Visualization Analysis & Design for Biology

Tamara Munzner

Department of Computer Science
University of British Columbia

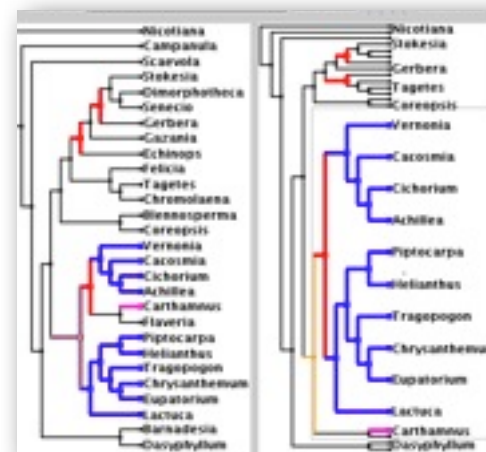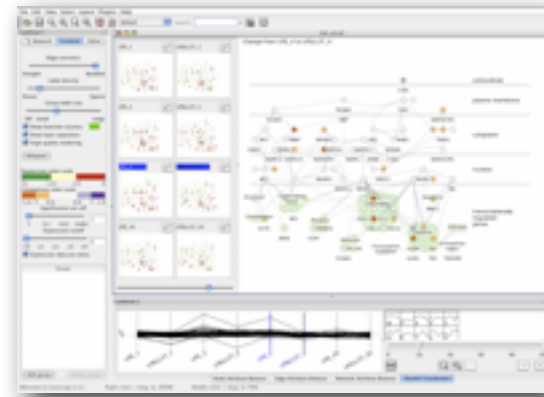*VanBUG: Vancouver Bioinformatics Users Group*
8 October 2015, Vancouver BC

http://www.cs.ubc.ca/~tmm/talks.html#vanbug15   **@tamaramunzner**

Outline

- introduction

- Cerebral

- MizBee

- TreeJuxtaposer

- wrapup

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.

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.

- many analysis problems ill-specified, not clear what questions to ask in advance
  - don't need vis when fully automatic solution exists and is trusted
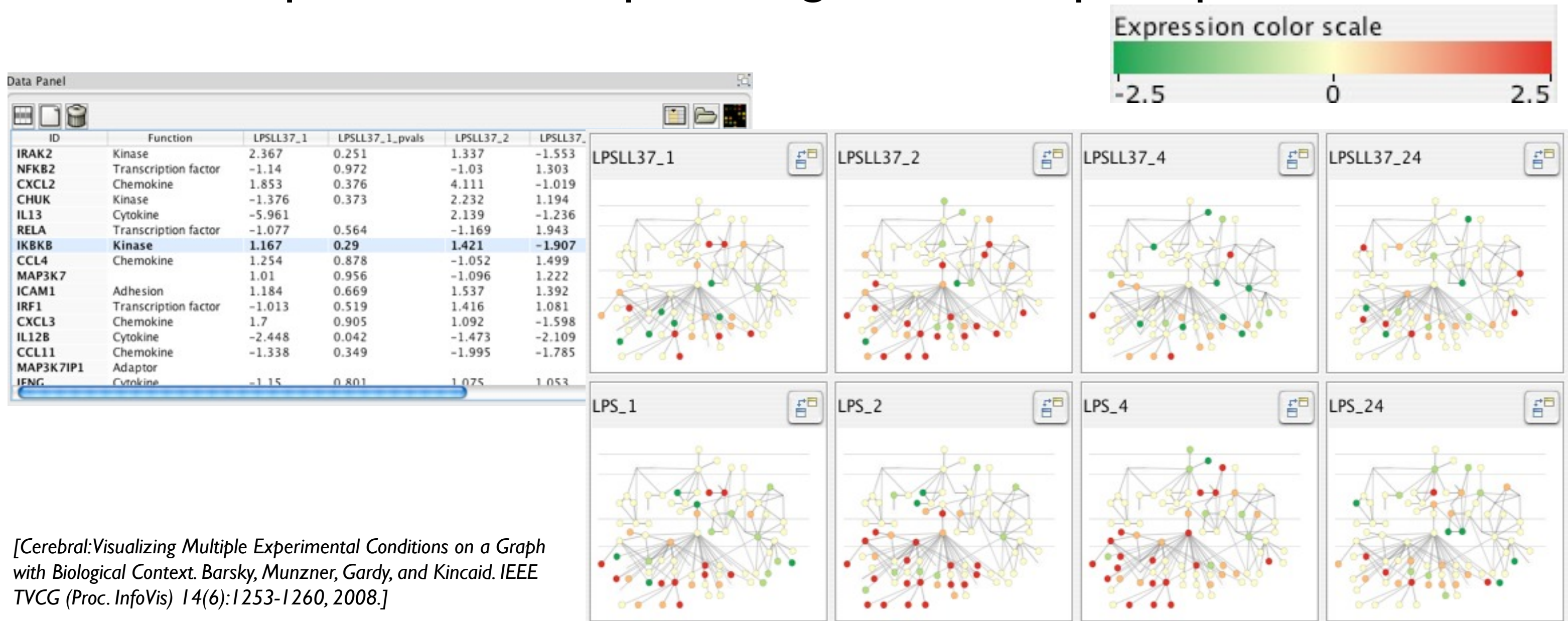
Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	8
y variance	4
x/y correlation	1
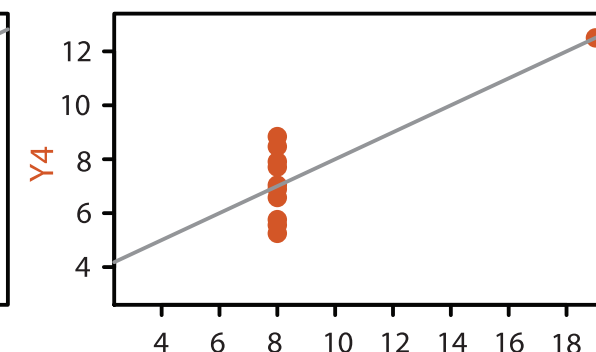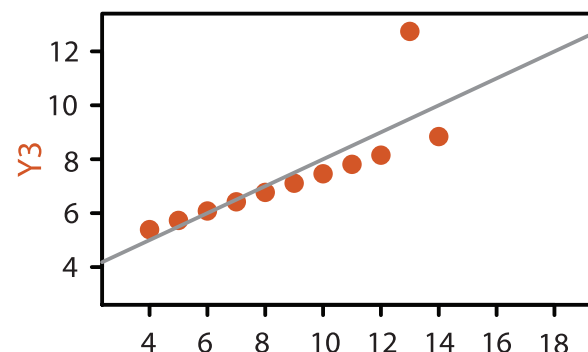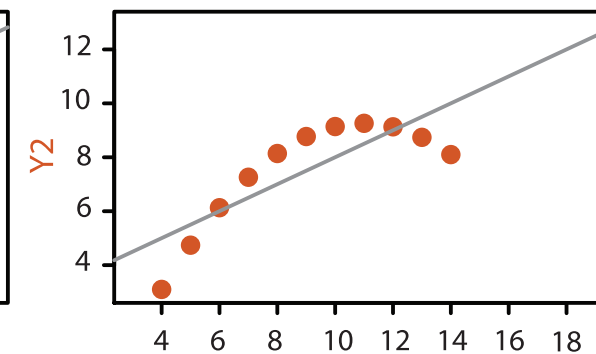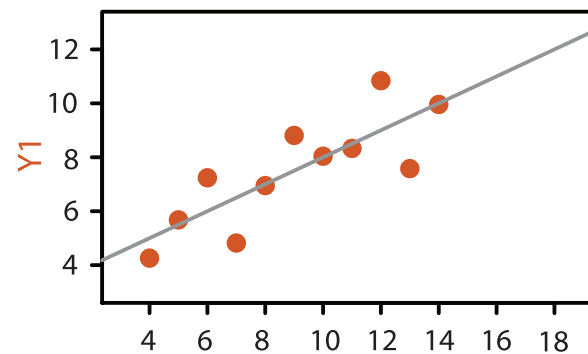
Why use an external representation?

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.]

Analysis framework: Four levels, three questions

- *domain* situation
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data abstraction**
      - often don't just draw what you're given: transform to new form
    - **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
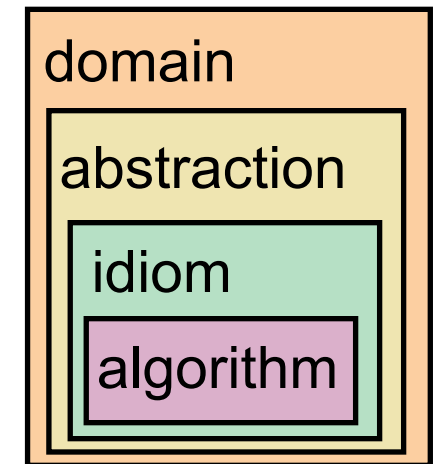
[A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]
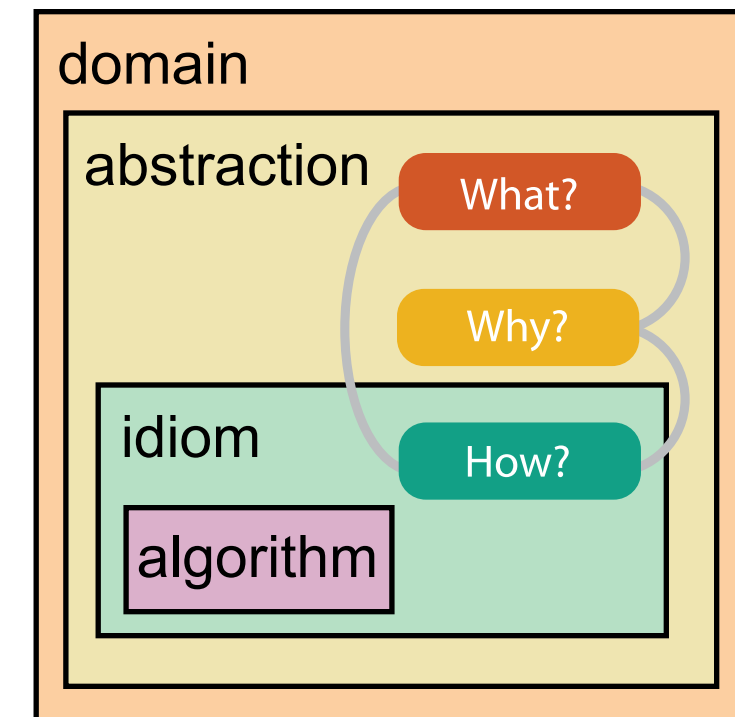
[A Multi-Level Typology of Abstract Visualization Tasks
Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

Validation methods from different fields for each level

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
work

technique-driven
work

- mismatch: cannot show idiom good with system timings
- mismatch: cannot show abstraction good with lab study

technique-driven work ↔ problem-driven work

theoretical foundations

evaluation
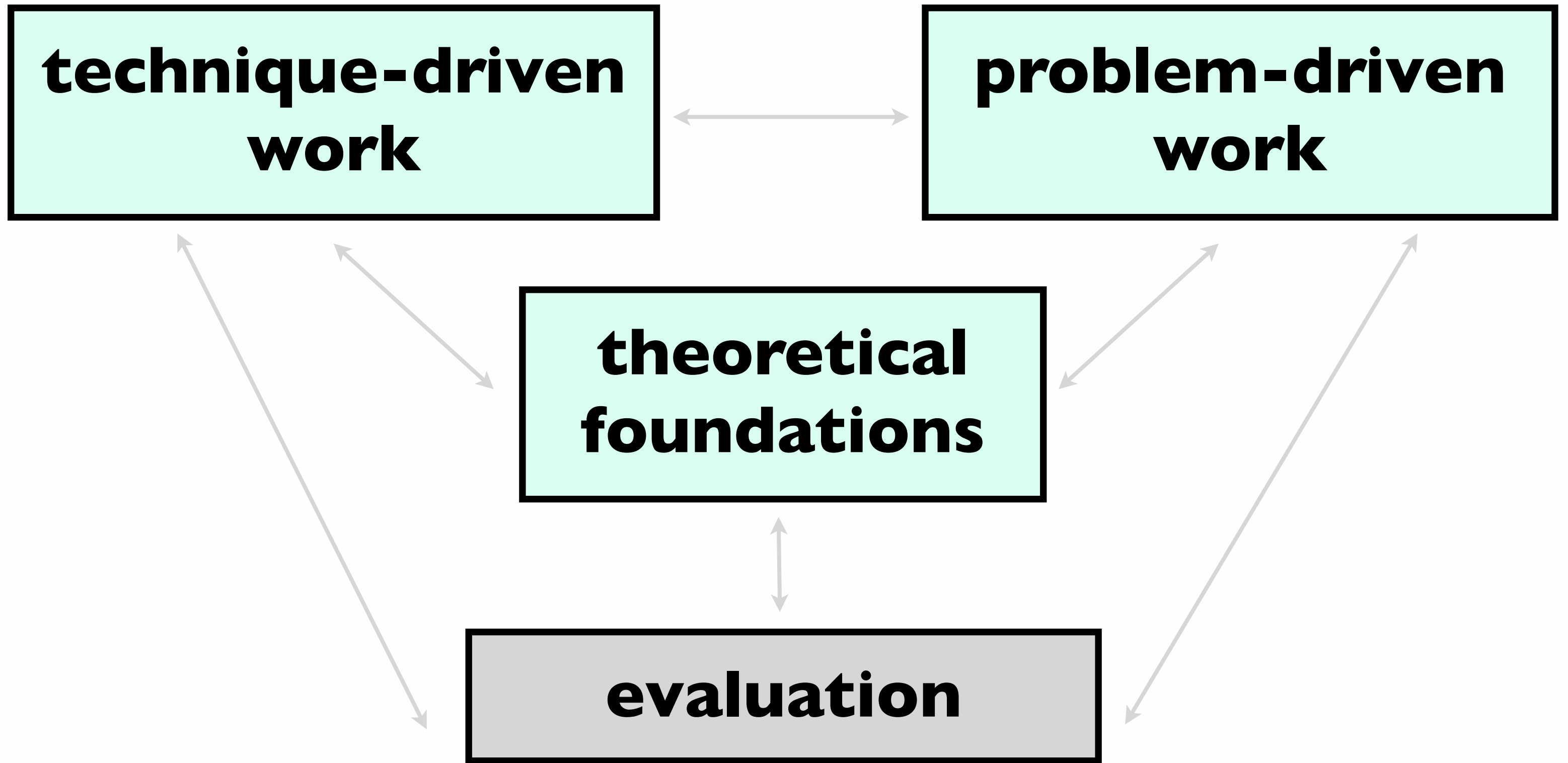
Technique-driven work: Networks

- scaling up networks
  - multilevel networks, 10K-100K nodes
    - topologically aware decomposition, layout, browsing
  - trees, millions of nodes
    - guaranteed visibility of semantically meaningful marks
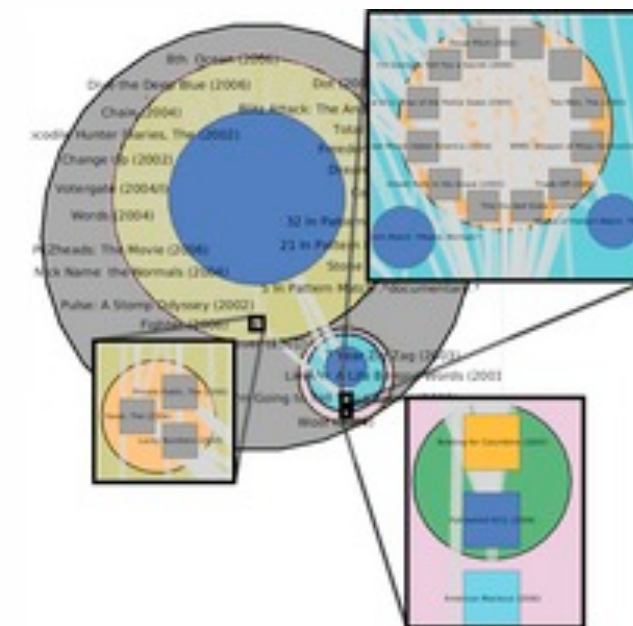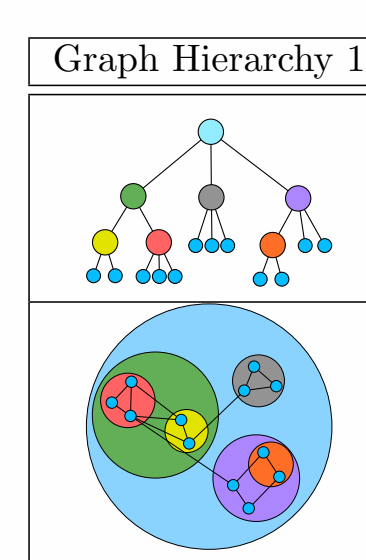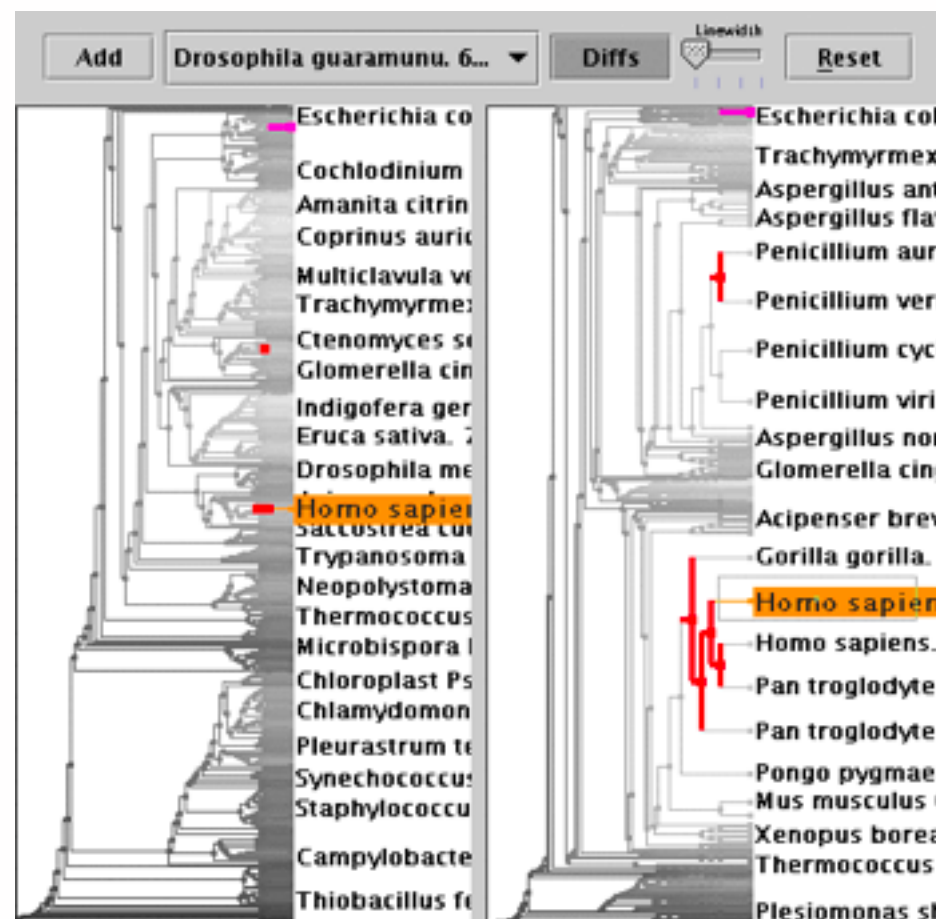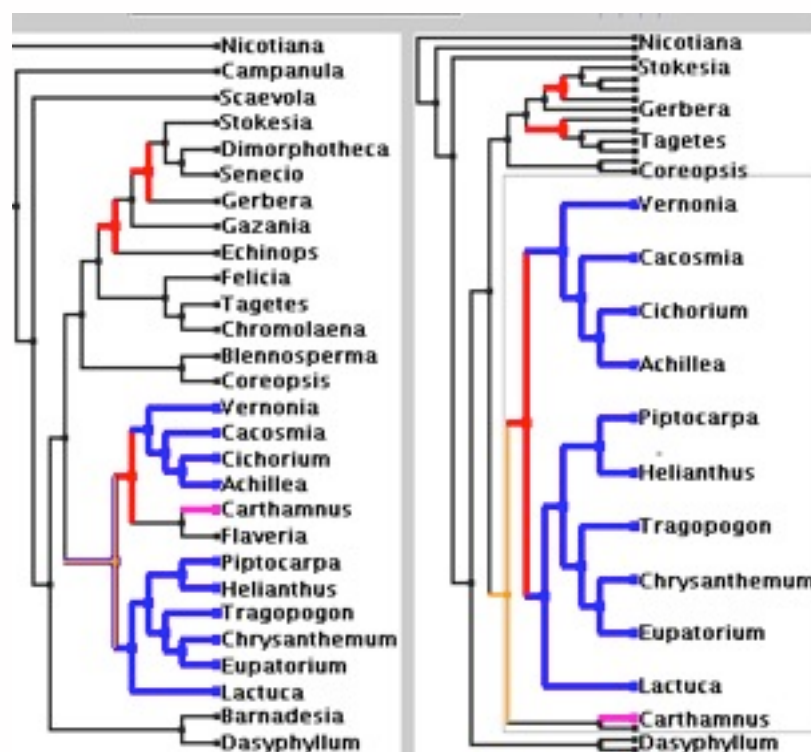
TopoLayout
Smashing Peacocks Further
**Grouse**
GrouseFlocks
TugGraph

http://youtu.be/t1Xbt6XOWp8

http://youtu.be/AWXAe8zvkt8

TreeJuxtaposer
PRISAD

http://youtu.be/fq8EIAOutvs

http://youtu.be/GdaPj8a9QEo

Technique-driven work: Dimensionality reduction

- close overlap with machine learning
  - Glimmer: MDS on the GPU
  - Glint: DR for costly distances
  - QSNE: sparse documents
    - high quality for millions of items

Glimmer
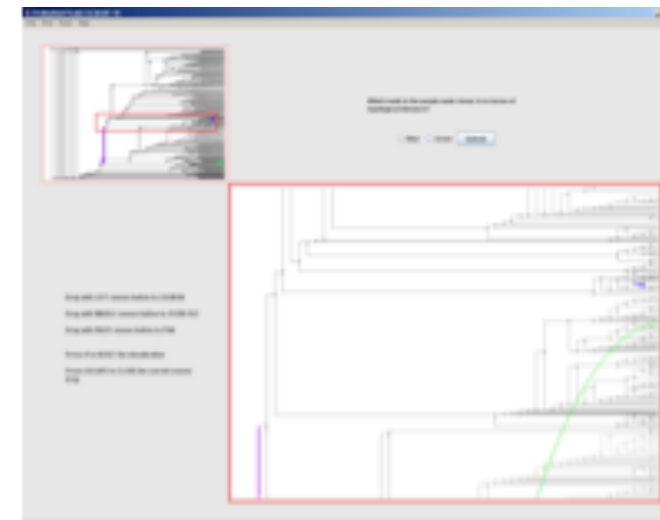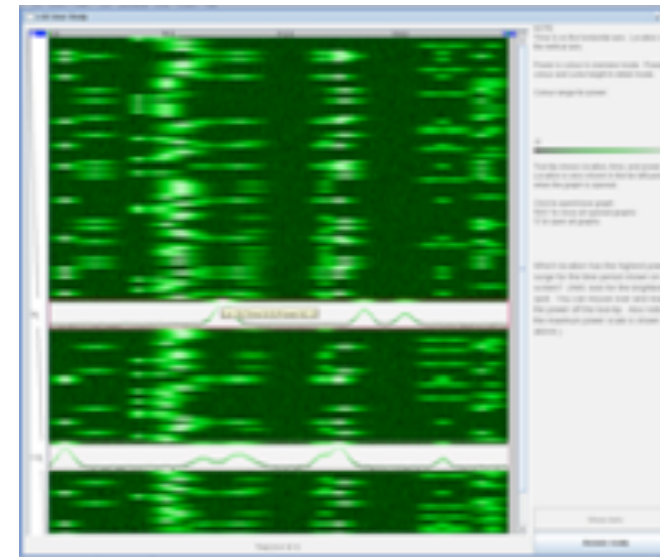
http://youtu.be/PLaBAPM6qLI
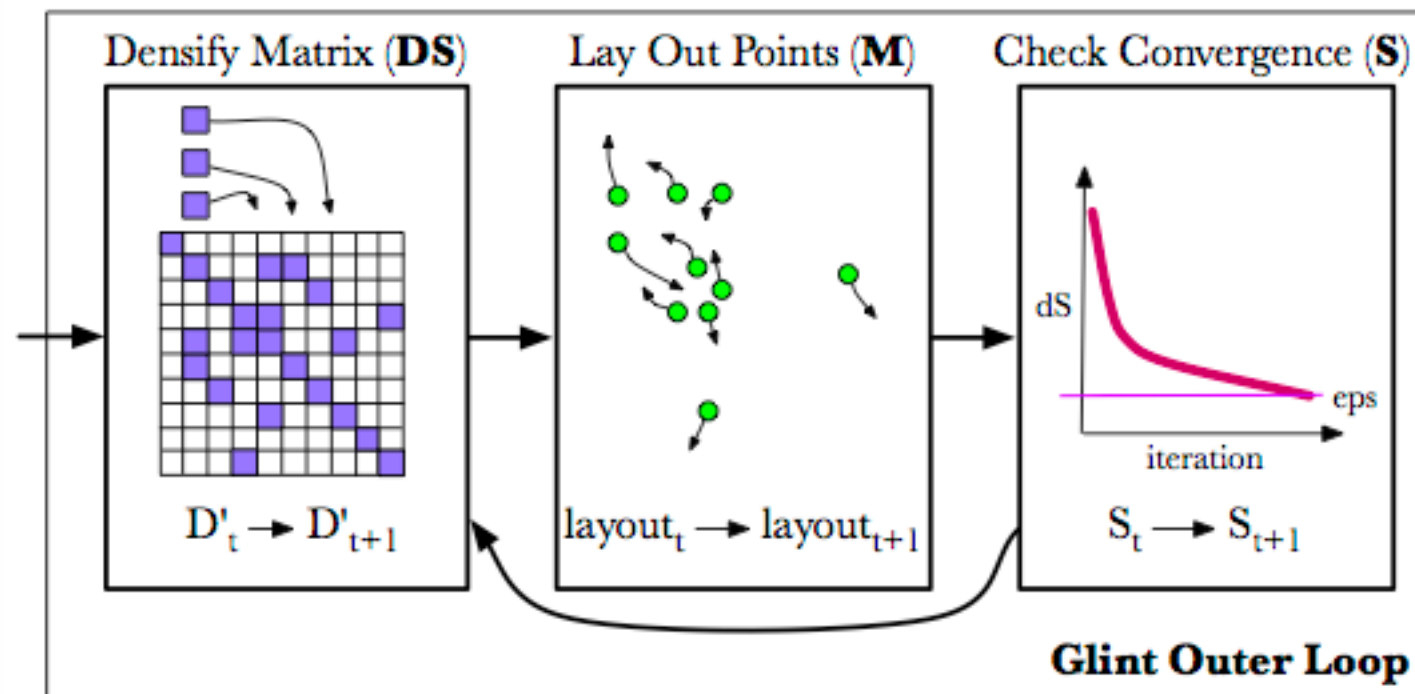
Glint
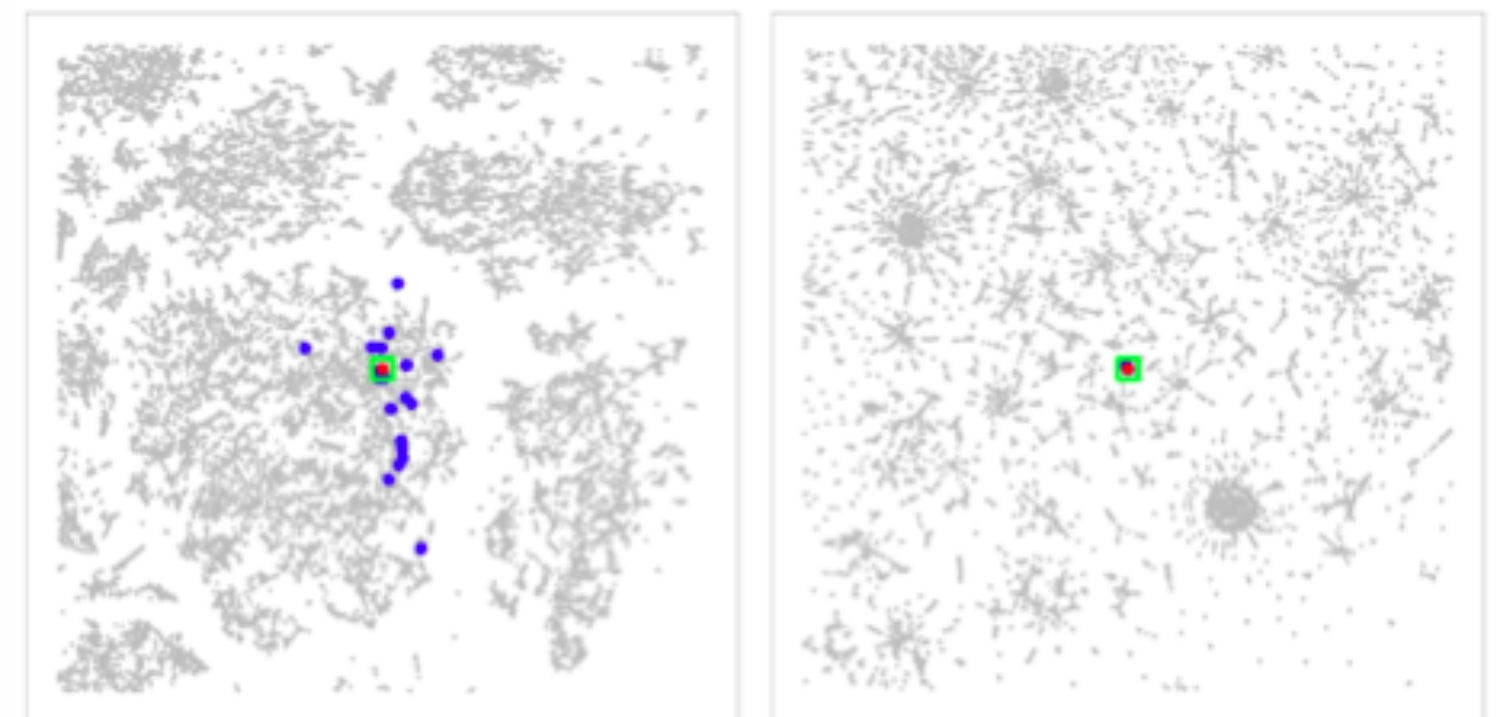
QSNE

Evaluation: Dimensionality Reduction

- guidance on scatterplot/DR choices

- taxonomy of cluster separation factors

- 2D points vs 3D landscapes
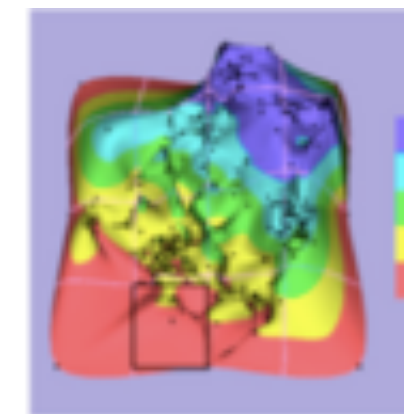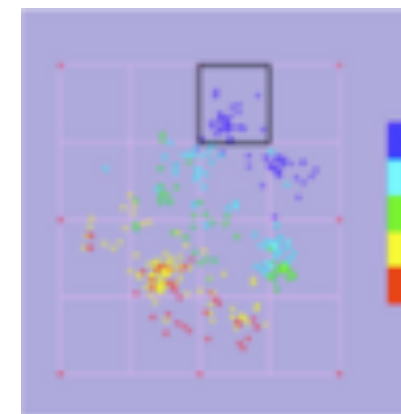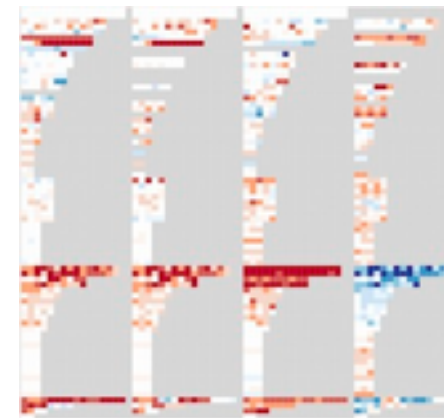
Evaluation: Focus+Context

- overviews: separate vs. integrated views

- navigation: stretch and squish vs. pan/ zoom navigation

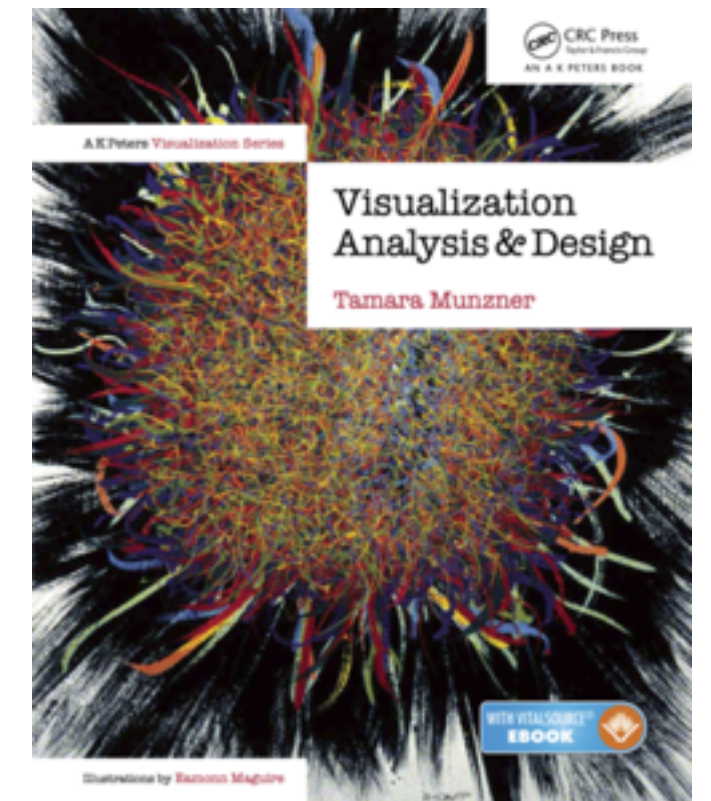- impact of distortion on visual search, visual memory
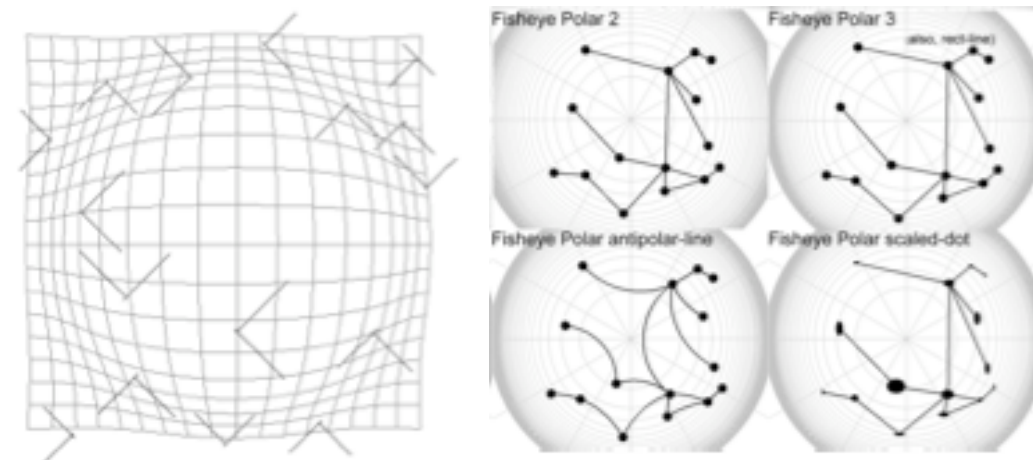
Theory/Models

- multi-level typology of abstract visualization tasks

- design study methodology

- nested model for vis design and validation

- papers process and pitfalls

- book: Visualization Analysis and Design
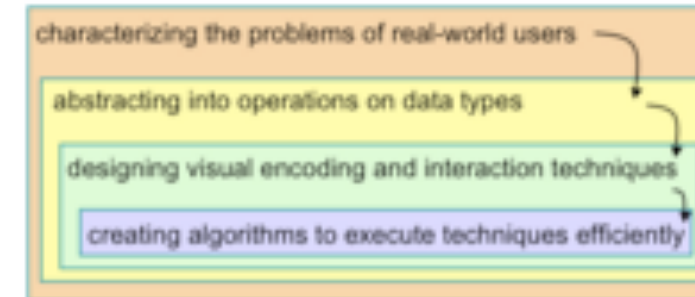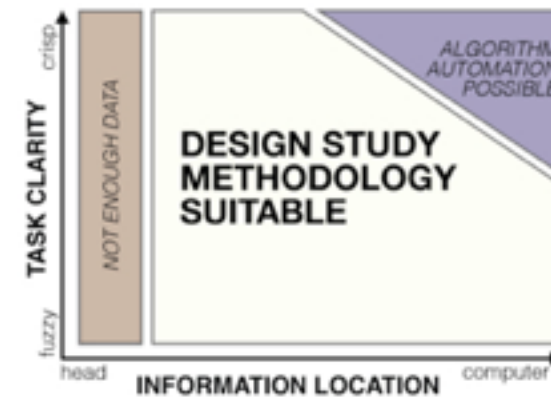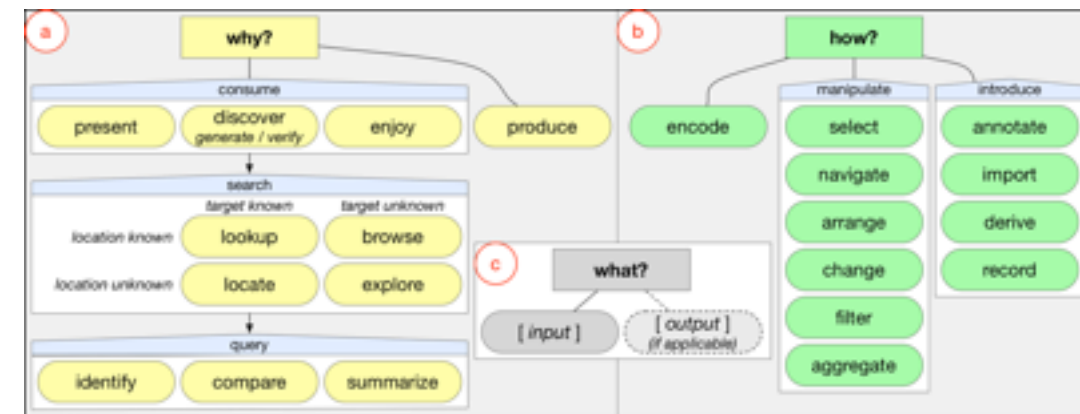
Problem-driven work: Many domains

http://youtu.be/ld0c3H0VSkw

LiveRAC: system management time-series
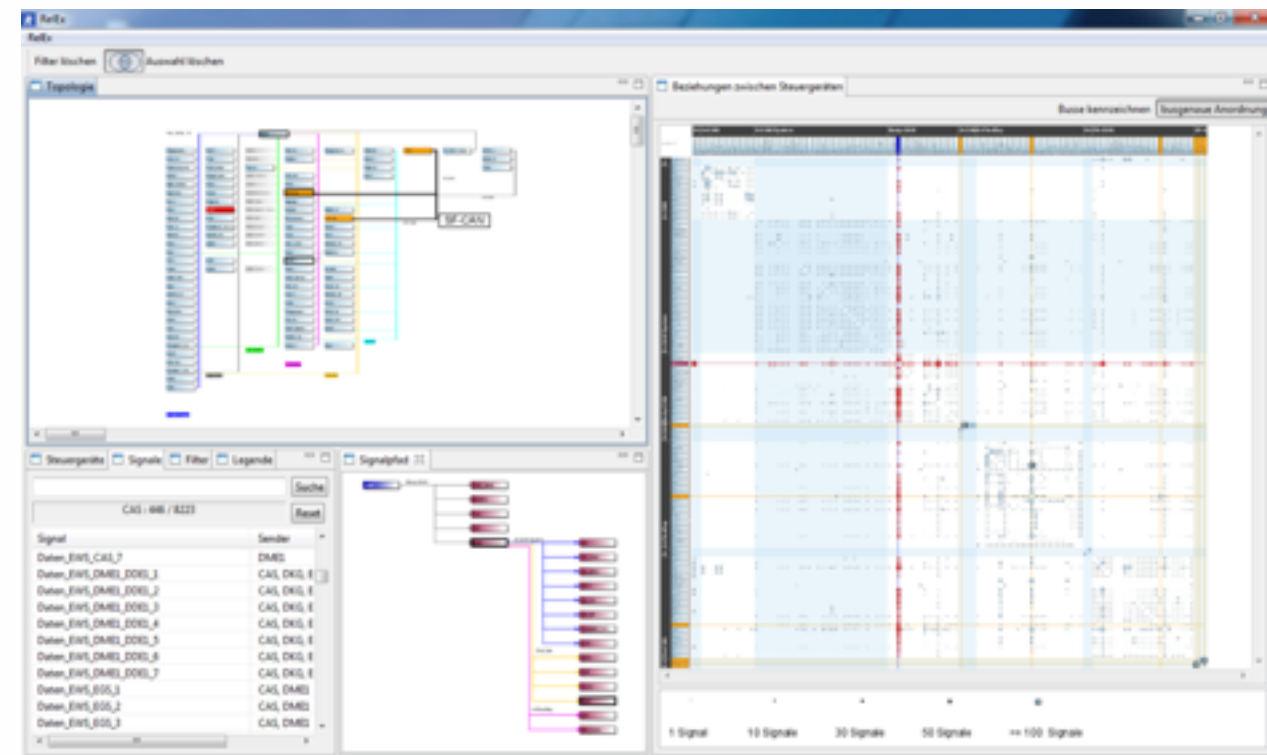
RelEx: in-car overlay networks http://youtu.be/89lsQXc6Ao4

Vismon: fisheries management http://youtu.be/h0kHoS4VYmk
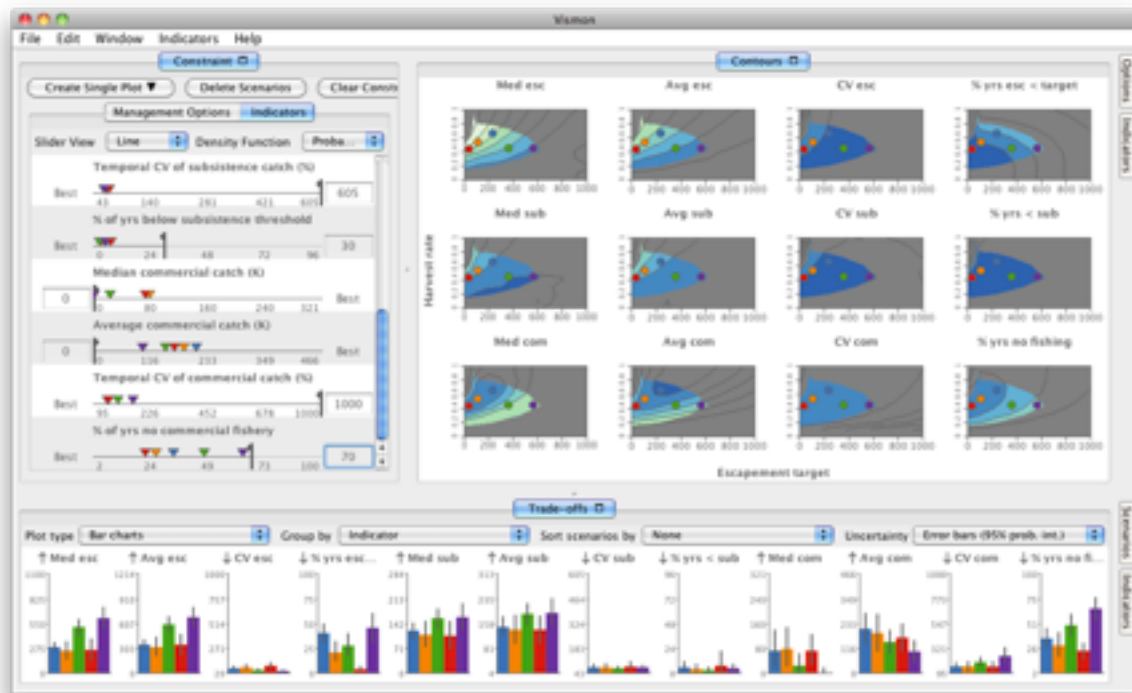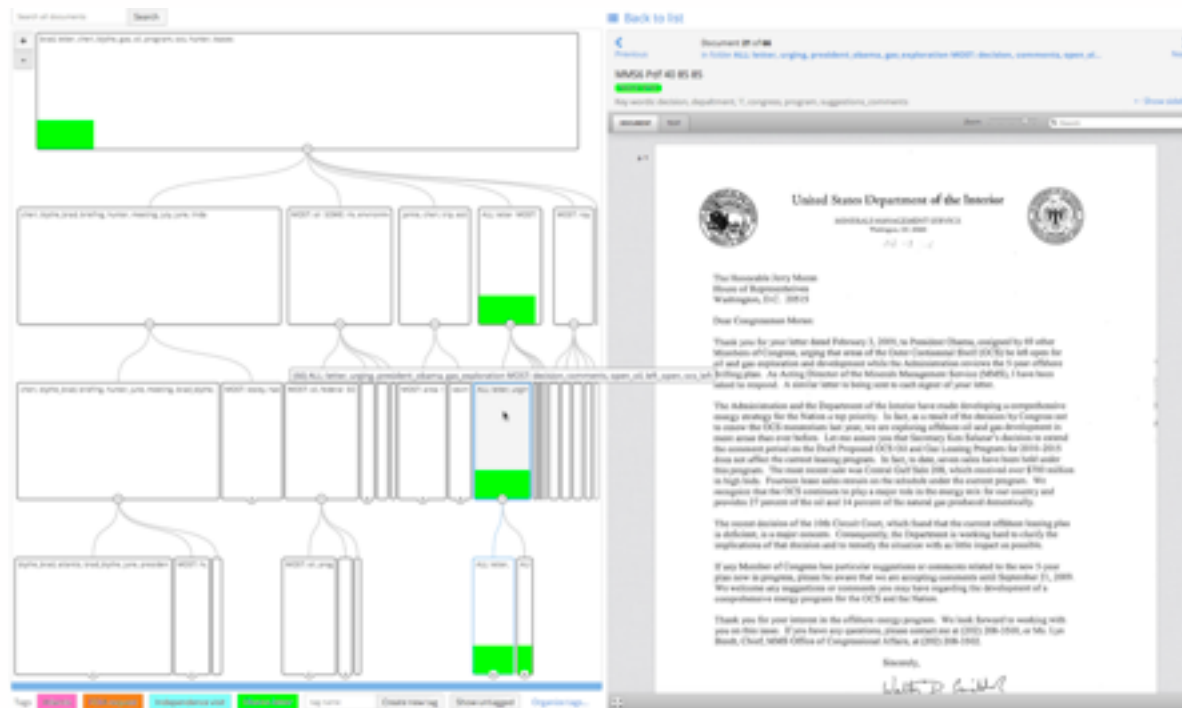
Overview: investigative journalism http://vimeo.com/71483614

14

Problem-driven work: Genomics

MizBee

http://youtu.be/86p7brwuz2g
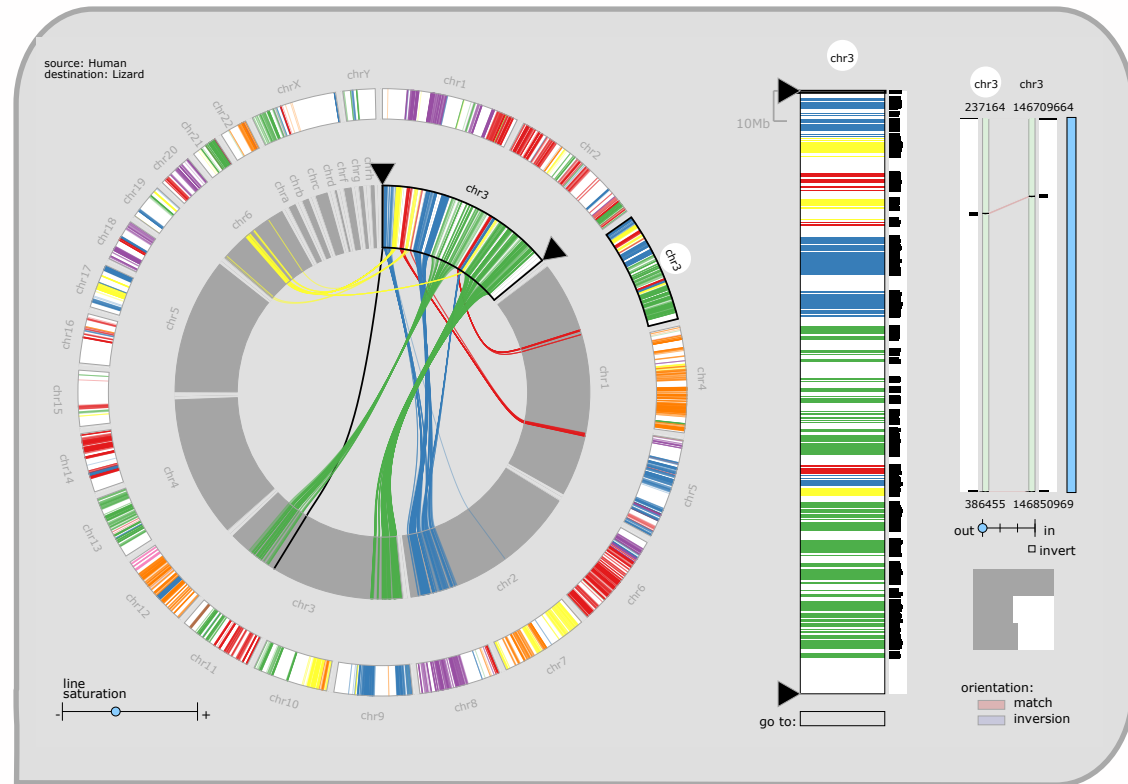
**MulteeSum
Pathline**
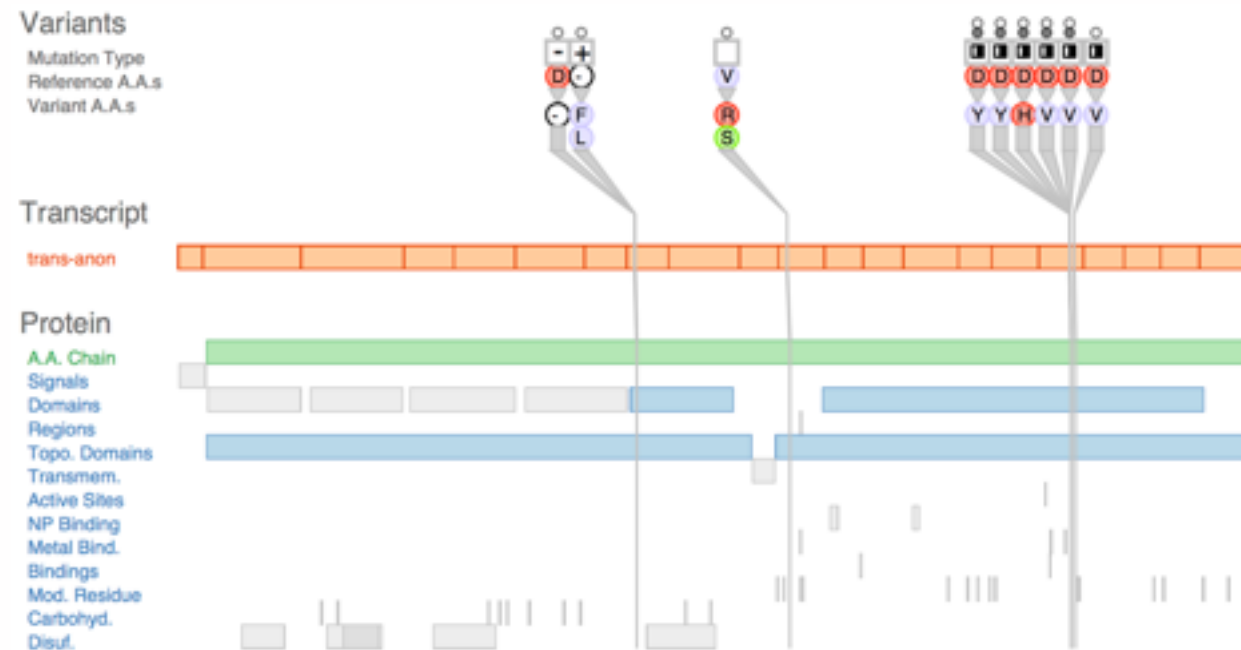
Variant View

http://youtu.be/AHDnv_qMXxQ
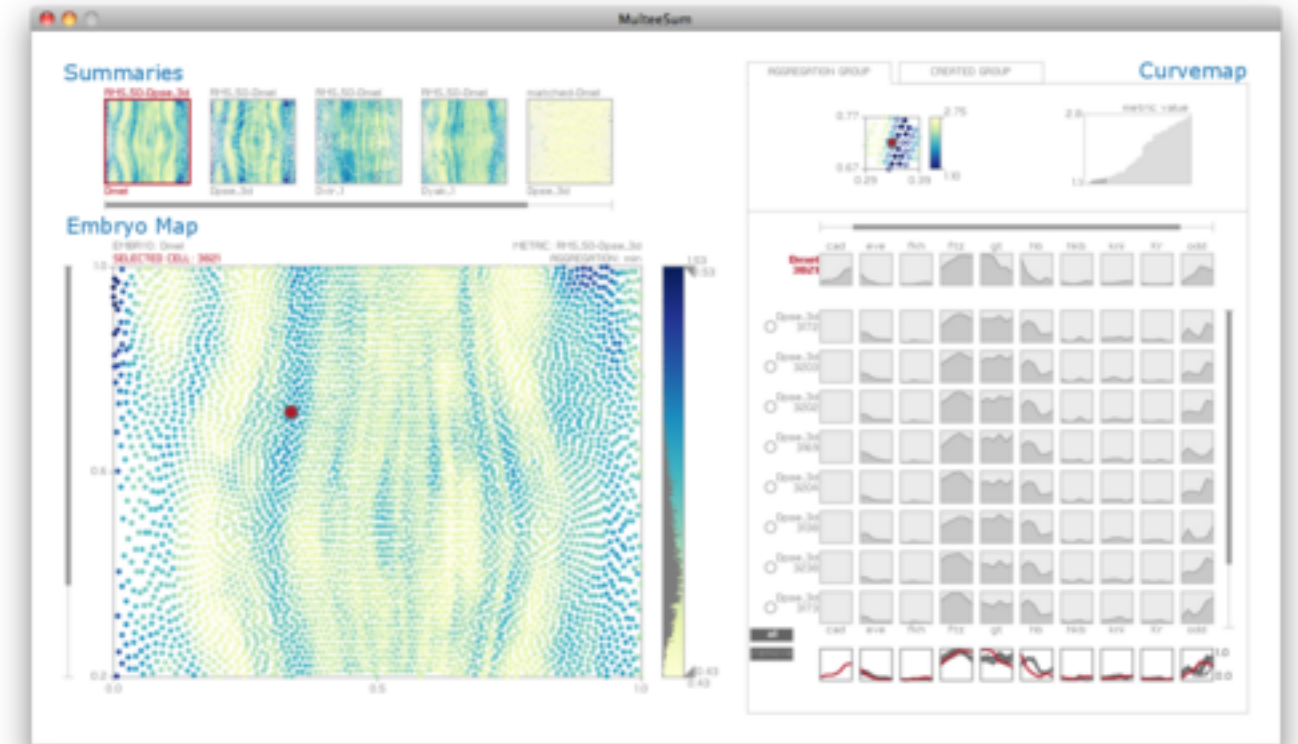
Cerebral

http://youtu.be/76HhG1FQngl
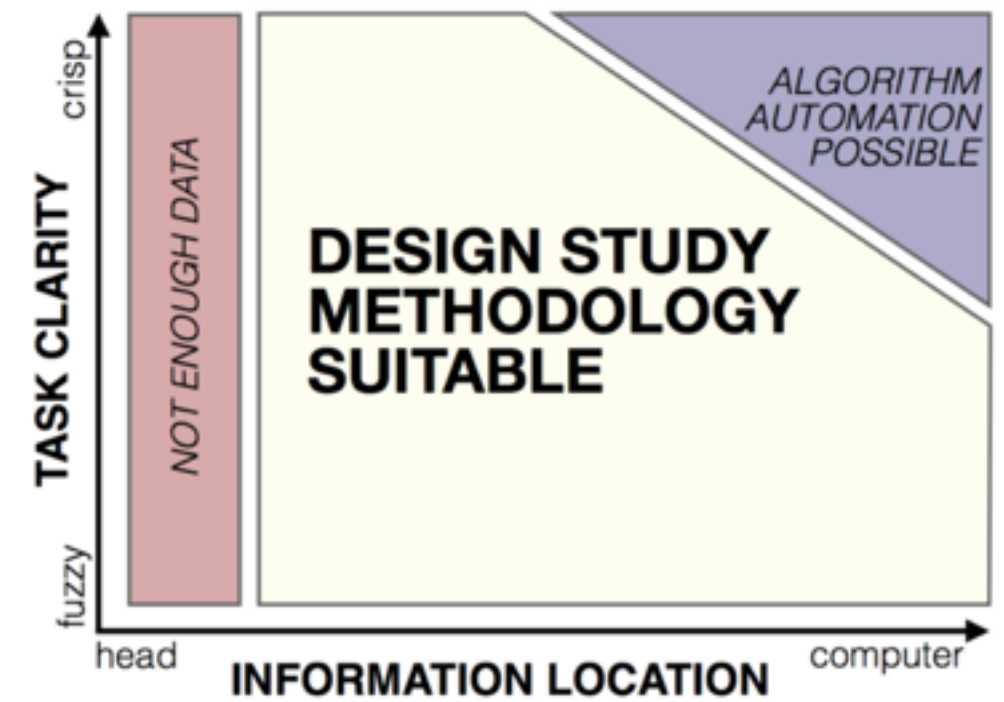
15
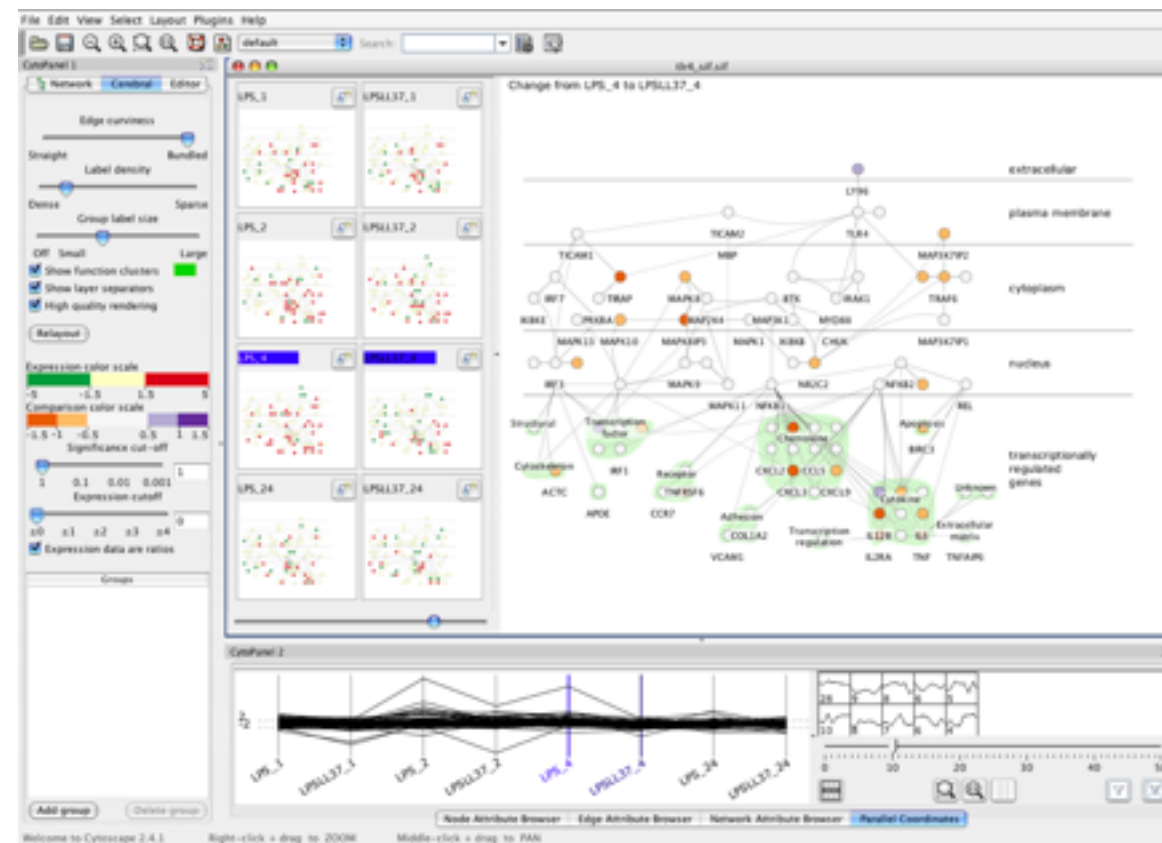
Design Study Methodology

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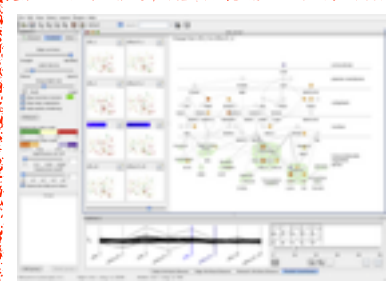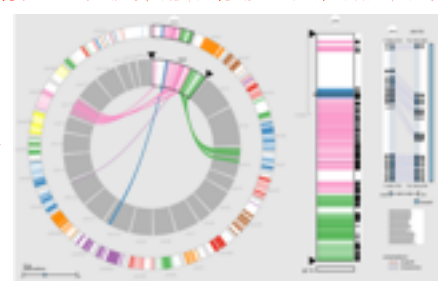
Design Study Methodology: Reflections from the Trenches and from the Stacks.
Sedlmair, Meyer, Munzner. IEEE Trans. Visualization and Computer Graphics 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

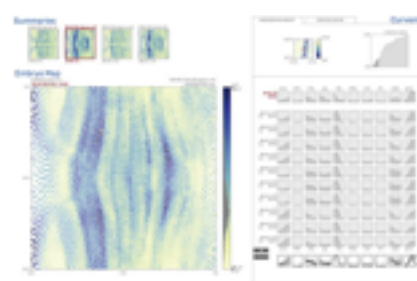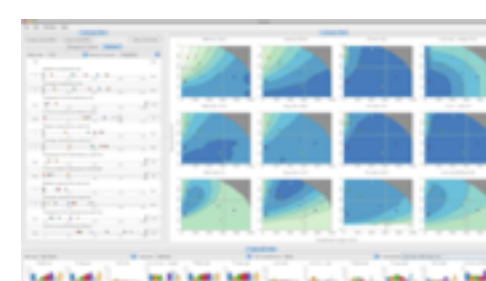# Design Studies: Lessons learned after 21 of them

Cerebral
genomics

MizBee
genomics

Pathline
genomics

MulteeSum
genomics

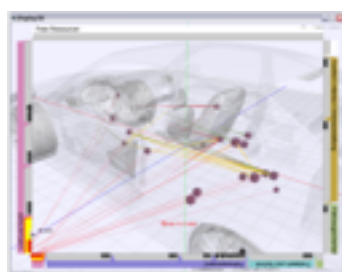*Vismon*
fisheries management

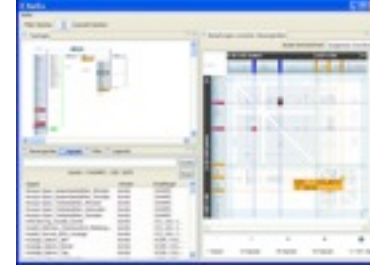*QuestVis*
sustainability

WiKeVis
in-car networks

MostVis
in-car networks

Car-X-Ray
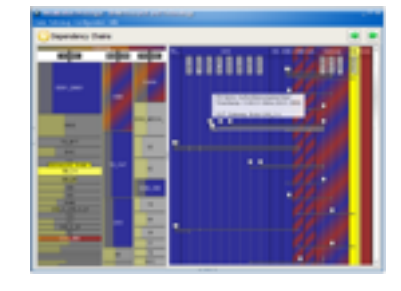in-car networks

ProgSpy2010
in-car networks

RelEx
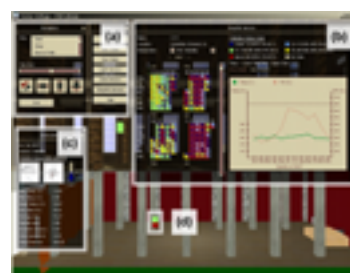in-car networks

Cardiogram
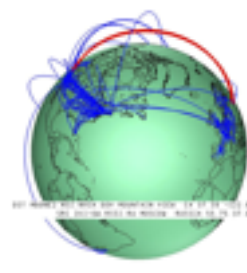in-car networks

AutobahnVis
in-car networks

VisTra
in-car networks

Constellation
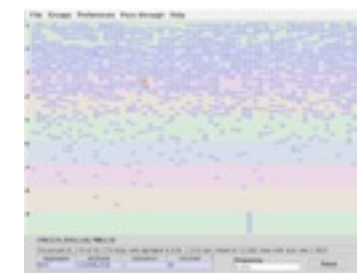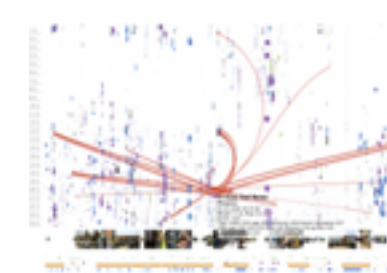linguistics

LibVis
cultural heritage

Caidants
multicast

SessionViewer
web log analysis

LiveRAC
server hosting

PowerSetViewer
data mining

LastHistory
music listening

TASK CLARITY

crisp

fuzzy

ALGORITHM AUTOMATION POSSIBLE

NOT ENOUGH DATA

DESIGN STUDY METHODOLOGY SUITABLE

head

computer

INFORMATION LOCATION

18
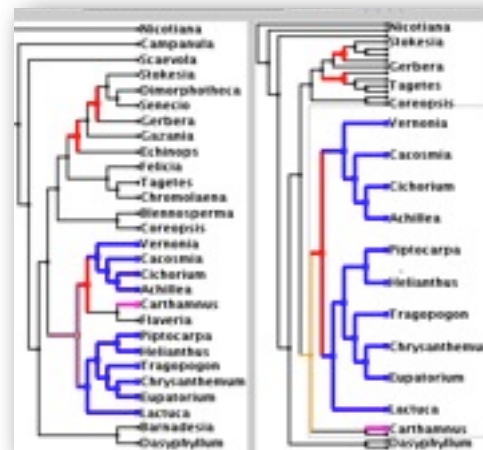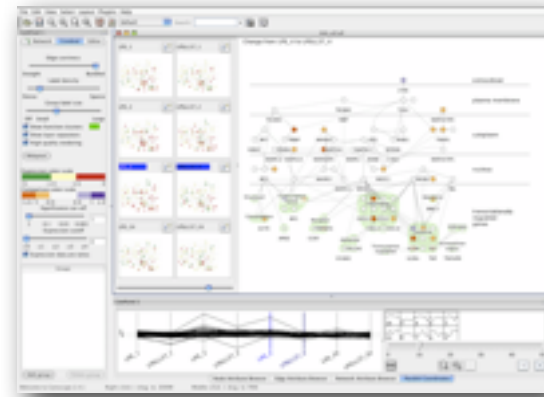
Problem-driven work and you

- enormous opportunity for bioinformatics
  - apply human-centered design methods beyond visualization!
  - task analysis: what do the target users really need?

Outline

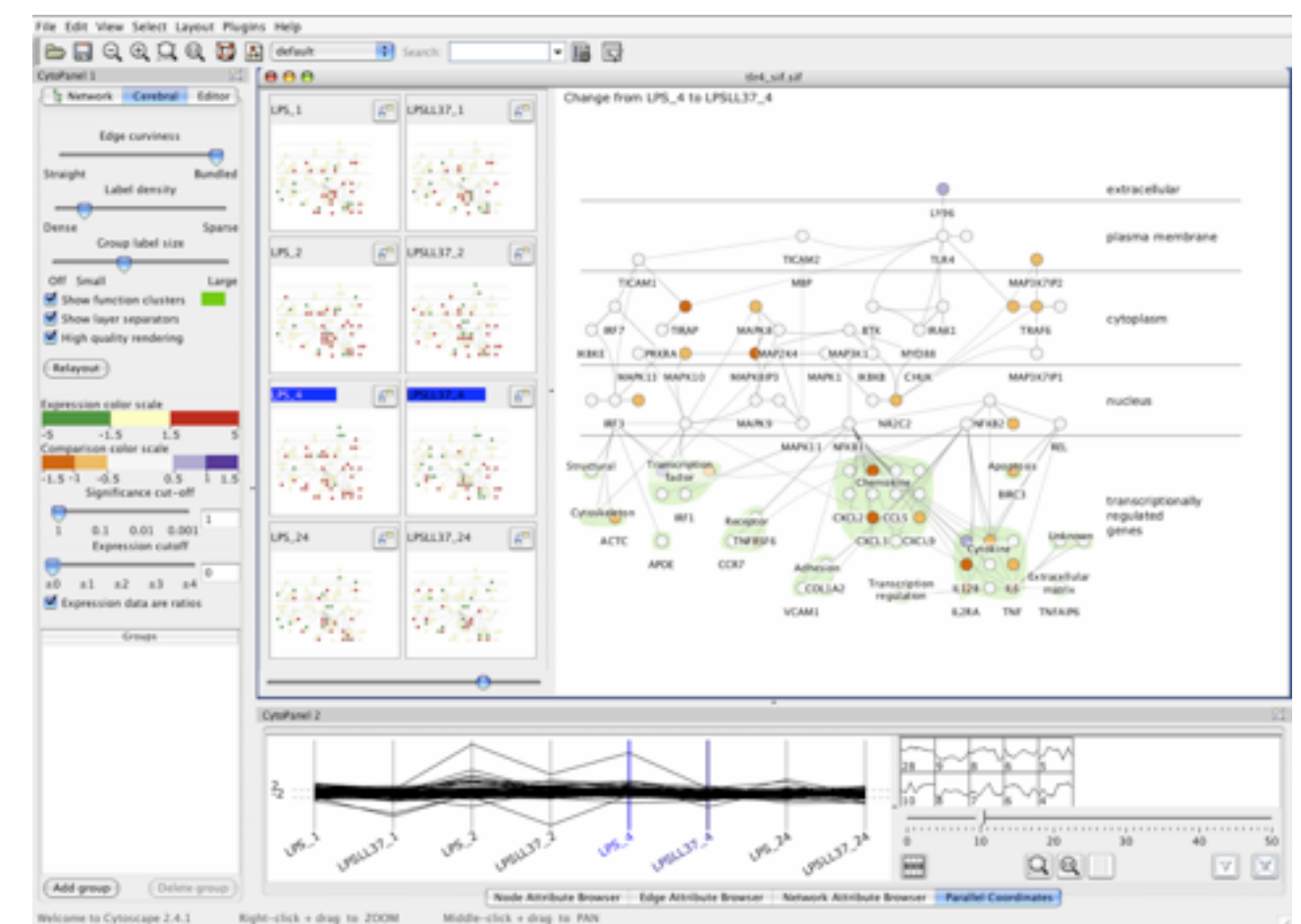- introduction

- Cerebral

- MizBee

- TreeJuxtaposer

- wrapup

Cerebral

Visualizing Multiple Experimental Conditions on a Graph with Biological Context
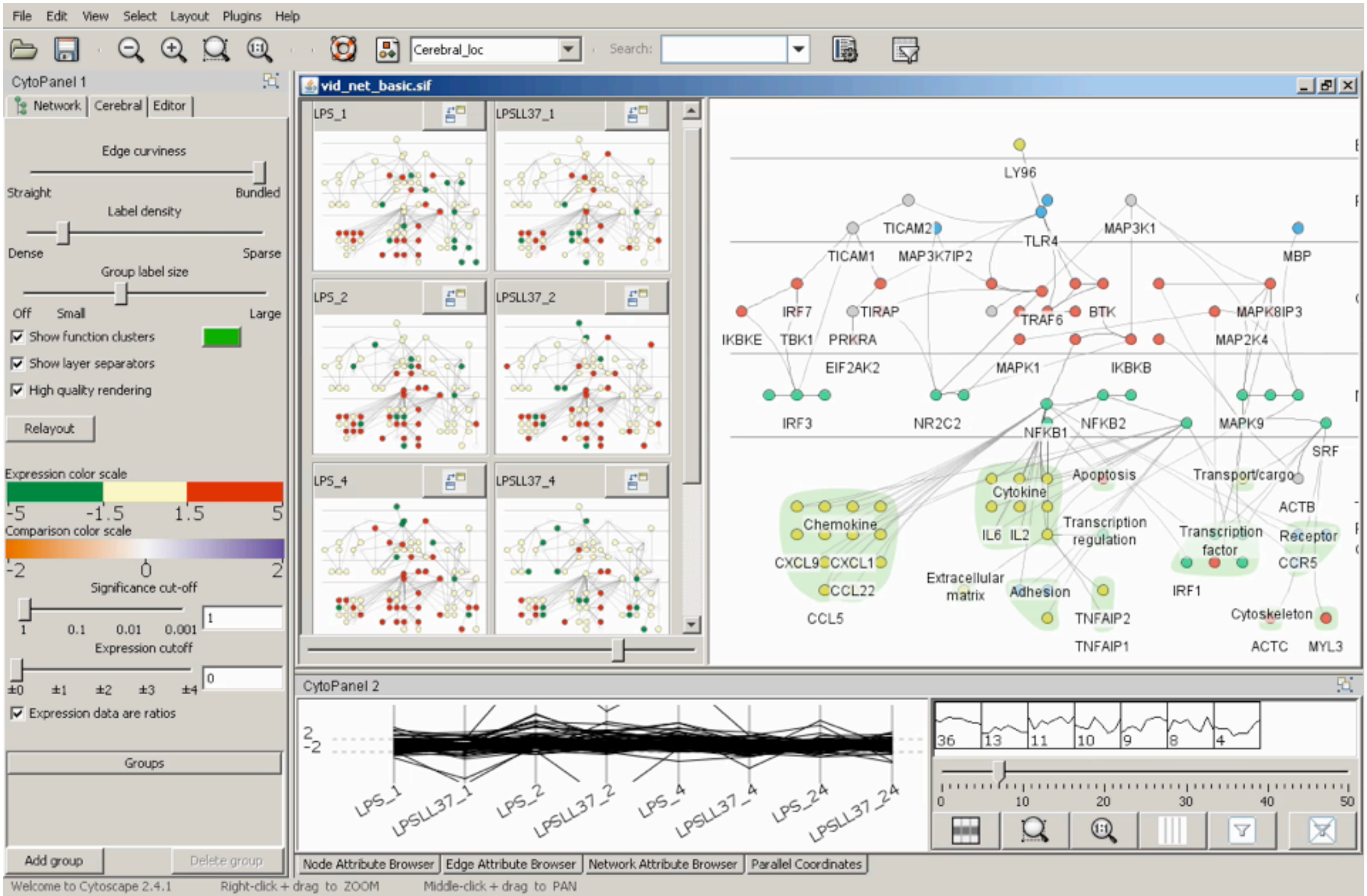
joint work with:

Aaron Barsky, Jennifer Gardy, Robert Kincaid

Cerebral video

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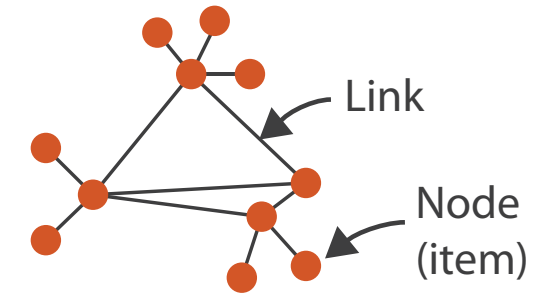
- **dataset types**
  - network
    - nodes: genes
    - links: known interactions between genes
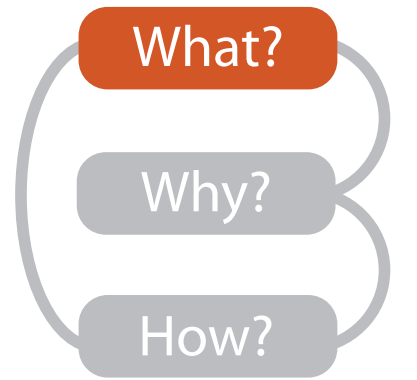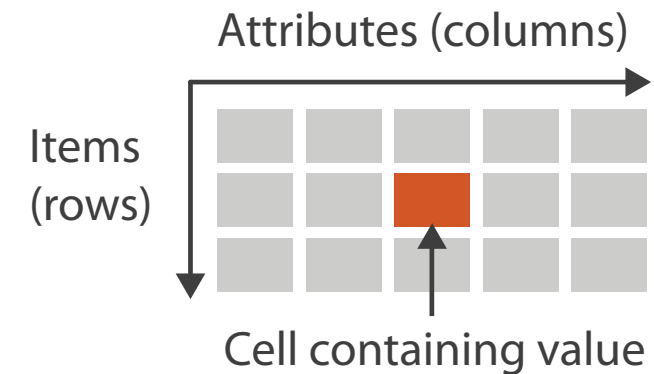 - table
    - quantitative attributes
      - gene expression levels for nodes across different experimental conditions
    - categorical attributes
      - subcellular location of interaction
      - functional groups

➔ Networks

Link

Node
(item)

➔ Tables

Attributes (columns)

Items
(rows)

Cell containing value

➔ **Attribute Types**
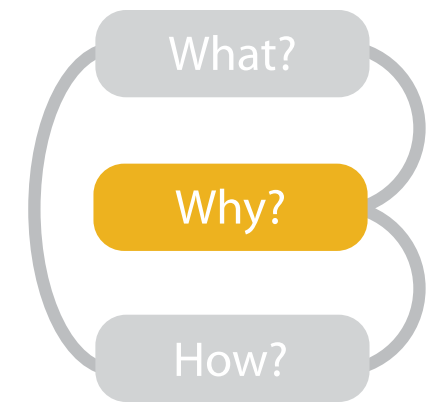
➔ Categorical

➔ Ordered

➔ *Ordinal*   ➔ *Quantitative*

Why: Task abstraction

- task: interpret microarray experiment results with respect to gene network
  - goal: accelerate existing discovery workflow
  - compare distributions between attributes
    - experimental conditions
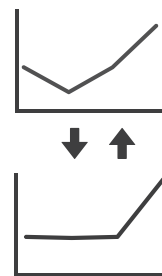 - interpret attributes in context of current network topological structure

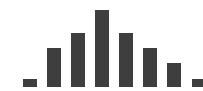➜ *Discover*

✍ Actions

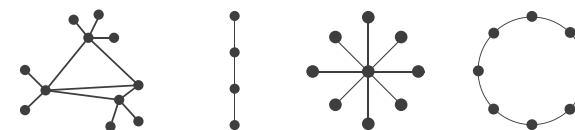➜ Compare

◎ Targets

➜ **Attributes**
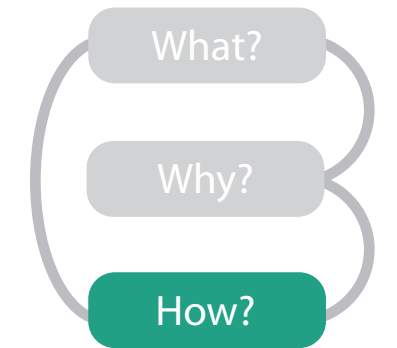
➜ One

➜ *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
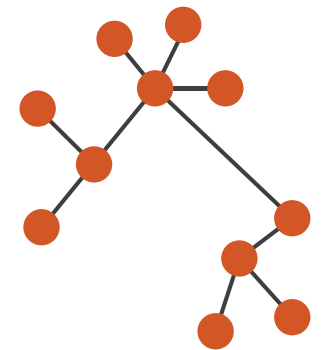 - cluster according to functional grouping

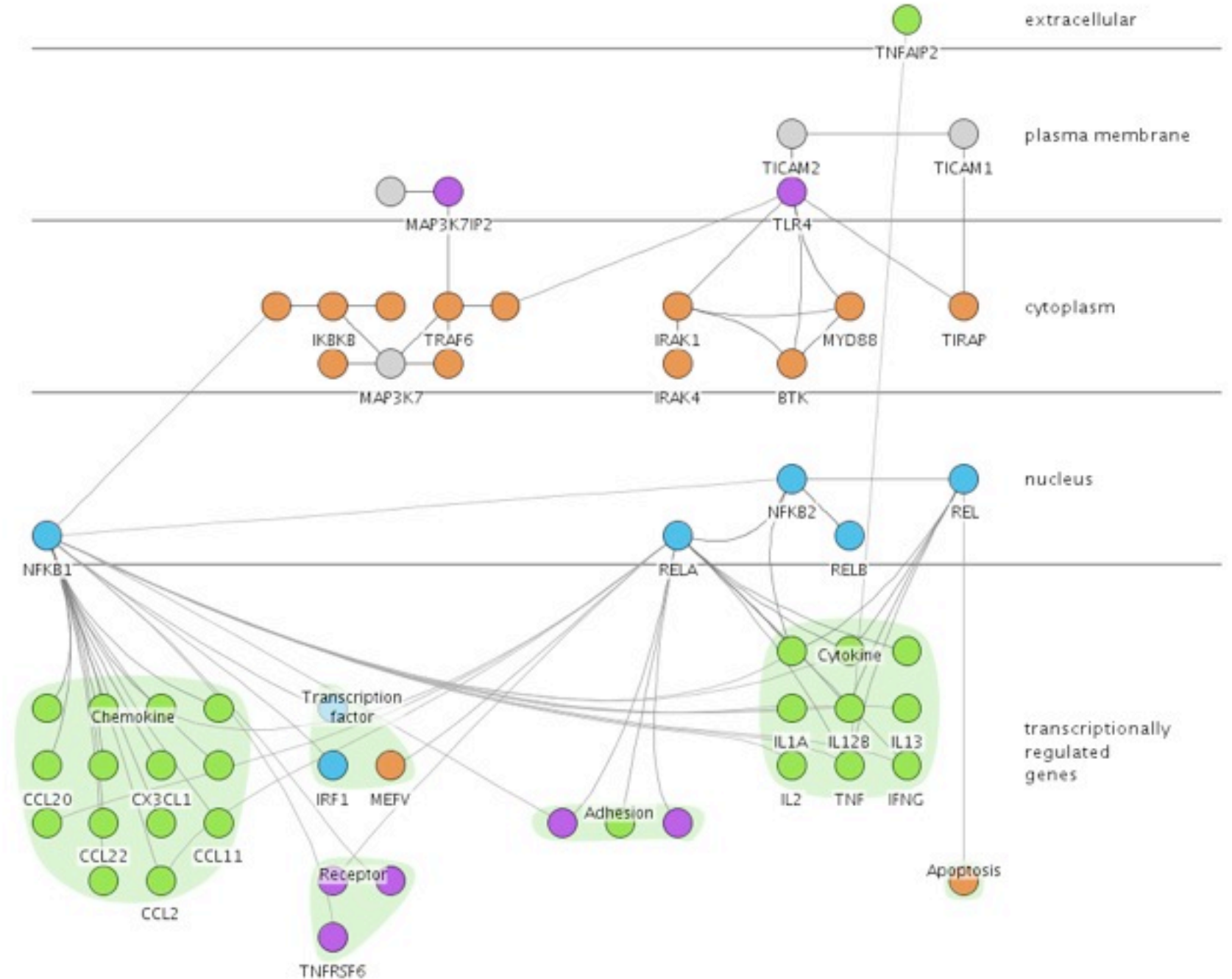**Arrange Networks And Trees**
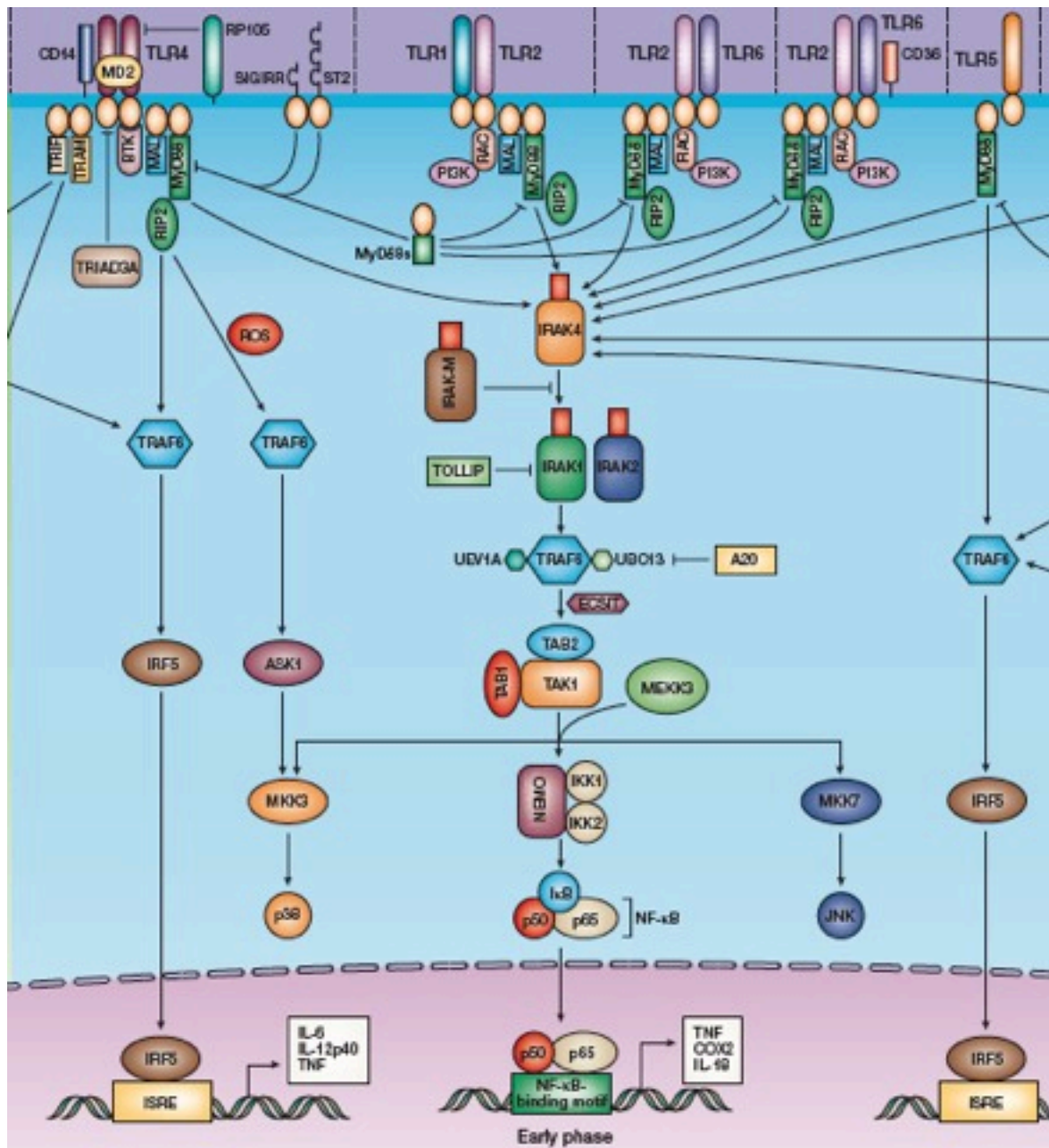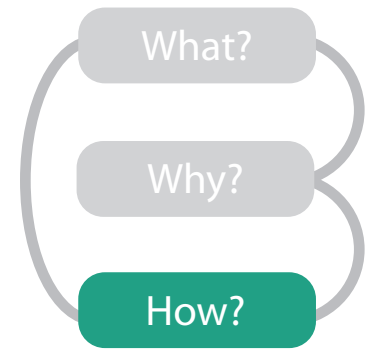
➔ **Node-link Diagrams**
Connections and Marks

✔ NETWORKS   ✔ TREES
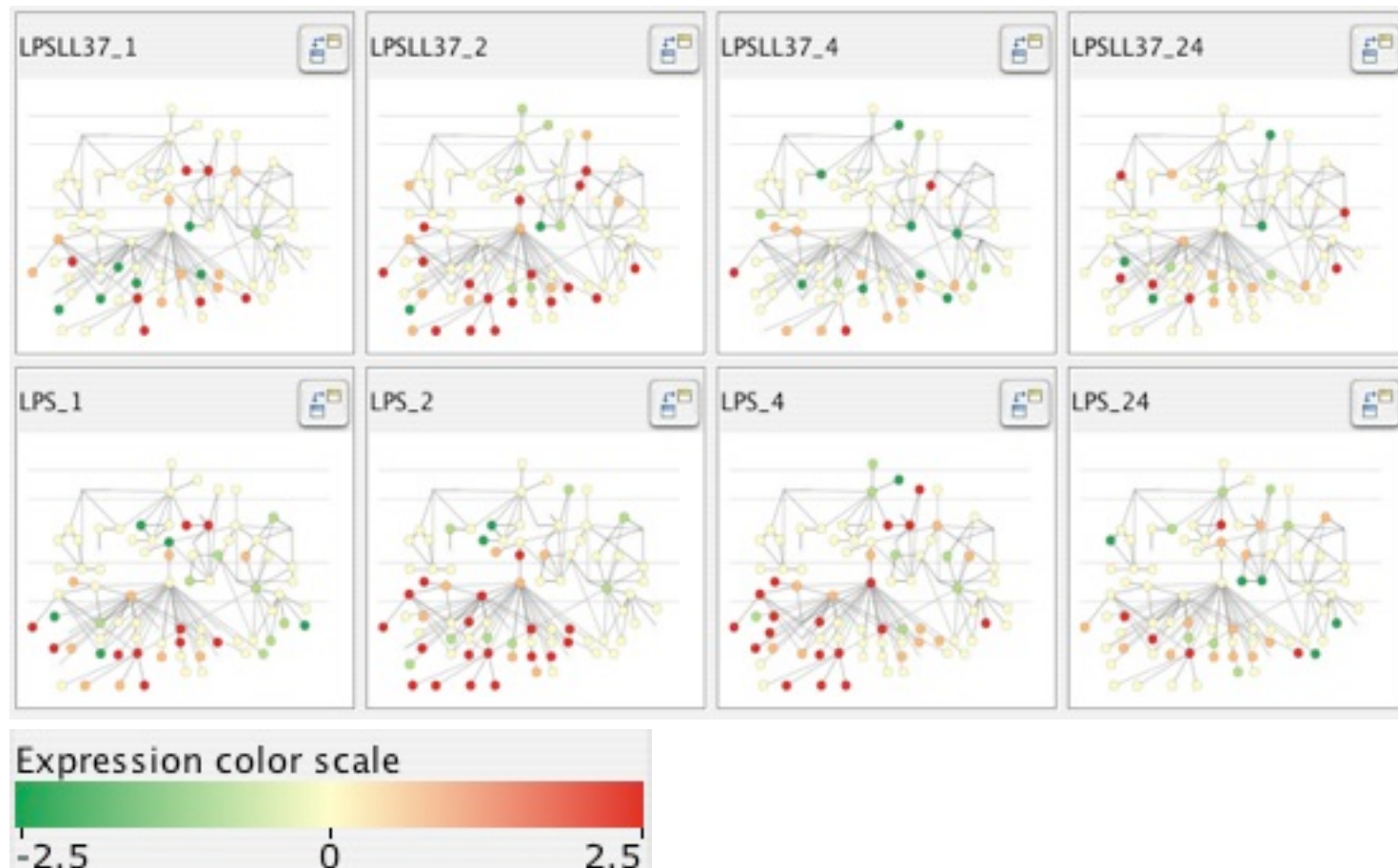
How: Arrange space

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

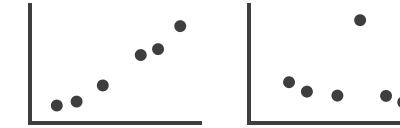http://www.nature.com/nri/focus/tlr/nri1397.html

How: Idiom design decisions

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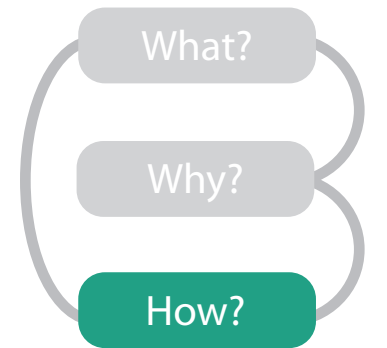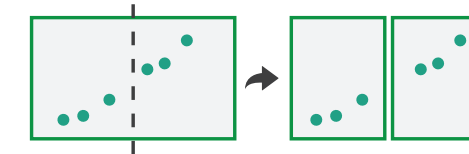   - nodes in each view colored by expression levels for experimental condition

Juxtapose

Partition

LPSLL37_1 LPSLL37_2 LPSLL37_4 LPSLL37_24

LPS_1 LPS_2 LPS_4 LPS_24

Expression color scale

-2.5   0   2.5

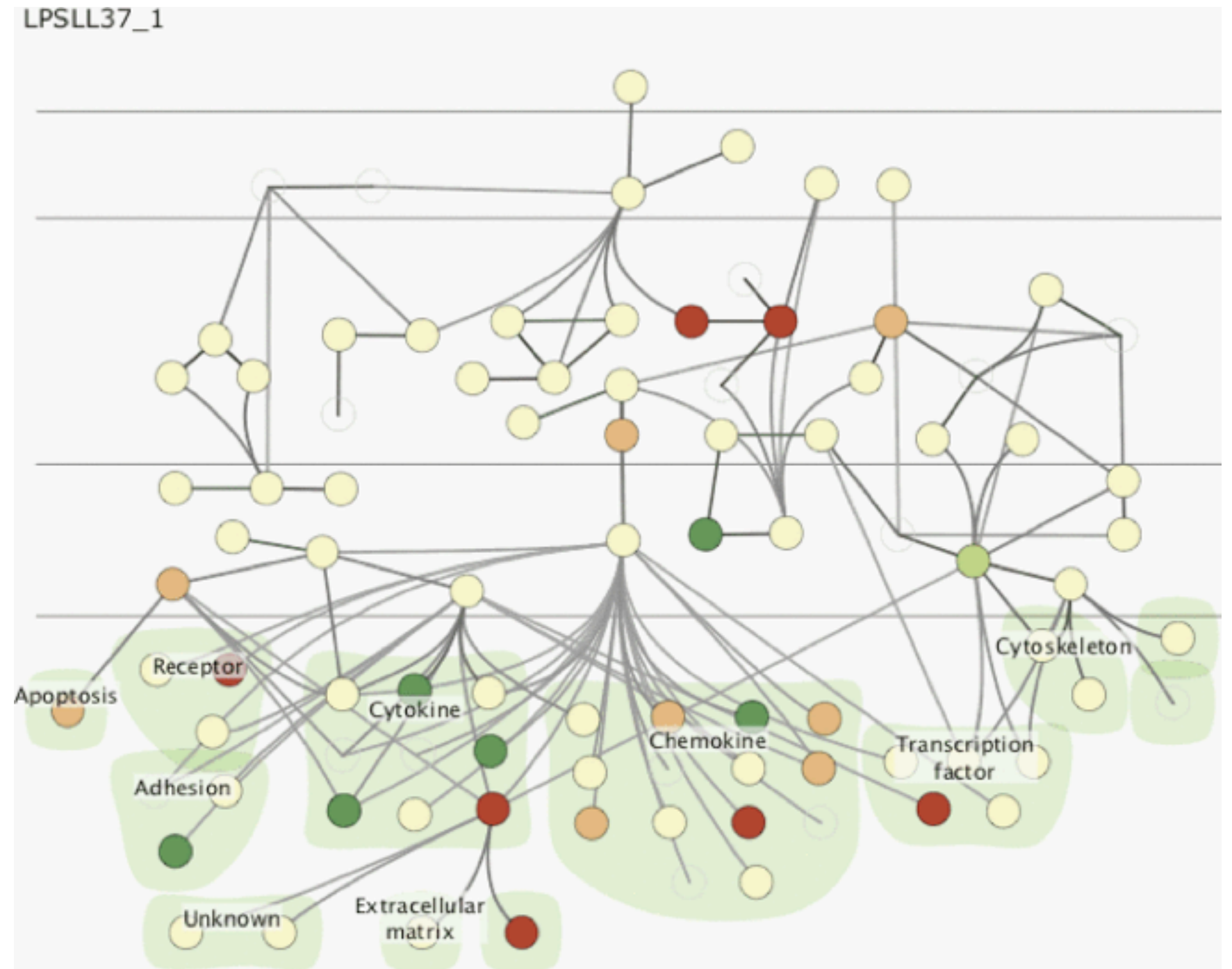|		Data		
		All	Subset	None
Encoding	Same	Redundant	*Overview/ Detail*	*Small Multiples*
	Different	*Multiform*	*Multiform, Overview/ Detail*	No Linkage

How: Juxtapose vs. animate

Manipulate

→ **Change**
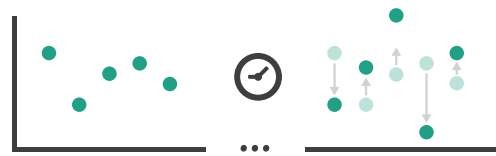
- comparison difficult across many frames with with many changes everywhere

- 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

Cerebral contributions

- multiple juxtaposed views support interactive comparison between gene expression level experimental data and network context

- automatic network layout algorithm in spirit of hand drawn diagrams
  - localization and functional group attributes affect spatial position

- open source
  - Cytoscape plugin
  - InnateDB database integration

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

Outline

- introduction

- Cerebral

- MizBee

- TreeJuxtaposer

- wrapup
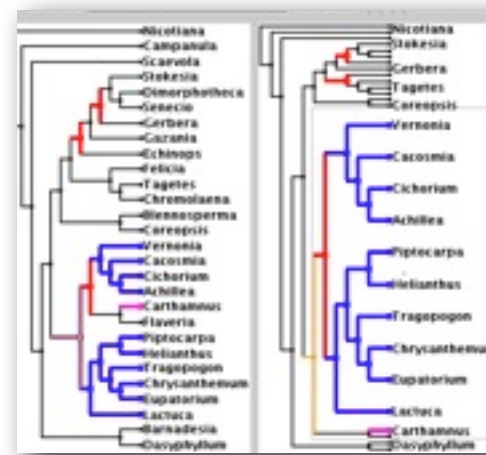
MizBee

A Multiscale Synteny Browser
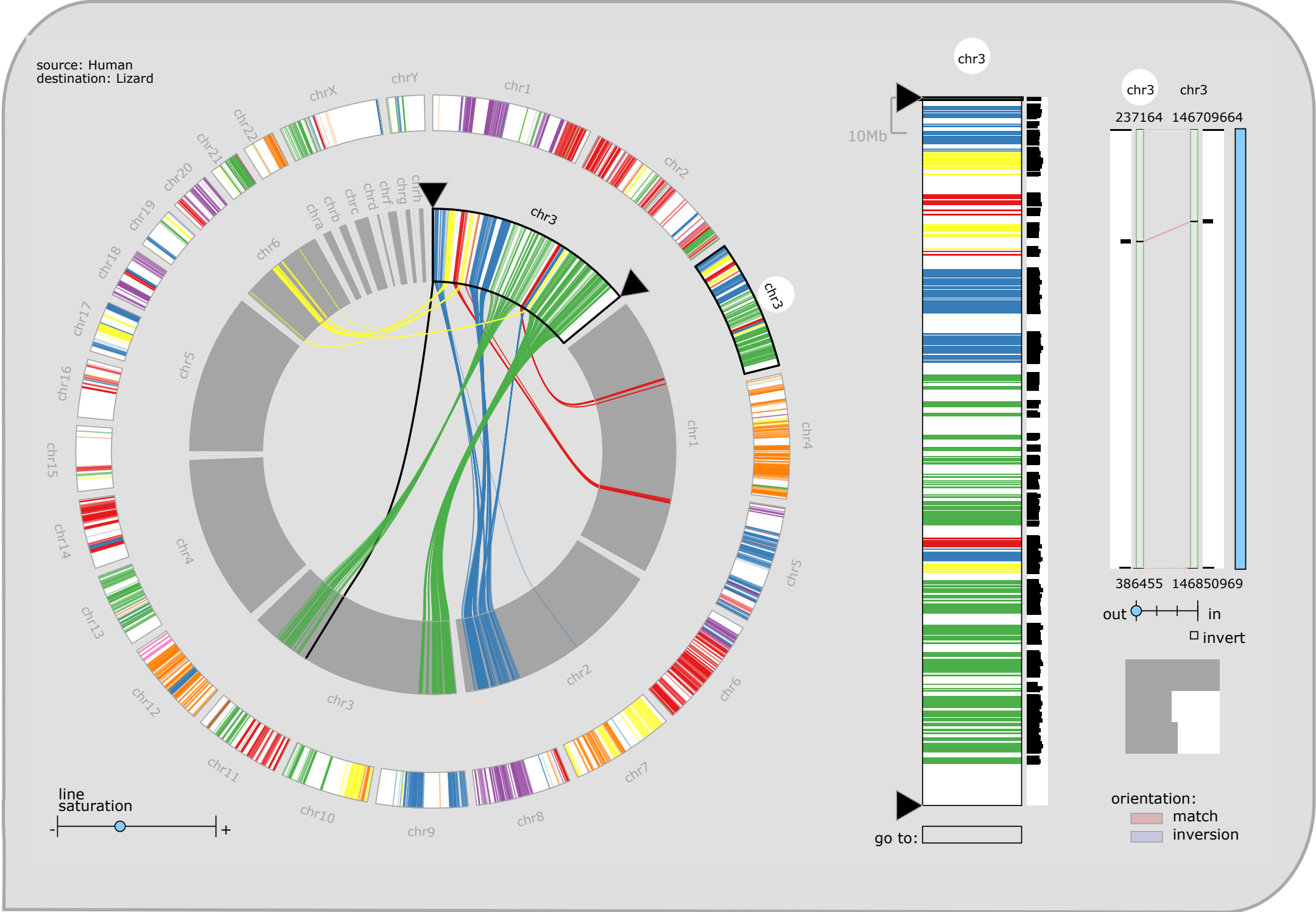
joint work with:

 Miriah Meyer, Hanspeter Pfister

http://www.cs.utah.edu/~miriah/mizbee

MizBee video

What: Data abstraction

- data: multiscale lists
  - features: hundreds of thousands
    - ordered attribute: position in chromosome sequence coordinates
    - categorical attributes: orientation, chromosome of matching feature
    - quantitative attributes: length, similarity score
  - syntenic blocks: thousands
    - contiguous sets of features on same chromosome
    - combine thresholded features if
      - destination chromosome and orientation match
      - close together
  - chromosomes: dozens
 - genomes: two
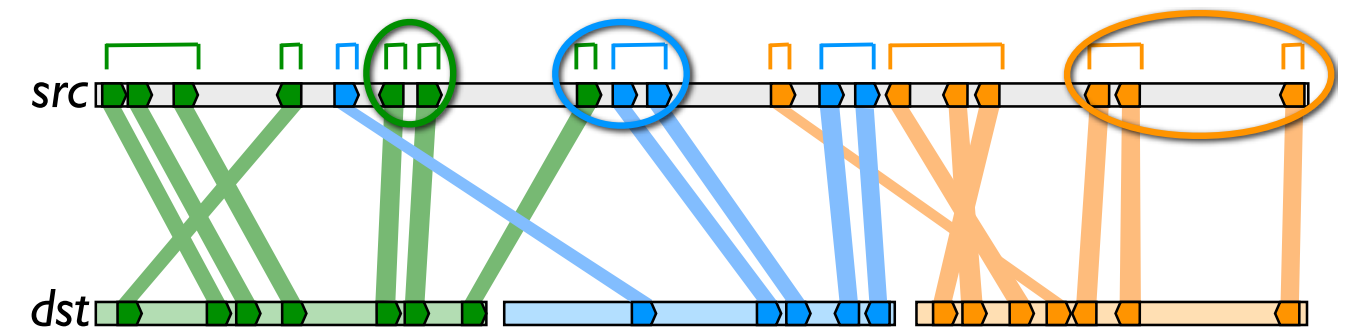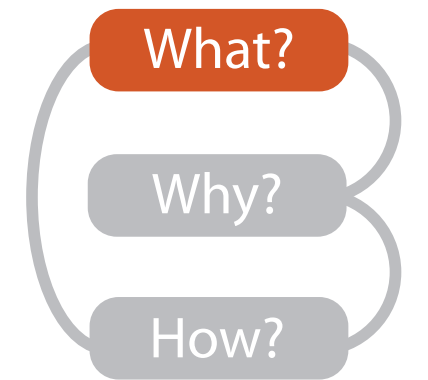
Why: Tasks in domain language

- analyze conservation (similarity) relationships between genomic features
  - high-level biology questions
    - evolution
      - how long ago did two species share common ancestor?
    - function
      - which segment of the genome is responsible for specific function in the cell?
    - ...
  - low-level data-centric questions
    - algorithm refinement
      - are paired features within a block contiguous?
      - which chromosomes share conserved blocks?
      - are similarity scores alike within block?
      - ...
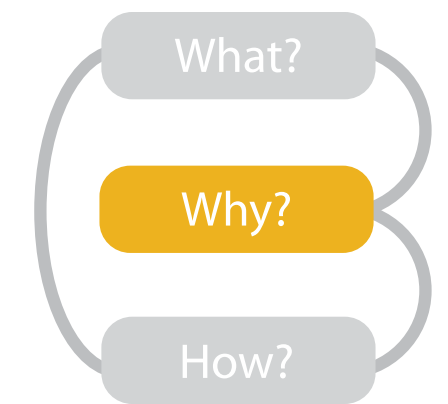
Why: Tasks abstraction
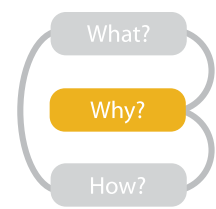
- relationship types: proximity, size, orientation, similarity
- data scales: genome, chromosome, block, feature
- topics: algorithm in/out, block reliability, high-level science

|	genome	chromosome	block	feature	proximity / location	size	orientation	similarity
Which chromosomes share conserved blocks?	X				X			
For one chromosome, how many other chromosomes does it share blocks with?	X	X			X			
What is the density of coverage and where are the gaps on: chromosomes? blocks?	X	X	X		X			
Where are the blocks: on chromosomes? around a specific location on a chromosome?	X	X			X			
What are the sizes and locations of other genomic features near a block?			X		X	X		
How large are the blocks?			X			X		
Do neighboring blocks go to the same: chromosomes? relative location on a chromosome?	X	X			X			
Are the orientations matched or inverted for: block pairs? feature pairs?		X	X				X	
Do the orientations match for pairs of: neighboring blocks? features within a block?		X	X				X	
Are similarity scores alike: with respect to neighboring blocks? within a block?		X	X					X
Are the paired features within a block contiguous?			X		X			
How large is a feature relative to other genes within a block?			X			X		
What are the sizes, locations, and names of features within a block?			X		X	X		
What are the differences between individual nucleotides of feature pairs?				X				X

relationship   scale

How: Idiom design choices

- encode match relationships between chromosome segments with both

  - color

  src
  dst

  - connection marks

  src
  dst
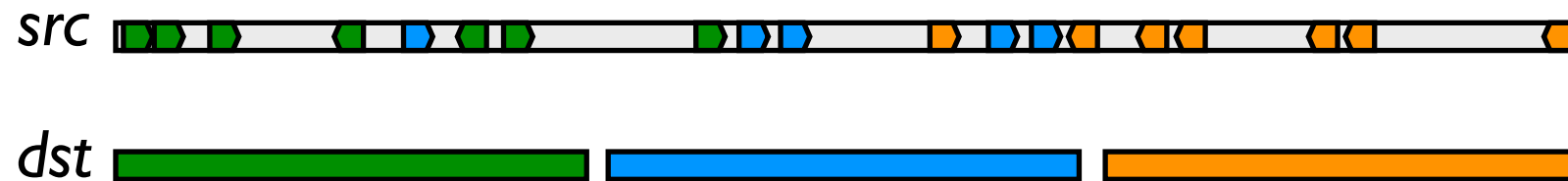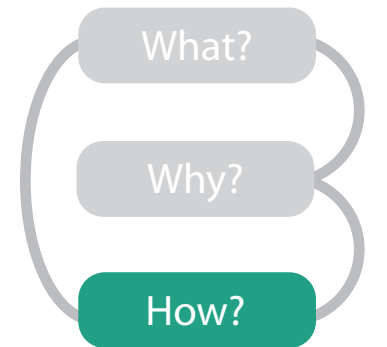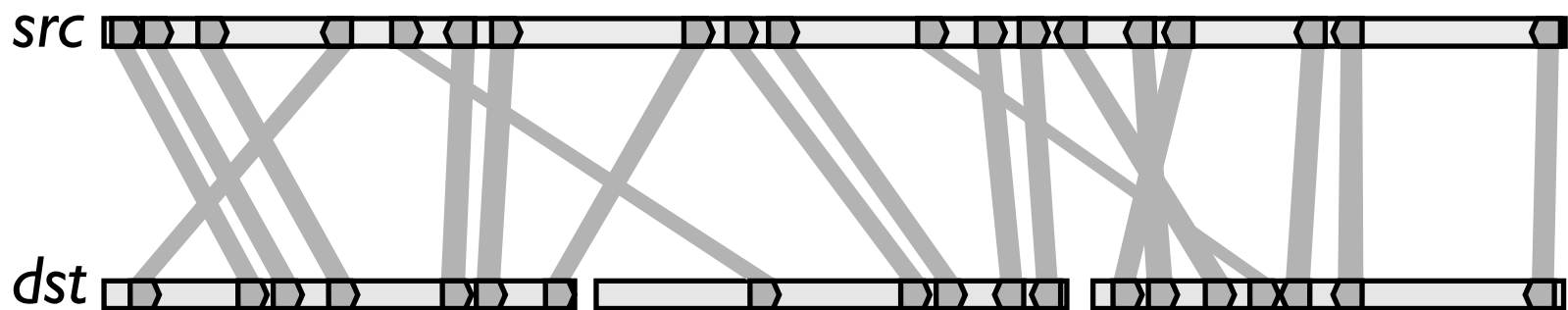
➔ Identity Channels: Categorical Attributes
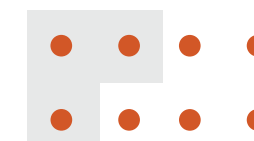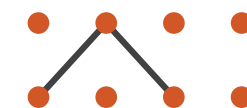
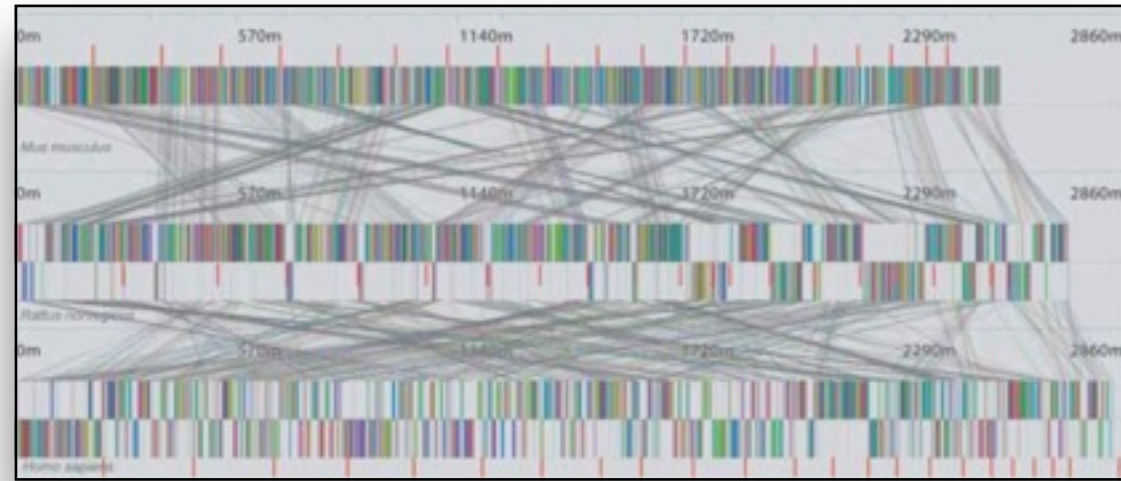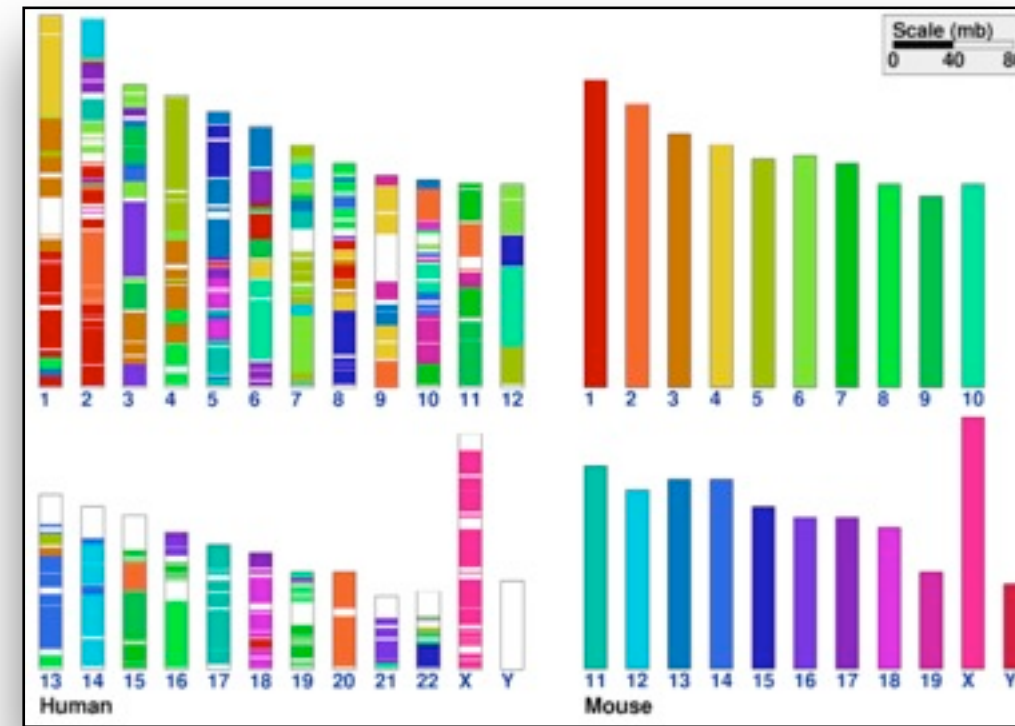Spatial region

Color hue

Motion

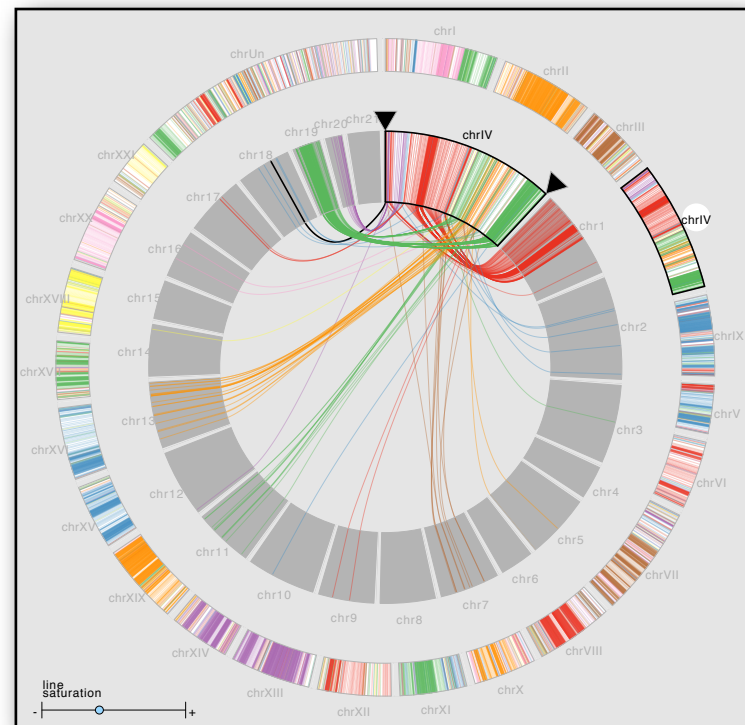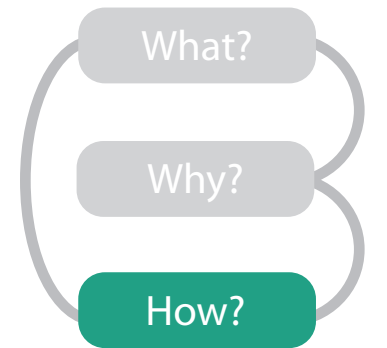Shape

Marks As Links

➔ Containment
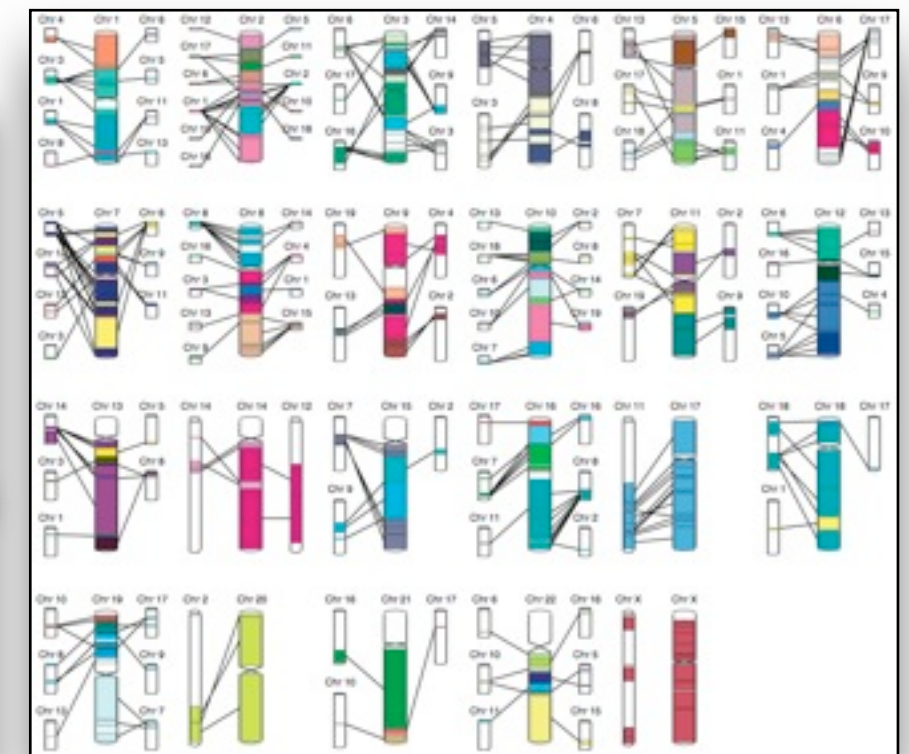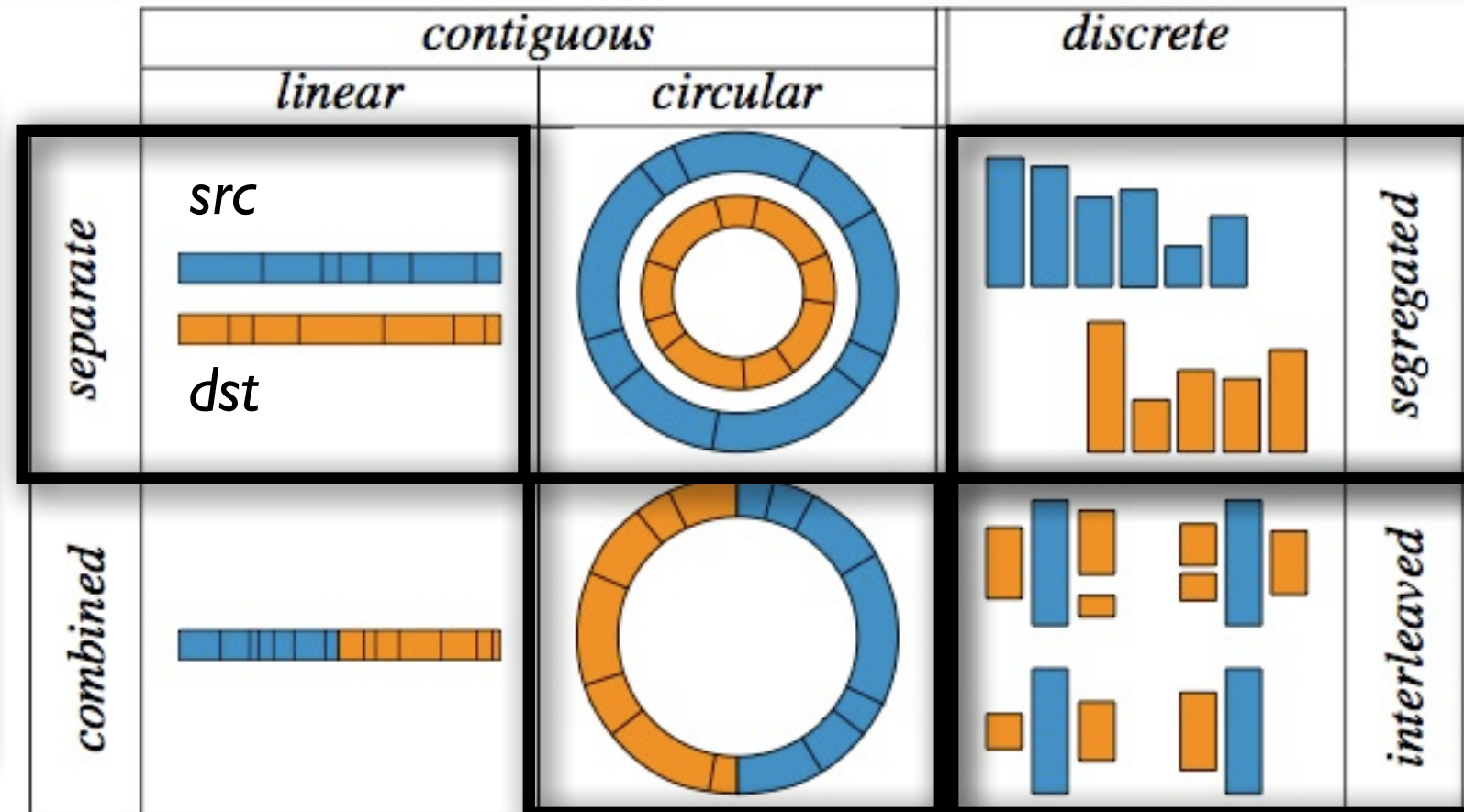
➔ Connection

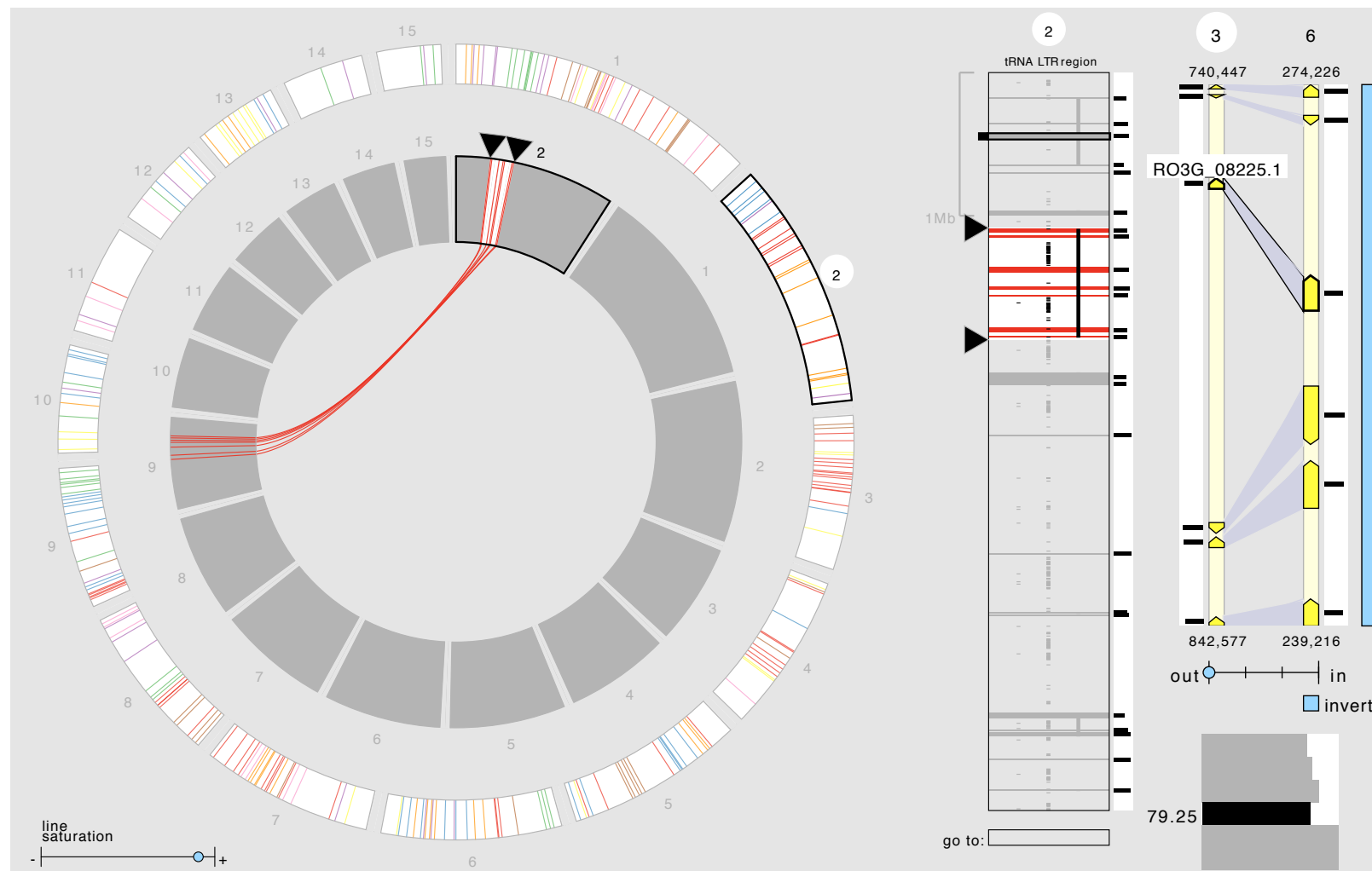# How: Arrange space

- design space of arrangements

Mauve [Darling04]

Cinteny

	contiguous		discrete	
	linear	*circular*		
separate	*src* *dst*			*segregated*
combined				*interleaved*

MizBee

Apollo [Lewis02]

37

How: Idiom design choices

- juxtapose linked views
  - *multiform overview-detail*
    - three views: genome, chromosome, block
    - different visual encoding in each

Facet

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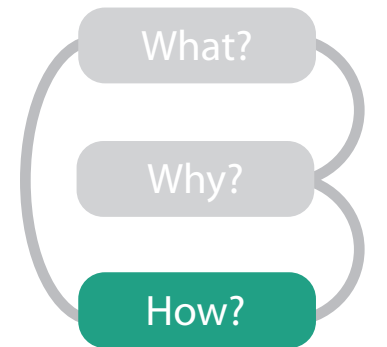
		Data		
		All	Subset	None
Encoding	Same	*Redundant*	*Overview/ Detail*	*Small Multiples*
	Different	*Multiform*	*Multiform, Overview/ Detail*	*No Linkage*

How: Idiom design choices

- axis orientation
  - radial: genome
  - rectilinear: chromosome, block
    - aligned position more accurate than angle
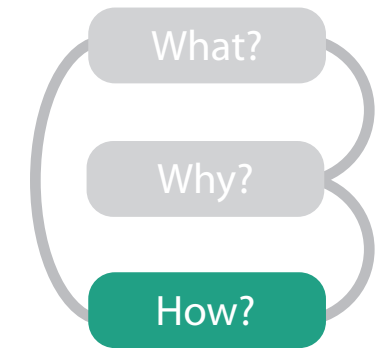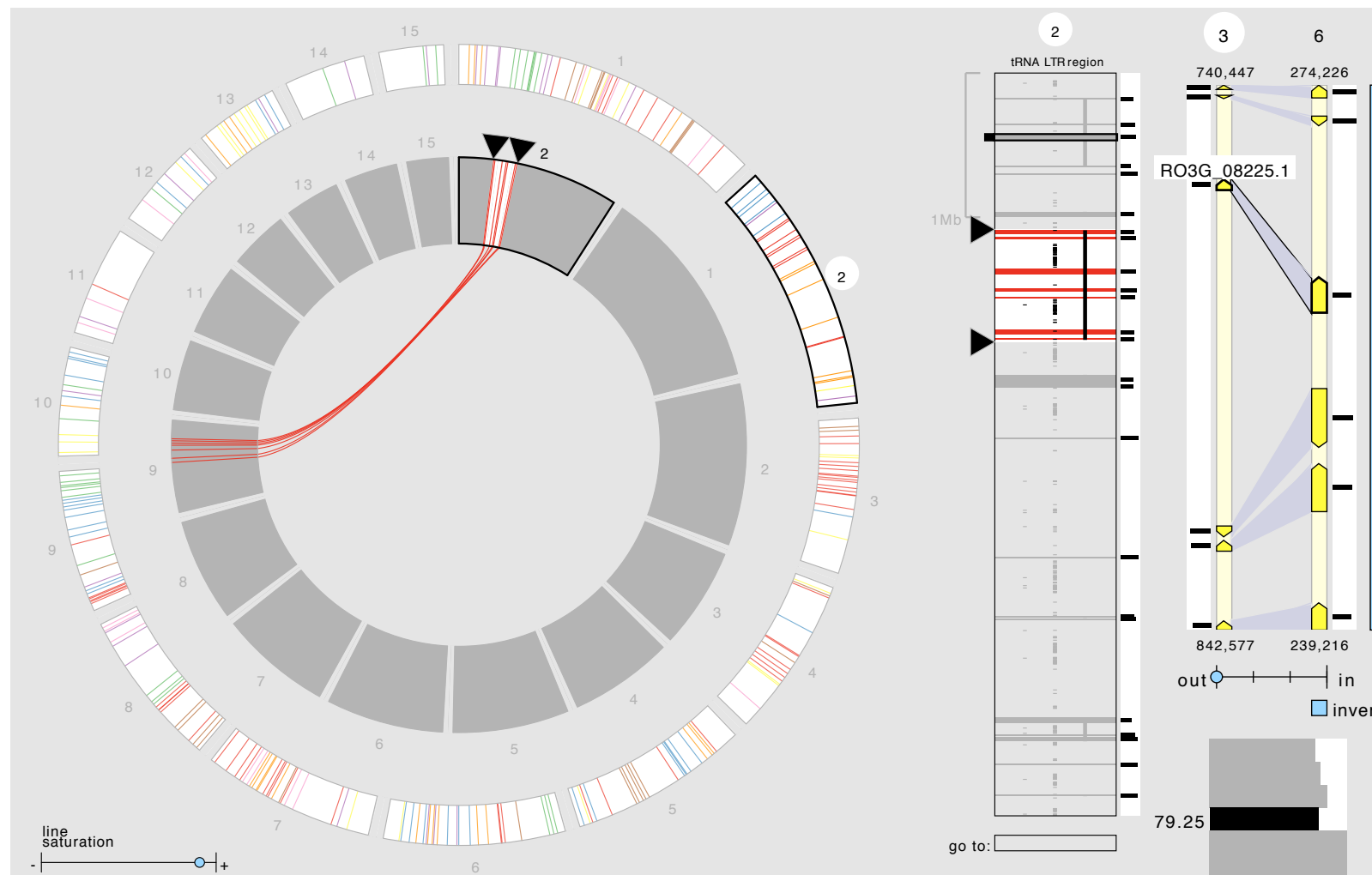
Arrange

→ **Axis Orientation**

→ Rectilinear   → Parallel   → Radial
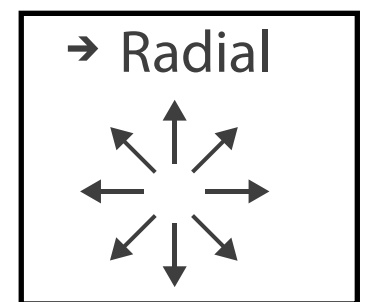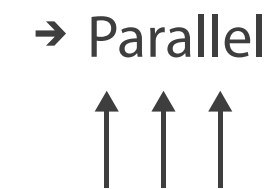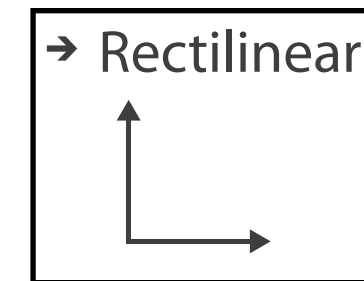
→ **Magnitude** Channels: **Ordered** Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

How: Idiom design choices

- filter
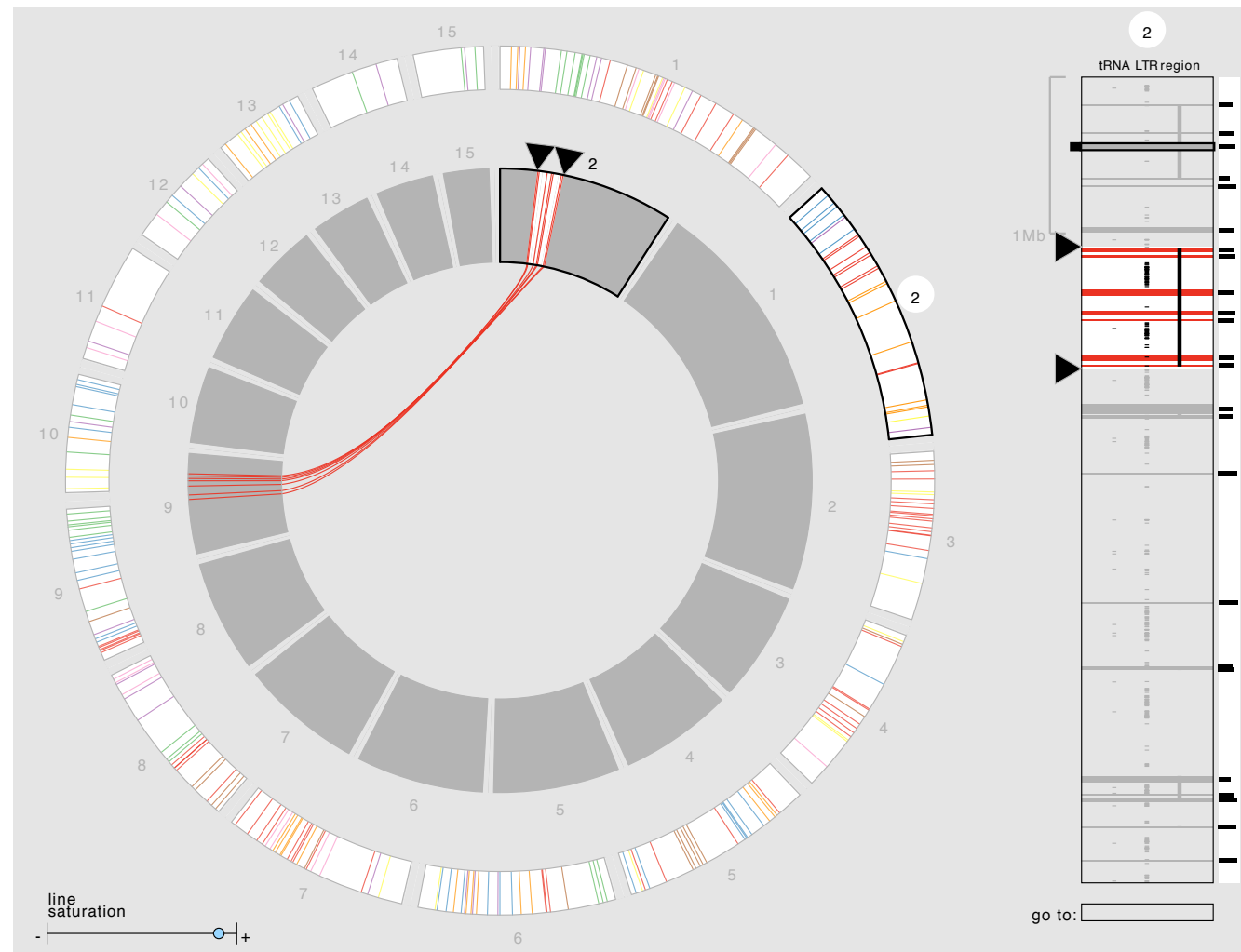
Reduce

→ **Filter**
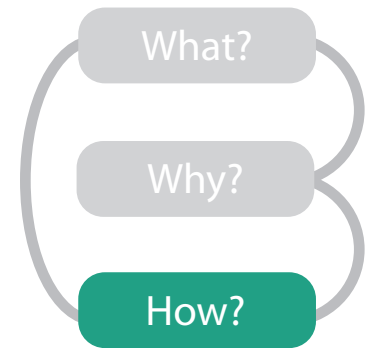
How: Idiom design choices

- outer ring: summarize relationships with color
  - select one chromosome from set of source chromosomes

- inner ring:
  - destination chromosomes around copy of selected source chromosome
  - show relationship details with connection marks as well as color

Manipulate

→ **Select**

Actions

→ **Query**

→ Identify   → Compare   → Summarise

What?

Why?

How?
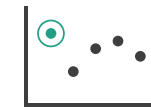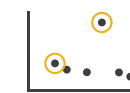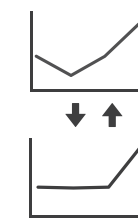
MizBee contributions

- first synteny browser with side-by-side linked views
  - across the range of scales
  - encoding all four conservation relationship types
    - proximity, size, orientation, similarity

- open source
  http://www.cs.utah.edu/~miriah/mizbee

Outline

- introduction

- Cerebral

- MizBee

- TreeJuxtaposer

- wrapup
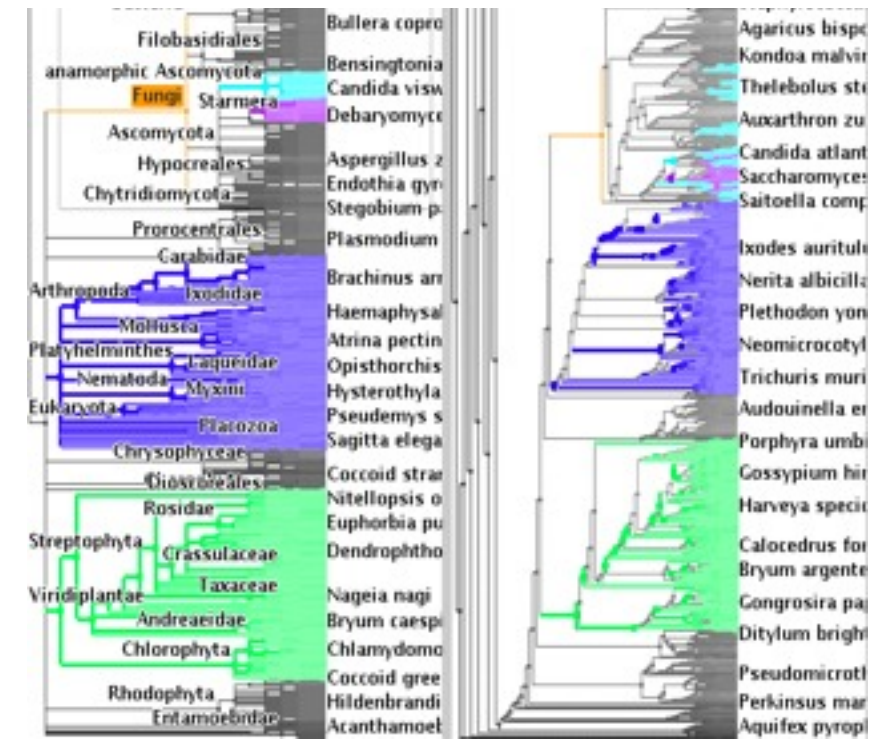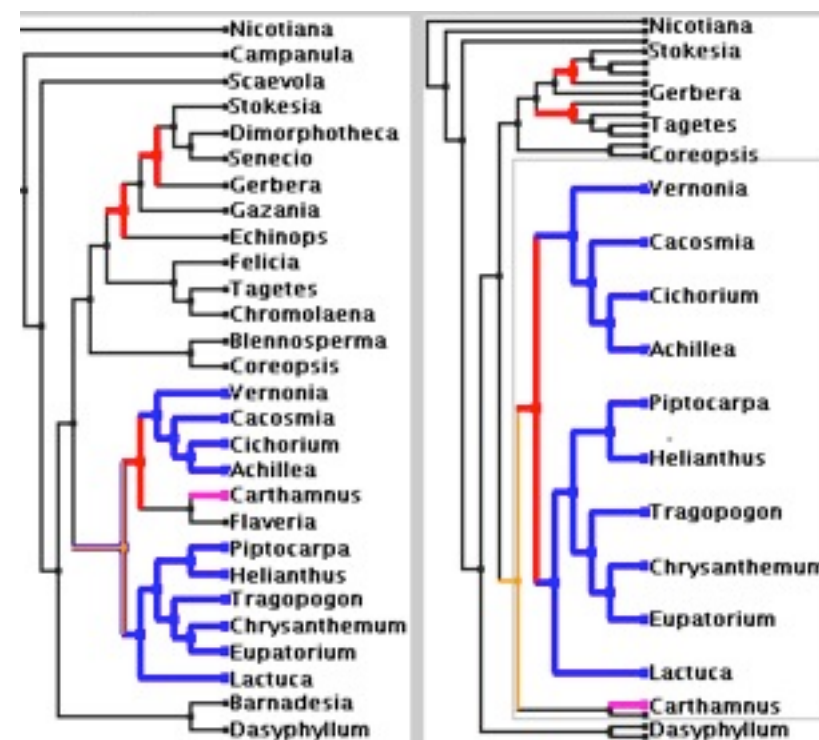
TreeJuxtaposer

Scalable Tree Comparison using Focus+Context with Guaranteed Visibility

joint work with:

François Guimbretière, Serdar Tasiran, Li Zhang, Yunhong Zhou

http://www.cs.ubc.ca/labs/imager/tr/2003/tj/

TreeJuxtaposer video
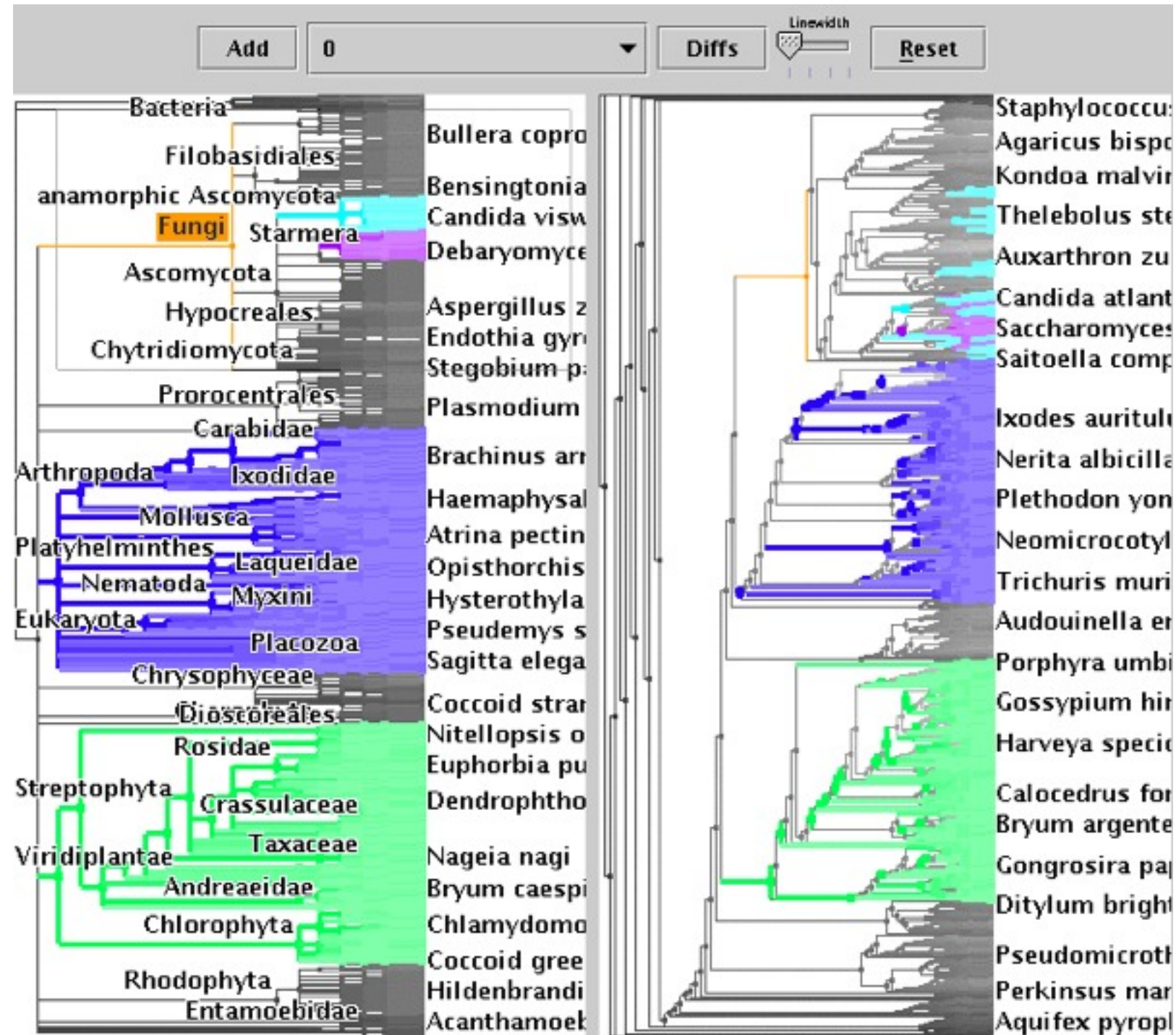
What and why: Data and task abstraction

- ## data: trees
  - phylogenetic tree reconstruction
    - siblings unordered, interior nodes inferred

- ## task: compare topological structure
  - larger query scopes require more explicit tool support
    - compare several is more difficult than identify/inspect one
      - even trickier: summarize all

- ## derived data: structural differences
  - best corresponding node in other tree

➔ **Dataset Types**

➔ *Trees*
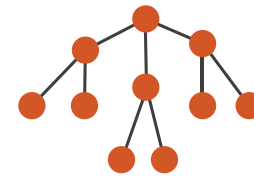
◎ **Targets**

➔ **Network Data**

➔ Topology
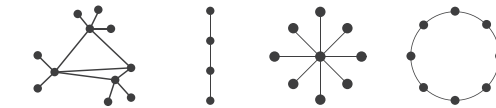
➔ *Paths*

✎ **Actions**

➔ **Query**

➔ Identify   ➔ Compare   ➔ Summarise

➔ *Derive*
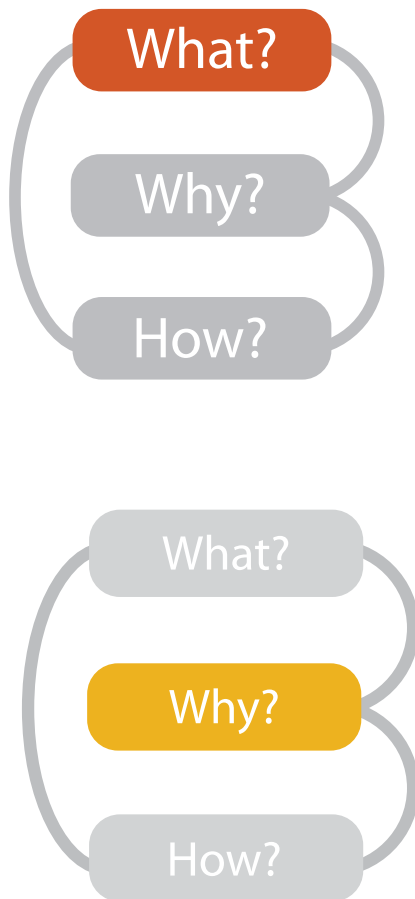
How: Idiom design decisions

- juxtapose linked views
  - show two tree layouts side by side
  - linked navigation

- encode with color: linked highlighting
  - structural differences
  - corresponding subtree (click select)
  - best corresponding node (hover select)

Facet

Manipulate

➔ **Juxtapose**

➔ **Select**

Facet

➔ **Juxtapose and Coordinate Views**

➔ Share Encoding: Same/Different

➔ *Linked Highlighting*

➔ Share Data: All/Subset/None

➔ Share Navigation
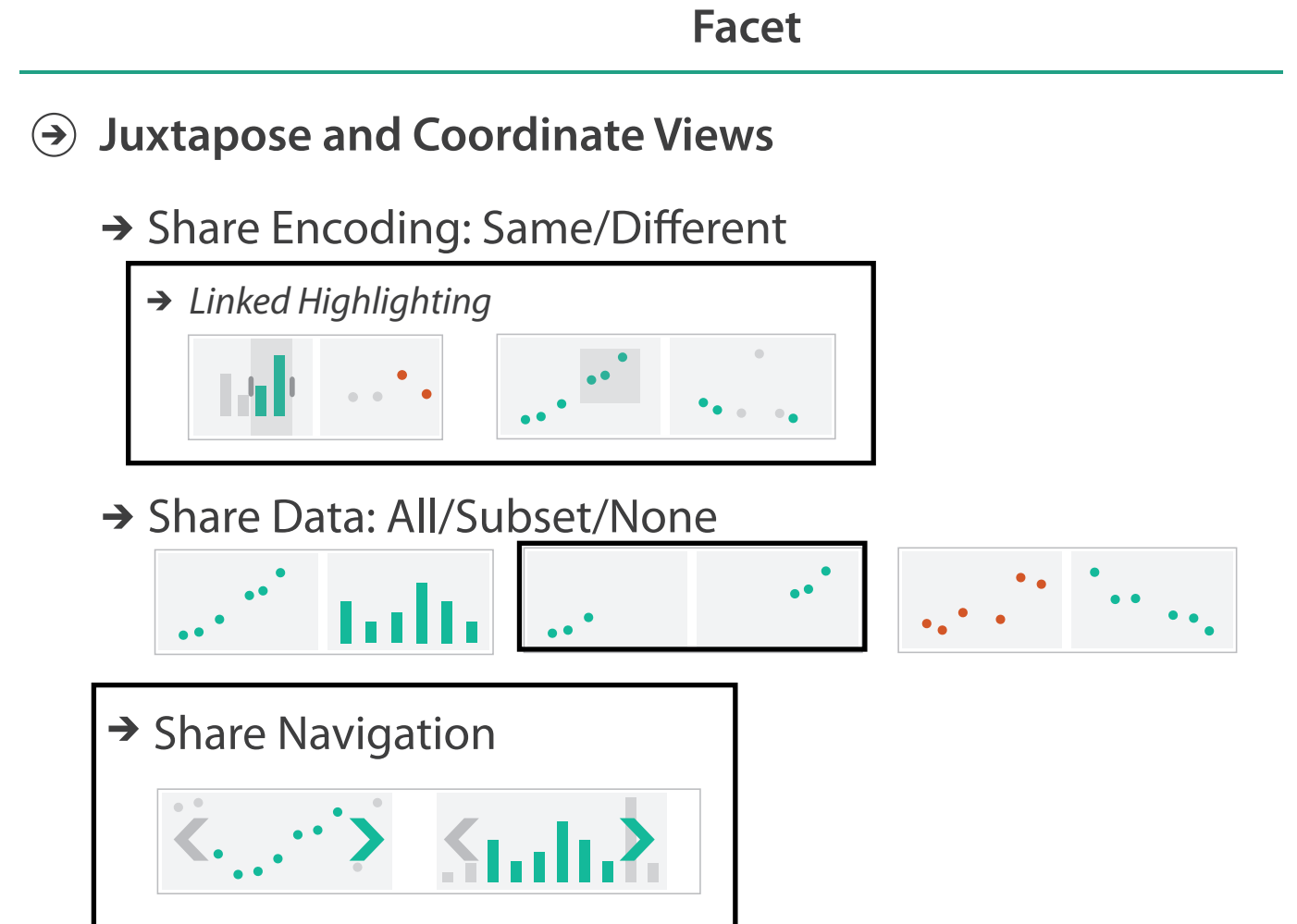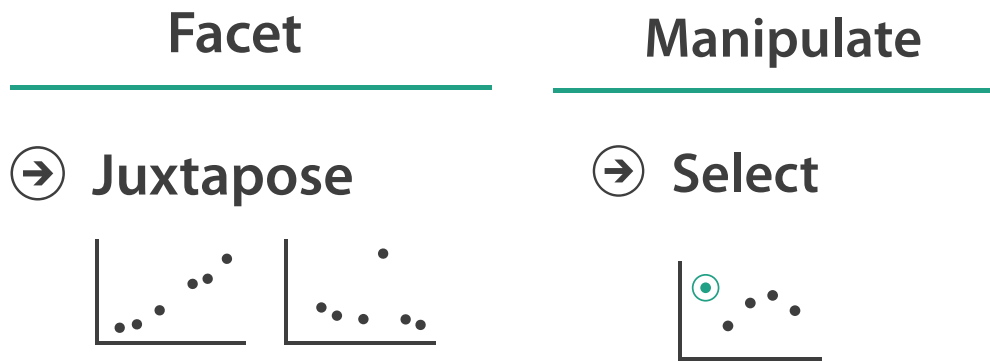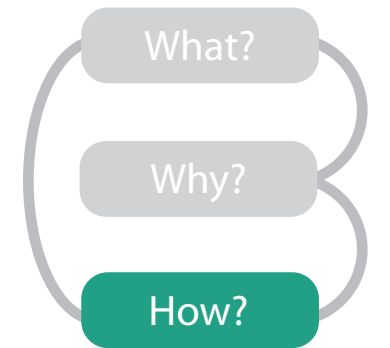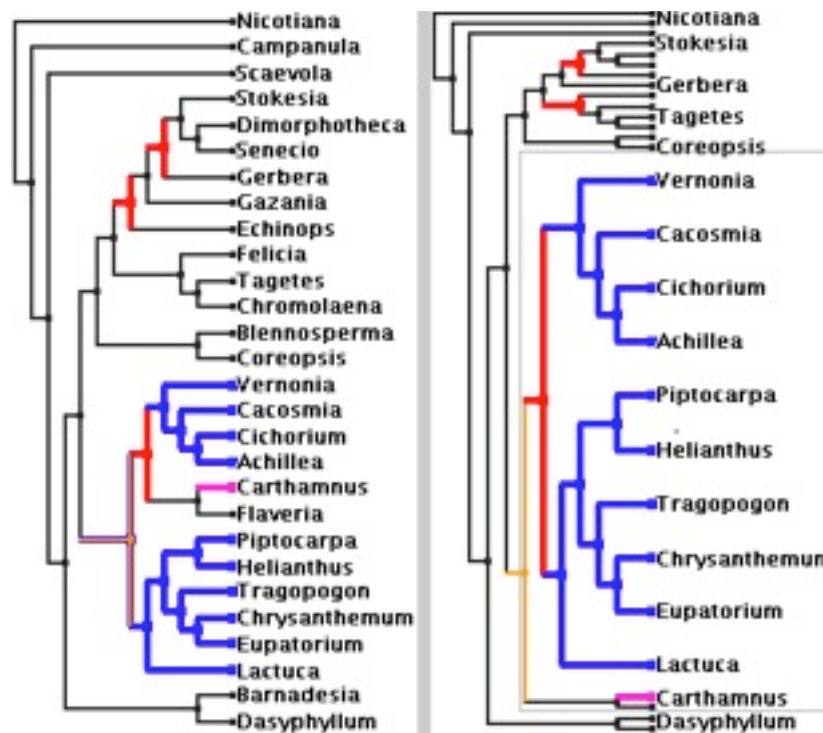
How: Idiom design decisions

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

Filter

Aggregate
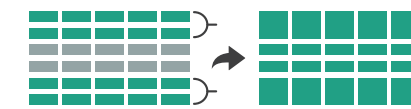
Embed

➜ Distort Geometry

What?

Why?

How?
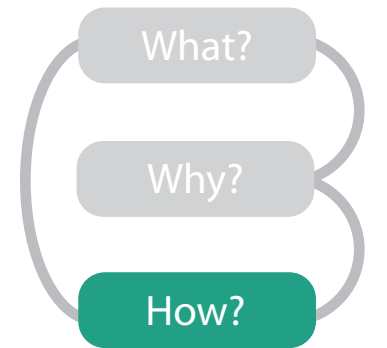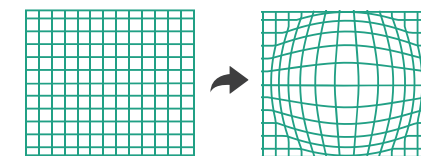
TreeJuxtaposer contributions

- first interactive tree comparison system
  - derive structural difference data to support comparison task
    - subquadratic algorithm: best corresponding node
  - juxtapose views with cross-dataset linked highlighting
- embed focus+context information in single view with stretch and squish navigation
  - sublinear algorithm: guaranteed visibility of structure marks even when squished

- open source
  http://olduvai.sf.net/tj

Outline

- introduction

- Cerebral

- MizBee

- TreeJuxtaposer

- wrapup
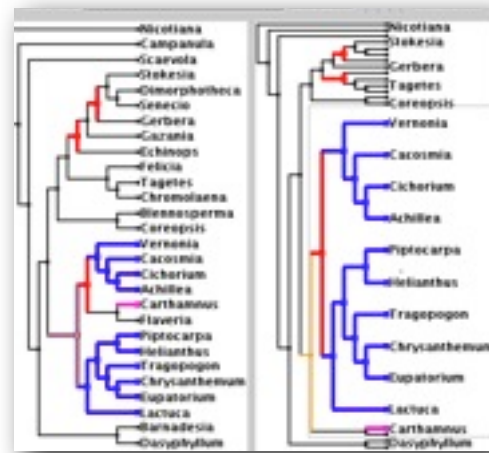
Visualization Analysis and Design
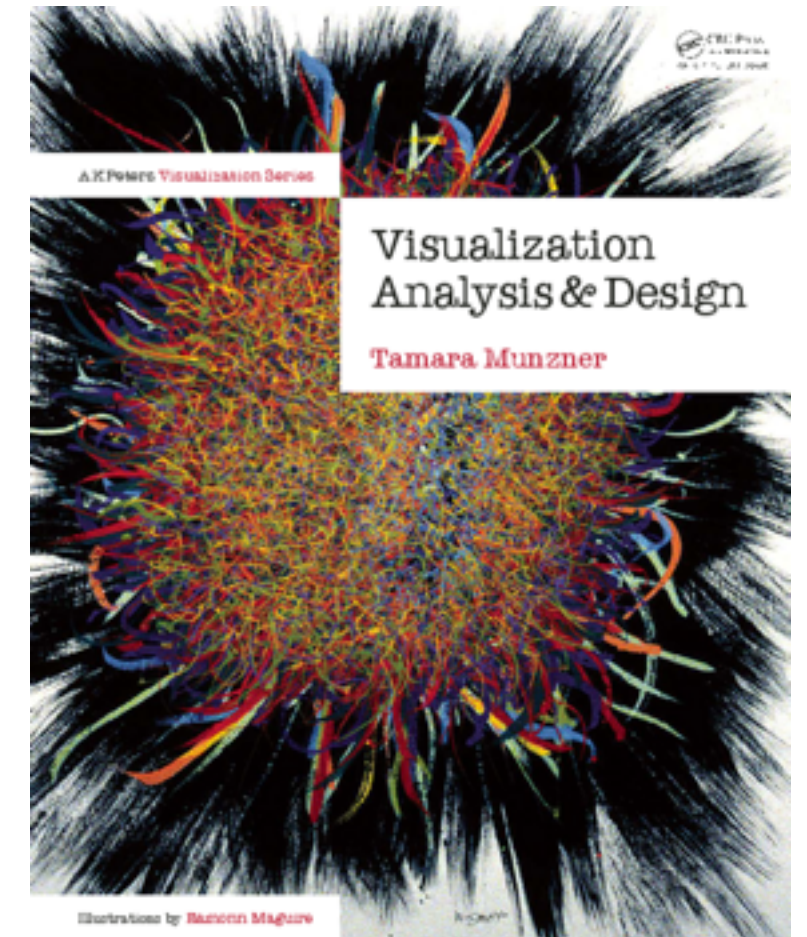
Visualization Analysis and Design
Munzner. AK Peters Visualization Series, CRC Press, 2014.

What?

Why?

How?
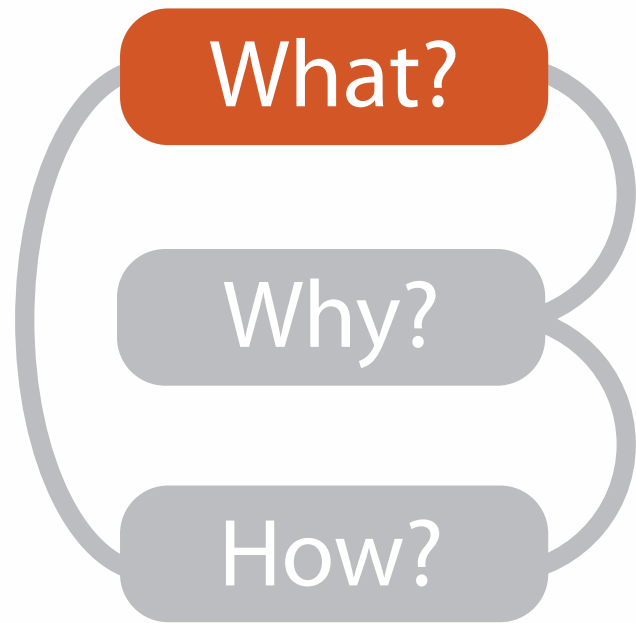
Datasets

Attributes

➔ **Data Types**

➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

➔ **Data and Dataset Types**

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

➔ **Dataset Types**

➔ Tables

Attributes (columns)

Items (rows)

Cell containing value

➔ Networks

Link

Node (item)

➔ Fields (Continuous)

Grid of positions

Cell

Attributes (columns)

Value in cell
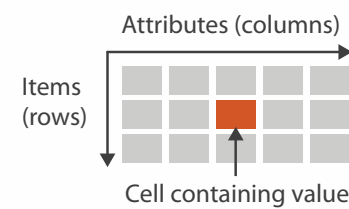
➔ *Multidimensional Table*

Key 1

Key 2

Value in cell

Attributes

➔ *Trees*

➔ Geometry (Spatial)

Position

➔ **Dataset Availability**

➔ Static

➔ Dynamic

➔ **Attribute Types**

➔ Categorical

➔ Ordered

➔ *Ordinal*

➔ *Quantitative*
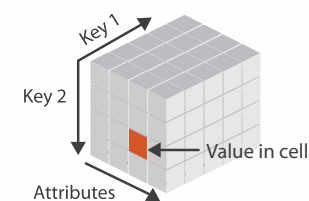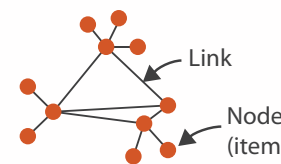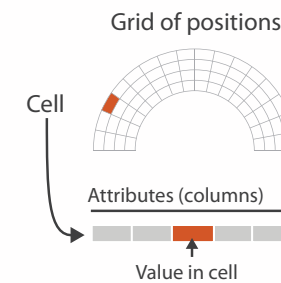
➔ **Ordering Direction**

➔ Sequential

➔ Diverging

➔ Cyclic

52

What?

Why?

How?

- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Why?

✎ Actions

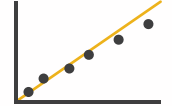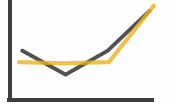**➔ Analyze**

➔ Consume

➔ *Discover* ➔ *Present* ➔ *Enjoy*

➔ Produce

➔ *Annotate* ➔ *Record* ➔ *Derive*

➔ Search

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

➔ Query

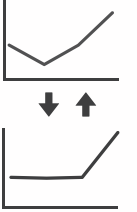➔ Identify ➔ Compare ➔ Summarize

◎ Targets

➔ All Data

➔ Trends ➔ Outliers ➔ Features

➔ Attributes

➔ One ➔ Many

➔ *Distribution* ➔ *Dependency* ➔ *Correlation* ➔ *Similarity*

➔ *Extremes*

➔ Network Data

➔ Topology

➔ *Paths*

➔ Spatial Data

➔ Shape

Encode

Arrange

➔ Express

➔ Separate

➔ Order

➔ Align

➔ Use

Map
from categorical and ordered attributes

➔ Color

➔ *Hue*

➔ *Saturation*
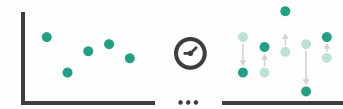
➔ *Luminance*

➔ Size, Angle, Curvature, ...

➔ Shape

➔ Motion
Direction, Rate, Frequency, ...
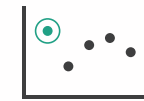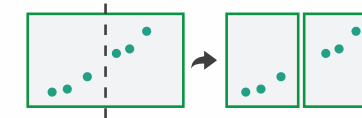
Manipulate

Change

Select

Navigate

Facet

Juxtapose

Partition

Superimpose

Reduce

Filter

Aggregate

Embed

What?
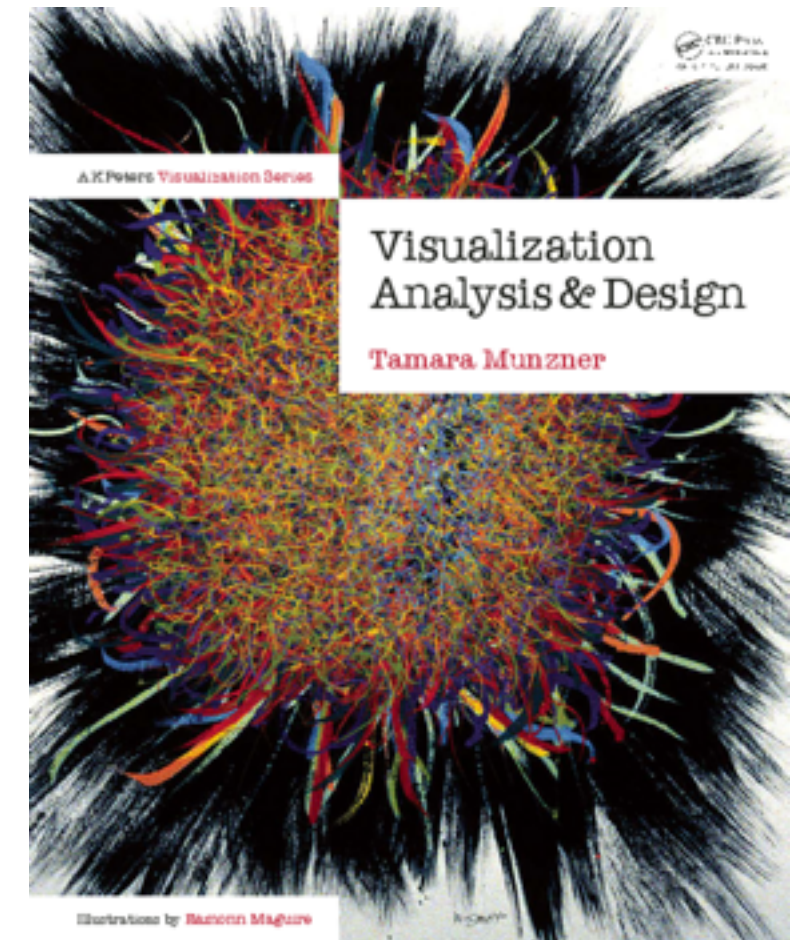
Why?

How?

More Information

- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#vanbug15

- 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

  - 20% promo code for book+ebook combo: HVN17

  - http://www.crcpress.com/product/isbn/9781466508910

- acknowledgements

@tamaramunzner