University of Konstanz
Data Analysis and Visualization Group

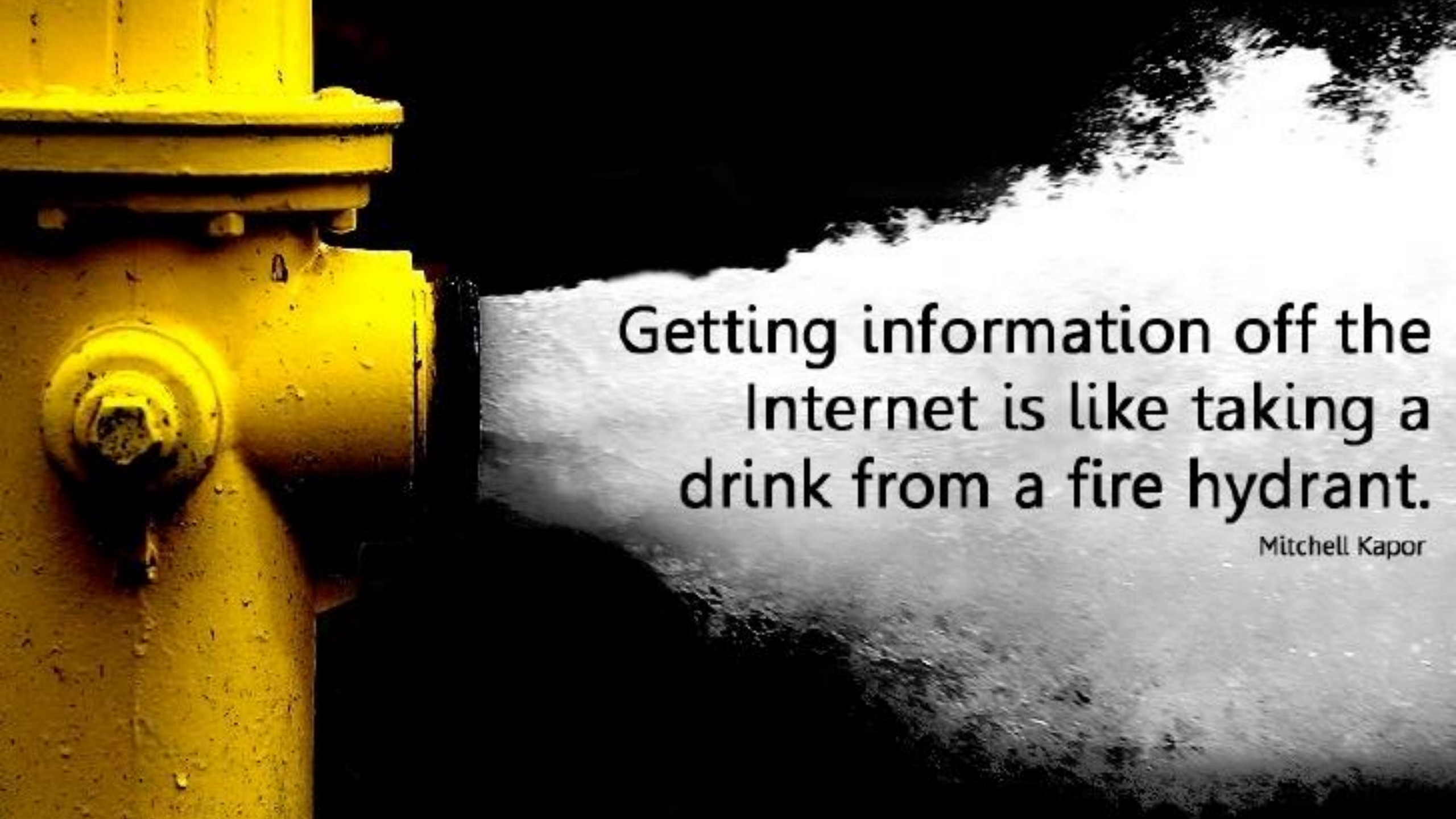
Visual Text Analytics
An Introduction

@manunna_91

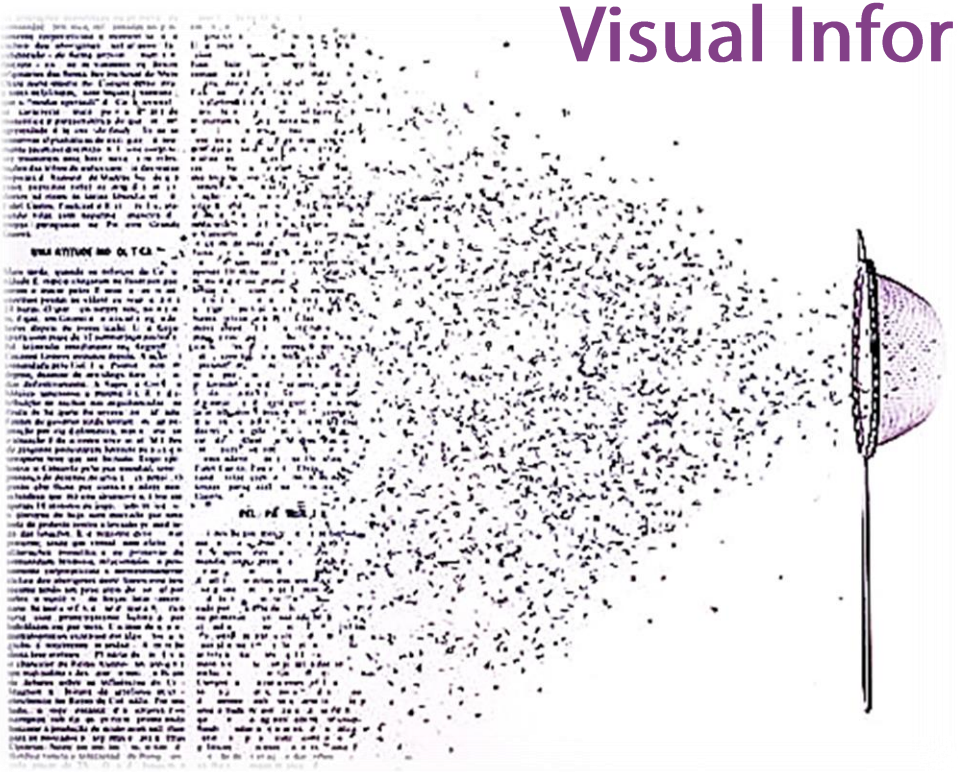
Mennatallah El-Assady

menna.el-assady@uni.kn

Today a person is subject to more new information **every day** than a person in the middle ages in their **entire lifetime**!


Getting information off the Internet is like taking a drink from a fire hydrant.
Mitchell Kapor

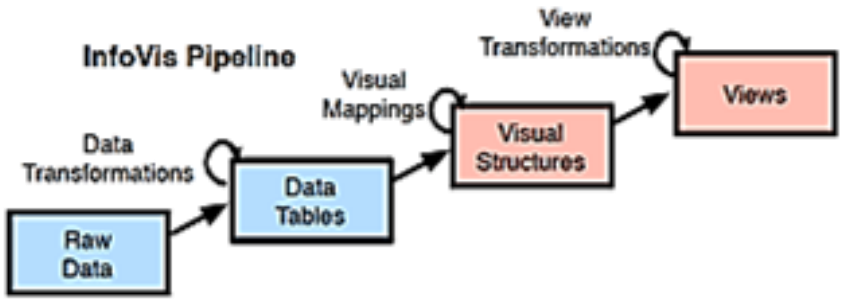
Visual Information Seeking Mantra



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

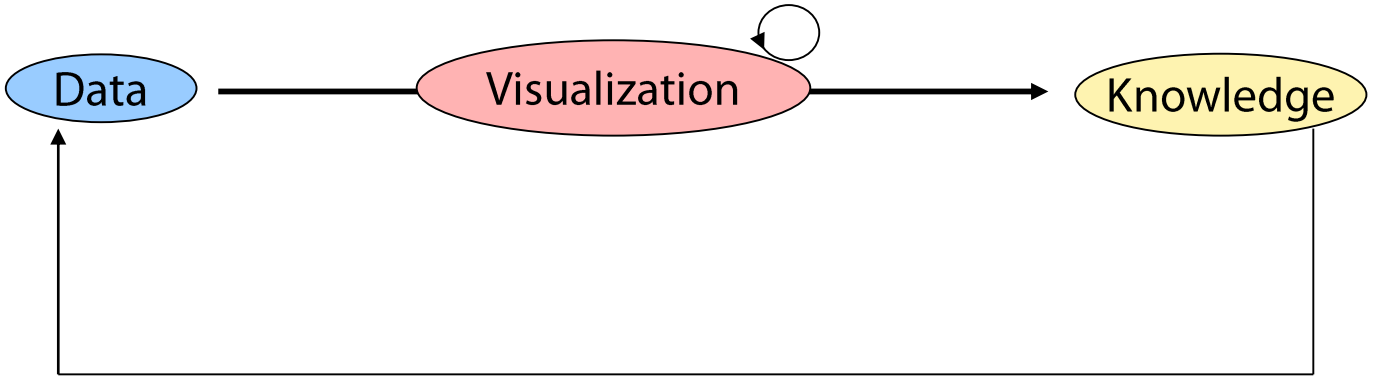
– Ben Schneiderman

What is Information Visualization?



Human Interaction is crucial for…

→ Exploration of Data
→ Generation of Hypotheses
→ Interpretation of Results
→ Steering of the Analysis
→ <u>Hypothesis Evaluation</u>



Feedback loop

What about a Fully Automated Analysis?

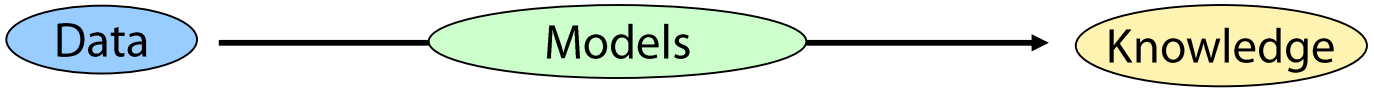
Fully Automated Methods only work under certain preconditions!

Preconditions:

→ Data is clearly structured
→ Data semantics is well-defined
→ Data is complete, correct, and not changing over time

AND

→ Problem is well-defined



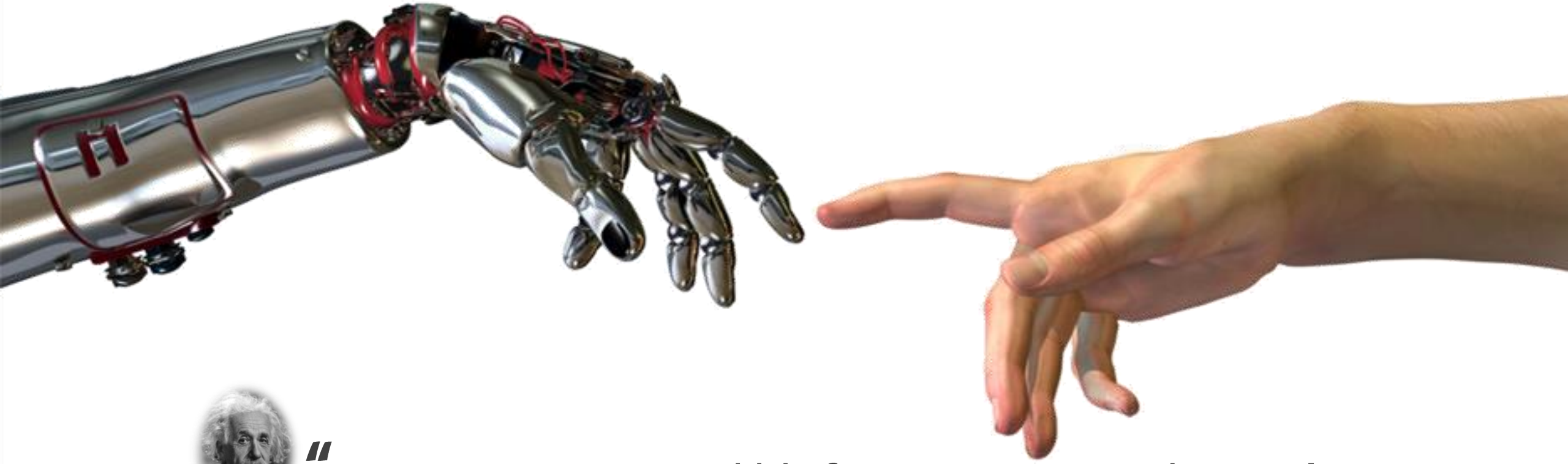
Automated techniques are not sufficient when…

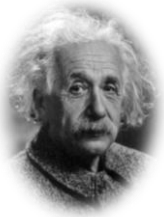
→ Data ambiguous and Incomplete

→ Complex Relationship

→ Semantic Gap / Domain- (and World-) Knowledge

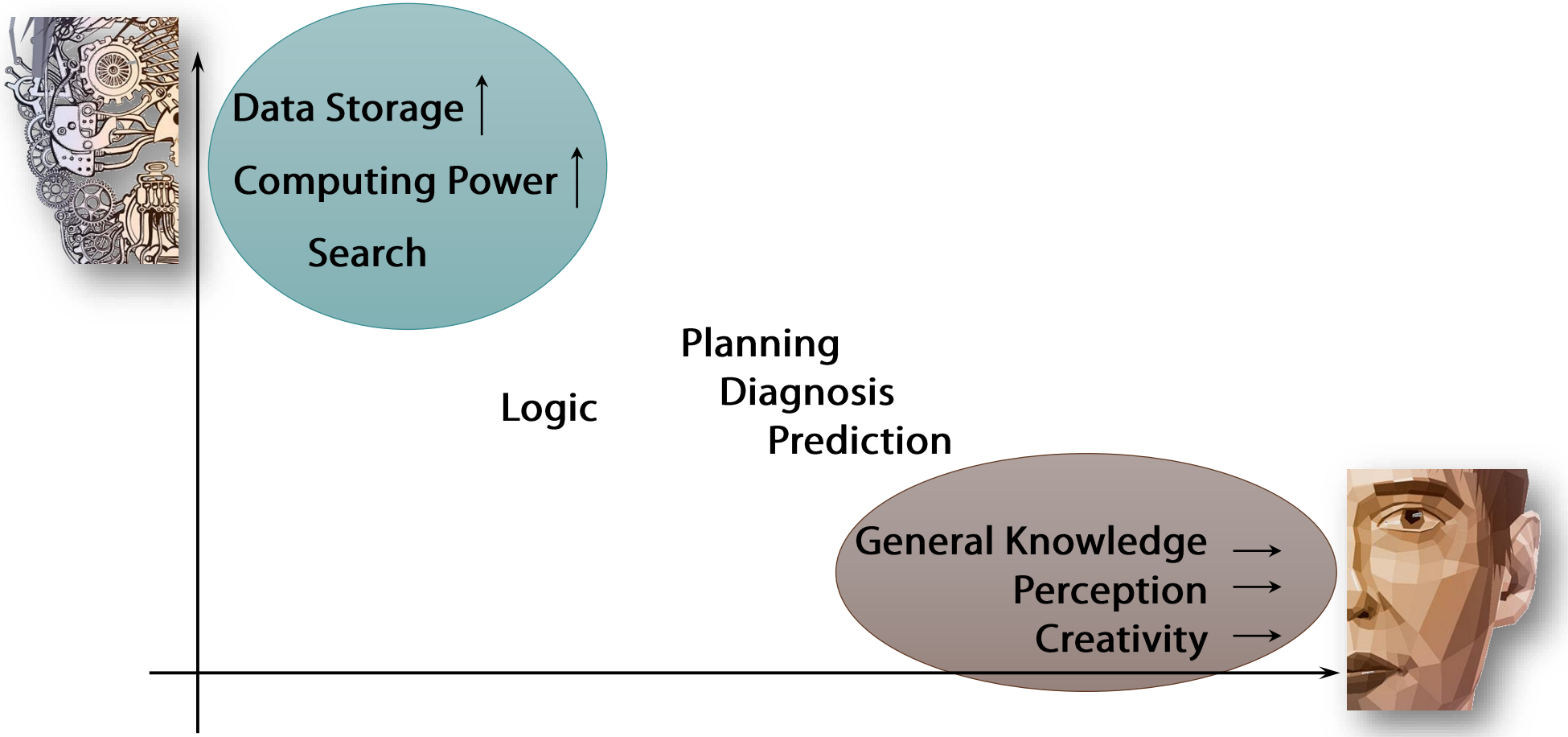
→ Limited Accuracy



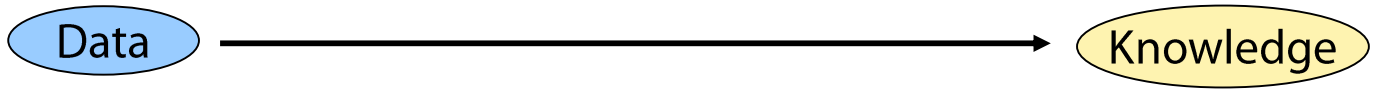


"Computers are incredibly **fast**, **accurate**, and **stupid**;
humans are incredibly **slow**, **inaccurate**, and **brilliant**;
together they are powerful beyond imagination."

attributed to **Albert Einstein**


Data Storage
Computing Power
Search
Logic
Planning
Diagnosis
Prediction
General Knowledge
Perception
Creativity

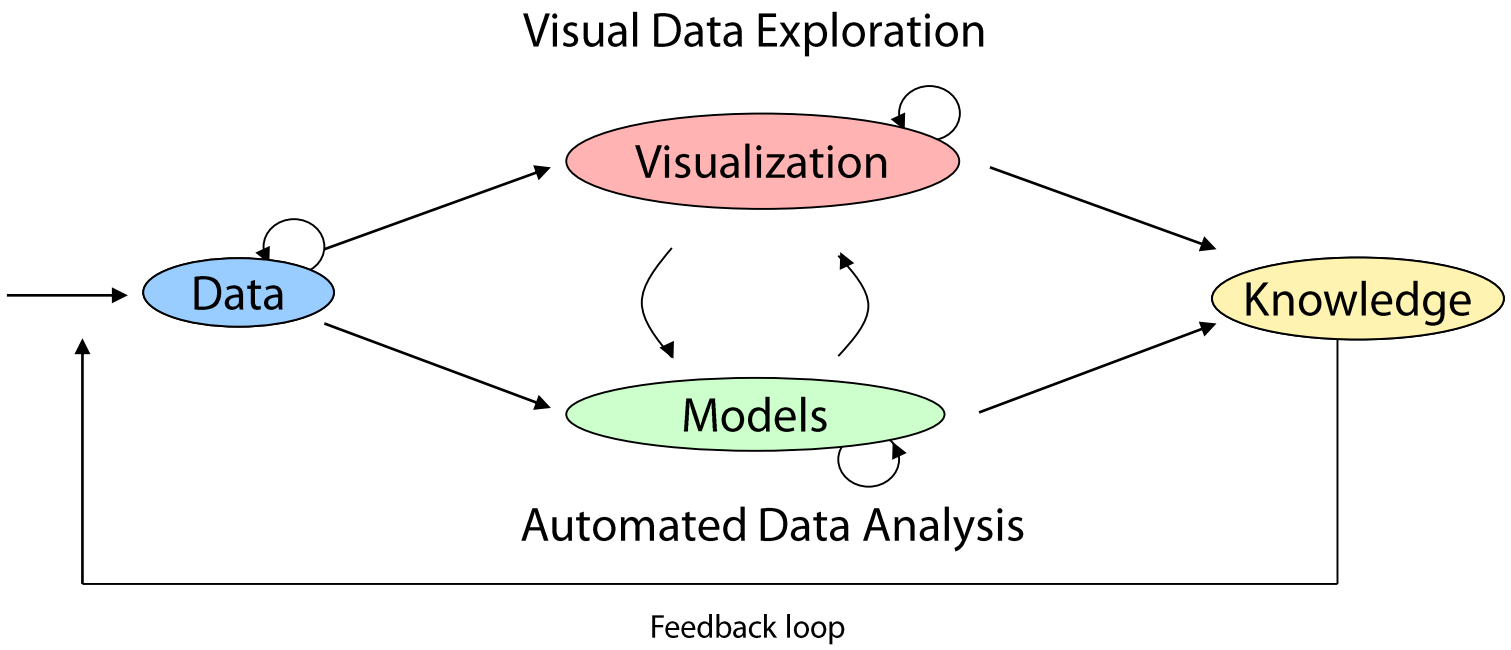
What is Visual Analytics?



Tight Integration of **Visual** and **Automatic Data Analysis Methods**

for **Information Exploration** and **Scalable Decision Support**

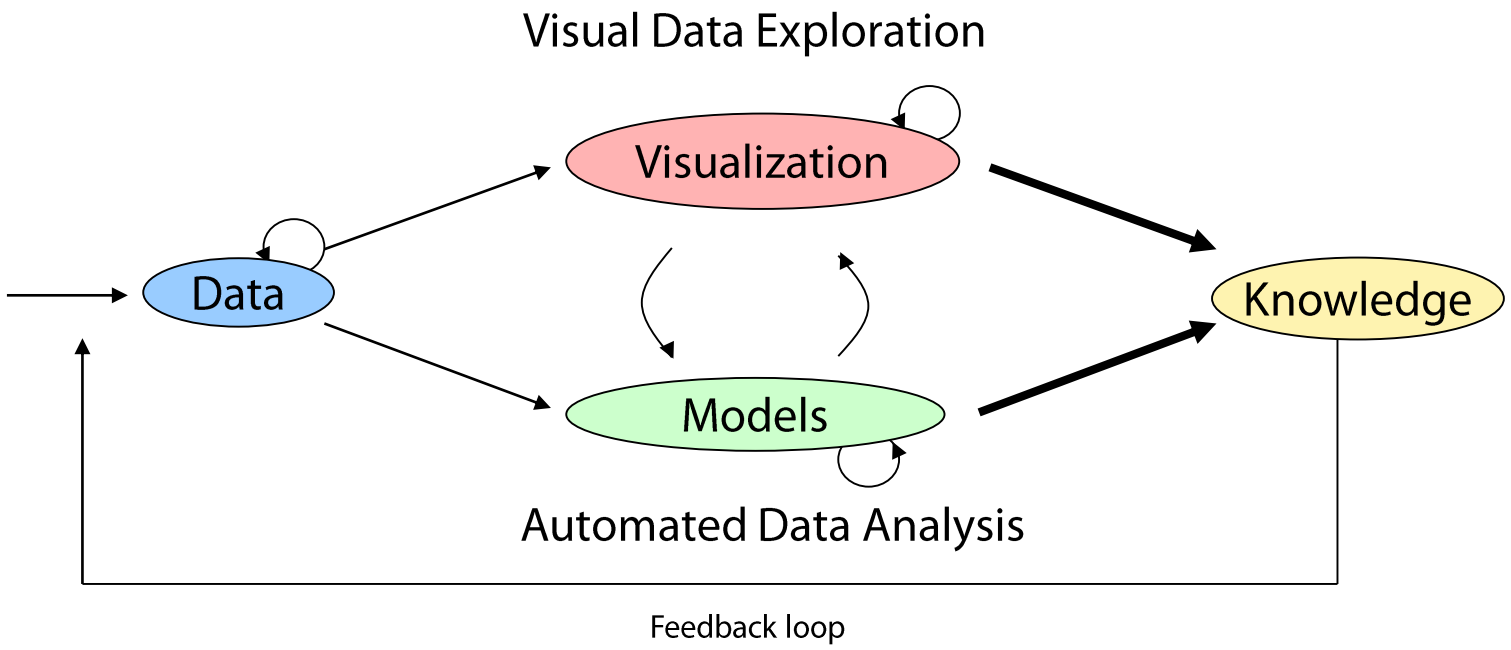
What is Visual Analytics?



Tight Integration of **Visual** and **Automatic Data Analysis Methods**

for **Information Exploration** and **Scalable Decision Support**

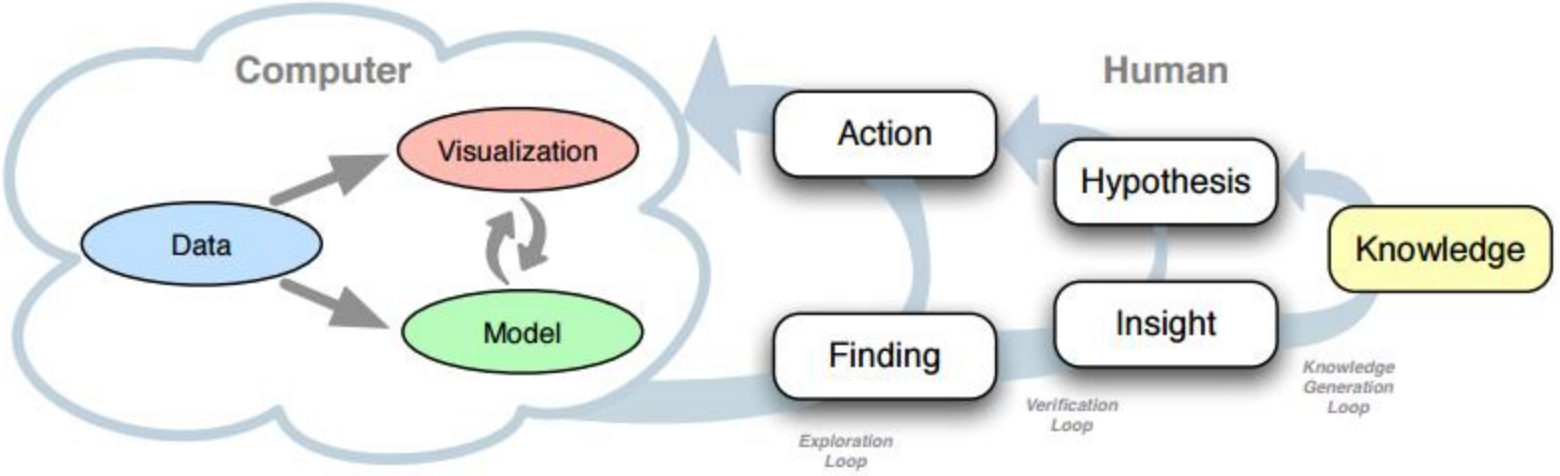
What is Visual Analytics?



Tight Integration of **Visual** and **Automatic Data Analysis Methods**

for **Information Exploration** and **Scalable Decision Support**

What is Visual Analytics?



Tight Integration of **Visual** and **Automatic Data Analysis Methods**

for **Information Exploration** and **Scalable Decision Support**

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

Resources

Introduction – Goals for Visualization

- **Presentation** (Communication)
  - Starting point: facts to be presented are fixed a priori
  - Process: choice of appropriate presentation techniques
  - Result: high-quality visualization of the data to present facts

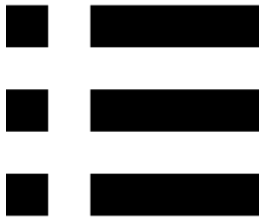
- **Confirmatory Analysis**
  - Starting point: hypotheses about the data
  - Process: goal-oriented examination of the hypotheses
  - Result: visualization of data to confirm or reject the hypotheses

- **Exploratory Analysis**
  - Starting point: no hypotheses about the data
  - Process: interactive, usually undirected search for structures, trends
  - Result: visualization of data to lead to hypotheses about the data

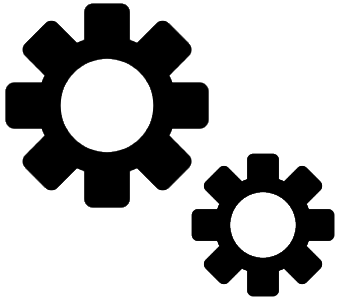
Introduction – Steps for Visualizing Data

Data Type

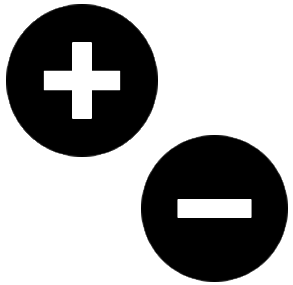
Questions



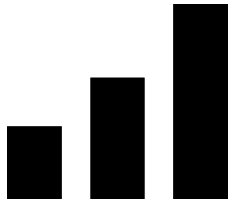
Requirements



Preprocessing



Pro/Cons



Design

Evaluation

Find out what kind of data you have.

Which questions shall be answered by the data?

Derive requirements from these tasks.

Preprocess data (missing data, outliers, transformations, …).

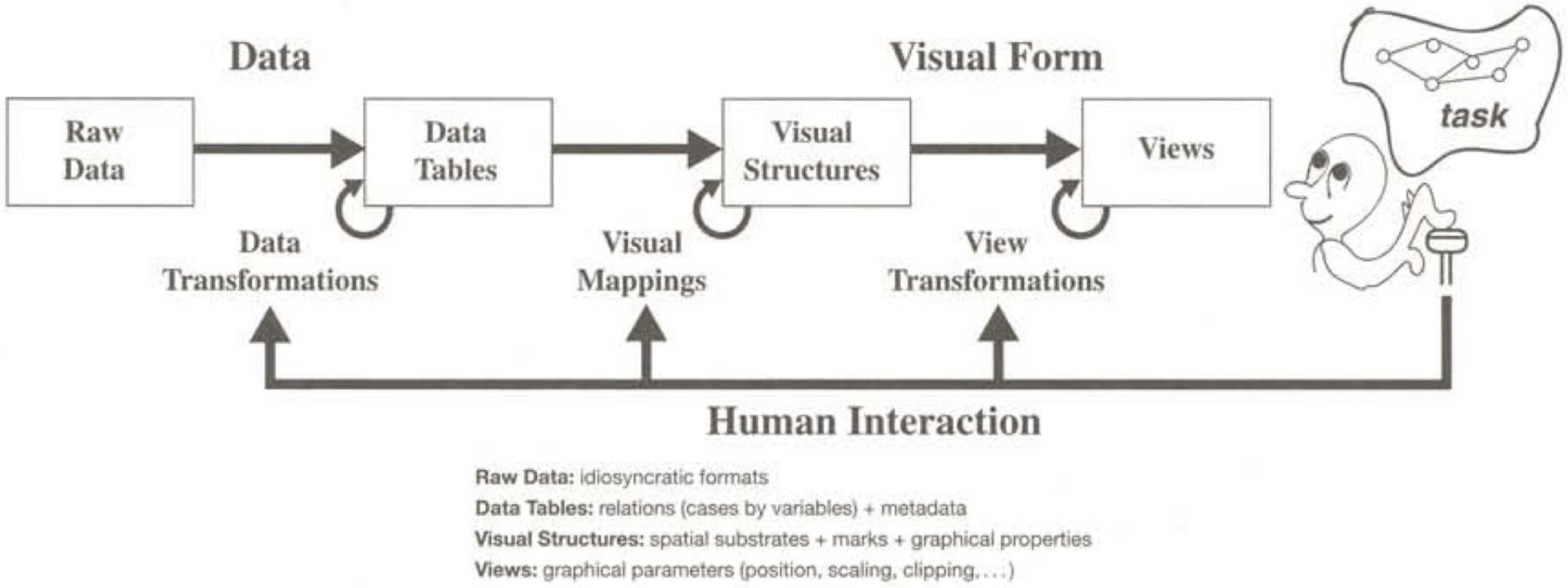
Consider the pros and cons of different visualization techniques (effectiveness, familiarity …)

Design a visualization system.

Test and improve it

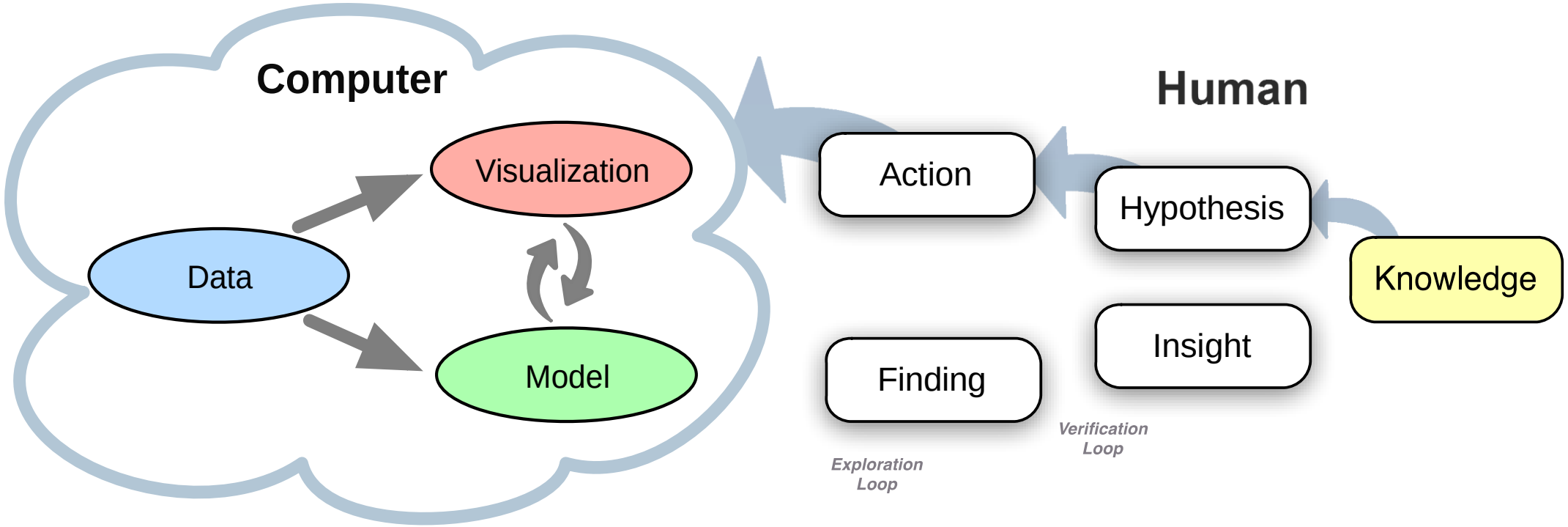
Introduction – Steps for Visualizing Data - Pipeline

"Reference Model for Information Visualization"



S. Card, J. Mackinlay, and B. Shneiderman. 1999. Readings in Information Visualization: Using Vision to Think. Chapter 1.

Knowledge Generation Model for Visual Analytics



D. Sacha, Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. A. .2014 . Knowledge Generation Model for Visual Analytics. IEEE Transactions on Visualization and Computer Graphics (Proceedings Visual Analytics Science and Technology 2014), 20(12):1604 – 1613.

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

Resources

Data Foundations – Data Types

Nominal

- No quantitative relationship between categories
- Classification without ordering

→ No intrinsic order

e.g. sports

Ordinal

- Attributes can be rank-ordered
- Distances between values do not have any meaning

→ Ordered in a sequence

e.g. places ordered by noise

Numeric

- Attributes can be rank-ordered
- Distances between values have a meaning
- Mathematical operations are possible

e.g. height of a person

References: C. Ware. 2012. Information Visualization: Perception for Design. Chapter 1.

Data Preprocessing – Real World Data is Dirty, What's Wrong?

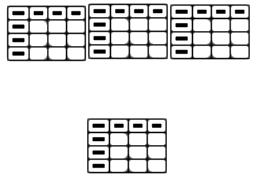
Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0		3	90 min
352	1	25	0	1h 43s
674	1	25	1	0,4 day
865	0	18	3	1h 50s
341	X	41	2	1h 33s

Data Foundations – Preprocessing

- Data Cleaning
  - Missing Values
  - Noise …

- Normalization
  - Common scale (0-1)
  - E.g., color mapping

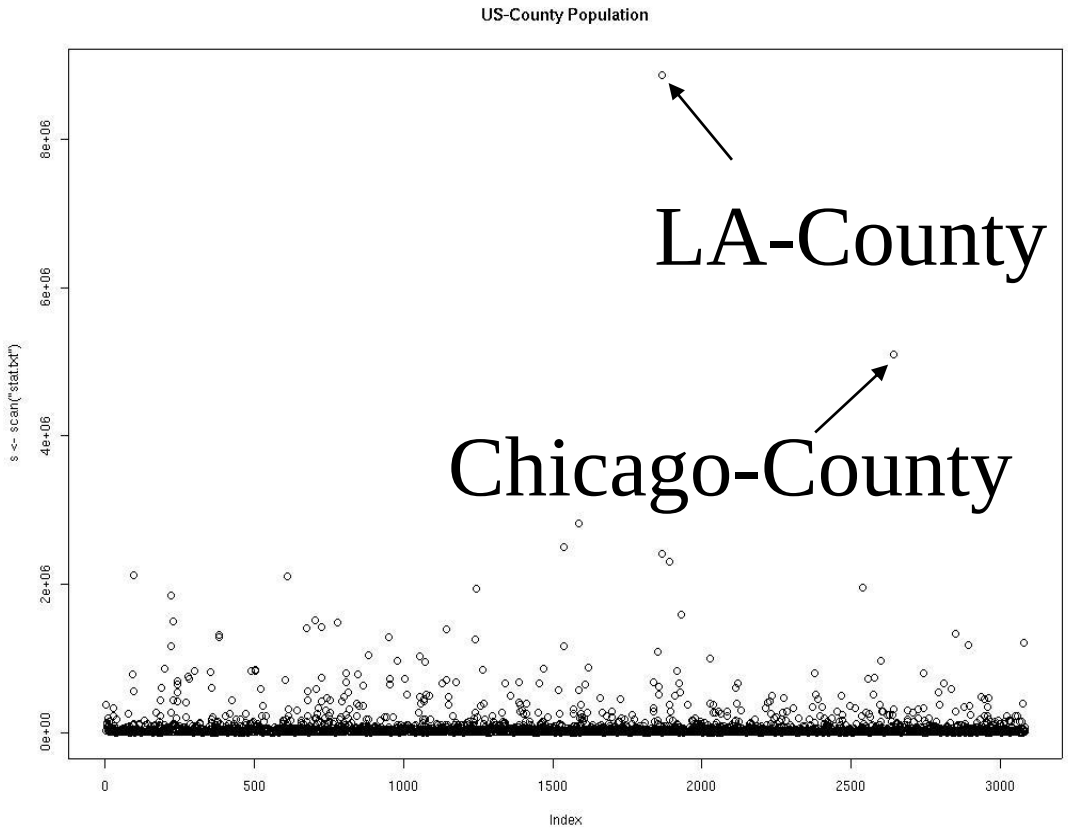
- Segmentation
  - Divide data in meaningful parts

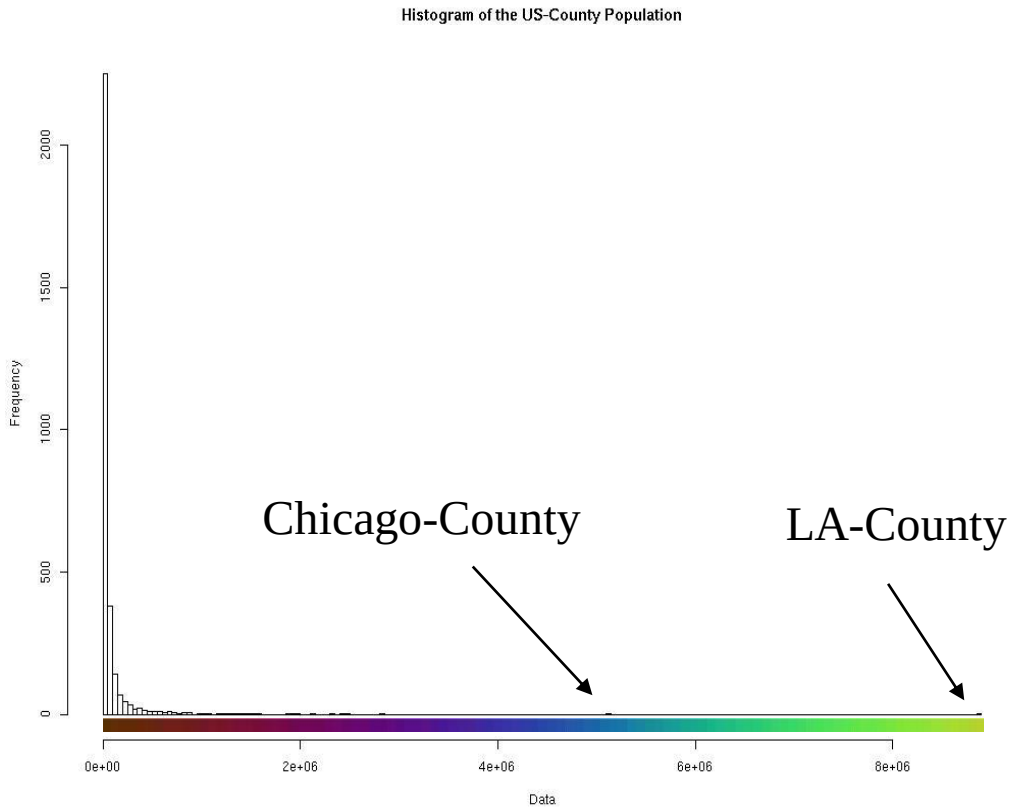


- Data Reduction
  - Sampling
  - Dim. Reduction

- ➢ Big part of our daily work!
- ➢ Methods have to be applied carefully
- ➢ Analysts needs to know which techniques have been applied! (Loss of Information)

Data Foundations – Normalization

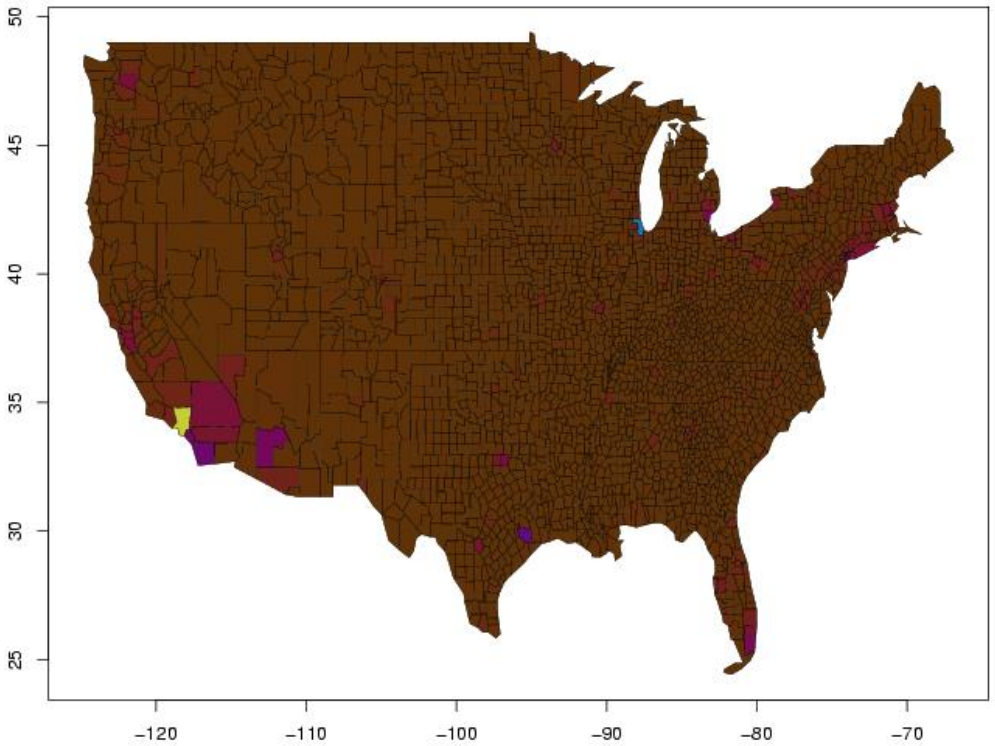


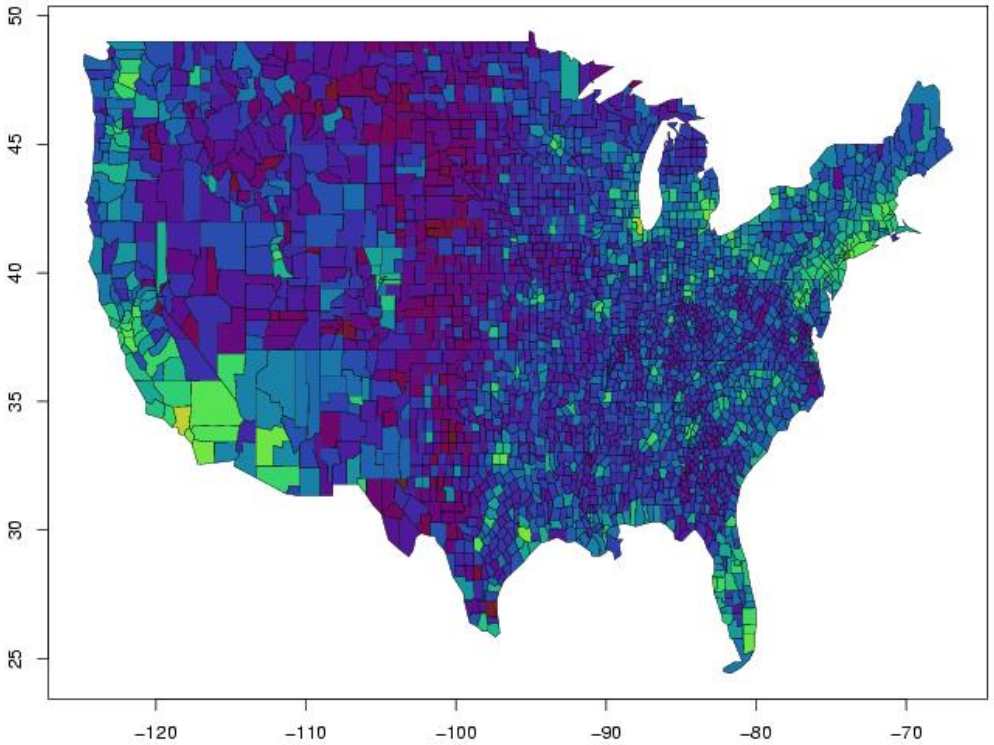


Data Foundations – Linear vs. Logarithmic Normalization

$$f_{lin}(v) = \frac{v - min}{max - min}$$

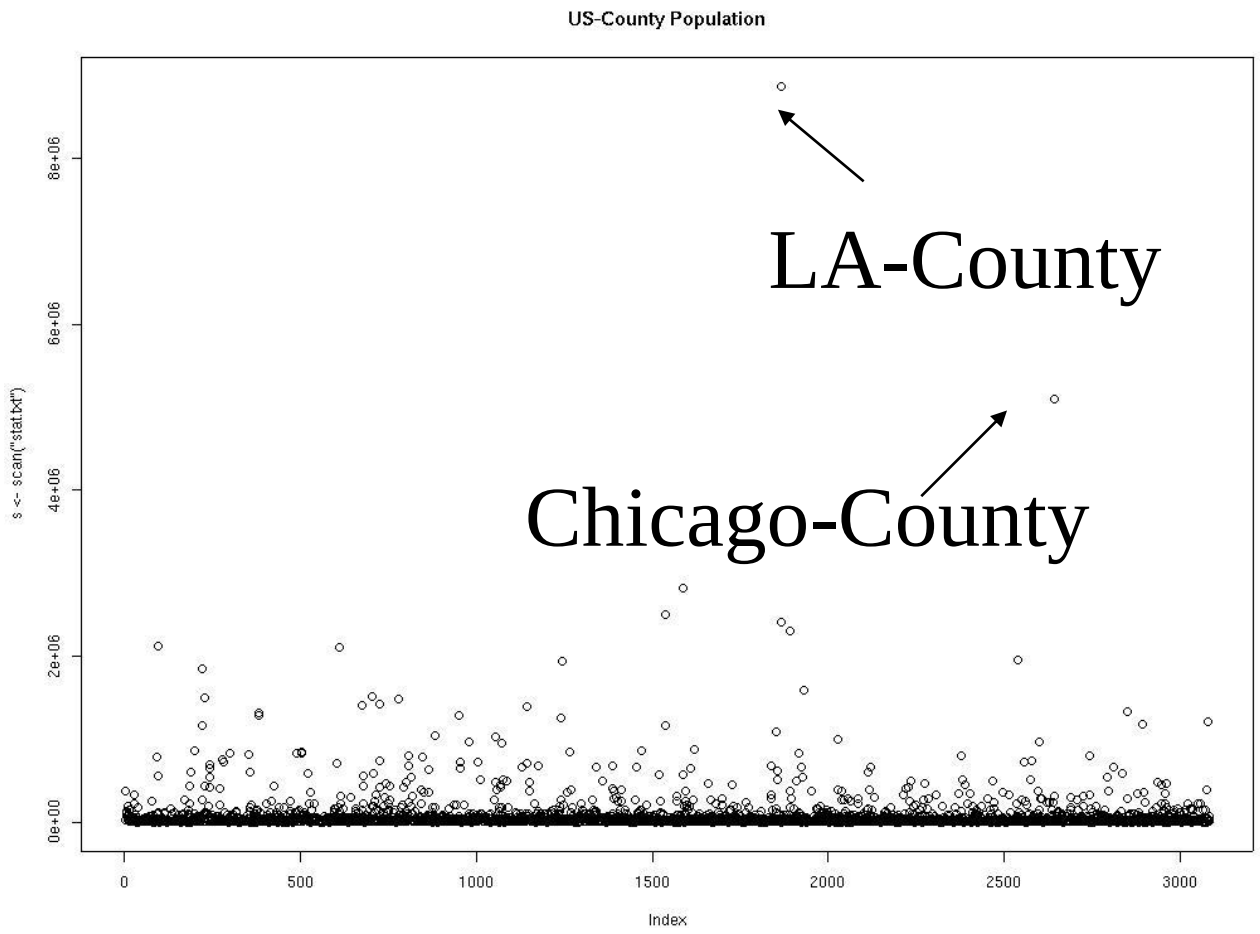
$$f_{\ln}(v) = \frac{\ln(v) - \ln(\min)}{\ln(\max) - \ln(\min)}$$

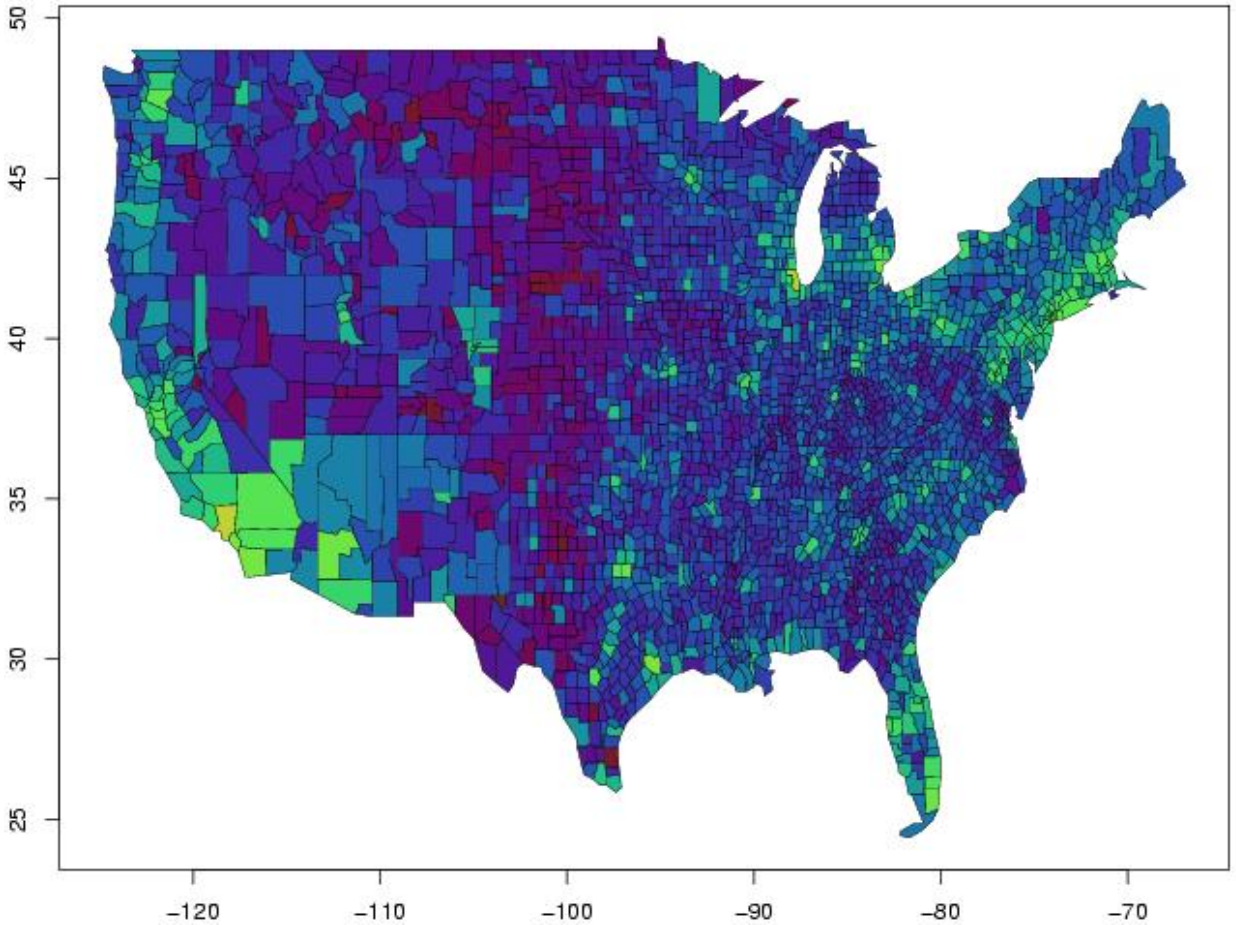




Data Foundations – Logarithmic Normalization

Advantages and Disadvantages?





Data Foundations – Practically Speaking

- Think carefully about appropriate preprocessing methods
  - Consider data type, class, tasks and data quality
  - “Tailor data to the analysis goals”
- Keep track and communicate which preprocessings have been applied and consider their impacts on the final data product
- Poorly designed preprocessings may create artificial patterns in the data leading to wrong conclusions or analysis results

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

Resources

Visualization Basics

- **Perceptual/Pre-Attentive Processing**
  - Certain low level visual aspects are recognized before *conscious* awareness
- **Gestalt Laws**
  - The *tendency* to **perceive** elements as belonging to a group, based on certain visual properties
- **Visual Variables**
  - The different visual aspects that can be used to **encode** information

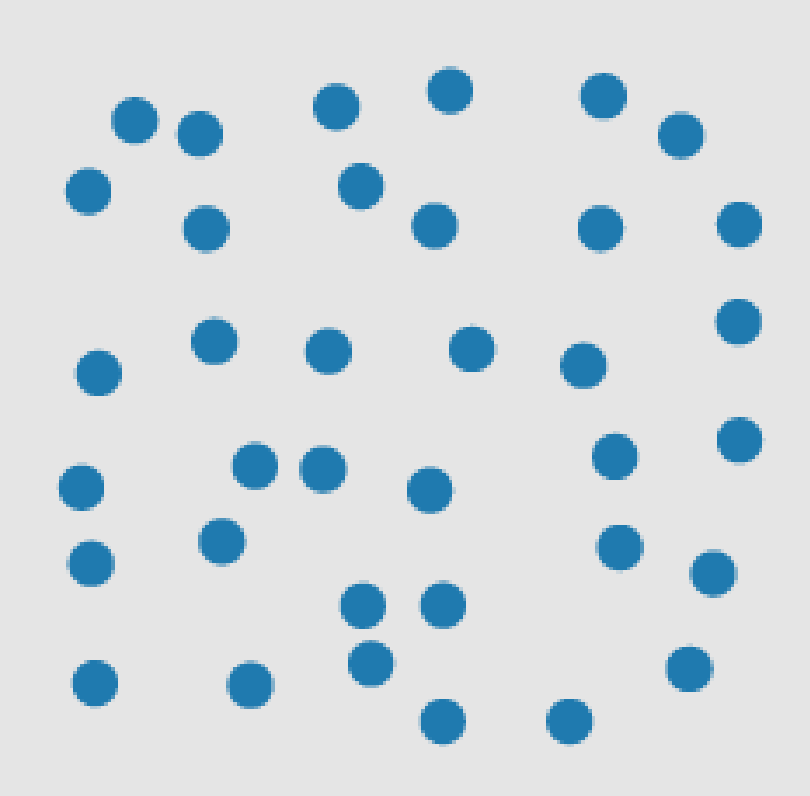
Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

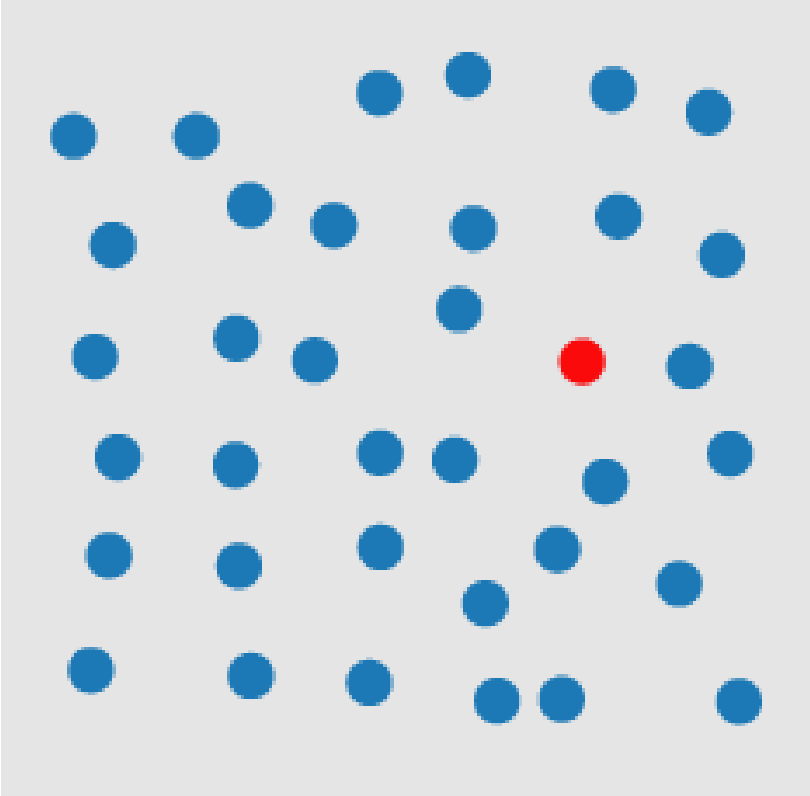
Preattentive Processing

- Perception of visual features managed by the low-level visual machinery

- Extremely fast: < 200 msec (the eyes take > 200 msec to initiate movement) => processed in parallel

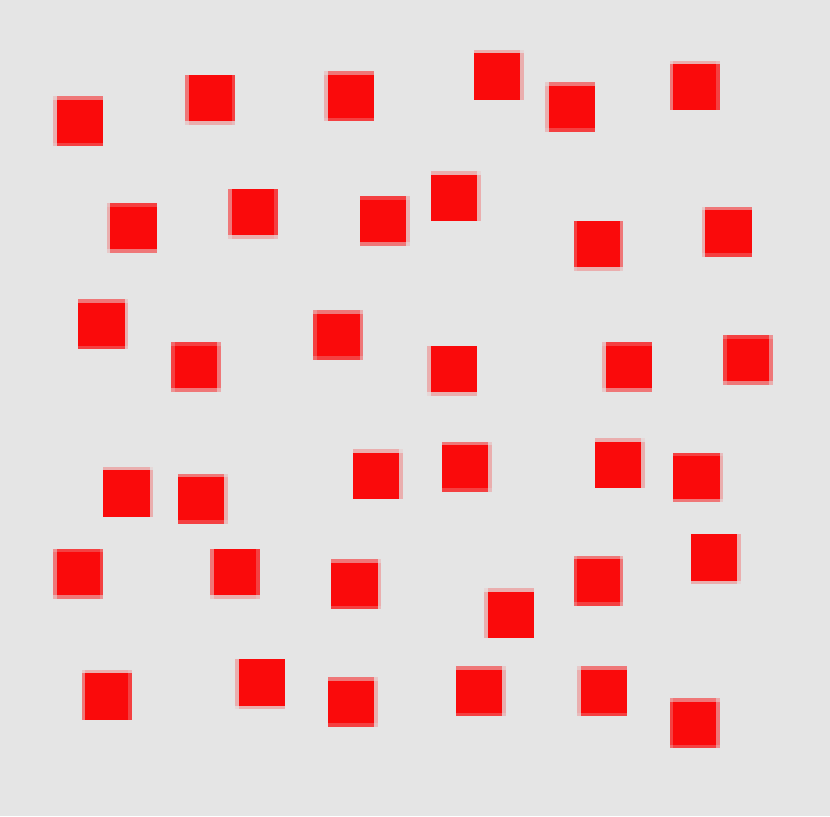
Preattentive Processing (color)

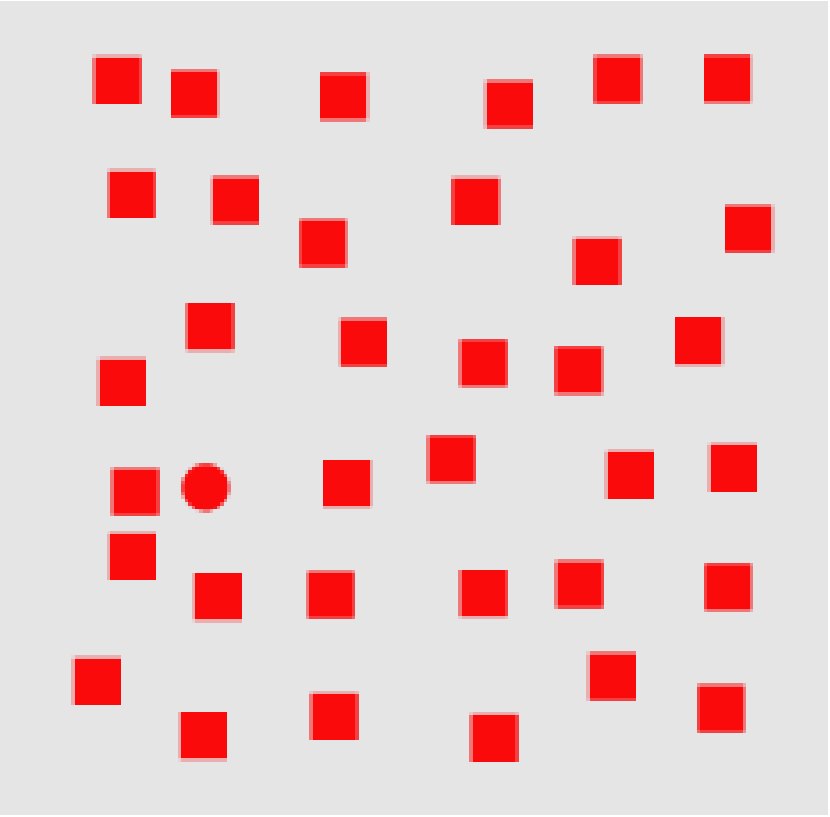




http://www.idvbook.com/

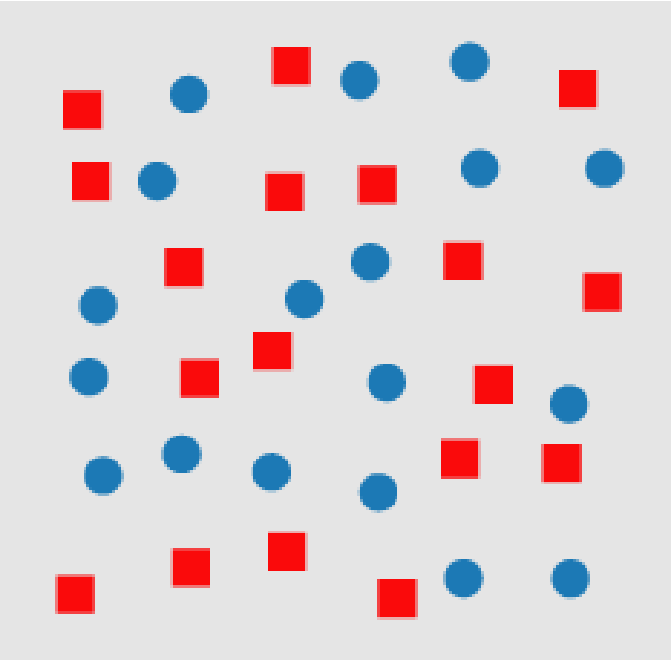
Preattentive Processing (shape)

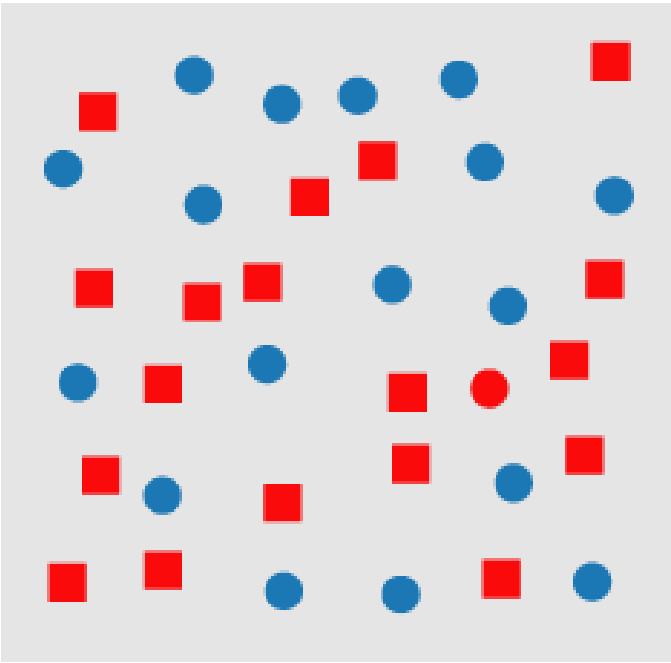




http://www.idvbook.com/

Serial Processing





http://www.idvbook.com/

- Finding the red circle requires serial processing
- The red circle is a conjuctive target (composed of non-unique features)

- Demo: http://www.csc.ncsu.edu/faculty/healey/PP/index.html

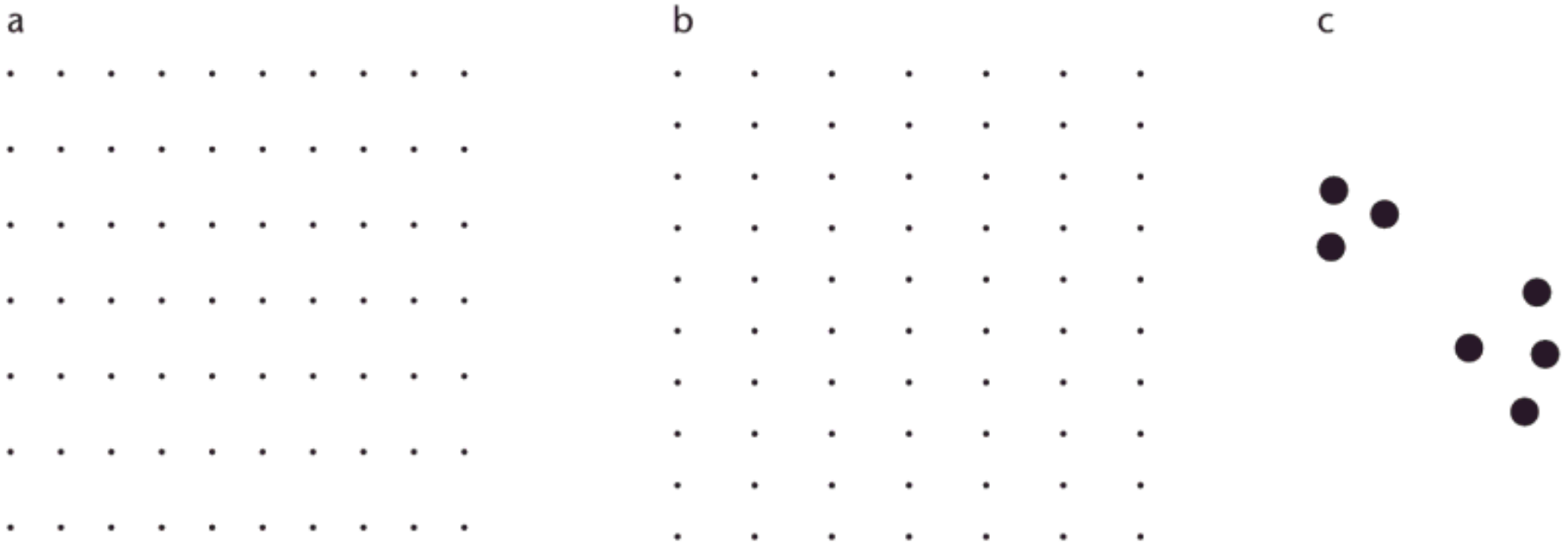
Gestalt Laws

Perceptual laws about how we **group** visual objects together to form visual entities.

1. Law of Proximity
2. Law of Similarity
3. Law of Connectedness
4. Law of Continuity
5. Law of Symmetry
6. Law of Closure and Common Region
7. Law of Figure and Ground

Gestalt Laws - Proximity

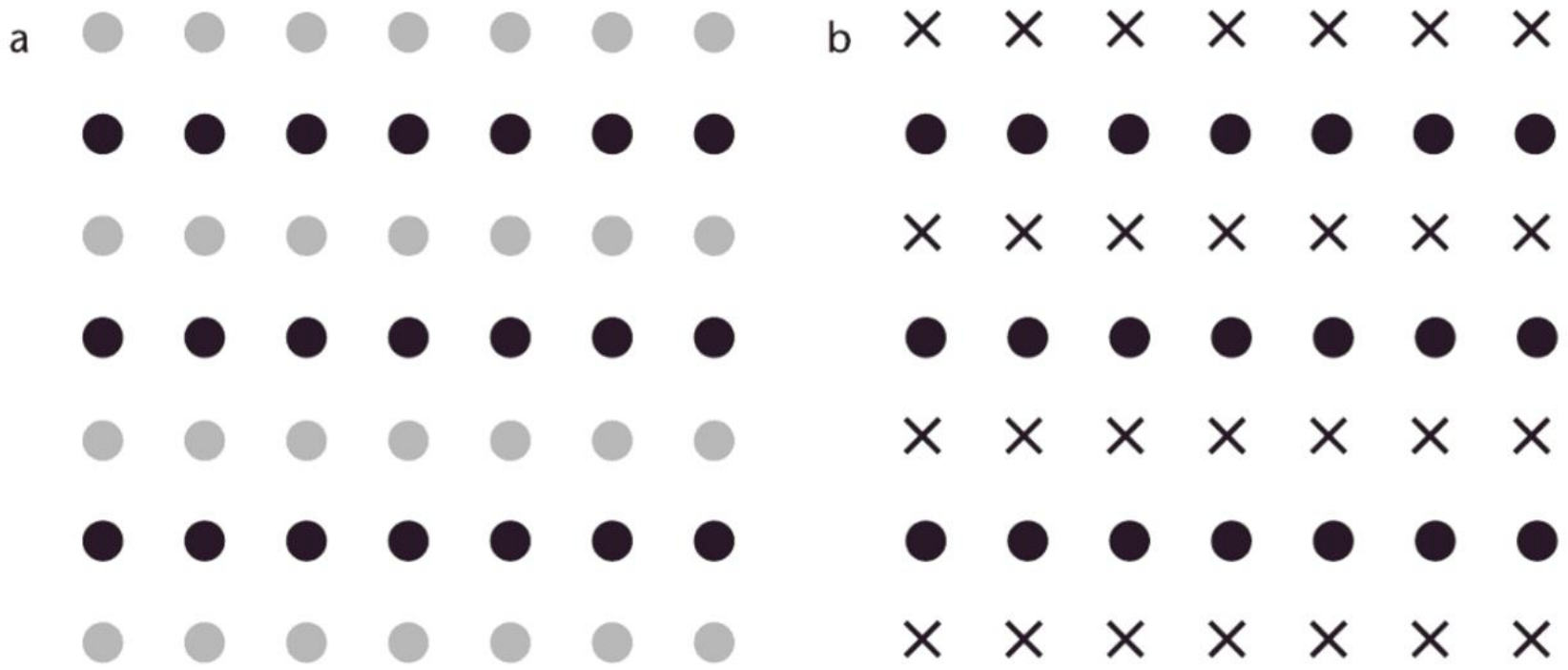
Close objects are perceptually grouped.



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws - Similarity

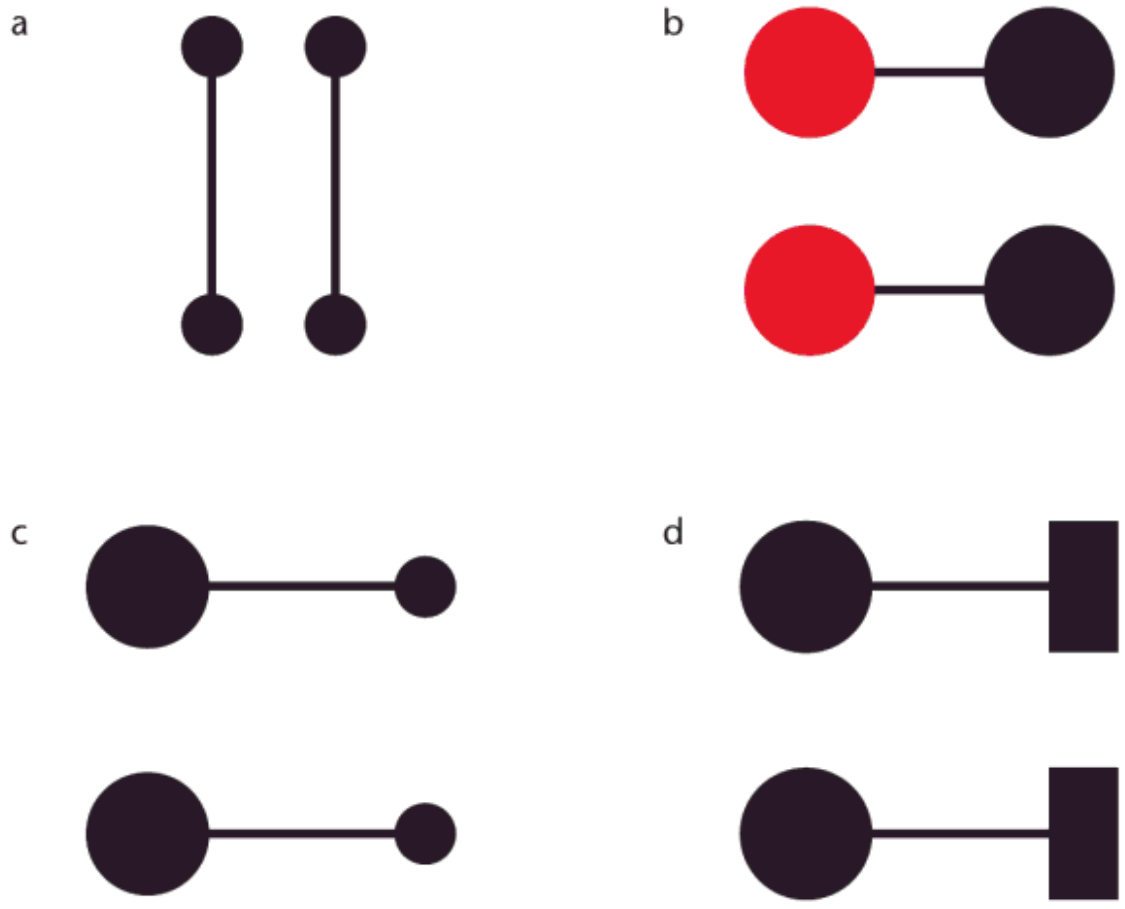
Similar objects are perceptually grouped.



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws - Connectedness

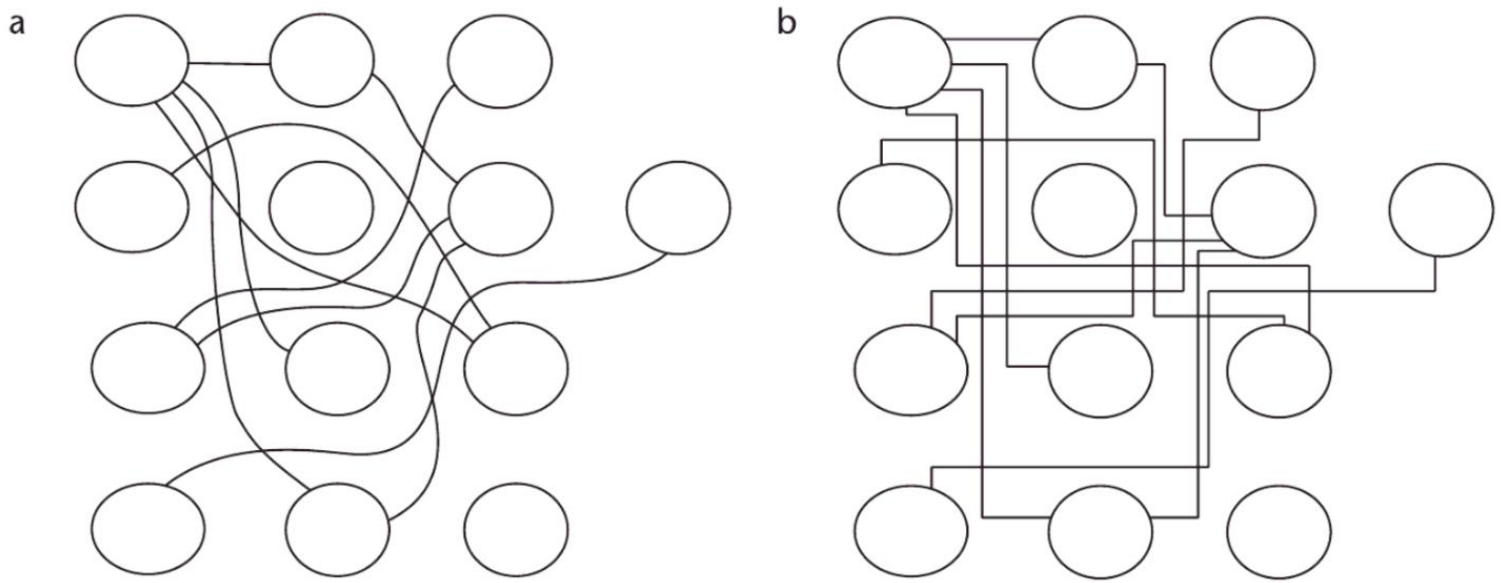
- Connected objects are perceptually grouped.
- Can be more powerful than proximity, color, size or shape.
- Used in node–link diagrams



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws - Continuity

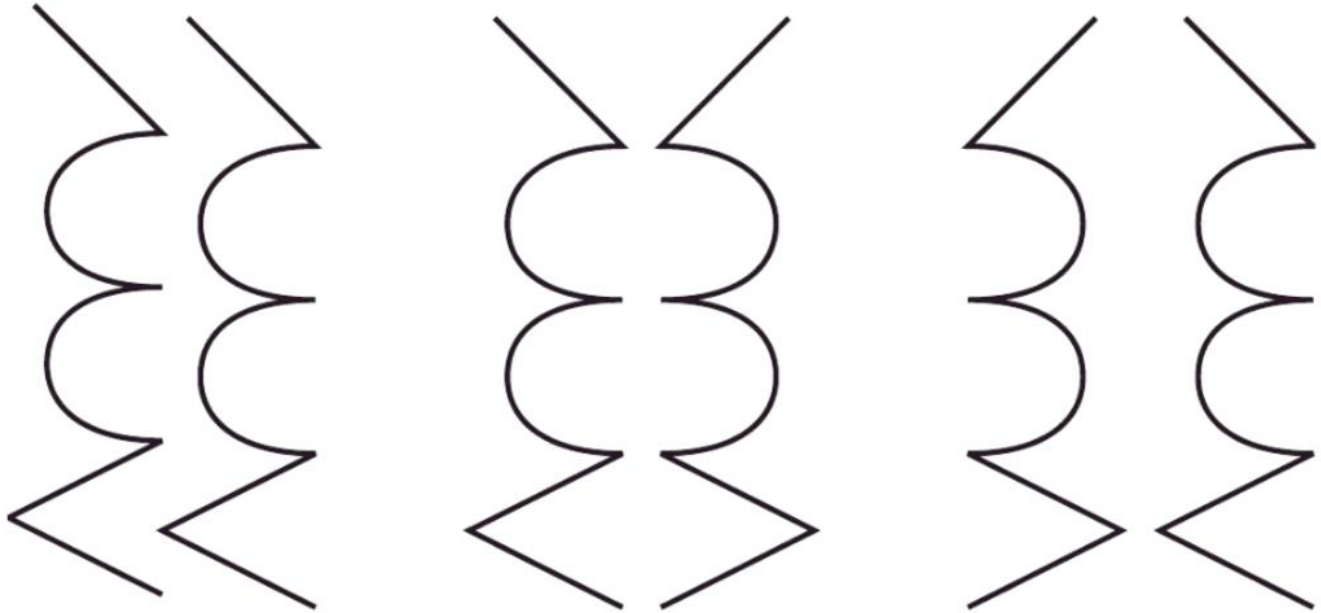
It is easier to construct visual objects out of visual elements if the connection is smooth and continuous.
Lines with abrupt changes in direction are harder to read.



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws - Symmetry

- powerful organization principle
- symmetric forms build a visual whole



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

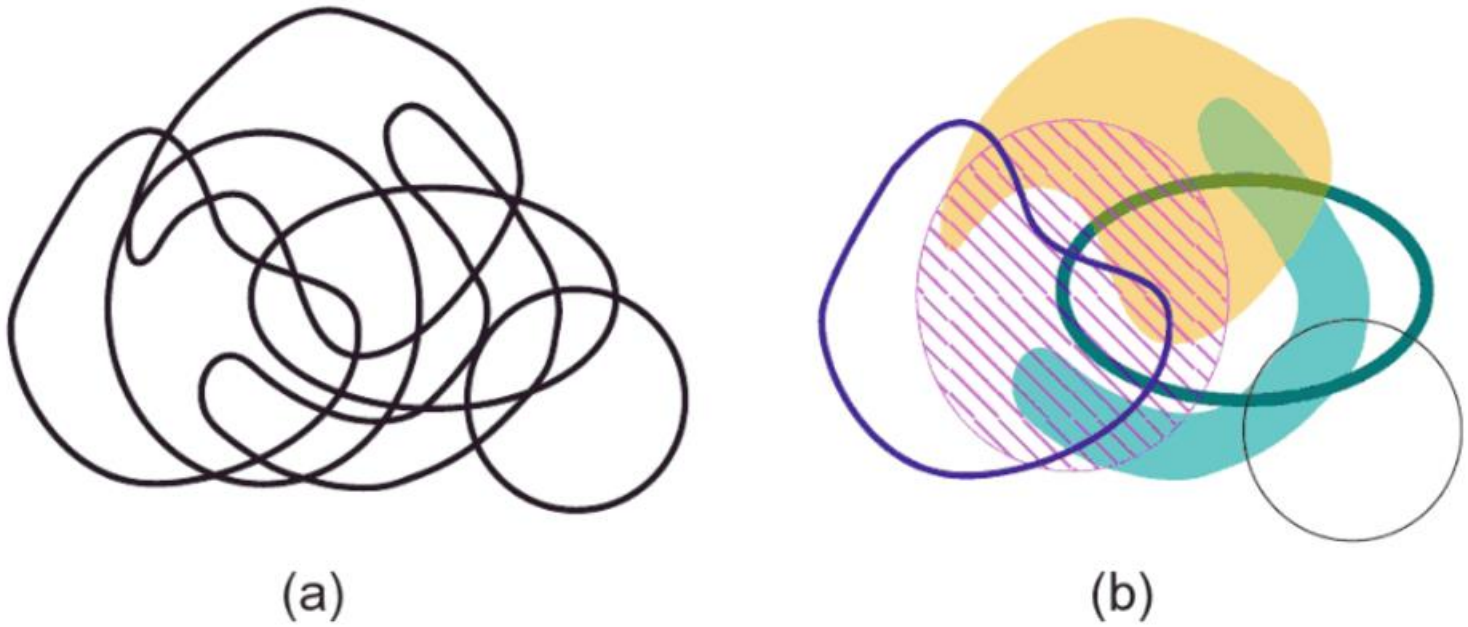
Gestalt Laws – Closure and Common Region

We prefer perceptual solutions with closed contours.
We see a ring behind a rectangle rather than a broken ring.

Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws – Closure and Common Region

Euler diagram to using color and texture to highlight contours.

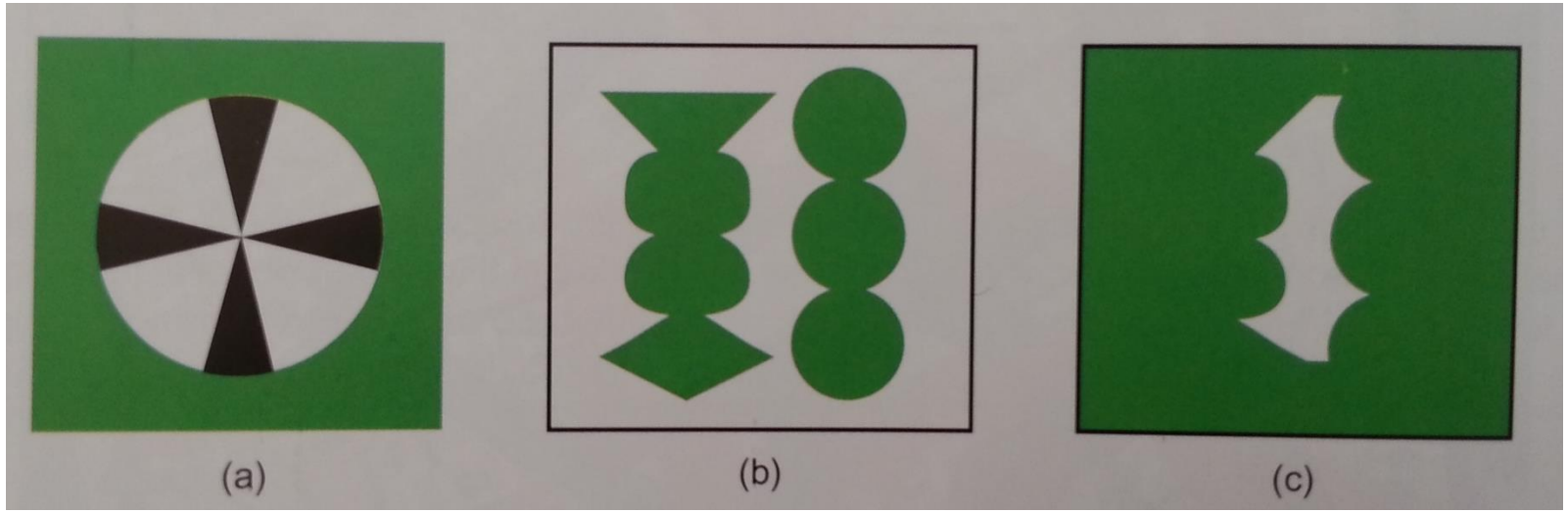


Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Gestalt Laws – Figure and Ground

- **figure**: object like, perceived to be in the foreground.
- **ground**: whatever lies behind the figure.

- fundamental to identify objects.
- all Gestalt laws contribute to create a figure



Colin Ware, Information Visualization: Perception for Design, Morgan Kaufmann

Some more examples you might know

Closure

Figure Ground

Proximity, Closure

https://de.wikipedia.org/wiki/WWF#/media/File:WWF_Logo.svg

https://en.wikipedia.org/wiki/Finder_%28software%29#/media/File:Mac_Finder_icon_%28OS_X_Yosemite%29.png

https://de.wikipedia.org/wiki/IBM#/media/File:IBM_logo.svg

Marks and Visual Variables

Jacques Bertin:

- A **mark** is "something that is visible and can be used ... to show relationships within sets of data"
- **Visual variables** are "the different ways that a mark can be varied"

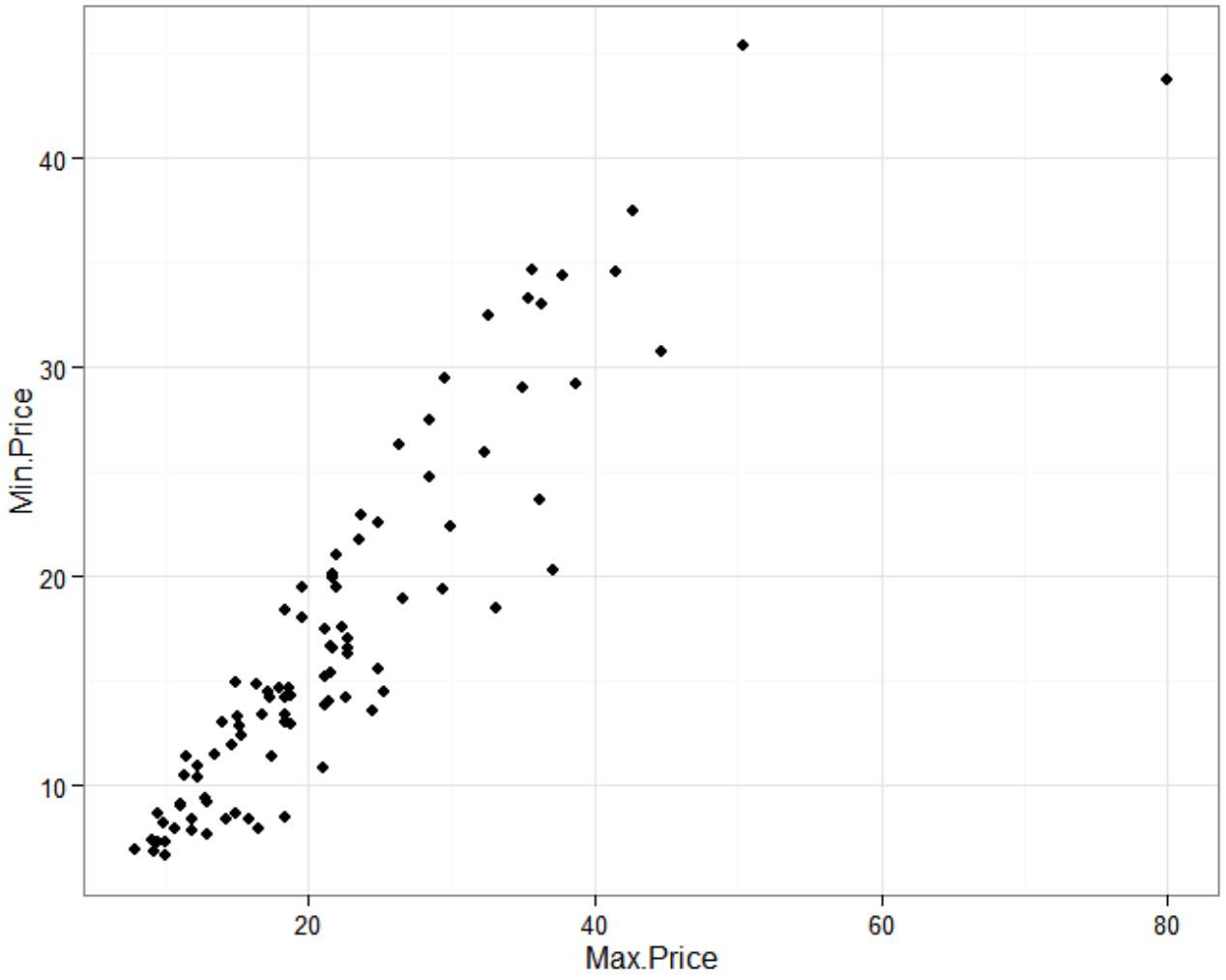
Reference:
S. Carpendale. 2003. Considering Visual Variables as a Basis for Visualization. Research report 2001-693-16, Department of Computer Science, University of Calgary.

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

The Eight Visual Variables

1. Position
2. Mark/Shape
3. Size (Length, Area and Volume)
4. Brightness
5. Color
6. Orientation
7. Texture
8. Motion

The Eight Visual Variables - Position

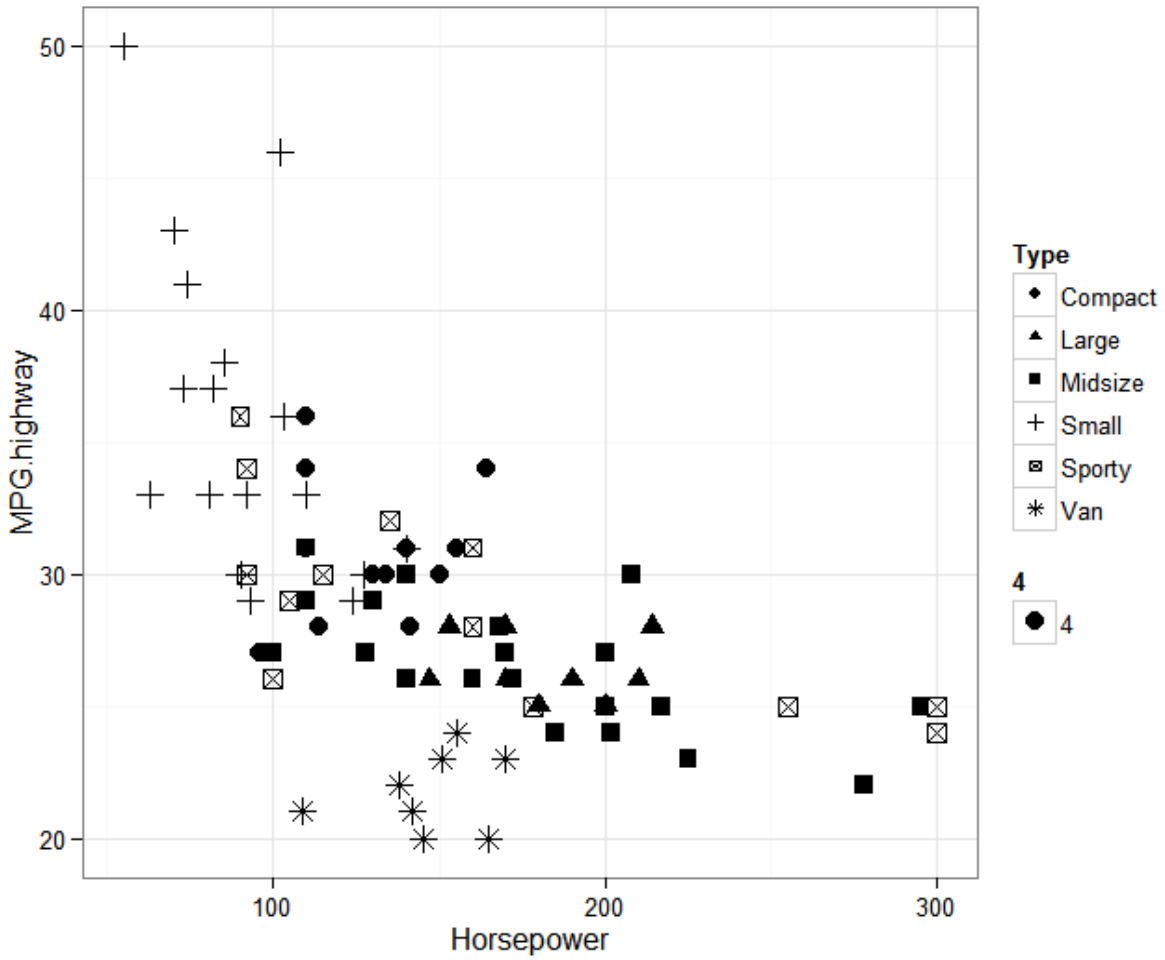


The Eight Visual Variables – Shape/Mark

Several examples of different marks or glyphs that can be used.

http://www.idvbook.com/

The Eight Visual Variables – Shape/Mark



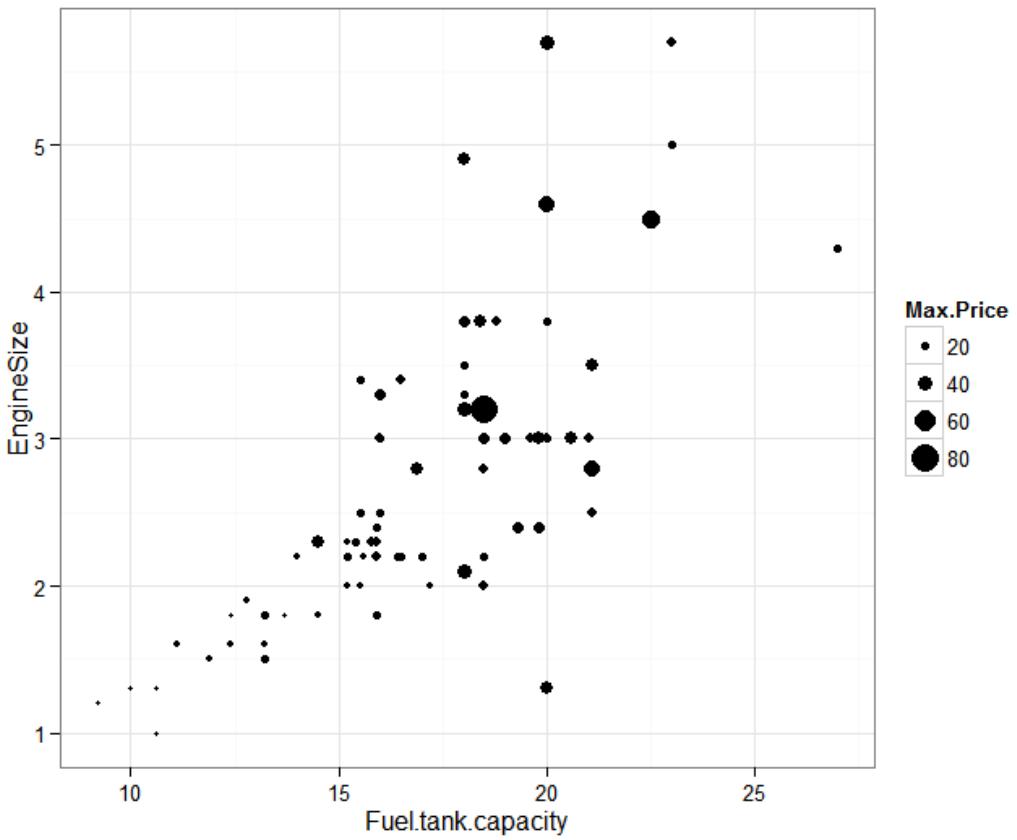
This visualization uses shapes to distinguish between different car types in a plot comparing highway MPG and horsepower. Clusters are clearly visible, as well as some outliers.

The Eight Visual Variables – Size (Length, Area and Volume)

Example sizes to encode data.

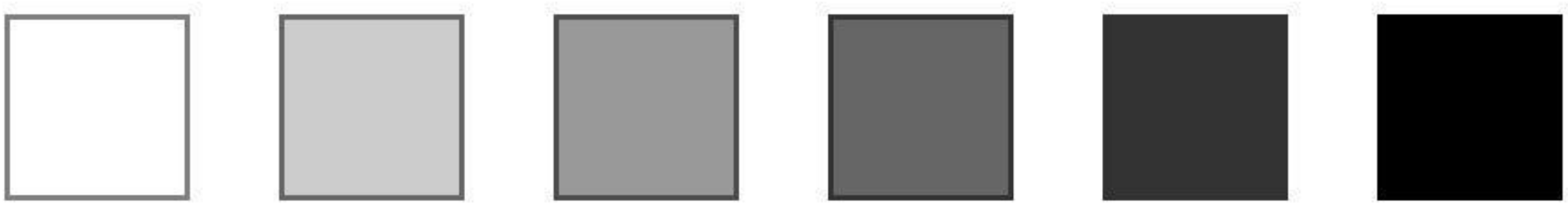
http://www.idvbook.com/

The Eight Visual Variables – Size (Length, Area and Volume)



This is a visualization of the 1993 car models data set, showing engine size versus fuel tank capacity. Size is mapped to maximum price charged.

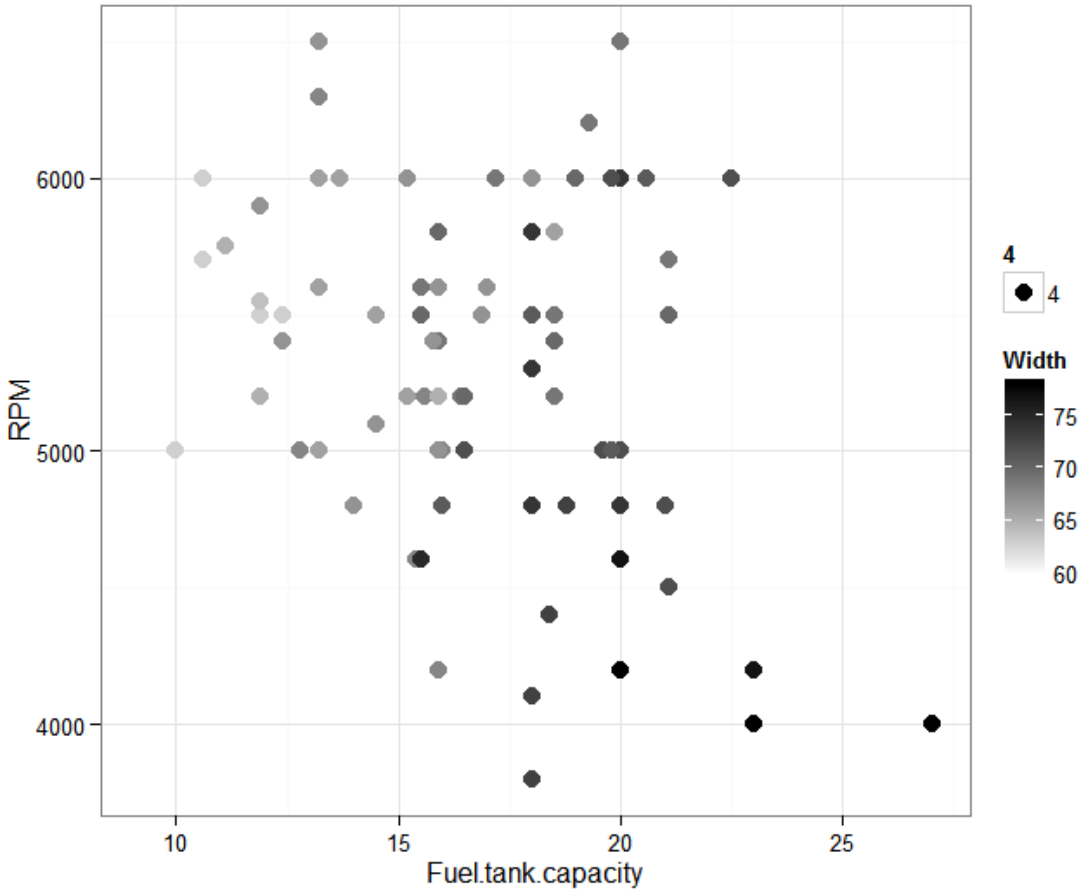
The Eight Visual Variables – Brightness



Brightness scale for mapping values to the display.

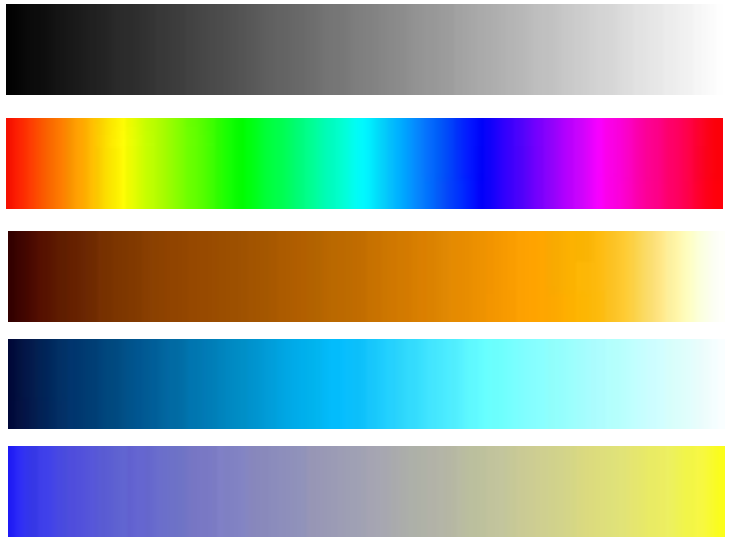
http://www.idvbook.com/

The Eight Visual Variables – Brightness



Another visualization of the 1993 car models data set, this time illustrating the use of brightness to convey car width (the darker the points, the wider the vehicle).

The Eight Visual Variables – Color



standard linear gray scale

rainbow

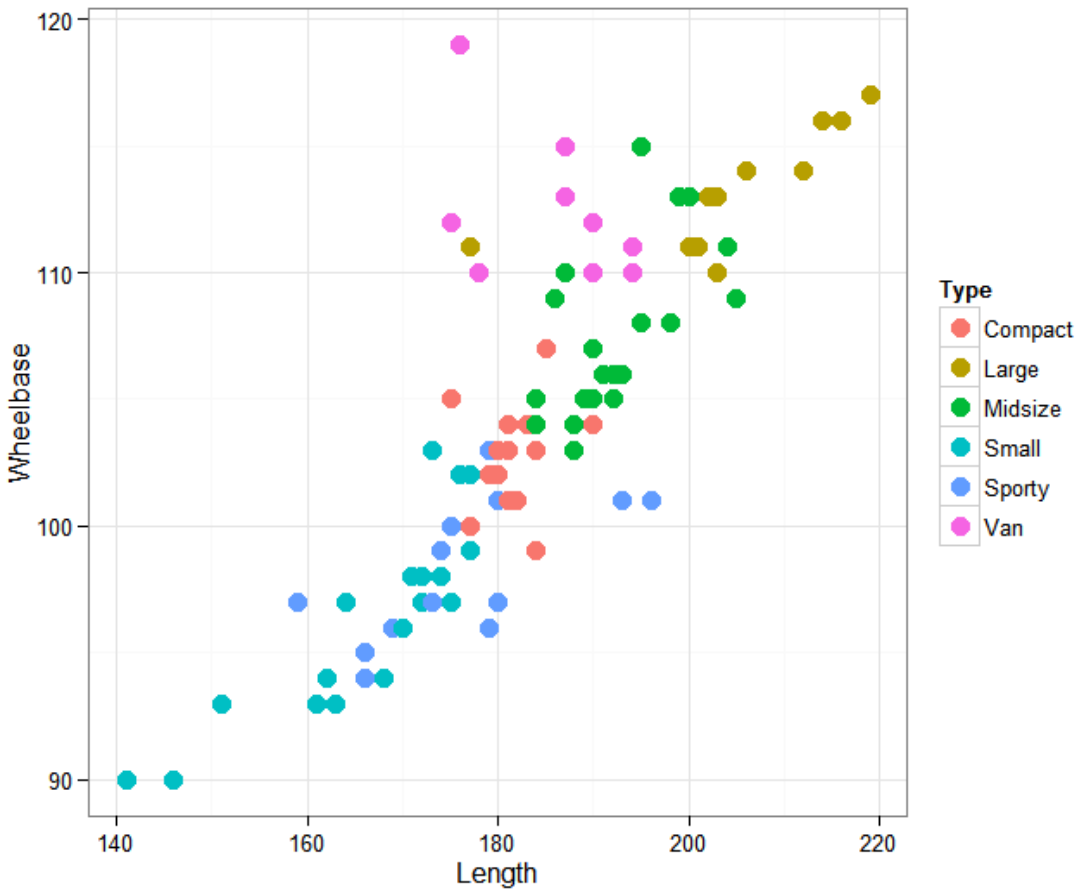
heated

blue to cyan

blue to yellow

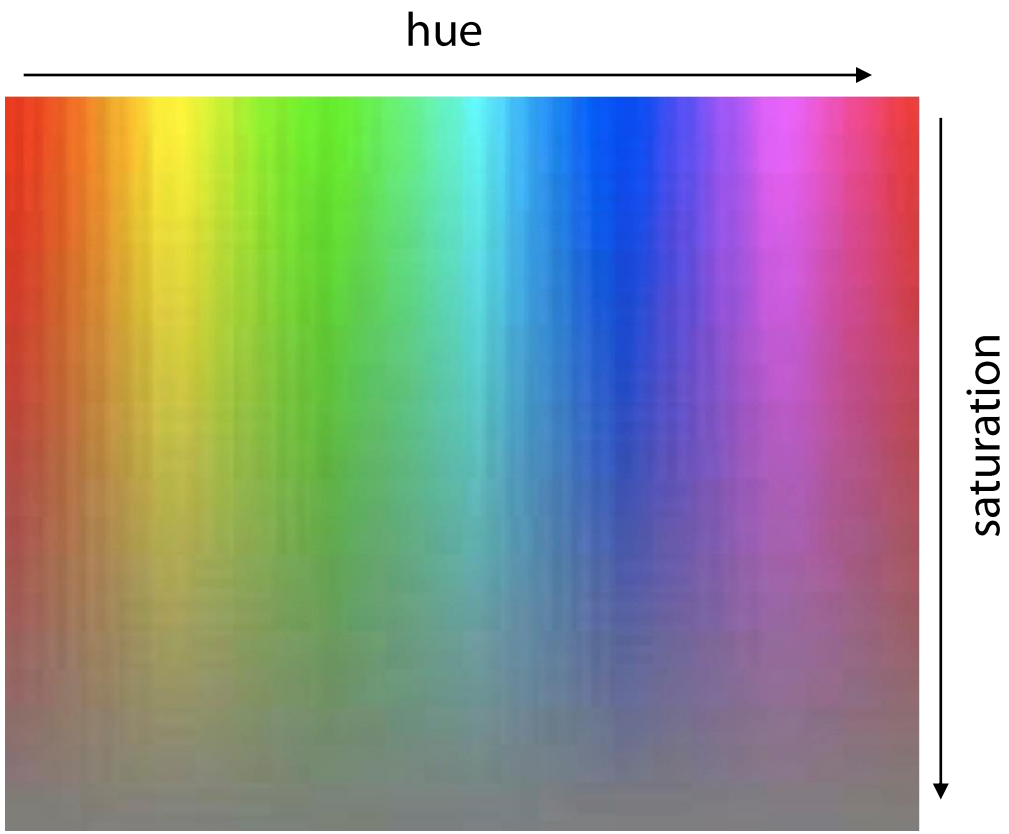
http://www.idvbook.com/

The Eight Visual Variables – Color



A visualization of the 1993 car models, showing the use of color to display the car's length. Here length is also associated with the *y*-axis and is plotted against wheelbase. In this figure, blue indicates a shorter length, while yellow indicates a longer length.

The Eight Visual Variables – Color



Microsoft hue/saturation color selector.

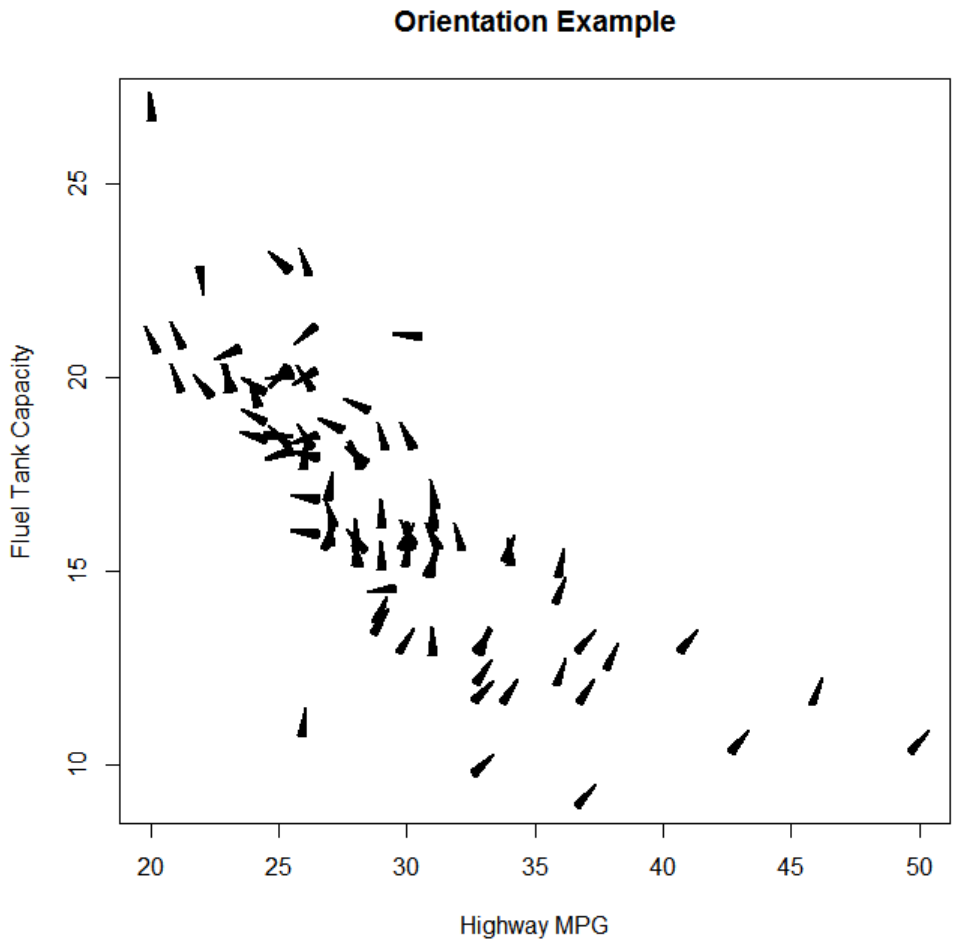
http://www.idvbook.com/

The Eight Visual Variables – Orientation

Example orientations of a representation graphic, where the lowest value maps to the mark pointing upward and increasing values rotate the mark in a clockwise rotation.

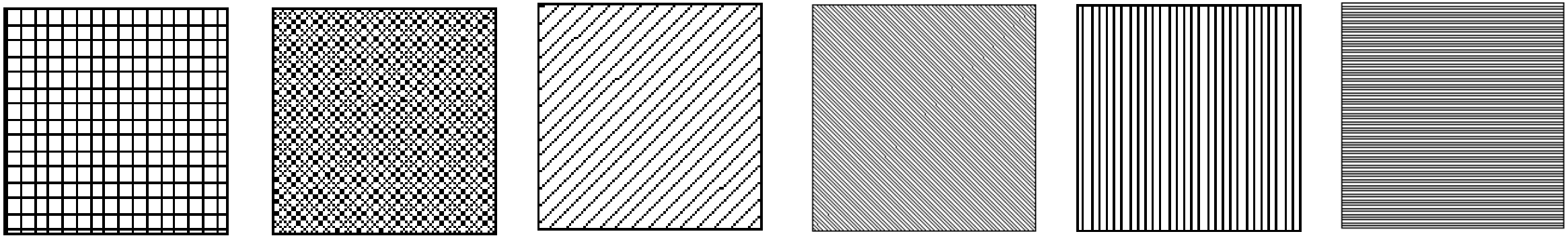
http://www.idvbook.com/

The Eight Visual Variables – Orientation



Sample visualization of the 1993 car models data set depicting using highway miles per-gallon versus fuel tank capacity (position) with the additional data variable, midrange price, used to adjust mark orientation.

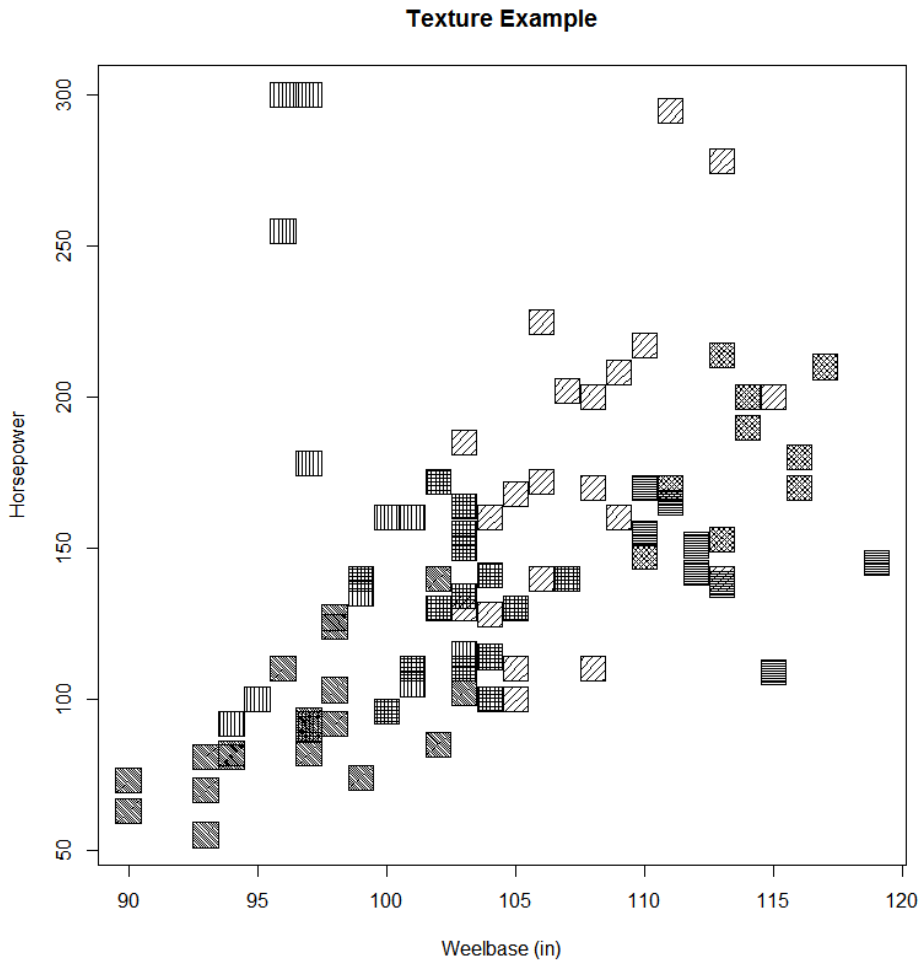
The Eight Visual Variables – Texture



Six possible example textures that could be used to identify different data values.

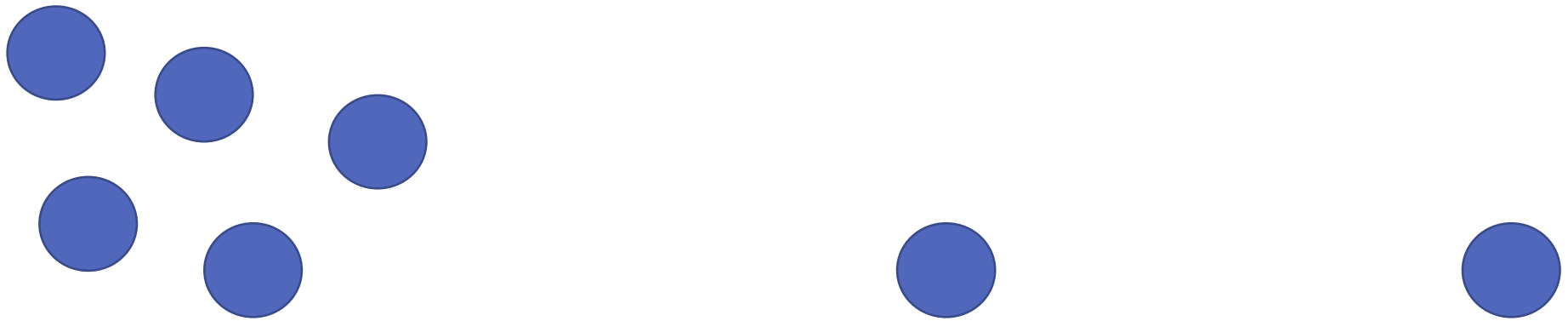
http://www.idvbook.com/

The Eight Visual Variables – Texture



Example visualization using texture to provide additional information about the 1993 car models data set, showing the relationship between wheelbase versus horsepower (position) as related to car types, depicted by different textures.

The Eight Visual Variables – Motion



- Can be associated with any of the other visual variables

Example : Gapminder

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

Resources

Characteristics of Visual Variables

- Selective: Is X different from the others?
- Associative: Is X like the others?
- Order: Is X more/greater/bigger/… than Y?
- Quantitative: How much is the difference between X and Y?
- Length: How many different categories can we represent with this variable for a task?

Reference: Summary of:
S. Carpendale. 2003. Considering Visual Variables as a Basis for Visualization. Research report 2001-693-16, Department of Computer Science, University of Calgary.

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

Characteristics of Visual Variables

	Selective	Associative	Quantitative	Order	Length
Position	✔	✔	✔	✔	✔
Size	✔	✔	≈	✔	(✔)
Shape	≈	≈	✗	✗	✔
Value	✔	✔	✗	✔	(✔)
Color (Hue)	✔	✔	✗	✗	(✔)
Orientation	(✔)	(✔)	✗	✗	(✔)
Grain	✔	(✔)	✗	✗	(✔)
Pattern	≈	≈	✗	✗	✔
Texture	✔	✔	✗	✗	✔
Motion	✔	✔	≈	✔	✔

Reference: Summary of:
S. Carpendale. 2003. Considering Visual Variables as a Basis for Visualization. Research report 2001-693-16, Department of Computer Science, University of Calgary.

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

Using Visual Variables - Practically Speaking

- Don't use ordered variables for nominal data
- Use position for the most important information
- Many lengths are ~5 — a limiting factor
- For quantity, use (vertical) position, length, size, brightness, saturation
- For many distinctions: use size, shape or brightness
- Redundant encoding *may* make features easier to interpret

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

Color– Some Notes

Color is powerful, but there are special considerations

- Hue is nominal, but can be given (cultural) orderings (e.g. temperature, elevation)
- Affective connotations of color vary across cultures
- Depending on who your users are, you may need to design for colorblindness or low-vision
  Allowing users to set colors / color schemes is the most flexible.

Resource:
Sim Daltonism: color blindness simulator for OS X (There other tools, and for other platforms)

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

Color - Some Guidelines

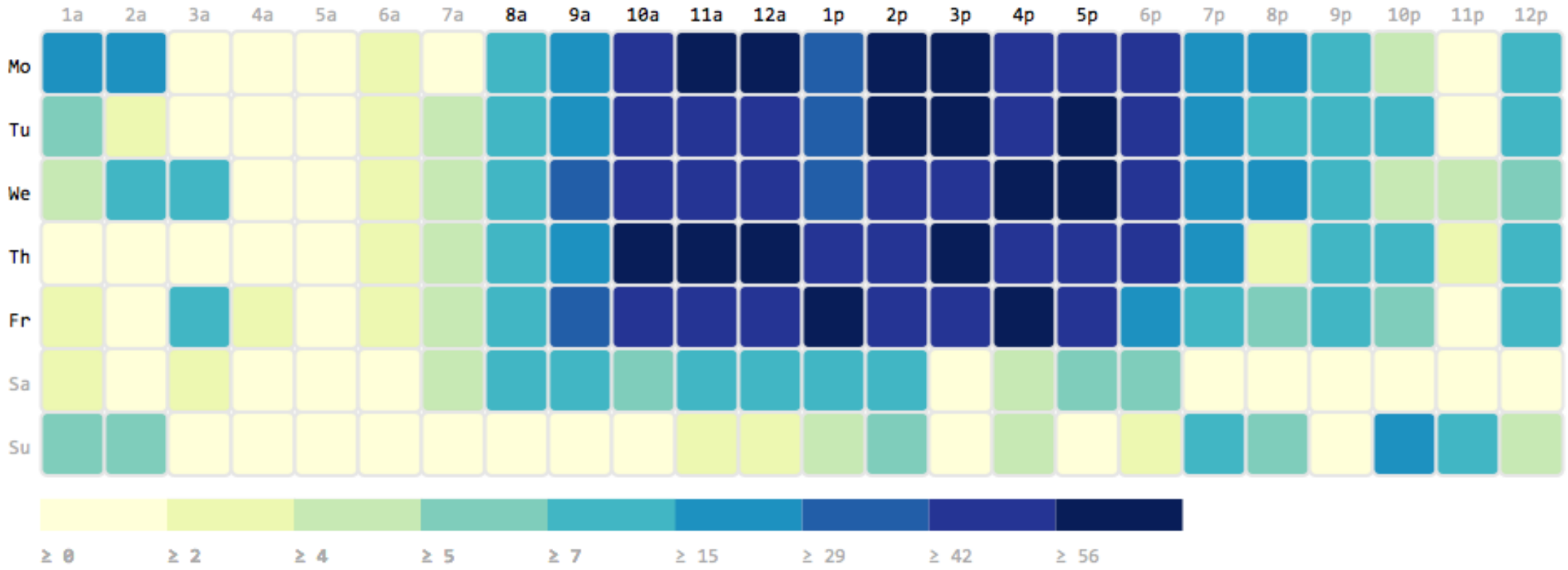
- Use lighter, less saturated colors for larger areas
- Use darker, more saturated colors for smaller areas and lines
- For ordinal data, use darker and more saturated for higher values (e.g. heat map)
- Color (hue) length is ~6-10, so with more categories, color may not be an appropriate visual variable.
- Saturation has a length of ~3
- Ensure a luminance contrast of 3:1 for text.

Resource:
Color Brewer

Reference: Slides provided by Chris Culy (http://ling.uni-konstanz.de/pages/home/butt/main/material/esslli14-vis/CuC_slides/reveal-based/theory.html#/title)

Heatmap Example

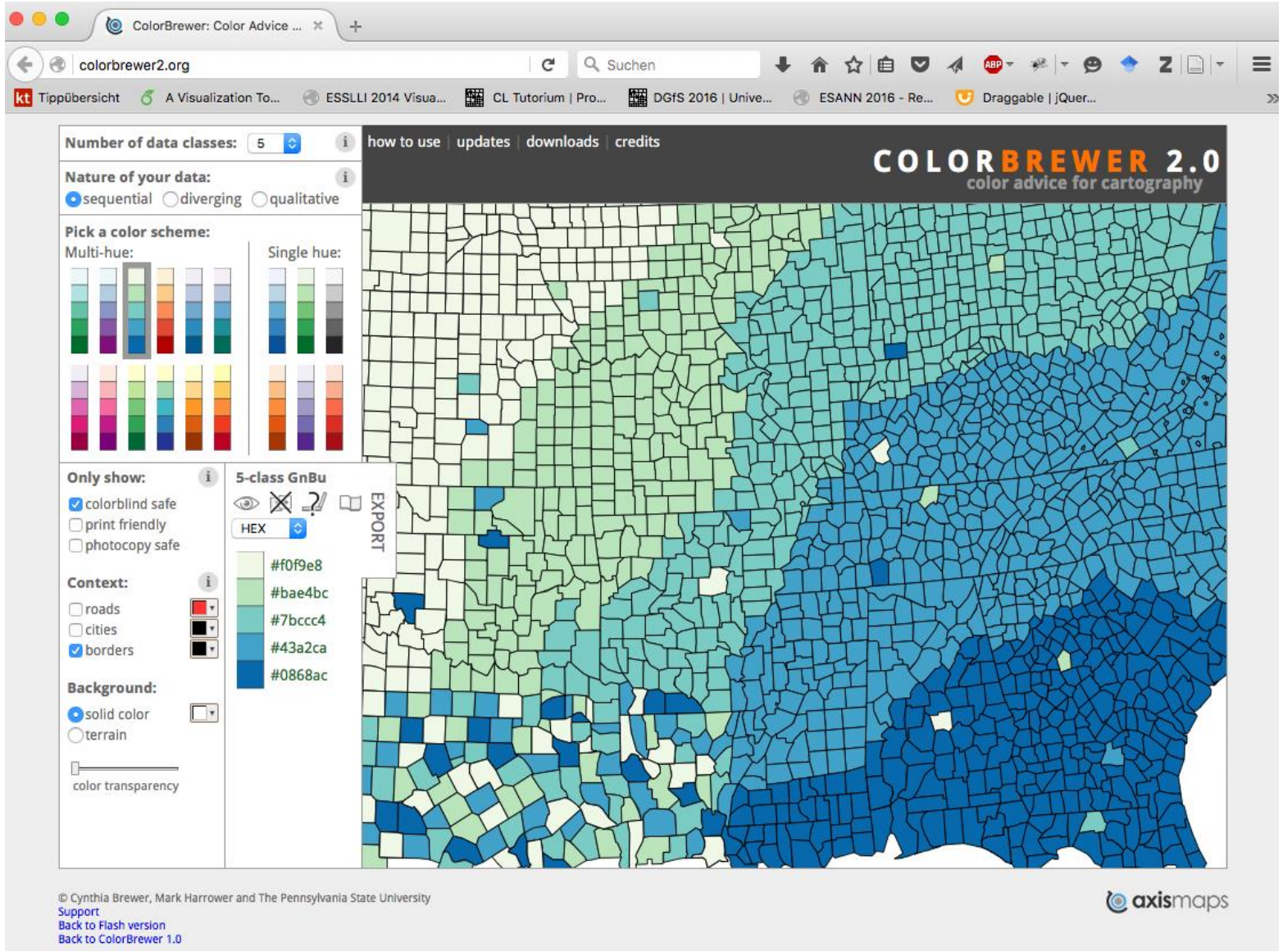
Example data shows concurrent user sessions over time, taken from a development environment.



http://bl.ocks.org/tjdecke/5558084

ColorBrewer

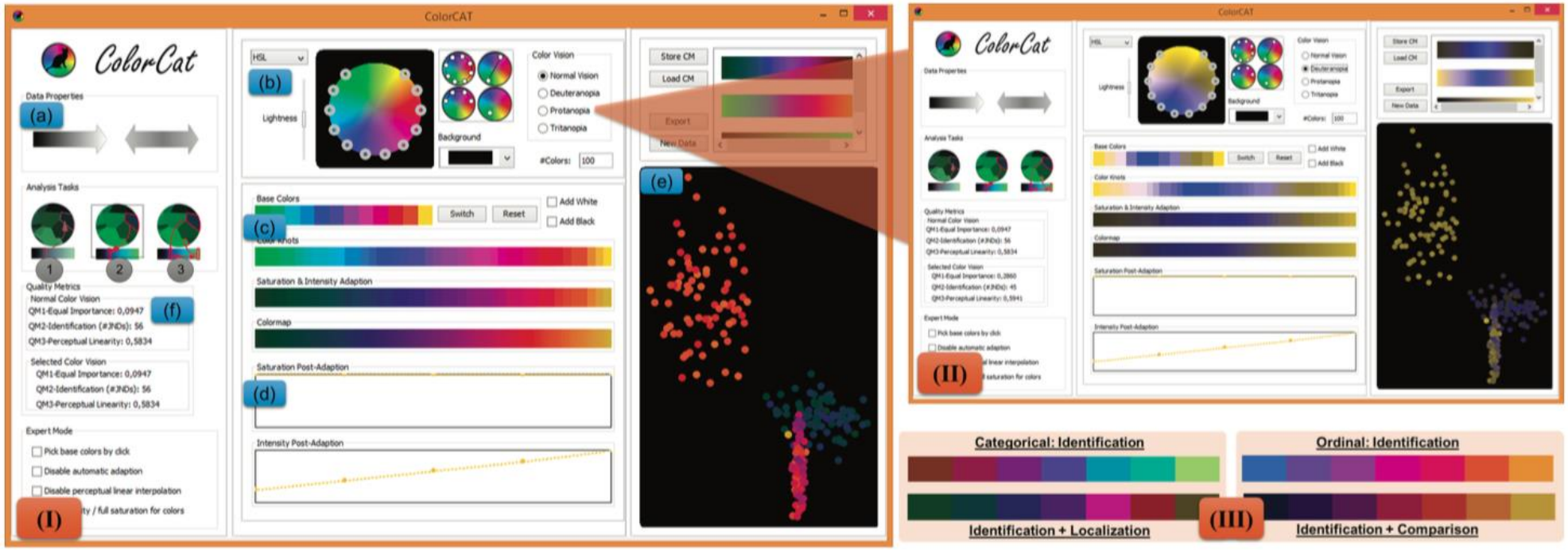
Color Brewer



http://colorbrewer2.org/

More Advanced - ColorCat

Colormaps for Combined **A**nalysis Tasks.



https://github.com/SebastianMittelstaedt/ColorCAT

S. Mittelstädt, D. Jäckle, F. Stoffel and D. A. Keim. ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks. Eurographics Conference on Visualization (EuroVis) - Short Papers, The Eurographics Association, DOI: 10.2312/eurovisshort.20151135, 2015.

Designing Visualizations

- Visualizations help:
  - To describe some **structure**, **patterns** or **anomaly** in the data.
  - To **explore** and **analyze** large datasets.
  - To make effective use of the **information overflow**.
  - To **communicate** information to people.

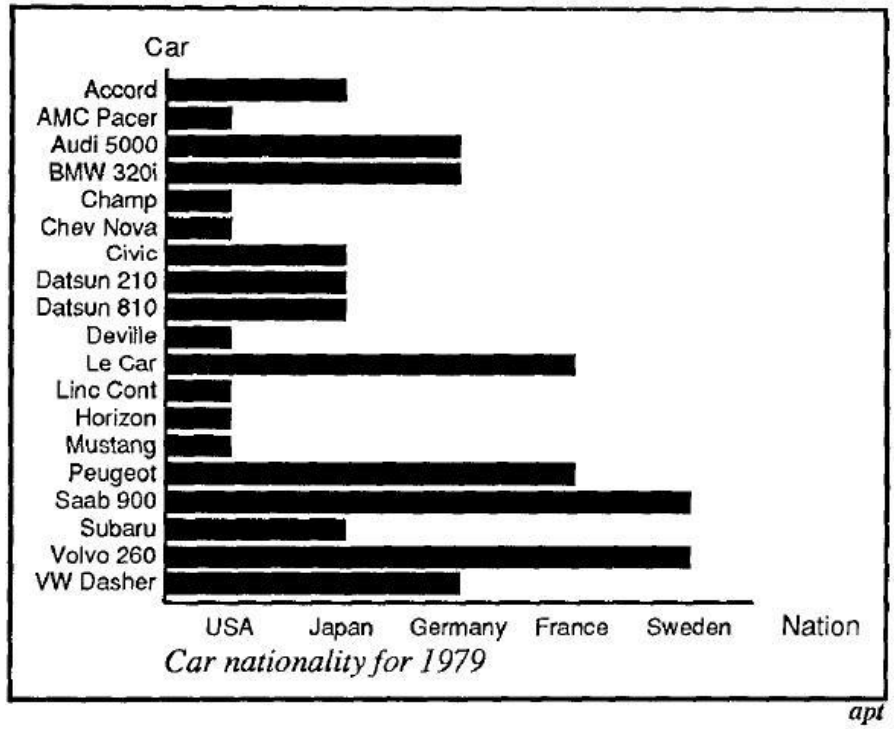
However, visualization can distort the "truth"!

Designing Visualizations

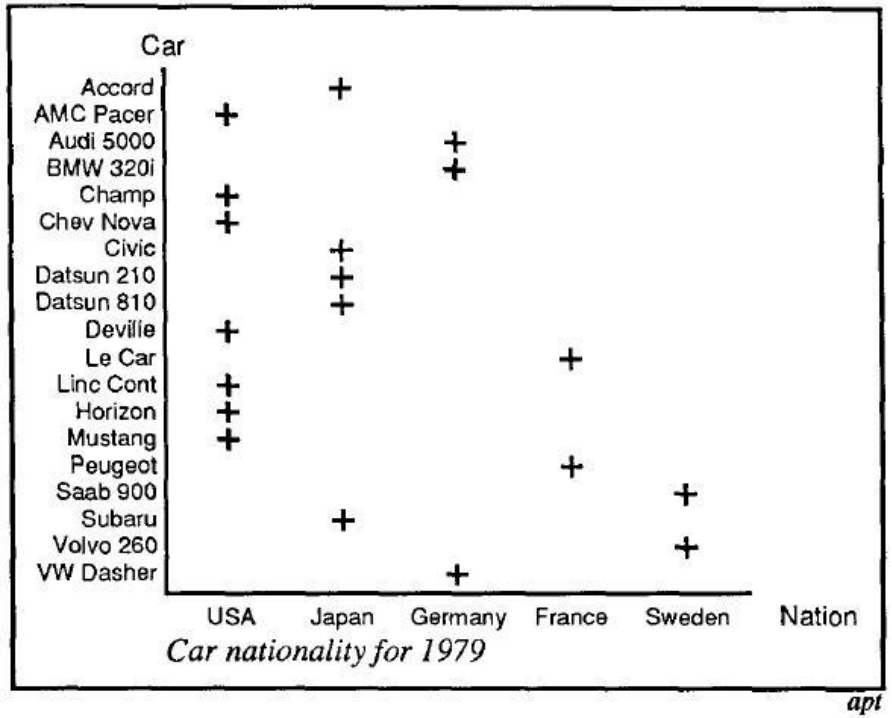
- **Expressiveness:** Visualization presents all the information and <u>only</u> the information.

- **Effectiveness:** Visualization is effective when it can be interpreted <u>accurately</u> and quickly and when it can be rendered in a cost-effective manner.

How to Visualize Badly



Poor use of a bar chart.

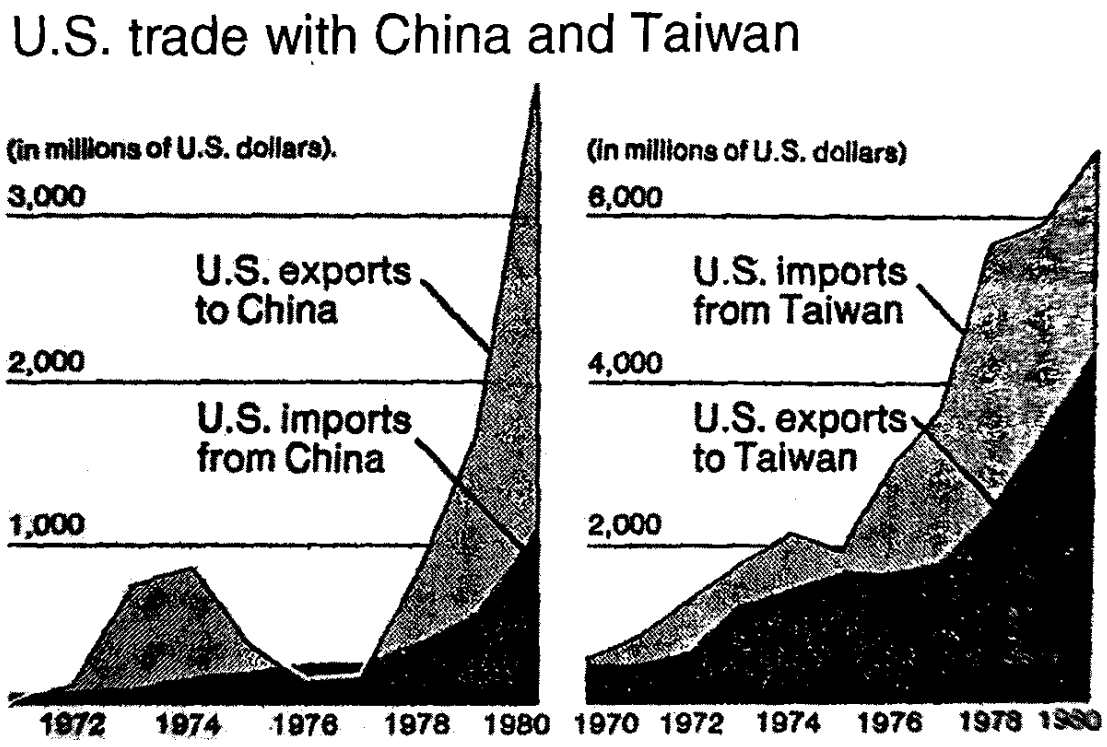


Better use of a scatterplot.

http://www.idvbook.com

How to Visualize Badly – What's Wrong?

(a)

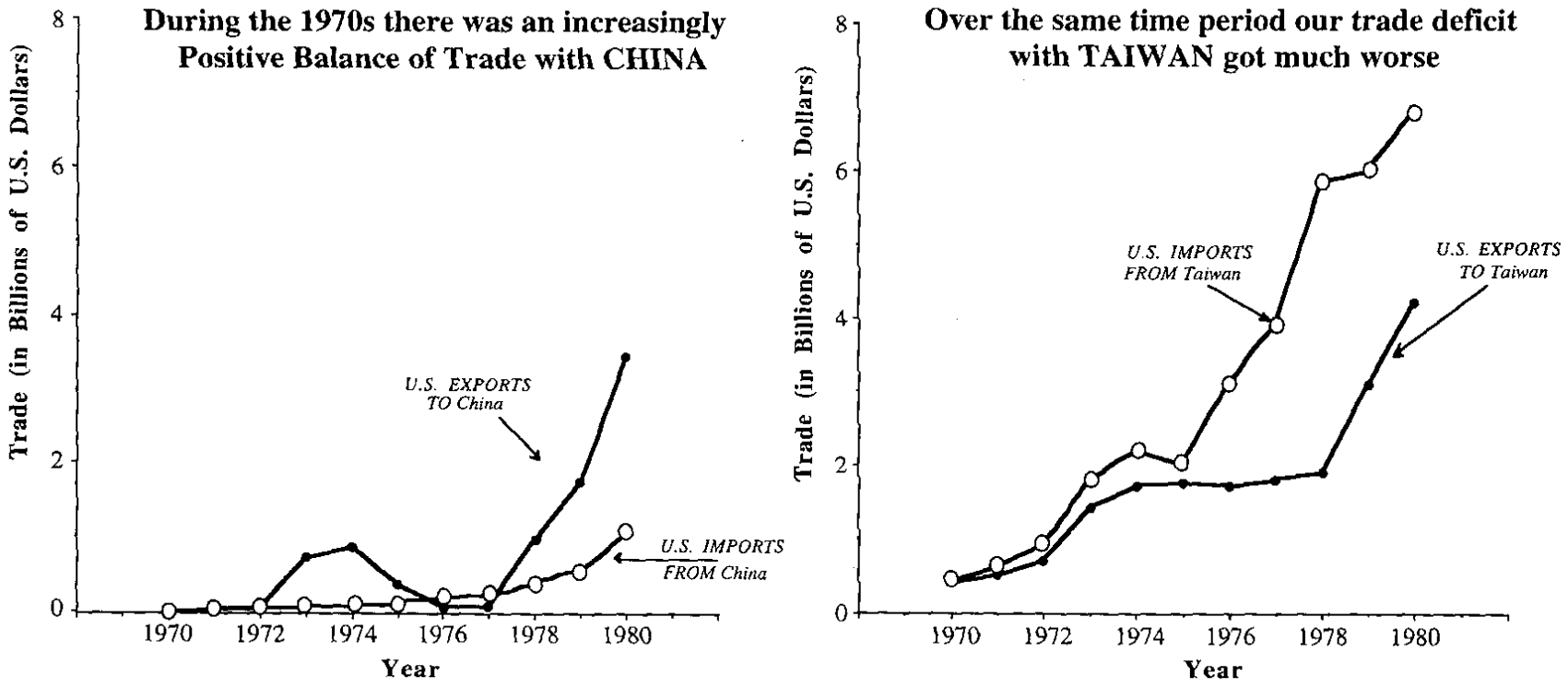


a) New York Times, Jun 14, 1981

Reversing the metaphor in mid-graphics while changing scales on both axis

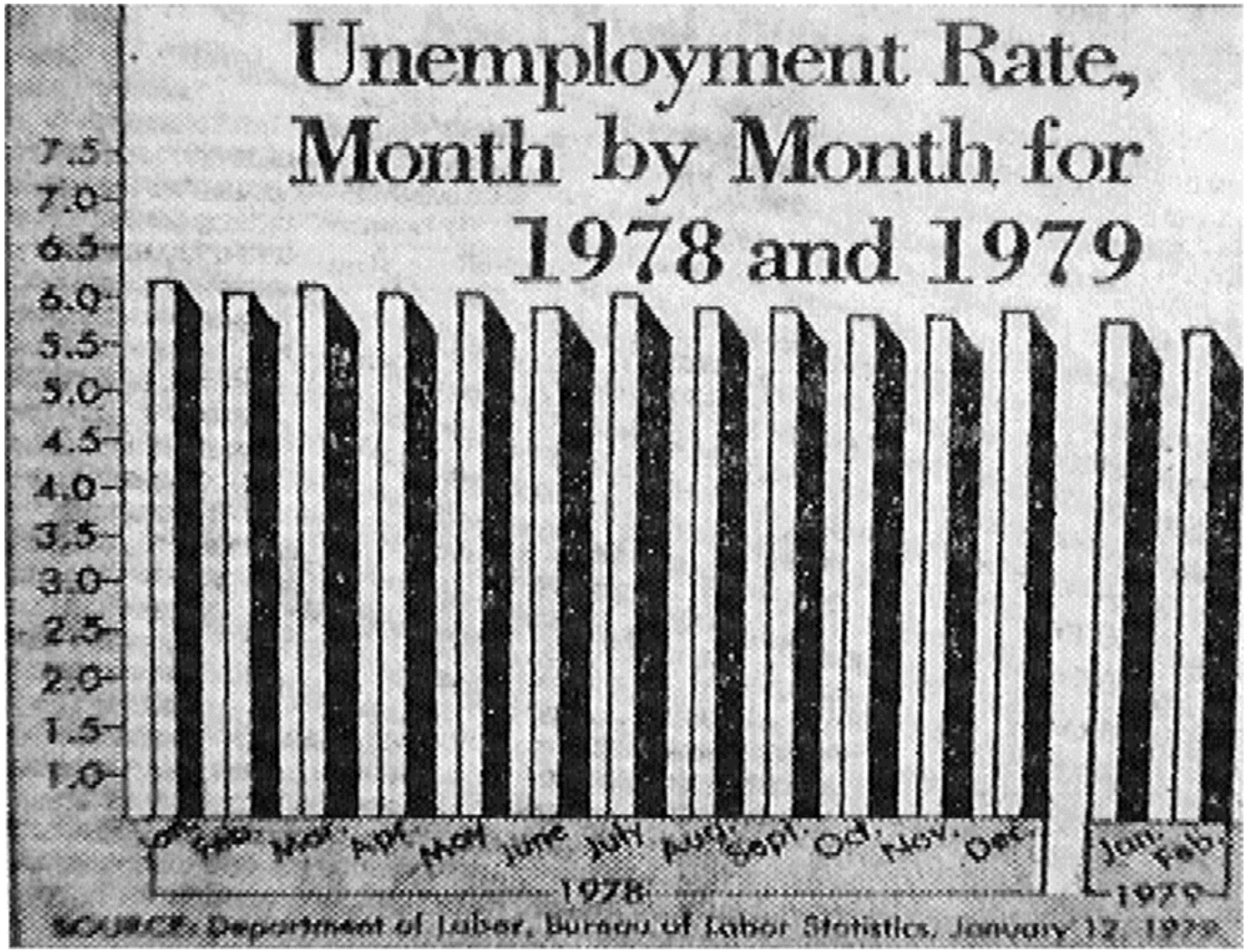
How to Visualize Badly – Improvements

(b)



(b) Redone with a consistent scale and visual metaphor

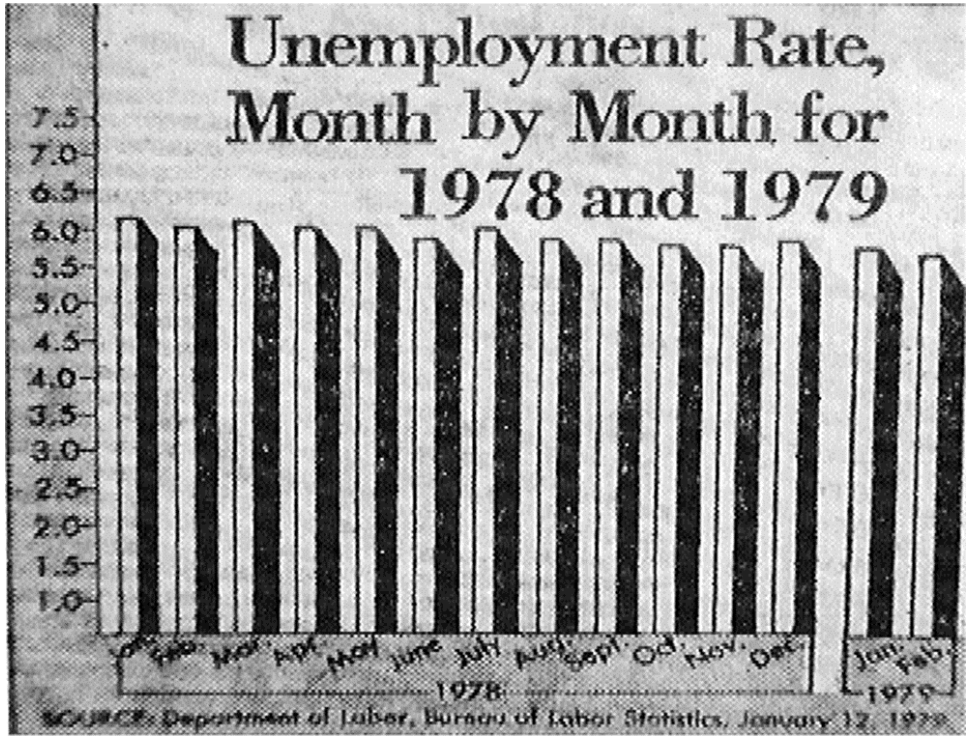
How to Visualize Badly – What's Wrong?



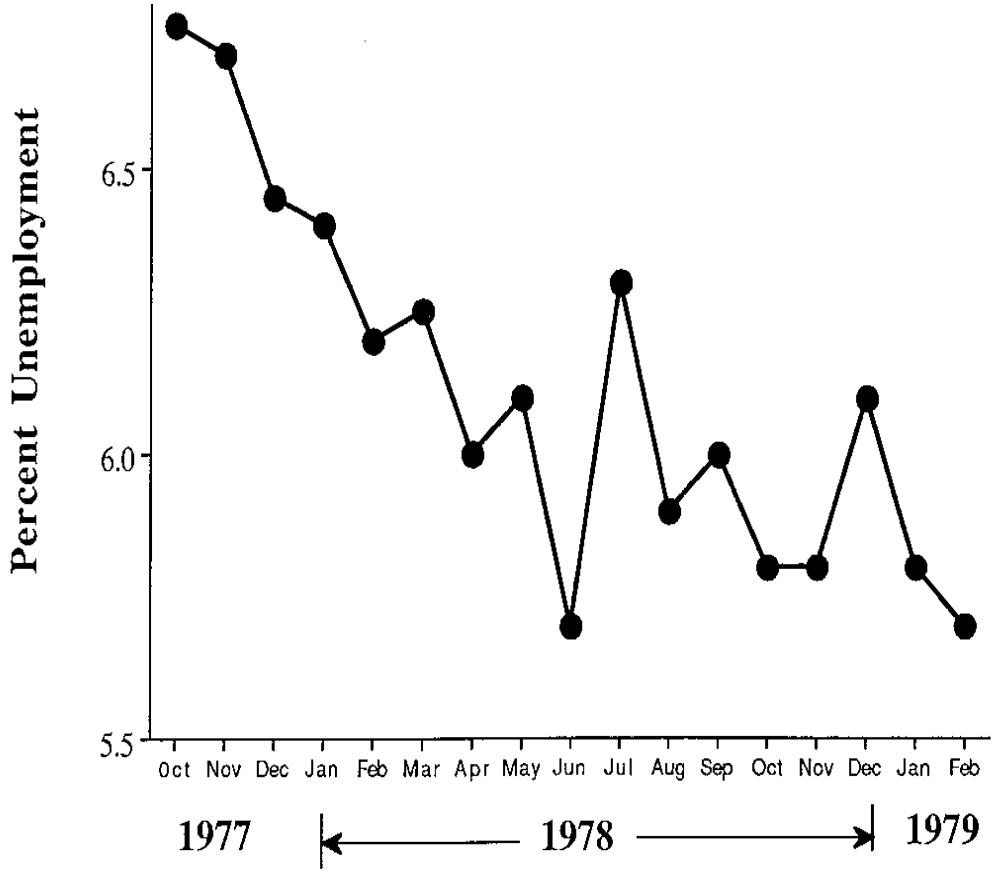
(a) Washington Post

How to Visualize Badly – Improvements

Hiding the effect by the careful choice of scale and origin

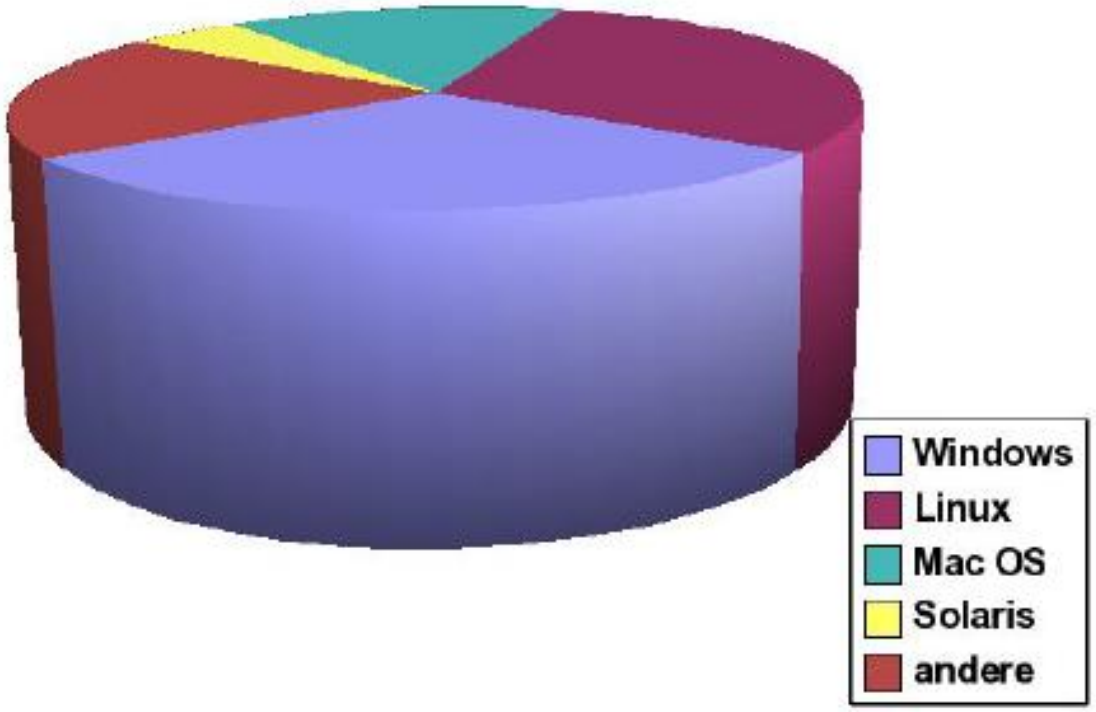


(a) Washington Post

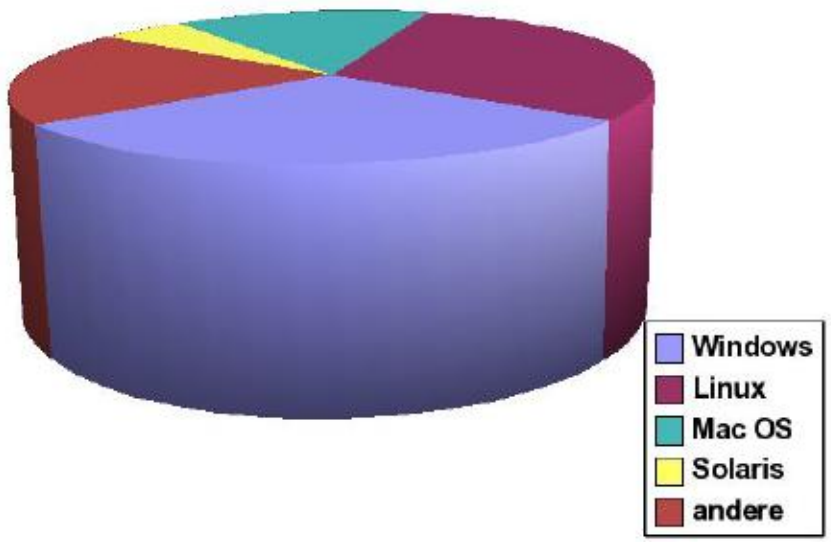


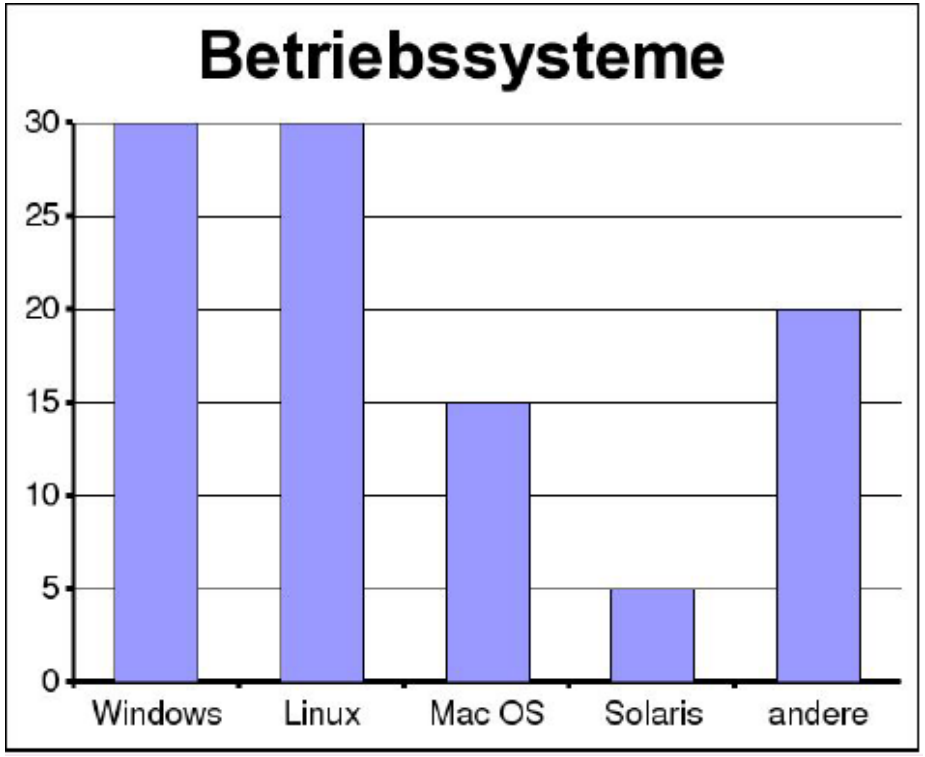
(b) Redone with expanded scale

How to Visualize Badly – What's Wrong?



How to Visualize Badly – Improvements





3-dimensional representation can make it difficult to directly compare sizes of objects

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

Resources

Interactive Analysis Process with Visualizations

- **Shneiderman's "Information Seeking Mantra"**

  "Overview first. Zoom and Filter. Details on Demand"

- **Keim's Visual Analytics Mantra**

  "Analyze first, show the important, zoom, filter and analyze further, details on demand"

References:
B. Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the 1996 IEEE Sympoium on Visual Languages, VL '96.
D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. 2006. Challenges in visual data analysis. Proceedings of the conference on Information Visualization, IV '06.

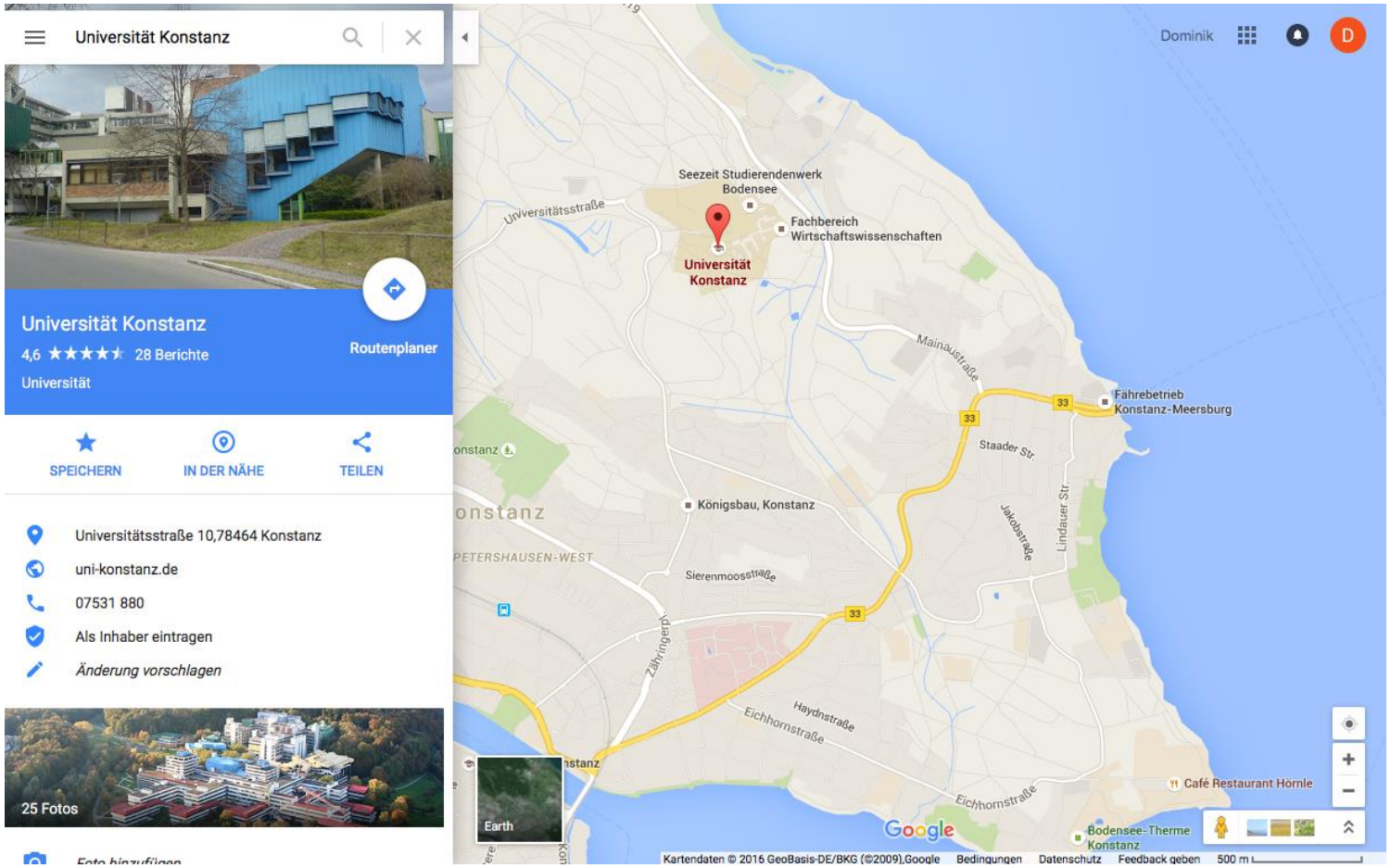
Interaction Types

- Select: mark something as interesting
- Explore: show me something else
- Reconfigure: show me a different arrangement [same type]
- Encode: show me a different representation [different type]
- Abstract/Elaborate: show me less or more detail
- Filter: show me something conditionally
- Connect: Show me related items
- Some other things: history, annotate, extract

References:
J.S. Yi, Y.A. Kange, J.T. Sasko, J.A. Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics. 13:6
B. Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the 1996 IEEE Sympoium on Visual Languages, VL '96.

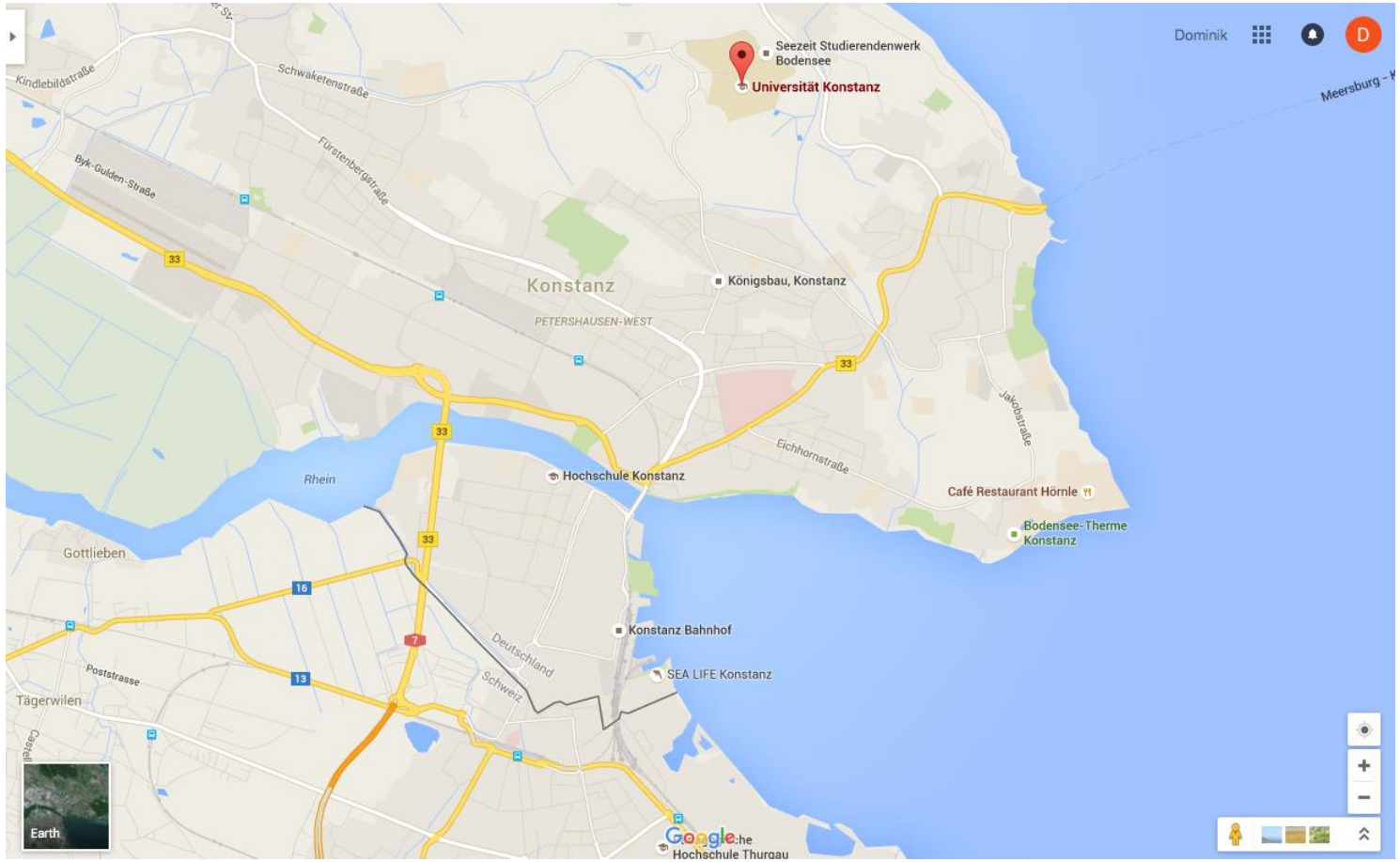
Interaction – Select

Example



https://www.google.de/maps/

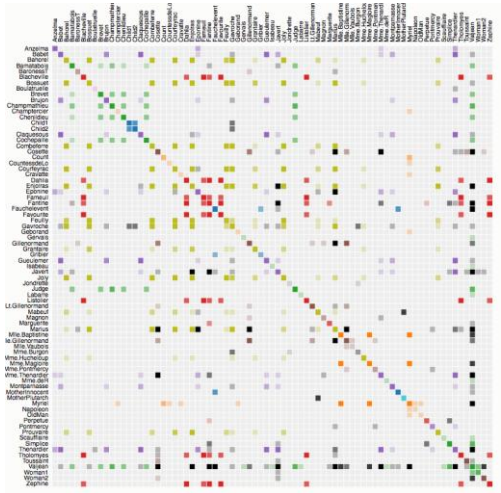
Interaction - Explore Example

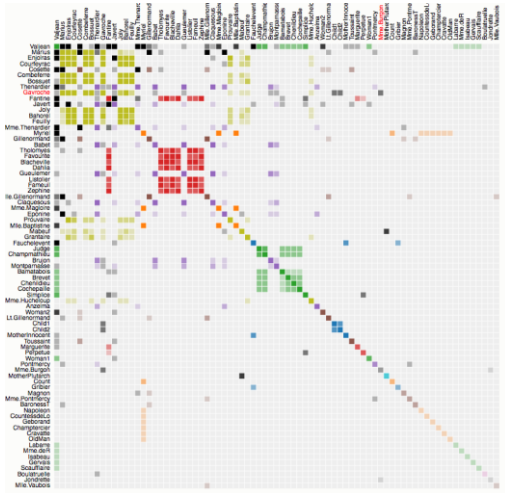


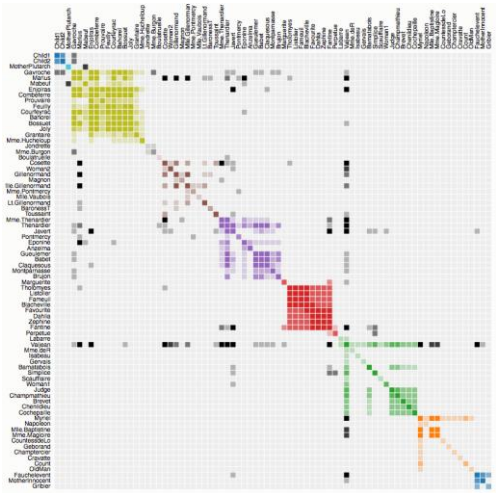
https://www.google.de/maps/

Interaction - Reconfigure

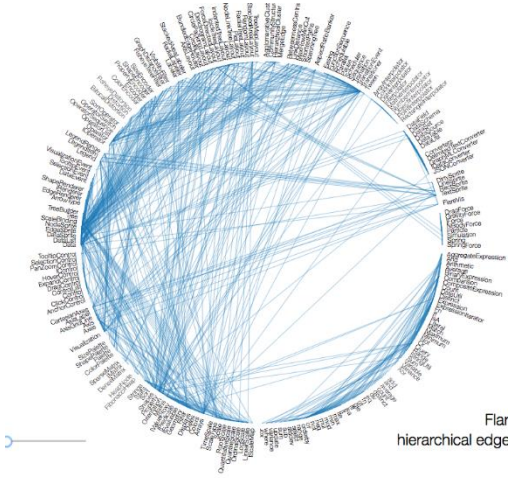
Example Example

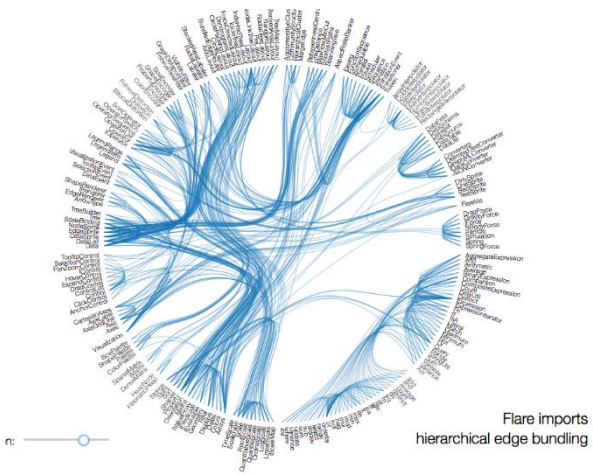






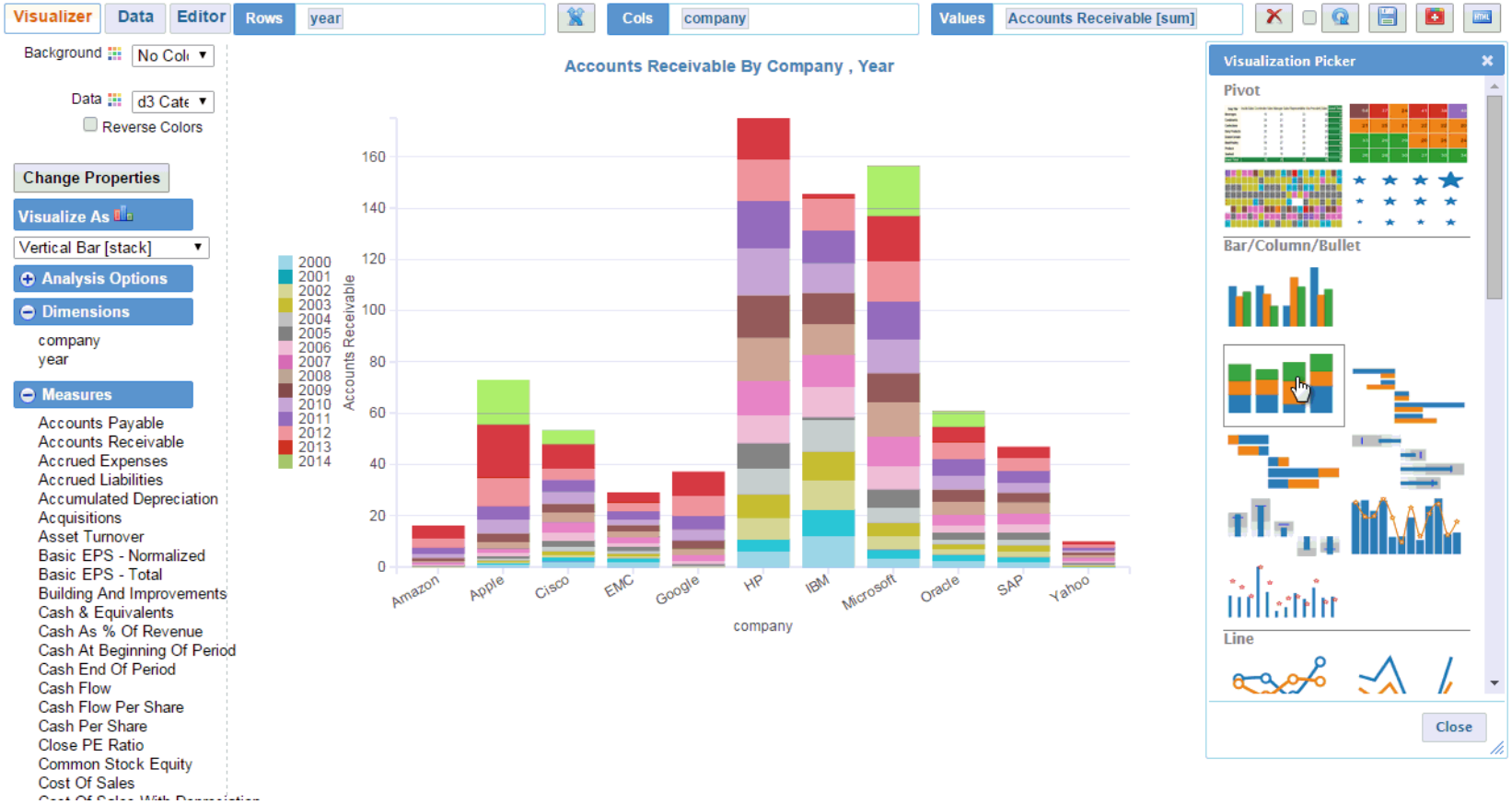
https://bost.ocks.org/mike/miserables/





http://mbostock.github.io/d3/talk/20111116/bundle.html

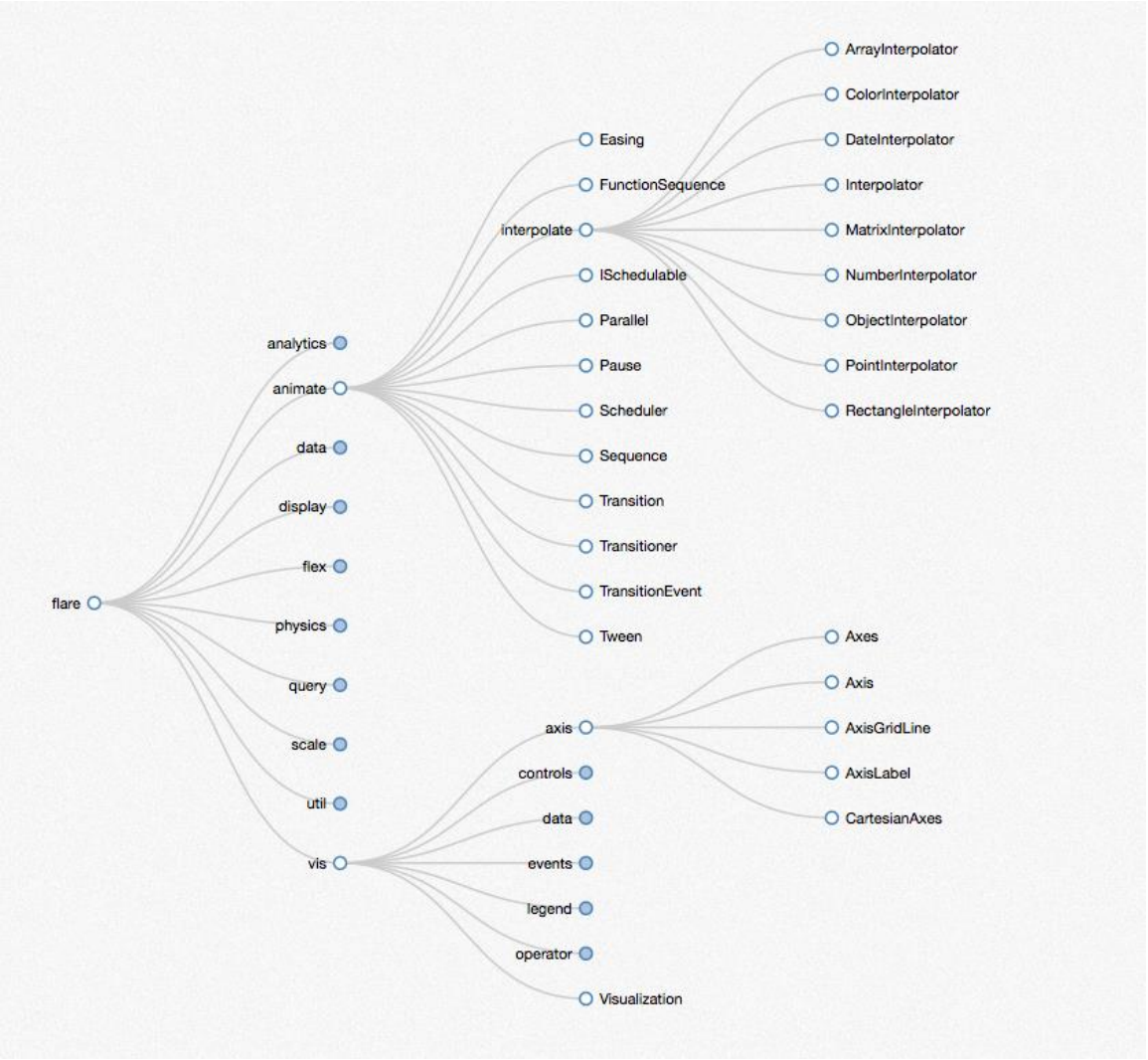
Interaction - Encode Example



http://www.infocaptor.com/

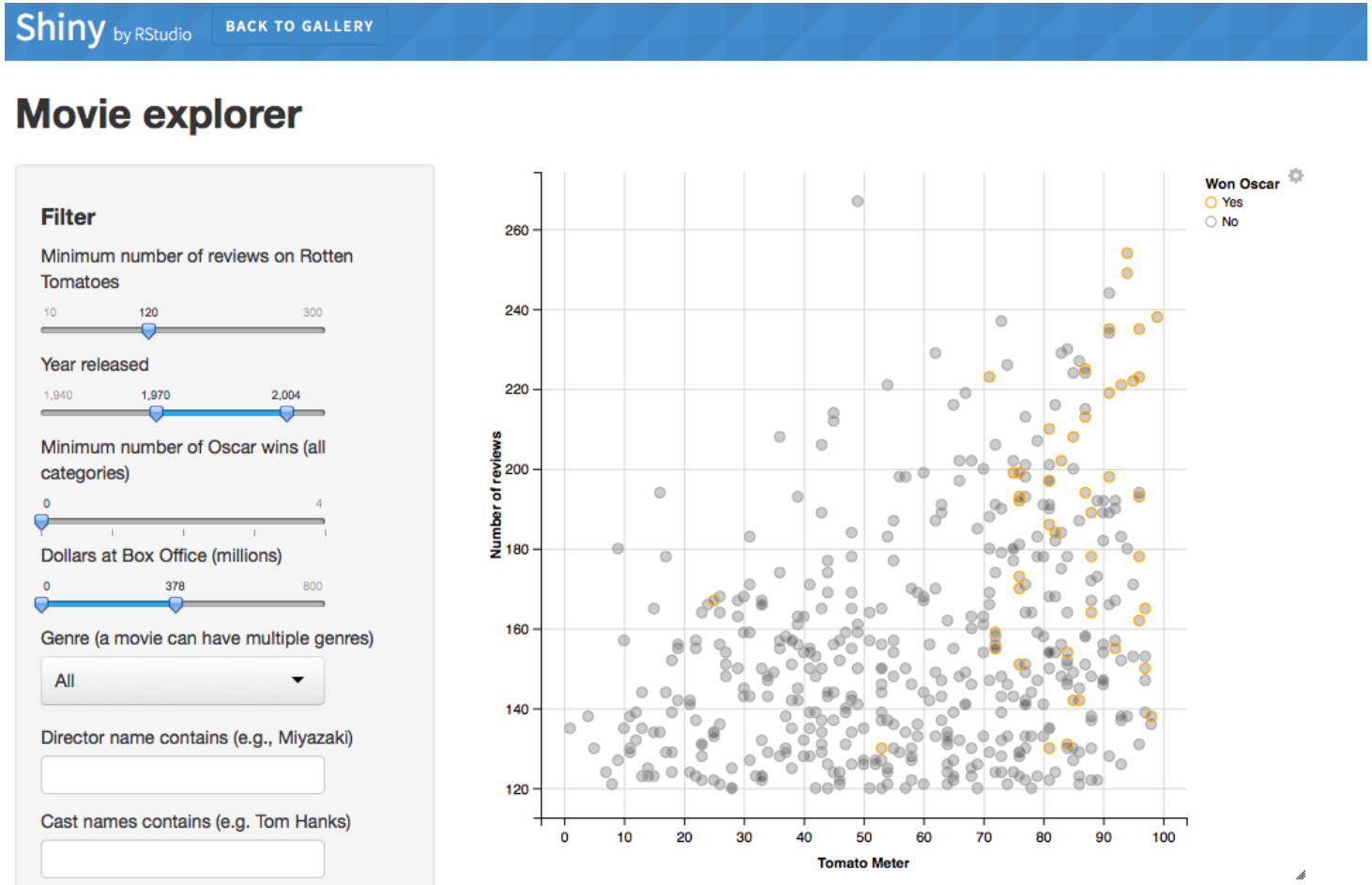
Interaction – Abstract/Elaborate

Example



http://mbostock.github.io/d3/talk/20111018/tree.html

Interaction – Filter Example

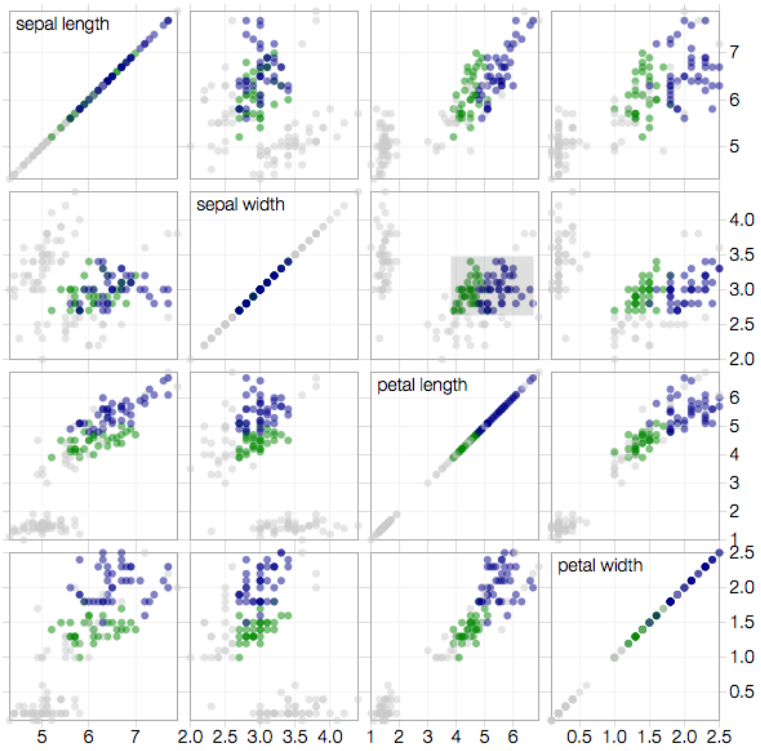


http://shiny.rstudio.com/gallery/movie-explorer.html

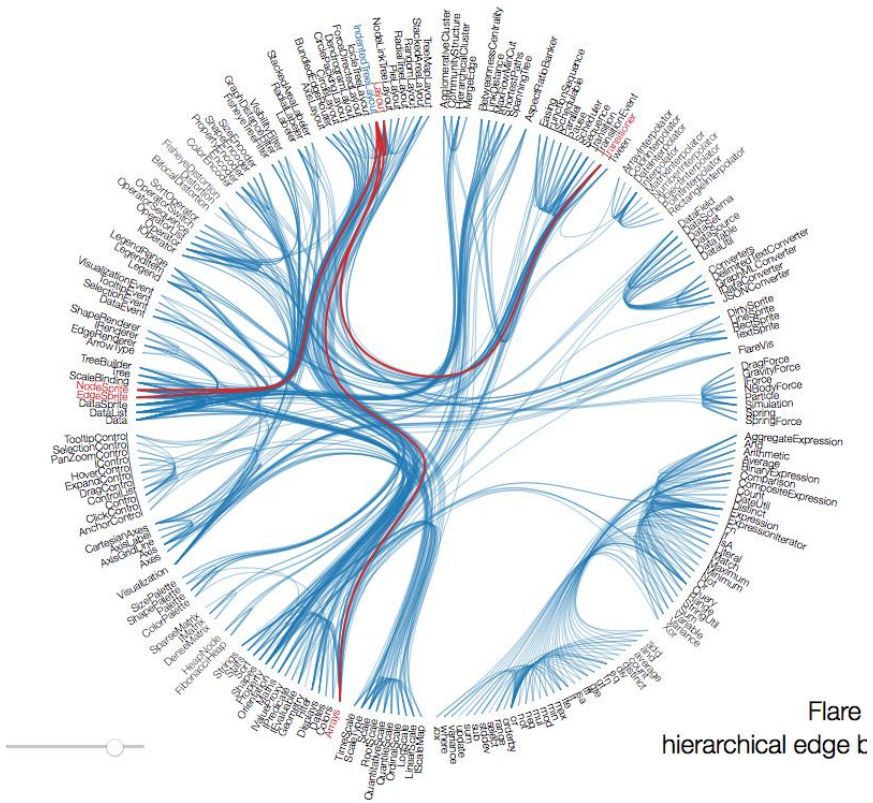
Interaction – Connect

Example

Example



http://mbostock.github.io/d3/talk/20111116/iris-splom.html



http://mbostock.github.io/d3/talk/20111116/bundle.html

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

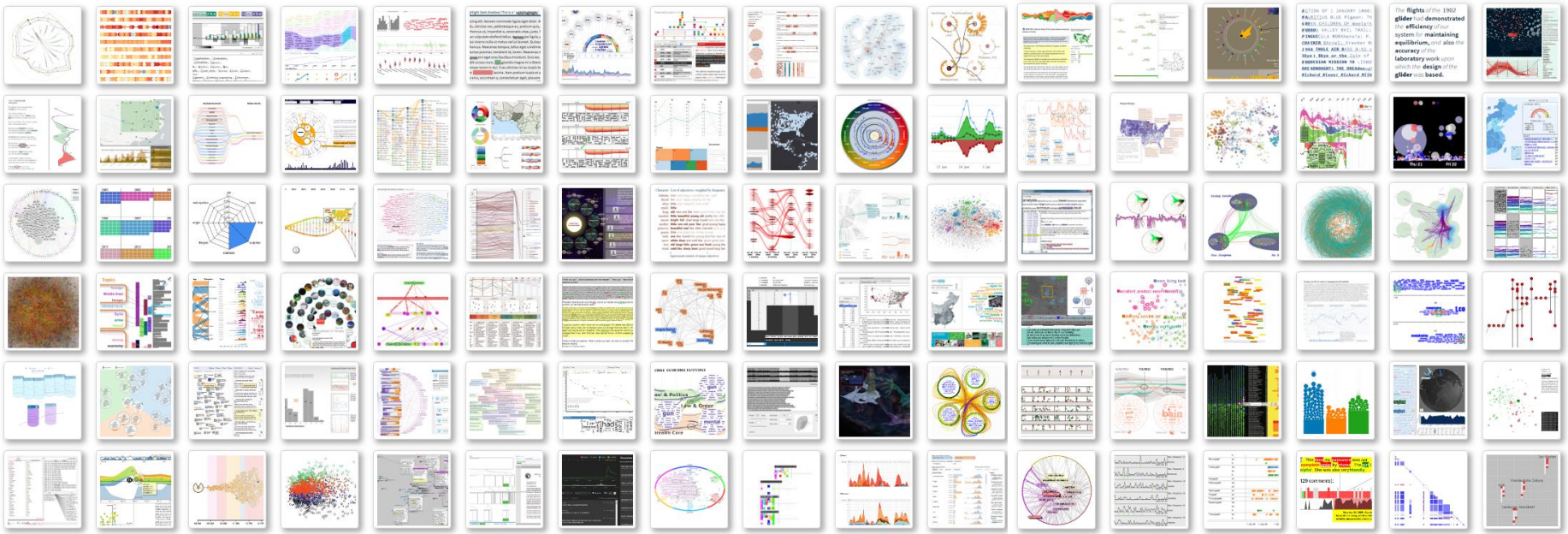
Designing Visualizations

Interactive Information Visualization

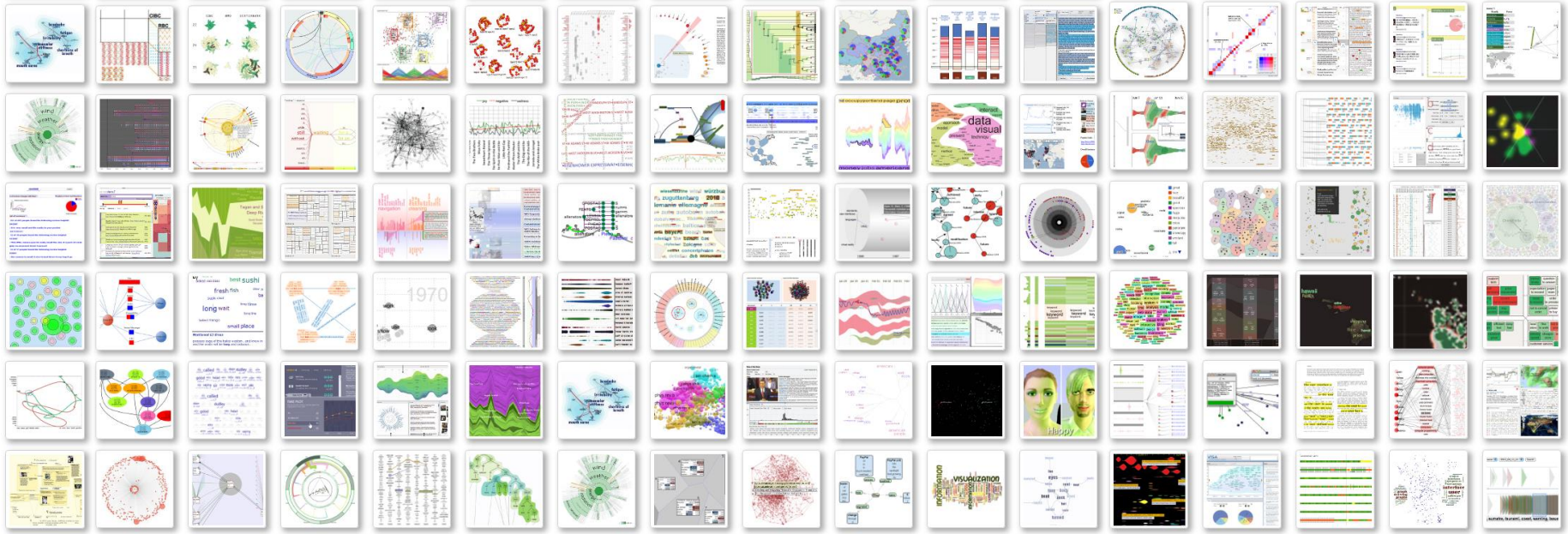
Linguistic Information Visualization

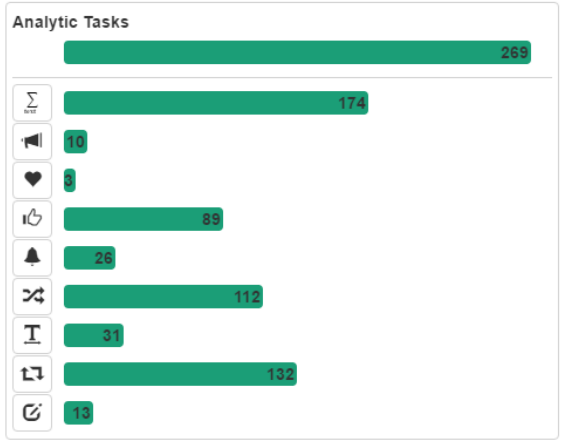
Open Research Challenges

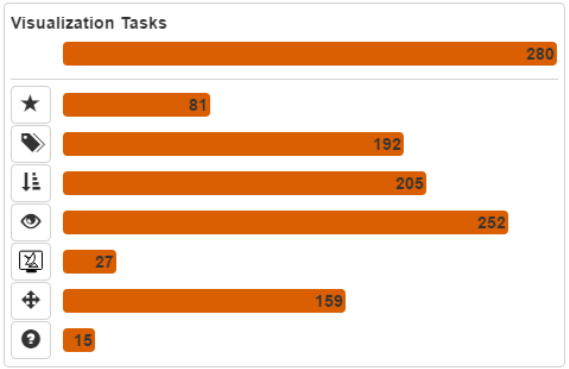
Resources

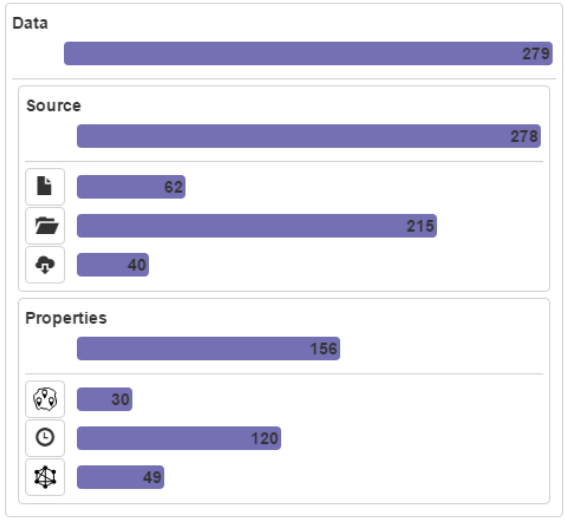


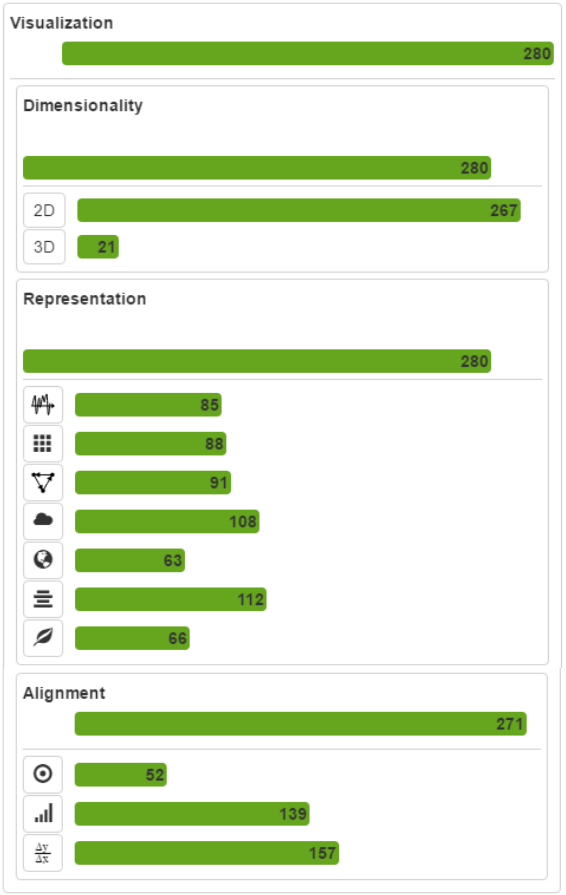
Linguistic Information Visualization

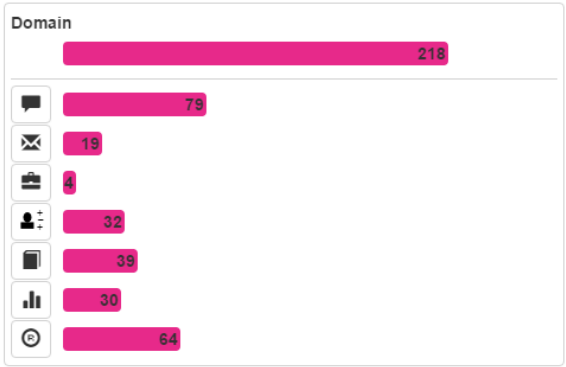












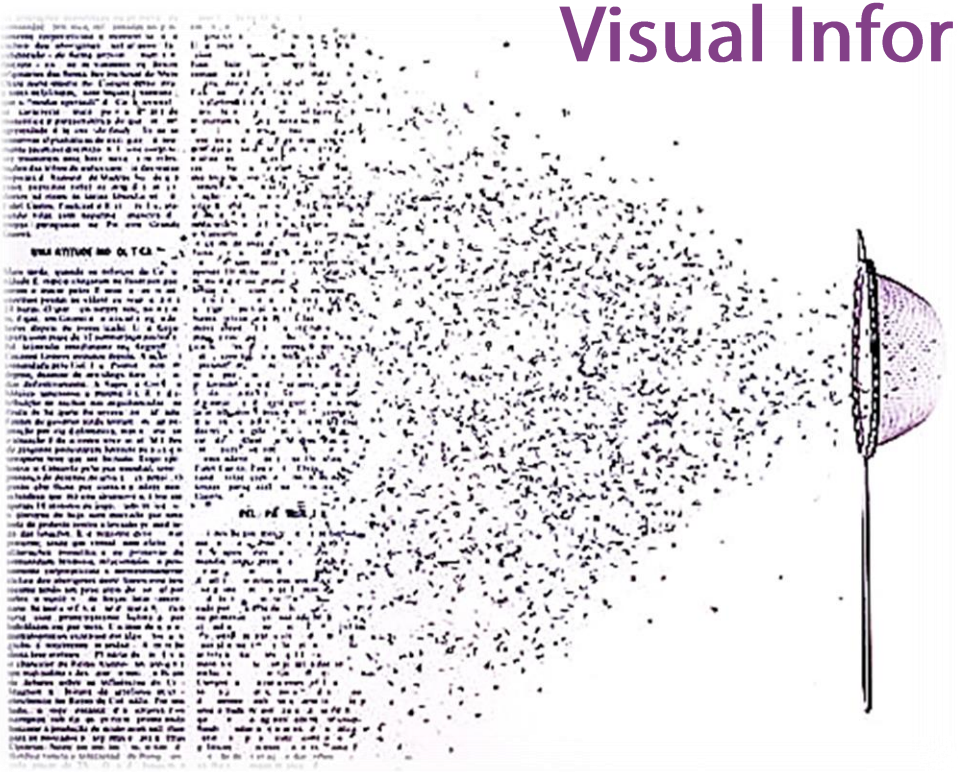
Text Visualization Techniques

http://textvis.lnu.se/

Text Visualization Browser

Developed by Kostiantyn Kucher and Andreas Kerren
ISOVIS group, Linnaeus University, Växjö, Sweden

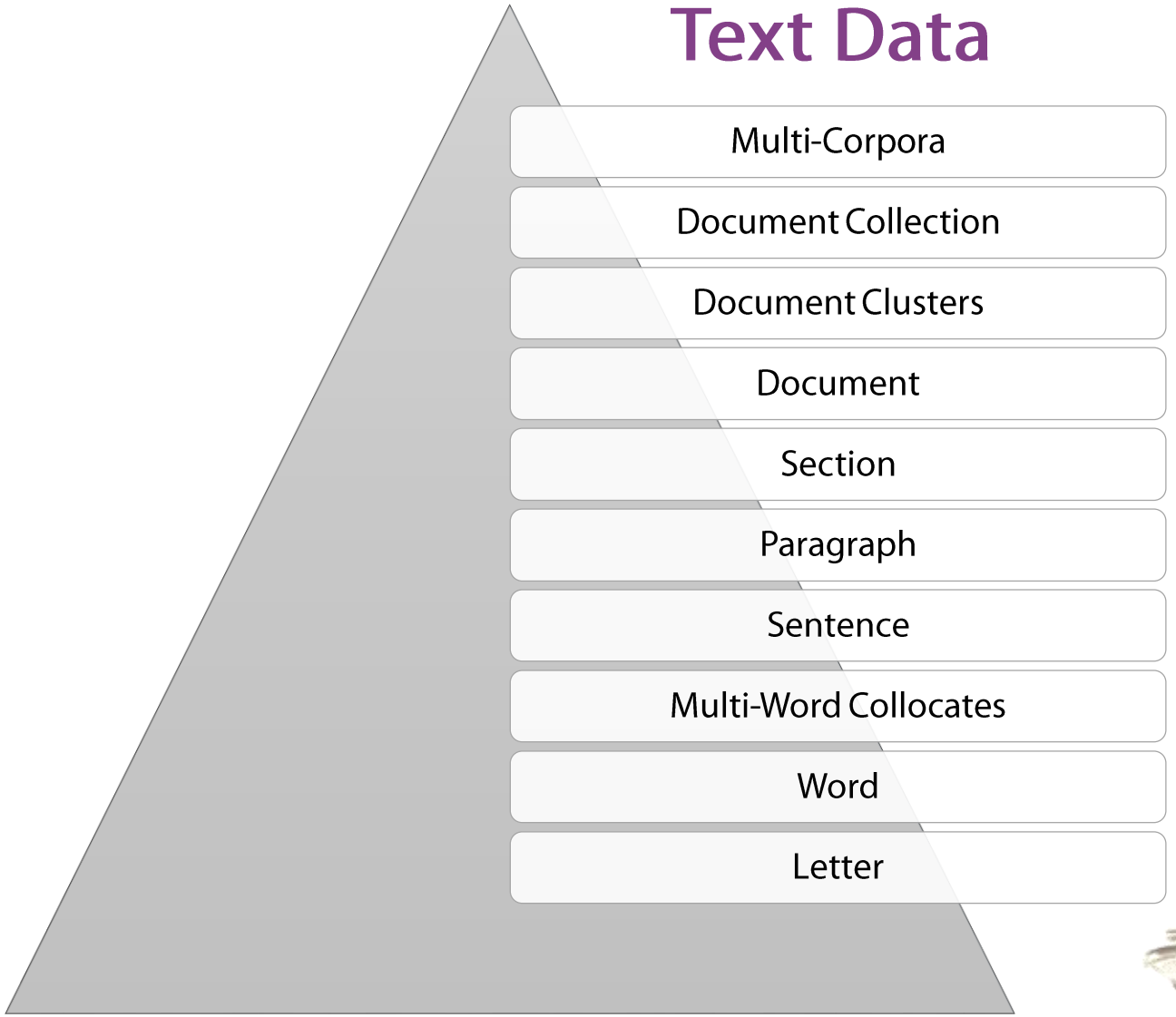
Visual Information Seeking Mantra



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

– Ben Schneiderman

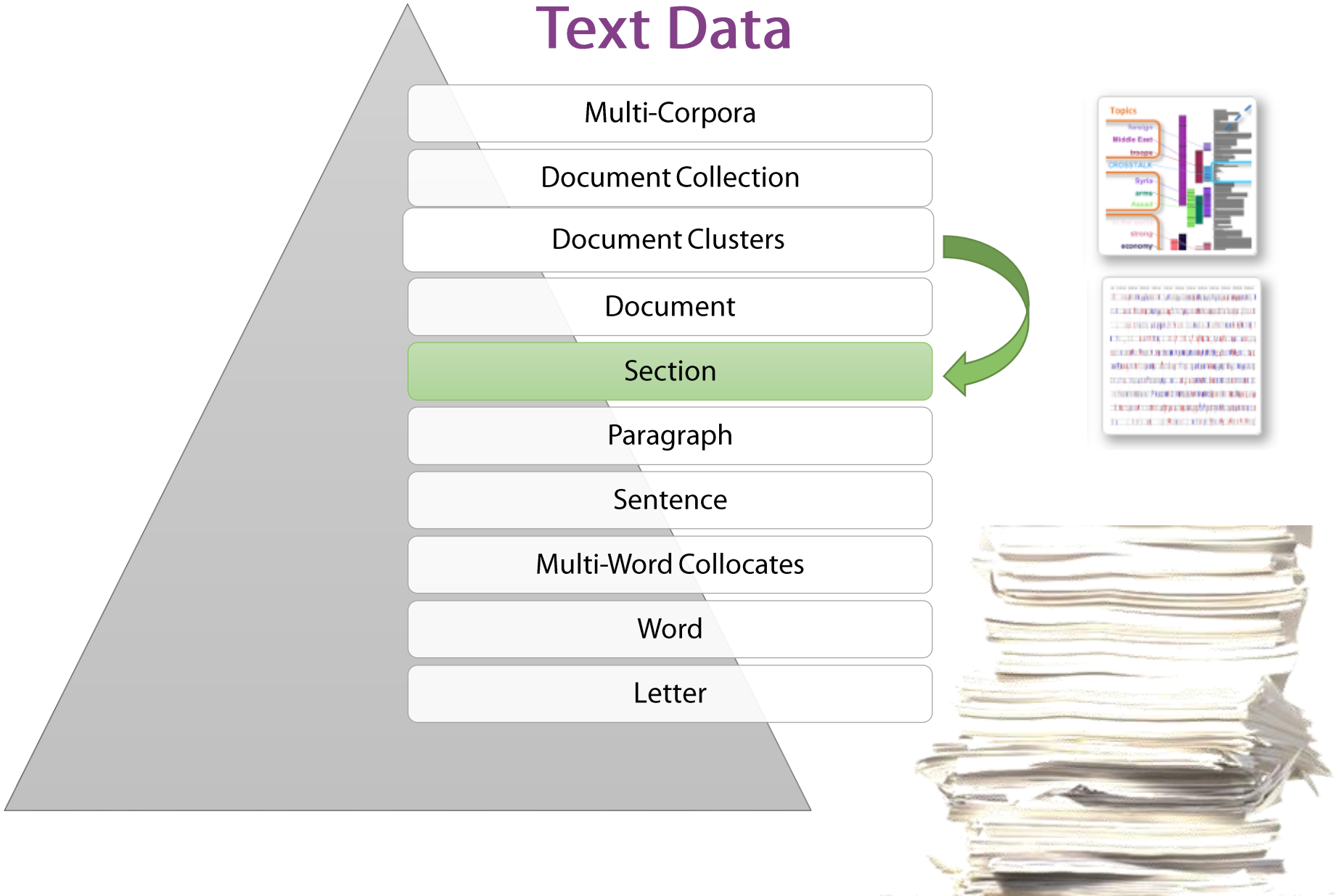
Text Data


Multi-Corpora
Document Collection
Document Clusters
Document
Section
Paragraph
Sentence
Multi-Word Collocates
Word
Letter

Text Data

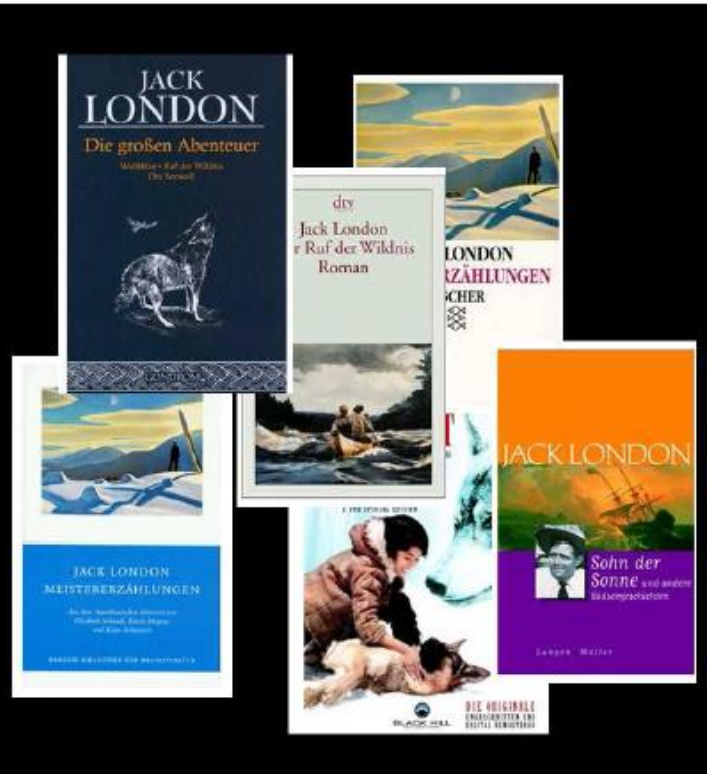
- Multi-Corpora
- Document Collection
- Document Clusters
- **Document**
- Section
- Paragraph
- Sentence
- Multi-Word Collocates
- Word
- Letter

Text Data


Multi-Corpora
Document Collection
Document Clusters
Document
Section
Paragraph
Sentence
Multi-Word Collocates
Word
Letter
Topics
CROSSTALK

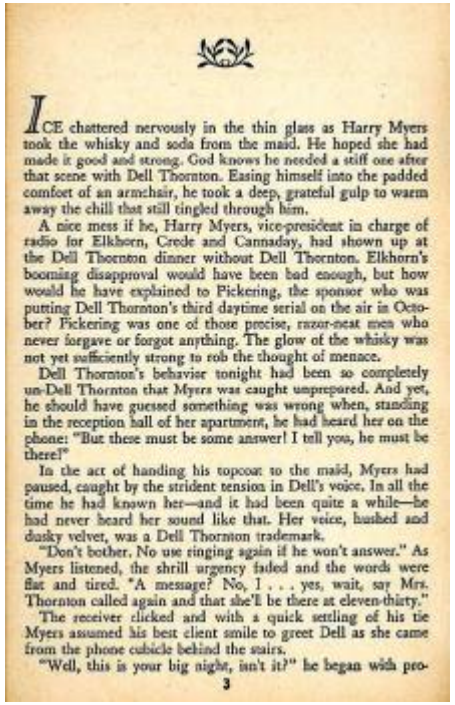
Autorship Attribution

Books of Mark Twain



Books of Jack London

Autorship Attribution



ICE chattered nervously in the thin glass as Harry Myers took the whisky and soda from the maid. He hoped she had made it good and strong. God knows he needed a stiff one after that scene with Dell Thornton. Easing himself into the padded comfort of an armchair, he took a deep, grateful gulp to warm away the chill that still tingled through him.

A nice mess if he, Harry Myers, vice-president in charge of radio for Elkhorn, Crede and Cannaday, had shown up at the Dell Thornton dinner without Dell Thornton. Elkhorn's booming disapproval would have been bad enough, but how would he have explained to Pickering, the sponsor who was putting Dell Thornton's third daytime serial on the air in October? Pickering was one of those precise, razor-neat men who never forgave or forgot anything. The glow of the whisky was not yet sufficiently strong to rob the thought of menace.

Dell Thornton's behavior tonight had been so completely un-Dell Thornton that Myers was caught unprepared. And yet, he should have guessed something was wrong when, standing in the reception hall of her apartment, he had heard her on the phone: "But these must be some answer! I tell you, he must be there!"

In the act of handing his topcoat to the maid, Myers had paused, caught by the strident tension in Dell's voice. In all the time he had known her—and it had been quite a while—he had never heard her sound like that. Her voice, hushed and dusky velvet, was a Dell Thornton trademark.

"Don't bother. No use singing again if he won't answer." As Myers listened, the shrill urgency faded and the words were flat and tired. "A message? No, I . . . yes, wait, say Mrs. Thornton called again and that she'll be there at eleven-thirty."

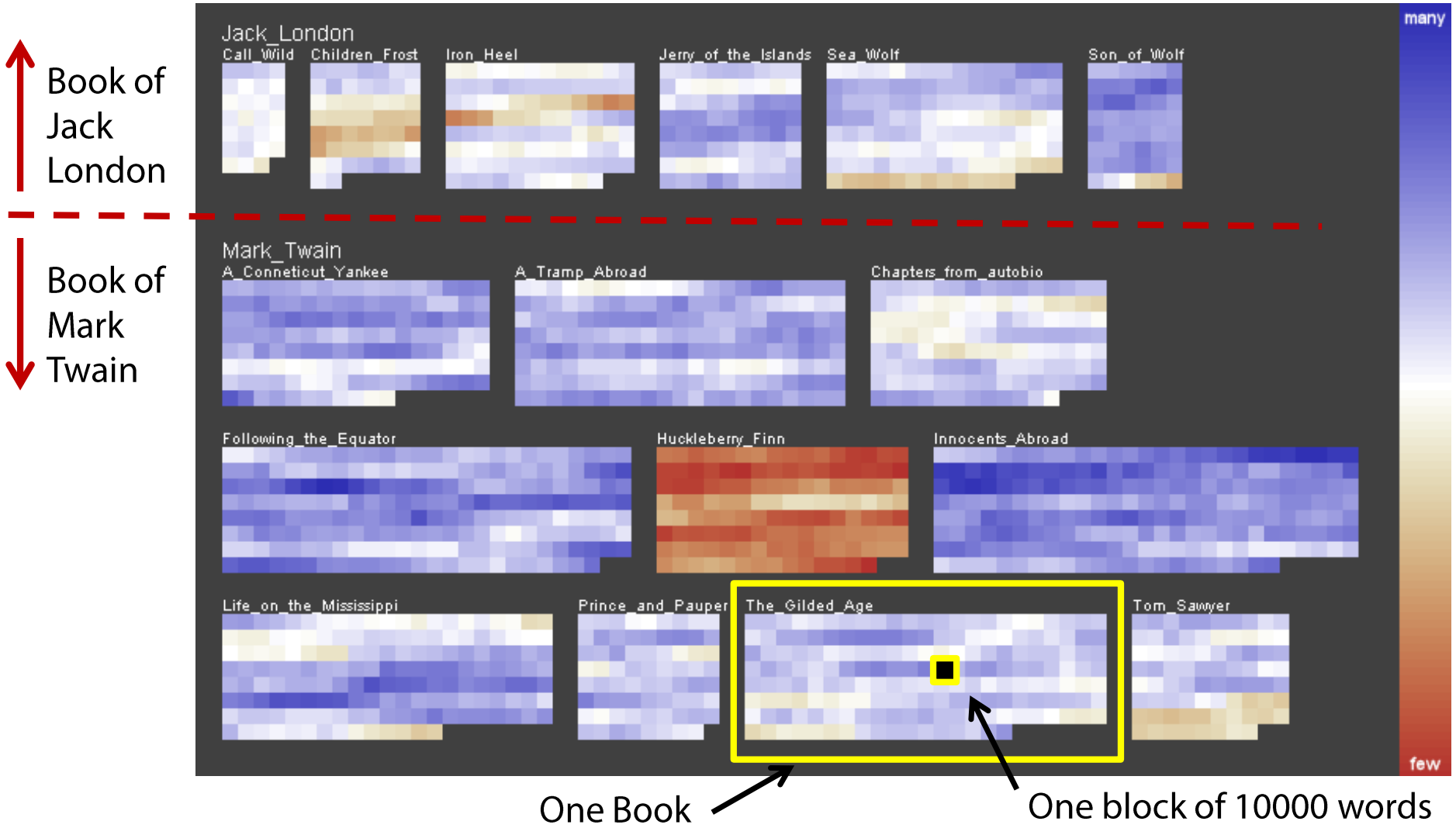
The receiver clicked and with a quick settling of his tie Myers assumed his best client smile to greet Dell as she came from the phone cubicle behind the stairs.

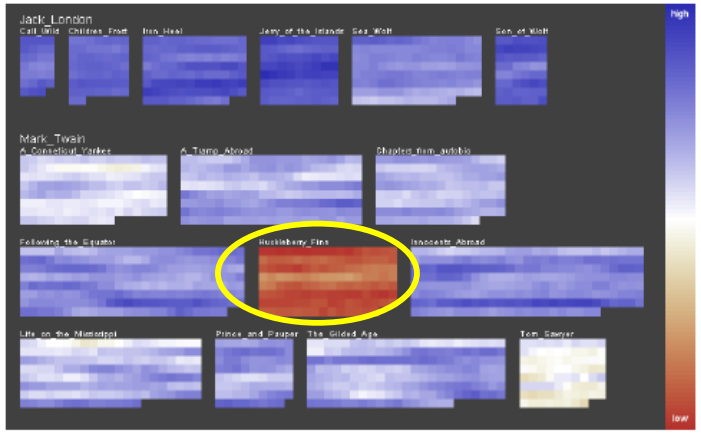
"Well, this is your big night, isn't it?" he began with pro-

3

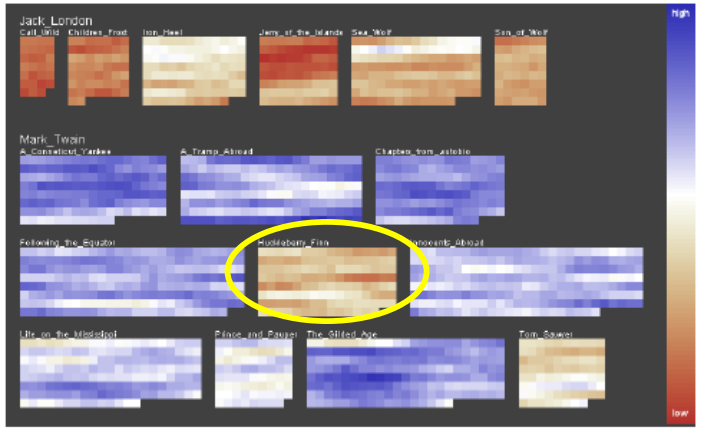
-4.81105499467848;-0.629981482830907;-2.82158710348232;0.864467863341222;-0.879489778398353;1.57891949482903;-1.09089131021939;-
0.259905421643884;0.0464150799579077;0.792075775807013;1.72929256021144;0.680213056027416;0.382737441718323;-0.374769983057510;-1.67480276712309;-
1.22852218796697;1.06096996519381;0.427147689189461;-0.124558036179387;-1.33485147434241;0.967821876528 5;-0.160298552062011;1.30576710708025;-
1.41012444824691;1.95746012752952;-0.116265953142018;-0.266903961839062;0.345574424741929;-0.146582593509817;0.230657536389546;-
0.357790583763730;0.245376276580946;0.533848123723 32;-0.463507192245968;-0.085194751337469 2;-0.068071965635272 8;0.451427601235678;0.443237033395699;-
0.115991305971826;0.204596871785712;0.303070769499014;0.476179531484601;0.129709378303450;-0.312422228547057;-0.819586364607007;-0.130447034808557;-
1.01048215393487;0.854966121234136;0.163594826008225;0.060478544292906 8;-0.140809923336609;-0.131763805088605;-0.130118204746893;0.069784916591852 6
-4.72619587305622;-0.625946193650 87;-2.76912115538271;0.920662215882416;-0.843228493513209;1.57943224746163;-1.00414981835522;-
0.252525954229222;0.117680500878398;0.795454000957187;1.68415312019892;0.689991757449287;0.374192012410939;-0.317945402612463;-1.71795524036913;-
1.29019982547521;1.02659004294441;0.361517069344435;-0.151844564067564;-1.32023363743146;0.906370506749 15;-0.156571534663971;1.32149294809112;-
1.39698786288930;1.91410304896837;-0.157440323847322;-0.225639336002252;0.347281599628439;-0.123805740225228;0.183338418435825;-
0.387582525546098;0.232620393646911;0.462210311105 86;-0.382061442244298;-0.023957245978215 1;-0.099359562084463 6;0.445670492666394;0.451488506062 36;-
0.152585024855877;0.186203671850665;0.309302185117181;0.389549293682638;0.145967379910122;-0.296632063148021;-0.801863626235 06;-0.084062754702204 9;-
0.974176667206306;0.852149991891077;0.245379569046821;0.065984560433762 5;-0.138558967577697;-0.151051756175594;-0.093701239882418 7;0.084007745200758 8
-4.53803451556225;-0.627440300863745;-2.42974348499624;1.01268635613751;-1.17277566026348;1.59684107236816;-0.563156592136839;-0.121589812251103;-
0.321926511652796;0.547872004952833;1.99981045735607;0.567578280369841;0.234485057162931;-0.377713568107434;-1.61762347490241;-
1.29675056407773;0.660303934051429;0.235185733552525;-0.062323354406978;-1.16817553512491;0.659538411599586;0.182411448205030;1.69948232981438;-
1.33908644901255;2.04369715249477;-0.215088311746 34;-0.196063840742395;0.414189493219527;-0.088116447503862 8;0.347748195017220;-
0.543439199428085;0.153819022839213;0.746263707976347;-0.095729806905964 1;-0.259720188716113;-
0.344597356765336;0.689996775468 22;0.490213703016975;0.228076915166204;0.118930042369515;0.235717238997765;0.580790636395948;0.407620535441267;-
0.229591928082285;-1.09001595522736;-0.116614999076637;-0.922886095206421;0.891412426417 06;0.045551052229331 1;0.165924004910293;-0.400208121177327;-
0.172533793857722;-0.129049102110246;0.161723777665322
-3.89479063986018;-0.461231913229098;-2.04875395916492;1.06045680548381;-1.44702164263929;1.78851354719597;-0.333457585484126;-0.528644168714319;-
0.562947243219 39;-0.109884062818805;2.13892002337232;0.605829859476453;-0.517082981397625;-0.660474980566 4;-1.64845993455231;-
1.42418546476728;0.321218418289759;-0.278224604723157;-0.378396421934162;-1.27631182773158;0.642705541642978;0.182762198751351;1.78592358263019;-
1.11299780016413;1.69358847547942;-0.545994901136617;-0.308620722517962;0.205128508596046;0.026907542674258 0;0.319086280618103;-0.460458510122498;-
0.365522039729986;0.735472368592393;0.181720001215043;-0.651703914723933;-0.062473265438567 3;0.789087125838315;0.061966135294490 9;0.036424995801957 8;-
0.269581900889604;0.274389603823092;0.707264417321 79;0.425887141090119;-0.037946162448910 4;-0.570221874024205;-0.269057089403 53;-
1.05181914500808;0.542422415780 93;0.199936495595751;0.154660903994125;-0.299738998058595;-0.459550779334414;-0.159966189650733;0.080947753677227
-3.87036054460171;-0.374884238982513;-1.67486772003841;1.02683065980590;-1.49836211884464;1.71211234764303;0.157829123266546;-0.633558846195596;-
0.270936441634708;-0.337577057025756;1.96676065620906;0.574338259108837;-0.813477900647415;-0.570034321071755;-1.45308434837561;-
1.81430447664568;0.275400448604515;-0.541296362525052;-0.486781463676837;-1.67533547259934;0.873954032421597;0.541660727076866;1.36203561225374;-
0.898936252627846;1.56135217187206;-0.594435448727 84;-0.198426384838783;0.160859388461496;0.166981604944151;0.368127156674806;-0.319849224597977;-
0.406483136709955;0.289423487531856;0.390148536735591;-0.482402223416843;0.320343592322035;0.993249878971298;-0.220042513838974;0.228577825581678;-
0.028393311849188 9;0.350283284140497;0.771245745420676;0.104501558480396;-0.115141102614263;-0.745824611409785;-0.326496069148749;-
1.25346963706088;0.335995945365706;0.320082682722975;0.464017075241565;-0.259408875341937;-0.462527317984896;-0.043149867790536 2;0.008500939628181 98
-3.66603806887673;0.040533402346 1462;-1.81214933456537;1.42842528582614;-1.71548434050041;1.62984917989802;0.207007973461431;-0.538856801875074;-
0.013972838991769 4;-0.383239601796243;2.30794350677442;0.333151817437956;-0.877884448234885;-0.743297435947398;-0.972312669153 88;-
1.68844492460254;0.579040119427 91;-0.717622142033 18;-0.838093523126091;-1.62245175495065;0.538635732231559;0.385783766968803;1.35612169347029;-
0.999035556448768;1.41439327622851;-0.811619700117558;-0.652628820792095;0.057704606724926 3;0.162134286704143;0.298343498479876;-0.158720117144092;-
0.443351319905787;0.196964664405961;0.287861904790051;-0.391722154219812;0.269296548922348;1.02374994575456;0.132216603958664;0.113205181743037;-
0.275610854806034;0.121123553120483;0.461547089884782;0.149506962616352;0.220785490563439;-0.431862221898506;-0.560446852418 97;-
1.15535504313704;0.424243198950193;0.220819855808734;0.284200832538836;-0.466988222702398;-0.385237952852 36;-0.132016421169319;-0.113681266419068
-3.83265937820526;-0.032257279962648 3;-1.91869839173992;0.822947377140584;-1.78264265978812;1.57815263445 87;0.598228666916479;-
0.654824599648395;0.096192931525926 8;-0.804103018512743;2.27297230699681;0.431018480126555;-0.848864748124 99;-0.815406275976604;-1.36460656984998;-
1.61169528052136;0.486965082451928;-0.720406346923 88;-0.756076236328314;-1.33849629747858;0.445397766204165;0.728196963985361;1.25108767858813;-
1.10797973956121;0.710503918509611;-1.18896756137538;-0.439015965553627;0.340186525934337;0.086329023646116 3;0.164612227248230;-0.058066755088450 6;-
0.583161291625632;0.205060133427212;0.609666247924545;-0.253681043029627;0.156582027299861;1.02264018034174;0.117658890942417;-0.198151660166062;-
0.545233522055342;0.227845817434955;0.256391923314255;0.322385578402513;0.279465112006521;-0.329436557970751;-0.581405406362526;-

Literature Fingerprinting

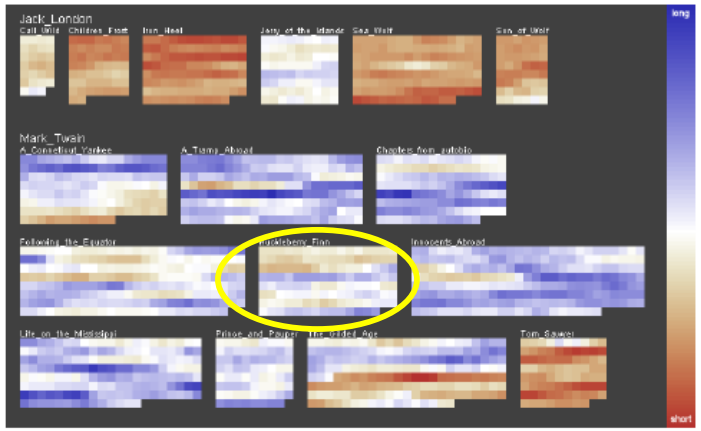




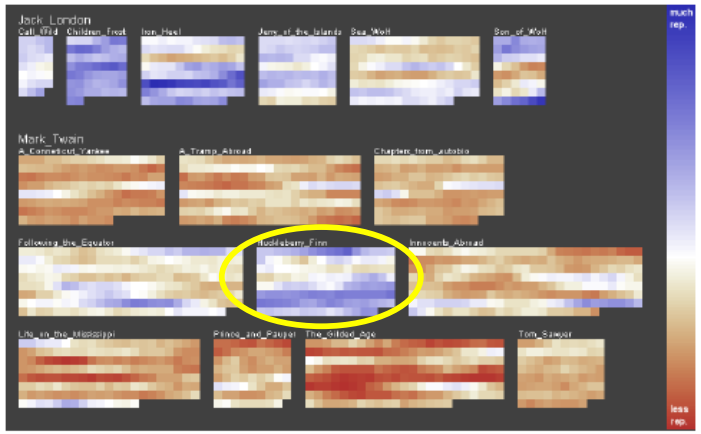
(a) Function words (First Dimension after PCA)



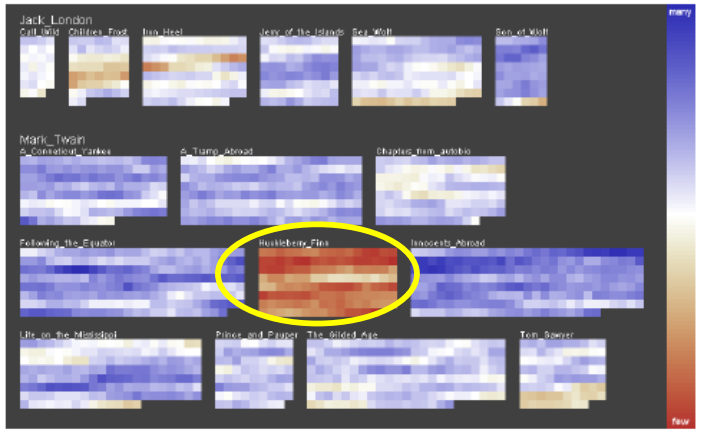
(b) Function words (Second Dimension after PCA)



(c) Average sentence length

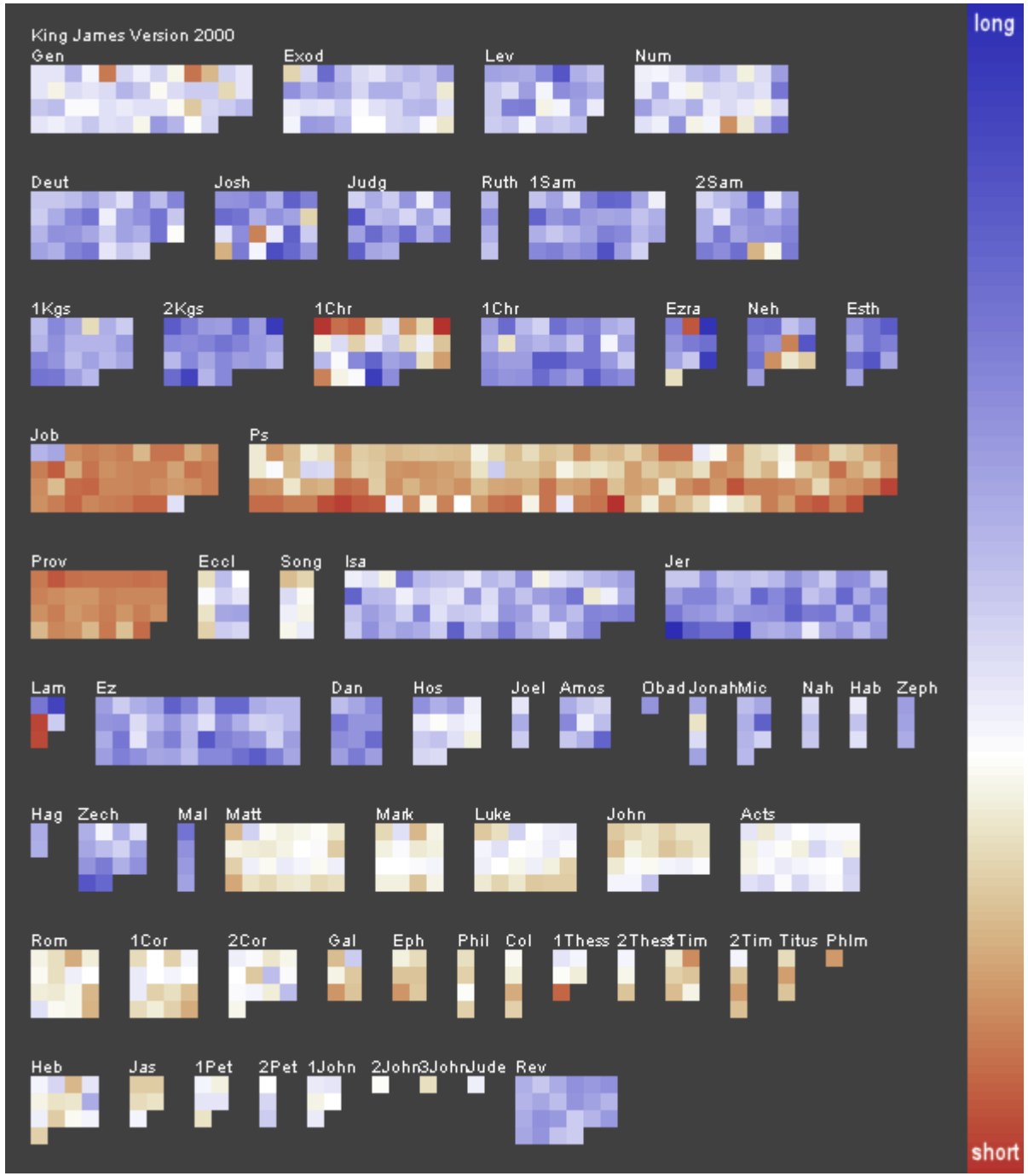


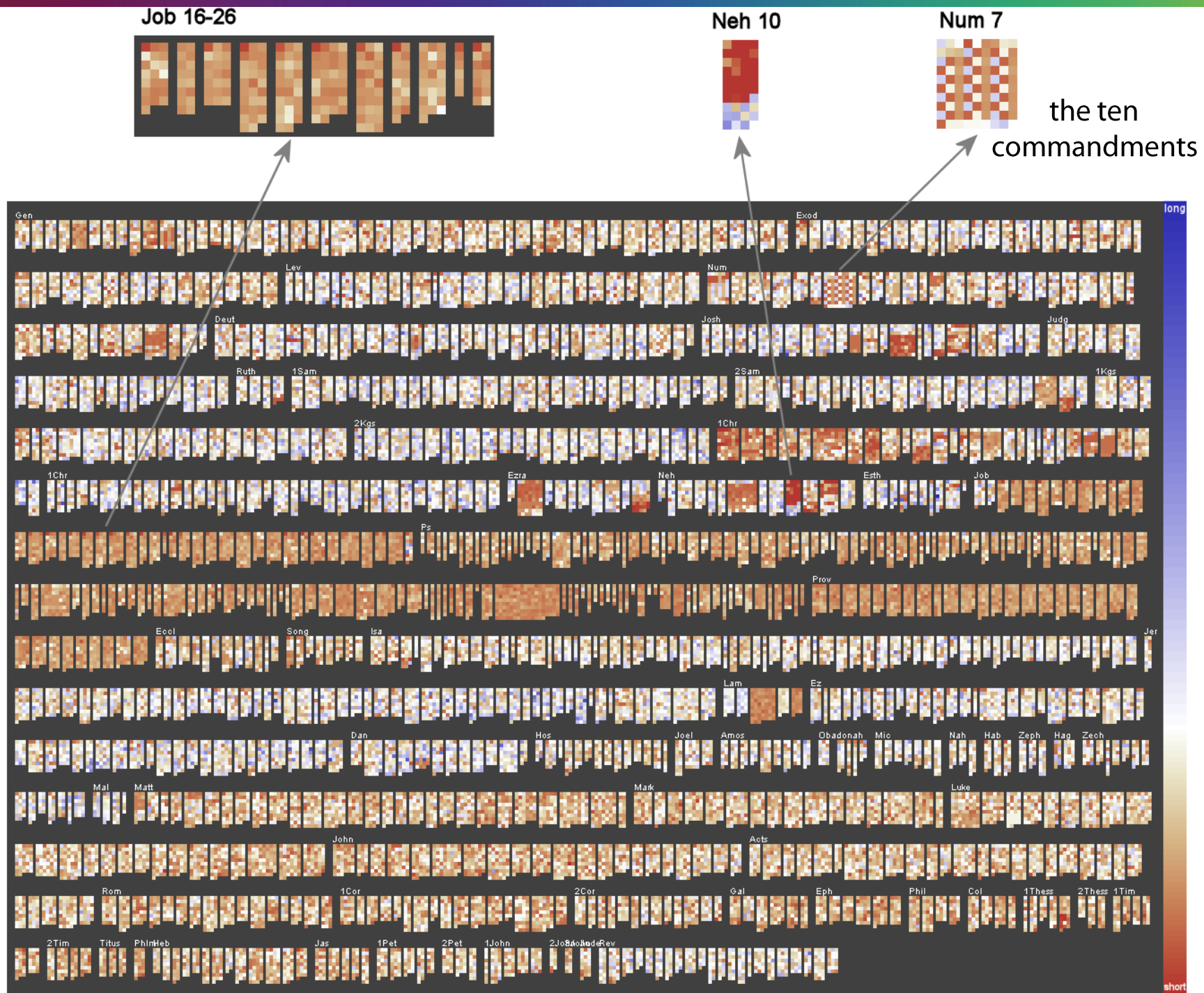
(d) Simpson's Index



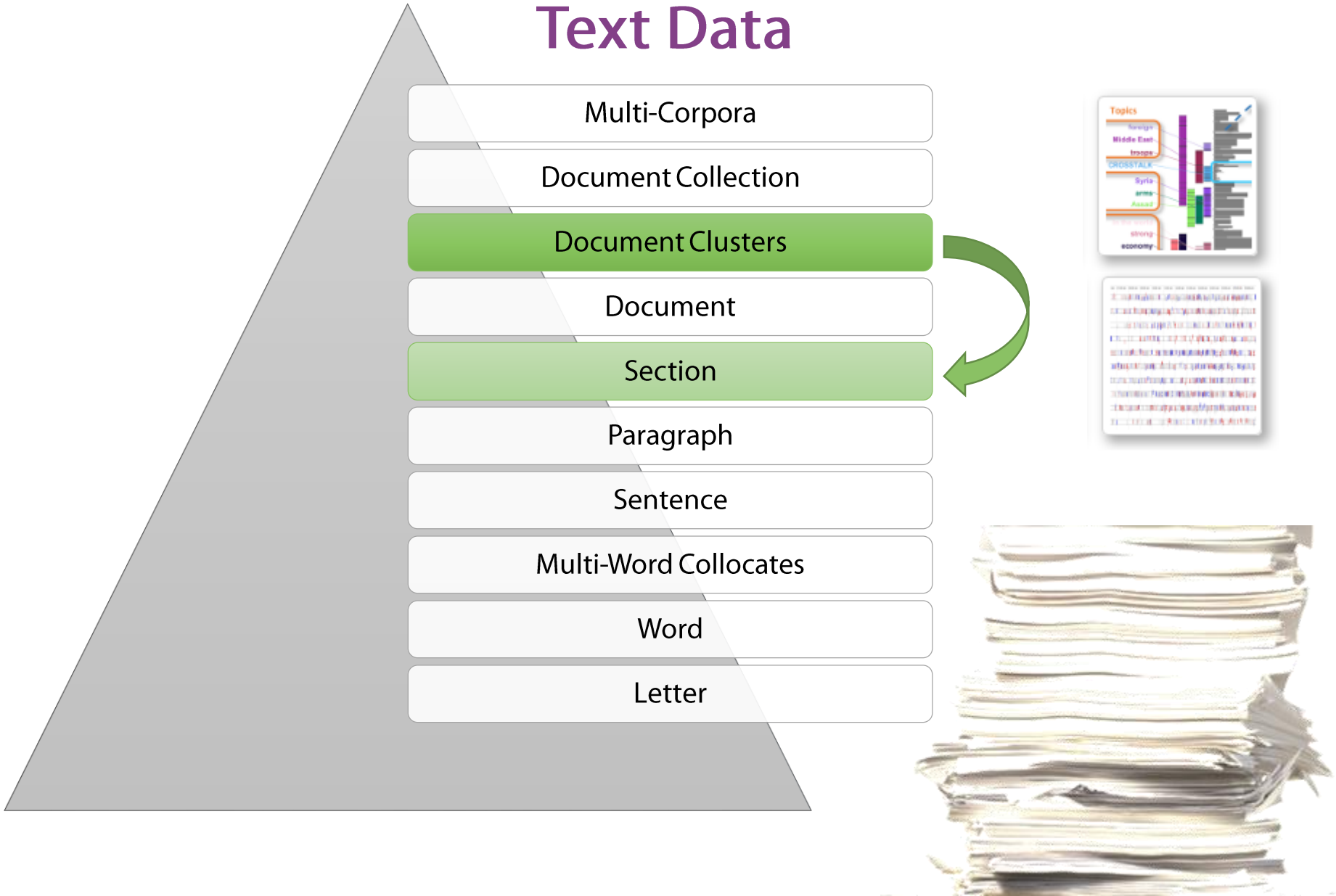
(e) Hapax Legomena

(f) Hapax Dislegomena

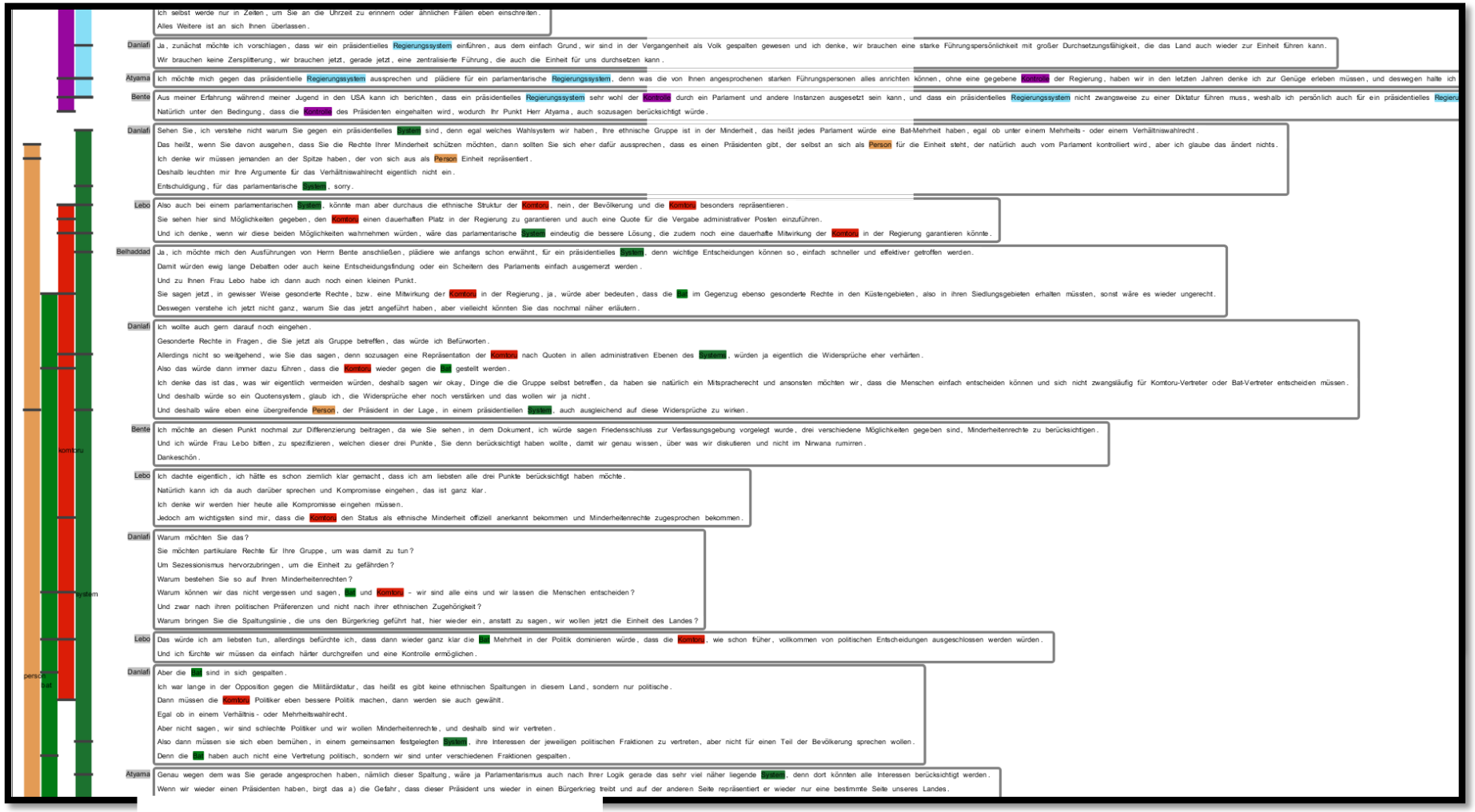

King James Version 2000
Gen
Exod
Lev
Num
Deut
Josh
Judg
Ruth
1Sam
2Sam
1Kgs
2Kgs
1Chr
1Chr
Ezra
Neh
Esth
Job
Ps
Prov
Eccl
Song
Isa
Jer
Lam
Ez
Dan
Hos
Joel
Amos
Obad
Jonah
Mic
Nah
Hab
Zeph
Hag
Zech
Mal
Matt
Mark
Luke
John
Acts
Rom
1Cor
2Cor
Gal
Eph
Phil
Col
1Thess
2Thess
1Tim
2Tim
Titus
Phlm
Heb
Jas
1Pet
2Pet
1John
2John
3John
Jude
Rev
long
short


Job 16-26
Neh 10
Num 7
the ten commandments
Gen
Exod
Lev
Num
Deut
Josh
Judg
Ruth
1Sam
2Sam
1Kgs
2Kgs
1Chr
1Chr
Ezra
Neh
Esth
Job
Ps
Prov
Eccl
Song
Isa
Jer
Lam
Ez
Dan
Hos
Joel
Amos
Obadonah
Mic
Nah
Hab
Zeph
Hag
Zech
Mal
Matt
Mark
Luke
John
Acts
Rom
1Cor
2Cor
Gal
Eph
Phil
Col
1Thess
2Thess
1Tim
2Tim
Titus
Phlm
Heb
Jas
1Pet
2Pet
1John
Rev
long
short

Text Data


Multi-Corpora
Document Collection
Document Clusters
Document
Section
Paragraph
Sentence
Multi-Word Collocates
Word
Letter
Topics
CROSSTALK

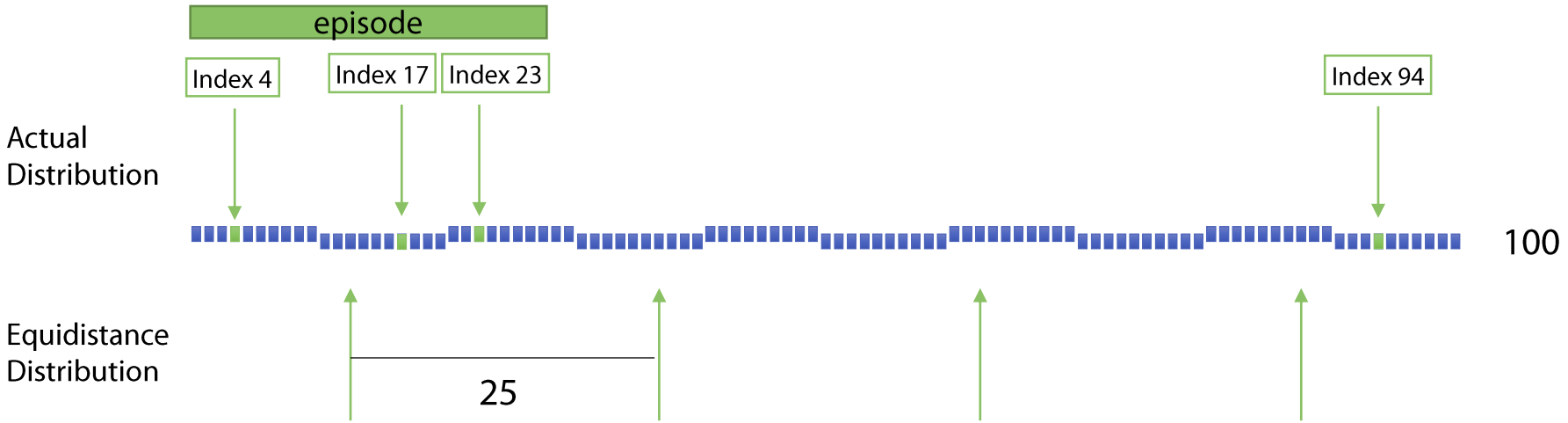
Lexical Episodes Plots



Close Reading

Lexical Episodes

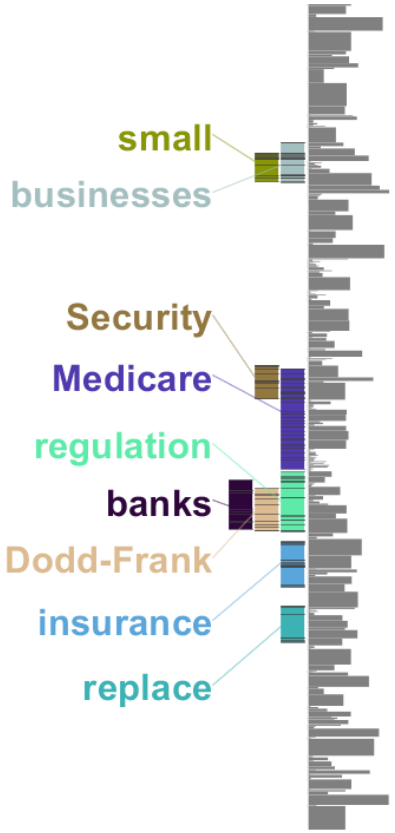
= portion within the word sequence of a corpus where a certain word appears more densely than expected from its frequency in the whole text.



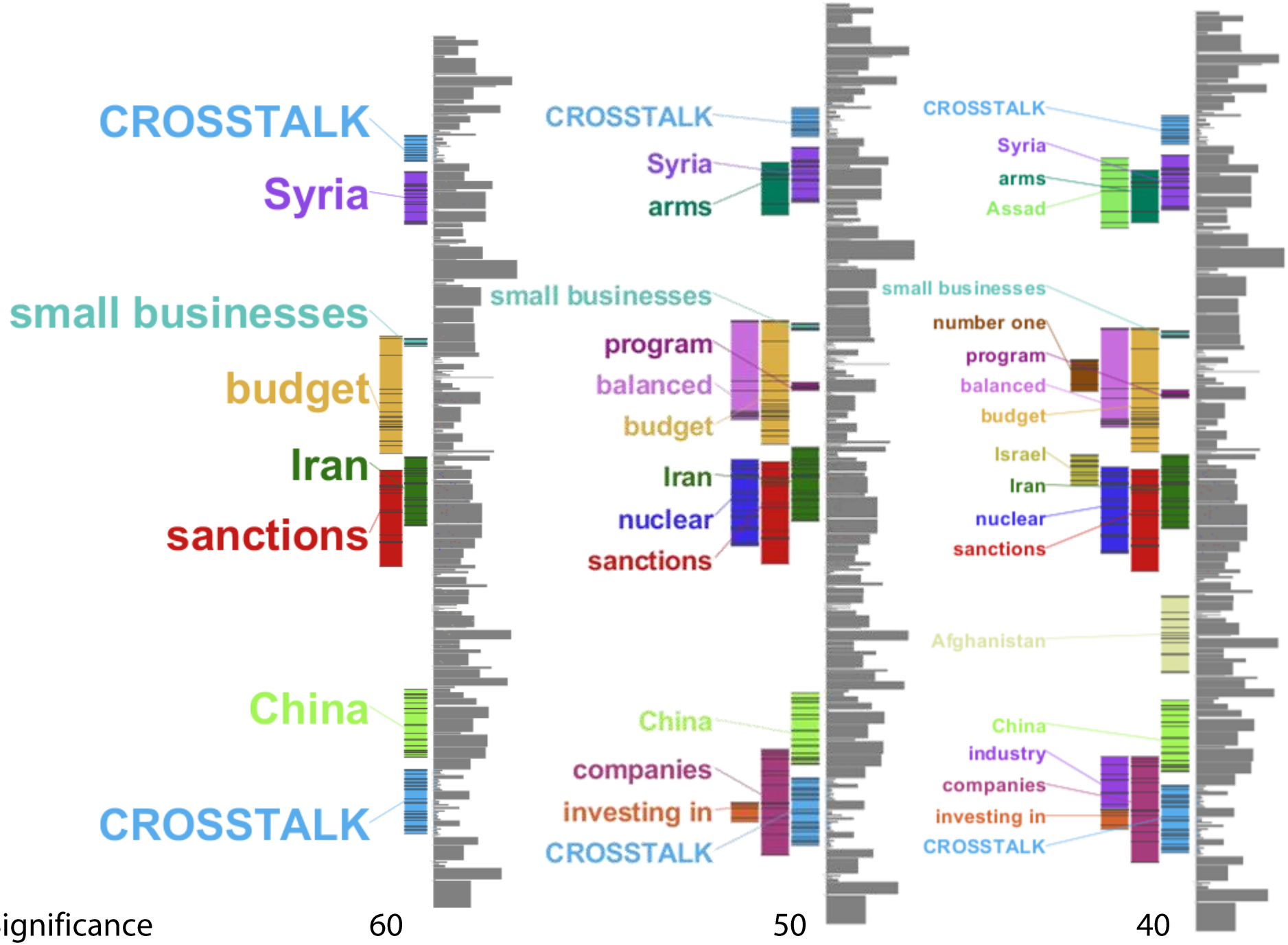
Use Case

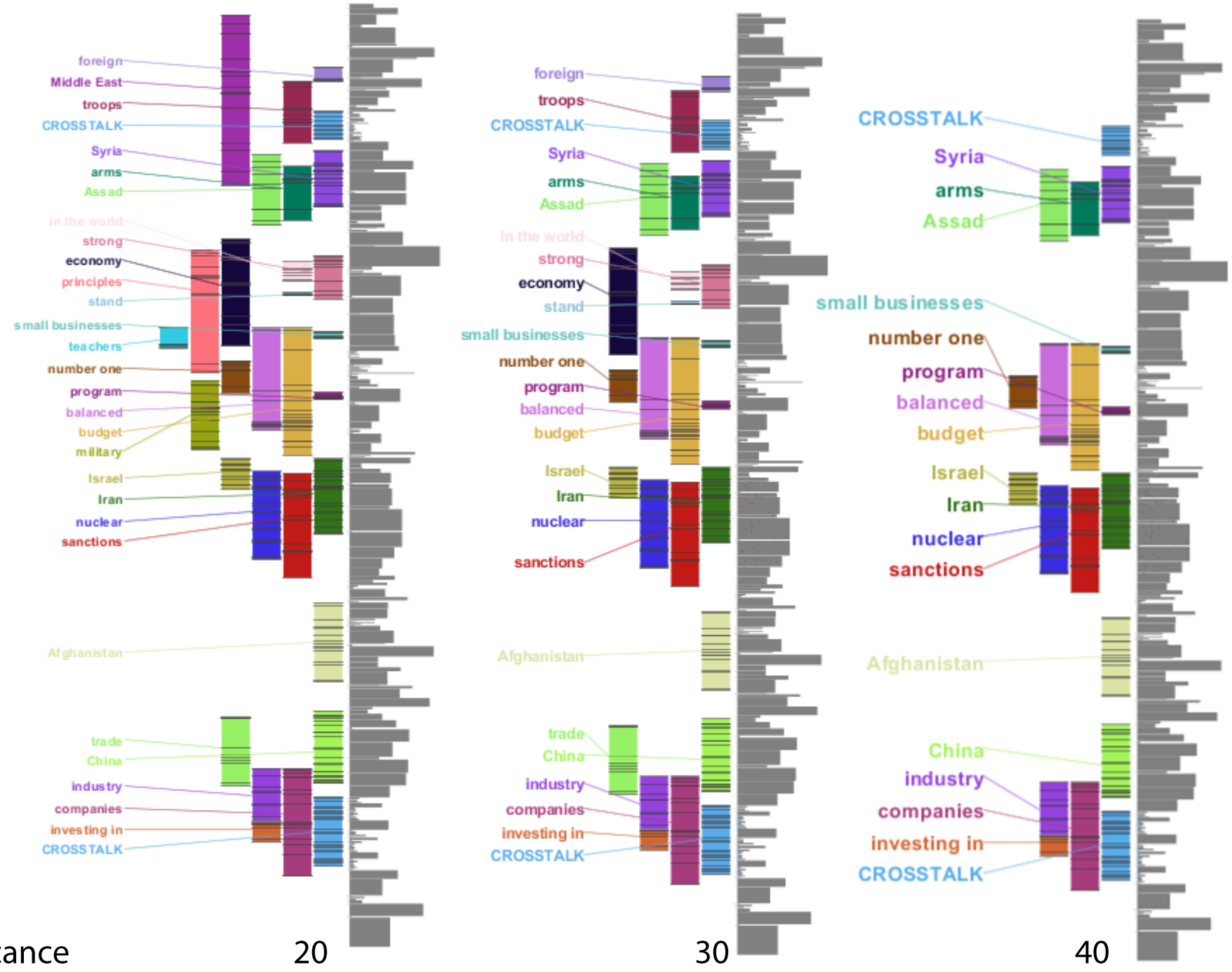
US Presidential Debates: Obama vs. Romney

Overview | 3 Debates



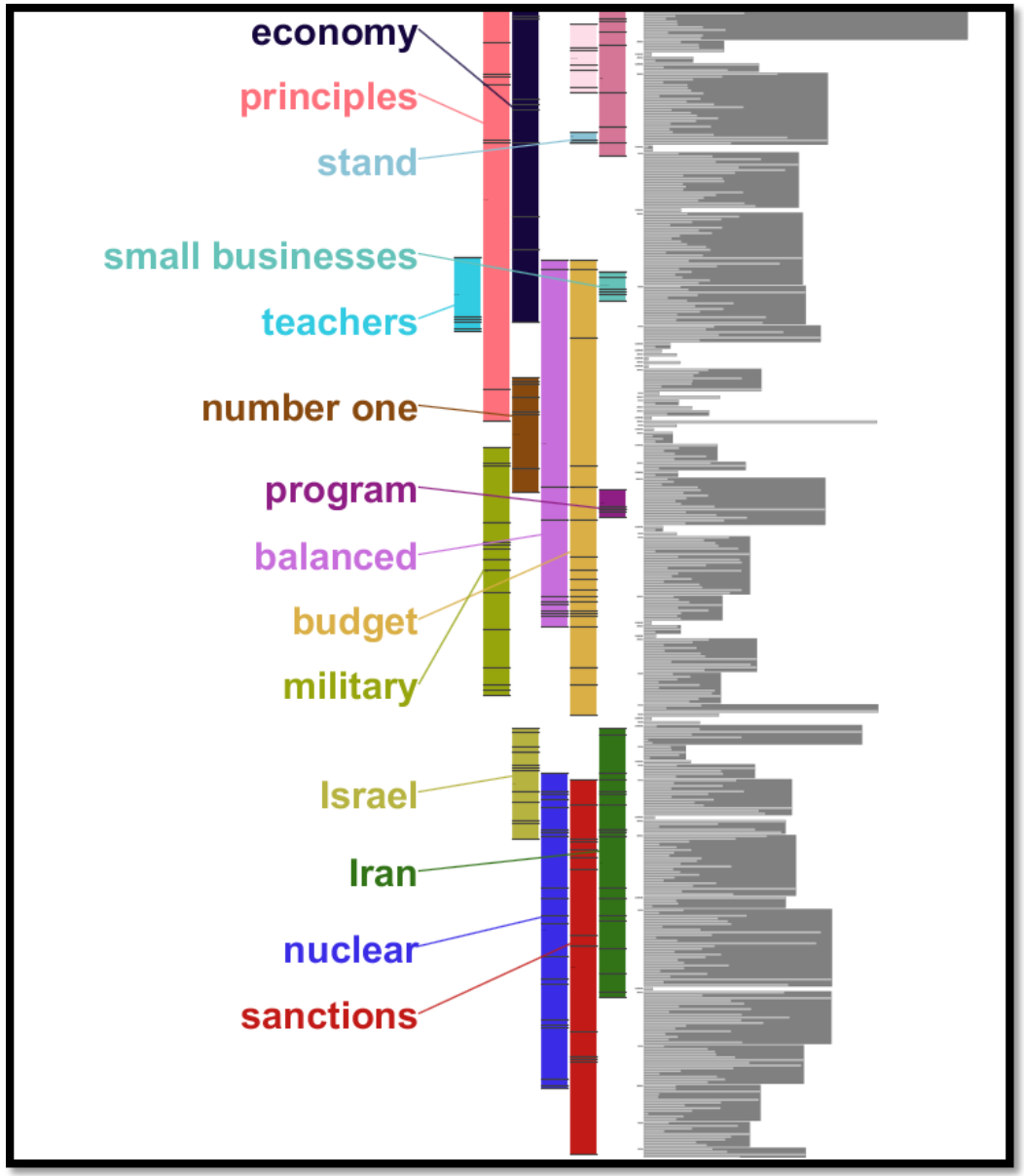
First Obama – Romney
Presidential Debate

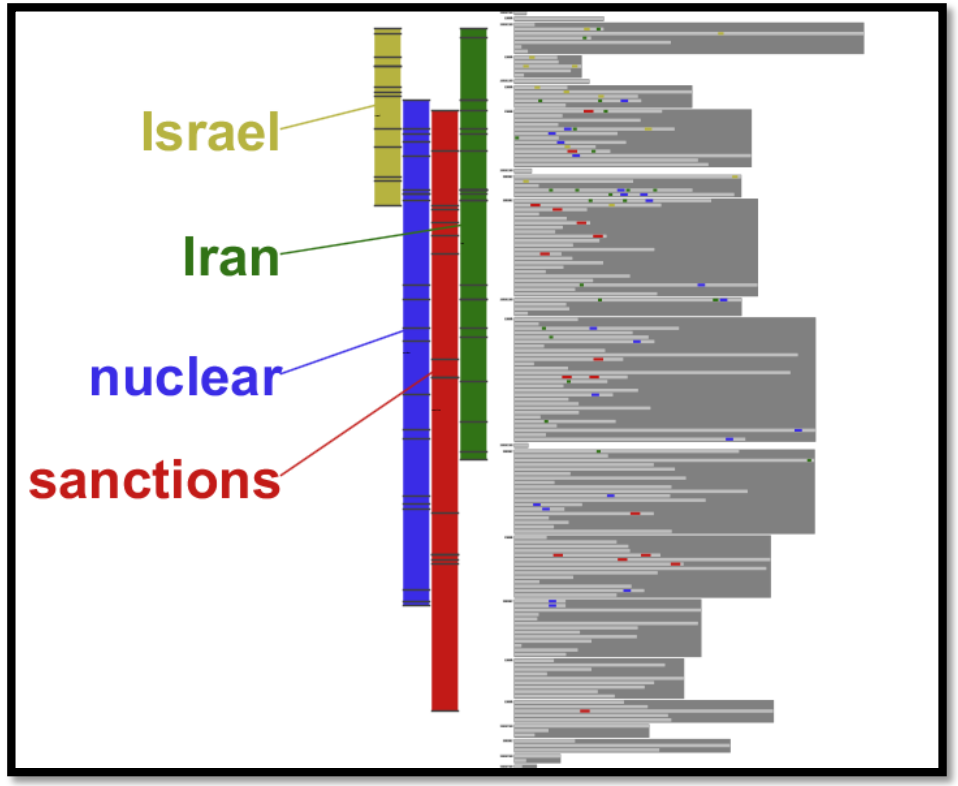

CROSSTALK
Syria
small businesses
budget
Iran
sanctions
China
CROSSTALK
CROSSTALK
Syria
arms
small businesses
program
balanced
budget
Iran
nuclear
sanctions
China
companies
investing in
CROSSTALK
CROSSTALK
Syria
arms
Assad
small businesses
number one
program
balanced
budget
Israel
Iran
nuclear
sanctions
Afghanistan
China
industry
companies
investing in
CROSSTALK
Significance
60
50
40

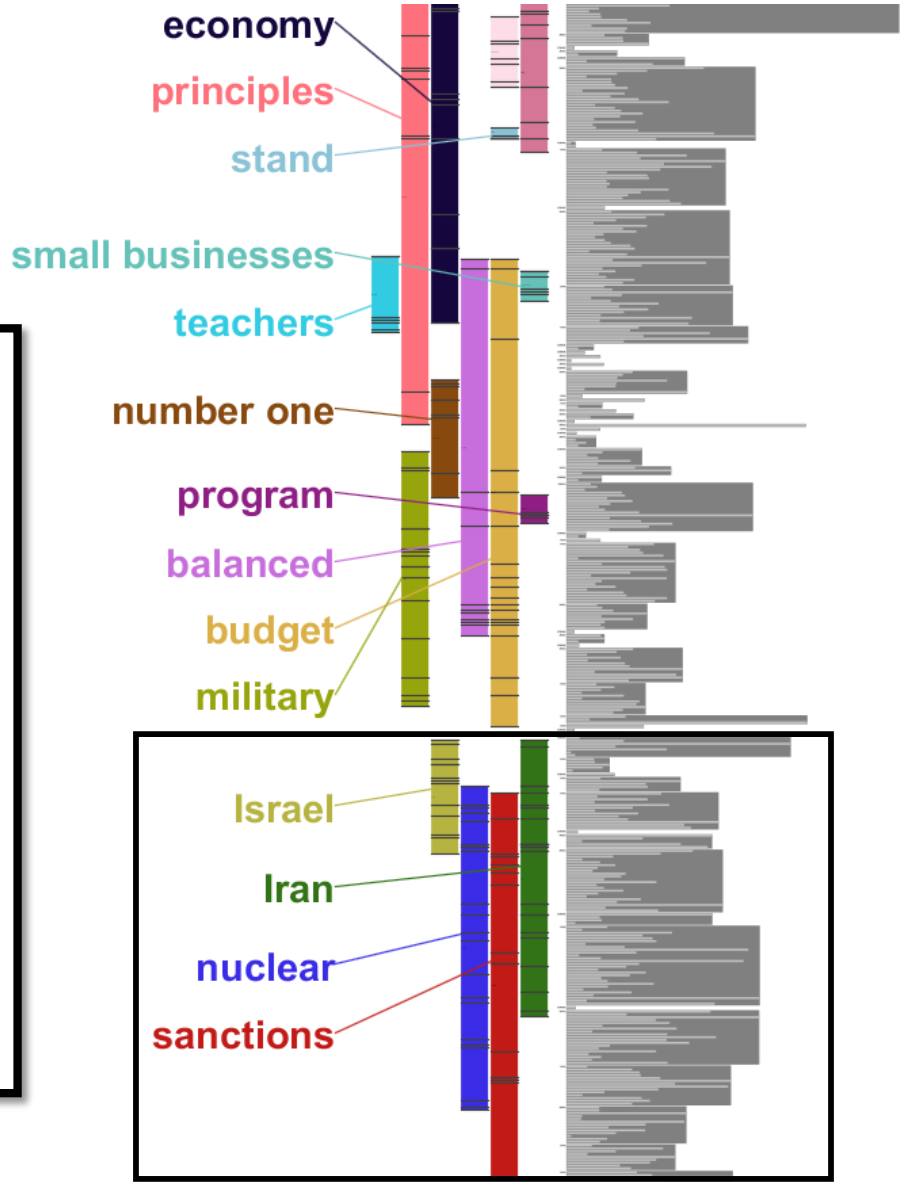


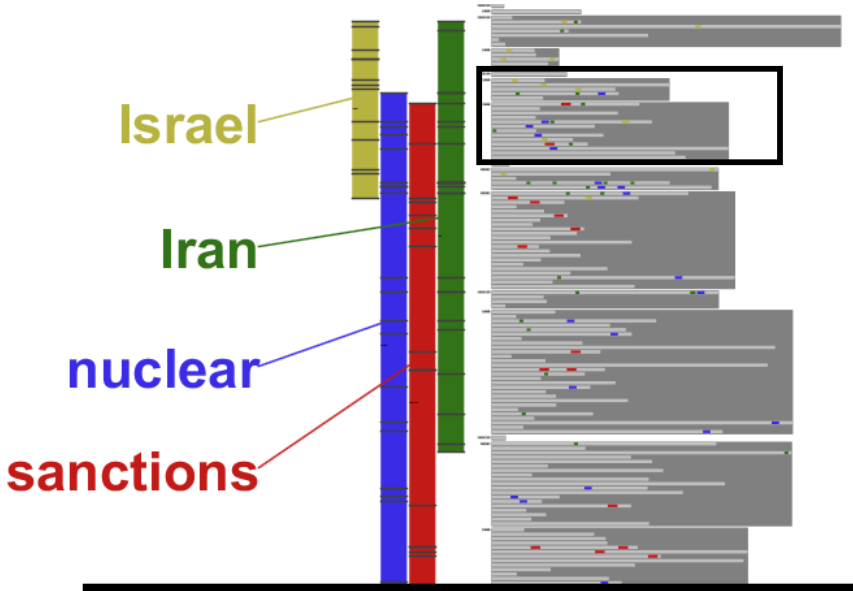
Significance 20 30 40


foreign
Middle East
troops
CROSSTALK
Syria
arms
Assad
in the world
strong
economy
principles
stand
small businesses
teachers
number one
program
balanced
budget
military
Israel
Iran
nuclear
sanctions
Afghanistan
trade
China
industry
companies
investing in
CROSSTALK


economy
principles
stand
small businesses
teachers
number one
program
balanced
budget
military
Israel
Iran
nuclear
sanctions


Israel
Iran
nuclear
sanctions


economy
principles
stand
small businesses
teachers
number one
program
balanced
budget
military
Israel
Iran
nuclear
sanctions

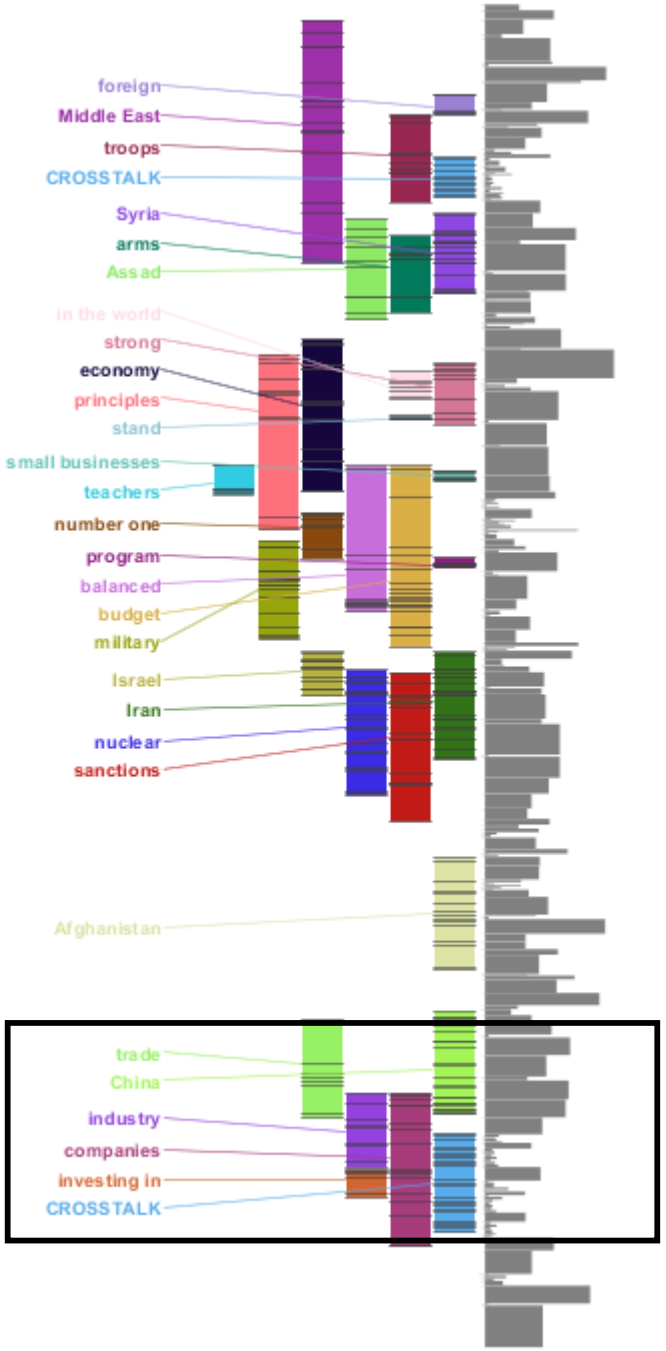


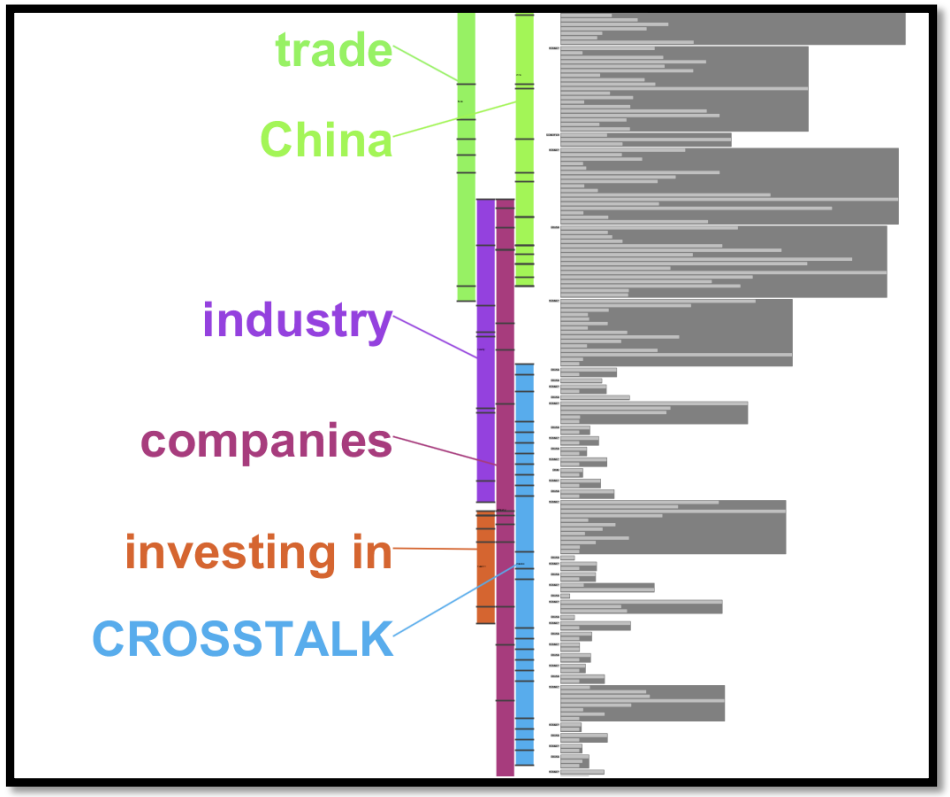
OBAMA First of all, Israel is a true friend.
It is our greatest ally in the region.
And if Israel is attacked, America will stand with Israel.
I've made that clear throughout my presidency.
And...

SCHIEFFER So you're -- you're saying we've already made that declaration.

OBAMA I will stand with Israel if they are attacked.
And this is the reason why, working with Israel, we have created the strongest military and intelligence cooperation between our two countries in history.
In fact, this week we'll be carrying out the largest military exercise with Israel in history, this very week.
But to the issue of Iran, as long as I'm president of the United States Iran will not get a nuclear weapon.
I made that clear when I came into office.

OBAMA We then organized the strongest coalition and the strongest sanctions against Iran in history, and it is crippling their economy.
Their currency has dropped 80 percent.
Their oil production has plunged to the lowest level since they were fighting a war with Iraq 20 years ago.
So their economy is in a shambles.
And the reason we did this is because a nuclear Iran is a threat to our national security, and it is a threat to Israel's national security.
We cannot afford to have a nuclear arms race in the most volatile region of the world.
Iran is a state sponsor of terrorism.
And for them to be able to provide nuclear technology to non-state actors, that's unacceptable.
And they have said that they want to see Israel wiped off the map.
So the work that we've done with respect to sanctions now offers Iran a choice.
They can take the diplomatic route and end their nuclear program or they will have to face a united world and a United States president, me, who said we're not going to take any options off the table.
The disagreement I have with Governor Romney is that, during the course of this campaign, he's often talked as if we should take premature military action.
I think that would be a mistake, because when I've sent young men and women into harm's way, I always understand that that is the last resort, not the first resort.


foreign
Middle East
troops
CROSSTALK
Syria
arms
Assad
In the world
strong
economy
principles
stand
small businesses
teachers
number one
program
balanced
budget
military
Israel
Iran
nuclear
sanctions
Afghanistan
trade
China
industry
companies
investing in
CROSSTALK


trade
China
industry
companies
investing in
CROSSTALK

ROMNEY ...if -- if you're...
(CROSSTALK)

OBAMA ...you've had the floor for a while.
(CROSSTALK)

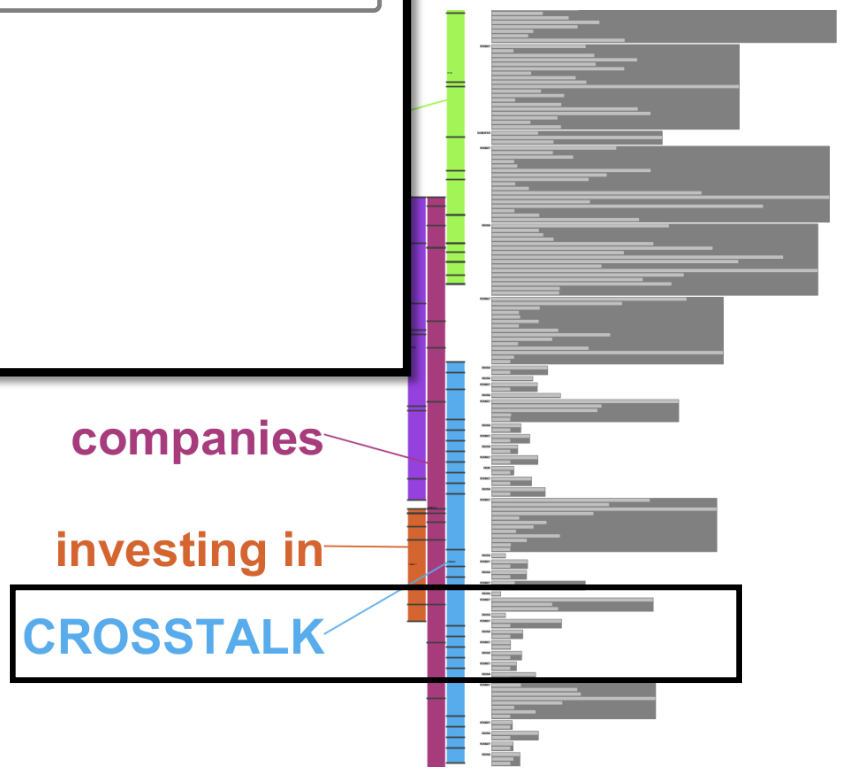
ROMNEY ...get someone else's.
OBAMA: The -- look, I think anybody out there can check the record.
Governor Romney, you keep on trying to, you know airbrush history here.
You were very clear that you would not provide, government assistance to the US. auto companies, even if they went through bankruptcy.
You said that they could get it in the private marketplace.
That wasn't true.
They would have gone through a...
(CROSSTALK)

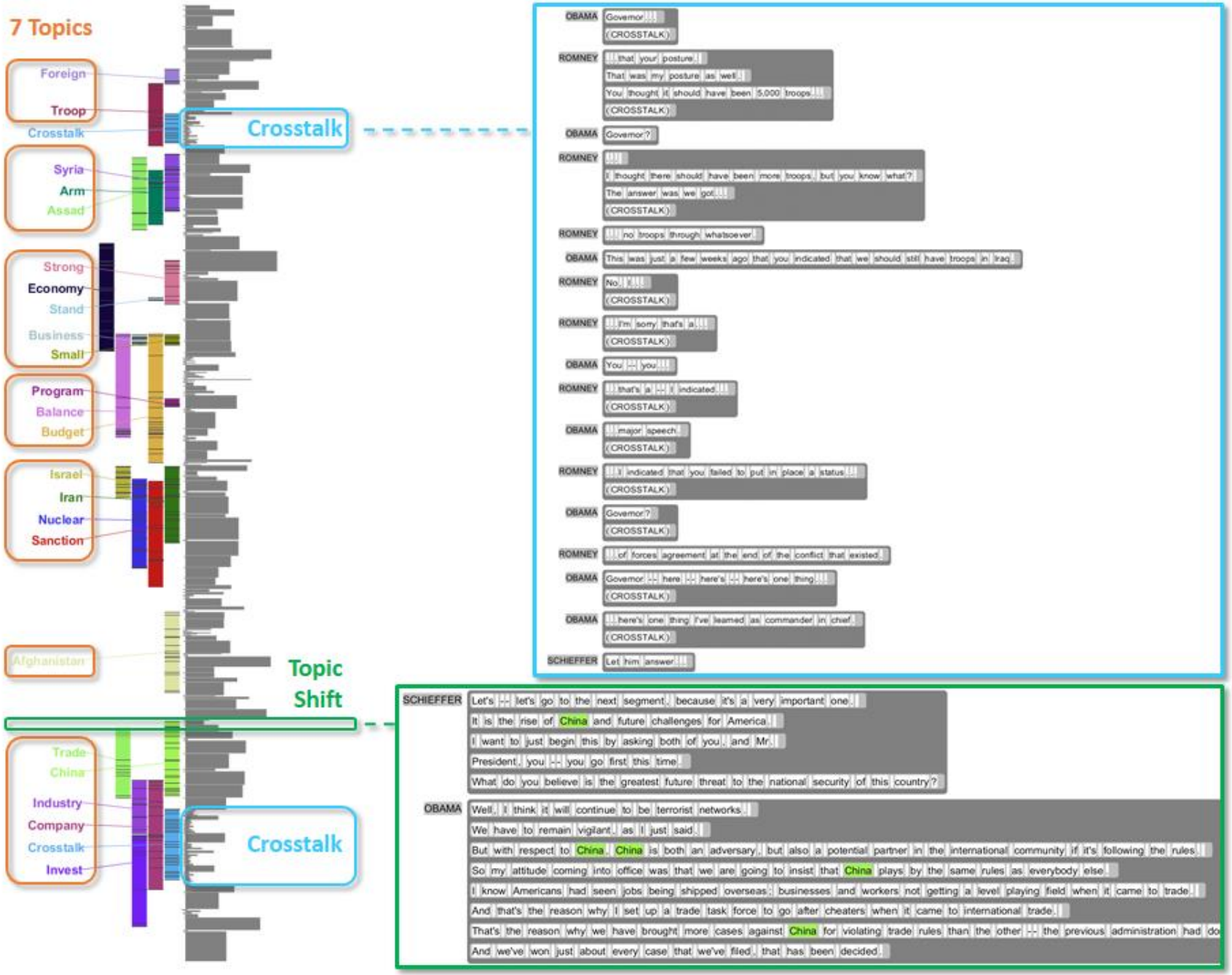
ROMNEY You're wrong...
(CROSSTALK)

OBAMA ...they would have gone through a...
(CROSSTALK)

ROMNEY ...you're wrong.
(CROSSTALK)

OBAMA No, I am not wrong.
I am not wrong.
(CROSSTALK)




7 Topics
Foreign
Troop
Crosstalk
Syria
Arm
Assad
Strong
Economy
Stand
Business
Small
Program
Balance
Budget
Israel
Iran
Nuclear
Sanction
Afghanistan
Trade
China
Industry
Company
Crosstalk
Invest
Crosstalk
Topic Shift
Crosstalk
OBAMA Governor...
(CROSSTALK)
ROMNEY ...that your posture.
That was my posture as well.
You thought it should have been 5,000 troops...
(CROSSTALK)
OBAMA Governor?
ROMNEY ...
I thought there should have been more troops, but you know what?
The answer was we got...
(CROSSTALK)
ROMNEY ... no troops through whatsoever.
OBAMA This was just a few weeks ago that you indicated that we should still have troops in Iraq.
ROMNEY No. Y...
(CROSSTALK)
ROMNEY ...I'm sorry that's a...
(CROSSTALK)
OBAMA You -- you...
ROMNEY ...that's a -- I indicated...
(CROSSTALK)
OBAMA ...major speech.
(CROSSTALK)
ROMNEY ...I indicated that you failed to put in place a status...
(CROSSTALK)
OBAMA Governor?
(CROSSTALK)
ROMNEY ...of forces agreement at the end of the conflict that existed.
OBAMA Governor -- here -- here's -- here's one thing...
(CROSSTALK)
OBAMA ...here's one thing I've learned as commander in chief.
(CROSSTALK)
SCHIEFFER Let him answer...
SCHIEFFER Let's -- let's go to the next segment, because it's a very important one.
It is the rise of China and future challenges for America.
I want to just begin this by asking both of you, and Mr.
President, you -- you go first this time.
What do you believe is the greatest future threat to the national security of this country?
OBAMA Well, I think it will continue to be terrorist networks.
We have to remain vigilant, as I just said.
But with respect to China, China is both an adversary, but also a potential partner in the international community if it's following the rules.
So my attitude coming into office was that we are going to insist that China plays by the same rules as everybody else.
I know Americans had seen jobs being shipped overseas; businesses and workers not getting a level playing field when it came to trade.
And that's the reason why I set up a trade task force to go after cheaters when it came to international trade.
That's the reason why we have brought more cases against China for violating trade rules than the other -- the previous administration had do
And we've won just about every case that we've filed, that has been decided.

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

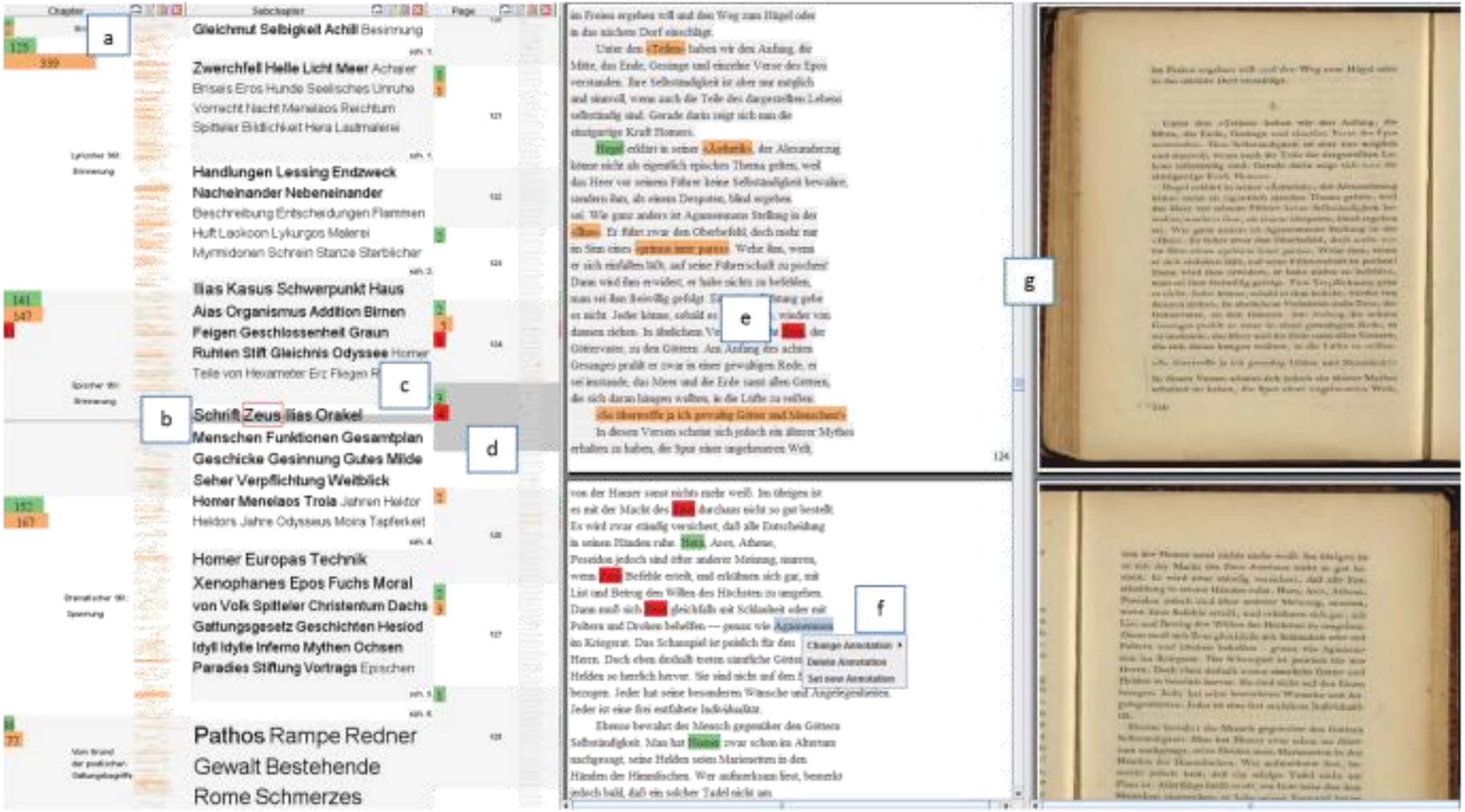
Interactive Information Visualization

Linguistic Information Visualization

Open Research Challenges

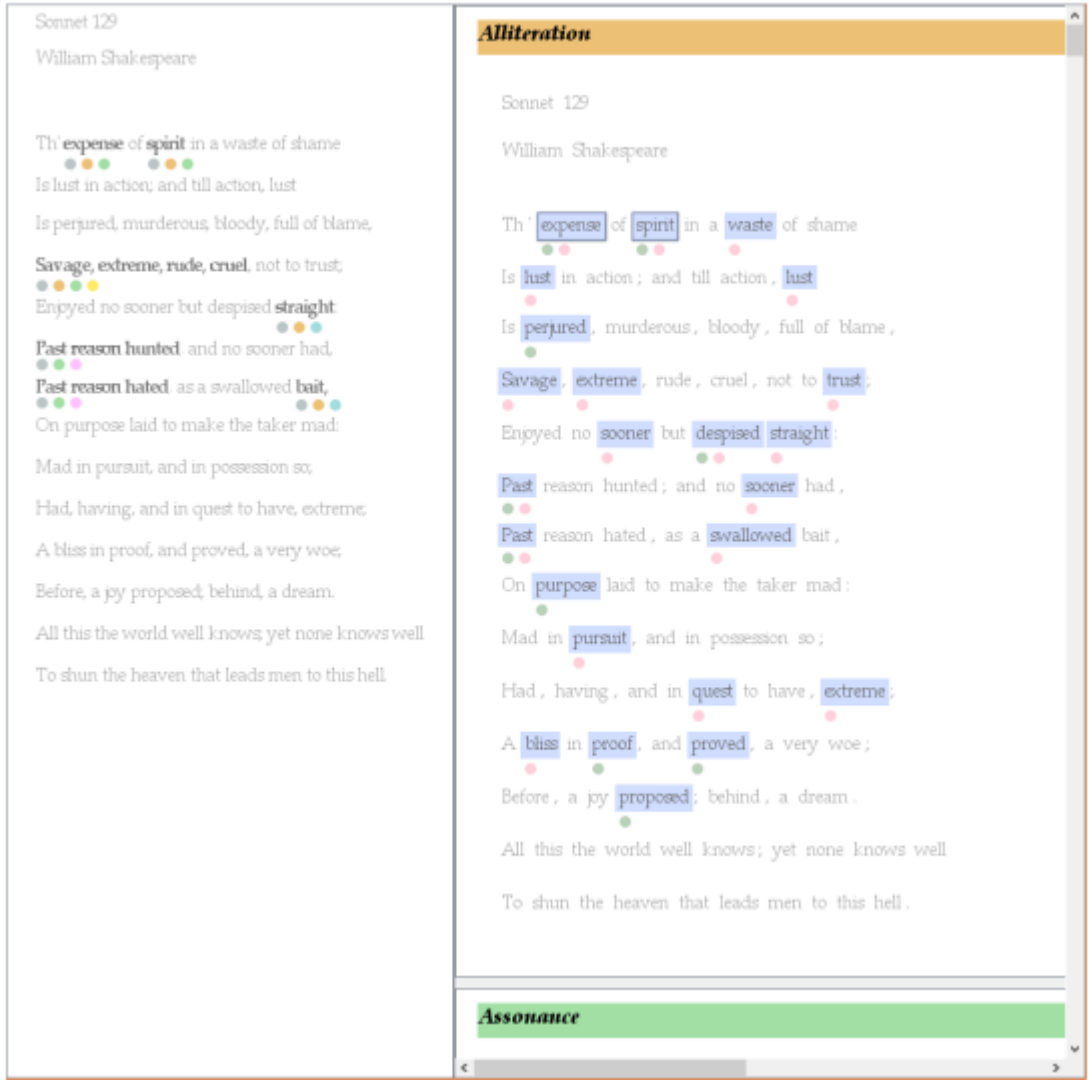
Resources

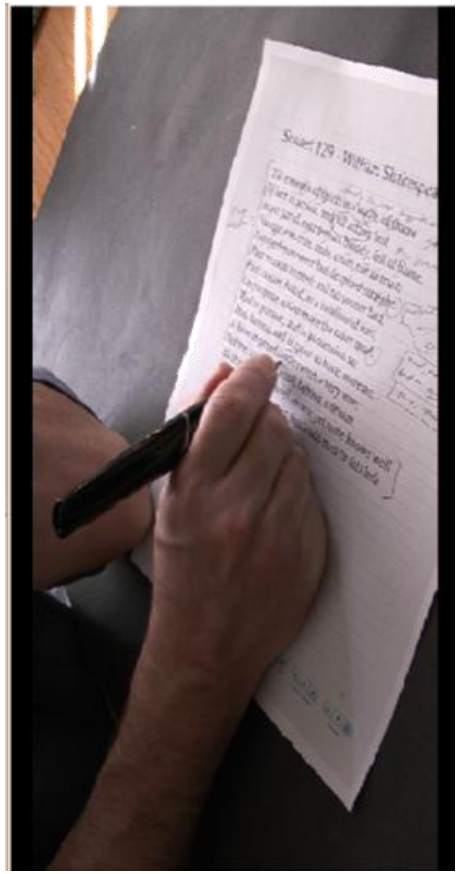
Unify Close and Distant Reading

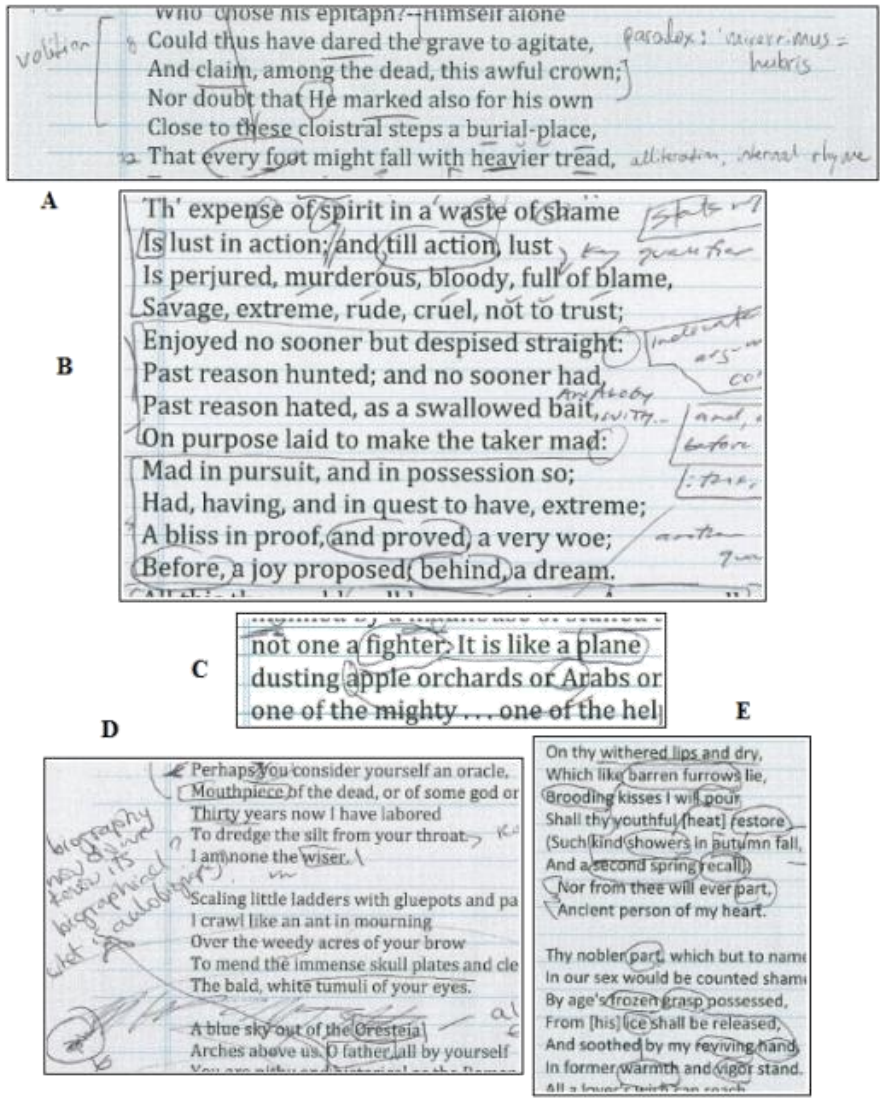


Koch et al., VarifocalReader — In-Depth Visual Analysis of Large Text Documents

“Prosthetic Reading” vs. “Slow Analysis”

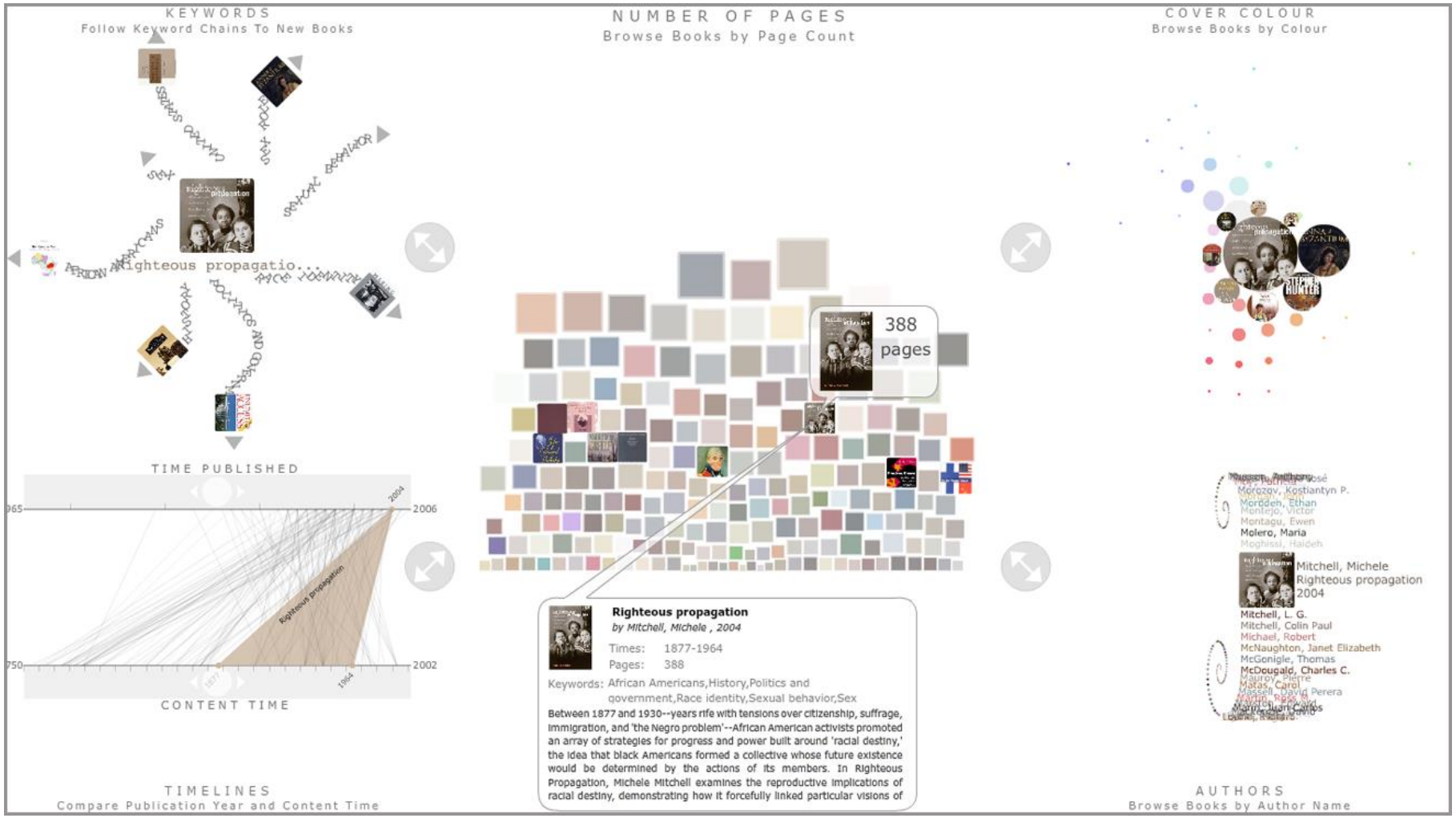






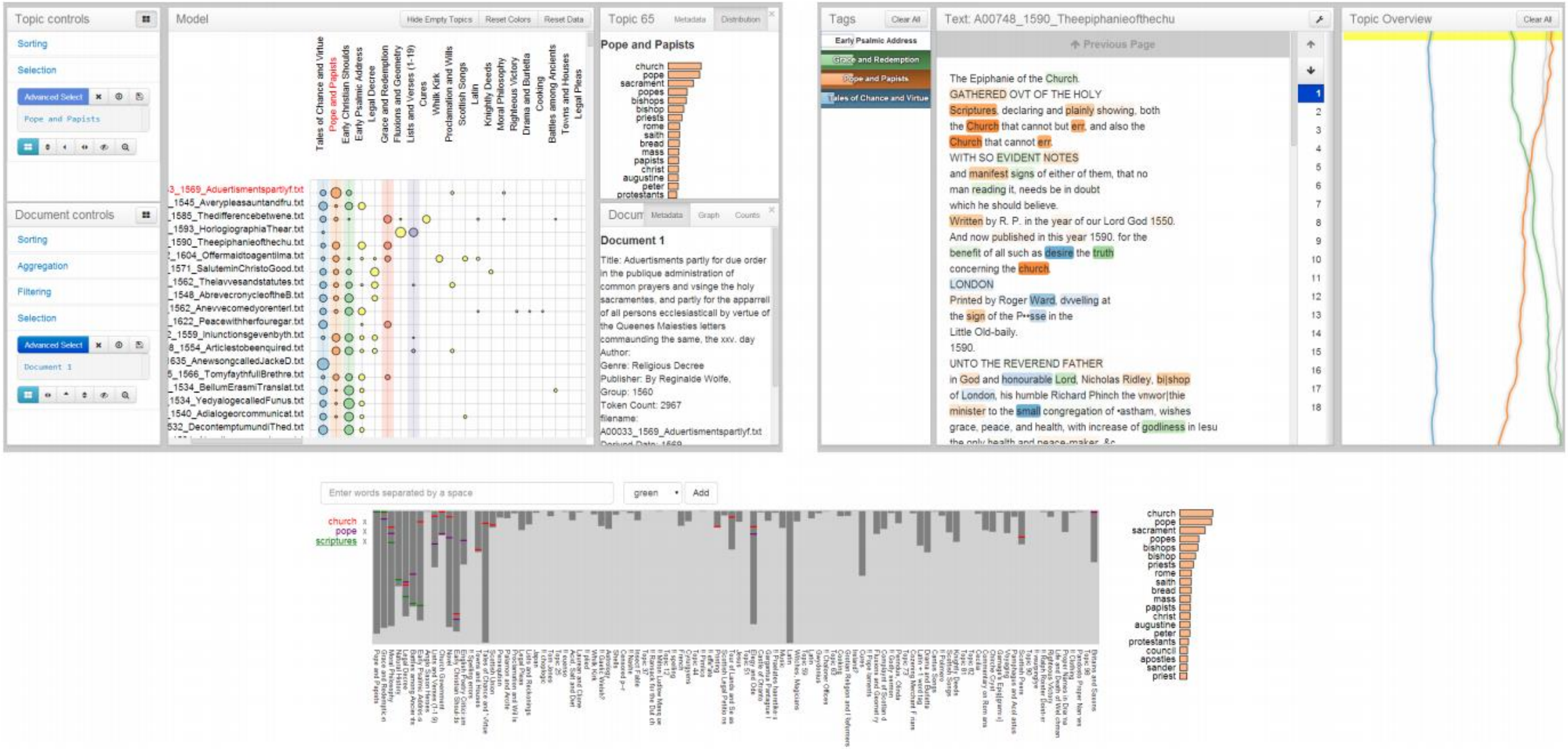
Mehta et al., MetaTation: Augmenting Free-Form Annotations to Bridge Close and Distant Reading of Poetry

Supporting Serendipitous and Guided Discovery



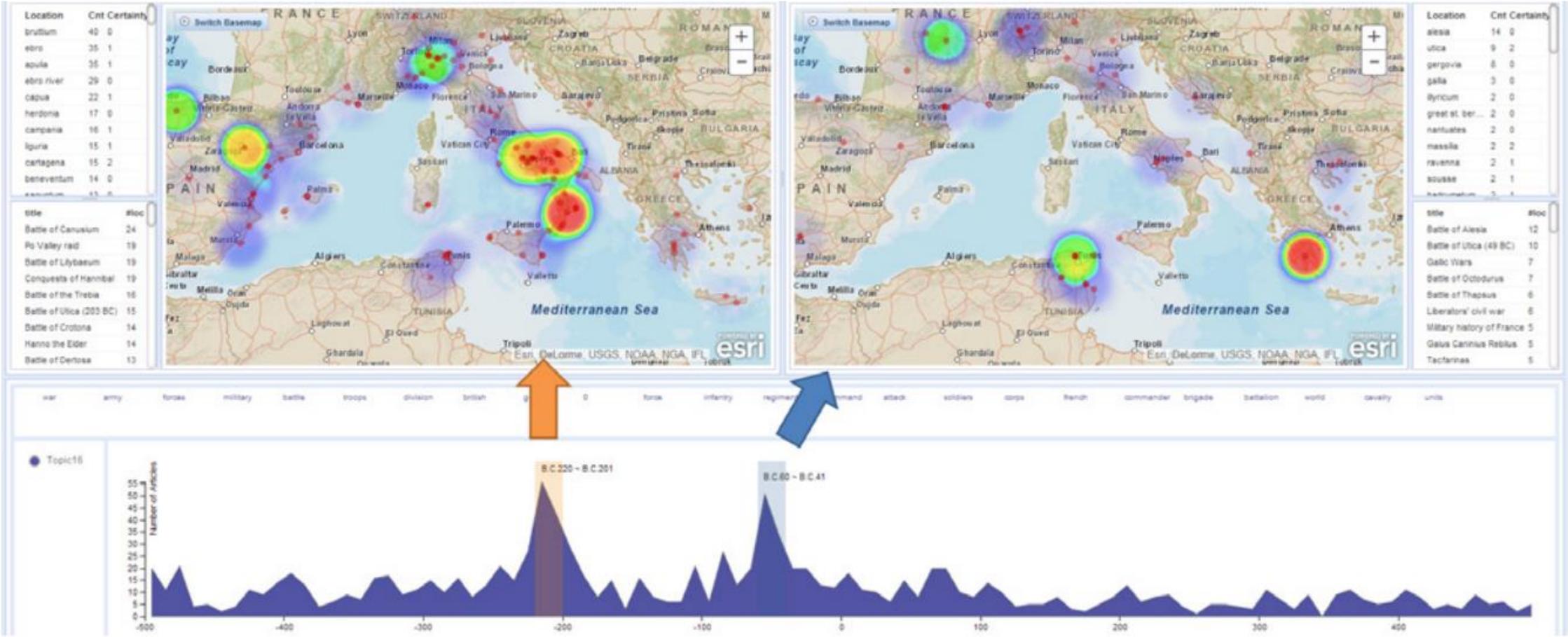
Thudt et al., The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization

Supporting Serendipitous and Guided Discovery



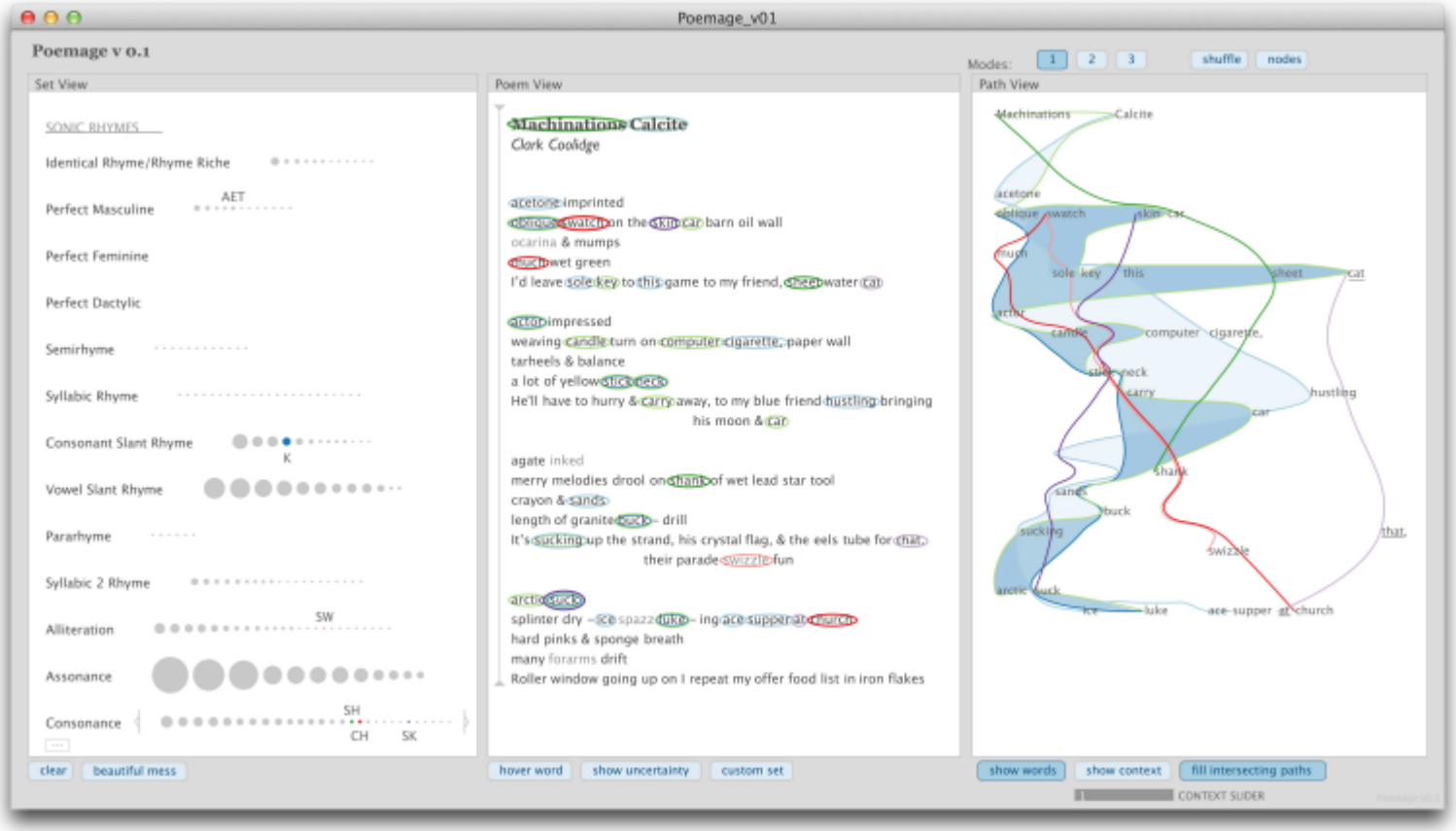
Alexander et al., Serendip: Topic Model-Driven Visual Exploration of Text Corpora

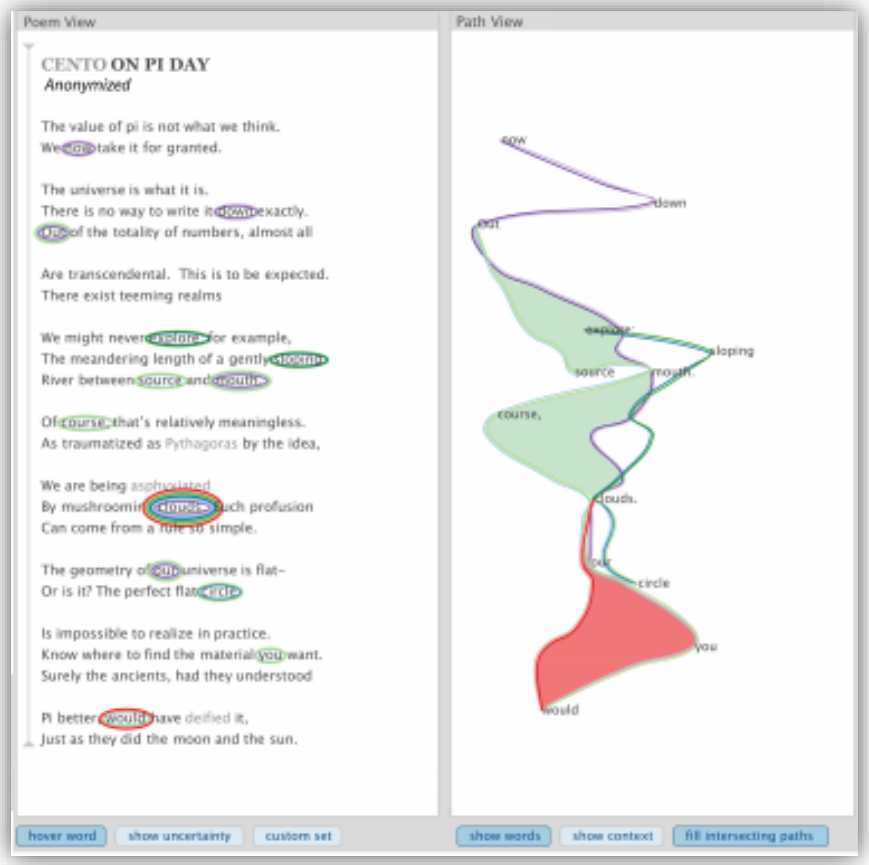
Linking Outside Resources into the Analysis



Cho et al., VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History

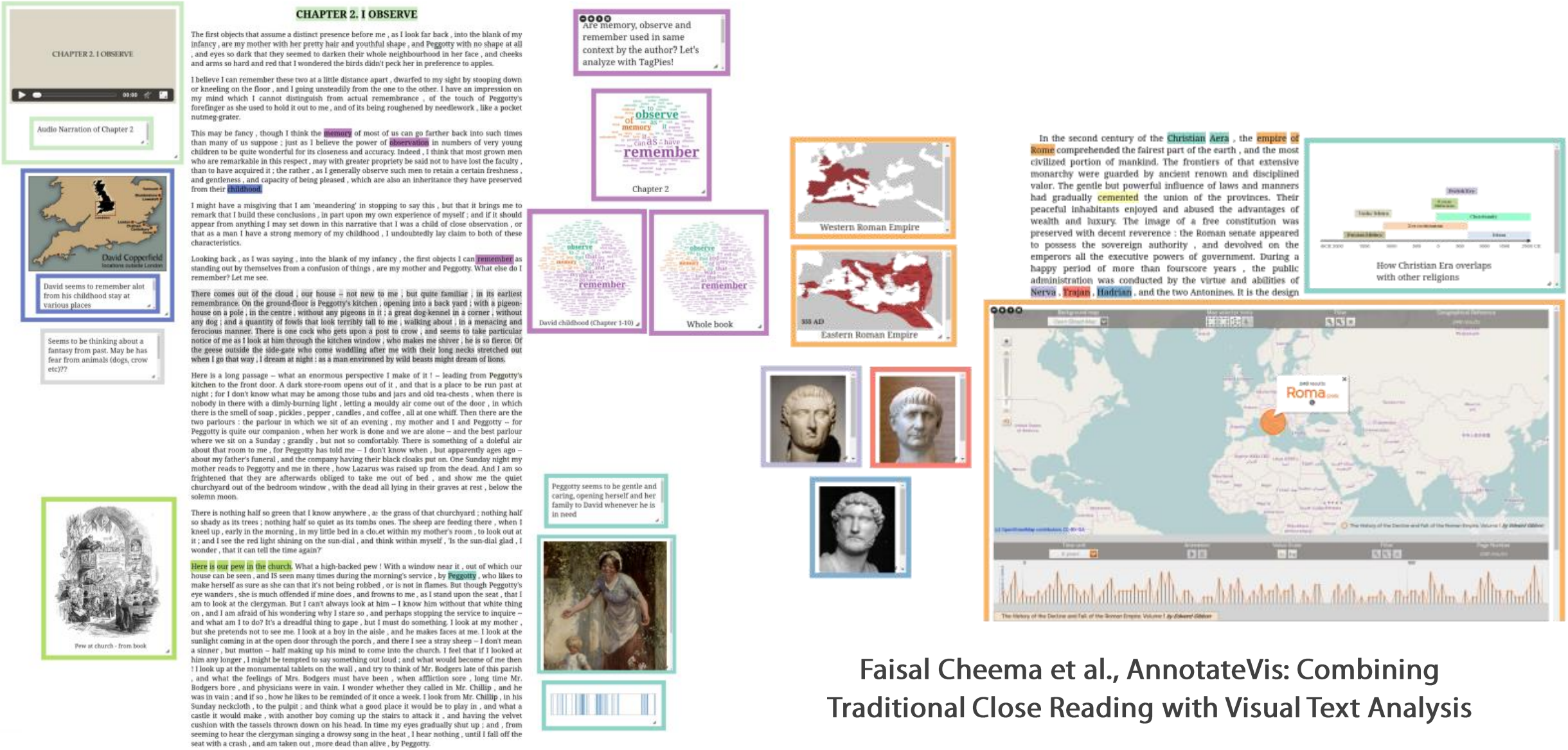
Aesthetic Experiences





McCurdy et al., Poemage: Visualizing the Sonic Topology of a Poem

Annotation of Resources



Faisal Cheema et al., AnnotateVis: Combining Traditional Close Reading with Visual Text Analysis

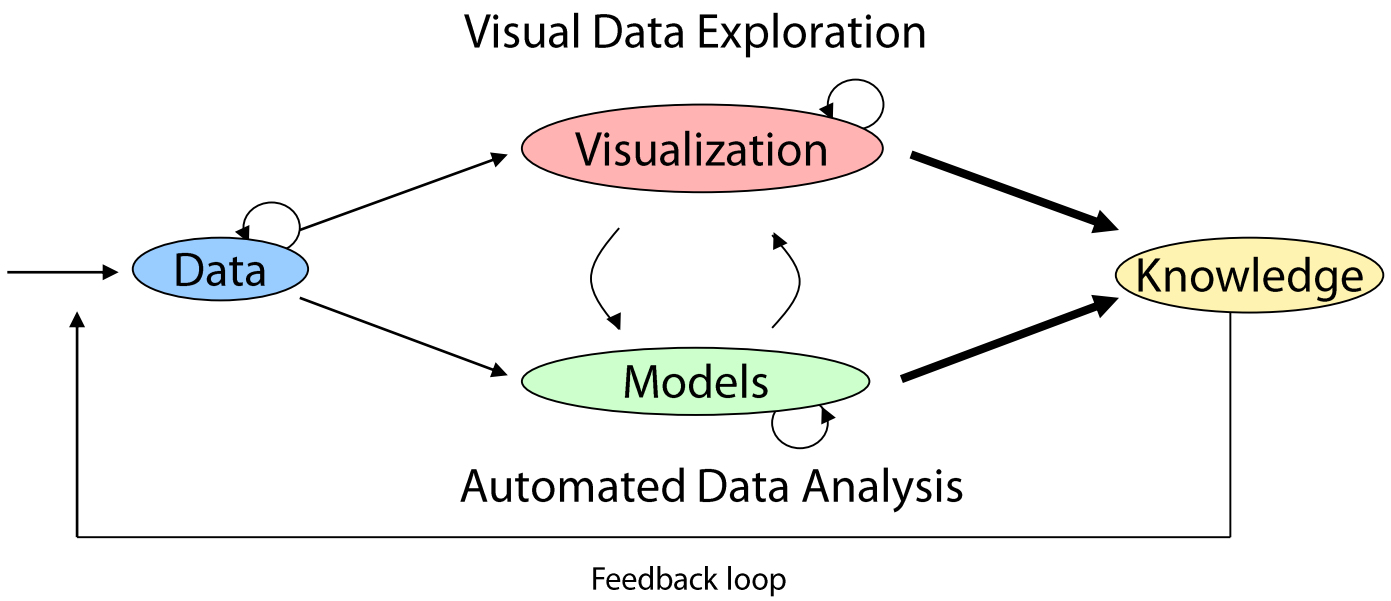
Take-Home Message…

Visual Analytics in the Linguistics is central to allow humans and computers to cooperate effectively!

Computers allow efficient automated processing of large data.

Humans interact and understand the data to generate knowledge.

However, …



- If automated analysis solves the problem:
  → **DO Automated Analysis!**
- If information visualization solves the problem:
  → **DO Information Visualization!**
- If combining both is the solution:
  → **DO Visual Analytics!**

Motivation

Introduction

Data Foundations

Key Aspects of Visualization

Designing Visualizations

Interactive Information Visualization

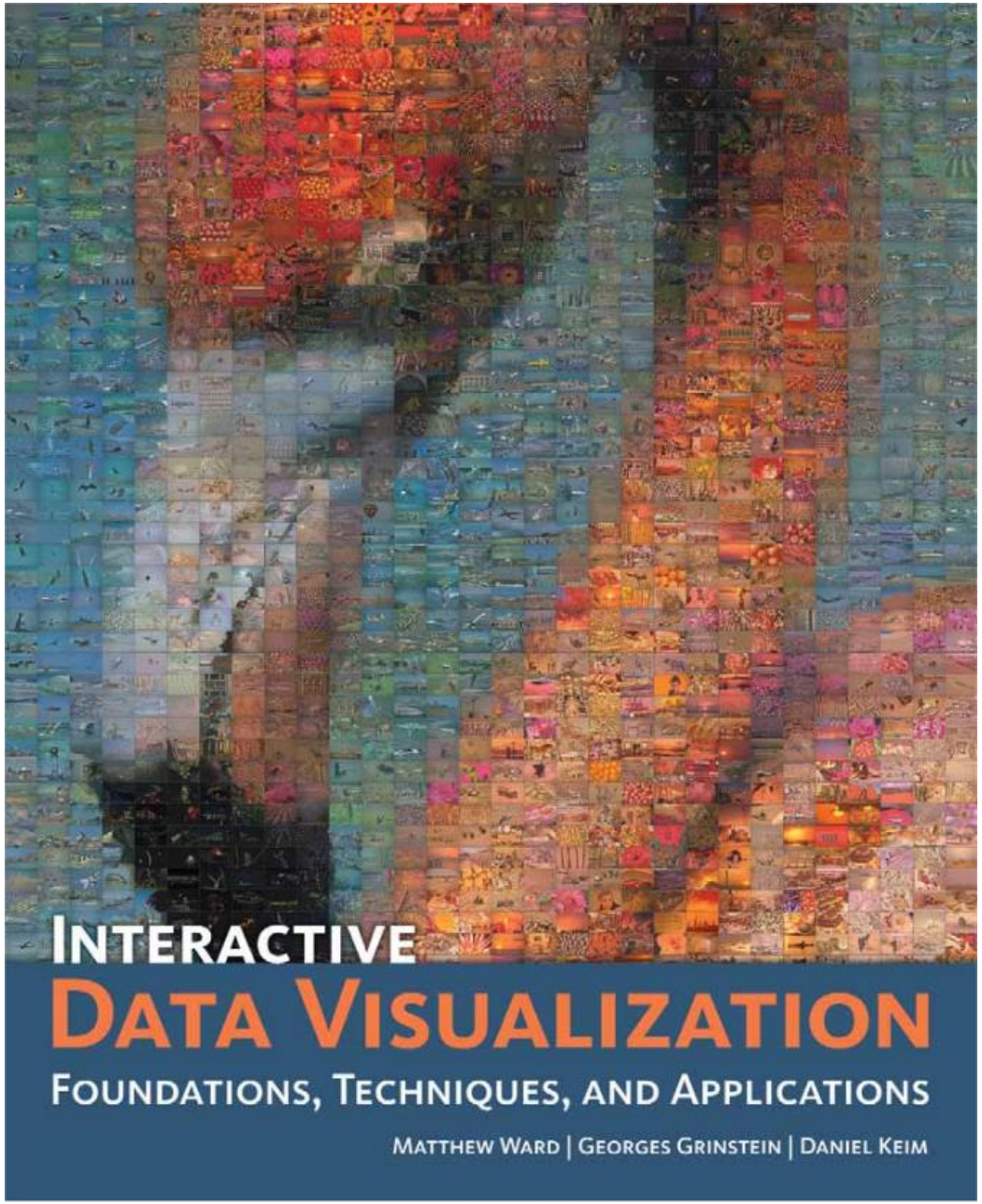
Linguistic Information Visualization

Open Research Challenges

Resources

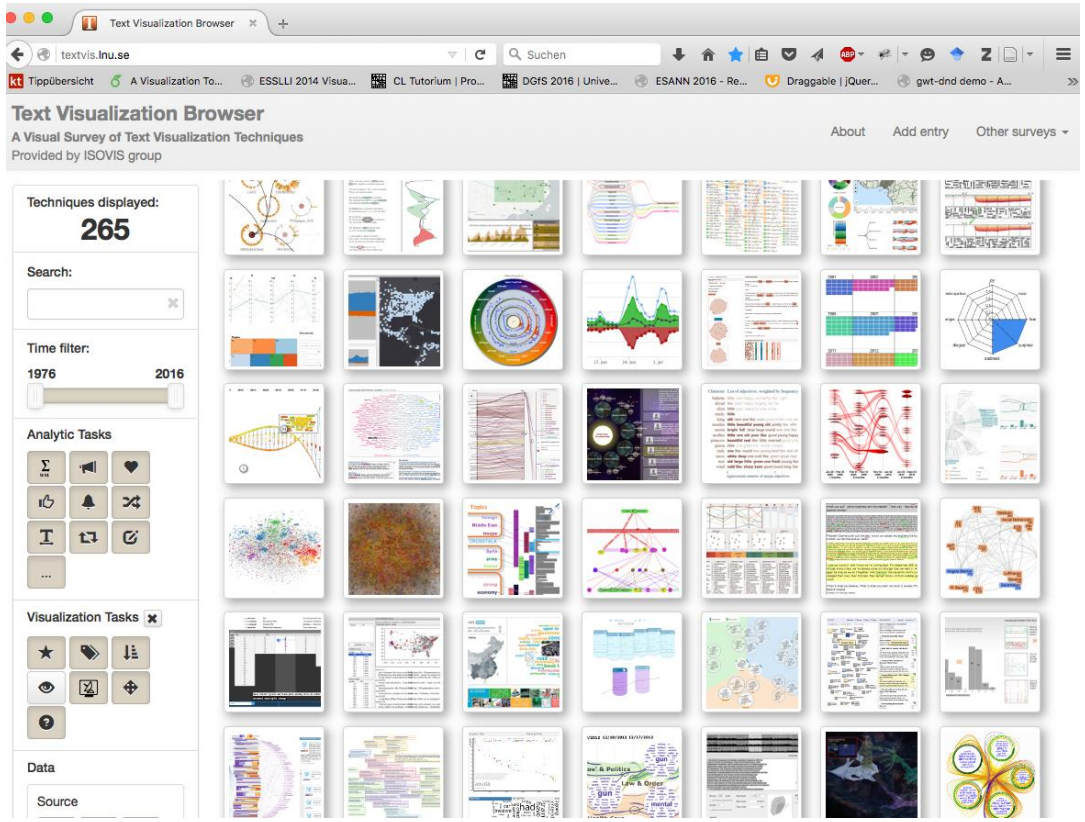
For More Detailed Information

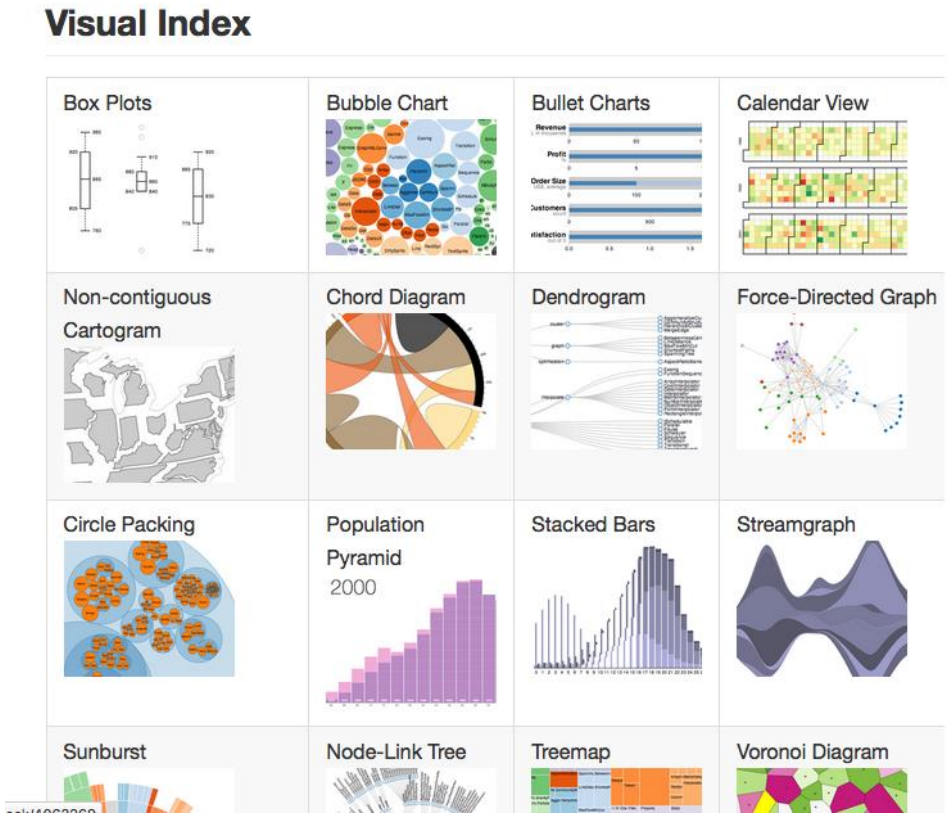
- IDV book
- Covers all important InfoVis aspects
- Good starting point
- Also used in our InfoVis lectures



Many Design Alternatives! – Checkout Existing Visualizations

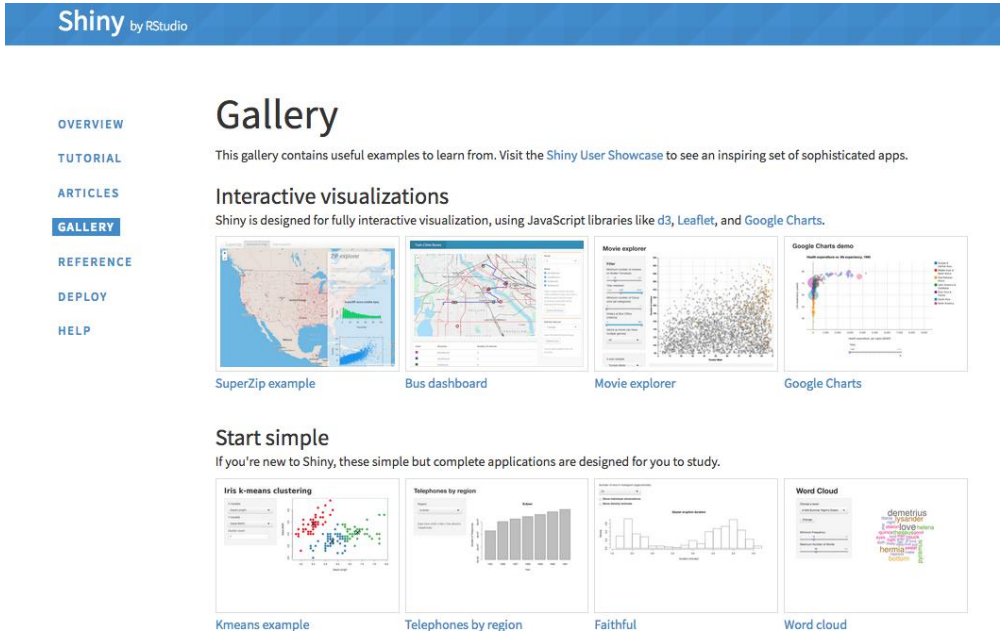
- http://textvis.lnu.se/
- https://github.com/mbostock/d3/wiki/Gallery



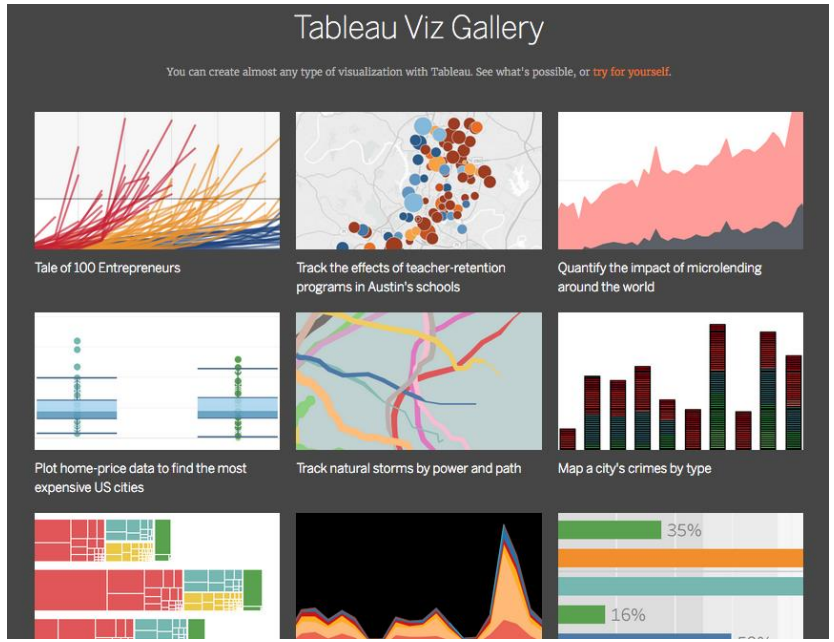


Many Design Alternatives! – Checkout Existing Visualizations

- http://shiny.rstudio.com/gallery/



- http://http://www.tableau.com/stories/gallery/



Available Datasets

- **VisArgue** data (available from instructors)
- **Project Gutenberg** Top 100 Books (available from instructors or from https://www.gutenberg.org/)
- Bulk Datasets from the **British Library**: https://data.bl.uk/
- Datasets from **NLTK** (see http://www.nltk.org/book/ch02.html)
- **Film Dialogue Data** https://github.com/matthewfdaniels/scripts/
- NRC **Emotion Lexicon** http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
- Public data extracted from the **Hathi Trust** (see https://goo.gl/ANOYuz)
- **WikiLeaks** data
- **Enron Emails** dataset
- **VisPubData** http://www.vispubdata.org/site/vispubdata/
- **Shakespeare Corpus** http://lexically.net/wordsmith/support/shakespeare.html
- Westbury Lab **Wikipedia Collection** http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

Questions, Comments, Feedback?