Native Network Intelligence, Fast & slow

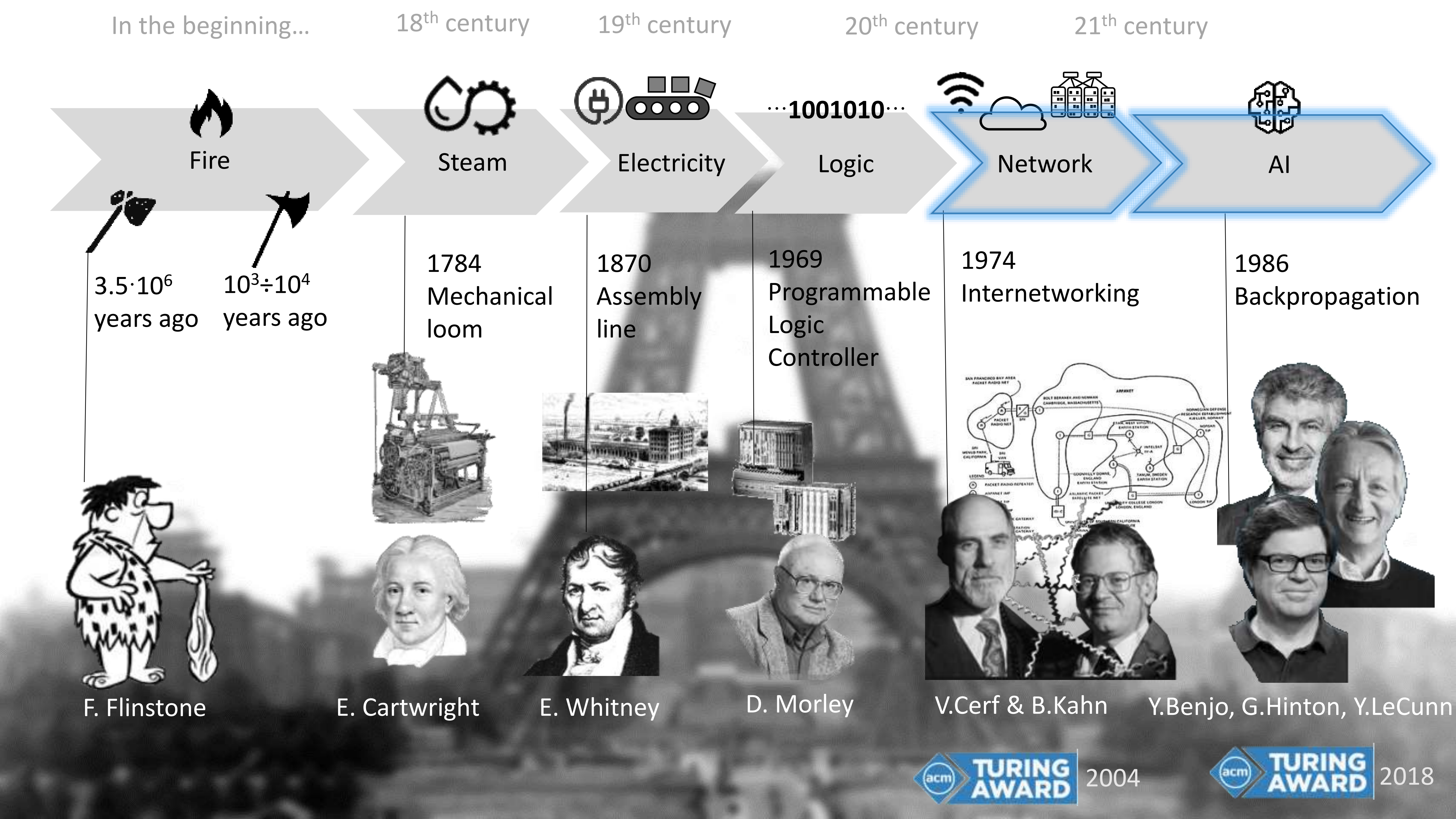
Prof. Dario Rossi
dario.rossi@huawei.com

Director, Huawei AI4NET Lab, Network Products & Solutions
Director, DataCom Lab, Paris Research Center

The past

(Paris, 1888)

*"The farther back you look,
the further ahead you can see"*
Winston Churchill


In the beginning...
18th century
19th century
20th century
21th century
Fire
Steam
Electricity
···1001010···
Logic
Network
AI
3.5·10^6 years ago
10^3÷10^4 years ago
1784 Mechanical loom
1870 Assembly line
1969 Programmable Logic Controller
1974 Internetworking
1986 Backpropagation
F. Flinstone
E. Cartwright
E. Whitney
D. Morley
V.Cerf & B.Kahn
Y.Benjo, G.Hinton, Y.LeCunn
acm TURING AWARD 2004
acm TURING AWARD 2018

AI viewpoint

Network viewpoint

- **1955** J. McCarthy "artificial intelligence"
- **1957** F. Rosenblatt Perceptron
- **1959** A. Samuel "machine learning"

- **1986** R. Dechter "deep learning"
- **1989** G. Piatetsky-Shapiro "data mining"

- **2000** I. Aizenberg "deep neural networks"
- **2012** AlexNet… re-starts the hype on DNN

Pioneering times

Independent evolution

Pioneering times

1st AI winter (late 70s)

Early cross pollination

Further independent development

2nd AI winter (late 80s)

.com bubble (late 90s)

Increased adoption

AI-Assisted Networking

- **1945** V. Bush "memex"
- **1964** P. Baran "block message"
- **1974** V. Cerf & B. Kahn Internetworking

- **1977** K. S. Narendra learning automata for telephone traffic control

- **1989** T. Berners Lee World Wide Web
- **1990s** dot com

- **2000s** all-IP
- **2010s** cloud-native & IoT

Future confluence:

AI-Native Networking

AI viewpoint

Network viewpoint

Meaning of AI-Native nets ? 1 *What*

Limits of AI-Assisted nets ? 2 *Why*

? Road toward AI-Native nets 3 *How*

FAST FWD TO THE FUTURE

BACK TO THE PRESENT

BACK TO THE FUTURE

AI-Native Networking

FAST FWD TO THE FUTURE

Let’s make a 1-slide trip to the future

1 What ?

AI Native Networking

AI as 1st class network citizen & starting point of the equation (AI+) instead of a later addendum (+AI)

- Networking paradigm where
  - exploiting AI is **seamless** and **straightforward**
  - AI is **pervasive** (and, if needed, **ubiquitous**)
  - AI brings **cost effective** irreplaceable **added value**

(as opposite to)
- Bespoke and fragile
- Suffer chicken & egg
- Increase cost

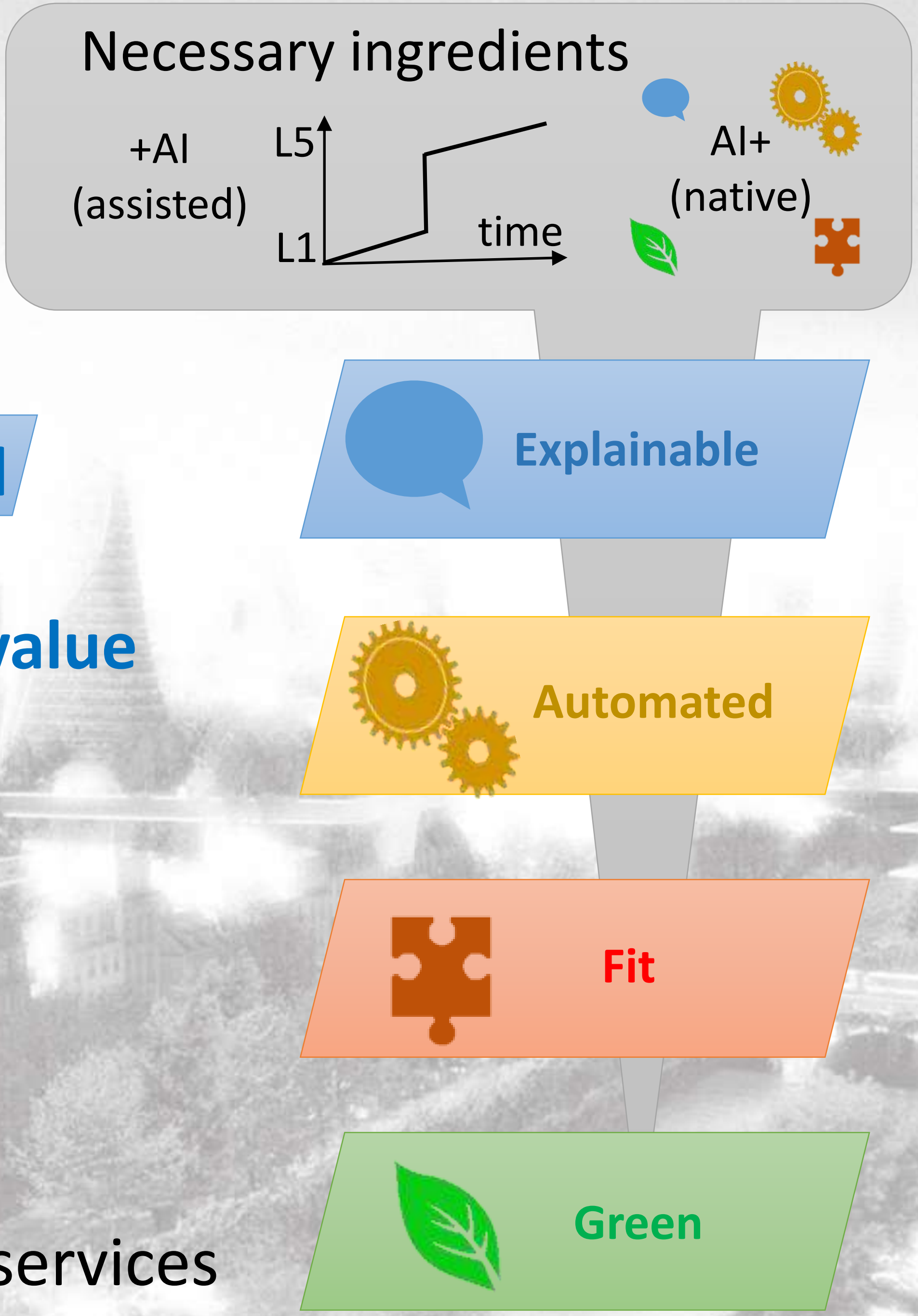
- Communication system
  - **tuned via** AI, systematically
  - **designed by** an AI, intuitively
  - **designed around** AI principles & techniques

(as opposite to)
- magic numbers
- heuristics
- model based

- Communication as a tool
  - to facilitate **interconnection of AI functions**/services

(extending)
- Human or M2M

AI Native Networking



- Networking paradigm where
  - exploiting AI is **seamless** and **straightforward**
  - AI is **pervasive** (and, if needed, **ubiquitous**)
  - AI brings **cost effective** irreplaceable **added value**
- Communication system
  - **tuned via** AI, systematically
  - **designed by** an AI, intuitively
  - **designed around** AI principles & techniques
- Communication as a tool
  - to facilitate **interconnection of AI functions**/services

BACK TO THE PRESENT

Why ? 2

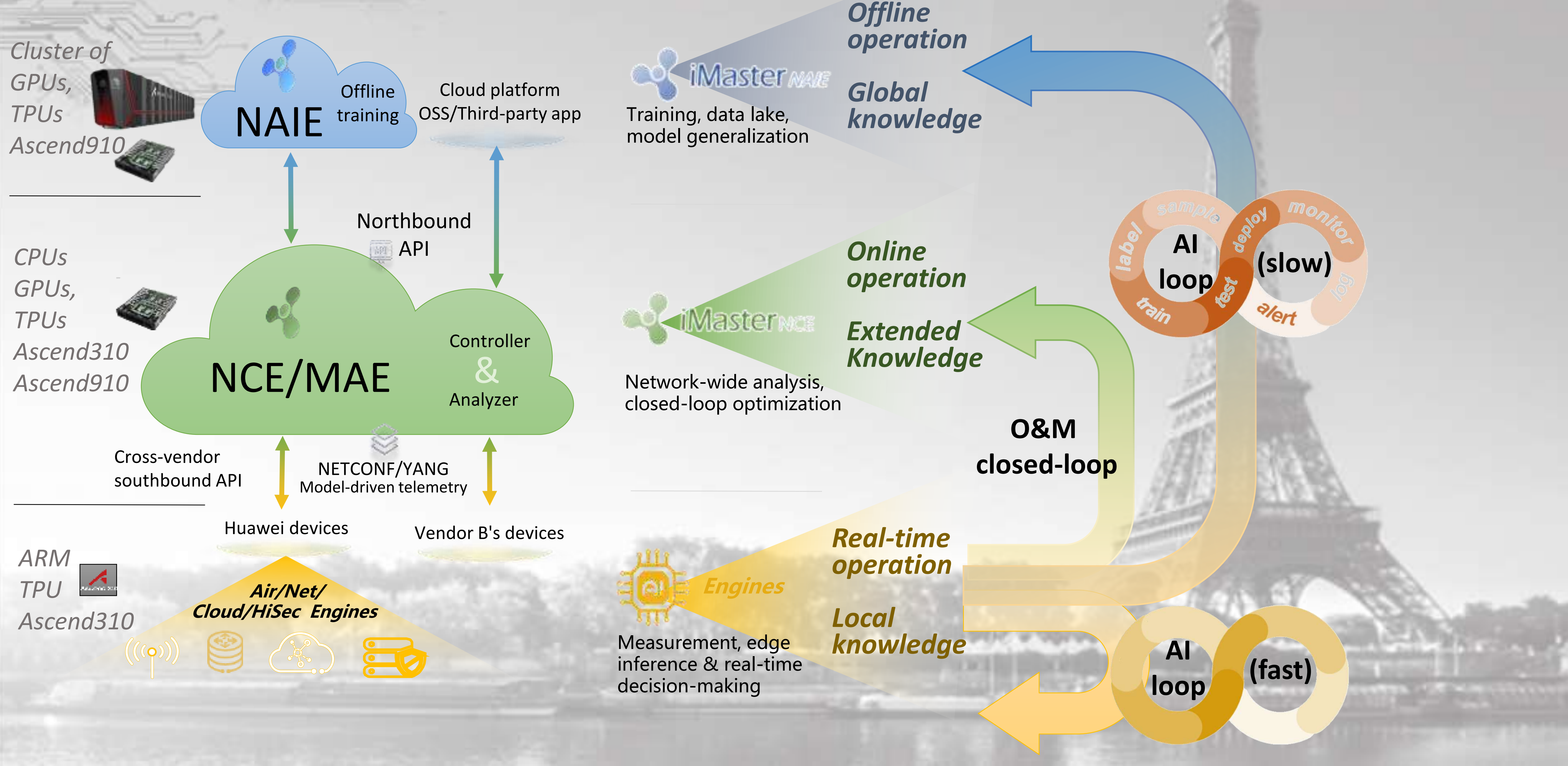
Examine AI-Assisted networking limits

Grasp AI-Native needs

Agenda | Past | **Present** | Future

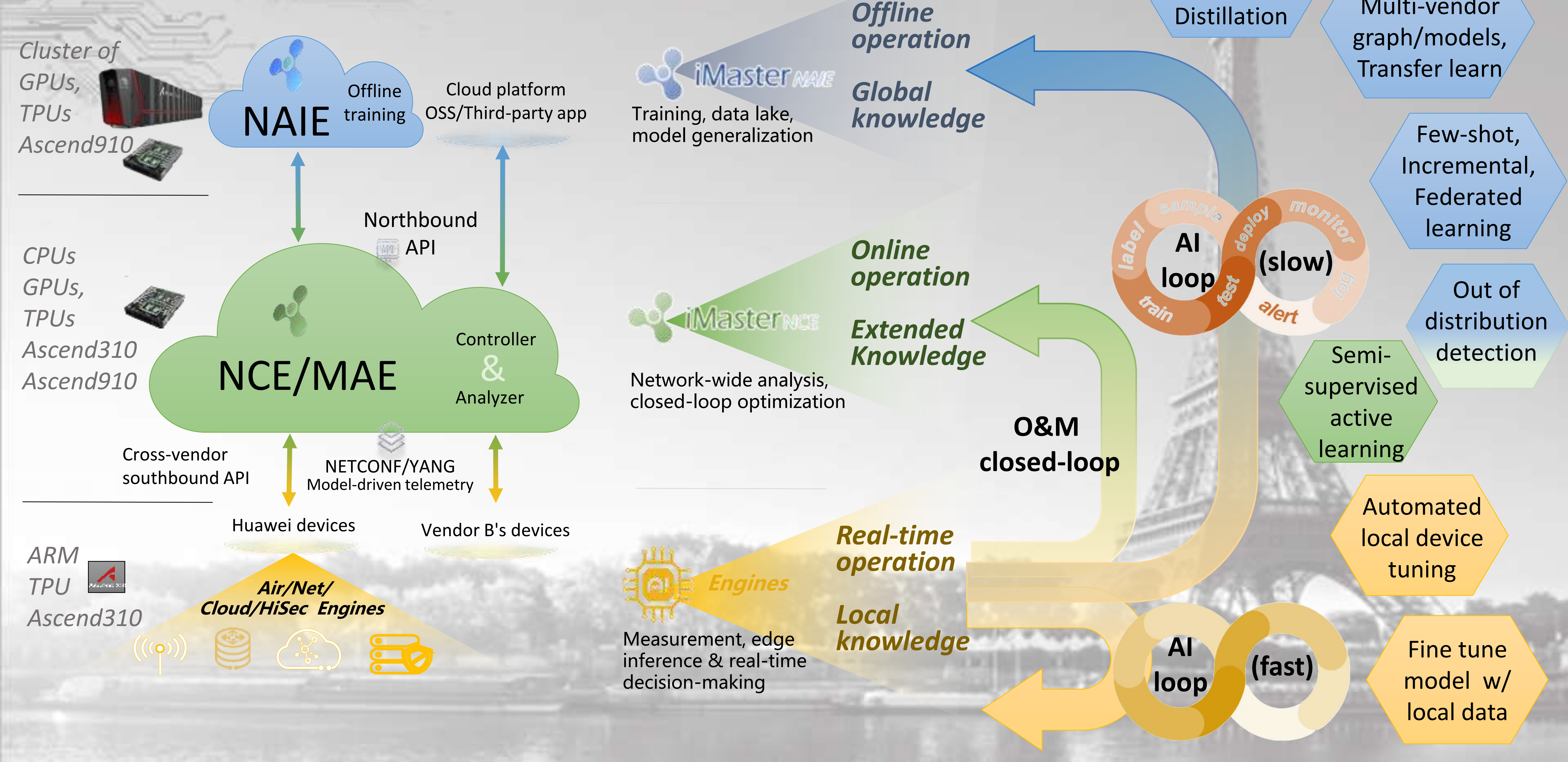
AI-Assisted networking

Evolved cloud-native 3-tiered architecture



AI-Assisted networking

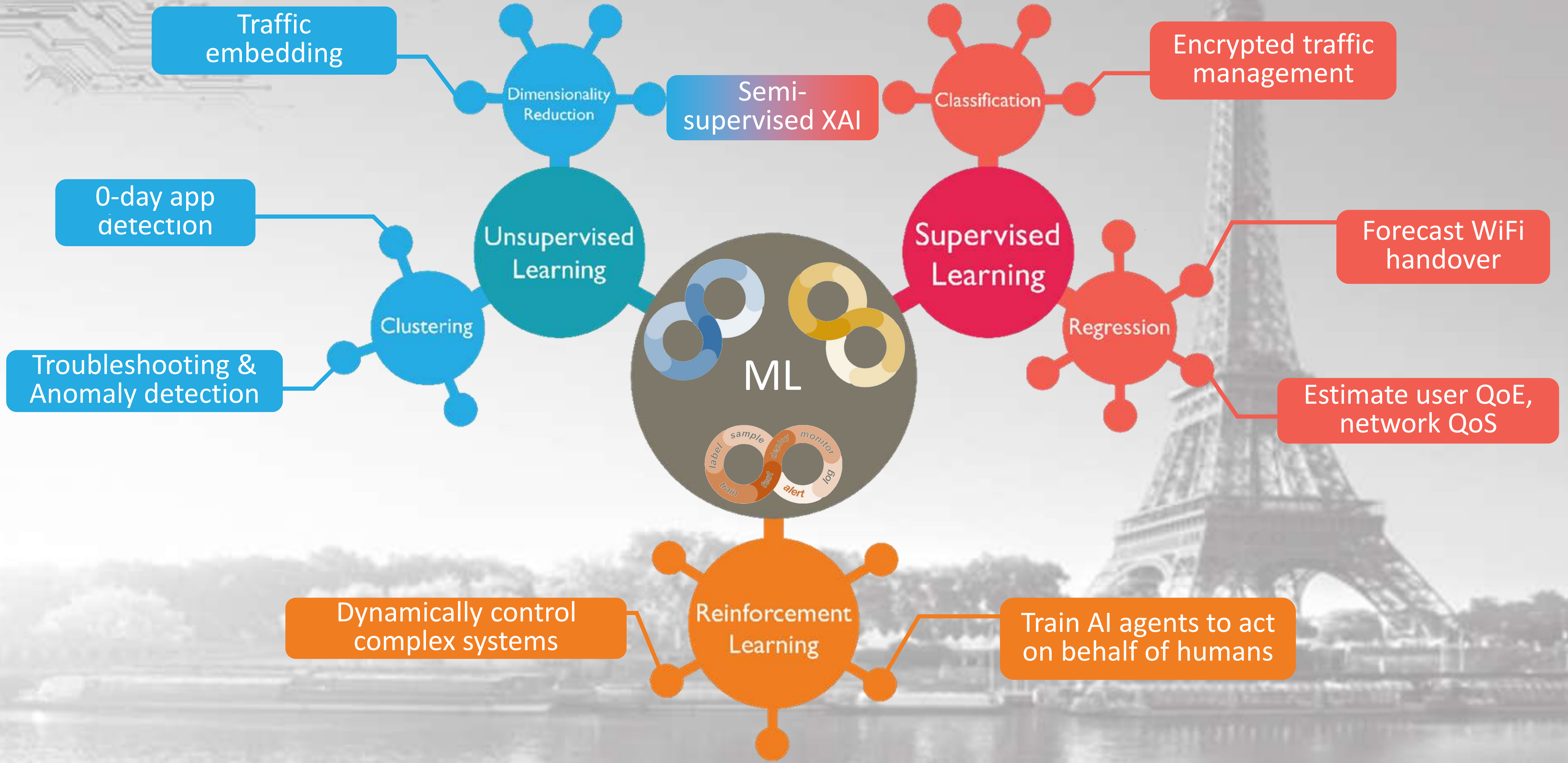
Evolved cloud-native 3-tiered architecture



AI-Assisted networking

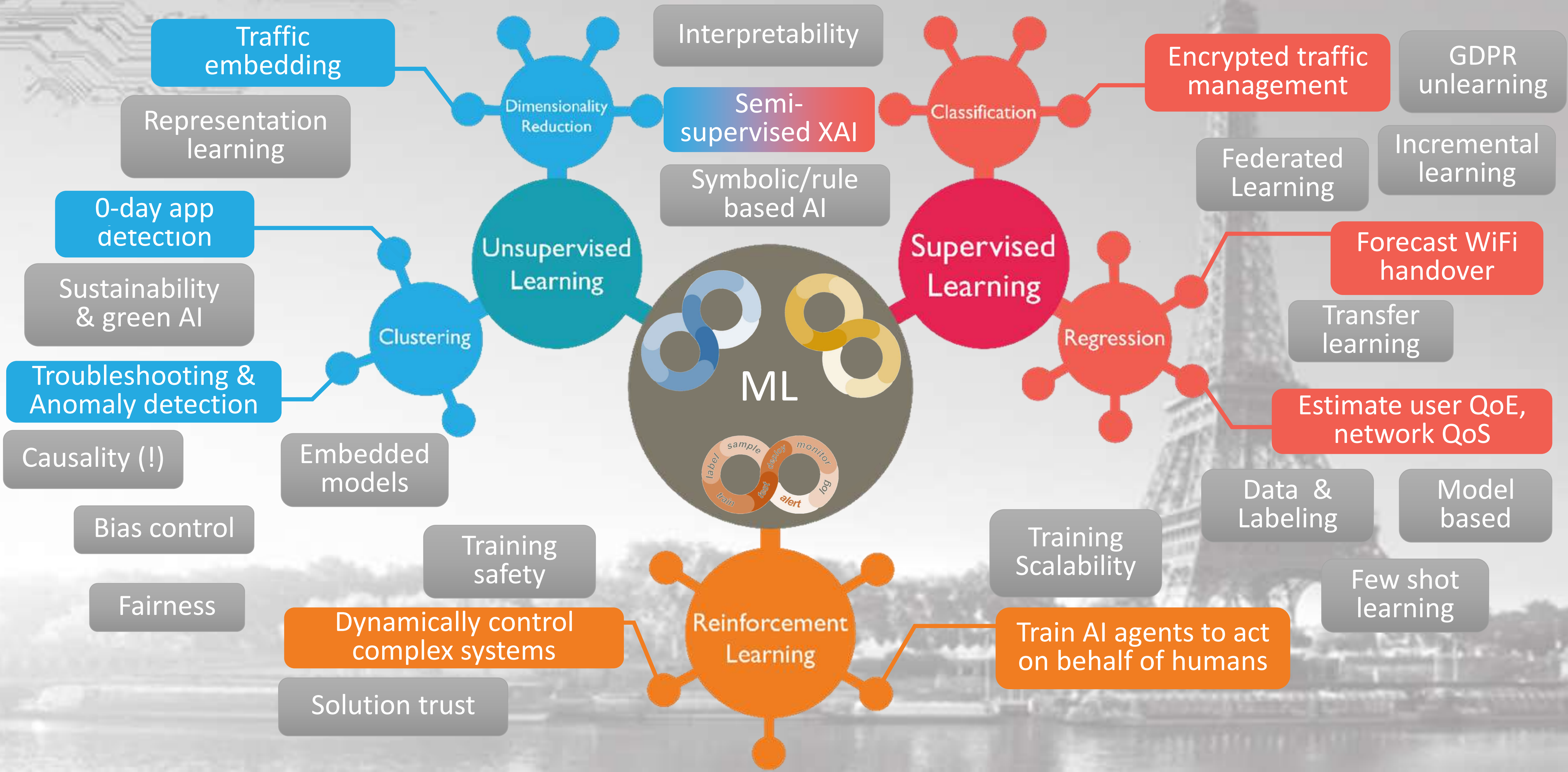
Sample of Lab research activities…

Use-cases



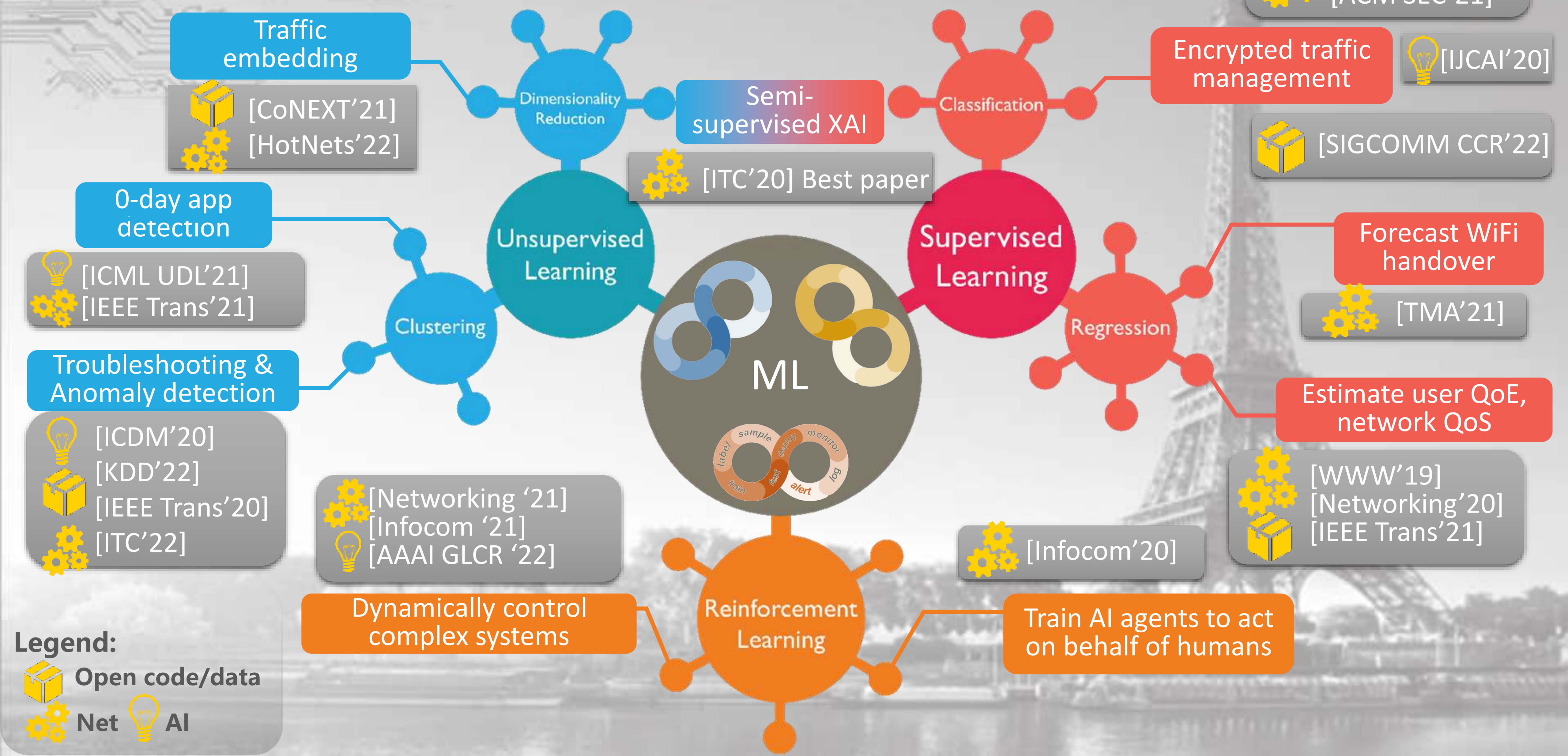
AI-Assisted networking

Sample of Lab research activities… and corresponding methods



AI-Assisted networking

Sample of Lab research activities + recent flagship publications, open code & dataset



…identify encrypted traffic application

Traffic classification

Supervised learning

Legal compliance

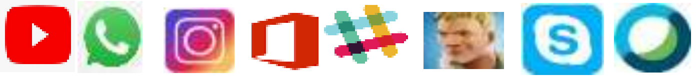
GDPR & AIAct

Product Engineer

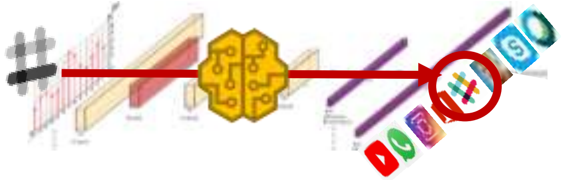
Enterprise +Residential
40TB, 10M flows
Commercial DPI

Data Engineer

Top-200 apps >95% traffic (out of >3000 labels)



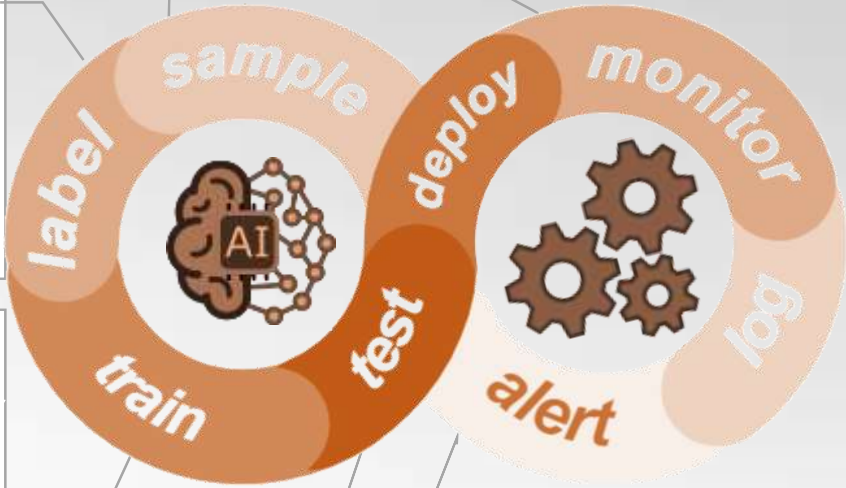
AI Scientist



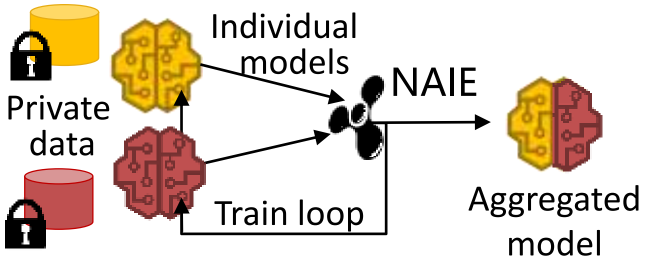
Recognize existing applications
>95% class & >90% app accuracy
[IEEE Trans'21]

>10x simpler & faster than state of the art
- 150us on single ARM core (A72)
- ARM Cache + Ascend310 support

[SEC'21][SIGCOMM'20 Demo][INFOCOM'22]

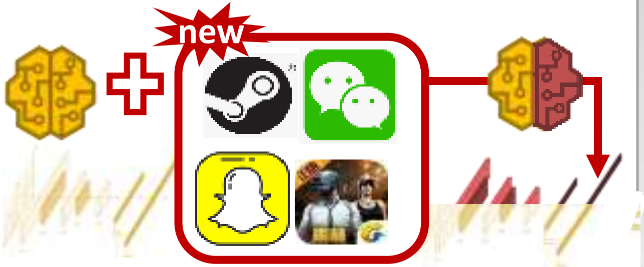


Distributed training, keeping data private



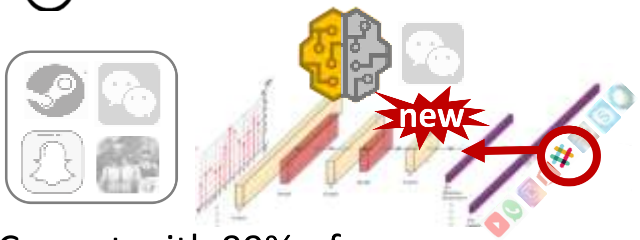
[EP20/061440] [IJCAI FL'20]

Train incrementally & from few samples



Old model + New samples = New model
[TMA'21]

Detect zero-day application



Correct with 99% of new apps
[EP21/057212][ICML UDL'21][ComMag'21]

Network troubleshooting

Unsupervised Learning

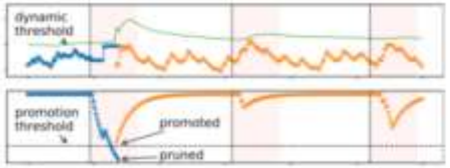
Unsupervised Outlier detection

Generic algorithms
Test several datasets, including network data

RHF (Batch) [ICDM'20]

$$RHT_i(p) = \log\left[\frac{1}{P_{Q_j}}\right], \quad p \in S(Q_j),$$

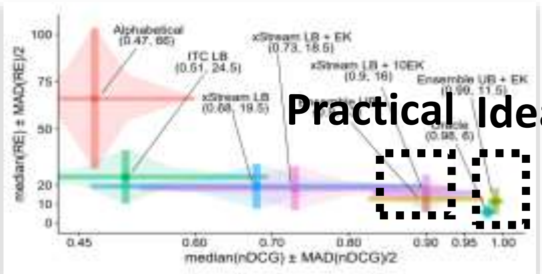
ODS (Streaming) [IEEE Trans'21]

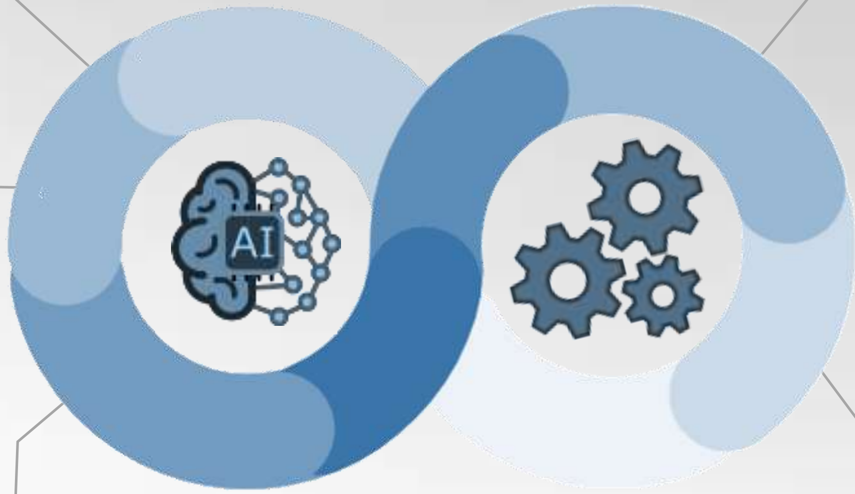


Semi-supervised XAI

Complementary expert knowledge

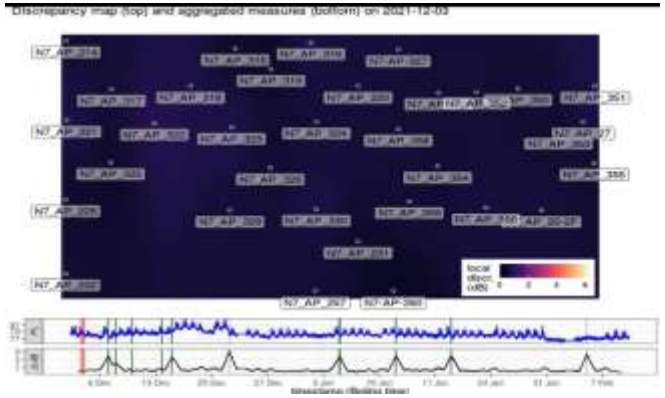
Best paper at [ITC32]





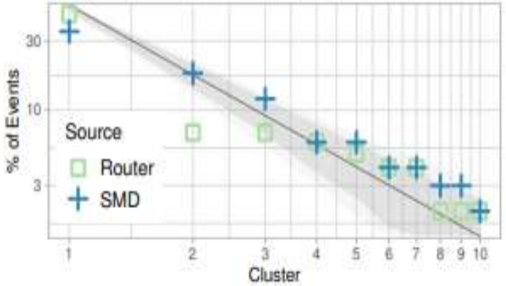
Changepoint detection

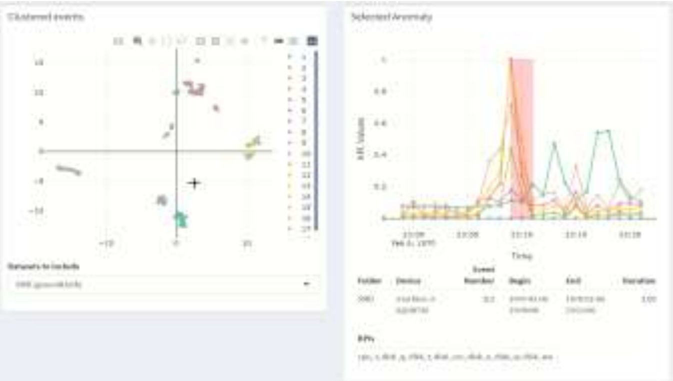
Focus on detection of impactful events



Clustering

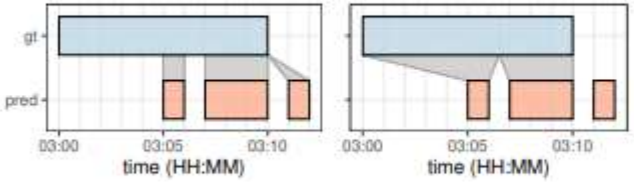
Business-specific anomaly DB





Unbiased evaluation [KDD'22]

Theoretically principled metrics, provably robust against adversarial prediction



Train on simulator

Validate on controlled settings

Test on real network

Actions = Channel, Power

Reward = *f*(coverage, interference)

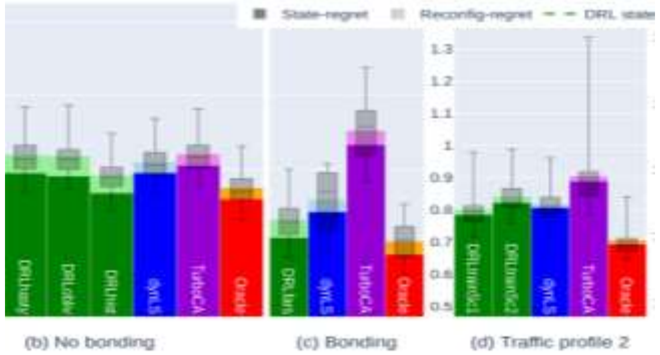
State = AP KPIs or 802.11k STA data

Autonomous network configuration

Reinforcement Learning

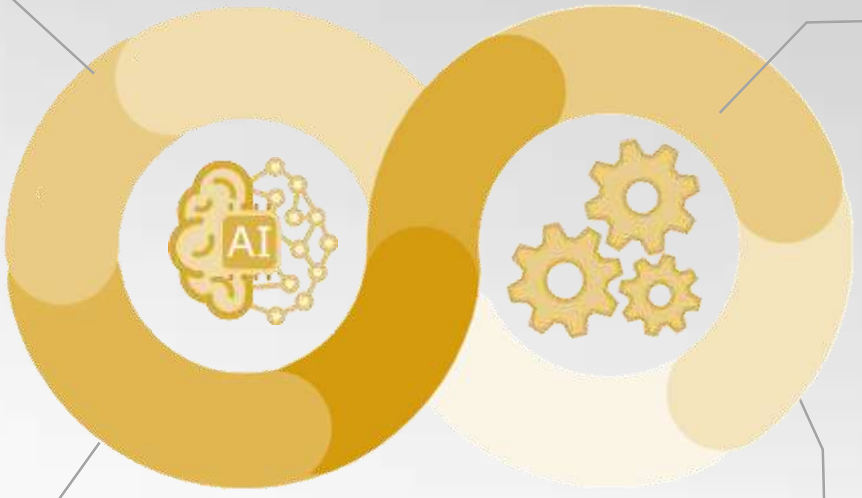
Simulation

Compare Deep RL vs Optimal oracle / TurboCA

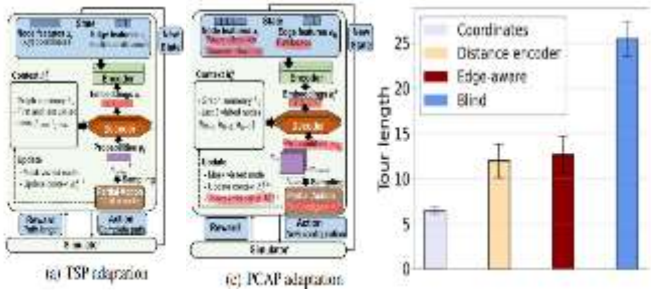


[Networking'21]

Principled *>100,000 states explored at training, unbiased academic-style comparison*



Innovate, and loop again



[AAAI GLRC'22]

Transformer-based (the T in GPT-3 stands for transformer)

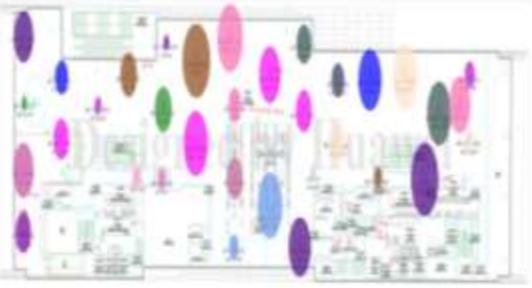
Controlled *Generalization ability validated on different settings w.r.t. training*

clear-cut-threshold

Noise mean [dB] \ Noise std. [dB]	0	1	2	4	8	16
-16	60.1%	58.0%	58.6%	55.6%	54.7%	47.5%
-8	22.1%	21.9%	21.5%	23.0%	25.7%	29.2%
-4	7.9%	8.0%	8.6%	10.1%	15.0%	23.3%
-2	4.4%	3.9%	4.2%	6.1%	11.3%	18.9%
-1	1.8%	2.2%	2.4%	4.5%	9.2%	18.2%
0	0.0%	0.6%	1.8%	3.3%	8.0%	16.3%
1	0.4%	0.4%	1.0%	2.5%	7.1%	14.8%
2	0.6%	0.7%	0.8%	2.0%	6.3%	14.2%
4	1.2%	1.4%	1.2%	1.7%	4.5%	11.4%
8	2.0%	2.0%	2.2%	2.2%	3.3%	8.9%
16	4.7%	4.6%	4.6%	4.4%	4.0%	6.0%

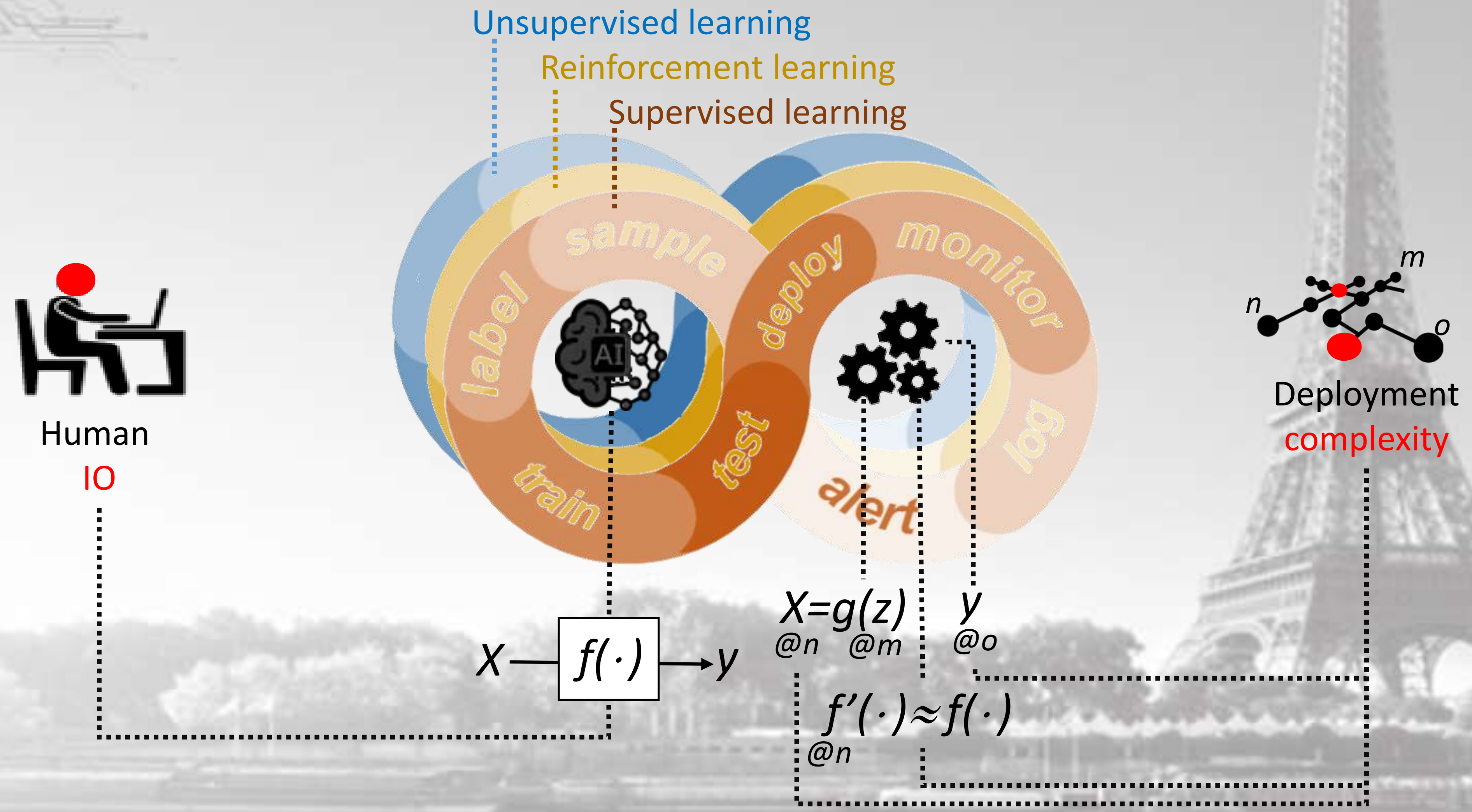
[ArXiv'22]

Real network Real deployment, Months of operation with >10,000 users



Demo@[Infocom'21]

Limits of AI-Assisted networking



Limits of AI-Assisted networking

Human IO
Input

Humans

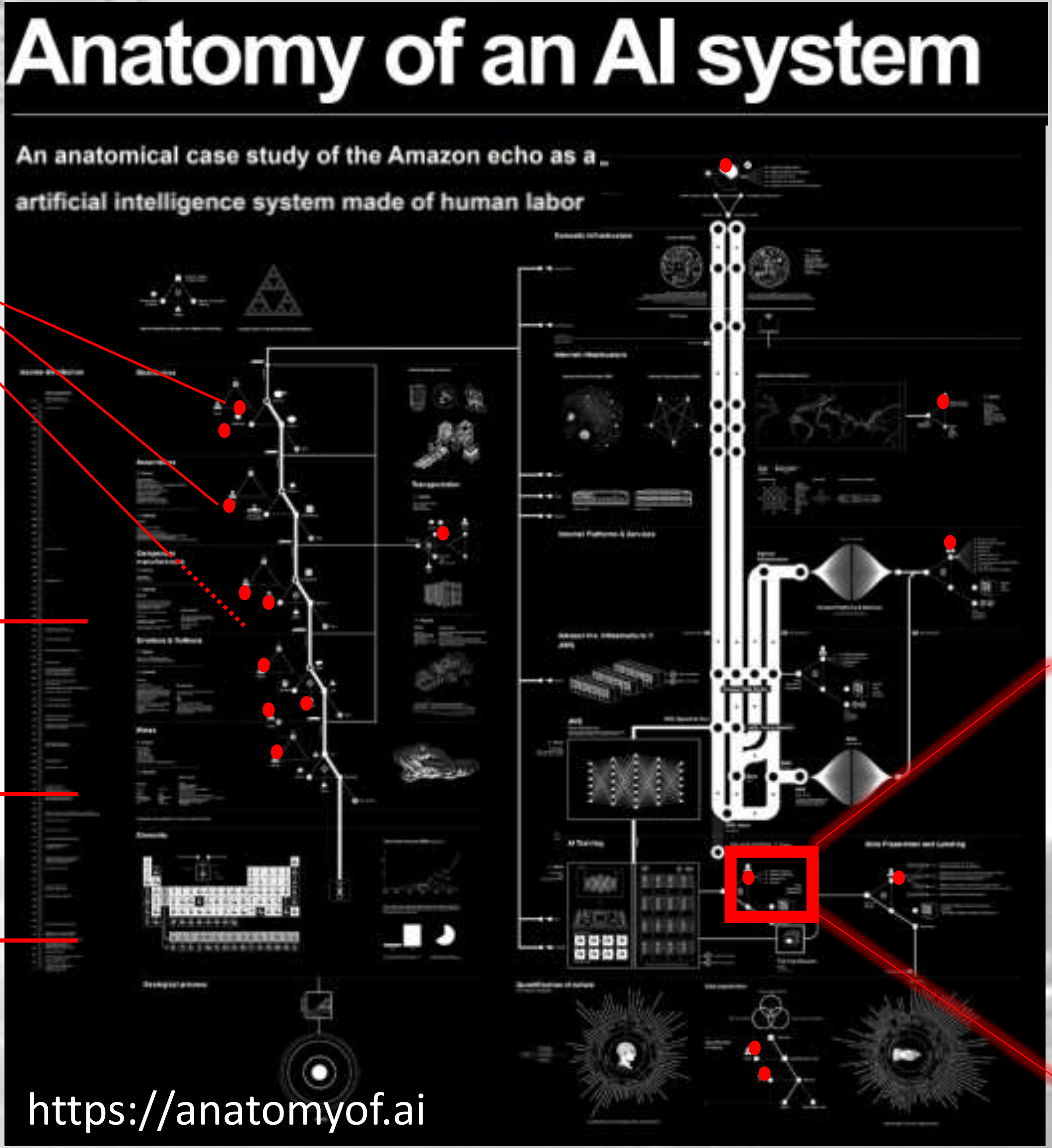
Salary USD/mo

10k

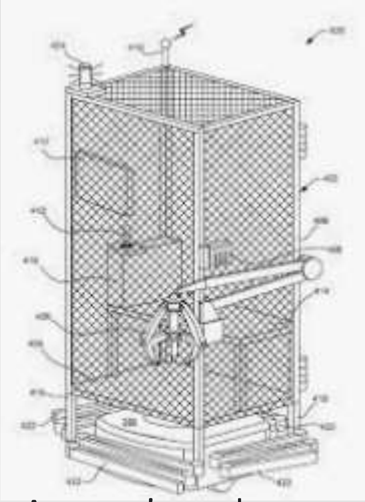
5k

1k

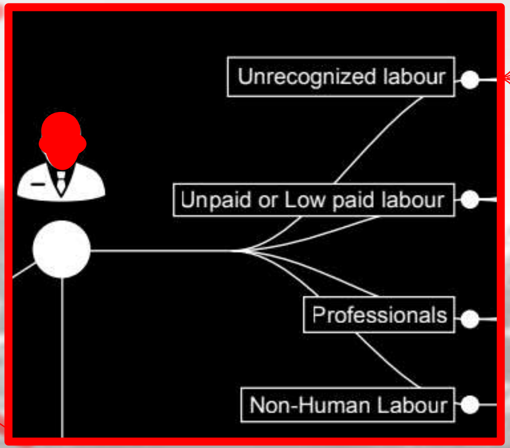
US poverty line

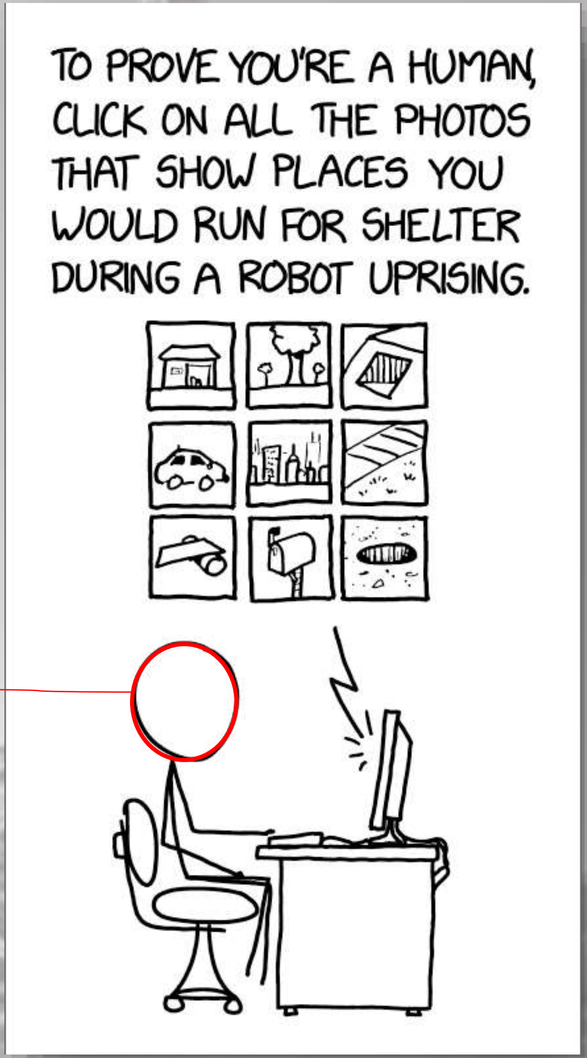


https://anatomyof.ai



Amazon's worker cage
[US Patent 9,280,157]





https://xkcd.com/2228/

Limits of AI-Assisted networking

Human IO

Output

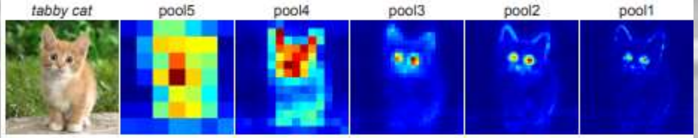
ML System → Output

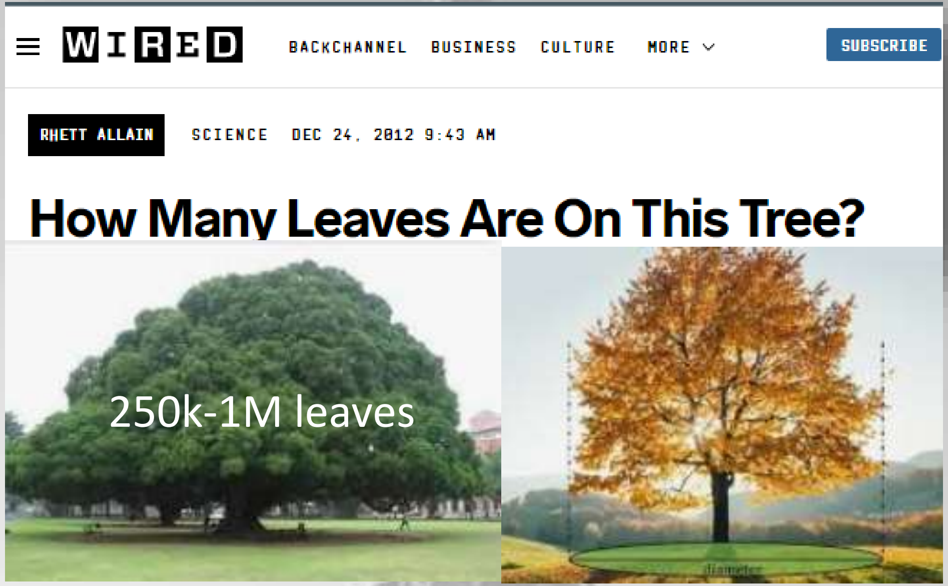
$X \rightarrow \boxed{f(\cdot)} \rightarrow y$

https://xkcd.com/1838/

"Clever hans" models, good (or bad) for the wrong reason

(a) Husky classified as wolf (b) Explanation





(I've seen RF models >>1M)

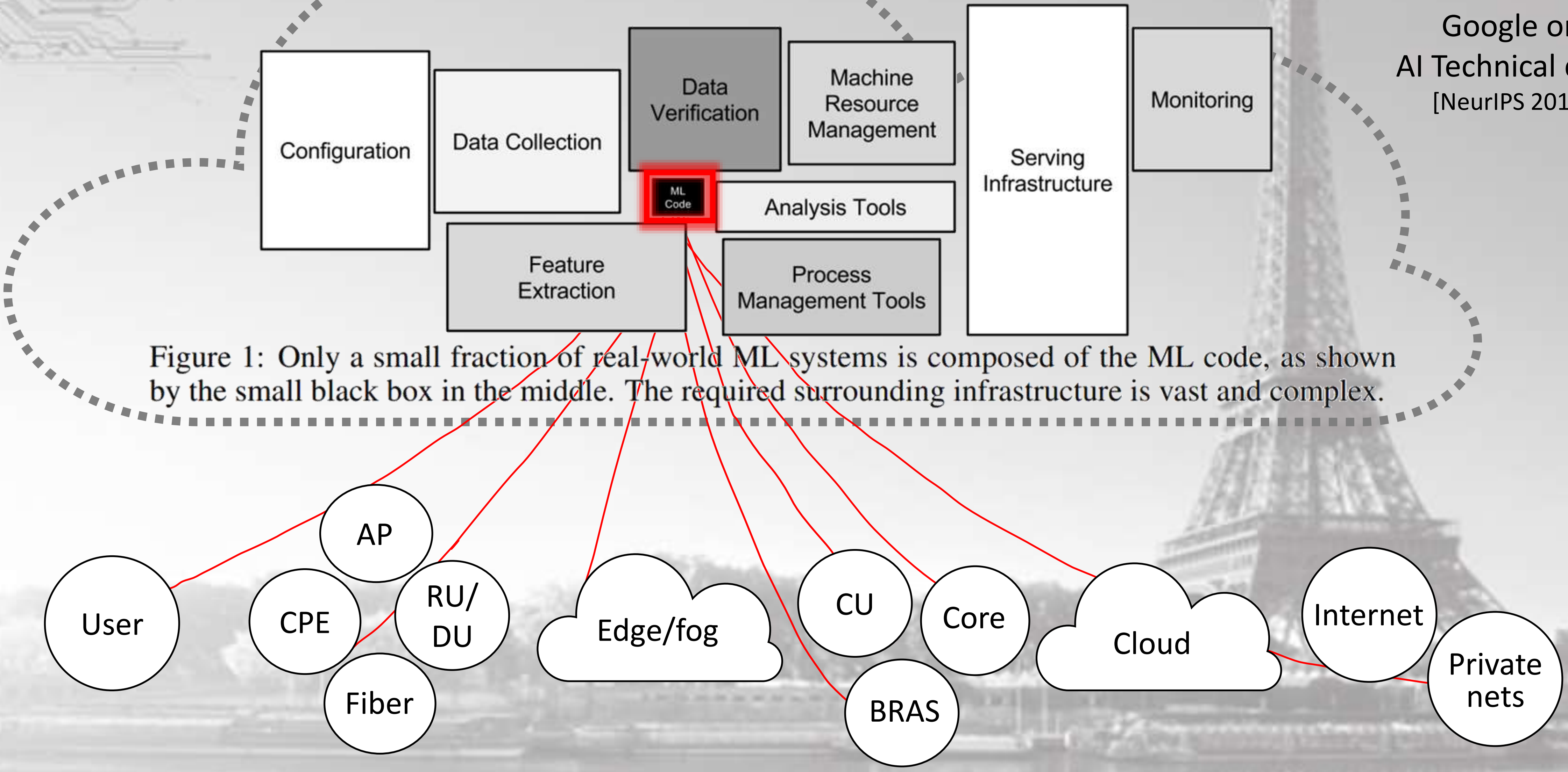
"Tree are interpretable"

Activation & feature maps, gradient back-propagation: interpretable only for images !

Limits of AI-Assisted networking

Deployment complexity

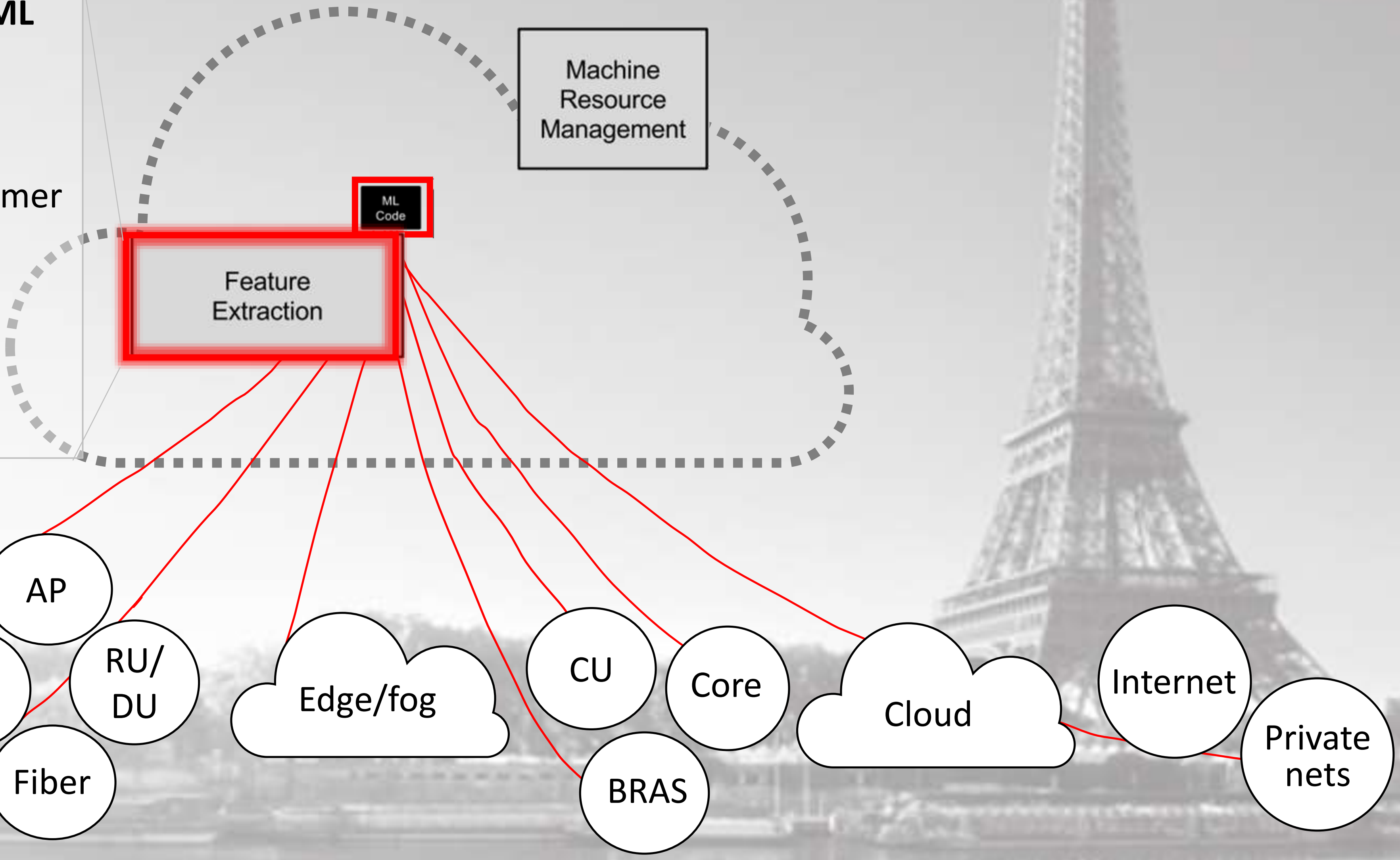
Google on
AI Technical debt
[NeurIPS 2015]



Limits of AI-Assisted networking

Heterogenity
complexity

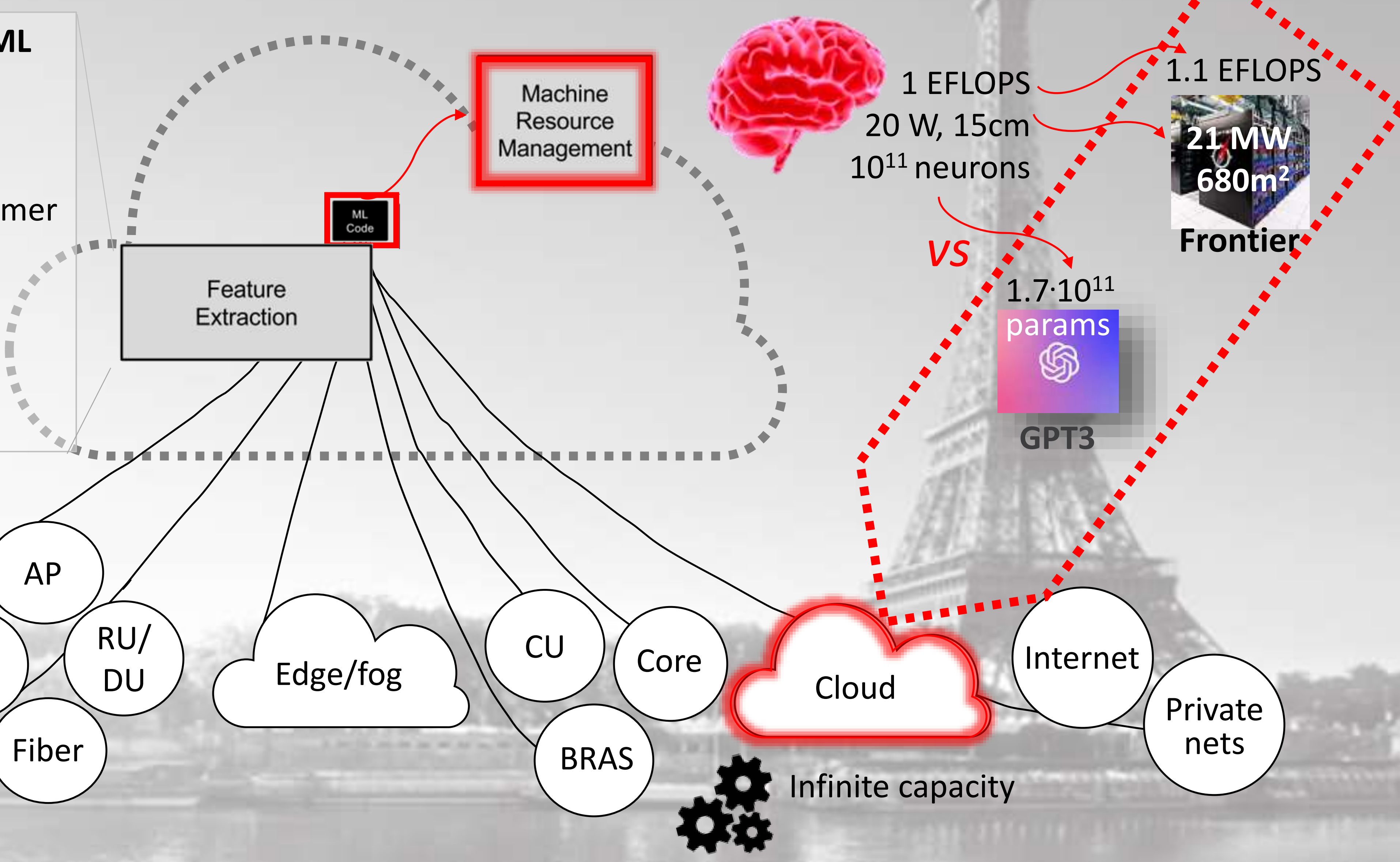
Raw data	appropriate ML
Tabular	RF/XGBoost
Image(Synth)	CNN(GAN)
Audio	CNN
Text	RNN/Transformer
Timeseries	LSTM
Graphs	GCN
Cyberphysical	multimodal
Networking	multimodal



Limits of AI-Assisted networking

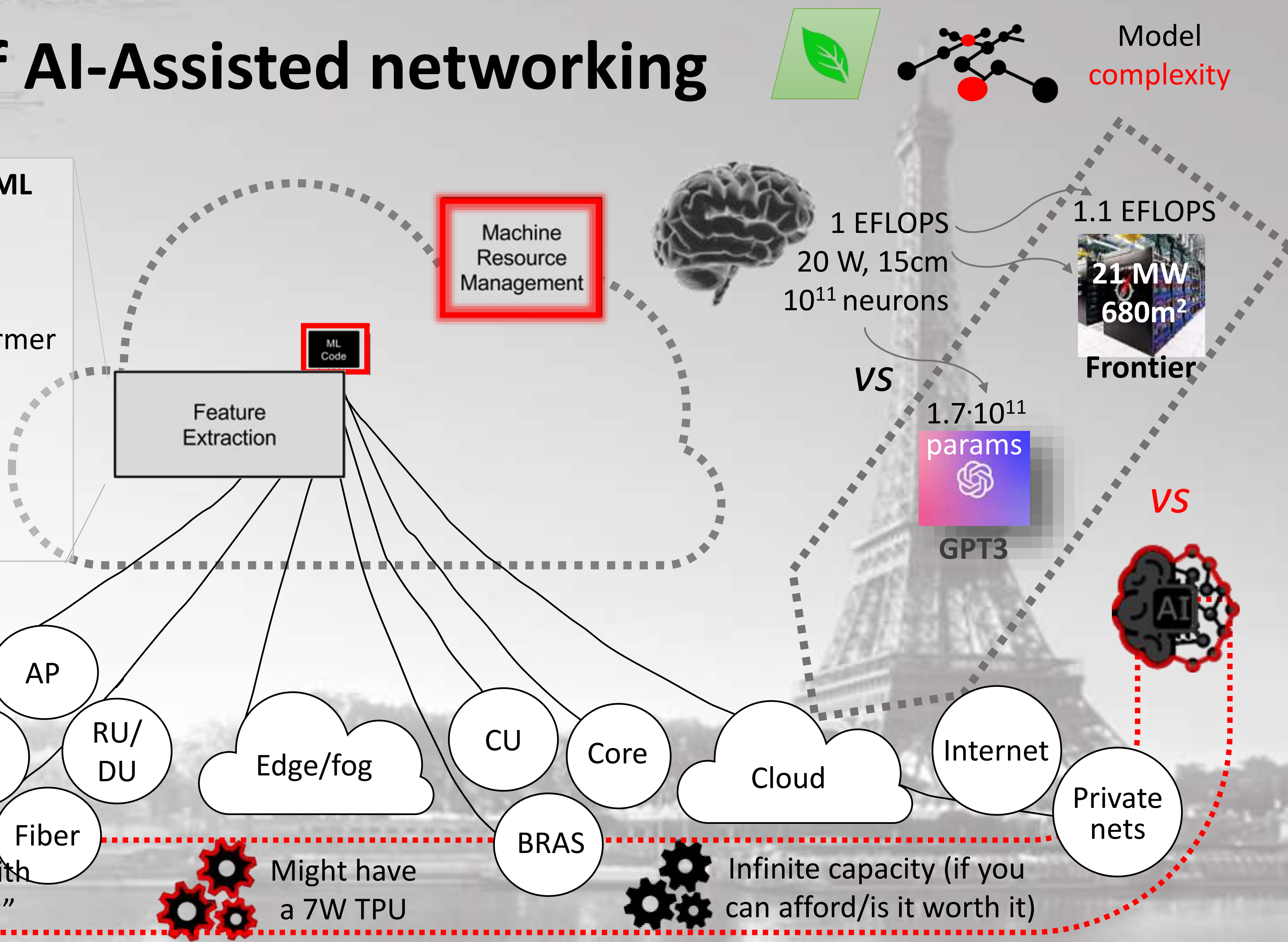
Model complexity

Raw data	**appropriate ML**
Tabular	RF/XGBoost
Image(Synth)	CNN(GAN)
Audio	CNN
Text	RNN/Transformer
Timeseries	LSTM
Graphs	GCN
Cyberphysical	multimodal
Networking	multimodal



Limits of AI-Assisted networking

Raw data	appropriate ML
Tabular	*RF/XGBoost*
Image	GAN
Audio	CNN
Text	RNN/Transformer
Timeseries	LSTM/GAN
Graphs	GCN
Cyberphysical	multimodal
Networking	multimodal



BACK TO THE FUTURE

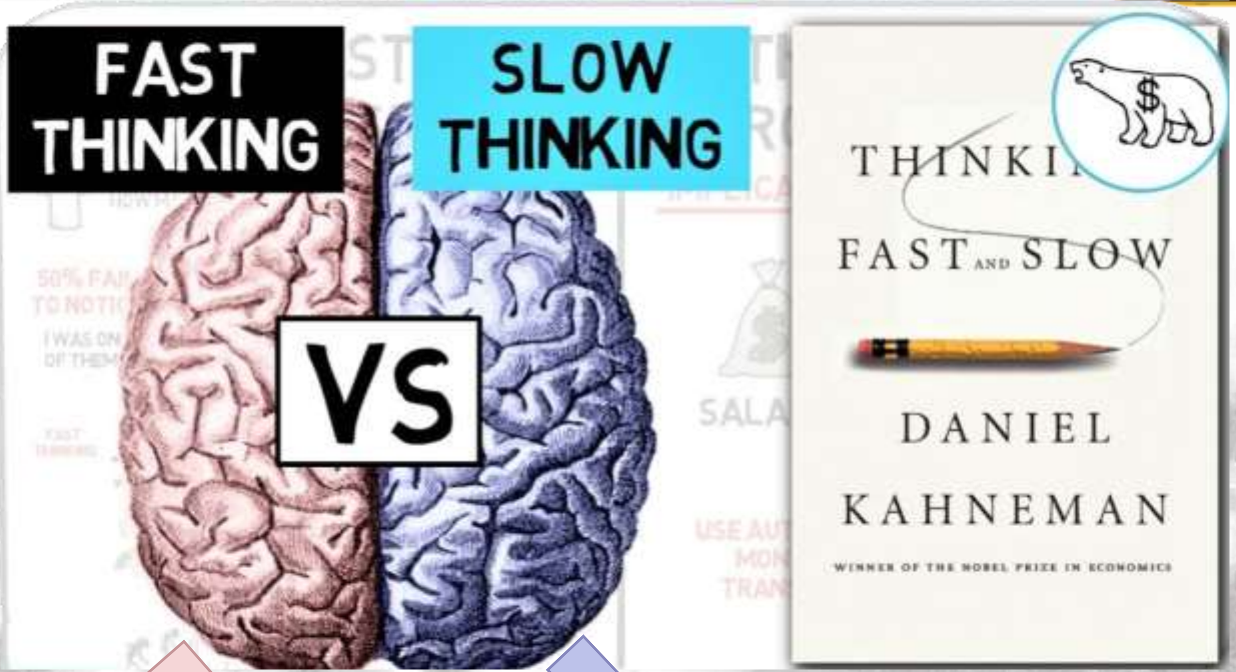
How 3

Principles of AI-Native networking

AI algorithms & compute

Highly inappropriate analogy
Do not interpret as first-degree

Explains irrational bias in human decision related to economics



FAST THINKING

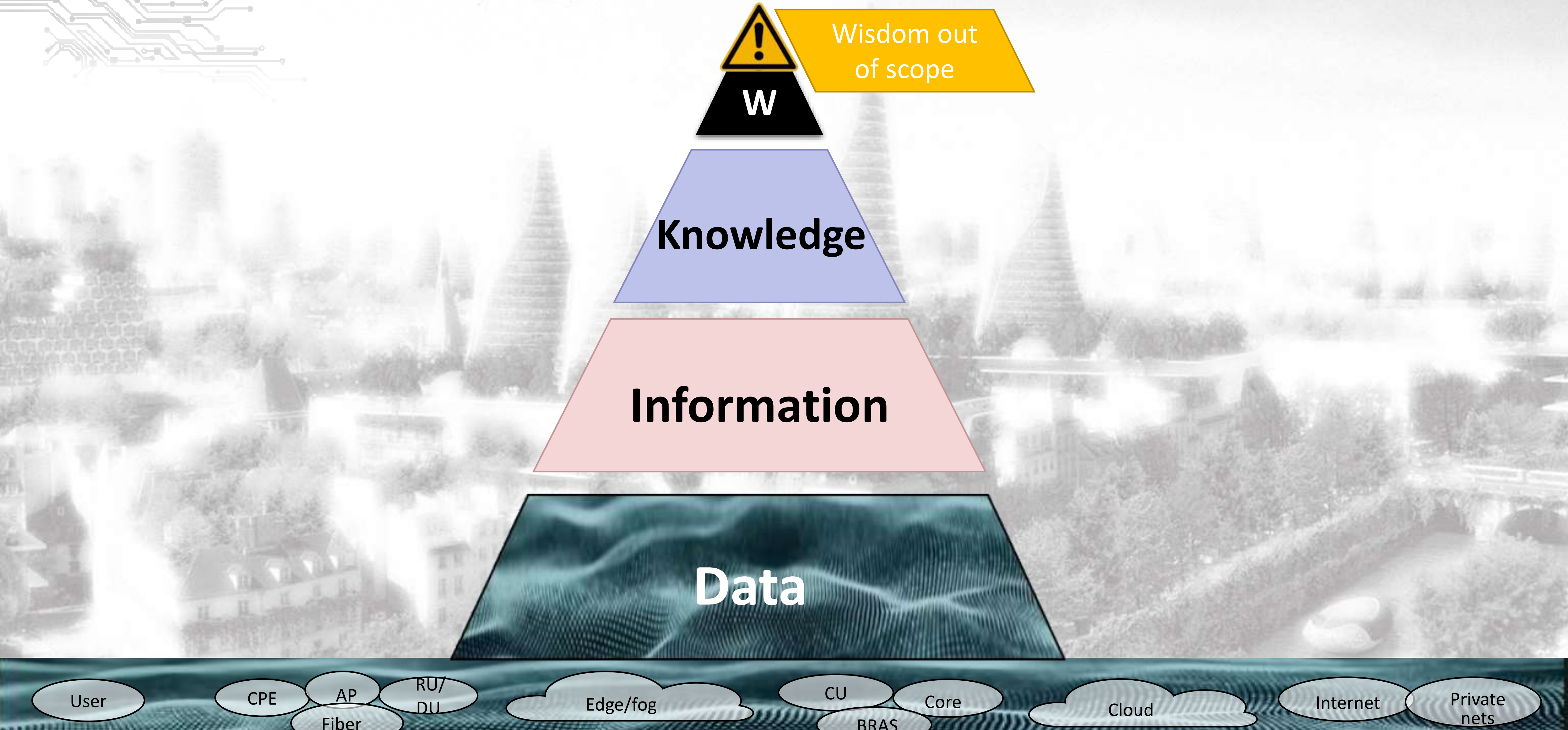
Correct on simple repetitive tasks.
Low cost but prone to bias and errors.

Advanced capabilities required for difficult tasks, or missing information.
Significant cognitive effort.

SLOW THINKING

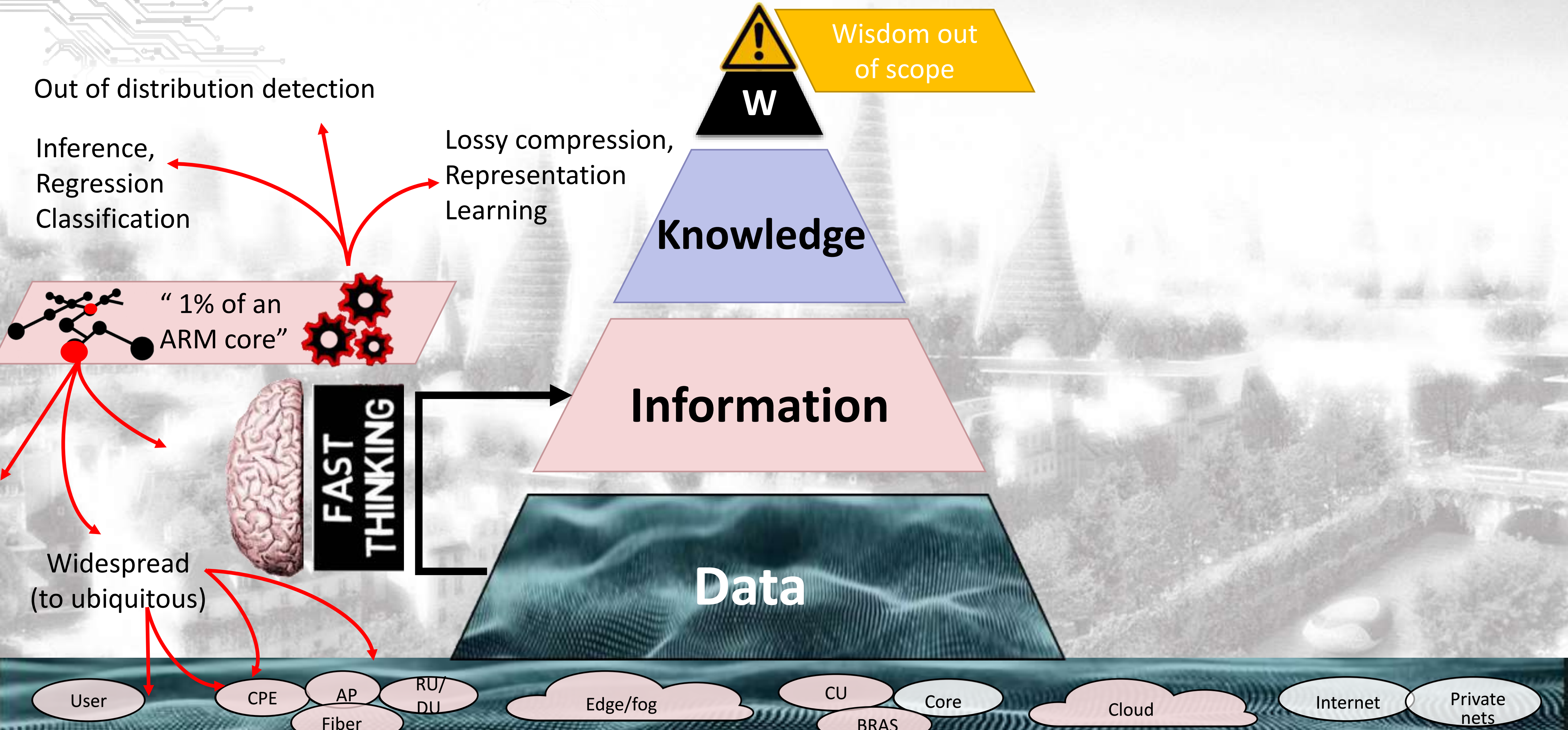
Principles of AI-Native networking

DIKW pyramid



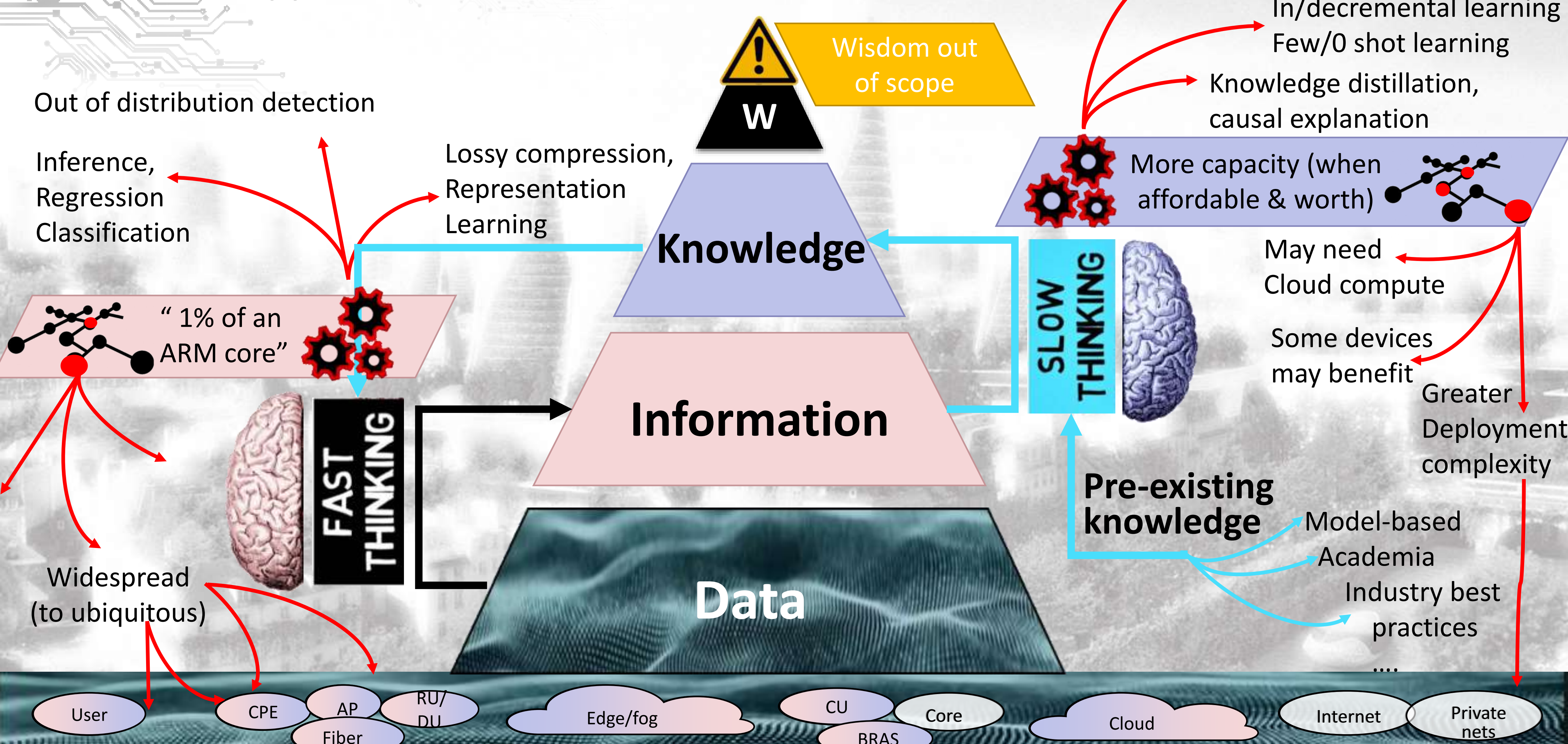
Principles of AI-Native networking

DIKW pyramid



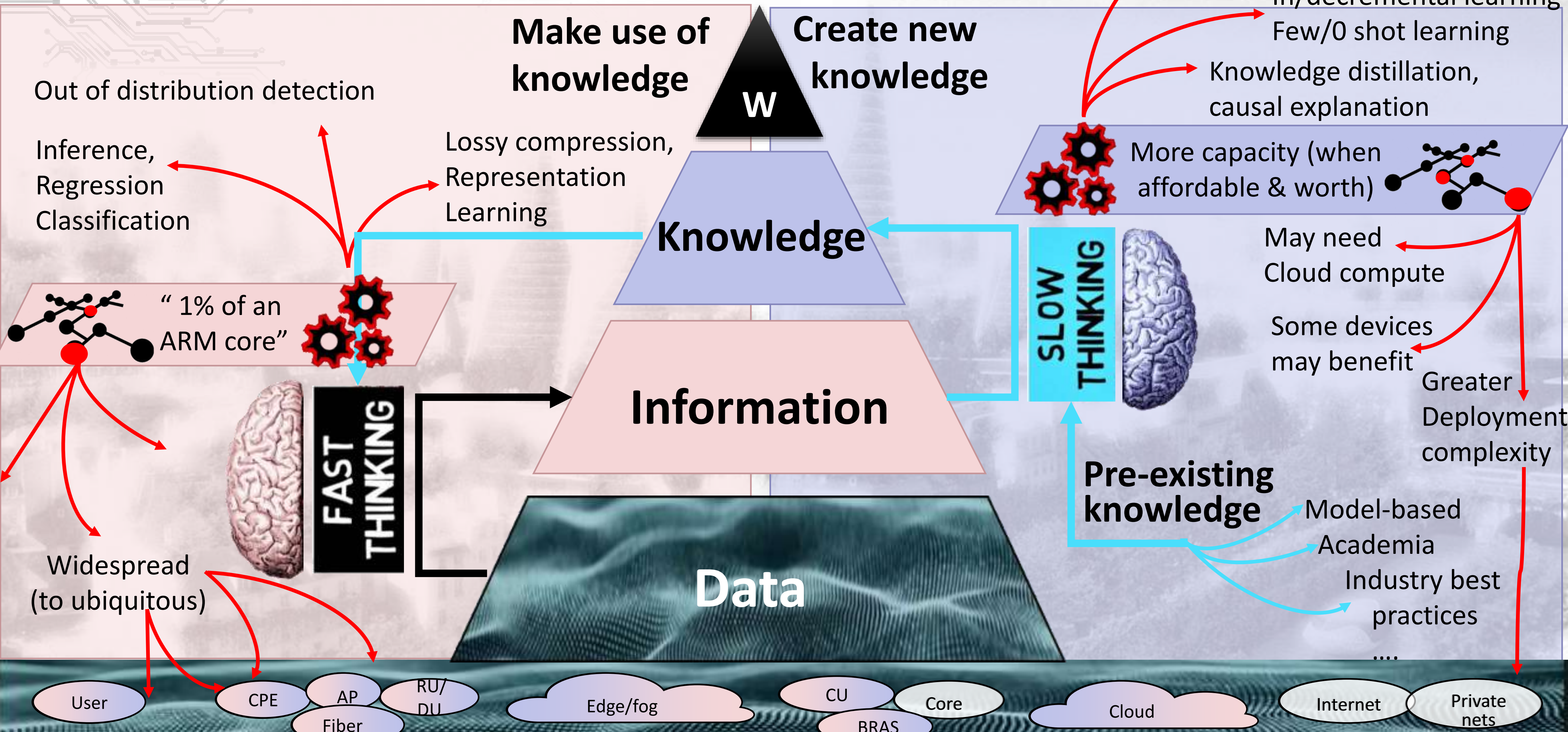
Principles of AI-Native networking

DIKW pyramid

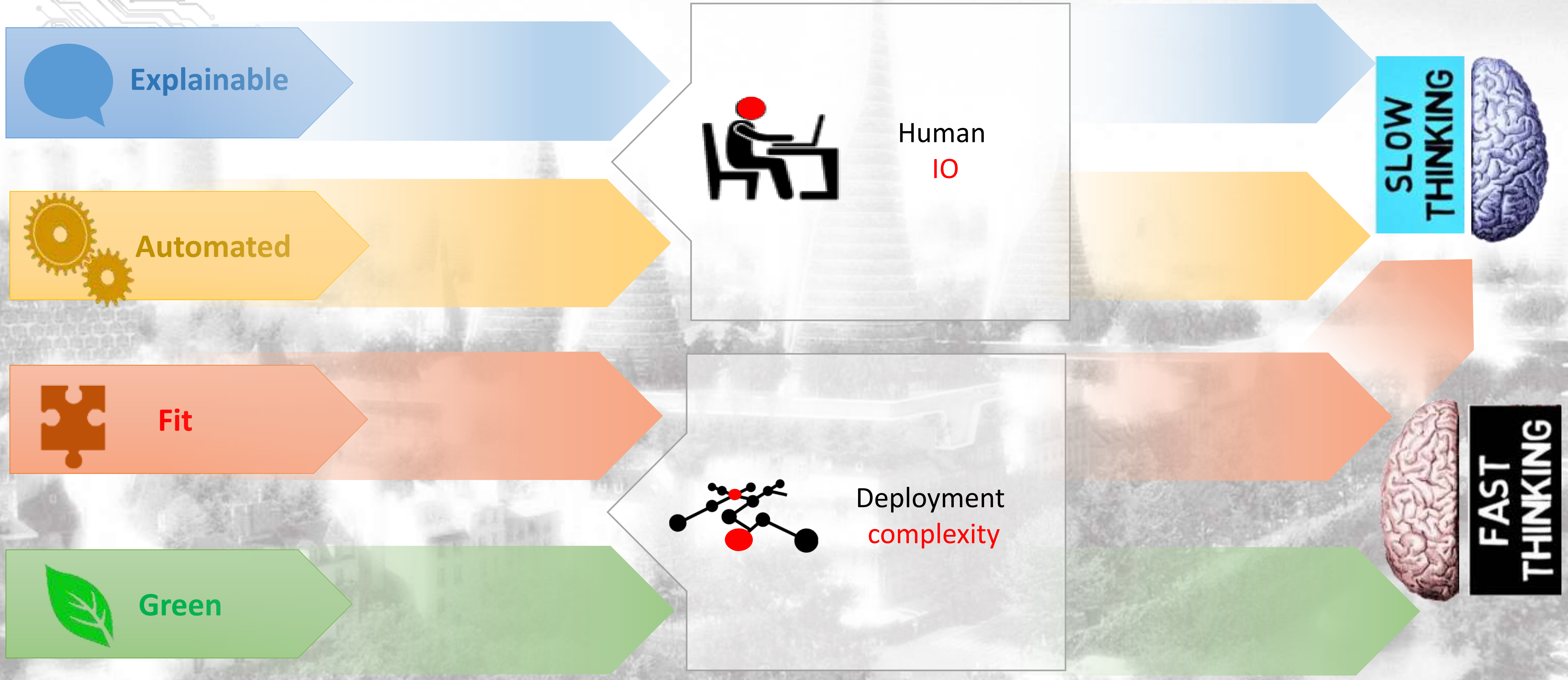


Principles of AI-Native networking

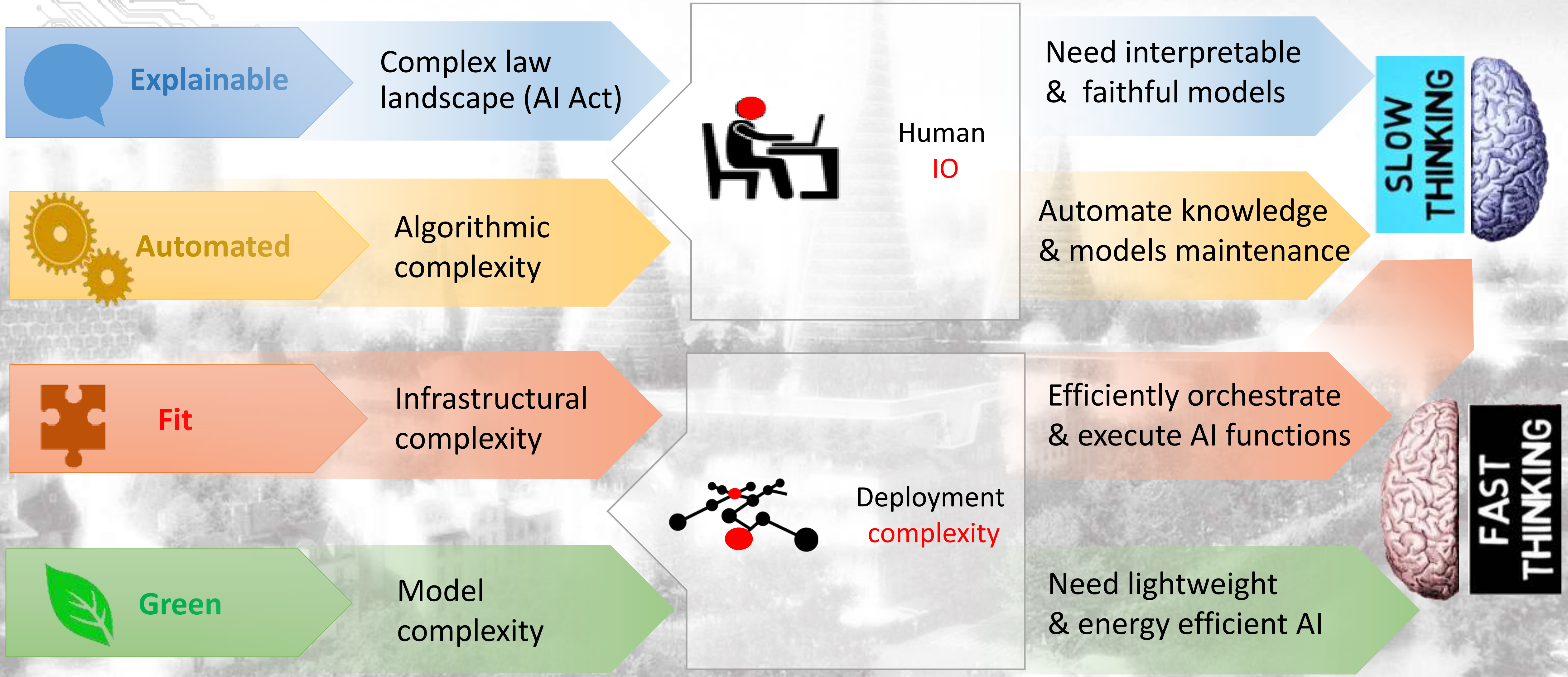
DIKW pyramid



Ingredients of AI-Native networking



Ingredients of AI-Native networking



Native

Explainable

Automated

Fit

Green

❑ Artificial Intelligence Act

New EU legislation ~GDPR extended to AI processing, primordial for *biometric* data (eg. Facial recognition) to avoid *bias* (eg. Racial discrimination) or privacy leaks

Complex law landscape (AI Act)

❑ Network AI issues/risks

- ***Bias***: ensure effective transferability of AI models
- ***Accountability***: business-policy explanation of AI decision
- ***Compliance***: legal aspects of accountability
- ***Interpretability***: AI decisions inherently less interpretable than human-made heuristics
- ***Verifiability:*** what can be proven, can be more easily trusted

❑ Need to explain models outputs (XAI)

- Explicit quantification of *confidence* in the model output
- *Step-by-step, multi-level* explanation
- *Faithfulness, i.e.,* explaination of actual model decision (vs surrogate)
- Explicit warnings about input *data quality*

❑ Cost of explainability

- May tradeoff with accuracy loss, or with increased complexity
- As for security, you need to budget the risk of *not having* faitfhful explainability

Native

❑ Artificial Intelligence Act

New EU legislation ~GDPR extended to AI processing, primordial for *biometric* data (eg. Facial recognition) to avoid *bias* (eg. Racial discrimination) or privacy leaks

Explainable

Automated

Fit

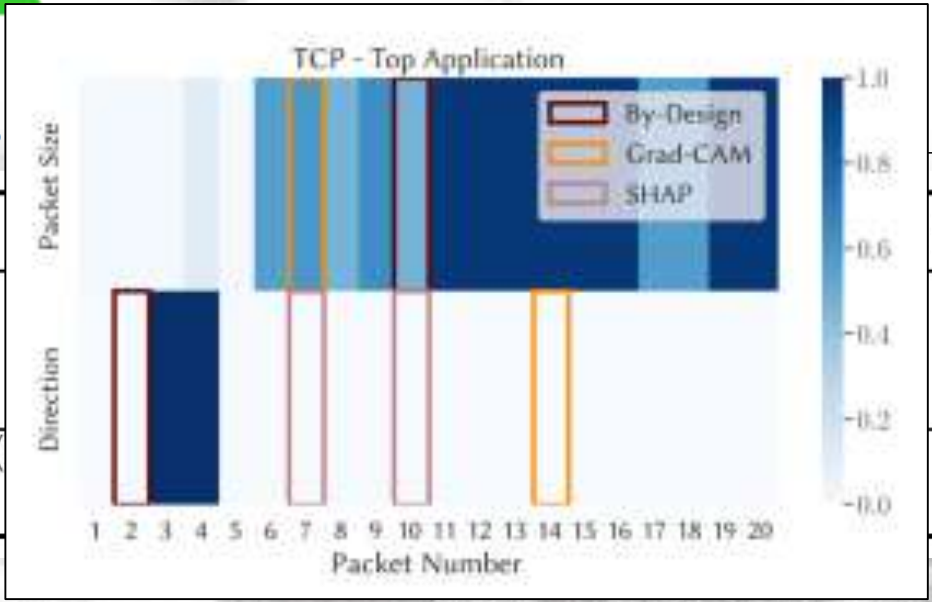
Green

Complex law landscape (AI Act)

❑Cost of explainability ?

Metric	ProtoPNet
Accuracy (%)	81.6
Inference GPU (μs/sample)	5.5
Inference CPU (μs/sample)	169.7
Number of Trainable Params (	201
FLOPs - Multiply-Adds (M)	

2 Prototypes /class



State of the art faithful XAI model

(Of course we have something better but it's not public :)

Grad-CAM	SHAP	XAI Correctness	ProtoPNet
8.7	6.1	Top 2 Accuracy (%)	
39.9	27.8	Top 10 Accuracy (%)	

100% By design

Native

Explainable

Automated

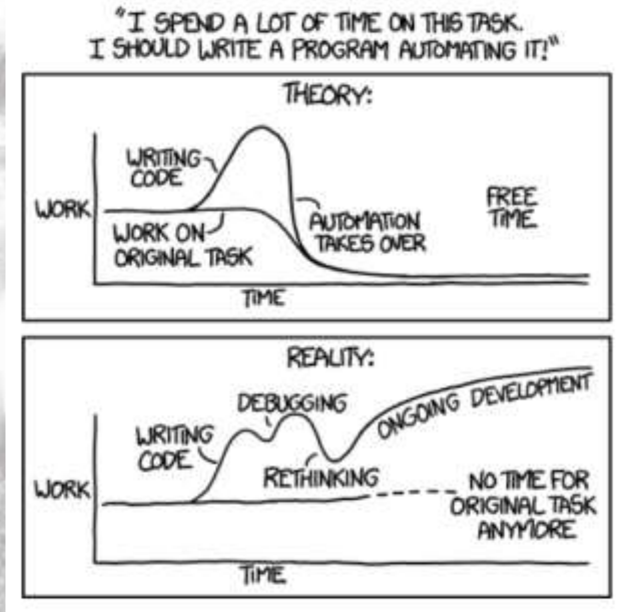
Fit

Green

Top-5 bottlenecks holding back AI adoption ?

https://www.oreilly.com/radar

Taming algorithmic complexity



https://xkcd.com/1319/

- **Compensate lack of skills**
  - Network Architecture Search (autoML/NAS) though inherently non-
  - Class incremental learning (CIL)
  - Automated hyperparameter selection for anomaly detection (autoAD/metaOD)
- **Compensate lack of data and label**
  - Few-shot learning (FSL)
  - Self-supervised learning (SSL)

Native

Explainable

Automated

Fit

Green

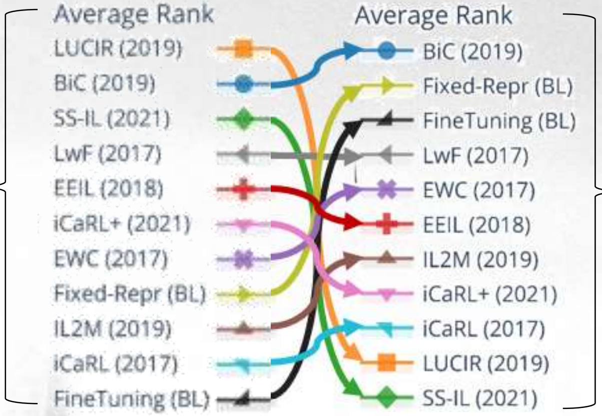
Supervised learning

Class Incremental Learning (CIL)

- ❑ Rank the state of the art
- ❑ Same dataset, difference in adding many classes once vs few classes several times

Image-style scenario *Network-like problem*

Rank



Totally different ranking

Taming algorithmic complexity

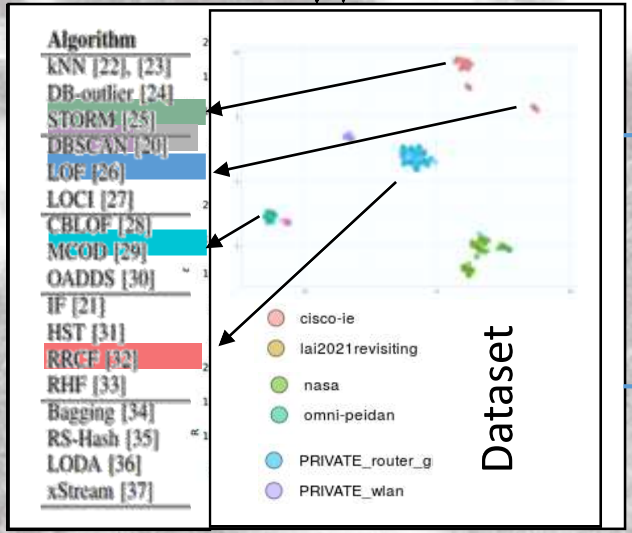
Unsupervised Learning

- ✓ No training
- ✓ Generalize better
- × Lots of algorithms
- × Hyperparametrization time-consuming even for AI experts

Unsupervised Meta-learning

Multi-variate KPI time-series

Computational budget



Auto Anomaly Detection (AutoAD)

=

Expert-level algorithm selection / ensembling

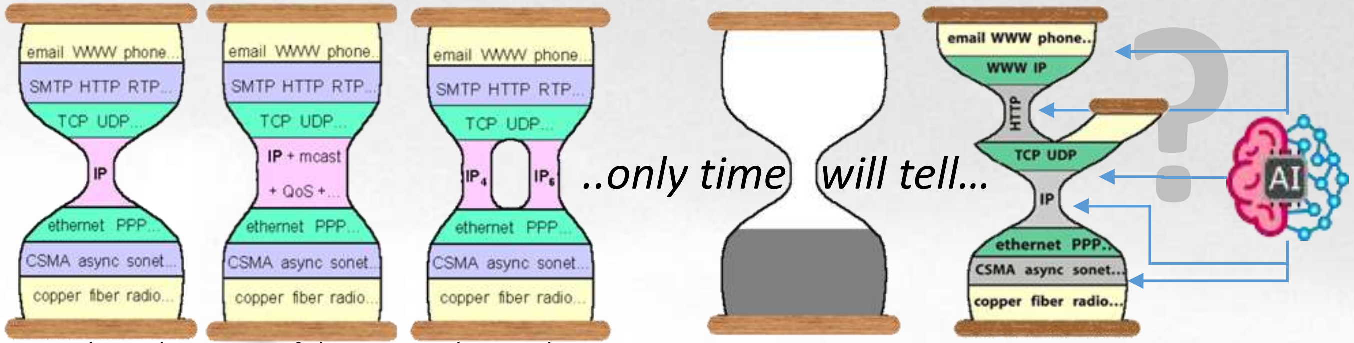
+

Hyper-parameter auto-tuning

Native

AI latest kids on the Networking block

IP Network originally designed for connectivity (neither for QoS, mobility, nor security… nor AI!)



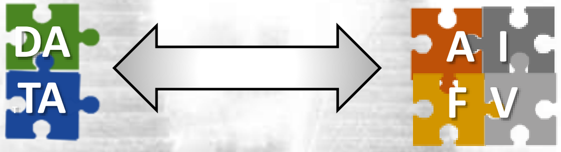
Watching the Waist of the Protocol Hourglass
Steve Deering @IETF51 (2001)

Evolution of the IP Model
Dave Thaler @IETF73 (2009)

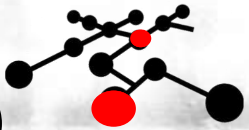
Explainable

Automated

Network data representation (& governance)

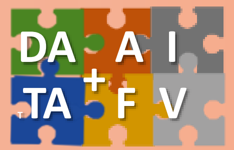


Networked AI functions (& architecture)

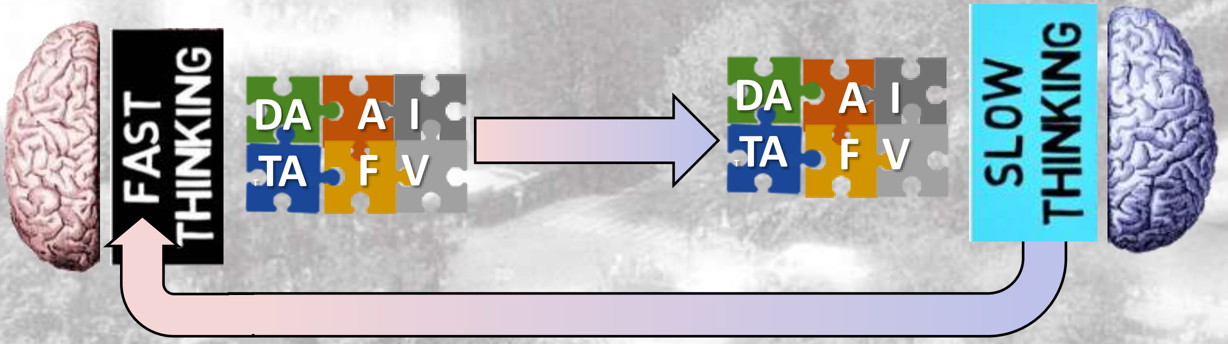


Fit

Taming infrastructural complexity



Green

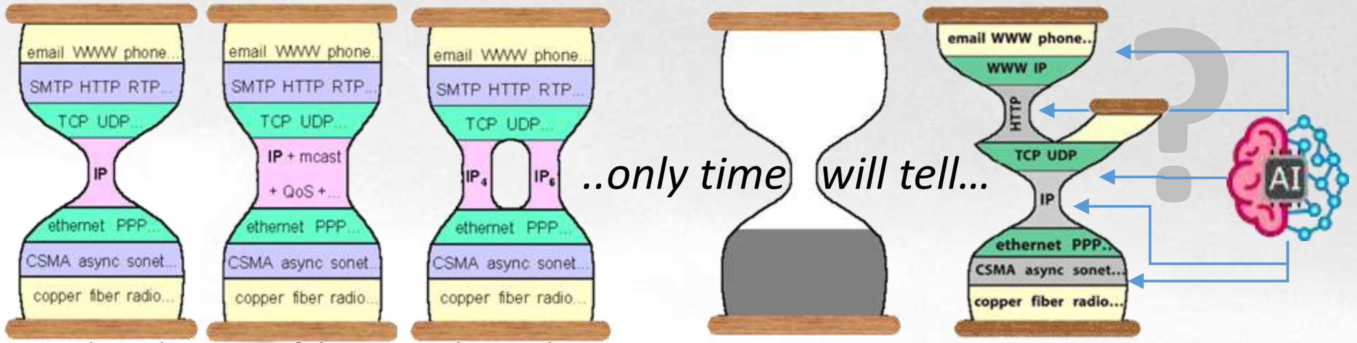


[CoNEXT-NNI'22]

Native

AI latest kids on the Networking block

IP Network originally designed for connectivity (neither for QoS, mobility, nor security… nor AI!)



Watching the Waist of the Protocol Hourglass
Steve Deering @IETF51 (2001)

Evolution of the IP Model
Dave Thaler @IETF73 (2009)

Explainable

Automated

Fit

Green

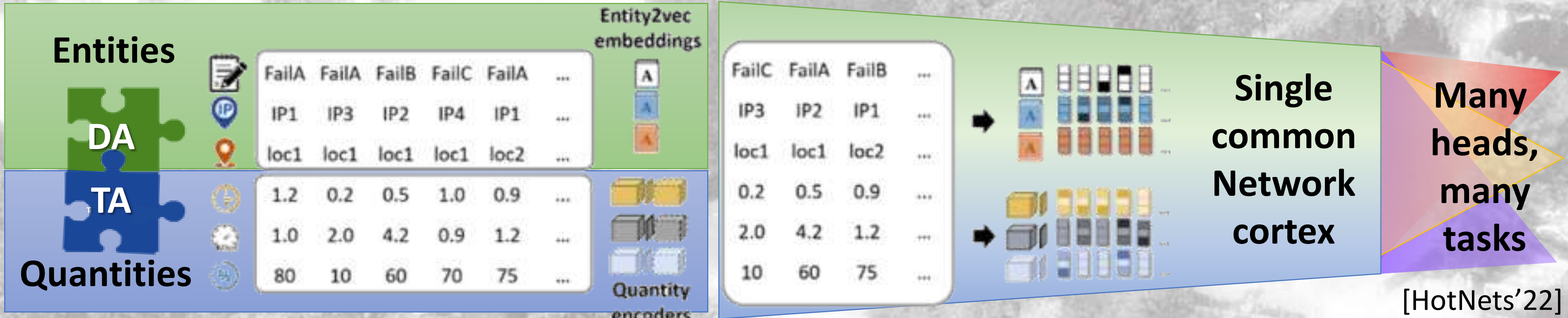
❑ Network data representation & governance

DA TA

- Universal network data representation (multimodal)
- Fit for AI processing (common network cortex, many AI tasks)
- Fit for AI Act & GDPR (eg. ACL, data boundaries, …)

Taming infrastructural complexity

DA A I
TA + F V

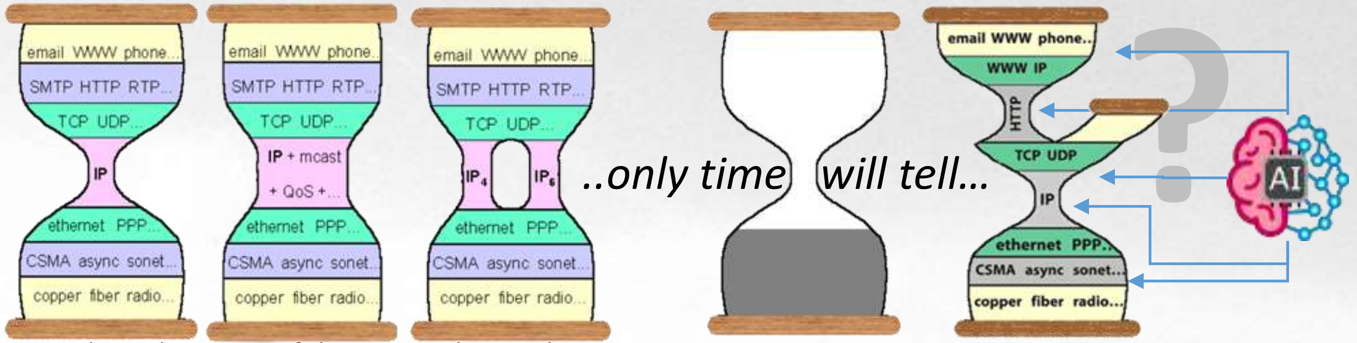


[HotNets'22]

AI Native

AI latest kids on the Networking block

IP Network originally designed for connectivity (neither for QoS, mobility, nor security… nor AI!)



Watching the Waist of the Protocol Hourglass
Steve Deering @IETF51 (2001)

Evolution of the IP Model
Dave Thaler @IETF73 (2009)

Explainable

Automated

Fit

Green

❑ Network data representation & governance

DATA

- Universal network data representation (multimodal)
- Fit for AI processing (common network cortex, many AI tasks)
- Fit for AI Act & GDPR (eg. ACL, data boundaries, …)

Taming infrastructural complexity

DATA + AIFV

❑ Network AI functions & architecture

AIFV

- Cloud-native architecture not enough, Edge/fog/serverless neither. P4/SDN at the other extreme. Reminiscense of Tennenhouse's Active networking (1996) ?
- Which network functions are best fit to be replaced/auto-tuned by AI ? How to systematically compose, execute AIFV ?

Native

Explainable

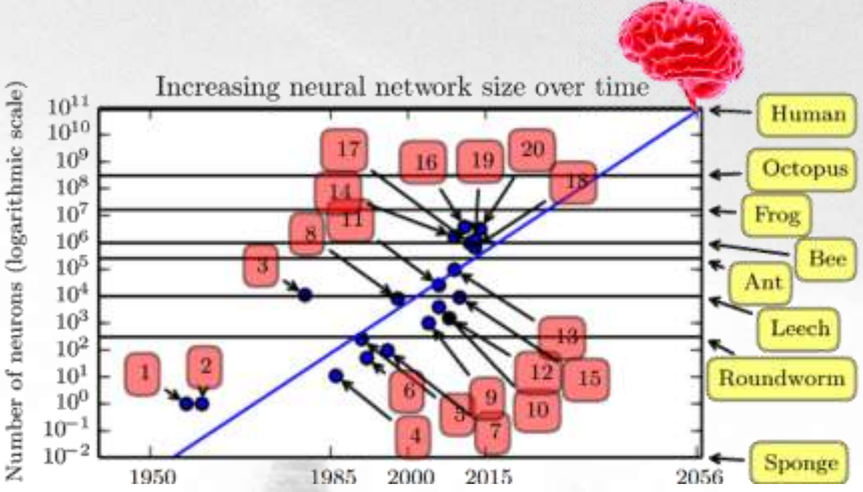
Automated

Fit

Green

AI models growth exceeds Moore law

Deep learning, MIT Press
https://deeplearningbook.org



Model size in 2020 exceeds 2015 forecast by *10,000x*!

Human brain scale reached in 2020 (30+ years earlied than expected)

❑ Need energy efficient AI models

- Raise awareness of computational complexity
- Set explicit “accuracy/joule” targets for certification (~“km/liter” for cars,or A-D energy labels)
- Applies to many AI aspects (training, inference, etc.)

❑ Small is beautiful (but not *too* small)

- Huge models race is for NLP/CVPR (like the quantum qubit race)
- AI researchers produce huge models, system researchers use tiny ones

Taming (unnecessary) model complexity

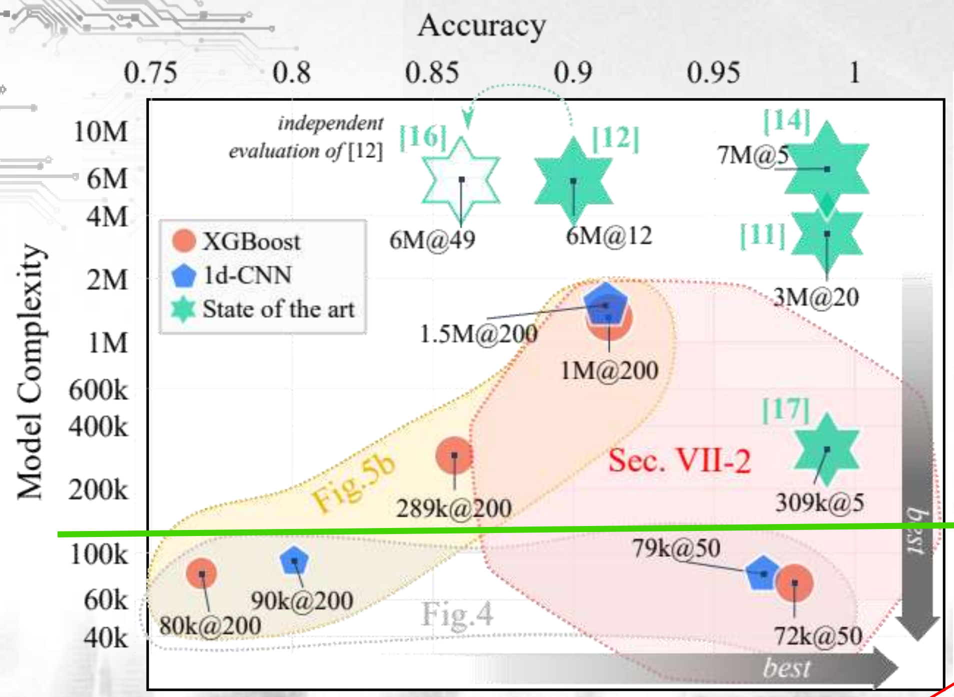
Native

Explainable

Automated

Fit

Green



AI researchers: Unnecessarily big !

	Method	$W\,[\times 1k]$	K	$W/K\,[\times 1k]$	Acc. [%]
State of the art	[14]	6,640	5	1,328	99
	[17]	309	5	61	99
	[12]	5,800	12	483	90
	[11]	3,270	20	163	99
	[16]	5,800	49	119	86
	1d-CNN	79	50	1.44	96
		90	200	0.450	81
		1,479	200	7.5	91
	XGBoost	72	50	1.44	96
		80	200	0.4	76
		289	200	1.44	85
		1,307	200	5	91

$O(10^6)$

~100k-large overall
~5k weights/class

[TNSM’21b] [SEC’21]
AI-viewpoint + system

O(10)

ASIC: 3 layers, 100 neurons (overall!)
SmartNIC: 50 neurons (overall!!)

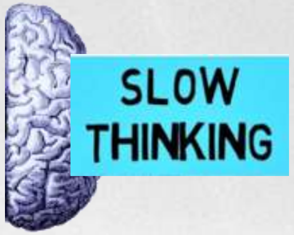
System researchers: too small !

❑ Small is beautiful (but not *too* small)

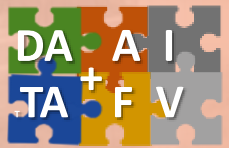
- Huge models race is for NLP/CVPR (like the quantum qubit race)
- AI researchers produce huge models, system researchers use tiny ones

Taming (unnecessary) model complexity

Native network intelligence, FAST THINKING & SLOW THINKING



Explainable	Complex law landscape (AI Act)	Need interpretable & faithful models
Automated	Algorithmic complexity	Tooling to automate ML model maintenance
Fit	Infrastructural complexity	Efficiently orchestrate & execute AI functions
Green	Model complexity	Need lightweight & energy efficient AI



Recent stuff

Internet: https://nonsns.github.io/
Intranet: https://frc-datacom.rnd-gitlab-eu-c.huawei.com/ai4net/

2022 20

- [CoNEXT-NNI-22a] Boffa, Matteo and Vassio, Luca and Drago, Idilio and Mellia, Marco and Milan, Giulia and Houidi, Zied Ben and Rossi, Dario, *"On Using Pretext Tasks to Learn Representations from Network Logs" ACM CoNext workshop on Native Network Intelligence (NNI)* dec. 2022, Conference Abstract Bibtex
- [CoNEXT-NNI-22b] Rossi, Dario and Liang, Zhang, *"Native Network Intelligence, Fast and Slow" ACM CoNext workshop on Native Network Intelligence (NNI)* dec. 2022, Conference Abstract Bibtex
- [CoNEXT-GNN-22] Fernandes, Danilo Marinho and Krolikowski, Jonatan and Houidi, Zied Ben and Chen, Fuxing and Rossi, Dario, *"Cross-network transferable neural models for WLAN interference estimation" ACM CoNext workshop on Graph Neural Networks (GNN)* dec. 2022, Conference Abstract Bibtex
- [AICCSA-22] Nesic, Stefan and Putina, Andrian and Bahri, Maroua and Huet, Alexis and Navarro, Jose Manuel and Rossi, Dario and Sozio, Mauro, *"StreamRHF: Tree-based unsupervised anomaly detection for data streams" 19th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2022)* dec. 2022, Conference Abstract Bibtex
- [HotNets-22] Houidi, Zied Ben and Azorin, Raphael and Gallo, Massimo and Finamore, Alessandro and Rossi, Dario, *"Towards a systematic multi-modal representation learning for network data" ACM HotNets* nov. 2022, Conference Abstract Bibtex
- [TNSM-22] Rossi, Dario and Zhang, Liang, *"Landing AI on Networks: An equipment vendor viewpoint on Autonomous Driving Networks" In IEEE Transactions on Network and Service Management (TNSM), Vol. 19,* sep. 2022, DOI 10.1109/TNSM.2022.3169988 Journal Abstract Bibtex
- [ITC34] Navarro, Jose Manuel and Huet, Alexis and Rossi, Dario, *"Rare Yet Popular: Evidence and Implications from Anomaly Detection Datasets" ITC34* sep. 2022, Conference Abstract Bibtex
- [SIGMETRICS-PER-22] Roberts, James and Rossi, Dario, *"Size-Based Scheduling vs Fairness for Datacenter Flows: A Queuing Perspective" In ACM SIGMETRICS Perform. Eval. Rev., Vol. 50, No. 2,* sep. 2022, Journal Abstract Bibtex
- [PATENT-PCT/EP2022/075646] YANG, Lixuan and FINAMORE, Alessandro and CHEN, Fuxing and ROSSI, Dario, *"A device and method for network traffic classification"* , sep. 2022, Patent
- [KDD-22] Huet, Alexis and Navarro, Jose Manuel and Rossi, Dario, *"Local Evaluation of Time Series Anomaly Detection Algorithms" ACM SIGKDD Conference on Knowledge Discovery and Data mining (KDD)* aug. 2022, Conference Abstract Bibtex
- [SIGCOMM-CCR-22] Wang, Chao and Finamore, Alessandro and Yang, Lixuan and Fauvel, Kevin and Rossi, Dario, *"AppClassNet: A commercial-grade dataset for application identification research" In ACM SIGCOMM Computer Communication Review, Vol. 52,* jul. 2022, DOI https://doi.org/10.1145/3561954.3561958 Journal Abstract Bibtex
- [ComCom-22] Houidi, Zied Ben and Rossi, Dario, *"Neural language models for network configuration: Opportunities and reality check" In Elsevier Computer Communication, Vol. (to appear),* jul. 2022, Journal Abstract Bibtex
- [ArXiV-22-DRL] Iacoboaiea, Ovidiu and Krolikowski, Jonatan and Houidi, Zied Ben and Rossi, Dario, *"From Design to Deployment of Zero-touch Deep Reinforcement Learning WLANs"* jul. 2022, arXiv Tech.Rep. Abstract Bibtex
- [ICML-22] Franzese, Giulio and Rossi, Simone and Yang, Lixuan and Finamore, Alessandro and Rossi, Dario and Filippone, Maurizio and Michiardi, Pietro, *"How much diffusion time is enough?" ICML 2022 workshop on Continuous time methods for machine learning* jun. 2022, Conference Abstract Bibtex
- [INFOCOM-22] Finamore, Alessandro and Roberts, James and Gallo, Massimo and Rossi, Dario, *"Accelerating Deep Learning Classification with Error-controlled Approximate-key Caching" IEEE INFOCOM* may. 2022, Conference Abstract Bibtex
- [PATENT-PCT/EP2022/059292] FINAMORE, Alessandro and YANG, Lixuan and ROSSI, Dario, *"Method to address extreme class imbalance in AI based classifiers"* , apr. 2022, Patent
- [PATENT-PCT/EP2022/057757] NAVARRO, Jose Manuel and HUET, Alexis and ROSSI, Dario, *"Aggregation of Anomalies in a Network"* , mar. 2022, Patent

AI Assisted

"Landing AI on Networks: An equipment vendor viewpoint on Autonomous Driving Networks" I
n IEEE TNSM, Vol. 19, sep. 2022, DOI 10.1109/TNSM.2022.3169988

AI Native

"Native Network Intelligence, Fast and Slow" ACM CoNext workshop on Native Network Intelligence (NNI) dec. 2022

Thanks!
// || ??

Prof. Dario Rossi
dario.rossi@huawei.com

Director, Huawei AI4NET Lab, Network Products & Solutions
Director, DataCom Lab, Paris Research Center