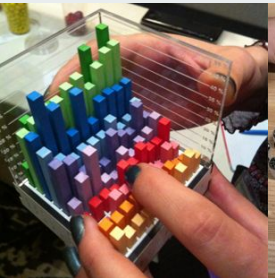
Aviz

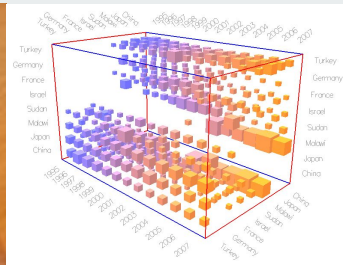
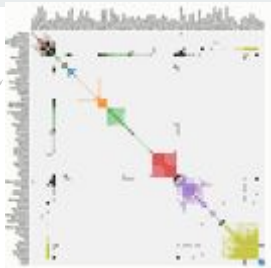
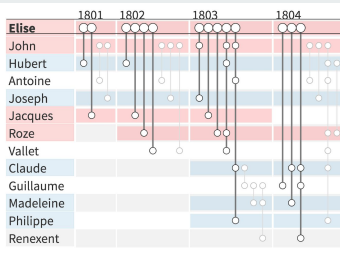
Visual Analytics Project

Inria

Practical Use Cases for Progressive Visual Analytics

Jean-Daniel Fekete, Inria
Qing Chen, Polytechnique & Inria
Yuheng Feng, Inria
Jonas Renault, CNRS



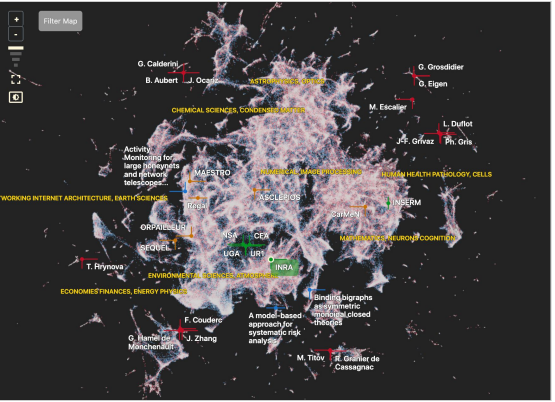





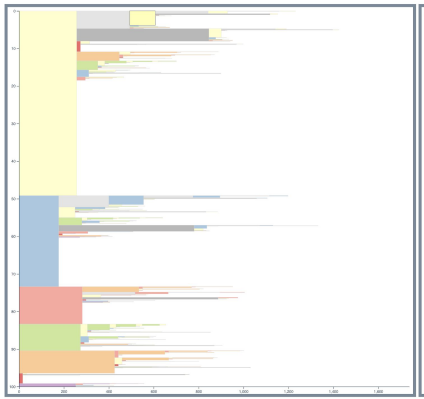
Take Away Messages

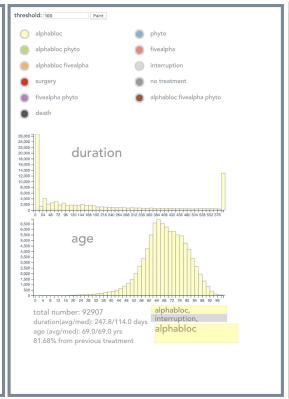
1. The visualization community should address the scalability challenge
2. The Progressive Data Analysis Paradigm (PVA) allows Visual Analytics applications to scale
3. PVA is emerging and will become mature in a few years
4. Meanwhile, many techniques and libraries can already be used to implement PVA applications
    - We will review a few of them here

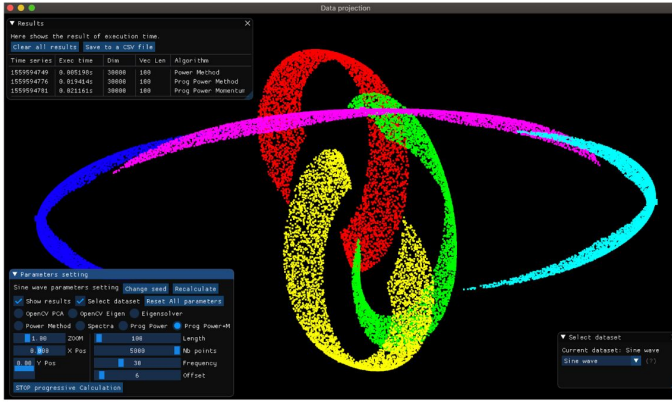
PVA: Semantically-Meaningful Partial Results, Improving over time, cConverting to the final solution.

3 Examples: Cartolabe, ParcoursVis, and PPCA





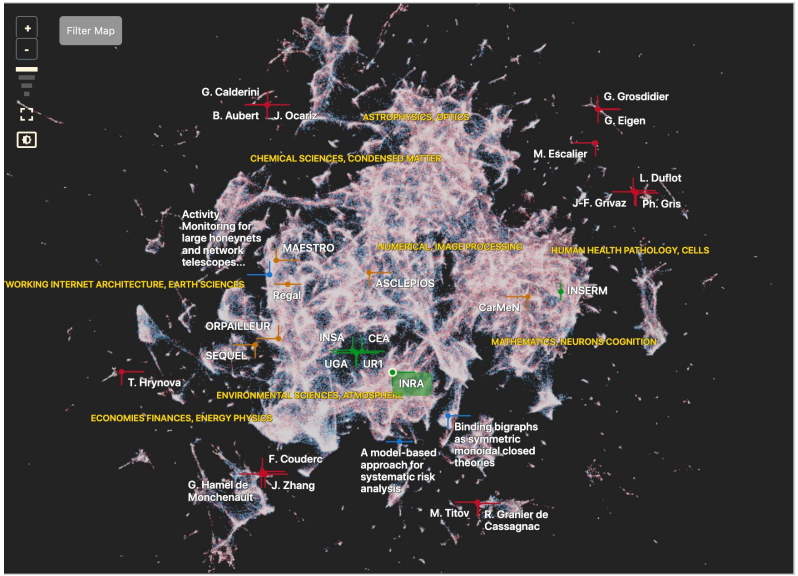




Cartolabe: PVA for Large Textual Corpora

Philippe Caillou Jean-Daniel Fekete Michèle Sebag Jonas Renault Anne-Catherine Letournel

- See https://cartolabe.fr
- Web-based visualization
  - HAL, Wikipedia, ArXiv, etc.
- Uses multidimensional projection (UMAP)
- From 500k to 5 Million entities visualized
- Use tiled images
- Tiles are pre-computed
- Filtering is done progressively and new filtered tiles are recomputed progressively



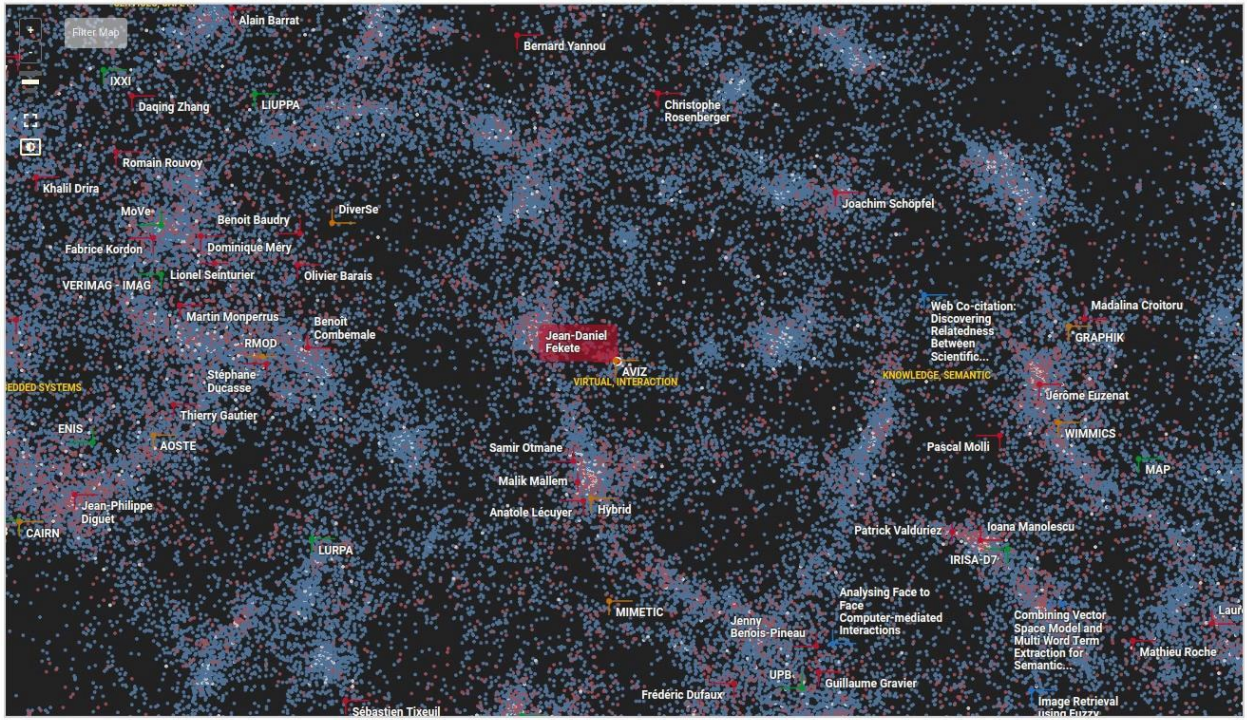
Cartolabe Map About

Search
Jean-Daniel Fekete ×

› Search Filters

Jean-Daniel Fekete
Author 132

Projection of entities extracted from HAL



Density maps Authors Articles

Labels Authors Articles Labs Teams Words Clusters

Developped by Inria cnrs Paris Sud

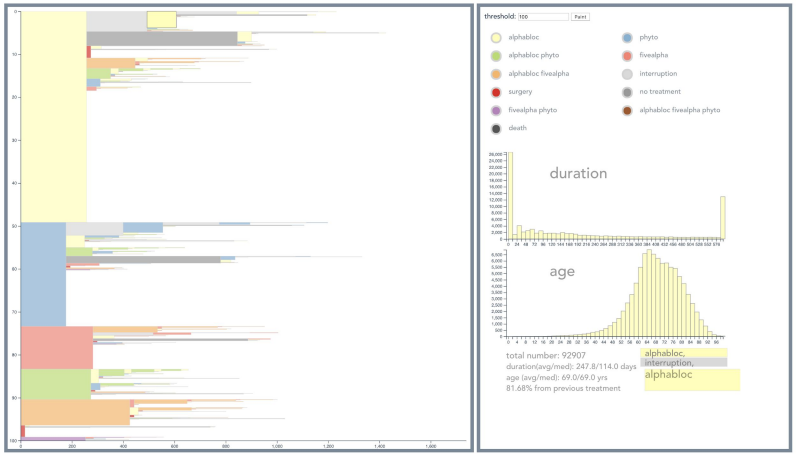
ParcoursVis: Exploring Patient Events at Scale

Qing Chen, Stéphane Gaïffas, Emmanuel Bacry, Jean-Daniel Fekete

EventFlow on Steroïds

- https://hcil.umd.edu/eventflow/

- French Social Security maintains a database of all reimbursement and medical acts
  - 100 million people for 50 years+
- We are exploring it for specific treatments
  - Male non-prostatic urinary problems
- 3 million people, 100+ events
  - Dynamic queries, details on demand

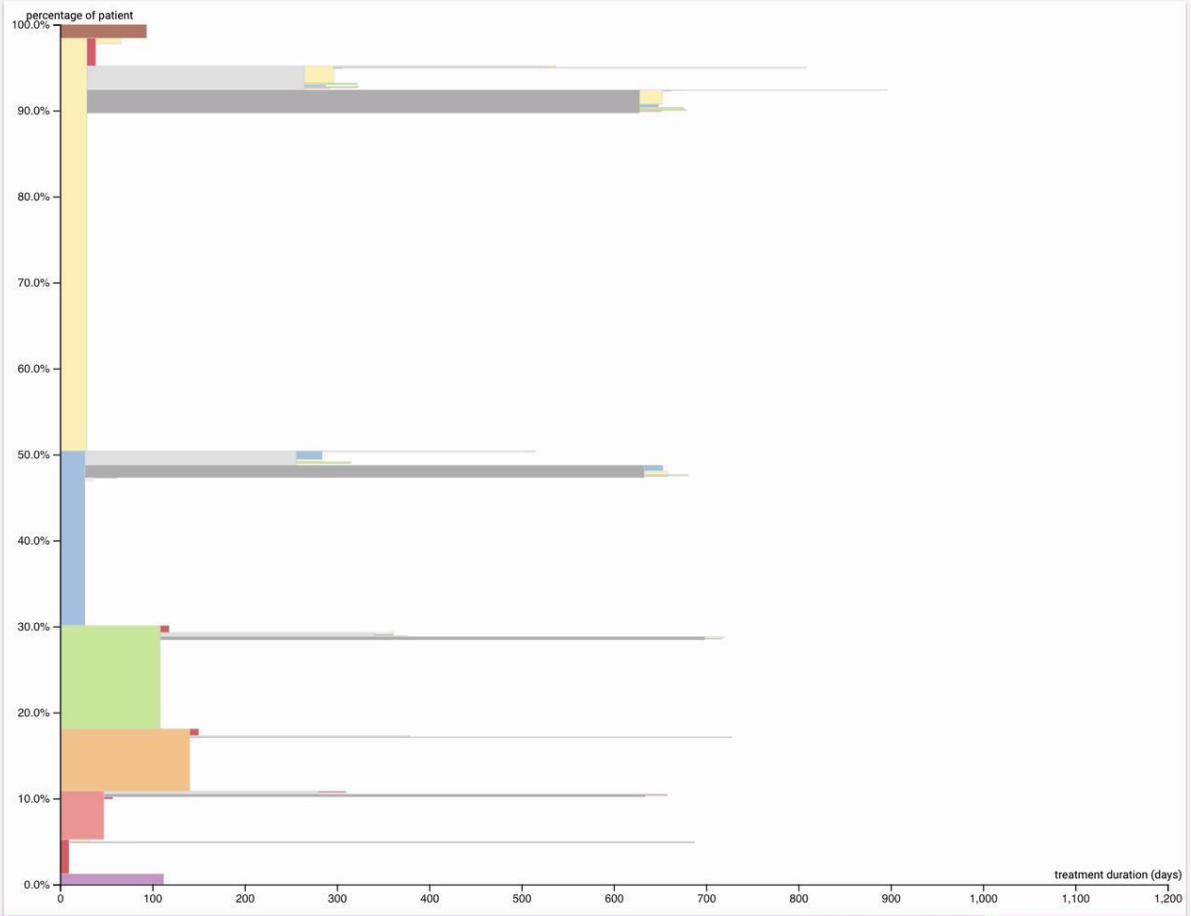


PARCOURSVIS

NUMBER OF PATIENTS: 2199914(PROCESSED)/ (TOTAL)

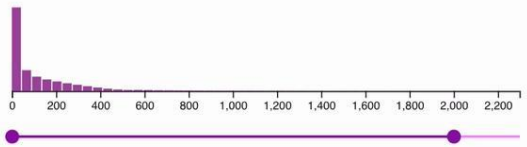
DATA INFO

HELP

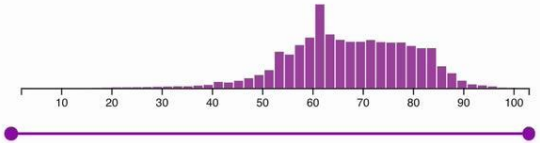


OVERVIEW DETAIL CONTROL HISTORY SETTING

Treatment Duration (days) 0 ~ 2298



Age (years)



Depth

5

No treatment duration

365

Interruption duration

180

Diseases

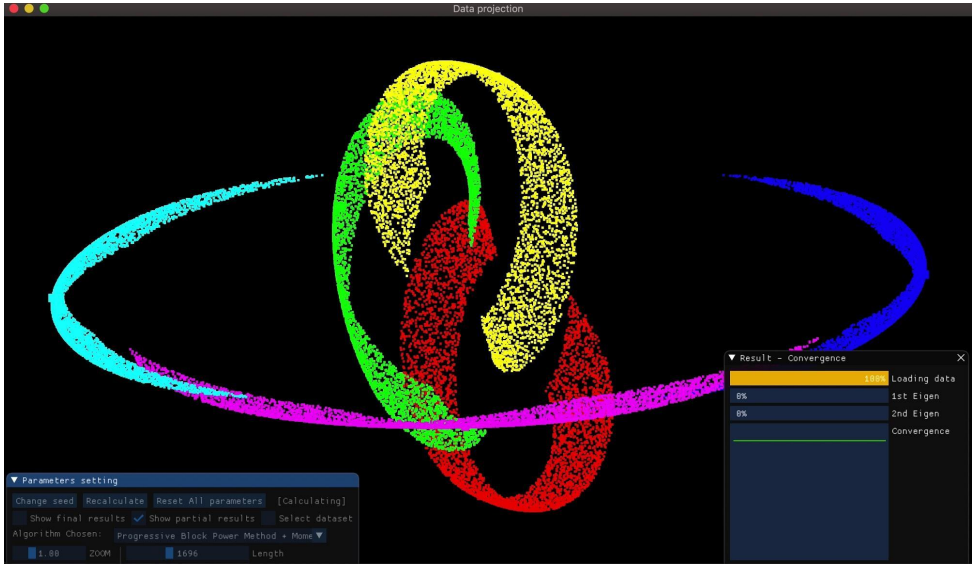
Diabete both yes no

2200000 / 4000000

Progressive PCA for Massive Time-Series

Yuhen Feng, Alejandro Ribes, Jean-Daniel Fekete

- EDF needs to forecast the behavior of complex hydraulic systems over time
- Ensemble simulations generate multiple time-series
  - Thousands of time-series
  - Tens of thousands of dimensions
- Exploring them with existing systems is not interactive



Existing Building Blocks

- No general framework for PVA yet
  - Still working on it… stay tuned
- But not bare-bone either
- Three technologies allow to scale up
  - Compressed Bitsets
  - Data Sketching
  - Online Algorithms

- Compressed Bitsets
  - AKA Compressed Bitmaps
  - Super fast, super small, super useful
- Data Sketching
  - Streaming algorithms to maintain approximate computations
  - Trade between speed, memory, accuracy
- Online Algorithms
  - Update their status when new data arrives
  - Exact but not real-time guarantees

Compressed Bitmaps

- Roaring Bitmap
  - http://roaringbitmap.org/ many languages
  - *Use Roaring for bitmap compression whenever possible. Do not use other bitmap compression methods* (Wang et al., SIGMOD 2017)
- Bitmagic
  - http://bitmagic.io/ C++
  - Check it for various unexpected and smart uses of bitsets
- FastBitSet.js
  - https://github.com/lemire/FastBitSet.js

Used for:

- compressed int lists (indices)
- Selection with fast boolean operations
- In Cartolabe
  - An index is a bitmap (e.g., Inria, CNRS)
  - A Year is a bitmap
  - A tile is a bitmap
  - Super fast intersections (e.g., tile & filter)
- In ParcoursVis
  - Each node contains a list of ids

Data Sketching

Streaming algorithms to maintain approximate computations (see https://datasketches.github.io)

- COUNT UNIQUE
- Quantile + Histogram
  - Seel also https://github.com/tdunning/t-digest
- Frequent items
- Uniform sampling
- SVD

- ParcoursVis should compute histograms
- On many attributes potentially
- Also compute quantiles

Not fully progressive though:

- Only new values can be added
- No filtering or removal
- More work to extend to PVA

Online Algorithms

Update their status when new data arrives
Exact but not real-time guarantees
Many Online algorithms are available in sklearn:

- Classification: Perceptron, SGD classifier, Naive bayes classifier
- Regression: SGD Regressor, Perceptron, Passive Aggressive regressor, SAG Logistic
- Clustering: Mini-batch k-means
- Feature extraction: Incremental PCA, Mini-batch dictionary learning

PPCA uses IncrementalPCA

- Real-time by tuning the # of items by batch

But:

- Online structures cannot be filtered or changed in general
- More research for PVA

Take Away Messages

1. The visualization community should address the scalability challenge
2. The Progressive Data Analysis Paradigm (PVA) allows Visual Analytics applications to scale
3. PVA is emerging and will become mature in a few years
4. Meanwhile, many techniques and libraries can already be used to implement PVA applications
    - We will review a few of them here