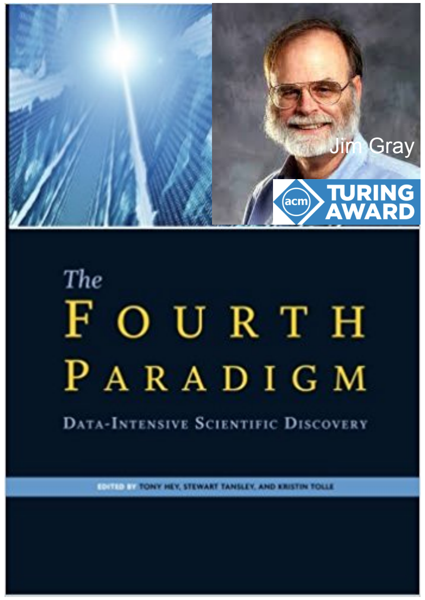
Data Intensive Science

Patrick Valduriez

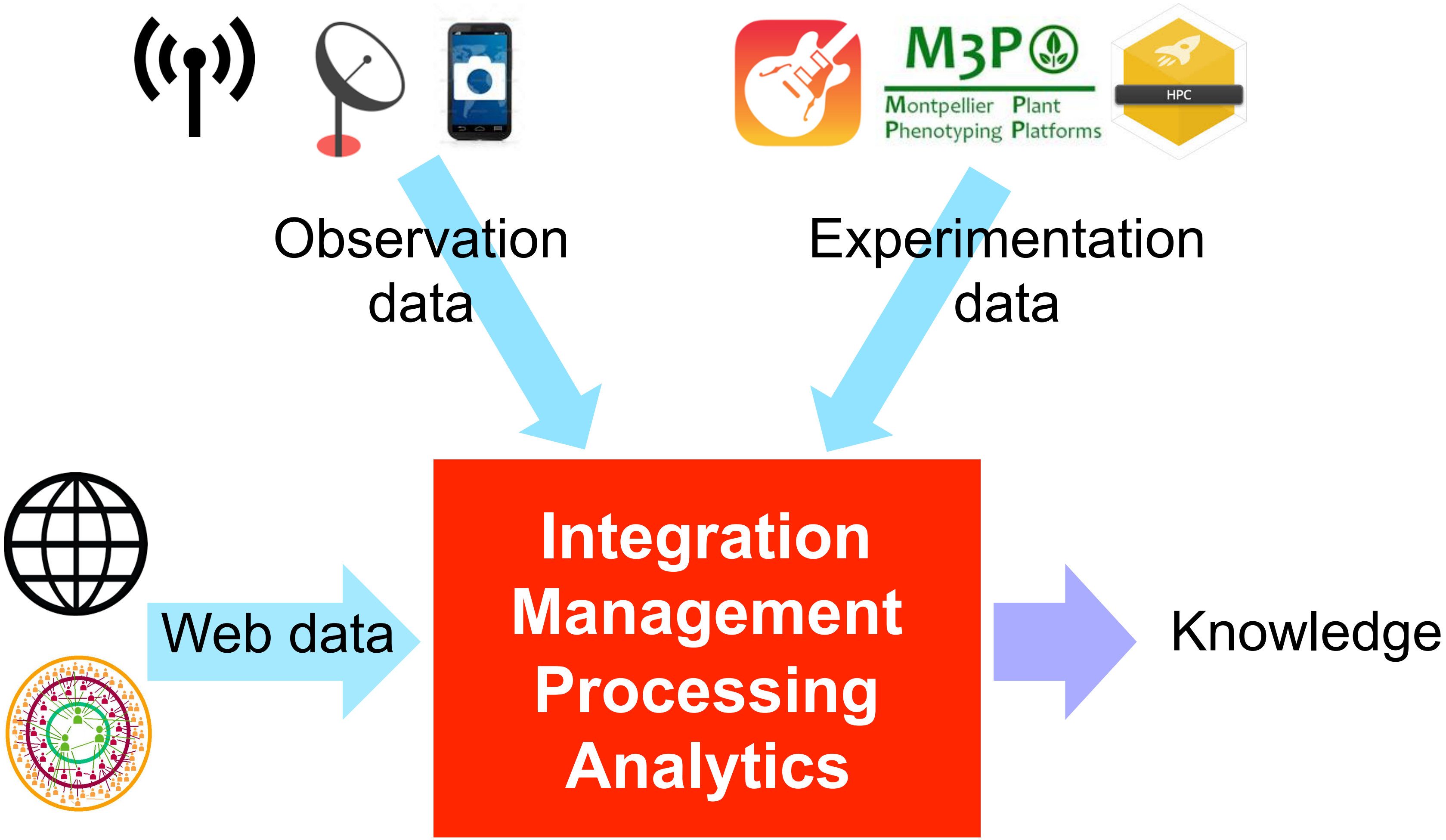
Data-intensive science


Jim Gray
acm TURING AWARD
The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

- Modern science such as astronomy, biology and computational engineering must deal with overwhelming amounts of data
  - Generated by sensors, scientific instruments or HPC simulation

- Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore these massive datasets

From Data to Knowledge



Observation data

Experimentation data

Web data

Integration
Management
Processing
Analytics

Knowledge

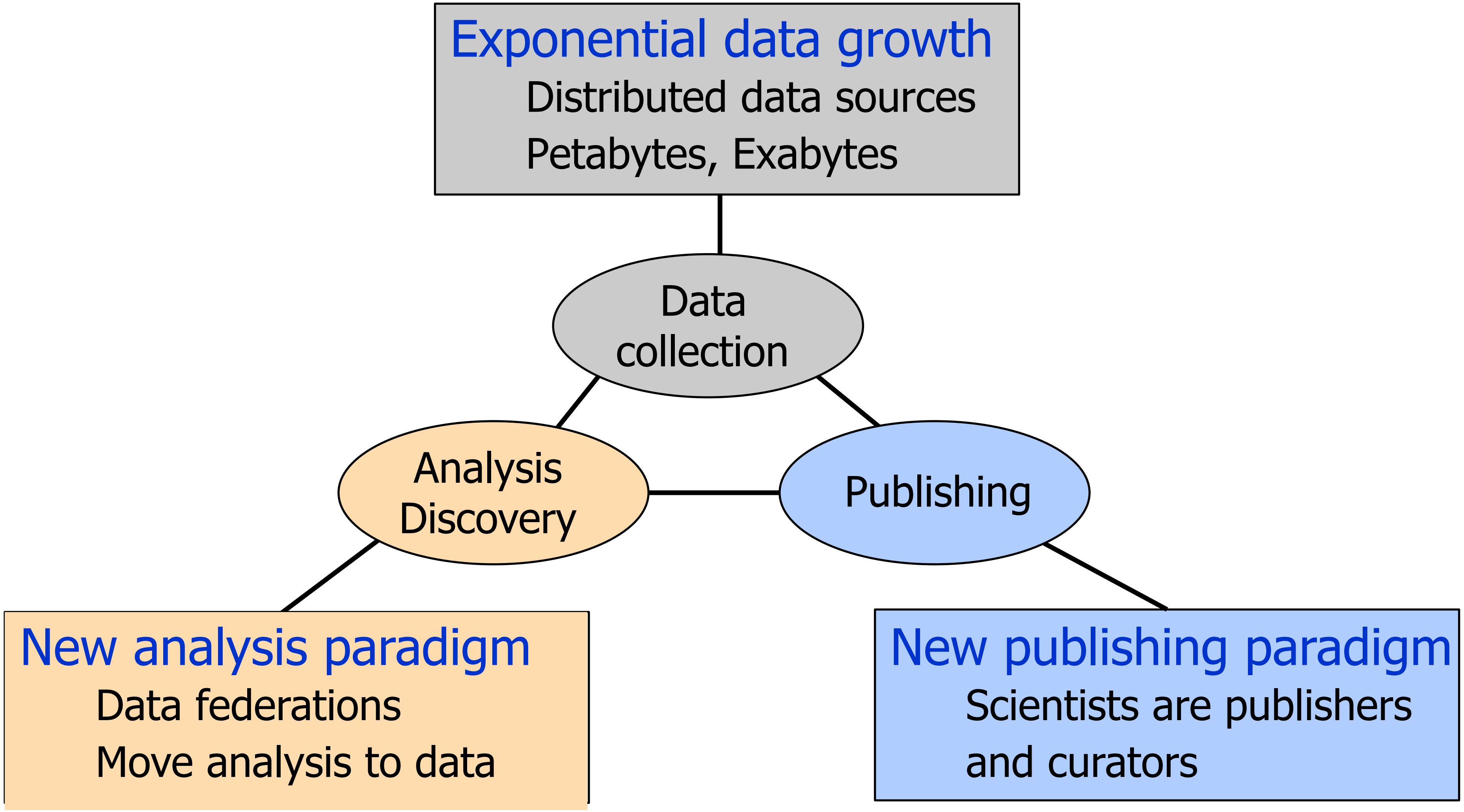
From Data to Knowledge

The major challenge

"*Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse*"

The Office Science Data Management Challenge
USA DoE 2004

New Paradigms for Scientists



Exponential data growth

Distributed data sources

Petabytes, Exabytes

Data collection

Analysis Discovery

Publishing

New analysis paradigm

Data federations

Move analysis to data

New publishing paradigm

Scientists are publishers and curators

Source: Jim Gray (ACM Turing Award)

Science Data Sharing

- Scientific databases
  - Astronomy (SkyServer), Biology (GenBank), etc.
- Web portals
  - HAL, GoogleScholar, DBLP, data.gouv.fr, AgroPortal, …
- Data storage & computing platforms
  - GENCI, LHC Computing Grid, Grid5000, PlanetLab, etc.
- Open science
  - Data papers
  - Overlay journals, e.g. episcience.org
  - Crowdsourcing platforms, e.g. GalaxyZoo, Telabotanica

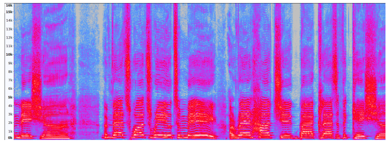
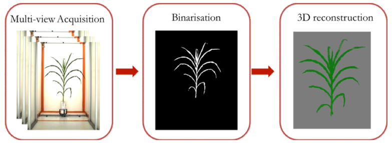
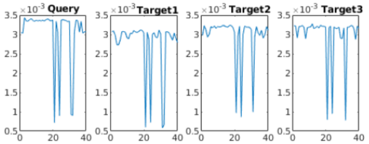
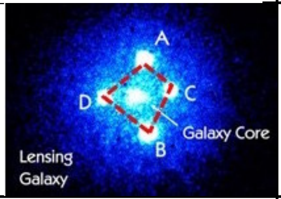
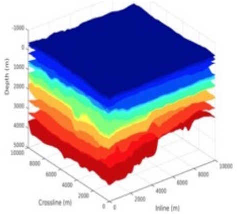
Impact on Scientific Practice

- Example in climate change
  - 97% of the papers in climate change research conclude that global warming is real
- But what about those 3% of papers that reach contrary conclusions?
- Answer
  - Learning from Mistakes in Climate Research. R.E. Benestad, et al. *Theoretical and Applied Climatology* 126: 699 (2016)
  - By the team of Katharine Hayhoe, director of the Climate Science Center at Texas Tech University

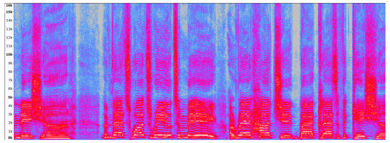
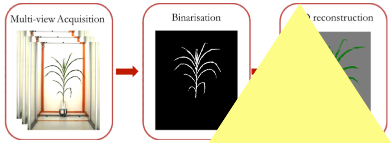
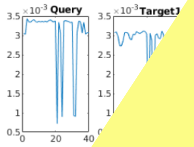
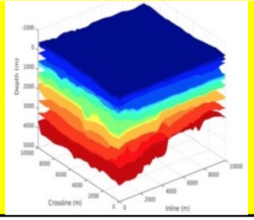
The Story

- The team tried to replicate the results of those 3% of sceptic papers
  - Looked at the 38 papers published in peer-reviewed journals in the last decade that denied global warming
- Findings
  - Every single one had an error, in their assumptions, methodology, or analysis
  - Many had cherry-picked the results that conveniently supported their conclusion, while ignoring other data
- Conclusion
  - Using the papers' data and after corrections of the errors, the results not only contradict the original papers but do support that global warming is real

Some Scientific Applications

Domain	Data	Examples	Partners	Contributions
Audio heritage	Audio & music recordings		Abbey Road, Musée de l'Homme	Music demixing Deep learning with time series
Agronomy – plant biology	Plant images, spatial data		CIRAD, INRA	Plant growth modeling and simulation with scientific workflows
Aircraft mechanic	Sensor data, continuous data		Safran	Indexing and querying of time series
Astronomy	Spatial data, geometrical forms	Einstein Galaxy core 	LNCC	Analysis of geometrical patterns with constellation queries
Biodiversity - botany	Plant images, descriptors, annotations		CIRAD, INRA, IRD, Telabotanica	Plant identification Crowd-sourced data production
Geoscience – oil & gas	Spatial data, measure-ments	3D Soil area 	LNCC, UFRJ, Petrobras, Repsol, Total	Simulation data analysis PDF computation

Some Scientific Applications

Domain	Data	Examples	Partners	Contributions
Audio heritage	Audio & music recordings		Abbey Road, Musée de l'Homme	Music demixing Deep learning with time series
Agronomy – plant biology	Plant images, spatial data		CIRAD, INRA	Plant growth modeling and simulation with scientific workflows
Aircraft mechanic	Sensor data, continuous data		Safran	Indexing and querying of time series
Astronomy	Spatial data, geometrical forms	Eins Ga c	S	Analysis of geometrical patterns with constellation queries
Biodiversity - botany	Plant images descriptors annotatior		A,	Plant identification Crowd-sourced data production
Geoscience – oil & gas	Spatial data, measure-ments	3D Soil area 	LNCC, UFRJ, Petrobras, Repsol, Total	Simulation data analysis PDF computation

Spotlight
in a moment

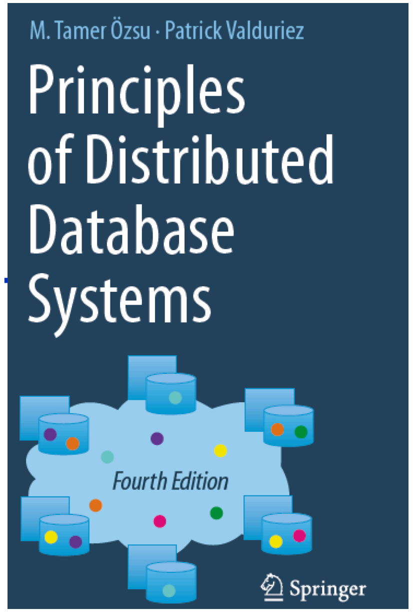
Scientific Data – *common features*

- Big data
  - Lots of data (both structured and unstructured)
  - Complex: multiscale, many dimensions, uncertainty
  - Heterogeneous: different formats, specific processes
- Processed through complex *workflows*
- Important *metadata* about experiments and their *provenance*
- Requires strong domain expertise
  - High-consequence interpretation errors

Approach

- Principles of distributed data management
  - Declarative languages, optimization, caching, indexing
  - Distributed and parallel data processing
- Highly distributed environments
  - HPC, cluster and cloud for scalability and performance
- Data science
  - Machine learning, statistics and data mining for high-dimensional data processing and analytics
- Extensive validation
  - By building high-quality software
  - Using real or synthetic datasets from application partners

Approach



M. Tamer Özsu · Patrick Valduriez

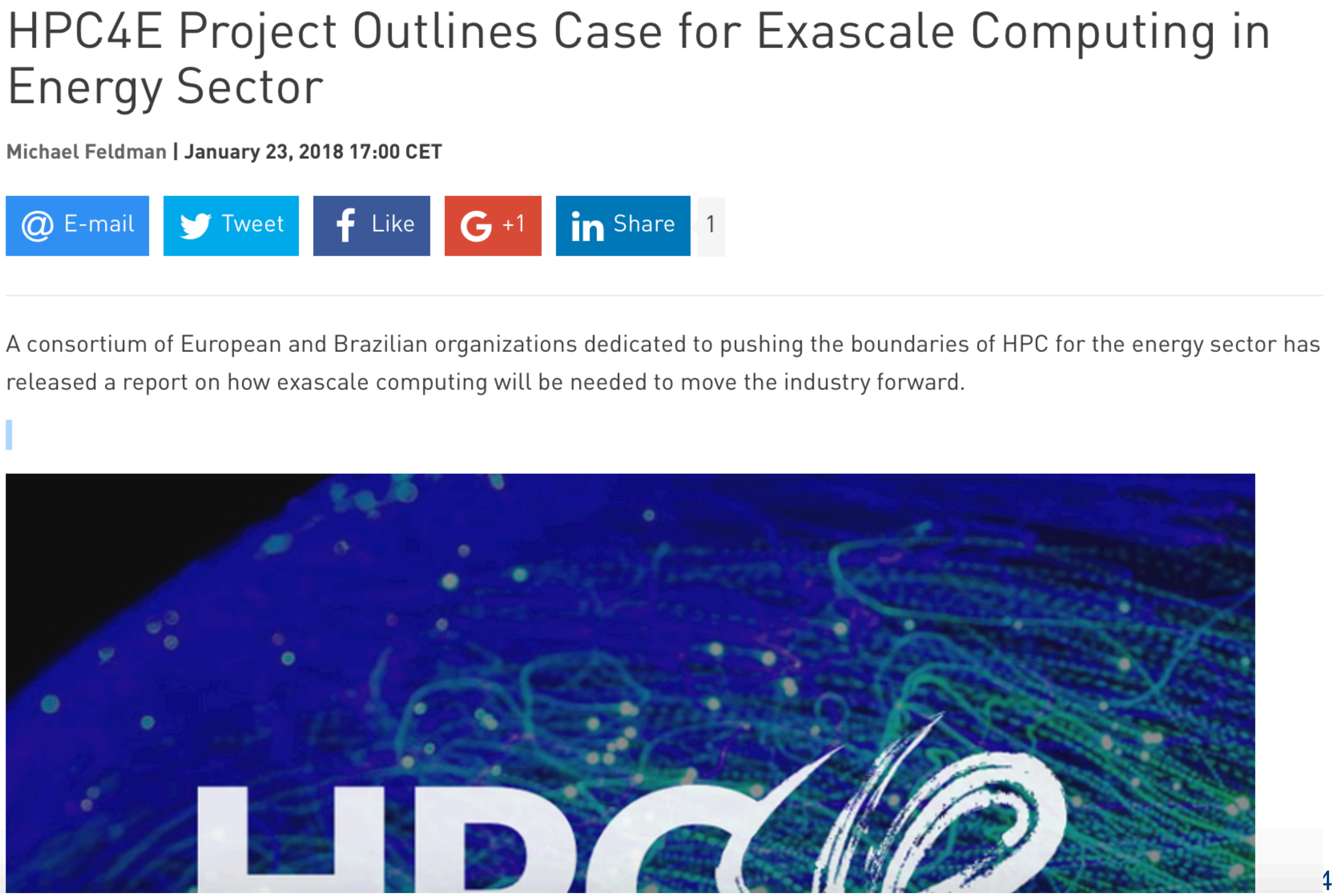
Principles of Distributed Database Systems

Fourth Edition

Springer

- Principles of distributed data managemen
  - Declarative languages, optimization, caching,
  - Distributed and parallel data processing
- Highly distributed environments
  - HPC, cluster and cloud for scalability and performance
- Data science
  - Machine learning, statistics and data mining for high-dimensional data processing and analytics
- Extensive validation
  - By building high-quality software
  - Using real or synthetic datasets from application partners

Zoom on HPC4E



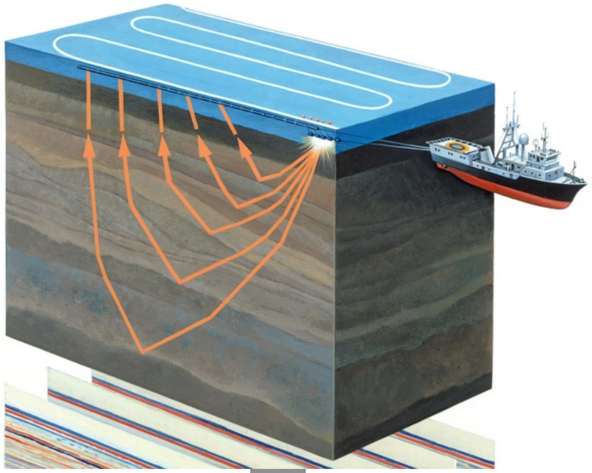
HPC4E Project Outlines Case for Exascale Computing in Energy Sector

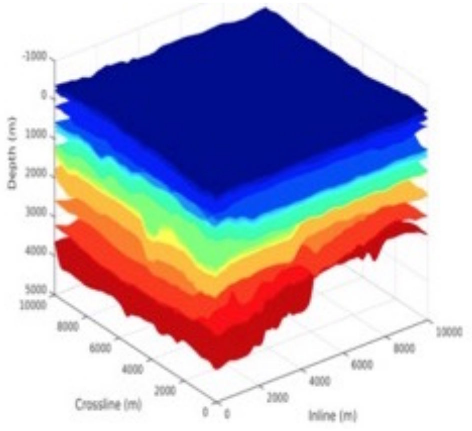
Michael Feldman | **January 23, 2018 17:00 CET**

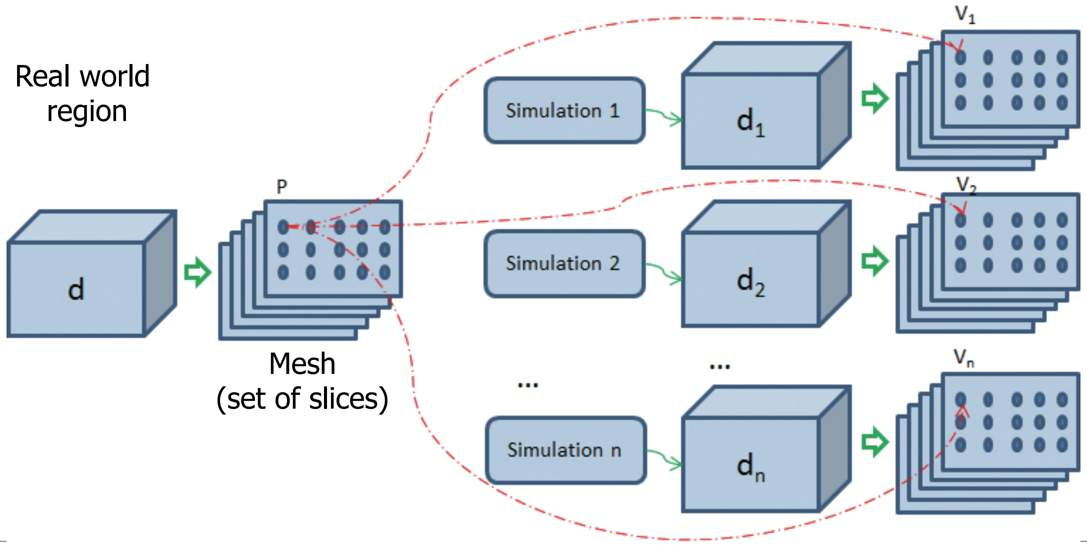
A consortium of European and Brazilian organizations dedicated to pushing the boundaries of HPC for the energy sector has released a report on how exascale computing will be needed to move the industry forward.

Context: seismic interpretation

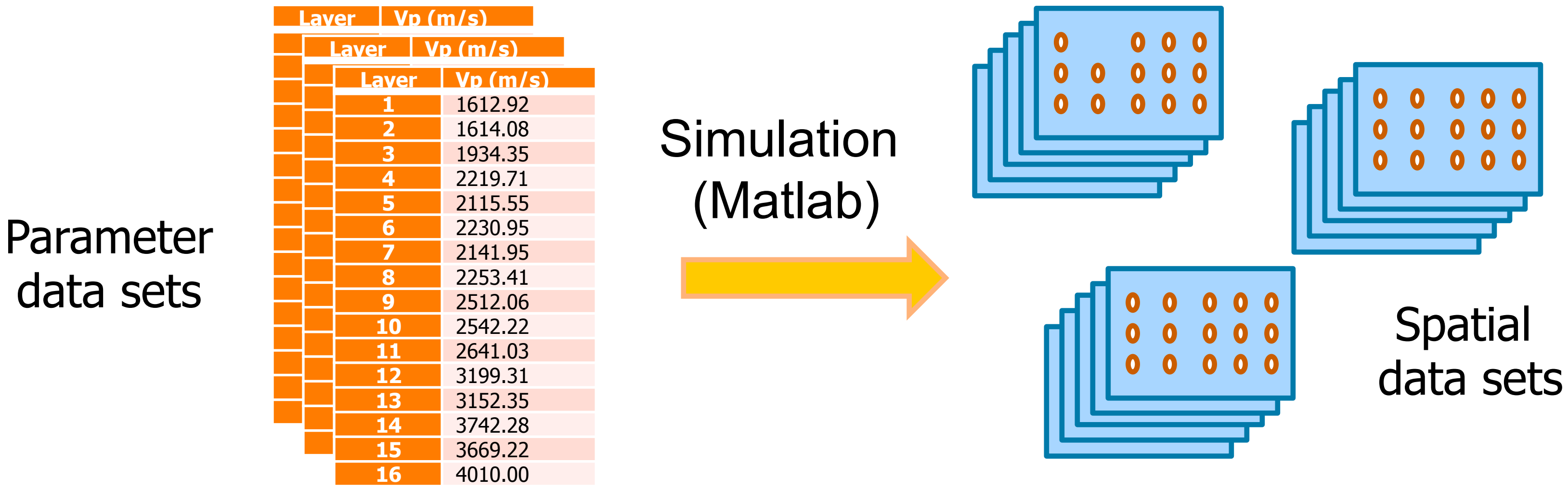
- Goal
  - Identify the different soil layers, e.g. oil reservoir, in an underground 3D area
- Observed data
  - From soil instruments that send signals to the underground and get back signals
  - Signals may have errors (noise)
- Solution
  - Uncertainty Quantification of the observed data using multiple simulations, with different parameters
  - Simulation data correspond to meshes that represent 3D soil areas







Simulation Data



Parameter data sets

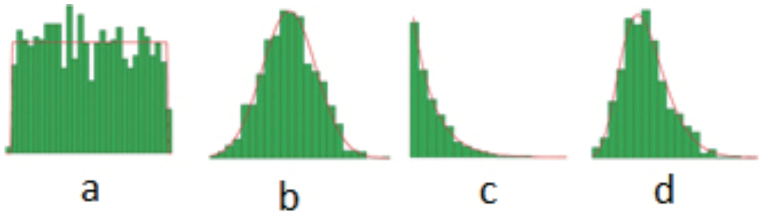
Layer	Vp (m/s)
1	1612.92
2	1614.08
3	1934.35
4	2219.71
5	2115.55
6	2230.95
7	2141.95
8	2253.41
9	2512.06
10	2542.22
11	2641.03
12	3199.31
13	3152.35
14	3742.28
15	3669.22
16	4010.00

Simulation (Matlab)

Spatial data sets

- One point in the cube area may correspond to different values in the spatial data sets
  - E.g. 3 values per point (see above)

- The set of values at a point may have four distribution types



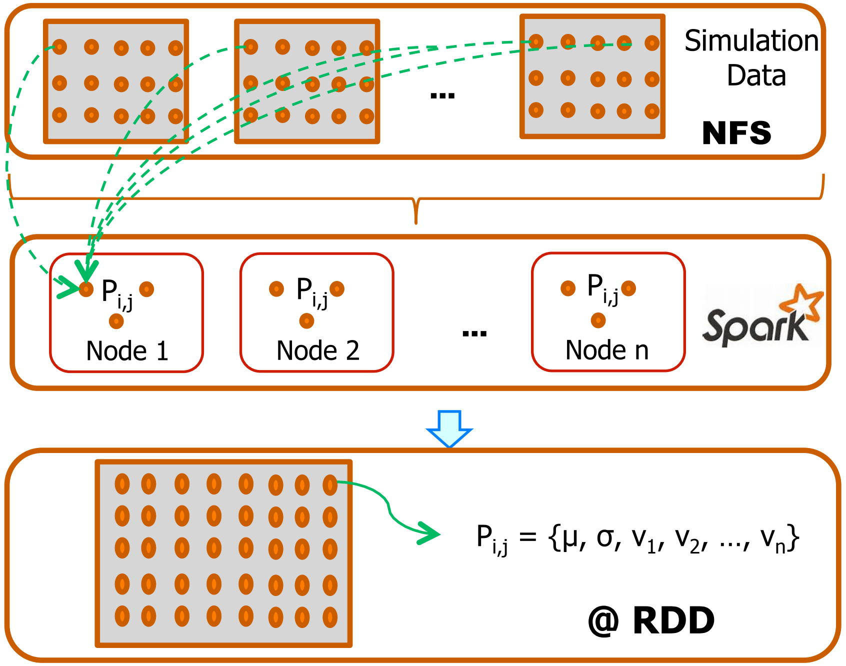
4 types of frequency per value distribution of a point

Probability Density Function (PDF)

- To quantify the error between the simulated data and the observed values, we calculate a PDF
  - A curve, with a distribution type and parameters
- Error = the difference between the PDF and the set of simulated values *V*
  - Computed by comparing the probability of the values in different intervals in *V* and the probability computed according to the PDF
- We need to calculate the PDF of each point per slice of the cube
  - Takes much time, e.g. days to process 2.4 TB data for an area of 10km (distance) * 10km (depth) * 5km (width)

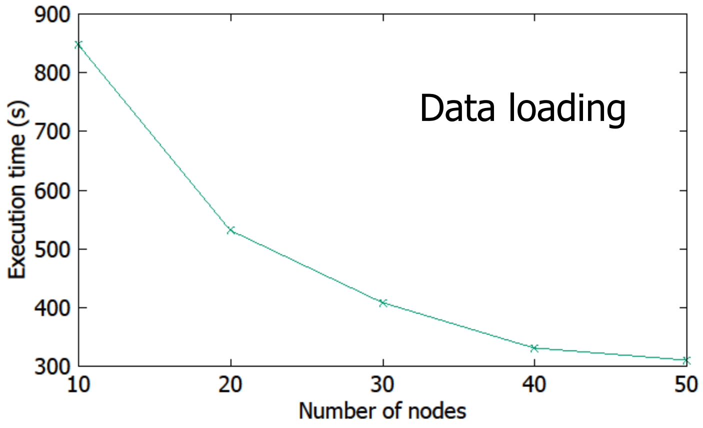
Approach with Spark

- Load the simulation data from NFS to a Spark cluster
  - Each point's values distributed at different nodes
- Parallel execution of different points
  - External programs (Java and R) to generate PDF of each point
- Optimizations
  - Aggregation of different points and reuse of results
  - Machine learning (decision tree) to predict distribution
  - Sampling and point aggregation to predict the distribution

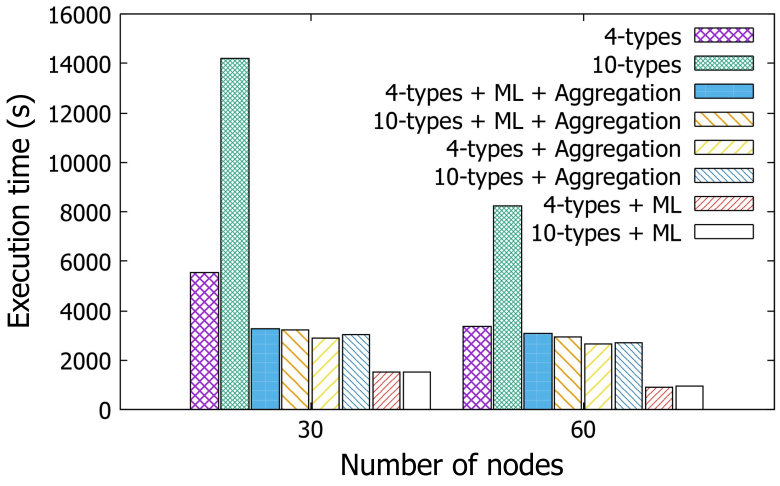


Results

- Validation with big simulation data (2.35 Terabytes)
  - Ji Liu, Noel Lemus, Esther Pacitti, Fabio Porto, Patrick Valduriez. Parallel Computation of PDFs on Big Spatial Data Using Spark, DAPD, 38 p, 2019.



- Experimental setting
  - Data from HPC4e benchmark
  - 10K simulations, which generate 10K values per point
  - Grid5000 cluster with 60 16-core nodes
- Experimental results
  - Linear scalability
  - Major performance improvement
    - Aggregation + ML up to 33
    - From 4 hours (base line) down to 15mn (best method)



Zenith