COPPE
UFRJ

Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

Marta Mattoso

Federal University of Rio de Janeiro (UFRJ), Brazil

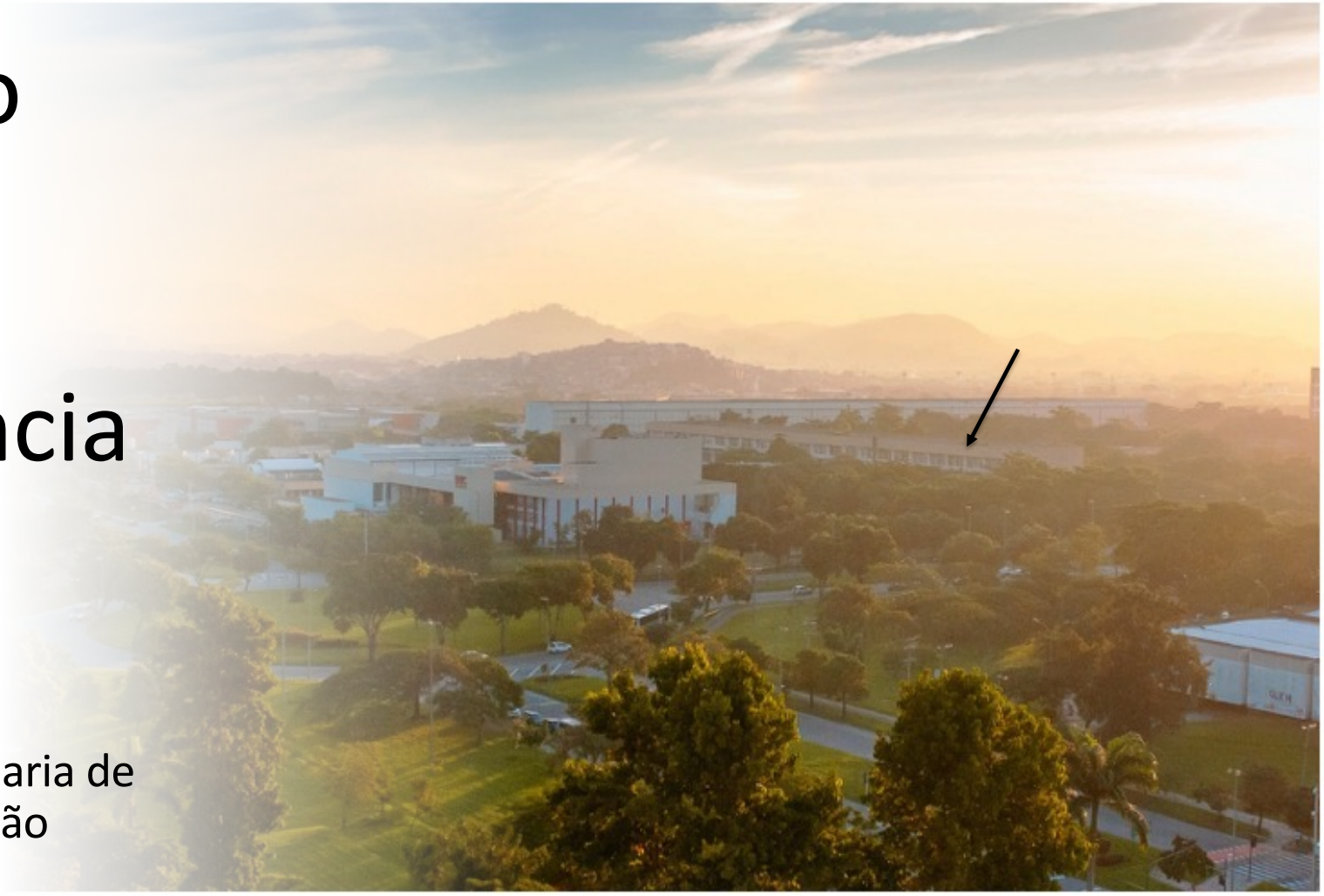
Introdução a ECI

"Ciência de Dados com transparência"

Um pouco sobre a minha proveniência

...

Professora Titular do
PESC/ COPPE/UFRJ
Programa de Engenharia de Sistemas e Computação



@foto: Ana Marina Coutinho (Coord Com UFRJ)

Um pouco sobre a minha proveniência …

- Orientação de alunos MSc e DSc no PESC/COPPE (>80)
- Bolsista de produtividade 1B do CNPq na Computação
- Coordenação de projetos financiados pelo CNPq, Faperj, Capes, Inria
- Membro de comitês no CNPq e Capes

Um pouco sobre a minha proveniência

- Editora associada de revistas: IEEE TBD, FGCS, PeerJCS, JIDM
- Mercator Fellow do DFG, Alemanha, projeto FoNDA
- Membro do corpo de especialistas do projeto WorkflowsRI nos E.U.A
- Trabalho em equipe voltado a problemas reais

HUB.RIO - CENTER OF EXCELLENCE IN DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE OF THE STATE OF RIO DE JANEIRO

GOAL:

"Integrate the competence and technical capacity of the scientific and business community of Rio de Janeiro and its partners to produce innovation involving digital transformation and artificial intelligence techniques to meet the challenges of the digital revolution."

Information about the Hub.Rio initiative:
Prof. Guilherme Horta Travassos
Systems and Computer Engineering Program
(+5521) 3938 8712 or ght@cos.ufrj.br

55 anos antecipando o futuro

Magritte – La Trahison des Images ; Centre Pompidou 2016

Por que Transparência?

ACESSO À INFORMAÇÃO

ENCONTRE NA LEI

DADOS ABERTOS

PERGUNTAS FREQUENTES

SERVIDOR MUNICIPAL

ESTRUTURA

CONSELHOS MUNICIPAIS

CONSULTA A PROCESSOS

CONTAS RIO

LICITAÇÃO

PAINEL DAS ORGANIZAÇÕES SOCIAIS

SISTEMA DE REGULAÇÃO SISREG

PLANOS MUNICIPAIS

OUTROS DADOS E DOCUMENTOS

CARTA DE SERVIÇOS

CONTRATOS

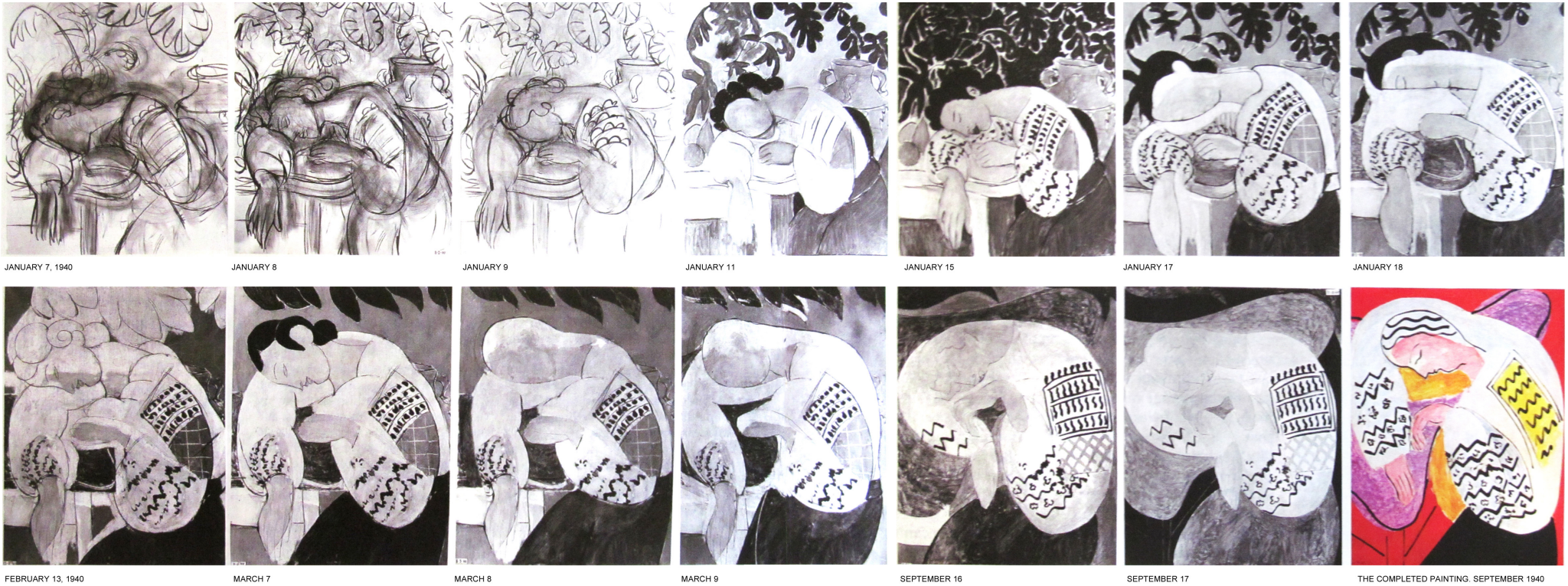
INFORMAÇÕES SOBRE O CORONAVÍRUS

COMBATE AO CORONAVÍRUS: GASTOS PÚBLICOS

PAINEL DA LAI

Metadata:
Title: “Le Rêve”
Painter: Henri Matisse
Year: 1940
Source: Matisses’ archive

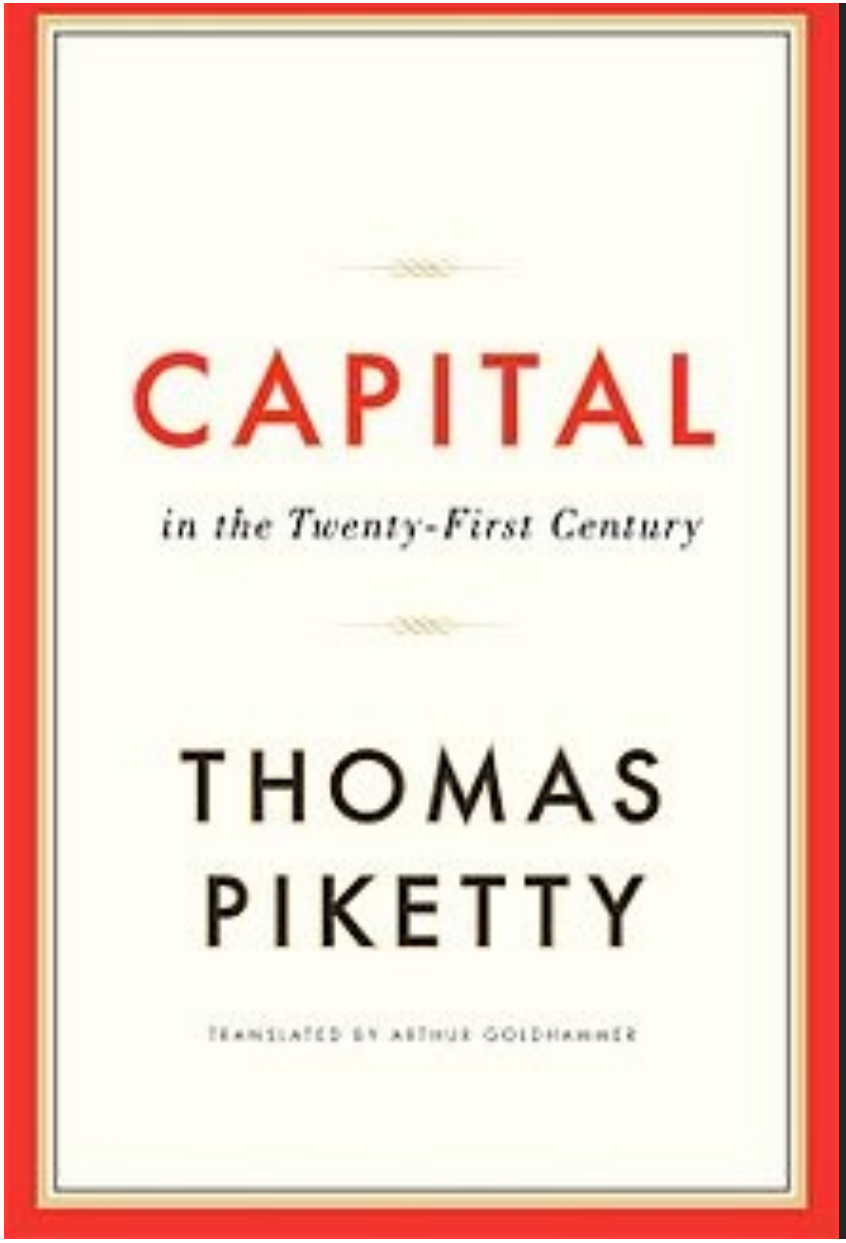
What is the provenance of the process?



Stage photos from “Le Rêve” - Nice, Hotel Régina - Source: Archive Henri Matisse ; Expo: “Pairs et Series” Centre George Pompidou, Paris

Metadata
Title: “The Dream”
Painter: Henri Matisse
Year: 1940
Source: Matisses’ archive

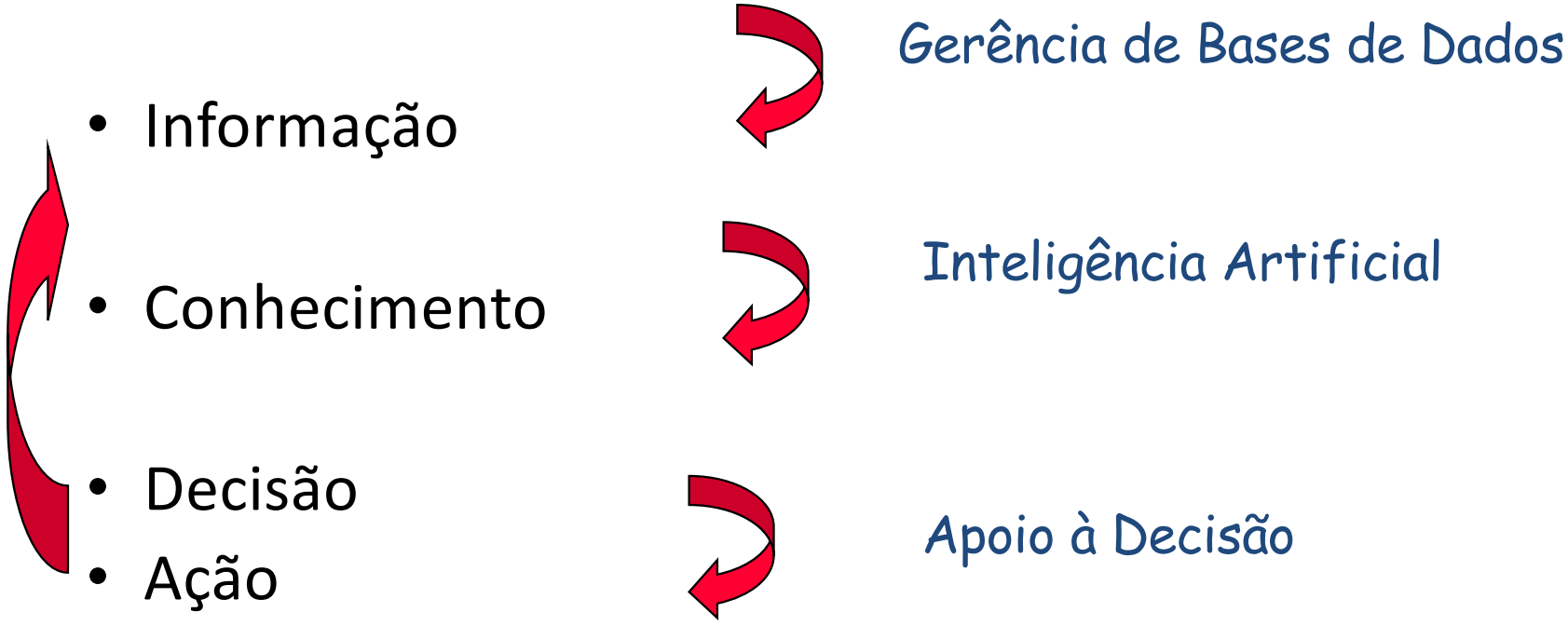
Provenance of the process



You may or not agree, but has “provenance” (source data and transformations are available)

Transformações em Ciência de Dados

- Dado
- Informação
- Conhecimento
- Decisão
- Ação

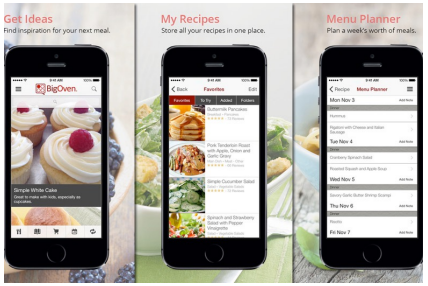


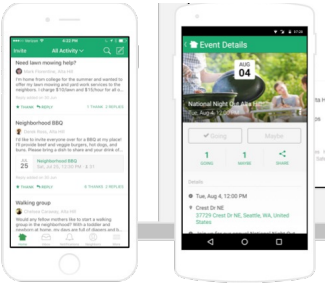
Transparência em Ciência de Dados

QUALIDADE

CONFIANÇA

EXPLICAÇÃO





How to trust these recommendations ?

- Felix Naumann, “Reflexions of a data skeptic” – ACM SIGMOD blog 2016

Aprendizado

- Entrada: x (atributos medidos do cliente)
- Saida: y (bom/mal cliente?)
- Target funtion: f : X → Y (ideal - desconhecida)
- Dados: (x1,y1),(x2,y2),··· ,(xN,yN) (histórico)

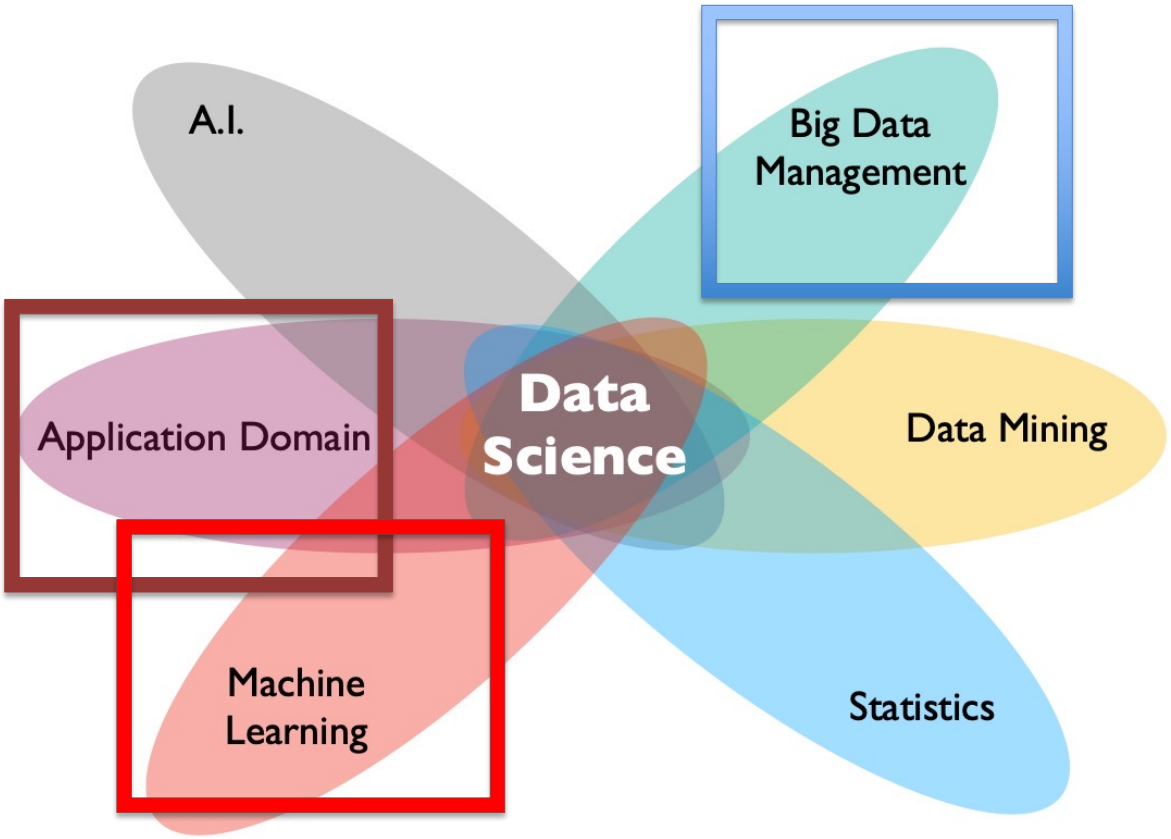
↓ ↓ ↓

- Hypothese: g : X → Y (formula que será usada)

Data Science ≠ ML

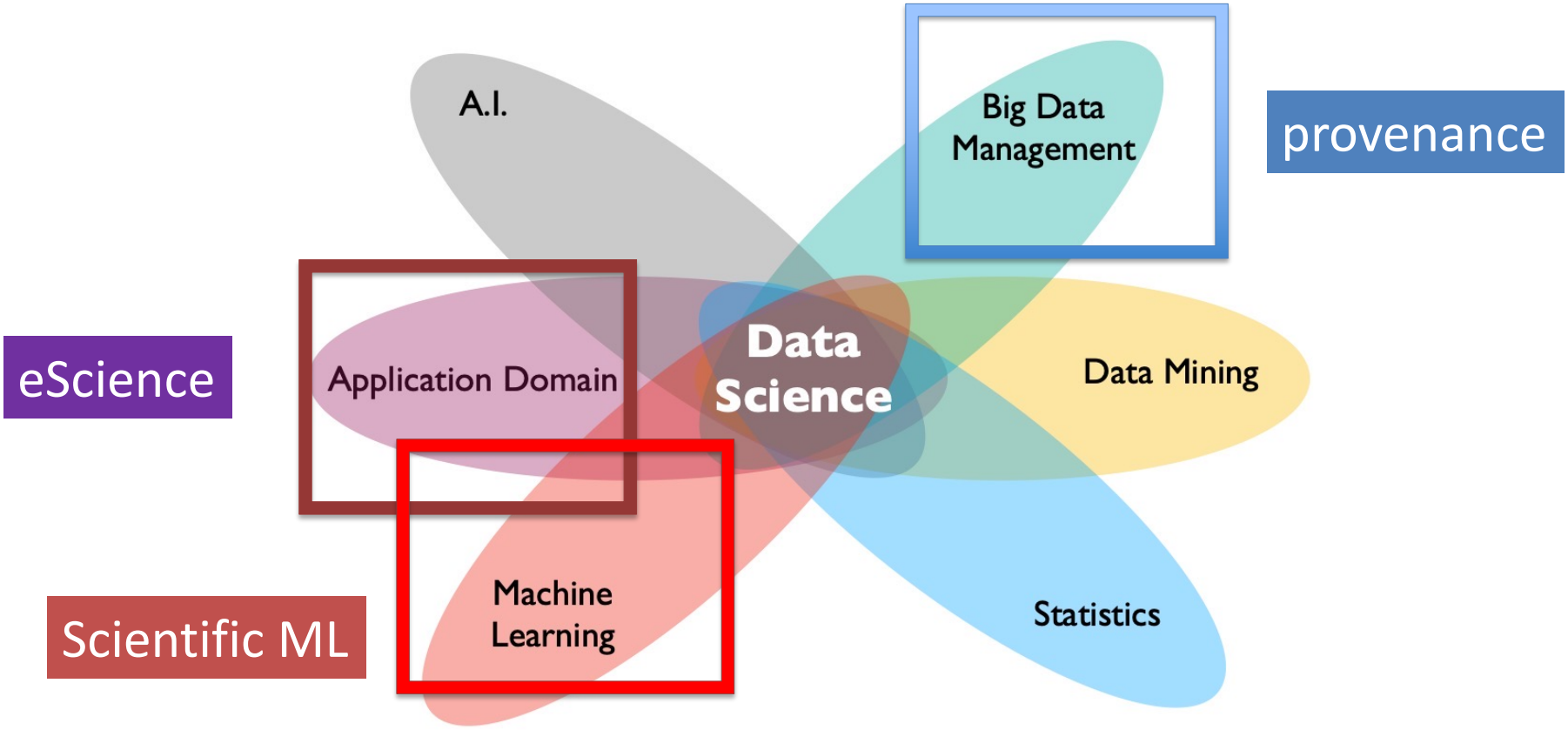
[Jens Dittrich 2018] - http://www.youtube.com/user/jensdit

Data Science ≠ Machine Learning [Jens Dittrich]



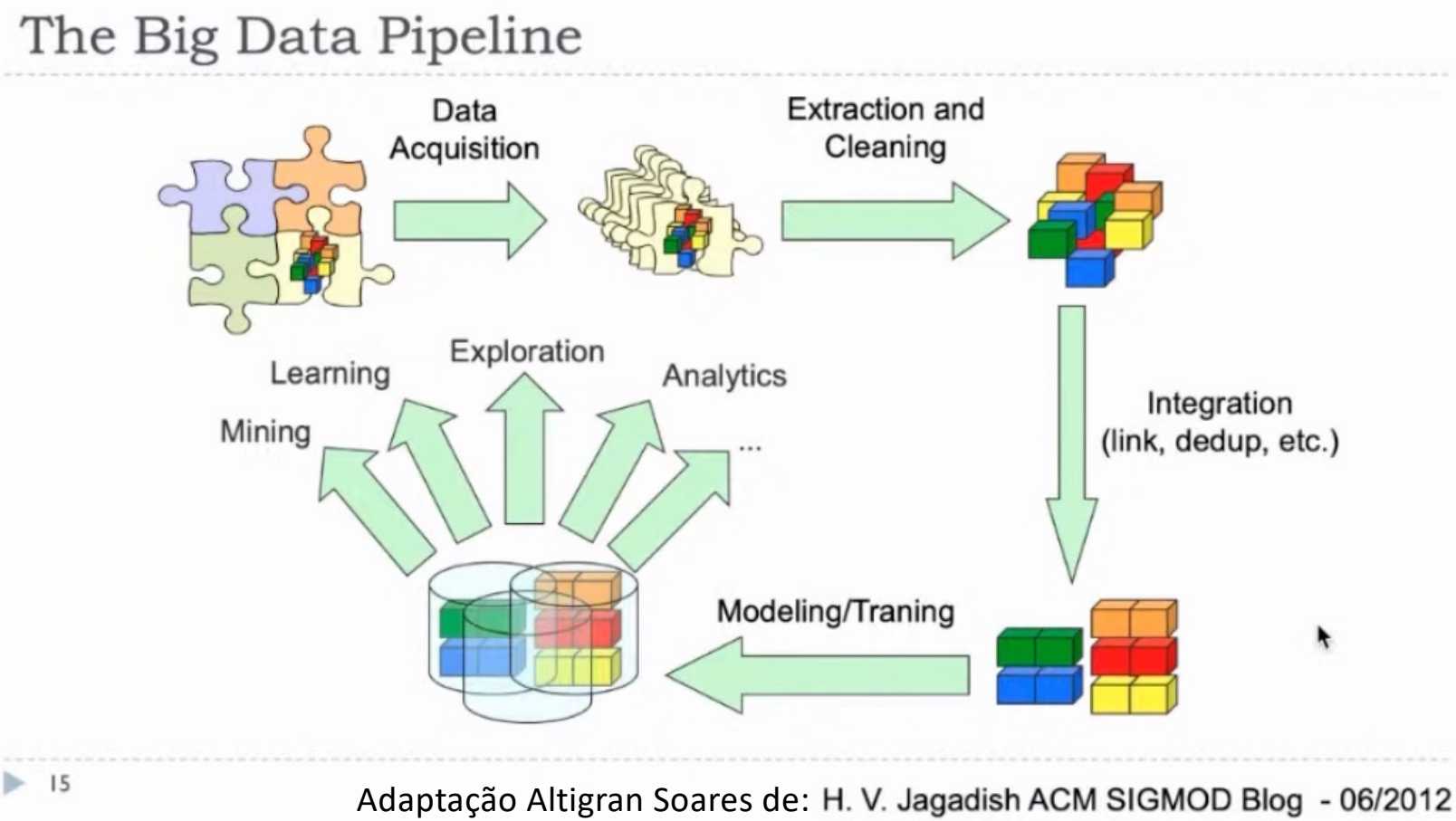
"Data Science ≠ Machine Learning: Some Thoughts on the Role of Data Management in the new AI-Tsunami" – Jens Dittrich Keynote DEEM@SIGMOD 2018, (DEEM: Workshop on Data Management for End-to-End Machine Learning)

Data Science ≠ Machine Learning and Provenance helps on this integration



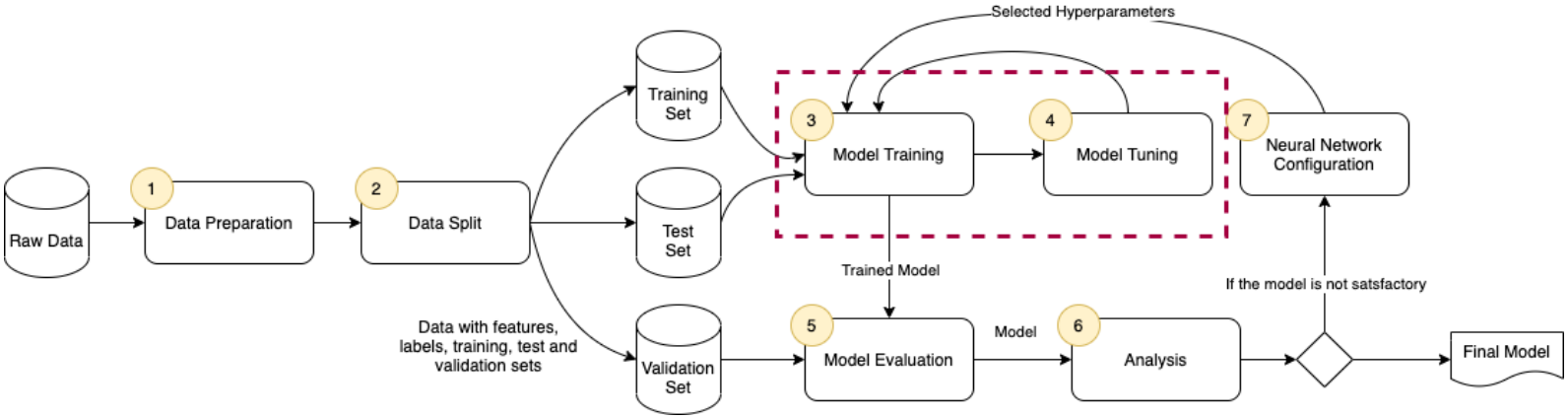
Jens Dittrich "Data Science ≠ Machine Learning: Some Thoughts on the Role of Data Management in the new AI-Tsunami" -- Keynote DEEM@SIGMOD 2018, June 2018

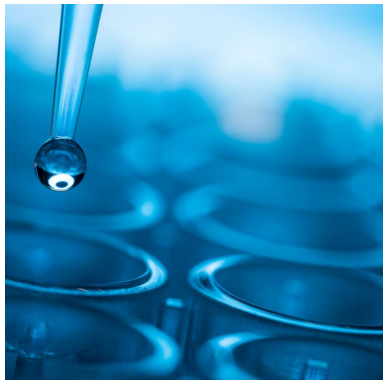
Palestra Prof. Altigran Soares (UFAM): https://www.youtube.com/watch?v=N43528a23xo&t=1279s



Adaptação Altigran Soares de: H. V. Jagadish ACM SIGMOD Blog - 06/2012

Deep Learning workflow





Final Model: trust? Explain?

Provenance Supporting Hyperparameter Analysis in Deep Neural Networks (*Débora Pina, Liliane Kunstmann, Daniel de Oliveira, Patrick Valduriez, Marta Mattoso*) **IPAW 2021**

Aprendizado de Máquinas

- ✓ É preciso **entender** o que se está fazendo, simplesmente apertar o botão do software não dá certo.
- ✓ É preciso relacionar com **conceitos estatísticos**.
- ✓ É preciso entender e **pré-tratar os dados**.
- ✓ É preciso escolher o **caminho mais simples** que resolva o problema e **não o caminho da moda**.
- ✓ É preciso partir **do problema ------> para a técnica** e nunca o inverso.

Transparency?

- The recipe for cooking a dish may be the same
- Recipe instructions execution (n variations)
  - How long took the first step ?
  - What was used to guarantee slow cooking ?
  - What substitutions were made ?
  - Were all steps followed?

Recipe is Transparency?

The recipe for cooking a dish may be the same

Recipe instructions execution (n variations)

- Ingredients
- Recipe
- Result

is not enough to answer provenance queries

Provenance queries:

- How long took the first step ?
- What was used to guarantee slow cooking ?
- What substitutions were made ?
- Were all steps followed?

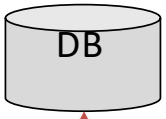
Provenance Data

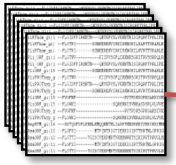
The recipe for cooking a dish may be the same

Recipe instructions execution (n variations)

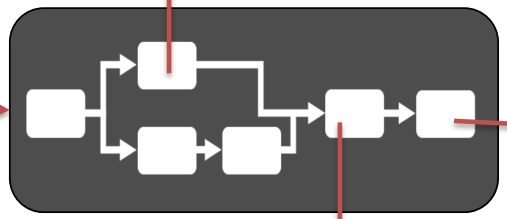
- The trace that registers the execution is the provenance
- The intermediate association between ingredients, their transformations and the final result is the provenance
- Choosing the best variation benefits from provenance data analysis
- The best variation decision is supported by provenance data record

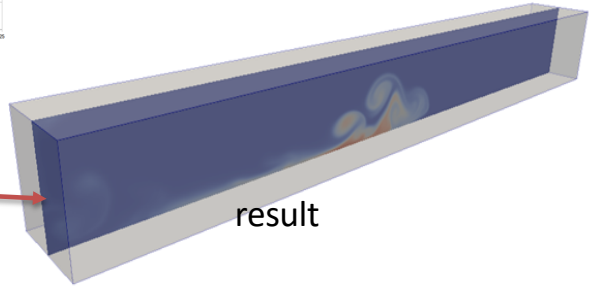
Scientific Experiment: agents, entities, activities



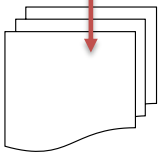


input data



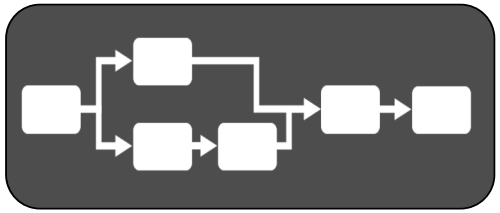


result



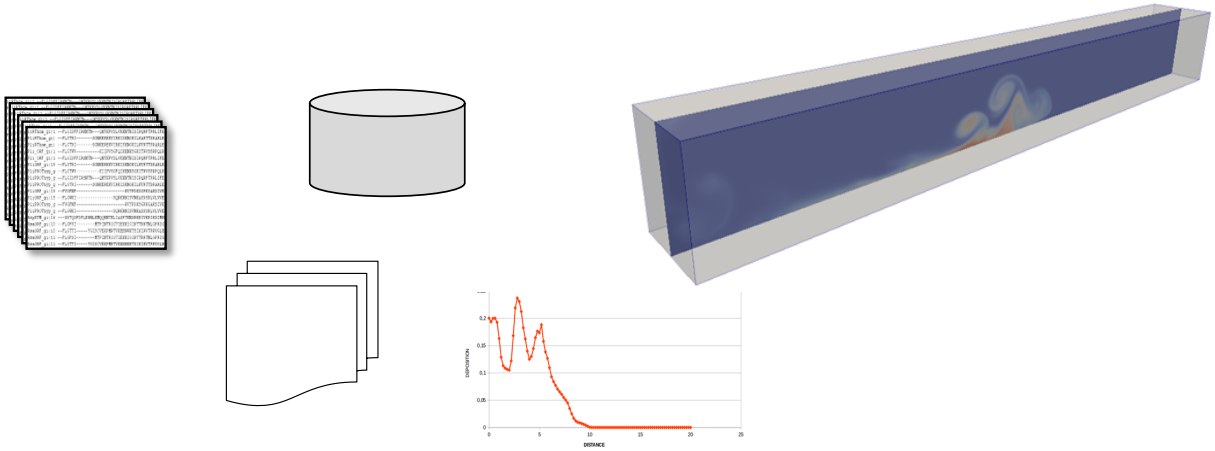
Intermediate data

Scientific Experiment: activities



- Processes
- Data Transformations
- Scripts
- Workflows

Scientific Experiment: entities

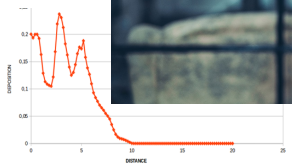


- Data
- Files
- Databases
- Documents
- Images
- Logs

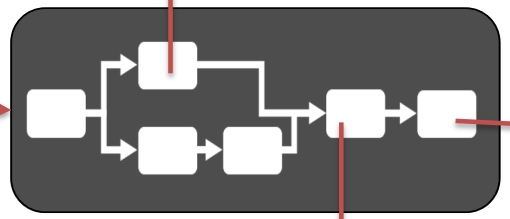
Scientific Experiment: agents

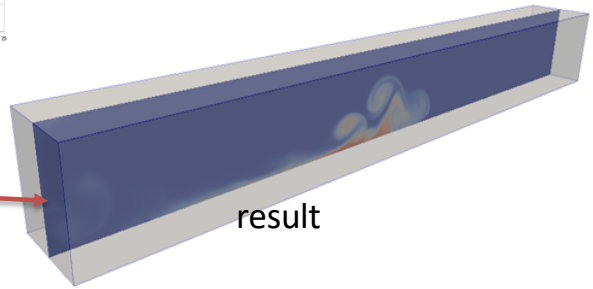
- People
- Teams
- Organizations

Scientific Experiment: agents, entities, activities

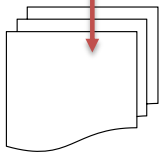


input data





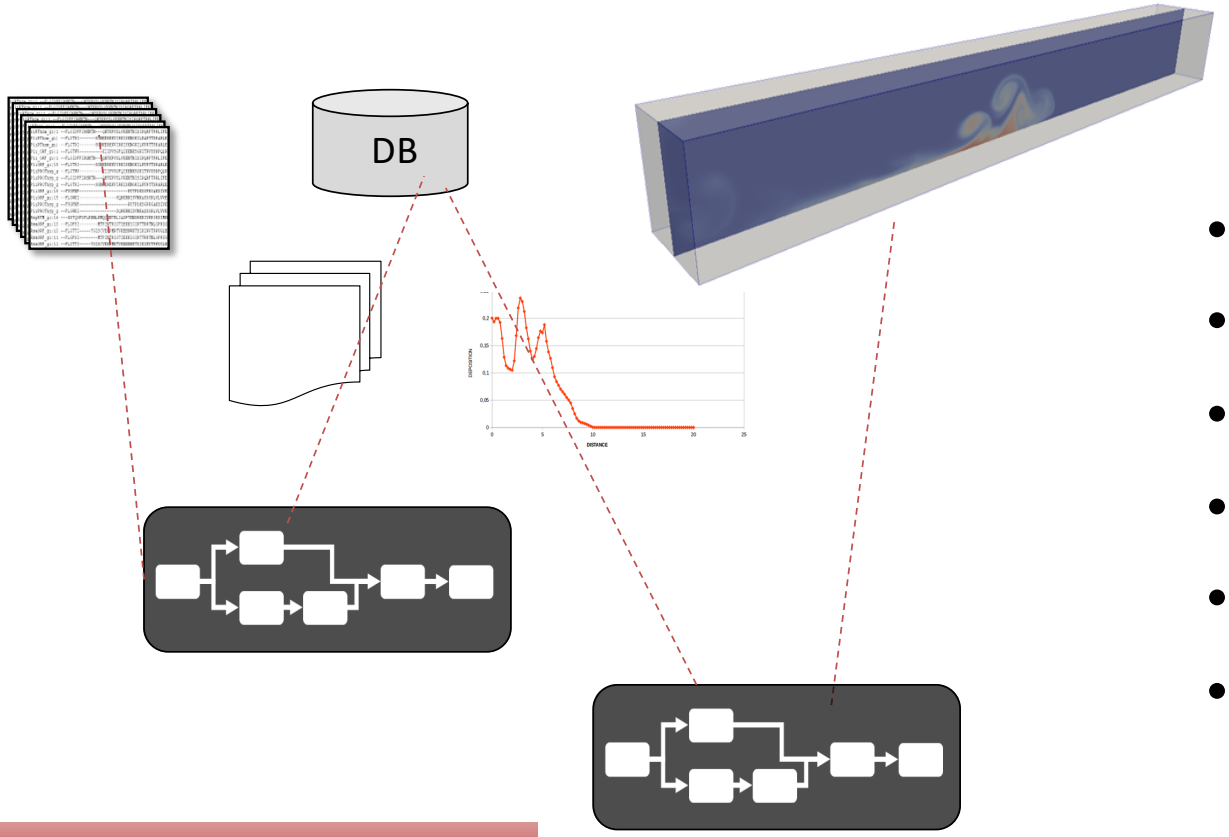
result



Intermediate data

Implicit relationship between objects

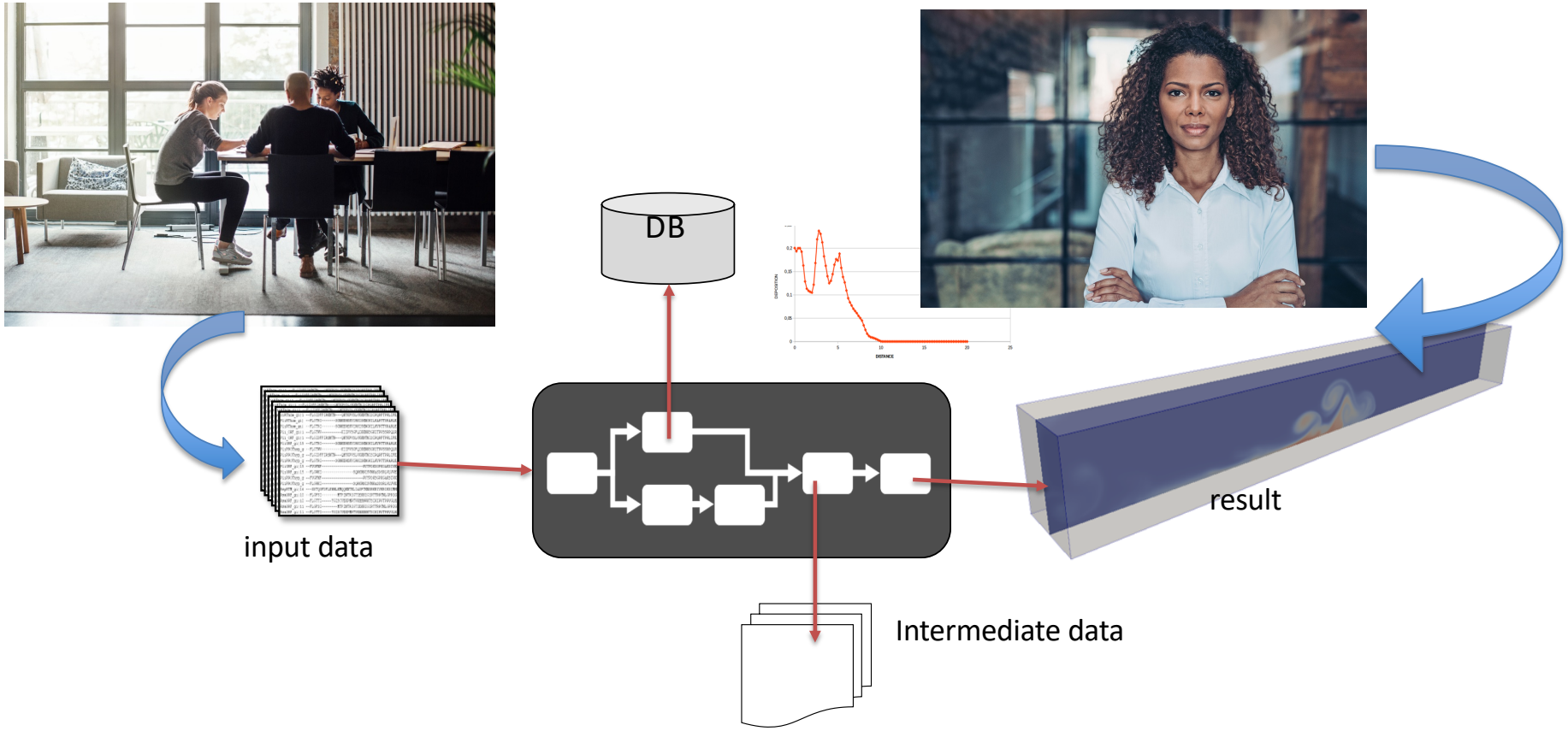
Data Science / management ignore activities



- Data cleaning
- Machine Learning
- KDD
- DBMS, key-value, text
- Data Lakes
- Polystores

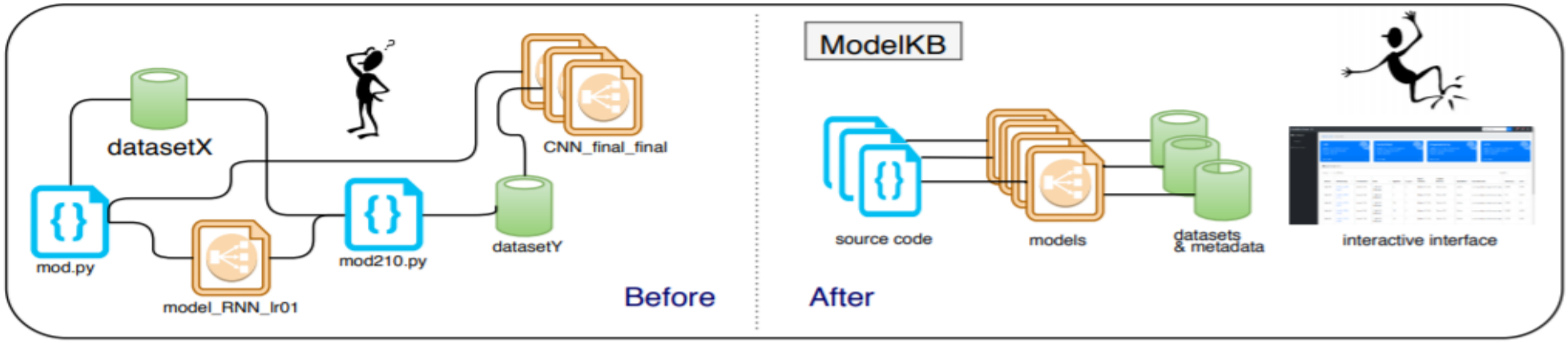
relationships are lost

Provenance: traces agents, entities, activities



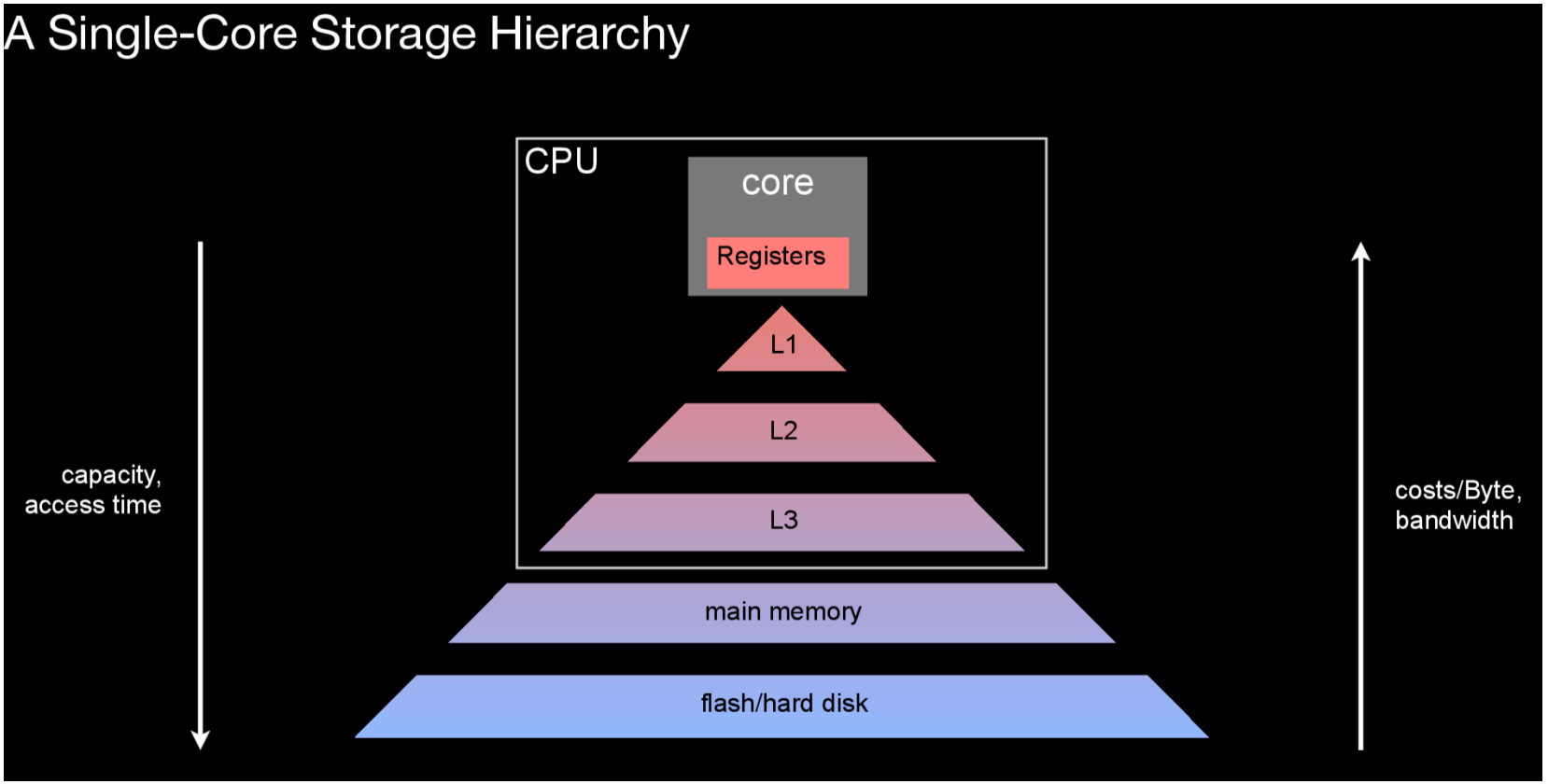
Automated Management of Deep Learning Experiments

Gharib Gharibi, Vijay Walunj, Rakan Alanazi, Sirisha Rella, Yugyung Lee
ggk89@mail.umkc.edu
University of Missouri-Kansas City
Kansas City, MO

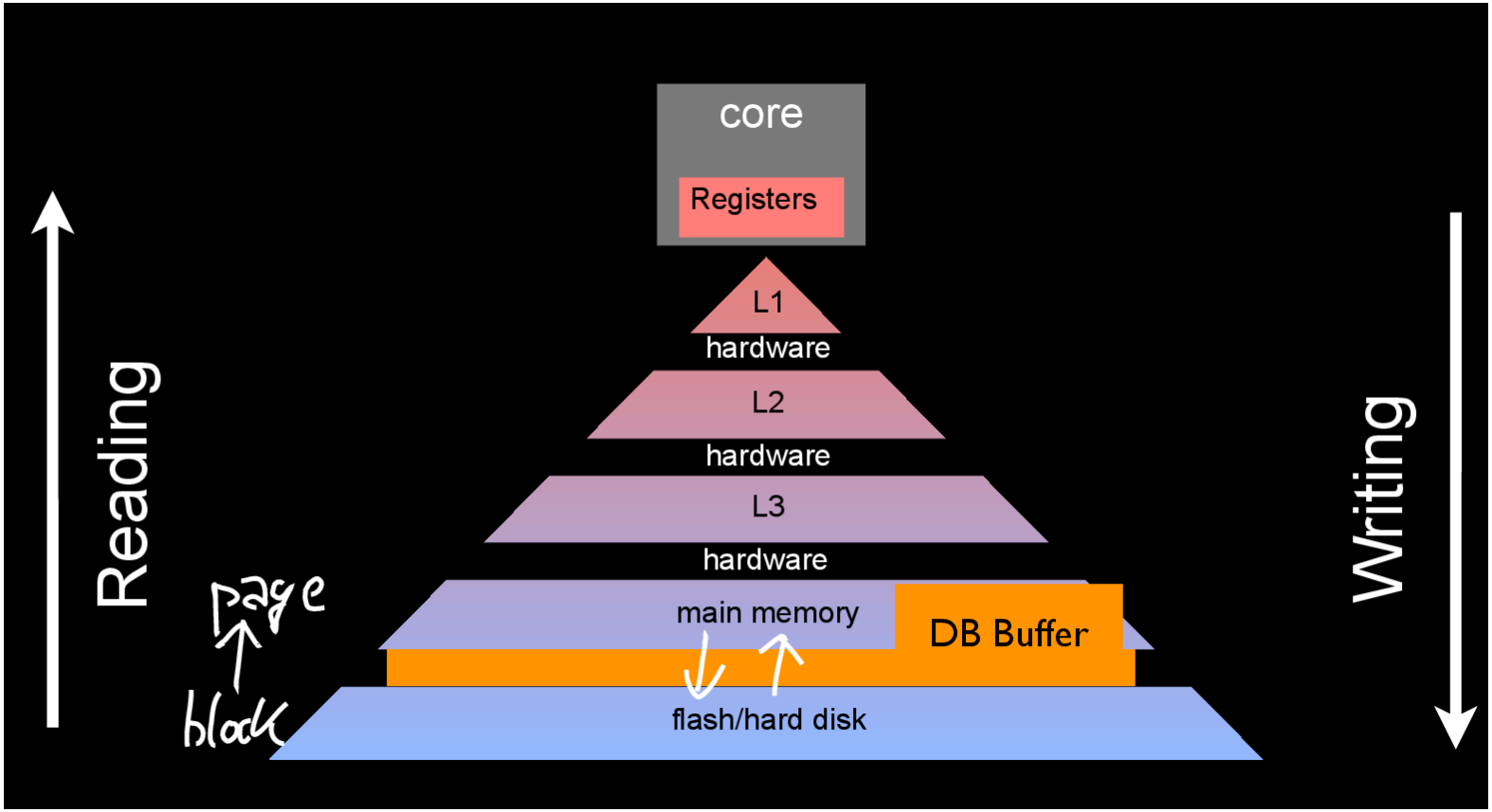


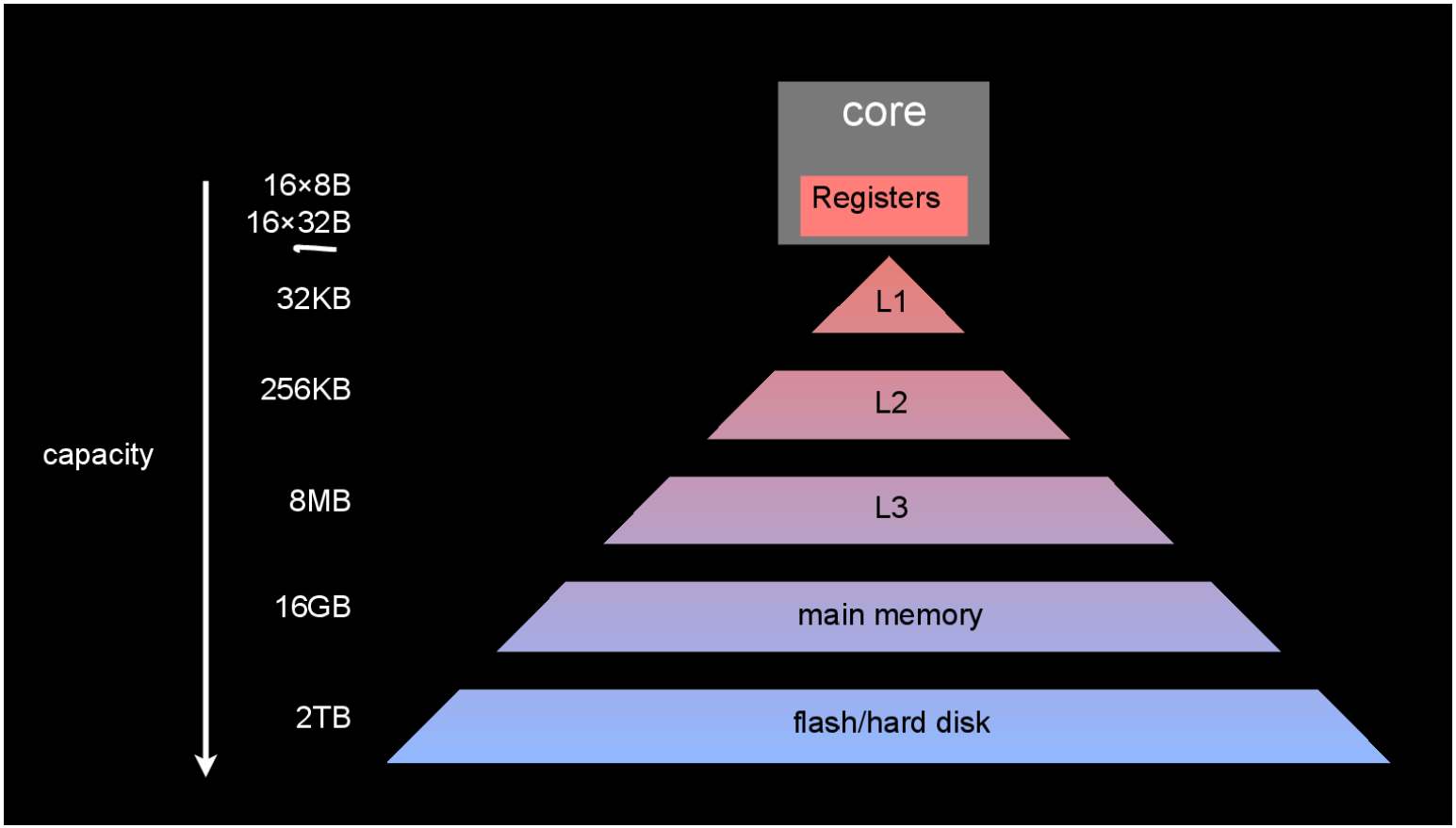
Por que fazer essa organização de dados não é trivial?

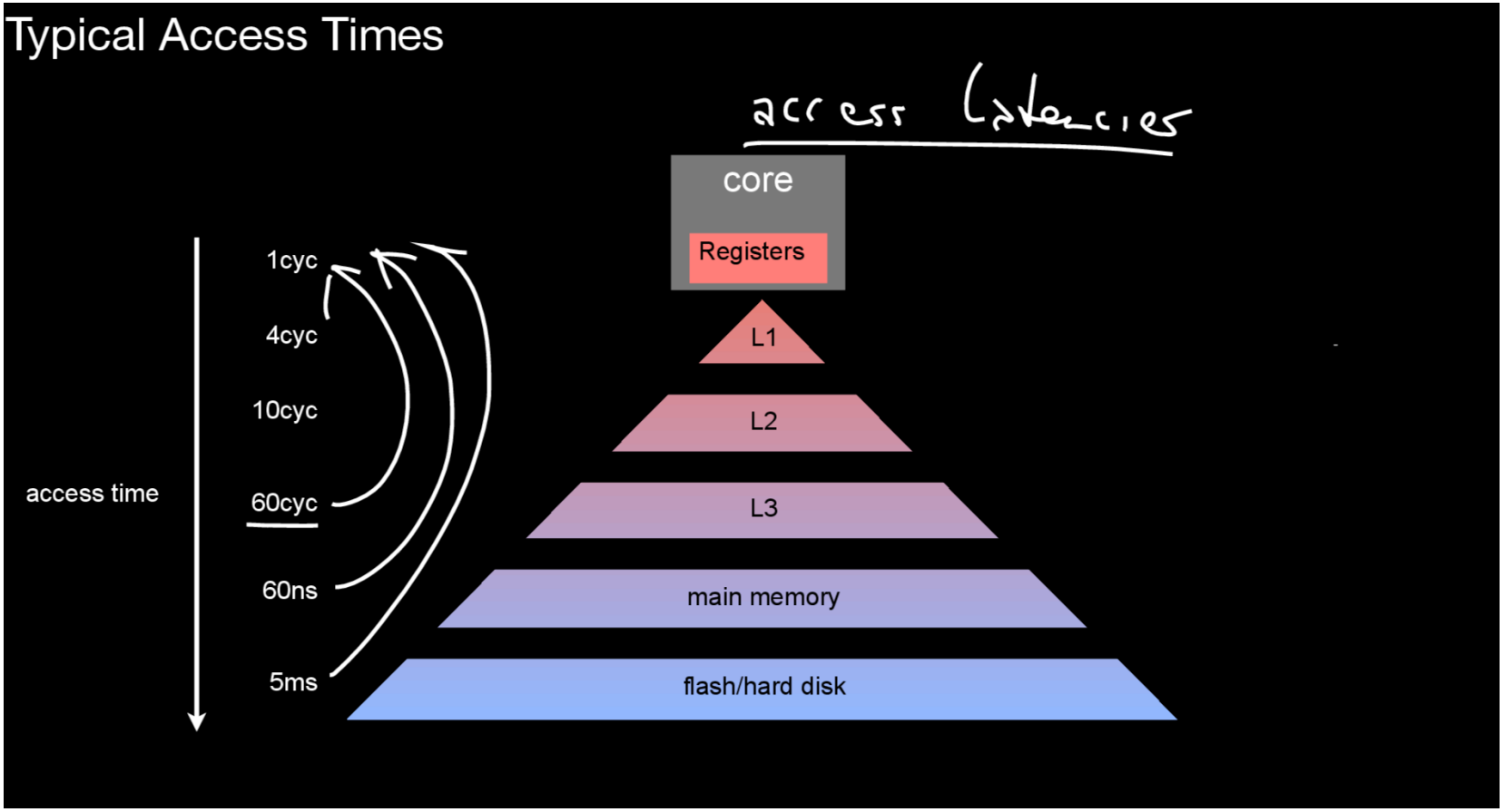
Hierarquia de memória



DB Buffer




core
Registers
16×8B
16×32B
32KB
256KB
8MB
16GB
2TB
capacity
L1
L2
L3
main memory
flash/hard disk



Relative Distances!

Factor 45

Factor 15

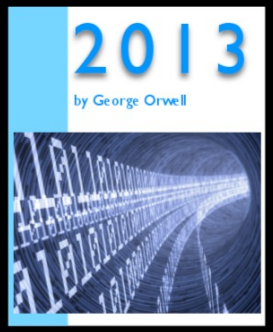
Factor 2.5

"L1 cache is like grabbing a piece of paper from your desk (2 second),

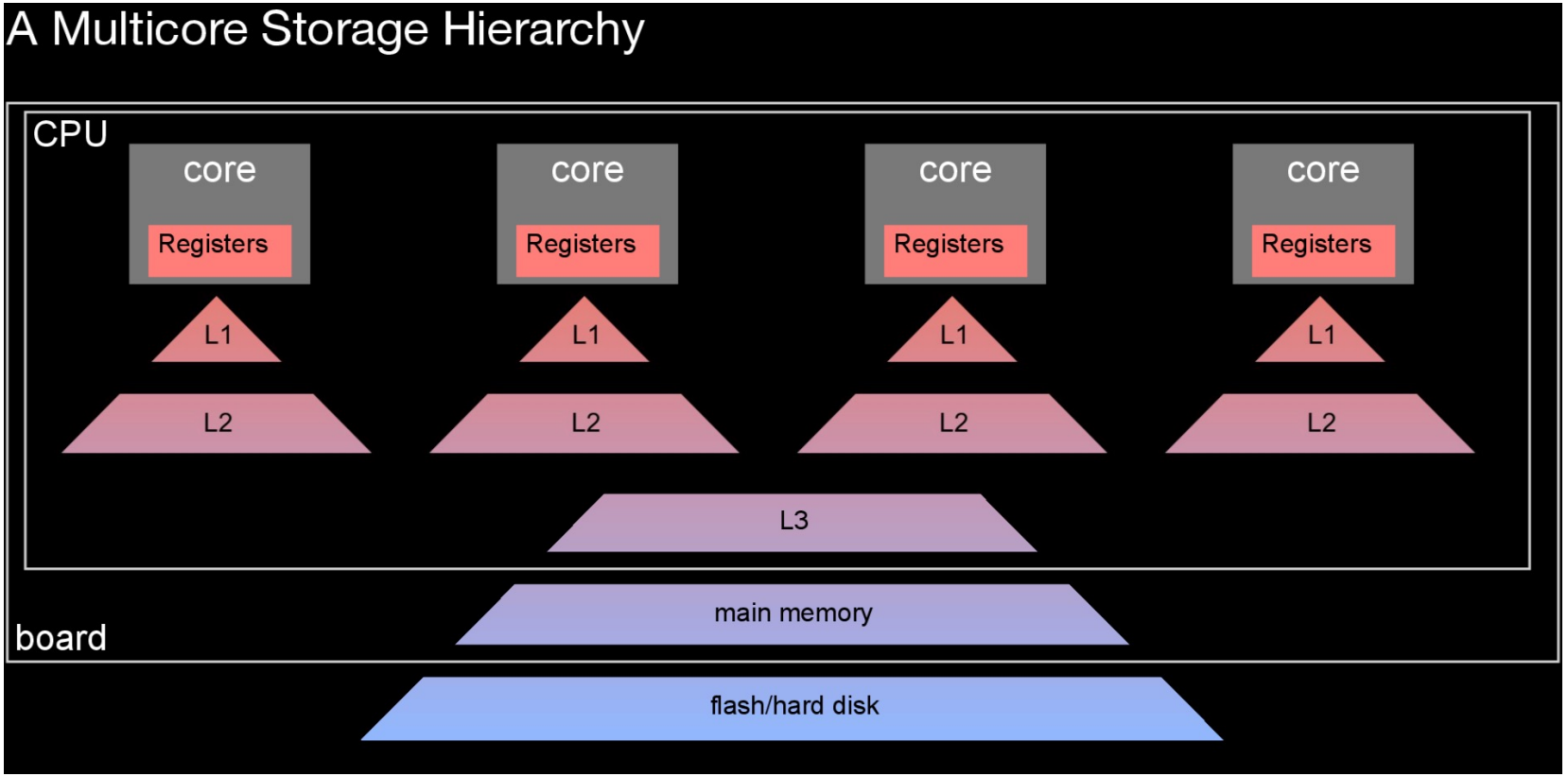
L2 cache is picking up a book from a nearby shelf (5 seconds),

L3 cache is picking up a book from the next room (30 seconds),

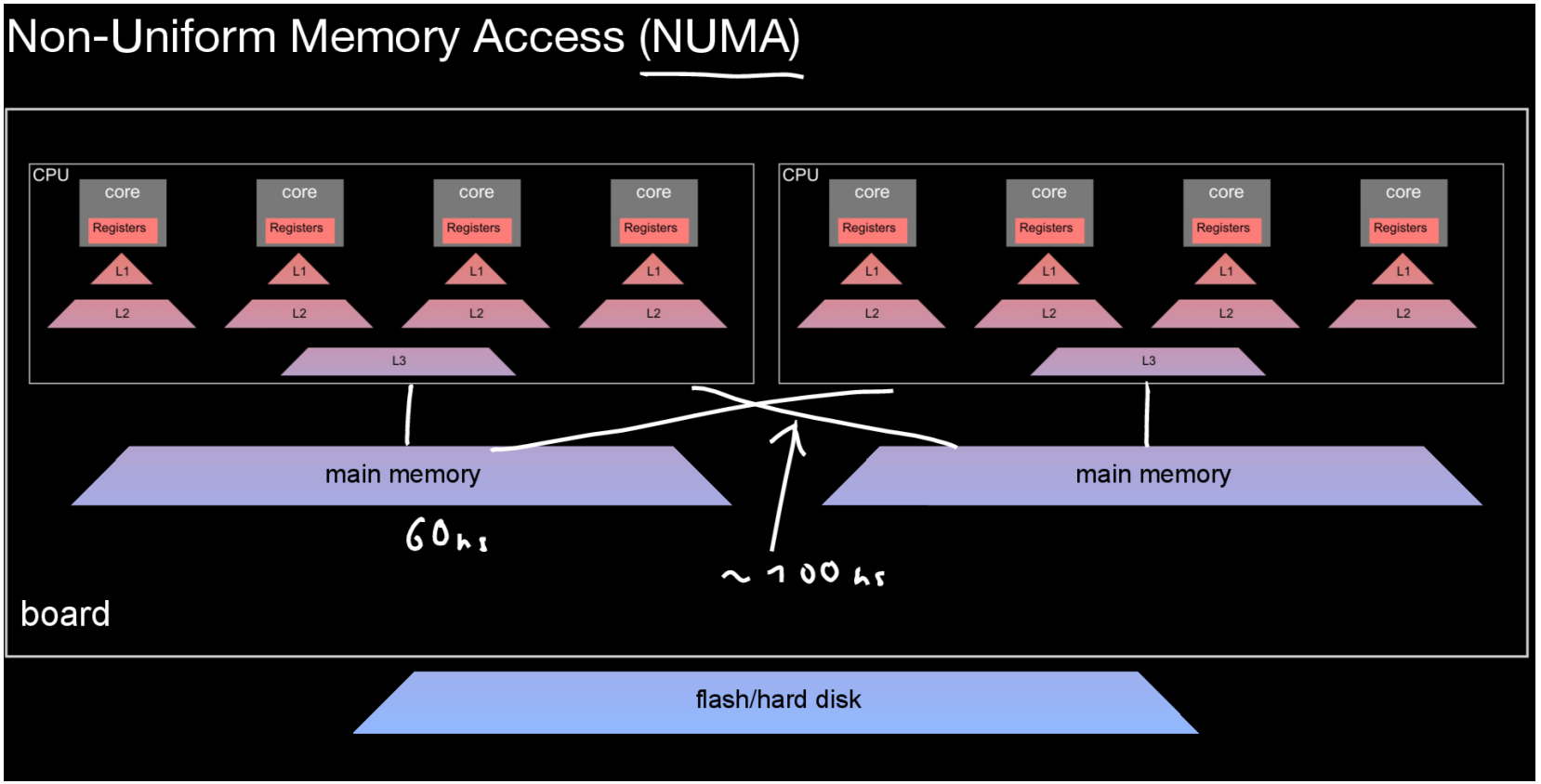
DRAM is taking a walk down the hall to buy a Twix bar (90 seconds)."

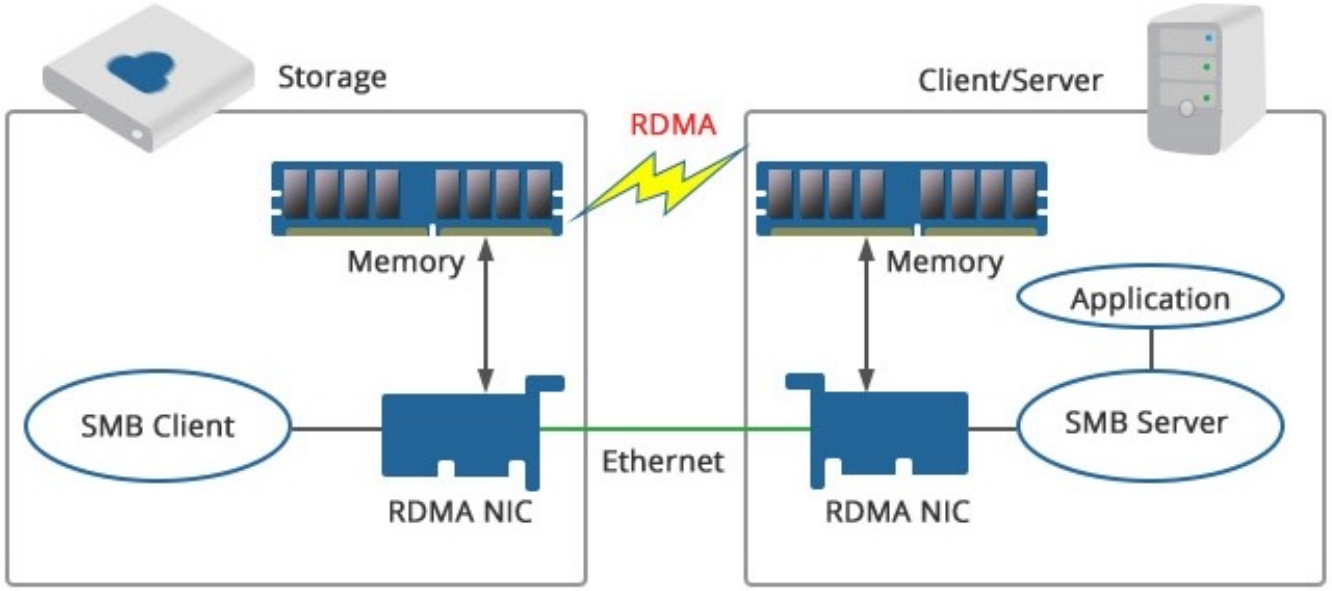


Hierarquia de memória



Hierarquia de memória

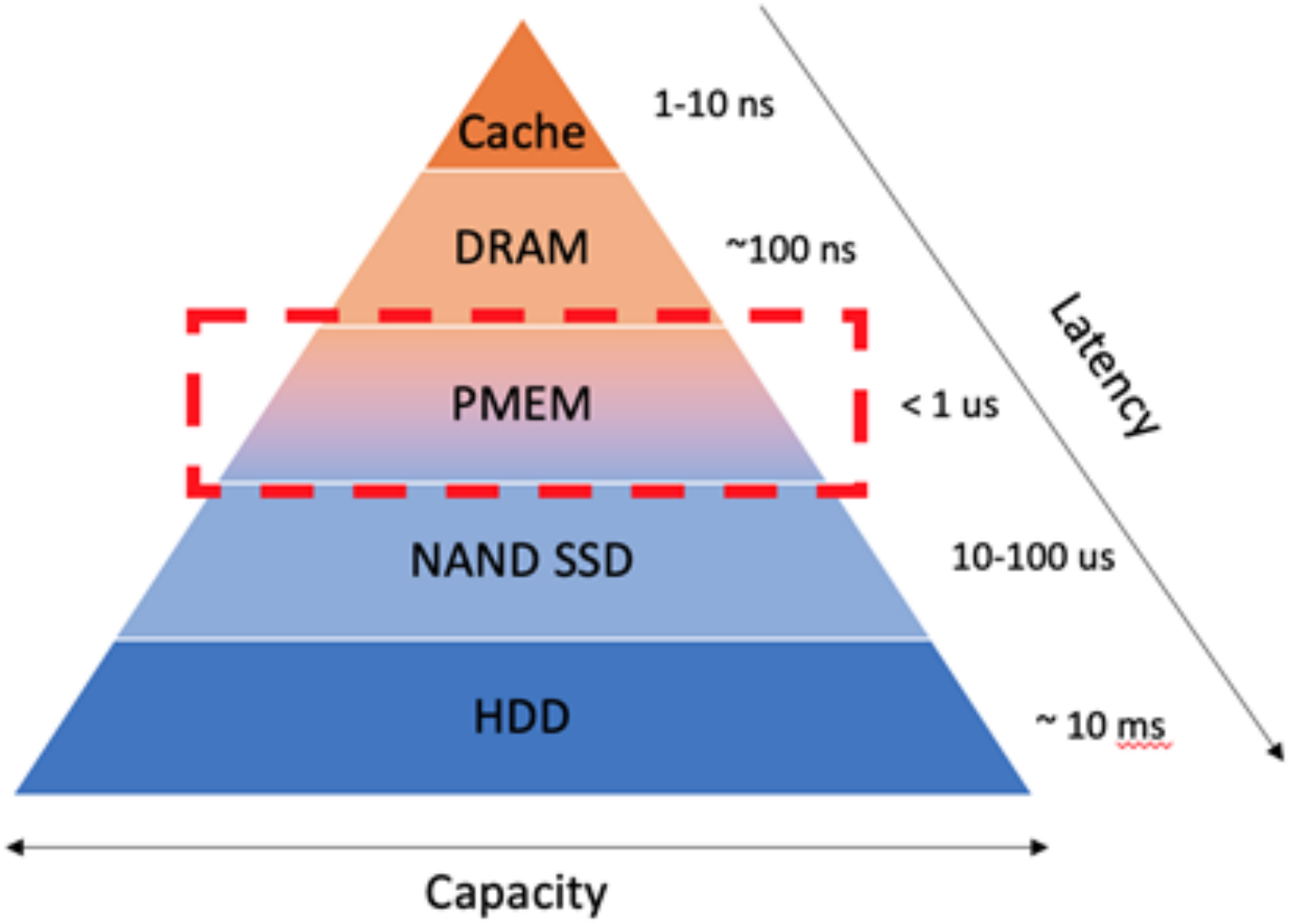




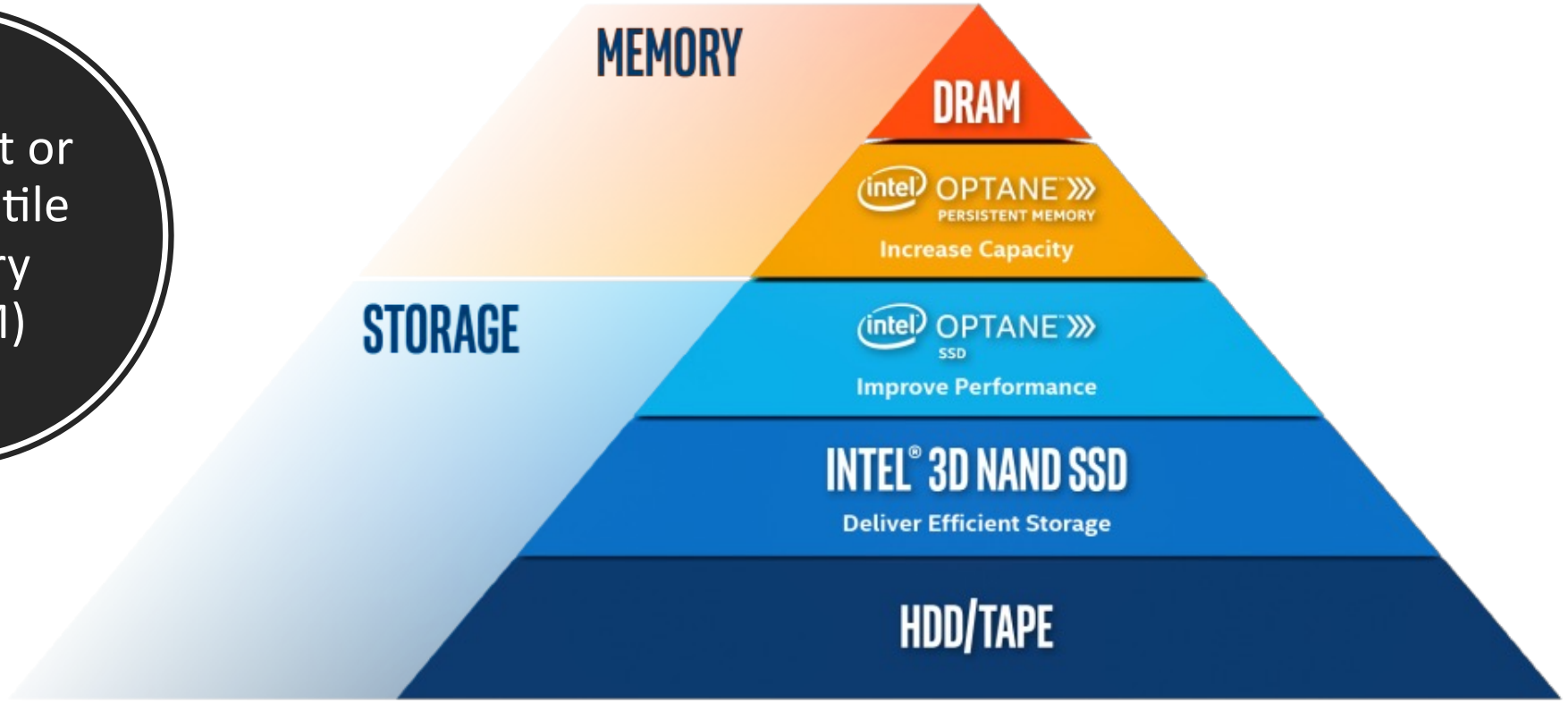
RDMA – Remote Direct Memory Access

- Enables direct memory access from the memory of one host or server to the memory of another host or server without involving the CPU.

Persistent or Non Volatile Memory (PMEM)



Persistent or Non Volatile Memory (PMEM)



TPU, TCU, NPU (tensor procs) & DBMS @Sigmod 2022

TCUDB: Accelerating Database with Tensor Processors

Yu-Ching Hu
University of California, Riverside
yhu130@ucr.edu

Yuliang Li
Megagon Labs
yuliang@megagon.ai

Hung-Wei Tseng
University of California, Riverside
htseng@ucr.edu

ABSTRACT

The emergence of novel hardware accelerators has powered the tremendous growth of machine learning in recent years. These accelerators deliver incomparable performance gains in processing high-volume matrix operators, particularly matrix multiplication, a core component of neural network training and inference. In this work, we explored opportunities of accelerating database systems using NVIDIA's Tensor Core Units (TCUs). We present TCUDB, a TCU-accelerated query engine processing a set of query operators including natural joins and group-by aggregates as matrix operators within TCUs. Matrix multiplication was considered inefficient in the past; however, this strategy has remained largely unexplored in conventional GPU-based databases, which primarily rely on vector or scalar processing. We demonstrate the significant performance gain of TCUDB in a range of real-world applications including entity matching, graph query processing, and matrix-based data analytics. TCUDB achieves up to 288× speedup compared to a baseline GPU-based query engine.

1 INTRODUCTION

the query engine's performance. Three major challenges must be addressed.

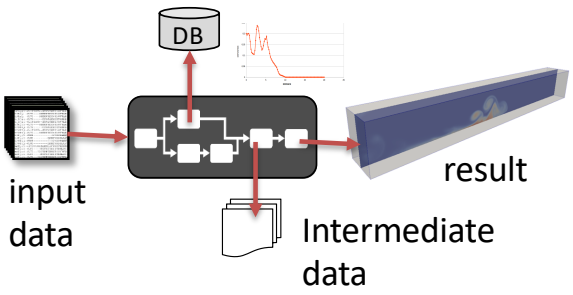
Challenges. First, the conventional GPU databases primarily implement the physical operators (e.g., the partitioned hash join algorithm [44]) in a non-matrix-friendly manner. These algorithms and operators typically do not operate on tensors directly. As a result, it is hard to modify them with the intent of taking advantage of TCUs' computation power.

Second, although DB operators such as joins can theoretically be encoded as matrix multiplications, executing all of them as dense multiplication might not always be beneficial. For example, the underlying data distributions can cause the two operands to be sparse matrices, which require a different data organization and APIs to achieve the best performance.

Next, a DB engine with TCUs must prevent itself from generating erroneous query results because of the low-precision nature of the tensor processors. The current tensor processors are limited in precision as AI/ML applications are error-tolerant because NVIDIA's TCUs only support 16-bit floating-point numbers while Google's TPUs only work on at most 8-bit integers. Moreover, these

Disciplinas ligadas a Engenharia/Bancos de dados

Why use provenance?



Provenance enables scientists to reason about results

- to assess how many trial-and-error paths produced a particular result
- how a given result was **derived**
- which **processes** led to a given result

Juliana Freire, David Koop, Emanuele Santos, and Cláudio T Silva. 2008. Provenance for computational tasks: A survey. Comput. Sci. Engineer. 10, 3 (2008), 11–21

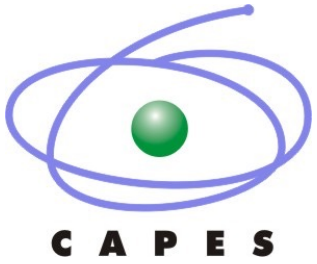
It is worth collecting provenance!

Provenance enables scientists to fine-tune during the experiment

- monitoring
- debugging
- adapting
- trust

Silva V, Souza R, Camata J, de Oliveira D, Valduriez P, Coutinho AL, Mattoso M. Capturing provenance for runtime data analysis in computational science and engineering applications. InInternational Provenance and Annotation Workshop (IPAW) 2018

Acknowledgements

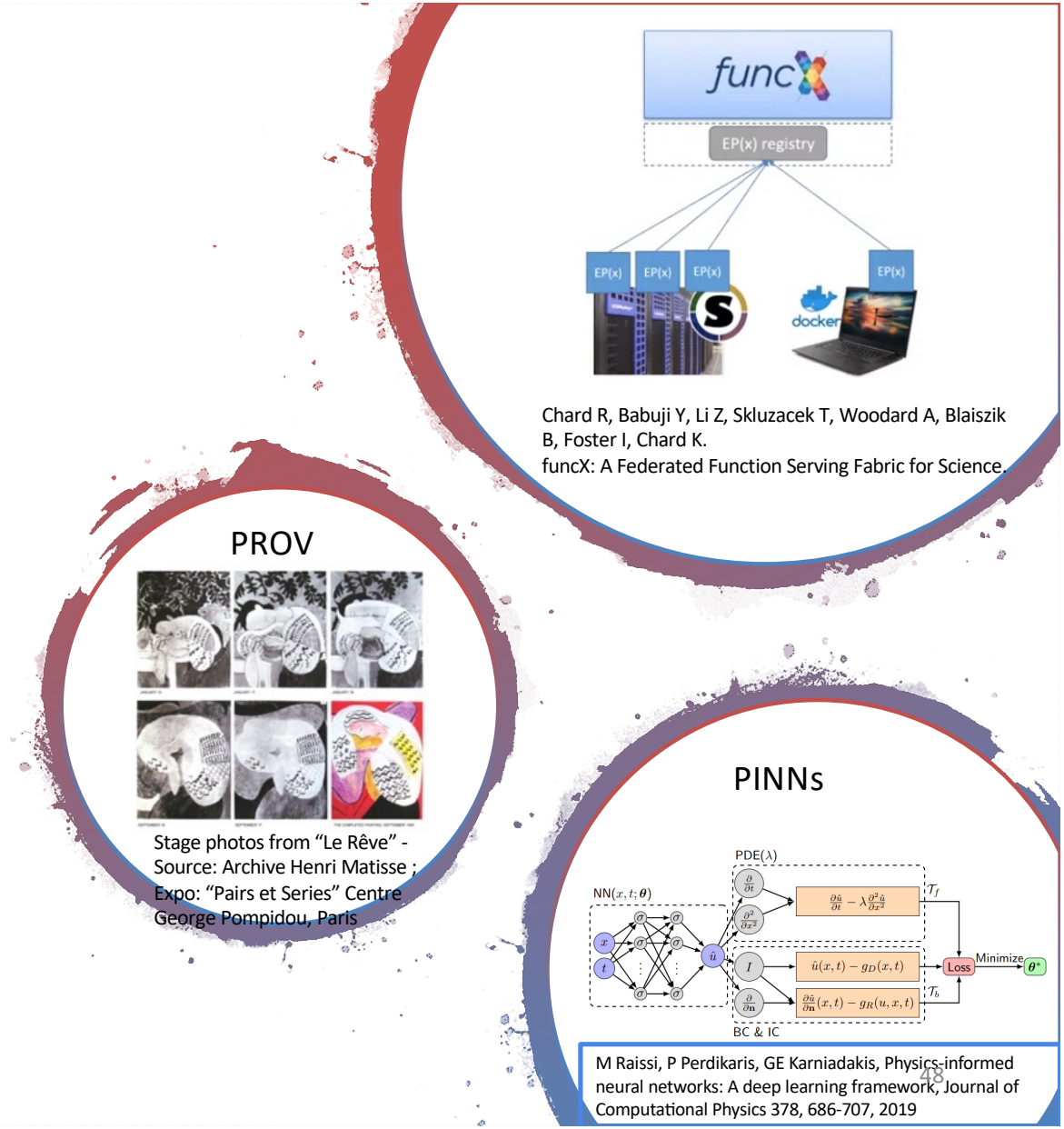


Obrigada!!!

PROV:
show me your data

Human-in-the-loop

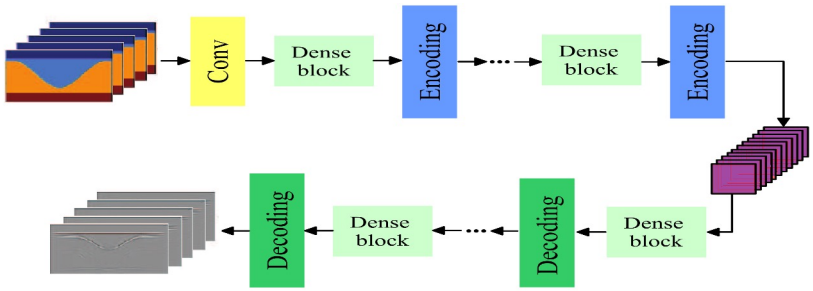
Computing in the continuum



Thanks to Provenance Collabs

The problem of model selection

Model architecture



Growth Rate, #layers in the dense block

Hyperparameters fine tuning during training

The configuration search space

Searching Configuration Space

AutoML

e.g.: grid Search, random, Bayesian

Interactive Human-In-the-Loop

The CEREBRO framework

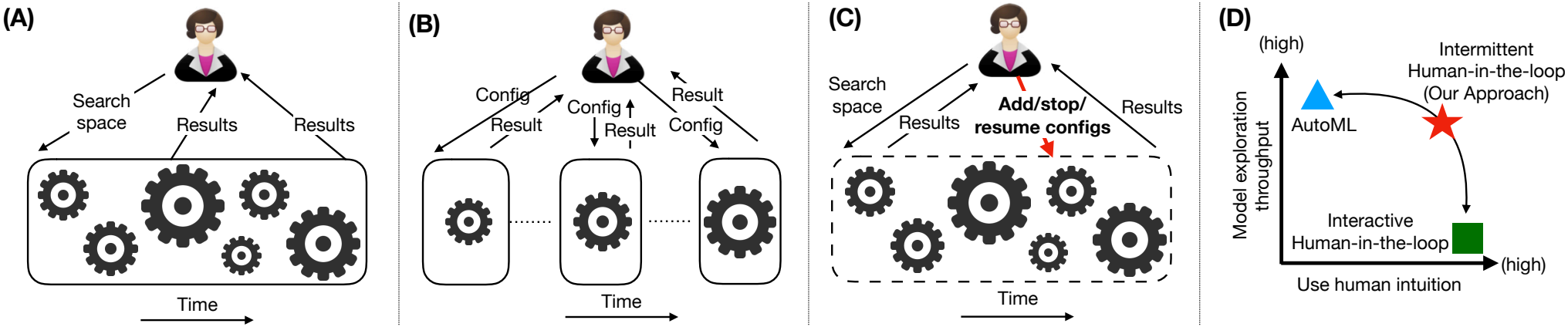


Figure 1: A) AutoML-based model selection. B) Interactive human-in-the-loop model selection. C) Our paradigm of *intermittent human-in-the-loop model selection.* D) Qualitative comparison of different paradigms.

Li L, Nakandala S, Kumar A. Intermittent human-in-the-loop model selection using cerebro: a demonstration. Proceedings of the VLDB Endowment. 2021 Jul 1;14(12):2687-90.