Intro and Lecture 1

8/25/2020

Intro

- A little bit about me
- Syllabus
- A little bit about you

How Much Data Is There?

- 2010: 1.2 zettabytes
- 2013: 4.4 zettabytes
- 2020: ~40 zettabytes

zettabyte ~= 1,000,000,000,000,000,000,000 or 1021

200x all words ever spoken by humans

Gantz et al 2011] http://www.6 jamesshuggins.com/h/tek1/how_big.htm

The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades... !

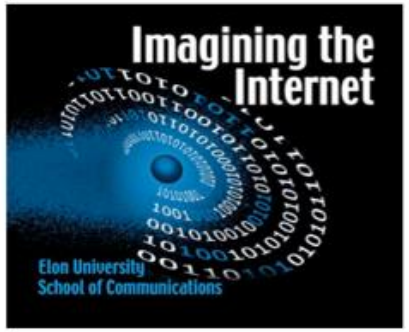
Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist

The McKinsey Quarterly, Jan 2009

PewResearchCenter

Pew Internet
Pew Internet & American Life Project



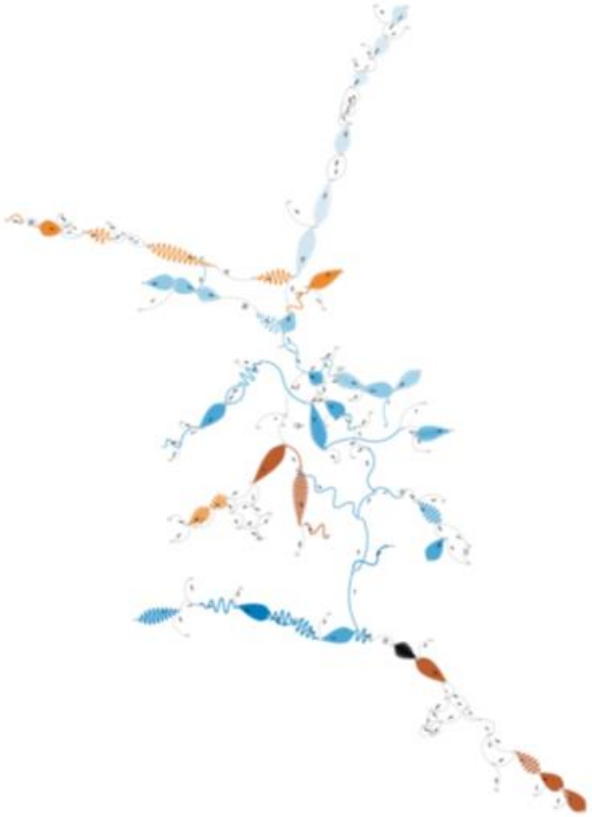
Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well

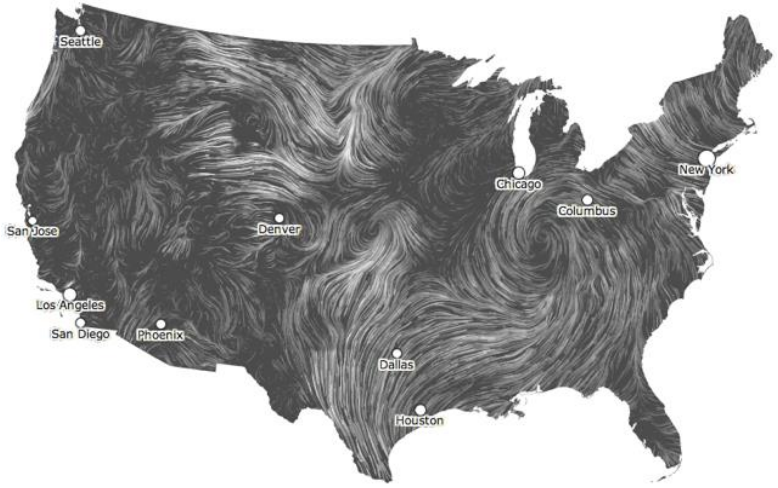
Tech experts believe the vast quantities of data that humans and machines will be creating by the year 2020 could enhance productivity, improve organizational transparency, and expand the frontier of the "knowable future." But they worry about "humanity's dashboard" being in government and corporate hands and they are anxious about people's ability to analyze it wisely

Janna Quitney Anderson, Elon University
Lee Rainie, Pew Research Center's Internet & American Life Project
July 20, 2012

Pew Research Center's Internet & American Life Project
An initiative of the Pew Research Center
1615 L St., NW – Suite 700
Washington, D.C. 20036
202-419-4500 | pewinternet.org

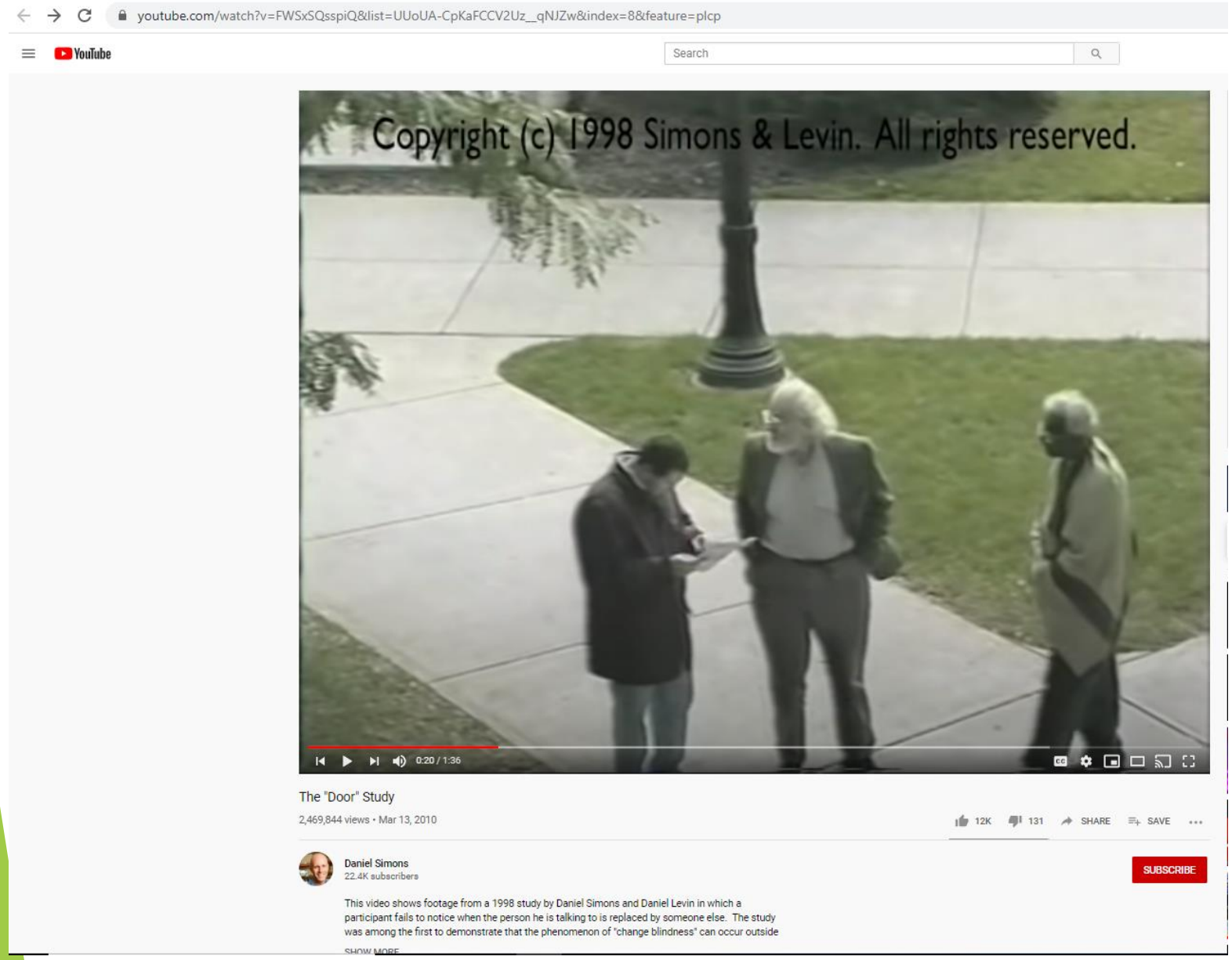
Variety of Data





Why Does Visualization Work?

- Cognition is limited



https://www.youtube.com/watch?v=FWSxSQsspiQ&list=UUoUA-CpKaFCCV2Uz__qNJZw&index=8&feature=plcp

Why Does Visualization Work?

- Cognition is limited
  - Change blindness: large changes go unnoticed when we are working on something else in our view
- Visual working memory is limited

Calculation example

- 34 * 28 = ?

Calculation example

- 34 * 28 = ?
- 79 * 16 = ?

Visualization

- Uses perception to point out interesting things

MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

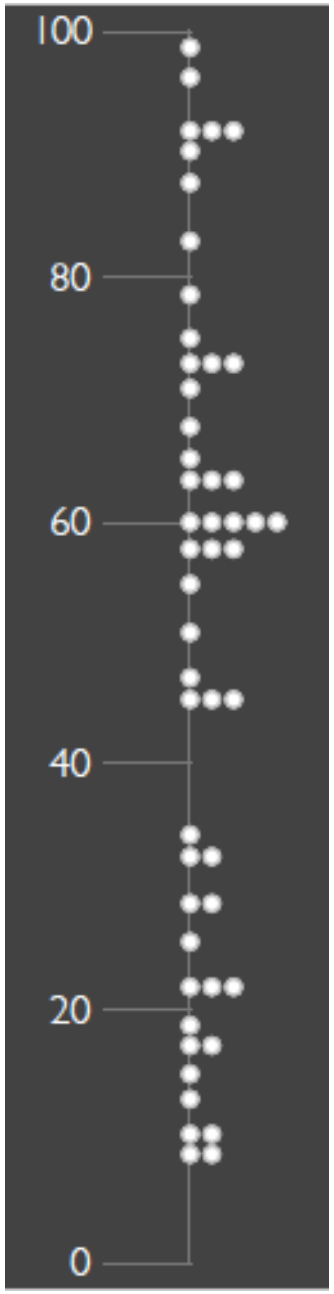
MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

Visualization

- Uses perception to point out interesting things
- Uses pictures to enhance working memory

15	19	60
33	11	75
57	34	79
18	51	92
73	22	13
71	60	22
17	10	68
73	18	55
65	46	29
60	73	22
46	92	97
10	58	46
57	17	83
26	99	33
88	92	60
91	29	57
96	12	47

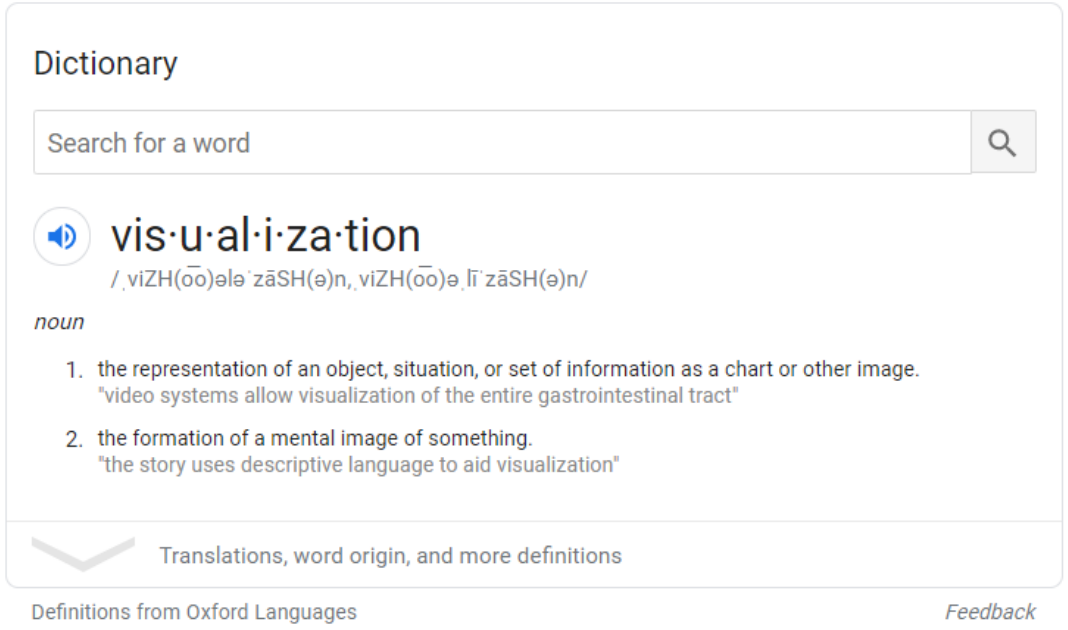
Given these 50 numbers,
what number appears most often?



Given these 50 numbers,
what number appears most often?

Visualization

Formal definition:



"Computer-based visualization systems provide visual representations of datasets intended to help people carry out tasks more effectively.."

Tamara Munzner

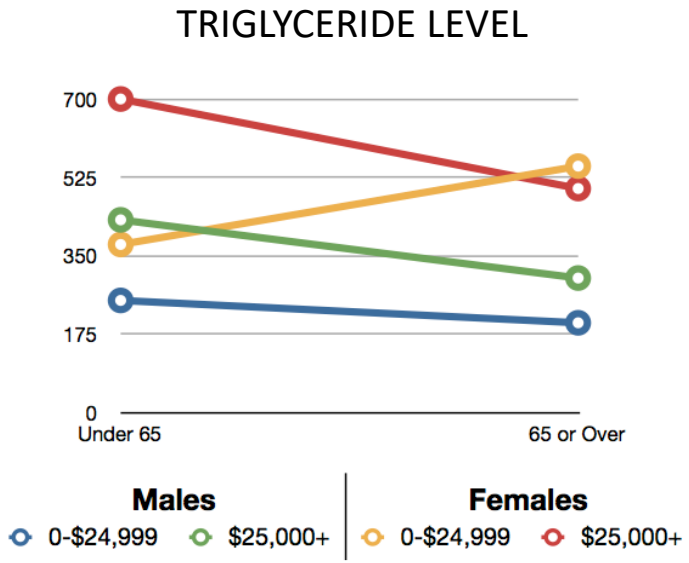
Why

- Query exercise

TRIGLYCERIDE LEVEL

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-$24,999	250	200	375	550
$25,000+	430	300	700	500

Question:
Which gender and income level shows a different effect of age on triglyceride levels?



Question:
Which gender and income level shows a different effect of age on triglyceride levels?

Why Visual…

Anscombe's Quartet: Raw Data

I		II		III		IV	
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

Descriptive Statistics
Graphing Data
Outlier Influence

[F. J. Anscombe]

Why Visual…

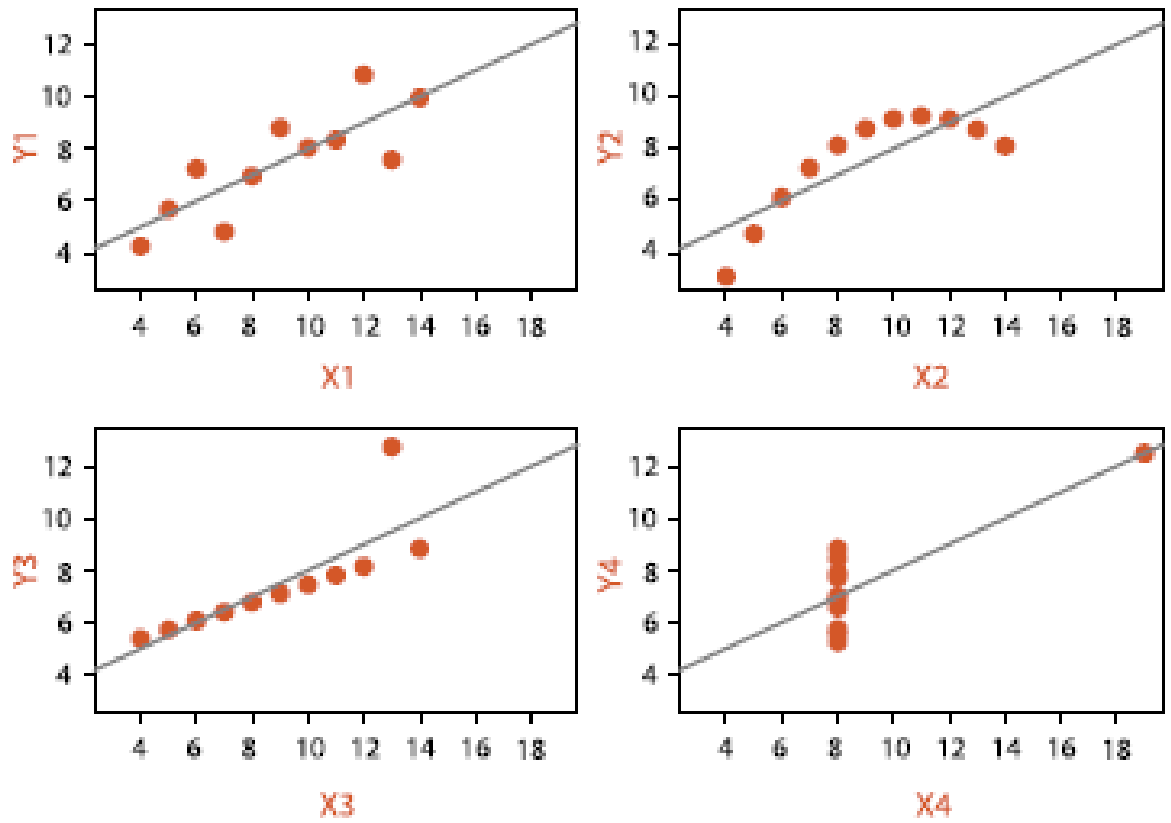
Anscombe's Quartet: Raw Data

	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	

Mean of x	9
Variance of x	11
Mean of y	7.50
Variance of y	4.122
Correlation	0.816

[F. J. Anscombe]

Why Visual…



Why create visualizations?

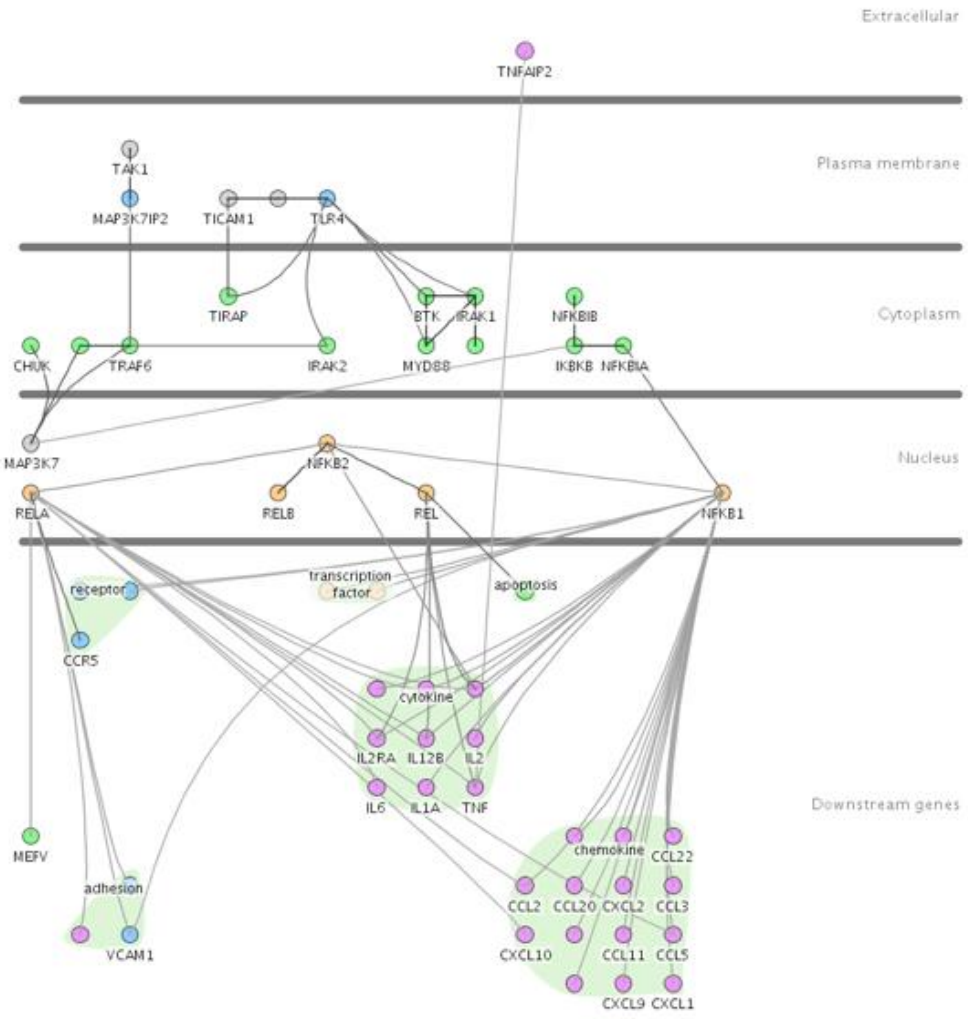
- Answer questions
- Generate hypothesis
- Make decisions
- See data in context
- Expand memory
- Support computational analysis
- Find patterns
- Tell a story
- Inspire

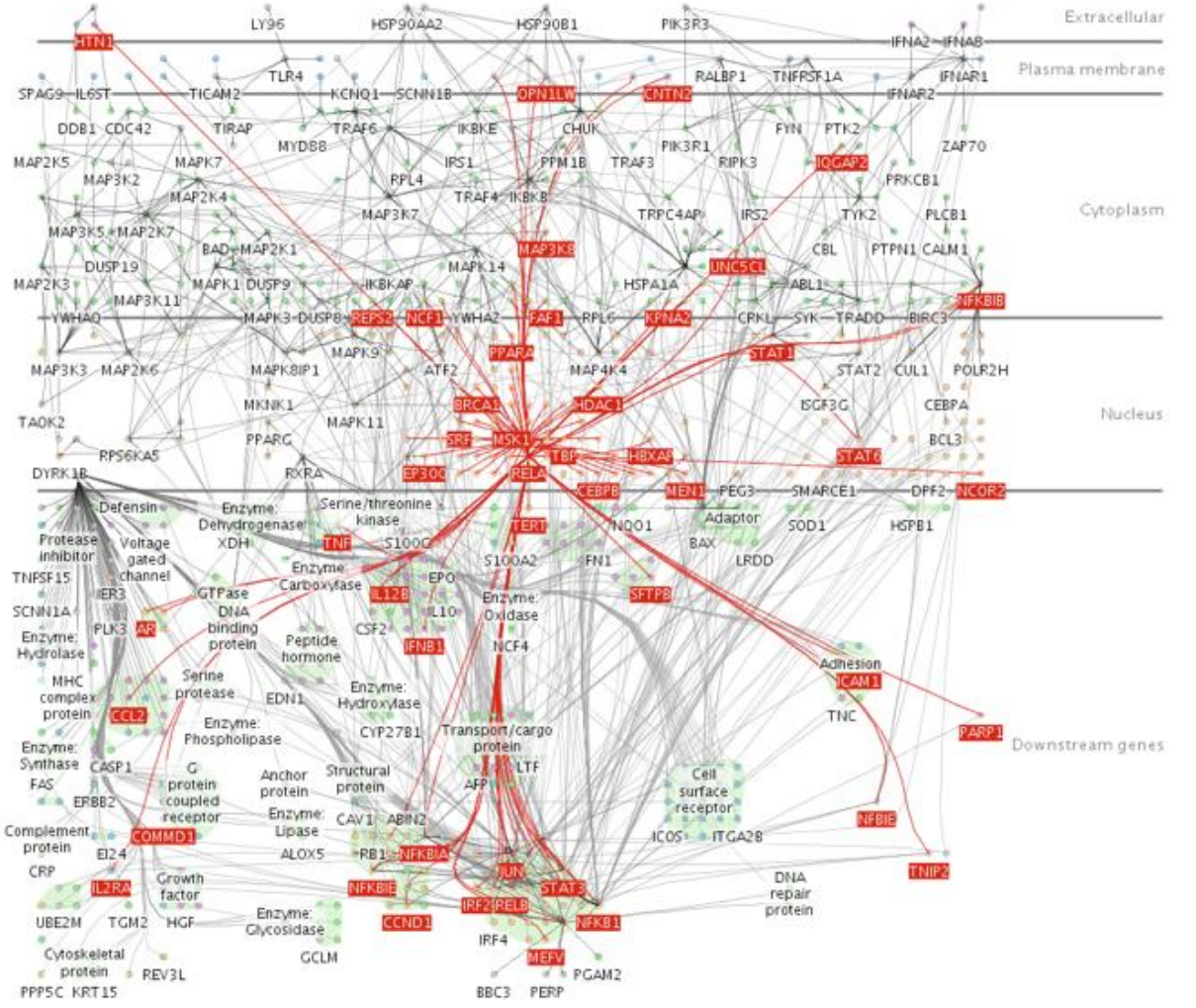
It is often easier to show someone something than to tell them a bunch of facts about the data (and let them explore it)

Visualization Goals

- Record information
- Analyze data
- Confirm hypothesis
- Communicate ideas to others

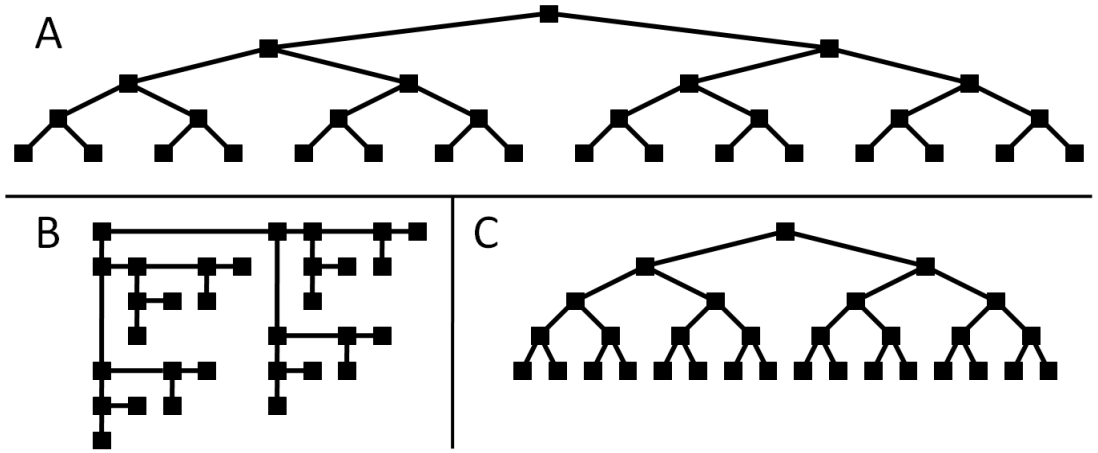
Why Computers?





Resource Limitations

- Memory and space constraints
- How many pixels do I have?
- Information Density

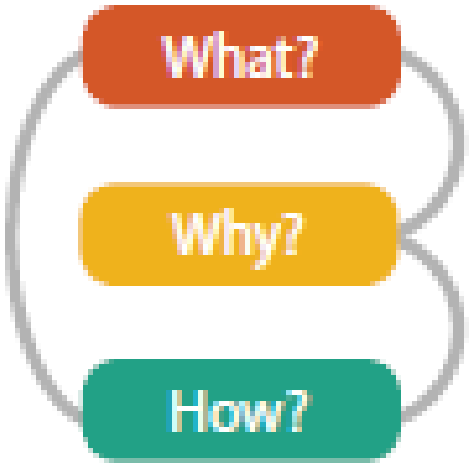


[McGuffin & Robert, 2010]

Effectiveness

- “It’s not just about pretty pictures”
- Any depiction of data requires the designer to make choices about how that data is visually represented
- Effectiveness measures how well the visualization helps a person with their tasks
  - Insight, engagement, efficiency
  - Benchmarks and user studies

Three-part analysis framework for a visualization



Sources of Visualization

- Excel
- Programming softwares
  - R
  - Python
- Other tools
  - Tableau
  - Sisence
  - Microsoft Power BI
  - SQL Server Reporting Services
  - IBM Cognos Analytics

Data

- Iris dataset found in R base package
- Excel
  - Pivot Table and Chart
- R
  - ggplot2 vignette
    - https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
  - Basic plots
    - Using the base packages
  - ggplot2 plotting

Sources/Credits

- Tamara Munzner, Visualization Analysis & Design, A K Peters Visualization Series, CRC Press, 2014.
- Utah, Miriah Meyer, Visualization (2014).
- UMass Dartmouth, David Koop, Data Visualization (2015).