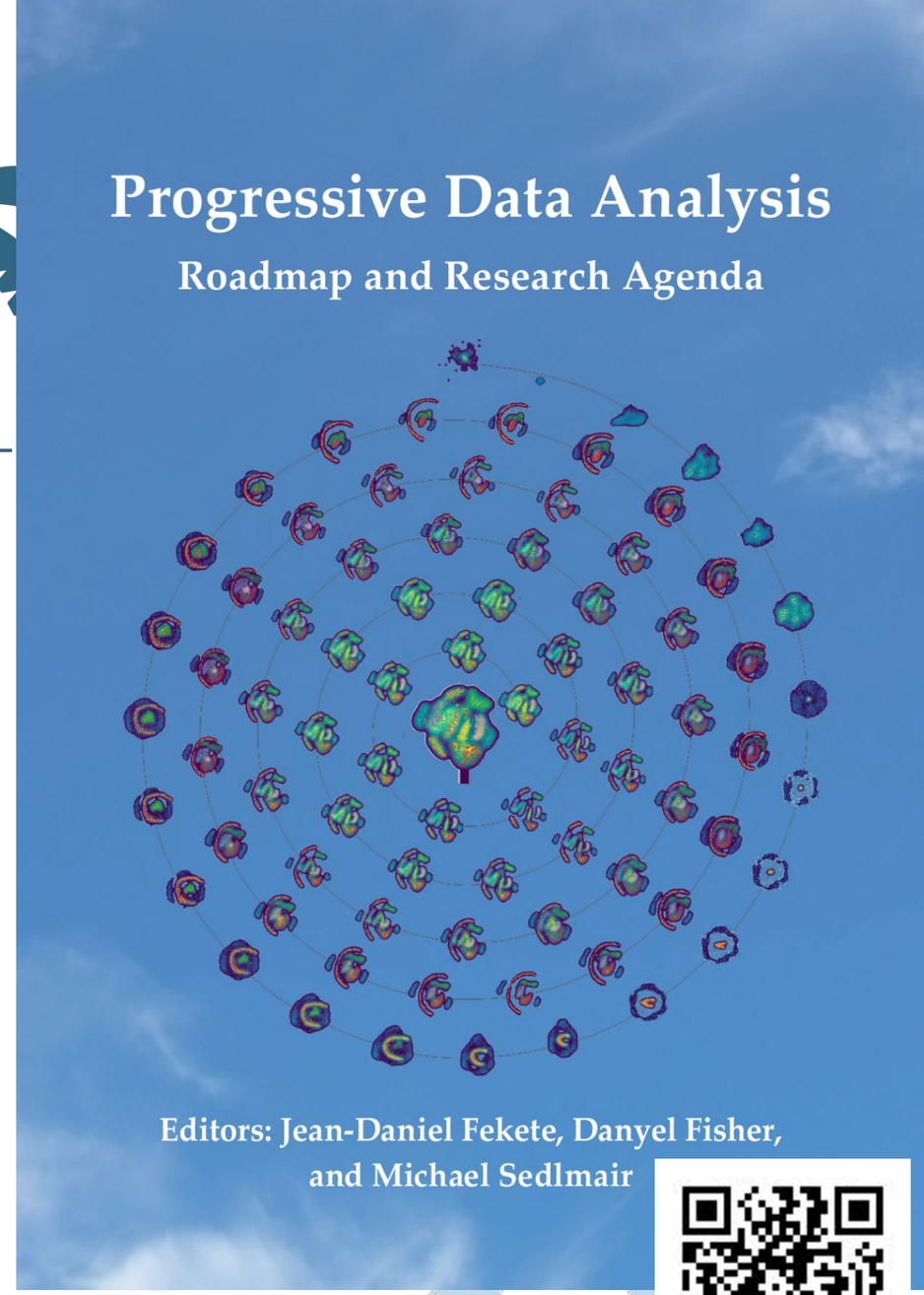
VIS2024

Progressive Data Analysis
Roadmap and Research Agenda

Eurographics Association (open book)
ISBN: 978-3-03868-270-7
https://doi.org/10.2312/pda.20242707



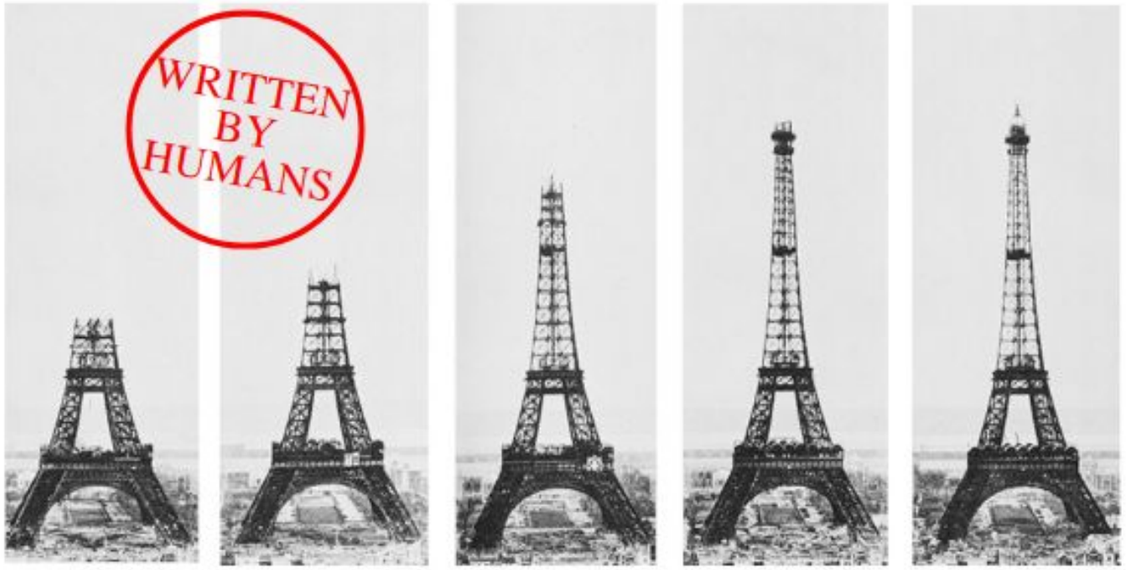
https://www.aviz.fr/Progressive/PDABook

Progressive Data Analysis

Roadmap and Research Agenda

Editors: Jean-Daniel Fekete,
Danyel Fisher, and
Michael Sedlmair

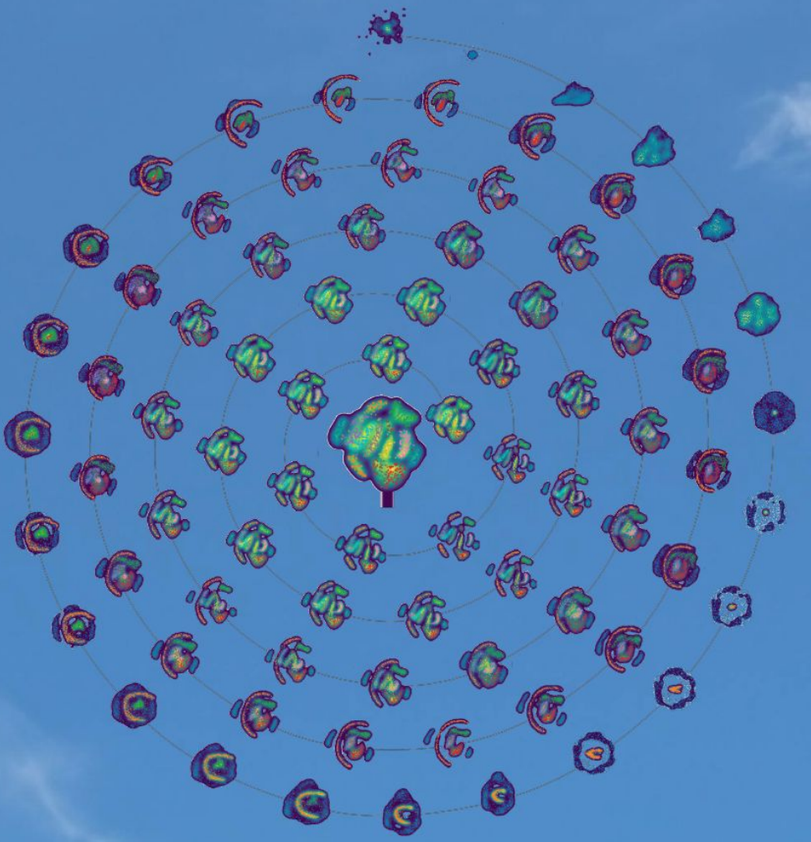
October 13, 2024



VIS2024

Progressive Data Analysis

Roadmap and Research Agenda



Editors: Jean-Daniel Fekete, Danyel Fisher,
and Michael Sedlmair

Book Authors

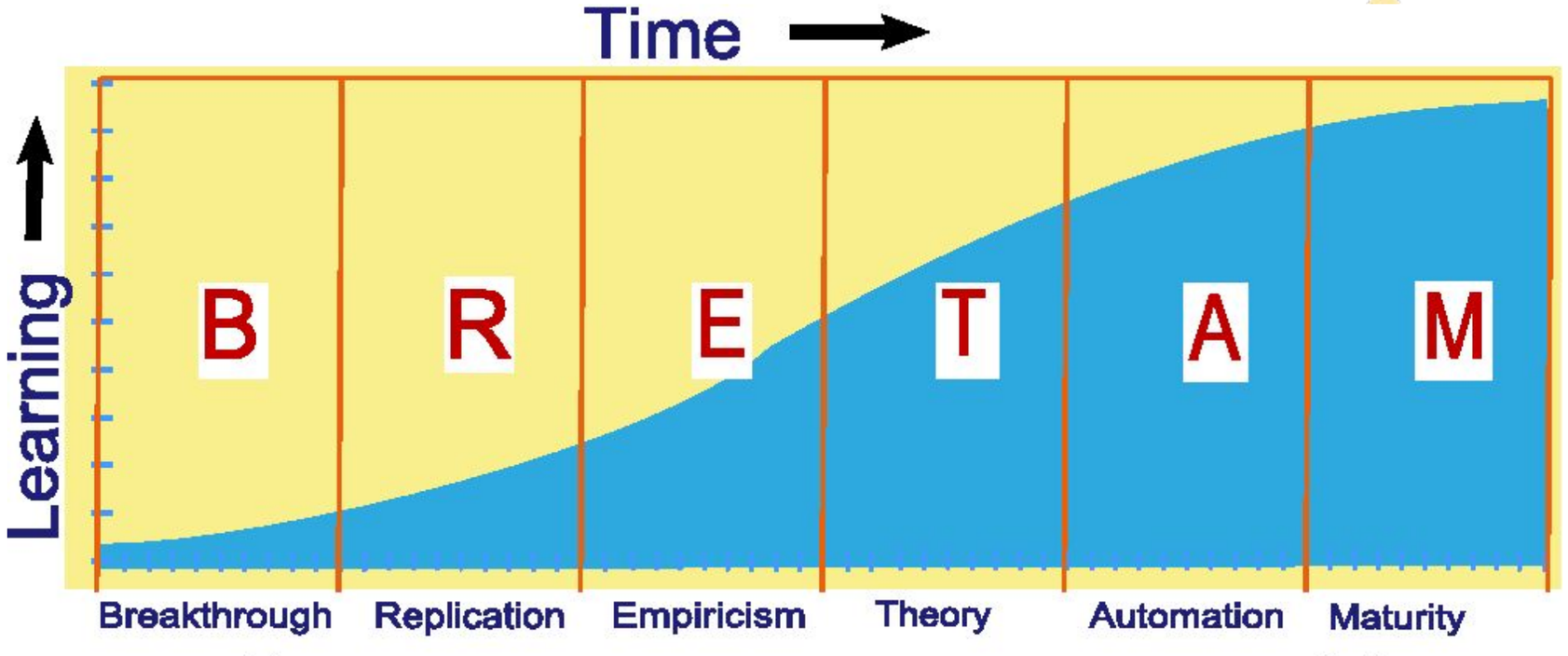
- Participants of [Dagstuhl Seminar 18411]
  - October 2018

Marco Angelini	Sapienze Univesità di Roma
Michaël Aupetit	Qatar Computing Research Institute
Sriram Karthik Badam	Apple Inc.
Carsten Binnig	Technical University of Darmstadt & DFKI
Jean-Daniel Fekete	Inria & Université Paris-Saclay
Danyel Fisher	
Barbara Hammer	CITEC, Bielefeld University
Jaemin Jo	Sungkyunkwan University
Nicola Pezzotti	Philips Cardiologs, TU/e, AI4MR
Gaëlle Richer	Inria & Université Paris-Saclay
Florin Rusu	University of California Merced
Giuseppe Santucci	Sapienze Univesità di Roma
Hans-Jörg Schulz	Aarhus University
Michael Sedlmair	University of Stuttgart
Hendrik Strobelt	IBM Research, MIT-IBM Watson AI Lab
Cagatay Turkay	University of Warwick
Anna Vilanova	TU/e
Chris Weaver	University of Oklahoma

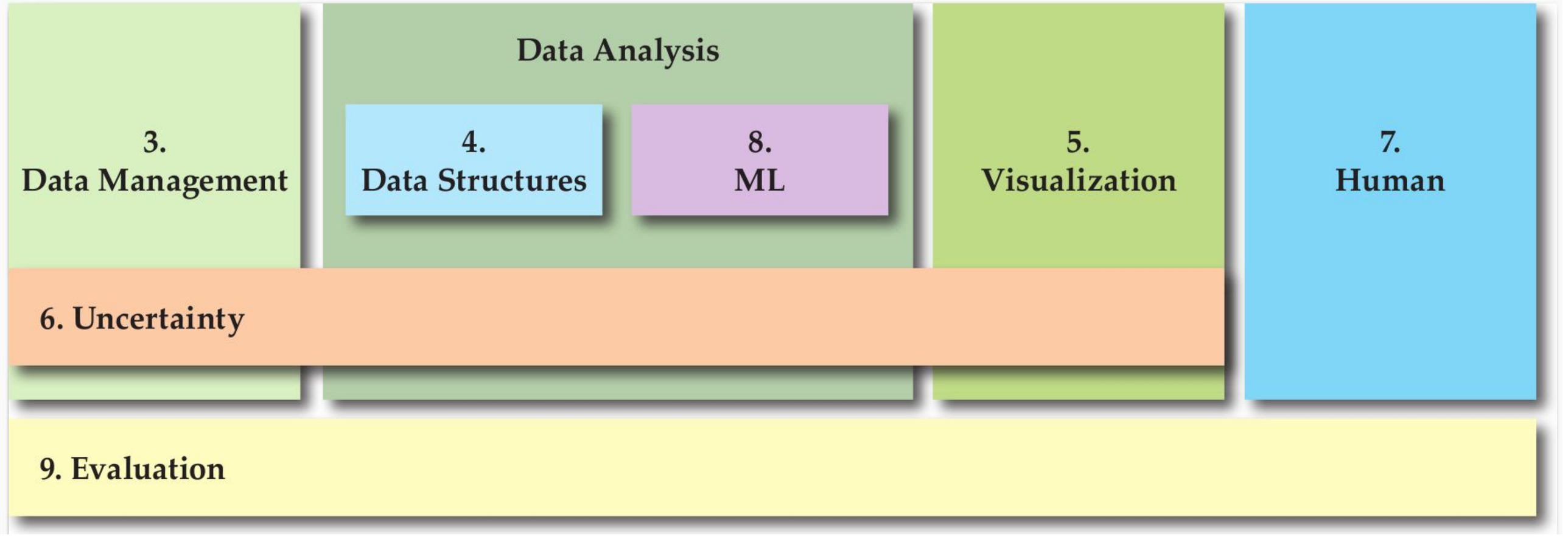
PDA+V: Where are we now?

Apply the BRETAM Model [Gaines 91]

1. Breakthrough
2. Replication
3. Empiricism
4. Theory
5. Automation
6. Maturity



Structure of the book



Chapters

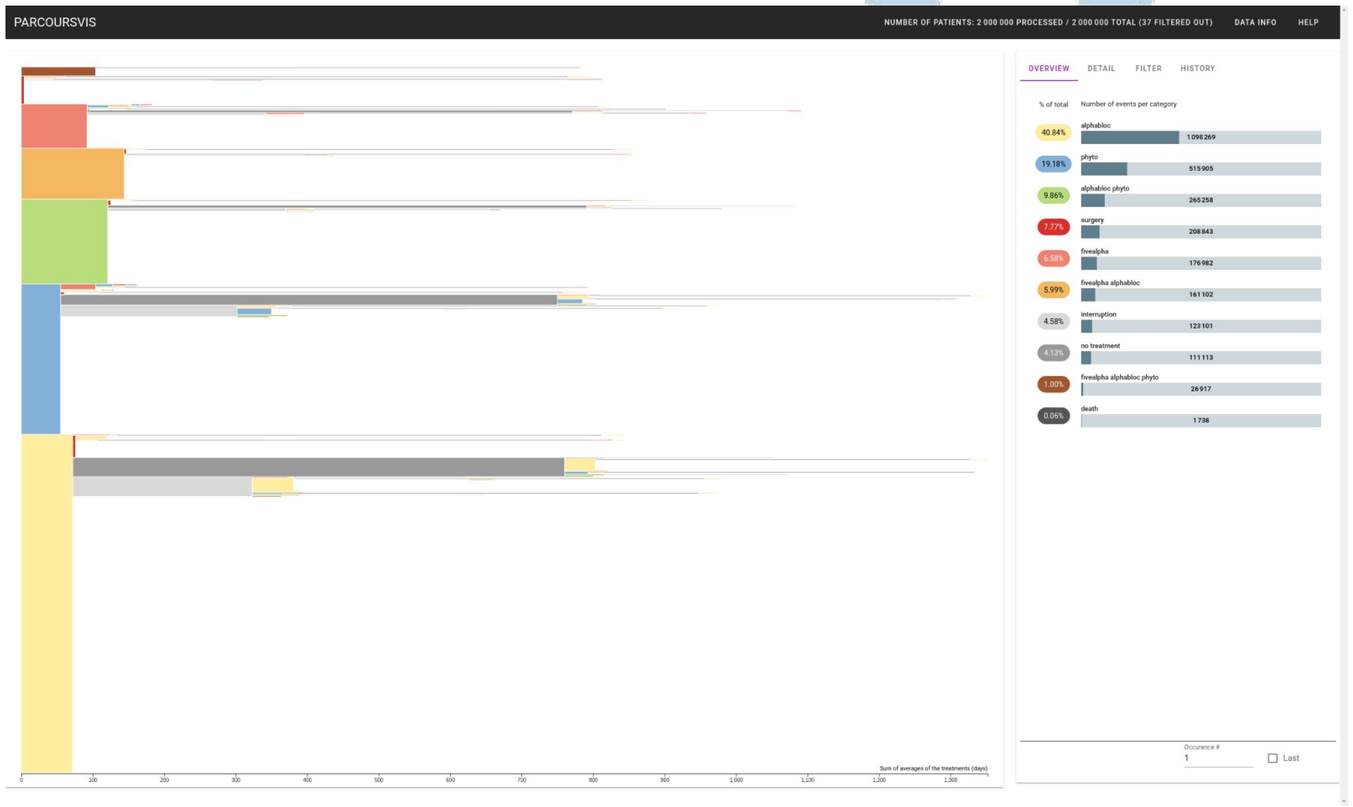
1. Introduction
2. Concepts and Definitions
3. Data Management
4. Data Structures and Algorithms
5. Visualization
6. Uncertainty and Quality
7. Human Aspects
8. Machine Learning
9. Evaluation
10. Challenges and Research Agenda

Highlights

- PDA is truly scalable with controlled latency!
  - But it requires feedback on quality
- PDA often confused with other concepts (streaming, online, …)
  - Misunderstandings from reviewers and colleagues
- Paradigm shift, incompatible with imperative programming
- No infrastructure/language yet
- Most algorithms can be adapted to PDA
- Most visualizations can too (come see our survey on Thursday)
  - Beware of stability!
- Uncertainty is challenging
- Many exciting scientific and technical challenges
  - No other paradigm can achieve the scalability needed to work on real problems

PDA is truly scalable with controlled latency!

- Many applications of PDA reach new orders of magnitude
  - E.g., billion medical events
- With smooth interaction



PDA often confused with other concepts

Chapter 2 of the book clarifies the differences

- Online
- Iterative
- Incremental
- Streaming
- Real-Time
- Anytime
- Progressive
- Approximate Query Processing

Make sure you clarify the term or it will steer misunderstandings!

Paradigm shift, incompatible with imperative programming

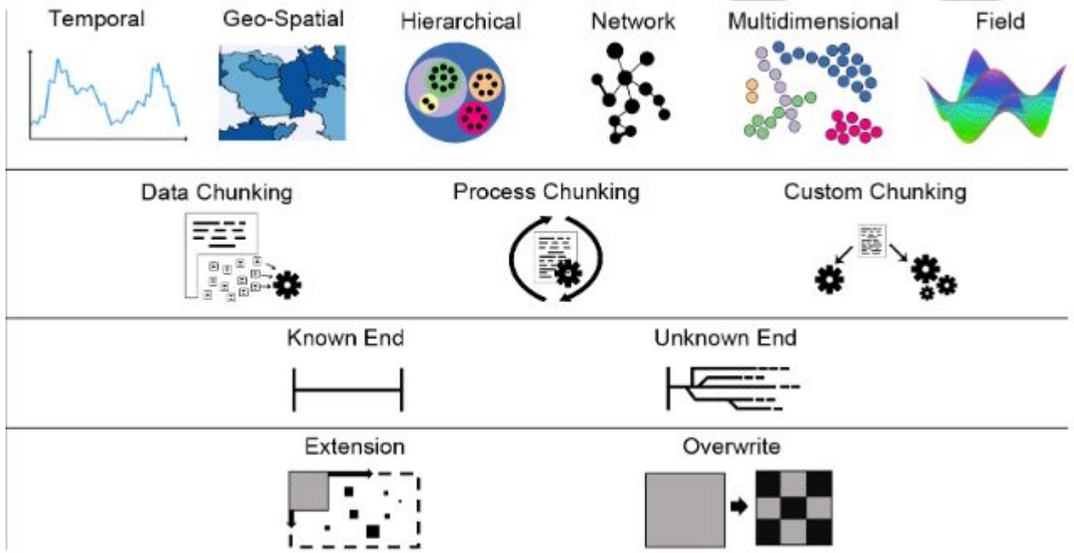
- It would be great to have PDA a another library
  - mix with the other ones and run everything progressively!
- But no, it is a programming paradigm shift
- Will need to build everything from the ground up
  - No infrastructure/language yet
- Not always difficult
  - but need to break down existing libraries to reach to lower-level code

Many algorithms can be adapted to PDA

- Increasing literature on progressive algorithms
- Trade offs are necessary
  - progressive by chunk vs. progressive by iteration
- Missing standard library of progressive data structures
  - no interoperability between algorithm implementations

Most visualizations can be adapted to PDA

- Come see our survey on Thursday
- Many have already been adapted
- Beware of stability!



Uncertainty is challenging

- Assessing the quality of a "progressive pipeline" is hard in general
- Quality and uncertainty are not easy to combine on pipeline stages
- Stability is a simple proxy of quality, but approximate
  - if the progressive computation is not stable, the results are not good
  - if it is stable … maybe
- More research is needed to generalize uncertainty computation over general progressive pipelines

Many exciting scientific and technical challenges

- No other paradigm can achieve the scalability needed to explore data with guaranteed latency
- Important challenges to reconnect with the scale of other fields
  - databases
  - AI/ML
  - simulation

Challenges and Research Agenda

PDA systems work on "pipelines" of processes, data flowing from input, computation, visualization, interaction, results, and reports.

- **C&RA**: We need to build infrastructures for managing this PDA pipeline, providing consistent mechanisms throughout
  - co-design with the database, the algorithms, and the visualization
- **C&RA**: In a steerable system, the user input must conversely flow backward through the system to drive different computations.
- **C&RA**: We need algorithms and user experiences that are aware of the concept of stability, and of the tradeoffs between stability and quality

Challenges and Research Agenda

- **RA**: Evaluation should consider three different deadlines:
  - the first meaningful progressive result,
  - the earliest result accurate for decision-making, and
  - the total computation time if needed
- **C**: Steering should also be considered for evaluation.
  There is currently no standard methodology to take it into account
- **C&RA**: PDA prefers shuffled data to ensure fair sampling,
  how can we do it for general big data?
- **RA:** New conventions and standards should be designed for progressive databases.
- **RA:** Query steering is important for exploration in PDA, but is far from standard in data management system

Agenda for Tomorrow

- To achieve progressive systems, we must agree on a common communication mechanism between progressive functions and modules. **Create it!**
- What is the next progressive SQL? **Create it!**
- Languages like Python and JavaScript do not support progressive data structures and algorithms. **Create them!**
- Visualization libraries and not designed for PDA. **Fix them!**
- What would a standard GUI/Notebook for PDA look like? **Create it!**
- Collaboration is required between DB, Stats, ML, Simulation, Visualization, and HCI to find suitable solutions and create working systems. **Break the boundaries between academic domains!**

https://www.aviz.fr/Progressive/PDABook