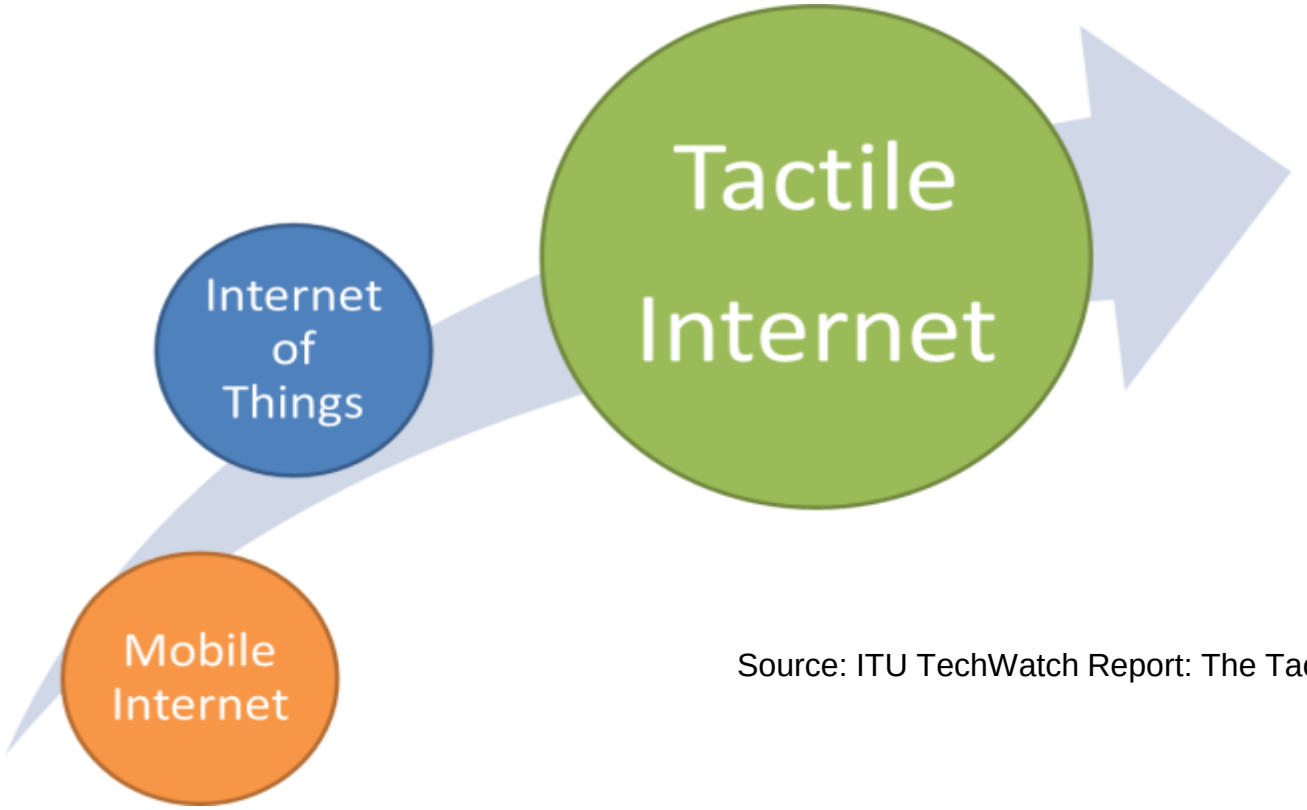
Machine Learning for 5G and Beyond

Slawomir Stanczak

Joint work with R.L.G. Cavalcante, S. Limmer and L. Miretti

Fraunhofer Heinrich Hertz Institute

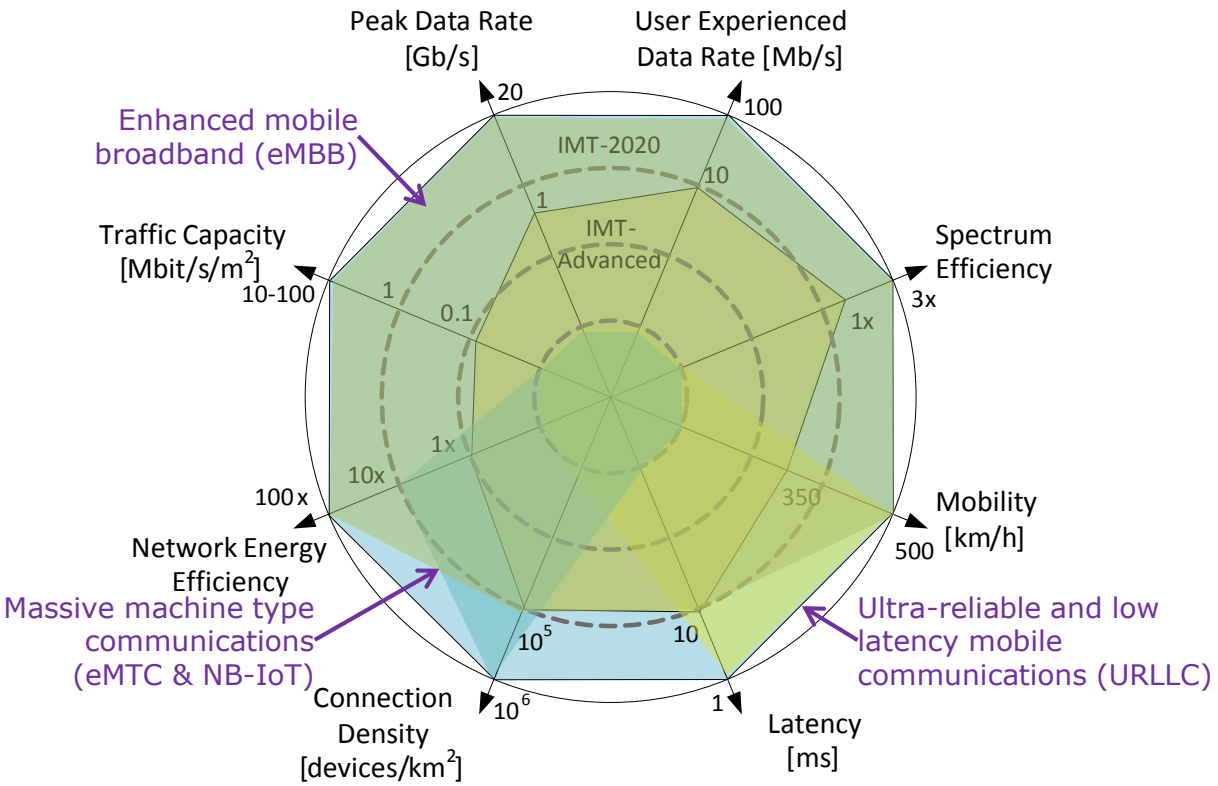
Revolutionary Leap of Tactile Internet



Source: ITU TechWatch Report: The Tactile Internet

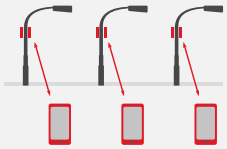
Connecting machines into control loops at humanoid reaction times of milliseconds and less

4G and 5G compared

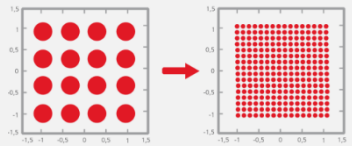


NB: Downlink metrics shown

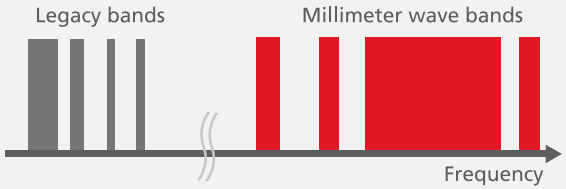
Mechanisms for Performance Enhancements



Cell densification



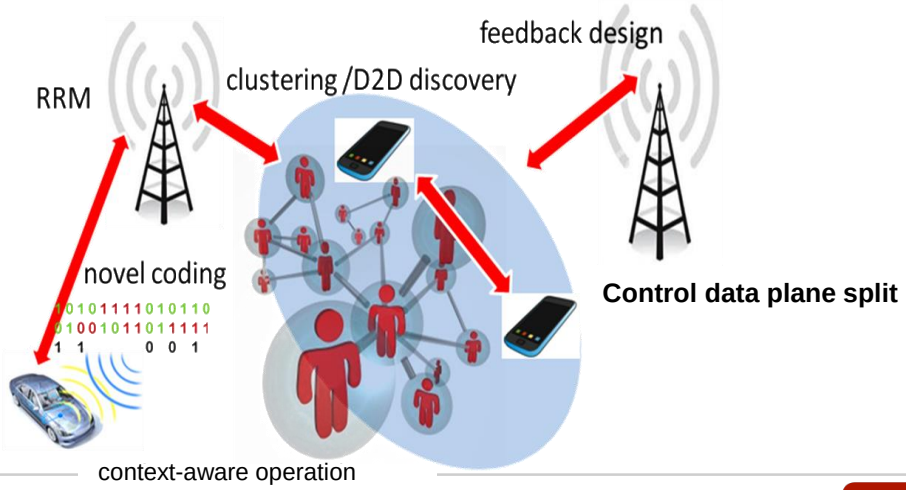
Spectral efficiency increase



Spectrum extension

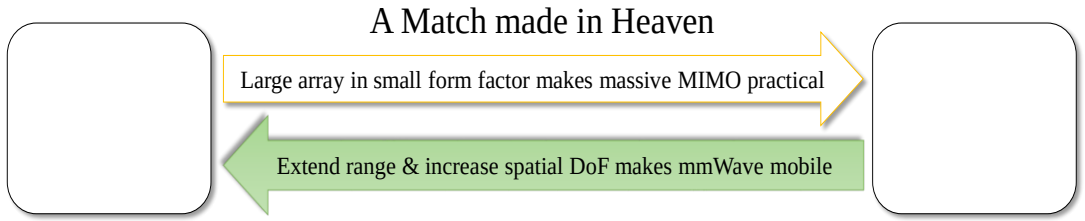
New architecture for mobile infrastructure

- **Higher density**
  - ➔ **New small cell architectures**
- **Higher spectral efficiency**
  - ➔ **Massive MIMO (M-MIMO)**
- **More bandwidth**
  - ➔ **Millimeter wave technologies**



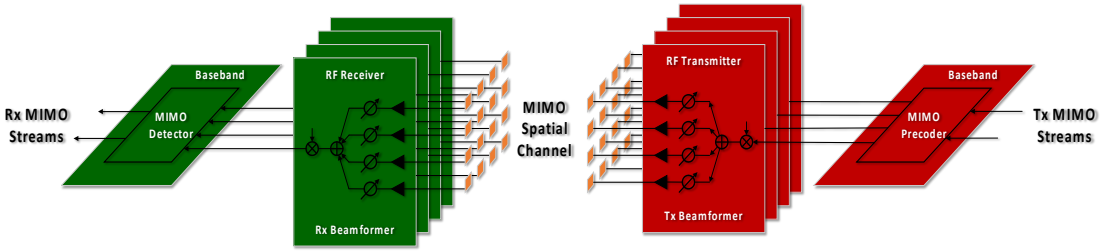
mmWave & M-MIMO: A Match Made in Heaven

- Several possible bands between 26 and 80 GHz
- Advantages:
  - Large bandwidth
  - Mobile M-MIMO
  - Inherent security
- Challenges:
  - High attenuation
  - Short coherence time
  - Baseband processing



Analog Digital Array Processing Transceiver (ADAPT)

- Analog Beamforming to track large-scale channel characteristics (AoD, AoA of dominant paths)
- Digital MIMO processing on the effective channel given analog beamforming



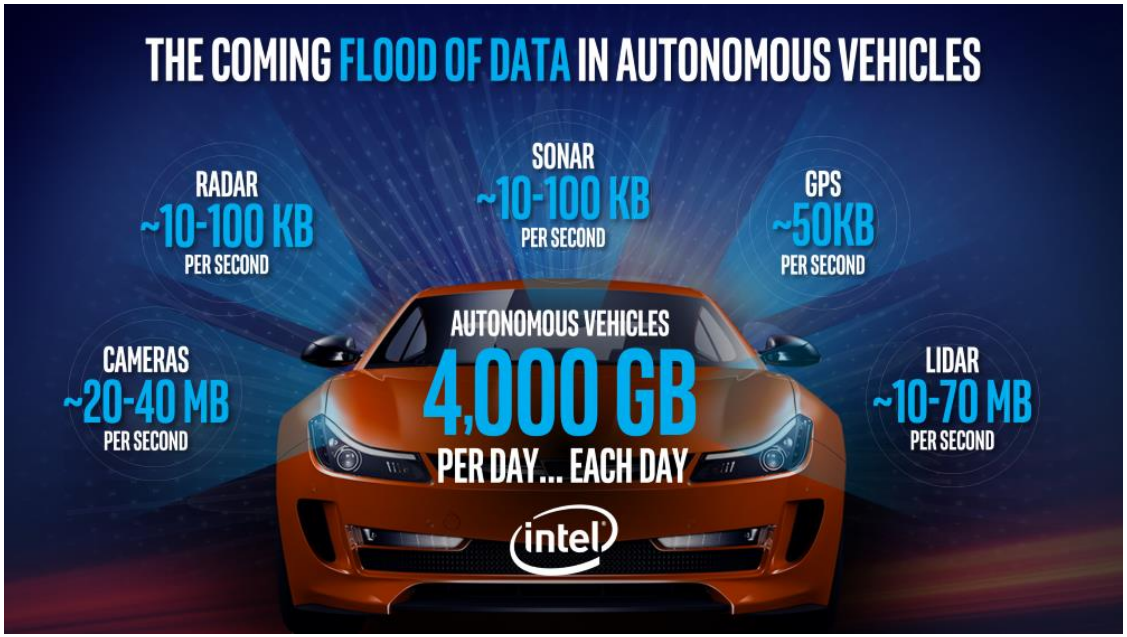
Source: A Straight Path towards 5G, Talk 3GPP RAN Workshop on 5G, Phoenix Arizona, Sept. 2015

mmW Radio for Indsutrial Communication

© Bosch

- Potential solution for production cells or distributed industrial plants
- Point to multi-point networking with directional transmission
- large bandwidth, low latency
- low interference to neighbouring systems
- Higher immunity to jamming and eavesdropping

mmW Radio for V2X Communication



© Intel

- Broadband data exchange between adjacent vehicles
- Application e.g. for platooning
- Low latency
- Low interference and high immunity to interference
- "Stand Alone" or "Non Stand Alone" in combination with 5G

What do we expect from ML?

- to help to cope with the massively increased complexity
  ➔diminish mismatch between model and reality
  ➔facilitating dense small-cell architecture
- to enhance efficiency and robustness (e.g. by reducing the number of measurements and facilitating robust decisions)
  ➔enabling massive MIMO and mmWave
- to make self-organizing feasible
  ➔cognitive network management
- to provide robust predictions
  ➔pro-active strategies, enhance outdated information
- **To provide hardware friendly, flexible and cost-effective approximations for complex models**

What are the challenges?

- High mobility
  - ➔ changes in network topology
  - ➔ wireless links exhibit ephemeral and dynamic nature
- Noisy capacity-limited transmission exposed to interference
  - ➔ wireless channel is error-prone and highly unreliable
- Stringent requirements of many 5G applications
  - ➔ Massive access
  - ➔ High reliability
  - ➔ Low-latency implementation
- Data is distributed at different locations
- Models, context information and expert knowledge are available
- There is a lot of structure in the channel, signals and functions

Massive or not massive? Does it matter?

Yes, it is essential whether you serve

- 100 devices at 1Mbps each or
- 10000 devices at 10 kbps each!

As the system uncertainty grows fast with # devices!

- # **unknowns** increases dramatically

➔ High protocol overhead for channel estimation, synchronization, user activity detection etc.

What about Reliability?

It makes a huge difference if you have to provide

- 100 Mbs 90% of the time or
- 100 kbps 99.9999999% of the time!

➔ In the high reliability regime, the system performance is very sensitive with respect to **unknowns**

Mobile networks are full of both known unknowns and unknown unknowns

Network Analytics

Reconstruction of Radio/Knowledge Maps

KPI analysis

- Assess the implications of changes of control parameters on KPIs
- Detect abnormal or highly suboptimal states

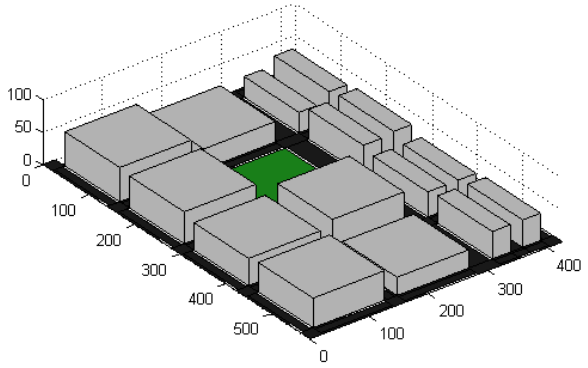
Robust Prediction of KPIs

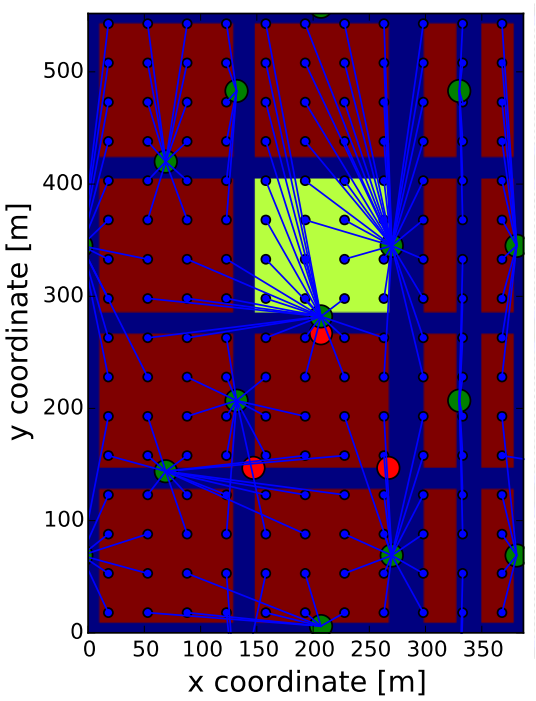
- Forecast KPIs as a function of control parameters or other information about the network (day of the week, etc.)
- Feature selection by robust dimensionality reduction
- Load prediction
- Relations among KPIs (e.g. new causality techniques)
- Anomaly detection (e.g. emergency detection in networks)

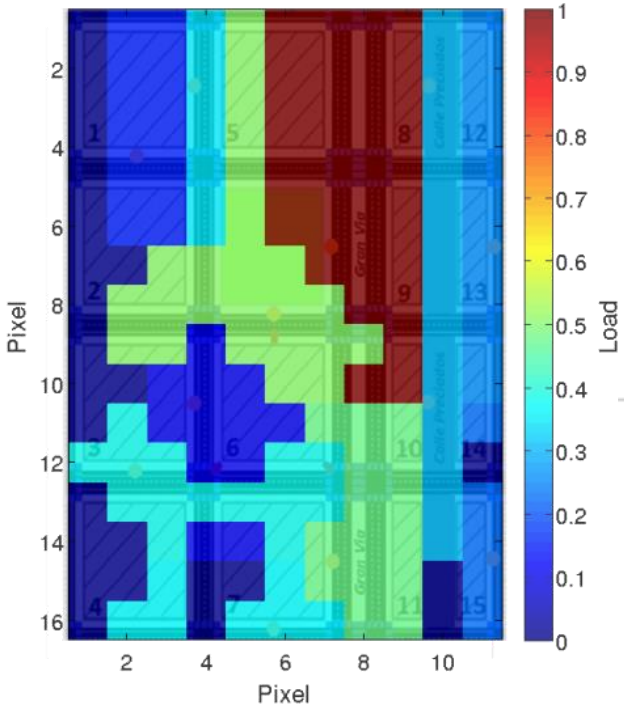
Framework for abstracting and sharing big data from infrastructures/hybrid clouds network/services/users enabling new services and businesses

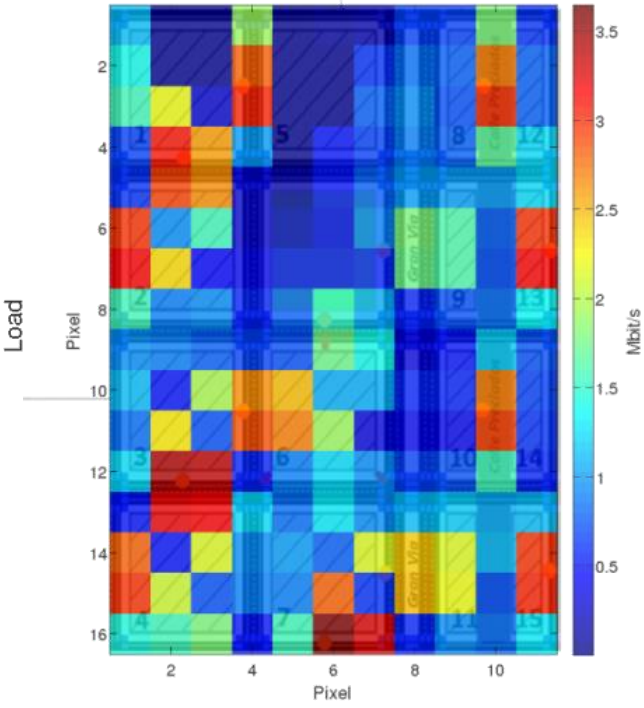
Madrid Scenario

- Madrid grid environmental model (METIS)

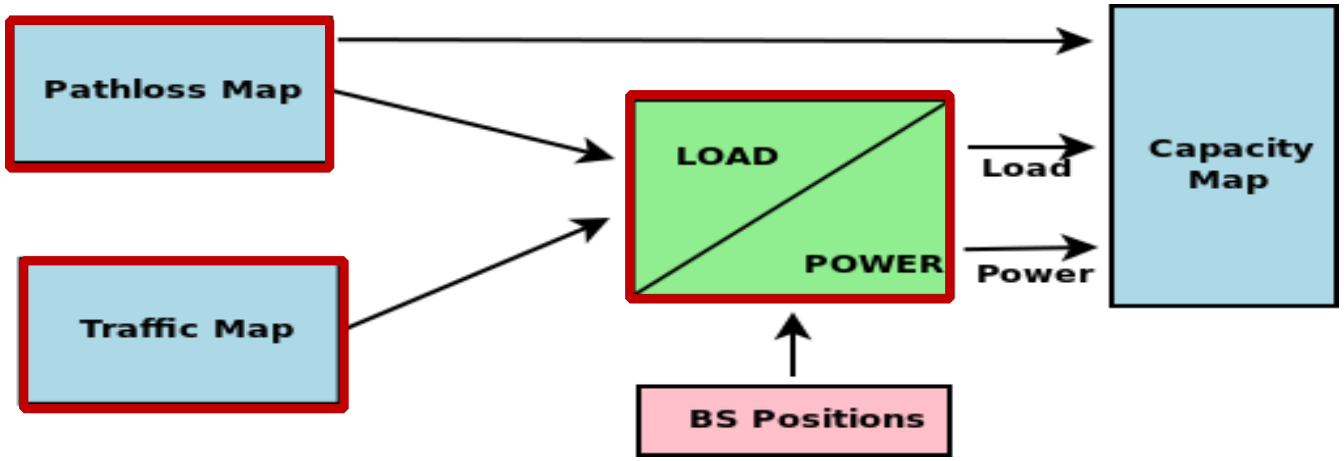








Learning Capacity Maps: Key Ingredients



- **Pathloss map**: adaptive projected sparse-aware multi-kernel approach, tensor completion (regularized approx. of rank minimization)
- **Traffic map**: Gaussian processes, Quantile estimation, context information
- **Load estimation**: hybrid-driven methods

Kasparick M., R. L. G. Cavalcante, S. Valentin, S. Stanczak, and M. Yukawa, "Kernel-Based Adaptive Online Reconstruction of Coverage Maps with Side Information," IEEE Transactions on Vehicular Technology, vol. 65, no. 7, pp. 5461-5473, July 2016

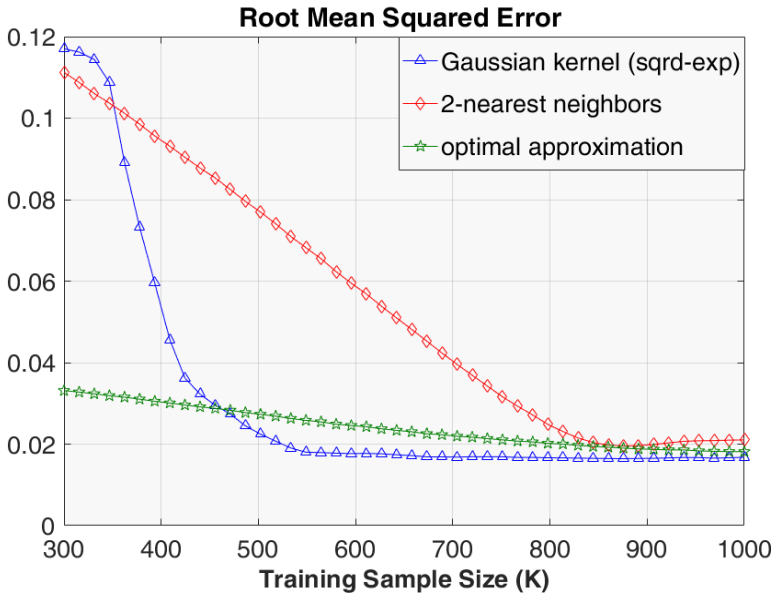
Madrid Scenario: Learning Load Maps

Objective: Given a power allocation for cells and the traffic demand for users, what is the load at each cell (fraction of the used resources)?

Challenge: The mapping relating rate to load is highly dynamic and nonlinear owing to the interference ➔ **training must be short**

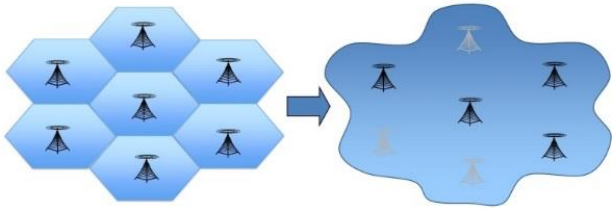
- The rate-load mapping has a **rich structure** (e.g., monotonicity) that is hard to exploit in typical machine learning tools

- New **hybrid-driven methods**: more robust and optimal, in a well-defined sense, in uncertain environments

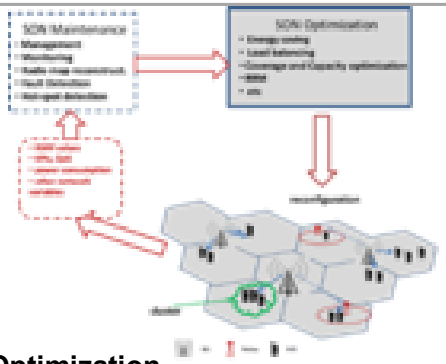


D. A. Awan, R. L. G. Cavalcante, and S. Stanczak, ``A robust machine learning method for cell-load approximation in wireless networks,´´ arXiv:1710.09318, 2017

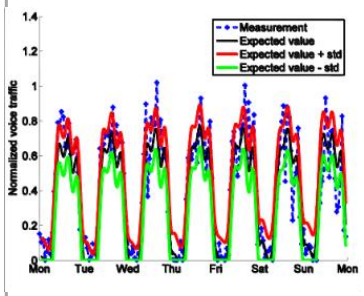
Network Planning and Optimization



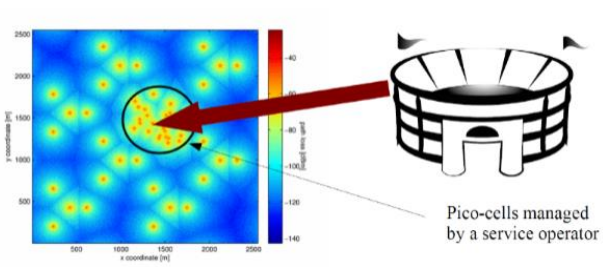
Load-dependent network configuration



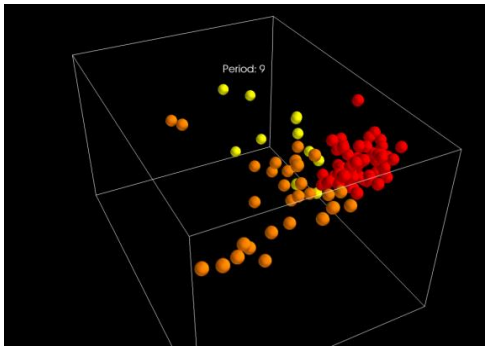
SON Optimization



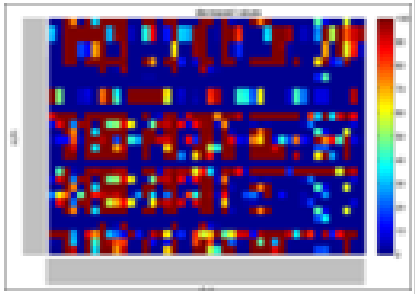
Traffic prediction



Anomaly detection, fault diagnosis, troubleshooting



Network Clustering



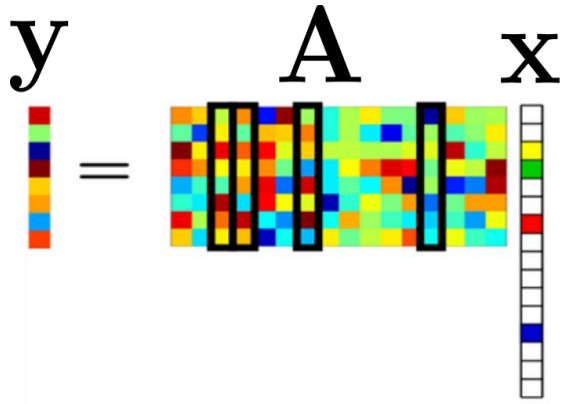
KPI Analysis

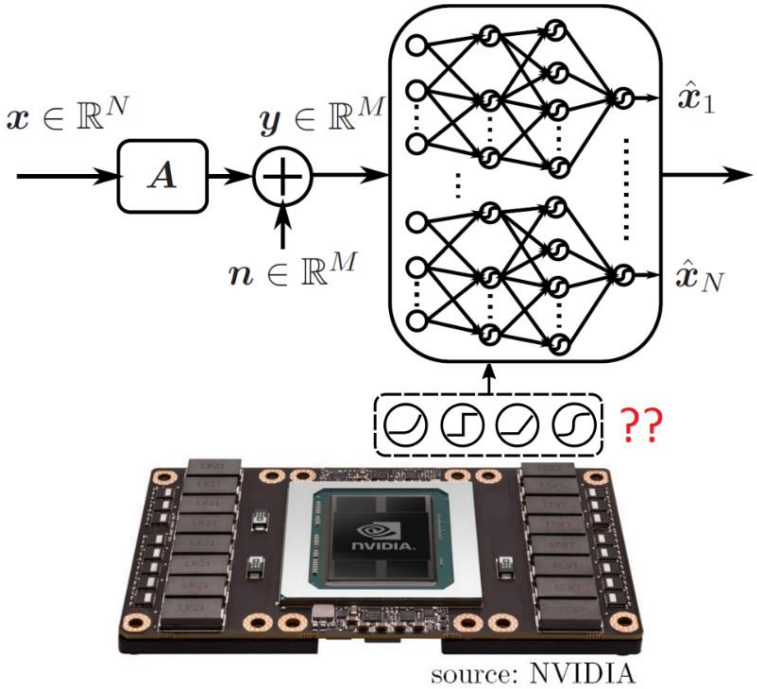
What is still missing?

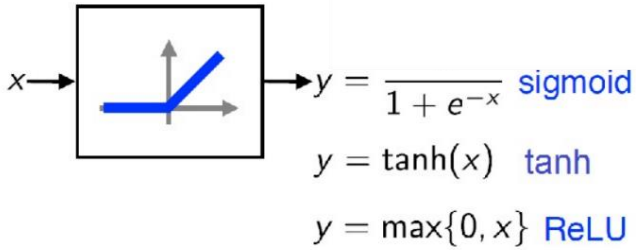
- Robust online ML with good tracking capabilities
  ➔ ML with small data sets and under latency constraints
- Exploitation of structures in signals and channels
  ➔ Dictionary learning
- Exploitation of context information and expert knowledge
  ➔ Hybrid-driven ML approaches
- Distributed learning for efficient usage of scarce resources
- New architectures for Big Data analytics
- Low-complexity, low-latency ML solutions (training networks still requires lots of resources)

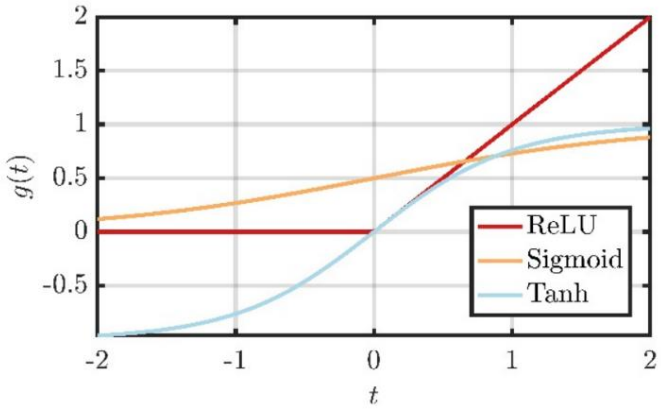
Optimal Neural Network for Sparse Recovery

- Compression: Linear dim. Reduction&quantization
- Recovery: non-linear estimator
- Need methods to chose architecture automatically
- Only require fine-tuning ➔ Very fast and robust









S. Limmer and S. Stanczak, “Optimal deep neural networks for sparse recovery via Laplace techniques“, Preprint available at arXiv, Sept. 2017

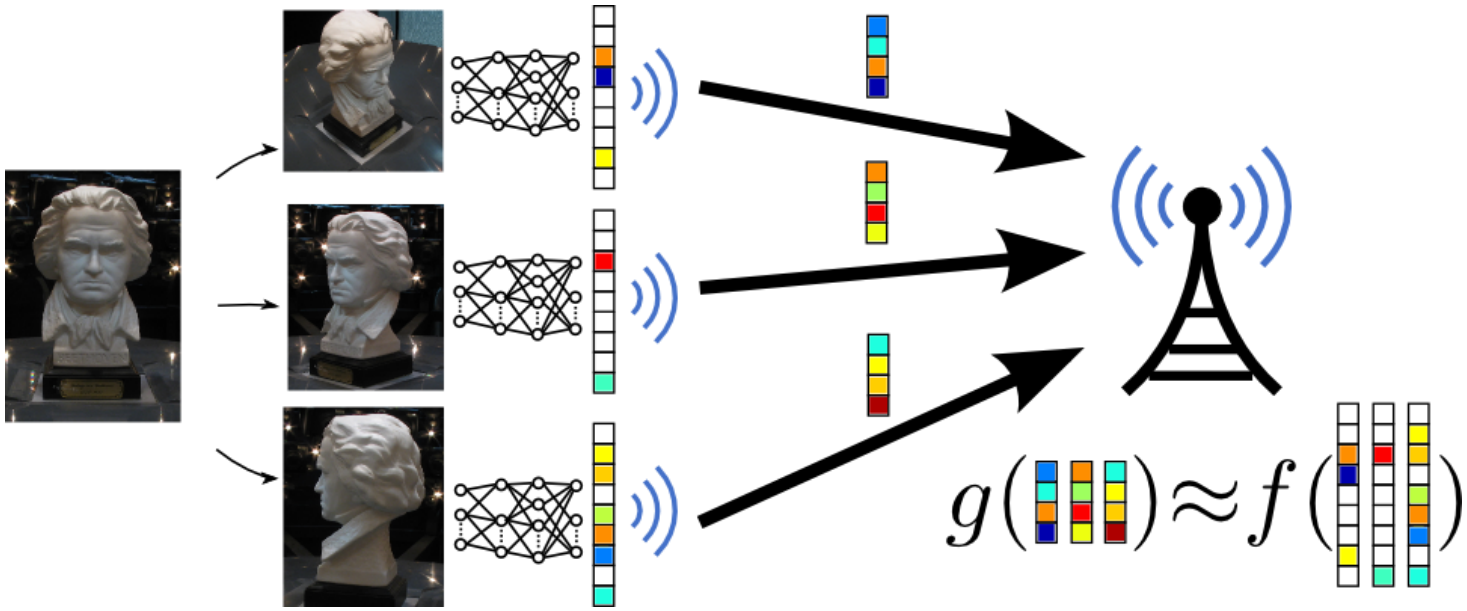
Cooperative Deep Learning

Setup: (currently 2x Jetson TX2 + PC Receiver)

1) Record multiview Images

2) Transmit compressed decisions

3) Fusion to recover original function (e.g. 'Beethoven')



Objectives (e.g. for industrial applications):

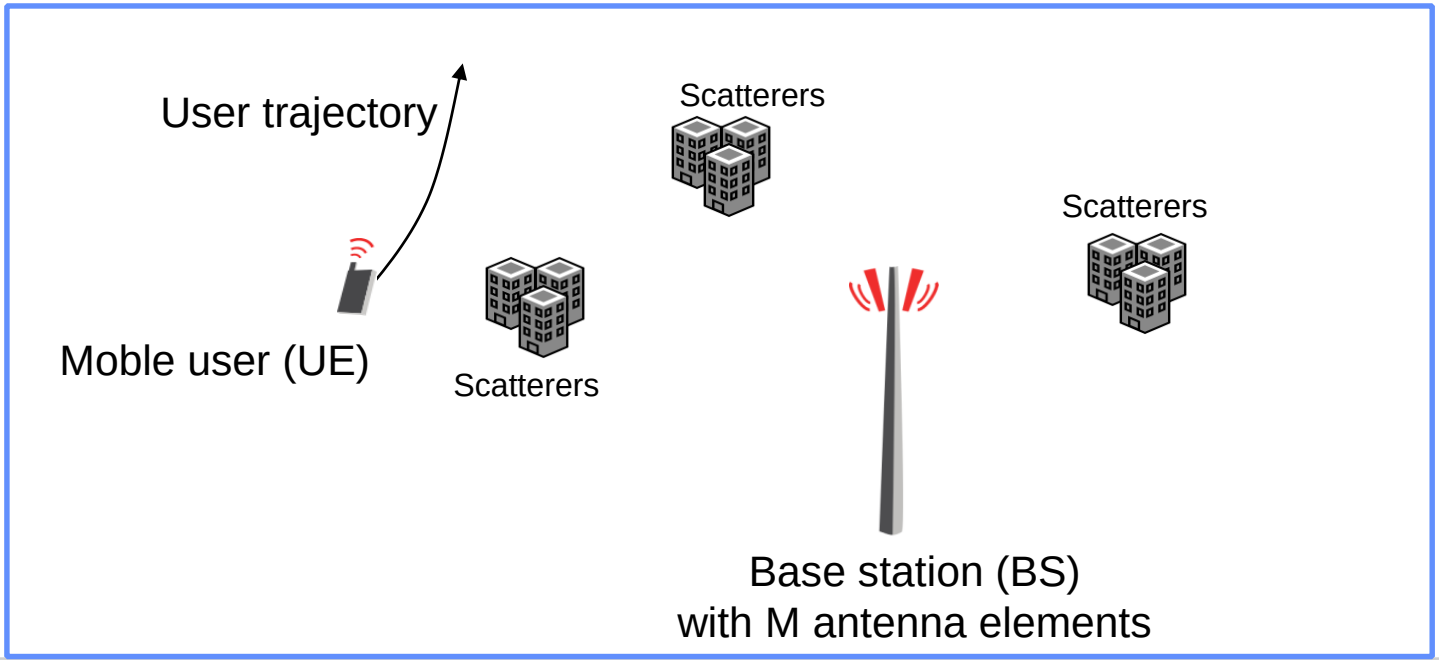
- High compression -> low latency
- Support different function: max, arith. mean, median, svm classifier

Conclusions

- Machine learning and mobile communications can be **a match made in heaven!**
- But there is a strong need for new ML methods
  - Learn feature insensitive to frequency bands, hardware implementation, signal phase …
- Hybrid-driven distributed machine learning
  - Robust online learning with good tracking capabilities
  - Distributed learning
  - Exploitation of structures, context and expert knowledge

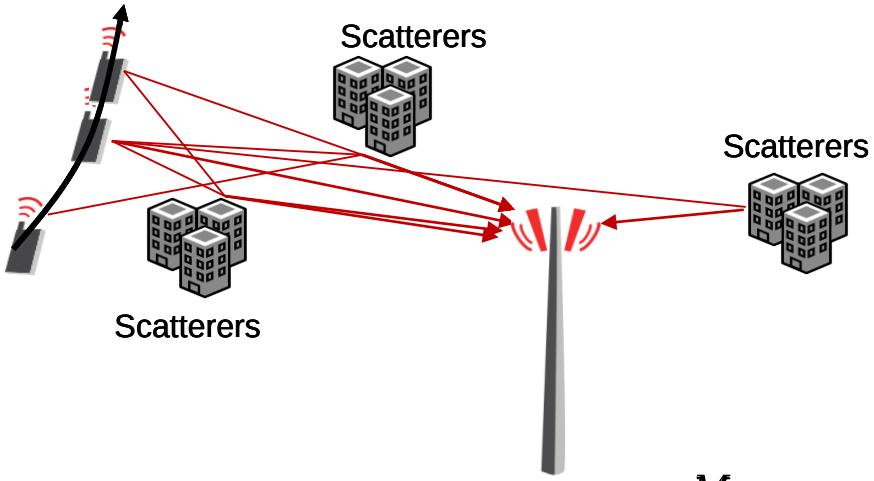
FDD Massive MIMO Channel Estimation

- Downlink and uplink use **different** frequencies (channels are different)
- It is easier to perform uplink measurements (massive MIMO regime)
- **Problem:**
  - Learn downlink covariance matrix
  - Reduce the number of measurements
  - **Downlink measurements in the massive MIMO regime**



FDD Massive MIMO Channel Estimation

- Uplink measurement (no noise):
  - UE transmits a pilot that is used by BS to estimate the channel vector



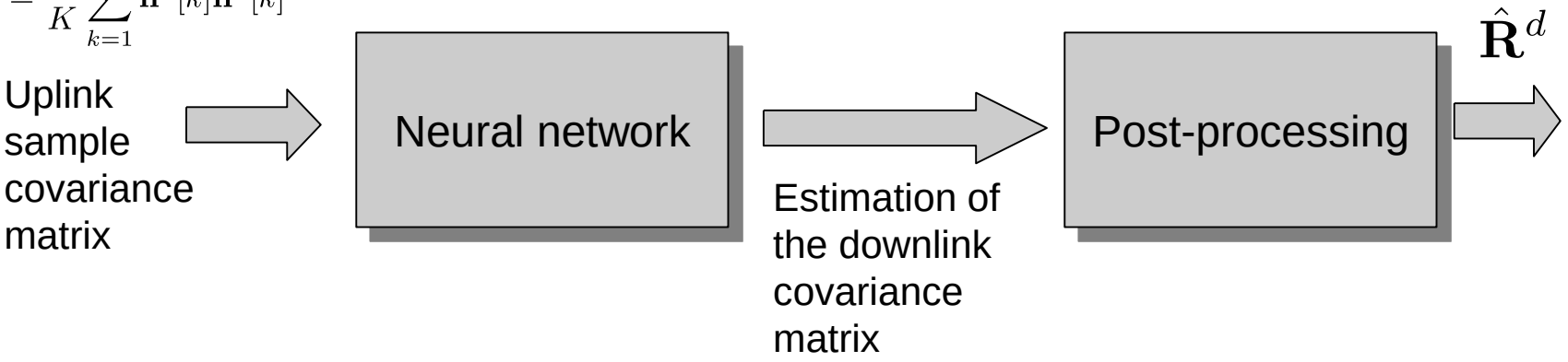
$$\mathbf{h}^u[k] \in \mathbb{C}^M$$

$$\mathbf{R}^U \approx \mathbf{C}^u = \frac{1}{K}\sum_{k=1}^{K} \mathbf{h}^u[k]\mathbf{h}^u[k]^H$$

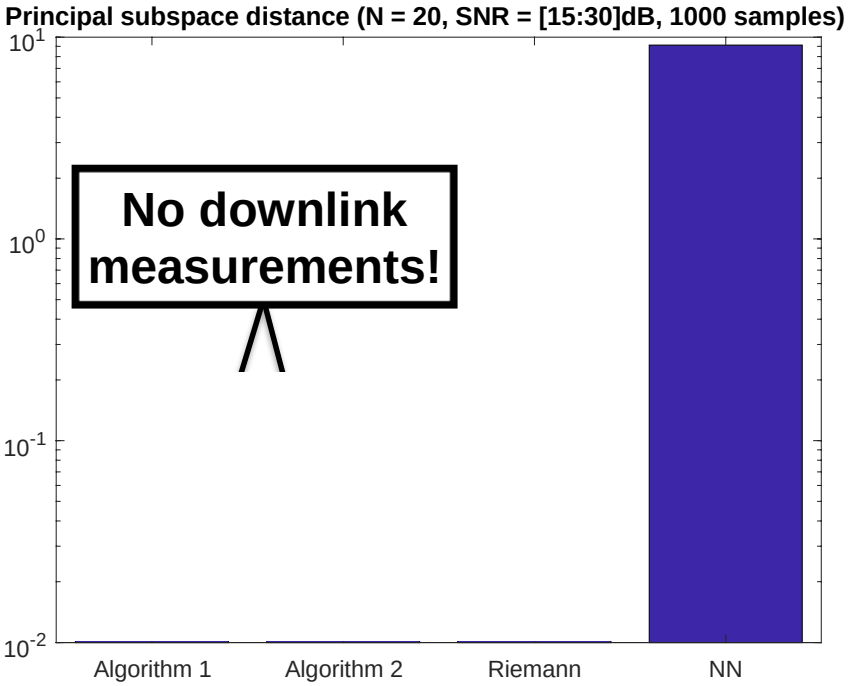
FDD Massive MIMO Channel Estimation

- Traditional ML approach: Neural Network (NN)
- Training phase based on uplink and **downlink** measurements
- After the training phase, the DNN is used for the estimation

$$\mathbf{C}^u = \frac{1}{K}\sum_{k=1}^{K} \mathbf{h}^u[k]\mathbf{h}^u[k]^H$$



FDD Massive MIMO Channel Estimation

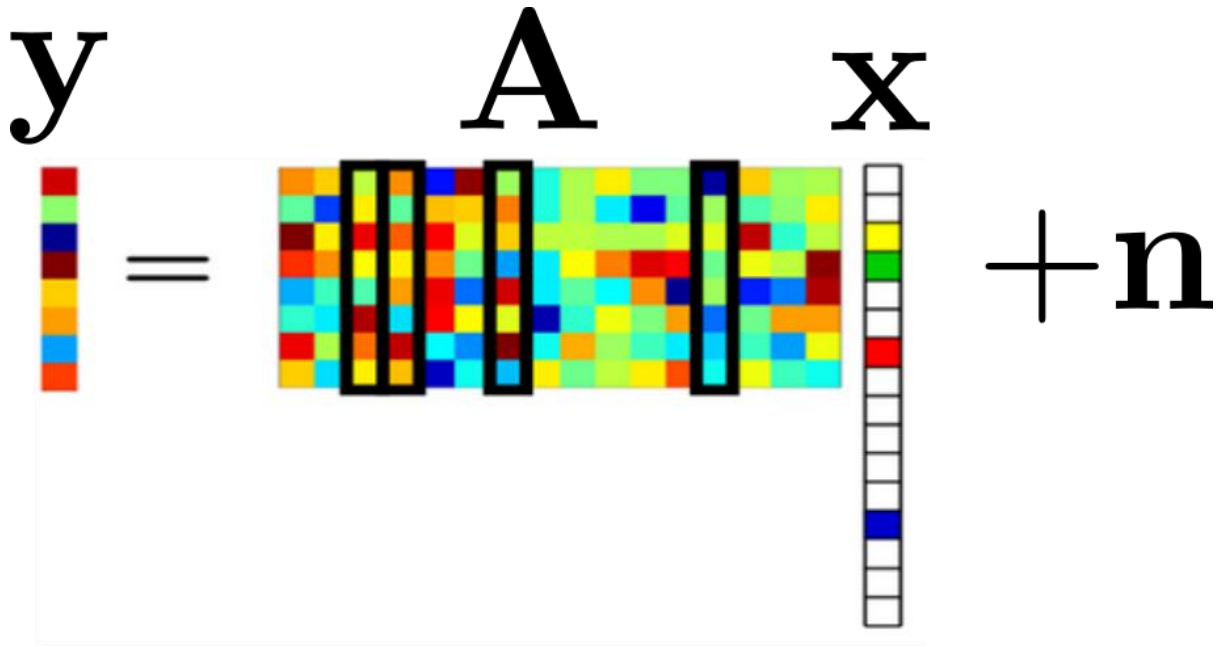


A. Decurninge, M. Guillaud, and D.T.M. Slock, "Channel covariance estimation in massive MIMO frequency division duplex systems," in IEEE Globecom, 2015

L. Miretti, R.L.G. Cavalcante and S. Stanczak, „FDD Massive Channel Spatial Covariance Using Projection Methods, Preprint, Oct. 2017

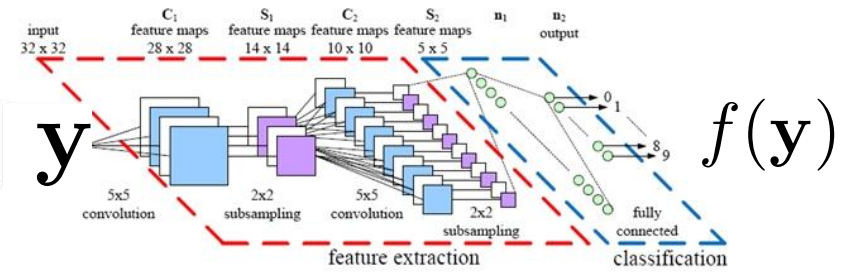
Problem: Transmission of sparse decisions

- Methods for transmission and compression of sparse decision vectors $\mathbf{x}$
- Recover sparse vector $\mathbf{x}$ from (noisy) dimension reduced $\mathbf{y}$



- Optimal solution requires solving combinatorial problem
- Recovery problems in communications require fast solutions (~1ms)
- Existing approaches are not suitable for low-latency applications

Mathematics of (Deep) Neural Networks



Source: M. Peemen et al., Speed Sign Detection and Recognition by Convolutional Neural Networks. 2011

$$f(\mathbf{y}) = \mathbf{W}_L \rho(\mathbf{W}_{L-1} \rho(\dots \rho(\mathbf{W}_1(\mathbf{y}))))$$

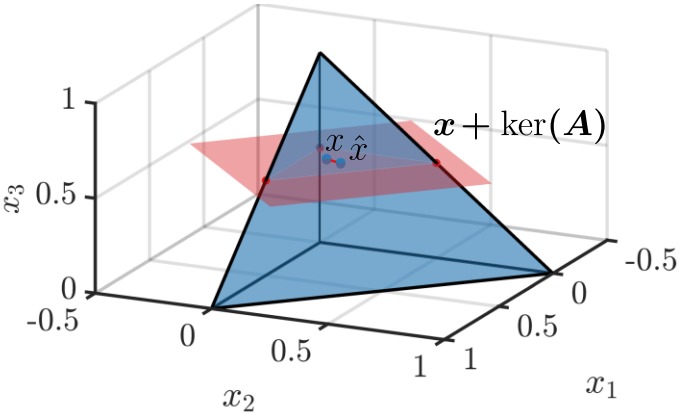
Affine linear maps

Nonlinear functions (rectifier)

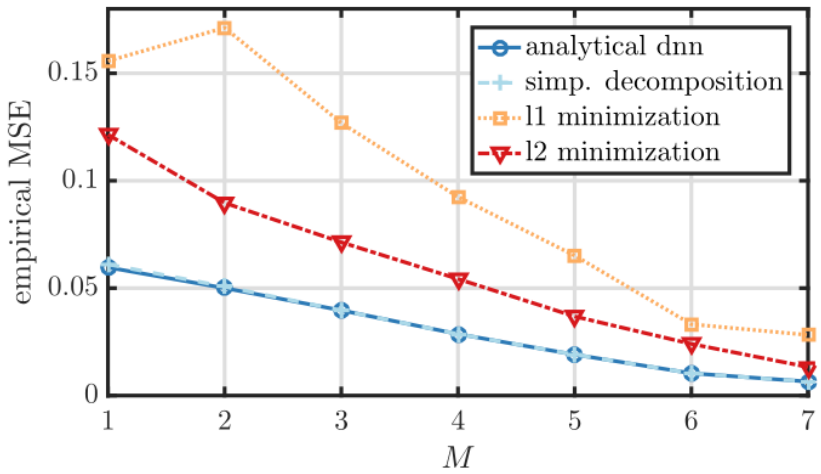
- Neural networks can approximate any function (Barron93, ‘Universal Ap proximation bounds’)
- Training networks requires lots of resources (nonconvex optimization)

Optimal Neural Network for Sparse Recovery

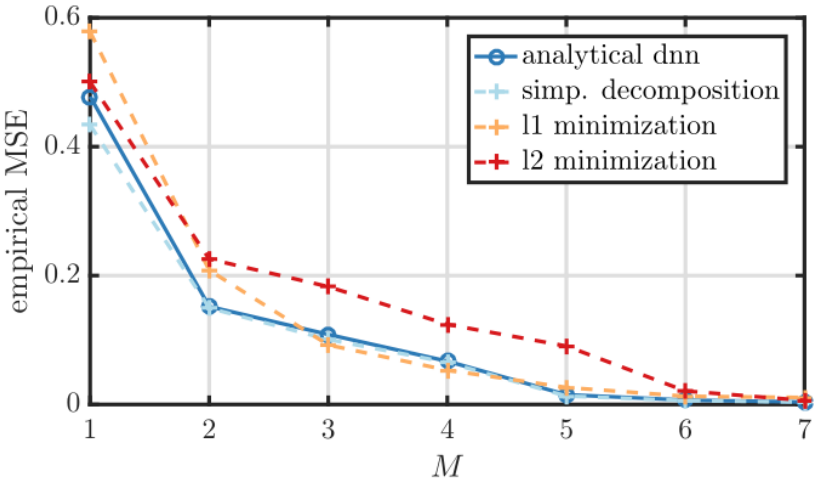
- Example: The conditional MMSE estimator is a polytope centroid under certain conditions.
- The problem reduces to:
  - Volume and moment computation
- **Implementable using the DL architecture**



Synthetic data



Real data



S. Limmer and S. Stanczak, “Optimal deep neural networks for sparse recovery via Laplace techniques“, Preprint available at arXiv, Sept. 2017