Tutorial: Information Visualization & Visual Analytics

AMW 2018 Summer School
Cali, Colombia

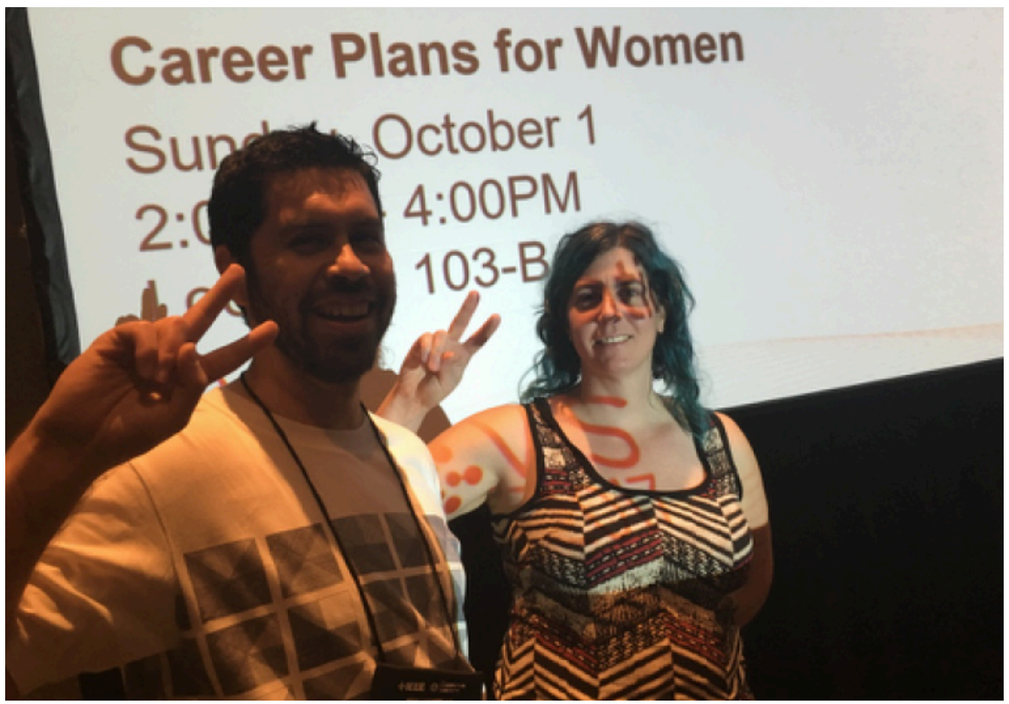
Denis Parra
Assistant Professor PUC Chile

Brief Biography Denis Parra

- PhD in Information Science, University of Pittsburgh (2008-2013)
- Faculty at CS Department, PUC Chile (2013 -)
- Researcher at Milenium Research Institute in Fundamentals of Data
- Leads SocVis Lab http://www.socvis.cl

Credits

- A good portion of this tutorial is based on Tamara Munzner material



Here, very glad with Tamara Munzner at IEEE Vis 2017, since she allowed me to use her material 🤓

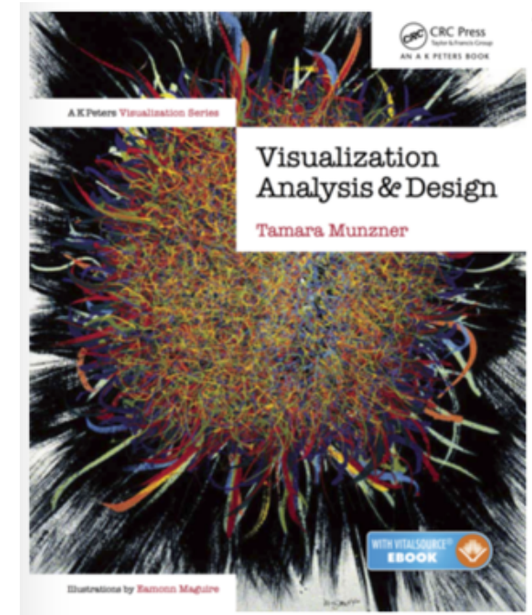
Also, with the support of my InfoVis class TAs

- Nebil Kawas
  nakawas@uc.cl

- Fernando Florenzano
  faflorenzano@uc.cl

Additional Material

- The *backbone is based on T. Munzner's* ***Visualization: Analysis & Design***.
- But it is also complemented with material from:
  - ***The Visual Display of Quantitative Information*** by Edward Tufte,
  - ***Information Visualization: Perception for Design*** by Colin Ware,
  - ***Visualize This*** by Nathan Yau,
  - Several IEEE Vis, ACM CHI and IUI conference papers



Outline

- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - **Introduction: Definitions**
  - Analysis: What, Why, How
  - Marks and Channels

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - Arrange Tables
  - Hands on: Altair
  - Rules of Thumb & Vis Research

What is Information Visualization?

Definition

- *"The use of visual representations of data … to amplify human cognition"*

[Card, Mackinlay, & Shneiderman 1999]

How Information Visualization has been used in the past ?

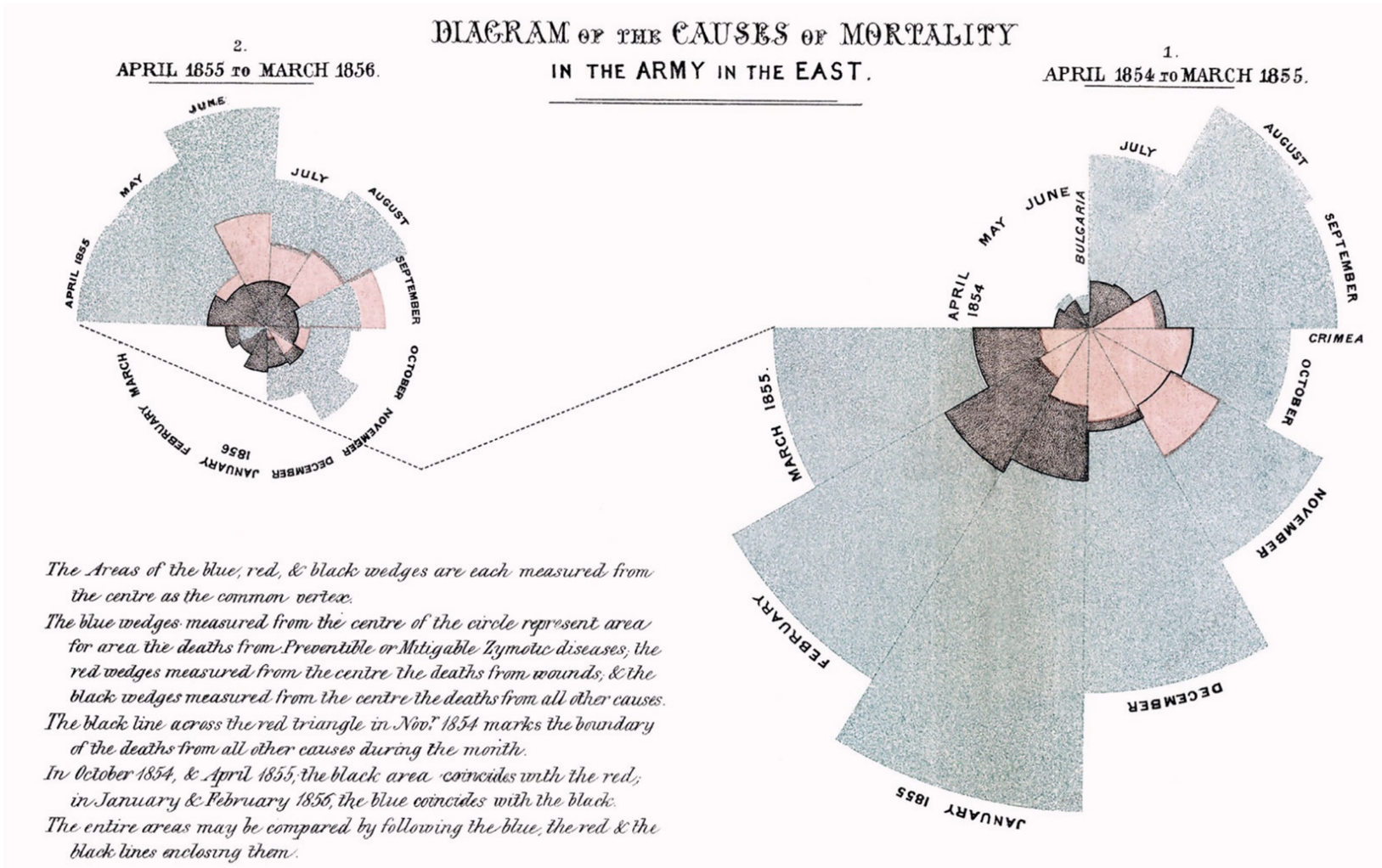
1854: Cholera in London

- Dr. Jon Snow uses spatial analysis to support his hypothesis

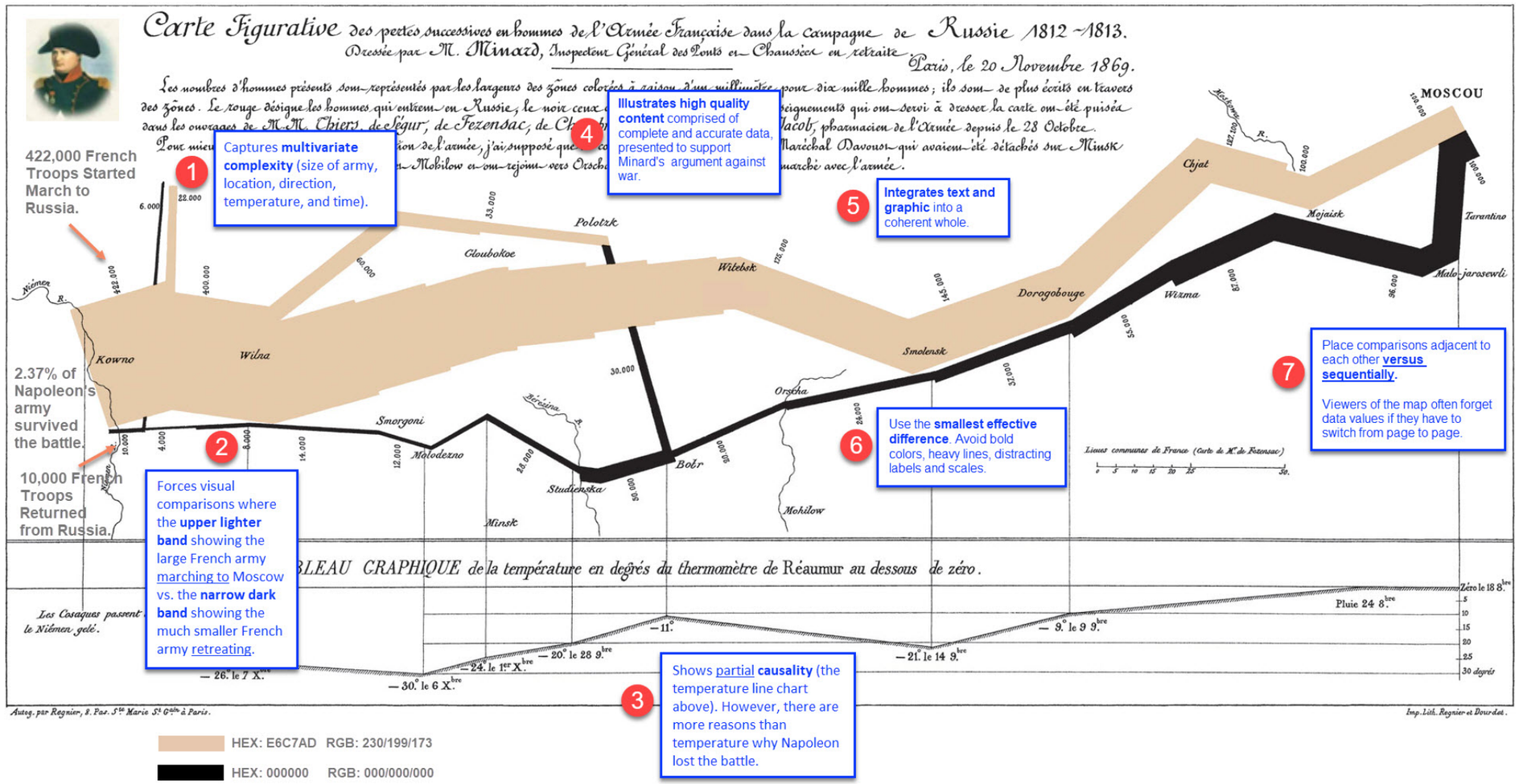
1854: Cholera in London

- … which is proven after a contaminated water pump is closed.

1859 F. Nightingale's Rose (Crimea's War)



1869: Minard's Napoleon Map of Russian Campaign



1. Captures **multivariate complexity** (size of army, location, direction, temperature, and time).
2. Forces visual comparisons where the **upper lighter band** showing the large French army marching to Moscow vs. the **narrow dark band** showing the much smaller French army retreating.
3. Shows partial **causality** (the temperature line chart above). However, there are more reasons than temperature why Napoleon lost the battle.
4. **Illustrates high quality content** comprised of complete and accurate data, presented to support Minard's argument against war.
5. **Integrates text and graphic** into a coherent whole.
6. Use the **smallest effective difference**. Avoid bold colors, heavy lines, distracting labels and scales.
7. Place comparisons adjacent to each other **versus sequentially**.

   Viewers of the map often forget data values if they have to switch from page to page.

https://datavizblog.com/2017/07/11/dataviz-as-history-annotating-minards-napoleon-map/

… but why do we need information visualization in this age ?

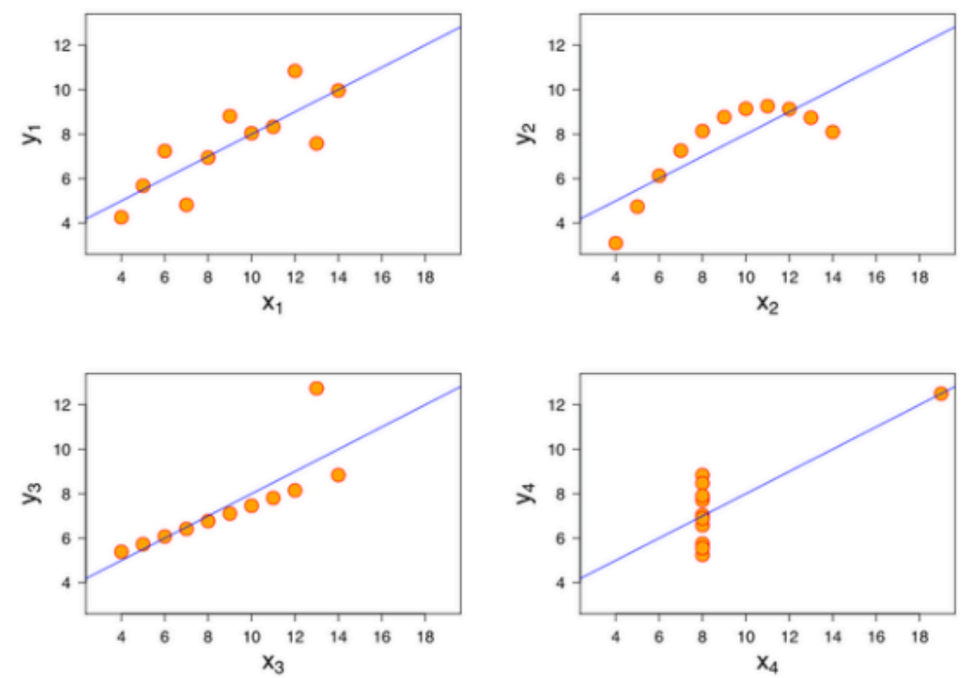
- We have powerful computers, databases and algorithms to crunch large amounts of data (Database and IR research)

- And we also have statistical learning models to make sense of these data (Machine Learning) and make decisions automatically

Anscombe's Quartet (1973)

- Sometimes the models do not tells us the whole story

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places



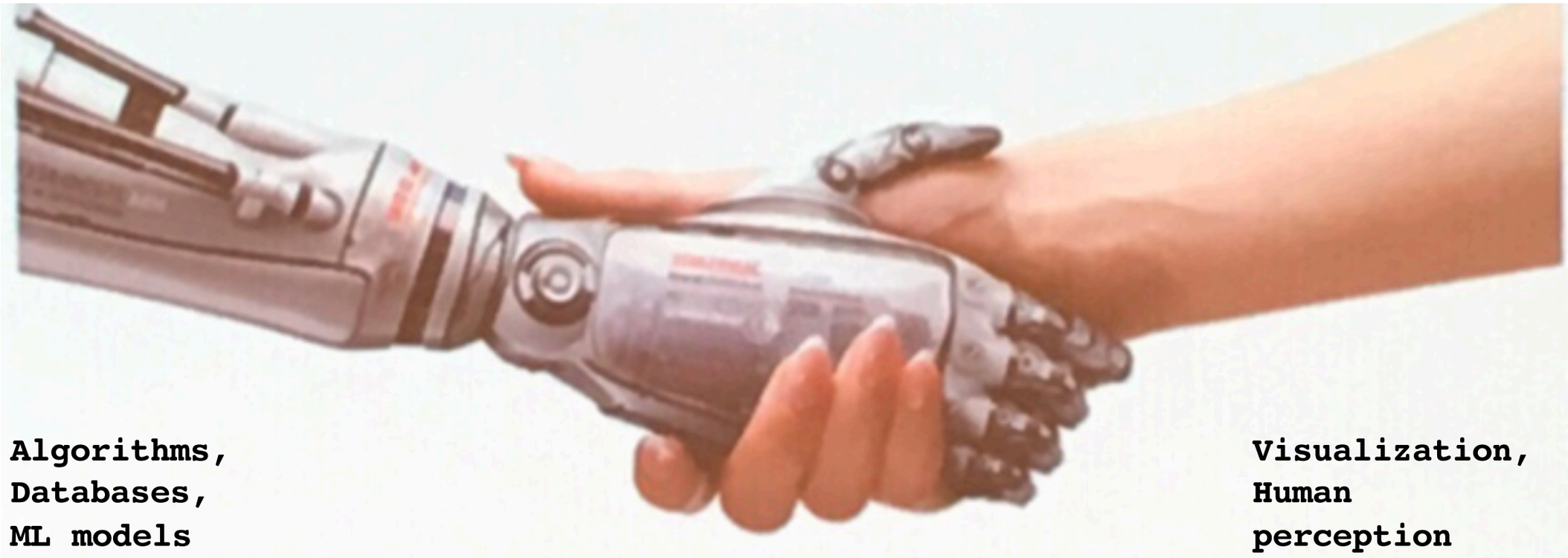
F. J. Anscombe (1973) Graphs in Statistical Analysis, The American Statistician, 27:1, 17-21, DOI: 10.1080/00031305.1973.10478966

Is this a competition then?

Algorithms,
Databases,
ML models

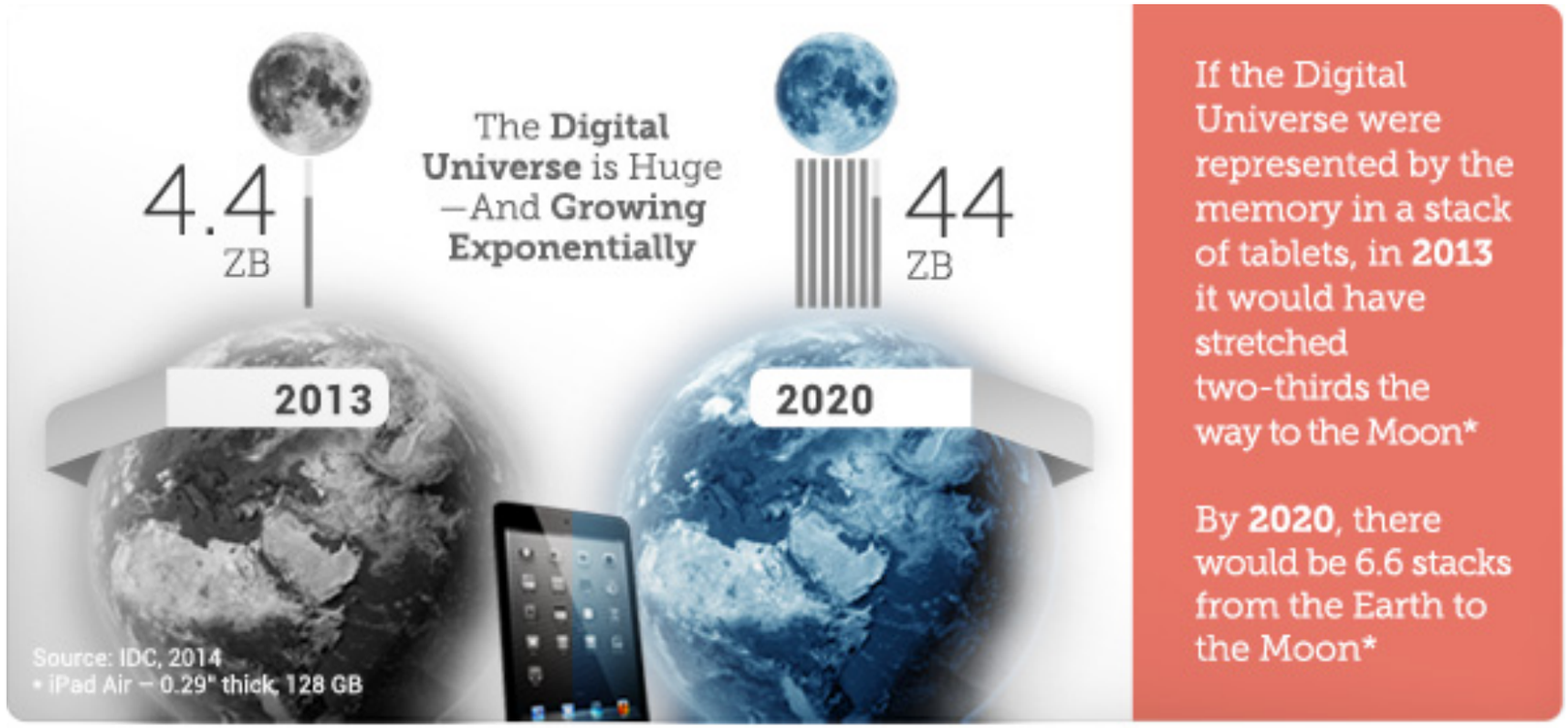
Visualization,
Human
perception

We actually need a collaboration



How much data have humans produced ?

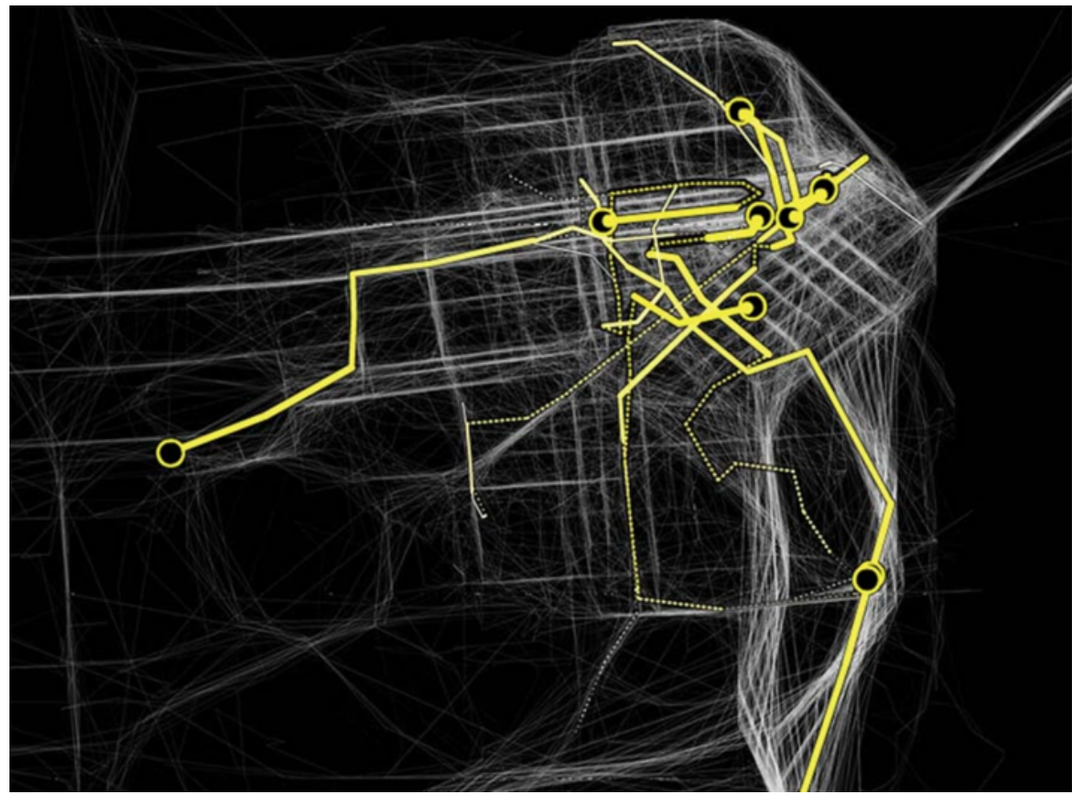
- By 2013: 4,4 zetabytes (1 zetabyte = 10^{21} bytes = 10^{9} terabytes)



https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

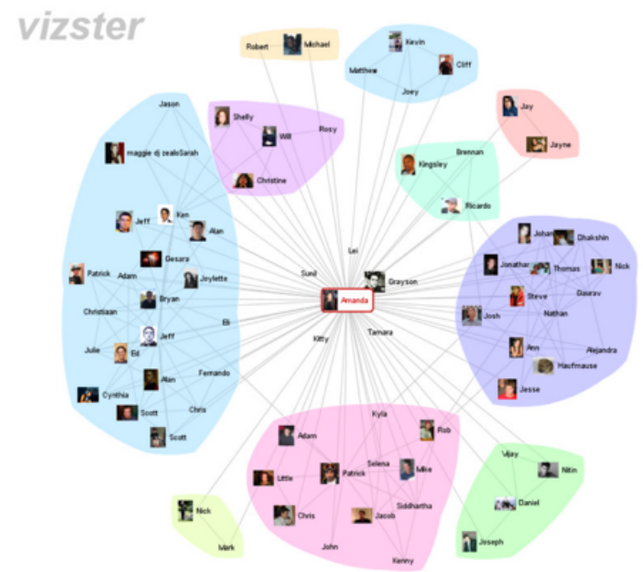
Where these data come from? Physical Sensors

- e.g., GPS data from NYC taxi cabs: cabspotting.org



Where these data come from? Human behavior online

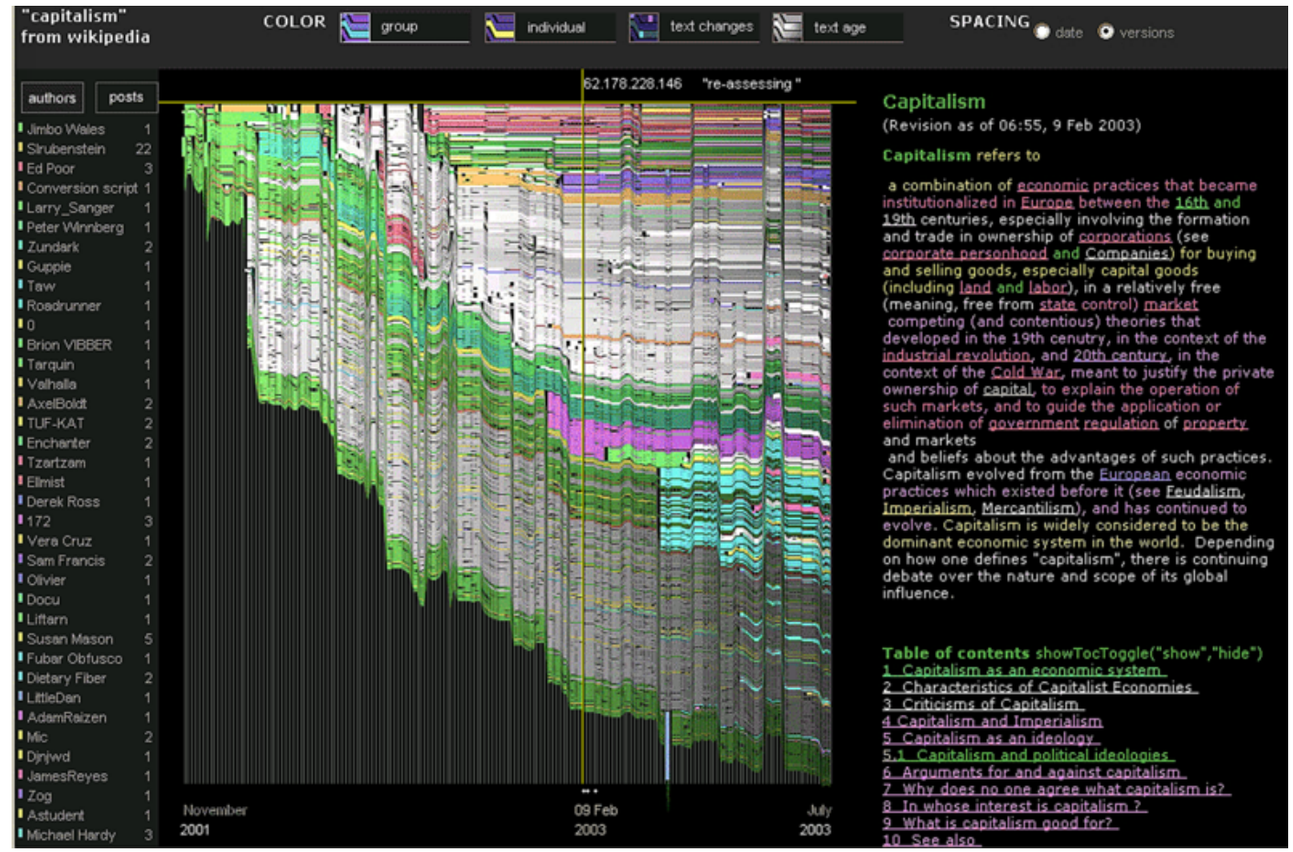
- Activity logs from social networks, online communities, media consumption, etc.



Heer, J., & Boyd, D. (2005, October). Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (pp. 32-39). IEEE.

Where these data come from? Human behavior online

- IBM History Flow (2004):
- Visualization of Wikipedia *edit wars*.



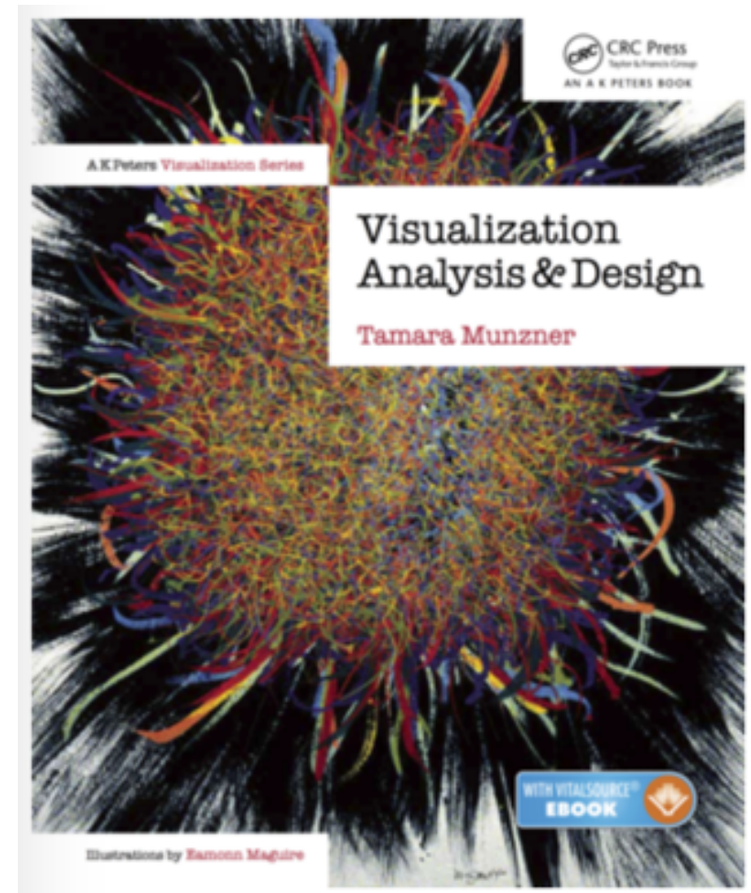
Information Overload

"What ***information consumes*** is rather obvious: it consumes the ***attention of its recipients***.
Hence a wealth of information creates a poverty of attention, ***and a need to allocate that attention efficiently*** among the overabundance of information sources that might consume it."

Herbert Simon
Nobel Prize in Economics
Turing Award

Let's review VAD definition

- Munzner's defines *Computer based visualization systems* as those systems that …



Defining visualization (vis)

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

Why?...

Visualization (vis) defined & motivated

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

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

- human in the loop needs the details & no trusted automatic solution exists
  - doesn't know exactly what questions to ask in advance
  - exploratory data analysis
    - *speed up* through human-in-the-loop visual data analysis
  - present known results to others
  - stepping stone towards automation
    - before model creation to provide understanding
    - during algorithm creation to refine, debug, set parameters
    - before or during deployment to build trust and monitor

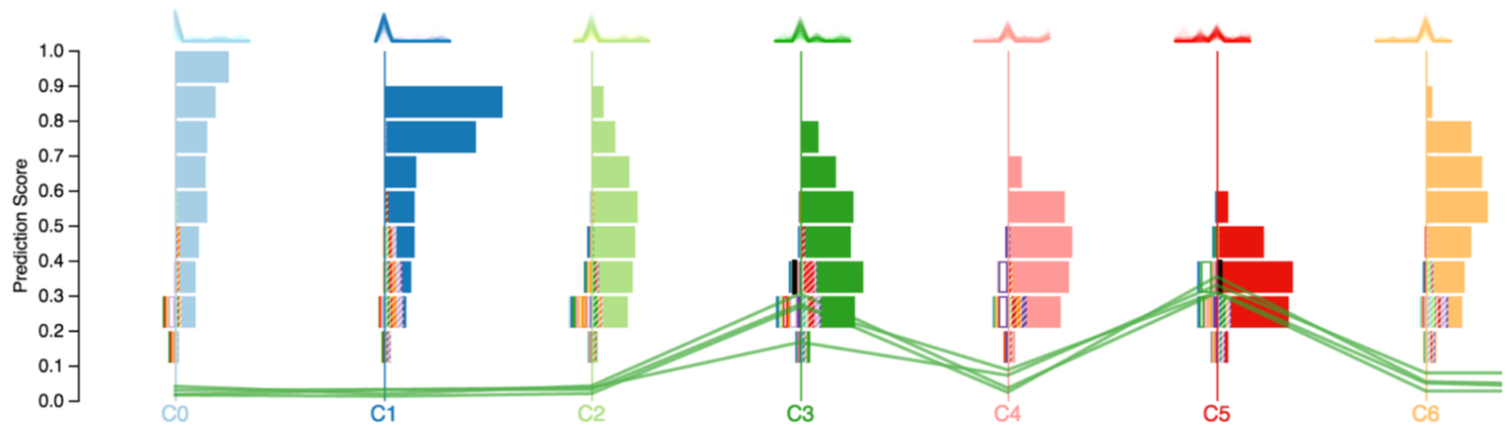
Why use an external representation?

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

- external representation: replace cognition with perception

- IEEE VIS 2016:

Ren et al.
Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers



Dataset: **scenario.randomforest.csv** | 10 Classes, 5000 Instances, 5000 Shown | Acc.: 0.87, Prec.: 0.87, Recall: 0.87, TP: 436.0, FP: 64.0, TN: 4436.0, FN: 64.0

★	Image	TRUE	Assigned	Correct	Prediction Score	C0	C1	C2
☆		C3	C5	0	0.3094	0.0426	0.0208	0.0437
☆		C3	C5	0	0.3561	0.0315	0.0339	0.0331
☆		C3	C5	0	0.3121	0.0180	0.0144	0.0218
☆		C3	C5	0	0.3320	0.0188	0.0326	0.0396
☆		C7	C7	1	0.5849	0.0103	0.0226	0.0646
☆		C1	C1	1	0.3576	0.0111	0.3576	0.1180

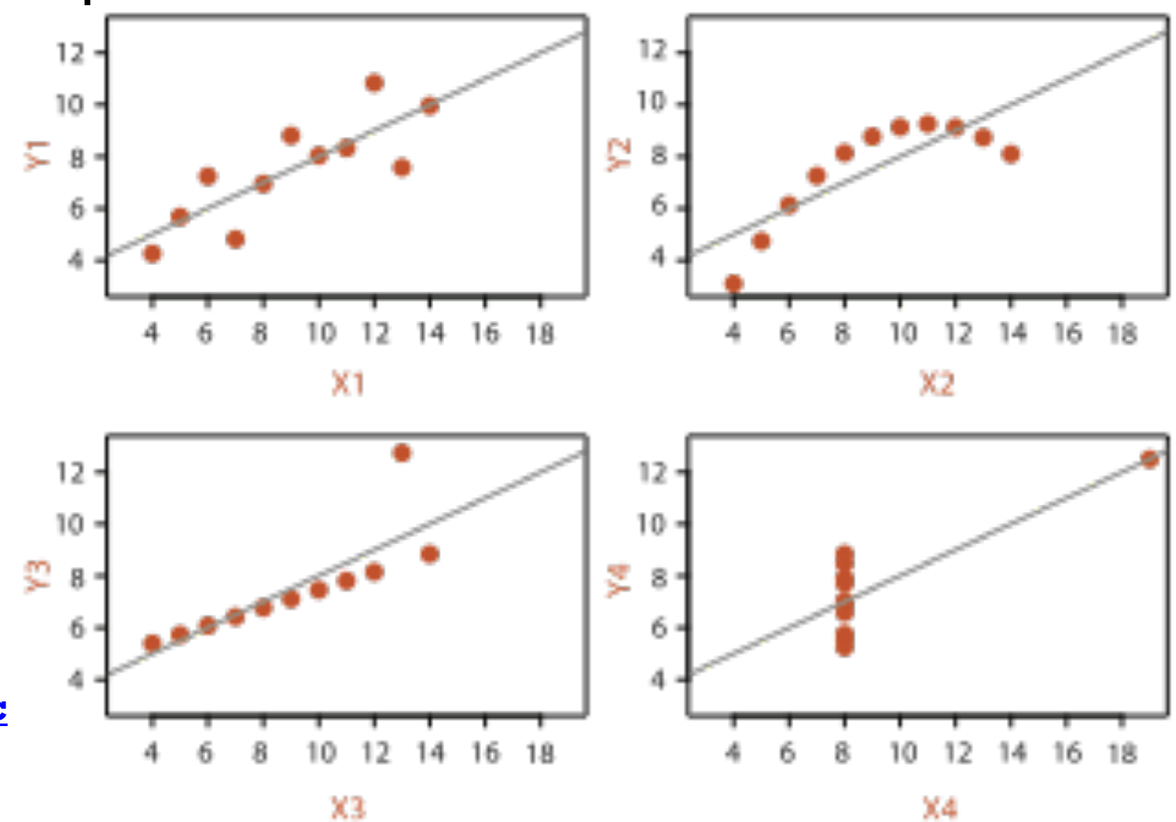
Why represent all the data?

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

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

Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816



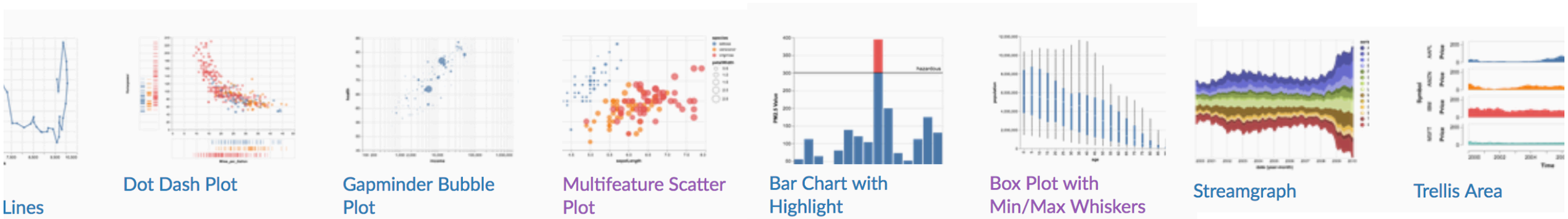
https://www.youtube.com/watch?v=DbJyPELmhJc

Same Stats, Different Graphs

Why focus on tasks and effectiveness?

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

- effectiveness requires match between data/task and representation
  - set of representations is huge
  - many are ineffective mismatch for specific data/task combo
  - increases chance of finding good solutions if you understand full space of possibilities



Why focus on tasks and effectiveness?

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

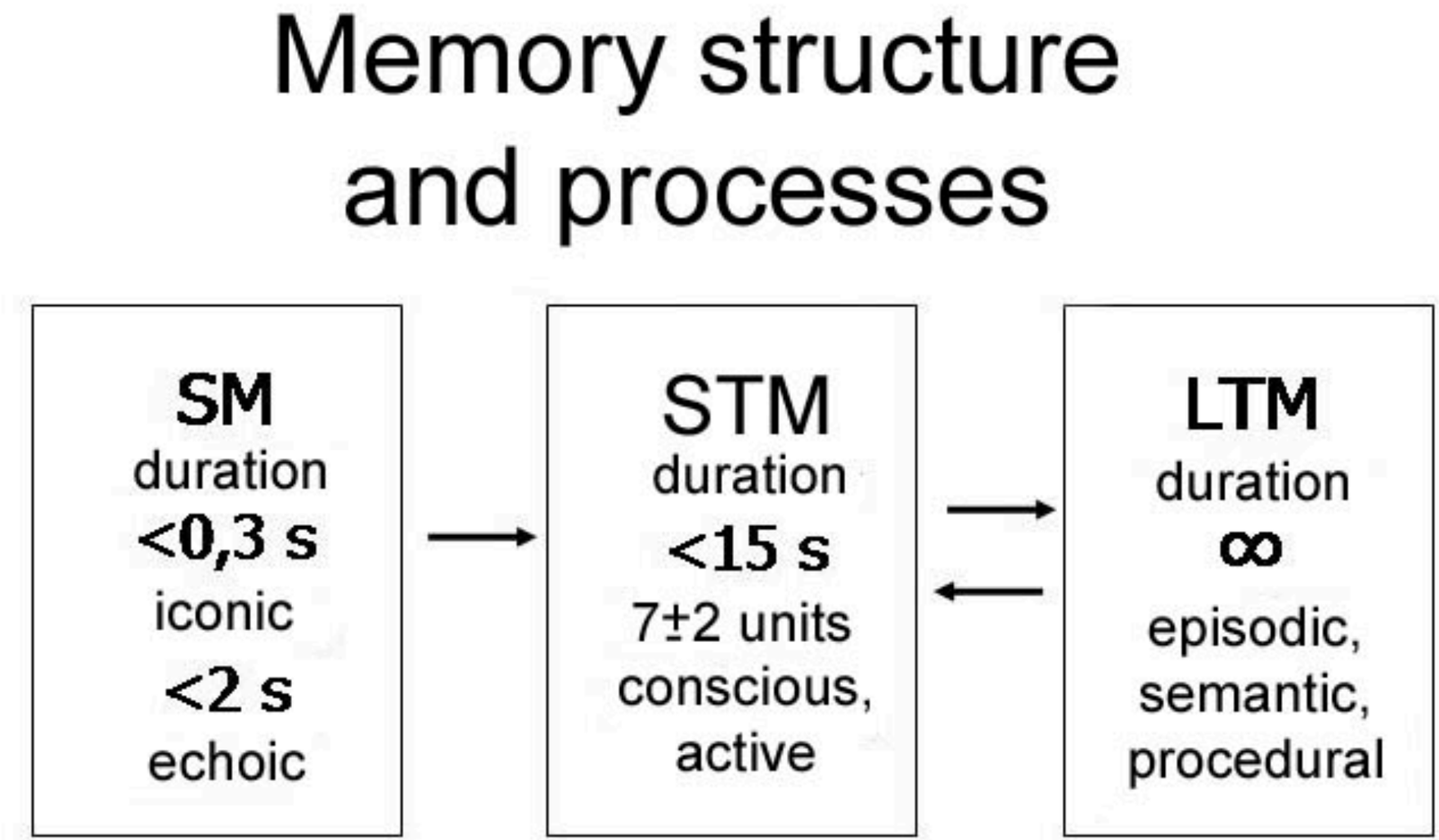
- effectiveness requires match between data/task and representation
  - set of representations is huge
  - many are ineffective mismatch for specific data/task combo
  - increases chance of finding good solutions if you understand full space of possibilities
- what counts as effective?
  - novel: enable entirely new kinds of analysis
  - faster: speed up existing workflows
- how to validate effectiveness
  - many methods, must pick appropriate one for your context

What resource limitations are we faced with?

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

Human memory limits

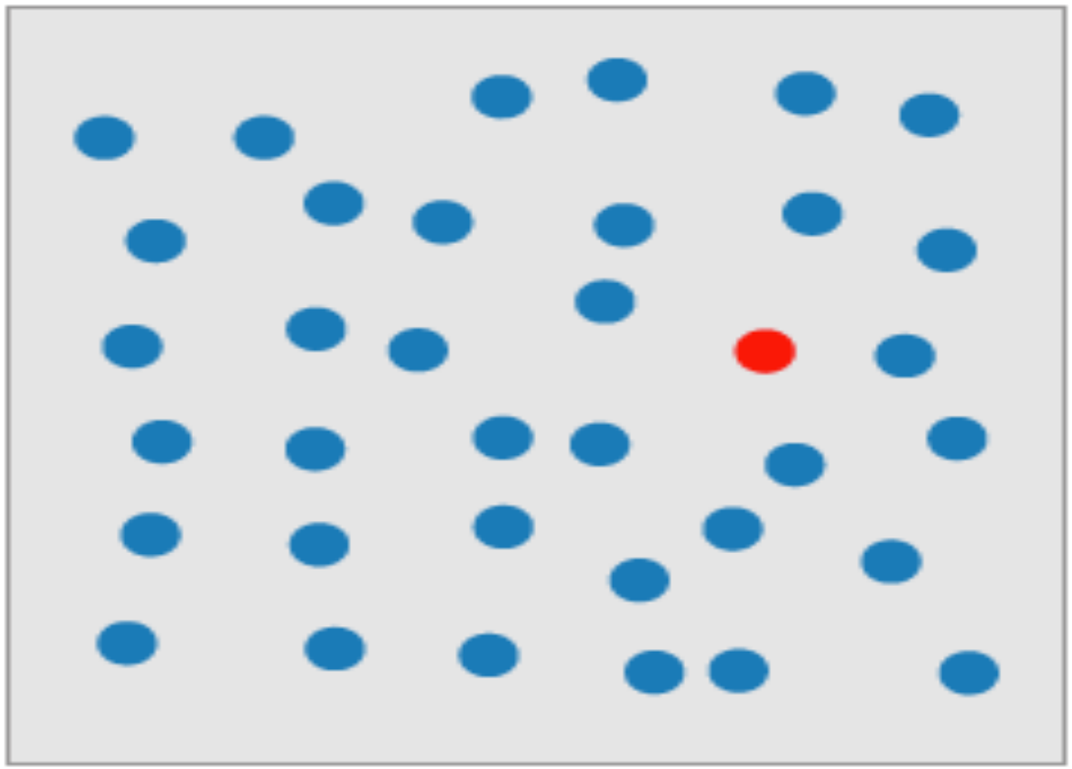


Pre-attentive Processing

"An understanding of what is processed pre-attentively is probably the most important contribution that visual science can make to data visualization" (Ware, 2004, p. 19)

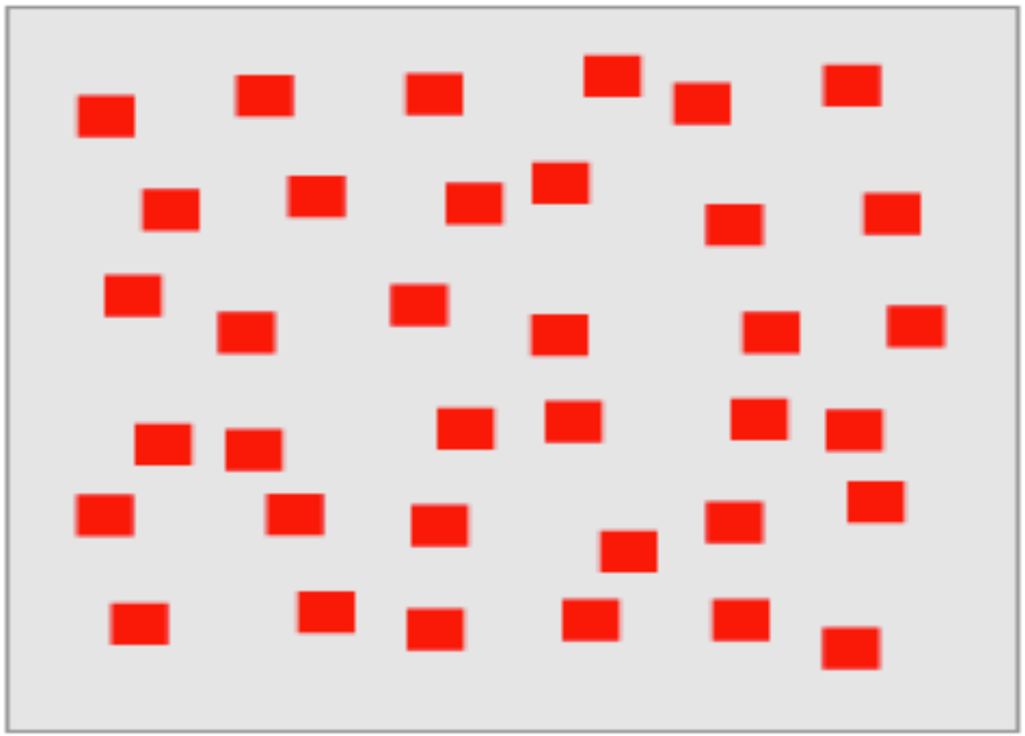
Taken from: http://www.slideshare.net/kverbert/information-visualization-perception-and-principles

Find the outlier element



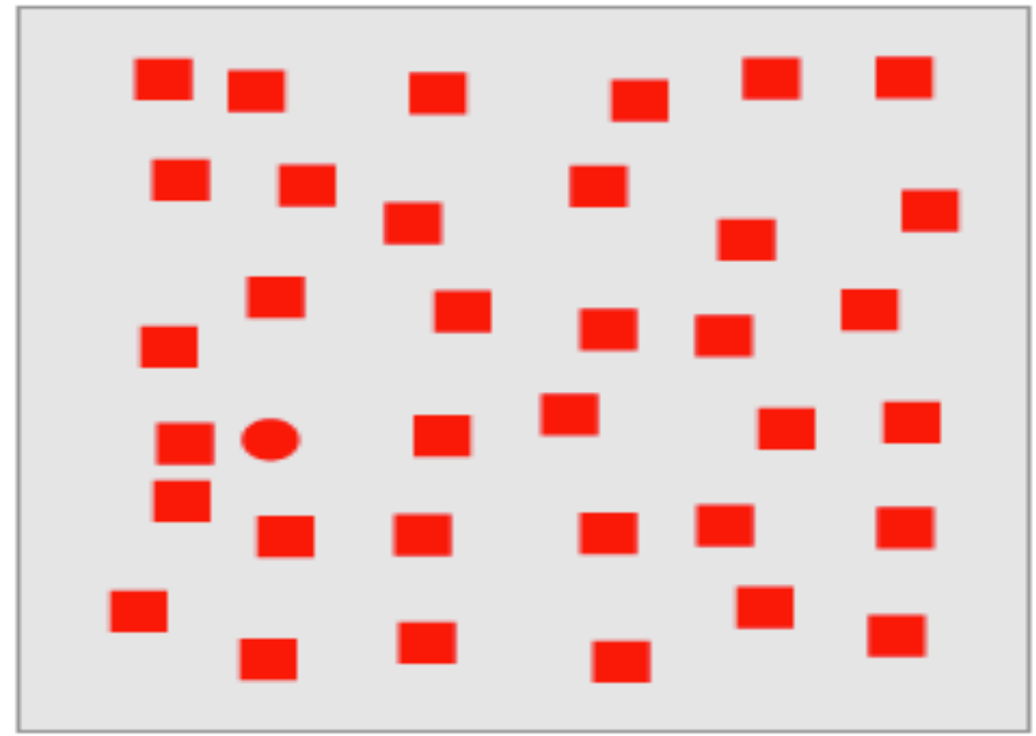
(a)

Find the outlier element



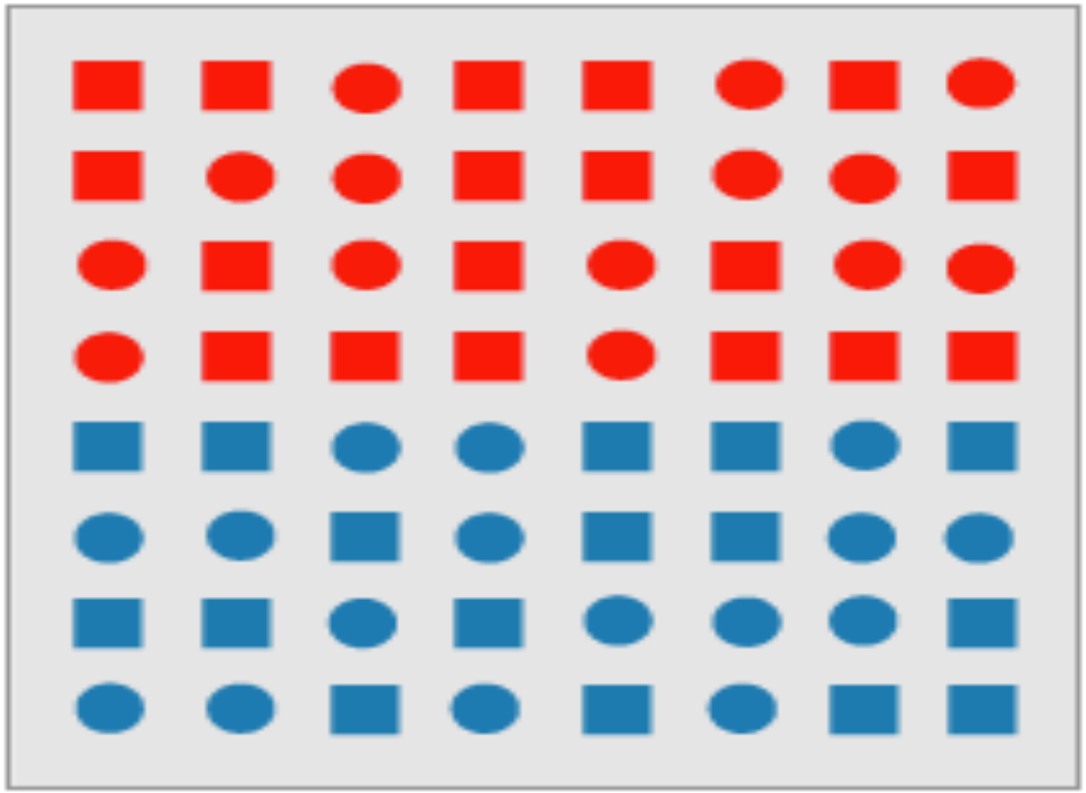
(a)

Find the outlier element



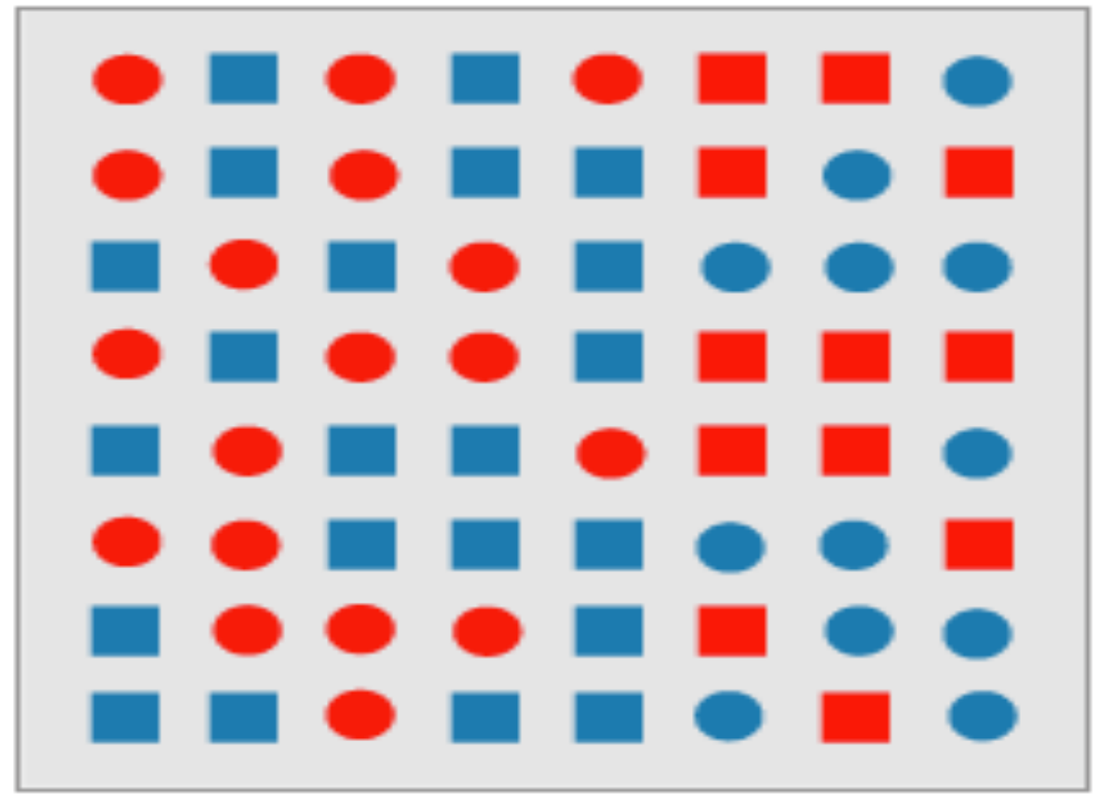
(b)

Draw a straight line to separate in 2 groups



(a)

Draw a straight line to separate in 2 groups



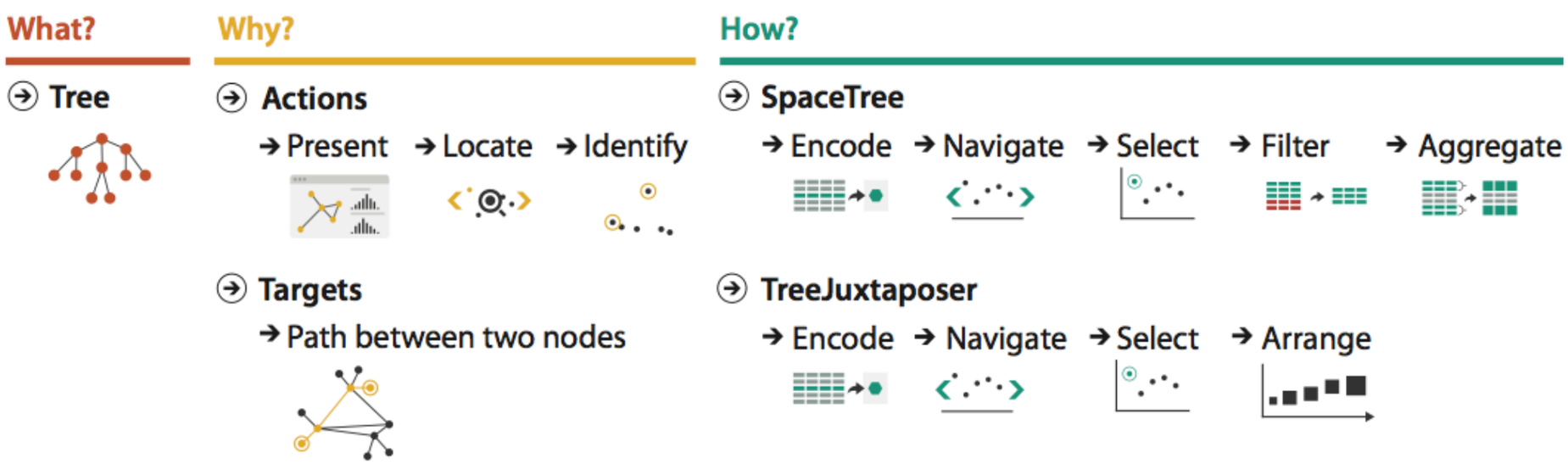
(b)

… the space of posibilities is huge!

- How are we going to find the right encoding for our task/data ?

A framework for analysis of visualizations

- Imposes structure on huge design space
  - scaffold to help you think systematically about choices
  - analyzing existing as stepping stone to designing new
  - most possibilities ineffective for particular task/data combination



Further reading

- Visualization Analysis and Design. Munzner. AK Peters Visualization Series, CRC Press, 2014.
  - *Chap 1: What's Vis, and Why Do It?*

Outline

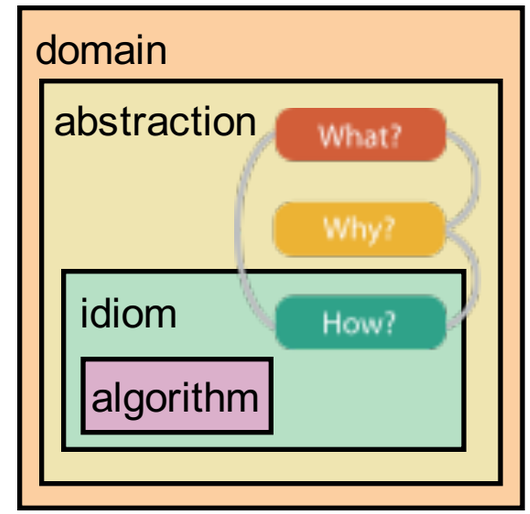
- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - Introduction: Definitions
  - **Analysis: What, Why, How**
  - Marks and Channels

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - Arrange Tables
  - Hands on: Altair
  - Rules of Thumb & Vis Research

Nested model: Four levels of vis design

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

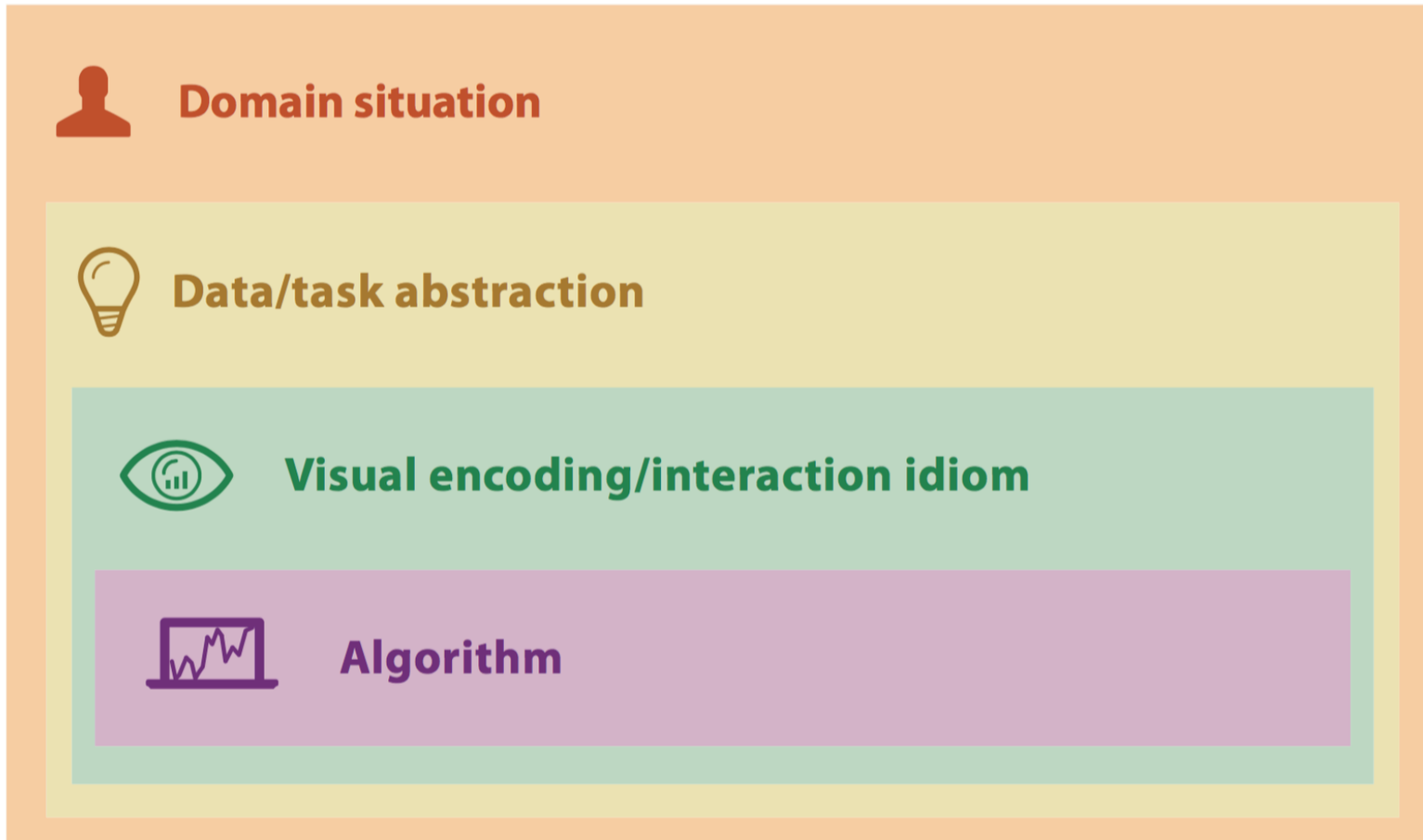
- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom*
  - **how** is it shown?
    - **visual encoding** idiom: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



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

Why is validation difficult?

- different ways to get it wrong at each level



Why is validation difficult?

- solution: use methods from different fields at each level

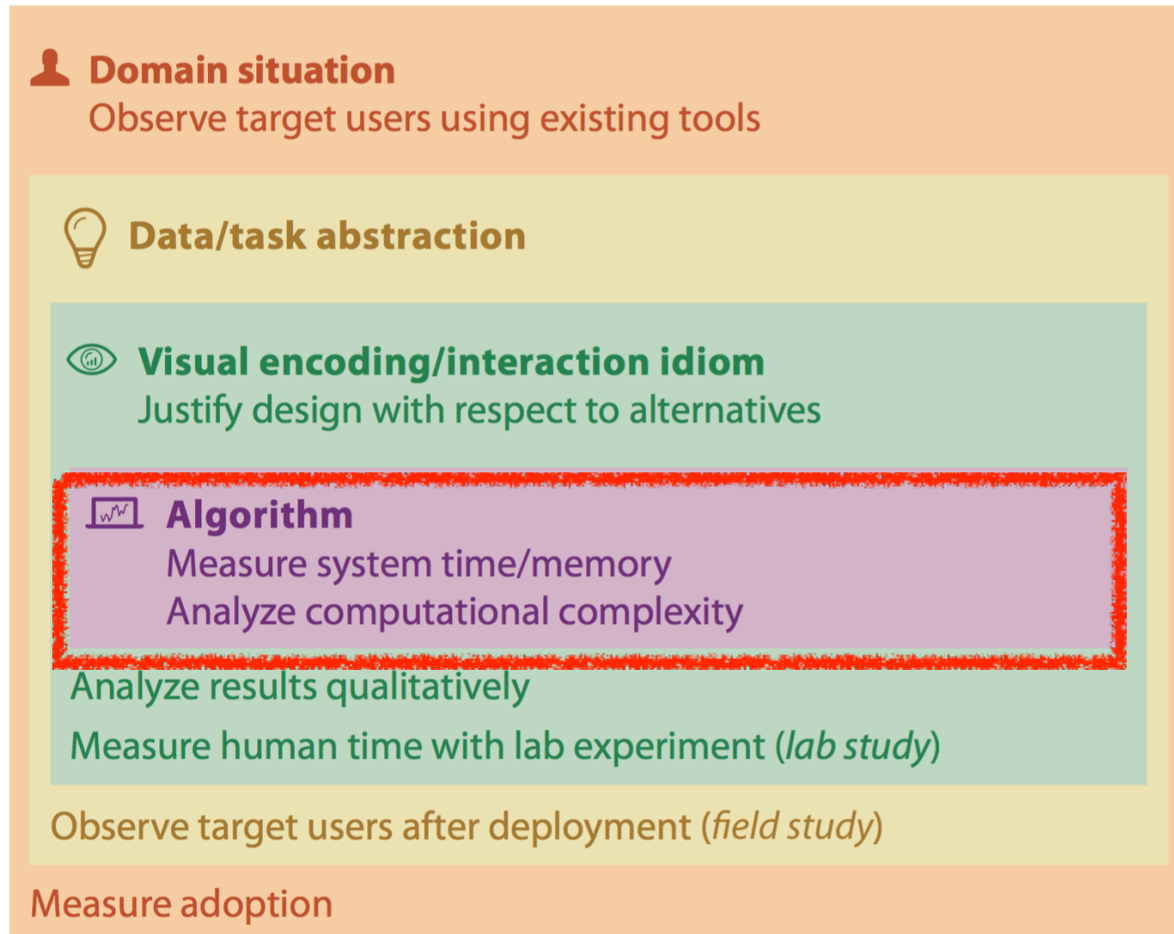
anthropology/ ethnography

design

computer science

cognitive psychology

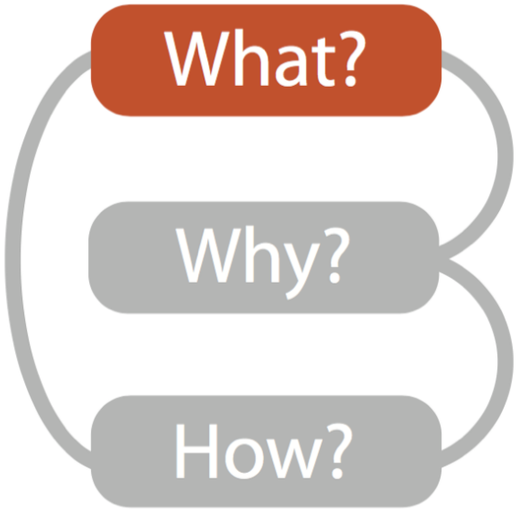
anthropology/ ethnography

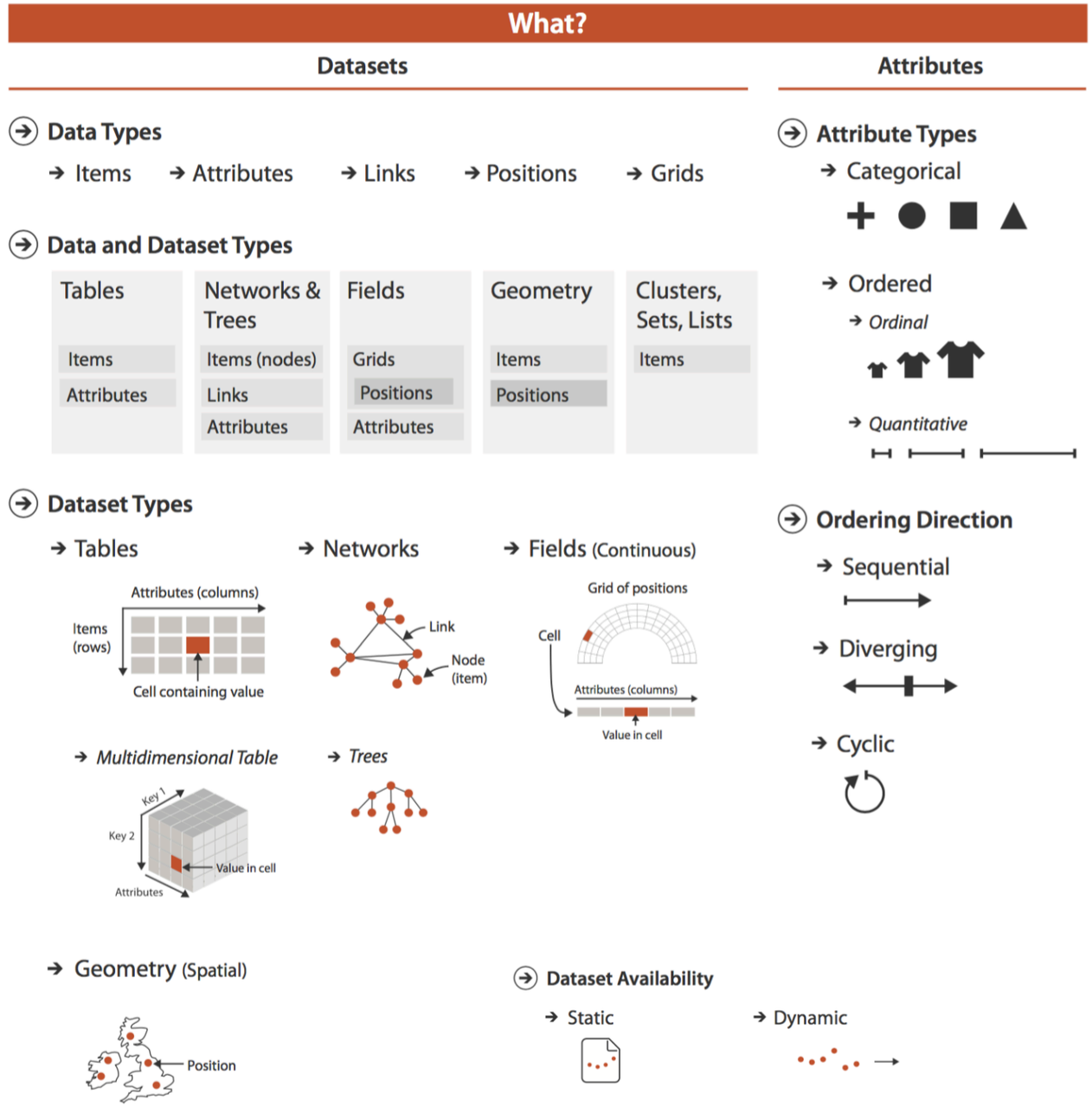


problem-driven work

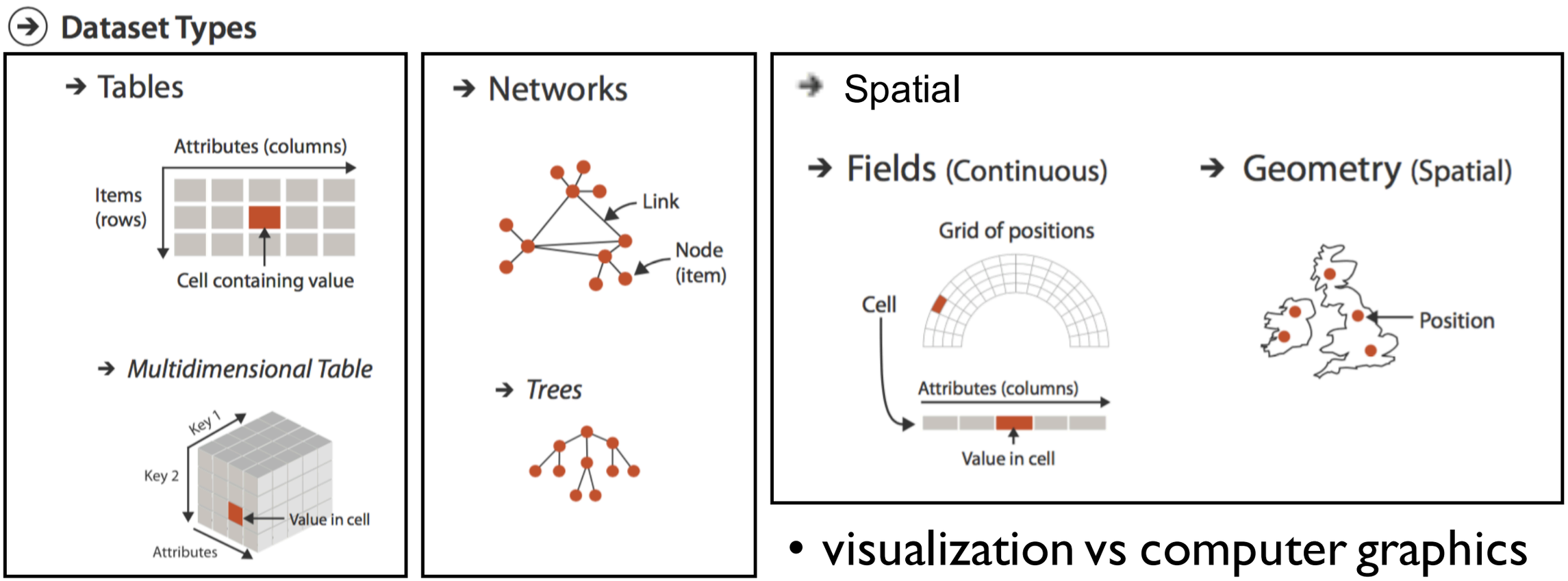
technique-driven work

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


What?
Why?
How?


What?
Datasets
Attributes
Data Types
Items
Attributes
Links
Positions
Grids
Data and Dataset Types
Tables
Items
Attributes
Networks & Trees
Items (nodes)
Links
Attributes
Fields
Grids
Positions
Attributes
Geometry
Items
Positions
Clusters, Sets, Lists
Items
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

Three major dataset types



- visualization vs computer graphics
  - geometry is design decision

Dataset and data types

Data and Dataset Types

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

Data Types

→ Items → Attributes → Links → Positions → Grids

Dataset Availability

→ Static

→ Dynamic

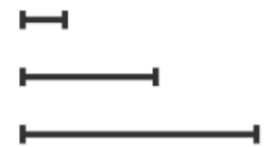
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*

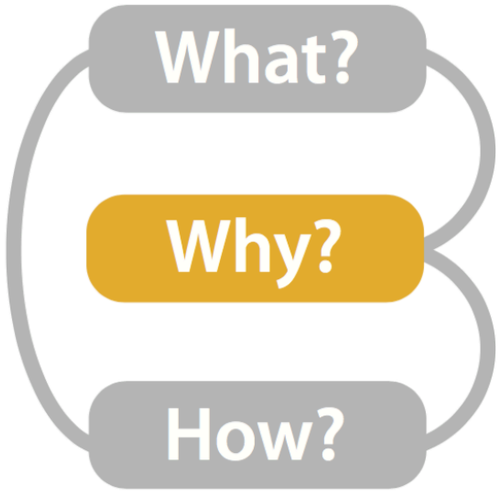


Ordering Direction

- Sequential

- Diverging

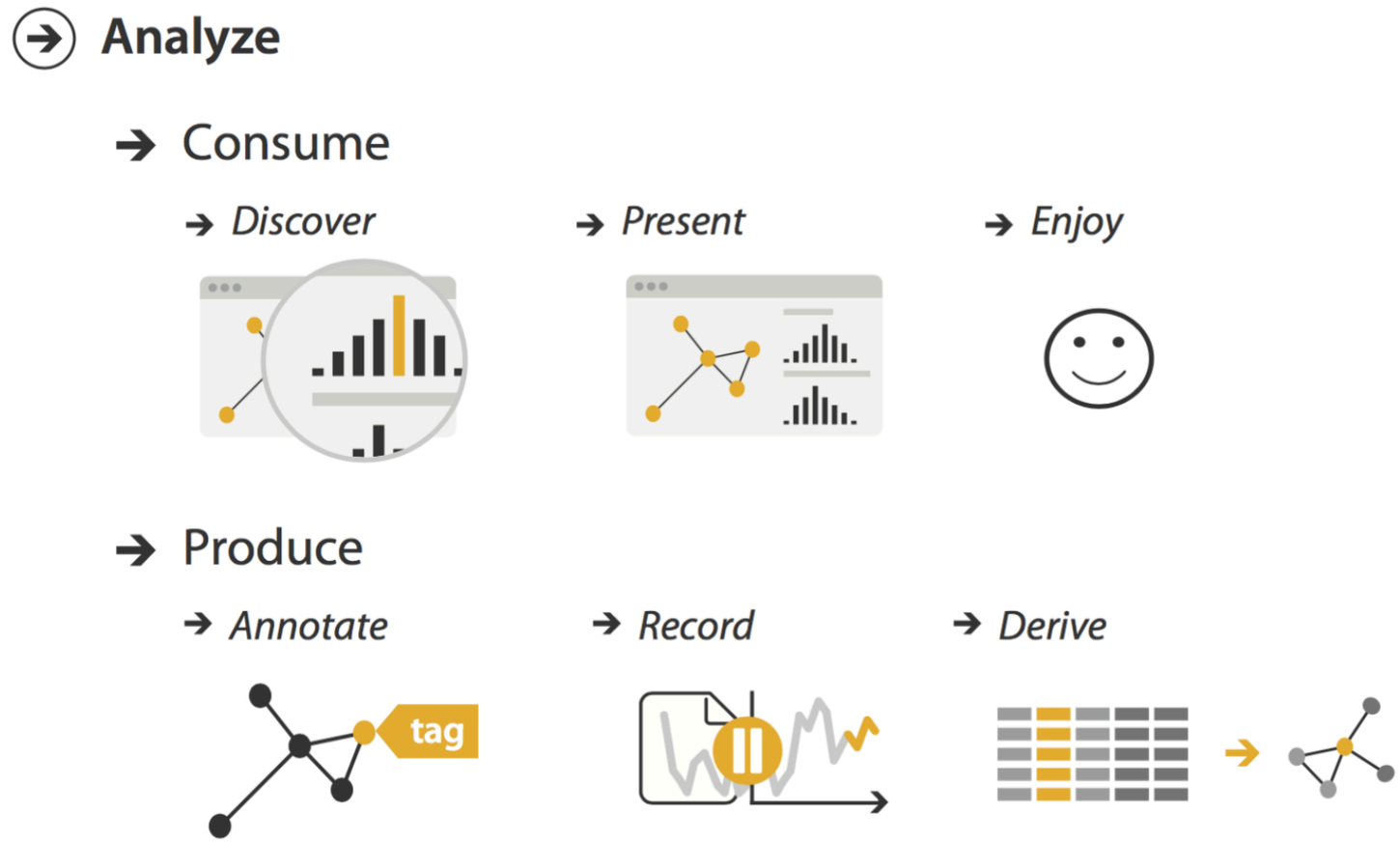
- Cyclic



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

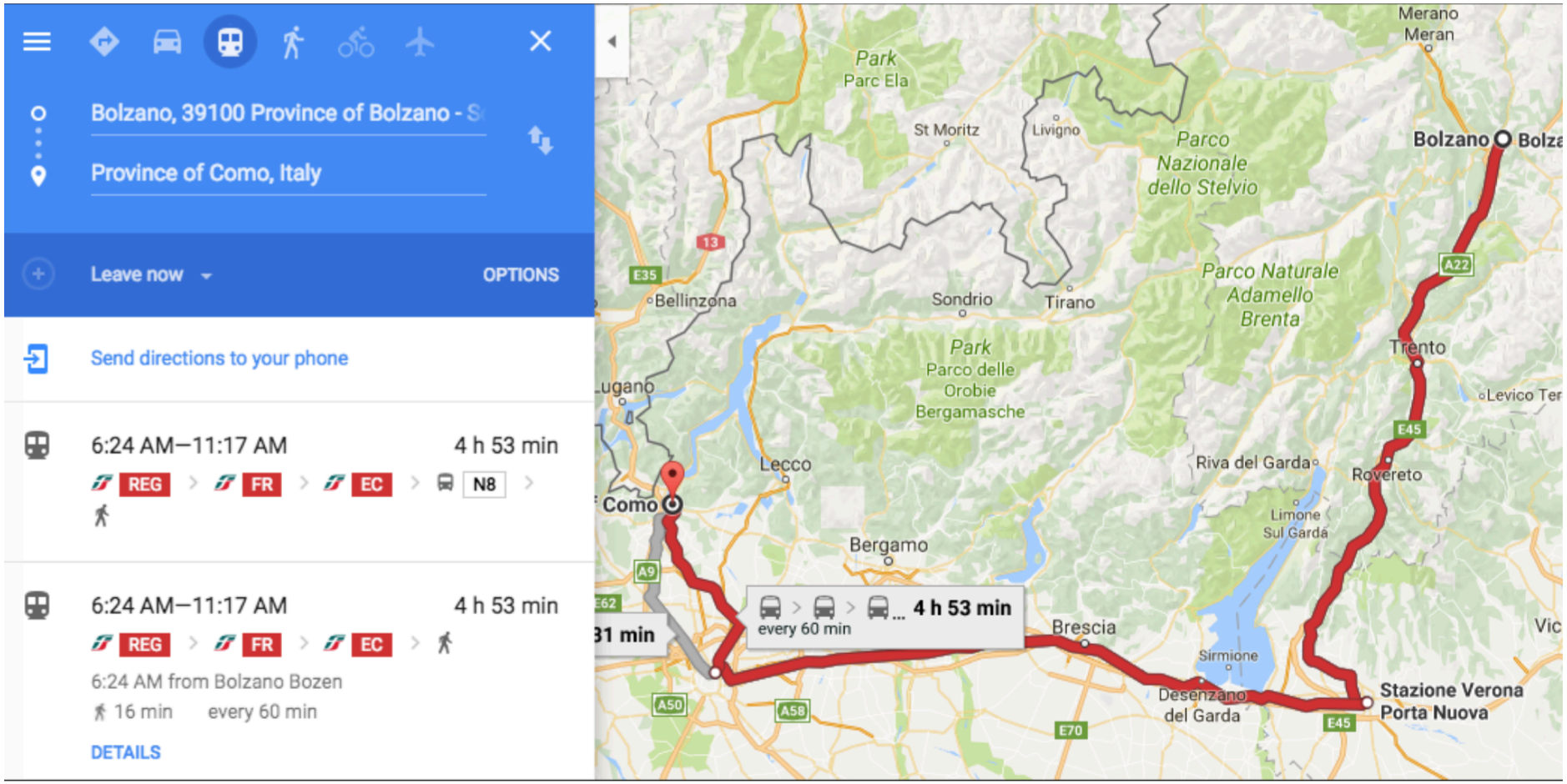
Actions: Analyze

- consume
  - discover vs present
    - classic split
    - aka explore vs explain
  - enjoy
    - newcomer
    - aka casual, social
- produce
  - annotate, record
  - derive
    - crucial design choice



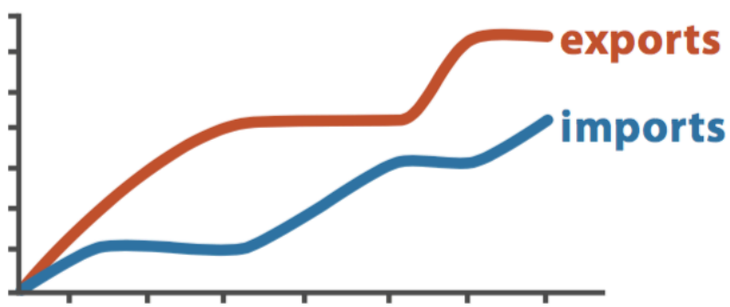
Produce > Anotate

- Using Google Maps

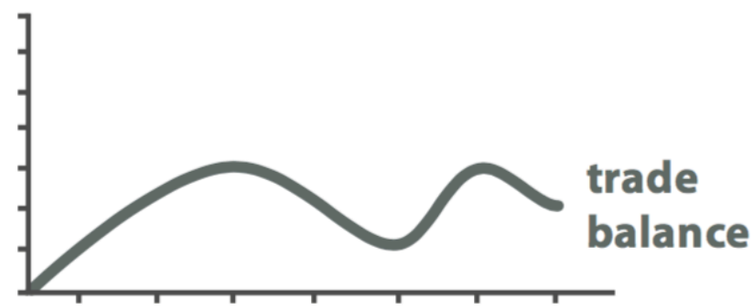


Produce > Derive

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



Original Data



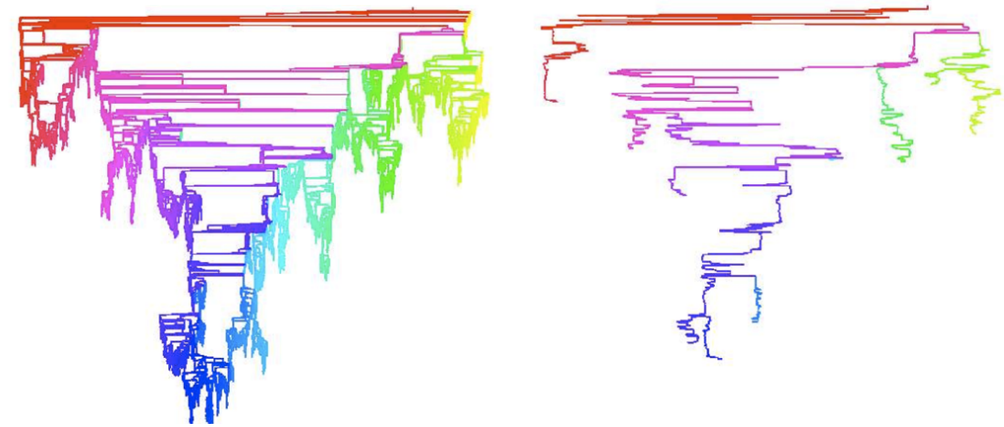
trade balance = *exports* − *imports*

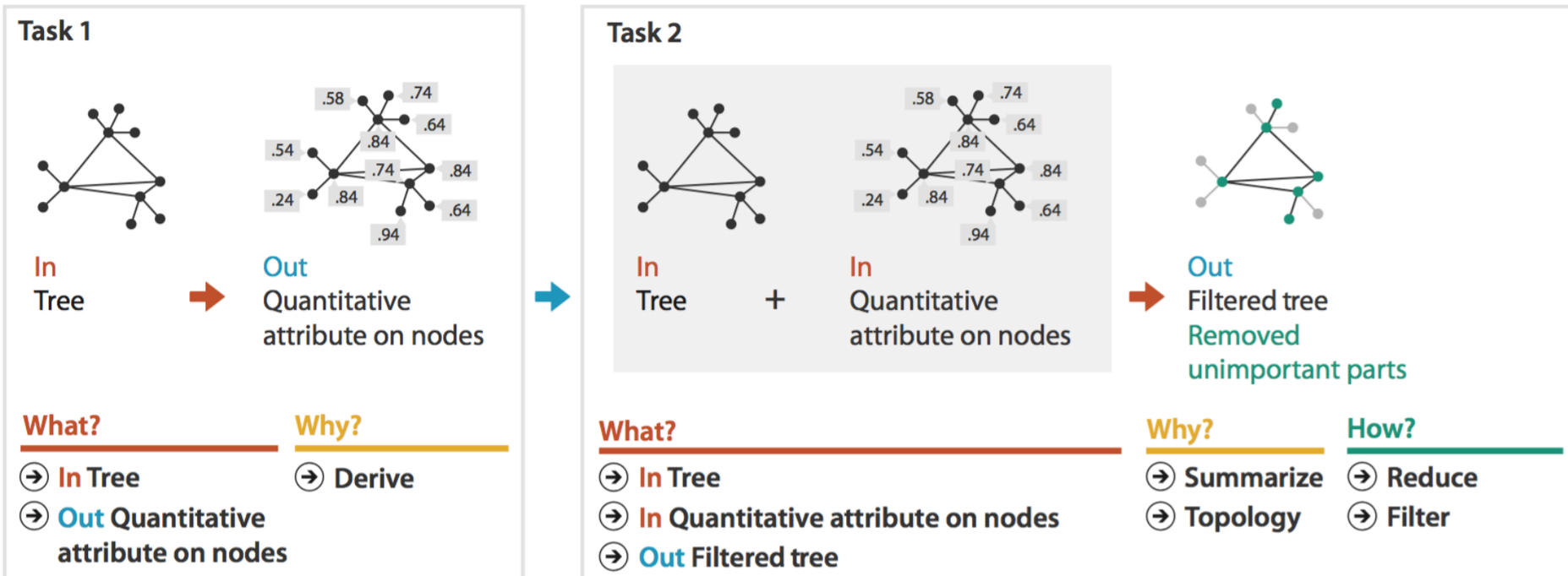
Derived Data

Analysis example: Derive one attribute

- Strahler number
  - centrality metric for trees/networks
  - derived quantitative attribute
  - draw top 5K of 500K for good skeleton

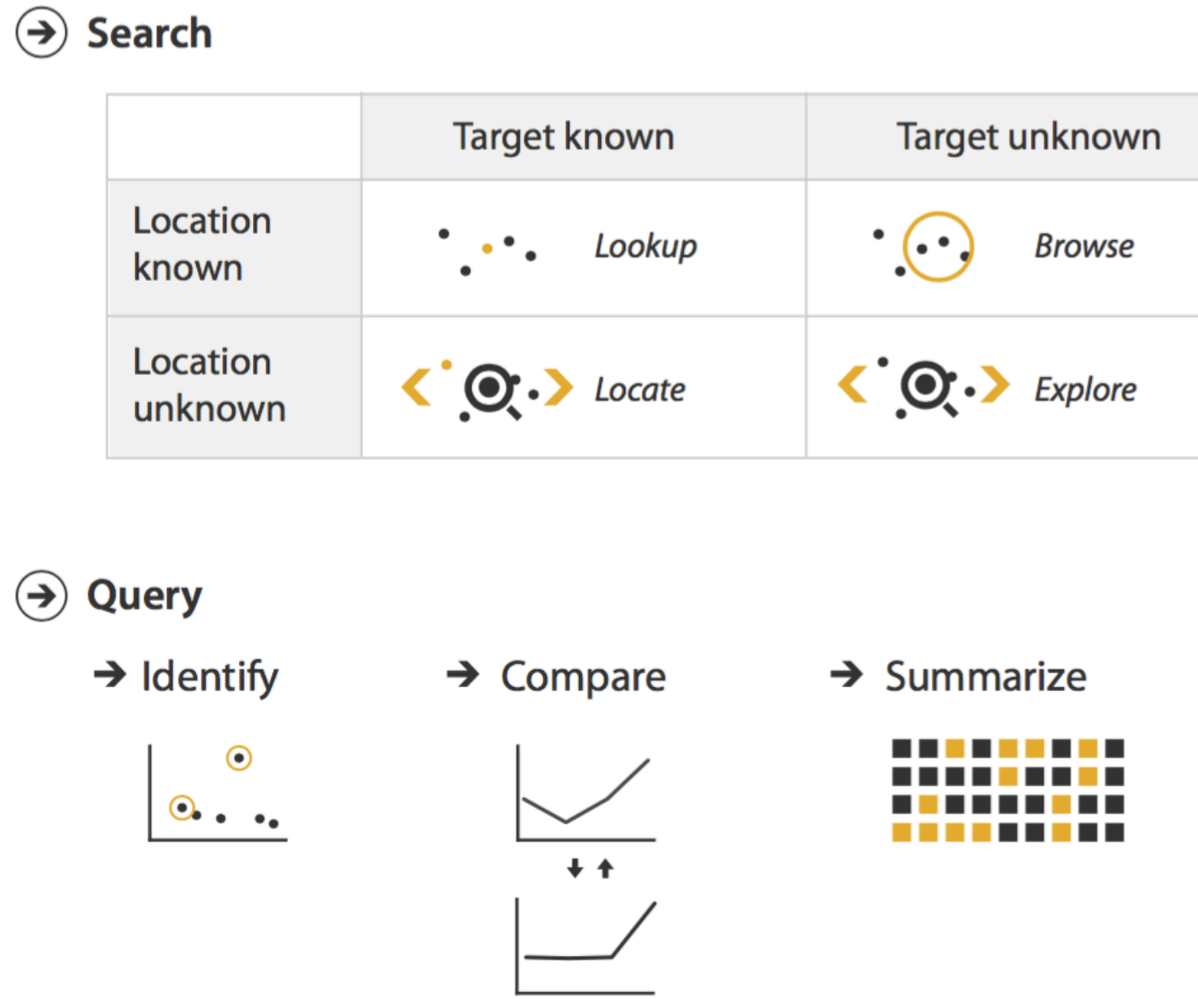
    [Using Strahler numbers for real time visual exploration of huge graphs. Auber. Proc. Intl. Conf. Computer Vision and Graphics, pp. 56–69, 2002.]





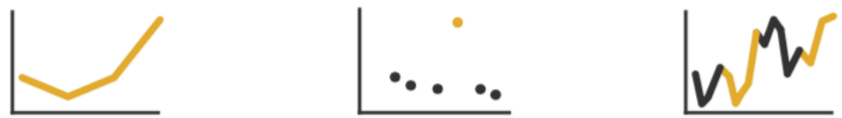
Actions: Search, query

- what does user know?
  - target, location
- how much of the data matters?
  - one, some, all
- independent choices for each of these three levels
  - analyze, search, query
  - mix and match



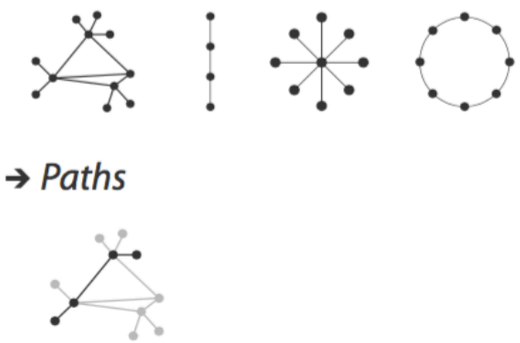
Why: Targets

- **All Data**
  - Trends
  - Outliers
  - Features

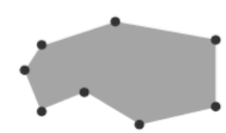


- **Attributes**
  - One
    - *Distribution*
    - *Extremes*
  - Many
    - *Dependency*
    - *Correlation*
    - *Similarity*

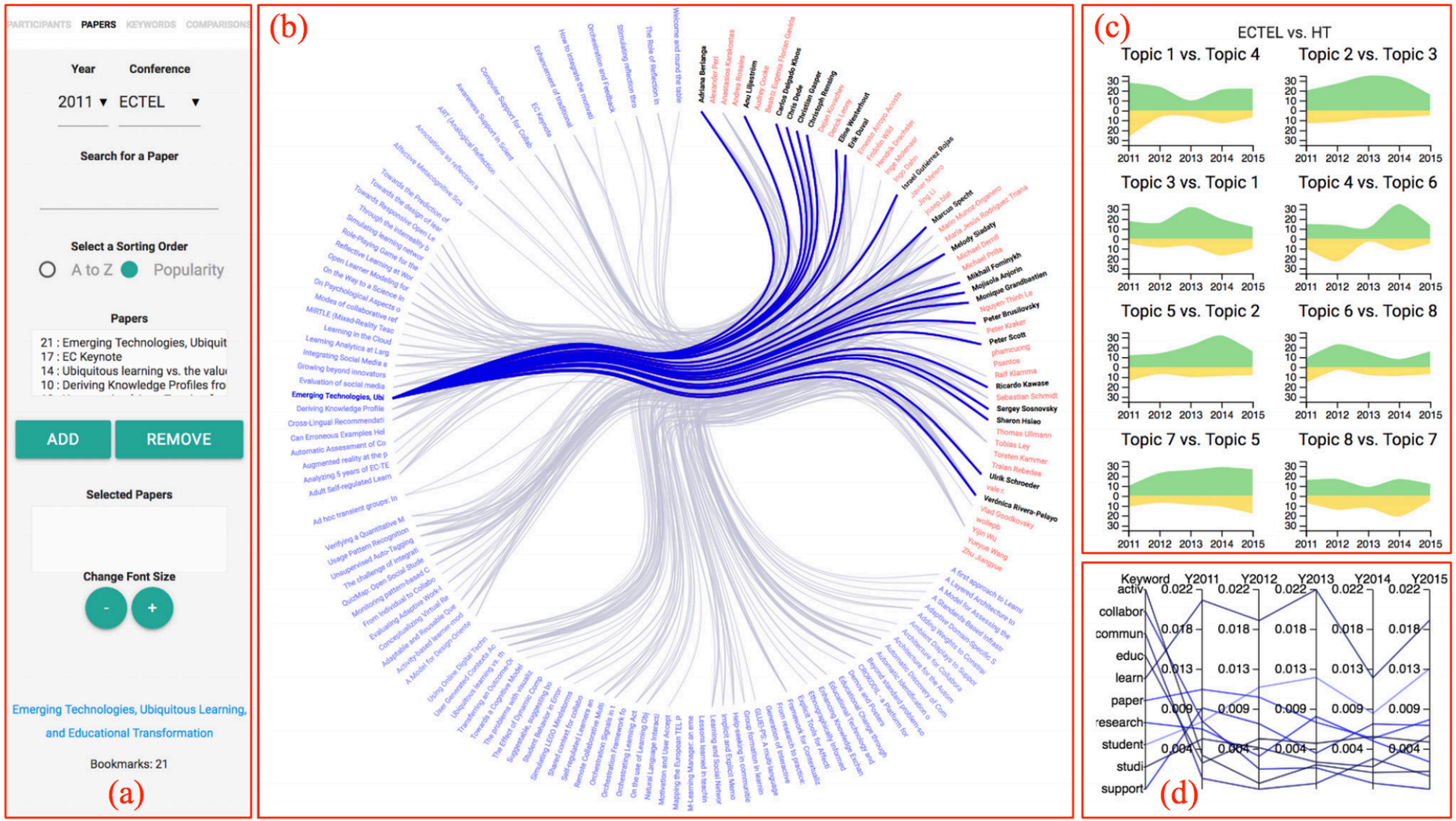
- **Network Data**
  - Topology
  - *Paths*



- **Spatial Data**
  - Shape



CNVis: Bailey et al. (2018)



Bailey, S. M., Wei, J. A., Wang, C., & Parra, D. CNVis: A Web-Based Visual Analytics Tool for Exploring Conference Navigator Data.

EpistAid: Donoso et al. (2018)

Donoso-Guzmán, I., & Parra, D. (2018, March). An Interactive Relevance Feedback Interface for Evidence-Based Health Care. In *23rd International Conference on Intelligent User Interfaces* (pp. 103-114). ACM.

How?

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

Activity

Download the following article
http://www.shixialiu.com/publications/TopicPanorama/index.htm

TopicPanorama: a Full Picture of Relevant Topics

IEEE VAST 2014

Shixia Liu[1] Xiting Wang[1,2] Jianfei Chen[2] Jun Zhu[2] Baining Guo[1,2]
[1]Microsoft Research Asia [2]Tsinghua University

Identify
- actions (why)
- targets
- how to implement

Further reading

- Visualization Analysis and Design. Munzner. AK Peters Visualization Series, CRC Press, 2014.
  - *Chap 2: What: Data Abstraction*
  - *Chap 3: Why: Task Abstraction*
- *A Multi-Level Typology of Abstract Visualization Tasks.* Brehmer and Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 19:12 (2013), 2376–2385.
- *Low-Level Components of Analytic Activity in Information Visualization.* Amar, Eagan, and Stasko. Proc. IEEE InfoVis 2005, p 111–117.
- *A taxonomy of tools that support the fluent and flexible use of visualizations.* Heer and Shneiderman. Communications of the ACM 55:4 (2012), 45–54.
- *Rethinking Visualization: A High-Level Taxonomy.* Tory and Möller. Proc. IEEE InfoVis 2004, p 151–158.
- Visualization of Time-Oriented Data. Aigner, Miksch, Schumann, and Tominski. Springer, 2011.

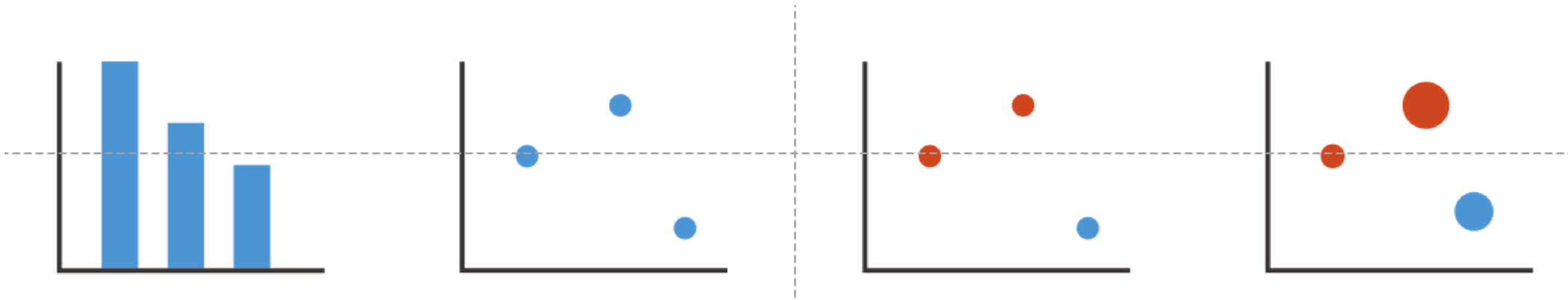
Outline

- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - Introduction: Definitions
  - Analysis: What, Why, How
  - **Marks and Channels**

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - Arrange Tables
  - Hands on: Altair
  - Rules of Thumb & Vis Research

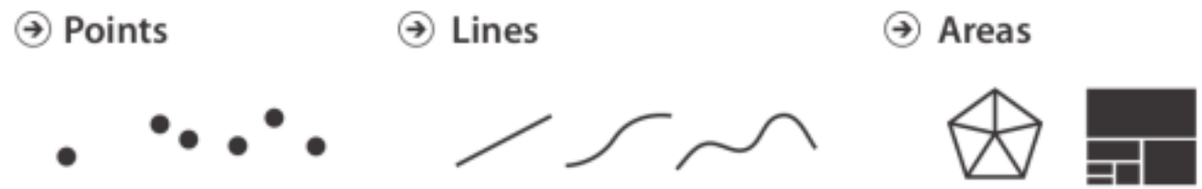
Visual encoding

- analyze idiom structure

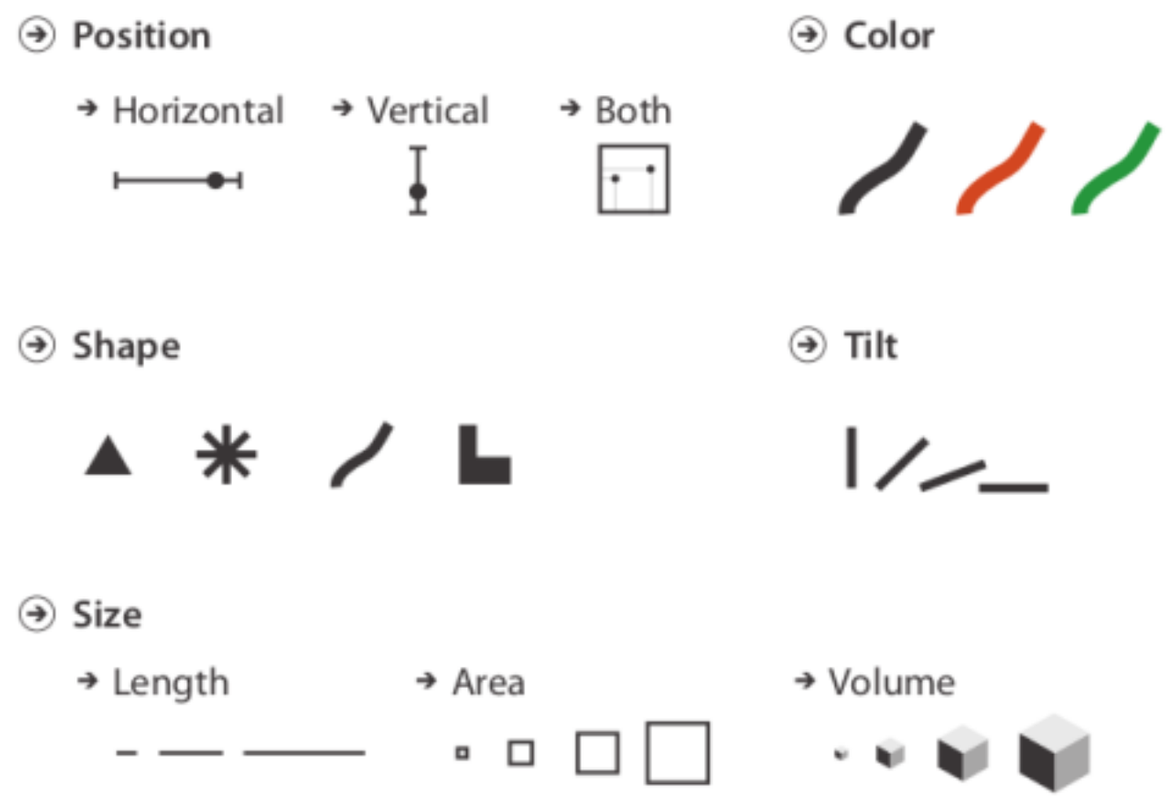


Definitions: Marks and channels

- marks
  - geometric primitives

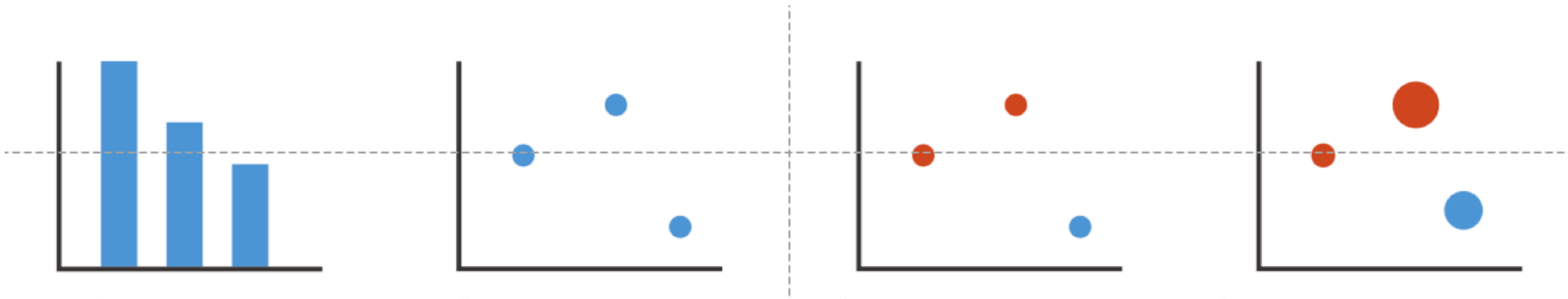


- channels
  - control appearance of marks
  - can redundantly code with multiple channels



Visual encoding

- analyze idiom structure
  - as combination of marks and channels



1:
vertical position

mark: line

2:
vertical position
horizontal position

mark: point

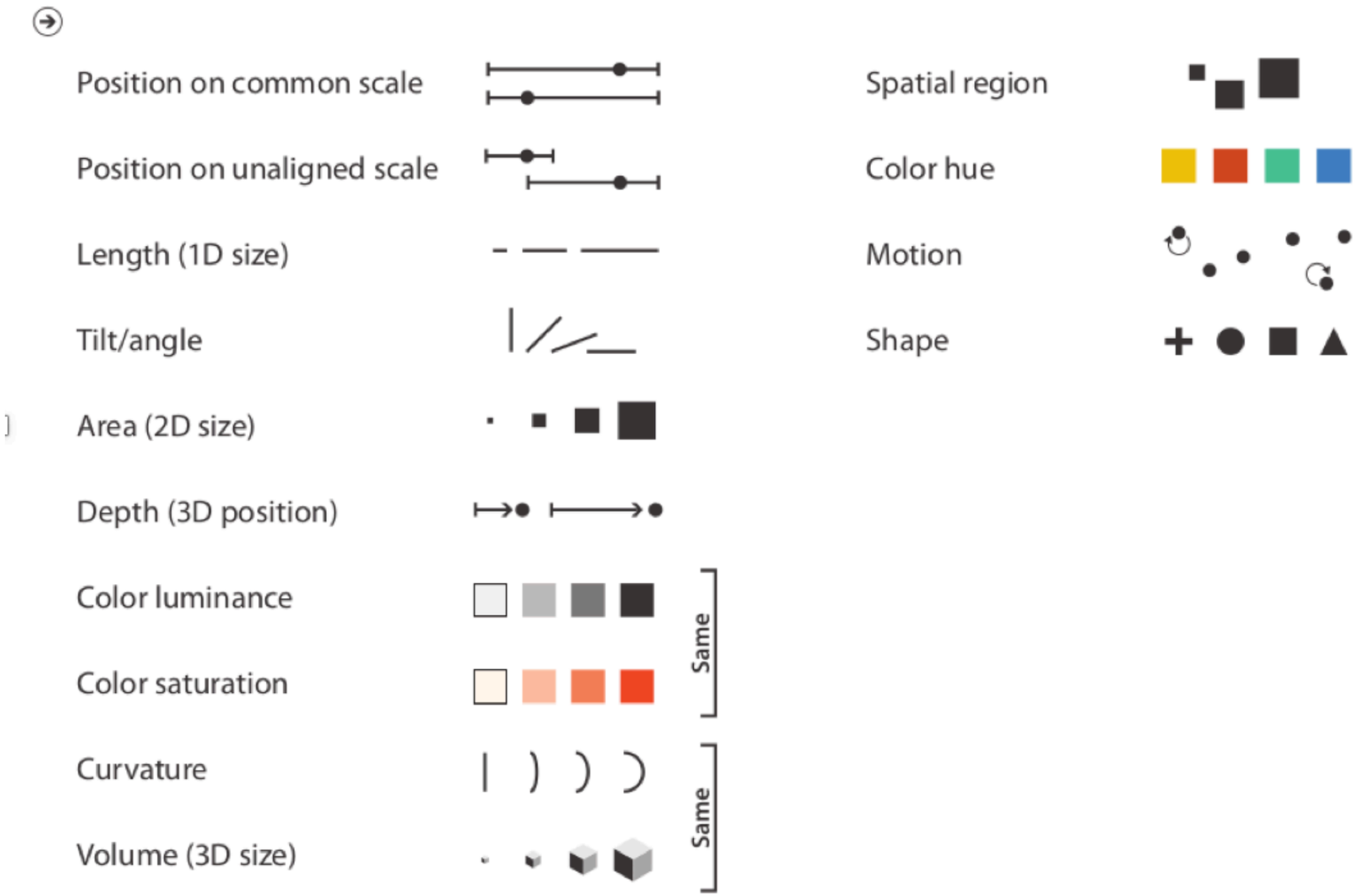
3:
vertical position
horizontal position
color hue

mark: point

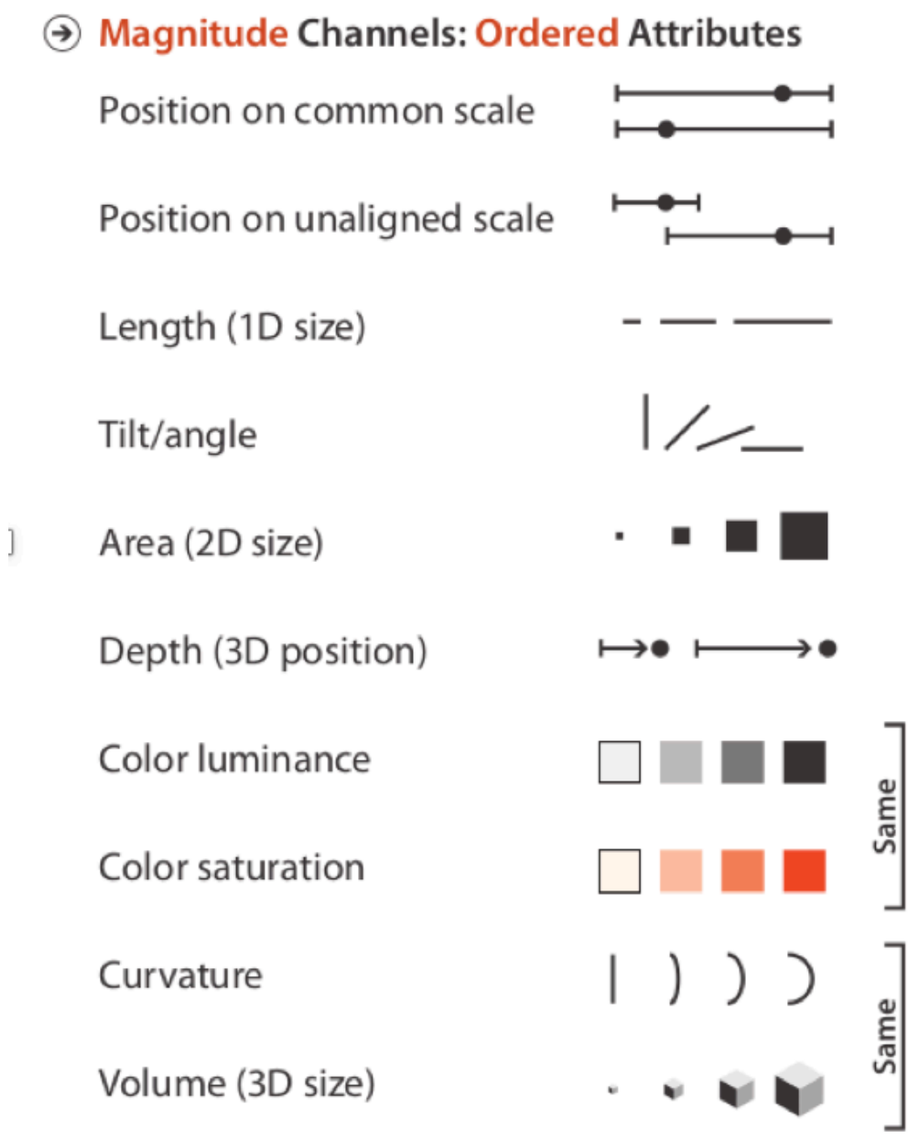
4:
vertical position
horizontal position
color hue
size (area)

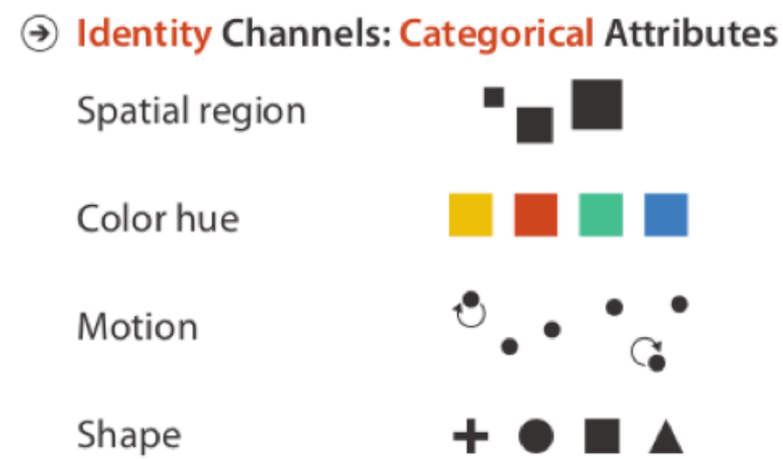
mark: point

Channels


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
Spatial region
Color hue
Motion
Shape

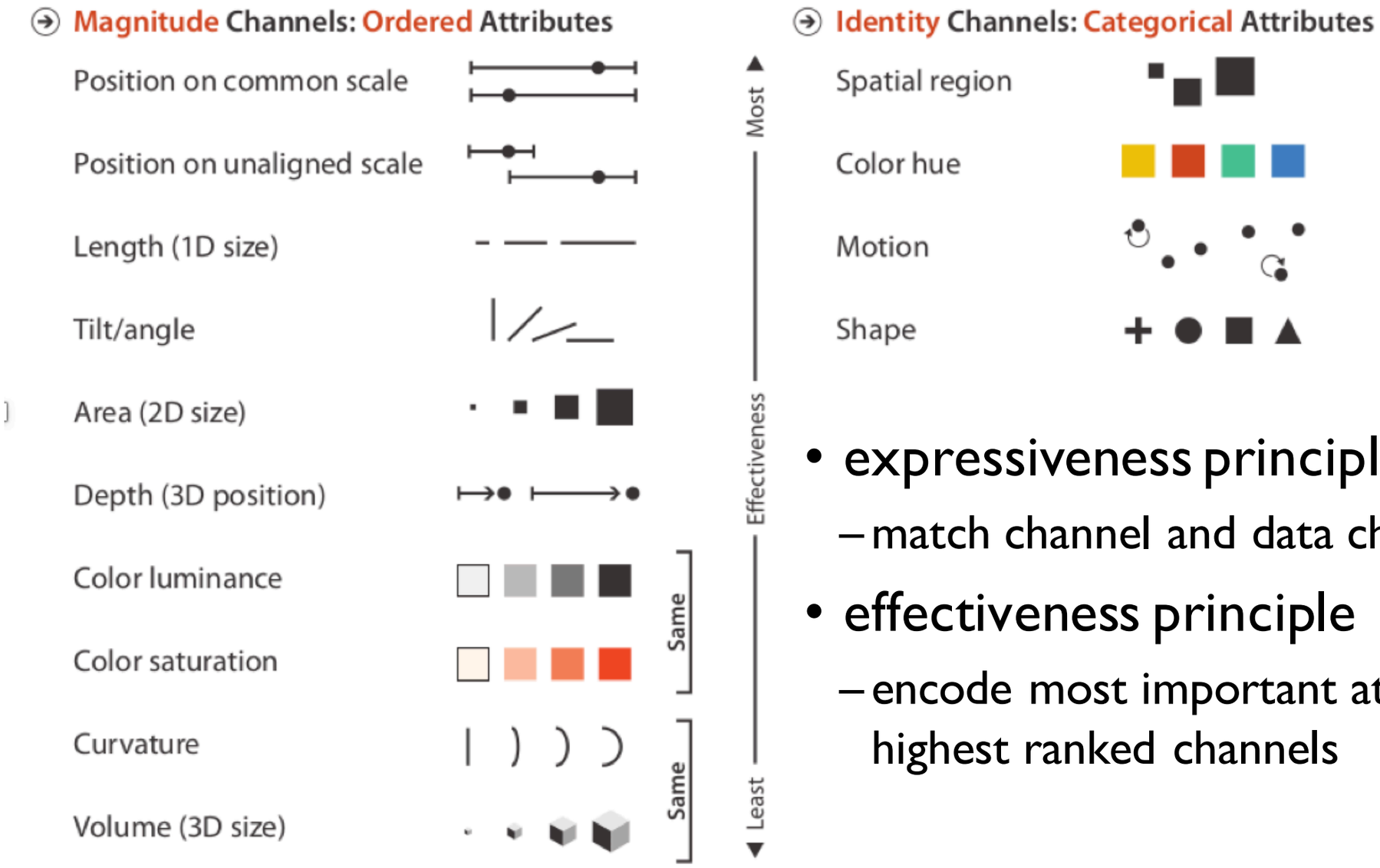
Channels: Matching Types





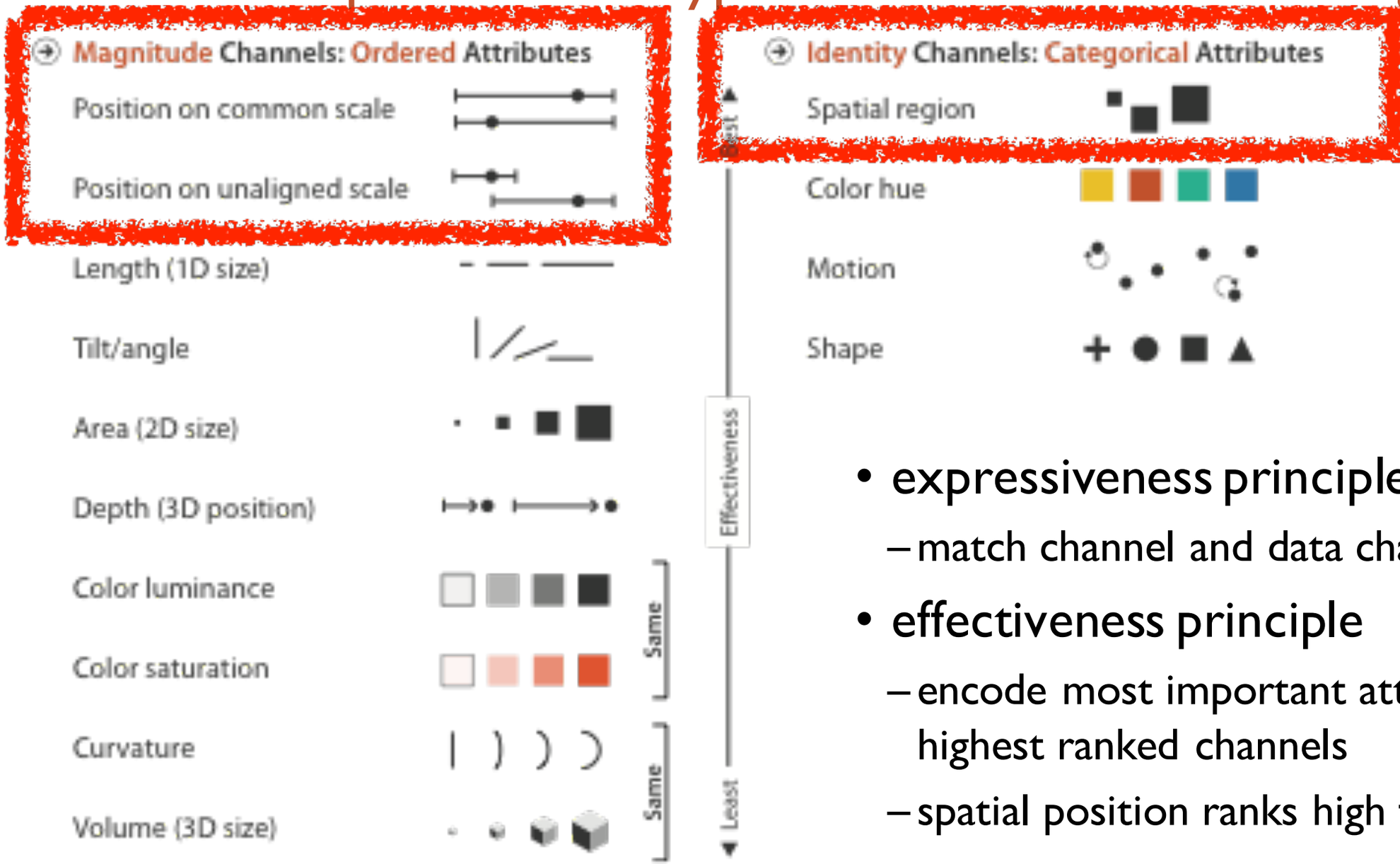
- expressiveness principle
  - match channel and data characteristics

Channels: Rankings



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

Channels: Expressiveness types and effectiveness rankings



- expressiveness principle
  - match channel and data characteristics
- effectiveness principle
  - encode most important attributes with highest ranked channels
  - spatial position ranks high for both

Accuracy: how to encode?

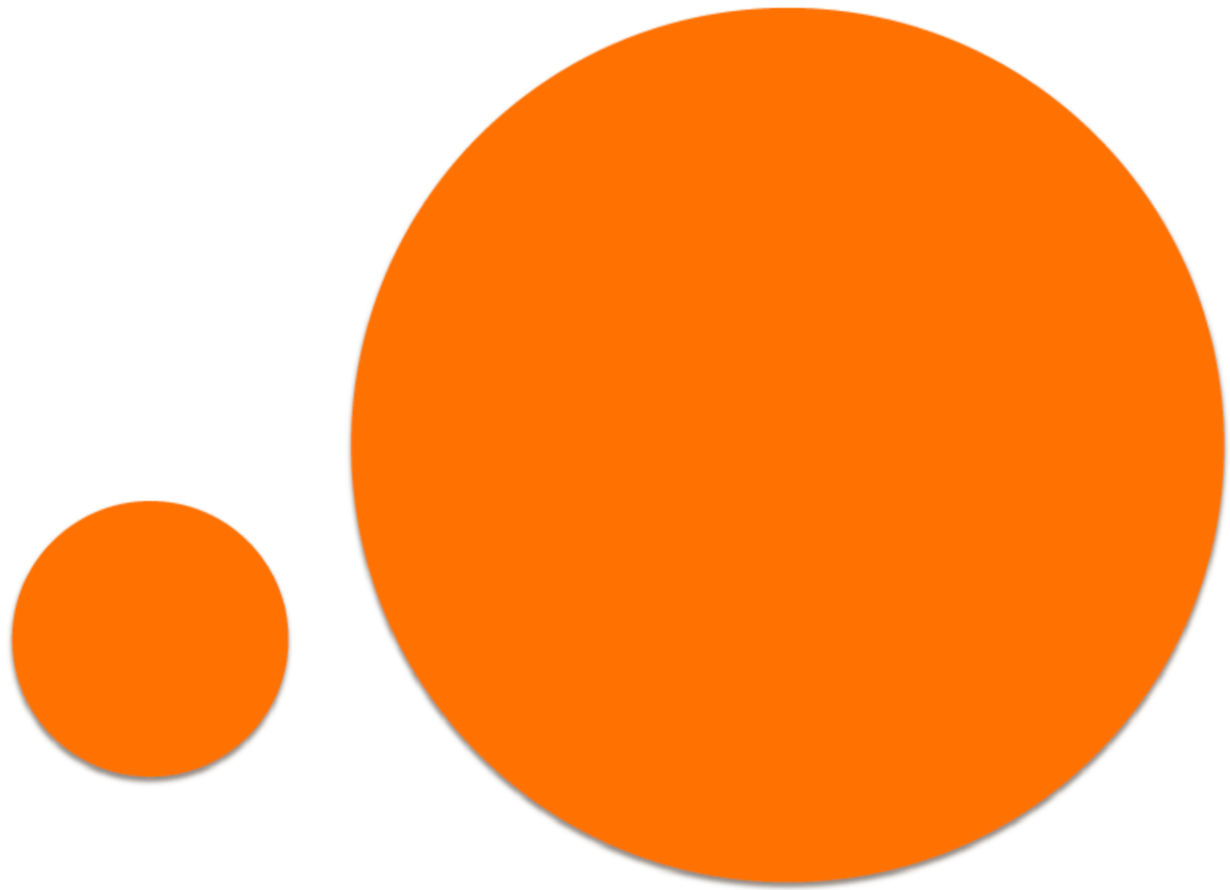
- How large is the second bar with respect to the first one?

Accuracy: how to encode?

- How large is the second bar with respect to the first one?

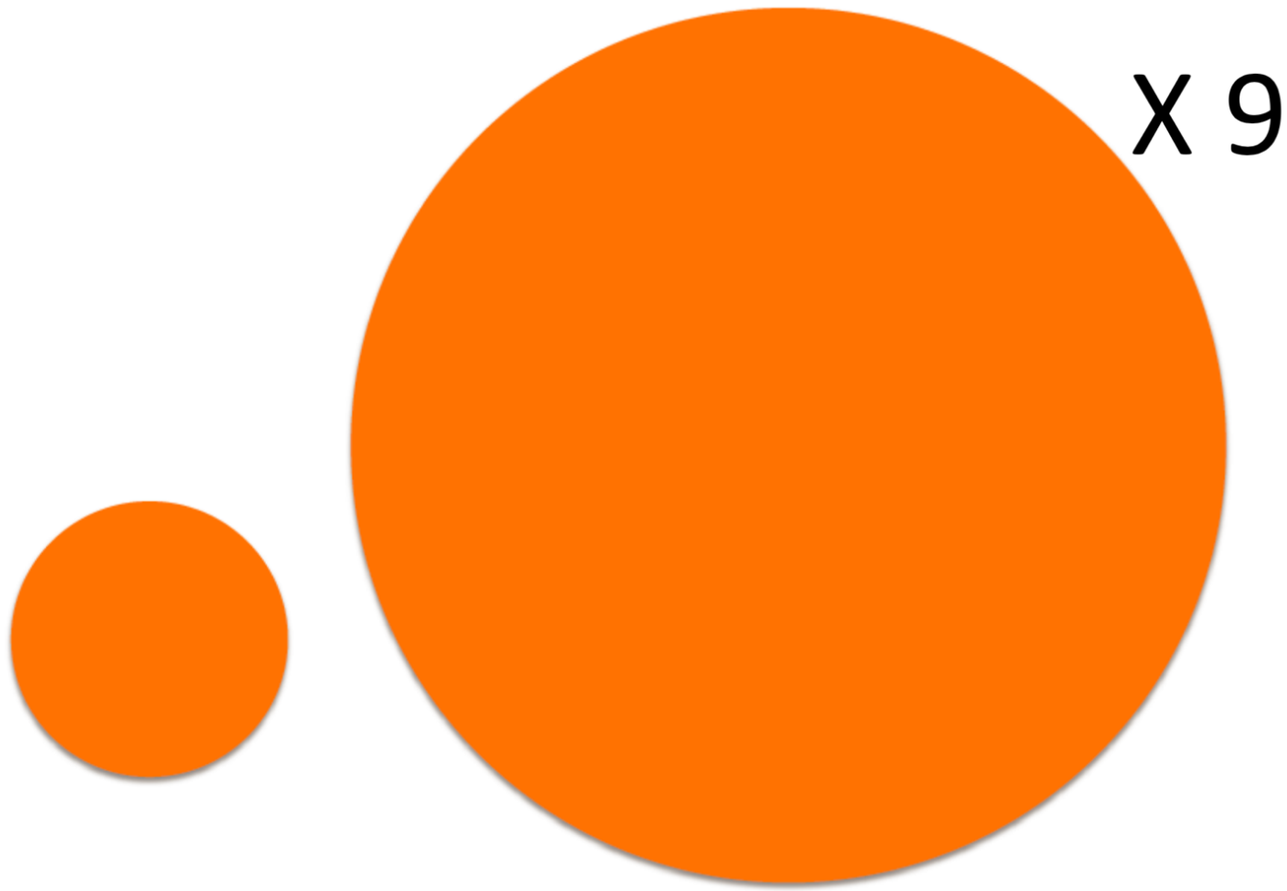
Accuracy: how to encode?

- How large is the second circle with respect to the first one?

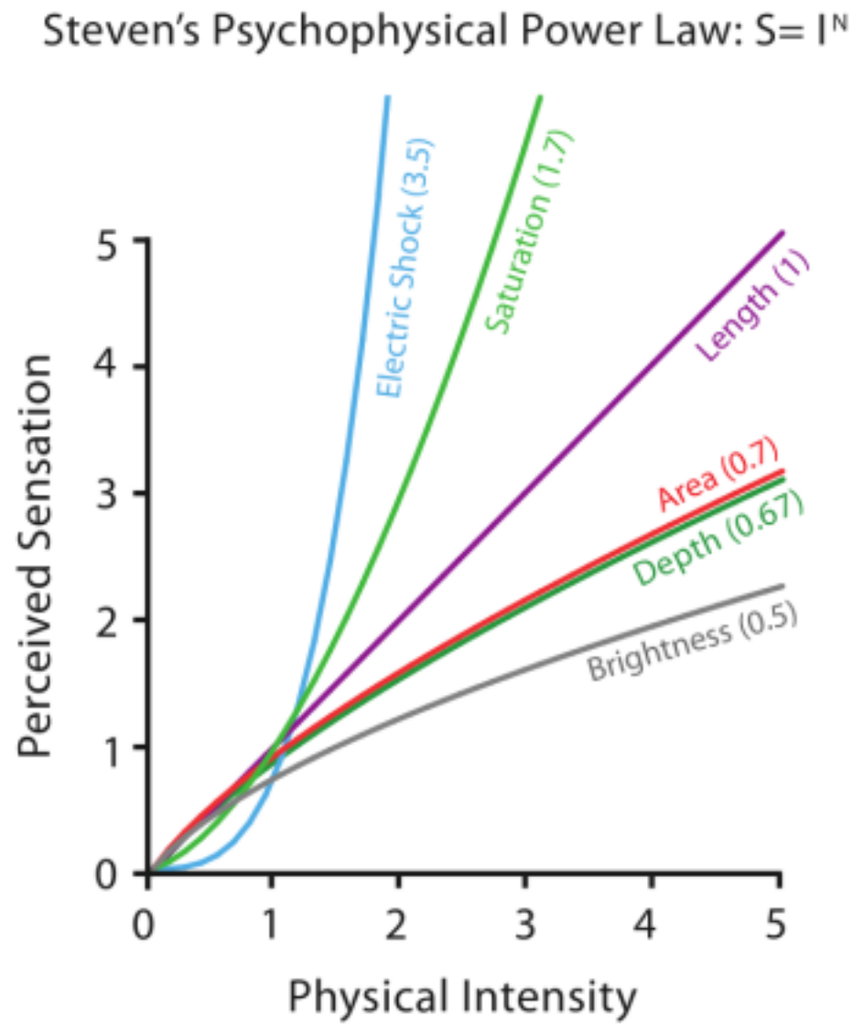


Accuracy: how to encode?

- How large is the second circle with respect to the first one?

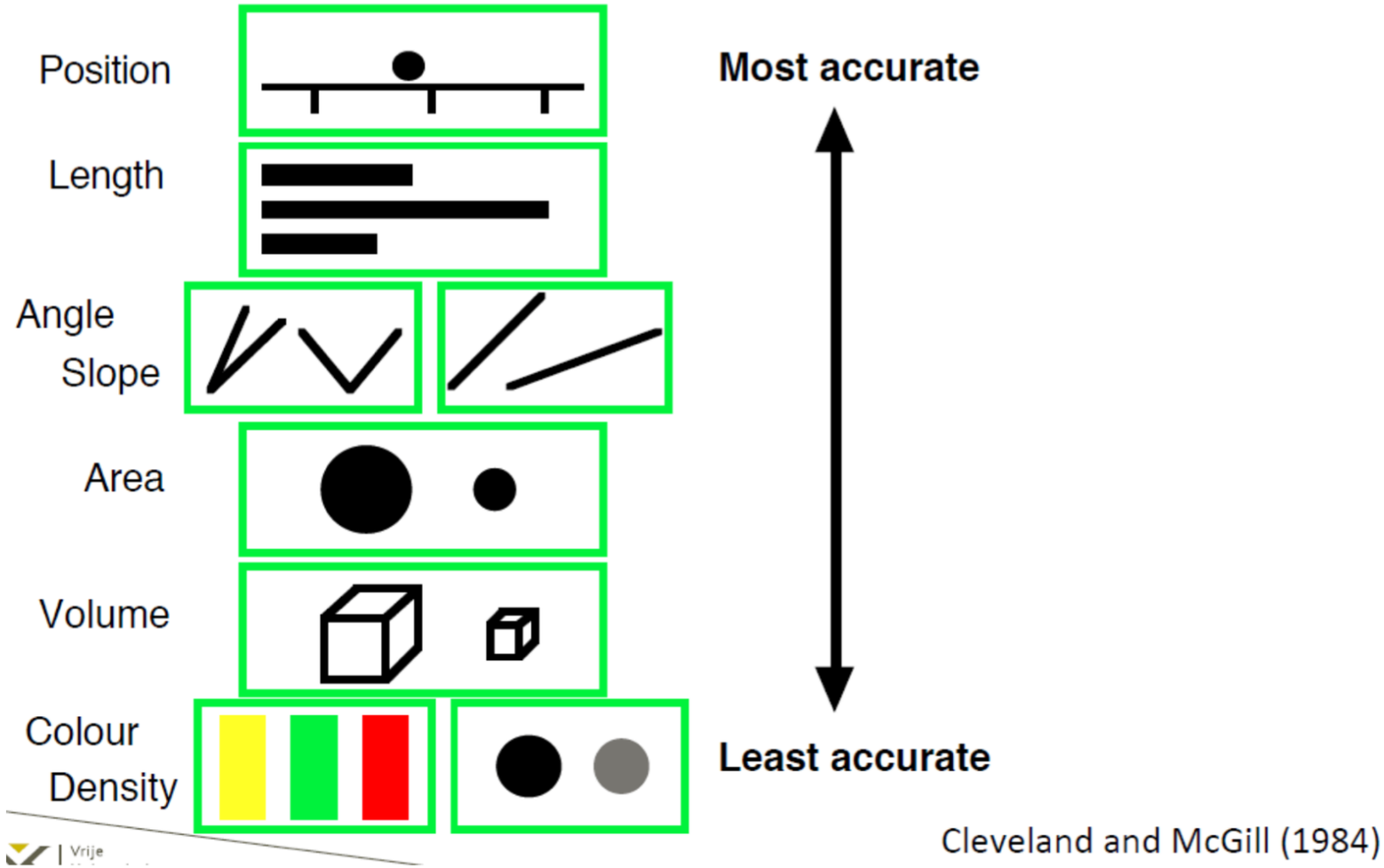


Accuracy: Fundamental Theory



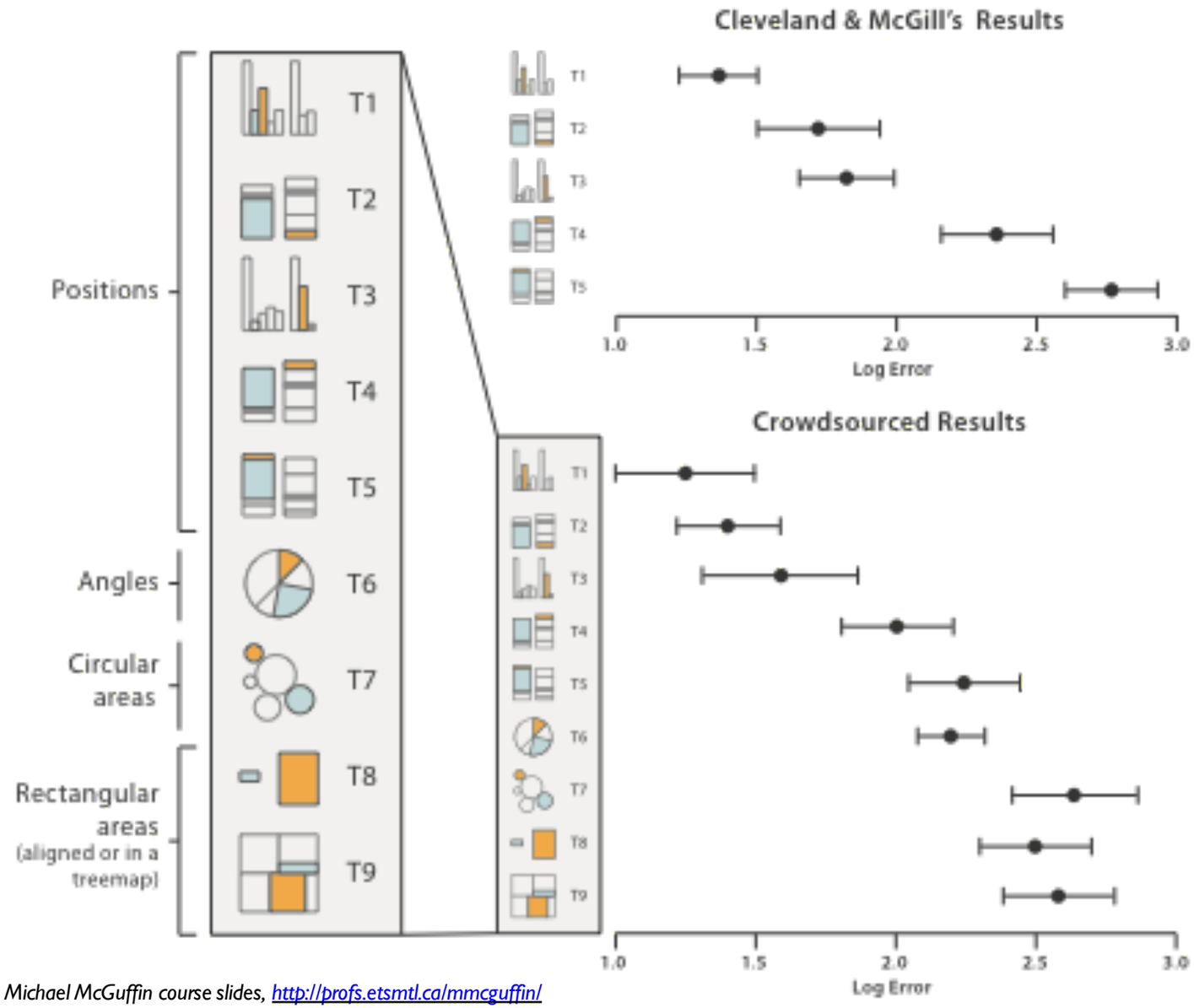
http://makingmaps.net/2007/08/28/perceptual-scaling-of-map-symbols/

What should we use to encode then ?



http://www.slideshare.net/kverbert/information-visualization-perception-and-principles

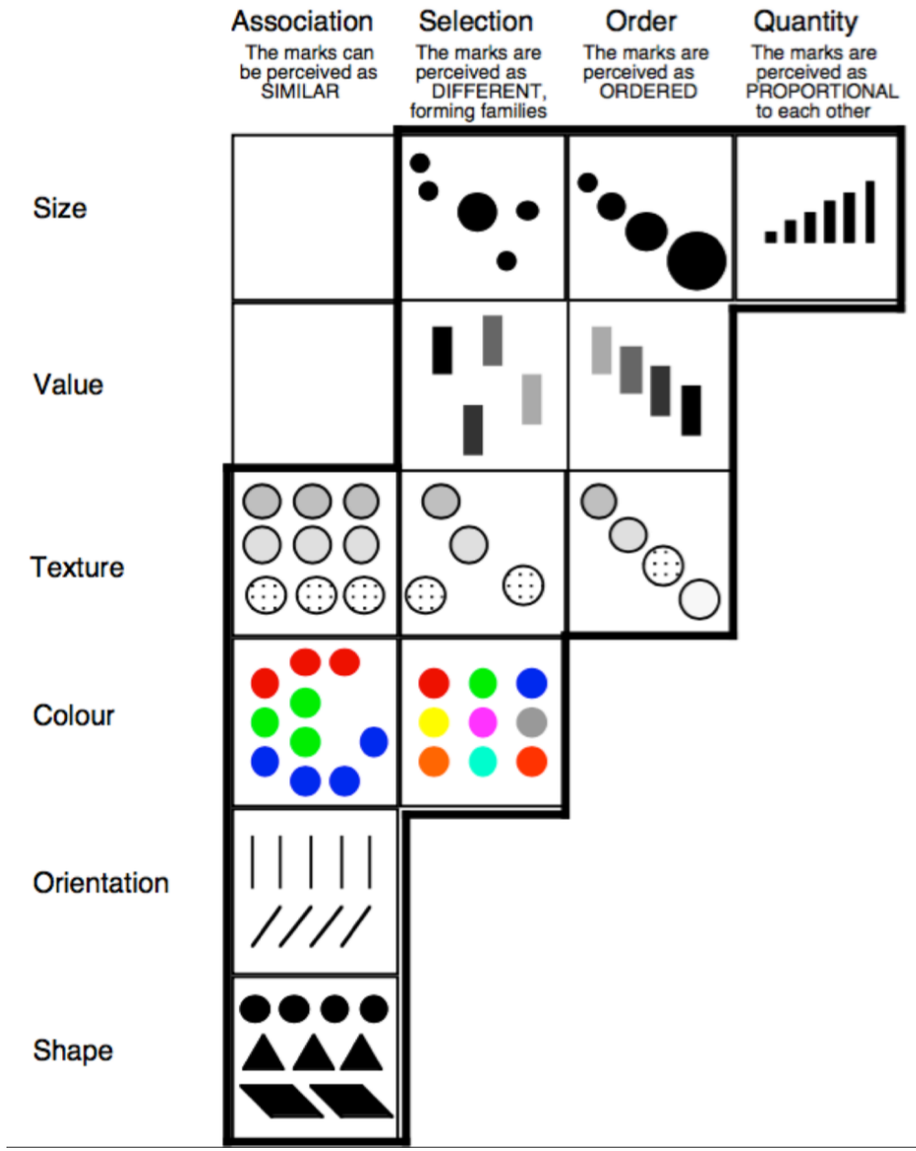
Accuracy: Vis experiments



after Michael McGuffin course slides, http://profs.etsmtl.ca/mmcguffin/

[Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. Heer and Bostock. Proc ACM Conf. Human Factors in Computing Systems (CHI) 2010, p. 203–212.]

J. Bertin guide for visual variables

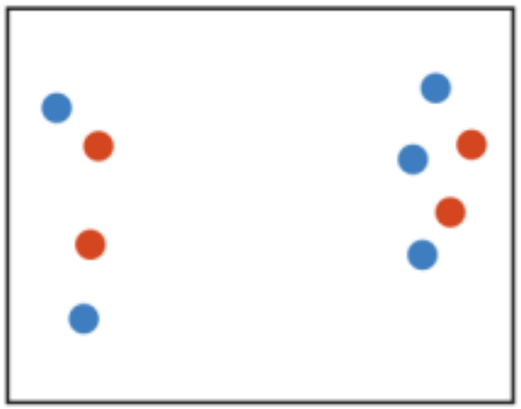


Choice of encoding

- Bertin's guidance
- suitability of various encoding methods
- to support common tasks

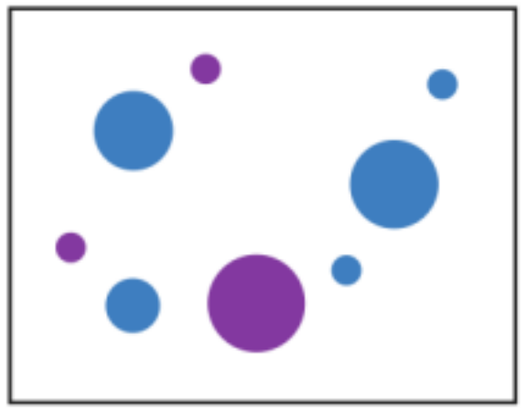
Separability vs. Integrality

Position
+ Hue (Color)



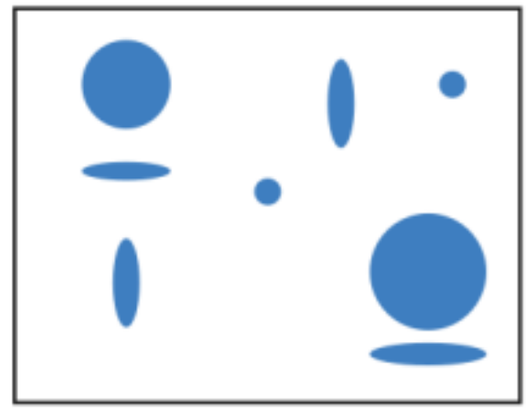
Fully separable

Size
+ Hue (Color)



Some interference

Width
+ Height



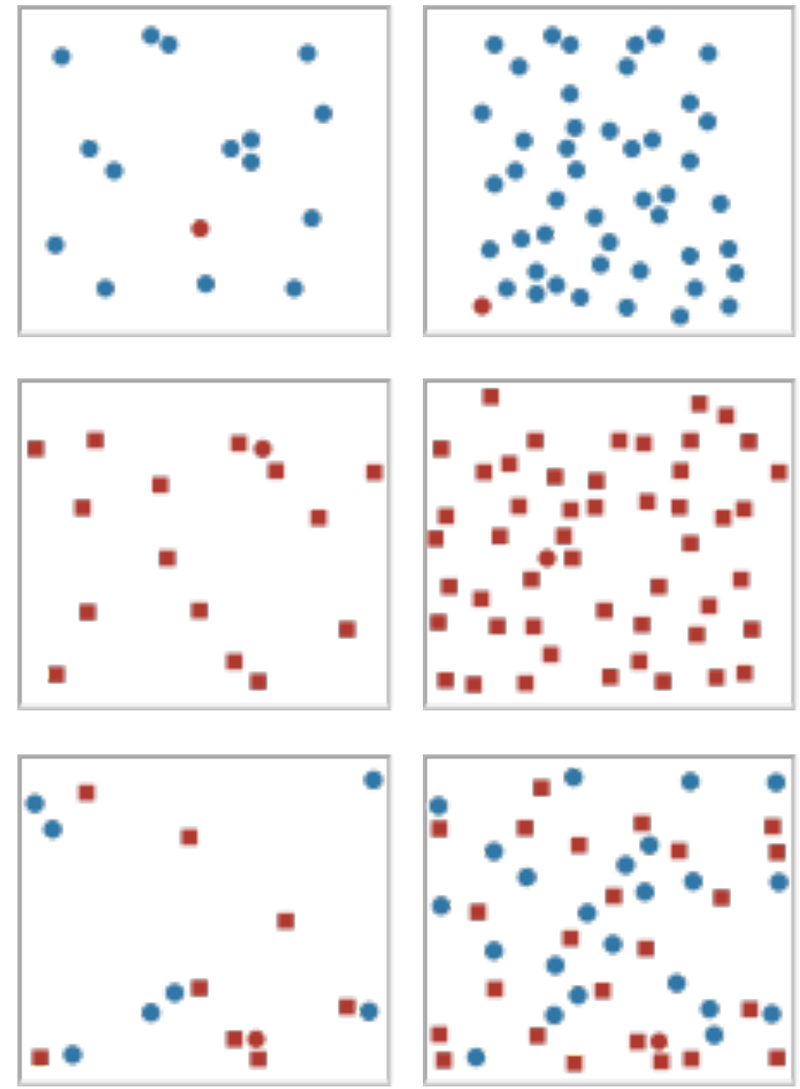
Some/significant interference

Red
+ Green

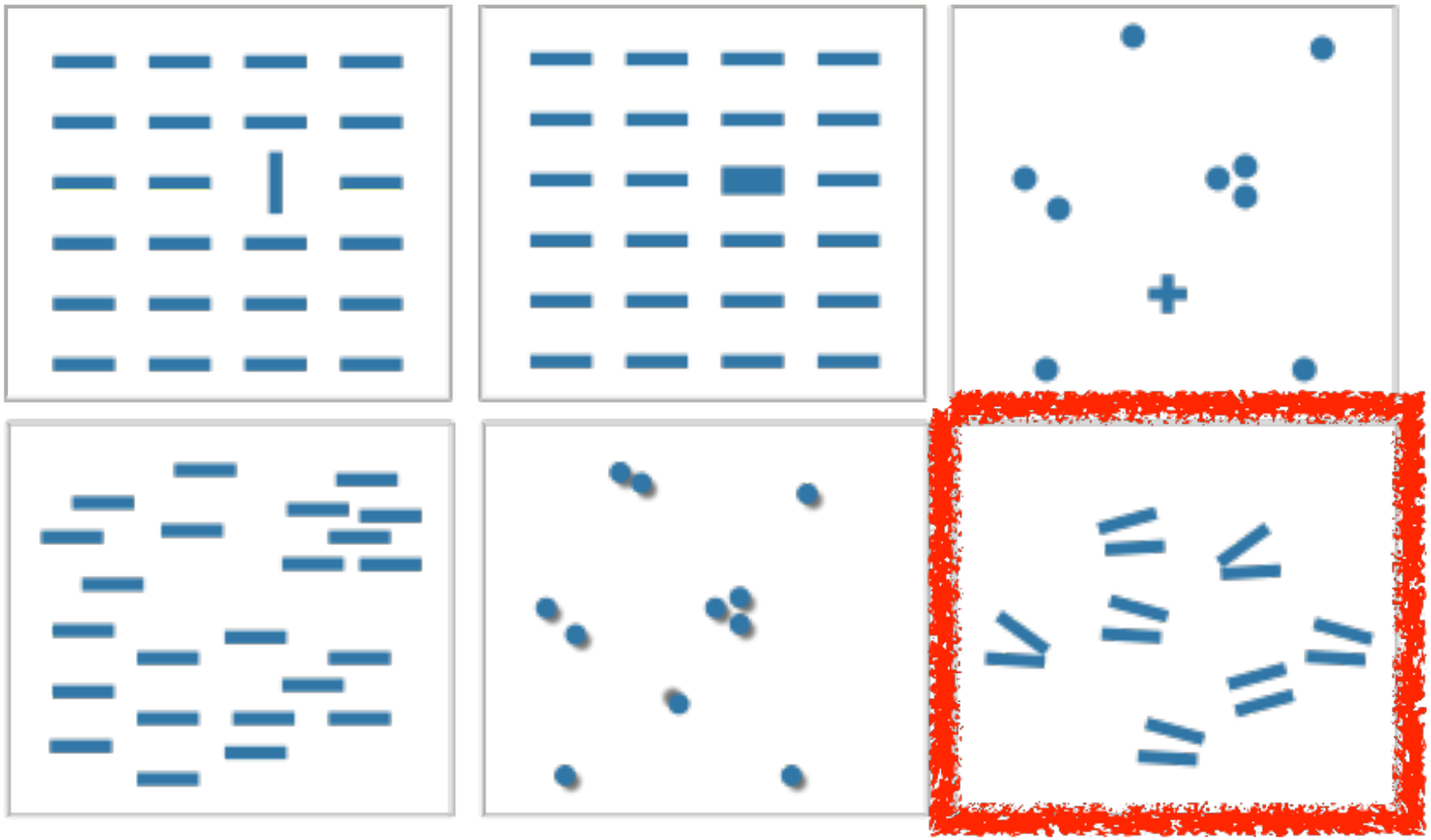
Major interference

Popout

- find the red dot
  - how long does it take?
- parallel processing on many individual channels
  - speed independent of distractor count
  - speed depends on channel and amount of difference from distractors
- serial search for (almost all) combinations
  - speed depends on number of distractors



Popout (pre-attentive processing)



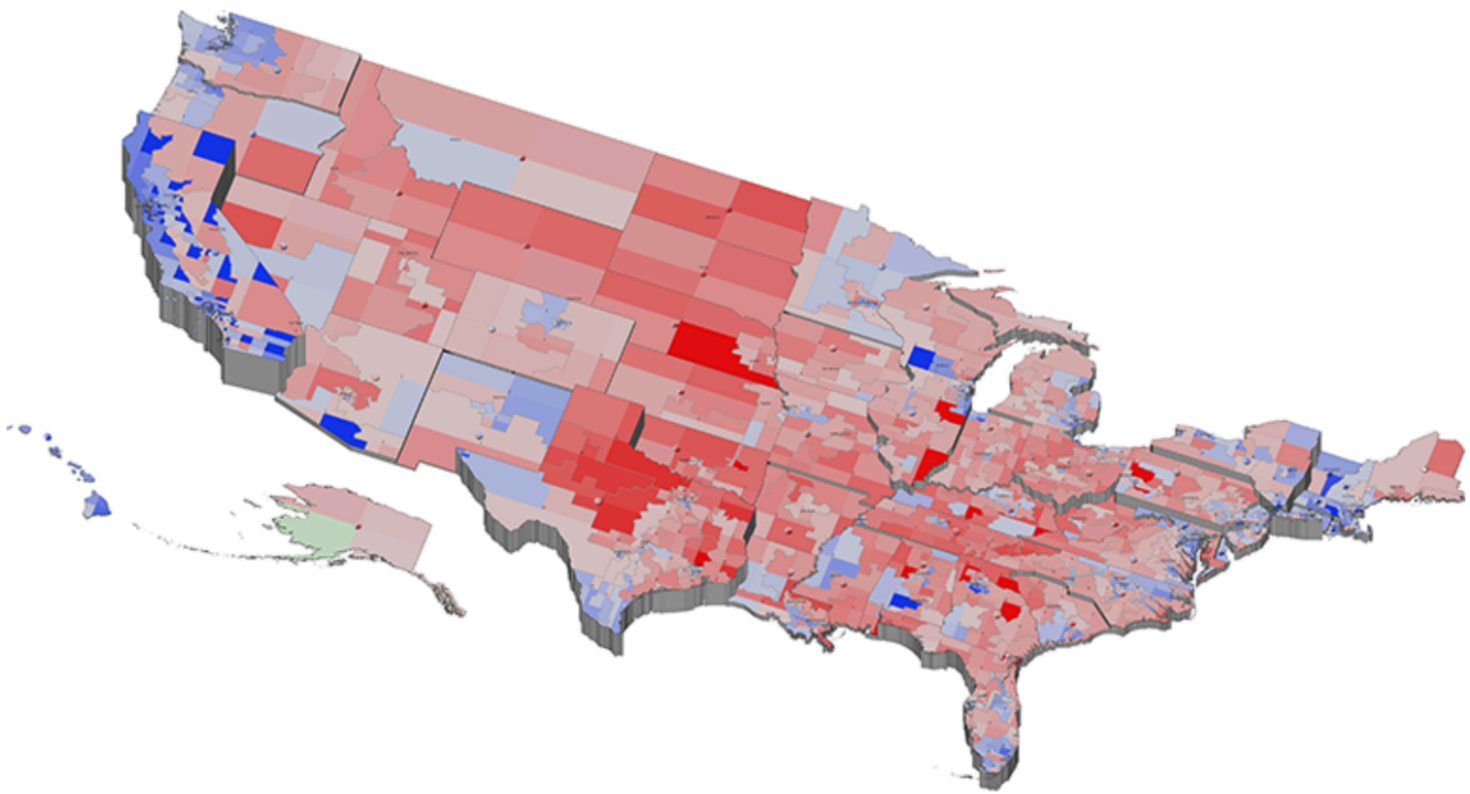
- many channels: tilt, size, shape, proximity, shadow direction, ...
- but not all! parallel line pairs do not pop out from tilted pairs

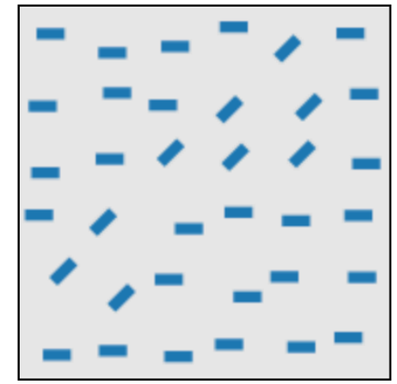
Further reading

- https://www.csc2.ncsu.edu/faculty/healey/PP/

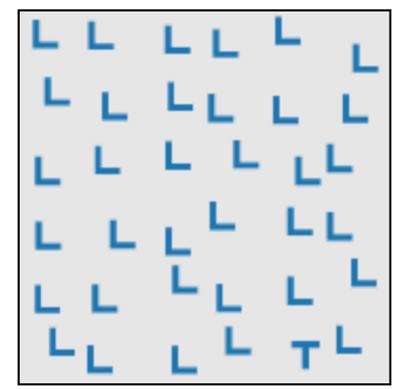
Perception in Visualization

Christopher G. Healey
Department of Computer Science, North Carolina State University





number, estimation
Sagi & Julész 85b; Healey et al. 93; Trick & Pylyshyn 94

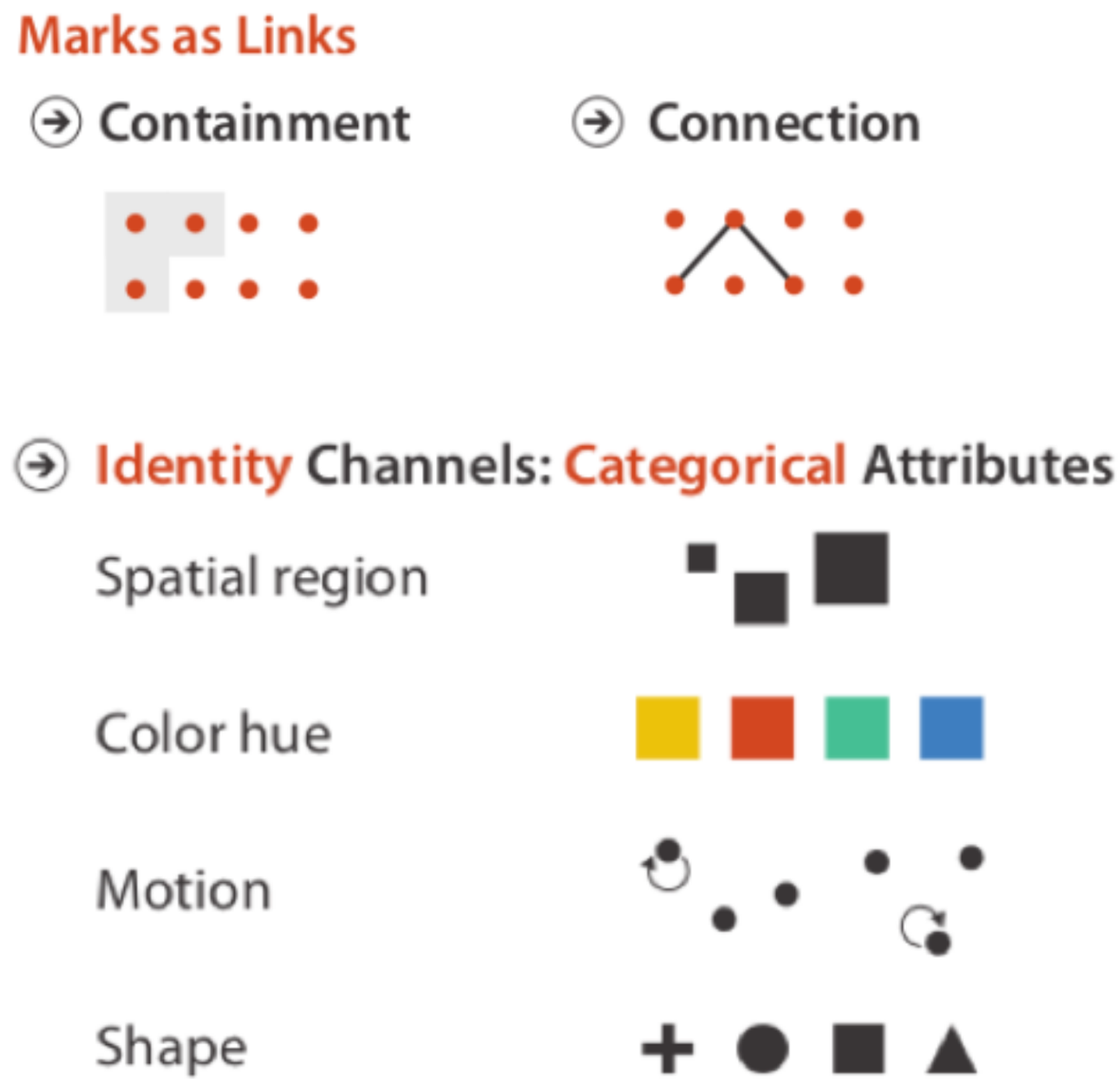


terminators
Julész & Bergen 83

Grouping

- containment
- connection

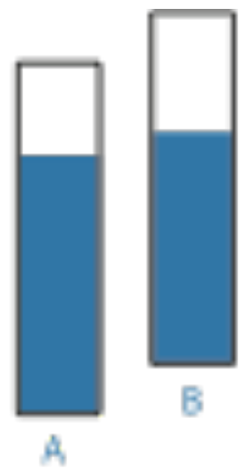
- proximity
  - same spatial region
- similarity
  - same values as other categorical channels



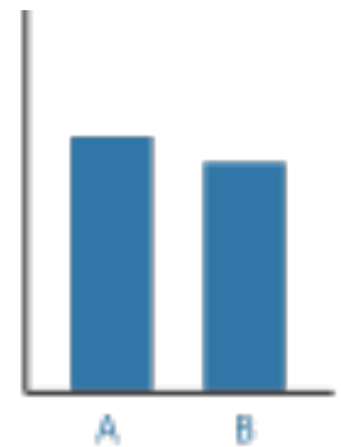
Relative vs. absolute judgements

- perceptual system mostly operates with relative judgements, not absolute
  - that's why accuracy increases with common frame/scale and alignment

length



position along unaligned common scale

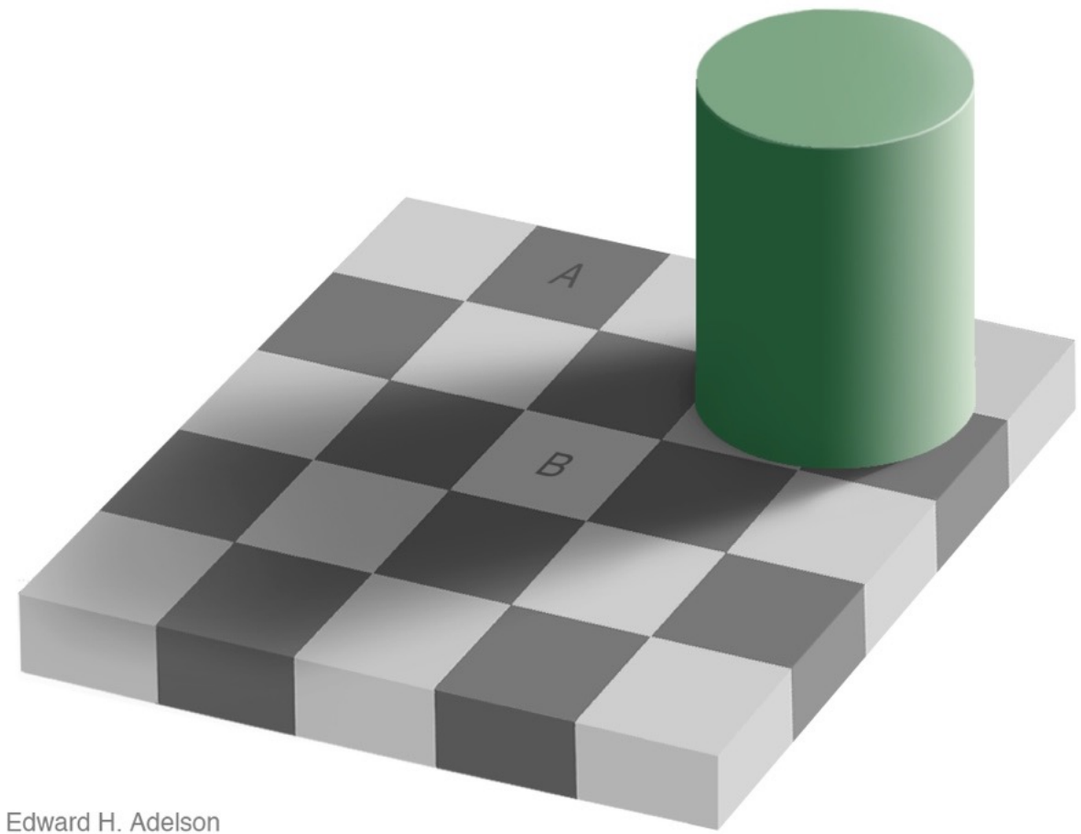


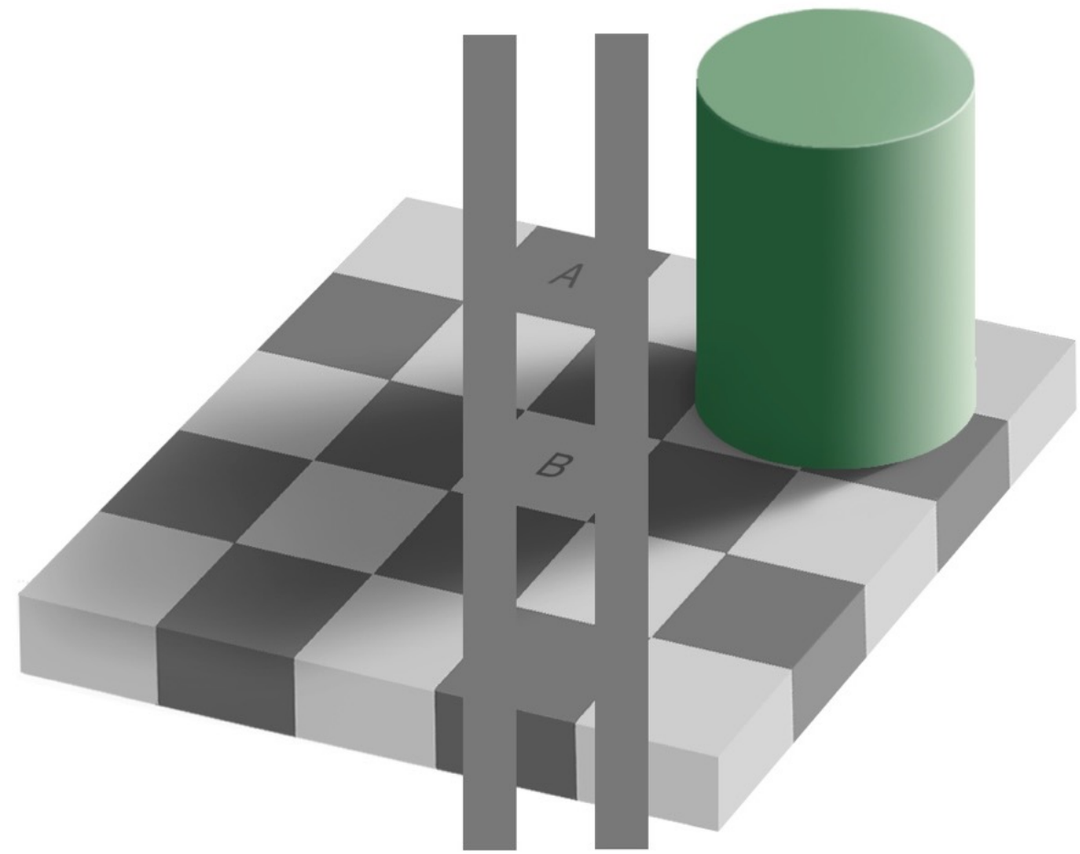
position along aligned scale

after [Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Cleveland and McGill. Journ. American Statistical Association 79:387 (1984), 531–554.]

Relative luminance judgements

- perception of luminance is contextual based on contrast with surroundings





http://persci.mit.edu/gallery/checkershadow

Further reading

- Visualization Analysis and Design. Munzner. AK Peters Visualization Series, CRC Press, 2014.
  - *Chap 5: Marks and Channels*
- *On the Theory of Scales of Measurement.* Stevens. Science 103:2684 (1946), 677–680.
- Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. Stevens. Wiley, 1975.
- *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods.* Cleveland and McGill. Journ. American Statistical Association 79:387 (1984), 531–554.
- *Perception in Vision*. Healey. http://www.csc.ncsu.edu/faculty/healey/PP
- Visual Thinking for Design. Ware. Morgan Kaufmann, 2008.
- Information Visualization: Perception for Design, 3rd edition. Ware. Morgan Kaufmann /Academic Press, 2004.

Outline

- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - **Arrange Tables**
  - Hands on: Altair
  - Rules of Thumb & Vis Research

How?

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

Arrange tables

Express Values

Separate, Order, Align Regions

- Separate

- Order

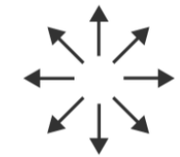
- Align

Axis Orientation

- Rectilinear

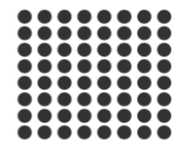
- Parallel

- Radial



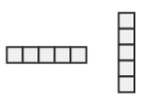
Layout Density

- Dense



- Space-Filling

- *1 Key*
  List



- *2 Keys*
  Matrix

- *3 Keys*
  Volume

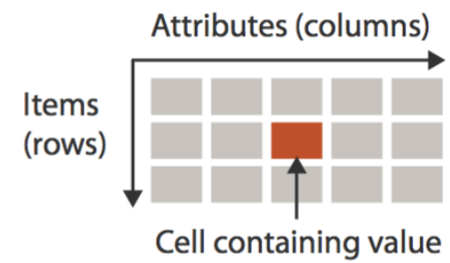
- *Many Keys*
  Recursive Subdivision

5

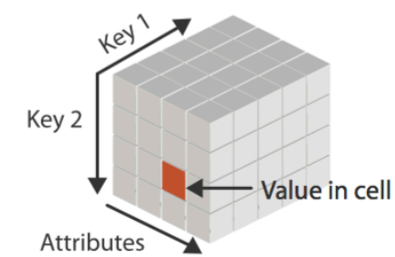
Keys and values

- key
  - independent attribute
  - used as unique index to look up items
  - simple tables: 1 key
  - multidimensional tables: multiple keys
- value
  - dependent attribute, value of cell
- classify arrangements by key count
  - 0, 1, 2, many...

➔ Tables

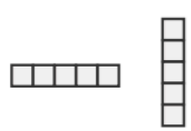


➔ *Multidimensional Table*

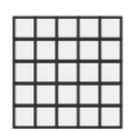


➔ **Express Values**

➔ *1 Key*
List



➔ *2 Keys*
Matrix

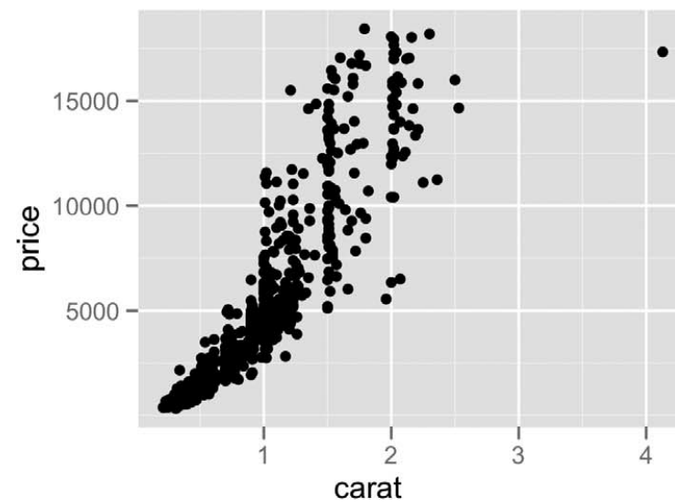


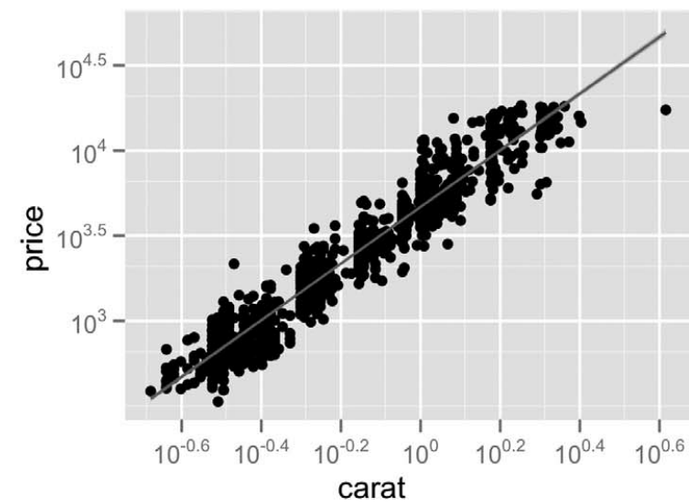
➔ *3 Keys*
Volume

➔ *Many Keys*
Recursive Subdivision

Idiom: **scatterplot**

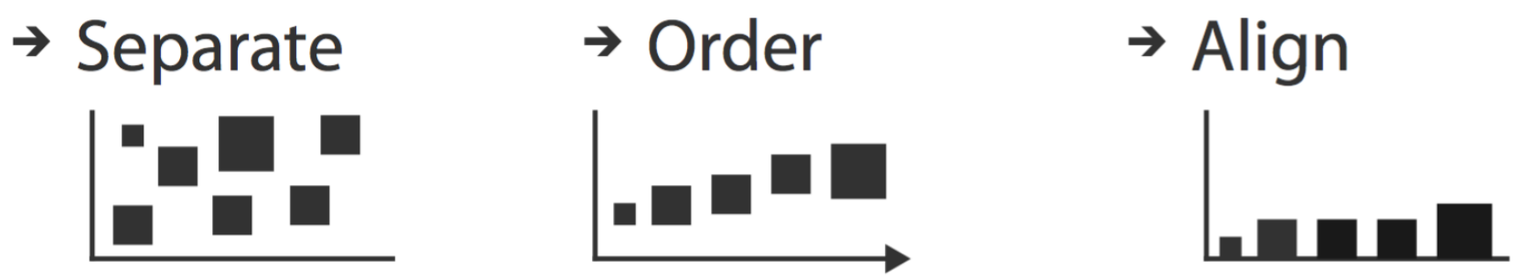
- ***express*** values
  - quantitative attributes
- no keys, only values
  - data
    - 2 quant attribs
  - mark: points
  - channels
    - horiz + vert position
  - tasks
    - find trends, outliers, distribution, correlation, clusters
  - scalability
    - hundreds of items



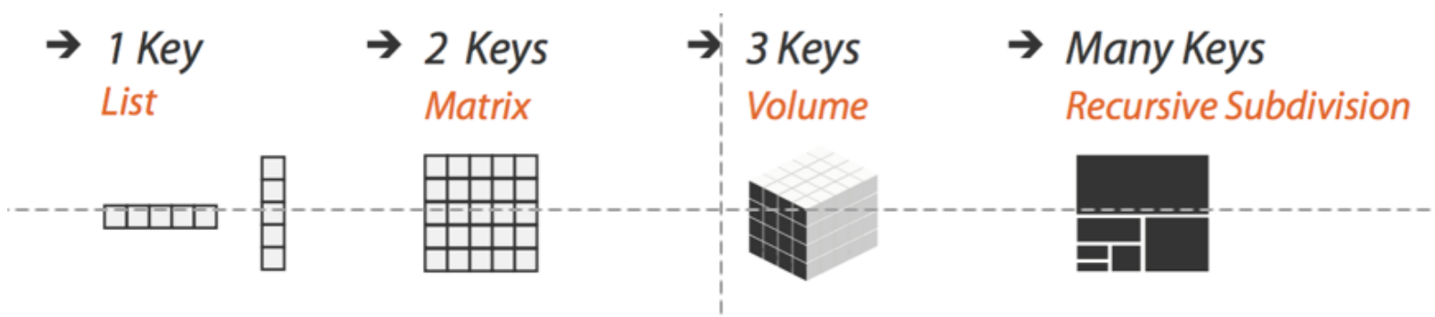


[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.]

Some keys: Categorical regions

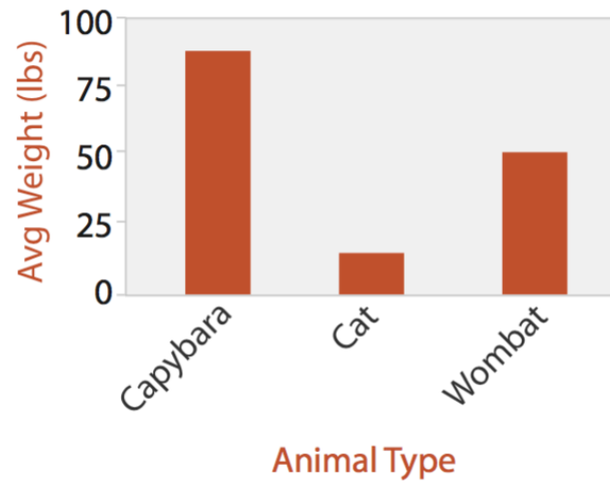


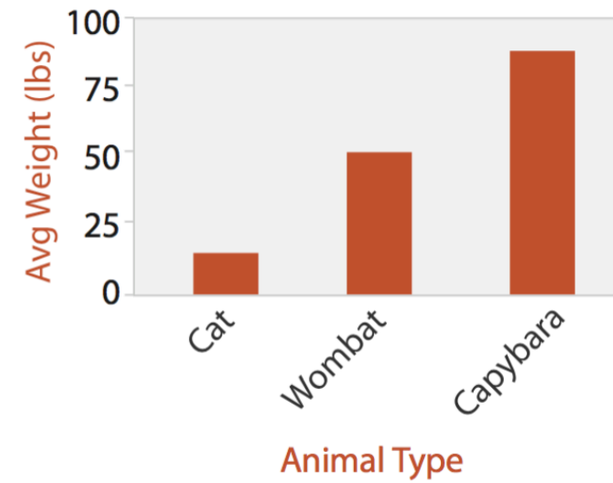
- **regions**: contiguous bounded areas distinct from each other
  - using space to ***separate*** (proximity)
  - following expressiveness principle for categorical attributes
- use ordered attribute to ***order*** and ***align*** regions



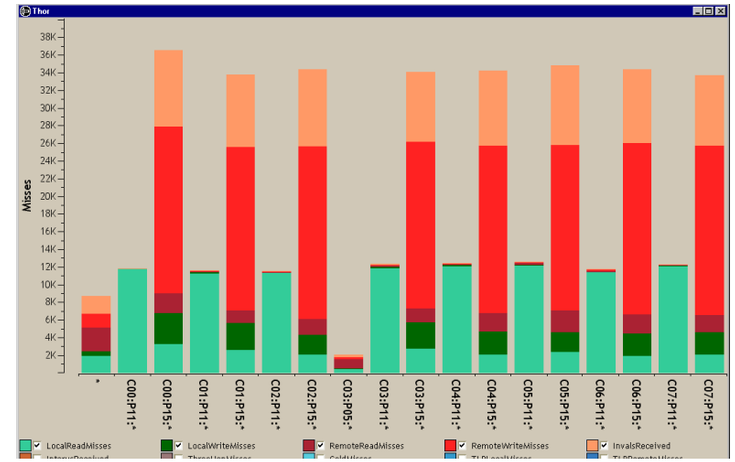
Idiom: **bar chart**

- one key, one value
  - data
    - 1 categ attrib, 1 quant attrib
  - mark: lines
  - channels
    - length to express quant value
    - spatial regions: one per mark
      - separated horizontally, aligned vertically
      - ordered by quant attrib
        - » by label (alphabetical), by length attrib (data-driven)
  - task
    - compare, lookup values
  - scalability
    - dozens to hundreds of levels for key attrib





Idiom: **stacked bar chart**

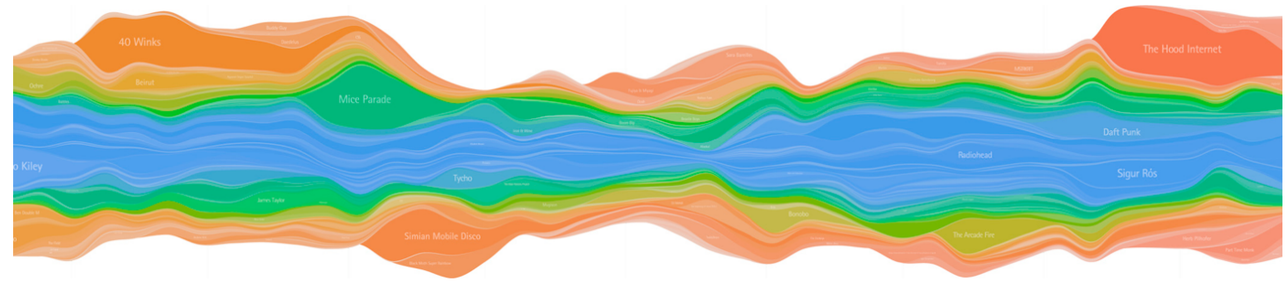


- one more key
  - data
    - 2 categ attrib, 1 quant attrib
  - mark: vertical stack of line marks
    - **glyph**: composite object, internal structure from multiple marks
  - channels
    - length and color hue
    - spatial regions: one per glyph
      - aligned: full glyph, lowest bar component
      - unaligned: other bar components
  - task
    - part-to-whole relationship
  - scalability
    - several to one dozen levels for stacked attrib

[Using Visualization to Understand the Behavior of Computer Systems. Bosch. Ph.D. thesis, Stanford Computer Science, 2001.]

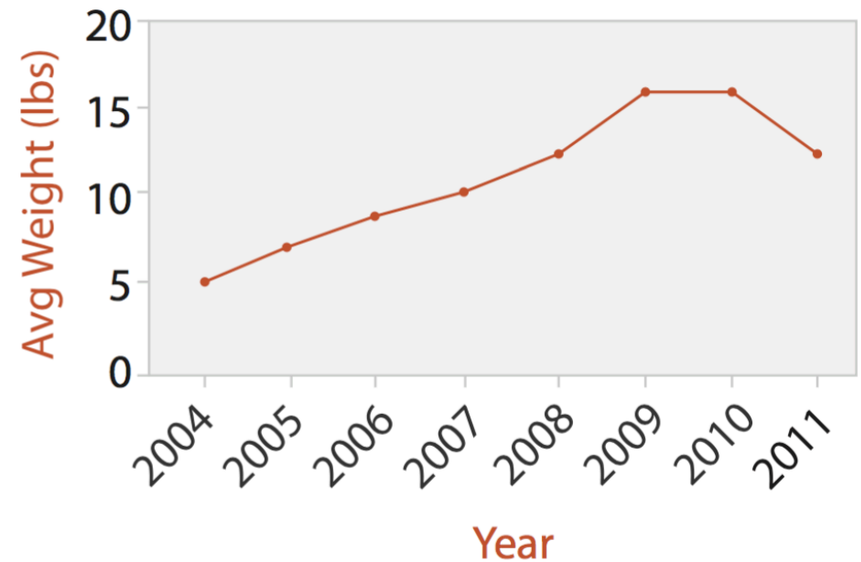
Idiom: **streamgraph**

- generalized stacked graph
  - emphasizing horizontal continuity
    - vs vertical items
  - data
    - 1 categ key attrib (artist)
    - 1 ordered key attrib (time)
    - 1 quant value attrib (counts)
  - derived data
    - geometry: layers, where height encodes counts
    - 1 quant attrib (layer ordering)
  - scalability
    - hundreds of time keys
    - dozens to hundreds of artist keys
      - more than stacked bars, since most layers don't extend across whole chart



[Stacked Graphs Geometry & Aesthetics. Byron and Wattenberg. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2008) 14(6): 1245–1252, (2008).]

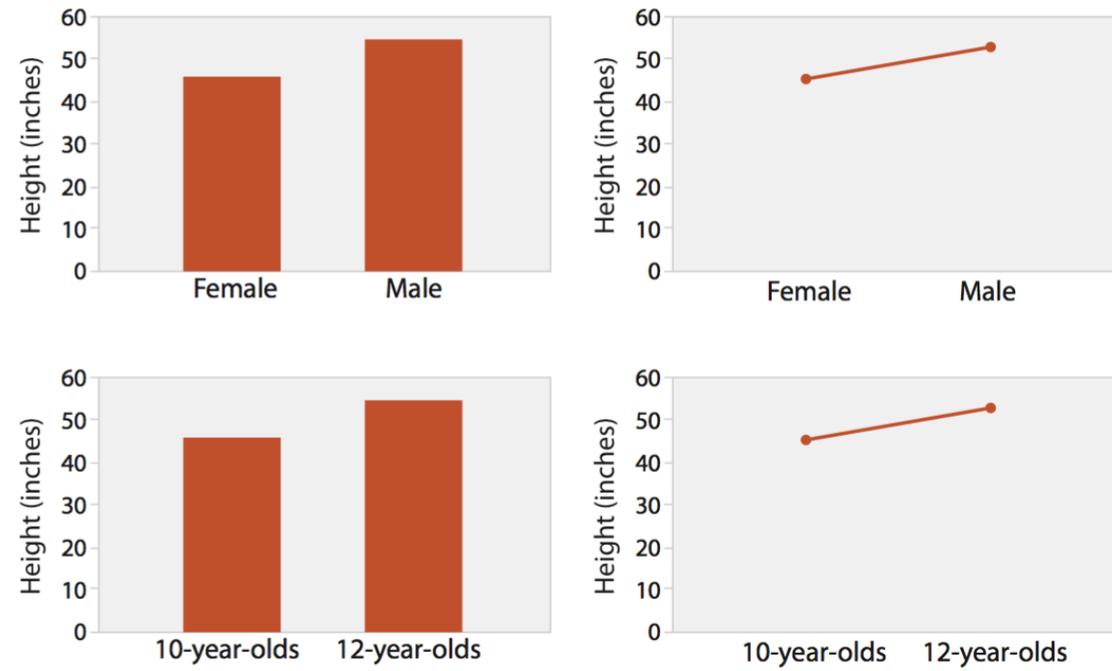
Idiom: **line chart**



- one key, one value
  - data
    - 2 quant attribs
  - mark: points
    - line connection marks between them
  - channels
    - aligned lengths to express quant value
    - separated and ordered by key attrib into horizontal regions
  - task
    - find trend
      - connection marks emphasize ordering of items along key axis by explicitly showing relationship between one item and the next

Choosing bar vs line charts

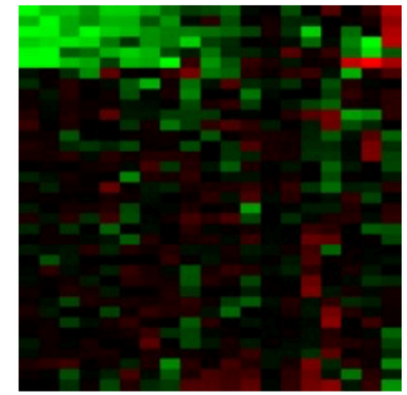
- depends on type of key attrib
  - bar charts if categorical
  - line charts if ordered
- do not use line charts for categorical key attribs
  - violates expressiveness principle
    - implication of trend so strong that it overrides semantics!
      - "The more male a person is, the taller he/she is"

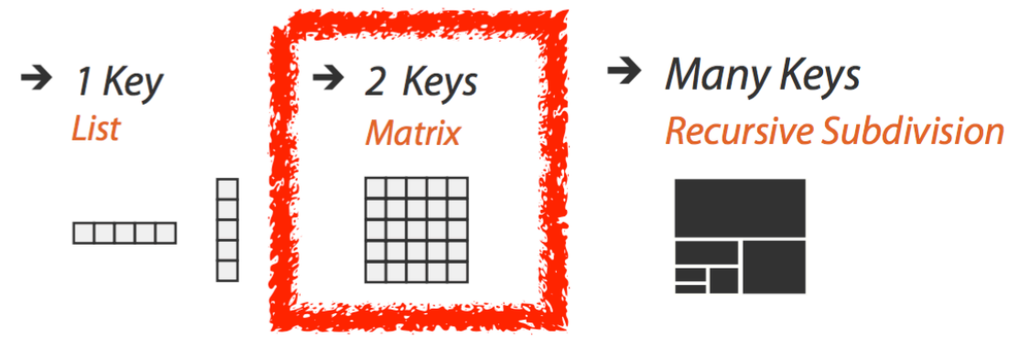


after [Bars and Lines: A Study of Graphic Communication. Zacks and Tversky. Memory and Cognition 27:6 (1999), 1073–1079.]

Idiom: **heatmap**

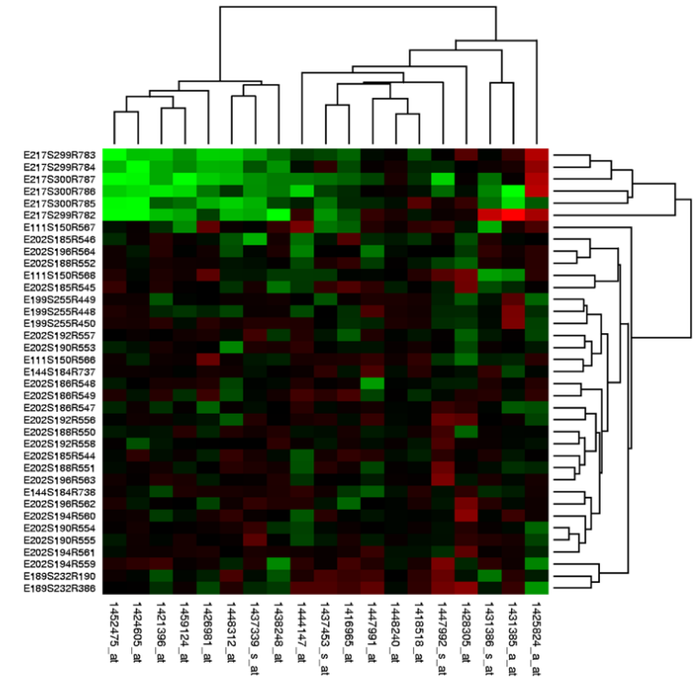
- two keys, one value
  - data
    - 2 categ attribs (gene, experimental condition)
    - 1 quant attrib (expression levels)
  - marks: area
    - separate and align in 2D matrix
      - indexed by 2 categorical attributes
  - channels
    - color by quant attrib
      - (ordered diverging colormap)
  - task
    - find clusters, outliers
  - scalability
    - 1M items, 100s of categ levels, ~10 quant attrib levels





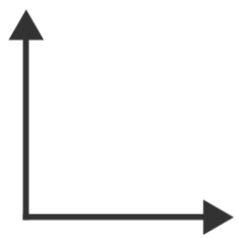
Idiom: **cluster heatmap**

- in addition
  - derived data
    - 2 cluster hierarchies
  - dendrogram
    - parent-child relationships in tree with connection line marks
    - leaves aligned so interior branch heights easy to compare
  - heatmap
    - marks (re-)ordered by cluster hierarchy traversal

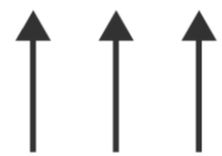


Axis Orientation

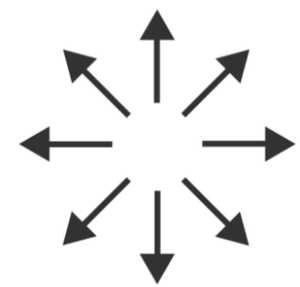
- Rectilinear



- Parallel



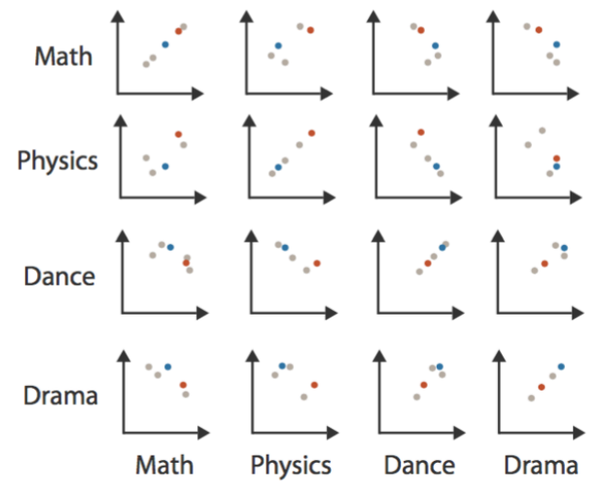
- Radial



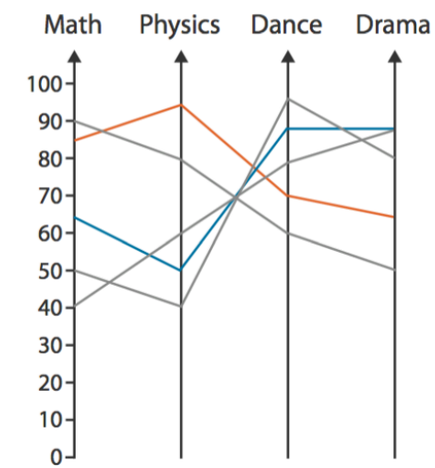
Idioms: **scatterplot matrix, parallel coordinates**

- scatterplot matrix (SPLOM)
  - rectilinear axes, point mark
  - all possible pairs of axes
  - scalability
    - one dozen attribs
    - dozens to hundreds of items
- parallel coordinates
  - parallel axes, jagged line representing item
  - rectilinear axes, item as point
    - axis ordering is major challenge
  - scalability
    - dozens of attribs
    - hundreds of items

Scatterplot Matrix



Parallel Coordinates



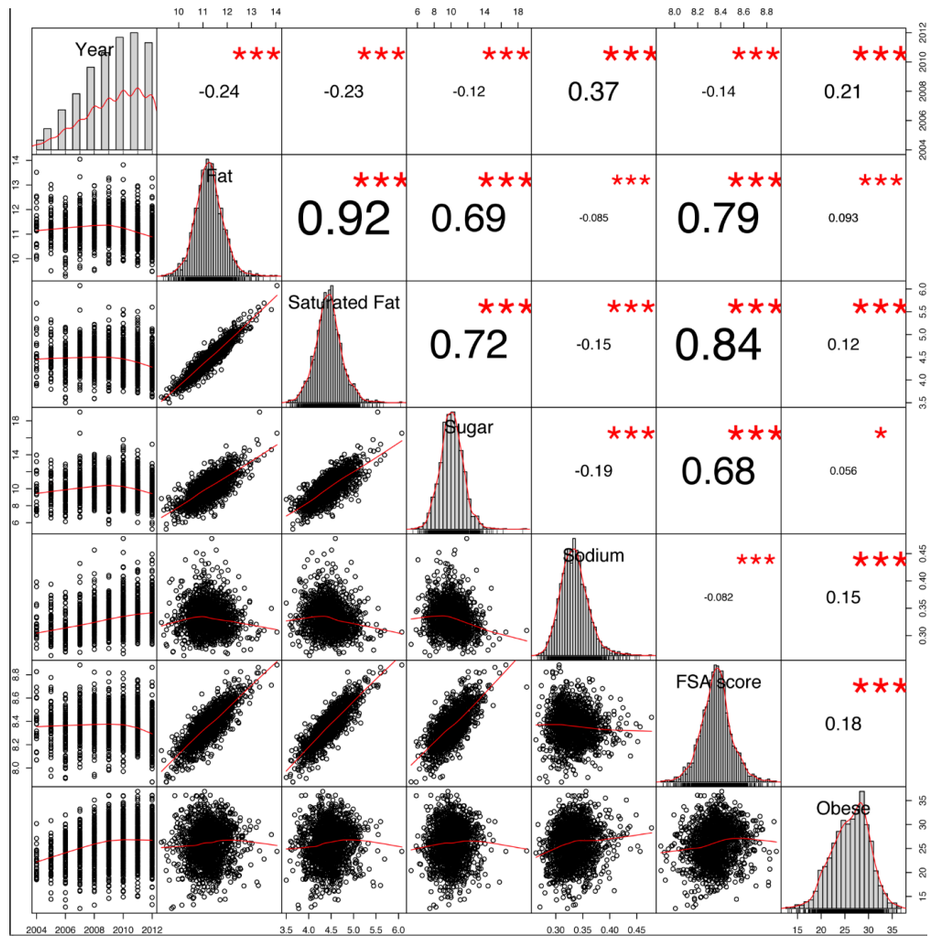
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

after [Visualization Course Figures. McGuffin, 2014. http://www.michaelmcguffin.com/courses/vis/]

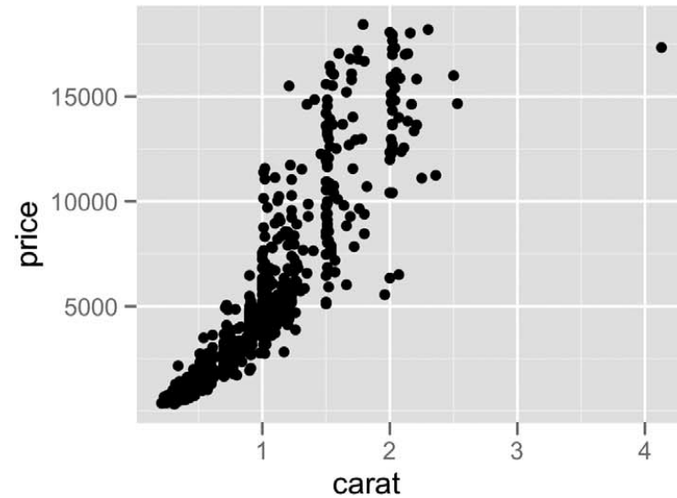
Scatterplot Matrix

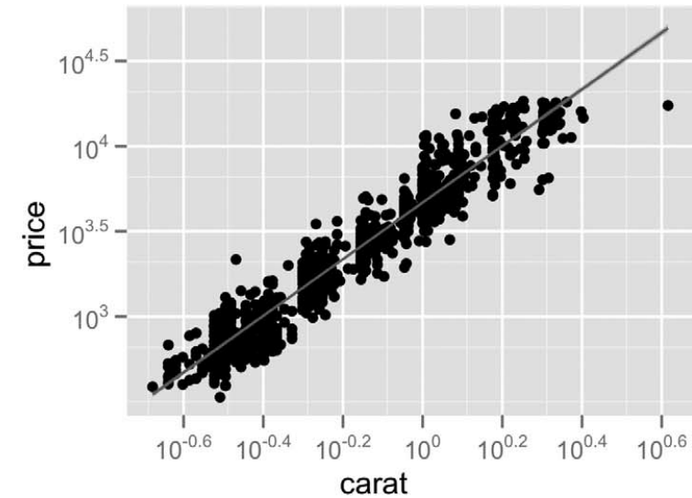
- Trattner, C., Parra, D., & Elsweiler, D. (2017). Monitoring obesity prevalence in the United States through bookmarking activities in online food portals. *PloS one*.



Task: Correlation

- scatterplot matrix
  - positive correlation
    - diagonal low-to-high
  - negative correlation
    - diagonal high-to-low
  - uncorrelated
- parallel coordinates
  - positive correlation
    - parallel line segments
  - negative correlation
    - all segments cross at halfway point
  - uncorrelated
    - scattered crossings





[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.]

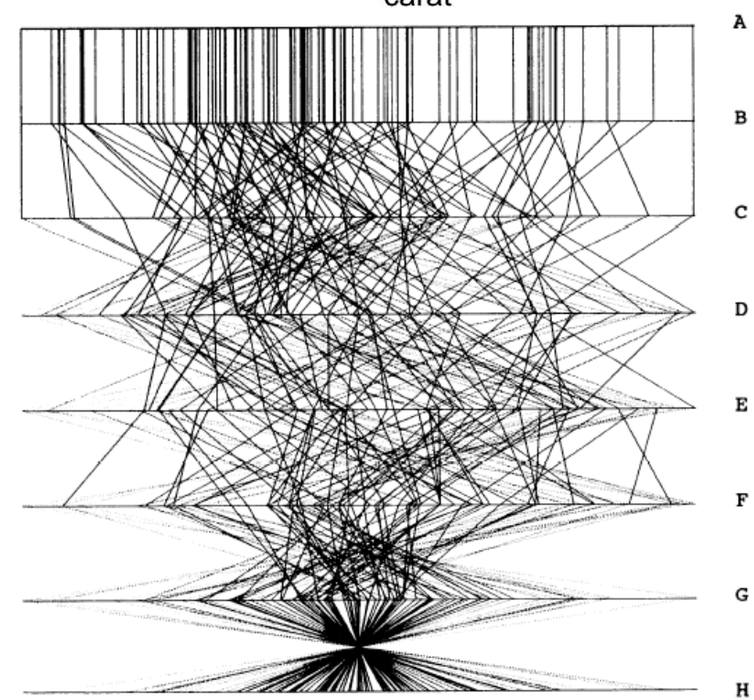
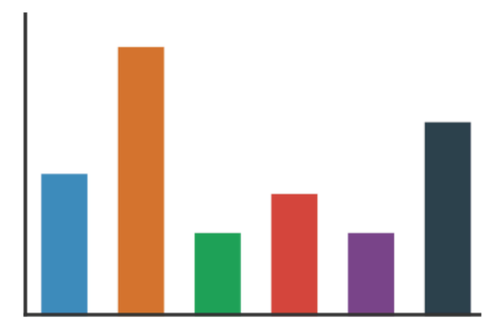


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of ρ = 1, .8, .2, 0, −.2, −.8, and −1.

[Hyperdimensional Data Analysis Using Parallel Coordinates. Wegman. Journ. American Statistical Association 85:411 (1990), 664–675.]

Idioms: **radial bar chart, star plot**

- radial bar chart
  - radial axes meet at central ring, line mark
- star plot
  - radial axes, meet at central point, line mark
- bar chart
  - rectilinear axes, aligned vertically

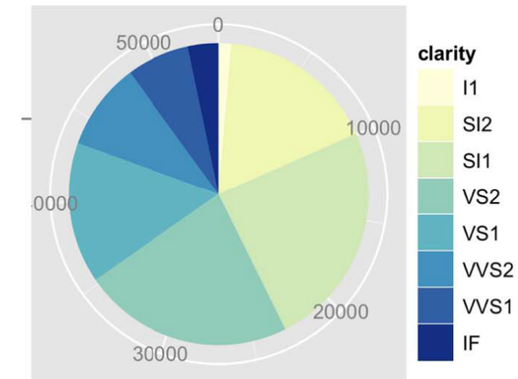


- accuracy
  - length unaligned with radial
    - less accurate than aligned with rectilinear

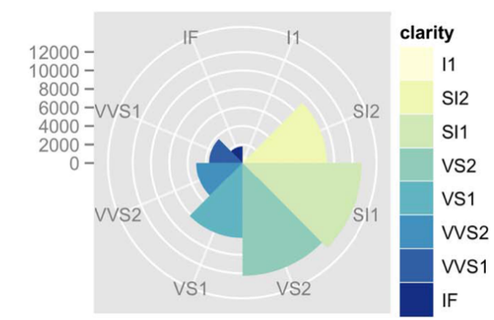
[Vismon: Facilitating Risk Assessment and Decision Making In Fisheries Management. Booshehrian, Möller, Peterman, and Munzner. Technical Report TR 2011-04, Simon Fraser University, School of Computing Science, 2011.]

Idioms: **pie chart, polar area chart**

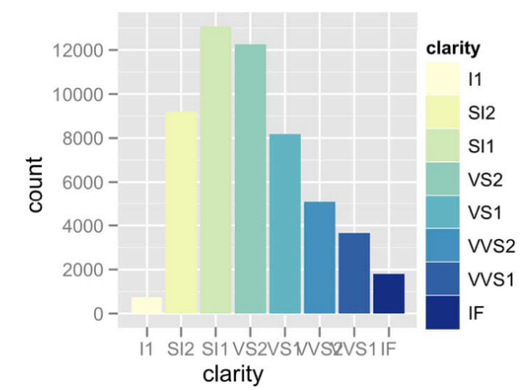
- pie chart
  - area marks with angle channel
  - accuracy: angle/area much less accurate than line length



- polar area chart
  - area marks with length channel
  - more direct analog to bar charts



- data
  - 1 categ key attrib, 1 quant value attrib
- task
  - part-to-whole judgements



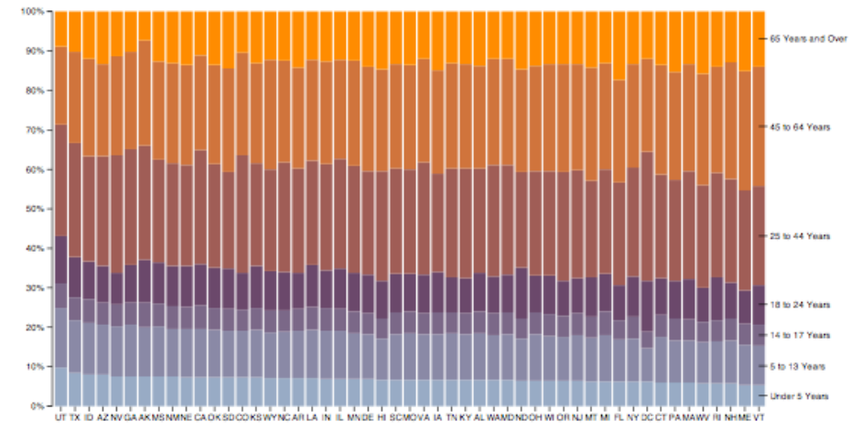
[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.]

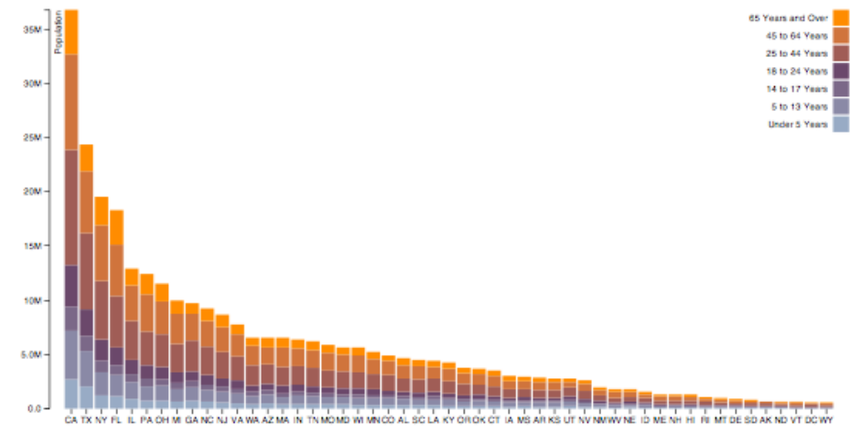
Idioms: **normalized stacked bar chart**

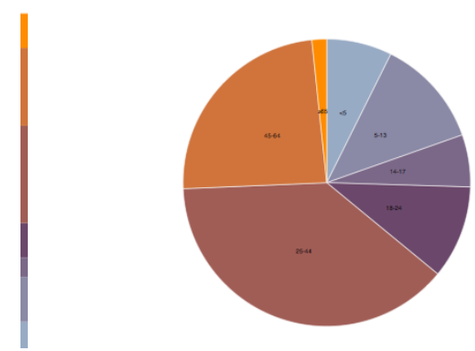
- task
  - part-to-whole judgements

- normalized stacked bar chart
  - stacked bar chart, normalized to full vert height
  - single stacked bar equivalent to full pie
    - high information density: requires narrow rectangle

- pie chart
  - information density: requires large circle





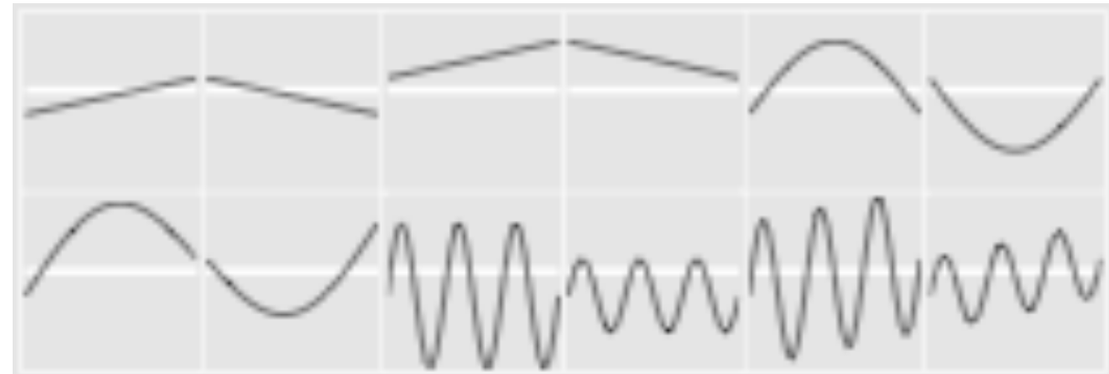


http://bl.ocks.org/mbostock/3887235, http://bl.ocks.org/mbostock/3886208,

http://bl.ocks.org/mbostock/3886394.

Idiom: **glyphmaps**

- rectilinear good for linear vs nonlinear trends



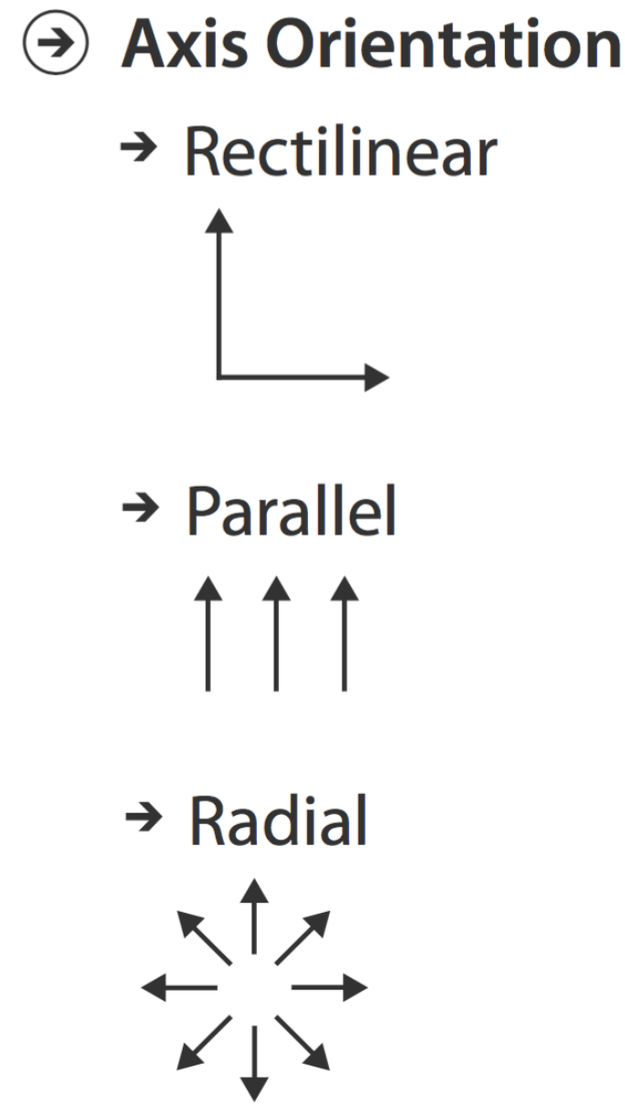
- radial good for cyclic patterns

[Glyph-maps for Visually Exploring Temporal Patterns in Climate Data and Models. Wickham, Hofmann, Wickham, and Cook. Environmetrics 23:5 (2012), 382–393.]

Orientation limitations

- rectilinear: scalability wrt #axes
  - 2 axes best
  - 3 problematic
    - more in afternoon
  - 4+ impossible
- parallel: unfamiliarity, training time
- radial: perceptual limits
  - angles lower precision than lengths
  - asymmetry between angle and length
    - can be exploited!

[Uncovering Strengths and Weaknesses of Radial Visualizations - an Empirical Approach. Diehl, Beck and Burch. IEEE TVCG (Proc. InfoVis) 16(6):935–942, 2010.]



Further reading

- Visualization Analysis and Design. Munzner. AK Peters / CRC Press, Oct 2014.
  - *Chap 7: Arrange Tables*
- Visualizing Data. Cleveland. Hobart Press, 1993.
- *A Brief History of Data Visualization*. Friendly. 2008. http://www.datavis.ca/milestones

Outline

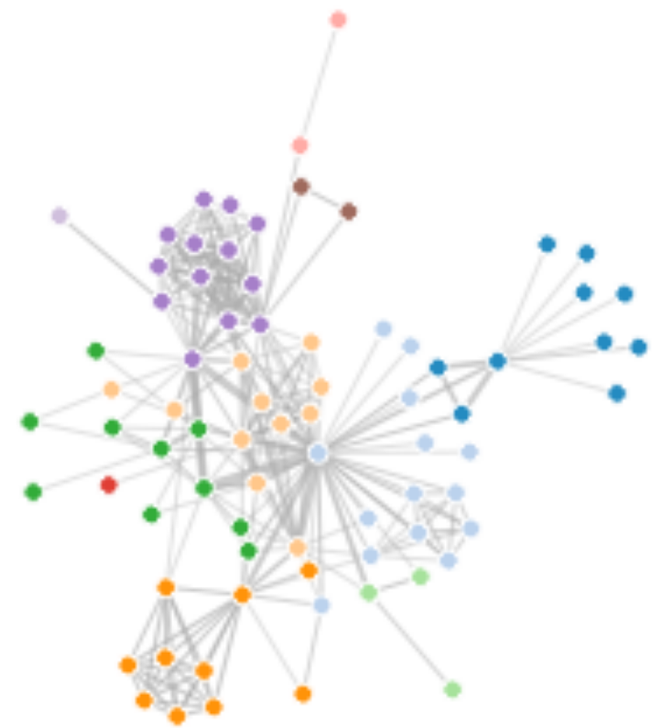
- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - **Arrange Networks & Trees (if time allows)**
  - Hands on: Altair
  - Rules of Thumb & Vis Research

Arrange networks and trees

Idiom: **force-directed placement**

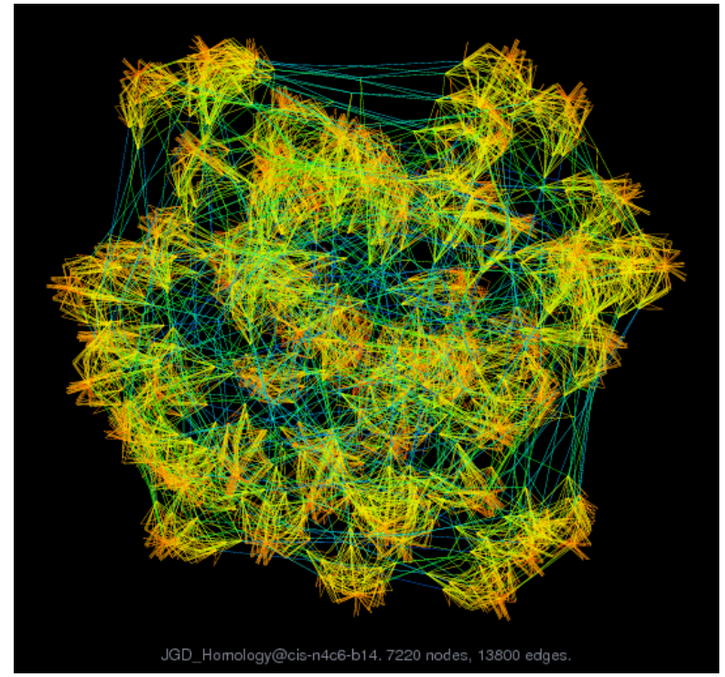
- visual encoding
  - link connection marks, node point marks
- considerations
  - spatial position: no meaning directly encoded
    - left free to minimize crossings
  - proximity semantics?
    - sometimes meaningful
    - sometimes arbitrary, artifact of layout algorithm
    - tension with length
      - long edges more visually salient than short
- tasks
  - explore topology; locate paths, clusters
- scalability
  - node/edge density E < 4N



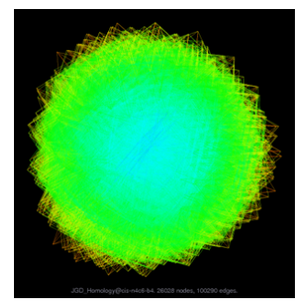
http://mbostock.github.com/d3/ex/force.html

Idiom: **sfdp** (multi-level force-directed placement)

- data
  - original: network
  - derived: cluster hierarchy atop it
- considerations
  - better algorithm for same encoding technique
    - same: fundamental use of space
    - hierarchy used for algorithm speed/quality but not shown explicitly
    - (more on algorithm vs encoding in afternoon)
- scalability
  - nodes, edges: 1K-10K
  - hairball problem eventually hits



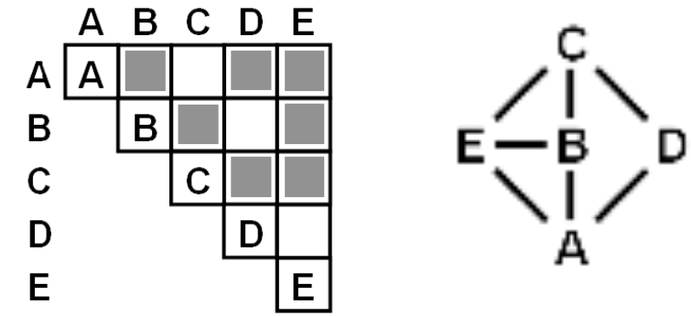
[Efficient and high quality force-directed graph drawing. Hu. The Mathematica Journal 10:37–71, 2005.]



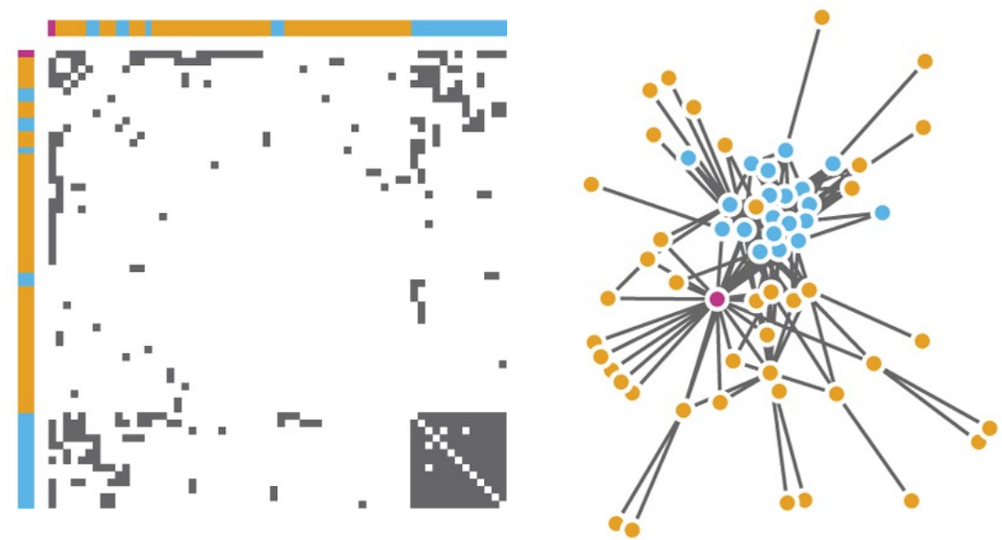
http://www.research.att.com/yifanhu/GALLERY/GRAPHS/index1.html

Idiom: **adjacency matrix view**

- data: network
  - transform into same data/encoding as heatmap
- derived data: table from network
  - 1 quant attrib
    - weighted edge between nodes
  - 2 categ attribs: node list x 2
- visual encoding
  - cell shows presence/absence of edge
- scalability
  - 1K nodes, 1M edges



[NodeTrix: a Hybrid Visualization of Social Networks. Henry, Fekete, and McGuffin. IEEE TVCG (Proc. InfoVis) 13(6):1302-1309, 2007.]

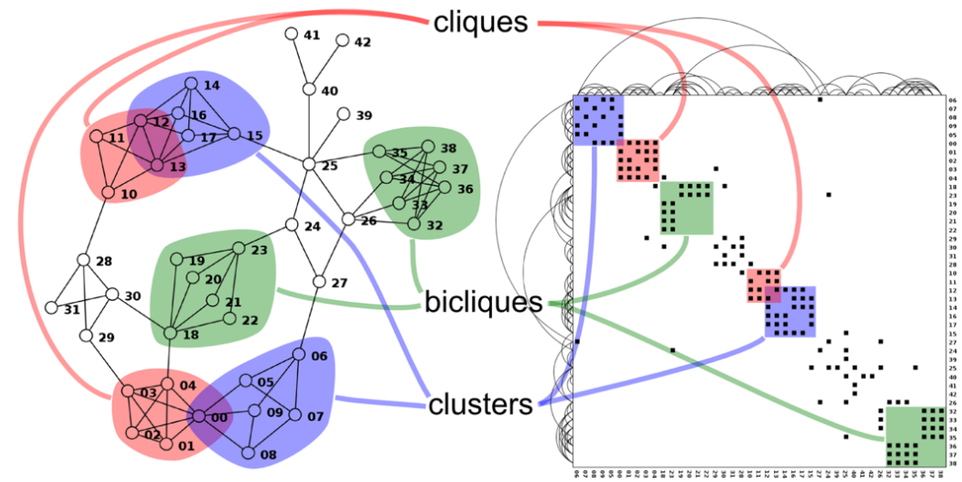


[Points of view: Networks. Gehlenborg and Wong. Nature Methods 9:115.]

Connection vs. adjacency comparison

- adjacency matrix strengths
  - predictability, scalability, supports reordering
  - some topology tasks trainable
- node-link diagram strengths
  - topology understanding, path tracing
  - intuitive, no training needed
- empirical study
  - node-link best for small networks
  - matrix best for large networks
    - if tasks don't involve topological structure!

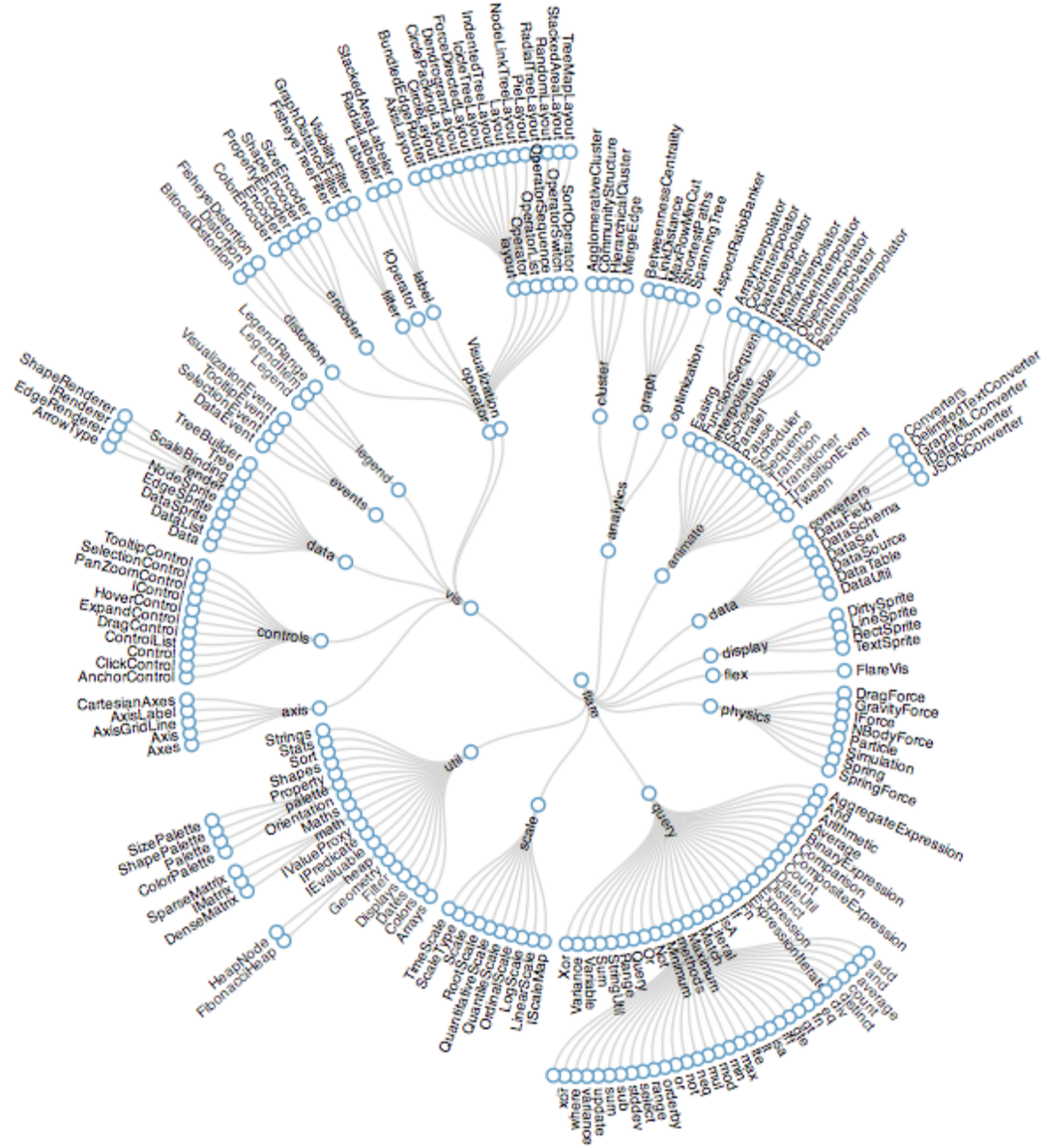
[On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Ghoniem, Fekete, and Castagliola. Information Visualization 4:2 (2005), 114–135.]



http://www.michaelmcguffin.com/courses/vis/patternsInAdjacencyMatrix.png

Idiom: **radial node-link tree**

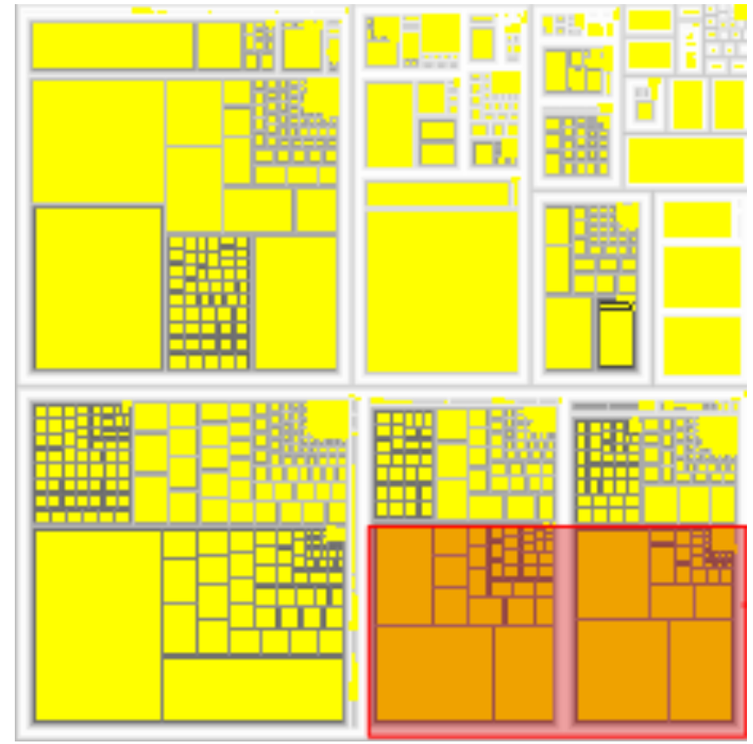
- data
  - tree
- encoding
  - link connection marks
  - point node marks
  - radial axis orientation
    - angular proximity: siblings
    - distance from center: depth in tree
- tasks
  - understanding topology, following paths
- scalability
  - 1K - 10K nodes



http://mbostock.github.com/d3/ex/tree.html

Idiom: **treemap**

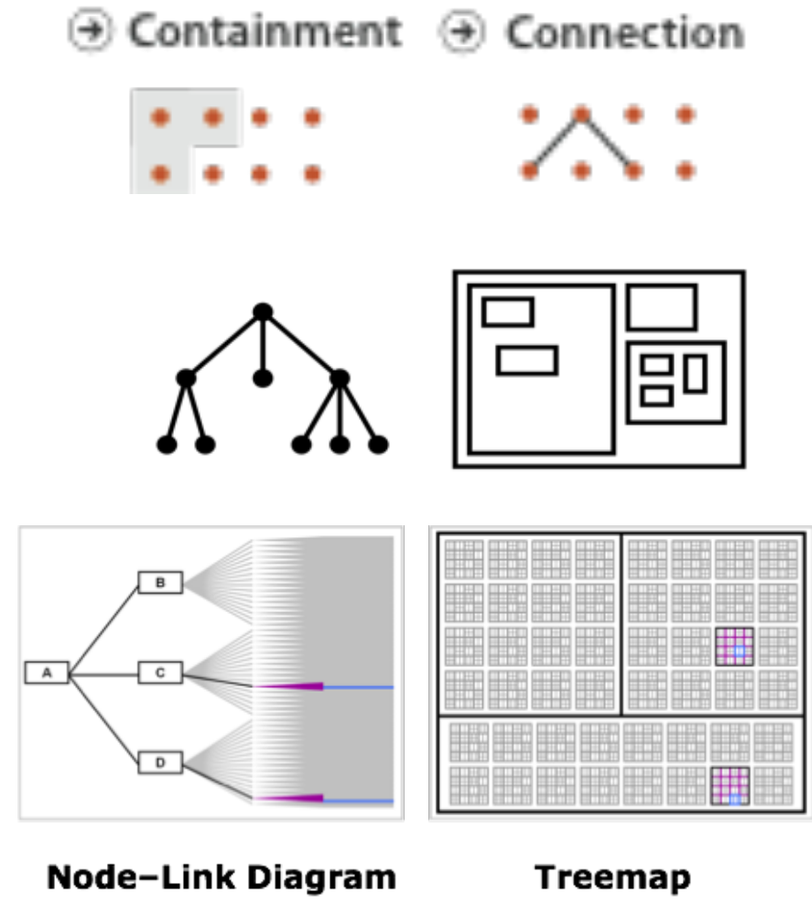
- data
  - tree
  - 1 quant attrib at leaf nodes
- encoding
  - area containment marks for hierarchical structure
  - rectilinear orientation
  - size encodes quant attrib
- tasks
  - query attribute at leaf nodes
- scalability
  - 1M leaf nodes



http://tulip.labri.fr/Documentation/3_7/userHandbook/html/ch06.html

Link marks: Connection and Containment

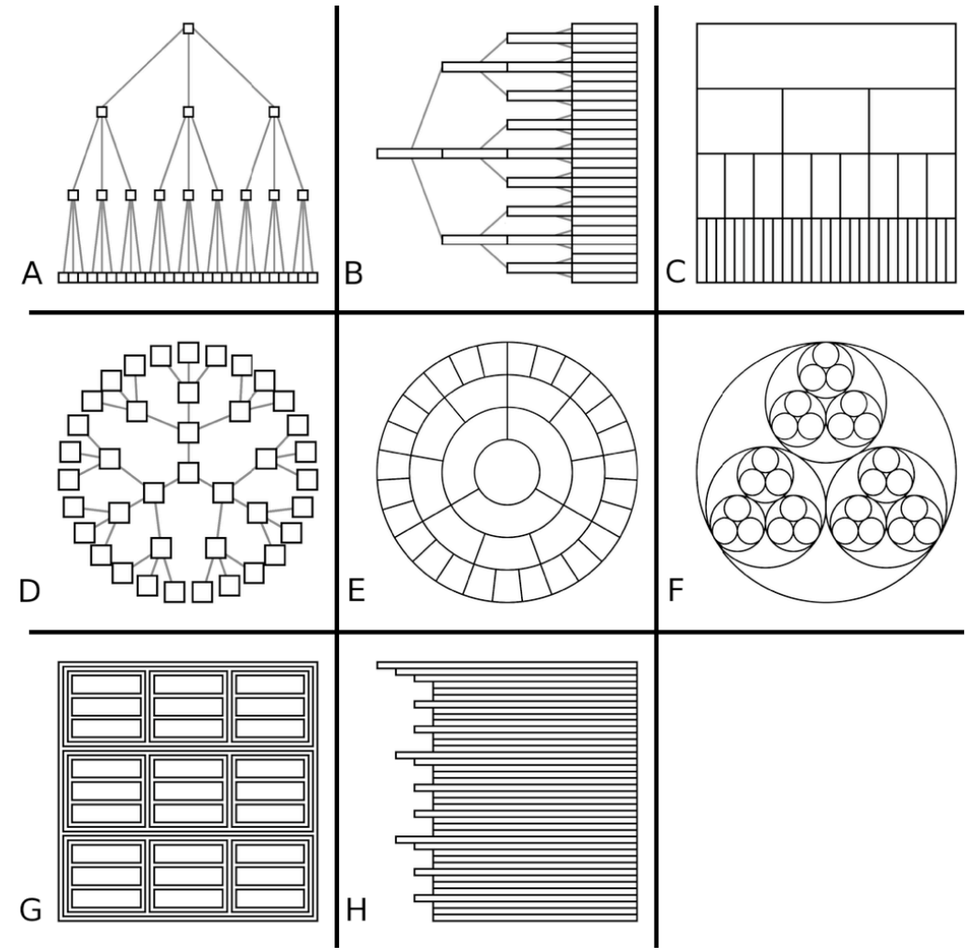
- marks as links (vs. nodes)
  - common case in network drawing
  - 1D case: connection
    - ex: all node-link diagrams
    - emphasizes topology, path tracing
    - networks and trees
  - 2D case: containment
    - ex: all treemap variants
    - emphasizes attribute values at leaves (size coding)
    - only trees



[Elastic Hierarchies: Combining Treemaps and Node-Link Diagrams. Dong, McGuffin, and Chignell. Proc. InfoVis 2005, p. 57-64.]

Tree drawing idioms comparison

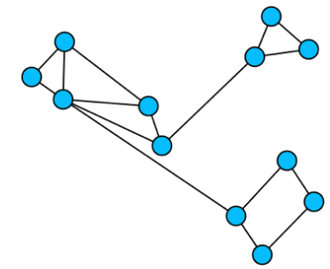
- data shown
  - link relationships
  - tree depth
  - sibling order
- design choices
  - connection vs containment link marks
  - rectilinear vs radial layout
  - spatial position channels
- considerations
  - redundant? arbitrary?
  - information density?
    - avoid wasting space

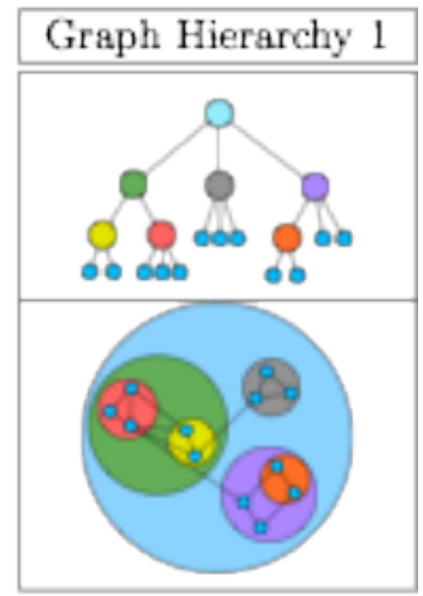


[Quantifying the Space-Efficiency of 2D Graphical Representations of Trees. McGuffin and Robert. Information Visualization 9:2 (2010), 115–140.]

Idiom: **GrouseFlocks**

- data: compound graphs
  - network
  - cluster hierarchy atop it
    - derived or interactively chosen
- visual encoding
  - connection marks for network links
  - containment marks for hierarchy
  - point marks for nodes
- dynamic interaction
  - select individual metanodes in hierarchy to expand/contract





[GrouseFlocks: Steerable Exploration of Graph Hierarchy Space. *Archambault, Munzner, and Auber. IEEE TVCG 14(4):900 913, 2008.*]

Further reading

- Visualization Analysis and Design. Munzner. AK Peters / CRC Press, Oct 2014.
  - *Chap 9: Arrange Networks and Trees*
- Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. von Landesberger et al. Computer Graphics Forum 30:6 (2011), 1719–1749.
- Simple Algorithms for Network Visualization: A Tutorial. McGuffin. Tsinghua Science and Technology (Special Issue on Visualization and Computer Graphics) 17:4 (2012), 383–398.
- Drawing on Physical Analogies. Brandes. In Drawing Graphs: Methods and Models, LNCS Tutorial, 2025, edited by M. Kaufmann and D. Wagner, LNCS Tutorial, 2025, pp. 71–86. Springer-Verlag, 2001.
- Treevis.net: A Tree Visualization Reference. Schulz. IEEE Computer Graphics and Applications 31:6 (2011), 11–15. http://www.treevis.net
- Perceptual Guidelines for Creating Rectangular Treemaps. Kong, Heer, and Agrawala. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 16:6 (2010), 990–998.

Outline

- Session 1 *10:0-11:15am*
  Visualization Analysis Framework
  - Introduction: Definitions
  - Analysis: What, Why, How
  - **Marks and Channels**

- Session 2 *11:45am-13:00pm*
  Spatial Layout
  - Arrange Tables
  - **Hands on: Altair**
  - Rules of Thumb & Vis Research

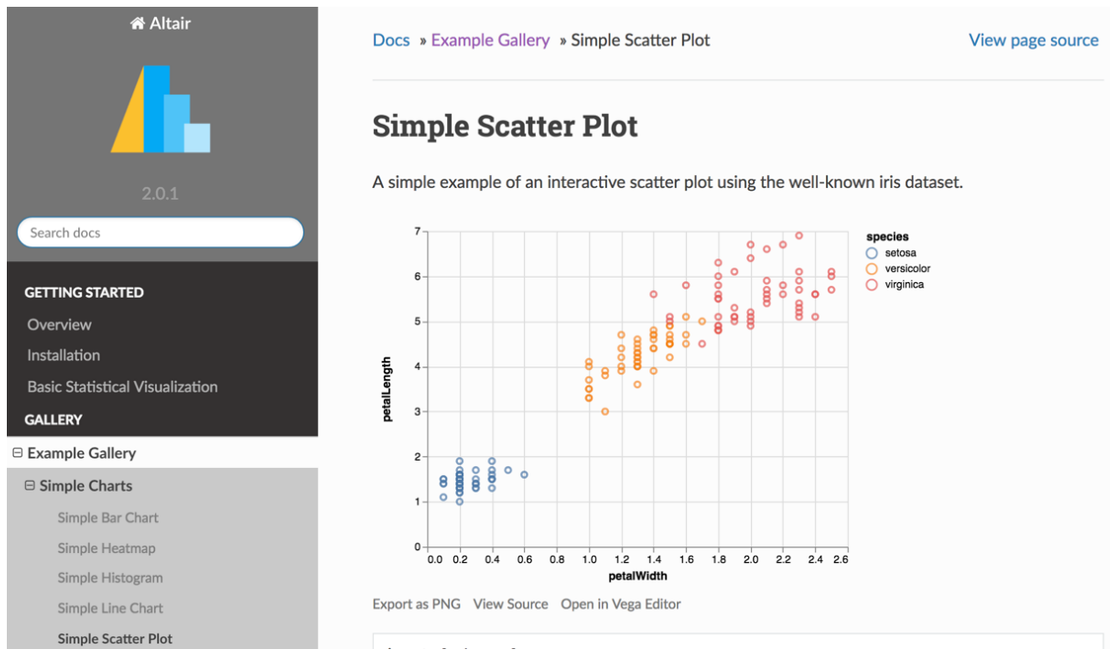
How to use these concepts in practice ?

- Some libraries like Vega and Vega-lite allow a simple transition

https://vega.github.io/

- If you are used to python, one alternative is altair:

https://altair-viz.github.io/

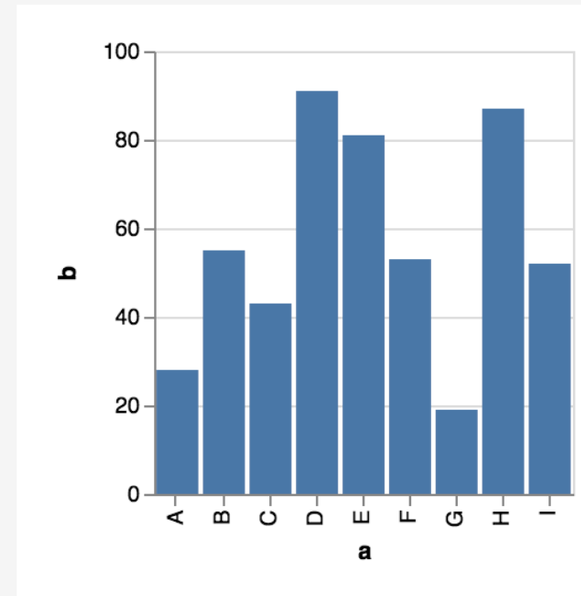


Example: Vega-Lite

Examples Gist New

Format

```
{
  "$schema": "https://vega.github.io/schema/vega-lite/v2.json",
  "description": "A simple bar chart with embedded data.",
  "data": {
    "values": [
      {"a": "A","b": 28}, {"a": "B","b": 55}, {"a": "C","b": 43},
      {"a": "D","b": 91}, {"a": "E","b": 81}, {"a": "F","b": 53},
      {"a": "G","b": 19}, {"a": "H","b": 87}, {"a": "I","b": 52}
    ]
  },
  "mark": "bar",
  "encoding": {
    "x": {"field": "a", "type": "ordinal"},
    "y": {"field": "b", "type": "quantitative"}
  }
}
```

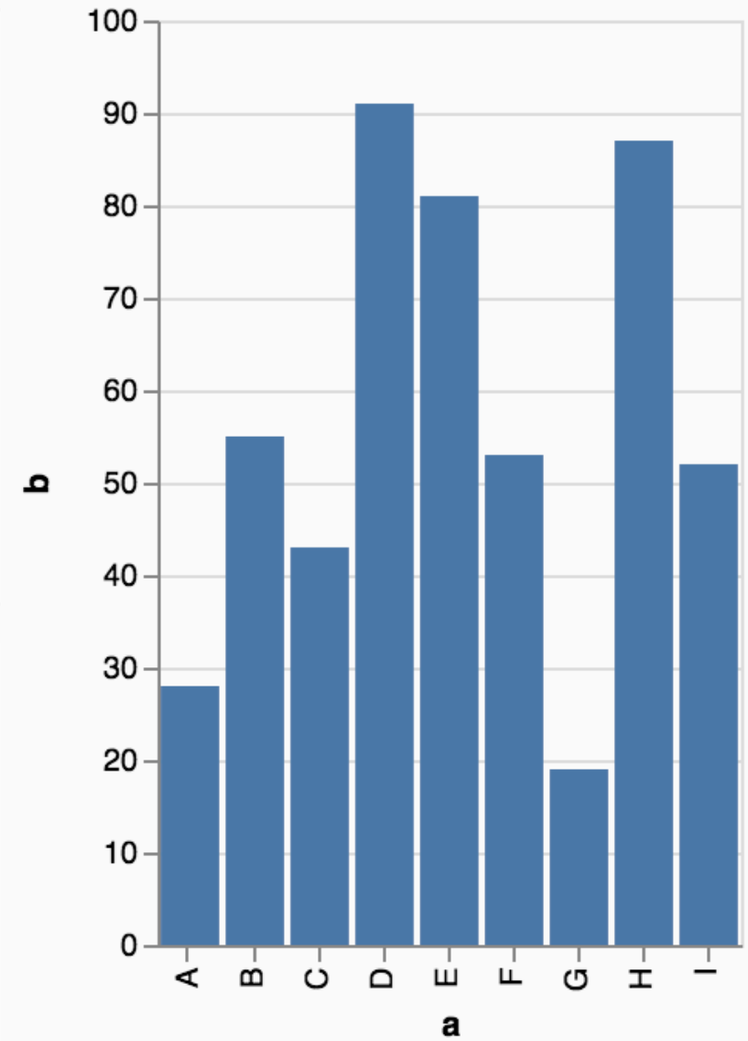


Example: Altair

```
import altair as alt
import pandas as pd

data = pd.DataFrame({
    'a': ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I'],
    'b': [28, 55, 43, 91, 81, 53, 19, 87, 52]
})

alt.Chart(data).mark_bar().encode(
    x='a',
    y='b'
)
```

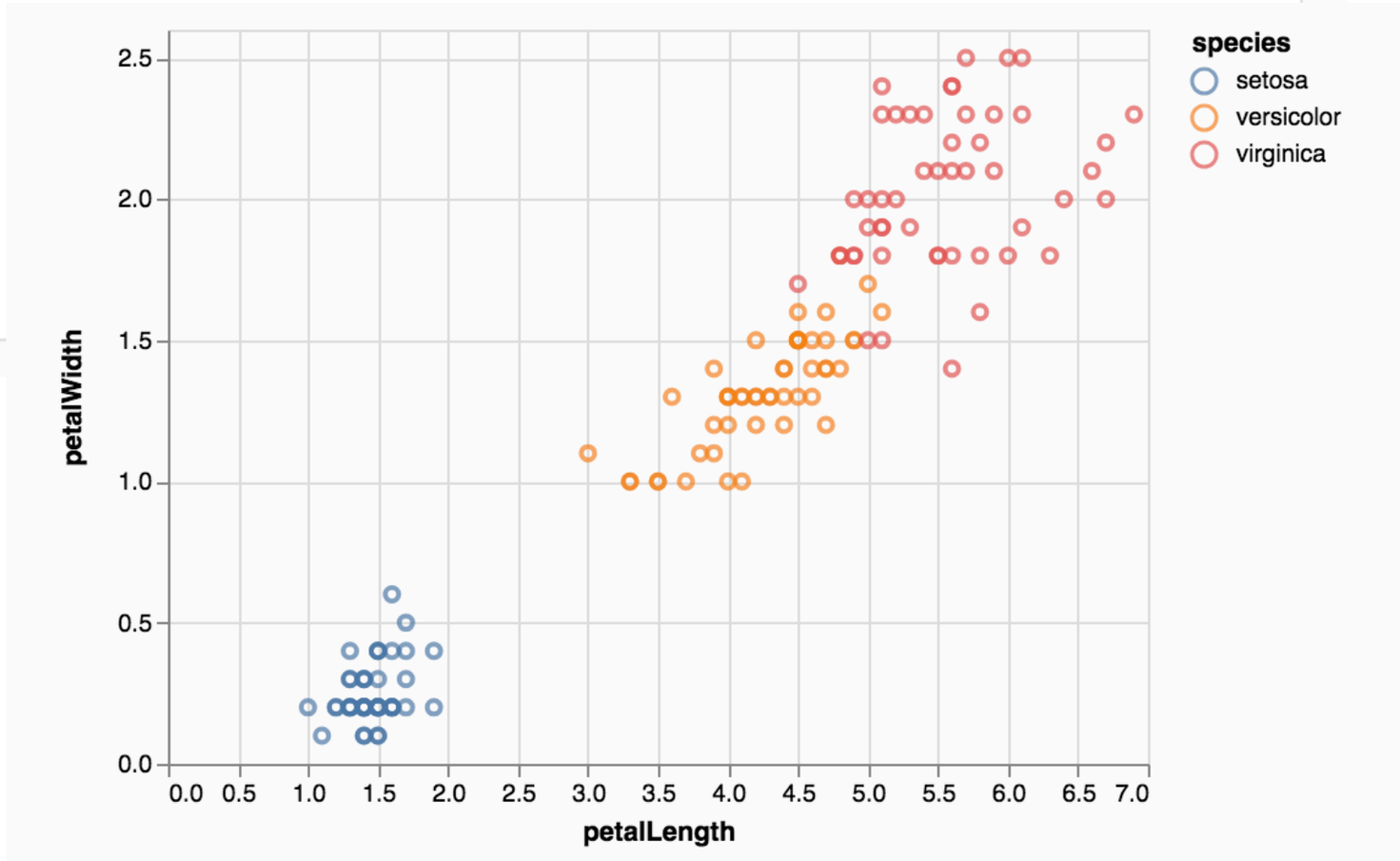


Example 2 : Altair

```
import altair as alt
from vega_datasets import data

iris = data.iris()

alt.Chart(iris).mark_point().encode(
    x='petalLength',
    y='petalWidth',
    color='species'
)
```



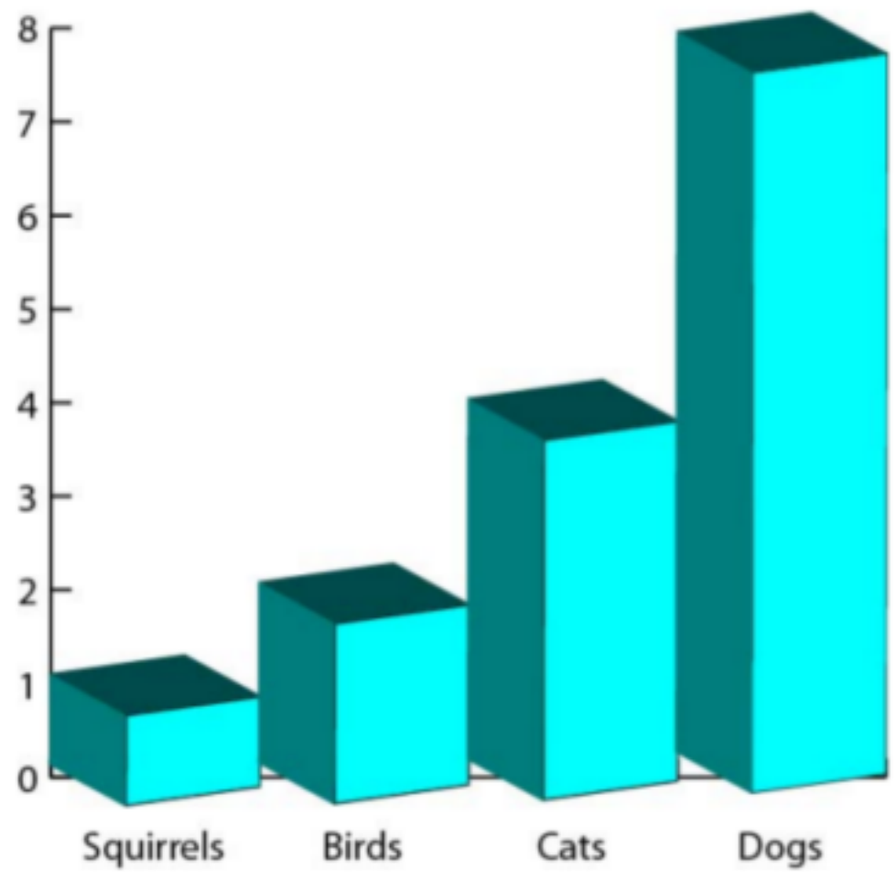
Some Rules of Thumb

Data ink ratio (Tufte)

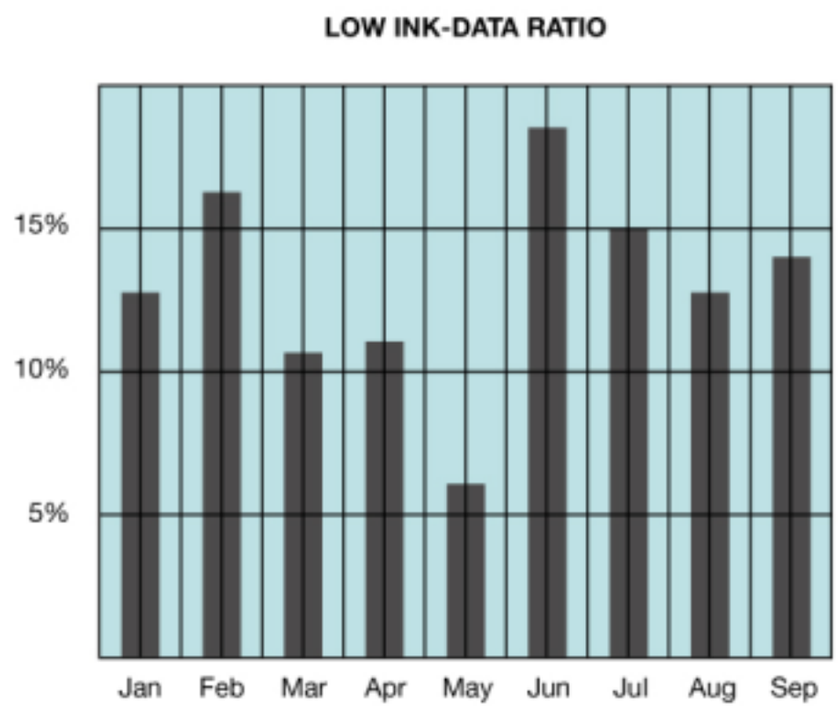
$$\text{data ink ratio} = \frac{\text{data ink}}{\text{total ink used}}$$

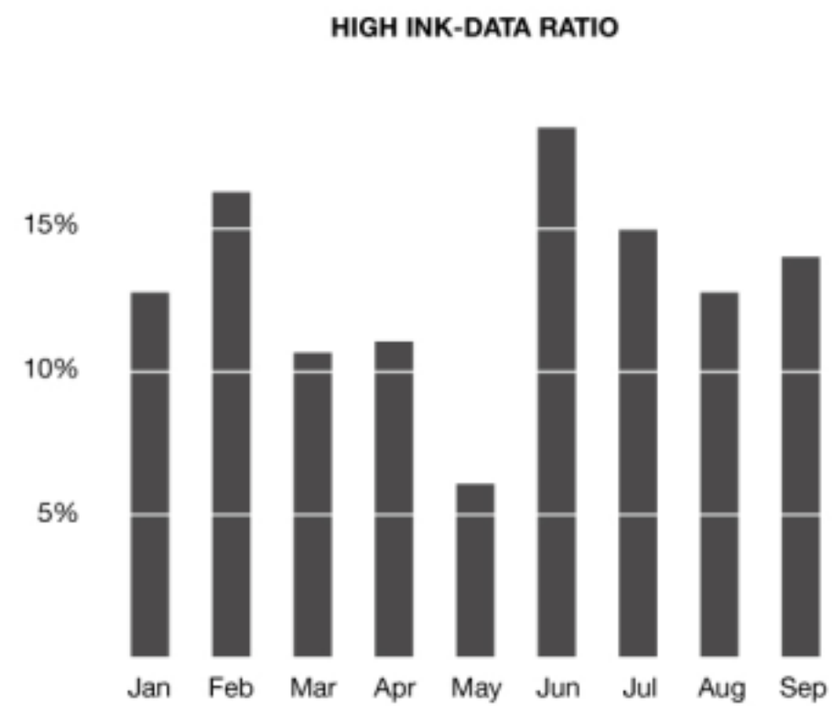
We try to maximize this ratio, in order to justify the use of any mark/channel. In an extreme case (i.e. Ratio = 1), each pixel must be justified.

Data ink ratio (Tufte)



Data ink ratio (Tufte)



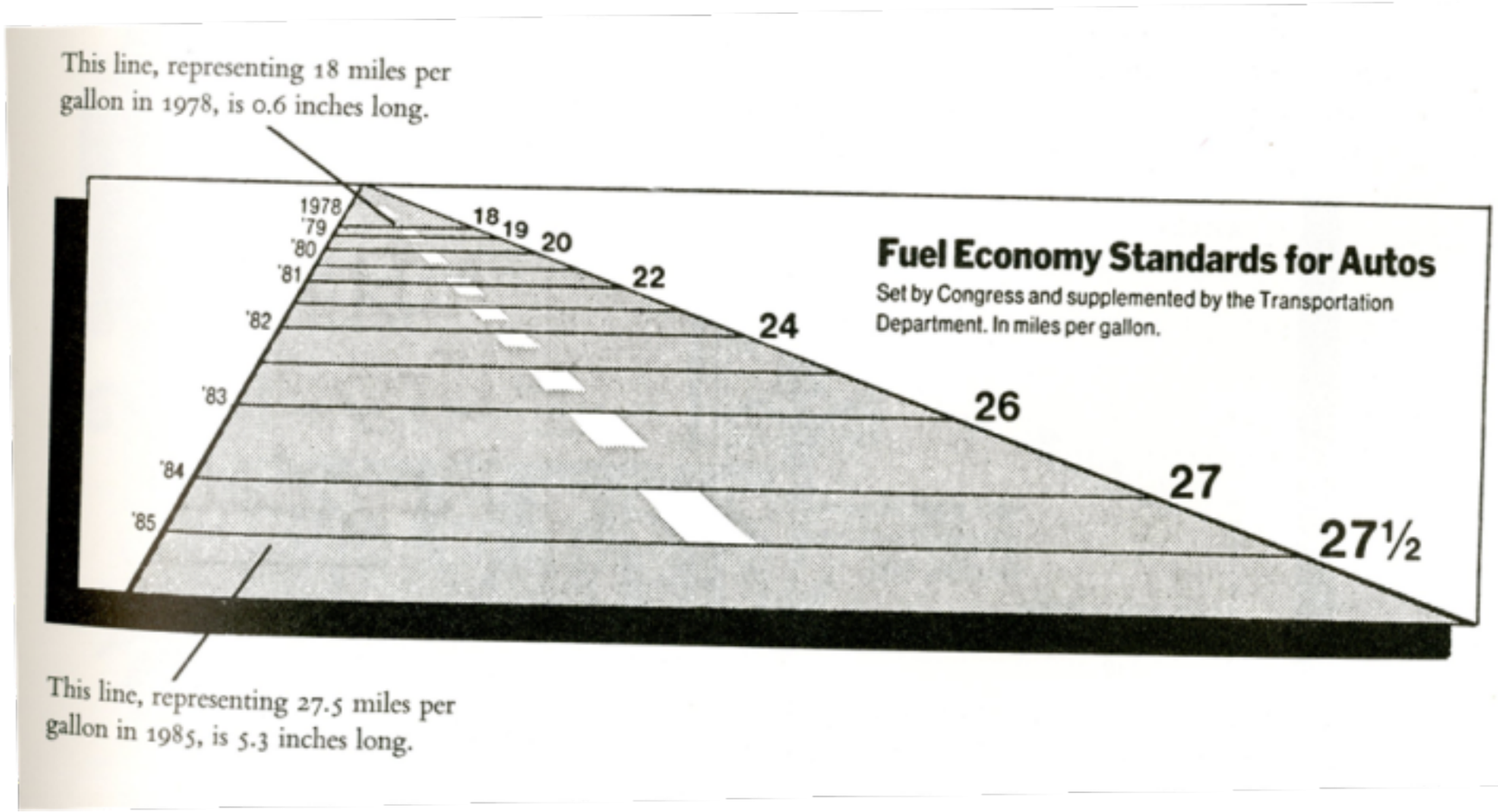


Lie factor (Tufte)

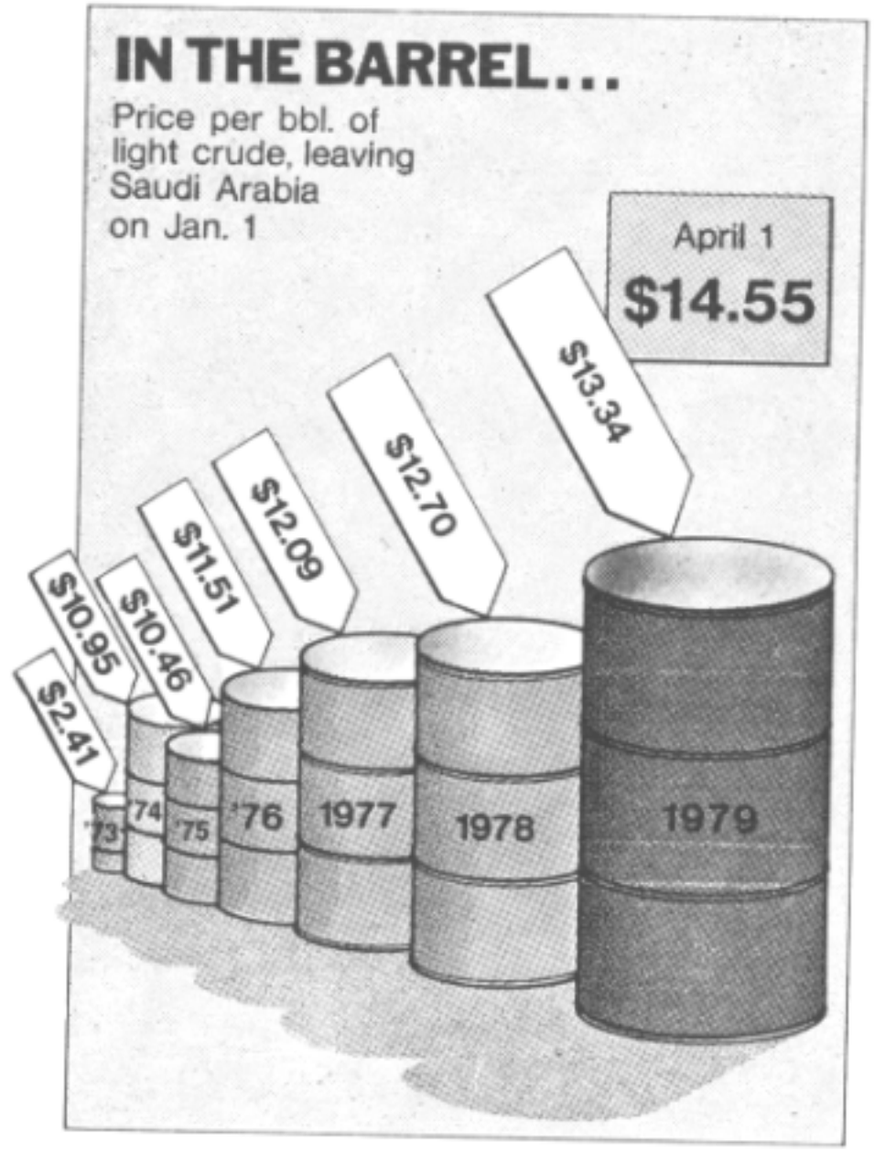
$$\text{lie factor} = \frac{\text{size effect in graphic}}{\text{size effect in data}}$$

- Data must be fully reflected by the effect in the graph.
- We should target to a lie factor = 1

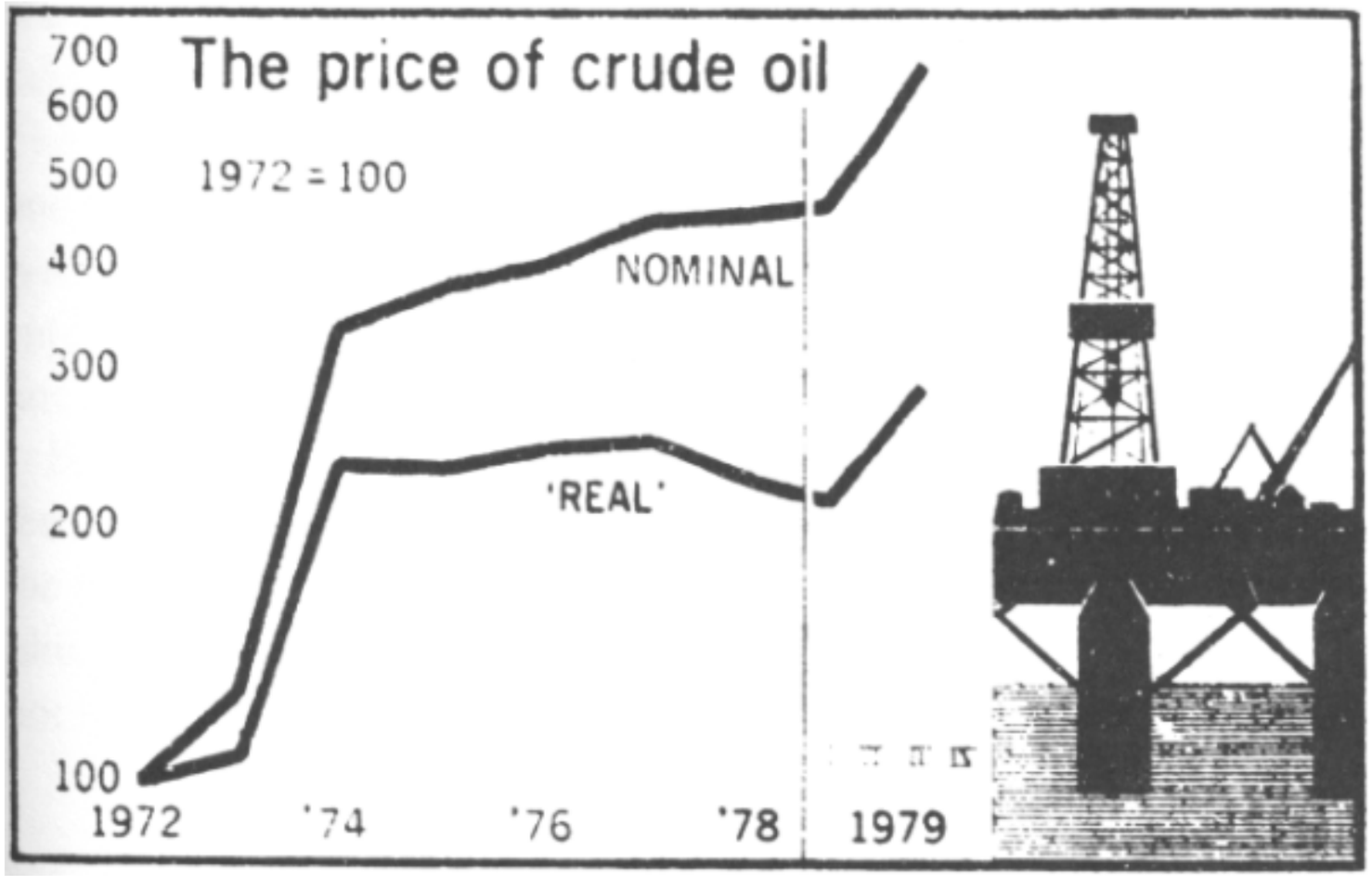
Lie factor (Tufte)



Lie factor (Tufte)

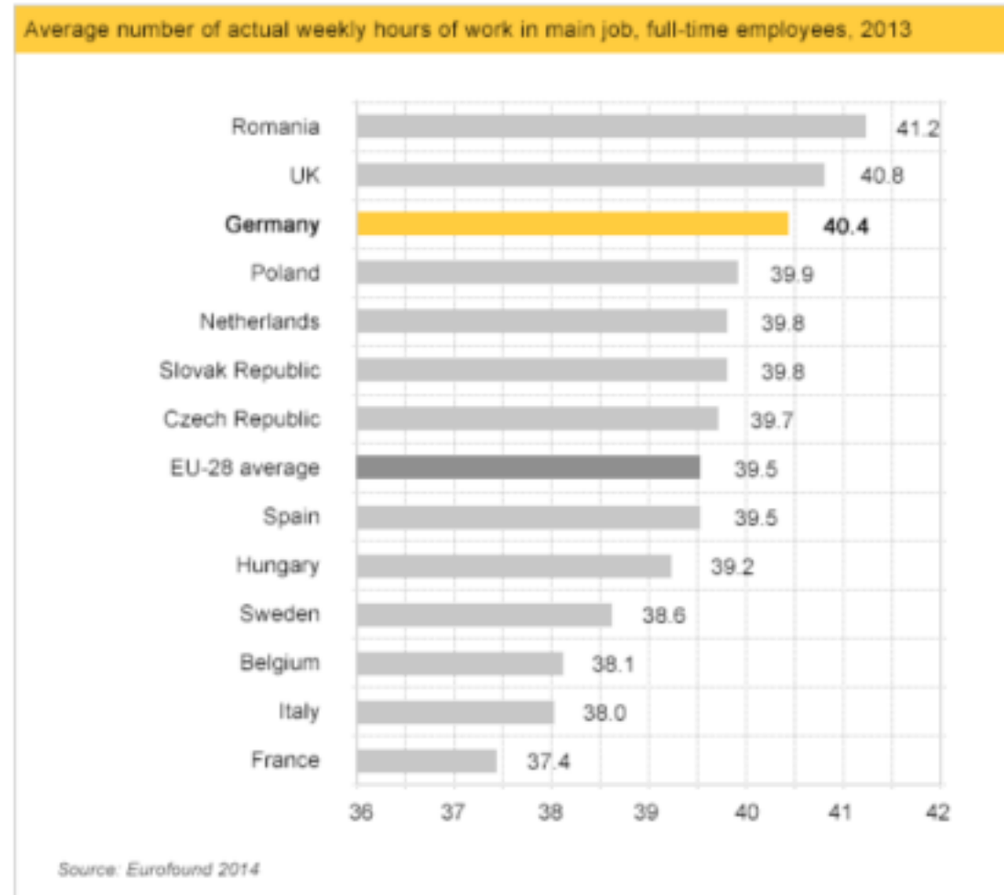


Lie factor (Tufte)



Deceiving axes I

- In this plot, axes do not start in 0.
- It seems that Germans works 3 times more than French, but the actual ratio is 1.08.

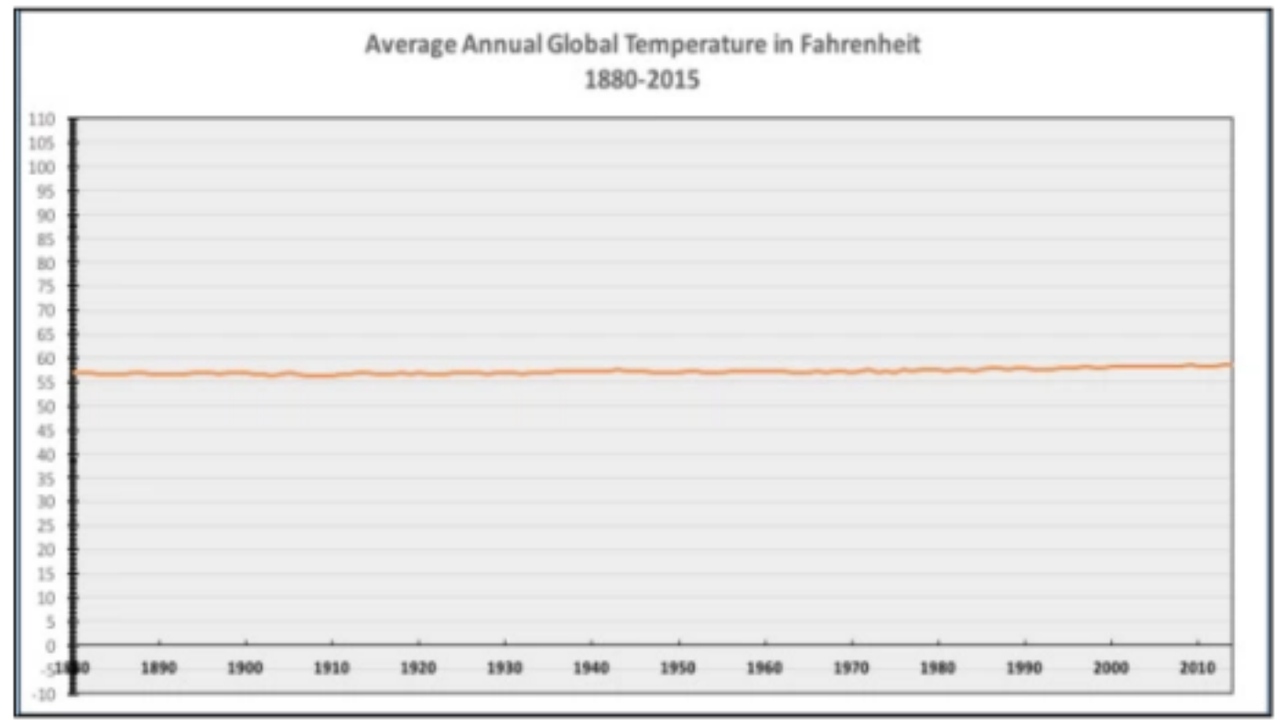


Deceiving axes I

- Same plot with axes starting in zero.

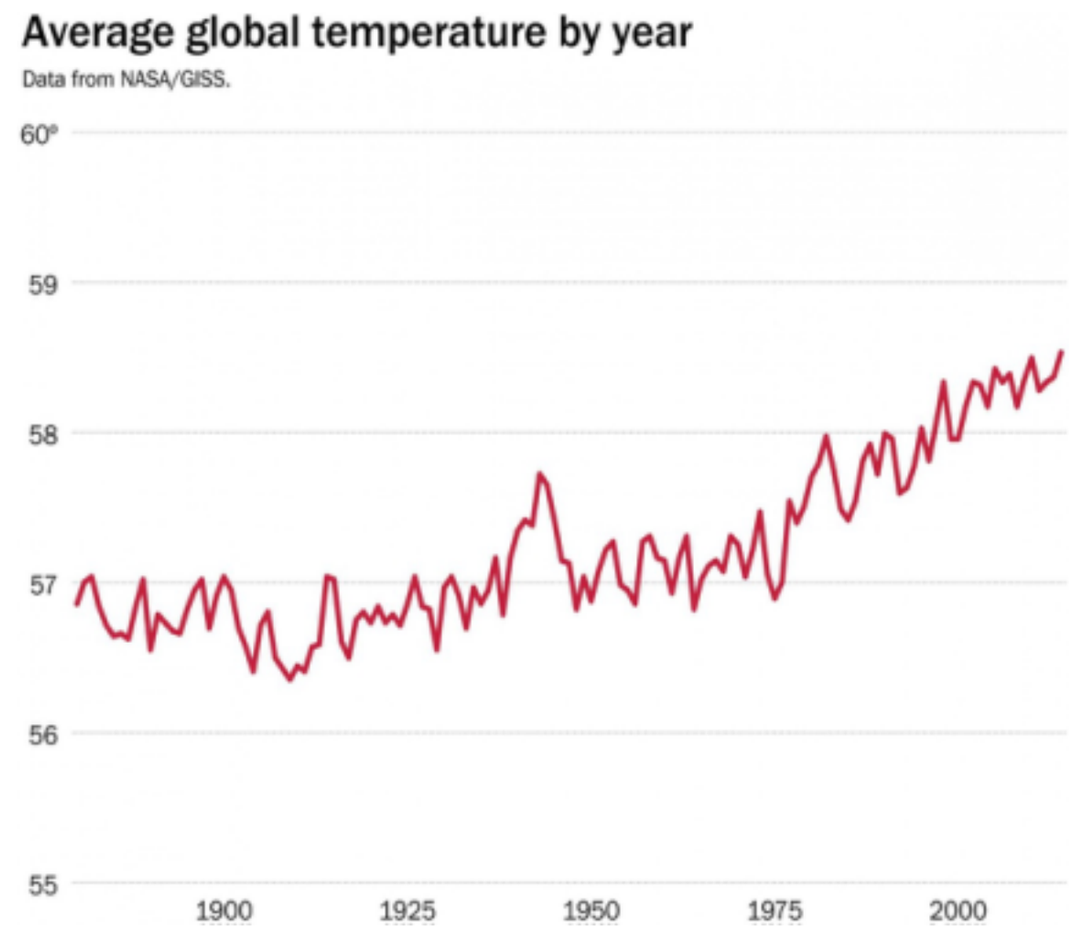
Deceiving axes II

- Sometimes adding the zero could be deceiving.
- Here the plot shows absolute magnitude for changes in temperatue in 1880-2015 (F)



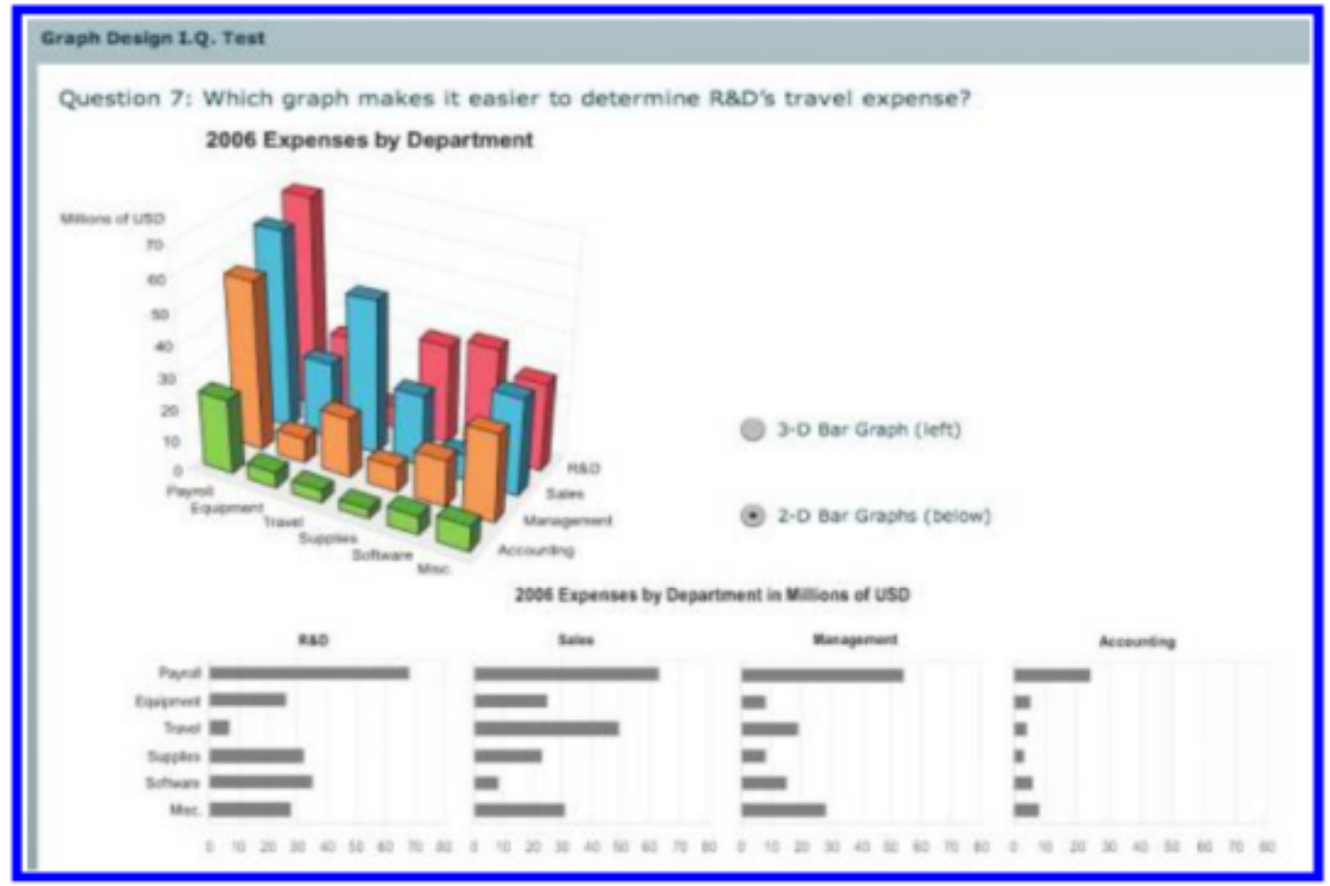
Deceiving axes II

- This is a more appropriate representation, considering that is the change what really matters and the fact that small changes in temperature can have a large effect.



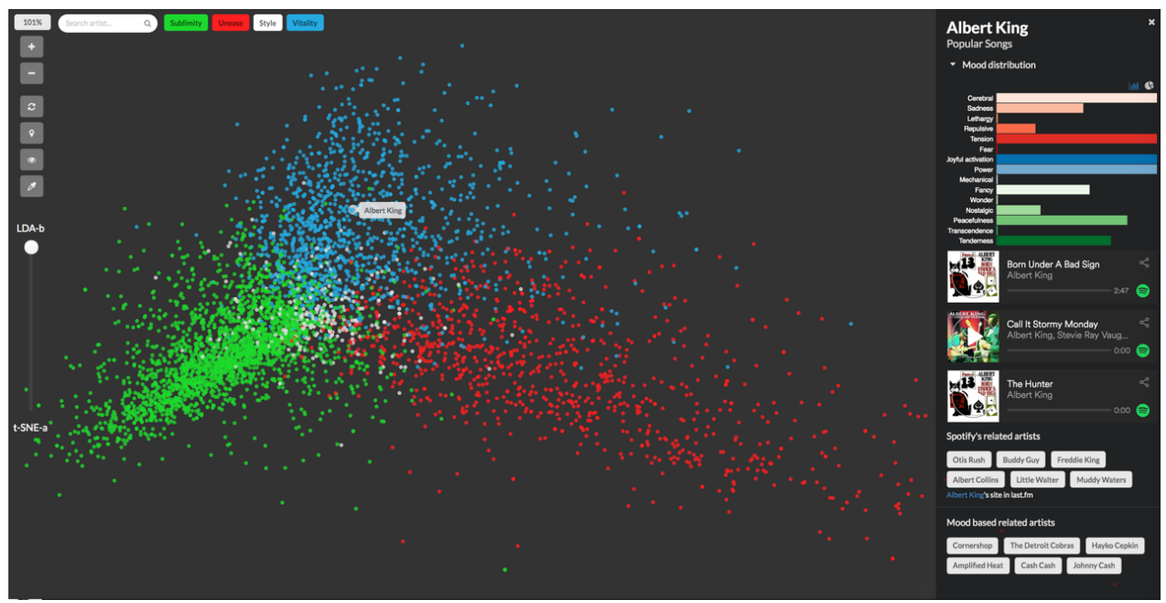
NO to unjustified 3D

- Unnecesary oclusion
- In this example, there are 2 categorical variables and 1 quantitative.
- Faceting could be a better option in this case.

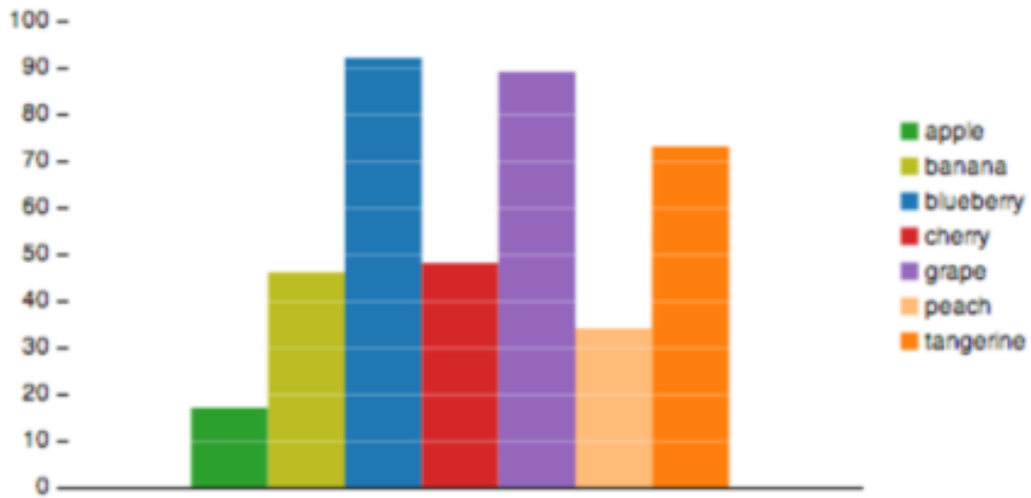


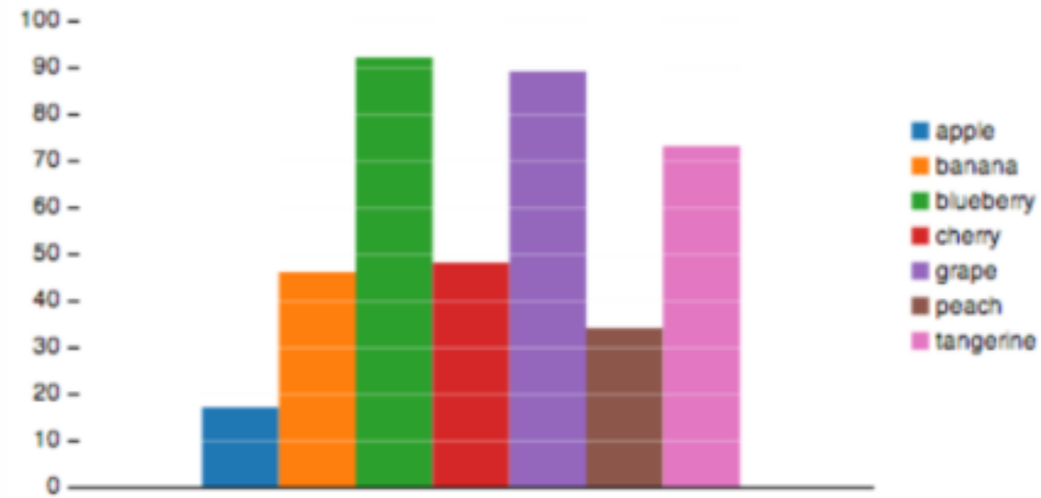
“Overview first, details on demand”

- Ben Shneiderman mantra [1996]. An idiom following it can provide the user conciousness about the whole information space, helping to summarize. Example: http://moodplay.pythonanywhere.com



Try to maintain semantic consitency

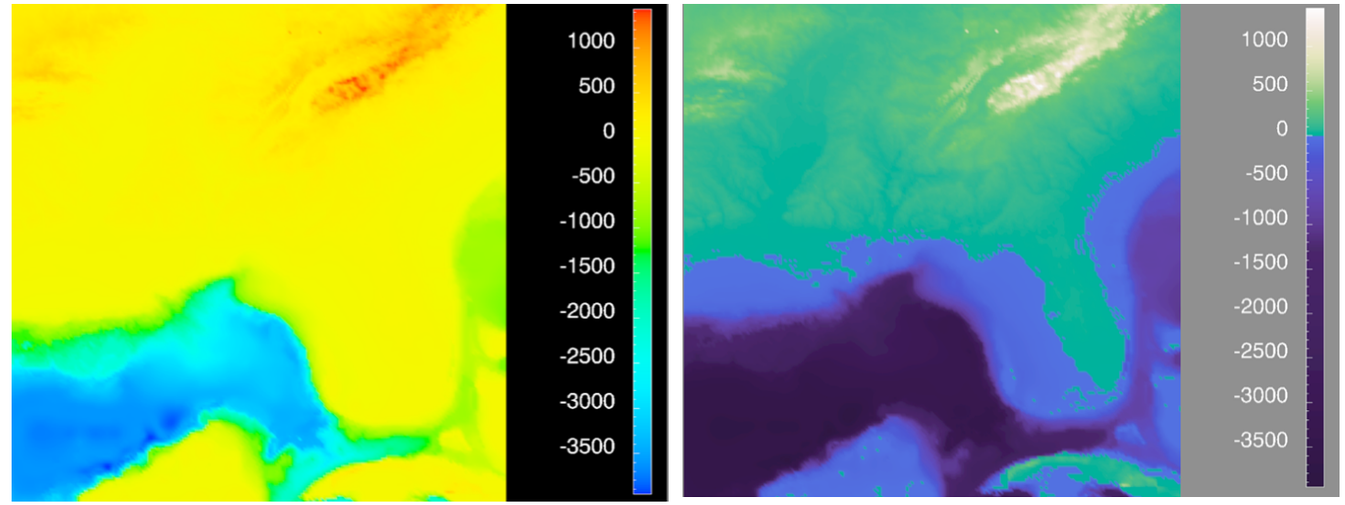




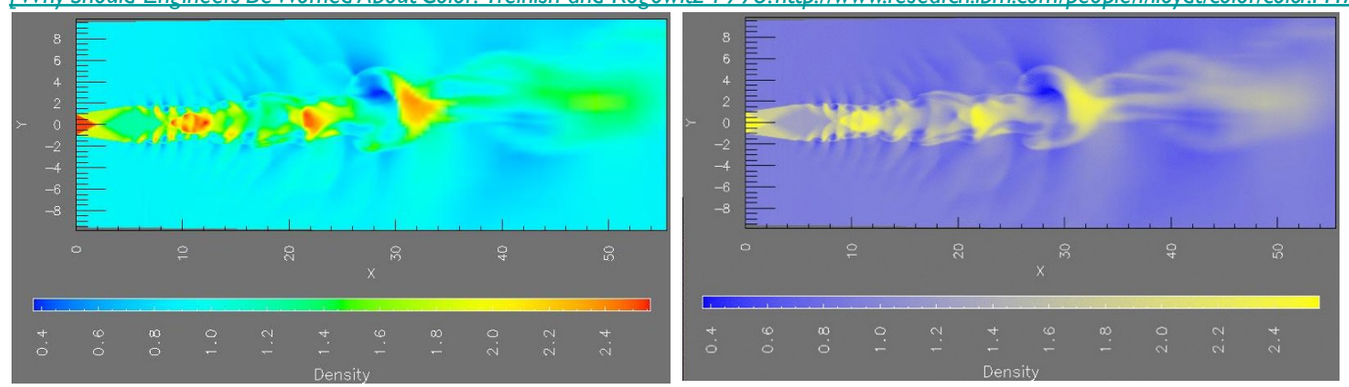
Selecting Semantically–Resonant Colors for Data Visualization — S. Lin et al.

Ordered color: Rainbow is poor default

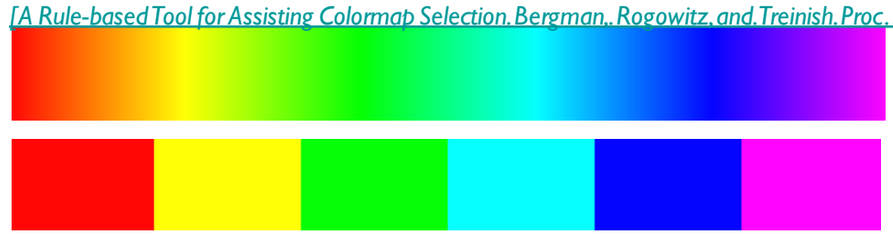
- problems
  - perceptually unordered
  - perceptually nonlinear
- benefits
  - fine-grained structure visible and nameable
- alternatives
  - fewer hues for large-scale structure
  - multiple hues with monotonically increasing luminance for fine-grained
  - segmented rainbows good for categorical, ok for binned



[Why Should Engineers Be Worried About Color? Treinish and Rogowitz 1998. http://www.research.ibm.com/people/l/lloydt/color/color.HTM]

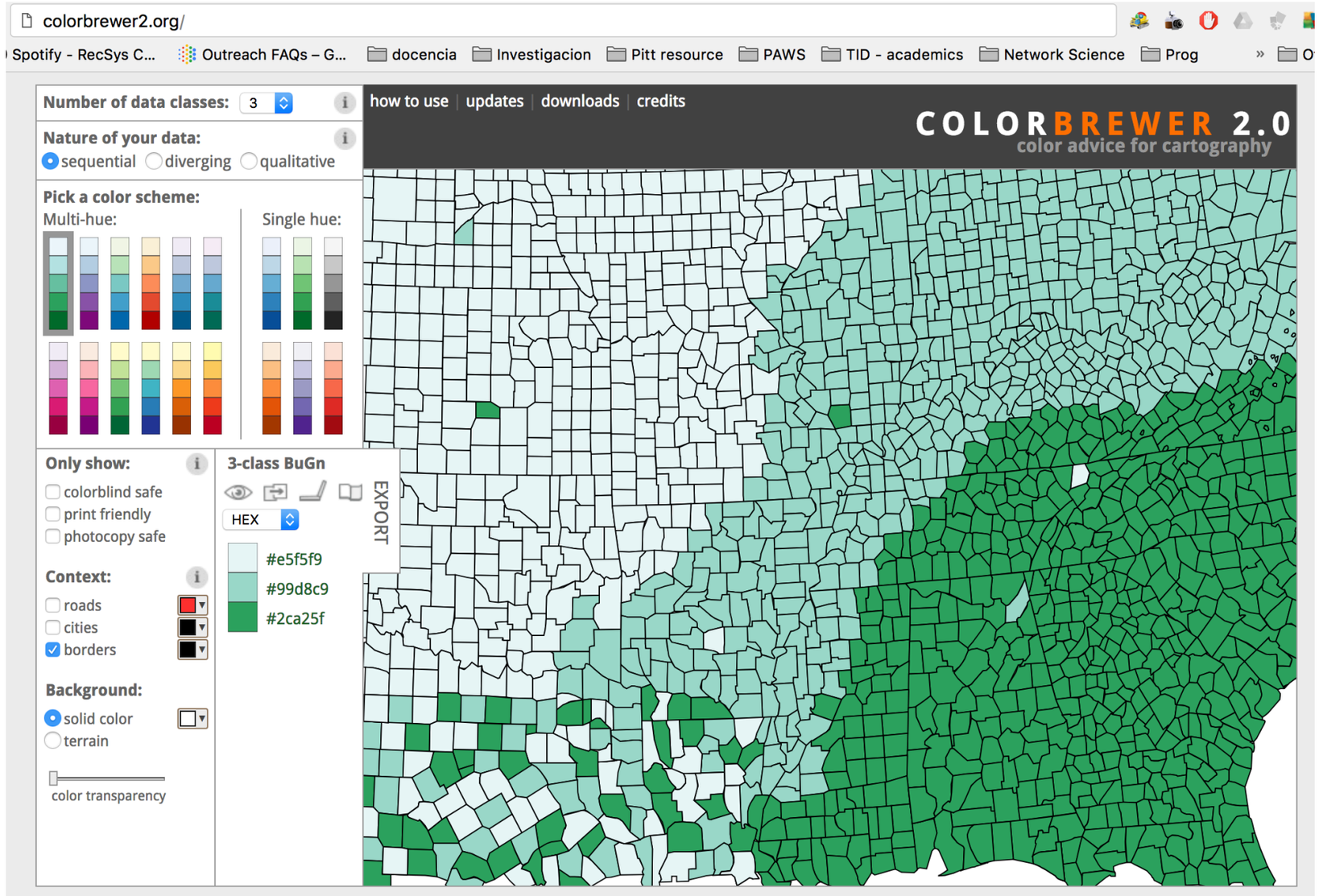


[A Rule-based Tool for Assisting Colormap Selection. Bergman,. Rogowitz, and Treinish. Proc. IEEE Visualization (Vis), pp. 118–125, 1995.]



[Transfer Functions in Direct Volume Rendering: Design, Interface, Interaction. Kindlmann. SIGGRAPH 2002 Course Notes]

Use colorbrewer2.org



More guidelines

- Units must be standardized for comparison (e.g. Currency)
- Plot dimensions should not exceed data ranges.
- Data should be presented in their context.
- Check always for color blindness.
- Check always for typography.
- Make plot labels and legens self-explaining.

Bibliography

- “The Visual Display of Quantitative Information” — Edward Tufte
- “VAD” — Tamara Munzner
- WTFViz
- Calling Bullshit: in the age of big data

Current Research on Information Visualization

What is being researched on Information Visualization

- Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J., & Möller, T. (2017). Visualization as seen through its research paper keywords. *IEEE transactions on visualization and computer graphics*, *23*(1), 771-780.

What is being researched on Information Visualization

- Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J., & Möller, T. (2017). Visualization as seen through its research paper keywords. *IEEE transactions on visualization and computer graphics*, *23*(1), 771-780.

  - Researchers analyzed co-occurrence of: a) Free keywords provided by authors b)Words from the taxonomy of the PCS systems.
  - Dataset: papers from IEEE Vis conference (VAST, InfoVis, Vis, SciVis)

Topic Clusters (author keywords)

Table 4. Resulting clusters of topic-coded keywords; represented by their two most frequent keywords. N = number of keywords per cluster.

ID	keywords (InfoVis, Vis/SciVis, VAST; 2004–2013)	N
T1	interaction techniques—general; graph/network data & techniques	15
T2	machine learning & statistics; uncertainty techn. & vis.	4
T3	timeseries, time-varying data & techn.; animation & motion	8
T4	evaluation general; cognition	7
T5	programming, algorithms, & data structures; geography, geospatial vis, cartography, terrain vis	6
T6	applications—general & other; text, document, topic analysis, data, & techn.	13
T7	abstraction, simplification, approximation; multiresolution techn.	4
T8	numerical methods/mathematics; meshes, grids, & lattices	10
T9	tensor data & techn.; tractography	2
T10	topology-based techn.; scalar field data & techn.	2
T11	flow vis., data, & techn.; vector fields, data, & techn.	4
T12	biology & bioinformatics; molecular science & chemistry	4
T13	large scale data & scalability; data acquisition & management	9
T14	displays—general; large & high-res displays	2
T15	volume rendering, modeling, & vis.; gpu-based techn.	5
T16	biomedical science & medicine; vis. techn. & tools—general	21

Topic Clusters (PCS taxonomy)

ID	keywords (InfoVis, Vis/SciVis, VAST; 2008–2015)	N
P1	volume rendering; biomedical and medical visualization	8
P2	time-varying data; vis. in physical sciences and engineering	8
P3	feature detection and tracking; vector field data	8
P4	visual knowledge discovery; graph/network data	13
P5	high-dimensional data; multidimensional data	5
P6	coordinated and multiple views; time series data	11
P7	collaborative and distributed vis.; large and high-res displays	4
P8	intelligence analysis; situational awareness	8
P9	scalability issues; streaming data	10
P10	multiresolution techniques; compression techniques	3
P11	scene perception; stereo displays	3
P12	data transformation and representation; data aggregation	9
P13	illustrative vis.; multimedia (image/video/music) vis.	10
P14	quantitative evaluation; laboratory studies	10
P15	user interfaces; interaction design	5
P16	visualization models; qualitative evaluation	12

Table 5. Resulting clusters of PCS keywords; represented by their two most frequent keywords. N = number of keywords per cluster.

Main trends

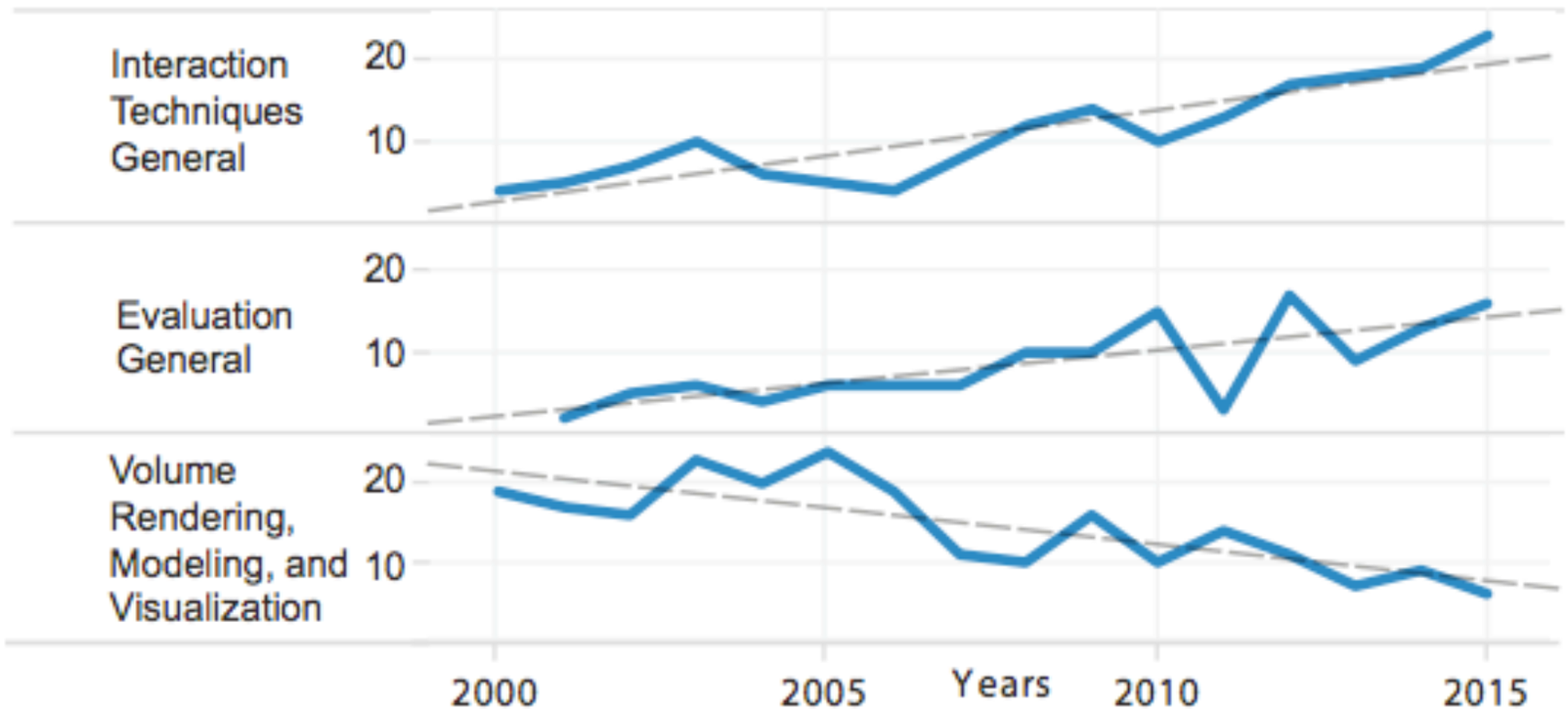


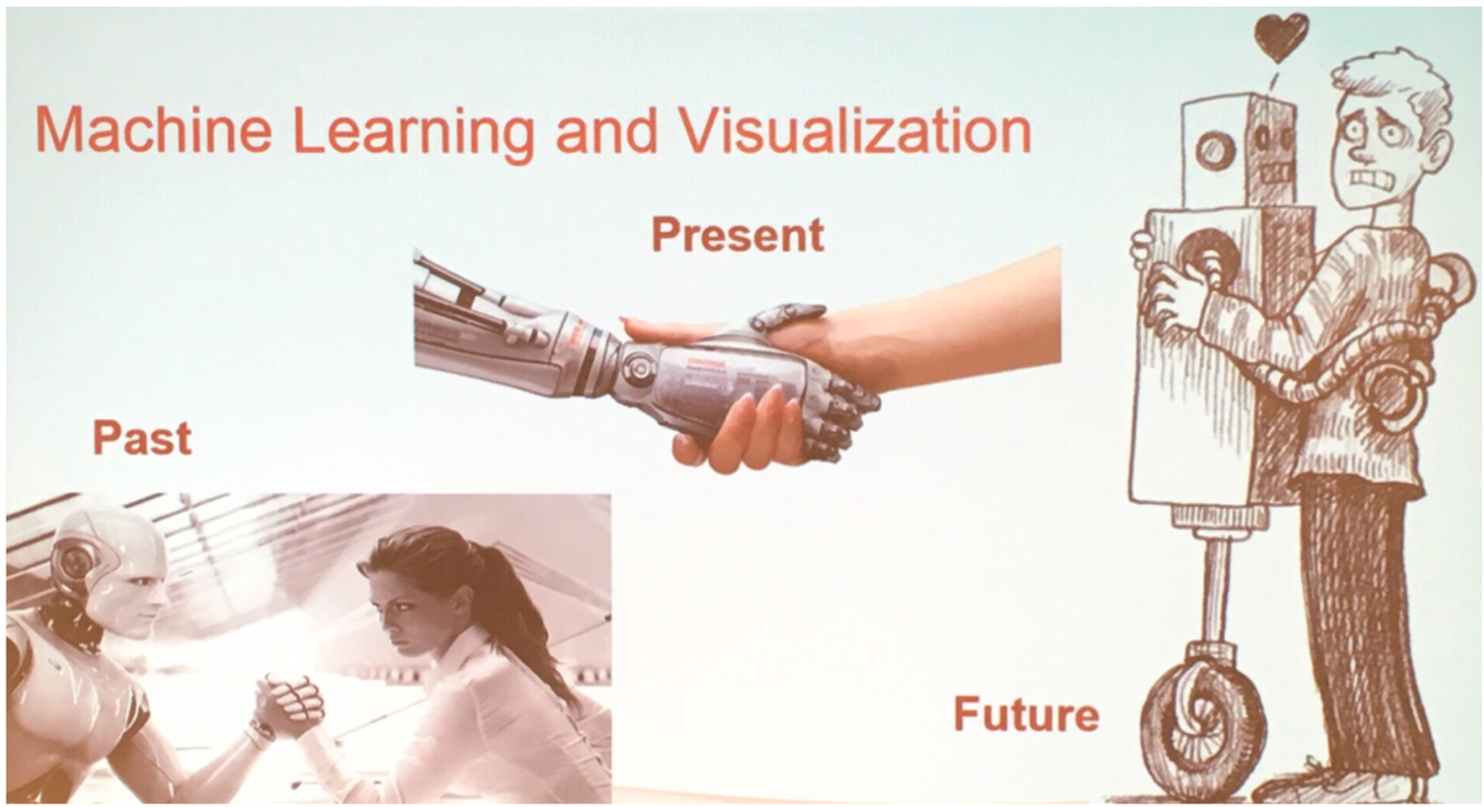
Fig. 7. The two strongest rising and the strongest falling topic keyword.

General Trends

topic (2000–2015)								PCS (2008–2015)							
keyword	chart	#	slope	SE	df	*p*-val.	*t*-val.	keyword	chart	#	slope	SE	df	*p*-val.	*t*-val
interaction techniques—general		175	1.12	0.15	14	**<.001**	7.503	**quantitative evaluation**		266	4.40	0.73	6	**.001**	7.457
evaluation general		128	0.81	0.20	13	**.001**	4.068	**visual knowledge discovery**		416	4.33	1.07	6	**.007**	6.296
machine learning and statistics		85	0.73	0.11	14	**<.001**	6.881	**time series data**		276	4.31	0.76	6	**.001**	6.190
timeseries, time-varying data and techniques		109	0.60	0.18	14	**.005**	3.294	**geographic/geospatial vis.**		269	4.20	1.05	6	**.007**	4.000
multidim./multivar./multifield data and techn.		109	0.57	0.12	14	**<.001**	4.555	**coordinated & multiple views**		334	4.17	1.47	6	**.030**	2.962
analysis process—general		110	0.47	0.18	12	**.024**	2.580	**data transf. and repres.**		267	3.75	1.20	6	**.020**	3.354
graph/network data and techniques		134	0.46	0.15	14	**.007**	3.169	interaction design		246	1.45	1.21	6	.275	1.951
visual encoding and layout—general		78	0.32	0.09	13	**.004**	3.455	graph/network data		373	0.89	1.03	6	.419	1.467
data clustering and aggregation		83	0.21	0.08	12	**.030**	2.465	multidimensional data		255	0.75	0.32	6	.056	2.633
visualization techniques and tools—general		82	0.18	0.10	13	.085	1.853	user interfaces		270	0.60	1.93	6	.768	1.983
biomedical science and medicine		122	-0.02	0.13	14	.855	-0.187	high-dimensional data		236	0.45	1.61	6	.788	1.572
flow visualization, data, and techniques		114	-0.11	0.16	14	.488	-0.713	biomedical and medical vis.		284	0.07	1.07	6	.949	0.204
numerical methods / mathematics		94	-0.42	0.15	14	**.016**	-2.752	vis. system and toolkit design		262	-0.31	1.78	6	.868	-0.174
meshes, grids, and lattices		86	-0.68	0.10	14	**<.001**	-6.869	time-varying data		263	-0.77	0.71	6	.315	0.486
volume rendering, modeling, and vis.		232	-0.92	0.19	14	**<.001**	-4.776	volume rendering		278	-2.52	1.33	6	.107	-1.870

Conclusions

- Information Visualization is a great tool to augment human capabilities by means of visual representations.
- Using a systematic procedure such as Munzner framework can help us deliver expressive and effective visualizations, making sure we decrease the risk of failing in different levels.
- Libraries like vega-lite or altair can help us to bridge the gap between this theoretical framework and the actual implementation of visualizations.

With respect to visualization and automatic decision making

IEEE VIS 2017 – Panel ML & Vis ?