Data Visualization

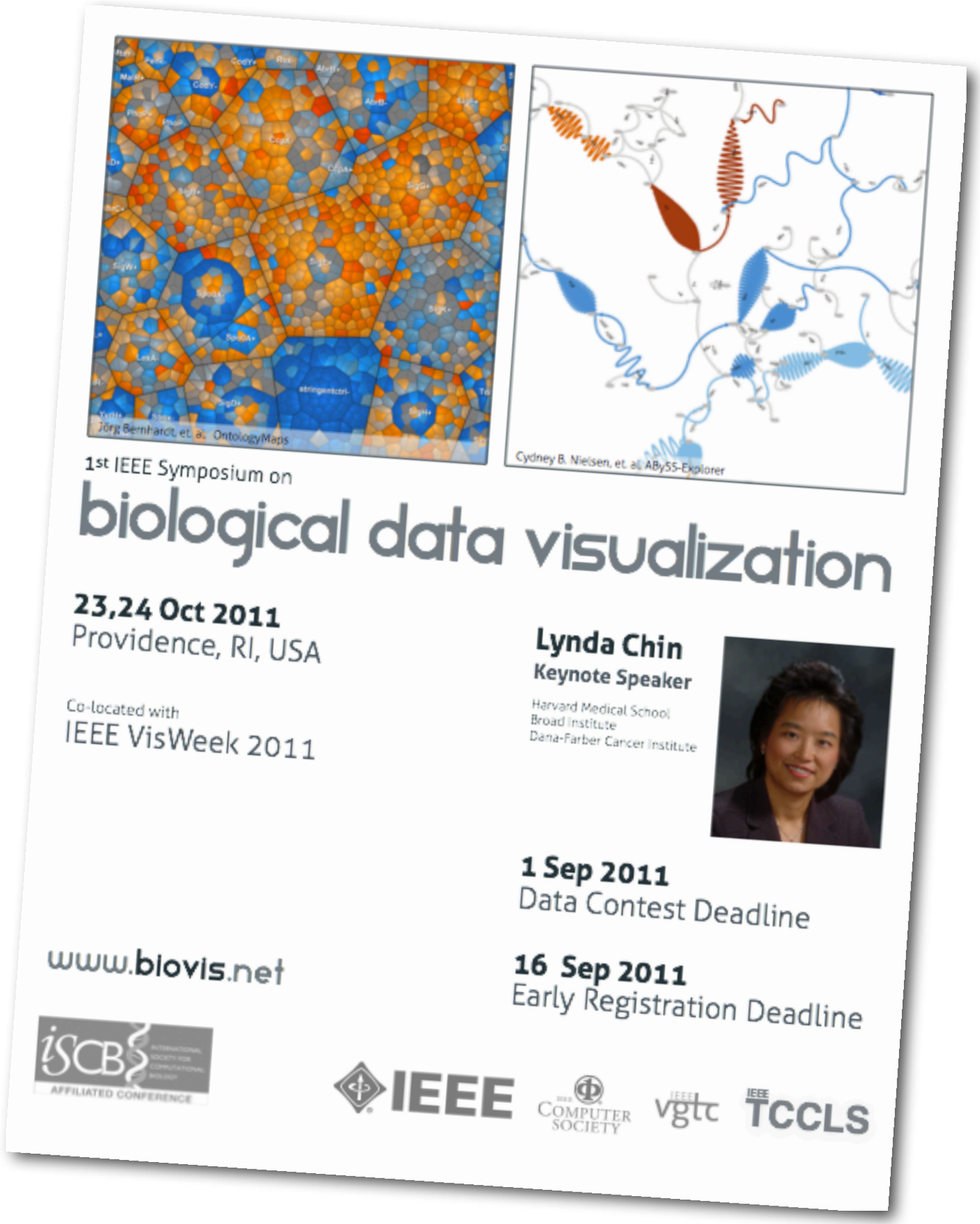
Nils Gehlenborg (nils@hms.harvard.edu)

Center for Biomedical Informatics / Harvard Medical School
Cancer Program / Broad Institute of MIT and Harvard

ISMB/ECCB 2011

http://www.biovis.net

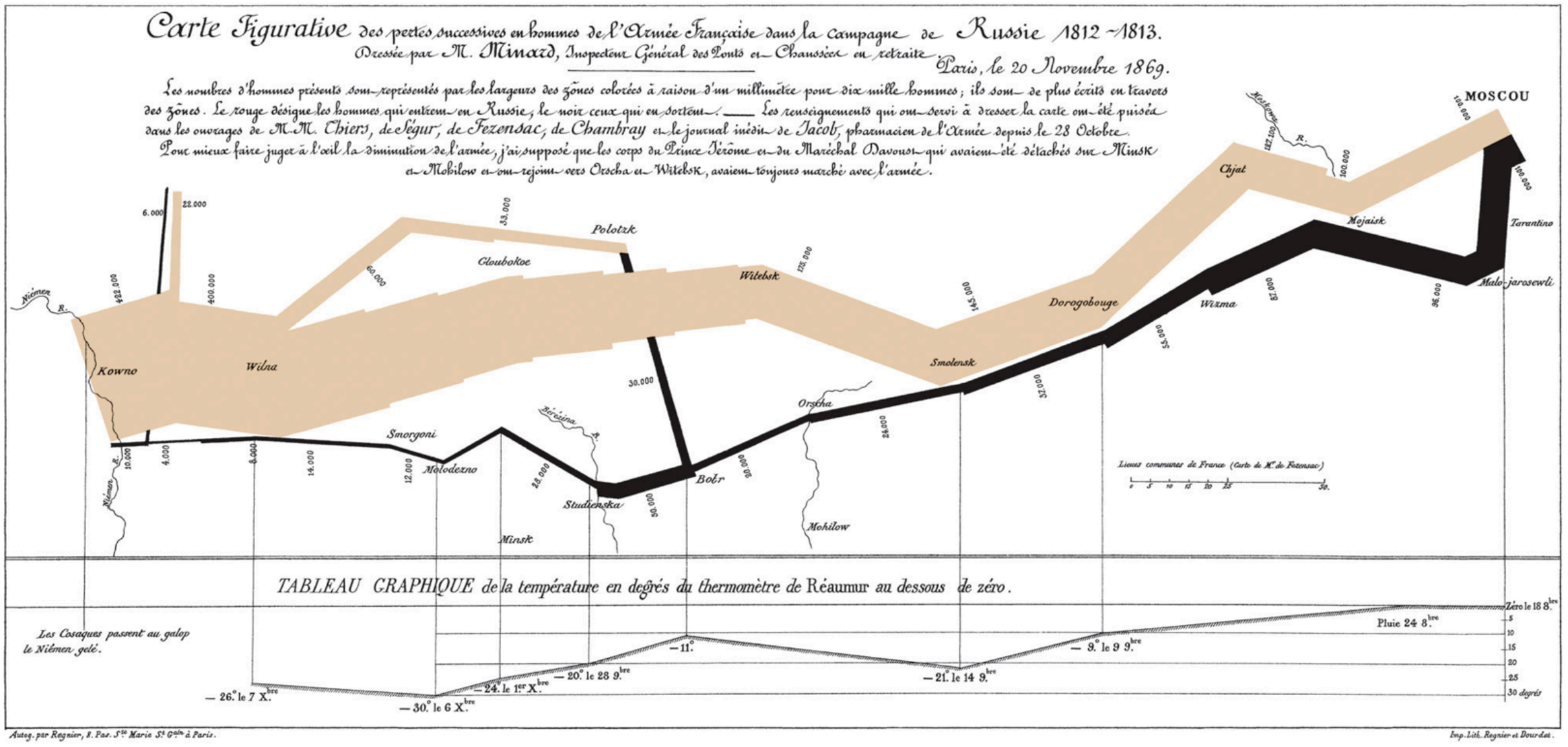
Flyers at ISCB booth!



A good sketch is better than a long speech.

Napoleon Bonaparte

Napoleon's March on Moscow



I believe when I see it.

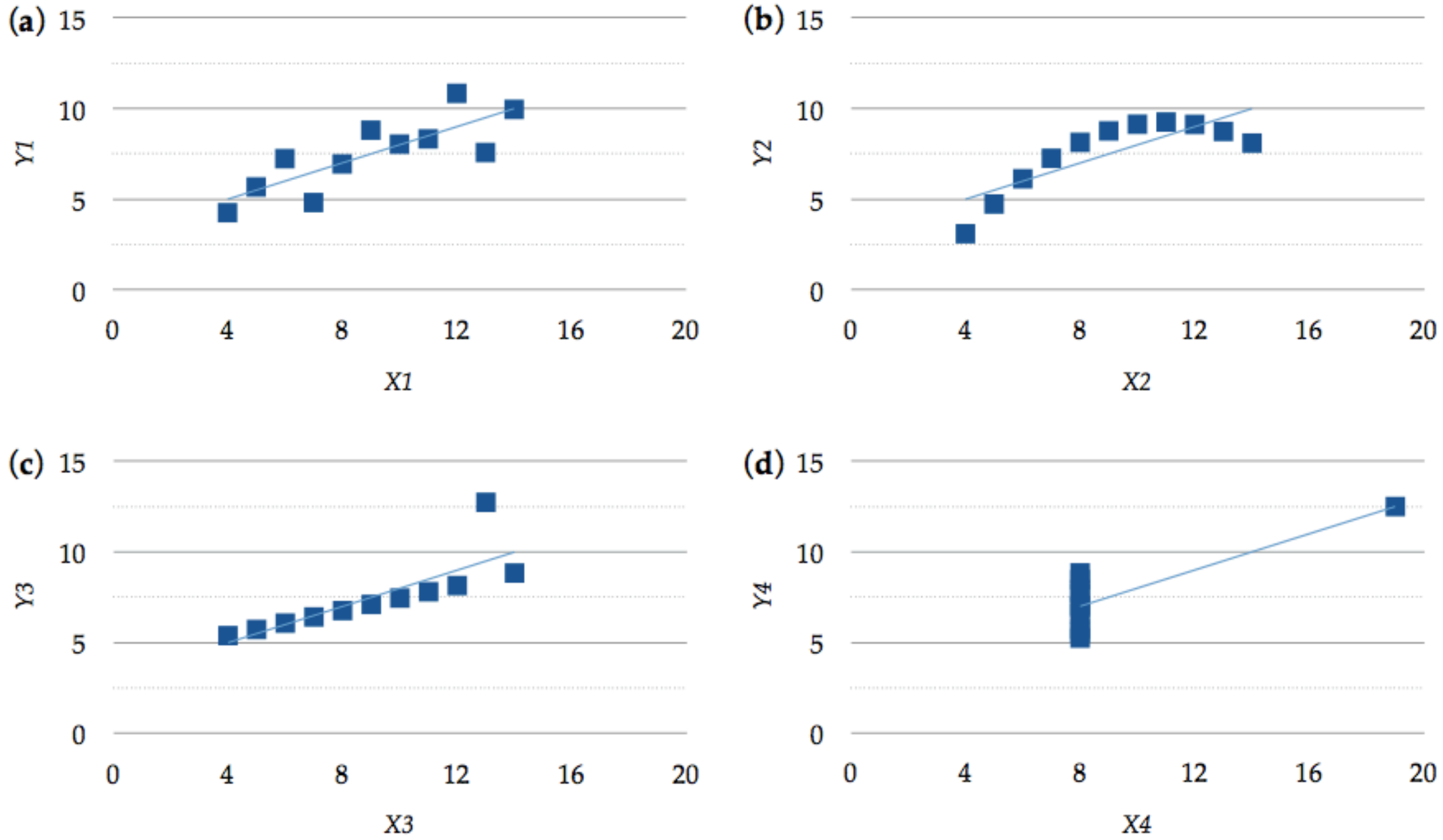
Unknown

Anscombe's Quartet

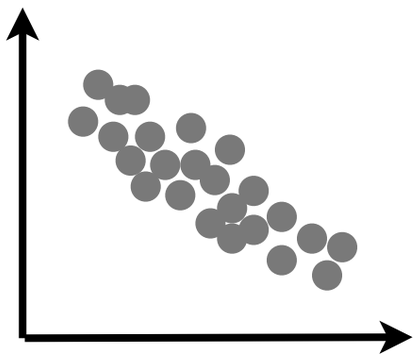
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

mean(X) = 9, var(X) = 11, mean(Y) = 7.5, var(Y) = 4.12,
cor(X,Y) = 0.816, linear regression line Y = 3 + 0.5*X

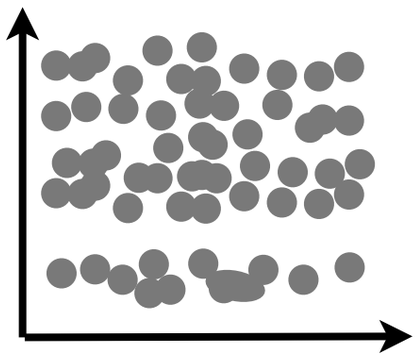
Anscombe's Quartet



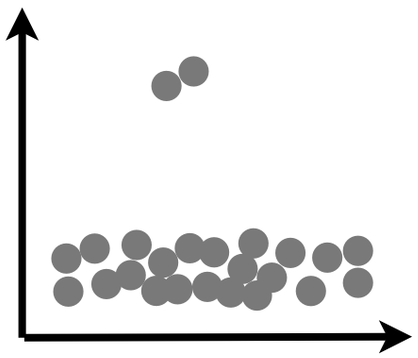
Exploration: Hypothesis Generation



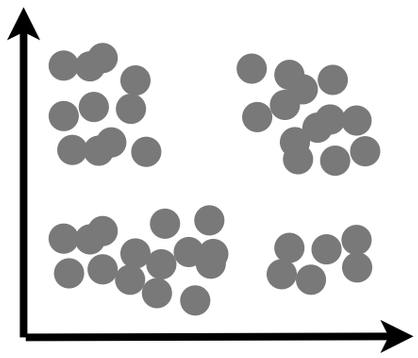
trends



gaps



outliers



clusters

- A large data set is given and the goal is to learn something about it.

- Visualization is employed to perform pattern detection using the human visual system.

- The goal is to generate hypotheses that can be tested with statistical methods or follow-up experiments.

Visualization Use Cases

Presentation

Confirmation

Exploration

Definition

The use of computer-supported, interactive, visual representations of data to amplify cognition.

Stu Card, Jock Mackinlay & Ben Shneiderman

Computer-based visualization systems provide visual representations of datasets intended to help people carry out some task more effectively.

Tamara Munzner

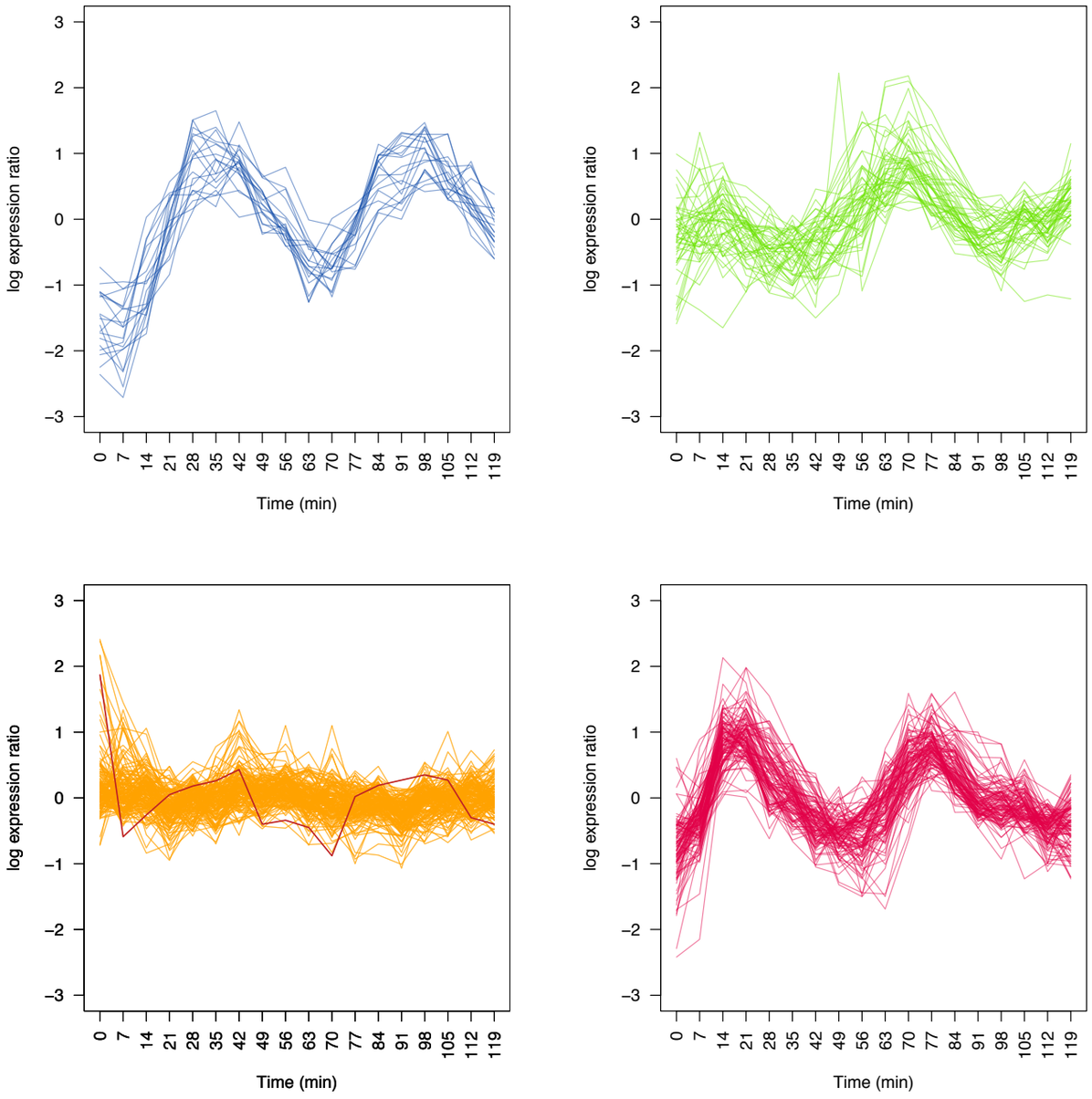
Tasks

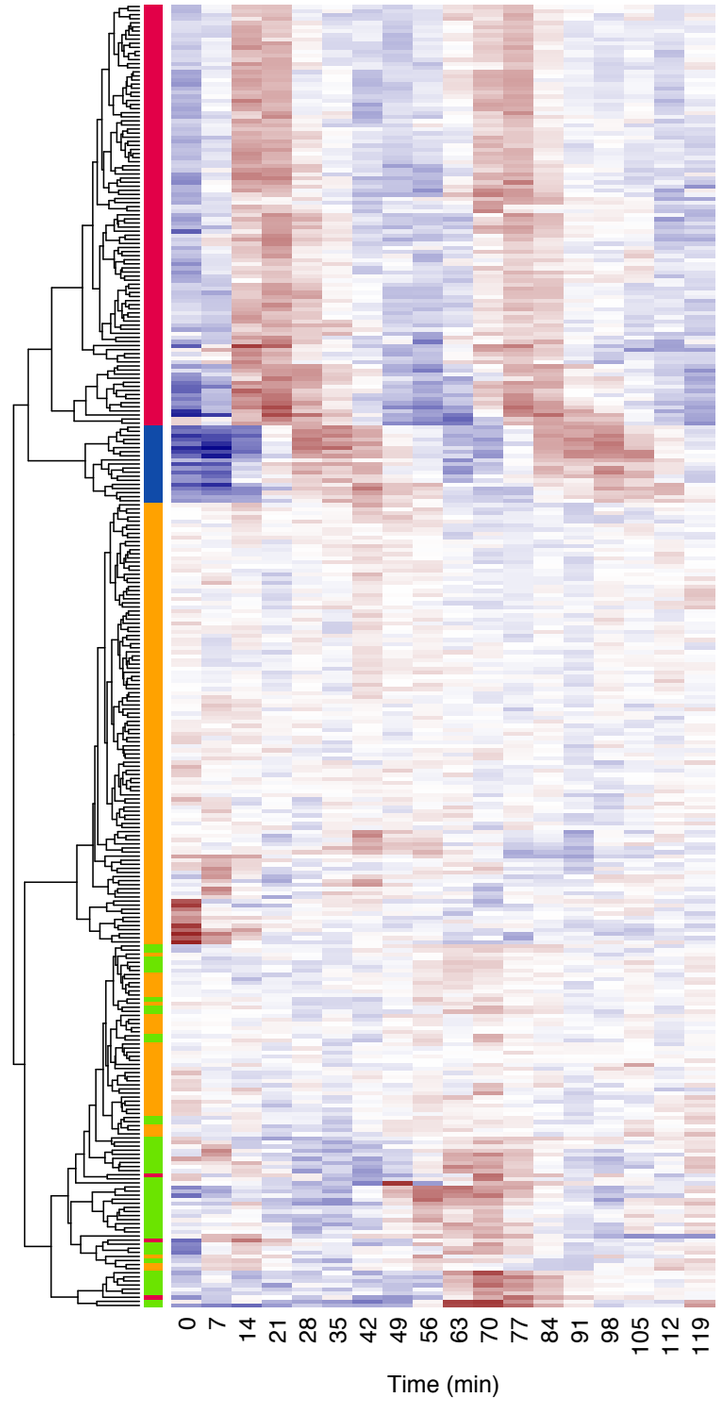
(Rule #1 - Know Your Users: User-centered design)

Tasks in Gene Expression Analysis

- Task 1: Determine expression level of a given gene in a given sample (*retrieve value*)
    - Visualization: Provide the whole profile as context for the particular measurement.
- Task 2: Determine the range of expression levels in a given profile and how much they vary across the profile (*extrema, range, characterize distribution*)
    - Present profile so that the range and distribution of expression levels can be evaluated efficiently.

Tasks in Gene Expression Analysis





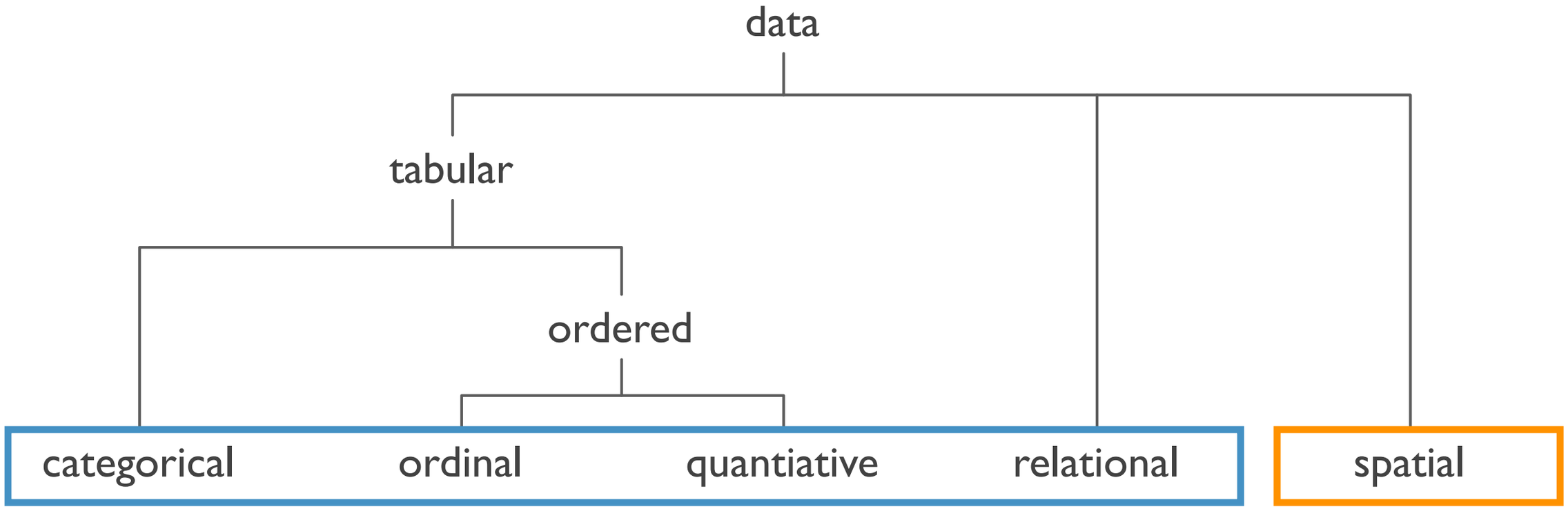
Visual Representation

Data Types

data

apples	small	10 inches	trees	intrinsic
oranges	medium	13 inches	networks	position
bananas	large	18.5 inches		

Data Types


data
tabular
ordered
categorical
ordinal
quantiative
relational
spatial
abstract

Marks: Geometric Primitives

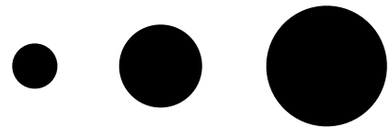
points

lines

areas

Visual Channels: Appearance of Marks

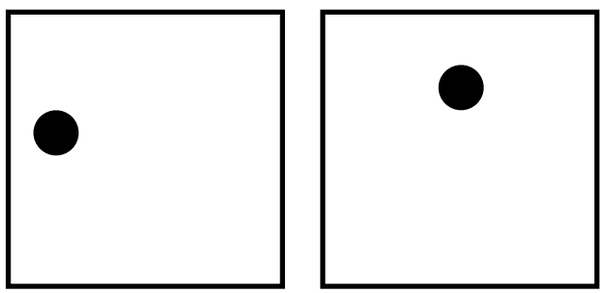
size



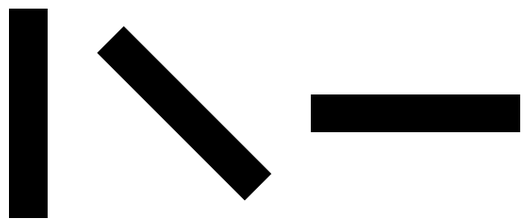
color

shape

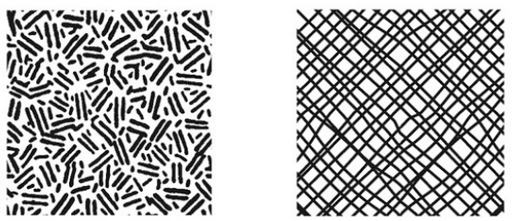
position



angle

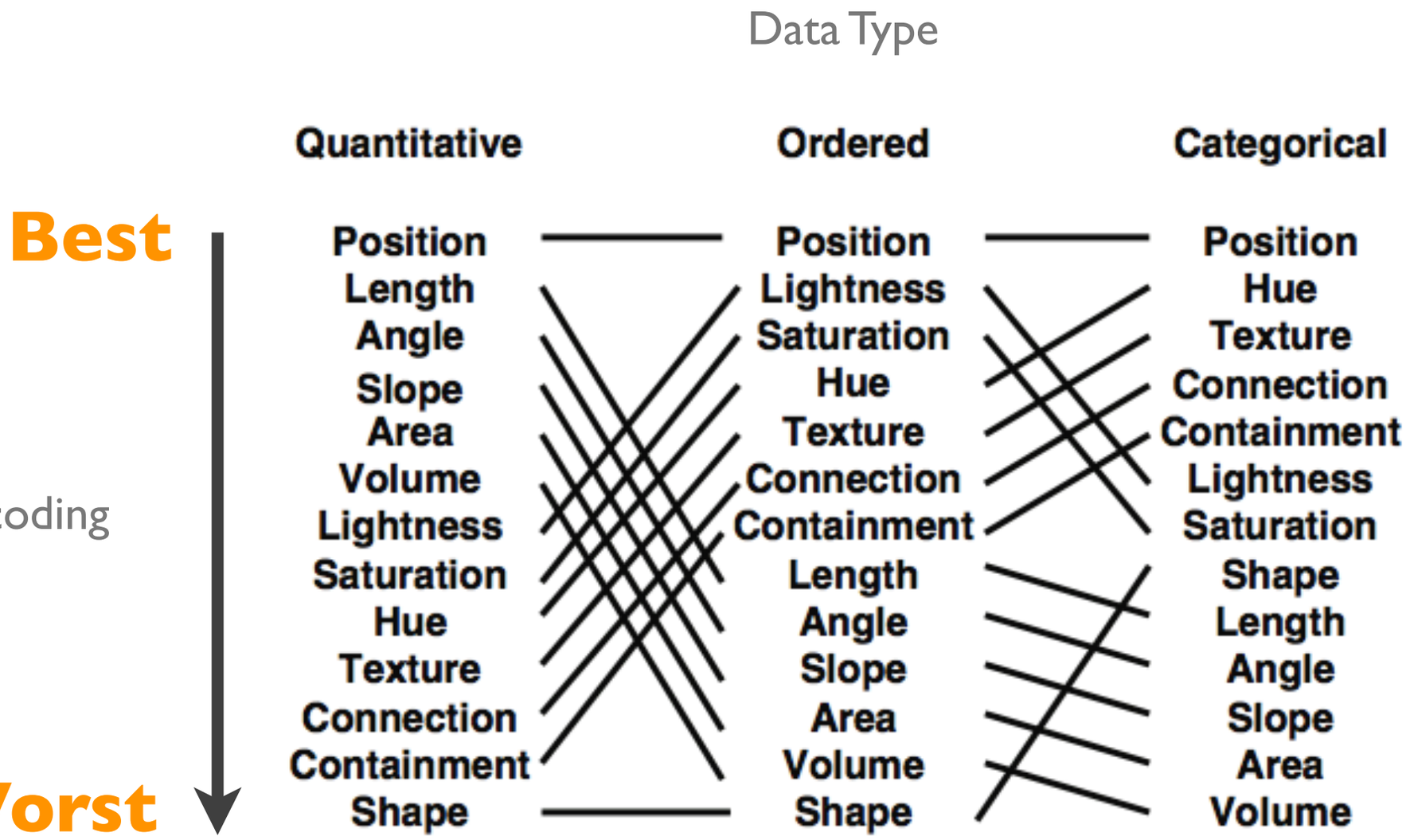


texture

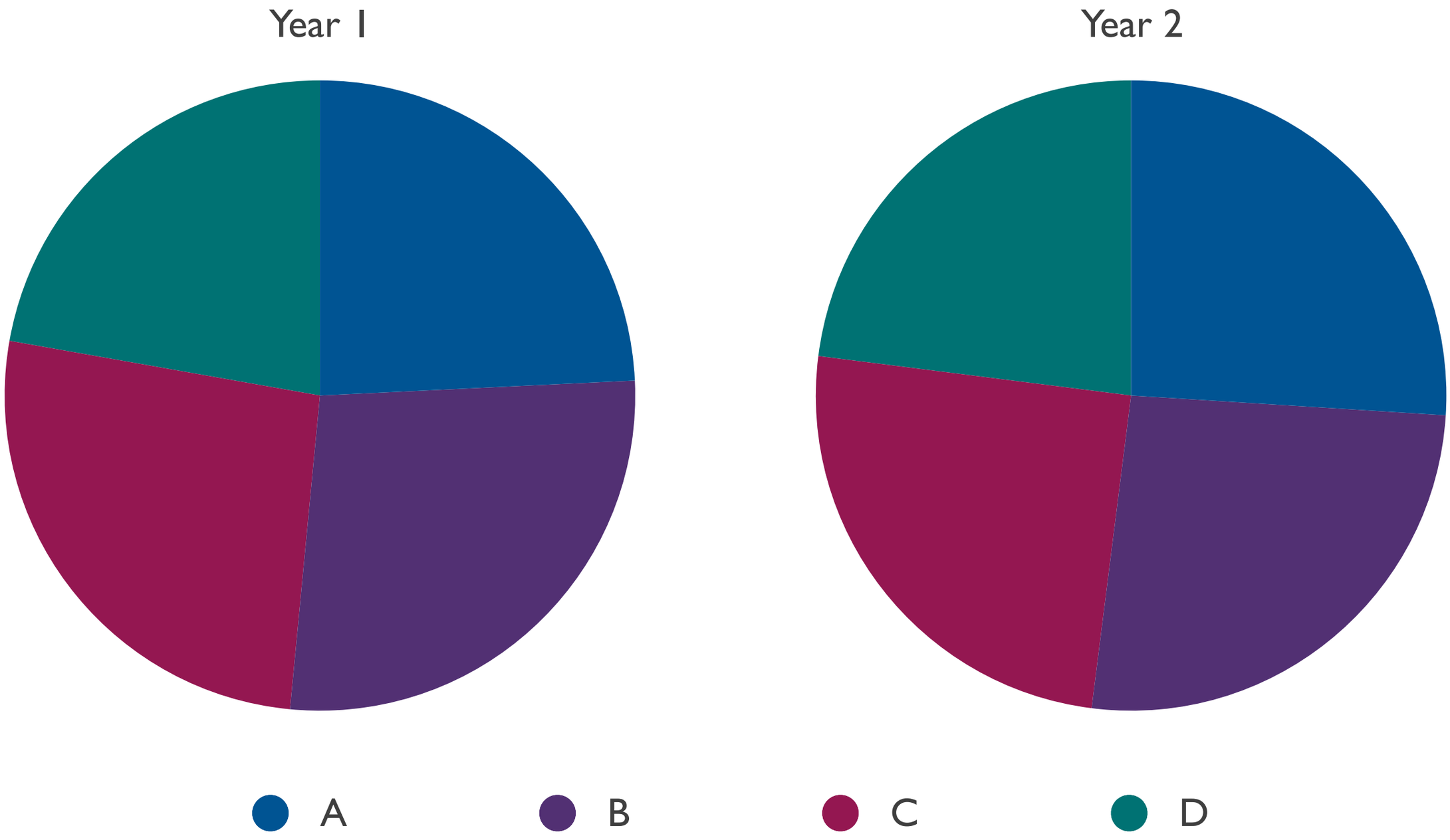


Munzner 2009, in *Fundamentals of Computer Graphics (redrawn from Mackinlay 1986)*

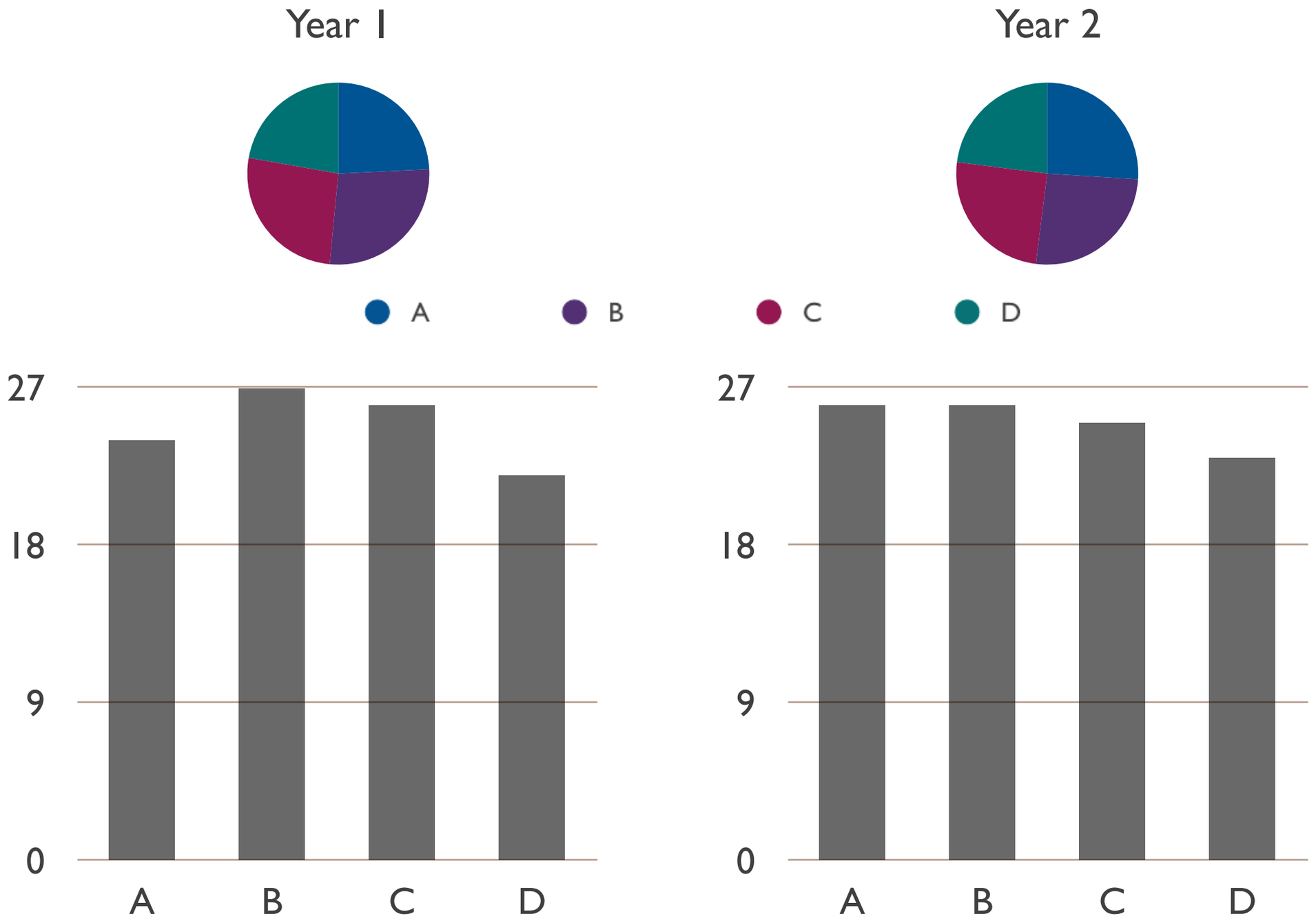
Ranking of Encodings



Rankings in Action



Rankings in Action



Ranking of Encodings

- How accurately can the data be read from the visualization?
- How many classes can be distinguished?
- Can the channels be separated from each other?
- Which channels are processed preattentively?

Principle of Importance Ordering (Mackinaly 1986):

Encode more important information more effectively.

Interaction

Shneiderman 1996, in Proceedings IEEE Symposium on Visual Languages

Information Seeking Mantra

- In explorative settings the user is normally dealing with large amounts of data.

- Impossible to grasp everything at once.

- Solution: Make visualizations **interactive** to support the user in exploring subsets of the data at different resolutions.

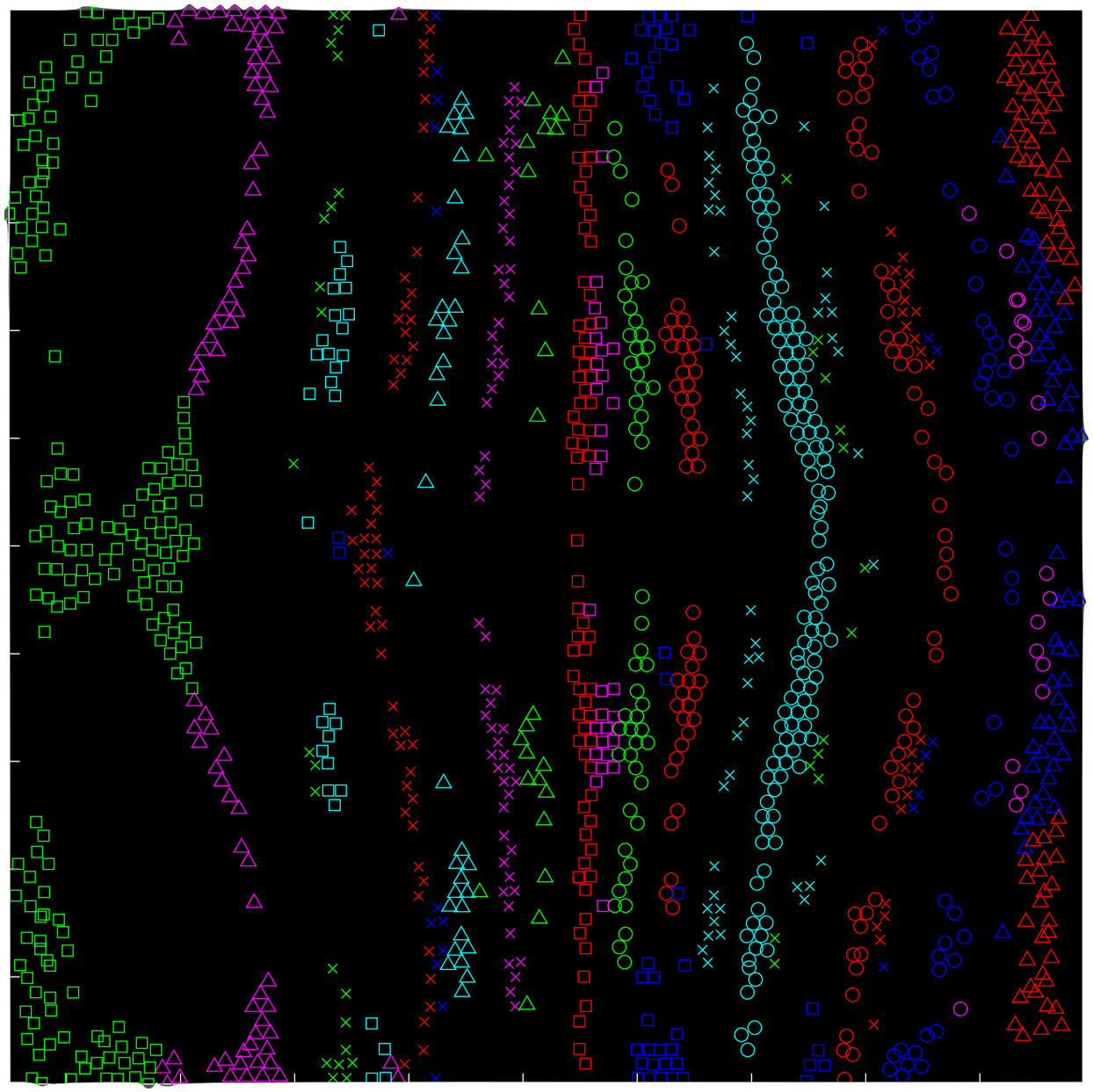
- Ben Shneiderman's Information Seeking Mantra:

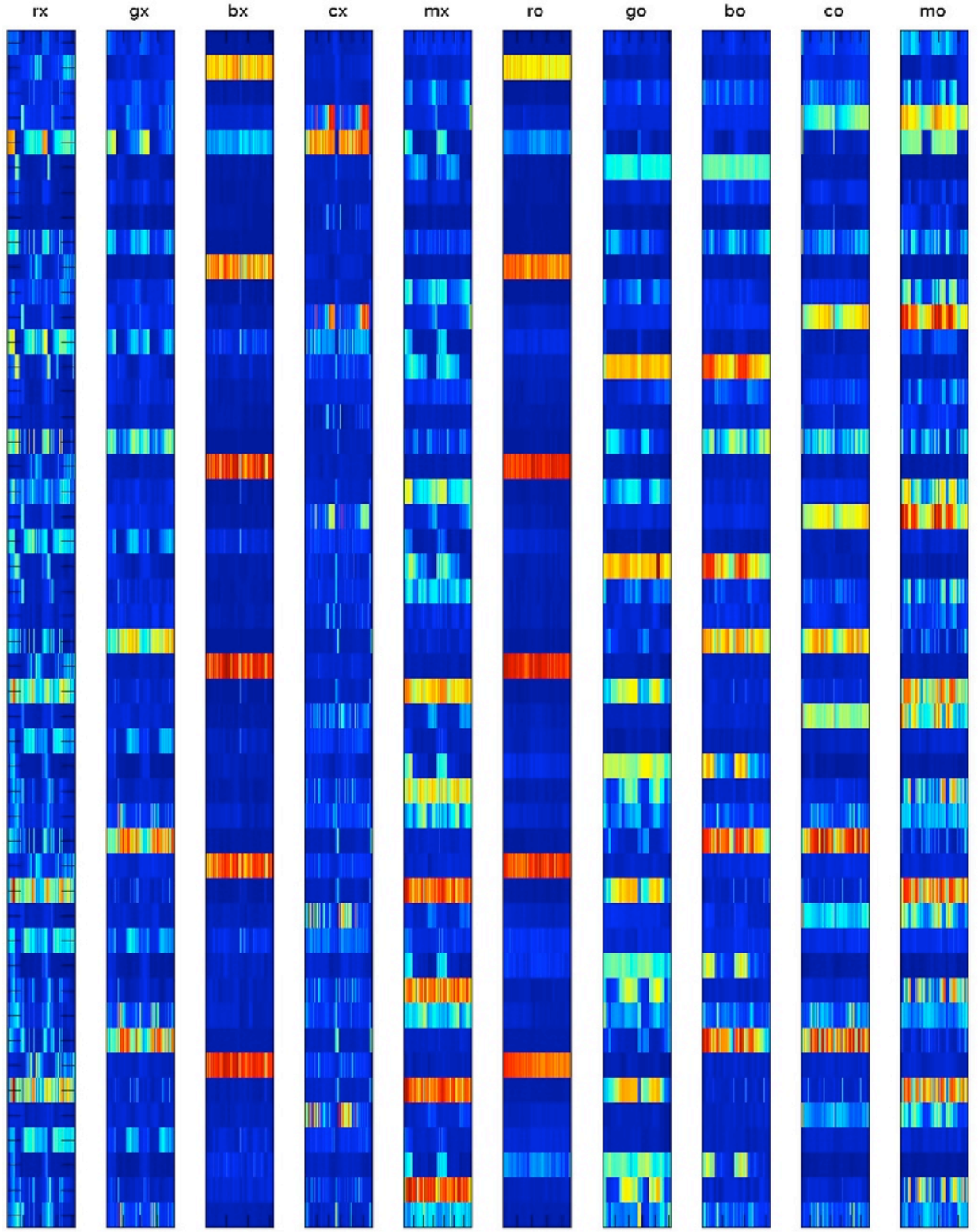
Overview first, zoom and filter, then details on demand.

Linked Views

- beyond static views, **multiple linked views**
- "allow the user to have a dialog with the data"
- technique that allows for data exploration
- interactive, multiple views of the data

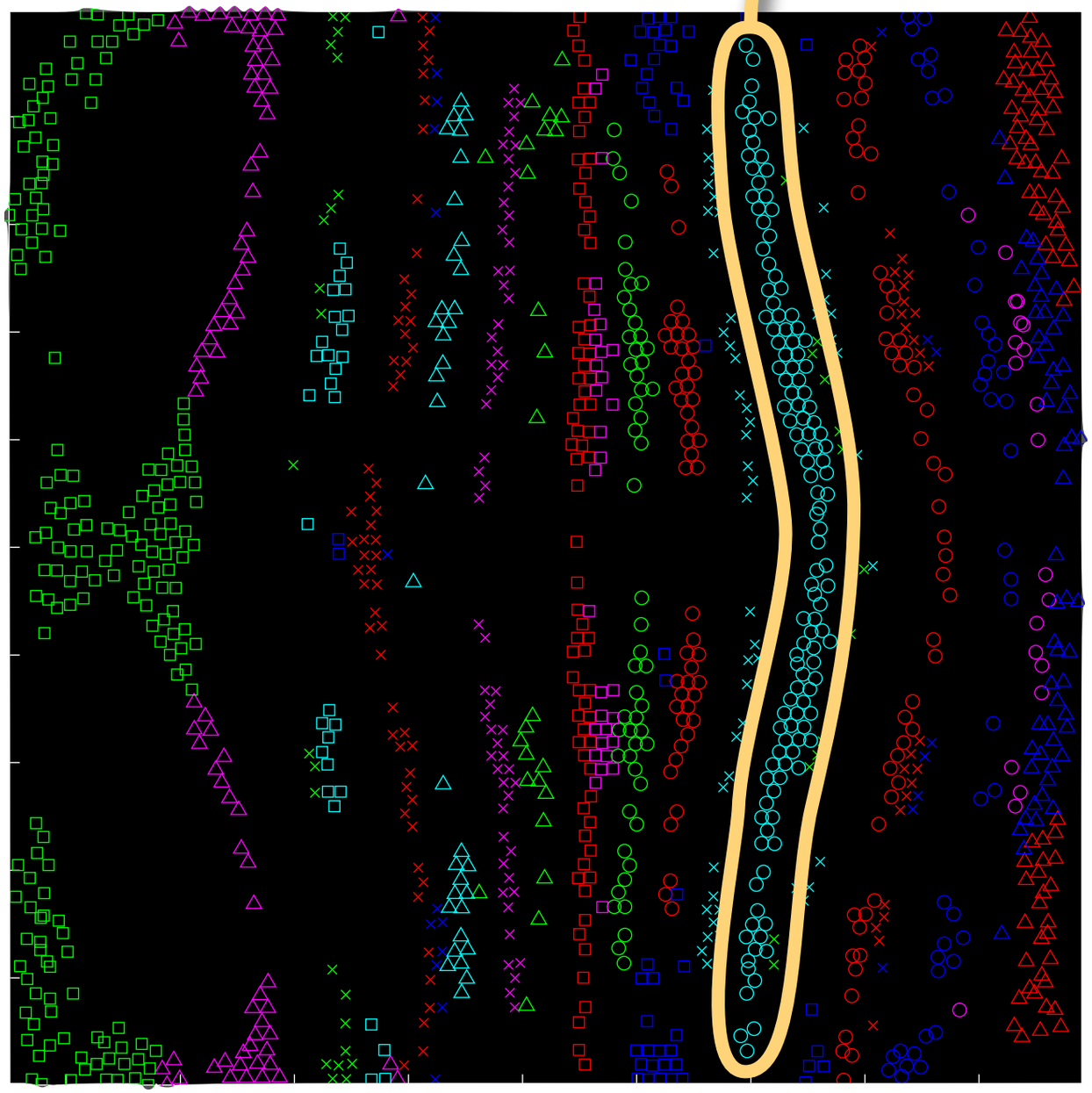
Slide from Miriah Meyer @ Utah

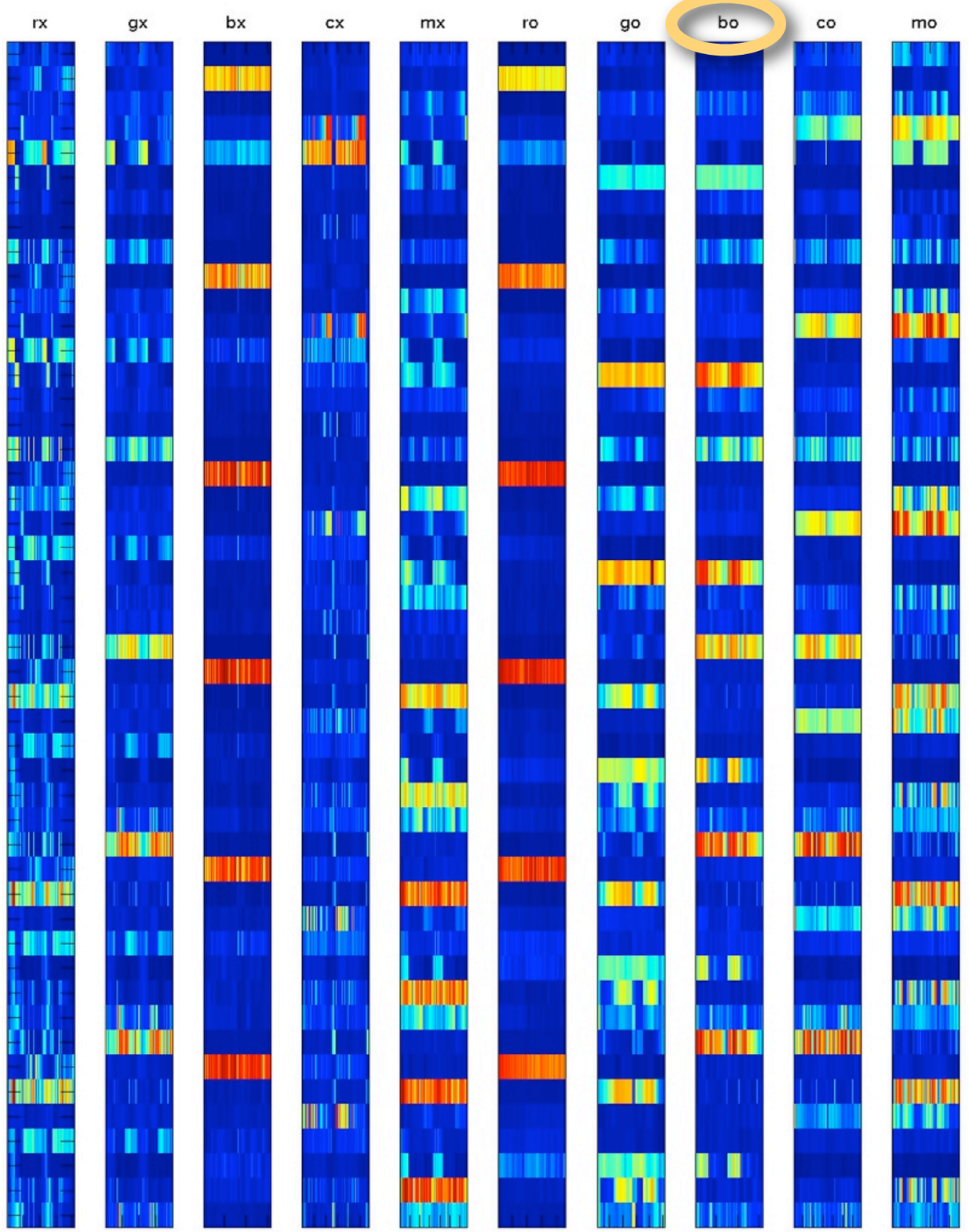




images courtesy of Angela DePace and Charles Fowlkes

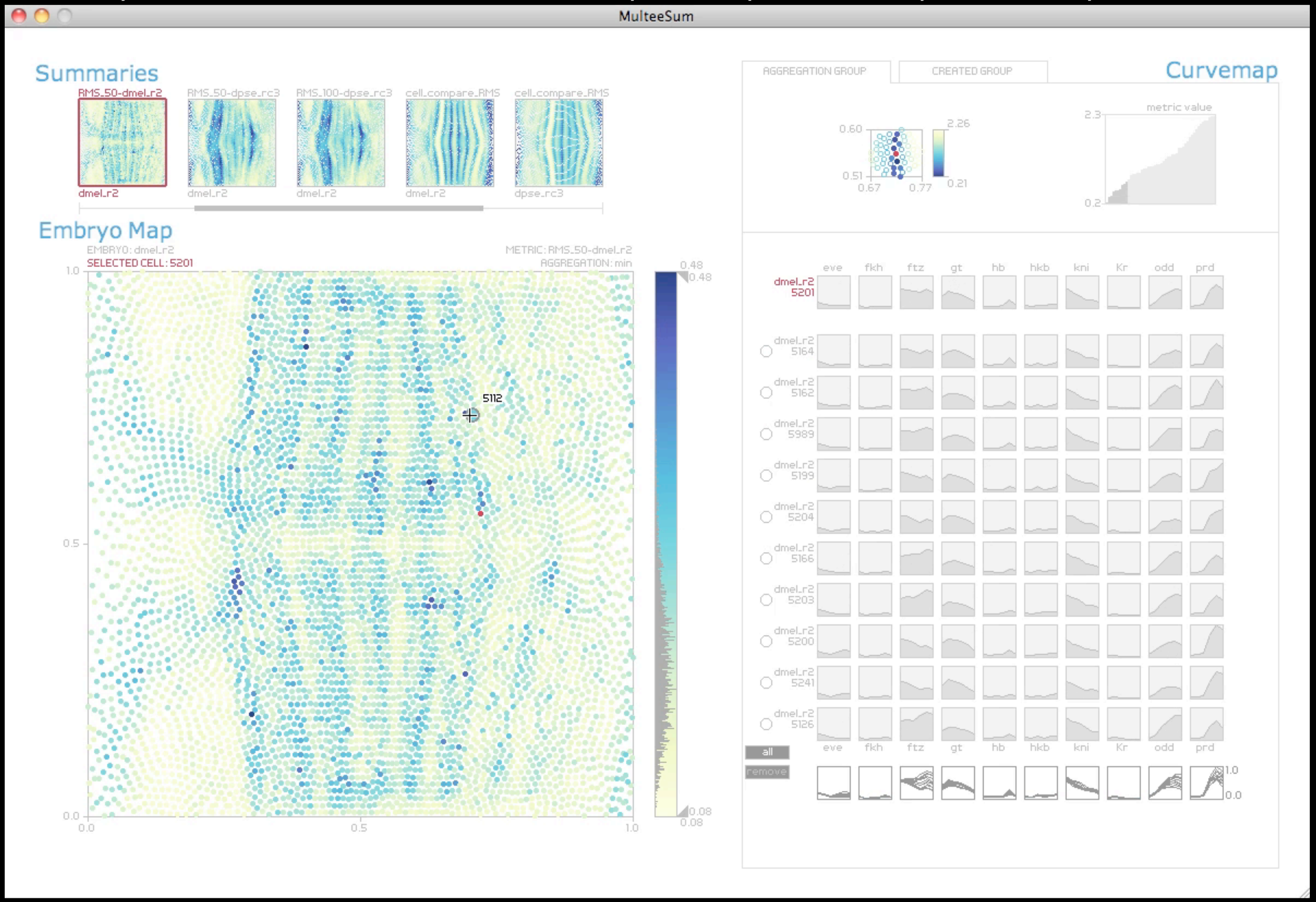
Slide from Miriah Meyer @ Utah





images courtesy of Angela DePace and Charles Fowlkes

Meyer et al. 2010, MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data



Implementation

Considerations

Rendering: Render loop or event-driven?

Platform: performance, deployment, UI support, libraries, plugin vs application

Backend: local or remote, database or files, web service, HTTP/FTP, in memory, etc.

Data summaries: precompute or compute on the fly?

Platforms and APIs

Java: Java 2D, Java 3D, JOGL (OpenGL)

JavaScript: SVG, HTML5 Canvas, WebGL (OpenGL "light")

Flash + Actionscript: hardware accelerated rendering in Version 11

others: deployment? UI support? cross-platform compatibility?

Visualization Toolkits for the Web

- Java applets: **Processing**, Prefuse*

- Flash: Flare*

- JavaScript

  - SVG: Google Chart Tools*, Flot*, **ProtoVis/D3.js***, Raphael, TheJIT

  - HTML5 Canvas: Three.js, ProcessingJS

  - WebGL: Three.js, PhiloGL

** indicates high-level visualization library*

Take Home Message

Carefully analyze the tasks that need to be supported.

Make informed decisions about your visual encodings.

Use interaction and multi-scale approaches to get a handle on the data size.

Choose your platform wisely.

Visualization is science, not art.

Acknowledgements: Slides & Ideas

Miriah Meyer (University of Utah)

Tamara Munzner (University of British Columbia)

Resources

Scientific and Information Visualization

- Scientific Visualization (“scivis”) and Information Visualization (“infovis”) are very ill-defined terms

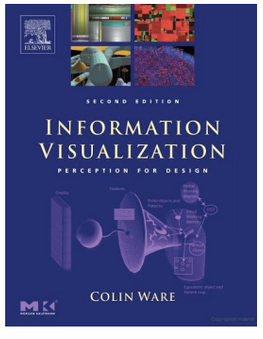
- Scientific Visualization is often used to describe visualization of data that is intrinsically spatial (such as medical imaging data, fluid flows or protein structures)

- Information Visualization is typically used to describe visualization of abstract data (such as gene expression data or interaction networks)

- there is plenty of overlap and the separation is quite arbitrary

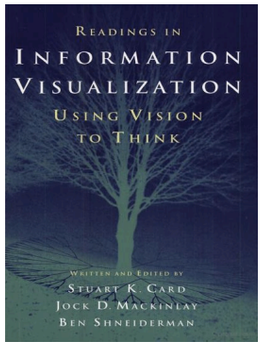
- both Scientific and Information Visualization are used to visualize scientific data

Recommended Books



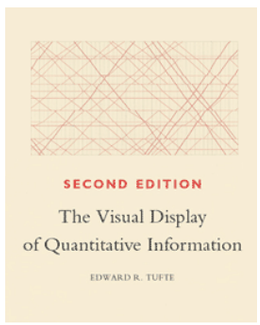
Information Visualization - Perception for Design

Colin Ware, Morgan Kaufmann, 2004



Information Visualization - Using Vision to Think

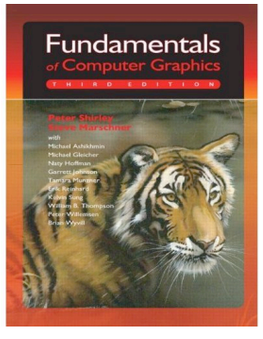
Stuart K Card, Jock D Mackinlay, Ben Shneiderman, Morgan Kaufmann, 1999



The Visual Display of Quantitative Information (2nd Edition)

Edward R Tufte, Graphics Press, 2001

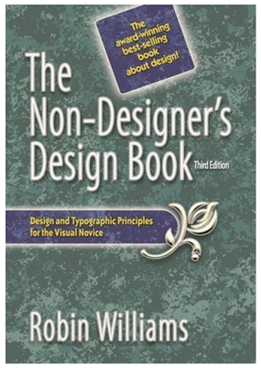
Recommended Books



Fundamentals of Computer Graphics (3rd Edition)

Peter Shirley, Steve Marschner, AK Peters Publishers, 2009

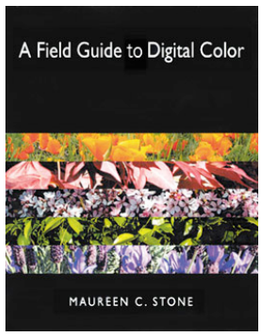
(in particular: "Chapter 27 - Visualization", also as free PDF from Tamara Munzner's website)



The Non-Designer's Design Book (3rd Edition)

Robin Williams, Peachpit Press, 2008

Recommended Resources on Color



A Field Guide to Digital Color

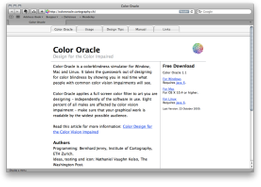
Maureen C Stone, AK Peters Publishers, 2003

ColorBrewer 2.0

Cynthia Brewer, Mark Harrower, http://www.colorbrewer2.org

VisCheck

http://www.vischeck.com



Color Oracle

http://colororacle.cartography.ch

Recommended Journals

Nature Methods *Special Issue on Visualizing Biological Data*
http://www.nature.com/nmeth/journal/v7/n3s
Nature Methods *Points of View* column by Bang Wong
http://bang.clearscience.info/?p=546

IEEE Transactions on Visualization and Computer Graphics
http://www.computer.org/portal/web/tvcg

IEEE Computer Graphics and Applications
http://www.computer.org/portal/web/cga/home

Recommended Meetings

IEEE Symposium on Biological Data Visualization - BioVis
http://www.biovis.net

Workshop on Visualizing Biological Data - VIZBI
http://www.vizbi.org

IEEE VisWeek with InfoVis, Vis and VAST Conferences
http://www.visweek.org

Tools for Interaction Network Visualization

Name	*Cost*	*Availability*	*Description*	*URL*
Stand-alone				
Arena 3D	Free	Win Mac Linux	Visualization of biological multi-layer networks in 3D	http://www.arena3d.org
BiNA	Free	Win Mac Linux	Exploration and interactive visualization of pathways	http://www.bnplusplus.org/bina
BioLayout Express 3D	Free	Win Mac Linux	Generation and cluster analysis of networks with 2D/3D visualization	http://www.biolayout.org
BiologicalNetworks 2	Free	Win Mac Linux	Analysis suite; visualizes networks and heat map; maps abundance data	http://www.biologicalnetworks.org
Cytoscape	Free	Win Mac Linux	Network analysis; extensive list of plug-ins for advanced visualization	http://www.cytoscape.org
GENeVis	Free	Win Mac Linux	Network and pathway visualization; abundance data	http://tinyurl.com/genevis
Medusa	Free	Win Mac Linux	Basic network visualization tool	http://coot.embl.de/medusa
NBrowse	Free	Win Mac Linux	Network visualization software for heterogeneous interaction data	http://www.gnetbrowse.org
NAViGaTOR	Free	Win Mac Linux	Visualization of large protein-protein interaction data sets; abundance data	http://tinyurl.com/navigator1
Ondex	Free	Win Mac Linux	Integrative workbench; large network visualizations; abundance data	http://www.ondex.org
Osprey	Free	Win Mac Linux	Tool for visualization of interaction networks	http://tinyurl.com/osprey1
Pajek	Free	Win	Generic network visualization and analysis tool	http://pajek.imfm.si
ProViz	Free	Win Mac Linux	Software for visualization and exploration of interaction networks	http://tinyurl.com/proviz
SpectralNET	Free	Win	Network visualizations; scatter plots for dimensionality reduction methods	http://tinyurl.com/spectralnet
Tulip	Free	Win Mac Linux	Generic visualization and analysis tool; extremely large networks; 3D support	http://tulip.labri.fr/TulipDrupal
VANTED	Free	Win Mac Linux	Combined visualization of abundance data and pathways	http://tinyurl.com/vanted
yEd	Free	Win Mac Linux	Generic network visualization software; offers many layout algorithms.	http://tinyurl.com/yEdGraph
Cytoscape Plug-ins				
BiNoM	Free	Win Mac Linux	Extensive support for common systems biology network formats	http://tinyurl.com/binom1
BioModules	Free	Win Mac Linux	Detects modules in networks; maps abundance data onto nodes and modules	http://tinyurl.com/biomodules
Cerebral	Free	Win Mac Linux	Biologically motivated layout algorithm; maps abundance data; clustering	http://tinyurl.com/cerebral1
MCODE	Free	Win Mac Linux	Network clustering algorithm; support for manual cluster refinement	http://preview.tinyurl.com/MCODE123
VistaClara	Free	Win Mac Linux	Mapping of abundance data to nodes and "heat strips"; provides heat map	http://www.cytoscape.org/plugins
Web-based				
Graphle	Free		Distributed client/server network exploration and visualization tool	http://tinyurl.com/graphle
Lichen	Free		Library for web-based visualization of network and abundance matrix data	http://tinyurl.com/Lichen1
MAGGIE Data Viewer	Free		Visualization of networks; abundance data in heat maps and profile plots	http://maggie.systemsbiology.net
STITCH 2	Free		Construction and visualization of networks from a wide range of sources	http://stitch.embl.de
VisANT	Free	Win Mac Linux	Analysis, mining and visualization of pathways and integrated omics data	http://visant.bu.edu

Tools for Pathway Visualization

Name	*Cost*	*Availability*	*Description*	*URL*
Stand-alone				
BioTapestry	Free	Win Mac Linux	Visualization of genetic regulatory networks, also with experimental data.	http://www.biotapestry.org
Caleydo	Free	Win Linux	Interactive framework for pathway and expression data; 3D "bucket" view	http://www.caleydo.org
CellDesigner	Free	Win Mac Linux	Drawing and simulation of pathways and models, supports SBGN	http://www.celldesigner.org
Edinburgh Pathway Editor	Free	Win Mac Linux	Construction and visualization of pathway diagrams, supports SBGN	http://tinyurl.com/EdinburghPE
GenMAPP 2	Free	Win	Pathway visualization and construction; abundance data	http://www.genmapp.org
IngenuityPathways	$	Win Mac Linux	Full analysis suite; network and pathway visualizations; abundance data.	http://tinyurl.com/IngenuityPath
JDesigner	Free	Win	Drawing and simulation of pathways and models	http://tinyurl.com/jdesigner
KaPPA View	Free	Win	Analysis and visualization of plant pathways and mapped abundance data	http://tinyurl.com/kappa-view
KEGG Atlas	Free	Win Mac Linux	Visualization of abundance data on interactive KEGG pathways	http://www.genome.jp/kegg
MetaCore	$	Win Mac Linux	Pathway, network and omics data analysis and visualization suite	http://www.genego.com
PathVisio	Free	Win Mac Linux	Visualization and editing pathways, supports mapping of omics data	http://www.pathvisio.org
VitaPad	Free	Win Mac Linux	Editing of pathway diagrams, integration of abundance data	http://tinyurl.com/vitapad
Web-based				
ArrayXPath	Free		Mapping of abundance data to pathway visualizations	http://tinyurl.com/ArrayXPath
GEPA	Free		Analysis suite; visualization of transcriptomics data on pathways maps	http://tinyurl.com/GEPAT1
iPath	Free		Visualization and exploration of combined KEGG pathways	http://pathways.embl.de
MapMan	Free		Application that visualizes abundance data on metabolic pathways	http://tinyurl.com/MapManApp
Omics Viewer	Free		Tool that maps abundance data to BioCyc pathway diagrams	http://www.biocyc.org
Pathway Explorer	Free		Visualization of abundance data on pathways	http://tinyurl.com/pathwayexp
PATIKA	Free		Extensive pathway visualization tool; good support for signaling pathways	http://www.patika.org
Payaologue	Free		Collaborative pathway annotation and visualization tool	http://celldesigner.org/payao
ProMeTra	Free		Maps abundance matrices of multiple omics data types on pathways	http://tinyurl.com/ProMeTra
Reactome SkyPainter	Free		Visualization of overrepresented pathways and reactions from gene lists	http://reactome.org
WikiPathways	Free		Wiki-based, community-driven pathway curation and visualization tool	http://www.wikipathways.org

Tools for Visualization of Multivariate Data

Name	*Cost*	*OS*	*Description*	*URL*
Stand-alone				
BicOverlapper	Free	Win Mac Linux	Visualization of biclusters combined with profile plots and heat maps	http://vis.usal.es/bicoverlapper/
BiGGEsTS	Free	Win Mac Linux	Heat map-based bicluster visualization	http://tinyurl.com/BiGGEsTS
Brain Explorer	Free	Win Mac	Visualization of 3D transcription data in the central nervous system	http://tinyurl.com/brainExplorer
Caryoscope	Free	Win Mac Linux	Abundance data mapped to chromosomal location	http://tinyurl.com/caryoscope
Data Matrix Viewer	Free	Win Mac Linux	Simple profile plot visualization; supports Gaggle	http://gaggle.systemsbiology.net
EXPANDER	Free	Win Linux	Heat maps, scatter plots and profile plots of cluster averages	http://acgt.cs.tau.ac.il/expander
GENESIS	Free	Win Mac Linux	Analysis suite; offers several interactive visualizations	http://genome.tugraz.at
GeneSpring GX	$	Win Mac Linux	Analysis suite; interactive and linked visualizations; also networks	http://tinyurl.com/genespring
GeneVAnD	Free	Win Mac Linux	Linked heat maps, dendrograms and 2D/3D scatter plots	http://tinyurl.com/GeneVAnD
geWorkbench	Free	Win Mac Linux	Modular suite; heat maps, dendrograms, profile and scatter plots	http://tinyurl.com/geWorkbench
Hierarchical Clustering Explorer	Free	Win	Linked heat map, profile and scatter plots; systematic exploration	http://tinyurl.com/HCExplorer
Java TreeView	Free	Win Mac Linux	Linked heat maps, karyoscopes, sequence alignments, scatter plots	http://jtreeview.sourceforge.net
Mayday	Free	Win Mac Linux	Modular suite; many linked visualizations; enhanced heat map113	http://tinyurl.com/maydaywp
MultiExperiment Viewer	Free	Win Mac Linux	Analysis suite; heat maps, dendrograms, profile and scatter plots	http://www.tm4.org
PointCloudXplore	Free	Win Mac Linux	Visualization of 3D transcription data in Drosophila embryos	http://tinyurl.com/PointCloudXplore
Spotfire Functional Genomics	$	Win	Analysis suite; many linked visualizations and exploration tools	http://spotfire.tibco.com
TimeSearcher	Free	Win	Exploration and analysis of time series; advanced profile plots	http://tinyurl.com/timesearcher
R/BioConductor				
Geneplotter	Free	Win Mac Linux	Karyoscope-style plots and other visualizations	http://www.bioconductor.org
Web-based				
ExpressionProfiler	Free		Transcriptomics data analysis suite with basic visualizations	http://tinyurl.com/expresspro
GenePattern	Free		Modular analysis platform; several visualization modules available	http://tinyurl.com/GenePatt