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

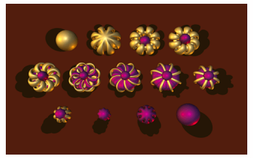
http://www.cs.ubc.ca/~tmm/talks.html#skytree14

About me: Geometry Center 1991-1995

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

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

Geomview http://geomview.org/

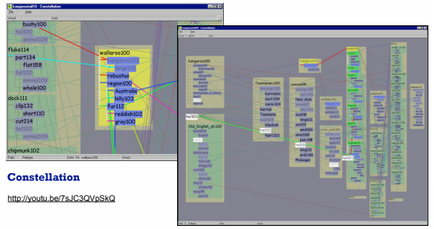


Outside In
http://youtu.be/sKqt6e7EcCs
http://youtu.be/x7d13SgqUXg

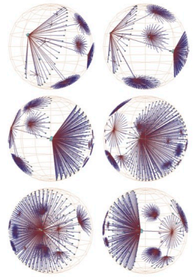
http://youtu.be/i6j4T7f49H3Y http://www.crcpress.com/product/isbn/9781568814537

About me: Stanford 1995-2000

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

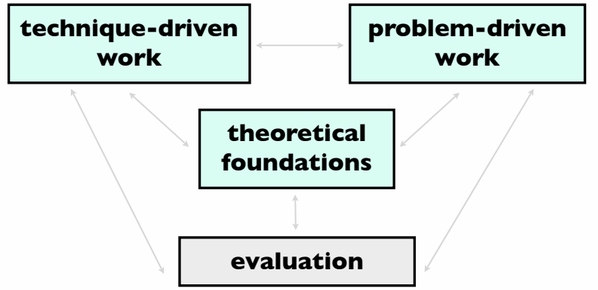


Constellation
http://youtu.be/7sJC3QVpSkQ



H3
http://youtu.be/fhbQy_NOwWI

About me: UBC 2002-

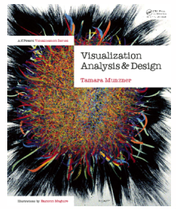


When to use visualization

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.
Munzner. AK Peters, 2014, to appear.

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

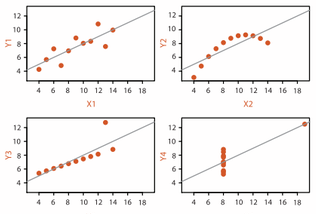
Identical statistics	
x mean	9
x variance	10
y mean	8
y variance	4
x/y correlation	1

Why show data to people?

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

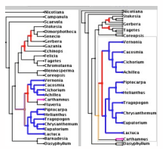
Anscombe's Quartet

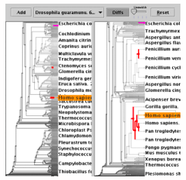
Identical statistics	
x mean	9
x variance	10
y mean	8
y variance	4
x/y correlation	1

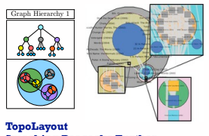


Technique-driven work: Networks

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







TopoLayout
Smashing Peacocks Further
Grouse
GrouseFlocks
TugGraph

http://youtu.be/t1Xbt6XOWp8

http://youtu.be/AWXAe8zv6t8

TreeJuxtaposer
PRISAD

http://youtu.be/fq8EIAOutvs

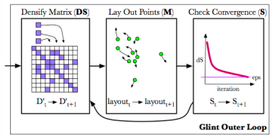
http://youtu.be/GdaPj8a9QEo

Technique-driven work: Dimensionality reduction

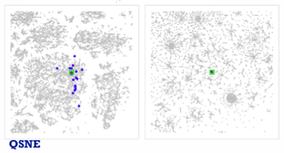
- closest overlap between vis and ML
  - Glimmer: MDS on the GPU
  - Glint: DR for costly distances
  - QSNE: sparse documents
    - high quality for millions of items

Glimmer

http://youtu.be/PLaBAPM6qLI

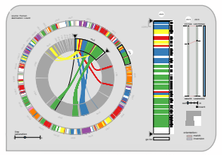


Glint

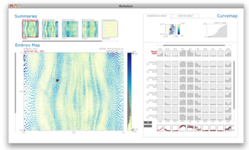


QSNE

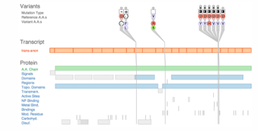
Problem-driven work: Genomics



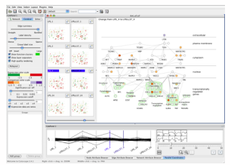
MizBee http://youtu.be/86p7brwuz2g



MulteeSum



Variant View http://youtu.be/AHDnv_qMXxQ

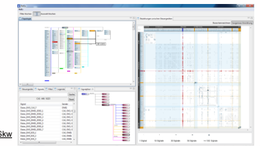


Cerebral http://youtu.be/76HhG1FQngI

Problem-driven work: Many domains

http://youtu.be/ld0c3H0VSkw

LiveRAC: system management time-series



RelEx: in-car overlay networks http://youtu.be/89lsQXc6Ao4

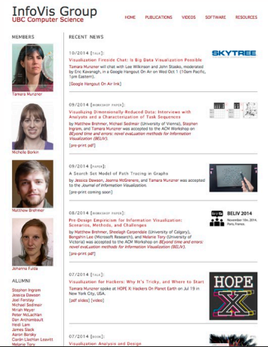
Vismon: fisheries management http://youtu.be/h0kHsoS4VYmk

Overview: investigative journalism http://vimeo.com/71483614

More info

http://www.cs.ubc.ca/group/infovis/

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



Overview design evolution

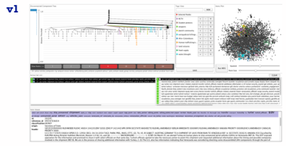
v4

Overview design evolution

v4

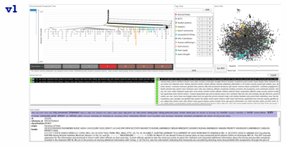
- how to find the needle in the haystack?
- how to convince that the haystack has no needles?

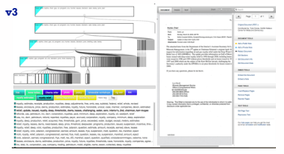
Overview design evolution



- how to find the needle in the haystack?
- how to convince that the haystack has no needles?

Overview design evolution



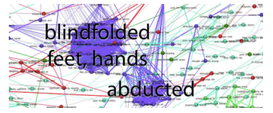


- how to find the needle in the haystack?
- how to convince that the haystack has no needles?

Overview origin story: WikiLeaks meets Glimmer

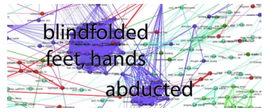
Overview origin story: WikiLeaks meets Glimmer

- 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

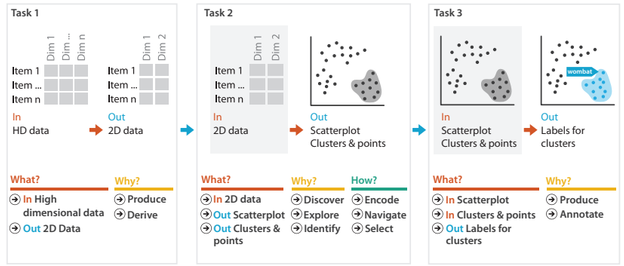
- WikiLeaks: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  - conjecture that existing label classification falls short of showing all meaningful structure in data
    - friendly action, criminal incident, ...
  - had some NLP, needed better vis tools



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

[Glimmer: Multilevel MDS on the GPU.
Ingram, Munzner, Olano. IEEE TVCG 15(2):249-261, 2009.]

Visual dimensionality reduction for document datasets



- more on visual DR: hour-long talk *Dimensionality Reduction from Several Angles*
  http://www.cs.ubc.ca/~tmm/talks.html#linz14

What/Why/How interplay

What/Why/How interplay

- why: understand clusters

What/Why/How interplay

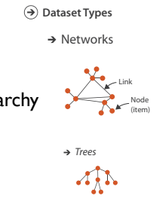
- why: understand clusters
- what: derive data of full cluster hierarchy

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!

Dataset Types → Networks → Link, Node (item) → Trees

Targets → Network Data → Topology → Paths

What? Why? How?

Produce → *Annotate* → tag

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

Dataset Types → Networks → Link, Node (item) → Trees

Targets → Network Data → Topology → Paths

What? Why? How?

Produce → *Annotate* → tag

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

Dataset Types → Networks → Link, Node (item) → Trees

Targets → Network Data → Topology → Paths

What? Why? How?

Produce → *Annotate* → tag

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
    - user-defined semantics

Dataset Types → Networks → Link, Node (item) → Trees

Targets → Network Data → Topology → Paths

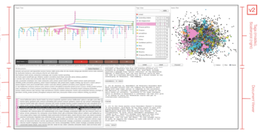
What? Why? How?

Produce → *Annotate* → tag

Arrange Networks And Trees → Node-link Diagrams, Connections and Marks → NETWORKS ✓ TREES ✓

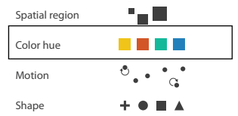
How: Idiom design decisions

- 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 → *Linked Highlighting*

Identity Channels: Categorical Attributes



What? Why? How?

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

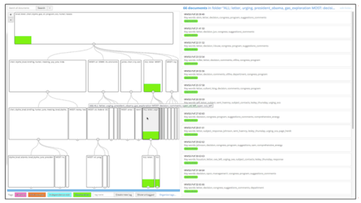
Path to adoption

- 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

What? Why? How?

Why: Task abstractions

- what's in this collection? (of leaked docs)

What? Why? How?

Why: Task abstractions

- what's in this collection? (of leaked docs)
  - *generate* hypothesis

What? Why? How?

Why: Task abstractions

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

What? Why? How?

Why: Task abstractions

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

What? Why? How?

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

→ *Discover*

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

→ *Discover*

→ Query
→ Identify → Compare → Summarise

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

→ *Discover*

→ Query
→ Identify → Compare → Summarise

→ Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

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

→ *Discover*

→ Query
→ Identify → Compare → Summarise

→ Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

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!

→ *Discover*

→ Query
→ Identify → Compare → Summarise

→ Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

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

→ *Discover*

→ Query
→ Identify → Compare → Summarise

→ Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

Now what?

Knight Foundation AP ASSOCIATED PRESS

Overview

- 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

https://www.overviewproject.org/

http://overview.ap.org/

- InfoVis14 paper

Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists. *Brehmer, Ingram, Stray, and, Munzner.*

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/