Foundation of Data System Research (OT1)

The first semester of your PhD (in data management)

Riccardo Tommasini

Who I Am

- **Riccardo Tommasini**
- Associate Professor at INSA Lyon, LIRIS
- Former Lecturer at UT
- Streaming Enthusiast!
- 🇮🇹->🇺🇸->🇪🇪->🇫🇷

DB Team
(a subset)

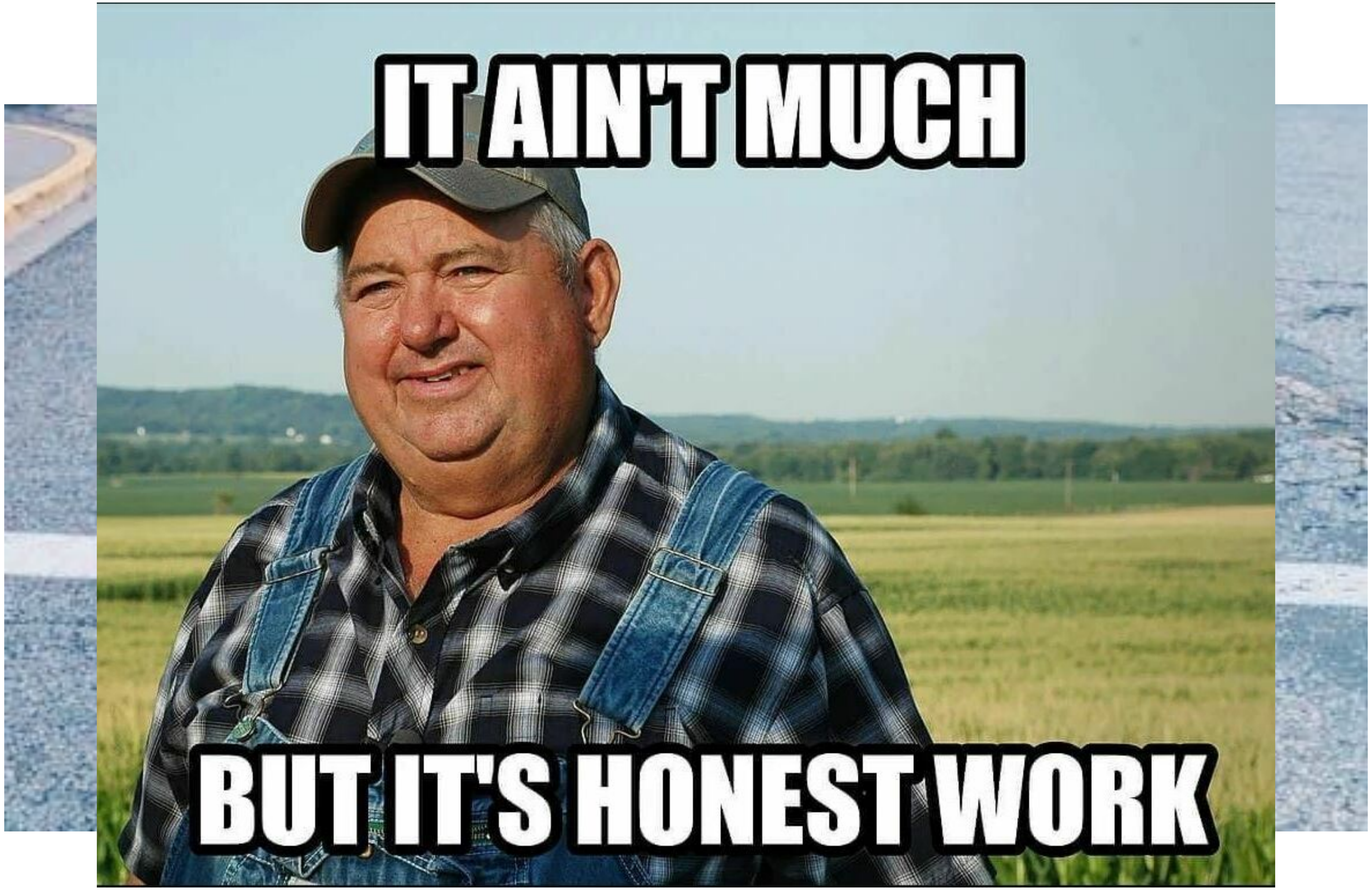
DB Team
(a subset)

Riccardo’s Pros and Cons

- I try to come to class as energetic as I can (8am permitting)
- I am stateless (and hopeless), it means i don’t hold the grudge
  - please remind me if I promise something in class
  - if i don’t follow up emails in 24h consider it lost in the pile
- I am Italian

- I can be intimidating sometimes, feel free to ask me to slow down
  - especially if I speak to fast (don’t wait until the end to tell me you don’t understand)
- Being stateless also means I forget about bad things, ask questions!
  - I never hold the grudge even if i have to repeat many times!

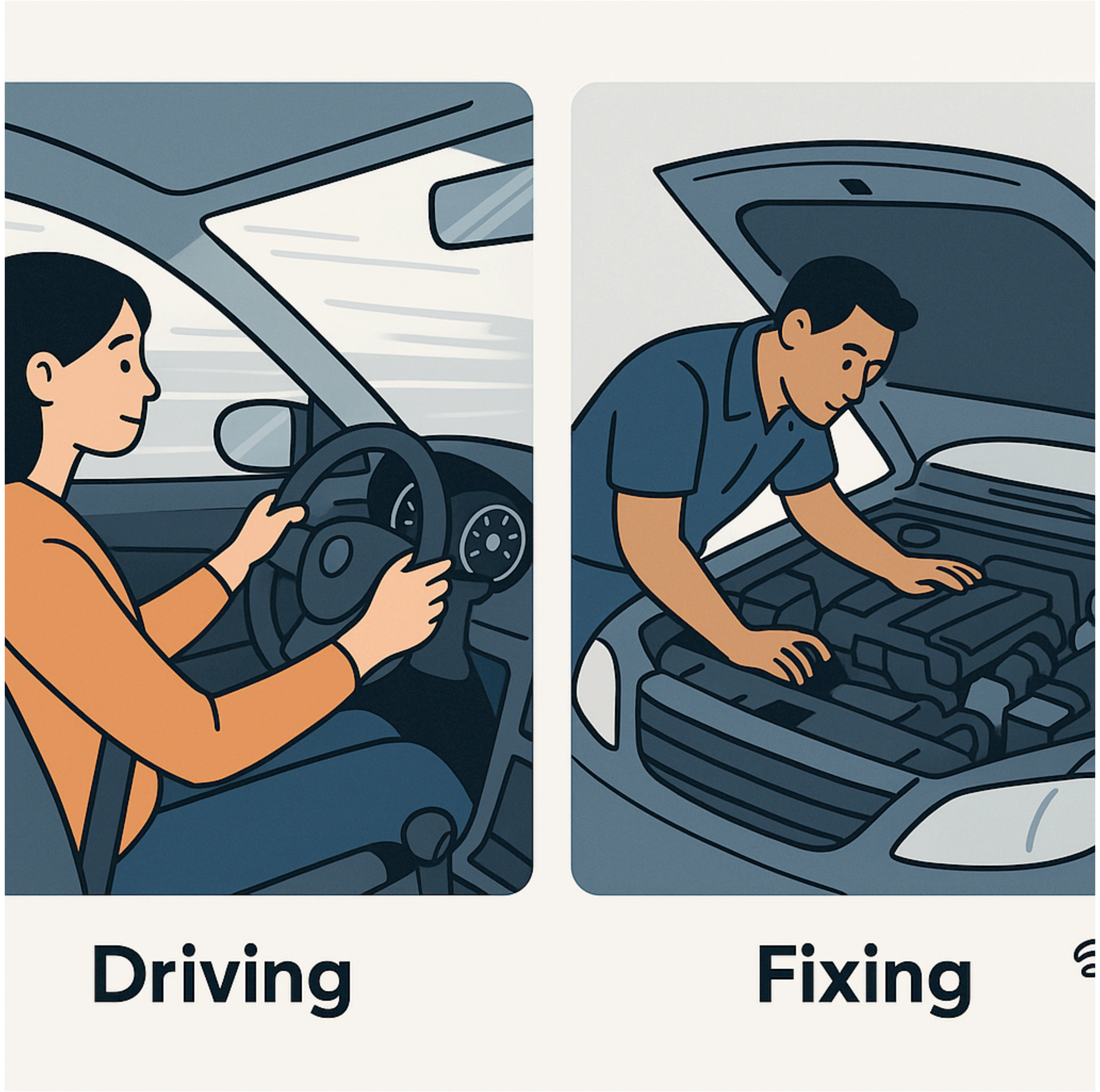
My Teaching Style



Data Systems

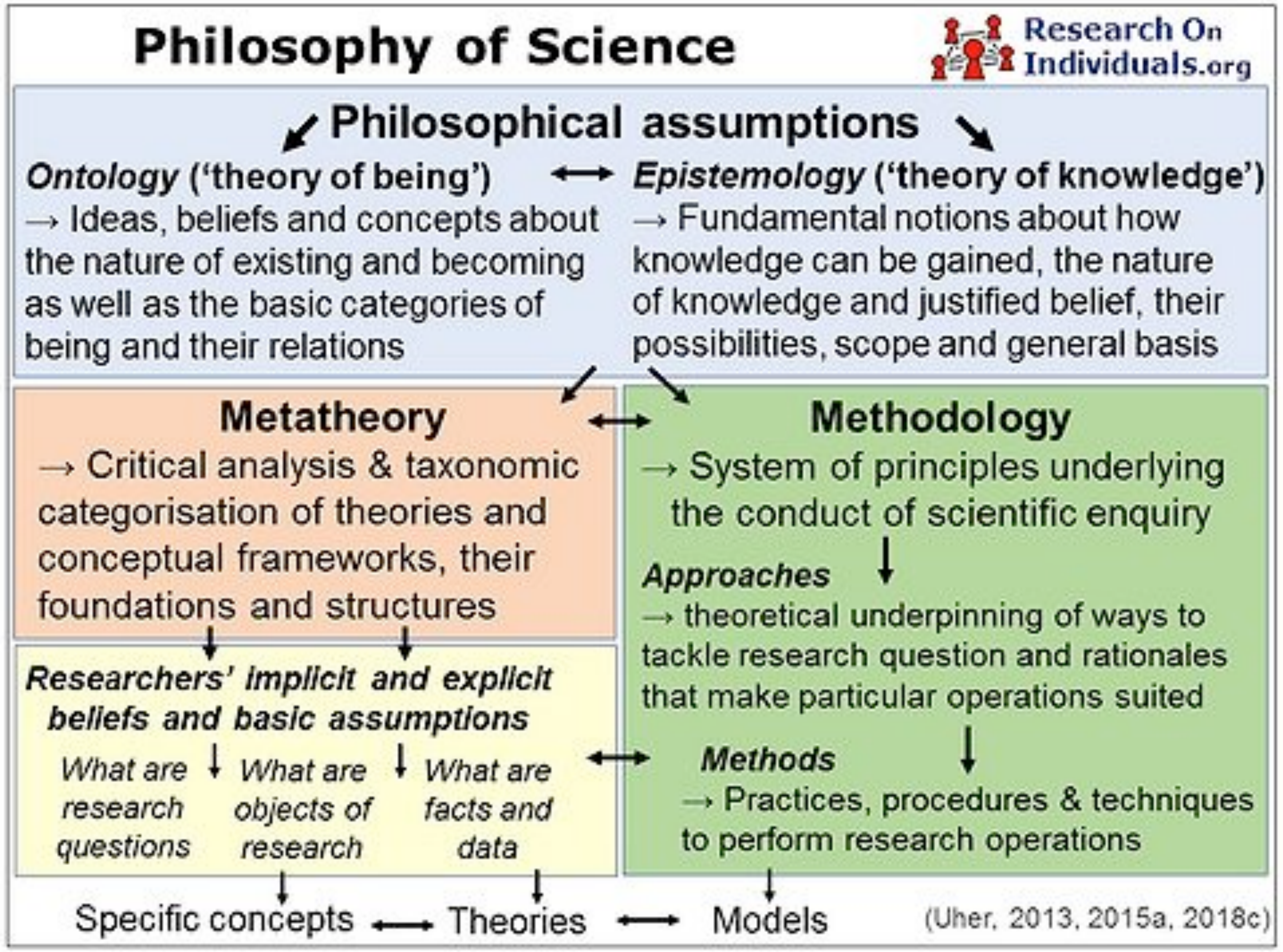
What are they?

- **Store & organize data** — Ingests data from sources, keeps it durably, and structures it (files/tables/indexes) so it can be found.
- **Process & make it usable** — Transforms and queries data with compute (batch/stream), optimizing for scale, latency, and reliability.
- **Govern & serve** — Enforces security, quality, and lineage, then exposes data via APIs, queries, dashboards, or ML services to people and apps.



Science

Philosophy of



Philosophy of Science

Research On Individuals.org

Philosophical assumptions

Ontology ('theory of being')

→ Ideas, beliefs and concepts about the nature of existing and becoming as well as the basic categories of being and their relations

Epistemology ('theory of knowledge')

→ Fundamental notions about how knowledge can be gained, the nature of knowledge and justified belief, their possibilities, scope and general basis

Metatheory

→ Critical analysis & taxonomic categorisation of theories and conceptual frameworks, their foundations and structures

Methodology

→ System of principles underlying the conduct of scientific enquiry

Approaches

→ theoretical underpinning of ways to tackle research question and rationales that make particular operations suited

Researchers' implicit and explicit beliefs and basic assumptions

What are research questions

What are objects of research

What are facts and data

Methods

→ Practices, procedures & techniques to perform research operations

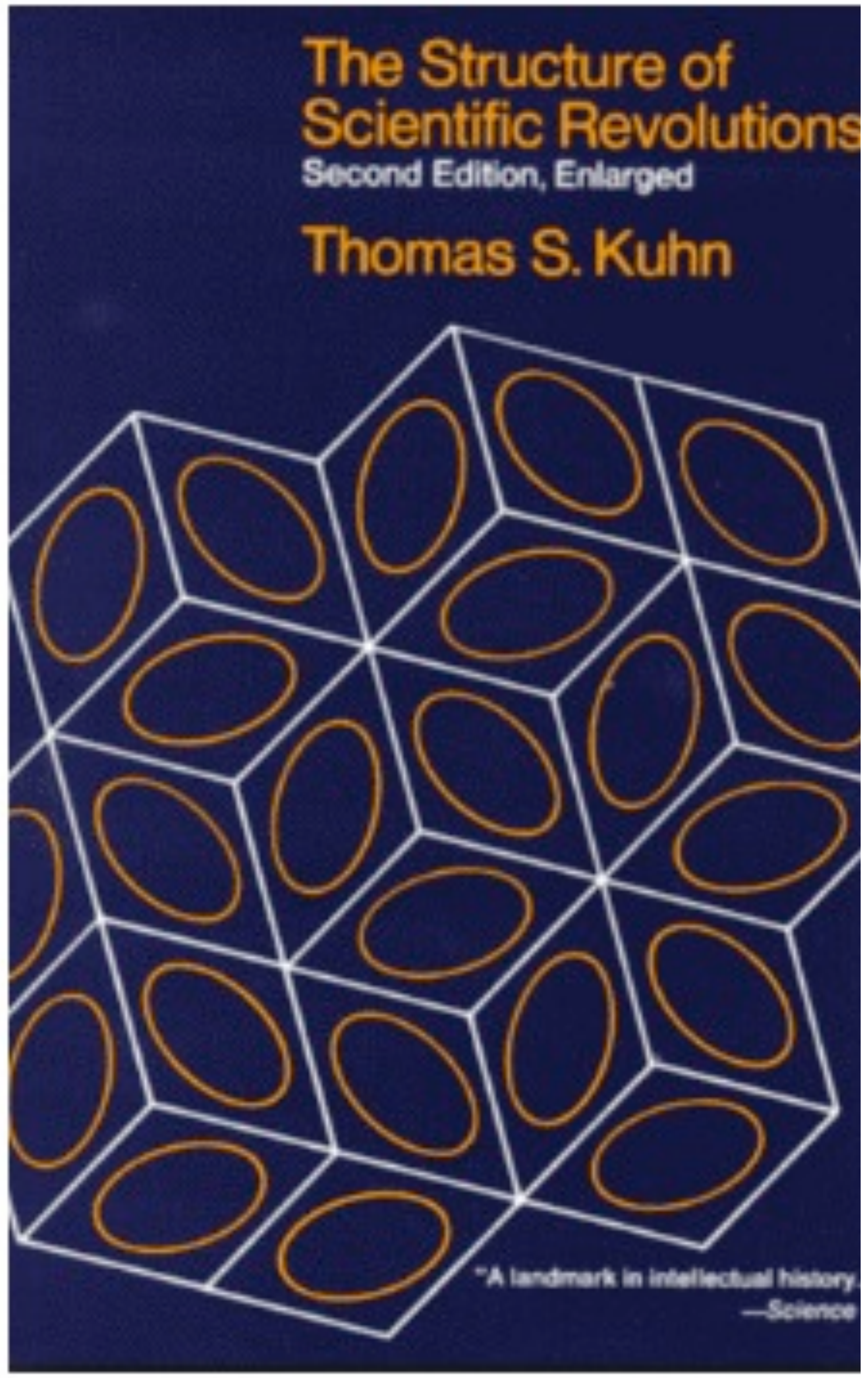
Specific concepts ⟷ Theories ⟷ Models

(Uher, 2013, 2015a, 2018c)

Thomas Khun

How does science work?

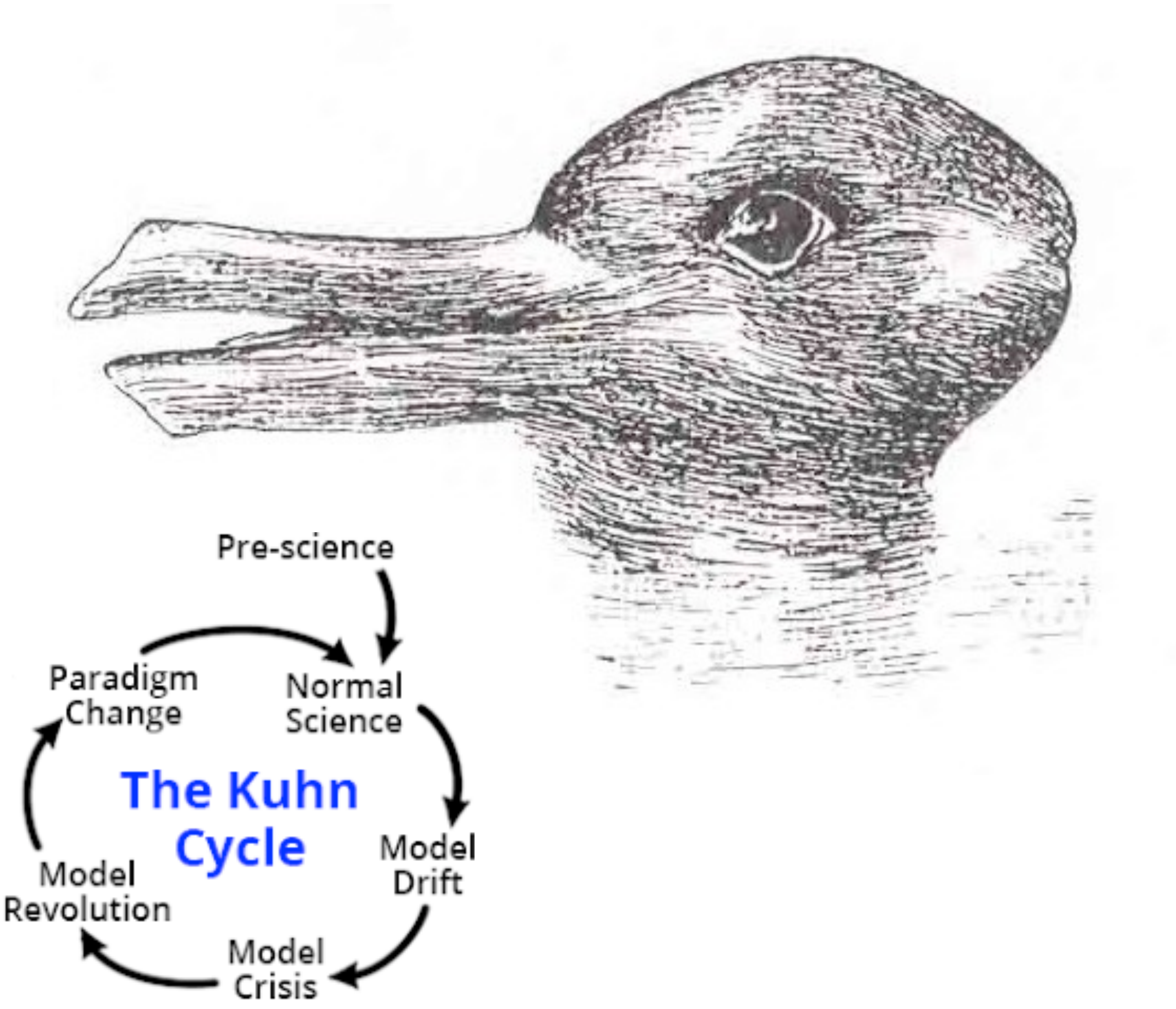
A scientific paradigm is characterized by a set of **theories** and **ideas** that define what is **possible** and **rational** to do, giving scientists a clear set of tools to approach certain **problems**.



Thomas Kun

Revolutionary View

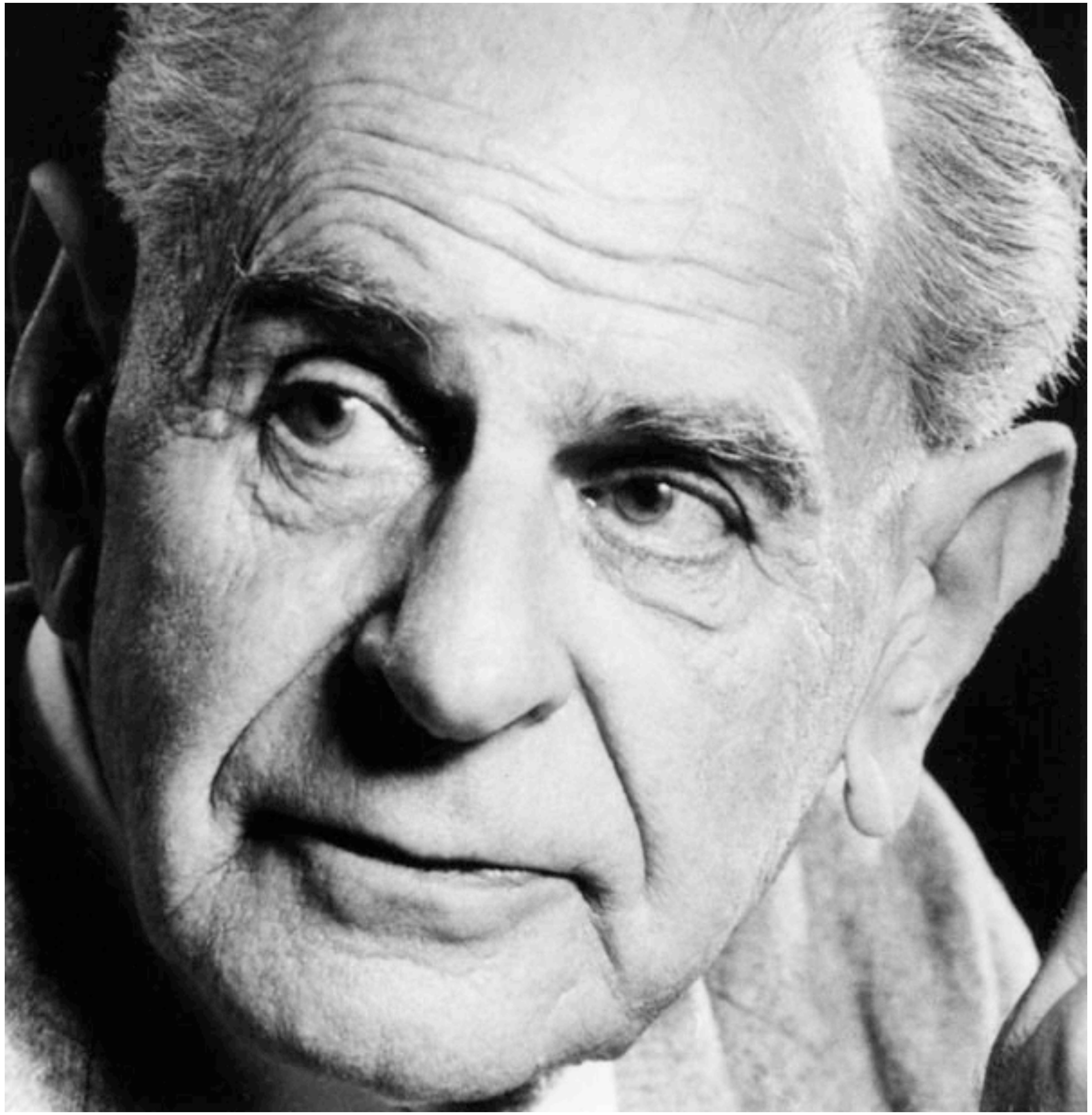
- Normal science – a dominant paradigm is active.
- Extraordinary research – the scientific discipline is thrown into a state of crisis.
- Adoption of a new paradigm .
- Aftermath of the scientific revolution



Karl Popper

How does science work?

Falsifiability is a standard of evaluation of scientific theories and hypotheses. A hypothesis is falsifiable if it belongs to a language or logical structure capable of describing an empirical observation that contradicts it.



Objectives

Understand

- Identify the main problem discussed and addressed in the paper
- Identify the main experimental hypothesis behind the devalutation
- judge if the evaluation is correctly design to test such hypotheses

Conceive

- Understand and re-elaborate a research question
- formulate the problem statement for your (ours) idea
- formalize the problem statement using simple math

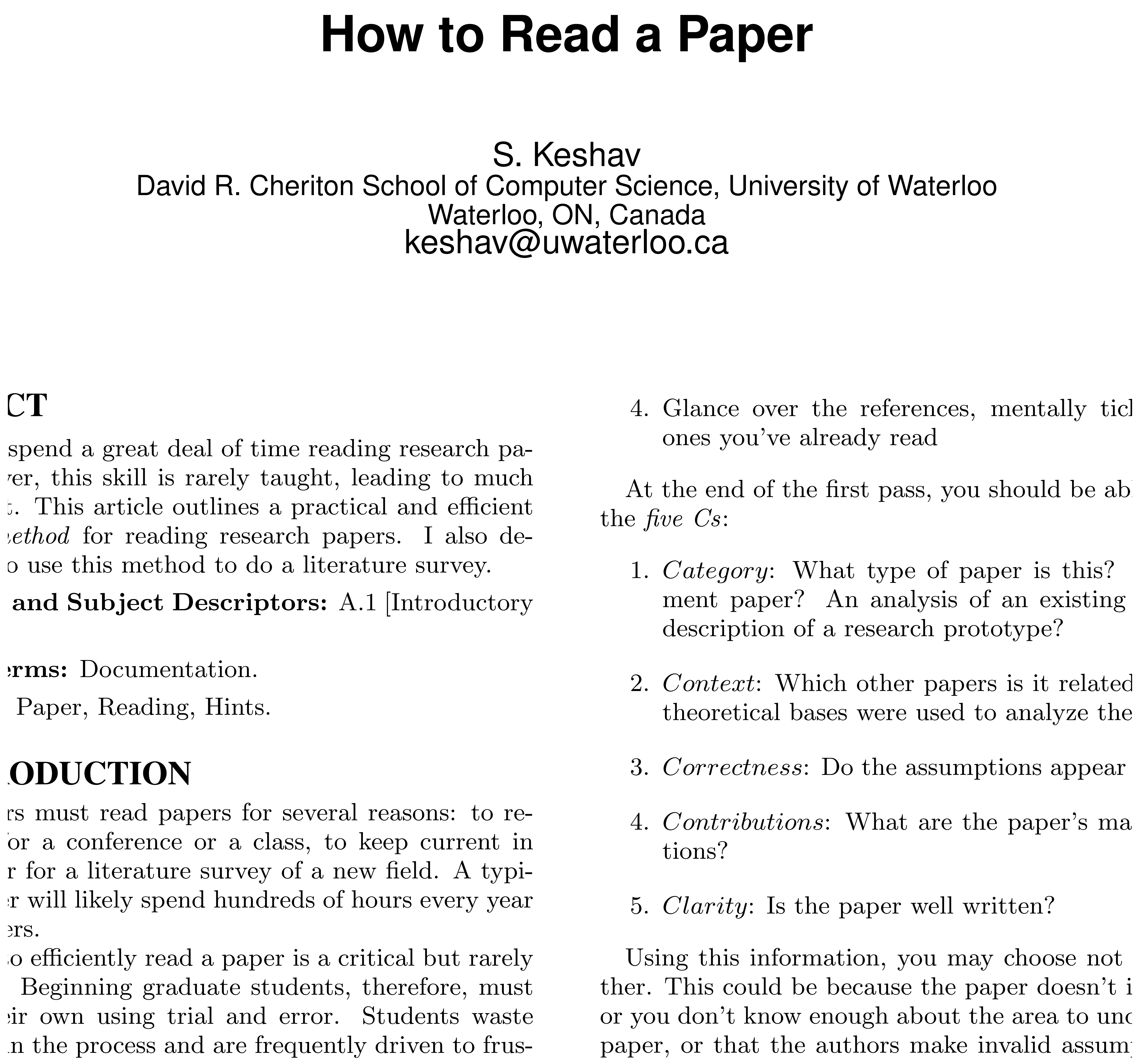
Assess

- **Validate** the idea by using your “engineering” skills
- **Evaluate** (see Popper) the resulting solution for the problem statement

How to Read a Research Paper

Understand

- We collected several research paper around two main topics
  - Stream Processing
  - Graph Processing
- You will need to read them and chose 1 in a group of 3 to replicate the evaluation



How to Read a Paper

S. Keshav
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

CT

spend a great deal of time reading research pa-
ver, this skill is rarely taught, leading to much
:. This article outlines a practical and efficient
ethod for reading research papers. I also de-
o use this method to do a literature survey.

and Subject Descriptors: A.1 [Introductory

rms: Documentation.

Paper, Reading, Hints.

ODUCTION

rs must read papers for several reasons: to re-
or a conference or a class, to keep current in
r for a literature survey of a new field. A typi-
r will likely spend hundreds of hours every year
rs.

o efficiently read a paper is a critical but rarely
Beginning graduate students, therefore, must
ir own using trial and error. Students waste
n the process and are frequently driven to frus-

4. Glance over the references, mentally ticl
   ones you've already read

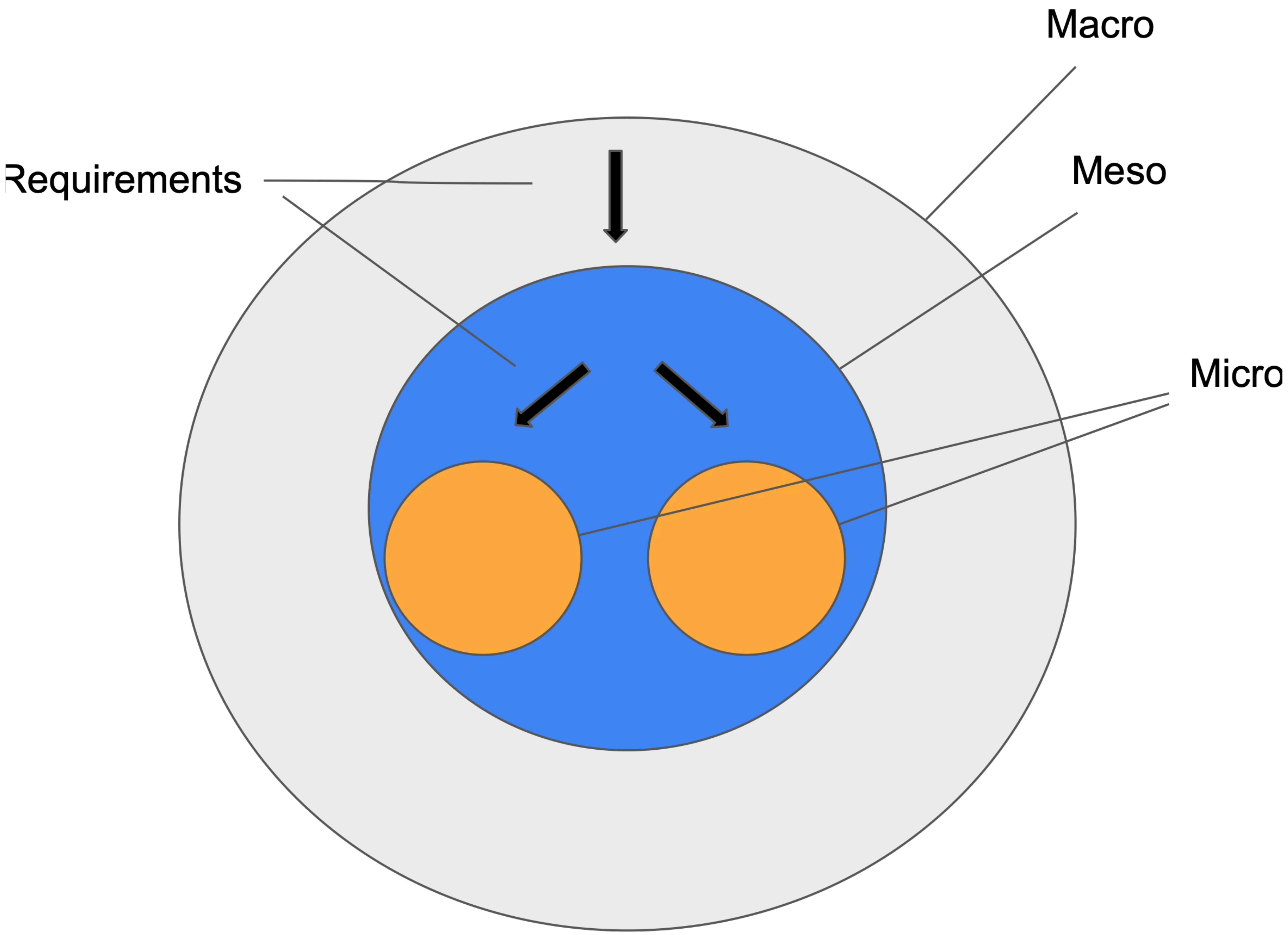
At the end of the first pass, you should be ab
the *five Cs*:

1. *Category*: What type of paper is this?
   ment paper? An analysis of an existing
   description of a research prototype?
2. *Context*: Which other papers is it related
   theoretical bases were used to analyze the
3. *Correctness*: Do the assumptions appear
4. *Contributions*: What are the paper's ma
   tions?
5. *Clarity*: Is the paper well written?

Using this information, you may choose not
ther. This could be because the paper doesn't i
or you don't know enough about the area to unc
paper, or that the authors make invalid assum

The Macro-Meso-Micro framework

Problem Formulation



Sandro Serpa and Carlos Ferreira. “Micro, Meso and Macro Levels of Social Analysis”. In: International Journal of Social Science Studies 7 (Apr. 2019), p. 120. doi: 10.11114/ijsss.v7i3.4223.

A framework to scope your problem statement

Macro: broad, complex, and unanswerable questions.

- What drives a research community

Meso: specific, but still unanswerable questions

- What drives (small) groups

Micro: where the actual research questions are formalized

- Works of single or small group of people

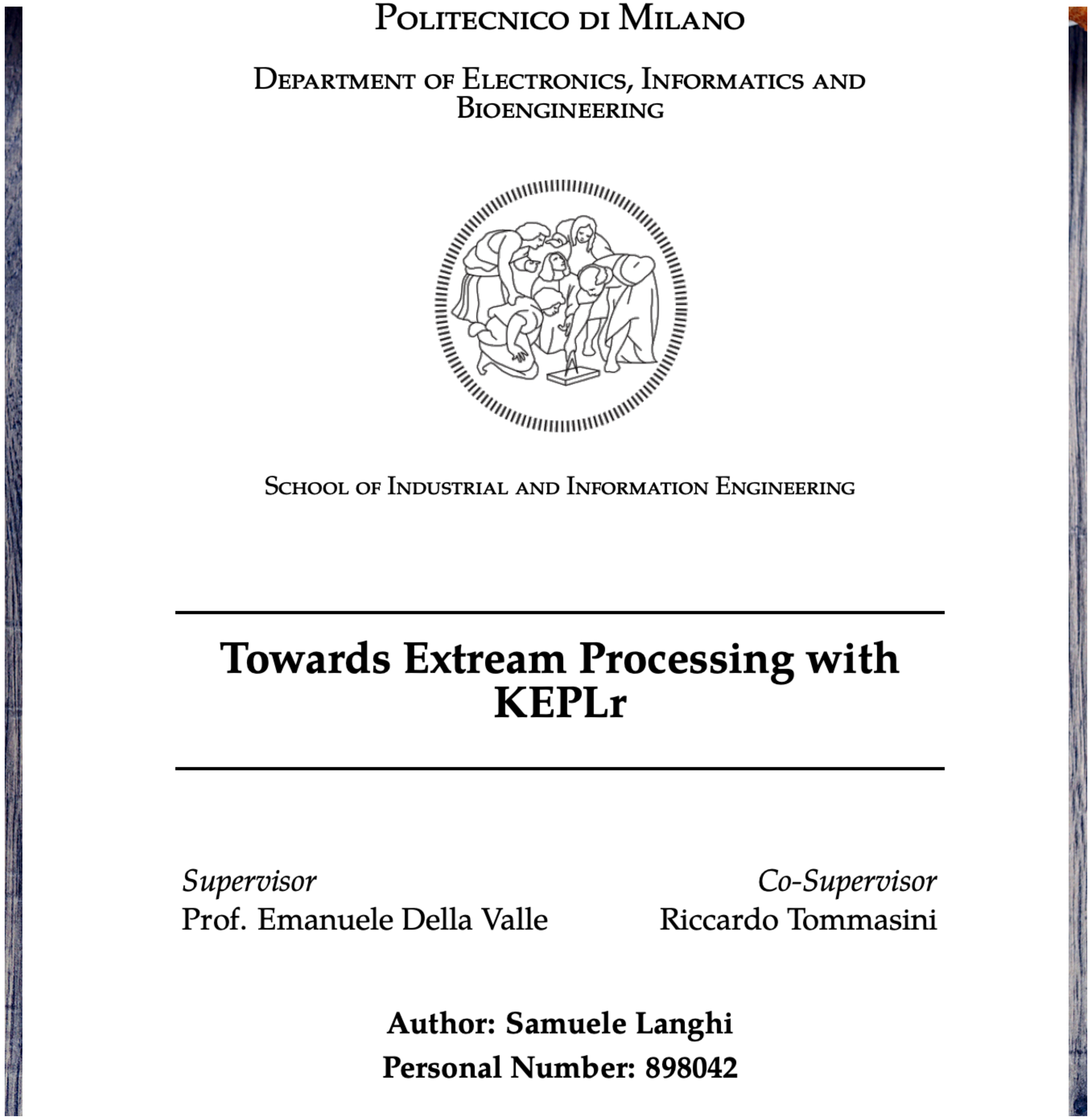
Example

Big-data challenges

- Volume
- Velocity
- Variety

Class of applications

- Big Data Stream Processing Engines (BigSPEs)
  - *eXtream Processing (XP)* system



POLITECNICO DI MILANO

DEPARTMENT OF ELECTRONICS, INFORMATICS AND BIOENGINEERING

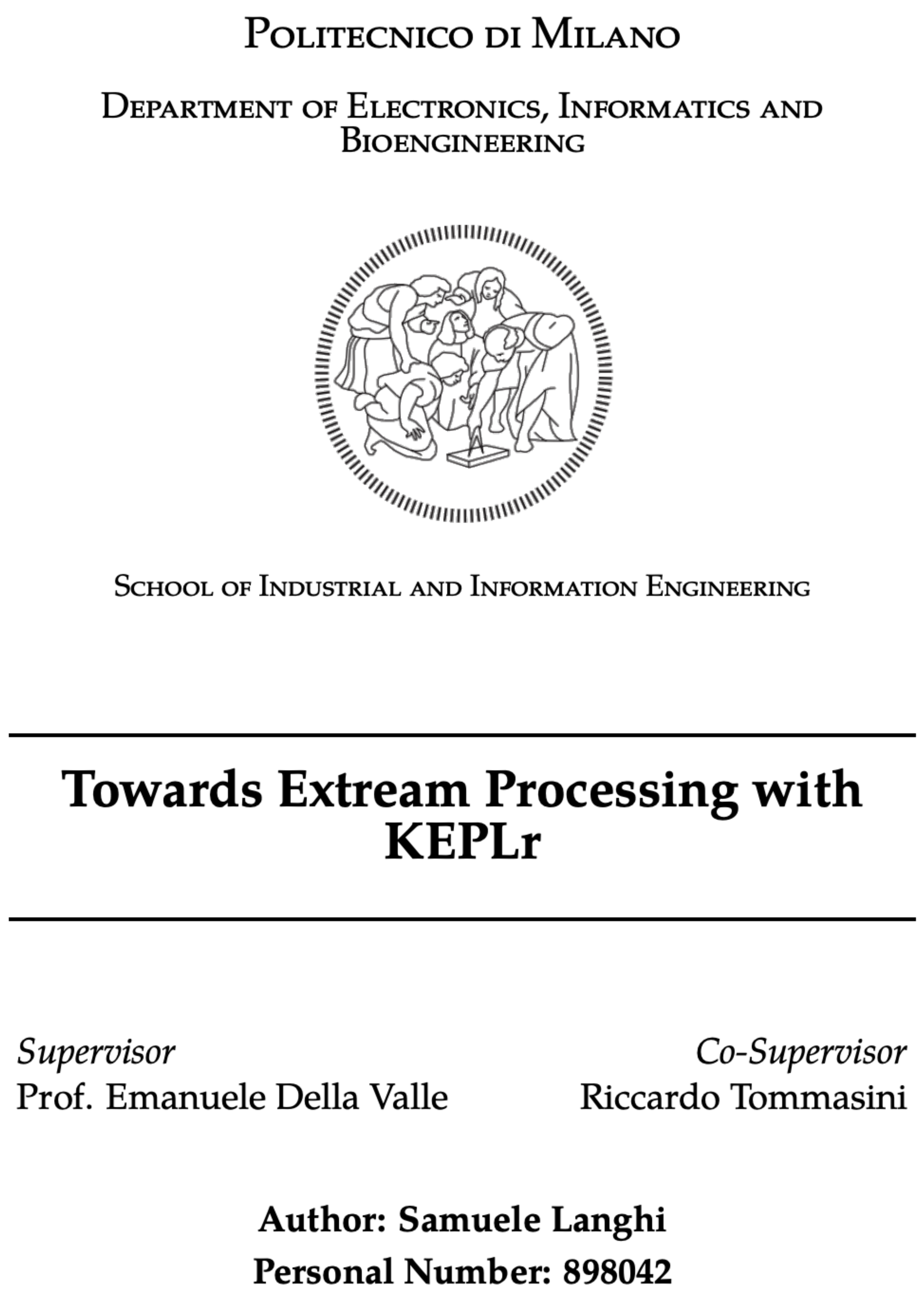
SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING

Towards Extream Processing with KEPLr

Supervisor
Prof. Emanuele Della Valle

Co-Supervisor
Riccardo Tommasini

Author: Samuele Langhi
Personal Number: 898042



Politecnico di Milano

Department of Electronics, Informatics and Bioengineering

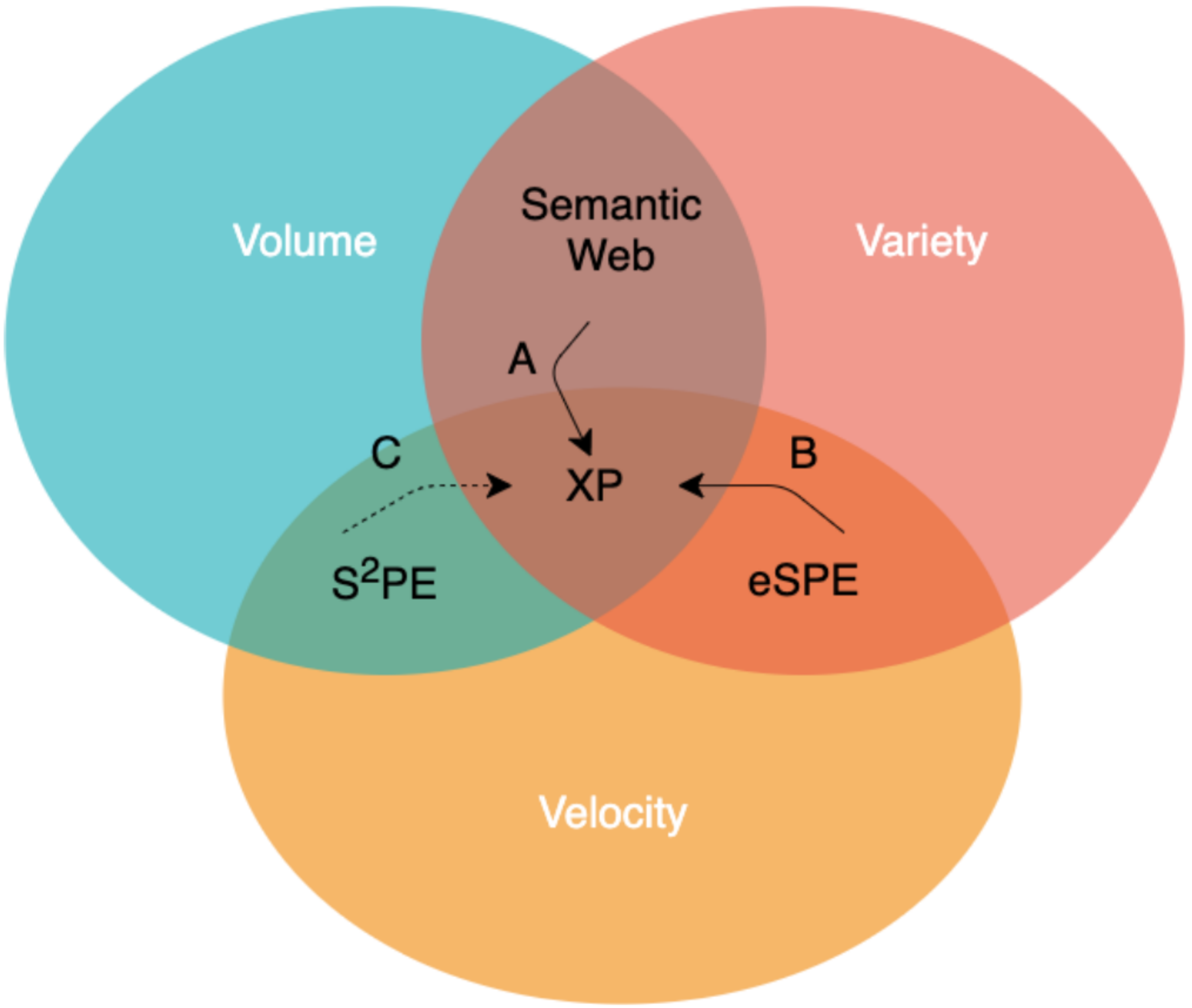
School of Industrial and Information Engineering

Towards Extream Processing with KEPLr

Supervisor
Prof. Emanuele Della Valle

Co-Supervisor
Riccardo Tommasini

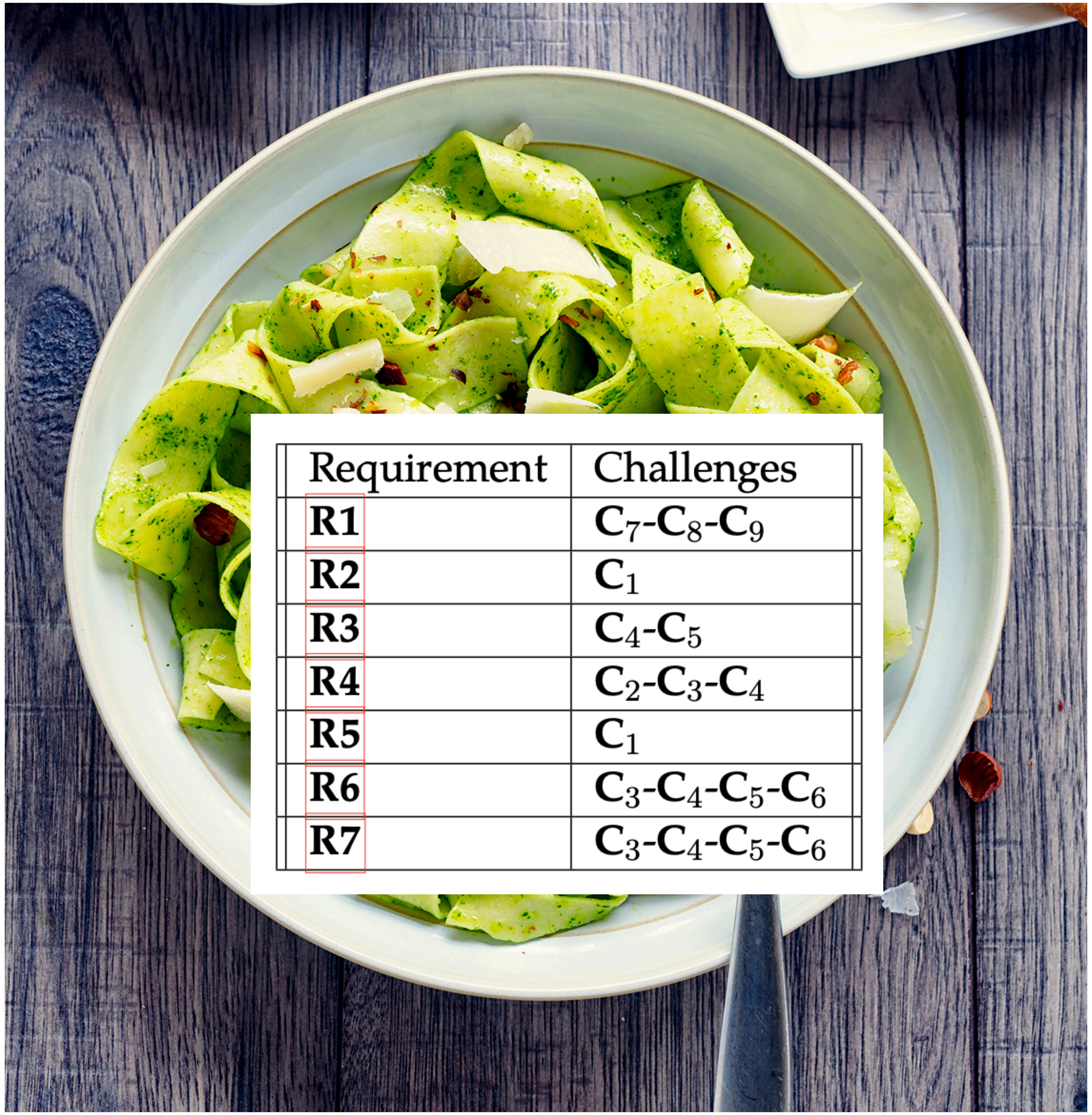
Author: Samuele Langhi
Personal Number: 898042



Macro level

(How) Is it possible to perform XP starting from BigSPEs?

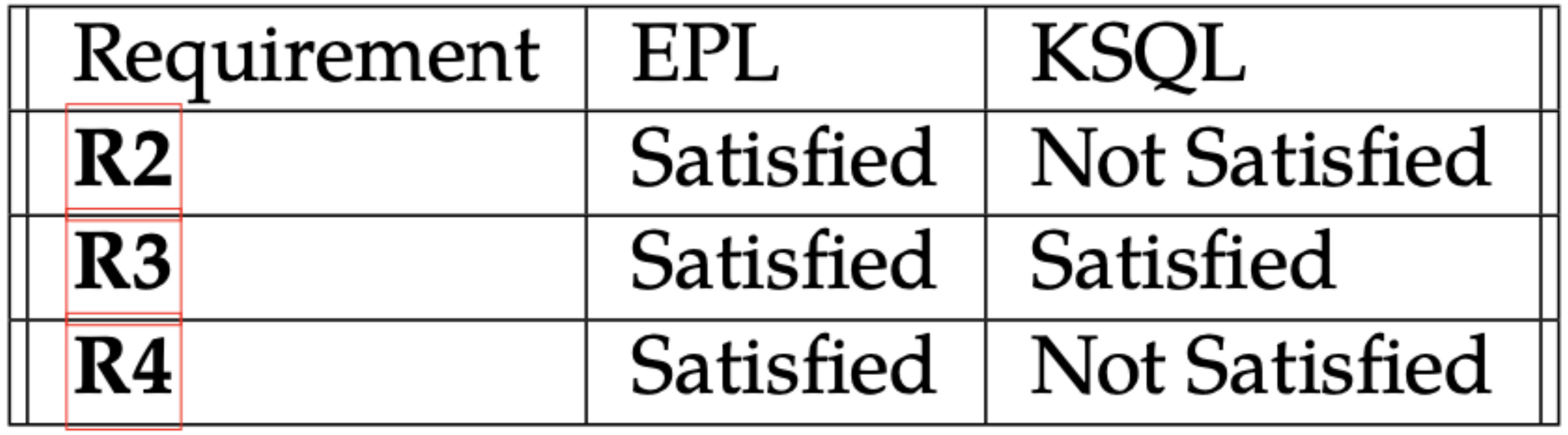
- Handle data unboundness
- Reactive processing
- Standard adoption
- Robustness
- Expressive Declarative Language
- Extended time model
- Fault tolerance
- Simple state management
- Layered data representation



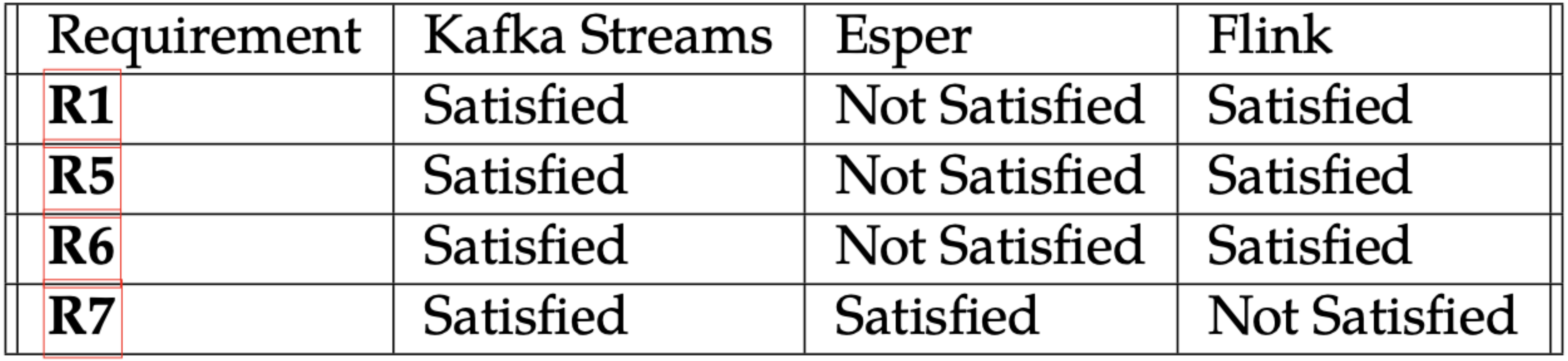
Requirement	Challenges
R1	C_7-C_8-C_9
R2	C_1
R3	C_4-C_5
R4	C_2-C_3-C_4
R5	C_1
R6	C_3-C_4-C_5-C_6
R7	C_3-C_4-C_5-C_6

Meso level

(How) Is it possible to perform CEP and DSMS operations starting from S2PE?



Requirement	EPL	KSQL
R2	Satisfied	Not Satisfied
R3	Satisfied	Satisfied
R4	Satisfied	Not Satisfied



Requirement	Kafka Streams	Esper	Flink
R1	Satisfied	Not Satisfied	Satisfied
R5	Satisfied	Not Satisfied	Satisfied
R6	Satisfied	Not Satisfied	Satisfied
R7	Satisfied	Satisfied	Not Satisfied

Micro level

How can we port EPL onto Kafka Streams?

↓

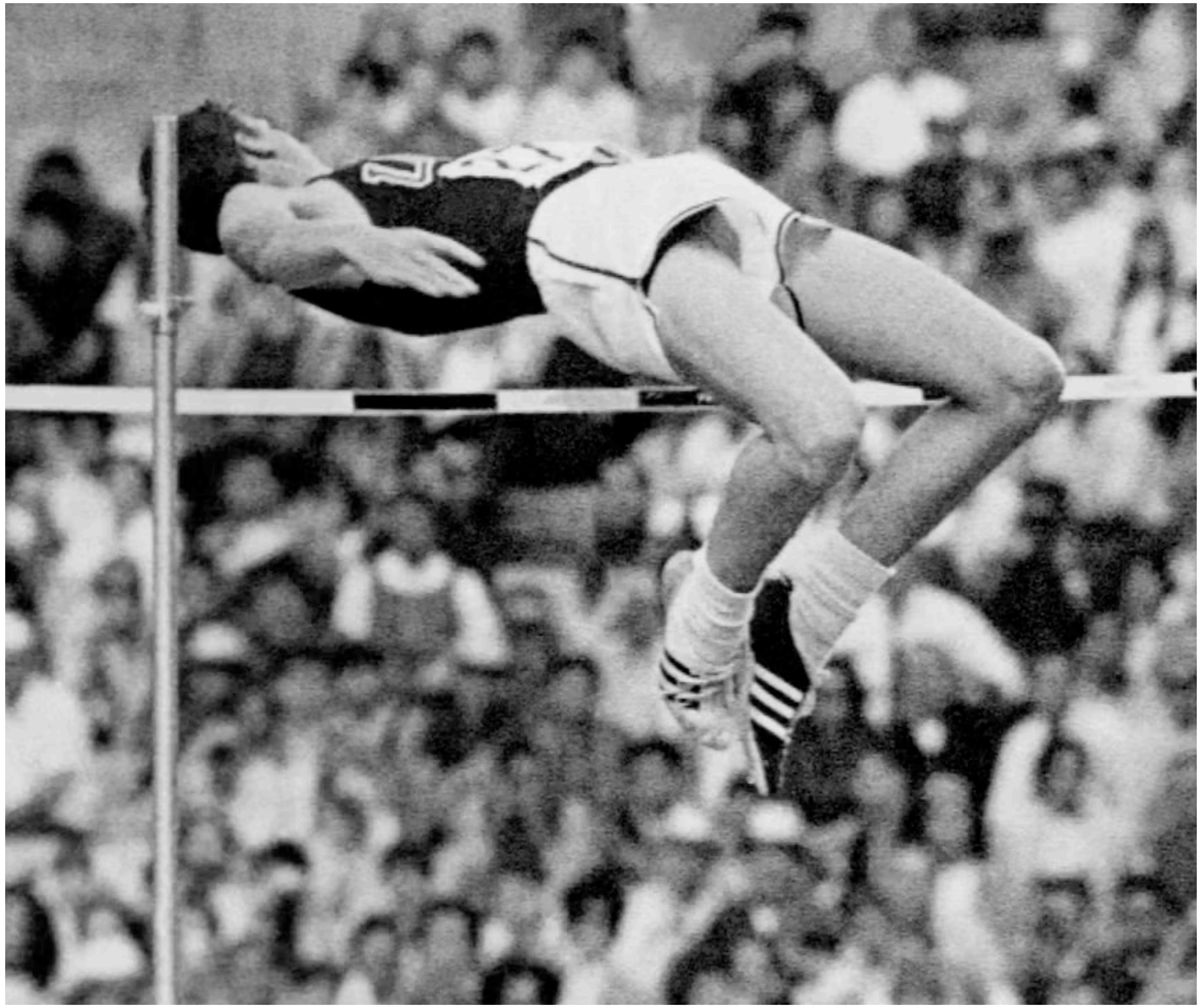
How can we enable extreme processing by porting EPL onto Kafka Streams?

What do you need to do?

- Come up with an idea (together)
- discuss it
- formulate it as a problem
- Implement it —> validation

The Role of Evaluation

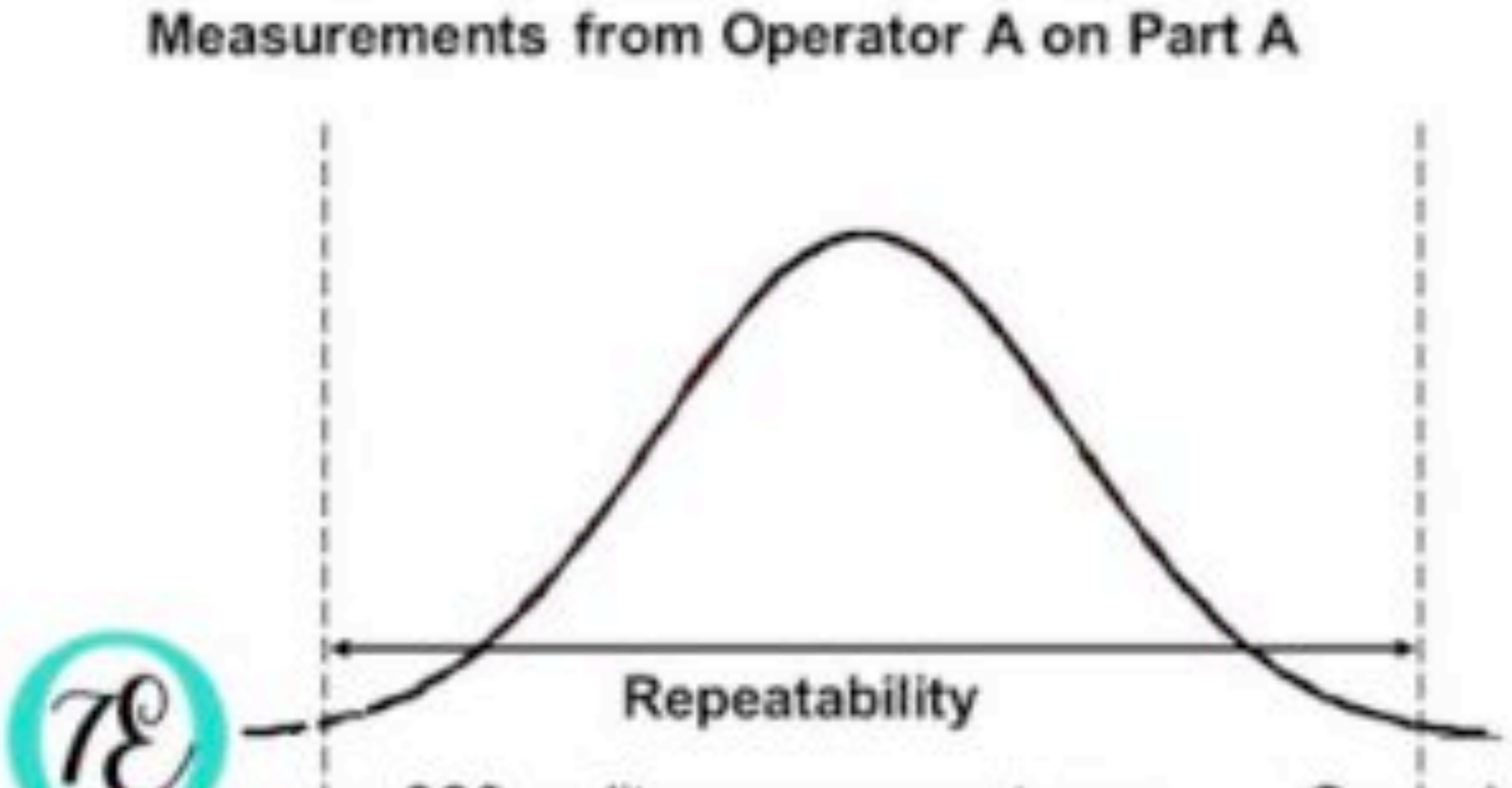
~~Assess the quality of the solution~~ Show the falsifiability of the theory



Repeatability

Same operator Same gage Same part

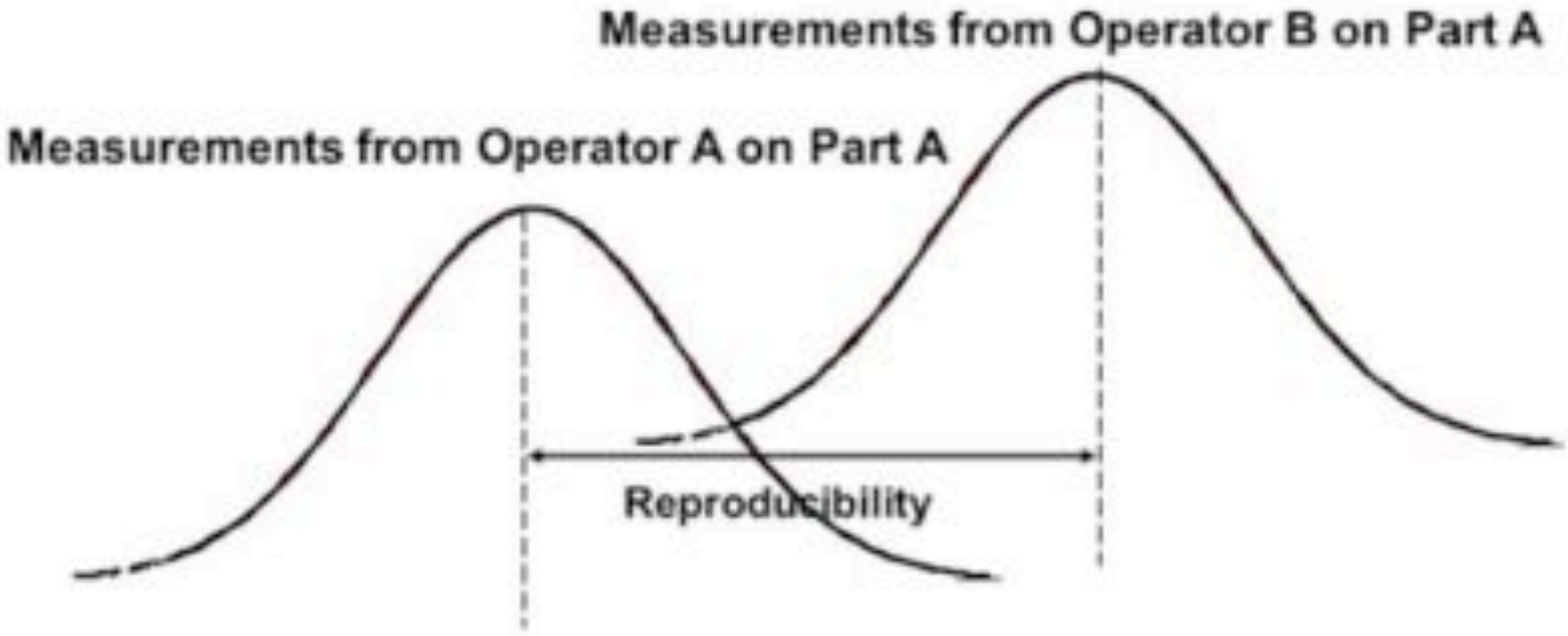
Can I measure the same thing more than once and get the same answer?



Reproducibility

Different operators Same gage Same part

Can I change the method of measurement, the observer, the location, the time (next day), and get the same answer?


Measurements from Operator B on Part A
Measurements from Operator A on Part A
Reproducibility

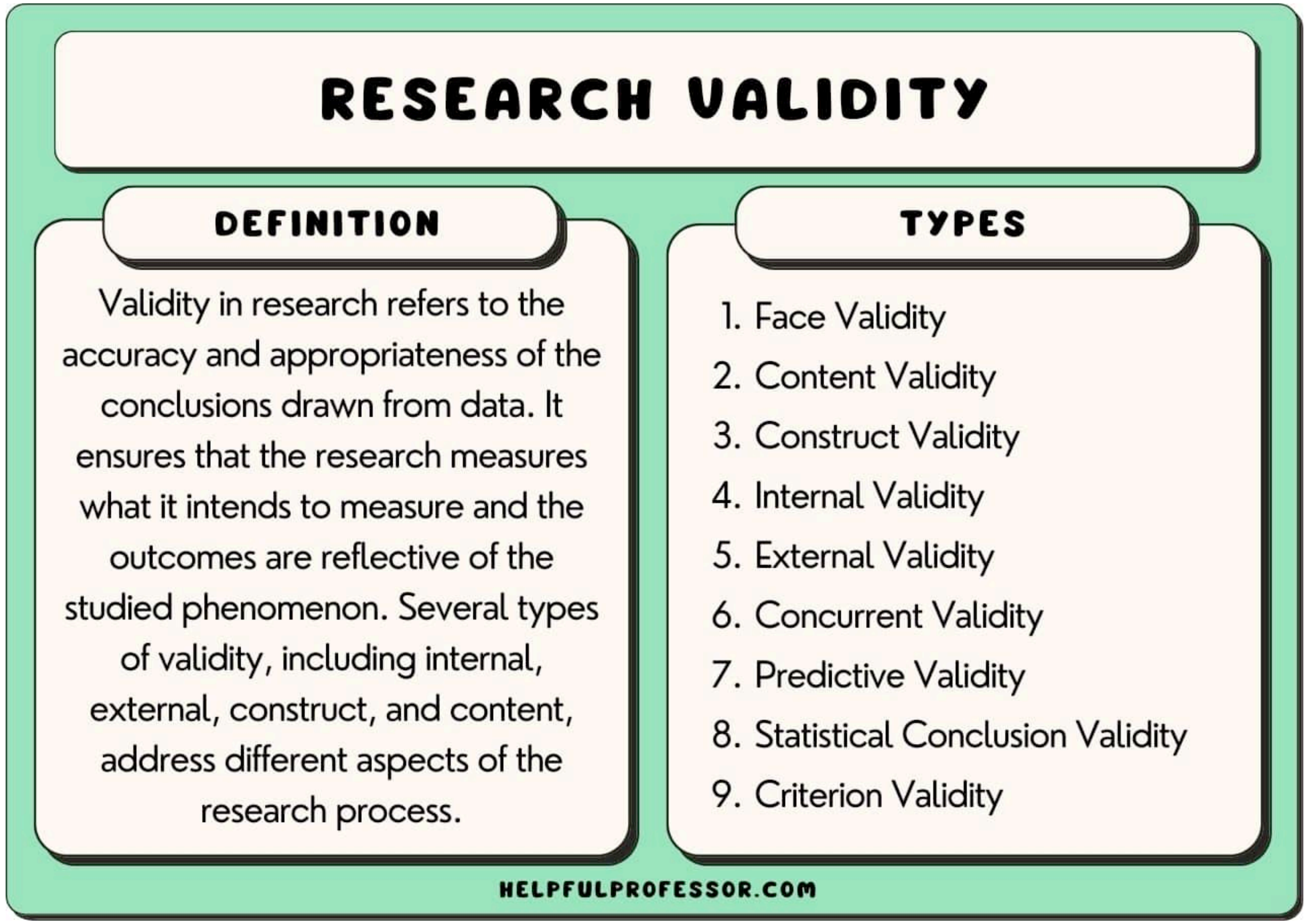
What do you need to do?

- take one of the paper we **understood** at the very beginning
- repeat (some part of) the evaluation changing experimental conditions
  - reimplement the code in a different language/systems
  - change datasets

The Role of Validation

Jump Example

- The minimimal viable solution that can addresses the problems statement
- comes BEFORE evaluation
- requires coding



RESEARCH VALIDITY

DEFINITION

Validity in research refers to the accuracy and appropriateness of the conclusions drawn from data. It ensures that the research measures what it intends to measure and the outcomes are reflective of the studied phenomenon. Several types of validity, including internal, external, construct, and content, address different aspects of the research process.

TYPES

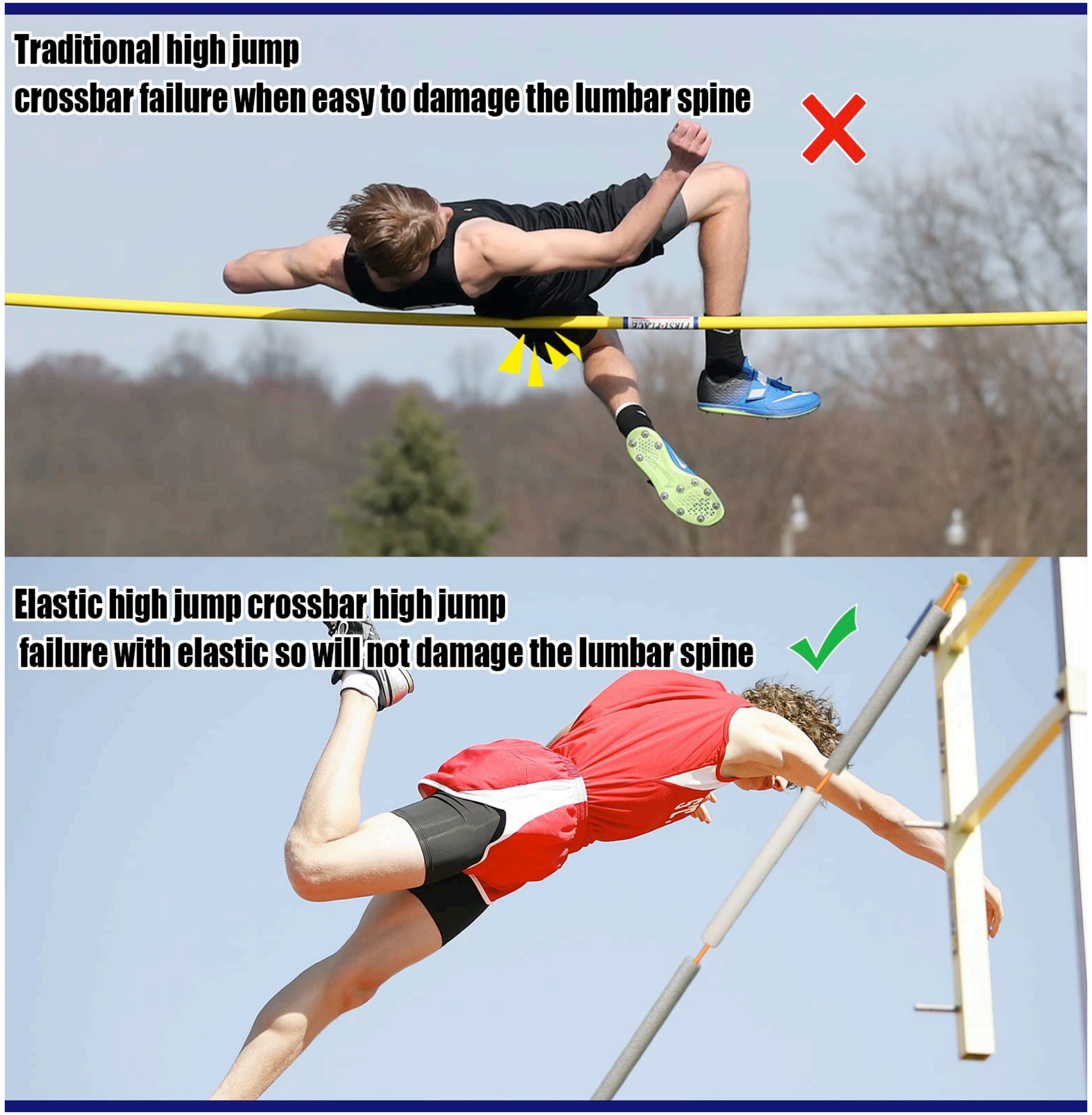
1. Face Validity
2. Content Validity
3. Construct Validity
4. Internal Validity
5. External Validity
6. Concurrent Validity
7. Predictive Validity
8. Statistical Conclusion Validity
9. Criterion Validity

HELPFULPROFESSOR.COM

The Role of Validation

Jump Example

- Jumpers must take off from one foot.
- A jump is considered a failure if
  - the jumper dislodges the bar or
  - touches the ground or
  - the jumper touches any object behind the bar before clearance.



Traditional high jump
crossbar failure when easy to damage the lumbar spine
✗
Elastic high jump crossbar high jump
failure with elastic so will not damage the lumbar spine
✓

What do you need to do?

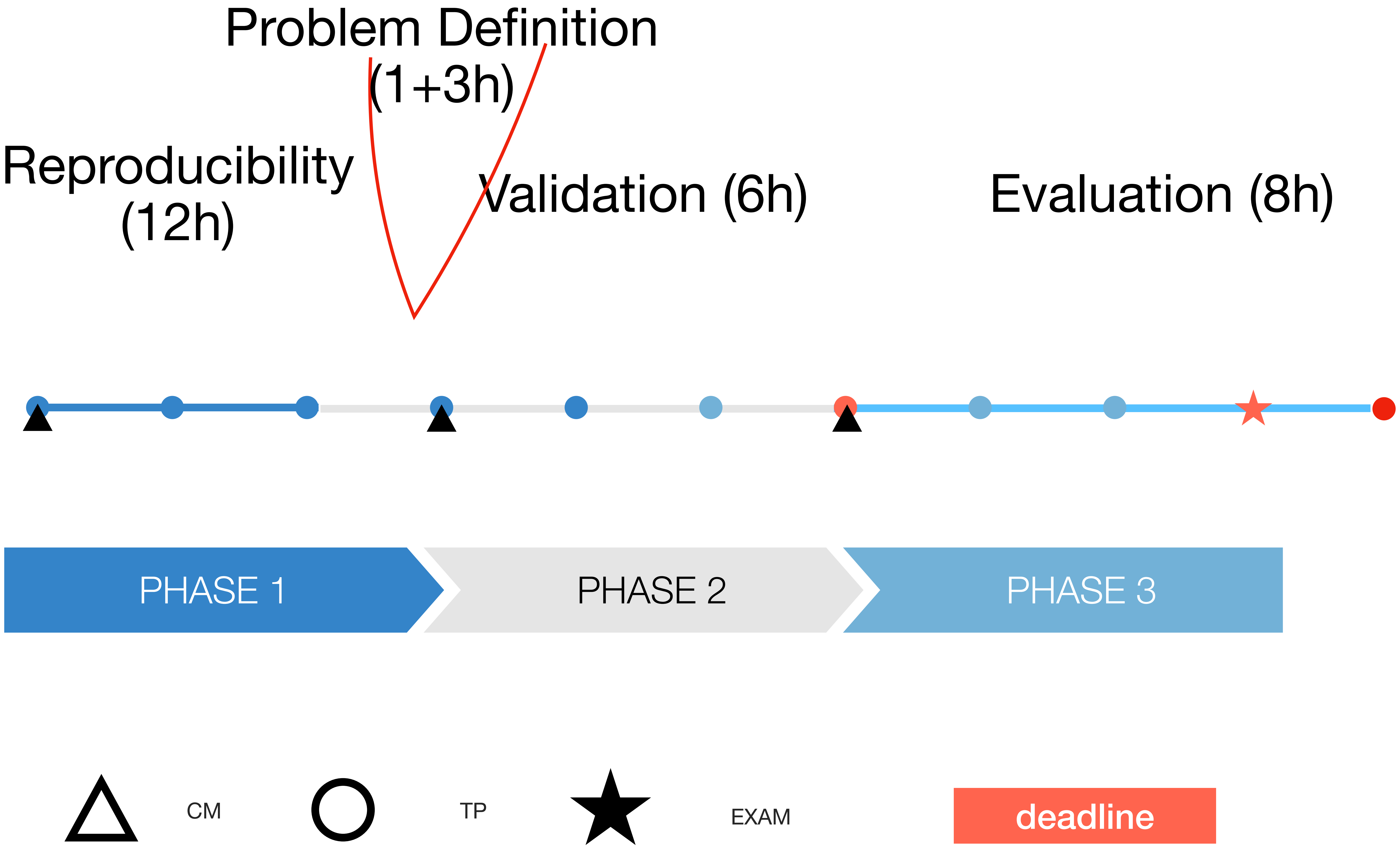
- Implement your idea **WITHIN** an existing data system
- quality its validity
  - which sometimes means reformulate the problem statement
  - which means prepare a **DEMO**

Course Schedule

Table 1

Class Topic	Practice/Theory	Milestone	Day	Date	When	Hours
Intro	📓		Monday	24 November 2025	11:00–13:00	2
Reproducibility	🖐🖐	Group Creation	Wednesday	26 November 2025	15:00–19:00	4
Reproducibility	🖐🖐		Monday	1 December 2025	15:00–19:00	4
Reproducibility	🖐🖐		Wednesday	3 December 2025	15:00–19:00	4
Problem Definition	📓+🖐	Using Cards	Monday	8 December 2025	15:00–19:00	4
Problem Submission	🏁	Slides	Friday	12 December 2025	00:00–23:59	0
Validation vs Evaluation	📓+🖐		Monday	5 January 2026	15:00–19:00	4
Prototyping (validation)	🖐🖐		Tuesday	6 January 2026	11:00–13:00	2
Prototyping (validation)	🖐🖐		Wednesday	7 January 2026	09:00–13:00	2
Prototyping (evaluation)	🖐🖐		Tuesday	13 January 2026	09:00–13:00	2
Prototyping (evaluation)	🖐🖐		Wednesday	14 January 2026	15:00–19:00	4
Presentaton/Posters	🎭		Wednesday	28 January 2026	15:00–19:00	4
Paper Submission	🏁		Tuesday	27 January 2026	00:00–23:59	0

Course Timeline (Three Phases)



Problem Definition
(1+3h)

Reproducibility
(12h)

Validation (6h)

Evaluation (8h)

PHASE 1

PHASE 2

PHASE 3

△ CM

○ TP

★ EXAM

deadline