Visualization Analysis and Design for Business Intelligence

Tamara Munzner
Department of Computer Science
University of British Columbia

Disney Research
20 July 2015, Glendale CA

http://www.cs.ubc.ca/~tmm/talks.html#disney15 **@tamaramunzner**

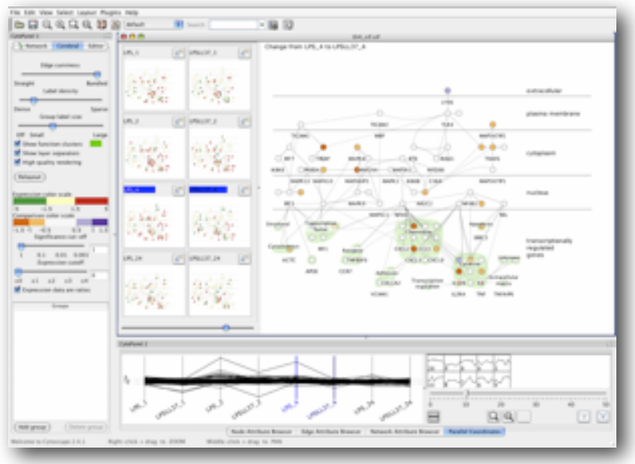
Outline

- interactive visual analysis
  - role and advantages

- LiveRAC
  - time-series data: managed web hosting (with AT&T)

- Cerebral
  - network of relationships: genes (with Agilent and UBC Immunology)

- wrapup



Defining visualization (vis)

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

- what counts as effective?
  - novel: enable entirely new kinds of analysis
  - faster: speed up existing workflows
    - most common case!

Why have a human in the loop?

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

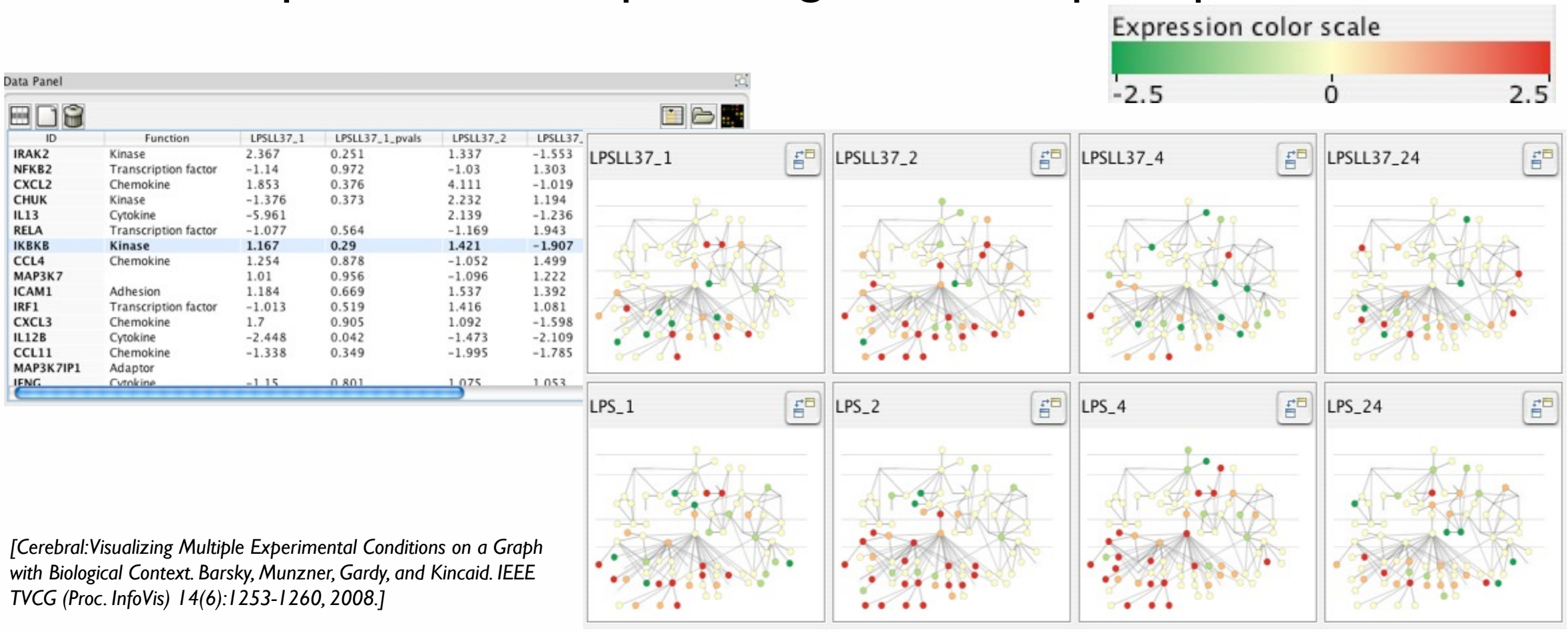
Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

- don't need vis when fully automatic solution exists and is trusted
- many analysis problems ill-specified
  - don't know exactly what questions to ask in advance
- possibilities
  - long-term use for end users (e.g. exploratory analysis of scientific data)
  - presentation of known results
  - stepping stone to better understanding of requirements before developing models
  - help developers of automatic solution refine/debug, determine parameters
  - help end users of automatic solutions verify, build trust

Why use an external representation?

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

- external representation: replace cognition with perception



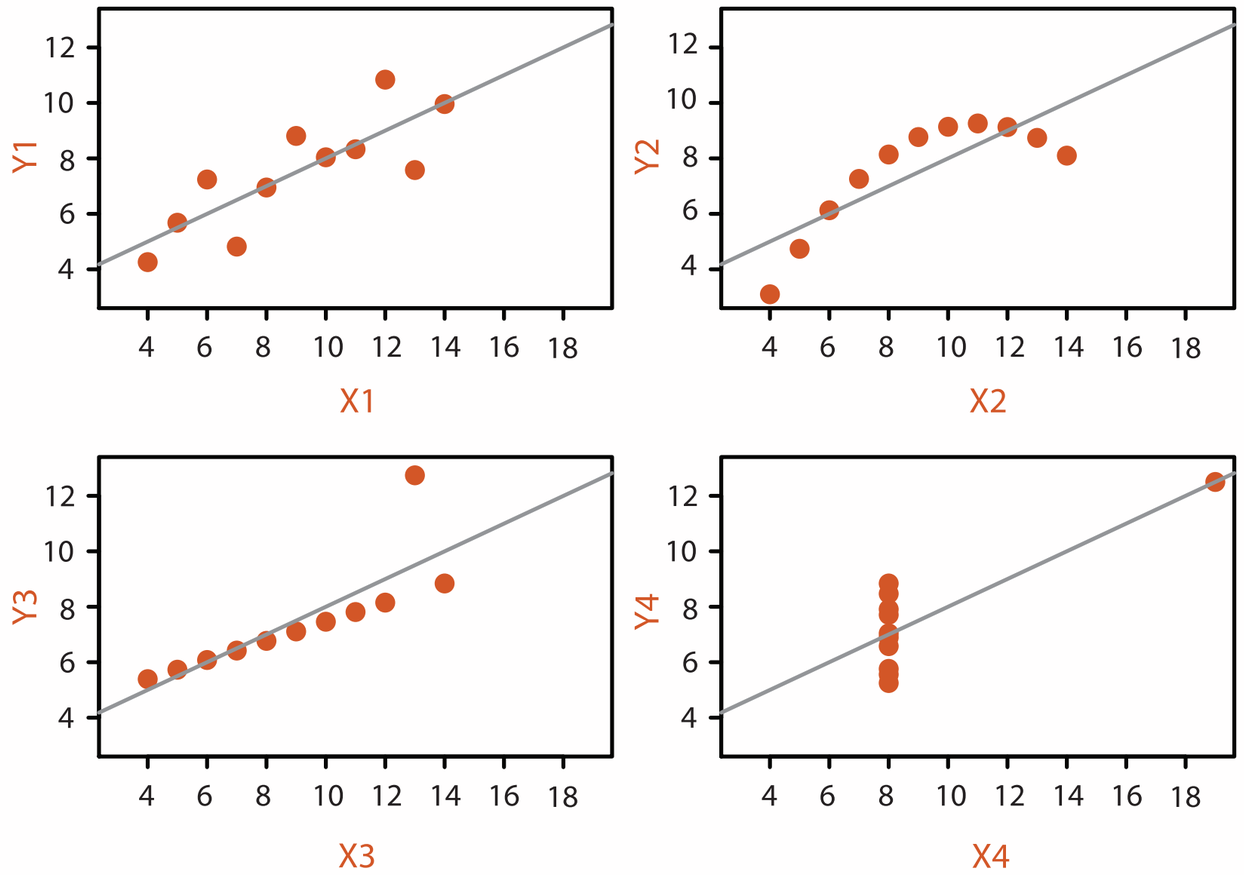
[Cerebral:Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky, Munzner, Gardy, and Kincaid. IEEE TVCG (Proc. InfoVis) 14(6):1253-1260, 2008.]

Why show the data in detail?

- 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



What are the resource limitations?

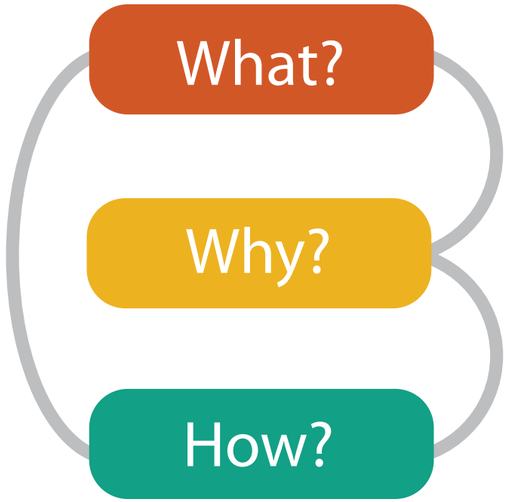
Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.

- computational limits
  - processing time
  - system memory
- human limits
  - human attention and memory
- display limits
  - pixels are precious resource, the most constrained resource
  - **information density**: ratio of space used to encode info vs unused whitespace
    - tradeoff between clutter and wasting space, find sweet spot between dense and sparse

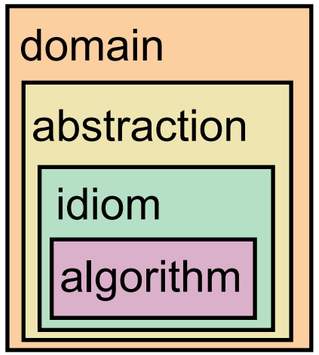
How to analyze vis design?

Vis usage can be analyzed in terms of what data is shown, why the user needs it, and how the idiom is designed.

- abstractions
  - translate from specifics of domain to vocabulary of vis
    - *data abstraction*: **what** to show
      - might not draw what you're given: transform data into form useful for task
    - *task abstraction*: **why** they're looking at it
- idioms
  - *visual encoding idiom*: **how** to draw
  - *interaction idiom*: **how** to manipulate
- analysis framework: scaffold to think systematically about design space
  - huge, and most possibilities ineffective for particular task/data combination



How to validate design?



anthropology/ ethnography

design

computer science

cognitive psychology

anthropology/ ethnography

Domain situation
Observe target users using existing tools

Data/task abstraction

Visual encoding/interaction idiom
Justify design with respect to alternatives

Algorithm
Measure system time/memory
Analyze computational complexity

Analyze results qualitatively

Measure human time with lab experiment (*lab study*)

Observe target users after deployment (*field study*)

Measure adoption

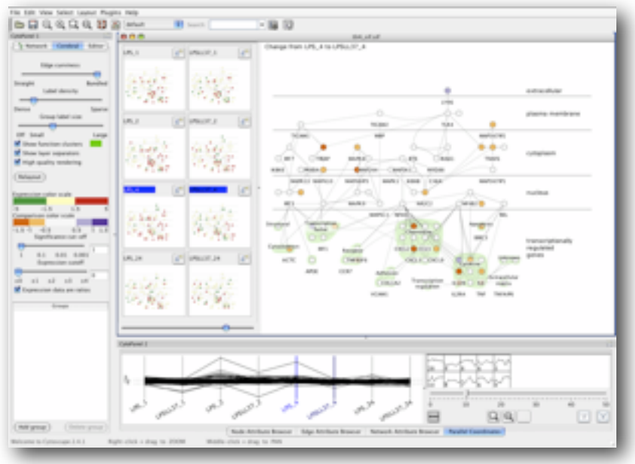
problem-driven work

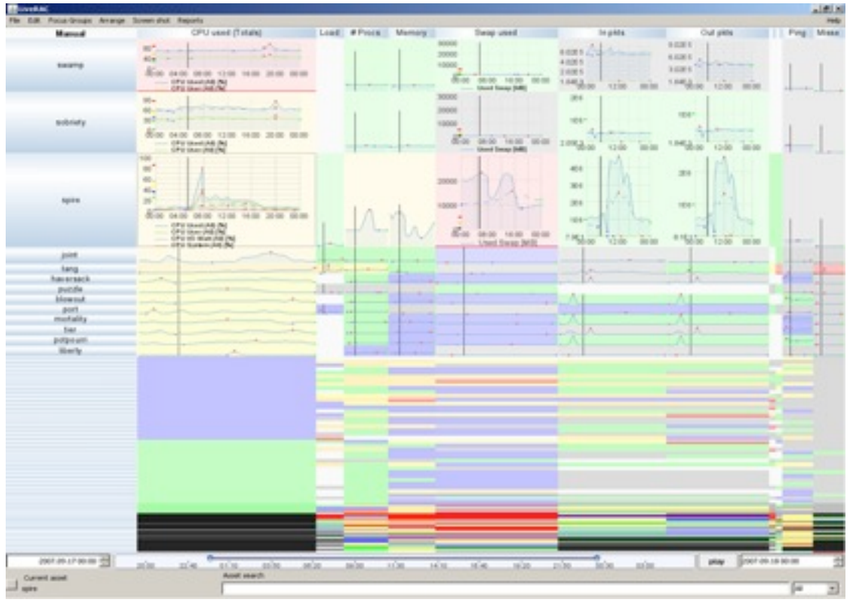
technique-driven work

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

Outline

- interactive visual analysis
  - role and advantages
- LiveRAC
  - time-series data: managed web hosting (with AT&T)
- Cerebral
  - network of relationships: genes (with Agilent and UBC Immunology)
- wrapup





LiveRAC

Interactive Visual Exploration of System Management Time-Series Data

joint work with:

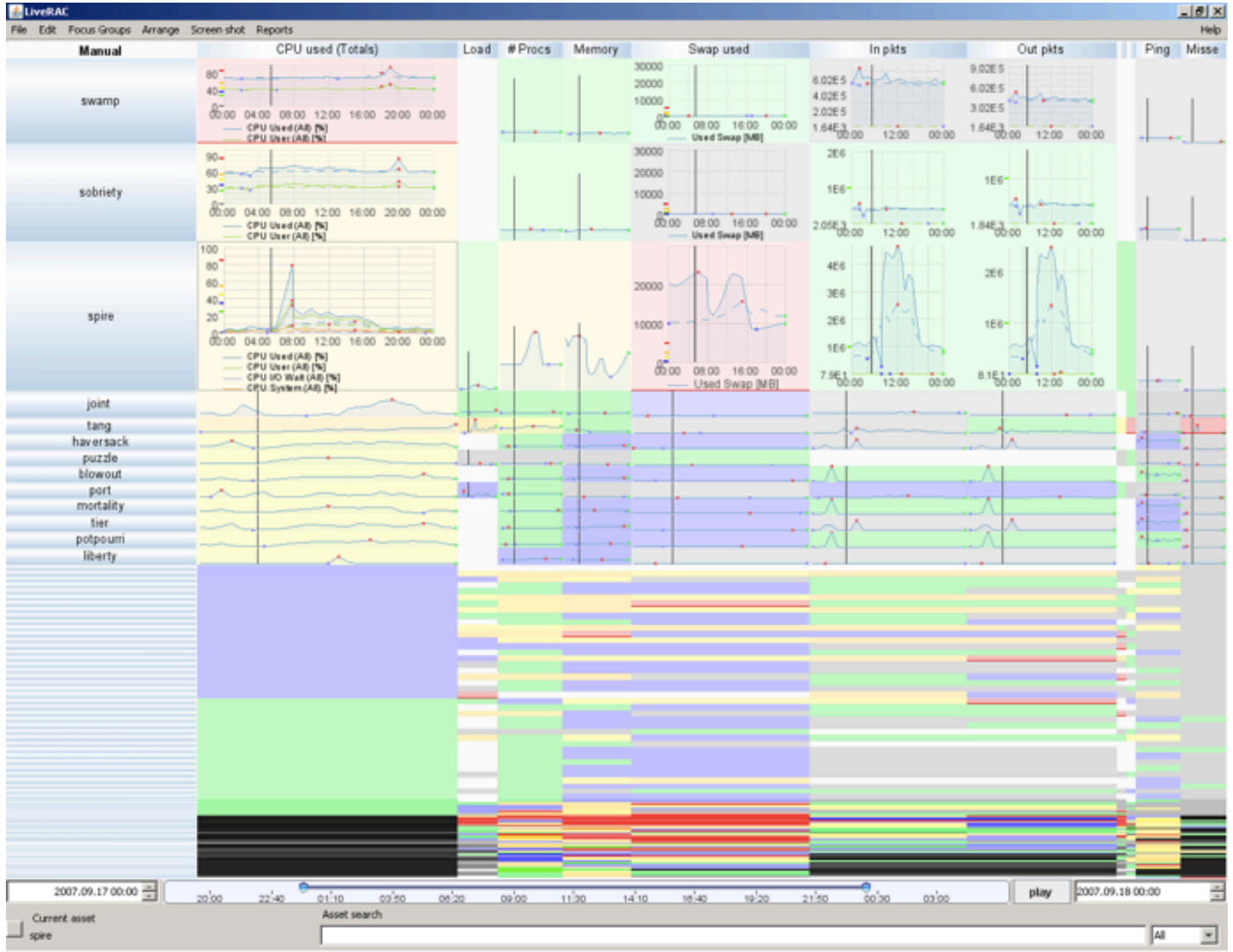
Peter McLachlan, Eleftherios Koutsofios, Stephen North.

http://www.cs.ubc.ca/labs/imager/tr/2008/liverac

LiveRAC - Interactive Visual Exploration of System Management Time-Series Data.

McLachlan, Munzner, Koutsofios, North. Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'08), p 1483-1492, 2008.

LiveRAC video

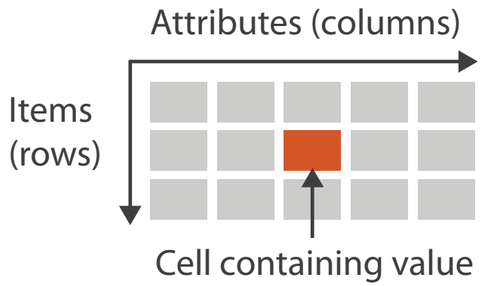


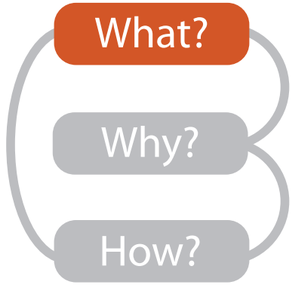
http://youtu.be/Id0c3H0VSkw

What: Data abstraction

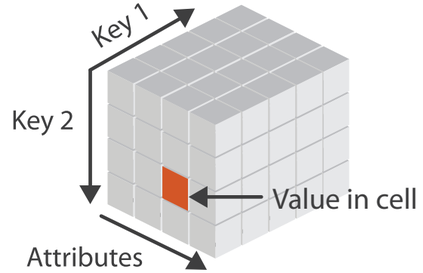
- multidimensional table: time series data
  - key attributes
    - time
      - 50,000: 5-minute intervals over 6 months
      - multiscale levels of interest
    - devices
      - 4000
    - parameters
      - 20
      - ex: CPU usage, memory load, network traffic, alarms, ...
  - value attributes
    - parameter value for device at time point
      - quantitative
    - device groups
      - categorical

➔ Tables





➔ *Multidimensional Table*



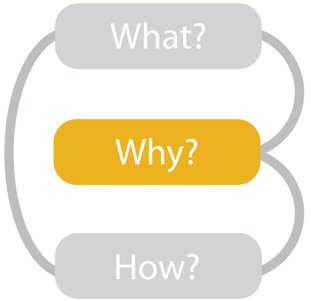
➔ **Attribute Types**

➔ Categorical

➔ Ordered

➔ *Quantitative*

Why: Tasks in domain language

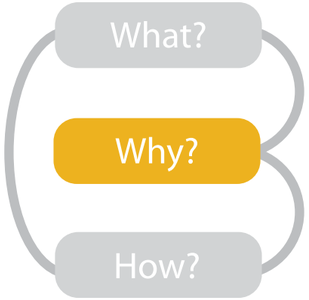


- interpret network environment status
- report generation
- capacity planning
- event investigation/forensics
- coordination
  - between customers, engineering, ops

Why: Task abstraction

- browse and correlate across combinations of parameter, device, time
  - correlate alarm attribute with other parameter attribs
  - find trends across groups of devices
  - summarize over different time intervals
  - identify devices at or beyond parameter thresholds
  - identify critical parameter values
  - compare device behavior at specific event times

→ *Discover*



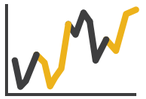
Targets

→ All Data

→ Trends

→ Outliers

→ Features

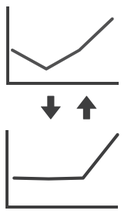


Actions

→ Query

→ Identify

→ Compare

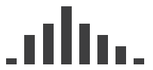


→ Summarise

→ Attributes

→ One

→ *Distribution*

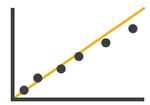


→ *Extremes*

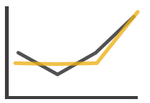
→ Many

→ *Dependency*

→ *Correlation*

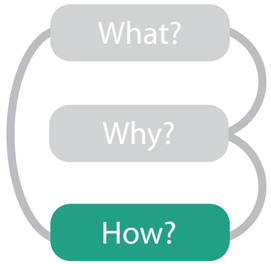


→ *Similarity*



How: Navigate

- semantic zooming: adapts to pixels available
  - many: superimposed line charts with full labeling
  - some: iconic line chart (sparkline)
  - few: color-coded box (heatmap)



Manipulate

Change View Over Time

Navigate

→ Item Reduction

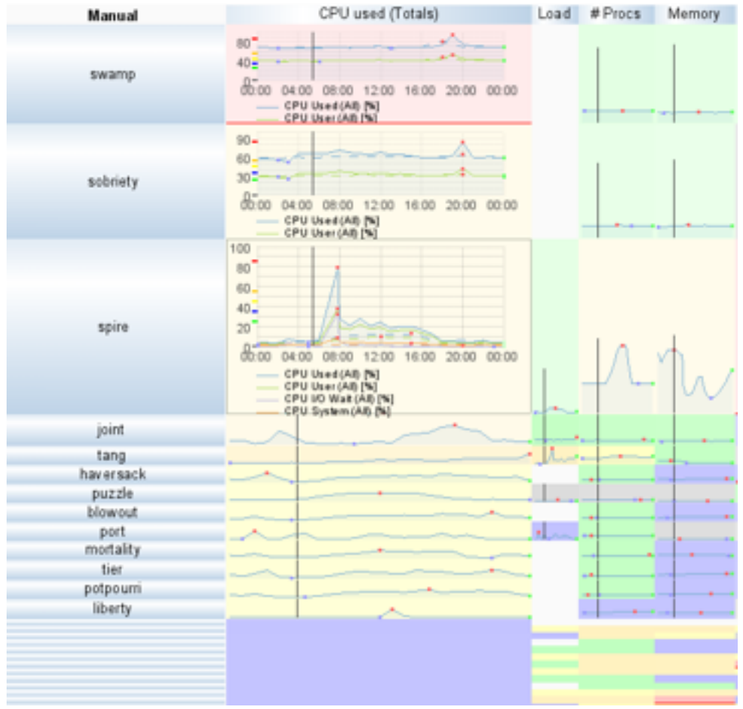
→ *Zoom*
Geometric or *Semantic*

→ *Pan/Translate*

→ *Constrained*

How: Reduce

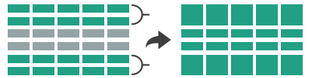
- reduce data shown with complex combination of filtering and aggregation
  - embed focus+context in single view
  - distort geometry
    - metaphor: stretch and squish navigation
    - shape: rectilinear
    - foci: multiple
    - impact: global



Reduce

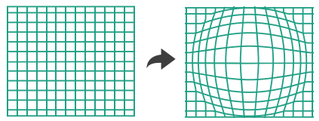
Filter

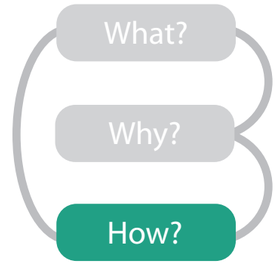
Aggregate



Embed

→ Distort Geometry





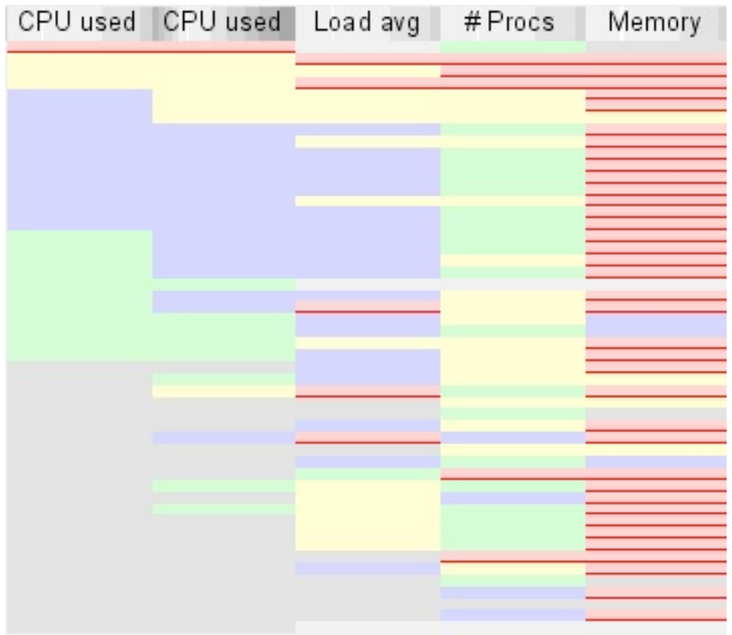
Manipulate

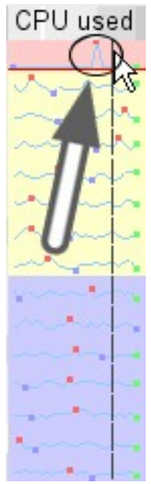
Navigate

→ *Constrained*

How: Reordering

- change spatial arrangement
  - resort by selected attribute
  - check for correlations between aligned attribute columns
    - ex: high load without high CPU, maybe I/O bound

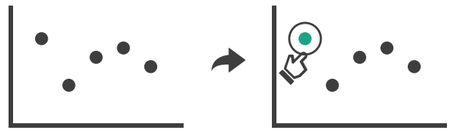




Manipulate

Change View Over Time

Select

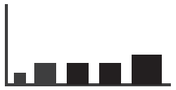


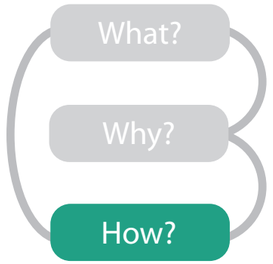
Encode

Arrange

→ Order

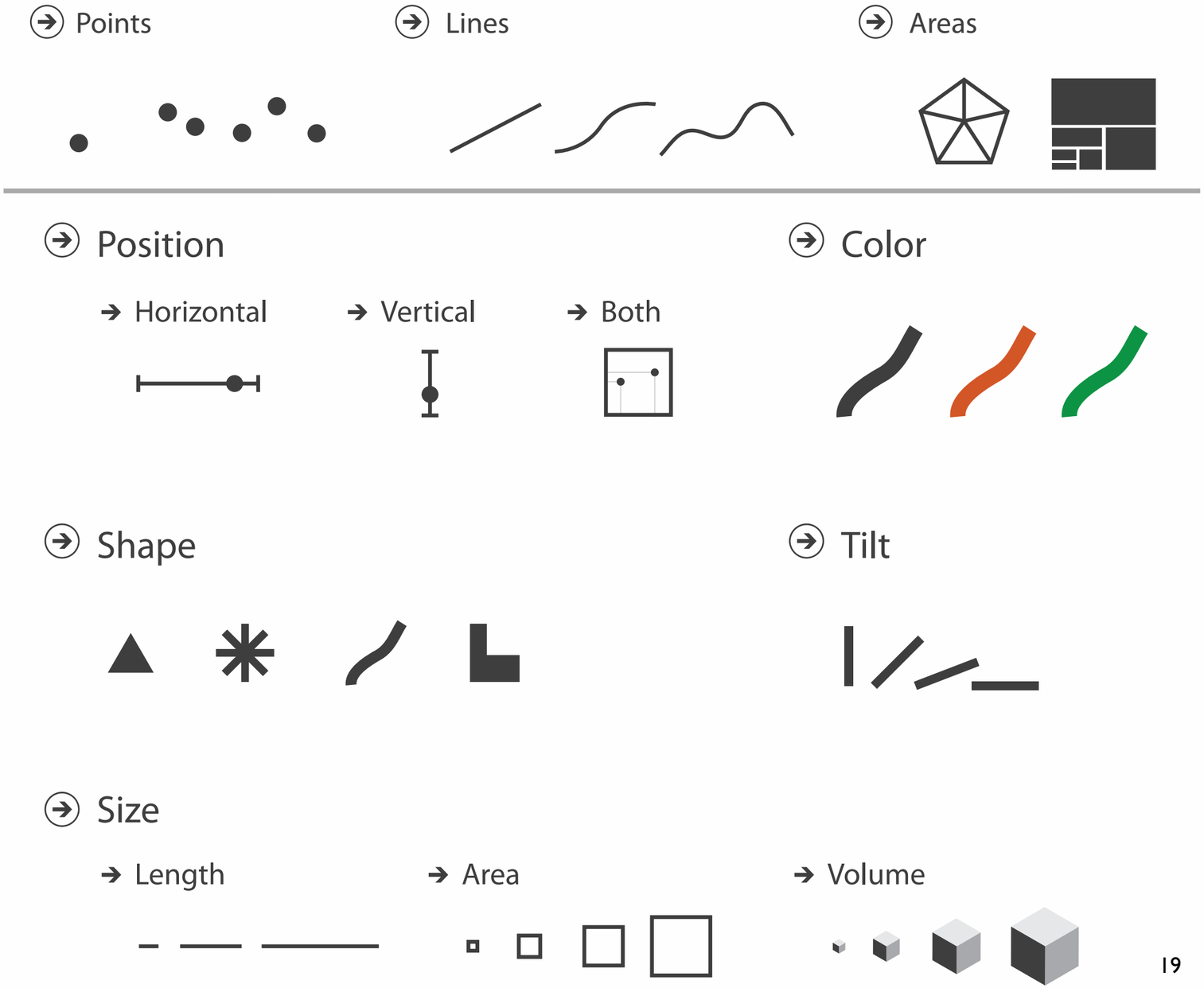
→ Align





Importance of arranging space: Underlying definitions

- marks
  - geometric primitives
- channels
  - control appearance of marks



Channels: Expressiveness types and effectiveness rankings

Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

Same

Same

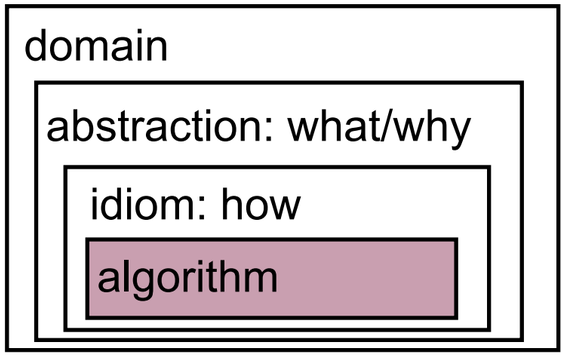
Effectiveness: Best ▲ ▼ Least

Identity Channels: Categorical Attributes

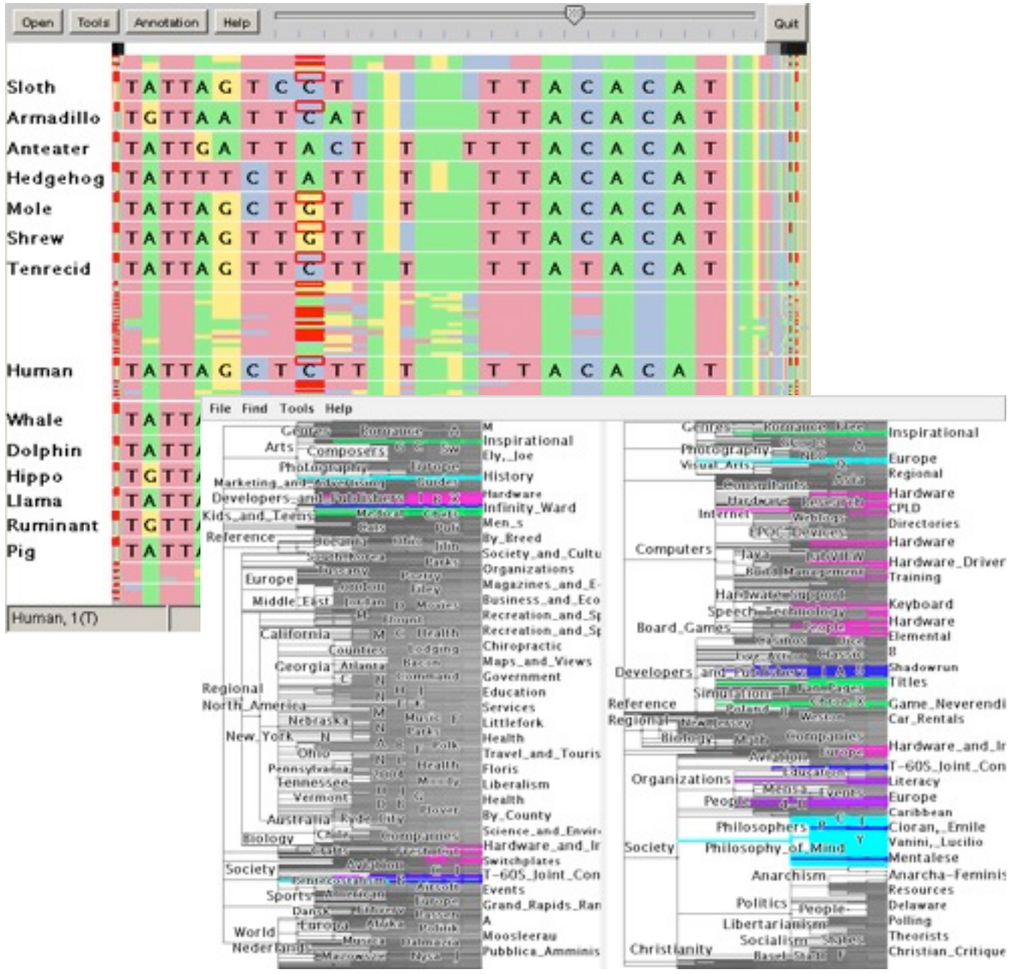
- Spatial region
- Color hue
- Motion
- Shape

- spatial position channels best in both cases
  - high accuracy
- more on channel rankings: hour-long talk *Visualization Principles*
  http://www.cs.ubc.ca/~tmm/talks.html#networkbio12

Algorithms



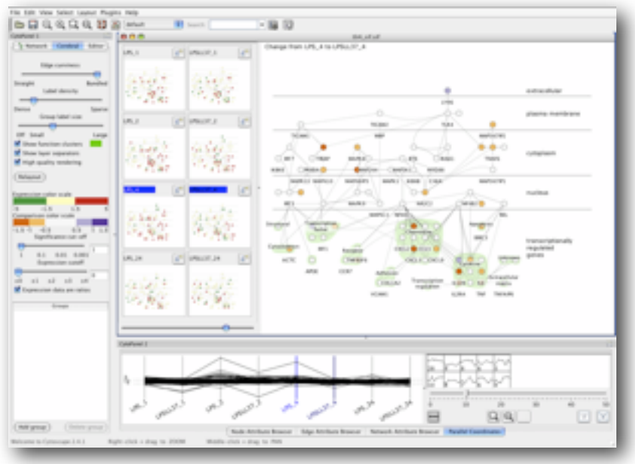
- back end: SWIFT server
- front end: PRISAD rendering
  - separate threads for render vs server update
  - guaranteed visibility of semantically important marks even when squished small
    - sublinear rendering: O(p) where p = pixel count
      - scalable for n of millions
      - generic framework
        - » time series charts, gene sequences, trees

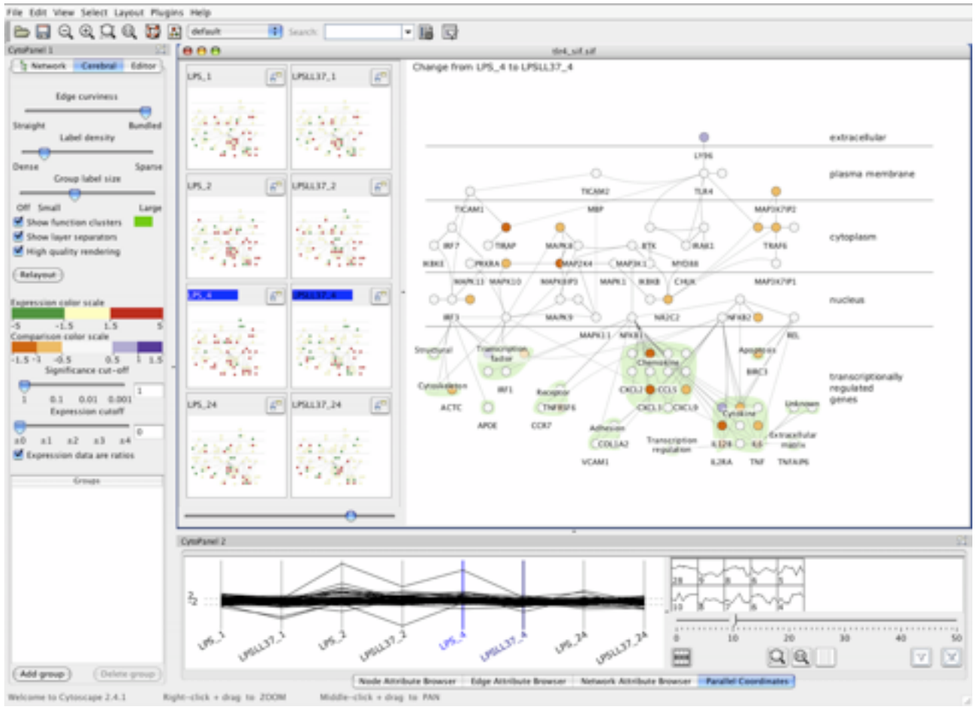


[Partitioned Rendering Infrastructure for Scalable Accordion Drawing (Extended Version). Slack, Hildebrand, and Munzner. Information Visualization, 5(2), p. 137-151, 2006.]

Outline

- interactive visual analysis
  - role and advantages
- LiveRAC
  - time-series data: managed web hosting (with AT&T)
- Cerebral
  - network of relationships: genes (with Agilent and UBC Immunology)
- wrapup





Cerebral

Visualizing Multiple Experimental Conditions on a Graph with Biological Context

joint work with:

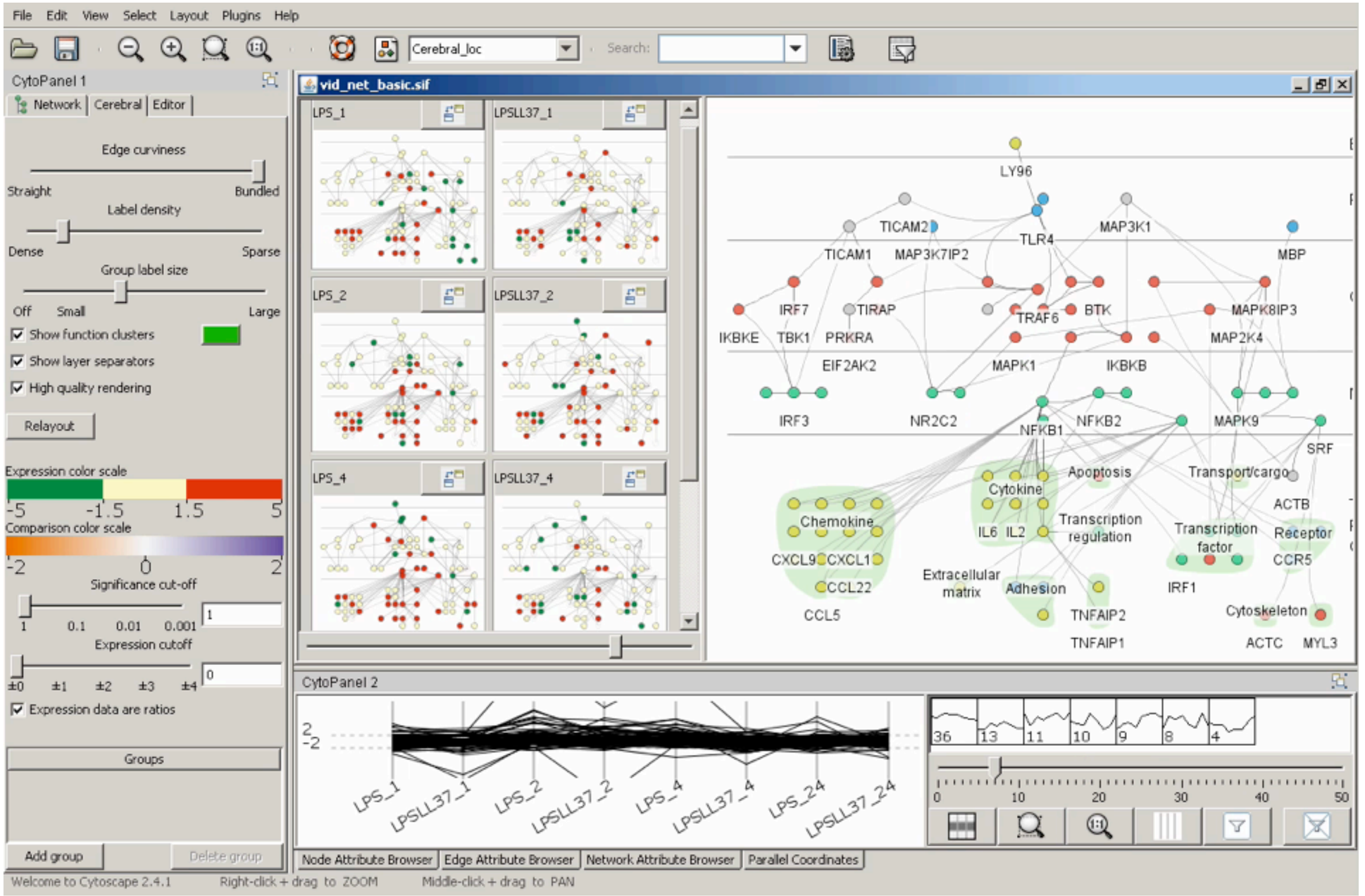
Aaron Barsky, Jennifer Gardy, Robert Kincaid

http://www.pathogenomics.ca/cerebral/

Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context.

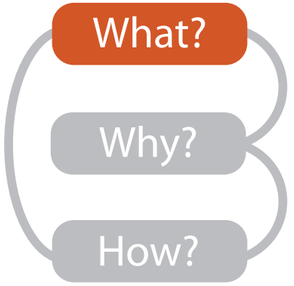
Barsky, Munzner, Gardy, Kincaid. IEEE Trans. Visualization and Computer Graphics 14(6):1253-1260 2008. (Proc. InfoVis 2008).

Cerebral video

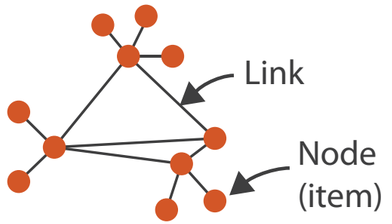


What: Data abstraction

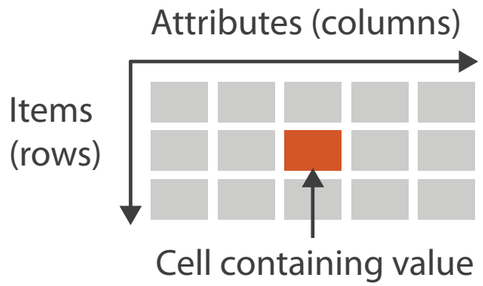
- dataset types
  - network
    - nodes: genes
    - links: known interactions between genes
  - table
    - quantitative attributes
      - gene expression levels for nodes across different experimental conditions
    - categorical attributes
      - subcellular location of interaction
      - functional groups



➔ Networks



➔ Tables



➔ **Attribute Types**

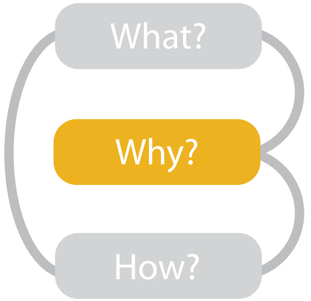
➔ Categorical

➔ Ordered

➔ *Ordinal*

➔ *Quantitative*

Why: Task abstraction

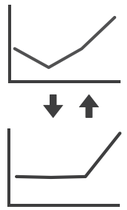


- task: interpret experiment results with respect to gene network
  - goal: accelerate existing discovery workflow
  - compare distributions between attributes
    - different experiments
  - interpret attributes in context of known network structure

➔ *Discover*

Actions

➔ Compare

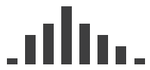


Targets

➔ Attributes

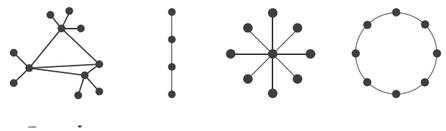
➔ One

➔ *Distribution*



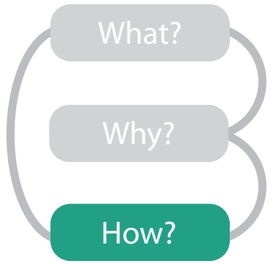
➔ Network Data

➔ Topology



How: Idiom design decisions

- arrange space for networks
  - custom node-link diagram layout
    - points for nodes
    - connection marks for links
  - vertical compartment according to subcellular location attribute
  - cluster according to functional grouping

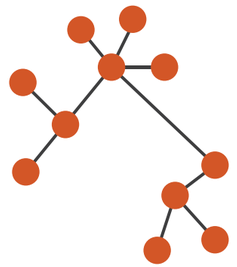


Arrange Networks And Trees

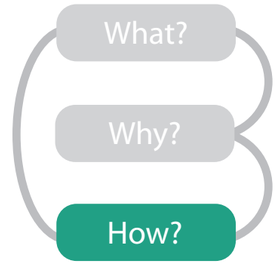
Node-link Diagrams

Connections and Marks

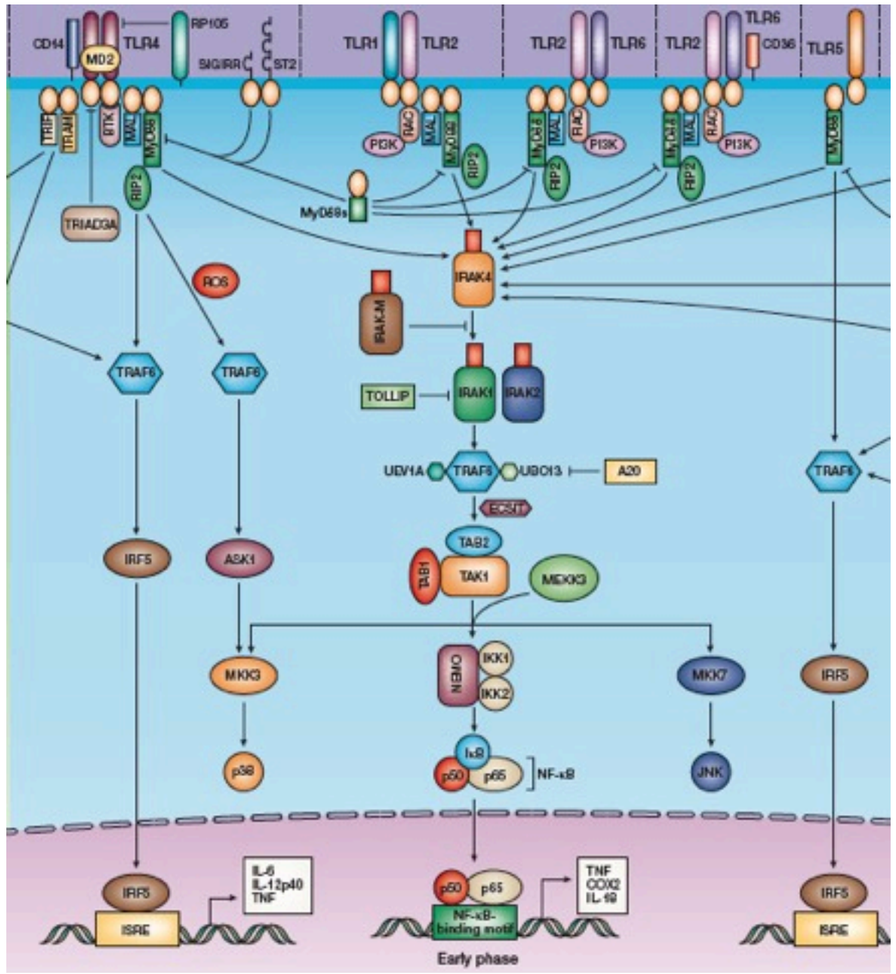
✔ NETWORKS ✔ TREES

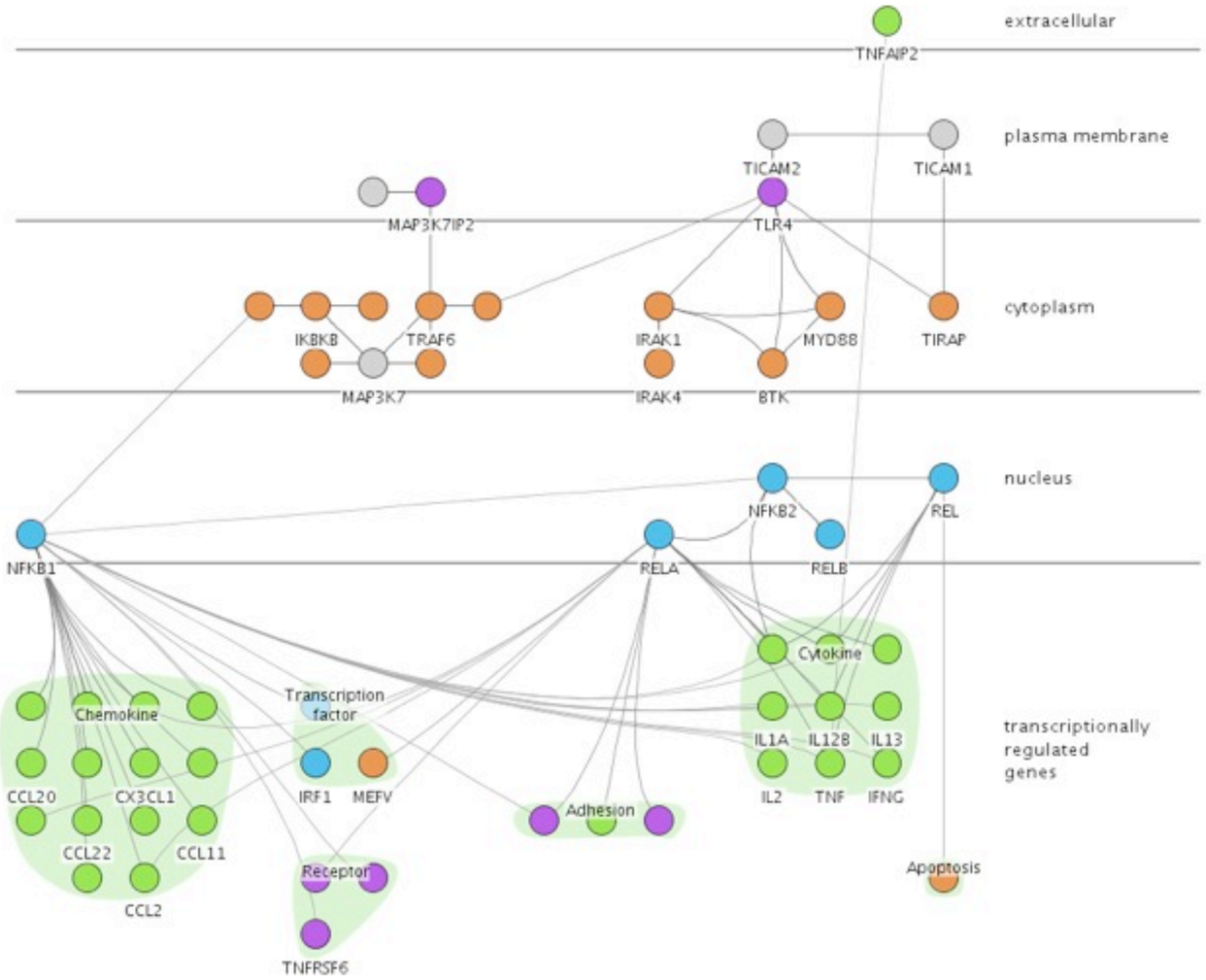


How: Arrange space



- automatic layout similar to hand-drawn diagrams
  - vertical compartment according to subcellular location attribute

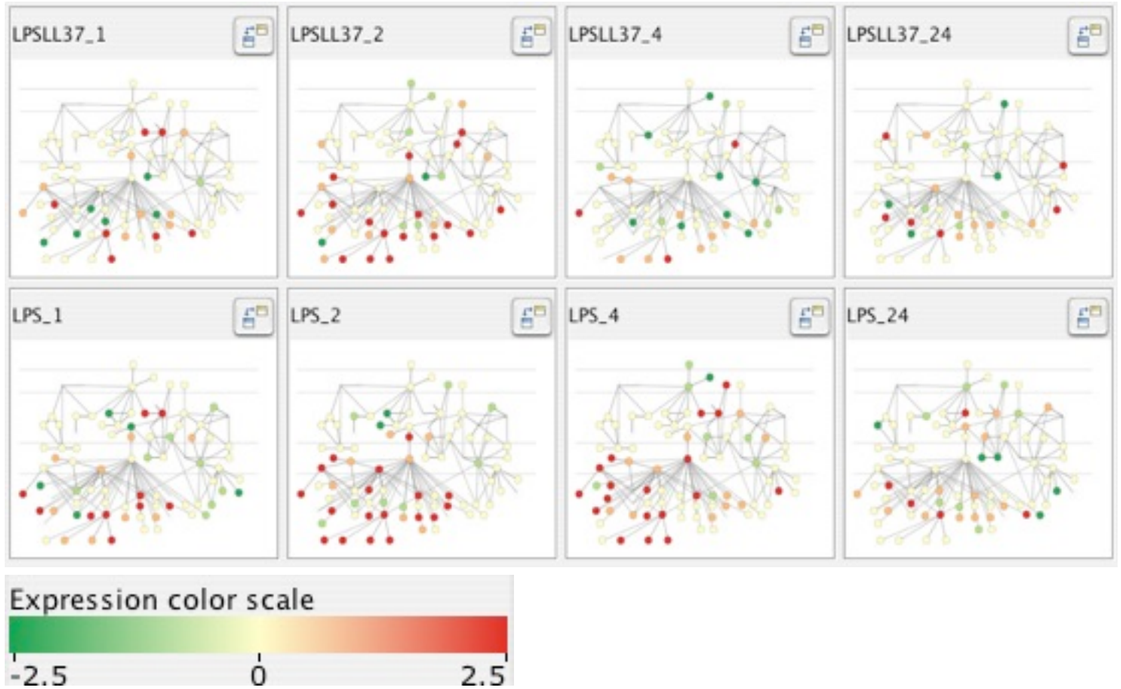




http://www.nature.com/nri/focus/tlr/nri1397.html

How: Idiom design decisions

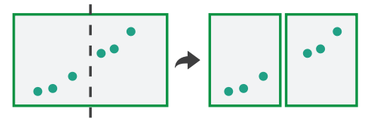
- facet: partition data into multiple views
  - juxtapose views side by side
    - same encoding, different data: *small multiples*
    - nodes in each view colored by expression levels for experimental condition

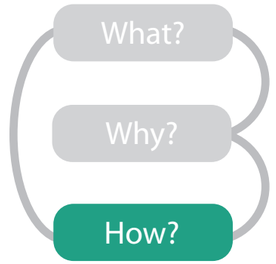


Facet

Juxtapose

Partition





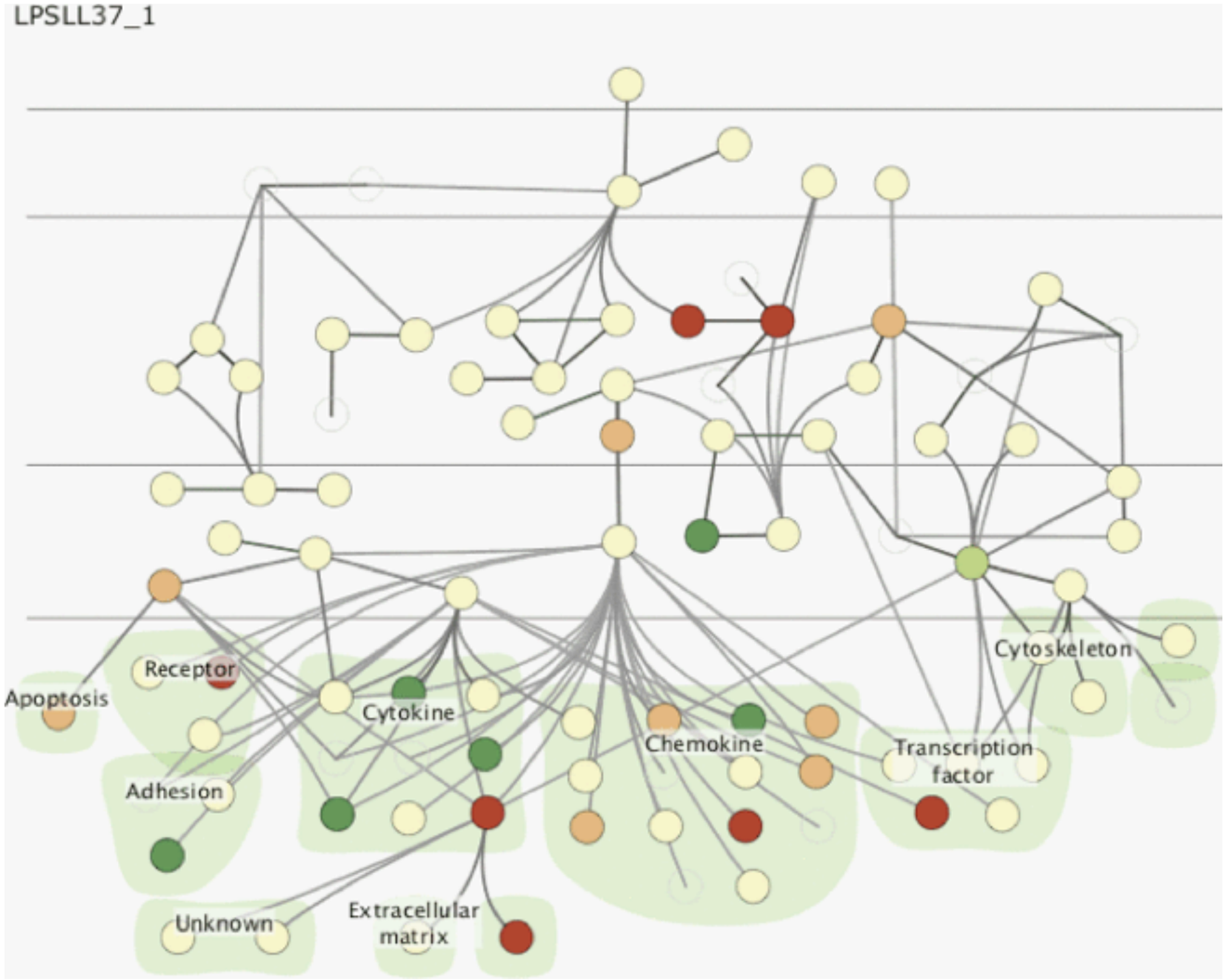
		Data		
		All	Subset	None
Encoding	Same	Redundant	*Overview/ Detail*	*Small Multiples*
	Different	*Multiform*	*Multiform, Overview/ Detail*	No Linkage

How: Juxtapose vs. animate

Manipulate

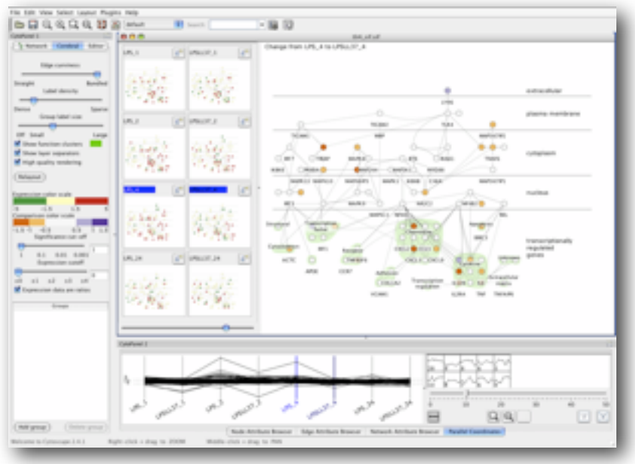
Change

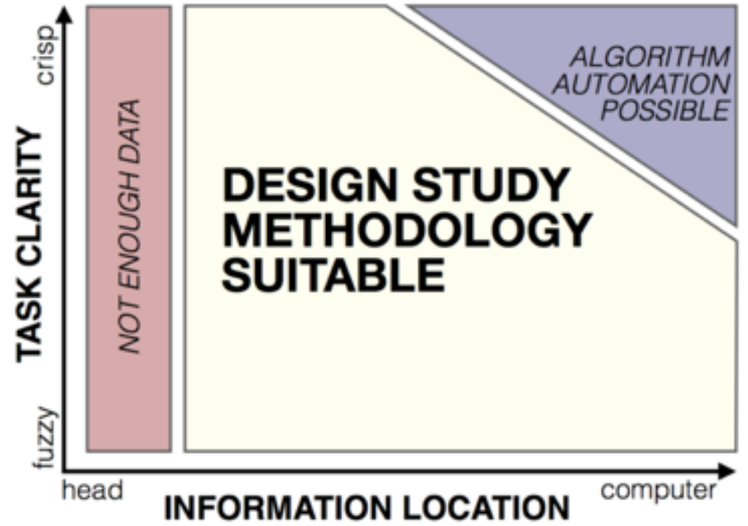
- comparison difficult across many frames with with many changes everywhere
- rule of thumb: eyes beat memory
  - principle: external cognition vs. internal memory
    - easy to compare by moving eyes between side-by-side views
    - harder to compare memory of what you saw to visible view



Outline

- interactive visual analysis
  - role and advantages
- LiveRAC
  - time-series data: managed web hosting (with AT&T)
- Cerebral
  - network of relationships: genes (with Agilent and UBC Immunology)
- wrapup





Design Study Methodology

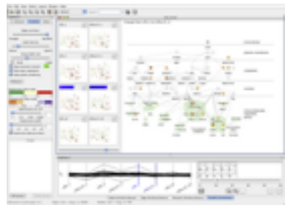
Reflections from the Trenches and from the Stacks

joint work with:
Michael Sedlmair, Miriah Meyer

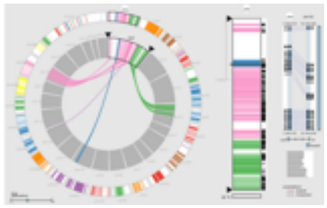
http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/

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

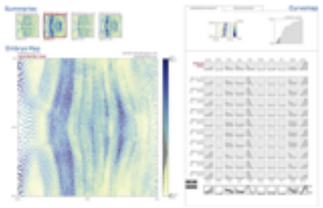


Cerebral
genomics

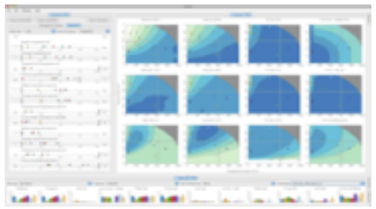


MizBee
genomics

Pathline
genomics



MulteeSum
genomics

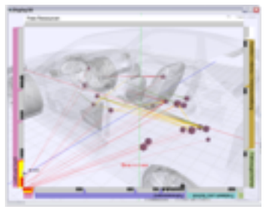


Vismon
fisheries management

QuestVis
sustainability

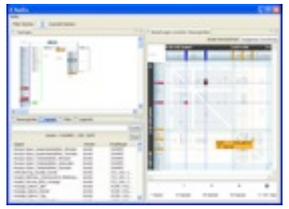
WiKeVis
in-car networks

MostVis
in-car networks

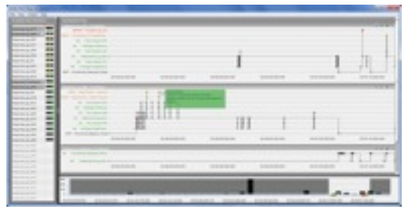


Car-X-Ray
in-car networks

ProgSpy2010
in-car networks

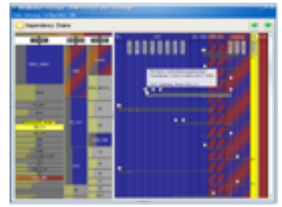


RelEx
in-car networks



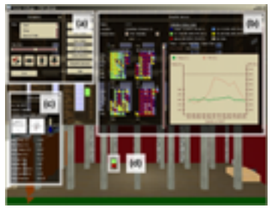
Cardiogram
in-car networks

AutobahnVis
in-car networks

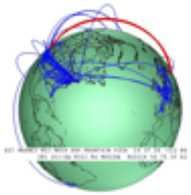


VisTra
in-car networks

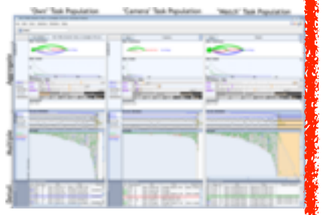
Constellation
linguistics



LibVis
cultural heritage

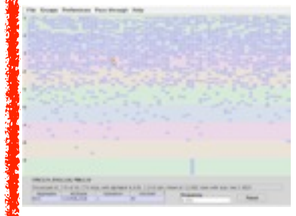


Caidants
multicast

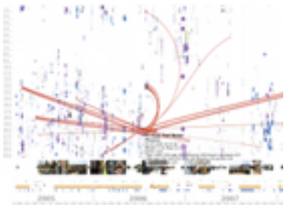


SessionViewer
web log analysis

LiveRAC
server hosting



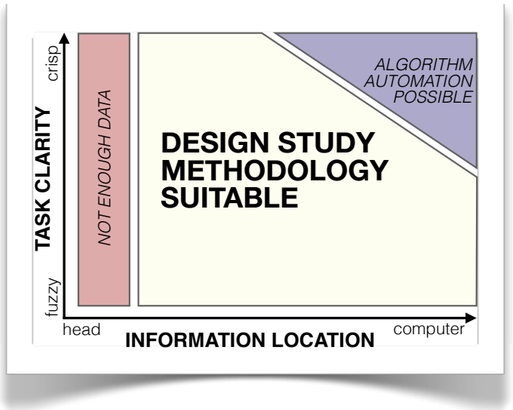
PowerSetViewer
data mining

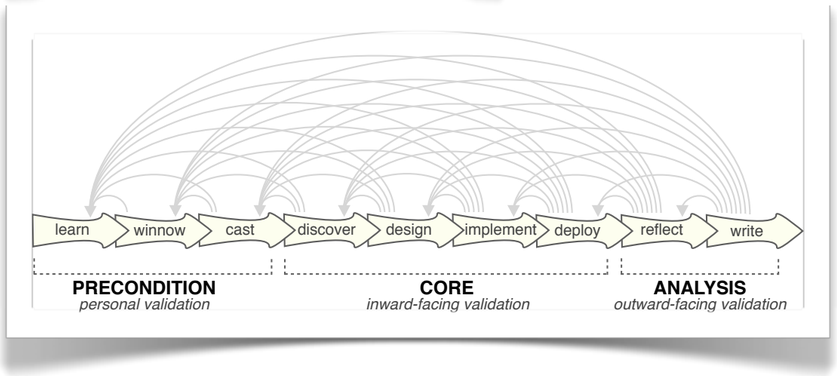


LastHistory
music listening

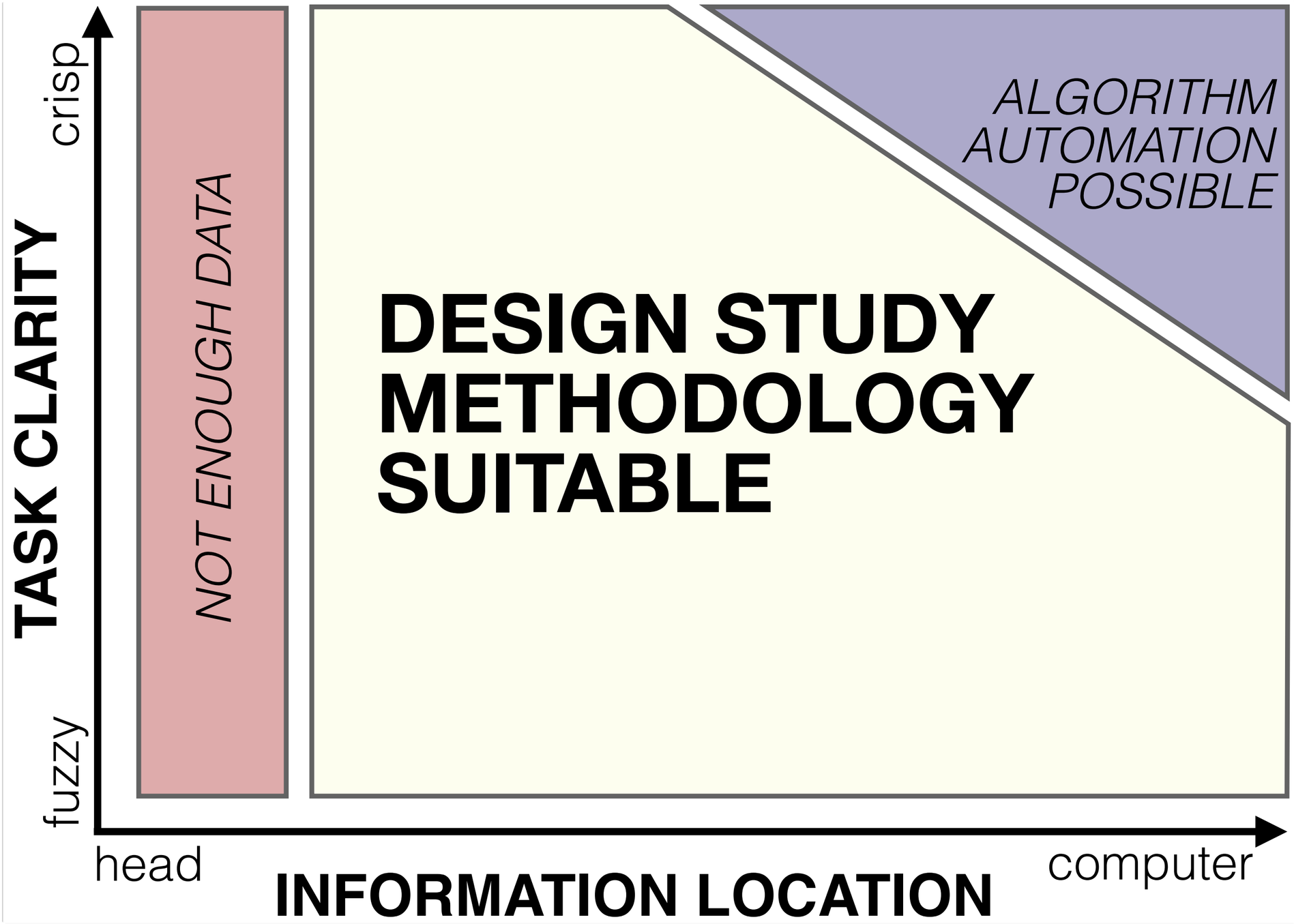
Methodology for Problem-Driven Work

- definitions
- 9-stage framework
- 32 pitfalls
  - and how to avoid them



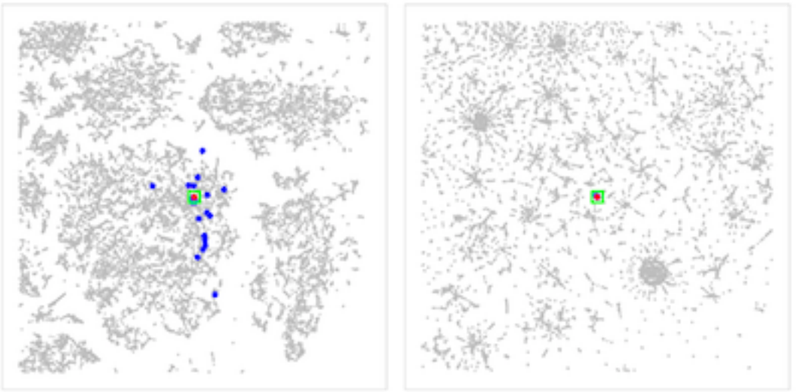


PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
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


TASK CLARITY
crisp
fuzzy
NOT ENOUGH DATA
ALGORITHM
AUTOMATION
POSSIBLE
DESIGN STUDY
METHODOLOGY
SUITABLE
head
computer
INFORMATION LOCATION

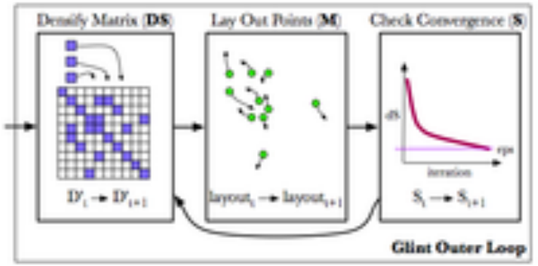
Techniques: Dimensionality Reduction

- reducing high-dimensional data to tractable low-dimensional form



  - Q-SNE: high-quality clusters for millions of documents

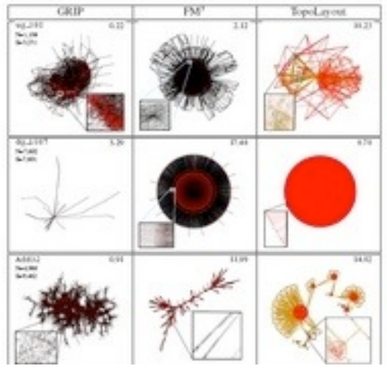
  - Glint: costly distance functions
    - incl. preferences elicited from people

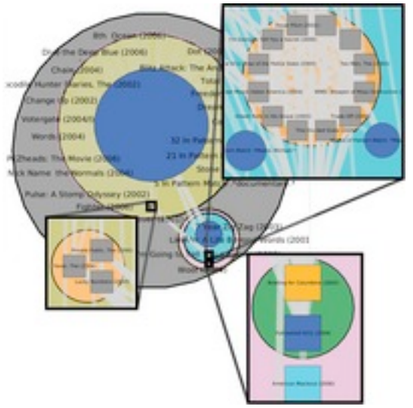


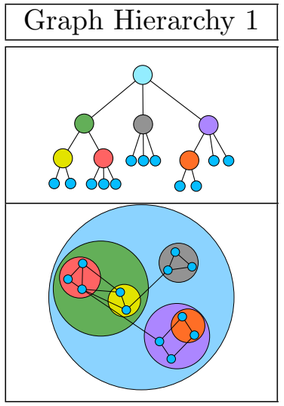
Techniques: Networks & Trees

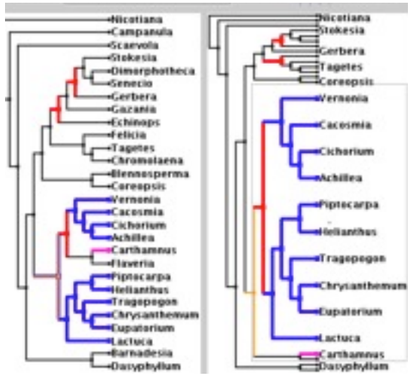
- large multi-level networks
  - layout
    - TopoLayout
  - interaction
    - Grouse
    - GrouseFlocks
    - TugGraph

- large tree comparison
  - TreeJuxtaposer



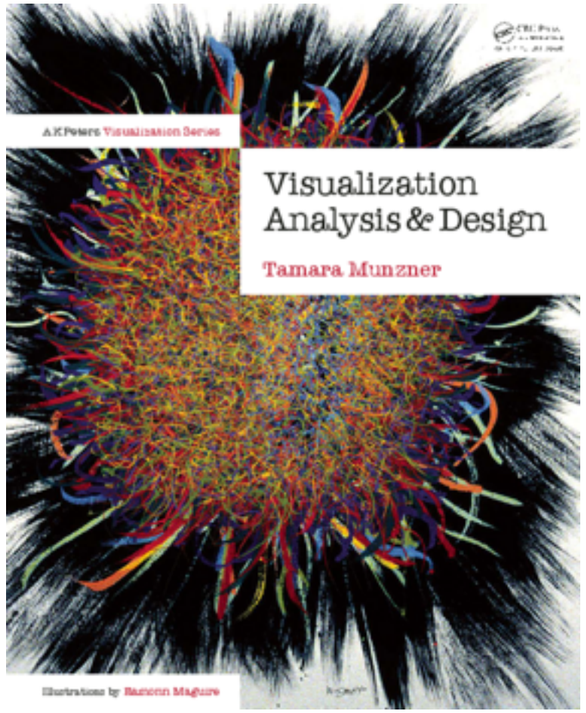






More Information

- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#disney15
- papers, videos, software, talks, courses
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm
- book
  http://www.cs.ubc.ca/~tmm/vadbook



- acknowledgements
  - funding: Agilent, AT&T, NSERC, NSF

@tamaramunzner