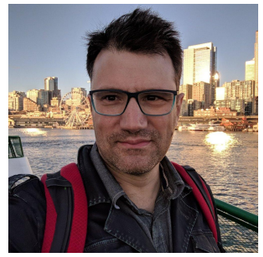
Trip Report: Dagstuhl Seminar on Progressive Data Analysis and Visualization

Jean-Daniel Fekete, Inria
Michael Sedlmair, Univ. of Stuttgart

Progressive Data Analysis and Visualization

- Seminar took place October 7 – 12 , 2018 (2 weeks ago)
- Organizers:
  - Jean-Daniel Fekete, Inria, FR
  - Danyel Fisher, Honeycomb, US
  - Arnab Nandi, Ohio State University, US
  - Michael Sedlmair, Univ. Stuttgart, DE
- About 30 participants from:
  - Visualization
  - Database
  - Machine-Learning

VIS 2018

https://www.dagstuhl.de/18411

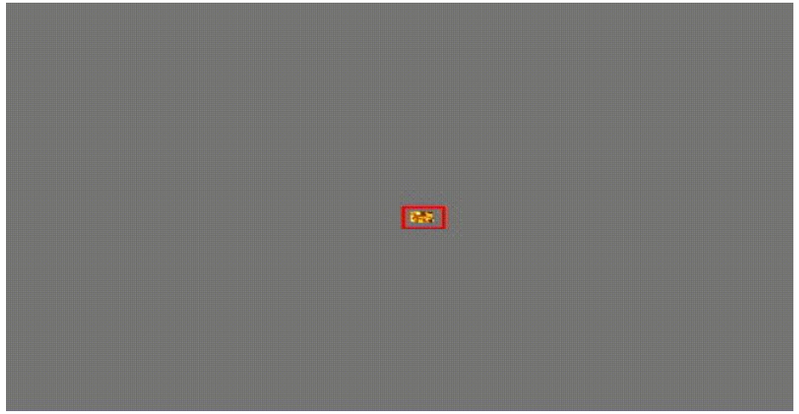
VIS 2018

What's the problem?

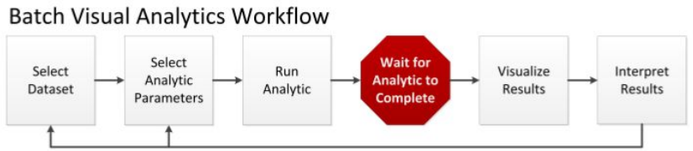
- New infrastructures allow big-data analysis at scale
  - but only for predictive computation
- Exploratory Data Analysis not well supported
  - Iterative loop should be kept under bounded latency
    - 0.1s for animation, 1s for interface reaction, 10s for results
  - When data grows or computations become more complex, existing systems exceed these latencies
- Progressive Data Analysis is meant to address the latency issue
  - Maintain the analyst’s attention
  - Allow early decision

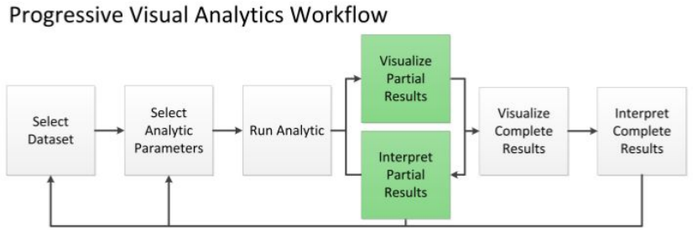
Progressive Computation: What is it?

- A computation
  - bounded on time and data
- which reports intermediate outputs
  - a result,
  - a measure of quality,
  - a measure of progress
- within bounded latency
- converging towards the true result
- controllable by a user during execution
  - abort, pause, etc.
  - parameters (steering)



Williams, M.; Munzner, T., "Steerable, Progressive Multidimensional Scaling," in *INFOVIS 2004.*





Charles D. Stolper, Adam Perer, and David Gotz. **Progressive Visual Analytics**. *IEEE TVCG* (Volume 20, Issue 12, 2014).

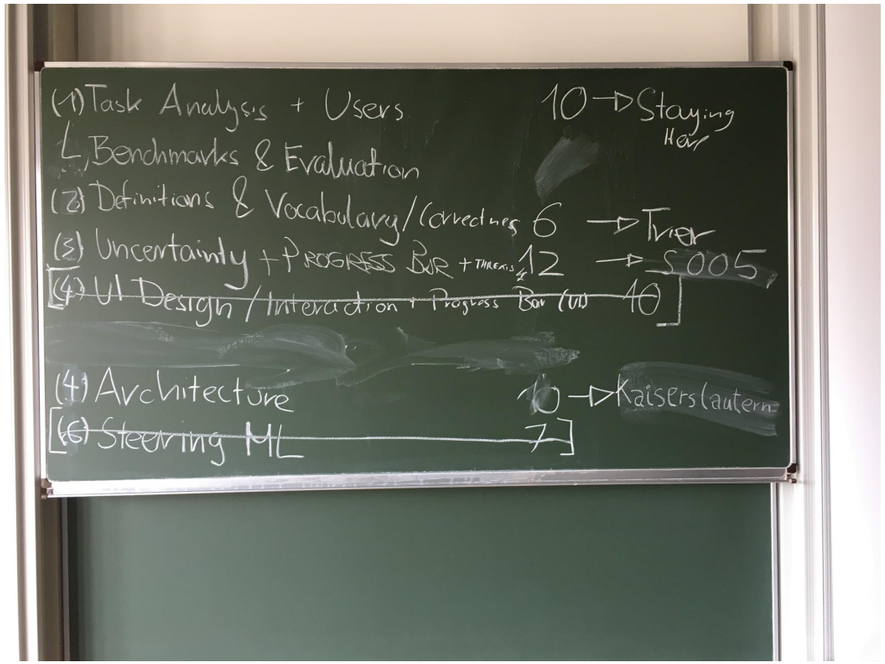
Organisation of the Seminar

- 5 Lightning Talks
- Short success stories presentations
- 2 sessions of breakout groups
  - 2 half days of group discussion per session
  - Report after each half day (except 1)
  - First session groups decided on day 1
  - Second session groups decided on day 3
- Final discussion
- Decision of outcomes

VIS 2018

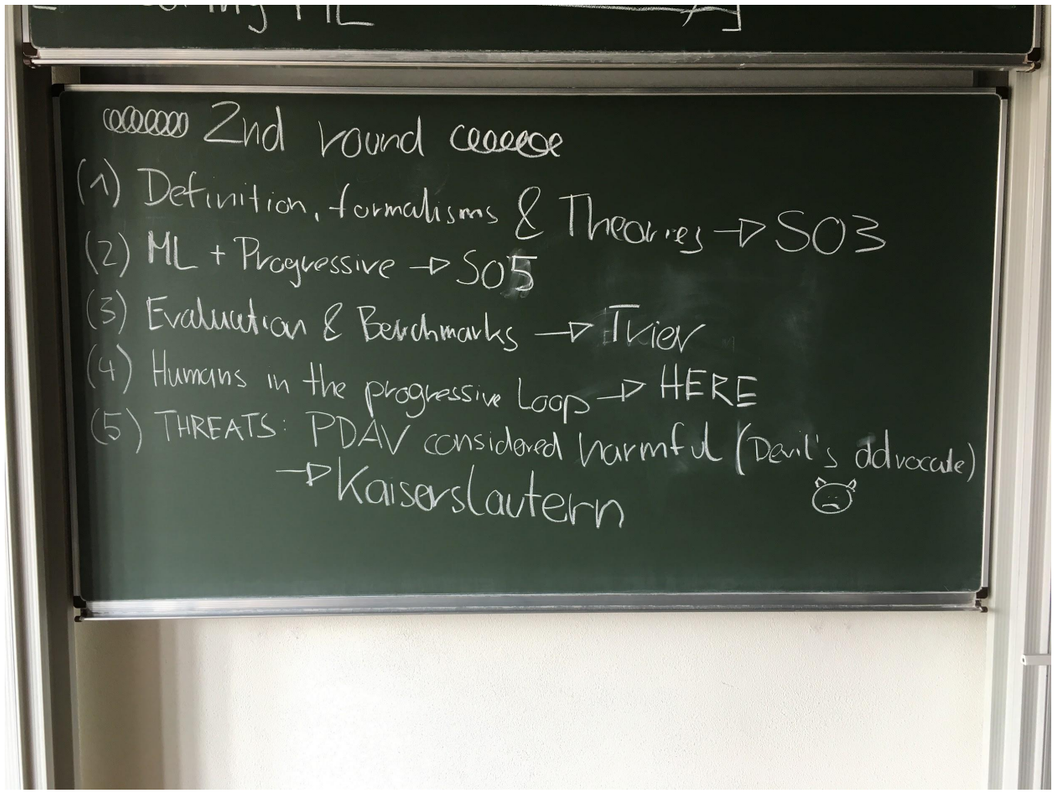
First Session Groups

- Definition & Vocabulary
- Task Analysis & Users
- Architecture
- Uncertainty



Second Session Groups

- Definition (cont.)
- Evaluation & Benchmarking
- Human in the Progressive Loop
- Machine Learning
- Threats

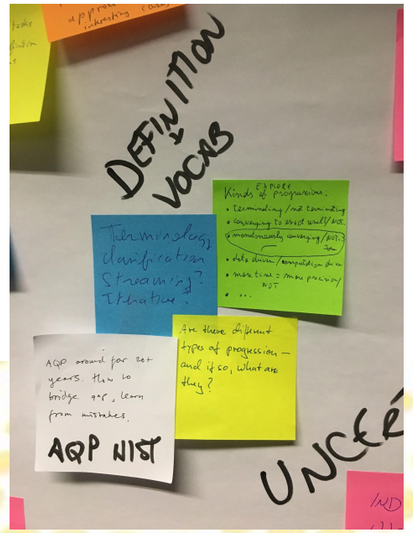


Definition & Vocabulary

Michael Aupetit, Christopher Jermaine, Jaemin Jo, Gaëlle Richer, Giuseppe Santucci, Hans-Jörg Schulz, Chad Stolper

	Input	Desirable Result R	Cmp Time t	Alg Params P	Quality Q	Progress π
Eager	$F_e(S, P, D)$	Exact R_D	t_D	Fixed	Baseline	Irrelevant
AQP	$F_a(S, P, D, \theta)$	Apx $d(R, R_D) \geq 0$	$\leq \theta$	Fixed	Unbounded	Irrelevant
Optimistic	Composed: AQP and eager	$R_2 = R_D$	$\leq t_D + \theta$	Fixed	Unbd→Baseline	Explicit
Online	$F_o(S_z, P, \bigcup C_{1,\ldots,z})$	Exact for $z = n_C$	$\ll t_D$	Fixed	Baseline	Implicit
Streaming	$F_s(S_z, P, C_z)$	Apx $d(R_z, R_D) \geq 0$	$\ll t_D$	Fixed	Unbounded	Implicit
Iterative	$F_i(S_z, P, D)$	$R_z \xrightarrow[z\to\infty]{} R_D$	Unbounded	Fixed	Unbounded	Implicit
Progressive	$F_p(S_z, P_z, D_z, \theta)$	$R_z \xrightarrow[z\to Z]{} R_D$	$\leq \theta$	Ctrled	Unbounded	Explicit

Table 1: Comparison of computation methods properties.
t_D is time taken on full dataset D.
$\bigcup_{z\in\{1\ldots n_C\}} C_z = D$ and $C_i \cap C_j = \emptyset$.
$D_i \subseteq D$.
θ is a user-defined upper-bound on latency
Output is always: $(S_{z+1}, R_z, t_z, Q_z, \pi_z)$



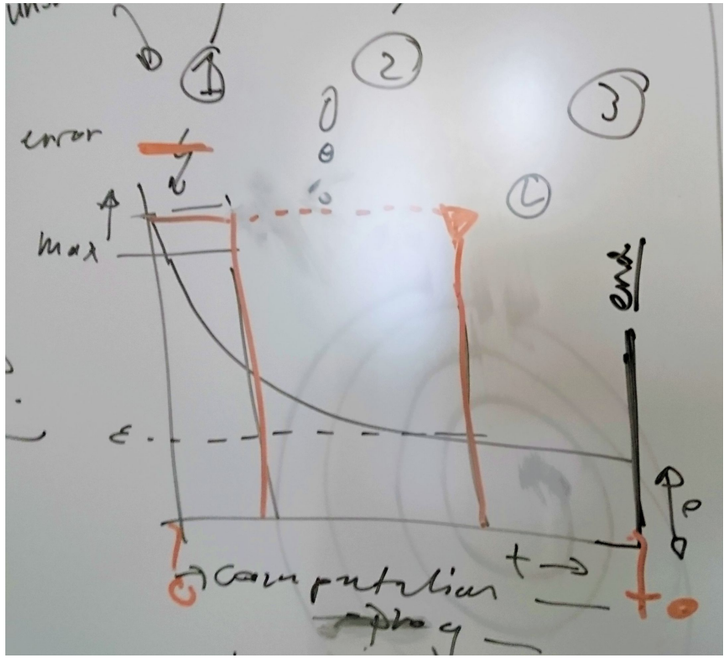
VIS 2018

Uncertainty

Marco Angelini, Sriram Karthik Badam, Barbara Hammer, Christopher M. Jermaine, Hannes Mühleisen, Cagatay Turkay, Frank van Ham, Anna Vilanova, Yunhai Wang

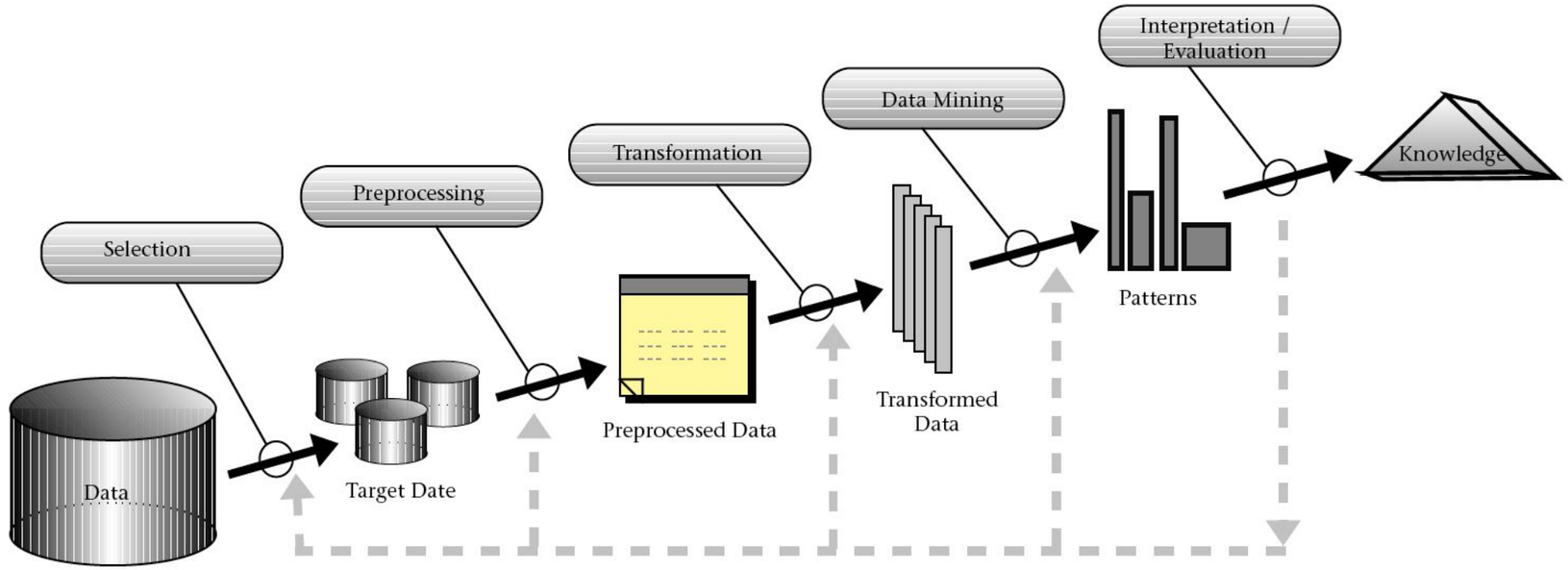
- 4 stages
  - don’t even look -- chaos
  - don’t trust -- might be interesting though
  - ready to make decision
  - waste of user & computer time

  [Angelini et al.]



Machine Learning

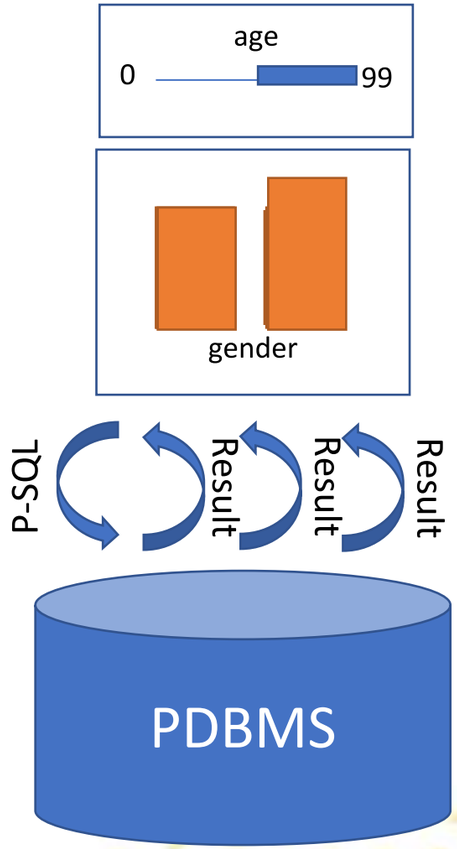
Nicola Pezzotti, Barbara Hammer, Daniel A. Keim, Cagatay Turkay, Yunhai Wang, Themis Palpanas, Florin Rusu, Carsten Binnig, Hendrik Strobelt



Architecture

Chad Stolper, Chris Weaver, Florian Rusu, Carsten Binnig, Stefan Manegold, Jörn Kohlhammer, Nicola Pezzotti

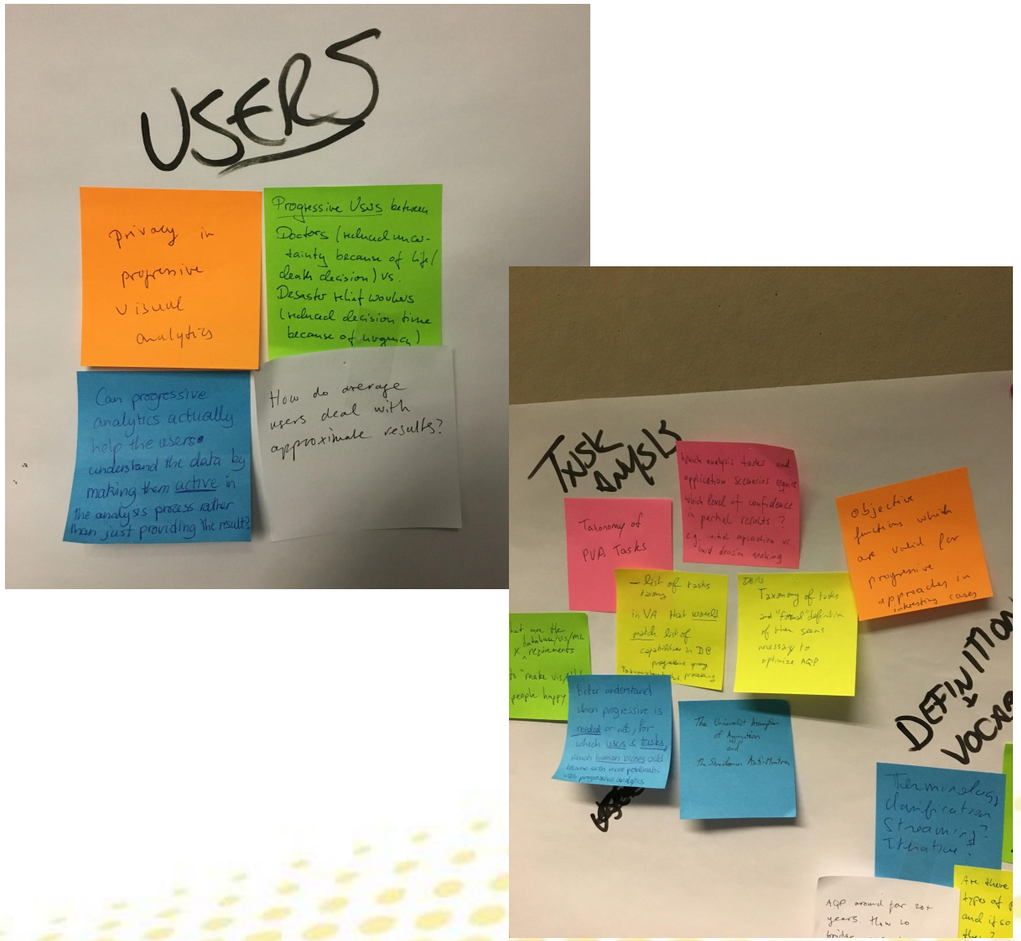
- 5 stages … current databases
  - give you all results at once
    → progressive results
  - no change of the query while it is running
    → want to steer queries
  - always compute from scratch
    → intelligently reuse
  - don’t predict
    → future queries
  - user context



Tasks & Users

Luana Micallef, Remco Chang, Hendrik Strobelt, Adam Perer, Dominik Moritz, Emanuel Zgraggen, Themis Palpanas, Michael Aupetit, Thomas Mühlbacher

- Developed a set of use cases/scenarios
- e.g. how progressive analytics could help in disaster scenarios



Humans in the Progressive Loop

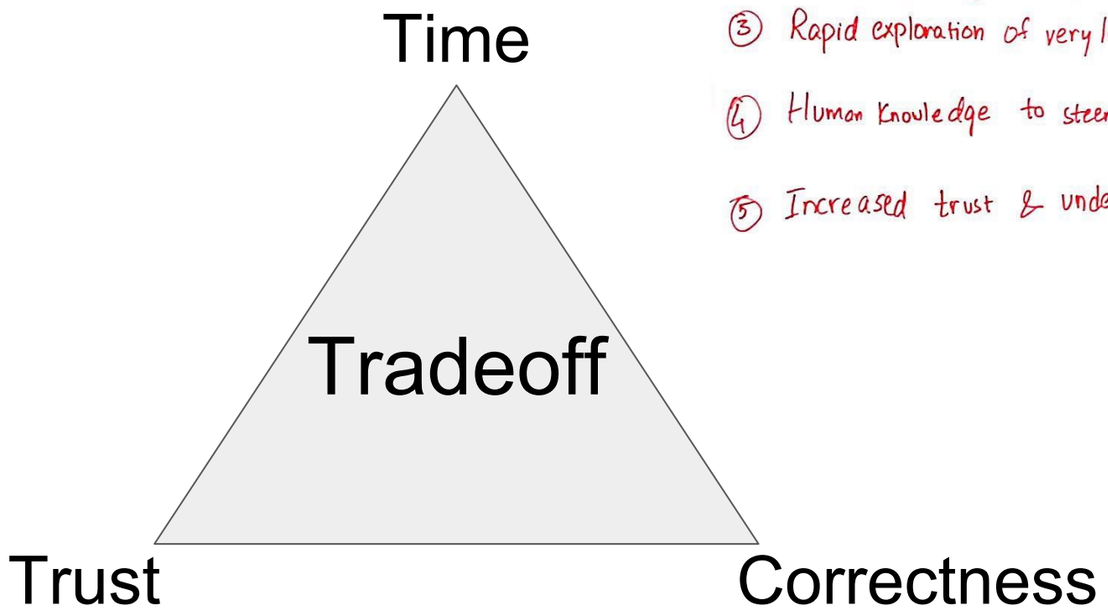
Marco Angelini, Remco Chang, Jörn Kohlhammer, Luana Micallef, Thomas Mühlbacher, Adam Perer, Giuseppe Santucci, Hans-Jörg Schulz, Chris Weaver

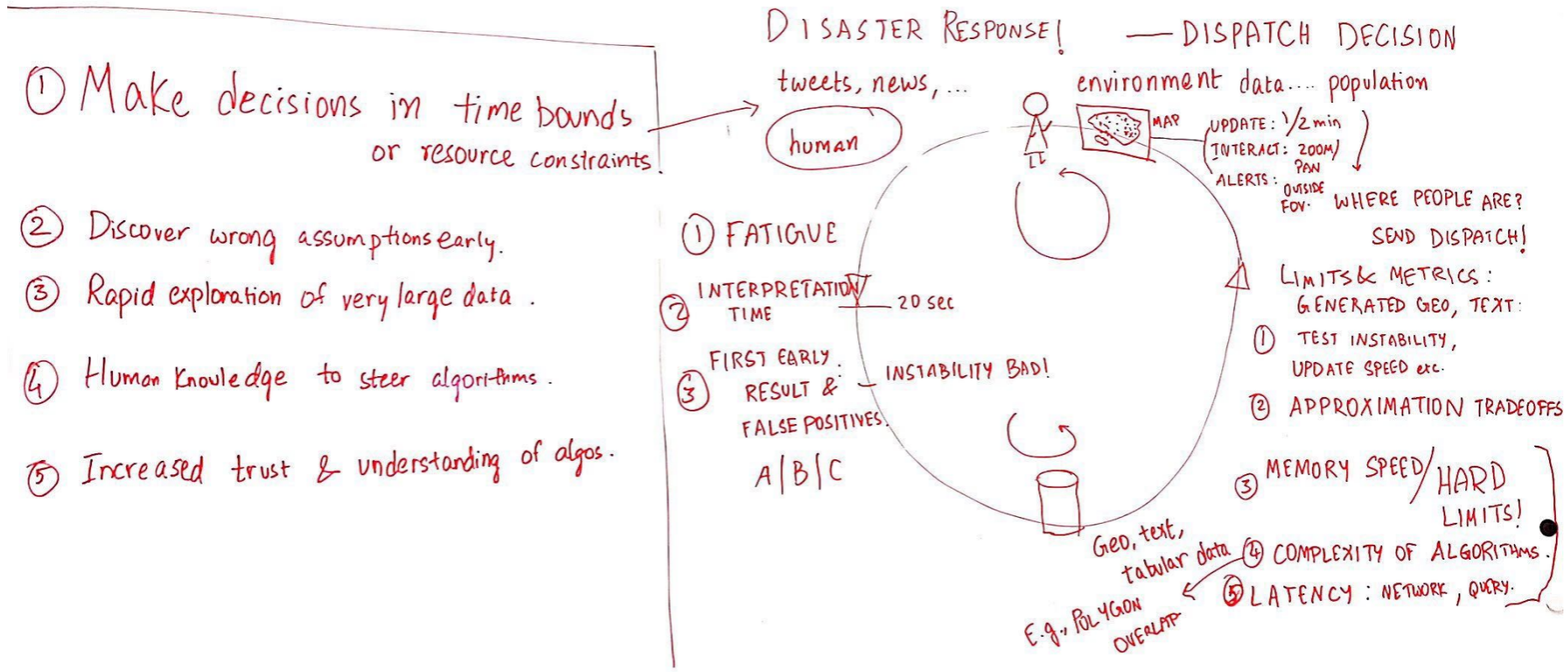
- Types of Users
  - Consumer
  - Domain Expert
  - Data Scientist
  - Learner / Apprentice
- User Strategies
- Cognitive Biases

Evaluation & Benchmarks

Hannes Mühleisen, Sriram Karthik Badam, Stefan Manegold, Jaemin Jo

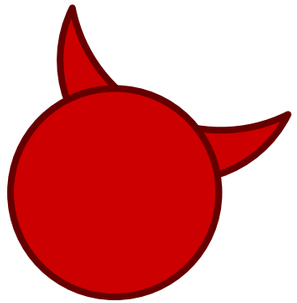
- User Studies
- System Benchmarks





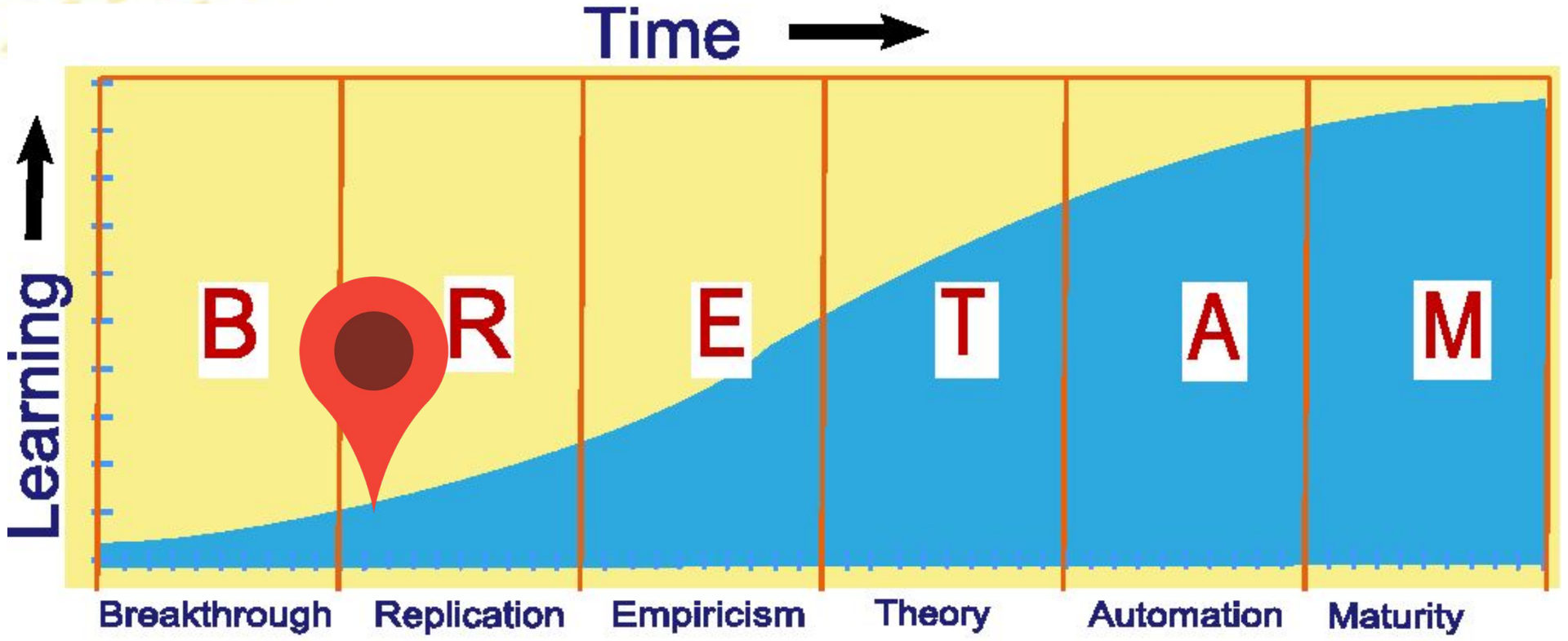
Threats

Christopher M. Jermaine, Dominik Moritz, Anna Vilanova, Emanuel Zgraggen



- Cost vs Benefits
- Uncertainty and Progressiveness
- What’s new?

- Visual Analytics is limited wrt big data and expensive algorithms
- Progressive Data Analysis and Visualization will make it scalable
  - probably with limitations



Brian R. Gaines, *Modeling and forecasting the information sciences*, Information Sciences, Volumes 57–58, 1991, Pages 3-22, ISSN 0020-0255, https://doi.org/10.1016/0020-0255(91)90066-4.