AI Botany: Open RAN

H. Lehmann | T-LABS | 10.09.2024

GROUP **TECHNOLOGY**

Agenda

01 Phenomenological Modelling of AI Capabilities

02 Why & How

03 Scales

04 Overarching Problems

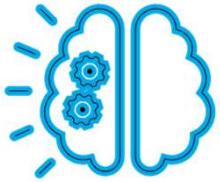
Phenomenological Modelling of AI Capabilities

Botany:

- describe appearances & phenomena
- cluster & relate
- no causal understanding

AI in (radio access) networks:

- Is it possible and sensible to develop an analogous approach?
- Lots of apparent reasons not to:
  - AI does not grow like plants…
  - the system which hosts the AI is not like nature (living, evolving, self-organizing)…
  - give up on all the exact math which is out there…
  - …



Why we yet try it: Some Motivation

In future networks, AI will be **a zoo of interacting distributed capabilities** with differing objectives, mechanics and governance.		The deterministic nature of these interactions is not accessible.
x/rApps are provided by a multitude of vendors who own the inherent (AI) algorithms. Hence, they are **opaque to the operator**.		Empirical description is the only way to model anything.
For the operator, AI in the network means a **techno-economical optimization problem** (where to put resources and how to equip them).		This is an aggregate problem which is decided on a coarse-grained level.
AI security in networks relies on **standardized and comprehensive approaches** (procedures and tests).		See, e.g., the state of the art in SW supply chain security: it's purely empirical.

How to ?

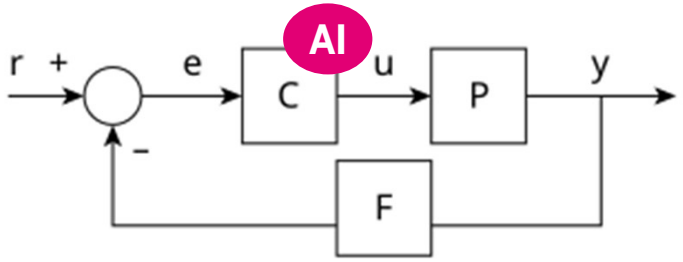
Identify description classes:

- **resource consumption descriptors**
- **system response descriptors**
  - borrow from physics:

$$x(t) = \int_{-\infty}^{t} dt' \, \chi(t - t') h(t') + \cdots .$$

AI

  - borrow from control theory:



Resource Consumption Modelling

- **computational weight**

  $\mathbf{cw} = \{cw_{training}, cw_{execution}\}$ with
  $cw_x = \{t_x, \text{compute load}_x\}$.

- **data sets – loci, volume and flows**

  $\mathbf{x}_{\text{data set}}^{(gen)}, \mathbf{x}_{\text{data set}}^{(stor)}$
  $\mathbf{V}_{\text{data set}}$ and $\boldsymbol{\nu}_{\text{data set}}$.

Summary description of computational effort and related data logistics per AI capability.

Be aware of some “botanical blurring” – i.e. the principal limits of phenomenology:

- algorithm variations may result in swifter convergence and a decreased $t_{training}$
- tuning of the training phase may bring down the required $compute\ load_{training}$.
- $t_{execution}$ will alter with configuration specifics
- algorithm advances may bring down the required $\mathbf{V}_{\text{data set}}$

System Response Modelling

- **induced latency**

$$\mathbf{t}_{\text{data flow}} = \max\{ t_{\text{data flow}}^{(1)}, t_{\text{data flow}}^{(2)} \ldots, t_{\text{data flow}}^{(n)}\}$$

This aggregate latency is accumulated over the consecutive architecture elements a data set item traverses. As data sets may be composed from different $\mathbf{x}_{\text{data set}}^{(\text{gen})}$, only the maximum is relevant.

“Botanical blurring”:

- stochastics of individual forwarding steps per network elements

System Response Modelling

- **action space and reach**

$$\{\mathbf{x}_{\text{action}}^{(i)}\} = \{\mathbf{x}_{\text{action direct}}^{(i)}, \mathbf{x}_{\text{action indirect}}^{(i)}\}$$

The subspace of the network affected by a specific capability *(i)*.

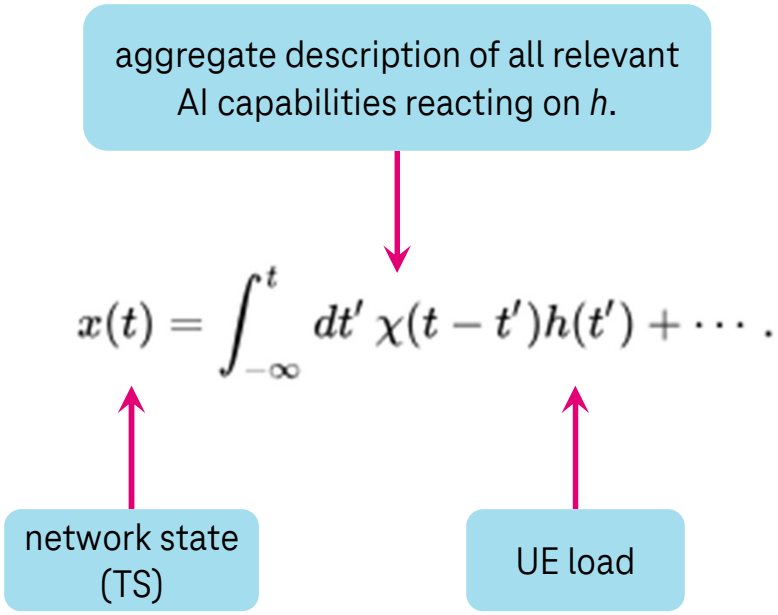
First, be clear on space concept:

- **disaggregation space** (relating to the ORAN functional architecture),
- **real space** which relates the network elements to geo-positions attributed with, e.g., territory characteristics or user number statistics, and
- **network space** (the mapping of one of the above on the other).

Very hard to quantify; range from $\{\mathbf{x}_{\text{action}}^{(i)}\} = \varnothing$ to the entire network.

Spatio-Temporal Scales as Ordering Principle

- **time**



Spatio-Temporal Scales as Ordering Principle

- **time**

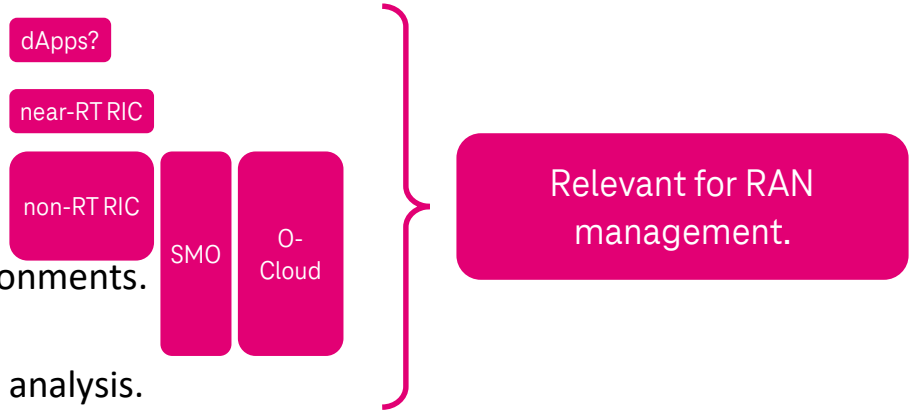
Time-ordering of the relevant AI capabilities in the network

$$x(t) = \int_{-\infty}^{t} dt' \, \chi(t - t') h(t') + \cdots .$$

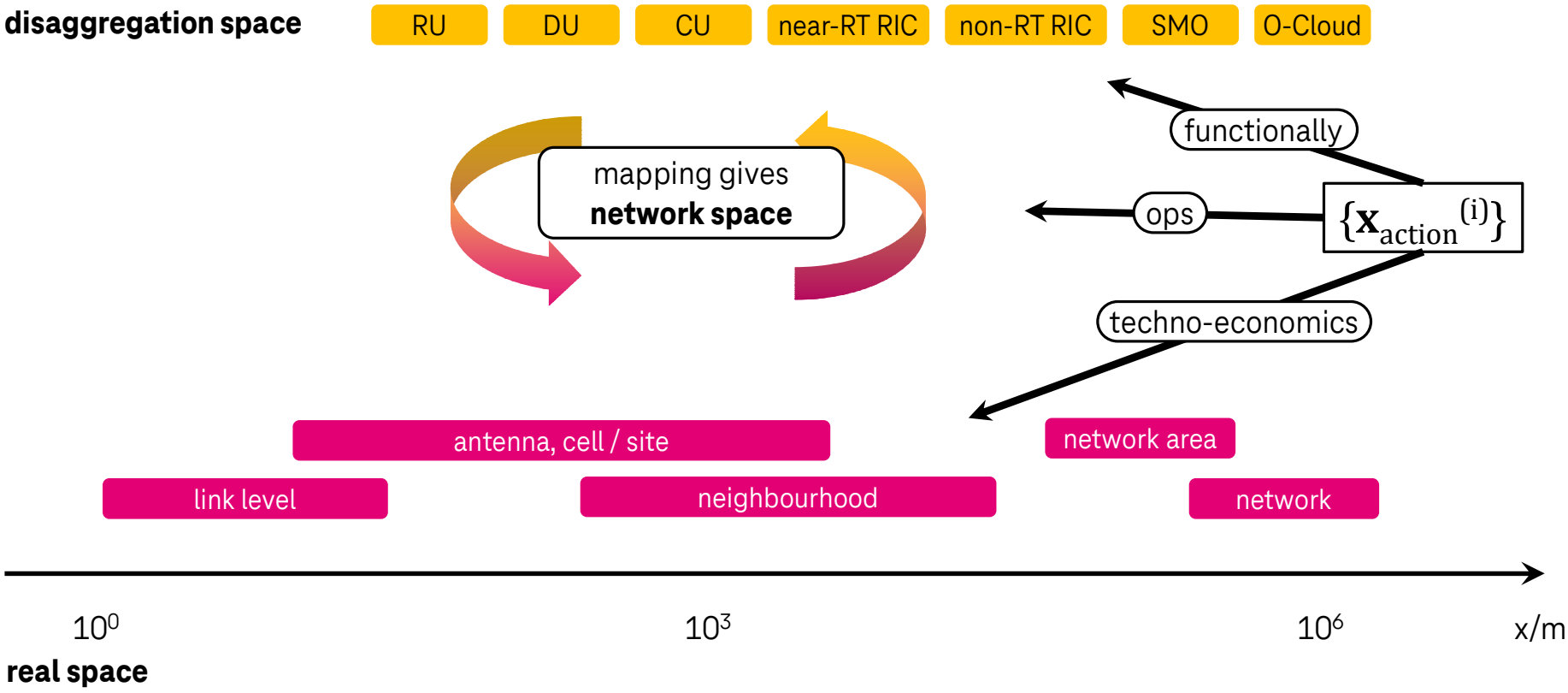
Separation of processes along physically, or conventionally given time scales drastically simplifies the quantification of the response function

Spatio-Temporal Scales as Ordering Principle

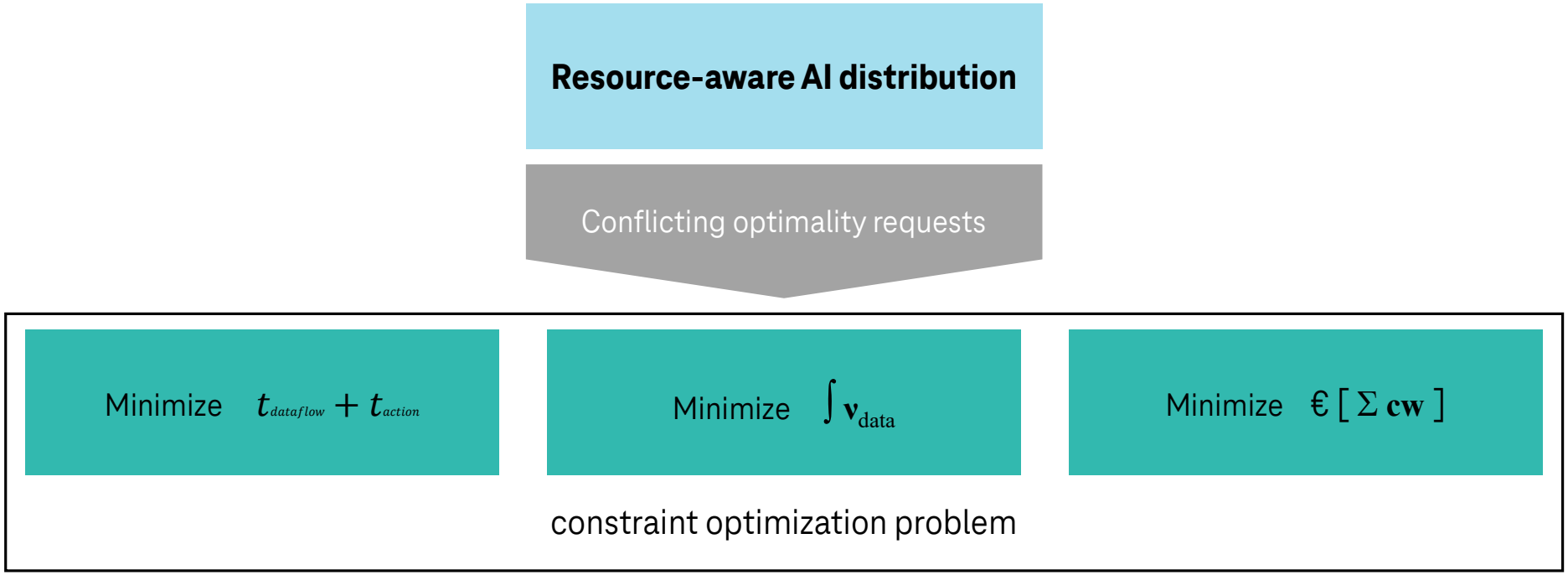
- **time**
  - Real time: 10 ms > *t*
  - Near-real time: 10 ms < *t* < 1 s
  - Non-real time: 1 s < *t*
  - Mobility scale: *t* ~ 1 hour,
    the typical scale for traffic variability in urban environments.
  - Norm week: *t* ~ 1 week,
    presenting the basis for (off-line) network planning analysis.
  - Seasonal alterations: *t* ~ several months,
    giving (in middle-European climate) the scale for, e.g. changes in the HVAC operation of central offices.
  - Budget year: *t* ~ a year,
    basis for budgetary decisions in the corporate context of the operator.
  - Lifecycle scale: *t* ~ a decade,
    the usual replacement cycle of telecommunication field elements



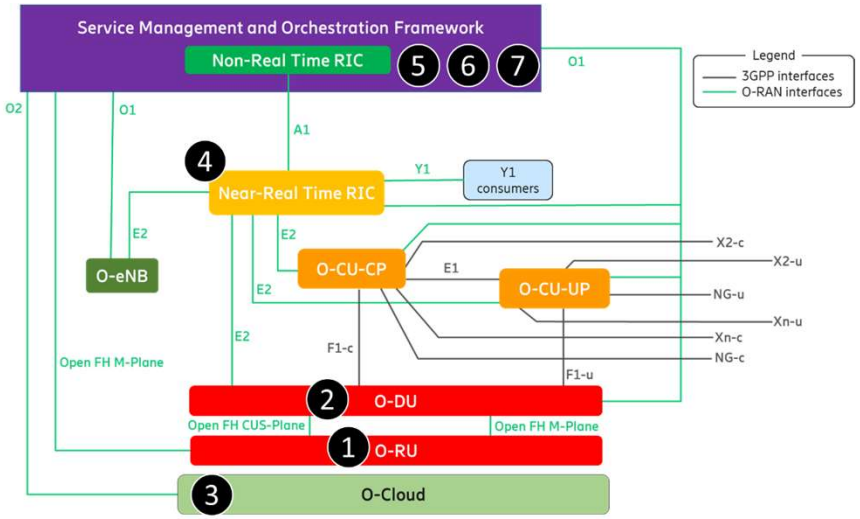
Spatio-Temporal Scales as Ordering Principle



Overarching Problems



Overarching Problems: Holistic Energy Management



- function approximation (energy consumption on load) ❶
- dApp ❷
- energy mgmt. per virtualization ❸

- **checking the formalism:**

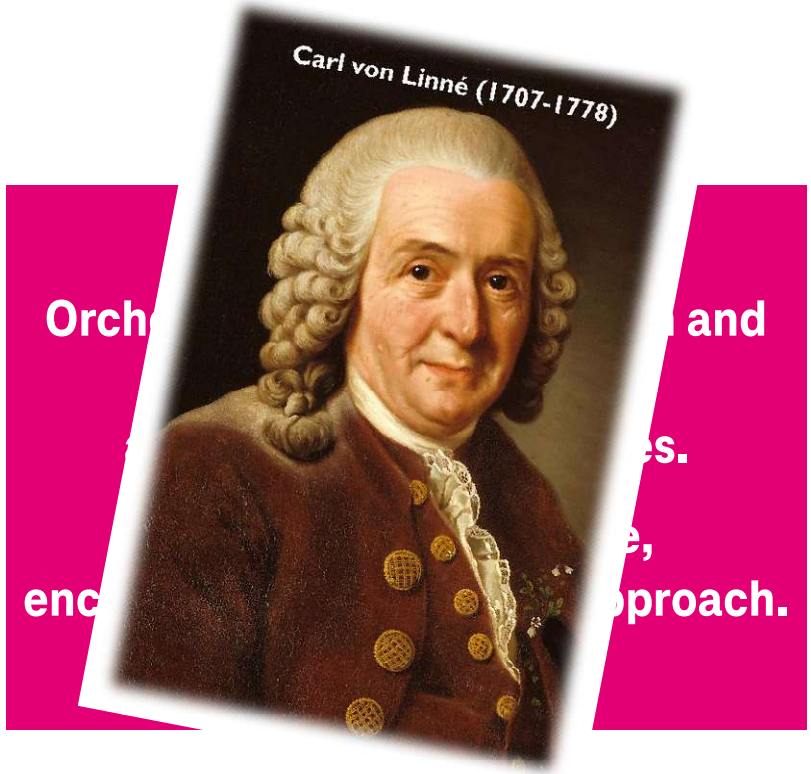
orch.

- $\mathbf{x}_{\text{data set}}^{(\text{gen})}$ as shown
- dedicated xApp (❹) with
  - $\{\mathbf{x}_{\text{action}}^{(i)}\}$ → O-RU (map on real space)
  - **cw** – moderate, $\mathbf{V}_{\text{data set}}$ – low, $\mathbf{t}_{\text{data flow}}$ – nearRT
- dedicated rApp (❺) (reinforcement learning) with
  - $\{\mathbf{x}_{\text{action}}^{(j)}\}$ → O-RU (map on real space)
  - **cw** – high, $\mathbf{V}_{\text{data set}}$ – mod, $\mathbf{t}_{\text{data flow}}$ – mobility scale
  - and auxiliary AI capability traffic forecast (❻)
  - and conflict detection / mitigation /resolution (❼)

Summing Up

AI in future networks is a heterogeneous ensemble.

Vendors think functions, operators think systems.



Thanks to:

Martin Stahn

Matthias Weh

Richard Weiler