Ciência de Dados ≠ Inteligência Artificial (ou Neural Networks)

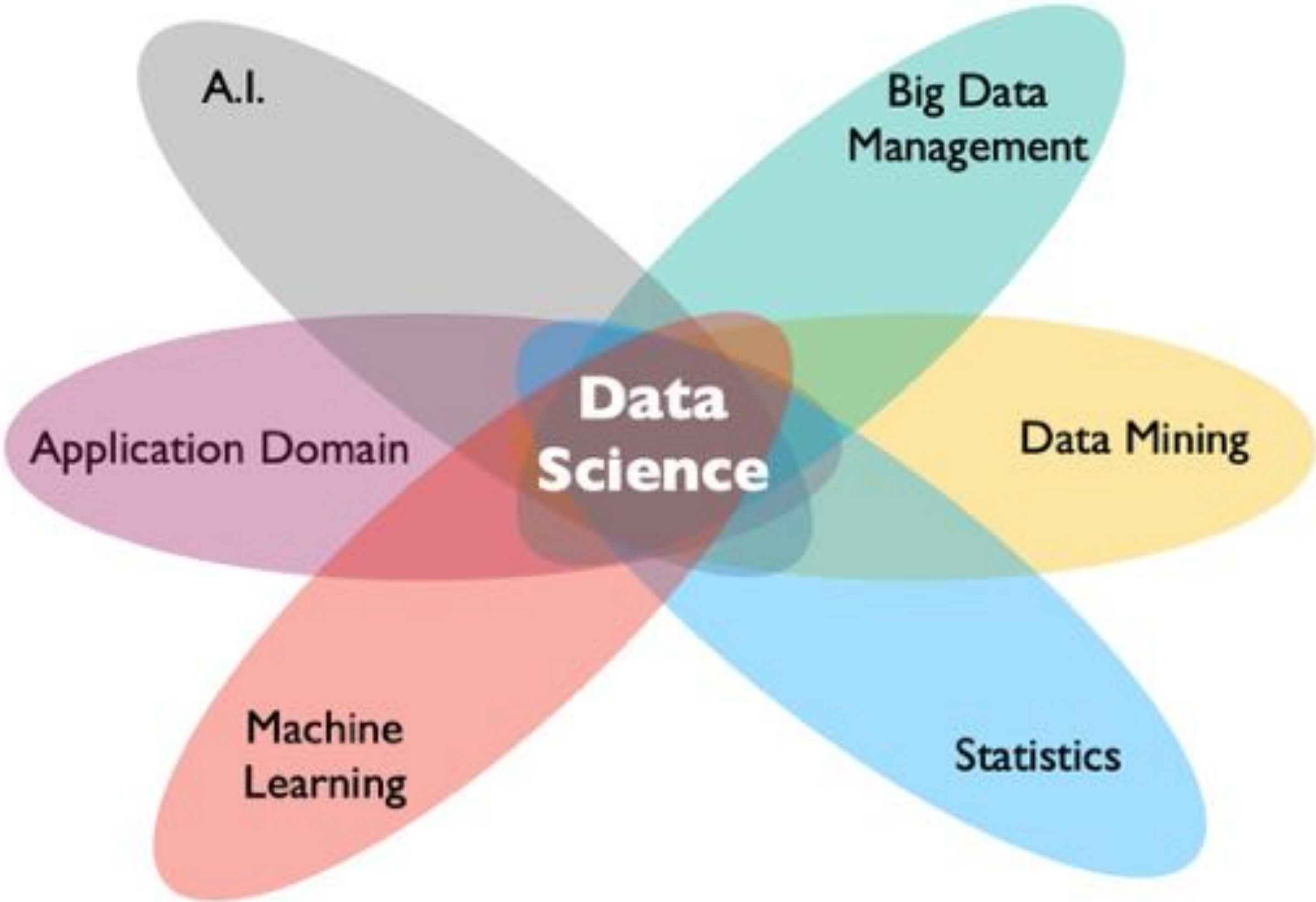
Marta Mattoso
EDC

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


big
cycle
talk
may
life
need
current
activities
features
analyses
Scientific
broader
explorations
Big
However
involve
issues
explore
lack
analysis
manage
part
scientific
process
analysts
consistent
many
experiments
typically
provide
provenance
steering
parallel
programs
Provenance
experiment
different
execution
results
workflows
useful
discuss
SWfMS
view
Workflow
DBMS
datasets
interference
HPC
reproducible
workflow
support
data
using
analyze
approaches

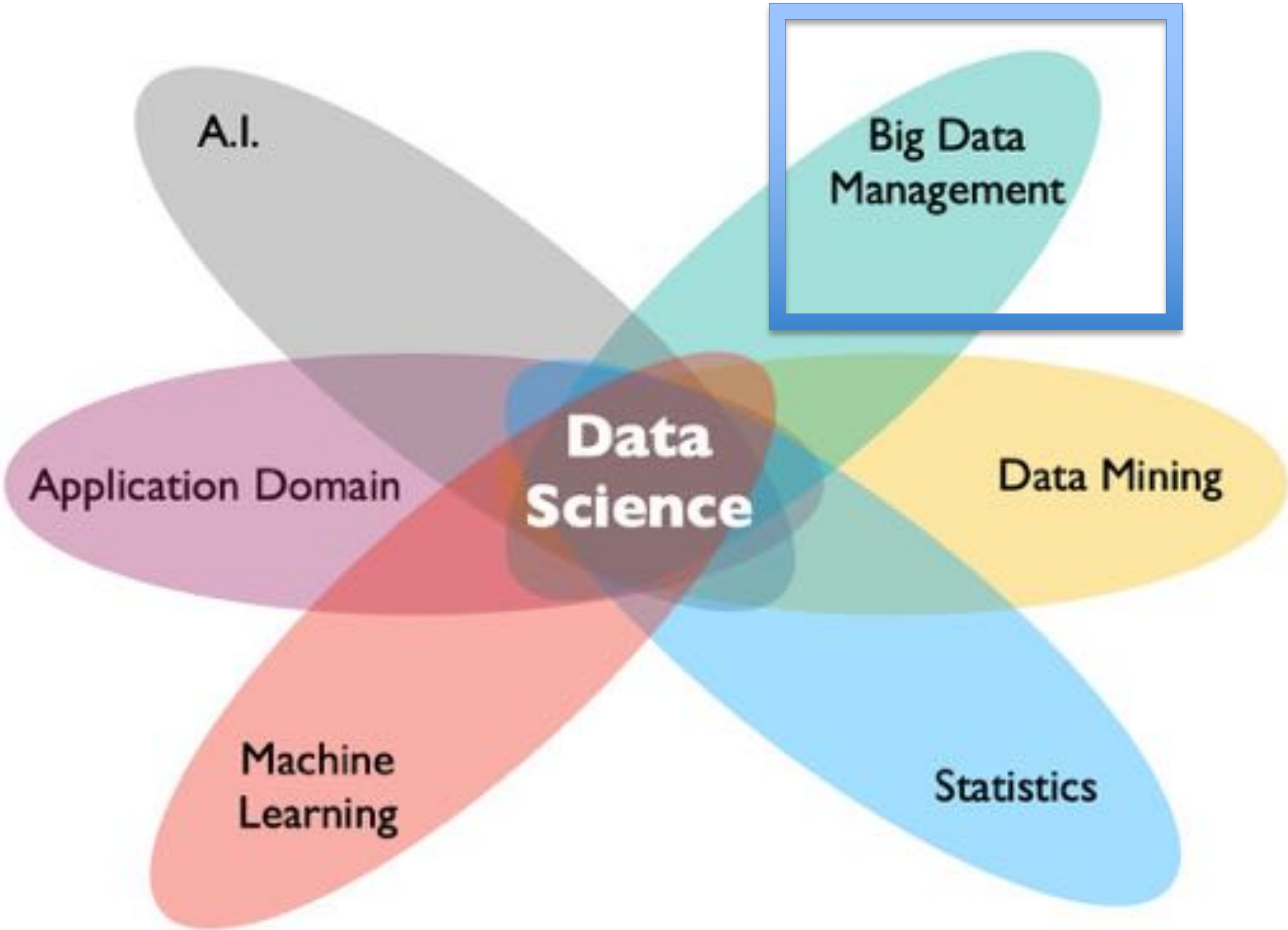

Scientific
Provenance
workflows
data
analysis
workflow
DataIntensive
scheduling
dynamic
Computational
Analyzing
simulation
Library
practical
multisite
Cloud
de
using
Efficient
dataflows
Science
Runtime
raw
visualization
capture
Learning
Towards
Engineering

Data Science ≠ Machine Learning

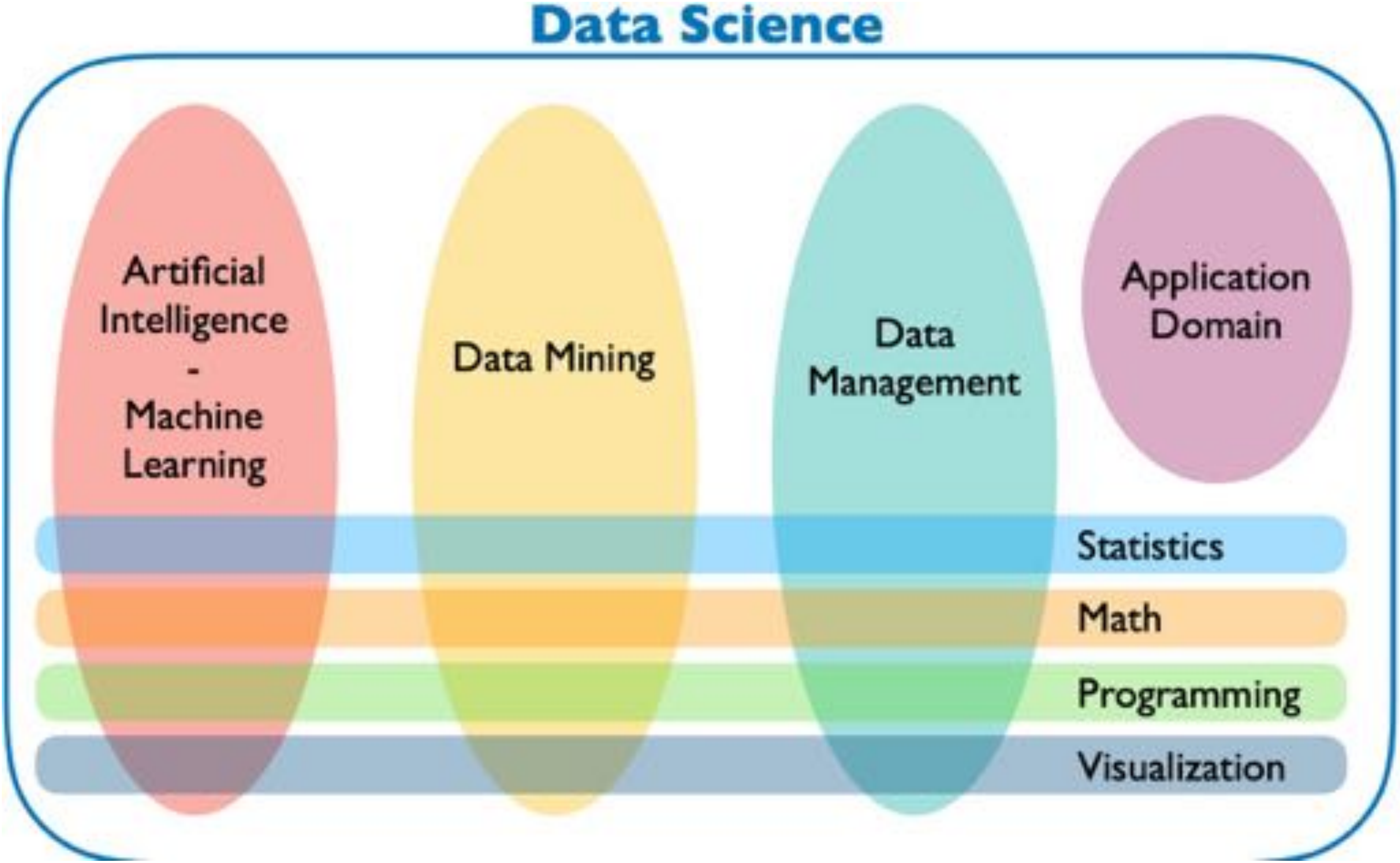


Jens Dittrich "Data Science ≠ Machine Learning: Some Thoughts on the Role of Data Management in the new AI-Tsunami" -- Keynote DEEM@SIGMOD 2018, June 2018

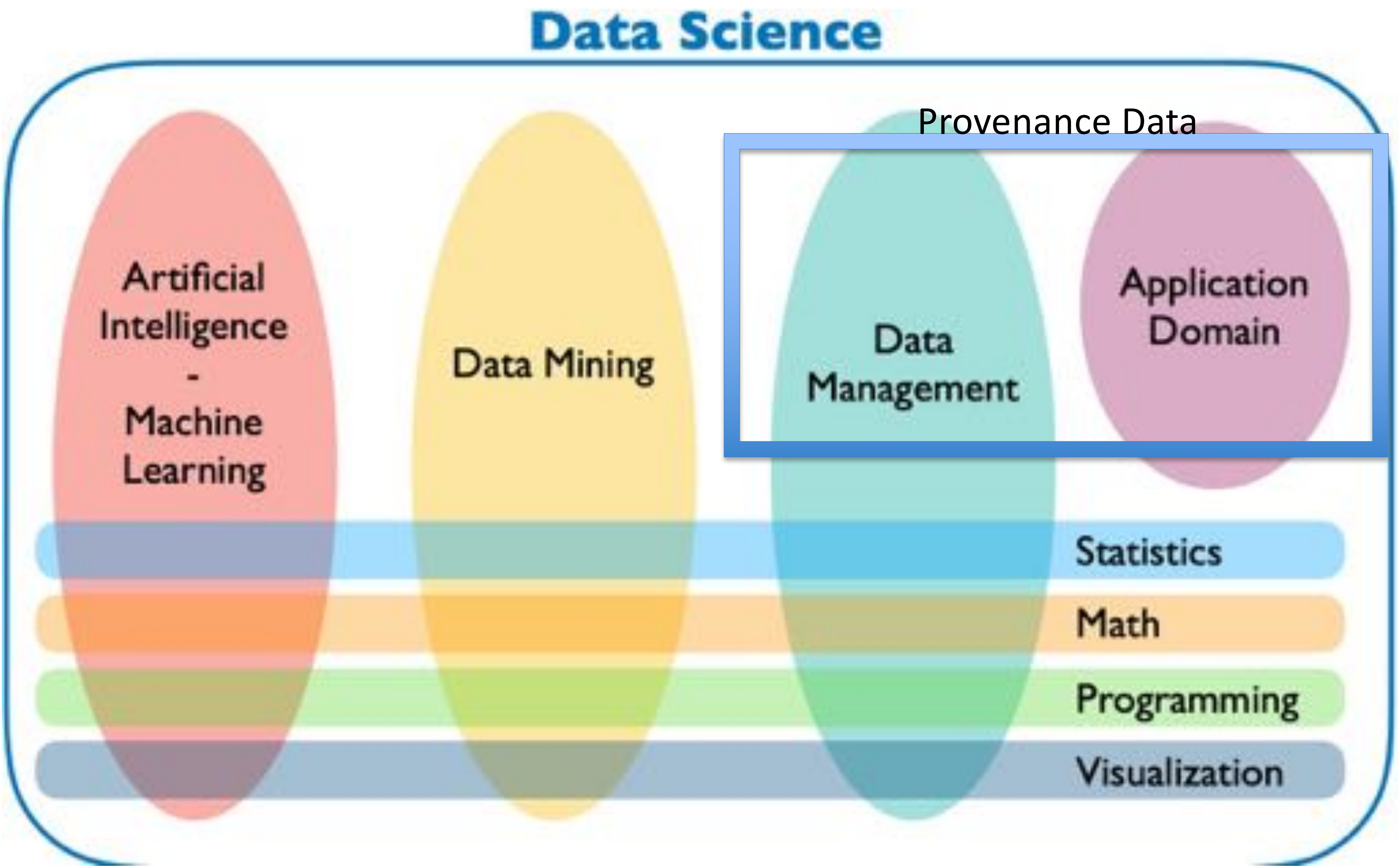
Data Science ≠ Machine Learning



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



Jens Dittrich, VLDB 2017 (invited talk: “Deep Learning (m)eats Databases”)



Jens Dittrich, VLDB 2017 (invited talk: “Deep Learning (m)eats Databases”)

Projetos de Pesquisa

- D-Interpret - Gerência de dados para auxiliar a explicação de resultados em aplicações de ciência de dados- CNPq Universal/Faixa C
- MonDataSim - Análise de dados de simulações computacionais por meio de monitoramento da execução, dados de proveniência e intervenções dinâmicas- Faperj Cientista RJ
- SciDISC - Scientific data analysis using Data-Intensive Scalable Computing- França-Brasil, INRIA Associate Teams - Patrick Valduriez
- Bolsa Produtividade de Pesquisa CNPq- 1B

Trabalho em Equipe

Orientações
Graduação > 30
Mestrado > 60
Doutorado > 20

ECI – aluno de IC 2008

Fernando Chirigati   Research   Awards and Honors   Publications   Talks   Professional Activities   Bio   CV

Fernando Chirigati

Postdoctoral Research Associate

NYU Tandon School of Engineering
VIDA Research Center
370 Jay Street, 11th Floor
Brooklyn, NY 11201

fchirigati [at] nyu [dot] edu

Twitter

LinkedIn

About Me

Currently, I'm a Postdoctoral Research Associate at NYU Tandon School of Engineering, under the supervision of Prof. Juliana Freire. I came to the beautiful - and very busy - city of New York in January 2012 to pursue a Ph.D. degree. Before, I worked as a Research Assistant at Federal University of Rio de Janeiro (UFRJ), under the supervision of Prof. Marta Mattoso. I have a Ph.D. in Computer Science from NYU, and a B.E. in Computer and Information Engineering from UFRJ. To check out my full CV, click here.

I come from the gorgeous city of Petrópolis, in Brazil, where almost all my family and friends still reside. I had the opportunity to study in Rio de Janeiro, the "Marvelous City," where I not only made a lot of good friends, but also started working with research in the database area.

Research

My research interests are mainly in the area of scientific data management, including **provenance management and analytics**, **large-scale data analytics**, **data science**, **data mining**, **reproducibility**, and **data visualization**.

ECI – aluno de MSc (2011) e DSc (2013)

Melhor Tese de Doutorado no Concurso de Teses do XXX Simpósio Brasileiro de Bancos de Dados, SBC.

ECI – aluno de IC (2012); MSc (2014) e DSc (2018)

Melhor Tese de Doutorado no Concurso de Teses do XXXIV Simpósio Brasileiro de Bancos de Dados, SBC.

Atualmente na Snap, Los Angeles

DCC/IM – aluno de MSc (2015) e DSc previsão defesa 2019

Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows

Authors	Daniel de Oliveira, Eduardo Ogasawara, Fernanda Baião, Marta Mattoso
Publication date	2010/7/5
Conference	2010 IEEE 3rd International Conference on Cloud Computing
Pages	378-385
Publisher	IEEE
Description	Most of the large-scale scientific experiments modeled as scientific workflows produce a large amount of data and require workflow parallelism to reduce workflow execution time. Some of the existing Scientific Workflow Management Systems (SWfMS) explore parallelism techniques - such as parameter sweep and data fragmentation. In those systems, several computing resources are used to accomplish many computational tasks in homogeneous environments, such as multiprocessor machines or cluster systems. Cloud computing has become a popular high performance computing model in which (virtualized) resources are provided as services over the Web. Some scientists are starting to adopt the cloud model in scientific domains and are moving their scientific workflows (programs and data) from local environments to the cloud. Nevertheless, it is still difficult for the scientist to express a parallel computing ...
Total citations	Cited by 189

Fonte: Google Scholar

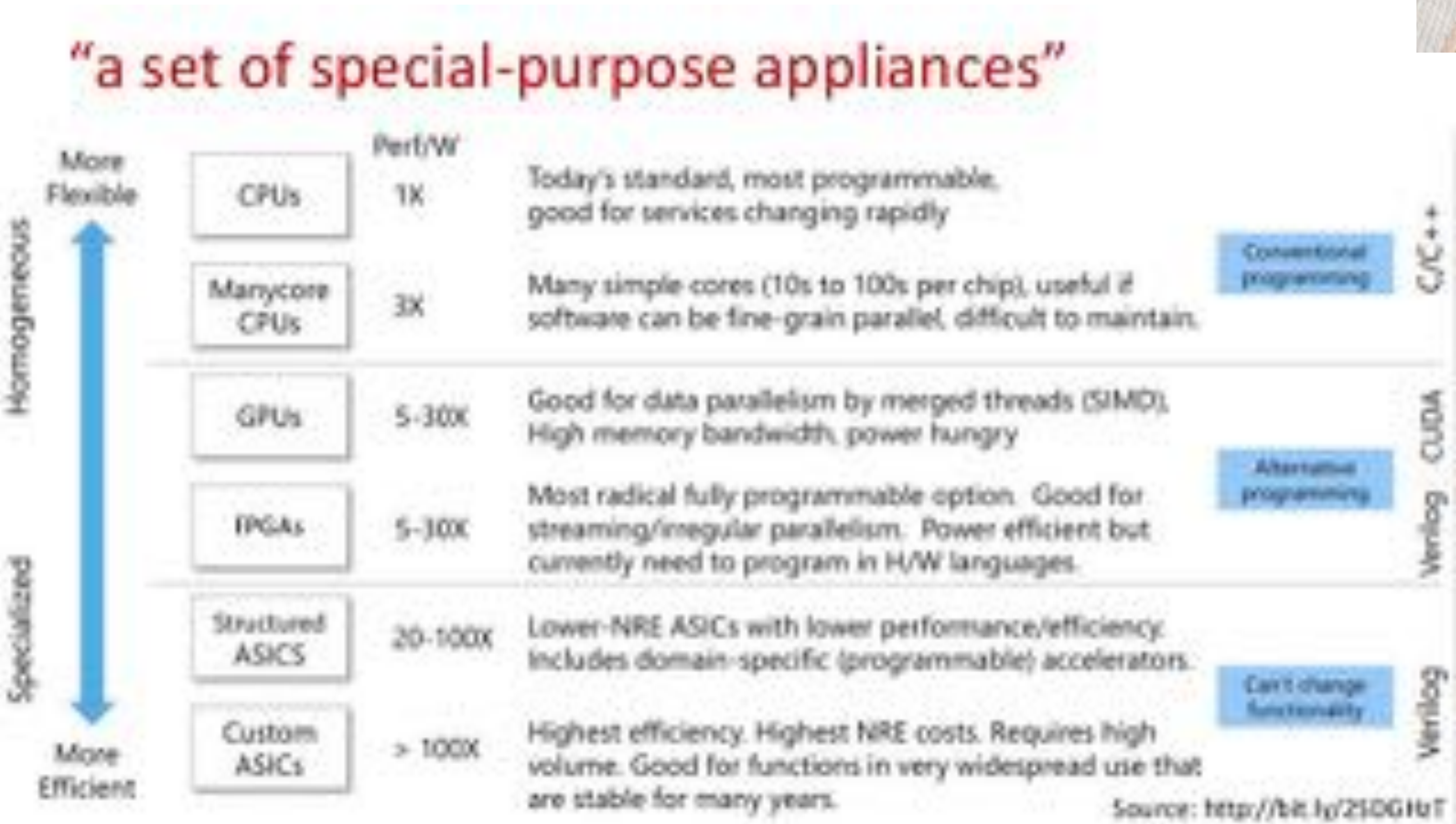
Data Challenges

Computing in Continuum

Human-In-the-Loop (HIL)

Data Integration

Especialização: extremos



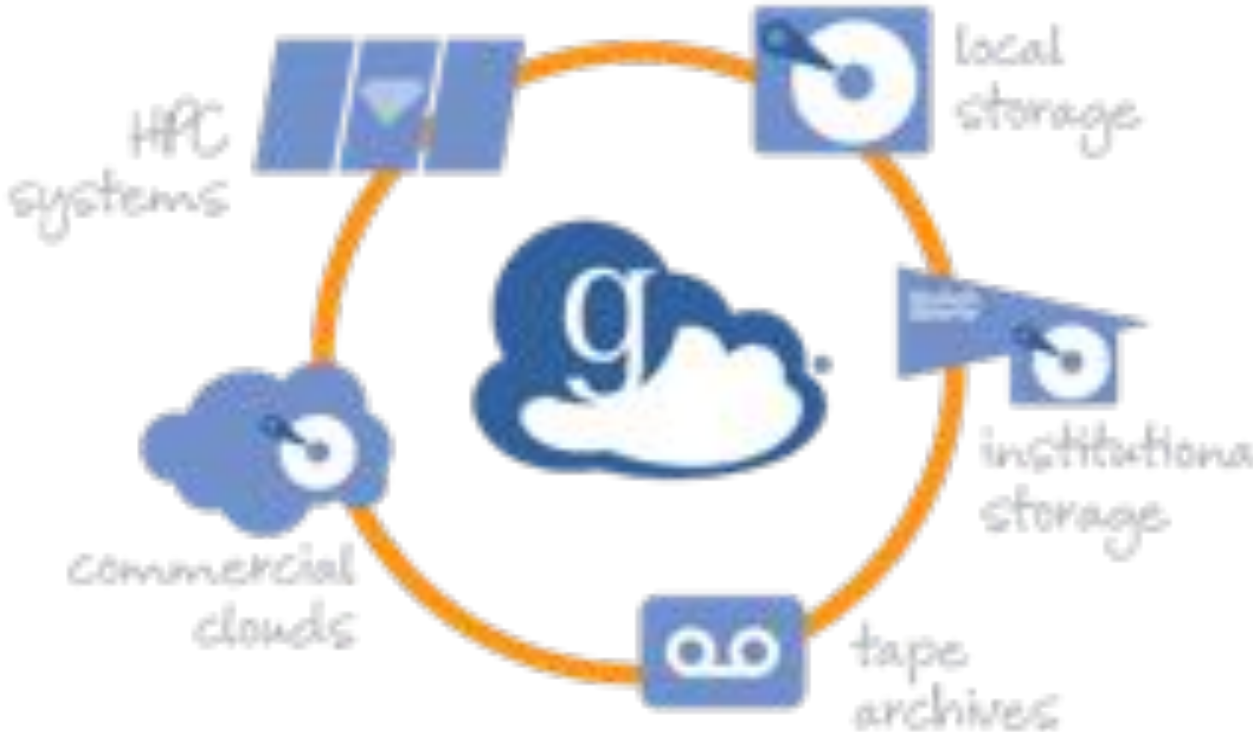
http://www.ianfoster.org/wordpress/presentations/ (Code in Continuum)

Exascale “sob medida”

- Google projetou sua própria unidade de processamento de tensores (TPU), projetada inicialmente para alta vazão de operações aritméticas de baixa precisão. Novas gerações de TPU chegam a petaflops além da Edge TPU
  - TPU aumenta o desempenho das redes neurais por trás de Google Search, Street View, Google Photos and Google Translate.

Code in continuum

Globus



Ian Foster
(IPDPS’2019)

https://www.slideshare.net/ianfoster/coding-the-continuum

Putting the human in the loop

"In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding."

Exploring the inherent technical challenges in realizing the potential of Big Data.

BY H.V. JAGADISH, JOHANNES GEHRKE, ALEXANDROS LABRINIDIS, YANNIS PAPAKONSTANTINOU, JIGNESH M. PATEL, RAGHU RAMAKRISHNAN, AND CYRUS SHAHABI

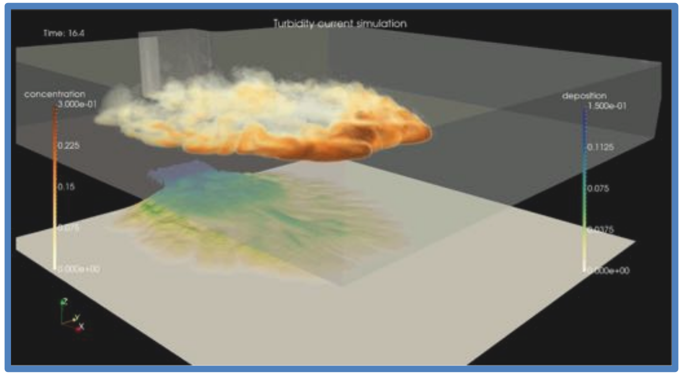
Big Data and Its Technical Challenges

Systems were not built to have humans in the loop

JULY 2014 | VOL. 57 | NO. 7 | **COMMUNICATIONS OF THE ACM**

PROVENANCE IN DATA ANALYTICS

Analyzing sedimentation solver data with provenance data

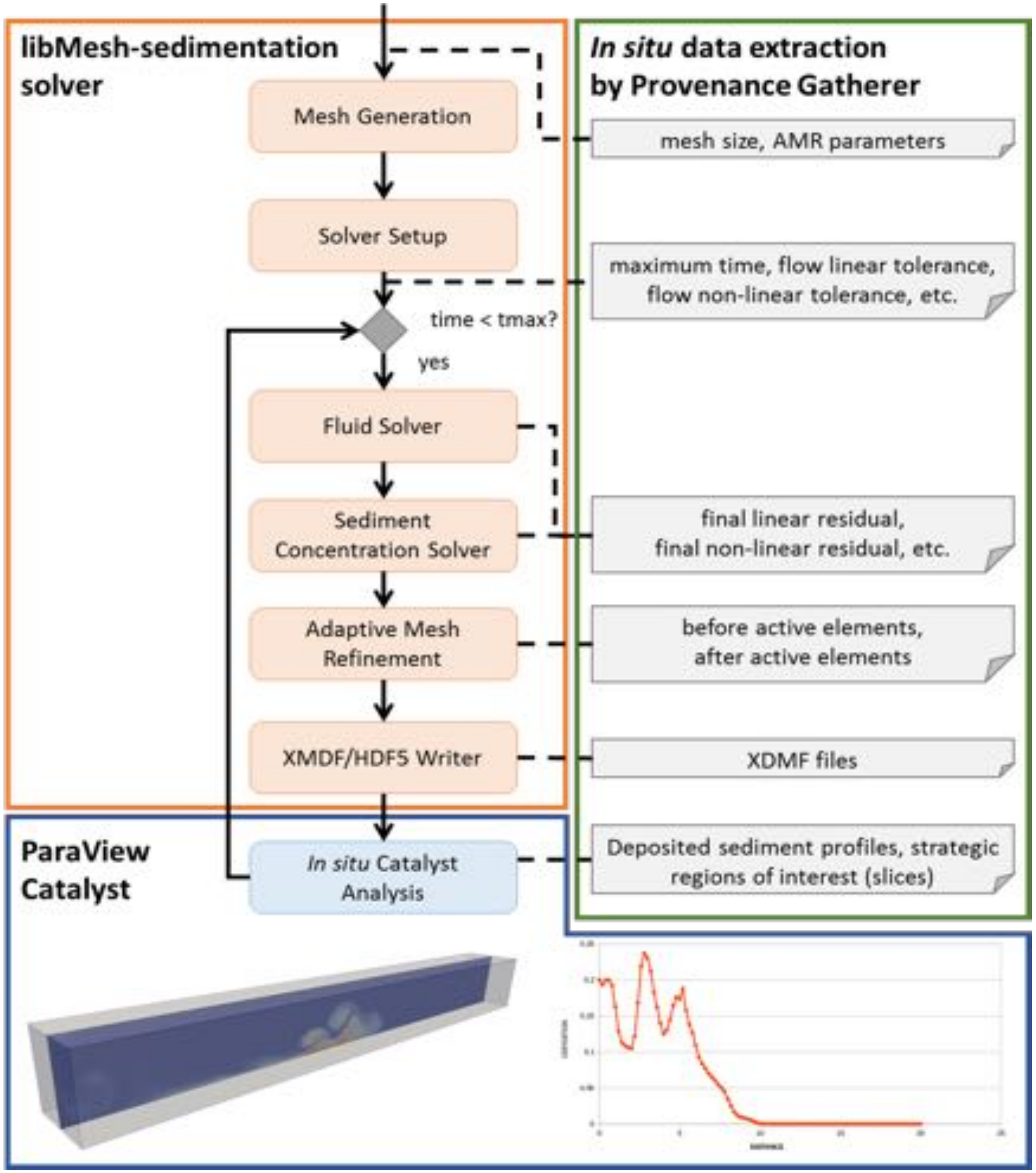


Real Case Study:

- DfAnalyzer tool
  - Provenance Gatherer
  - Simulation Data Extractor
  - Extract data from libMesh
  - Query and visualize at runtime
- TACC computer using 1,040 cores
- Simulation elapsed time: 137.75 min

  Solver: 136.80 min → 99.31% of total time
  DfAnalyzer w. Catalyst: 0.95 min → 0.69%

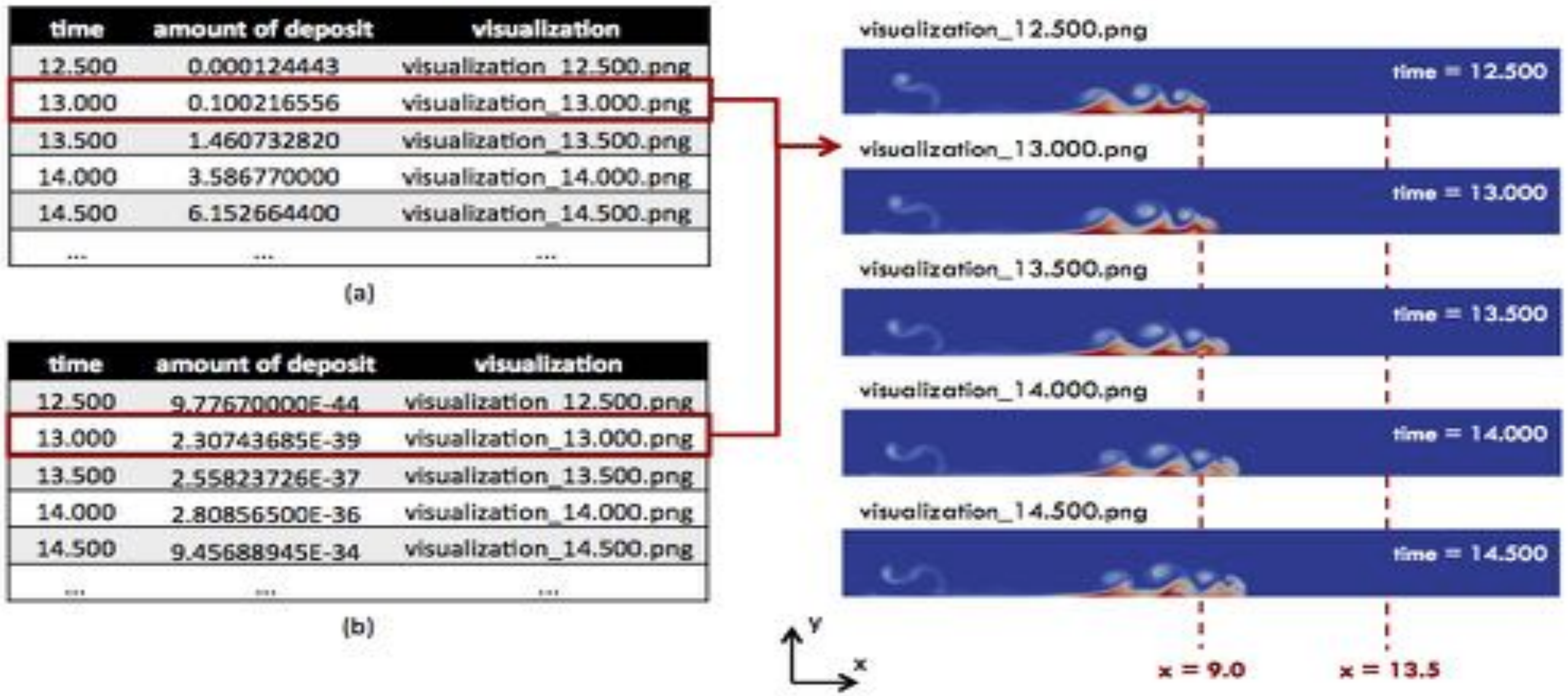
  Initial input mesh: 480 x 80 x 80 hexahedra
  Total time steps: 200



Silva, V., de Oliveira, D., Valduriez, P., & Mattoso, M. DfAnalyzer: Runtime Dataflow Analysis of Scientific Applications using Provenance
PVLDB 2018.

Sedimentation provenance data analysis complementing viz tools

- DfAnalyzer registers deposition along time at predefined locations and pointers to viz files.
- We can query online with a negligible time (< 500 ms).



time	amount of deposit	visualization
12.500	0.000124443	visualization_12.500.png
13.000	0.100216556	visualization_13.000.png
13.500	1.460732820	visualization_13.500.png
14.000	3.586770000	visualization_14.000.png
14.500	6.152664400	visualization_14.500.png
...	...	...

(a)

time	amount of deposit	visualization
12.500	9.77670000E-44	visualization_12.500.png
13.000	2.30743685E-39	visualization_13.000.png
13.500	2.55823726E-37	visualization_13.500.png
14.000	2.80856500E-36	visualization_14.000.png
14.500	9.45688945E-34	visualization_14.500.png
...	...	...

(b)

Figure: Sediment deposition monitoring at five time instants at $x = 9.0$ (a) and $x = 13.5$ (b) combining data with in-situ visual information

Camata, J.J., Silva, V., Valduriez, P., Mattoso, M., Coutinho, A.L.G.A.: In situ visualization and data analysis for turbidity currents simulation. Computers & Geosciences v.110, 23–31 (2018).

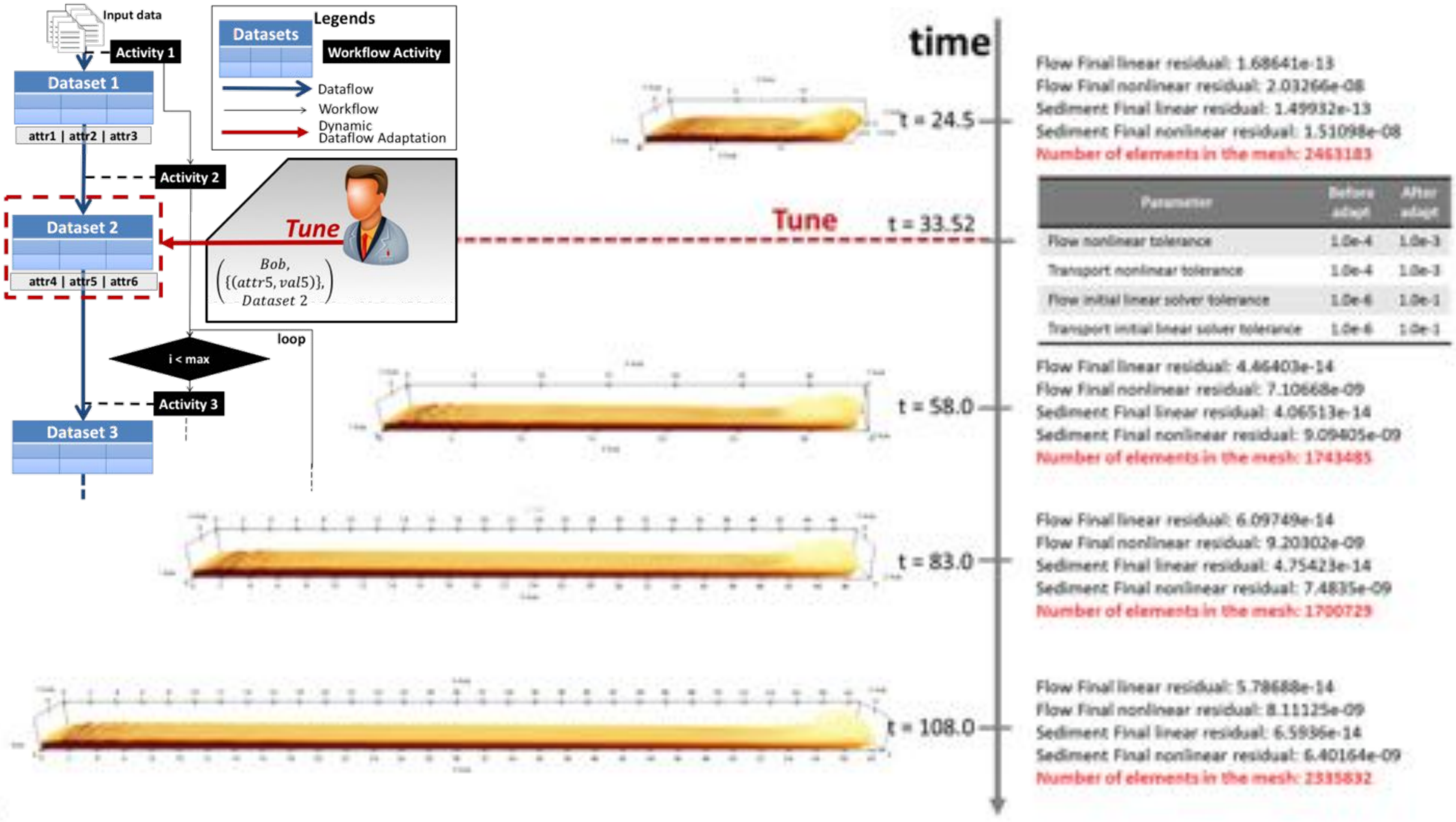
Query processing

- Analytical queries for...
  - **Monitoring:**
    - The appearance of sediments in the domain bottom layer for a specific time step
  - **Debugging and user steering:**
    - Analysis of the algorithm output parameters after the convergence of the solver in the fluid and sediments loops in a specific execution of the sedimentation solver

time step	x	y	z	d
1	0.000	1.000	0.000	2.00E-04
1	0.180	1.000	0.000	2.00E-04
1	0.360	1.000	0.000	2.00E-04
1	0.540	1.000	0.000	2.00E-04
1	0.720	1.000	0.000	1.99E-04
1	0.900	1.000	0.000	1.19E-04
1	1.080	1.000	0.000	3.04E-08
…	…	…	…	…

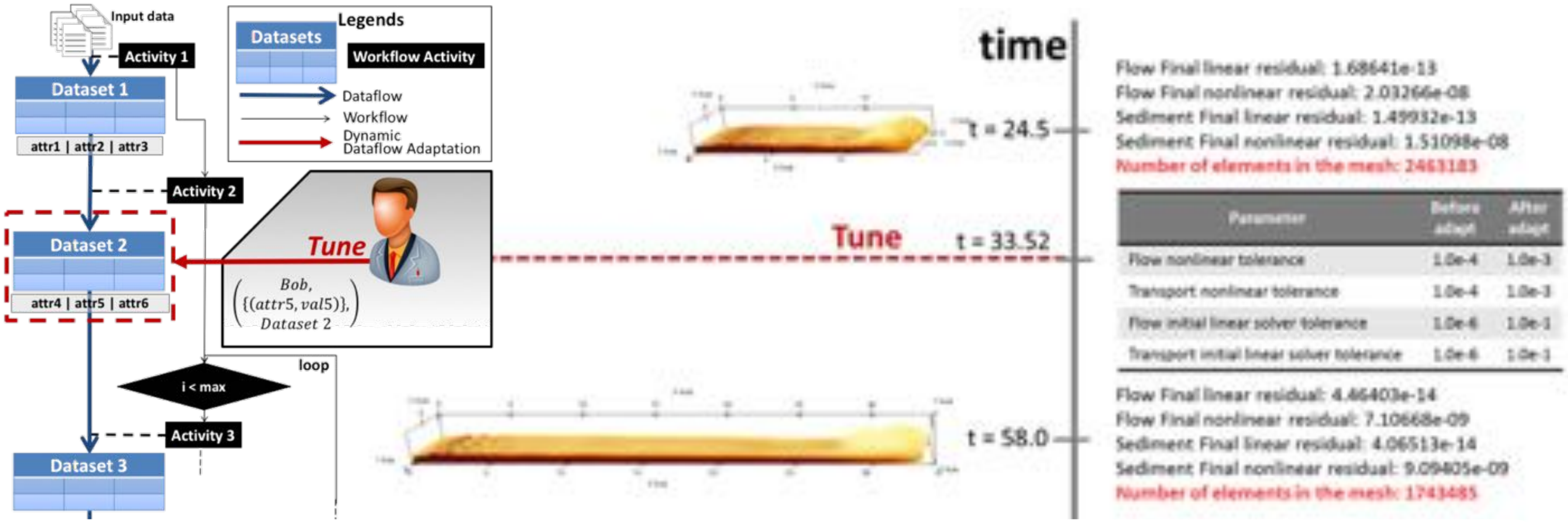
Fluid		**Sediments**	
linear residual norm	nonlinear residual norm	linear residual norm	nonlinear residual norm
3.98E-06	13.54823207	8.66E-06	0.004445721
4.30E-06	0.390224835	2.00E-09	0.002435432
4.30E-06	0.390224835	1.31E-05	0.016144017
7.00E-09	0.002712742	2.00E-09	0.002435432
7.00E-09	0.002712742	1.31E-05	0.016144017
…	…	…	…

Turbidity currents simulation data analysis with provenance

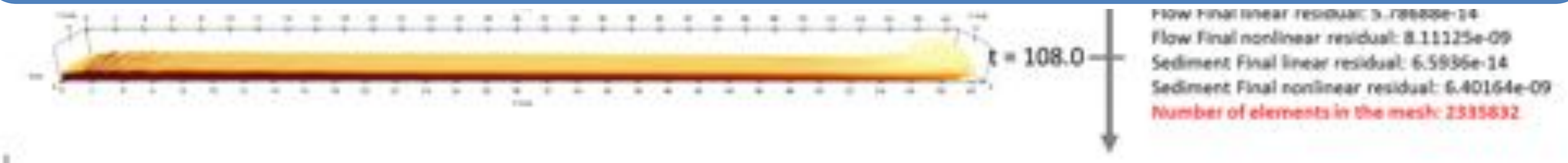


Renan Souza et al. Tracking of Online Parameter Fine-tunings in Scientific Workflows, *WORKS workshop in ACM/IEEE Supercomputing*, 2017. Extended to FGCS Special Issue 2019

Turbidity currents simulation data analysis with provenance



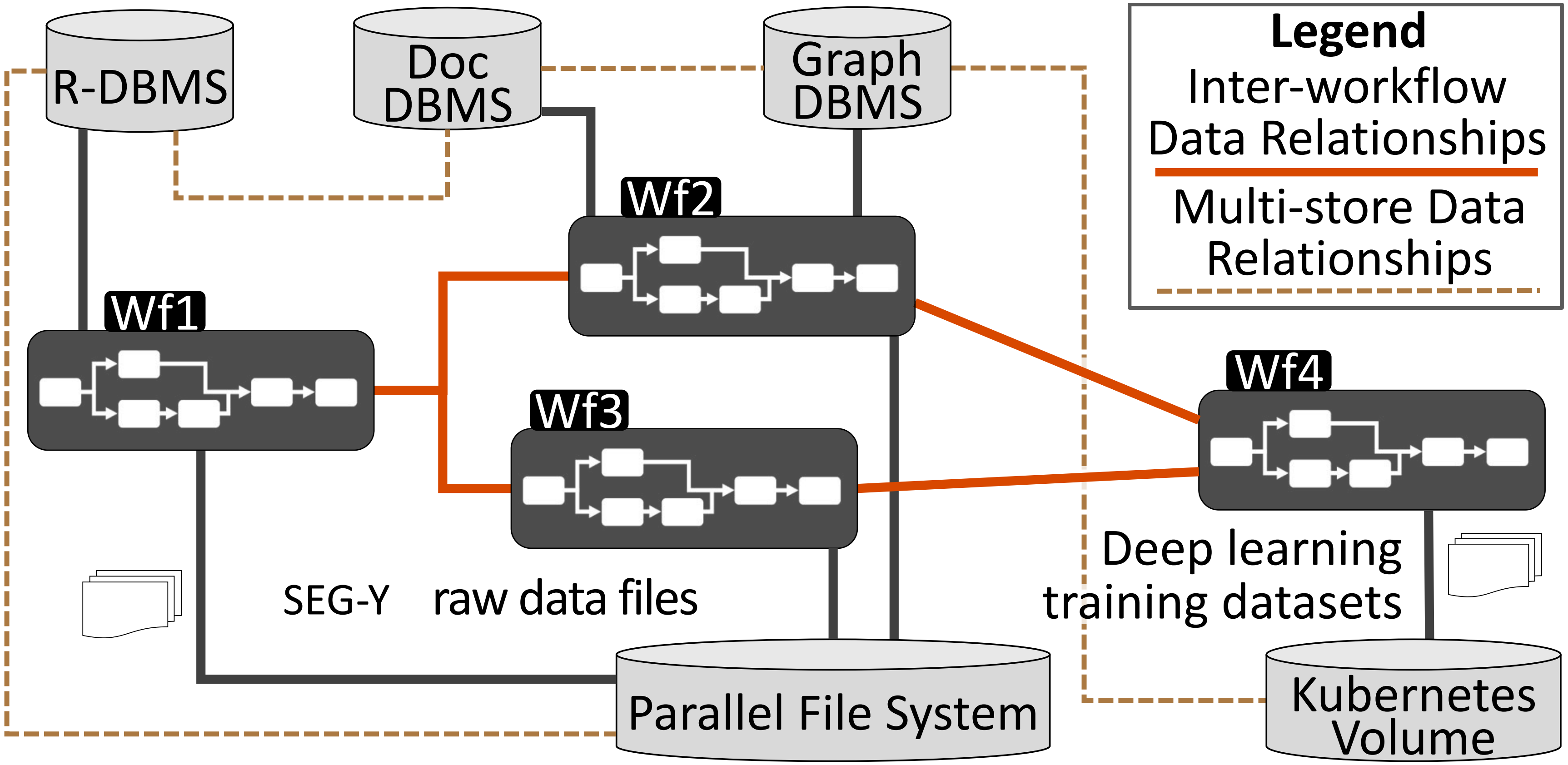
Parameter-tuning reduced the execution time by 10 days (37%), overhead < 1% and allowed job to finish successfully



Renan Souza et al. Tracking of Online Parameter Fine-tunings in Scientific Workflows, *WORKS workshop in ACM/IEEE Supercomputing*, 2017.

Provenance relating data for DL

(IEEE eScience 2019)



How the geographic coordinates were extracted from the SEG-Y file that is being used to produce training and validation files?
What is the spatial resolution between slices in the seismic data?

+ R. Souza, ongoing Ph.D. thesis at COPPE/UFRJ

Big data need big theory too

Peter V. Coveney[1], Edward R. Dougherty[2] and Roger R. Highfield[3]

[1]Centre for Computational Science, University College London, Gordon Street, London WC1H 0AJ, UK
[2]Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77843-31283, USA
[3]Science Museum, Exhibition Road, London SW7 2DD, UK

PVC, 0000-0002-8787-7256

The current interest in big data, machine learning and data analytics has generated the widespread impression that such methods are capable of solving most problems without the need for conventional scientific methods of inquiry. Interest in these methods is intensifying, accelerated by the ease with which digitized data can be acquired in virtually all fields of endeavour, from science, healthcare and cybersecurity to economics, social sciences and the humanities.

How BD might assist with the struggle of the human mind to overcome three notorious barriers:

- nonlinearity,
- non-locality and
- hyperdimensional spaces.

Big data: the end of the scientific method?

Sauro Succi[1,2] and Peter V. Coveney[3,4]

[1]Center for Life Nano Sciences at La Sapienza, Istituto Italiano di Tecnologia, viale R. Margherita, 265, 00161, Roma, Italy
[2]Institute for Applied Computational Science, J. Paulson School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, USA
[3]Centre for Computational Science, Department of Chemistry, University College London, London, UK
[4]Yale University, New Haven, USA

PVC, 0000-0002-8787-7256

For it is not the abundance of knowledge, but the interior feeling and taste of things, which is accustomed to satisfy the desire of the soul.

2019, Phil.Trans.R.Soc.A -Special issue ‘Multiscale modelling, simulation andcomputing: from the desktop to the exascale’.

How BD might assist with the struggle of the human mind to overcome three notorious barriers:

- nonlinearity,
- non-locality and
- … mensional spaces.

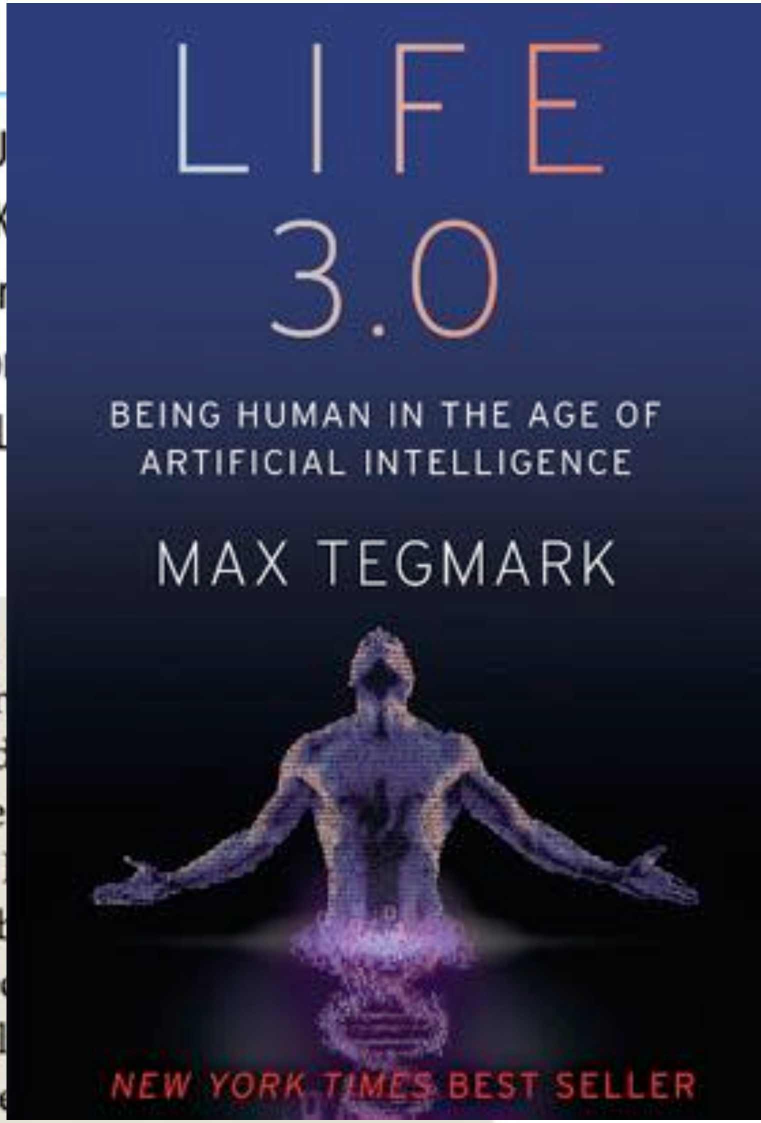
Big data need big theory too

Peter V. Coveney[1], Edward R. Dougherty[2] and Roger R. Highfield[3]

[1]Centre for Computational Science, U…
Gordon Street, London WC1H 0AJ, UK
[2]Center for Bioinformatics and Genor…
Texas A&M University, College Statio…
[3]Science Museum, Exhibition Road, L…

PVC, 0000-0002-8787-7256

The current interest in big … and data analytics has ger… impression that such method… most problems without the… scientific methods of inquiry. … is intensifying, accelerated b… digitized data can be acquire… endeavour, from science, heal… to economics, social science…



…ta: the end of the …ific method?

…i[1,2] and Peter V. Coveney[3,4]

…fe Nano Sciences at La Sapienza, Istituto Italiano di …iale R. Margherita, 265, 00161, Roma, Italy
… Applied Computational Science, J. Paulson School of …and Applied Sciences, Harvard University, 29 Oxford …ridge, USA
…mputational Science, Department of Chemistry, …llege London, London, UK
…ity, New Haven, USA

…-0002-8787-7256

… it is not the abundance of knowledge, but … interior feeling and taste of things, which …ccustomed to satisfy the desire of the soul. …nt Ignatius of Loyola).

…ue that the boldest claims of big data (BD)

Data Science ≠ ML

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

Acknowledgements

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

NACAD
High Performance Computing Center

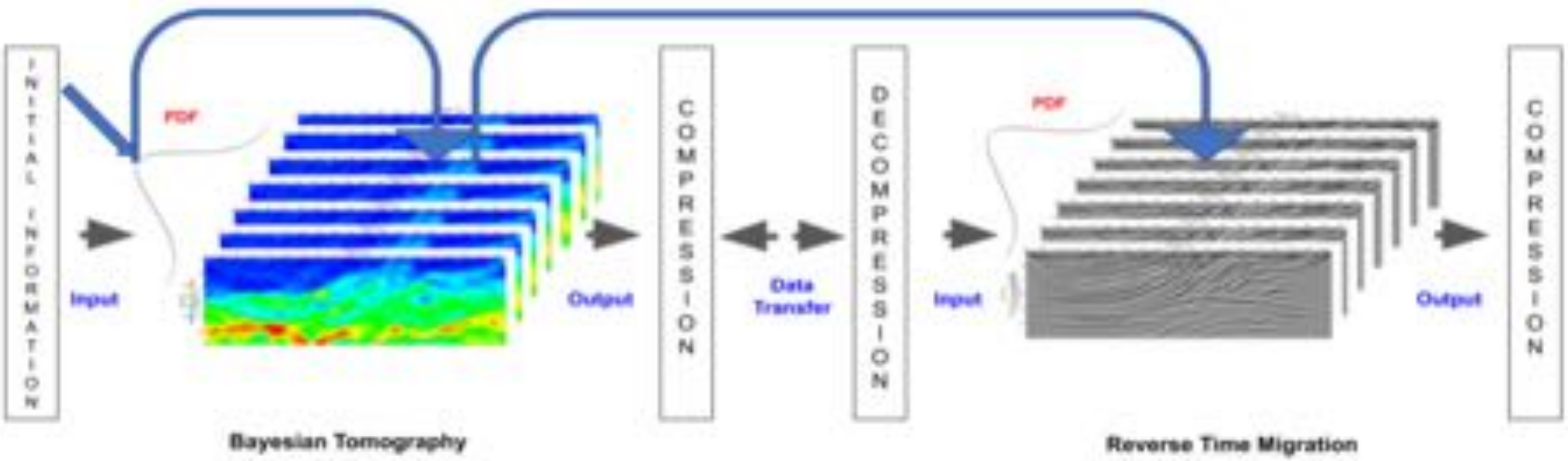
CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Inria
informatics mathematics

FAPERJ
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

CAPES

Semana PESC / Ciência de Dados - 17 de outubro de 2019



Obrigada!

Marta Mattoso

PESC
Programa de Engenharia de Sistemas e Computação

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