SkyTree Visualization Fireside Chat
Is Big Data Visualization Possible?

Tamara Munzner
Department of Computer Science
University of British Columbia

Google Hangout on Air
October 1 2014

http://www.cs.ubc.ca/~tmm/talks.html#skytree14
About me: Geometry Center 1991-1995

- geometry and topology vis
  - 3D, 4D, non-Euclidean

Geomview
http://geomview.org/

The Shape of Space
http://youtu.be/-gLNIC_hQ3M

Outside In
http://youtu.be/sKqt6e7EcCs
http://youtu.be/x7d13SgqUXg

http://youtu.be/6j4T7l49H3Y
http://www.crcpress.com/product/isbn/9781568814537
About me: Stanford 1995-2000

• infovis: network vis
  – 3D hyperbolic trees/networks
  – computational linguistics network

Constellation
http://youtu.be/7sJC3QVpSkQ

H3
http://youtu.be/fhbQy_NCwWI
About me: UBC 2002-

technique-driven work

problem-driven work

theoretical foundations

evaluation
When to use visualization

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

• human in the loop needs the details
  – doesn't know exactly what questions to ask in advance
  – longterm analysis
  – automation stepping stone, refining, trustbuilding
  – presentation
• external representation: perception vs cognition
• intended task, measurable definitions of effectiveness

more at:
Visualization Analysis and Design, Chapter 1.
Why show data to people?

• summaries lose information
  – confirm expected and find unexpected patterns
  – assess validity of statistical model
Why show data to people?

• summaries lose information
  – confirm expected and find unexpected patterns
  – assess validity of statistical model

**Anscombe’s Quartet**

<table>
<thead>
<tr>
<th>Identical statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x mean</td>
<td>9</td>
</tr>
<tr>
<td>x variance</td>
<td>10</td>
</tr>
<tr>
<td>y mean</td>
<td>8</td>
</tr>
<tr>
<td>y variance</td>
<td>4</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>1</td>
</tr>
</tbody>
</table>
Why show data to people?

- summaries lose information
  - confirm expected and find unexpected patterns
  - assess validity of statistical model

Anscombe’s Quartet

<table>
<thead>
<tr>
<th>Identical statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x mean</td>
<td>9</td>
</tr>
<tr>
<td>x variance</td>
<td>10</td>
</tr>
<tr>
<td>y mean</td>
<td>8</td>
</tr>
<tr>
<td>y variance</td>
<td>4</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>1</td>
</tr>
</tbody>
</table>
Technique-driven work: Networks

• scaling up networks
  – multilevel networks, 10K-100K nodes
    • topologically aware decomposition, layout, browsing
  – trees, millions of nodes
    • guaranteed visibility of semantically meaningful marks

[Figure 7.25: GrouseFlocks uses containment to show graph hierarchy structure. (a) Original graph. (b) Several alternative hierarchies built from the same graph. The hierarchy alone is shown in the top row. The bottom row combines the graph encoded with connection with a visual representation of the hierarchy using containment. From [Archambault et al. 08], Figure 3.]

TopoLayout
Smashing Peacocks Further
Grouse
GrouseFlocks
TugGraph

http://youtu.be/T1Xbt6XOWp8
http://youtu.be/AWXAe8zykI8

TreeJuxtaposer
PRISAD

http://youtu.be/fq8EIAOutvs
http://youtu.be/GdaPj8a9QEo
Technique-driven work: Dimensionality reduction

- closest overlap between vis and ML
  - Glimmer: MDS on the GPU
  - Glint: DR for costly distances
  - QSNE: sparse documents
    - high quality for millions of items

Glimmer

http://youtu.be/PLaBAPM6qLI

Glint

QSNE
Problem-driven work: Genomics

MizBee
http://youtu.be/86p7brwuz2g

Variant View
http://youtu.be/AHDnv_qMXxQ

Cerebral
http://youtu.be/76HhG1FQngI

MulteeSum
Problem-driven work: Many domains

LiveRAC: system management time-series

RelEx: in-car overlay networks

Vismon: fisheries management

Overview: investigative journalism

http://youtu.be/h0kHoS4VYmk

http://youtu.be/ld0c3H0VSkw

http://youtu.be/89lsQXc6Ao4

http://vimeo.com/71483614
http://www.cs.ubc.ca/group/infovis/

http://www.cs.ubc.ca/~tmm/talks.html#skytree14
Overview design evolution
Overview design evolution

• how to find the needle in the haystack?
• how to convince that the haystack has no needles?
Overview design evolution

• how to find the needle in the haystack?

• how to convince that the haystack has no needles?
• how to find the needle in the haystack?
• how to convince that the haystack has no needles?
Overview origin story: WikiLeaks meets Glimmer
Overview origin story: WikiLeaks meets Glimmer

• WikiLeaks: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  – conjecture that existing label classification falls short of showing all meaningful structure in data
    • friendly action, criminal incident, ...
  – had some NLP, needed better vis tools
Overview origin story: WikiLeaks meets Glimmer

- **WikiLeaks**: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  - conjecture that existing label classification falls short of showing all meaningful structure in data
    - friendly action, criminal incident, ...
  - had some NLP, needed better vis tools

- **Glimmer**: multilevel dimensionality reduction algorithm
  - scalability to 30K documents and terms

Visual dimensionality reduction for document datasets

### Task 1
In HD data | Out 2D data
---|---
**What?** | In High dimensional data | Produce | Derive |
**Why?** | Out 2D Data |

### Task 2
In 2D data | Out Scatterplot Clusters & points
---|---
**What?** | In 2D data | Out Scatterplot Clusters & points |
**Why?** | Discover | Explore | Identify |
**How?** | Encode | Navigate | Select |

### Task 3
In Scatterplot Clusters & points | Out Labels for clusters
---|---
**What?** | In Scatterplot Clusters & points | Out Labels for clusters |
**Why?** | Produce | Annotate |

- more on visual DR: hour-long talk *Dimensionality Reduction from Several Angles*
  
  [http://www.cs.ubc.ca/~tmm/talks.html#linz14](http://www.cs.ubc.ca/~tmm/talks.html#linz14)
What/Why/How interplay
What/Why/How interplay

• why: understand clusters
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings
What/Why/How interplay

• **why:** understand clusters

• **what:** derive data of full cluster hierarchy
  – explore space of possible clusterings
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
  – arrange space: node-link
What/Why/How interplay

• **why**: understand clusters

• **what**: derive data of full cluster hierarchy
  – explore space of possible clusterings

• **how**: show cluster hierarchy
  – arrange space: node-link

• **how**: support tagging clusters/docs
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
  – arrange space: node-link

• how: support tagging clusters/docs
What/Why/How interplay

- **why**: understand clusters

- **what**: derive data of full cluster hierarchy
  - explore space of possible clusterings

- **how**: show cluster hierarchy
  - arrange space: node-link

- **how**: support tagging clusters/docs
  - following or cross-cutting hierarchy!
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
  – arrange space: node-link

• how: support tagging clusters/docs
  – following or cross-cutting hierarchy!
  • simple annotation

Dataset Types

Networks

Targets

Network Data

Produce

Annotate

Arrange Networks And Trees

Node-link Diagrams

Connections and Marks

✔ NETWORKS ✔ TREES
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
  – arrange space: node-link

• how: support tagging clusters/docs
  – following or cross-cutting hierarchy!

  • simple annotation
  • progress tracking
What/Why/How interplay

• why: understand clusters

• what: derive data of full cluster hierarchy
  – explore space of possible clusterings

• how: show cluster hierarchy
  – arrange space: node-link

• how: support tagging clusters/docs
  – following or cross-cutting hierarchy!
    • simple annotation
    • progress tracking
    • user-defined semantics
How: Idiom design decisions

- **facet: juxtapose linked views**
  - linked color coding
    - cluster hierarchy tree
    - DR scatterplot
    - tags
  - reading text/keywords
    - cluster list
    - doc reader

- **Juxtapose and Coordinate Views**
  - Share Encoding: Same/Different
  - Share Data: All/Subset/None
  - Linked Highlighting

- **Identity Channels: Categorical Attributes**
  - Spatial region
  - Color hue
  - Motion
  - Shape
Overview video (version 1)

http://www.cs.ubc.ca/labs/imager/tr/2012/modiscotag/
Path to adoption

• version 1
  – fast cluster hierarchy construction for sparse data
  – research prototype by PhD student
  – positive initial assessment from AP Caracas bureau chief
    • barrier to adoption: difficult install/load process
Path to adoption

• version 1
  – fast cluster hierarchy construction for sparse data
  – research prototype by PhD student
  – positive initial assessment from AP Caracas bureau chief
    • barrier to adoption: difficult install/load process

• version 2
  – web deployment, DocumentCloud integration, usability
    • many months of engineering
      – Knight Foundation funding to the rescue!
    • published story by unaffiliated reporter: police corruption in Tulsa
Path to adoption

• even more rounds of what/why/how interplay
  – which views needed? what should they show? how should they show it?
  – usability and utility

• version 3
  – published story: VP candidate Ryan asked for federal help even as championed cuts
  – published story: gun control debate

• version 4
  – followup investigation: government corruption in Texas
  – published story: police misconduct in New York (Pulitzer prize finalist!)
Overview video v4

http://vimeo.com/71483614
Overview video v4

http://vimeo.com/71483614

• versions 3 and 4
  – no DR scatterplot
  – tree arrangement emphasizing nodes not links
  – combined doc/cluster viewer
Why: Task abstractions
Why: Task abstractions

• what’s in this collection?
  (of leaked docs)
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters
Why: Task abstractions

• what’s in this collection?
  (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence
  (within FOIA dump)
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
  – locate clusters/documents
Why: Task abstractions

• what’s in this collection?  
  (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence  
  (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
  – locate clusters/documents
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – *generate* hypothesis
  – *summarize* clusters
  – *explore* clusters

• locate evidence (within FOIA dump)
  – *verify* hypothesis
  – *identify* clusters/documents
  – *locate* clusters/documents
Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
  – locate clusters/documents

Why: Task abstractions

- what’s in this collection? (of leaked docs)
  - generate hypothesis
  - summarize clusters
  - explore clusters

- locate evidence (within FOIA dump)
  - verify hypothesis
  - identify clusters/documents
  - locate clusters/documents

- prove non-existence of evidence

Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
  – locate clusters/documents

• prove non-existence of evidence
  – even harder!

Why: Task abstractions

• what’s in this collection? (of leaked docs)
  – generate hypothesis
  – summarize clusters
  – explore clusters

• locate evidence (within FOIA dump)
  – verify hypothesis
  – identify clusters/documents
  – locate clusters/documents

• prove non-existence of evidence
  – even harder!
  – exhaustive reading vs filtering out irrelevant

Now what?

• continuing adoption
  – food stamp distribution delays in North Carolina
  – Surprise! Many credit card agreements allow repossession
  – The brilliance of Louis C.K.'s emails: He writes like a politician
  – Private memo reveals winding tale involving John McCain, the NRA, and... condors

• continuing development
  – Knight Foundation funds v5: named entity recognition, plugin API

• InfoVis'14 paper
  [Link](http://www.cs.ubc.ca/labs/imager/tr/2014/Overview/)
Algorithm: Spinoff series

• dimensionality reduction for huge text collections
  – great algorithm problem in its own right!
  – QSNE: fast and high-quality DR for millions of documents
    • key feature: handle sparseness appropriately

[Dimensionality Reduction for Documents with Nearest Neighbor Queries. Ingram and Munzner. Neurocomputing (Special Issue on Visual Analytics using Multidimensional Projections), to appear 2014.]

http://www.cs.ubc.ca/labs/imager/tr/2014/QSNE/