The Synergy Between Quality of Experience and Deep Reinforcement Learning for Uncoordinated Multi-Agent Resource Allocation in Cognitive Radio Network

Andres Kwasinski

Rochester Institute of Technology

ak@mail.rit.edu

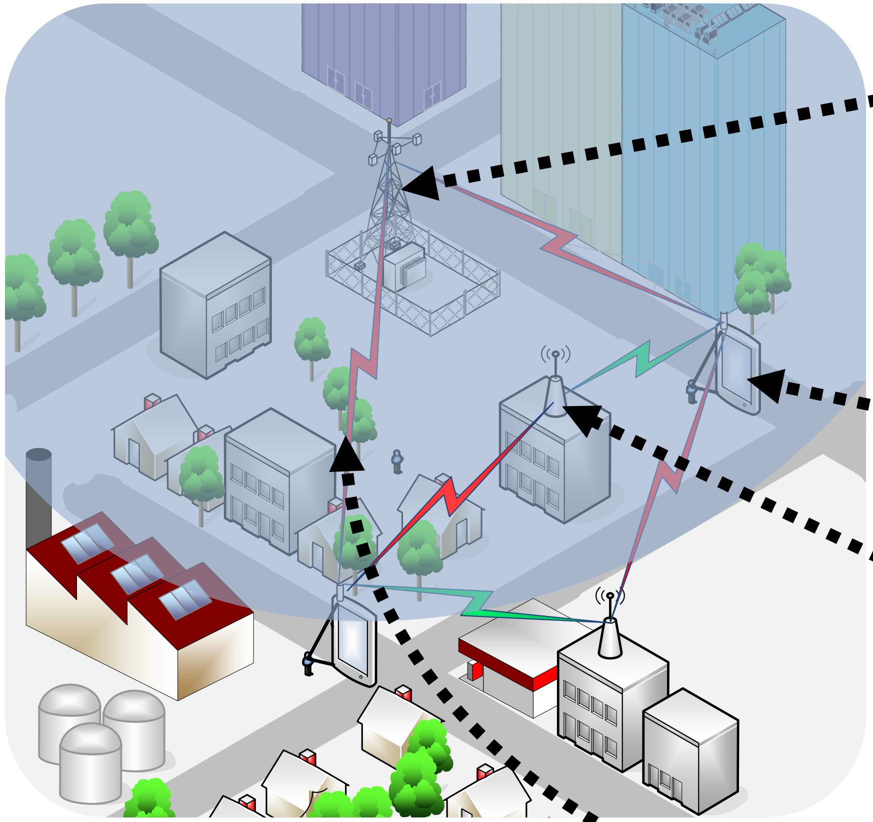
Outline

- Cognitive radio resource allocation in underlay dynamic spectrum access.
- Deep Q-learning (DQL) solution.
- Dissimilar traffic integration and performance measure → The Mean Opinion Score.
- Accelerating learning → Why using Mean Opinion Score is important.
- Conclusions.
- Bonus short talk about forming researchers.

Acknowledgment

Fatemeh Shah-Mohammadi

Setup: Underlay Dynamic Spectrum Access (DSA)



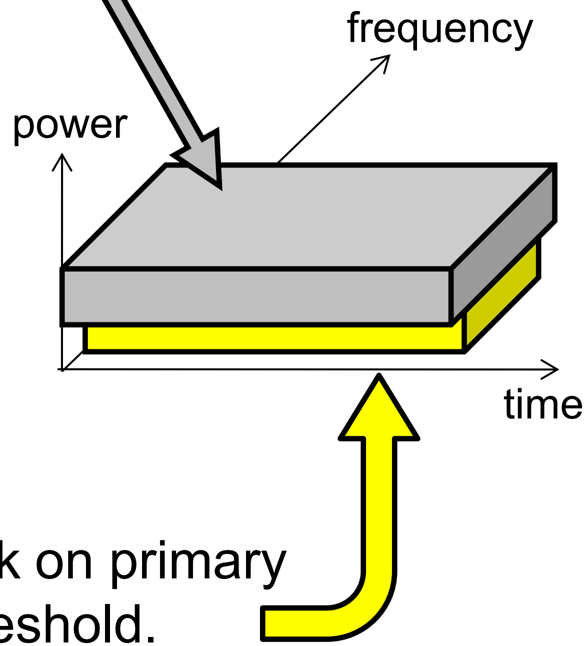
- A primary network (PN) owns a portion of the spectrum.

  Primary network access point.

- A secondary network (SN) transmits over the same portion of the spectrum.

  Secondary network terminal (secondary user – SU) – These are cognitive radios (CRs).

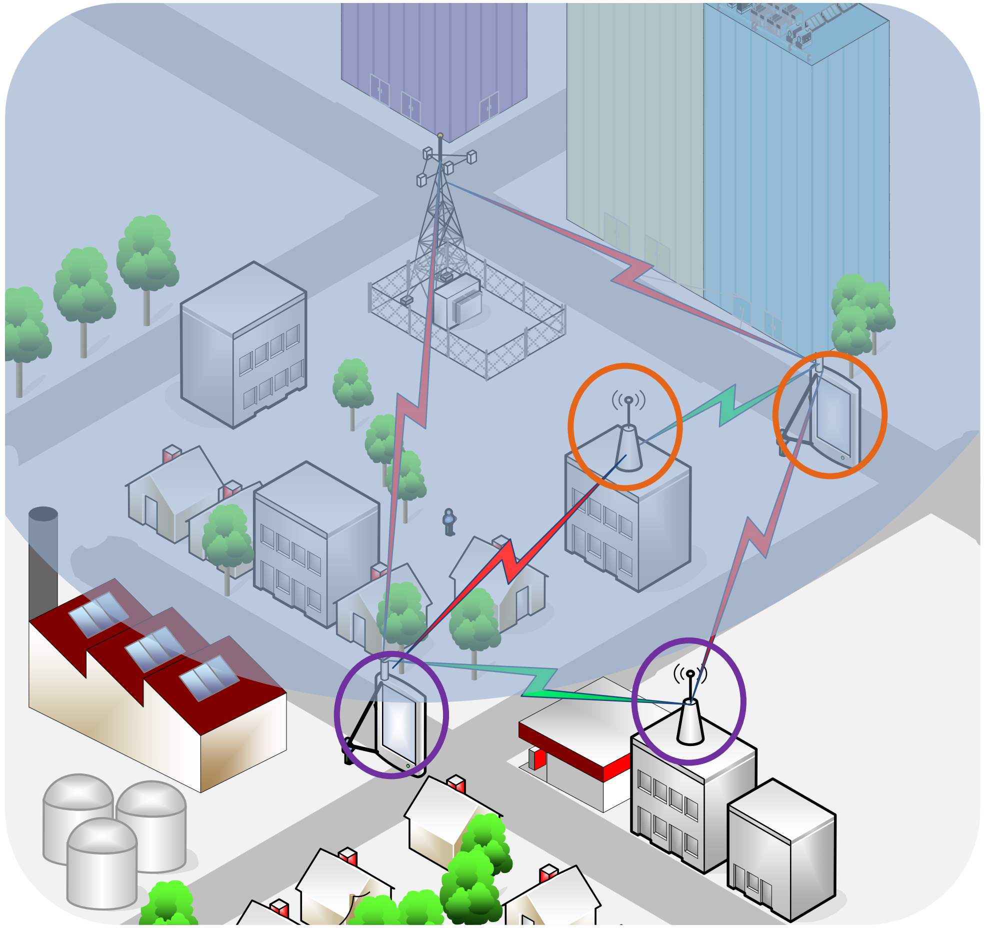
  Secondary network access point.



- Interference created by secondary network on primary network needs to be below a tolerable threshold.

  Interference from secondary network to primary network.

Setup: Cognitive Radio Operation



- Each transmitter find its best transmit power that keeps interference to the PN below threshold.
    - Interference to PN is estimated by leveraging link adaptation → no information exchange between PN and SN.
- Quality of Experience (QoE) requirements determine a minimum Signal-to-Interference-plus-Noise (SINR) ratio at each CR link.
- Because of the use of link adaptation (adaptive modulation and coding) adjusting transmit power and rate is equivalent to adjusting the target SINR for the link.

Deep Q-Learning (DQL) Solution

- Reinforcement learning is a natural fit to solve the problem:
  - Model free approach (we assumed no prior knowledge of the environment).
  - Autonomous learning to adapt behavior to the environment realized following the Observe-Orient-Decide-Act framework:
    - Observe current wireless environment state x_t.
    - Do action a_t on the environment ← **There is no coordination between CRs for this**
    - Observe reward R_t from taken action and new state x_{t+1}.
    - Keep trying actions until identifying the one that for a given state, maximizes expected cumulative reward.
      - → Policy: mapping from state to actions.
      - → Q-values: Estimated expected cumulative reward when taking a certain action when at a given state.
      - → Deep Q-Learning: We use an artificial neural network to estimate Q-values.

Modern Resource Management

→ Today’s networks are expected to carry a broad variety of traffic:

- 5G is the first cellular standard designed with the consideration for different types of traffic: Ultra-Reliable and Low Latency Communications (URLLC), Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (mMTC).
- The challenge has always been that different types of traffic have different service requirements.

→ With 5G there has been a growing attention on end-to-end resource management:

- This implies that service quality is becoming end user-centered.

Measuring Traffic Performance

- → "Traditional" approach to traffic performance measurement:
  - – Based on Quality-of-Service (QoS)
    - ○ **Objective** metrics on individual links (e.g., bit rate, error rate, delay).
- → The end user-centered approach:
  - – Based on Quality-of-Experience (QoE)
    - ○ Metrics that measure **perceptual** (subjective) level of user satisfaction with a service.
- → How to keep track of QoE on live traffic streams?
  - – There exists an extensive body of R&D work that has developed models to calculate QoE metrics from QoS measurements.

The Mean Opinion Score (MOS) Metric

- → QoE metric that rates perceived quality in a scale from 1 (bad) to 5 (excellent)
  - – Originates from the testing of old telephone systems using a panel of human subjects.
  - – Key benefit: there exists models to calculate MOS *from QoS measurements* for practically all traffic types of interest → **Using the same scale to measure different types of traffic is key when doing integrated resource allocation across dissimilar traffic**.
    - ○ Some of the models are ITU standards.

MOS	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The Mean Opinion Score (MOS) Metric

→ In this talk we'll focus on two types of traffic: video and delay-tolerant data.

- They account for a majority of today's traffic volume.

→ Data (FTP) MOS:

$$Q_D = a\log_{10}(b\, r_d(1 - p_{e2e}))$$

end-to-end packet loss probability

data stream bit rate

→ Video MOS:

$$Q_V = \frac{c}{1 + e^{(d(\mathrm{PSNR} - h))}}$$

$$\mathrm{PSNR} = k\log r_i + p$$

(objective video coding quality

video transmission bit rate

MOS Use In The DQL Reward

- DRL aims at finding the action at each state that maximizes the expected cumulative reward.
- The reward r_t measured at each step is a measure of the fitness of the taken action to the goal of the DRL agent.
- In this case we combine MOS with a queuing delay-related reward.

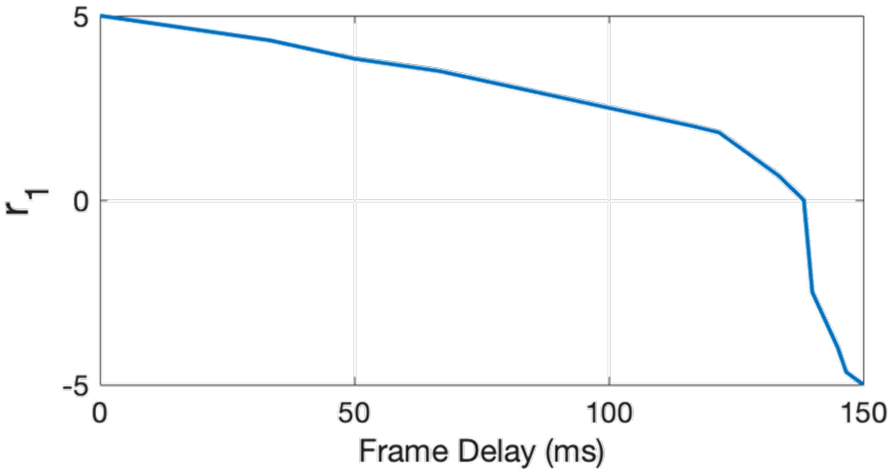
negative constant

$$R_t = \begin{cases} J, & \text{if interference or delay constraint violation,} \\ w_1\ r_1 + w_2\ Q_{(D \text{ or } V)}, & \text{else.} \end{cases}$$

r_1 queueing delay-related delay

r_1 for video traffic (measured on P most recent packets)

(r_1 equals a constant > 0 for data traffic)



DQL State Space and Action Space

- Action space: $\mathcal{A} = \{\tilde{\beta}_1, \ldots, \tilde{\beta}_{|\mathcal{A}|}\}$, where $\tilde{\beta}_i$ is a target link SINR.

- State space: state at time t is $S_t = (I_t, L_t, O_t)$, where

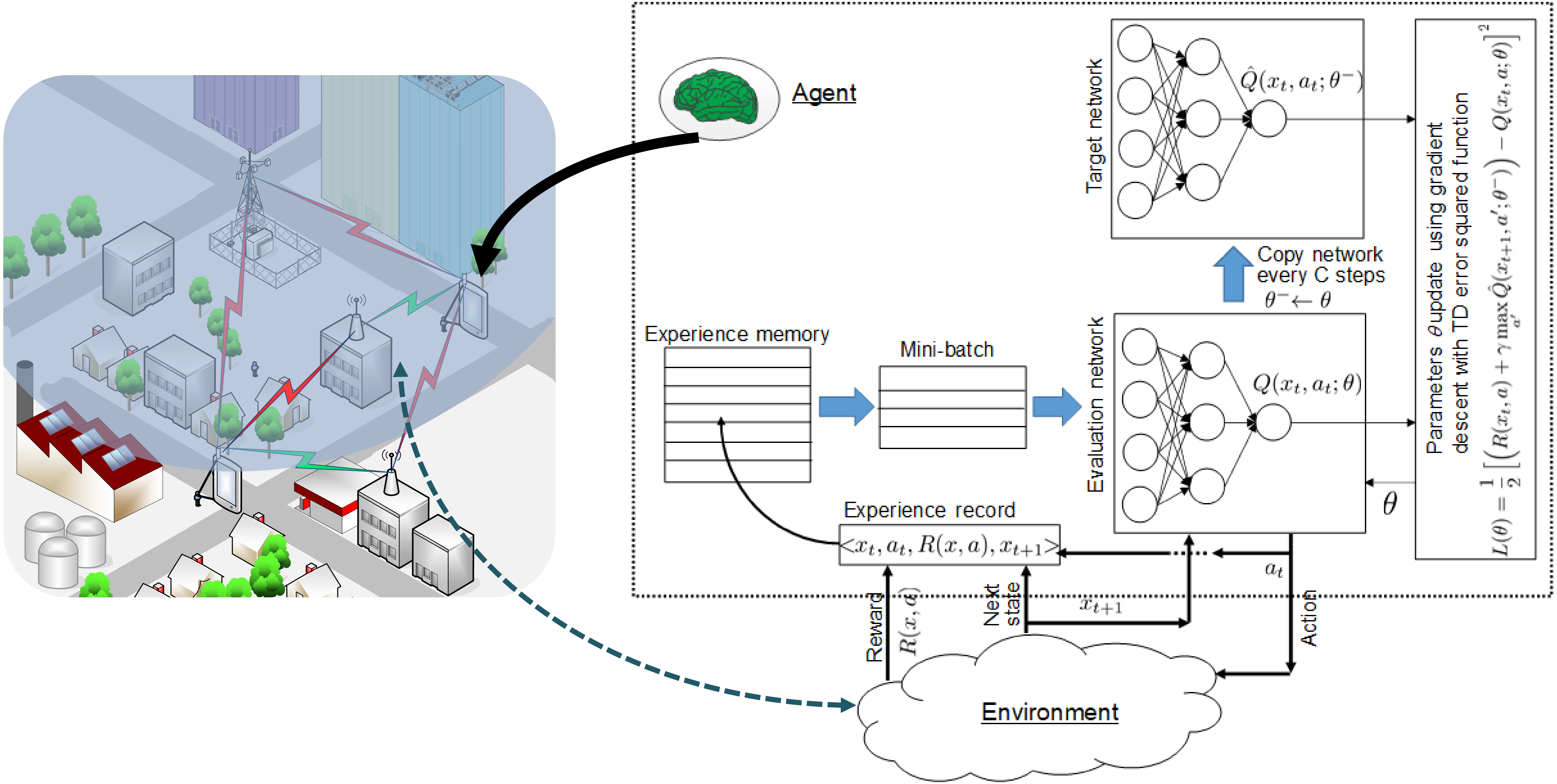
$$I_t = \begin{cases} 0, & \text{if SINR constraint on PN is satisfied,} \\ 1, & \text{else.} \end{cases}$$

$$L_t = \begin{cases} 0, & \text{if SINR constraint in SN is satisfied,} \\ 1, & \text{else.} \end{cases}$$

$$O_t^i = \begin{cases} 0, & \text{if end} - \text{to} - \text{end delay requirement is met,} \\ 1, & \text{else.} \end{cases}$$

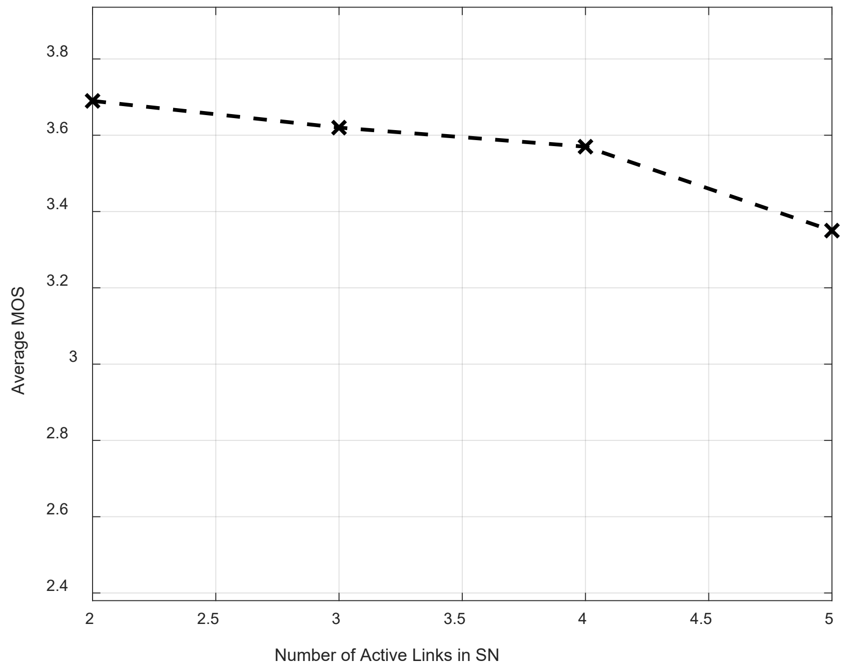
$O_t^i = 0$ for data traffic

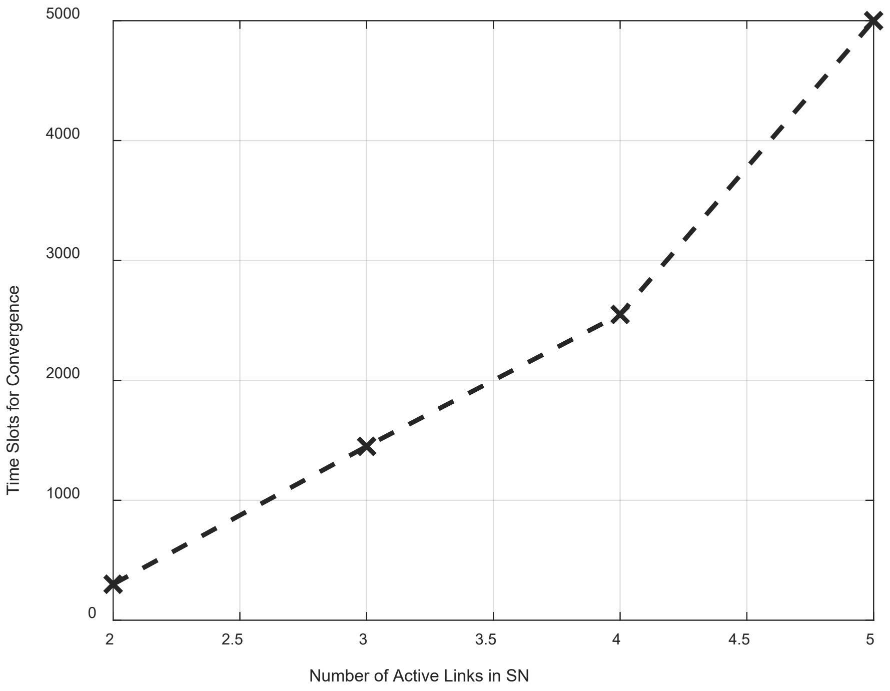
Cognitive Engine Block Diagram


Agent
Target network
$\hat{Q}(x_t, a_t; \theta^-)$
Copy network every C steps
$\theta^- \leftarrow \theta$
Experience memory
Mini-batch
Evaluation network
$Q(x_t, a_t; \theta)$
Parameters θ update using gradient descent with TD error squared function
$L(\theta) = \frac{1}{2}\left[\left(R(x_t, a) + \gamma \max_{a'} \hat{Q}(x_{t+1}, a'; \theta^-)\right) - Q(x_t, a; \theta)\right]^2$
θ
Experience record
$<x_t, a_t, R(x, a), x_{t+1}>$
a_t
Reward
$R(x, a)$
Next state
x_{t+1}
Action
Environment

Cognitive Engine Learning Performance

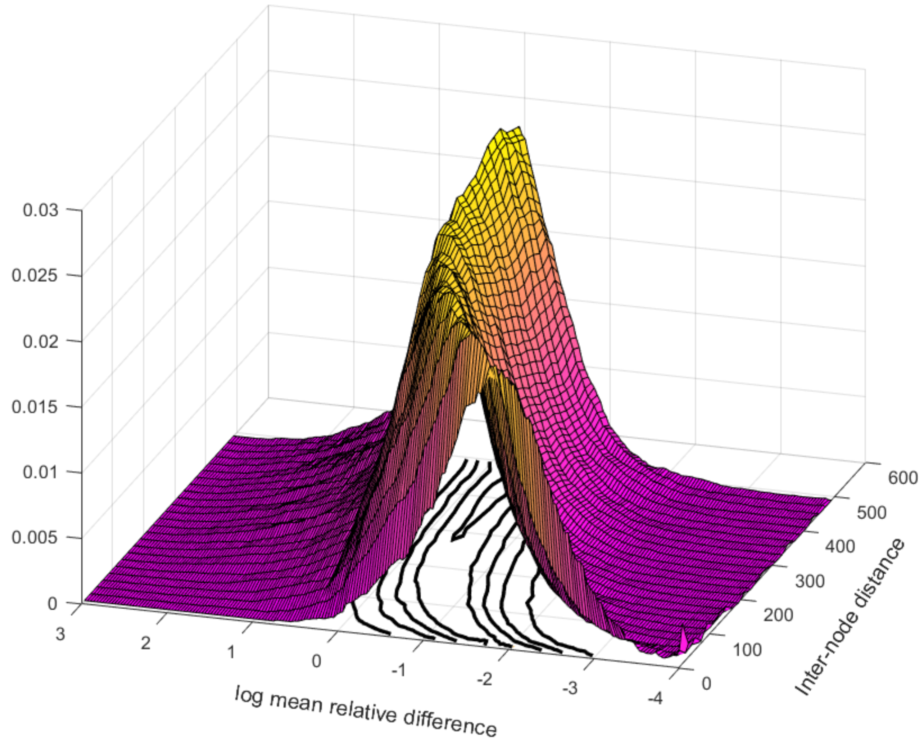
- We see that the cognitive engine is able to learn a policy for the resource allocation policy.
  - However, learning takes too long.
    - ○ This is even as DQL does learn faster than the traditional table-based approach.



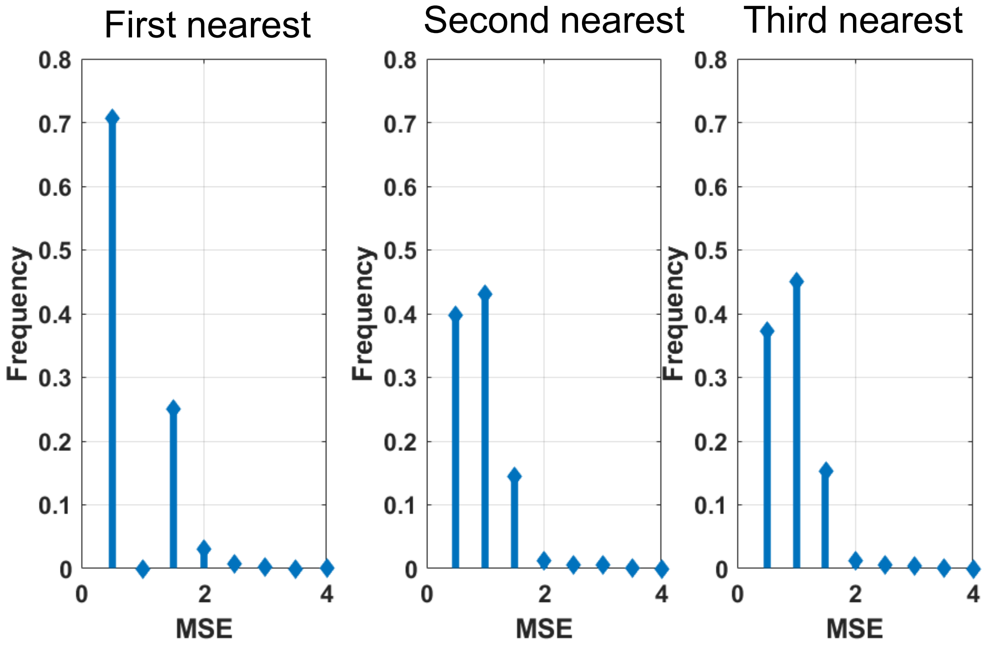


Learning Faster

- How could a CR learn faster?
  - Let’s take a deeper dive into what is learned and how it is represented:



Average of the relative difference in the Q-values between two CRs



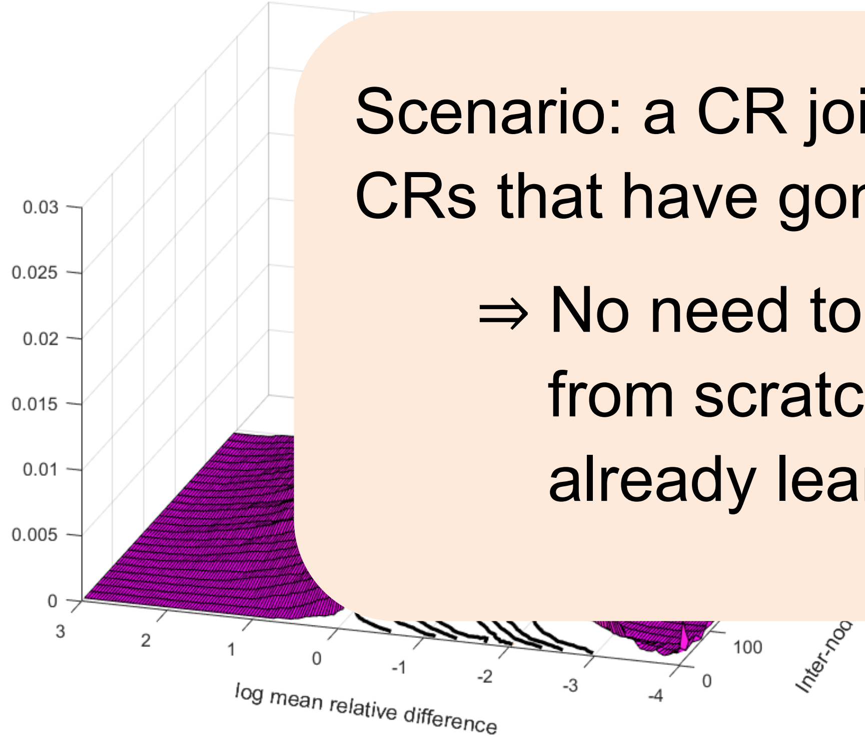
Difference in the DQL parameters θ between a CR and its first, second, and third nearest neighbor CR

Learning Faster

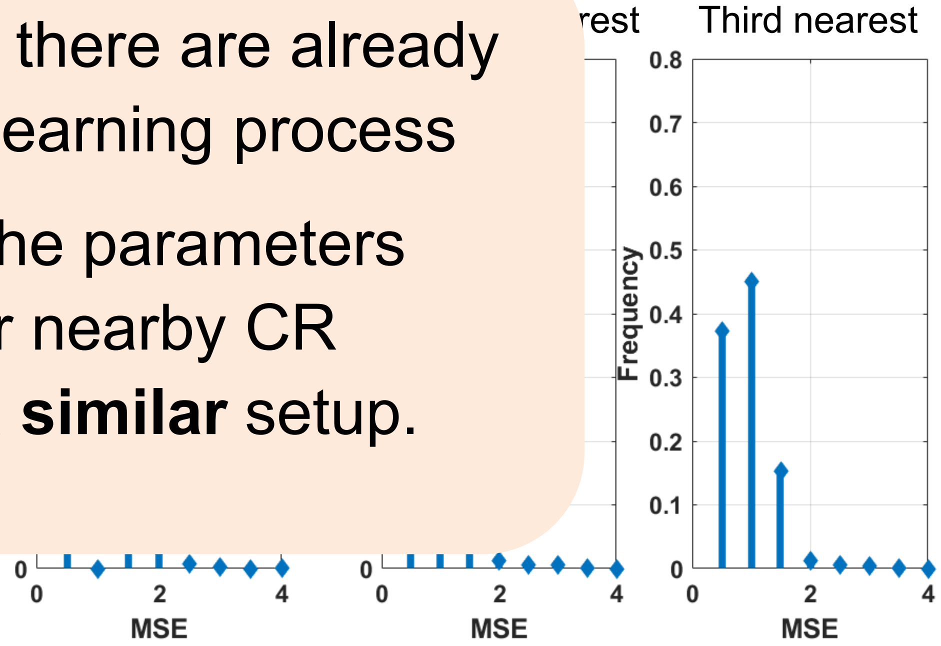
- How could a CR learn faster?
  - Let’s take a deeper dive into what is learned and how it is represented:

Scenario: a CR joins a SN where there are already CRs that have gone through the learning process

⇒ No need to start learning the parameters from scratch when another nearby CR already learned them for a **similar** setup.



Average of the relative difference in the Q-values between two CRs



Difference in the DQL parameters θ between a CR and its first, second, and third nearest neighbor CR

Learning Faster

⇒ No need to start learning the parameters from scratch when another nearby CR already learned them for a **similar** setup.

- The new CR has no experience
- The existing CR has experienced

The existing nearby CR could be a “teacher” to the new CR (the “student”), transferring the experience it already has so that the “student” CR does not need to learn from scratch.

θ
A MASTER AND AN APPRENTICE,
ALWAYS TWO THERE ARE, NO LESS, NO MORE

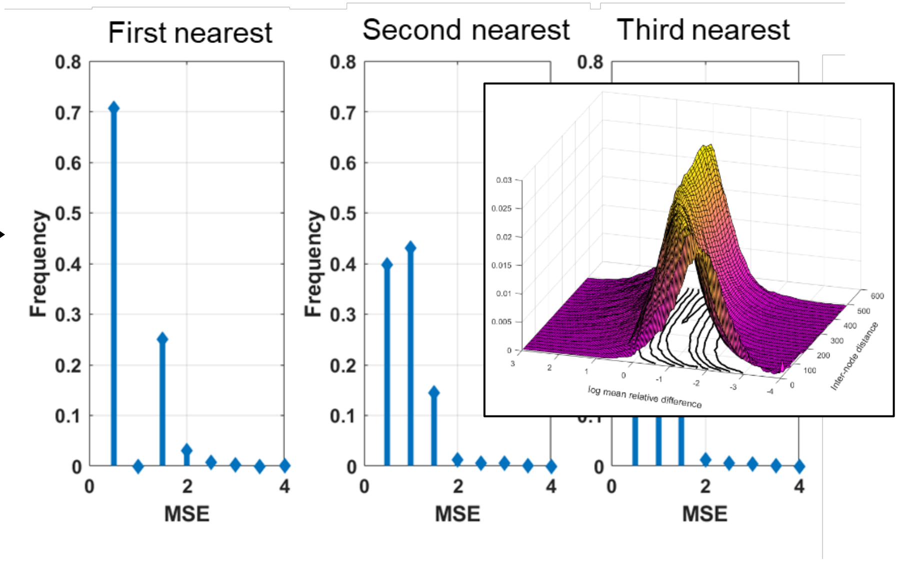
Image created using Microsoft Copilot

Achieved by transferring the parameters θ from the teacher to the student

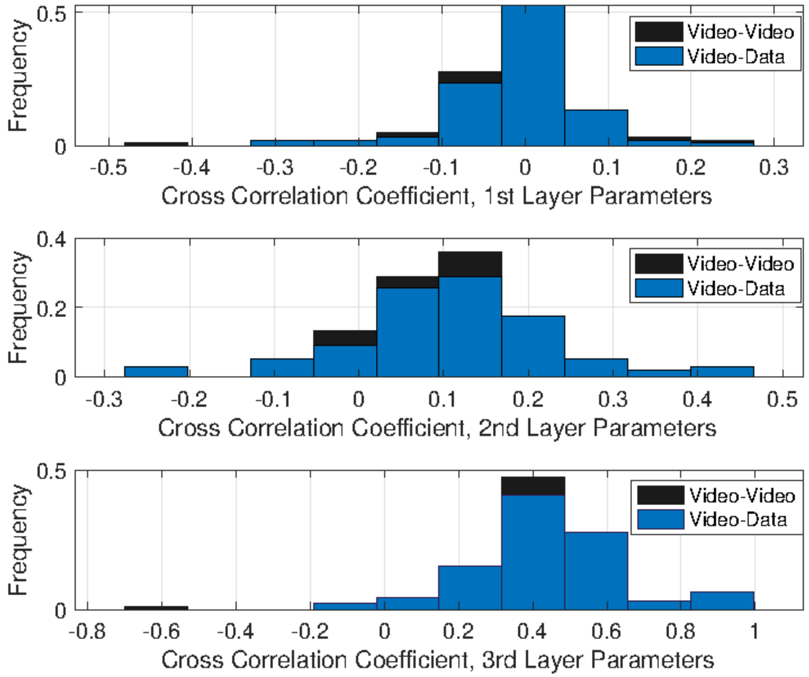
(followed by a stage of student learning to fine tune the parameters)

Implementing the Transfer of θ

- What CR acts as “teacher” to a newcomer CR?
  - We already know the answer to this: pick the **nearest** CR.



- Addressing the overhead from the transfer of θ:



- Studying the cross-correlation between teacher and student parameters θ at each layer of the DQL neural network:
  - Parameters θ after learning converges.
  - Teacher-student distance is the mean of the distance random variable.

⇒ Third layer shows the largest correlation

Mechanism Is Agnostic To Traffic Type

- The distribution of the cross-correlation coefficients shows little difference whether the teacher and student carry the same or different types of traffic.
  - **This is a consequence of using MOS to assess reward (using the same scale to measure action fitness for all traffic types).**

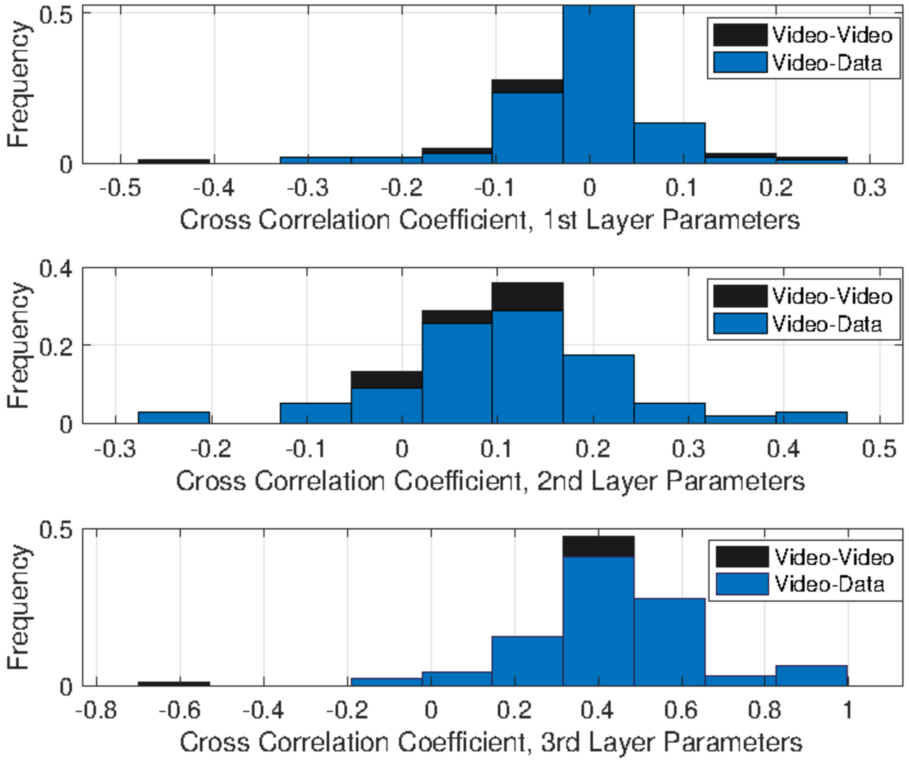
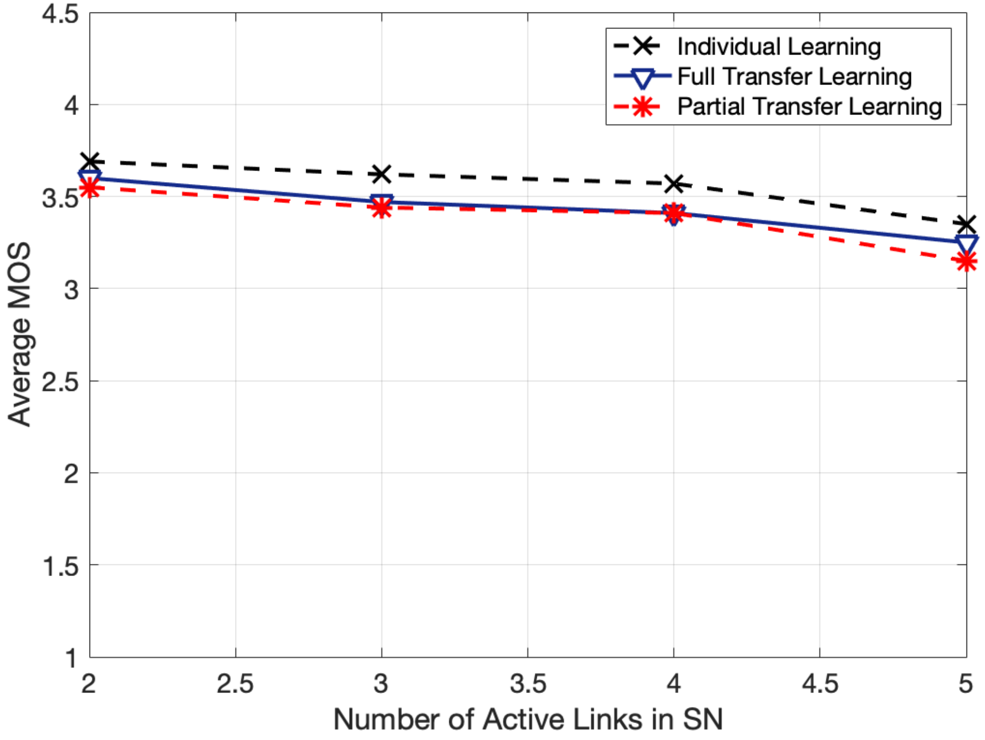


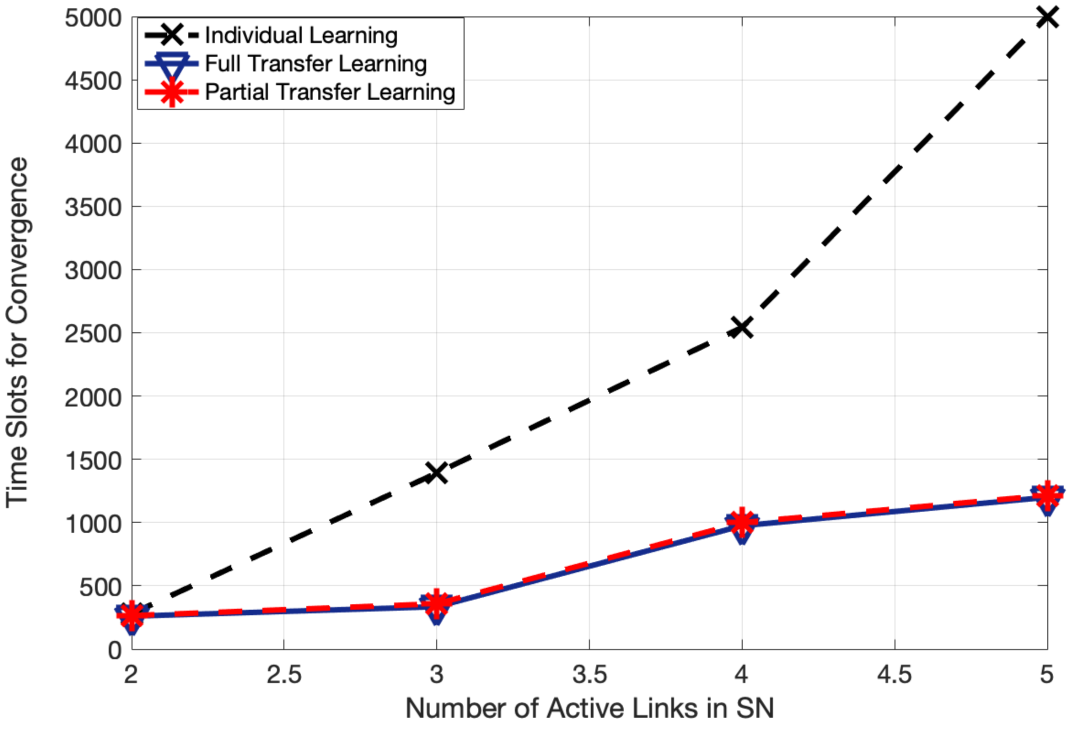
Image created using Microsoft Copilot

=

Image created using Microsoft Copilot

Results





- Individual learning: random initialization of all 31 parameters.
- Full transfer learning: All 31 parameters are initialized from the teacher’s transfer.
- Partial transfer learning: Last layer (3 parameters) from the teacher; other parameters are initialized at random.

Conclusions

- Discussed CR learning:
  - » Underlay dynamic spectrum access/sharing.
  - » Network **supports multiple types of traffic** (video and data – FTP- considered).
  - » Deep Reinforcement Learning for resource allocation.
  - » Reward using MOS → End-to-end QoE approach, **common scale for all traffic types**.
- Learning faster:
  - » A CR joining the network (the "student") receives the DQL parameters from the nearest neighbor (the "teacher").
    - **The parameters are agnostic to the traffic type (because of the reward design)**.
    - 80% learning steps reduction with negligible sacrifice in average QoE performance.
    - 90% parameter transfer overhead reduction by transferring only the parameters in the last layer – no change in QoE from full transfer.

Bonus! Thoughts About Educating Researchers

- Where I’m coming from (what occupies me most of the day): I’m the ECE PhD Director at RIT
  - » My main job is to create an ecosystem that forms elite researchers.
  - » One of my main concerns: the quantity-vs-quality pendulum.
    - The struggle has always existed but I think that decades ago the pendulum was more on the quality side.
    - Today the quantity-over-quality ethos has become a wave too enormous to be able to counter as individual researchers (many factors for this, but the evidence is clear → see number of conferences, submission statistics, growth of arXiv, etc.)
    - A realistic perspective: I need to admit that PhD students will need to be within the quantity wave if they are to succeed in their career as a researcher.

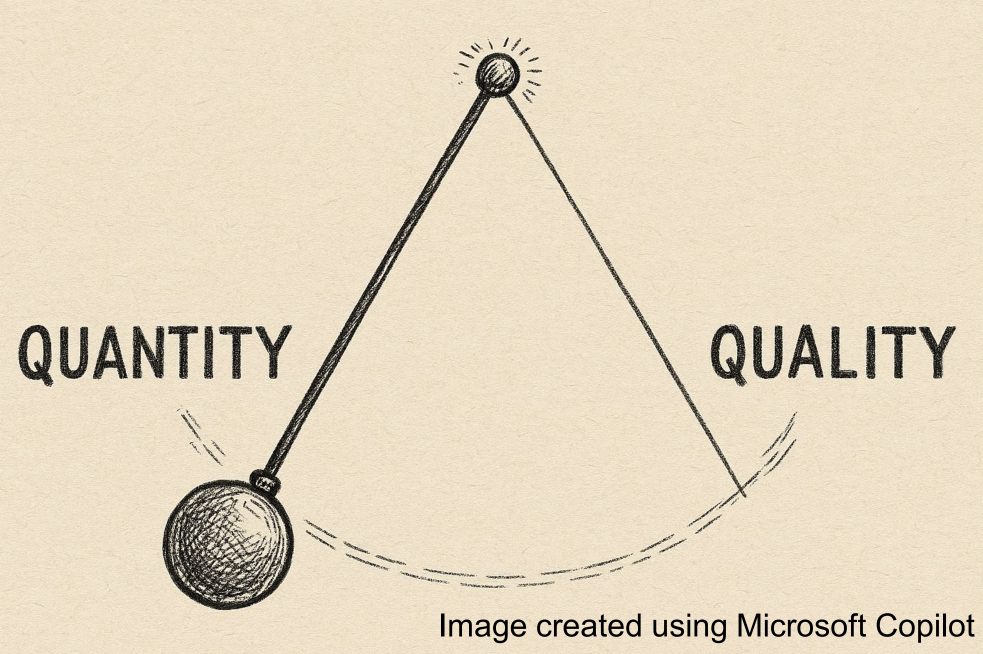


Image created using Microsoft Copilot

Quantity Does Little to Bridge Knowledge Gaps

- While in the abstract, the reasons of why “quality over quantity” appear obvious, the reality is that there are multiple pressures that drive the researcher's mindset towards a “quantity over quality” mentality.
- Quality over quantity → a useful metaphor:
  - » Research outcomes lay down stepping stones in the path of advancing knowledge.
  - » High-quality, high-impact research is like a stepping stone in the form of a large stone slab → Compared to this, low-quality research would look like a pebble.
  - » If our path of innovation and discovery gets to a point where we need to build a bridge across a fast-flowing river, what kind of stepping stones will get us over the gap?

QUANTITY
QUALITY

Image created using Microsoft Copilot

What To Do With the Quantity Wave?

- I tell the students to think themselves as a surfer:
  - » They don’t want to be caught so much in the quantity-over-quality wave that brings them to the shore tumbling and rolling,
  - » I want them to be the ***expert*** surfer that gracefully rides the wave to the shore with ***quality*** moves,
  - » But to achieve this, they need to know the research techniques (the “surfing techniques”) and what is that they are doing when using them.

Image created using Microsoft Copilot

A Unique Research Methods Course

- “Research Methods in ECE”:
  - » A course conceived for ECE students to learn typical techniques used in our research and to **think** about how they work.
  - » It is not the conventional research methods course → At RIT another course teaches, for example, how to write a paper.
  - » Some topics covered: The dovetail of critical and creative thinking in research, inductive and deductive reasoning, argumentation in informal logic, formal logic, deductive systems, fallacies, abstraction (and its relation with system modeling), assumptions, trade-offs.

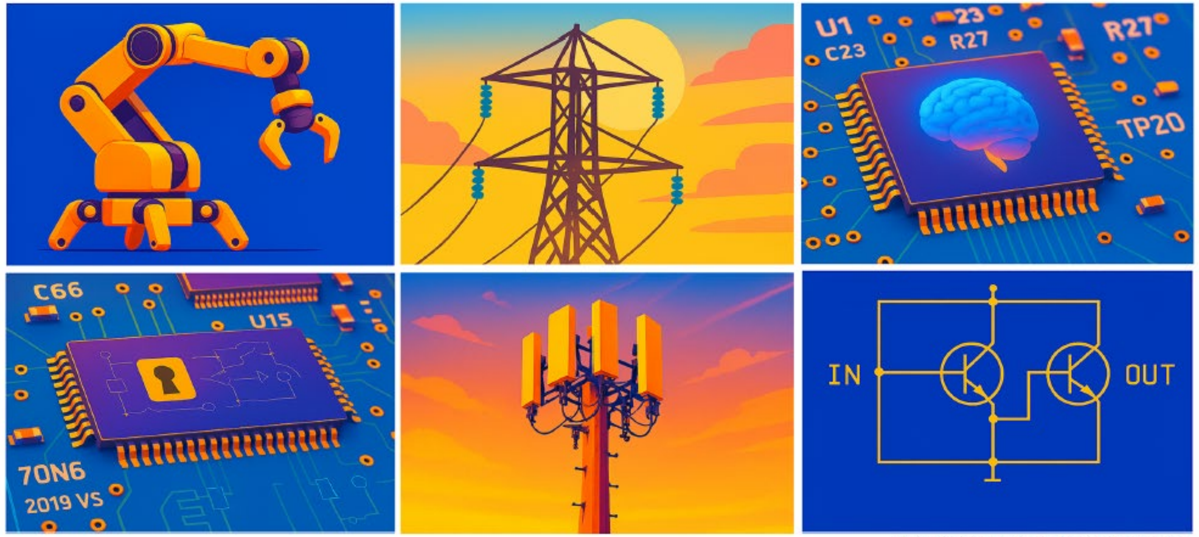


Image created using Microsoft Copilot.

https://people.rit.edu/axkeec/research_methods_ece_notes.pdf

To download the course notes →

Thank You!