Visualization & Journalism: Four Vignettes

Tamara Munzner
Department of Computer Science
University of British Columbia

Computation and Journalism Symposium 2016
October 2016, Stanford CA

http://www.cs.ubc.ca/~tmm/talks.html#cj16 **@tamaramunzner**

Four vignettes

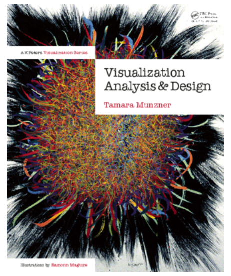
- a tale of two tools created for journalistic use
  - shared frameworks of interdisciplinary methods from my research group
    - thinking about collaboration
      - roles & rewards, for computer scientists & journalists
    - reasoning about visualization design
      - beyond pretty pictures
  - divergent goals & audiences
    - Overview: investigation / exploratory
    - TimeLineCurator: presentation / explanatory
- two cautionary tales with actionable advice
  - lessons we've learned in vis
    - challenges of color
    - difficulties of depth

Visualization (vis) defined & motivated

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 exploratory analysis
  - presentation of known results
  - stepping stone towards automation: refining, trustbuilding
- external representation: perception vs cognition
- intended task, measurable definitions of effectiveness



more at:
Visualization Analysis and Design, Chapter 1.
Munzner. AK Peters Visualization Series, CRC Press, 2014.

Vignette 1:
Vis Tool for Investigative Reporting

Overview

The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists

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

https://www.overviewdocs.com

Matthew Brehmer
@mattbrehmer

Stephen Ingram
@FroweFace

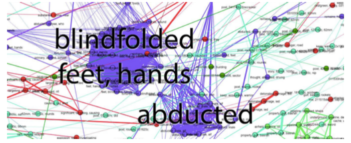
Jonathan Stray
@jonathanstray

Tamara Munzner
@tamaramunzner

Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists.
Brehmer, Ingram, Stray, and, Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2014), 20(12):2271-2280, 2014.

Origin story: WikiLeaks meets Glimmer

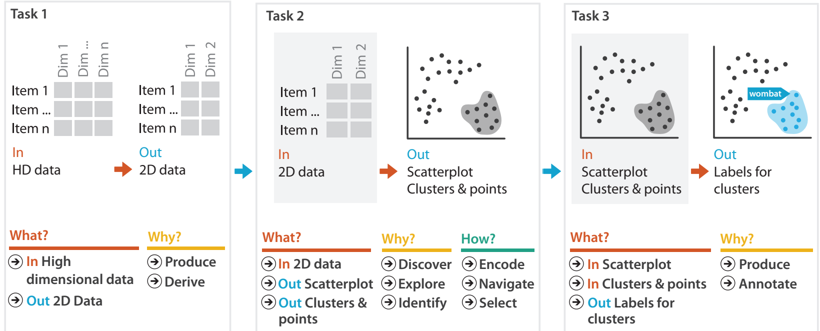
- WikiLeaks: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  - one instance of general problem: Too Many Documents
  - conjectured that existing label classification falls short of showing all meaningful structure in data
    - friendly action, criminal incident, ...
  - he had some NLP, needed better vis tools



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

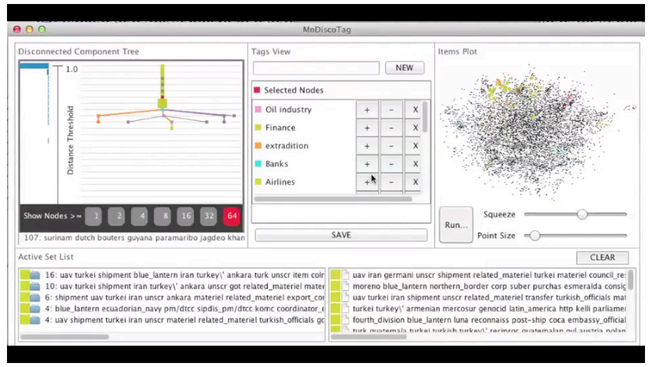
[Glimmer: Multilevel MDS on the GPU. Ingram, Munzner, Olano. IEEE TVCG 15(2):249-261, 2009.]

Starting point: Dimensionality reduction for document datasets



- more on DR: hour-long talk *Dimensionality Reduction from Several Angles*
  http://www.cs.ubc.ca/~tmm/talks.html#kelowna16

Overview: Early version



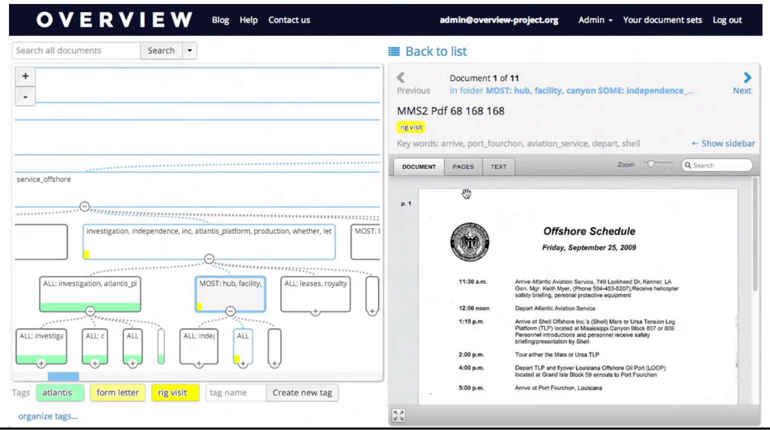
107: surinam dutch bouters guyana paramaribo jagdeo khan

16: uav turkei shipment blue_lantern iran turkey\' ankara turk unscr item colr
10: uav turkei shipment iran turkey\' ankara unscr got related_materiel mater
6: shipment uav turkei iran unscr ankara materiel related_materiel export_co
4: blue_lantern ecuadorian_navy pm/dtcc sipdis_pm/dtcc komc coordinator_e
4: uav shipment turkei iran unscr materiel related_materiel turkish_officials gc

uav iran germani unscr shipment related_materiel turkei materiel council_re
moreno blue_lantern northern_border corp suber purchas esmeralda consic
uav turkei iran shipment unscr related_materiel transfer turkish_officials mat
turkei turkey\' armenian mercosur genocid latin_america http kelli parliame
fourth_division blue_lantern luna reconnaiss post-ship coca embassy_official

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

Overview: current version



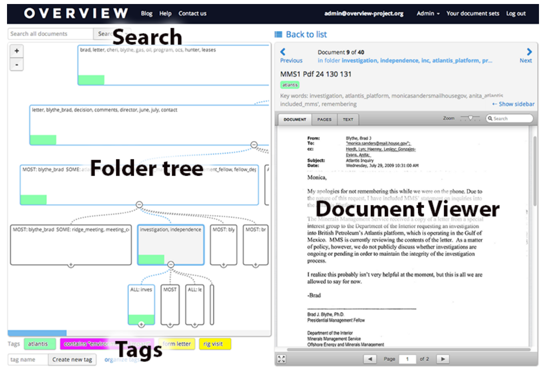
Document 1 of 11
in folder MOST: hub, facility, canyon SOME: independence_...
MMS2 Pdf 68 168 168
rig visit
Key words: arrive, port_fourchon, aviation_service, depart, shell

p. 1

Offshore Schedule
Friday, September 25, 2009

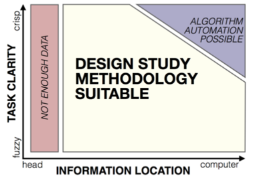
11:30 a.m.	Arrive Atlantic Aviation Service, 749 Lockheed Dr, Kenner, LA Gen. Mgr. Keith Myer, (Phone 504-453-6207);Receive helicopter safety briefing, personal protective equipment
12:00 noon	Depart Atlantic Aviation Service
1:15 p.m.	Arrive at Shell Offshore Inc.'s (Shell) Mars or Ursa Tension Leg Platform (TLP) located at Mississippi Canyon Block 807 or 809. Personnel introductions and personnel receive safety briefing/presentation by Shell.
2:00 p.m.	Tour either the Mars or Ursa TLP
4:00 p.m.	Depart TLP and flyover Louisiana Offshore Oil Port (LOOP) located at Grand Isle Block 59 enroute to Port Fourchon
5:00 p.m.	Arrive at Port Fourchon, Louisiana

Overview evolution: rationale?



Deploy in the real world

Case Study	#1	#2	#3	#4	#5	#6
Document Collection	4,500 pages from FOIA	5,996 emails from FOIA	8,680 pages from FOIA	1,278 survey comments	4,653 emails from FOIA	1,680 bills
Question	*What did security contractors do during Iraq war?*	*Were municipal police funds mismanaged?*	*Were Paul Ryan's campaign statements hypocritical?*	*What is the gun ownership debate about?*	*Was gov't response to emergency incident effective?*	*Did gov't fail to pass bills addressing police misconduct?*



Design Study Methodology

Reflections from the Trenches and from the Stacks

http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/

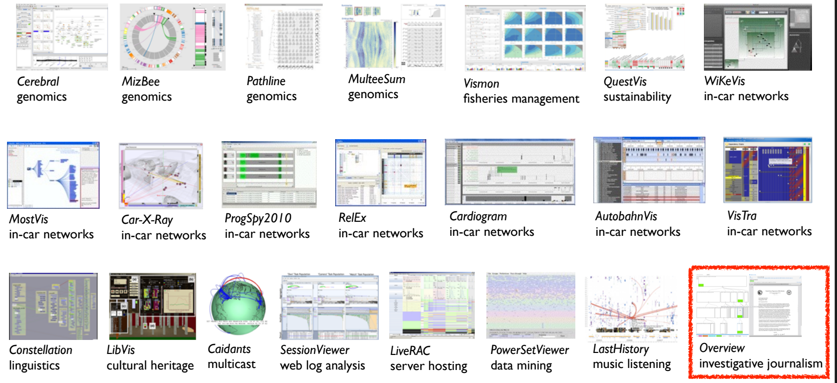
Michael Sedlmair

Miriah Meyer

Tamara Munzner
@tamaramunzner

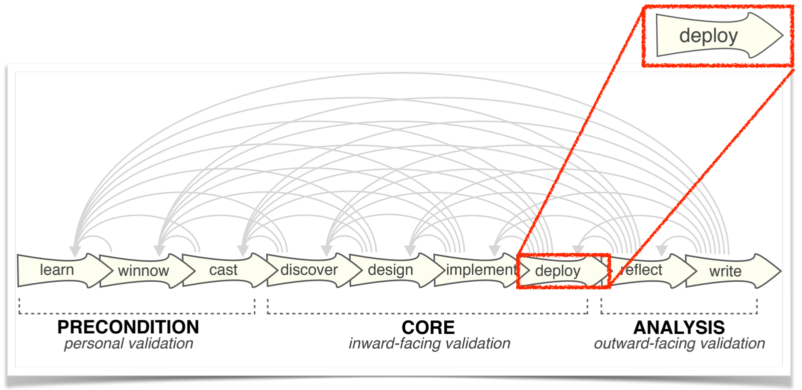
Design Study Methodology: Reflections from the Trenches and from the Stacks.
Sedlmair, Meyer, Munzner. IEEE Trans. Visualization and Computer Graphics 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

Design Studies: Lessons learned after 21 of them

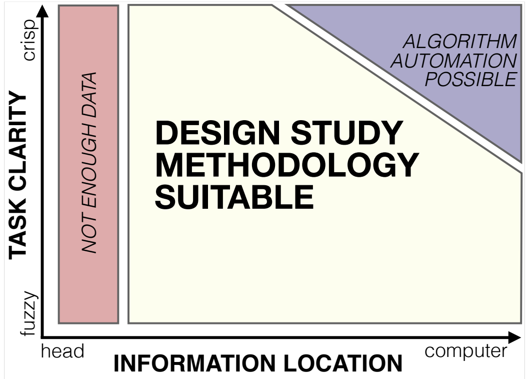


Cerebral genomics
MizBee genomics
Pathline genomics
MulteeSum genomics
Vismon fisheries management
QuestVis sustainability
WiKeVis in-car networks
MostVis in-car networks
Car-X-Ray in-car networks
ProgSpy2010 in-car networks
RelEx in-car networks
Cardiogram in-car networks
AutobahnVis in-car networks
VisTra in-car networks
Constellation linguistics
LibVis cultural heritage
Caidants multicast
SessionViewer web log analysis
LiveRAC server hosting
PowerSetViewer data mining
LastHistory music listening
Overview investigative journalism

Design study methodology: 9-stage framework



Design study methodology: definitions



Design study methodology: 32 Pitfalls

- and how to avoid them

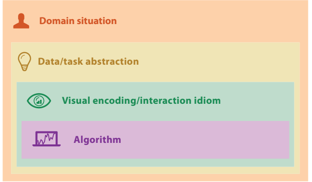
PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

Collaboration incentives

- why do CS/vis people need to understand journalism's problems?
  - we work with you to understand your driving problems
  - we build tools intended to help
    - only works out if we understood the problems deeply enough
  - we observe how you use them
    - if they're good enough
      - CS win: research success stories
      - journalist win: access to better tools
  - we develop guidelines on how to build better tools in general
    - CS win: research progress in visualization

Deploy in the real world, understand user goals

Case Study	#1	#2	#3	#4	#5	#6
Document Collection	4,500 pages from FOIA	5,996 emails from FOIA	8,680 pages from FOIA	1,278 survey comments	4,653 emails from FOIA	1,680 bills
Question	*What did security contractors do during Iraq war?*	*Were municipal police funds mismanaged?*	*Were Paul Ryan's campaign statements hypocritical?*	*What is the gun ownership debate about?*	*Was gov't response to emergency incident effective?*	*Did gov't fail to pass bills addressing police misconduct?*
		find the needle in the haystack				prove haystack contains no needles!

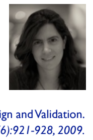


A Nested Model

for Visualization Design and Validation

www.cs.ubc.ca/labs/imager/tr/2009/NestedModel

Tamara Munzner
@tamaramunzner

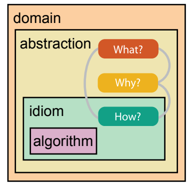


A Nested Model for Visualization Design and Validation.
Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 09), 15(6):921-928, 2009.

Nested model: Four levels of vis design

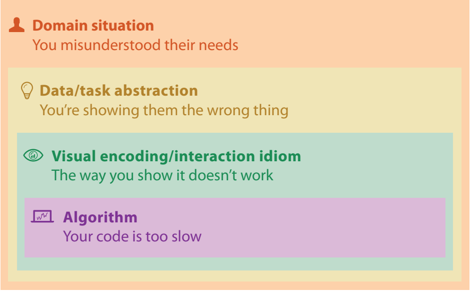
[A Nested Model of Visualization Design and Validation.
*Munzner. IEEE TVCG 15(6):921-928, 2009
(Proc. InfoVis 2009).]*

- *domain situation*
  - who are the target users?
    - CS: domain = journalism; journ: domain = story topic
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient computation



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

Threats to validity differ at each level



Domain situation
You misunderstood their needs

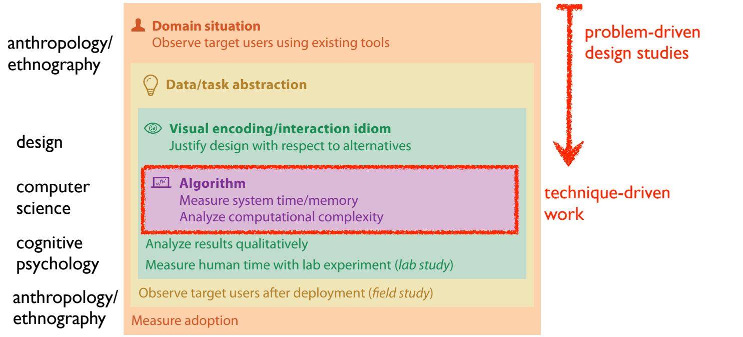
Data/task abstraction
You're showing them the wrong thing

Visual encoding/interaction idiom
The way you show it doesn't work

Algorithm
Your code is too slow

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Evaluate success at each level with methods from different fields



anthropology/ ethnography — **Domain situation**: Observe target users using existing tools

Data/task abstraction

design — **Visual encoding/interaction idiom**: Justify design with respect to alternatives

computer science — **Algorithm**: Measure system time/memory; Analyze computational complexity

cognitive psychology — Analyze results qualitatively; Measure human time with lab experiment (*lab study*)

anthropology/ ethnography — Observe target users after deployment (*field study*); Measure adoption

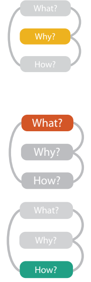
problem-driven design studies

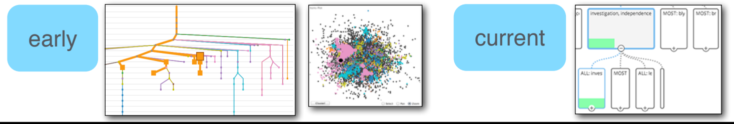
technique-driven work

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

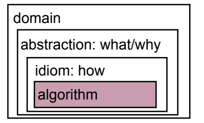
Evolution across levels

- evolution of task abstraction
  - task 1: generate hypotheses → explore → summarize
    - *obviously you can't read everything; speed up with tool for categorizing and counting*
  - task 2: verify hypotheses → locate → identify
    - *you really do read each doc; speed up with tool to keep track of findings*
- evolution of data abstraction & idioms
  - arrange cluster tree to emphasize nodes vs links
  - new vis insight: DR scatterplot less effective than cluster tree vis + tagging
    - better affordance for systematic traversal of document cluster hierarchy

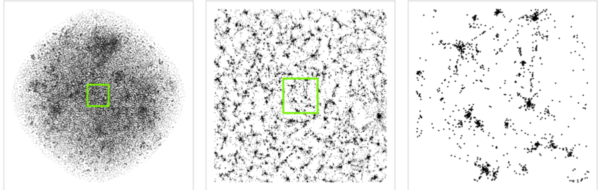




Algorithm: Spinoff series



- 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), Volume 150 Part B, p 557-569, 2015.]

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

Vignette 2:
Vis Tool for Journalistic Presentation

TimeLineCurator

Interactive Authoring of Visual Timelines from Unstructured Text

http://about.timelinecurator.org

http://timelinecurator.org

Johanna Fulda
@jofu_

Matthew Brehmer
@mattbrehmer

Tamara Munzner
@tamaramunzner

TimeLineCurator: Interactive Authoring of Visual Timelines from Unstructured Text.
Fulda, Brehmer, Munzner. IEEE Trans. Visualization and Computer Graphics (Proc IEEE VAST 2015) 22(1):300-309, 2015.

Origin story: Tedium in the newsroom

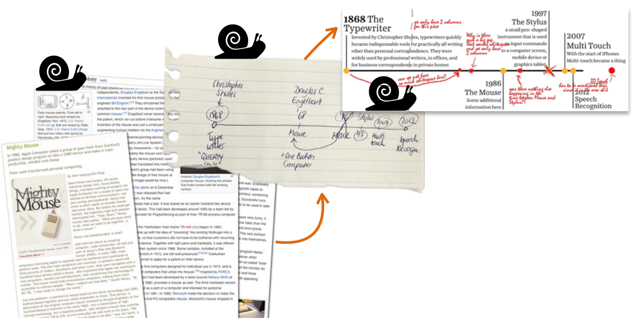
- Johanna Fulda: interactive infographics developer, Sueddeutsche Zeitung
  - then Munich CS master's student, visiting UBC
- what pain point could we address with interactive visualization?
  - plus some NLP
- sound familiar?…

TimeLineCurator
visual & browser-based

https://vimeo.com/jofu/tlc

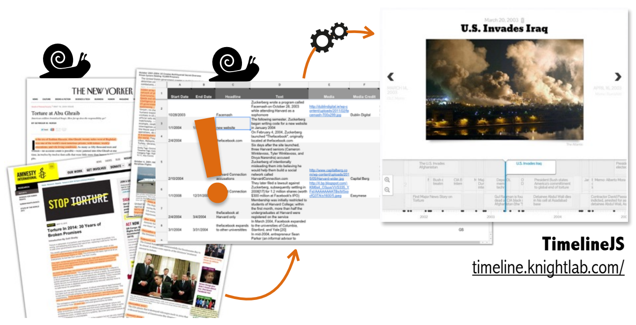
Manual creation process

Browse Extract Format Show Update



Structured creation process

Browse Extract Format Show Update



TimelineJS
timeline.knightlab.com/

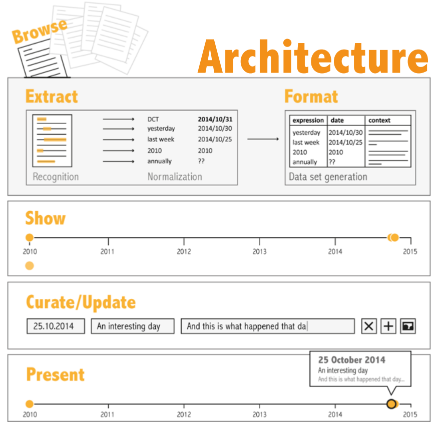
Timeline authoring model

- time required for each task

	Browse	Extract	Format	Show	Update
Manual Drawing	slow	slow		slow	slow
Structured Creation	slow	slow	slow	automated	fast
TimeLine Curator	fast	automated	automated	fast	fast

The general case for curation

- build for human in the loop as continuing need
  - automatic processing to accelerate not replace
  - assume computational results good but not perfect
    - for the indefinite future!
  - visual feedback to accelerate



The importance of being brisk

- sexy use case: eureka moment
  - enable what was impossible before
  - vis tools for new insights & discoveries
- workhorse use case: workflow speedup
  - vis tools to accelerate what you're already doing
    - sometimes enables the previously infeasible
- TLC use cases
  - started with speedup use case, for presentation
    - make this doc into a timeline now!
  - two other use cases nudge towards exploration
    - comparison between multiple timelines
    - speculative browsing

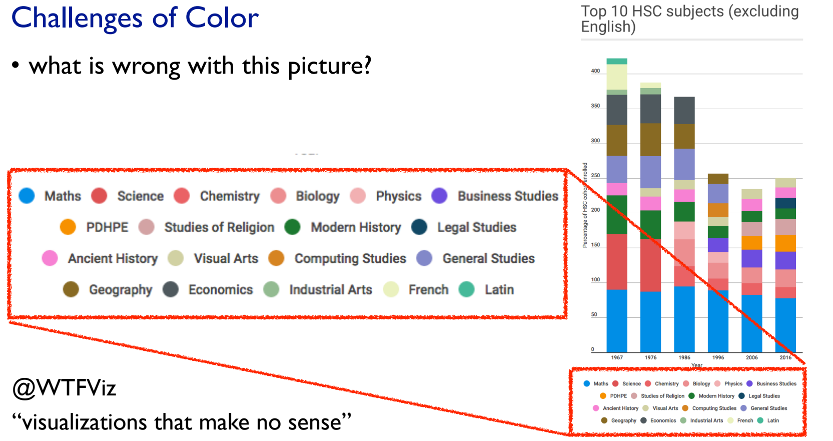
TimeLineCurator: Speculative Browsing

https://vimeo.com/jofu/tlc

Vignette 3: Challenges of Color (A Cautionary Tale)

Challenges of Color

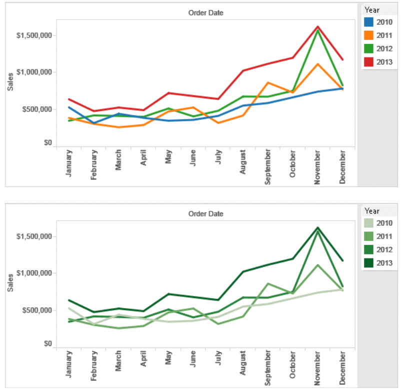
- what is wrong with this picture?

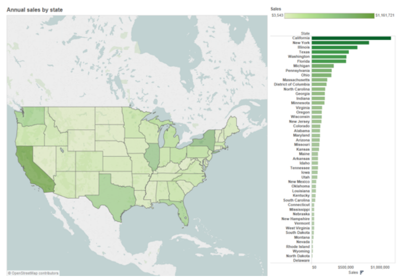


@WTFViz
"visualizations that make no sense"

http://viz.wtf/post/150780948819/maths-enrolments-drop-to-lowest-rate-in-50-years

Categorical vs ordered color

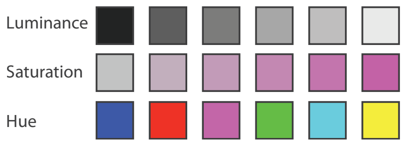




[Seriously Colorful: Advanced Color Principles & Practices. Stone. Tableau Customer Conference 2014.]

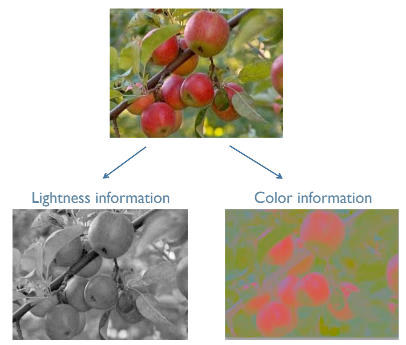
Decomposing color

- first rule of color: do not talk about color!
  - color is confusing if treated as monolithic
- decompose into three channels
  - ordered can show magnitude
    - luminance
    - saturation
  - categorical can show identity
    - hue
- channels have different properties
  - what they convey directly to perceptual system
  - how much they can convey: how many discriminable bins can we use?



Luminance

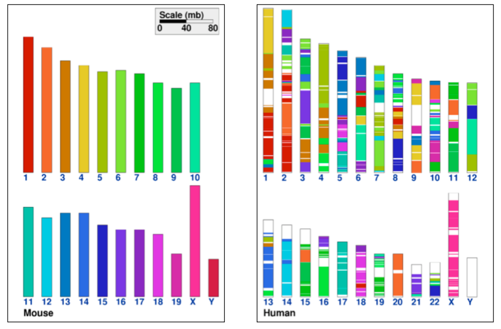
- need luminance for edge detection
  - fine-grained detail only visible through luminance contrast
  - legible text requires luminance contrast!
- intrinsic perceptual ordering



[Seriously Colorful: Advanced Color Principles & Practices. Stone. Tableau Customer Conference 2014.]

Categorical color: limited number of discriminable bins

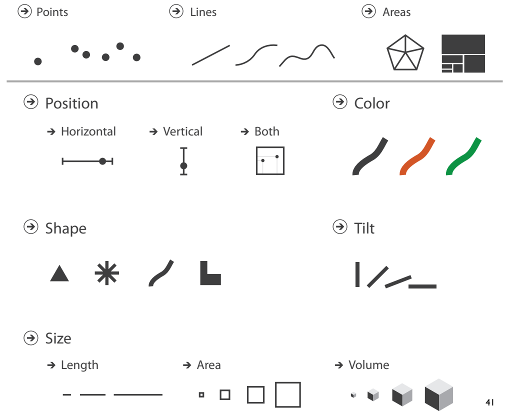
- human perception built on relative comparisons
  - great if color contiguous
  - surprisingly bad for absolute comparisons
- noncontiguous small regions of color
  - fewer bins than you want
  - rule of thumb: 6-12 bins, including background and highlights
  - so what can we do instead?



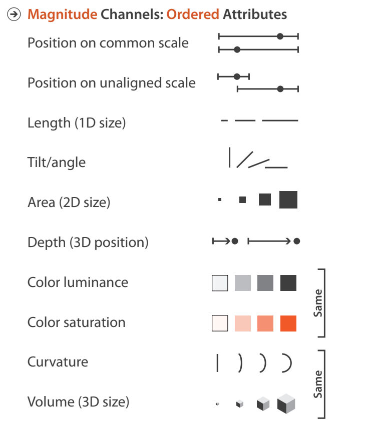
[Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. Sinha and Meller. BMC Bioinformatics, 8:82, 2007.]

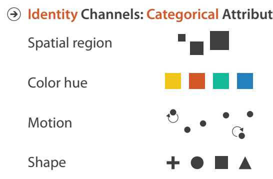
Analyzing visual encoding via marks and channels

- marks
  - geometric primitives
- channels
  - control appearance of marks
  - channel properties differ
    - type & amount of information that can be conveyed to human perceptual system
      - number of discriminable bins
      - show magnitude vs. identity
      - accuracy of perception

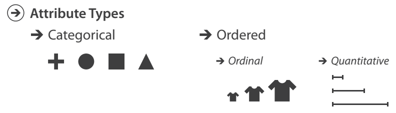


Channels: Matching expressiveness

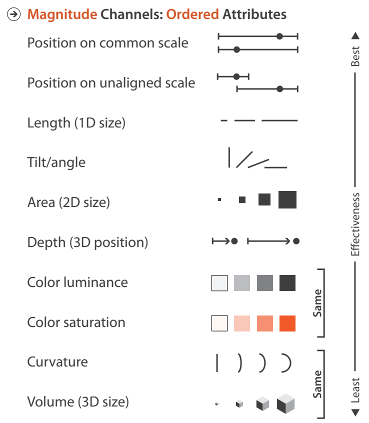


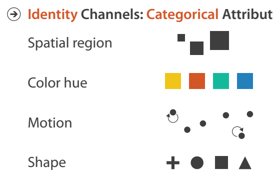


- expressiveness principle
  - match channel and data characteristics



Channels: Ranking effectiveness

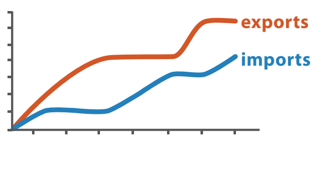




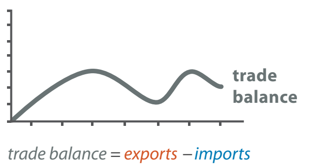
- expressiveness principle
  - match channel and data characteristics
- effectiveness principle
  - encode most important attributes with highest ranked channels

Derive

- don't just draw what you're given!
  - decide what the right thing to show is
  - create it with a series of transformations from the original dataset
  - draw that
- one of the four major strategies for handling complexity



Original Data



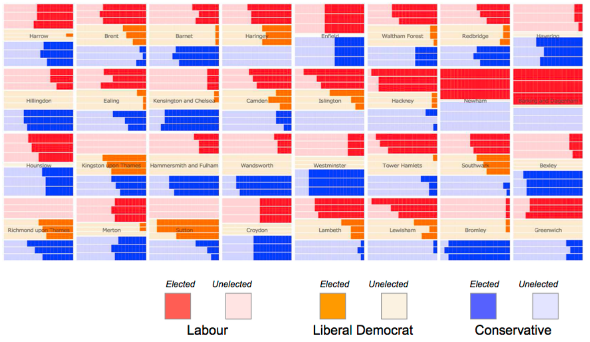
Derived Data

BallotMaps

- ballots in the UK are alphabetically ordered
  - govt: not sufficient to affect electoral outcome
  - vis researcher hunch: it matters!
- vis project
  - task: compare geographic regions of voting and spatial position of candidate name on ballot paper
  - data: Greater London elections 2010
    - geographic location, candidate name, alphabetical position in ballot, # candidate votes, party, elected/lost
  - color coding alone will not save the day!
  - derived new data
    - alphabetical position within the party
    - vote order within party

BallotMaps: deriving data

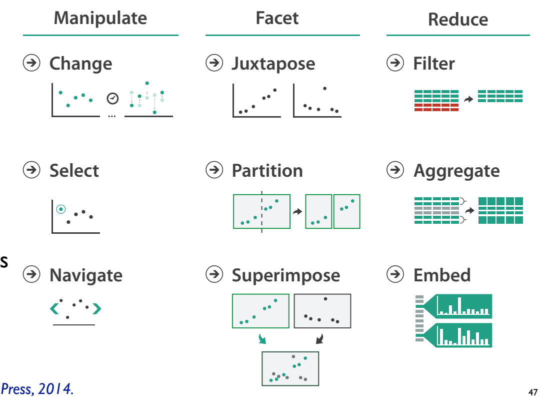
- bias exists in regions where systematic structure in bar lengths visible
  - yes in some
  - no in others



[BallotMaps: Detecting name bias in alphabetically ordered ballot papers. Wood, Badawood, Dykes, Slingsby. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2011), 17(12): 2384-2381, 2011]

Four strategies to handle complexity

- derive new data to show within view
- change view over time
- facet across multiple views
- reduce items/attributes within single view



more at:
Visualization Analysis and Design.
Munzner. AK Peters Visualization Series, CRC Press, 2014.

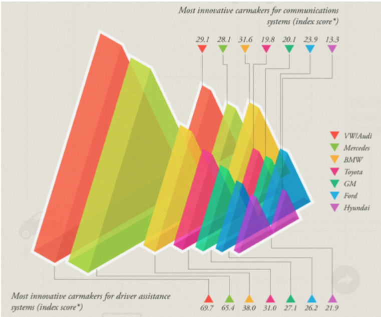
Vignette 4: Difficulties of Depth (Another Cautionary Tale)

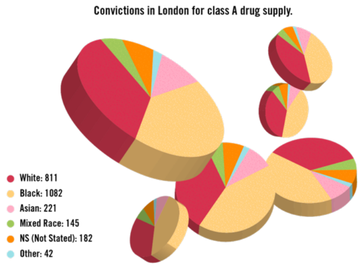
Visual encoding: 2D vs 3D

- 2D good, 3D better?
  - not so fast…

http://amberleyromo.com/images/Bookcover/Animal-Farm.png

Unjustified 3D all too common, in the news and elsewhere



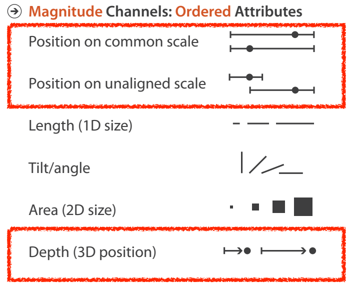


http://viz.wtf/post/137826497077/eye-popping-3d-triangles

http://viz.wtf/post/139002022202/designer-drugs-ht-ducqn

Depth vs power of the plane

- high-ranked spatial position channels: **planar** spatial position
  - not depth!



Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

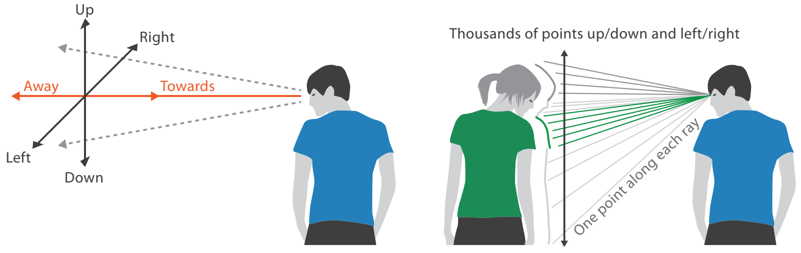
Tilt/angle

Area (2D size)

Depth (3D position)

Life in 3D?…

- we don't really live in 3D: we see in 2.05D
  - acquire more info on image plane quickly from eye movements
  - acquire more info for depth slower, from head/body motion

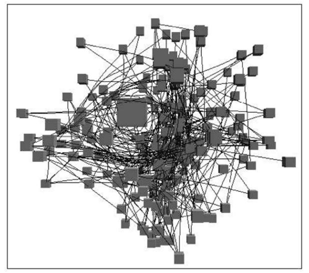


We can only see the outside shell of the world

[adapted from Visual Thinking for Design. Ware. Morgan Kaufmann 2010.]

Occlusion hides information

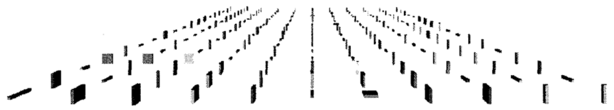
- occlusion
- interaction complexity



[Distortion Viewing Techniques for 3D Data. Carpendale et al. InfoVis1996.]

Perspective distortion loses information

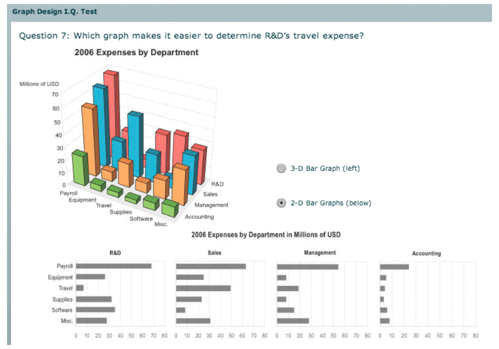
- perspective distortion
  - interferes with all size channel encodings
  - power of the plane is lost!



[Visualizing the Results of Multimedia Web Search Engines. Mukherjea, Hirata, and Hara. InfoVis 96]

3D vs 2D bar charts

- 3D bars never a good idea!



Graph Design I.Q. Test

Question 7: Which graph makes it easier to determine R&D's travel expense?

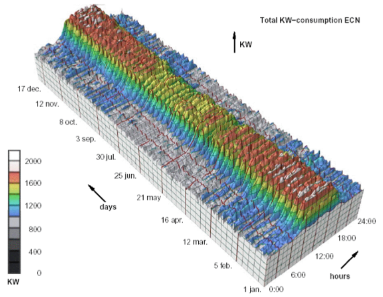
○ 3-D Bar Graph (left)

○ 2-D Bar Graphs (below)

[http://perceptualedge.com/files/GraphDesignIQ.html]

No unjustified 3D example: Time-series data

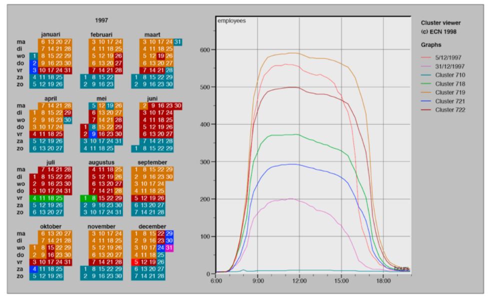
- extruded curves: detailed comparisons impossible



[Cluster and Calendar based Visualization of Time Series Data. van Wijk and van Selow, Proc. InfoVis 99.]

No unjustified 3D example: Transform for new data abstraction

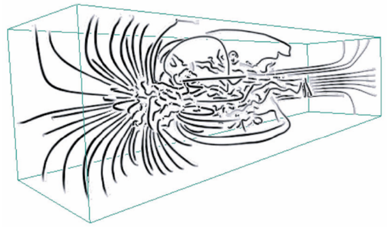
- derived data: cluster hierarchy
- juxtapose multiple views: calendar, superimposed 2D curves



[Cluster and Calendar based Visualization of Time Series Data. van Wijk and van Selow, Proc. InfoVis 99.]

Justified 3D: shape perception

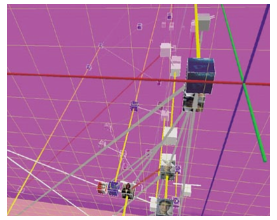
- benefits outweigh costs when task is shape perception for 3D spatial data
  - interactive navigation supports synthesis across many viewpoints



[Image-Based Streamline Generation and Rendering. Li and Shen. IEEE Trans. Visualization and Computer Graphics (TVCG) 13:3 (2007), 630–640.]

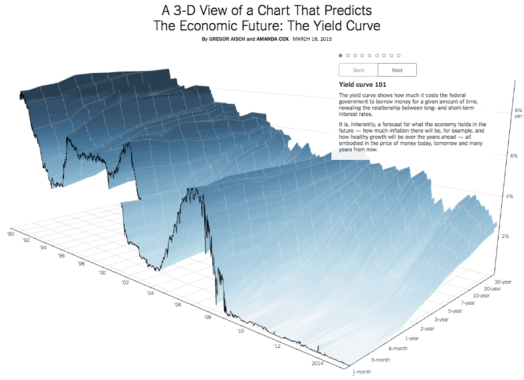
No unjustified 3D

- 3D legitimate for true 3D spatial data
- 3D needs very careful justification for abstract data
  - enthusiasm in 1990s, but now skepticism
  - be especially careful with 3D for point clouds or networks



[WEBPATH-a three dimensional Web history. Frecon and Smith. Proc. InfoVis 1999]

Justified 3D: Economic growth curve

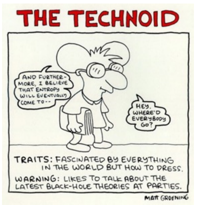


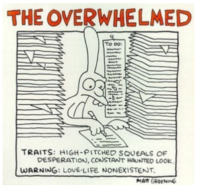
A 3-D View of a Chart That Predicts The Economic Future: The Yield Curve

http://www.nytimes.com/interactive/2015/03/19/upshot/3d-yield-curve-economic-growth.html

Wrap-up

- a tale of two tools
  - exploration: Overview
    - collaboration between CS and journalism: methods & rewards
    - reasoning about four levels of vis design
  - presentation: TimeLineCurator
    - visual curation of imperfect computational results
    - the importance of being brisk: speedup vs eureka moment
- two cautionary tales
  - guidance on color & 3D from vis literature

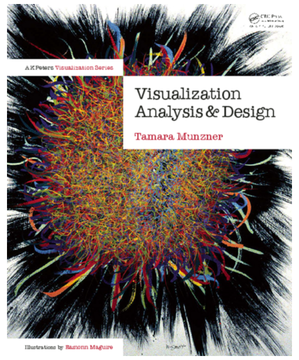

THE TECHNOID
TRAITS: FASCINATED BY EVERYTHING IN THE WORLD BUT HOW TO DRESS.
WARNING: LIKES TO TALK ABOUT THE LATEST BLACK-HOLE THEORIES AT PARTIES.


THE OVERWHELMED
TRAITS: HIGH-PITCHED SQUEALS OF DESPERATION, CONSTANT HAUNTED LOOK.
WARNING: LOVE LIFE NONEXISTENT.

More Information

@tamaramunzner

- this talk
  www.cs.ubc.ca/~tmm/talks.html#cj16
- book
  http://www.cs.ubc.ca/~tmm/vadbook
  - 20% off promo code, book+ebook combo: HVN17
  - http://www.crcpress.com/product/isbn/9781466508910
- papers, videos, software, talks, courses
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm


A K Peters Visualization Series
Visualization Analysis & Design
Tamara Munzner
Illustrations by Eamonn Maguire

Visualization Analysis and Design.
Munzner. A K Peters Visualization Series, CRC Press, Visualization Series, 2014.