

Data Analysis: Discussion

Chris Williams
School of Informatics, University of Edinburgh

March 2010

General points

- ▶ Blind man and the elephant

General points

- ▶ Blind man and the elephant
- ▶ Data-centric research/Data-driven science/Data-intensive research

General points

- ▶ Blind man and the elephant
- ▶ Data-centric research/Data-driven science/Data-intensive research
- ▶ Goal: doing science (data diarrhoea, information constipation!)
- ▶ Data collection, organization and analysis

General points

- ▶ Blind man and the elephant
- ▶ Data-centric research/Data-driven science/Data-intensive research
- ▶ Goal: doing science (data diarrhoea, information constipation!)
- ▶ Data collection, organization and analysis
- ▶ Huge need for and opportunities for analysis (old and new)

General points

- ▶ Blind man and the elephant
- ▶ Data-centric research/Data-driven science/Data-intensive research
- ▶ Goal: doing science (data diarrhoea, information constipation!)
- ▶ Data collection, organization and analysis
- ▶ Huge need for and opportunities for analysis (old and new)
- ▶ Requires domain and data experts to collaborate
Provide time and incentives (Kersten)

General points

- ▶ Blind man and the elephant
- ▶ Data-centric research/Data-driven science/Data-intensive research
- ▶ Goal: doing science (data diarrhoea, information constipation!)
- ▶ Data collection, organization and analysis
- ▶ Huge need for and opportunities for analysis (old and new)
- ▶ Requires domain and data experts to collaborate
Provide time and incentives (Kersten)
- ▶ In a MSc or centre for doctoral training (CDT) one should have core first year courses on data analysis

Dimensions

- ▶ Scale
- ▶ Data complexity (e.g. 'omics data)
- ▶ Model complexity

Dimensions

- ▶ Scale
- ▶ Data complexity (e.g. 'omics data)
- ▶ Model complexity
- ▶ Alex Szalay: power law of data set sizes

Dimensions

- ▶ Scale
- ▶ Data complexity (e.g. 'omics data)
- ▶ Model complexity
- ▶ Alex Szalay: power law of data set sizes
- ▶ Throwing away data ...

Dimensions

- ▶ Scale
- ▶ Data complexity (e.g. 'omics data)
- ▶ Model complexity
- ▶ Alex Szalay: power law of data set sizes
- ▶ Throwing away data ...
- ▶ Infrastructure work: clearly necessary for big data, but need to ensure relevance to end users. Sociological factors

Types of data

- ▶ “Rectangular data” (records and fields)
- ▶ Images
- ▶ Text
- ▶ Remote sensing
- ▶ Sensor data streams
- ▶ 'Omics data

- ▶ Commonalities across disciplines

Types of Data Analysis

- ▶ Exploratory data analysis
 - ▶ Impressive tools for interactive visualization in earth sciences (Haines), geospatial data (Batty)

Types of Data Analysis

- ▶ Exploratory data analysis
 - ▶ Impressive tools for interactive visualization in earth sciences (Haines), geospatial data (Batty)
- ▶ Descriptive data analysis
 - ▶ e.g. Discovering patterns in text and relational data (McCallum)

Types of Data Analysis

- ▶ Exploratory data analysis
 - ▶ Impressive tools for interactive visualization in earth sciences (Haines), geospatial data (Batty)
- ▶ Descriptive data analysis
 - ▶ e.g. Discovering patterns in text and relational data (McCallum)
- ▶ Predictive data analysis
 - ▶ Graepel, Rougier

Other Issues

- ▶ Data cleaning and quality
- ▶ Complexity and prior knowledge
- ▶ Analysis of simulation data (Szalay)
- ▶ Parameter estimation for (complex) dynamical models/simulations
- ▶ Circuit/network inference
- ▶ Scientific computational environments + toolboxes (vs specific tools)
- ▶ Data-centric thinking is a state of mind
Provide time and incentives for training young (and not-so-young) researchers