The University of Nottingham
UNITED KINGDOM • CHINA • MALAYSIA

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

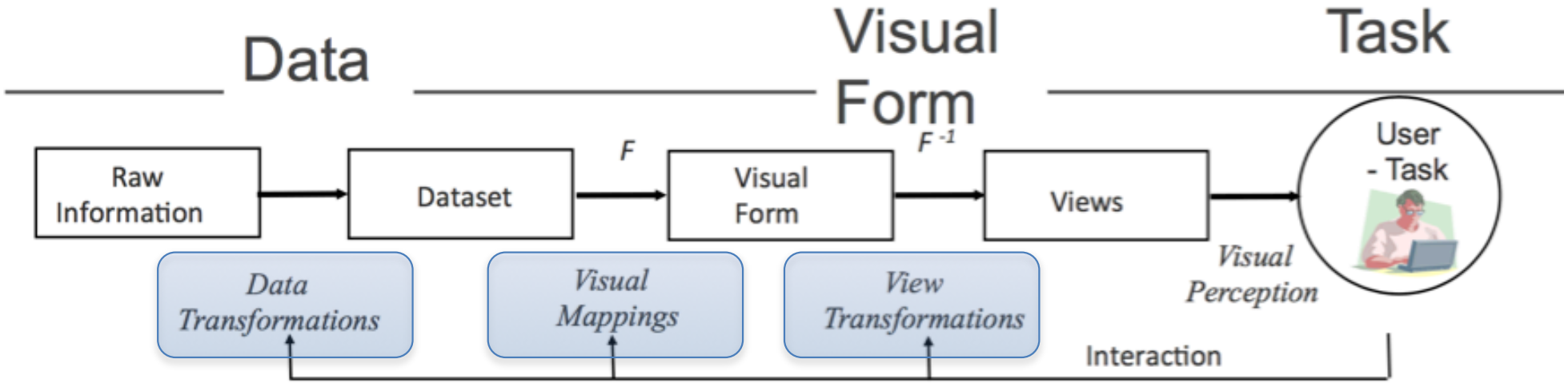
Summary of Reasons

- **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

Key Applications of IV

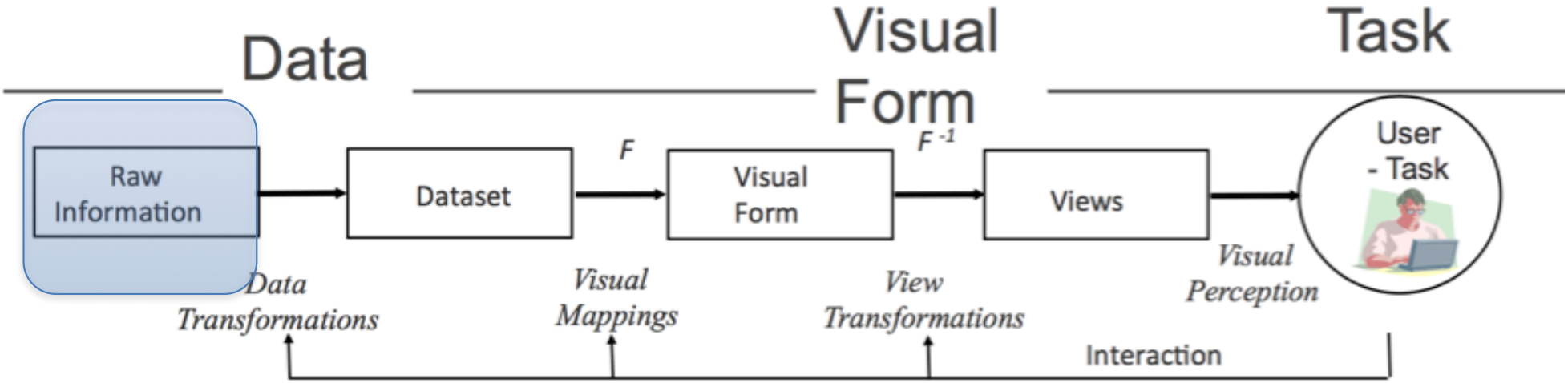
- I. **Record Information**
- II. **Communications (Presentation)**
  - Communicate data and ideas
  - Explain and inform
  - Provide evidence and support
  - Influence and persuade

- III. **Reasoning (Analysis)**
  - Explore the data
  - Assess a situation
  - Determine how to proceed
  - Decide what to do



- 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

Seven Stages: Acquire

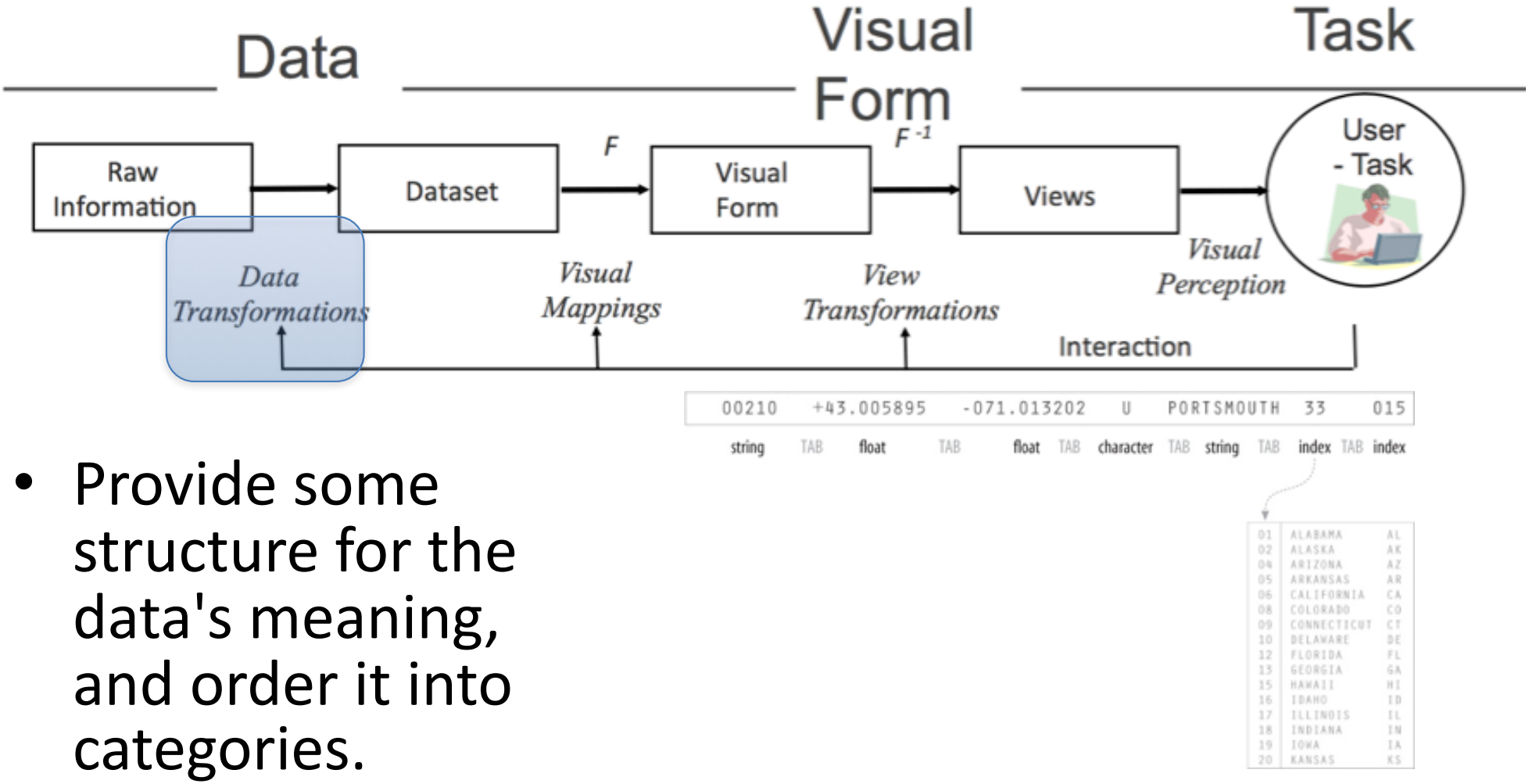


- Obtain the data, whether from a file on a disk or a source over a network

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

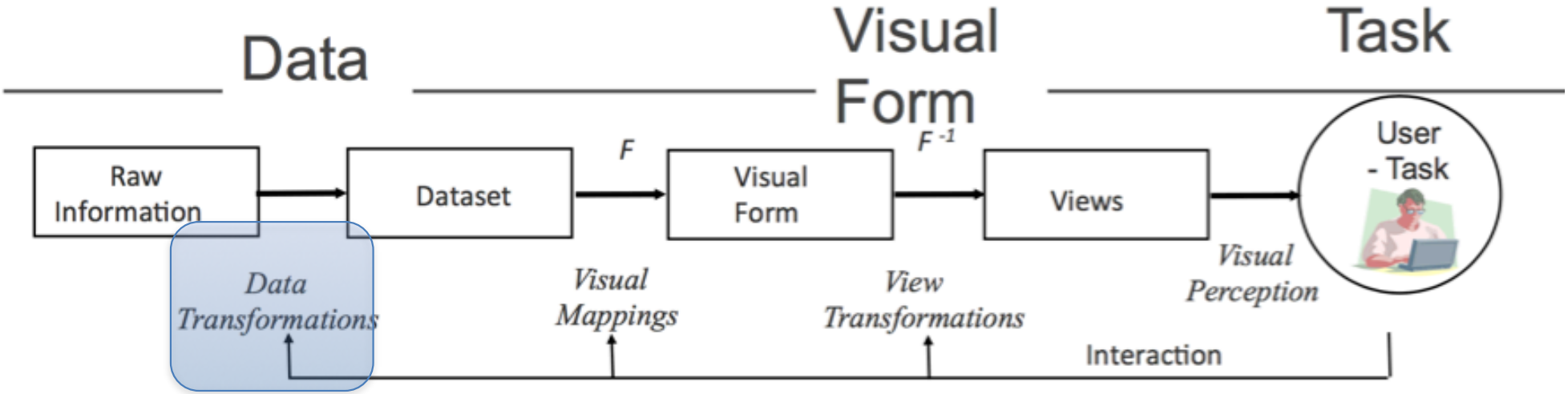
Seven Stages: Parse



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

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

Seven Stages: Filter

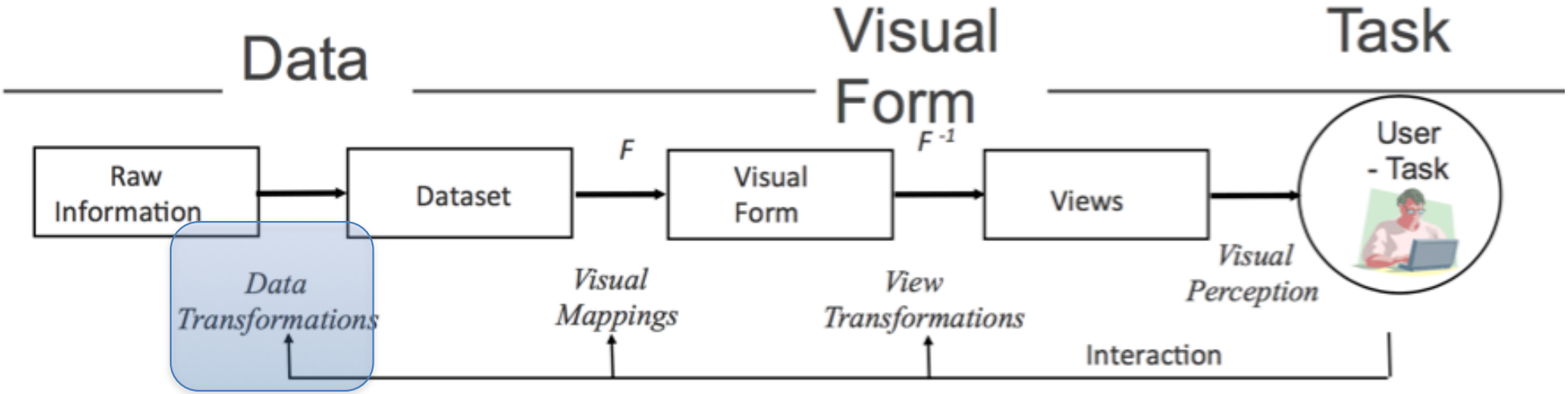


- Remove all but the data of interest.

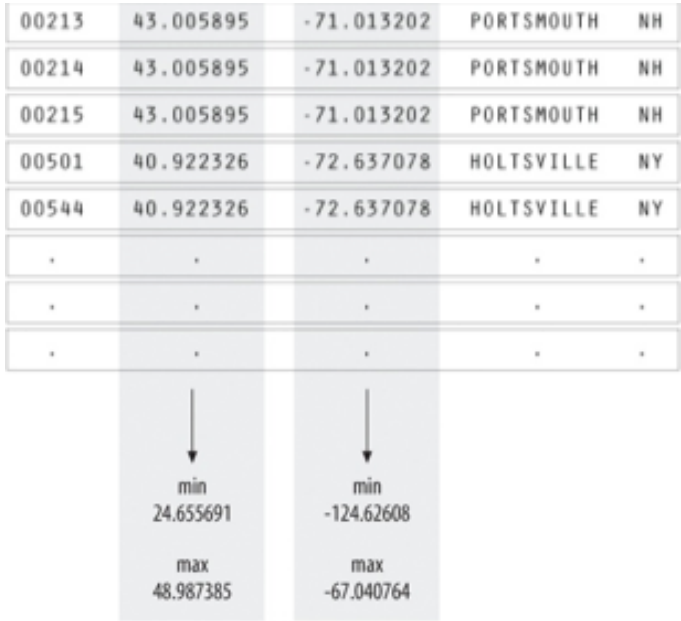
00210	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00211	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00212	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00213	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00214	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00215	+43.005895	-071.013202	U	PORTSMOUTH	33	015
00501	+40.922326	-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

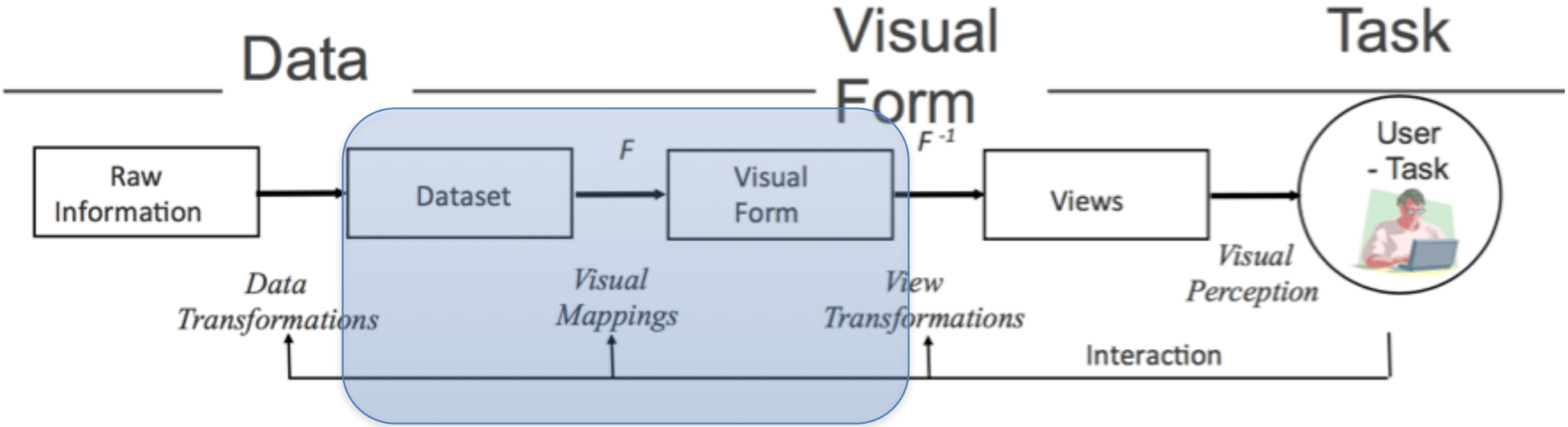
Seven Stages: Mine



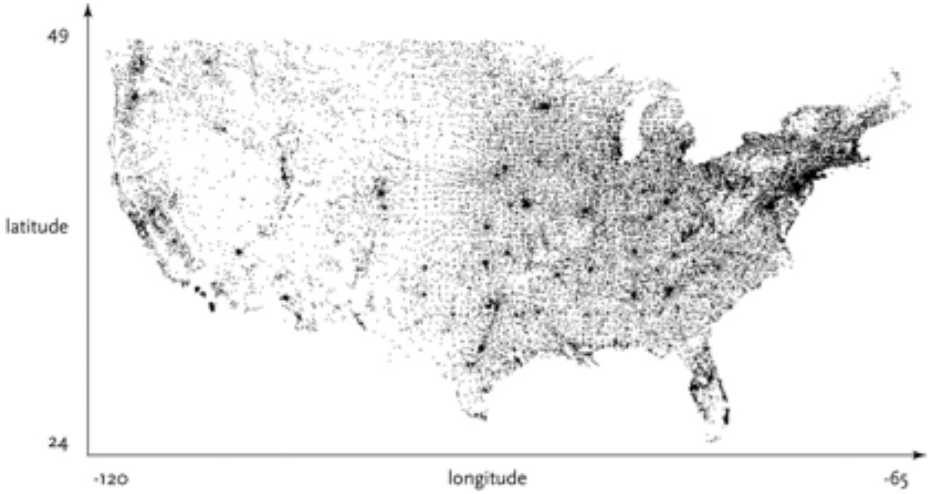
- Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.



Seven Stages: Represent

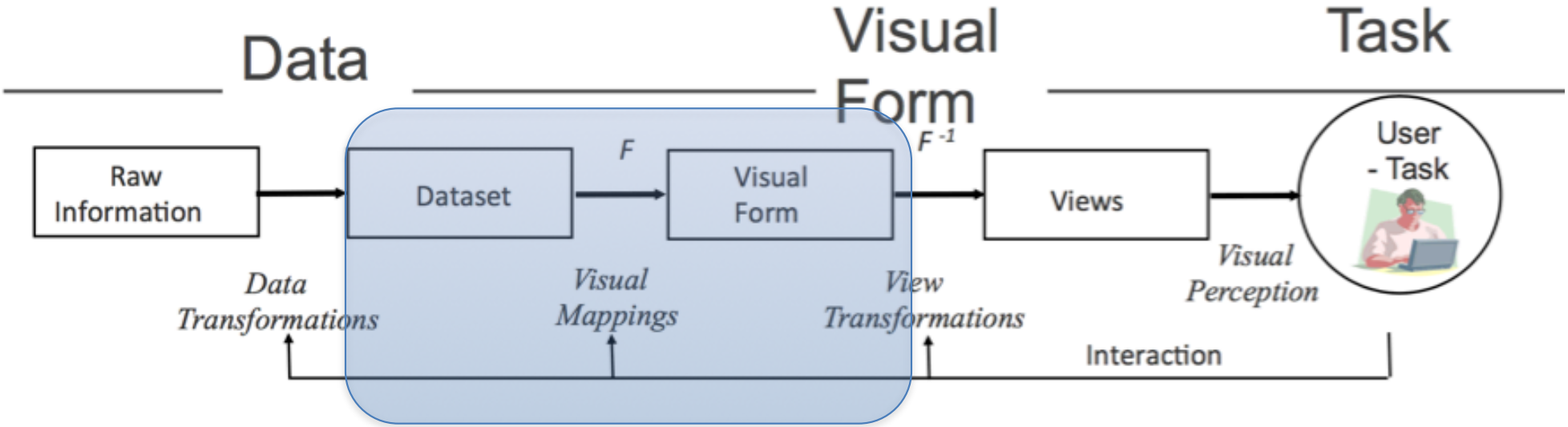


- Choose a visual model, such as a bar graph, list, or tree.

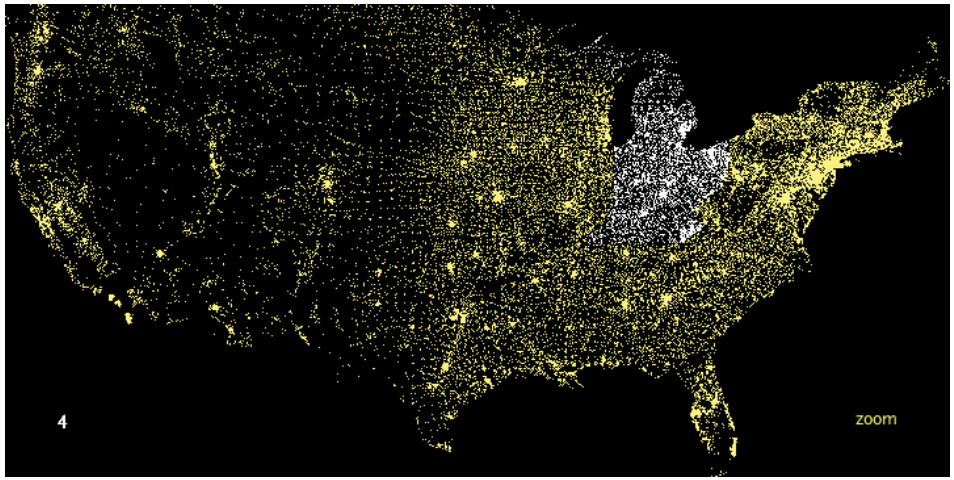


Basic visual representation of zip code data

Seven Stages: Refine



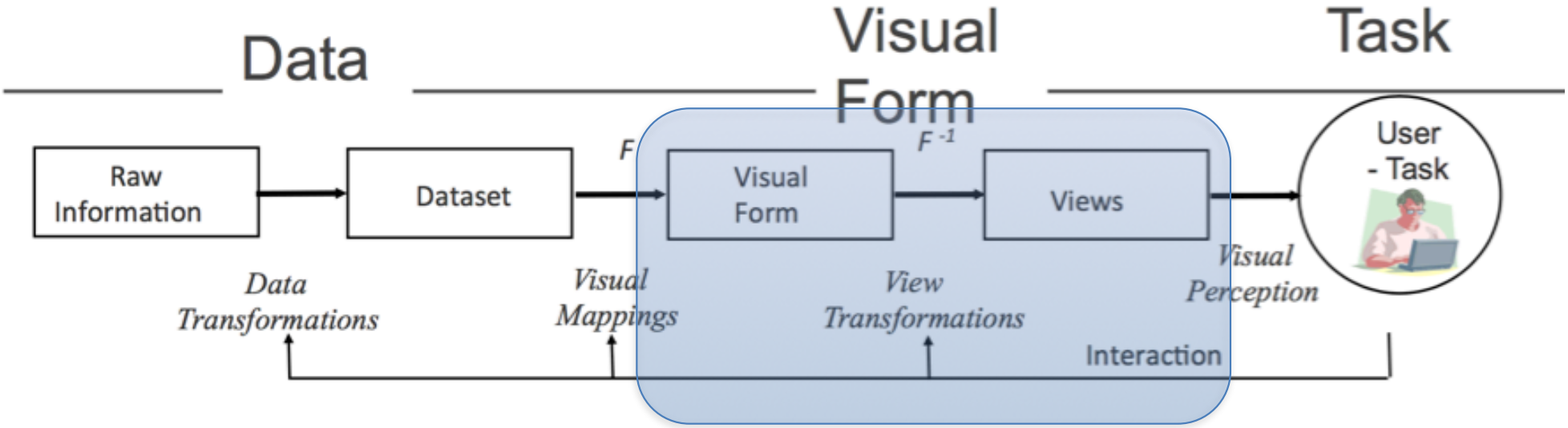
- Improve the basic representation to make it clearer and more visually engaging.



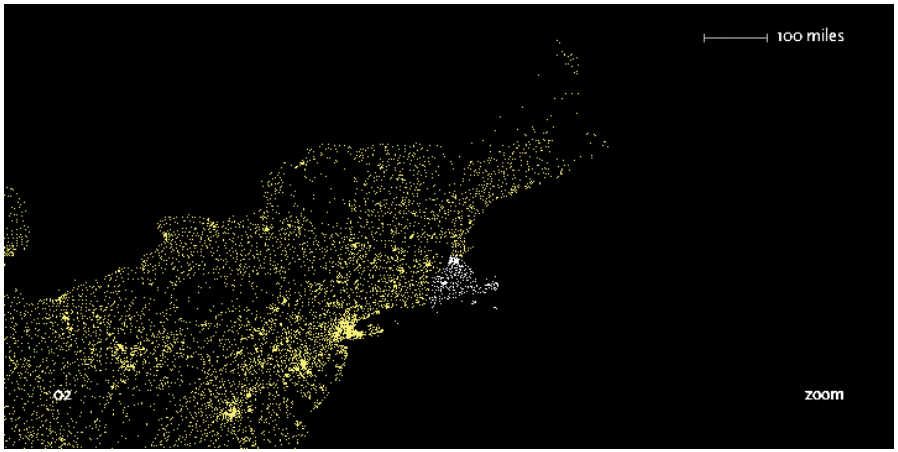
Using color to refine the representation

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

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)

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

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.

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

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.

Taxonomy of Data Types

- 1D (sets and sequences)
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Temporal
- Trees (hierarchies)
- Networks (graphs)
- 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: =, ≠

  e.g. math, art (course)

- O – Ordered
  - Operations: =, ≠, <, >

  e.g. A, A-, B+, B (grade)

- Q - Interval (location of zero arbitrary)
  - Operations: =, ≠, <, >, -
  - Can measure distances or spans

  e.g. (3.23, -1.2) (GPS)

- Q - Ratio (zero fixed)
  - Operations: =, ≠, <, >, -, %
  - Can measure ratios or proportions

  e.g. 20, 19, 22, 21 (age)

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

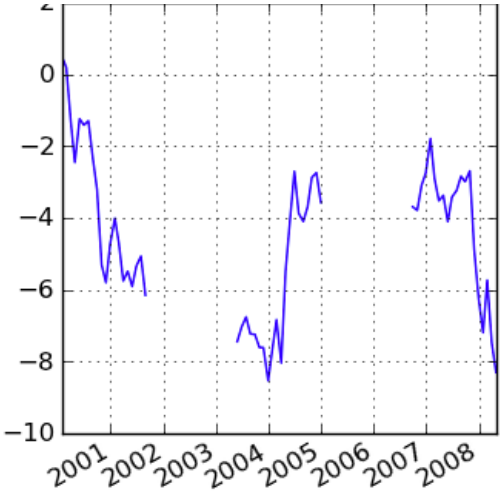
independent variables

	Math	Art	Sport
Avg Age	19.5	22	21

dependent variables

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.

Reading: https://en.wikipedia.org/wiki/Relational_algebra

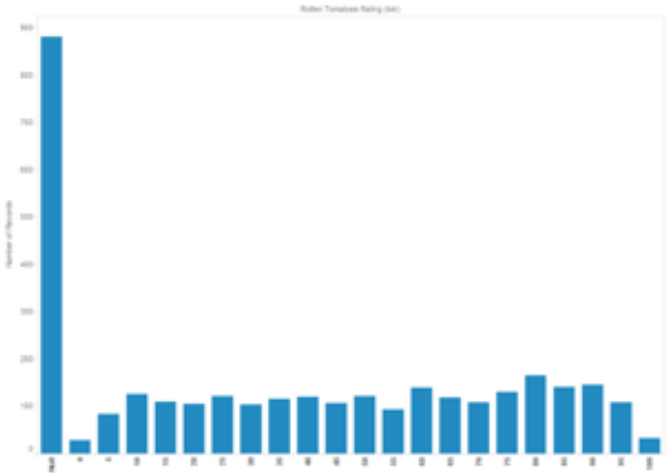
Data Cleaning and Filtering

- 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!

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

- 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: 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

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

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

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 express (1) all the facts in the set of data, and (2) only the facts in the data.

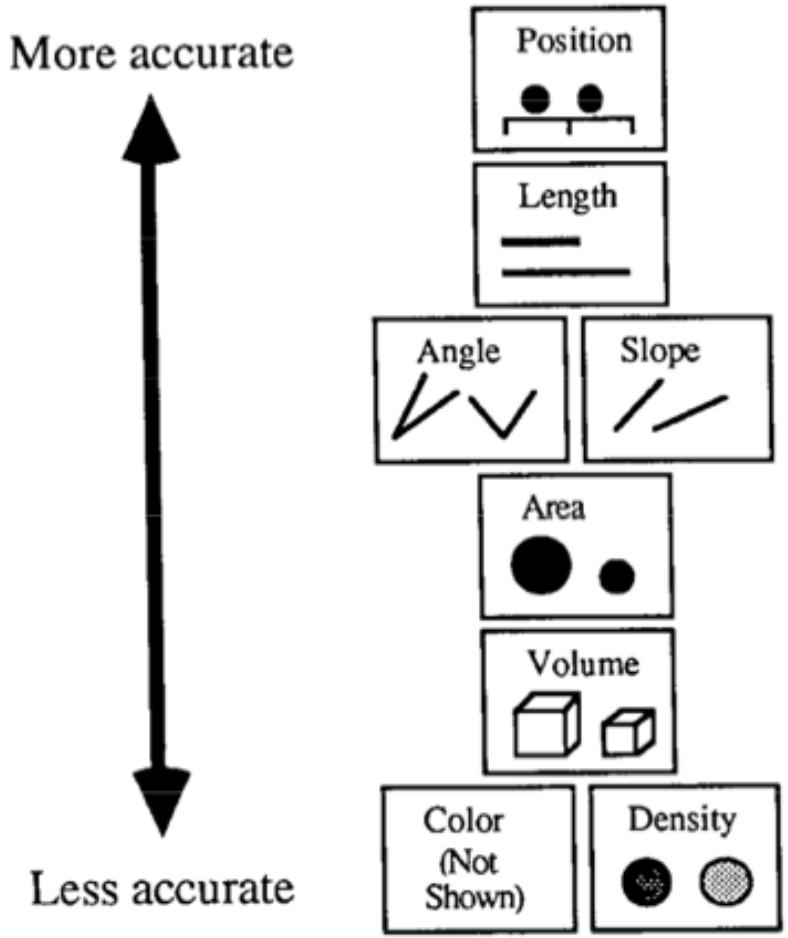
Tell the truth

- **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.

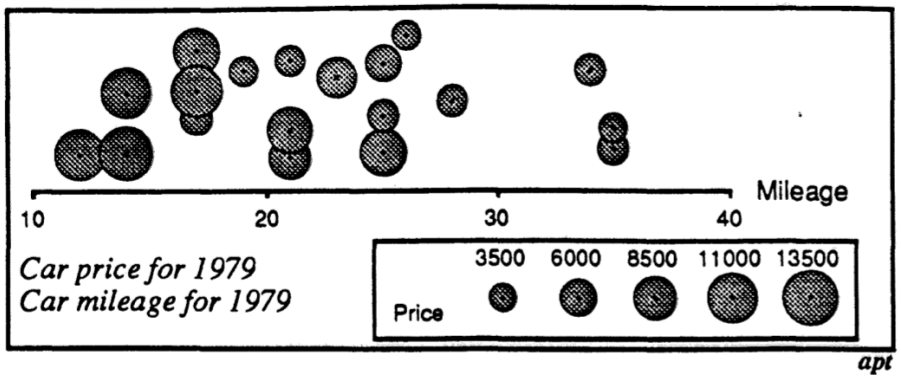
Effectiveness: Accuracy Ranking for Quantitative Information

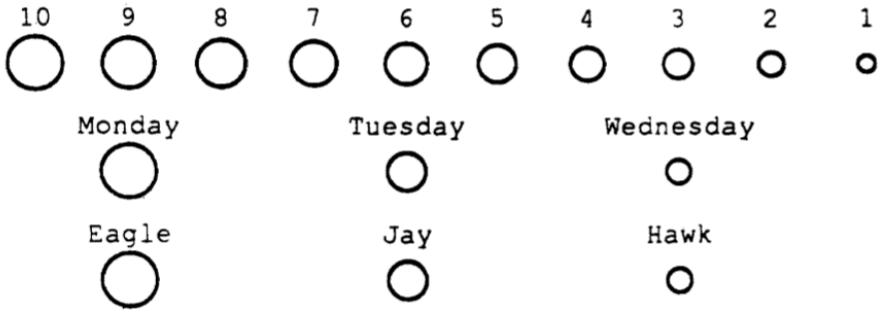


Mackinlay, Automating the design of graphical presentations of relational information, 1986.

Effectiveness: Accuracy Ranking for Nominal/ Ordinal Information?

Area Encoding





Quantitative

We can use, but not so accurate.

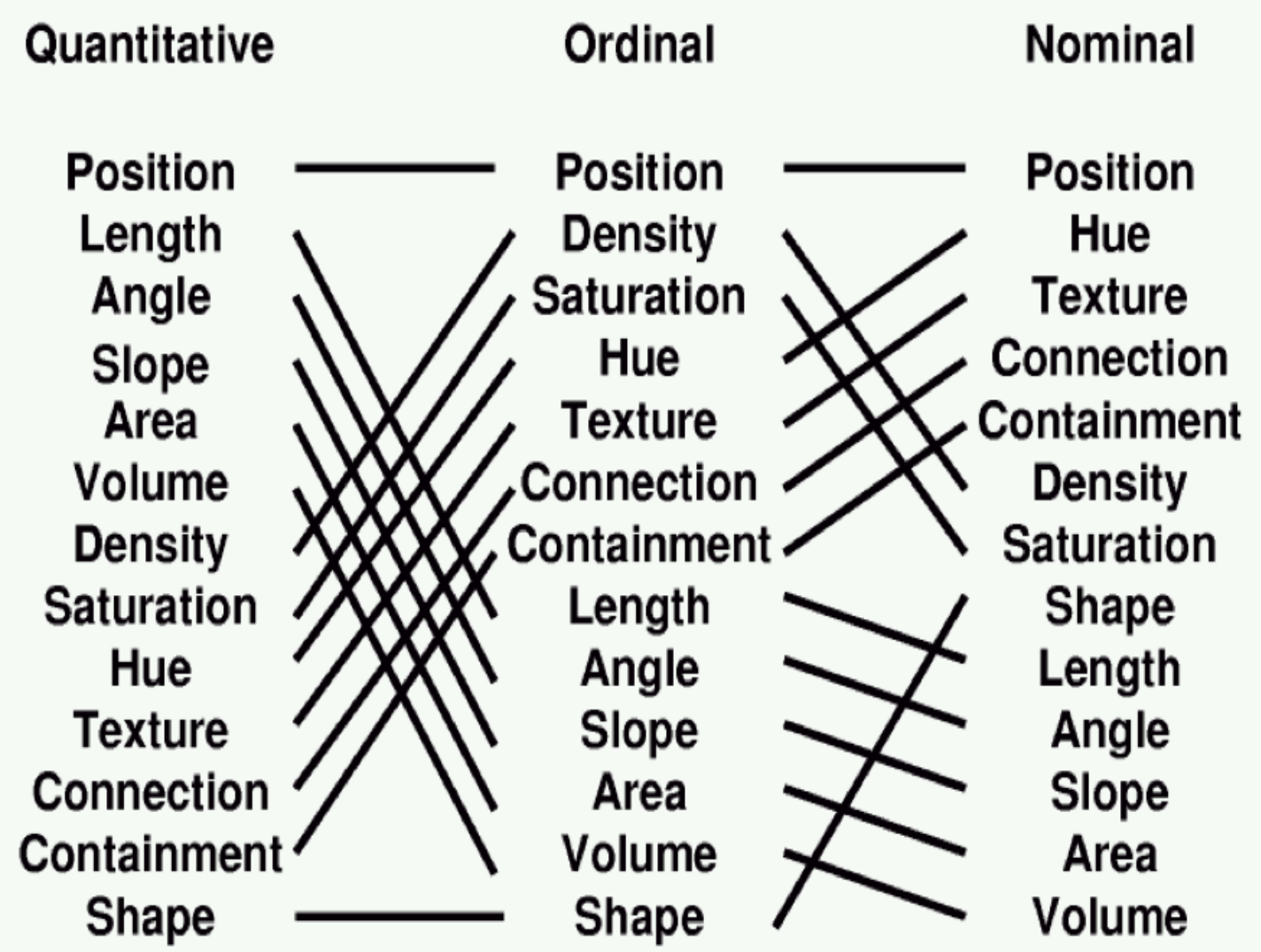
Nominal

- 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.

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

- Data Dimensions
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data

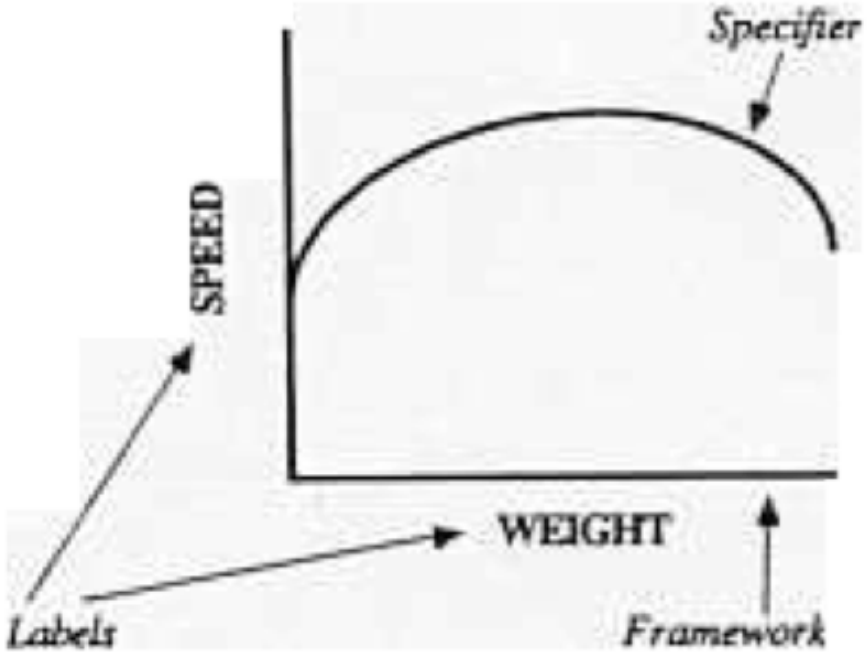
We mainly focus on uni, bi and tri-variate data for the rest of the lecture.

- Data Types
  - Nominal, Ordinal, Quantitative

- Visualization Representations
  - Points, Lines, Bars, Boxes

Components of Graphs

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



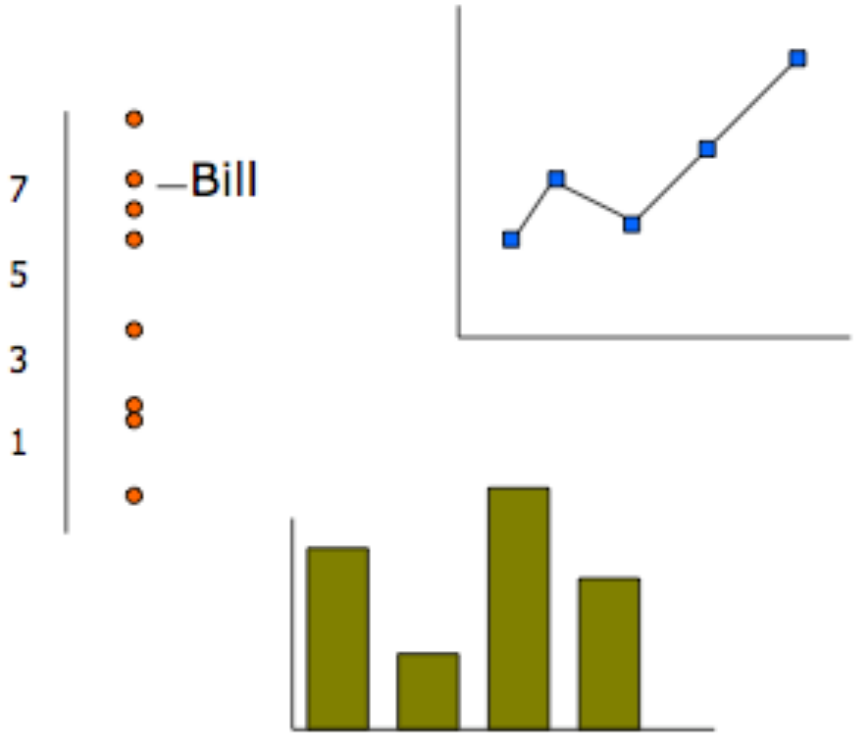
Points, Lines, Bars, Boxes

- 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

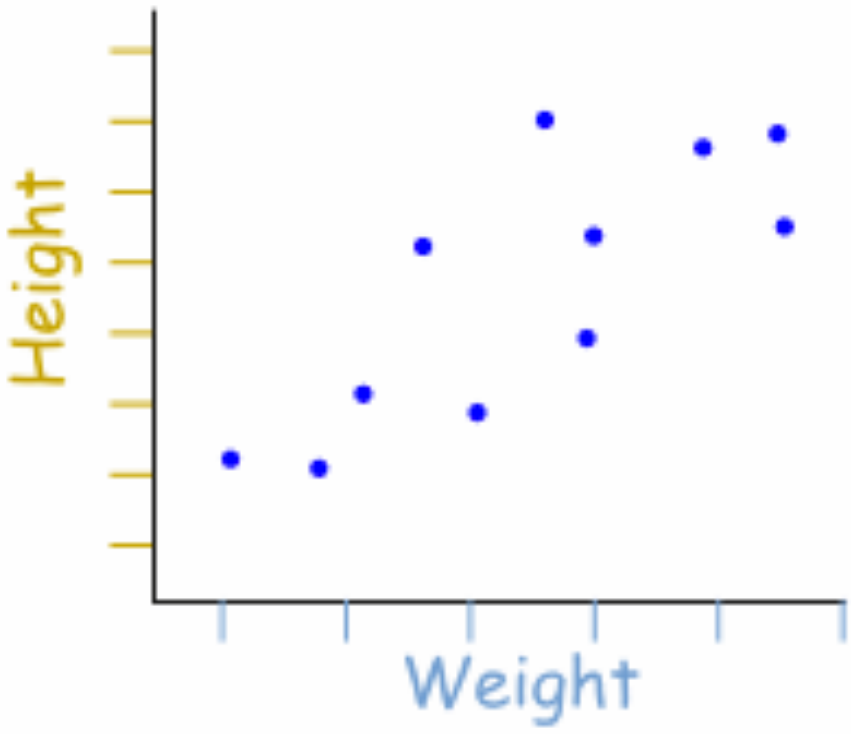
- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another.

- Statistical view
  - Independent variable on x-axis (data case)
  - Track dependent variable along y-axis (value)



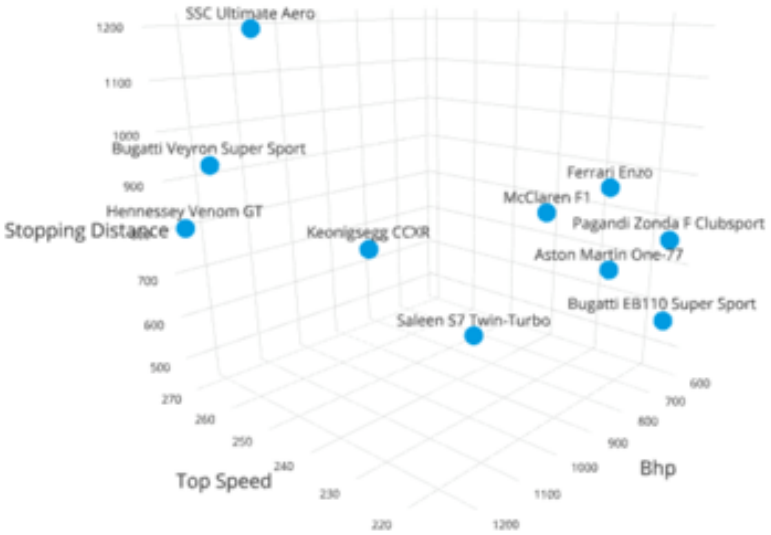
Bivariate Data

- 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?

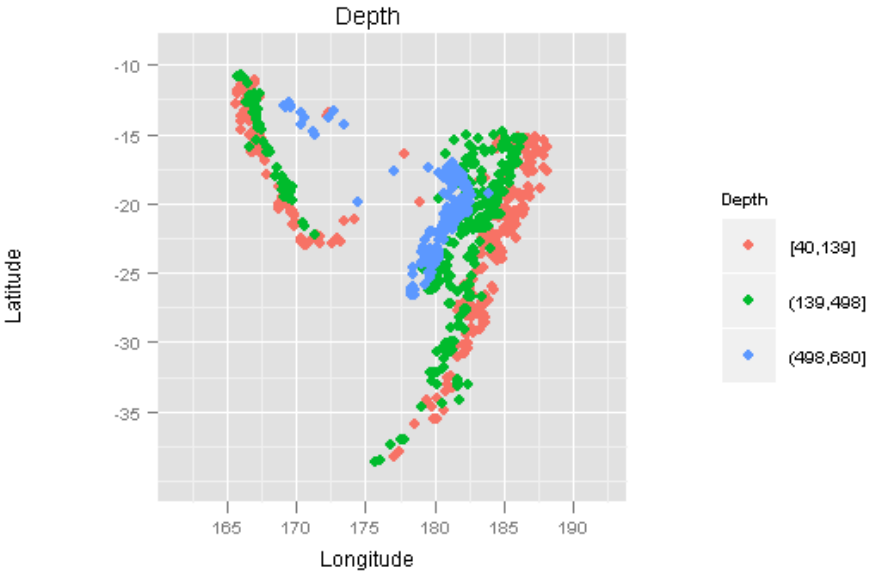


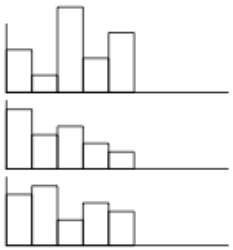
Trivariate Data

3D scatter plot



2D + mark property

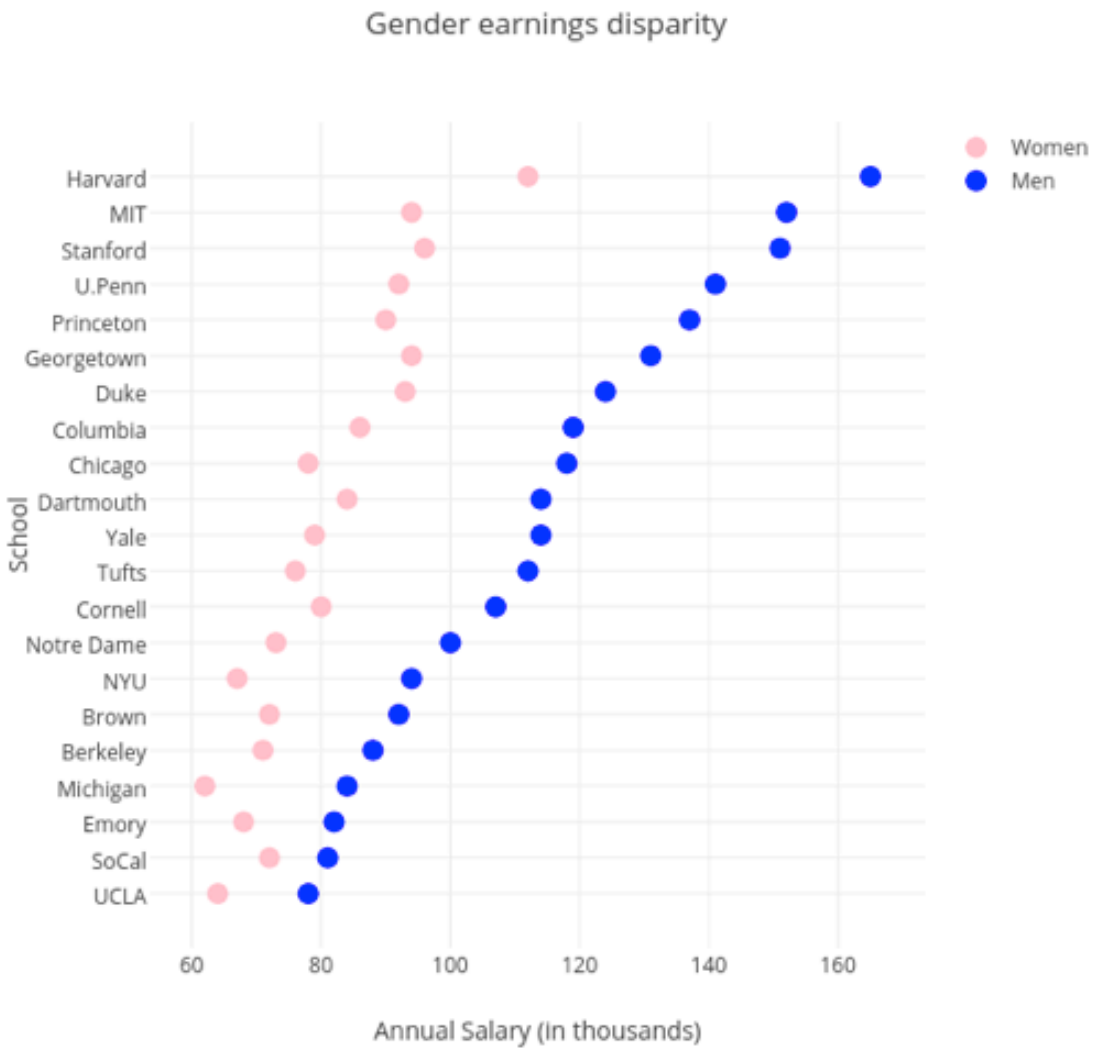




Represent each variable in its own explicit way

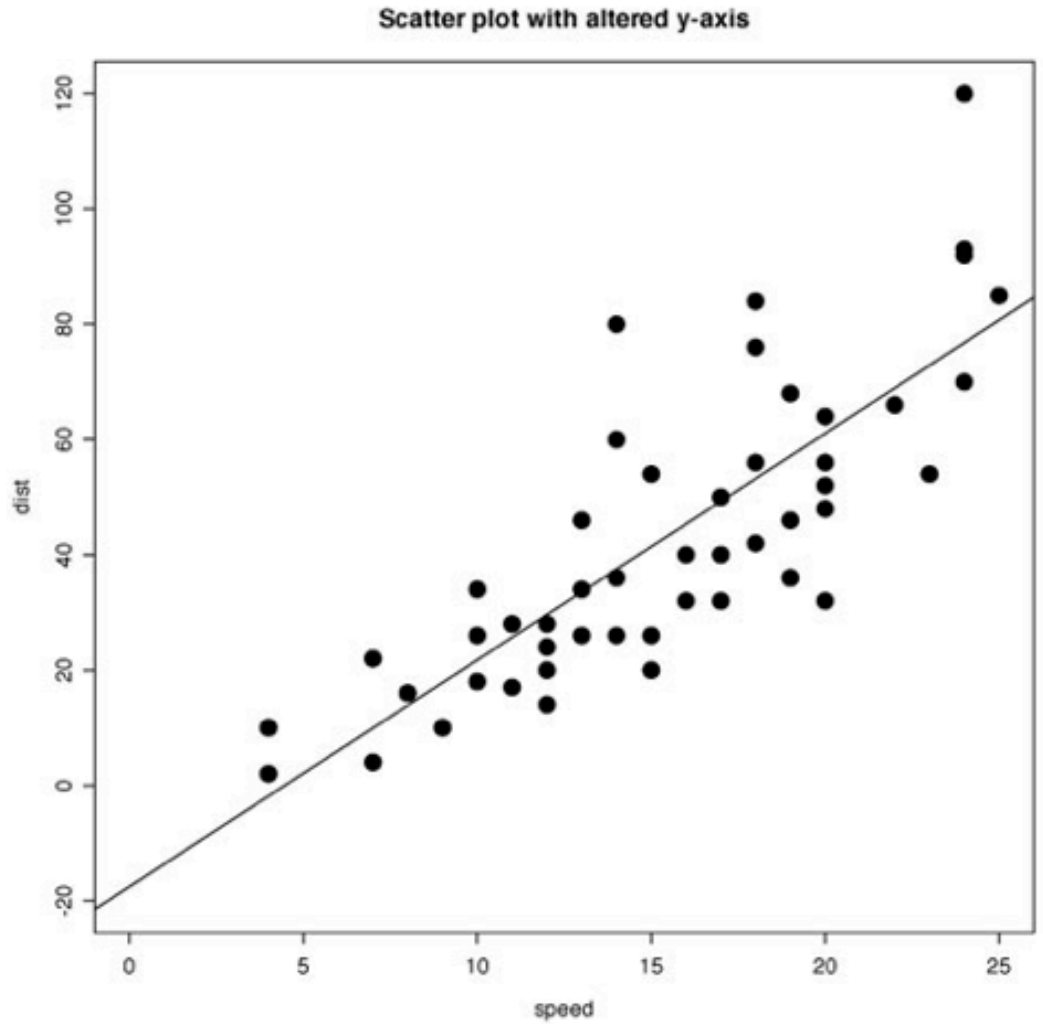
Dot Plots

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



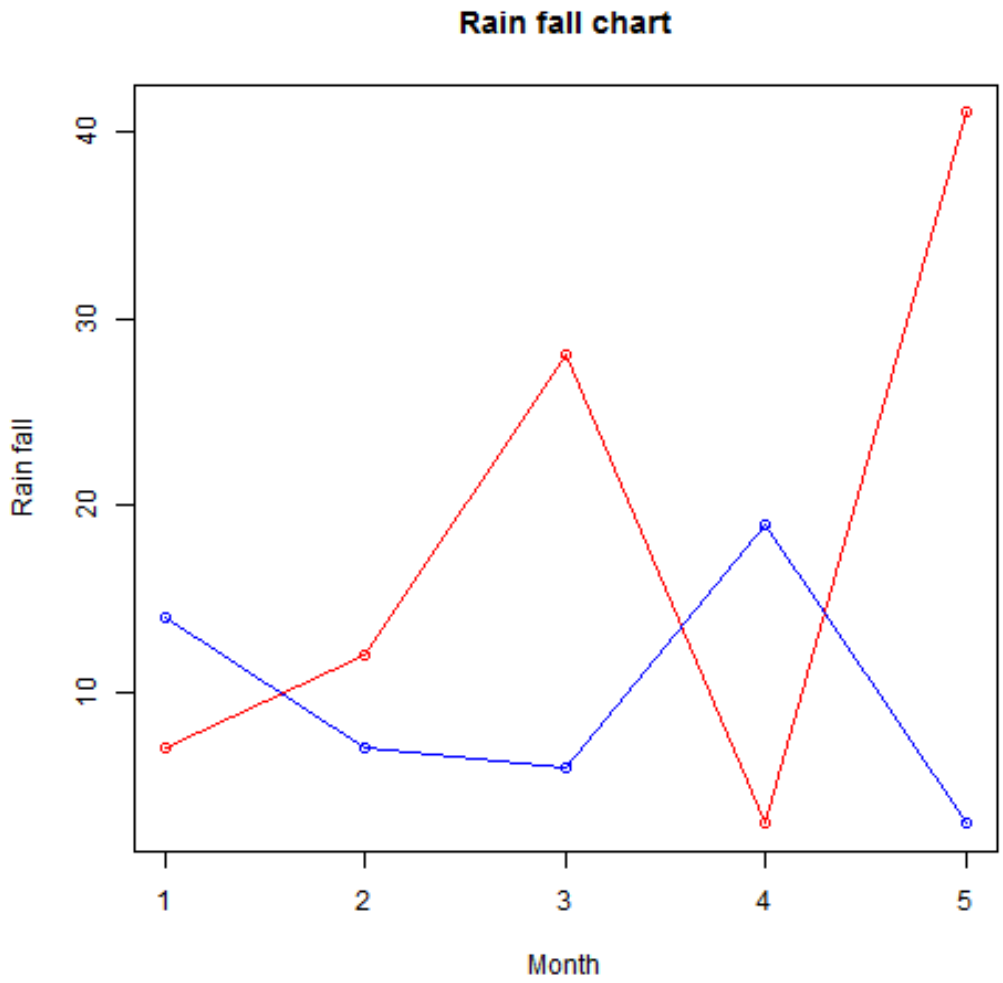
Scatter Plot

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



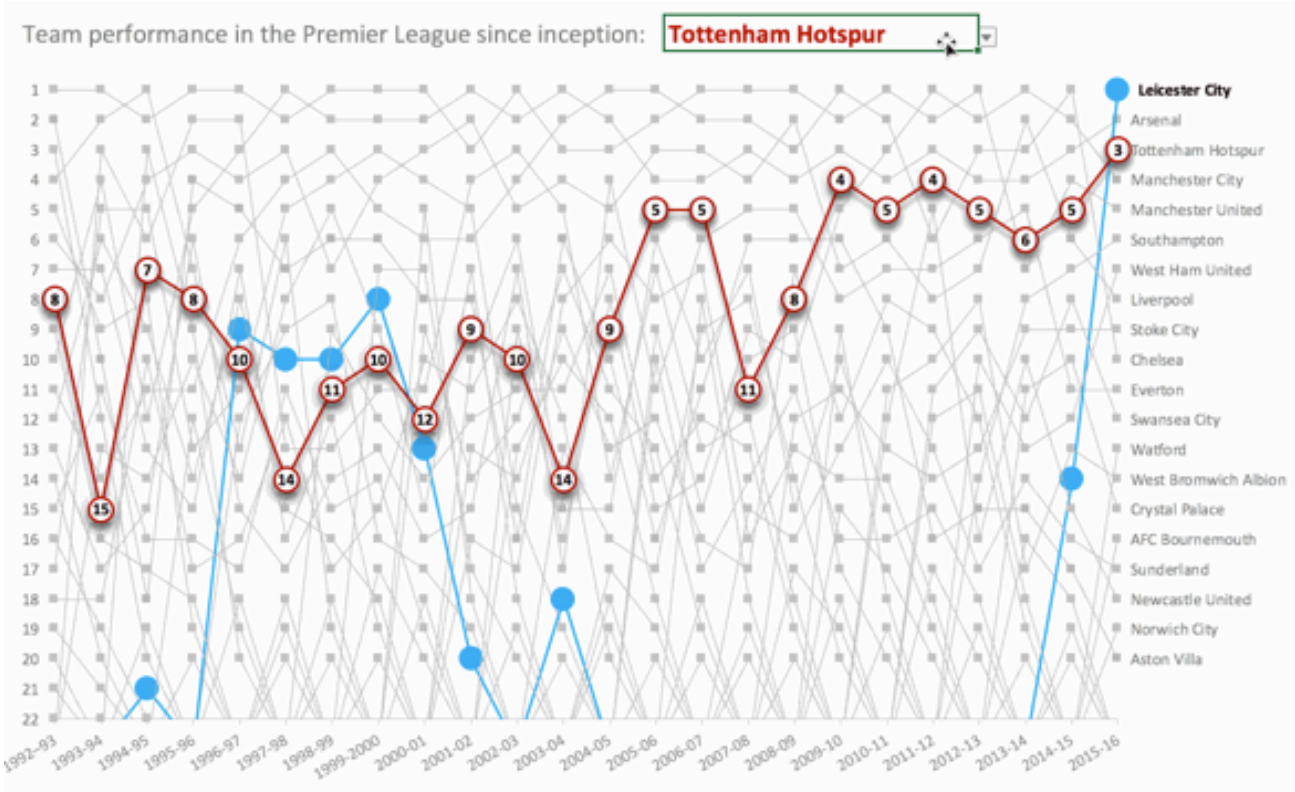
Line Graphs

- 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



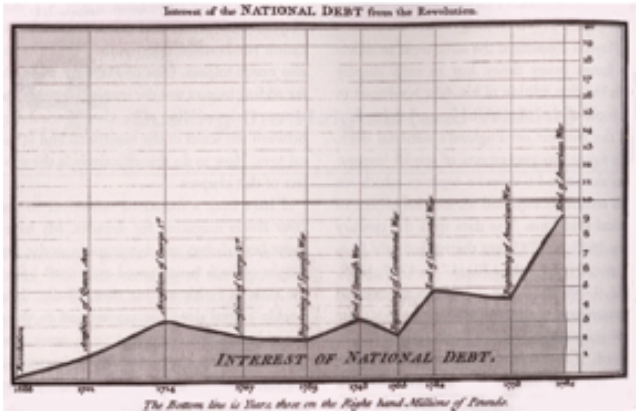
Bump Chart

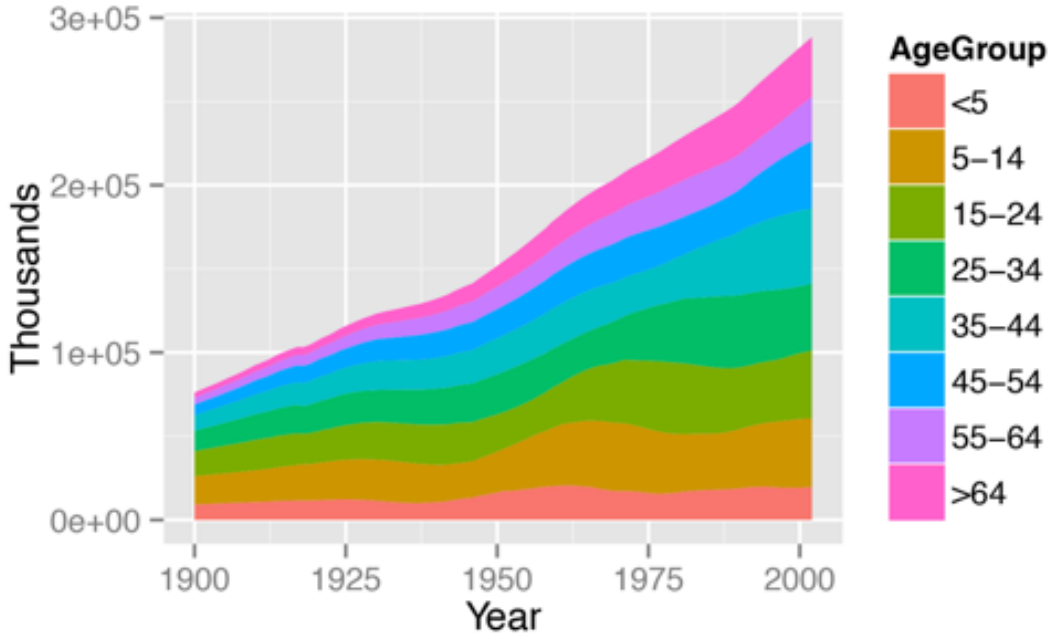
- When to use:
  - Similar to line graph
  - Y-axis: rank rather than (continuous) values



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.

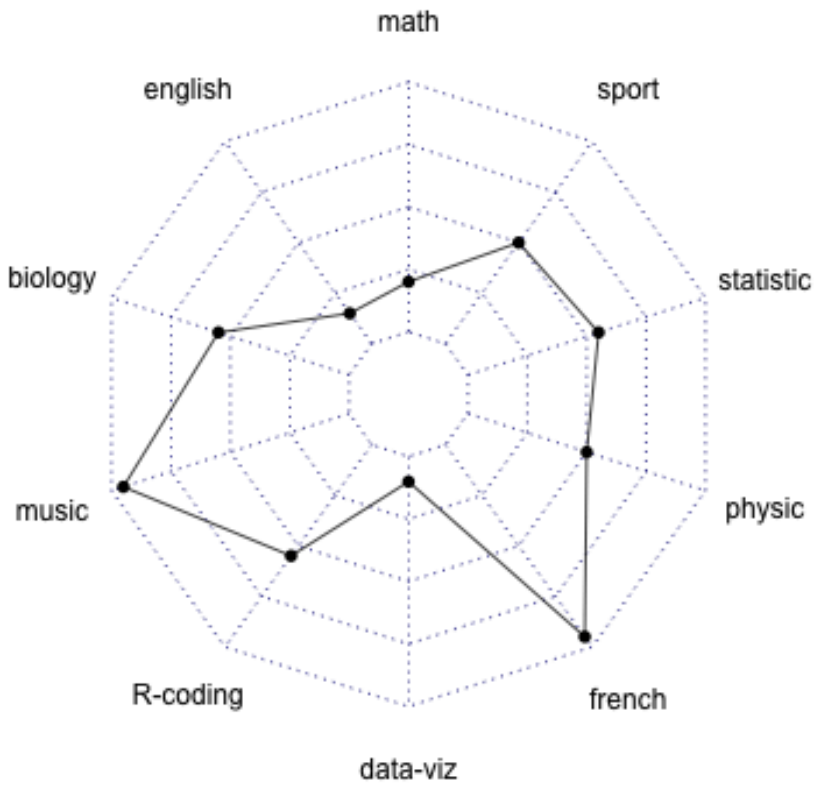




Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

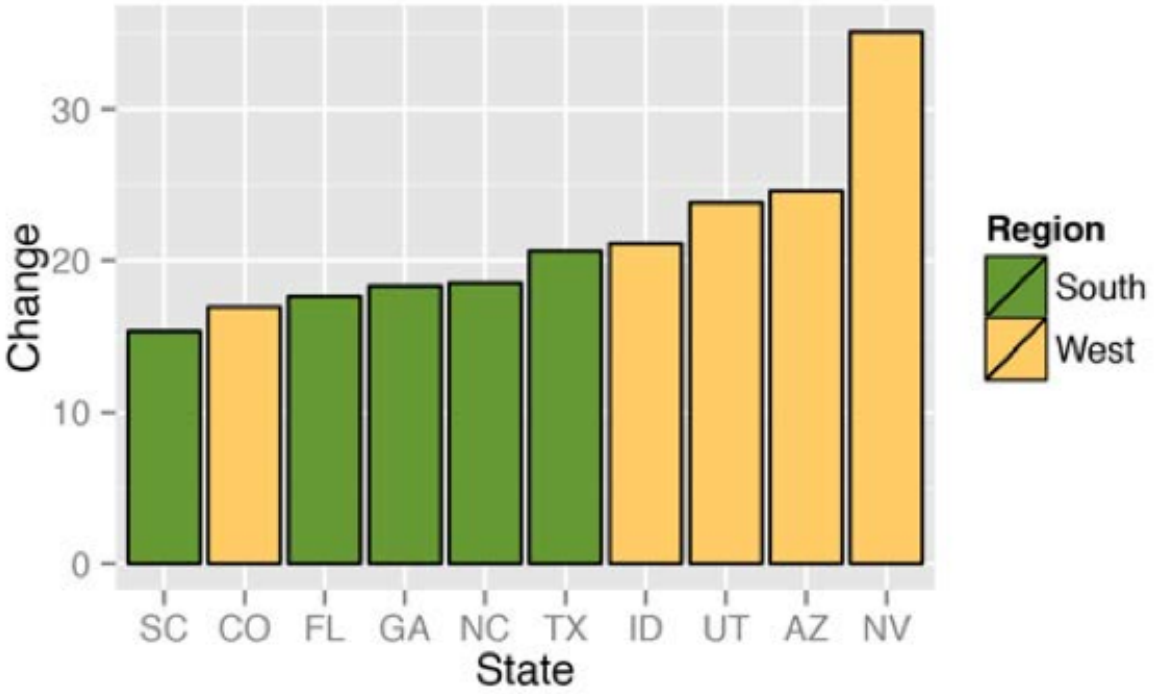
Radar Graphs

- 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



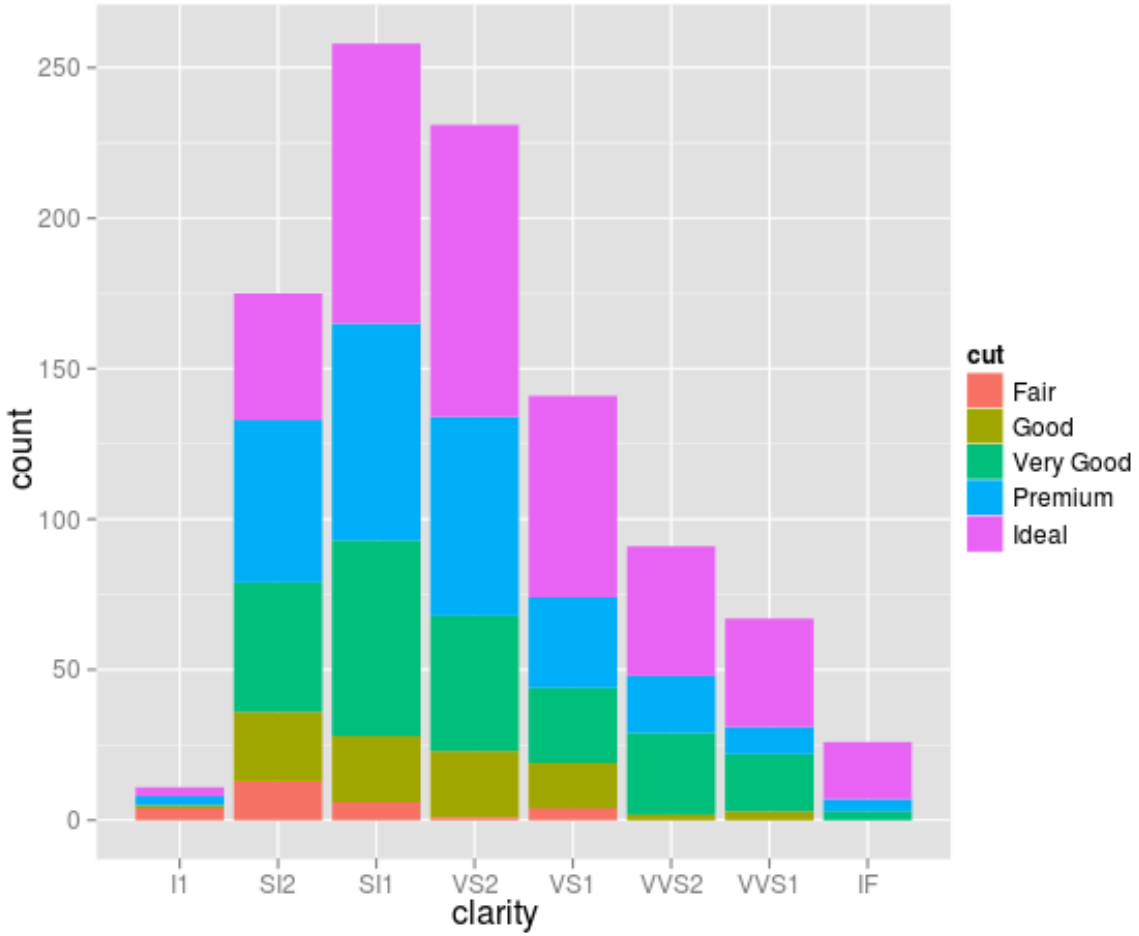
Bar Graphs

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



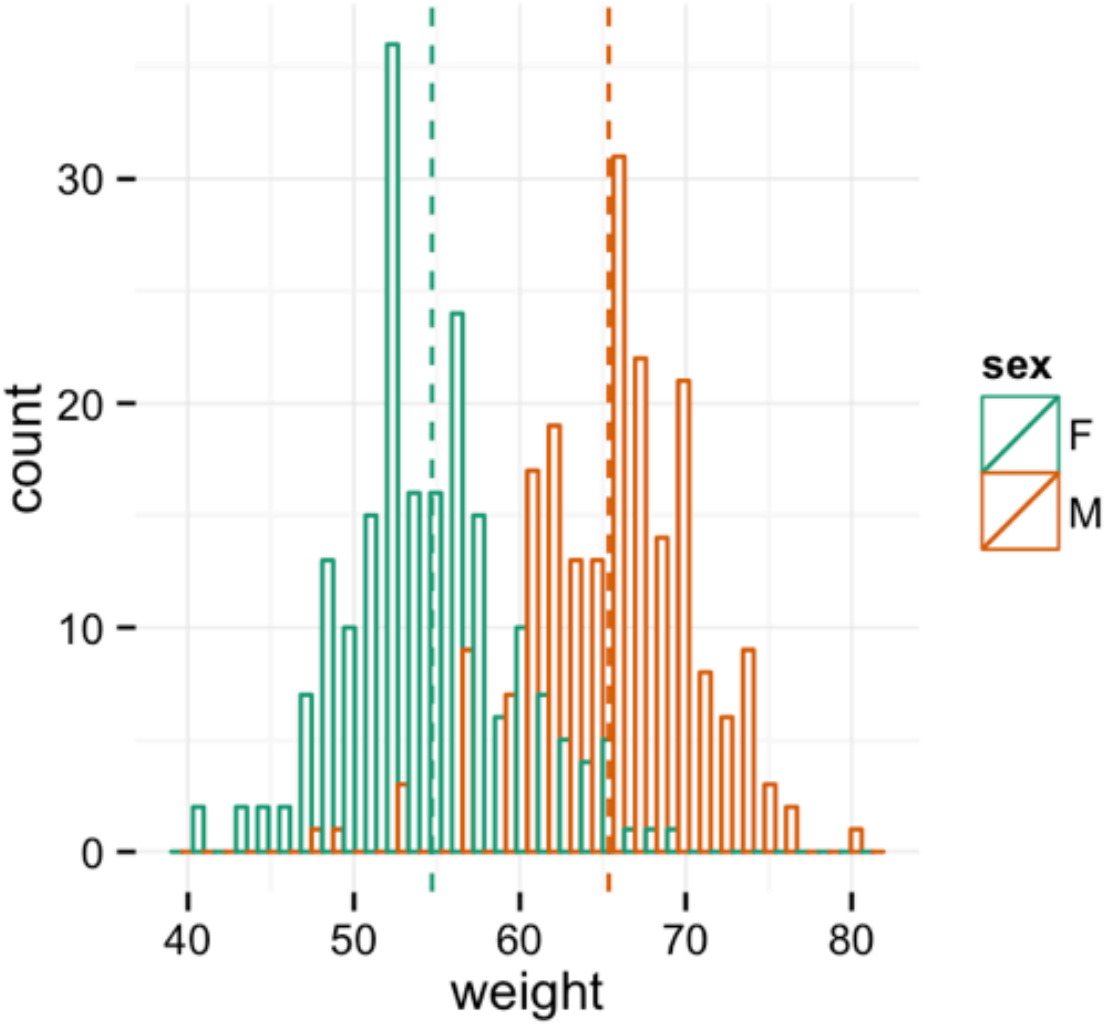
Stacked Bar Chart

- 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



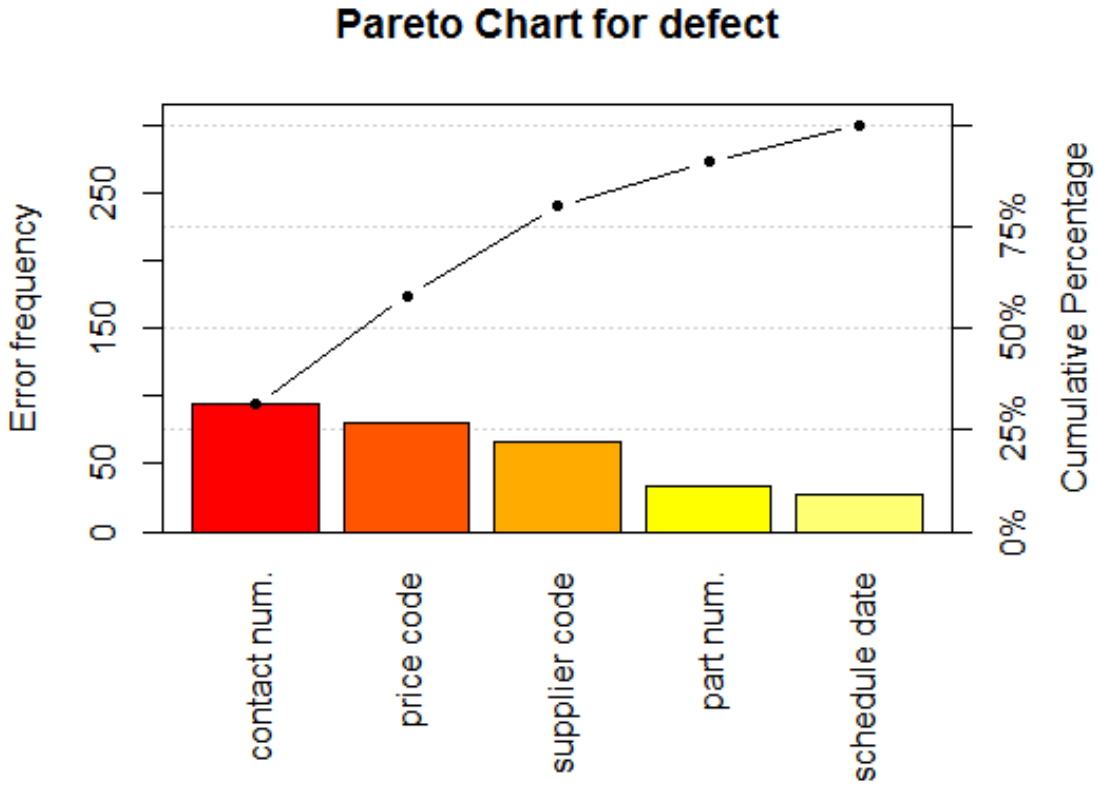
Histogram

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



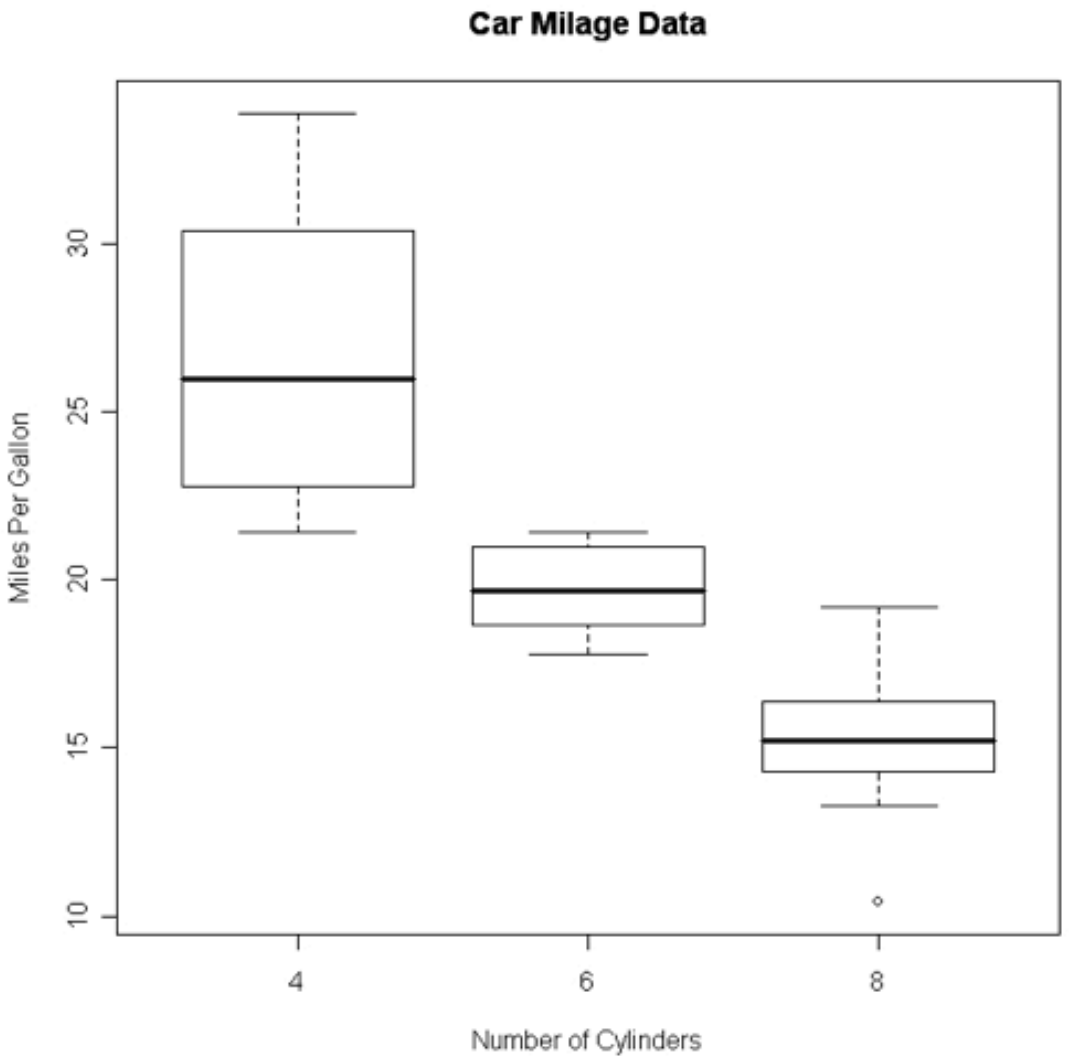
Pareto chart

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



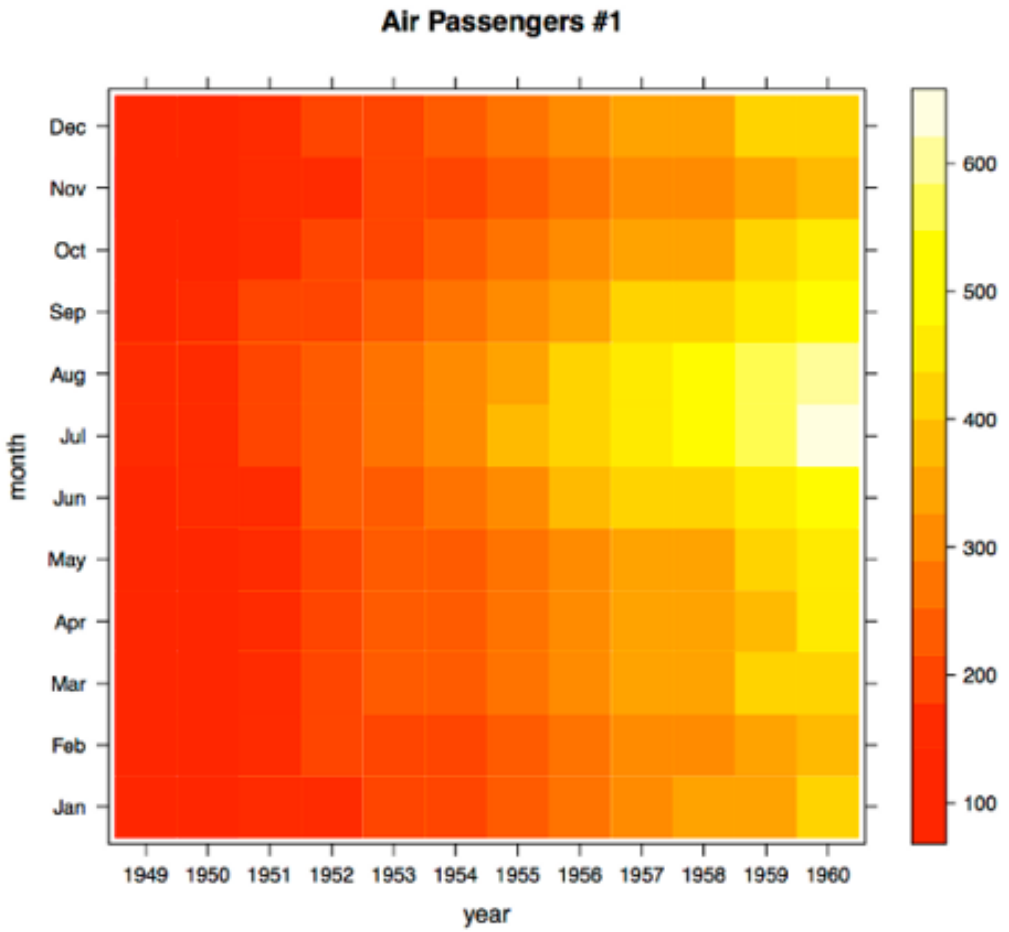
Box Plots

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



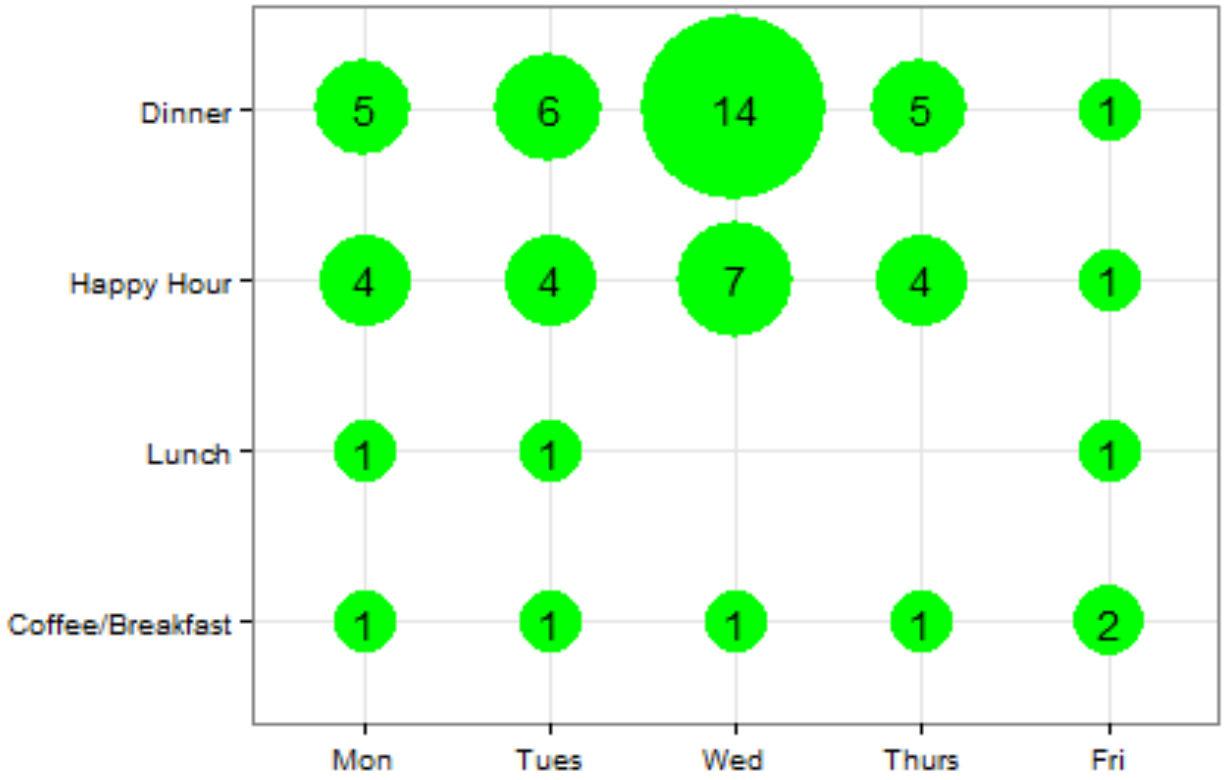
Heat Maps

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



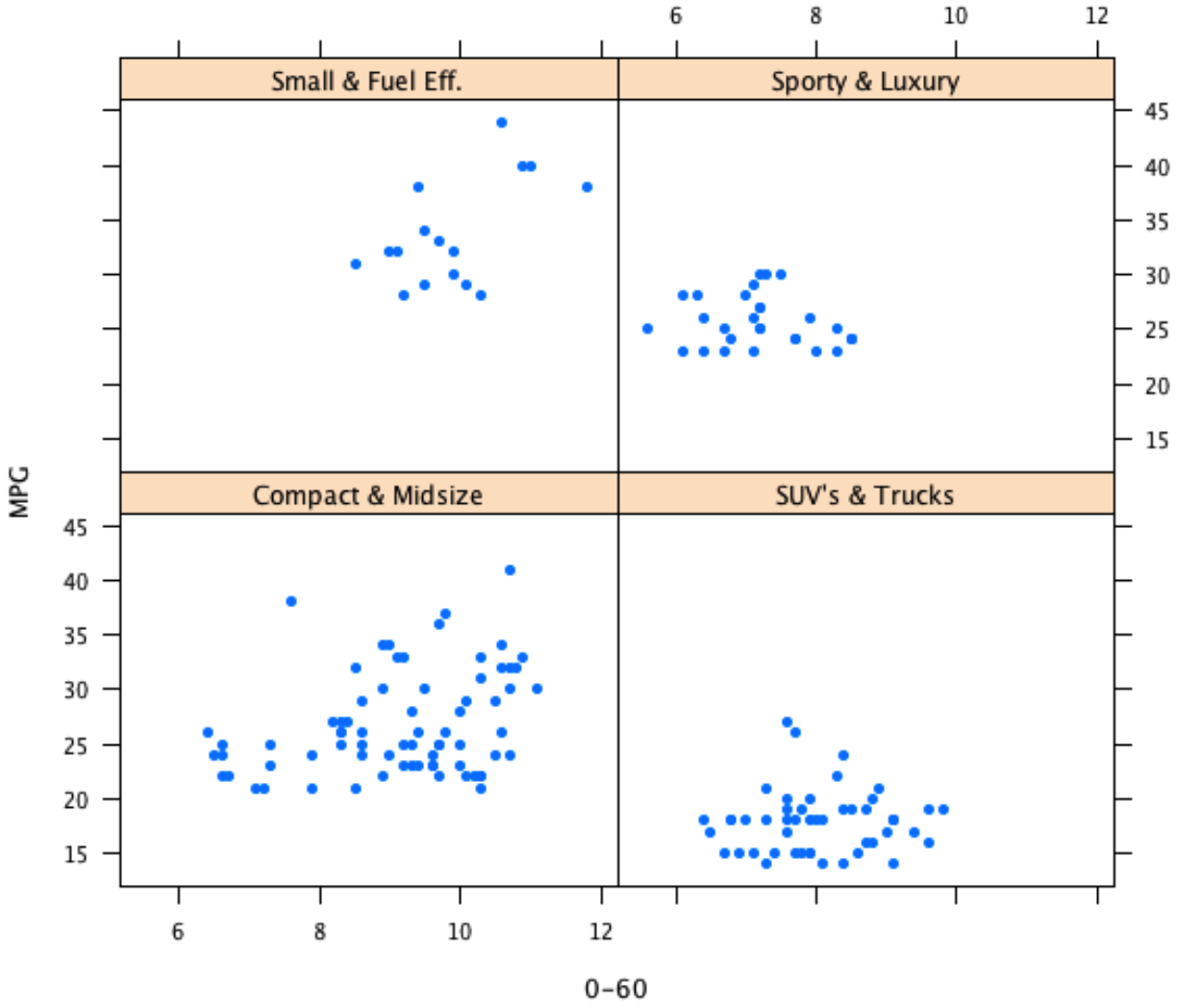
Crosstab Plot

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



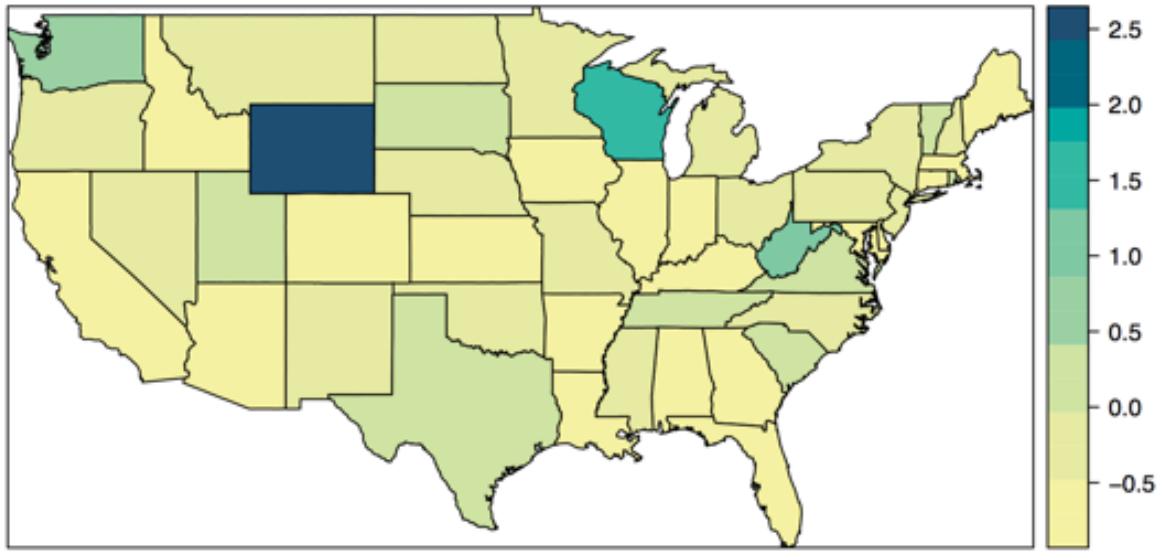
Trellis Display

- 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



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

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 240D	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

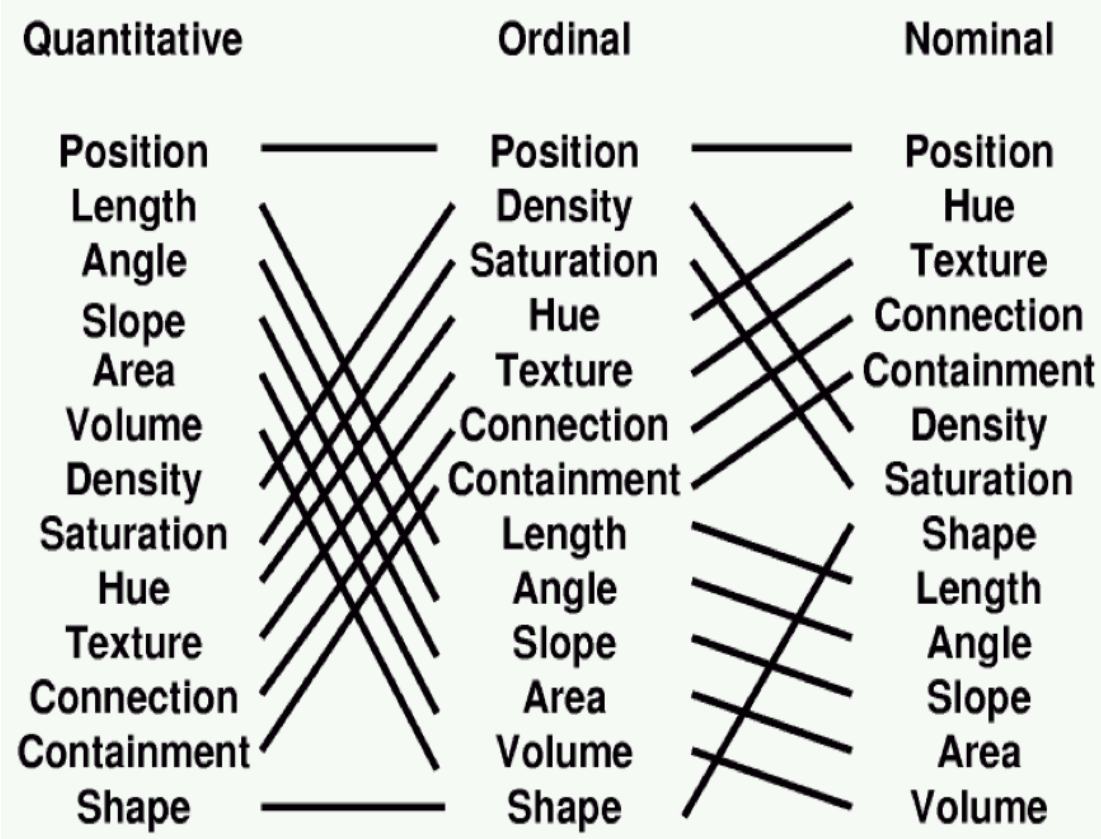
When to use?

- 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.

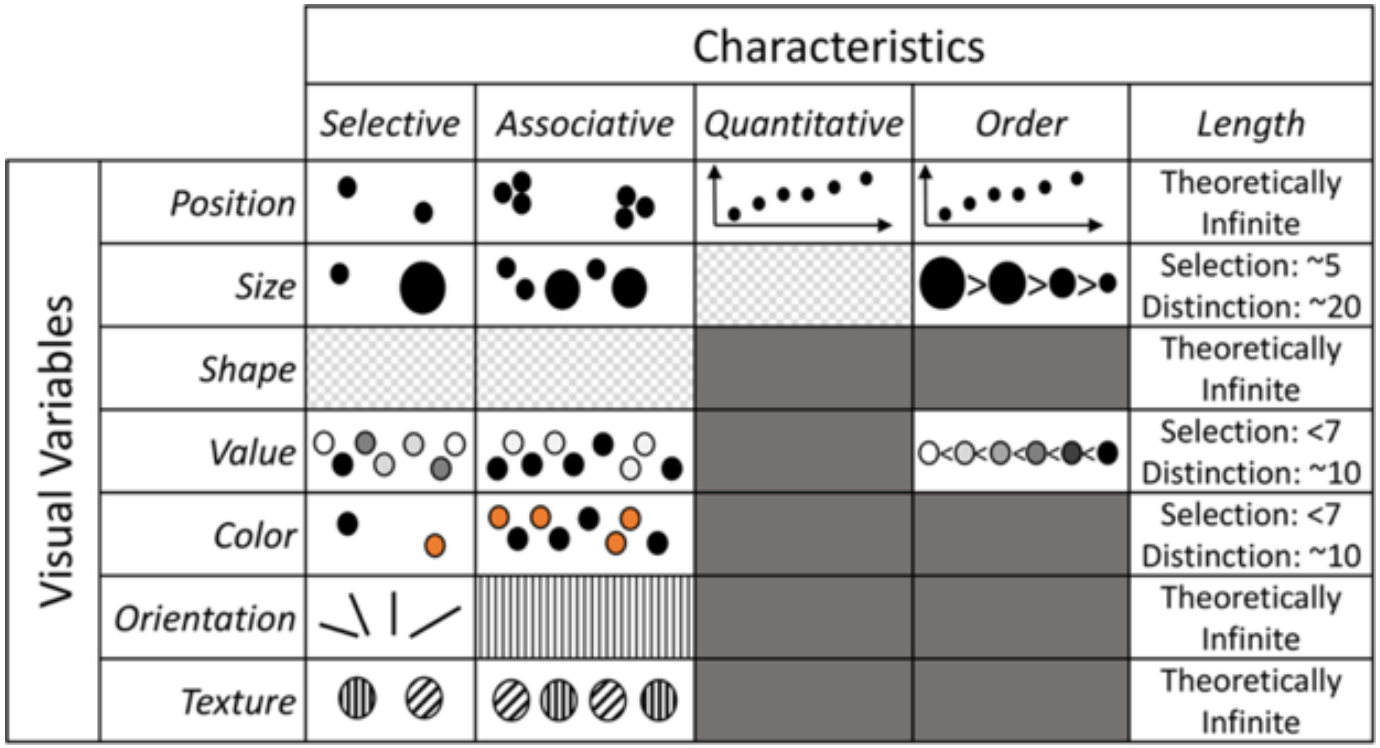
Design Challenge

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



Multivariate Data Visualization

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



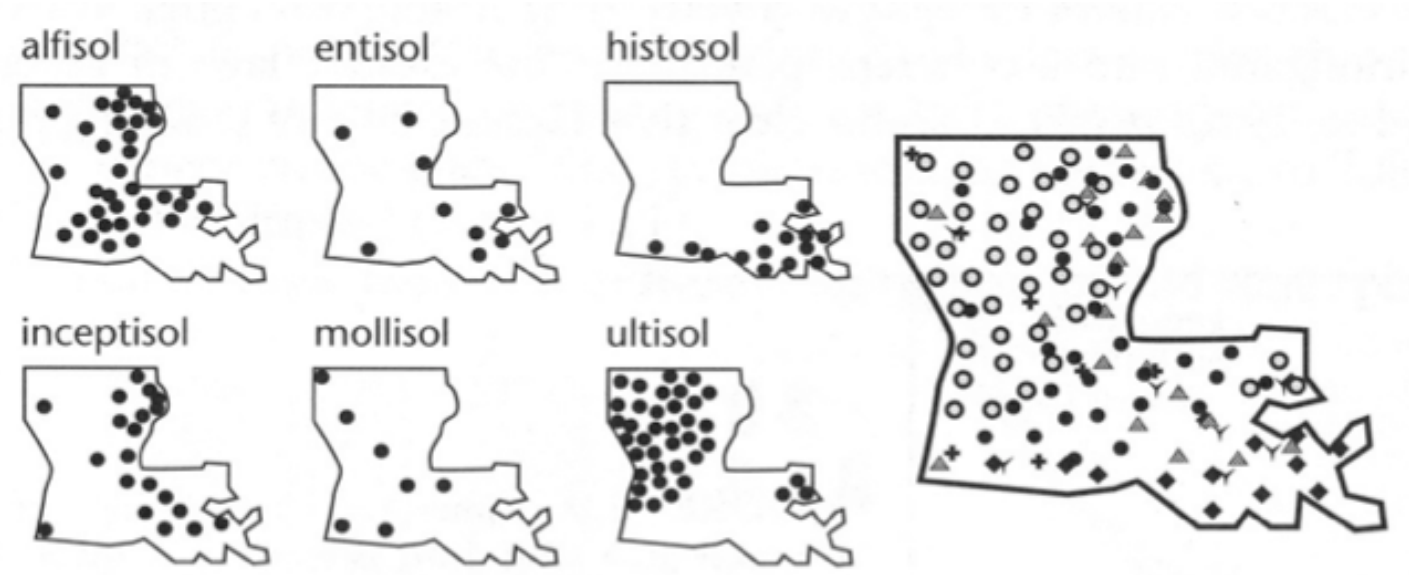
Visual Variables	Characteristics				
	Selective	*Associative*	*Quantitative*	*Order*	*Length*
Position					Theoretically Infinite
Size					Selection: ~5 Distinction: ~20
Shape					Theoretically Infinite
Value					Selection: <7 Distinction: ~10
Color					Selection: <7 Distinction: ~10
Orientation					Theoretically Infinite
Texture					Theoretically Infinite

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

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. "

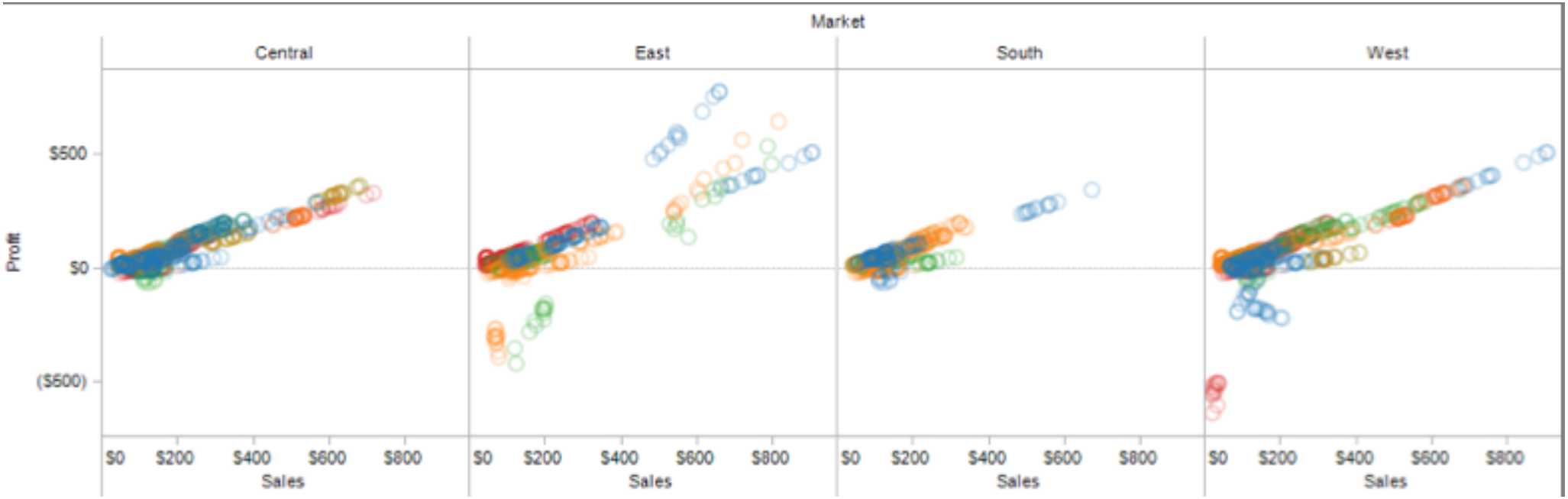
Tufte, Envisioning Information



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

Trellis Display (Small Multiples)

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



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.

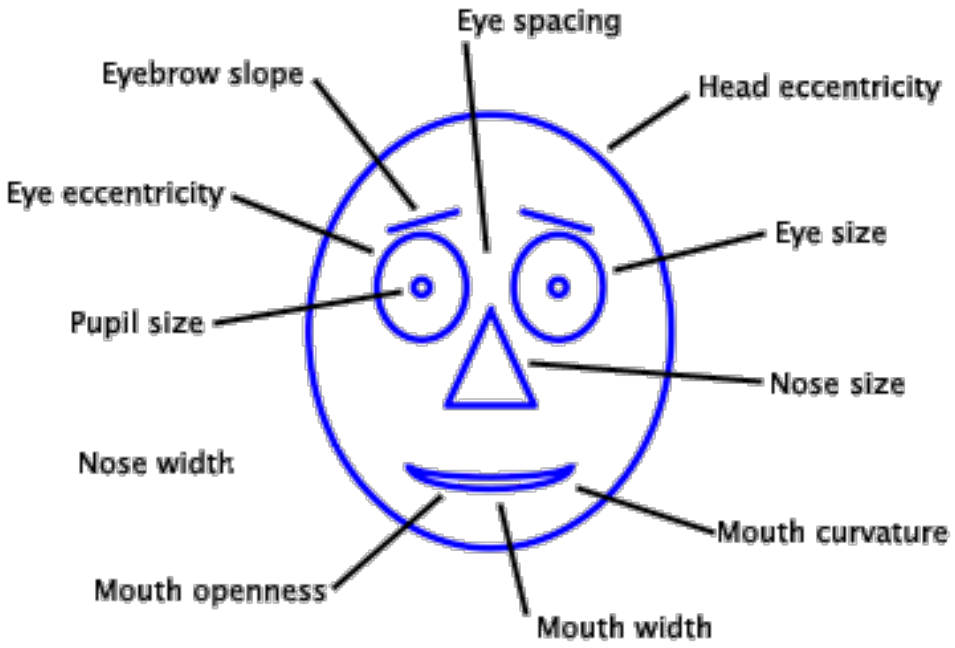
Common Multivariate Data Visualization Techniques

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

Chernoff Faces

- 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.

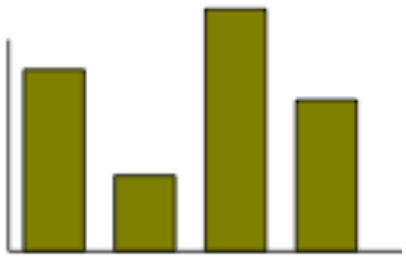
Visual Mapping

- Basic idea: Change quantitative values to bars

- What do you do for nominal data?

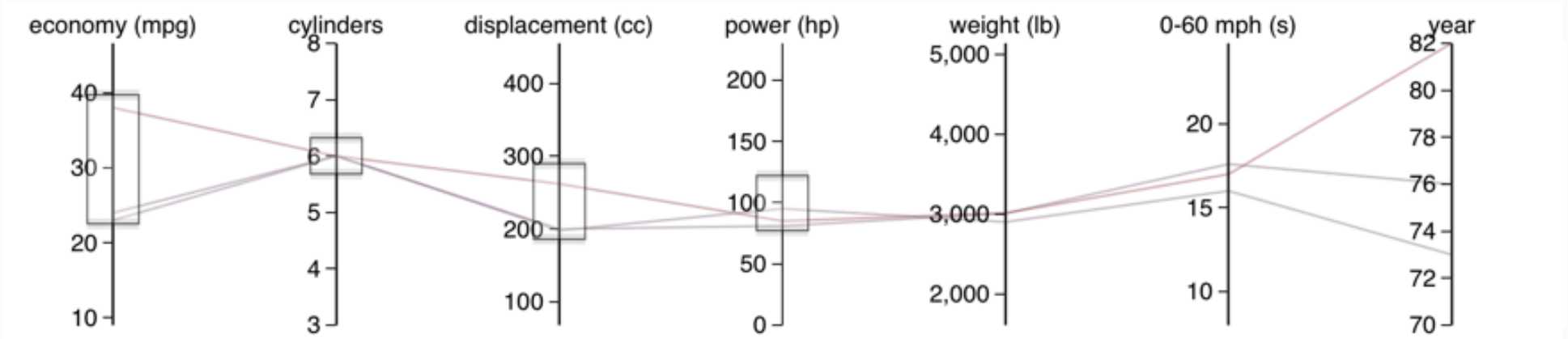
a1-cereals [Read-Only] [Compatibility Mode]

	A	B	C	D	E	F	G	H	I
1	Cereal	Manufactu	Type	Calories	Protein	Fat	Sodium	Fiber	Carbol
2	Frosted Mini-Wheats	K	C	100	3	0	0	3	
3	Raisin Squares	K	C	90	2	0	0	2	
4	Shredded Wheat	N	C	80	2	0	0	3	
5	Shredded Wheat 'n'Bran	N	C	90	3	0	0	4	
6	Shredded Wheat spoon s	N	C	90	3	0	0	3	
7	Puffed Rice	Q	C	50	1	0	0	0	
8	Puffed Wheat	Q	C	50	2	0	0	1	
9	Maypo	A	H	100	4	1	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	150	4	3	95	3	



Parallel Coordinates

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



name	economy (mpg)	cylinders	displacement (cc)	power (hp)	weight (lb)	0-60 mph (s)	year
Ford Maverick	24	6	200	81	3012	17.6	76
Oldsmobile Cutlass Ciera (Diesel)	38	6	262	85	3015	17	82
Plymouth Duster	23	6	198	95	2904	16	73

A. Inselberg, Multidimensional Detective, Proc. IEEE InfoVis, 1997

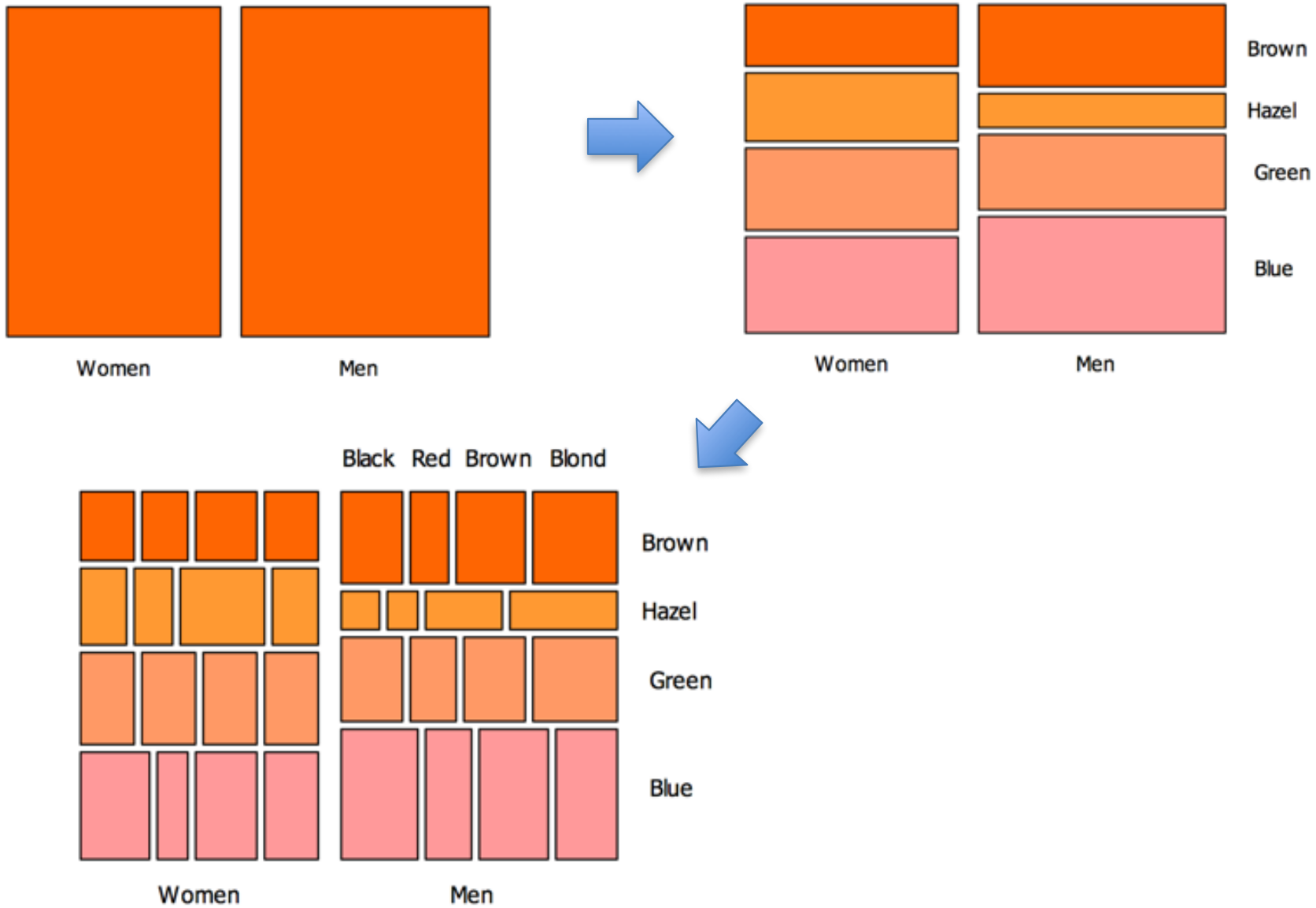
Multivariate Categorical Data

- 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, ...

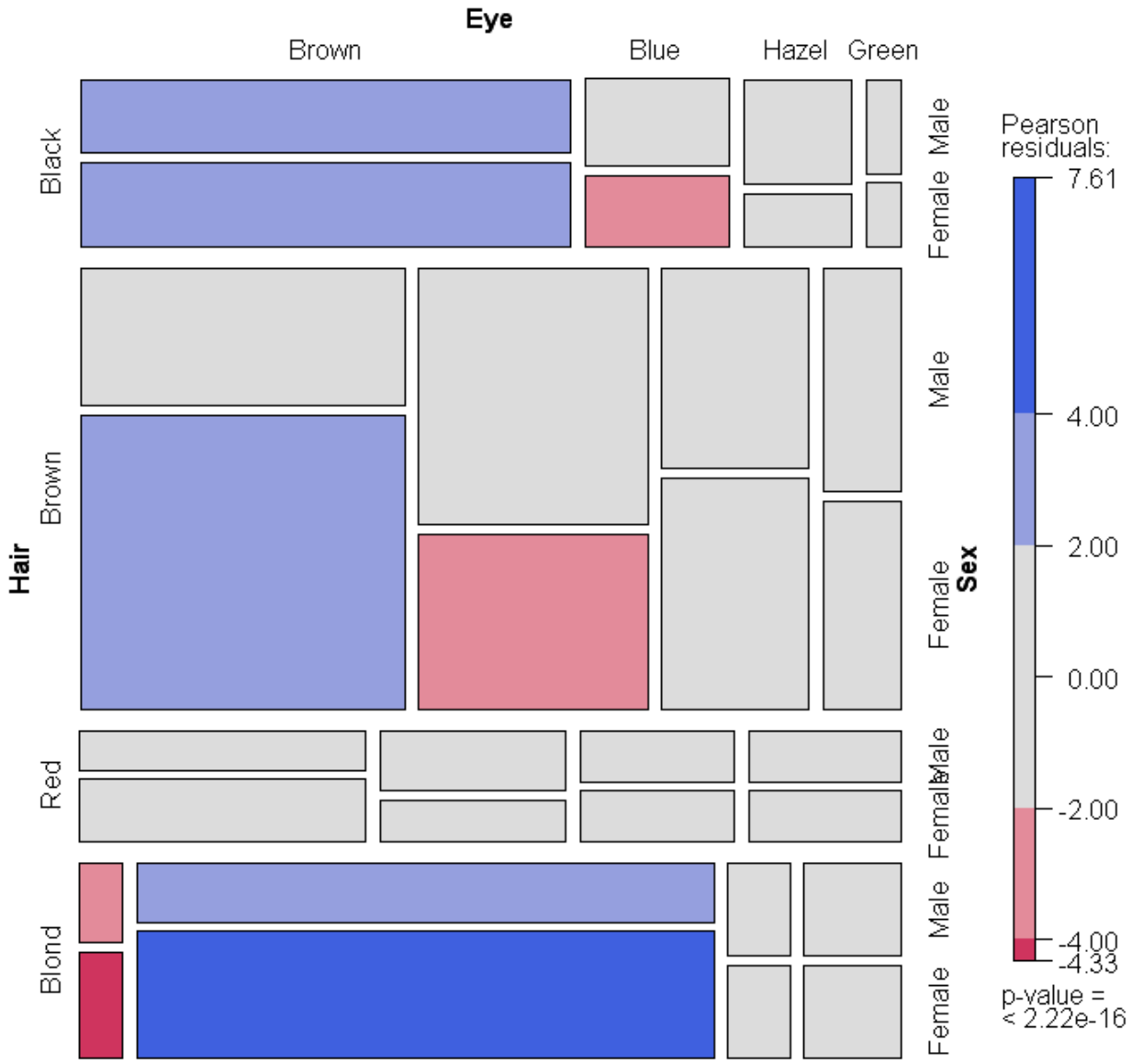
Friendly, Michael. "Mosaic displays for multi-way contingency tables."
Journal of the American Statistical Association 89.425 (1994): 190-200.

Mosaic Plot Decomposition



Mosaic Plot

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

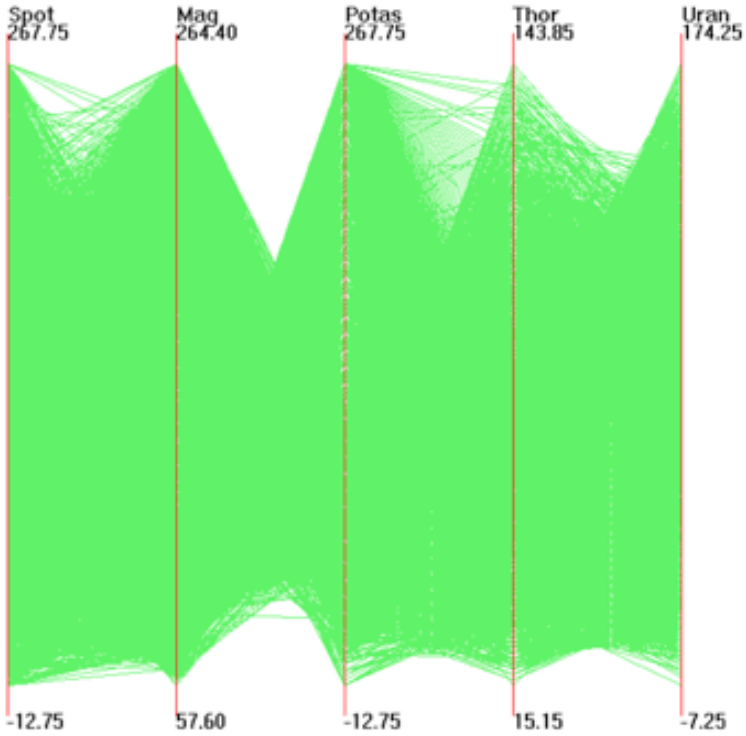


Data Overload

- 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?

We will address this in other lectures.



Out5d dataset(5 dimensions, 16384 items)

The University of Nottingham
UNITED KINGDOM • CHINA • MALAYSIA

G53FIV: Fundamentals of Information Visualization

Lecture 6: Visualization with R - Fundamentals

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

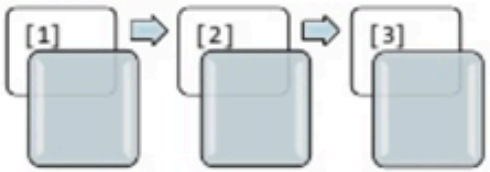
https://moodle.nottingham.ac.uk/course/view.php?id=68644

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.

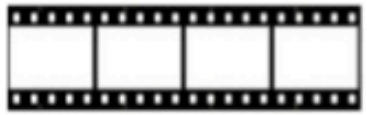
MATRIX

- 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: ist 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.

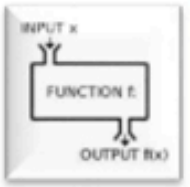
```
> weekdays=c("Monday","Tuesday","Wednesday","Thursday","Friday")
> wf <- factor(weekdays)
[1] Monday    Tuesday   Wednesday Thursday  Friday
Levels: Friday Monday Thursday Tuesday Wednesday
```

Used to group and summarize data:

```
WeekDaySales <- (DailySalesVector, wf, sum)
# Sum daily sales figures by M,T,W,Th,F
```

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.

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"
```

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.

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

R is a tool for...

Data Manipulation

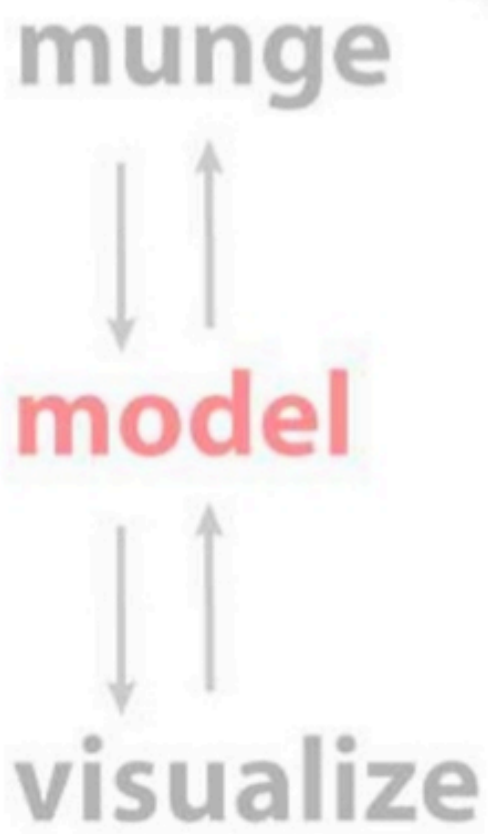
- connecting to data sources
- slicing & dicing data

Modeling & Computation

- statistical modeling
- numerical simulation

Data Visualization

- visualizing fit of models
- composing statistical graphics



Transform Data: A Swiss-Army Knife

	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

- 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),]**

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

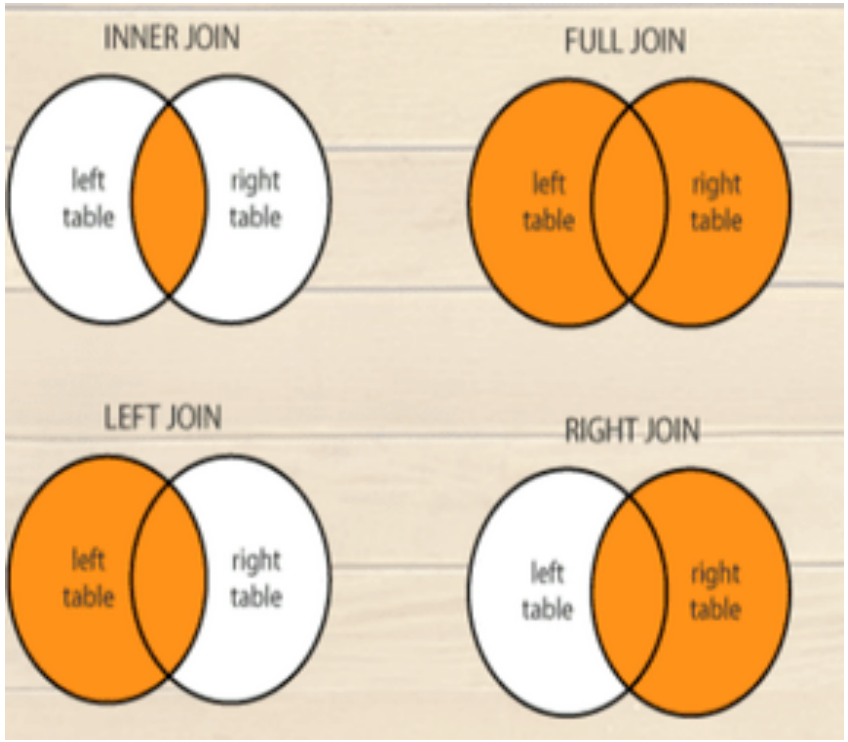
```
cut(Insurance$Claims, breaks=c(-
1,100,Inf), labels=c('lo','hi'))
```

- 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



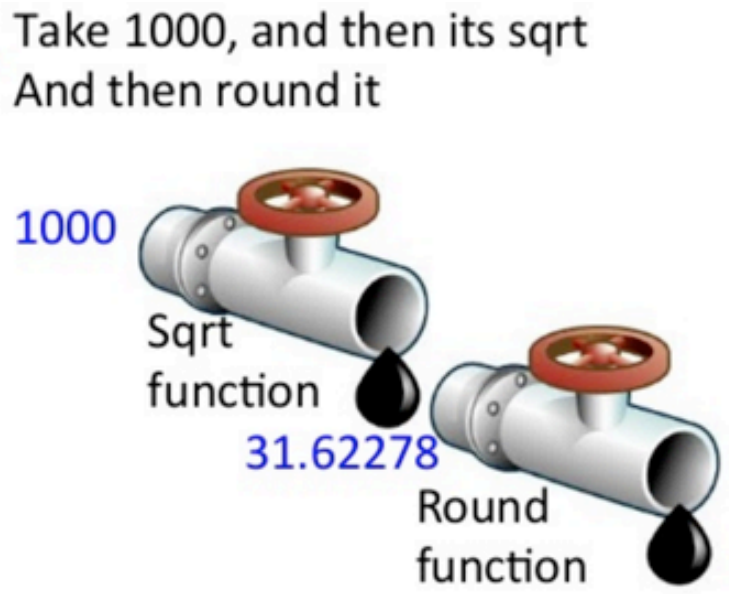
- merge(x=df1, y=df2, by.x="id", by.y="bid")

Pipe Operator

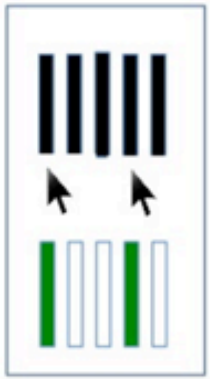
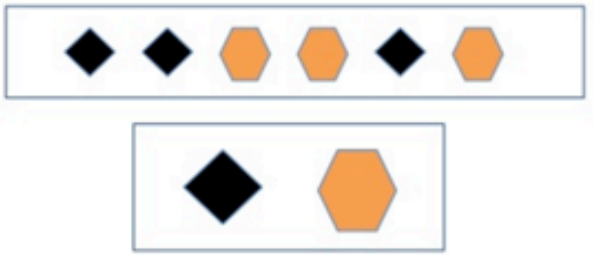
- Library(maggritr)
  - A R package launched on Jan 2014
  - A “magic” operator called the PIPE was introduced
  - %>%
  - i.e. “AND THEN”, “PIPE TO”

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

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



5 Basic Verbs

- FILTERows 
- SELECT Column Types 
- ArRange Rows (SORT) 
- Mutate (into something new) 
- **Summarize** by Groups 

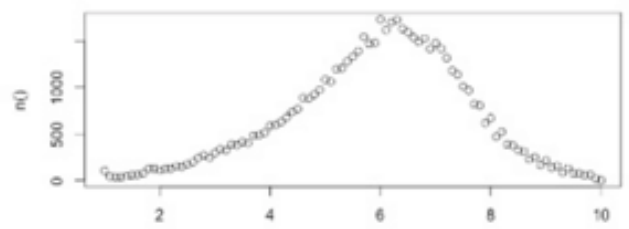
Chain the "Verbs" Together

- 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
```



The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

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

Visualization Tools

Chart Typologies
Excel, Many Eyes, Google Charts

Visual Analysis Grammars
VizQL, ggplot2

Visualization Grammars
Protovis, 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

Visualization Grammars
Protovis, D3.js

Declarative Languages

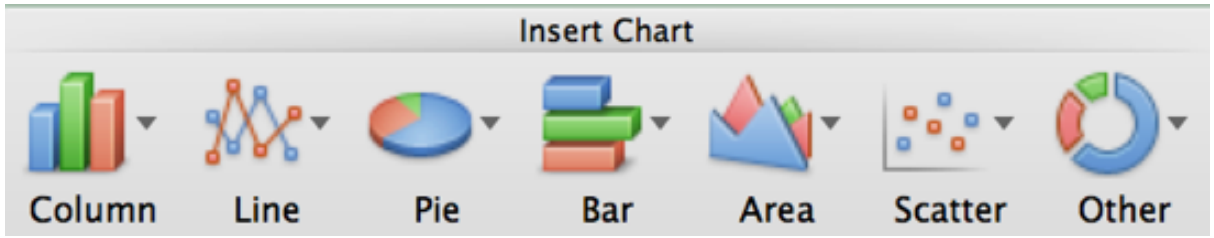
Component Architectures
Prefuse, Flare, Improvise, VTK

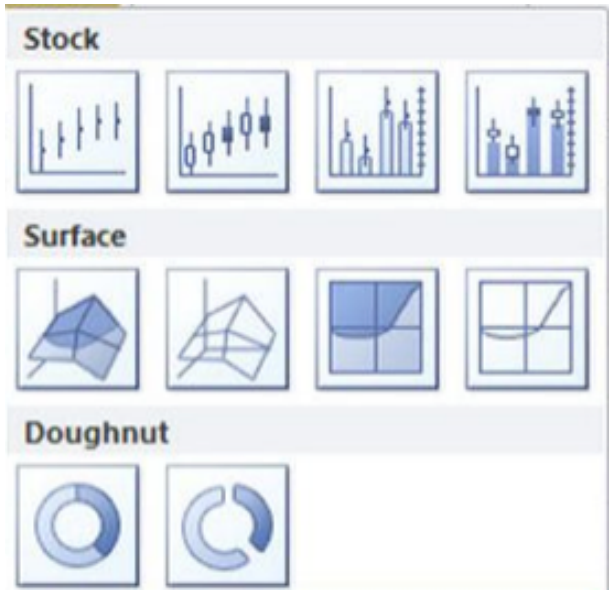
Graphics APIs
Processing, OpenGL, Java2D

Programming Toolkits

Chart Typology (Charting Tools)

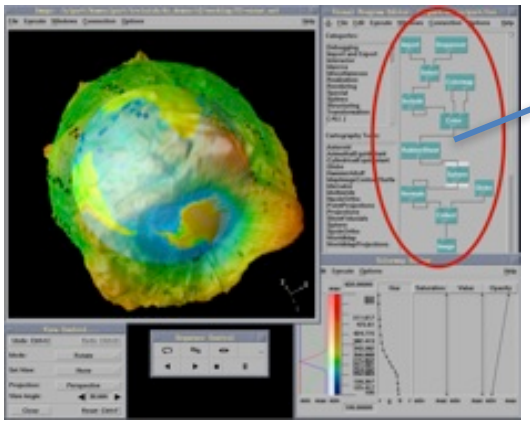
- Pick from a stock of templates
- Easy-to-use but limited expressiveness
- Prohibits novel designs, new data types

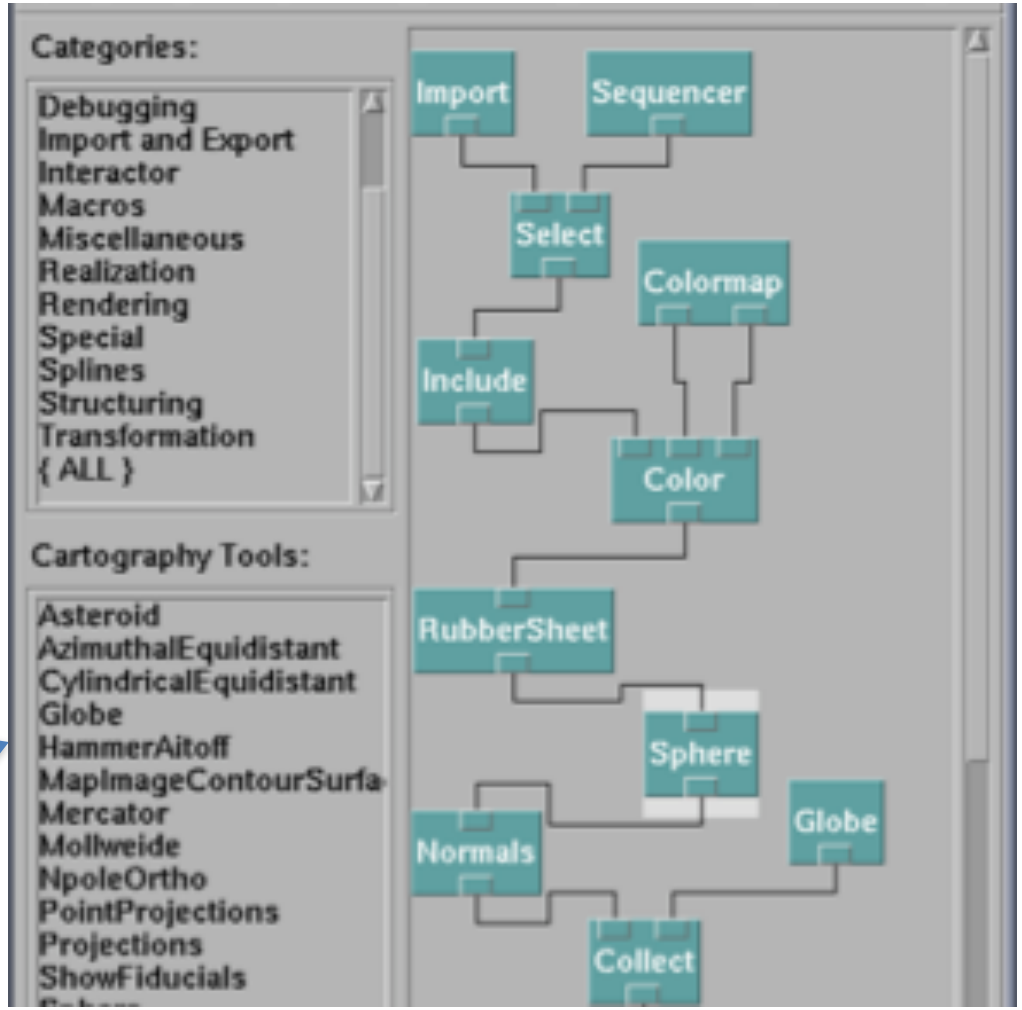




Component Architectures (Programming Toolkits)

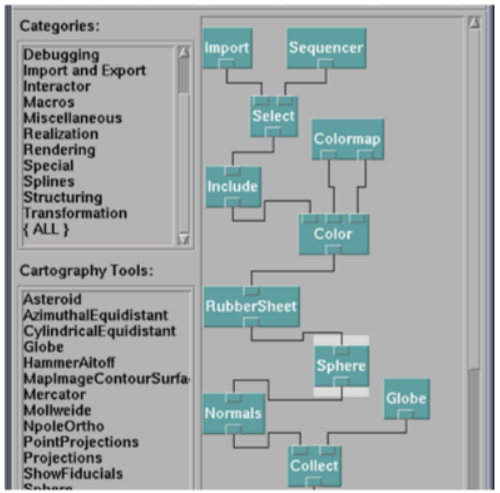
- Permits more combinatorial possibilities
- Novel views require new operators, which requires software engineering.





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

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.

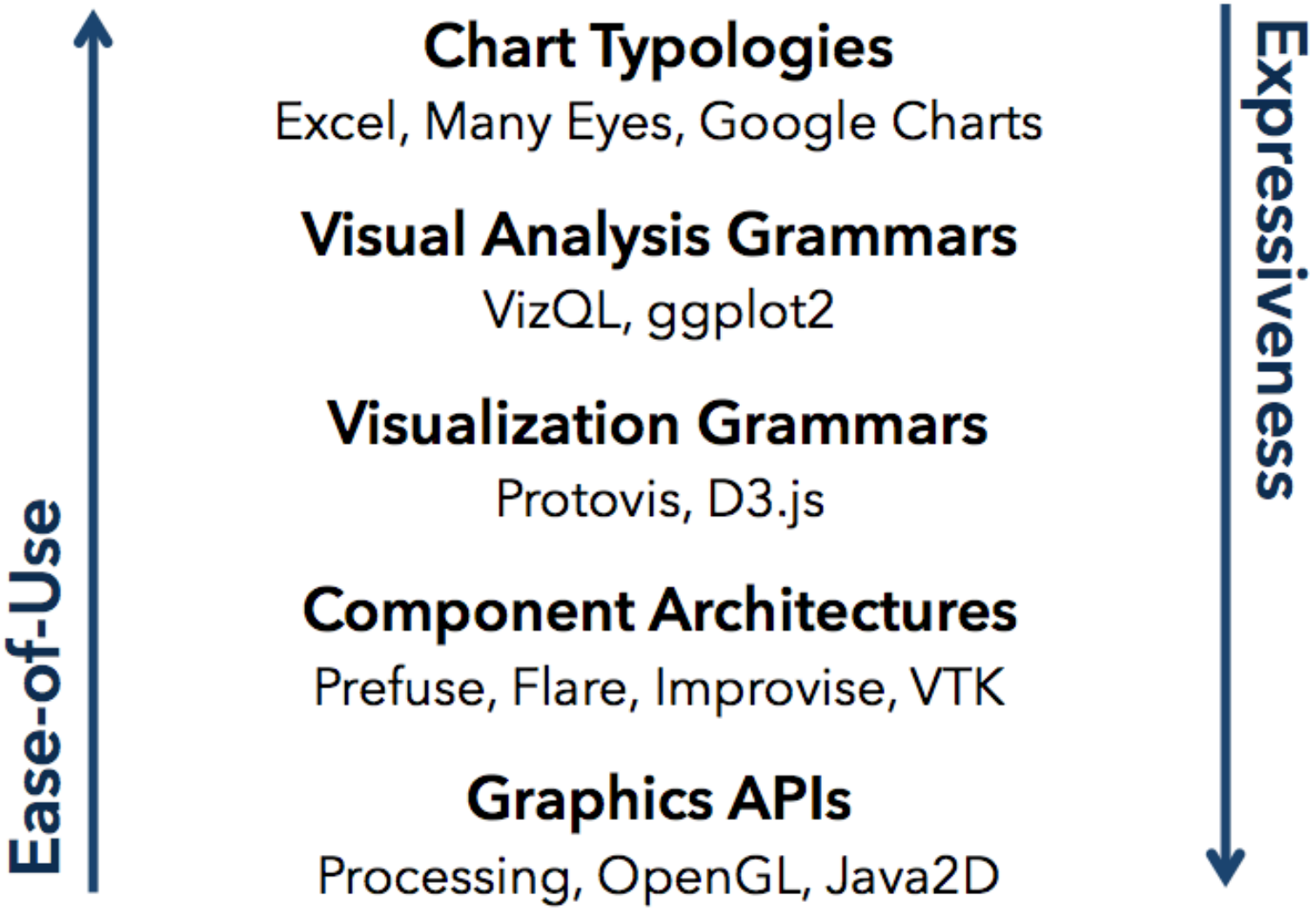
The Advantages of Declarative Languages

- **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.

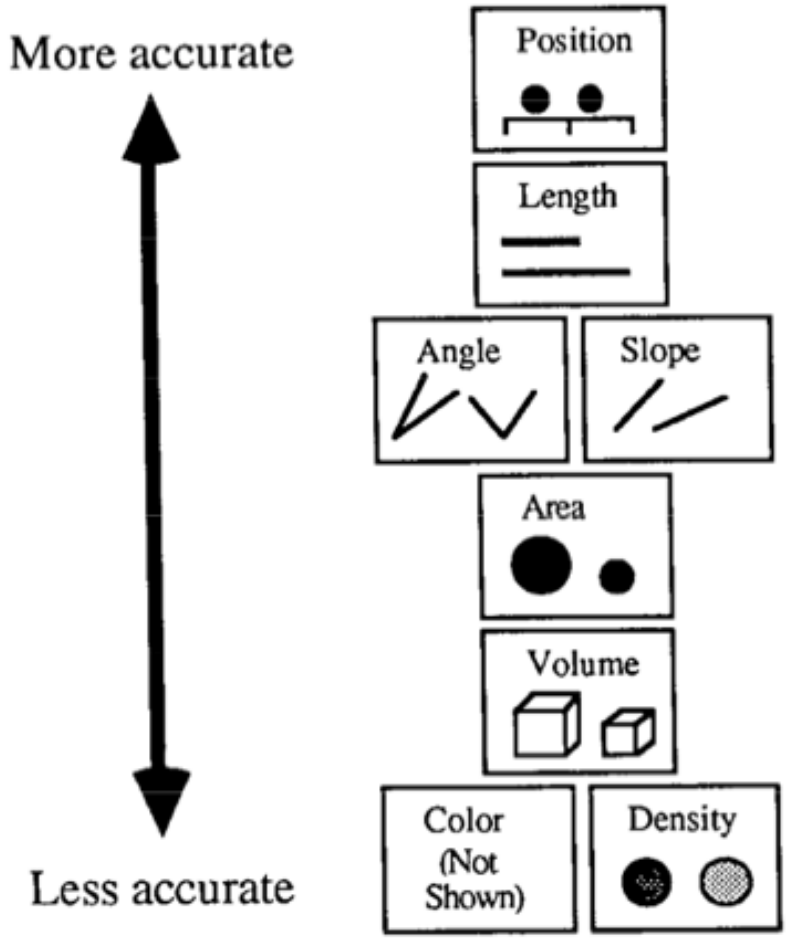
Tools Tradeoffs

- 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

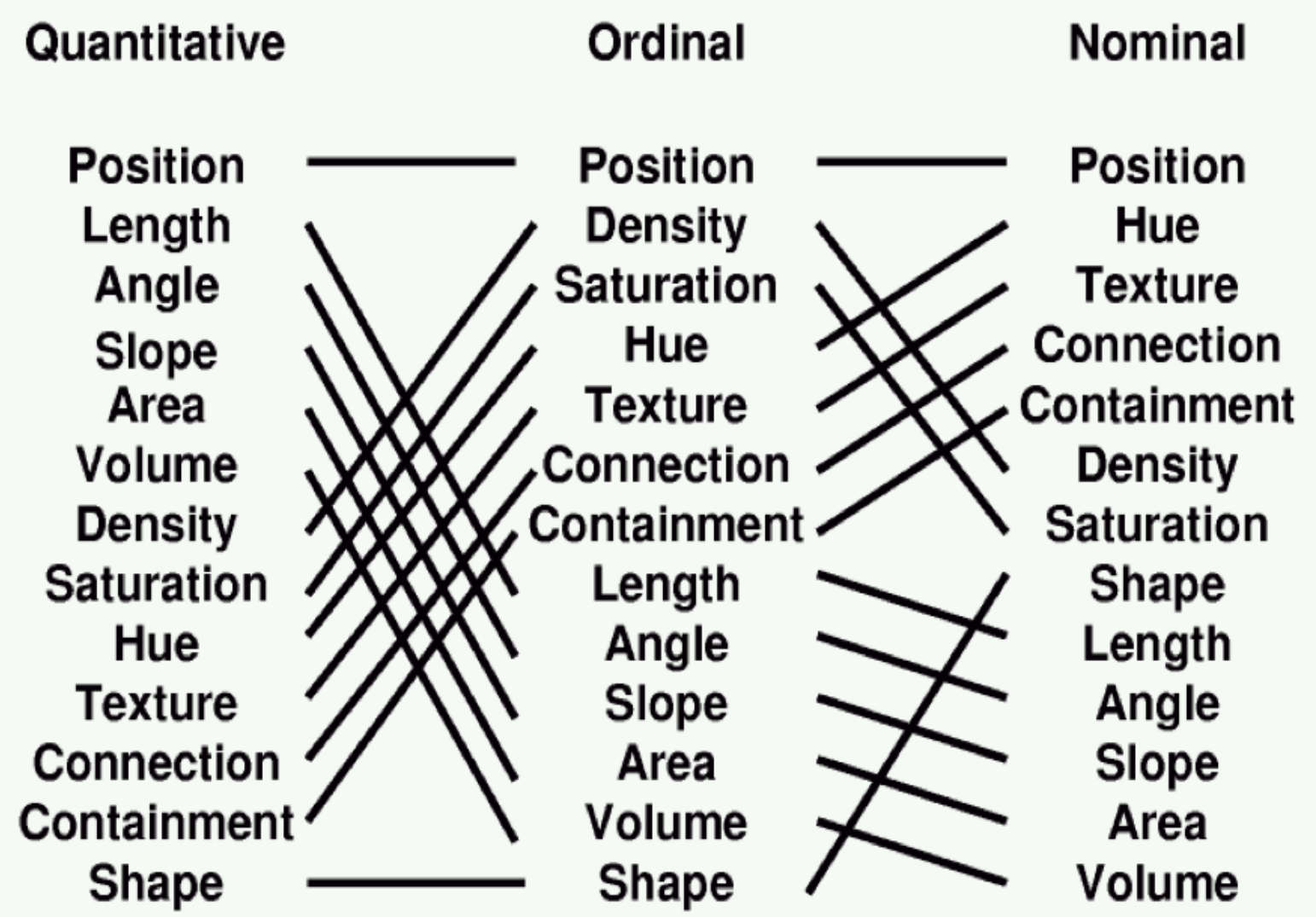


Effectiveness: Accuracy Ranking



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

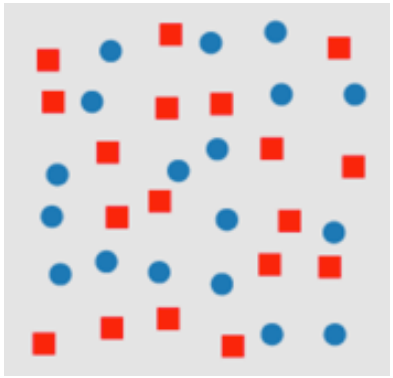
Conjectured Effectiveness of Encodings by Data Type



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

Perceptual Processing Model

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



Stage 1

Early, parallel detection of color, texture, shape, spatial attributes

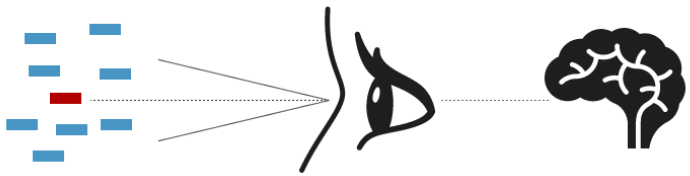
Stage 2

Serial processing of object identification (using memory) and spatial layout, action

Information visualization: Perception for Design (3rd Edition), Colin Ware

Stage 1: Pre-attentive Processing - Low-level, Parallel

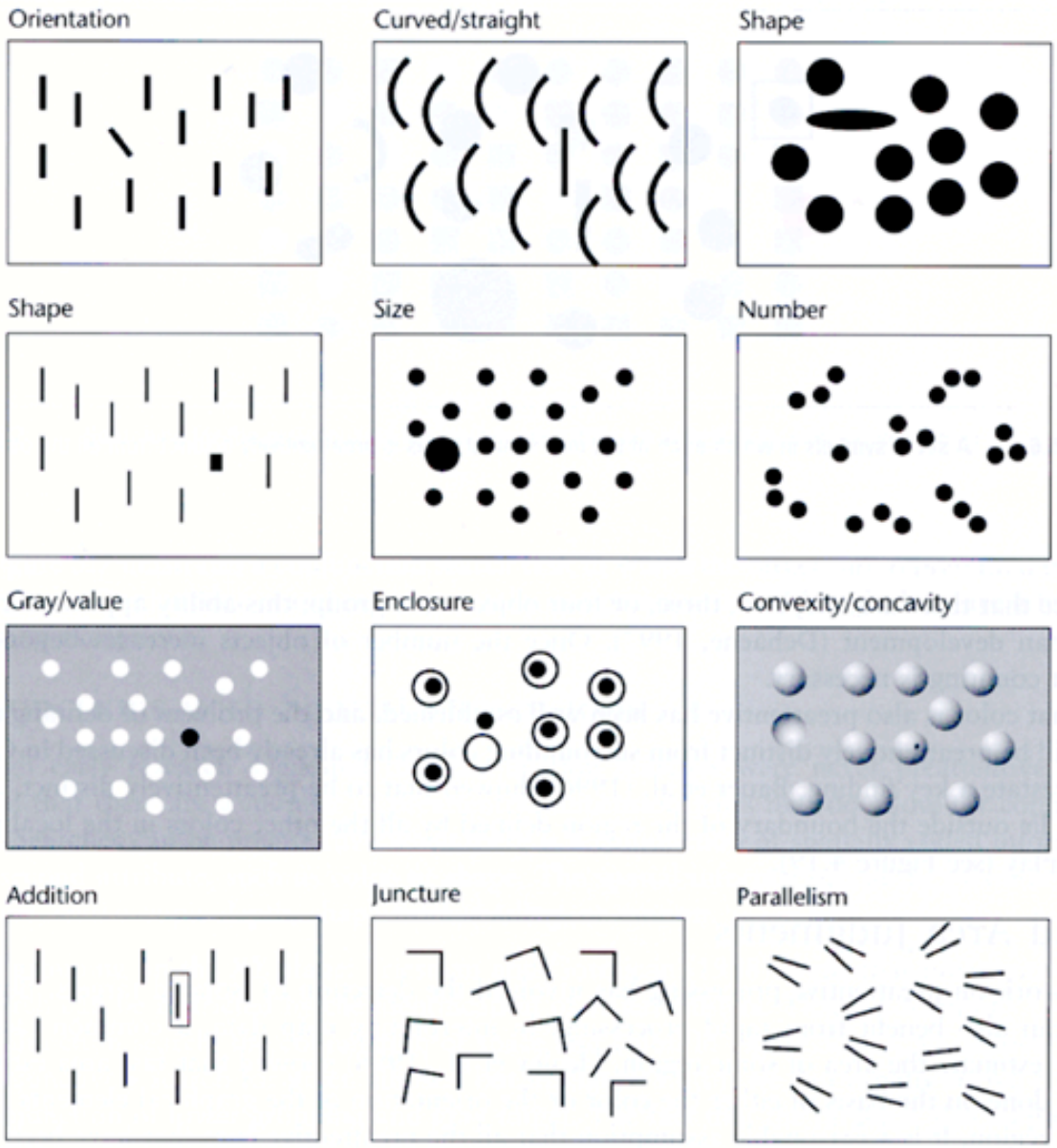
- Neurons in eye & brain responsible for different kinds of information
  - Orientation, color, texture, movement, etc.
- Arrays of neurons work in parallel, occurs "automatically" and rapidly
  - Generally less than 200-250 msecs
- Information is transitory, briefly held in iconic store
- Bottom-up 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 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

Information visualization: Perception for Design (3rd Edition), Colin Ware

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

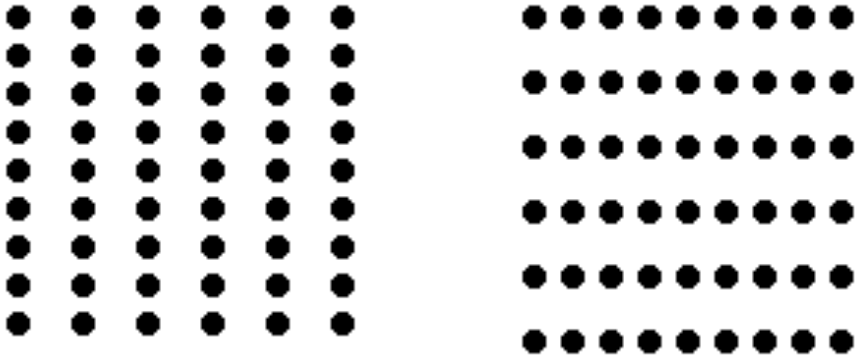
"All else being equal, elements that are related by X tend to be grouped perceptually into higher-order units."

— Stephen Palmer

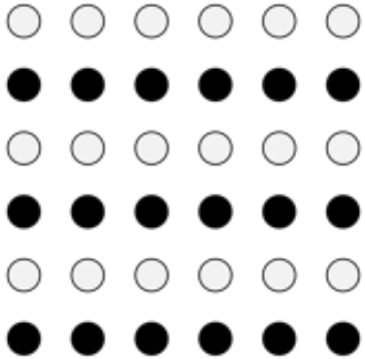
- Proximity
- Similarity
- Connectedness
- Continuity
- Symmetry
- Closure
- Figure/Ground
- Common Fate

Gestalt Grouping Principles

- Proximity
  - Things close together are perceptually grouped together



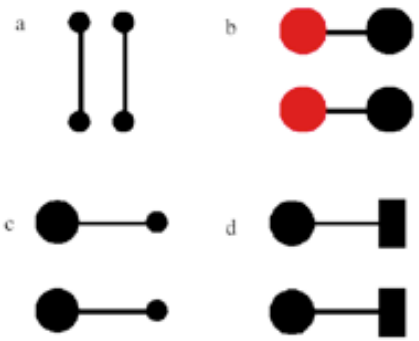
- Similarity
  - Similar elements get grouped together



Rows dominate due to similarity

Gestalt Grouping Principles

- Connectedness
  - Connecting different objects by lines unifies them

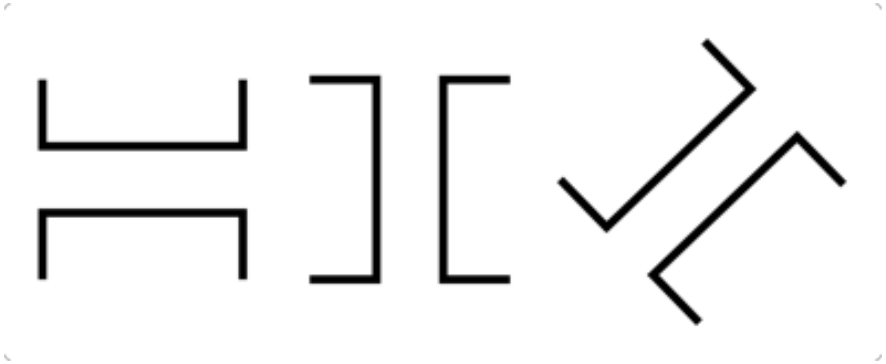


Connectedness overrules proximity, size, color shape

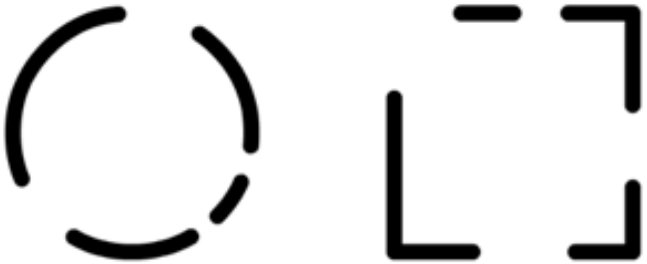
- Continuity
  - More likely to construct visual entities out of smooth, continuous visual elements

Gestalt Grouping Principles

- Symmetry
  - Symmetrical patterns are perceived more as a whole

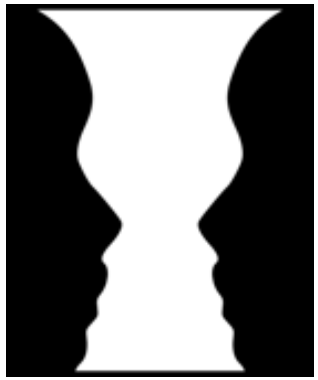


- Closure
  - A closed contour is seen as an object

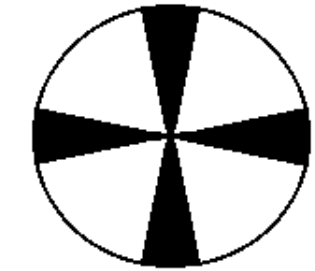


Gestalt Grouping Principles

- Figure/Ground
  - Figure is foreground, ground is behind



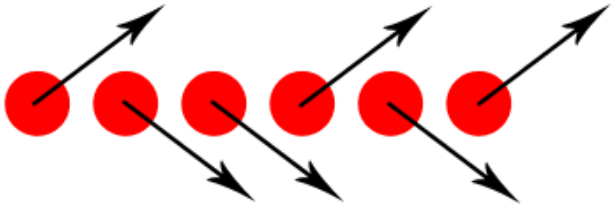
ambiguous



Relative size

surroundedness

- Common Fate (Synchrony)
  - Elements that move in the same direction are perceived as more related



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

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

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

Categorization based on User Intent

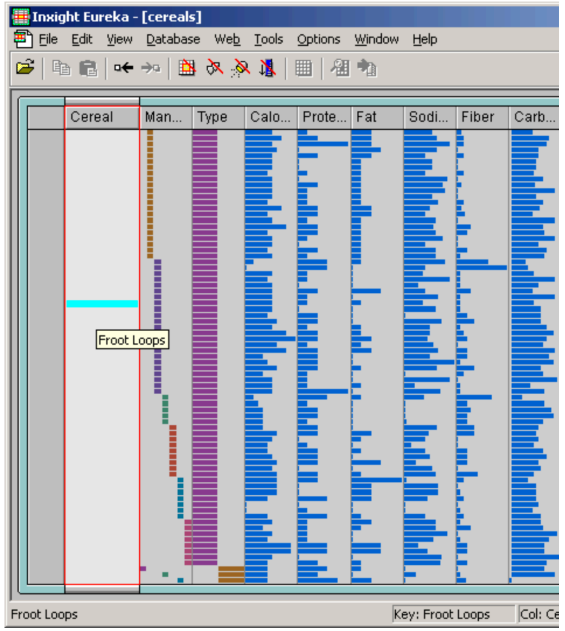
- Don't focus so much on particular interactive operations and how they work
- Interaction is ultimately being done by a person for a purpose
  - Seeking more information, solving a problem
  - Fundamental aspect of exploratory, analytic discourse
- Taxonomy based on **User Intent**
  - What a user wants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

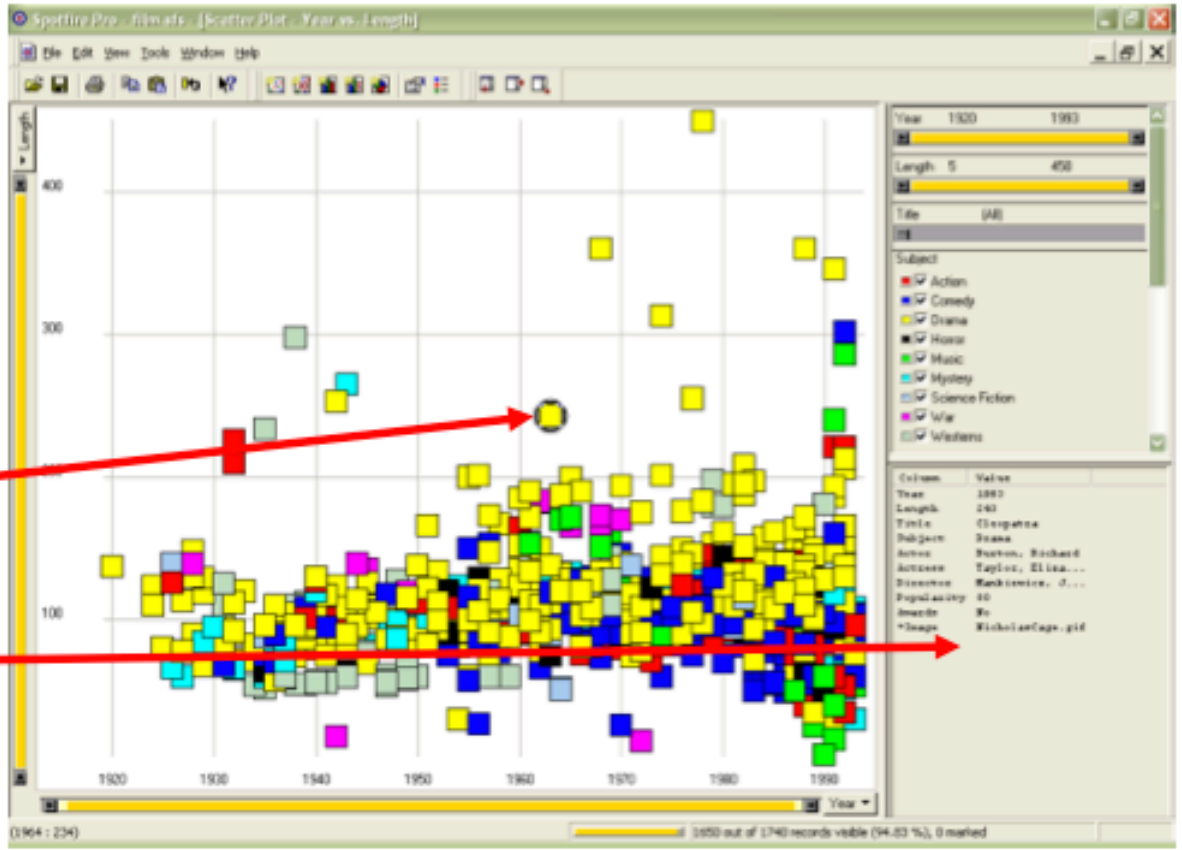


Mouse Selection

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

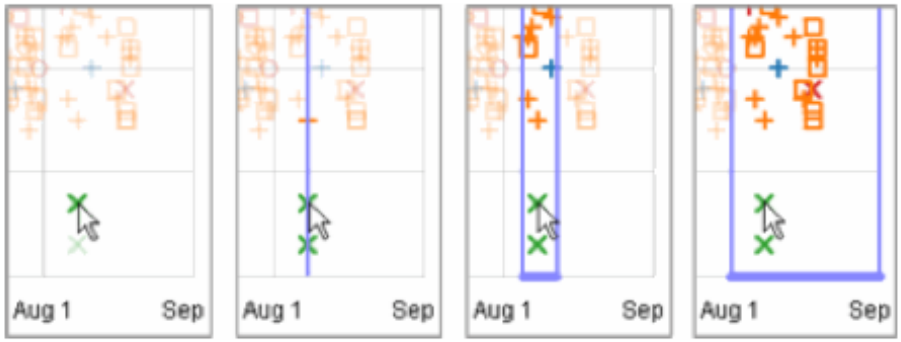
Selected item

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/

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

Heer, Jeffrey, Maneesh Agrawala, and Wesley Willett. "Generalized selection via interactive query relaxation." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008.

2. Explore

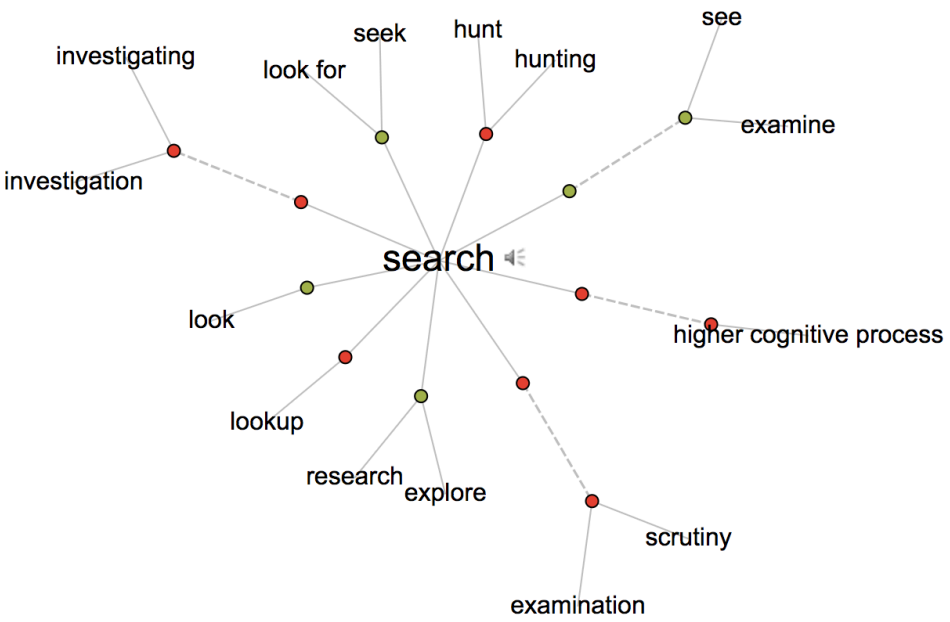
- "Show me something different"

- 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

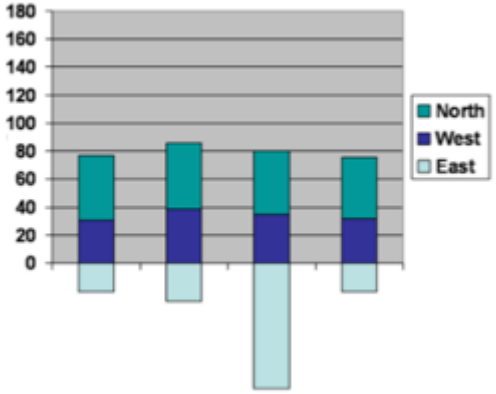
- Linkages between cases
- Exploring one may lead to another
- Example:
  - Visual Thesaurus



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

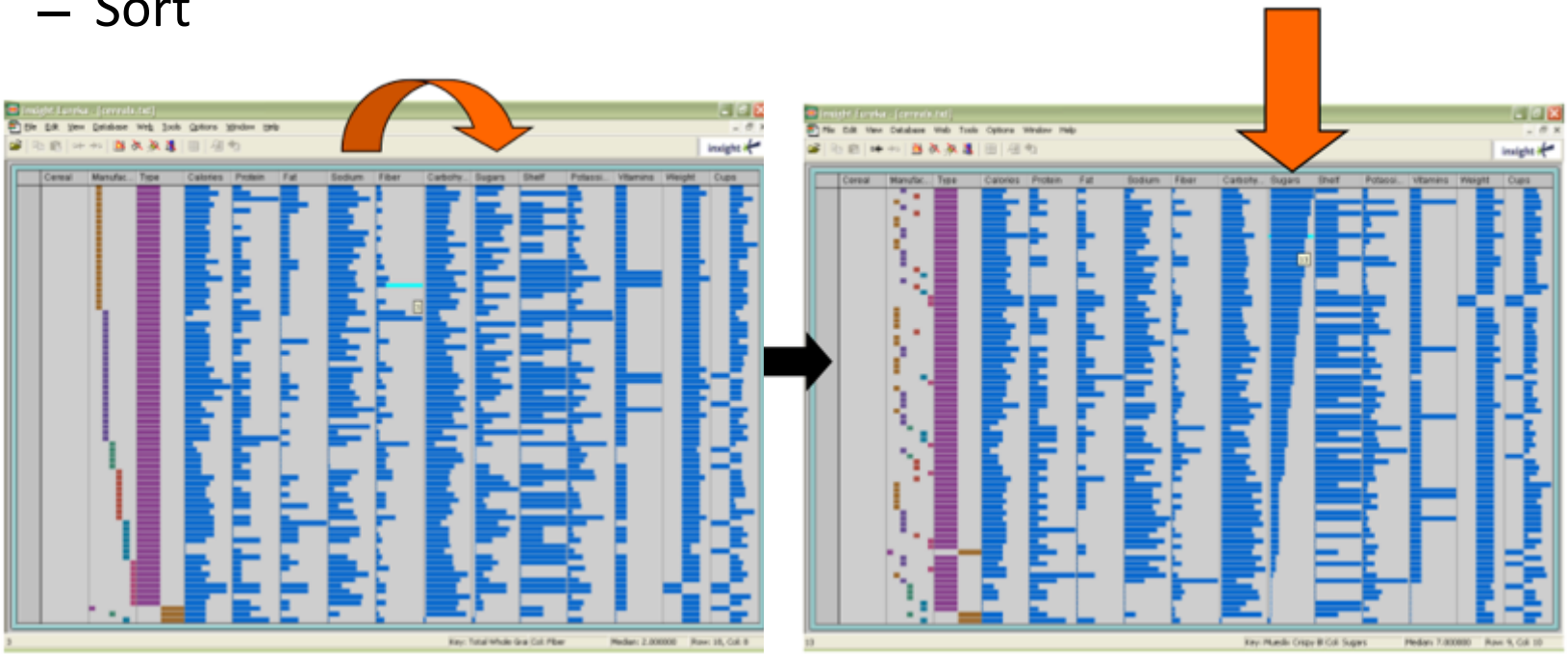
3. Reconfigure

- "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



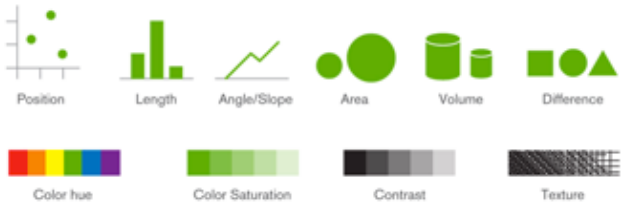
Rearrange View and Sorting

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



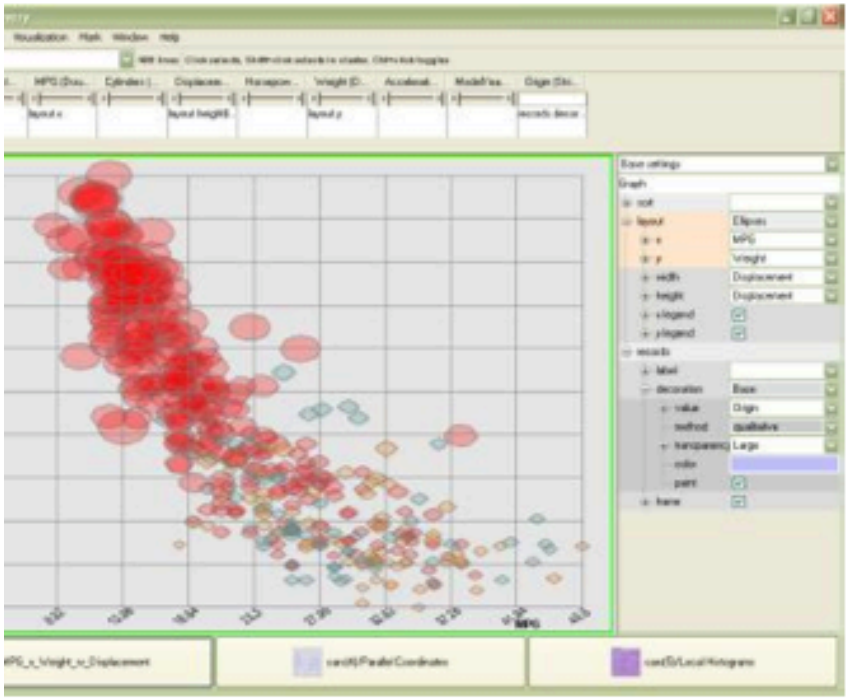
4. Encode

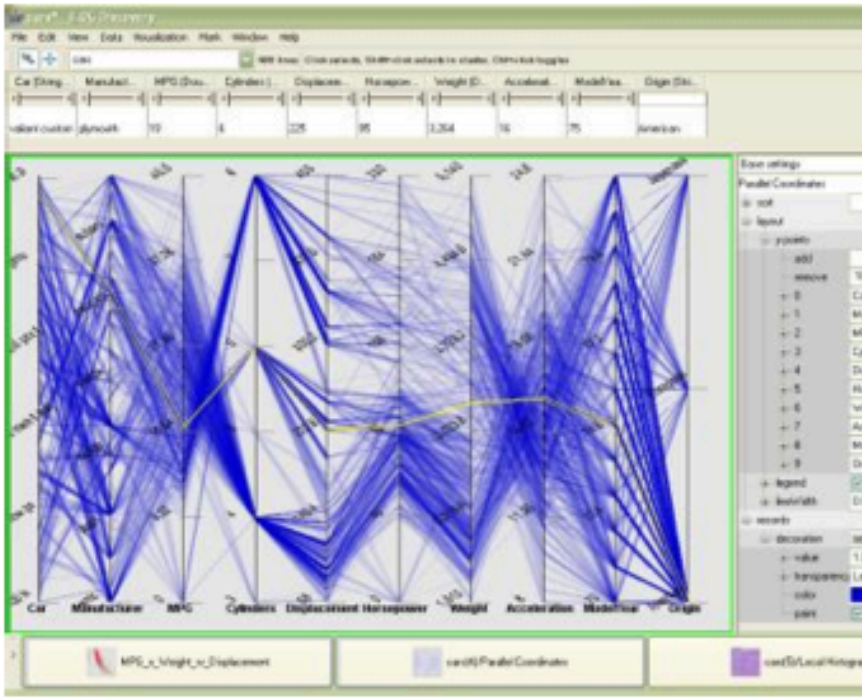
- "Show me a different representation"
- Change visual appearances
  - color encoding, size, orientation, font, shape



- 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

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

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

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

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

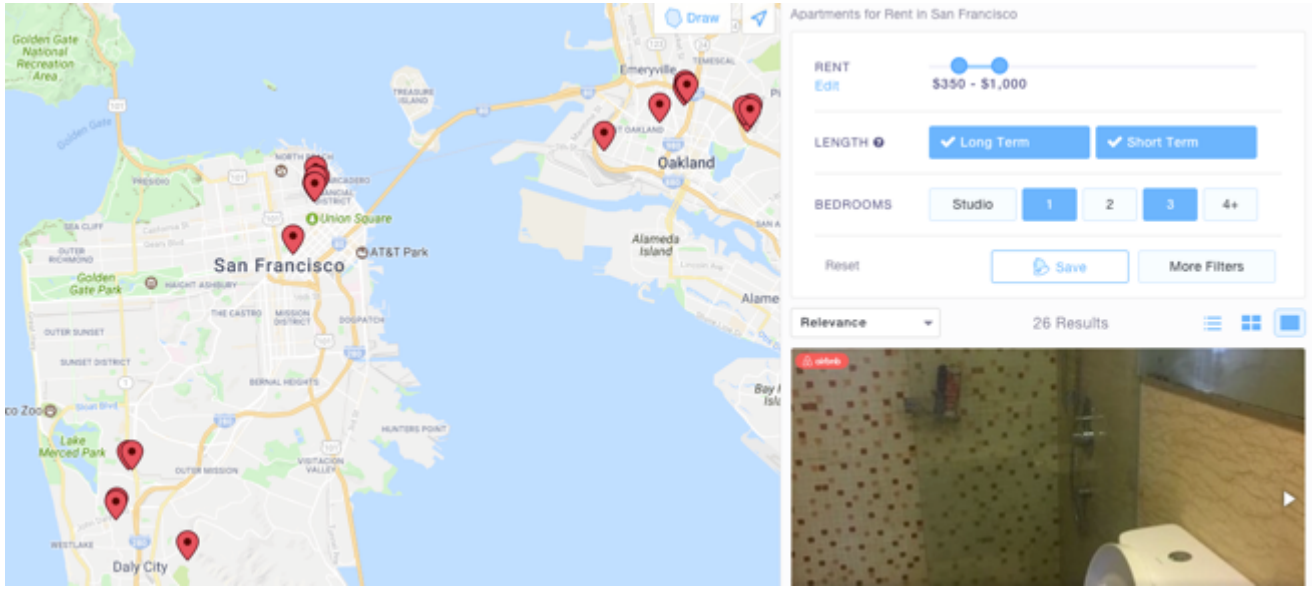
- Specifying a query brings immediate display of results
- Responsive interaction (< .1 sec) with data, concurrent presentation of solution
- "Fly through the data", promote exploration, make it a much more "live" experience
  - Timesharing vs. batch

- 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

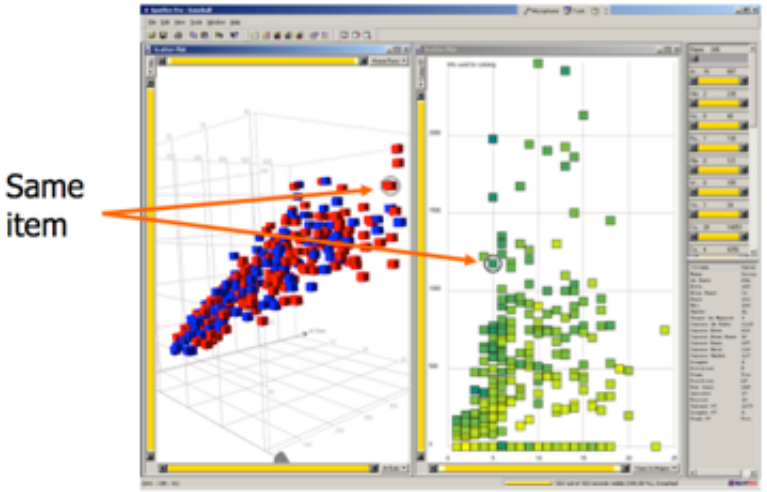
- "Show me related items"

- 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



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

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

G53FIV: Fundamentals of Information Visualization

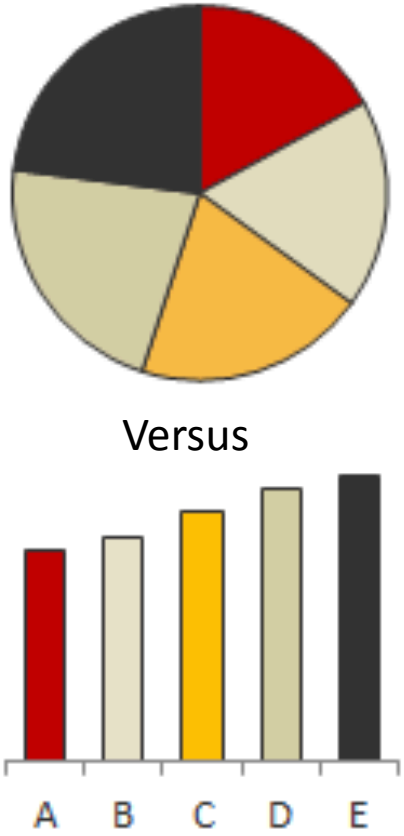
Lecture 10: Evaluation

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

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

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

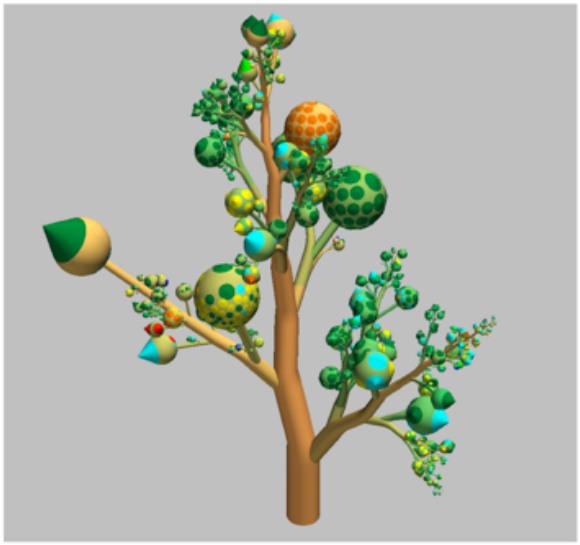


Figure 5: Botanic visualization contents of a hard disk [10, 27]. Useful or just a nice picture?

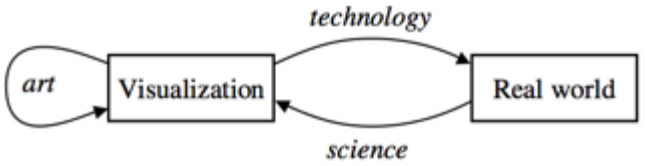


Figure 6: Views on visualization

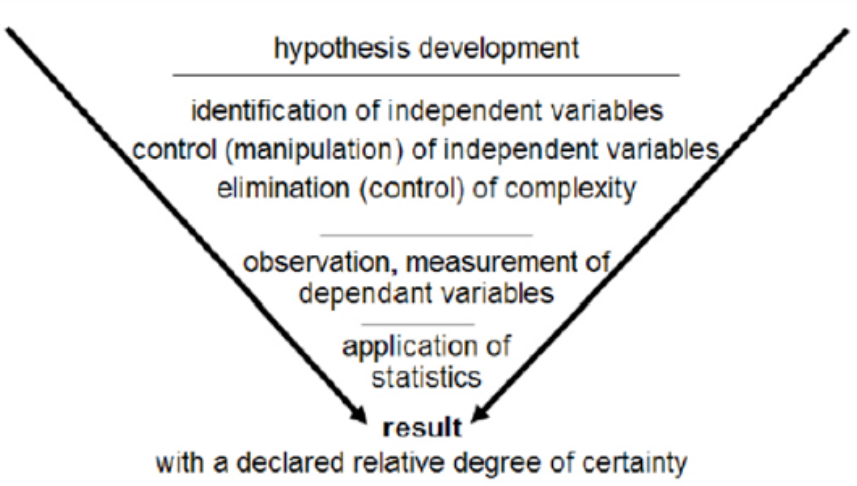
Evaluation Approaches

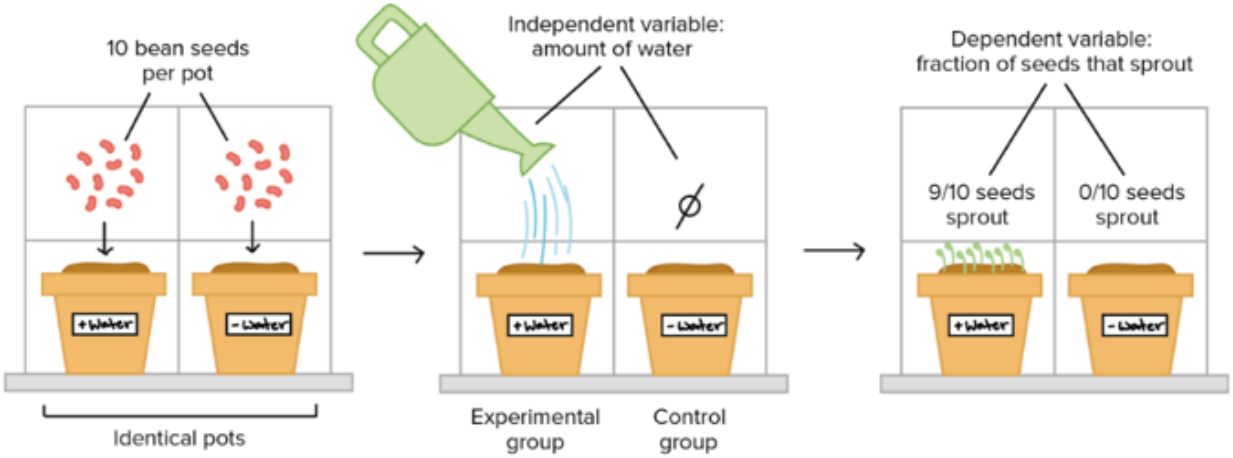
- Many different Forms
  - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, ...

- 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, …





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

- 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

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?

Qualitative Challenges

- Sample sizes
- Subjectivity
- Analyzing qualitative data

Meta-evaluate Evaluation Approaches

- Desirable features
  - Generalizability
  - Precision
  - Realism

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	Unstructure, 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

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

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

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.

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?

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, @nottinghamforestfball
  - 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

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

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} = count(t) in d
  - Can take log frequency: $\log(1 + tf_{td})$
  - Can normalize to show proportion ($tf_{td} / \Sigma_t tf_{td}$)

- 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

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

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

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

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

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

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

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

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

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

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

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

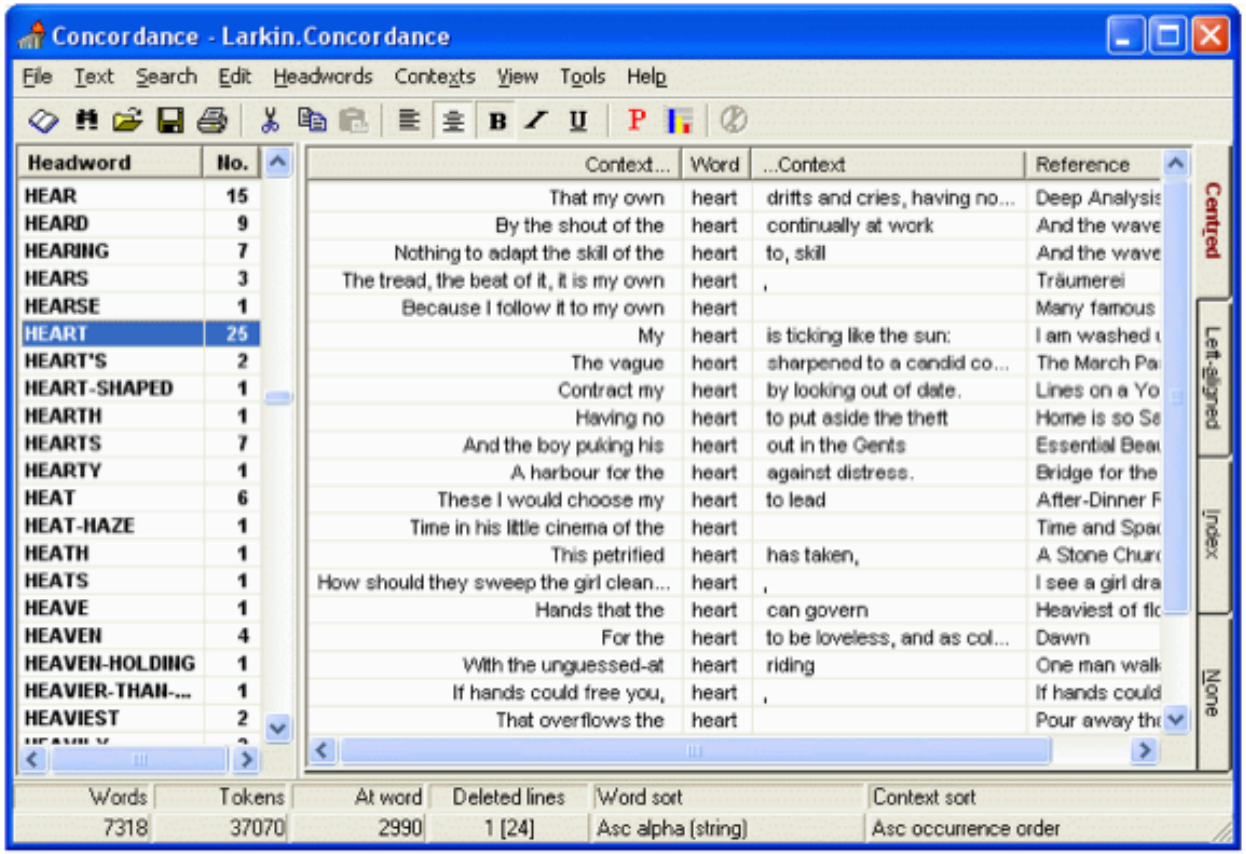
Descriptive Phrases

- 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/

Word / Phrase Context

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



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

Phrase Nets

- Concordances show local, repeated structure, but what about other types of patterns?
  - Lexical: <A> at <B>
  - 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

van Ham et al TVCG (InfoVis) ‘09

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

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

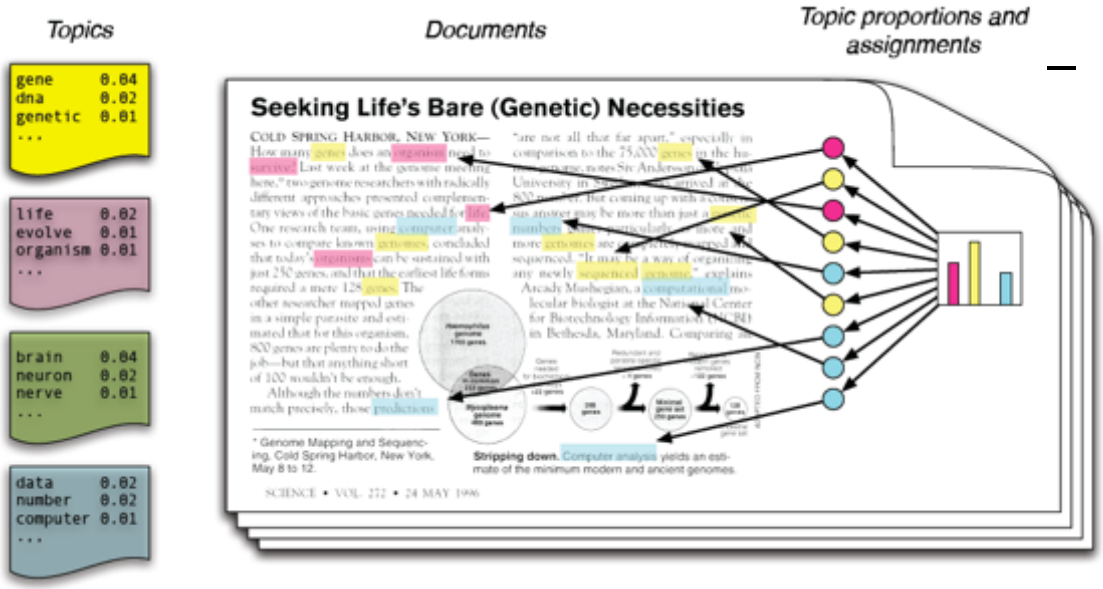
Visualizing Document Collections
Named Entity Recognition

- 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?

Similarity & Clustering

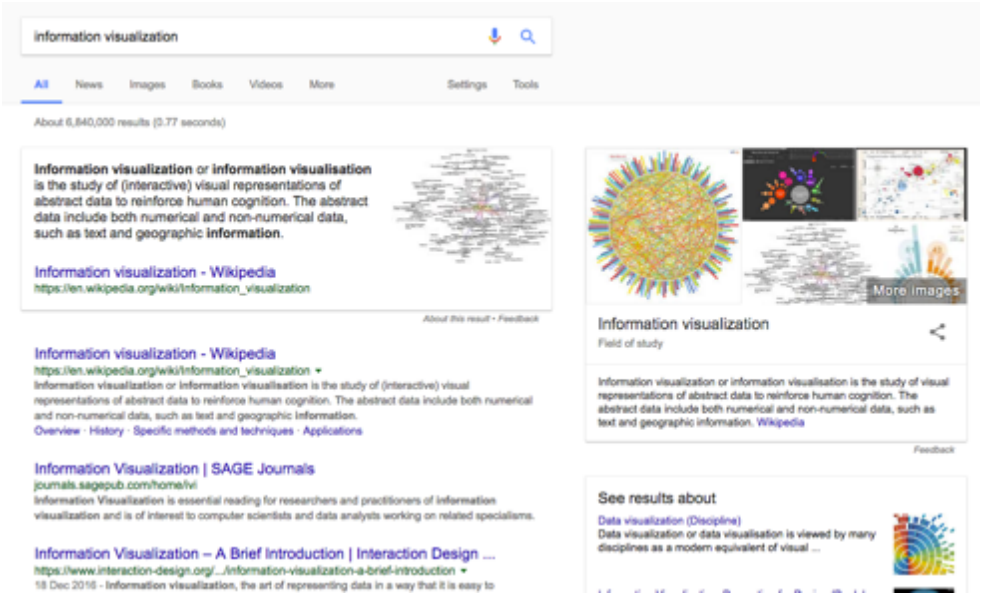
- Compute vector distance among docs
  - For TF.IDF, typically cosine distance Similarity measure can be used to cluster

- 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



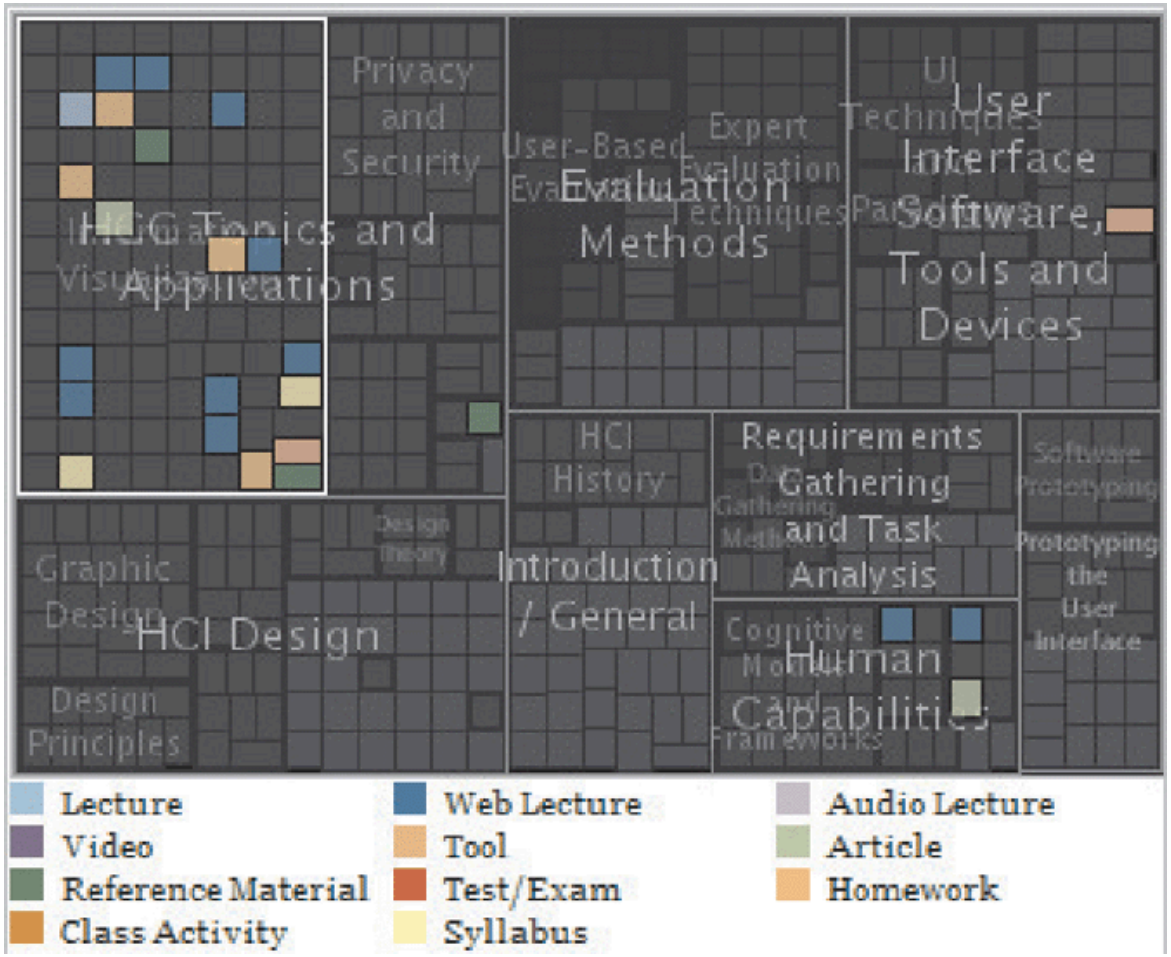
Information Retrieval (IR)

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



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

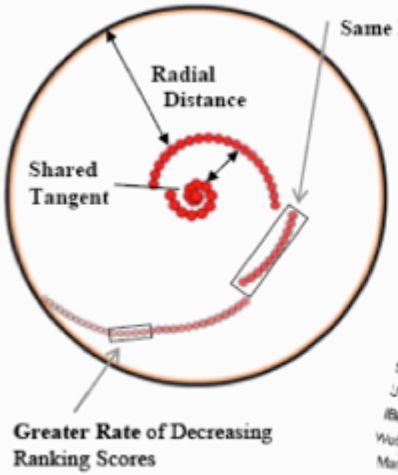
Result Maps



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

Clarkson, Desai & Foley InfoVis'09

RankSpiral



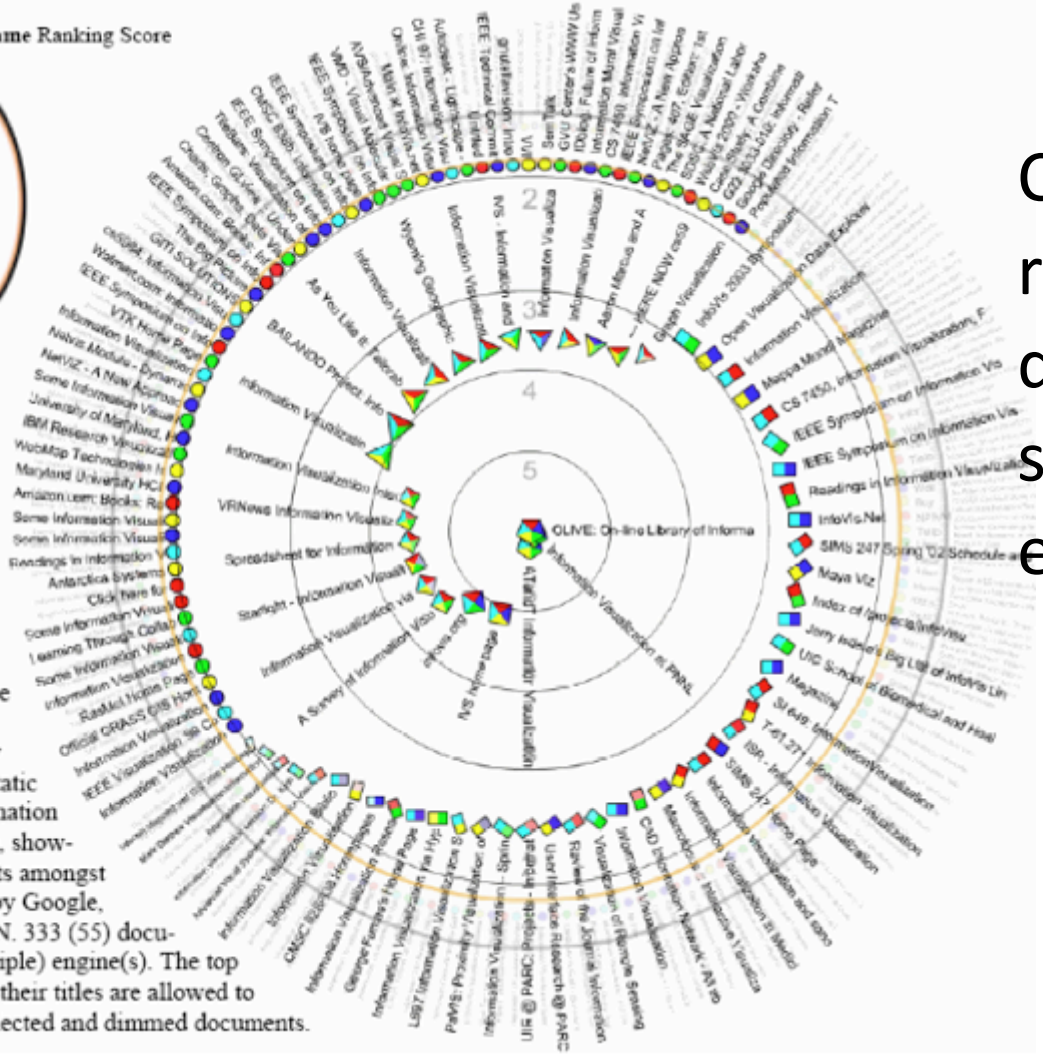


Figure 1: (Top) *RankSpiral* places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases to-ward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral that maximizes information density and minimizes occlusions, showing here the 388 unique documents amongst the top 100 documents retrieved by Google, Teoma, AltaVista, Lycis and MSN. 333 (55) documents were found by single (multiple) engine(s). The top 100+ documents are selected and their titles are allowed to extend across the remaining unselected and dimmed documents.

Color represents different search engines

Spoerri InfoVis ’04 poster

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

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

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

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

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)

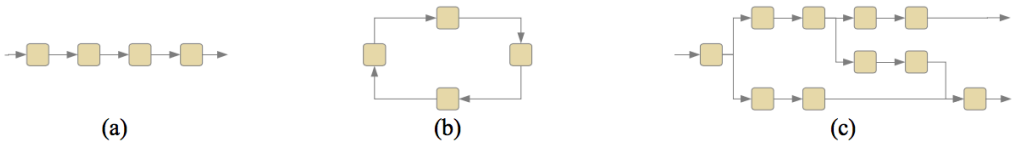
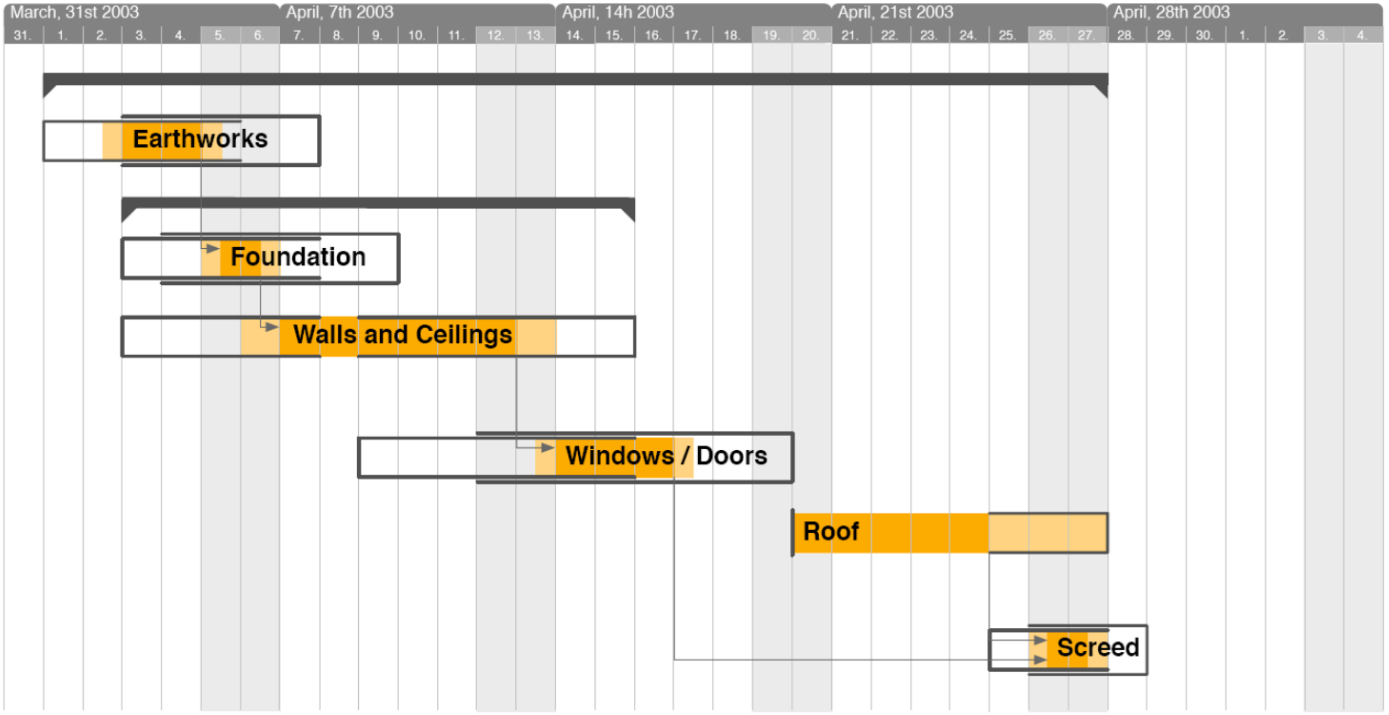


Fig. 2. Structure of time: (a) Linear time; (b) Cyclic time; (c) Branching time.

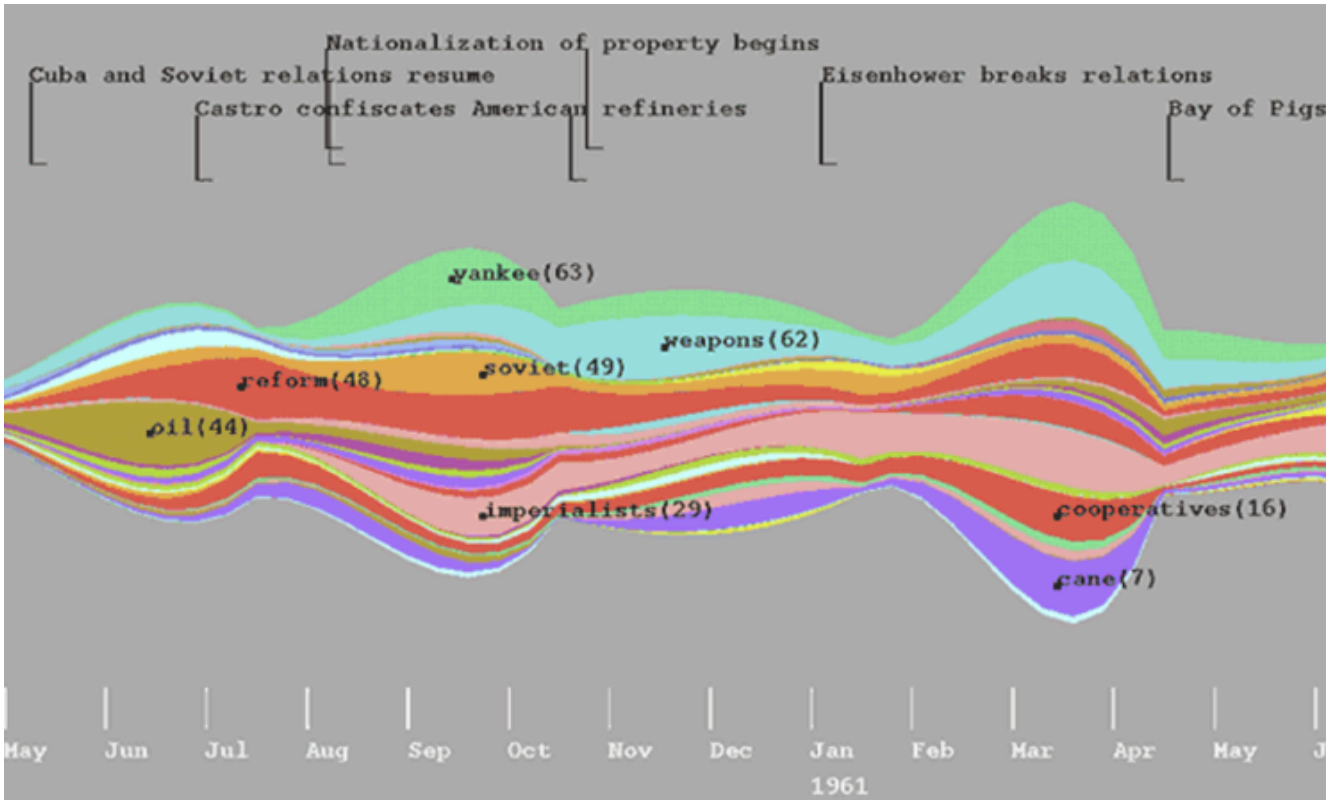
Aigner, Wolfgang, et al. "Visualizing time-oriented data - a systematic view." Computers & Graphics 31.3 (2007): 401-409.

Gantt Chart



Time	Temporal primitives	Intervals
	Structure of time	Linear, Branching
Data	Frame of reference	Abstract
	Number of variables	Univariate
	Level of abstraction	Data
Representation	Time dependency	Static
	Dimensionality	2D

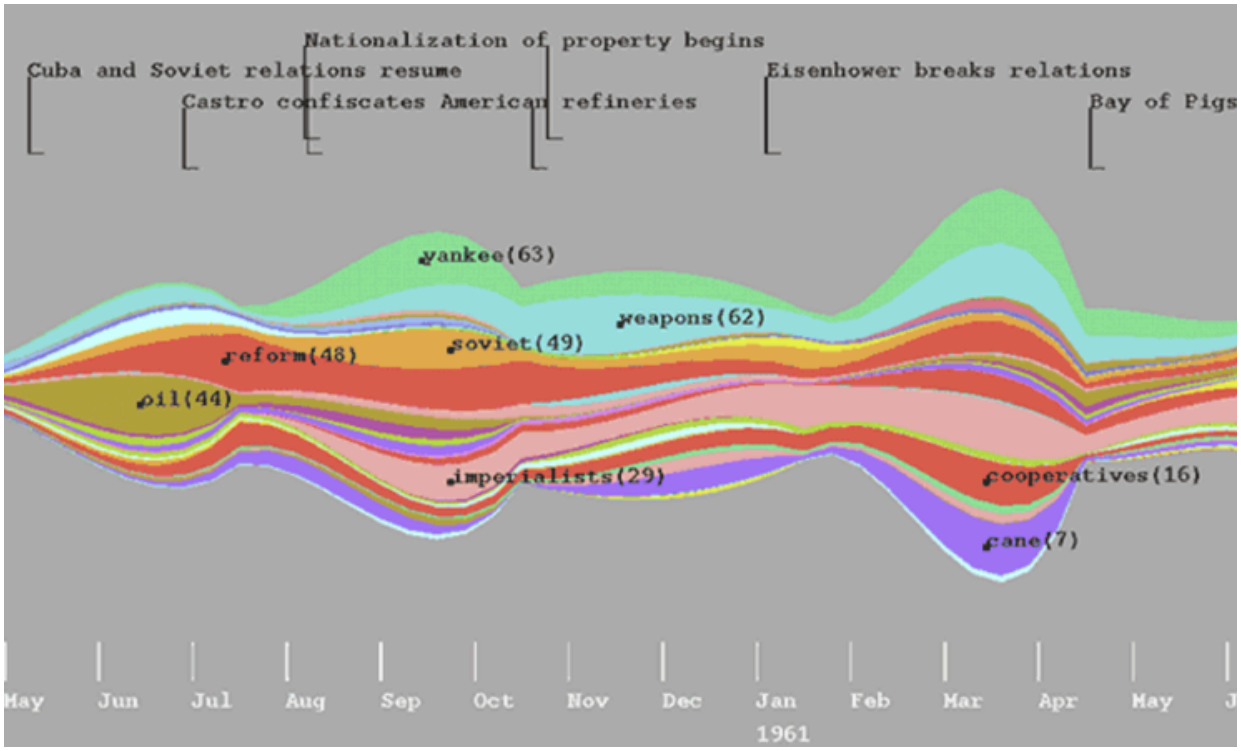
Theme River
(Stream Graphs / Stacked Area Charts)



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) .

(Optional) Havre, Susan, et al. "Themeriver: Visualizing thematic changes in large document collections." IEEE transactions on visualization and computer graphics 8.1 (2002): 9-20.

Theme River



Time	Temporal primitives	Time Points
	Structure of time	Linear
Data	Frame of reference	Abstract
	Number of variables	Multivariate
	Level of abstraction	Data
Representation	Time dependency	Static
	Dimensionality	2D

Havre, Susan, et al. "Themeriver: Visualizing thematic changes in large document collections." IEEE transactions on visualization and computer graphics 8.1 (2002): 9-20.

Cyclic Patterns

- Time data are often cyclic
  - Spiral displays are good to bring out cyclic patterns
  - One period per loop (for example, a year)

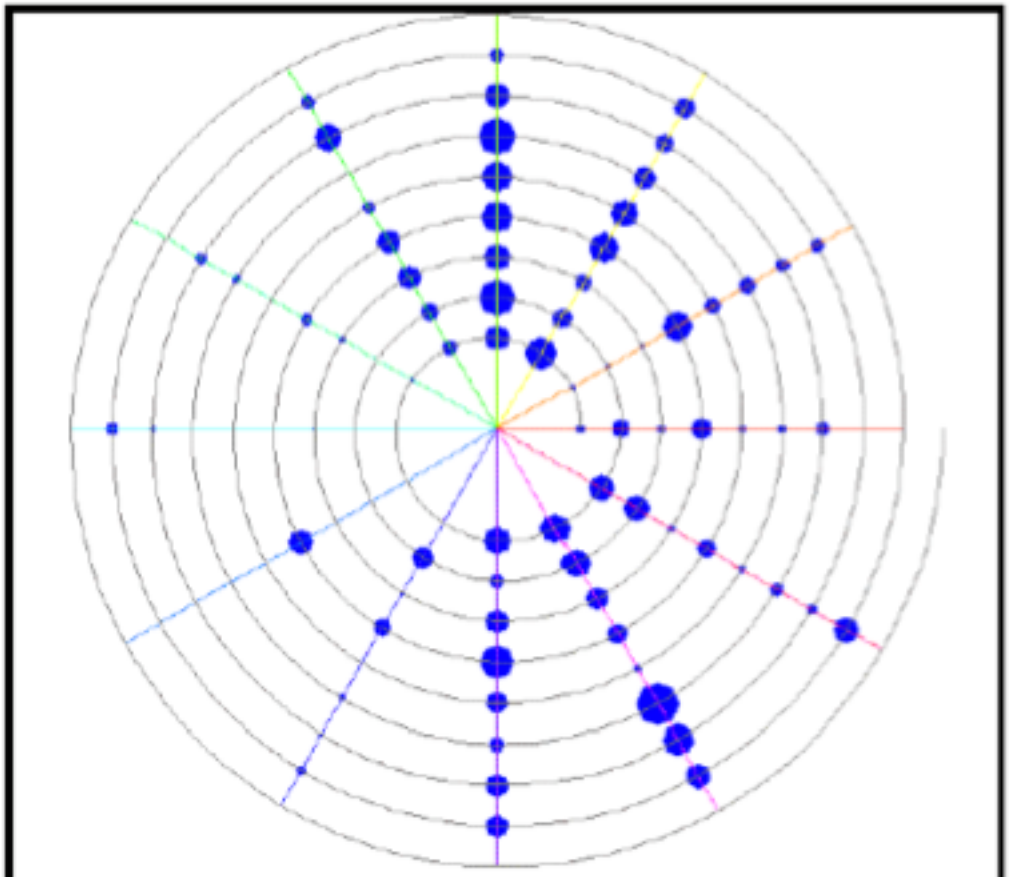


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Carlis & Konstan UIST '98

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

Cyclic Patterns

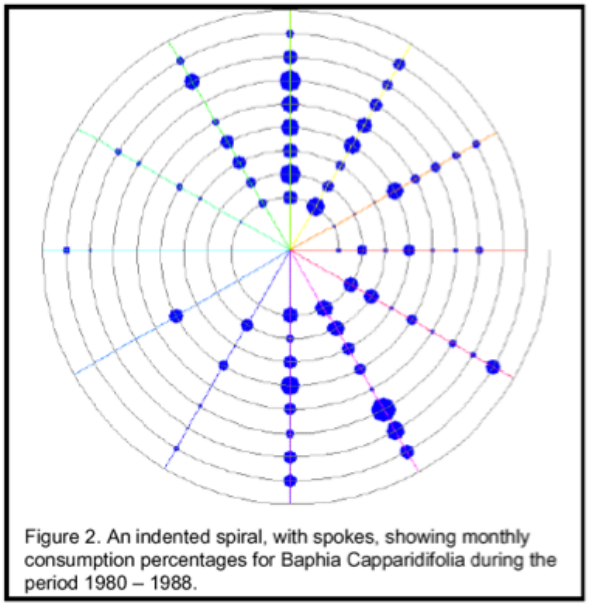


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Time	**Temporal primitives**	**Time Points**
	Structure of time	**Cyclic**

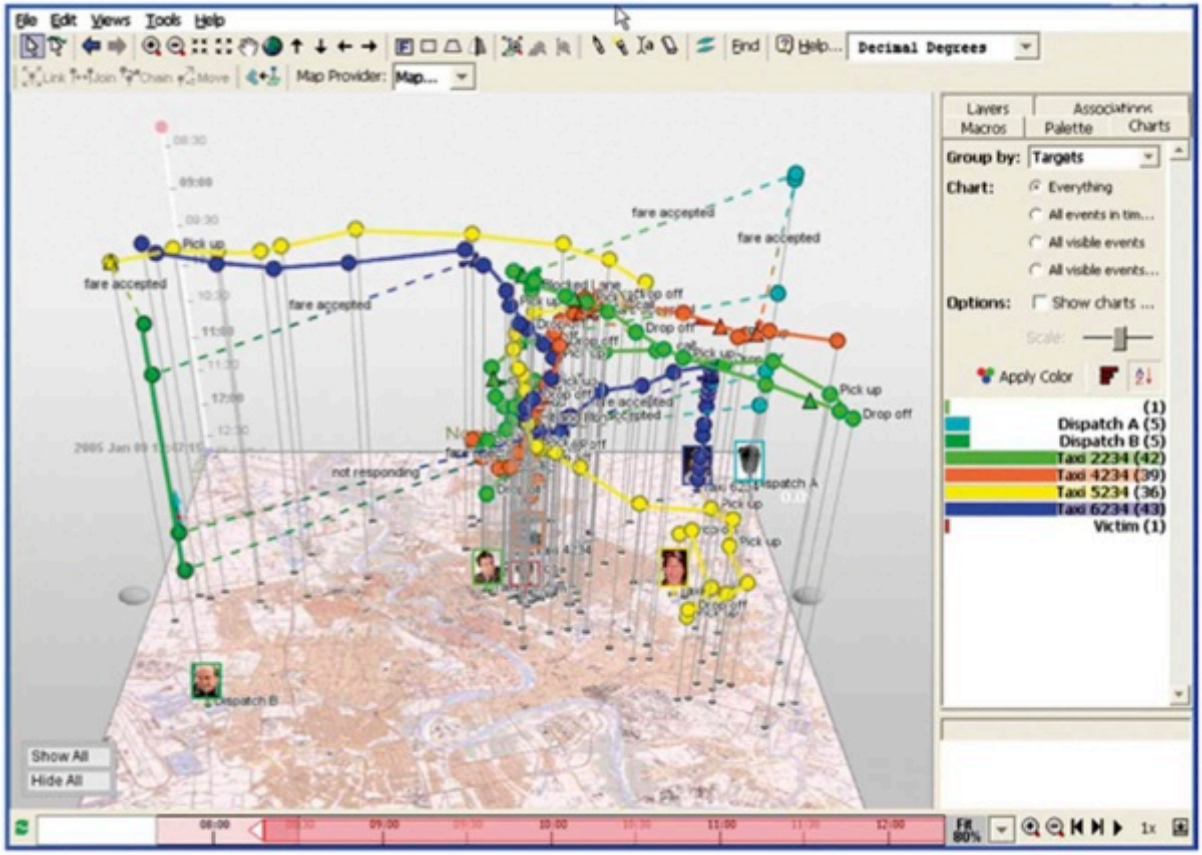
Data	**Frame of reference**	**Abstract**
	Number of variables	**Univariate**
	Level of abstraction	**Data**

Representation	**Time dependency**	**Static**
	Dimensionality	**2D**

Carlis & Konstan UIST '98

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.

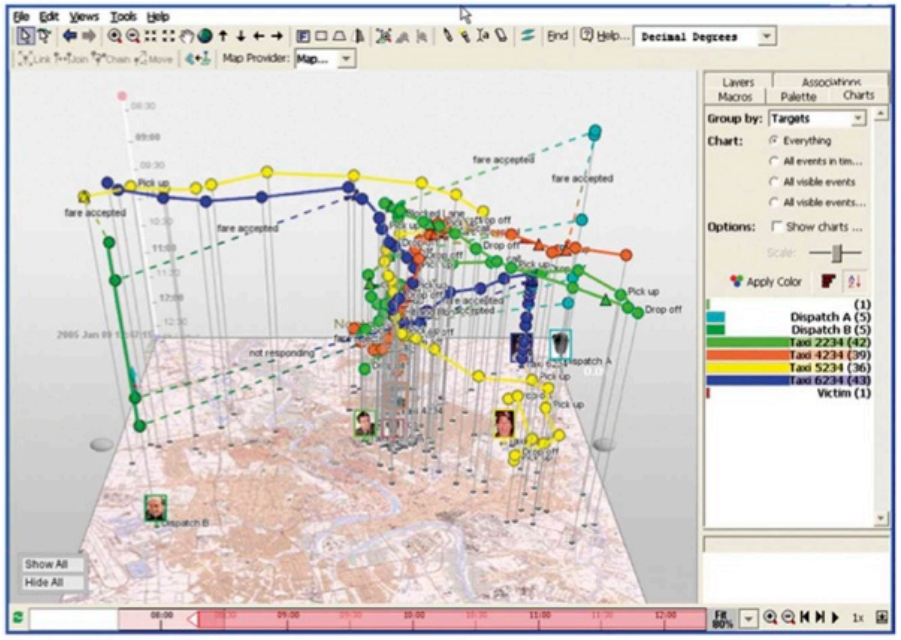


Video: https://www.youtube.com/watch?v=P9-lpE47oWc

Kapler & Wright InfoVis '04

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

Combining Space and Time

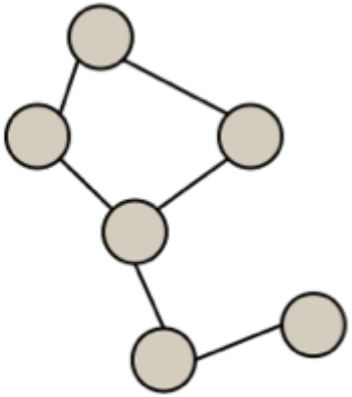


Time	**Temporal primitives**	**Time Points**
	Structure of time	**Linear**
Data	**Frame of reference**	**Spatial**
	Number of variables	**Multivariate**
	Level of abstraction	**Data**
Representation	**Time dependency**	**Static, Dynamic**
	Dimensionality	**3D**

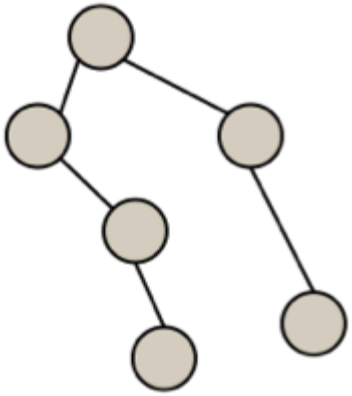
Kapler & Wright InfoVis '04

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.

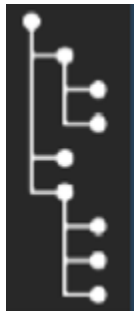
- Often the goal is to effectively depict the graph structure:
  - Connectivity, path-following
  - Network distance
  - Clustering
  - Ordering (e.g., hierarchy level)

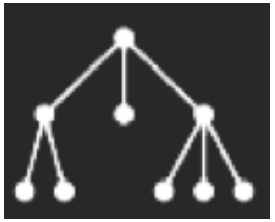
Many Applications

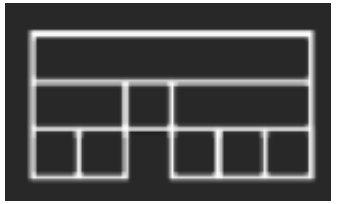
- Tournaments
- Organization Charts
- Biological Interactions (Genes, Proteins)
- Computer Networks
- Social Networks
- Integrated Circuit Design

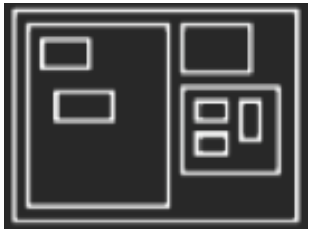
Tree Visualizations

- 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



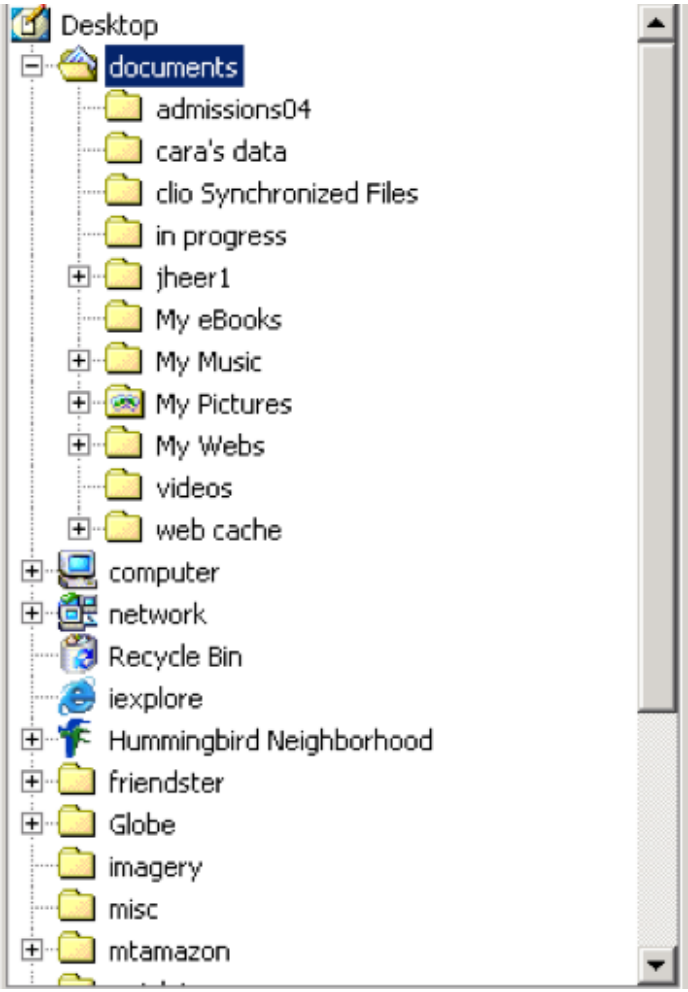






Indented List

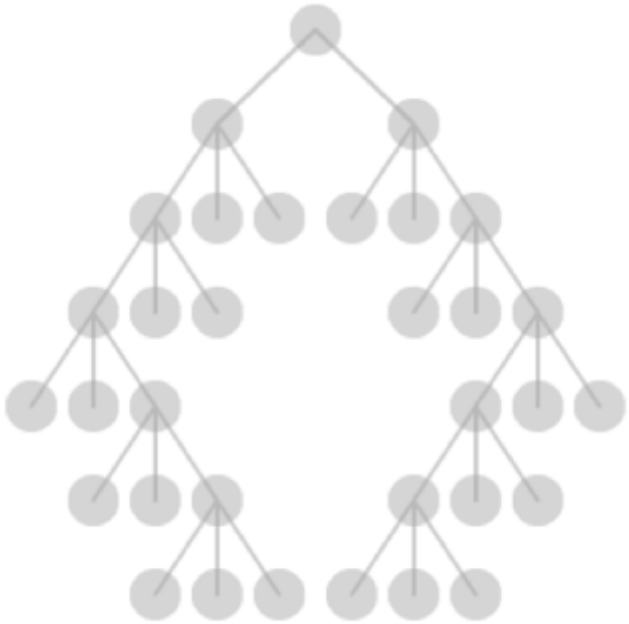
- Places all items along vertically spaced rows
- Indentation used to show parent/child relationships
- Commonly used as a component in an interface
- Breadth and depth contend for space
- Often requires a great deal of scrolling



Node-Link Trees

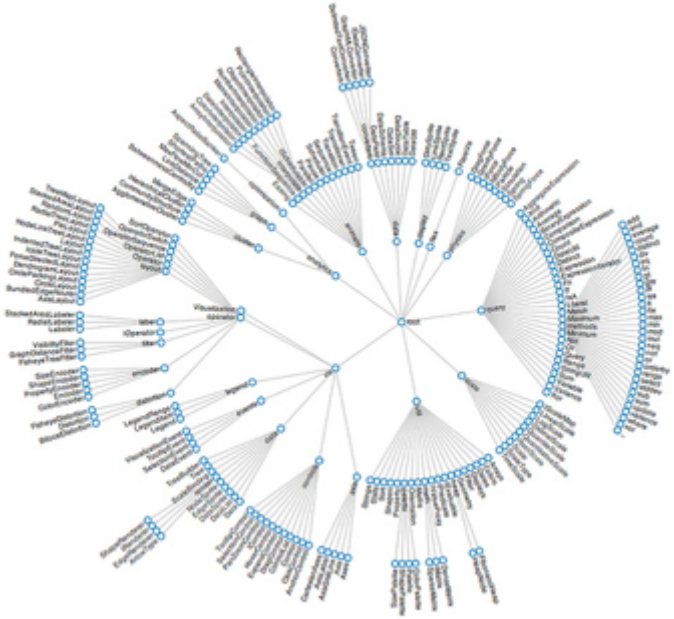
- Nodes are distributed in space, connected by straight or curved lines.
- Typical approach is to use 2D space to break apart breadth and depth.
- Often space is used to communicate hierarchical orientation (e.g., towards authority or generality)

- Reingold-Tilford algorithm can achieve linear time in presenting a compact layout

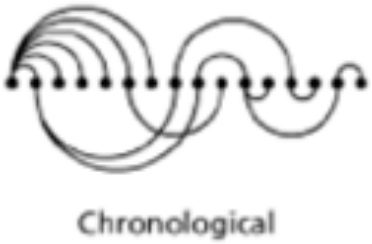


Other Node-Link Trees

- Radial layout places the root in the center.
  - The radius encodes the depth.



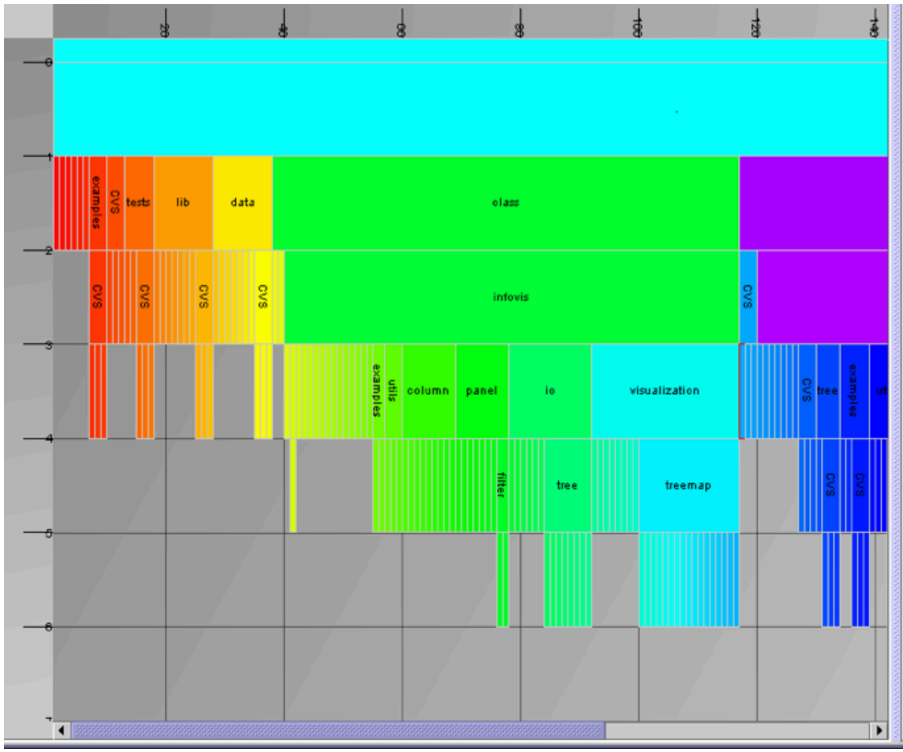
- ThreadArcs
  - combine the chronology of messages with the branching tree structure in a mixed-model visualization



(Optional Reading) Kerr, Bernard. "Thread arcs: An email thread visualization." Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on. IEEE, 2003.

Layered Diagrams

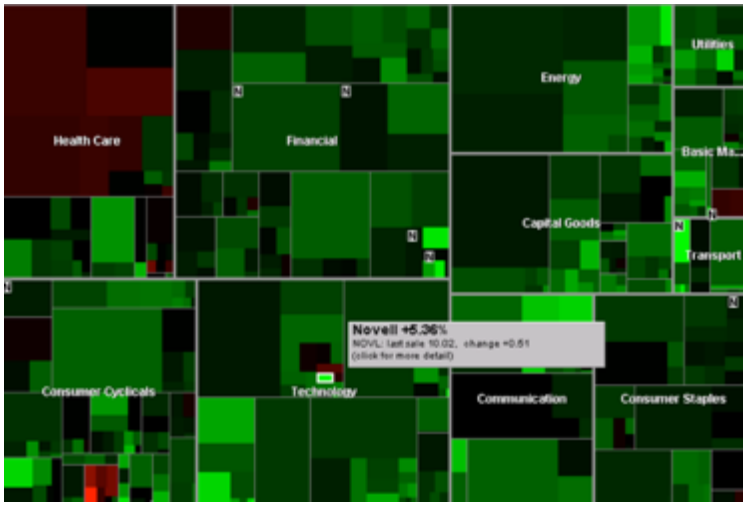
- Signify tree structure using
  - Layering
  - Adjacency
  - Alignment
- Involves recursive sub-division of space
- Higher-level nodes get a larger layer area, whether that is horizontal or angular extent.
- Child levels are layered, constrained to parent's extent

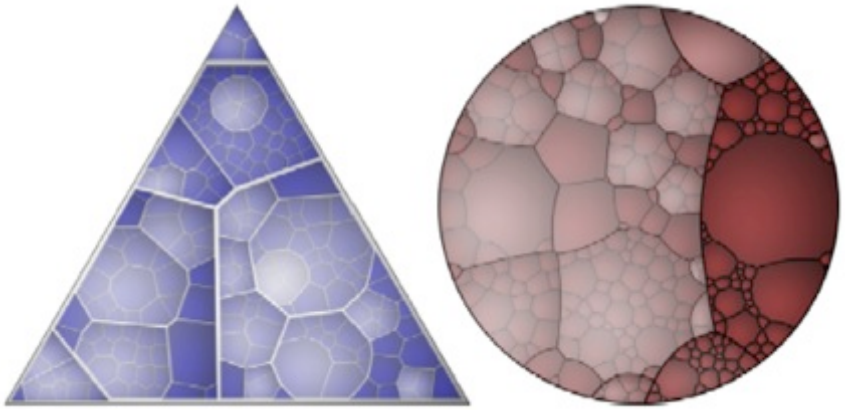


Icicle Trees

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 tesselation.





http://www.cs.umd.edu/hcil/treemap-history/

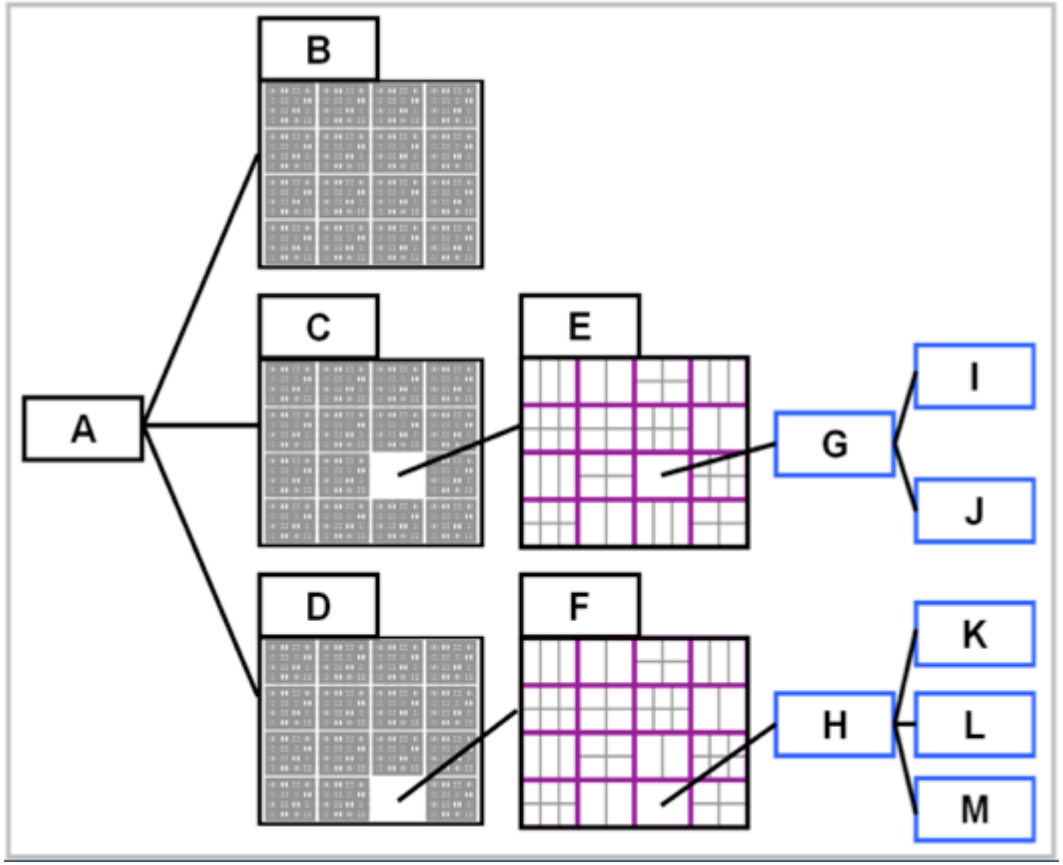
Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Hybrids

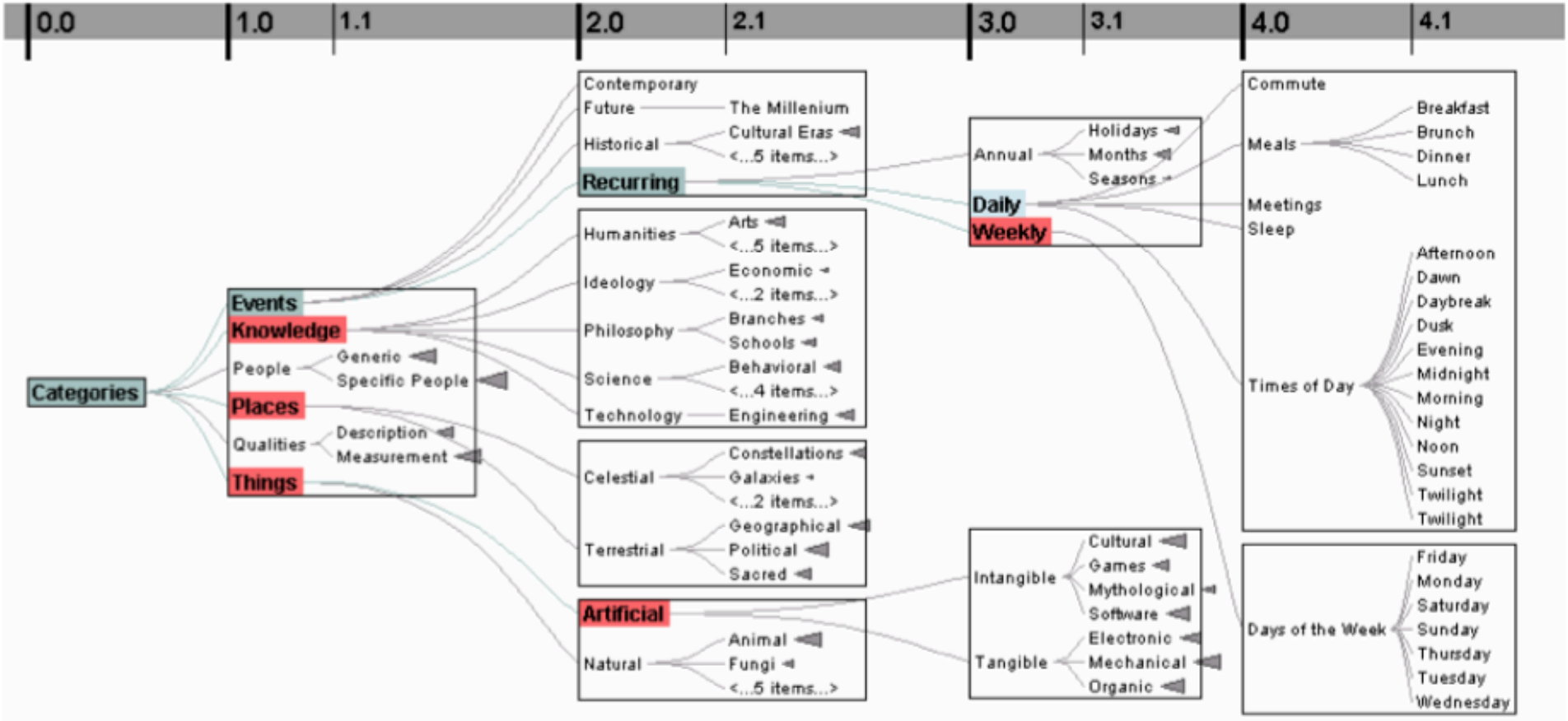
- Elastic Hierarchies
  - Node-link diagram with treemap nodes.

- Video:

https://www.youtube.com/watch?v=nvsIqYQ75yA



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.

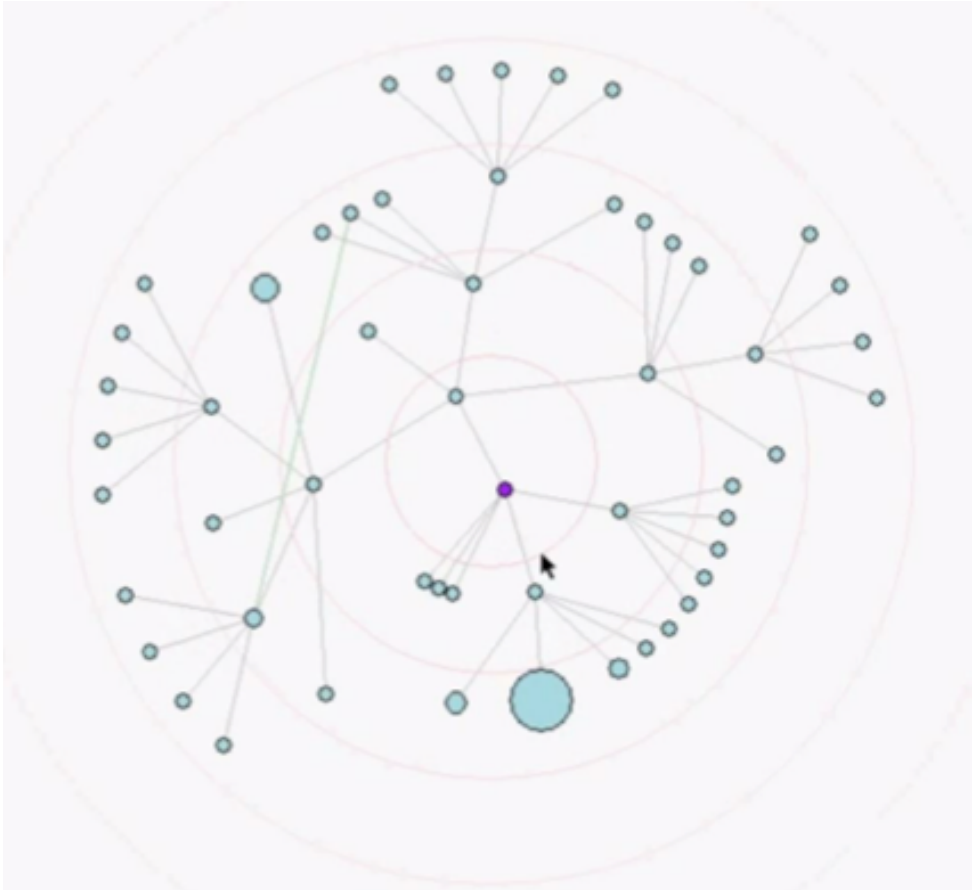
Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

Graph Visualization

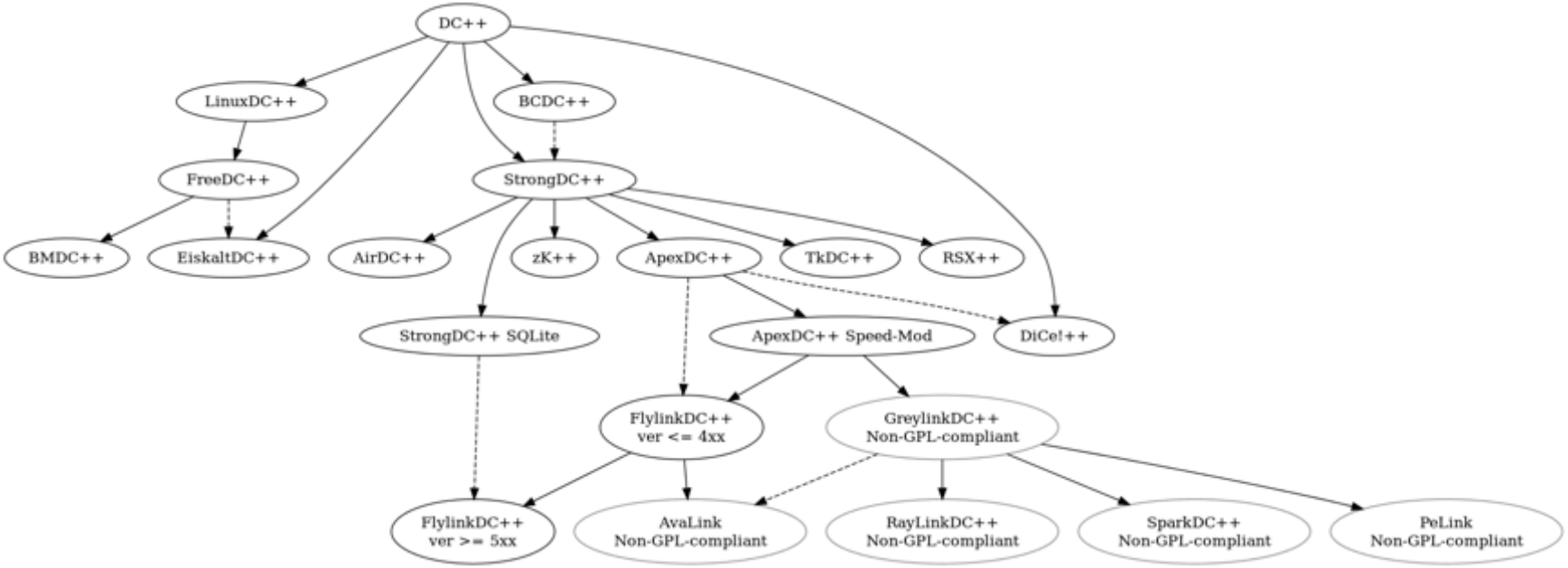
- Two representations:
  - Node-link diagrams
  - Matrices

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



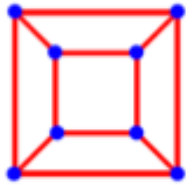
Hierarchical Graph Layout

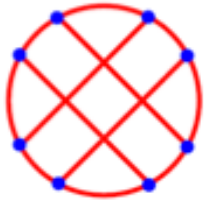


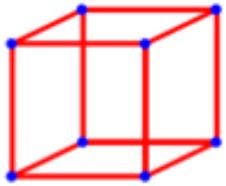
- Evolution of the DC++ tool
- Layered graph drawing
- Layout of a Direct Acyclic Graph
- Hierarchical layering based on descent

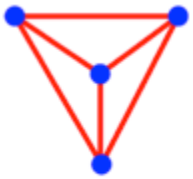
Optimization Techniques

- 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







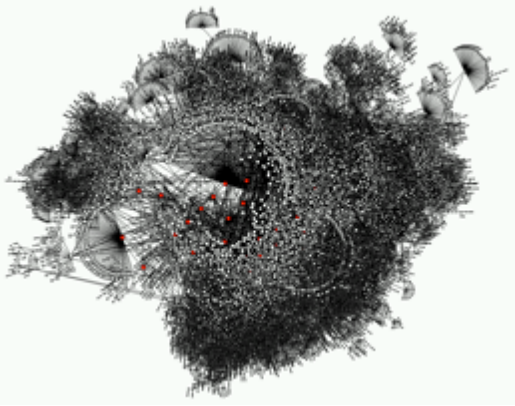


min # crossings

max symmetries

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



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**

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

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

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: =, ≠

  e.g. math, art (course)

- O – Ordered
  - Operations: =, ≠, <, >

  e.g. A, A-, B+, B (grade)

- Q - Interval (location of zero arbitrary)
  - Operations: =, ≠, <, >, -
  - Can measure distances or spans

  e.g. (3.23, -1.2) (GPS)

- Q - Ratio (zero fixed)
  - Operations: =, ≠, <, >, -, %
  - 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.

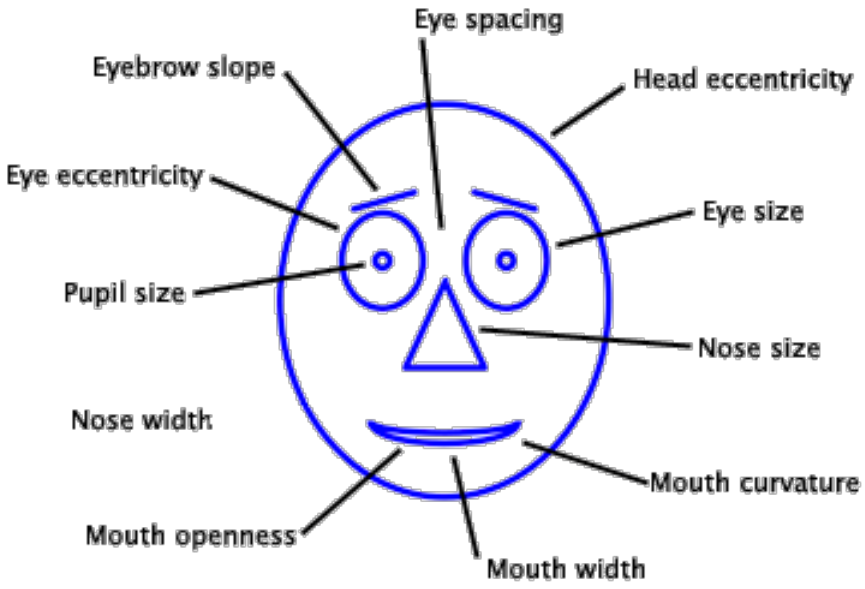
	Student 1	Student 2	Student 3	Student 4
Name (N)	Tom	Jim	Mary	Jane
Age (Q)	20	19	22	21
Grade (O)	A	B	A-	B+
Course (N)	Math	Math	Art	Sport
Entry Year (Q)	1997	1998	1995	1996

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

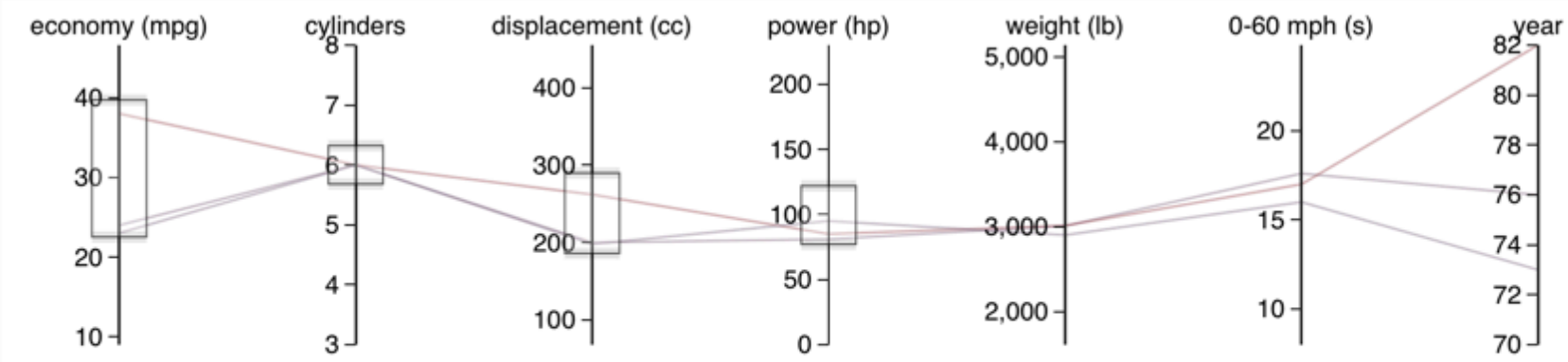
- Observation: We have evolved a sophisticated ability to interpret faces.
- Idea: Encode different variables' values in characteristics of human face



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



name	economy (mpg)	cylinders	displacement (cc)	power (hp)	weight (lb)	0-60 mph (s)	year
Ford Maverick	24	6	200	81	3012	17.6	76
Oldsmobile Cutlass Ciera (Diesel)	38	6	262	85	3015	17	82
Plymouth Duster	23	6	198	95	2904	16	73

A. Inselberg, Multidimensional Detective, Proc. IEEE InfoVis, 1997

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

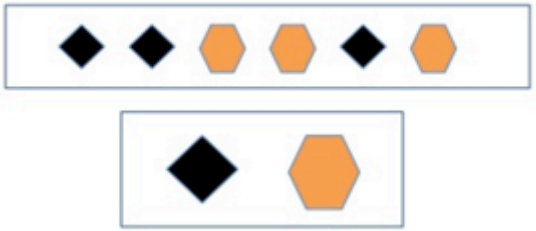
Expected Answer

- 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

- FILTERows 
- 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?
`filter()`	Select a subset of ROWS by conditions
`arrange()`	Reorders ROWS in a data frame
`select()`	Select the COLUMNS of interest
`mutate()`	Create new columns based on existing columns (mutations!)
`summarise()`	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.)

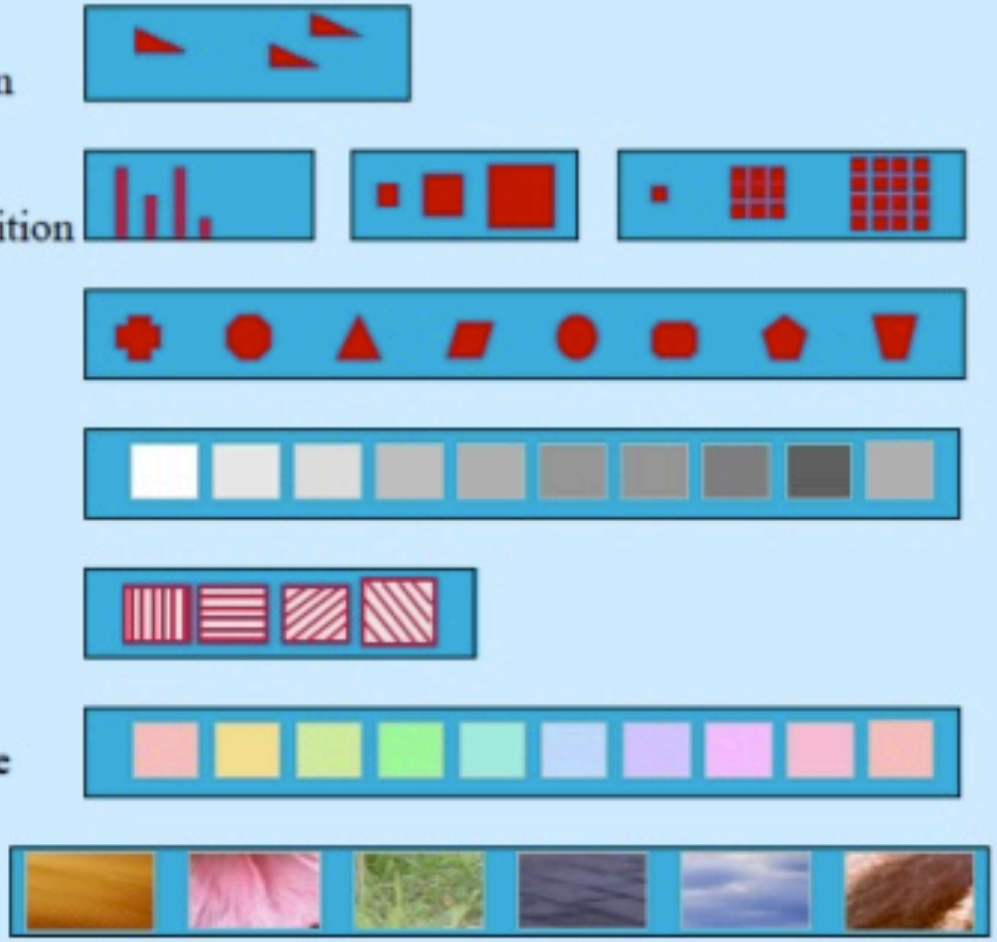
Visualization

- 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

Levels of Organization

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

✓ = Good
~ = OK
✗ = Bad

Expected Answer

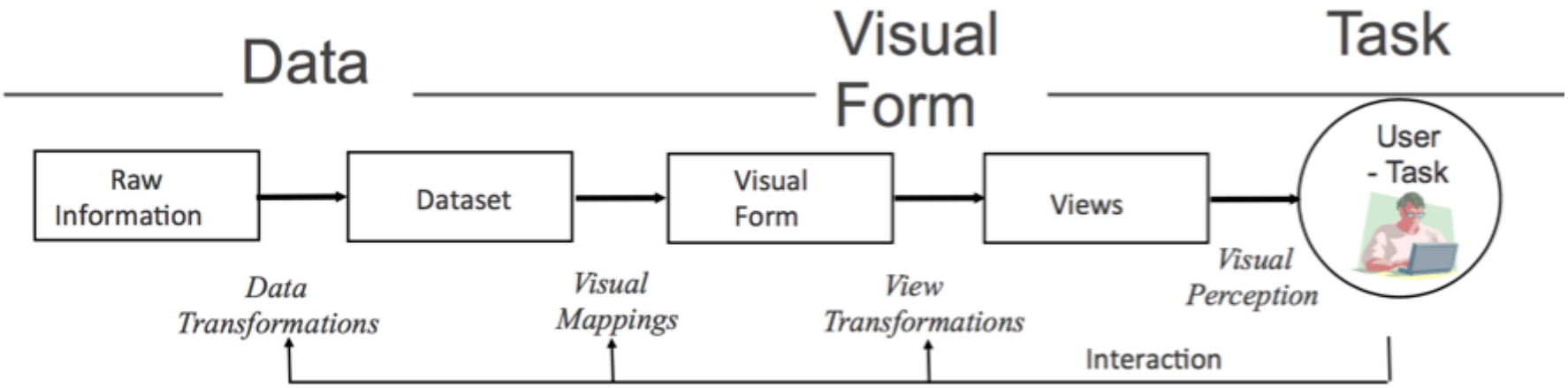
- Two visual encodings are used in this visualization: position and color.
- The x and y positions represent respectively the school (nominal data) and annual salary (quantitative data).
- The color Hue demonstrates different gender, which is of nominal data type.

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.

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