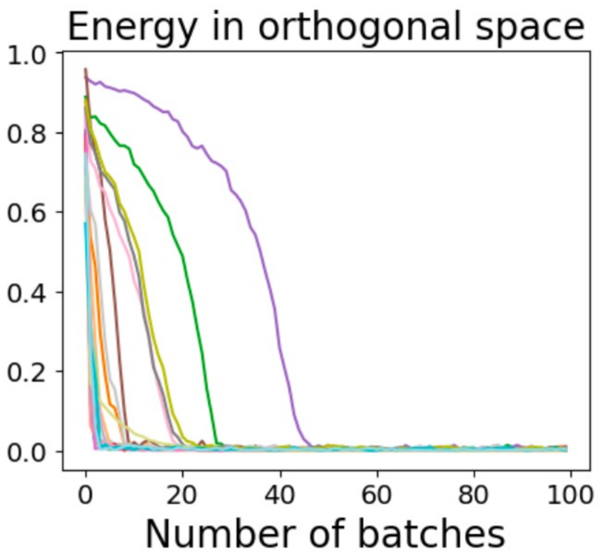
Smoothened "divisive normalization"
Eliminates vulnerability due to "excessive linearity"

Set "data noise" for inference higher than for training
➔ Increased robustness despite training with clean data

The geometry of TEXP via a simplified example

A 2D Gaussian signal hiding in 10D Gaussian noise

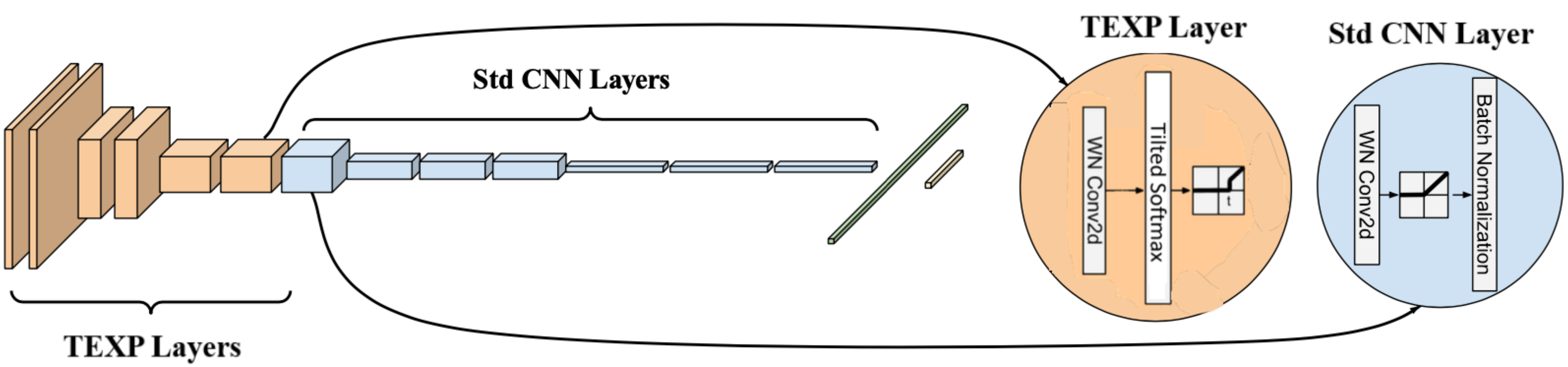




Neurons learnt via TEXP hone in on 2D "signal subspace"
Energy of neurons in 8D orthogonal "noise subspace" falls off as we train

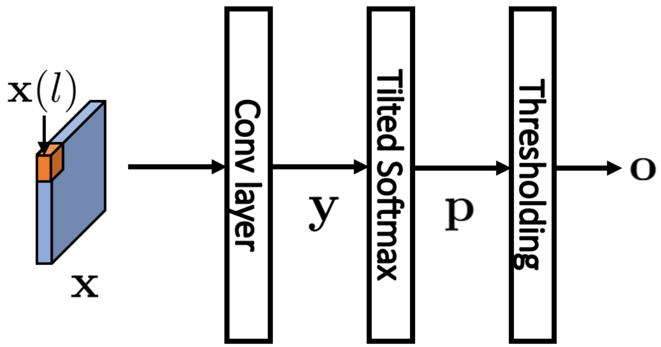
TEXP in CNNs

- Training: supplement end-to-end cost with TEXP-based layerwise costs
  - Smooth objective leading to Hebbian learning
- Inference: replace ReLU + batch norm by tilted softmax
  - Implicit normalization and thresholding
  - Tilted softmax is a form of divisive normalization
- Outperforms our prior work on Hebbian/anti-Hebbian (HaH) learning



Puranik et al, 2023 ICML Workshop Adv ML, AISTATS 2024

Why TEXP is expected to increase robustness



- Gradient of TEXP objective promotes strong activations
- Tilted softmax inference ➔ nonlinearity attenuating perturbations
- Smaller tilt for inference than for training ➔ robustness
- Neuron-specific thresholding ➔ denoising

Many open questions before TEXP can become a generic layer

- How do we choose the tilt parameters?
- How do we weight the layerwise objectives?

TEXP provides broad spectrum robustness

Performs well for both common corruptions and mild adversarial attacks

Model	Clean	Noise $\nu = 0.1$	Min/Avg corruptions	Min/Avg severity level: 5	Autoattack ℓ_2 adv, $\epsilon = 0.25$	Autoattack ℓ_∞ adv, $\epsilon = 2/255$
VGG-16	92.26 ± 0.04	24.80 ± 1.24	$46.86 \pm 1.26/72.28 \pm 0.26$	$19.56 \pm 0.73/54.70 \pm 0.40$	13.34 ± 0.14	10.30 ± 0.21
HaH (Cekic et al. (2022))	87.72 ± 0.15	62.76 ± 0.40	$59.56 \pm 0.42/77.02 \pm 0.21$	$49.06 \pm 0.88/67.80 \pm 0.27$	26.30 ± 0.52	20.04 ± 0.38
TEXP-VGG-16	88.28 ± 0.12	75.14 ± 0.20	$73.68 \pm 0.22/80.40 \pm 0.07$	$52.38 \pm 0.81/72.56 \pm 0.14$	50.90 ± 0.16	41.50 ± 0.21
VGG-16 + AugMix	92.98 ± 0.06	62.92 ± 0.74	$65.12 \pm 0.35/83.58 \pm 0.09$	$42.12 \pm 0.79/74.00 \pm 0.16$	18.16 ± 0.15	13.60 ± 0.17
TEXP-VGG-16 + AugMix	88.84 ± 0.21	78.90 ± 0.04	$77.28 \pm 0.20/83.54 \pm 0.05$	$62.94 \pm 0.53/78.30 \pm 0.07$	52.20 ± 0.23	42.52 ± 0.20
VGG-16 + RandAug	93.32 ± 0.11	43.32 ± 0.72	$63.24 \pm 0.45/80.68 \pm 0.17$	$39.98 \pm 1.01/66.96 \pm 0.30$	18.38 ± 0.47	14.30 ± 0.37
TEXP-VGG-16 + RandAug	89.90 ± 0.08	74.26 ± 0.07	$75.48 \pm 0.09/82.86 \pm 0.02$	$57.52 \pm 0.19/75.78 \pm 0.07$	50.82 ± 0.24	40.02 ± 0.34
VGG-16 + AutoAug	93.50 ± 0.03	46.54 ± 0.54	$59.84 \pm 0.52/81.58 \pm 0.14$	$37.08 \pm 0.23/70.66 \pm 0.18$	13.50 ± 0.23	9.78 ± 0.20
TEXP-VGG-16 + AutoAug	90.06 ± 0.10	72.66 ± 0.46	$71.98 \pm 0.24/82.58 \pm 0.12$	$54.14 \pm 0.89/75.50 \pm 0.18$	46.96 ± 0.31	35.00 ± 0.32
VGG-16 + Adv Tr	88.04 ± 0.12	78.78 ± 0.45	$50.52 \pm 0.66/79.44 \pm 0.12$	$17.60 \pm 0.39/70.64 \pm 0.13$	72.60 ± 0.23	72.82 ± 0.23
TEXP-VGG-16 + Adv Tr	86.38 ± 0.07	81.08 ± 0.28	$67.72 \pm 0.73/80.38 \pm 0.14$	$37.08 \pm 0.85/74.02 \pm 0.22$	71.02 ± 0.40	66.76 ± 0.29

Detailed performance report under common corruptions (highest severity level)

Corruptions →	Noise				Weather					Blur					Digital				
Models ↓	Gauss.	Shot	Speck.	Imp.	Snow	Frost	Fog	Brig.	Spat.	Defoc.	Gauss.	Glass	Motion	Zoom	Cont.	Elas.	Pixel.	JPEG	Satur.
VGG-16	24.3	31.8	38.4	19.1	73.3	62.0	63.8	87.9	67.3	50.8	39.8	47.6	60.0	61.5	19.9	75.6	54.6	77.4	82.4
HaH (Cekic et al., 2022)	61.7	61.7	59.2	46.3	73.8	72.3	62.8	83.2	76.7	64.3	58.4	53.2	65.1	68.9	76.0	74.0	60.5	79.3	79.6
TEXP-VGG-16	75.3	76.5	75.5	61.3	76.4	76.8	51.8	83.2	76.1	68.9	63.4	68.6	65.0	74.2	66.0	75.2	80.8	82.9	78.8
VGG-16 + AugMix	60.7	68.1	71.3	44.9	80.2	75.3	76.5	89.7	81.7	84.8	80.8	59.6	81.4	84.0	40.0	79.5	69.4	82.0	86.9
TEXP-VGG-16 + AugMix	78.9	79.5	79.0	67.7	78.4	79.0	62.2	83.8	78.8	81.5	79.8	72.4	77.1	82.6	75.5	78.6	83.6	83.7	81.6
VGG-16 + RandAug	44.7	53.5	57.5	40.0	78.6	72.8	71.0	90.9	85.3	63.6	52.9	61.0	67.8	71.7	48.3	79.9	56.9	81.7	88.5
TEXP-VGG-16 + RandAug	74.1	75.1	72.7	57.1	79.1	78.7	60.3	88.6	81.3	73.4	68.7	70.7	70.8	77.4	83.5	78.3	79.4	84.5	85.8
VGG-16 + AutoAug	45.7	53.1	56.7	37.1	77.2	69.8	81.1	91.9	81.1	79.1	75.2	51.8	75.2	81.1	80.0	76.5	50.4	80.2	90.2
TEXP-VGG-16 + AutoAug	72.3	72.5	70.8	53.1	76.9	76.1	62.9	88.3	77.5	76.1	72.9	65.6	72.4	79.8	86.0	76.5	77.4	84.5	86.2
VGG-16 + Adv Tr	79.8	81.1	80.3	62.7	74.3	73.3	33.2	76.8	77.7	71.1	66.8	76.0	69.1	74.9	18.3	78.4	82.6	84.8	76.6
TEXP-VGG-16 + Adv Tr	81.6	82.3	81.9	74.8	71.9	75.8	39.0	76.9	78.5	75.9	72.8	76.8	73.1	78.3	52.9	78.6	83.2	84.0	76.3

Outperforms HaH, plays well with augmentation

Parting Thoughts

- Pure reliance on end-to-end training can only lead to black boxes
  - Limits the possibility of performance guarantees and interpretability
- Layer-wise feature control is a potential robustifier
  - Shaping layer outputs to be sparse and strong enhances resilience
- Preliminary results promising, but most of the work remains…
  - More efficient training, guidance on hyperparameters
  - Additional shaping design guidelines and theoretical foundations
  - Different learning modalities (self-supervised, unsupervised, RL,…)
  - Enhanced interpretability?