Interactive Visual Data Analysis in the Times of Big Data

Cagatay Turkay

* giCentre, City University London

Who?

- Lecturer (Asst. Prof.) in Applied Data Science
- Started December 2013
- @ the giCentre (gicentre.net)
- PhD @ VisGroup at Univ. of Bergen, Norway
  - with Helwig Hauser
- MSc @ CGLab at Sabanci University, Istanbul, Turkey
  - with Selim Balcisoy
giCentre

• 6 academics
• 3 postdocs
• 5 PhDs

Research to date

• Integrating **Computational** Tools in **Interactive Visual** Analysis Methods
• **Perceptually** Optimized Visualization
• **Information Theoretical** Approach to **Crowd Simulation**
Interactive Visual Data Analysis in the Times of Big Data

Data supported science

- **Data analysis** in almost all scientific fields
  - Biology, medicine, astronomy, psychology, ...
- **Data driven** science
- Research in **several fields**
  - Visualization
  - Data Mining
  - Machine Learning
  - Statistics
The “times” of Big Data

http://www.ibmbigdatahub.com/infographic/four-v-big-data

More V’s added

• Volume
• Velocity
• Variety
• Veracity
• Visualisation
• Value

“big data” on Google Trends, as of today
http://www.google.com/trends/explore#q=big%20data
Getting the best of “Big data”

- Limited value in size!
- Real value in **complexity**
  - Heterogeneous
  - Several modalities
  - Many variables
  - Spatially & Temporally varying

Visual data analysis can help!
Visualization?

“Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.”
[Tamara Munzner, 2014]

“The use of computer-generated, interactive, visual representations of data to amplify cognition”[Card, Mackinlay, & Shneiderman 1999]

Visualization – a mature field
Interactive Visual Analysis (a.k.a. Visual Analytics)

• “... the science of analytical reasoning facilitated by interactive visual interfaces.” [Thomas and Cook, 2005]
• “... combining visualization, human factors and data analysis.” [Keim et al., 2006]
• Combine the best of two worlds: human capabilities and computing power

[Image: Interactive Visual Analysis examples]

How visual analysis can help?

• Ease of interpretation
• Ease of communication
• Flexibly varying & relating multiple aspects
• Compare multiple computational outputs & communicate uncertainties
• Seamless integration of computation

[Note: This talk!]

[Image: Visual Analysis Examples]
1- **Flexibly vary & relate multiple aspects**

**Basics -- Multiple linked views**

- Several visualizations
- **Linking + brushing**
- **Combining** selections
- Commercial success

![Tableau Software]
Example:
Analyzing Crime Statistics through Selection Combinations and Data Derivation

Analyzing Crime Statistics

- Crime statistics over Birmingham
- Crime records
  - Crime types
  - Offender location
  - Crime location
- Want to identify:
  - Hot spots
  - Groups
  - Trends / patterns

VALCRI
A key aspect – Iterative process

- Learn **through the process**
- Evaluate & re-perform

2- **Compare** multiple computational outputs & communicate **uncertainties**
Issues in computational analysis

- **Computational tools** have helped
  - Several tasks: find **structures**, **summarize**, trends…
  - Rich set of methods, PCA, LDA, Clustering …

- **Have issues**, unfortunately
  - Not flexible
  - Limited **interaction**
  - **Reliability** (noisy data, assumptions)

- **Visual analysis to:**
  - **Compare** multiple computational outputs & communicate **uncertainties**

---

Example:

**Comparing** and **Characterizing**

Clustering Results in Cancer Subtype Analysis

[ Turkay, Cagatay, et al.," Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX
IEEE computer graphics and applications 34.2 (2014): 38-47. ]
Cluster analysis

- **Cluster** is a group of "similar" entities
- Several algorithms – important to compare
- Challenging to evaluate/interpret

Thanks to Alexander Lex for the slides

Multiple Clusterings (& Datasets)

- Cluster A1
- Cluster A2
- Cluster A3

- Clustering 1, e.g., k-means
- Clustering 2, e.g., hierarchical

Thanks to Alexander Lex for the slides
Multiple Clusterings (& Datasets)

Cluster A1  
Cluster A2  
Cluster A3  

C B1  
C B2  

Clustering 1, e.g., k-means  
Clustering 2, e.g., hierarchical

Thanks to Alexander Lex for the slides

Multiple Clusterings (& Datasets)

Cluster A1  
Cluster A2  
Cluster A3  

C B1  
C B2  
Dep. C1  
Dep. C2

Clustering 1, e.g., k-means  
Clustering 2, e.g., hierarchical  
Meta Data, e.g., clinical data

Thanks to Alexander Lex for the slides
Cancer types are **not** homogeneous, i.e., **subtypes**

**Subtypes** are identified by **clustering** datasets, e.g.,
- based on an **expression pattern**
- a **mutation status**
- a **copy number alteration**
- a **combination** of these

http://caleydo.org/

Multiple Clustering Results

Sample Overlaps

Gene Overlaps ??

Many shared Patients

Clustering 1

Clustering 2

Thanks to Alexander Lex for the slides
Integrated detail visualisations

Only the member samples visualized

Communicating (Un)Certainty

Only the member samples used in computations

Genes significantly different for a cluster
Some observations

• Compare many results to get **support** and **increase trust**
• Identify **how and why** results change
• Communicate **uncertainties**
• Applied to many other problems

---

Booshehrian et al., 2012

---

Prediction

Simulations

Waser, 2012
3- Seamless integration of computation

Seamless integration

• Best of human + computer – the ultimate goal of VA!
• Computational algorithms implicit
• Embedded within interaction
• Basic interaction methods, e.g., select, pan, zoom, etc.
Example: Dynamically generated visual (statistical) summaries for geographical data analysis

Visualization to understand how things change with geography
visualizations to look at many variables...

and

computational support to generate comparative summaries for the visualizations
Example: Analysis of UK OA data

UK Census of Population in 2001 and 2011 for the 181,000 Output Areas (OA) for 41 indicators

Integrated summary computation

Variable comparison baseline (e.g., country or city avg.)

Each value is computed only for the selection

Dynamic computations to compare
Along the Southwest England coast

An interactive transect through London
More examples of integrations

**Multi-dimensional projection**

Janicke et al., 2008

**Regression analysis**

Muhlbacher, T., Piringer, 2013

**Finding structures**

Oliver et al., 2010

Time to wrap-up!
Lessons learned

- All starts with **problems** and related **data**
  - Need to earn users’ trust!
- Visual analysis as a means to increase **data’s value**
- Enhanced analysis through **informed** use of computation
- Interactive visual methods improve **reliability and interpretability**
- Tight integration enables quick **hypothesis prototyping**
- Important to communicate the **certainty** of the findings

Looking into the future

- Understanding the **real value of** data
- **Scalable analysis** processes
- Improved methods for **visual guidance**
- Enable **interactive response** in seamless integrations
- Adapt to changing information characters & problems
  - **Dynamic data**, e.g., data streams
  - **Prediction**
Thank you!

- Helwig Hauser
- Peter Filzmoser
- giCentre folks: Aidan Slingsby, Jason Dykes, Jo Wood
- Arvid Lundervold, Astri Lundervold, Nathalie Reuter, the Caleydo Team