CS 109: Data Science

Process, Data, and Visual Attributes

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

This Week

- HW0 - due next Tuesday (not graded)
  - Install Anaconda & IPython frameworks
  - Sign up for Piazza and introduce yourself
  - Fill out survey
- Friday lab **10-11:30 am** in MD G115
  - *Intro to Python* with Ian Stokes-Rees
  - Make sure to have IPython installed and ready
- Readings - post comments on Piazza

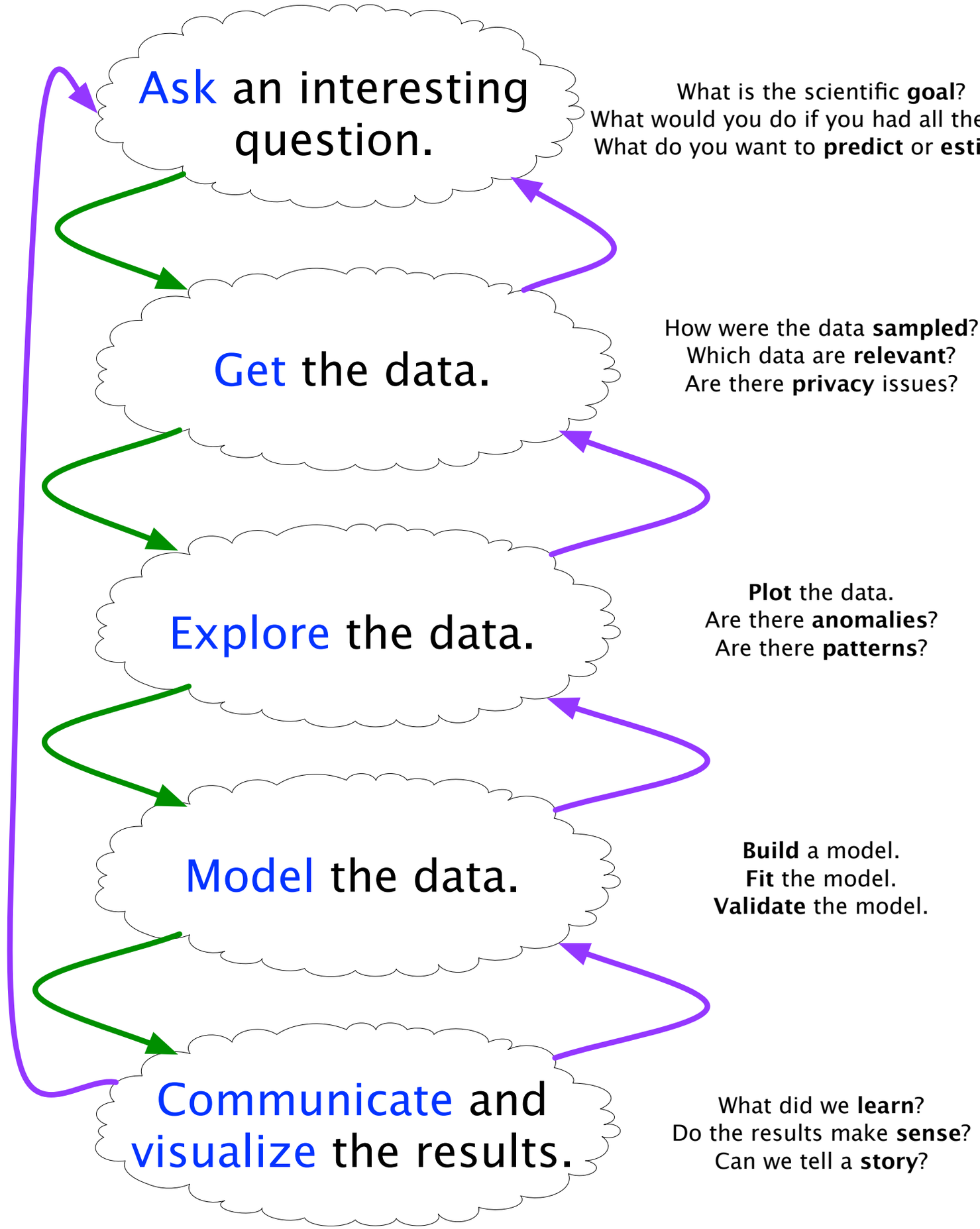
Outline

- Process & Process Books
- What makes visualizations effective?
- Data Sources & Data Cleanup

Process

Data Exploration

Not always sure what we are looking for (until we find it)


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

What do analysts do?

Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

1 Introduction

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

ery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note

What do analysts do?

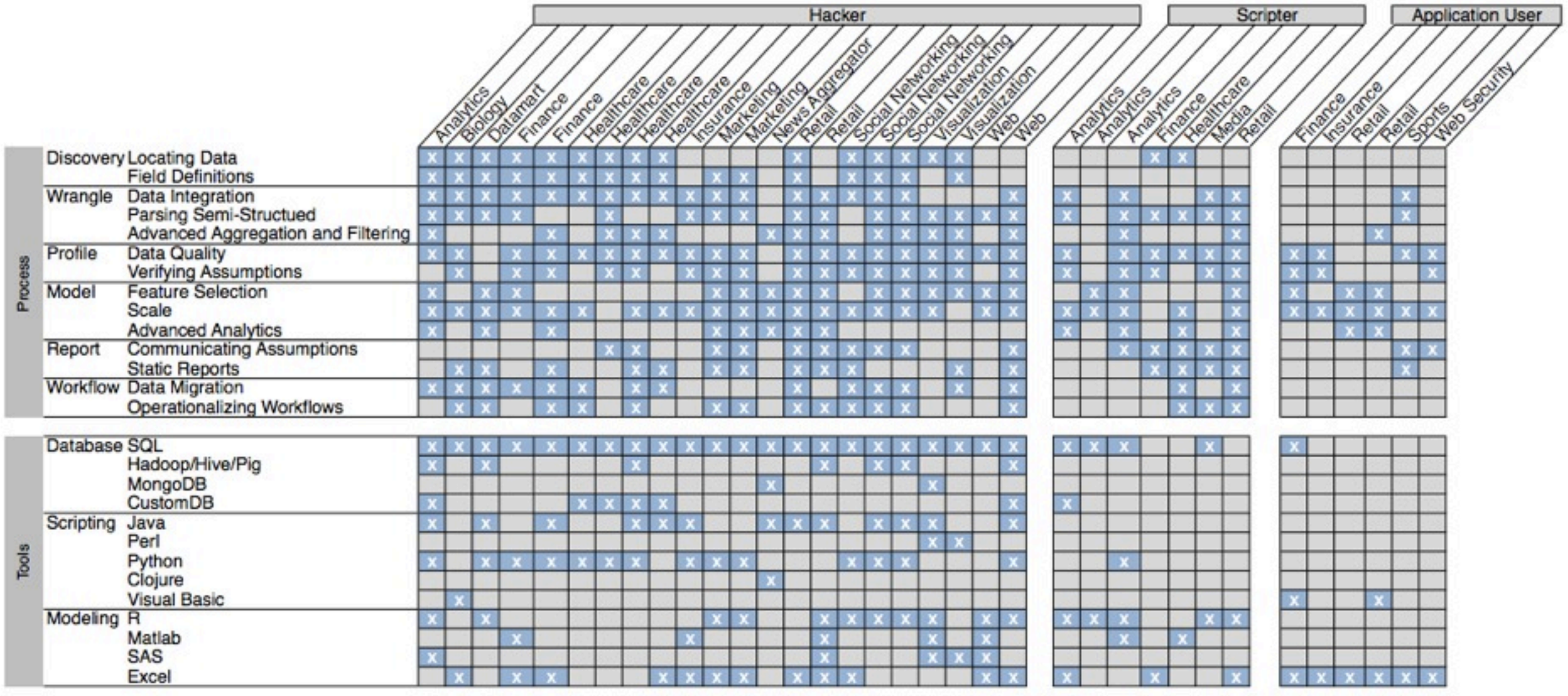


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

What do analysts do?

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis. Most of the time once you transform the data you just do an average... the insights can be scarily obvious. It's fun when you get to do something somewhat analytical.

Exploratory Data Analysis

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

John Tukey

Ascombe's Quartet

Same mean, variance, correlation, and linear regression line

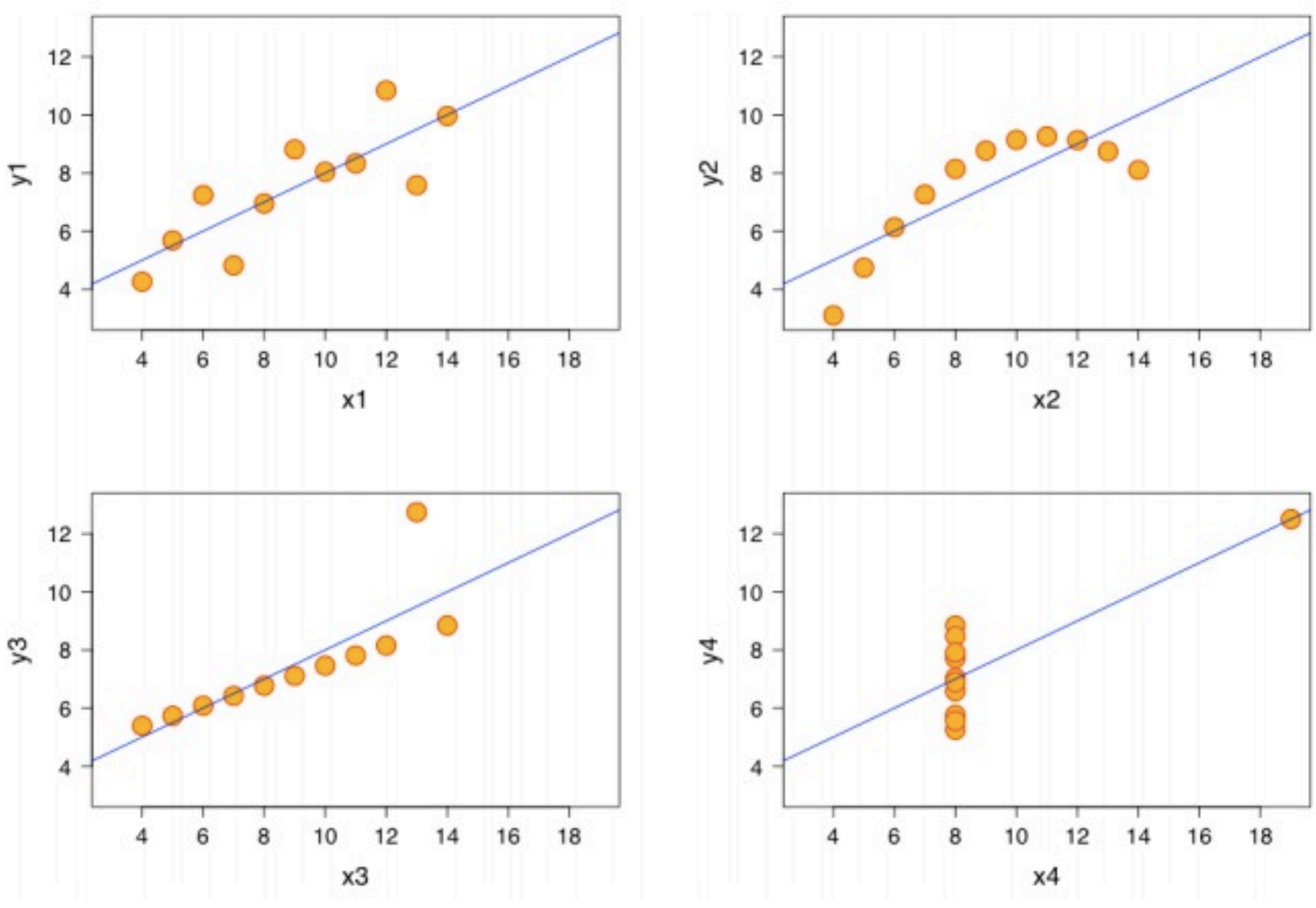
Anscombe's Quartet: Raw Data

	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
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

Anscombe '73

Ascombe's Quartet

Same mean, variance, correlation, and linear regression line



Anscombe '73

Example: Antibiotics

Will Burtin, 1951

Effectiveness of Antibiotics

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

Data & Questions

- What are the data types?
- What are possible questions?

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

Data

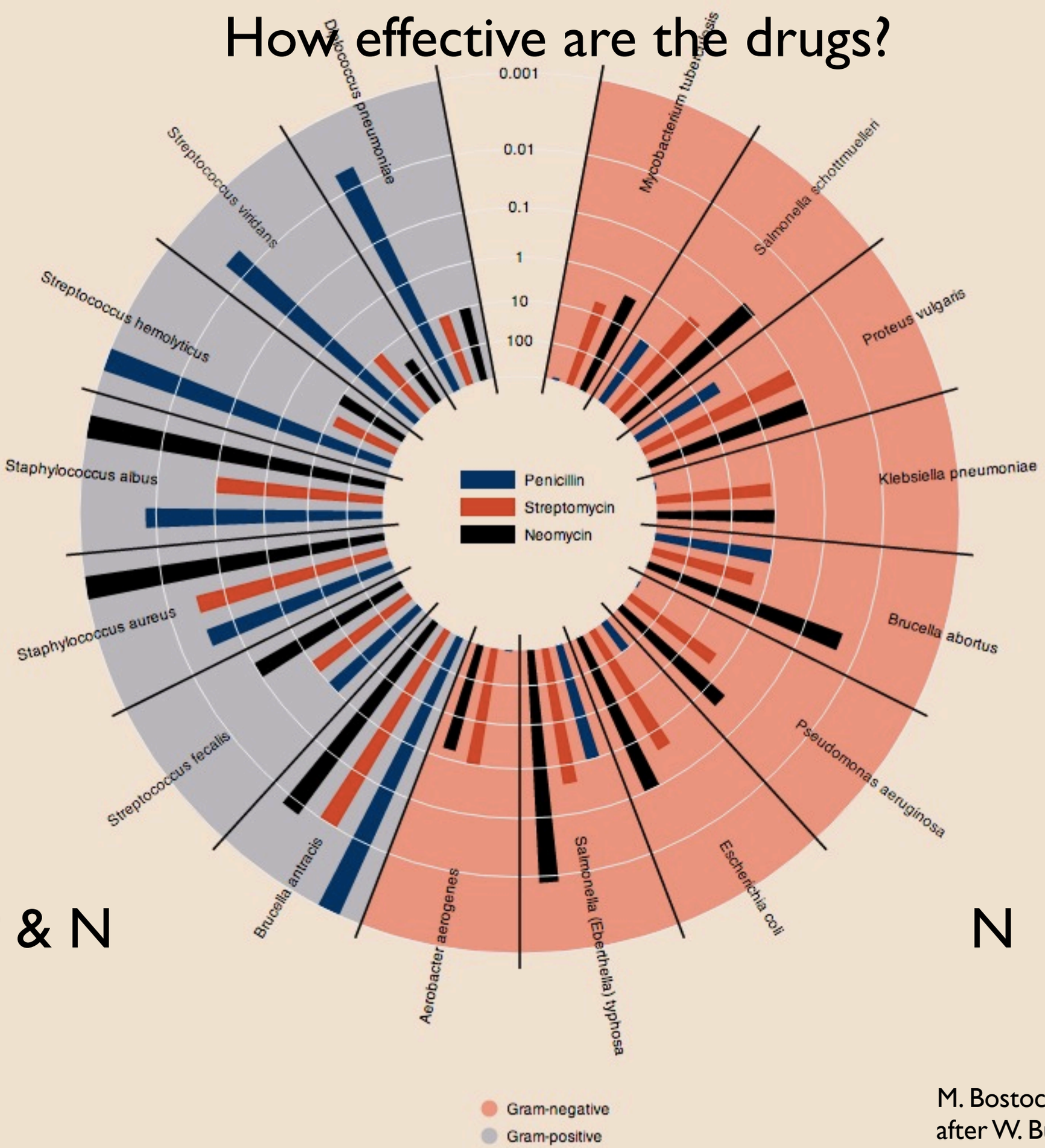
- Genus & species of bacteria [string]
- Antibiotic name [string]
- Gram staining? [pos/neg]
- Minimum inhibitory concentration (mg/ml) [float]

(lower == more effective)

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+

What Questions?

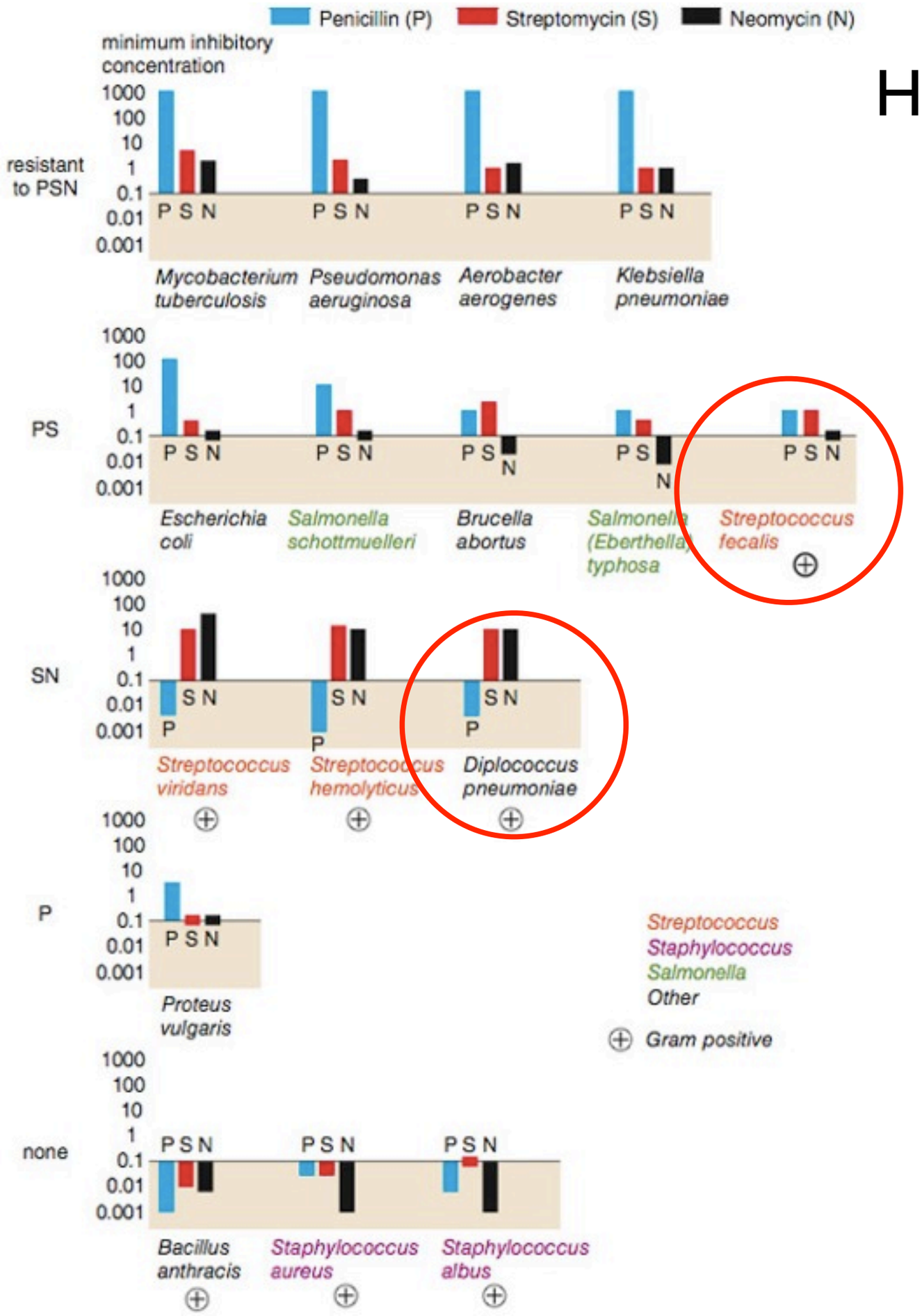
How effective are the drugs?



P & N

N

M. Bostock, Protovis
after W. Burtin, 1951



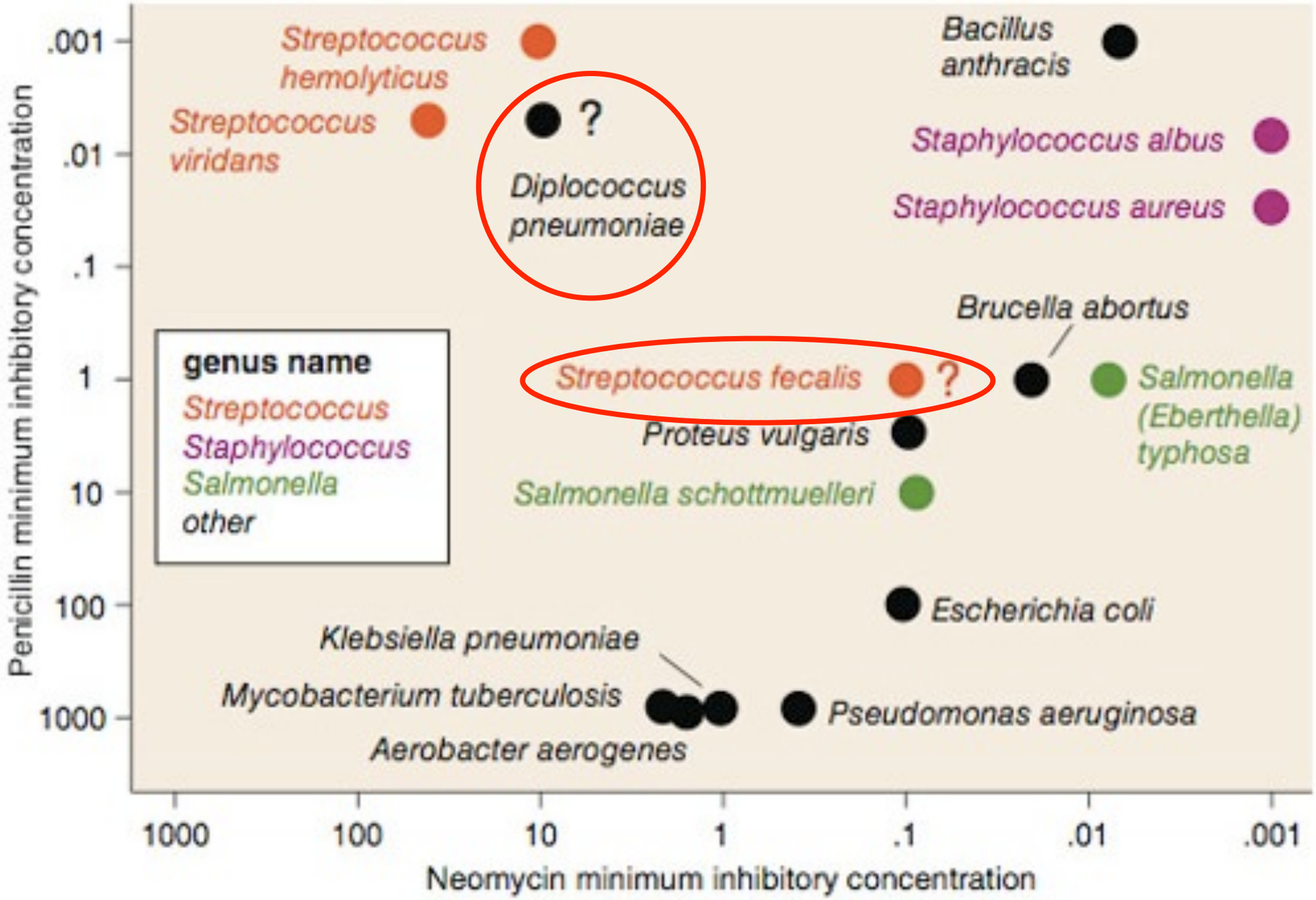
How do the bacteria compare?

Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

How do the bacteria compare?



Wainer & Lysen, "That's funny..."
American Scientist, 2009

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

John Tukey

Process Books

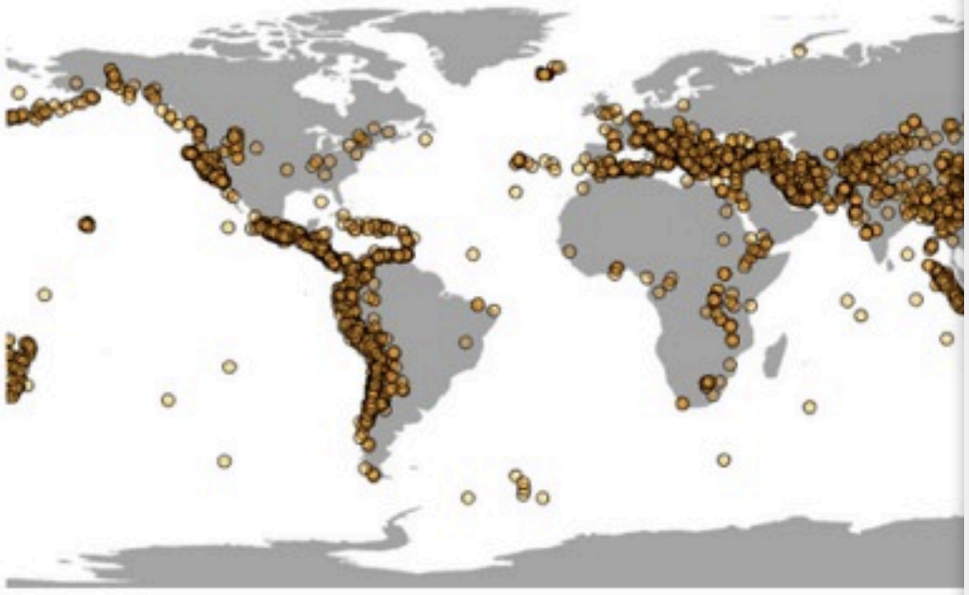
Process Books

[Blake Walsh, Gabriel Trevino, Antony Bett, CS171, 2013]

More Data Cleaning and Automation of Point Plotting in Prototype
[entry by btwalsh, 04.01]

Today, now that I have confirmed that my implementation of a few sample points is functional, I took our XLS file of earthquake data and began to clean it up some more with Google Refine specifically for the purpose of easily integrating it with our existing Processi the cleaning that was performed, I renamed the columns to read more sens entries that did not indicate a latititude and longitude for the location, and th rows of data that similarly did not actually report a magnitude.

Now, we have implemented a way to automatically map our latitudinal/longi onto our grid of x-y pixels, as demonstrated below:



[note by gtrevino 04.01]

Thankfully we got this to work (after dealing with some null pointer exceptio work on our data density, i.e. how to deal with multiple points overlapping, a especially important as we implement **user interaction**. This will probably r

[Varun Bansal, Cici Cao, Sofia Hou, CS171, 2013]

project process book

Users

The target audience is the general public, lovers of music, or simply those who want to learn about music. To appeal to the eclectic tastes of each individual, we have included nine main genres of music ranging from classical to country to Rap. This is not a visualization on the fundamentals of music theory so everyone has equal access to it - both for education and personal enjoyment!

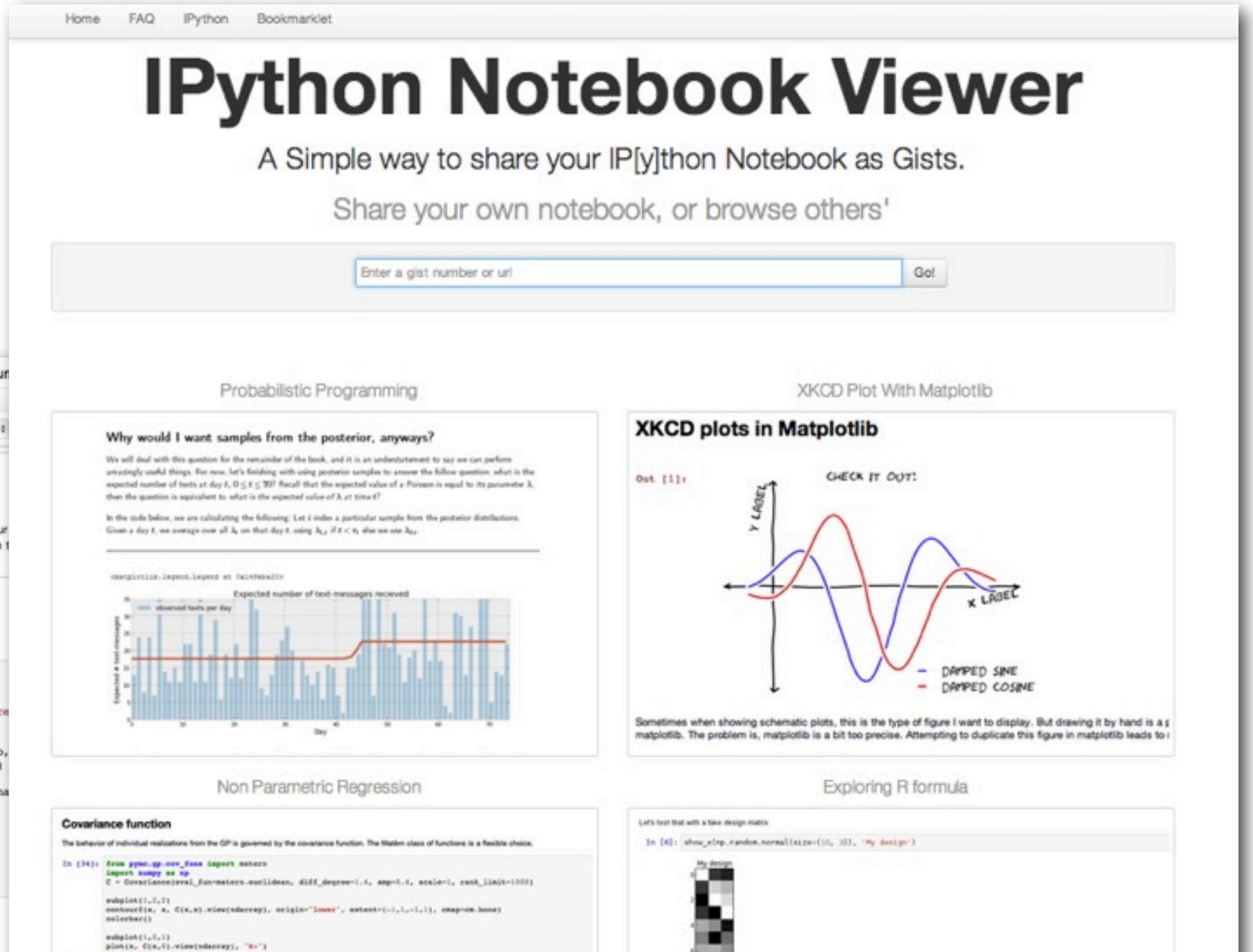
Related Work

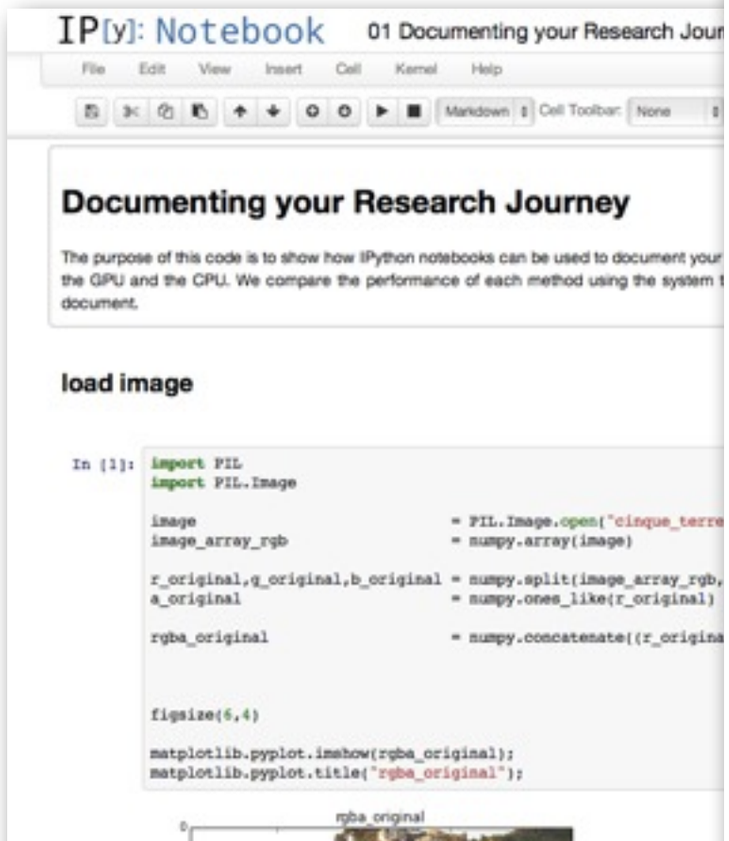
Everyone loves music, and with the physics of sound the data on music is limitless. Each individual piece of song can have a vast amount of frequency, amplitude, and beats data. However, we were inspired to analyze music after seeing a project named the Shape of a Song (http://bewitched.com/song.html)in which the author visualized, through arch diagrams, similar patterns within different types of songs. This provided analysis on two levels: song and genre specific. On the song specific level, the arches showcased observations such as how the beginning of the song was similar to the end. On the genre specific level, comparing the arch diagrams of songs from different genres (i.e. modern techno or pop versus classical) shows characteristic patterns of each genre, with modern electronic and synthesized music displaying more repetition and similarity throughout the piece.

Chopin, Mazurka in F# Minor
The image illustrates the complex, nested structure of the piece.

IPython Notebooks

http://nbviewer.ipython.org/





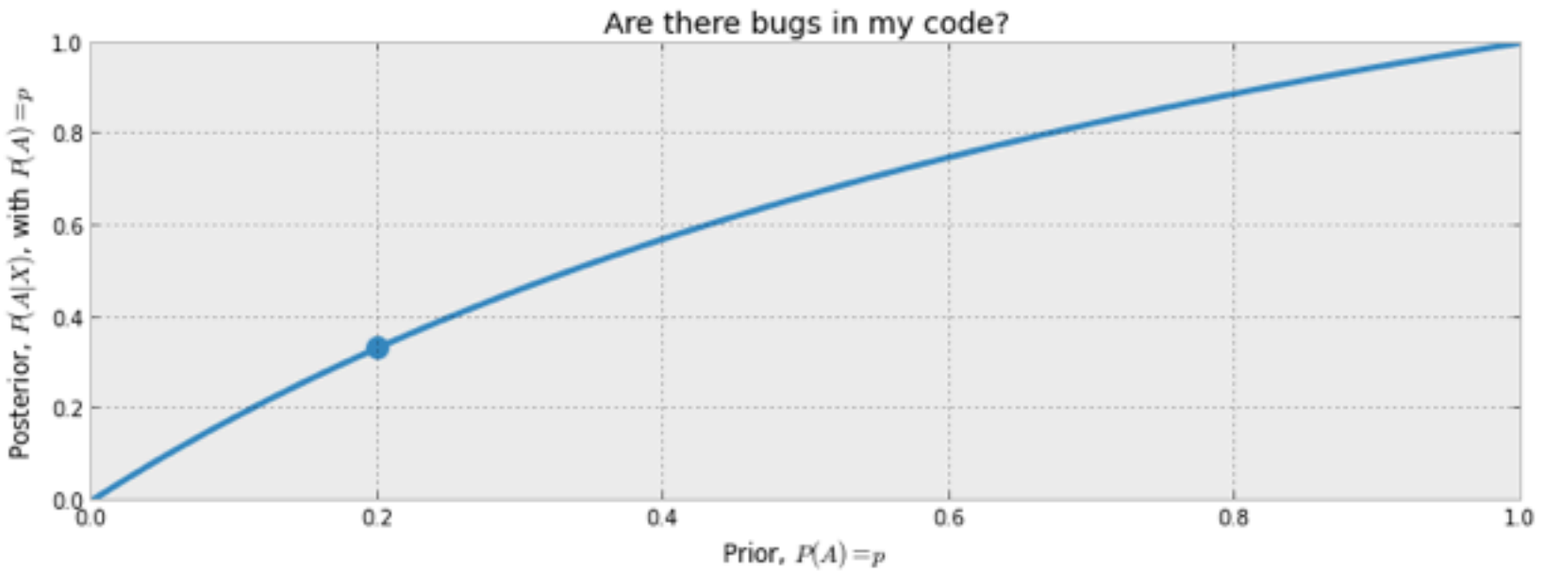
IPython is Great

(for large-scale computation, data exploration, and creating reproducible research artifacts)

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)}$$
$$= \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

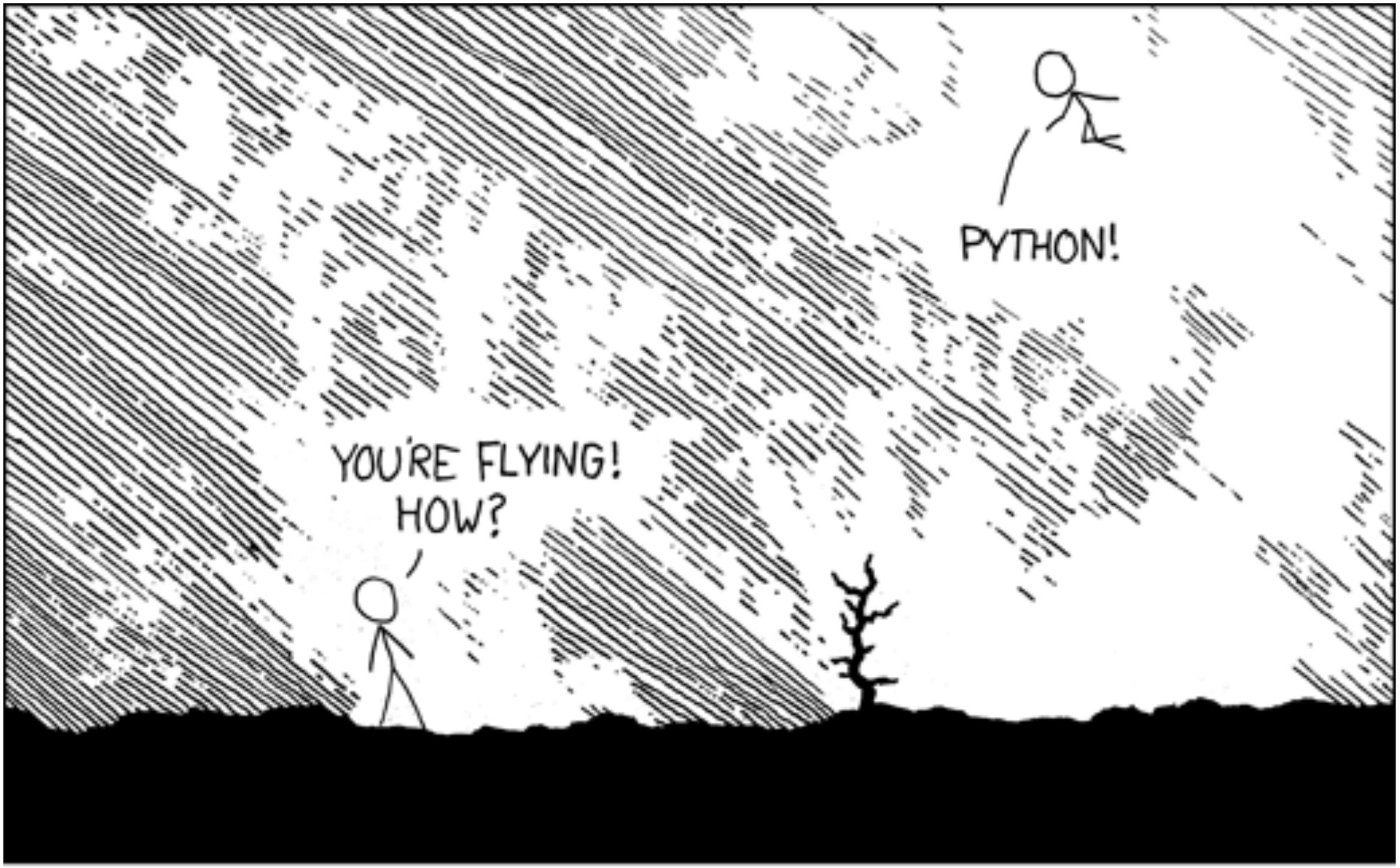
```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```



[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

Mike Roberts, Stanford University

Python is great.



[http://xkcd.com/353/]

E.g.: https://github.com/jrjohansson/scientific-python-lectures

Home FAQ IPython Bookmarklet Download Notebook

Introduction to Python programming

J.R. Johansson (robert@riken.jp) http://dml.riken.jp/~rob/

The latest version of this IPython notebook lecture is available at http://github.com/jrjohansson/scientific-python-lectures.

The other notebooks in this lecture series are indexed at http://jrjohansson.github.com.

Python program files

- Python code is usually stored in text files with the file ending ".py":

  ```
  myprogram.py
  ```

- Every line in a Python program file is assumed to be a Python statemen
  - The only exception is comment lines, which start with the charact
    Comment lines are usually ignored by the Python interpreter.
- To run our Python program from the command line we use:

  ```
  $ python myprogram.py
  ```

- On UNIX systems it is common to define the path to the interpreter on the first line of the program (note that this is a comment line as far as the Python interpreter is concerned):

  ```
  #!/usr/bin/env python
  ```

If we do, and if we additionally set the file script to be executable, we can run the program like this:

```
$ myprogram.py
```

Example:

```
In [1]: ls scripts/hello-world*.py
scripts/hello-world-in-swedish.py  scripts/hello-world.py
```

Intro to Python Lab this Friday!! 10-11:30 am, MD G115

You probably all know the default Python interpreter.

Don't bother with

```
Last login: Mon May 20 17:53:32 on ttys000
dn0a2100e6:~ mike$ python
Enthought Python Distribution -- www.enthought.com
Version: 7.3-2 (32-bit)

Python 2.7.3 |EPD 7.3-2 (32-bit)| (default, Apr 12 2012, 11:28:34)
[GCC 4.0.1 (Apple Inc. build 5493)] on darwin
Type "credits", "demo" or "enthought" for more information.
>>> 2 + 2
4
>>>
```

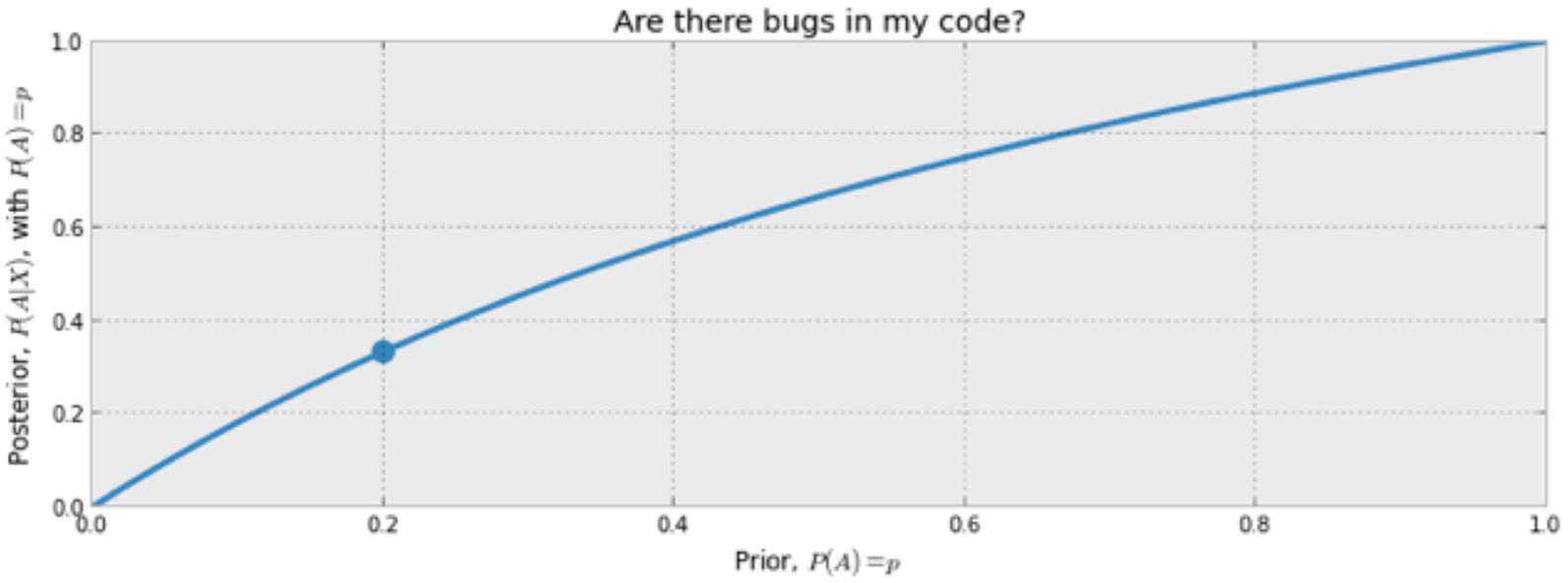
IPython is a more powerful interactive Python interpreter.

write and execute Python code in *snippets*

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)} = \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```



[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

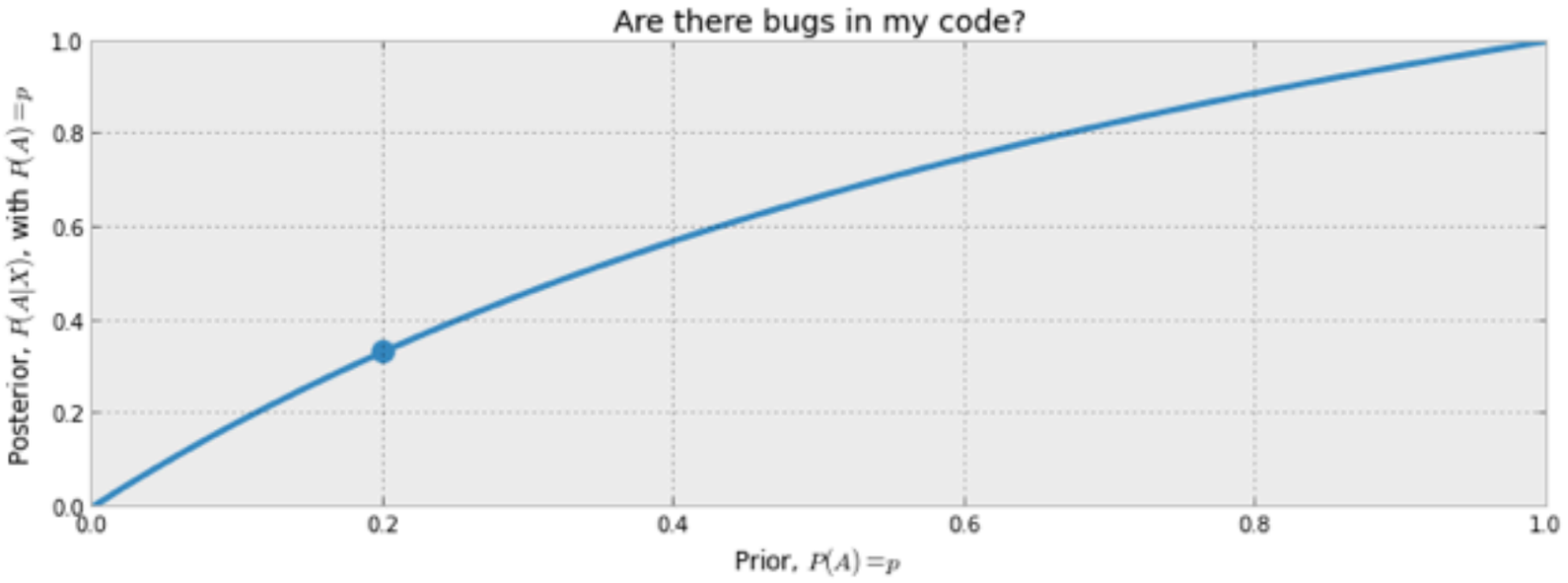
IPython is a more powerful interactive Python interpreter.

comments can include Latex math

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)} = \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```



[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

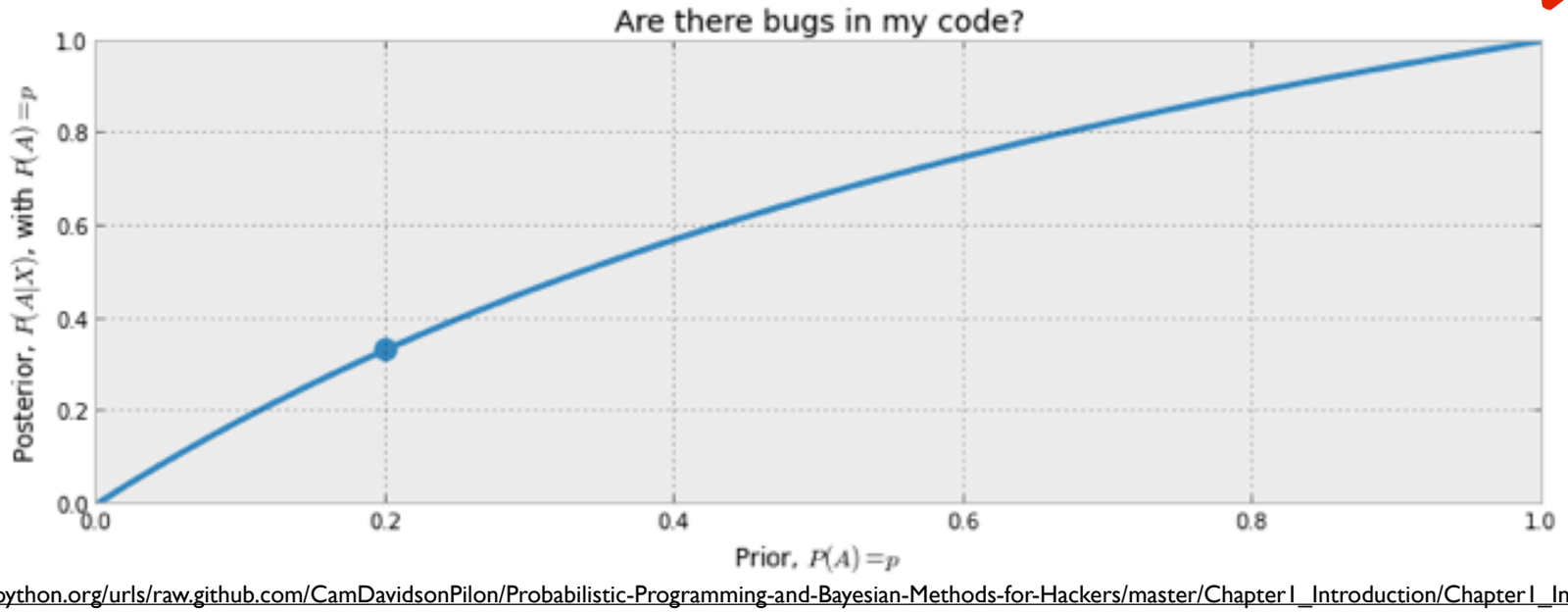
IPython is a more powerful interactive Python interpreter.

any plots generated by your code are displayed inline

$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)} = \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5,4)
p = np.linspace( 0,1, 50)
plt.plot( p, 2*p/(1+p), color = "#348ABD", lw = 3 )
#plt.fill_between( p, 2*p/(1+p), alpha = .5, facecolor = ["#A60628"])
plt.scatter( 0.2, 2*(0.2)/1.2, s = 140, c ="#348ABD"  )
plt.xlim( 0, 1)
plt.ylim( 0, 1)
plt.xlabel( "Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title( "Are there bugs in my code?");
```



[http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter1_Introduction/Chapter1_Introduction.ipynb]

Write Python code interactively in a web browser instead of a terminal window.

IPy IPython Dashboard | IPy HW2 (Gaussian Blur)

127.0.0.1:8888/d795d6fe-e6bb-44a4-a11c-f947e19a2560

IP[y]: Notebook HW2 (Gaussian Blur) Last saved: May 21 11:24 AM

File Edit View Insert Cell Kernel Help

Code

common gaussian blur code

We begin by defining the non-normalized Gaussian Function $G(x, y)$ as follows:

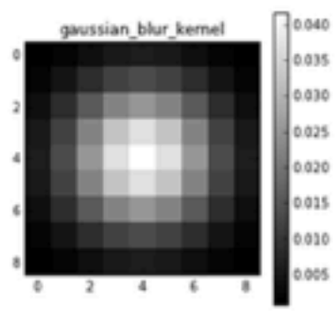
$$G(x, y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$$

We then define the normalized Gaussian Function as $N(x, y) = cG(x, y)$, where c is a normalization constant:

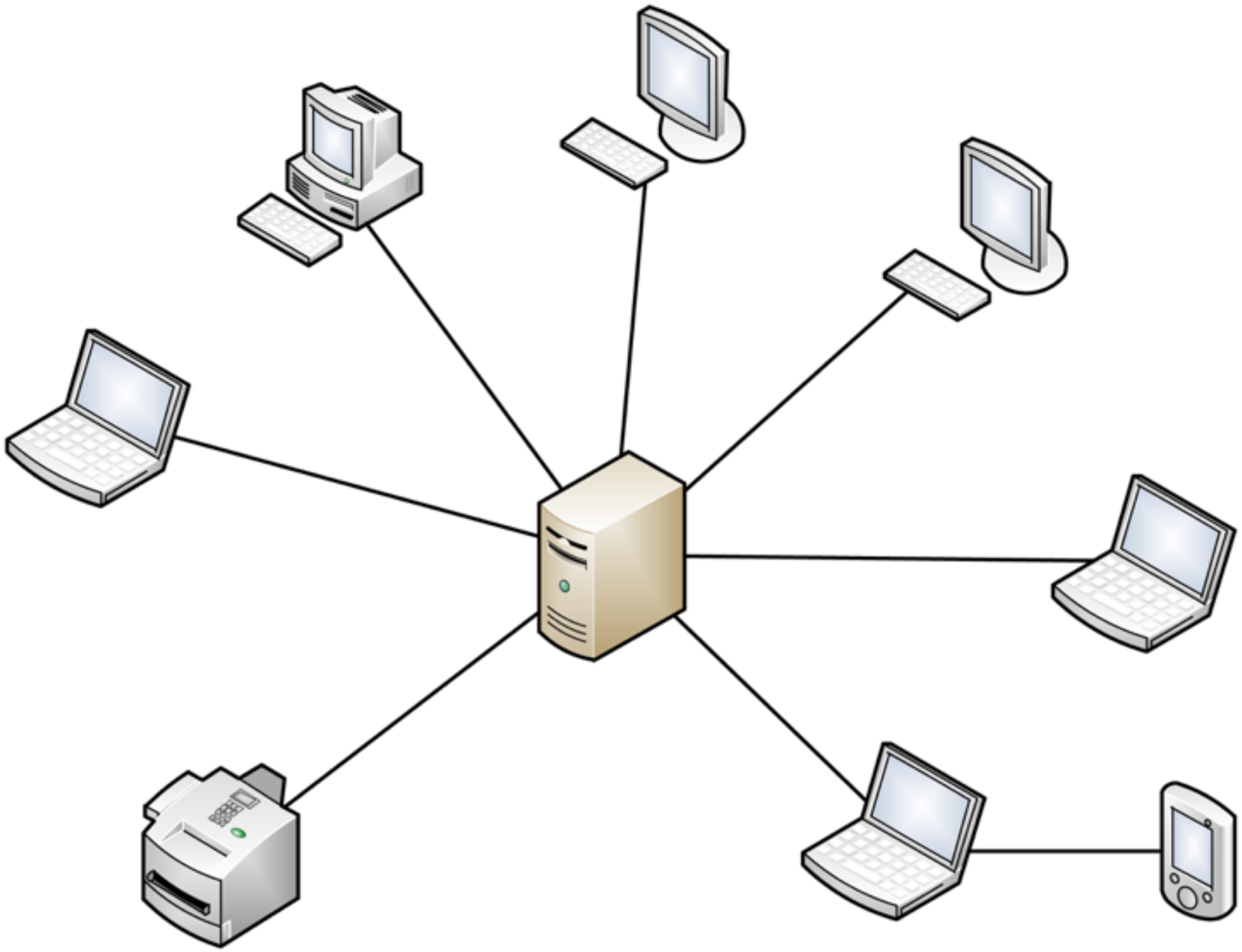
$$c = \frac{1}{\int\int G(x, y)\mathrm{d}x\mathrm{d}y}$$

To perform a Gaussian Blur, we use $N(x, y)$ as a convolution kernel.

In [4]:
```
import scipy

gaussian_blur_kernel_width      = numpy.int32(9)
gaussian_blur_kernel_half_width = numpy.int32(4)
gaussian_blur_sigma             = numpy.float32(2)

y, x = \
    scipy.mgrid[-gaussian_blur_kernel_half_width:gaussian_blur_kernel_half_width+1,-gaussian_blur_kernel_half_width:gaussian_blur_kernel_half_width+1]

gaussian_blur_kernel_not_normalized = numpy.exp( ( - ( x**2 + y**2 ) ) / ( 2 * gaussian_blur_sigma**2 ) )
normalization_constant              = numpy.float32(1) / numpy.sum(gaussian_blur_kernel_not_normalized)
gaussian_blur_kernel                = (normalization_constant * gaussian_blur_kernel_not_normalized).astype(numpy.float32)



matplotlib.pyplot.imshow(gaussian_blur_kernel, cmap="gray", interpolation="nearest");
matplotlib.pyplot.title("gaussian_blur_kernel");
matplotlib.pyplot.colorbar();
```

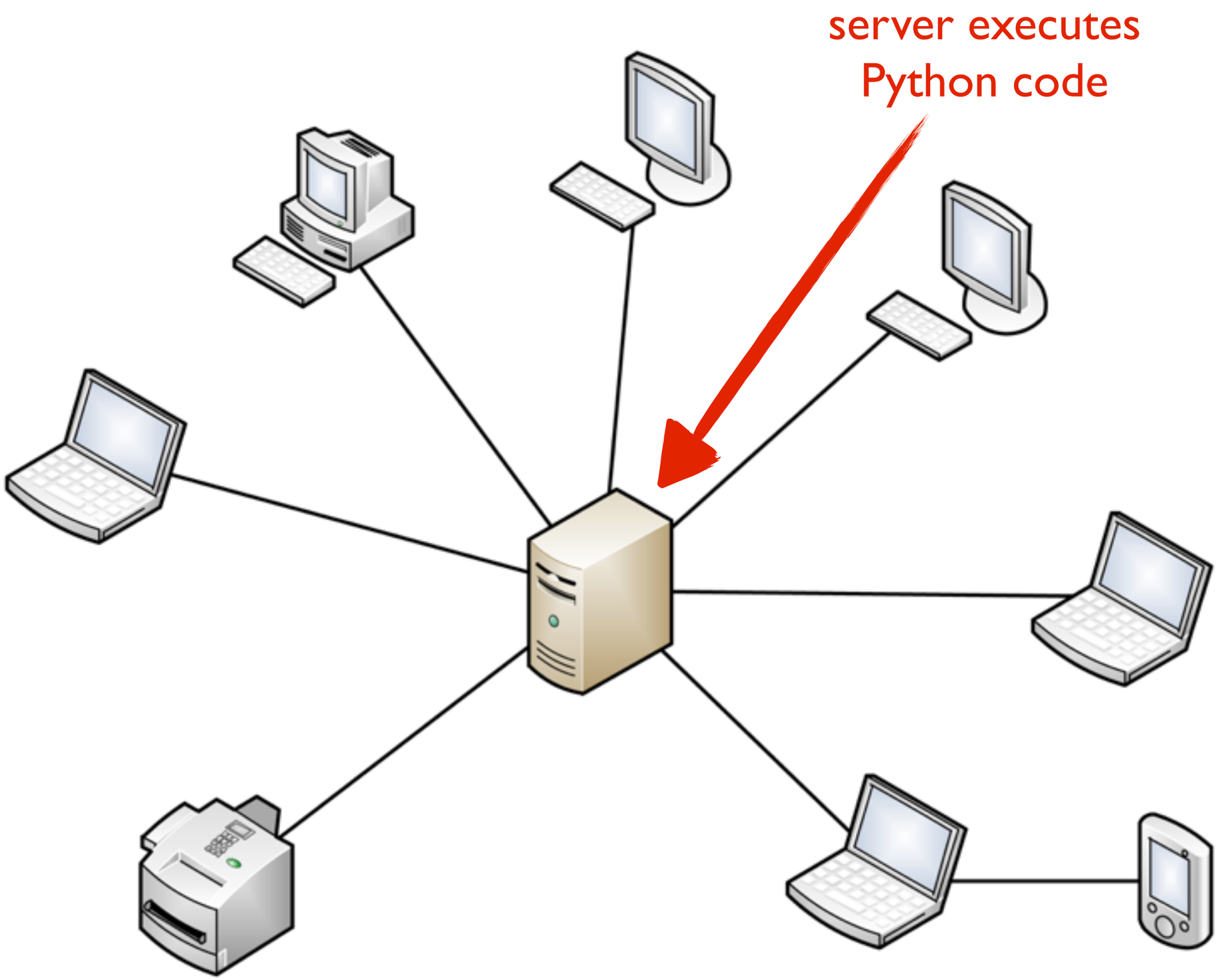


IPython has a decoupled client-server architecture.



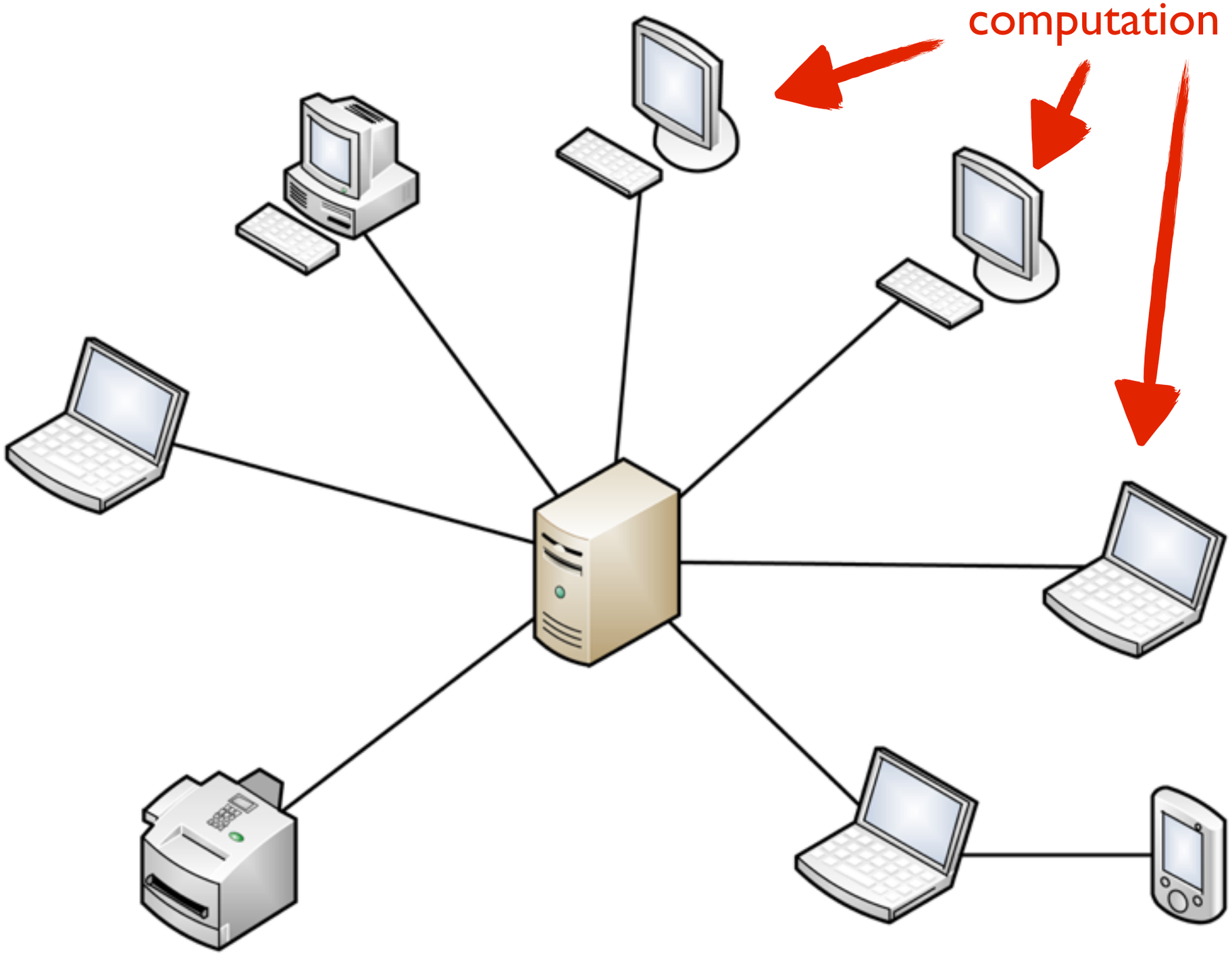
[http://communities.intel.com/community/datastack/blog/2011/05/02/top-10-reasons-to-setup-a-client-server-network]

IPython has a decoupled client-server architecture.



[http://communities.intel.com/community/datastack/blog/2011/05/02/top-10-reasons-to-setup-a-client-server-network]

IPython has a decoupled client-server architecture.

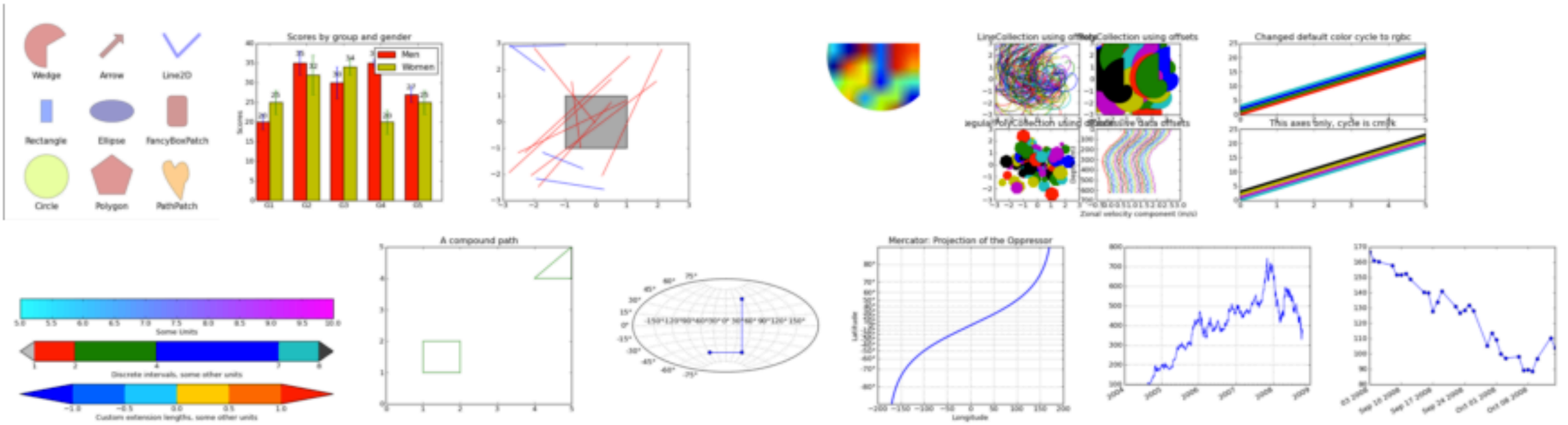


[http://communities.intel.com/community/datastack/blog/2011/05/02/top-10-reasons-to-setup-a-client-server-network]

You can stay productive on any computer with an internet connection and a web browser.

[image source unknown]

IPython integrates seamlessly with Matplotlib, making it well-suited for data exploration.



[http://matplotlib.org/gallery.html]

Examples

[image source unknown]

Resources

- http://nbviewer.ipython.org/4542975
- https://github.com/mroberts3000/IPythonIsGreat
- http://nbviewer.ipython.org/
- https://github.com/jrjohansson/scientific-python-lectures
- http://scipy-lectures.github.io/index.html
- Many more online and on the course web site

Good Practices

- IPython notebooks to document your process
- Visualizations for data exploration
- Comment your code!
- Modularity - breaking down code into small functional, composable pieces
- Array-oriented computing
- Using assert statements and tests
- Version control (svn, git, github)

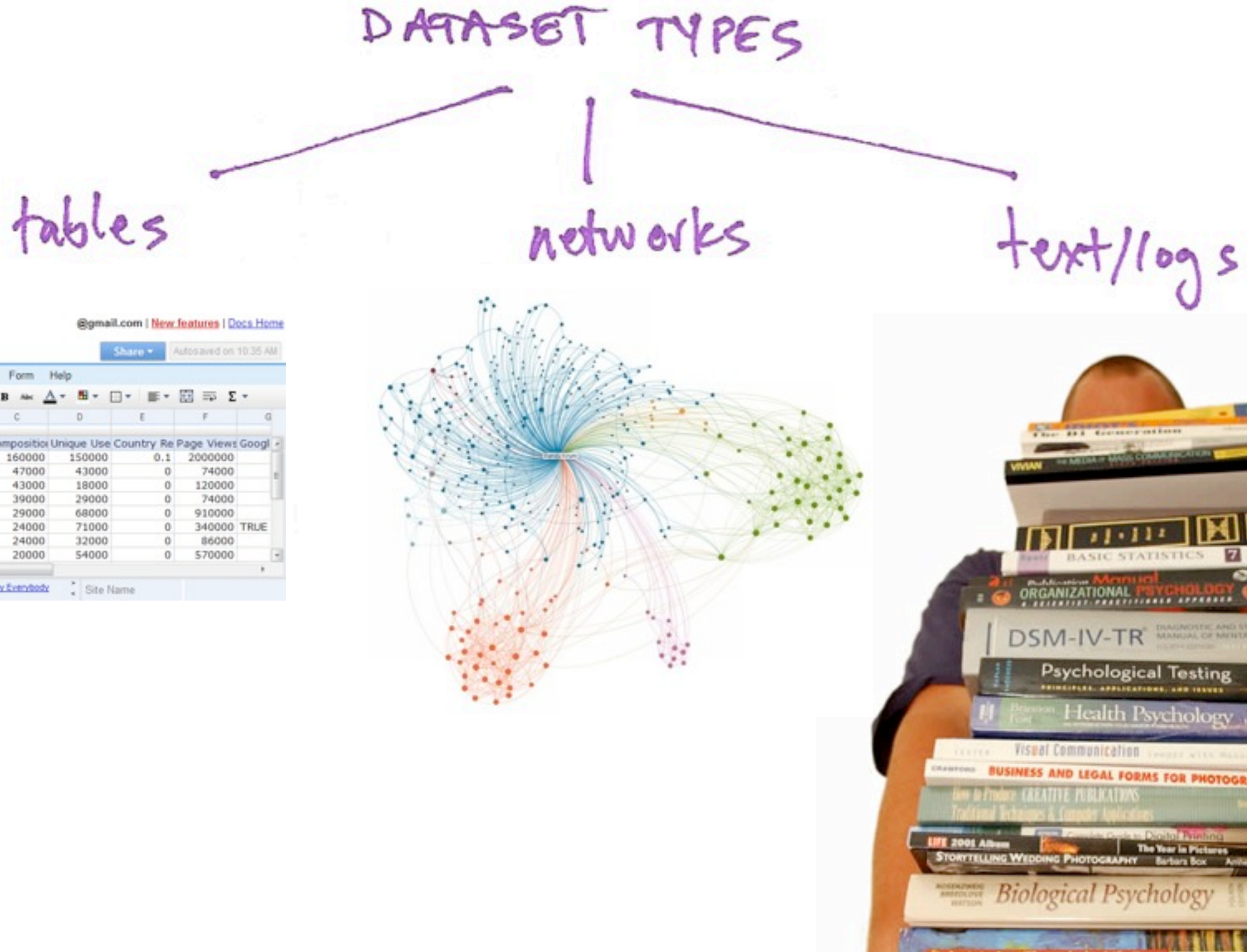
Data Types

Ben Shneiderman, 1996

- 1D (sequences)
- Temporal
- 2D (maps)
- 3D (shaped)
- nD (relational)
- Trees (hierarchical)
- Networks (graphs)
- Others?

The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization [Shneiderman, 96]

Tamara Munzner, 2013



Semantics vs. Types

- Data Semantics: The real-world meaning

  e.g., company name, day of the month, person height, etc.

- Data Type: Interpretation in terms of scales of measurements

  e.g., quantity or category, sensible mathematical operations, data structure, etc.

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
Nominal Categorical Qualitative	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
Interval	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

On the theory of scales and measurements [S. Stevens, 46]

Data Types

- Nominal (Categorical) (N)

  Are = or ≠ to other values

  Apples, Oranges, Bananas,...

- Ordinal (O)

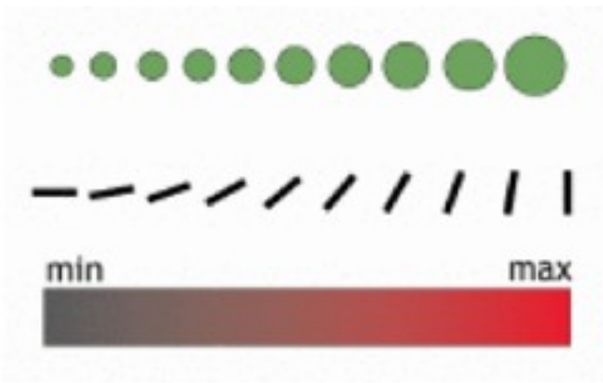
  Obey a < relationship

  Small, medium, large

- Quantitative (Q)

  Can do arithmetic on them

  10 inches, 23 inches, etc.



On the theory of scales and measurements [S. Stevens, 46]

Data Types

- Q - Interval (location of zero arbitrary)

  Dates: Jan 19; Location: (Lat, Long)

  Like a geometric point. Cannot compare directly.

  Only differences (i.e., intervals) can be compared

- Q - Ratio (zero fixed)

  Measurements: Length, Mass, Temp, ...

  Origin is meaningful, can measure ratios & proportions

  Like a geometric vector, origin is meaningful

On the theory of scales and measurements [S. Stevens, 46]

Data Types

- N - Nominal (labels)

  Operations: =, ≠

- O - Ordinal (ordered)

  Operations: =, ≠, >, <

- Q - Interval (location of zero arbitrary)

  Operations: =, ≠, >, <, +, − (distance)

- Q - Ratio (zero fixed)

  Operations: =, ≠, >, <, +, −, ×, ÷ (proportions)

On the theory of scales and measurements [S. Stevens, 46]

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Semantics

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Item

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack		2/22/08
32	7/16/07	2-High	Small Pack		7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag		10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Attribute aka Feature

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

1 = Quantitative

2 = Nominal

3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Speci		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

1 = Quantitative
2 = Nominal
3 = Ordinal

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05			44	6/6/05
69	6/4/05			.6	6/6/05
70	12/18/06			59	12/23/06
70	12/18/06			82	12/23/06
96	4/17/05			55	4/19/05
97	1/29/06			38	1/30/06
129	11/19/08			37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Nominal /Ordinal = Dimensions

Describe the data, independent variables

Quantitative = Measures

Numbers to be analyzed, dependent variables

Data vs. Conceptual Model

- Data Model: Low-level description of the data

  Set with operations, e.g., floats with +, -, /, *

- Conceptual Model: Mental construction

  Includes semantics, supports reasoning

Data	Conceptual
1D floats	temperature
3D vector of floats	space

Data vs. Conceptual Model

- From data model...

  32.5, 54.0, -17.3, … (floats)

- using conceptual model...

  Temperature

- to data type

  Continuous to 4 significant figures (Q)

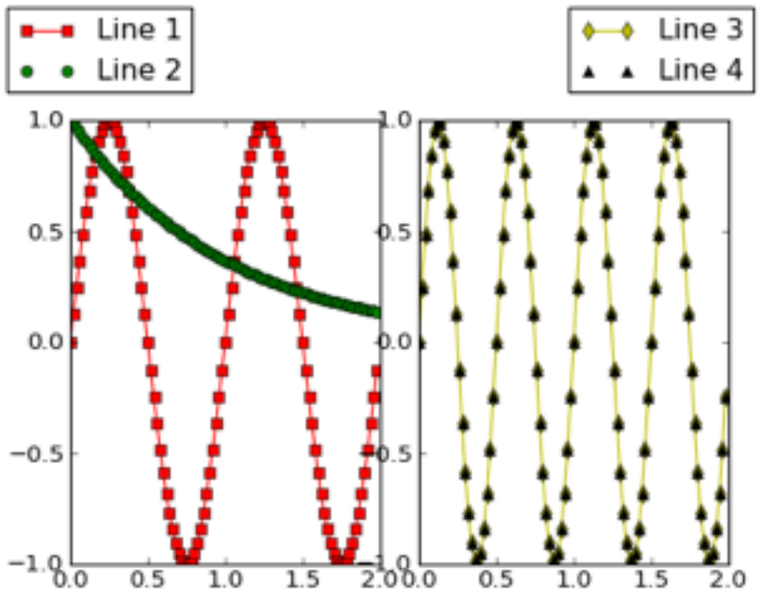
  Hot, warm, cold (O)

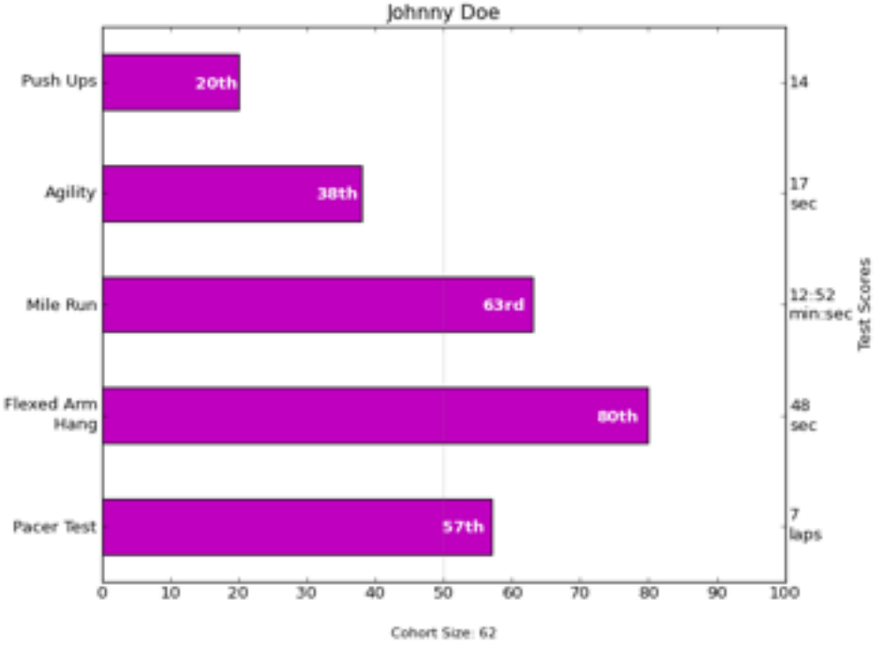
  Burned vs. Not burned (N)

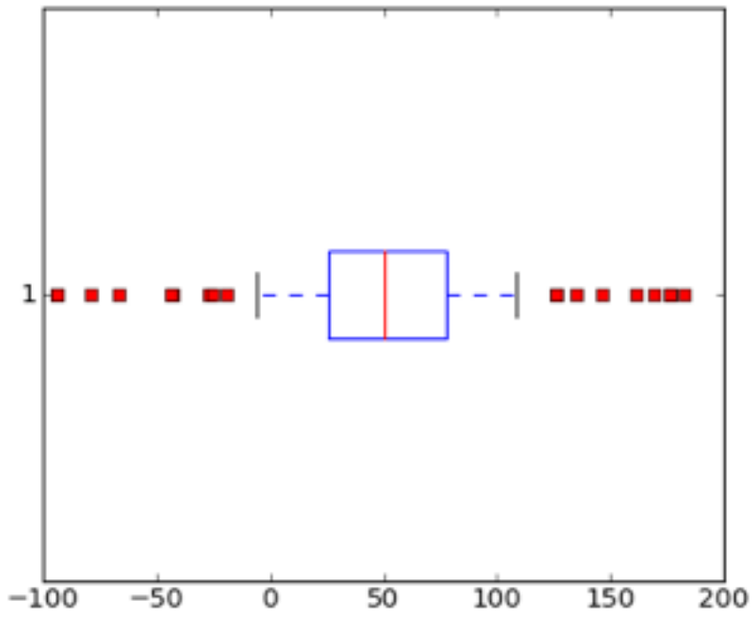
Based on slide from Munzner

Data Dimensions

Univariate Data



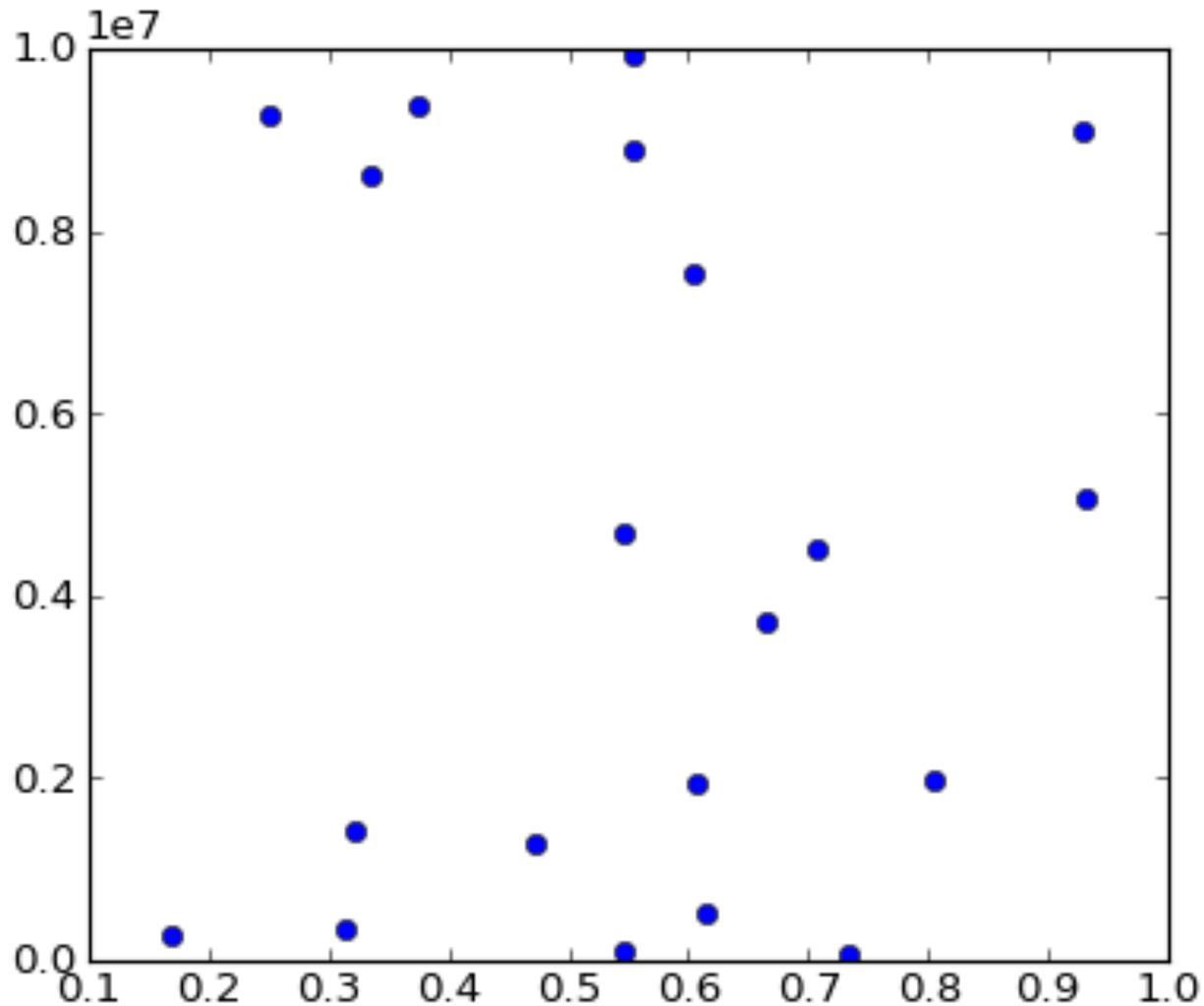




Based on slide from M. Agrawala

Bivariate Data

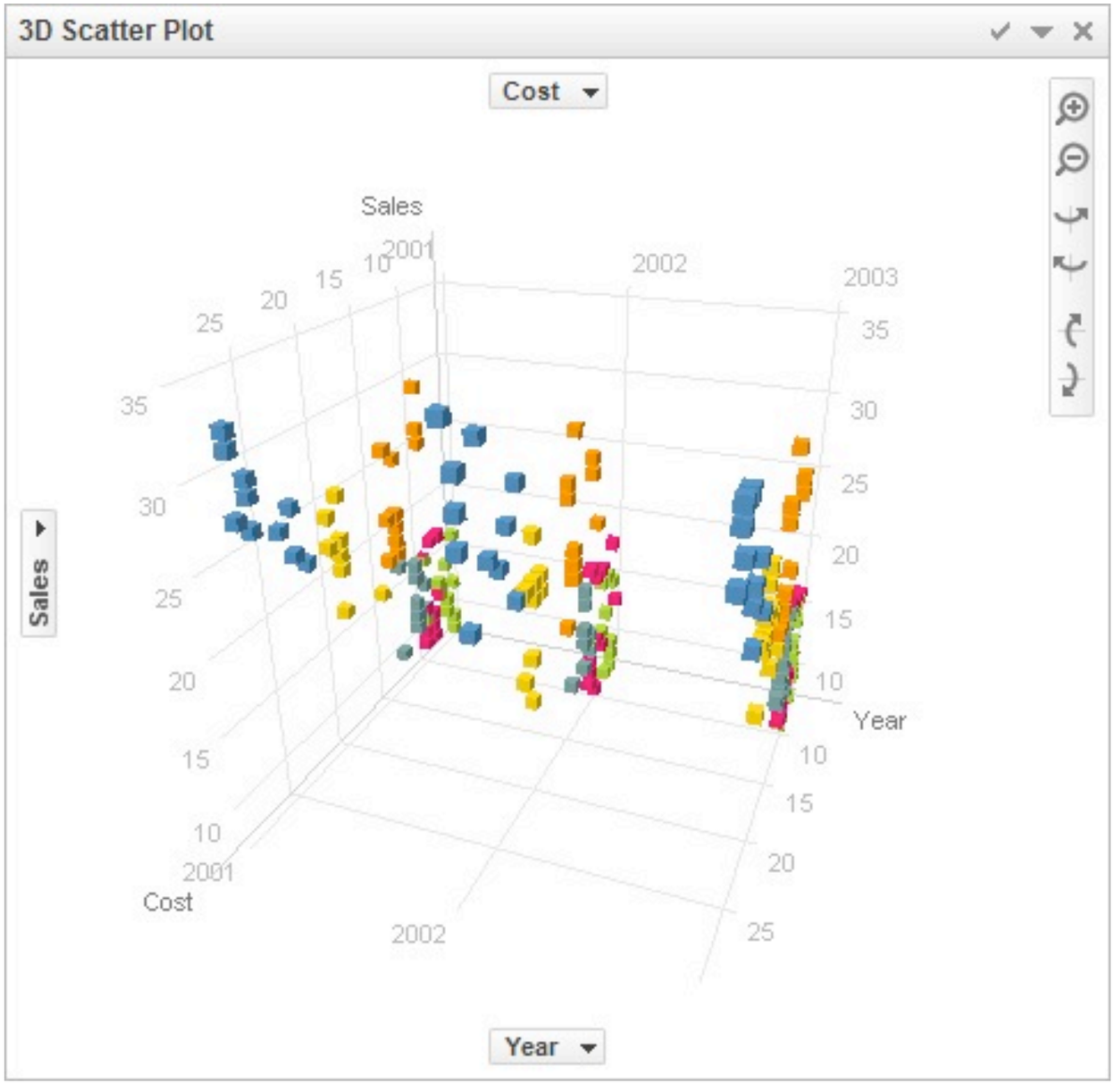
Scatterplot is common



Based on slide from M. Agrawala

Trivariate Data

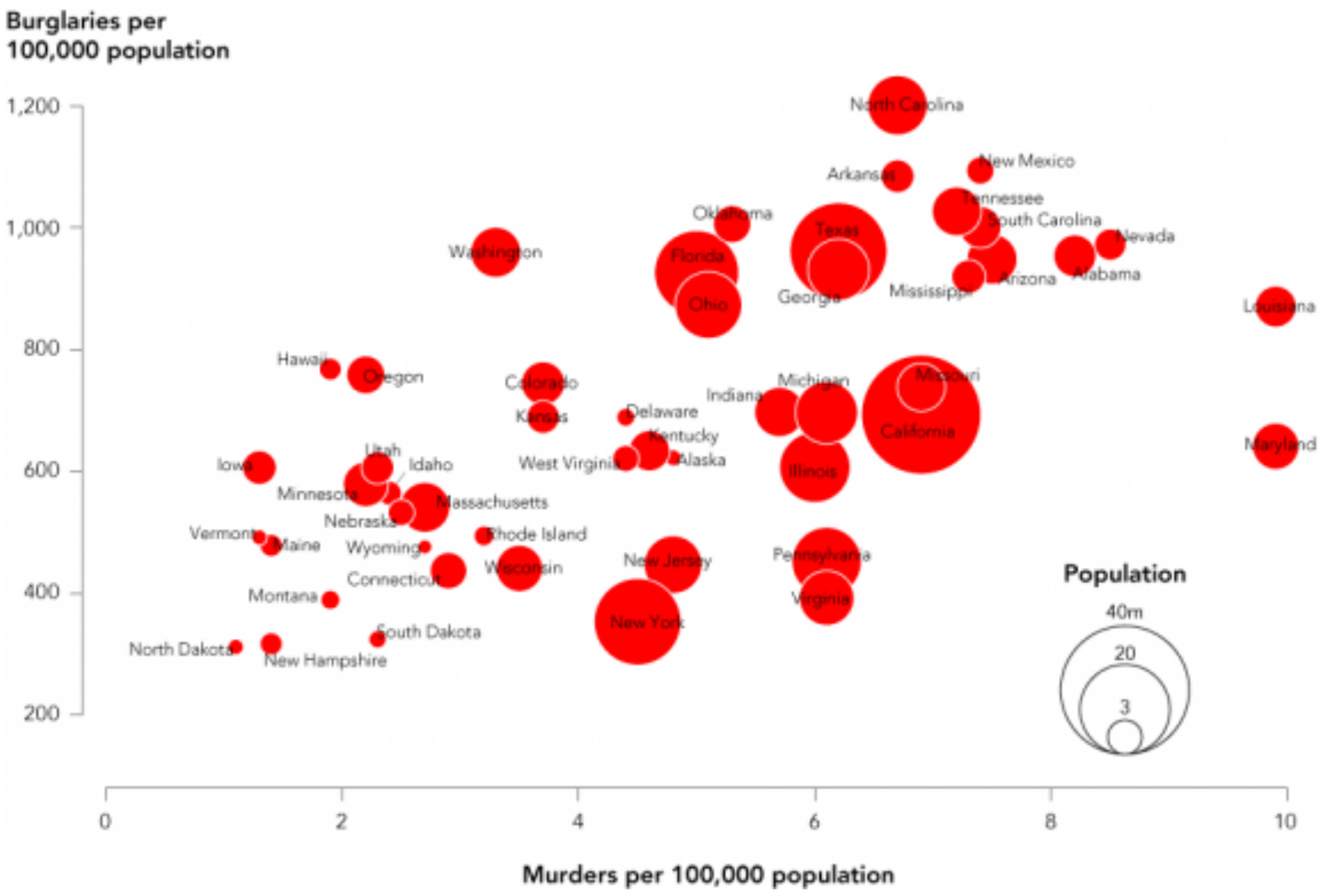
Do NOT use 3D scatterplots!



Based on slide from M. Agrawala

Trivariate Data

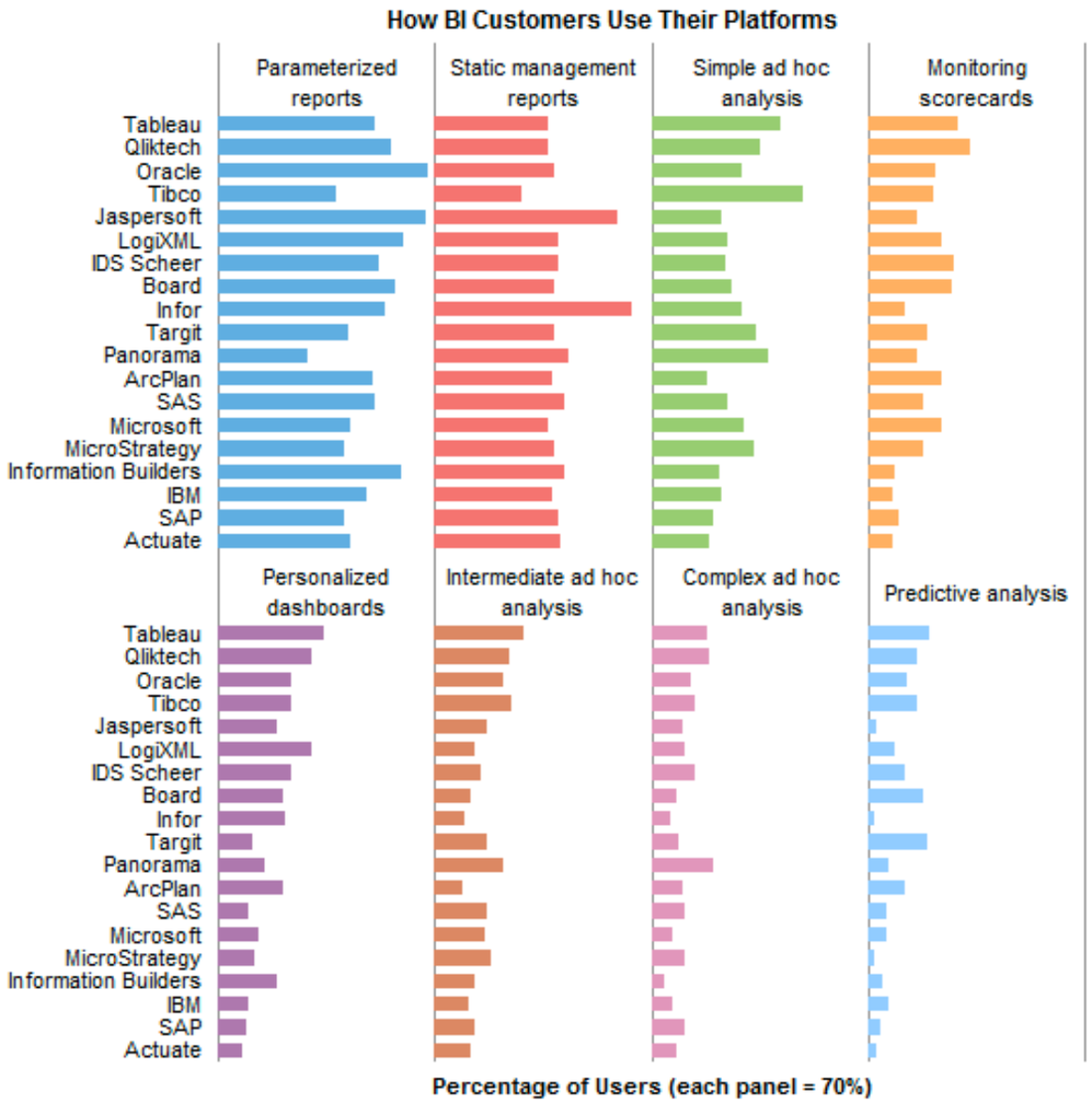
Map the third dimension to some other visual attribute



Based on slide from M. Agrawala

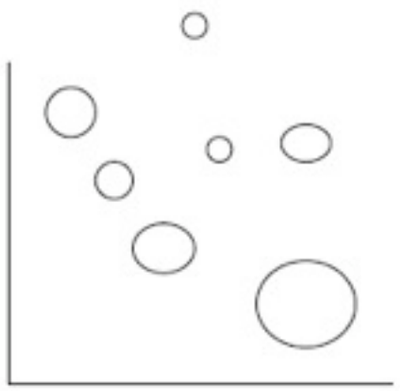
Multivariate Data

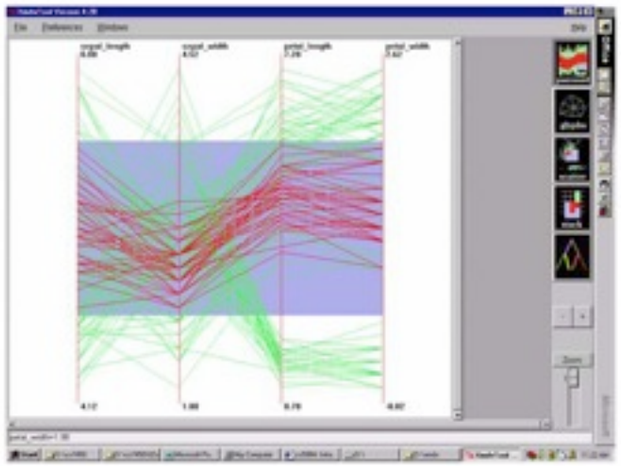
Give each attribute its own display (small multiples)

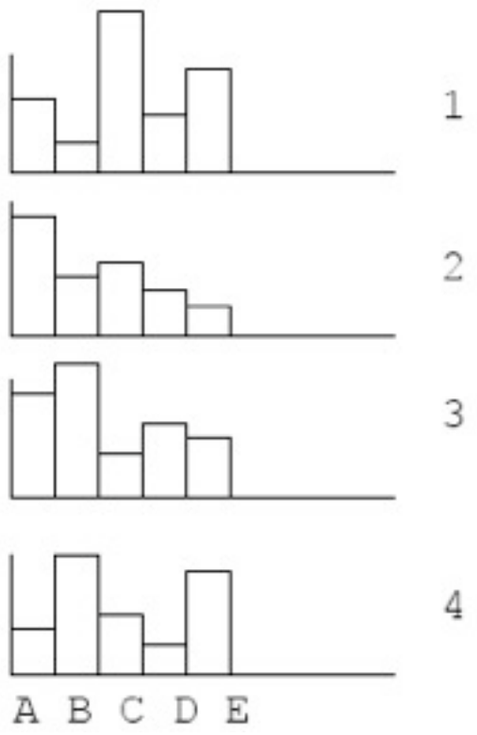


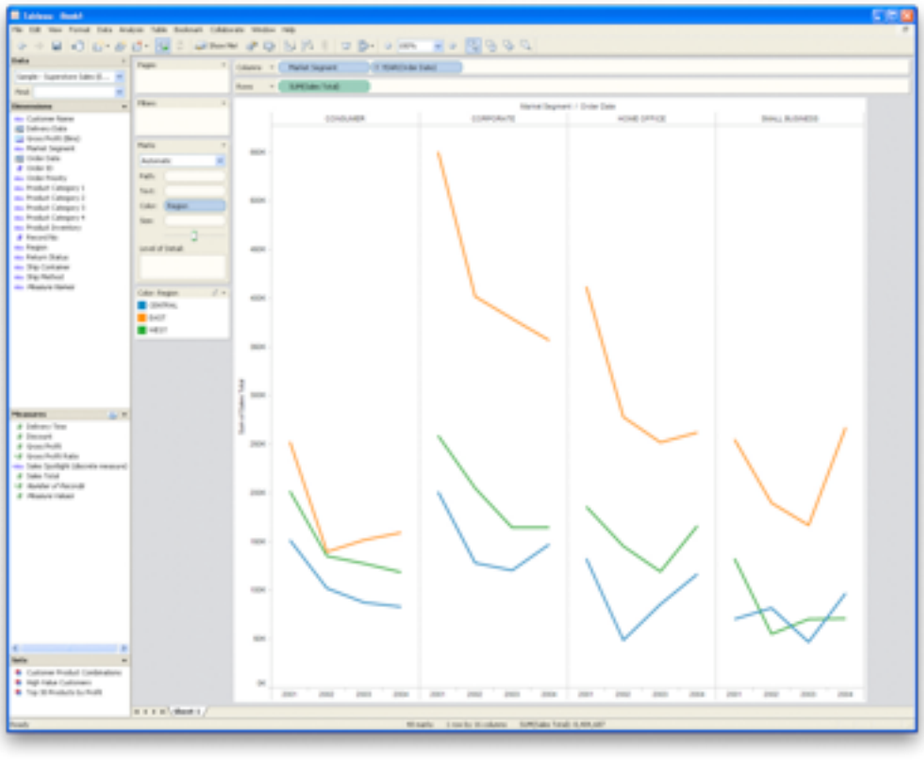
Based on slide from M. Agrawala

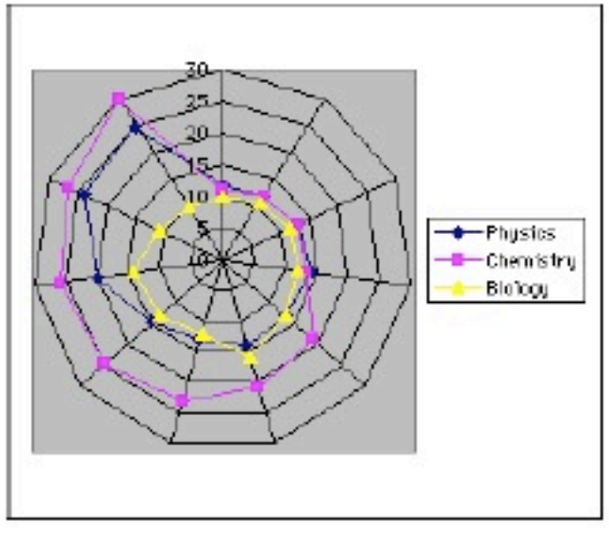
Multivariate Data Representations











Data Reduction

- Filtering: Eliminate some items or attributes

  e.g., select range of interest, zoom in, remove outliers, etc.

- Aggregation: Represent a group of elements by a new derived element

  e.g., take average, min, max, count, sum

  Attribute aggregation a.k.a. dimensionality reduction

Mapping Data to Images

Jacques Bertin

- French cartographer [1918-2010]
- Semiology of Graphics [1967]
- Theoretical principles for visual encodings

Bertin's Visual Attributes

Marks Points Lines Areas

Channels

Position

Size

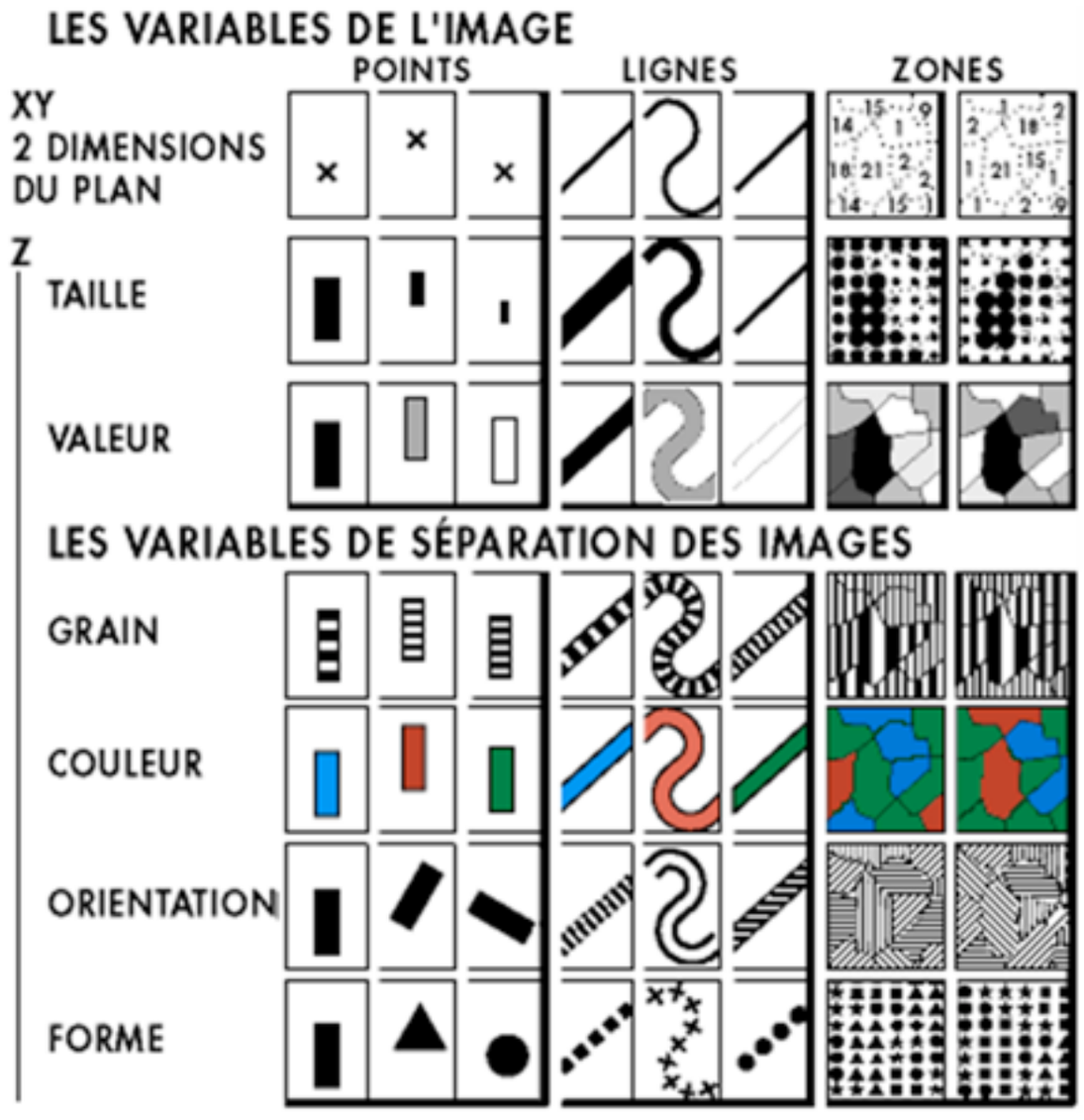
(Grey)Value

Texture

Color

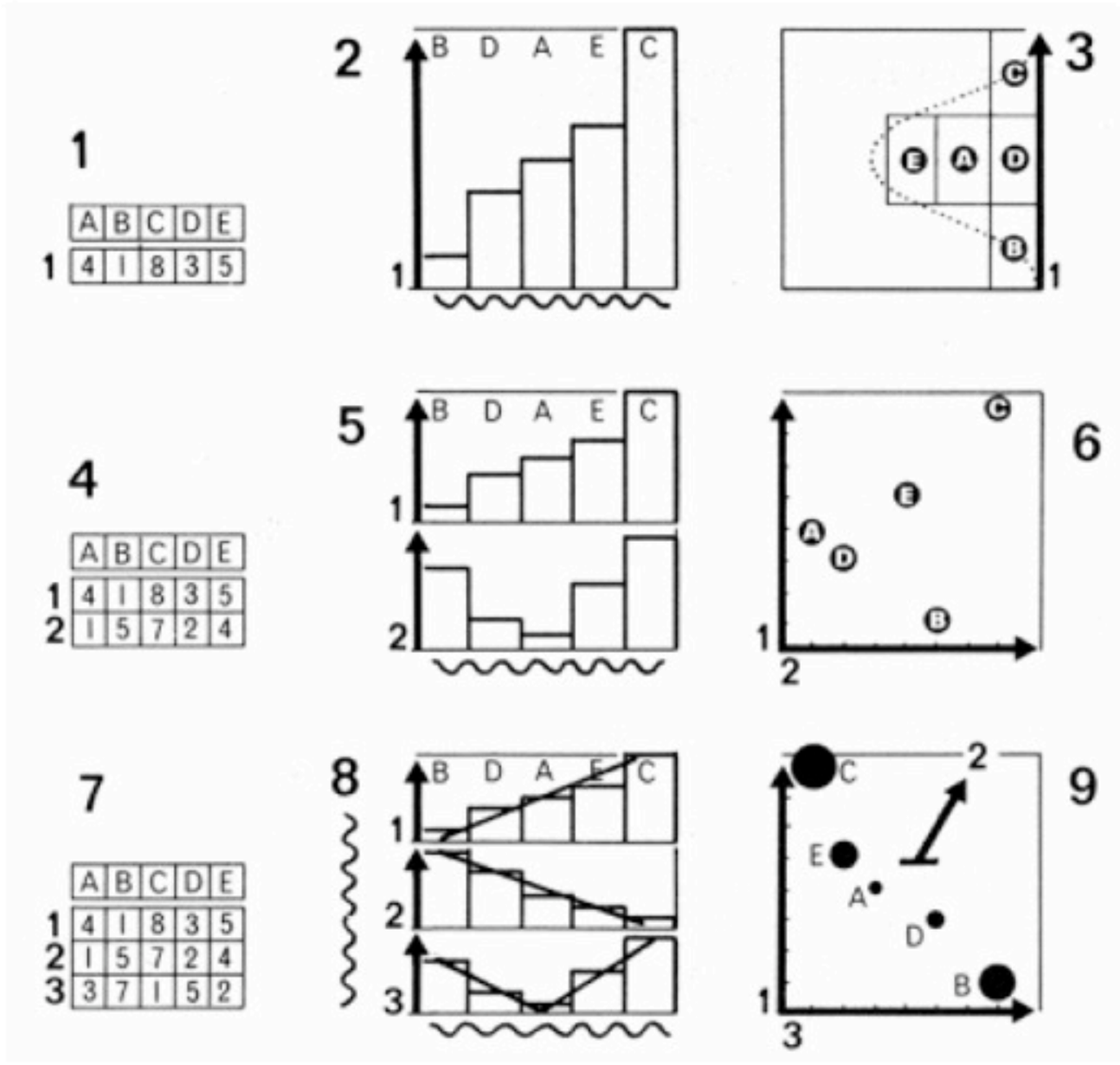
Orientation

Shape



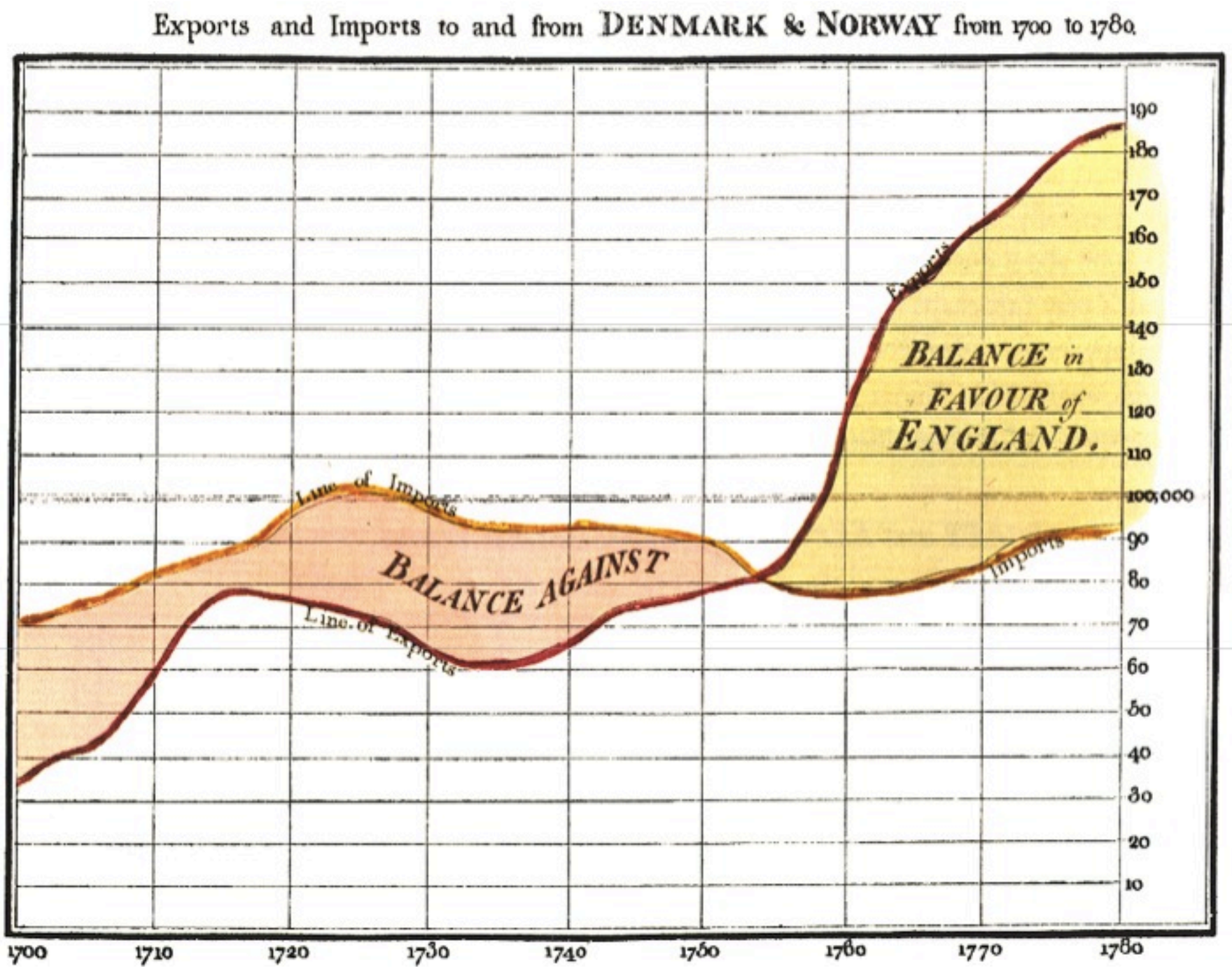
Bertin, Semiology of Graphics, 1967

Large Design Space (Visual Metaphors)



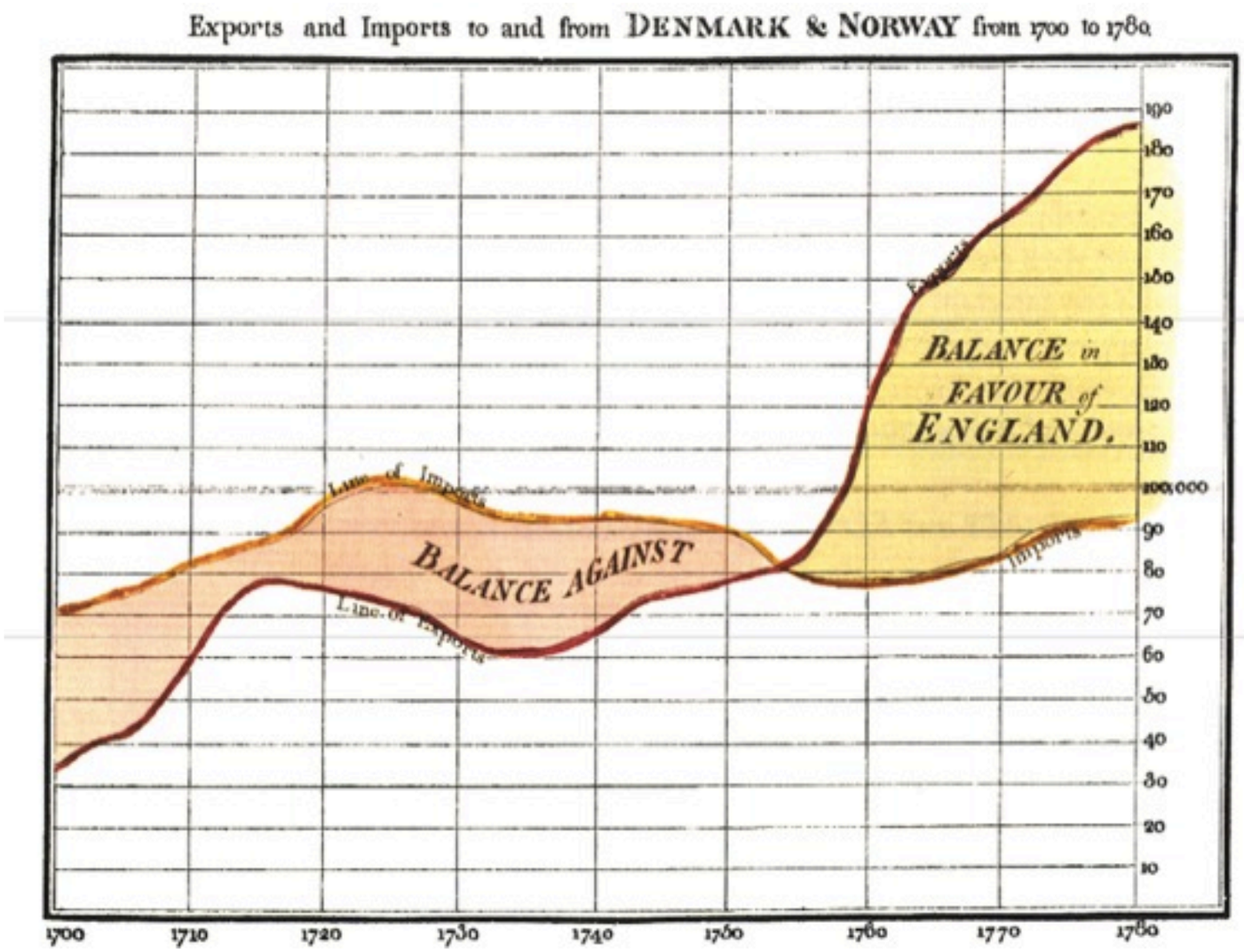
J. Bertin, "Graphics and Graphic Information Processing", 1981

Example



W. Playfair, 1786

Example

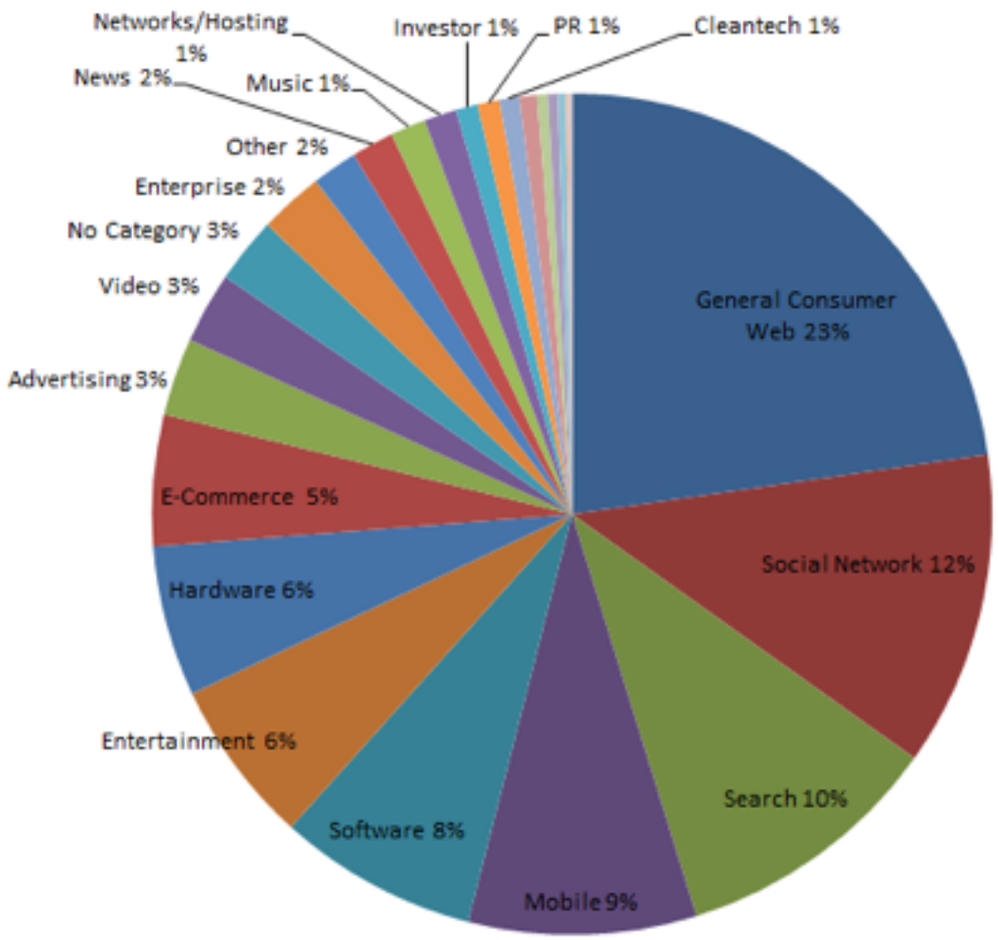


x-axis: Year (Q)
y-axis: Currency (Q)
Color: Imports / Exports (N, O)

W. Playfair, 1786

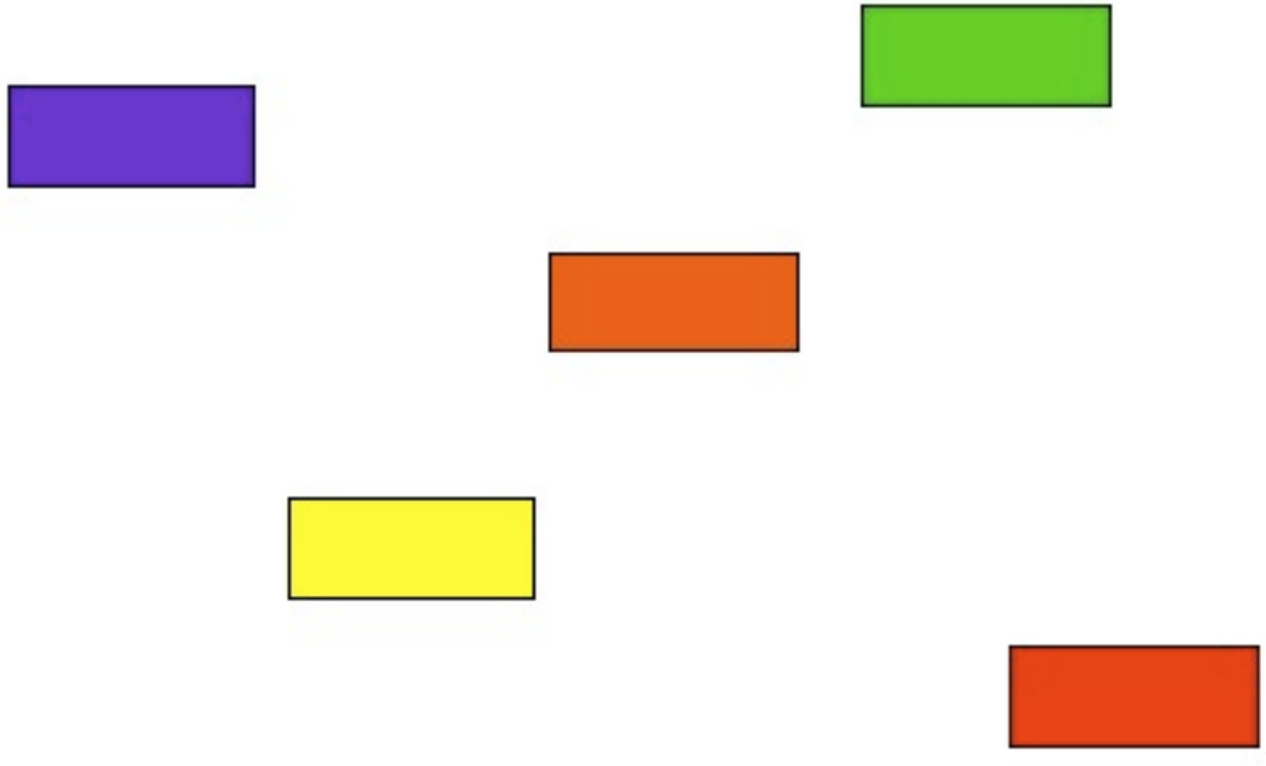
Effective Visual Attributes

Compare These Values



Cole Nussbaumer

Order These Colors

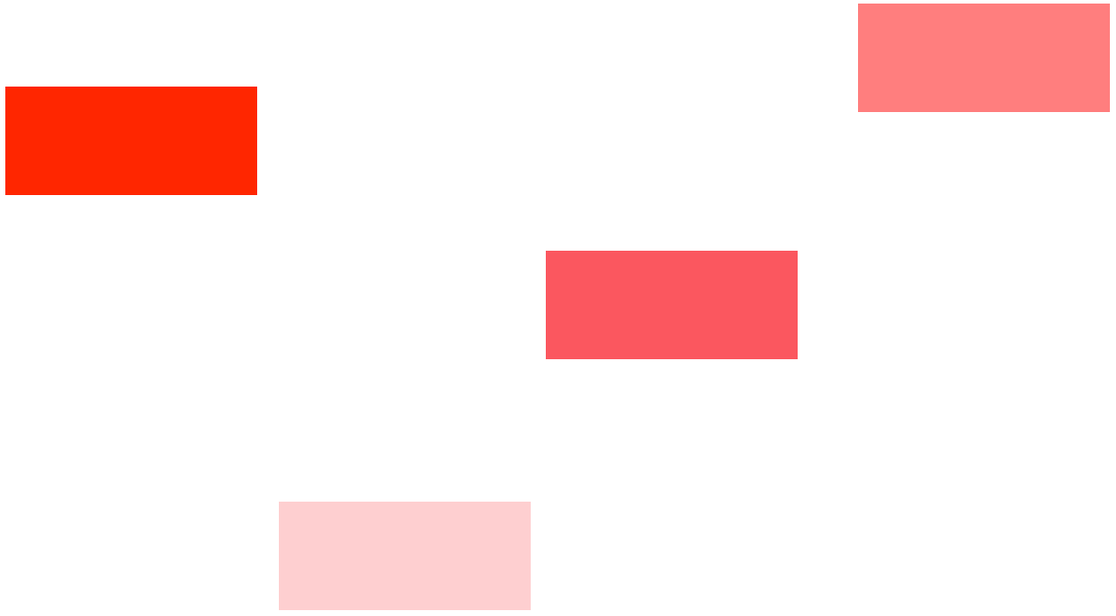


Based on slide from Stasko

Order These Colors

Based on slide from Stasko

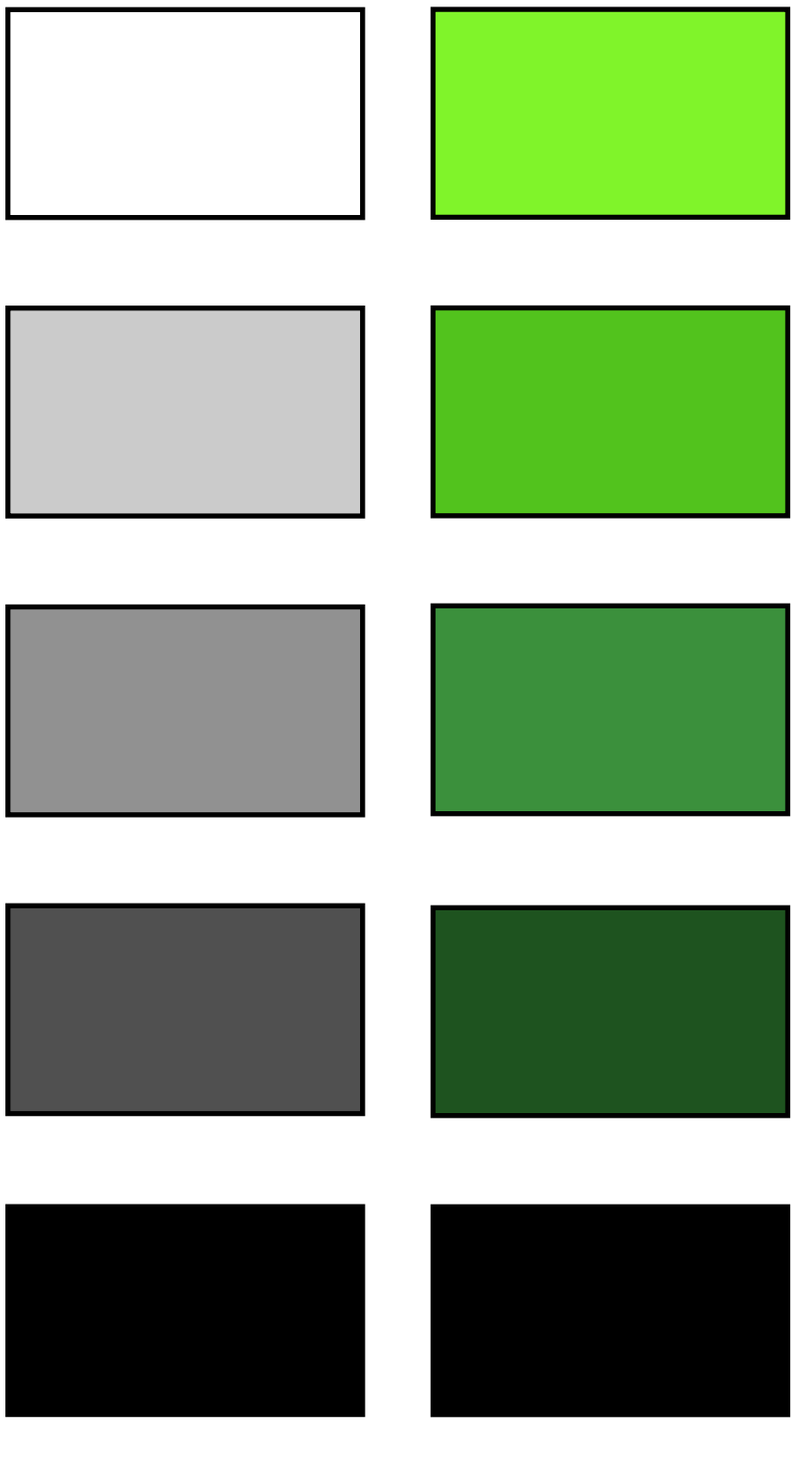
Order These Colors



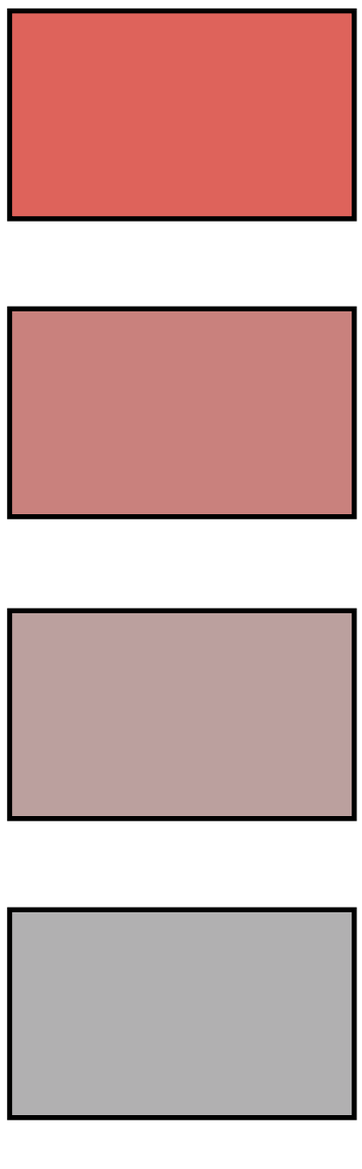
Based on slide from Stasko

Perceived as Ordered

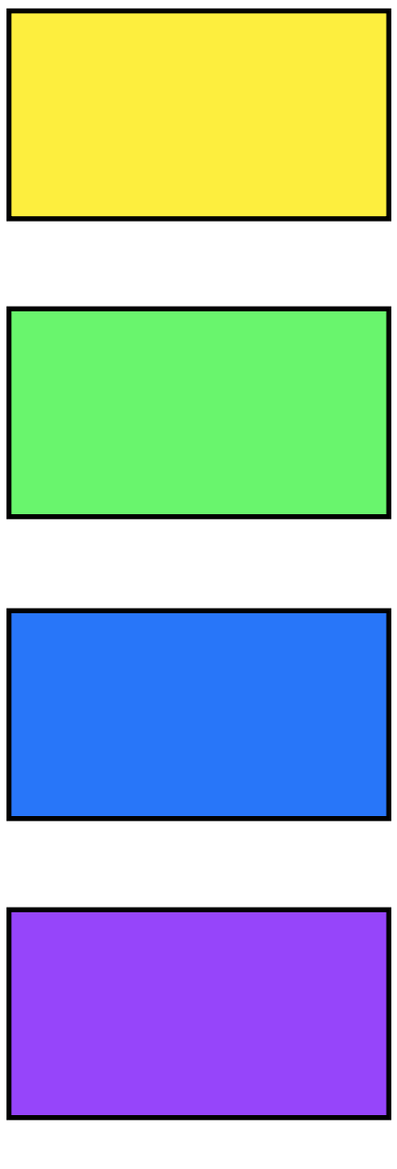
Brightness



Saturation



Hue: not as much



Visual Attributes per Data Type

Bertin, 1967

	Categories	Ordinal	Quantitative
Position	✔	✔	✔
Length	✔	✔	✔
Brightness	✔	✔	~
Texture	✔	~	✖
Color	✔	~	✖
Angle	✔	✖	✖
Shape	✔	✖	✖

✔ = Good
~ = OK
✖ = Bad

Bertin, Semiology of Graphics, 1967

Cleveland / McGill, 1984

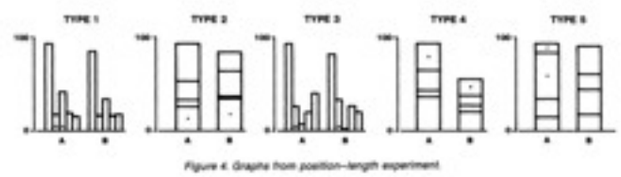


Figure 4. Graphs from position–length experiment.

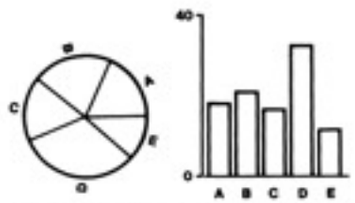
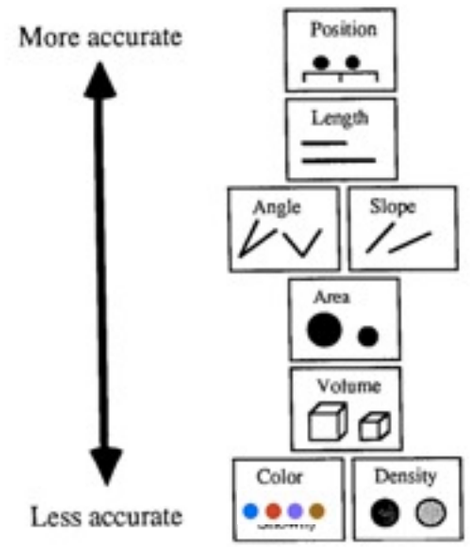


Figure 3. Graphs from position–angle experiment.

William S. Cleveland; Robert McGill , "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." 1984

Mackinlay, 1986



Jock Mackinlay "Automating The Design of Graphical Presentations." 1986

Most Efficient

Position

Length

Slope

Angle

Quantitative

Area

Intensity

Ordinal

Least Efficient

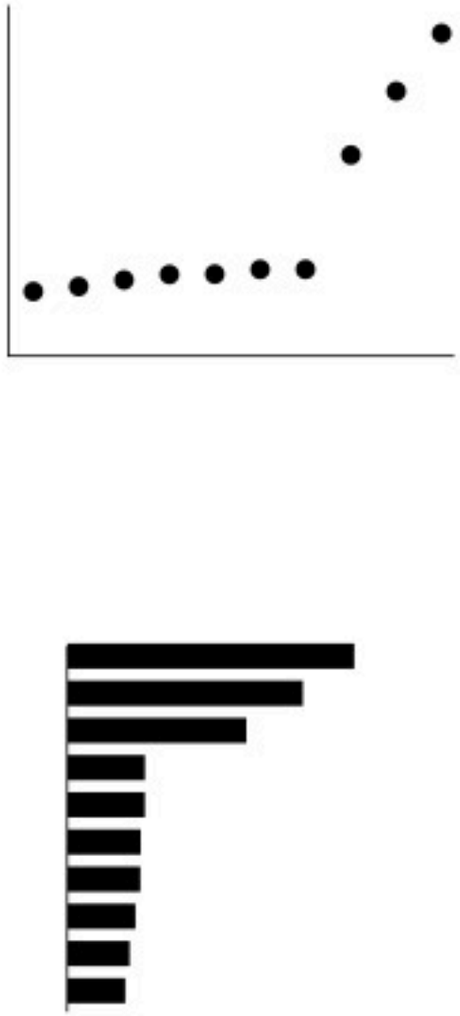
Color

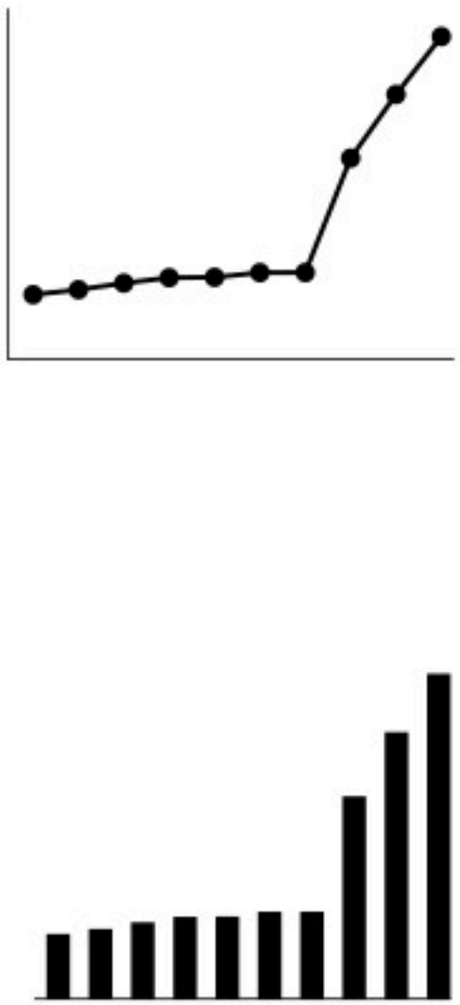
Shape

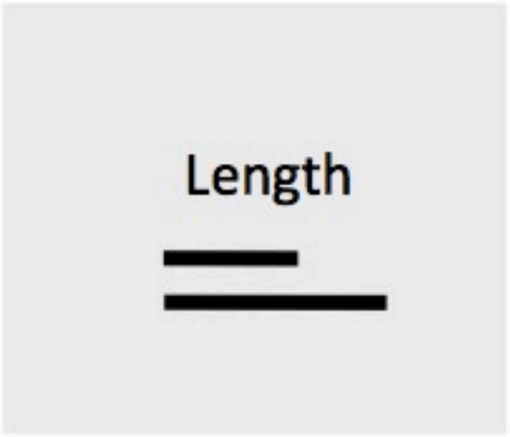
Nominal

C. Mulbrandon
VisualizingEconomics.com

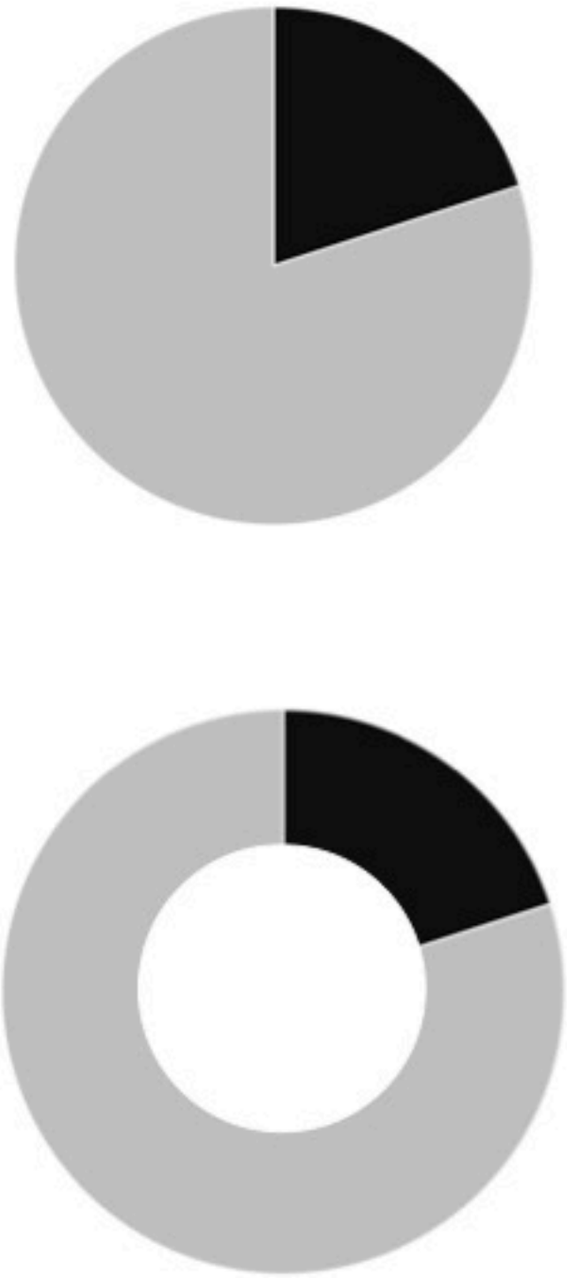
Most Effective

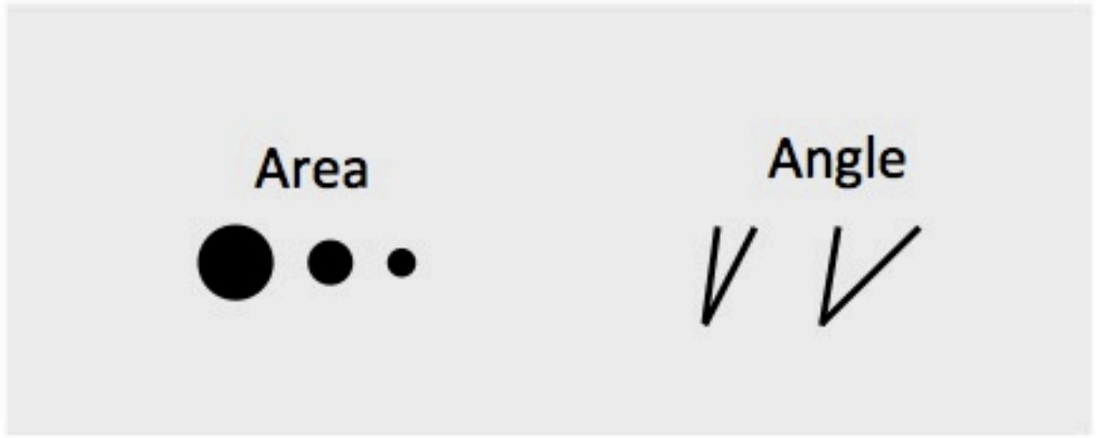


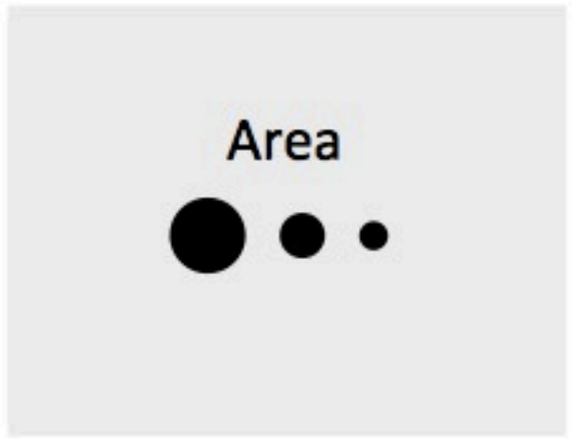




Less Effective







Least Effective

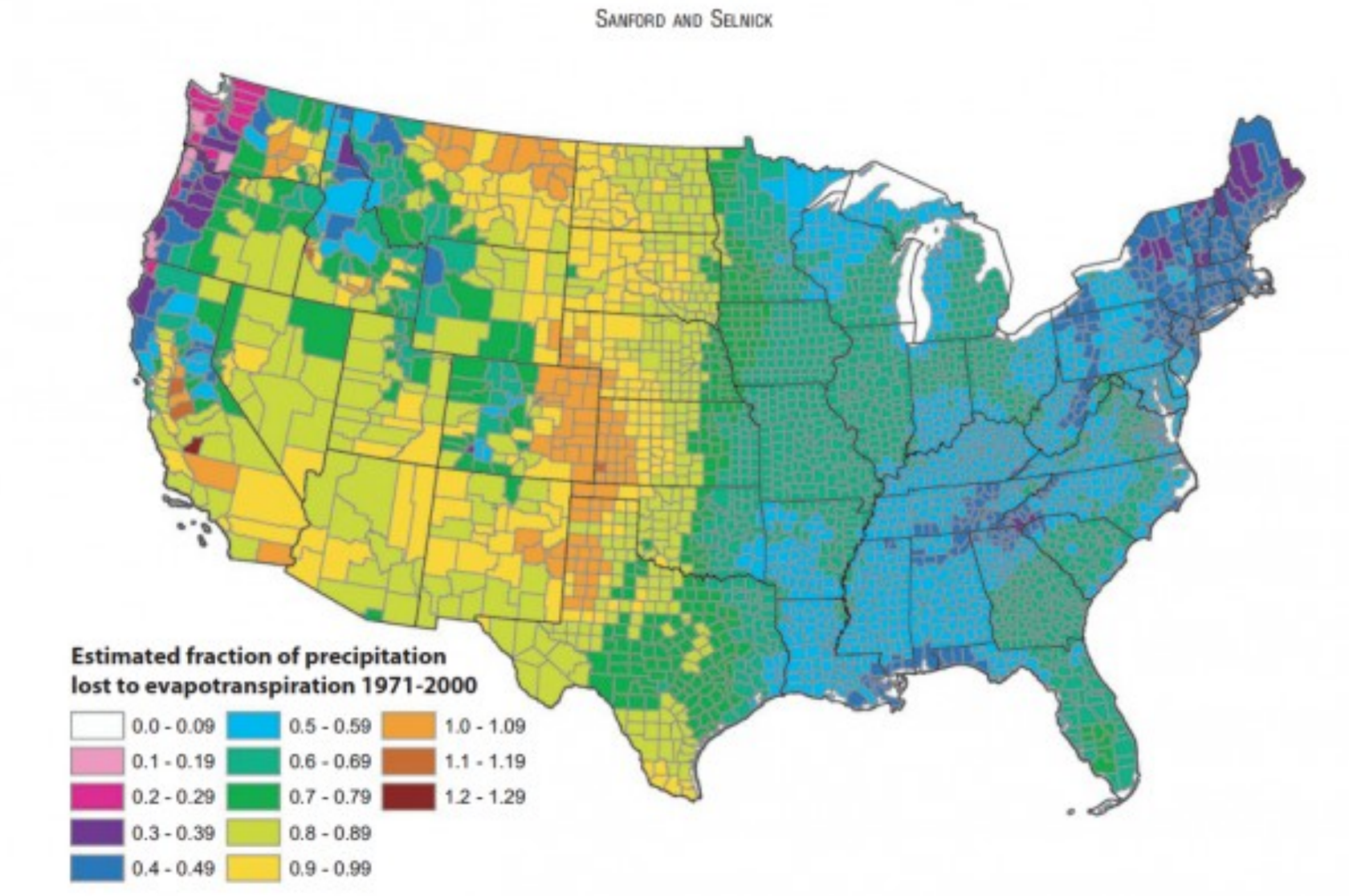
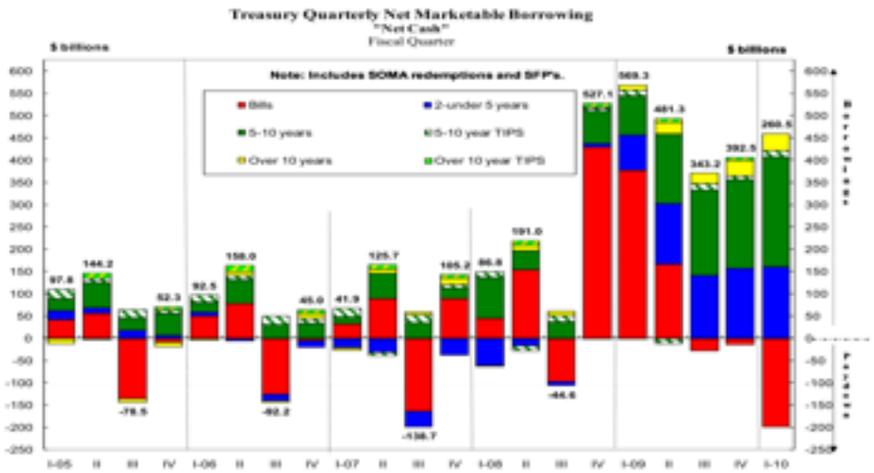


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (*P*) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/*P* were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Cliff Mass

Effective Visualizations

Not Effective....



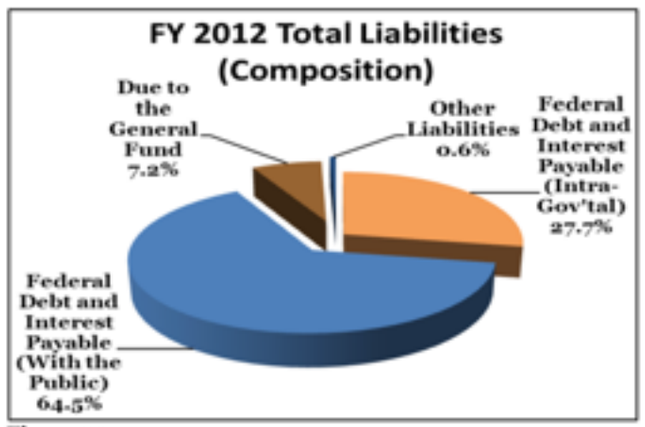
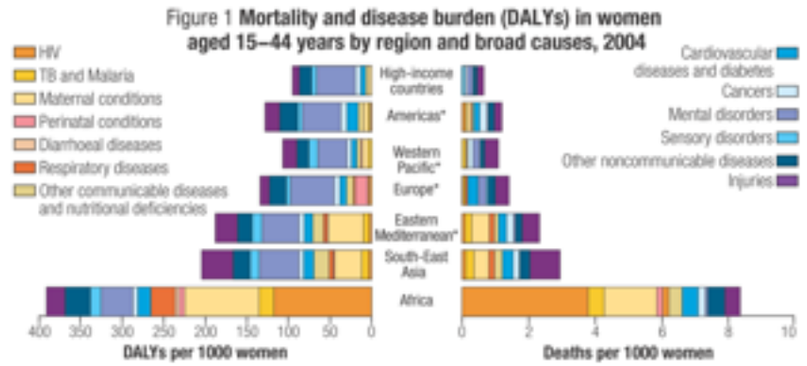
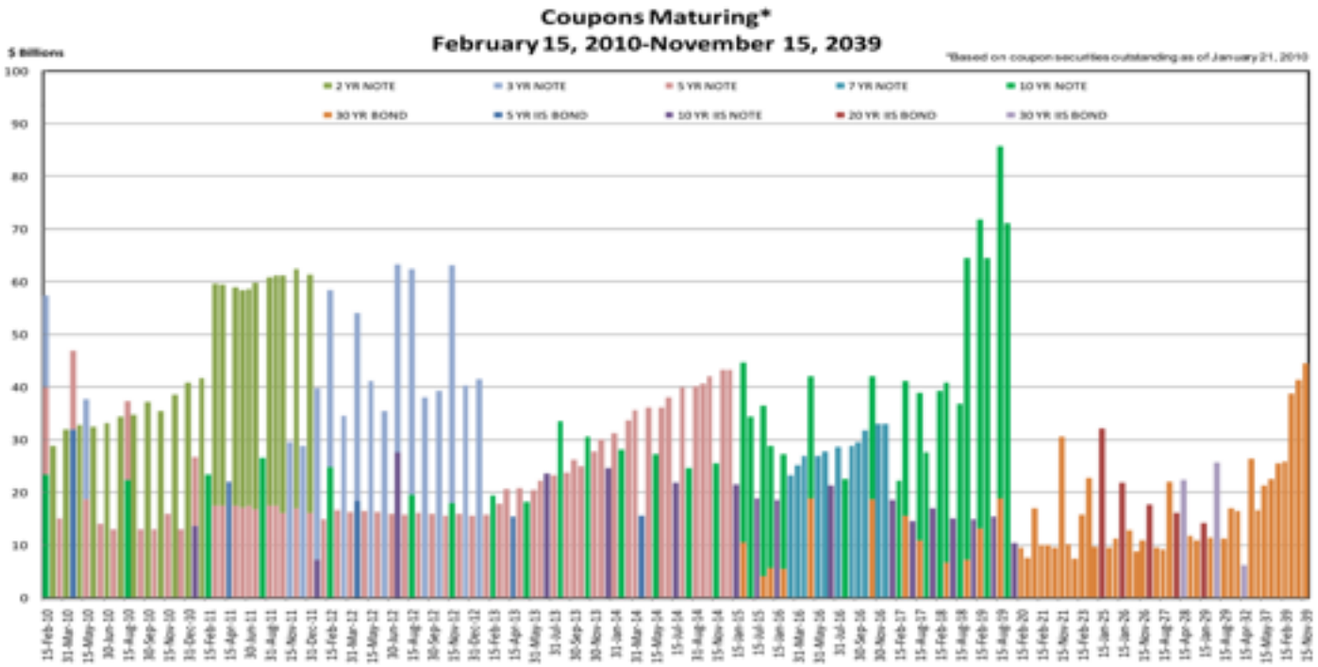


Figure 10





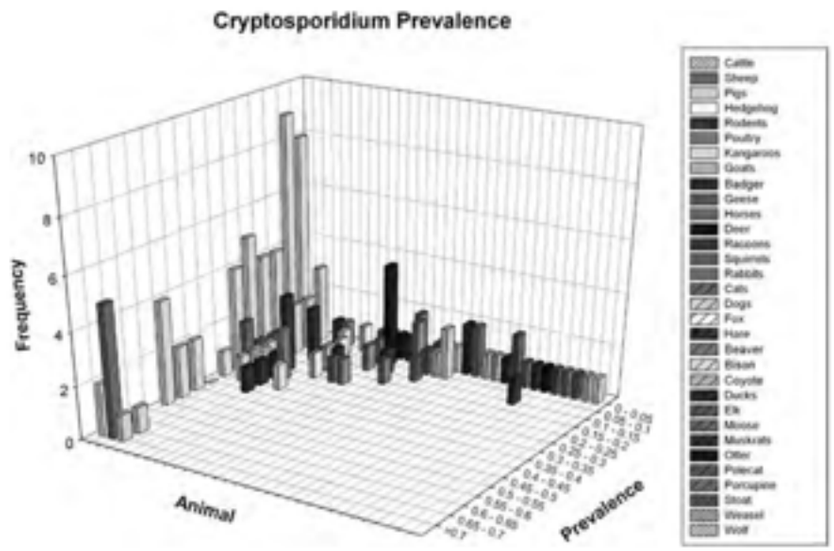
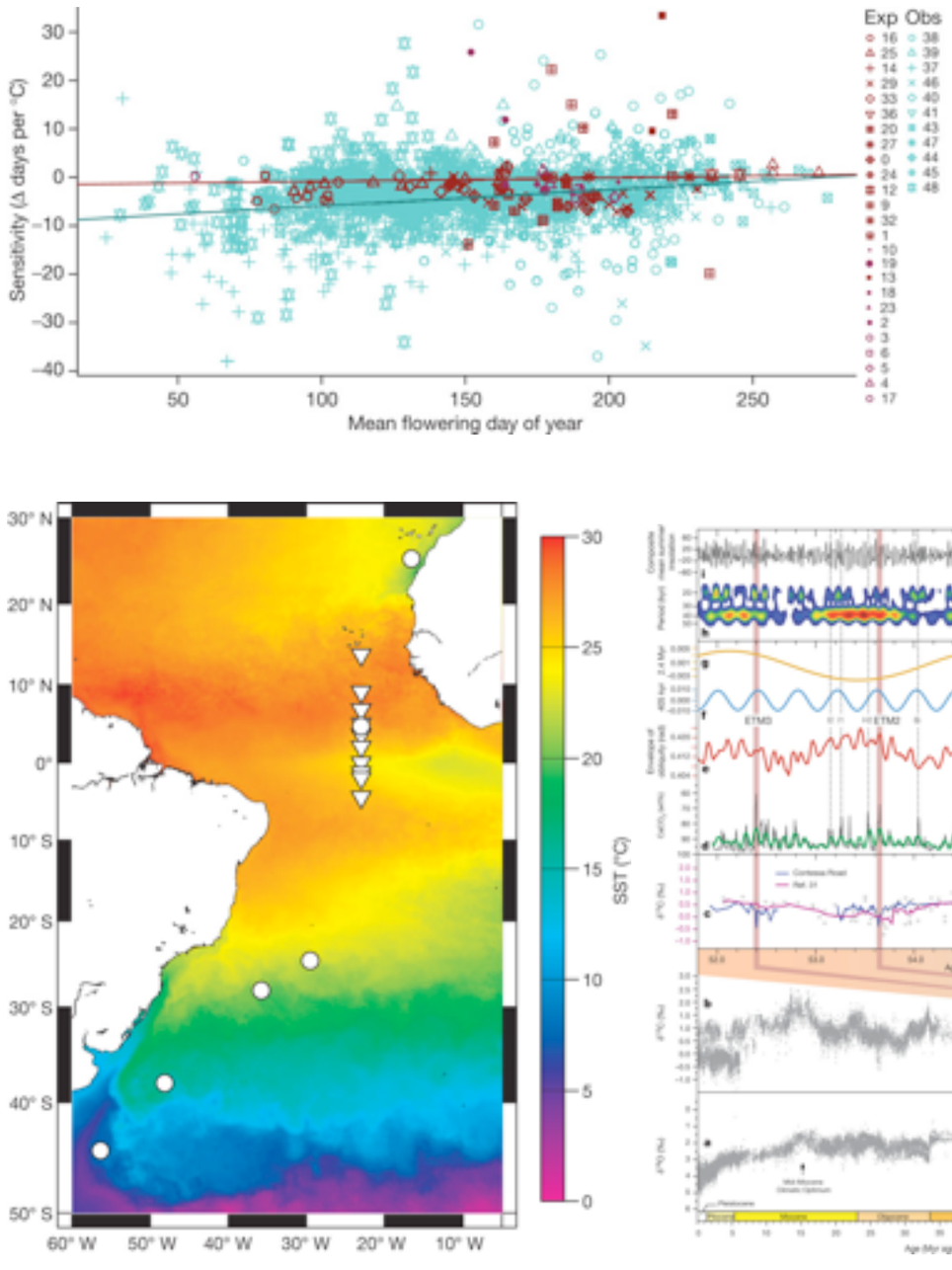
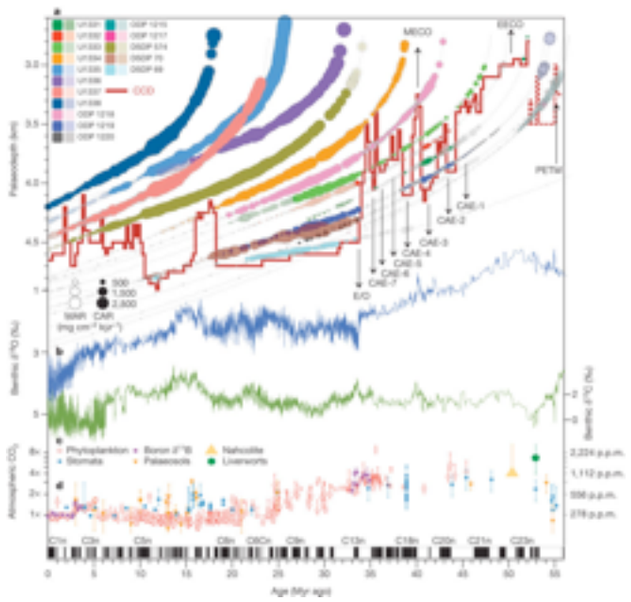


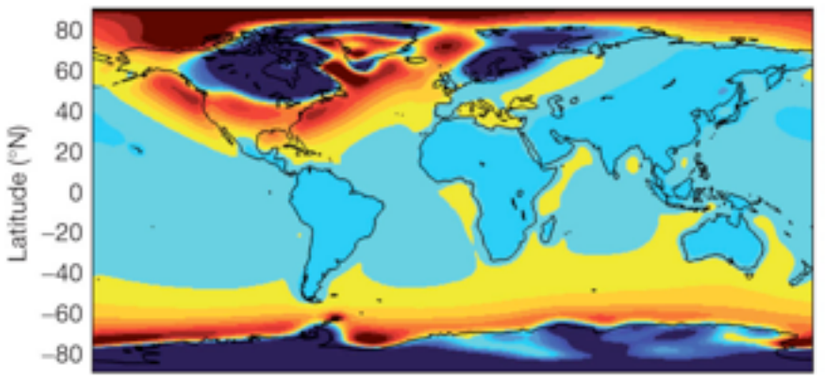
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

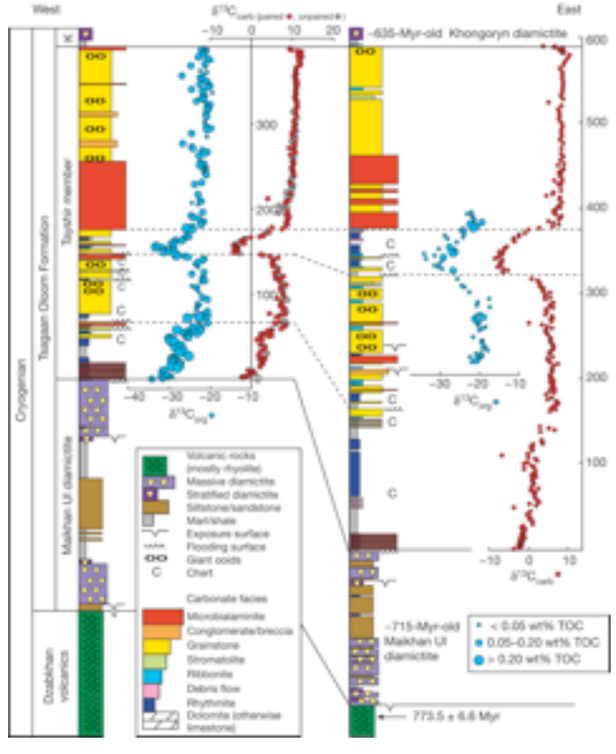
Sources: US Treasury and WHO reports

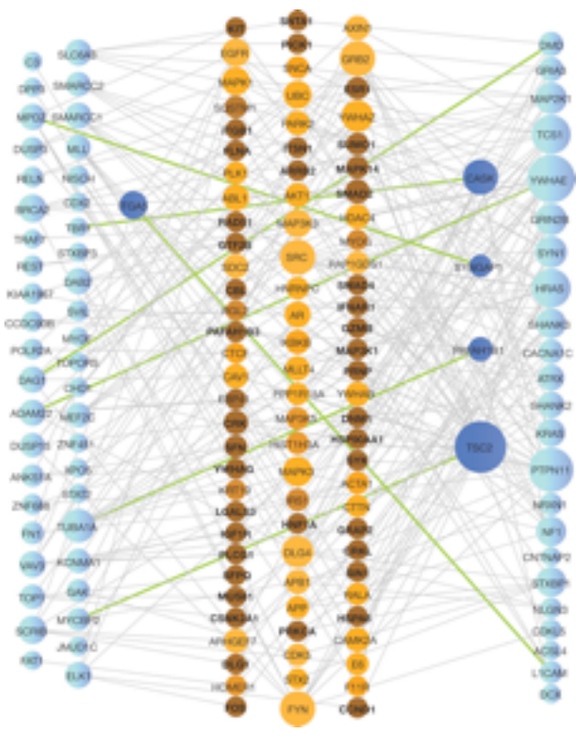
Also *not* effective....





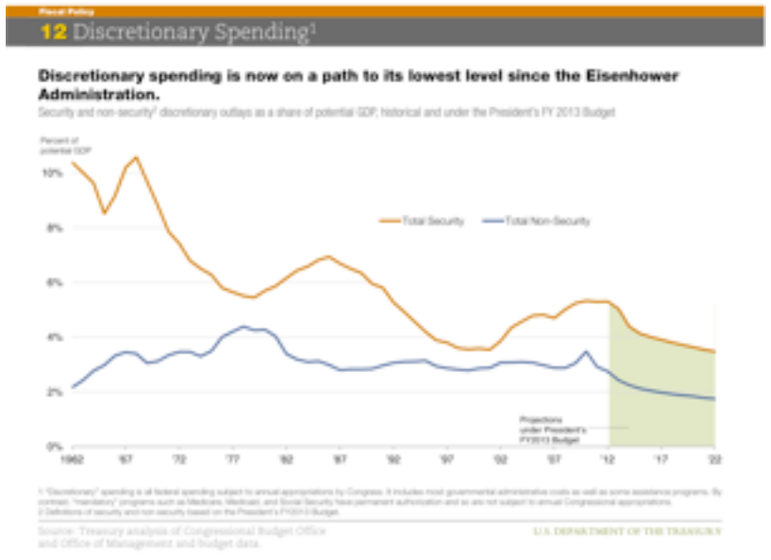


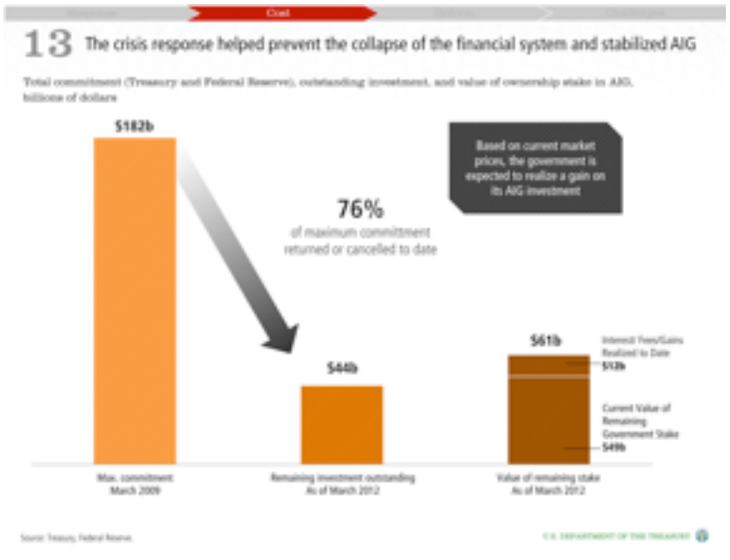


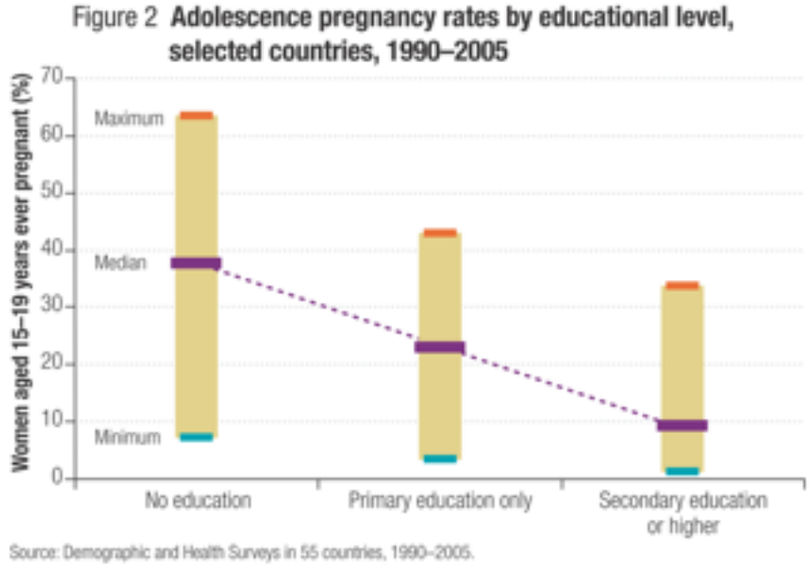


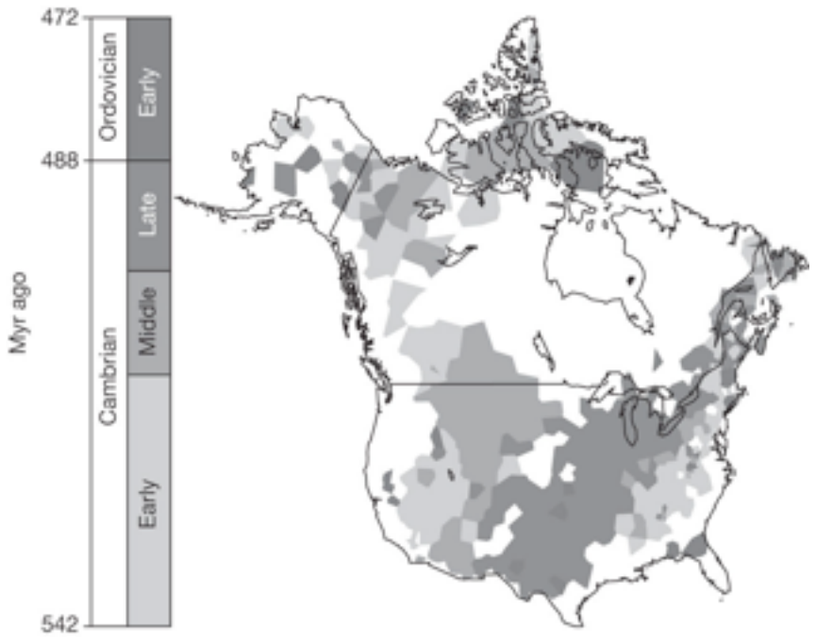
Source: Nature

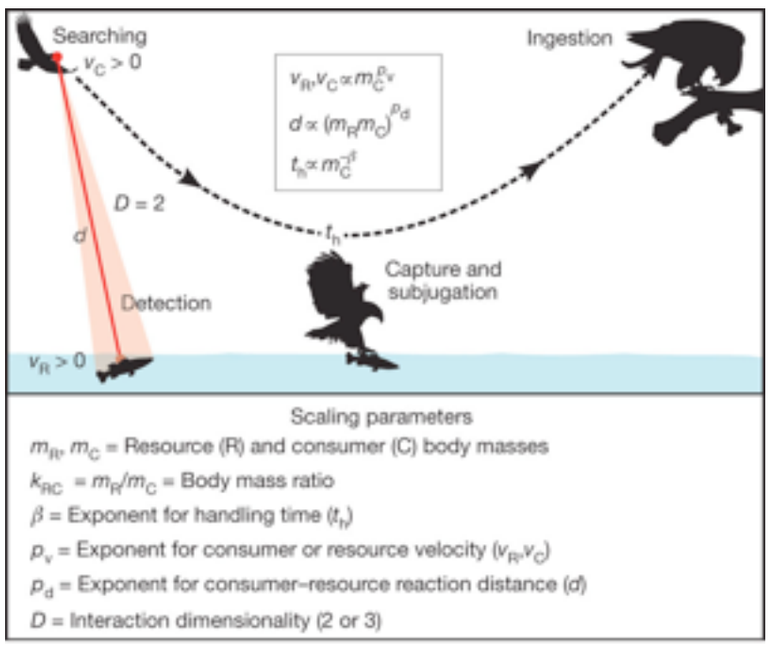
Much better...

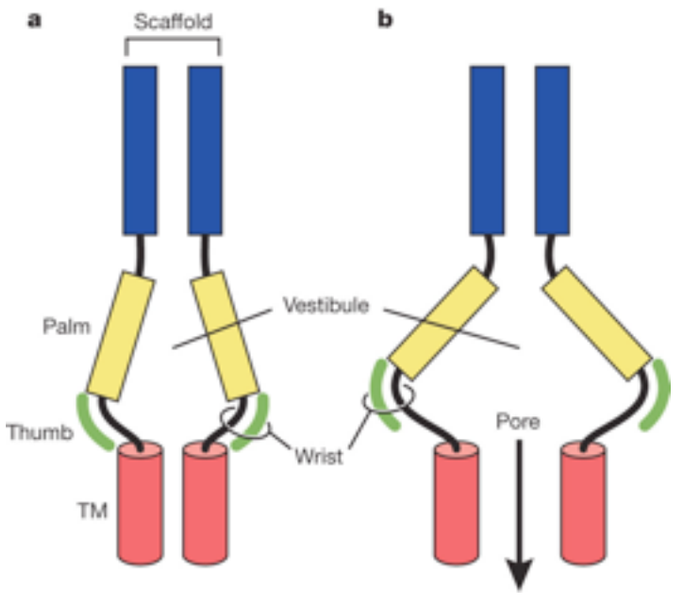












Sources: US Treasury, WHO, Nature