

G53FIV: Fundamentals of Information Visualization

Lecture 1: Introduction

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Contact Information

- Contact Information:
 - Name: Ke Zhou
 - E-mail: Ke.Zhou@nottingham.ac.uk
 - Office: B50
 - Office hour: Monday 14:00 – 15:00 PM
- Course Website:
 - Moodle: <https://moodle.nottingham.ac.uk/course/view.php?id=68644>
 - Personal homepage: <http://cs.nott.ac.uk/~pszkz/>

Overview

- Motivation for the Module
 - Background
 - Examples
- Module Information (both G53FIV and G53IVP)
 - Objective
 - Structure
 - Schedule

Online in 60 Seconds



Information Overload



The Key Challenge

- How to make use of the data
 - How do we avoid being overwhelmed?
 - How do we make sense of the data?
 - How do we harness this data in decision-making processes?



Objective

- Transform the data into information (understanding, insight) thus making it useful



What is Information Visualization?

- Definitions
 - “... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]
 - “The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]

The Best of Both Sides

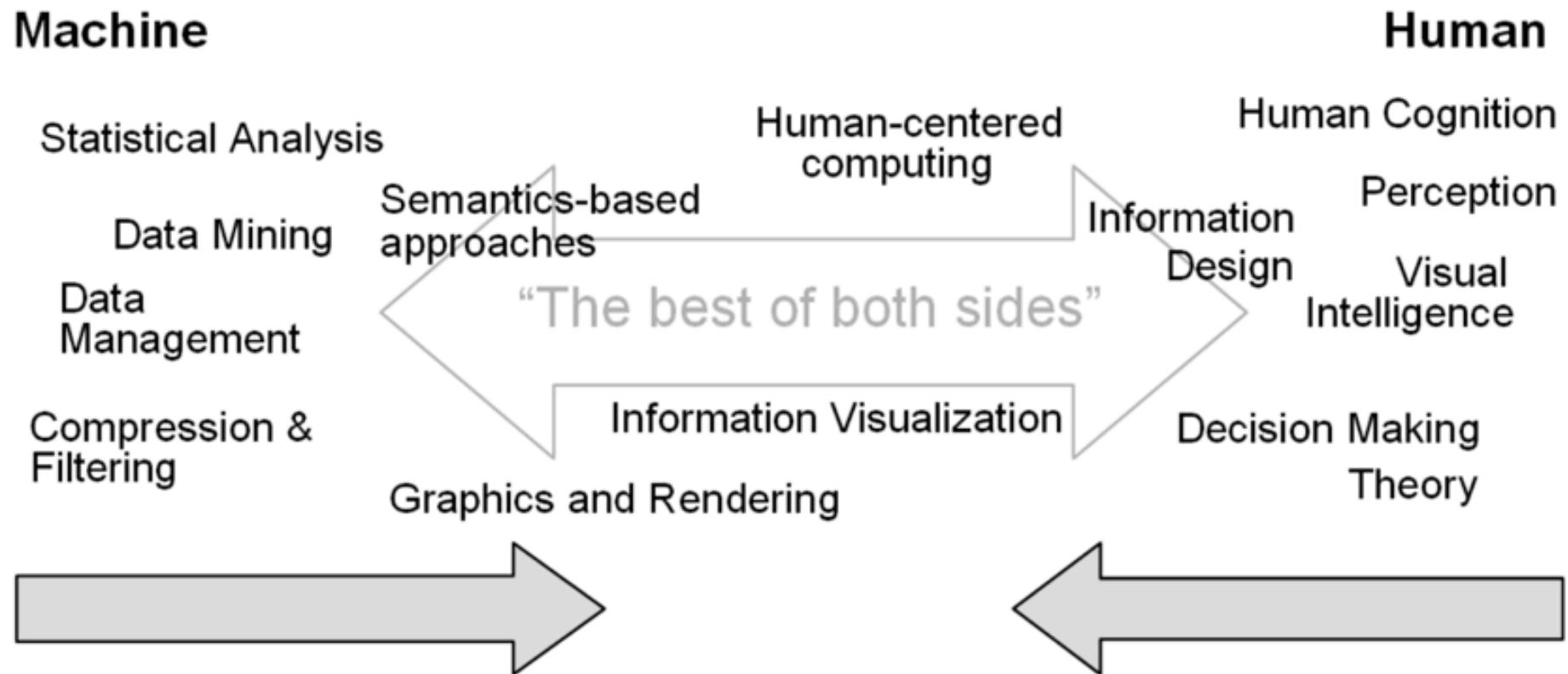


Fig. 2. Visual analytics integrates scientific disciplines to improve the division of labor between human and machine.

Anscombe's Quartet

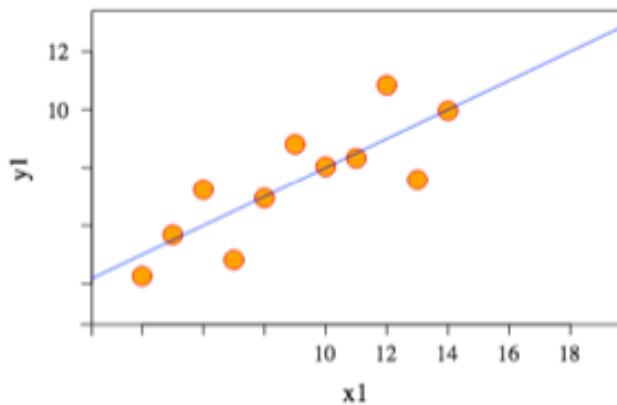
- Four data sets

	Set A		Set B		Set C		Set D	
	X	Y	X	Y	X	Y	X	Y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
corr	0.82		0.82		0.82		0.82	
lin. reg.	$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$	

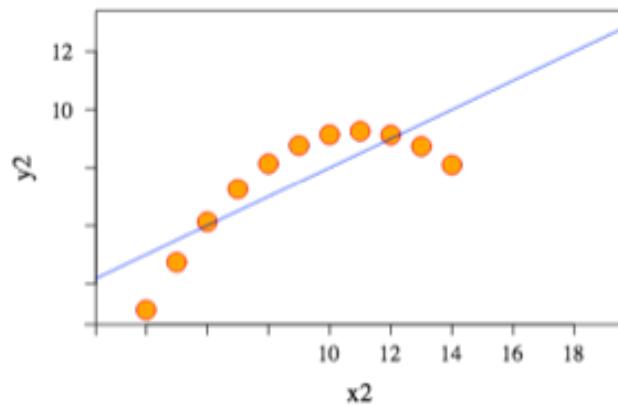
[Anscombe, 1973]

Visualization of Anscombe's Quartet

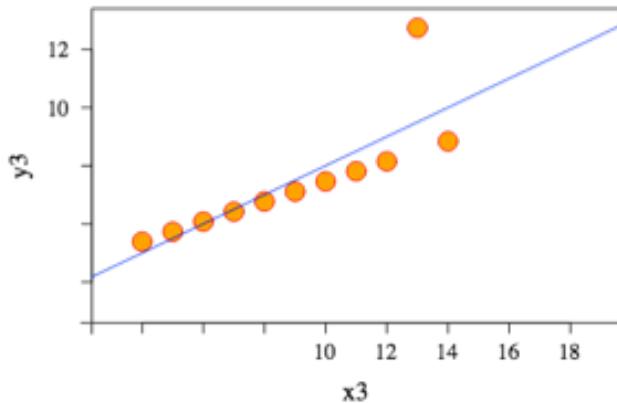
"what most people would see in their mind's eye [for a linear relationship with some unexplained variation]"



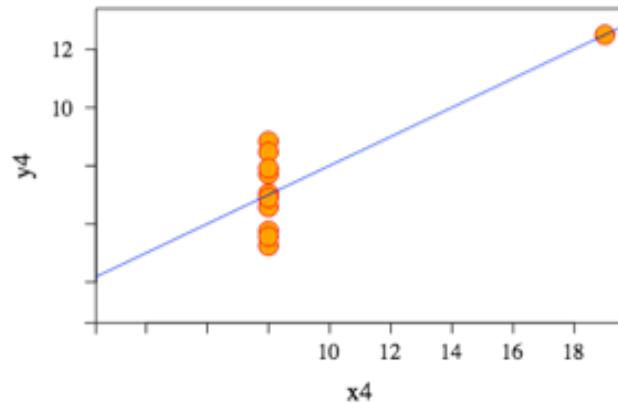
" y has a smooth curved relation with x , possibly quadratic, and there is little residual variability"



"all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation)"



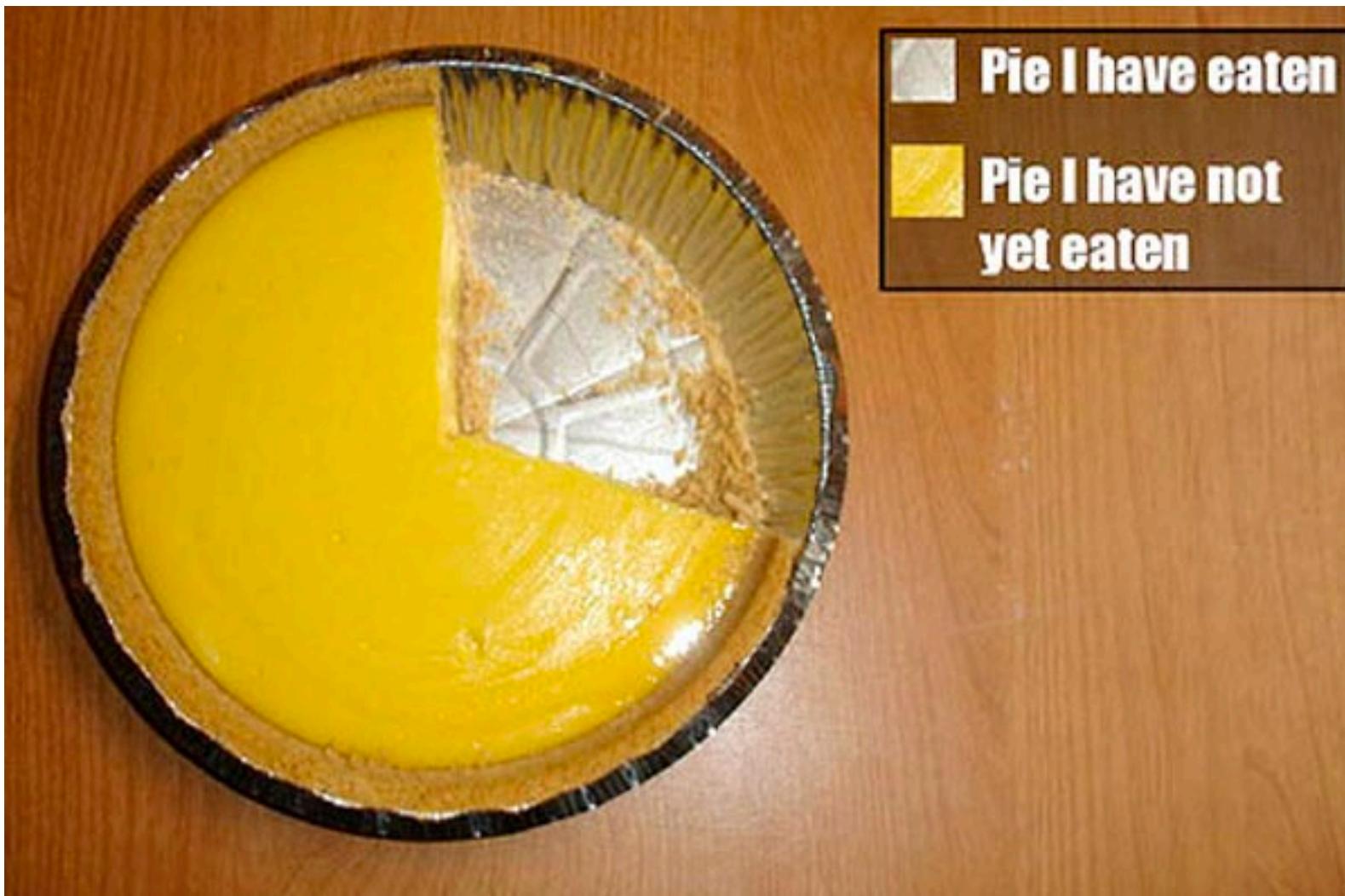
"all the information about the slope of the regression line resides in one observation"



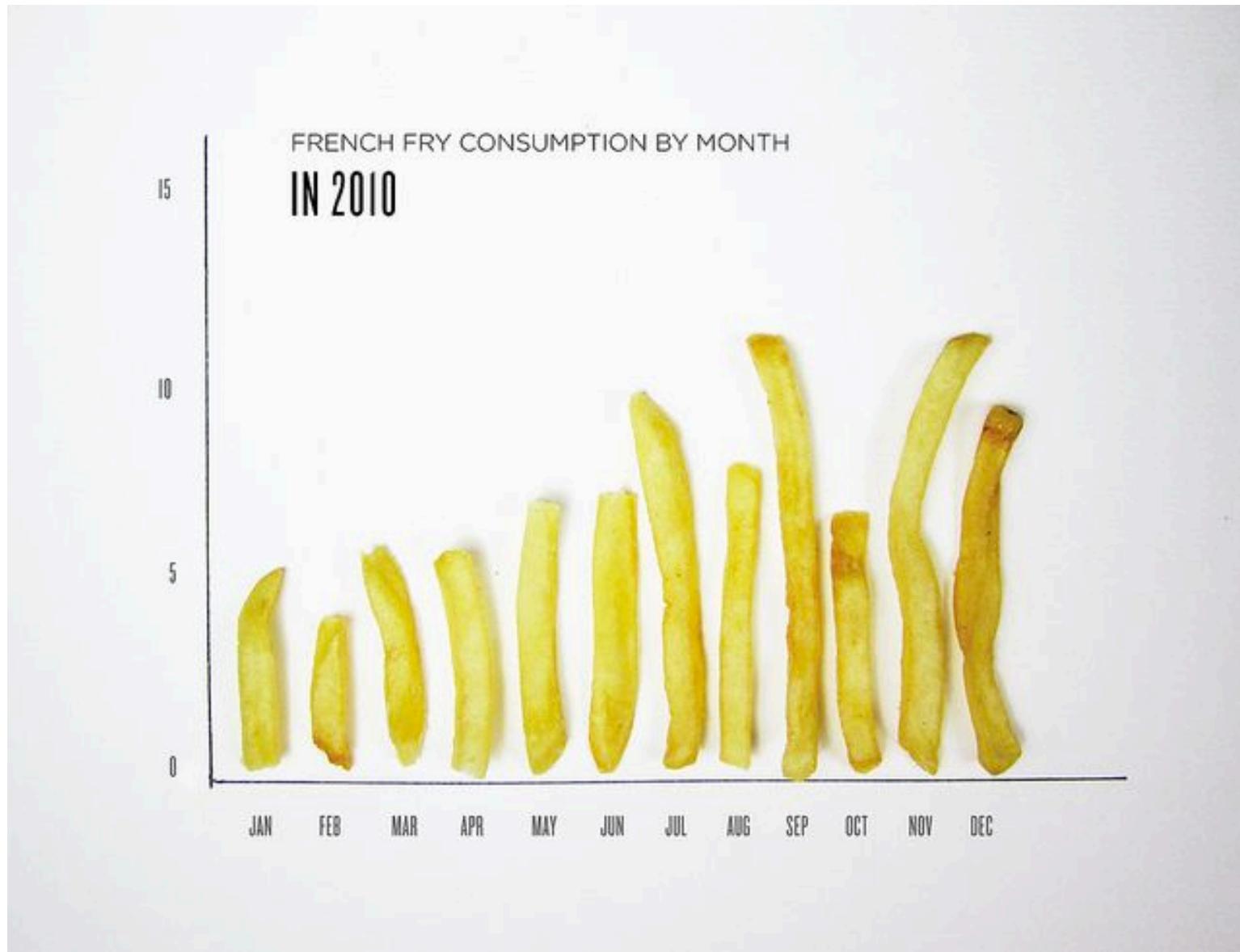
Common Information Visualizations

- What are the common information visualizations that you can think of?

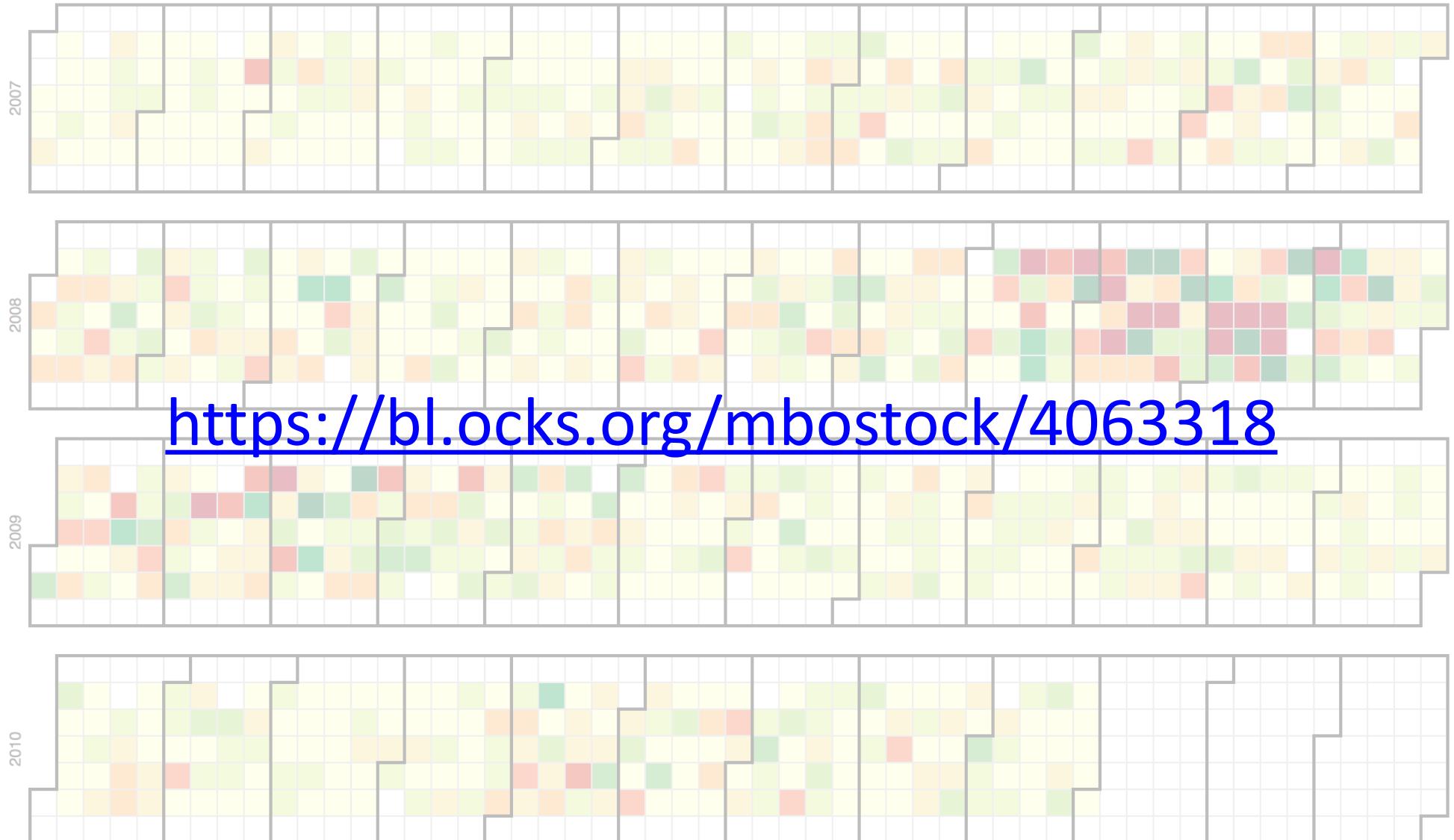
Pie Chart



Bar Chart



Calendar View



Wikipedia Page



WIKIPEDIA
The Free Encyclopedia

Article Talk

Read Edit View history

Search Wikipedia



Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

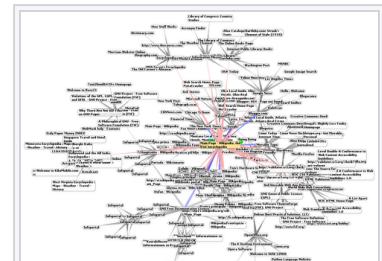
Information visualization

From Wikipedia, the free encyclopedia

Information visualization or **information visualisation** is the study of (interactive) visual representations of abstract data to reinforce human cognition. The abstract data include both numerical and non-numerical data, such as text and geographic information. However, information visualization differs from **scientific visualization**: "it's infovis [information visualization] when the spatial representation is chosen, and it's scivis [scientific visualization] when the spatial representation is given".^[1]

Contents [hide]

- [1 Overview](#)
- [2 History](#)
- [3 Specific methods and techniques](#)
- [4 Applications](#)
- [5 Organization](#)
- [6 See also](#)
- [7 References](#)
- [8 Further reading](#)
- [9 External links](#)



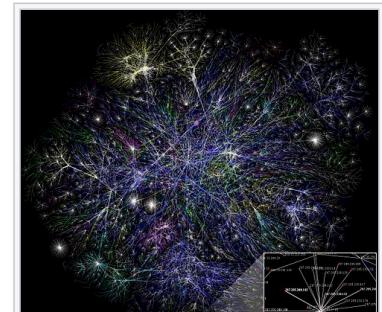
Graphic representation of a minute fraction of the [WWW](#), demonstrating hyperlinks

Overview [edit]

The field of information visualization has emerged "from research in [human-computer interaction](#), [computer science](#), [graphics](#), [visual design](#), [psychology](#), and [business methods](#). It is increasingly applied as a critical component in scientific research, [digital libraries](#), [data mining](#), financial data analysis, market studies, manufacturing [production control](#), and [drug discovery](#)".^[2]

Information visualization presumes that "visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once. Information visualization focused on the creation of approaches for conveying abstract information in intuitive ways."^[3]

Data analysis is an indispensable part of all applied research and problem solving in industry. The most fundamental data analysis approaches are visualization (histograms, scatter plots, surface plots, tree maps, parallel coordinate plots, etc.), [statistics](#) ([hypothesis test](#), [regression](#), [PCA](#), etc.), [data mining](#) ([association mining](#), etc.), and [machine learning](#) methods ([clustering](#), [classification](#), [decision trees](#), etc.).



Wikipedia Edit Evolution

Not logged in Talk Contributions Create account Log in

Article [Talk](#) [Read](#) [Edit](#) View history

Information visualization: Revision history

[View logs for this page](#)

Search for revisions

From year (and earlier): From month (and earlier): Tag filter: Show

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help>Edit summary](#).

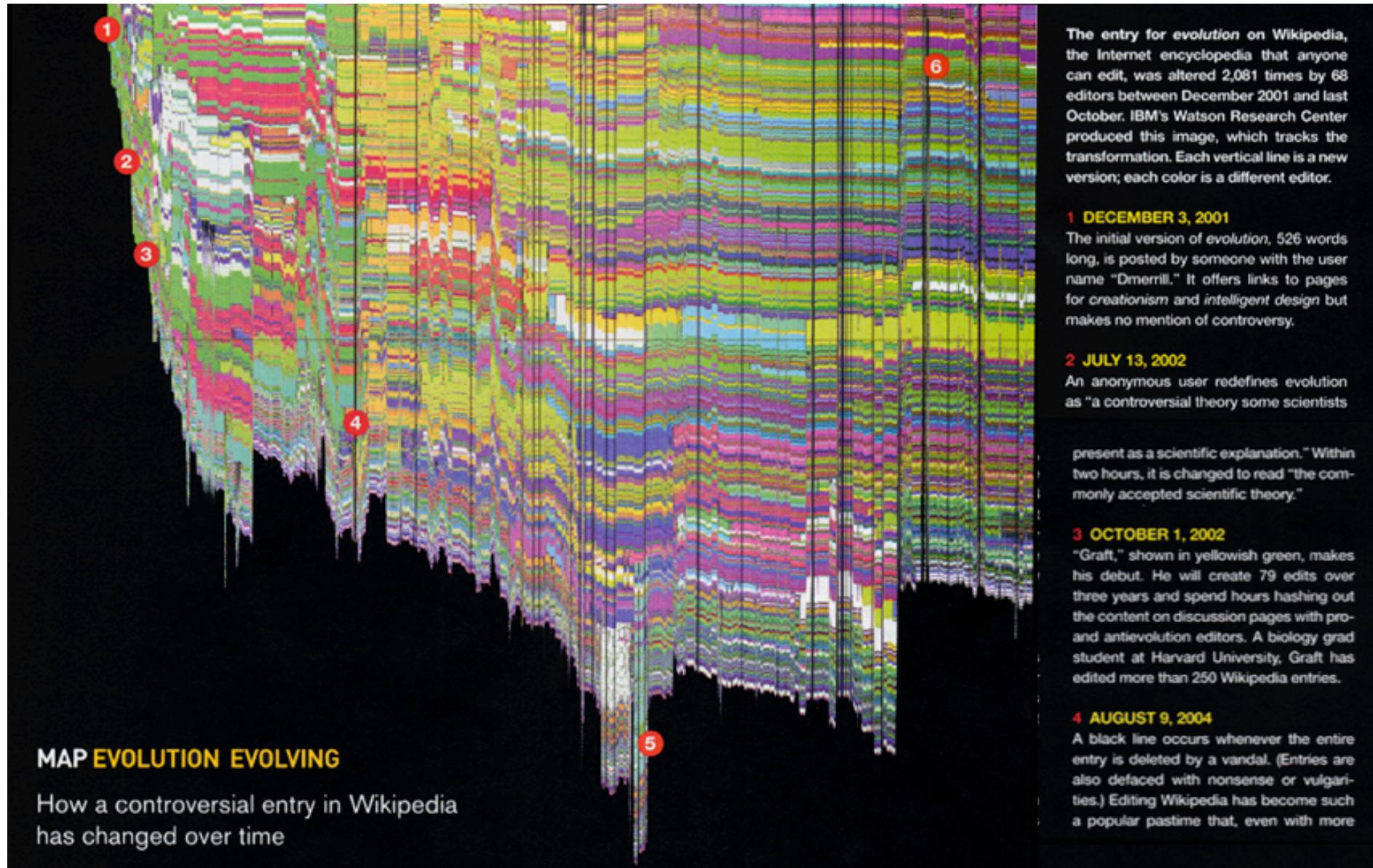
External tools: [Revision history statistics](#) · [Revision history search](#) · [Edits by user](#) · [Number of watchers](#) · [Page view statistics](#) · [Fix dead links](#)

(cur) = difference from current version, (prev) = difference from preceding version, m = minor edit, → = section edit, ← = automatic edit summary
[\(newest\)](#) [\(oldest\)](#) [View](#) ([newer 50](#) | [older 50](#)) ([20](#) | [50](#) | [100](#) | [250](#) | [500](#))

[Compare selected revisions](#)

- (cur | prev) 09:14, 19 December 2017 Billhpike (talk | contribs) . . (9,942 bytes) (-17) . . (→See also) ([undo](#))
- (cur | prev) 21:57, 13 November 2017 InternetArchiveBot (talk | contribs) . . (9,959 bytes) (+11) . . (Rescuing 1 sources and tagging 0 as dead. #IABot (v1.6.1)) ([undo](#))
- (cur | prev) 16:33, 5 November 2017 14.139.157.209 (talk) . . (9,848 bytes) (+16) . . (→Specific methods and techniques) ([undo](#))
- (cur | prev) 04:25, 1 September 2017 KolbertBot (talk | contribs) m . . (9,832 bytes) (+3) . . (Bot: [HTTP→HTTPS](#)) ([undo](#))
- (cur | prev) 23:51, 15 August 2017 Philafrenzy (talk | contribs) . . (9,829 bytes) (+4) . . (Link.) ([undo](#))
- (cur | prev) 17:49, 16 June 2017 Magic links bot (talk | contribs) m . . (9,825 bytes) (+8) . . (Replace [magic links](#) with templates per [local RfC](#) and [MediaWiki RfC](#)) ([undo](#))
- (cur | prev) 22:39, 5 June 2017 Blueclaw (talk | contribs) . . (9,817 bytes) (+694) . . (→Organization: reorganized section) ([undo](#))
- (cur | prev) 22:27, 5 June 2017 Blueclaw (talk | contribs) . . (9,123 bytes) (-2) . . (→Organization: removed defunct and irrelevant pages, added more organizations) ([undo](#))
- (cur | prev) 04:12, 27 May 2017 Dhjnavy (talk | contribs) m . . (9,125 bytes) (+39) . . (add this page into the category - [Information Visualization.](#)) ([undo](#))
- (cur | prev) 01:29, 10 February 2017 Omnipaedista (talk | contribs) . . (9,086 bytes) (+2) . . (→References) ([undo](#))
- (cur | prev) 20:50, 2 February 2017 178.4.223.168 (talk) . . (9,084 bytes) (+20) . . ([undo](#))
- (cur | prev) 20:40, 2 February 2017 178.4.223.168 (talk) . . (9,064 bytes) (+22) . . ([undo](#))
- (cur | prev) 14:58, 29 January 2017 MrOllie (talk | contribs) . . (9,042 bytes) (-387) . . (→External links: [WP:EL](#)) ([undo](#))
- (cur | prev) 14:02, 8 January 2017 45.247.69.160 (talk) . . (9,429 bytes) (+3) . . (→History) ([undo](#))
- (cur | prev) 15:58, 19 November 2016 174.118.59.81 (talk) . . (9,426 bytes) (-31) . . (→Specific methods and techniques) ([undo](#))

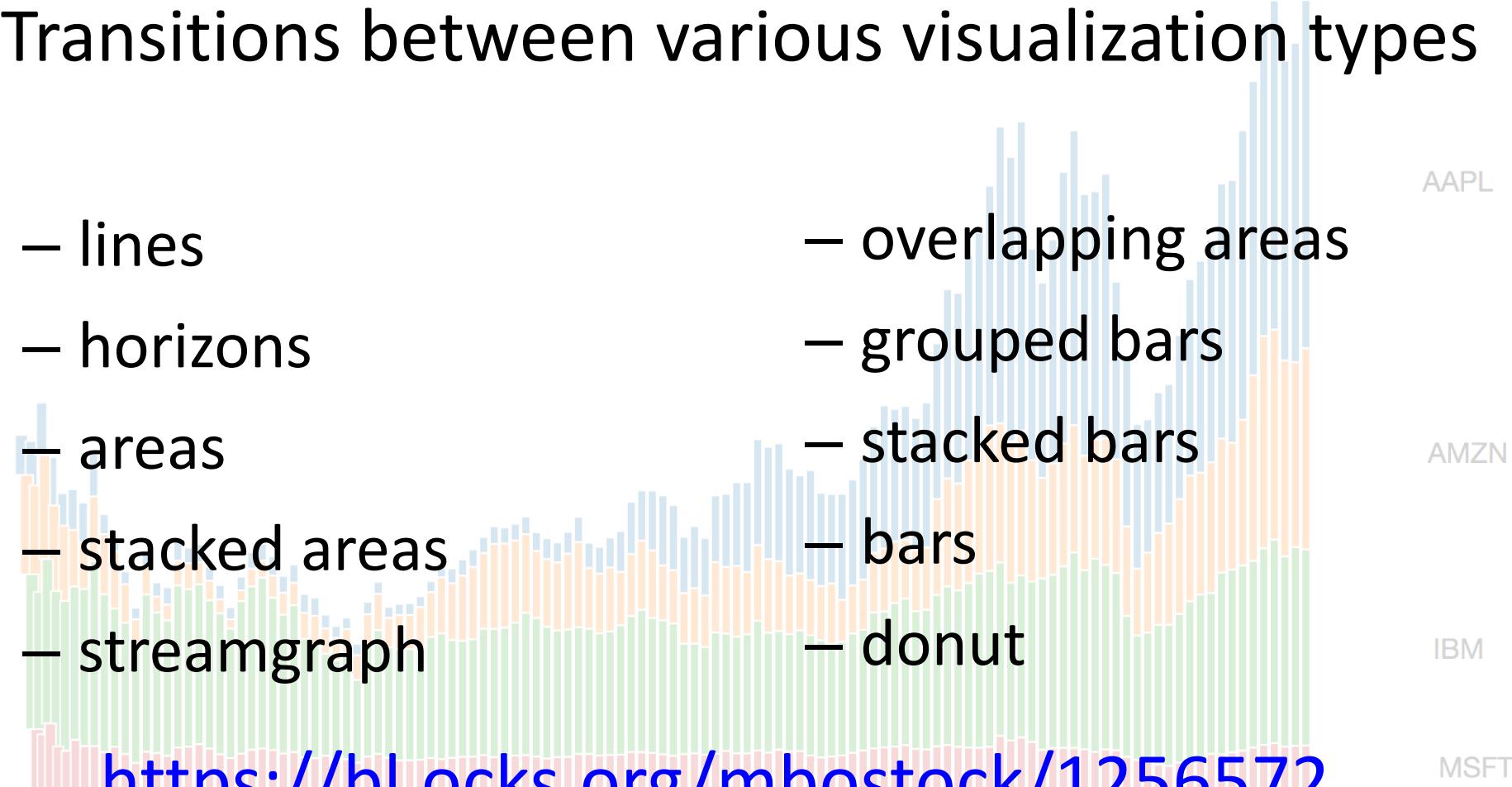
Visualizing Wikipedia Edit Evolution



D3's Show Reel

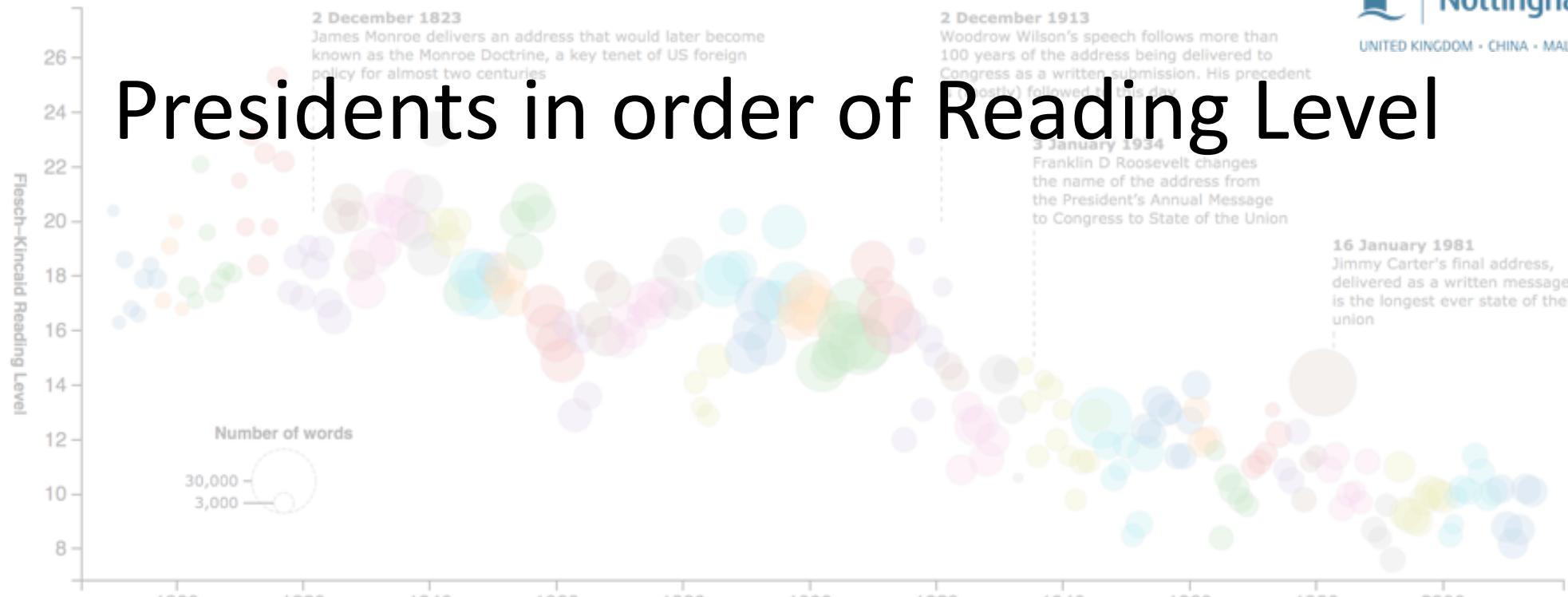
- Transitions between various visualization types

- lines
- horizons
- areas
- stacked areas
- streamgraph
- overlapping areas
- grouped bars
- stacked bars
- bars
- donut



<https://bl.ocks.org/mbostock/1256572>

Presidents in order of Reading Level



<https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level>



Information Visualization Examples



<https://d3js.org/>

Module Objectives

- Fundamental understanding on how visualizations convey information and how humans perceive
- Master an essential set of visualization techniques
- Practical experience in visualizing real-world data

Module Structure

- Lectures (8 weeks)
 - 2 x 1 hours / week
 - Monday 12:00 - 14:00
 - Business South A25, Jubilee Campus
- Assessment
 - 75% written examination
 - Contents from all lectures, core texts and paper handouts examinable.
 - 25% course work
 - Implementing a simple visualization
 - A written report, due on April 8, 2019

Contact Information

- Contact Information:
 - Name: Ke Zhou
 - E-mail: Ke.Zhou@nottingham.ac.uk
 - Office: B50
 - Office hour: Monday 14:00 – 15:00 PM
- Course Website:
 - Moodle: <https://moodle.nottingham.ac.uk/course/view.php?id=68644>
 - Personal homepage: <http://cs.nott.ac.uk/~pszkz/>

Course Materials

- Core text:
 - [The Visual Display of Quantitative Information](#) (2nd Edition). E. Tufte. Graphics Press, 2001 [available in the library].
 - [R Graphics Cookbook](#), Winston Chang, O'Reilly Media, 2013 [you can find it online by googling].
 - [Paper Handouts](#) (available on moodle in additional materials session per week)
- Other resources:
 - Moodle (Optional)

Lecture Schedule

Week	Topic (A25, Bus-South)	Topic (A25, Bus-South)	Lab (Optional, CS-A32)
1 (w19)	Introduction	The Value of Visualization	NONE
2 (w20)	Data and Image Models	Graphs and Charts	NONE
3 (w21)	Multivariate Data Visualization	Visualization with R	NONE
4 (w22)	Advanced R and Visualization Tools	Visual Perception	Course Work Case Study (Optional)
5 (w23)	Interaction	Evaluation	Lab (Optional)
6 (w24)	Text and Document	Time Series Data Visualization	Lab (Optional)
7 (w25)	Trees and Graphs	Recap of Fundamentals	Lab (Optional)
8 (w34)	Review	Demo	NONE

G53IVP

- <https://moodle.nottingham.ac.uk/course/view.php?id=68656>
- You will gain practical experience of how to **design, implement** and **evaluate** a distinctive interactive visualization which presents information gathered from a complex and interesting data source.
- Assessments
 - 80%: written report, documentation and code repositories
 - 20%: presentation

G53IVP

- Initial meeting
 - Feb 11th 15:00, 17:00 (third week of G53FIV)
 - B50, School of Computer Science
 - Discuss the general format and available resources
- Proposal development
 - Feb 25th 11:00 (fifth week of G53FIV)
- Report, documentation, code due
 - May 10th 11:00
- Presentation
 - May 13th 10:00

Break

- Next:
- Topic: The Value of Visualization



G53FIV: Fundamentals of Information Visualization

Lecture 2: The Value of Visualization

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Overview

- What are the Key Values of IV?
 - Record
 - Communicate
 - Reason
- What is the Process of IV?

What are the Key Values of IV?

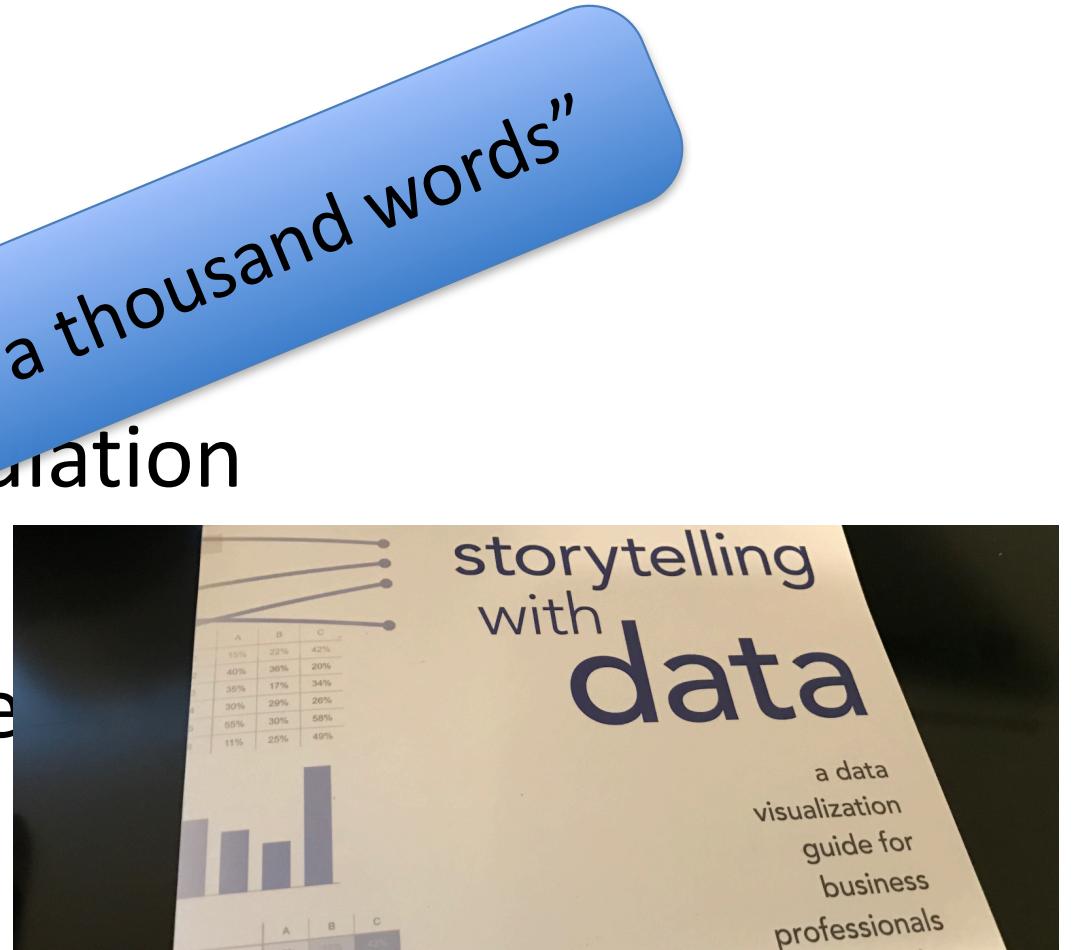
Why Create Visualization?

Why Create Visualization?

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present argument or tell a story
- Inspire

Why Create Visualization?

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support argument or persuasion
- Find patterns “A picture is worth a thousand words”
- Present argument or tell story
- Inspire



Summary of Reasons

- **Record** information
 - Blueprints, photographs, seismographs, ...
- **Communicate** information to others
 - Share and persuade
 - Collaborate and revise
- Analyze data to **support reasoning**
 - Find patterns / Discover errors in data
 - Expand memory
 - Develop and assess hypotheses

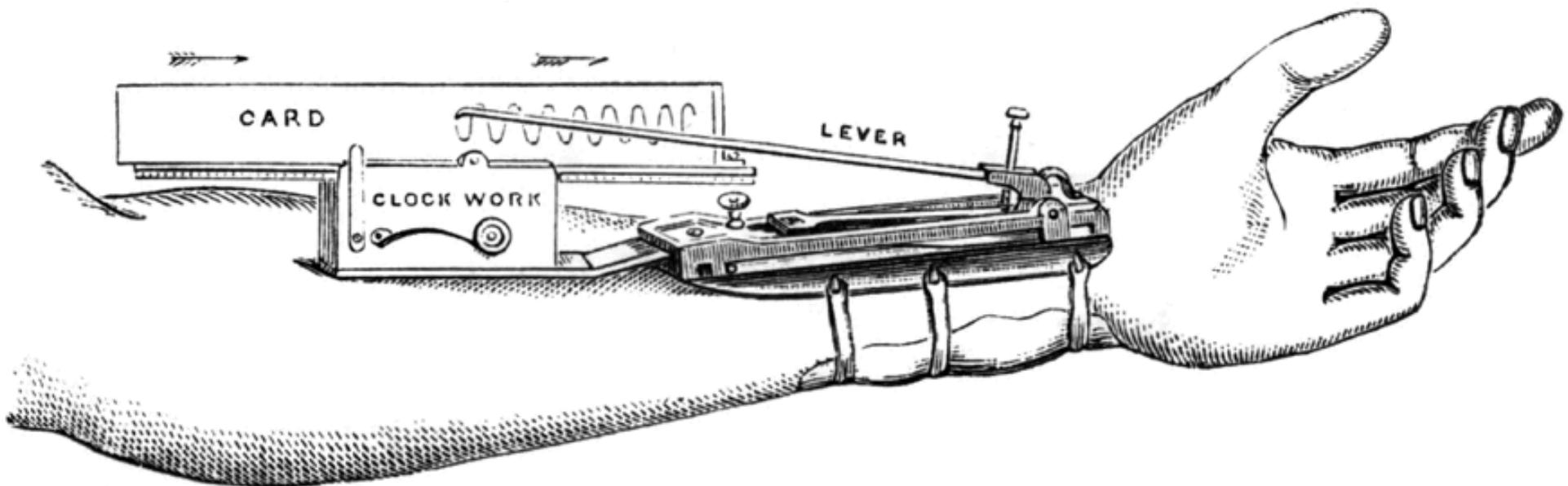
Record Information

- Egyptian hieroglyphs



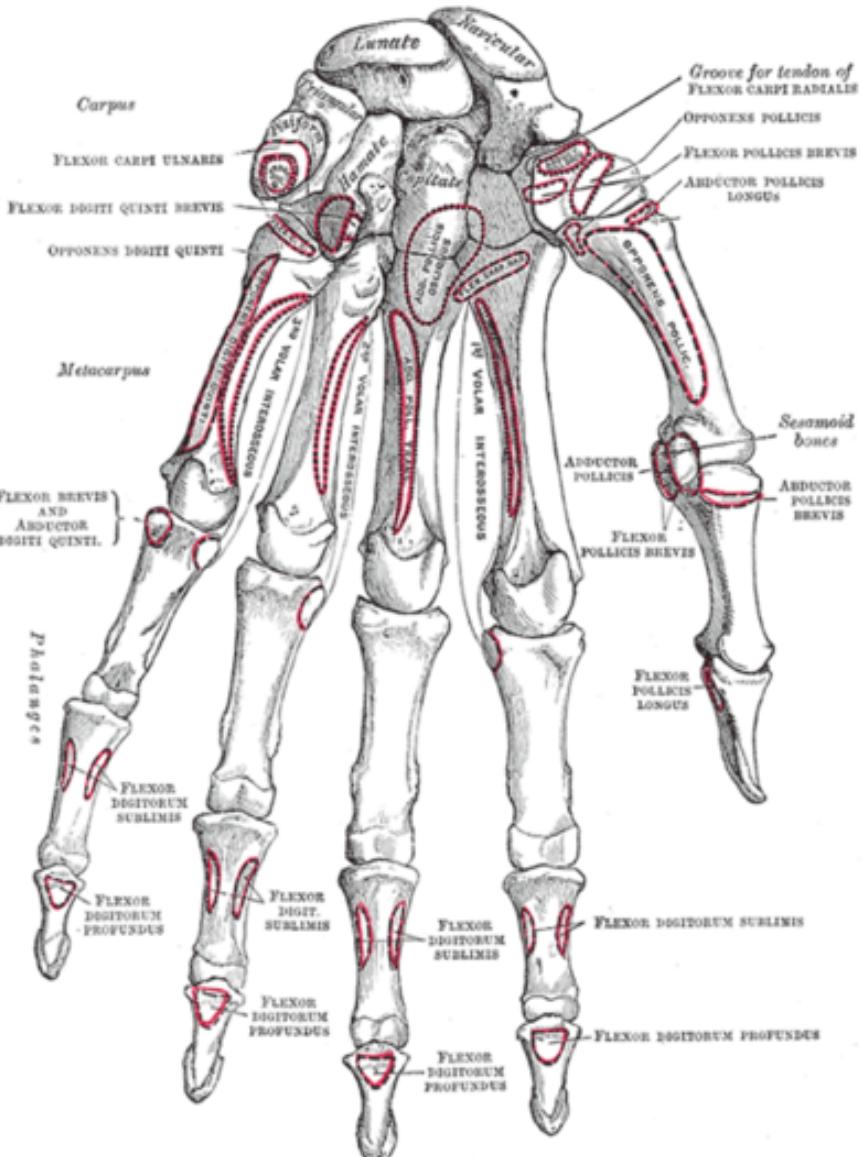
Record Information

- E.J. Marey's sphygmograph (1854)
 - an instrument which produces a line recording the strength and rate of a person's pulse.

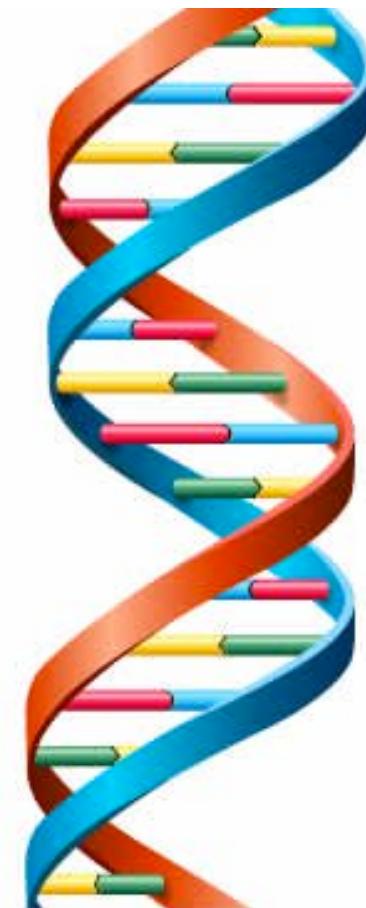


Communicate: convey information to others

Share and collaborate



Bones in hand (drawn in 1918)



DNA Helix

Persuade: Nightingale's Graph

2.
APRIL 1855 to MARCH 1856.

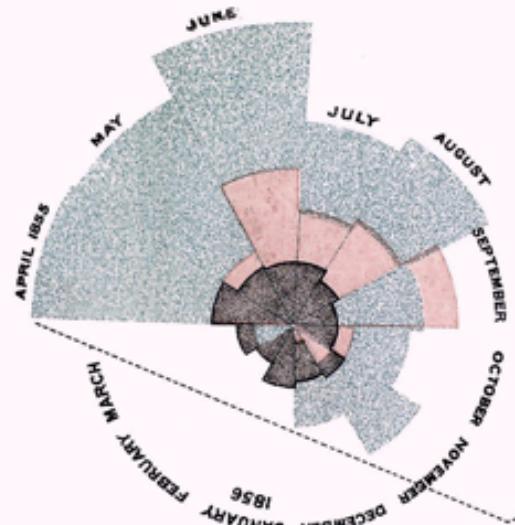
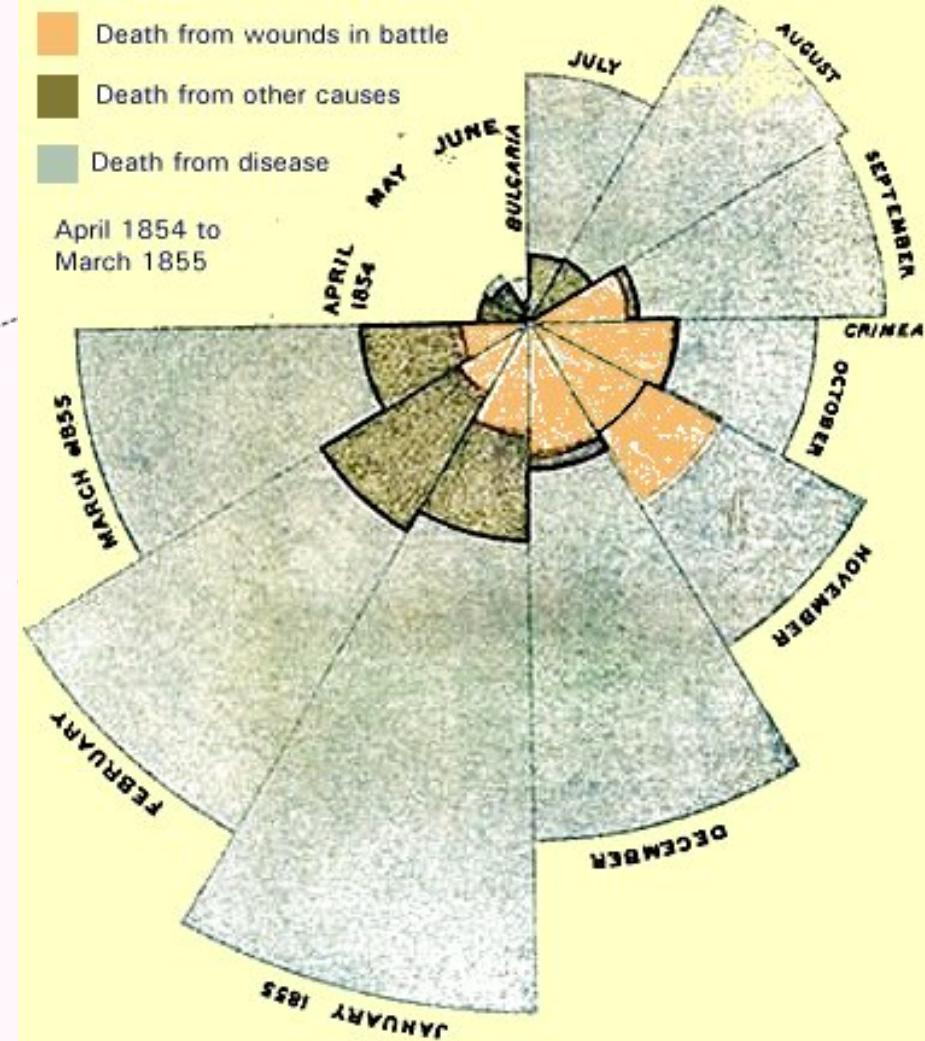


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

1.
APRIL 1854 to MARCH 1855.



The areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

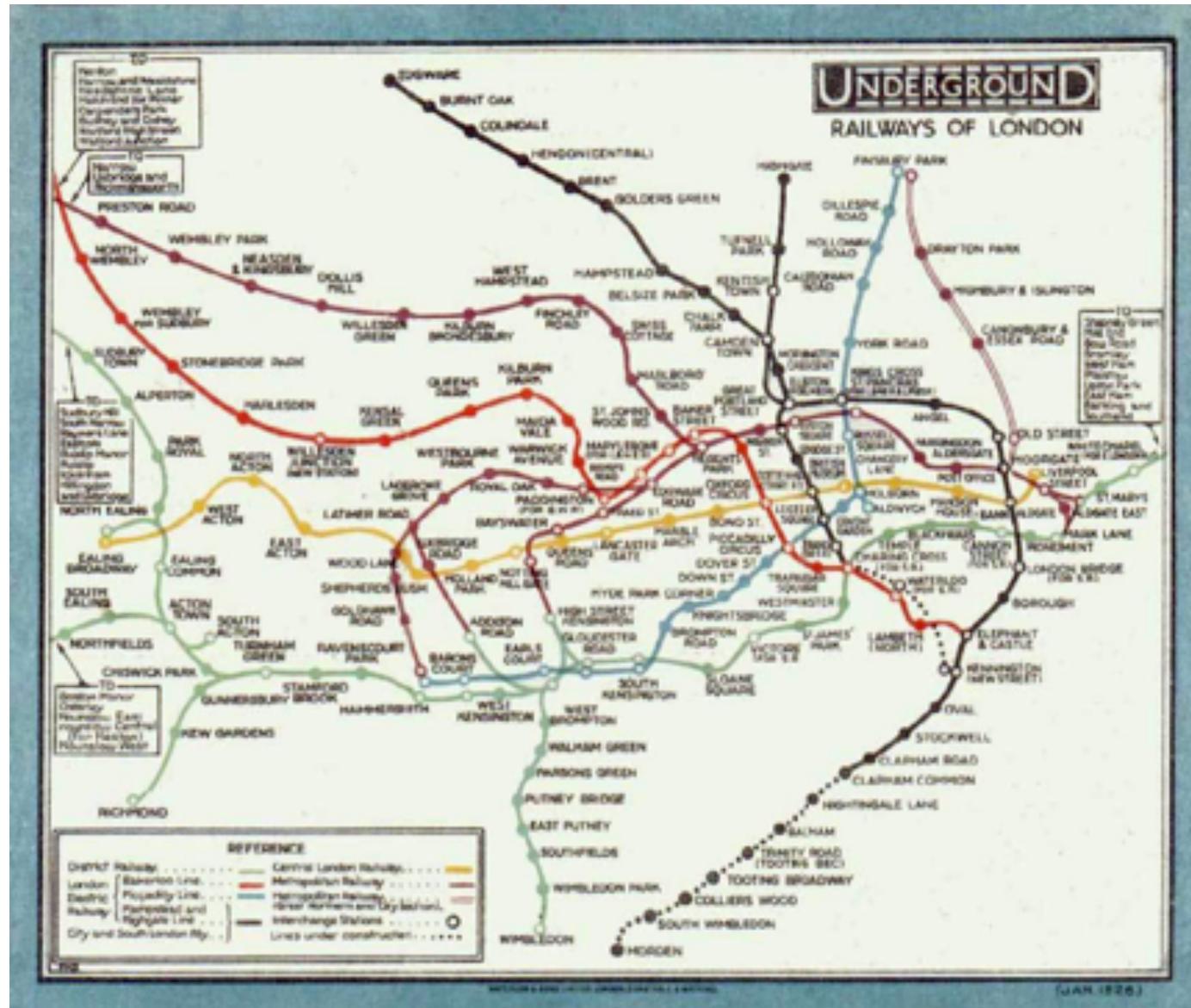
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Clarify/Revise: London's underground map

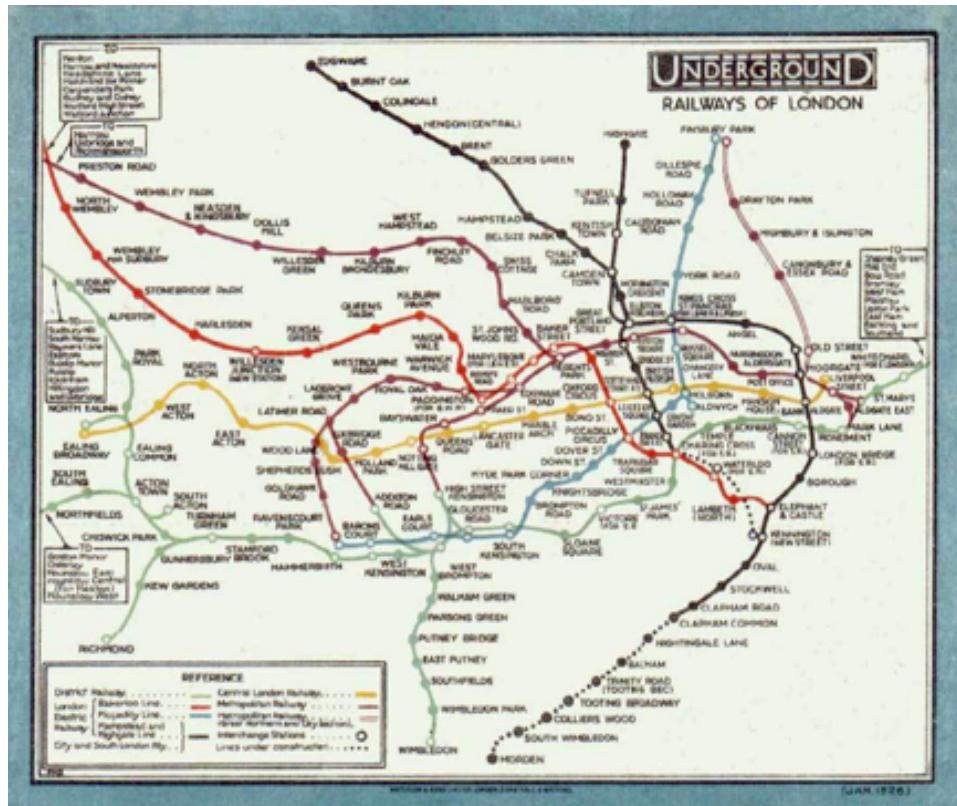


1926

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Clarify/Revise: London's underground map

- Horizontal, vertical and 45° segments
- Key insight: topology and relative location of stations



1926

vs.



1987

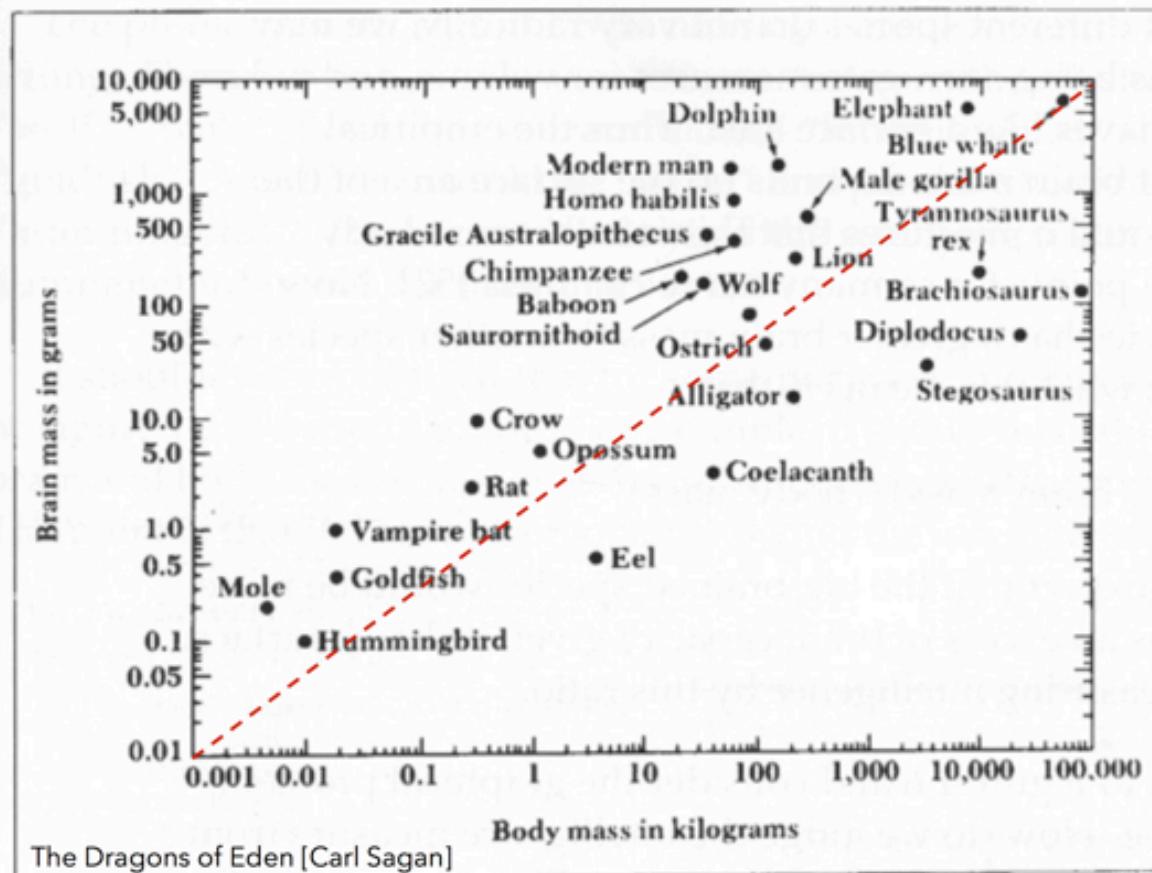
Beer Infographics



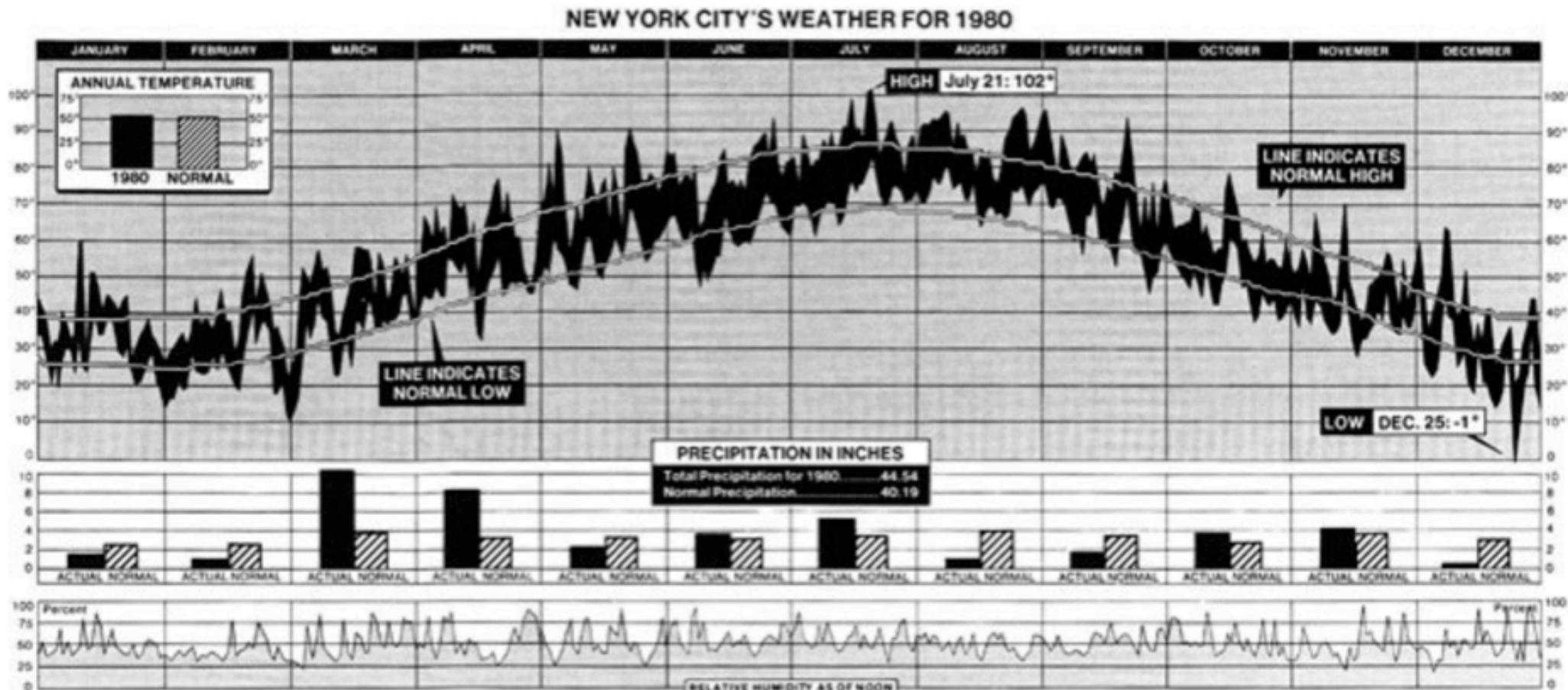
Support reasoning

Find Patterns: the most powerful brain

ID	Name	Body Weight	Brain Weight
1	Lesser Short-tailed Shrew	5	0.14
2	Little Brown Bat	10	0.25
3	Mouse	23	0.3
4	Big Brown Bat	23	0.4
5	Musk Shrew	48	0.33
6	Star Nosed Mole	60	1
7	Eastern American Mole	75	1.2
8	Ground Squirrel	101	4
9	Tree Shrew	104	2.5
10	Golden Hamster	120	1
11	Mole Rat	122	3
12	Galago	200	5
13	Rat	280	1.9
14	Chinchilla	425	6.4
15	Desert Hedgehog	550	2.4
16	Rock Hyrax (a)	750	12.3
17	European Hedgehog	785	3.5
18	Tenrec	900	2.6
19	Arctic Ground Squirrel	920	5.7
20	African Giant Pouched Rat	1000	6.6
21	Guinea Pig	1040	5.5
22	Mountain Beaver	1350	8.1
23	Slow Loris	1400	12.5
24	Genet	1410	17.5
25	Phalanger	1620	11.4



Find Patterns: NYC weather



2200 data points

Expand Memory

- Class Exercise

34

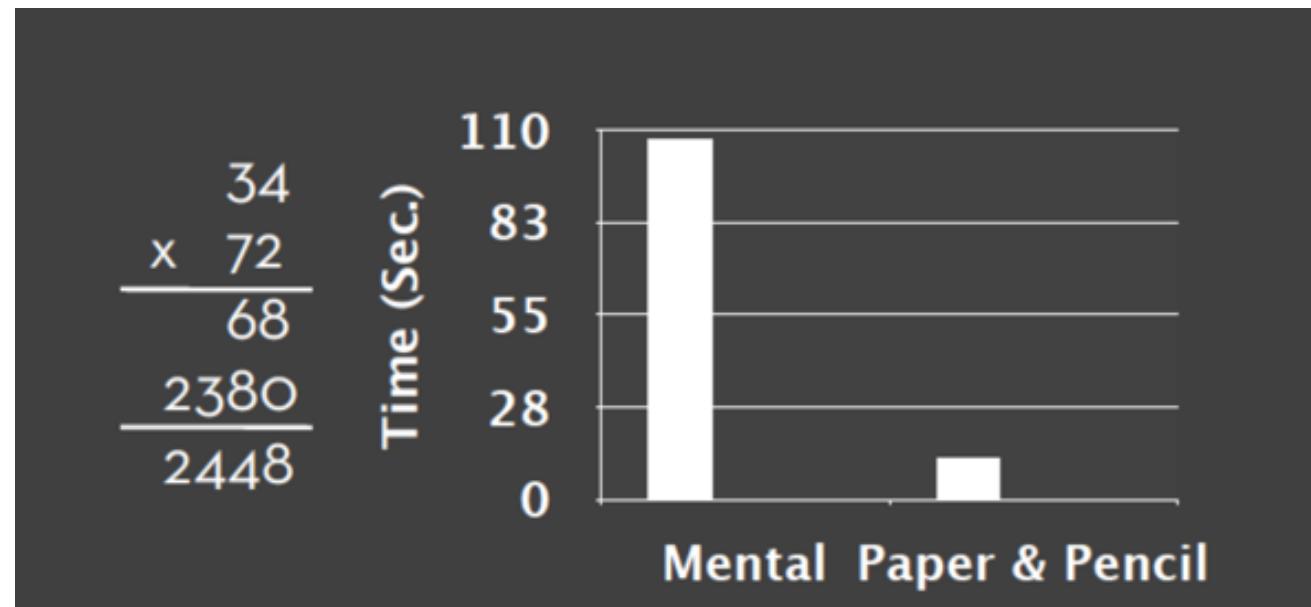
x 72

Expand Memory

- Class Exercise

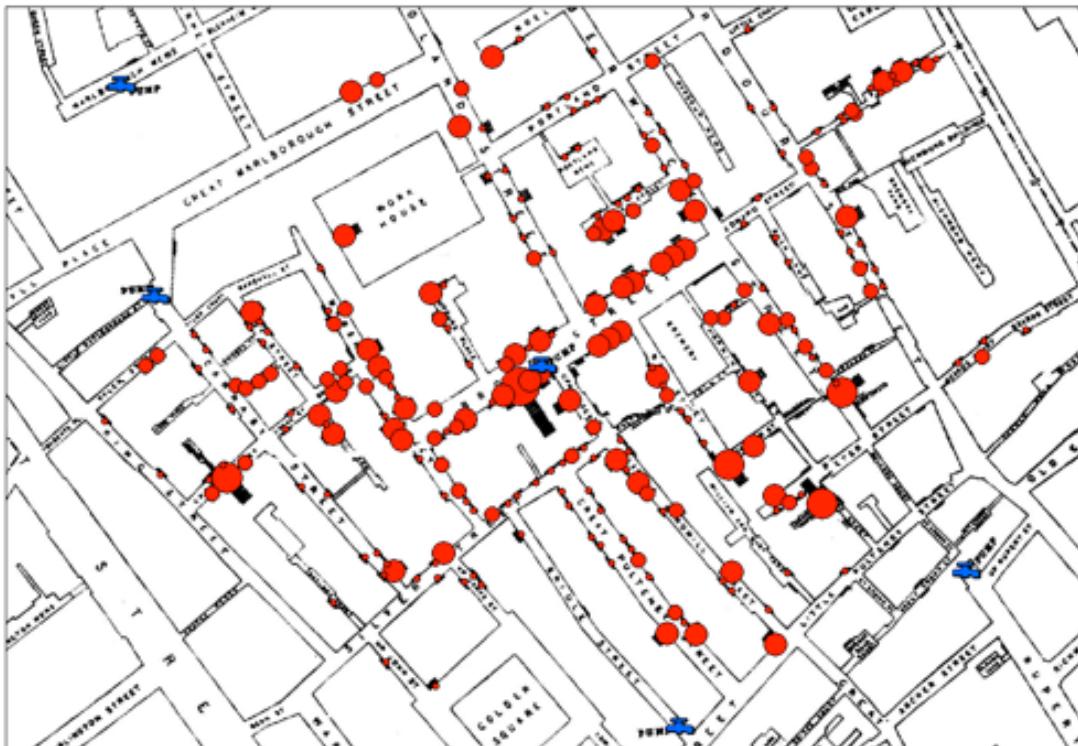
$$\begin{array}{r} 34 \\ \times 72 \\ \hline \end{array}$$

$$\begin{array}{r} 34 \\ \times 72 \\ \hline 68 \\ 2380 \\ \hline 2448 \end{array}$$



Develop and Assess Hypothesis

London Cholera Map



London Cholera Map
Visualization by John Snow, 1854.

- The closer to the Broad Street water pump, the greater the number of deaths.
- The information helped convince the public a true sewage system was needed.

Surprises in Data

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

John Tukey, 1977

“Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data...”

W. Cleveland
The Elements of Graphing Data

Reasoning / Exploration

“If you can articulate very precisely what you’re seeking, visualization likely isn’t your best approach”

J. Stasko, EuroVis’14

Exploration

- Don’t know what you’re looking for
- Don’t have a priori questions
- Want to know what questions to ask

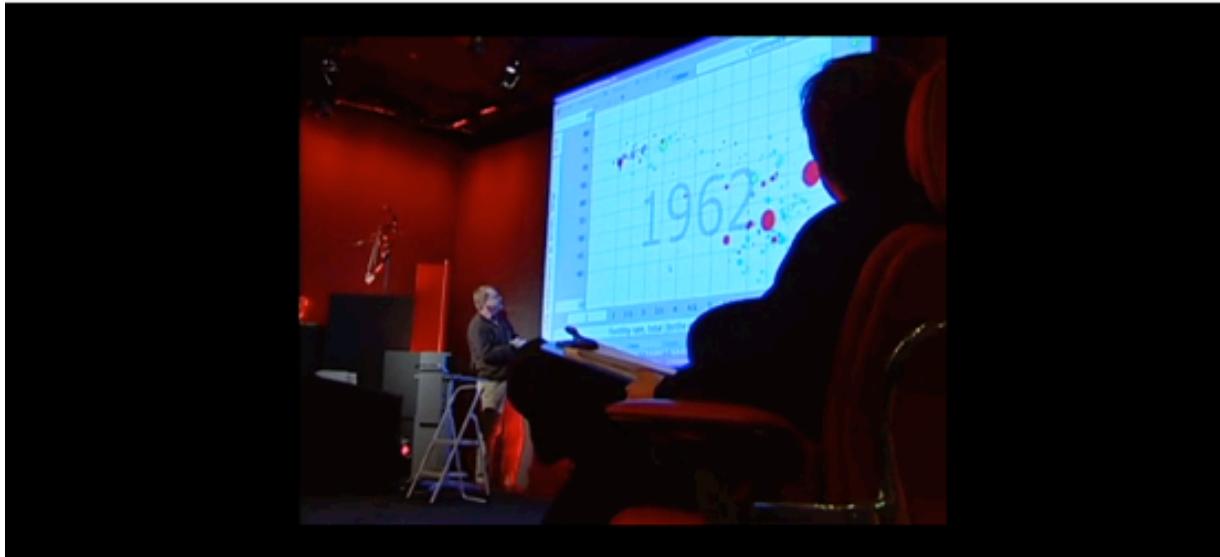


Hans Rosling's TED talk



TED

LOG IN



Hans Rosling:

The best stats you've ever seen

TED2006 · 19:50 · Filmed Feb 2006

 48 subtitle languages •

View interactive transcript



List



Download



Rate



Link



Share

11,146,687 Total views

http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

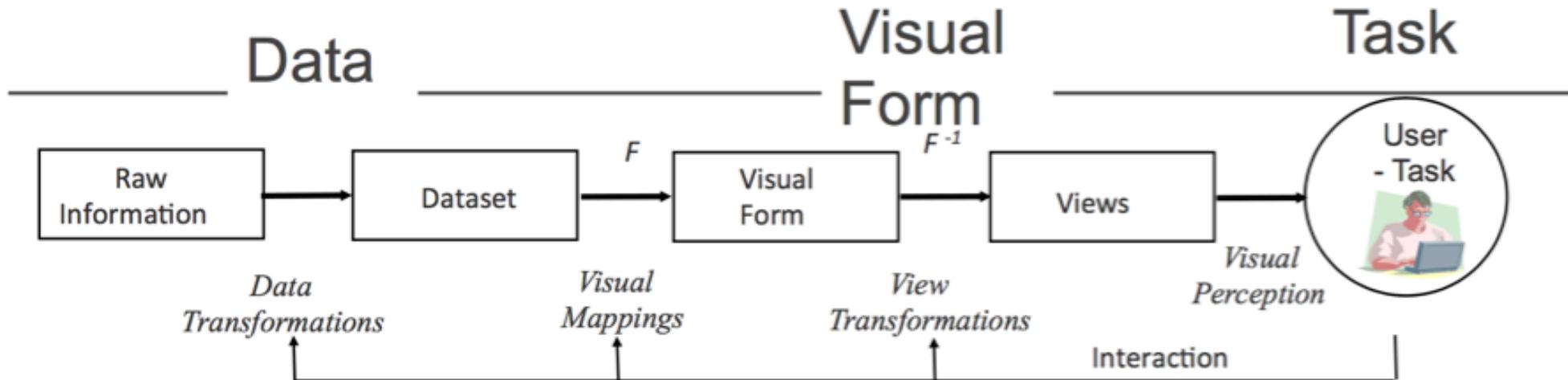
Key Applications of IV

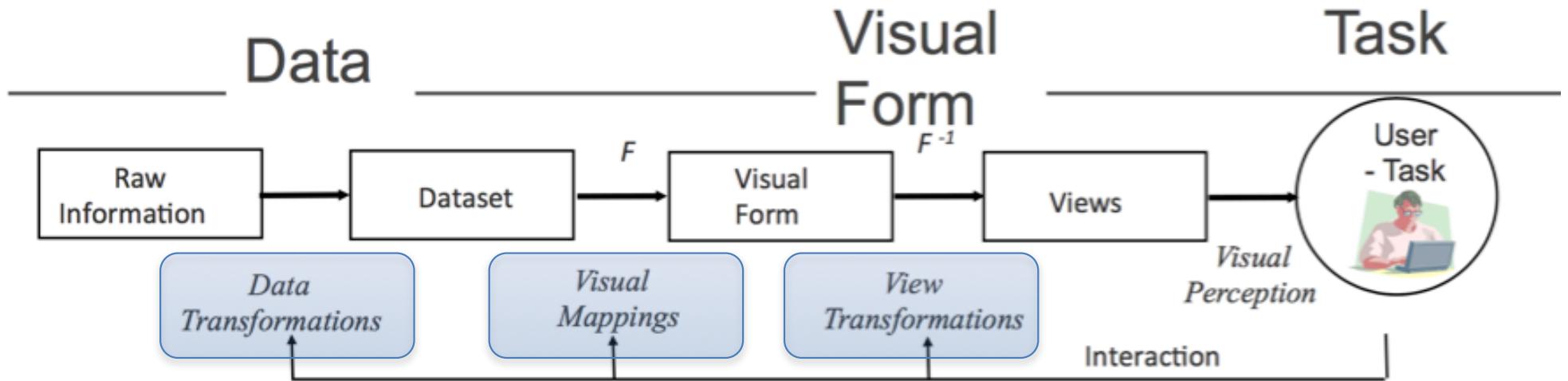
- I. Record Information
- II. Communications (Presentation)
 - Communicate data and ideas
 - Explain and inform
 - Provide evidence and support
 - Influence and persuade
- III. Reasoning (Analysis)
 - Explore the data
 - Assess a situation
 - Determine how to proceed
 - Decide what to do



The Visualization Process

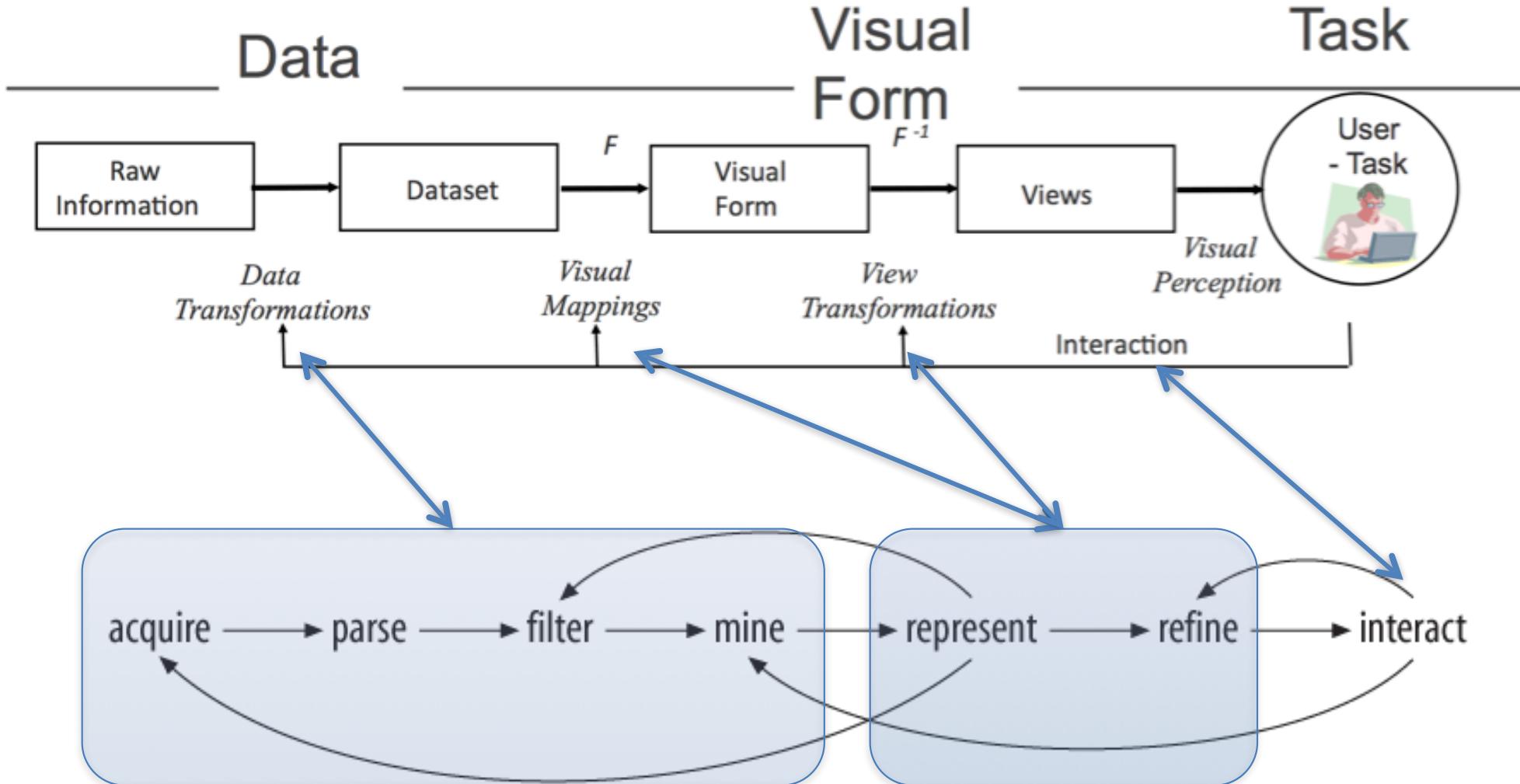
Different Stages of Visualization



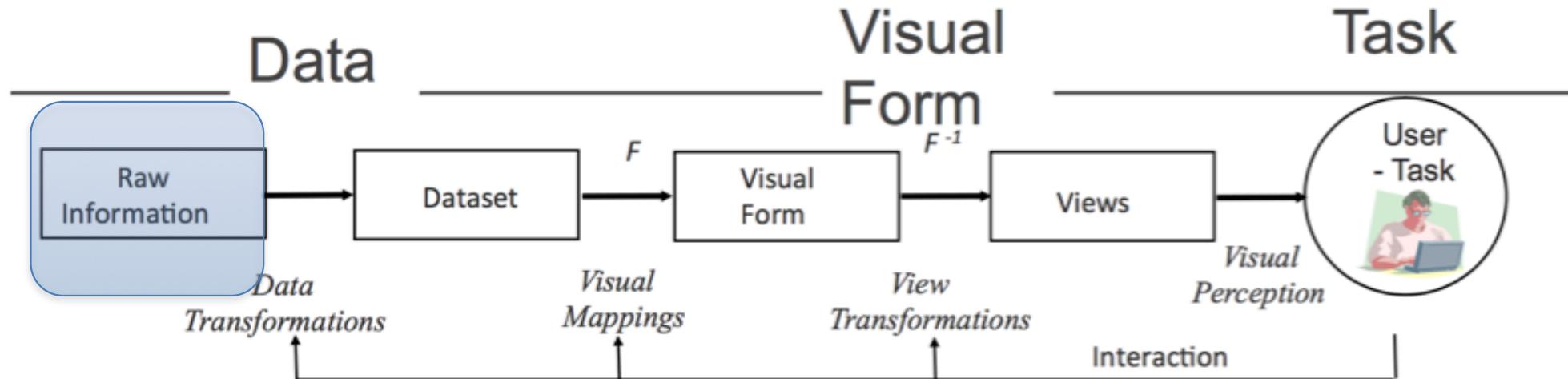


- Data transformation
 - create a structural model (schema), mapping raw data into data tables
- Visual mapping
 - create a visual spatial model, transforming data tables into visual structures
- View Transformations
 - Create views of the Visual Structures by specifying graphical parameters such as position, scaling, and clipping

Different Stages of Visualization



Seven Stages: Acquire

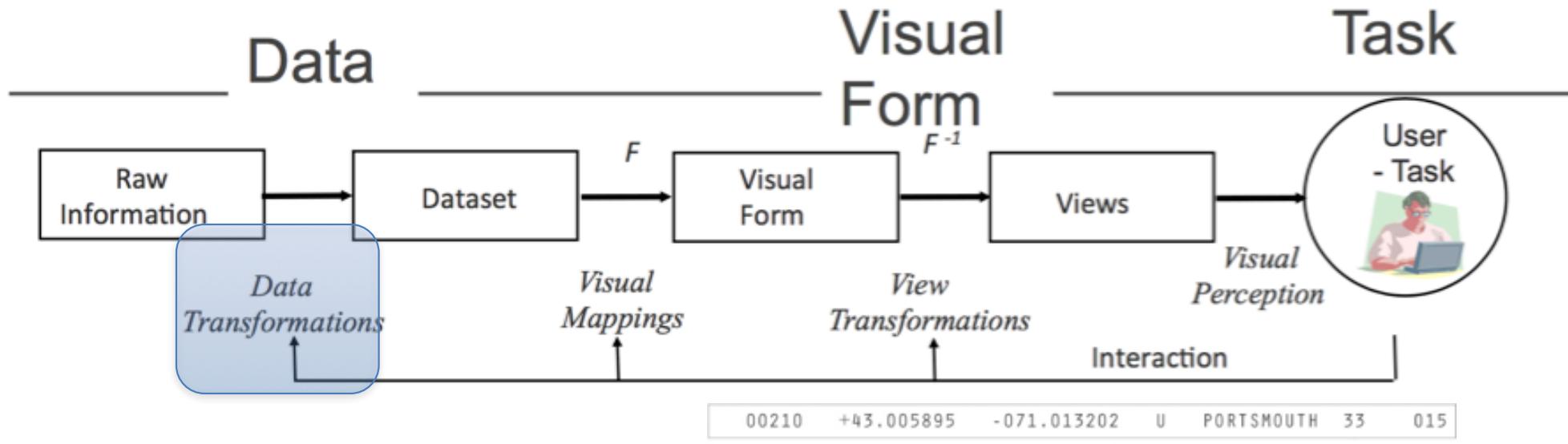


- Obtain the data, whether from a file on a disk or a source over a network

00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00544	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00601	+18.165273	-066.722583		ADJUNTAS	72	001
00602	+18.393103	-067.180953		AGUADA	72	003
00603	+18.455913	-067.145780		AGUADILLA	72	005
00604	+18.493520	-067.135883		AGUADILLA	72	005
00605	+18.465162	-067.141486	P	AGUADILLA	72	005
00606	+18.172947	-066.944111		MARICAO	72	093
00610	+18.288685	-067.139696		ANASCO	72	011

Zip codes in the format provided by the U.S. Census Bureau

Seven Stages: Parse

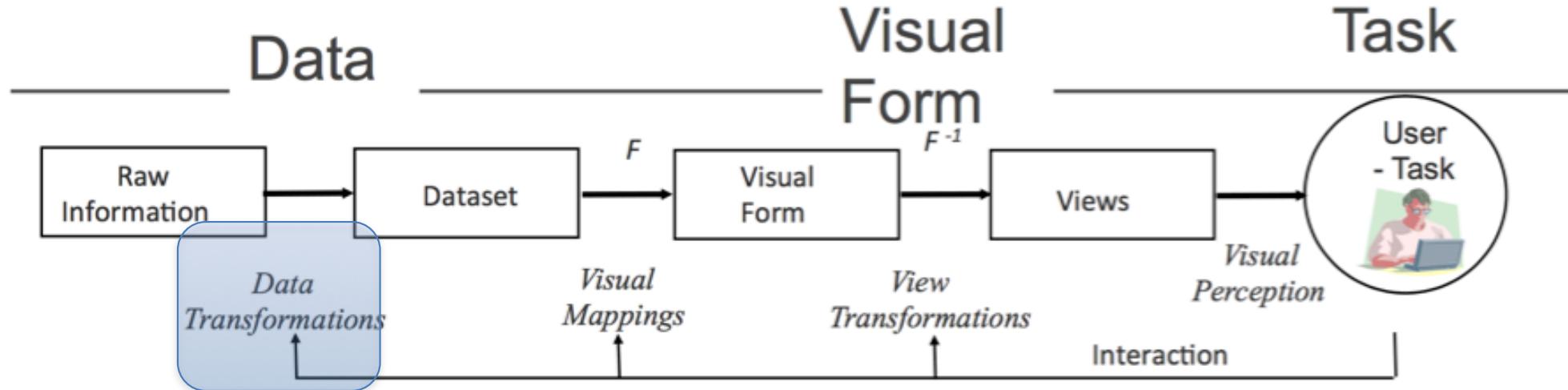


- Provide some structure for the data's meaning, and order it into categories.

01	ALABAMA	AL
02	ALASKA	AK
04	ARIZONA	AZ
05	ARKANSAS	AR
06	CALIFORNIA	CA
08	COLORADO	CO
09	CONNECTICUT	CT
10	DELAWARE	DE
12	FLORIDA	FL
13	GEORGIA	GA
15	HAWAII	HI
16	IDAHO	ID
17	ILLINOIS	IL
18	INDIANA	IN
19	IOWA	IA
20	KANSAS	KS

Structure of acquired data, formatted as a data type that we'll handle in a conversion program

Seven Stages: Filter

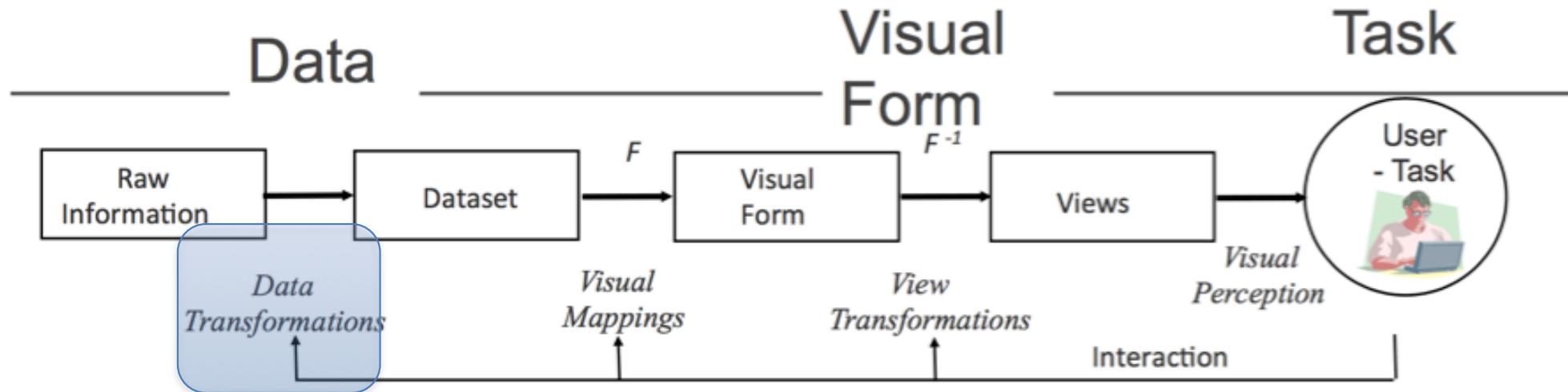


- Remove all but the data of interest.

00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00544	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00601	+18.165273	-066.722583		ADJUNTAS	72	001
00602	+18.393103	-067.180953		AGUADA	72	003
00603	+18.455913	-067.145780		AGUADILLA	72	005
00604	+18.493520	-067.135883		AGUADILLA	72	005
00605	+18.465162	-067.141486	P	AGUADILLA	72	005
00606	+18.172947	-066.944111		MARICAO	72	093
00610	+18.288685	-067.139696		ANASCO	72	011

Filter out some data points
remain only some data fields

Seven Stages: Mine



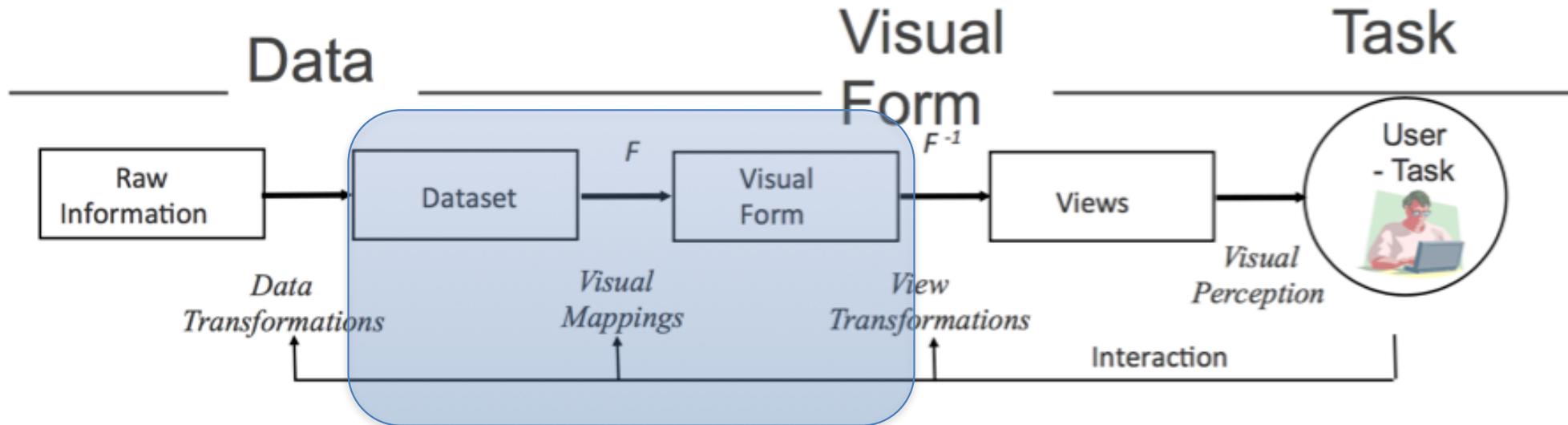
- Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.

00213	43.005895	-71.013202	PORTSMOUTH	NH
00214	43.005895	-71.013202	PORTSMOUTH	NH
00215	43.005895	-71.013202	PORTSMOUTH	NH
00501	40.922326	-72.637078	HOLTSVILLE	NY
00544	40.922326	-72.637078	HOLTSVILLE	NY
+	+	+	+	+
+	+	+	+	+
+	+	+	+	+

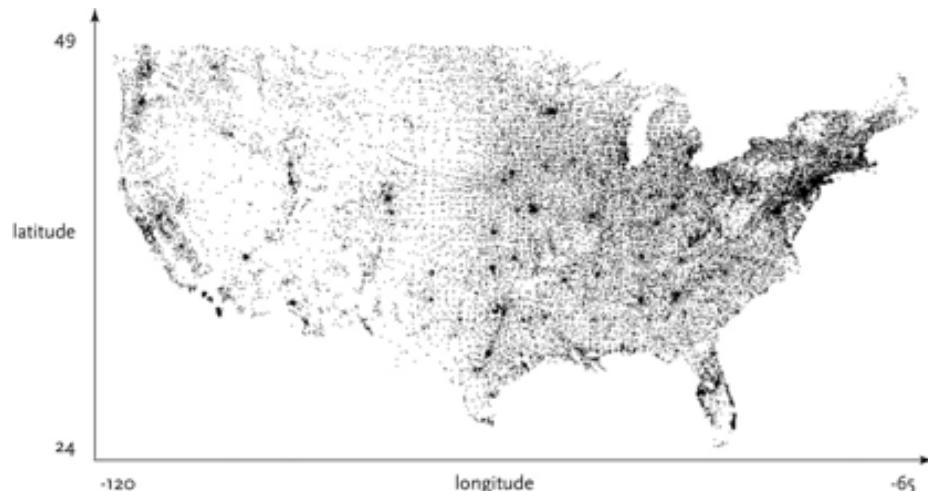
↓ ↓

min 24.655691	min -124.62608
max 48.987385	max -67.040764

Seven Stages: Represent

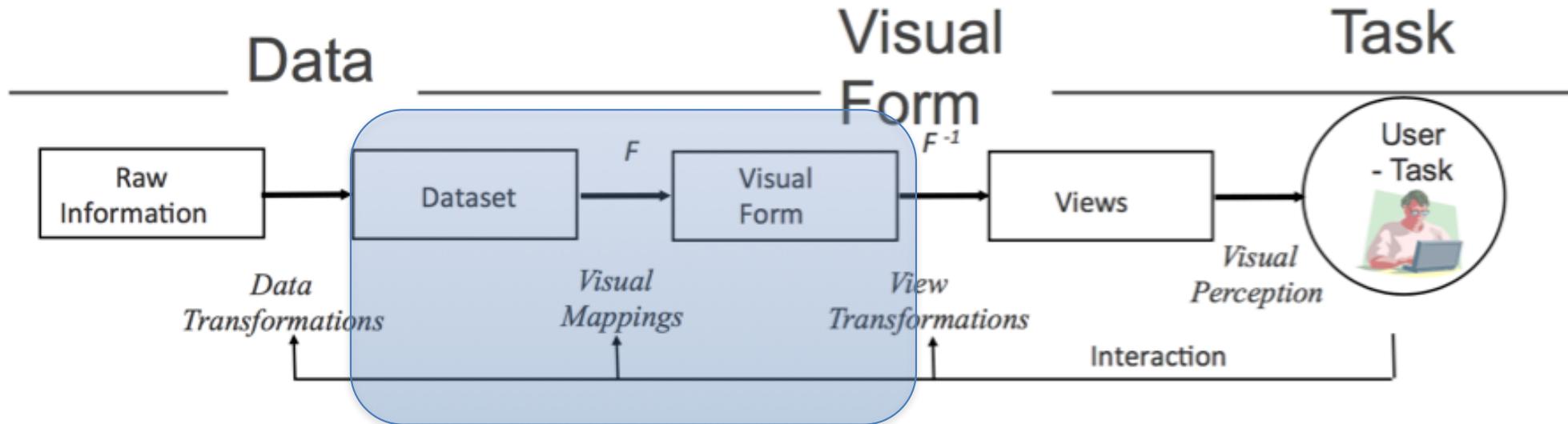


- Choose a visual model, such as a bar graph, list, or tree.

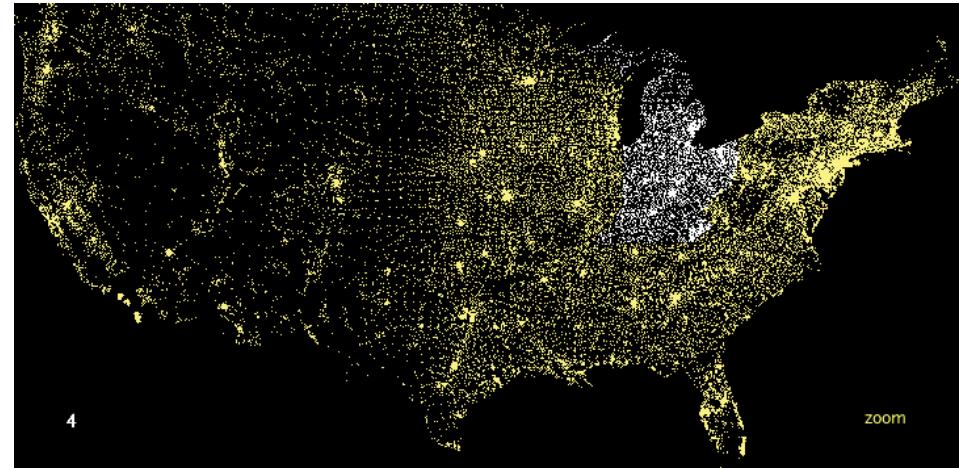


Basic visual representation of zip code data

Seven Stages: Refine

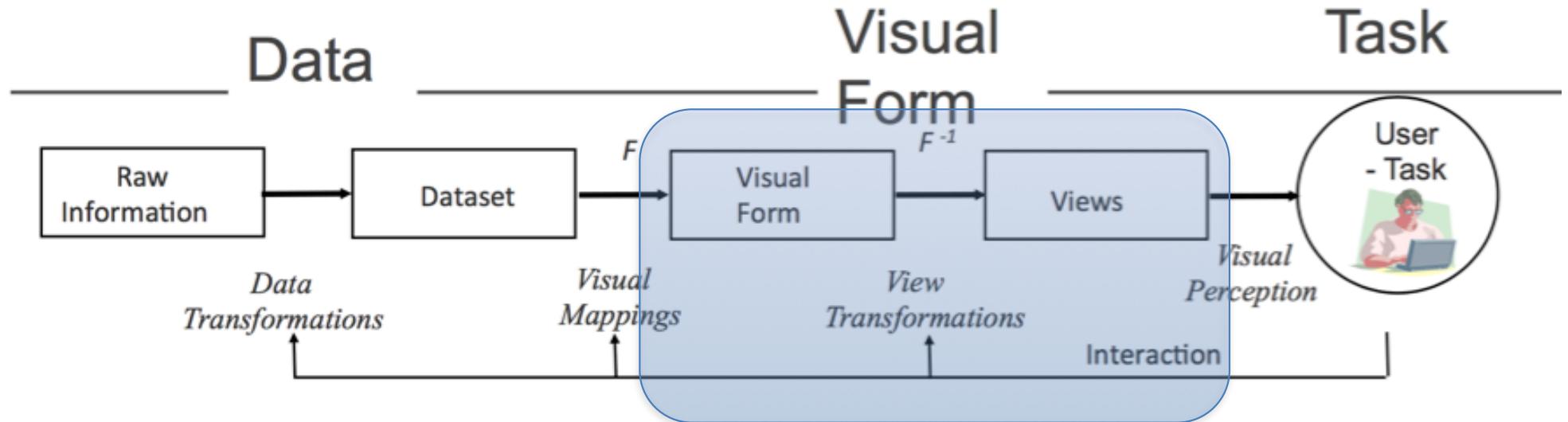


- Improve the basic representation to make it clearer and more visually engaging.

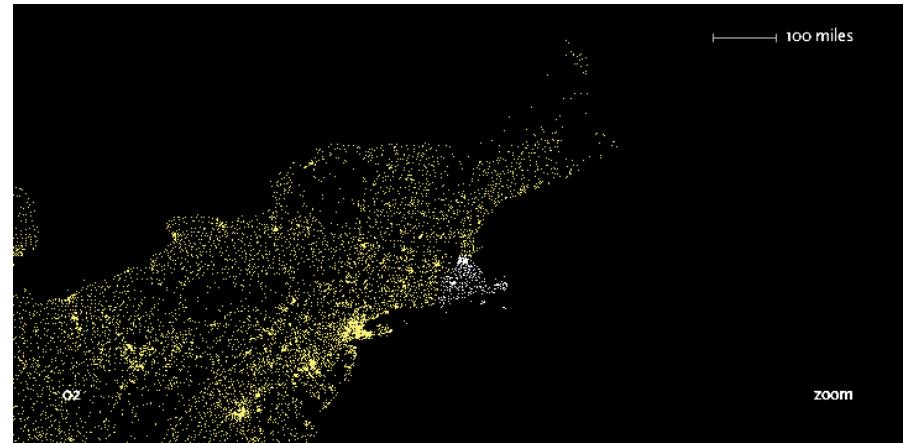


Using color to refine the representation

Seven Stages: Interact



- Add methods for manipulating the data or controlling what features are visible.



Zooming in with two digits of the post code (02)

Interaction is Vital for Exploration

- Engage in a dialog with your data
- Employ interaction in a more fundamental manner to strengthen the power of visualization
- Possible Actions
 - Select
 - Explore
 - Reconfigure
 - Encode
 - Abstract/Elaborate
 - Filter
 - Connect

Yi, et al. "Toward a deeper understanding of the role of interaction in information visualization." 2007.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Next Lecture

- Topic: Data and Image Models
 - Process data
 - Encode information
- Next Monday (4 Feb)
 - 12:00 – 14:00
 - A25, Business South, Jubilee Campus

LES VARIABLES DE L'IMAGE			12 14
	POINTS	LIGNES	
XY 2 DIMENSIONS DU PLAN	x x x	/\ / \ /	OQ ≠
Z TAILLE	.	/\ / \ /	OQ ≠
VALEUR	.	/\ / \ /	○ ≠
LES VARIABLES DE SÉPARATION DES IMAGES			13
GRAIN		/\ / \ /	○ ≠
COULEUR		/\ / \ /	= ≠
ORIENTATION		/\ / \ /	= ≠

G53FIV: Fundamentals of Information Visualization

Lecture 3: Data and Image

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Last Lecture

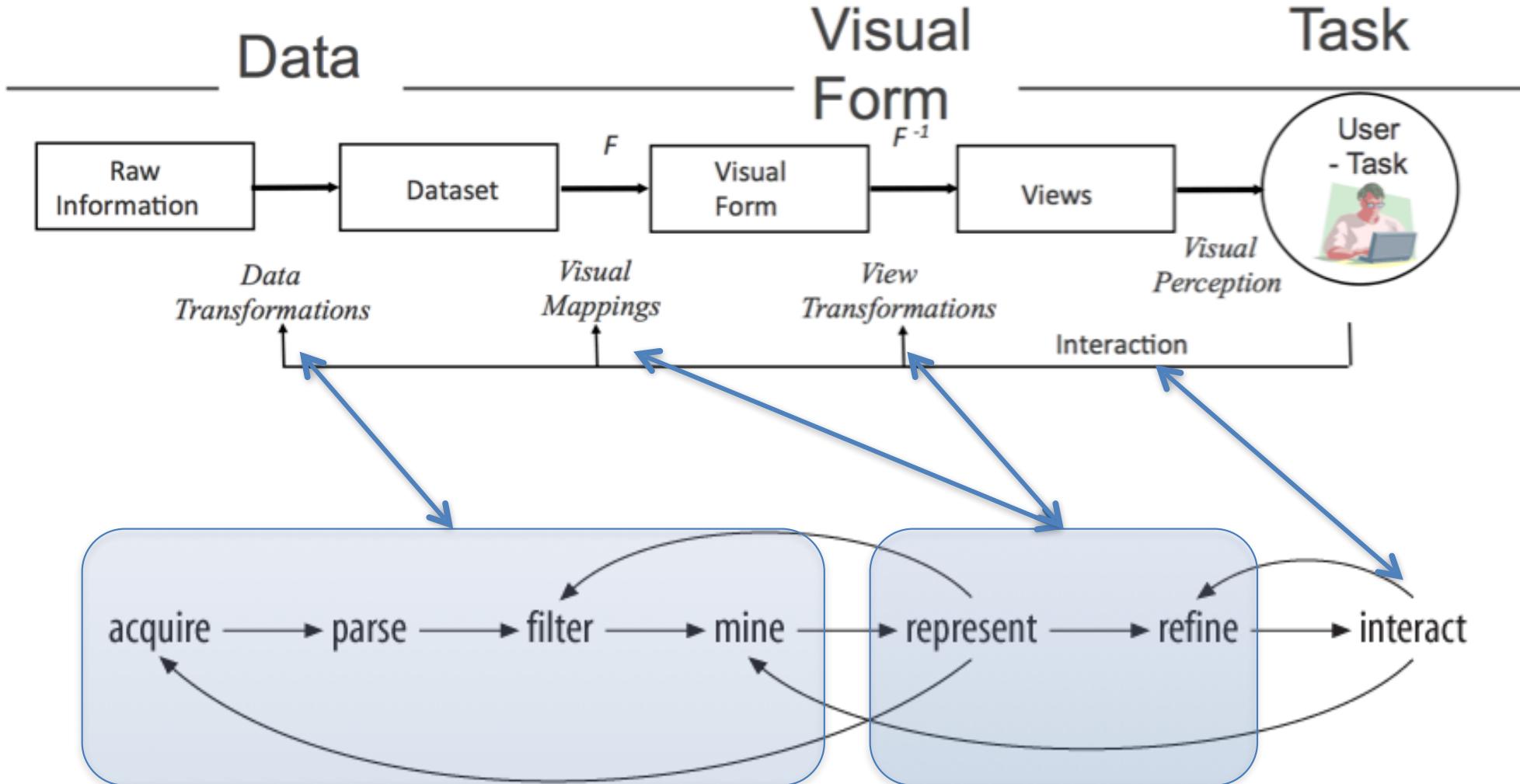
The Value of Visualization

Key Values of Visualizations

- **Record** information
 - Blueprints, photographs, seismographs, ...
- **Communicate** information to others
 - Share and persuade
 - Collaborate and revise
- Analyze data to **support reasoning**
 - Find patterns / Discover errors in data
 - Expand memory
 - Develop and assess hypotheses



Different Stages of Visualization



Overview

- How to process **data**?
 - Data models
 - Processing algorithms
- How to encode the data using **images** (the visual channel)?
 - Visual encoding (mapping)

Administrivia: Module Expectation

- 10 credits = 100 hours
- Around 20 hours of lectures
- 80 hours of **self-study**
 - 5 hours per week during term time, i.e. 1 hour per day
 - 20 hours revision
- Activities
 - Readings
 - Practice (course work)



Administrivia: G53FIV Coursework

- Objective: implementing a visualization with R
 - Pick a dataset of your interest
 - Pose the initial questions (3 to 5) that you would like to answer
 - Assess the fitness of the data
 - Answer the questions by visualizing the dataset using R in an exploratory fashion
 - Further refine/propose questions and produce the visualization for those refined/proposed (more exploratory) questions (<= 10 questions in total).
 - It is a bonus if you can make your visualization interactive
 - You can also try other visualization tools for the ultimate visualization if you want (optional, e.g., to make it more interactive). However, using R for the initial exploratory analysis is required.
 - You should work closely with the “R Graphics Cookbook”.

Administrivia: G53FIV Coursework

- Written report
 - Description of your data
 - The description with the initial questions
 - For each question, a description of your visualization strategies, including data cleaning, transformation, visual encoding, etc.
 - An explanation of the exploratory process of generating new questions and visualizations.
 - Critical discussion of your visualization design (e.g. why you pick these encodings or this visualization)
 - A reflection on the development process
 - Upload your R codes as well

Administrivia: G53IVP Project

- Goal: hands-on experience in **designing, implementing, and evaluating** a **new** visualization method, algorithm or tool.
- Some examples*:
 - <http://courses.ischool.berkeley.edu/i247/s16/>
- A written report
 - Introduction
 - Related work
 - Methods/Design (storyboard, etc.)
 - Results (Visualizations)
 - Evaluation (user study)
 - Discussions
 - Conclusions
- Demo
 - A poster covers the main visualizations
 - A presentation
- Code repositories

* Those examples are for inspiration purpose only. They are from a different course format.

Administrivia: G53IP Project

- More Examples from last year

G53IVP First Meeting

- First meeting: Feb 11th Monday 15:00 or 17:00
 - third week of G53FIV, i.e. **next Monday**
- B50, School of Computer Science
- Discuss the general format and available resources
- Doodle link to fill in (to send via email later)
- Next: Proposal development
 - Feb 25th 11:00 (fifth week of G53FIV)

Data

Data Models

- Data models are formal descriptions
- Characterize data through three components
 - Objects (Items of Interest)
 - Students, courses, semesters
 - Attributes (properties of data)
 - Name, age, id, date, score
 - Relations (how two or more objects relate)
 - Student takes course, course during semester, etc.

Example (Data Table)

cases



	Student 1	Student 2	Student 3	Student 4
Name	Tom	Jim	Mary	Jane
Age	20	19	22	21
Grade	A	B	A-	B+
Course	Math	Math	Art	Sport
Entry Year	1997	1998	1995	1996



variables

Taxonomy of Data Types

- 1D (sets and sequences) • Temporal
- 2D (maps) • Trees (hierarchies)
- 3D (shapes) • Networks (graphs)
- nD (relational) • Others?

Optional reading: The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

Nominal, Ordinal & Quantitative

- N - Nominal (labels or categories)
 - Operations: $=, \neq$ e.g. math, art (course)
- O – Ordered
 - Operations: $=, \neq, <, >$ e.g. A, A-, B+, B (grade)
- Q - Interval (location of zero arbitrary)
 - Operations: $=, \neq, <, >, -$ e.g. (3.23, -1.2) (GPS)
 - Can measure distances or spans
- Q - Ratio (zero fixed)
 - Operations: $=, \neq, <, >, -, \%$ e.g. 20, 19, 22, 21 (age)
 - Can measure ratios or proportions

Nominal, Ordinal & Quantitative

- N - Nominal (labels or categories)
 - Operations: $=, \neq$ e.g. math, art (course)
- O – Ordered
 - Operations: $=, \neq, <, >$ e.g. A, A-, B+, B (grade)
- Q - Interval (location of zero arbitrary)
 - Operations: $=, \neq, <, >, -$ e.g. (3.23, -1.2) (GPS)
 - Can measure distances or spans
- Q - Ratio (zero fixed)
 - Operations: $=, \neq, <, >, -, \%$ e.g. 20, 19, 22, 21 (age)
 - Can measure ratios or proportions

Nominal, Ordinal & Quantitative

- N - Nominal (labels or categories)
 - Operations: $=, \neq$ e.g. math, art (course)
- O – Ordered
 - Operations: $=, \neq, <, >$ e.g. A, A-, B+, B (grade)
- Q - Interval (location of zero arbitrary)
 - Operations: $=, \neq, <, >, -$ e.g. (3.23, -1.2) (GPS)
 - Can measure distances or spans
- Q - Ratio (zero fixed)
 - Operations: $=, \neq, <, >, -, \%$ e.g. 20, 19, 22, 21 (age)
 - Can measure ratios or proportions

Nominal, Ordinal & Quantitative

- N - Nominal (labels or categories)
 - Operations: $=, \neq$ e.g. math, art (course)
- O – Ordered
 - Operations: $=, \neq, <, >$ e.g. A, A-, B+, B (grade)
- Q - Interval (location of zero arbitrary)
 - Operations: $=, \neq, <, >, -$ e.g. (3.23, -1.2) (GPS)
 - Can measure distances or spans
- Q - Ratio (zero fixed)
 - Operations: $=, \neq, <, >, -, \%$ e.g. 20, 19, 22, 21 (age)
 - Can measure ratios or proportions

Example

cases



	Student 1	Student 2	Student 3	Student 4
Name (N)	Tom	Jim	Mary	Jane
Age (Q)	20	19	22	21
Grade (O)	A	B	A-	B+
Course (N)	Math	Math	Art	Sport
Entry Year (Q)	1997	1998	1995	1996

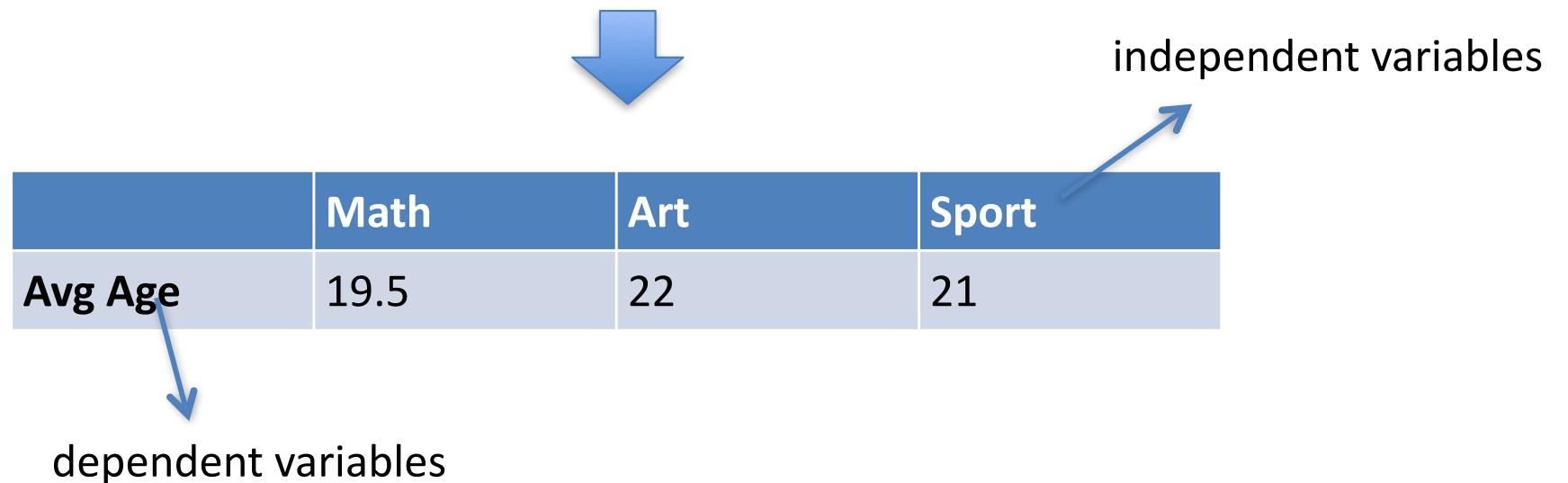
variables



Dimensions and Measures

- Dimensions (independent variables)
 - Discrete variables describing data (N, O)
 - Categories, dates, binned quantities
- Measures (dependent variables)
 - Data values that can be aggregated (Q)
 - Numbers to be analyzed
 - Aggregate as sum, count, avg, std. dev...

	Student 1	Student 2	Student 3	Student 4
Name (N)	Tom	Jim	Mary	Jane
Age (Q)	20	19	22	21
Grade (O)	A	B	A-	B+
Course (N)	Math	Math	Art	Sport
Entry Year (Q)	1997	1998	1995	1996



Exercises

- N, O, Q?
- Dimension or Measure?

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080

Exercises

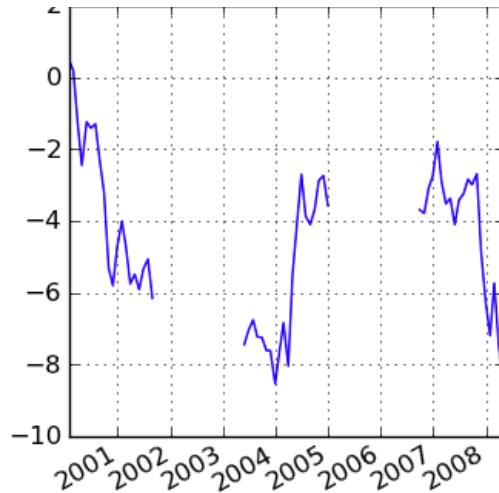
- N, O, Q?
- Dimension or Measure?

– Year	Q-Internal (O)	Dimension
– Age	Q-Ratio (O)	Depends
– Marital	N	Dimension
– Sex	N	Dimension
– People	Q-Ratio	Measure

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080

Data Processing

- Data cleaning and filtering
 - for quality control
 - Remove (Outlier, missing data)
 - Modify (conversion of format, etc.)
- Data adjustment
 - Depends on your task and questions to ask
 - Relational algebra:
 - e.g. Aggregation, mean, sort, projection
 - Reformatting and Integration



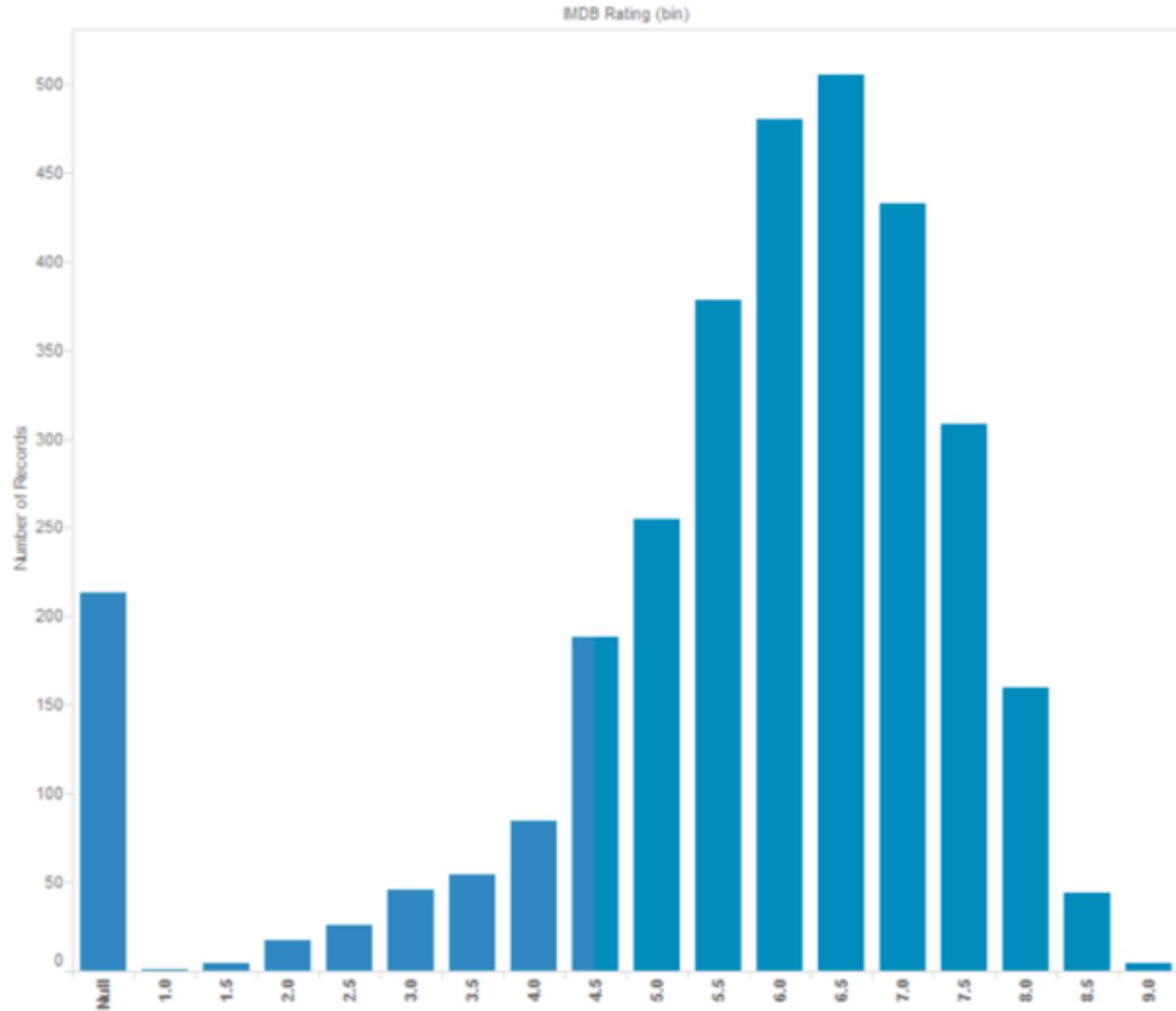
*We will learn later how
to do these in R.*

Data Cleaning and Filtering

- Missing Data
 - no measurements, redacted, ...?
- Erroneous Values
 - misspelling, outliers, ...?
- Type Conversion
 - e.g., zip code to lat-lon
- Entity Resolution
 - diff. values for the same thing?
- Data Integration
 - effort/errors when combining data
- Anticipate problems with your data. Many research problems around these issues!

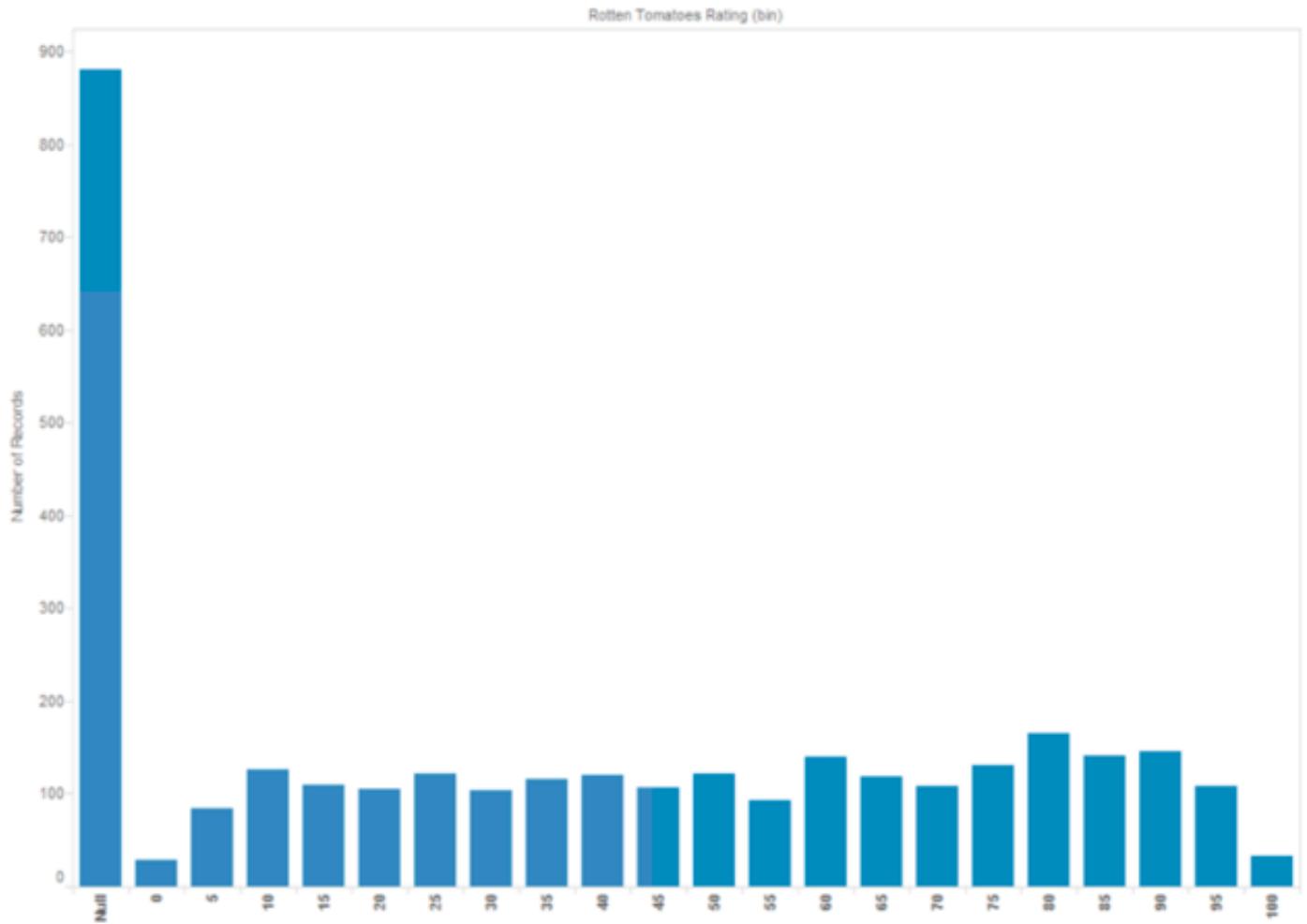
An Example

- Movie rating data
 - IMDB ratings



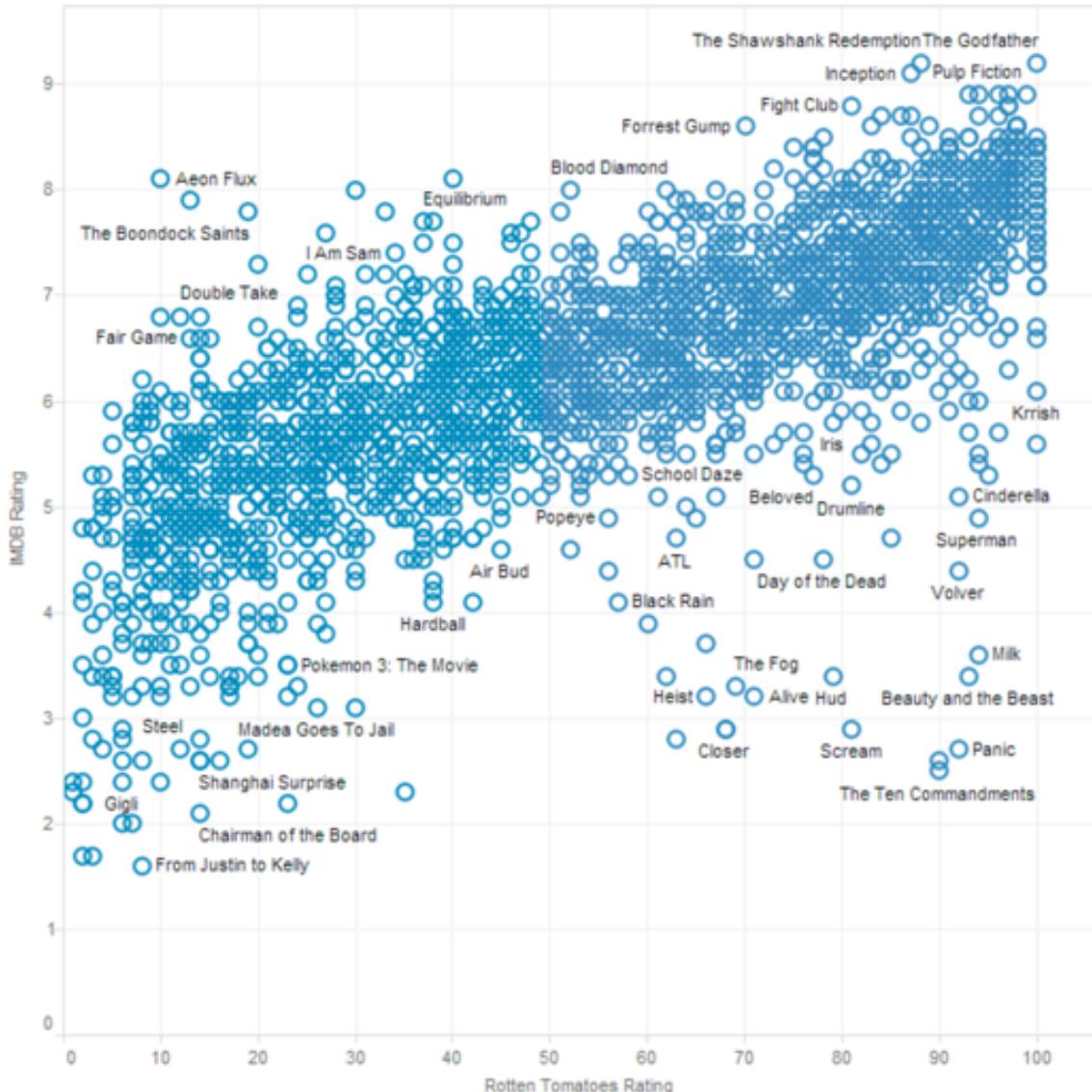
An Example

- Movie rating data
 - Rotten Tomato Ratings
- Many data ratings as null.



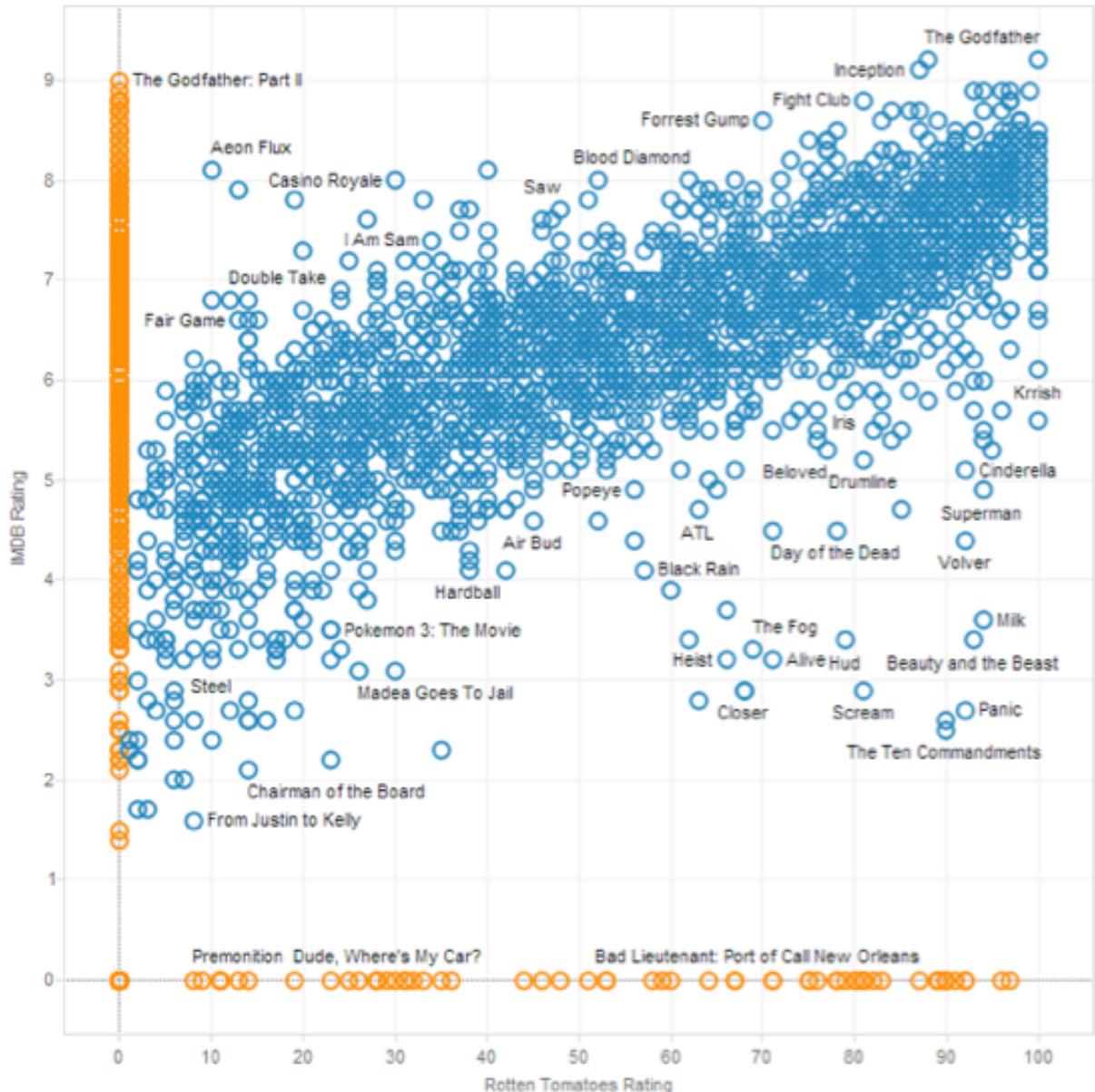
An Example

- Movie rating data scatter plot

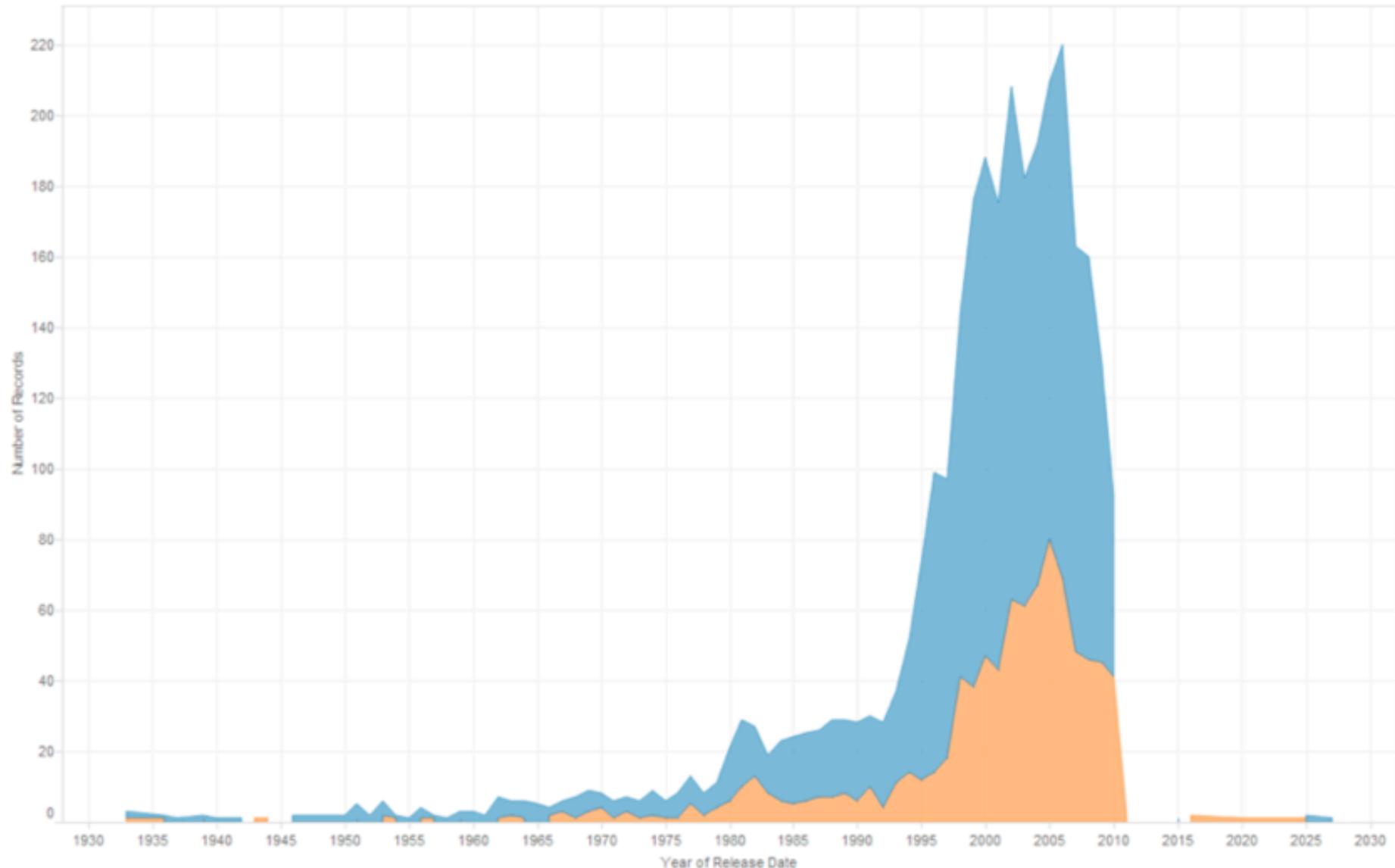


An Example

- Movie rating data scatter plot
- Many data ratings as null/missing (orange)

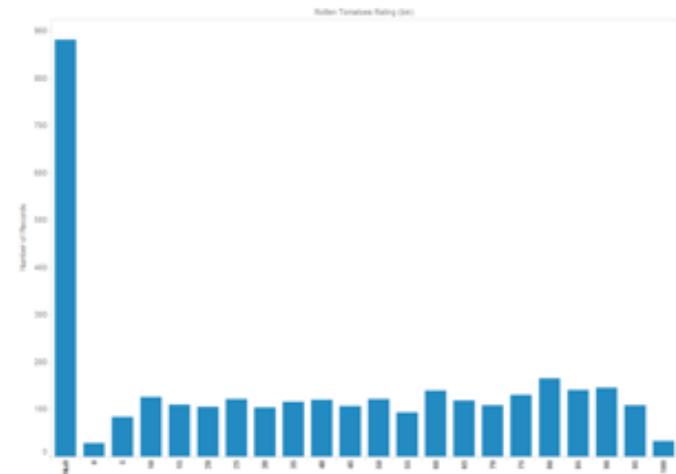


An Example



Data Cleaning and Filtering

- Exercise Skepticism
- Check data quality and your assumptions.
- Start with univariate summaries, then start to consider relationships among variables.
- Avoid premature fixation!

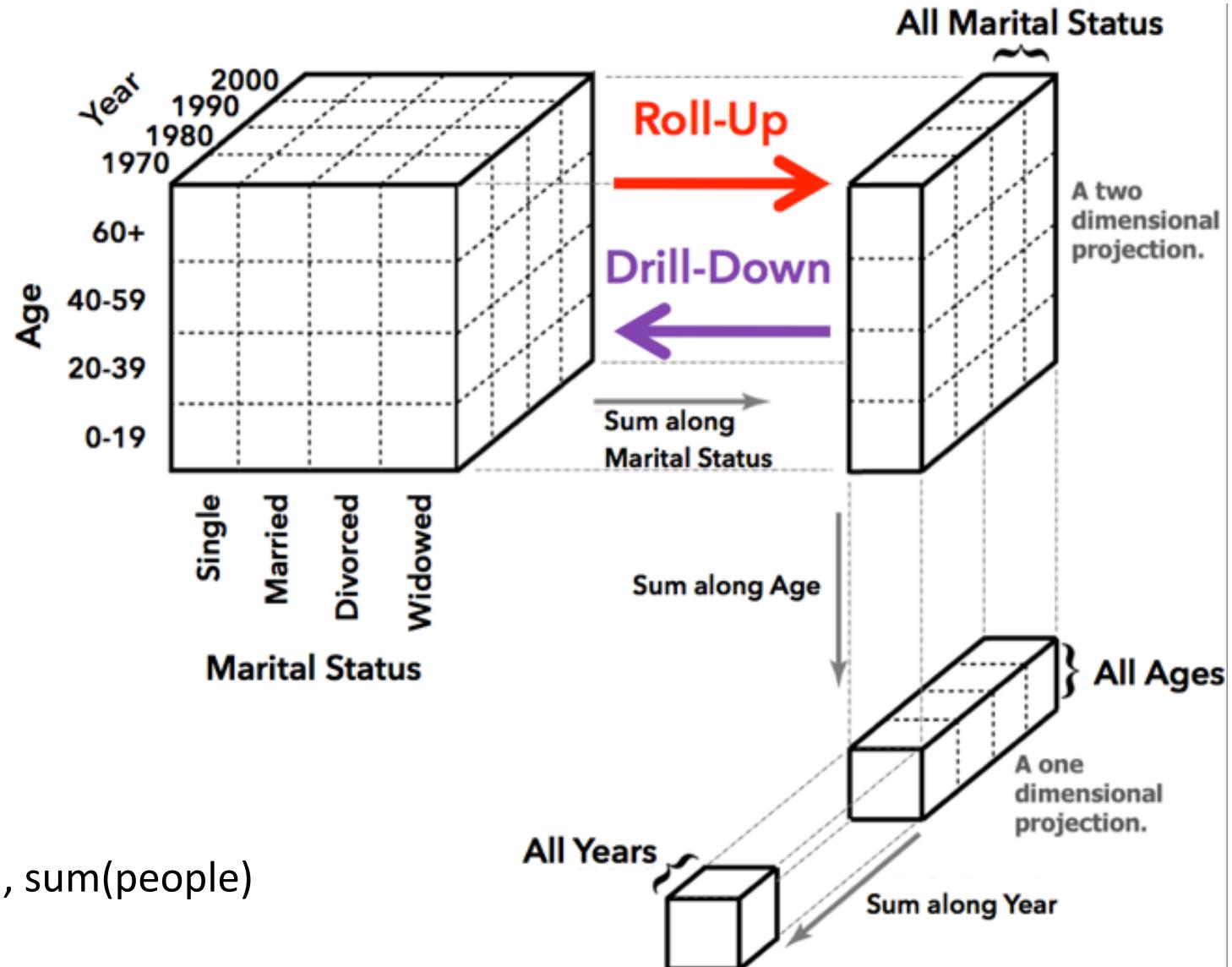


Data Adjustment: Relational Algebra

- Relational Data Model
- Data Transformations (SQL)
 - Projection (select) - selects columns
 - Selection (where) - filters rows
 - Sorting (order by)
 - Aggregation (group by, sum, min, max, ...)
 - Combine relations (union, join, ...)

Data Adjustment

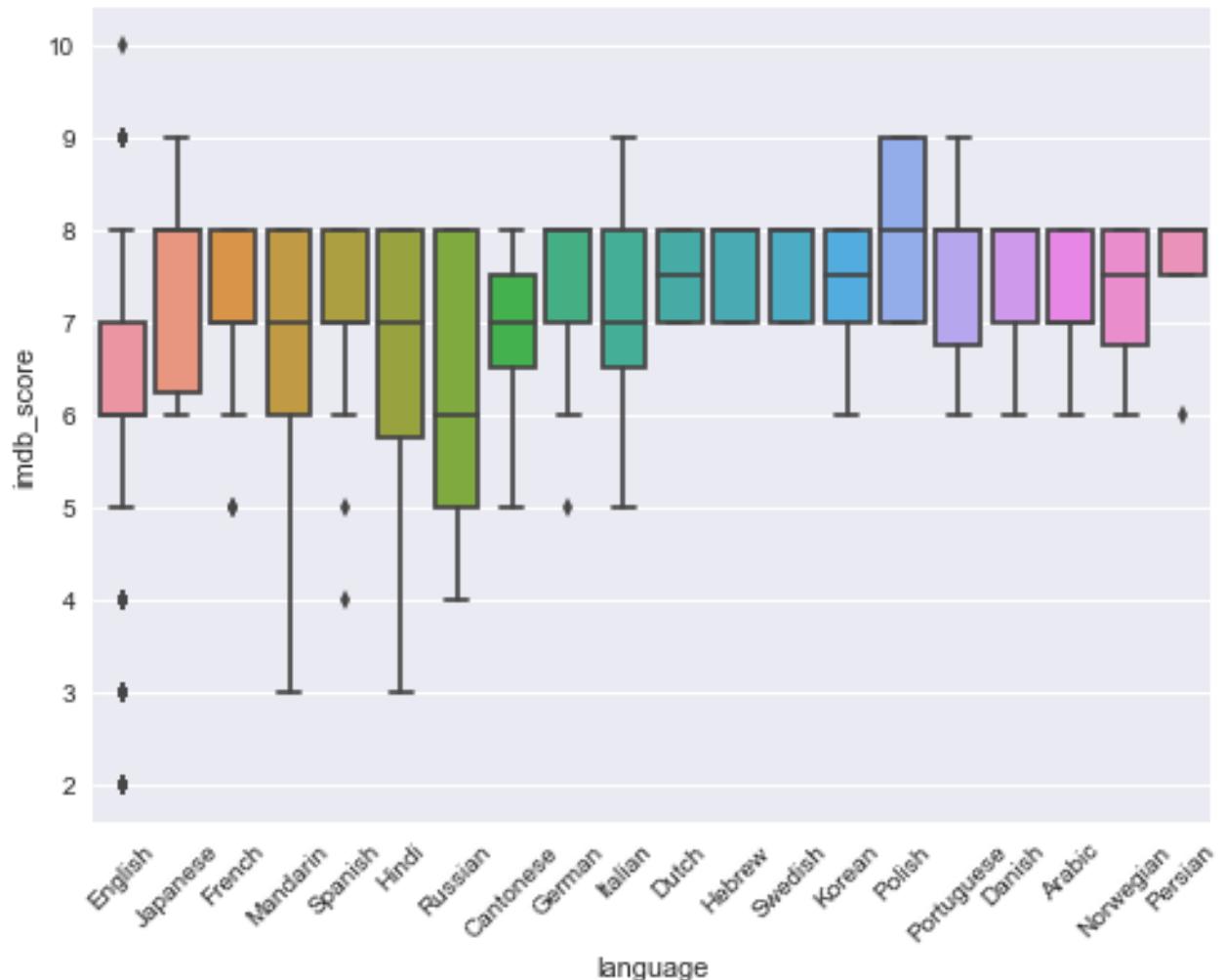
- Roll-up
- Drill-down



```
SELECT year, age, marital, sum(people)  
FROM census  
GROUP BY year, age, marital;
```

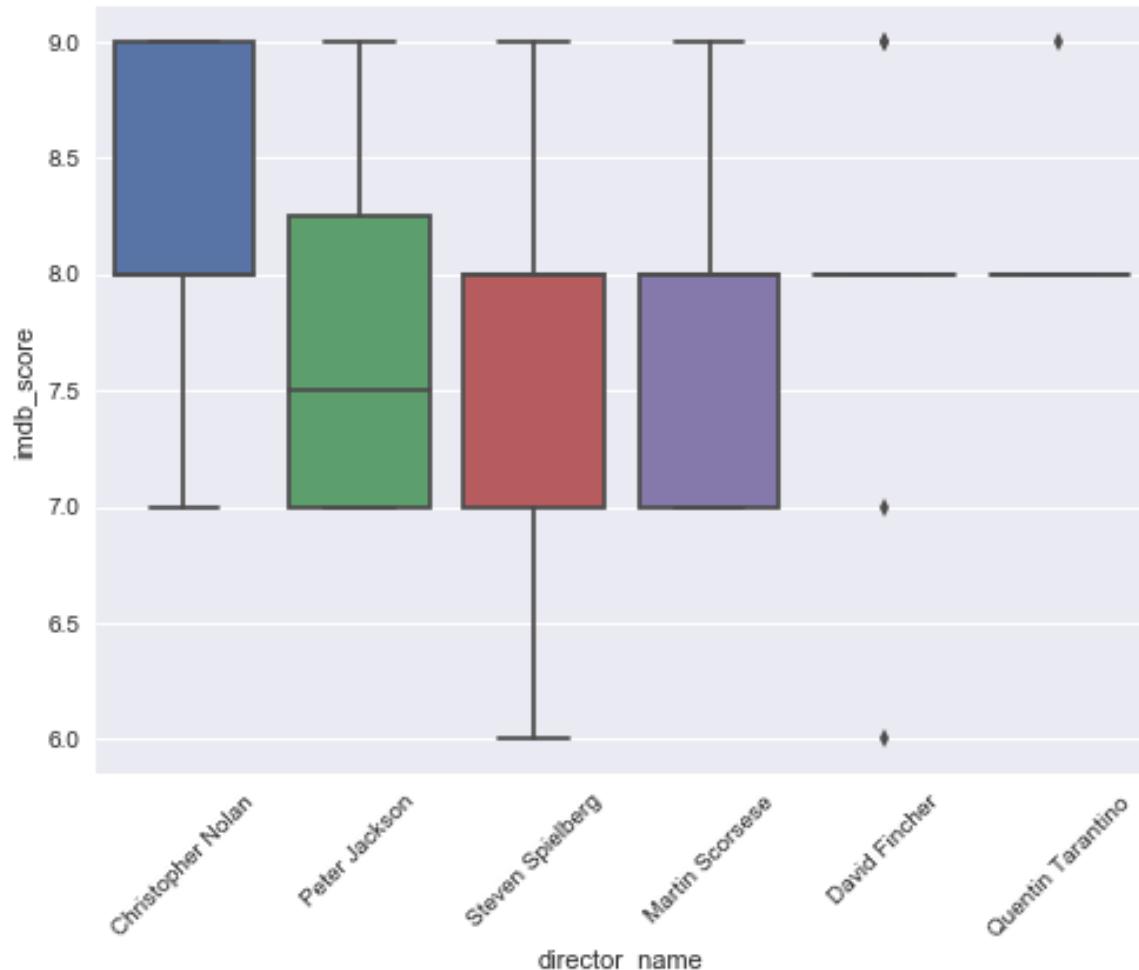
An Example

- IMDB movie rating by language



An Example

- IMDB movie rating by director



Data Adjustment

- Additional readings:
 - Relational algebra
 - database (SQL)
- You need to think carefully about what questions to answer in order to decide how you adjust the data.
- We will learn some basics when we process data using R.

Image

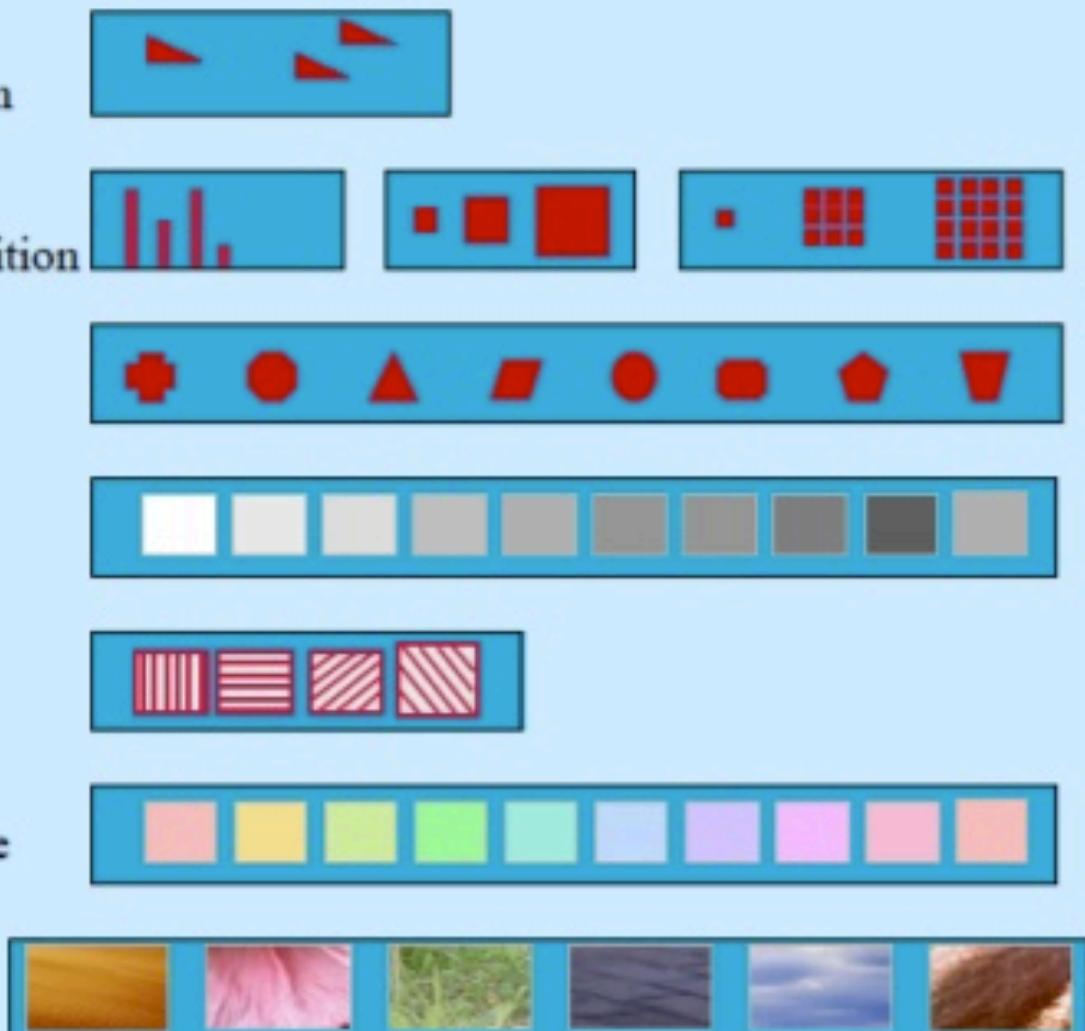
Image: Visual Language

- Visual Language is a Sign System
 - Images perceived as a set of signs
 - Sender encodes information in signs
 - Receiver decodes information from signs
- "Resemblance, order and proportion are the three sign fields in graphics."
 - Jacques Bertin

Visual Encoding Variables

Bertin's Semiology of Graphics (1967)

- **position**
 - changes in the x, y, (z) location
- **size**
 - change in length, area or repetition
- **shape**
 - infinite number of shapes
- **value**
 - changes from light to dark
- **orientation**
 - changes in alignment
- **colour**
 - changes in hue at a given value
- **texture**
 - variation in pattern
- **motion**



Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Information in Hue and Value

- Value is perceived as ordered
 - Encode ordinal variables (O)



- Encode continuous variables (Q) [not as well]



- Hue is normally perceived as unordered
 - Encode nominal variables (N) using color



Bertin's Levels of Organization

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

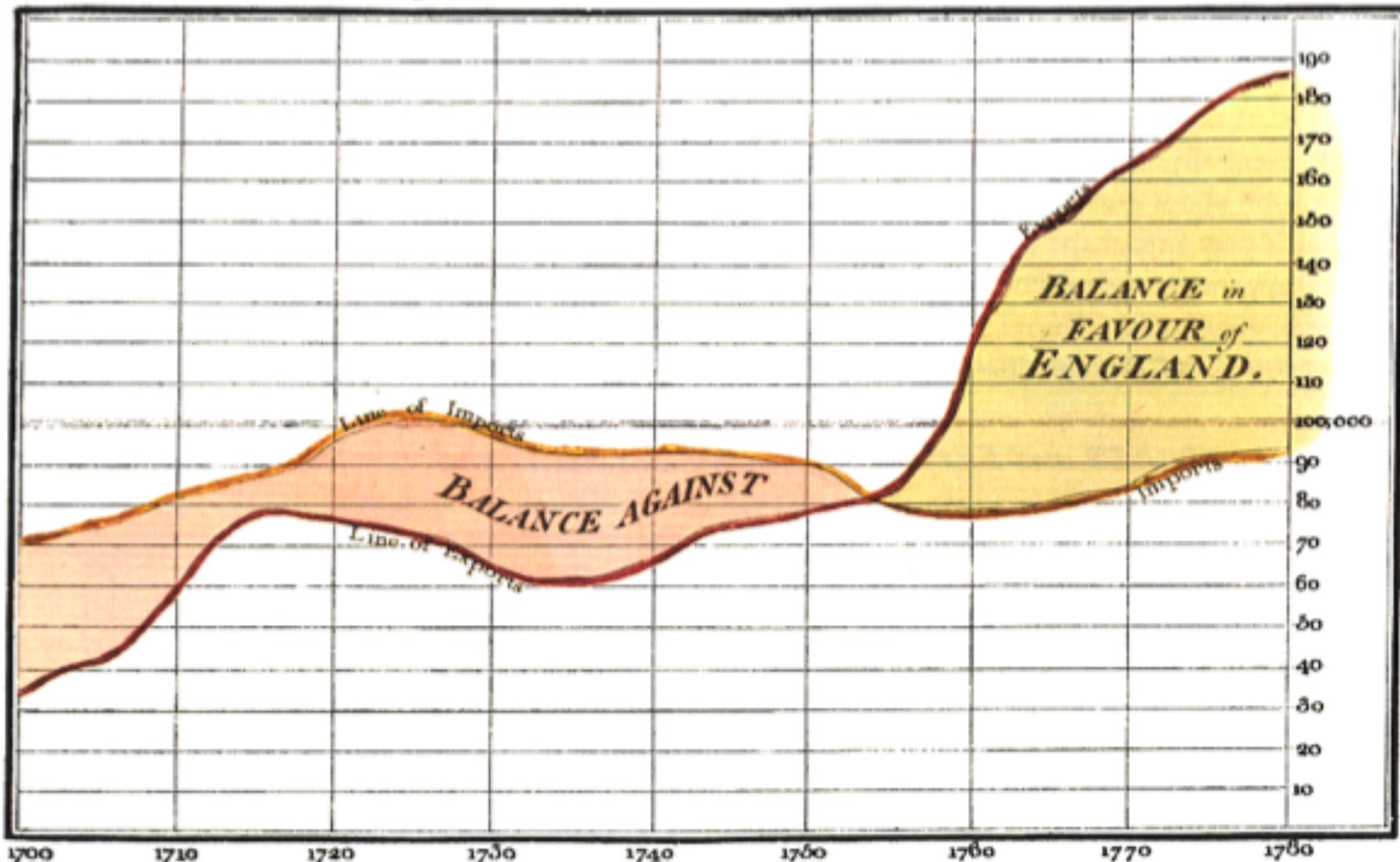
✓ = Good

~ = OK

✗ = Bad

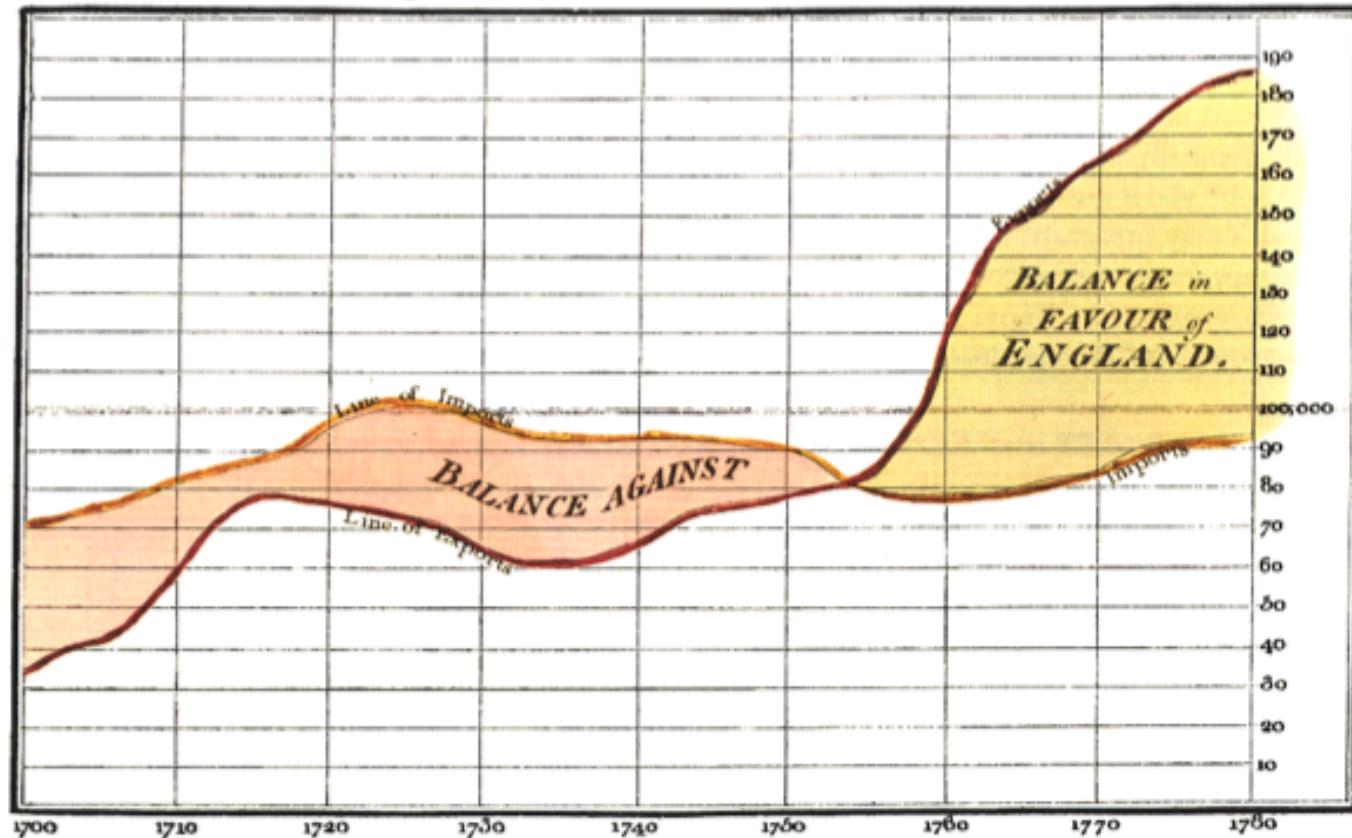
Examples

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



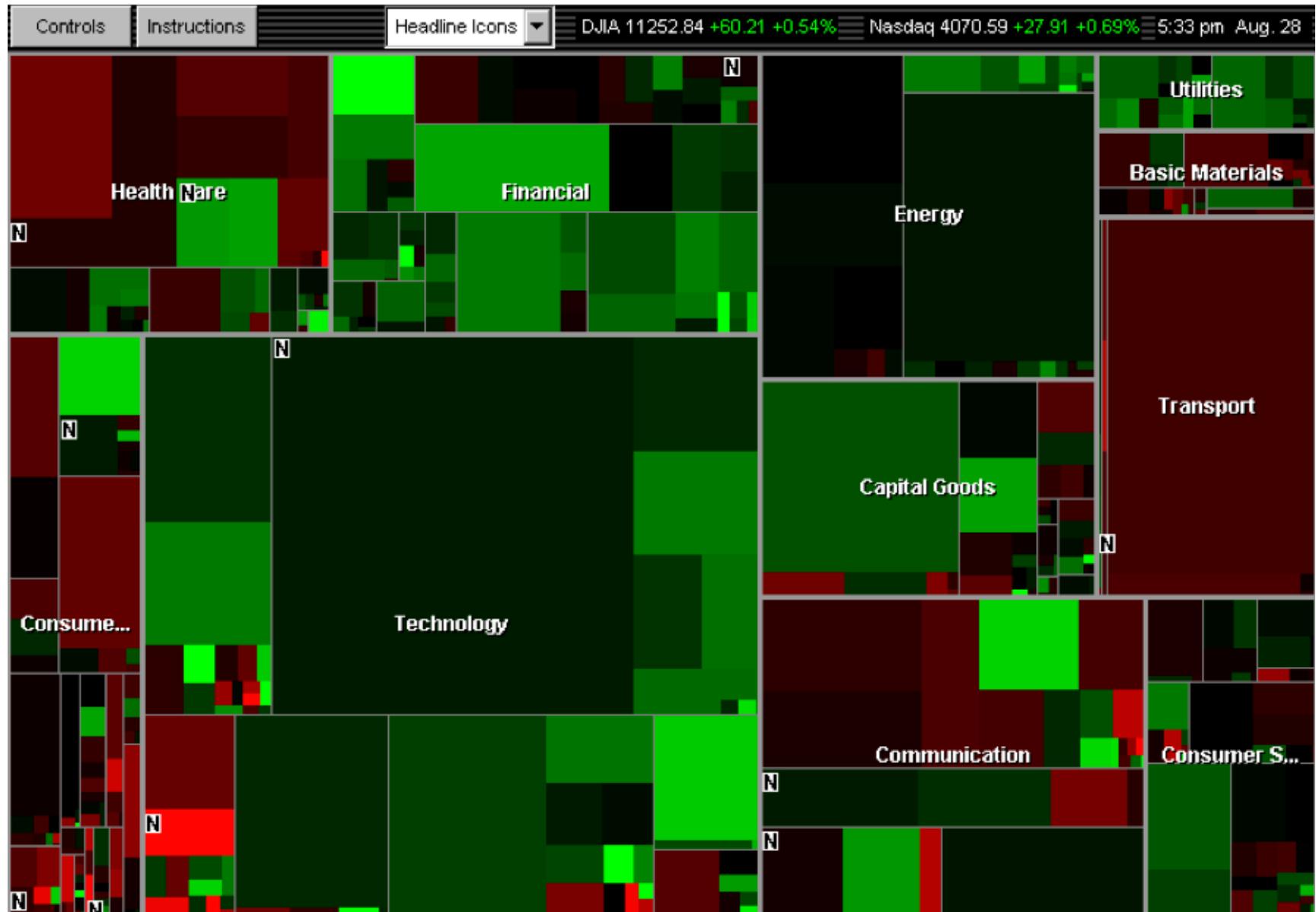
Examples

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

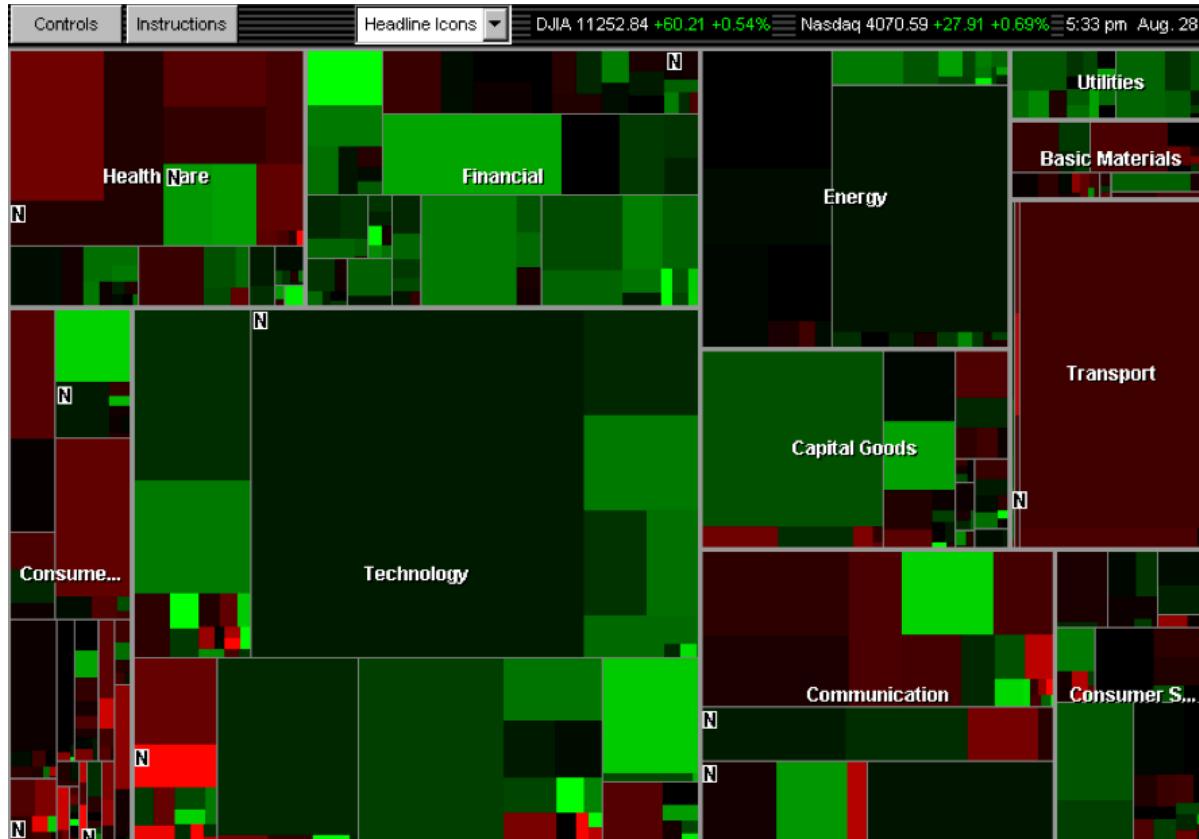


- X-axis: year (Q); Y-axis: currency (Q)
- Color: imports/exports (N, O)

Examples



Examples



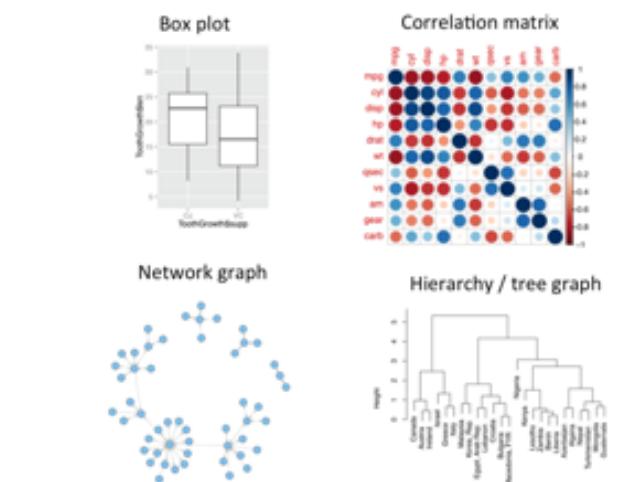
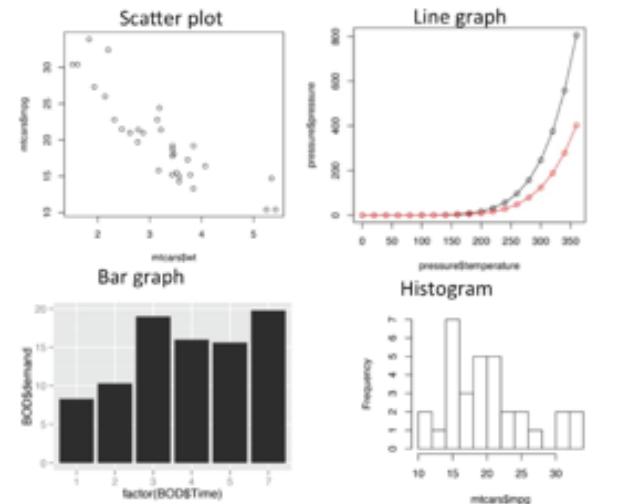
- Rectangle area: market cap (Q);
- Rectangle position: market sector (N)
- Color Hue: loss vs. gain (N, O)
- Color Value: magnitude of loss or gain (Q)

How do we choose visual encodings?

What design criteria should we follow?

Next Lecture

- Topic: Design and Graphs
 - Design Principles
 - Fundamental graphs and charts



G53FIV: Fundamentals of Information Visualization

Lecture 4: Design and Graphs

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Last Lecture

Data and Image

Overview

- Design Criteria
- Graphs
 - for uni, bi and tri-variate data

What design criteria should we follow?

Choosing Visual Encodings

- Assume k visual encodings and n data attributes.
We would like to pick the “best” encoding among a combinatorial set of possibilities of size $(n+1)^k$
- Principle of Consistency
 - The properties of the image (visual variables) should match the properties of the data.
- Principle of Importance Ordering
 - Encode the most important information in the most effective way.

What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language (1) **express all the facts** in the set of data, and (2) **only the facts** in the data.

Tell the truth

- **Effectiveness**

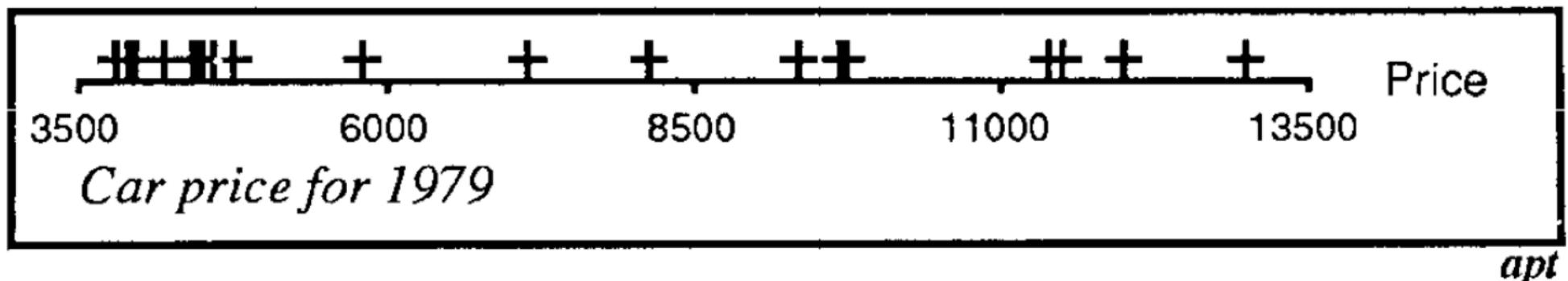
- A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Use proper encoding

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Expressiveness

- Unable to express all facts in a layout (fails first criterion)



Price : Cars → [3500, 13000]

Mileage : Cars → [10, 40]

Weight : Cars → [1500, 5000]

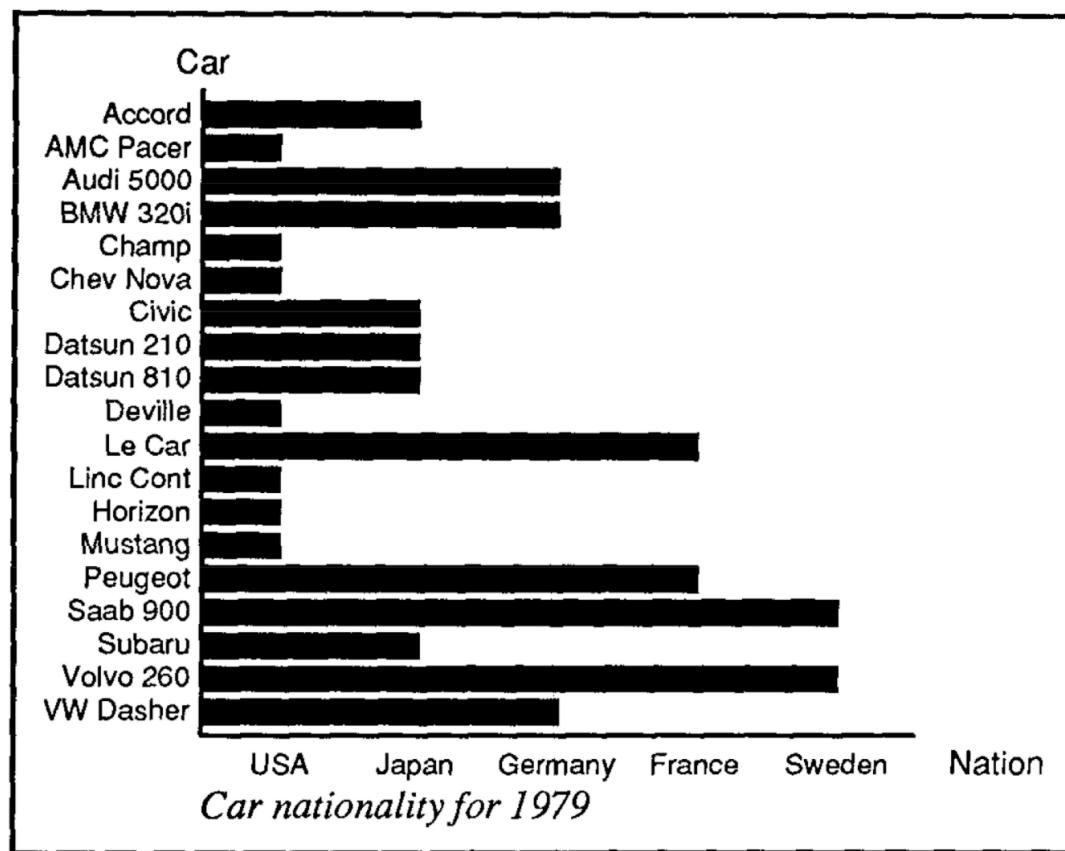
Repair : Cars → {Great, Good, OK, Bad, Terrible}

Nation : Cars → {USA, Germany, France, ...}

Cars = {Accord, AMC Pacer, Audi 5000, BMW 320i, ...}

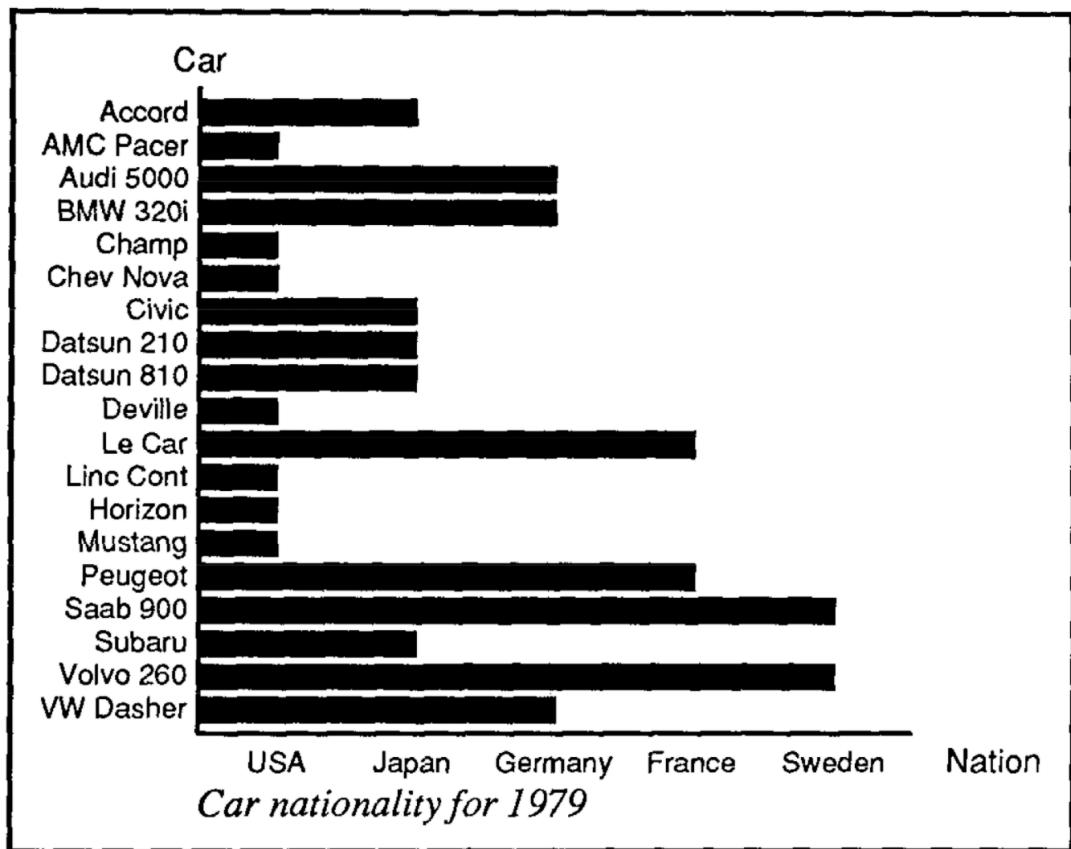
Expressiveness

- Expresses information not inherent in the dataset (fails second criterion)

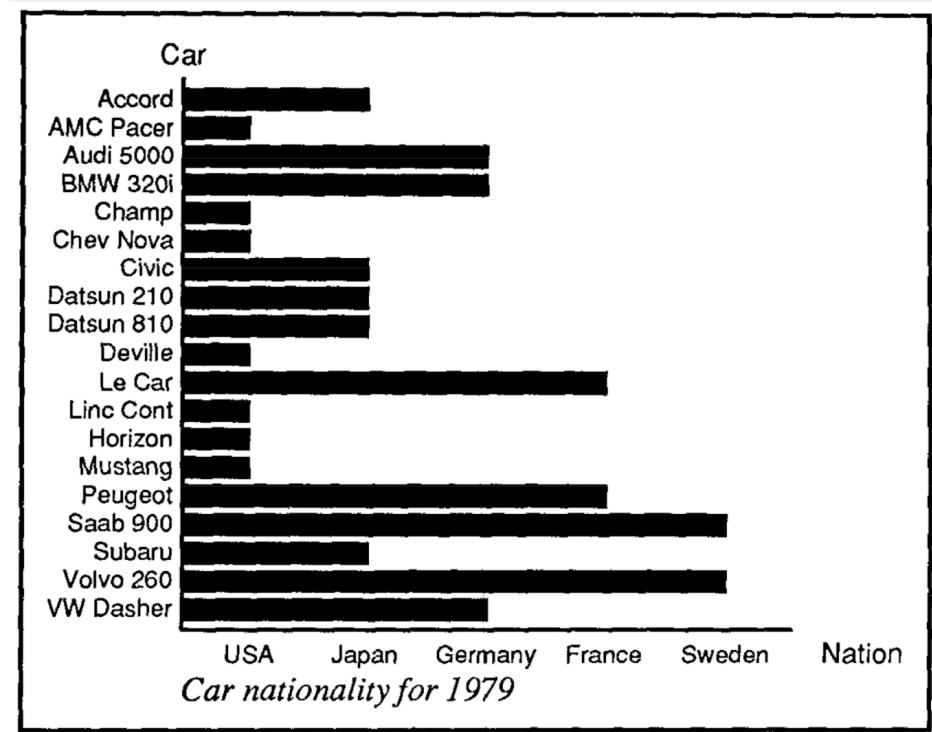
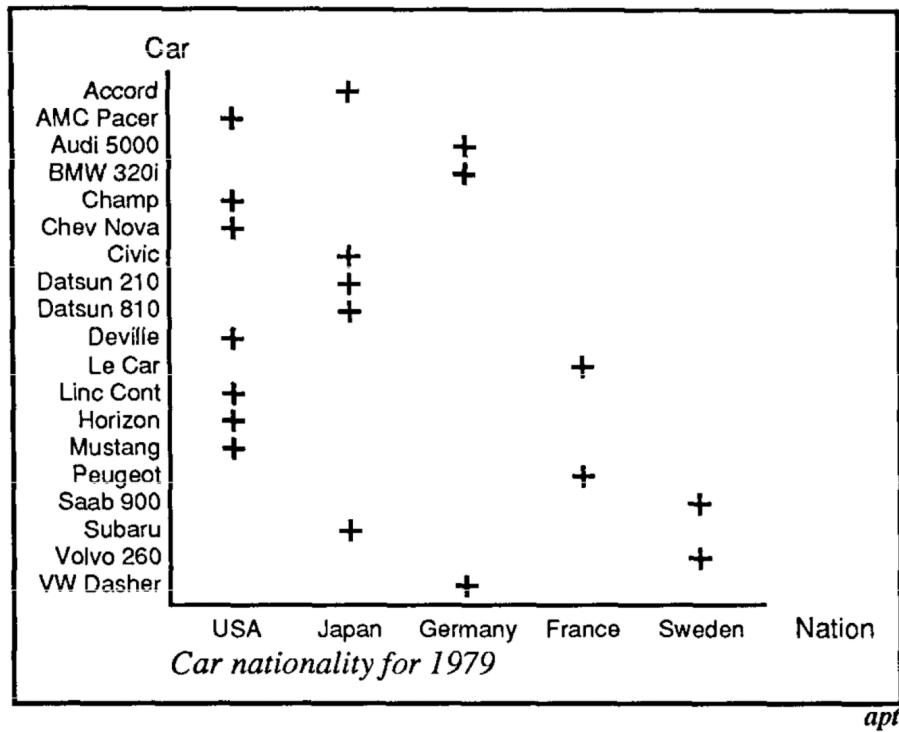


Expressiveness

- Expresses information not inherent in the dataset (fails second criterion)
- A length is interpreted as a quantitative value.



Expressiveness



An Alternative

What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express (1) all the facts in the set of data, and (2) only the facts in the data.

Tell the truth

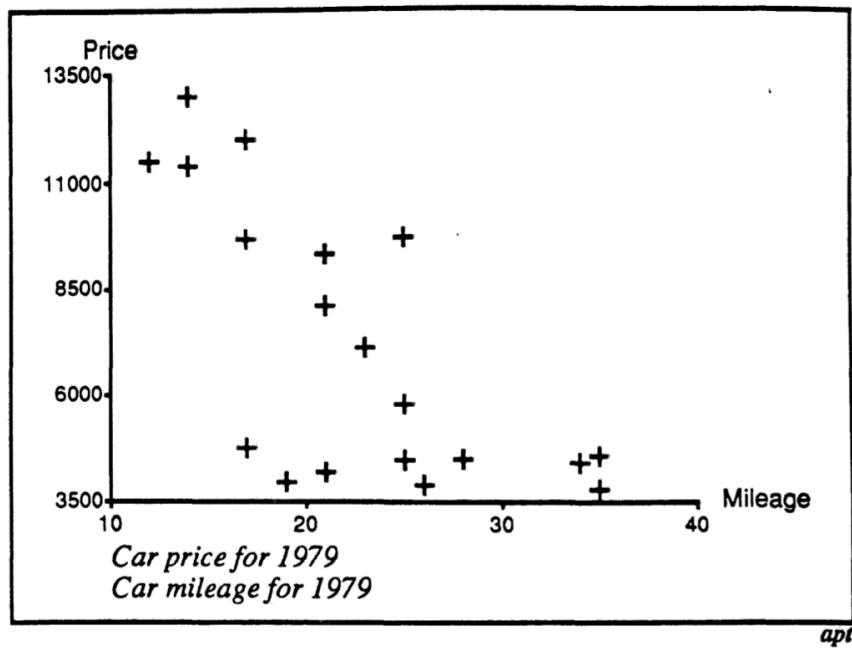
- **Effectiveness**

- A visualization is more effective than another visualization if the information conveyed by one visualization **is more readily perceived** than the information in the other visualization.

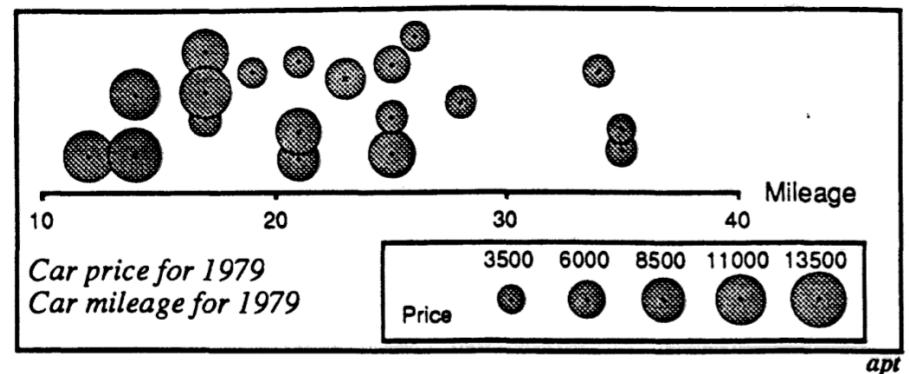
*Use proper
encoding*

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Effectiveness



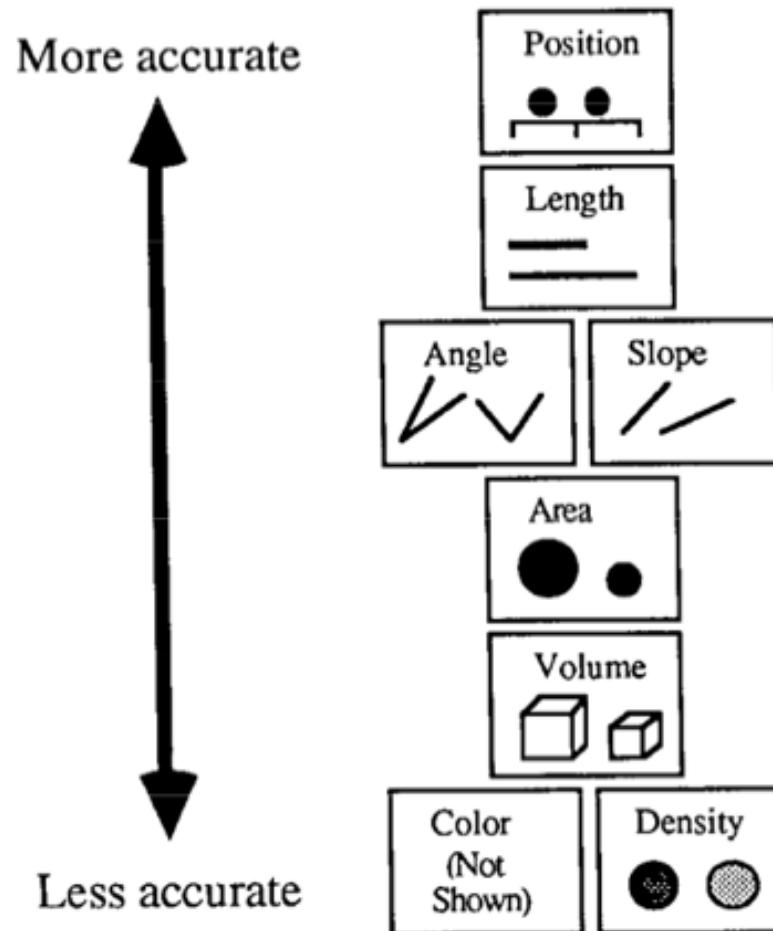
vs.



Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

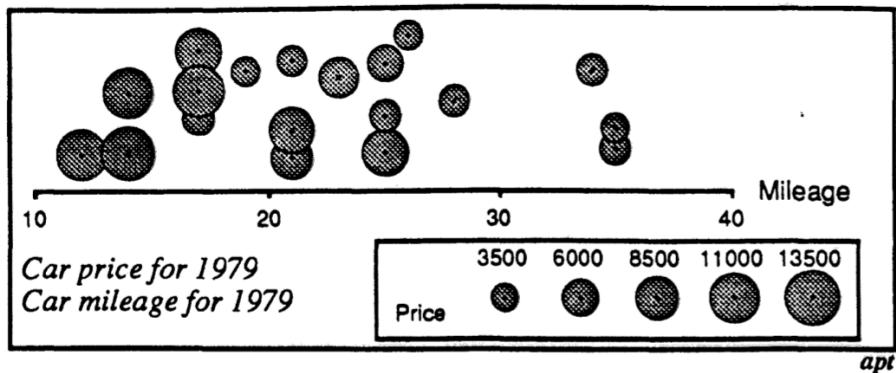
Effectiveness: Accuracy Ranking for Quantitative Information



Mackinlay, Automating the design of graphical presentations of relational information, 1986.

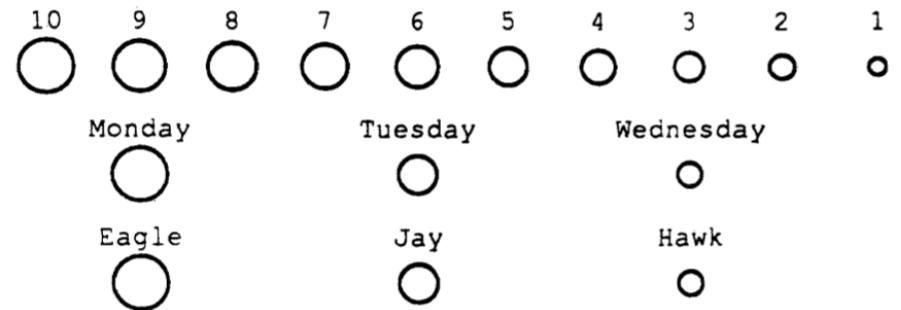
Effectiveness: Accuracy Ranking for Nominal/ Ordinal Information?

Area Encoding



Quantitative

We can use, but not so accurate.

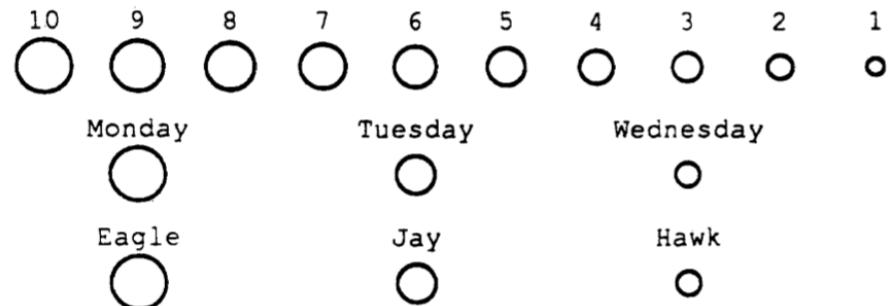
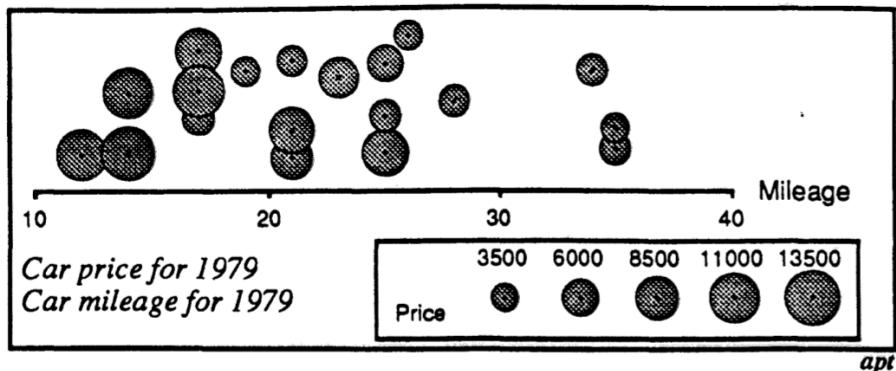


Nominal

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Effectiveness: Accuracy Ranking for Nominal/ Ordinal Information?

Area Encoding



Quantitative

We can use, but not so accurate.

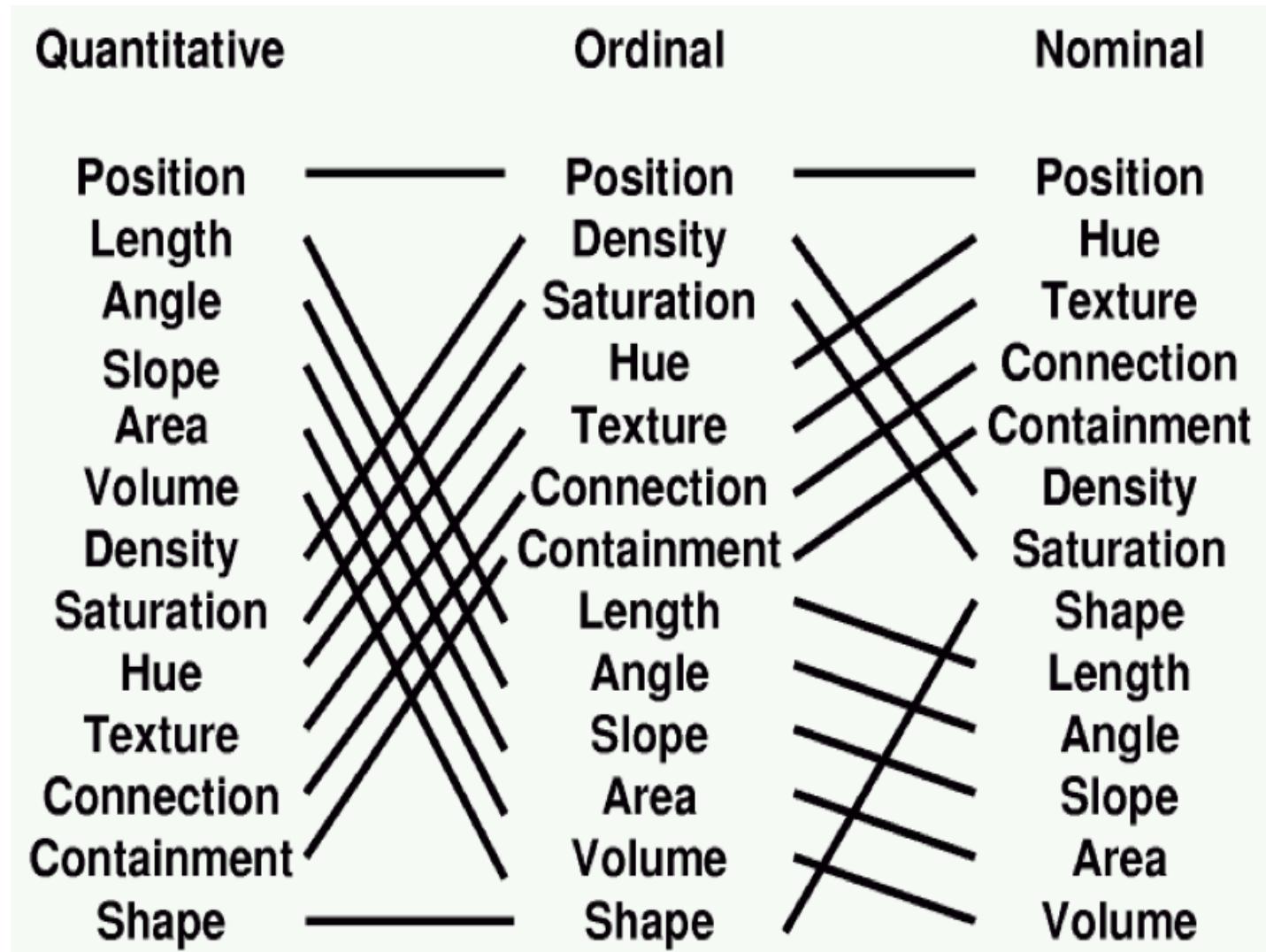
Nominal

- Problematic if there are too many categories;
- Can be expected to encode ordinal information

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Conjectured Effectiveness of Encodings by Data Type

- Nominal/Ordinal variables: detect differences
- Quantitative variables: estimate magnitudes



Mackinlay, Automating the design of graphical presentations of relational information, 1986.

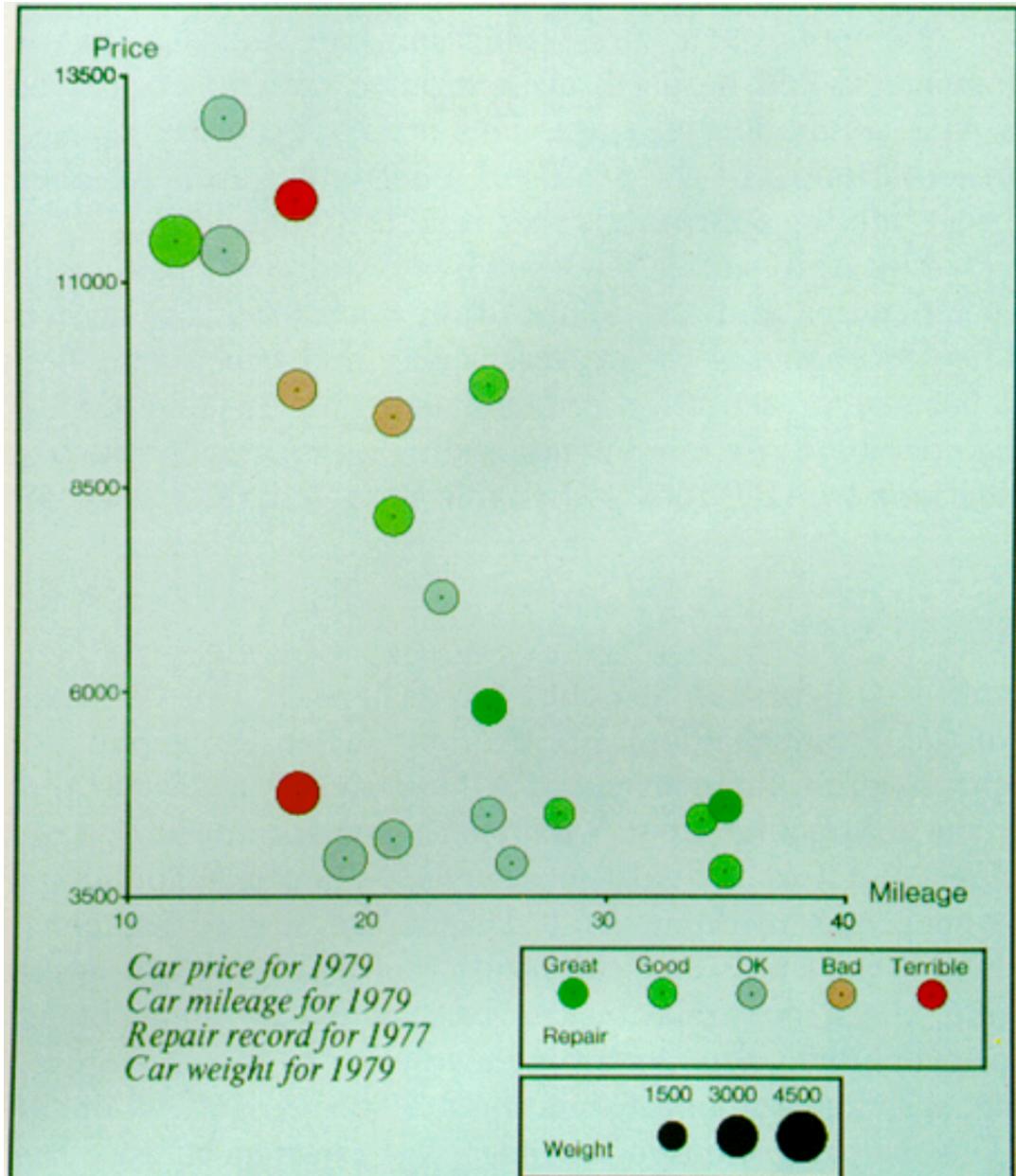
Mackinlay's Design Algorithm

- APT - “A Presentation Tool”, 1986
- User formally specifies data model and type
 - Input: ordered list of data variables to show
- APT searches over design space
 - Test expressiveness of each visual encoding Generate encodings that pass test
 - Rank by perceptual effectiveness criteria
- Output the “most effective” visualization

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

APT

- Automatically generate chart for car data
- Input variables:
 - Price
 - Mileage
 - Repair
 - Weight



Limitations of APT?

Limitations of APT?

- Does not cover many visualization techniques
 - Networks, hierarchies, maps, diagrams
 - Also: 3D structure, animation, illustration, ...
- Does not consider interaction
- Does not consider semantics / conventions
- Assumes single visualization as output

Summary of Design Criteria

- Choose expressive and effective encodings
 - Rule-based tests of expressiveness
 - Perceptual effectiveness rankings
 - Prioritizes encodings that are most easily/accurately interpreted
 - Principle of Importance Ordering: Encode more important information more effectively (Mackinlay)
- Question: how do we establish effectiveness criteria?
 - Subject of the visual perception lecture...

Graphs

How Many Variables?

- Data sets of dimensions 1, 2, 3 are common
- Number of variables per class
 - 1 - Univariate data
 - 2 - Bivariate data
 - 3 - Trivariate data
 - >3 - Hypervariate data

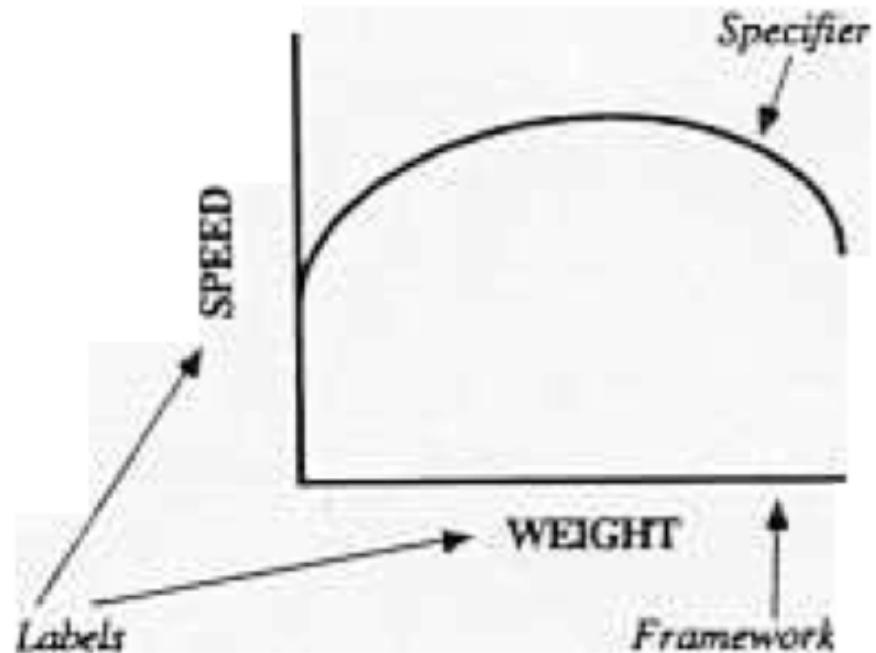
Graphs

- Data Dimensions
 - 1 - Univariate data
 - 2 - Bivariate data
 - 3 - Trivariate data
 - >3 - Hypervariate data
- Data Types
 - Nominal, Ordinal, Quantitative
- Visualization Representations
 - Points, Lines, Bars, Boxes

We mainly focus on uni, bi and tri-variate data for the rest of the lecture.

Components of Graphs

- Framework
 - Measurement types, scale
- Content (Specifier)
 - Marks, lines, points
- Labels
 - Title, axes, ticks

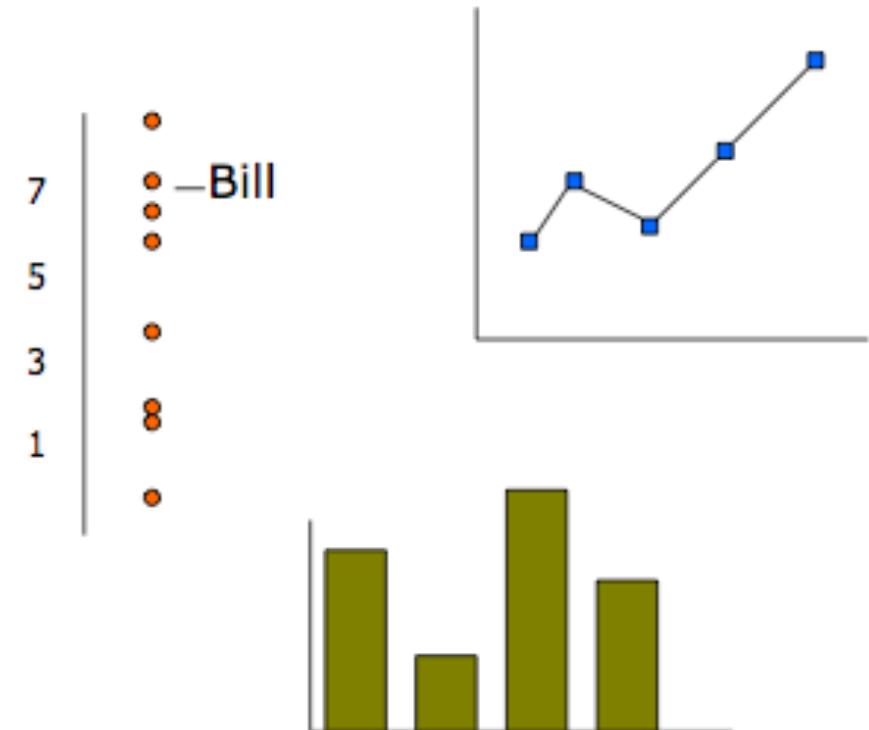


Points, Lines, Bars, Boxes

- Points
 - Useful in scatterplots for 2-values
 - Can replace bars when scale doesn't start at 0
- Lines
 - Connect values in a series
 - Show changes, trends, patterns
 - Not for a set of nominal or ordinal values
- Bars
 - Emphasizes individual values
 - Good for comparing individual values
- Boxes
 - Shows a distribution of values

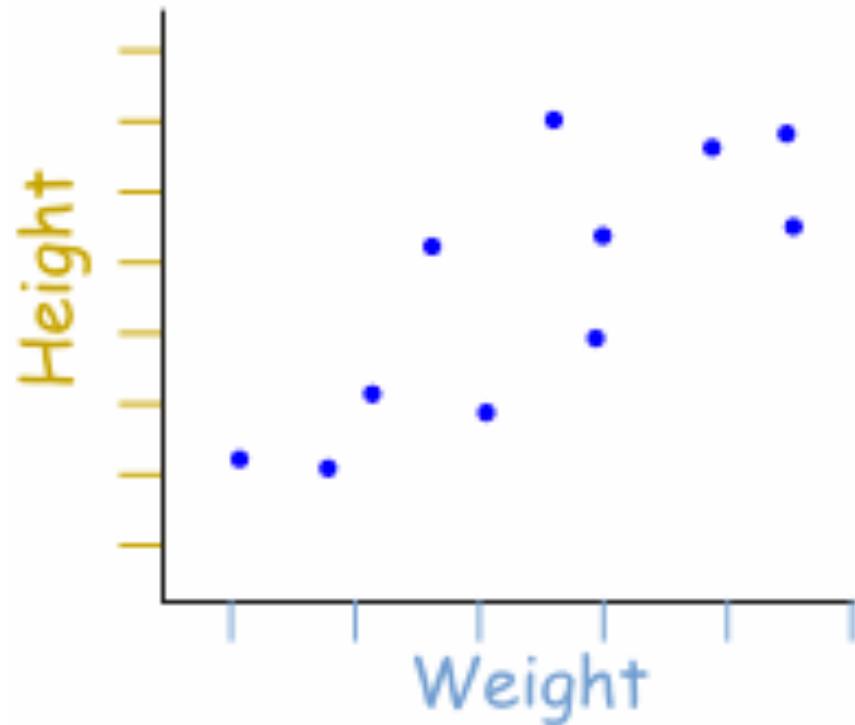
Univariate Data

- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another.
- Statistical view
 - Independent variable on x-axis (data case)
 - Track dependent variable along y-axis (value)



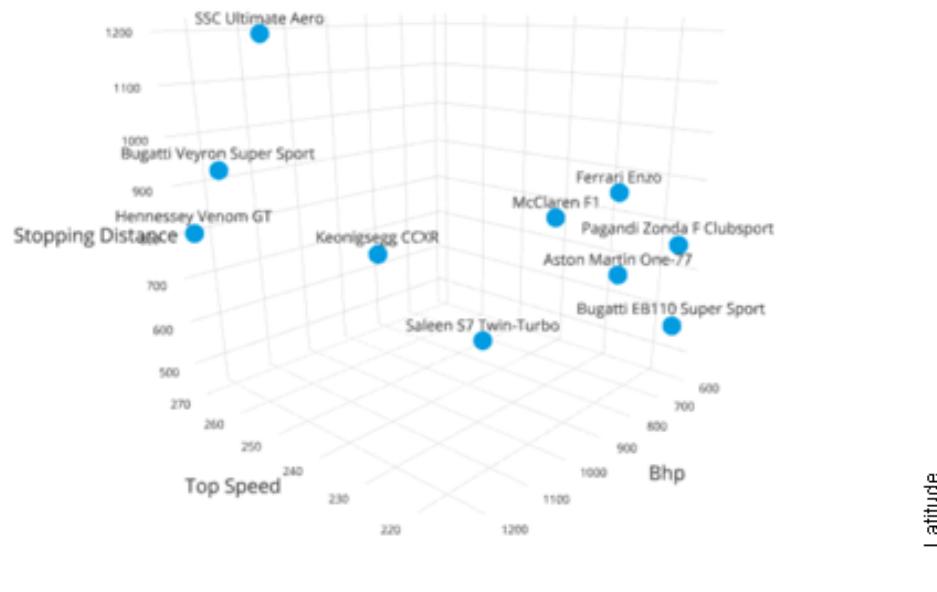
Bivariate Data

- Scatter plot is commonly used
- Each mark is now a data case
- Objective:
 - Two variables, want to see relationship
 - Is there a linear, curved or random pattern?

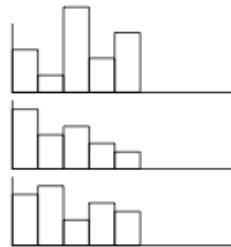
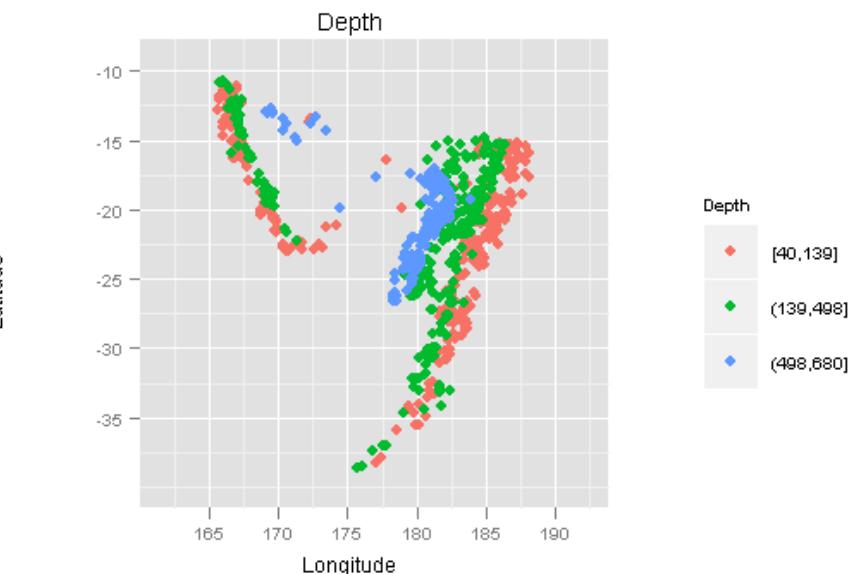


Trivariate Data

3D scatter plot

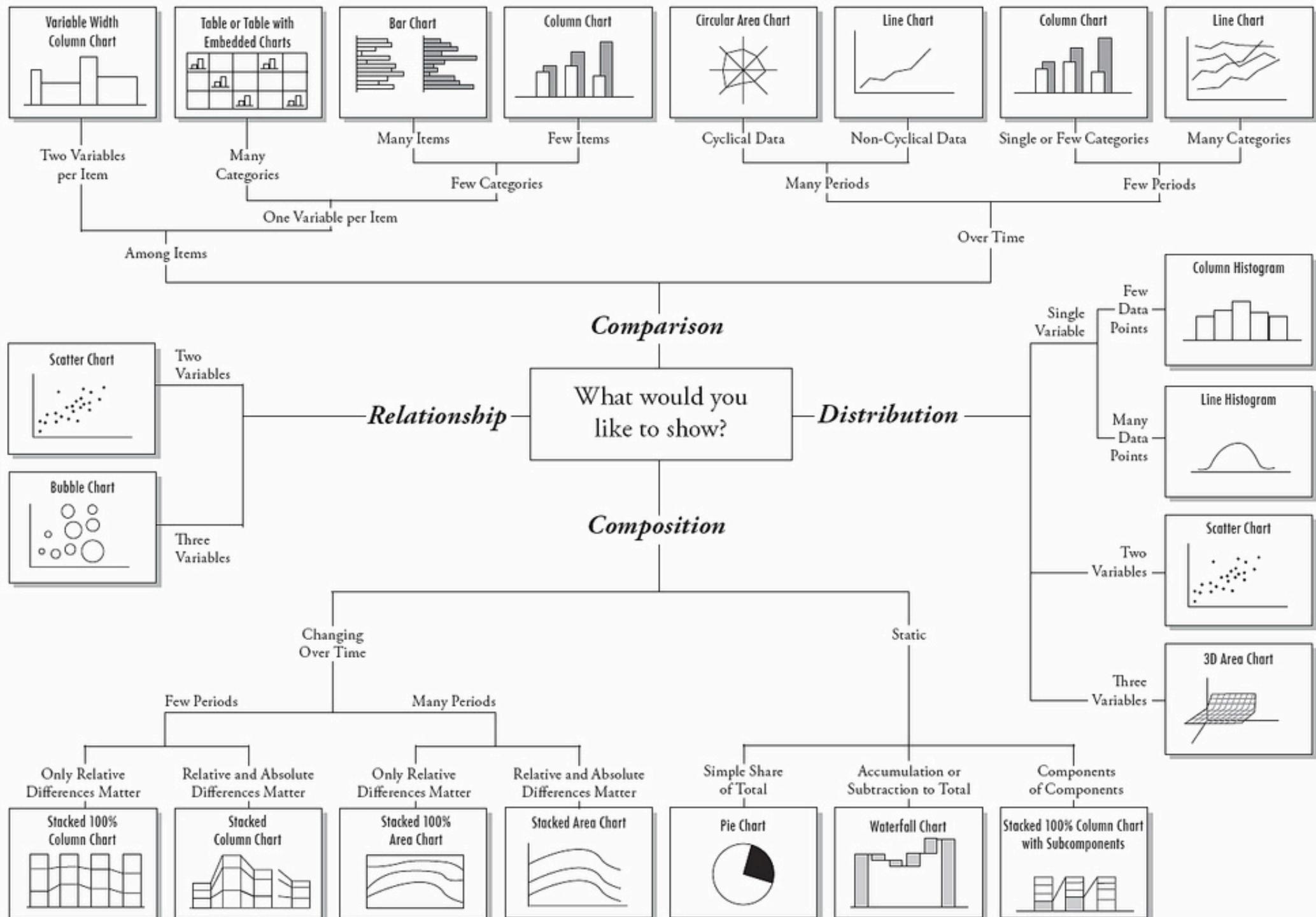


2D + mark
property



Represent each
variable in its own
explicit way

Chart Suggestions—A Thought-Starter



Data Visualization with ggplot2 :: CHEAT SHEET

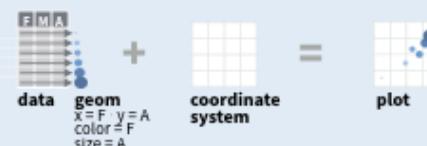


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data set**, a **coordinate system**, and **geoms**—visual marks that represent data points.



To display values, map variables in the data to visual properties of the geom (**aesthetics**) like **size**, **color**, and **x** and **y** locations.



Complete the template below to build a graph.

```
ggplot(data = <DATA>) + <GEOM_FUNCTION>
(mapping = aes(<MAPPINGS>),
stat = <STAT>, position = <POSITION>) +
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>
```

ggplot(data = mpg, aes(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

qplot(x = cty, y = hwy, data = mpg, geom = "point") Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.



Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

```
a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))
```

a + geom_blank()
(Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1,
xend = long + 1, curvature = z)) -> x, yend, y, yend,
alpha, angle, color, curvature, linetype, size

a + geom_path(linend = "butt", linejoin = "round",
linemetre = 1)
x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group))
x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin = lat, xmax =
long + 1, ymax = lat + 1)) -> xmin, ymin, ymax,
ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin = unemploy - 900,
ymax = unemploy + 900)) -> x, ymax, ymin,
alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

b + geom_abline(aes(intercept = 0, slope = 1))
b + geom_hline(aes(yintercept = lat))
b + geom_vline(aes(xintercept = long))

b + geom_segment(aes(yend = lat + 1, xend = long + 1))
b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian")
x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot()
x, y, alpha, color, fill

c + geom_frequpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight

c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

d <- ggplot(mpg, aes(fl))

d + geom_bar()
x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

e + geom_label(aes(label = cty), nudge_x = 1,
nudge_y = 1, check_overlap = TRUE) x, y, label,

alpha, angle, color, family, fontface, hjust,

lineheight, size, vjust

A
B
C

e + geom_jitter(height = 2, width = 2)
x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size

e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1,
nudge_y = 1, check_overlap = TRUE) x, y, label,

alpha, angle, color, family, fontface, hjust,

lineheight, size, vjust

discrete x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

f + geom_col(), x, y, alpha, color, fill, group, linetype, size

f + geom_boxplot(), x, y, lower, middle, upper,

ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight

f + geom_dotplot(binaxis = "y", stackdir =

"center")

x, y, alpha, color, fill, group

f + geom_violin(scale = "area"), x, y, alpha, color, fill, group, linetype, size, weight

discrete x , discrete y

g <- ggplot(diamonds, aes(cut, color))

g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

THREE VARIABLES

seals\$z <- with(seals, sqrt(delta_long^2 + delta_lat^2)) l <- ggplot(seals, aes(long, lat))

l + geom_contour(aes(z = z))
x, y, z, alpha, colour, group, linetype, size, weight

l + geom_raster(aes(fill = z)), hjust = 0.5, vjust = 0.5,

interpolate = FALSE)

x, y, alpha, fill

l + geom_tile(aes(fill = z)), x, y, alpha, color, fill, linetype, size, width

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

h + geom_bin2d(binwidth = c(0.25, 500))
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

continuous function

i <- ggplot(economics, aes(date, unemploy))

i + geom_area()
x, y, alpha, color, fill, linetype, size

i + geom_line()
x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv")
x, y, alpha, color, group, linetype, size

visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit - se, ymax = fit + se))

j + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, group, linetype, size

j + geom_errorbar(), x, y, ymax, ymin, alpha, color, group, linetype, size
(also **geom_errorbarh()**)

j + geom_linerange()
x, y, ymin, ymax, alpha, color, group, linetype, size

j + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size, weight

visualizing error

data <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
state = tolower(rownames(USArrests)))

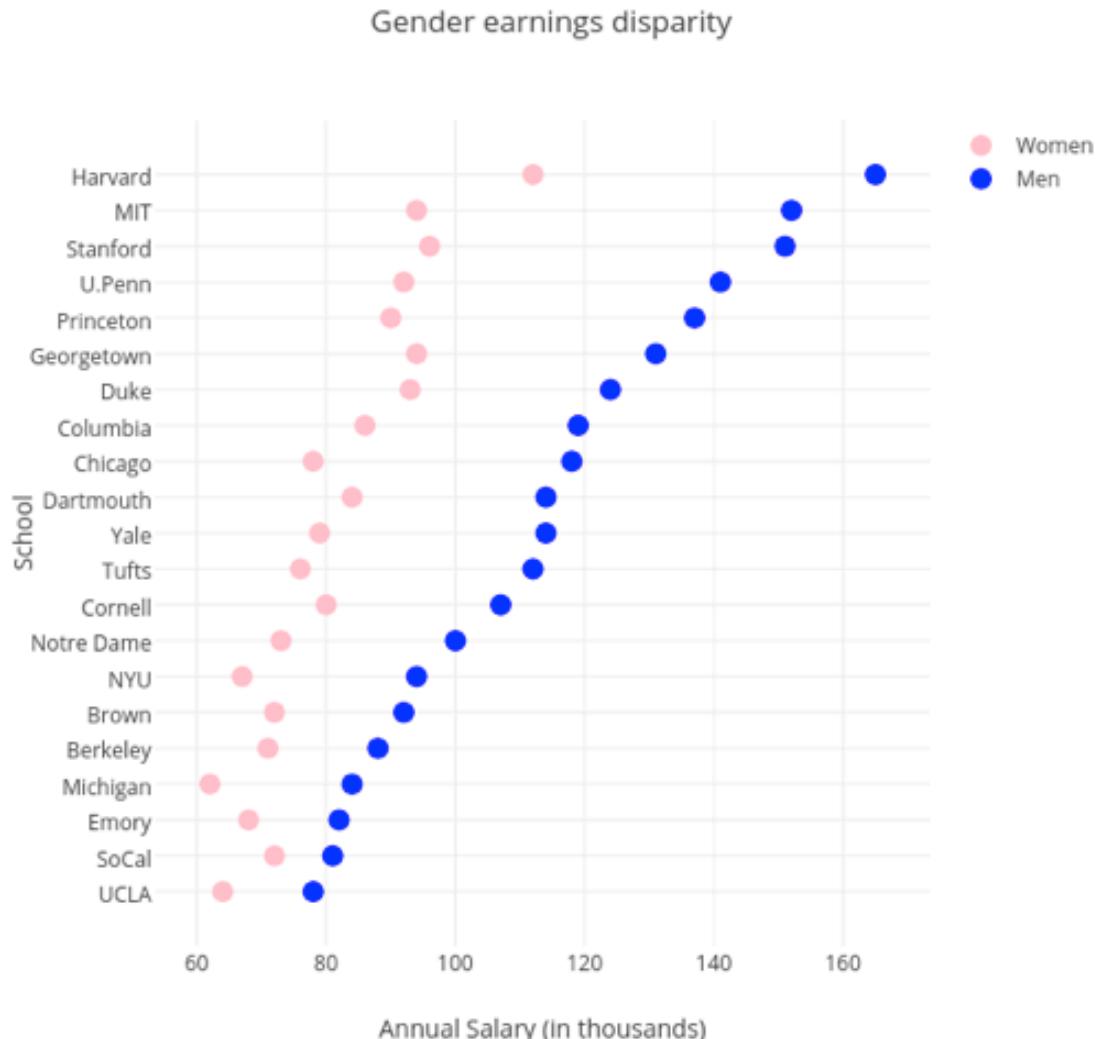
map <- map_data("state")

k <- ggplot(data, aes(fill = murder))

k + geom_map(aes(map_id = state), map = map)
+ expand_limits(x = map\$long, y = map\$lat), map_id, alpha, color, fill, linetype, size

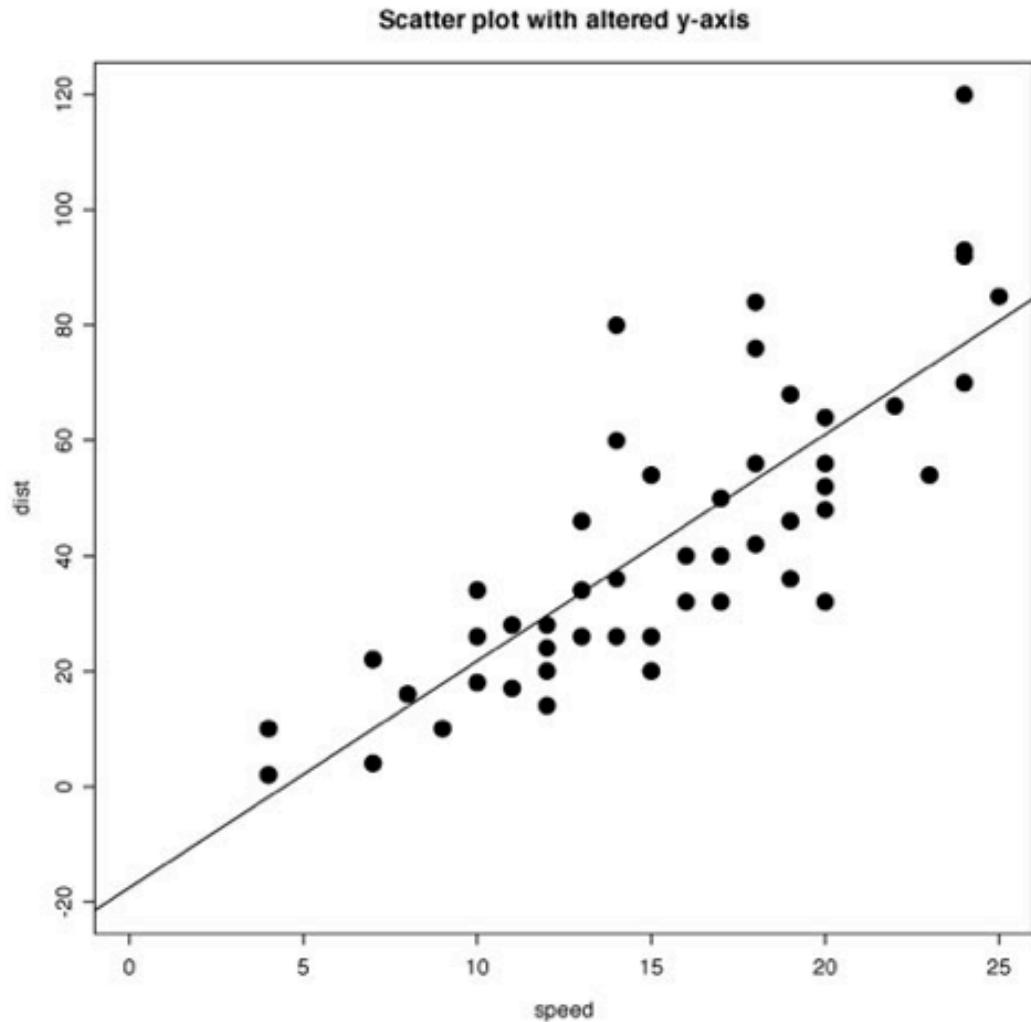
Dot Plots

- When to use:
 - When analyzing values that are spaced at irregular intervals
 - continuous, quantitative, univariate data



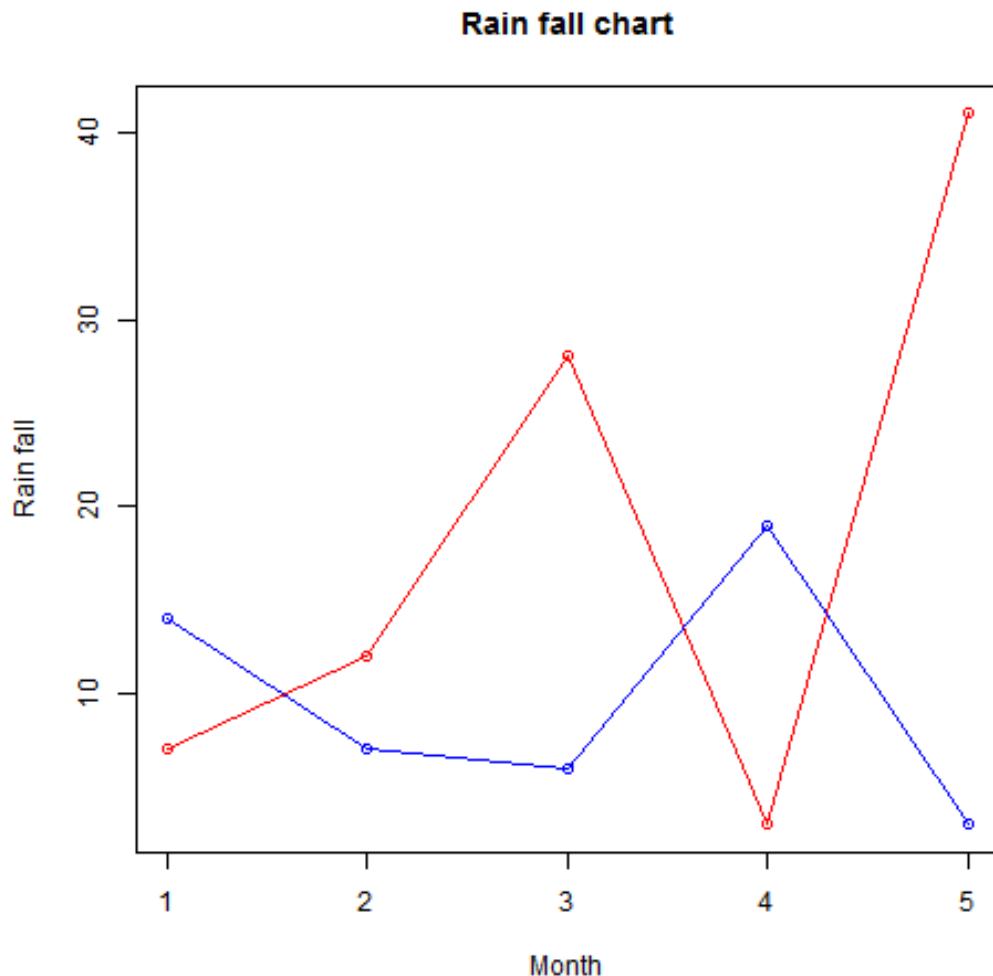
Scatter Plot

- When to use:
 - To compare how two quantitative variables change
 - continuous, quantitative, bivariate data
 - relationships for two variables



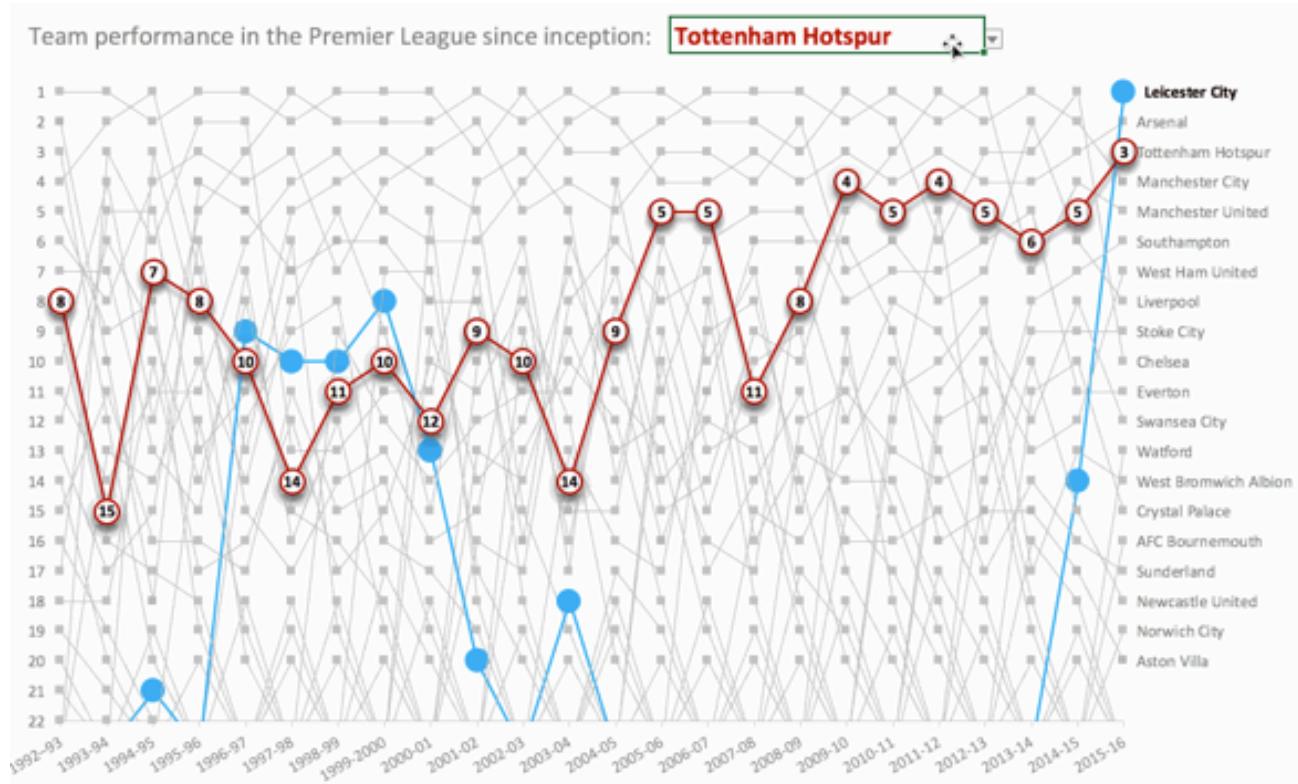
Line Graphs

- When to use:
 - When quantitative values change during a continuous period of time (for more than one group)
 - Time series data
 - Non-cyclical data over time



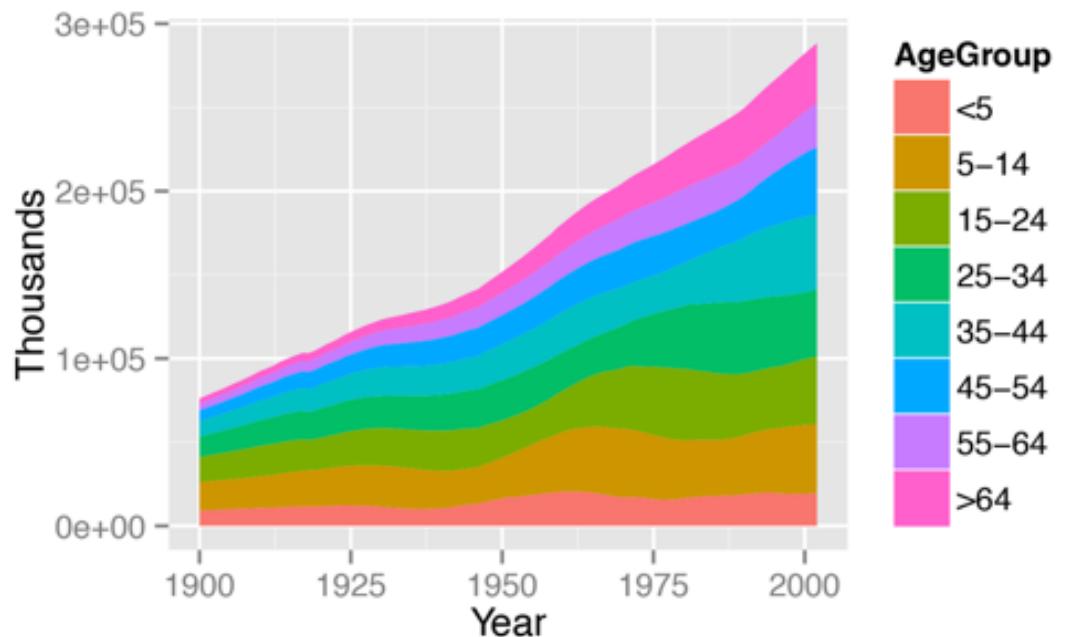
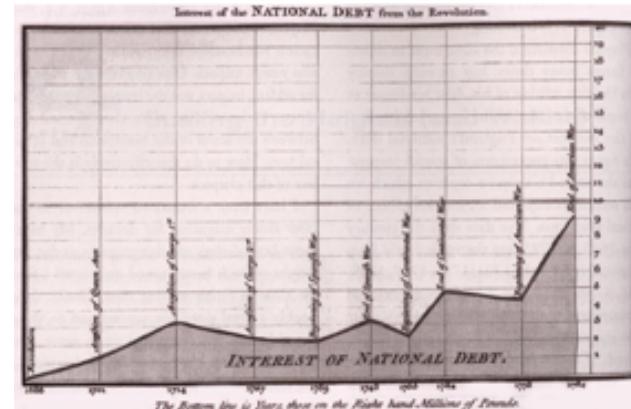
Bump Chart

- When to use:
 - Similar to line graph
 - Y-axis: rank rather than (continuous) values



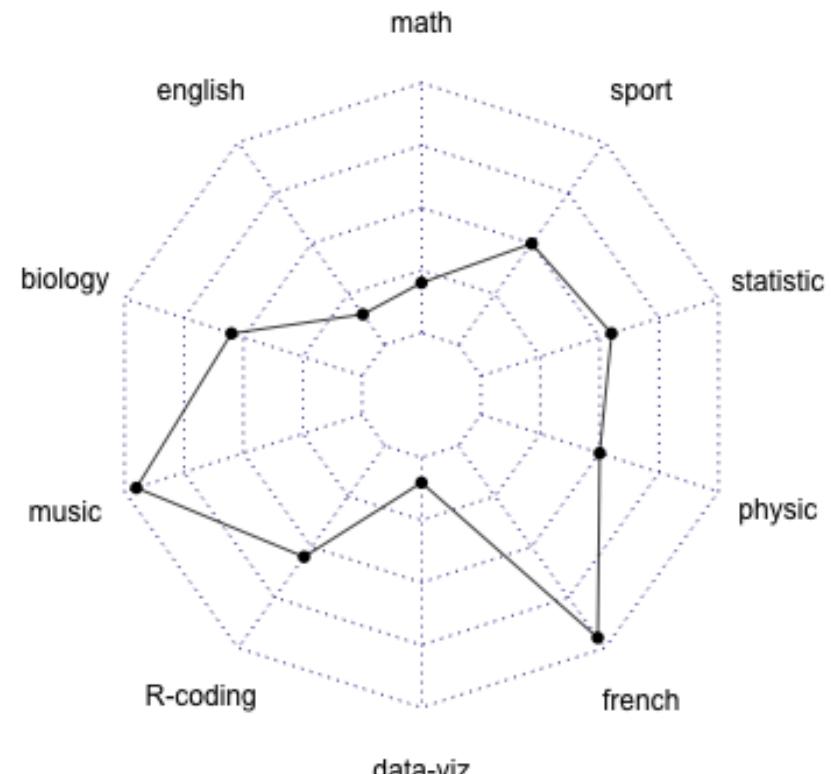
Area Graph

- When to use:
 - Commonly one compares with an area chart two or more quantities.
 - The area between axis and line are commonly emphasized with colors and textures.



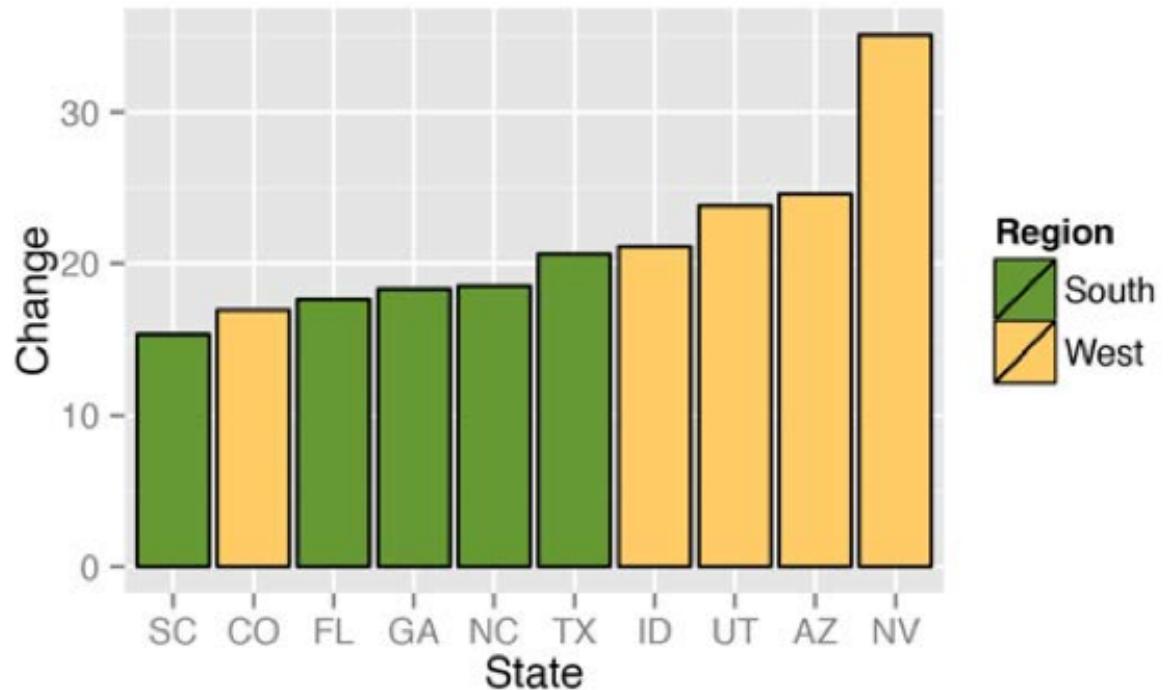
Radar Graphs

- When to use:
 - When you want to represent data across the cyclical nature of time
 - A two-dimensional chart of three or more quantitative variables represented on axes starting from the same point



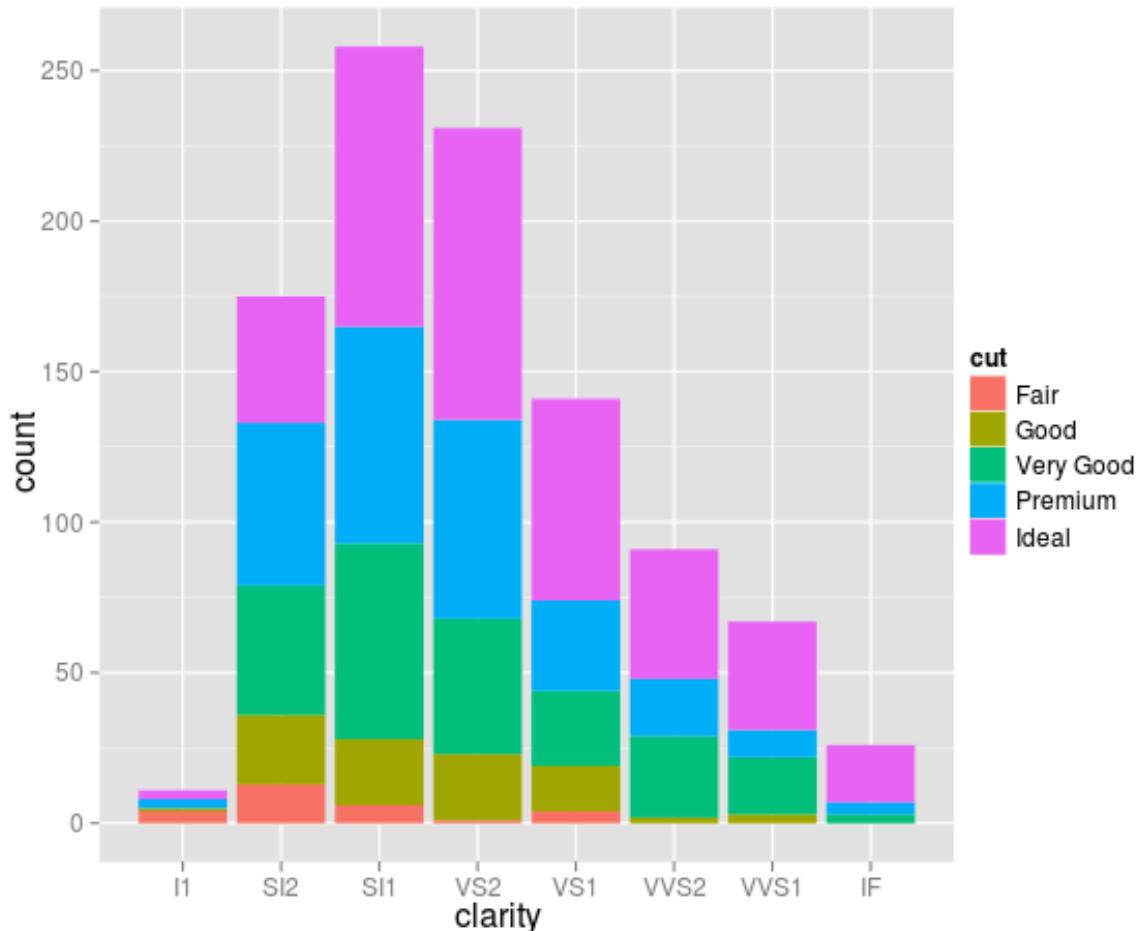
Bar Graphs

- When to use:
 - When you want to support the comparison of individual values between different groups
 - Can run vertically or horizontally



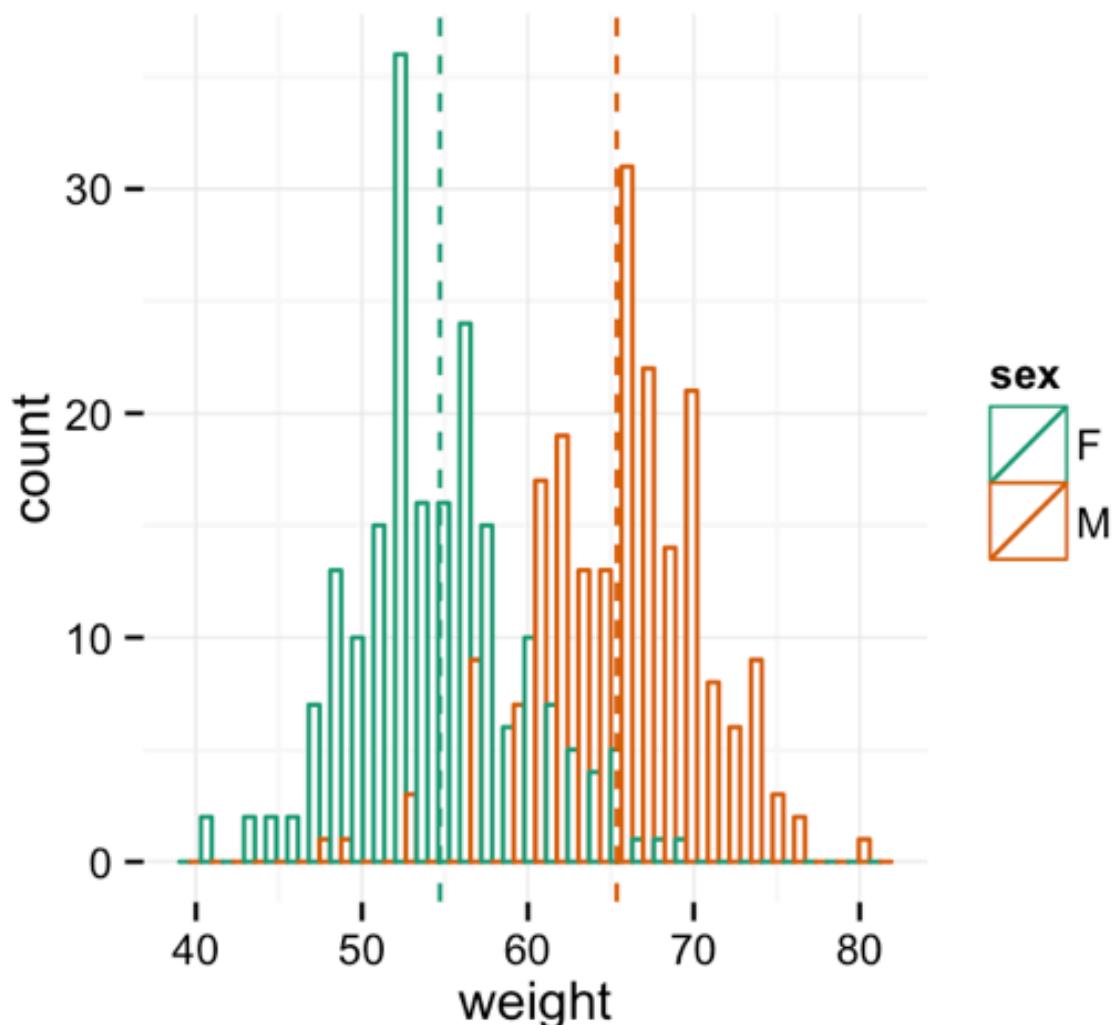
Stacked Bar Chart

- When to use:
 - When you want to present the total in a clear way while comparing part-to-whole relationship between different groups
 - Harder to compare the size of each categories



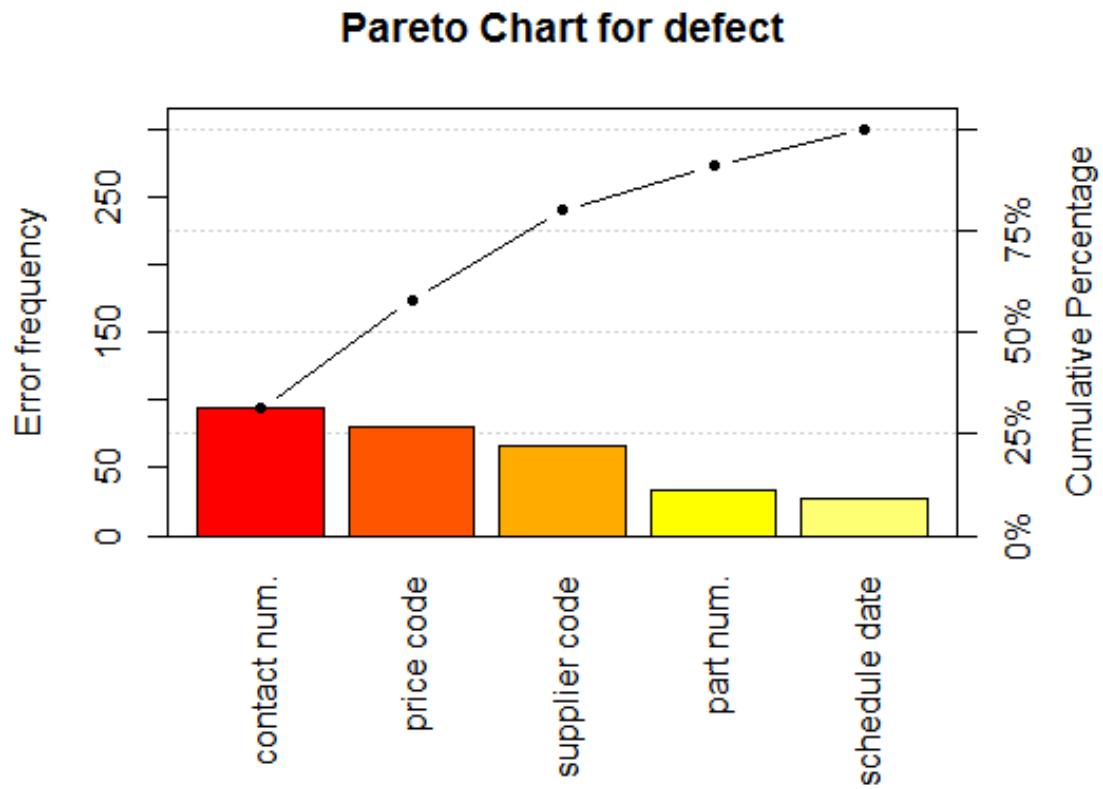
Histogram

- When to use:
 - the most commonly used graph to show frequency distributions
 - Continuous, quantitative, univariate data



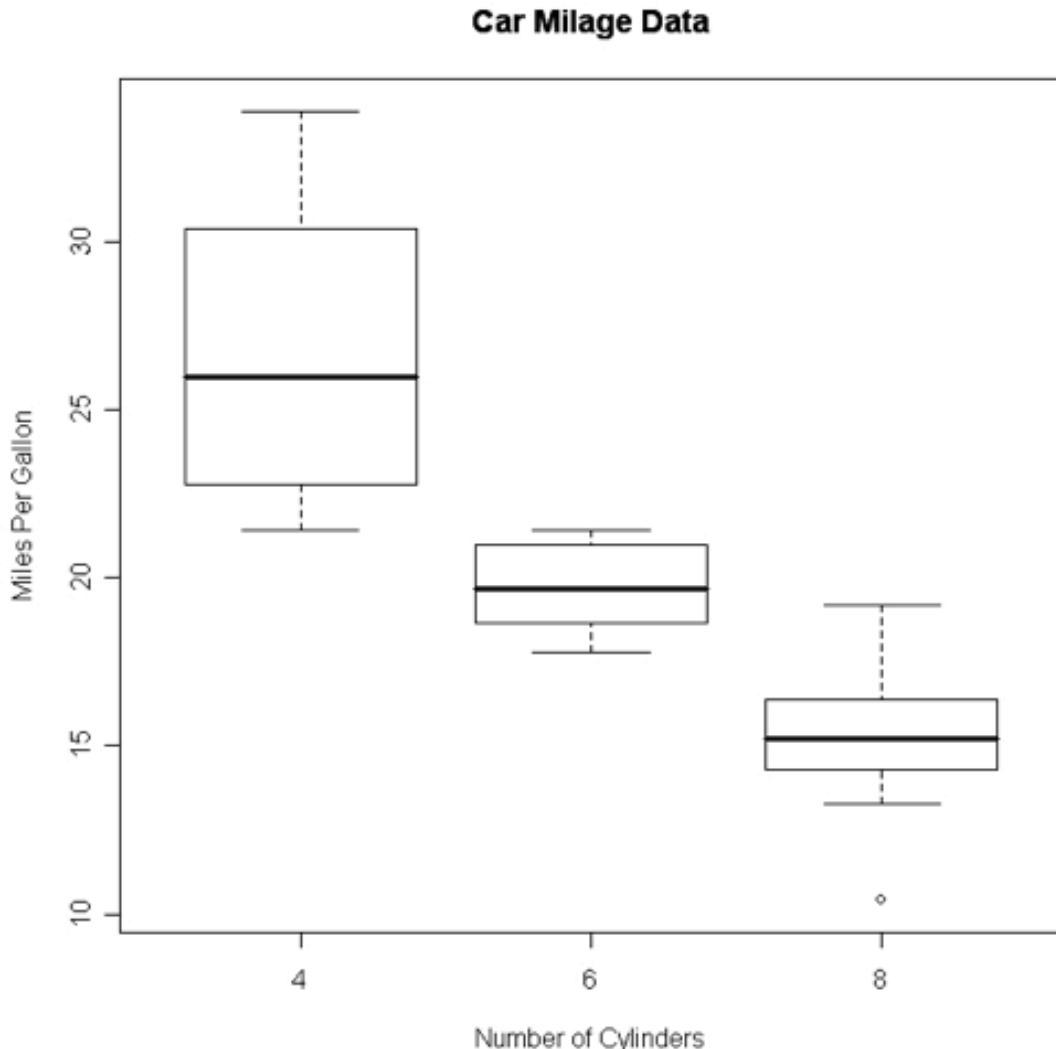
Pareto chart

- When to use:
 - When analyzing data about the frequency of problems or causes in a process.
 - containing both bars and a line graph



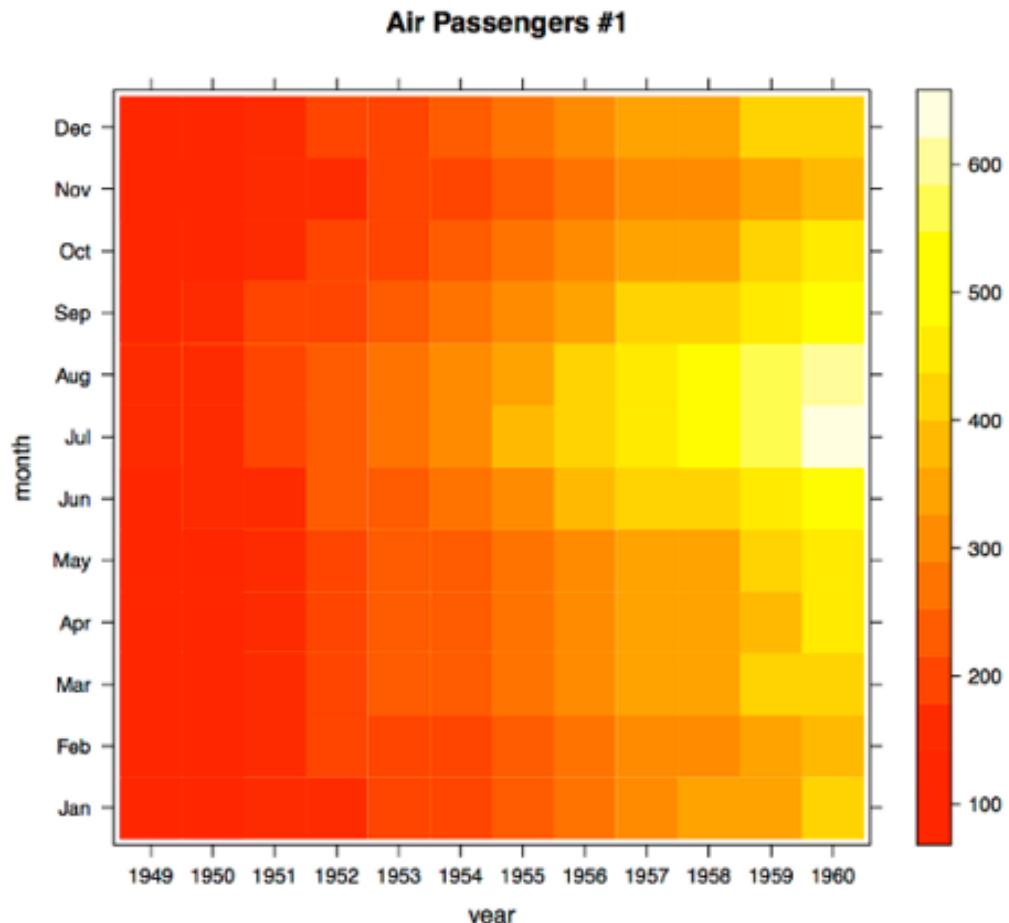
Box Plots

- When to use:
 - You want to show allow for comparison of data from different categories
 - graphically depicting groups of numerical data through their quartiles



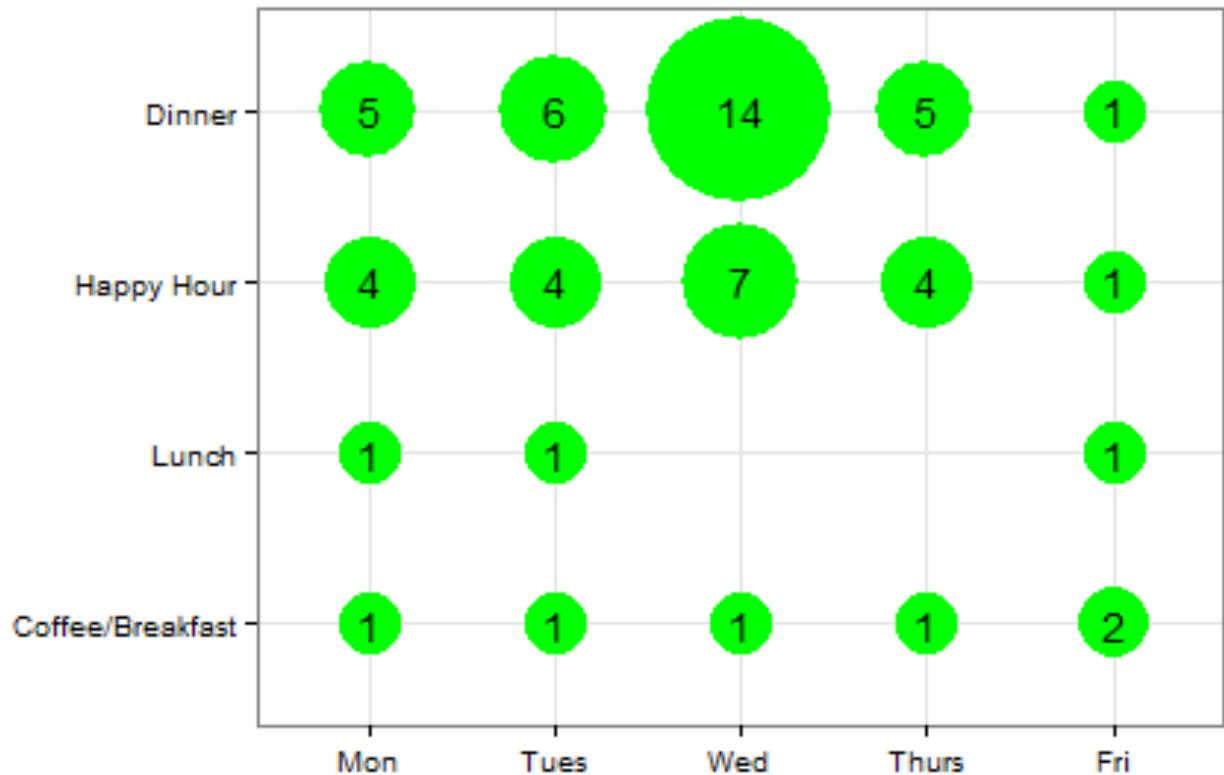
Heat Maps

- When to use:
 - When you want to display a large quantity of cyclical data (too much for radar)
 - Color choices: grayscales, rainbow, etc.



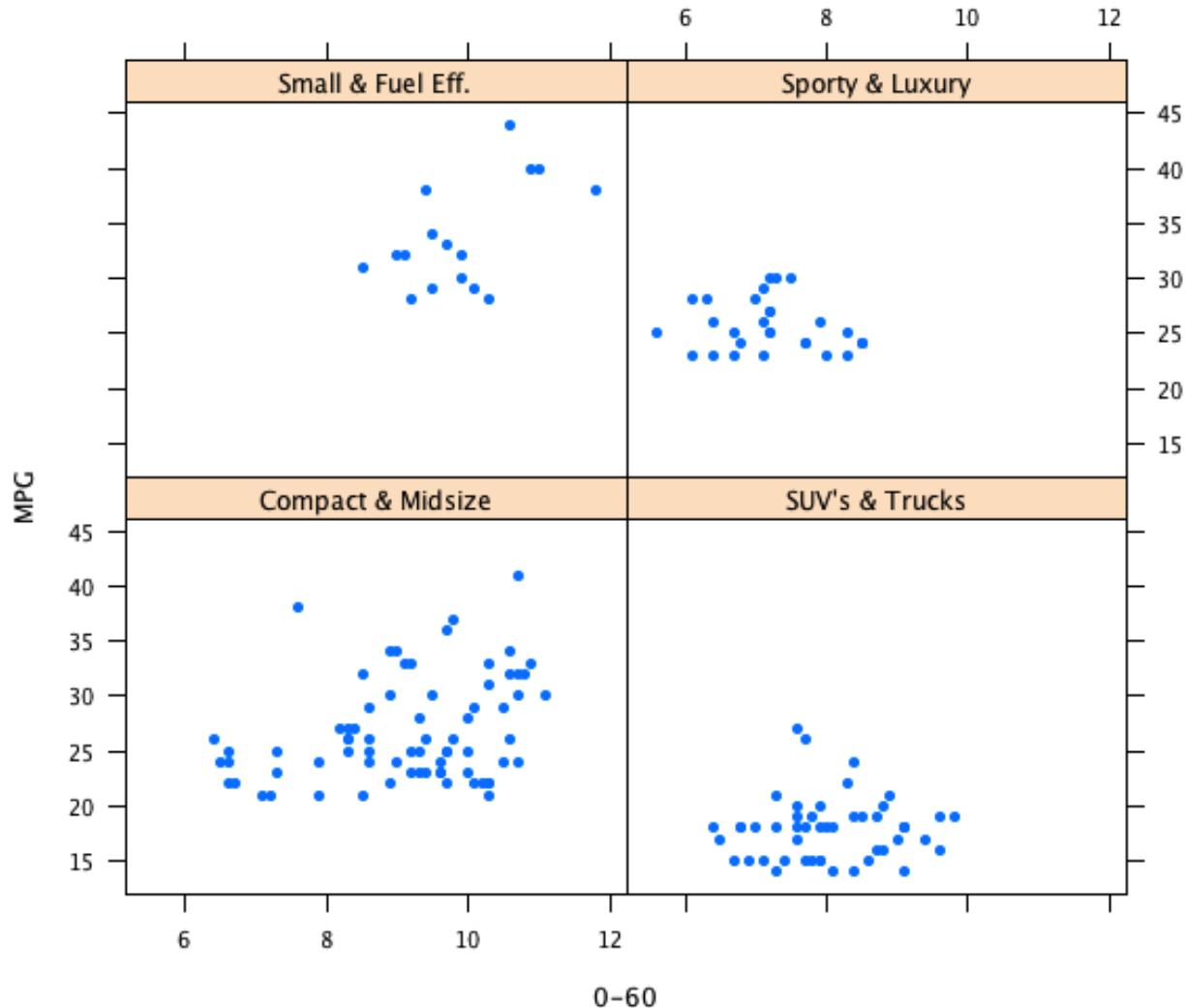
Crosstab Plot

- When to use:
 - Comparing different groups while presenting values (count)
 - Similar to heatmap



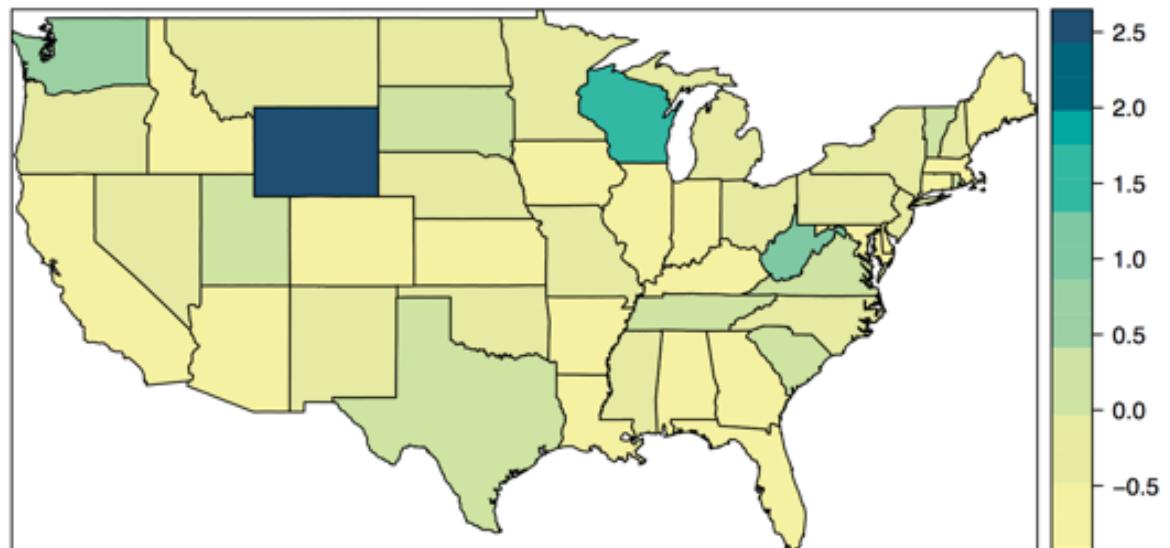
Trellis Display

- When to use:
 - Typically varies on one variable
 - Distribute different values of that variable across views

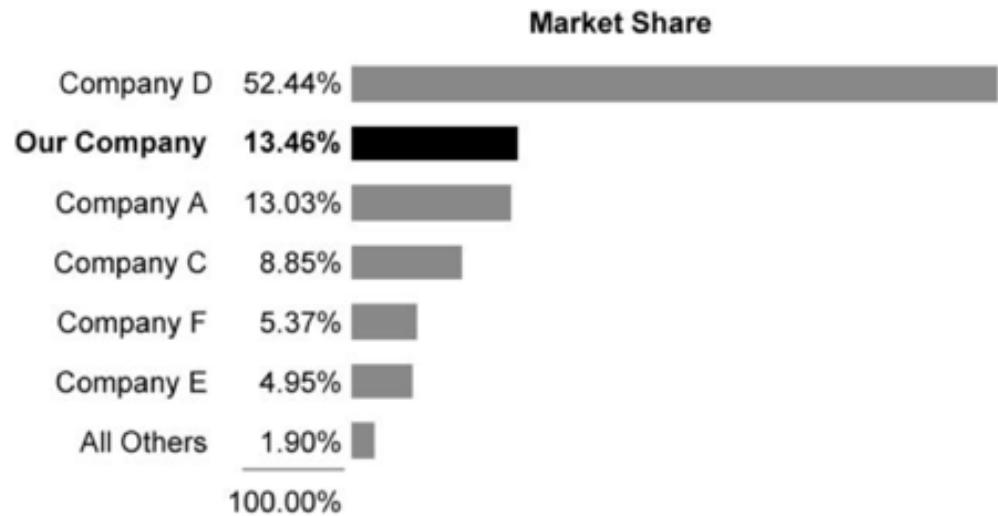
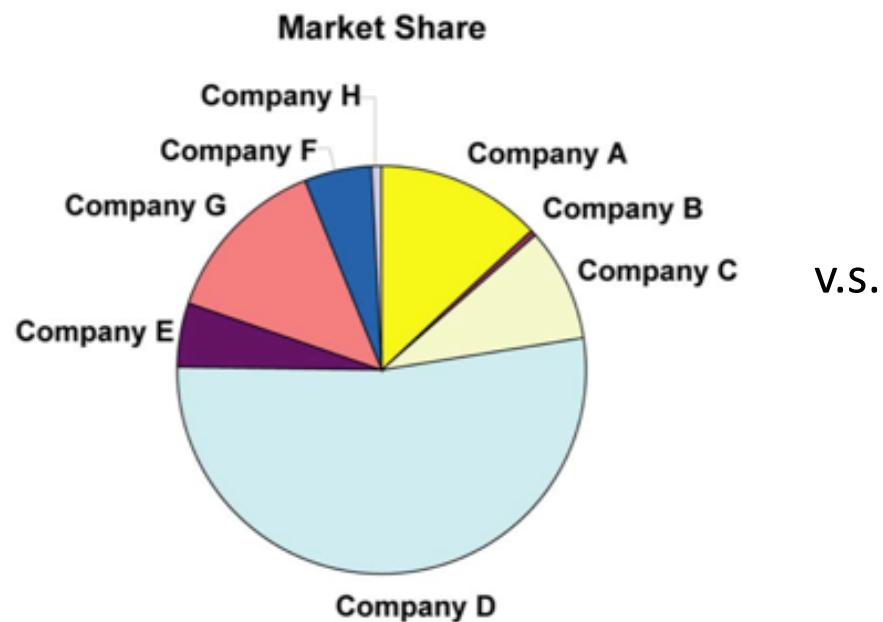


Hybrid: Map based Heatmap

- When to use:
 - When you want to display a large quantity of cyclical data over different geo-locations

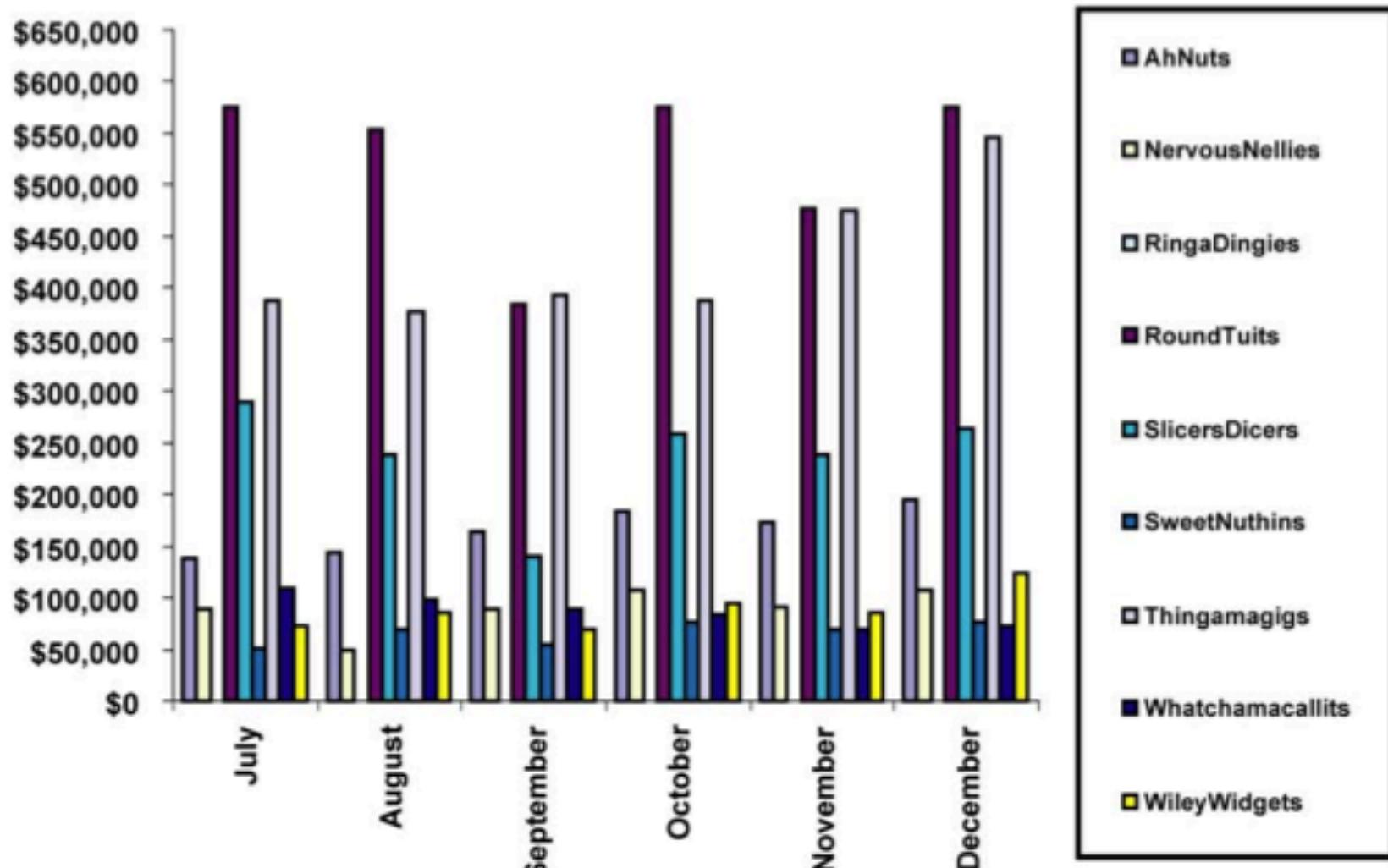


Comparisons

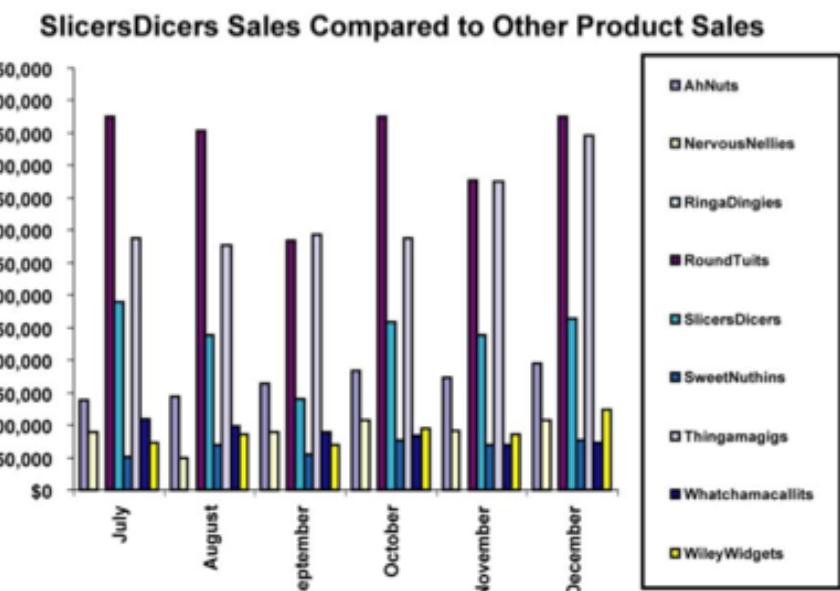


Comparisons

SlicersDicers Sales Compared to Other Product Sales



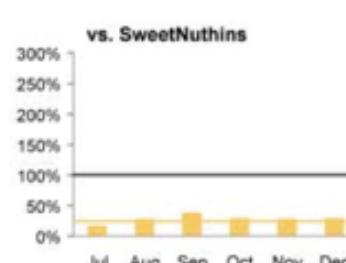
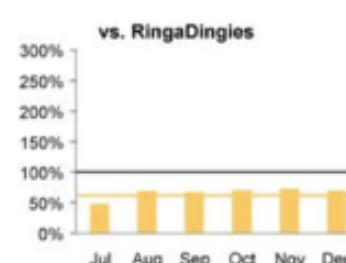
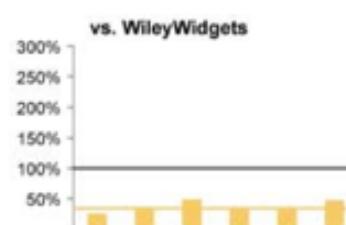
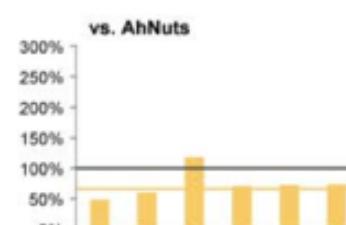
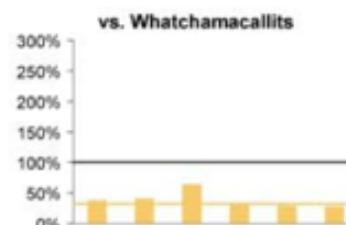
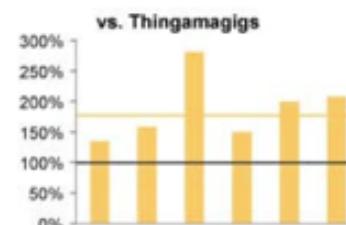
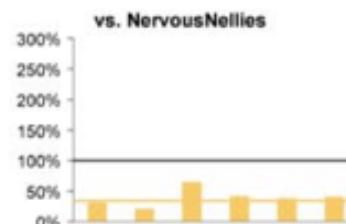
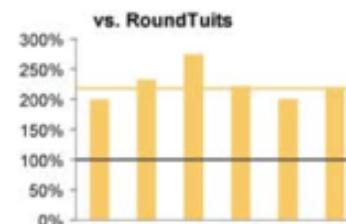
Comparisons



V.S.

Sales of SlicersDicers Compared to Sales of Other Products
July - December, 2003

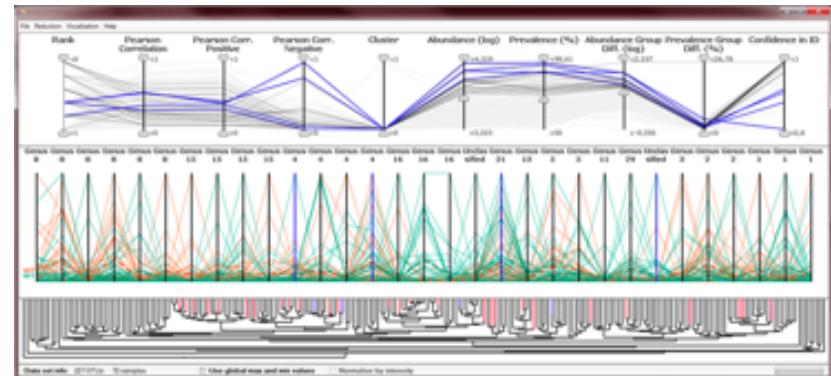
(SlicersDicers' sales are displayed as black reference lines of 100%; the orange lines represent the average monthly sales for July through December.)



Next Lecture

- Topic: Multivariate Data Visualization

- Next Monday (11 Feb)
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus



G53FIV: Fundamentals of Information Visualization

Lecture 5: Multivariate Data Visualization

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Announcement

- Change of Time on the Optional Lab Sessions from next week
- **Monday, 9:00 - 10:00, A32 Computer Science**
- Feb 18: Course Work Case Study (Optional)
 - one session at 9:00 - 10:00 at CS-A32.
 - one session at 14:00 - 15:00 as well at Business South, A25.
- Feb 25, Mar 4, Mar 11
 - Lab Computing Sessions (Optional) in CS-A32

Announcement

- Issue of Course Work
- Please check Moodle for details
- Due date: **April 8 2019**
 - report of maximum 10 pages (3000 words).
 - R codes

Last Lecture

Data and Image

Data Models

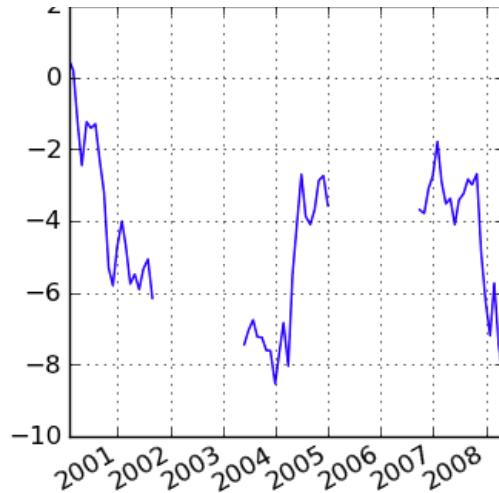
- N, O, Q?
- Dimension or Measure?

– Year	Q-Internal (O)	Dimension
– Age	Q-Ratio (O)	Depends
– Marital	N	Dimension
– Sex	N	Dimension
– People	Q-Ratio	Measure

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080

Data Processing

- Data cleaning and filtering
 - for quality control
 - Remove (Outlier, missing data)
 - Modify (conversion of format, etc.)
- Data adjustment
 - Depends on your task and questions to ask
 - Relational algebra:
 - e.g. Aggregation, mean, sort, projection
 - Reformatting and Integration

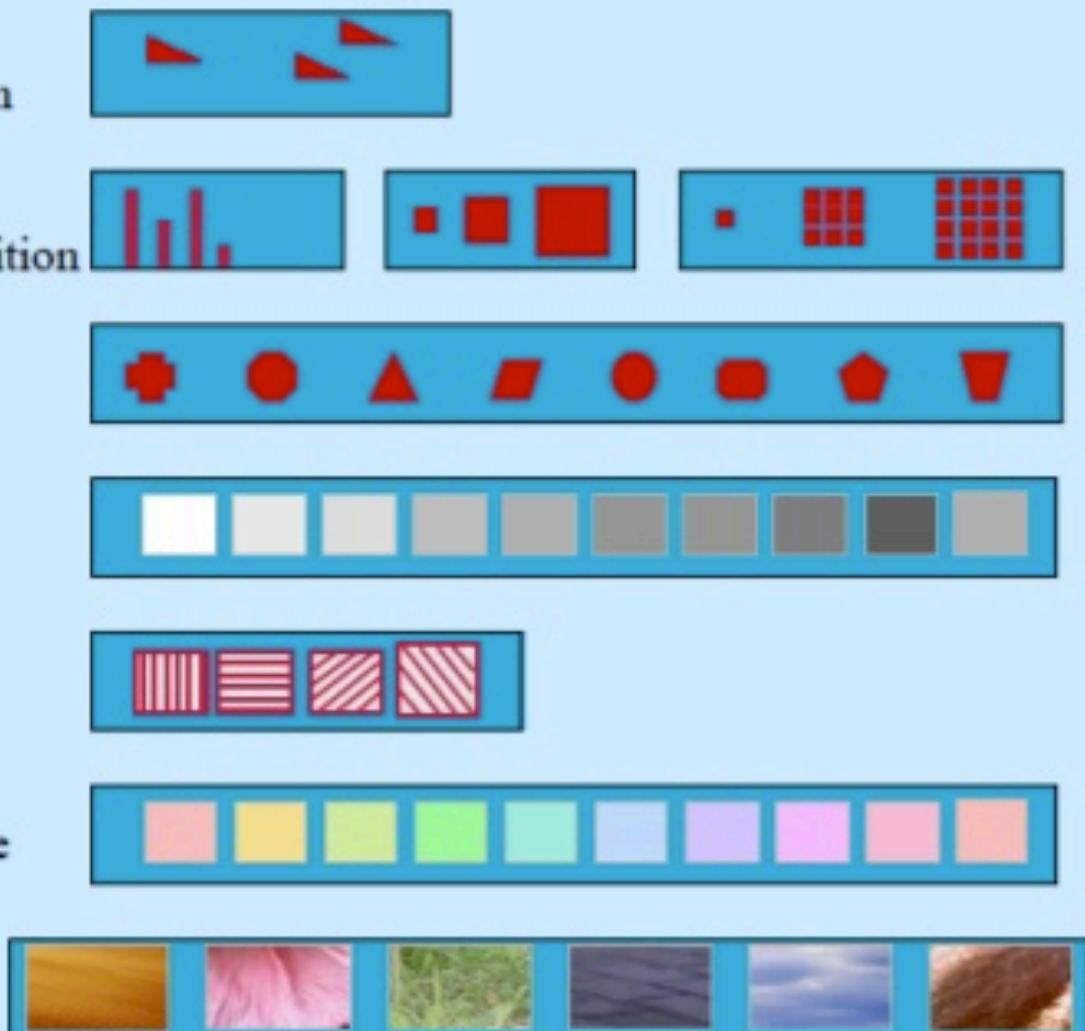


*We will learn later how
to do these in R.*

Image: Visual Encoding Variables

Bertin's Semiology of Graphics (1967)

- **position**
 - changes in the x, y, (z) location
- **size**
 - change in length, area or repetition
- **shape**
 - infinite number of shapes
- **value**
 - changes from light to dark
- **orientation**
 - changes in alignment
- **colour**
 - changes in hue at a given value
- **texture**
 - variation in pattern
- **motion**



Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Levels of Organization

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good

~ = OK

✗ = Bad

Last Lecture

Design and Graphs

What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language **express** (1) **all the facts** in the set of data, and (2) **only the facts** in the data.

Tell the truth

- **Effectiveness**

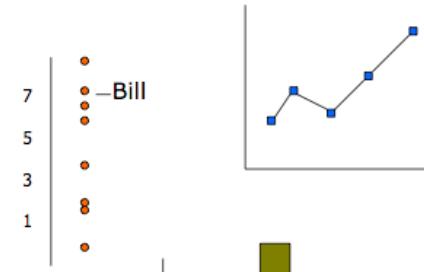
- A visualization is more effective than another visualization if the information conveyed by one visualization **is more readily perceived** than the information in the other visualization.

Use proper encoding

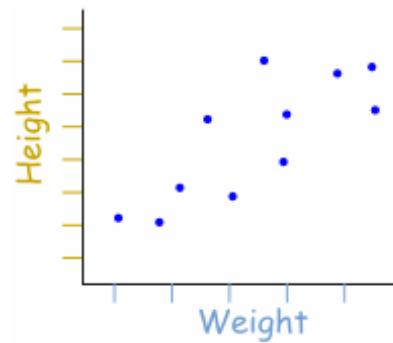
Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Graphs

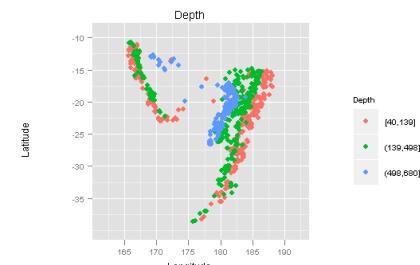
- Data Dimensions
 - 1 - Univariate data
 - 2 - Bivariate data
 - 3 - Trivariate data
 - >3 - Hypervariate data
- Data Types
 - Nominal, Ordinal, Quantitative
- Visualization Representations
 - Points, Lines, Bars, Boxes



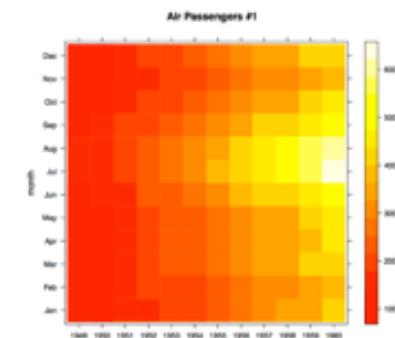
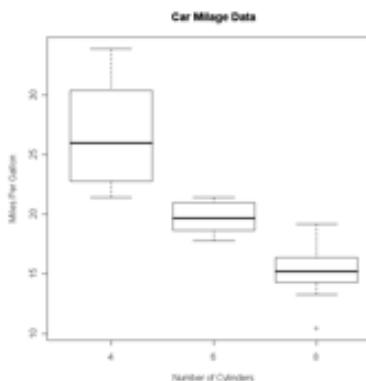
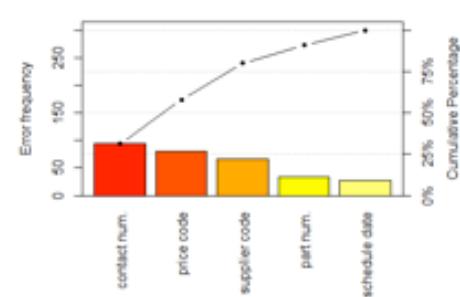
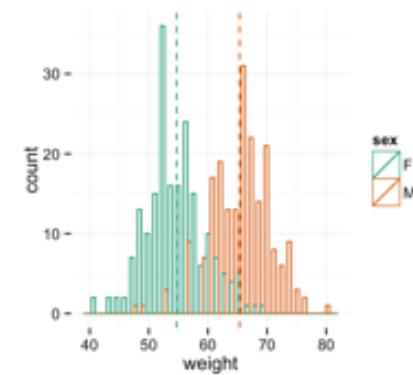
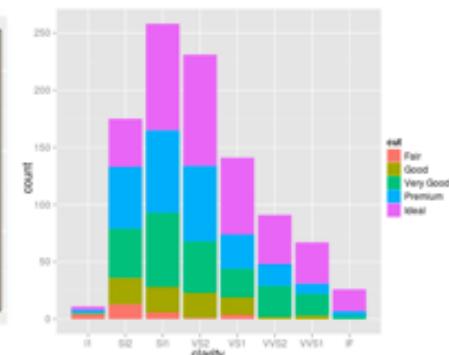
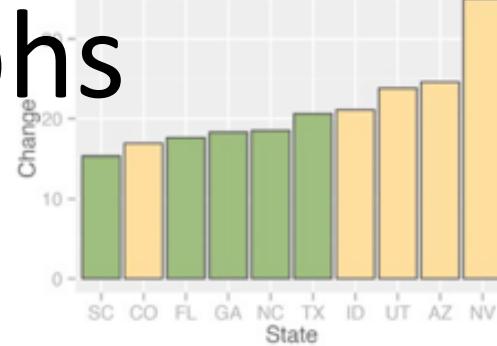
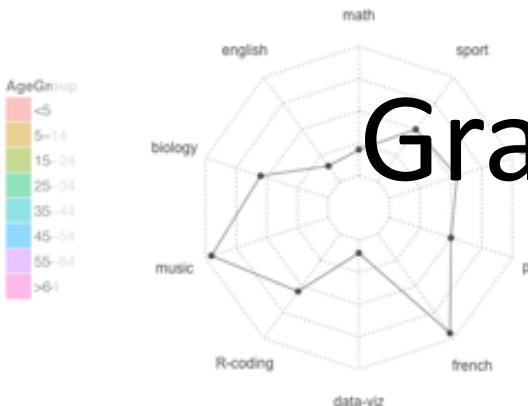
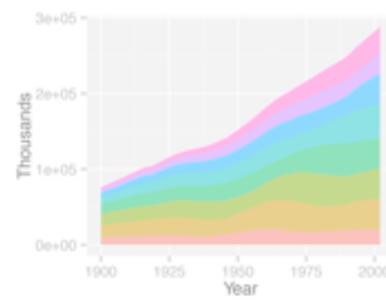
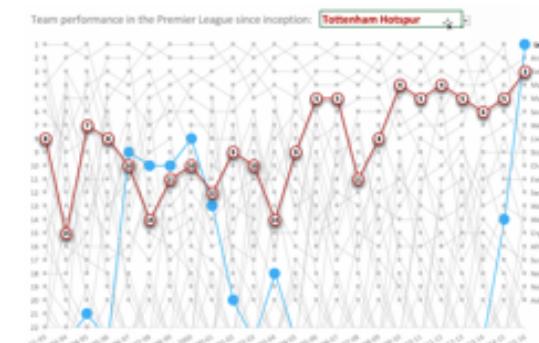
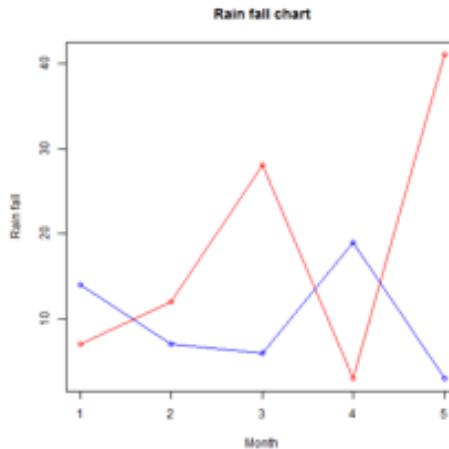
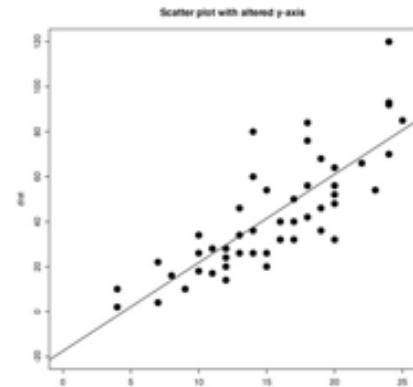
Univariate



Bivariate



Trivariate



Overview

- Multivariate Data Visualization Design Challenge
 - hypervariate data – our focus
- Common Multivariate Data Visualization Techniques

Design Challenge

- Data about dogs (hypervariate data)
 - Variety N
 - Group N
 - Size O
 - Smartness N
 - Popularity Q
 - Ranking Q
- Design a visualization for this multivariate data

Intuition

- Fundamentally, we have 2 geometric (position) display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane

Representation

- What are two main ways of presenting multivariate data sets?
 - Directly (textually): Tables
 - Symbolically (pictures): Graphs
- When use which?

Table / Spreedsheet

- A spreadsheet (table) already does that
 - Each variable is positioned into a column
 - Data cases in rows
 - This is a projection (mapping)

Name	Economy	Cylinders	Displacement	Horsepower
Mazda RX4	21	6	160	110
Mazda RX4 Wag	21	6	160	110
Datsun 710	22.8	4	108	93
Hornet 4 Drive	21.4	6	258	110
Hornet Sportabout	18.7	8	360	175
Valiant	18.1	6	225	105
Duster 360	14.3	8	360	245
Merc 2400	24.4	4	146.7	62
Merc 230	22.8	4	140.8	95
Merc 280	19.2	6	167.6	123
Merc 280C	17.8	6	167.6	123
Merc 450SE	16.4	8	275.8	180
Merc 450SL	17.3	8	275.8	180
Merc 450SLC	15.2	8	275.8	180
Cadillac Fleetwood	10.4	8	472	205
	---	-	---	---

Limitations

- Occupy large space
- Difficult to understand the relationships
- Hard to see the overall picture, focus and see the context
- Less effective in amplifying human perception and cognition

When to use?

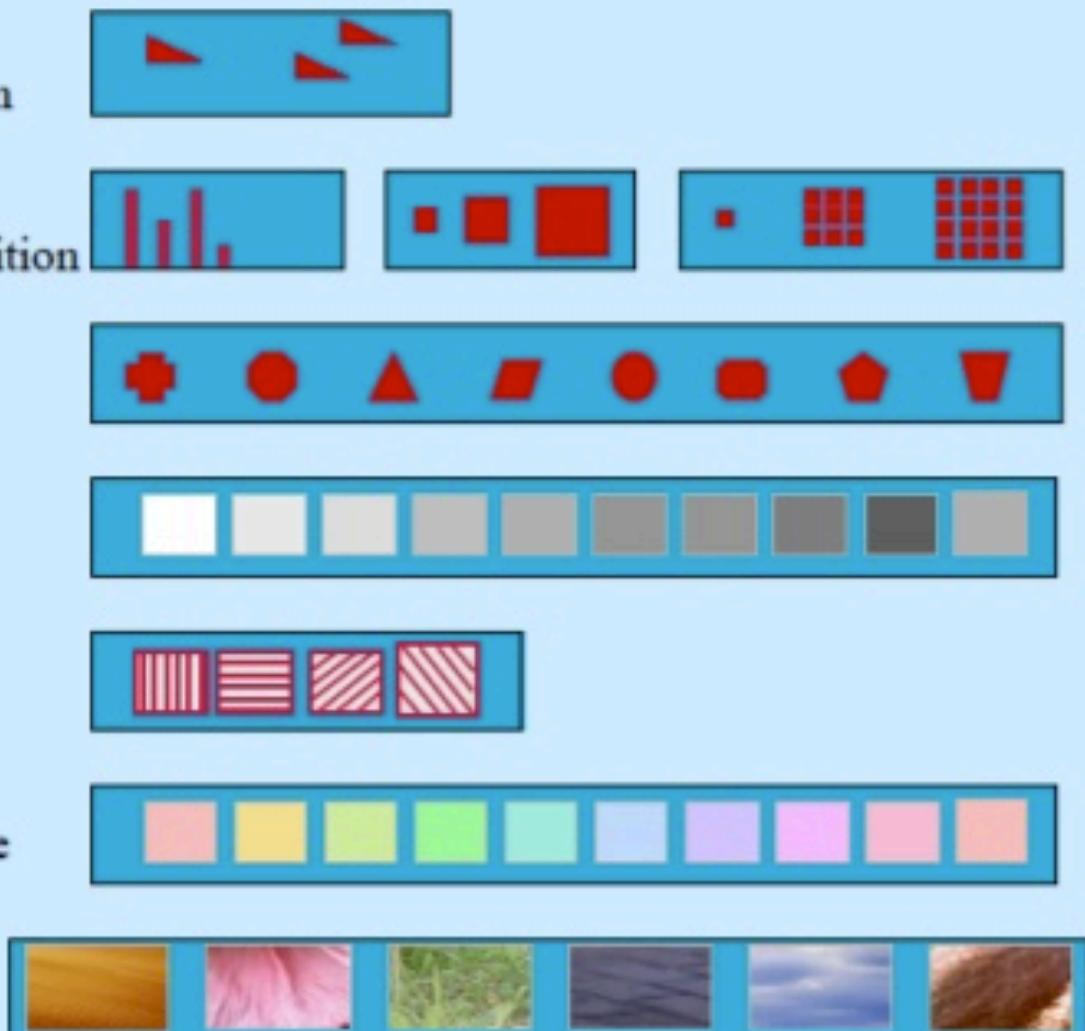
- Use tables when
 - The document will be used to **look up individual values**
 - The document will be used to **compare individual values**
 - **Precise values** are required
 - The quantitative info to be communicated involves **more than one unit of measure**
- Use graphs when
 - The message is contained in the **shape** of the values
 - The document will be used to **reveal relationships** among values
 - Especially useful when **the number of data points is huge**

(Optional Reading) Stephen Few. 2012. Show Me the Numbers: Designing Tables and Graphs to Enlighten (2nd ed.). Analytics Press, , USA.

Image: Visual Encoding Variables

Bertin's Semiology of Graphics (1967)

- **position**
 - changes in the x, y, (z) location
- **size**
 - change in length, area or repetition
- **shape**
 - infinite number of shapes
- **value**
 - changes from light to dark
- **orientation**
 - changes in alignment
- **colour**
 - changes in hue at a given value
- **texture**
 - variation in pattern
- **motion**

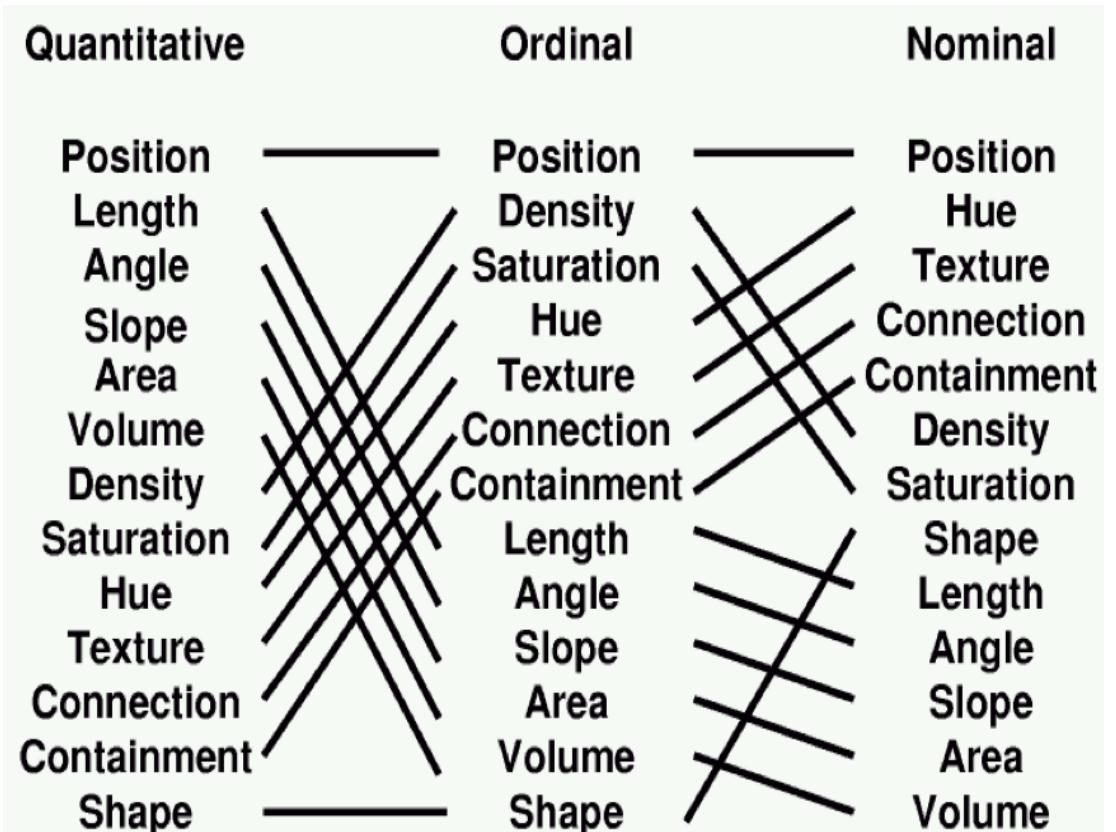


Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

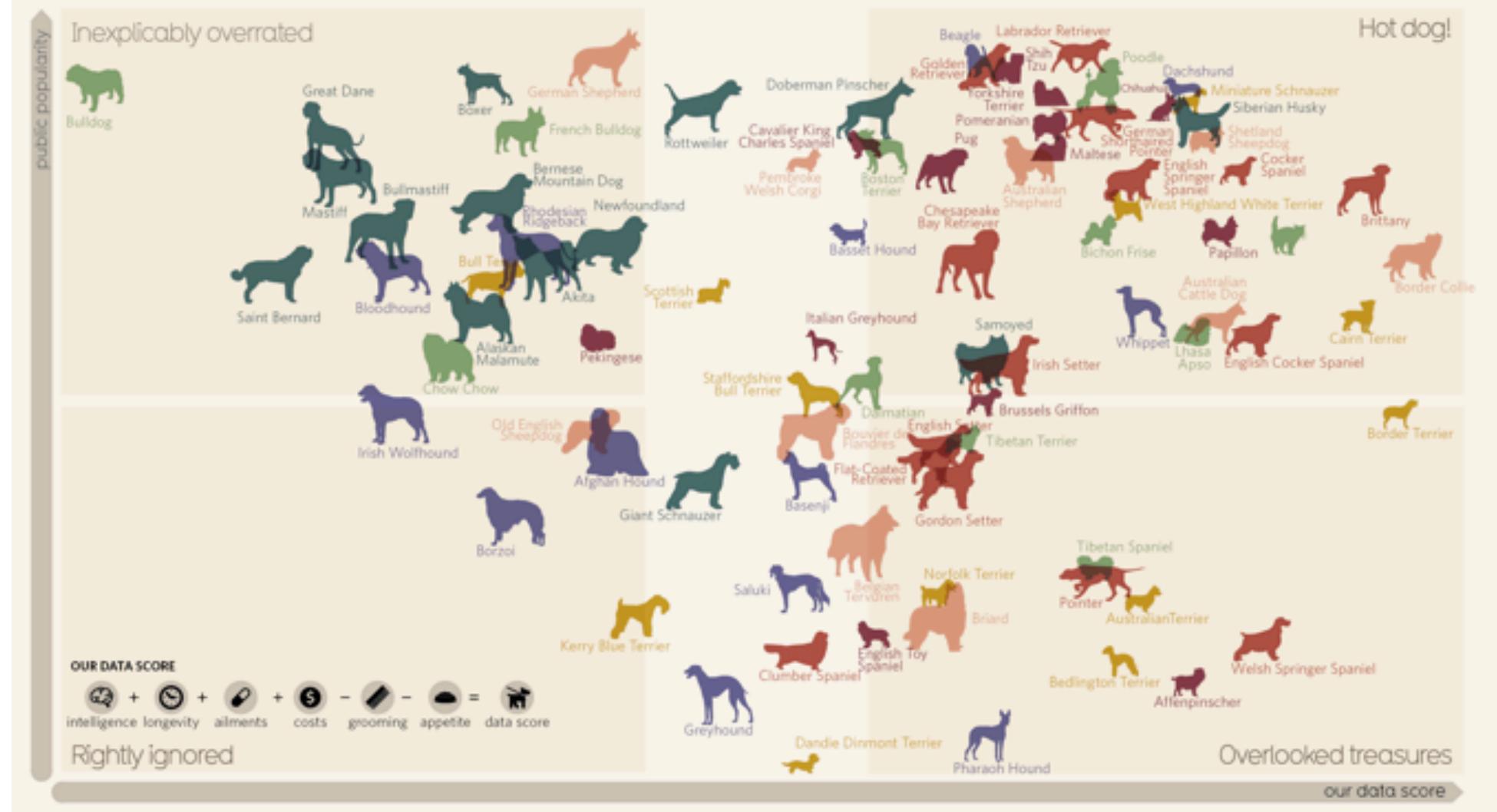
Design Challenge

- Data about dogs (hypervariate data)
 - Variety N
 - Group N
 - Size O
 - Smartness N
 - Popularity Q
 - Ranking Q
- Design a visualization



Best in Show

The ultimate data-dog



- Iconic Representations: Glyph (graphical object) represents a data case
- Visual properties of glyph represent different variables

Multivariate Data Visualization

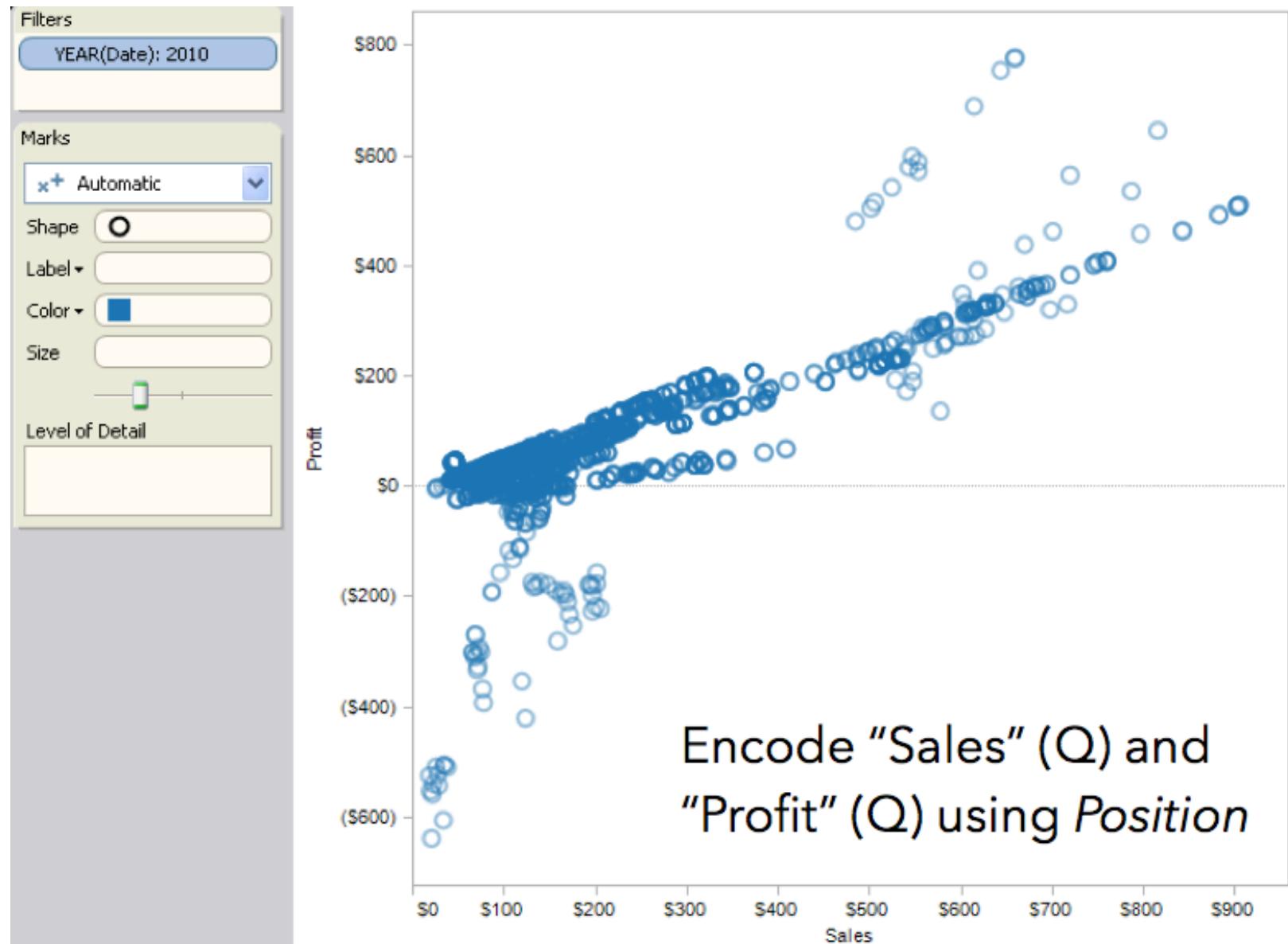
- Visual Encodings: 8 dimensions?
- Focus: techniques can generally handle all data sets

Visual Variables	Characteristics				
	Selective	Associative	Quantitative	Order	Length
<i>Position</i>	• .	••• .••	↑	↑	Theoretically Infinite
<i>Size</i>	• ●	•●●●●		●●●●●●●●●●	Selection: ~5 Distinction: ~20
<i>Shape</i>					Theoretically Infinite
<i>Value</i>	○●○○○○○	○○●●○○●●		○○○○○○○●●●●	Selection: <7 Distinction: ~10
<i>Color</i>	• ○	●○●●○●●●			Selection: <7 Distinction: ~10
<i>Orientation</i>	\\ /				Theoretically Infinite
<i>Texture</i>	○○○○	○○○○○○○○			Theoretically Infinite

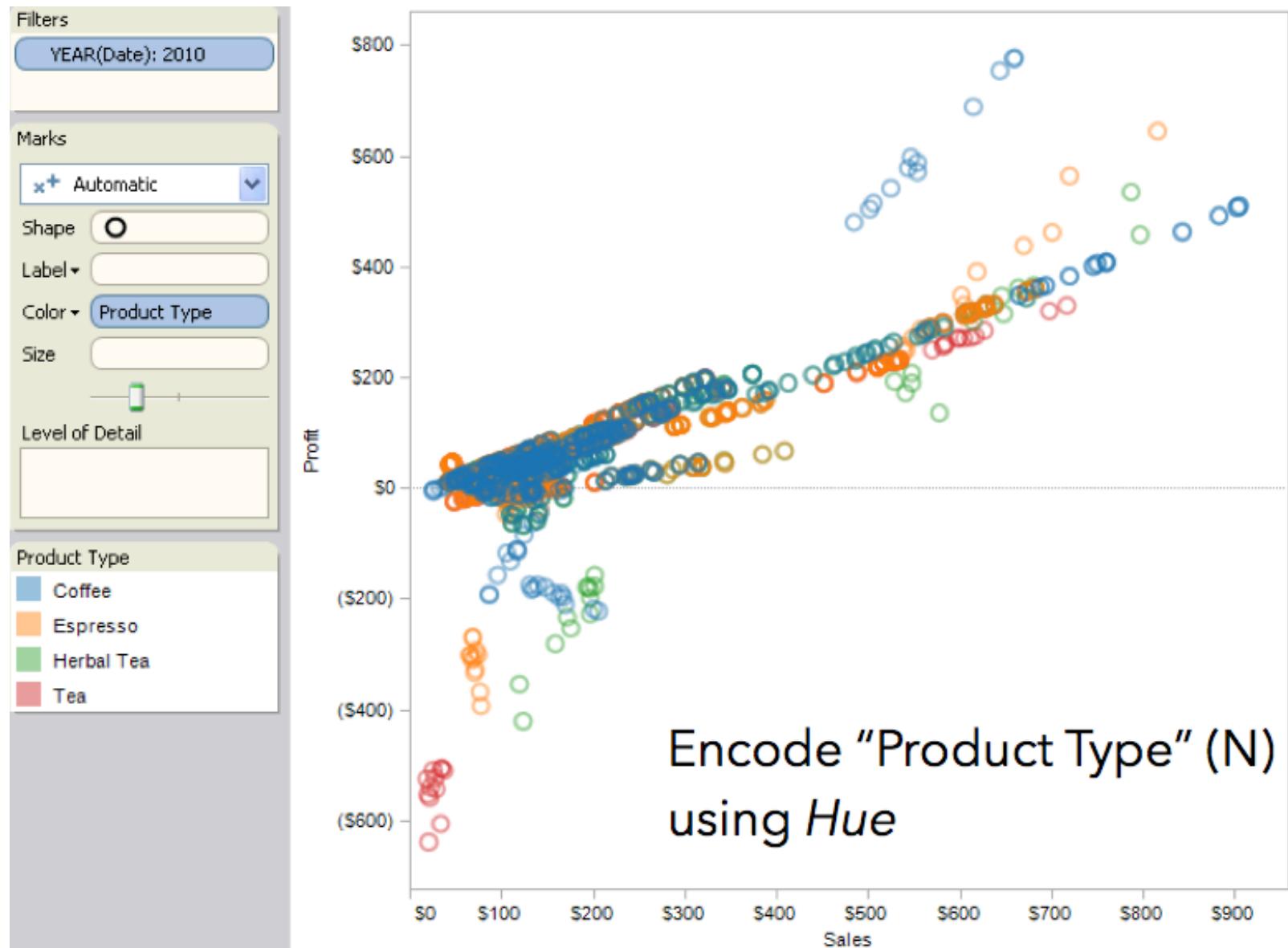
An Example: Coffee Sales

- Sales: Q-Ratio
- Profit: Q-Ratio
- Marketing: Q-Ratio
- Product Type: N {Coffee, Espresso, Herbal Tea, Tea}
- Market: N {Central, East, South, West}

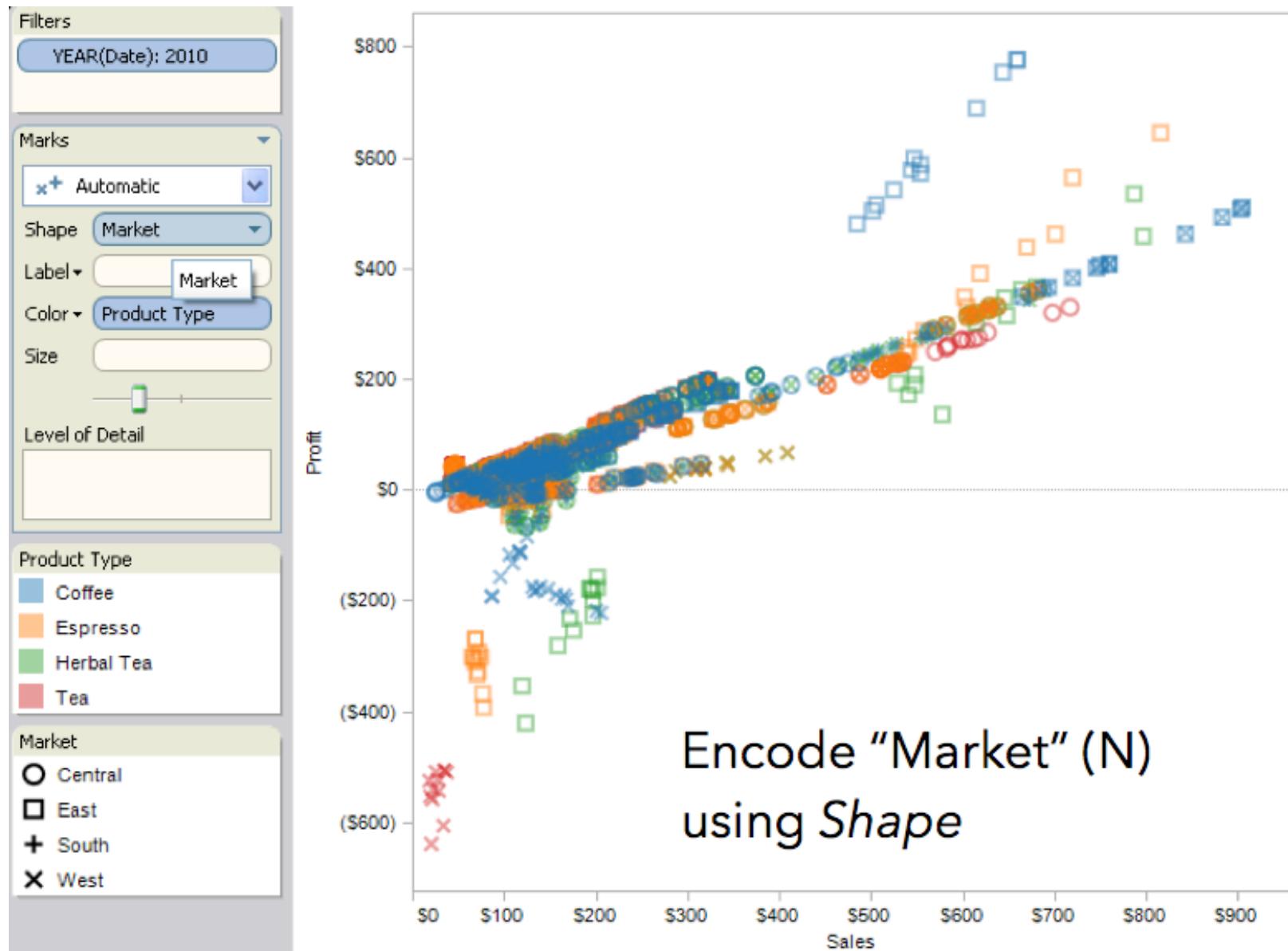
First Two Variables



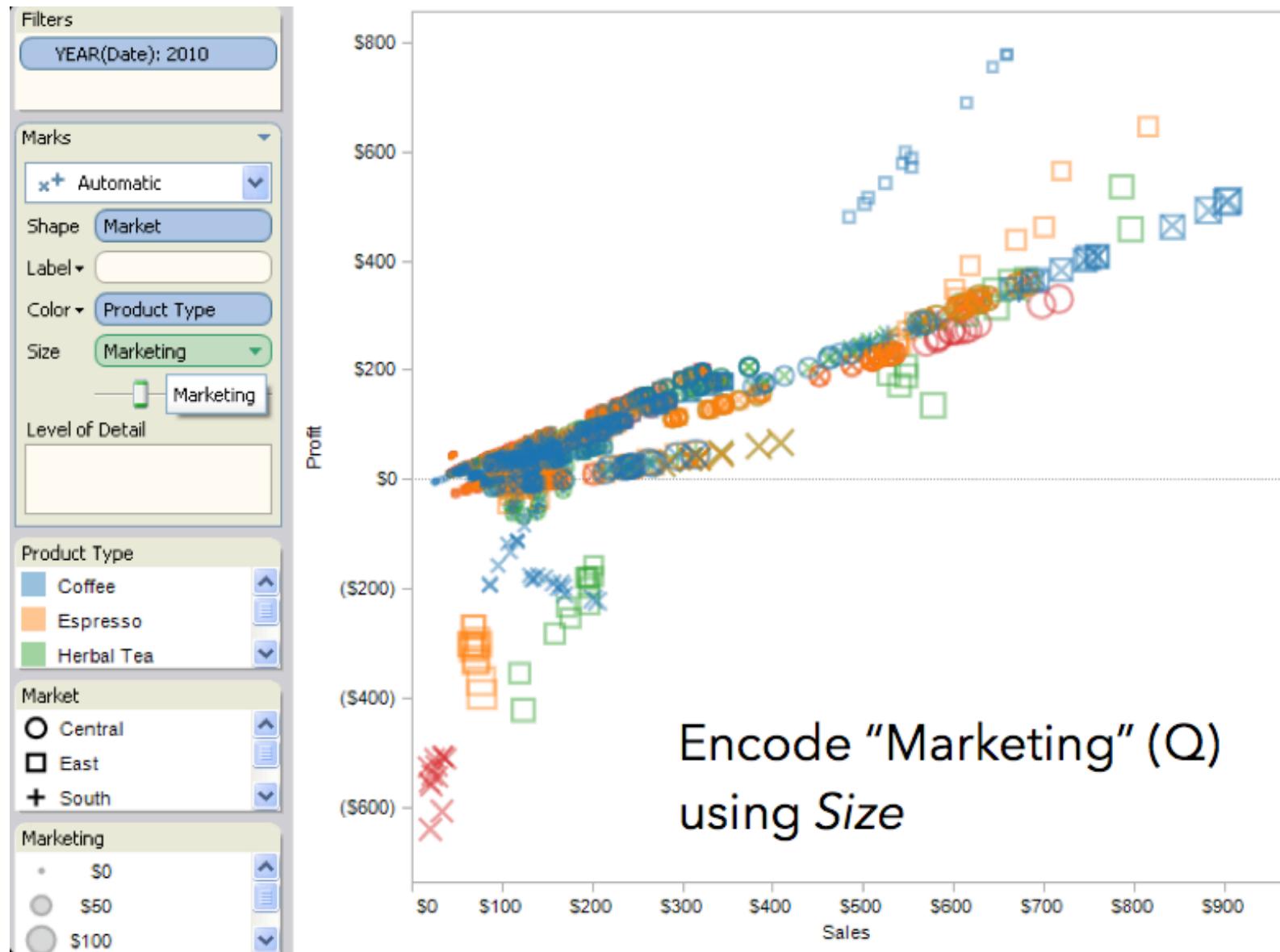
Third Variable



Fourth Variable



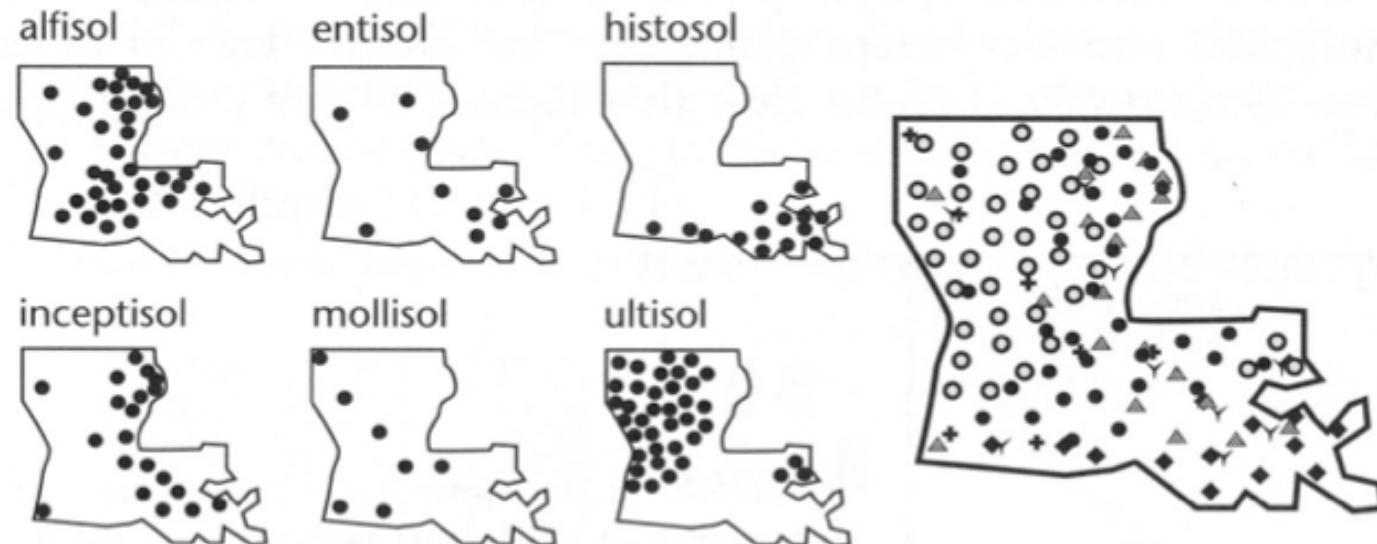
Fifth Variable



Small Multiples

“At the heart of quantitative reasoning is a single question: Compared to what? Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution.”

Tufte, Envisioning Information

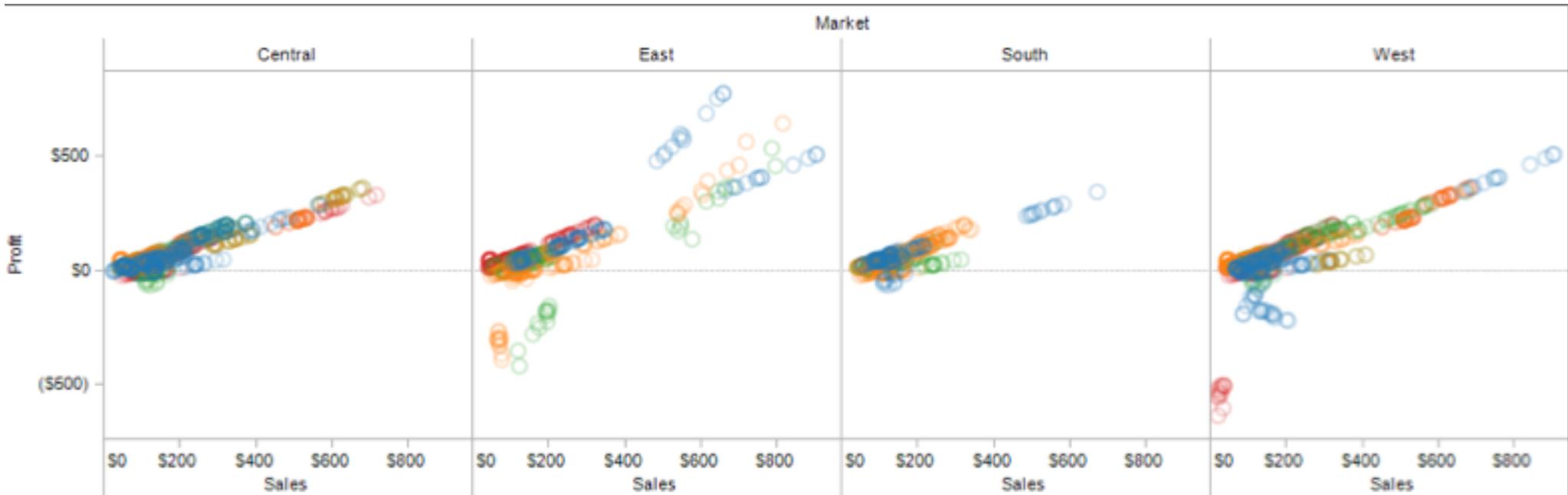


In The Visual Display of Quantitative Information (Textbook, Chapter 8)

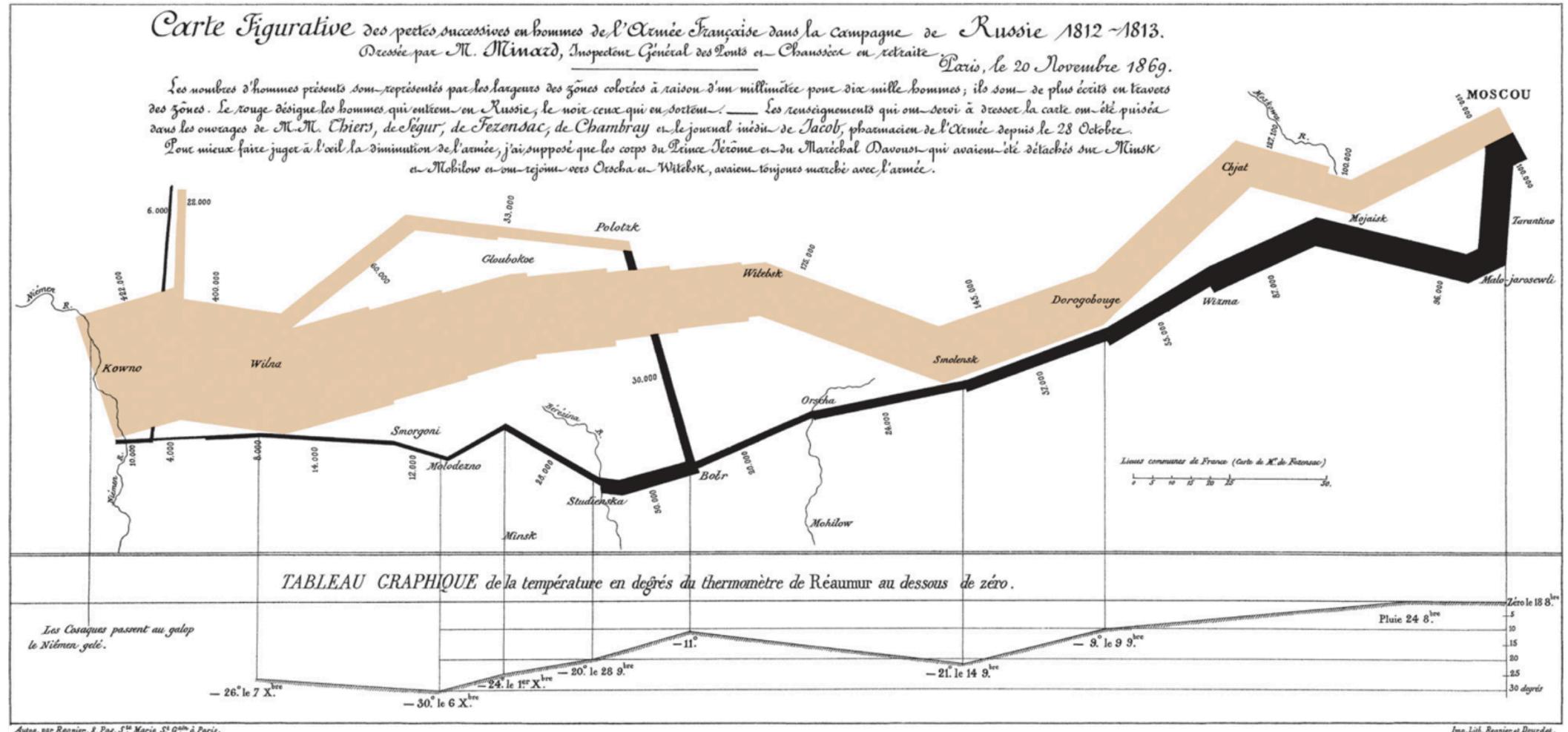
Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Trellis Display (Small Multiples)

- It subdivides space to enable comparison across multiple plots.
- Typically nominal or ordinal variables are used as dimensions for subdivision.

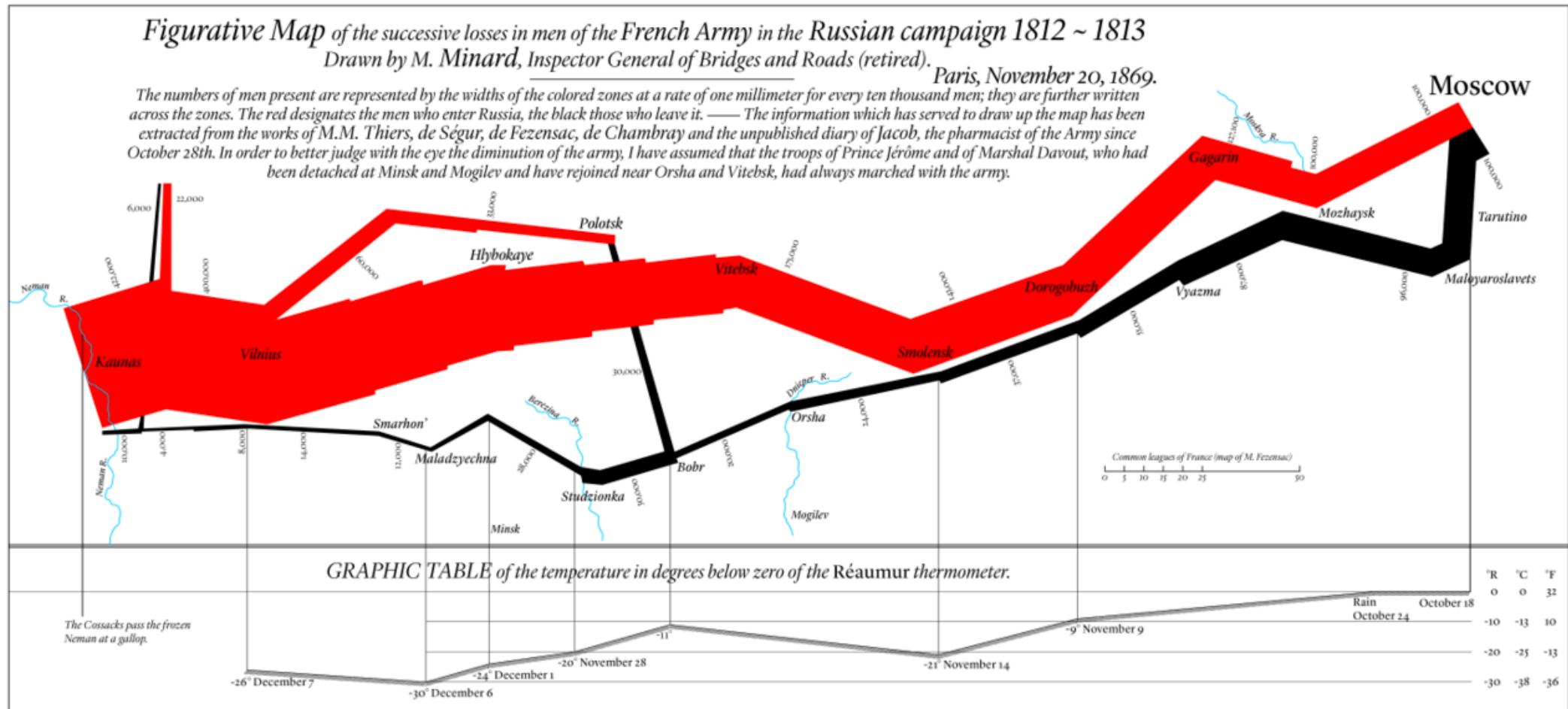


Minard 1869: Napoleon's March



In The Visual Display of Quantitative Information (Textbook, Chapter 1)

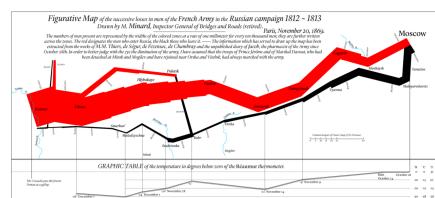
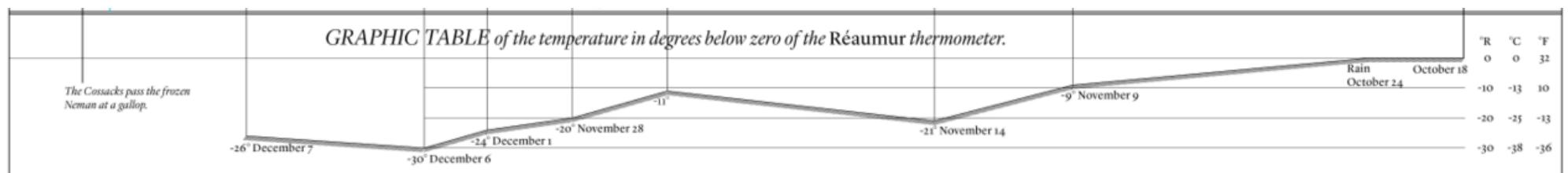
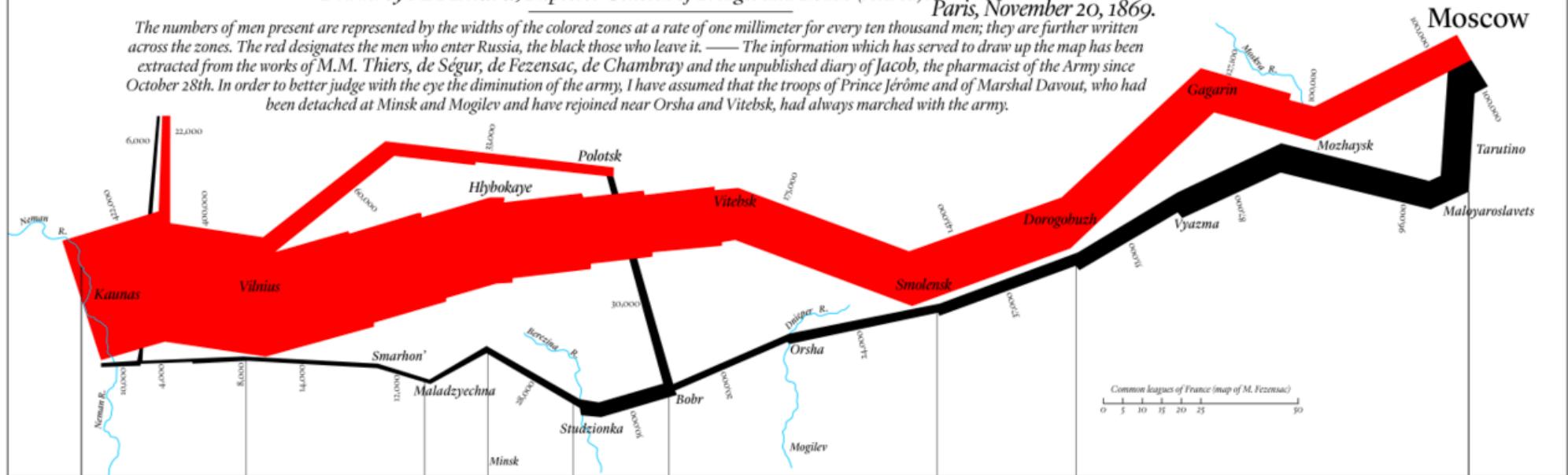
Minard 1869: Napoleon's March



Decomposition

*Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813
 Drawn by M. Minard, Inspector General of Bridges and Roads (retired).*

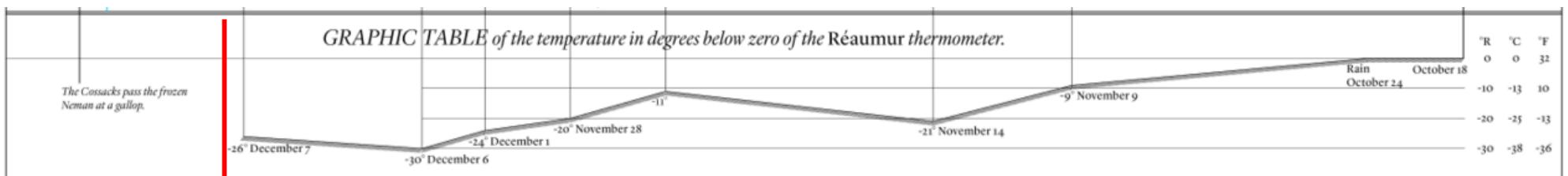
Paris, November 20, 1869.
 The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.



Decomposition

Y-axis: temperature (Q)

X-axis: longitude (Q) / time (O)



Temperature over space/time (Q x Q)

Decomposition

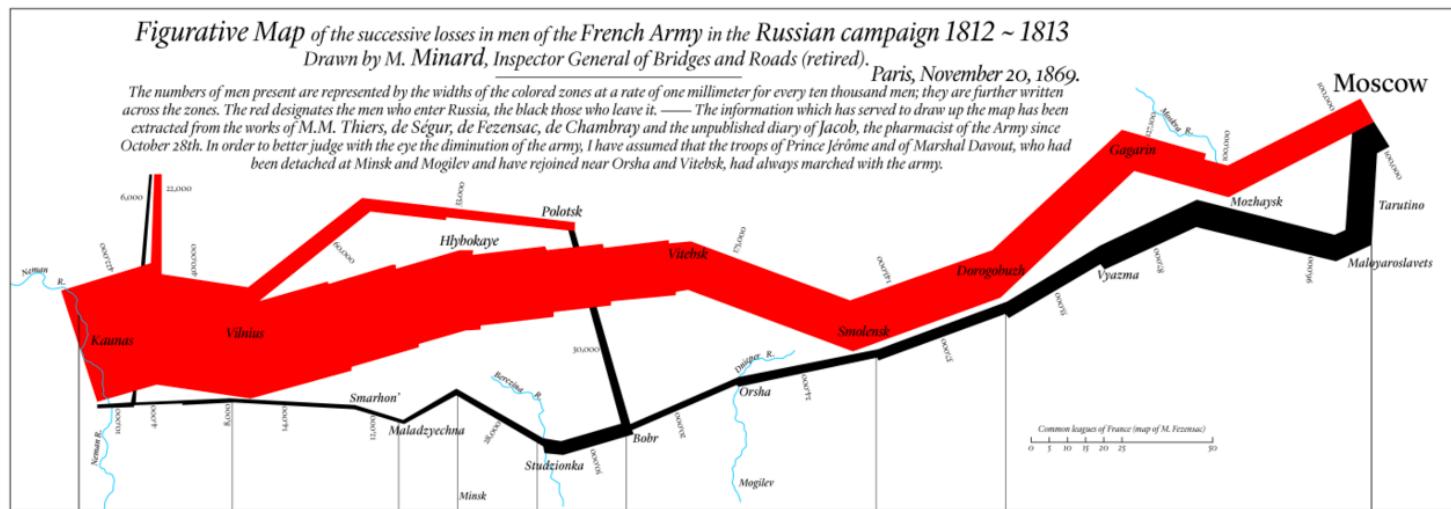
Y-axis: longitude (Q)



X-axis: latitude (Q)

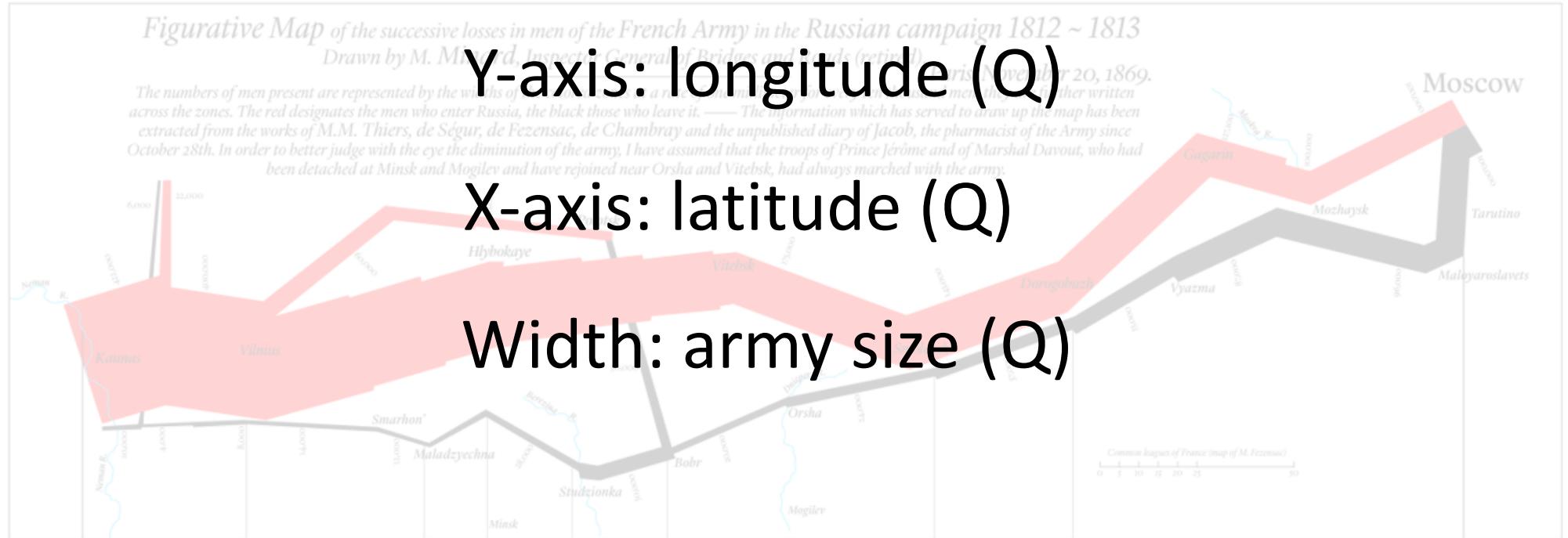


Width: army size (Q)



Army position ($Q \times Q$) and army size (Q)

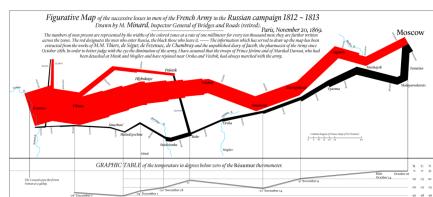
Minard 1869: Napoleon's March



+

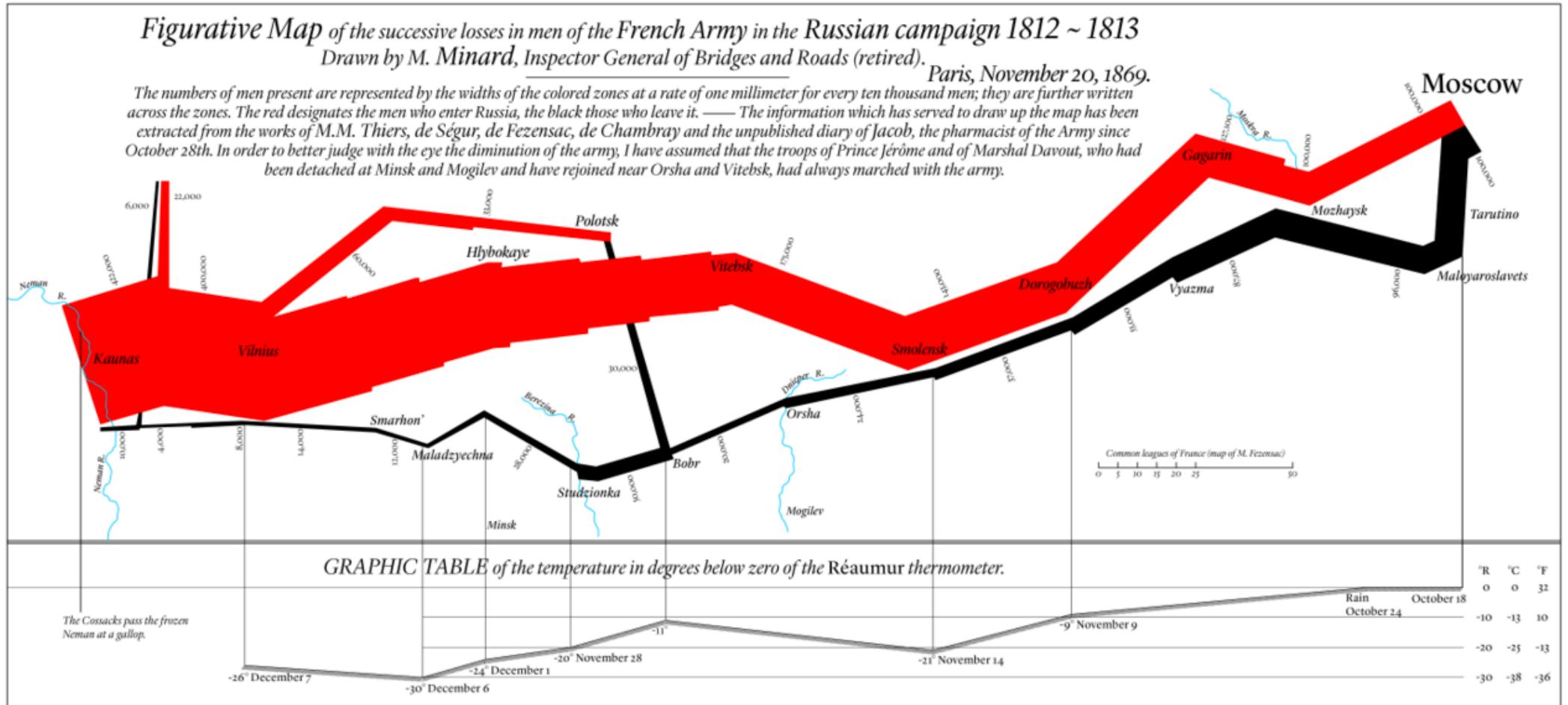


=



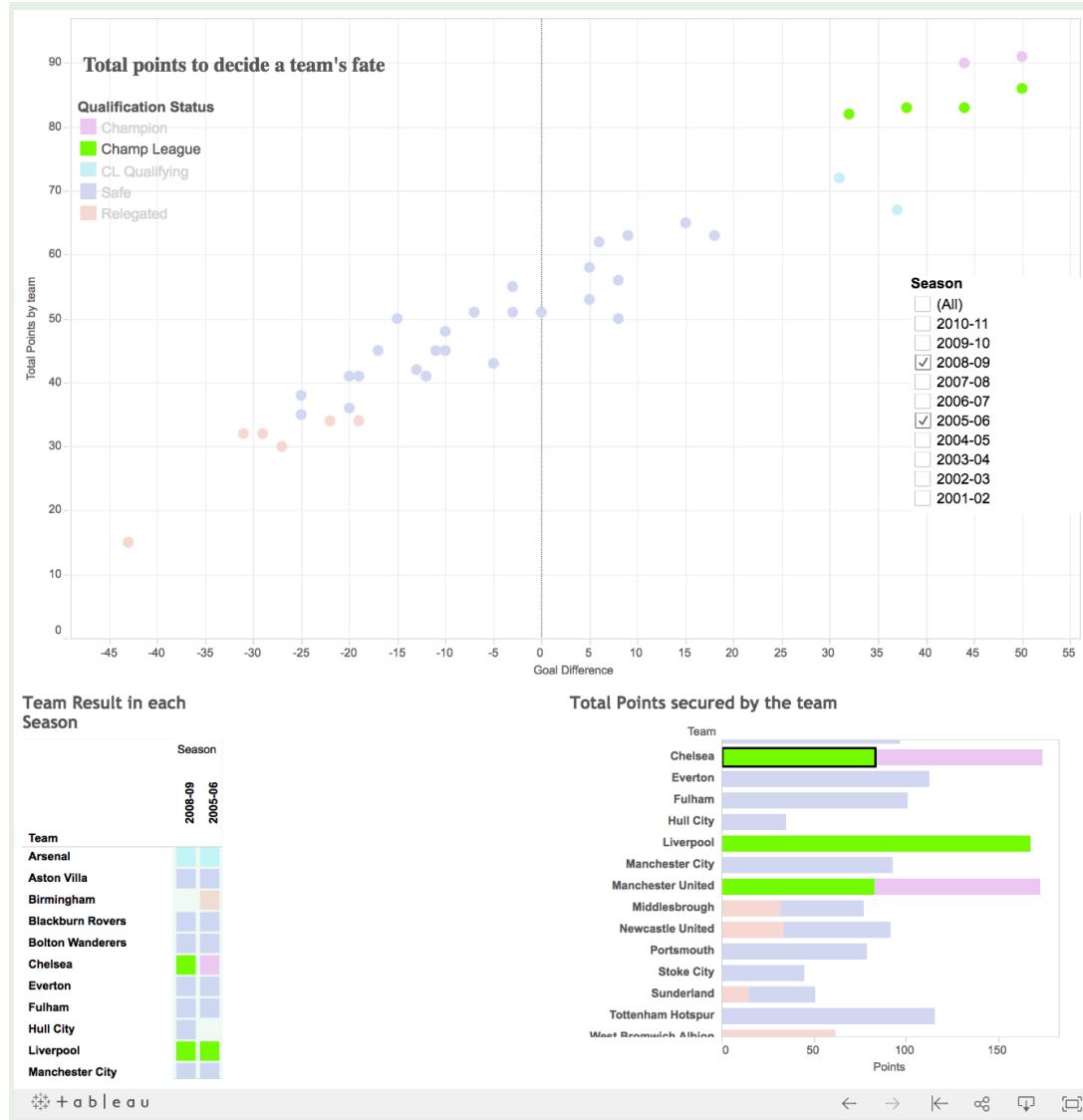
Depicts at least 5 quantitative variables.

Minard 1869: Napoleon's March



Depicts at least 5 quantitative variables. Any others?

Multiple Coordinated Views



Multivariate Data Visualization

- Strategies:
 - Avoid “over-encoding”
 - Use space and small multiples intelligently
 - Reduce the problem space
 - Use interaction to generate relevant views
- Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is key.

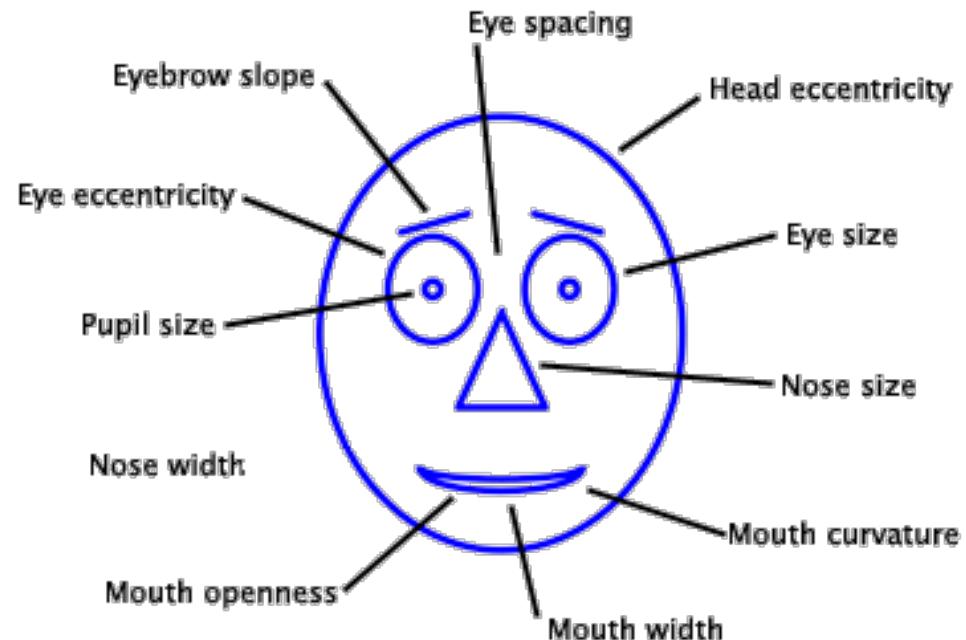
Common Multivariate Data Visualization Techniques

- Chernoff Faces
- Table Lens
- Parallel Coordinates
- Mosaic Plot

Chernoff Faces

Chernoff Faces

- Observation: We have evolved a sophisticated ability to interpret faces.
- Idea: Encode different variables' values in characteristics of human face



In The Visual Display of Quantitative Information (Textbook, Chapter 7)

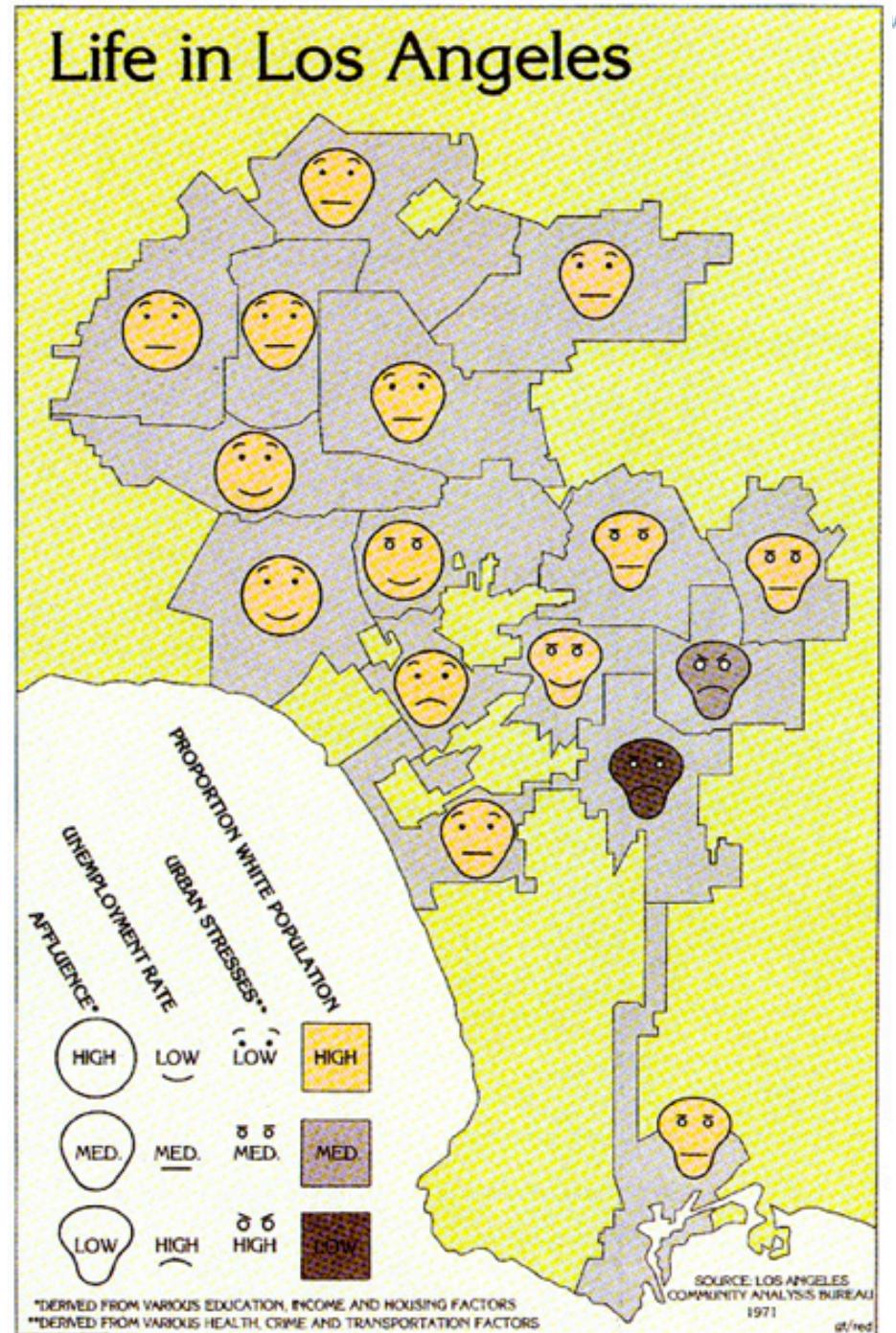
(Optional Reading) Chernoff, Herman. "The use of faces to represent points in k-dimensional space graphically." Journal of the American Statistical Association 68.342 (1973): 361-368.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Example

“It is probably one of the most interesting maps I’ve created because the expressions evoke an emotional association with the data.”

Eugene Turner



Critiques

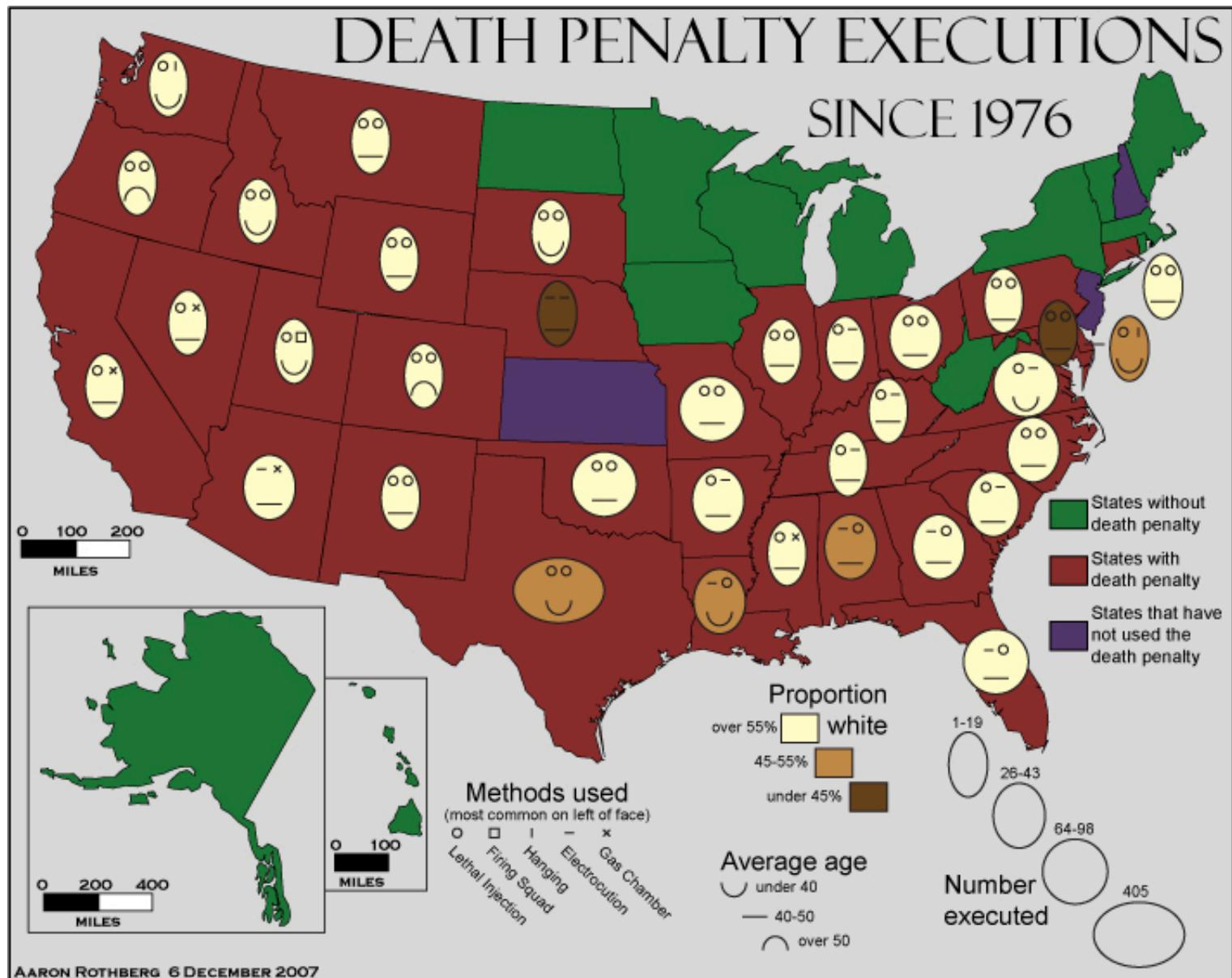


Table Lens

Table Lens

- Spreadsheet is certainly one hypervariate data presentation
- Idea: Make the text more visual and symbolic
- Just leverage basic bar chart idea

Rao, Ramana, and Stuart K. Card. "The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 1994.

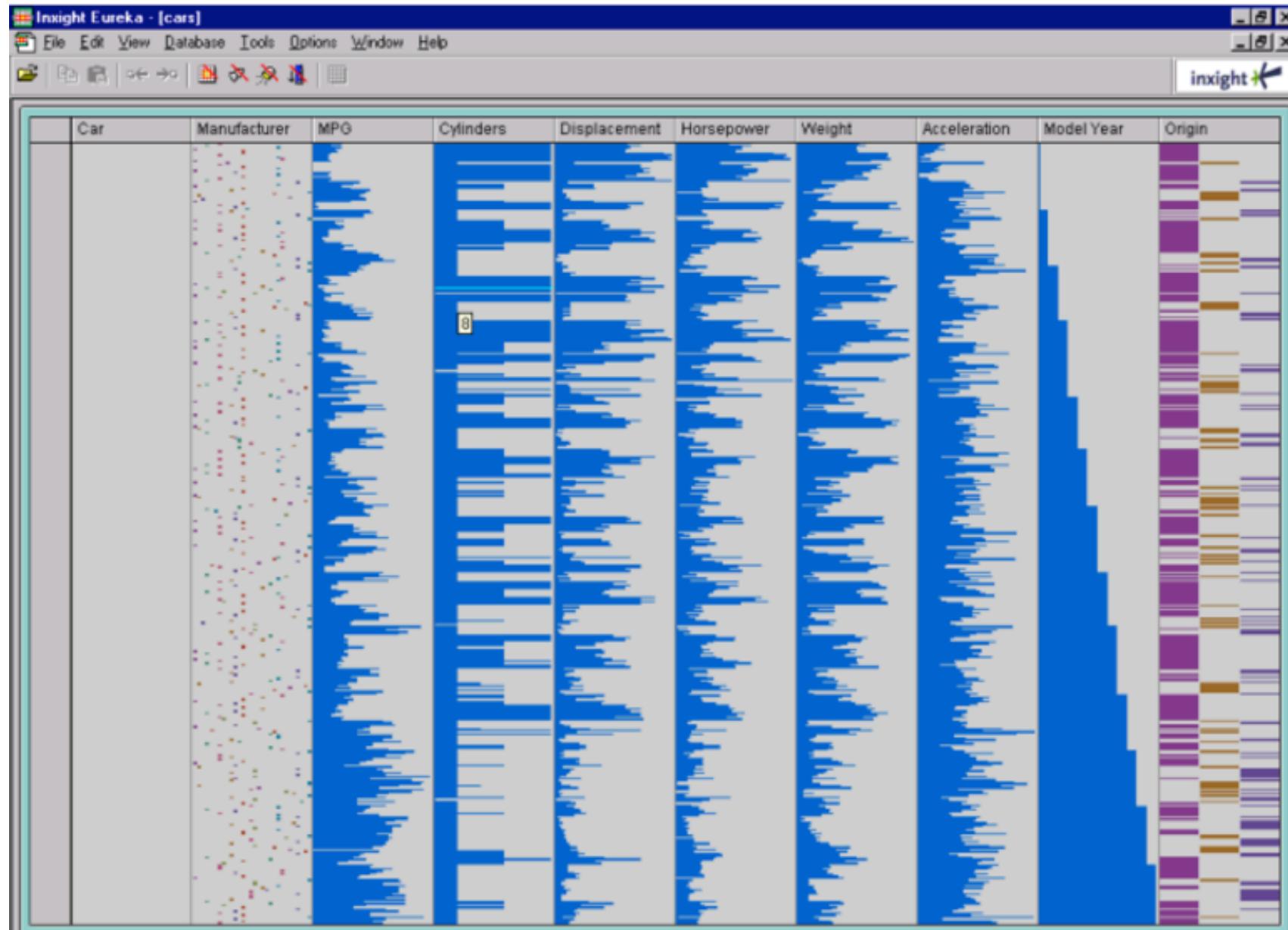
Visual Mapping

- Basic idea:
Change
quantitative
values to bars
- What do you
do for nominal
data?

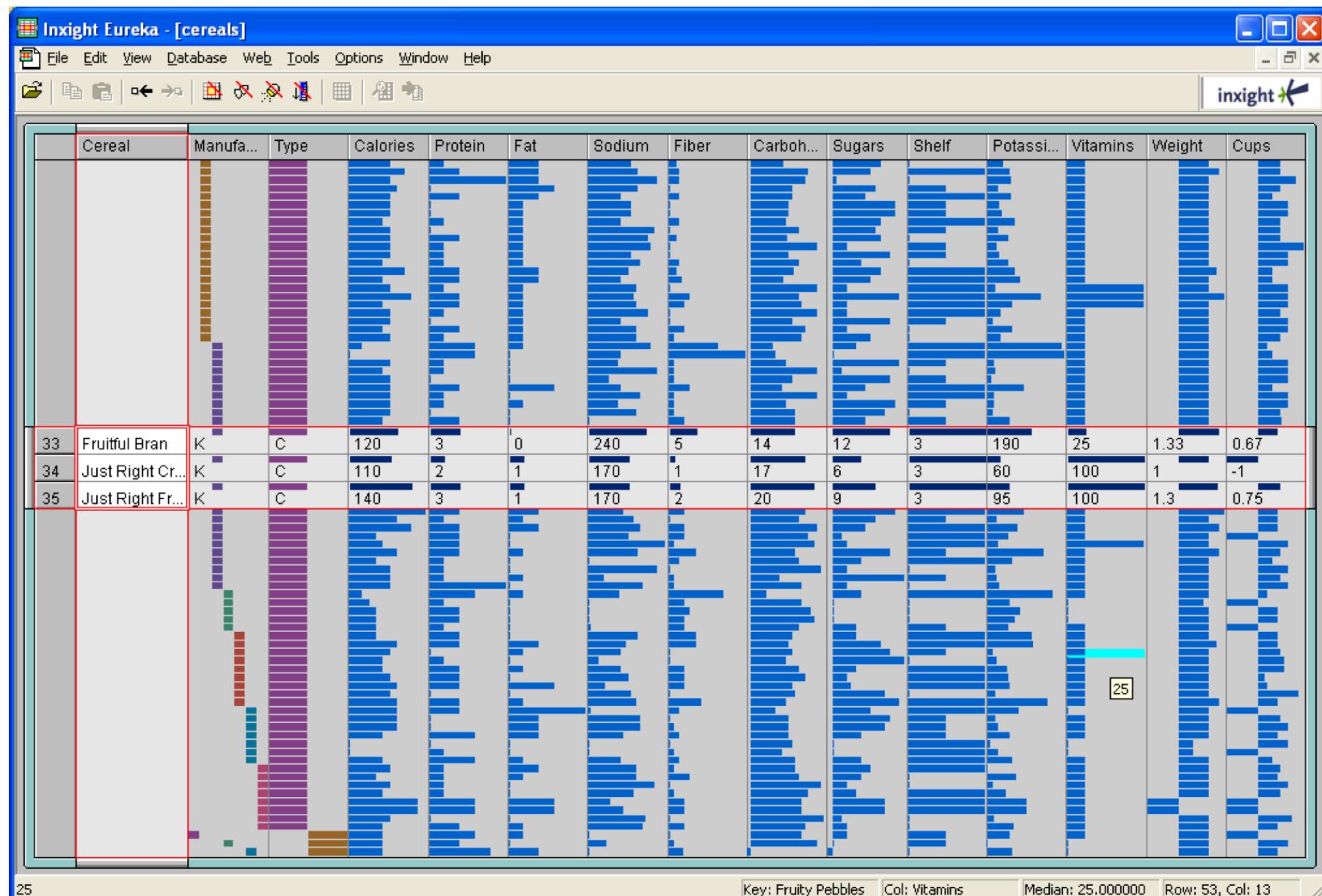
	A	B	C	D	E	F	G	H	I
1	Cereal	Manufactur	Type	Calories	Protein	Fat	Sodium	Fiber	Carbo
2	Frosted Mini-Wheats	K	C	100	3	0	0	0	3
3	Raisin Squares	K	C	90	2	0	0	0	2
4	Shredded Wheat	N	C	80	2	0	0	0	3
5	Shredded Wheat 'n'Bran	N	C	90	3	0	0	0	4
6	Shredded Wheat spoon s	N	C	90	3	0	0	0	3
7	Puffed Rice	Q	C	50	1	0	0	0	0
8	Puffed Wheat	Q	C	50	2	0	0	0	1
9	Maypo	A	H	100	4	1	0	0	0
10	Quaker Oatmeal	Q	H	100	5	2	0	2.7	
11	Strawberry Fruit Wheats	N	C	90	2	0	15	3	
12	100% Natural Bran	Q	C	120	3	5	15	2	
13	Golden Crisp	P	C	100	2	0	45	0	
14	Smacks	K	C	110	2	1	70	1	
15	Great Grains Pecan	P	C	120	3	3	75	3	
16	Cream of Wheat (Quick)	N	H	100	3	0	80	1	
17	Corn Pops	K	C	110	1	0	90	1	
18	Muesli Raisins, Dates, & R	C	C	150	4	3	95	3	
19	Apple Jacks	K	C	110	2	0	125	4	



Example of Table Lens



Focus and Context



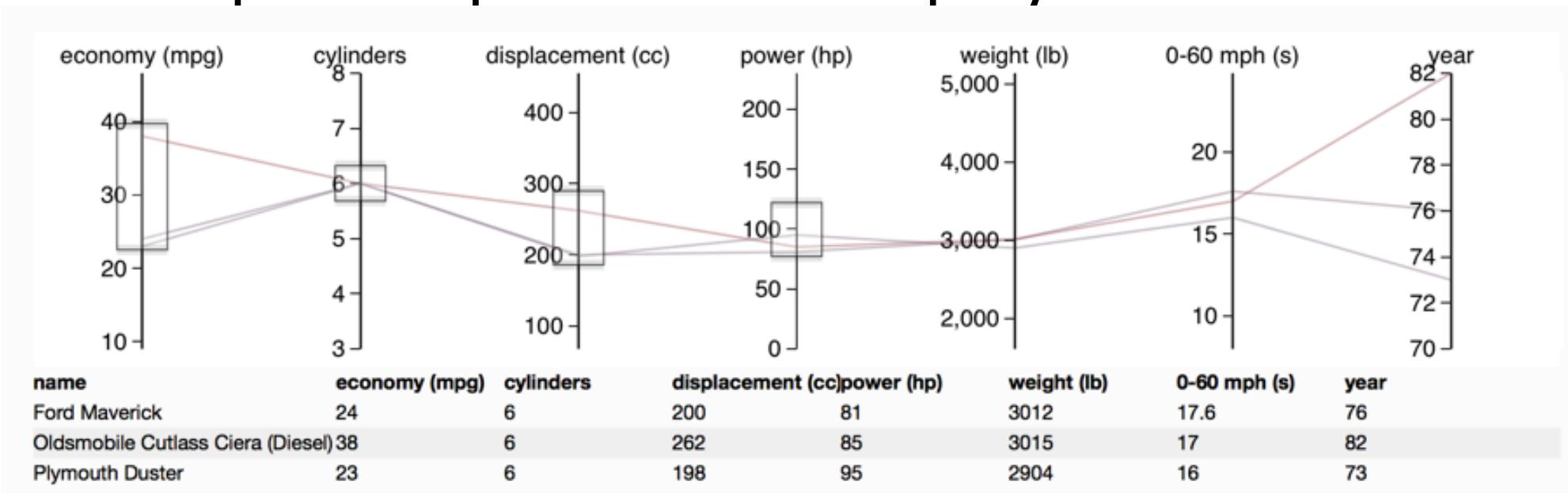
Video Demo

- [http://www.open-video.org/details.php?
videoid=8304](http://www.open-video.org/details.php?videoid=8304)
- Space advantage
- Fluid navigation
- Direct exploration

Parallel Coordinates

Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline



Live Demo

<https://syntagmatic.github.io/parallel-coordinates>

To learn more:

Heinrich, Julian, and Daniel Weiskopf. "State of the Art of Parallel Coordinates." Eurographics (STARs). 2013.

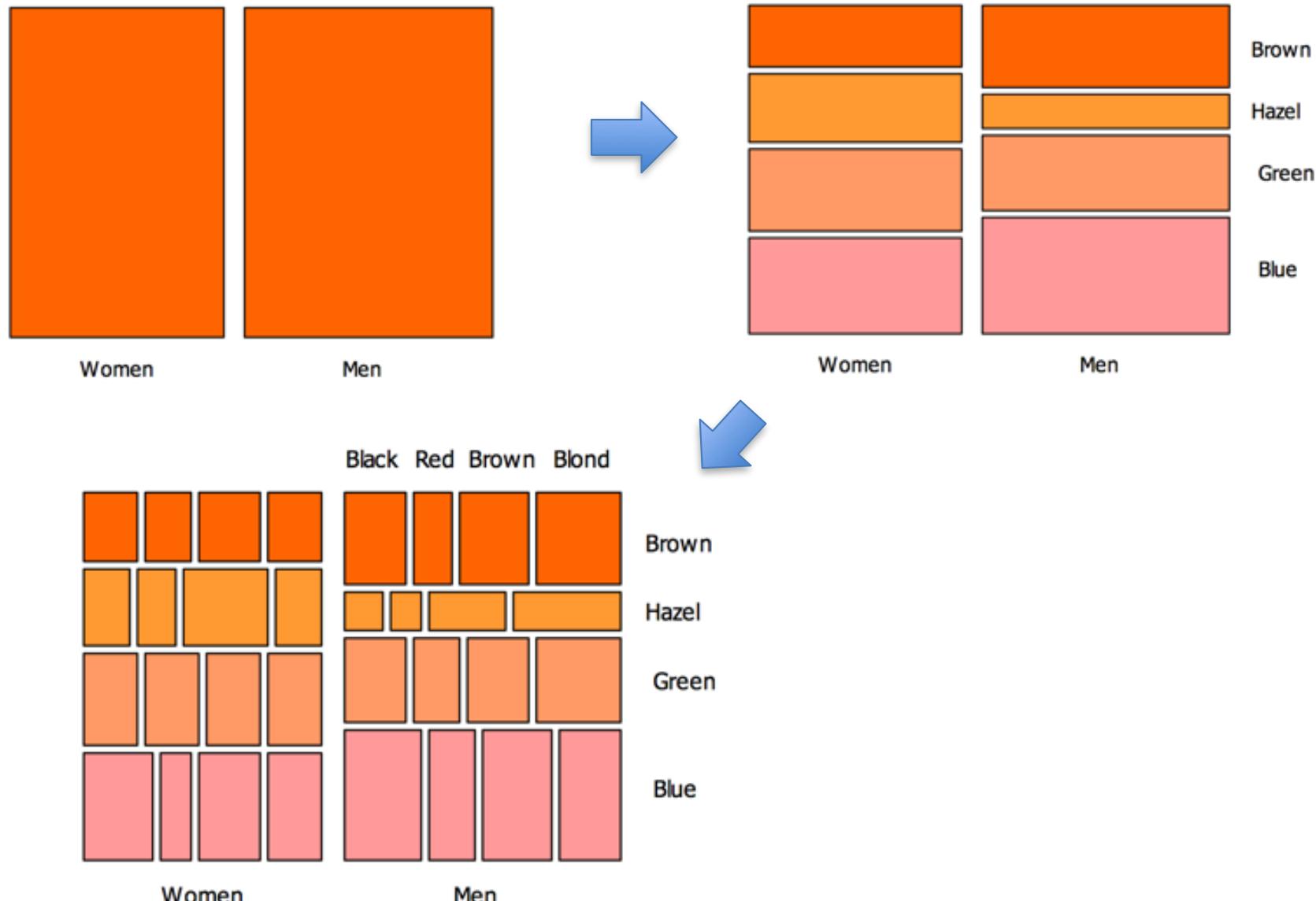
<http://www.parallelcoordinates.de>

Mosaic Plot

Multivariate Categorical Data

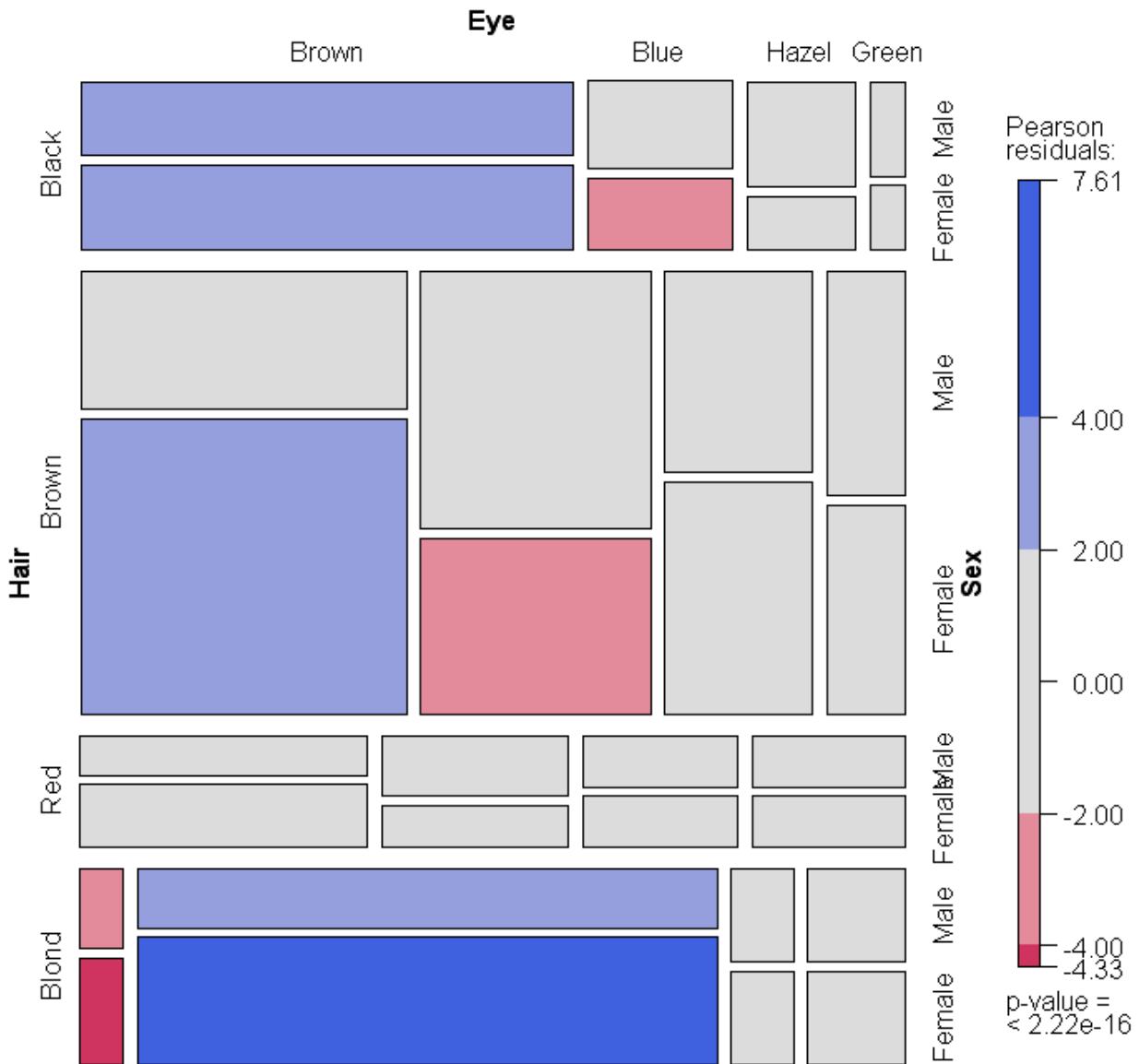
- How about multivariate categorical data?
- Students
 - Gender: Female, male
 - Eye color: Brown, blue, green, hazel
 - Hair color: Black, red, brown, blonde, gray
 - Home country: USA, China, Italy, India, ...

Mosaic Plot Decomposition



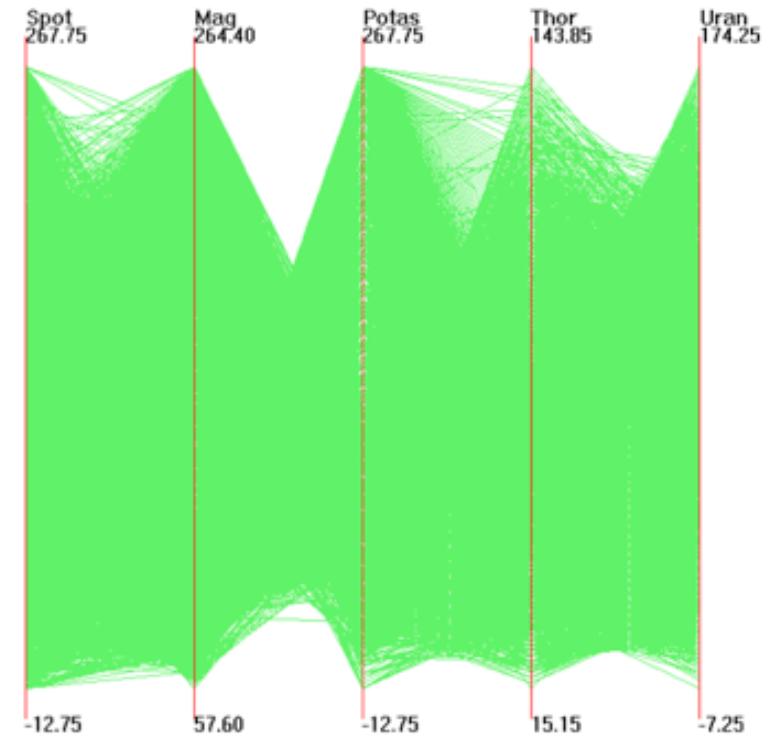
Mosaic Plot

- Hair
- Sex
- Eye
- Level of the Pearson residual



Data Overload

- Most of the techniques we've examined work for a modest number of data cases or variables
- What happens when you have lots and lots of data cases and/or variables?



Out5d dataset(5 dimensions, 16384 items)

We will address this in other lectures.

Summary

- Table vs. Graphs
- Visual encodings (Bertin's semiology)
 - Limitation of possible number of variables
- Reduce Problem Space
 - Small Multiples
 - Multiple Views
- Common Visualizations

G53FIV: Fundamentals of Information Visualization

Lecture 6: Visualization with R - Fundamentals

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Overview

- R Basics
- Visualization using R

R Basics

What is ?

- GNU project developed by John Chambers @ Bell Lab (<https://www.r-project.org/>)
- Free software environment for **statistical computing** and **graphics**
- Functional programming language written primarily in C, Fortran
- A lot of data scientists working in the company (such as Google) use R.
- IDE: R Studio (www.rstudio.com)

R is a tool for...

Data Manipulation

- connecting to data sources
- slicing & dicing data

Modeling & Computation

- statistical modeling
- numerical simulation

Data Visualization

- visualizing fit of models
- composing statistical graphics

munge

model

visualize

CRAN



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Contributed Packages

Available Packages

Currently, the CRAN package repository features 10093 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual [R Installation and Administration](#) (also contained in the R base sources) explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 34 views are available.

- install a package from the command line:
 - `install.package("ggplot2", dependencies = TRUE)`

<http://cran.r-project.org>

Getting Help with R

- Embedded “help” function in R
 - `help(func)`, `?func`
- For a topic
 - `help.search(topic)`, `??topic`
- `demo(is.things)`

- `search.r-project.org`
- Stack Overflow:
 - <http://stackoverflow.com/tags/R>

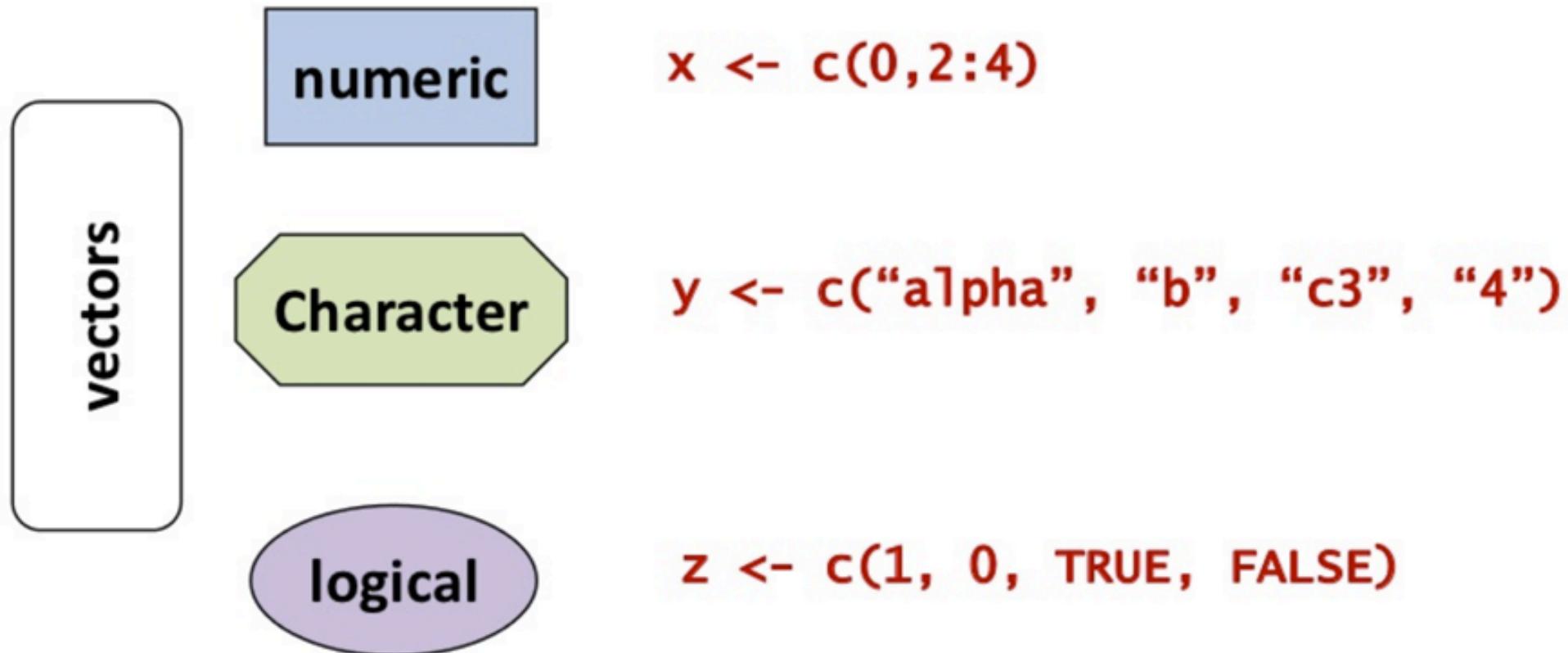
Bring Data into R

- Create csv file
- Name your variables well
 - Self-explanatory, unique, lowercase, short-ish, one-word name
- In R, set the working directory
 - `setwd("/users/you/R/tutorial")`
 - What is the working directory? `getwd()`
 - What is in the working directory? `dir()`
- Read in data
- Write data

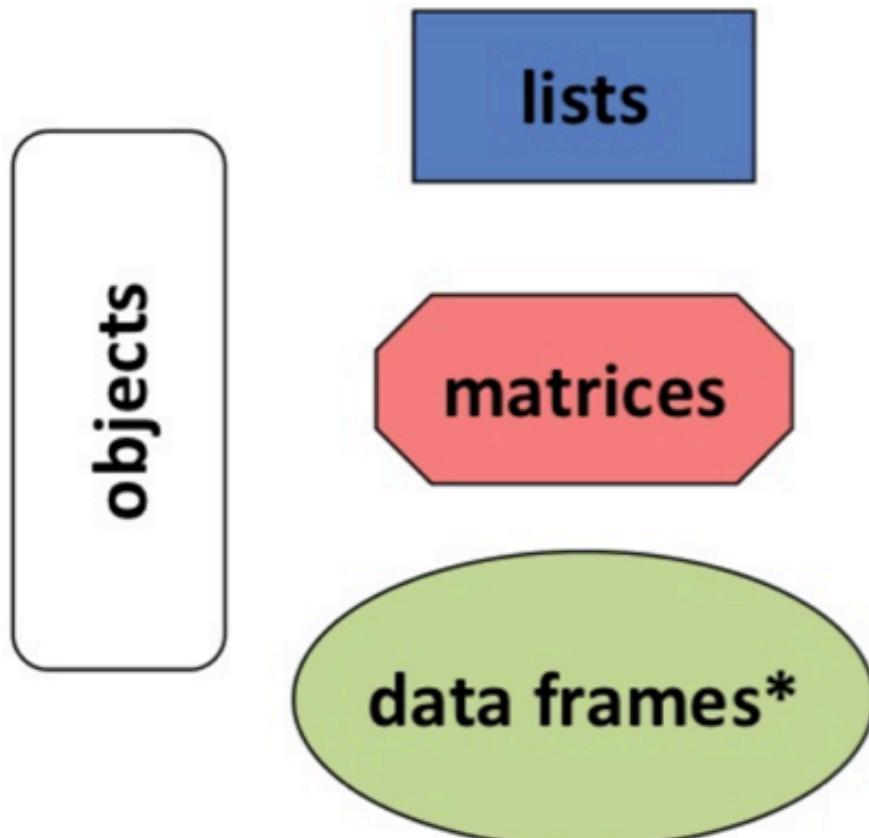
Read and Write

- Read in data
 - CSV files: `iris.df <- read.csv("iris.csv", header=T)`
 - Clipboard: `read.csv("clipboard")` – like cutting and pasting it
 - From web: `read.csv(http://url/1.csv)`
 - From excel files (using the XLConnect package):
 - `iris.df <- readWorksheetFromFile("iris.xlsx", sheet="Sheet1")`
 - From R object: `load("iris.Rdata")`
- Write data
 - To CSV: `write.csv(iris.df, "iris_dataframe.csv")`
 - To R objects: `save(iris, "iris.RData")`
 - To databases:
 - `con <- dbConnect(dbdriver, user, password, host, dbname)`
 - `dbWriteTable(con, "iris", iris.df)`

R Data Structures



R Data Structures



`lst <- list(x,y,z)`

`M <- matrix(rep(x,3),ncol=3)`

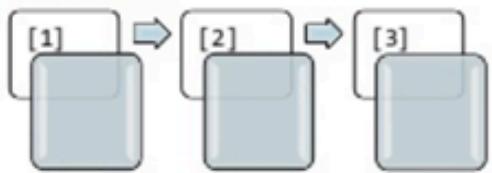
`df <- data.frame(x,y,z)`

R Data Structures

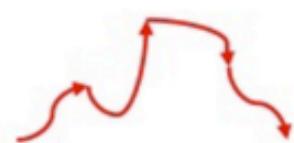
	Linear	Rectangular
Homogeneous	vectors	matrices
Heterogeneous	lists	data frames*

R Data Structures: more details

VECTOR



- 1 row, N columns.
- One data type only (numeric, character, date, OR logical).
- Uses: track changes in a single variable over time.
- Examples: stock prices, hurricane path, temp readings, disease spread, financial performance, sports scores.



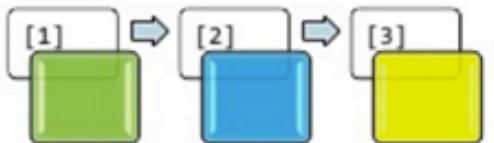
MATRIX



3	1	5	9	6	9
0	7	0	7	6	8
0	7	2	8	9	0
3	8	5	0	3	4
6	0	8	4	9	0
6	5	5	2	5	8
7	8	9	7	9	8

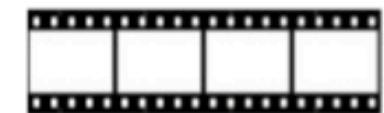
- N row, N columns.
- One data type only (any combination of numeric, character, date, logical).
- Basically, a collection of vectors.

LIST



- 1 row, N columns. Multiple data types.
- Uses: list detailed information for a person/place/thing/concept.
- Examples: Listing for real estate, book, movie, contact, country, stock, company, etc. Or, a "snapshot" or observation of an event or phenomenon such as stock market, or scientific experiment.

DATA FRAME



- N rows, N columns.
- Multiple data types.
- Basically, a collection of lists or snapshots which when assembled together provide a "bigger picture."

Other Important R Concepts

FACTORS

Stores each distinct value only once, and the data itself is stored as a vector of integers. When a factor is first created, all of its levels are stored along with the factor.

```
> weekdays=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
> wf <- factor(weekdays)
[1] Monday   Tuesday  Wednesday Thursday Friday
Levels: Friday Monday Thursday Tuesday Wednesday
Used to group and summarize data:
WeekDaySales <- (DailySalesVector, wf, sum)
# Sum daily sales figures by M,T,W,Th,F
```

PACKAGES, FUNCTIONS, DATASETS

```
> search() # Search for installed packages & datasets
[1] ".GlobalEnv"        "mtcars"           "tools:rstudio"
[4] "package:stats"     "package:graphics" "package:grDevices"

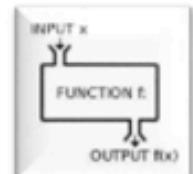
> library(ggplot2) # load package ggplot2
Attaching package: 'ggplot2'

> data() # List available datasets

> attach(iris) # Attach dataset "iris"
```

USER-DEFINED FUNCTIONS

```
> f <- function(a) { a^2 }
> f(2)
[1] 4
```



- Functions can be passed as arguments to other functions.
- Function behavior is defined inside the curly brackets { }.
- Functions can be nested, so that you can define a function inside another.
- The return value of a function is the last expression evaluated.

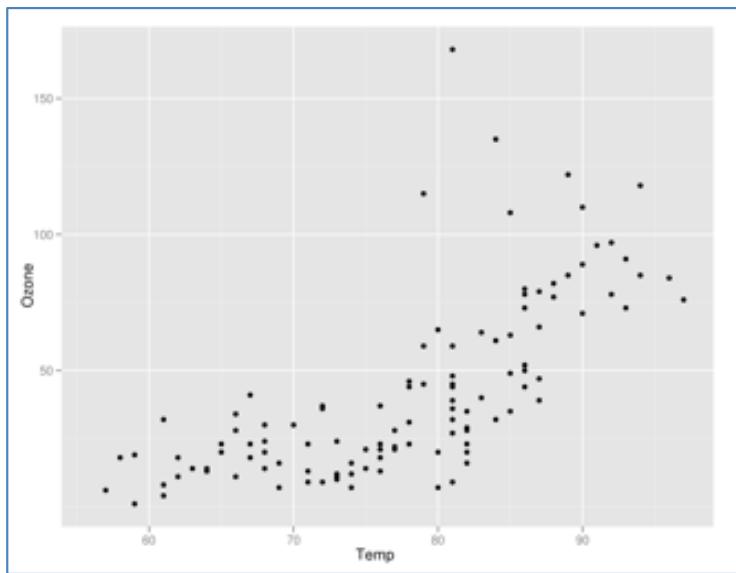
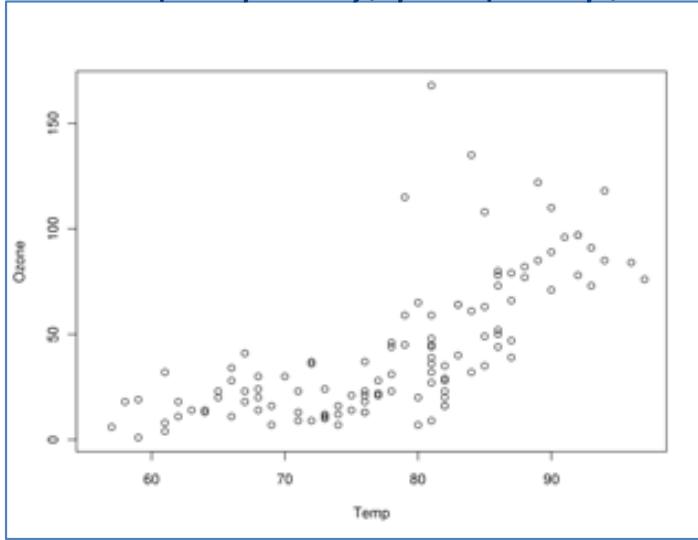
SPECIAL VALUES

- **pi=3.141593**. Use lowercase "pi"; "Pi" or "PI" won't work
 - **inf=1/0 (Infinity)**
 - **NA=Not Available**. A logical constant of length 1 that means neither TRUE nor FALSE. Causes functions to barf.
 - Tell function to ignore NAs: `function(args, na.rm=TRUE)`
 - Check for NA values: `is.na(x)`
 - **NULL=Empty Value**. Not allowed in vectors or matrixes.
 - Check for NULL values: `is.null(x)`
 - **NaN=Not a Number**. Numeric data type value for undefined (e.g., 0/0).
- See [this](#) for NA vs. NULL explanation.

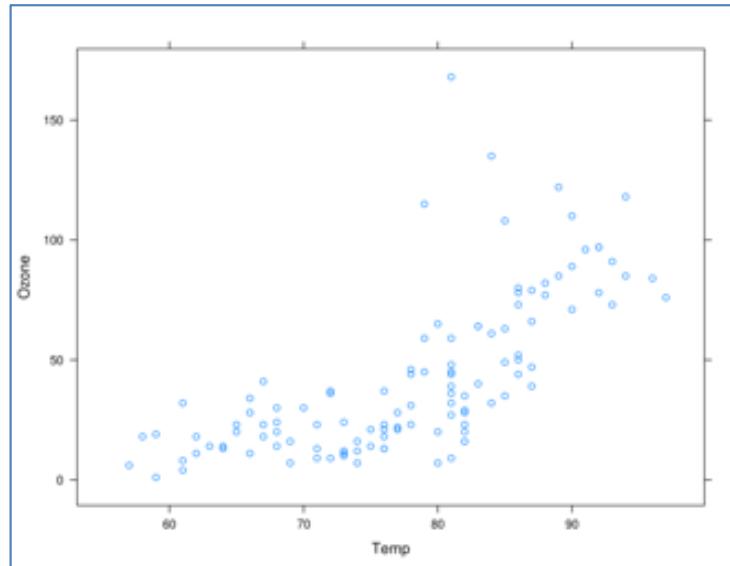
R Fundamental Visualization

R Graphics – 3 Main “Dialects”

base: `with(airquality, plot(Temp, Ozone))`



lattice: `xypplot(Ozone ~ Temp, airquality)`



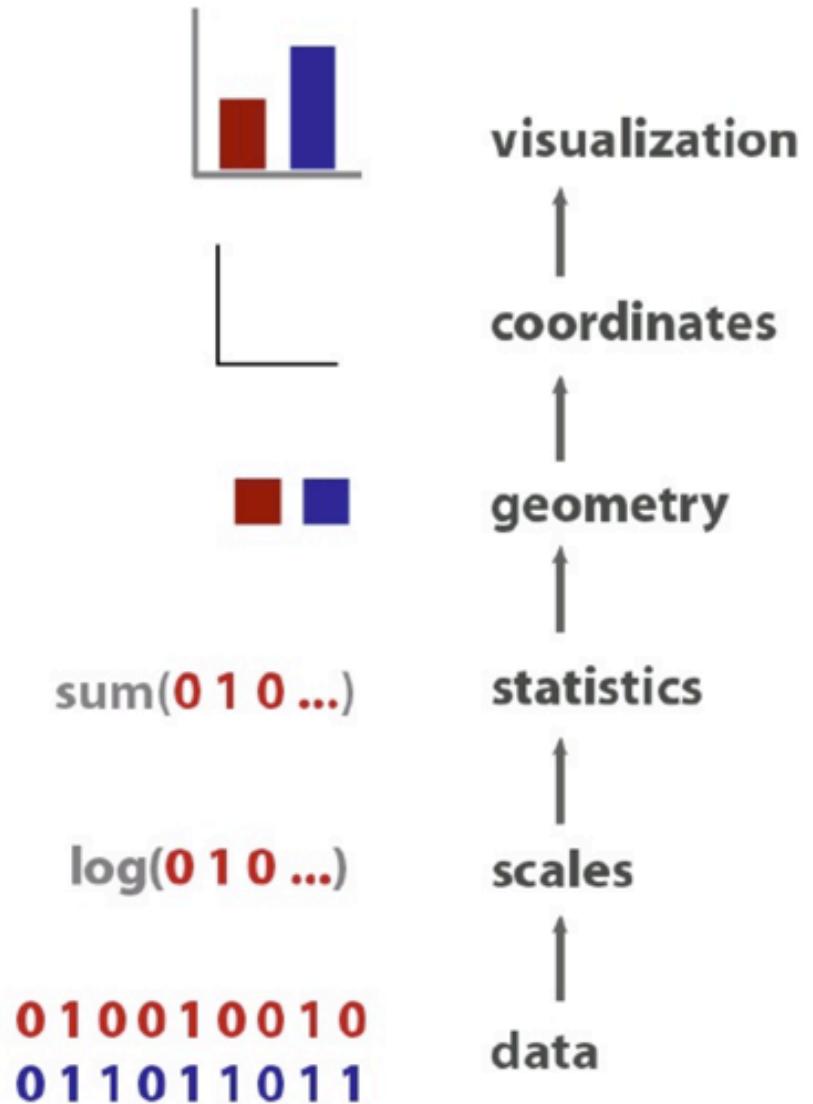
ggplot2: `ggplot(airquality, aes(Temp, Ozone)) + geom_point()`

Our focus: ggplot2

- More elegant and compact code than with base graphics
- More aesthetically pleasing defaults than lattice
- Very powerful for exploratory data analysis

ggplot2

- ‘gg’ is for ‘grammar of graphics’ (term by Lee Wilkinson)
- A set of terms that defines the basic components of a plot
- Used to produce figures using coherent, consistent syntax
- Easy to get started, plenty of power for complex figures



Building a Plot in ggplot2

data to visualize (a data frame)

map variables to ***aesthetic*** attributes

geometric objects – what you see (points, bars, etc)

scales map values from data to aesthetic space

faceting subsets the data to show multiple plots

statistical transformations – summarize data

coordinate systems put data on plane of graphic

Data

- Must be a data frame, pulled into the ggplot() object
- Example: the iris dataset
 - A multivariate dataset introduced by Fisher (1936)



Aesthetics (aes)

- How your data are represented visually
 - i.e. mapping
 - Which data on the x
 - Which data on the y
 - But also: color, size, shape, transparency

```
myplot <- ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width))
summary(myplot)

## # data: Sepal.Length, Sepal.Width, Petal.Length,
##   Petal.Width, Species [150x5]
## # mapping: x = Sepal.Length, y = Sepal.Width
## # faceting: facet_null()
```

Geometry (geom)

- The geometric objects in the plot
- Points, lines, polygons, etc.
- Shortcut functions
 - `geom_point()`
 - `geom_bar()`
 - `geom_line()`

Building a Plot in ggplot2

data to visualize (a data frame)

map variables to ***aesthetic*** attributes

geometric objects – what you see (points, bars, etc)

scales map values from data to aesthetic space

```
ggplot(iris) + geom_point(aes(x = Sepal.Length, y = Sepal.Width))
```

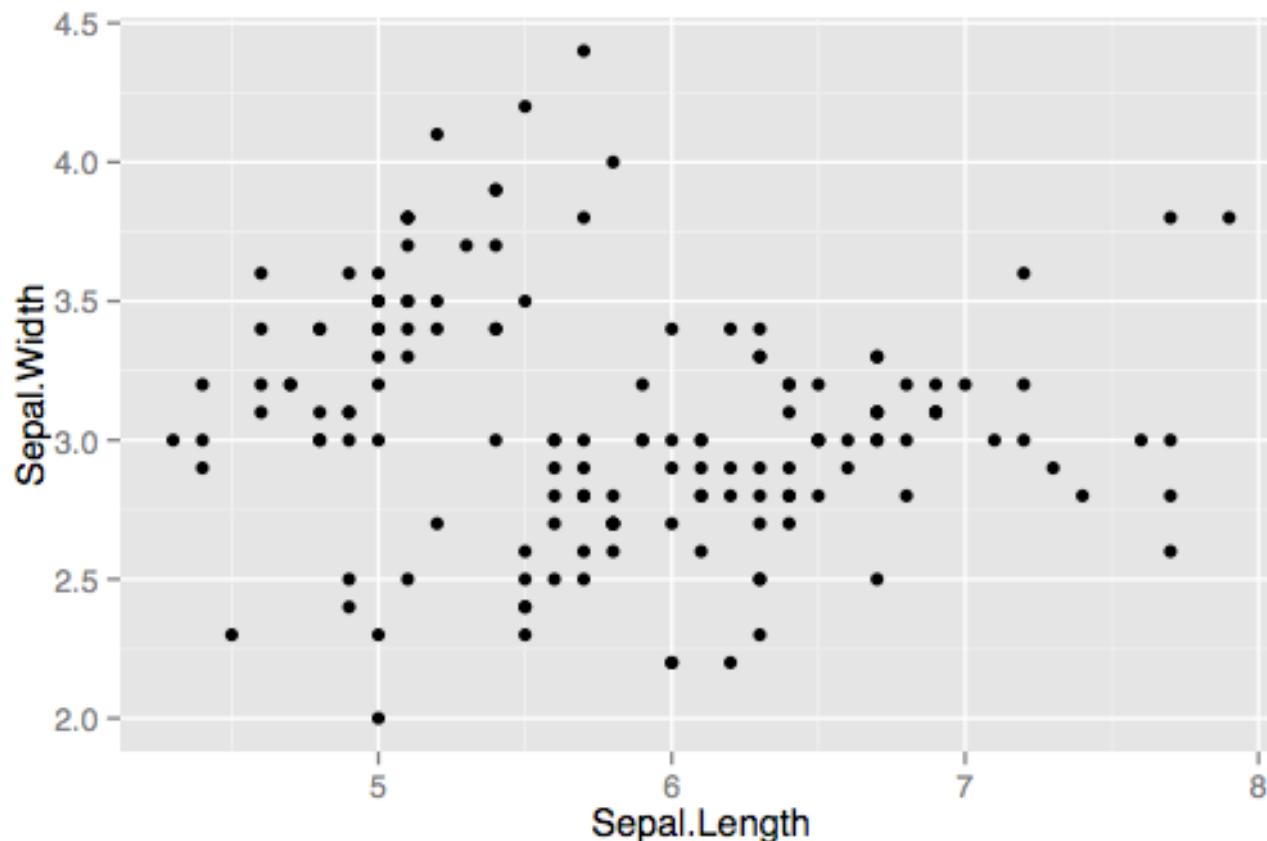
↑
Data

↑
Geometric objects to display

↑
Aesthetics map variables to scales

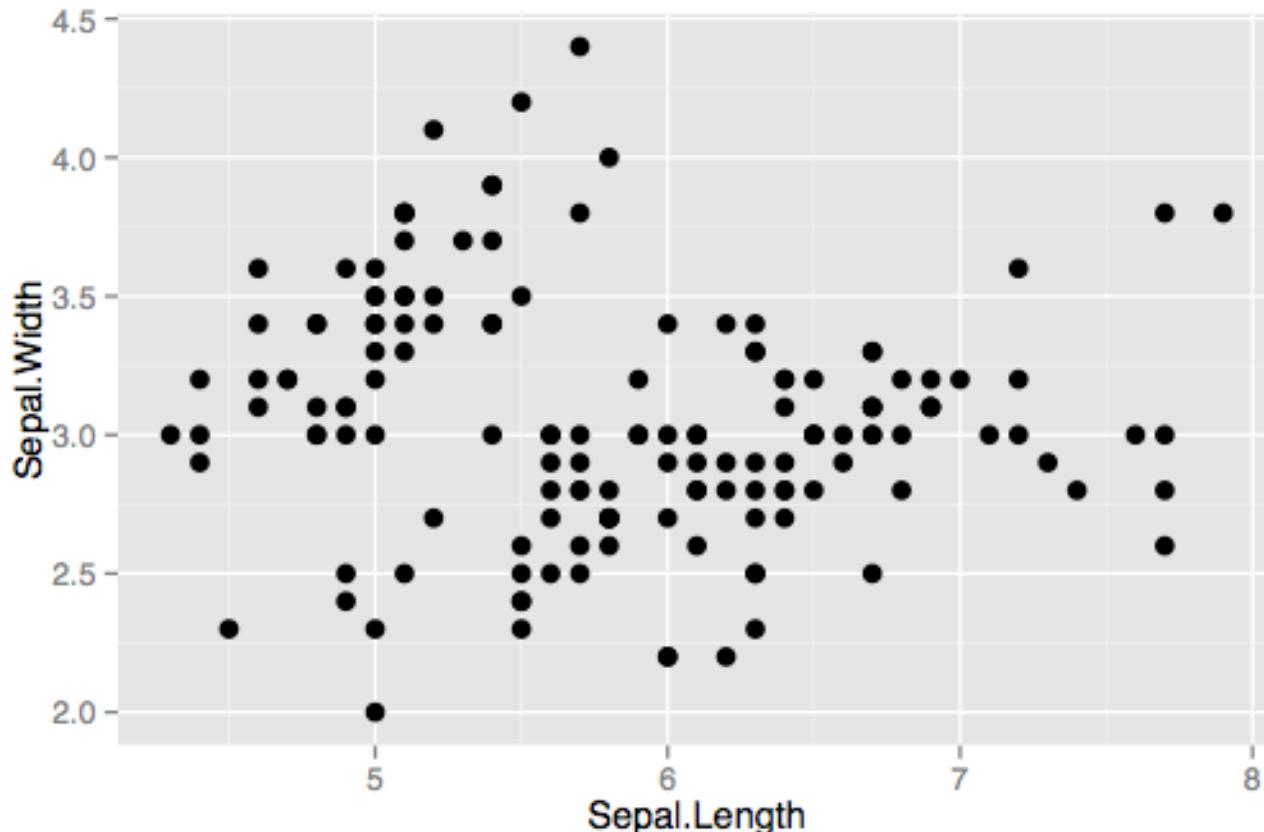
An Example: Visualizing iris Data

- `ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width)) +
geom_point()`



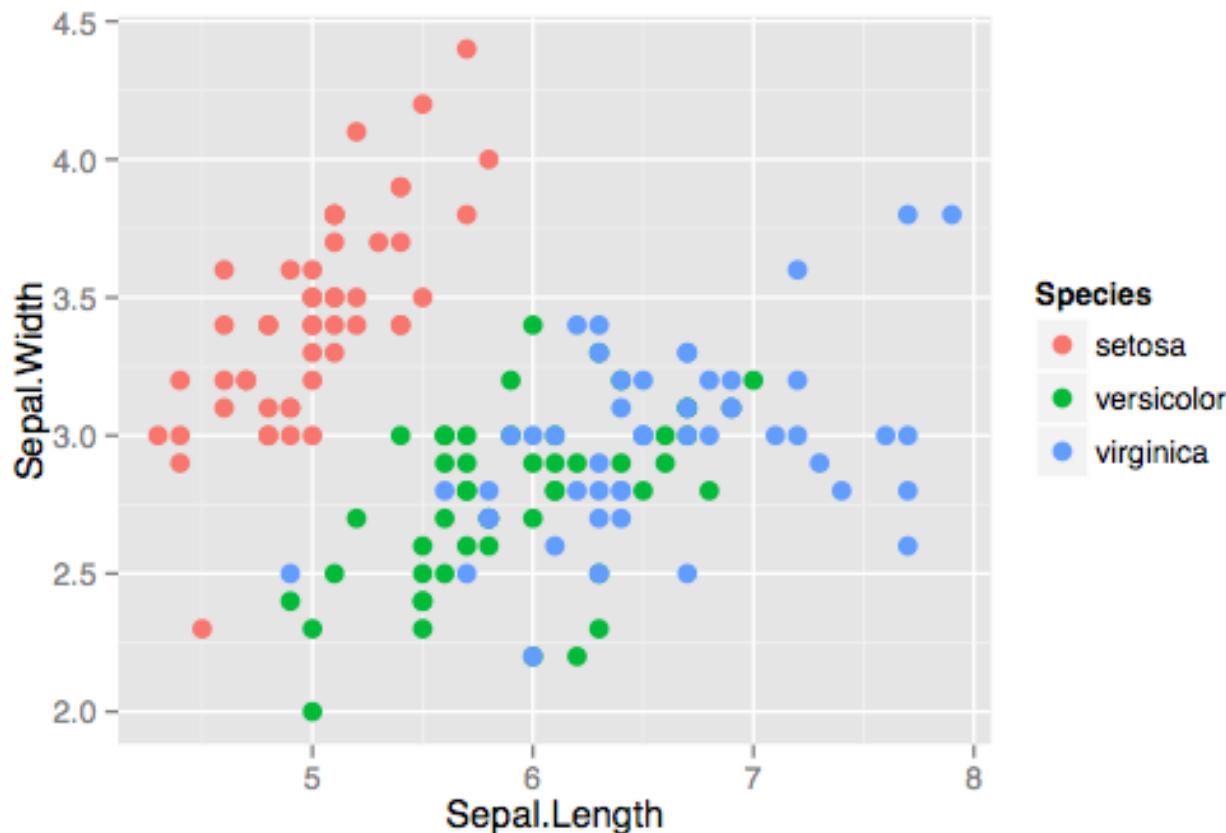
Changing the Aesthetics of a geom: increase the size of points

- `ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width)) +
geom_point(size = 3)`



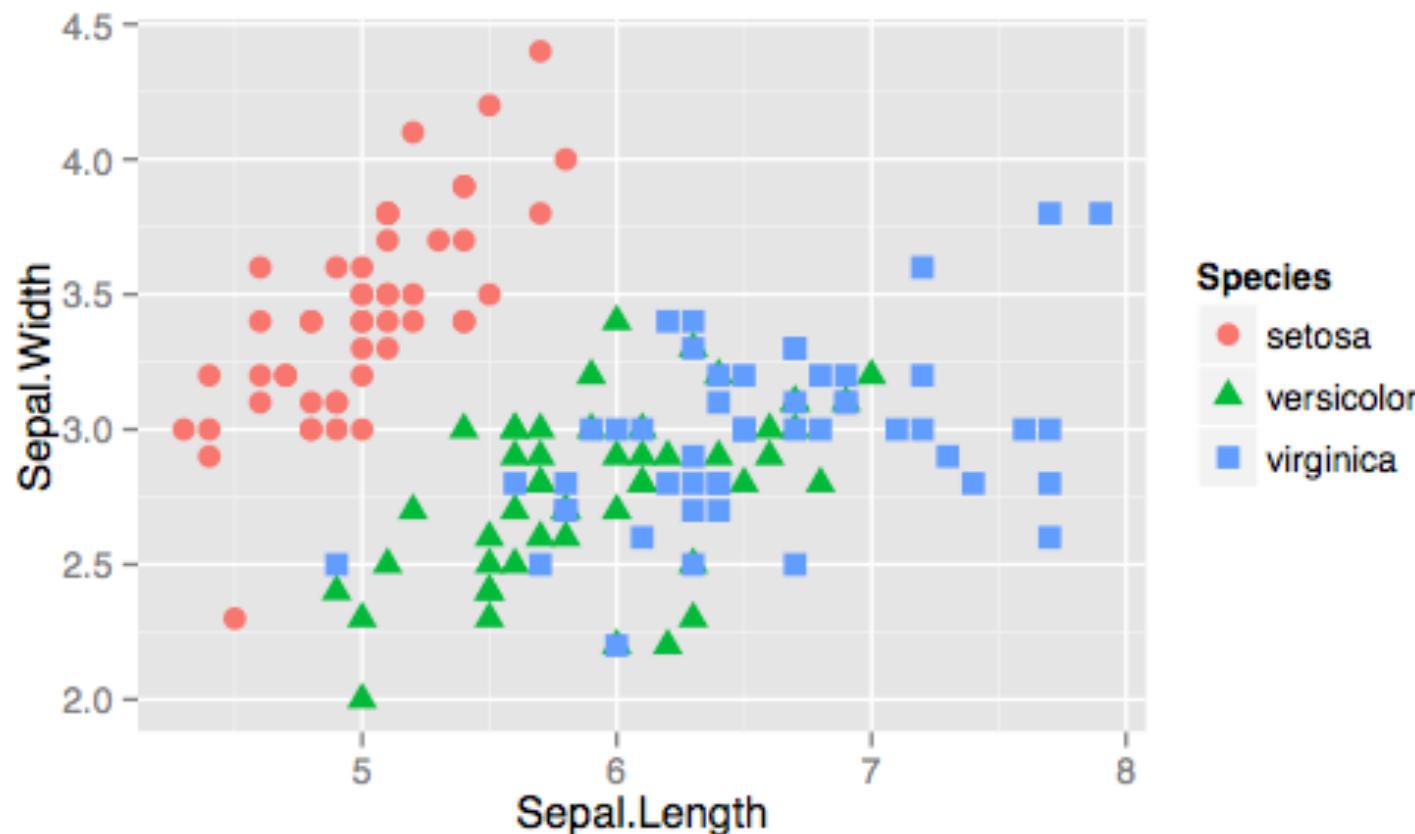
Changing the aesthetics of a geom: Add some color

- `ggplot(iris, aes(Sepal.Length, Sepal.Width, color = Species)) +
geom_point(size = 3)`



Changing the aesthetics of a geom: Differentiate points by shape

- `ggplot(iris, aes(Sepal.Length, Sepal.Width, color = Species)) +
geom_point(aes(shape = Species), size = 3)`

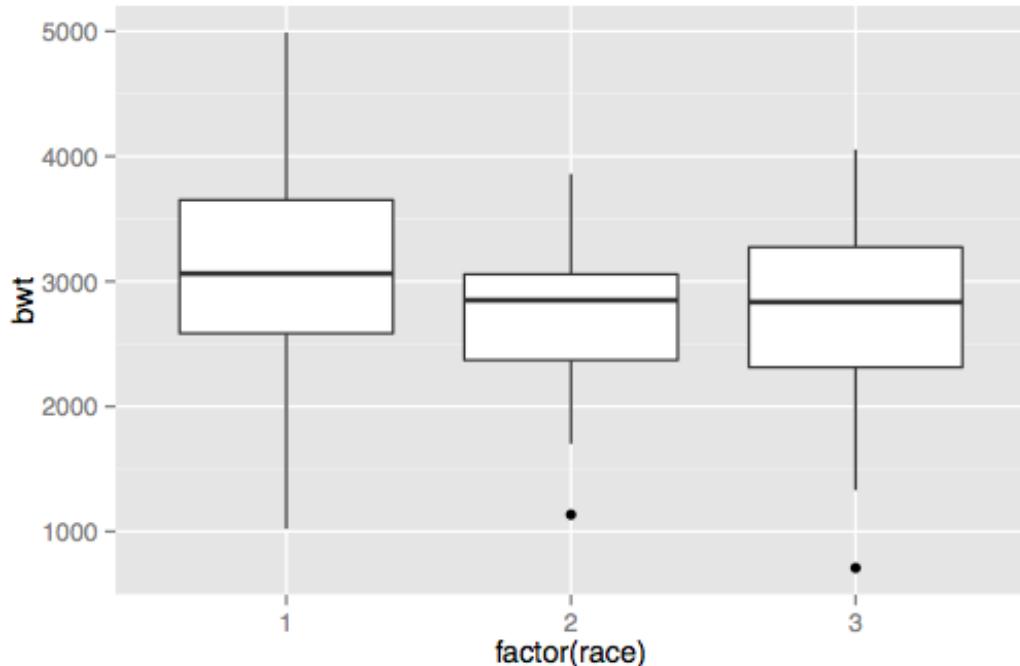


Stats (stat)

- Statistical transformations and data summary
 - All geoms have associated default stats, and vice versa
 - e.g. binning for a histogram or fitting a linear model

Example: boxplots

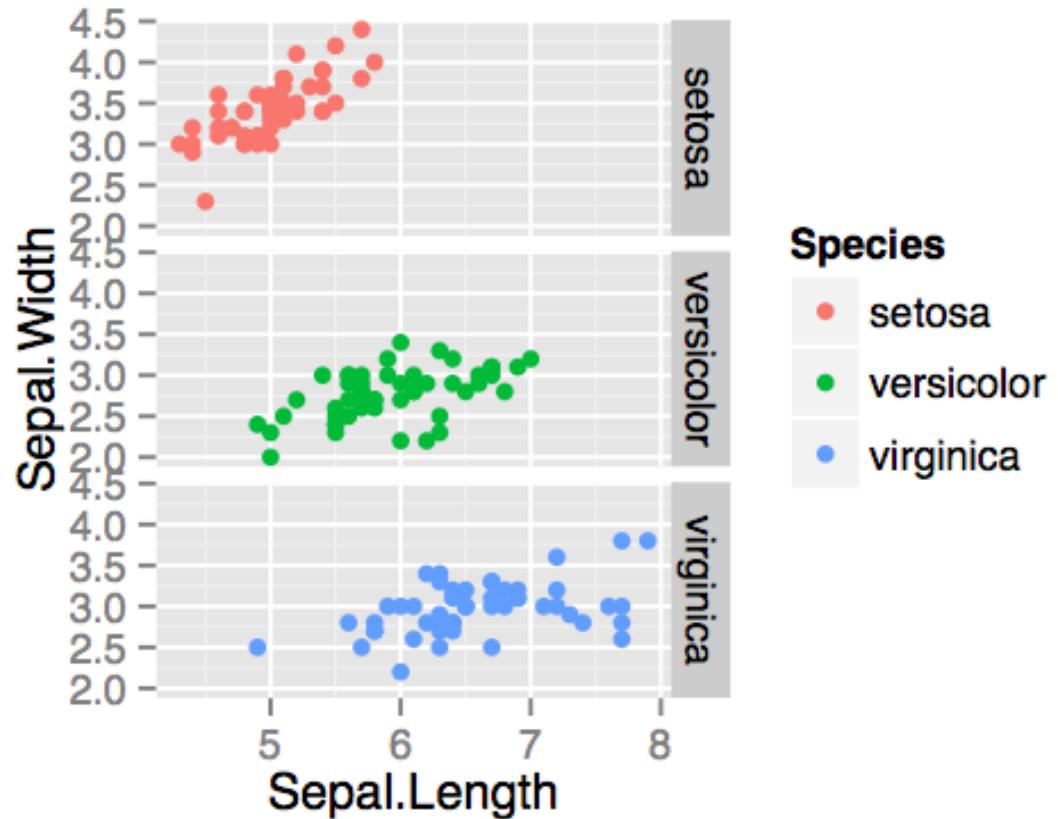
```
library(MASS)
ggplot(birthwt, aes(factor(race),
bwt)) + geom_boxplot()
```



Facets (facet)

- Subsetting data to make lattice plots
- An example: single column, multiple rows

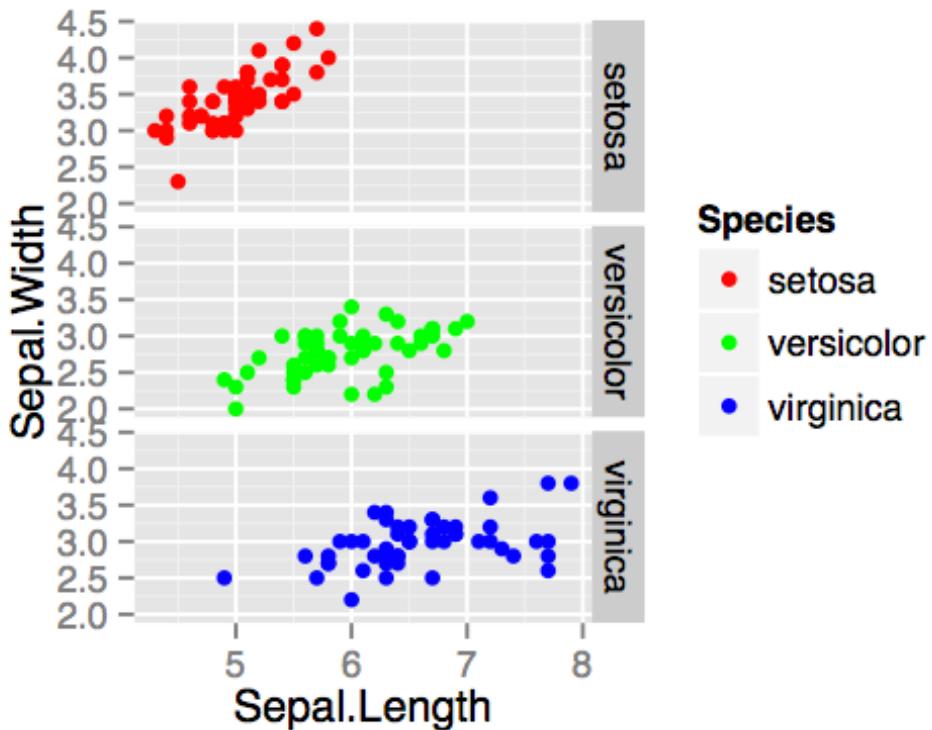
```
ggplot(iris, aes(Sepal.Length,  
Sepal.Width, color = Species)) +  
geom_point()  
+ facet_grid(Species ~ .)
```



Scales (scale)

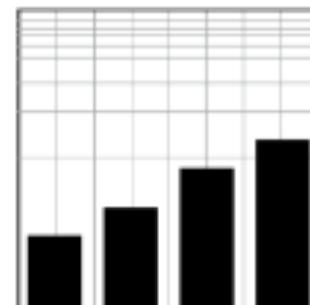
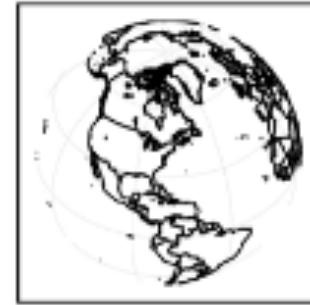
- Control the mapping from data to aesthetics
 - Often used for adjusting color mapping
- An example: manual color scale

```
ggplot(iris, aes(Sepal.Length,  
Sepal.Width, color = Species)) +  
  geom_point()  
+ facet_grid(Species ~.)  
+ scale_color_manual(values =  
c("red", "green", "blue"))
```



Coordinates (coord)

- put data on plane of graphic
 - e.g. polar coordinate plots
- Shortcut functions
 - `coord_cartesian`
 - `coord_polar()`
 - `coord_map()`
 - `coord_trans()`
- Will not cover this in detail



ggplot2 Help Topics

Help topics

Geoms

Geoms, short for geometric objects, describe the type of plot you will produce.

- [geom_abline](#) (geom_hline, geom_vline)
Lines: horizontal, vertical, and specified by slope and intercept.
- [geom_bar](#) (stat_count)
Bars, rectangles with bases on x-axis
- [geom_bin2d](#) (stat_bin2d, stat_bin_2d)
Add heatmap of 2d bin counts.
- [geom_blank](#)
Blank, draws nothing.
- [geom_boxplot](#) (stat_boxplot)
Box and whiskers plot.
- [geom_contour](#) (stat_contour)
Display contours of a 3d surface in 2d.
- [geom_count](#) (stat_sum)
Count the number of observations at each location.
- [geom_crossbar](#) (geom_errorbar, geom_linerange, geom_pointrange)
Vertical intervals: lines, crossbars & errorbars.
- [geom_density](#) (stat_density)
Display a smooth density estimate.
- [geom_density_2d](#) (geom_density2d, stat_density2d, stat_density_2d)
Contours from a 2d density estimate.
- [geom_dotplot](#)
Dot plot
- [geom_errorbarh](#)
Horizontal error bars
- [geom_freqpoly](#) (geom_histogram, stat_bin)
Histograms and frequency polygons.
- [geom_hex](#) (stat_bin_hex, stat_binhex)
Hexagon binning.



Write Functions for Day to Day Plots

- Call your function to generate a plot. It's a lot easier to fix one function that do it over and over for many plot

```
my_custom_plot <- function(df, title = "", ...) {  
  ggplot(df, ...) +  
  ggttitle(title) +  
  whatever_geoms() +  
  theme(...)  
}  
plot1 <- my_custom_plot(dataset1, title = "Figure 1")
```

Publication Quality Figures

- ▶ If the plot is on your screen

```
ggsave("~/path/to/figure/filename.png")
```

- ▶ If your plot is assigned to an object

```
ggsave(plot1, file = "~/path/to/figure/filename.png")
```

- ▶ Specify a size

```
ggsave(file = "/path/to/figure/filename.png", width = 6,  
height =4)
```

- ▶ or any format (pdf, png, eps, svg, jpg)

```
ggsave(file = "/path/to/figure/filename.eps")  
ggsave(file = "/path/to/figure/filename.jpg")  
ggsave(file = "/path/to/figure/filename.pdf")
```

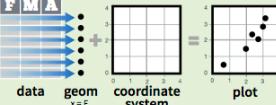
Data Visualization with ggplot2

Cheat Sheet

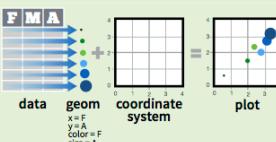


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Build a graph with **qplot()** or **ggplot()**

aesthetic mappings **data** **geom**

```
qplot(x = cty, y = hwy, color = cyl, data = mpg, geom = "point")
```

Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

ggplot(data = mpg, aes(x = cty, y = hwy))

Begins a plot that you finish by adding layers to. No defaults, but provides more control than qplot().

data

```
ggplot(mpg, aes(hwy, cty)) +  
  geom_point(aes(color = cyl)) +  
  geom_smooth(method = "lm") +  
  coord_cartesian() +  
  scale_color_gradient() +  
  theme_bw()
```

add layers, elements with +

layer = geom + default stat + layer specific mappings

additional elements

Add a new layer to a plot with a **geom_***() or **stat_***() function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

last_plot()

Returns the last plot

ggsave("plot.png", width = 5, height = 5)

Saves last plot as 5'x5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

One Variable

Continuous

```
a <- ggplot(mpg, aes(hwy))
```



a + geom_area(stat = "bin")

x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")

a + geom_density(kernel = "gaussian")

x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..density..))

a + geom_dotplot()

x, y, alpha, color, fill



a + geom_freqpoly()

x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))

a + geom_histogram(binwidth = 5)

x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

Discrete

```
b <- ggplot(mpg, aes(fl))
```



b + geom_bar()

x, alpha, color, fill, linetype, size, weight

Graphical Primitives

```
c <- ggplot(map, aes(long, lat))
```



c + geom_polygon(aes(group = group))

x, y, alpha, color, fill, linetype, size



d + geom_path(lineend = "butt",

linejoin = "round", linemetre = 1)

x, y, alpha, color, linetype, size



d + geom_ribbon(aes(ymin = unemploy - 900,

ymax = unemploy + 900))

x, ymax, ymin, alpha, color, fill, linetype, size



e + geom_segment(aes(

xend = long + delta_long,

yend = lat + delta_lat))

x, xend, y, yend, alpha, color, linetype, size



e + geom_rect(aes(xmin = long, ymin = lat,

xmax = long + delta_long,

ymax = lat + delta_lat))

xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

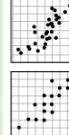
Two Variables

Continuous X, Continuous Y

```
f <- ggplot(mpg, aes(cty, hwy))
```



f + geom_blank()



f + geom_jitter()

x, y, alpha, color, fill, shape, size



f + geom_point()

x, y, alpha, color, fill, shape, size



f + geom_quantile()

x, y, alpha, color, linetype, size, weight



f + geom_rug(sides = "bl")

alpha, color, linetype, size



f + geom_smooth(model = lm)

x, y, alpha, color, fill, linetype, size, weight

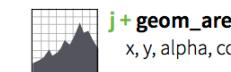


f + geom_text(aes(label = cty))

x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

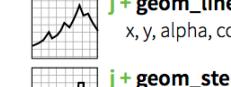
Continuous Function

```
j <- ggplot(economics, aes(date, unemploy))
```



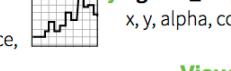
j + geom_area()

x, y, alpha, color, fill, linetype, size



j + geom_line()

x, y, alpha, color, linetype, size

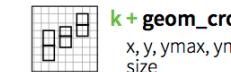


j + geom_step(direction = "hv")

x, y, alpha, color, linetype, size

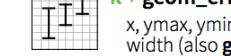
Visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
k <- ggplot(df, aes(grp, fit, ymin = fit - se, ymax = fit + se))
```



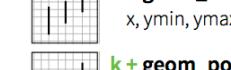
k + geom_crossbar(fatten = 2)

x, y, ymax, ymin, alpha, color, fill, linetype, size



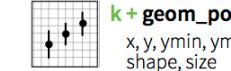
k + geom_errorbar()

x, ymax, ymin, alpha, color, linetype, size, width (also **geom_errorbarh()**)



k + geom_linerange()

x, ymin, ymax, alpha, color, linetype, size



k + geom_pointrange()

x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

Maps

```
data <- data.frame(murder = USArrests$Murder,
state = tolower(rownames(USArrests)))
map <- map_data("state")
l <- ggplot(data, aes(fill = murder))
```



l + geom_map(aes(map_id = state), map = map) + expand_limits(x = map\$long, y = map\$lat)

map_id, alpha, color, fill, linetype, size

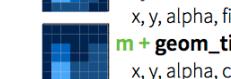
Three Variables

```
seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))
m <- ggplot(seals, aes(long, lat))
```



m + geom_raster(aes(fill = z), hjust = 0.5,

vjust = 0.5, interpolate = FALSE)



m + geom_contour(aes(z = z))

x, y, z, alpha, colour, linetype, size, weight

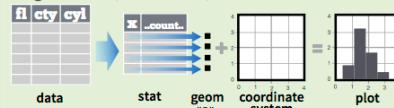


m + geom_tile(aes(fill = z))

x, y, alpha, color, fill, linetype, size

Stats - An alternative way to build a layer

Some plots visualize a **transformation** of the original data set. Use a **stat** to choose a common transformation to visualize, e.g. `a + geom_bar(stat = "bin")`



Each stat creates additional variables to map aesthetics to. These variables use a common `..name..` syntax.

stat functions and geom functions both combine a stat with a geom to make a layer, i.e. `stat_bin(geom="bar")` does the same as `geom_bar(stat="bin")`

```
stat function      layer specific mappings      variable created by transformation
i + stat_density2d(aes(fill = ..level..), geom = "polygon", n = 100)
geom for layer    parameters for stat
```

`a + stat_bin(binwidth = 1, origin = 10)` 1D distributions
`x, y | ..count.., ..ncount.., ..density.., ..ndensity..`
`a + stat_bindot(binwidth = 1, binaxis = "x")`
`x, y | ..count.., ..ncount..`
`a + stat_density(adjust = 1, kernel = "gaussian")`
`x, y | ..count.., ..density.., ..scaled..`

`f + stat_box2d(bins = 30, drop = TRUE)` 2D distributions
`x, y, fill | ..count.., ..density..`
`f + stat_hexbin(bins = 30)`
`x, y, fill | ..count.., ..density..`
`f + stat_density2d(contour = TRUE, n = 100)`
`x, y, color, size | ..level..`

`m + stat_contour(aes(z = z))` 3 Variables
`x, y, z, order | ..level..`
`m + stat_spoke(aes(radius = z, angle = z))`
`angle, radius, x, xend, y, yend | ..x.., ..xend.., ..y.., ..yend..`
`m + stat_summary_hex(aes(z = z), bins = 30, fun = mean)`
`x, y, z, fill | ..value..`
`m + stat_summary2d(aes(z = z), bins = 30, fun = mean)`
`x, y, z, fill | ..value..`

`g + stat_boxplot(coef = 1.5)` Comparisons
`x, y | ..lower.., ..middle.., ..upper.., ..outliers..`
`g + stat_ydensity(adjust = 1, kernel = "gaussian", scale = "area")`
`x, y | ..density.., ..scaled.., ..count.., ..n.., ..violinwidth.., ..width..`

`f + stat_ecdf(n = 40)` Functions
`x, y | ..x.., ..y..`
`f + stat_quantile(quantiles = c(0.25, 0.5, 0.75), formula = y ~ log(x), method = "rq")`
`x, y | ..quartile.., ..x.., ..y..`
`f + stat_smooth(method = "auto", formula = y ~ x, se = TRUE, n = 80, fullrange = FALSE, level = 0.95)`
`x, y | ..se.., ..x.., ..y.., ..ymin.., ..ymax..`

`ggplot() + stat_function(aes(x = -3:3), fun = dnorm, n = 101, args = list(sd = 0.5))` General Purpose
`x | ..y..`
`f + stat_identity()`
`ggplot() + stat_qq(aes(sample = 1:100), distribution = qt, dparams = list(df = 5))`
`sample, x, y | ..x.., ..y..`
`f + stat_sum()`
`x, y, size | ..size..`
`f + stat_summary(fun.data = "mean_cl_boot")`
`f + stat_unique()`

Scales

Scales control how a plot maps data values to the visual values of an aesthetic. To change the mapping, add a custom scale.



General Purpose scales

Use with any aesthetic:
`alpha, color, fill, linetype, shape, size`

`scale_*_continuous()` - map cont' values to visual values
`scale_*_discrete()` - map discrete values to visual values
`scale_*_identity()` - use data values as visual values
`scale_*_manual(values = c())` - map discrete values to manually chosen visual values

X and Y location scales

Use with `x` or `y` aesthetics (`x` shown here)

`scale_x_date(labels = date_format("%m/%d"), breaks = date_breaks("2 weeks"))` - treat `x` values as dates. See `?strptime` for label formats.
`scale_x_datetime()` - treat `x` values as date times. Use same arguments as `scale_x_date()`.
`scale_x_log10()` - Plot `x` on log10 scale
`scale_x_reverse()` - Reverse direction of `x` axis
`scale_x_sqrt()` - Plot `x` on square root scale

Color and fill scales

<p>Discrete</p> <p><code>n <- b + geom_bar(aes(fill = fl))</code></p> <p><code>n + scale_fill_brewer(palette = "Blues")</code> For palette choices: library(RColorBrewer); display.brewer.all()</p> <p><code>n + scale_fill_grey(start = 0.2, end = 0.8, na.value = "red")</code></p>	<p>Continuous</p> <p><code>o <- a + geom_dotplot(aes(fill = ...))</code></p> <p><code>o + scale_fill_gradient(low = "red", high = "yellow")</code></p> <p><code>o + scale_fill_gradient2(low = "red", high = "blue", mid = "white", midpoint = 25)</code> Also: rainbow(), heat.colors(6), topo.colors(), cm.colors(), RColorBrewer::brewer.pal()</p> <p><code>o + scale_fill_gradientn(colours = terrain.colors(6))</code></p>
--	--

Shape scales

<p>Manual shape values</p> <p><code>p <- f + geom_point(aes(shape = fl))</code></p> <p><code>p + scale_shape(solid = FALSE)</code></p> <p><code>p + scale_shape_manual(values = c(3:7))</code> Shape values shown in chart on right</p>	<p><code>p + geom_point(aes(size = cyl))</code></p> <p><code>q + scale_size_area(max = 6)</code> Value mapped to area of circle (not radius)</p>
--	--

Size scales

Coordinate Systems

`r <- b + geom_bar()`
`r + coord_cartesian(xlim = c(0, 5))`
`xlim, ylim`

The default cartesian coordinate system

`r + coord_fixed(ratio = 1/2)`
`ratio, xlim, ylim`

Cartesian coordinates with fixed aspect ratio between `x` and `y` units

`r + coord_flip()`
`xlim, ylim`

Flipped Cartesian coordinates

`r + coord_polar(theta = "x", direction = 1)`
`theta, start, direction`

Polar coordinates

`r + coord_trans(ytrans = "sqrt")`
`xtrans, ytrans, limx, limy`

Transformed cartesian coordinates. Set `extras` and `strains` to the name of a window function.

`z + coord_map(projection = "ortho", orientation = c(41, -74, 0))`
`projection, orientation, xlim, ylim`

Map projections from the mapproj package

(mercator (default), azequalarea, lagrange, etc.)

Facets divide a plot into subplots based on the values of one or more discrete variables.

`t <- ggplot(mpg, aes(cty, hwy)) + geom_point()`

`t + facet_grid(. ~ fl)`
facet into columns based on `fl`

`t + facet_grid(year ~ .)`
facet into rows based on `year`

`t + facet_grid(year ~ fl)`
facet into both rows and columns

`t + facet_wrap(~ fl)`
wrap facets into a rectangular layout

Set `scales` to let axis limits vary across facets

`t + facet_grid(y ~ x, scales = "free")`
`x and y axis limits adjust to individual facets`

- `"free_x"` - `x` axis limits adjust
- `"free_y"` - `y` axis limits adjust

Set `labeler` to adjust facet labels

<code>t + facet_grid(~ fl, labeller = label_both)</code>	<code>fl: c</code>	<code>fl: d</code>	<code>fl: e</code>	<code>fl: p</code>	<code>fl: r</code>
<code>t + facet_grid(~ fl, labeller = label_bquote(alpha ^ .(x)))</code>	<code>alpha^c</code>	<code>alpha^d</code>	<code>alpha^e</code>	<code>alpha^p</code>	<code>alpha^r</code>
<code>t + facet_grid(~ fl, labeller = label_parsed)</code>	<code>c</code>	<code>d</code>	<code>e</code>	<code>p</code>	<code>r</code>

Position Adjustments

Position adjustments determine how to arrange geoms that would otherwise occupy the same space.

`s <- ggplot(mpg, aes(fl, fill = drv))`

`s + geom_bar(position = "dodge")`
Arrange elements side by side

`s + geom_bar(position = "fill")`
Stack elements on top of one another, normalize height

`s + geom_bar(position = "stack")`
Stack elements on top of one another

`f + geom_point(position = "jitter")`
Add random noise to `X` and `Y` position of each element to avoid overplotting

Each position adjustment can be recast as a function with manual `width` and `height` arguments

`s + geom_bar(position = position_dodge(width = 1))`

Use scale functions to update legend labels

Labels

`t + ggtitle("New Plot Title")`

Add a main title above the plot

`t + xlab("New X label")`

Change the label on the `X` axis

`t + ylab("New Y label")`

Change the label on the `Y` axis

`t + labs(title = "New title", x = "New x", y = "New y")`
All of the above

Legends

`t + theme(legend.position = "bottom")`

Place legend at "bottom", "top", "left", or "right"

`t + guides(color = "none")`

Set legend type for each aesthetic: colorbar, legend, or none (no legend)

`t + scale_fill_discrete(name = "Title", labels = c("A", "B", "C"))`

Set legend title and labels with a scale function.

Themes

`r + theme_bw()`
White background with grid lines

`r + theme_classic()`
White background no gridlines

`r + theme_grey()`
Grey background (default theme)

`r + theme_minimal()`
Minimal theme

`ggthemes` - Package with additional ggplot2 themes

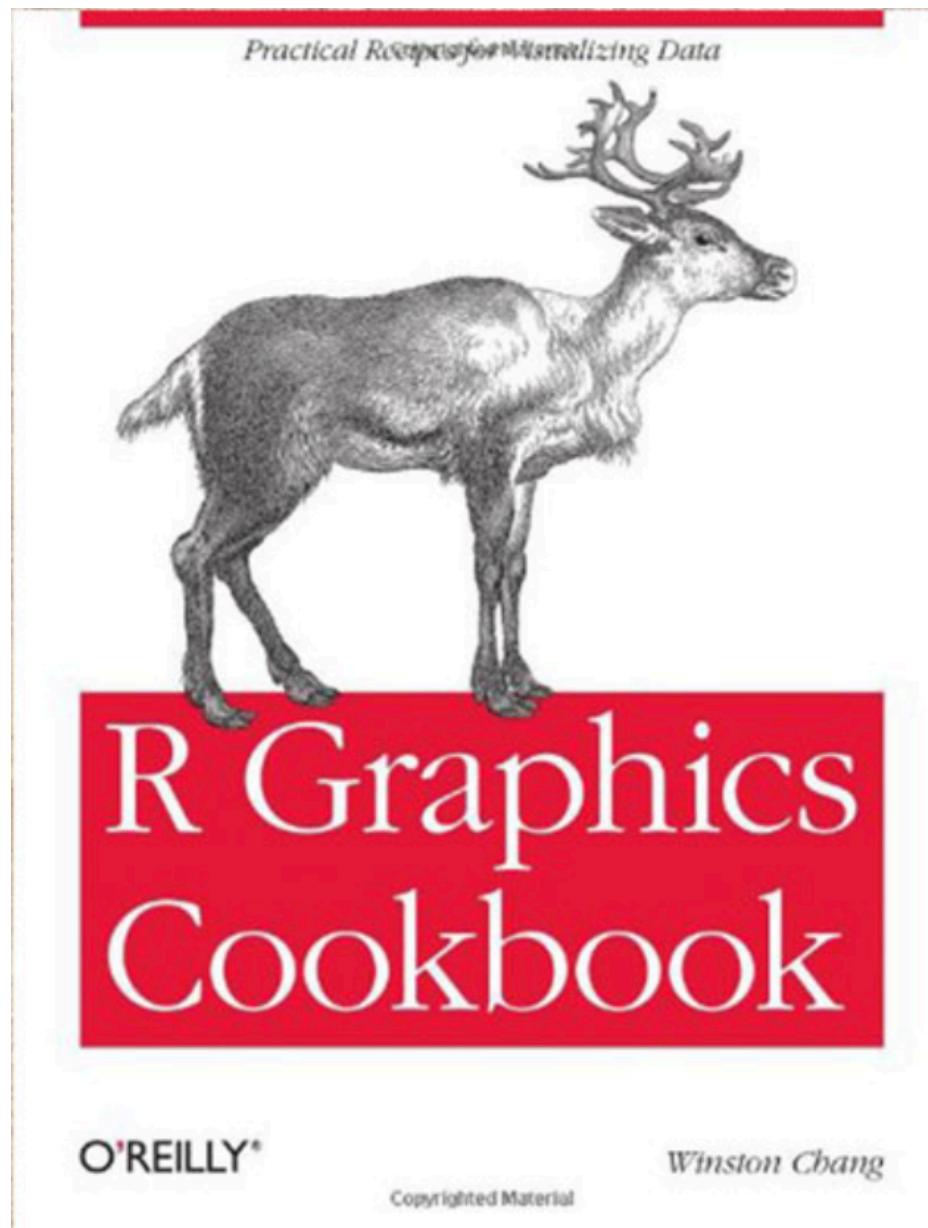
Without clipping (preferred)

`t + coord_cartesian(xlim = c(0, 100), ylim = c(10, 20))`

With clipping (removes unseen data points)

`t + xlim(0, 100) + ylim(10, 20)`

`t + scale_x_continuous(limits = c(0, 100)) + scale_y_continuous(limits = c(0, 100))`



Basic Plots

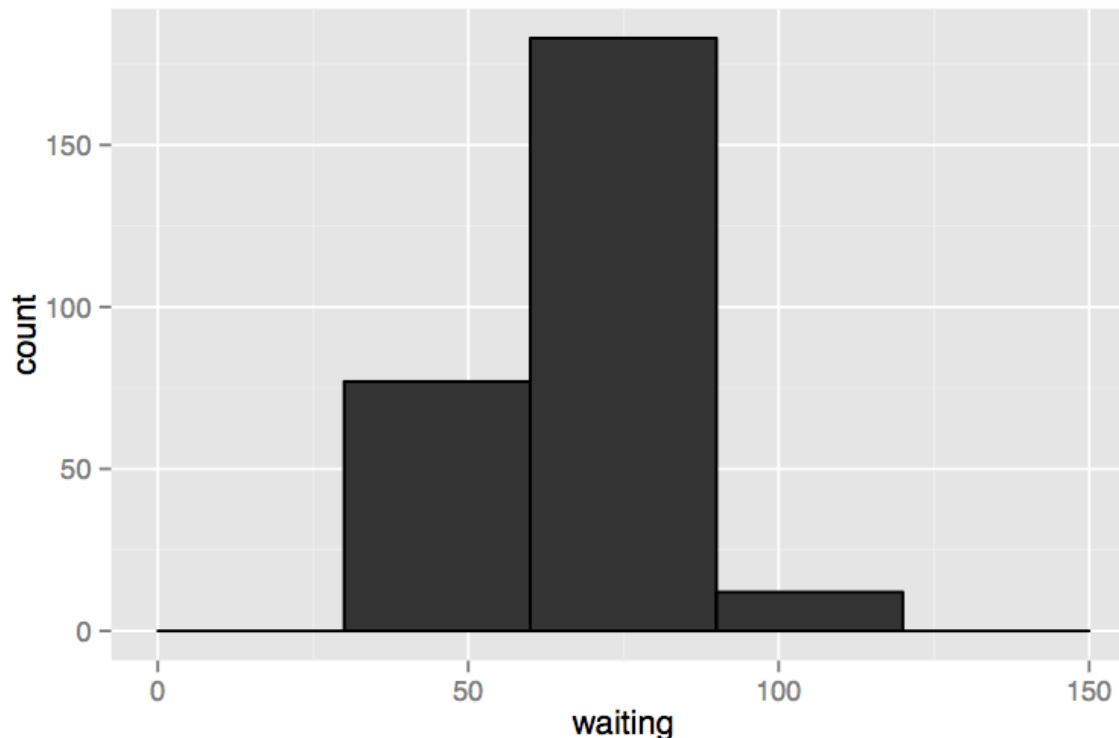
Histograms and Bar Plots

x axis is	Height of bar represents	Common name
<i>Continuous</i>	<i>Count</i>	Histogram
<i>Discrete</i>	<i>Count</i>	Bar graph
<i>Continuous</i>	<i>Value</i>	Bar graph
<i>Discrete</i>	<i>Value</i>	Bar graph

Histograms

- See `?geom_histogram` for list of options

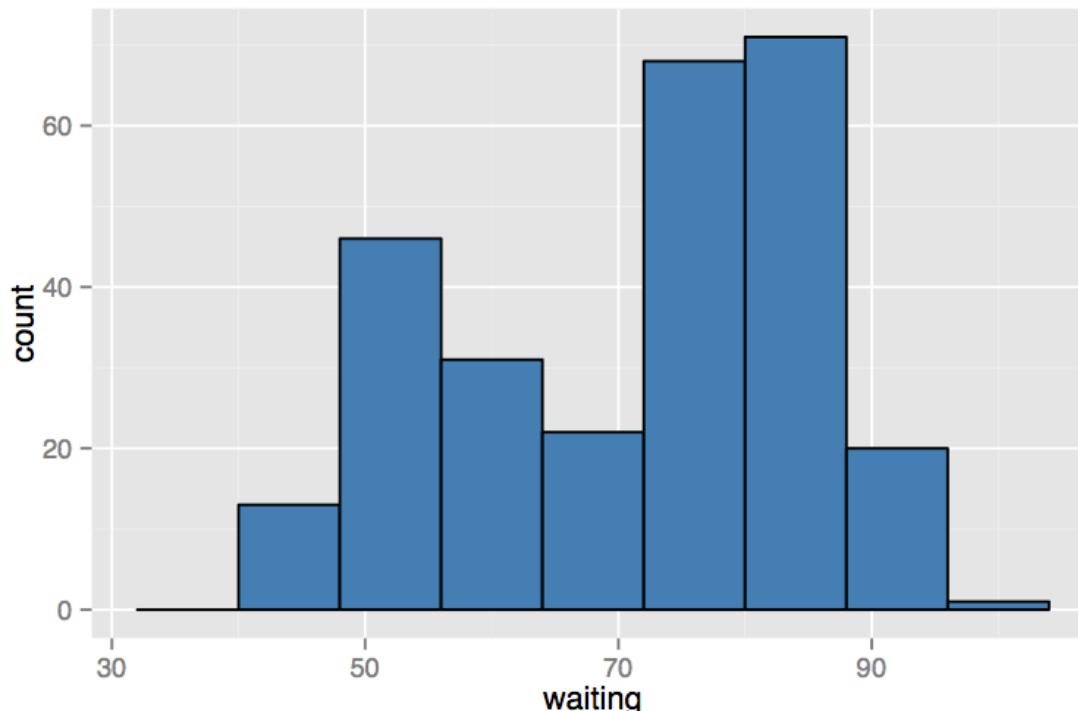
```
h <- ggplot(faithful, aes(x = waiting))  
h + geom_histogram(binwidth = 30, colour = "black")
```



Histograms

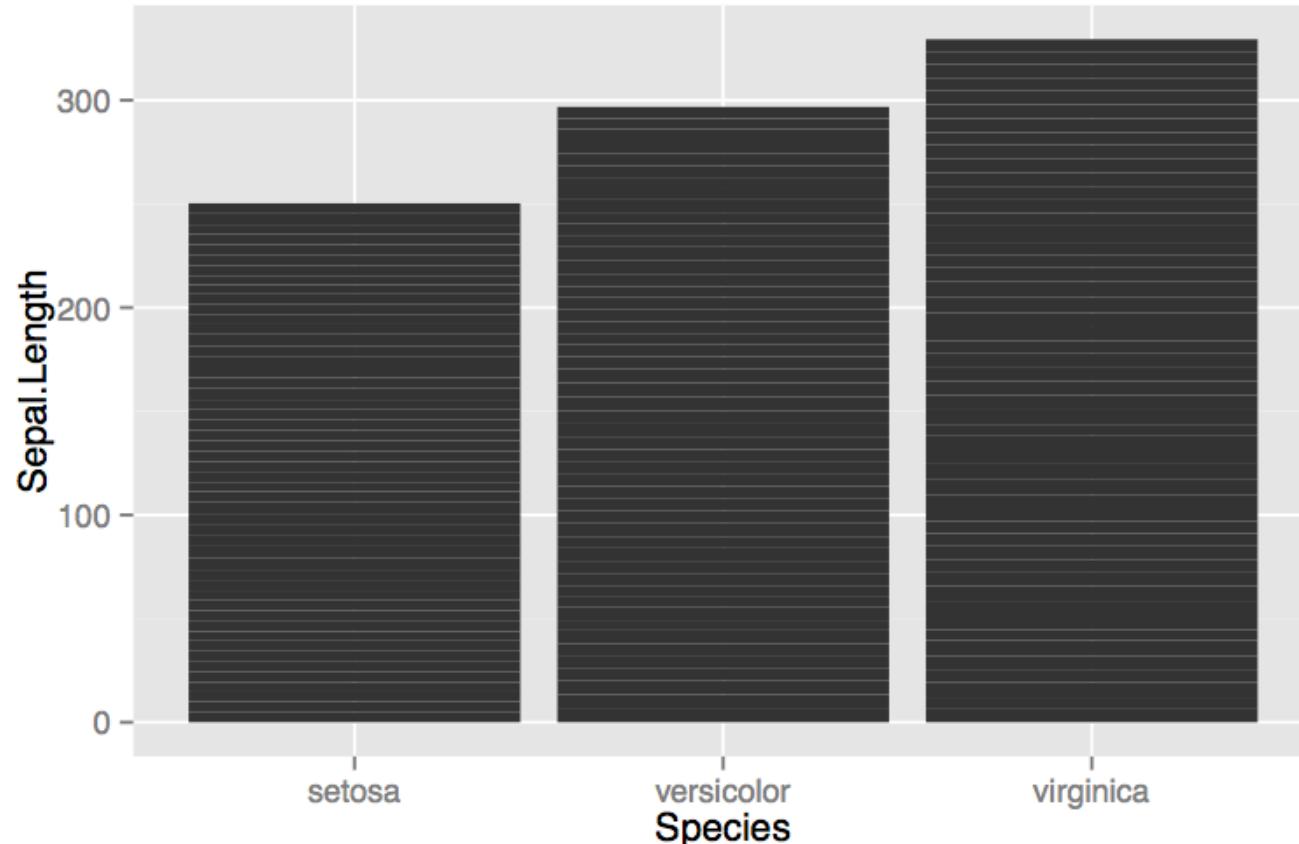
- See `?geom_histogram` for list of options

```
h <- ggplot(faithful, aes(x = waiting))  
h + geom_histogram(binwidth = 8, fill = "steelblue",  
colour = "black")
```



Bar Plots

```
ggplot(iris, aes(Species, Sepal.Length)) +  
  geom_bar(stat = "identity")
```



Bar Plots

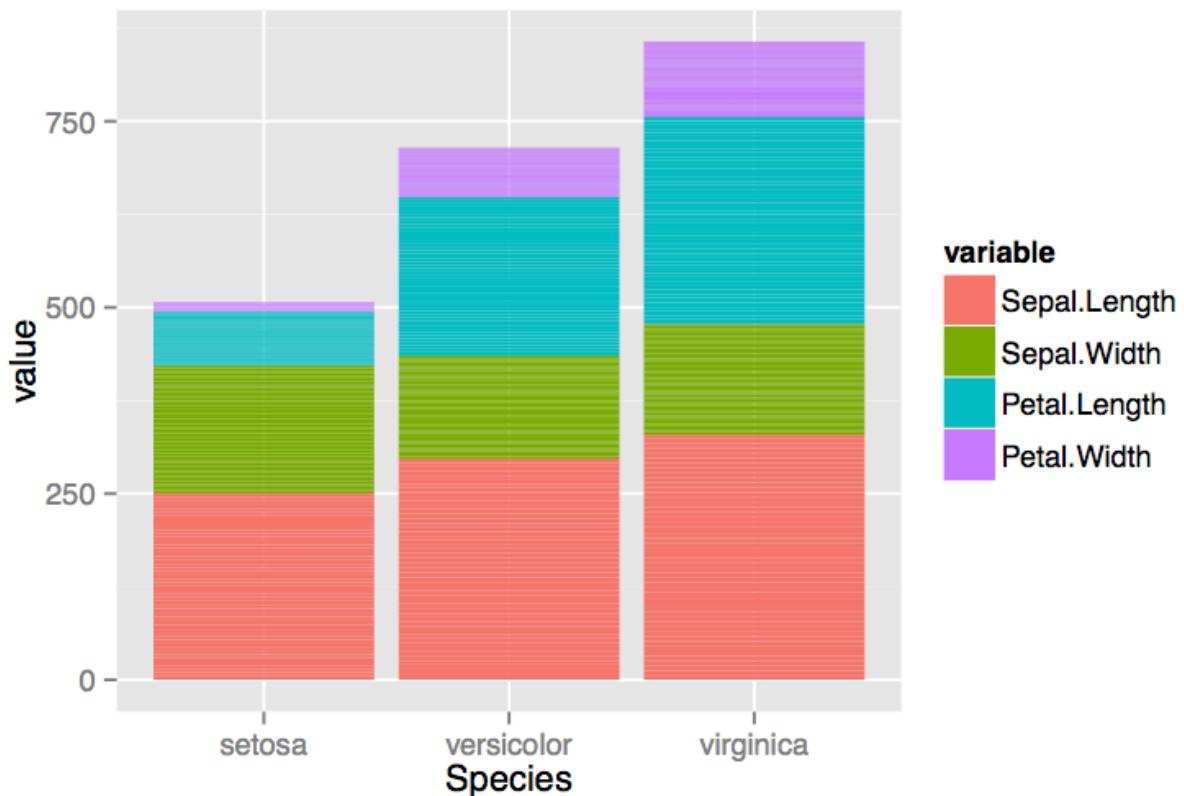
```
df <- melt(iris, id.vars = "Species")
ggplot(df, aes(Species, value, fill = variable)) +
  geom_bar(stat = "identity")
```

		id	time	x1	x2
1	1	1	1	5	6
1	2	1	2	3	5
2	1	2	1	6	1
2	2	2	2	2	4

melt(dat,
id=c("id","time"))

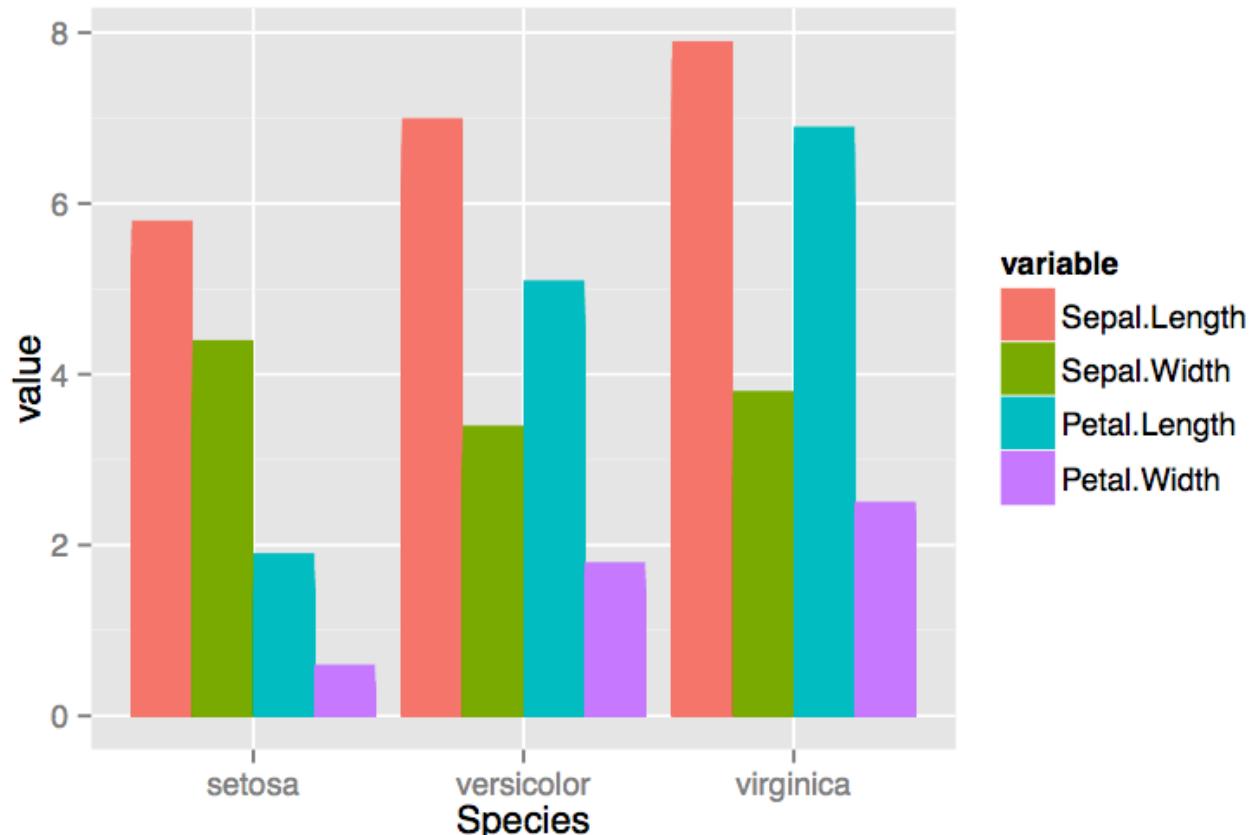


		id	time	variable	value
1	1	1	1	x1	5
1	2	1	2	x1	3
2	1	2	1	x1	6
2	2	2	2	x1	2
1	1	1	2	x2	6
1	2	1	2	x2	5
2	1	2	1	x2	1
2	2	2	2	x2	4



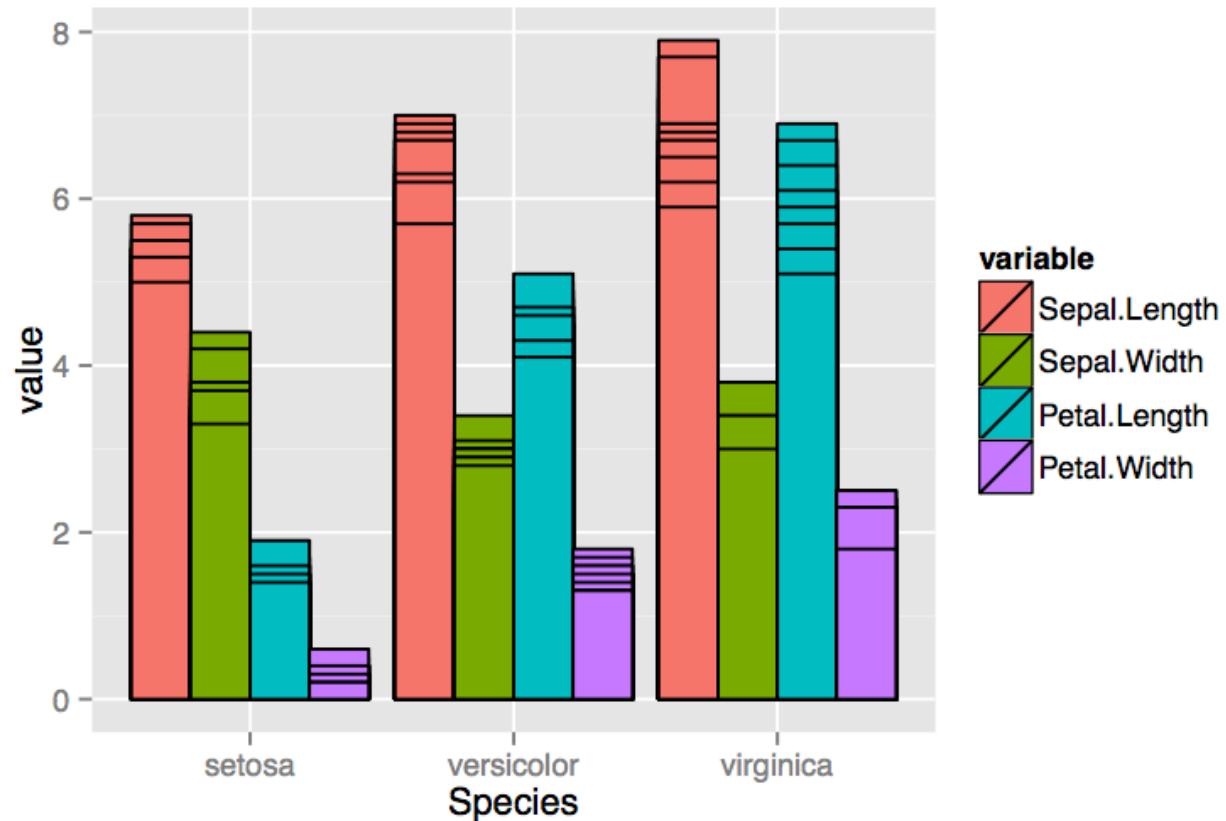
Bar Plots

```
ggplot(df, aes(Species, value, fill = variable)) +  
  geom_bar(stat = "identity", position = "dodge")
```



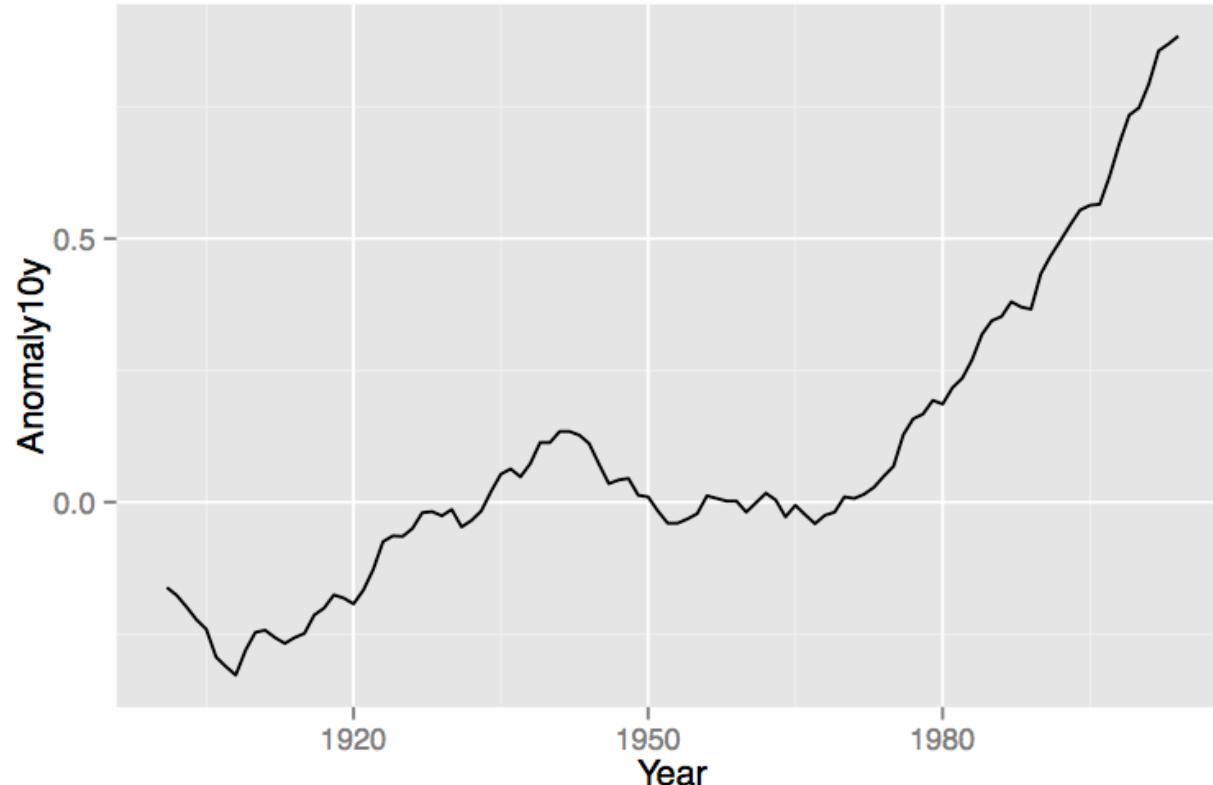
Bar Plots

```
ggplot(df, aes(Species, value, fill = variable)) +  
  geom_bar(stat = "identity", position="dodge", color="black")
```



Line Graphs

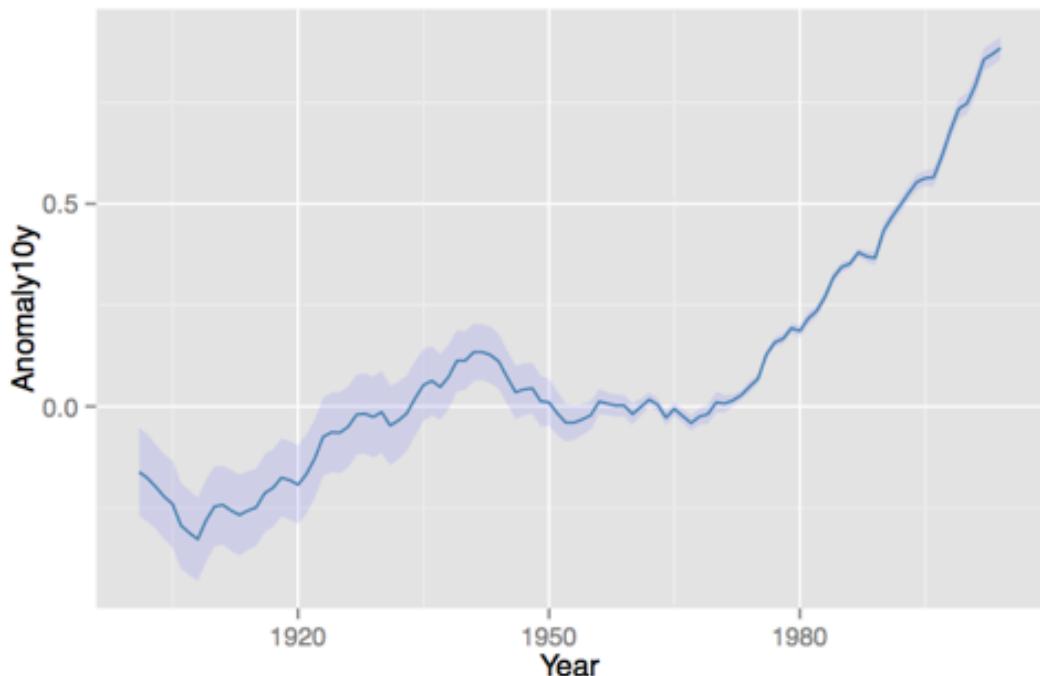
```
climate <- read.csv("data/climate.csv", header = T)
ggplot(climate, aes(Year, Anomaly10y)) +
  geom_line()
```



Line Graphs

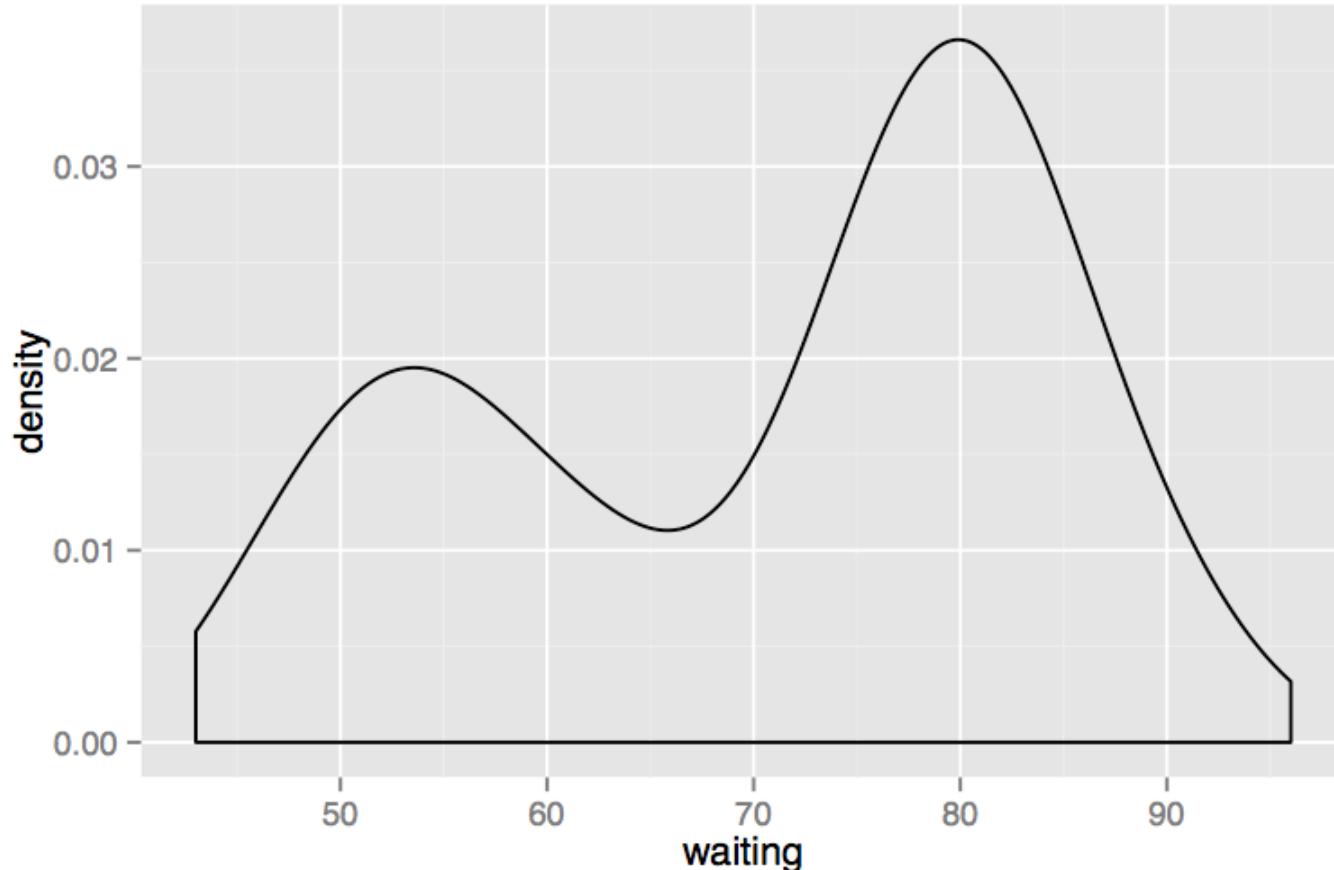
- Plot confidence regions

```
climate <- read.csv("data/climate.csv", header = T)
ggplot(climate, aes(Year, Anomaly10y)) +
  geom_ribbon(aes(ymin = Anomaly10y - Unc10y,
                  ymax = Anomaly10y + Unc10y),
              fill = "blue", alpha = .1) +
  geom_line(color = "steelblue")
```



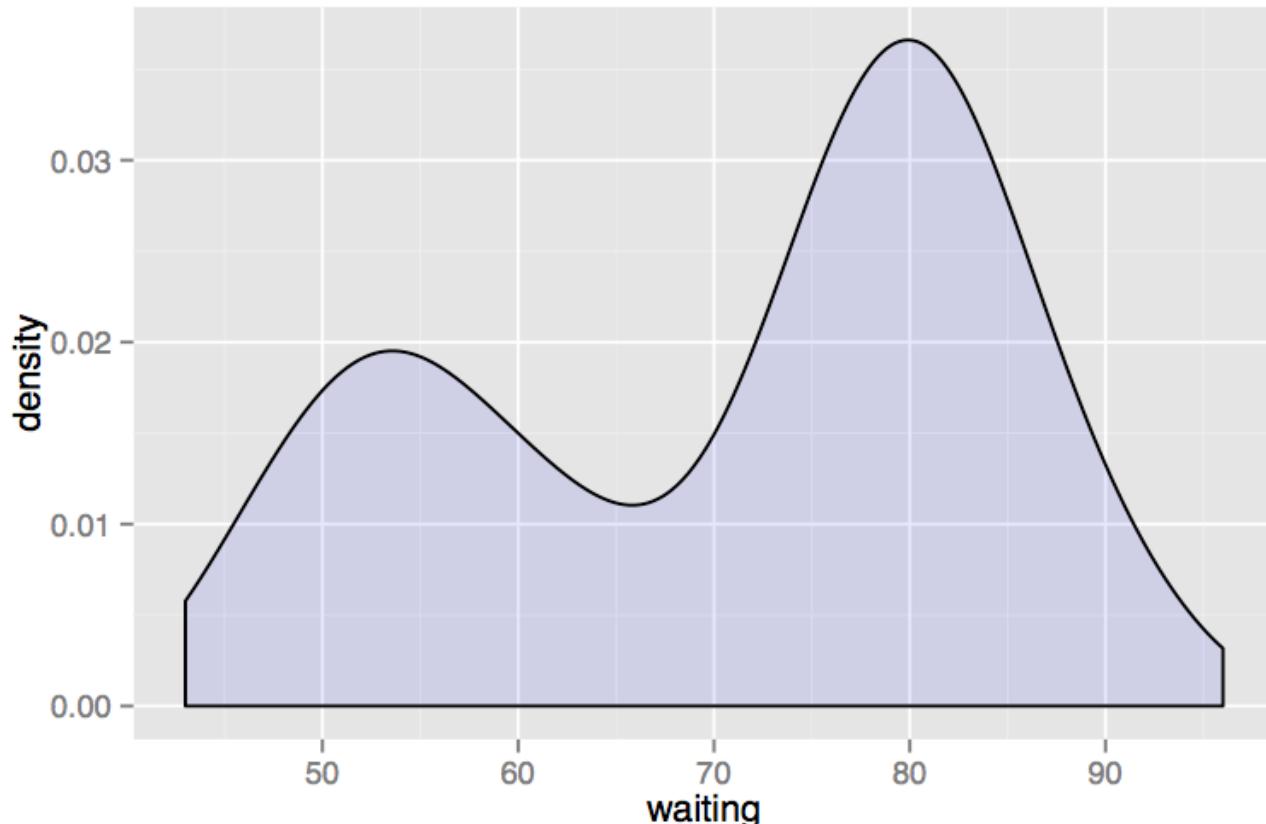
Density Plots

```
ggplot(faithful, aes(waiting)) + geom_density()
```



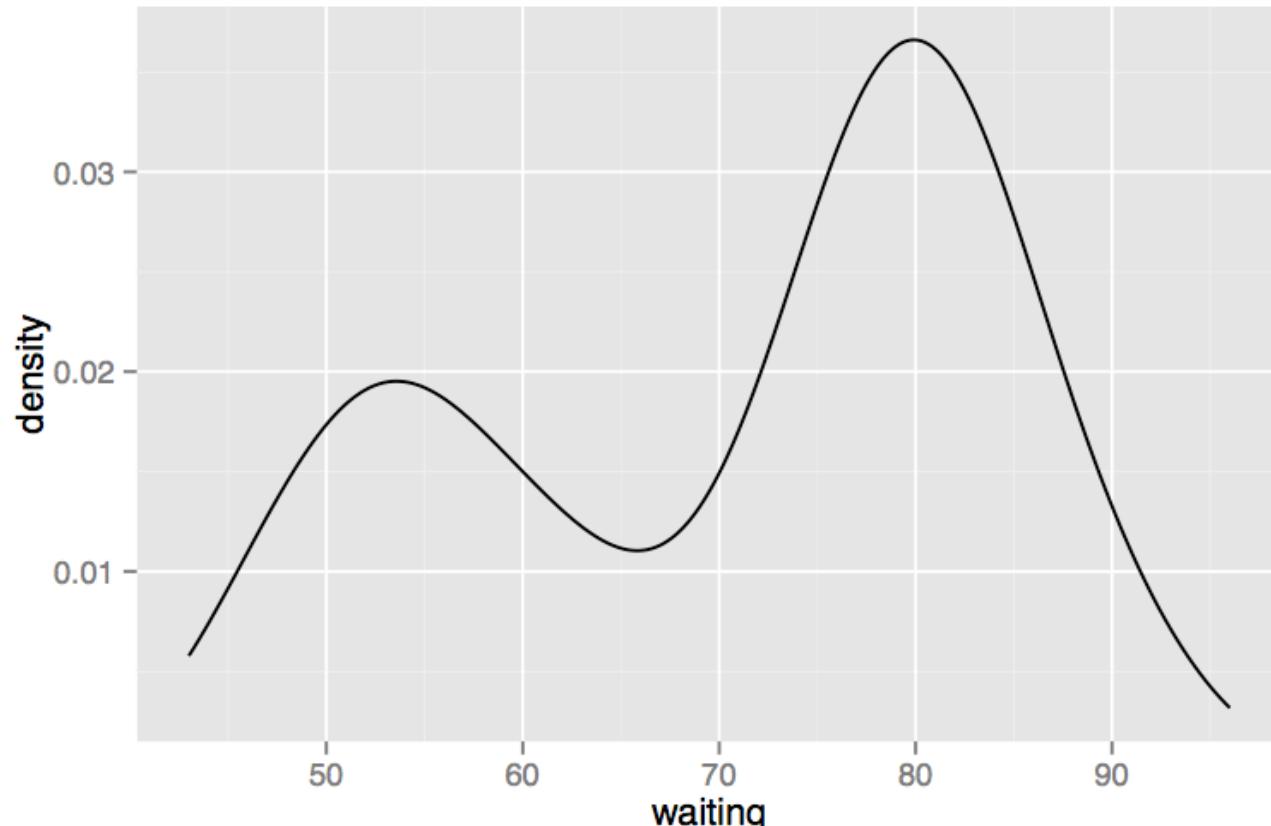
Density Plots

```
ggplot(faithful, aes(waiting)) +  
  geom_density(fill = "blue", alpha = 0.1)
```



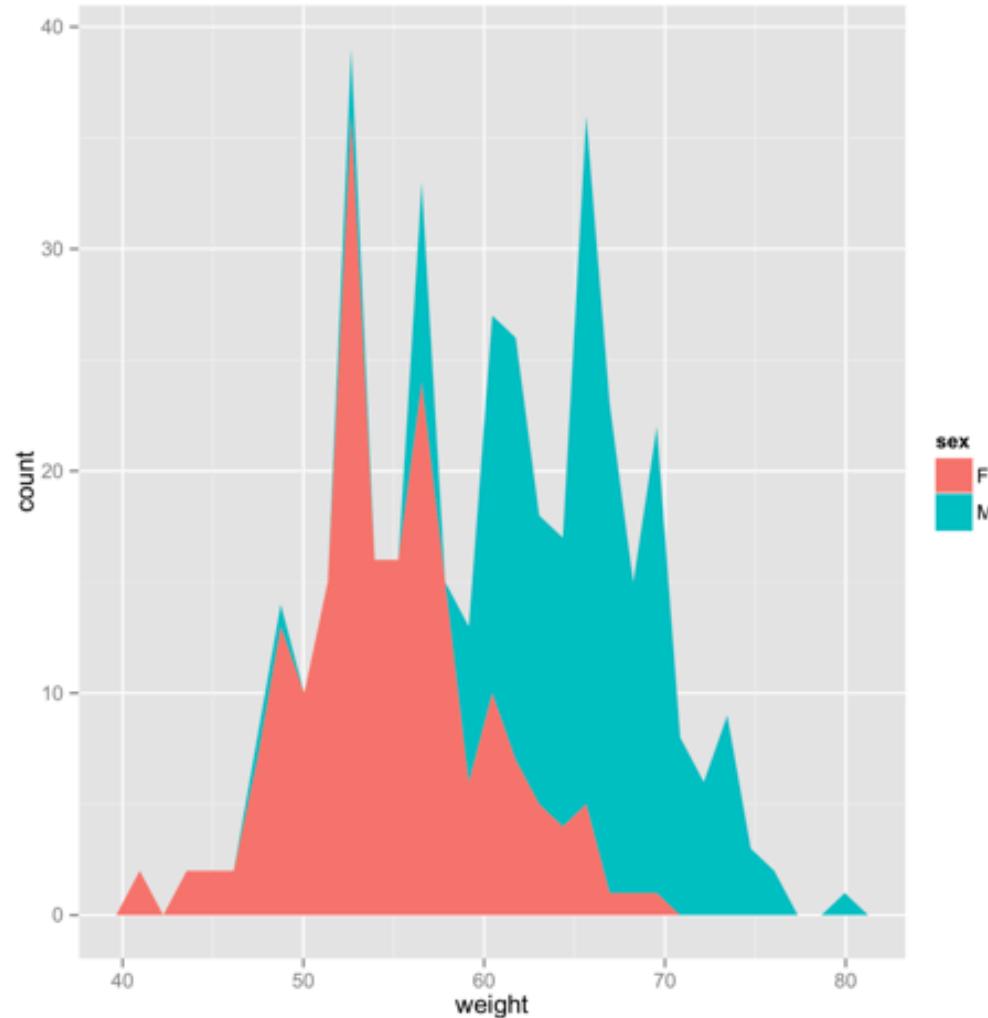
Density Plots

```
ggplot(faithful, aes(waiting)) +  
  geom_line(stat = "density")
```



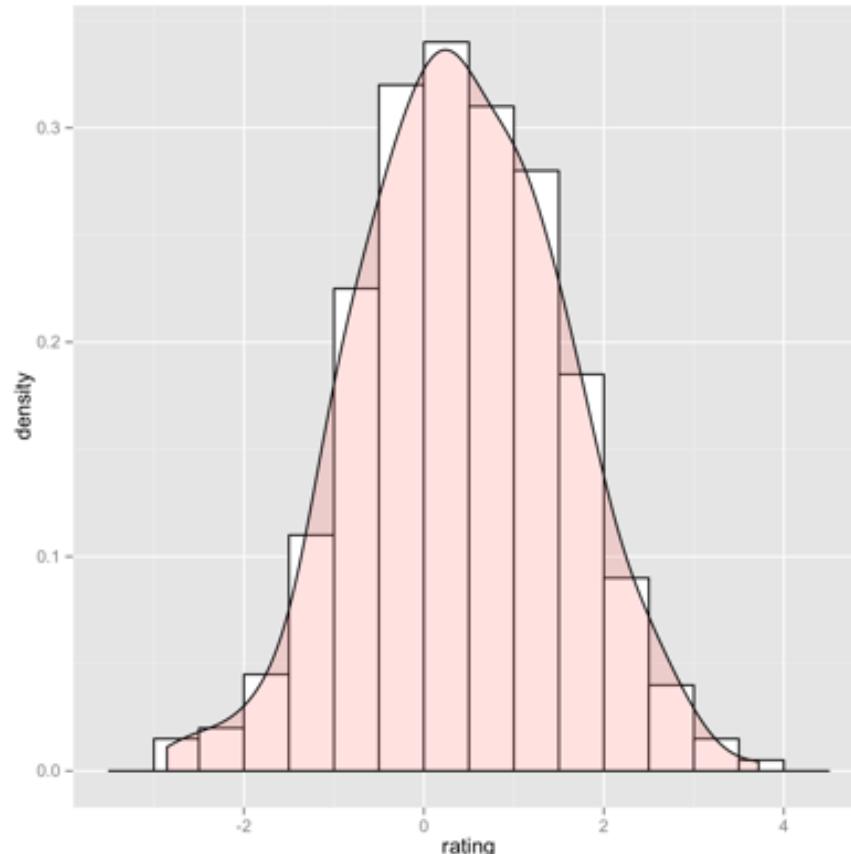
Area Graphs

```
ggplot(df, aes(x=weight, fill=sex)) + geom_area(stat="bin")
```



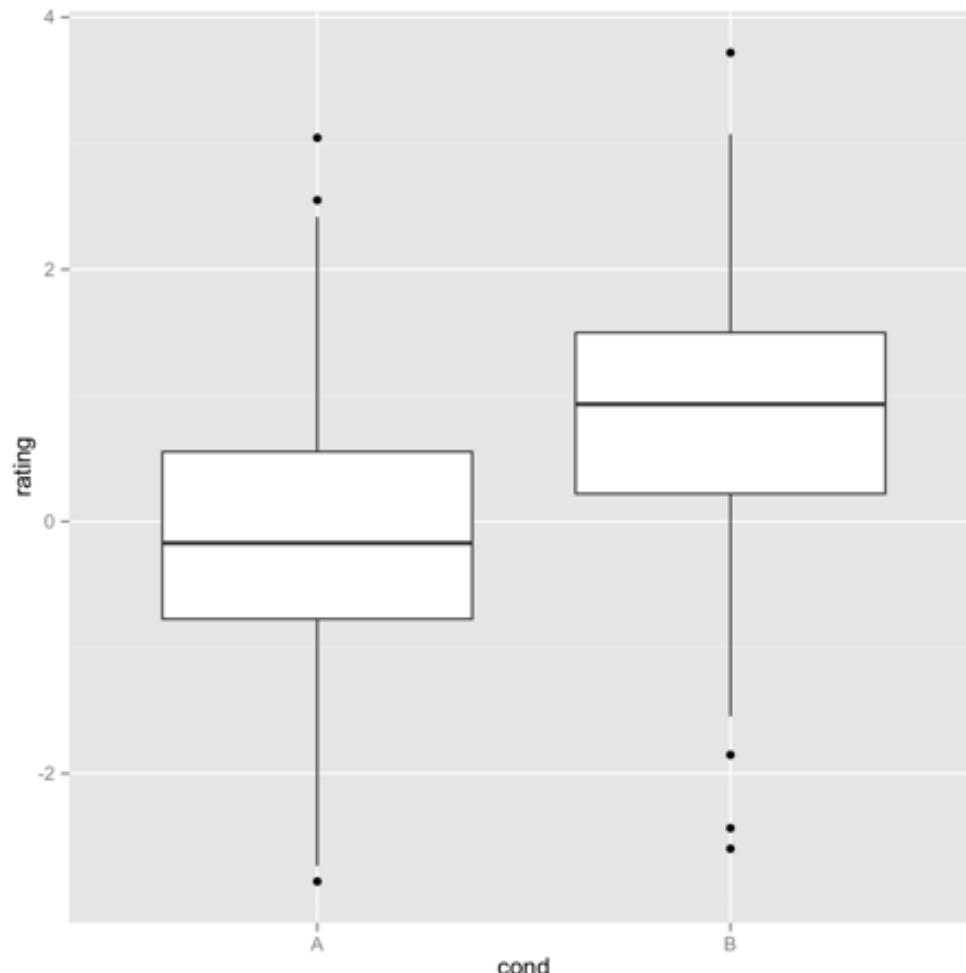
Histogram with Density Curve

```
ggplot(dat, aes(x=rating)) +  
  geom_histogram(aes(y=..density..), binwidth=.5, colour="black", fill="white") +  
  geom_density(alpha=.2, fill="#FF6666")
```



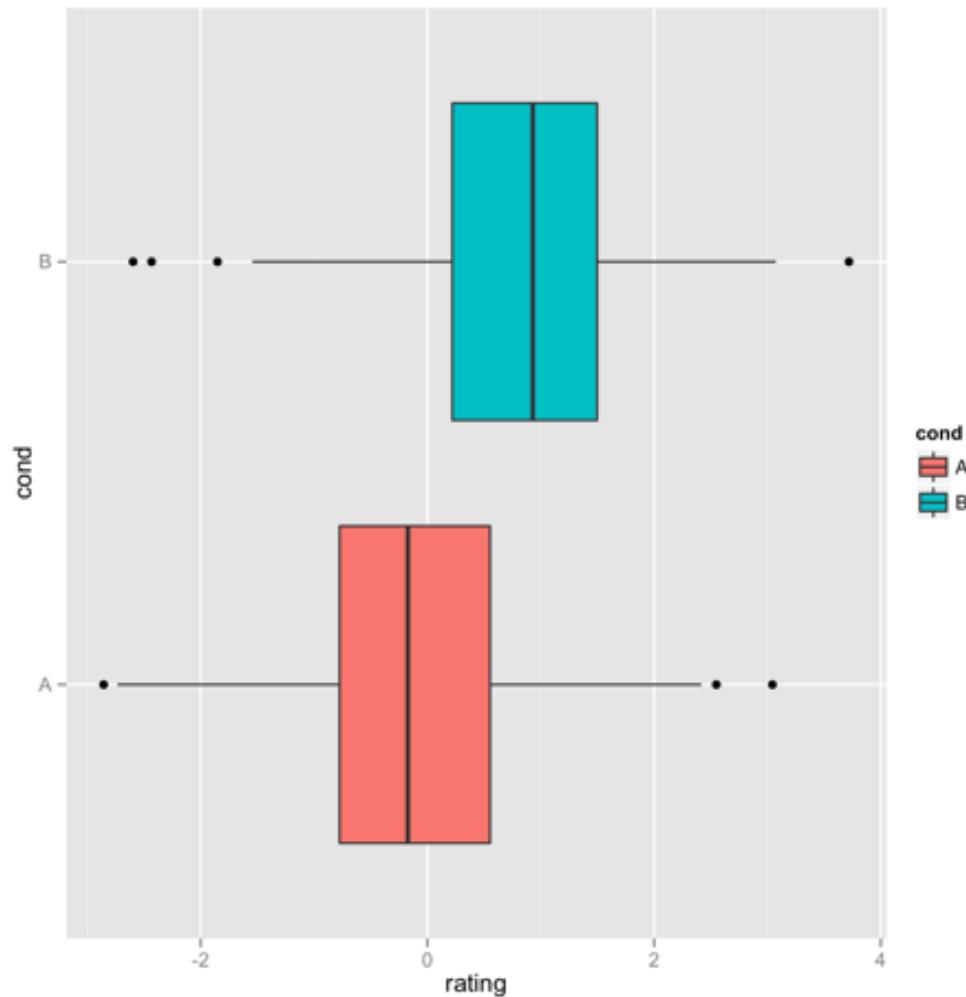
Box Plots

```
ggplot(dat, aes(x=cond, y=rating)) + geom_boxplot()
```



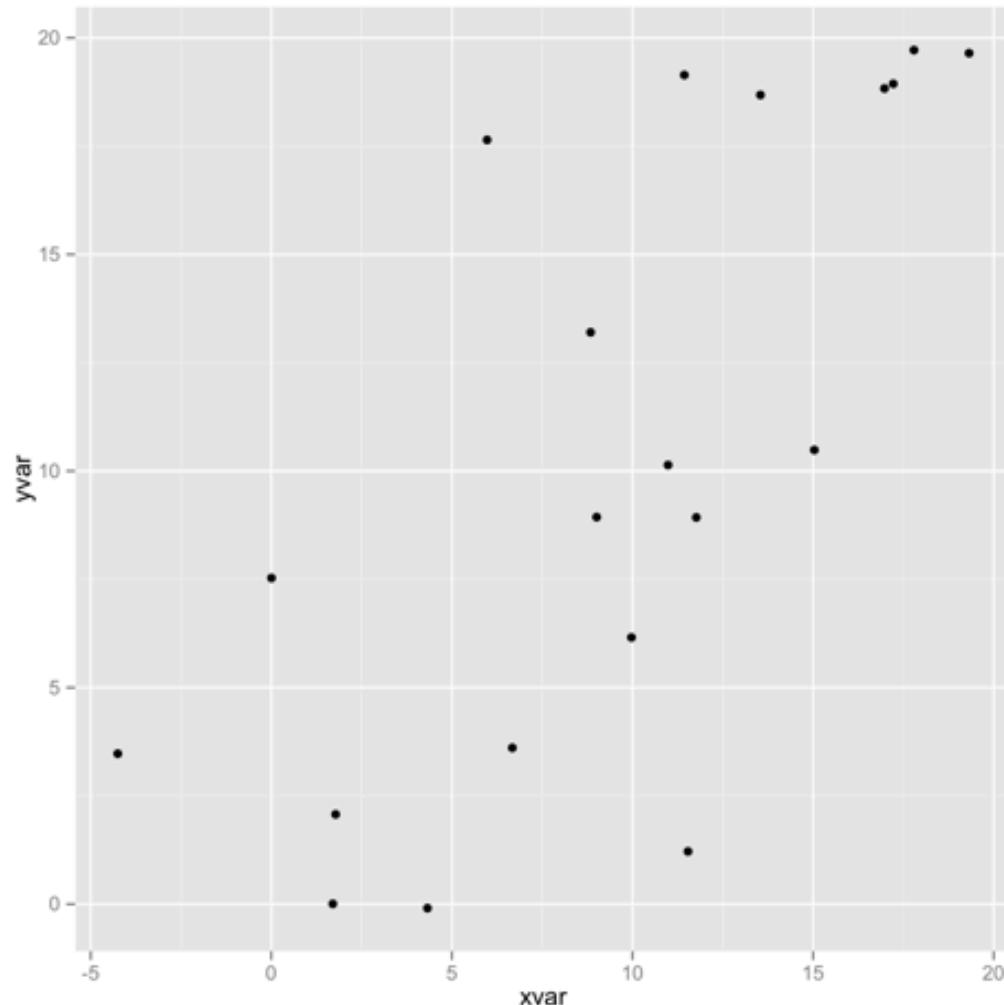
Box Plots

```
ggplot(dat, aes(x=cond, y=rating, fill=cond)) + geom_boxplot() + coord_flip()
```



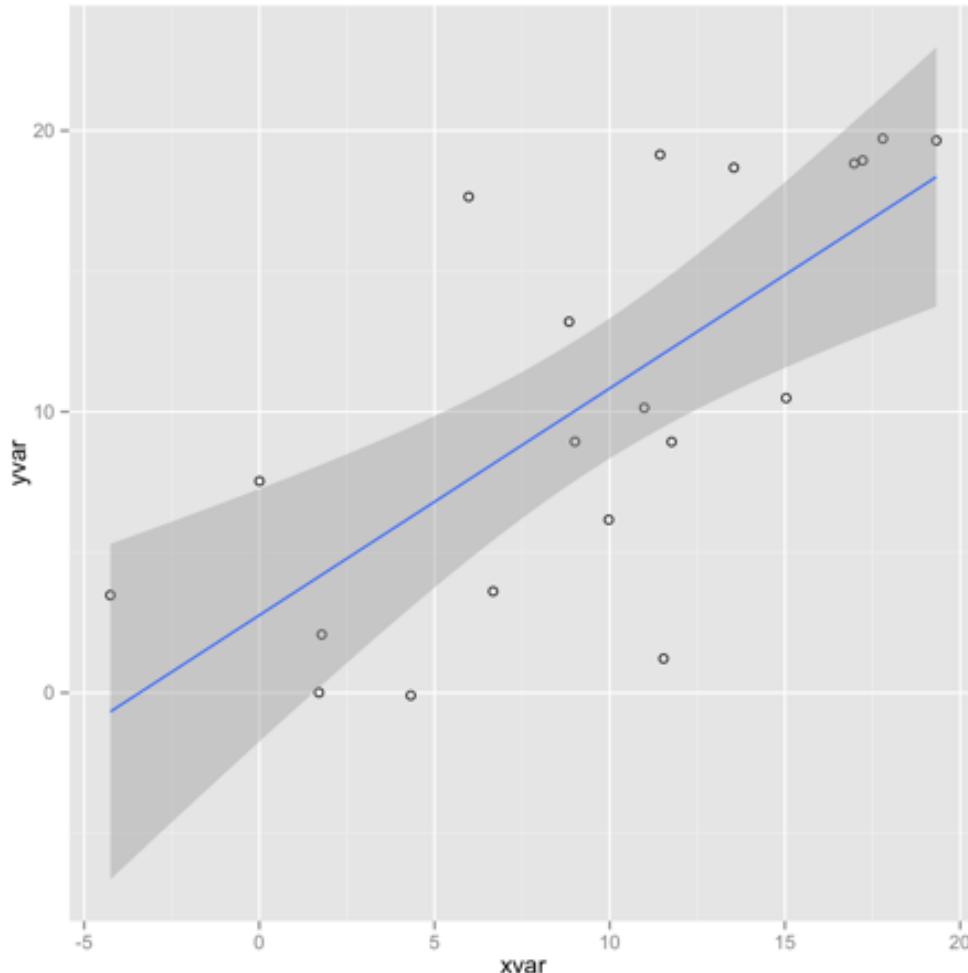
Scatter Plots

```
ggplot(dat, aes(x=xvar, y=yvar)) + geom_point()
```



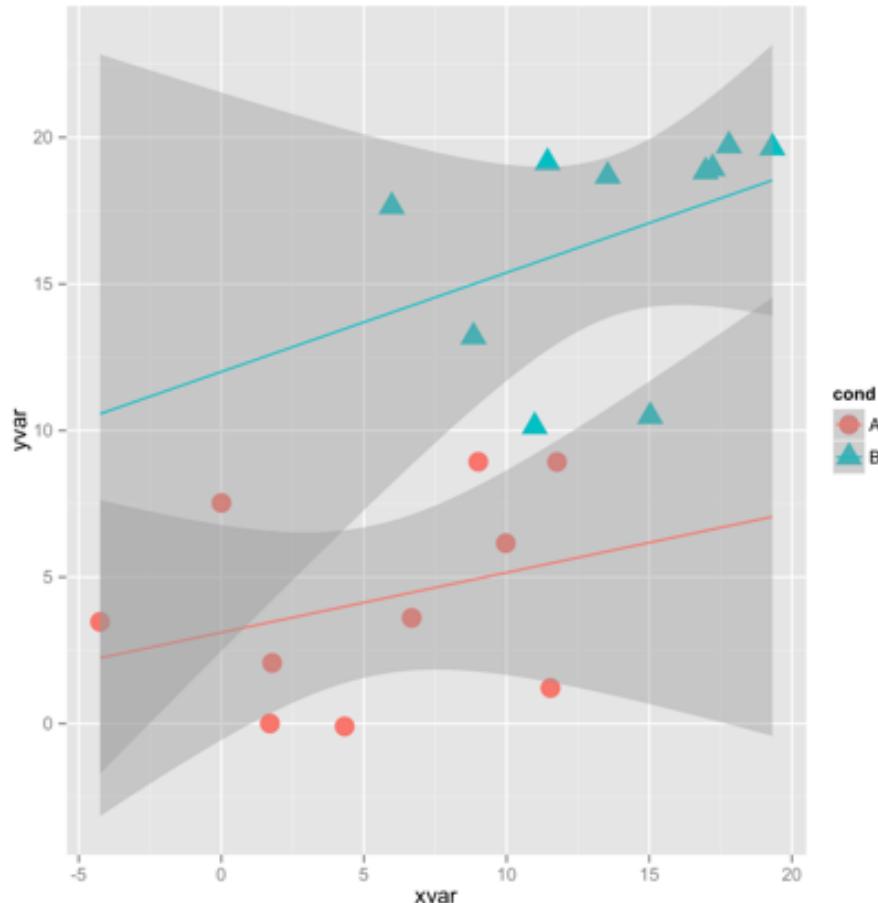
Scatter Plots

```
ggplot(dat, aes(x=xvar, y=yvar)) + geom_point(shape=1) + geom_smooth(method=lm)
```



Scatter Plots

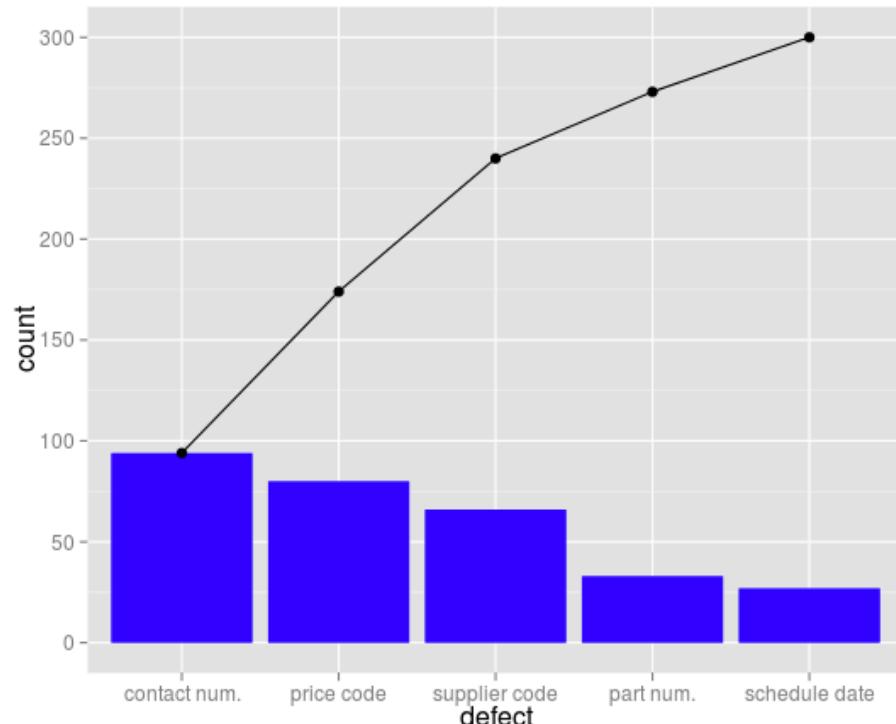
```
ggplot(dat, aes(x=xvar, y=yvar, colour=cond, shape=cond)) + geom_point(size=5) +  
geom_smooth(method=lm, fullrange=TRUE)
```



Pareto Chart

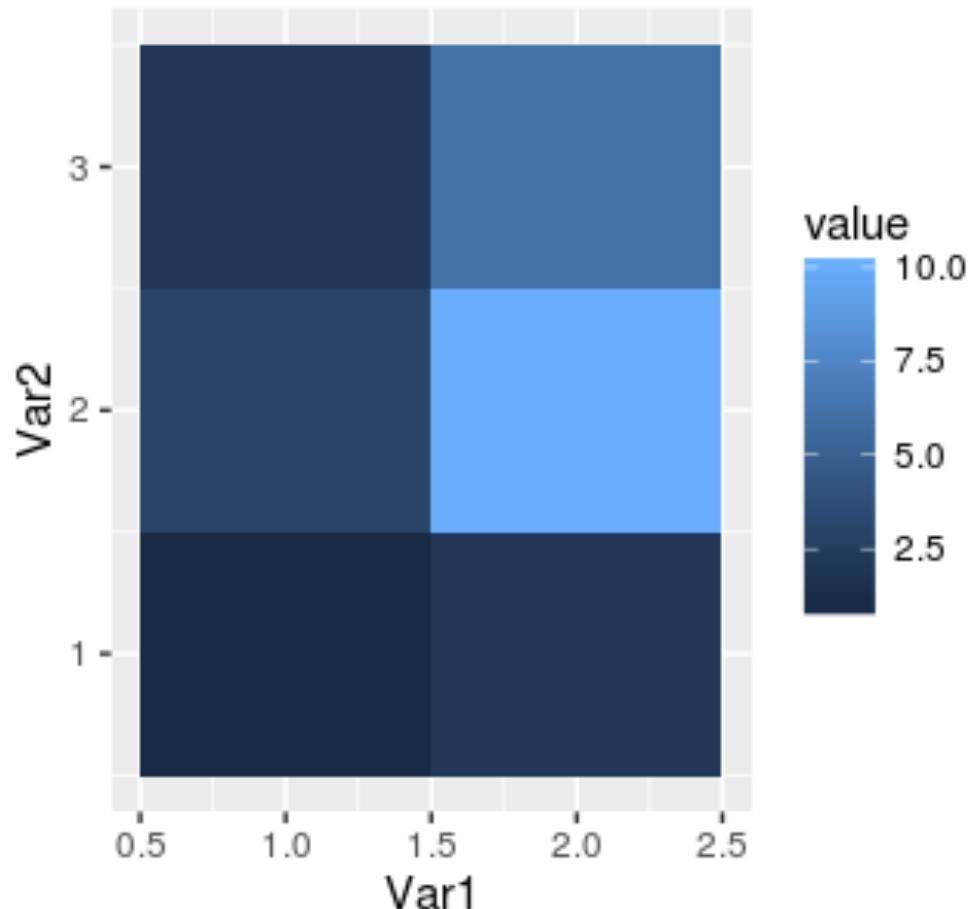
```
dat <- dat[order(dat$count, decreasing=TRUE), ]  
dat$defect <- factor(dat$defect, levels=dat$defe  
Dat$cum <- cumsum(dat$count)
```

```
ggplot(dat, aes(x=defect)) +  
  geom_bar(aes(y=count), fill="blue",  
stat="identity") +  
  geom_point(aes(y=cum)) +  
  geom_path(aes(y=cum, group=1))
```



Heat Map

```
ggplot(dat, aes(x=xvar, y=yvar, fill=value)) + geom_tile()
```



Complex Plots

ggplot2 Extensions

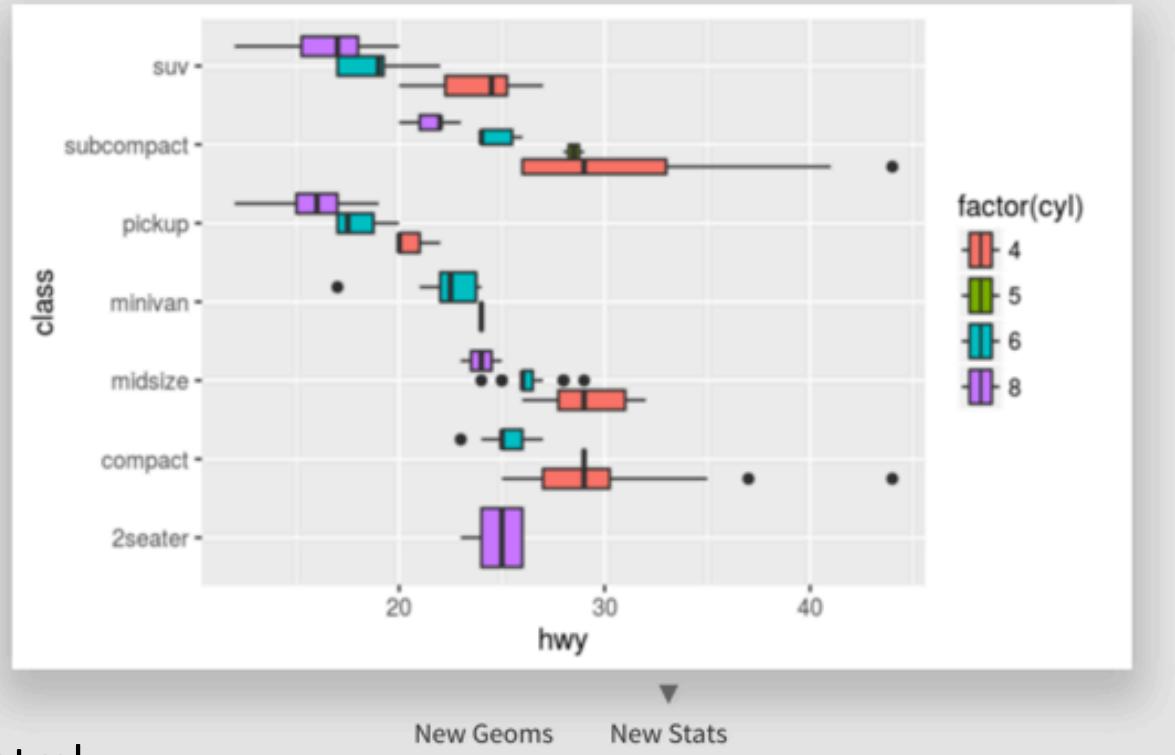
ggplot2 extensions

[Home](#)[Gallery](#)[Extensions](#)[GitHub](#)

A List of ggplot2 extensions

This site tracks and lists **ggplot2** extensions developed by R users in the community.

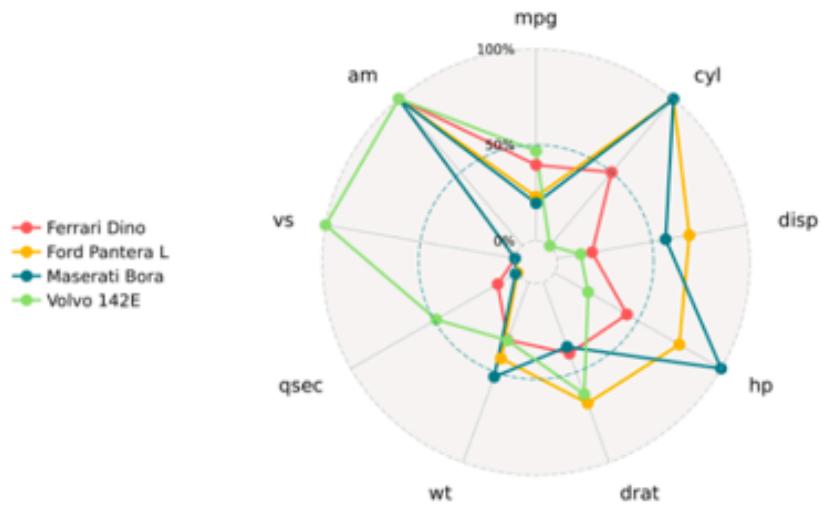
The aim is to make it easy for R users to find developed extensions.



<https://www.ggplot2-exts.org/ggiraph.html>

<http://www.ggplot2-exts.org/gallery/>

ggplot2 Extensions: Radar Graphs



ggradar

ggradar allows you to build radar charts with ggplot2.

■ **author:** ricardo-bion

■ **tags:** visualization, general

■ **js libraries:**

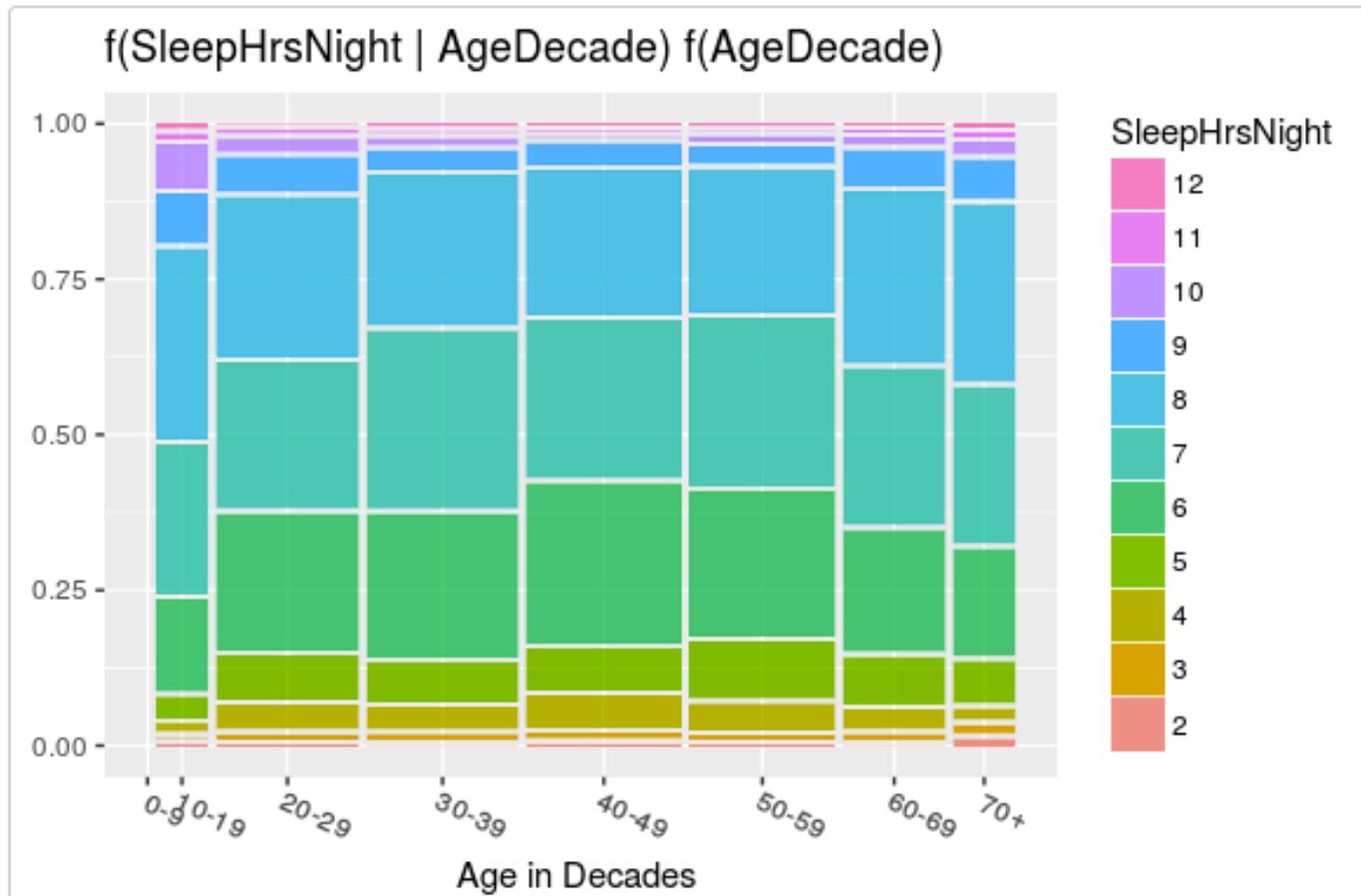
mtcars %>%

```
add_rownames( var = "group" ) %>%
  mutate_each(funs(rescale), -group) %>%
  tail(4) %>% select(1:10) -> mtcars_radar
```

ggradar(mtcars_radar)

ggplot2 Extensions: Mosaic Plots

```
ggplot(data = NHANES) +  
  geom_mosaic(aes(weight = Weight, x = product(SleepHrsNight, AgeDecade), fill=factor(SleepHrsNight)),  
  na.rm=TRUE) + theme(axis.text.x=element_text(angle=-25, hjust=.1)) + labs(x="Age in Decades ",  
  title='f(SleepHrsNight | AgeDecade) f(AgeDecade)') + guides(fill=guide_legend(title = "SleepHrsNight",  
  reverse = TRUE))
```



ggplot2 Extensions

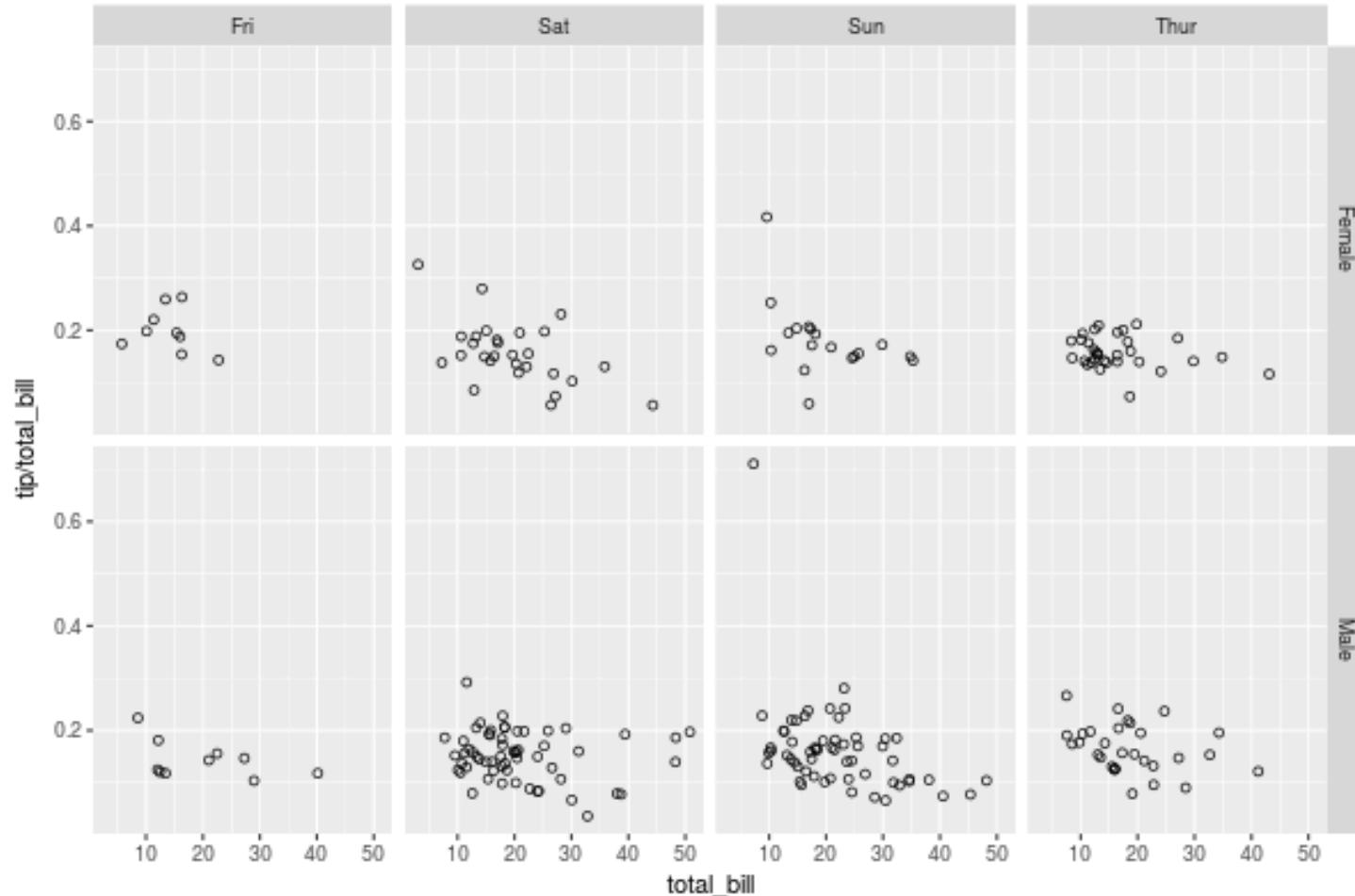
- Many more...

<http://www.ggplot2-exts.org/geomnet.html>



Trellis Display

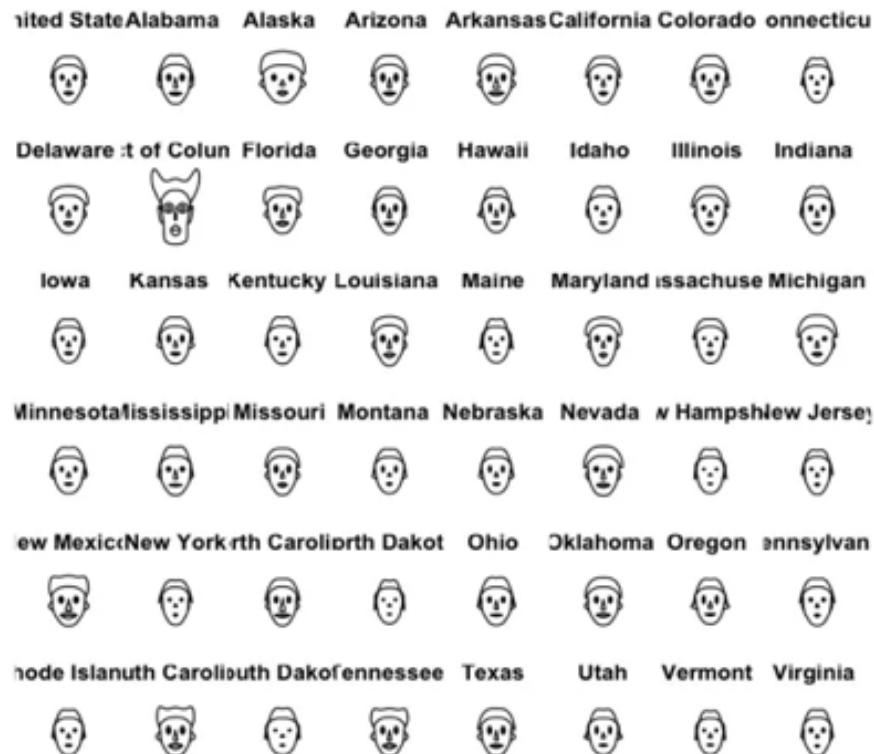
```
ggplot(tips, aes(x=total_bill, y=tip/total_bill)) + geom_point(shape=1) +  
+ facet_grid(sex ~ day)
```



Chernoff Faces

library(aplpack)

```
faces(crime_filled[,2:8], labels=crime_filled$state)
```

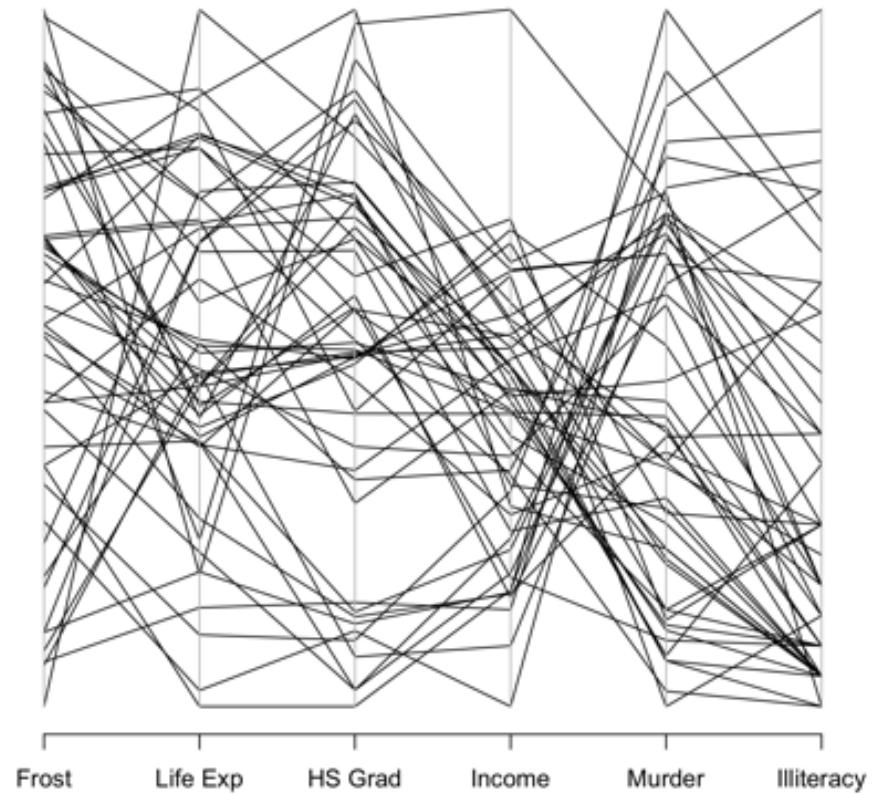


Parallel Coordinates

library(MASS)

parcoord(state.x77[, c(7, 4, 6, 2, 5, 3)])

```
> head(state.x77)
   Population Income Illiteracy Life Exp Murder HS Grad Frost Area
Alabama      3615    3624     2.1    69.05   15.1    41.3    20 50708
Alaska        365    6315     1.5    69.31   11.3    66.7   152 566432
Arizona       2212    4530     1.8    70.55    7.8    58.1    15 113417
Arkansas      2110    3378     1.9    70.66   10.1    39.9    65 51945
California    21198   5114     1.1    71.71   10.3    62.6    20 156361
Colorado      2541    4884     0.7    72.06    6.8    63.9   166 103766
```



<https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/parcoord.html>

<https://www.safaribooksonline.com/blog/2014/03/31/mastering-parallel-coordinate-charts-r/>

Table Lens

- ggplot2 and R may not be the best tool to achieve that.
- Detailed codes can be found in the reference



<http://simondorfman.com/create-table-lens-display-with-r-and-ggplot2>

Take Home Exercises

- You've just scratched the surface with R and ggplot2.
- Read the “R Graphics Cookbook”
- Practice
- Some codes on ggplot2 for iris data:
 - https://www.mailman.columbia.edu/sites/default/files/media/fdawg_ggplot2.html
 - <https://rpubs.com/karagawa/ggplot2>

More Resources

- [http://tutorials.iq.harvard.edu/R/Rgraphics/
Rgraphics.html](http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html)
- [http://r-statistics.co/Complete-Ggplot2-
Tutorial-Part1-With-R-Code.html](http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html)
- [https://www.statmethods.net/advgraphs/
ggplot2.html](https://www.statmethods.net/advgraphs/ggplot2.html)
- [http://r-statistics.co/ggplot2-Tutorial-With-
R.html](http://r-statistics.co/ggplot2-Tutorial-With-R.html)

Next Lecture

- Topic:
 - Advanced R and Visualization Tools
 - Chart Typologies**
Excel, Many Eyes, Google Charts
 - Visual Analysis Grammars**
VizQL, ggplot2
 - Visualization Grammars**
Protopis, D3.js
 - Component Architectures**
Prefuse, Flare, Improvise, VTK
 - Graphics APIs**
Processing, OpenGL, Java2D
- Next Monday (18 Feb)
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus

G53FIV: Fundamentals of Information Visualization

Lecture 7: Visualization with R – Advanced

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Last Lecture

Visualization with R

R is a tool for...

Data Manipulation

- connecting to data sources
- slicing & dicing data

Modeling & Computation

- statistical modeling
- numerical simulation

Data Visualization

- visualizing fit of models
- composing statistical graphics

munge

model

visualize

Building a Plot in ggplot2

data to visualize (a data frame)

map variables to **aes**thetic attributes

geometric objects – what you see (points, bars, etc)

scales map values from data to aesthetic space

faceting subsets the data to show multiple plots

statistical transformations – summarize data

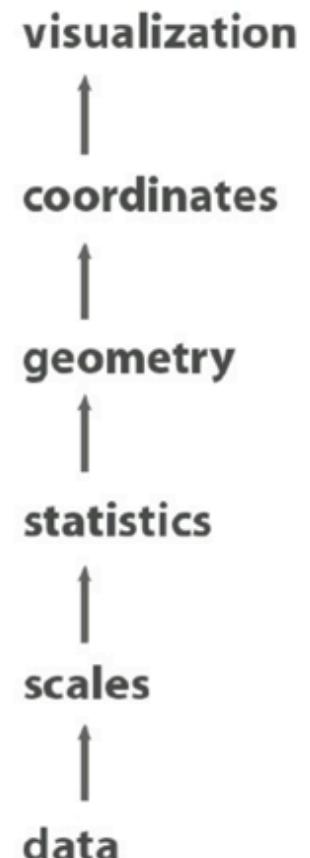
coordinate systems put data on plane of graphic

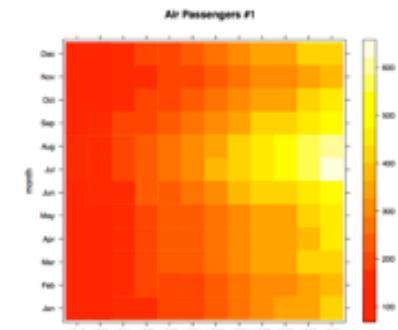
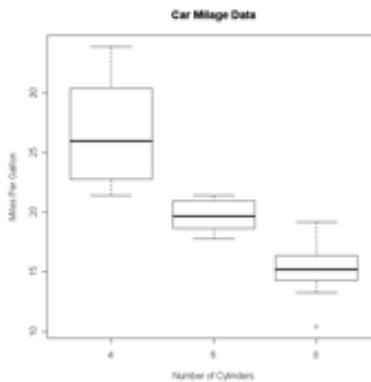
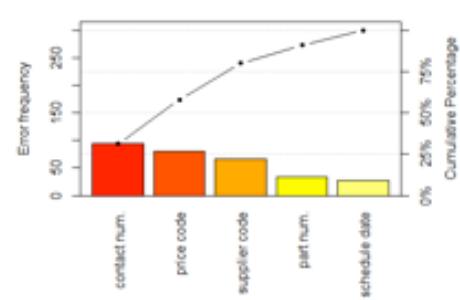
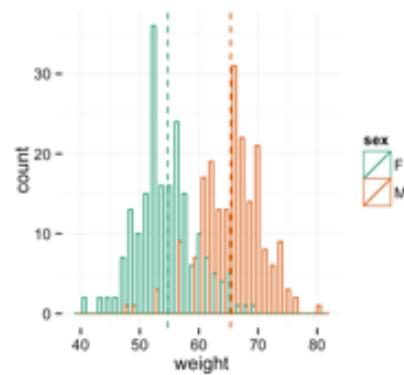
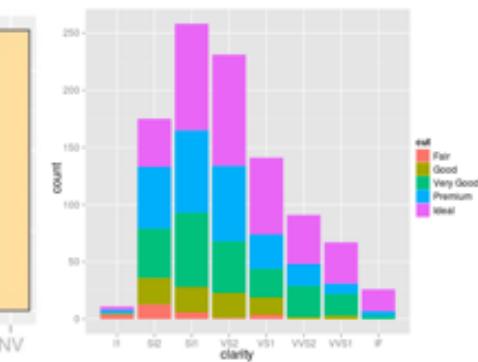
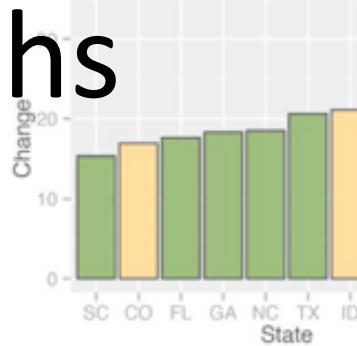
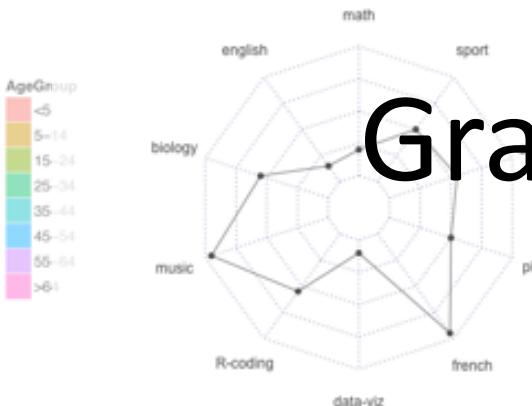
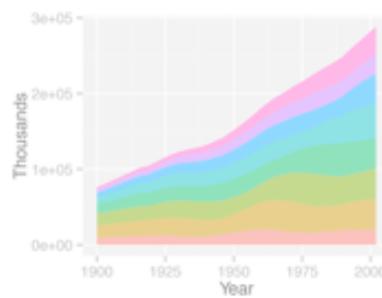
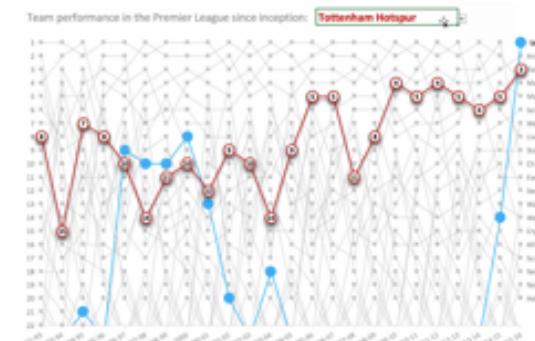
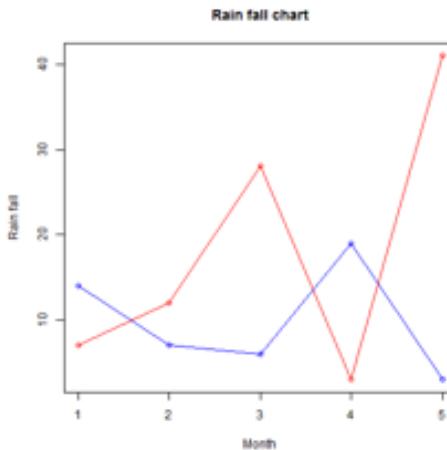
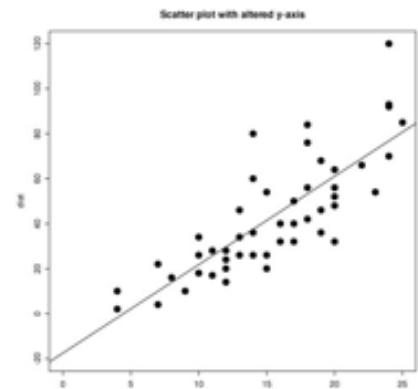


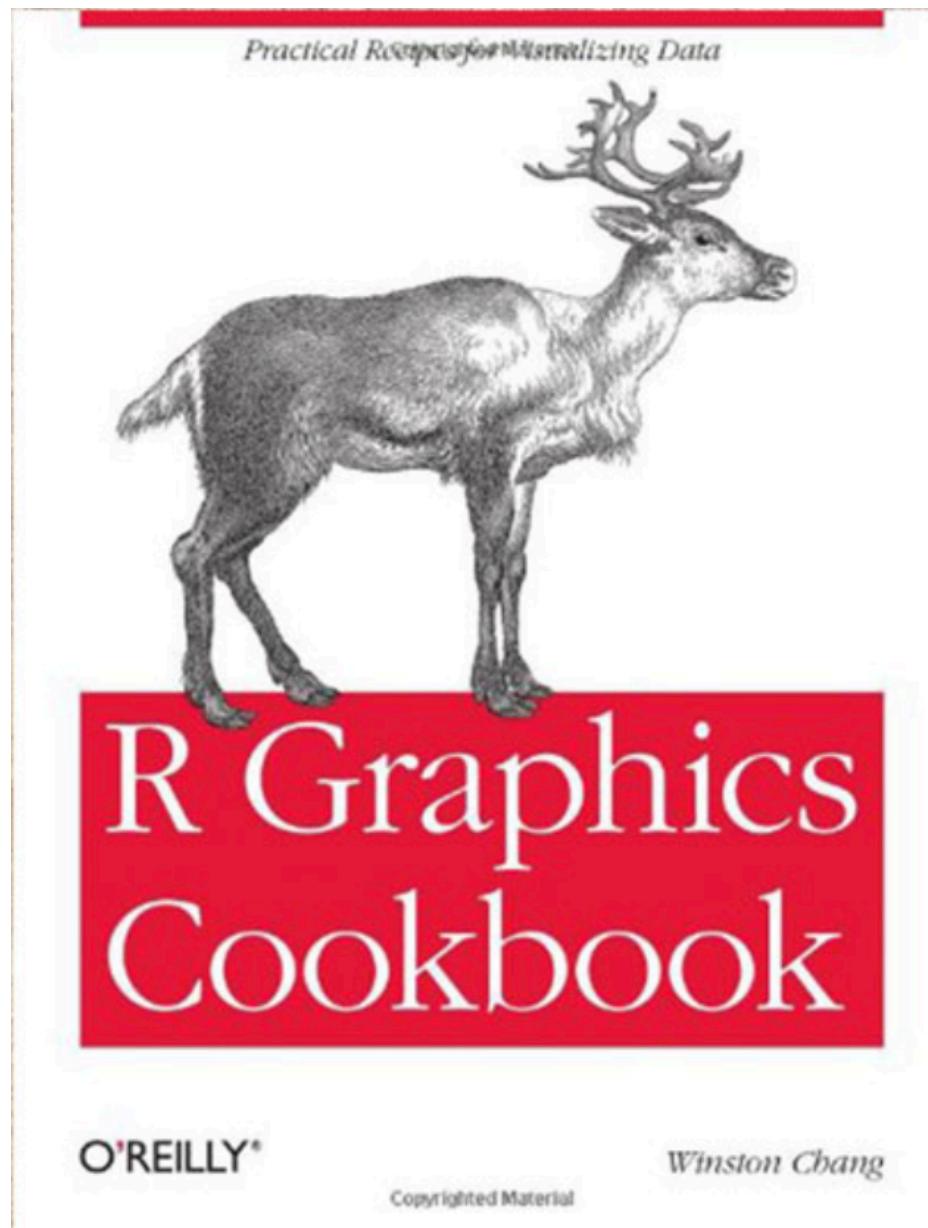
sum(0 1 0 ...)

log(0 1 0 ...)

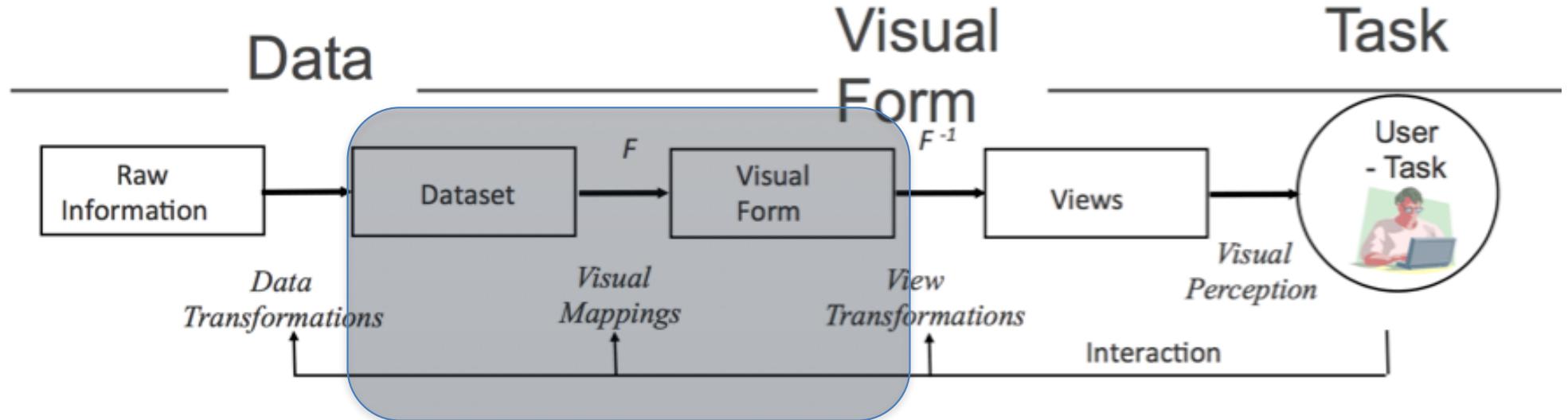
0 1 0 0 1 0 0 1 0
0 1 1 0 1 1 0 1 1



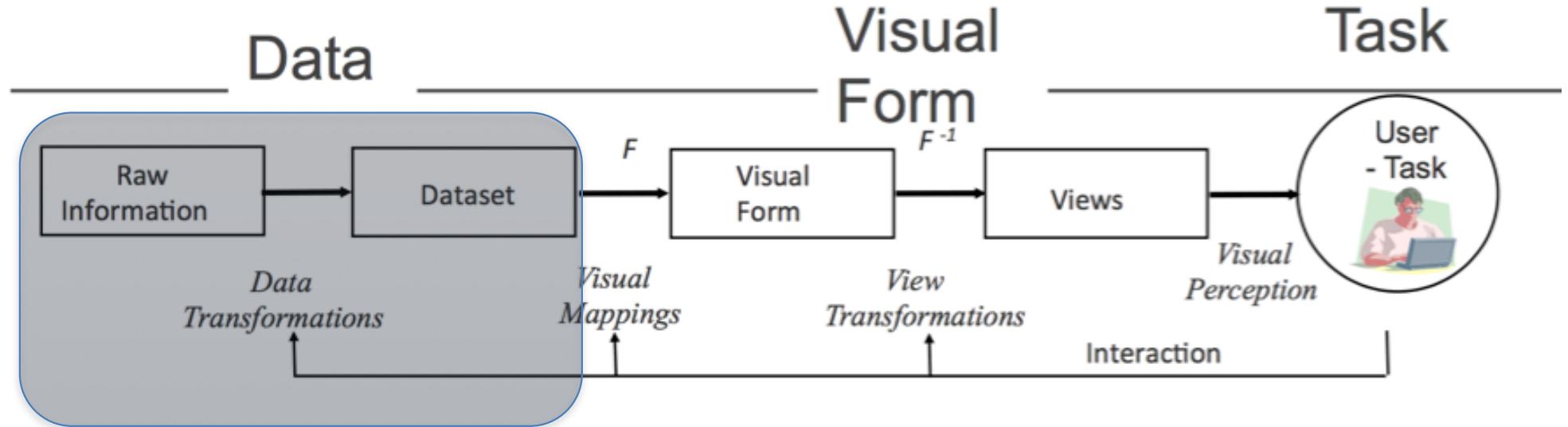




Seven Stages of Visualization



Seven Stages of Visualization



G53FIV Early Module Feedback

- Survey (on Moodle)
 - Anonymous
 - Seeking for constructive feedback
 - Two compulsory multiple questions
 - A few open-ended questions
 - If you have specific feedback or suggestions for improvement.
- A summary of the survey (and action points for improvements) will be presented.

Overview

- Data Manipulations with R
- Brief (Coursework) Case Study with R

Data Manipulations with R

Transform Data: A Swiss-Army Knife

- Indexing
- Three ways to index into a data frame
 - Array of integer indices
 - Array of character names
 - Array of logical Booleans
- Examples:
 - `df[1:3,]`
 - `df[c("New York", "Chicago"),]`
 - `df[c(TRUE, FALSE, TRUE, TRUE),]`

	A	B	C	D
1	year	age	marst	sex
2	1850	0	0	1
3	1850	0	0	2
4	1850	5	0	1
5	1850	5	0	2
6	1850	10	0	1
7	1850	10	0	2
8	1850	15	0	1
9	1850	15	0	2
10	1850	20	0	1
11	1850	20	0	2
12	1850	25	0	1
13	1850	25	0	2
14	1850	30	0	1
15	1850	30	0	2
16	1850	35	0	1
17	1850	35	0	2
18	1850	40	0	1
19	1850	40	0	2
20	1850	45	0	1
21	1850	45	0	2
22	1850	50	0	1
23	1850	50	0	2
24	1850	55	0	1



Transform Data: A Swiss-Army Knife

- **subset** – extract subsets meeting some criteria

```
subset(Insurance, District==1)  
subset(Insurance, Claims < 20)
```

- **transform** – add or alter a column of a data frame

```
transform(Insurance, Propensity=Claims/Holders)
```

- **CUT** – cut a continuous value into groups

```
cut(Insurance$Claims, breaks=c(-1,100,Inf), labels=c('lo','hi'))
```

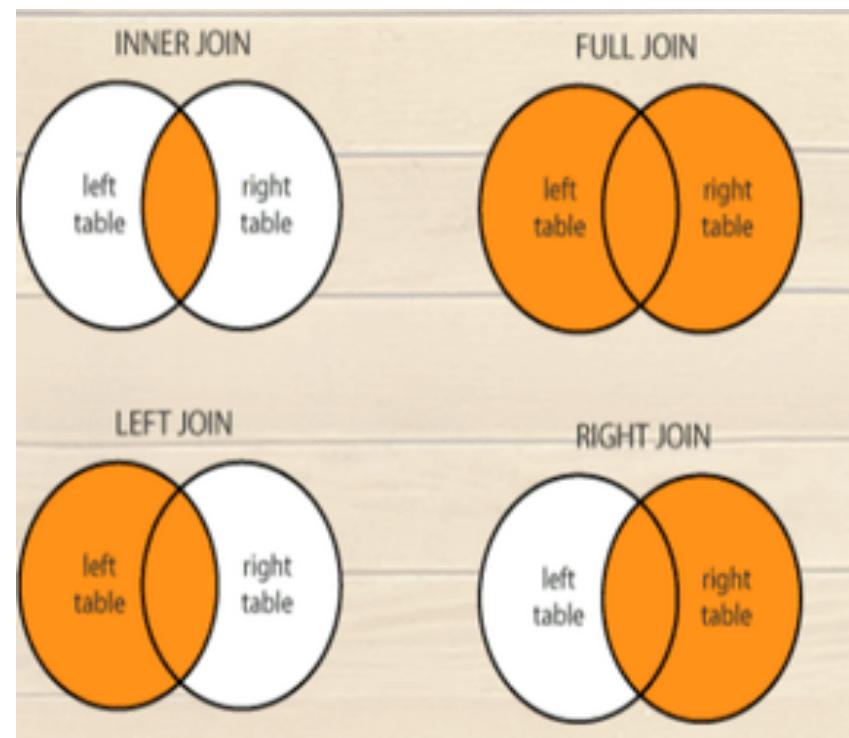
- Put it all together: create a new, transformed data frame

```
transform(subset(Insurance, District==1),  
ClaimLevel=cut(Claims, breaks=c(-1,100,Inf),  
labels=c('lo','hi')))
```



Joining Two Data Frames

- `inner_join(df1, df2, by = "common_column")`
- `?join`
 - `Left_join`, `right_join`
 - `Inner_join`, `outer_join`
- `merge(x=df1, y=df2, by.x="id", by.y="bid")`



More about Data Manipulations

- Packages
 - plyr
 - data.table
 - reshape2
 - doBY
 - sqldf
 - and many more

dplyr: A Grammar of Data Manipulation

- Very intuitive, once you understand the basics
- Very fast
 - Created with execution times in mind
- Easy for those migrating from the SQL world
- When written well, your code reads like a “recipe”
- “Code the way you think”

<https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html>

Pipe Operator

- Library(magrittr)
 - A R package launched on Jan 2014
 - A “magic” operator called the PIPE was introduced
 - %>%
 - i.e. “AND THEN”, “PIPE TO”

```
round(sqrt(1000), 3)

library(magrittr)
1000 %>% sqrt %>% round()
1000 %>% sqrt %>% round(., 3)
```

Take 1000, and then its sqrt
And then round it



dplyr

- dplyr takes the `%>%` operator and uses it to great effect for manipulating data frames
 - Works only with data frames
 - 5 basic “verbs” work for 90% of data manipulations

Verbs	What does it do?
<code>filter()</code>	Select a subset of ROWS by conditions
<code>arrange()</code>	Reorders ROWS in a data frame
<code>select()</code>	Select the COLUMNS of interest
<code>mutate()</code>	Create new columns based on existing columns (mutations!)
<code>summarise()</code>	Aggregate values for each group, reduces to single value

5 Basic Verbs

- FILTER Rows



- SELECT Column Types



- ArRANGE Rows (SORT)

Z
A



- Mutate (into something new)



- Summarize by Groups



Movies dataset

title	year	budget	votes	length	Docume ntary	rating	...
Titanic	1997	200,000,000	1000	195	0	7.8	...
Leon	1994	16,000,000	500	90	0	8.6	...
McQueen	2018	52,000,000	200	91	1	7.9	...

Filter()



- Usage:
filter(data, condition)
 - Returns a subset of rows
 - Multiple conditions can be supplied.
 - They are combined with an AND

```
movies_with_budgets <- filter(movies_df, !is.na(budget))
filter(movies, Documentary==1)
filter(movies, Documentary==1) %>% nrow()
good_comedies <- filter(movies, rating > 9, Comedy==1)
dim(good_comedies) #171 movies

# Let us say we only want highly rated comedies, which a lot
# of people have watched, made after year 2000.
movies %>%
  filter(rating >8, Comedy==1, votes > 100, year > 2000)
```

Select()

- Usage:

```
select(data, columns)
```

```
movies_df <-tbl_df(movies)
select(movies_df, title, year, rating) #Just the columns we want to see
select(movies_df, -c(r1:r10)) #we don't want certain columns

#You can also select a range of columns from start:end
select(movies_df, title:votes) # All the columns from title to votes
select(movies_df, -c(budget, r1:r10, Animation, Documentary, Short, Romance))

select(movies_df, contains("r")) # Any column that contains 'r' in its name
select(movies_df, ends_with("t")) # All vars ending with "t"

select(movies_df, starts_with("r")) # Gets all vars staring with "r"
#The above is not quite what we want. We don't want the Romance column
select(movies_df, matches("r[0-9]")) # Columns that match a regex.
```



Arrange()



Usage: **arrange(data, column_to_sort_by)**

- Returns a reordered set of rows
- Multiple inputs are arranged from left-to-right

```
movies_df <-tbl_df(movies)
arrange(movies_df, rating) #but this is not what we want
arrange(movies_df, desc(rating))
#Show the highest ratings first and the latest year...
#Sort by Decreasing Rating and Year
arrange(movies_df, desc(rating), desc(year))
```

What's the difference between these two?

```
arrange(movies_df, desc(rating), desc(year))
arrange(movies_df, desc(year), desc(rating))
```

Mutate()



- Usage:

```
mutate(data, new_col = func(olcolumns))
```

- Creates new columns, that are functions of existing variables

```
mutate(iris, aspect_ratio = Petal.Width/Petal.Length)

movies_with_budgets <- filter(movies_df, !is.na(budget))
mutate(movies_with_budgets, costPerMinute = budget/length) %>%
  select(title, costPerMinute)
```



Group_by() and Summarize()

```
group_by(data, column_to_group) %>%  
  summarize(function_of_variable)
```

- Group_by creates groups of data
- Summarize aggregates the data for each group

```
by_rating <- group_by(movies_df, rating)
```

```
by_rating %>% summarize(n())
```

```
avg_rating_by_year <-  
  group_by(movies_df, year) %>%  
  summarize(avg_rating = mean(rating))
```

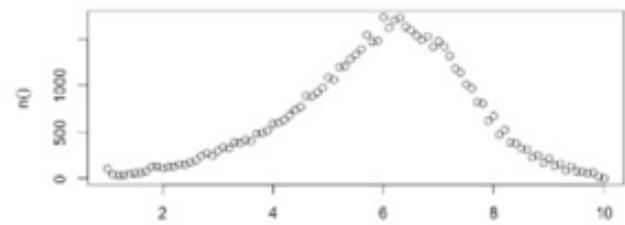
Chain the “Verbs” Together

- Chain them together

```
producers_nightmare <-
  filter(movies_df, !is.na(budget)) %>%
  mutate(costPerMinute = budget/length) %>%
  arrange(desc(costPerMinute)) %>%
  select(title, costPerMinute)
```

- Can also be fed to a “plot” command

```
movies %>%
  group_by(rating) %>%
  summarize(n()) %>%
  plot() # plots the histogram of movies by Each value of rating
```



Practice

- Find all the post-2000 comedy movies with over 1,000,000 budget, rank them by rating in the decreasing order, and output their title and rating

A Case Study: House Price Visualization

All materials are available on Moodle.

Pick a dataset of your interest

- <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads>
- Our Price Paid Data includes information on all property sales in England and Wales that are sold for full market value and are lodged with us for registration.
- Format:
 - tid,price,date,postcode,type,age,tenure,paon,saon,street,locality,city,district,county,ppdcategory,recordstatus

Pose the initial questions (3 to 5) that you would like to answer

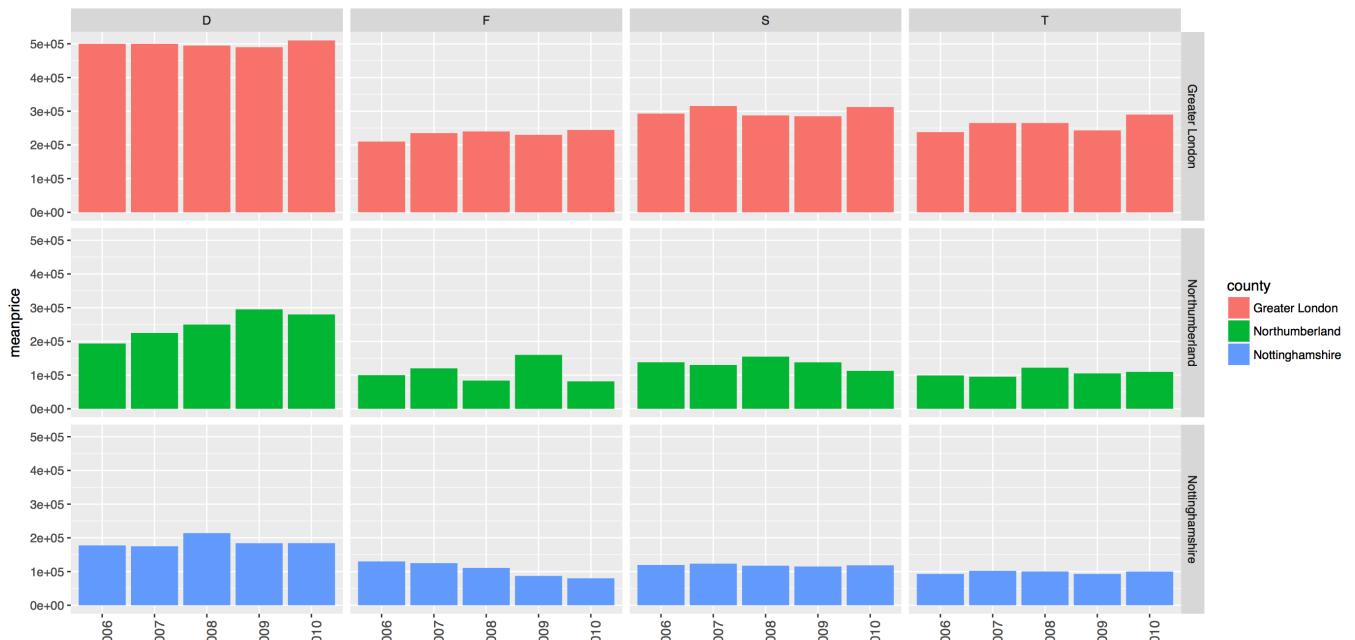
- RQ1: When do property sales generally happen? Which month is the best month to sell?
- RQ2: How do property sales change according to different property types and age?
- RQ3: How do property sale price compare across different counties?

Assess the fitness of the data

- Can we answer all those questions by using only the land registry data?
- RQ3: How do property sale price compare across different counties?
 - Although we have area data, we need the county data, in order to compare across different counties.
 - County-postcode mapping data:
 - [https://github.com/Gibbs/uk-postcodes/blob/master/
postcodes.csv](https://github.com/Gibbs/uk-postcodes/blob/master/postcodes.csv)

Answer the initial questions by visualizing the dataset using R

- Demo
- Load data
- Manipulate data
- Visualization



Further refine/propose questions

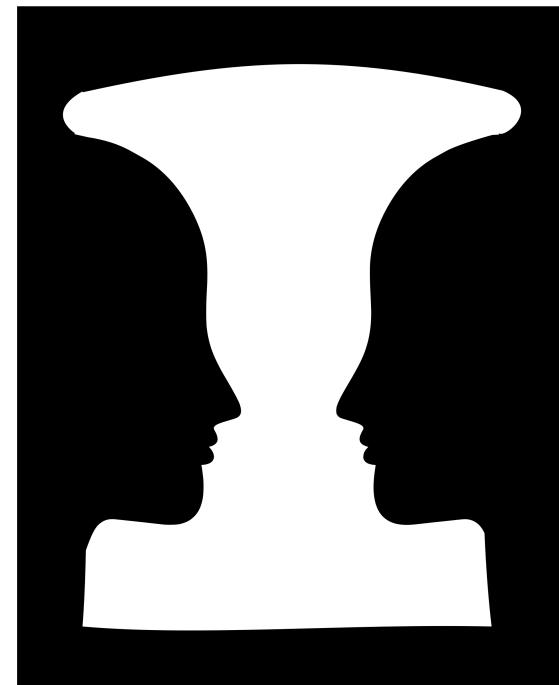
- Example question
 - RQ4: Which types of properties and in which county suffer the most from the economical crisis?
- How to manipulate/process data?
 - E.g. need to calculate year (month) over year (month) change
- What visualization techniques to use? What if there are too many variables?

Course Work Deadline

- April 8th 2018
- Written report
 - 3000 words, max 10 pages
 - R codes as well

Next Lecture

- Topic:
 - Visualization Tools and Visual Perception



G53FIV: Fundamentals of Information Visualization

Lecture 8: Visualization Tools and Visual Perception

Ke Zhou

School of Computer Science

Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Overview

- Visualization Tools
- Visual Perception

Visualization Tools

Visualization Tools

Chart Typologies

Excel, Many Eyes, Google Charts

Visual Analysis Grammars

VizQL, ggplot2

Visualization Grammars

Protopis, D3.js

Component Architectures

Prefuse, Flare, Improvise, VTK

Graphics APIs

Processing, OpenGL, Java2D

Visualization Tools

Chart Typologies

Excel, Many Eyes, Google Charts

Charting
Tools

Visual Analysis Grammars

VizQL, ggplot2

Declarative
Languages

Visualization Grammars

Protopis, D3.js

Programming
Toolkits

Component Architectures

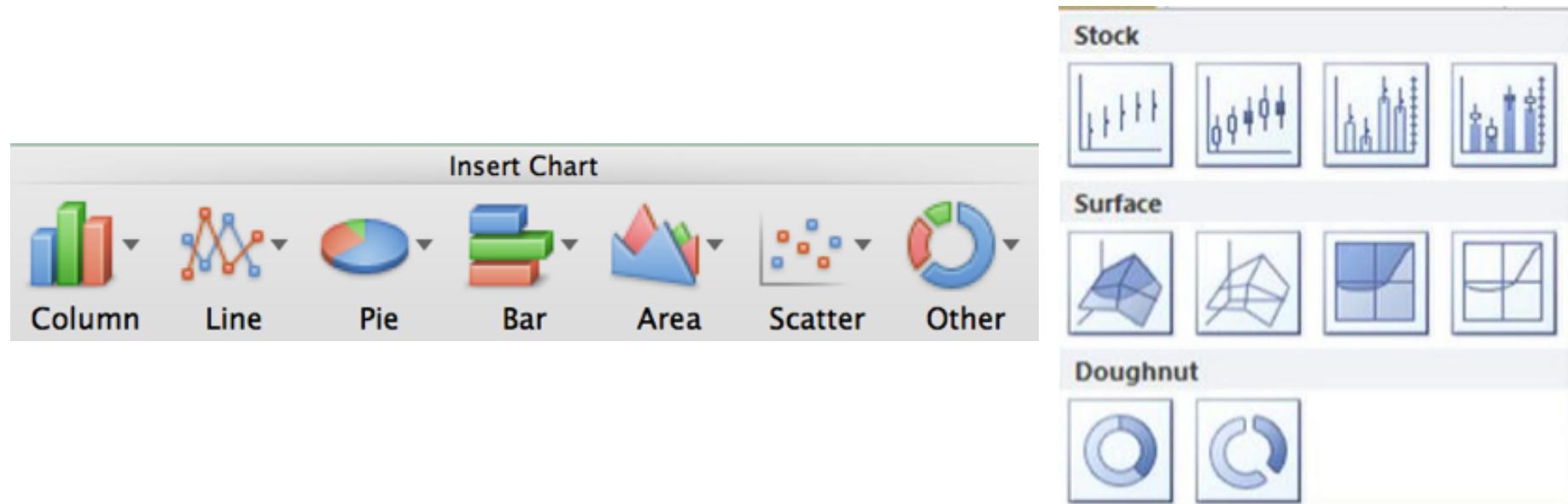
Prefuse, Flare, Improvise, VTK

Graphics APIs

Processing, OpenGL, Java2D

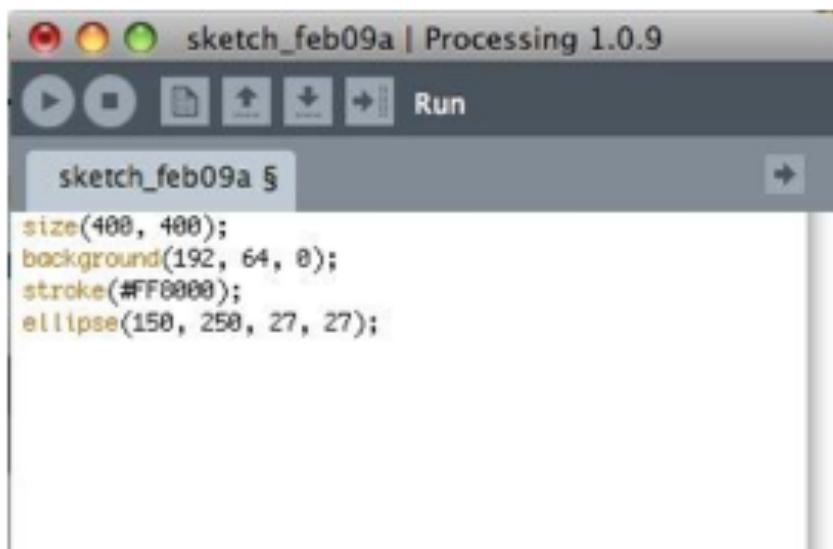
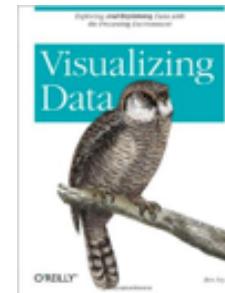
Chart Typology (Charting Tools)

- Pick from a stock of templates
- Easy-to-use but limited expressiveness
- Prohibits novel designs, new data types

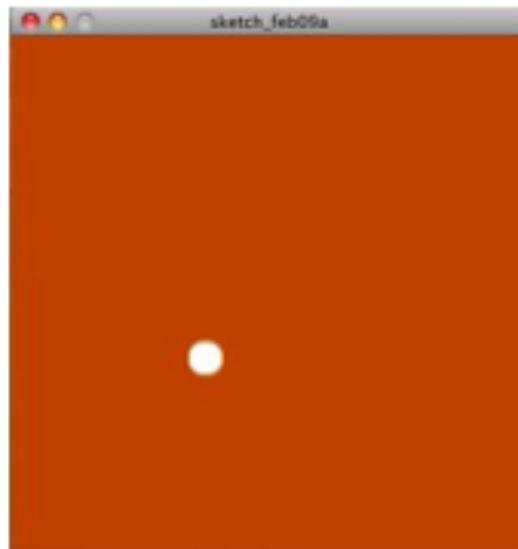


Graphics APIs (Programming Toolkits)

- Processing.org
 - Java based
 - not specifically designed for InfoVis
 - Well documented, lots of tutorials (even books)



```
sketch_feb09a | Processing 1.0.9
Run
sketch_feb09a 5
size(400, 400);
background(192, 64, 0);
stroke(#FF0000);
ellipse(150, 250, 27, 27);
```



Graphics APIs can be very powerful

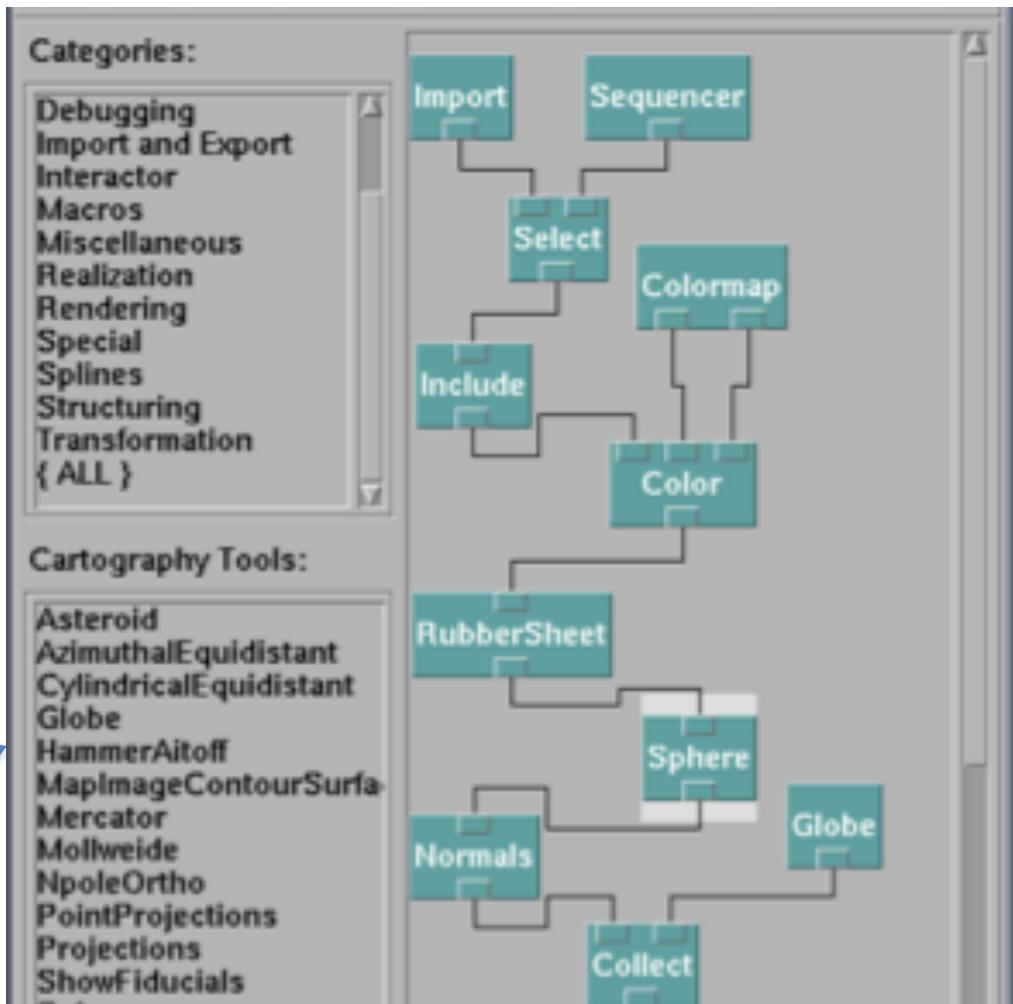
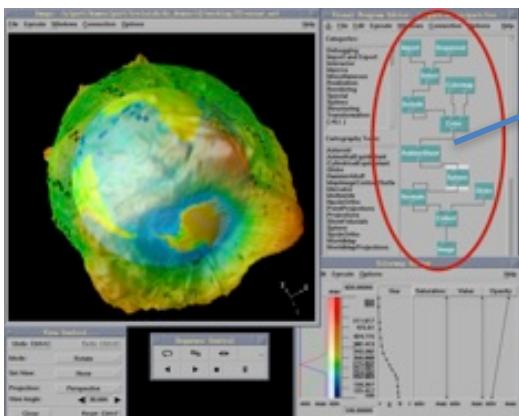


US Air Traffic Visualization

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

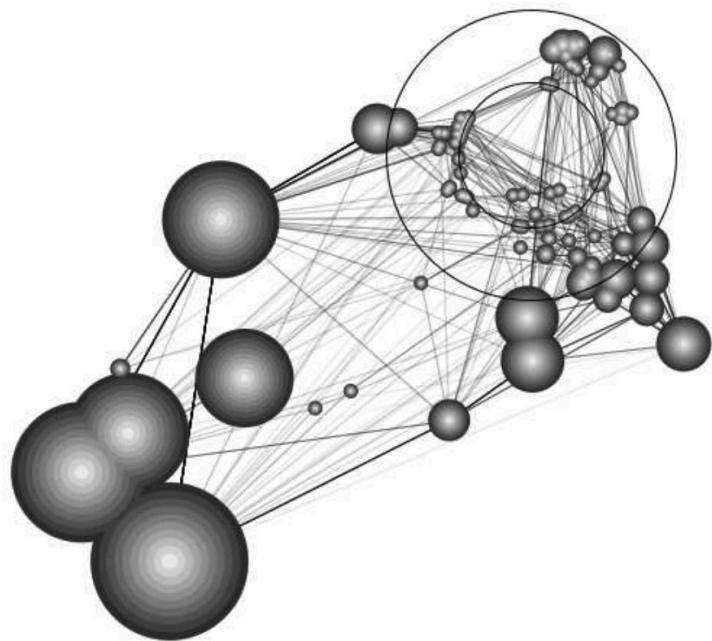
Component Architectures (Programming Toolkits)

- Permits more combinatorial possibilities
- Novel views require new operators, which requires software engineering.

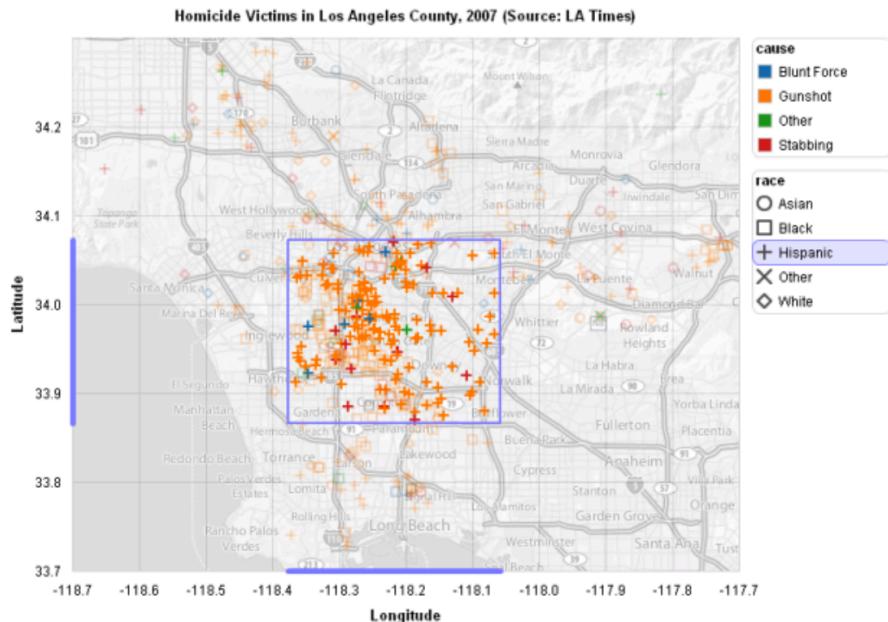


Prefuse & Flare

- Operator-based toolkits for visualization design
Vis = (Input Data -> Visual Objects) + Operators

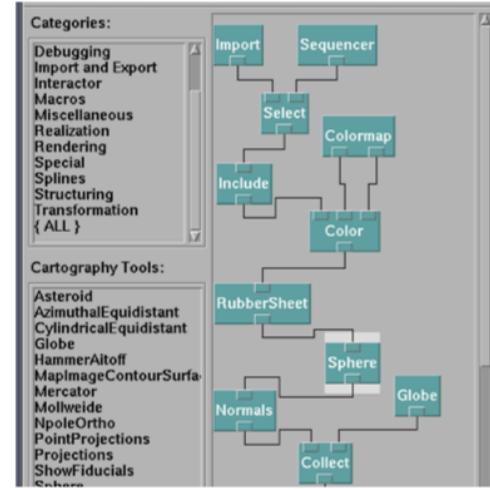


Prefuse (<http://prefuse.org>)



Flare (<http://flare.prefuse.org>)

Comparison



- **Chart Typology**

- Pick from a stock of templates
- Easy-to-use but limited expressiveness
- Prohibits novel designs, new data types

- **Component Architecture**

- Permits more combinatorial possibilities
- Novel views require new operators, which requires software engineering.

The Grammar of Graphics (Declarative Languages)

- Programming by describing what, not how
- Separate specification (what you want) from execution (how it should be computed)
- In contrast to imperative programming, where you must give explicit steps.

```
d3.selectAll("rect")
  .data(my_data)
  .enter().append("rect")
  .attr("x", function(d) { return xscale(d.foo); })
  .attr("y", function(d) { return yscale(d.bar); })
```

Building a Plot in ggplot2

data to visualize (a data frame)

map variables to **aes**thetic attributes

geometric objects – what you see (points, bars, etc)

scales map values from data to aesthetic space

faceting subsets the data to show multiple plots

statistical transformations – summarize data

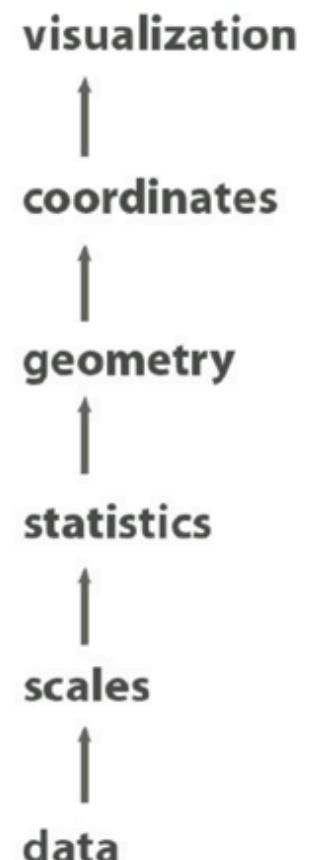
coordinate systems put data on plane of graphic



sum(0 1 0 ...)

log(0 1 0 ...)

0 1 0 0 1 0 0 1 0
0 1 1 0 1 1 0 1 1



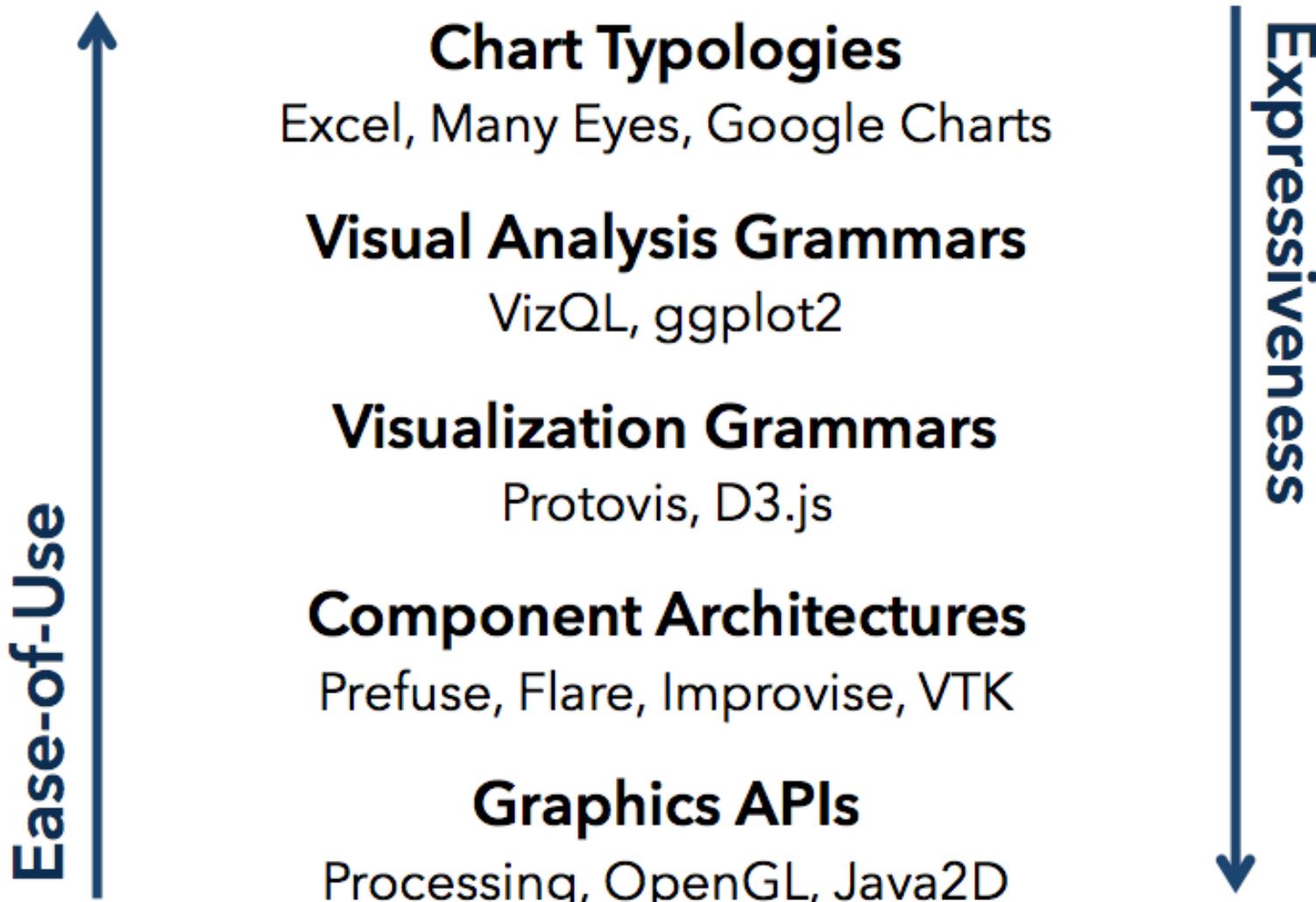
The Advantages of Declarative Languages

- **Faster iteration.** Less code. Larger user base.
- **Better visualization.** Smart defaults.
- **Reuse.** Write-once, then re-apply.
- **Performance.** Optimization, scalability.
- **Portability.** Multiple devices, renderers, inputs.
- **Programmatic generation.** Write programs which output visualizations. Automated search & recommendation.

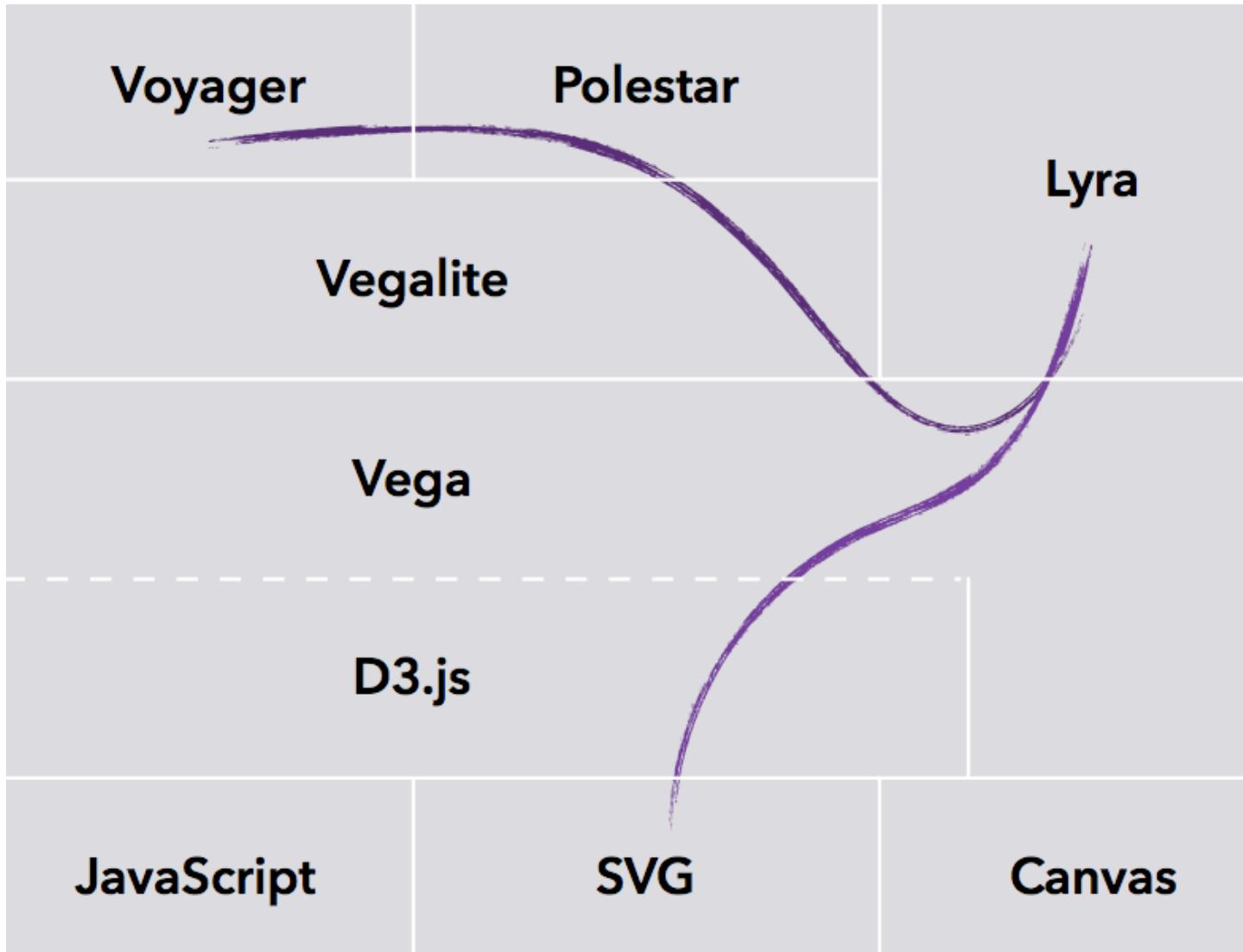
Tools Tradeoffs

- InfoVis-focused
 - Many fundamental techniques built-in
 - Can be faster to get something going
 - Often more difficult to implement something “different”
 - Documentation?
- Generic Graphics
 - More flexible
 - Can customize better
 - Big learning curve
 - Doc is often better
 - Can take a long time to (re)implement basic techniques

Visualization Tools



Many Tools Developed by Prof. Jeffrey Heer, University of Washington



This is just a reference point.
You should try those information
visualization tools out
(optional for those who don't take G53IVP)!

(Optional) Resources

- D3 tutorial: <https://uwdata.github.io/d3-tutorials/>
- Vega tutorial: <https://github.com/vega/vega/wiki/Tutorial>
- Please start working on the course work using R.

Visual Perception

The ability of viewers to interpret visual encodings of information and thereby decode information in graphs.

Related Disciplines

- Psychophysics
 - Applying methods of physics to measuring human perceptual systems
 - How fast must light flicker until we perceive it as constant?
 - What change in brightness can we perceive?
- Cognitive psychology
 - Understanding how people think, here, how it relates to perception

Effectiveness Ranking

Detecting Brightness



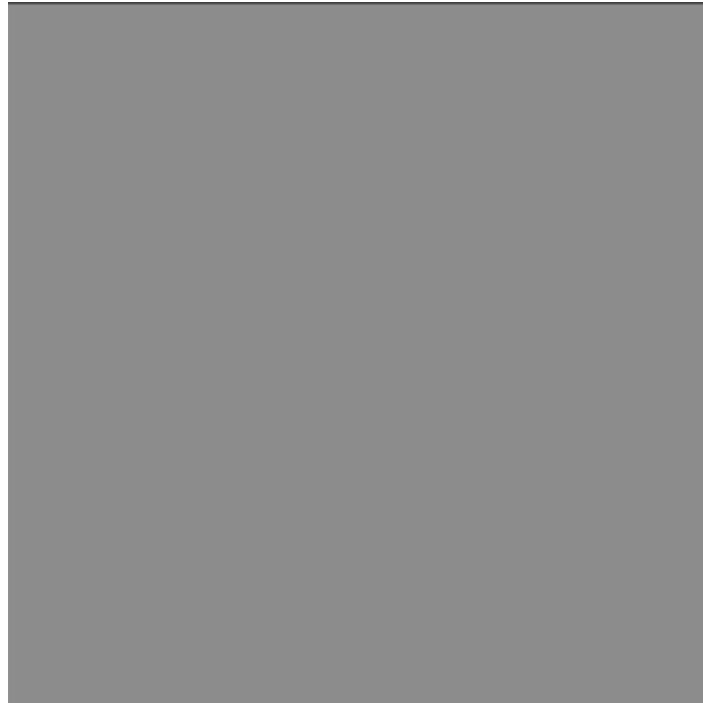
Which one is brighter?

Detecting Brightness

(128,128,128)



(144,144,144)



Which one is brighter?

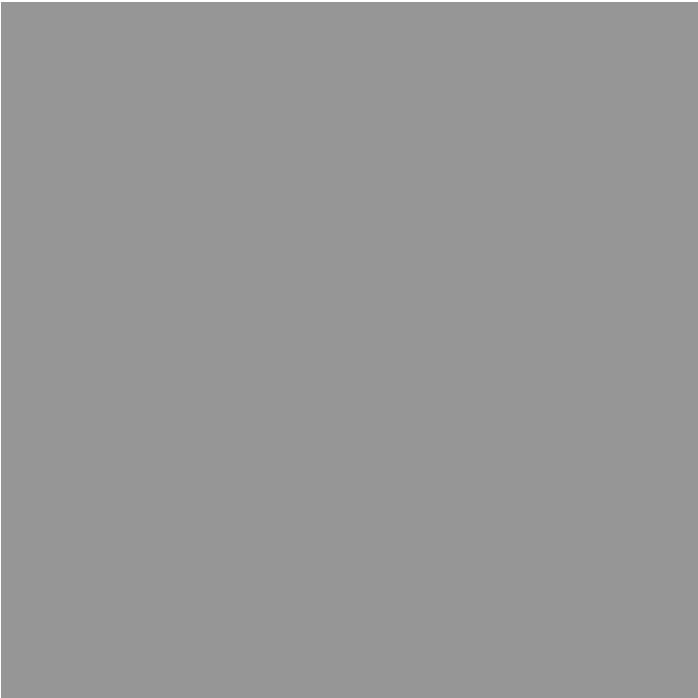
Detecting Brightness



Which one is brighter?

Detecting Brightness

(134,134,134)



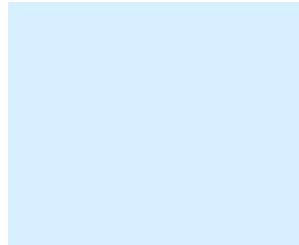
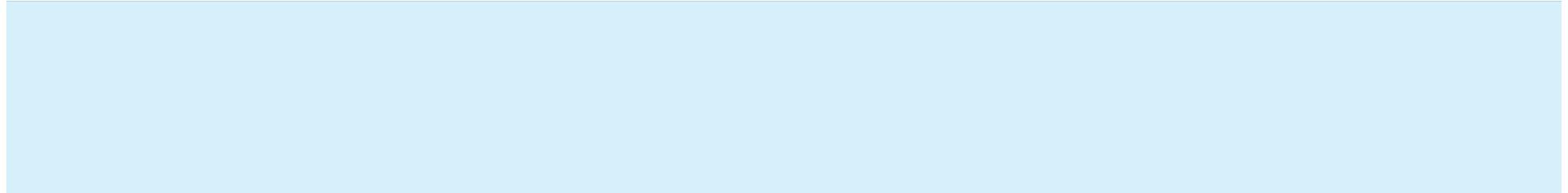
(128,128,128)



- Ratios more important than magnitude
- Most continuous variation in stimuli are perceived in discrete steps



Estimating Magnitude



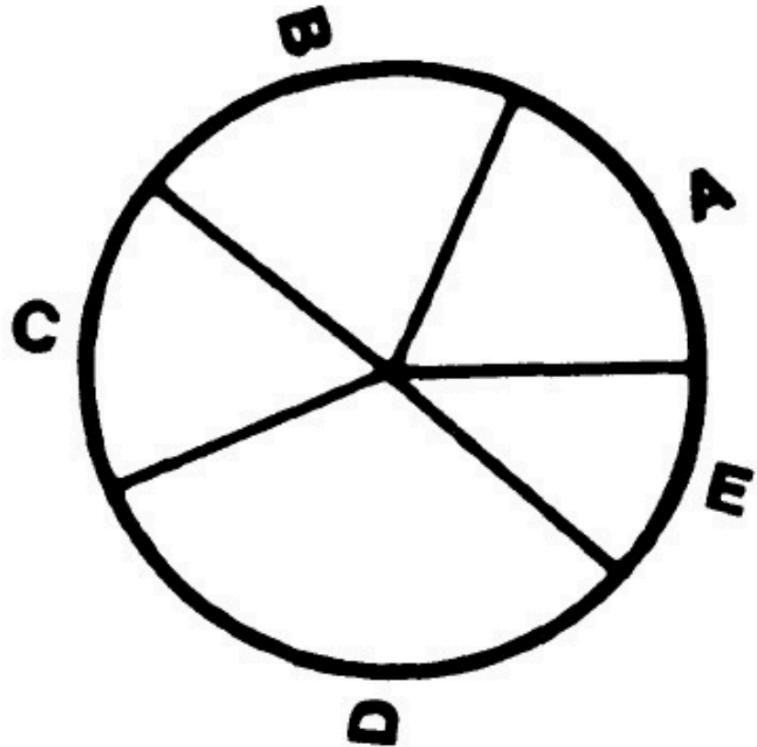
Compare length of bars

Estimating Magnitude



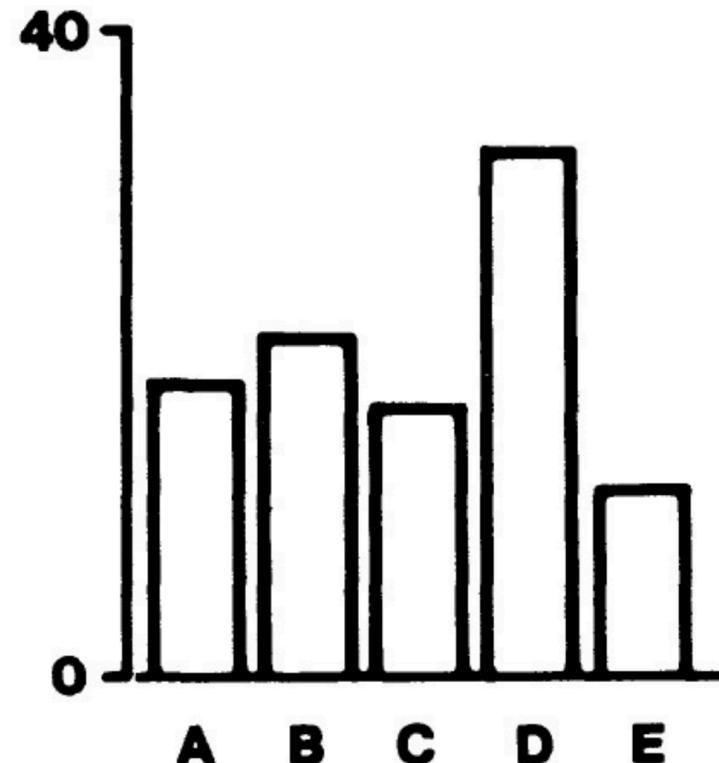
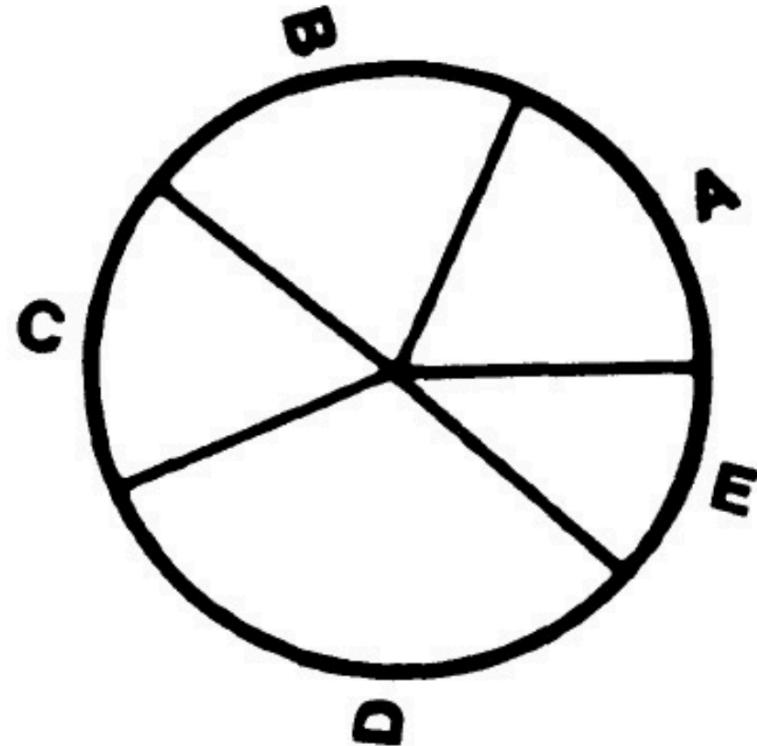
Compare length of bars

Estimating Magnitude



Which section is bigger? A or C?

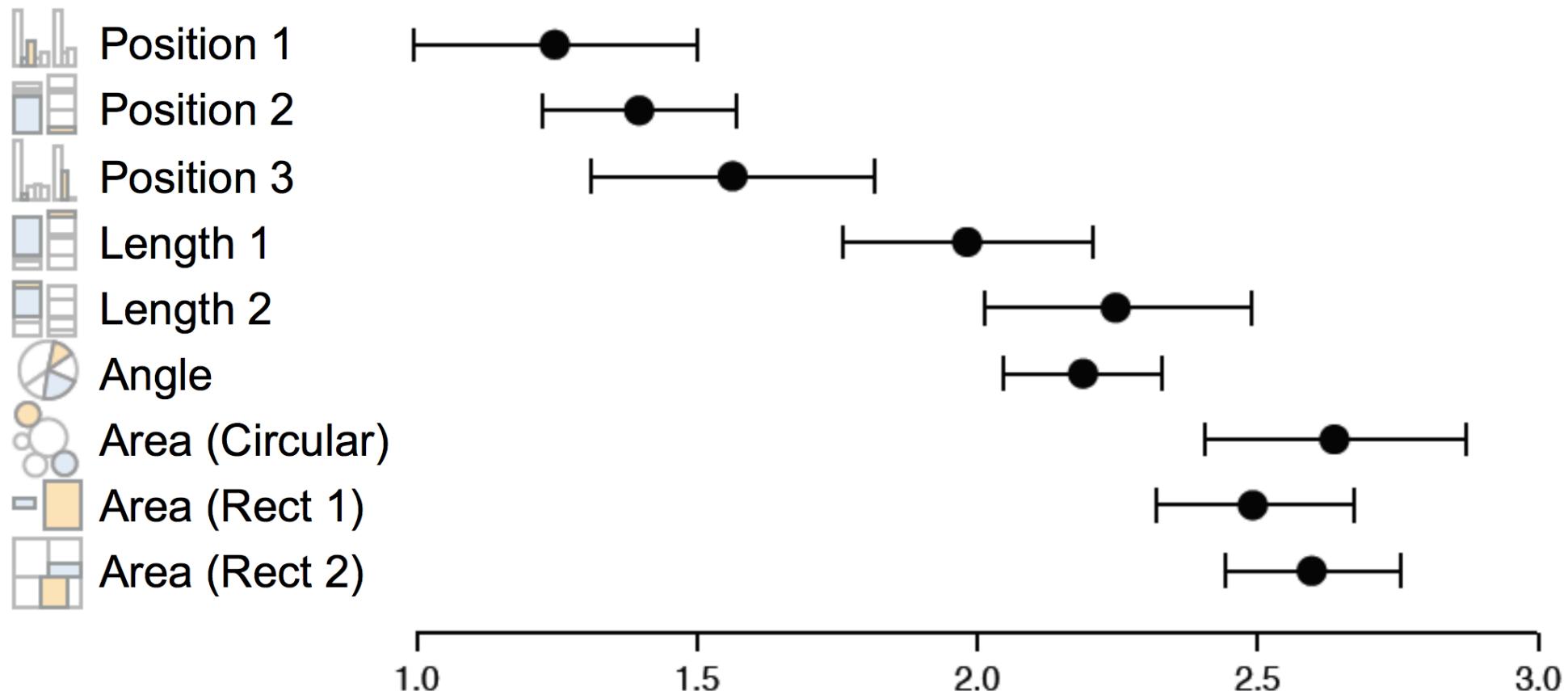
Estimating Magnitude



Which section is bigger? A or C?

Graphical Perception Experiments

- Empirical estimates of encoding effectiveness



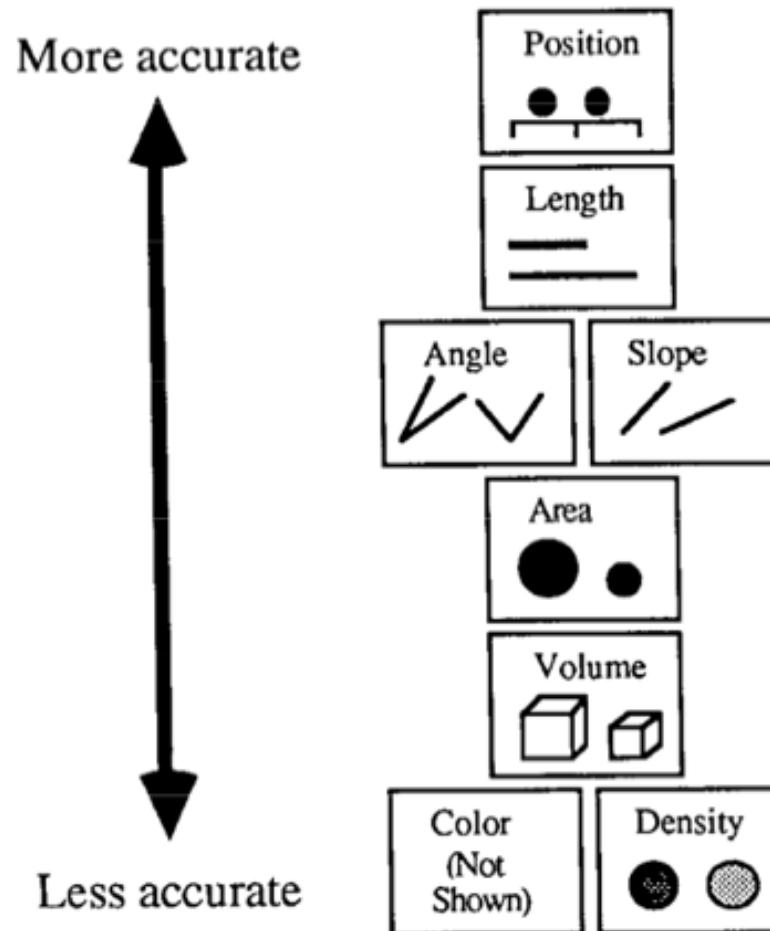
Heer & Bostock '10

(Optional Reading) Crowdsourcing Graphical Perception:
Using Mechanical Turk to Assess Visualization Design

Log Absolute Estimation Error

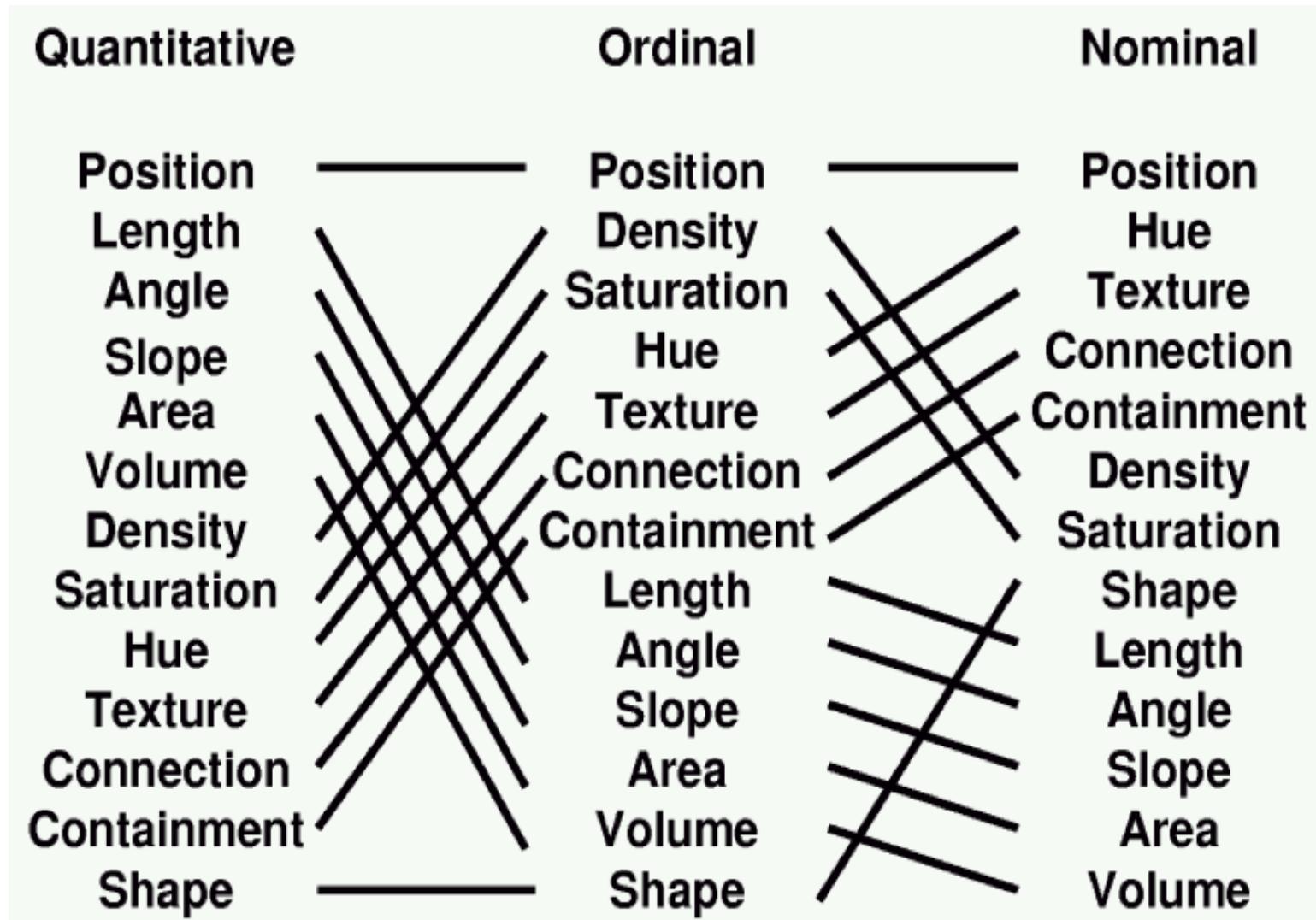
Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Effectiveness: Accuracy Ranking



Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Conjectured Effectiveness of Encodings by Data Type

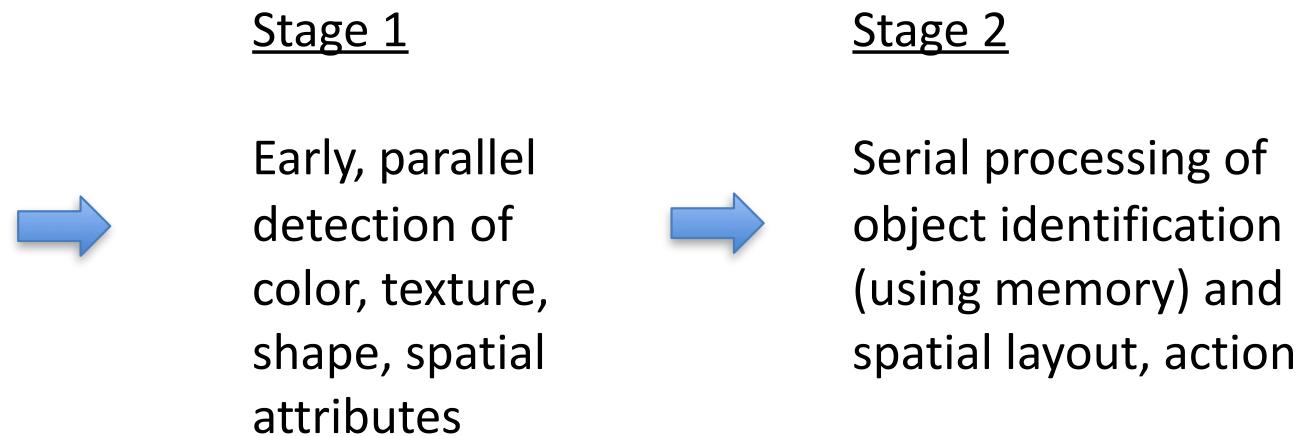
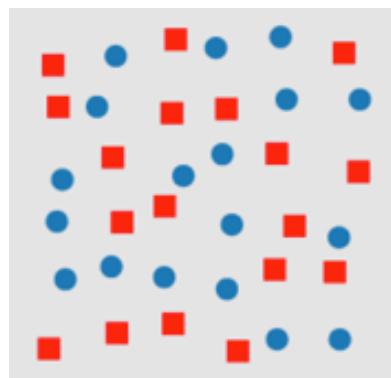


Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Perceptual Processing

Perceptual Processing Model

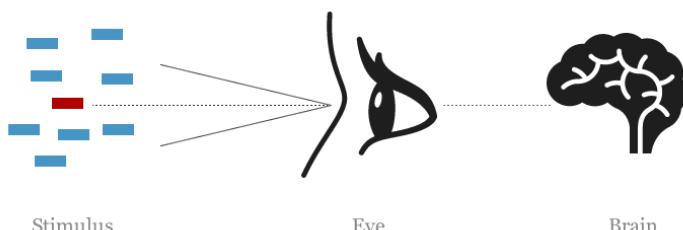
- Two stage process
 - Parallel extraction of low-level properties of scene
 - Sequential goal-directed processing



Stage 1: Pre-attentive Processing

- Low-level, Parallel

- Neurons in eye & brain responsible for different kinds of information
 - Orientation, color, texture, movement, etc.
- Arrays of neurons work in parallel, occurs “automatically” and rapidly
 - Generally less than 200-250 msecs
- Information is transitory, briefly held in iconic store
- Bottom-up data-driven model of processing
- Often called “pre-attentive” processing, i.e. without the need for focused attention



Stage 2 - Sequential, Goal-Directed

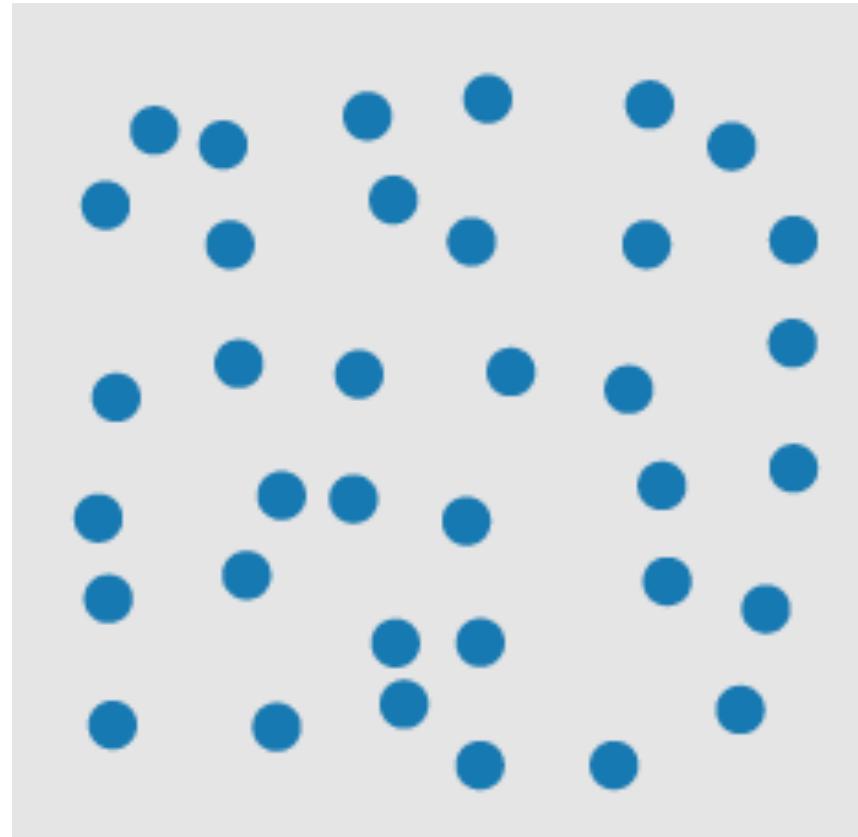
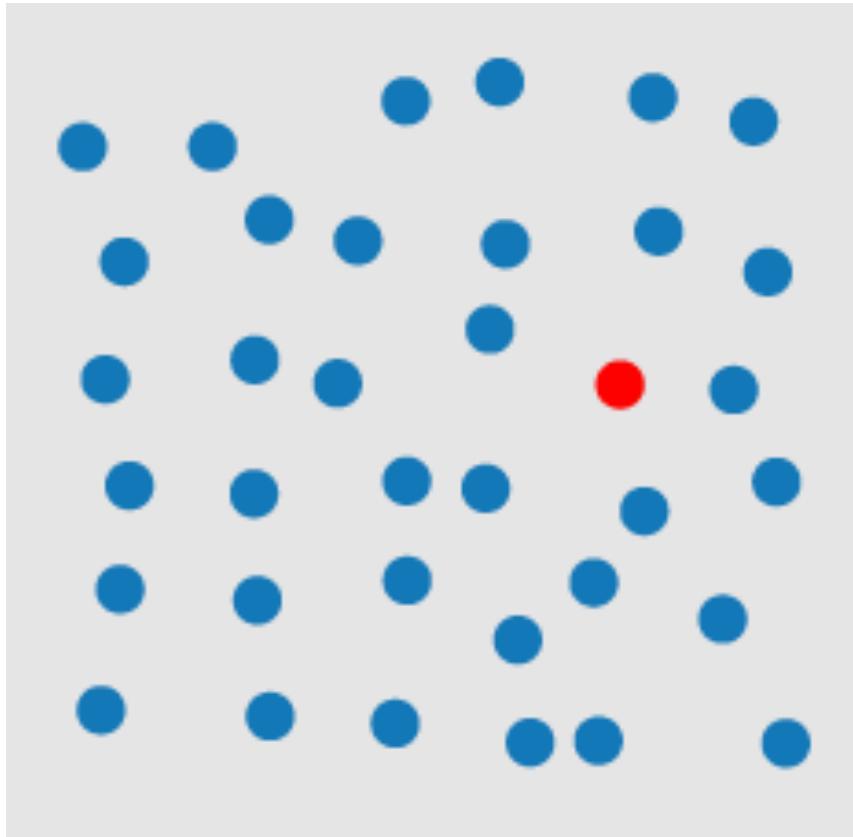
- Splits into subsystems for object recognition and for interacting with environment
- Increasing evidence supports independence of systems for symbolic object manipulation and for locomotion & action
- First subsystem then interfaces to verbal linguistic portion of brain, second interfaces to motor systems that control muscle movements
- Slow serial processing
- Involves working and long-term memory

Pre-attentive Processing

How many 3's?

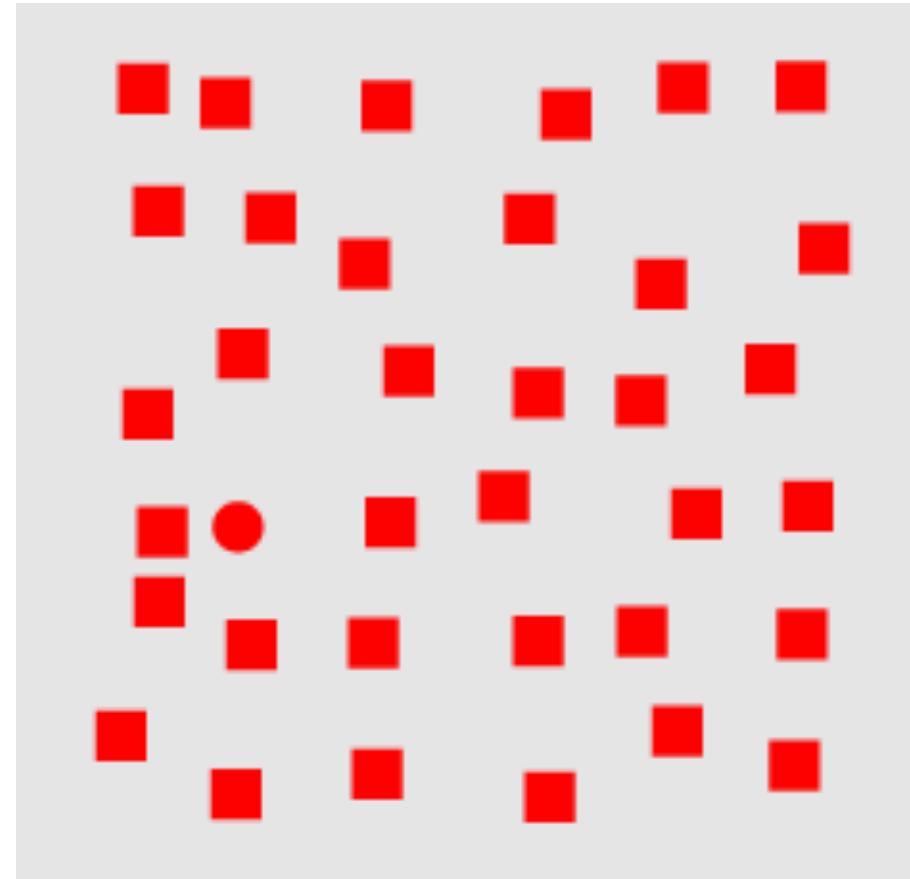
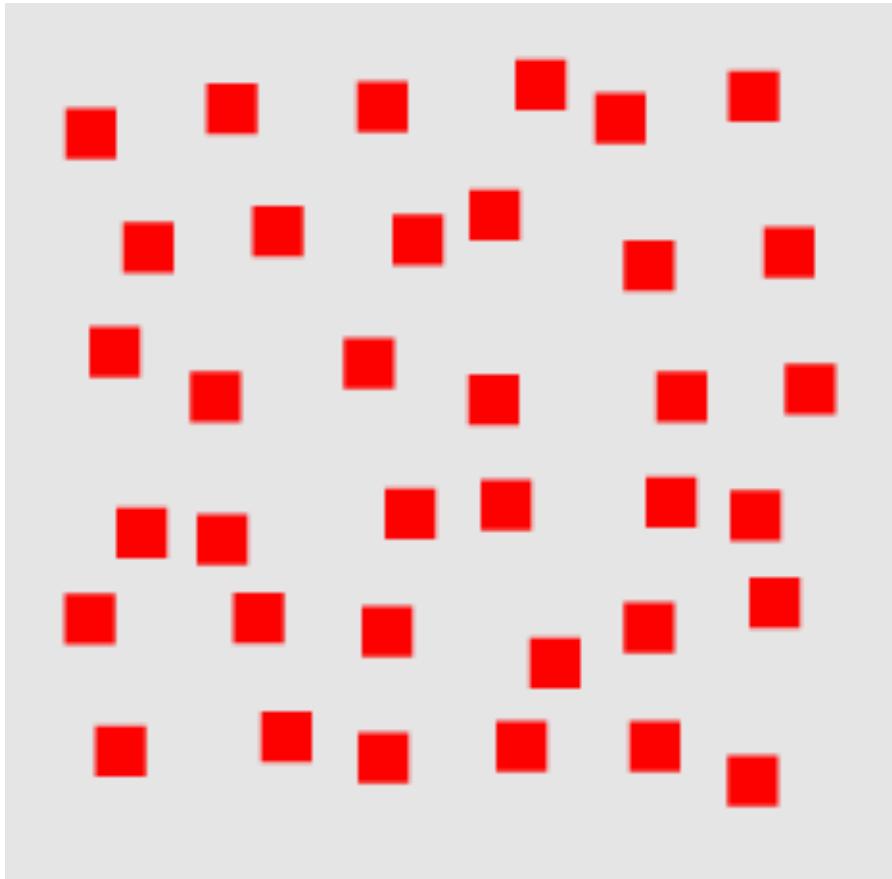
12817687561**3**8976546984506985604982826762
980985845822450985645894509845098094**3**585
90910**3**02099059595772564675050678904567
8845789809821677654876**3**64908560912949686

Visual Pop-Out: Color (Hue)



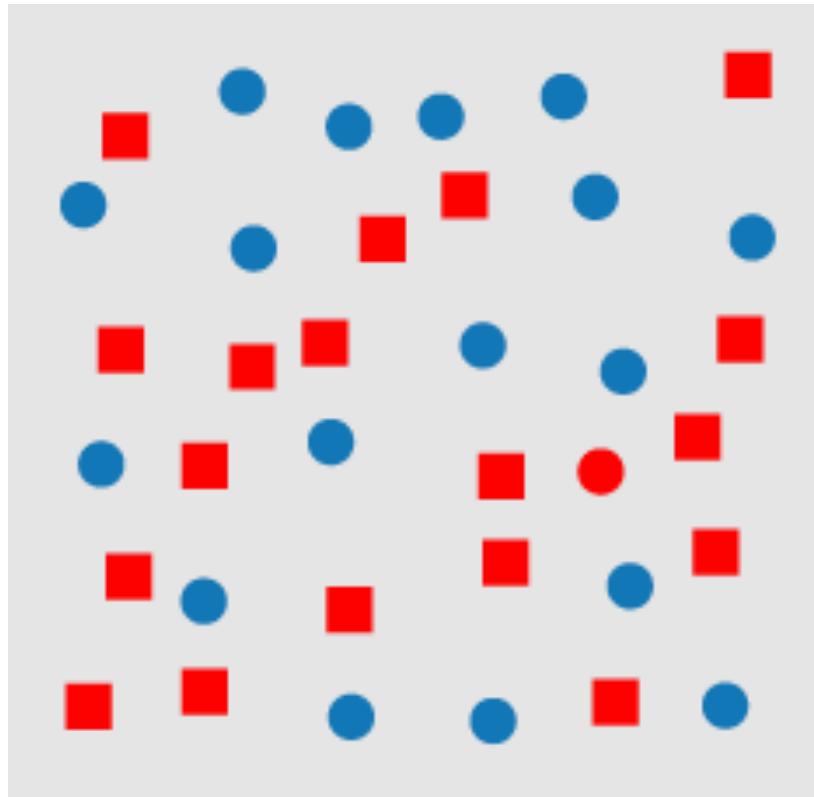
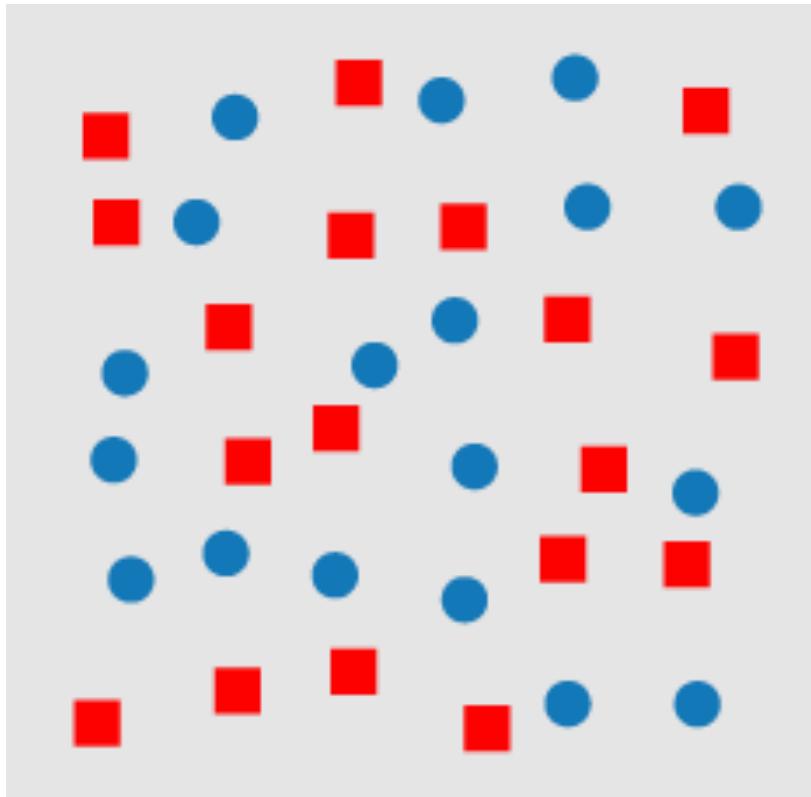
Can be done rapidly (preattentively) by people
Surrounding objects called “distractors”

Visual Pop-Out: Shape

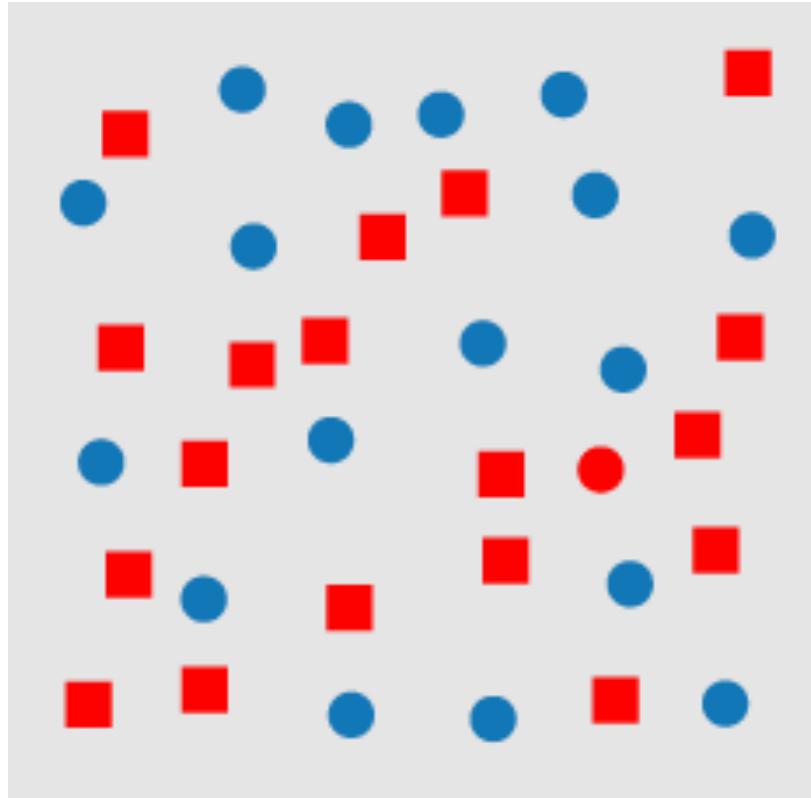
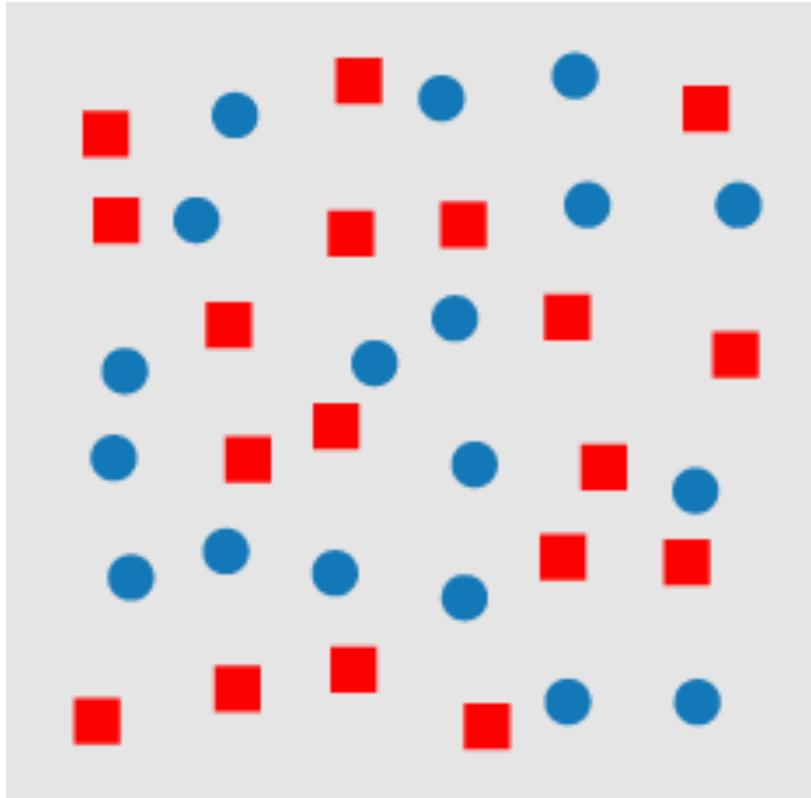


Can be done preattentively by people

Feature Conjunctions: Color and Shape

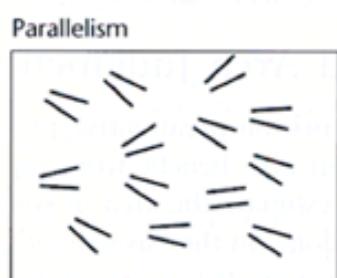
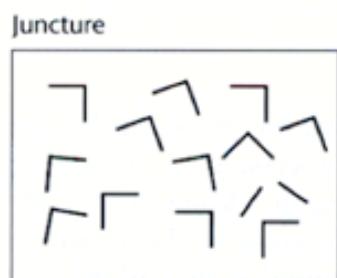
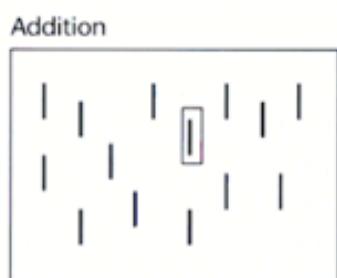
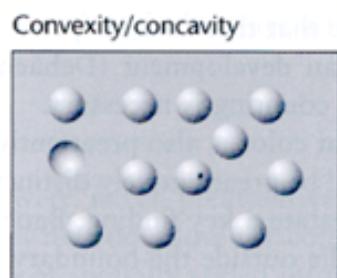
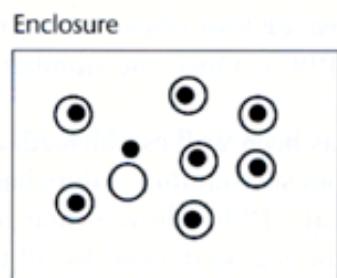
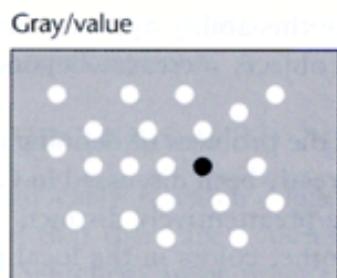
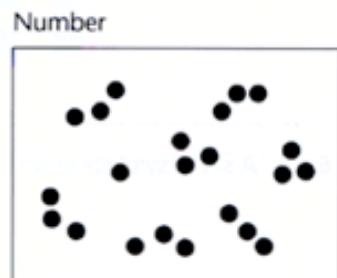
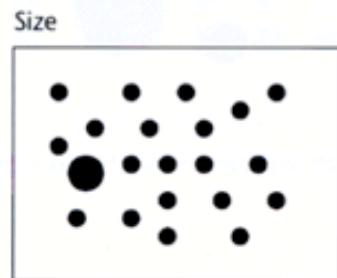
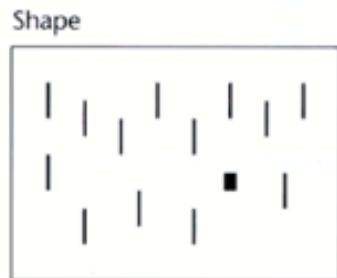
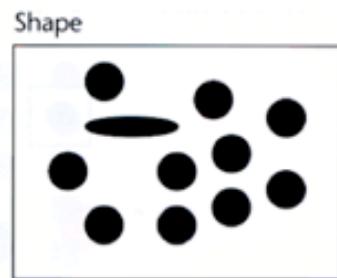
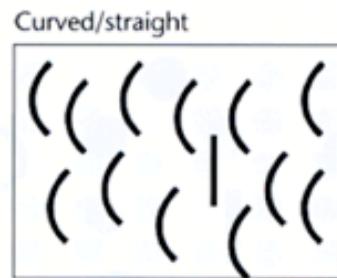
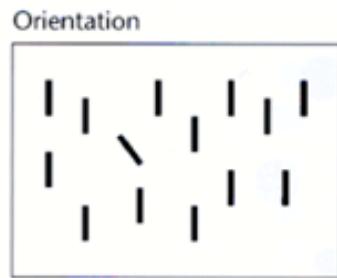


Feature Conjunctions: Color and Shape



- Cannot be done preattentively
- Must perform a sequential search
- Conjunction of features (shape and hue) causes it

Pre-Attentive Features



- length
- width
- size
- curvature
- number
- terminators
- intersection
- closure
- hue
- intensity
- flicker
- direction of motion
- binocular lustre
- stereoscopic depth
- 3-D depth cues
- lighting direction

Pre-Attentive Feature Conjunctions

- Spatial conjunctions are often pre-attentive
- Motion and 3D disparity
- Motion and color
- Motion and shape
- 3D disparity and color
- 3D disparity and shape
- Most conjunctions are not pre-attentive

Gestalt Grouping Principles

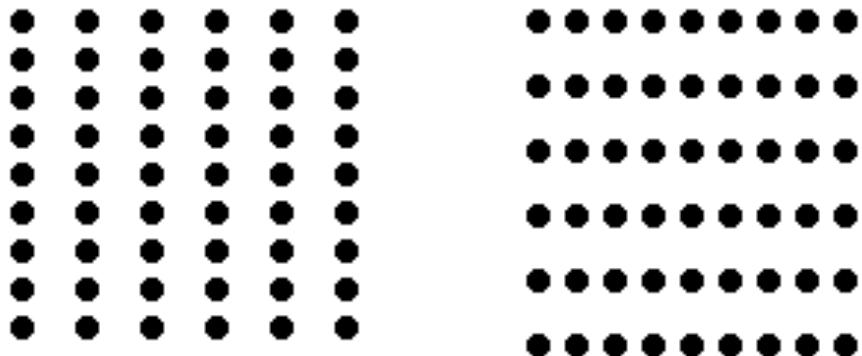
“All else being equal, elements that are related by X tend to be grouped perceptually into higher-order units.”

— Stephen Palmer

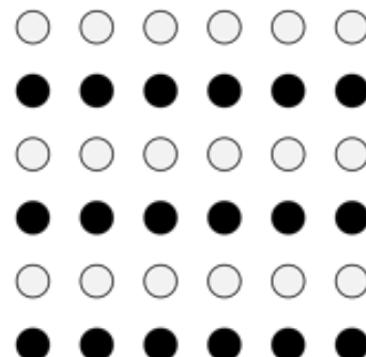
- Proximity
- Similarity
- Connectedness
- Continuity
- Symmetry
- Closure
- Figure/Ground
- Common Fate

Gestalt Grouping Principles

- Proximity
 - Things close together are perceptually grouped together



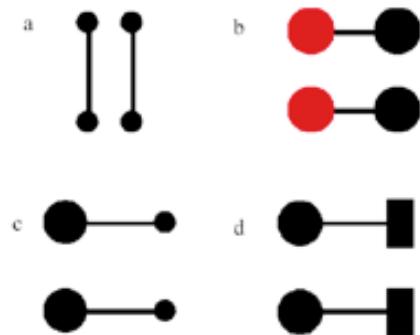
- Similarity
 - Similar elements get grouped together



Rows dominate due to similarity

Gestalt Grouping Principles

- Connectedness
 - Connecting different objects by lines unifies them
- Continuity
 - More likely to construct visual entities out of smooth, continuous visual elements



Connectedness
overrides
proximity, size,
color shape



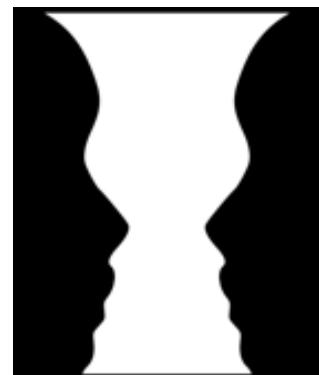
Gestalt Grouping Principles

- Symmetry
 - Symmetrical patterns are perceived more as a whole
- Closure
 - A closed contour is seen as an object

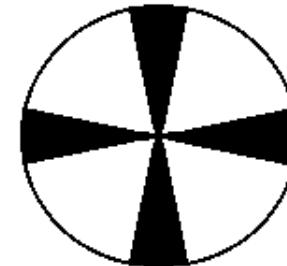


Gestalt Grouping Principles

- Figure/Ground
 - Figure is foreground, ground is behind
- Common Fate (Synchrony)
 - Elements that move in the same direction are perceived as more related



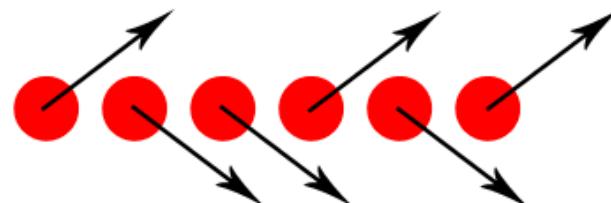
ambiguous



Relative size



surroundedness



Change Blindness

- We don't always see everything that is there!
- Is the viewer able to perceive changes between two scenes?
 - If so, may be distracting
 - Can do things to minimize noticing changes
- Video: <http://www.simonslab.com/videos.html>

Next Lecture

- Topic:
 - Interaction
- Next Monday (25 Feb)
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus



G53FIV: Fundamentals of Information Visualization

Lecture 9: Interactions

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

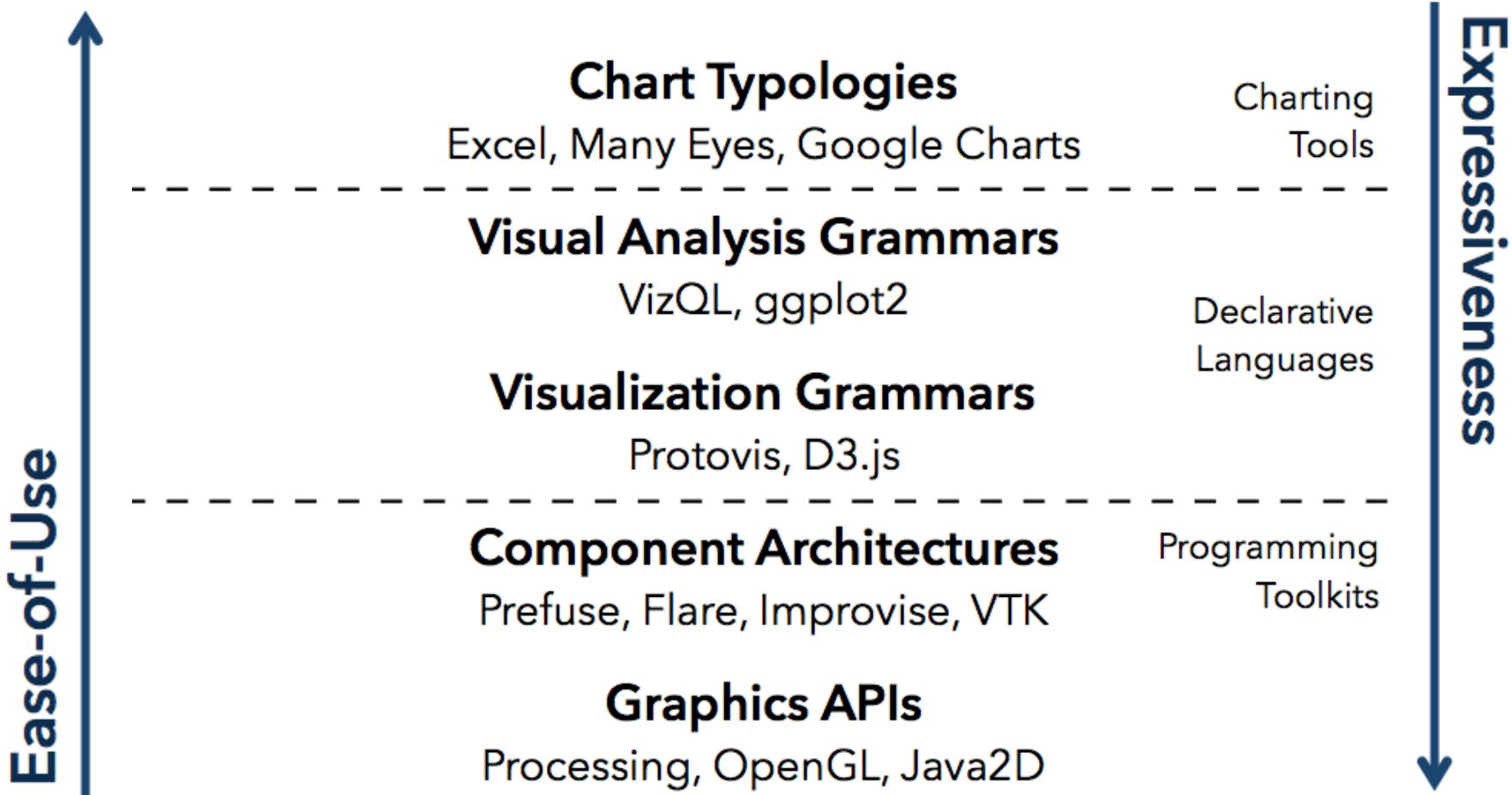
Administria

- Course work issue sheet made available on Moodle, with marking criteria
- Second optional lab session at 2:00 - 3:00 PM Monday in A32
 - for those who can not make it on the 9:00 - 10:00 AM Monday session

Last Lecture

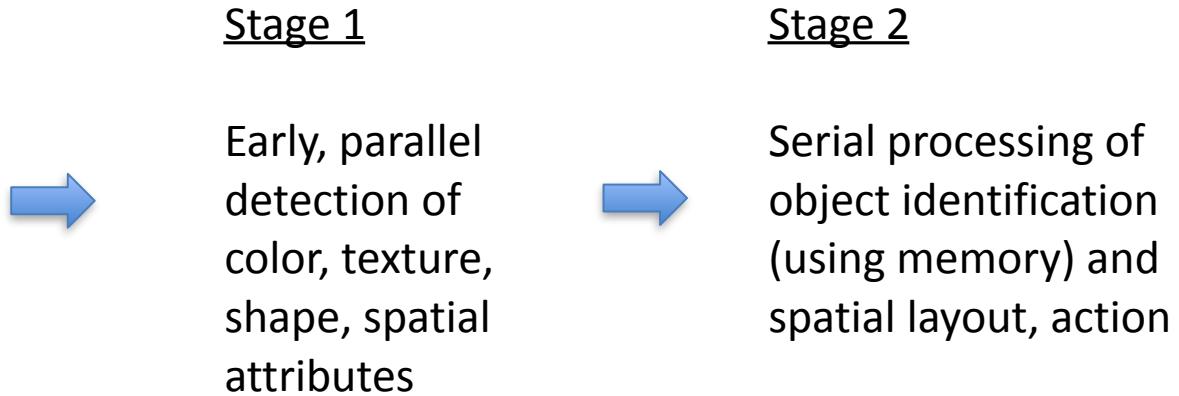
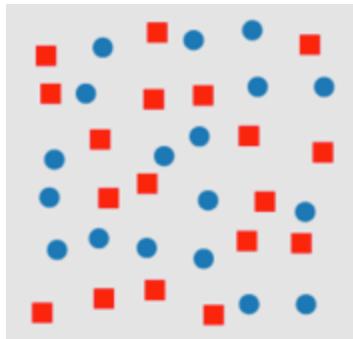
Visualization Tools and Visual Perception

Visualization Tools

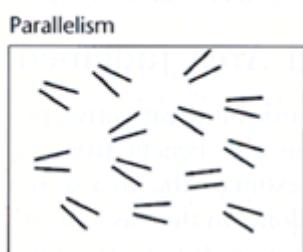
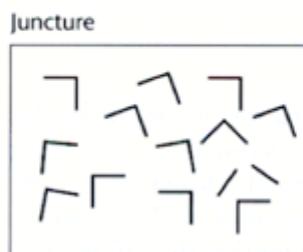
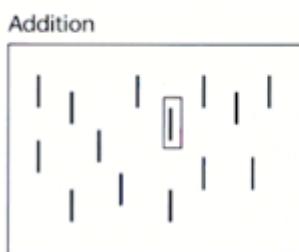
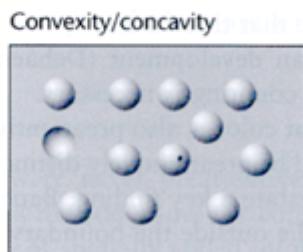
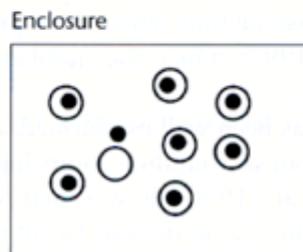
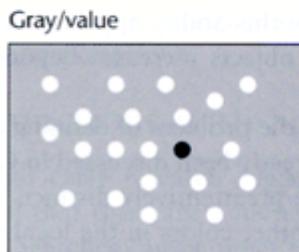
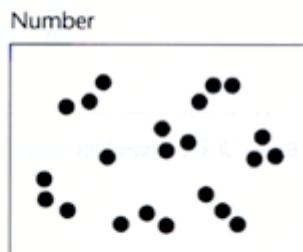
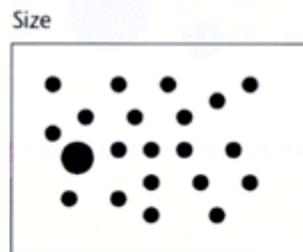
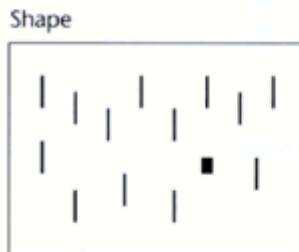
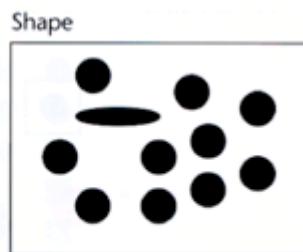
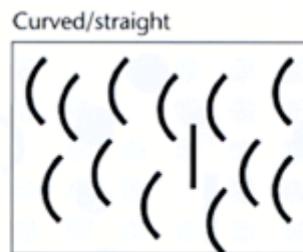
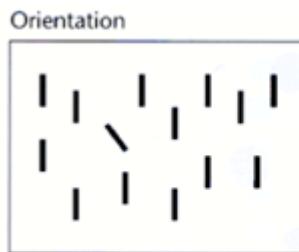


Perceptual Processing Model

- Two stage process
 - Parallel extraction of low-level properties of scene
 - Sequential goal-directed processing



Pre-Attentive Features



- length
- width
- size
- curvature
- number
- terminators
- intersection
- closure
- hue
- intensity
- flicker
- direction of motion
- binocular lustre
- stereoscopic depth
- 3-D depth cues
- lighting direction

Gestalt Grouping Principles

“All else being equal, elements that are related by X tend to be grouped perceptually into higher-order units.”

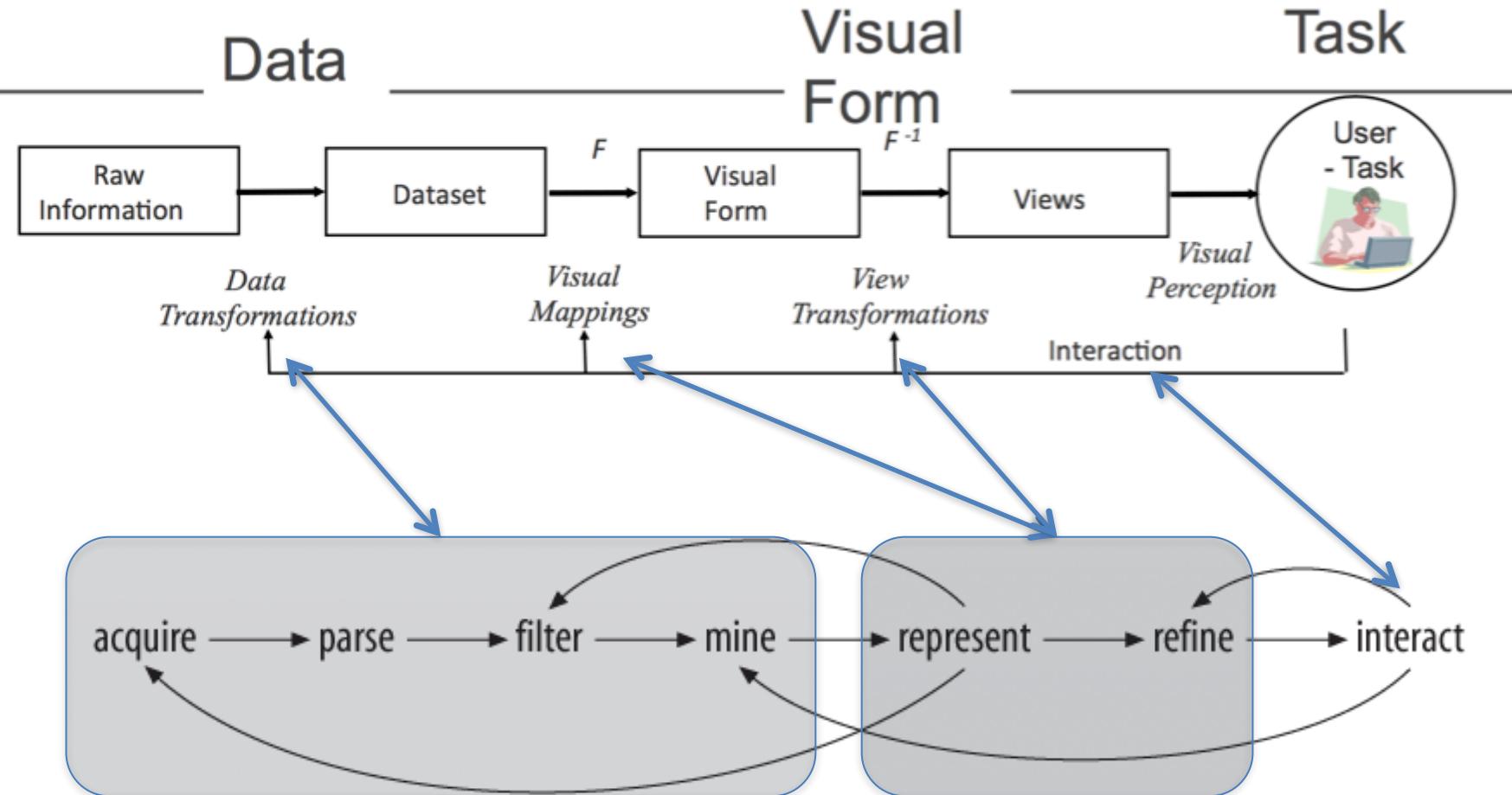
— Stephen Palmer

- Proximity
- Similarity
- Connectedness
- Continuity
- Symmetry
- Closure
- Figure/Ground
- Common Fate

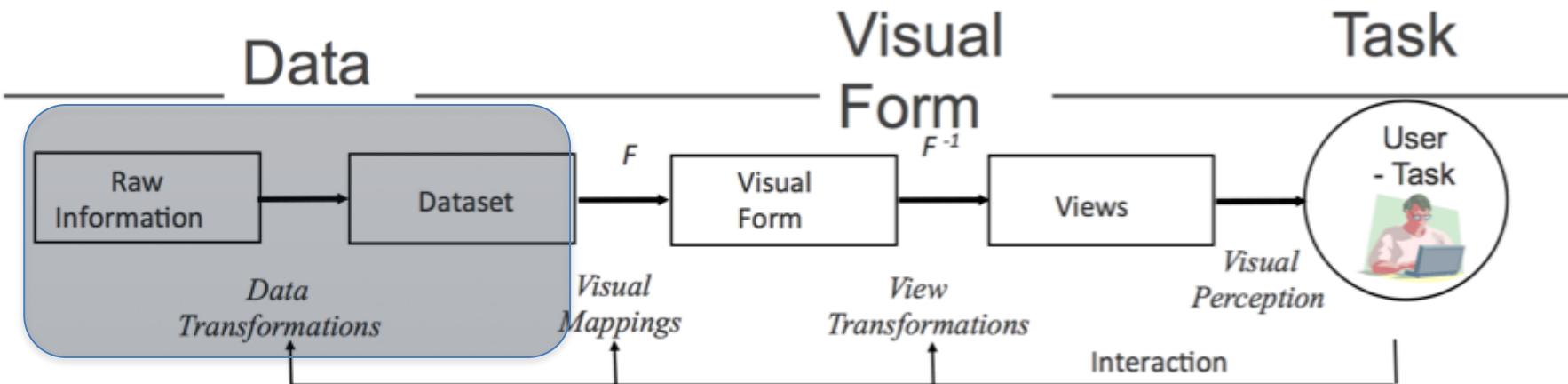
Overview

- Interaction
 - Definition
 - Taxonomy of Interactions
 - Pros and Cons

Different Stages of Visualization



Different Stages of Visualization



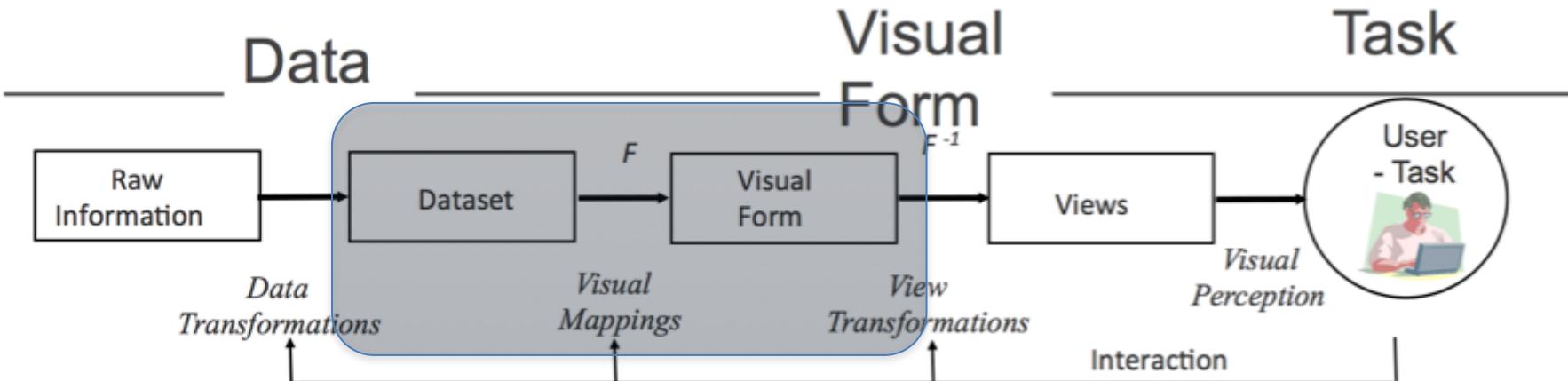
00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00544	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00601	+18.165273	-066.722583		ADJUNTAS	72	001
00602	+18.393103	-067.180953		AGUADA	72	003
00603	+18.455913	-067.145780		AGUADILLA	72	005
00604	+18.493520	-067.135883	P	AGUADILLA	72	005
00605	+18.465162	-067.141486		AGUADILLA	72	005
00606	+18.172947	-066.944111		MARICAO	72	093
00610	+18.288685	-067.139696		ANASCO	72	011



- FILTER Rows
- SELECT Column Types
- ArRANGE Rows (SORT)
- Mutate (into something new)
- Summarize by Groups



Different Stages of Visualization

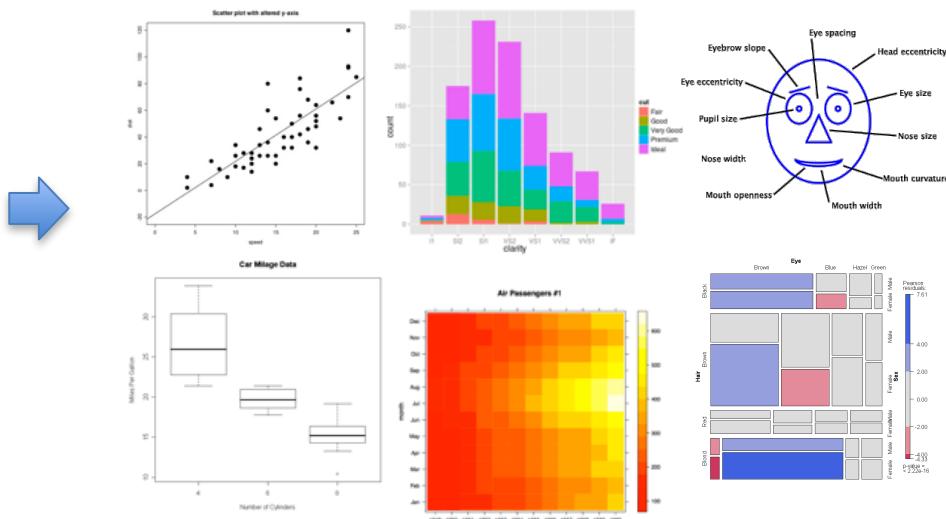


```

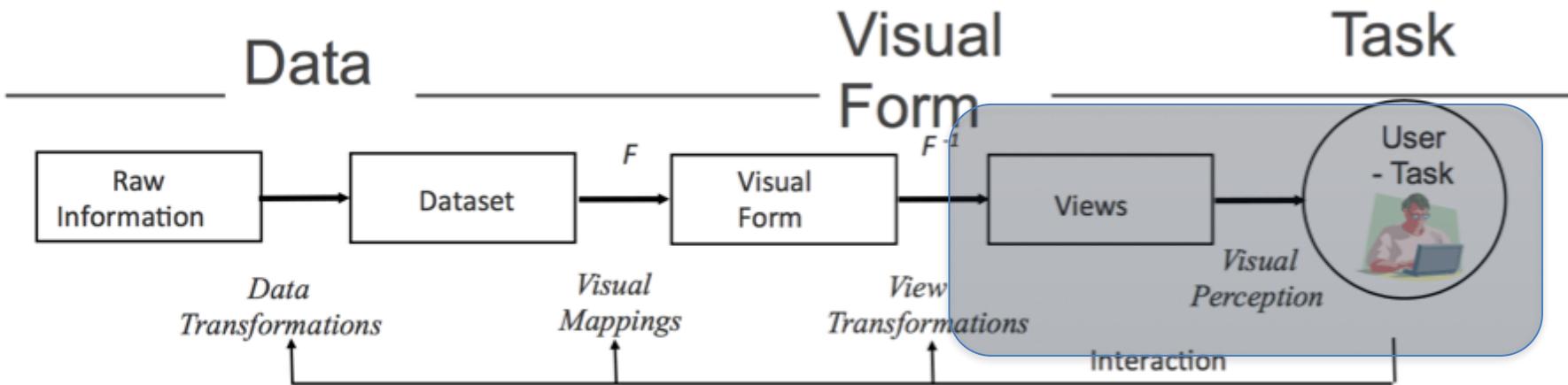
00210 +43.005895 -071.013202 U PORTSMOUTH 33 015
00211 +43.005895 -071.013202 U PORTSMOUTH 33 015
00212 +43.005895 -071.013202 U PORTSMOUTH 33 015
00213 +43.005895 -071.013202 U PORTSMOUTH 33 015
00214 +43.005895 -071.013202 U PORTSMOUTH 33 015
00215 +43.005895 -071.013202 U PORTSMOUTH 33 015
  
```

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good
 ~ = OK
 ✗ = Bad



Different Stages of Visualization



“The effectiveness of information visualization hinges on two things: its ability to clearly and accurately **represent** information and our ability to **interact** with it to figure out what the information means.”

S. Few, <Now you see it>

Representation and Interaction

- Two main components of information visualization
- Very challenging to come up with innovative, new visual representations
- But can do interesting work with how user interacts with the view or views
 - Analysis is a process, often iterative with different interactions

Why we need interaction?

- For larger data, there is simply too much to show in a coherent manner
 - With more variables, more data cases, it will be hard for users to perceive everything in one go.
 - Limited screen, limited cognitive ability, limited time, etc.
- Interaction helps us address that challenge
 - We want to help users to better accomplish their tasks

What is “interactive”?

- Can be captured and measured by the response time
 - .1 sec
 - animation, visual continuity, sliders
 - 1 sec
 - system response, conversation break
 - 10 sec
 - cognitive response



An Example

- Dust and Magnet

Video: [Dust & Magnet](#)

<https://www.youtube.com/watch?v=wLXwL38xek0>

Let's be interactive 😊

An Exercise

- List the different “categories” of interaction in information visualization
- Work in pairs

Taxonomy of Interactions

- Dix and Ellis (1998)
 - Highlighting and focus;
 - accessing extra info;
 - overview and context;
 - same representation, changing parameters;
 - Linking representations
- Keim (2002)
 - Projection
 - Filtering
 - Zooming
 - Distortion
 - Linking and brushing
- Few's Principles
 - Comparing
 - Sorting
 - Adding variables
 - Filtering
 - Highlighting
 - Aggregating
 - Re-expressing
 - Re-visualizing
 - Zooming and panning
 - Re-scaling
 - Accessing details on demand
 - Annotating
 - Bookmarking

Challenges

- Interaction seems to be a difficult thing to pin down and characterize
- User-centered versus system-centered characterizations
 - User intent: what a user wants to achieve through a specific interaction technique

A Summary of Interactions

- Survey
 - 59 papers
 - Papers introducing new interaction systems
 - Well-known papers in subareas of information visualization
 - 51 systems
 - Commercial Infovis Systems (SeeIT, Spotfire, TableLens, InfoZoom, etc.)
 - Collected 311 individual interaction techniques
- Affinity Diagram Method

Yi, Ji Soo, Youn ah Kang, and John Stasko. "Toward a deeper understanding of the role of interaction in information visualization." IEEE transactions on visualization and computer graphics 13.6 (2007): 1224-1231.

Categorization based on User Intent

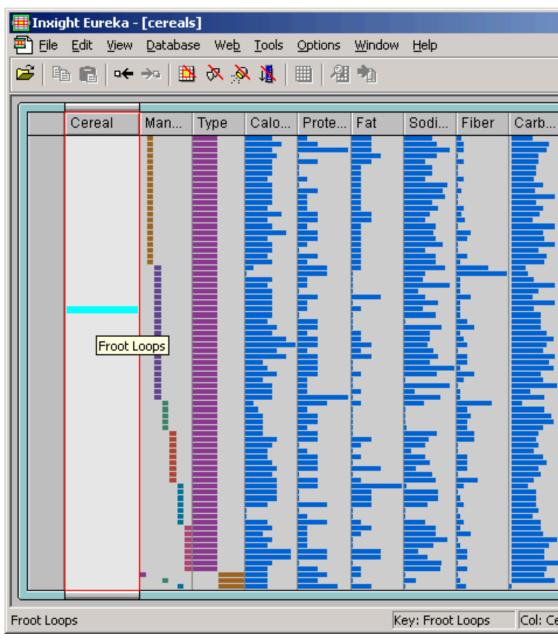
- Don't focus so much on particular interactive operations and how they work
- Interaction is ultimately being done by a person for a purpose
 - Seeking more information, solving a problem
 - Fundamental aspect of exploratory, analytic discourse
- Taxonomy based on **User Intent**
 - What a user wants to achieve through a specific interaction technique

Taxonomy of Interactions based on User Intent - 7 Categories

- Select
- Explore
- Reconfigure
- Encode
- Abstract/Elaborate
- Filter
- Connect

1. Select

- “Mark something as interesting”
- Mark items of interest to keep track
- Seems to often work as a preceding action to subsequent operations.
- Selecting a placemark in Google Map
- The Focus feature in TableLens

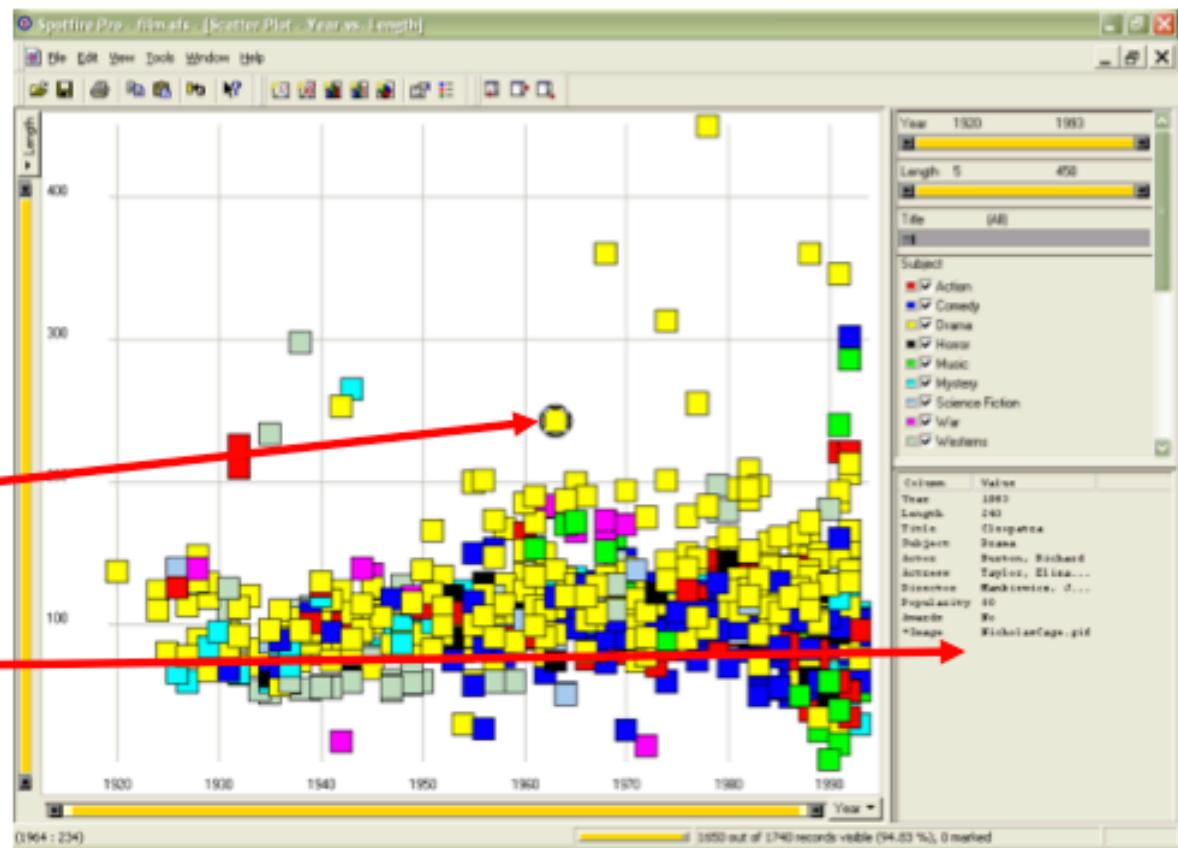


Mouse Selection

Clicking on an item selects it and attributes of the data point are shown

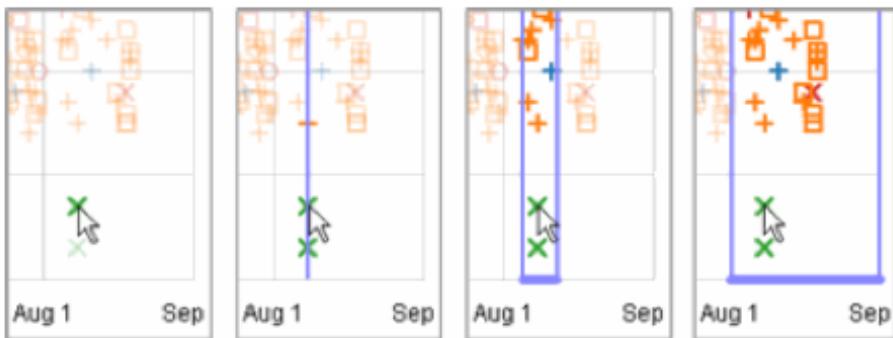
Selected item

Attributes



Generalized Selection

- The idea: you want to select items matching some attribute(s) of that item (rather than caring only about the precise item)



Video: [http://vis.berkeley.edu/
papers/generalized_selection/](http://vis.berkeley.edu/papers/generalized_selection/)

- As you dwell on your mouse pick, the selection criteria broaden and you can choose sets of items

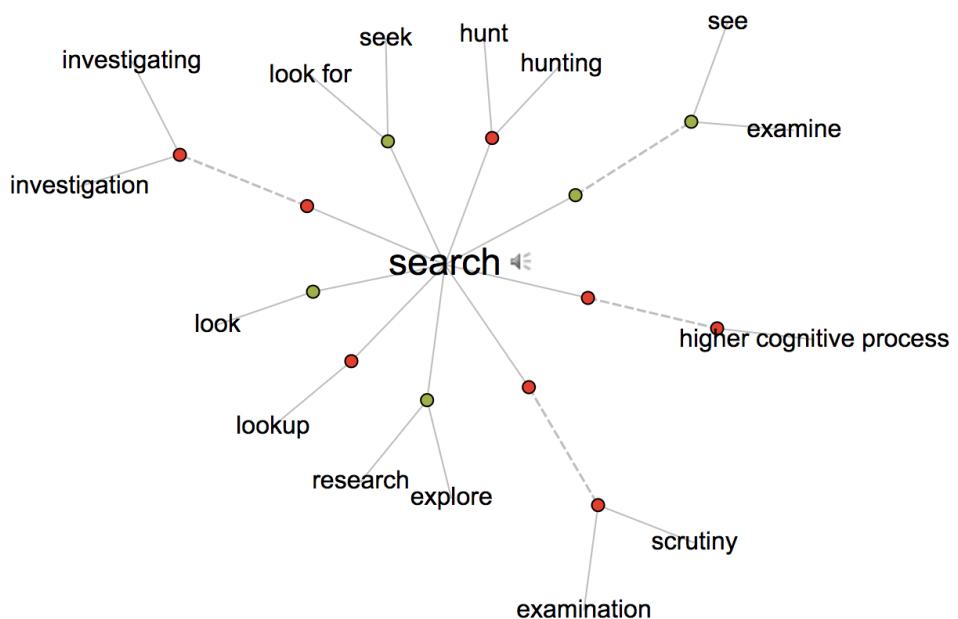
2. Explore

- “Show me something different”
- Enable users to examine a different subset of data
- Overcome the limitation of display size
- Panning in Google Earth
- Direct Walking in Visual Thesaurus



Direct Walk

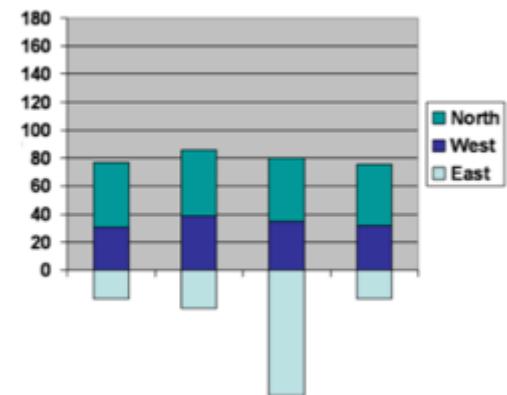
- Linkages between cases
- Exploring one may lead to another
- Example:
 - Visual Thesaurus



Demo: <https://www.visualthesaurus.com/app/view>

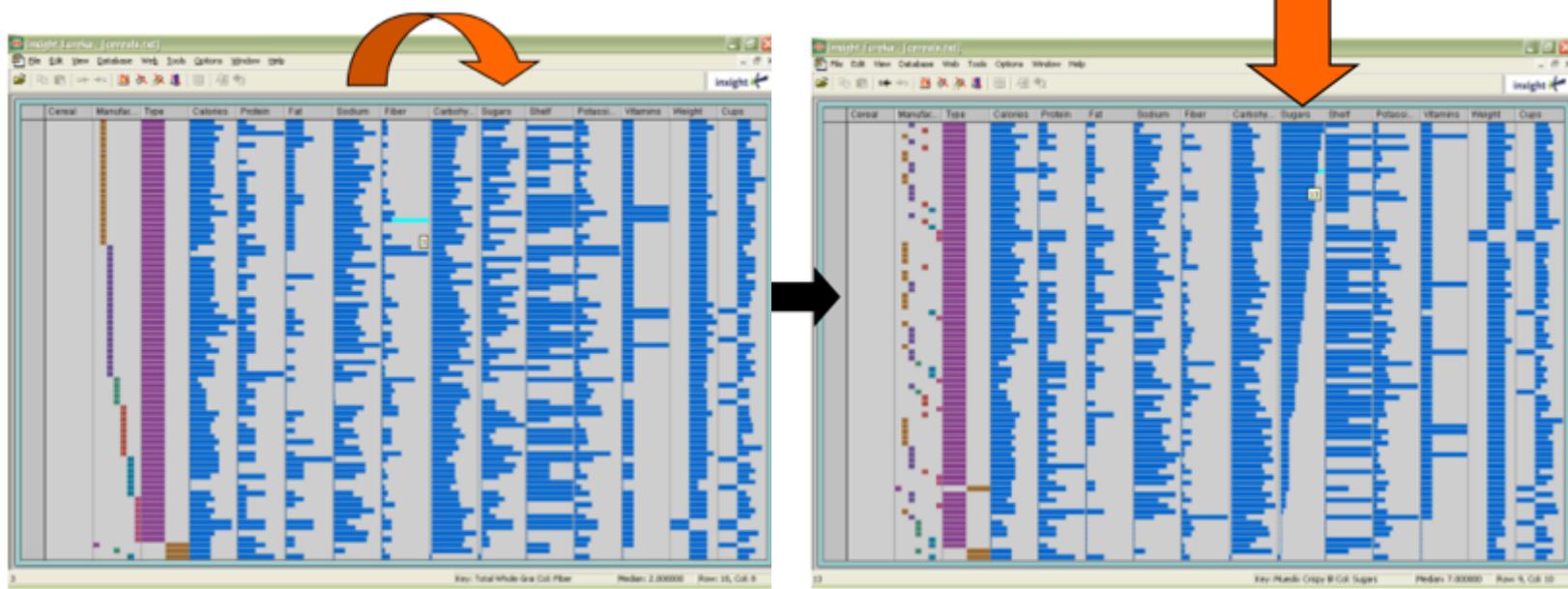
3. Reconfigure

- “Show me a different arrangement”
- Provide different perspectives by changing the spatial arrangement of representation
- Sorting and rearranging columns in TableLens
- Changing the attributes in a scatter plot
- The baseline adjustment feature in Stacked Histogram:
 - <http://meandeviation.com/dancing-histograms/>
- The “Spread Dust” feature in Dust & Magnet



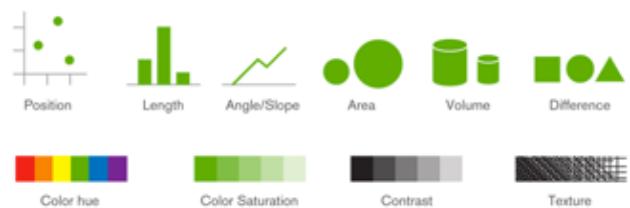
Rearrange View and Sorting

- Keep same fundamental representation and what data is being shown, but rearrange elements
 - Alter positioning
 - Sort

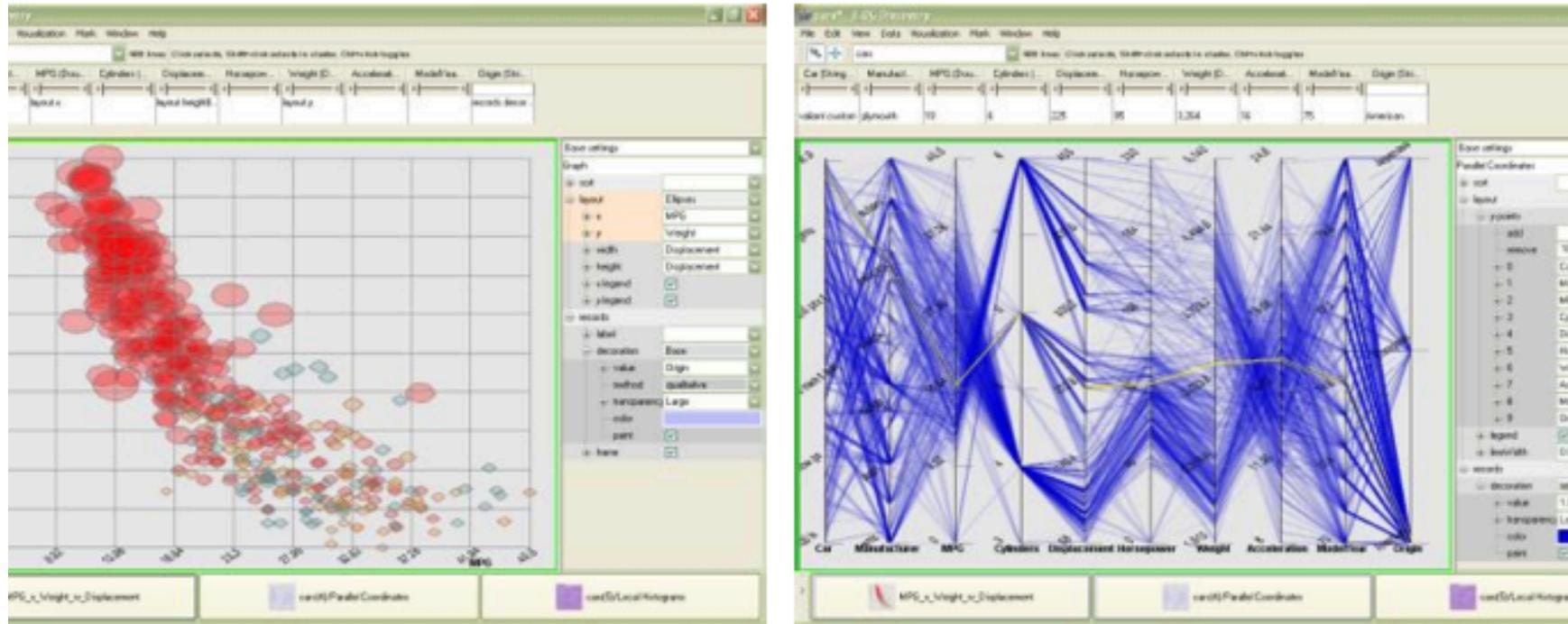


4. Encode

- “Show me a different representation”
- Change visual appearances
 - color encoding, size, orientation, font, shape
- May interactively change entire data representation
 - Looking for new perspective
 - Limited real estate may force change



Looking for New Perspective



Selecting different representation from options at bottom

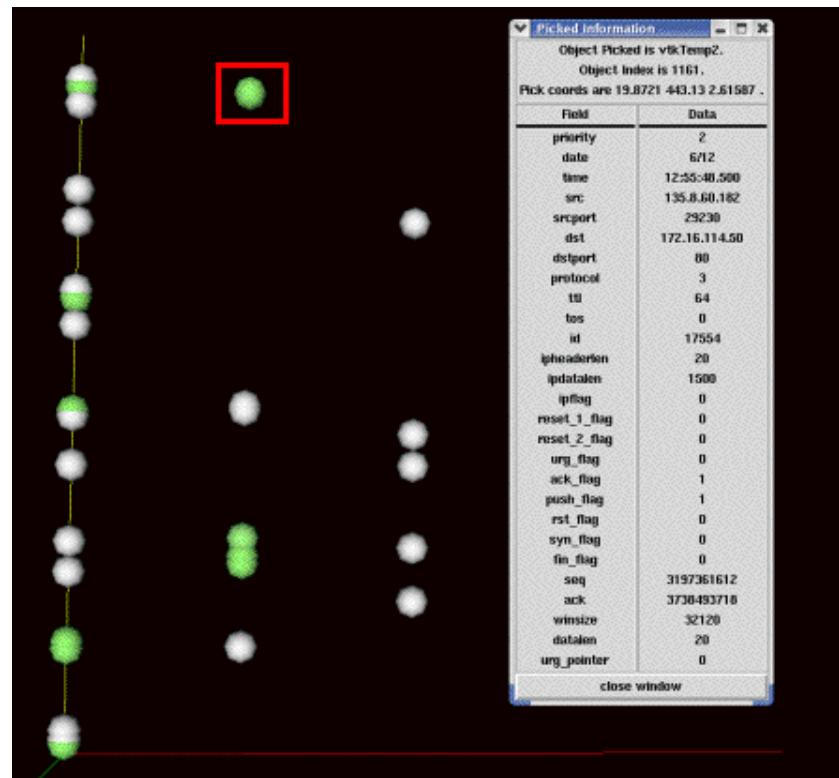
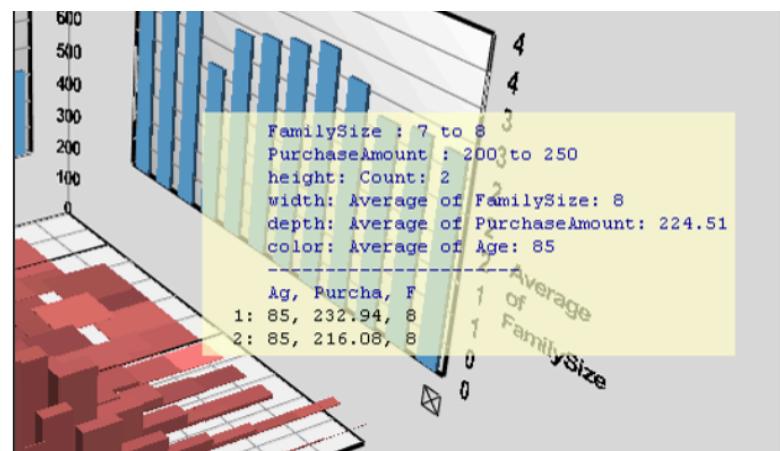
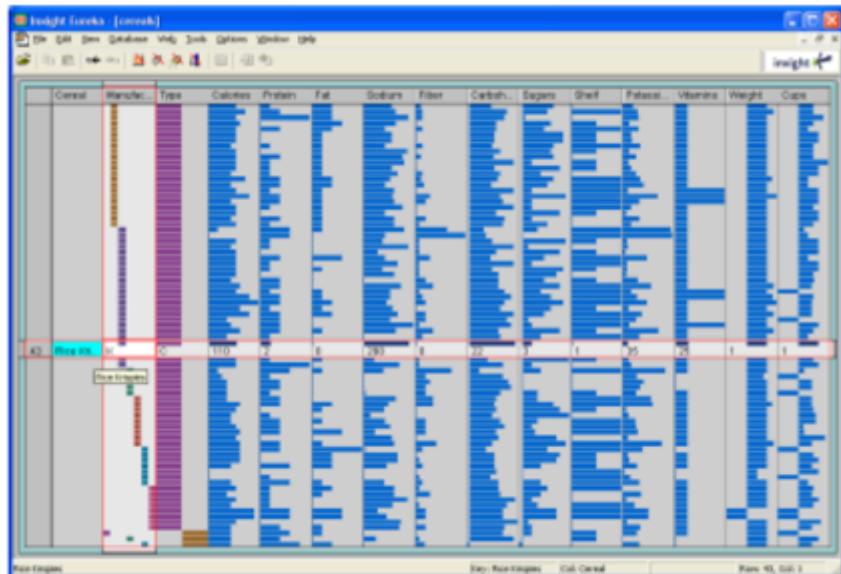
5. Abstract/Elaborate

- “Show me more or less detail”
- Adjust the level of abstraction (overview and details)
- Details-on-demand
- Unfolding sub-categories in an interactive pie chart
- Drill-down in Treemap
- Zooming (geometric zooming)

Details on Demand

- Term used in information visualization when providing viewer with more information/details about data case or cases
- May just be more information about a case
- May be moving from aggregation view to individual view
 - May not be showing all the data due to scale problem
 - May be showing some abstraction of groups of elements
 - Expand set of data to show more details, perhaps individual cases

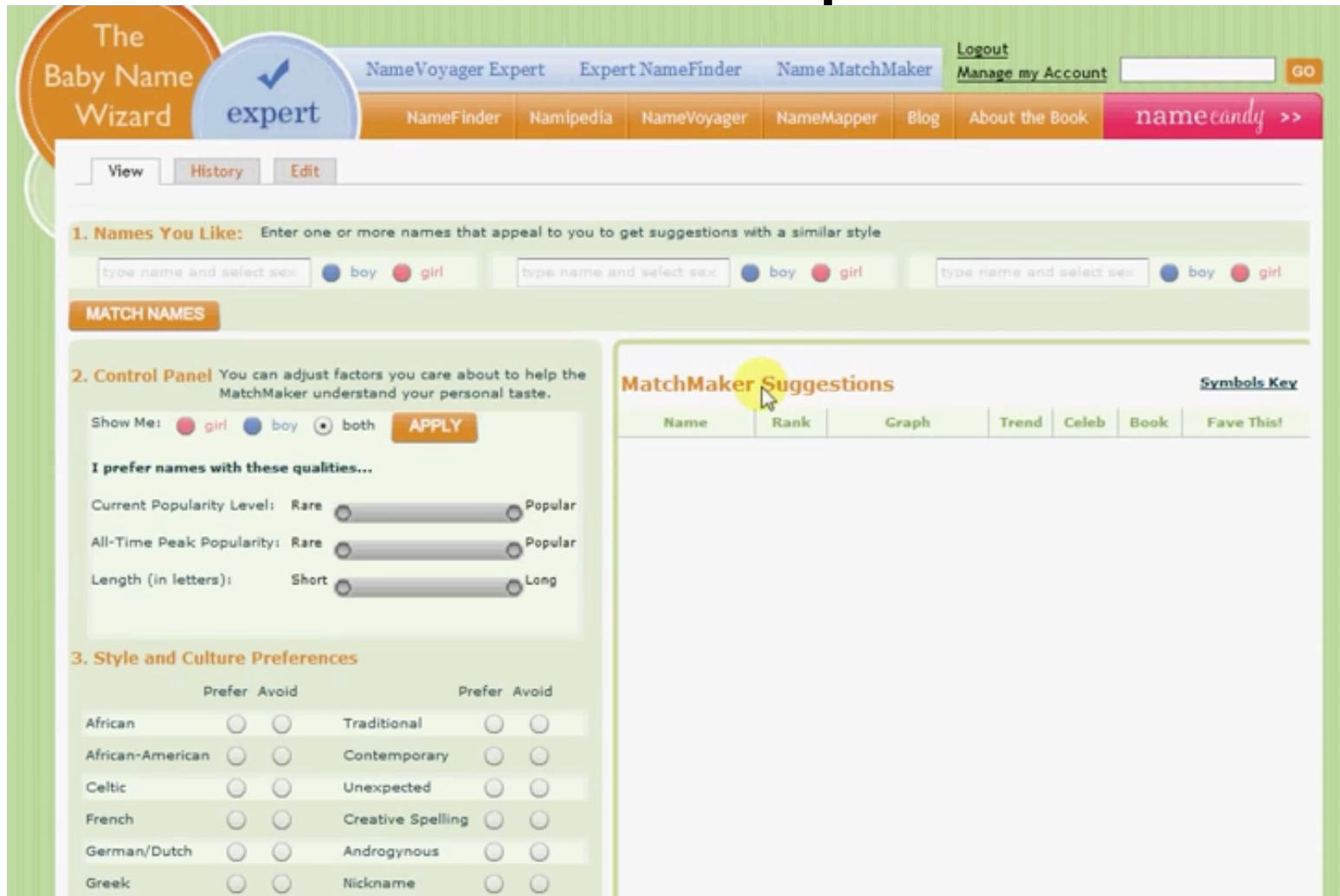
Example



6. Filter

- “Show me something conditionally”
- Change the set of data items being presented based on some specific conditions.
- Fundamental interactive operation in information visualization is changing the set of data cases being presented
 - Focusing
 - Narrowing/widening

An Example



The screenshot shows the Baby Name Wizard interface. On the left, there's a sidebar with sections for "Names You Like" (three input fields for boy/girl names), "Match Names" (a button), "Control Panel" (radio buttons for Show Me: girl, boy, both, with an "APPLY" button), and "Style and Culture Preferences" (a grid of 12 items with "Prefer" and "Avoid" checkboxes). On the right, the main area is titled "MatchMaker Suggestions" with a "Symbols Key" header. It features a table with columns for Name, Rank, Graph, Trend, Celeb, Book, and Fave This!. The first row of data in the table is partially visible.

Name	Rank	Graph	Trend	Celeb	Book	Fave This!
[Data]						

Video: <http://www.babynamewizard.com/baby-names-expert-upgrade-video-1>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Dynamic Query

- Dynamic Query
 - Probably best-known and one of most useful infovis techniques
- Database query: SQL
 - **Select** house-address
From atl-realty-db
Where price >= 200,000 **and** price <= 400,000
- Pros
 - Powerful, flexible
- Cons
 - Must learn language
 - Only shows exact matches
 - Don't know magnitude of results
 - No helpful context is shown
 - Reformulating to a new query can be slow

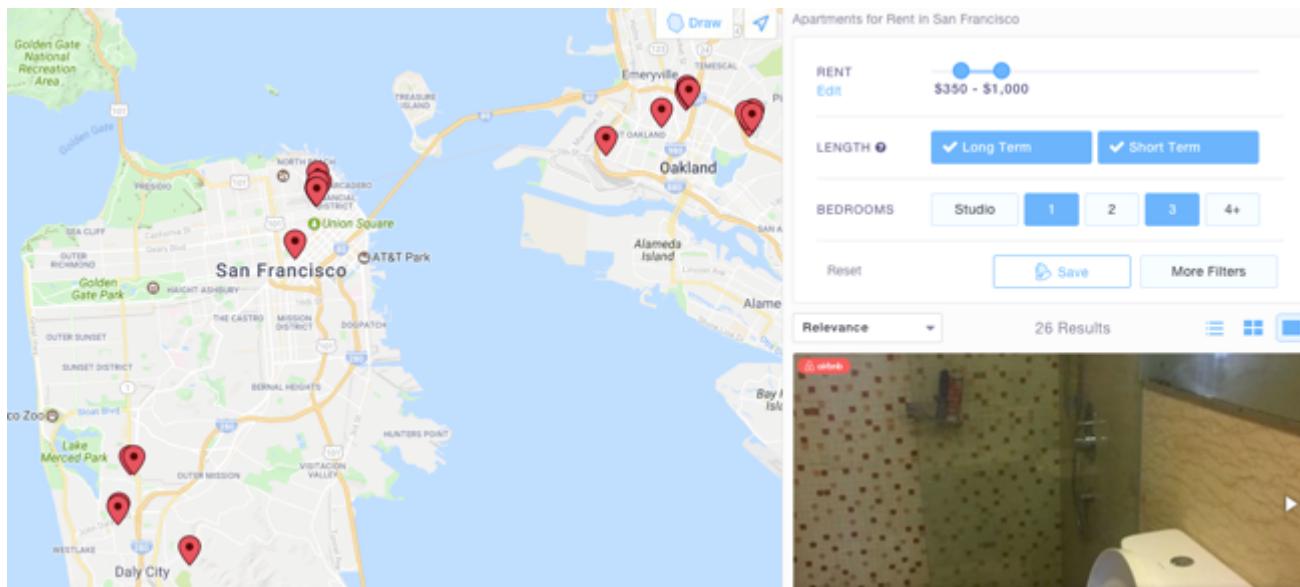
Dynamic Query

- Specifying a query brings immediate display of results
- Responsive interaction (< .1 sec) with data, concurrent presentation of solution
- “Fly through the data”, promote exploration, make it a much more “live” experience
 - Timesharing vs. batch
- There often simply isn’t one perfect response to a query
- Want to understand a set of tradeoffs and choose some “best” compromise

Example: <https://www.padmapper.com>

Query Control vs. Variable Type

- Binary nominal – Buttons
- Nominal with low cardinality - Radio buttons
- Ordinal, quantitative - Sliders



Dynamic Query: Pros and Cons

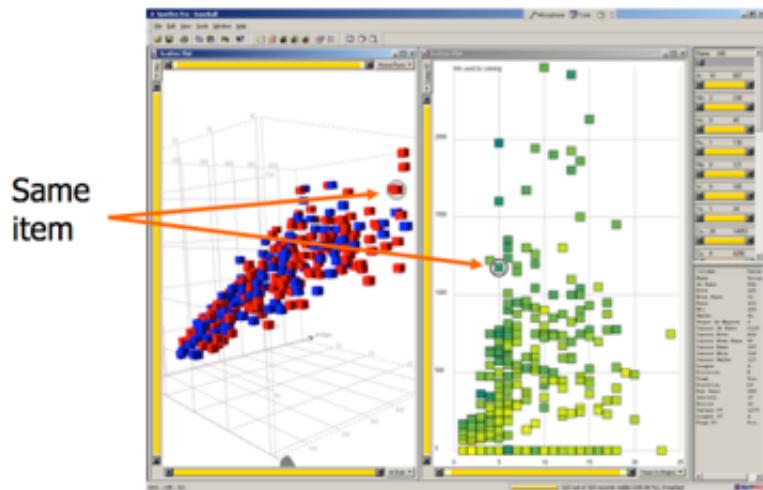
- Pros
 - Work is faster
 - Promote reversing, undo, exploration
 - Very natural interaction
 - Shows the data
- Cons
 - Operations are fundamentally conjunctive
 - Less flexible (can not formulate any boolean expression)
 - Controls are global in scope
 - Controls must be fixed in advance
 - Big data vs. real-time (more challenging)

7. Connect

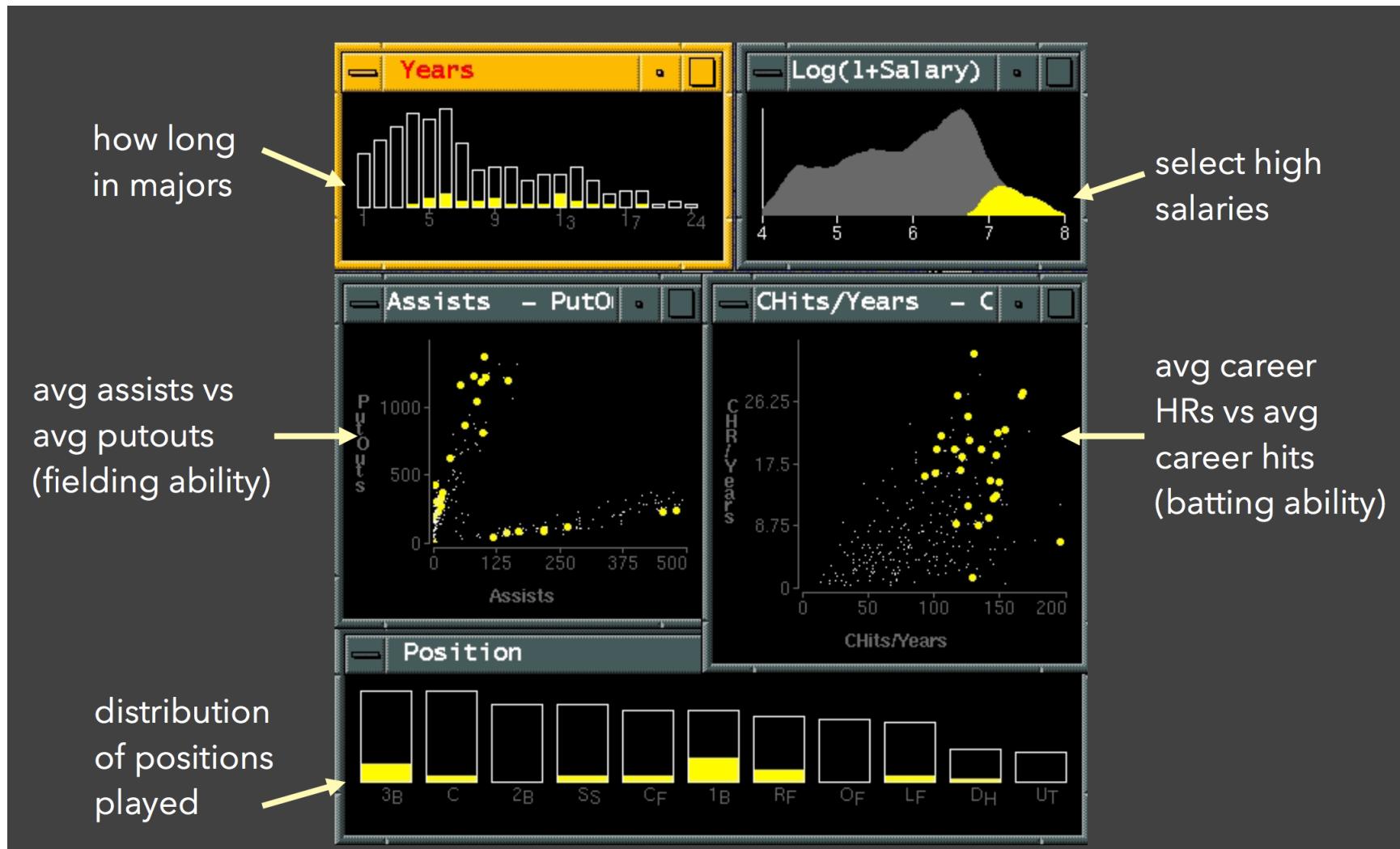
- “Show me related items”
- Highlight associations and relationships
- Show hidden data items that are relevant to a specified item
- Viewer may wish to
 - examine different attributes of a data case simultaneously
 - view data case under different perspectives or representations

Brushing

- Very common technique in Information Visualization
- Applies when you have multiple views of the same data
- Selecting or highlighting a case in one view generates highlighting the case in the other views



An Example: Baseball Statistics



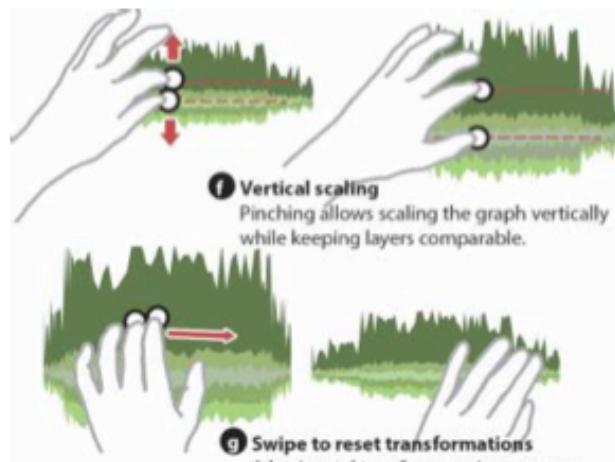
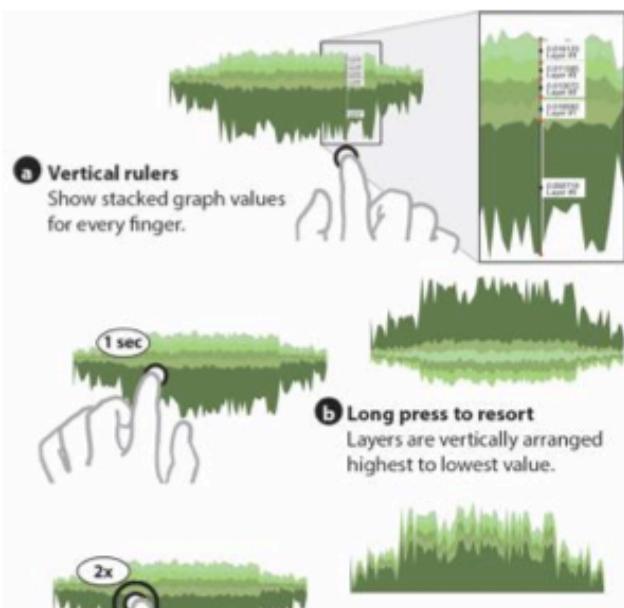
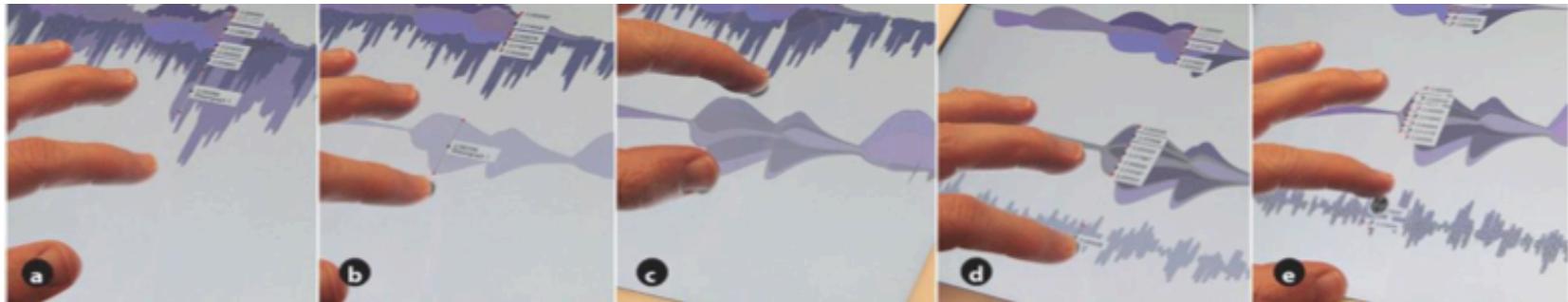
Interaction to Support Representation

- Interaction in many cases is vital to representation
 - Provides useful perspective
 - Many, many examples:
 - Parallel coords, InfoZoom, anything 3D
 - Necessary for clarifying representation
 - Dust & Magnet

Video: Dust & Magnet

[https://www.youtube.com/watch?
v=wLXwL38xek0](https://www.youtube.com/watch?v=wLXwL38xek0)

Other Interactions



Baur, Dominikus, Bongshin Lee, and Sheelagh Carpendale. "TouchWave: kinetic multi-touch manipulation for hierarchical stacked graphs." Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces. ACM, 2012.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Summary

- Interaction facilitates a dialog between the user and the visualization system
- Multiple views amplify importance of interaction
- Interaction often helps when you just can't show everything you want

Next Lecture

- Topic:
 - Evaluation



G53FIV: Fundamentals of Information Visualization

Lecture 10: Evaluation

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

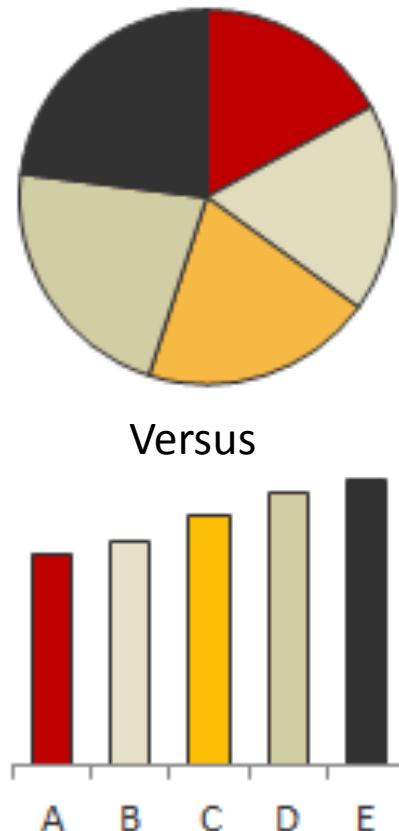
<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Overview

- Evaluation Methodologies
 - Controlled experiments
 - Subjective assessments
- Examples of evaluation

How do We Evaluate Visualizations?

- How do we evaluate visualizations?
 - Usability vs. Utility
- What evaluation techniques should we use?
- What do we measure?
 - What data do we gather?
 - What metrics do we use?



Evaluating Information Visualization in General

- Very difficult to compare “apples to apples”
 - Hard to compare System A to System B
 - Different tools were built to address different user tasks
- UI can heavily influence utility and value of visualization technique
- Utility vs. Aesthetics

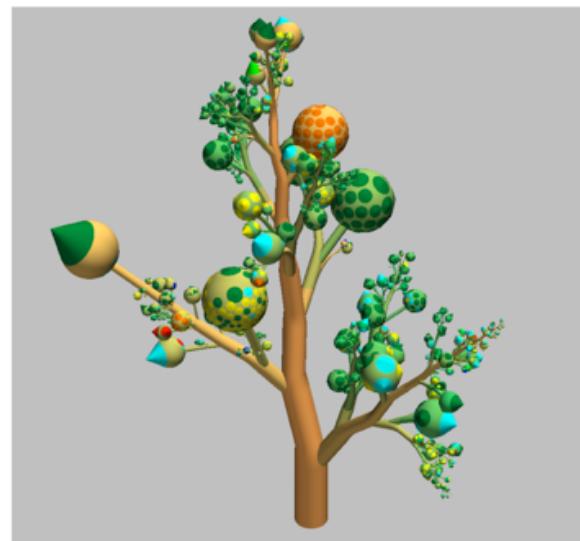


Figure 5: Botanic visualization contents of a hard disk [10, 27]. Useful or just a nice picture?

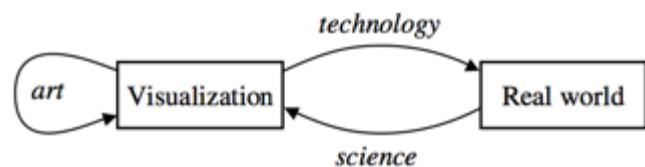


Figure 6: Views on visualization

Carpendale '08

- Challenges in information visualization evaluation
- Choosing an evaluation approach

Evaluating Information Visualizations

Sheelagh Carpendale

Department of Computer Science, University of Calgary,
2500 University Dr. NW, Calgary, AB, Canada T2N 1N4
sheelagh@ucalgary.ca

1 Introduction

Information visualization research is becoming more established, and as a result, it is becoming increasingly important that research in this field is validated. With the general increase in information visualization research there has also been an increase, albeit disproportionately small, in the amount of empirical work directly focused on information visualization. The purpose of this paper is to increase awareness of empirical research in general, of its relationship to information visualization in particular; to emphasize its importance; and to encourage thoughtful application of a greater variety of evaluative research methodologies in information visualization.

One reason that it may be important to discuss the evaluation of information visualization, in general, is that it has been suggested that current evaluations are not convincing enough to encourage widespread adoption of information visualization tools [57]. Reasons given include that information visualizations are often evaluated using small datasets, with university student participants, and using simple tasks. To encourage interest by potential adopters, information visualizations need to be tested with real users, real tasks, and also with large and complex datasets. For instance, it is not sufficient to know that an information visualization is usable with 100 data items if 20,000 is more likely to be the real-world case. Running evaluations with full data sets, domain specific tasks, and domain experts as participants will help develop much more concrete and realistic evidence of the effectiveness of a given information visualization. However, choosing such a realistic setting will make it difficult to get a large enough participant sample, to control for extraneous variables, or to get precise measurements. This makes it difficult to make definite statements or generalize from the results. Rather than looking to a single methodology to provide an answer, it will probably will take a variety of evaluative methodologies that together may start to approach the kind of answers sought.

The paper is organized as follows. Section 2 discusses the challenges in evaluating information visualizations. Section 3 outlines different types of evaluations and discusses the advantages and disadvantages of different empirical methodologies and the trade-offs among them. Section 4 focuses on empirical laboratory experiments and the generation of quantitative results. Section 5 discusses qualitative approaches and the different kinds of advantages offered by pursuing this type of empirical research. Section 6 concludes the paper.

Evaluation Approaches

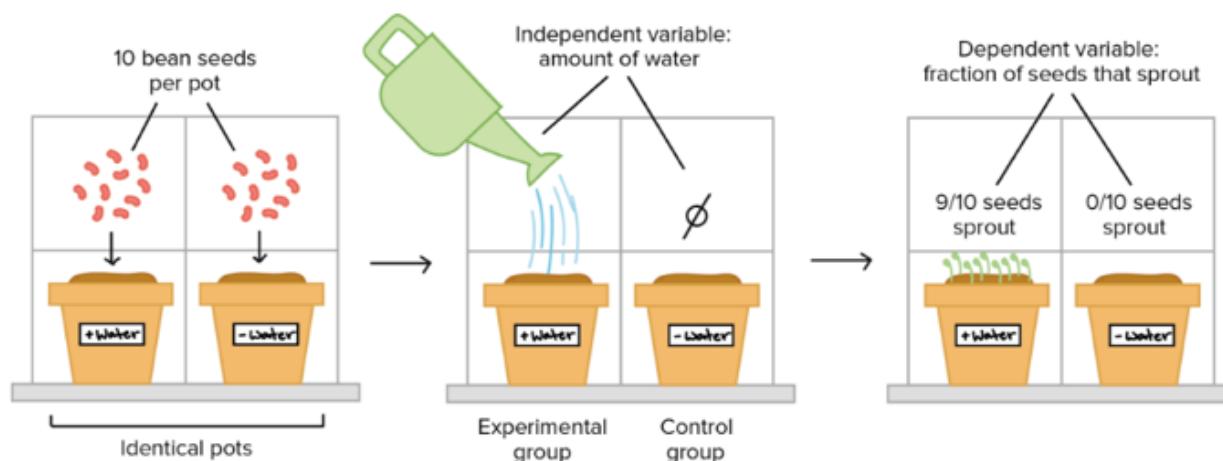
- Many different Forms
 - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, ...
- Two popular methodologies
 - Controlled experiments (Quantitative)
 - Subjective assessments (Qualitative)

Quantitative Methods

Quantitative Methods: Controlled Experiments

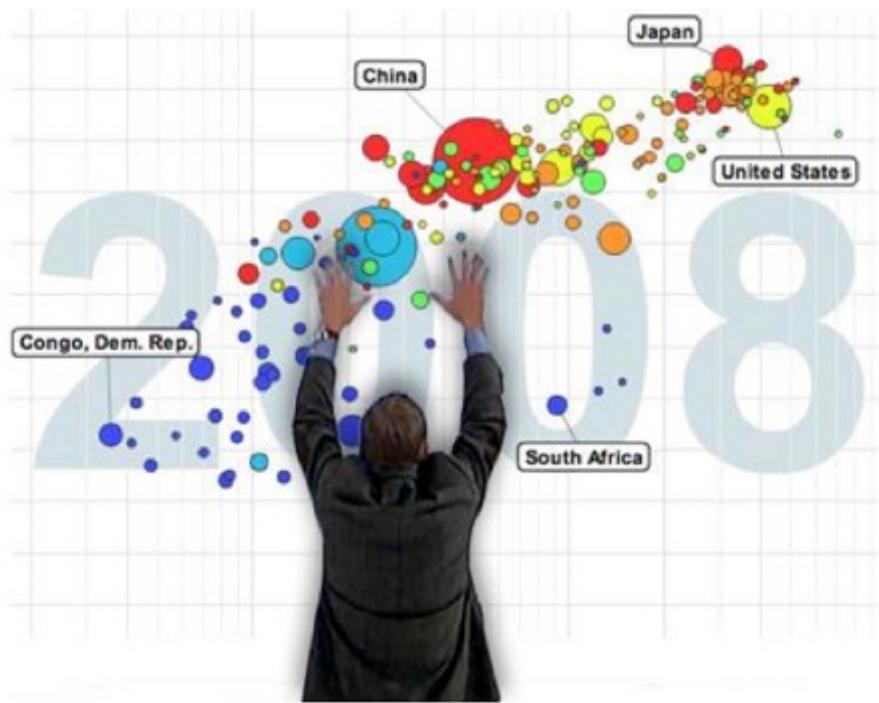
- Good for measuring performance or comparing multiple techniques
- What do we measure?
 - Performance, time, errors,

...



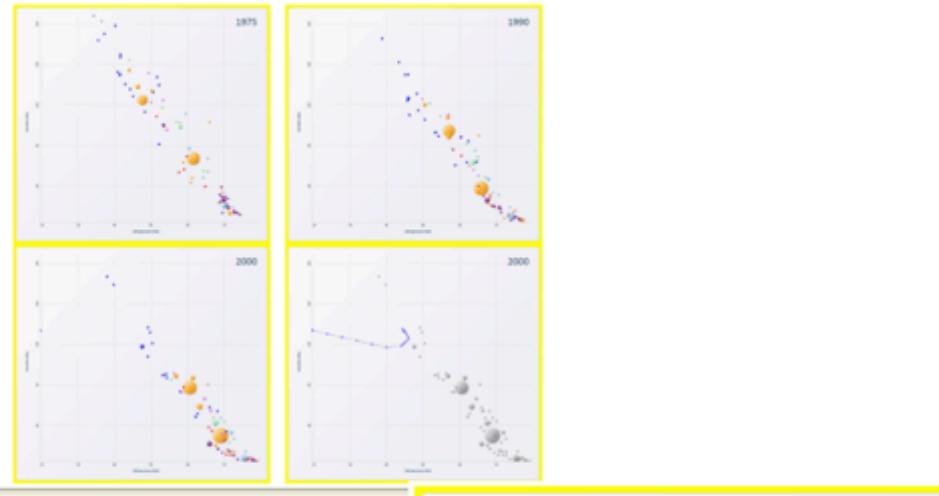
An Example: Controlled Experiment

- Run an experiment to evaluate three visualization strategies
 - Animation
 - Small multiples
 - Traces
- Especially interested in examining whether animated bubble charts are beneficial for analysis and presentation



*Do you remember Hans Rosling's TED talk?
(Lecture 2)*

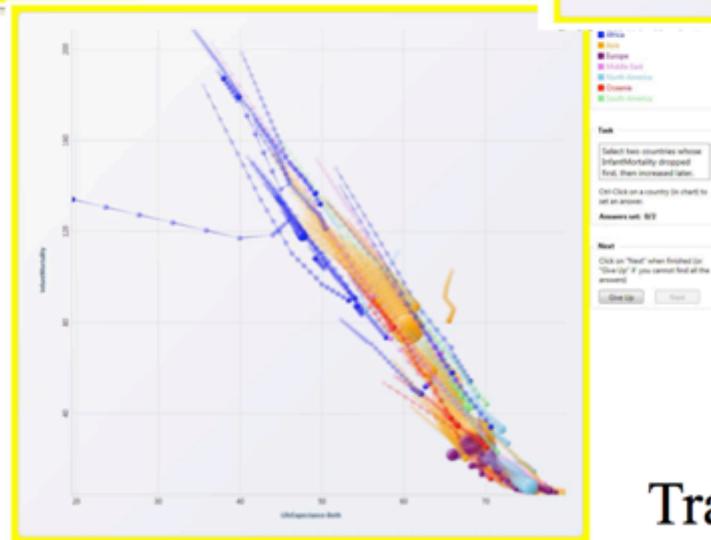
Three Visualizations



Animation

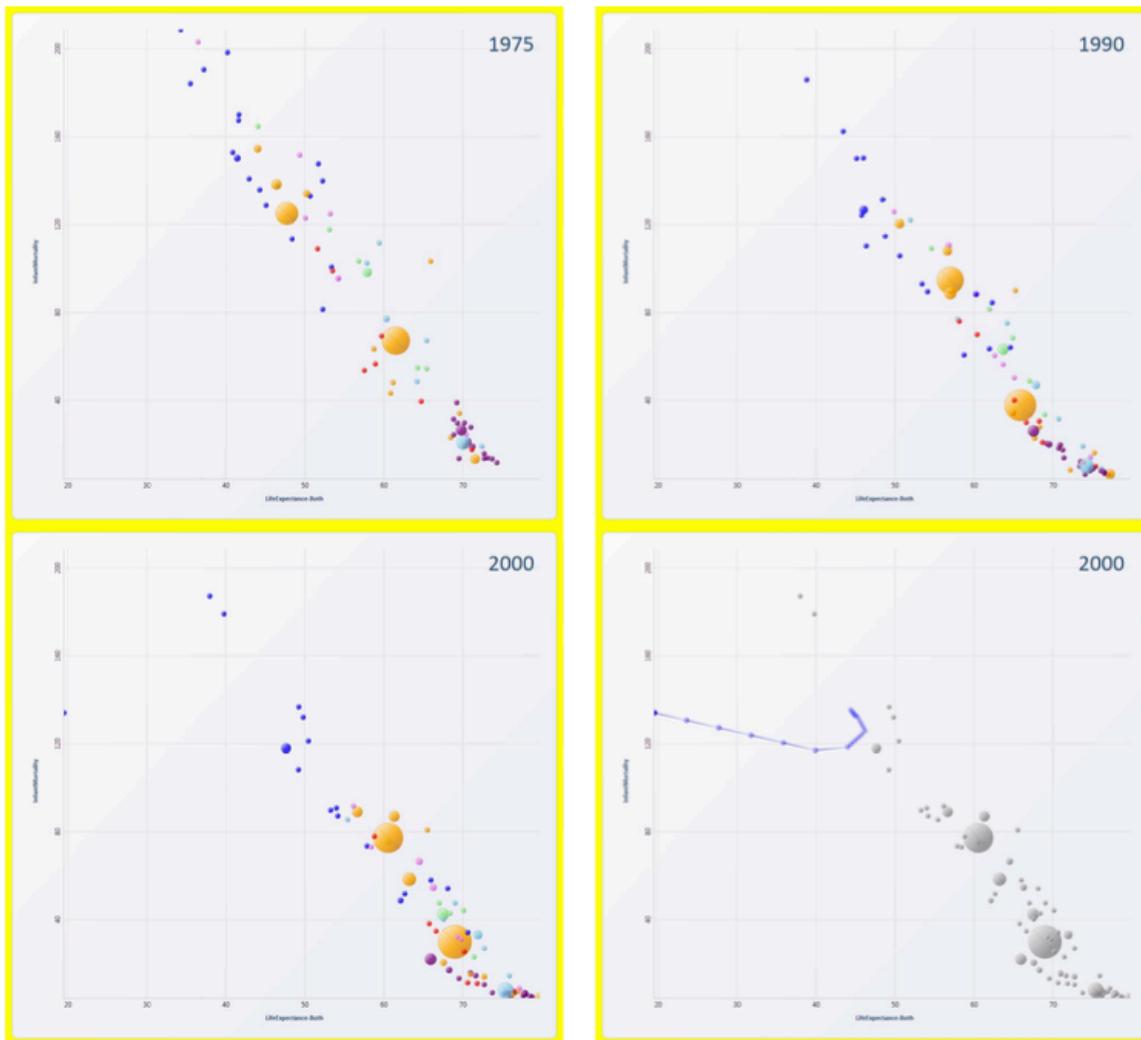


Small multiples

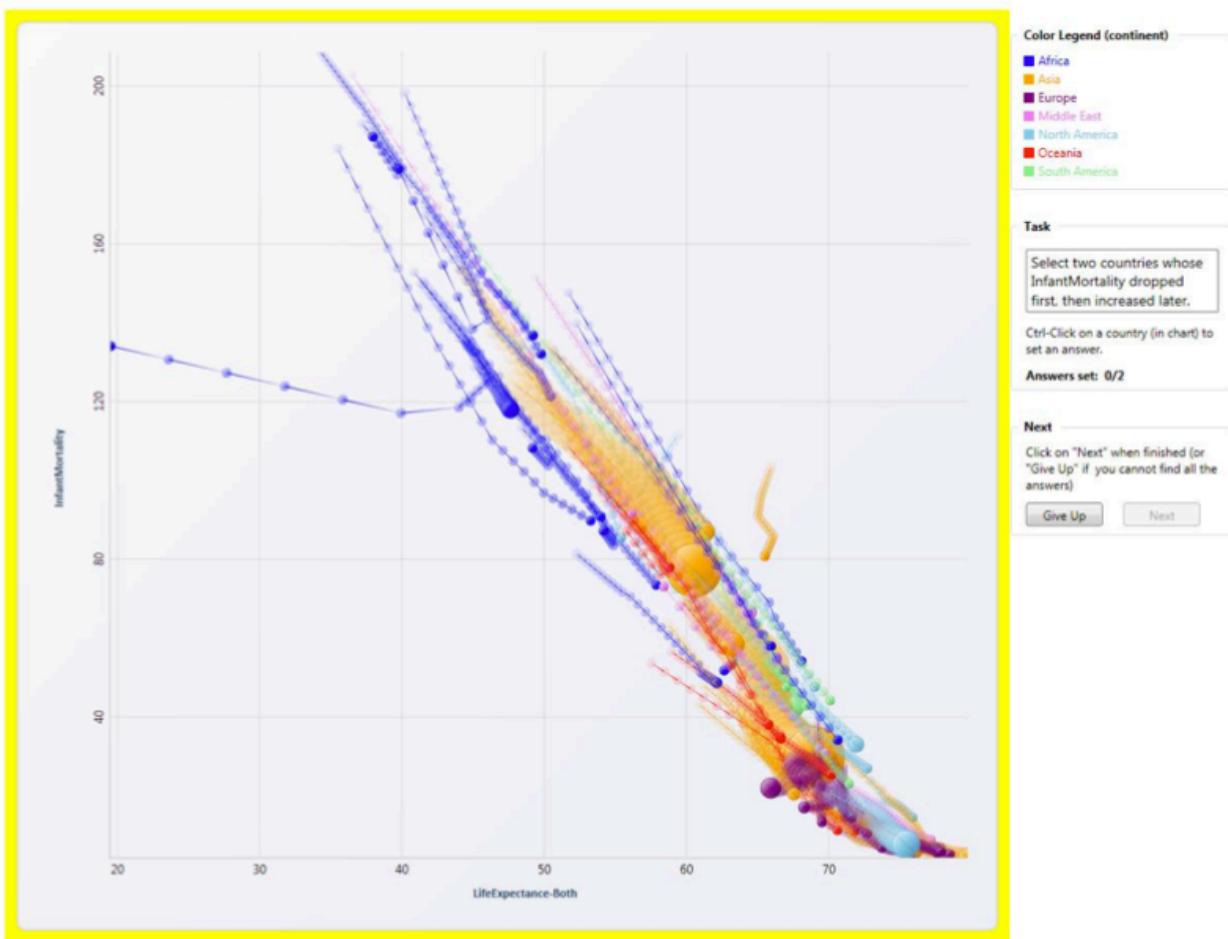


Traces

Three Visualizations: Animation



Three Visualizations: Traces



Three Visualizations: Small Multiples



Experimental Design

- 3 (visualization types) x 2 (data size: small & large) x 2 (presentation vs. analysis)
 - Presentation vs analysis – between subjects
 - Others – within subjects
- Animation has 10-second default time, but user could control time slider

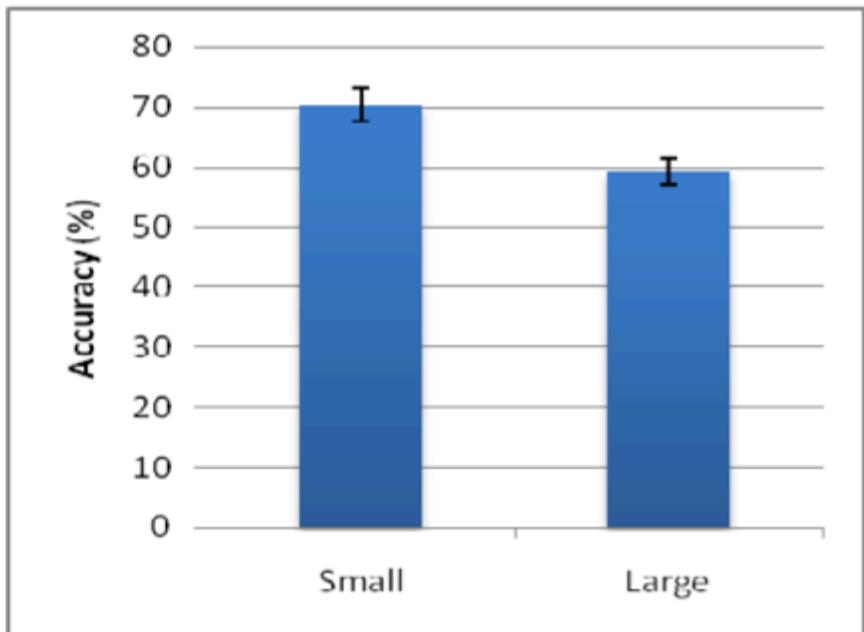
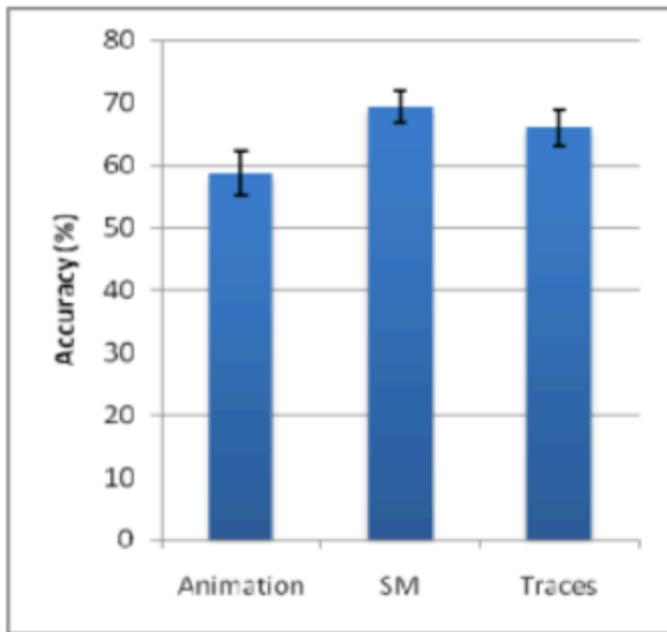
Experimental Design

- Data
 - Union Nations Common data about countries
- Tasks
 - 24 tasks, 1-3 requires answers per
 - Select 3 countries whose rate of energy consumption was faster than their rate of GDP per capita growth
 - Select 2 countries with significant decreases in energy consumption
 - Which continent had the least changes in GDP per capita

Conditions

- Analysis – straightforward, interactive
- Presentation
 - 6 participants at a time
 - Presenter described a trend relevant to task, but different
 - No interaction with system
 - In animation condition, participants saw last frame of animation (no interaction)

Results: Accuracy



- Summary
 - Small multiple is better than animation
 - Small data size is more accurate than large

Results: Speed

- Presentation
 - Animation faster than small multiples & traces 15.8 secs vs. 25.3 secs vs. 27.8 secs.
- Analysis
 - Animation slower than small multiples & traces 83.1 secs. vs. 45.69 secs. vs. 55.0 secs

Results: Subjective Ratings

Table 3. Average ratings for seven questions for each visualization.

* indicates significant differences ($p < .05$).

	Animation	SM	Traces
Q1. The visualization was helpful to me in answering the questions.	4.6 *Traces	4.2	4.1
Q2. For the smaller dataset, I found the tasks easy using this visualization.	4.6 *SM	4.2	4.5
Q3. For the larger dataset, I found the tasks easy using this visualization.	2.6	3.4 *Traces	2.3
Q4. I enjoyed using this visualization.	4.3 *SM *Traces	3.7	3.5
Q5. I found this visualization exciting.	4.3 *SM *Traces	3.1	3.0
Q6. For the smaller dataset, I found the screen too cluttered.	1.8	1.5	2.0
Q7. For the larger dataset, I found the screen too cluttered.	4.4	2.8 *Animation *Traces	4.7

Table 4. Average ratings for a few general questions.

	Presentation	Analysis	Overall
G1. I found the Traces view enjoyable.	3.8	2.9	3.4
G3. I found the Small Multiples view enjoyable.	4.1	3.4	3.7
G5. I found the Animation view enjoyable.	4.6	5.0	4.8
G7. The animation went too fast for me.	3.2	2.8	3.0
G8. The animation went too slow for me.	1.6	1.3	1.4
G9. I lost track of some data points as they moved.	4.9	4.6	4.8

Presentation, small: Animation (9) > SM (6) > Traces (3)

Presentation, large: Traces (8) > SM (6) > Animation (4)

Analysis, small: Animation (7) > SM (6) > Traces (5)

Analysis, large: Animation (8) > SM (6) > Traces (4)

Likert: 0-strongly disagree, 6-strongly agree

Summary of Results

- People rated animation more fun, but small multiples was more effective.
- As data grows, accuracy becomes an issue
 - Traces & animation get cluttered
 - Small multiple gets tiny
- Animation:
 - “fun”, “exciting”, “emotionally touching”
 - Confusing, “the dots flew everywhere”

Controlled Experiments at Large

- Online A/B testing in the commercial world
 - <https://www.coursera.org/learn/ui-testing/lecture/pMhKt/industry-practice-massive-a-b-testing-interview-with-ronny-kohavi>
- A very widely used methods in evaluating developed new systems/algorithms

Quantitative Challenges

- Conclusion Validity
 - Is there a relationship?
- Internal Validity
 - Is the relationship causal?
- Construct Validity
 - Can we generalize to the constructs (ideas) the study is based on?
- External Validity
 - Can we generalize the study results to other people/places/times?
- Ecological Validity
 - Does the experimental situation reflect the type of environment in which the results will be applied?

Qualitative Methods

Qualitative Methods

- Types
 - Nested methods
 - Experimenter observation, think-aloud protocol, collecting participant opinions
 - Inspection evaluation methods
 - Heuristics to judge
- Observational context
 - In situ, laboratory, participatory
 - Contextual interviews is important

Subjective Assessments

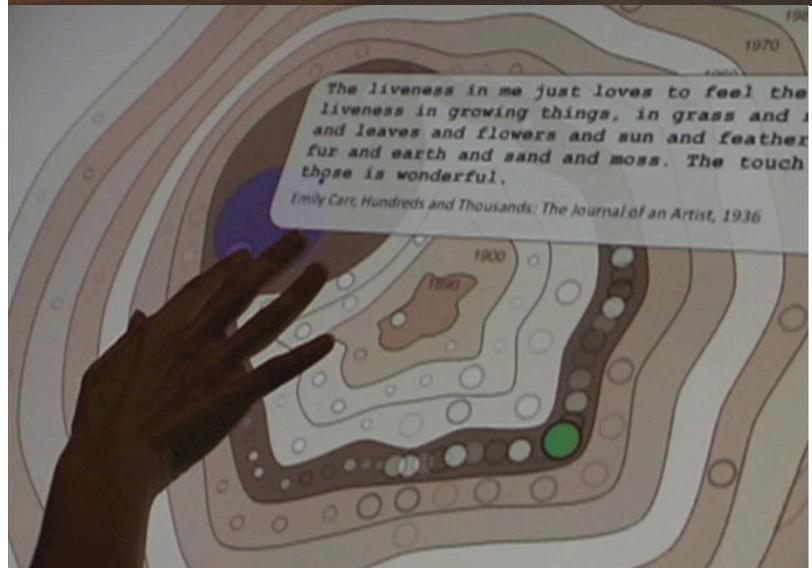
- Learn people's subjective views on tool
 - Was it enjoyable, confusing, fun, difficult, ...?
- This kind of personal judgment strongly influence use and adoption, sometimes even overcoming performance deficits

- Pros and Cons?
 - Compared to controlled experiments

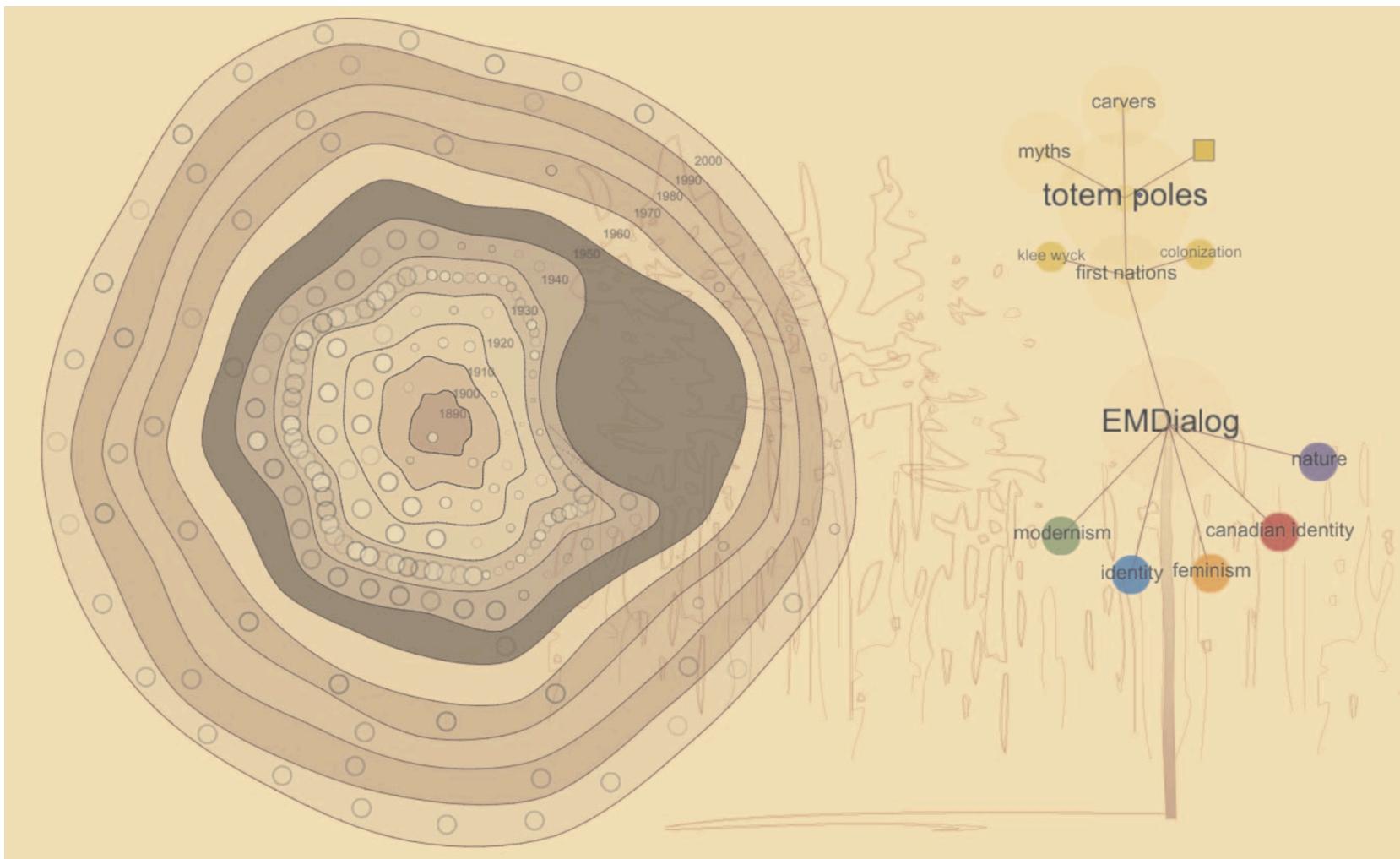


An Example: Subjective Assessments

- Evaluating a newly developed visualization system (EMDialog) at the museum (Emily Carr exhibit)
- Discourse visualization
 - Time
 - Context



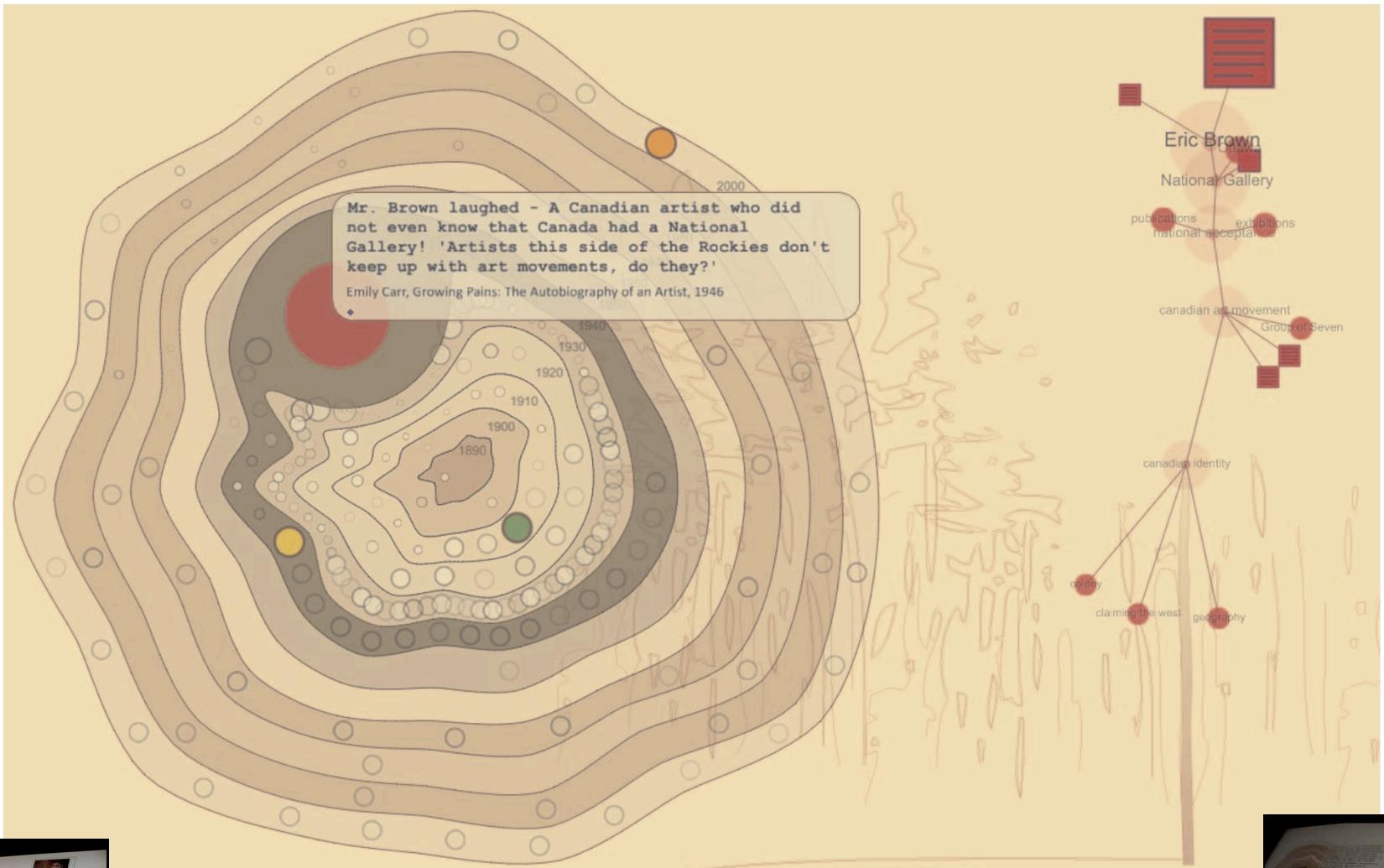
Visualization



Cut section

Tree section

Visualization



Cut section

Tree section

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~zhou/>)

Evaluation Goals

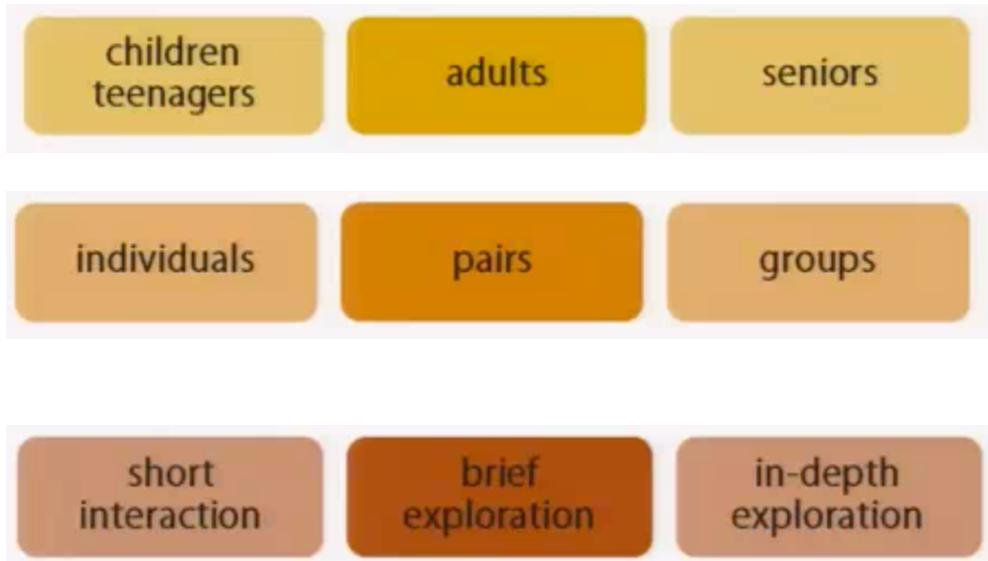
- Approach
 - How would people approach EMDialog? What would draw them toward the installation?
- Exploration techniques
 - How would they explore the information visualizations?
- Acceptance
 - What would visitors generally think of this type of information presentation in the museum context?

Experimental Design

- Emily Carr exhibition floor at the Glenbow Museum for around a month
- Open observation
 - Non-intrusive observation
 - Field notes
- Open-ended Questionnaires (voluntarily)
 - What participants liked or disliked about the installation

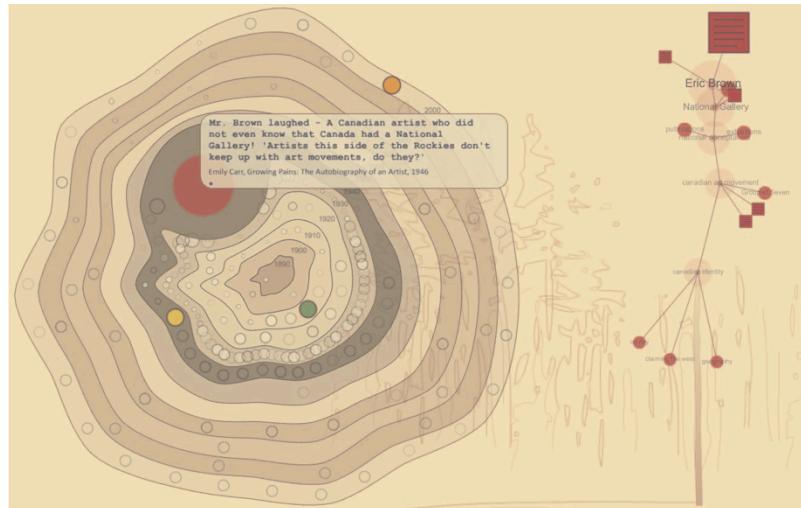
Approach

- Age
- Number
- Interaction time
- Motivation
 - Physical setup
 - Interaction of other people



Exploration Styles

- All information at once
- Broad exploration



- Structured
- Interest-based exploration

- Switch between two sections
 - Visual cues
 - Missing feedback

Performance

- Visible interaction
 - Evokes curiosity
 - Teaches interaction techniques
- Awareness of being observed
 - Performing in front of strangers
 - Taking over control



I like watching other people interacting!

I am uncertain about the performance aspect, I'm kind of an introvert!

I felt guilty interacting with the display not knowing whether or not someone was in the middle of reading the projected screen!

Performance: mixed visitor reaction



- [EMDialog] allowed me to **put Carr's work into context.**
- Enhanced the museum's experience by linking chronology and concept
- It allowed me to **focus on one aspect / period of her work.**



- It **took me a while to get the idea** (and resist fatigue after spending two hours in the exhibit) but **it quickly engaged me and was really neat and fun to use.**



- **Too much reading / not enough pictures.**
- **Totally confusing**

Take-Away for EMDialog

- Appealing information representation
- Interactivity
 - Short- and long-term exploration
 - Collaborative information exploration
 - Various exploration styles
- Leave traces in the visualization

Qualitative Challenges

- Sample sizes
- Subjectivity
- Analyzing qualitative data

Meta-evaluate Evaluation Approaches

- Desirable features
 - Generalizability
 - Precision
 - Realism

Methodology vs. Desirable Features

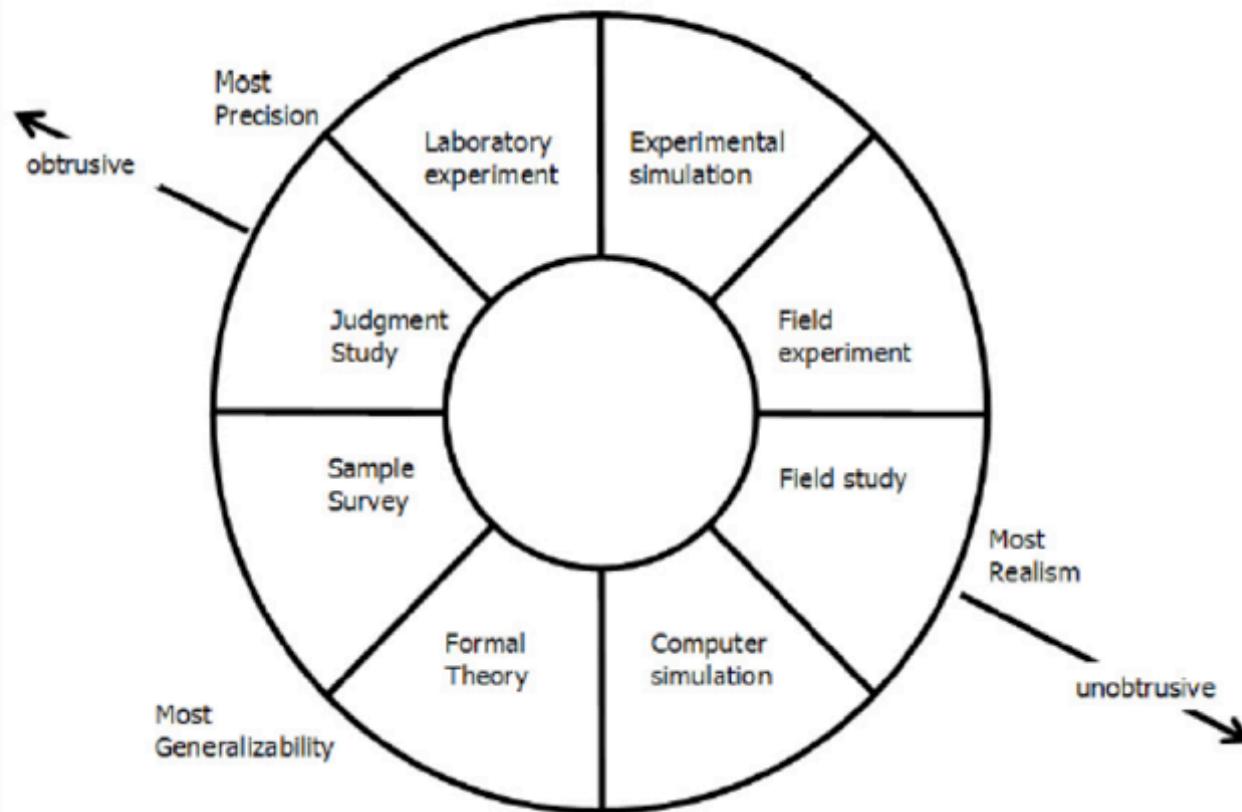


Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

Summary

Research Aspect	Quantitative	Qualitative
Common Purpose	Test Hypotheses or Specific Research Questions	Discover Ideas, used in Exploratory Research with General Research Objects
Approach	Measure and Test	Observe and Interpret
Data Collection Approach	Structured Response Categories Provided	Unstructured, Free-Form
Research Independence	Researcher Uninvolved Observer. Results Are Objective.	Researcher Is Intimately Involved. Results Are Subjective.
Samples	Large Samples to Produce Generalizable Results	Small Samples – Often in Natural Settings
Most Often Used	Descriptive and Causal Research Designs	Exploratory Research Designs

Crowd-Based Evaluation



- e.g. Amazon Mechanical Turk
- Emerging Method that enables scale
- Lots of issues

How to Conduct Evaluation Studies

- You can learn more about those in the following courses.
 - (G52HCI) Introduction to Human Computer Interaction
 - (G54HCI) Individual Project: Human-Computer Interaction

Summary

- Why do evaluation of InfoVis systems?
 - We need to be sure that new techniques are really better than old ones
 - We need to know the strengths and weaknesses of each tool; know when to use which tool
- Challenges
 - There are no standard benchmark tests or methodologies to help guide researchers
 - Defining the tasks is crucial
 - What about individual differences?
 - Controlled experiments vs. subjective assessments

Next Lecture

- Topic:
 - Text and Document
 - The next Monday (4 Mar)
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus



G53FIV: Fundamentals of Information Visualization

Lecture 11: Visualizing Text and Documents

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

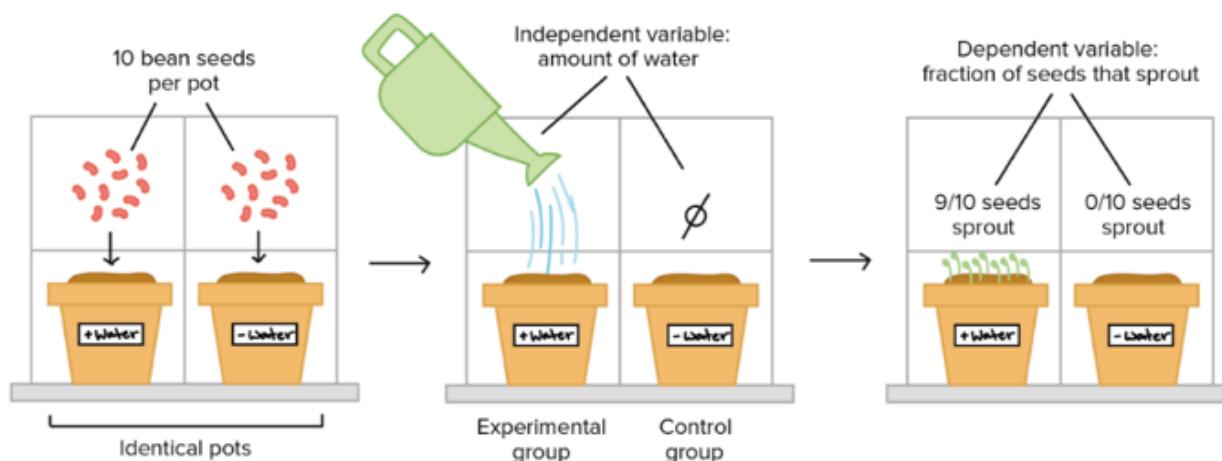
Last Lecture

Evaluation

Quantitative Methods: Controlled Experiments

- Good for measuring performance or comparing multiple techniques
- What do we measure?
 - Performance, time, errors,

...



An Example: Controlled Experiment

- Run an experiment to evaluate three visualization strategies
 - Animation
 - Small multiples
 - Traces
- Especially interested in examining whether animated bubble charts are beneficial for analysis and presentation



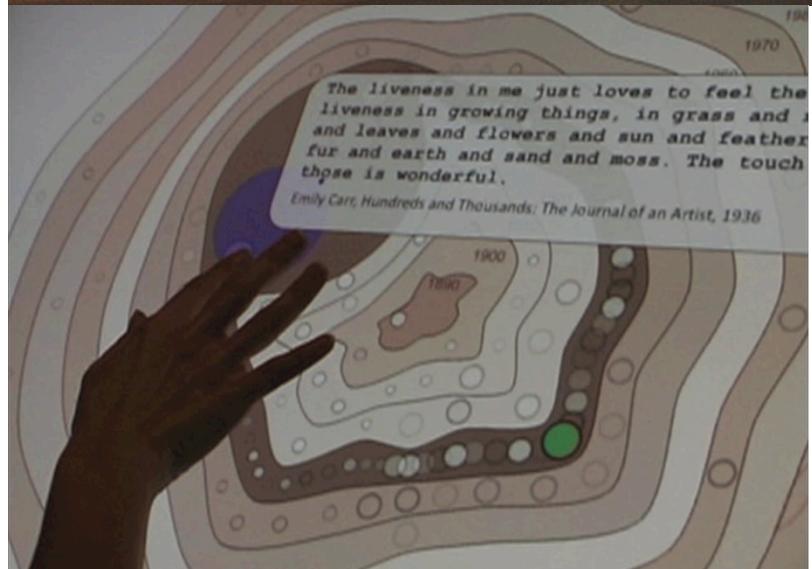
*Do you remember Hans Rosling's TED talk?
(Lecture 2)*

Qualitative Methods

- Types
 - Nested methods
 - Experimenter observation, think-aloud protocol, collecting participant opinions
 - Inspection evaluation methods
 - Heuristics to judge
- Observational context
 - In situ, laboratory, participatory
 - Contextual interviews is important

An Example: Subjective Assessments

- Evaluating a newly developed visualization system (EMDialog) at the museum (Emily Carr exhibit)
- Discourse visualization
 - Time
 - Context



Methodology vs. Desirable Features

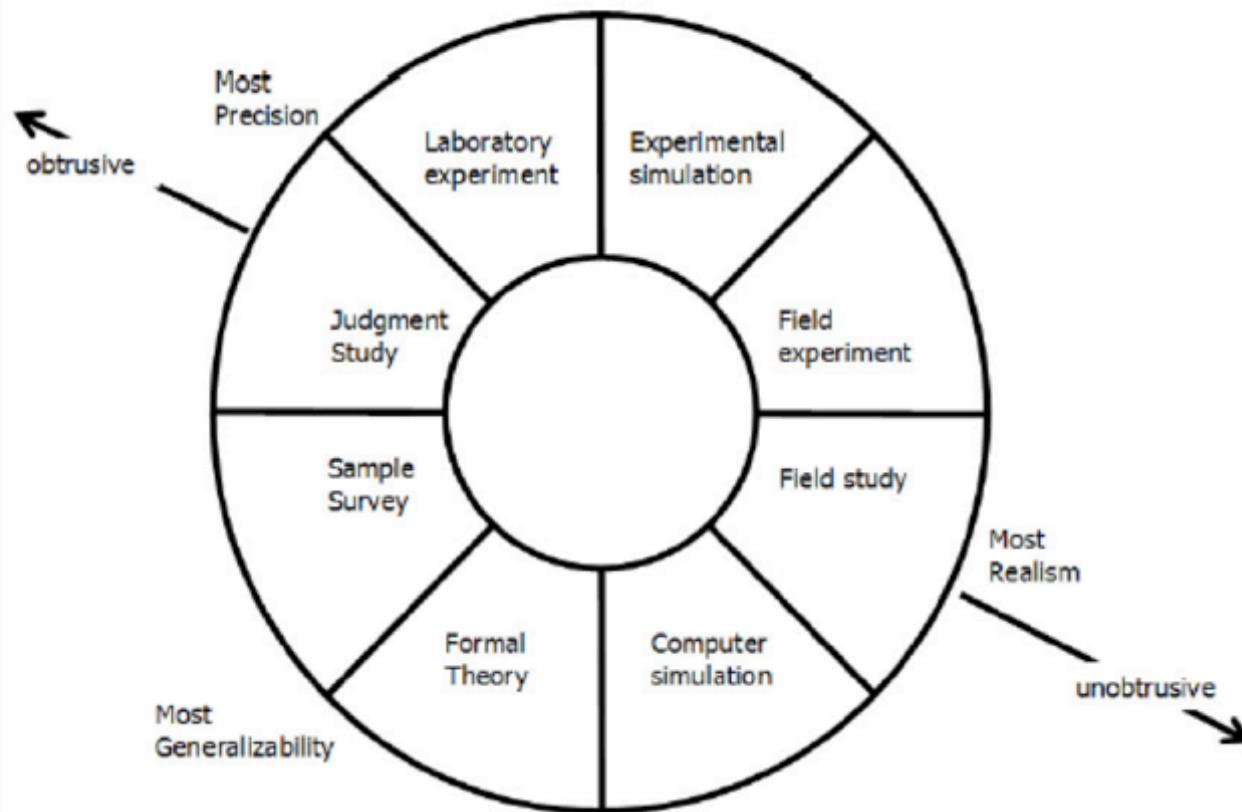


Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

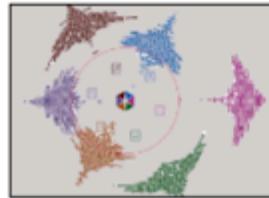
Overview



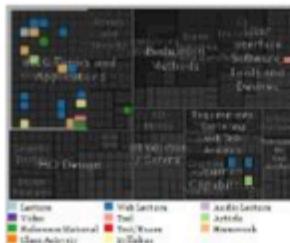
Visualizing text
Showing words,
phrases, and
sentences



- Content
- Context
- Relationship to others



Visualization for IR
Helping search



Why Visualize Text?

- What can information visualization provide to help users in understanding and gathering information from text and document collections?
- **Understanding**
 - get the “gist” of a document
- **Grouping**
 - cluster for overview or classification
- **Comparison**
 - compare document collections, or inspect evolution of collection over time
- **Correlation**
 - compare patterns in text to those in other data, e.g., correlate with social network

Challenges

- High Dimensionality
 - Where possible use text to represent text...
 - ... which terms are the most descriptive?
- Context and Semantics
 - Provide relevant context to aid understanding.
 - Show (or provide access to) the source text.
- Modeling Abstraction
 - Determine your analysis task.
 - Understand abstraction of your language models.
 - Match analysis task with appropriate tools and models.

Visualizing Text

How do we represent the words, phrases, and sentences in a document or set of documents?



Visualizing text
Showing words, phrases, and sentences



An Example

- Health care speech transcripts
 - Clinton in 1993
 - Obama in 2009
- What questions might you want to answer?
- What visualizations might help?

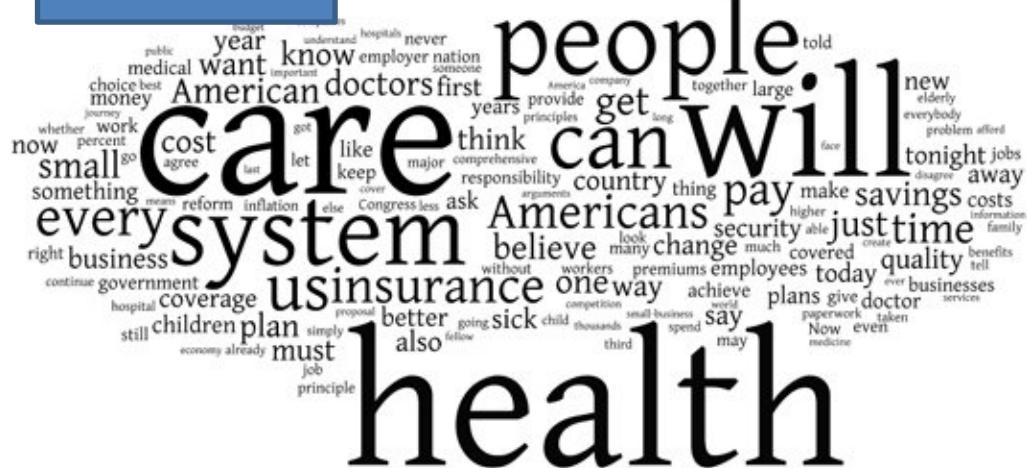
Bill Clinton on Health Care, 1993

Now we are in a time of profound change and opportunity – the end of the cold war, the information age, and the global economy have brought us both opportunity and hope and strife and uncertainty. Our purpose in this dynamic age must be to change – to make change our friend and not our enemy.

To achieve that goal, we must face all our challenges with confidence, with faith and with discipline, whether we're reducing the deficit, creating tomorrow's jobs and training our people to fill them, converting from a high-tech defense to a high-tech domestic economy, expanding trade, reinventing government, making our streets safer, or rewarding work over idleness, all of these challenges require us to change.

Tag Clouds: Word Count

Clinton 1993



Obama 2009

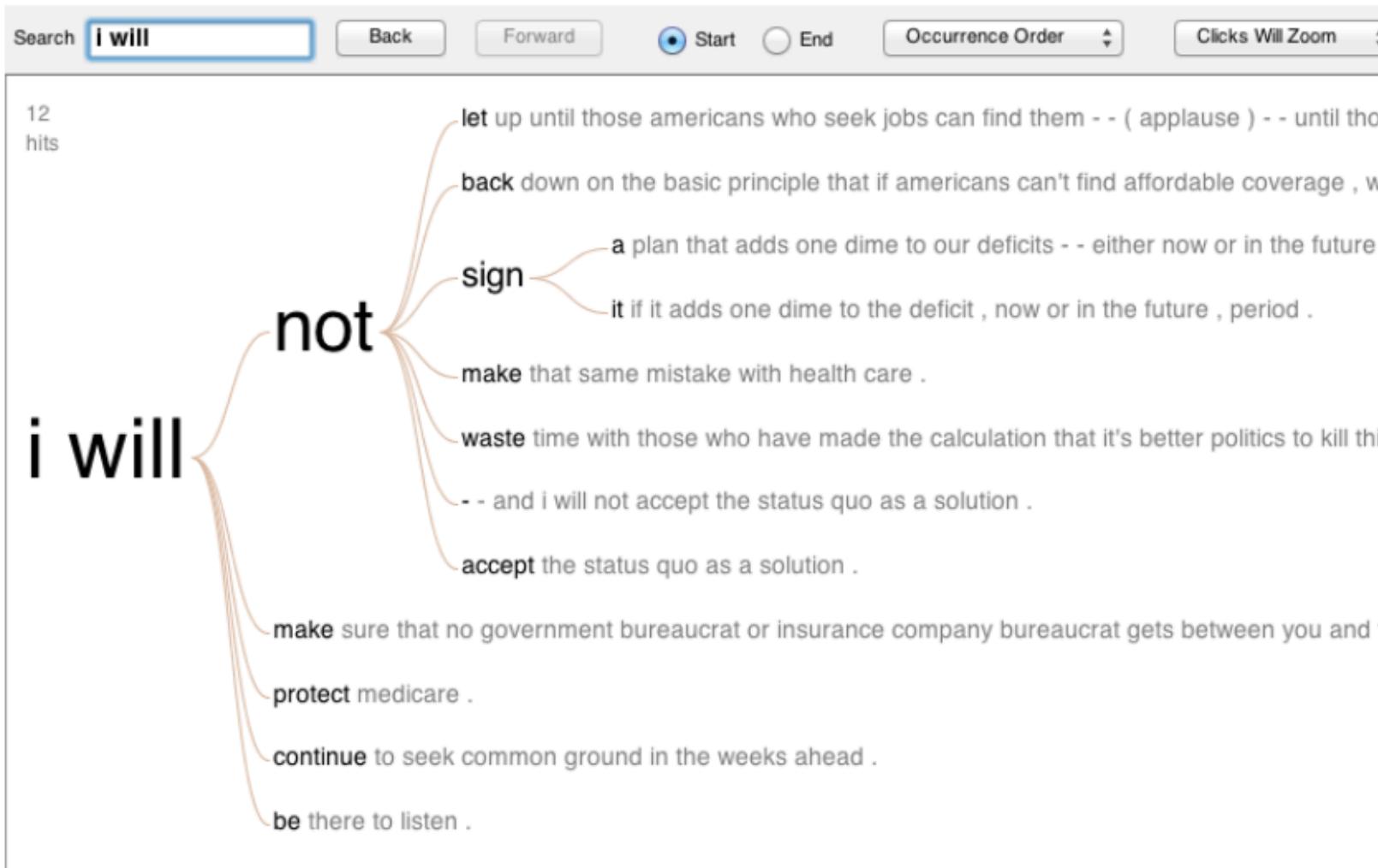


<https://economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Word Tree: Word Sequences

Visualizations : Word Tree President Obama's Address to Congress on Health Care



Language Model

- Many text visualizations do not represent the text directly. They represent the output of a language model (word counts, word sequences, etc.).
- Can you interpret the visualization? How well does it convey the properties of the model?
- Do you trust the model? How does the model enable us to reason about the text?

Challenges

- **High Dimensionality**
 - Where possible use text to represent text...
 - ... which terms are the most descriptive?
- **Context and Semantics**
 - Provide relevant context to aid understanding.
 - Show (or provide access to) the source text.
- **Modeling Abstraction**
 - Determine your analysis task.
 - Understand abstraction of your language models.
 - Match analysis task with appropriate tools and models.

Words as nominal data?

- High dimensional (10,000+)
- Words have meanings and relations
 - Correlations: Hong Kong, San Francisco, Bay Area
 - Order: April, February, January, June, March, May
 - Membership: Tennis, Running, Swimming, Hiking, Piano
 - Hierarchy, antonyms & synonyms, entities, ...

Text Processing Pipeline

- Tokenization
 - Segment text into terms.
 - Remove stop words? [a](#), [an](#), [the](#), [of](#), [to](#), [be](#)
 - Numbers and symbols? [#gocard](#), [@nottinghamforestfbball](#)
 - Entities? [Nottingham](#), [Trump](#).
- Stemming
 - Group together different forms of a word.
 - Porter stemmer? [visualization\(s\)](#), [visualize\(s\)](#), [visually](#) -> [visual](#)
 - Lemmatization? [goes](#), [went](#), [gone](#) -> [go](#)
- Ordered list of terms

Content

Bag of Words Model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
 - Each dimension corresponds to a term (10,000+)
 - Each value represents the relevance
 - For example, simple term counts
- Aggregate into a document-term matrix
 - Document vector space model

Document Term Matrix

- Each document is a vector of term weights
- Simplest weighting is to just count occurrences

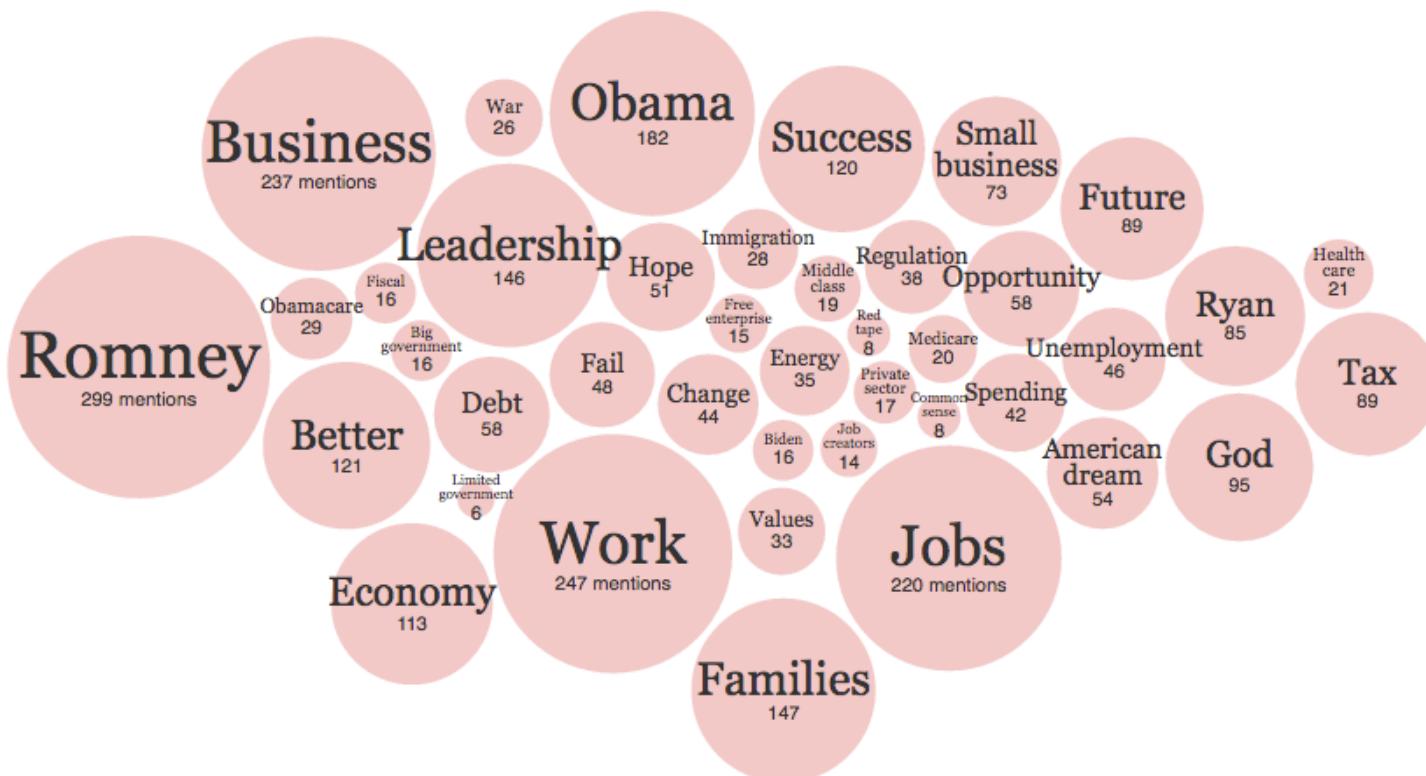
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Word Counts

At the Republican Convention, the Words Being Used

A look at how often speakers at the Republican National Convention have used certain words and phrases so far, based on an analysis of transcripts from the Federal News Service.

Add word or phrase



Word Counts



<http://www.wordcount.org>

Tag/Word Clouds



Wordle Tag Clouds

- Tightly packed words, sometimes vertical or diagonal
- Word size is linearly correlated with frequency (typically square root in cloud)
- Multiple color palettes
- User gets some control

Viegas, Wattenberg, &
Feinberg TVCG (InfoVis) '09

Tag Clouds: Pros and Cons

- Strengths
 - Can help with gisting and initial query formation.
- Weaknesses
 - Sub-optimal visual encoding (size vs. position)
 - Inaccurate size encoding (long words are bigger)
 - May not facilitate comparison (unstable layout)
 - Term frequency may not be meaningful
 - Does not show the structure of the text

Descriptive Words

- Given a text, what are the best descriptive words?

Keyword Weighting

- Term Frequency
 - $tf_{td} = \text{count}(t) \text{ in } d$
 - Can take log frequency: $\log(1 + tf_{td})$
 - Can normalize to show proportion $(tf_{td} / \sum_t tf_{td})$
- TF.IDF: Term Freq by Inverse Document Freq
 - $tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$
 - df_t = # docs containing t;
 - N = # of docs

An Example

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}("this", d_1) = \frac{1}{5} = 0.2$$

$$\text{tf}("this", d_2) = \frac{1}{7} \approx 0.14$$

$$\text{idf}("this", D) = \log\left(\frac{2}{2}\right) = 0$$

$$\text{tfidf}("this", d_1) = 0.2 \times 0 = 0$$

$$\text{tfidf}("this", d_2) = 0.14 \times 0 = 0$$

$$\text{tf}("example", d_1) = \frac{0}{5} = 0$$

$$\text{tf}("example", d_2) = \frac{3}{7} \approx 0.429$$

$$\text{idf}("example", D) = \log\left(\frac{2}{1}\right) = 0.301$$

$$\text{tfidf}("example", d_1) = 0 \times 0.301 = 0$$

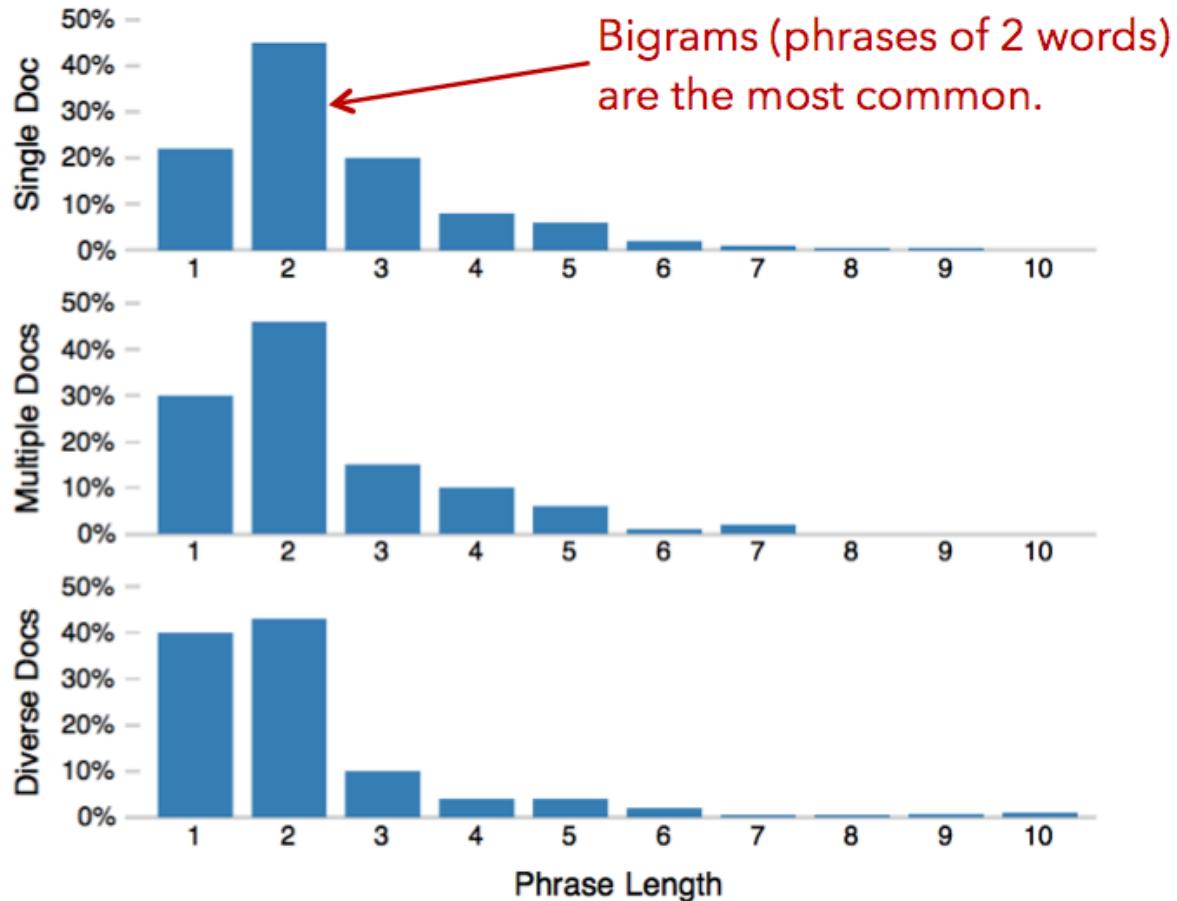
$$\text{tfidf}("example", d_2) = 0.429 \times 0.301 \approx 0.13$$

Limitations of Frequency Statistics

- Typically focus on unigrams (single terms)
- Often favors frequent (TF) or rare (IDF) terms
 - Not clear that these provide best description
- A “bag of words” ignores additional information
 - Grammar / part-of-speech
 - Position within document
 - Recognizable entities

How do people describe text?

- 69 subjects (graduate students) were asked to read and describe dissertation abstracts.



Word vs. Phrases

A fighter jet rain check

Story and video by [Chamila Jayaweera](#)

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out moreabout how the process works and watch the action unfold, click above to see the video story.



CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.



Word vs. Phrases

Word (e.g. TFIDF)

fighter

F/A

Hornet

Super

Boeing

-18

rain

St.

jet

Louis

15-inch-per-hour

douse

hangar

water-tight

Check

Baxter

sea-based

aircraft

Rich

seep

click

Navy

sure

Water

moisture

watch

enormous

stay

Key Phrase Extraction

Super Hornet

F/A -18

fighter jet

Boeing engineers

special process

rain check

electronics suite

Program experts

simulated rain

ultimate customers

enormous hangar

water-tight integrity

Rich Baxter

15-inch-per-hour rate

video story

aircraft

U.S. Navy fighter pilots

Super Hornet final assembly manager



Descriptive Phrases

- Understand the limitations of your language model.
 - Bag of words:
 - Easy to compute
 - Single words
 - Loss of word ordering
- Select appropriate model and visualization
 - Generate longer, more meaningful phrases
 - Adjective-noun word pairs for reviews
 - Show keyphrases within source text

(Optional Reading) Automatic Keyphrase Extraction: A Survey of the State of the Art: <http://www.aclweb.org/anthology/P/P14/P14-1119.xhtml>

<http://bdewilde.github.io/blog/2014/09/23/intro-to-automatic-keyphrase-extraction/>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Context

Challenges

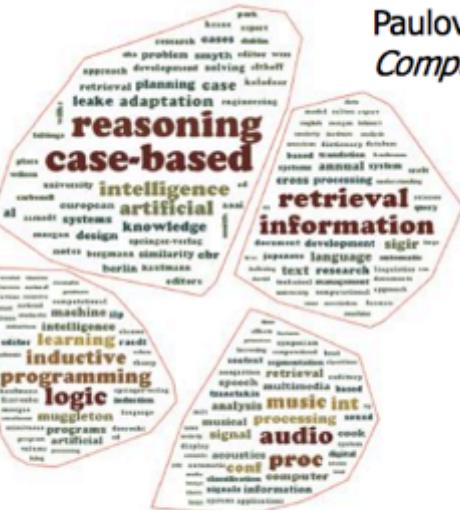
- High Dimensionality
 - Where possible use text to represent text...
 - ... which terms are the most descriptive?
- Context and Semantics
 - Provide relevant context to aid understanding.
 - Show (or provide access to) the source text.
- Modeling Abstraction
 - Determine your analysis task.
 - Understand abstraction of your language models.
 - Match analysis task with appropriate tools and models.

Semantic/Context Word Clouds

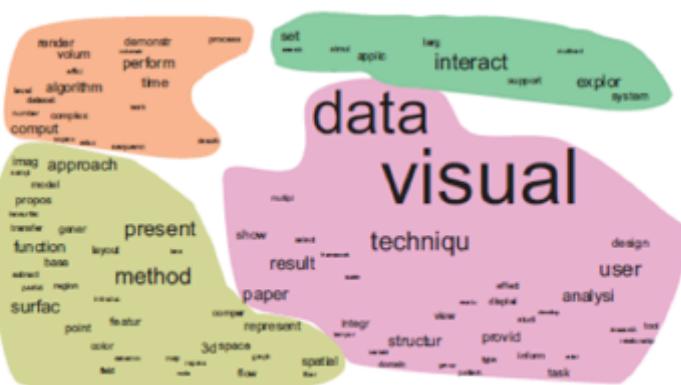
Wang et al
Graphics Interface '14



Paulovich et al
Computer Graphics Forum '12

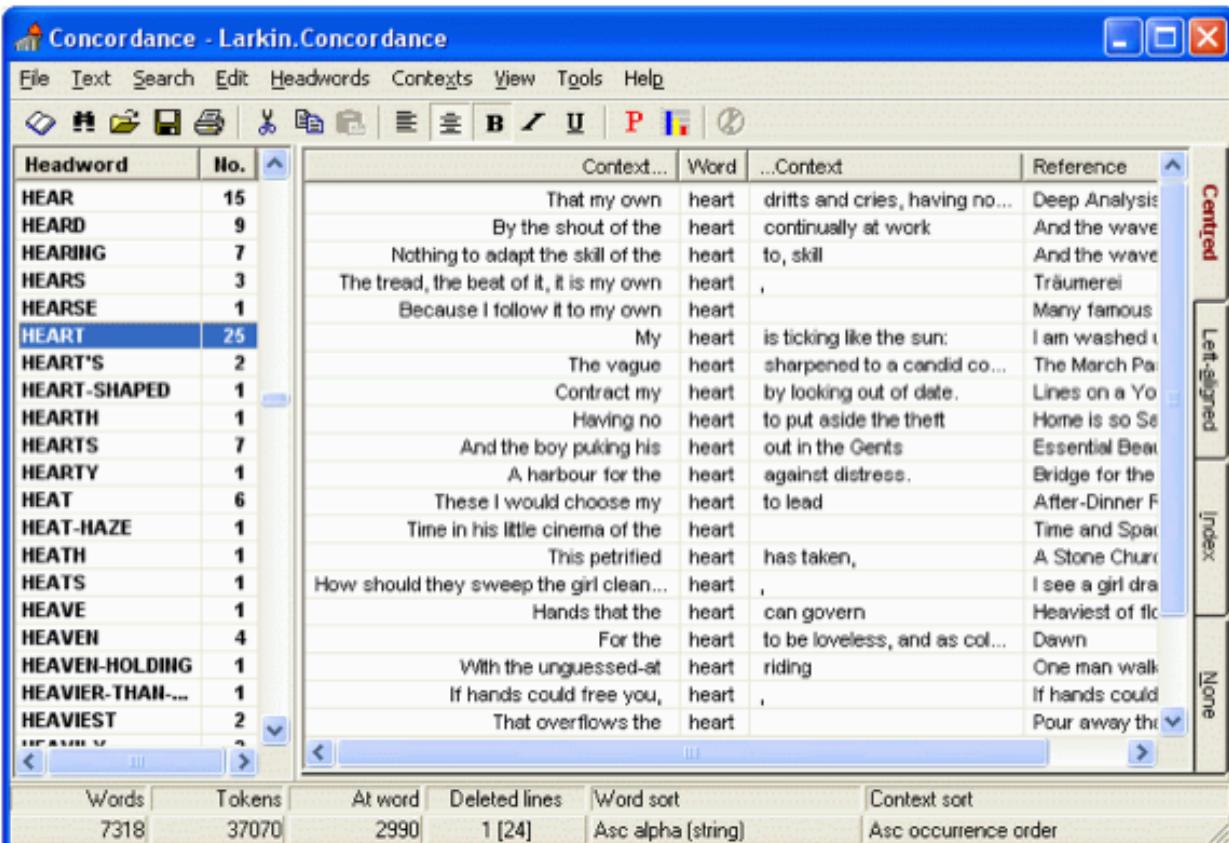


Wu et al
Computer Graphics Forum '11



Word / Phrase Context

- Concordance
 - What is the common local context of a term?

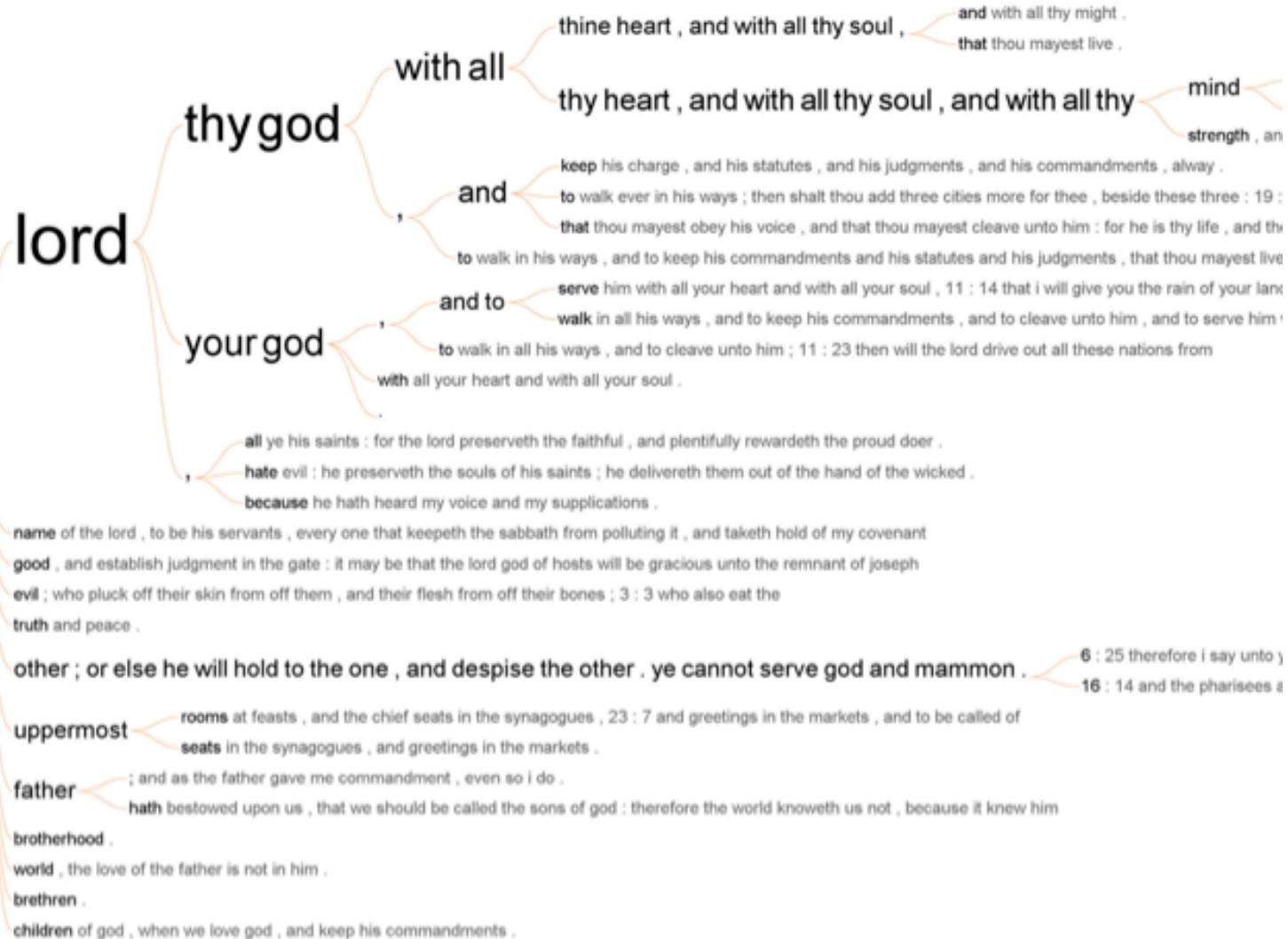


The screenshot shows the Larkin.Concordance software interface. On the left is a list of headwords with their frequencies (e.g., HEAR 15, HEART 25). The main window displays a grid of contexts for the word 'HEART'. Each row shows a context snippet, the word 'heart', and a reference line. The right side of the interface has dropdown menus for alignment ('Centred', 'Left-aligned', 'Index', 'None') and sorting ('Context sort', 'Word sort', '...Context', 'Reference').

Headword	No.	Context...	Word	...Context	Reference
HEAR	15	That my own	heart	dritts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart	,	Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed u...
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa...
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo...
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se...
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea...
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F...
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spac...
HEATH	1	This petrified	heart	has taken,	A Stone Churc...
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra...
HEAVE	1	Hands that the	heart	can govern	Heaviest of flo...
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk...
HEAVIER-THAN-...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away th...

Word Tree

love the

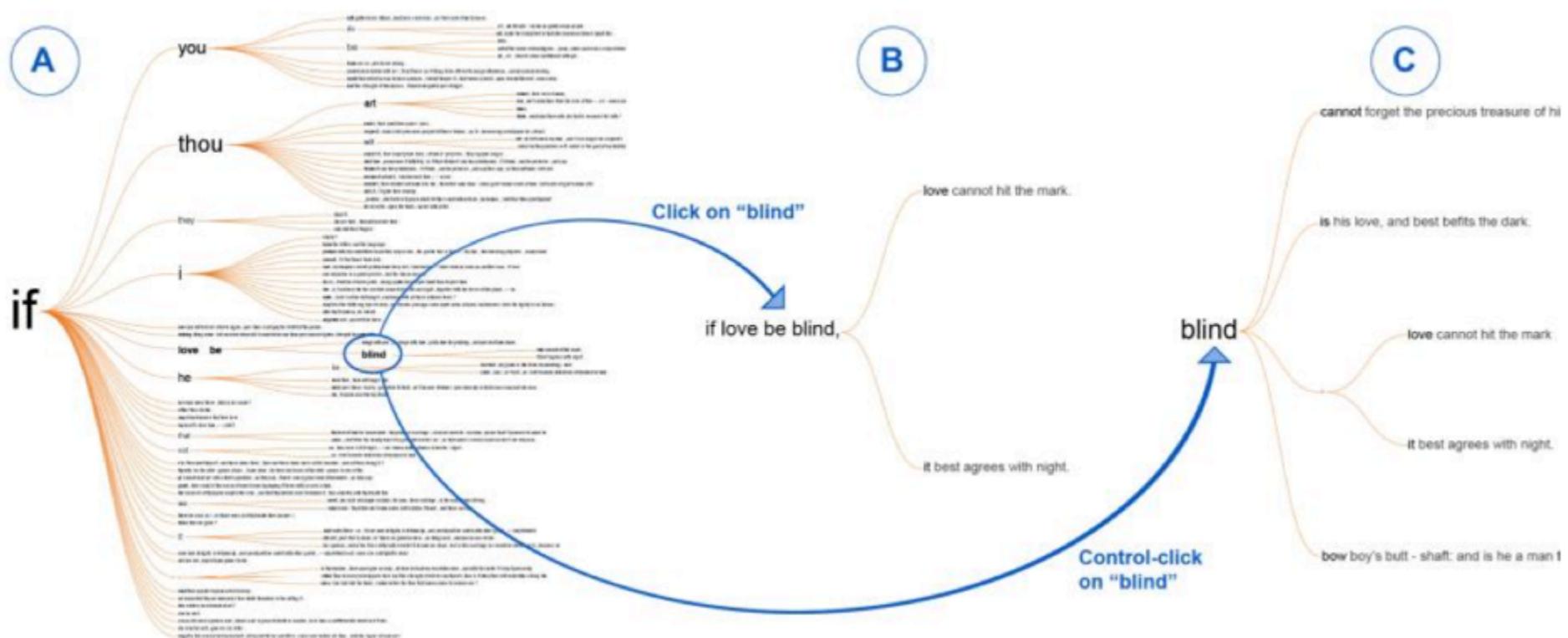


Word Tree

- Shows context of a word or words
 - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

(Optional) Wattenberg & Viégas TVCG (InfoVis) '08

Word Tree Interaction



Phrase Nets

- Concordances show local, repeated structure, but what about other types of patterns?
 - Lexical: <A> at
 - Syntactic: <Noun> <Verb> <Object>
- Look for specific linking patterns in the text:
 - ‘A and B’, ‘A at B’, ‘A of B’, etc
 - Could be output of regexp or parser.
- Visualize patterns in a node-link view
 - Occurrences -> Node size
 - Pattern position -> Edge direction

Phrase Nets

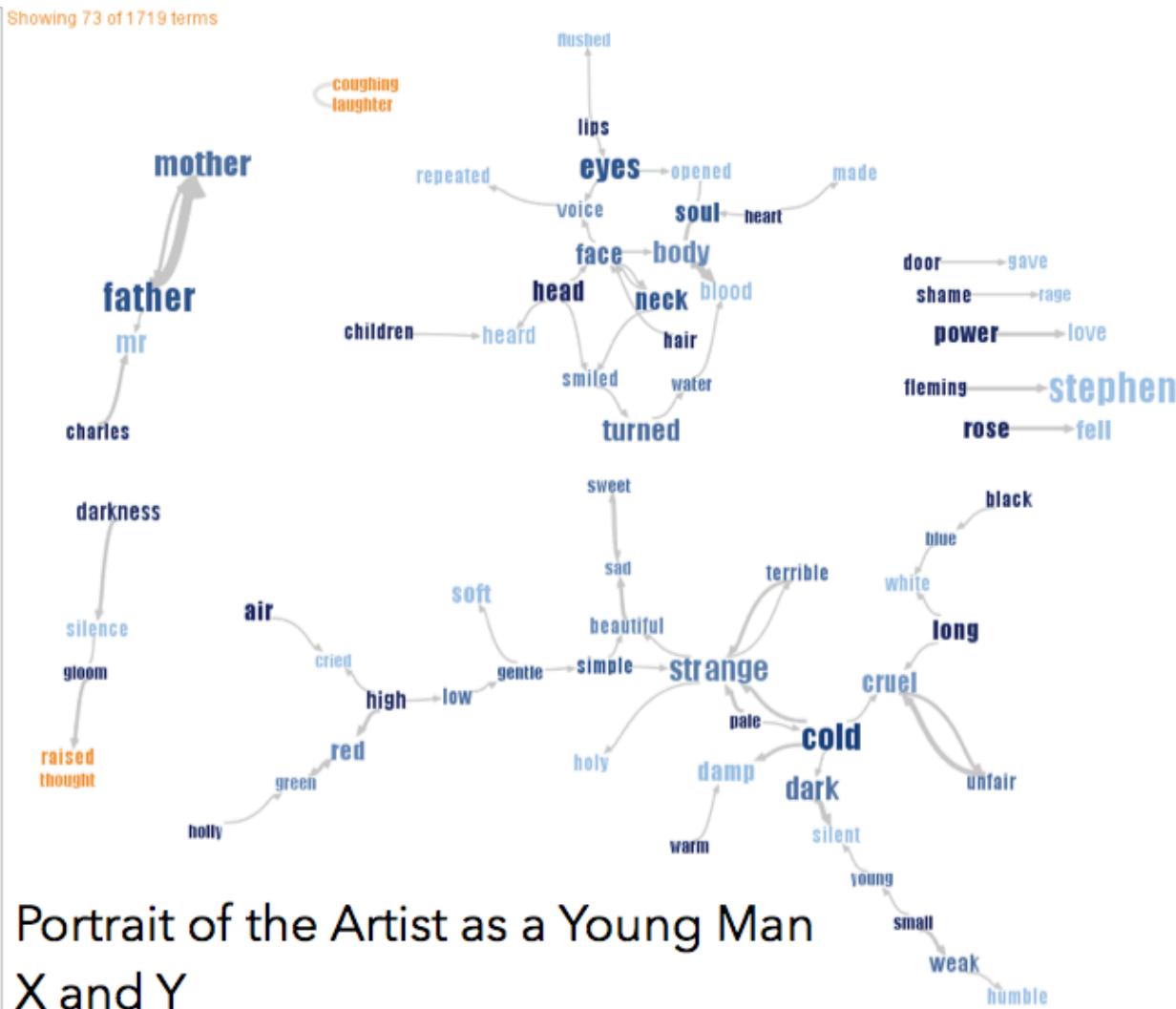
Select a phrase

word1	and	word2
word1	's	word2
word1	of the	word2
word1	the	word2
word1	a	word2
word1	at	word2
word1	is	word2
word1	[space]	word2

or enter your own
 * and *

Filters
 Show top: 100
 Hide common words

Zoom

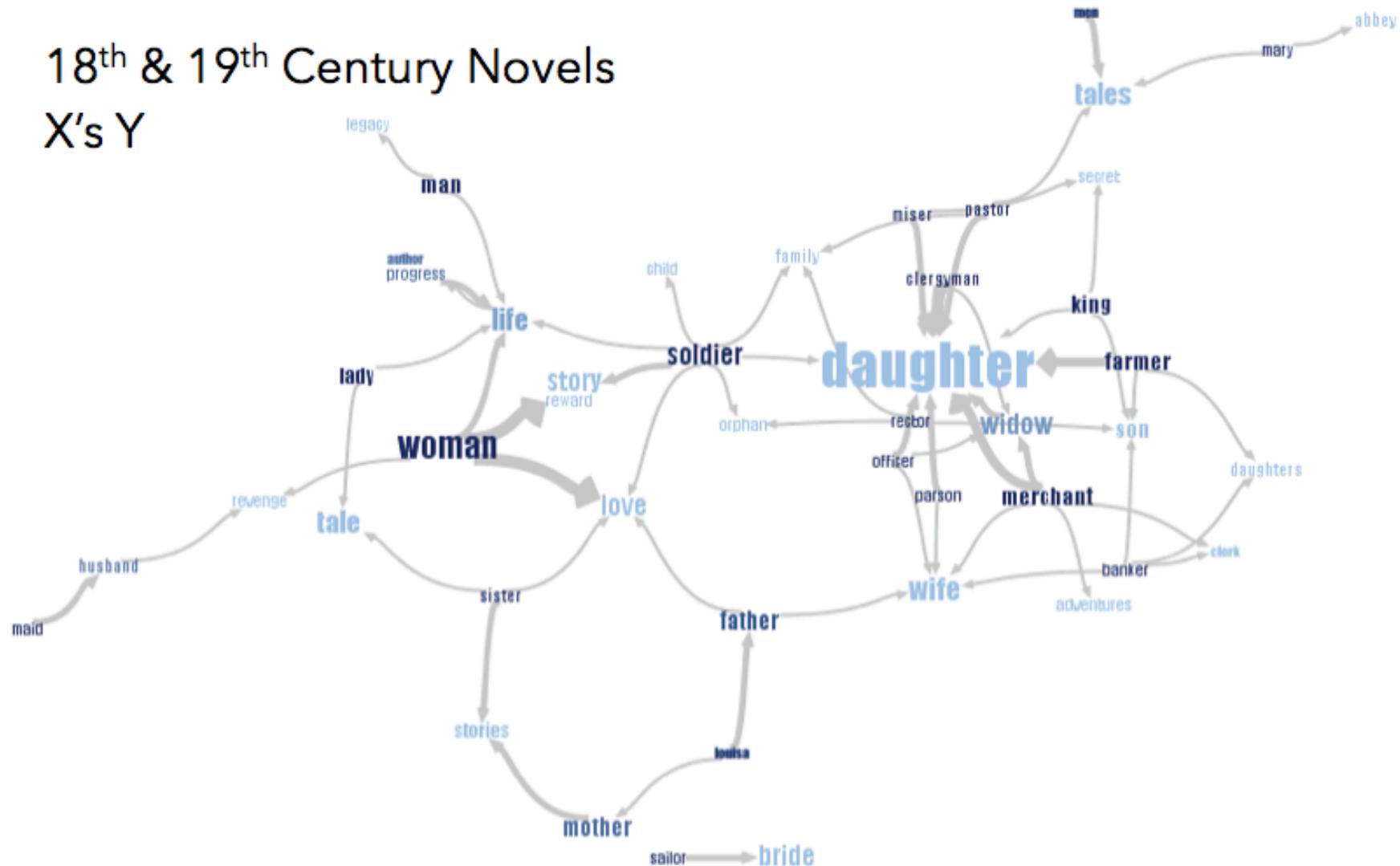


Portrait of the Artist as a Young Man
X and Y

Phrase Nets

18th & 19th Century Novels

X's Y



SentenTree

- Elements of word clouds and word trees
 - Highlight keywords using size
 - Show sentence fragments
 - Provide a summary of the dataset
 - Enable drill-down into details



Hu, et al. TVCG '17 (InfoVis '16)

Summary of 189,450 tweets (108,702 unique) posted in a 15 minute time window around the first goal of the opening game of the 2014 Soccer World Cup

Relationship to Other Texts

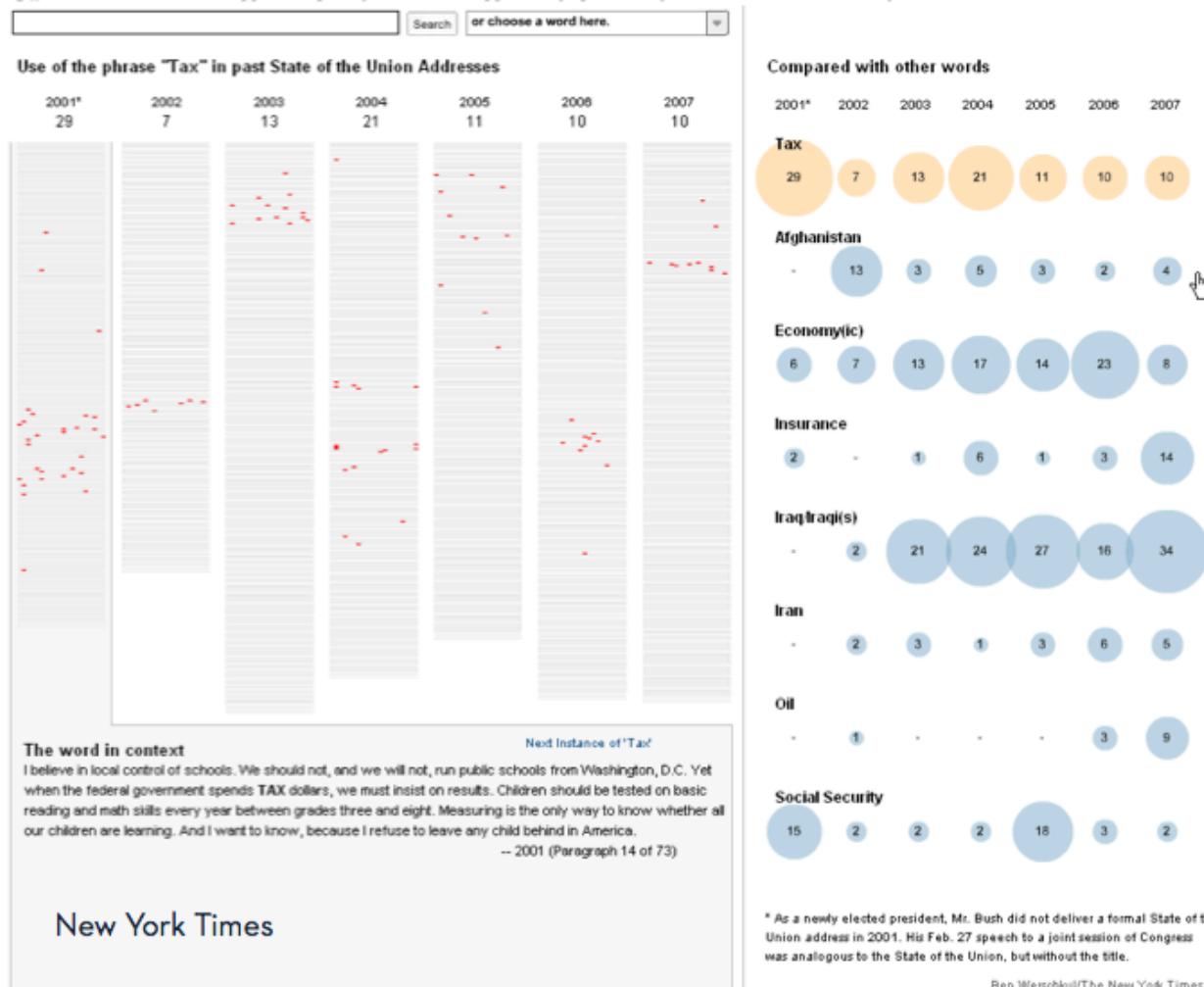
Compare to Others

THE WORDS THAT WERE USED

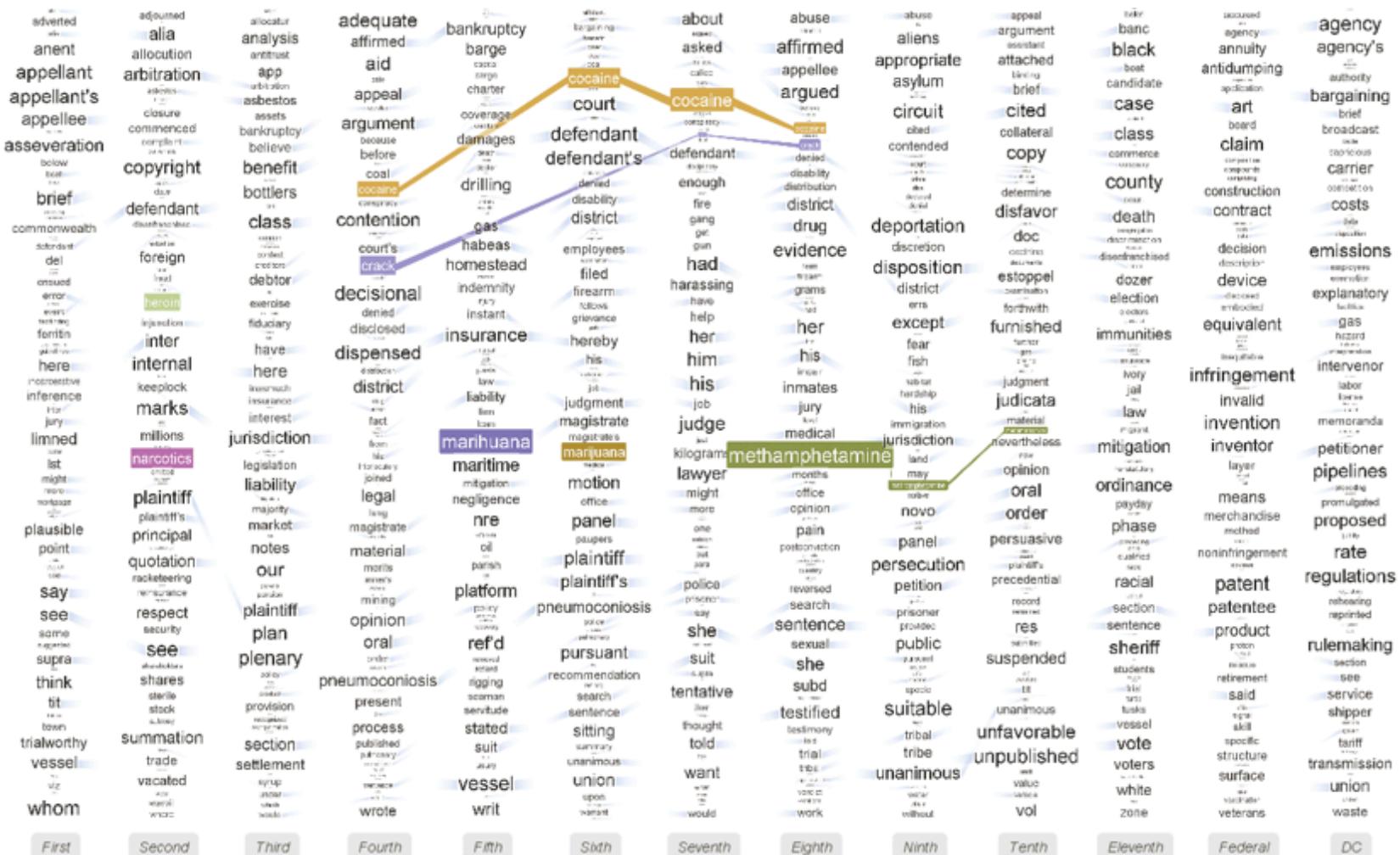
The 2007 State of the Union Address

READ 2007 SPEECH | FEEDBACK

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.



Parallel Tag Clouds



Collins et al VAST '09

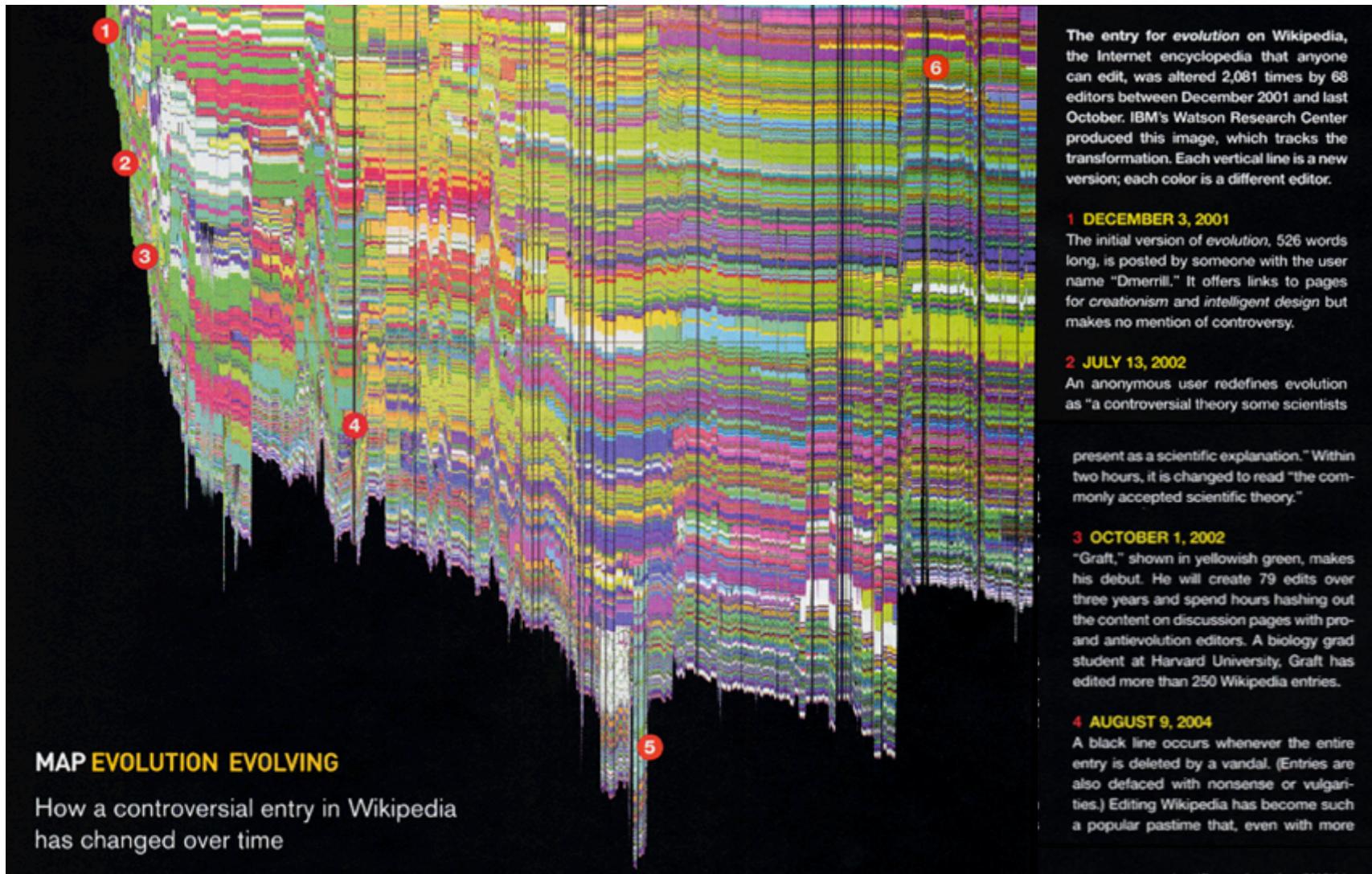
Video: <https://www.youtube.com/watch?v=rL3Ga6xBgLw>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Animated Traces: evolving documents



Wikipedia Edit Evolution

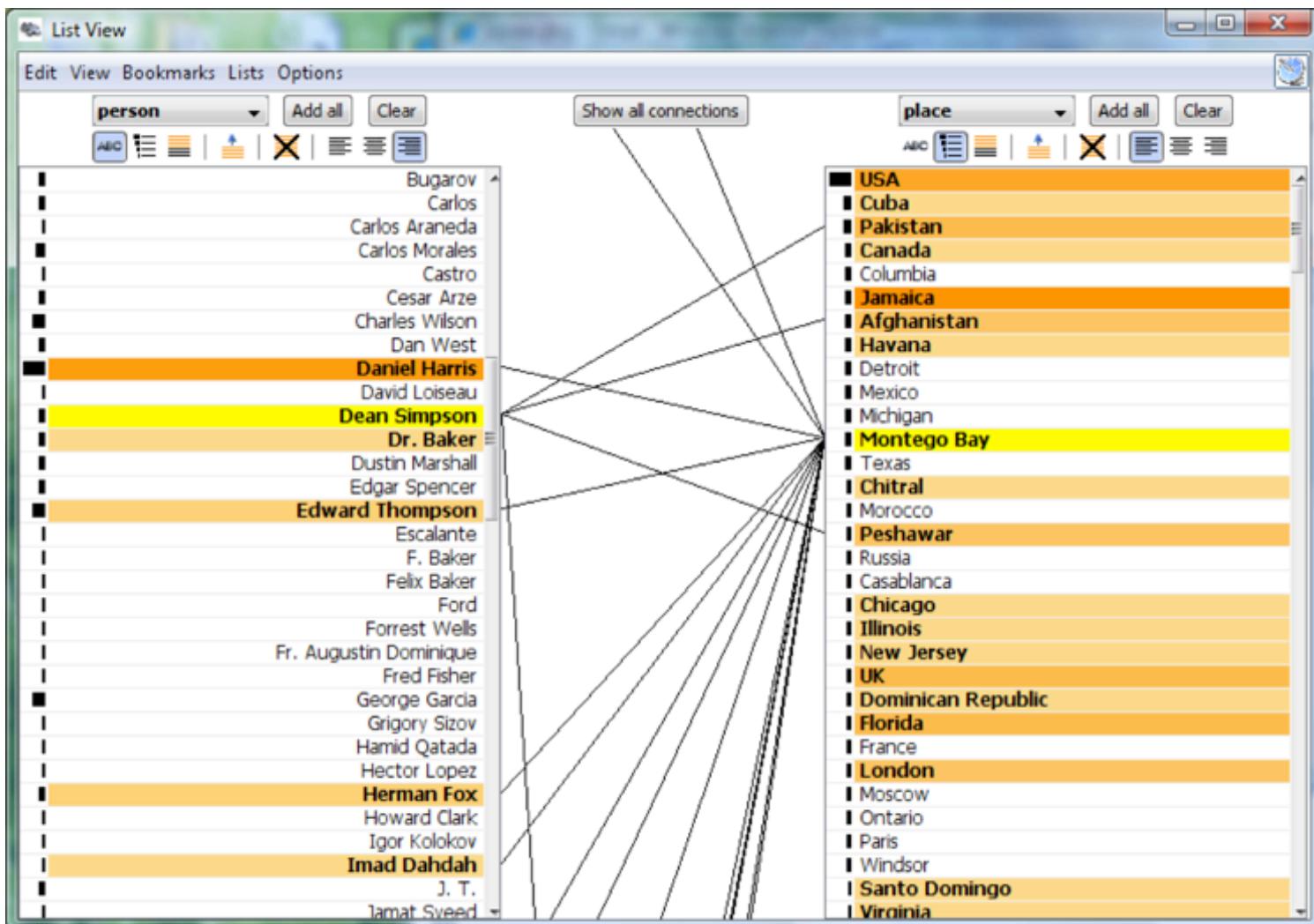


Visualizing Document Collections

Named Entity Recognition

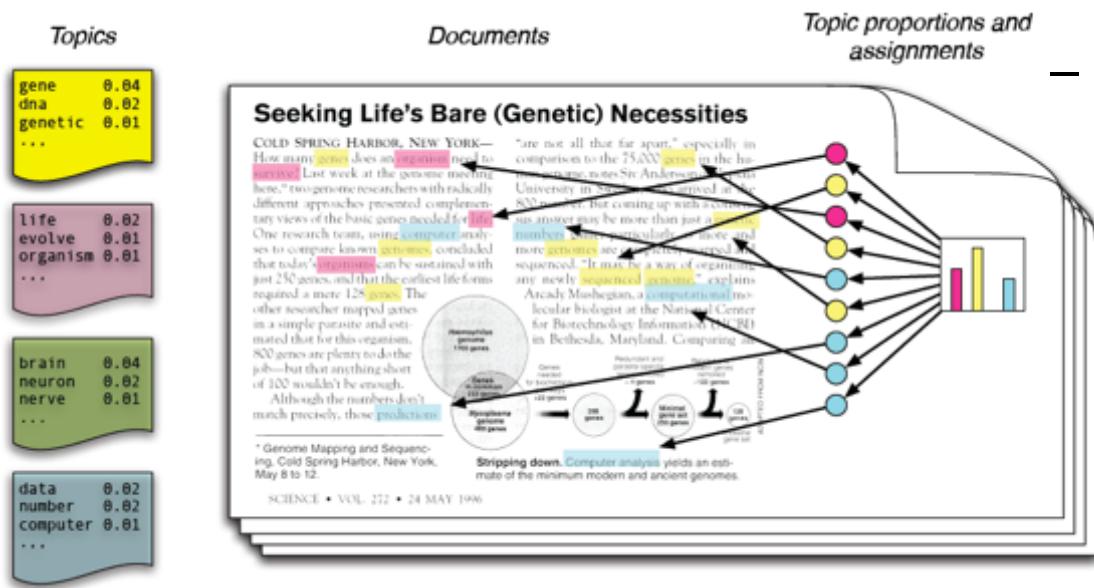
- Label named entities in text:
 - John Smith -> PERSON
 - Soviet Union -> COUNTRY
 - 353 Serra St -> ADDRESS
 - (555) 721-4312 -> PHONE NUMBER
- Entity relations: how do the entities relate?
- Simple approach: do they co-occur in a small window of text?

Entity Linkage



Similarity & Clustering

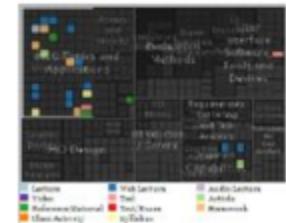
- Compute vector distance among docs
 - For TF.IDF, typically cosine distance Similarity measure can be used to cluster
- Topic modeling
 - Assume documents are a mixture of topics
 - Topics are (roughly) a set of co-occurring terms
 - Latent Semantic Analysis (LSA): reduce term matrix
 - Latent Dirichlet Allocation (LDA): statistical model



Visualization for Information Retrieval

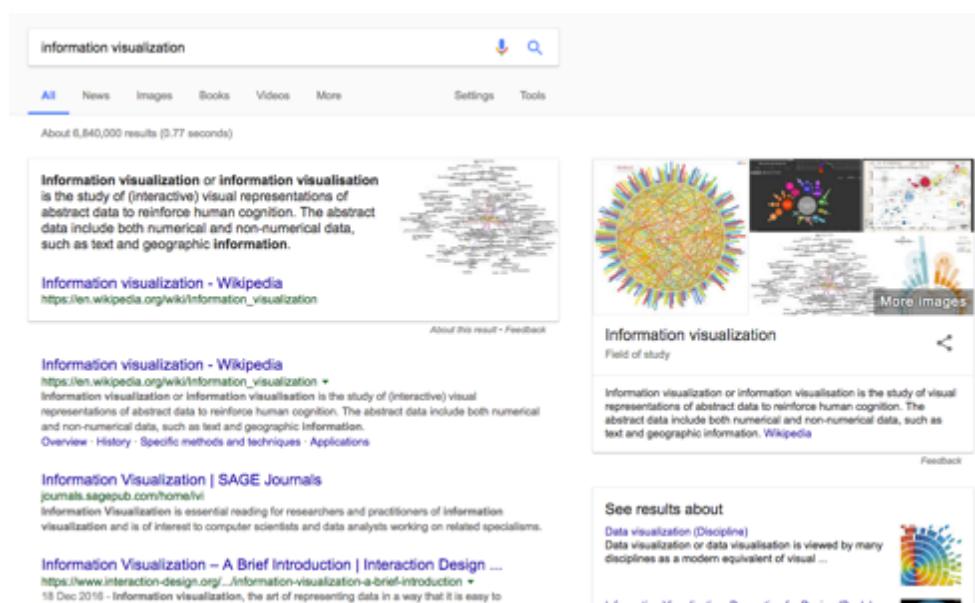


Visualization for IR



Information Retrieval (IR)

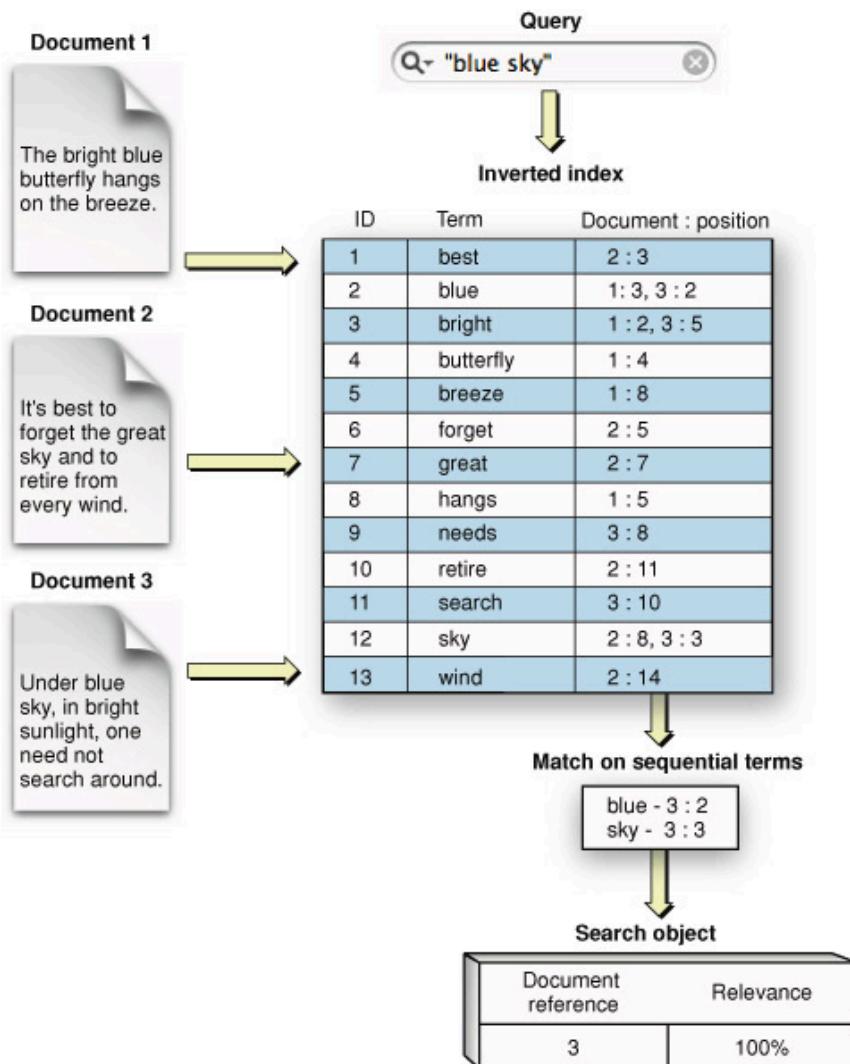
- Search for documents
- Visualization to contextualize query-doc matching results
- Can Information visualization help IR?
 - A (large) set of documents



A screenshot of a Google search results page for the query "information visualization". The search bar at the top contains the query. Below it, a navigation bar offers options: All, News, Images, Books, Videos, More, Settings, and Tools. The main search results area shows approximately 6,840,000 results found in 0.77 seconds. The first result is a summary of "Information visualization or information visualisation" from Wikipedia, featuring a complex network graph visualization. Below this is another Wikipedia entry for "Information visualization - Wikipedia" with a similar network diagram. Further down are links to "Information Visualization | SAGE Journals" and "Information Visualization – A Brief Introduction | Interaction Design ...". On the right side of the results, there's a sidebar titled "Information visualization" under "Field of study" with a link to the Wikipedia page. At the bottom right, there's a "Feedback" link.

Paper Handout: <http://searchuserinterfaces.com/book/>, Chapter 10

Information Retrieval



Ten Blue Links to More Complex Data and Visualization

Google search results for "homebrew podcast":

- Basic Brewing™ : Home Brewing Beer Podcast and DVD ...**
www.basicbrewing.com/radio/ ▾
 We visit Modern Times Brewing in San Diego to talk to Jacob McKean and Michael Tonsmeire about starting a brewery based on homebrew, roasting and ...
 Basic Brewing Radio™ 2014 - Basic Brewing Radio™ 2006
- The Brewing Network | Beer Radio for Brewers and Beer ...**
www.thebrewingnetwork.com/ ▾
 We went (like true homebrewing patriots) and partied at Club Night as sponsored by ...
 Podcast (jami-show): Play in new window | Download (Duration: 1:02:45 ...
 Brew Strong - Shows - Beer Forum - Dr. Homebrew
- BeerSmith Home Brewing Podcast**
www.beersmith.com ▾ BeerSmith Home Brewing Forum ▾ Our Web Sites ▾
 BeerSmith Home Brewing Podcast. ... by BeerSmith - All Podcast Episodes available on iTunes now! ... Episode #96 - Mastering Homebrew with Randy Mosher.
- Homebrew Podcast | Homebrew Podcasts**
www.hogtownbrewers.org/podcasts.cfm ▾
 Hosted by homebrew jockeys Ron & John, Homebrew Talk is a monthly local Gainesville podcast about all things homebrewing. Get your brew on by exploring ...
- Brewing Podcasts - Brew Your Own**
<https://byo.com/hops/item/302-brewing-podcasts> ▾ Brew Your Own ▾
 Homebrewing podcasts run the gamut from offering tips on the basics of homebrewing, to walking you through award-winning recipes, to attempting innovative ...
- Become a Better Brewer with the 5 Best Homebrewing ...**
blog.kegoutlet.com/become-a-better-brewer-with-the-5-best-homebrew... ▾
 Here are the 5 best homebrewing podcasts that I have found and listen to regularly. These 5 podcasts will make you a better brewer. Subscribe to any or all of ...



Google search results for "nottingham":

- The University of Nottingham - a world top 1% university**
<https://www.nottingham.ac.uk/> ▾
 The University of Nottingham ... Celebrating MRI in Nottingham. Marking 25 years of the Sir Peter Mansfield Magnetic Resonance Imaging Centre ...
- Nottingham - Wikipedia**
<https://en.wikipedia.org/wiki/Nottingham> ▾
 Nottingham is a city and unitary authority area in Nottinghamshire, England, located 30 miles (48 km) south of Sheffield and 30 miles (48 km) north of Leicester. Nottinghamshire: University of Nottingham - List of people from Nottingham - Arnold
- The Top 10 Things to Do in Nottingham 2017 - TripAdvisor**
<https://www.tripadvisor.co.uk/.../England-Nottingshire-Nottingham> ▾
 Top Things to Do in Nottingham, Nottinghamshire - Nottingham Attractions, ... Nottingham weather essentials ... Historic Sites, Science Museums, Points of Interest & Landmarks.
- Nottingham Post: Nottingham News, Sports & Events**
www.nottinghampost.com ▾
 Get the latest news from the Nottingham Post online. Plus breaking news updates, sports, events and local businesses in Nottinghamshire.

Top stories



Six things Bros told us before cancelling Nottingham concert

Nottingham Post · 17 hours ago



This is what it's like to work in Nottingham's oldest sex shop

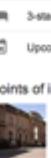
Nottingham Post · 19 hours ago



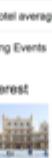
New cave is discovered under Nottingham city centre

Nottingham Post · 52 mins ago

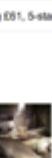
Points of interest



Galleries of Justice Museum



Wollaton Hall



City of Caves



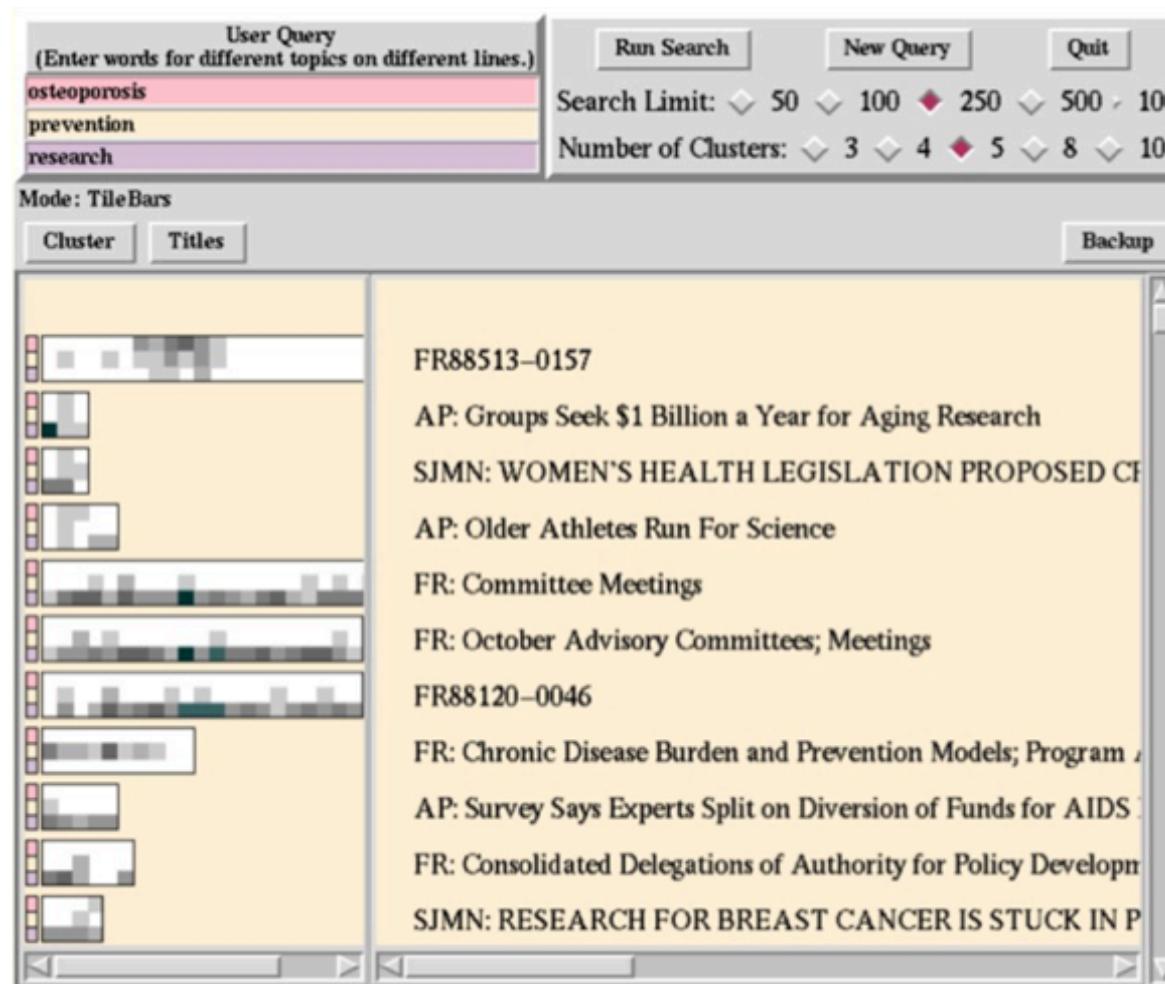
Nottingham Contemporary



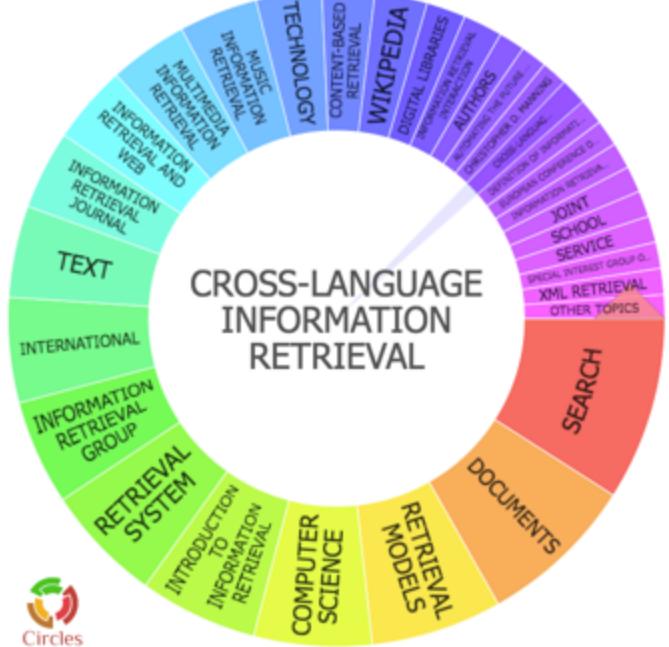
Lace Market

[View 10+ more](#)

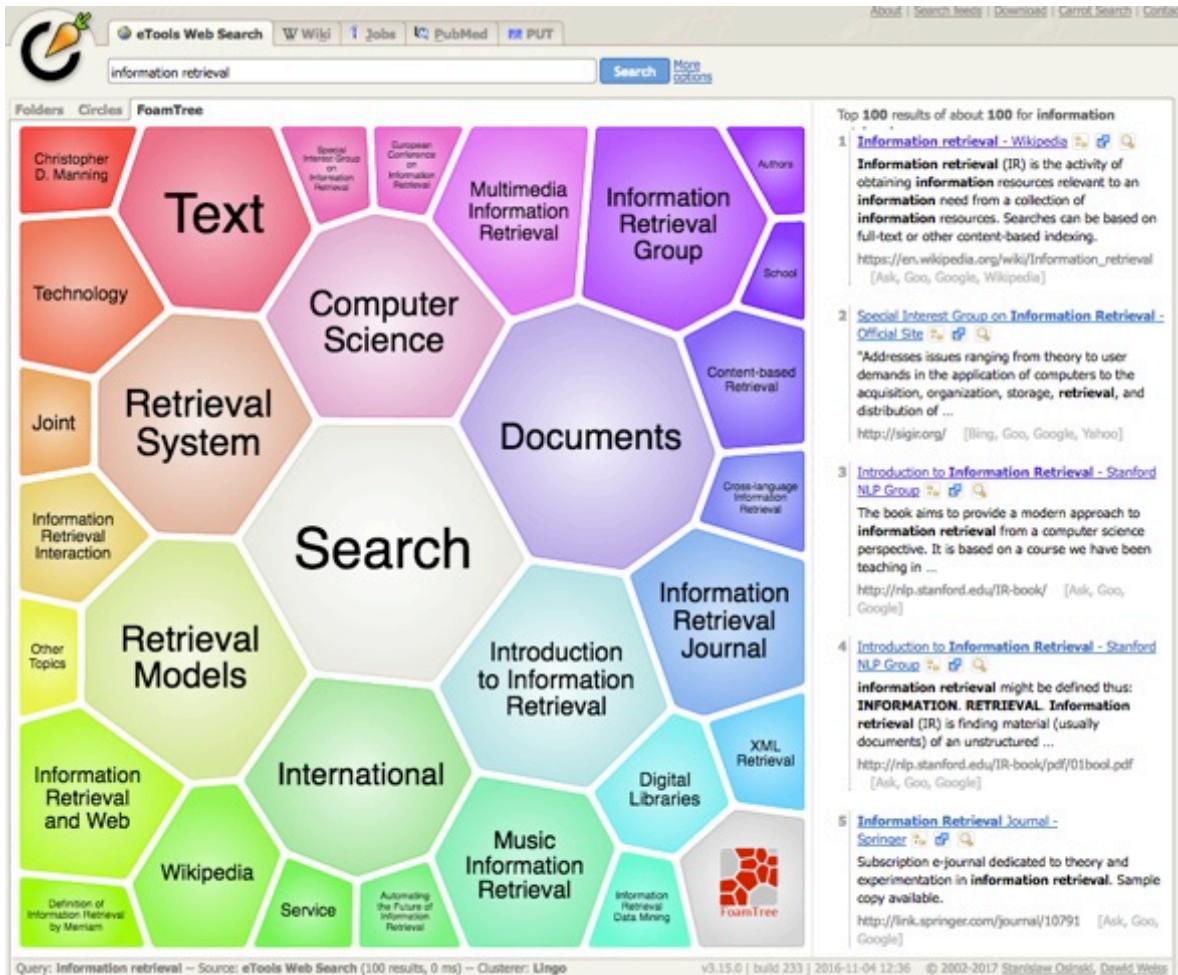
TileBars



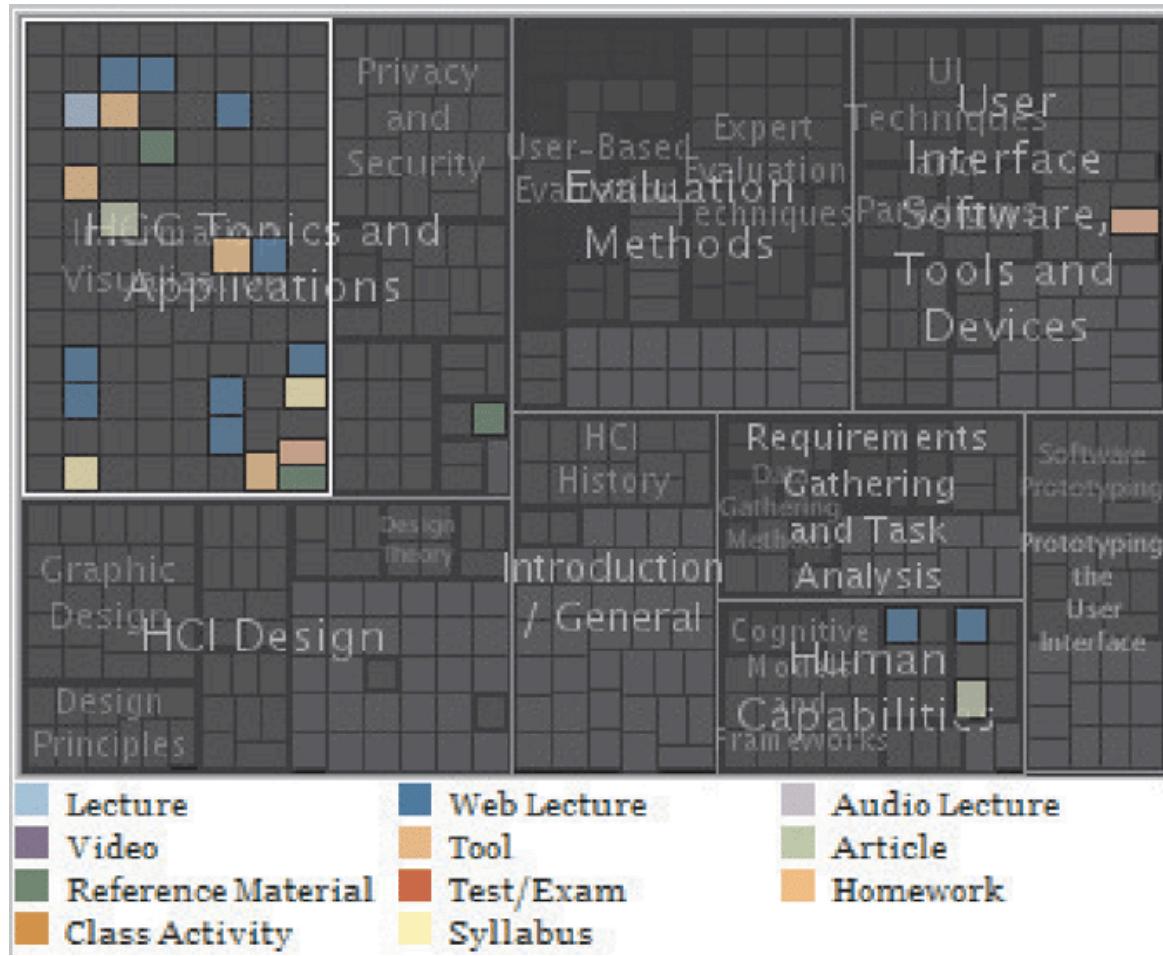
Cluster Based Search Visualization



<http://search.carrot2.org/>

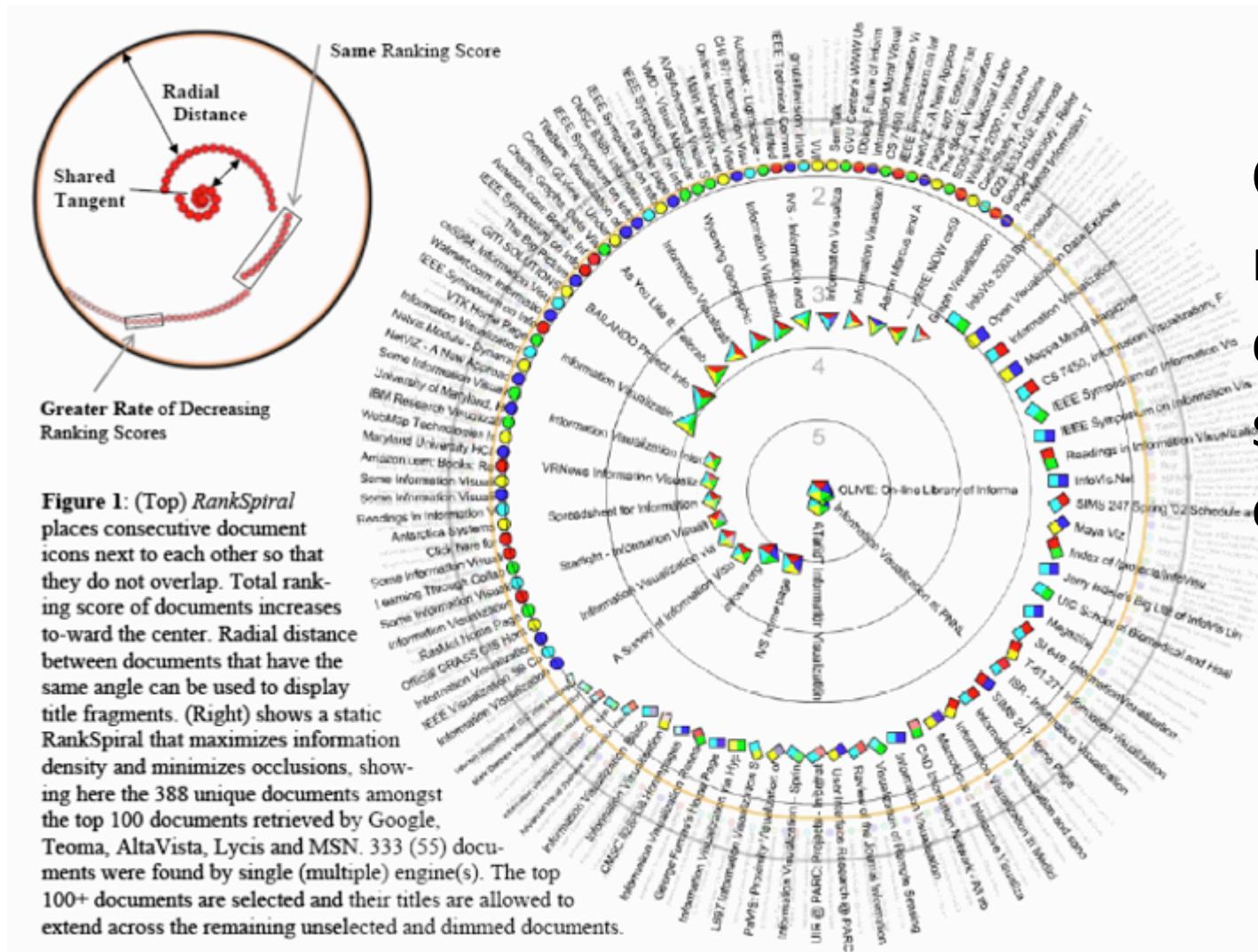


Result Maps



Treemap-style visualization for showing query results in a digital library

RankSpiral



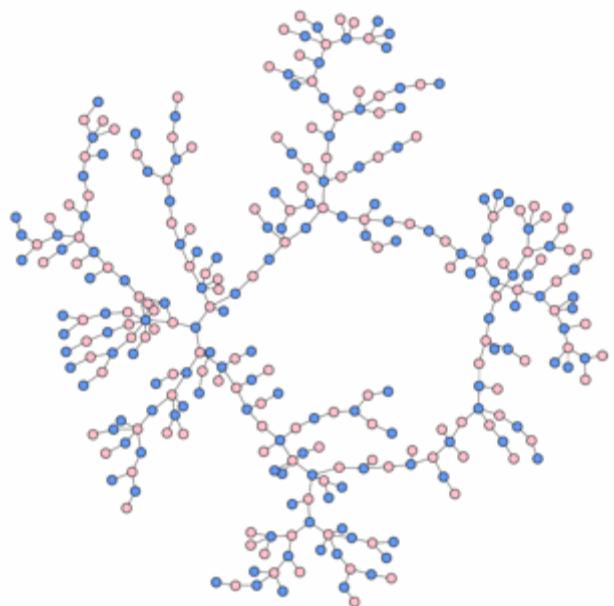
Color
represents
different
search
engines

Summary

- High Dimensionality
 - Where possible use text to represent text...
 - ... which terms are the most descriptive?
- Context & Semantics
 - Provide relevant context to aid understanding.
 - Show (or provide access to) the source text.
- Modeling Abstraction
 - Understand abstraction of your language models.
 - Match analysis task with appropriate tools and models.
 - From bag-of-words to vector space embeddings

Next Lecture

- Topic:
 - Visualizing Time Series,
Trees and Graphs



G53FIV: Fundamentals of Information Visualization

Lecture 12: Visualizing Time Series, Trees and Graphs

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

Overview

- Visualizing Time Series Data
- Trees and Graphs

Visualizing Time Series Data

Traditional Time Series Visualization



NVIDIA stock vs. NASDAQ (from Yahoo! finance)

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Challenges

- Temporal relationships can be highly complex
 - temporal ordering is a serious issue
 - event may occur in spatially disjoint locations
 - what came before what – cause and effect
 - what time shifts are acceptable/plausible?
- To understand temporal relationships, an analyst
 - might need to reread the paragraph many times
 - needs to cognitively make inferences between pieces of information

Tasks

- Often asked questions:
 - when was something greatest/least?
 - is there a pattern? are two series similar?
 - does a data element exist at time t, and when?
 - how long does a data element exist and how often?
 - how fast are data elements changing
 - in what order do data elements appear?
 - do data elements exist together?

(Optional Reading) Müller, Wolfgang, and Heidrun Schumann. "Visualization for modeling and simulation: visualization methods for time-dependent data-an overview." Proceedings of the 35th conference on Winter simulation: driving innovation. Winter Simulation Conference, 2003.

Taxonomy

<i>Time</i>	Temporal primitives	time points (a) (b) (c) (d) (e) (f) (g) (i)		time intervals (g) (h)
	Structure of time	linear (a) (b) (c) (d) (f) (g) (h) (i)		cyclic (e)
<i>Data</i>	Frame of reference	abstract (c) (d) (f) (g) (h) (i)		spatial (a) (b) (e) (i)
	Number of variables	univariate (a) (b) (f) (g) (h)		multivariate (c) (d) (e) (i)
	Level of abstraction	data (a) (b) (c) (d) (e) (f) (g) (h) (i)		data abstractions (b) (g) (i)
<i>Representation</i>	Time dependency	static (c) (d) (e) (g) (h) (i)		dynamic (a) (b) (f) (i)
	Dimensionality	2D (a) (c) (d) (g) (h) (i)		3D (b) (e) (f) (i)

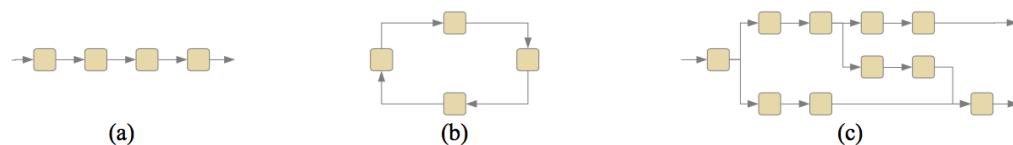
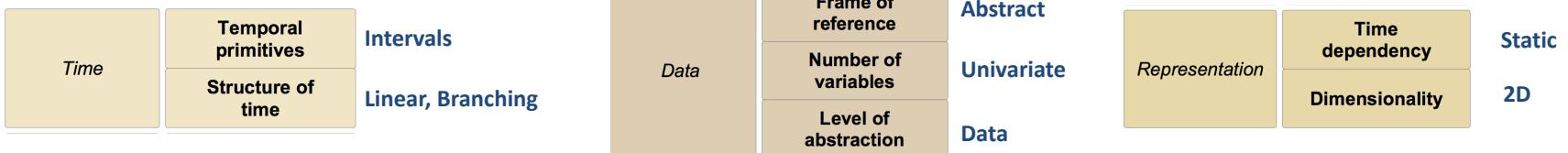
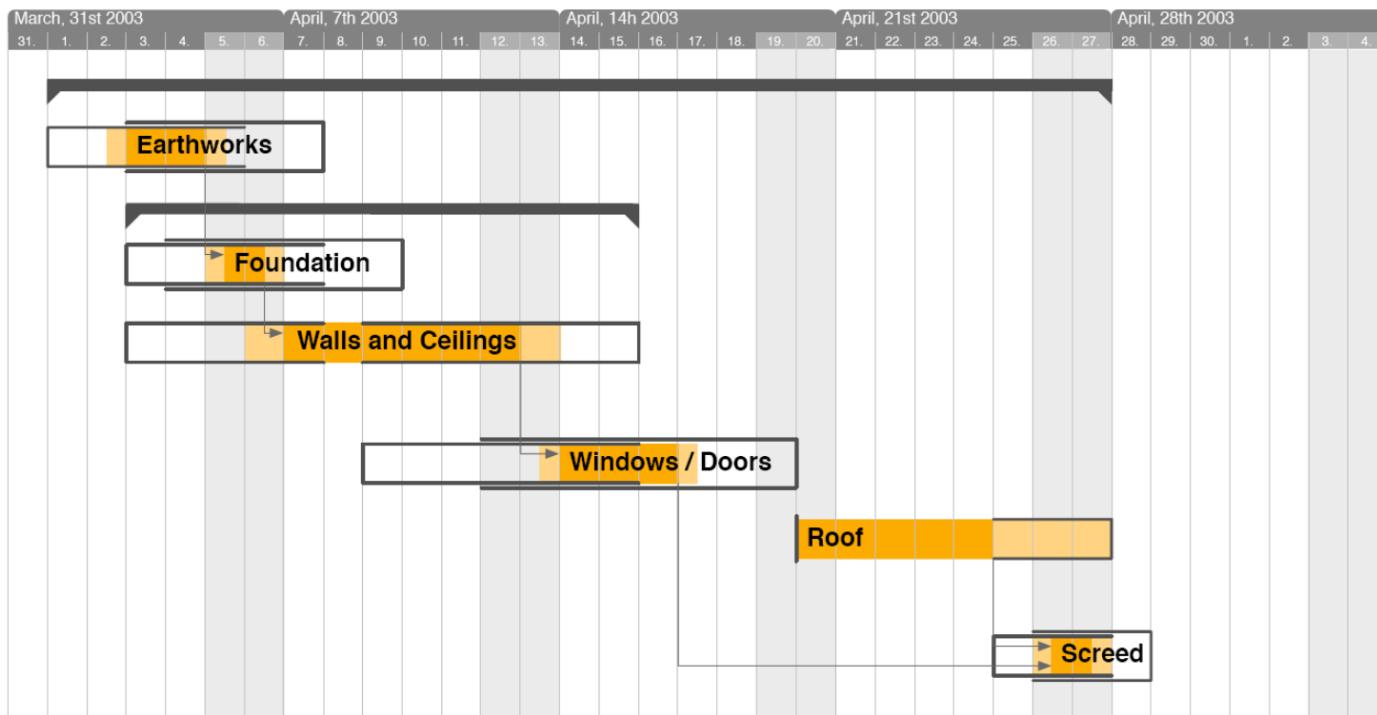
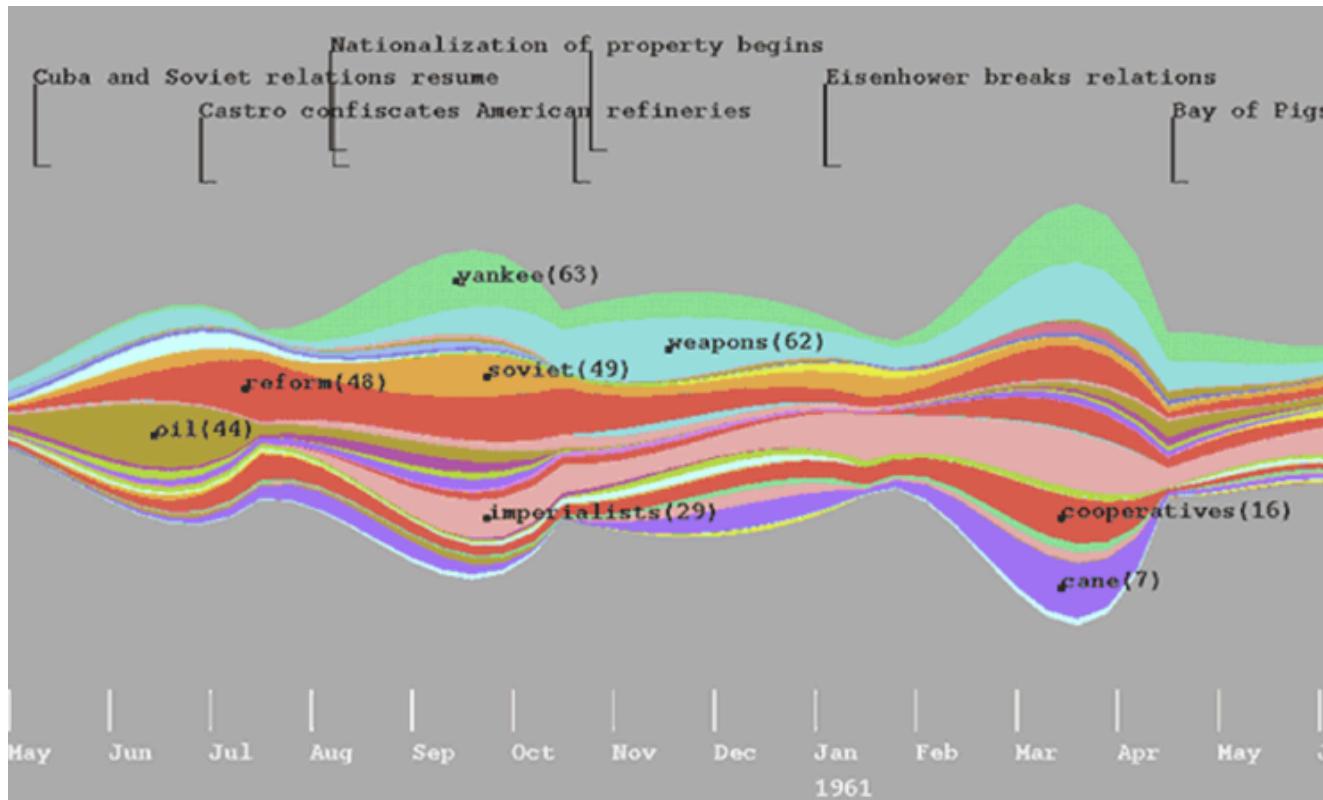


Fig. 2. Structure of time: (a) Linear time; (b) Cyclic time; (c) Branching time.

Gantt Chart

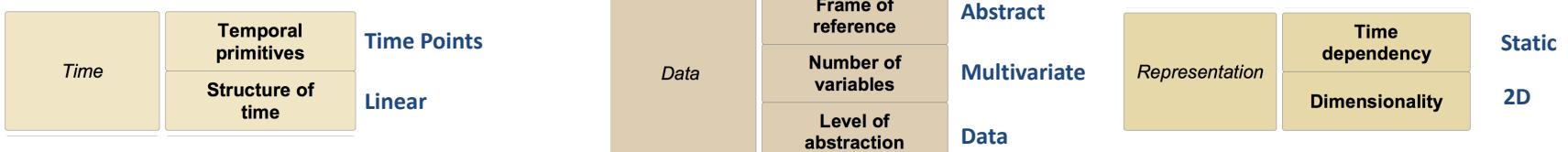
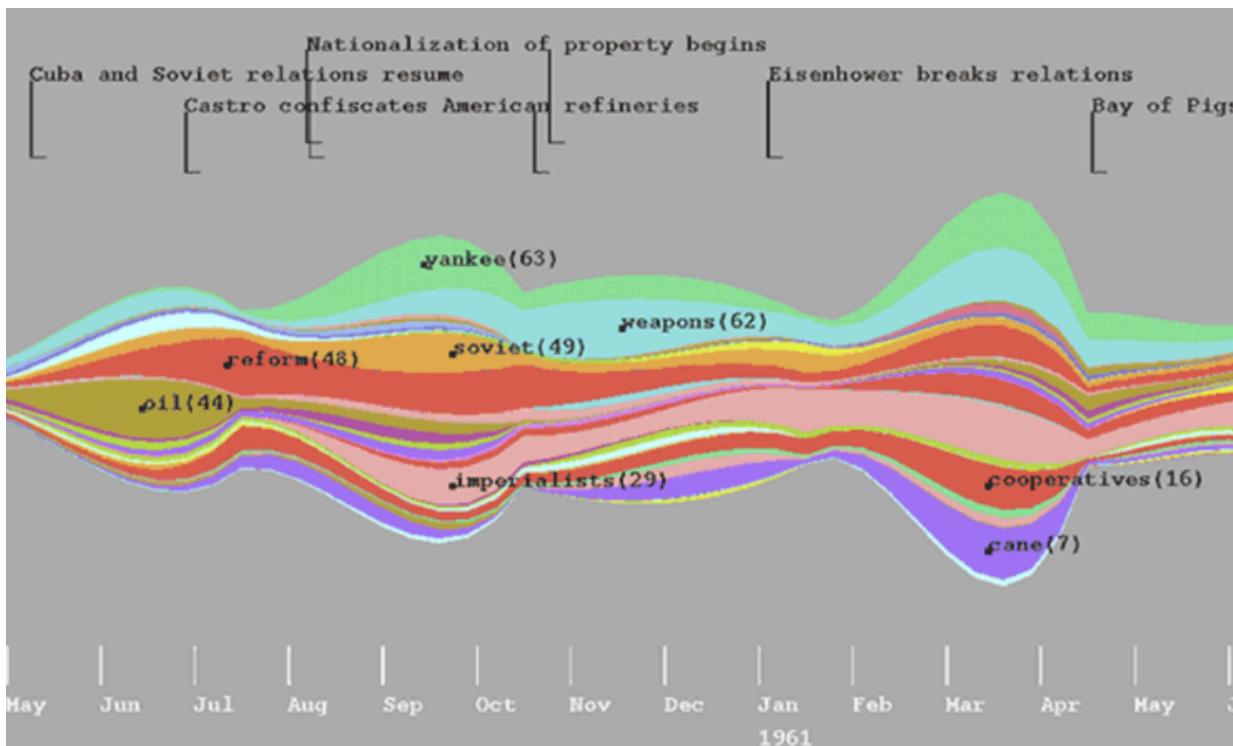


Theme River (Stream Graphs / Stacked Area Charts)



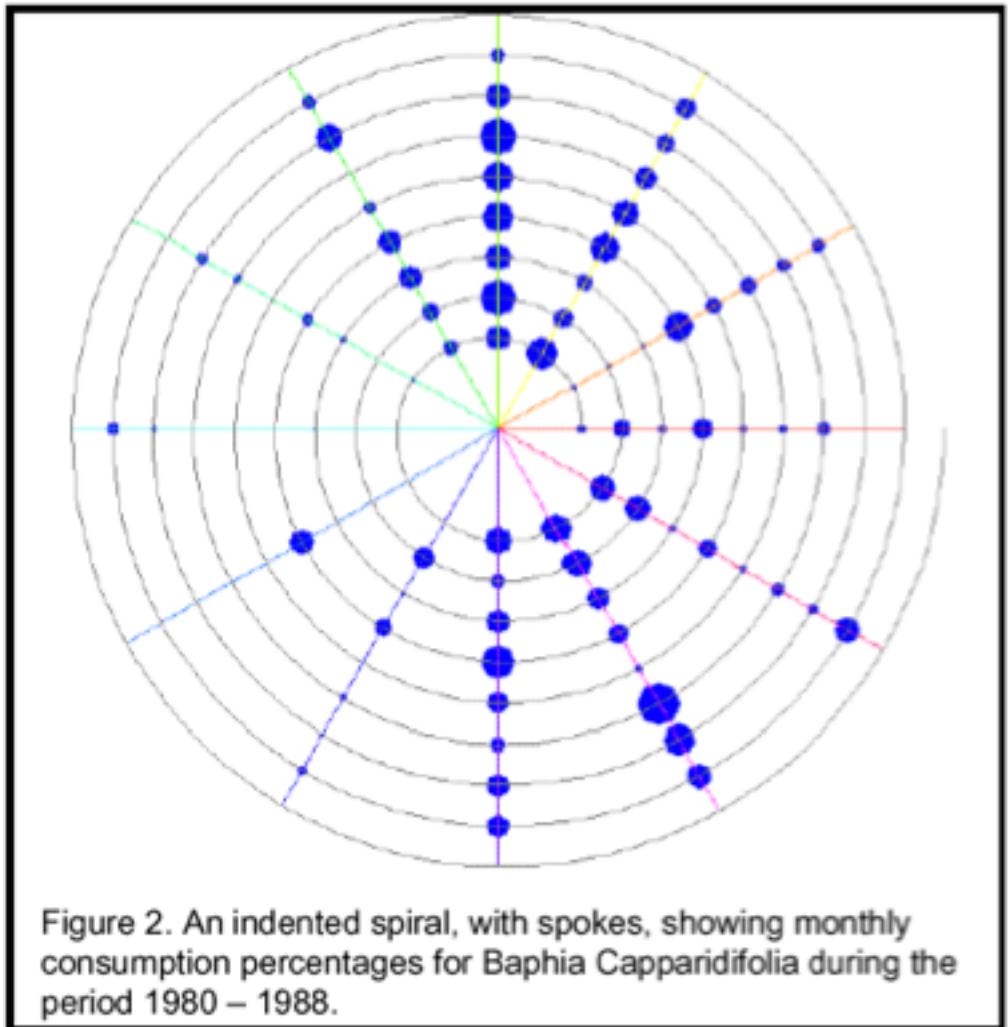
River widens or narrows to depict changes in the collective strength of selected themes in the underlying documents. Individual themes are represented as colored "currents" flowing within the river (example: Cuban Missile crisis) .

Theme River

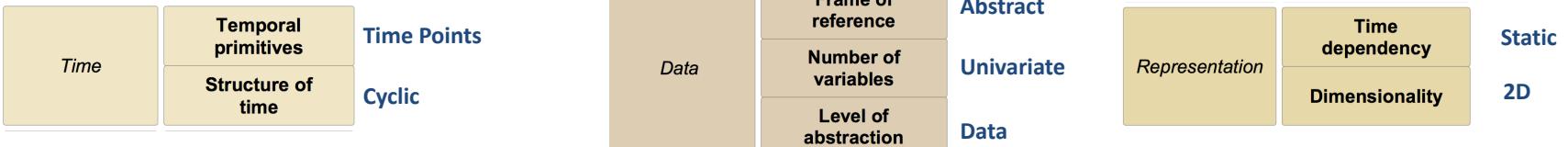
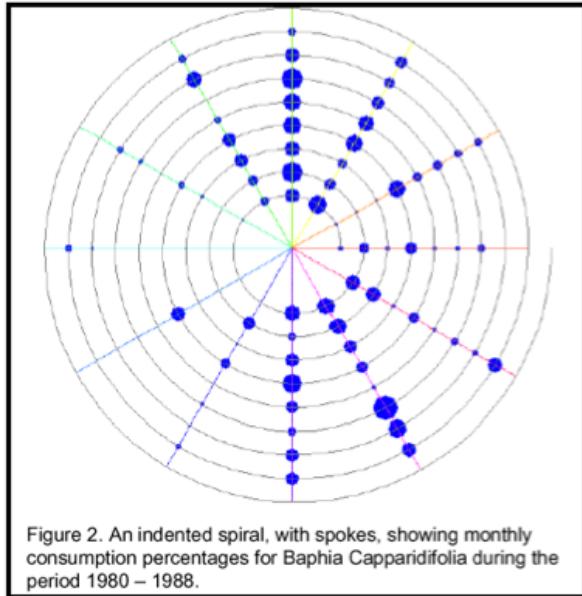


Cyclic Patterns

- Time data are often cyclic
 - Spiral displays are good to bring out cyclic patterns
 - One period per loop (for example, a year)

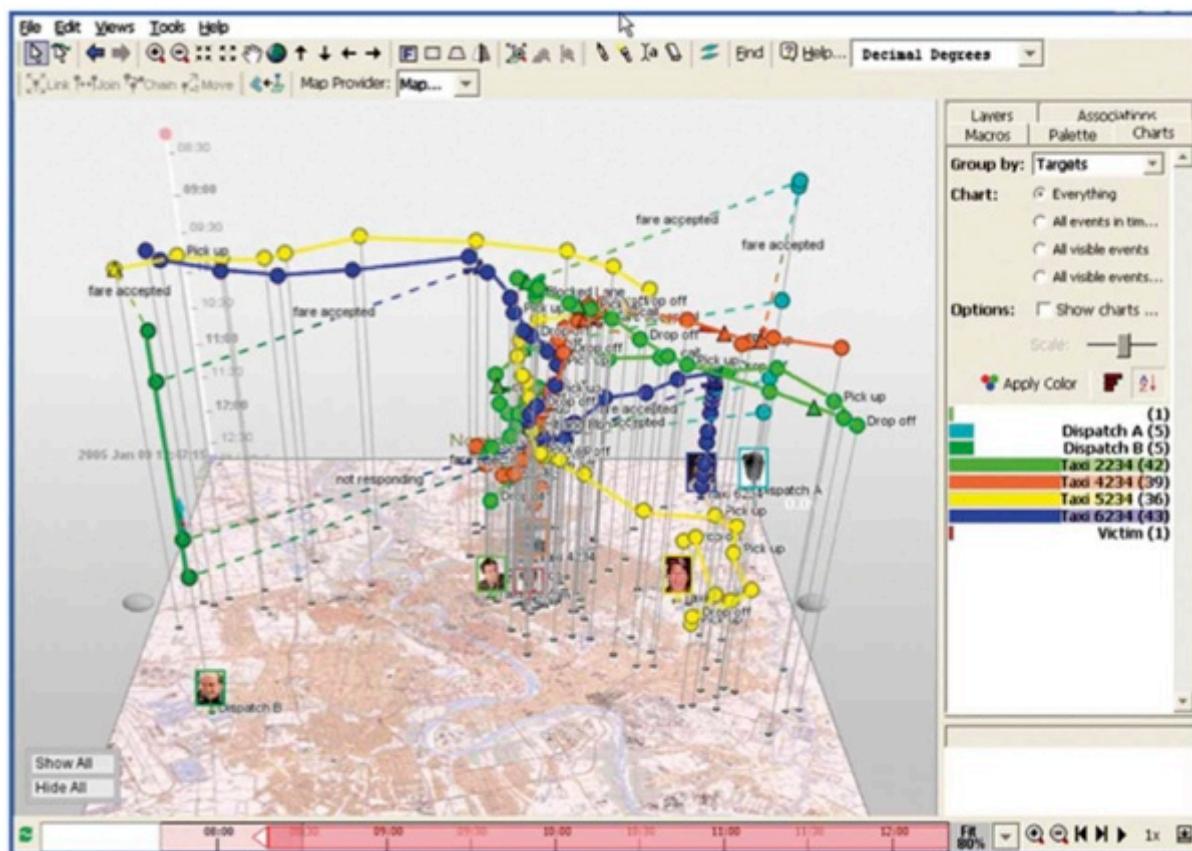


Cyclic Patterns



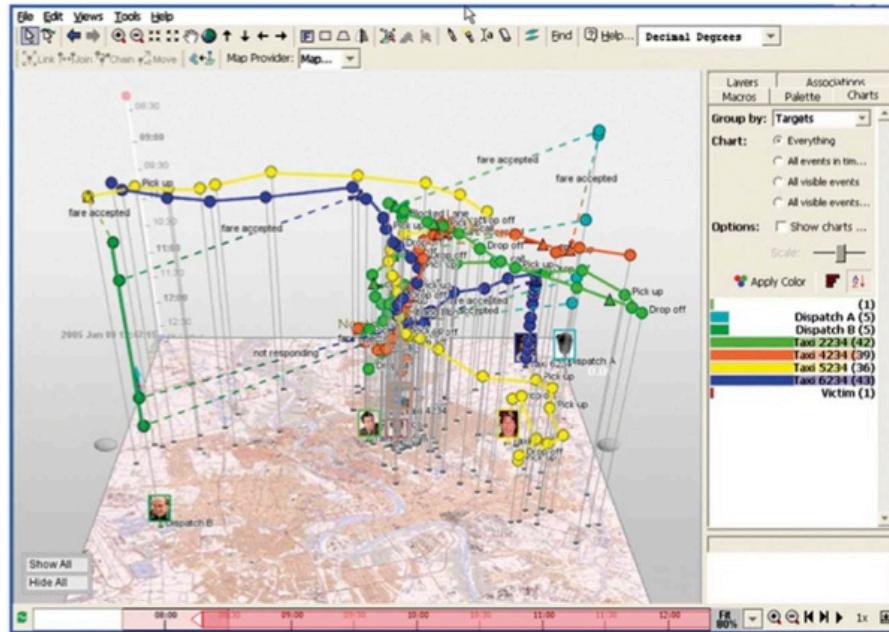
Combining Space and Time

- OculusInfo Geotime application
 - events are represented in an X,Y,T coordinate space
 - the X,Y plane shows geography
 - the vertical T axis represents time
 - events animate in time vertically through the 3-D space as the time slider bar is moved.

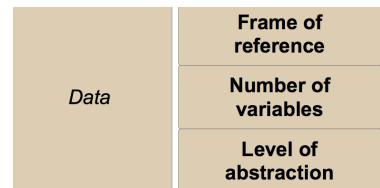


Video: <https://www.youtube.com/watch?v=P9-lpE47oWc>

Combining Space and Time



Time Points
Linear



Spatial
Multivariate
Data



Static, Dynamic
3D

Visualizing Time-oriented Data

A Systematic View

The TimeViz Browser
 A Visual Survey of Visualization Techniques for Time-Oriented Data
 by Christian Tominski and Wolfgang Aigner

of Techniques: 115

Search:

How to use filters:

- Want: Show me!
- Indifferent: I don't care.
- Hide: I'm not interested!

Data

Frame of Reference

- Abstract
- Spatial

Number of Variables

- Univariate
- Multivariate

Time

Arrangement

- Linear
- Cyclic

Time Primitives

- Instant
- Interval

Visualization

Mapping

- Static
- Dynamic



Aigner, Wolfgang, et al. "Visualizing time-oriented data—a systematic view." Computers & Graphics 31.3 (2007): 401-409.

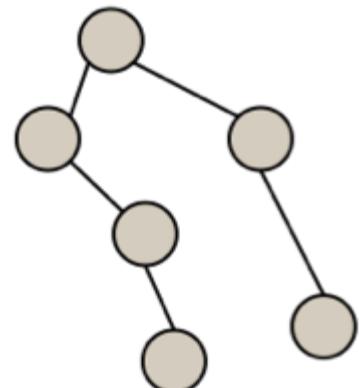
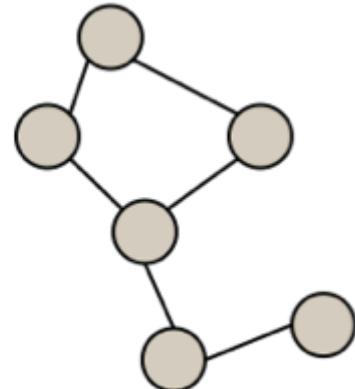
<http://www.timeviz.net/>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Visualizing Trees and Graphs

Graphs and Trees

- Graphs
 - Model relations among data
 - Nodes and edges
- Trees
 - Graphs with hierarchical structure
 - Connected graph with $N-1$ edges
 - Nodes as parents and children



Spatial Layout

- A primary concern of graph drawing is the spatial arrangement of nodes and edges.
- Often the goal is to effectively depict the graph structure:
 - Connectivity, path-following
 - Network distance
 - Clustering
 - Ordering (e.g., hierarchy level)

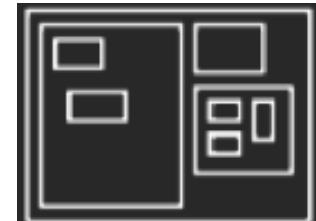
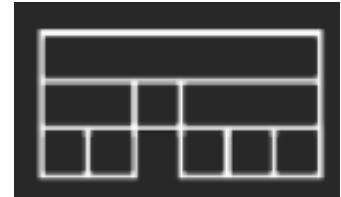
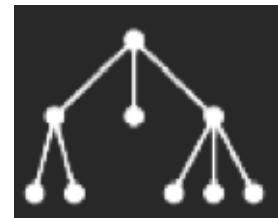
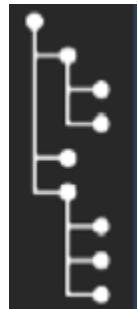
Many Applications

- Tournaments
- Organization Charts
- Biological Interactions (Genes, Proteins)
- Computer Networks
- Social Networks
- Integrated Circuit Design

Visualizing Trees

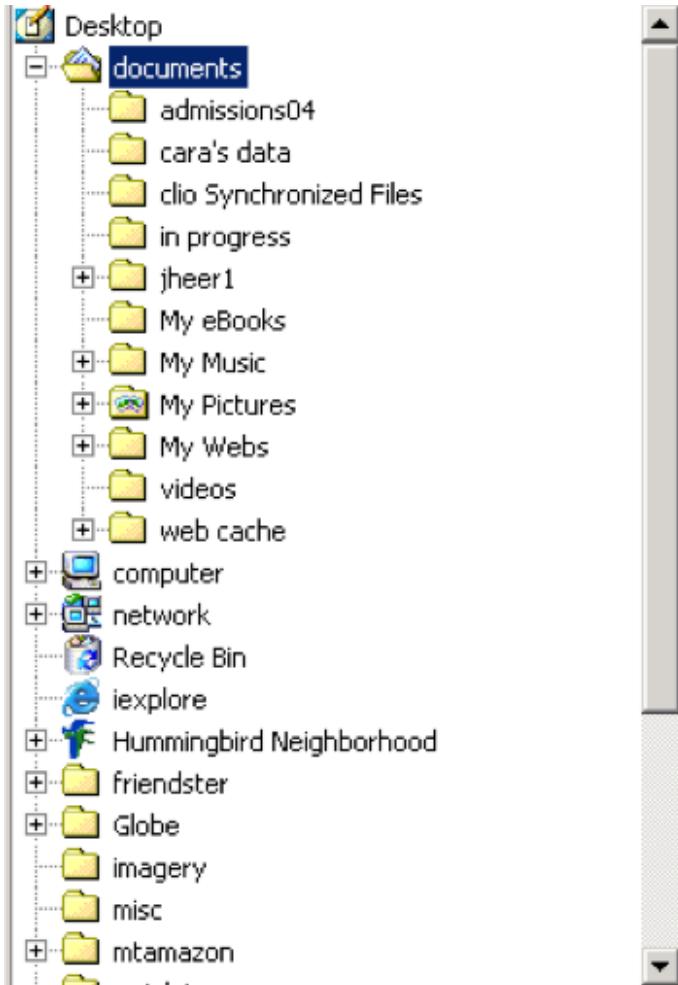
Tree Visualizations

- Indented lists
 - Linear list, indentation encodes depth
- Node-link trees
 - Nodes connected by lines/curves
- Layered diagrams
 - Relative position and alignment
- Treemaps (Enclosure diagrams)
 - Represent hierarchy by enclosure



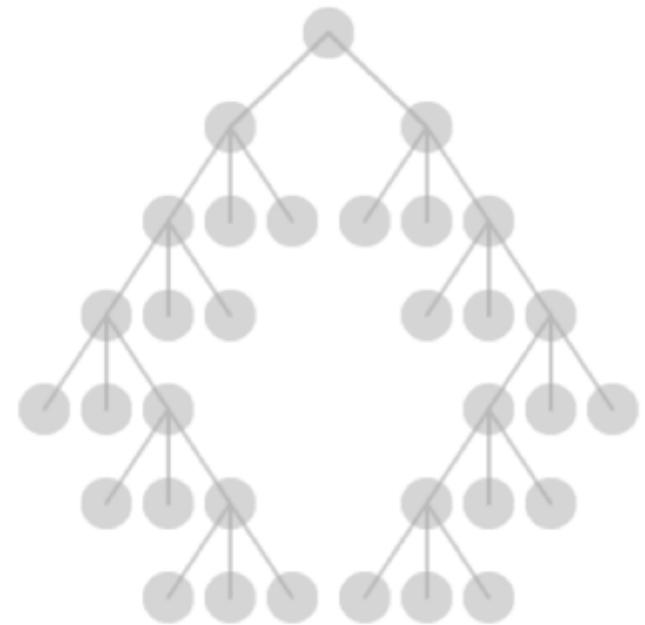
Indented List

- Places all items along vertically spaced rows
- Indentation used to show parent/child relationships
- Commonly used as a component in an interface
- Breadth and depth contend for space
- Often requires a great deal of scrolling



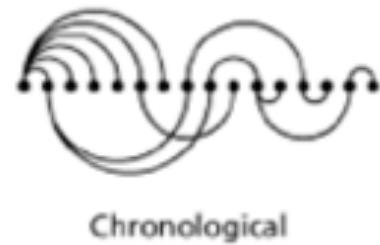
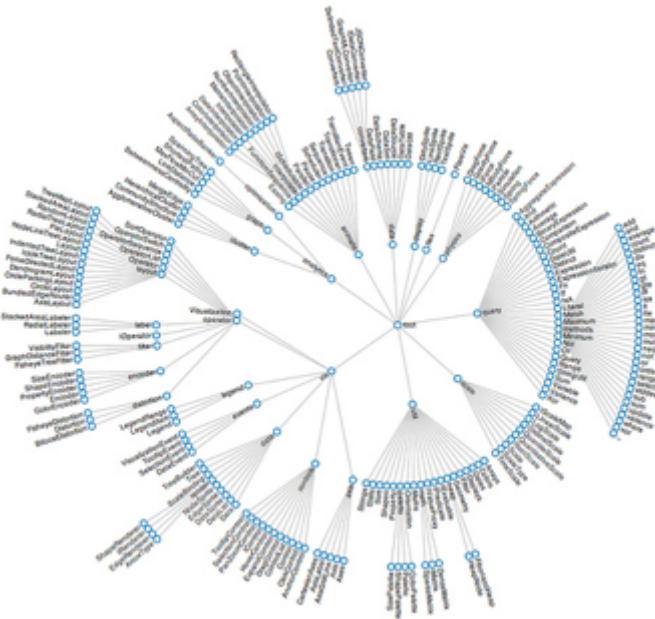
Node-Link Trees

- Nodes are distributed in space, connected by straight or curved lines.
- Typical approach is to use 2D space to break apart breadth and depth.
- Often space is used to communicate hierarchical orientation (e.g., towards authority or generality)
- Reingold-Tilford algorithm can achieve linear time in presenting a compact layout



Other Node-Link Trees

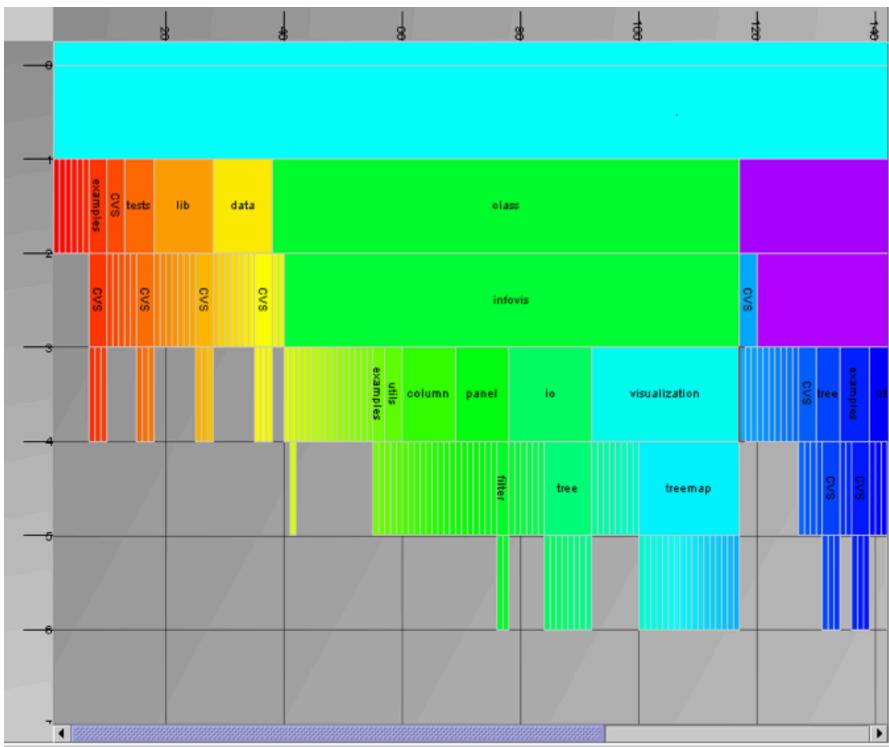
- Radial layout places the root in the center.
 - The radius encodes the depth.
- ThreadArcs
 - combine the chronology of messages with the branching tree structure in a mixed-model visualization



(Optional Reading) Kerr, Bernard. "Thread arcs: An email thread visualization." Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on. IEEE, 2003.

Layered Diagrams

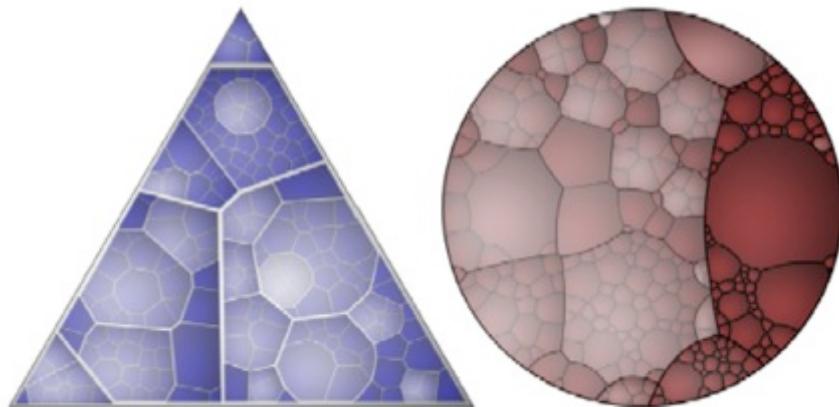
- Signify tree structure using
 - Layering
 - Adjacency
 - Alignment
- Involves recursive sub-division of space
- Higher-level nodes get a larger layer area, whether that is horizontal or angular extent.
- Child levels are layered, constrained to parent's extent



Icicle Trees

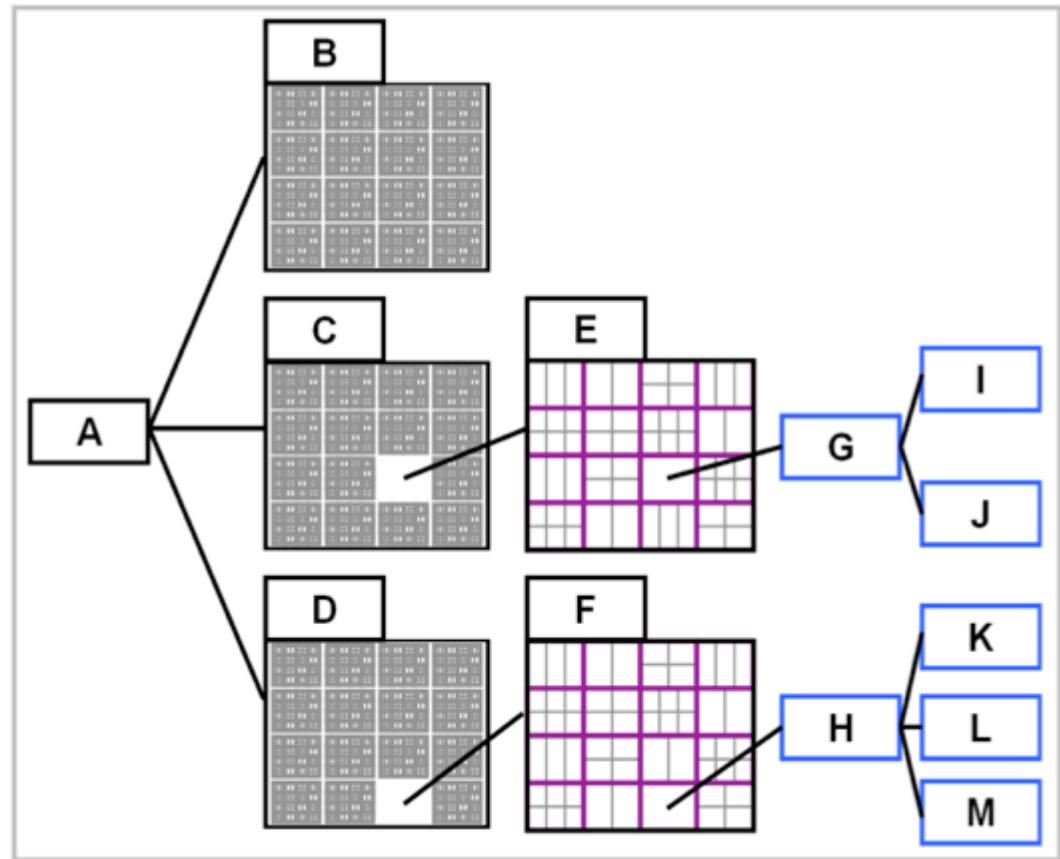
Treemaps (Enclosure Diagrams)

- Recursively fill space.
Enclosure signifies hierarchy.
- Additional measures can be taken to control aspect ratio of cells.
- Often uses rectangles, but other shapes are possible, e.g., iterative Voronoi tessellation.

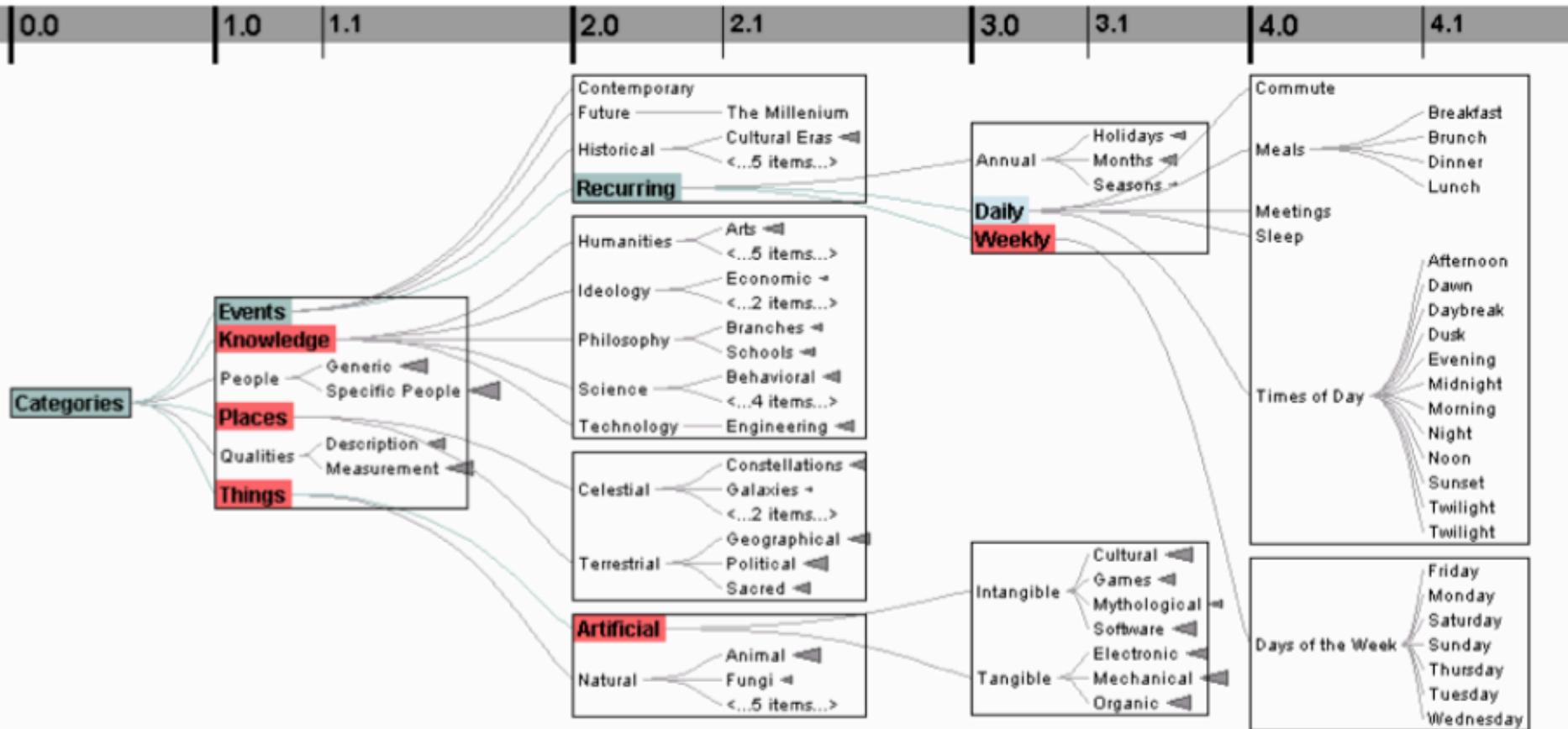


Hybrids

- Elastic Hierarchies
 - Node-link diagram with treemap nodes.
- Video:
<https://www.youtube.com/watch?v=nvslqYQ75yA>



Interactive: Degree-of-interest Trees



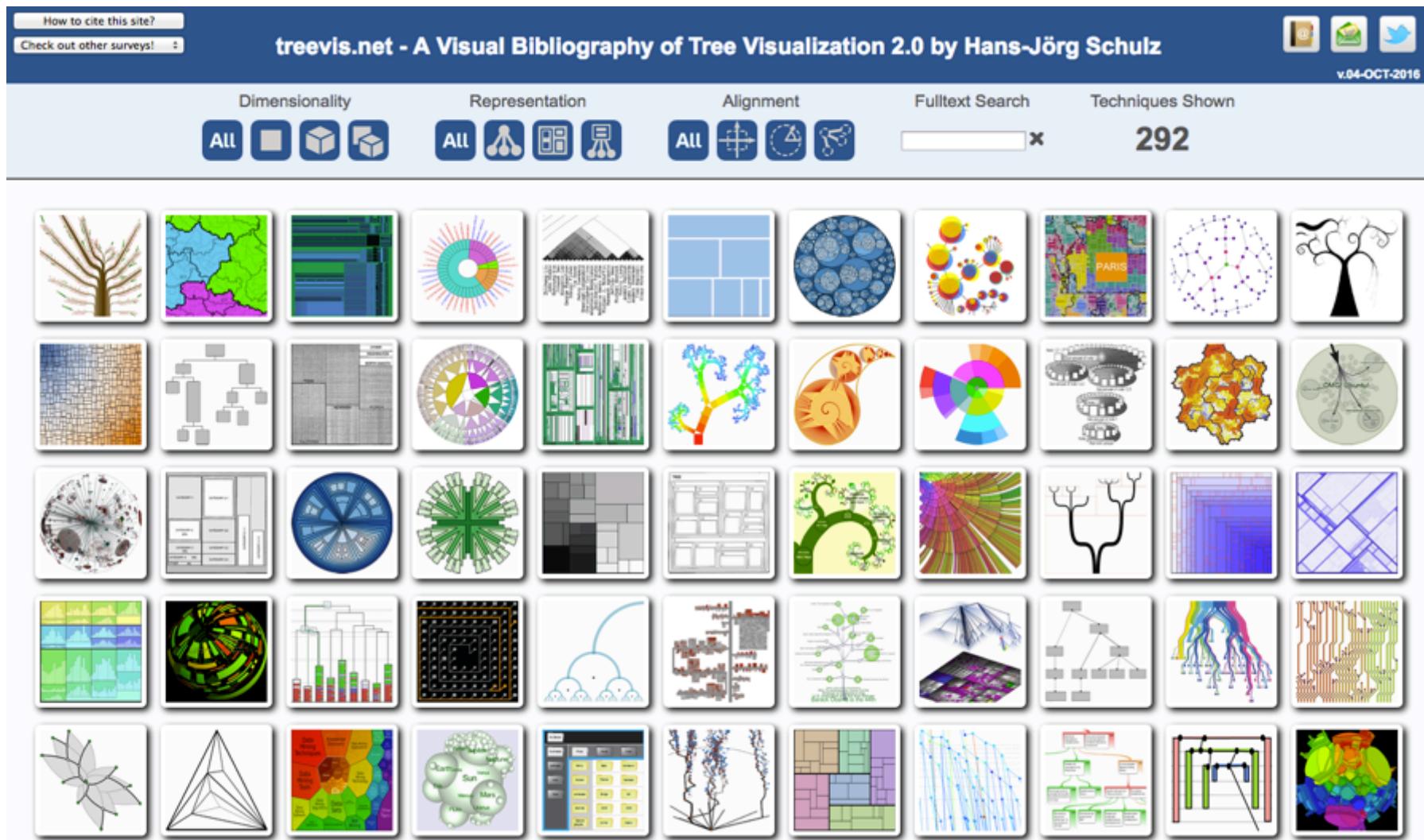
- Cull “un-interesting” nodes on a per block basis until all blocks on a level fit within bounds.
- Attempt to center child blocks beneath parents.

Treevis.net

How to cite this site?
Check out other surveys! ▾

treevis.net - A Visual Bibliography of Tree Visualization 2.0 by Hans-Jörg Schulz v.04-OCT-2016

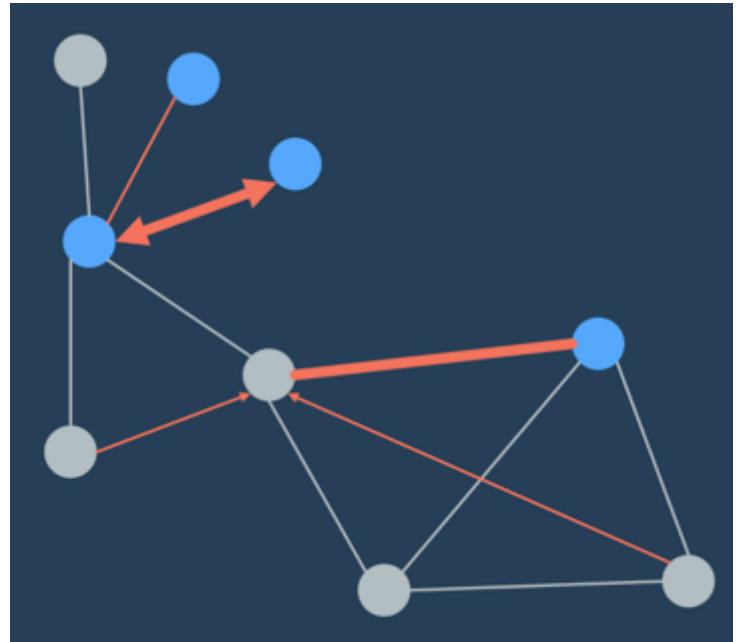
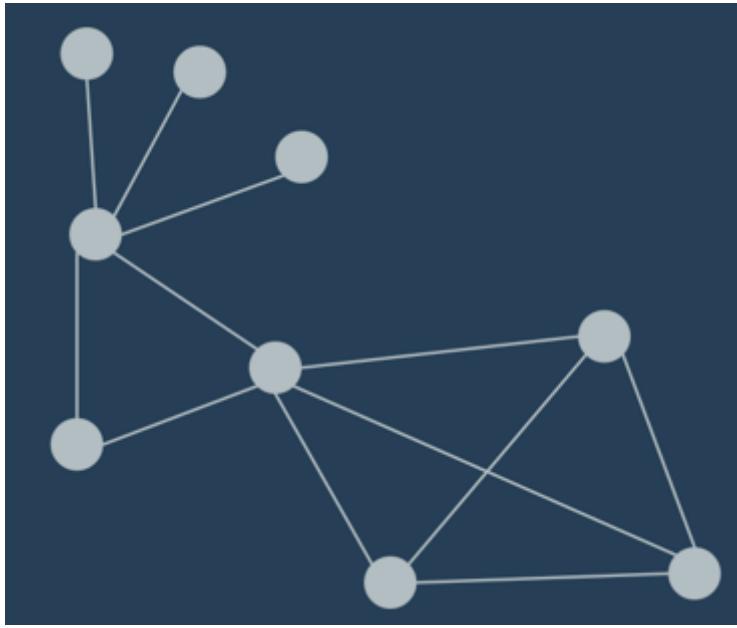
Dimensionality Representation Alignment Fulltext Search Techniques Shown 292



The website displays a collection of 292 tree visualization techniques, each represented by a small thumbnail image. The thumbnails are arranged in a grid format. The first row includes icons for citation and other surveys. The second row contains search and navigation buttons. The third row shows the total count of techniques. The remaining rows are filled with diverse visualizations such as hierarchical trees, sunburst charts, treemaps, and network graphs.

Visualizing Graphs

What's in a Graph?

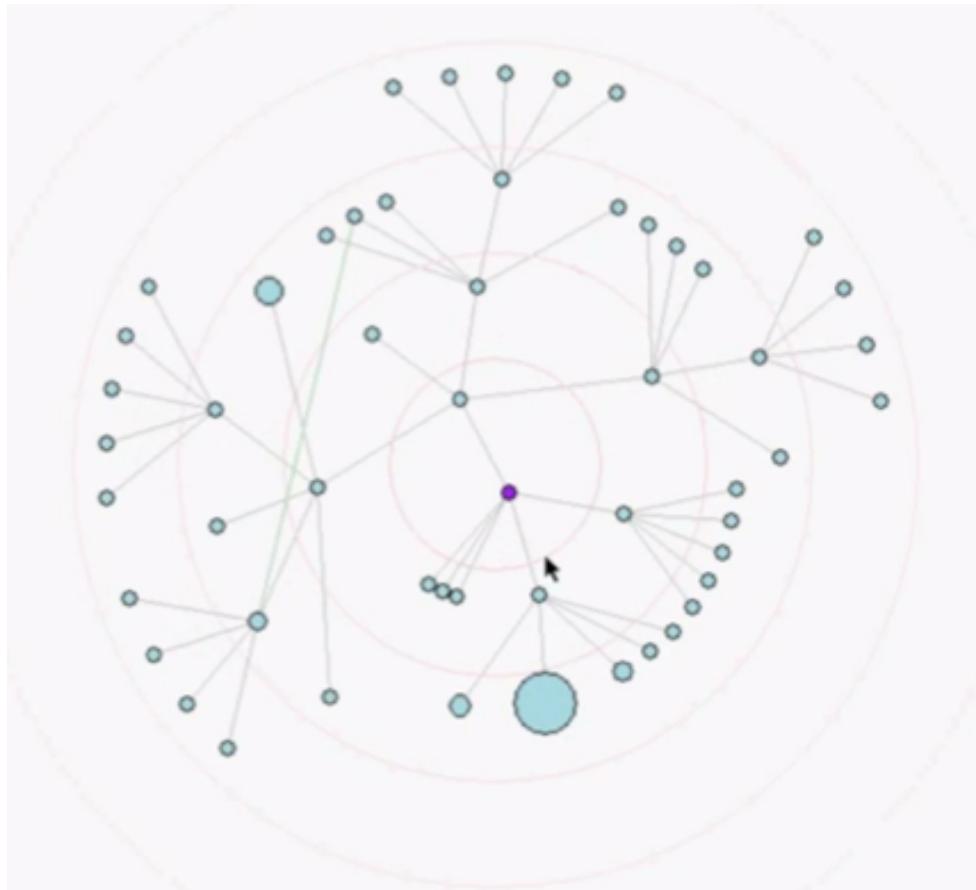


Graph Visualization

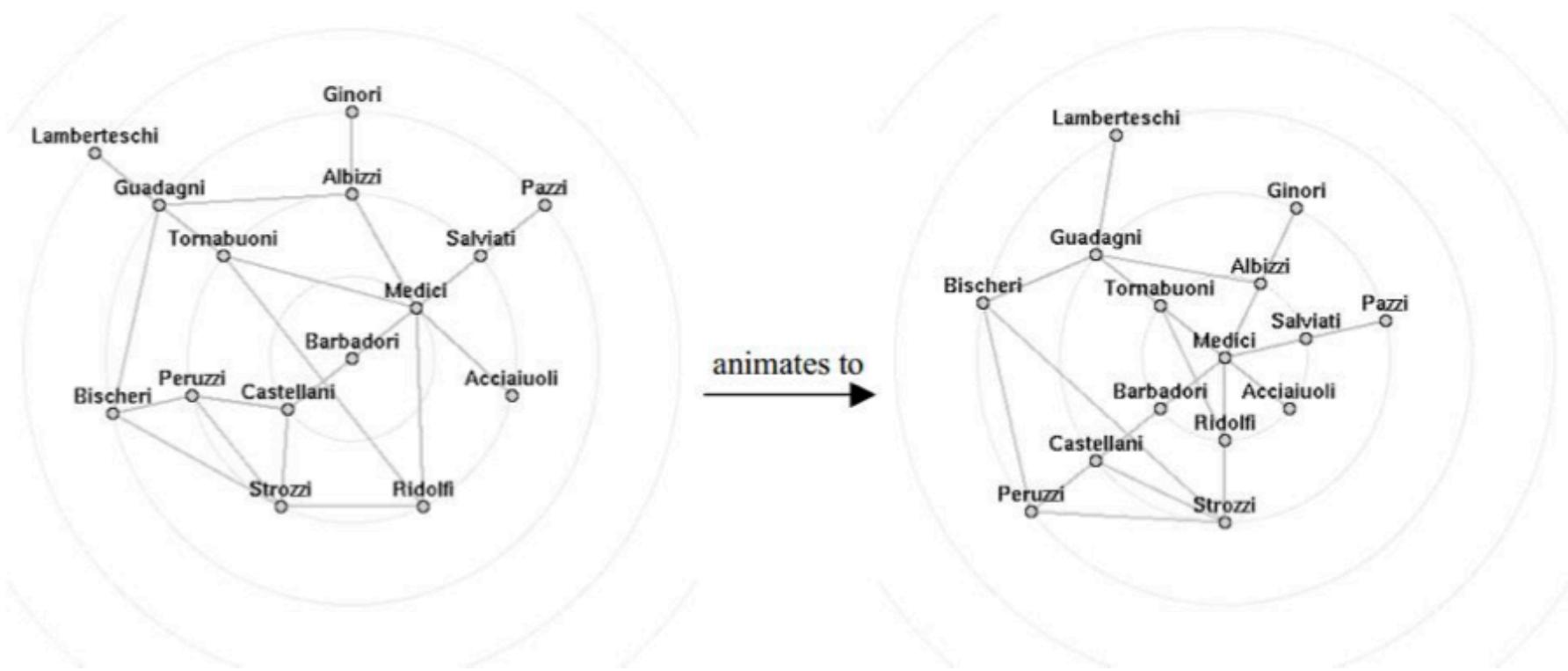
- Two representations:
 - Node-link diagrams
 - Matrices

Tree in the Graph

- Many graphs are tree-like or have useful spanning trees
- Spanning trees lead to arbitrary roots
- Fast tree layouts allow graph layouts to be recalculated at interactive rates

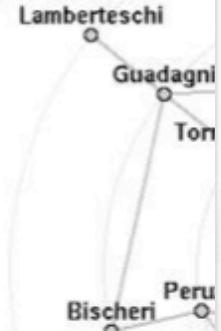


Tree in the Graph



animates to

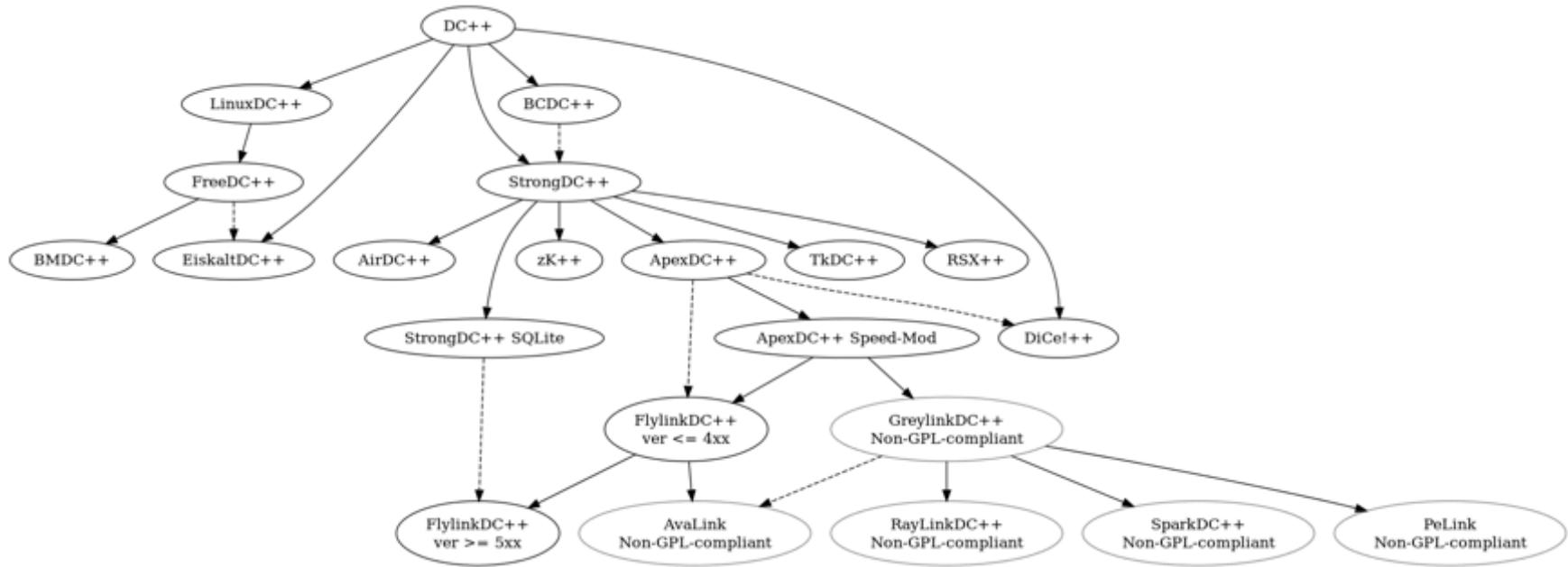
- Animated Graphs with Radial Layout
- Video: http://www.youtube.com/watch?v=OPX5iGro_IA



Lamberteschi
Guadagni
Torri
Bischeri
Peru
Pazzi
i

- Anir
- Vide

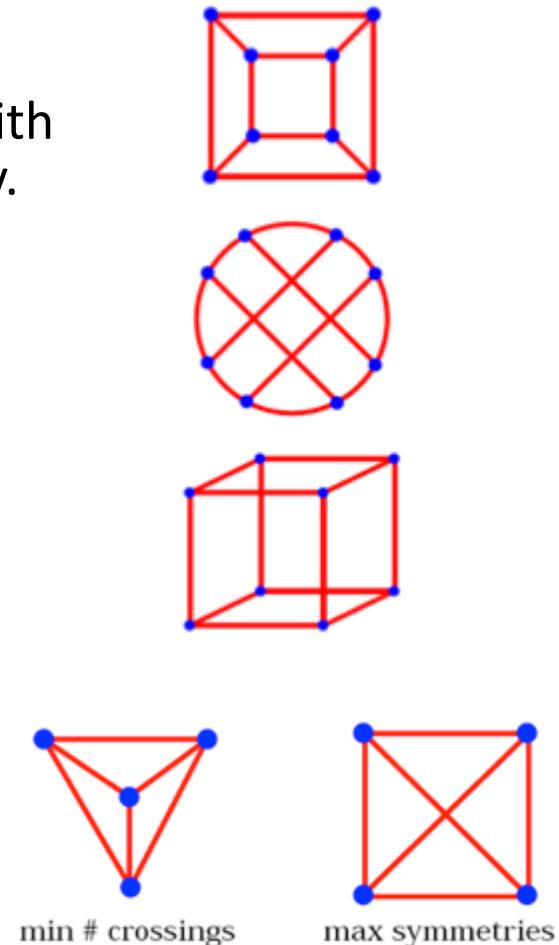
Hierarchical Graph Layout



- Evolution of the DC++ tool
- Layered graph drawing
- Layout of a Direct Acyclic Graph
- Hierarchical layering based on descent

Optimization Techniques

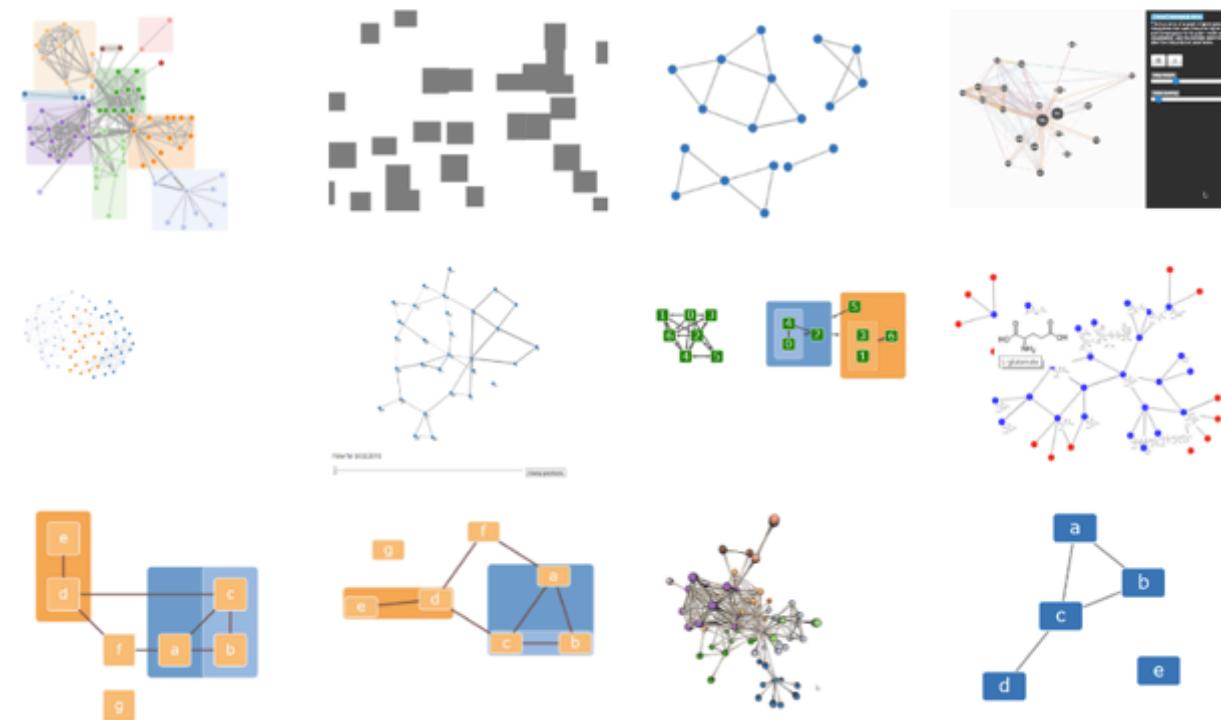
- Treat layout as an optimization problem
 - Define layout using an energy model along with constraints: equations the layout should obey.
 - Use optimization algorithms to solve
- Commonly posed as a physical system
 - Charged particles, springs, drag force, ...
- Different constraints can be introduced
 - Minimize edge crossings
 - Minimize area
 - Minimize line bends
 - Minimize line slopes
 - Maximize smallest angle between edges
 - Maximize symmetry



Constraint-based Optimization and Layout

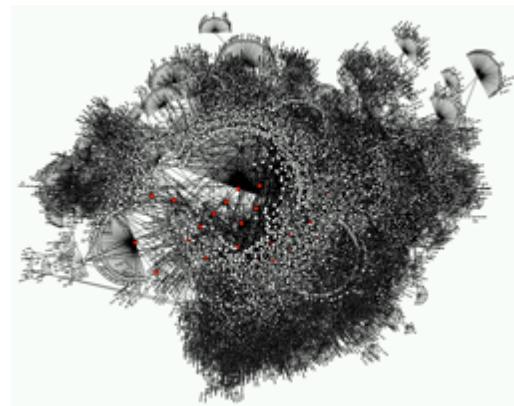
cola.js

Constraint-Based Layout in the Browser

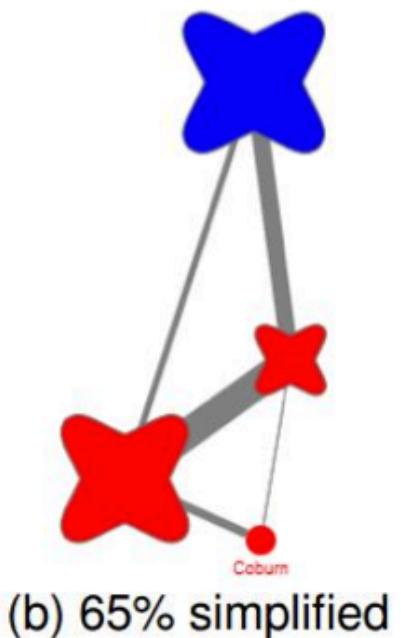
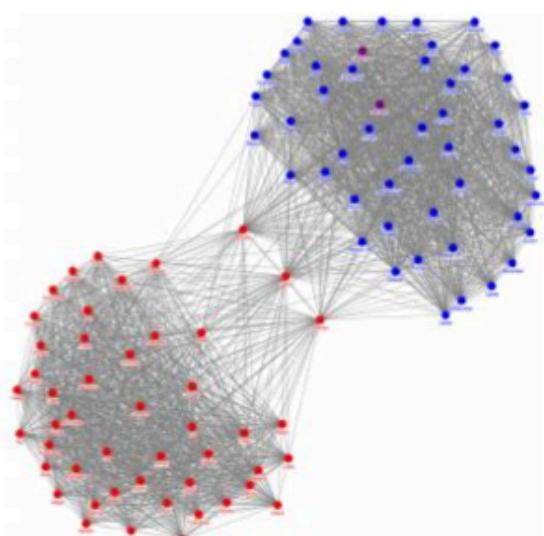
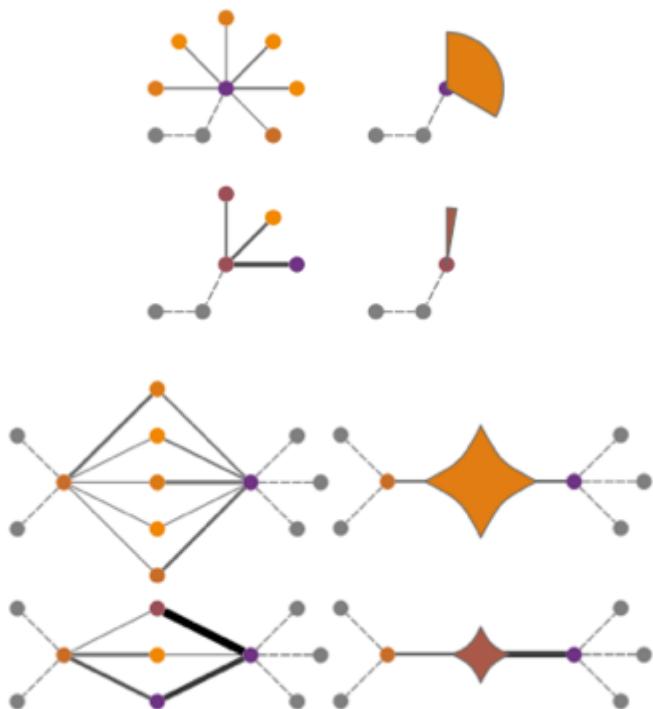


Scalability Issue

- Need to cope with messiness
- Solutions
 - Extracting network motifs
 - Taking advantage of node attributes
 - Degree-of-Interest graphs
 - Use the alternative representation: matrix



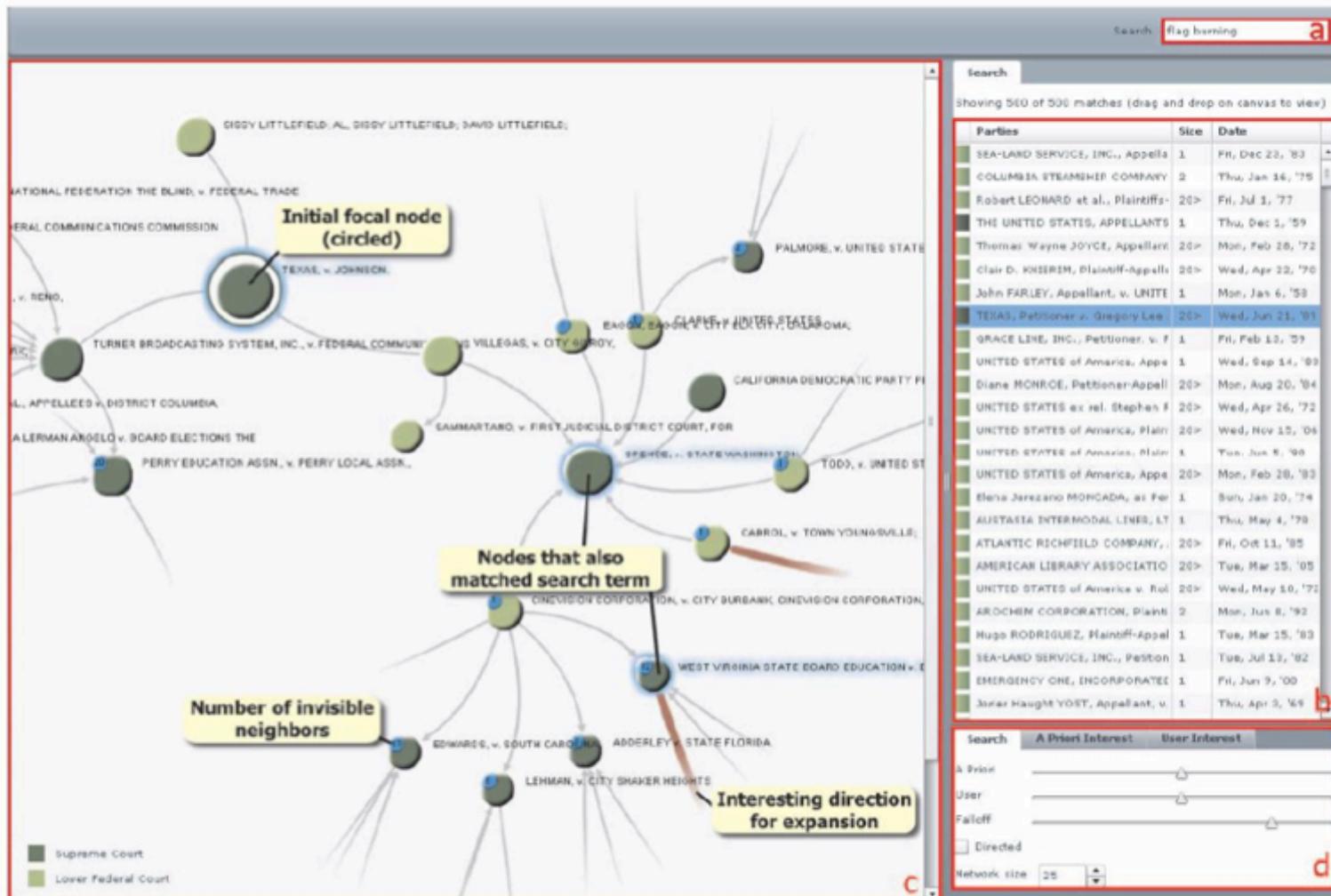
Motifs



Dunne, Cody, and Ben Shneiderman. "Motif simplification: improving network visualization readability with fan, connector, and clique glyphs." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

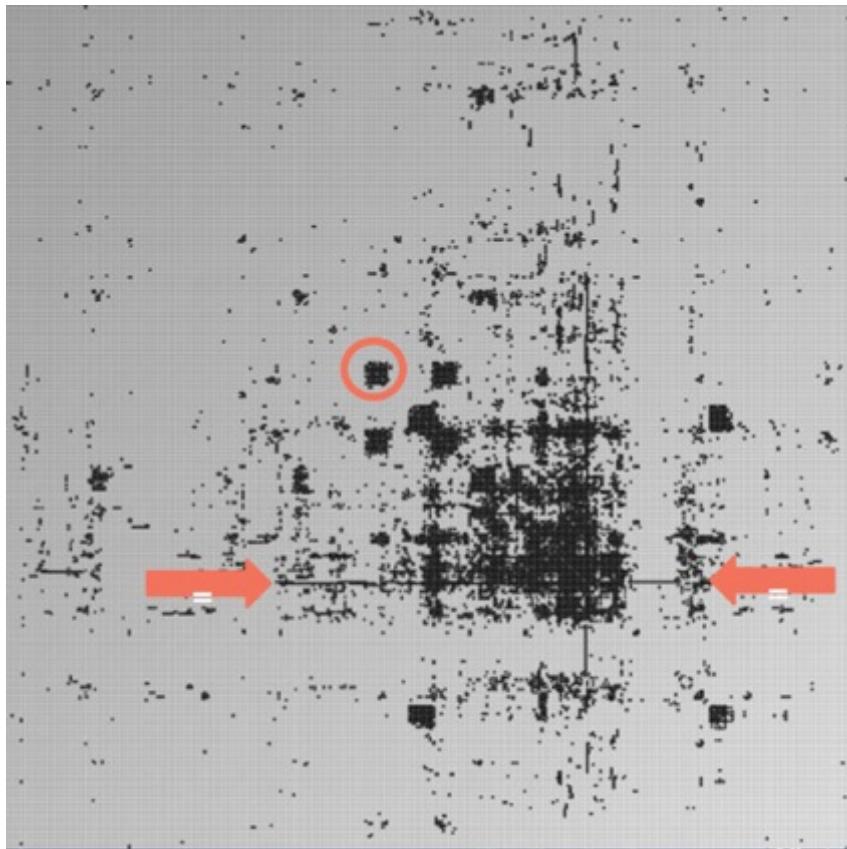
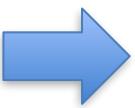
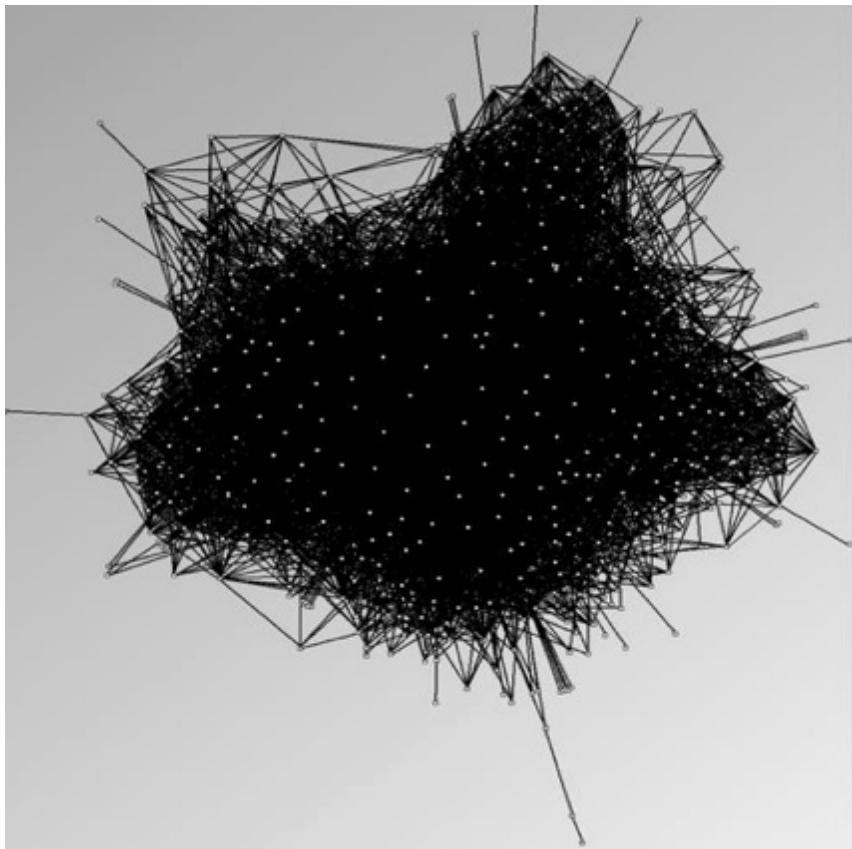
Interactive Degree-of-Interest Graphs



Van Ham, Frank, and Adam Perer. "Search, show context, expand on demand": supporting large graph exploration with degree-of-interest." IEEE Transactions on Visualization and Computer Graphics 15.6 (2009).

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

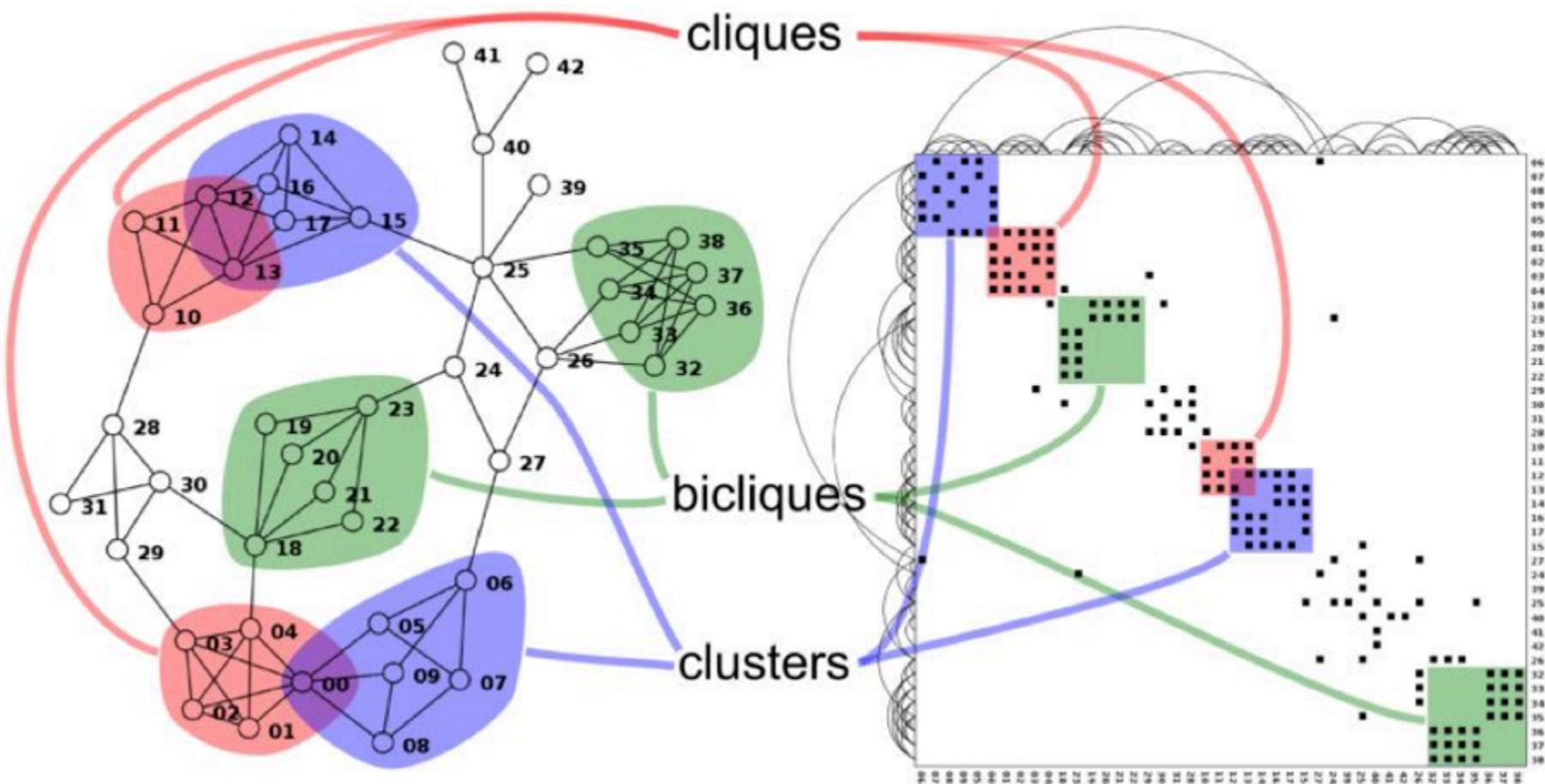
Matrices



Matrix vs. Node-link

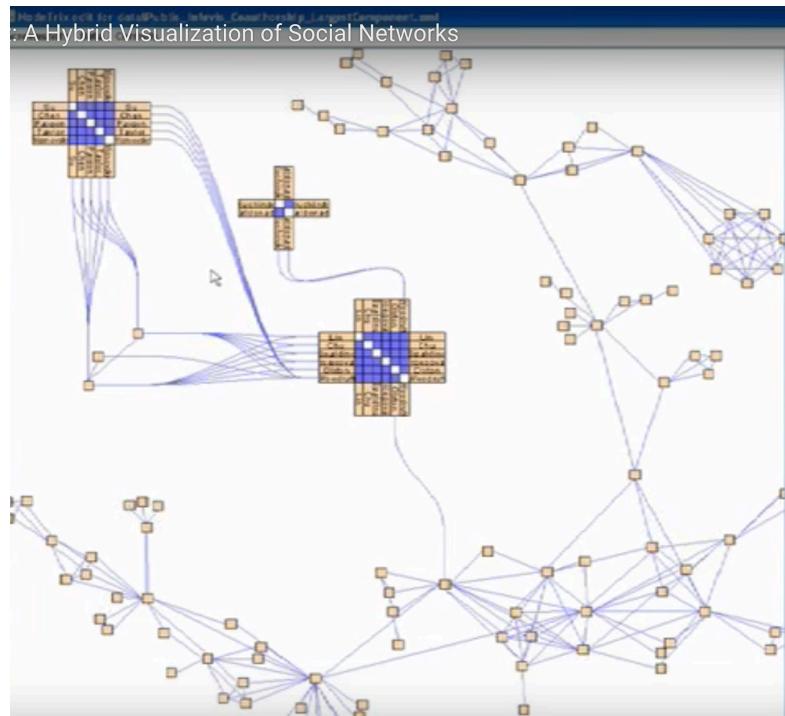
Matrix	Node-link
Require learning	Familiar
No overlap	Node overlap
No crossings	Link crossing
Use a lot of space	More compact
Dense graphs	Sparse graphs

Node Link to Matrix



Hybrid

- Merging Node-link with Matrices is possible.



Video: <https://www.youtube.com/watch?v=7G3MxyOcHKQ>

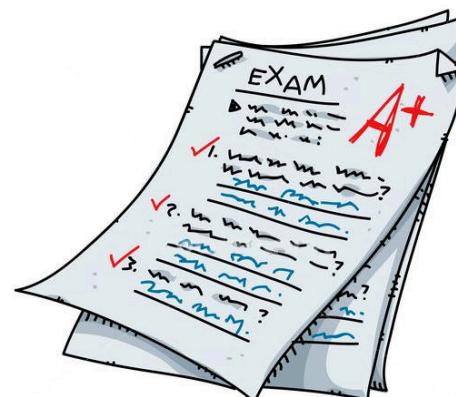
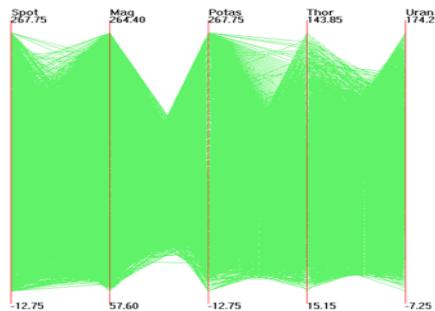
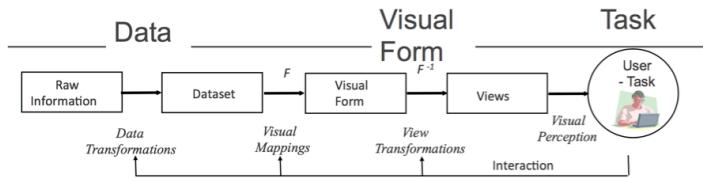
Henry, Nathalie, Jean-Daniel Fekete, and Michael J. McGuffin.
"NodeTrix: a hybrid visualization of social networks." IEEE transactions
on visualization and computer graphics 13.6 (2007): 1302-1309.

Summary

- Visualizing Tree Layout
 - Indented / Node-Link / Enclosure / Layers
- Visualizing Graph Layout
 - Tree in graph / Hierarchical graph layout
 - Layout Optimization
 - Scalability issue: motif, degree of interests, matrix
 - Matrix

Next Lecture

- Topic:
 - Recap of Fundamentals
 - Exam Review
- Next Monday (11 Mar)
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus



G53FIV: Fundamentals of Information Visualization

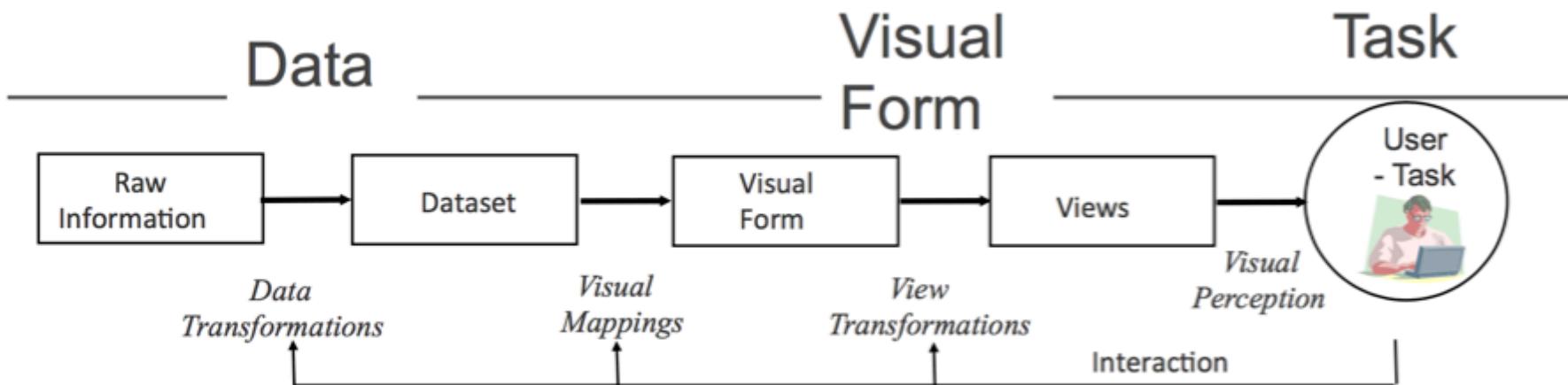
Lecture 13: Recap of Fundamentals

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

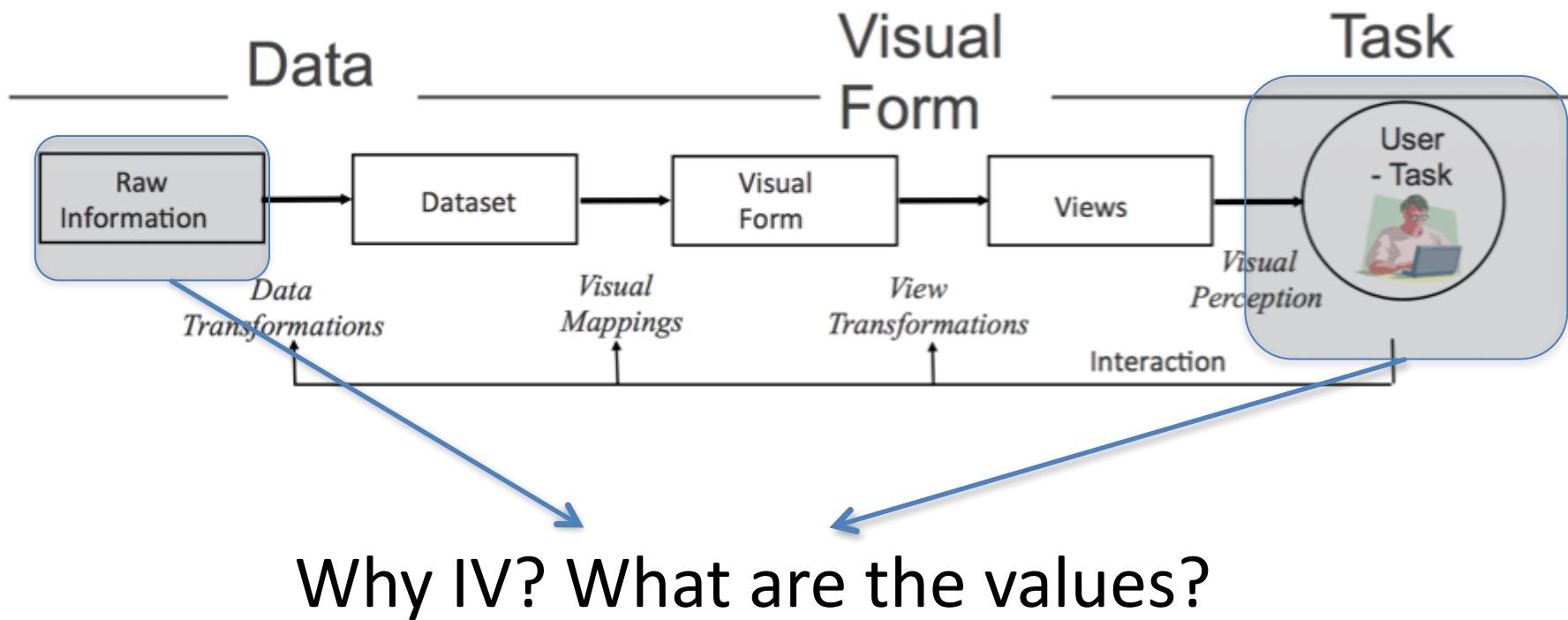
<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

A Recap of Fundamentals

Information Visualization



Information Visualization



Information Overload



Objective

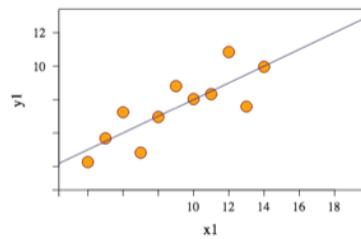
- Transform the data into information (understanding, insight) thus making it useful



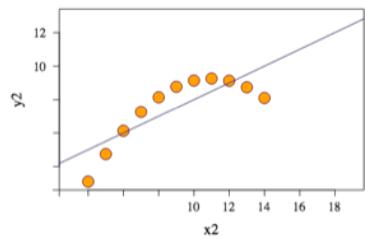
Anscombe's Quartet

	Set A		Set B		Set C		Set D	
	X	Y	X	Y	X	Y	X	Y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
corr	0.82		0.82		0.82		0.82	
lin. reg.	$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$	

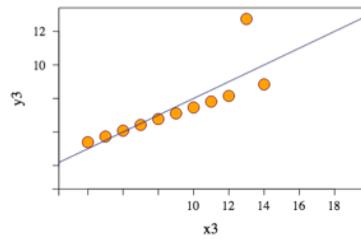
"what most people would see in their mind's eye [for a linear relationship with some unexplained variation]"



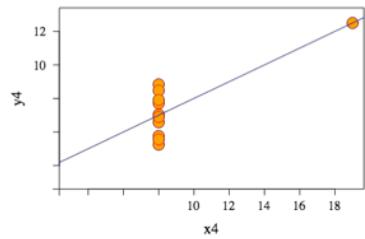
"y has a smooth curved relation with x, possibly quadratic, and there is little residual variability"



"all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation)"



"all the information about the slope of the regression line resides in one observation"



[Anscombe, 1973]

The Best of Both Sides

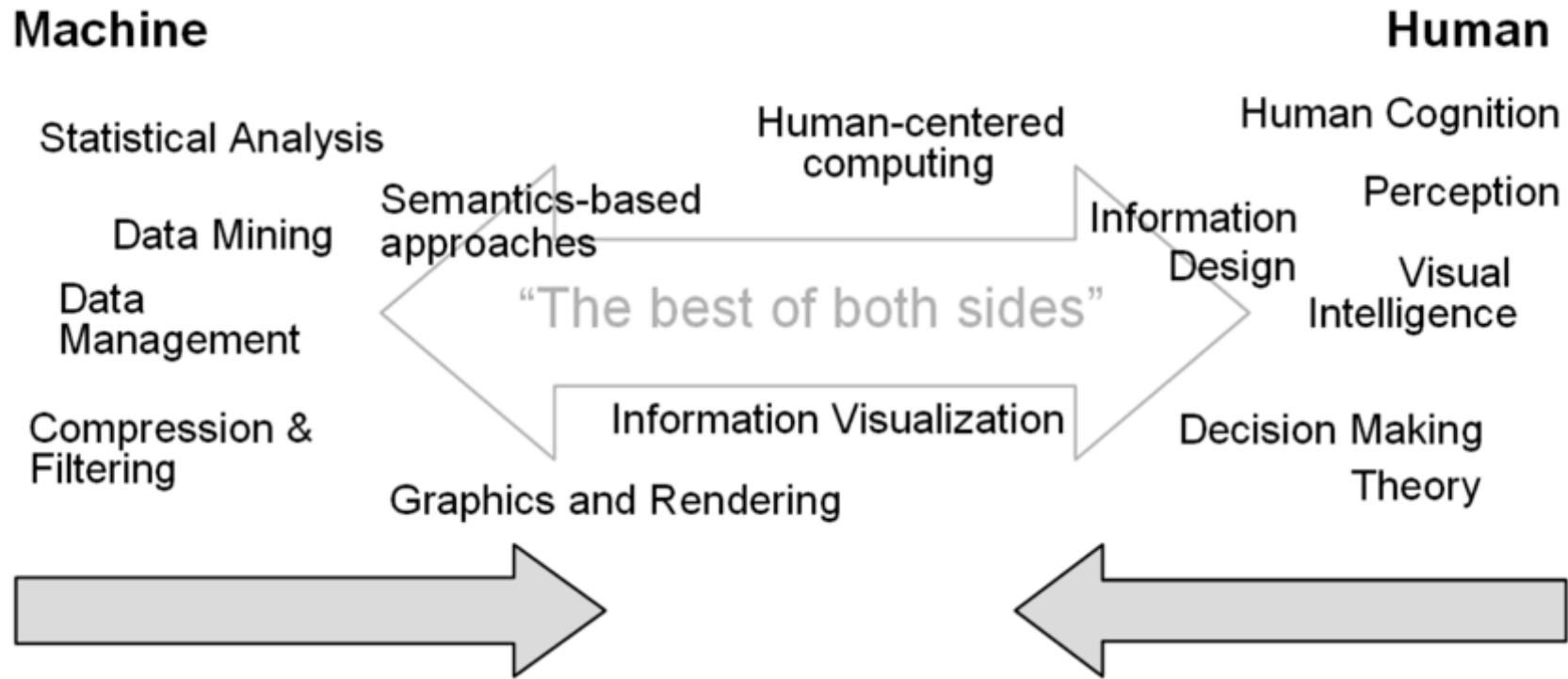


Fig. 2. Visual analytics integrates scientific disciplines to improve the division of labor between human and machine.

Key Values of Visualizations

- **Record** information
 - Blueprints, photographs, seismographs, ...
- **Communicate** information to others
 - Share and persuade
 - Collaborate and revise
- Analyze data to **support reasoning**
 - Find patterns / Discover errors in data
 - Expand memory
 - Develop and assess hypotheses

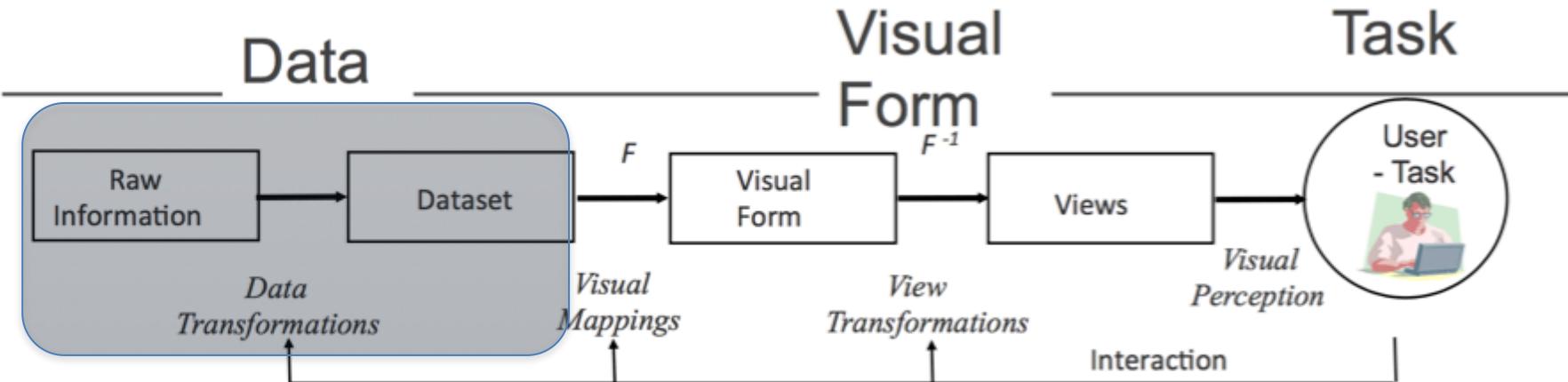


Data and Initial Tasks: Before the Actual Visualization

- Pick a dataset of your interest
- Pose the initial questions/tasks that you would like to answer/accomplish
- Assess the fitness of the data
- **Visualization**
- Refine your questions/tasks

Remember our course work and the house price case study?

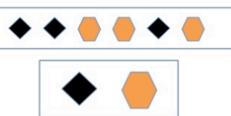
Information Visualization



00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00544	+40.922326	-072.637078	U	HOLTSVILLE	36	103
00601	+18.165273	-066.722583		ADJUNTAS	72	001
00602	+18.493103	-067.180953		AGUADA	72	003
00603	+18.455913	-067.145780		AGUADILLA	72	005
00604	+18.493520	-067.135883		AGUADILLA	72	005
00605	+18.465162	-067.141486	P	AGUADILLA	72	005
00606	+18.172947	-066.944111		MARICAO	72	093
00610	+18.288685	-067.139696		ANASCO	72	011



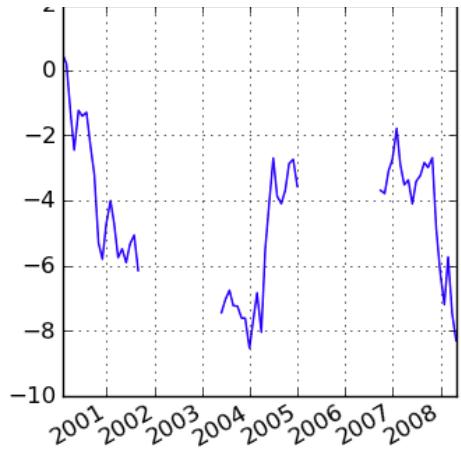
- Data transformation:** create a structural model (schema), mapping raw data into data tables

- FILTER Rows 
- SELECT Column Types 
- ArRANGE Rows (SORT) 
- Mutate (into something new) 
- Summarize by Groups 



Data Processing

- Data cleaning and filtering
 - for quality control
 - Remove (Outlier, missing data)
 - Modify (conversion of format, etc.)
- Data adjustment
 - Depends on your task and questions to ask
 - Relational algebra:
 - e.g. Aggregation, mean, sort, projection
 - Reformatting and Integration



R is a tool for...

Data Manipulation

- connecting to data sources
- slicing & dicing data

Modeling & Computation

- statistical modeling
- numerical simulation

Data Visualization

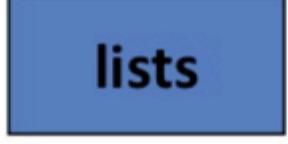
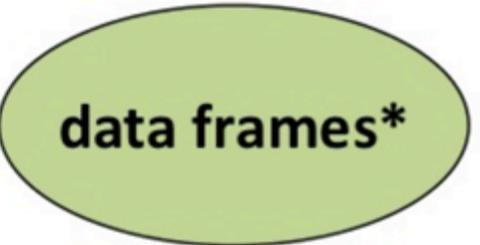
- visualizing fit of models
- composing statistical graphics

munge

model

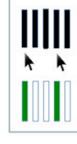
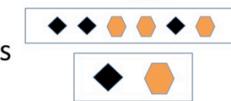
visualize

R Data Structures

	Linear	Rectangular
Homogeneous	? 	
Heterogeneous		

dplyr

- dplyr takes the `%>%` operator and uses it to great effect for manipulating data frames
 - Works only with data frames
 - 5 basic “verbs” work for 90% of data manipulations manipulations

Verbs	What does it do?	
<code>filter()</code>	Select a subset of ROWS by conditions	
<code>arrange()</code>	Reorders ROWS in a data frame	
<code>select()</code>	Select the COLUMNS of interest	
<code>mutate()</code>	Create new columns based on existing columns (mutations!)	
<code>summarise()</code>	Aggregate values for each group, reduces to single value	

Pipe Operator

- **library(magrittr)**
 - A R package launched on Jan 2014
 - A “magic” operator called the PIPE was introduced
 - `%>%`
 - i.e. “AND THEN”, “PIPE TO”

```
round(sqrt(1000), 3)  
  
library(magrittr)  
1000 %>% sqrt %>% round()  
1000 %>% sqrt %>% round(., 3)
```

Take 1000, and then its sqrt
And then round it



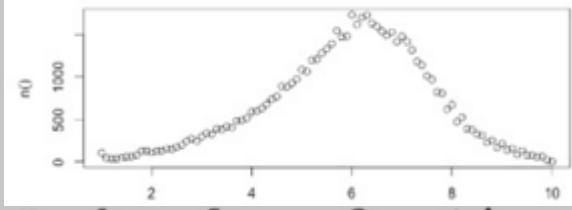
Chain the “Verbs” Together

- Chain them together

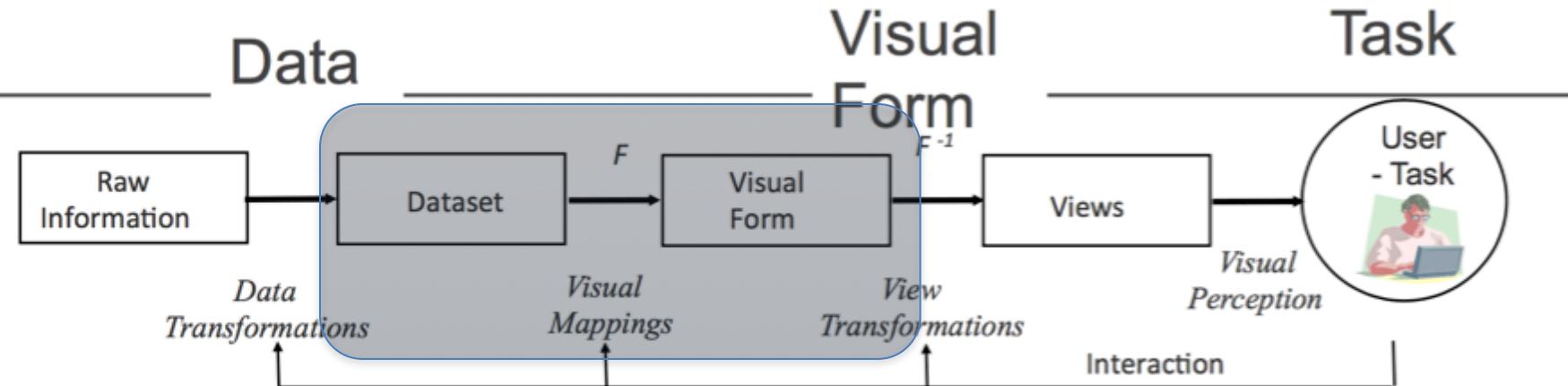
```
producers_nightmare <-  
  filter(movies_df, !is.na(budget)) %>%  
  mutate(costPerMinute = budget/length) %>%  
  arrange(desc(costPerMinute)) %>%  
  select(title, costPerMinute)
```

- Can also be fed to a “plot” command

```
movies %>%  
  group_by(rating) %>%  
  summarize(n()) %>%  
  plot() # plots the histogram of movies by Each value of rating
```



Information Visualization

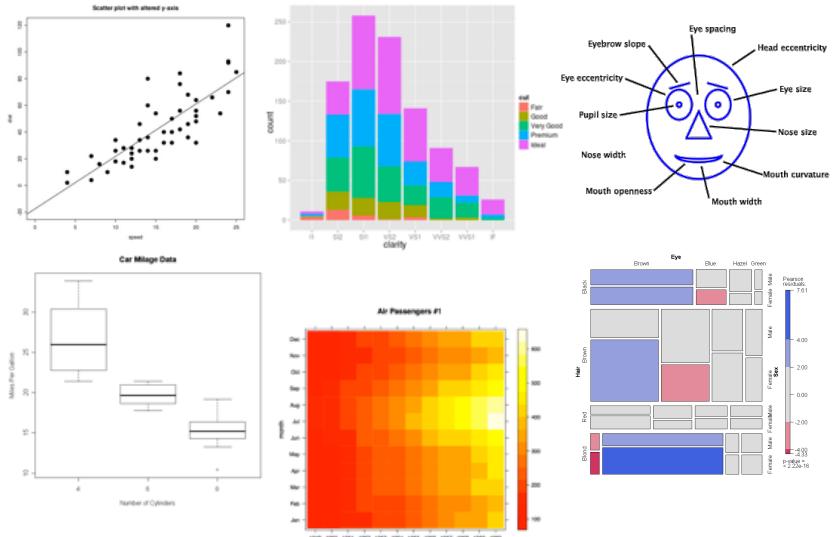


00210	+43.005895	-071.013202	U	PORTRSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTRSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTRSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTRSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTRSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTRSMOUTH	33	015

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good
 ~ = OK
 ✗ = Bad

Visual mapping: create a visual spatial model, transforming data tables into visual structures



Data, Image and Design

Data Models

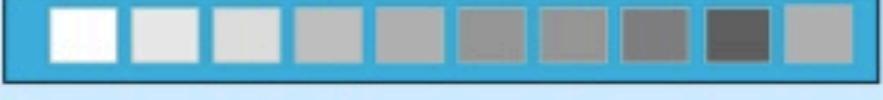
- Nominal, Ordinal, Quantitative?
- Dimension or Measure?

– Year	Q-Internal (O)	Dimension
– Age	Q-Ratio (O)	Depends
– Marital	N	Dimension
– Sex	N	Dimension
– People	Q-Ratio	Measure

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080

Image: Visual Encoding Variables

Bertin's Semiology of Graphics (1967)

- **position**
 - changes in the x, y, (z) location
- **size**
 - change in length, area or repetition
- **shape**
 - infinite number of shapes
- **value**
 - changes from light to dark
- **orientation**
 - changes in alignment
- **colour**
 - changes in hue at a given value
- **texture**
 - variation in pattern
- **motion**

Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Levels of Organization

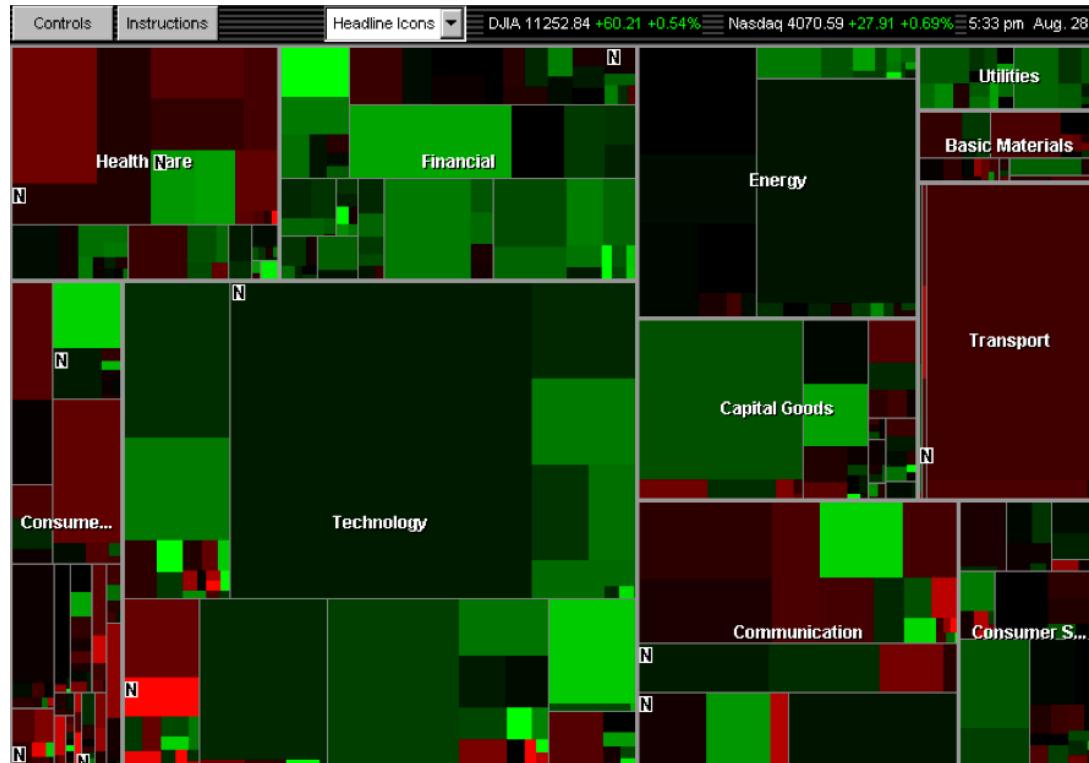
	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good

~ = OK

✗ = Bad

Example: Map of the Market

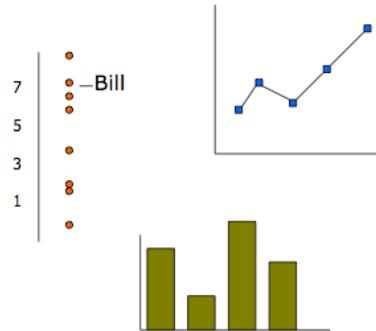


- Rectangle area: market cap (Q);
- Rectangle position: market sector (N)
- Color Hue: loss vs. gain (N, O)
- Color Value: magnitude of loss or gain (Q)

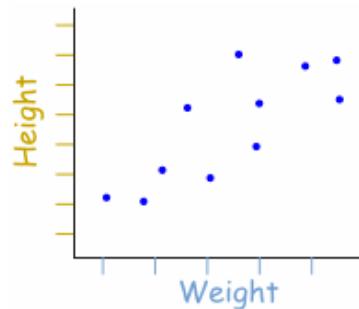
Graphs and Charts

Graphs

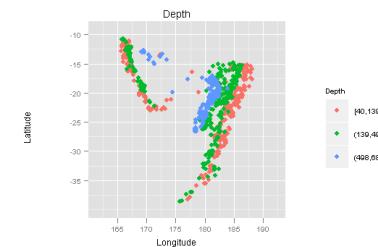
- Data Dimensions
 - 1 - Univariate data
 - 2 - Bivariate data
 - 3 - Trivariate data
 - >3 - Hypervariate data
- Data Types
 - Nominal, Ordinal, Quantitative
- Visualization Representations
 - Points, Lines, Bars, Boxes



Univariate



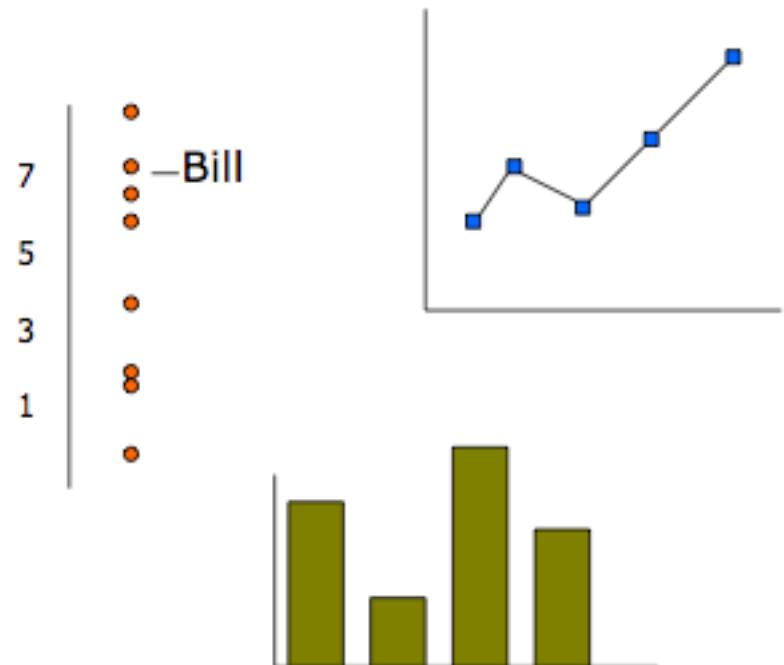
Bivariate



Trivariate

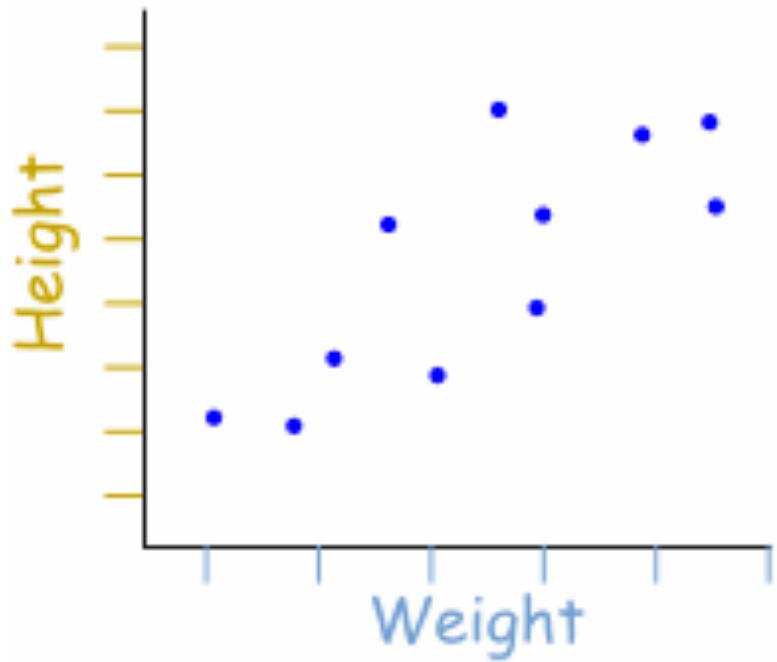
Univariate Data

- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another.
- Statistical view
 - Independent variable on x-axis (data case)
 - Track dependent variable along y-axis (value)



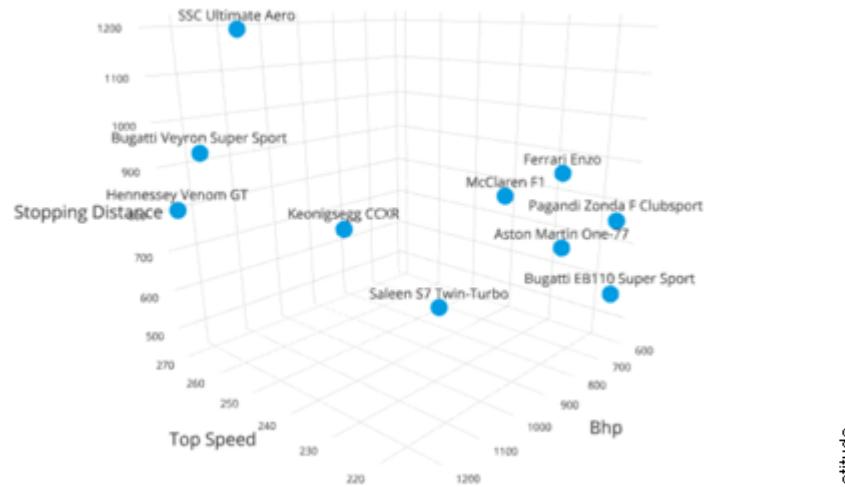
Bivariate Data

- Scatter plot is commonly used
- Each mark is now a data case
- Objective:
 - Two variables, want to see relationship
 - Is there a linear, curved or random pattern?

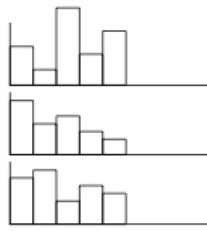
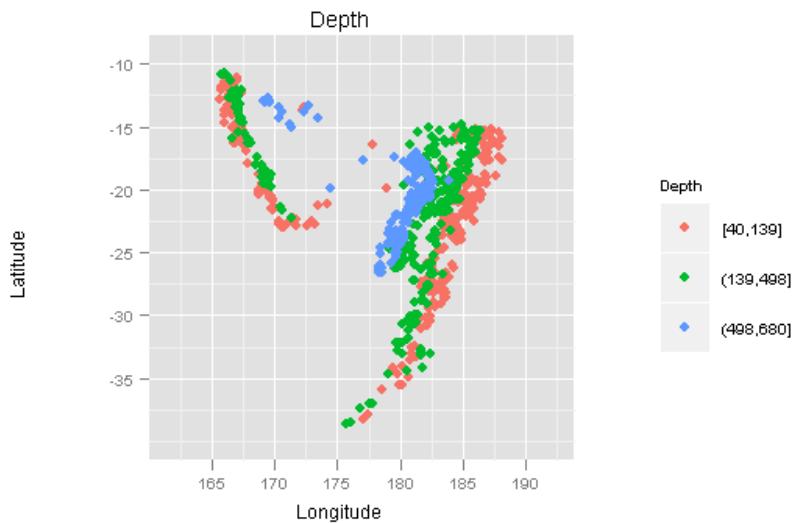


Trivariate Data

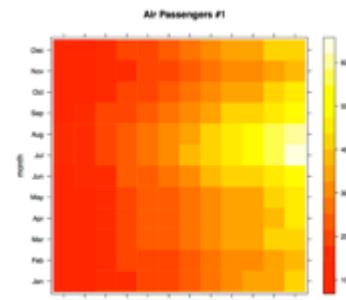
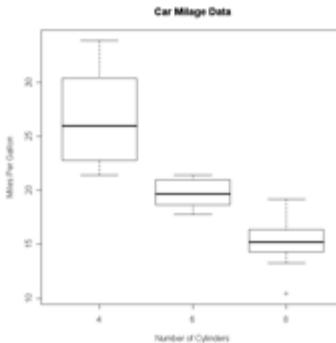
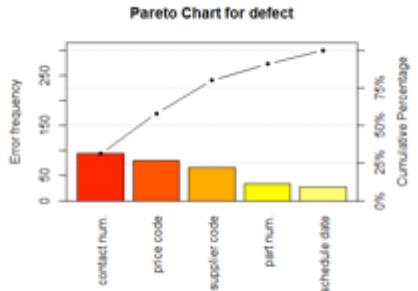
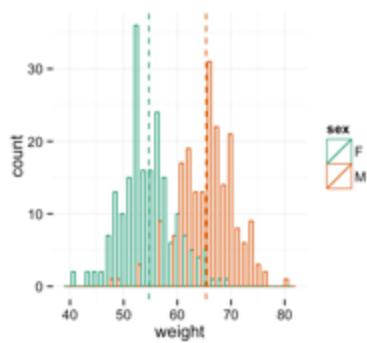
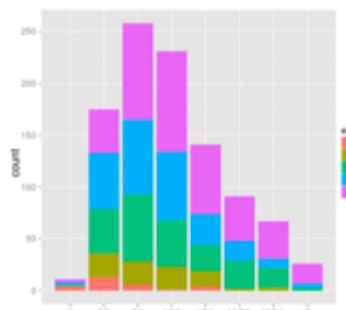
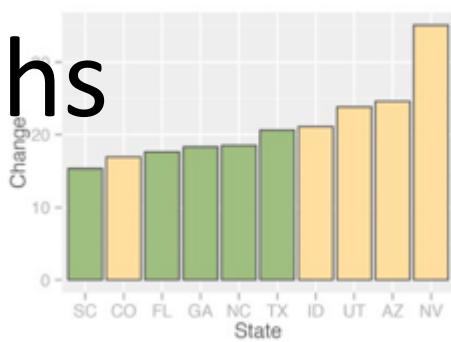
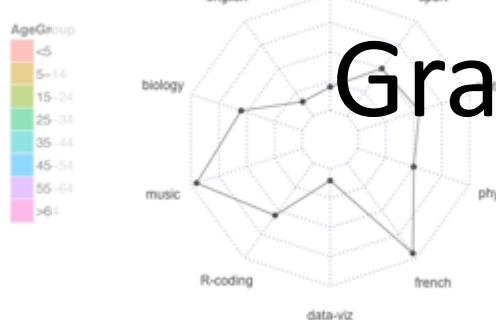
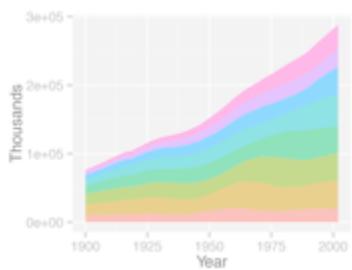
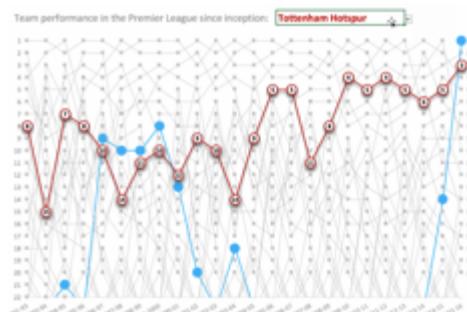
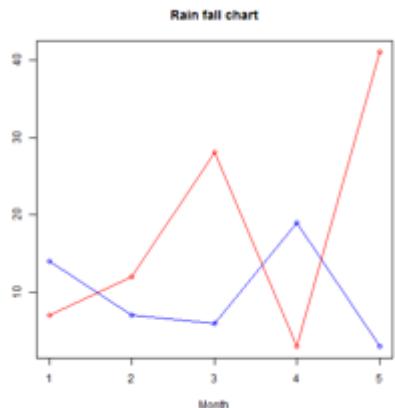
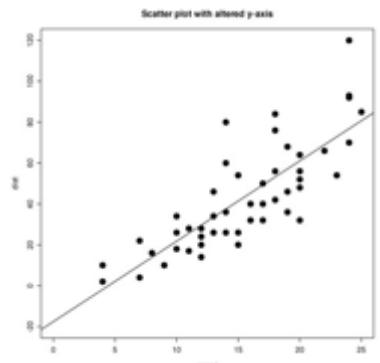
3D scatter plot



2D + mark
property



Represent each
variable in its own
explicit way



Multivariate Data Visualization

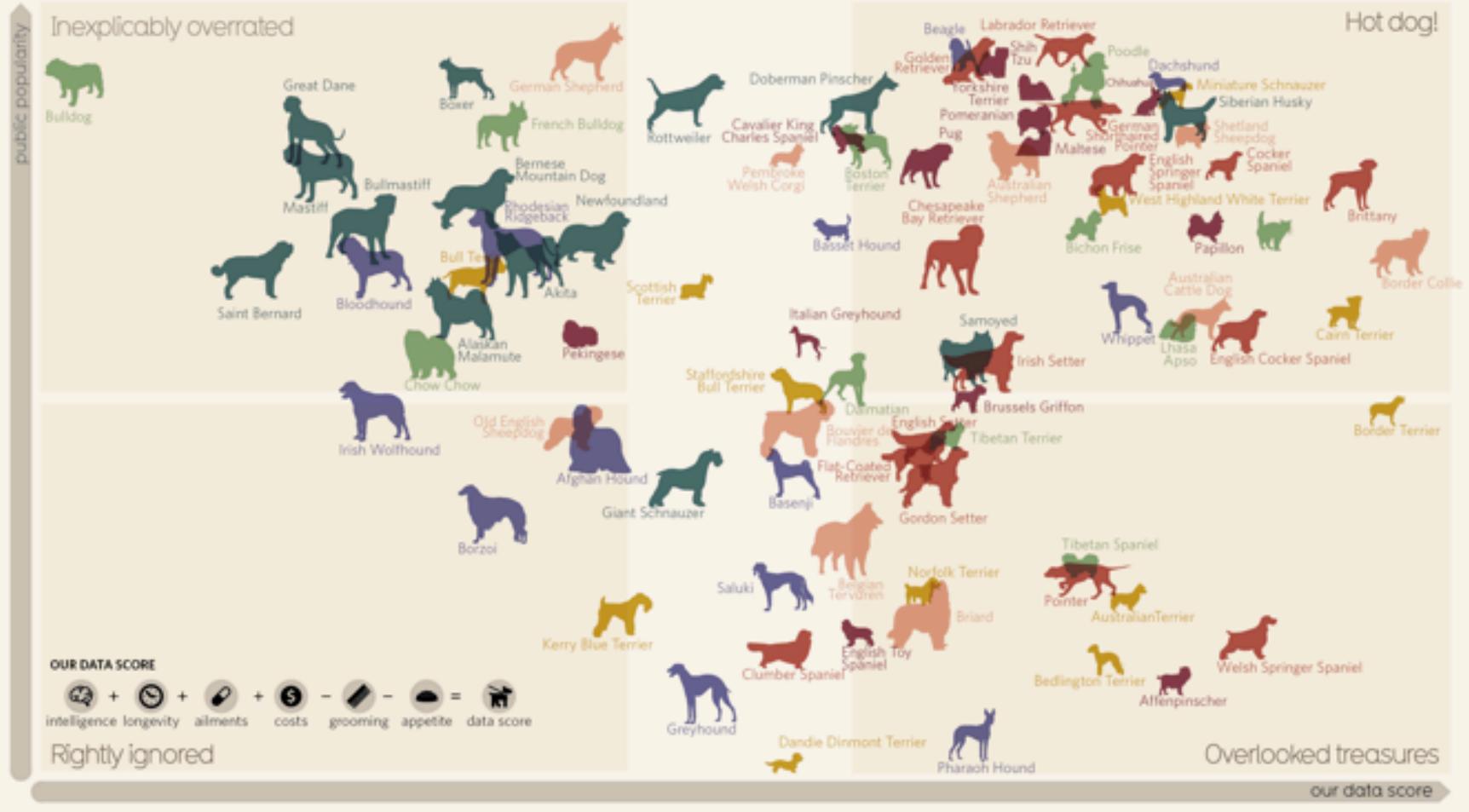
When to use?

- Use tables when
 - The document will be used to **look up individual values**
 - The document will be used to **compare individual values**
 - **Precise values** are required
 - The quantitative info to be communicated involves **more than one unit of measure**
- Use graphs when
 - The message is contained in the **shape** of the values
 - The document will be used to **reveal relationships** among values
 - Especially useful when the number of data points is huge

(Optional Reading) Stephen Few. 2012. Show Me the Numbers: Designing Tables and Graphs to Enlighten (2nd ed.). Analytics Press, , USA.

Best in Show

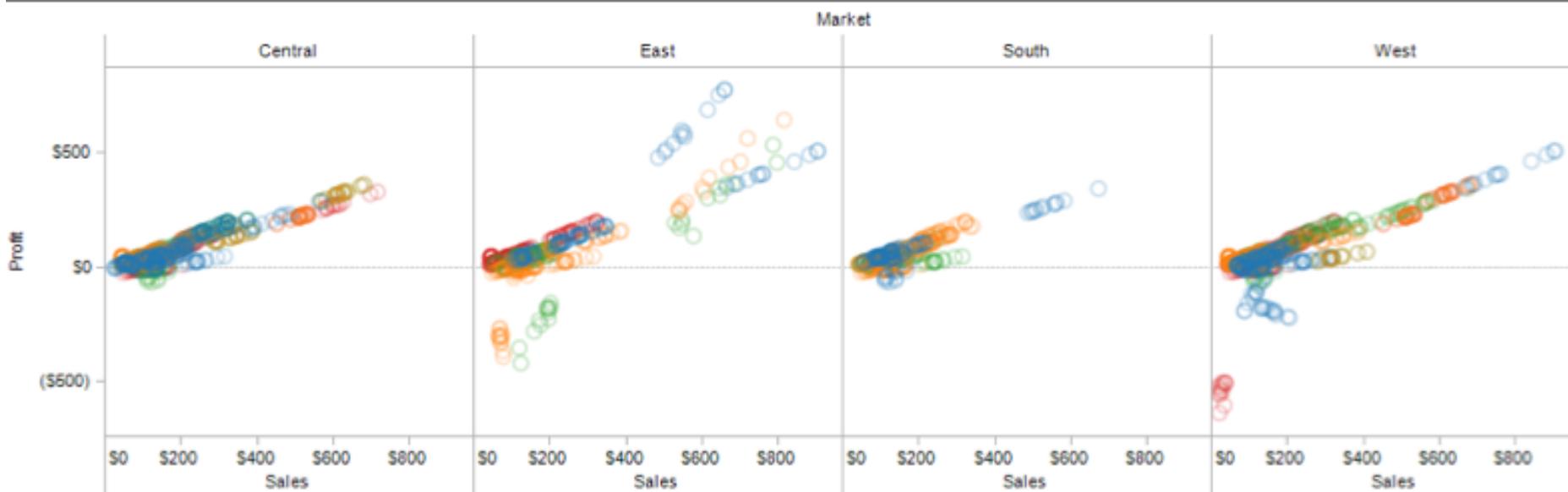
The ultimate data-dog



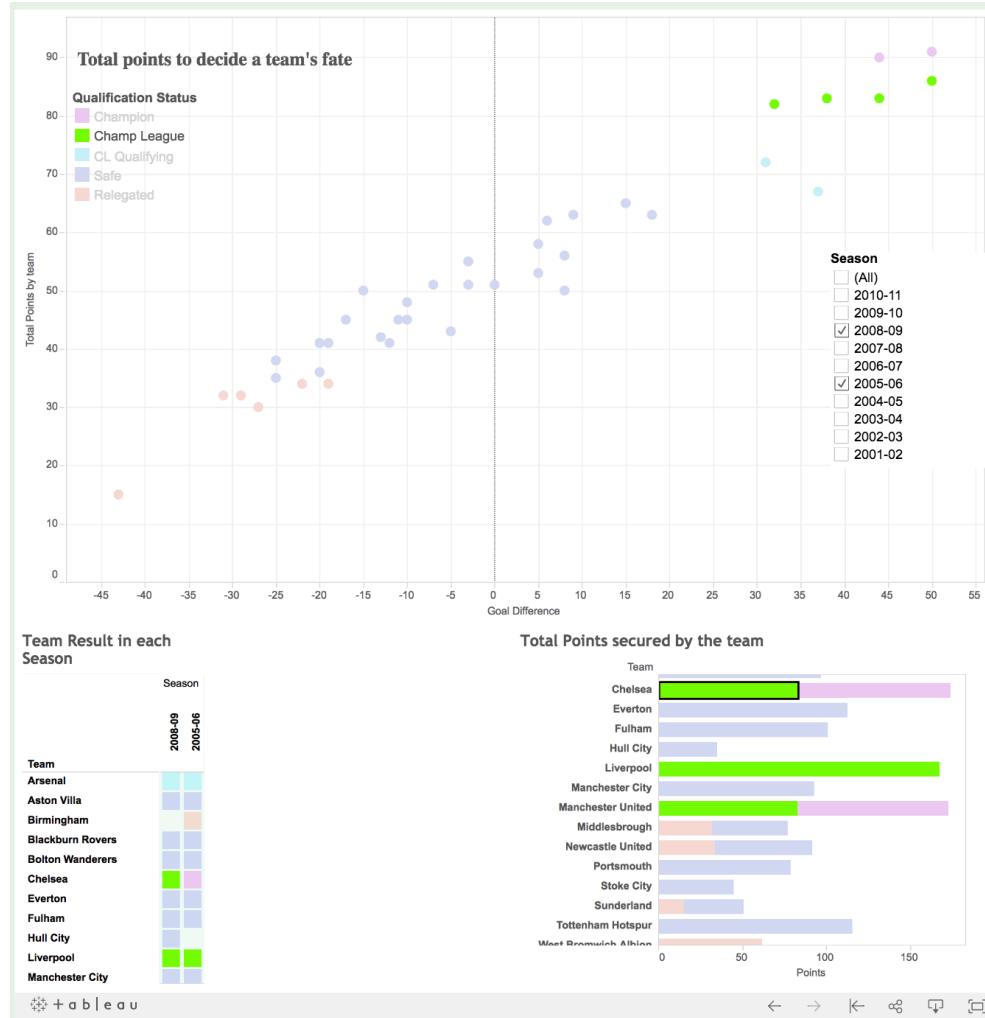
- Iconic Representations: Glyph (graphical object) represents a data case
- Visual properties of glyph represent different variables

Trellis Display (Small Multiples)

- It subdivides space to enable comparison across multiple plots.
- Typically nominal or ordinal variables are used as dimensions for subdivision.



Multiple Coordinated Views



Minard 1869: Napoleon's March

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813

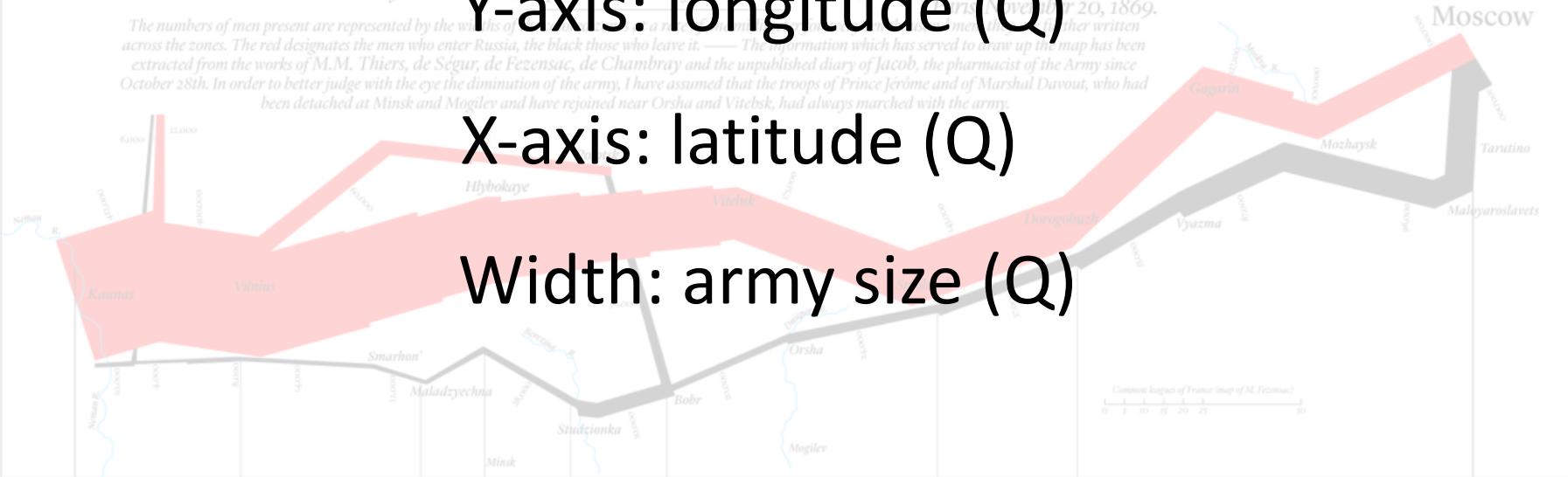
Drawn by M. Minard, Inspector General of Bridges and Roads (retired), Paris November 20, 1869.

The numbers of men present are represented by the widths of the lines which form a ribbon, the dimensions of which have been measured in the map itself after written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségrur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.

Y-axis: longitude (Q)

X-axis: latitude (Q)

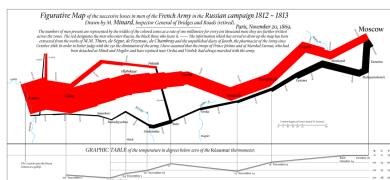
Width: army size (Q)



Y-axis: temperature (Q)

X-axis: longitude (Q) / time (O)

Depicts at least 5 quantitative variables.



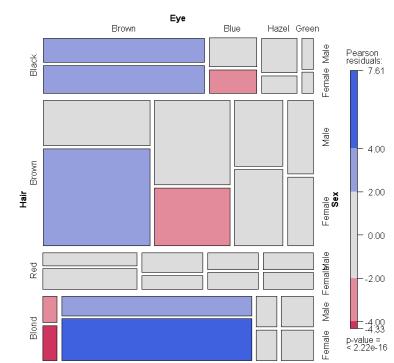
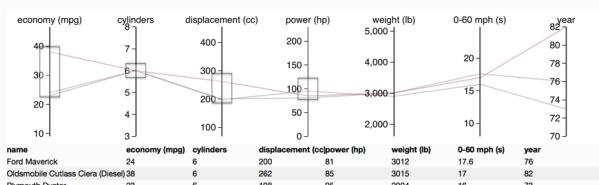
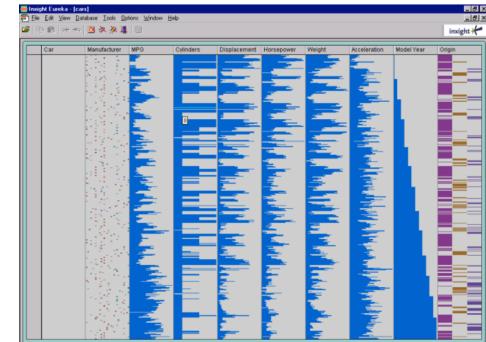
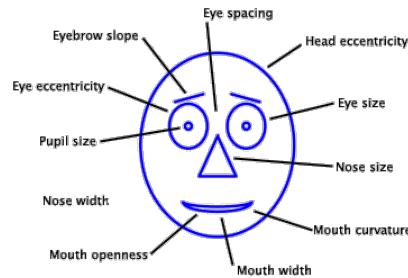
Multivariate Data Visualization

- Visual Encodings: 8 dimensions?
- Focus: techniques can generally handle all data sets

	Characteristics				
	Selective	Associative	Quantitative	Order	Length
Position	• •	•• ••	↑	↑	Theoretically Infinite
Size	• ●	••●●•●		●>●>●>●	Selection: ~5 Distinction: ~20
Shape					Theoretically Infinite
Value	○●○○○○	●○○●○○●		○<○<○<●<●<●	Selection: <7 Distinction: ~10
Color	• ○	●○●●○●●			Selection: <7 Distinction: ~10
Orientation	/ \ /				Theoretically Infinite
Texture	●●	●●●●●●			Theoretically Infinite

Common Multivariate Data Visualization Techniques

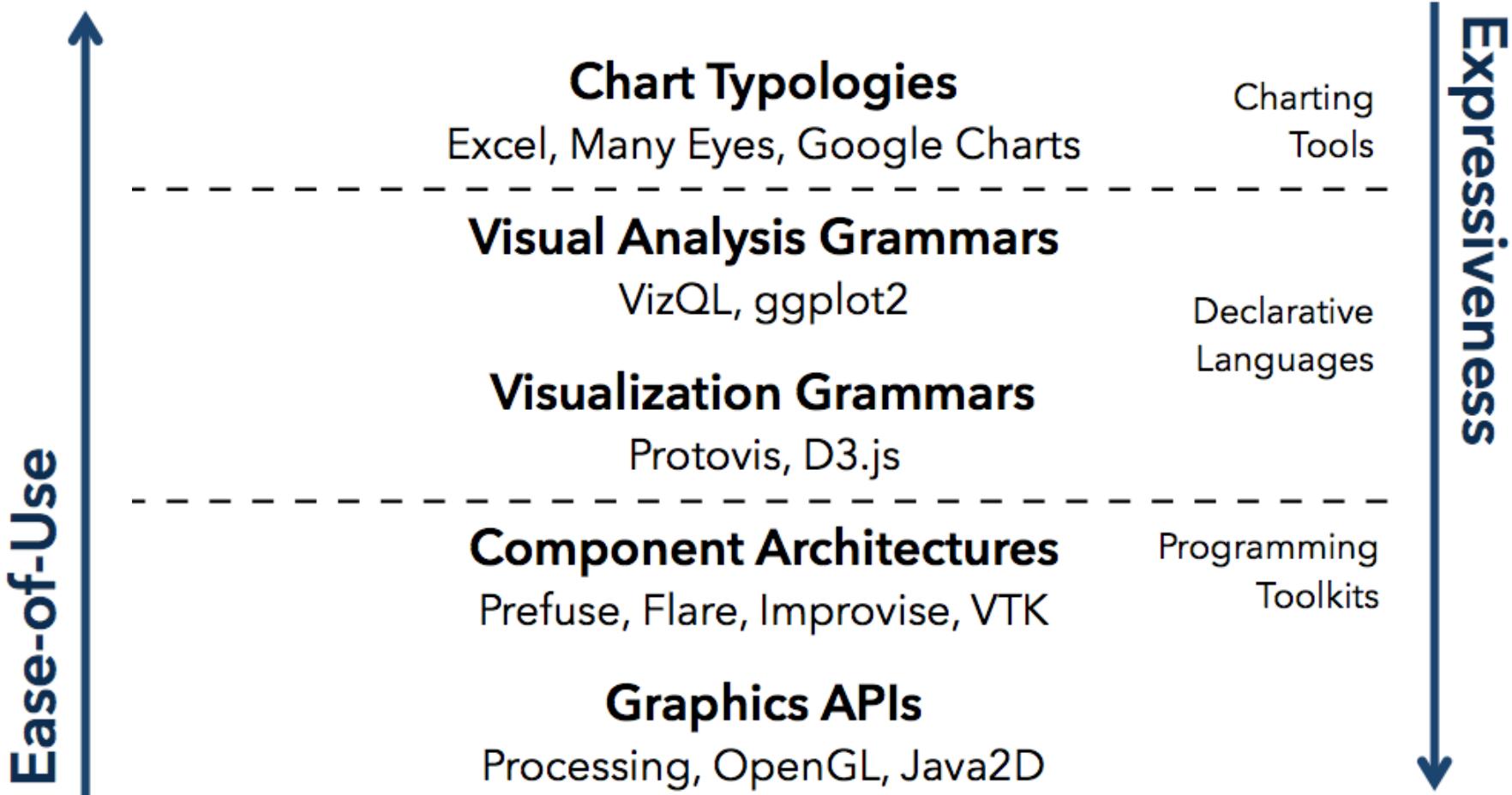
- Chernoff Faces
- Table Lens
- Parallel Coordinates
- Mosaic Plot



Multivariate Data Visualization

- Strategies:
 - Avoid “over-encoding”
 - Use space and small multiples intelligently
 - Reduce the problem space
 - Use interaction to generate relevant views
- Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is key.

Visualization Tools



The Advantages of Declarative Languages

- **Faster iteration.** Less code. Larger user base.
- **Better visualization.** Smart defaults.
- **Reuse.** Write-once, then re-apply.
- **Performance.** Optimization, scalability.
- **Portability.** Multiple devices, renderers, inputs.
- **Programmatic generation.** Write programs which output visualizations. Automated search & recommendation.

Building a Plot in ggplot2

data to visualize (a data frame)

map variables to **aes**thetic attributes

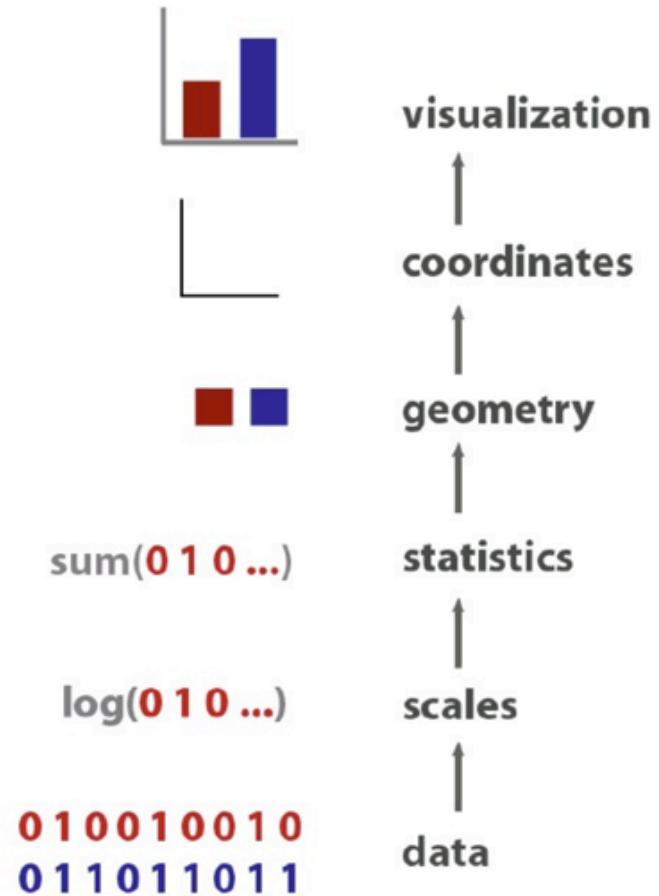
geometric objects – what you see (points, bars, etc)

scales map values from data to aesthetic space

faceting subsets the data to show multiple plots

statistical transformations – summarize data

coordinate systems put data on plane of graphic



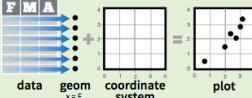
Data Visualization with ggplot2

Cheat Sheet

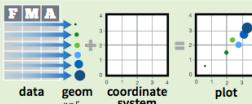


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Build a graph with **qplot()** or **ggplot()**

aesthetic mappings **data** **geom**

```
qplot(x = cty, y = hwy, color = cyl, data = mpg, geom = "point")
```

Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

ggplot(data = mpg, aes(x = cty, y = hwy))

Begins a plot that you finish by adding layers to. No defaults, but provides more control than **qplot()**.

data

```
ggplot(mpg, aes(hwy, cty)) +  
  geom_point(aes(color = cyl)) +  
  geom_smooth(method = "lm") +  
  coord_cartesian() +  
  scale_color_gradient() +  
  theme_bw()
```

add layers, elements with +

layer = geom + default stat + layer specific mappings

additional elements

Add a new layer to a plot with a **geom_***() or **stat_***() function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

last_plot()

Returns the last plot

ggsave("plot.png", width = 5, height = 5)

Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

One Variable

Continuous

```
a <- ggplot(mpg, aes(hwy))
```



a + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")



a + geom_density(kernel = "gaussian")
x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..count..))



a + geom_dotplot()
x, y, alpha, color, fill



a + geom_freqpoly()
x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))



a + geom_histogram(binwidth = 5)
x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

Discrete

```
b <- ggplot(mpg, aes(fl))
```



b + geom_bar()
x, alpha, color, fill, linetype, size, weight

Graphical Primitives

```
c <- ggplot(map, aes(long, lat))
```



c + geom_polygon(aes(group = group))
x, y, alpha, color, fill, linetype, size



d <- ggplot(economics, aes(date, unemploy))
d + geom_path(lineend = "butt", linejoin = "round", linemetre = 1)
x, y, alpha, color, linetype, size
d + geom_ribbon(aes(ymin = unemploy - 900, ymax = unemploy + 900))
x, ymax, ymin, alpha, color, fill, linetype, size



e <- ggplot(seals, aes(x = long, y = lat))
e + geom_segment(aes(xend = long + delta_long, yend = lat + delta_lat))
x, xend, y, yend, alpha, color, linetype, size



e + geom_rect(aes(xmin = long, ymin = lat, xmax = long + delta_long, ymax = lat + delta_lat))
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

Two Variables

Continuous X, Continuous Y

```
f <- ggplot(mpg, aes(cty, hwy))
```



f + geom_blank()



f + geom_jitter()
x, y, alpha, color, fill, shape, size



f + geom_point()
x, y, alpha, color, fill, shape, size



f + geom_quantile()
x, y, alpha, color, linetype, size, weight



f + geom_rug(sides = "bl")
alpha, color, linetype, size



f + geom_smooth(model = lm)
x, y, alpha, color, fill, linetype, size, weight



f + geom_text(aes(label = cty))
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust



AB
Discrete X, Continuous Y

```
g <- ggplot(mpg, aes(class, hwy))
```



g + geom_bar(stat = "identity")
x, y, alpha, color, fill, linetype, size, weight



g + geom_boxplot()
lower, middle, upper, x, ymax, ymin, alpha, color, fill, linetype, shape, size, weight



g + geom_dotplot(binaxis = "y", stackdir = "center")
x, y, alpha, color, fill



g + geom_violin(scale = "area")
x, y, alpha, color, fill, linetype, size, weight



Discrete X, Discrete Y
h <- ggplot(diamonds, aes(cut, color))



h + geom_jitter()
x, y, alpha, color, fill, shape, size

Three Variables

Continuous Bivariate Distribution

```
i <- ggplot(movies, aes(year, rating))
```



i + geom_hex2d(binwidth = c(5, 0.5))
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight



i + geom_density2d()
x, y, alpha, colour, linetype, size



i + geom_bin2d(binwidth = c(5, 0.5))
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight

Continuous Function

```
j <- ggplot(economics, aes(date, unemploy))
```



j + geom_area()
x, y, alpha, color, fill, linetype, size



j + geom_line()
x, y, alpha, color, linetype, size



j + geom_step(direction = "hv")
x, y, alpha, color, linetype, size

Visualizing error

```
df <- data.frame(grp = c("A", "B"), fit = 4.5, se = 1.2)
```



k + geom_crossbar(fatten = 2)
x, y, ymax, ymin, alpha, color, fill, linetype, size



k + geom_errorbar()
x, ymax, ymin, alpha, color, linetype, size, width (also **geom_errorbarh()**)



k + geom_linerange()
x, ymin, ymax, alpha, color, linetype, size



k + geom_pointrange()
x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

Maps

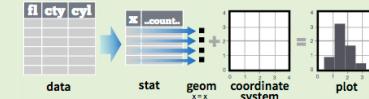
```
data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
```



l <- ggplot(data, aes(fill = murder))
l + geom_map(aes(map_id = state), map = map) + expand_limits(x = map\$long, y = map\$lat)
map_id, alpha, color, fill, linetype, size

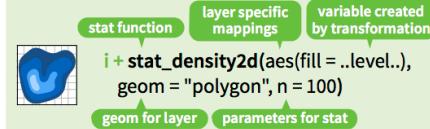
Stats - An alternative way to build a layer

Some plots visualize a **transformation** of the original data set. Use a **stat** to choose a common transformation to visualize, e.g. `a + geom_bar(stat = "bin")`



Each stat creates additional variables to map aesthetics to. These variables use a common `.name.` syntax.

stat functions and geom functions both combine a stat with a geom to make a layer, i.e. `stat_bin(geom="bar")` does the same as `geom_bar(stat="bin")`



`a + stat_bin(binwidth = 1, origin = 10)` 1D distributions
`x, y | .count, ..n, .count, ..density, ..ndensity..`
`a + stat_bindot(binwidth = 1, binaxis = "x")`
`x, y | .count, ..n, .count..`
`a + stat_density(adjust = 1, kernel = "gaussian")`
`x, y | .count, ..density, ..scaled..`

`f + stat_bin2d(bins = 30, drop = TRUE)` 2D distributions
`x, y, fill | .count, ..density..`
`f + stat_bineq(bins = 30)`
`x, y, fill | .count, ..density..`
`f + stat_density2d(contour = TRUE, n = 100)`
`x, y, color, size | ..level..`

`m + stat_contour(aes(z = z))` 3 Variables
`x, y, z, order | ..level..`
`mt + stat_speke(aes(radius = z, angle = z))`
`angle, radius, x, end, y, yend | ..x, ..xend, ..y, ..yend..`
`mt + stat_summary_hex(aes(z = z), bins = 30, fun = mean)`
`x, y, z, fill | ..value..`
`mt + stat_summary2d(aes(z = z), bins = 30, fun = mean)`
`x, y, z, fill | ..value..`

`g + stat_boxplot(coef = 1.5)` Comparisons
`x, y | ..lower, ..middle, ..upper, ..outliers..`
`g + stat_ydensity(adjust = 1, kernel = "gaussian", scale = "area")`
`x, y | ..density, ..scaled, ..count, ..n, ..violinwidth, ..width..`

`f + stat_ecdf(n = 40)` Functions
`x, y | ..x, ..y..`
`f + stat_quantile(quintiles = c(0.25, 0.5, 0.75), formula = y ~ log(x), method = "rq")`
`x, y | ..quintile, ..x, ..y..`
`f + stat_smooth(method = "auto", formula = y ~ x, se = TRUE, n = 80, fullrange = FALSE, level = 0.95)`
`x, y | ..se, ..x, ..y, ..ymin, ..ymax..`

`ggplot() + stat_function(es(x = -3), fun = dnorm, n = 101, args = list(sd = 0.5))` General Purpose
`x | ..y..`
`f + stat_identity()`
`ggplot() + stat_qq(aes(sample = 1:100), distribution = qt, dparams = list(df = 5))`
`sample, x, y | ..x, ..y..`
`f + stat_sum()`
`x, y, size | ..size..`
`f + stat_summary(fun.data = "mean_cl_boot")`
`f + stat_unique()`

Scales

Scales control how a plot maps data values to the visual values of an aesthetic. To change the mapping, add a custom scale.



General Purpose scales

Use with any aesthetic:
alpha, color, fill, linetype, shape, size

`scale_*_continuous()` - map cont' values to visual values
`scale_*_discrete()` - map discrete values to visual values
`scale_*_identity()` - use data values as visual values
`scale_*_manual(values = c())` - map discrete values to manually chosen visual values

X and Y location scales

Use with x or y aesthetics (x shown here)

`scale_x_date(labels = date_format("%m/%d"), breaks = date_breaks("2 weeks"))` - treat x values as dates. See `?strptime` for label formats.
`scale_x_datetime()` - treat x values as date times. Use same arguments as `scale_x_date()`.
`scale_x_log10()` - Plot x on log10 scale
`scale_x_reverse()` - Reverse direction of x axis
`scale_x_sqrt()` - Plot x on square root scale

Color and fill scales

Discrete
`n <- b + geom_bar(aes(fill = fl))`
`n + scale_fill_brewer(palette = "Blues")`
For palette choices:
library(Brewer)
display.brewer.all()

Continuous
`o <- a + geom_dotplot(aes(fill = ...))`
`o + scale_fill_gradient(low = "red", high = "yellow")`
`o + scale_fill_gradient2(low = "red", high = "blue", mid = "white", midpoint = 25)`
`o + scale_fill_gradientn(colours = terrain.colors(6))`
Also: rainbow(), heat.colors(), topo.colors(), cm.colors(), RColorBrewer::brewer.pal()

Shape scales

Manual shape values
`p <- f + geom_point(aes(shape = fl))`
`p + scale_shape(solid = FALSE)`
`p + scale_shape_manual(values = c(3:7))`
Shape values shown in chart on right

Size scales

`q <- f + geom_point(aes(size = cyl))`
`q + scale_size_area(max = 6)`
Value mapped to area of circle (not radius)

Coordinate Systems

`r <- b + geom_bar()`
`r + coord_cartesian(xlim = c(0, 5))`
`xlim, ylim`
The default cartesian coordinate system

`r + coord_fixed(ratio = 1/2)`
ratio, xlim, ylim
Cartesian coordinates with fixed aspect ratio between x and y units

`r + coord_flip()`
xlim, ylim
Flipped Cartesian coordinates

`r + coord_polar(theta = "x", direction = 1)`
theta, start, direction
Polar coordinates

`r + coord_trans(trans = "sqrt")`
xtrans, ytrans, xlim, ylim
Transformed cartesian coordinates. Set extras and strains to the name of a window function.

`z + coord_map(projection = "ortho", orientation = c(41, -74, 0))`
projection, orientation, xlim, ylim
Map projections from the mapproj package (mercator (default), azequalarea, lagrange, etc.)

Position Adjustments

Position adjustments determine how to arrange geoms that would otherwise occupy the same space.

`s <- ggplot(mpg, aes(fl, fill = drv))`
`s + geom_bar(position = "dodge")`
Arrange elements side by side

`s + geom_bar(position = "fill")`
Stack elements on top of one another, normalize height

`s + geom_bar(position = "stack")`
Stack elements on top of one another

`f + geom_point(position = "jitter")`
Add random noise to X and Y position of each element to avoid overplotting

Each position adjustment can be recast as a function with manual width and height arguments

`s + geom_bar(position = position_dodge(width = 1))`

Themes

`r + theme_bw()`
White background with grid lines

`r + theme_classic()`
White background no gridlines

`r + theme_grey()`
Grey background (default theme)

`r + theme_minimal()`
Minimal theme

ggthemes - Package with additional ggplot2 themes

Faceting

Facets divide a plot into subplots based on the values of one or more discrete variables.

`t <- ggplot(mpg, aes(cty, hwy)) + geom_point()`

`t + facet_grid(. ~ fl)`
facet into columns based on fl

`t + facet_grid(year ~ .)`
facet into rows based on year

`t + facet_grid(year ~ fl)`
facet into both rows and columns

`t + facet_wrap(~ fl)`
wrap facets into a rectangular layout

Set scales to let axis limits vary across facets

`t + facet_grid(y ~ x, scales = "free")`
x and y axis limits adjust to individual facets

- "free_x" - x axis limits adjust
- "free_y" - y axis limits adjust

Set labeller to adjust facet labels

<code>t + facet_grid(. ~ fl, labeller = label_both)</code>	<code>fl: c</code>	<code>fl: d</code>	<code>fl: e</code>	<code>fl: p</code>	<code>fl: r</code>
<code>t + facet_grid(. ~ fl, labeller = label_bquote(alpha ^ .(x)))</code>	<code>alpha^c</code>	<code>alpha^d</code>	<code>alpha^e</code>	<code>alpha^p</code>	<code>alpha^r</code>
<code>t + facet_grid(. ~ fl, labeller = label_parsed)</code>	<code>c</code>	<code>d</code>	<code>e</code>	<code>p</code>	<code>r</code>

Labels

`t + ggtitle("New Plot Title")`

Add a main title above the plot

`t + xlab("New X label")`

Change the label on the X axis

`t + ylab("New Y label")`

Change the label on the Y axis

`t + labs(title = "New title", x = "New x", y = "New y")`
All of the above

Use scale functions to update legend labels

Legends

`t + theme(legend.position = "bottom")`
Place legend at "bottom", "top", "left", or "right"

`t + guides(color = "none")`

Set legend type for each aesthetic: colorbar, legend, or none (no legend)

`t + scale_fill_discrete(name = "Title", labels = c("A", "B", "C"))`

Set legend title and labels with a scale function.

Zooming

Without clipping (preferred)

`t + coord_cartesian(xlim = c(0, 100), ylim = c(10, 20))`

With clipping (removes unseen data points)

`t + xlim(0, 100) + ylim(10, 20)`

`t + scale_x_continuous(limits = c(0, 100)) + scale_y_continuous(limits = c(0, 100))`

Visualizing Text

Challenges

- High Dimensionality
 - Where possible use text to represent text...
 - ... which terms are the most descriptive?
- Context and Semantics
 - Provide relevant context to aid understanding.
 - Show (or provide access to) the source text.
- Modeling Abstraction
 - Determine your analysis task.
 - Understand abstraction of your language models.
 - Match analysis task with appropriate tools and models.

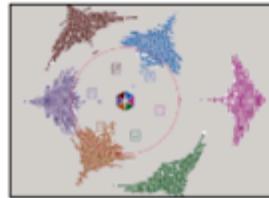
Text Visualization



Visualizing text
Showing words,
phrases, and
sentences



- Content
- Context
- Relationship to others



Visualization for IR
Helping search



Text Processing Pipeline

- Tokenization
 - Segment text into terms.
 - Remove stop words? [a](#), [an](#), [the](#), [of](#), [to](#), [be](#)
 - Numbers and symbols? [#gocard](#), [@nottinghamforestfbball](#)
 - Entities? [Nottingham](#), [Trump](#).
- Stemming
 - Group together different forms of a word.
 - Porter stemmer? [visualization\(s\)](#), [visualize\(s\)](#), [visually](#) -> [visual](#)
 - Lemmatization? [goes](#), [went](#), [gone](#) -> [go](#)
- Ordered list of terms

Bag of Words Model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
 - Each dimension corresponds to a term (10,000+)
 - Each value represents the relevance
 - For example, simple term counts
- Aggregate into a document-term matrix
 - Document vector space model

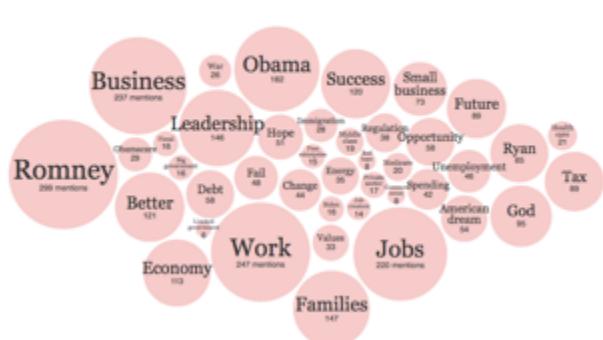
Keyword Weighting

- Term Frequency
 - $tf_{td} = \text{count}(t) \text{ in } d$
 - Can take log frequency: $\log(1 + tf_{td})$
 - Can normalize to show proportion $(tf_{td} / \sum_t tf_{td})$
- TF.IDF: Term Freq by Inverse Document Freq
 - $tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$
 - df_t = # docs containing t;
 - N = # of docs

Word Counts and Tag Cloud

At the Republican Convention, the Words Being Used

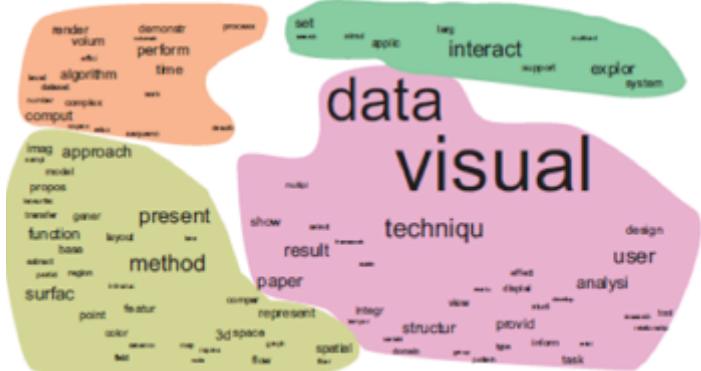
A look at how often speakers at the Republican National Convention have used certain words and phrases so far, based on an analysis of transcripts from the Federal News Service.



<http://www.nytimes.com/interactive/2012/08/28/us/politics/convention-word-counts.html>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

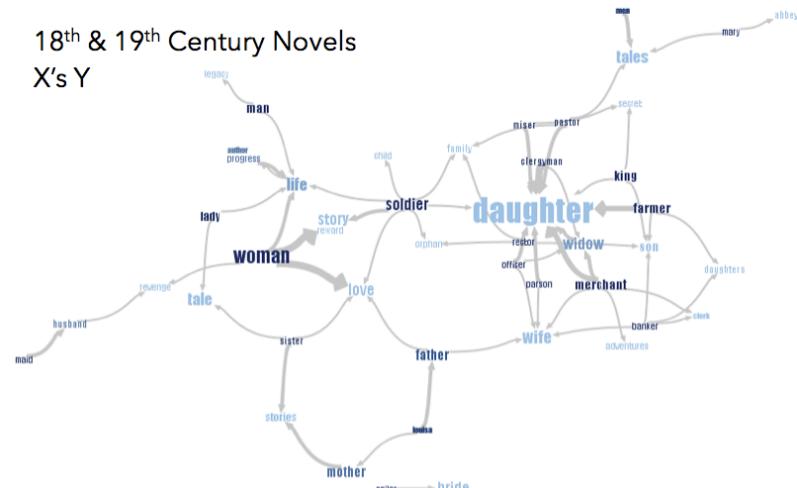
Context and Semantics



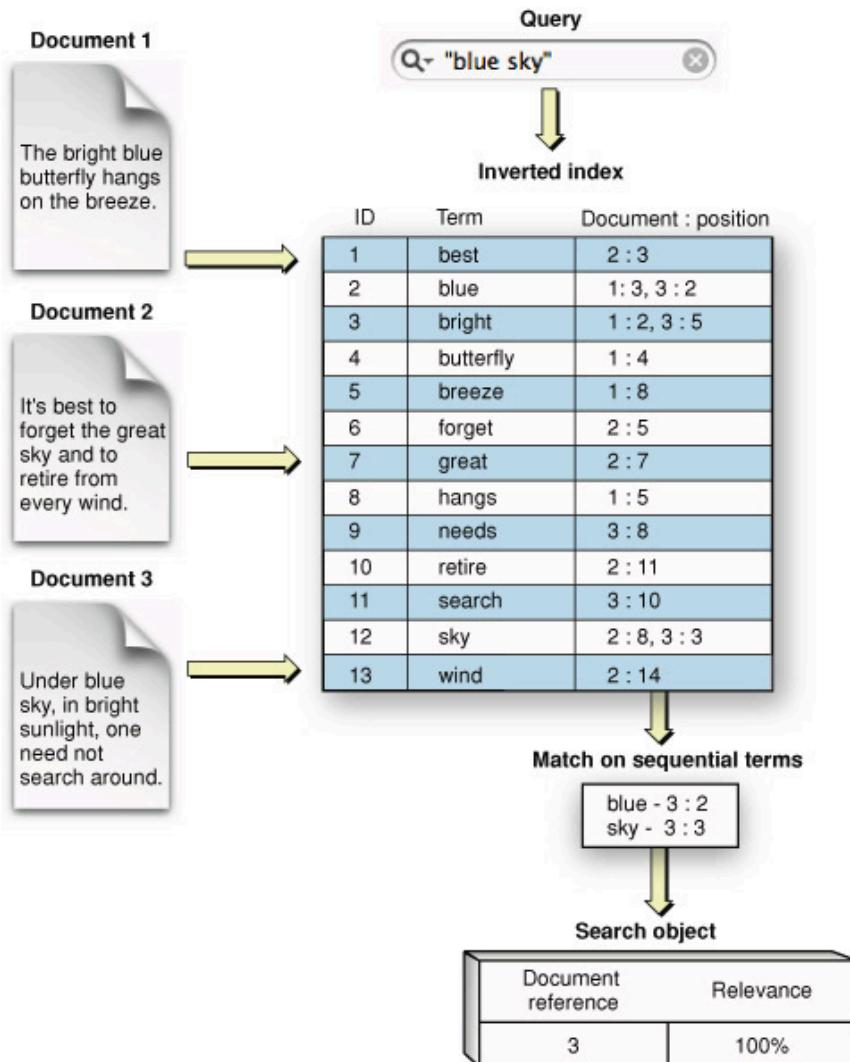
The screenshot shows the Concordance - Larkin.Concordance application window. The menu bar includes File, Text, Search, Edit, Headwords, Contexts, View, Tools, and Help. Below the menu is a toolbar with icons for opening files, saving, printing, and other functions. A vertical sidebar on the right contains buttons for 'Context', 'Text', 'Table', 'List', 'Index', and 'Note'. The main area displays a table of search results:

Headword	No.	Context...	Word	...Context	Reference
HEAR	15		That my own	heart	dritts and cries, having no...
HEARD	9		By the shout of the	heart	continually at work
HEARING	7	Nothing to adapt the skill of the	heart	to, still	And the wave And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	.	Träumerlei
HEARSE	1	Because I follow it to my own	heart	.	many famous
HEART	25		My	is tickling like the sun:	I was washed i...
HEART'S	2		heart	sharpened to a candid co...	The March Pa...
HEART-SHAPED	1		Contract my	heart	Lines on a Yo
HEARTH	1		Having no	heart	Home is so Se...
HEARTS	7	And the boy pulling his	heart	out in the Gents	Essential Bea...
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the...
HEAT	6	These I would choose my	heart	to lead	After-Dinner F...
HEAT-HAZE	1	Time in his little cinema of the	heart	.	Time and Spac...
HEATH	1	This petrified	heart	has taken,	A Stone Ch...
HEATS	1	How should they sweep the girl clean	heart	.	I see a girl dra...
HEAVE	1	Hands that the	heart	can govern	Heaviest of th...
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessable	heart	riding	One man walk...
HEAVIER-THAN...	1	If hands could free you,	heart	.	If hands could
HEAVIEST	2	That overflows the	heart	.	Pour away thi...

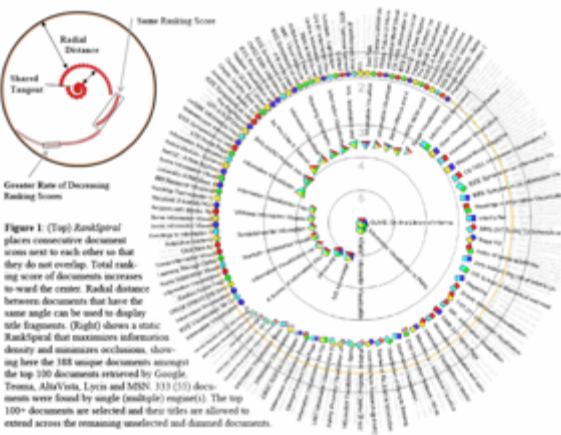
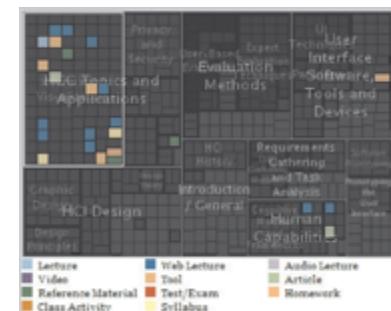
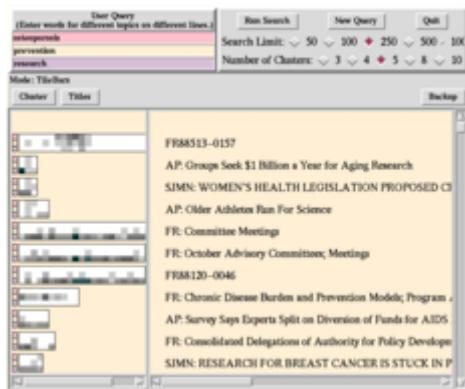
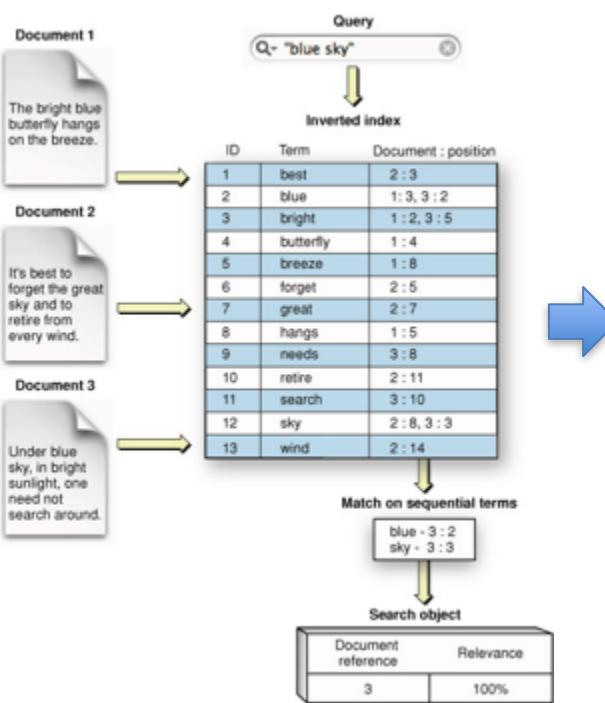
At the bottom, there are buttons for 'Words' (7318), 'Tokens' (37070), 'At word' (2990), 'Deleted lines' (1 [24]), 'Word sort' (Asc alpha [string]), 'Context sort' (Asc occurrence order), and a large 'Next' button.



Information Retrieval (IR)



Visualizing for IR



Visualizing Time Series Data

Tasks

- Often asked questions:
 - when was something greatest/least?
 - is there a pattern? are two series similar?
 - does a data element exist at time t, and when?
 - how long does a data element exist and how often?
 - how fast are data elements changing
 - in what order do data elements appear?
 - do data elements exist together?

(Optional Reading) Müller, Wolfgang, and Heidrun Schumann. "Visualization for modeling and simulation: visualization methods for time-dependent data—an overview." Proceedings of the 35th conference on Winter simulation: driving innovation. Winter Simulation Conference, 2003.

Visualizing Time-oriented Data

A Systematic View

The TimeViz Browser
 A Visual Survey of Visualization Techniques for Time-Oriented Data
 by Christian Tominski and Wolfgang Aigner

of Techniques: 115

Search:

How to use filters:

- Want: Show me!
- Indifferent: I don't care.
- Hide: I'm not interested!

Data

Frame of Reference

- Abstract
- Spatial

Number of Variables

- Univariate
- Multivariate

Time

Arrangement

- Linear
- Cyclic

Time Primitives

- Instant
- Interval

Visualization

Mapping

- Static
- Dynamic



Aigner, Wolfgang, et al. "Visualizing time-oriented data—a systematic view." Computers & Graphics 31.3 (2007): 401-409.

<http://www.timeviz.net/>

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Taxonomy

Time	Temporal primitives	time points (a) (b) (c) (d) (e) (f) (g) (i)									time intervals (g) (h)			
	Structure of time	linear (a) (b) (c) (d) (f) (g) (h) (i)					cyclic (e)			branching (h)				
Data	Frame of reference	abstract (c) (d) (f) (g) (h) (i)							spatial (a) (b) (e) (i)					
	Number of variables	univariate (a) (b) (f) (g) (h)							multivariate (c) (d) (e) (i)					
	Level of abstraction	data (a) (b) (c) (d) (e) (f) (g) (h) (i)							data abstractions (b) (g) (i)					
Representation	Time dependency	static (c) (d) (e) (g) (h) (i)							dynamic (a) (b) (f) (i)					
	Dimensionality	2D (a) (c) (d) (g) (h) (i)							3D (b) (e) (f) (i)					

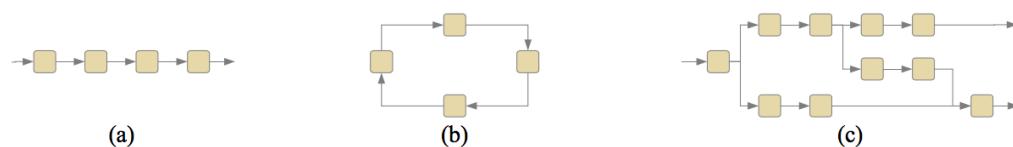


Fig. 2. Structure of time: (a) Linear time; (b) Cyclic time; (c) Branching time.

Time Series Visualization

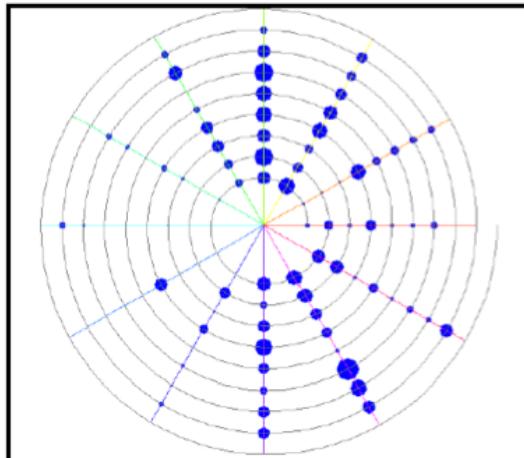
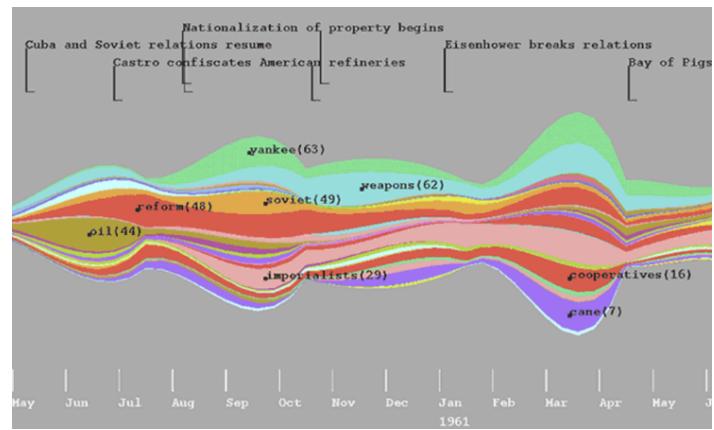
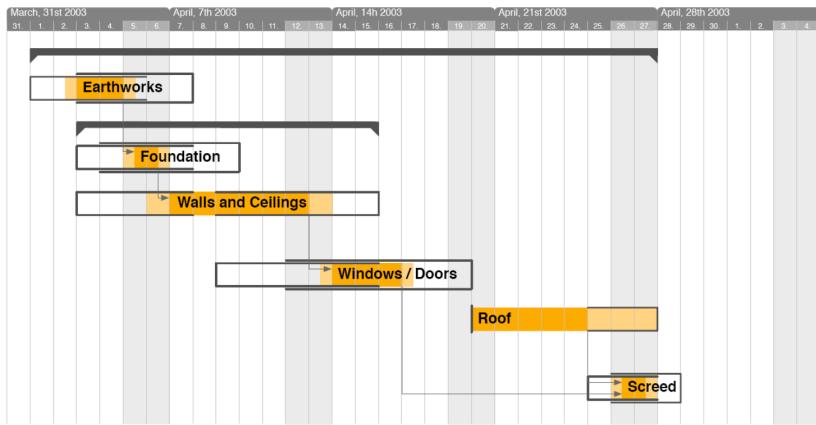
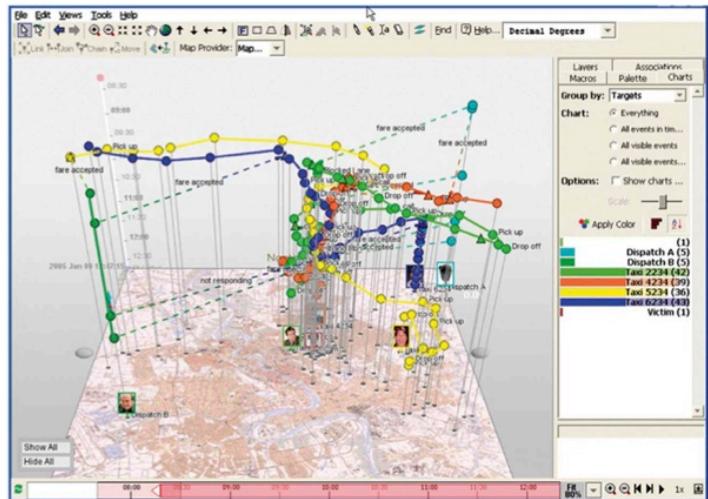


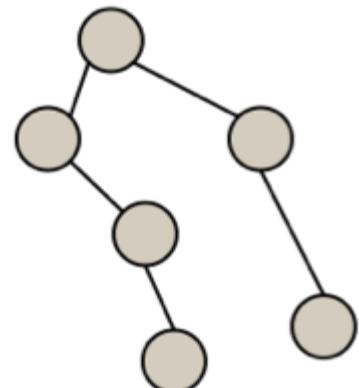
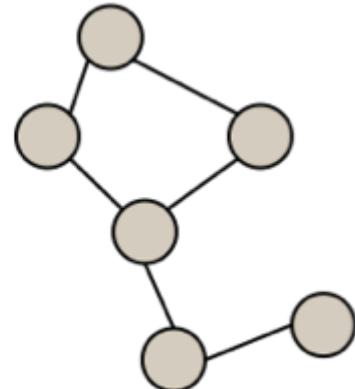
Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for *Baphia Capparidifolia* during the period 1980 – 1988.



Visualizing Trees and Graphs

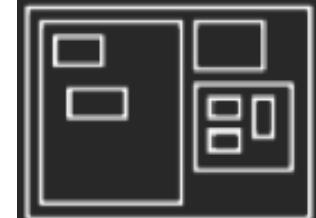
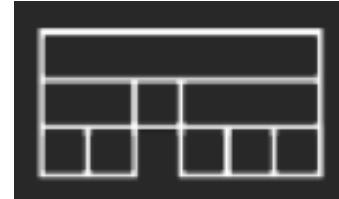
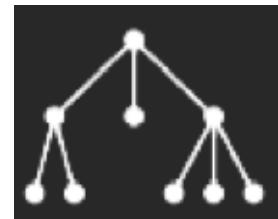
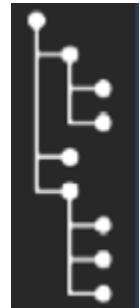
Graphs and Trees

- Graphs
 - Model relations among data
 - Nodes and edges
- Trees
 - Graphs with hierarchical structure
 - Connected graph with $N-1$ edges
 - Nodes as parents and children



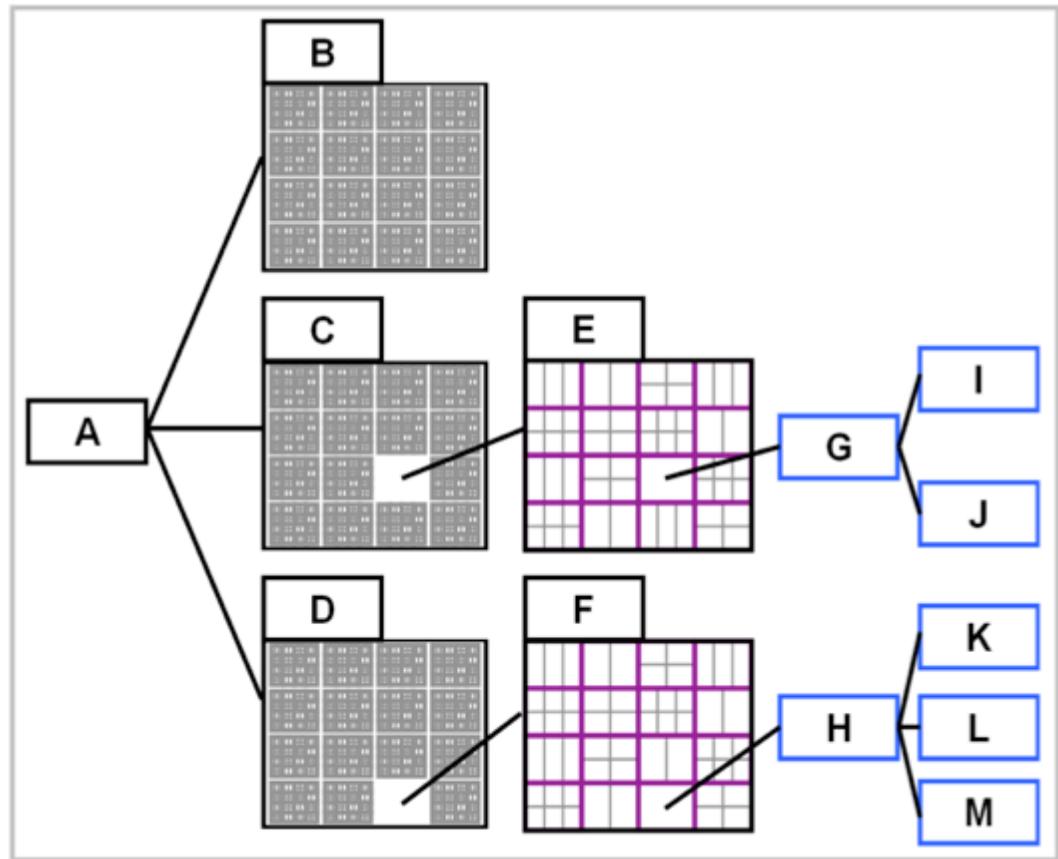
Tree Visualizations

- Indented lists
 - Linear list, indentation encodes depth
- Node-link trees
 - Nodes connected by lines/curves
- Layered diagrams
 - Relative position and alignment
- Treemaps
 - Represent hierarchy by enclosure

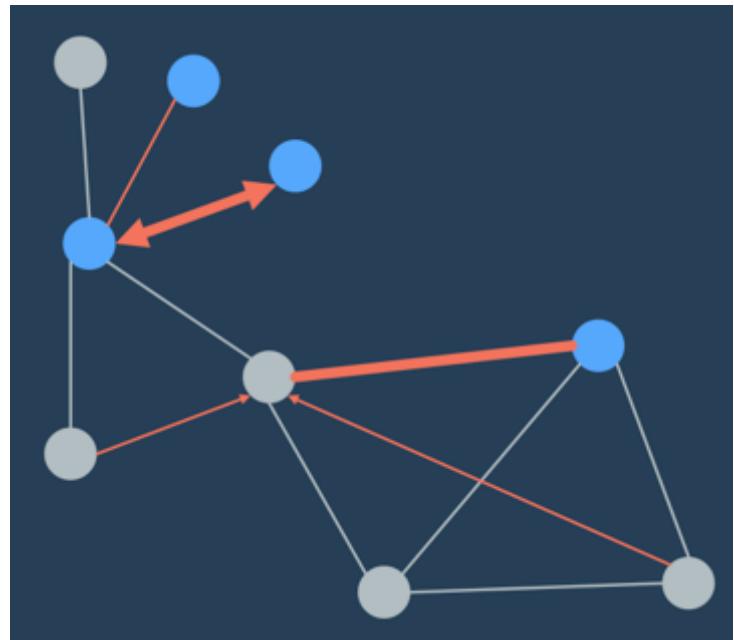
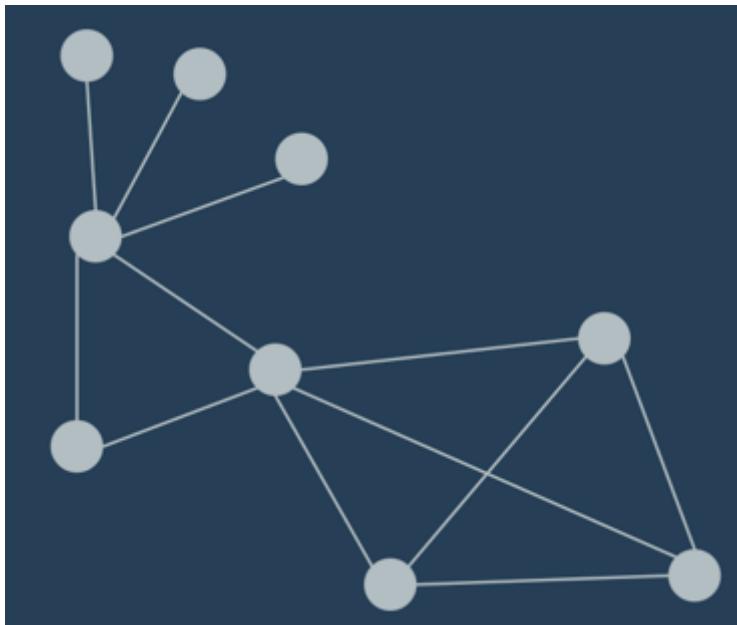


Hybrids

- Elastic Hierarchies
 - Node-link diagram with treemap nodes.
- Video:
<https://www.youtube.com/watch?v=nvslqYQ75yA>



What's in a Graph?



Graph Visualization

- Two representations:
 - Node-link diagrams
 - Matrices



- Major Node-Link Layouts

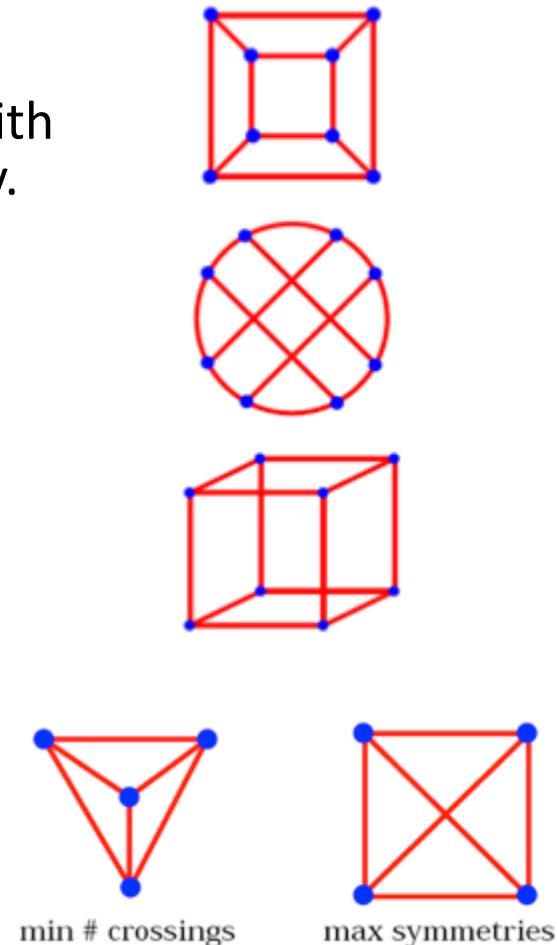
Tree in the Graph



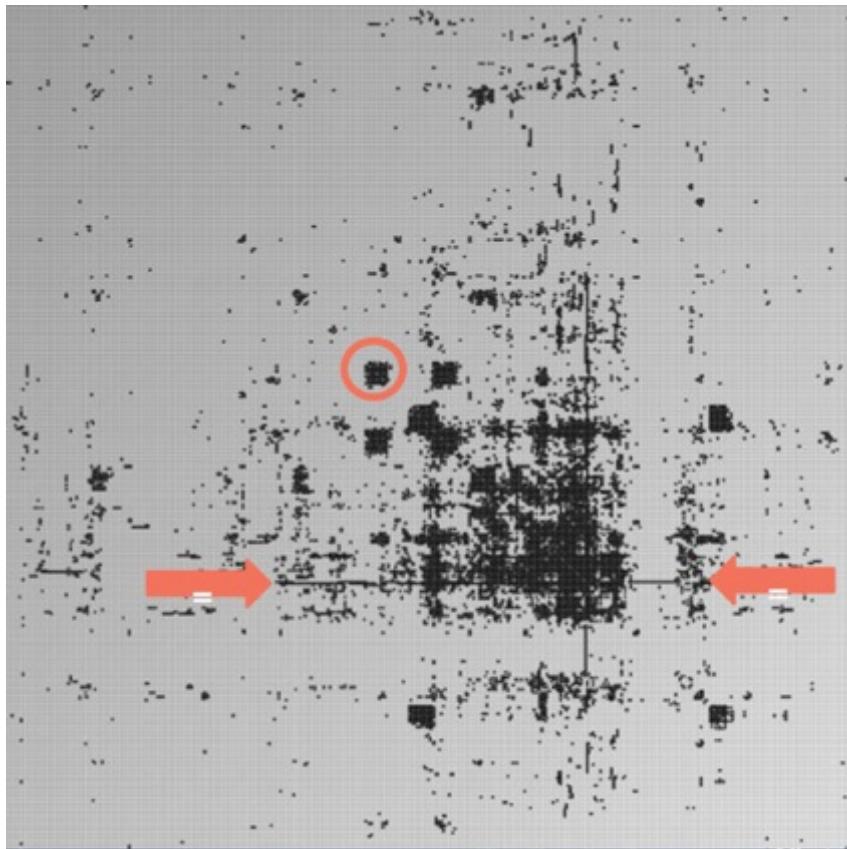
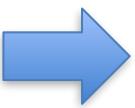
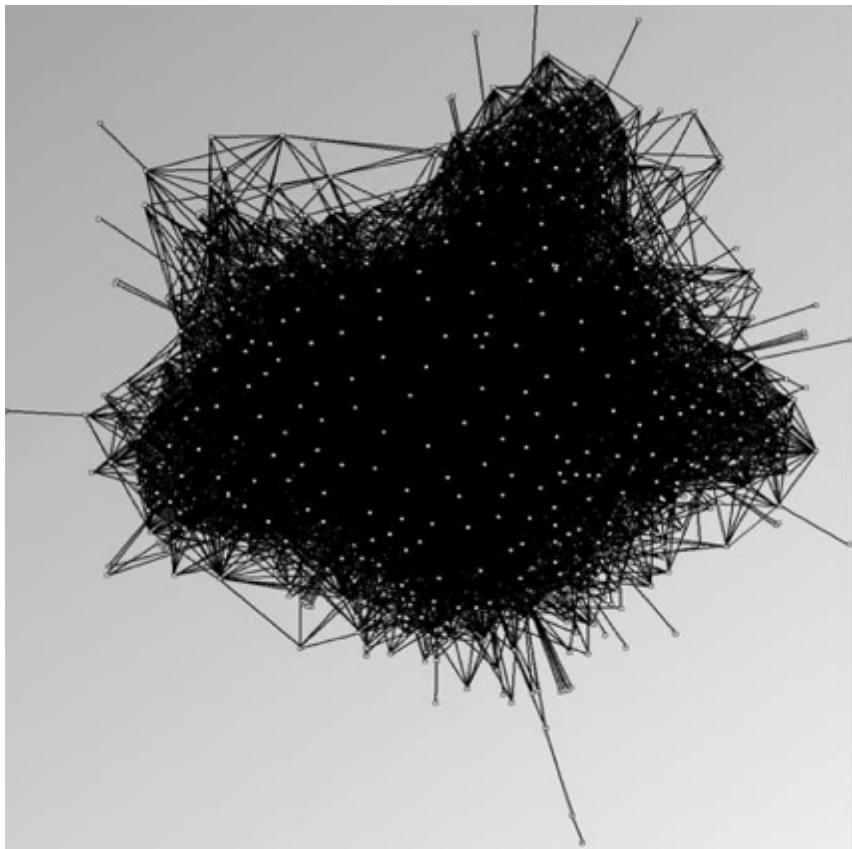
Hierarchical Graph Layout

Optimization Techniques

- Treat layout as an optimization problem
 - Define layout using an energy model along with constraints: equations the layout should obey.
 - Use optimization algorithms to solve
- Commonly posed as a physical system
 - Charged particles, springs, drag force, ...
- Different constraints can be introduced
 - Minimize edge crossings
 - Minimize area
 - Minimize line bends
 - Minimize line slopes
 - Maximize smallest angle between edges
 - Maximize symmetry



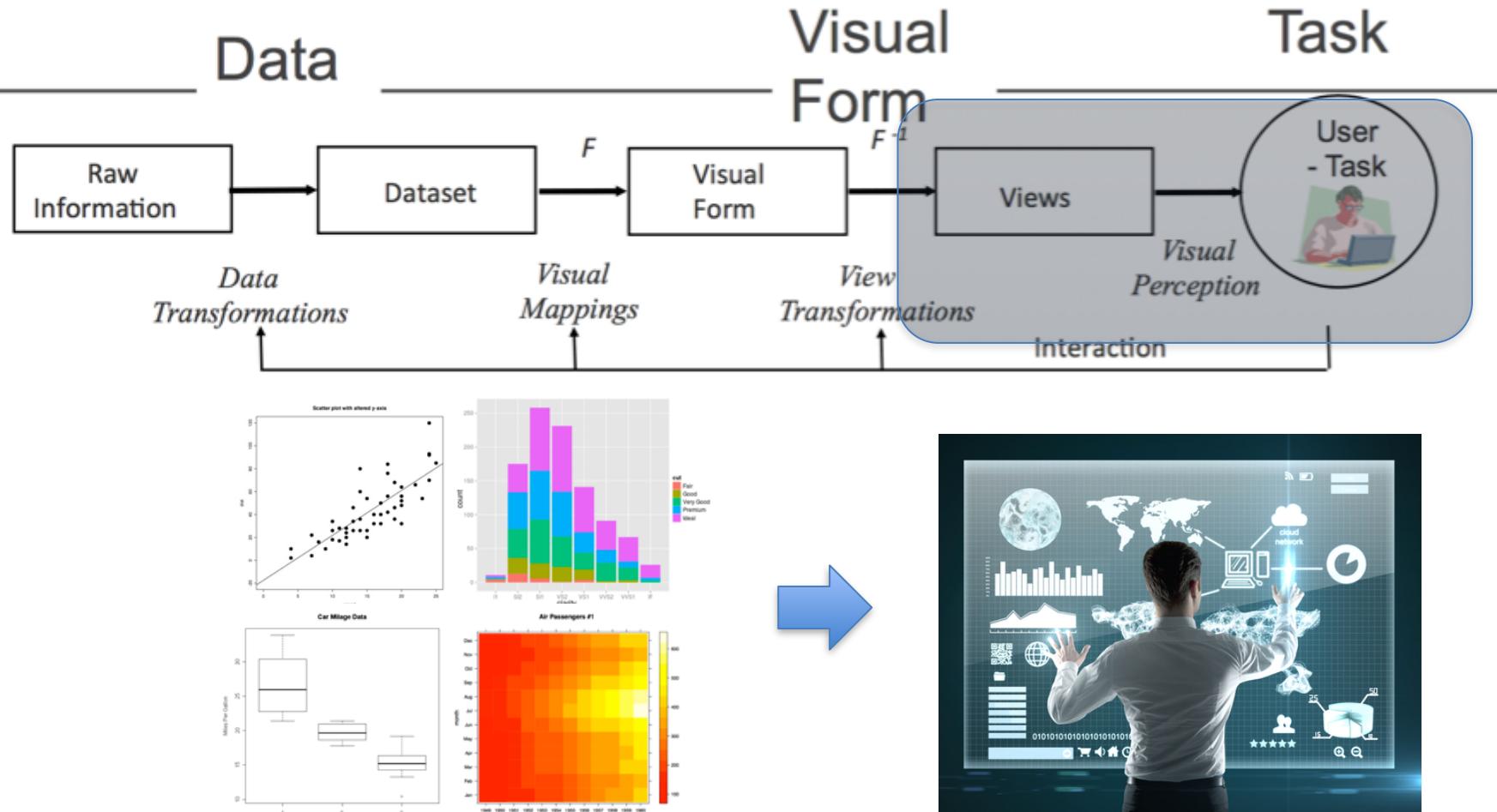
Matrices



Matrix vs. Node-link

Matrix	Node-link
Require learning	Familiar
No overlap	Node overlap
No crossings	Link crossing
Use a lot of space	More compact
Dense graphs	Sparse graphs

Information Visualization



Design: how can we design the visualization best for human to accomplish their tasks?

What design criteria should we follow?

What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language (1) **express all the facts** in the set of data, and (2) **only the facts** in the data.

Tell the truth

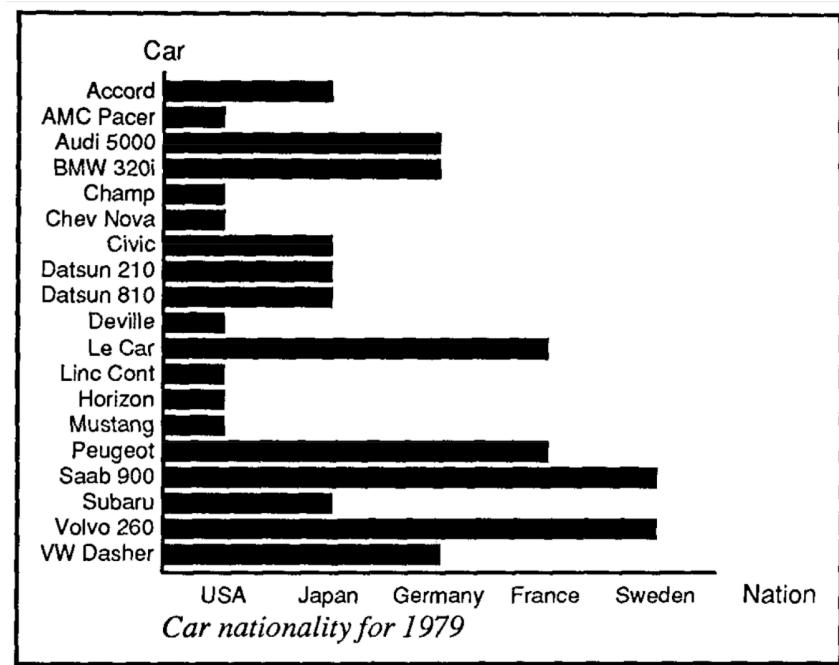
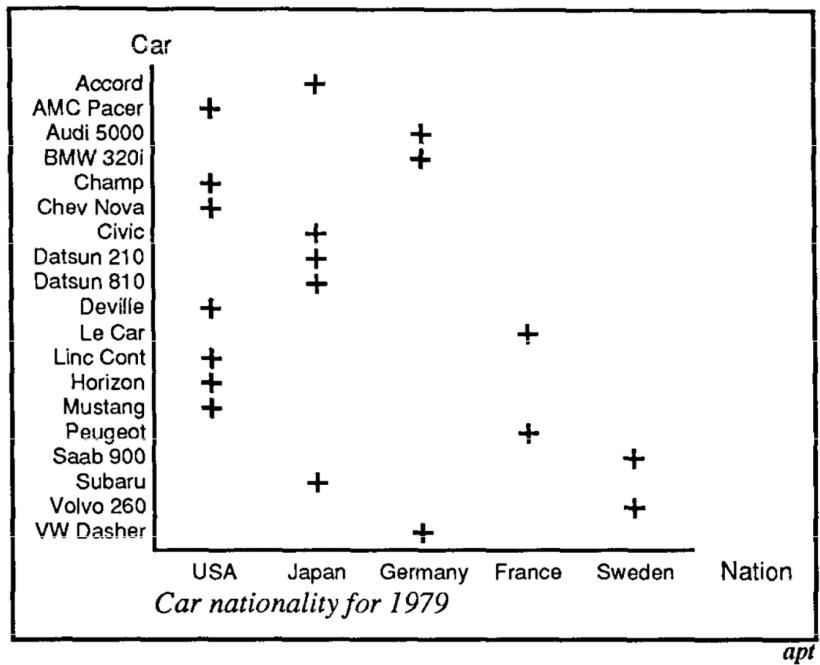
- **Effectiveness**

- A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Use proper encoding

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Expressiveness



An Alternative

What Design Criteria to Follow?

- **Expressiveness**

- A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express (1) all the facts in the set of data, and (2) only the facts in the data.

Tell the truth

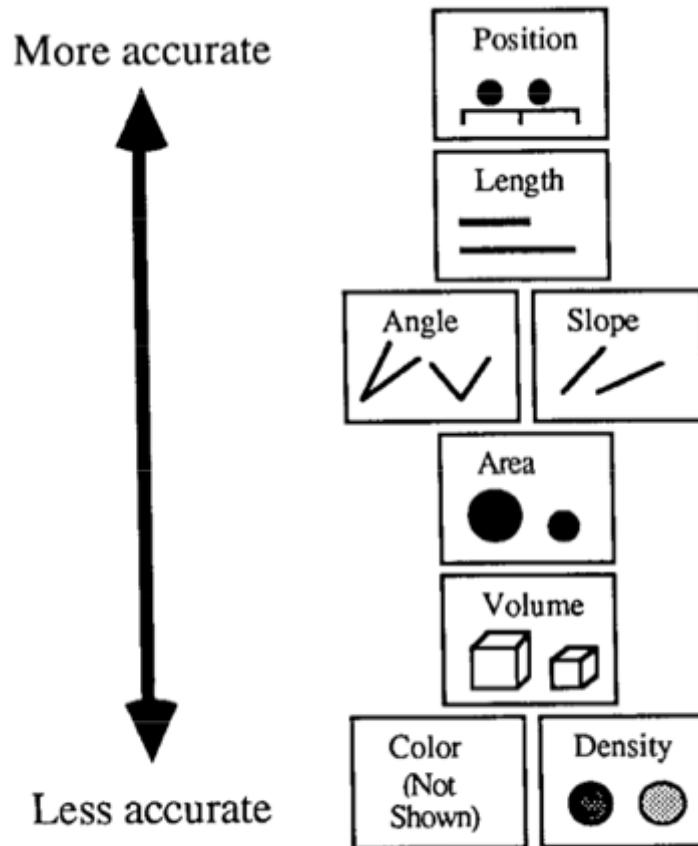
- **Effectiveness**

- A visualization is more effective than another visualization if the information conveyed by one visualization **is more readily perceived** than the information in the other visualization.

Use proper encoding

Mackinlay, Automating the design of graphical presentations of relational information, 1986.

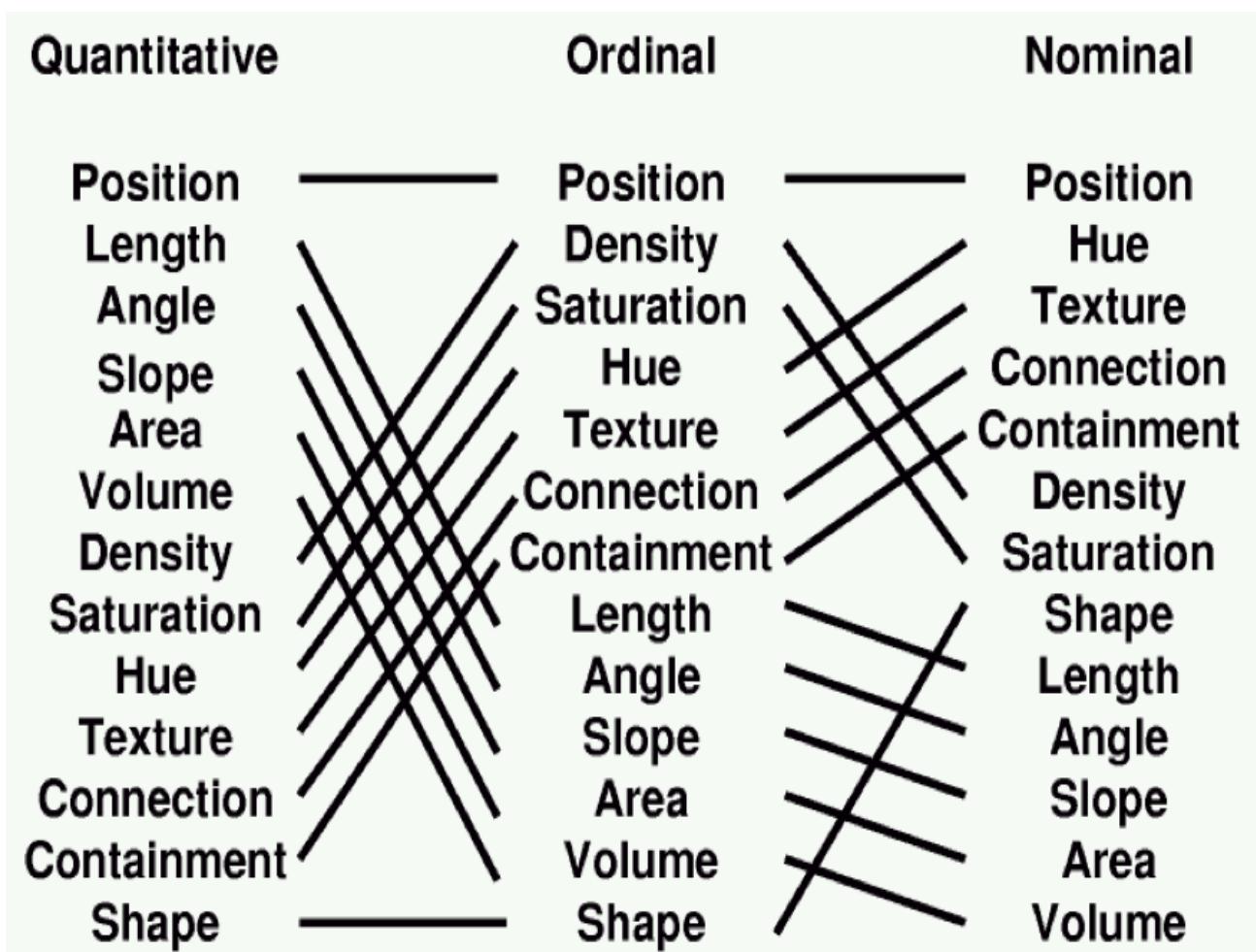
Effectiveness: Accuracy Ranking for Quantitative Information



Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Conjectured Effectiveness of Encodings by Data Type

- Nominal/
Ordinal
variables:
detect
differences
- Quantitative
variables:
estimate
magnitudes

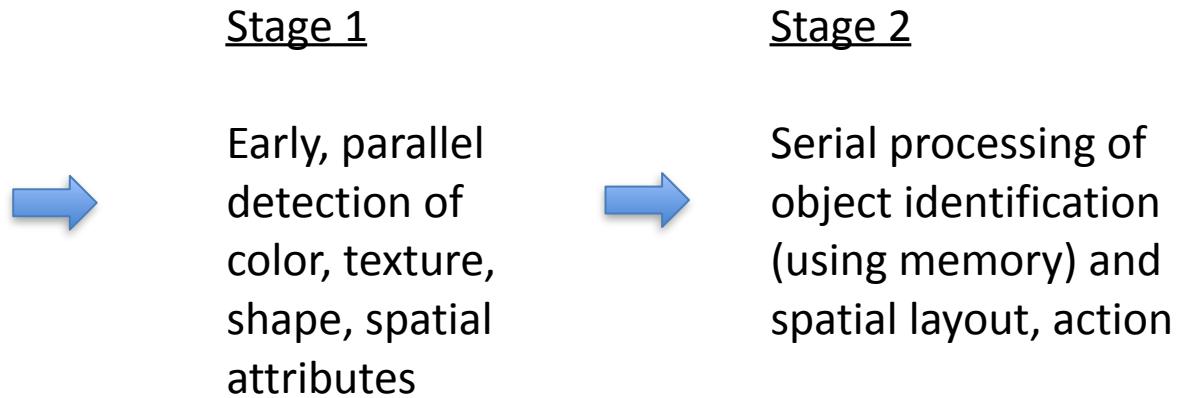
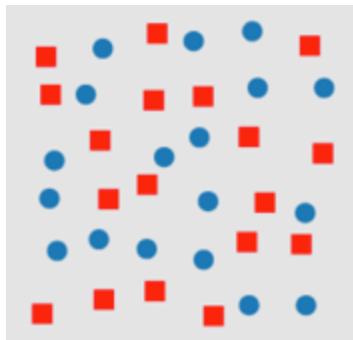


Mackinlay, Automating the design of graphical presentations of relational information, 1986.

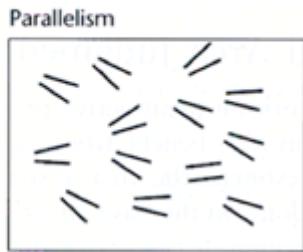
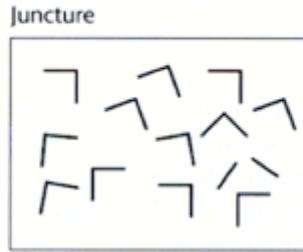
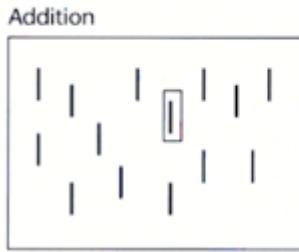
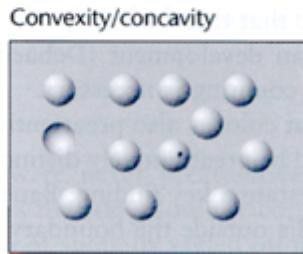
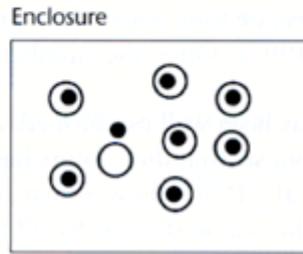
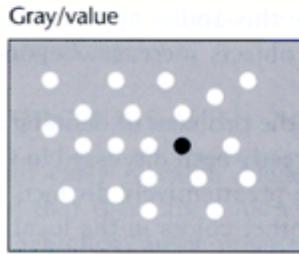
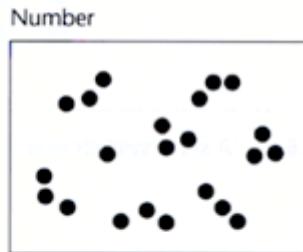
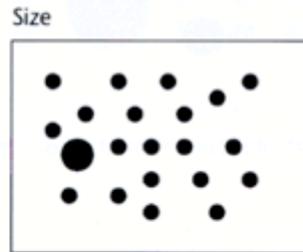
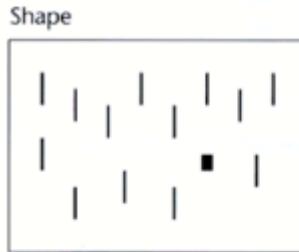
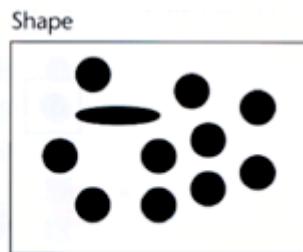
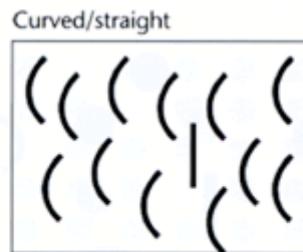
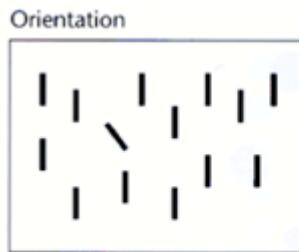
Visual Perception

Perceptual Processing Model

- Two stage process
 - Parallel extraction of low-level properties of scene
 - Sequential goal-directed processing



Pre-Attentive Features



- length
- width
- size
- curvature
- number
- terminators
- intersection
- closure
- hue
- intensity
- flicker
- direction of motion
- binocular lustre
- stereoscopic depth
- 3-D depth cues
- lighting direction

Pre-Attentive Feature Conjunctions

- Spatial conjunctions are often pre-attentive
- Motion and 3D disparity
- Motion and color
- Motion and shape
- 3D disparity and color
- 3D disparity and shape
- Most conjunctions are not pre-attentive

Gestalt Grouping Principles

“All else being equal, elements that are related by X tend to be grouped perceptually into higher-order units.”

— Stephen Palmer

- Proximity
- Similarity
- Connectedness
- Continuity
- Symmetry
- Closure
- Figure/Ground
- Common Fate

Change Blindness

- We don't always see everything that is there!
- Is the viewer able to perceive changes between two scenes?
 - If so, may be distracting
 - Can do things to minimize noticing changes
- Video: <http://www.simonslab.com/videos.html>

Summary of Design Criteria

- Choose expressive and effective encodings
 - Rule-based tests of expressiveness
 - Perceptual effectiveness rankings
 - Prioritizes encodings that are most easily/accurately interpreted
 - Principle of Importance Ordering: Encode more important information more effectively (Mackinlay)

Interaction

Representation and Interaction

- Two main components of information visualization
- Very challenging to come up with innovative, new visual representations
- But can do interesting work with how user interacts with the view or views
 - Analysis is a process, often iterative with different interactions

“The effectiveness of information visualization hinges on two things: its ability to clearly and accurately represent information and our ability to interact with it to figure out what the information means.”

S. Few, <Now you see it>

Taxonomy of Interactions

- Dix and Ellis (1998)
 - Highlighting and focus;
 - accessing extra info;
 - overview and context;
 - same representation, changing parameters;
 - Linking representations
- Keim (2002)
 - Projection
 - Filtering
 - Zooming
 - Distortion
 - Linking and brushing
- Few's Principles
 - Comparing
 - Sorting
 - Adding variables
 - Filtering
 - Highlighting
 - Aggregating
 - Re-expressing
 - Re-visualizing
 - Zooming and panning
 - Re-scaling
 - Accessing details on demand
 - Annotating
 - Bookmarking

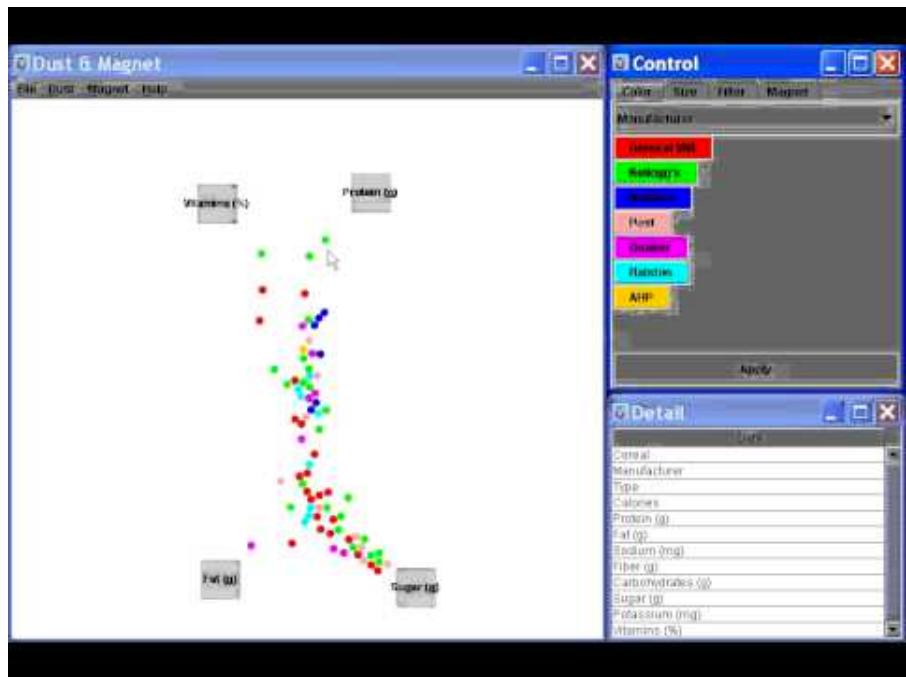
A Summary of Existing Taxonomy

- Survey
 - 59 papers
 - Papers introducing new interaction systems
 - Well-known papers in subareas of information visualization
 - 51 systems
 - Commercial Infovis Systems (SeeIT, Spotfire, TableLens, InfoZoom, etc.)
 - Collected 311 individual interaction techniques
- Affinity Diagram Method

Yi, Ji Soo, Youn ah Kang, and John Stasko. "Toward a deeper understanding of the role of interaction in information visualization." IEEE transactions on visualization and computer graphics 13.6 (2007): 1224-1231.

Taxonomy of Interactions based on User Intent

- 7 Categories
- Select
- Explore
- Reconfigure
- Encode
- Abstract/Elaborate
- Filter
- Connect



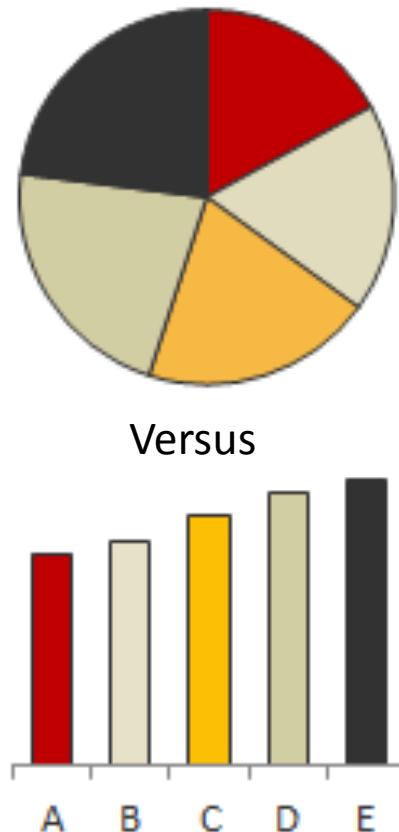
Interaction is Vital for Exploration

- Interaction facilitates a dialog between the user and the visualization system
- Multiple views amplify importance of interaction
- Interaction often helps when you just can't show everything you want

Evaluation

How do We Evaluate Visualizations?

- How do we evaluate visualizations?
 - Usability vs. Utility
- What evaluation techniques should we use?
- What do we measure?
 - What data do we gather?
 - What metrics do we use?



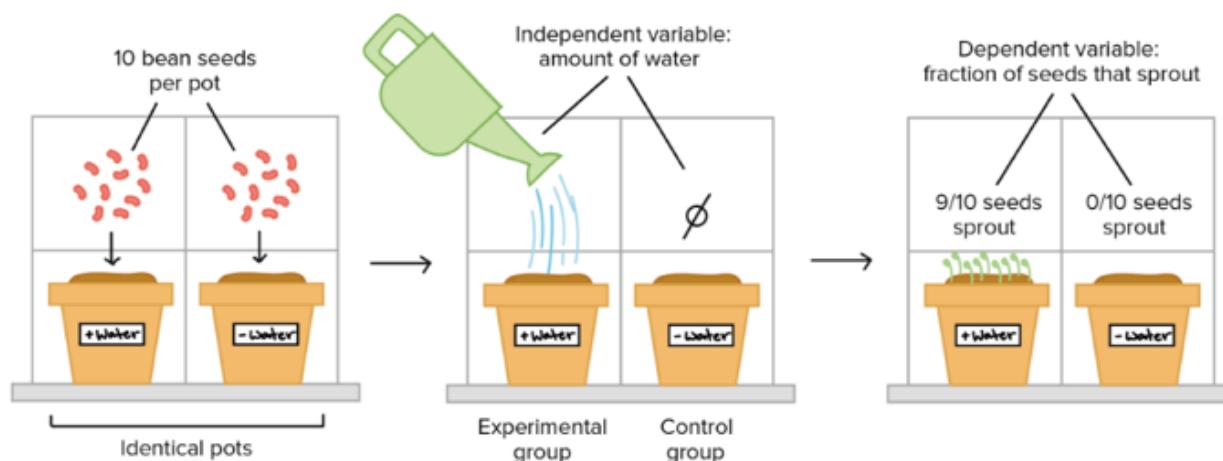
Evaluation Approaches

- Many different Forms
 - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, ...
- Two popular methodologies
 - Controlled experiments (Quantitative)
 - Subjective assessments (Qualitative)

Quantitative Methods: Controlled Experiments

- Good for measuring performance or comparing multiple techniques
- What do we measure?
 - Performance, time, errors,

...



An Example: Controlled Experiment

- Run an experiment to evaluate three visualization strategies
 - Animation
 - Small multiples
 - Traces
- Especially interested in examining whether animated bubble charts are beneficial for analysis and presentation



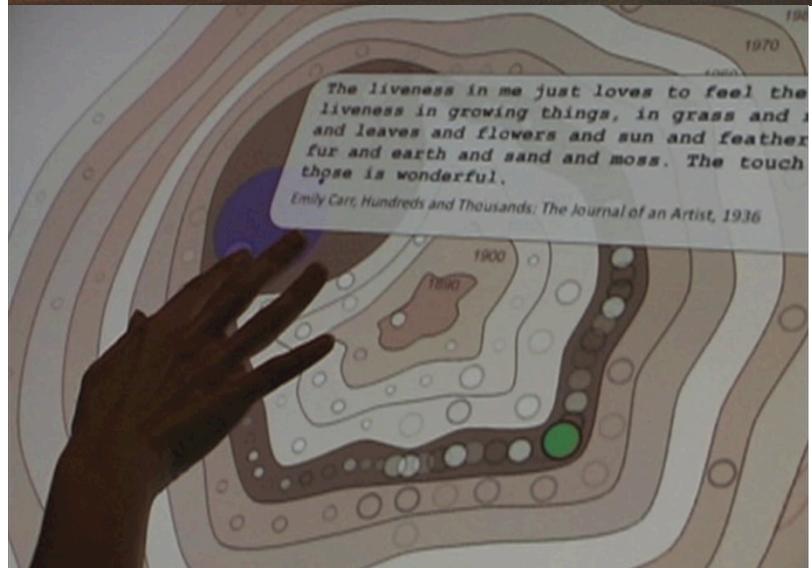
*Do you remember Hans Rosling's TED talk?
(Lecture 2)*

Qualitative Methods

- Types
 - Nested methods
 - Experimenter observation, think-aloud protocol, collecting participant opinions
 - Inspection evaluation methods
 - Heuristics to judge
- Observational context
 - In situ, laboratory, participatory
 - Contextual interviews is important

An Example: Subjective Assessments

- Evaluating a newly developed visualization system (EMDialog) at the museum (Emily Carr exhibit)
- Discourse visualization
 - Time
 - Context



Methodology vs. Desirable Features

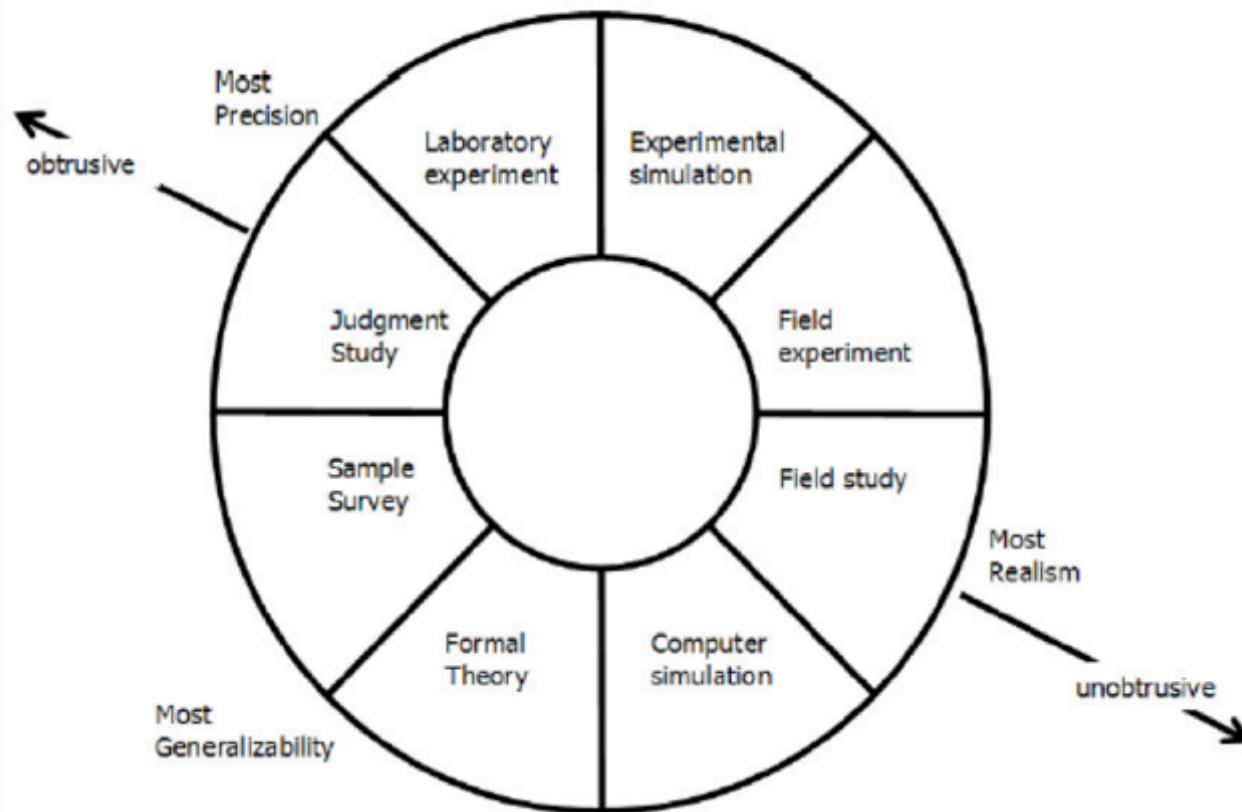
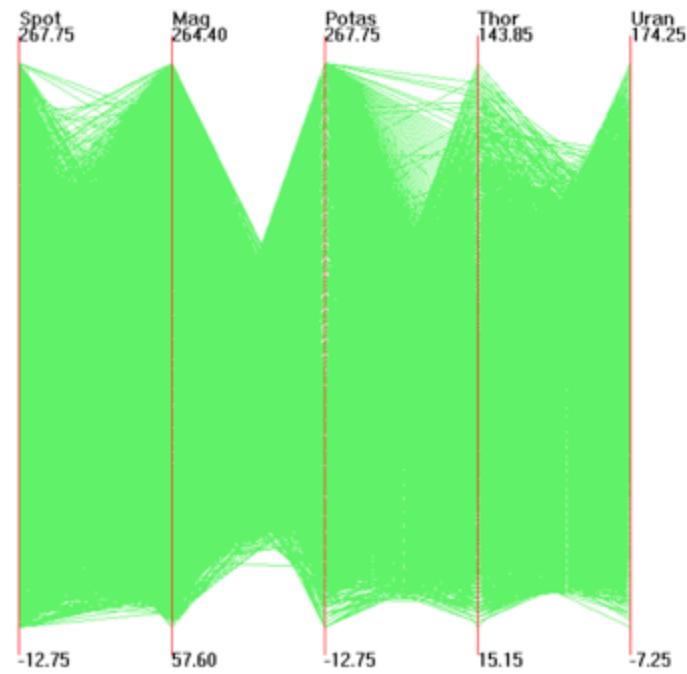


Fig. 1. Types of methodologies organized to show relationships to precision, generalizability and realism. (adapted, simplified from McGrath 1995)

Data Overload

Data Overload

- Most of the techniques we've examined work for a modest number of data cases or variables
- What happens when you have lots and lots of data cases and/or variables?



Out5d dataset(5 dimensions, 16384 items)

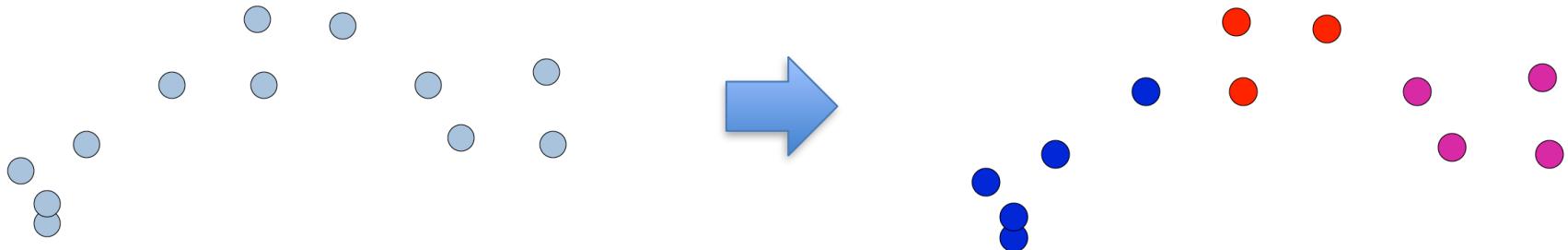
General Solution

- Data that is similar in most dimensions ought to be drawn together
 - Cluster at high dimensions
- Need to project the data down into the plane and give it some ultra-simplified representation
- Or perhaps only look at certain aspects of the data at any one time

Clustering and Dimensionality Reduction

- There exist many techniques for clustering high-dimensional data with respect to all those dimensions (too many data points)
 - Affinity propagation
 - k-means
 - Expectation maximization
 - Hierarchical clustering
- There exist many techniques for projecting n-dimensions down to 2-D (dimensionality reduction, too many variables)
 - Multi-dimensional scaling (MDS)
 - Principal component analysis (PCA)
 - Linear discriminant analysis
 - Factor analysis

K-means Clustering

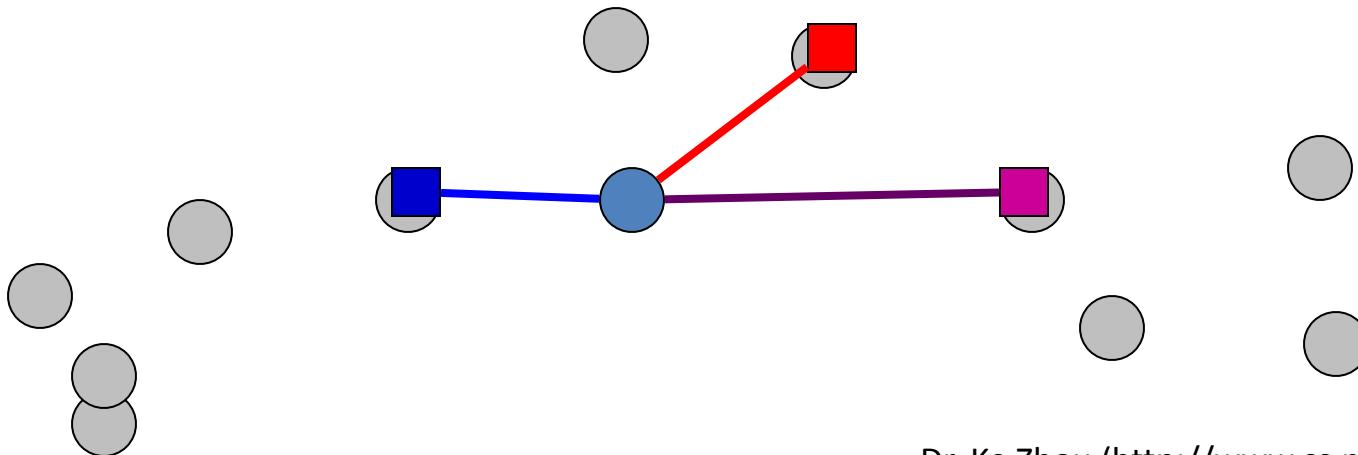


- The most well-known and popular clustering algorithm

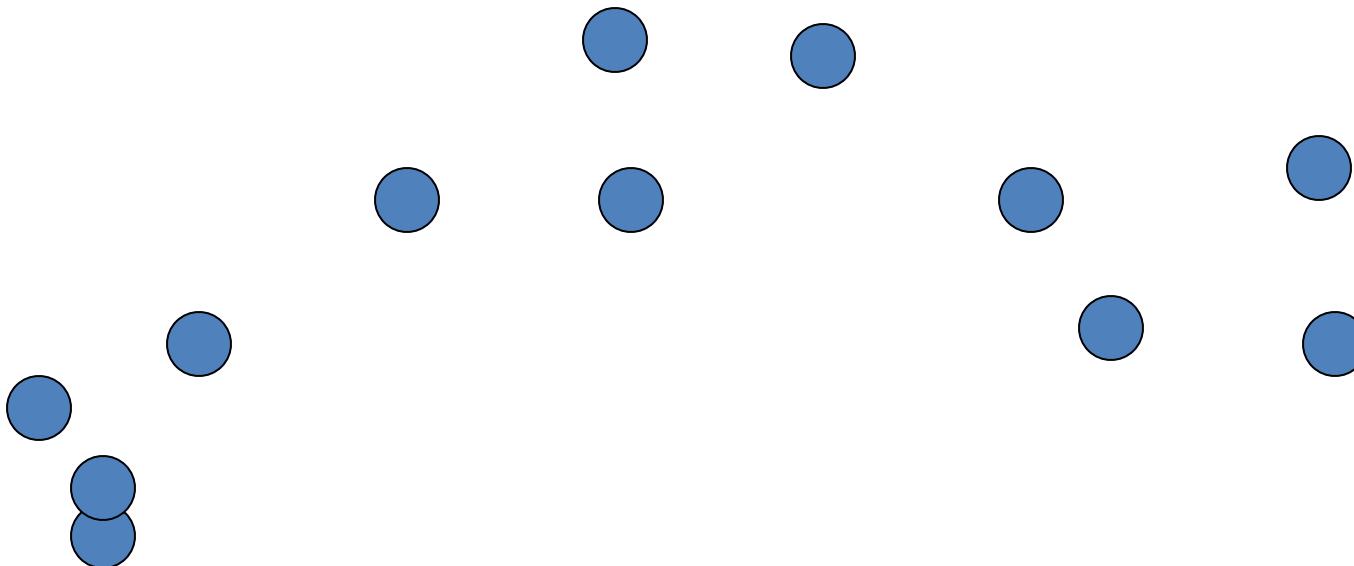
K-means Algorithm

Iterate:

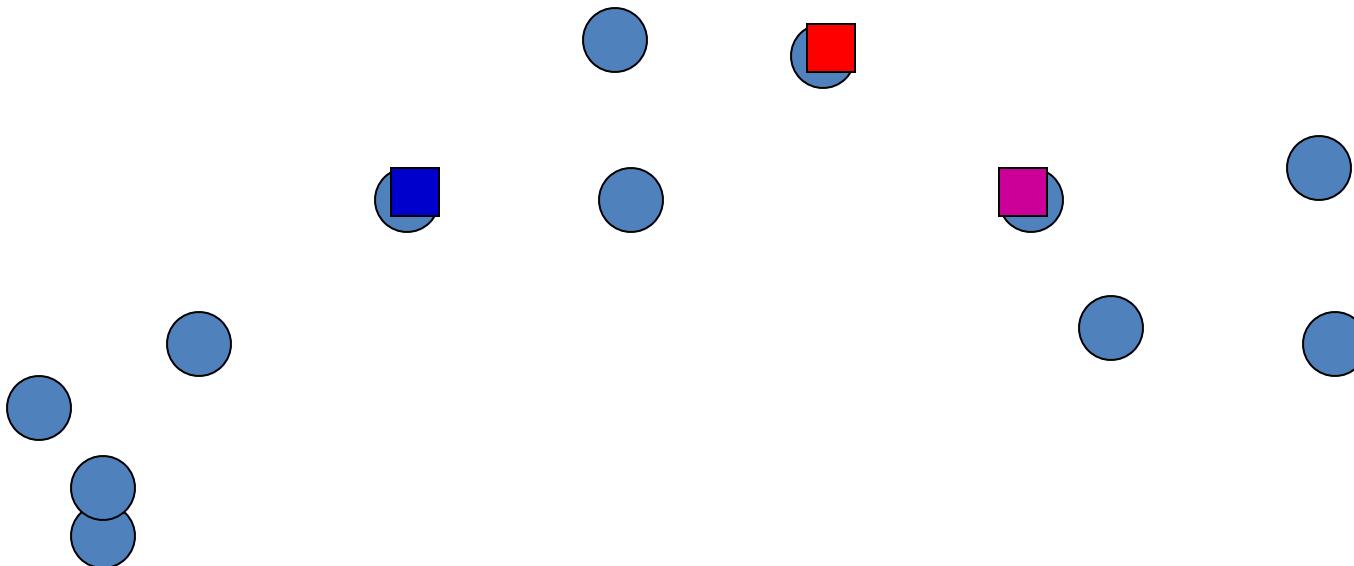
- Start with some initial cluster centers
- Assign/cluster each example to closest center
 - iterate over each point:
 - get distance to each cluster center
 - assign to closest center (hard cluster)
- Recalculate centers as the mean of the points in a cluster



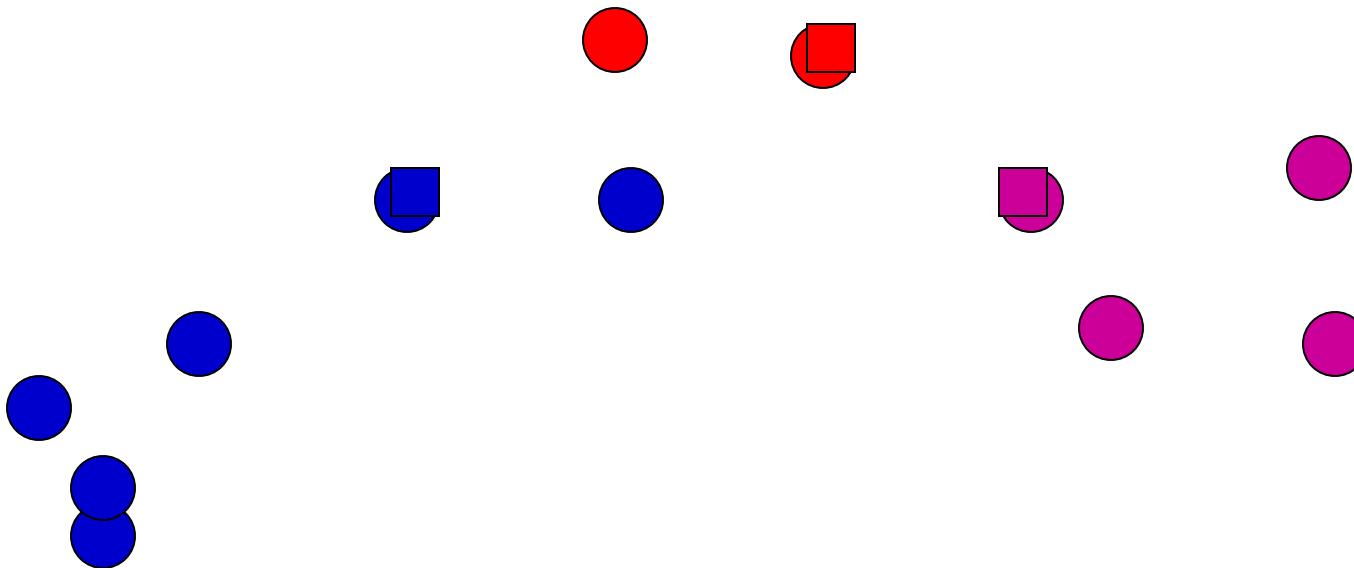
K-means: an example



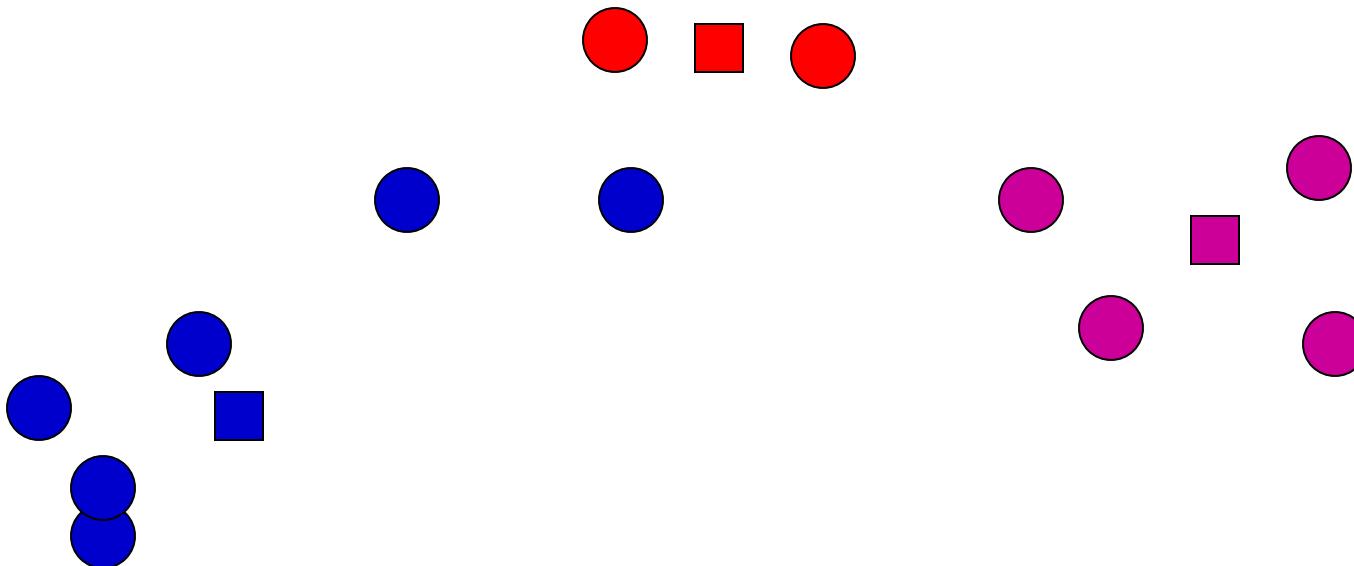
K-means: Initialize centers randomly



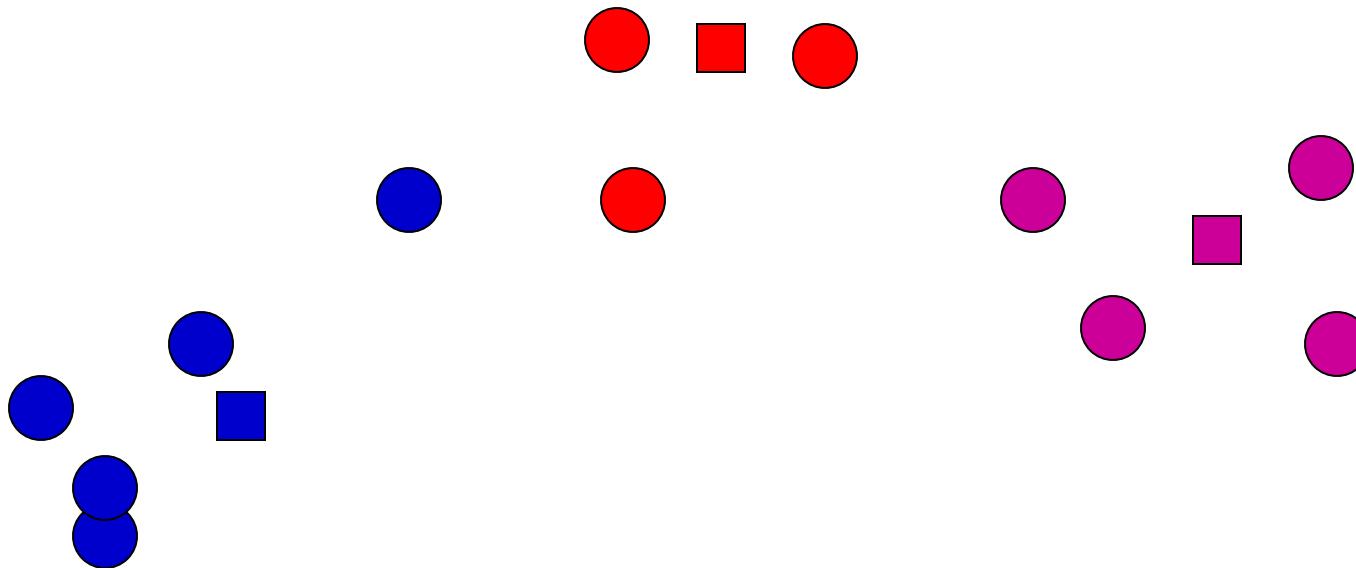
K-means: assign points to nearest center



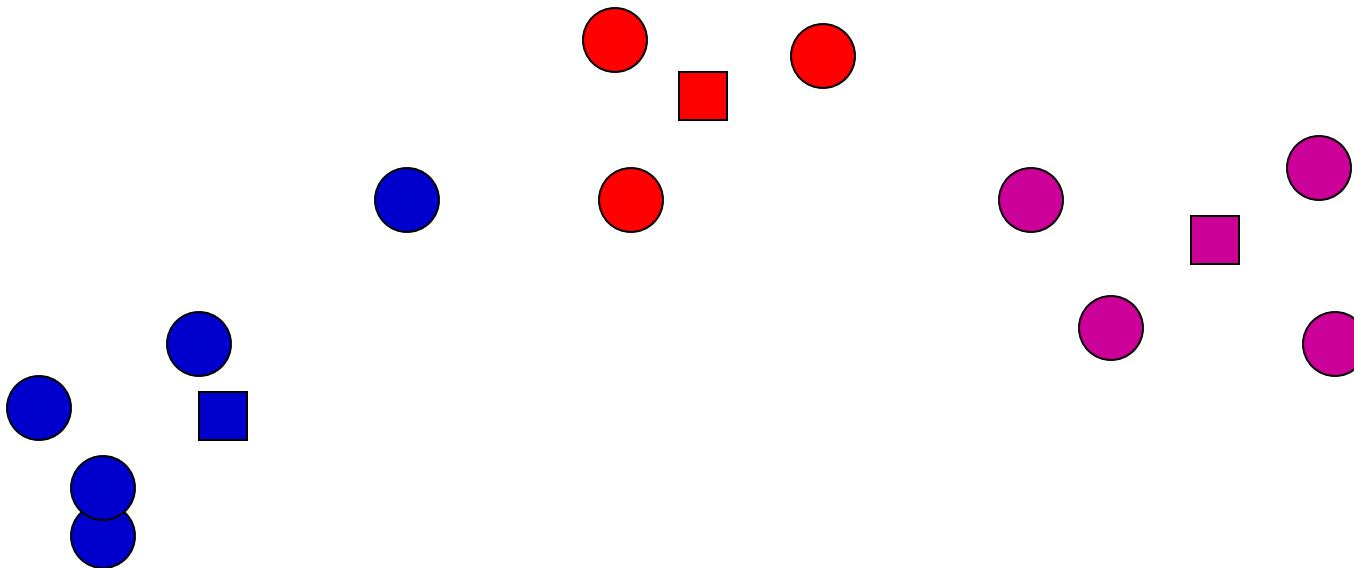
K-means: readjust centers



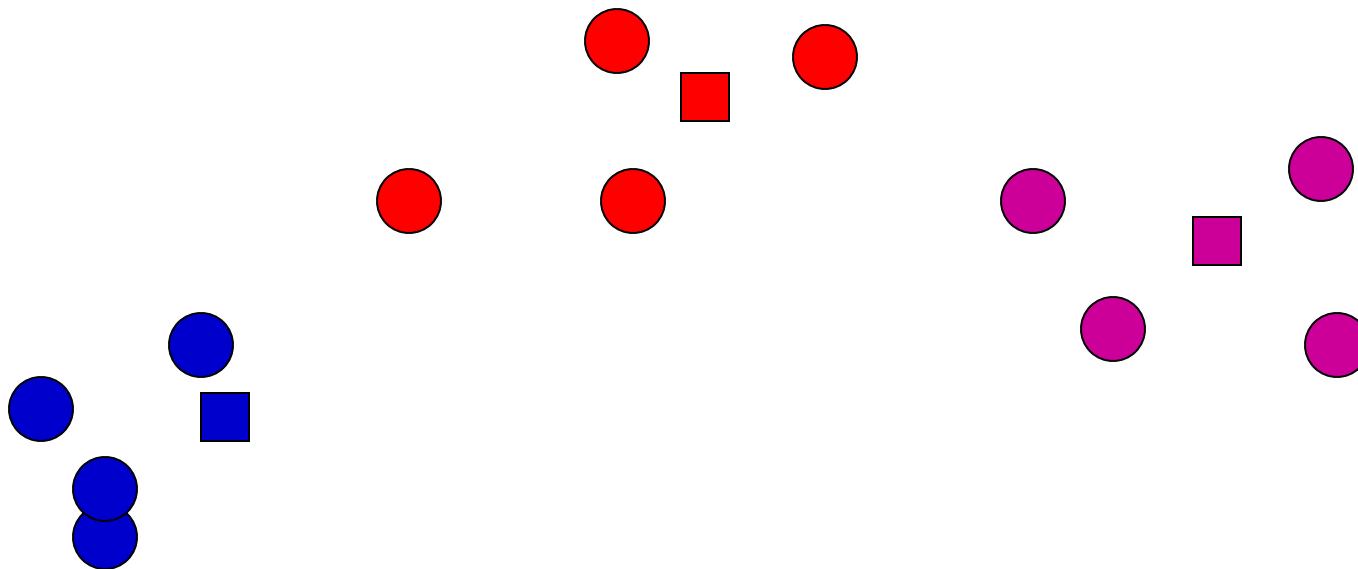
K-means: assign points to nearest center



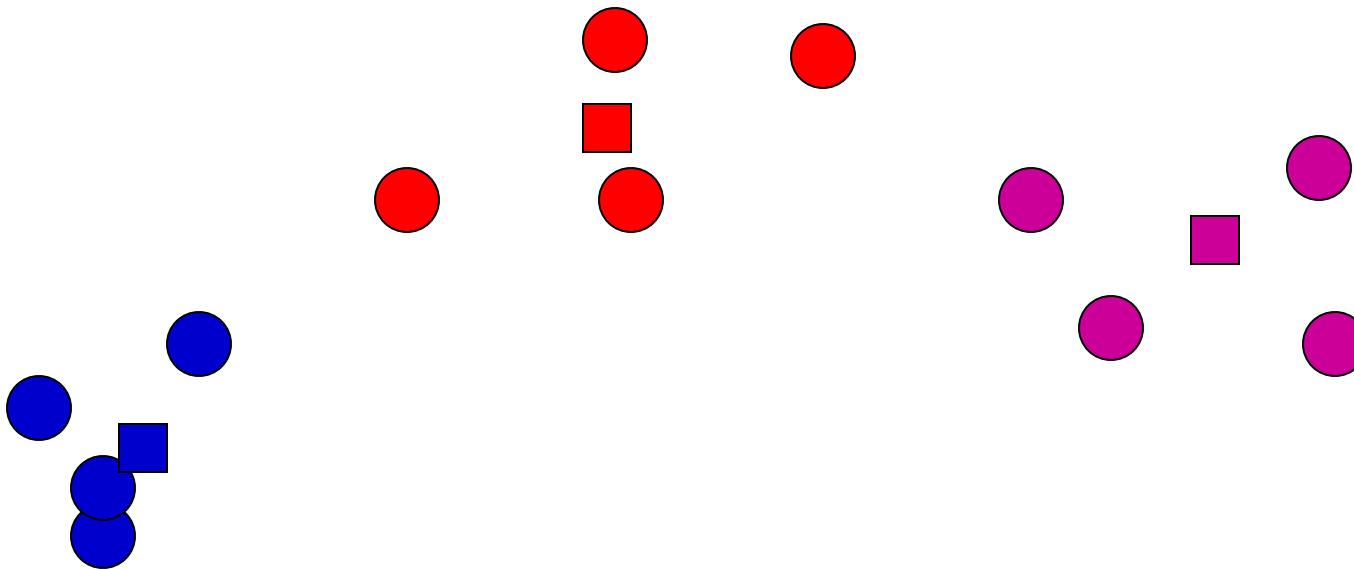
K-means: readjust centers



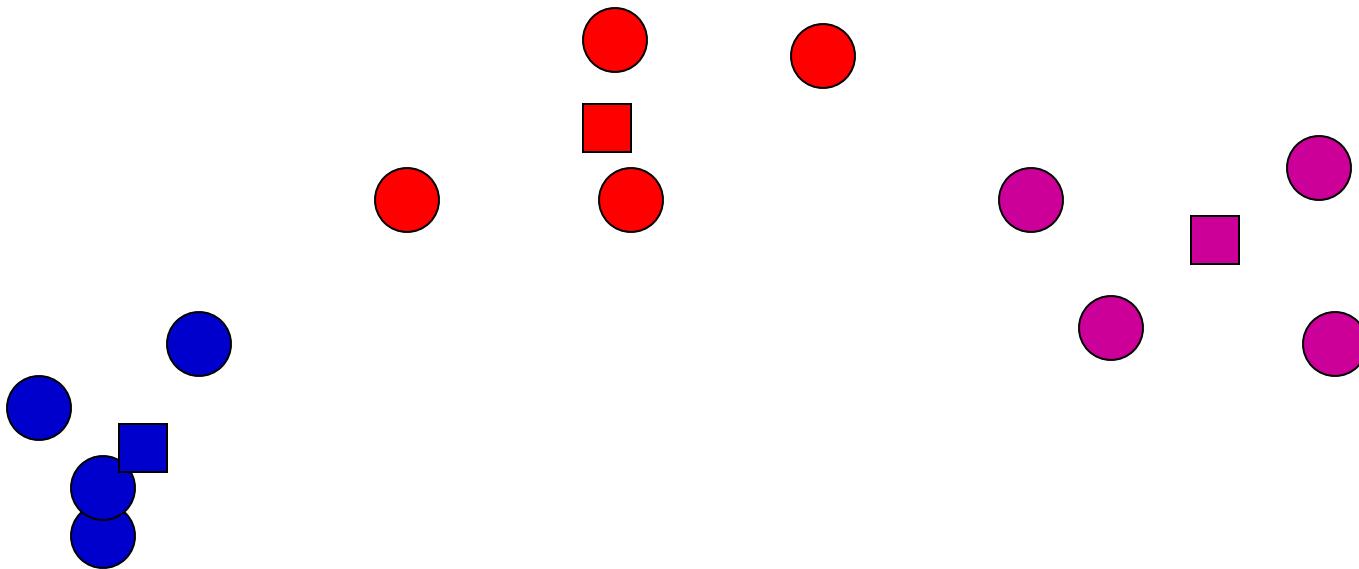
K-means: assign points to nearest center



K-means: readjust centers



K-means: assign points to nearest center



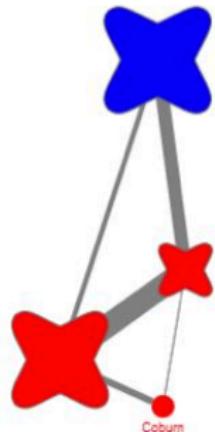
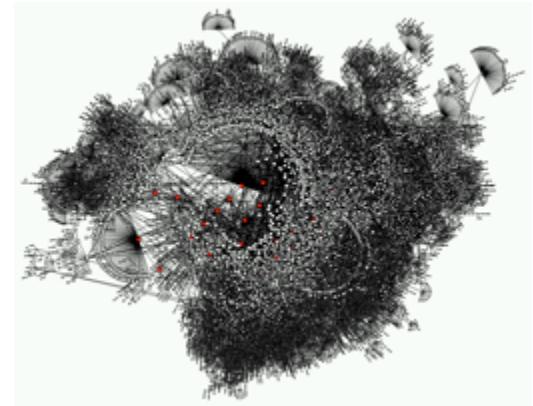
No changes: Done

Other Reduction Techniques

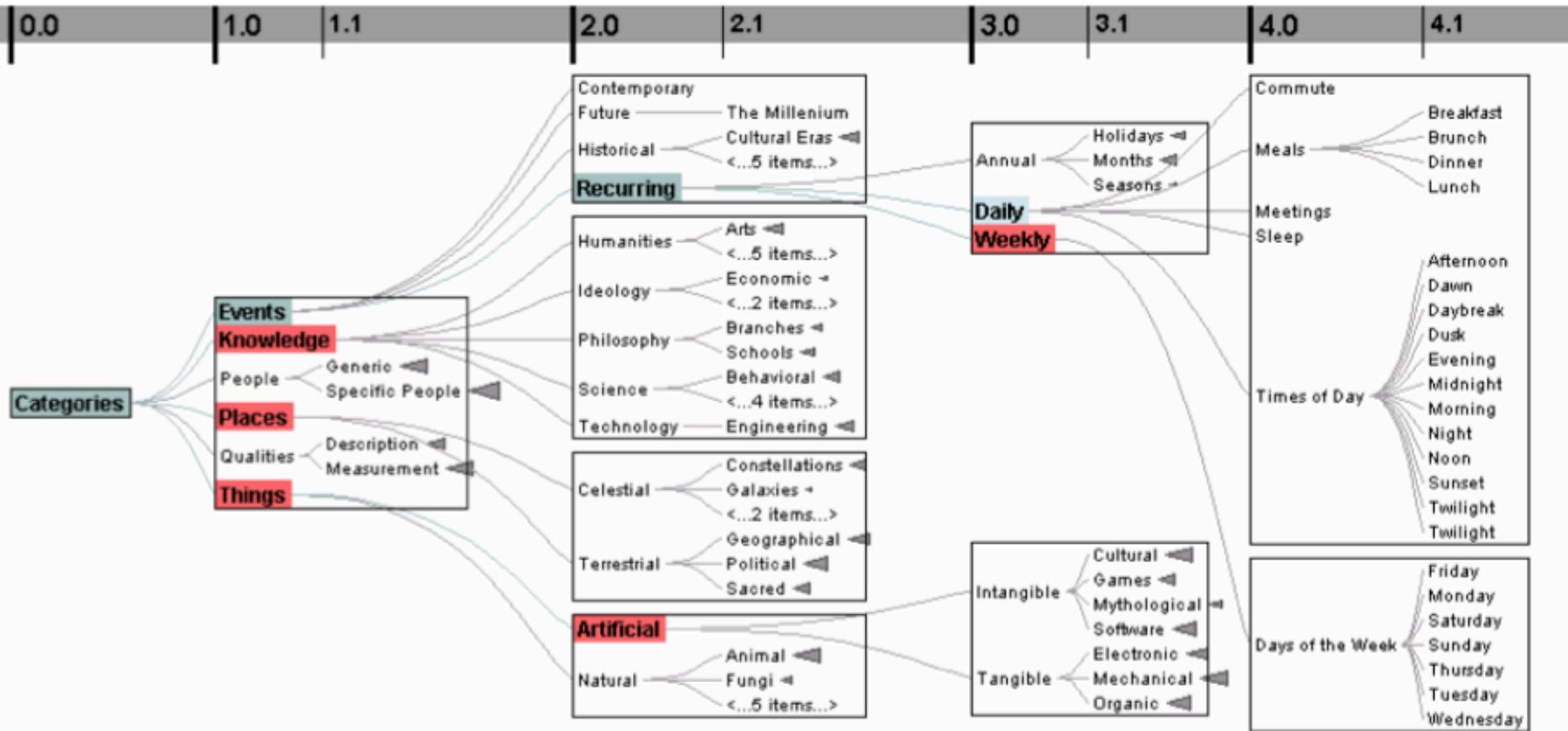
- Other techniques exist to manage scale
 - Sampling: only including every so many data cases or variables
 - Aggregation: combining many data cases or variables
- Interaction
 - Employ user interaction rather than special renderings to help manage scale

Application: Scalability Issue of Graphs

- Need to cope with messiness
- Solutions
 - Extracting network motifs
 - Taking advantage of node attributes
 - Degree-of-Interest graphs
 - Use the alternative representation: matrix

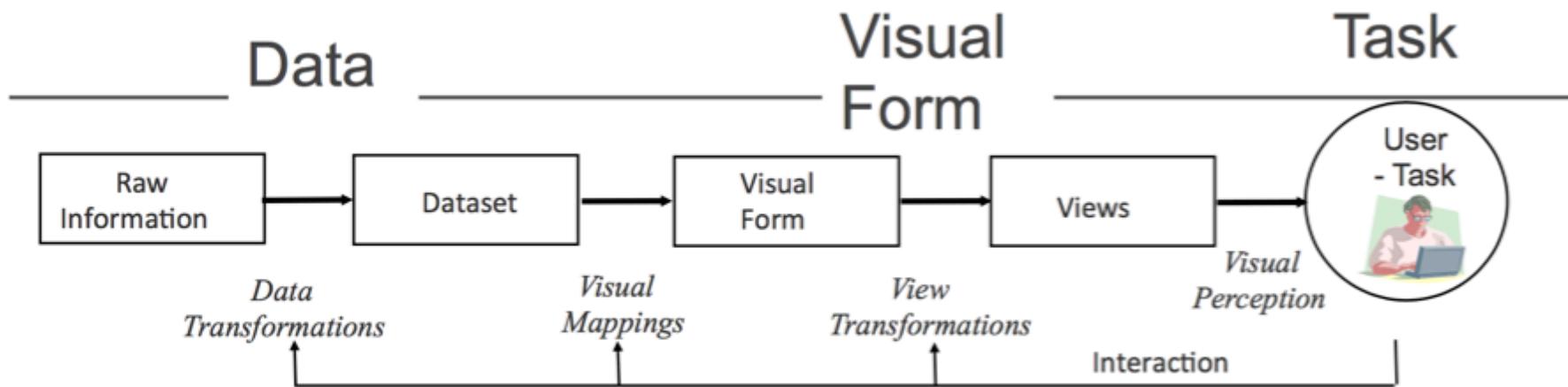


Interactive: Degree-of-interest Trees/Graphs



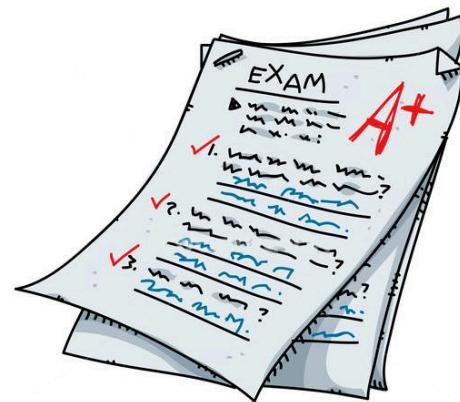
- Cull “un-interesting” nodes on a per block basis until all blocks on a level fit within bounds.
- Attempt to center child blocks beneath parents.

Information Visualization



Next Lecture

- Topic:
 - Review



G53FIV: Fundamentals of Information Visualization

Lecture 14: Review

Ke Zhou
School of Computer Science
Ke.Zhou@nottingham.ac.uk

<https://moodle.nottingham.ac.uk/course/view.php?id=68644>

A Bit About the Exam

What to Learn

- Lecture slides
- Selected chapters from the core text books and additional reading papers (all available on Moodle)
 - [The Visual Display of Quantitative Information](#) (2nd Edition). E. Tufte. Graphics Press, 2001 [available in the library].
 - [R Graphics Cookbook](#), Winston Chang, O'Reilly Media, 2013 [you can find it online by googling].
 - [Paper handouts](#).

Lecture Schedule

Week	Topic	Topic
1 (w19)	Introduction	The Value of Visualization
2 (w20)	Data and Image Models	Graphs and Charts
3 (w21)	Multivariate Data Visualization	Visualization with R - Fundamentals
4 (w22)	Visualization with R - Advanced	Visualization Tools and Visual Perception
5 (w24)	Interaction	Evaluation
6 (w25)	Visualizing Text and Documents	Visualizing Time Series, Trees and Graphs
7 (w26)	Recap of Fundamentals	Review
Break		
8 (w33)	Demo	Demo

Exam Format

- Two hours
- The written exam accounts for 75% of the whole module
- 4 questions relating to different aspects of information visualization (learned within the module)
 - Different sub-questions
- Bring your pen

Question Types

- Examples
 - Describe the definition or the key concept
 - Compare and assess different information visualizations
 - How to manipulate data
 - And others...

Past Paper

- The past paper is available on Moodle.
- You should practice with it after the revision.

Practice Questions

Describe a Key Concept

- Describe the three basic data types. Assess each column of the table on the corresponding data type.

	Student 1	Student 2	Student 3	Student 4
Name	Tom	Jim	Mary	Jane
Age	20	19	22	21
Grade	A	B	A-	B+
Course	Math	Math	Art	Sport
Entry Year	1997	1998	1995	1996

Nominal, Ordinal & Quantitative

- N - Nominal (labels or categories)
 - Operations: $=, \neq$
e.g. math, art (course)
- O – Ordered
 - Operations: $=, \neq, <, >$
e.g. A, A-, B+, B (grade)
- Q - Interval (location of zero arbitrary)
 - Operations: $=, \neq, <, >, -$
– Can measure distances or spans
e.g. (3.23, -1.2) (GPS)
- Q - Ratio (zero fixed)
 - Operations: $=, \neq, <, >, -, \%$
– Can measure ratios or proportions
e.g. 20, 19, 22, 21 (age)

Expected Answer

- There are three basic data types: nominal (N), ordinal (O) and quantitative (Q).
- With respect to the data in the table, each row represents a data case. The column “name” denotes nominal (N) data; “age” represents quantitative (Q) data; “grade” denotes ordinal (O) data; “course” represents nominal (N) data; and “entry year” denotes quantitative (Q) data.

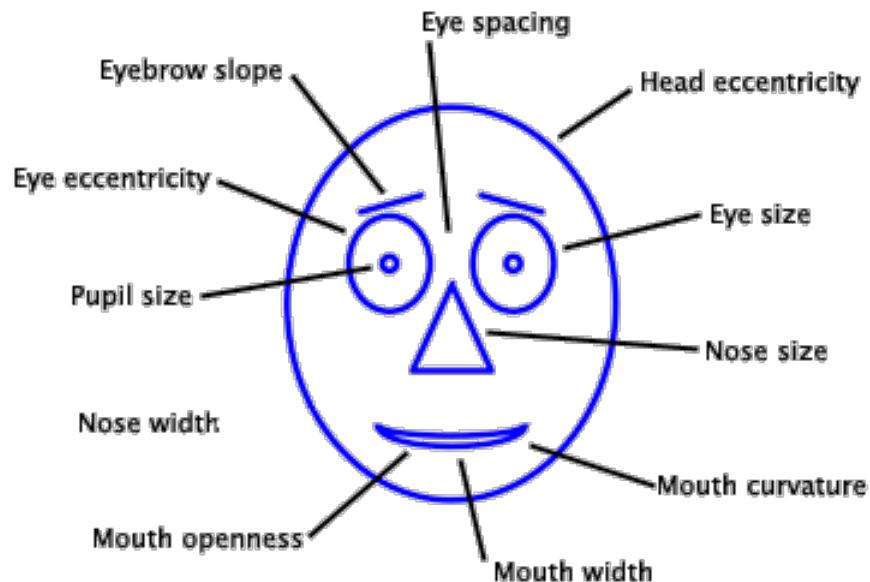
	Student 1	Student 2	Student 3	Student 4
Name (N)	Tom	Jim	Mary	Jane
Age (Q)	20	19	22	21
Grade (O)	A	B	A-	B+
Course (N)	Math	Math	Art	Sport
Entry Year (Q)	1997	1998	1995	1996

Compare Different Visualizations

- Compare and contrast two common techniques for visualizing multivariate data: Chernoff Faces and Parallel coordinates.
 - Explain Chernoff Faces and Parallel coordinates.
 - Identify the strengths and weaknesses in terms of “Find value of data case”

Chernoff Faces

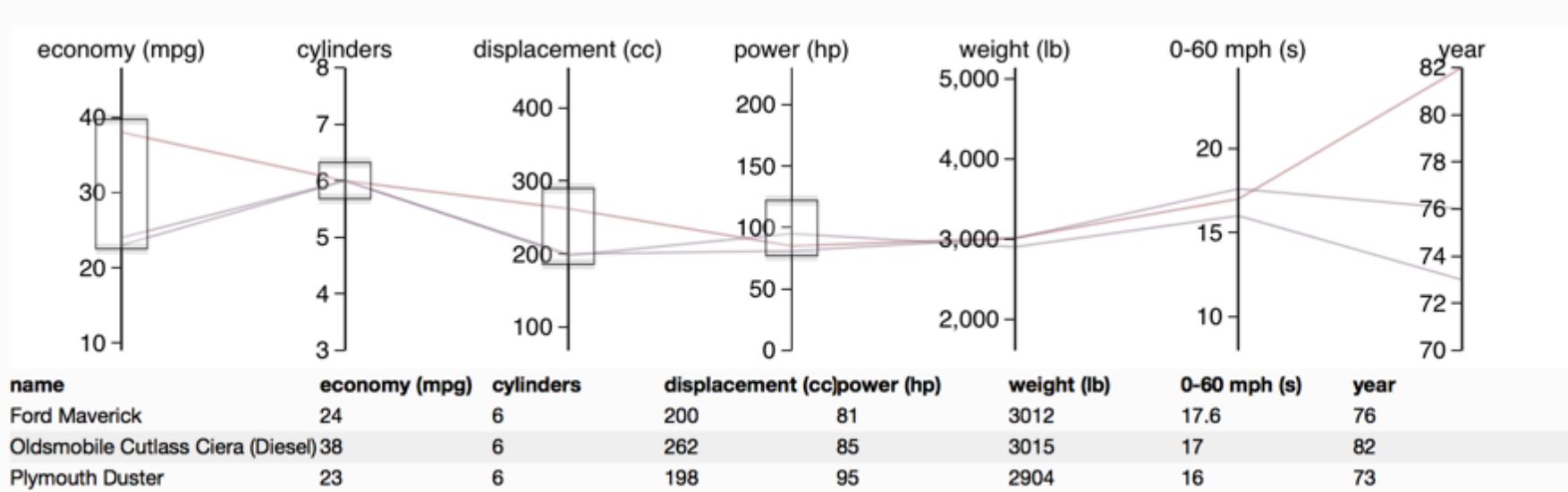
- Observation: We have evolved a sophisticated ability to interpret faces.
- Idea: Encode different variables' values in characteristics of human face



Chernoff, Herman. "The use of faces to represent points in k-dimensional space graphically." Journal of the American Statistical Association 68.342 (1973): 361-368.

Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline



Expected Answer

- Explain Chernoff Faces and Parallel coordinates
 - Chernoff faces exploits the individual parts, such as eyes, ears, and nose of the face to represent values of the variables.
 - In a parallel coordinates plot, the axes are placed in parallel and each data point is represented as a series of line segments intersecting the axes at the corresponding values.
- Find value of data case
 - Parallel coordinates are more suitable for finding value of data case when the data is of high dimension;
 - It is more difficult to find value in Chernoff faces, but it is easier to recognize differences between data cases.

Data Manipulations

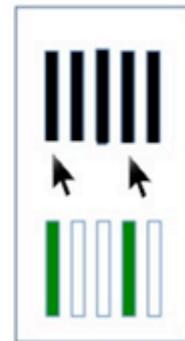
- List and describe the five most common data manipulation techniques.

5 Basic Verbs

- FILTER Rows



- SELECT Column Types



- ArRANGE Rows (SORT)



- Mutate (into something new)



- Summarize by Groups



dplyr

- dplyr takes the `%>%` operator and uses it to great effect for manipulating data frames
 - Works only with data frames
 - 5 basic “verbs” work for 90% of data manipulations

Verbs	What does it do?
<code>filter()</code>	Select a subset of ROWS by conditions
<code>arrange()</code>	Reorders ROWS in a data frame
<code>select()</code>	Select the COLUMNS of interest
<code>mutate()</code>	Create new columns based on existing columns (mutations!)
<code>summarise()</code>	Aggregate values for each group, reduces to single value

Expected Answer

- Filter: select a subset of data cases by a given condition.
- Arrange: reorder the data cases.
- Select: select a subset of the variables of interest.
- Mutate: create new variables of interest based on existing variables.
- Summarize: aggregate values for each group, reducing to single value.
- (Other answers may be correct as well, such as joining, etc.)

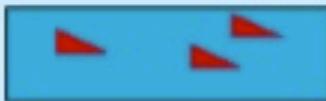
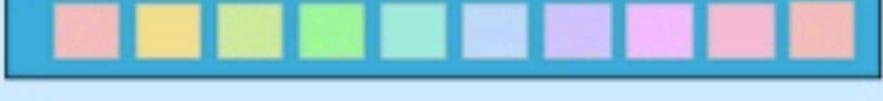
Visualization

- List and explain the visual encodings and their corresponding data types in this visualization.



Visual Encoding Variables

Bertin's Semiology of Graphics (1967)

- **position**
 - changes in the x, y, (z) location
- **size**
 - change in length, area or repetition
- **shape**
 - infinite number of shapes
- **value**
 - changes from light to dark
- **orientation**
 - changes in alignment
- **colour**
 - changes in hue at a given value
- **texture**
 - variation in pattern
- **motion**

Graphic by: Sheelagh Carpendale

Dr. Ke Zhou (<http://www.cs.nott.ac.uk/~pszkz/>)

Levels of Organization

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good

~ = OK

✗ = Bad

Expected Answer

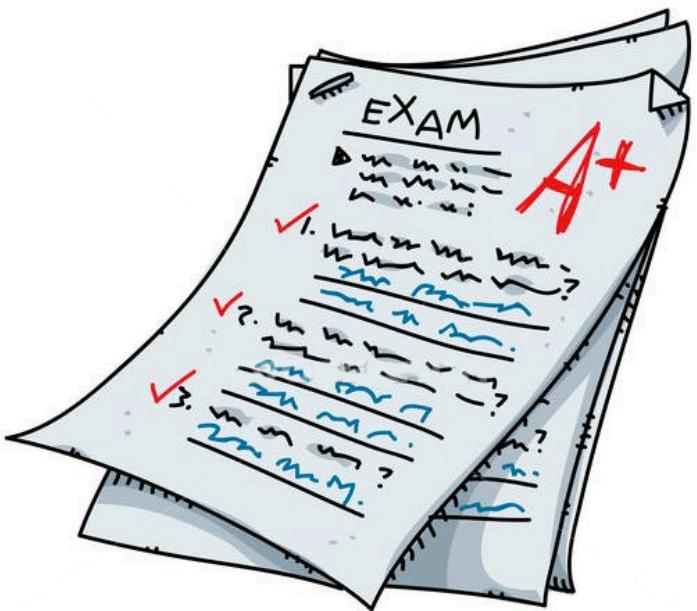
- Two visual encodings are used in this visualization: position and color.
- The x and y positions represent respectively the school (nominal data) and annual salary (quantitative data).
- The color Hue demonstrates different gender, which is of nominal data type.

Review Tips

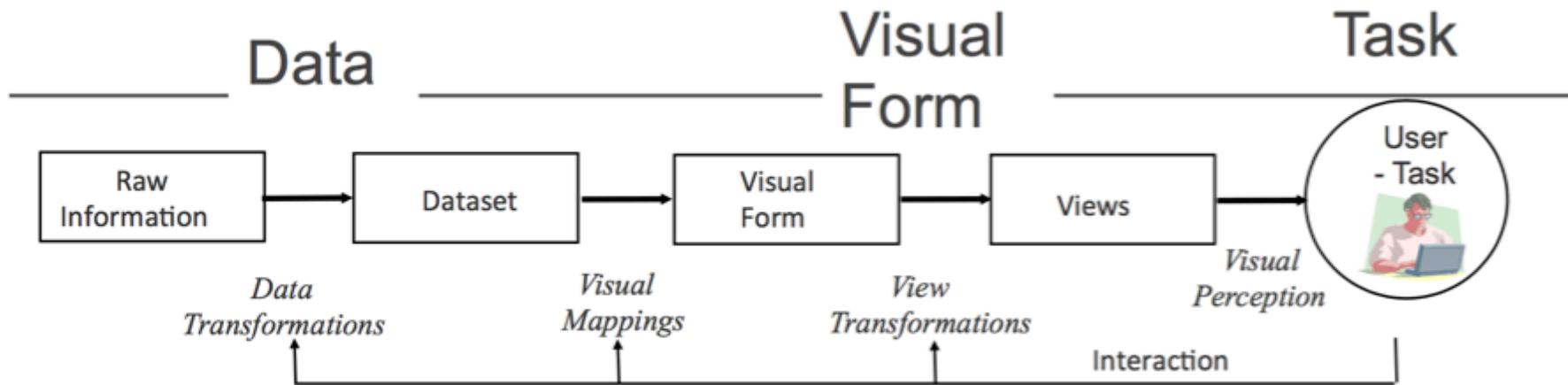
- You can find most of the key concepts or visualizations in the “recap of fundamentals” slides (Lecture 13).
 - A quick overview
- Review the lecture slides, the core texts and paper handouts.

Past Paper

- The past paper is available on Moodle.
- Let us go over it to see how you can get 100%.



Information Visualization



- Fundamental understanding on how visualizations convey information and how humans perceive
- Master an essential set of visualization techniques
- Practical experience in visualizing real-world data

SET/SEM Survey

SET/SEM Survey

- Official campus course evaluation
- Distributed and completed online. Your opinion is valued!



Module	Survey Type
Fundamentals of Information Visualisation	SET 
Fundamentals of Information Visualisation	SEM 

- Thanks for a great semester!

Next Lecture

- Topic:
 - Demo
- The Monday on 13 May
 - 12:00 - 14:00
 - A25, Business South, Jubilee Campus

