CSI3005

Advanced Data Visualization Techniques

IBM Predicts Demand For Data Scientists Will Soar 28% By 2020

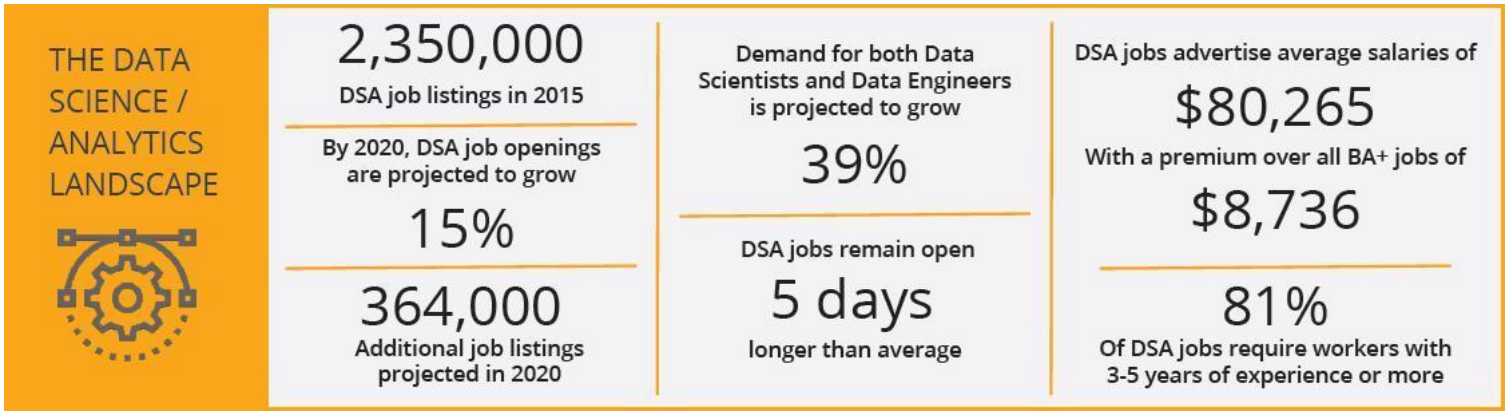
Louis Columbus, CONTRIBUTOR
FULL BIO
Opinions expressed by Forbes Contributors are their own.

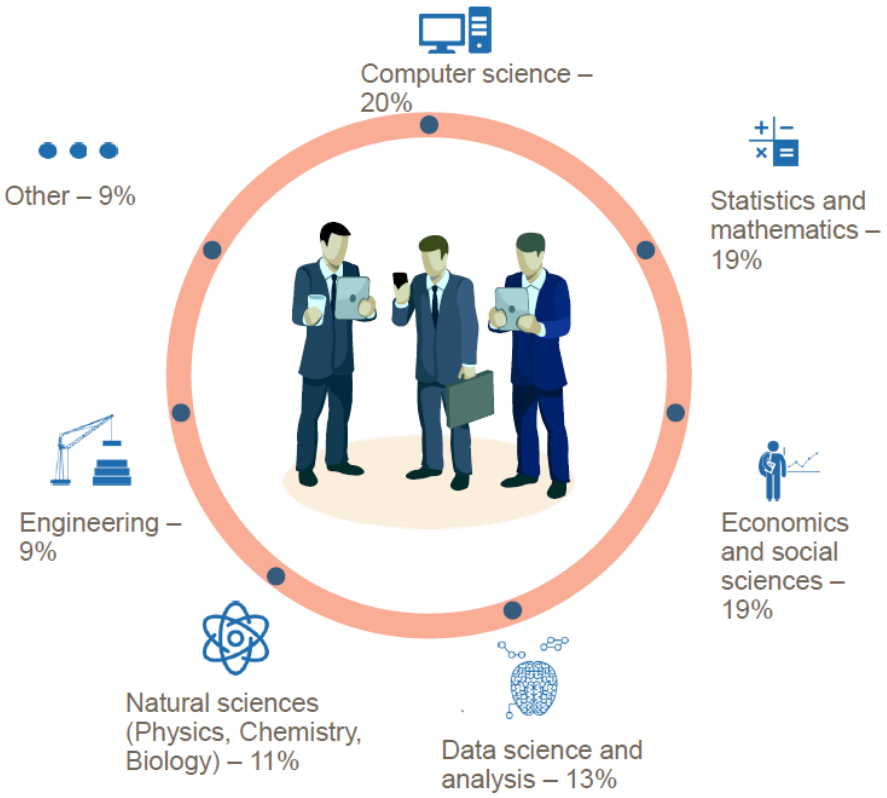
- Jobs requiring machine learning skills are paying an average of $114,000. Advertised data scientist jobs pay an average of $105,000 and advertised data engineering jobs pay an average of $117,000.
- 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.

By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM.

- **59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.** DSA jobs factor most prominently in the Finance and Insurance industry, where they account for 19% of all openings. The Professional Services and IT industries follow with 18% and 17% relative demand for DSA jobs, respectively. The following graphic provides an analysis of DSA job category demand by industry.

- **By 2020 the number of Data Science and Analytics job listings is projected to grow by nearly 364,000 listings to approximately 2,720,000** The following summary graphic from the study highlights how in-demand data science and analytics skill sets are today and are projected to be through 2020.





There is a lot of diversity in terms of the academic backgrounds of current data scientists.

The sample, for example, is richly populated with people coming from **Computer Science (20%), Statistics and Mathematics (19%)**, and Economics and Social Sciences (19%).

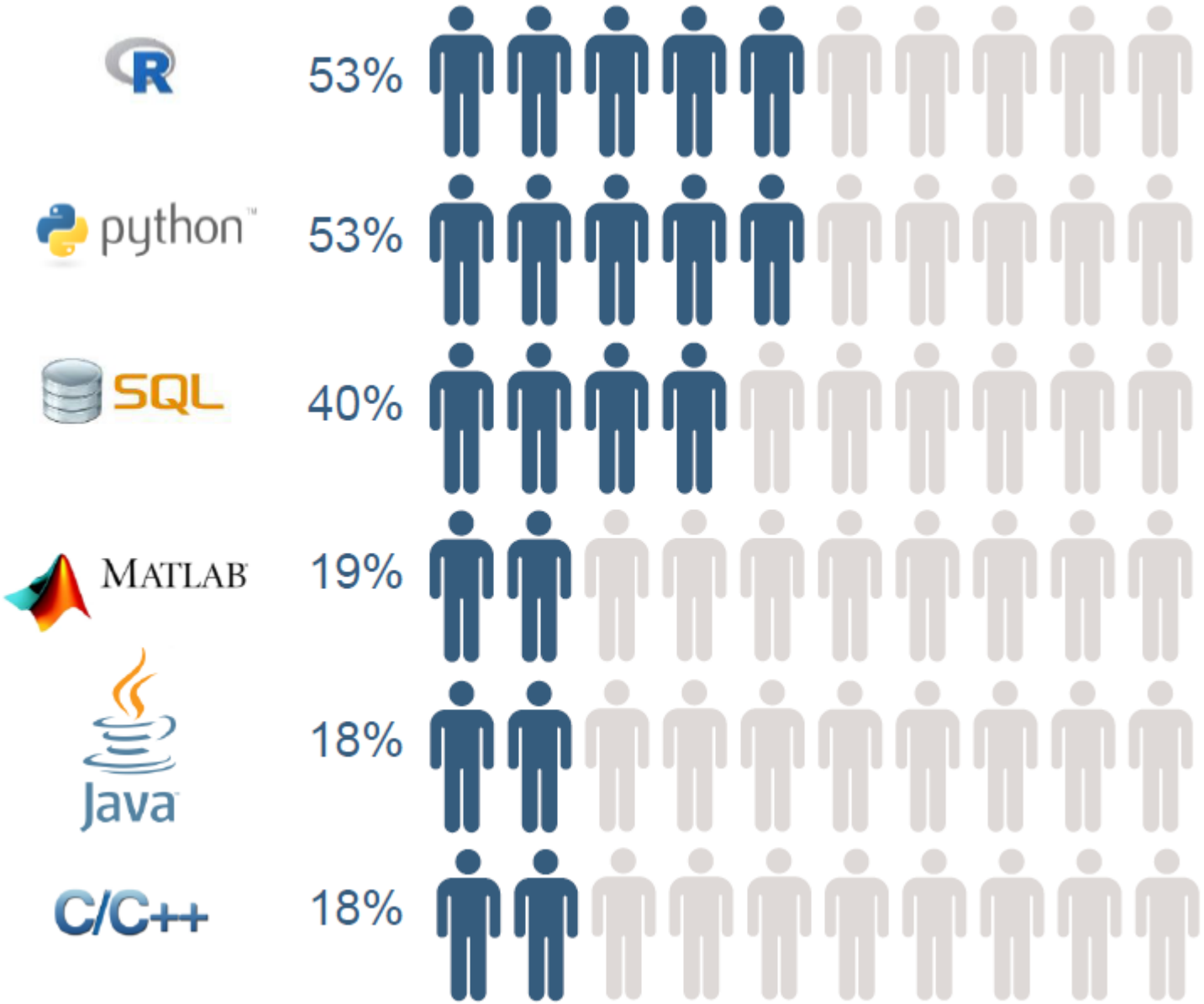

R
53%
python
53%
SQL
40%
MATLAB
19%
Java
18%
C/C++
18%

Table 10. Highest Paying Analytical Skills (with at Least 7,500 Postings)

Skill Name	Average Salary
MapReduce	$115,907
PIG	$114,474
Machine Learning	$112,732
Apache Hive	$112,242
Apache Hadoop	$110,562
Big Data	$109,895
Data Science	$107,287
NoSQL	$105,053
Predictive Analytics	$103,235
MongoDB	$101,323

Table 4. Share of DSA Category Demand by Industry

DSA Framework Category	Professional Services	Finance & Insurance	Manufacturing	Information	Health Care & Social Assistance	Retail Trade
Data-Driven Decision Makers	23%	17%	16%	10%	6%	6%
Functional Analysts	23%	34%	9%	5%	8%	4%
Data Systems Developers	41%	14%	14%	10%	5%	3%
Data Analysts	34%	25%	9%	6%	7%	3%
Data Scientists & Advanced Analysts	31%	23%	12%	10%	6%	4%
Analytics Managers	21%	41%	9%	9%	6%	3%

Module 1

Introduction to Data Visualization

- ❑ **Overview of data visualization**
- ❑ **Data Abstraction**
- ❑ **Task Abstraction**
- ❑ **Analysis: Four Levels for Validation**

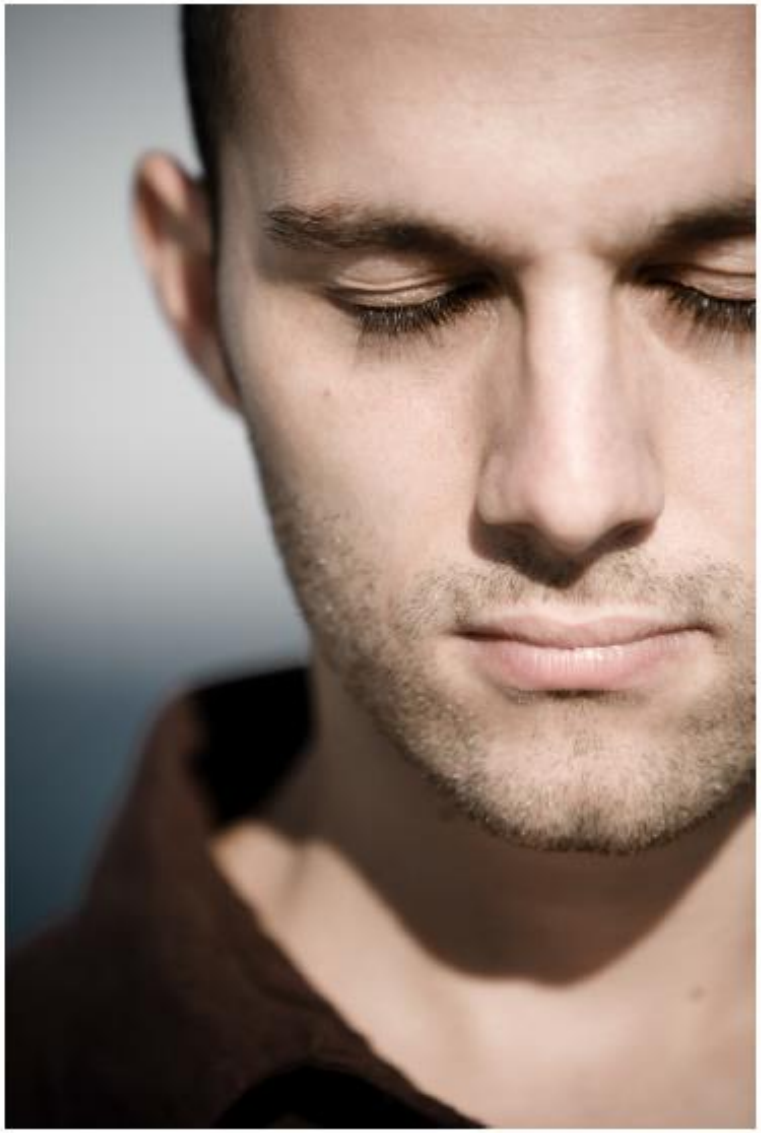
Text Book

Tamara Munzer**, Visualization Analysis and Design** -, CRC Press 2014 . **(Chapter 1, 2,3 and 4)**

Topic Objectives

- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- Explain at a high-level the "what-why-how" framework for analyzing visualization use.
- Describe at least one historical visualization and explain its impact.
- Differentiate between R, D3, and Tableau and describe the type of tasks for which each tool might be most appropriate.

Overview of data visualization

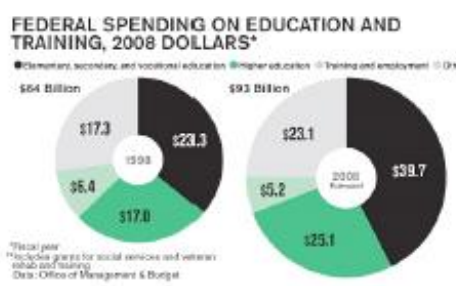


vi · su · al · ize

1. To form a mental image of
2. To make visible

Visualization

To convey information through visual representations



Visualization

- To convey information through visual representations

Map	Clarify
Record	Interact
Abstract	Communicate
Discover	Inspire

Visualisation

The broader field of visualisation has three main sub-fields:

- *SciVis*: Scientific Visualisation (SciVis) typically involves concrete (3d) objects, for example a medical scan of part of the body, or a simulation of air flow around an aircraft wing. SciVis visualisations often depict flows, volumes, and surfaces in (3d) space.
- *GeoVis*: Geographic Visualisation (GeoVis) is map-based. The data typically has inherent 2d or 3d spatial coordinates, and is generally shown in relation to a map.
- *InfoVis*: Information Visualisation (InfoVis) deals with *abstract* information structures, such as hierarchies, networks, or multidimensional spaces.

Data Visualisation (DataVis) = InfoVis + GeoVis.

Visual Analytics = DataVis (frontend) + Analytics (backend).

Why Visualise?

Table vs Line Chart

Compare the table of numbers in Table 1.1 with the visual representation (a line chart) of the same data in Figure 1.1.

- It is much easier to see trends and patterns in the visual representation.
- It is easier to make comparisons in the visual representation.
- It is easier to read off exact data values in the tabular representation (although you could, for example, display exact values upon mouseover in an interactive version of the line chart).

Why Visualise?

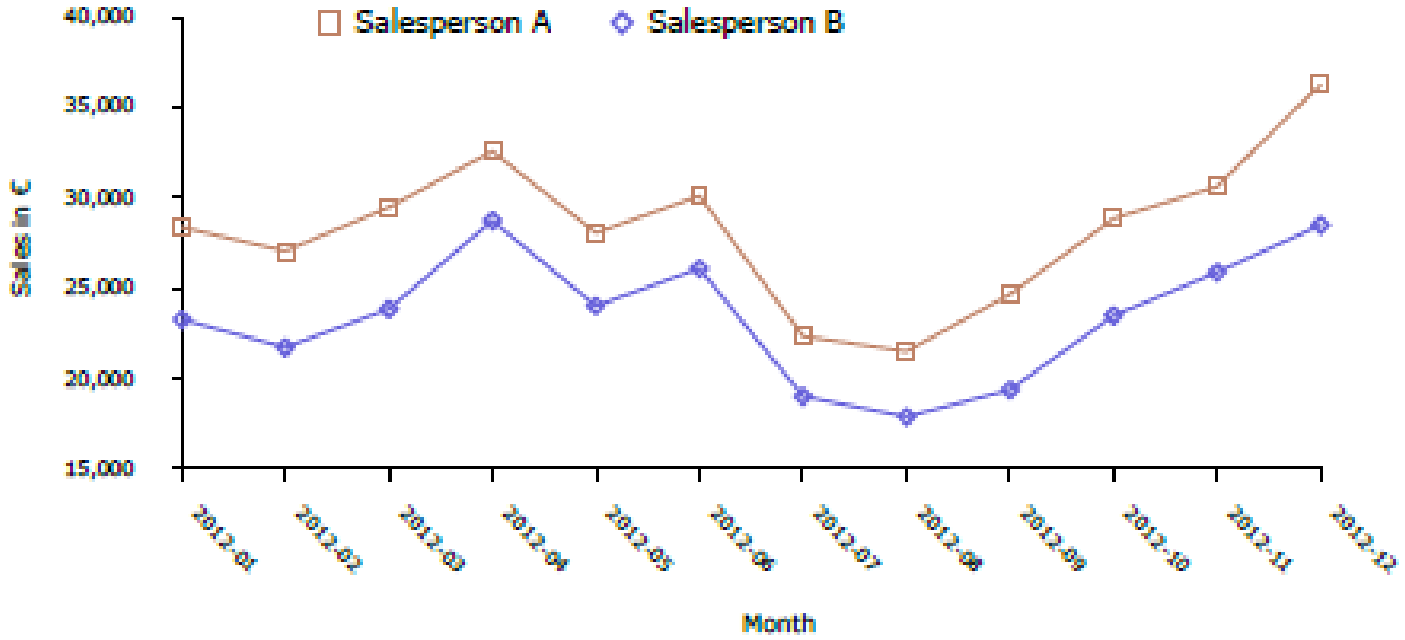


Figure 1.1: Sales for 2012 in € by salesperson. Line chart of the same sales data. It is much easier to see the trends and compare the data, when it is presented visually.

What is visualization?

- "The communication of information using graphical representations"
  - Ward, Grinstein, Keim
- "The use of computer-supported *interactive* visual representations of data to amplify *cognition*"
  - Card, Mackinlay, Shneiderman, *Readings in Information Visualization: Using Vision to Think*
- "The purpose of visualization is insight, not pictures."
  - Ben Shneiderman

What is Visualization

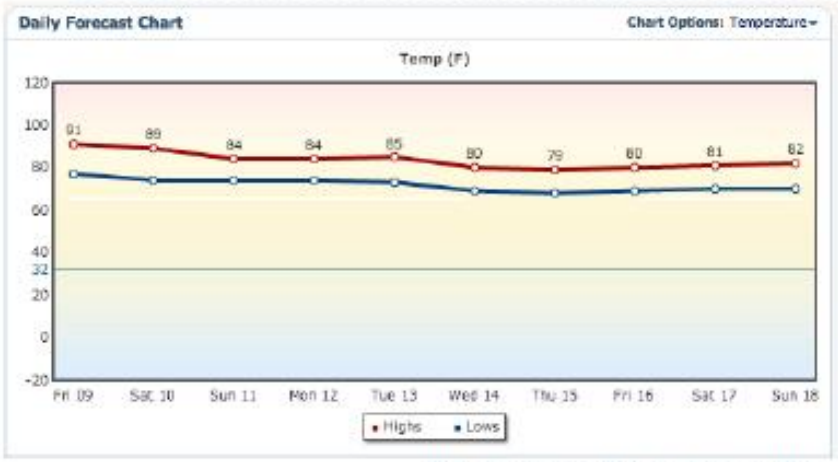
Visual representation of data sets designed to help people carry out tasks more effectively

- Augmenting people's ability rather than replacing
- The design space is huge, with many trade offs
- Validating design is both necessary and challenging
- Need to address three resource limitations
  - Computers
  - Human
  - Displays

We analyzing visualization by answering the following questions:

Why - people need it
What - data are shown
How - the idiom is designed

Where have you seen a visualization today?



http://hamptonroads.com/weather

http://hamptonroads.com/traffic

Select stock category from this list: Biggest Gainers

Data Through 08-09-2013 Time 10:26:54 AM ET

New York Stock Exchange (NYSE)

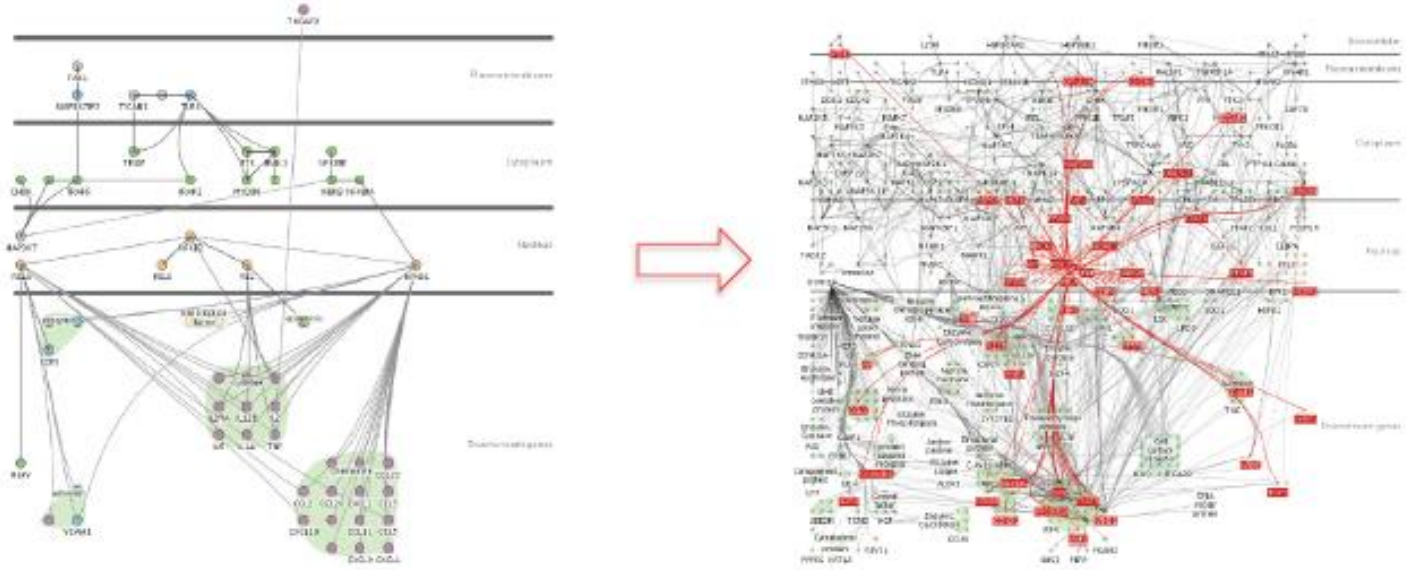
Ticker	Name	Price $	Price Change $	Price Change %
CHC/WS	China Hydroelectric Corp Wt	0.03	0.01	34.15
DATA	Tableau Software	73.19	12.25	20.10
FSS	Federal Signal Corp	11.11	1.55	16.21
NOAH	Noah Holdings	14.81	1.81	13.92
BITA	Bitauto Holdings (ADS)	15.84	1.46	10.15
LITB	LightInTheBox Holding Co Ltd	21.00	1.70	8.81
ZX	China Zenix Auto International ADS	3.18	0.25	8.53
OIBR/C	Brasil Telecom S/A ADS	2.05	0.16	8.47
WG	Willbros Group Inc	8.04	0.60	8.06
RAX	Rackspace Hosting	47.69	3.47	7.85

Why human in the loop

- When people do not know exactly what questions to ask in advance
  - They do not know how to approach the problem
- Putting human in the loop – enhance their ability rather than replace
- Three uses of visualization:
  - Transitional use: help designers with future solutions that are completely computational
  - Long term use: exploratory data analysis
  - Presentation use: explain something that is already known

Why computer in the loop

- People are unlikely to move beyond tiny datasets
- Complex drawings cannot be done by human

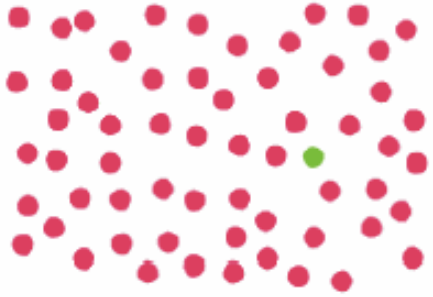


Why use external representations

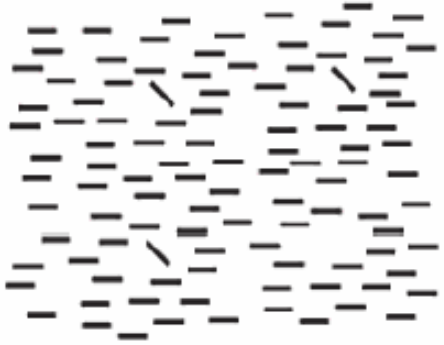
- Augment human's capacity by surpassing our internal cognition and memory limitations
- External representation (diagram)= external memory
  - Organize information in spatial locations
  - Accelerate search and cognition
  - Grouping relevant information in nearby locations

Why depend on vision

- The visual system provides a high bandwidth channel
- Process information in parallel
- Popout occurs even when the number of objects is large
- Sound is not as effective (sequential stream)
- Technological limitations rule out other senses



The green dot pops out



The oblique lines pop out

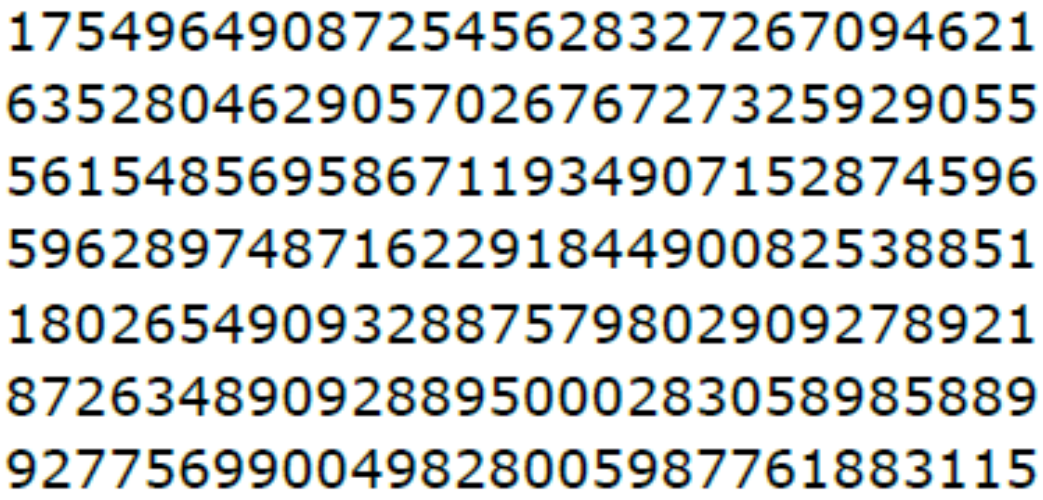


Figure 1.2: Count the number of 3s. Attentive processing requires conscious effort and proceeds serially.

- Certain visual attributes can be processed pre-attentively, which happens without conscious effort and in parallel (fast).

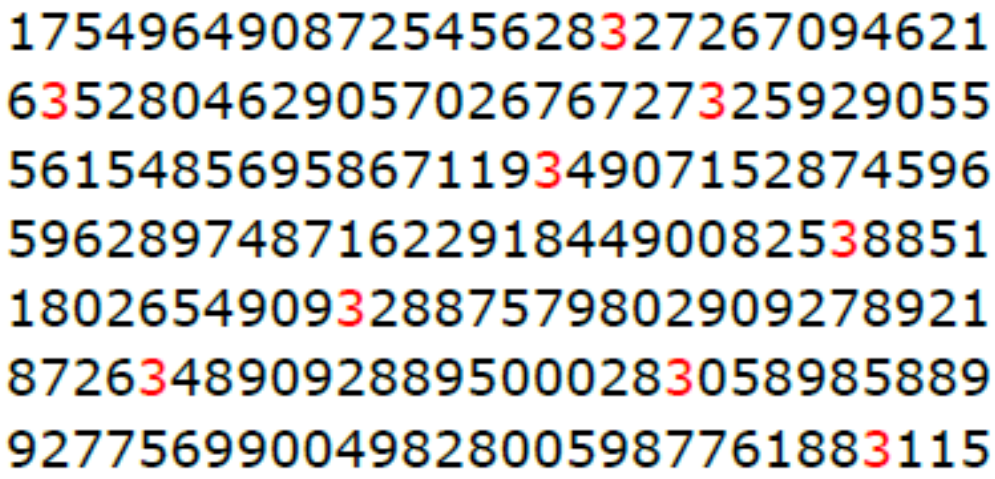


Figure 1.3: Count the number of 3s. Colour is a pre-attentive attribute. By encoding the target 3s in red, they can be rapidly processed by the human visual system pre-attentively. Pre-attentive processing occurs without conscious effort and in parallel.

Why show data in detail

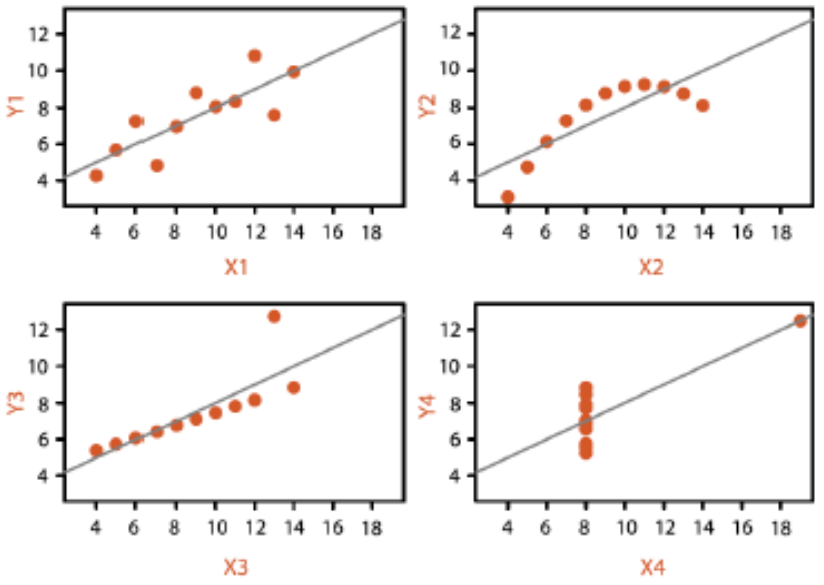
- Exploring data to find patterns
  - Confirm expected ones and discover unexpected ones
  - Assess the validity of statistical models
    - Identical statistics does not mean similar data

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

Why show data in detail

- Exploring data to find patterns
  - Confirm expected ones and discover unexpected ones
  - Assess the validity of statistical models
    - Identical statistics does not mean similar data

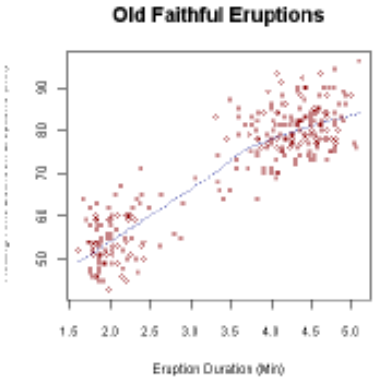


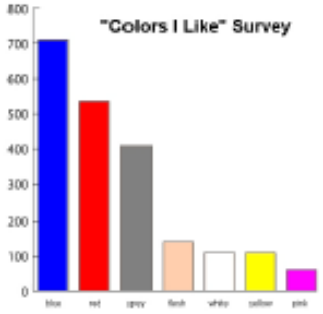
Why interactivity

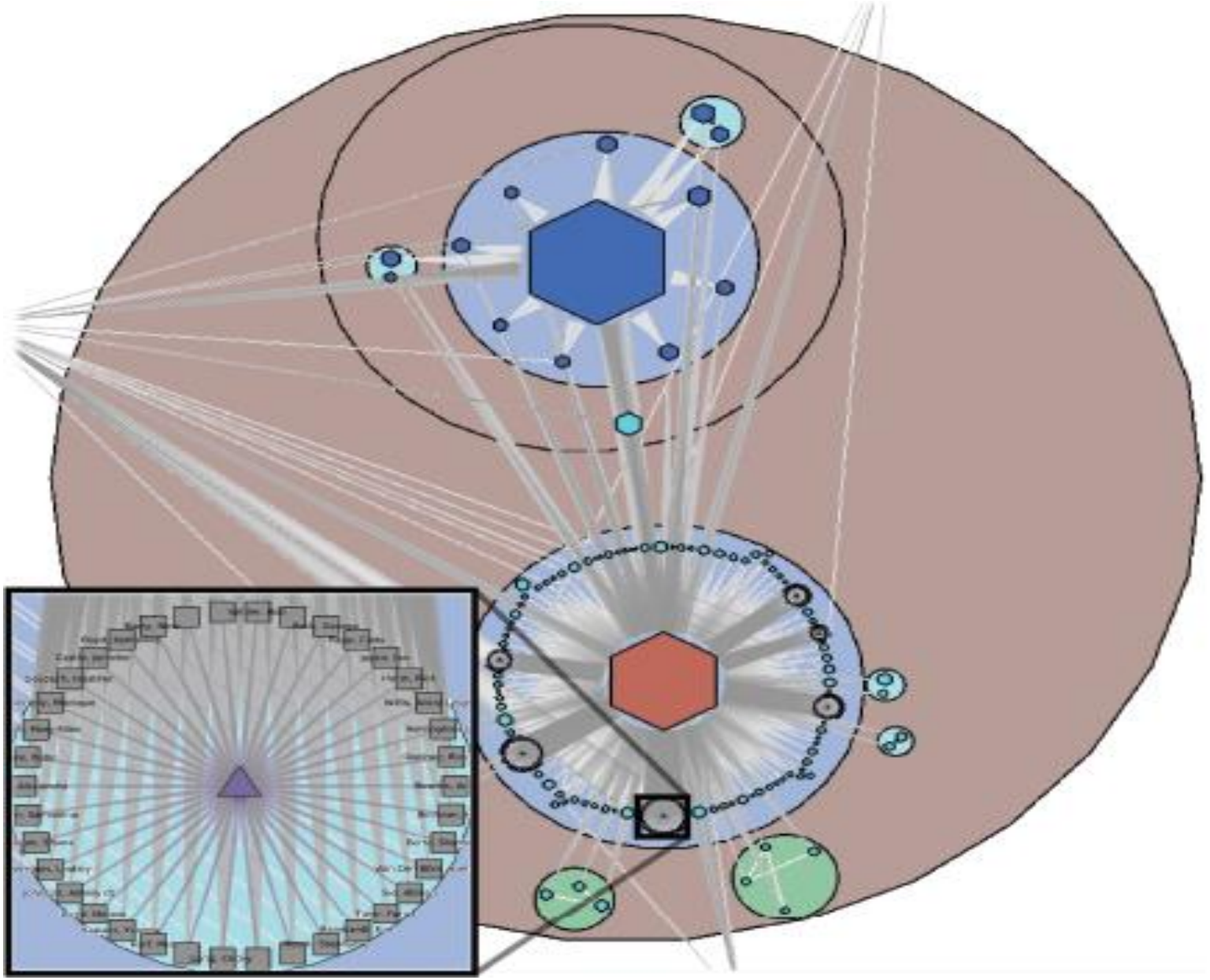
- Crucial for handling data complexity
- Overcome the limitations of people and display – not possible to show all the data at once
- Nee to show multiple aspects of the data
- Change displays to support many queries

Why vis idiom design space is huge

- There exist many ways to create visual encoding
  - Become even bigger when considering interaction
- Idioms: scatter plots, line charts, bar charts , etc., or the linking between them





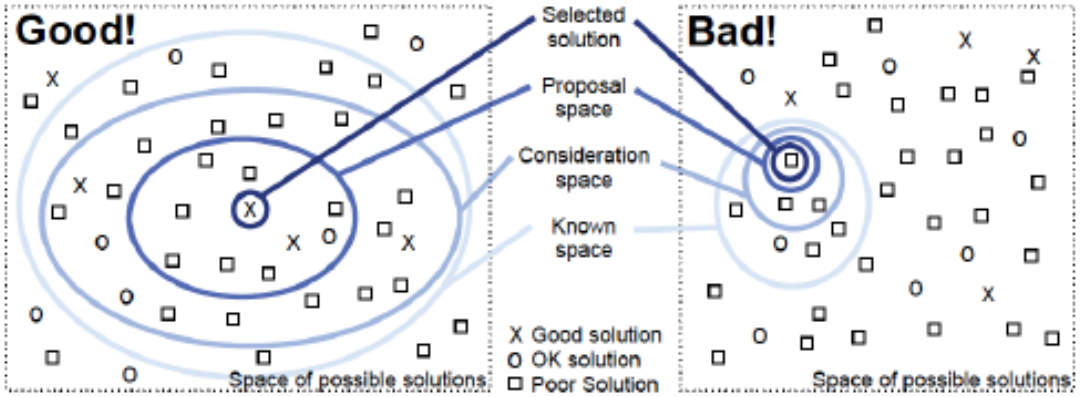


Why focus on tasks and effectiveness

- Tasks
  - a tool good for one tasks can be bad for another
  - Abstract tasks allow one to consider the similar or different needs across multiple fields
- Effectiveness
  - Leads to concerns of correctness, accuracy, and truth
  - A vast majority of design is ineffective for any special context
    - Optimize vs. Satisfy
  - Use a big consideration space, i.e., consider multiple alternatives

Why are most designs ineffective?

- Poor match with the properties of human perception and cognitive system
- Bad match for the intended task
- We should satisfy instead of optimizing

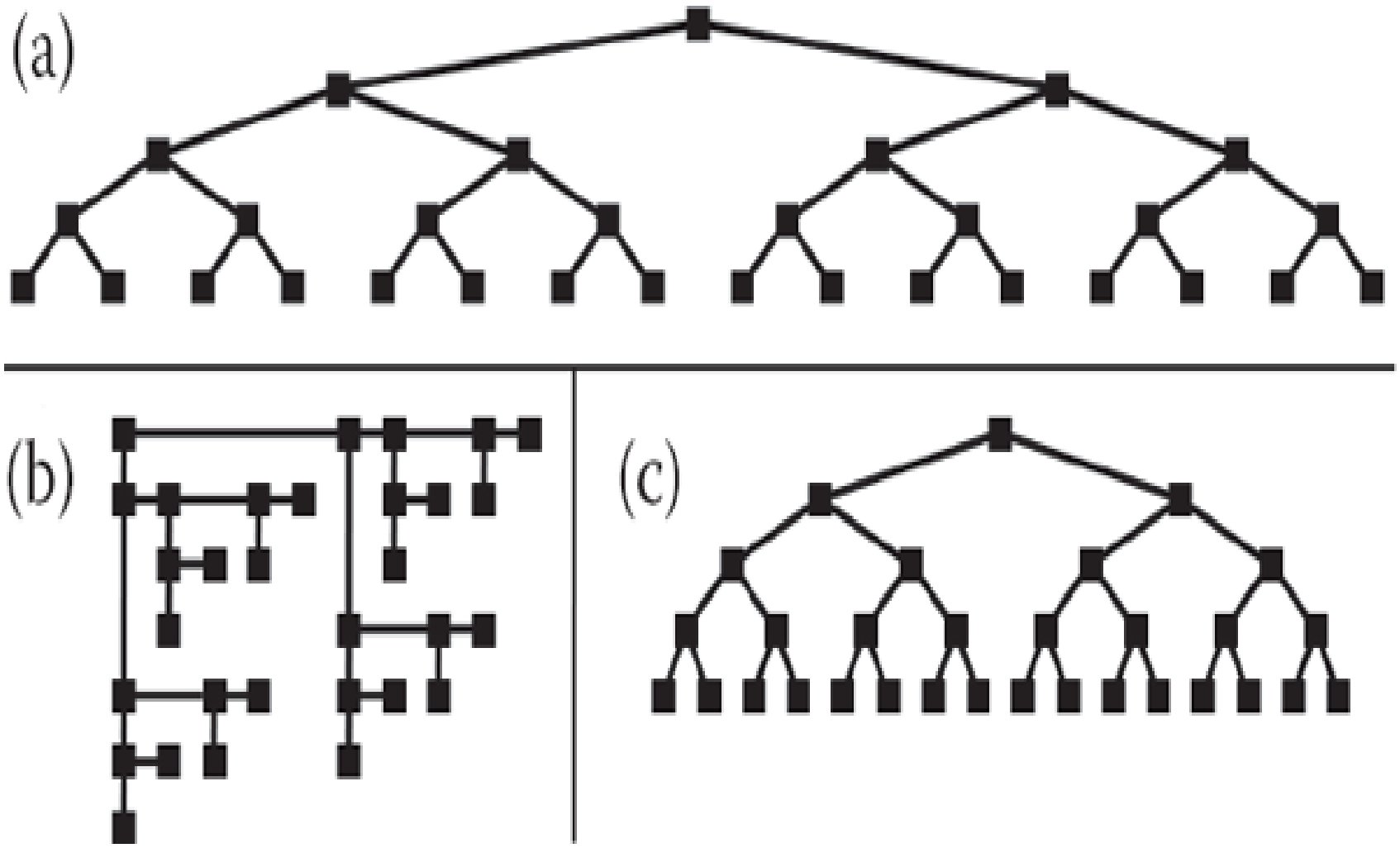


Why is validating difficult

- There are so many questions that you can ask
  - What does better mean?
  - What does effective mean?
  - How to measure insight/engagement?
  - Automatically or manually?
  - Who is the user?
  - What benchmark data and tasks to use?
  - How to measure the quality of images?
  - What is the algorithm's scalability wrt. data and image sizes?
  - …

Why resource limitations

- Limitations: Computation, Human Cognition, Display
- Especially for large data sets (scalability issues)
- Computation: memory and compute time
- Human: memory and attention
- Display: run out of pixels
  - Information density


(a)
(b)
(c)

Why analyzing existing techniques

- Impose a structure on the enormous design space

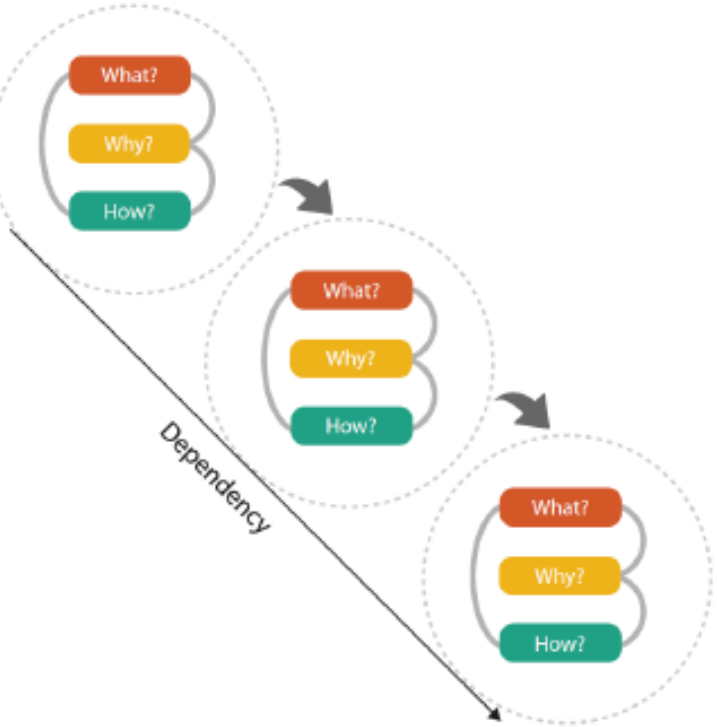
What data the user sees

Why the user intends to use vis tools

How the visual encoding and interaction idoms are constructed

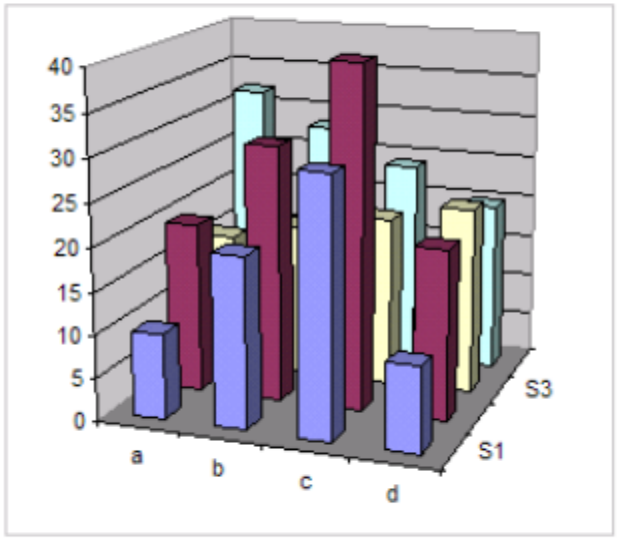
Why analyzing existing techniques

- Complex vis tools are often a sequence of what-why-how instances chained together (input/output dependencies)

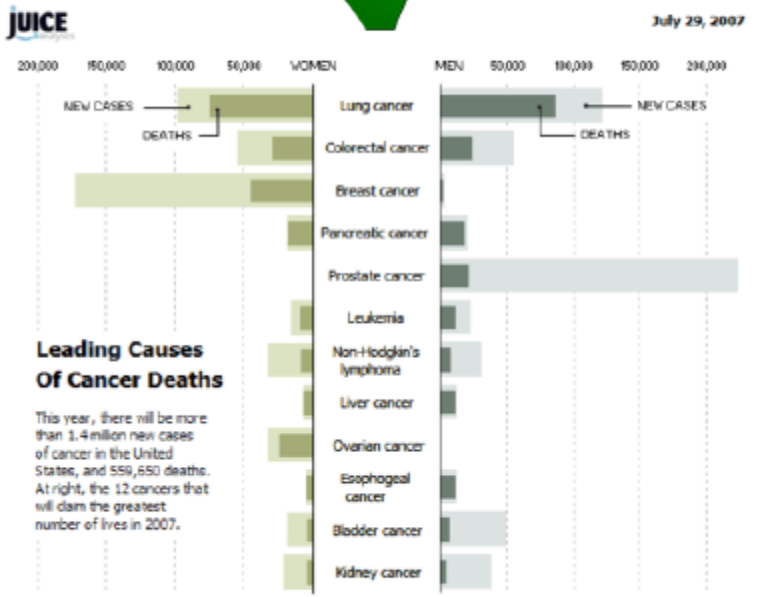


Tools

Excel

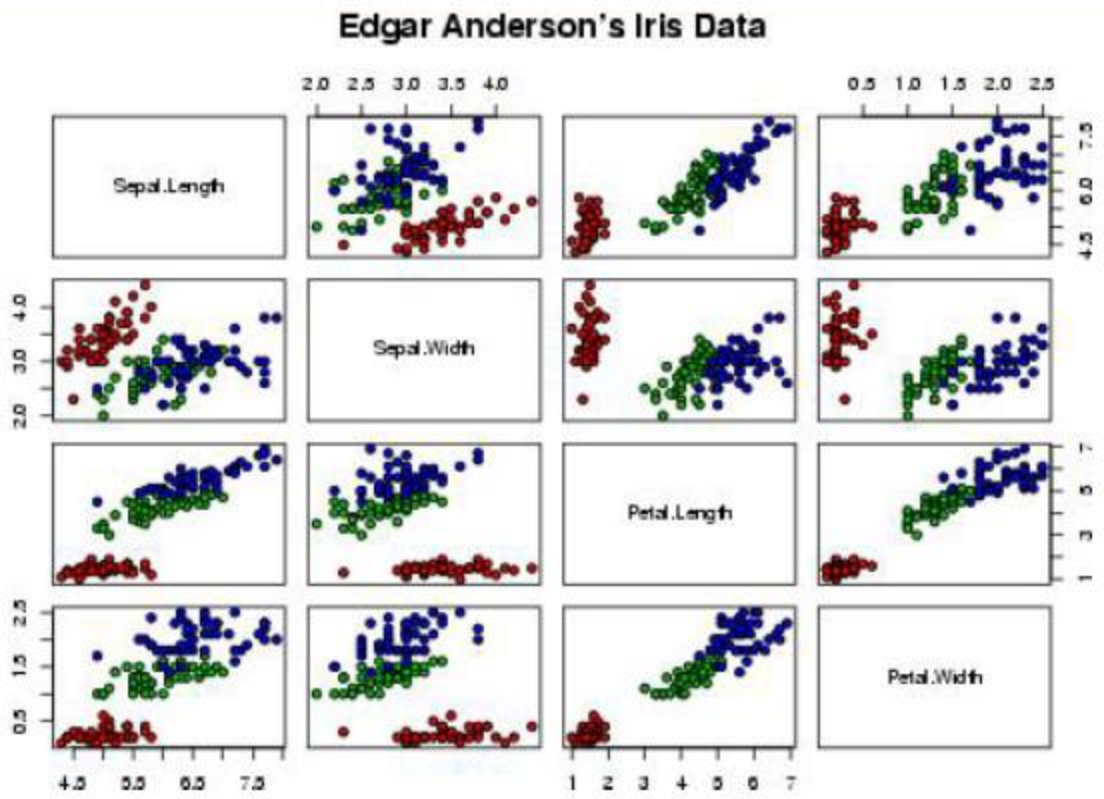


http://chandoo.org/wp/2008/09/03/6-charts-to-never-use/



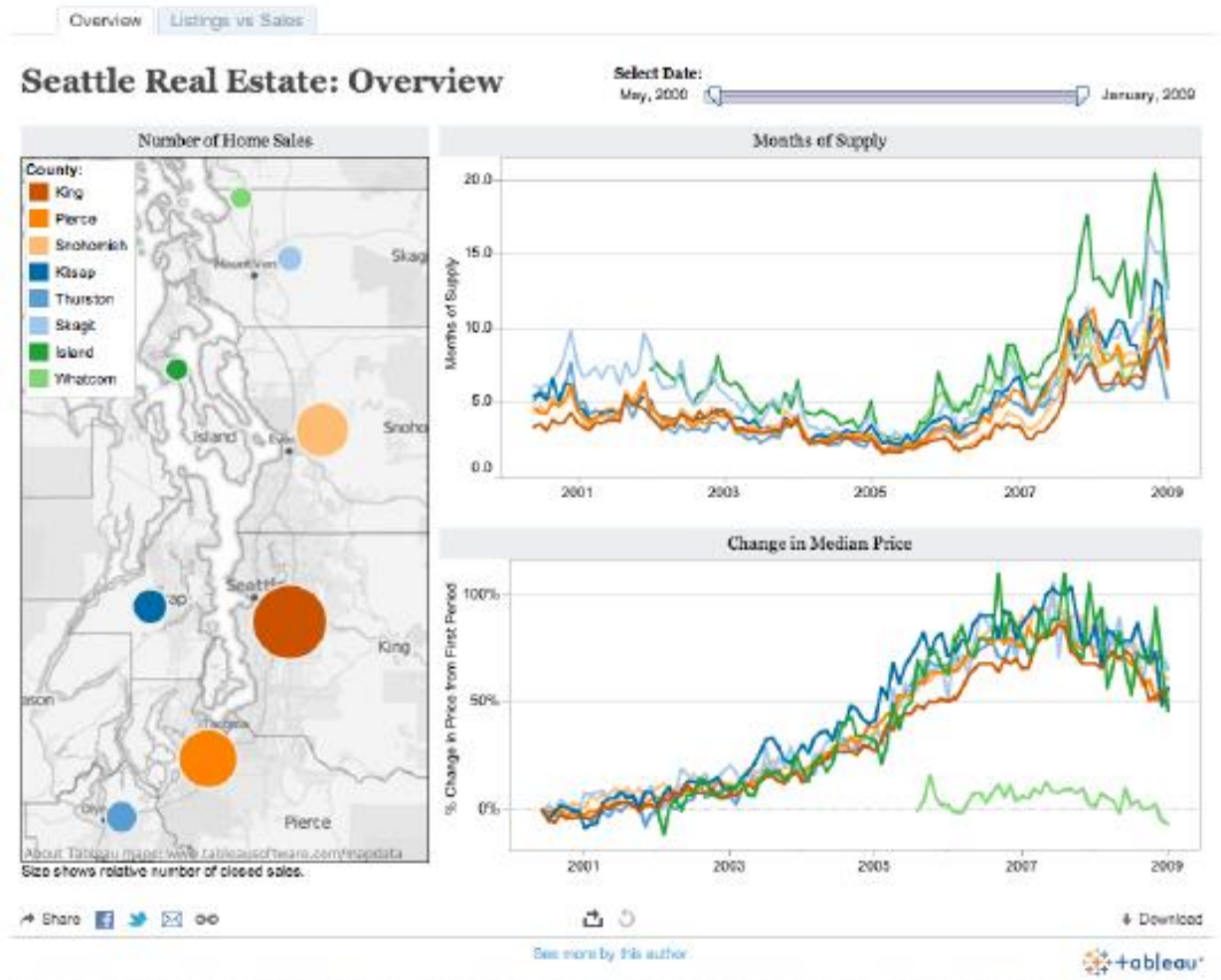
http://www.juiceanalytics.com/writing/recreating-ny-times-cancer-graph/

http://www.r-project.org



Tableau

http://www.tableausoftware.com



D3

http://d3js.org

Data-Driven Documents