Stat 215A: Lecture 4
Exploratory Data Analysis (EDA): data visualization

Instructor: Bin Yu binyu@stat.berkeley.edu,
Office hours: 10-11 T/Th, 409 Evans Hall

Summary of Lec 2—3: important concepts

- What is the question? What does each number measure? Why could we generalize? To what population? BE THERE.
- Data sampling model (making randomness explicit)

  Population (describe with as much detail as one can) vs sample
- Controlled experiment vs observational study
- Confounding factors
- Neyman-Rubin model for causal inference: counterfactuals/potential outcome

Randomization is an important

idea in experimental design

or the science of data collection

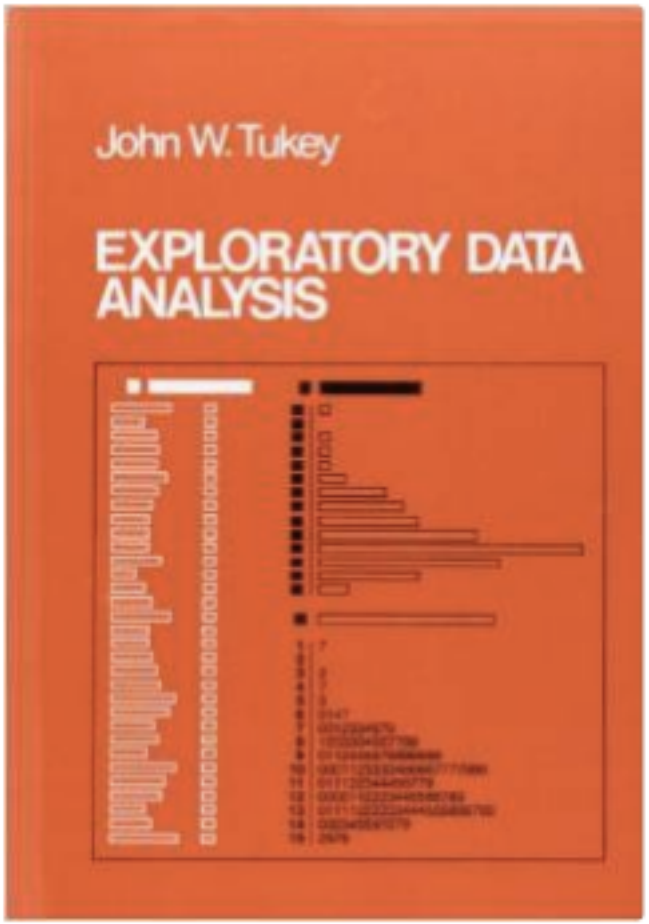
Classic books by Fisher, Cochran/Cox, …

Other uses of randomization:

Compressed sensing, randomized algorithms for big data,

Shannon's randomized channel code.

EDA by Tukey



Exploratory

Data Analysis

EDA through

Visualization (static or dynamic)

on

raw data, summaries/transformations, and

modeling/algorithm results

Visualization

should be your first choice

for lab reports/papers

for this class and for life

In 1801, Giuseppe Piazzi observed Ceres…

Pizazzi had three sightings of a new "planet", recorded positions on the sky and timings.

Tennenbaum and Director, 1997

In 1801, Giuseppe Piazzi observed Ceres…

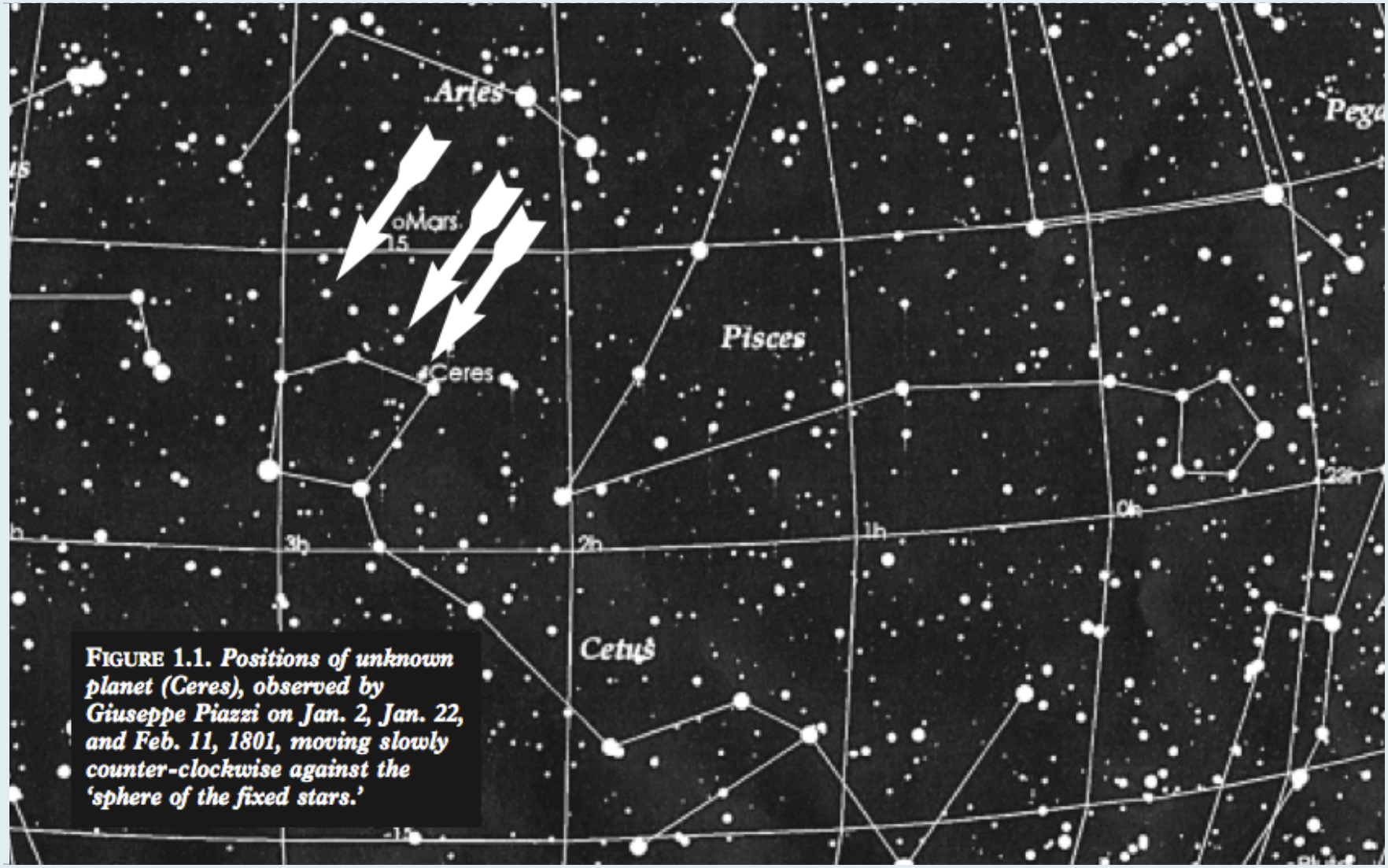


FIGURE 1.1. ***Positions of unknown planet (Ceres), observed by Giuseppe Piazzi on Jan. 2, Jan. 22, and Feb. 11, 1801, moving slowly counter-clockwise against the 'sphere of the fixed stars.'***

Tennenbaum and Director, 1997

Giuseppe Piazzi

(July 16, 1746 – July 22, 1826)

Piazzi was an Italian Catholic priest of the Theatine order, mathematician, and astronomer. He supervised the compilation of the Palermo Catalogue of stars, containing 7,646 star entries with unprecedented precision. Piazzi discovered Ceres, today known as the largest member of the asteroid belt. -- Wikepedia

Giuseppe Piazzi

(July 16, 1746 – July 22, 1826)

Piazzi was an Italian Catholic priest of the Theatine order, mathematician, and astronomer. He supervised the compilation of the Palermo Catalogue of stars, containing 7,646 star entries with unprecedented precision. Piazzi discovered Ceres, today known as the largest member of the asteroid belt. -- Wikepedia

Carl Friedrich Gauss

(April 30, 1777 – February 23, 1855)

Gauss was a German mathematician and physical scientist who contributed significantly to many fields, including number theory, algebra, statistics, analysis, differential geometry, geodesy, geophysics, electrostatics, astronomy and optics. -- Wikipedia

Carl Friedrich Gauss

(April 30, 1777 – February 23, 1855)

Gauss was a German mathematician and physical scientist who contributed significantly to many fields, including number theory, algebra, statistics, analysis, differential geometry, geodesy, geophysics, electrostatics, astronomy and optics. -- Wikipedia

Johannes Kepler

(December 27, 1571 – November 15, 1630)

Kepler was a German mathematician, astronomer and astrologer.

A key figure in the 17th century scientific revolution, he is best known for his laws of planetary motion... -- Wikipedia

Johannes Kepler

(December 27, 1571 – November 15, 1630)

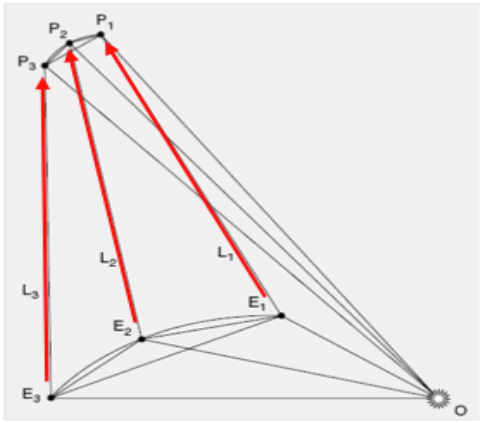
Kepler was a German mathematician, astronomer and astrologer.

A key figure in the 17th century scientific revolution, he is best known for his laws of planetary motion... -- Wikipedia

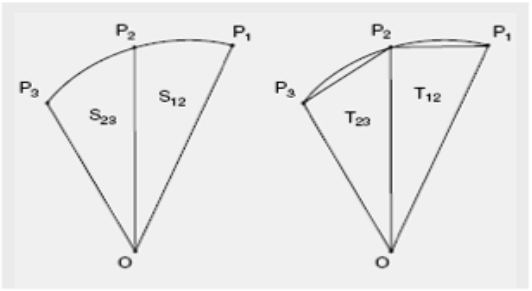
Gauss predicted Ceres' trajectory or future locations accurately with 3 data points

He used Kepler's second law of planetary motion, geometric relationships, approximations, and corrections,…

Piazzi's data: lines of sight L_1, L_2, L_3 and elapsed times between observations



Sectoral areas swept out by orbit are proportional to elasped times



$$\frac{S_{12}}{S_{23}} = \frac{t_2 - t_1}{t_3 - t_2} = 0.94952\,,$$

$$\frac{S_{12}}{S_{13}} = \frac{t_2 - t_1}{t_3 - t_1} = 0.48705\,,$$

$$\frac{S_{23}}{S_{13}} = \frac{t_3 - t_2}{t_3 - t_1} = 0.51295\,.$$

Approximate sectoral areas with triangular areas

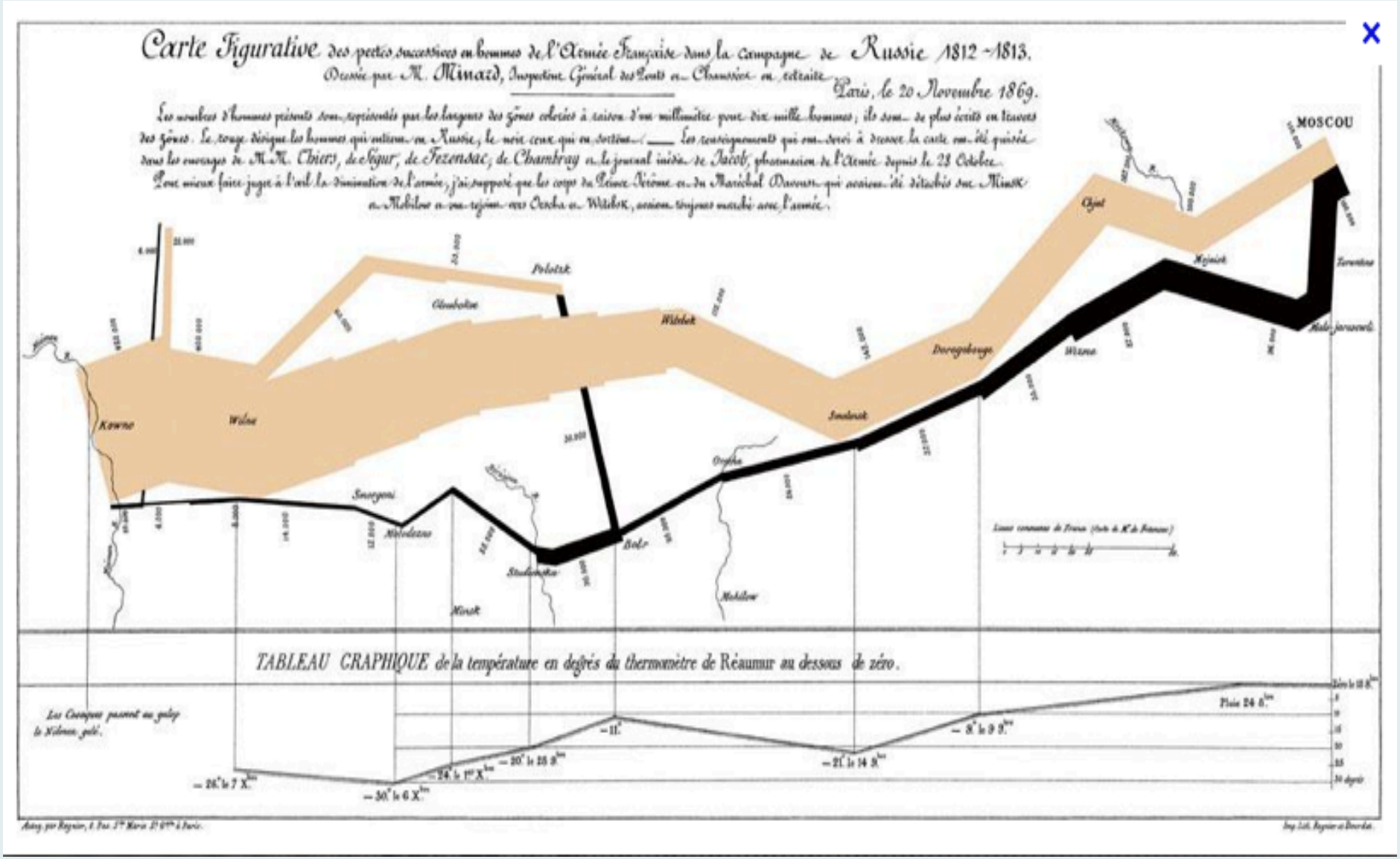
$$\frac{T_{23}}{T_{13}} = \text{(approximately)}\ \frac{S_{23}}{S_{13}} = 0.513\,, \qquad \text{="c"}$$

$$\frac{T_{12}}{T_{23}} = \text{(approximately)}\ \frac{S_{12}}{S_{23}} = 0.487\,. \qquad \text{="d"}$$

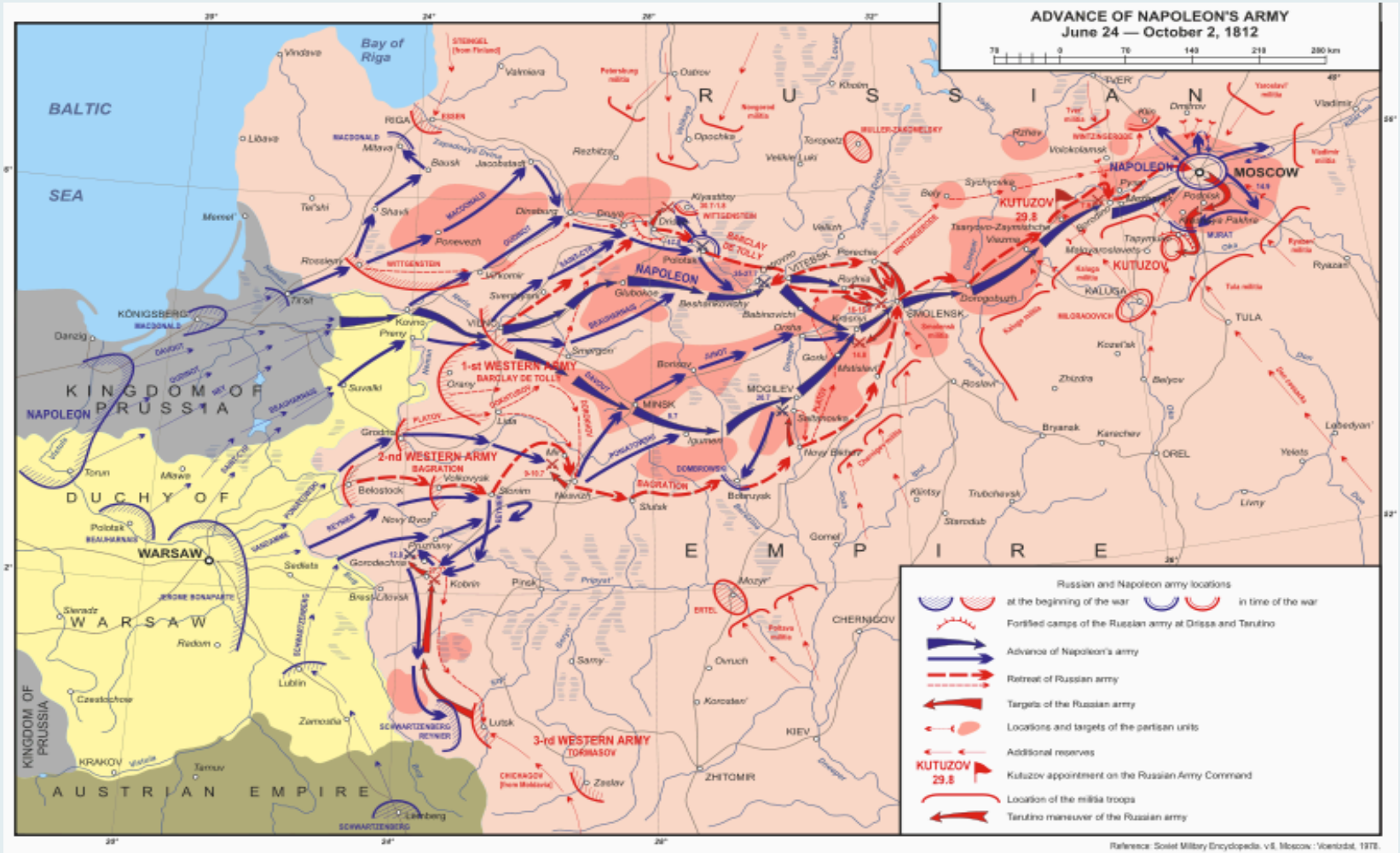
Tennenbaum and Director, 1997; Le Corvec, Donatelli and Hunt, 2007 Math 221 project, UCB

A good graph is worth of a thousand words

by Charles Joseph Minard



How about this one?



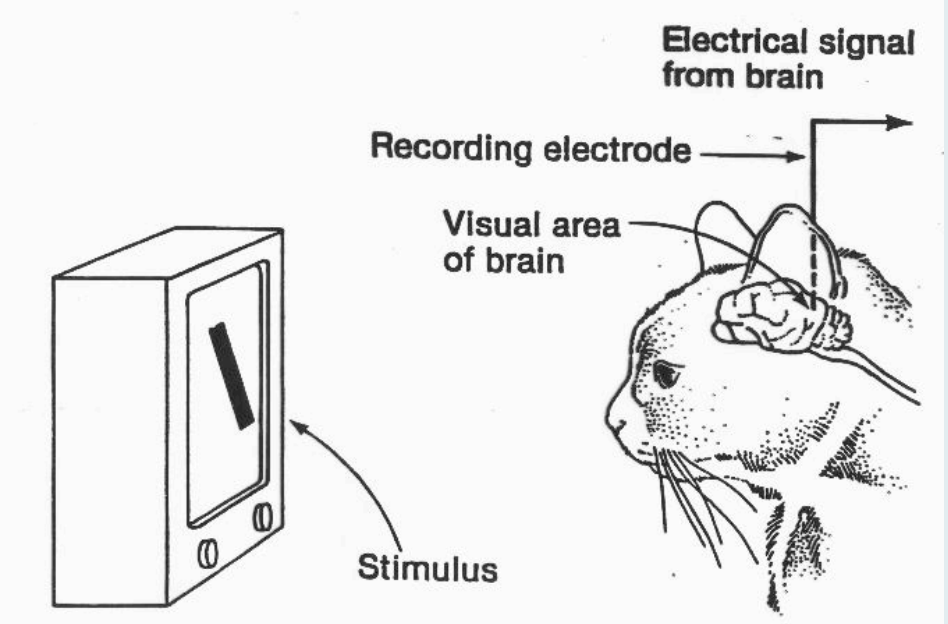
ize of this preview: 800 × 538 pixels. Other resolutions: 320 × 215 pixels I 640 × 430 pixels.

“Kepler”=Hubel and Wiesel (1959)

They discovered, in neuron cells of the primary cortex area V1,

orientation and location selectivity, and

excitatory and inhibitory regions .



Hubel and Wiesel (Nobel Prize, 1981)

"The signal message that the eye sends to the brain can be regarded as a secret code to which only the brain possesses the key and can interpret the message. Hubel and Wiesel have succeeded in breaking the code."

-- Presentation Speech by Professor David Ottoson of the Karolinska Institute

at the Award Ceremony for

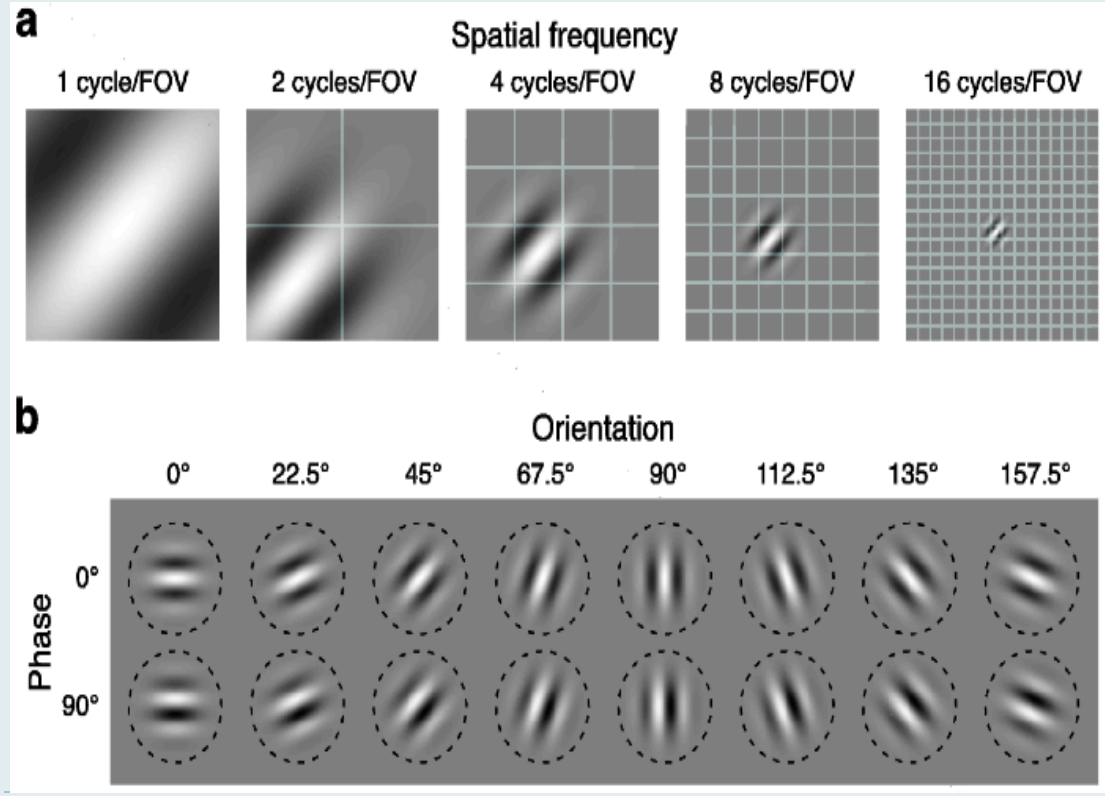
The Nobel Prize in Physiology or Medicine 1981

Roger W. Sperry, David H. Hubel, Torsten N. Wiesel

Modern Description of Hubel-Wiesel work: Early Visual Area V1

- Preprocessing an image:
  - Gabor filters corresponding to particular spatial frequencies, locations, orientations (Hubel and Wiesel, 1959,…)

Sparse representation after Gabor Filters, static or dynamic



"Collective Piazzi" = Gallant Lab Movie-fMRI Data

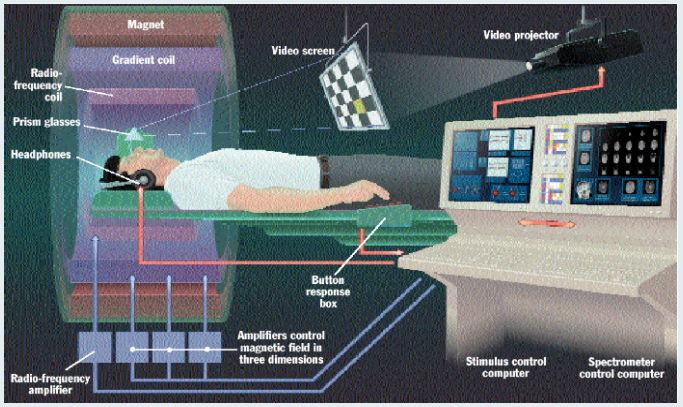
- Non invasive and indirect recording technique
- Measures oxygenated blood flow – correlate of neural activity
- Low temporal resolution, a few seconds
- High spatial resolution (voxels of 1x1x1 mm cubes)
  - > 10,000 voxels in early visual areas (V1, V2 and V3)
  - each voxel covers > 100,000 neurons

- Can watch videos inside the machine:

3 subjects

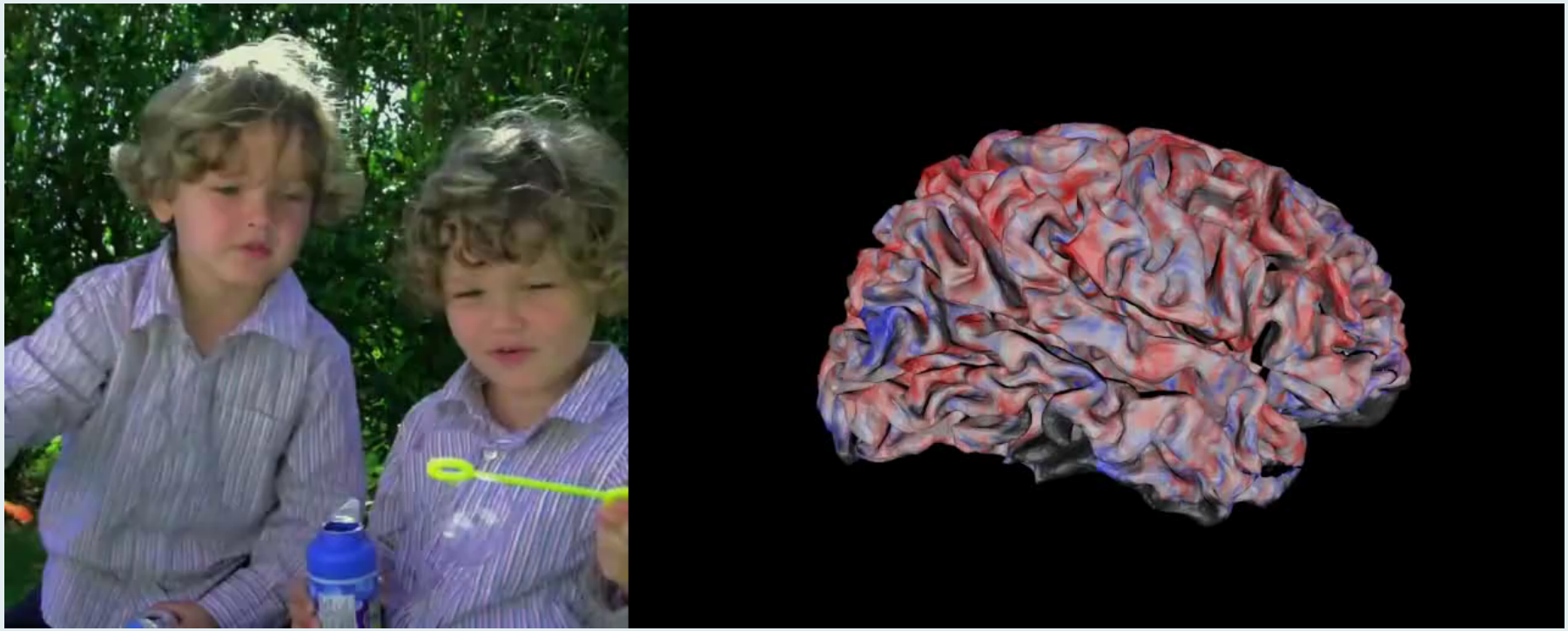
7200s training (1 replicate)

and 5400s test (10 replicates).



What is our movie-fMRI Data?

7200s training (1 replicate) and 5400s test (10 replicates).



References

The future of data analysis, by Tukey (1962).

Envisioning information, by Edward R. Tufte (1990).

Visualizing data, by William S. Cleveland (1993).

…

Visualizing data

One third of human brain is devoted to visual processing.
We can utilize our knowledge on visual perception to our advantage in order to find trends/patterns/outliers in data.

Visualizing data

Go to website:
http://queue.acm.org/detail.cfm?id=1805128

A tour through the visualization zoo

By by Jeffrey Heer, Michael Bostock, Vadim Ogievetsky | May 1, 2010 in "Queue"

Visualizing data

Well-designed visual representations can replace cognitive calculations with simple perceptual inferences and improve comprehension, memory, and decision making. By making data more accessible and appealing, visual representations may also help engage more diverse audiences in exploration and analysis. The challenge is to create effective and engaging visualizations that are appropriate to the data.

- A Tour through the Visualization Zoo

by Jeffrey Heer, Michael Bostock, Vadim Ogievetsky, 2010.

Visualizing data

The question is how to map attributes of data into visual representation elements.

Static representations of data rely on a 2-dim representation with added help from color and shading. But motion (dynamic representation) brings another dimension as well.

…

Basic facts about perception

- Continuous values are better mapped into gradient scales.
- Colors are often interpreted as different categories.
- Bright colors (e.g. red) catch attention and areas of color matter (small areas obtain attention with a bright color).
- Transparency and point size should be utilized for large data sets so more information about data gets through.
- A smoothing line is useful to fit through a 2-dim plot such as scatter or time series plots, especially when the data plots are over-plotted on top of each other.
- Brushing can bring in other variables effectively.
- Motion (movie) can bring out structures that static representation can't.

"Visualizing data" by Bill Cleveland

"Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones."

"There are two components to visualizing the structure of statistical data – graphing and fitting. Graphs are needed, of course, because visualization implies a process in which information is encoded on visual displays. Fitting mathematical functions to data is needed too. Just graphing raw data, without fitting them and without graphing the fits and residuals, often leaves important aspects of data undiscovered."

Visualizing data

Often, the most effective visualization tools are simple ones.

We review some in this class and give some analysis under generative model assumptions

1. Kernel density estimator (histogram)
2. Box-plot, pie-chart
3. Scatter plot, lowess (smoothing a data cloud)
4. Q-Q plot

Summarizing/describing a list of data points

Example: performances "scores" for prediction models based on two model selection methods (Lasso+CV, Lasso+SSCV) to related natural stimuli (video clips) to fMRI brain signals over different voxels in the primary visual cortex of one human subject. One score is the size of the model and the other is the correlation of predicted values and observed fMRI values on a validation set.

Summarizing/describing a list of data points

Given a list of data points, it is important to make sure they are comparable in terms of scale and meaning and can be put together in one list.

There are numerical summaries of a data list:

mean (median)

standard deviation (interquartile range)

– they are not adequate if the data list has multiple modes.

Graphical summaries of one-dim list

- Stem-and-leaf
- Histogram (smoothing-parameter: equal-bin-size, variable-bin-size)

> summaryCV[,1] (sizes of the CV-models for 2000 voxels)

```
 [1]  76  89  33  86  42  73  61  82  89  62  75  47  71 116  59 136  58  38  60 137  54
  34 109  73  56  21  55 132  63  55  42
[32]  90  78 114  47  71  79  81  98  72  61  70 150  70  59  90  65  73  47  20  23  51
  55  52  37  51  51  69  54  35  49  51
[63]  55  56  67  97  35  84  63  60  54 134  76  19  77  67  66  87  64  34  65 111  95
  67  90  58  42  66 158  47  64  99  81
[94]  53  50 108  61  76  64  44  83  70 135  62  74  99  55  33  40  62  80  57  72 …
```

Graphical summaries of one-dim list

```
>summarySSCV[,1]
 [1] 20 29 15 27 18 18 27 22 26 32 28 20 31 35 37 29
 6 13 22 65 27 19 19 37 36 15 32 20 23 17 32 39 47 42
 42 20 70 29 36 21 27
[42] 21 45 18 12 13 19 17 29 20 23 42 36 22 24 26 26
 21 18 17 24 15 26 21 17 16 15 16  9 26 15 27 19 10 26
 21 17 18 17 11 20 16
[83] 13 12 13 27 23 26 27 24 20 33 11…
```

Impressions from the lists:

SSCV selects smaller models on average than CV.

Graphical summaries of one-dim list

> summaryCV[,6] (correlation prediction scores)

[1] 0.46492314 0.60122322 0.24294175 0.55085573
0.39745474 0.49414930 0.53183917 0.45804443
0.56753623 0.45891439

[11] 0.59330310 0.22770444 0.56311166
0.56758193 0.35013004 0.47387326 0.58750394
0.47850106 0.37218576 0.53115991

[21] 0.55721982 0.39688628 0.54342919
0.47873061 0.44913212 0.21333301 0.39974994
0.64192666 0.49774619 0.40603347

[31] 0.38552492 0.23207142 …

Graphical summaries of one-dim list

> summaryCV[,6] (correlation prediction scores)

[1] 0.46492314 0.60122322 0.24294175 0.55085573
0.39745474 0.49414930 0.53183917 0.45804443
0.56753623 0.45891439

[11] 0.59330310 0.22770444 0.56311166
0.56758193 0.35013004 0.47387326 0.58750394
0.47850106 0.37218576 0.53115991

[21] 0.55721982 0.39688628 0.54342919
0.47873061 0.44913212 0.21333301 0.39974994
0.64192666 0.49774619 0.40603347

[31] 0.38552492 0.23207142 …

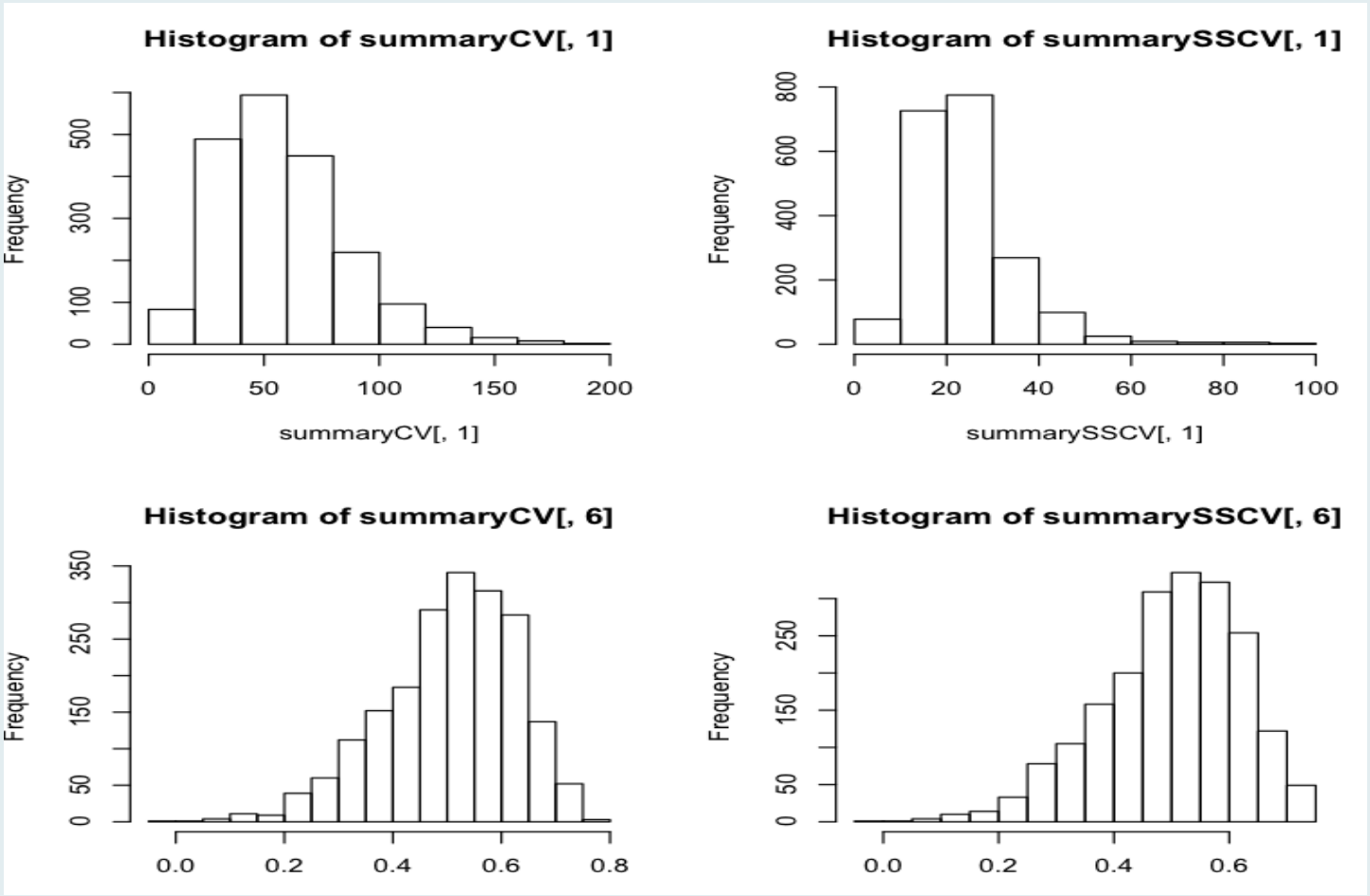
Graphical summaries of one-dim list

```
> summaryCV[,6]
[1]  0.46492314  0.60122322  0.24294175  0.55085573
  0.39745474  0.49414930  0.53183917  0.45804443
  0.56753623  0.45891439
 [11]  0.59330310  0.22770444  0.56311166
  0.56758193  0.35013004  0.47387326  0.58750394
  0.47850106  0.37218576  0.53115991
 [21]  0.55721982  0.39688628 …
```

Impressions: both lists are non-negative, but hard to know which list has larger numbers over all…

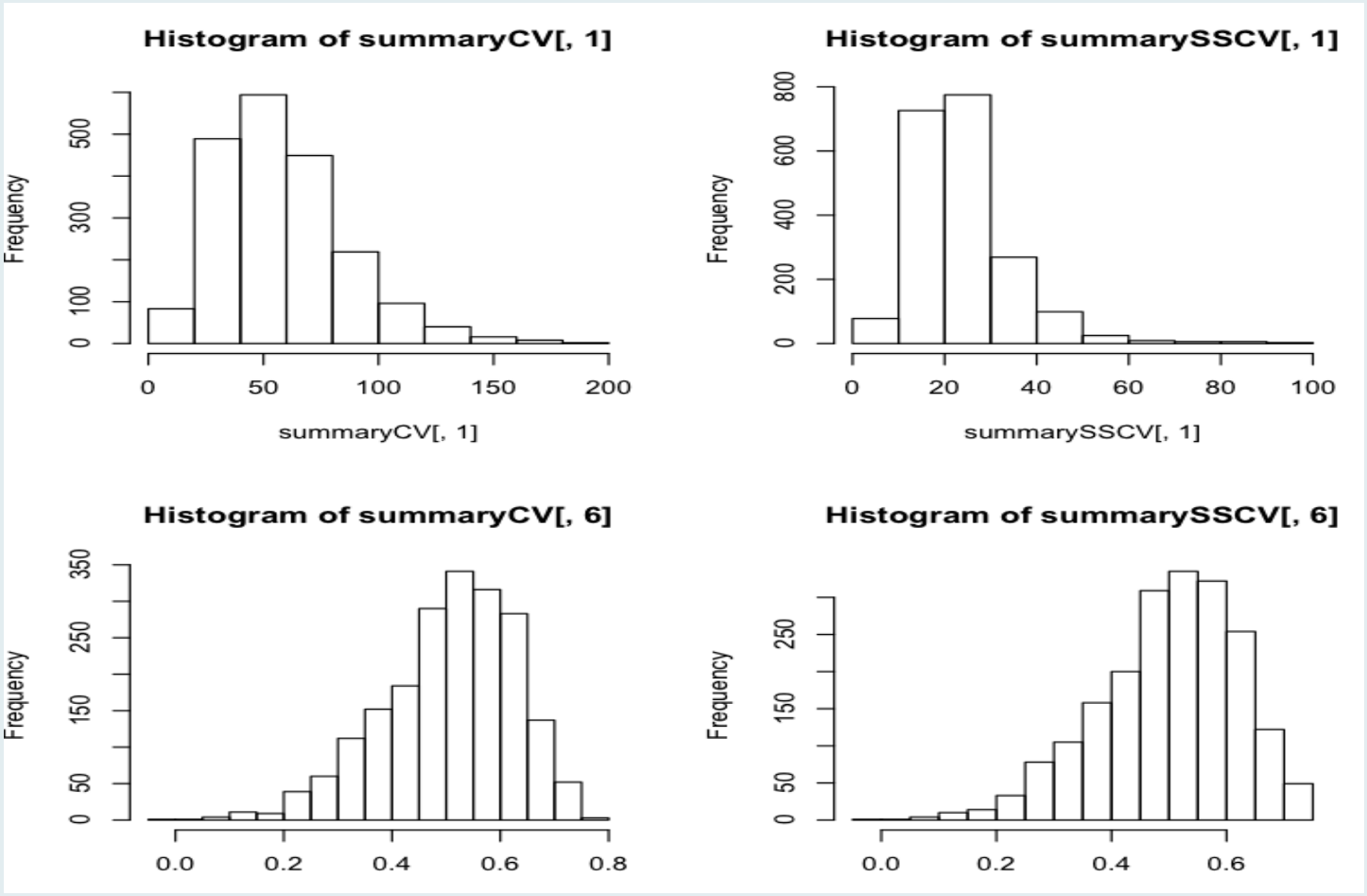
Graphical summaries of one-dim list

First row: model size; Second row: correlation



Graphical summaries of one-dim list

First row: model size; Second row: correlation



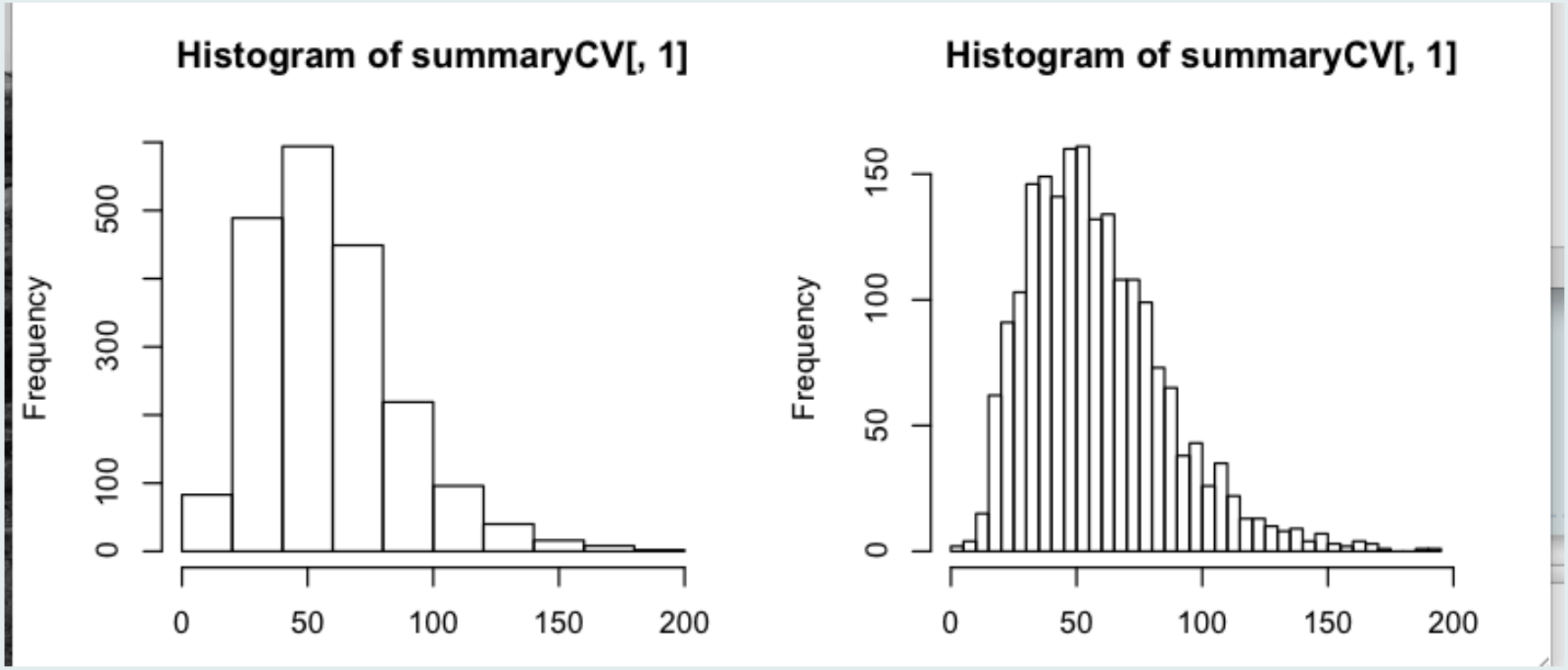
Graphical summaries of one-dim list

Why did I bother plot both of the previous two pages? They used the same data…

Based on plots on page 22, SSCV gives much smaller models, but similar prediction scores – it is preferred based on Occam's Razor.

Graphical summaries of one-dim list

The following plots are produced by the R-commands:
>hist(summaryCV[,1], xlim=c(0,200))
>hist(summaryCV[,1], xlim=c(0,200),nclass=45)
Are the modes in the right histogram real? Is a statistical test necessary. No.



Take a look at CS294-10 by Maneesh Agrawala

Thanks to Eugene Yedvabny in our class, who wrote to me and Ryan:

"Today's lecture on visualization is a subject of an entire course in the CS department taught by Maneesh Agrawala (CS294-10):

http://vis.berkeley.edu/courses/cs294-10-fa14/wiki/index.php/Main_Page <- This year's class, which updates after each lecture. To access readings use vis2014/vis2014Readings

http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/index.php/Main_Page <- Last year's class, so has the full slide deck, all readings, etc.

As far as I can tell he's teaching the same material this year. To access the readings use vis2013/vis2013Readings"

Reading:

1. Tukey (1962)

2. Go through visualization zoo at
http://queue.acm.org/detail.cfm?id=1805128

3. Browse CS294-10 website (see last slide)