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=96914

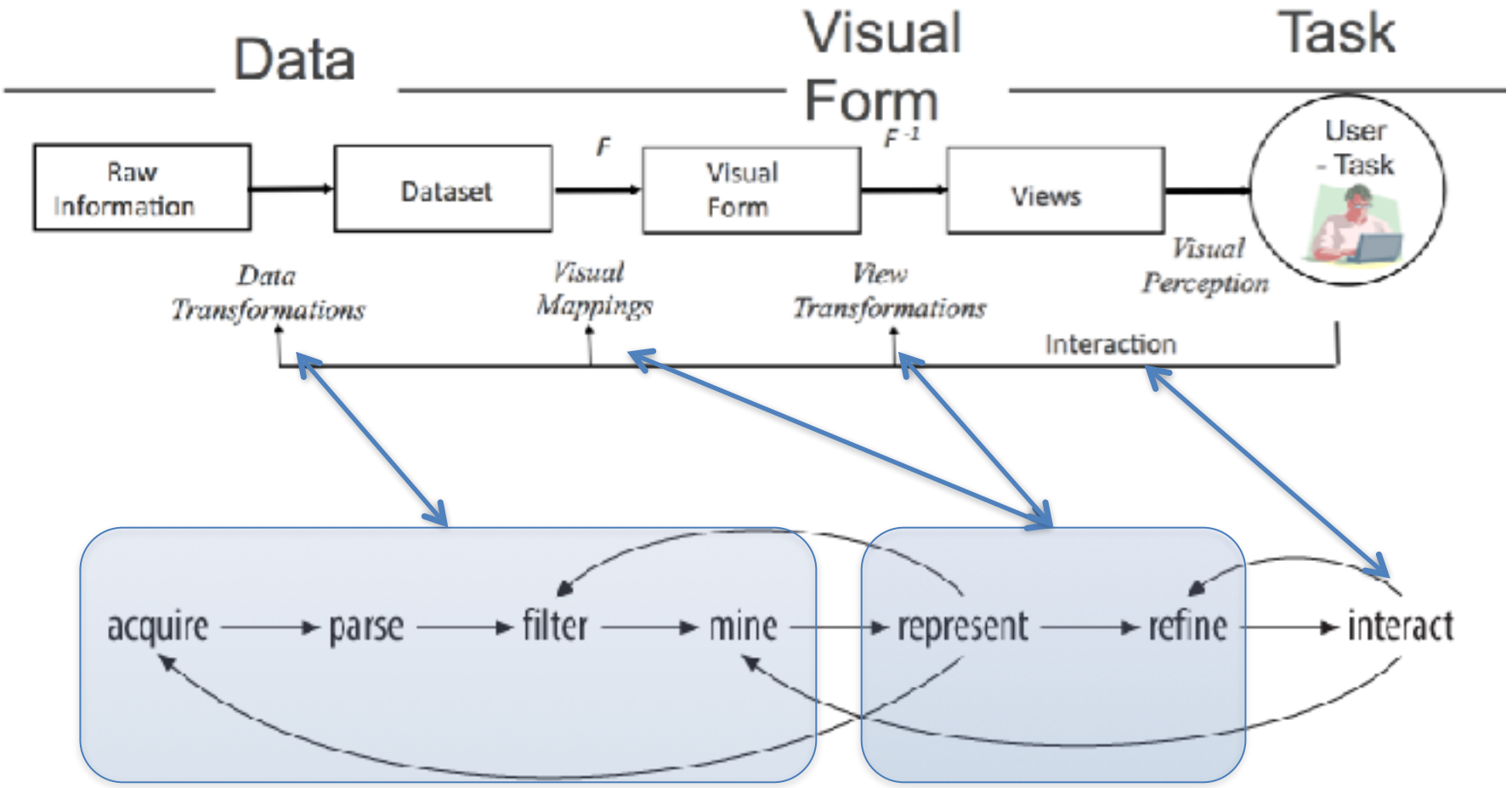
Last Lecture

The Value of Visualization

Key Values of Visualizations

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

Different Stages of Visualization



Visualizing Data by Ben Fry

Overview

- How to process **data**?
  - Data models
  - Processing algorithms

- How to encode the data using **images** (the visual channel)?
  - Visual encoding (mapping)

Administrivia: Module Expectation

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

Administrivia: G53FIV Coursework

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

  - You should work closely with the “R Graphics Cookbook”.

Administrivia: G53FIV Coursework

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

Administrivia: G53IVP Project

- Goal: hands-on experience in designing, implementing, and evaluating a new visualization method, algorithm or tool.
- Some examples*:
  - http://courses.ischool.berkeley.edu/i247/s16/

- A written report
  - Introduction
  - Related work
  - Methods/Design (storyboard, etc.)
  - Results (Visualizations)
  - Evaluation (user study)
  - Discussions
  - Conclusions

- Demo
  - A poster covers the main visualizations
  - A presentation

- Code repositories

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

G53IVP First Meeting

- First meeting: Today at 15:00 or 17:00
- A25 Business South
- Discuss the general format and available resources

- Next: Proposal development
  - Feb 28th 11:00

Data

Data Models

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

Example (Data Table)

cases

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

variables

Taxonomy of Data Types

- 1D (sets and sequences)
- 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)

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)

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)

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

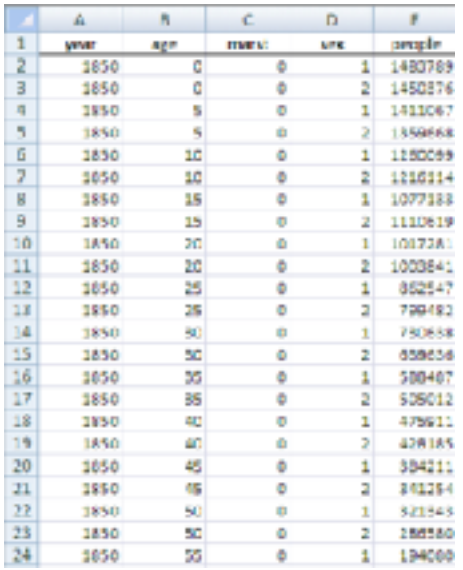
Exercises

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

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

Census data

Exercises



	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1480789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017381
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	750618
15	1850	30	0	2	655636
16	1850	35	0	1	580487
17	1850	35	0	2	505012
18	1850	40	0	1	479611
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341284
22	1850	50	0	1	321545
23	1850	50	0	2	286580
24	1850	55	0	1	194089

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

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

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

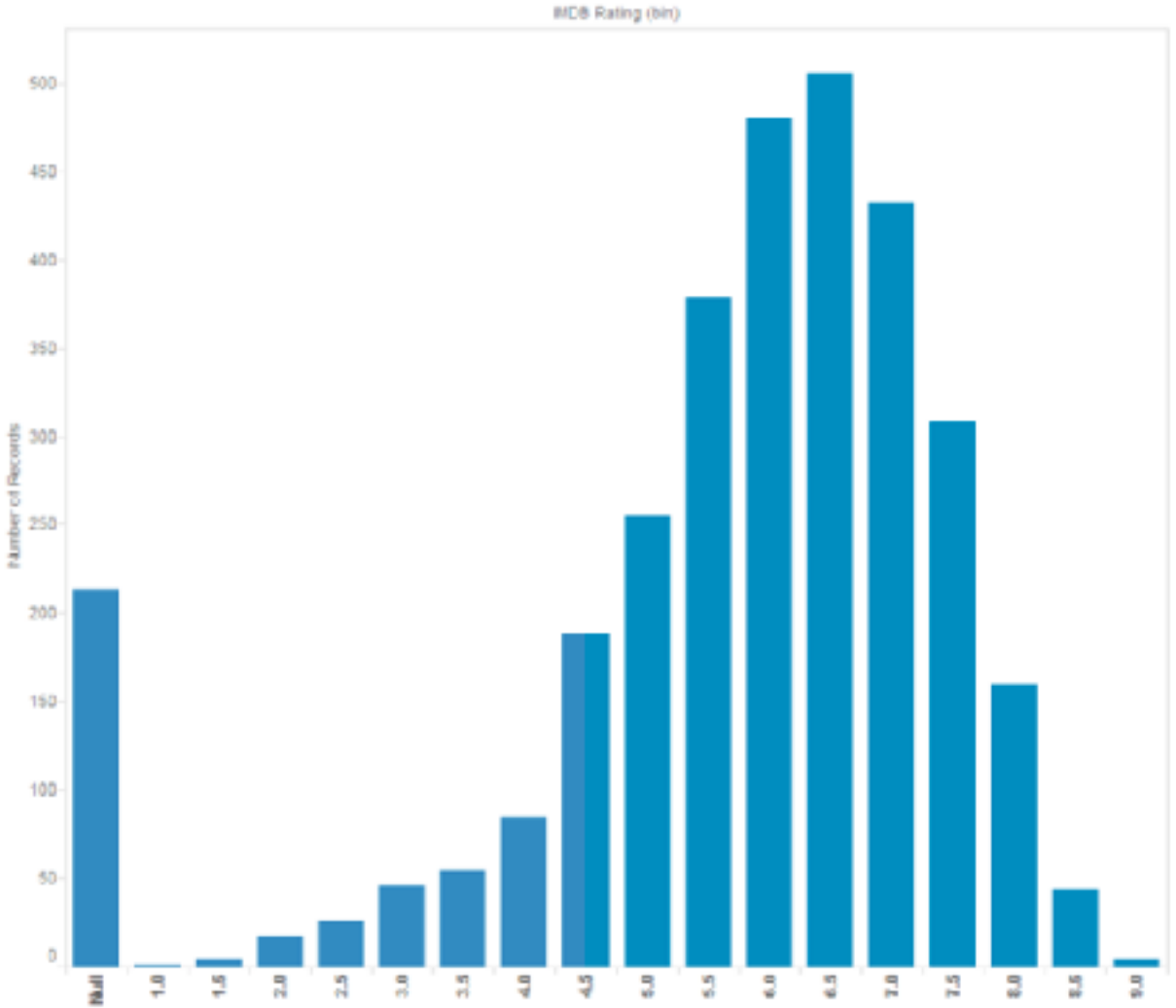
Data Cleaning and Filtering

- 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!

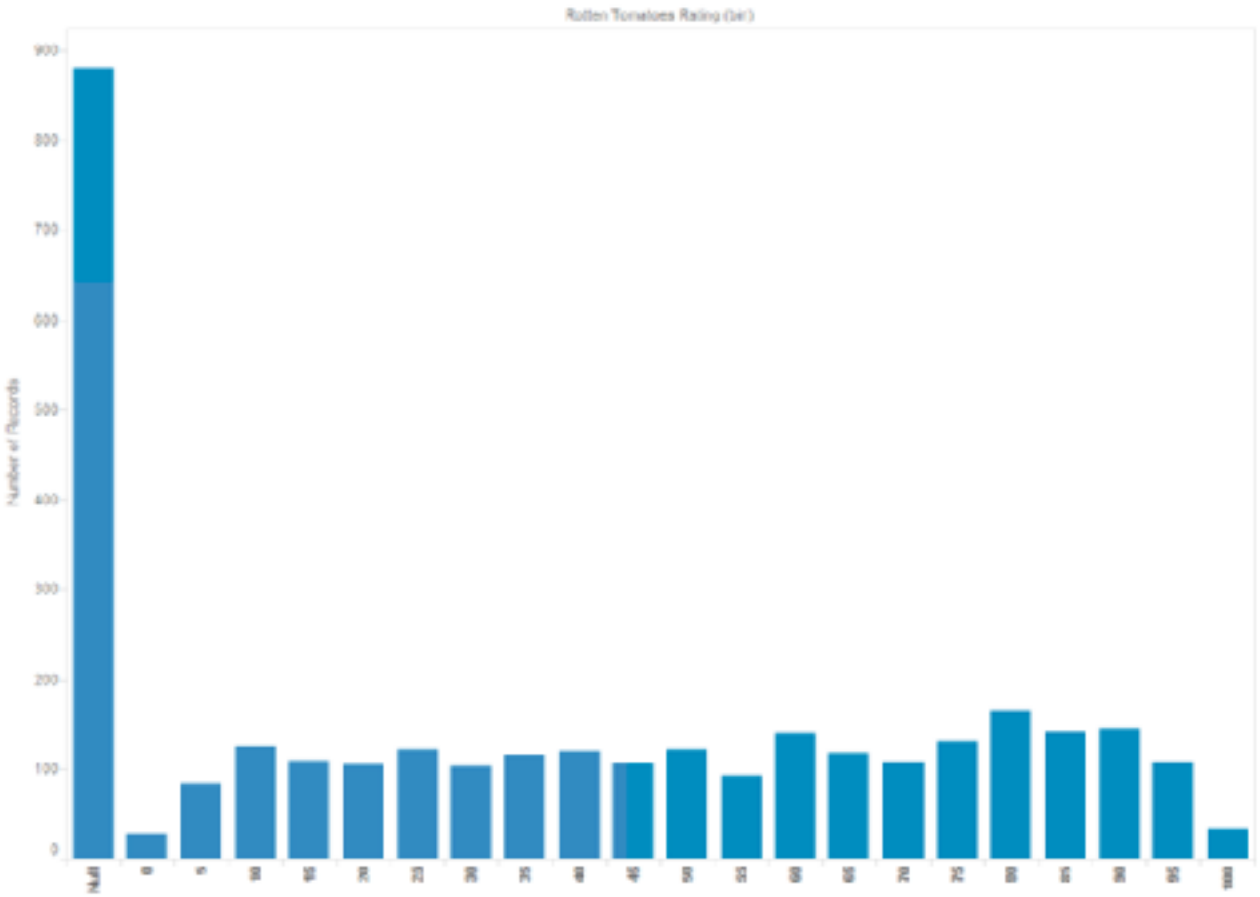
An Example

- Movie rating data
  - IMDB ratings



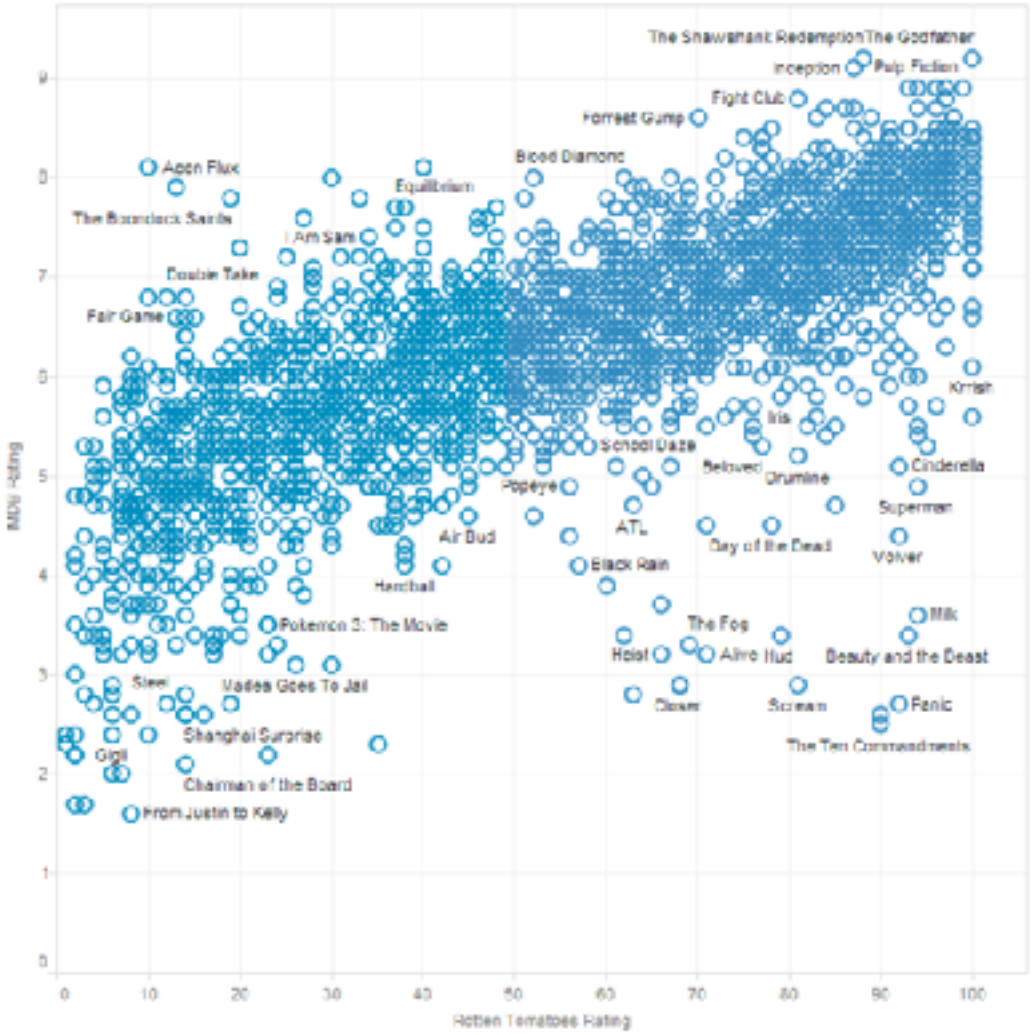
An Example

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



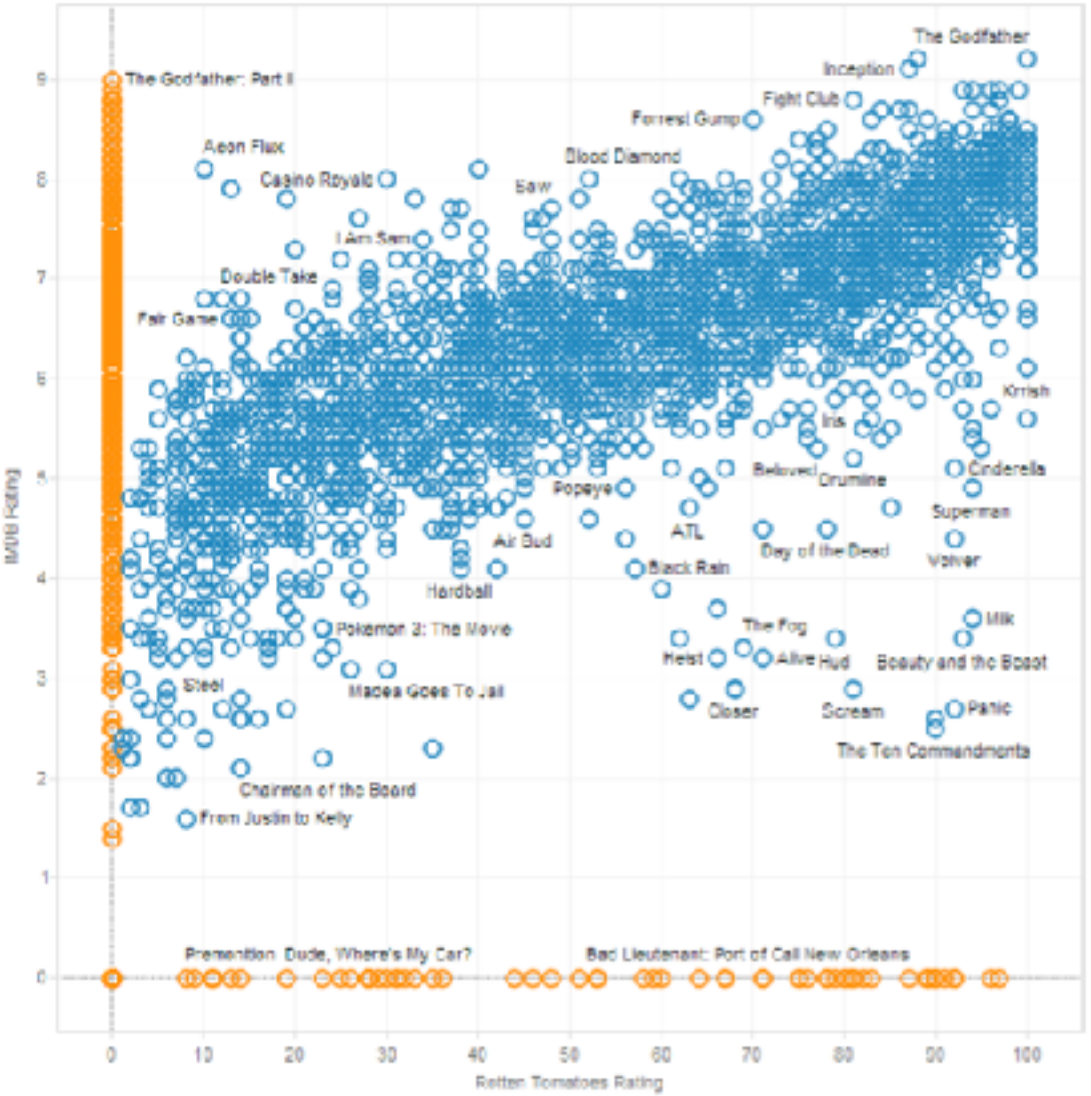
An Example

- Movie rating data scatter plot



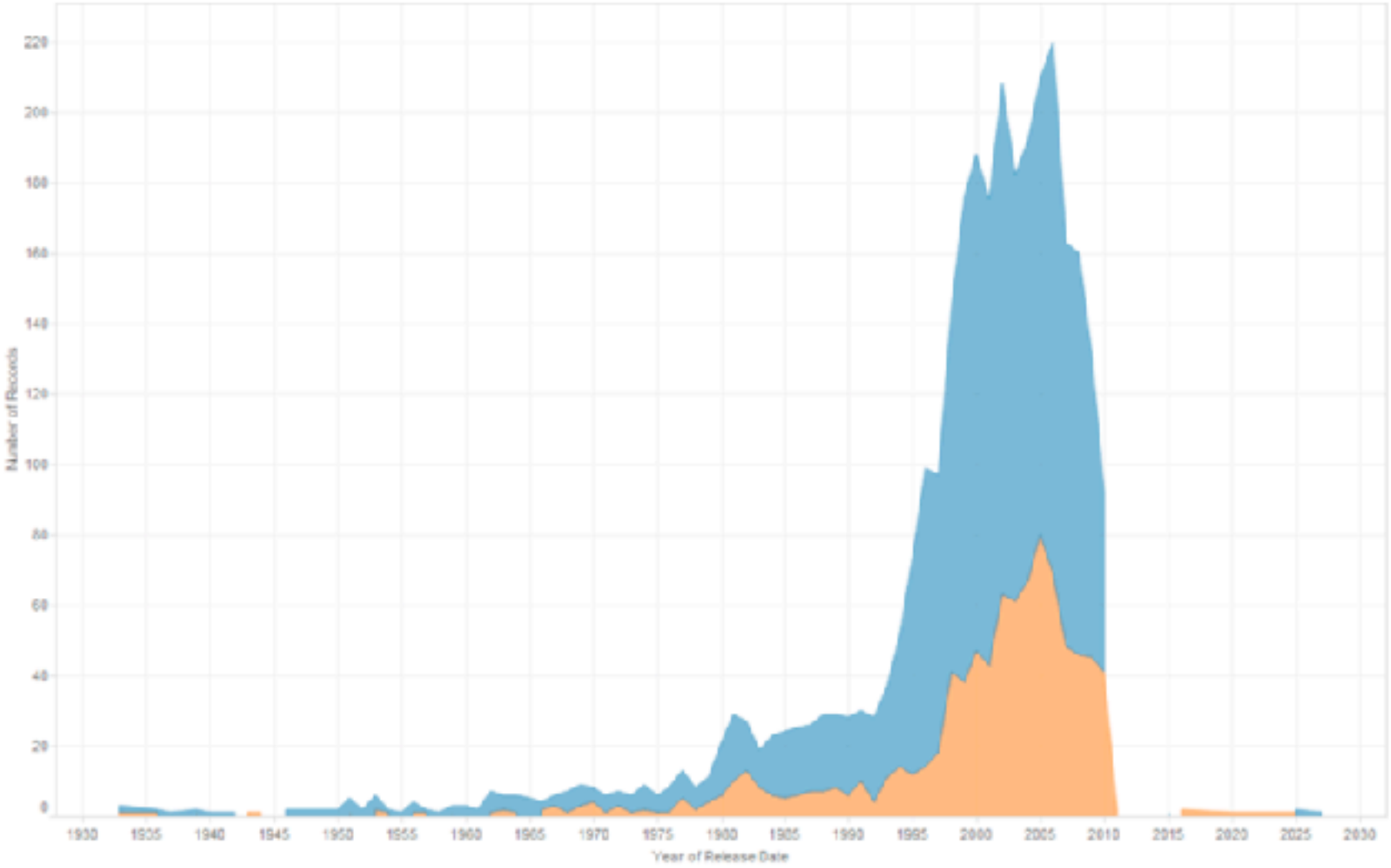
An Example

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



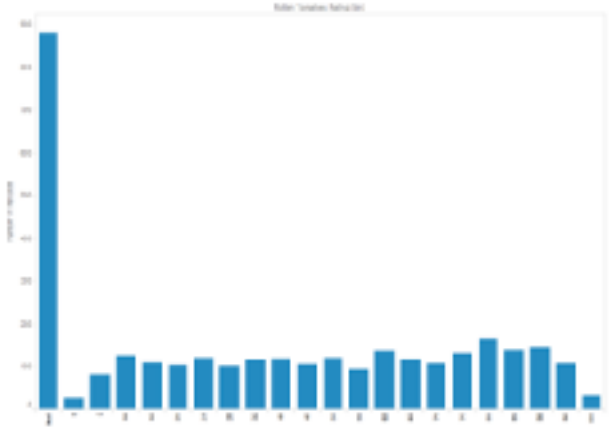
Dr. Ke Zhou (http://www.cs.nott.ac.uk/~pszkz/)

An Example


Number of Records
220
200
180
160
140
120
100
80
60
40
20
0
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date

Data Cleaning and Filtering

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

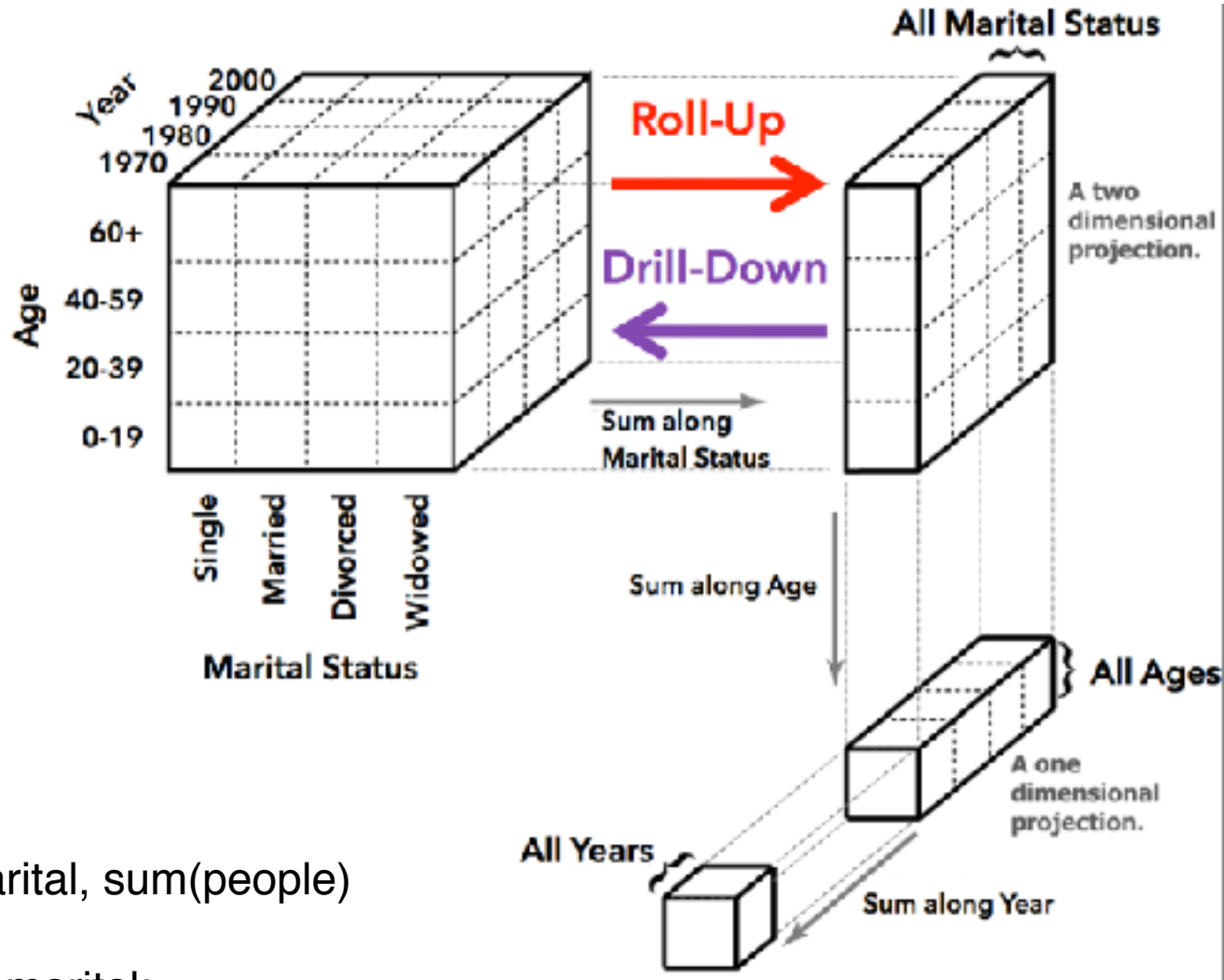


Data Adjustment: Relational Algebra

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

Data Adjustment

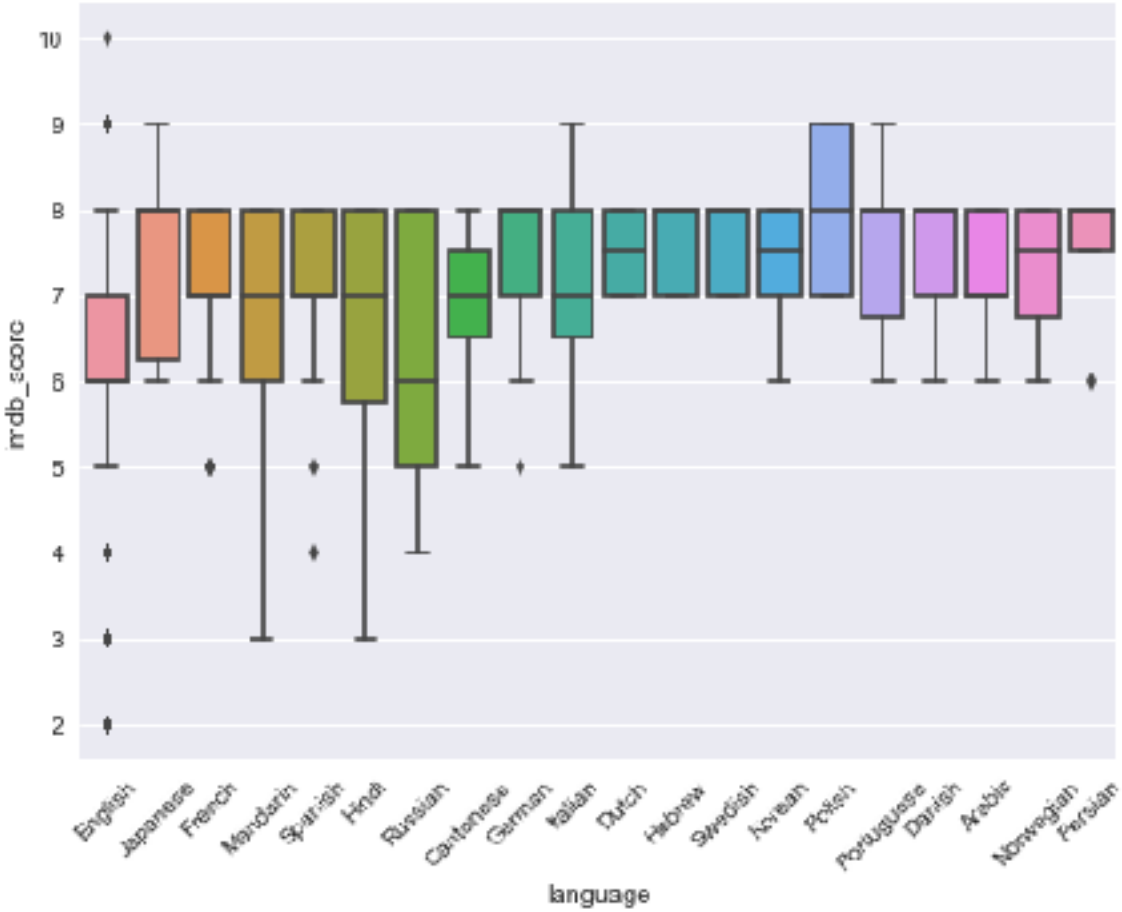
- Roll-up
- Drill-down




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

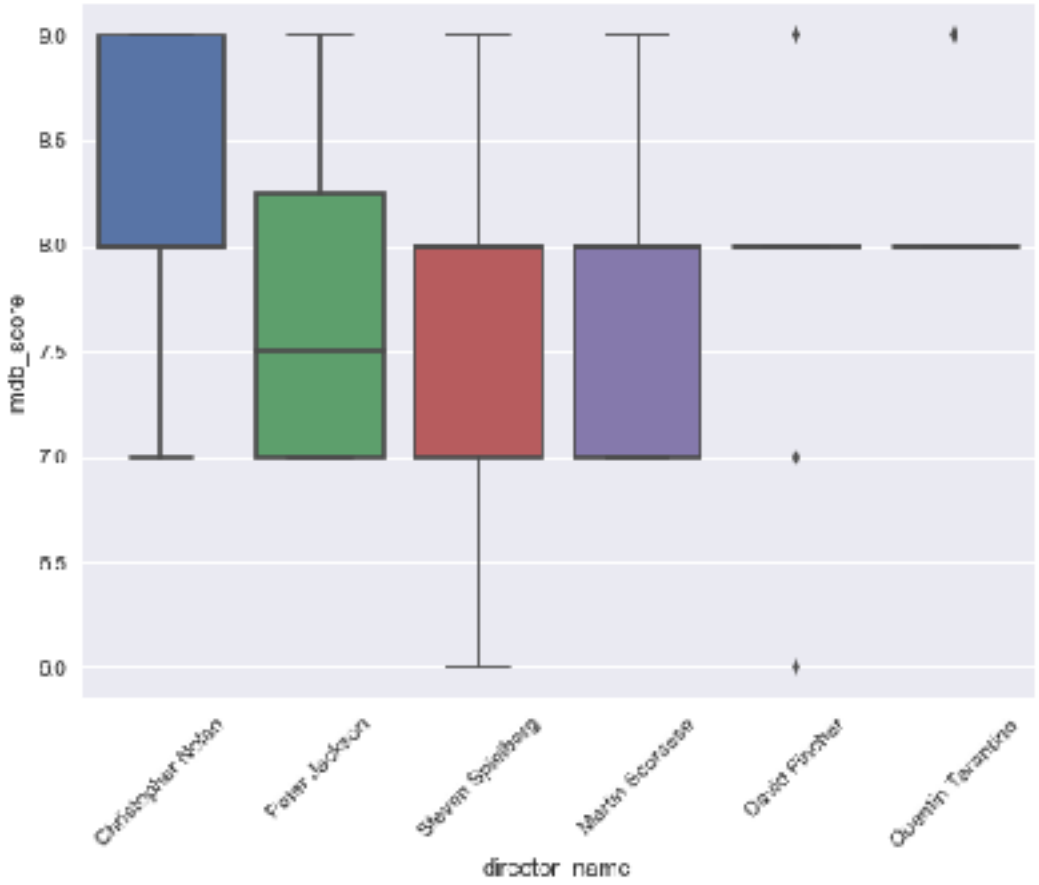
An Example

- IMDB movie rating by language



An Example

- IMDB movie rating by director

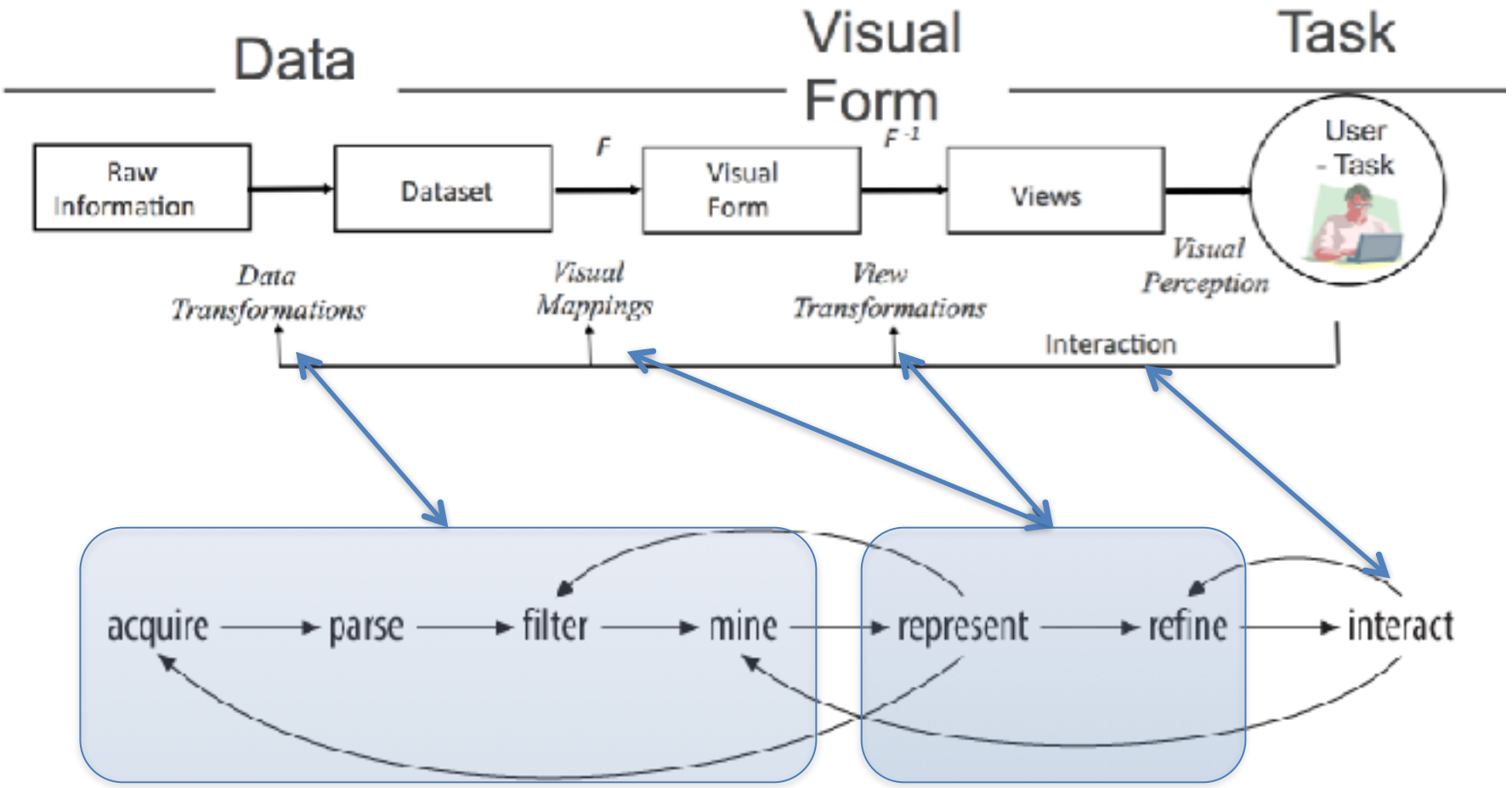


Data Adjustment

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

Image

Different Stages of Visualization



Visualizing Data by Ben Fry

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

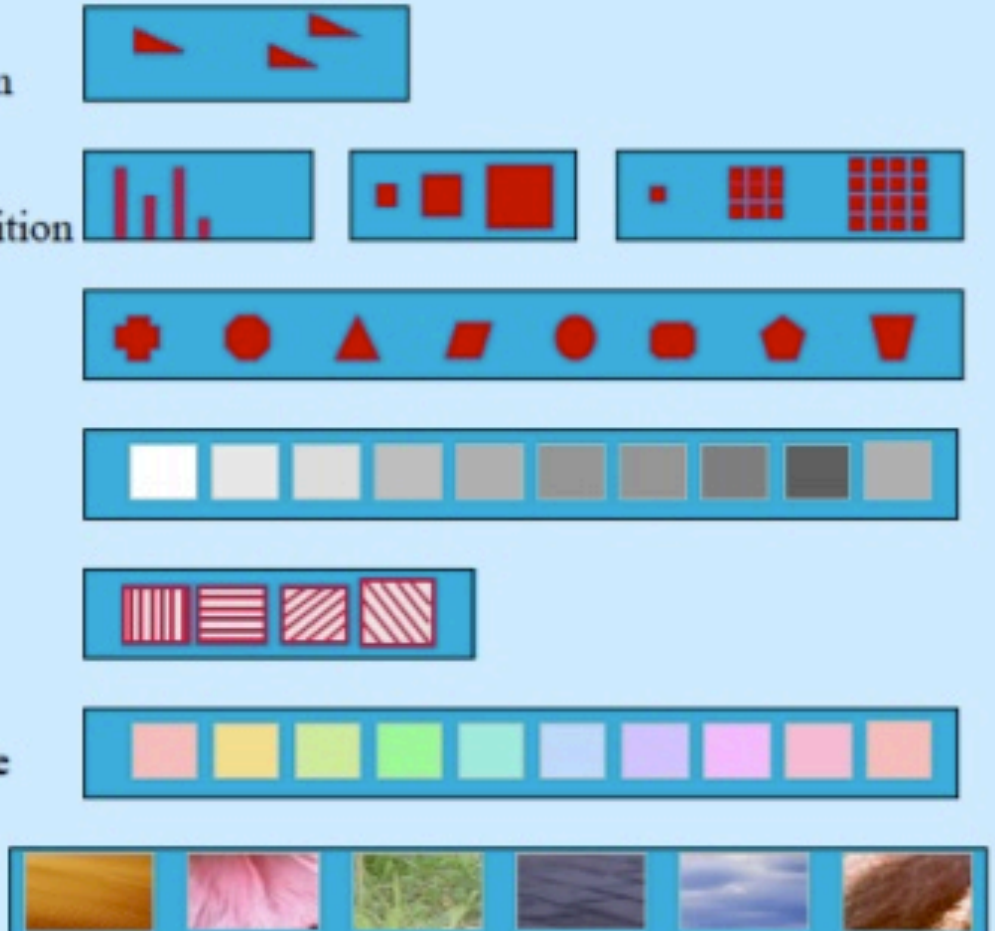
- "Resemblance, order and proportion are the three sign fields in graphics."

  – 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



Information in Hue and Value

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

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

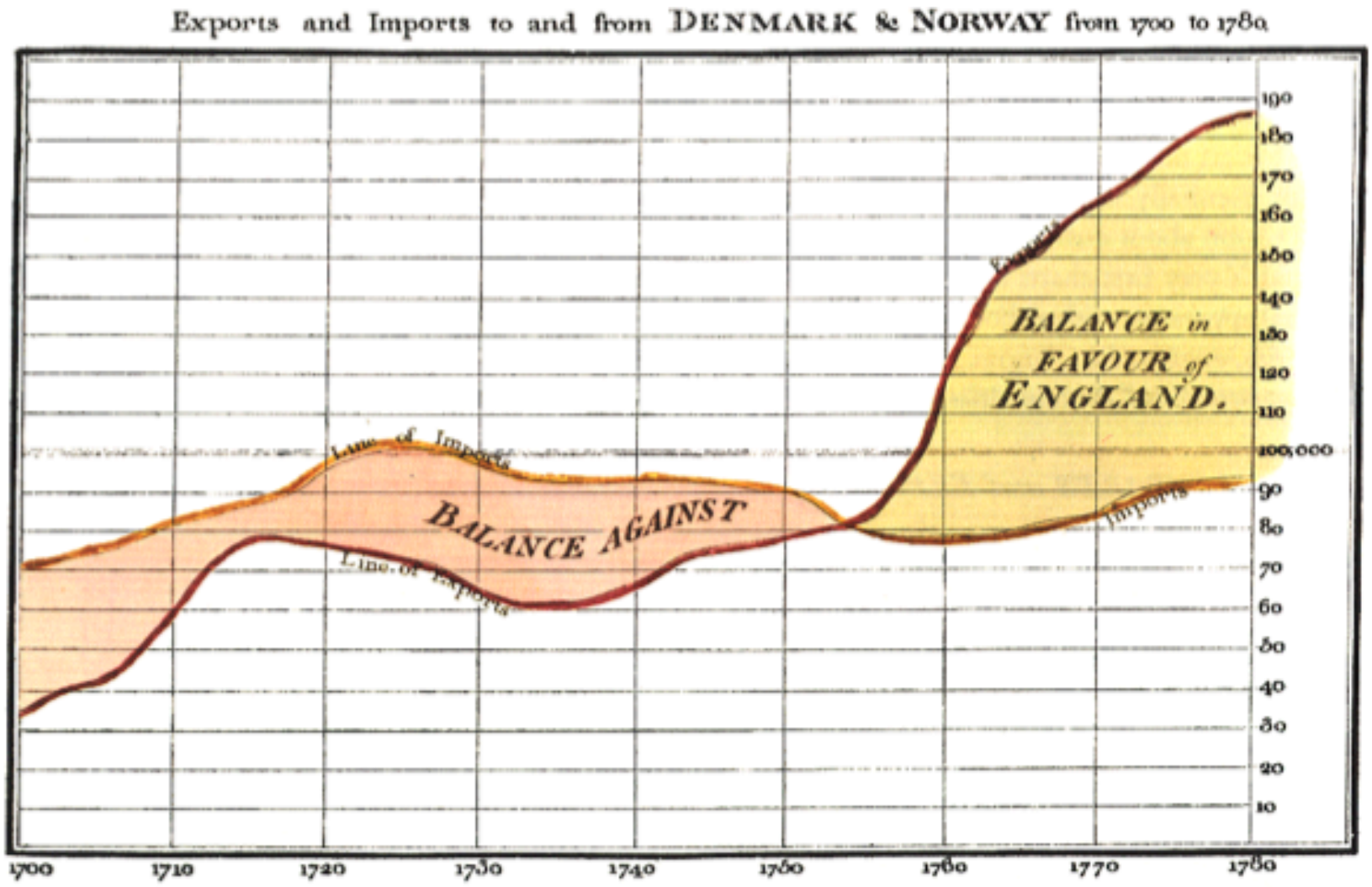
- Hue is normally perceived as unordered
  - Encode nominal variables (N) using color

Bertin's Levels of Organization

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

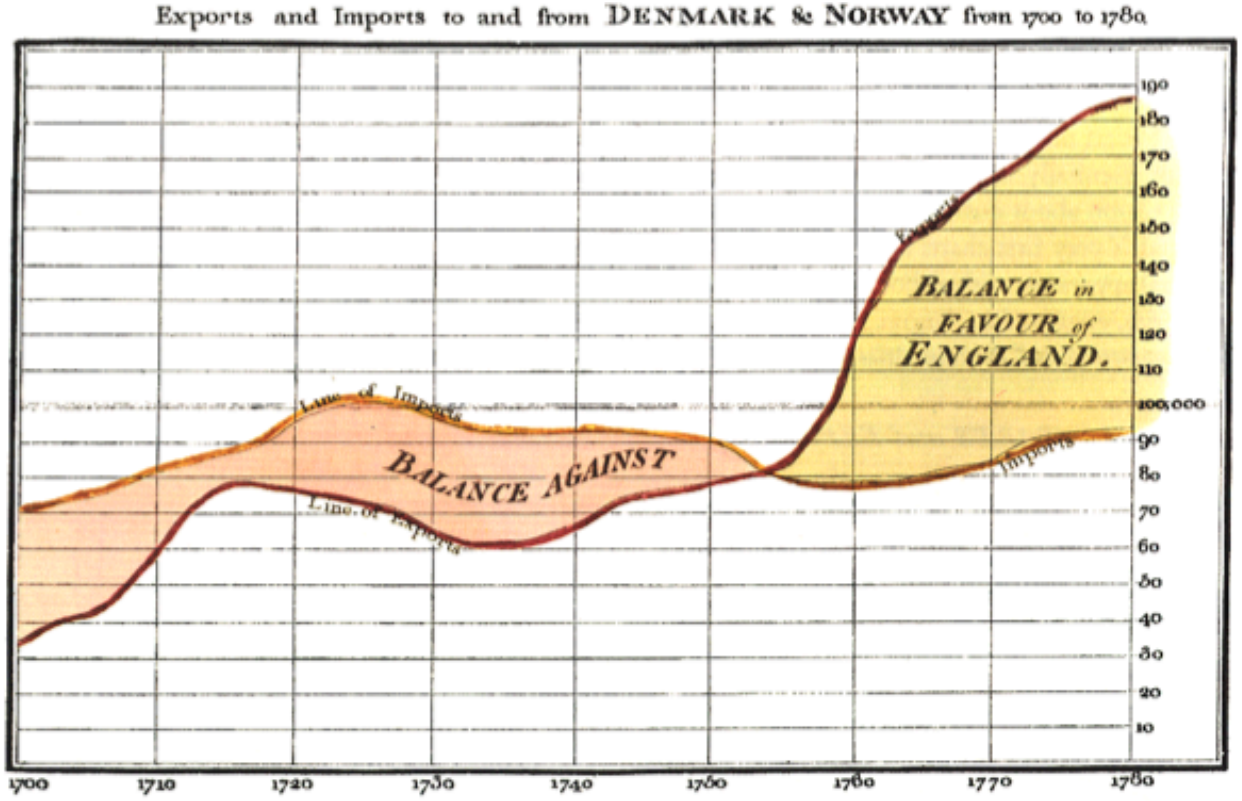
✓ = Good
~ = OK
✗ = Bad

Examples



The Visual Display of Quantitative Information (Textbook). Tufte.

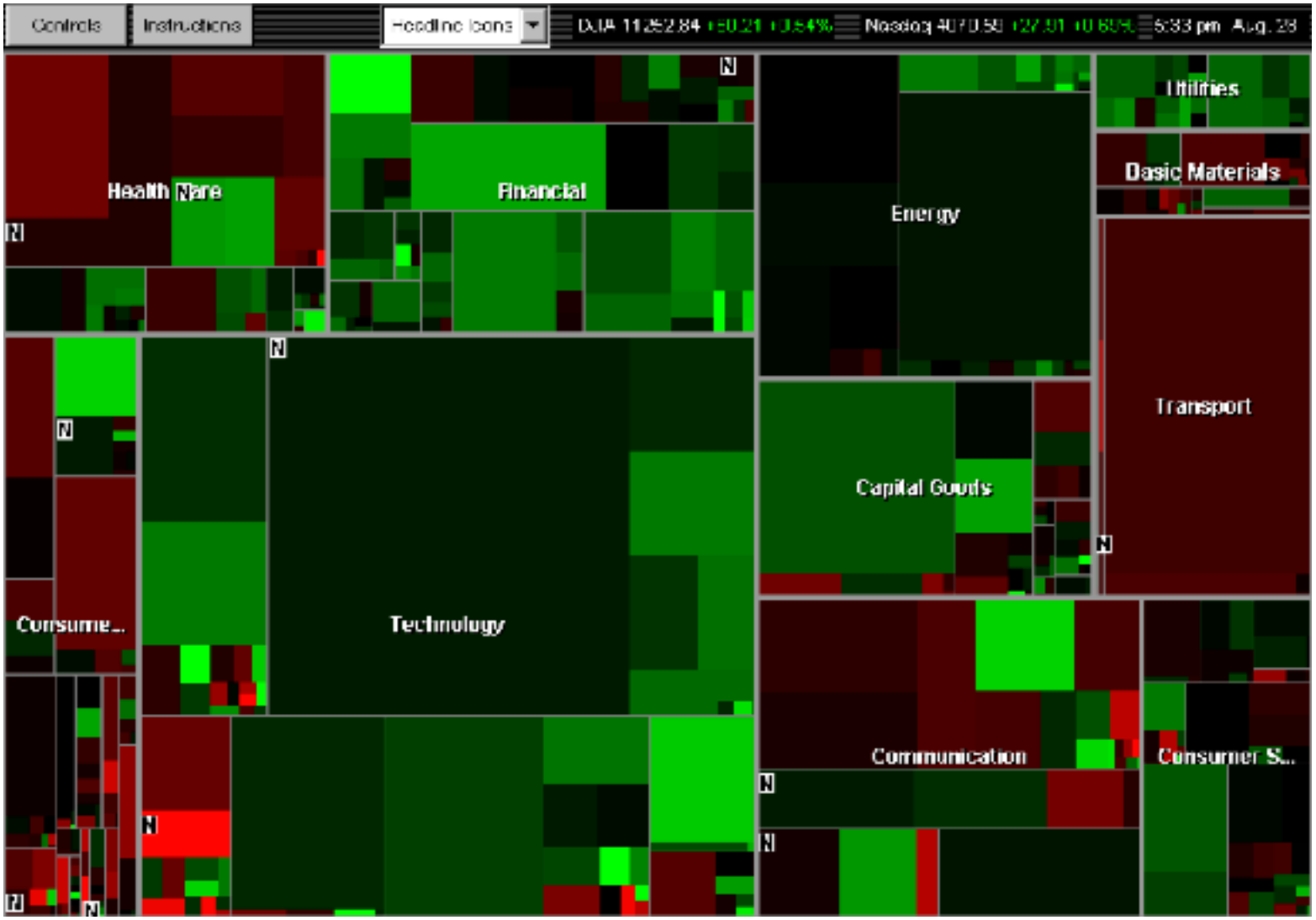
Examples



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

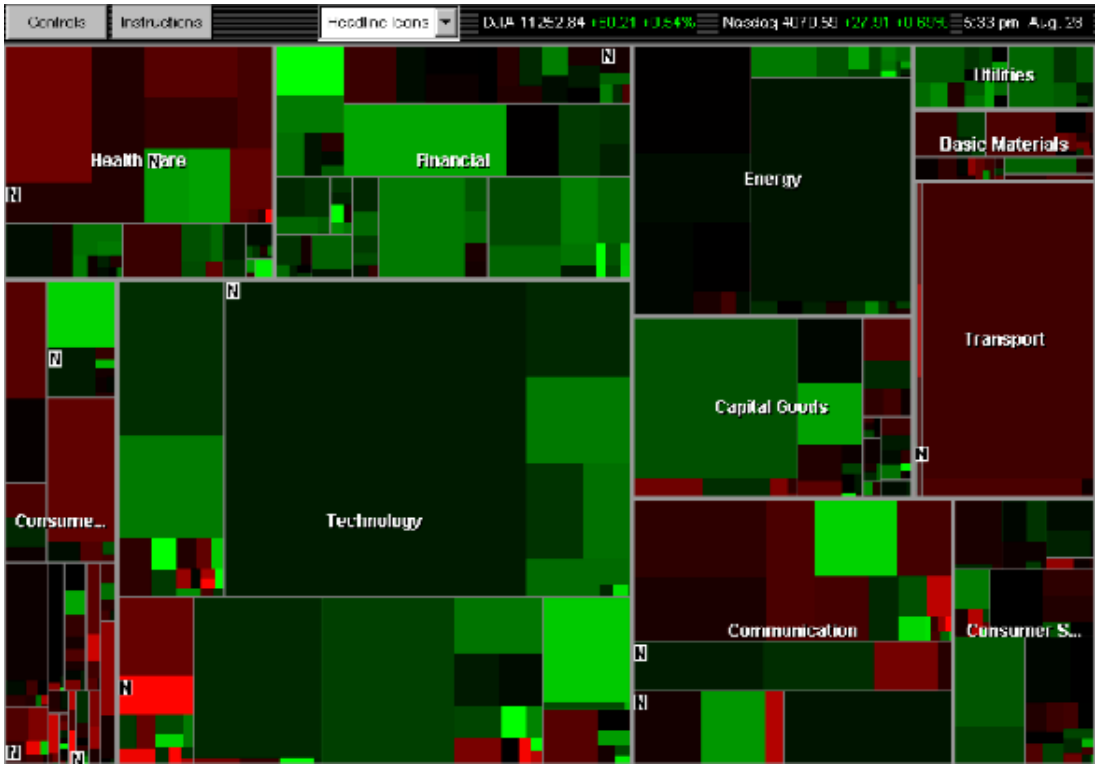
The Visual Display of Quantitative Information (Textbook). Tufte.

Examples



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

Examples



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

How do we choose visual encodings?

What design criteria should we follow?

Next Lecture

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