Supplemental Resource: Brain and Cognitive Sciences
Statistics & Visualization for Data Analysis & Inference
January (IAP) 2009

Statistics and Visualization for Data Analysis and Visualization

Mike Frank & Ed Vul

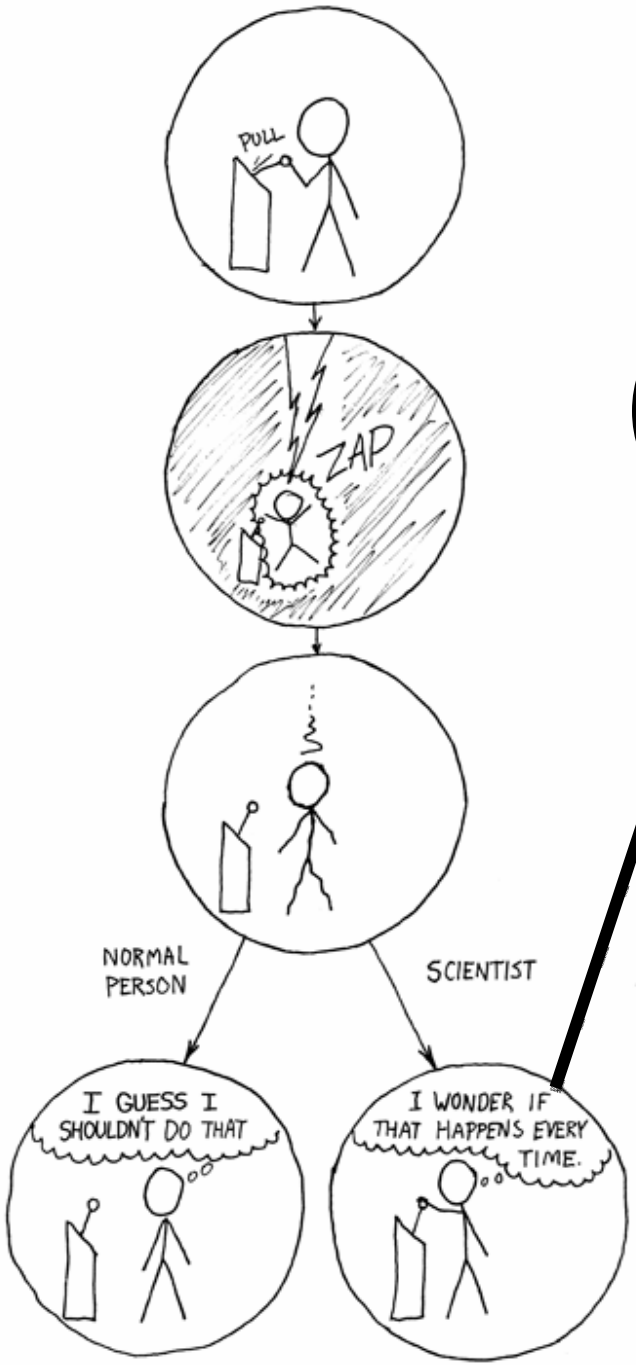
IAP 2009

Who we are

Ed Vul – 4th year grad student in the Kanwisher lab. Interested in optimal decision-making, resource allocation, and visual attention.

Mike Frank – 4th year grad student in the Gibson lab. Interested in language acquisition, interactions of language and cognition.

Why this course?



now what?

statistics and data analysis

Courtesy of xkcd.org

Our goals

- summarize and discuss an approach
  - contrast the “null-hypothesis significance testing” framework
  - to a “model-driven” framework
- link data to theory
  - “statistical models are models of data”
  - acknowledge scientific practice in data analysis and theory development
- get feedback from all of you

Your goals

- Name
- What you work on / where you work
- Your statistical background
- (optional) a question that bothers you sometimes when you're analyzing data

Approach

data visualization/data modeling: look at your data and try to understand where it came from

1. **visualization**
   - creating appropriate and informative pictures of a dataset
   - iterative exploration of data
2. **modeling**
   - don't just test for differences, try to understand the factors
   - not just "looking for interactions," main effects too
   - emphasis on effect size, not significance
   - use appropriate computational tools, don't rely on simple analytic approximations (e.g. t-tests) if they don't fit
3. **experimental design**
   - choose designs and measures that test questions
   - don't choose designs based on arbitrary statistical frameworks (e.g. ANOVA)

Classes

1. **Visualization** – how can I see what my data show?
2. **Resampling** – how do I estimate the uncertainty of my measures?
3. **Distributions** – how do I summarize what I believe about the world?
4. **The Linear Model** – how can I create a simple model of my data?
5. **Bayesian Modeling** – how can I describe the processes that generated my data?

Classes

1. **Visualization** – how can I see what my data show?
2. **Resampling** –how do I estimate the uncertainty of my measures?
3. **Distributions** – how do I summarize what I believe about the world?
4. **The Linear Model** – how can I create a simple model of my data?
5. **Bayesian Modeling** – how can I describe the processes that generated my data?

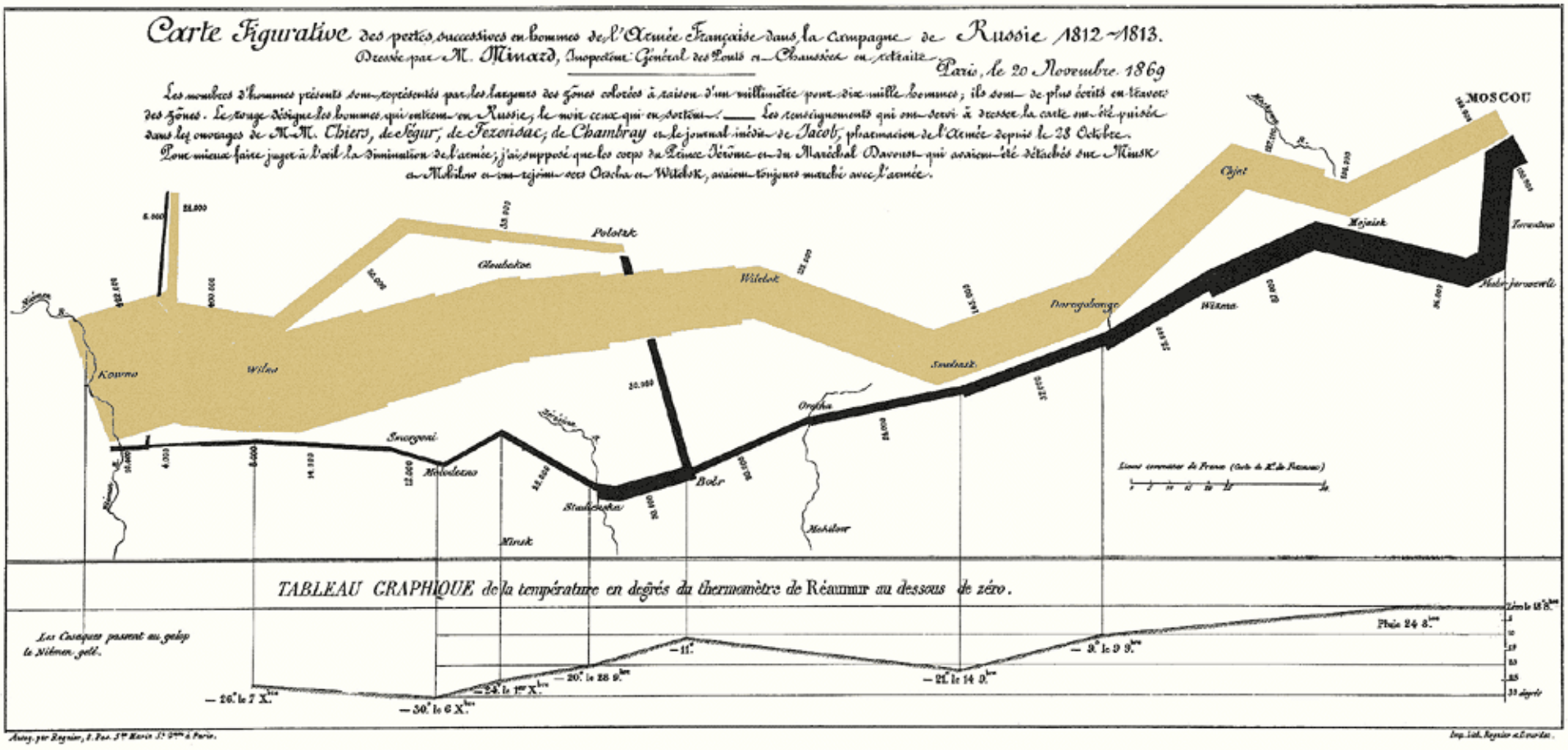
VISUALIZATION

Outline

- Why visualize?
  - to understand data
  - a worked example
- The visual vocabulary
  - elements
  - perceptual motivations
- Conventional modes of combination
  - taxonomy of visualization
- Tips & Tricks, Tradeoffs, & Trouble

(many slides courtesy of Chris Collins, U of T)

Example: Movements of the French Army



Minard, 1861; Tufte, 2001

Three principles for visualization:

1. **be true to your research** – design your display to illustrate a particular point
2. **maximize information, minimize ink** –use the simplest possible representation for the bits you want to convey
3. **organize hierarchically** – what should a viewer see first? what if they look deeper?

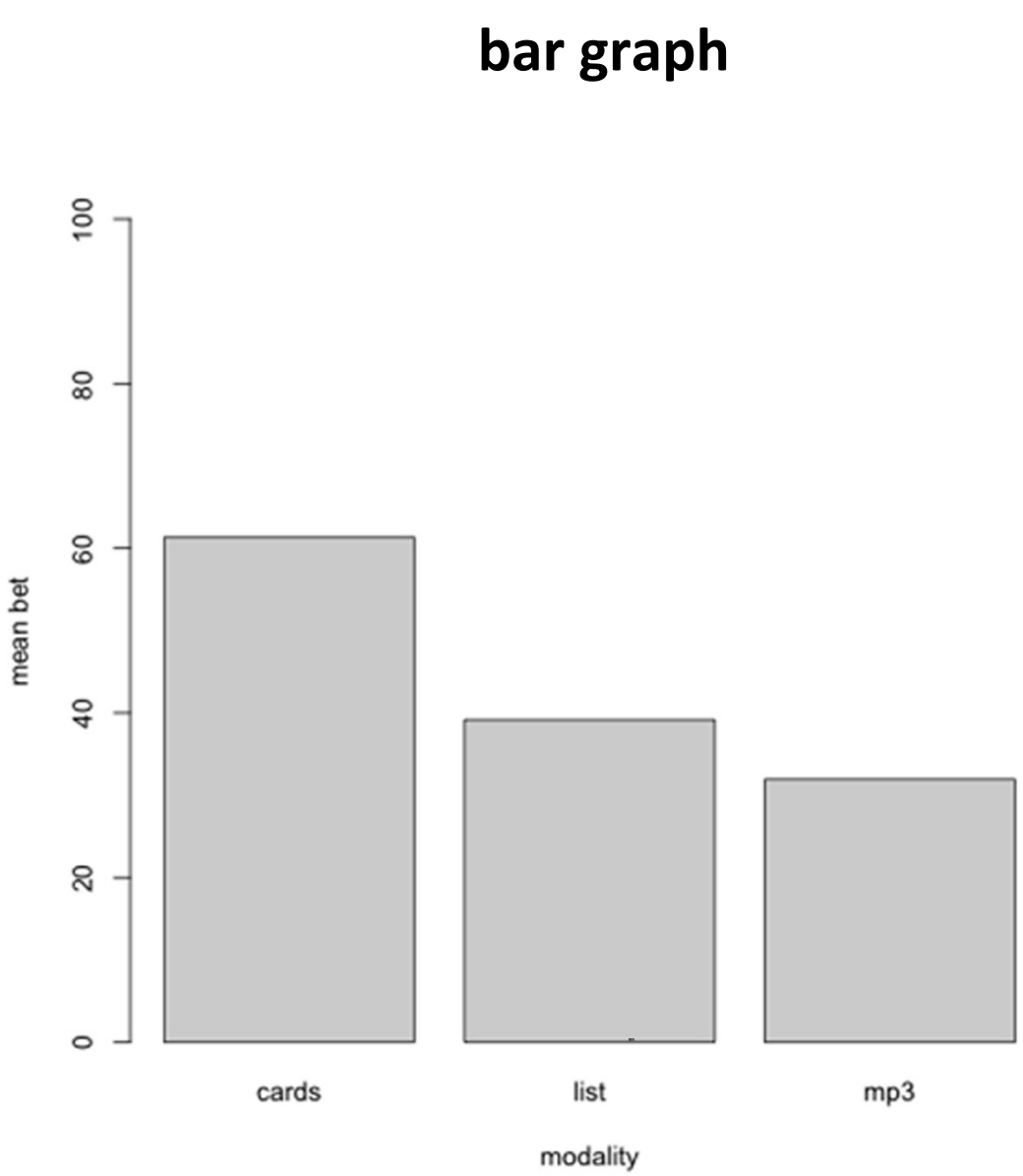
Worked example

- Participants heard examples from an artificial language
- Three different presentation methods for examples
  - index cards, list of sentences, mp3 files on ipod
- Task was to spread $100 of “bets” across different continuations for a new example
- Dependent measure was bet on the correct answer

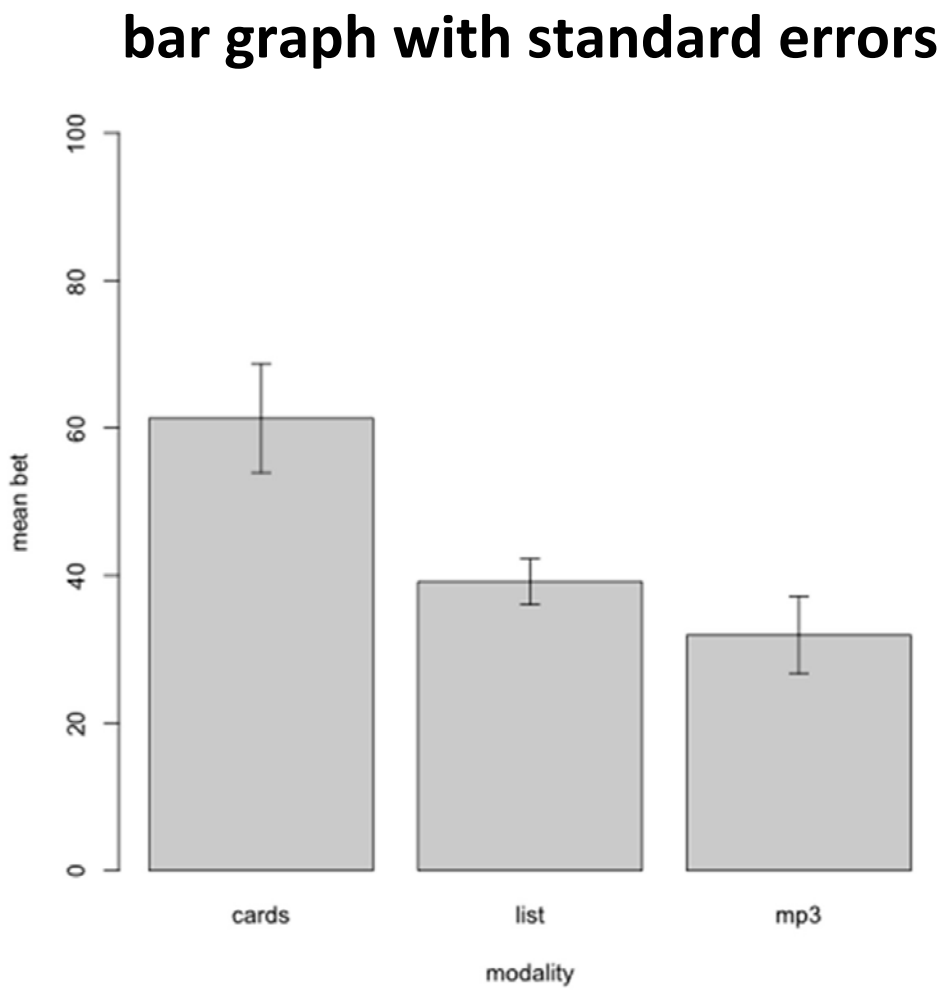
Worked example

trial.num	bet	trial	sub	expt	modality
1	80	1	S1	MNPQ	mp3
2	5	2	S1	MNPQ	mp3
3	90	3	S1	MNPQ	mp3
4	25	4	S1	MNPQ	mp3
5	0	1	S2	MNPQ	mp3
6	0	2	S2	MNPQ	mp3
7	50	3	S2	MNPQ	mp3
8	33	4	S2	MNPQ	mp3
9	0	1	S3	MNPQ	mp3
10	40	2	S3	MNPQ	mp3
11	60	3	S3	MNPQ	mp3
12	40	4	S3	MNPQ	mp3
…	…	…	…	…	…
191	40	3	S48	MNPQ	list
192	50	4	S48	MNPQ	list

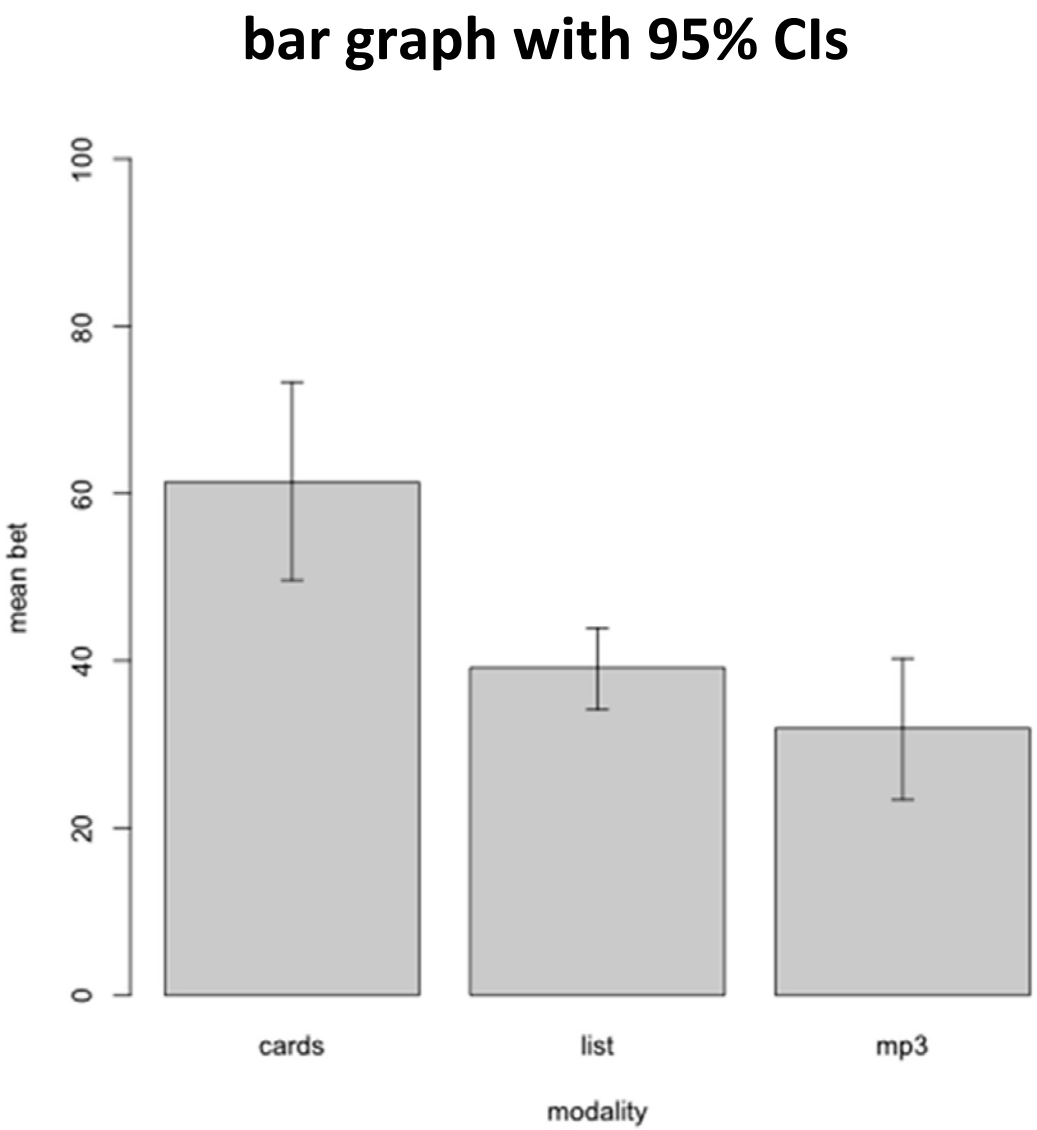
Worked example



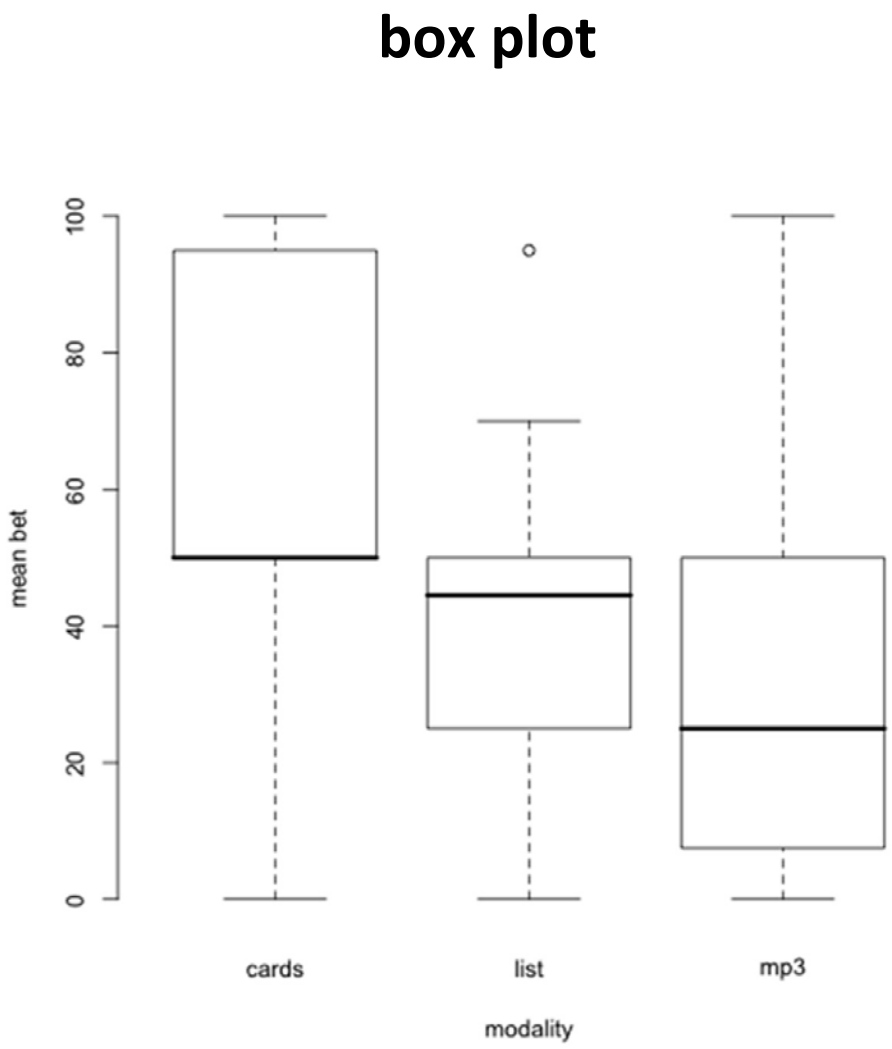
Worked example



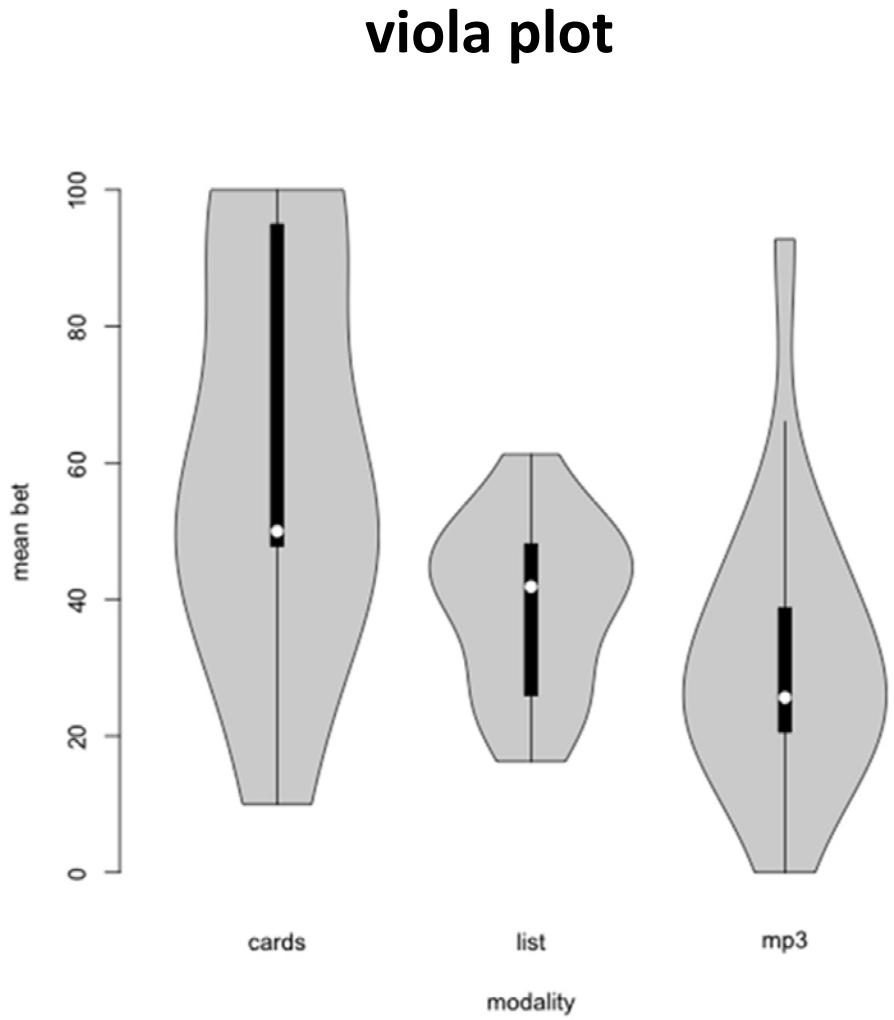
Worked example



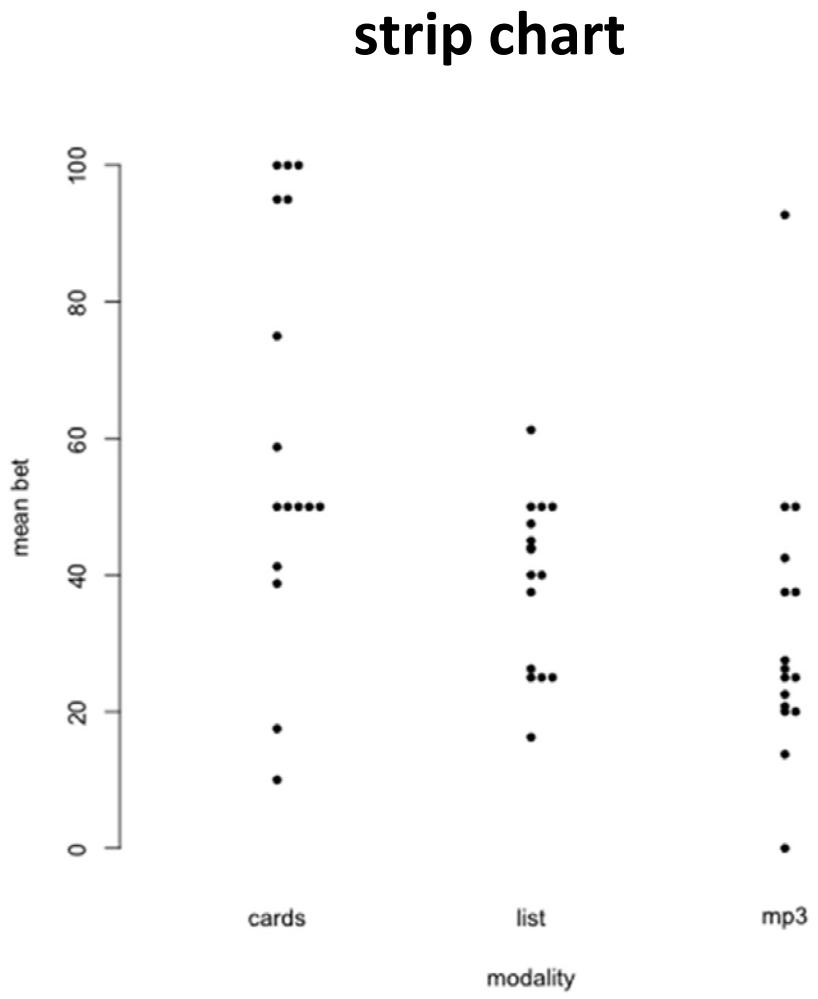
Worked example



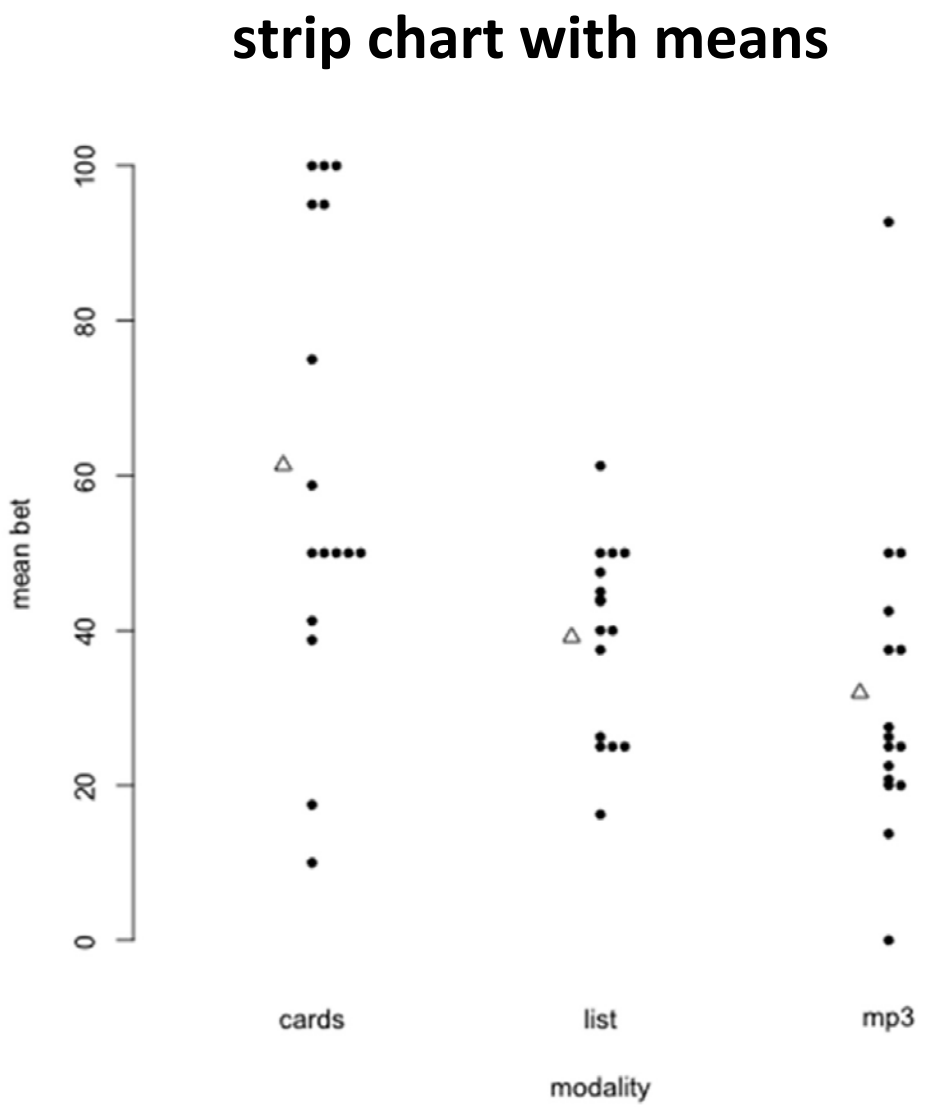
Worked example



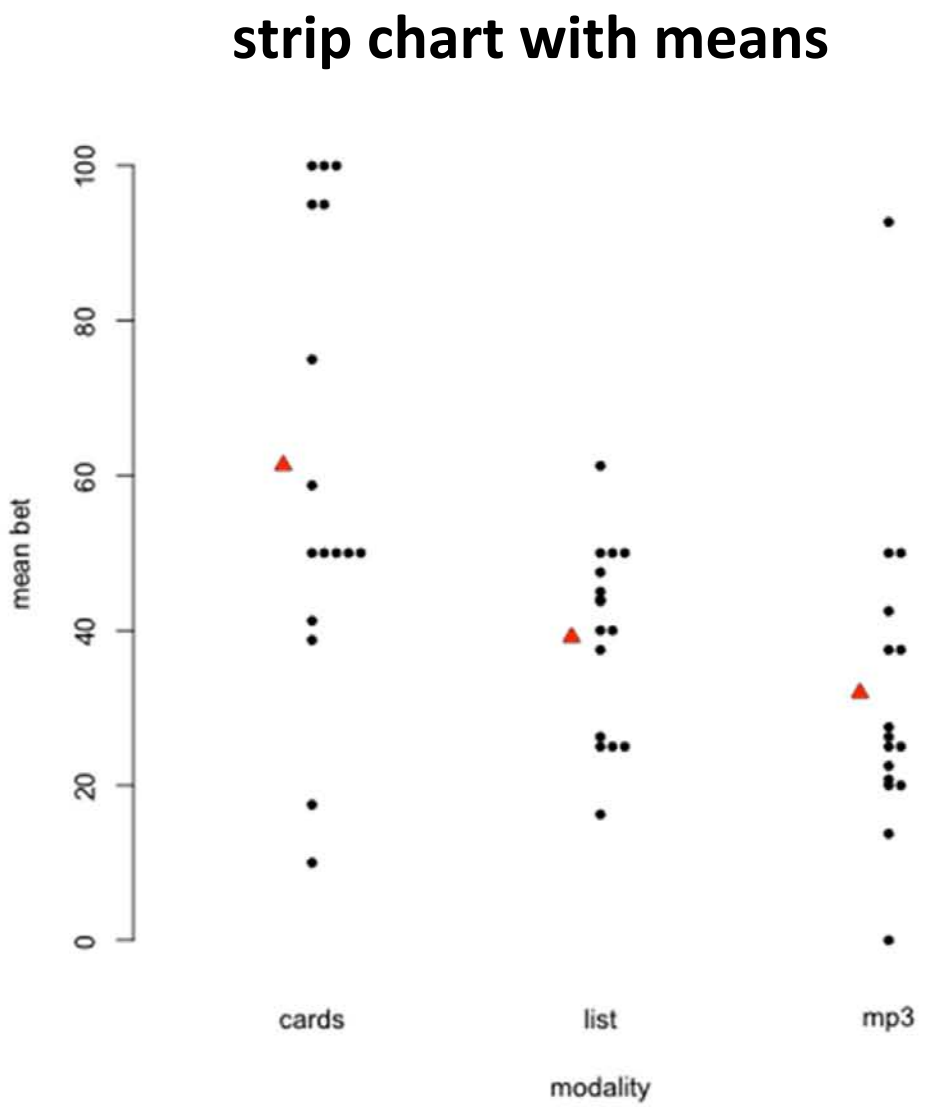
Worked example



Worked example



Worked example



Morals of the example

- Summary statistics
  - almost always necessary
  - but at what level of analysis?
- Distribution is important
  - what is the form of the data?
  - is your summary misleading?
- Fancier is not always better
  - pretty pictures are awesome
  - but not if they obscure the data

Anscombe's Quartet

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

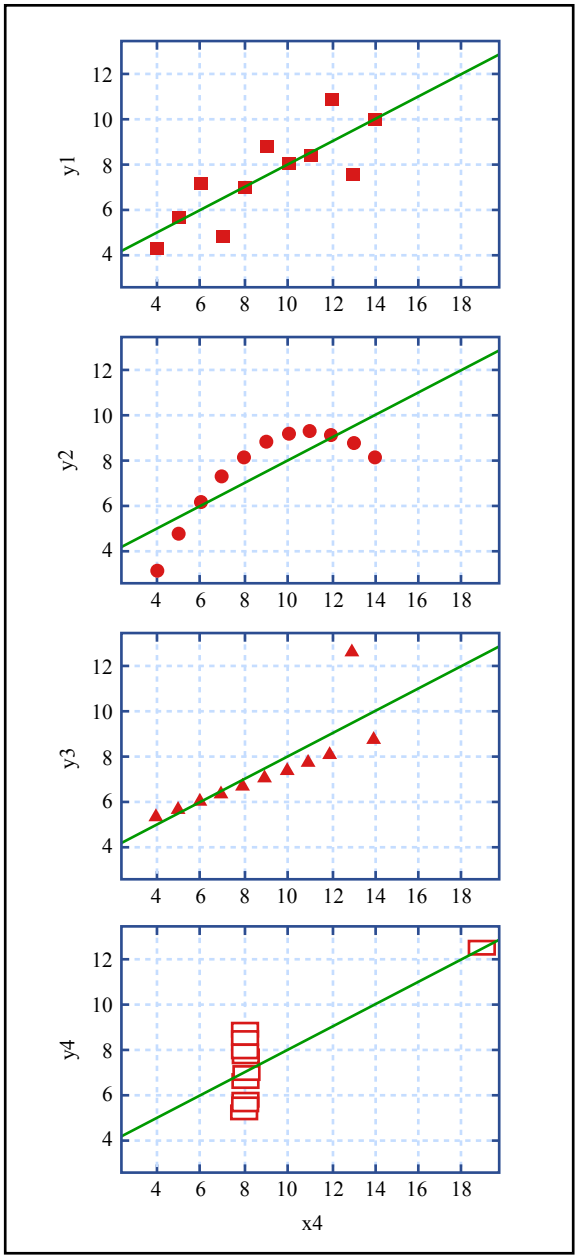
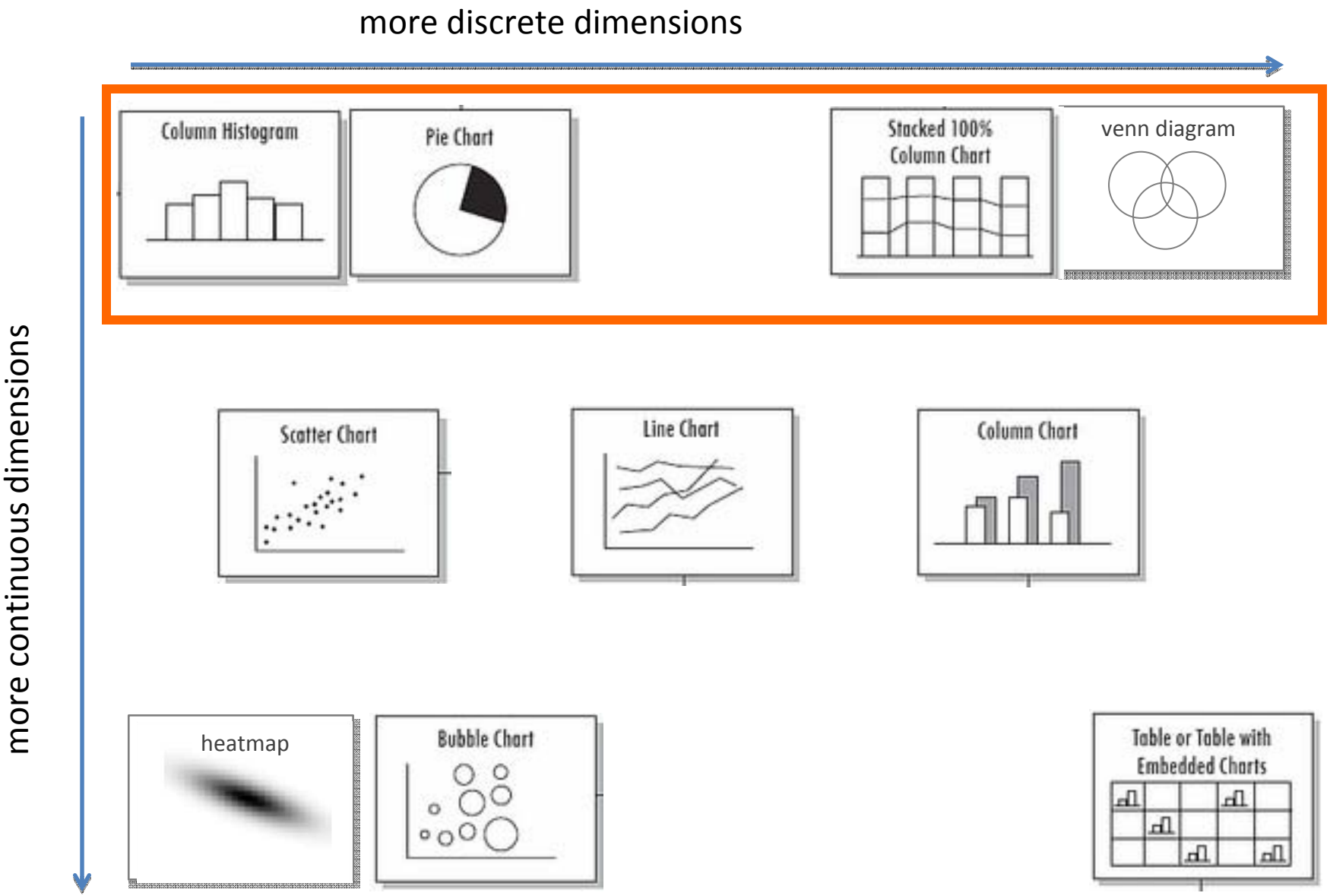


Figure by MIT OpenCourseWare.

F. J. Anscombe, 1973

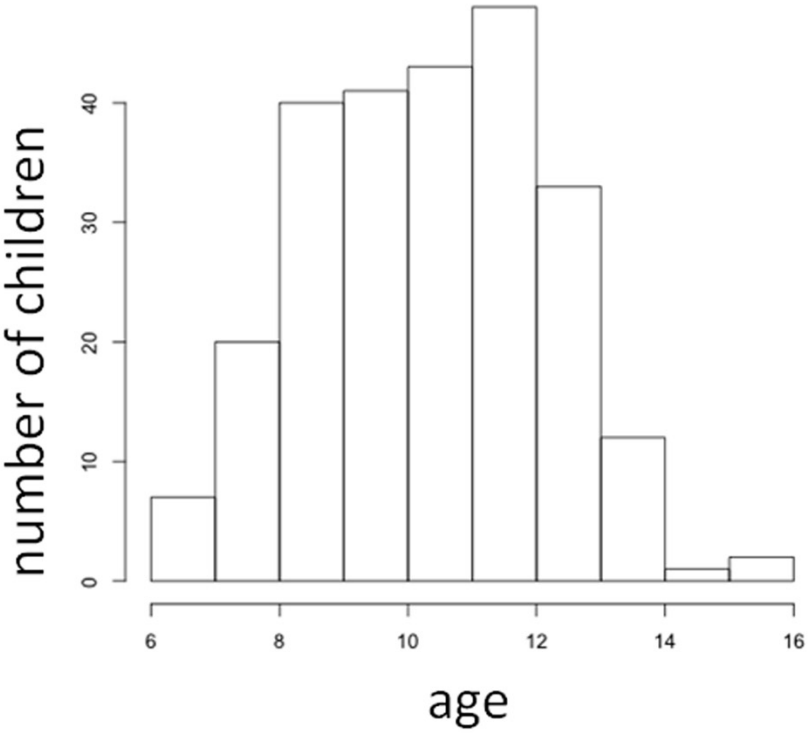
Conventional visualizations



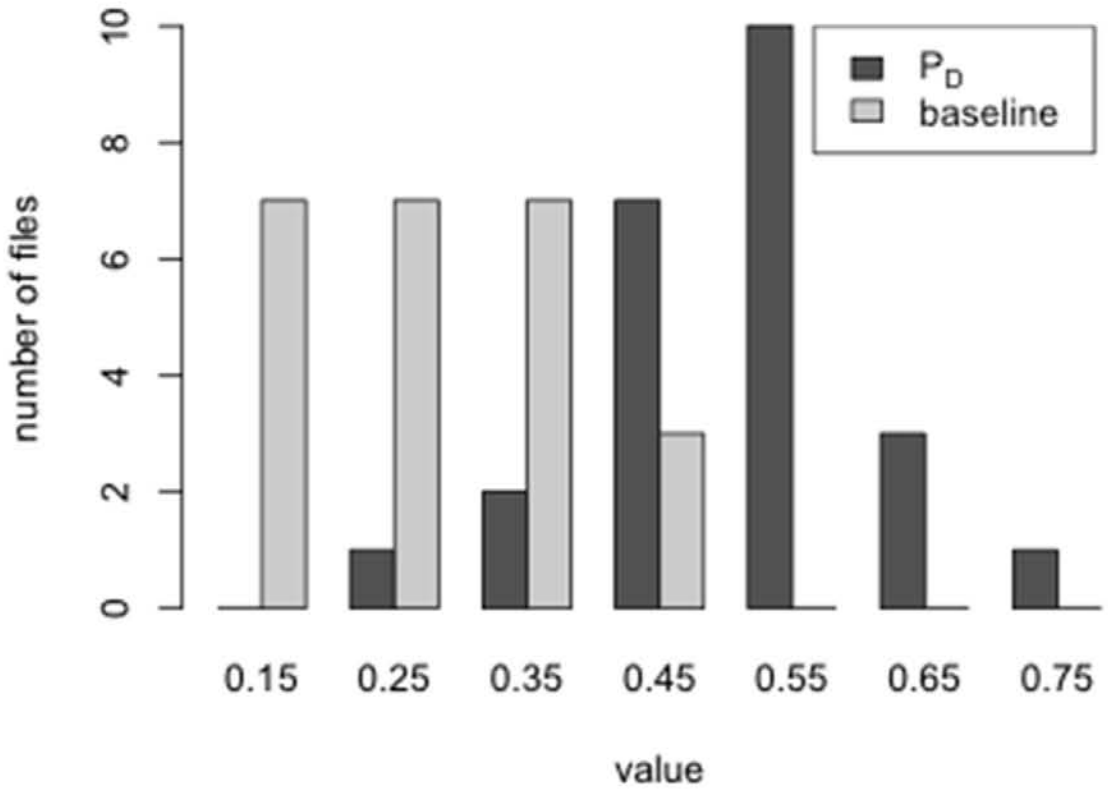
Courtesy of Amit Agarwal. Used with permission.

Histogram

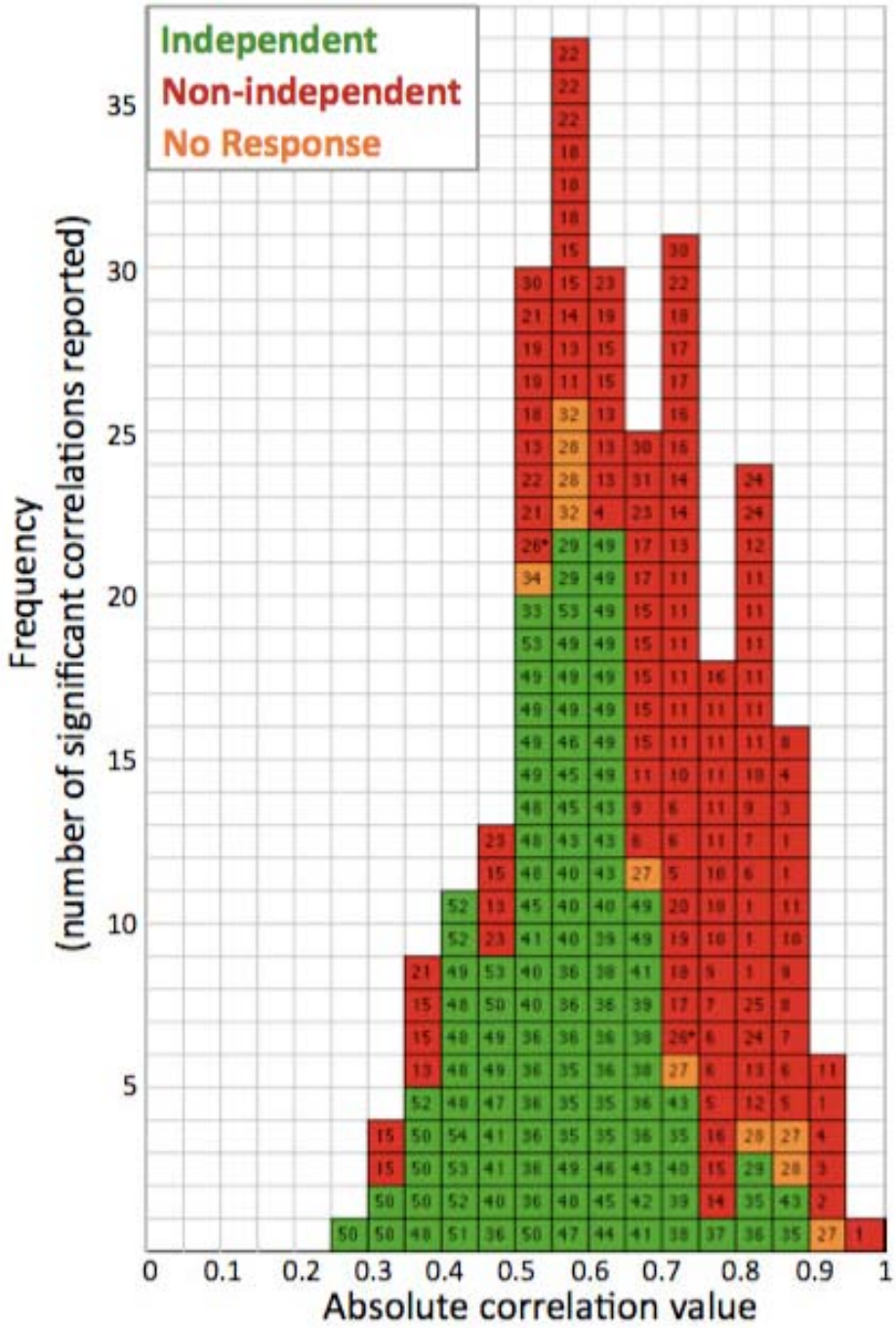
- Important first way of looking at your data
- One dimensional
- Shows shape by binning a continuous distribution



Grouped histogram



Grouped histogram



Pie chart

- A whole split into parts
- Emphasizes that all parts sum to a constant
- Single dimension with discrete categories

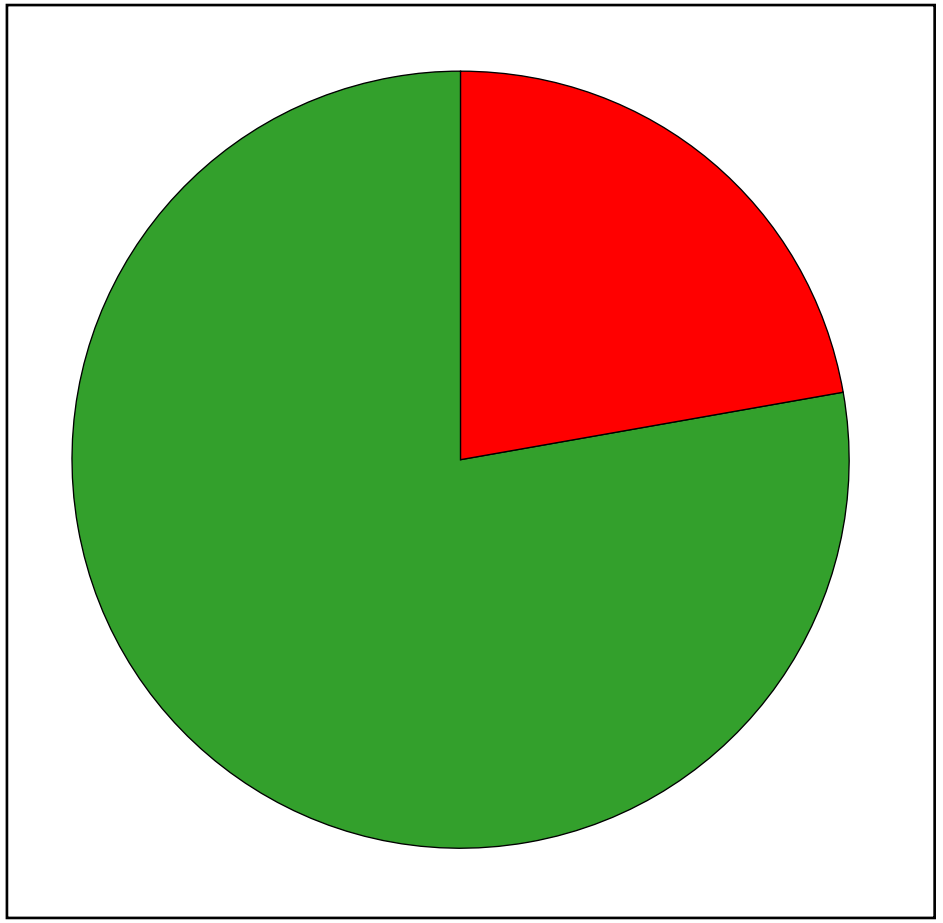
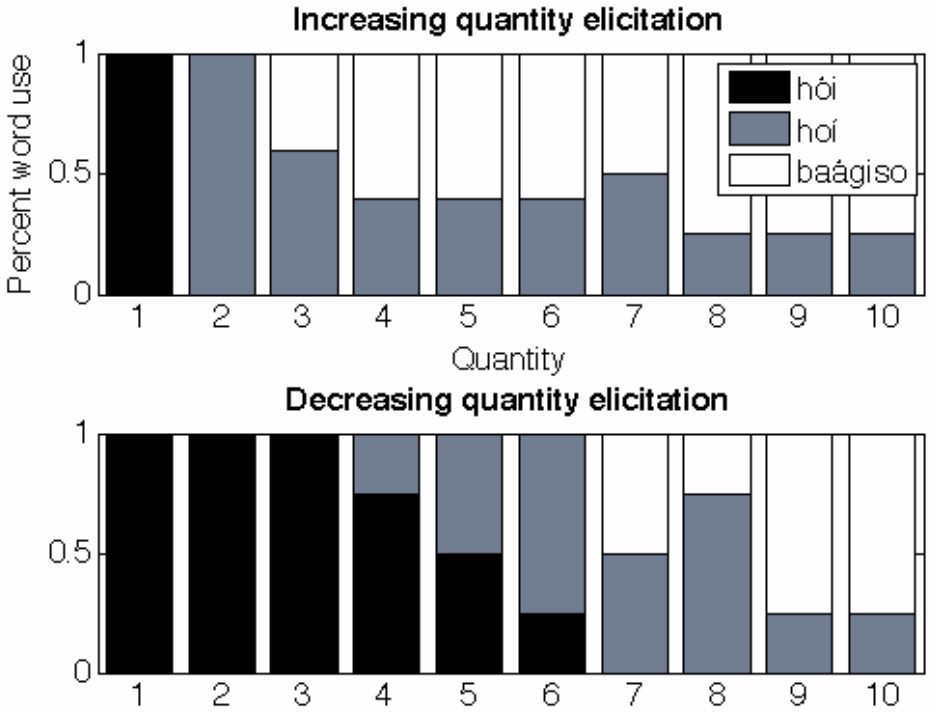


Figure by MIT OpenCourseWare.

Stacked bar graph

- Wholes split into parts
- Easy to compare
  - often better than pie chart
- Can have multiple discrete dimensions



Venn diagram

- Shows overlap between discrete groups
- Sometimes the only way to display overlapping sets
- Unintuitive
  - no “popout”

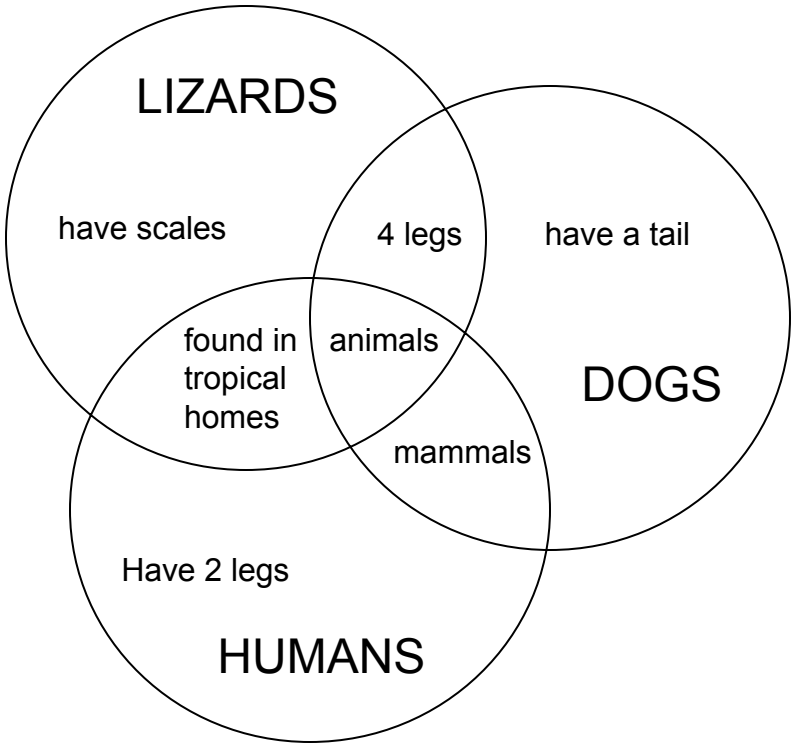
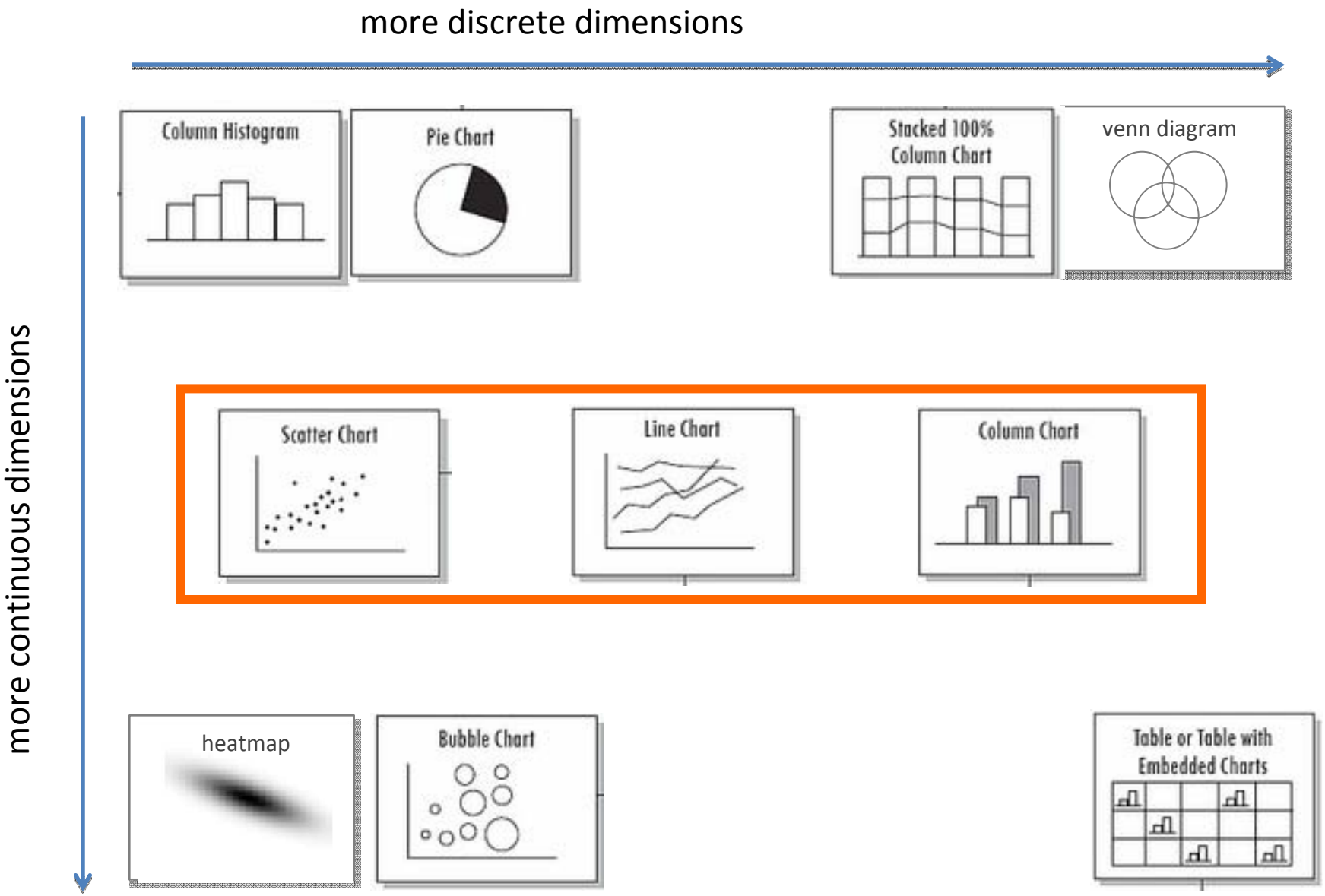


Figure by MIT OpenCourseWare.

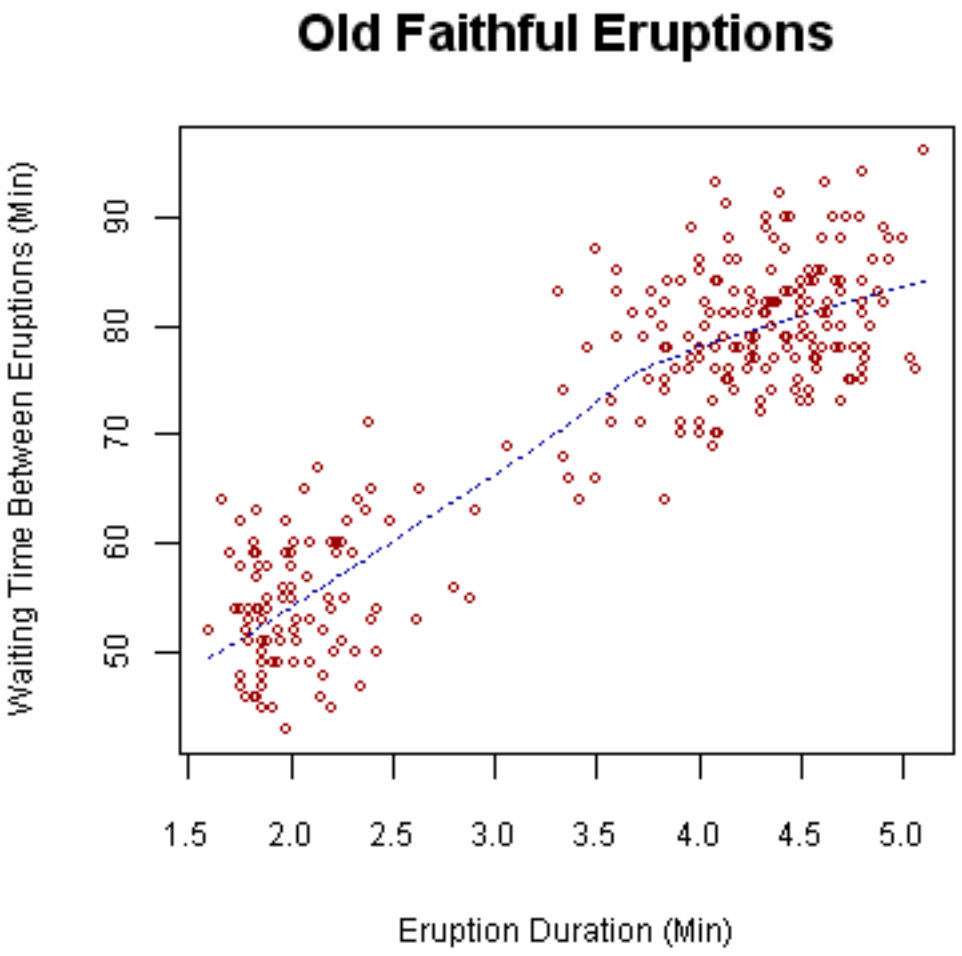
Conventional visualizations



Courtesy of Amit Agarwal. Used with permission.

Scatter plot

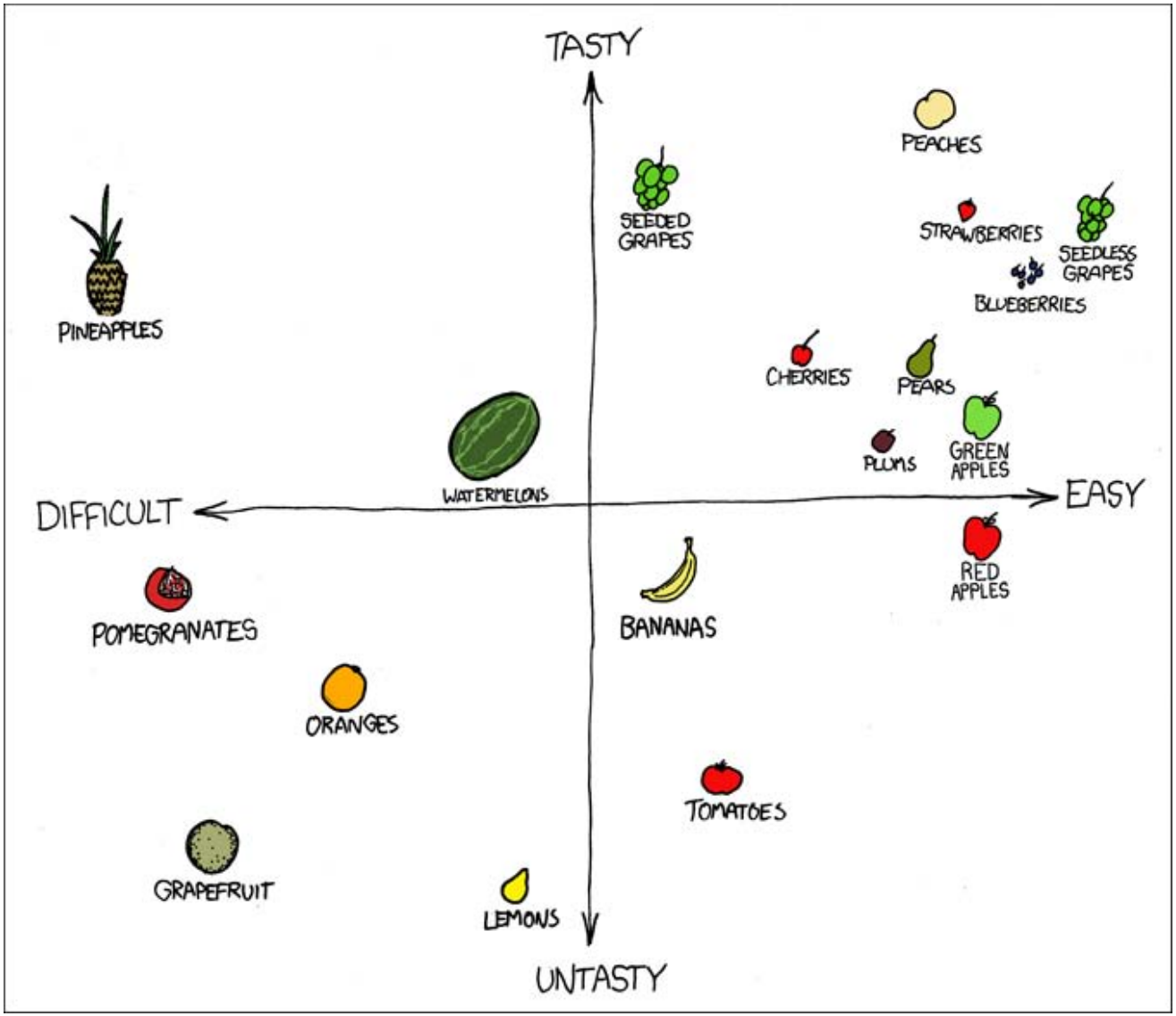
- Relationship between observations on two continuous dimensions
- Can show multiple groups
- Can show trend lines etc.
- Uninformative with too much data



Courtesy of xkcd.org.

Scatter plot

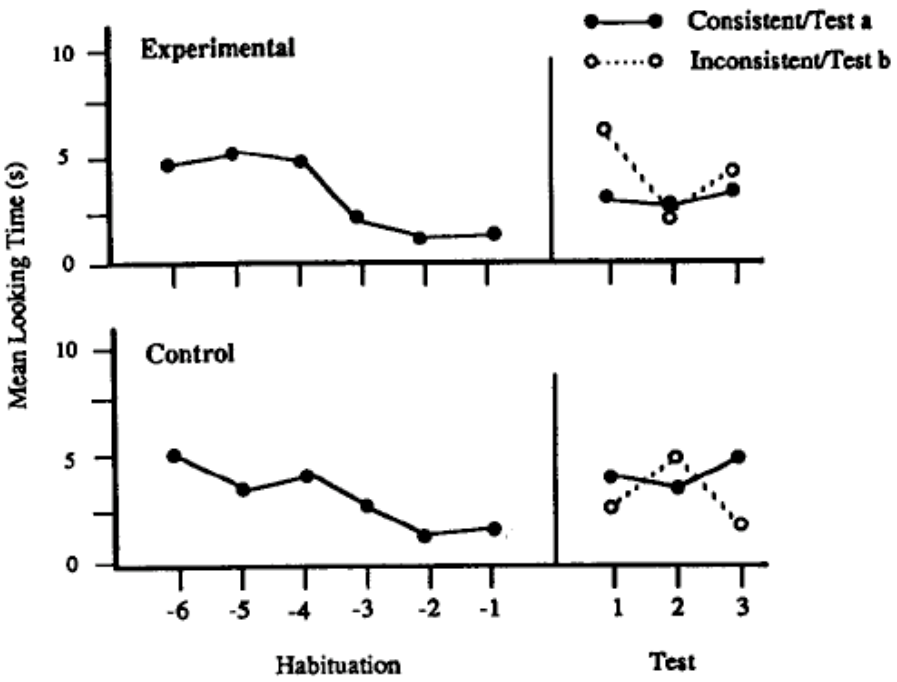
… with many discrete items (identity as a dimension)



XKCD

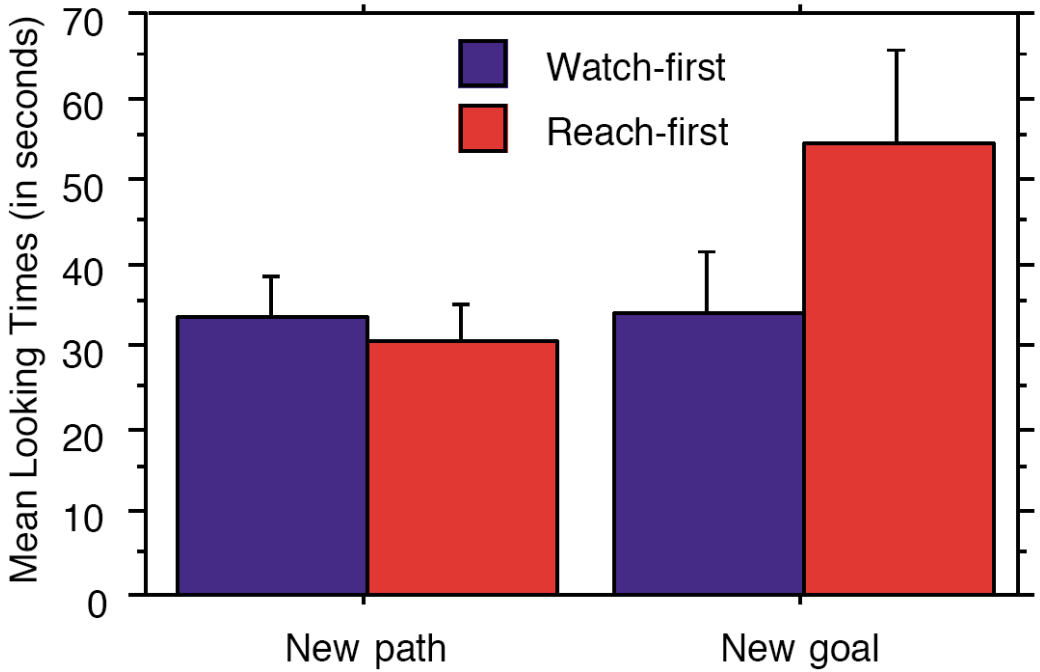
Line graph

- Also ubiquitous!
- Good for showing one variable (e.g., time) as continuous even though you have discrete measures
- Can compare several discrete groups



Bar graph

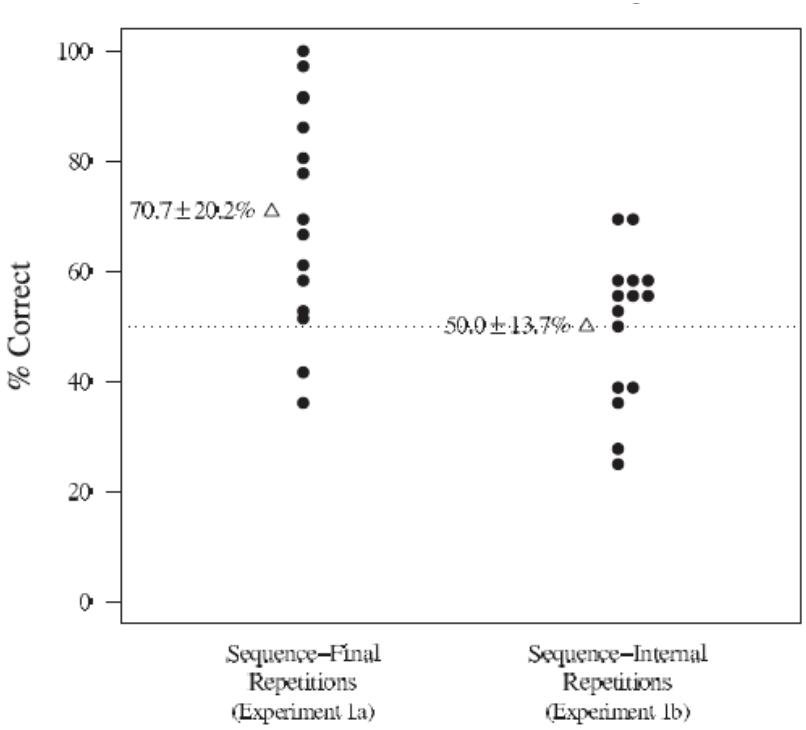
- aka “dynamite plot”
- Ubiquitous!
- Can be used for lots of discrete grouping factors
- Natural semantics of grouping
- Conceals data



Sommervilleet al. (2005)

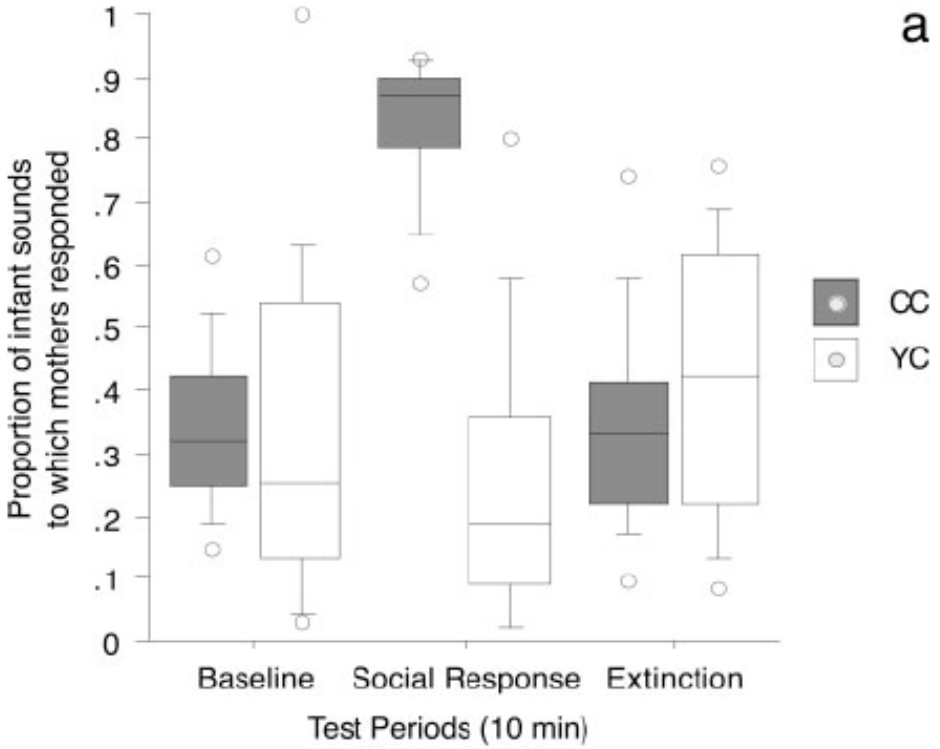
More bar graphs

Strip chart



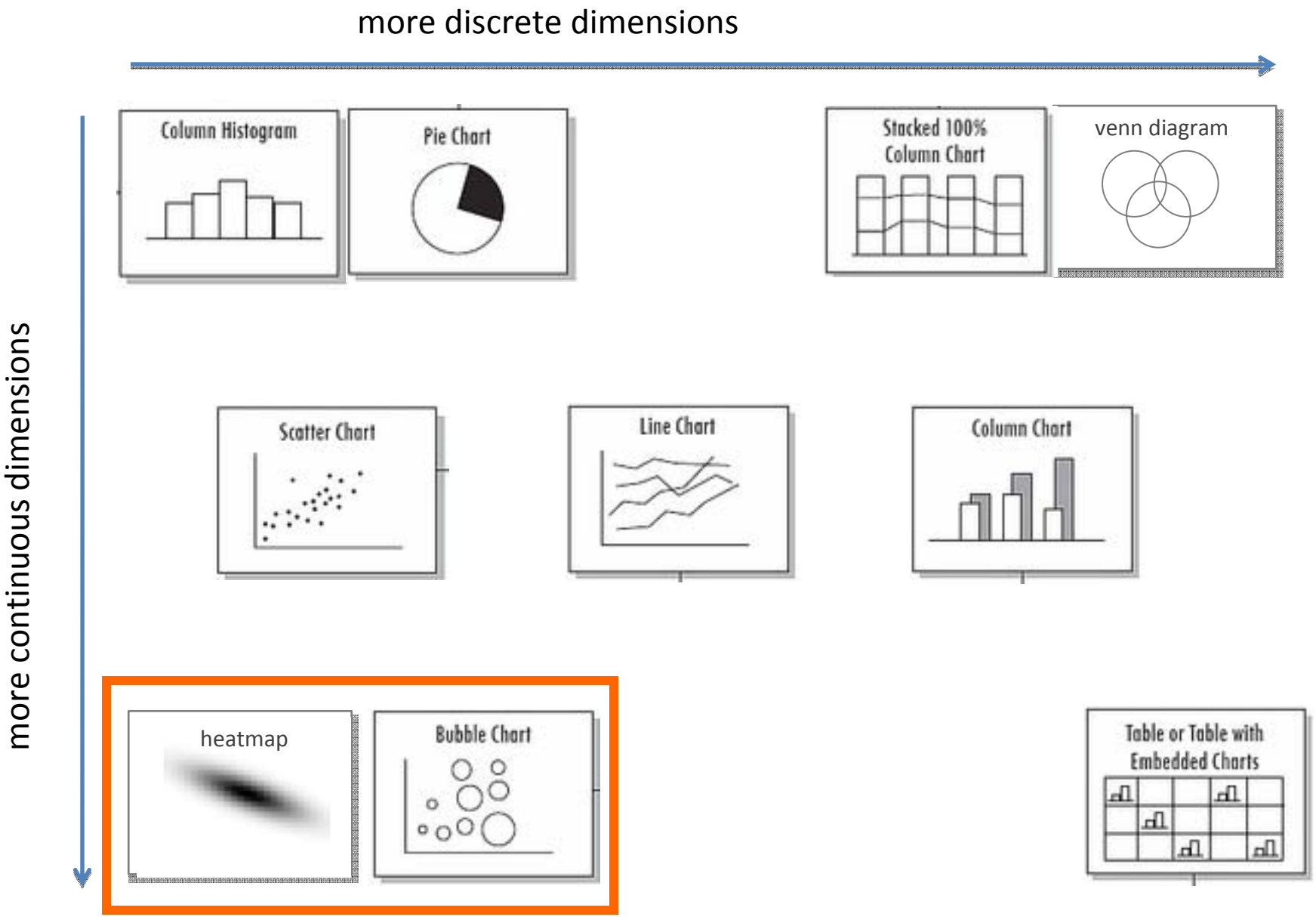
Very useful for showing individual subject means

Box plot



Shows the shape of distribution but not focused on individual subjects

Conventional visualizations

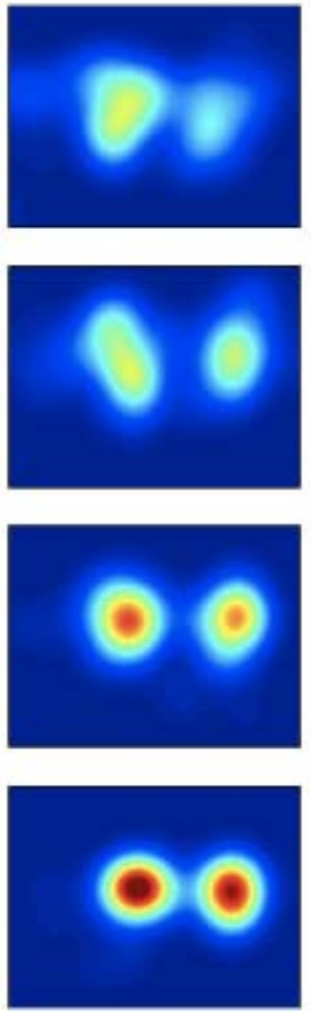


Courtesy of Amit Agarwal. Used with permission.

Heat map

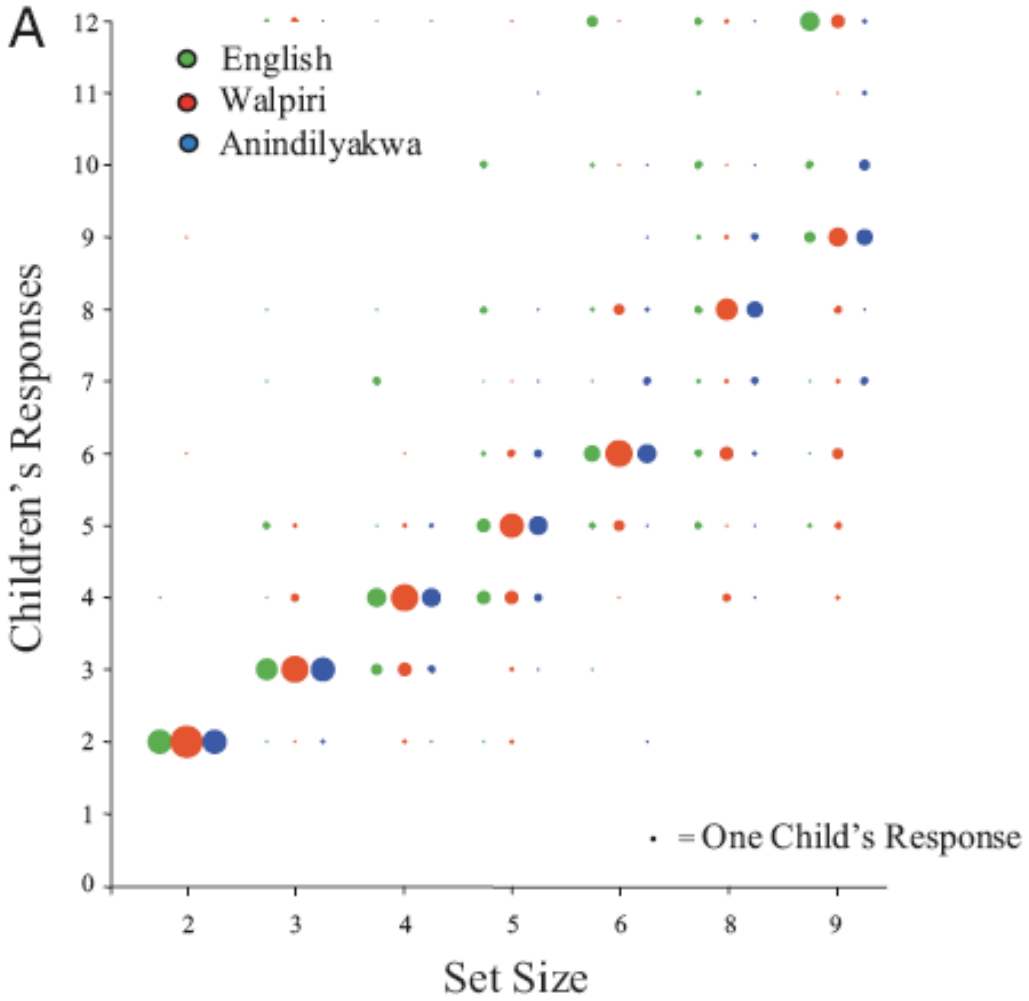
- Worksvery well when there are natural semantics
- Color mapping can be problematic
  - grayscale usually fine
- Can be unintuitive

Image removed due to copyright restriction.
http://tedlab.mit.edu/~mcfrank/papers/FVJ-cognition.pdf

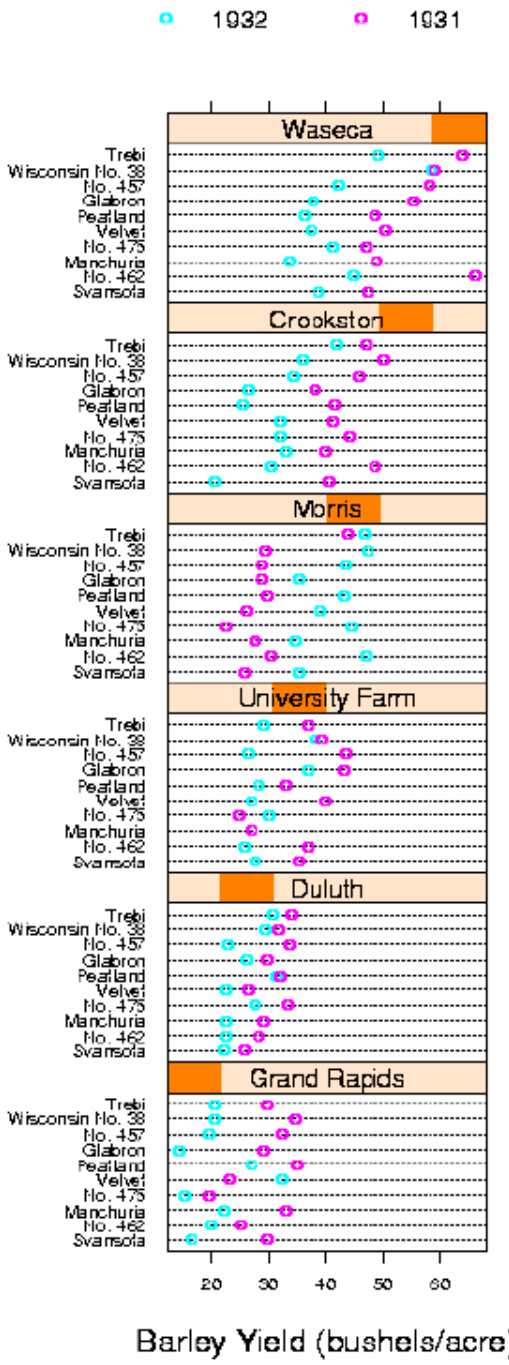


Bubble plot

- Can be very intuitive
- Size is not perfectly quantitative



Trellis plots

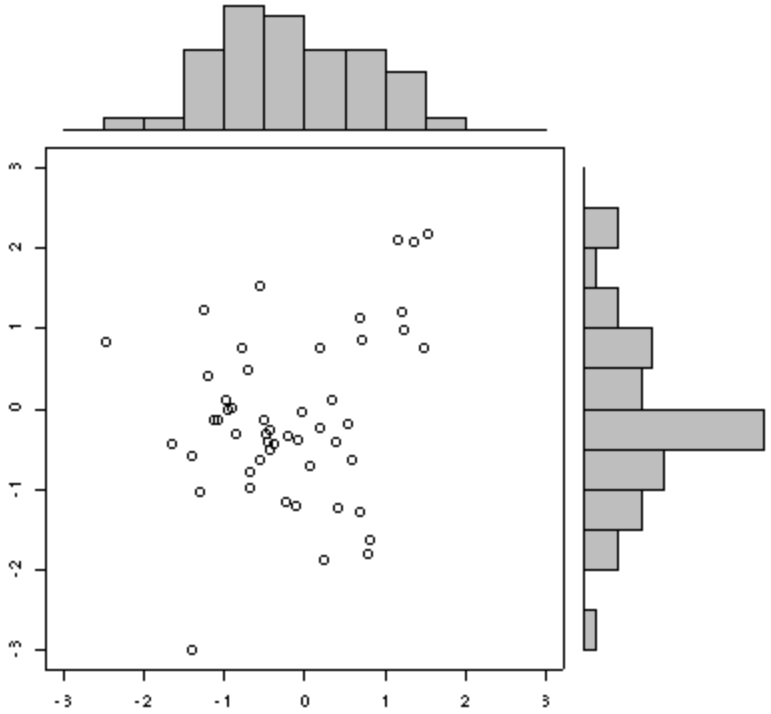


TIPS AND TRICKS

Three tricks for doing more with less

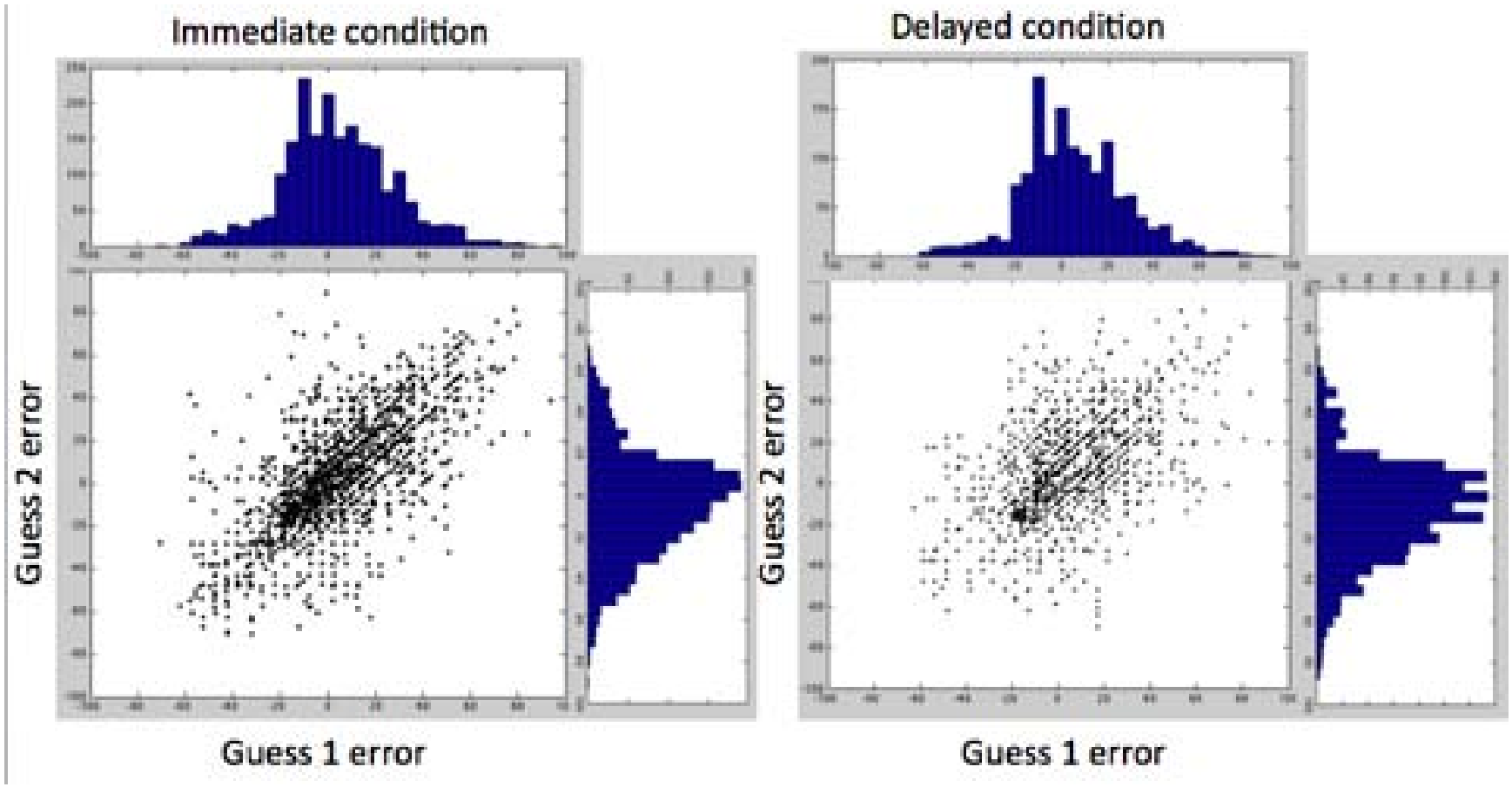
- Multiple plots
  - simple, easily interpretable subplots
  - can be beautiful but overwhelming
- Hybrid plots
  - a scatter plot of histograms
  - or a venn-diagram of histograms, etc.
- Multiple axes
  - plot two (or more) different things on one graph

Hybrid plots

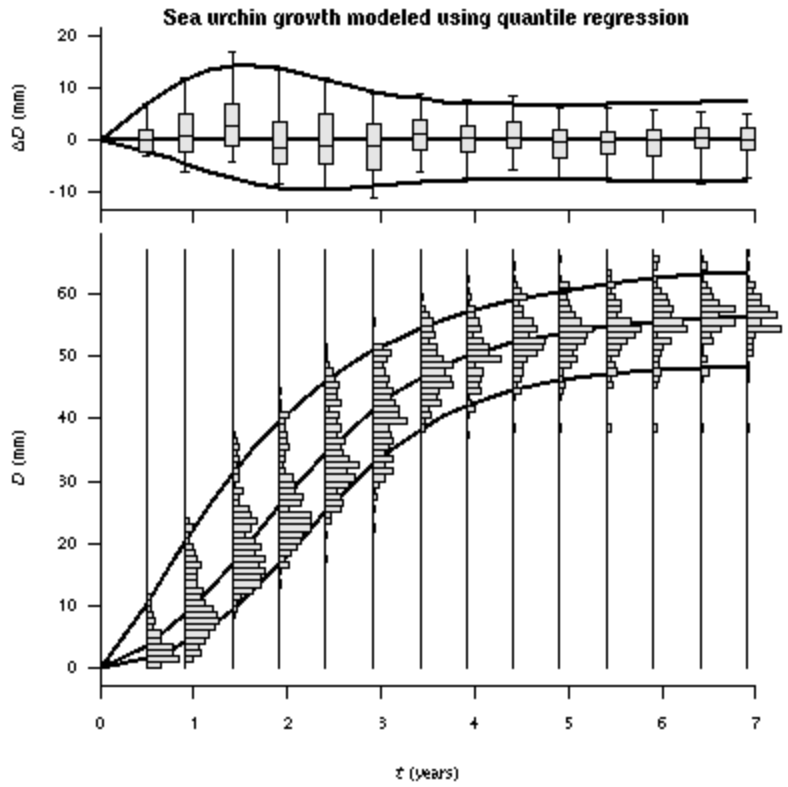


Courtesy of http://addictedtor.free.fr/graphiques/addNote.php?graph=78

Hybrid plots



Hybrid plots



Courtesy of http://addictedtor.free.fr/graphiques/addNote.php?graph=109

Multiple plots

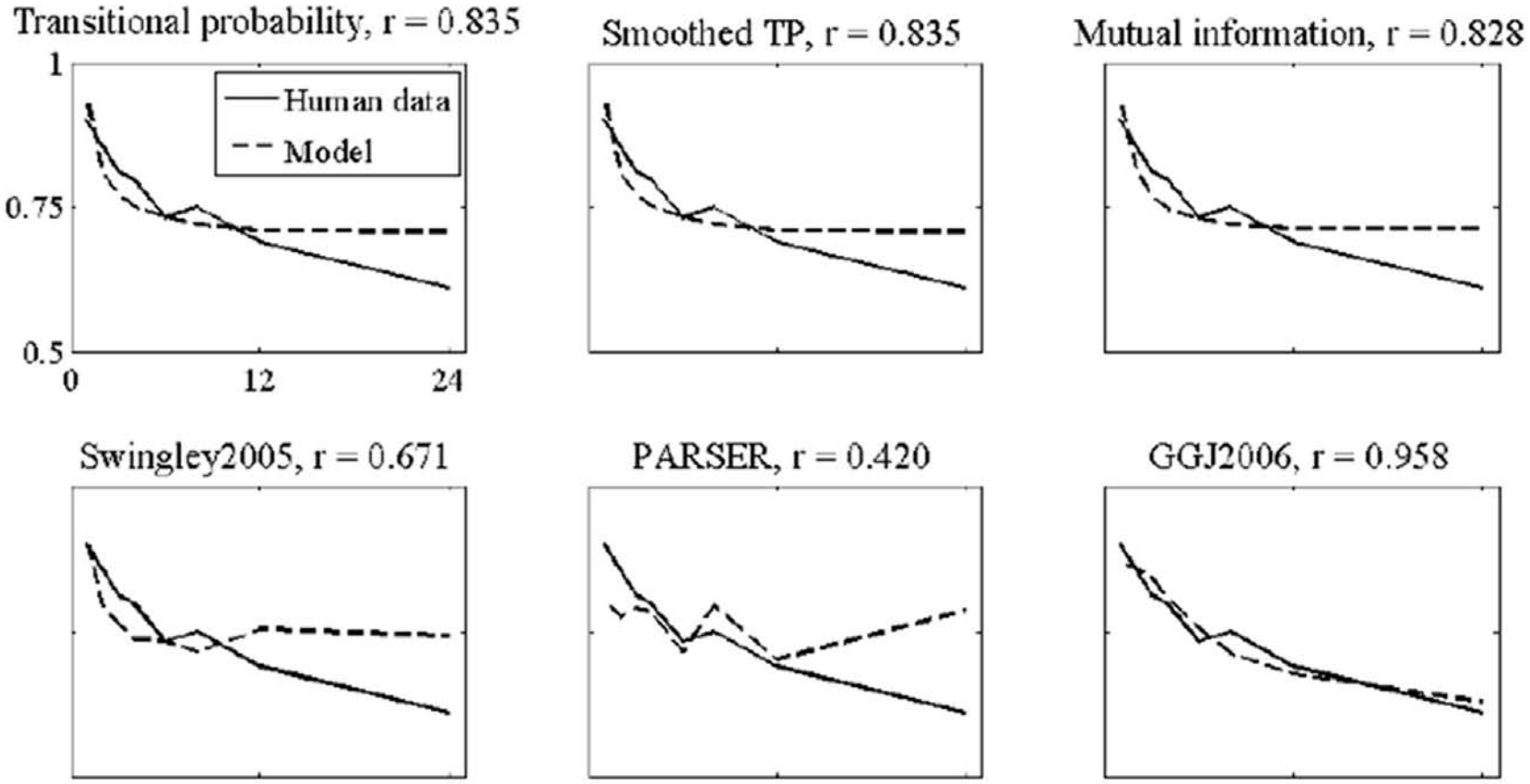


Figure 2. Best linear fit of each model's performance to human data, graphed by sentence length. The vertical axis represents decision probabilities for models and percentage correct for human data; the horizontal, sentence length.

Multiple plots

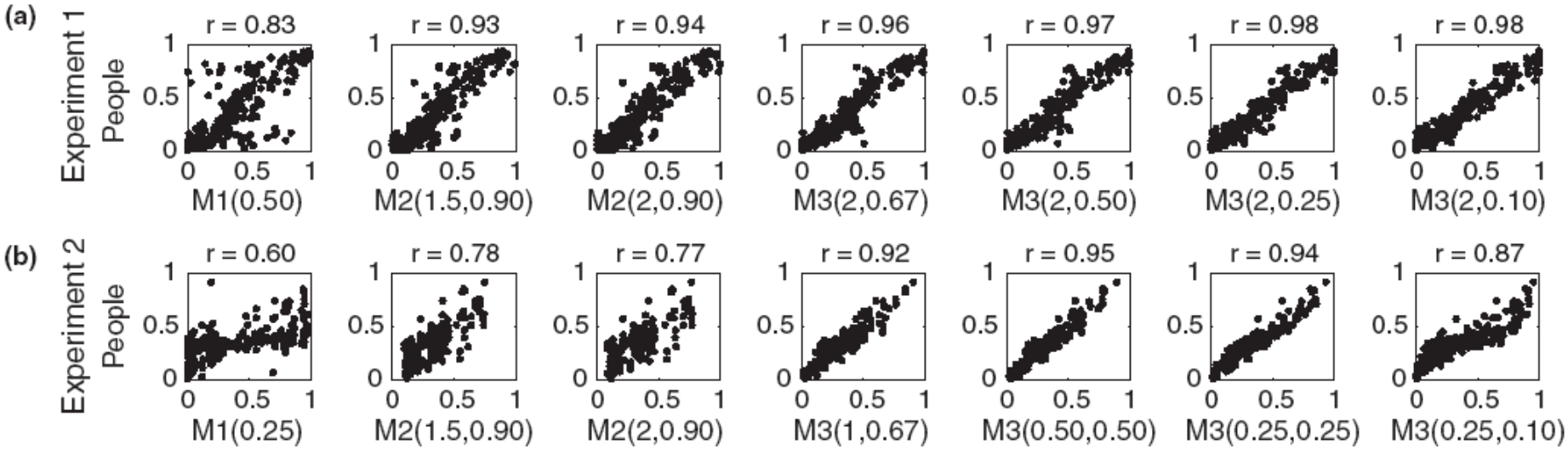
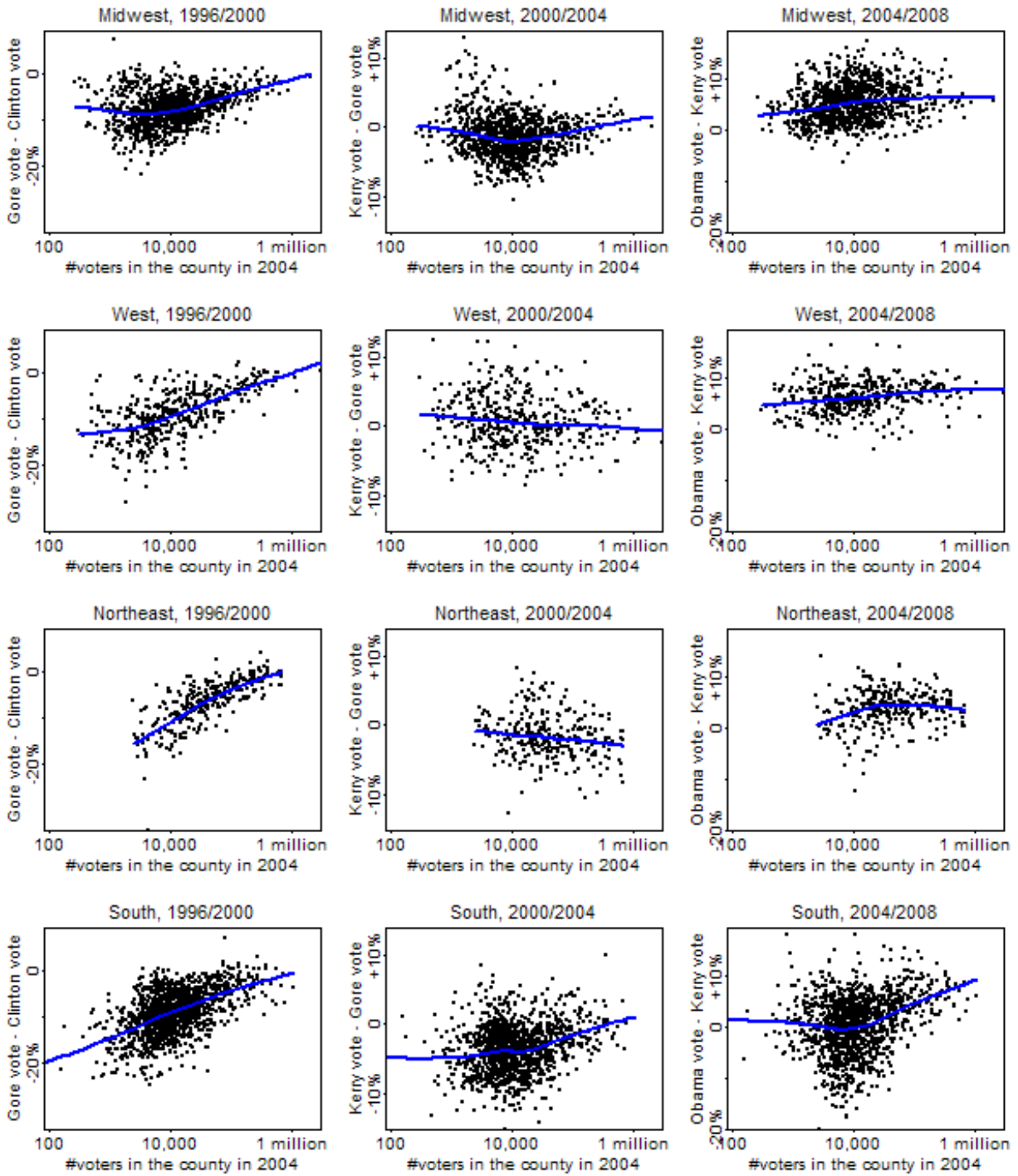
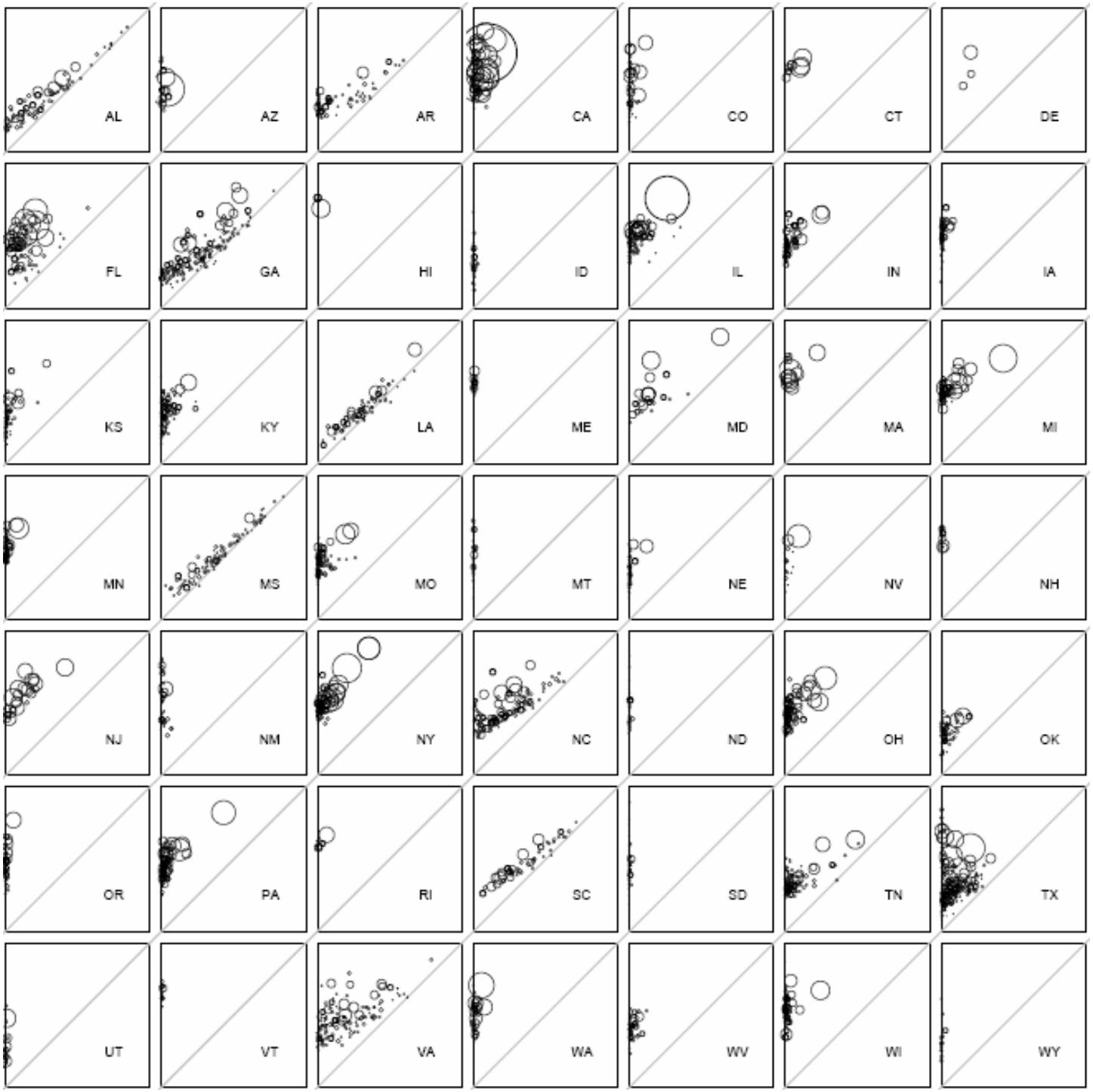


Figure 3: Example scatter plots of model predictions against subject ratings. Plots of model predictions use the parameter settings with the highest correlation from each model column of Tables 1 and 2. (a) Experiment 1 results. (b) Experiment 2 results.

Baker, Tenenbaum, & Saxe (2007)

Multiple plots





Courtesy of Andrew Gelman. Used with permission.

Multiple axes

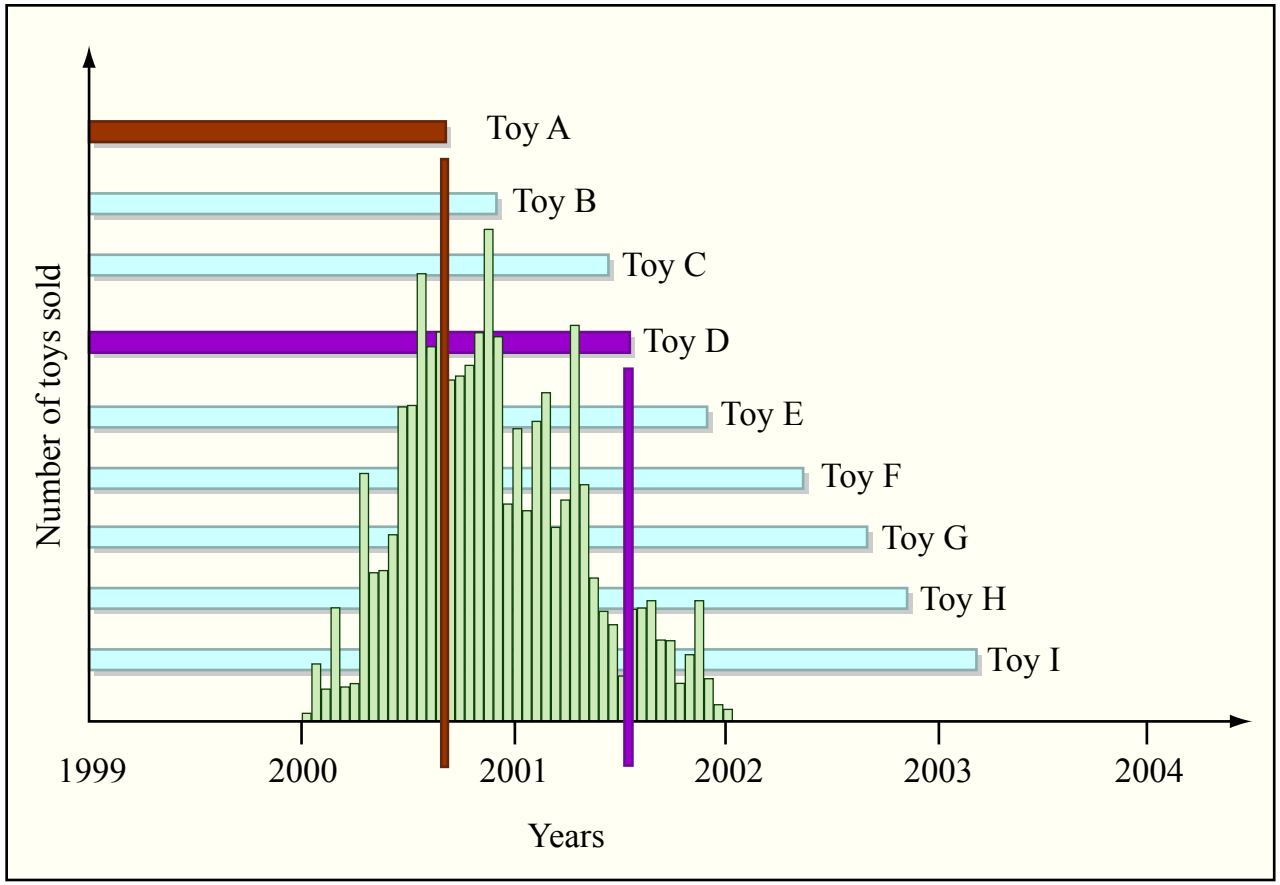
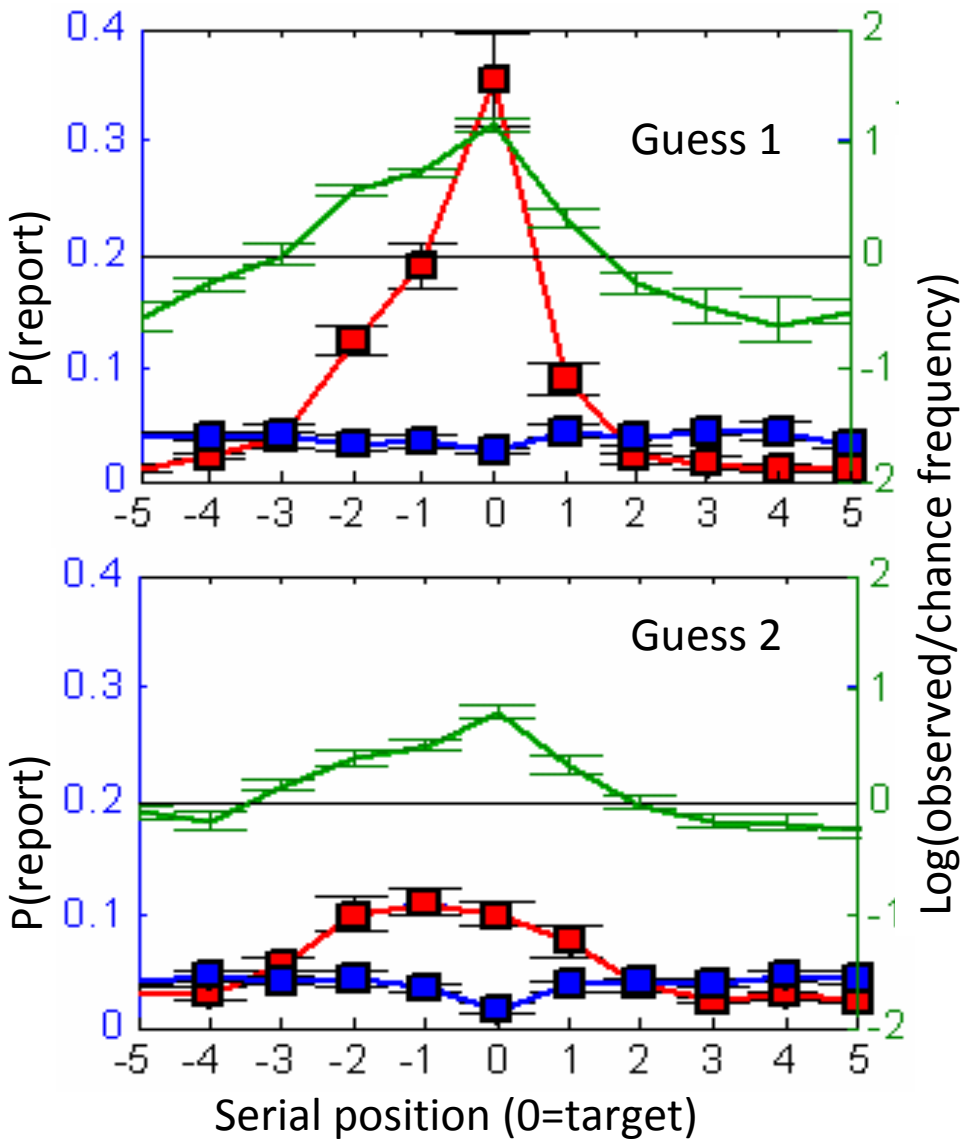


Figure by MIT OpenCourseWare.

Multiple axes



TWO TRADEOFFS

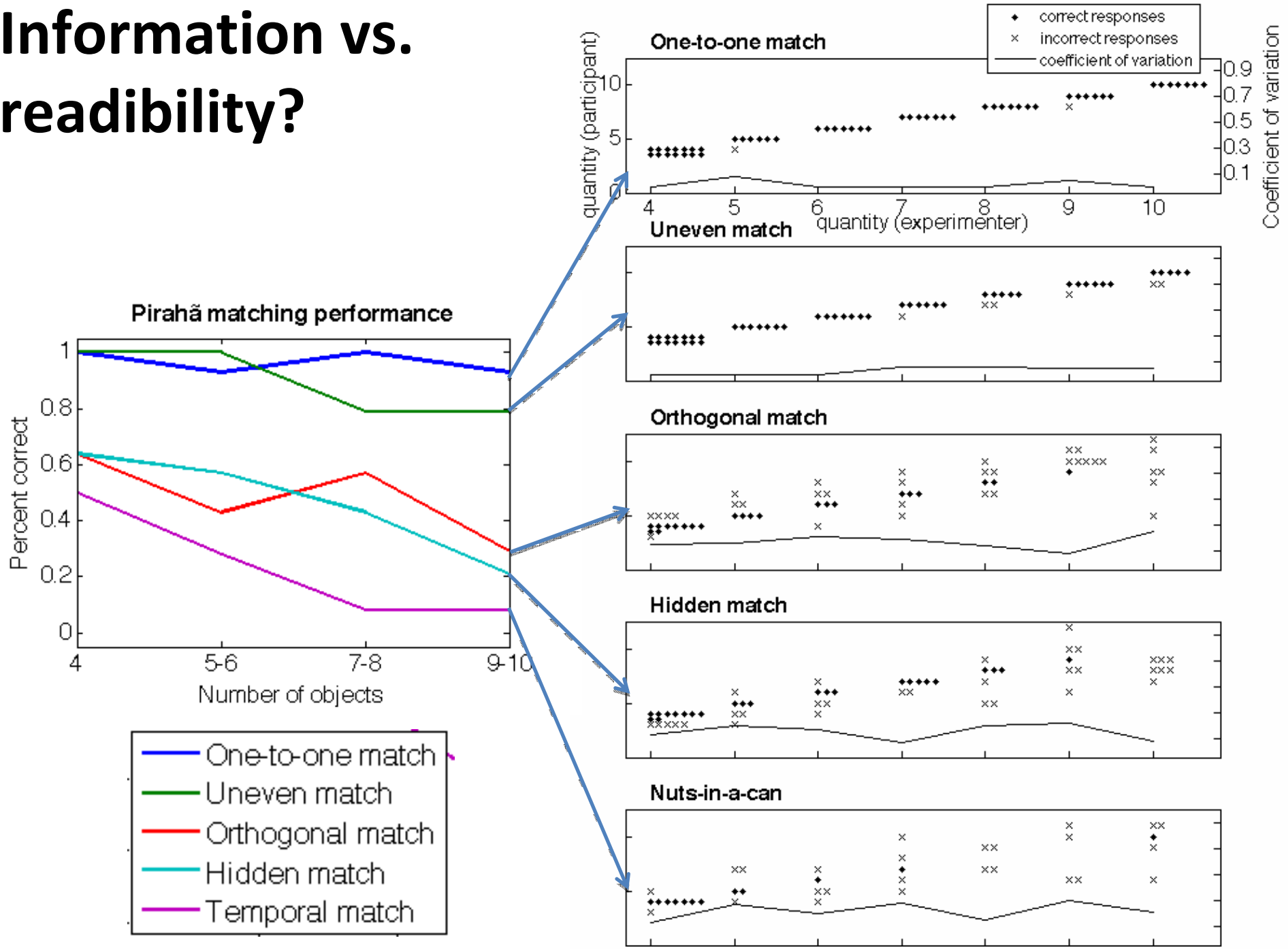
Two tradeoffs

- Informativeness vs. readability
  - Too little information can conceal data
  - But too much information can be overwhelming
  - Possible solution: hierarchical organization?
- Data-centric vs. viewer-centric
  - Viewers are accustomed to certain types of visualization
  - But novel visualizations can be truer to data

Information vs. readability

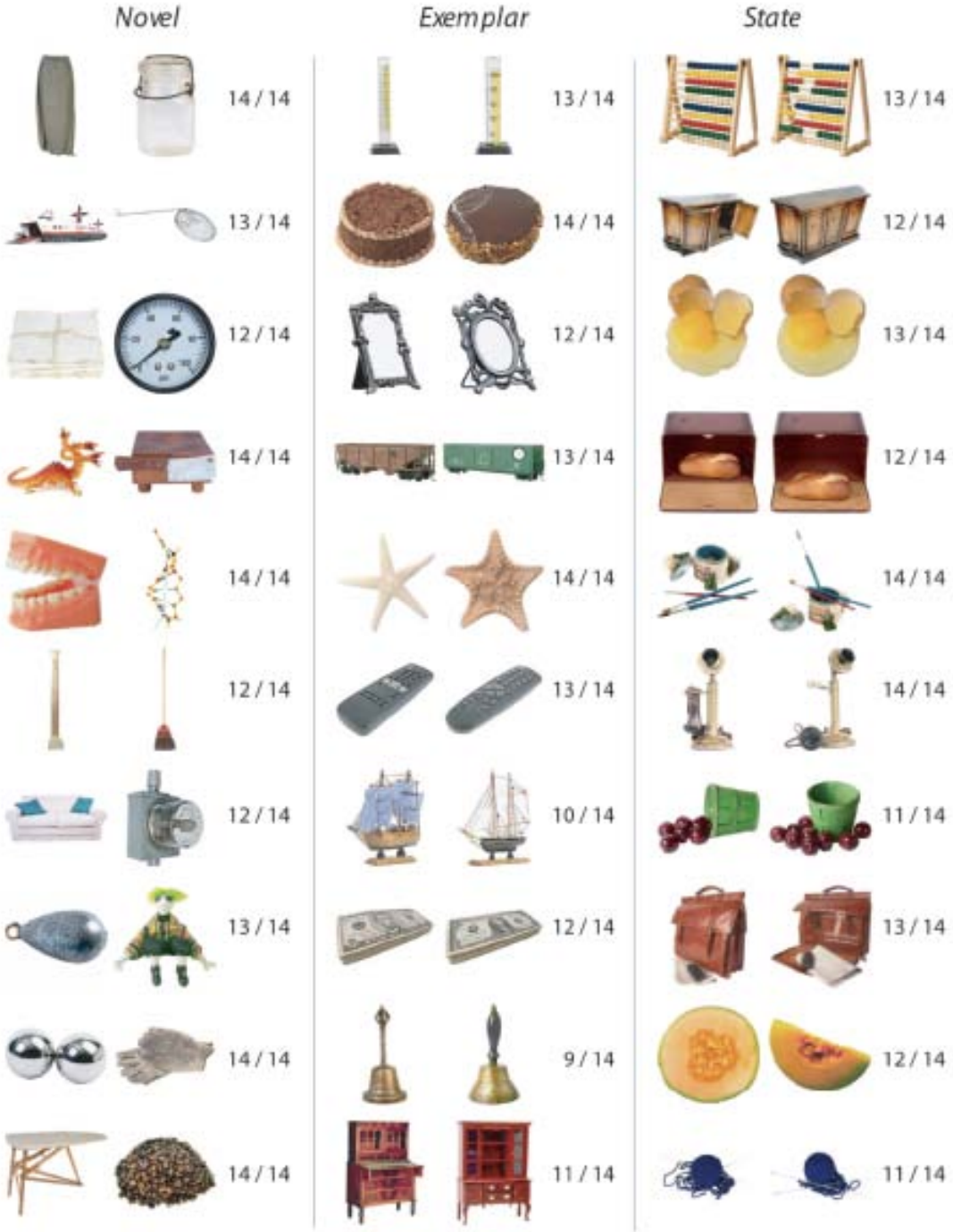
- Pirahã people of Brazil
  - Isolated indigenous group
  - No words for numbers
- Previous research suggested that they were unable to do simple matching games
- Five matching games, 14 participants, quantities 4-10 (split among participants)

Information vs. readibility?

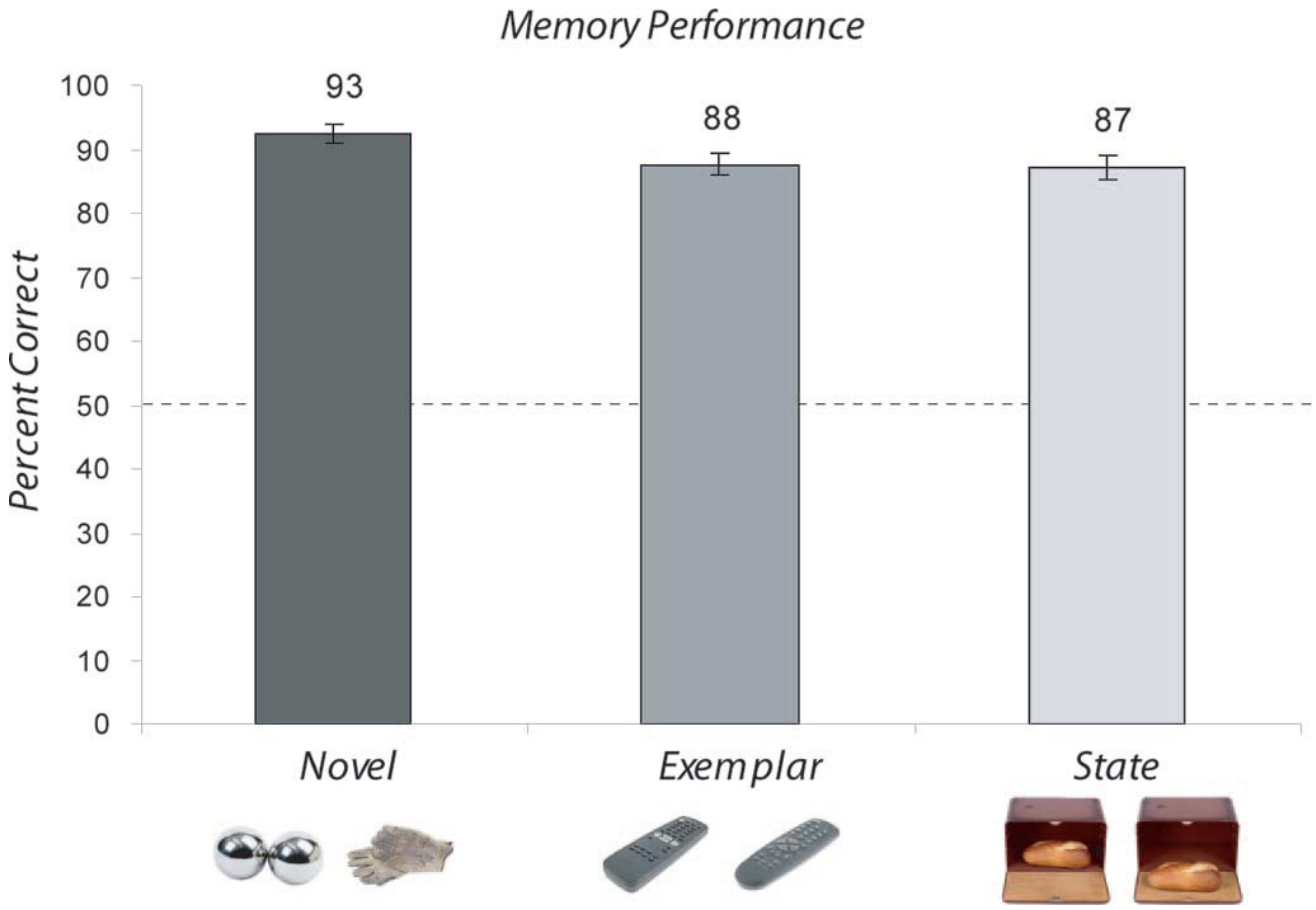


Information vs. readability

Brady, Konkle, Alvarez, Oliva (2008)



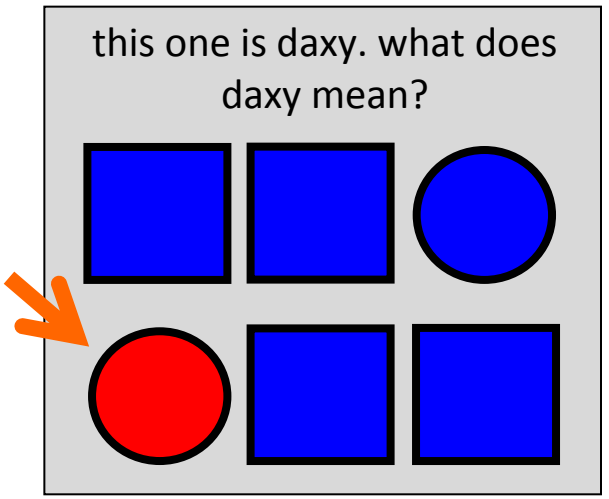
Information vs. readability



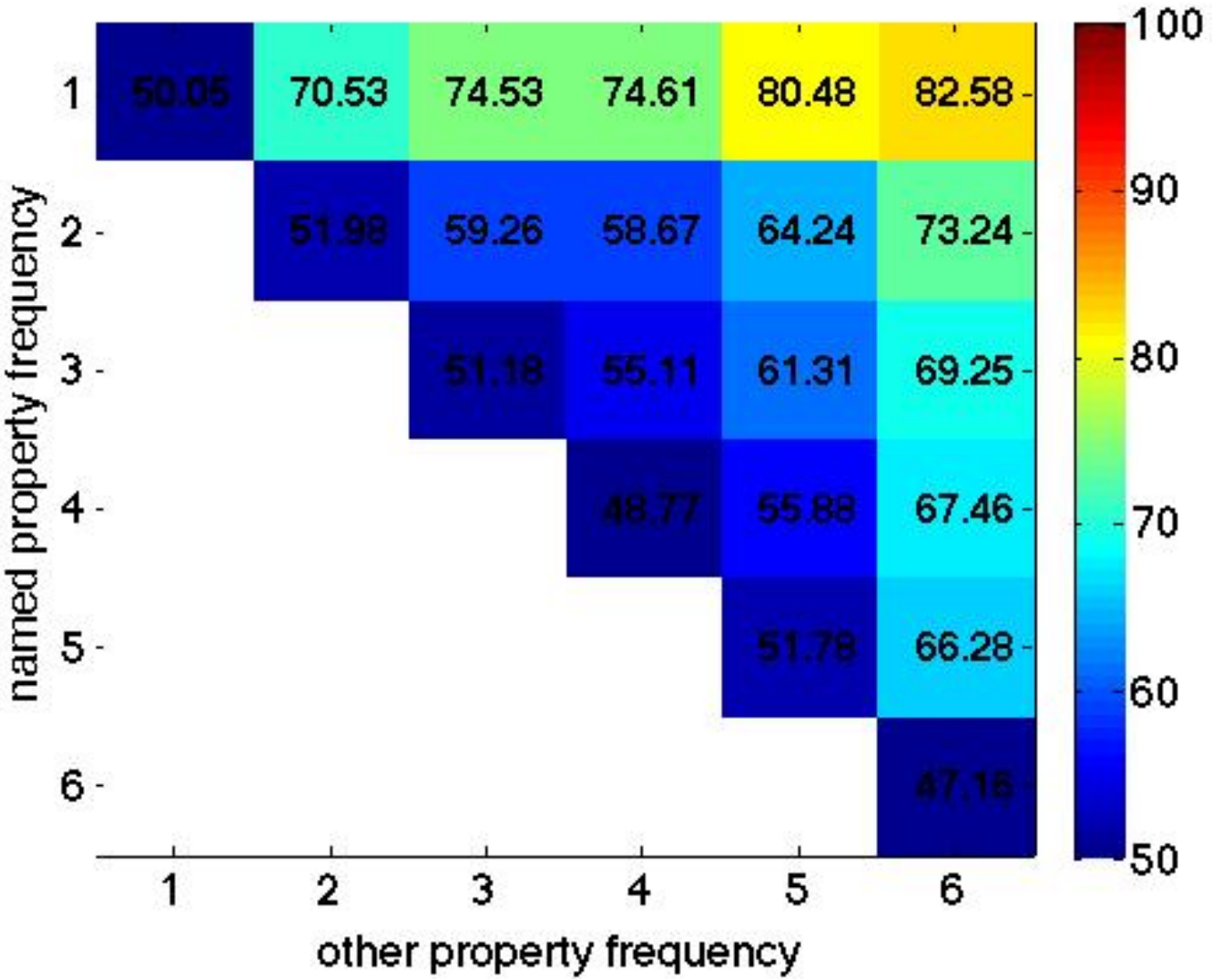
Brady, Konkle, Alvarez, Oliva (2008)

Data-centric vs. viewer-centric

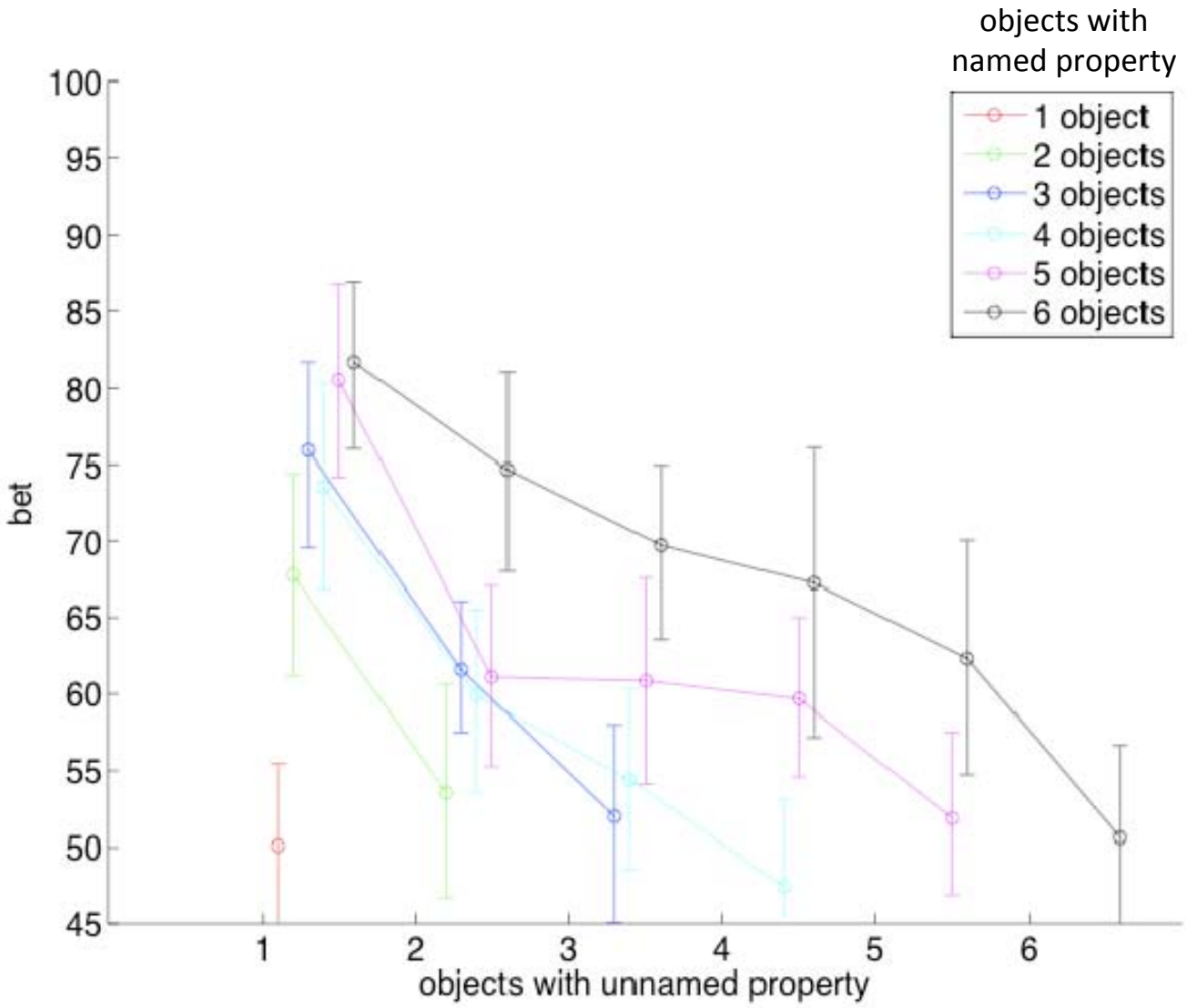
- Web study of word learning
  - n=700
  - lots of noise
- varied number of objects with different properties
  - asked for bets
- had a model that predicted performance



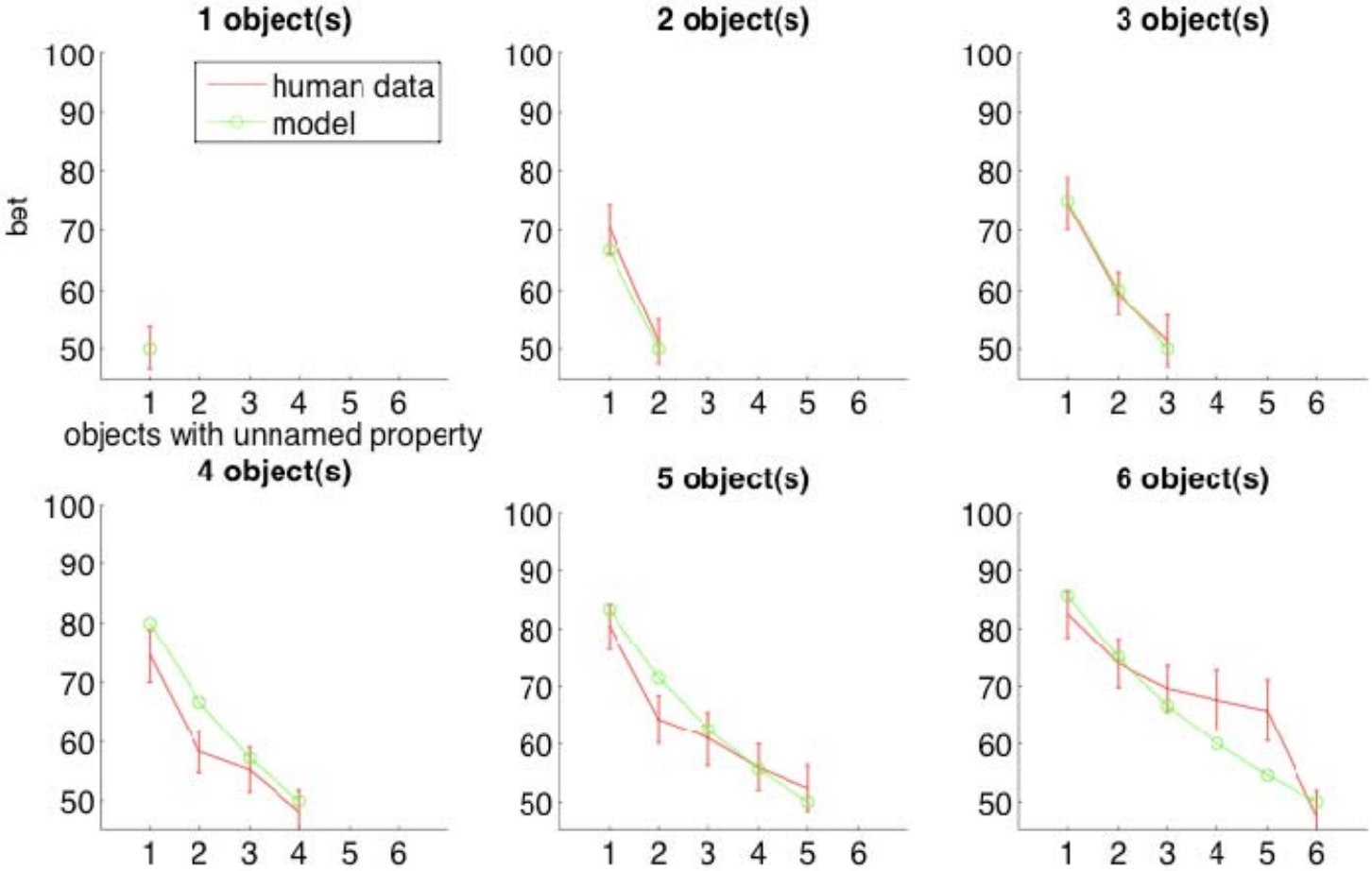
Data-centric vs. viewer-centric

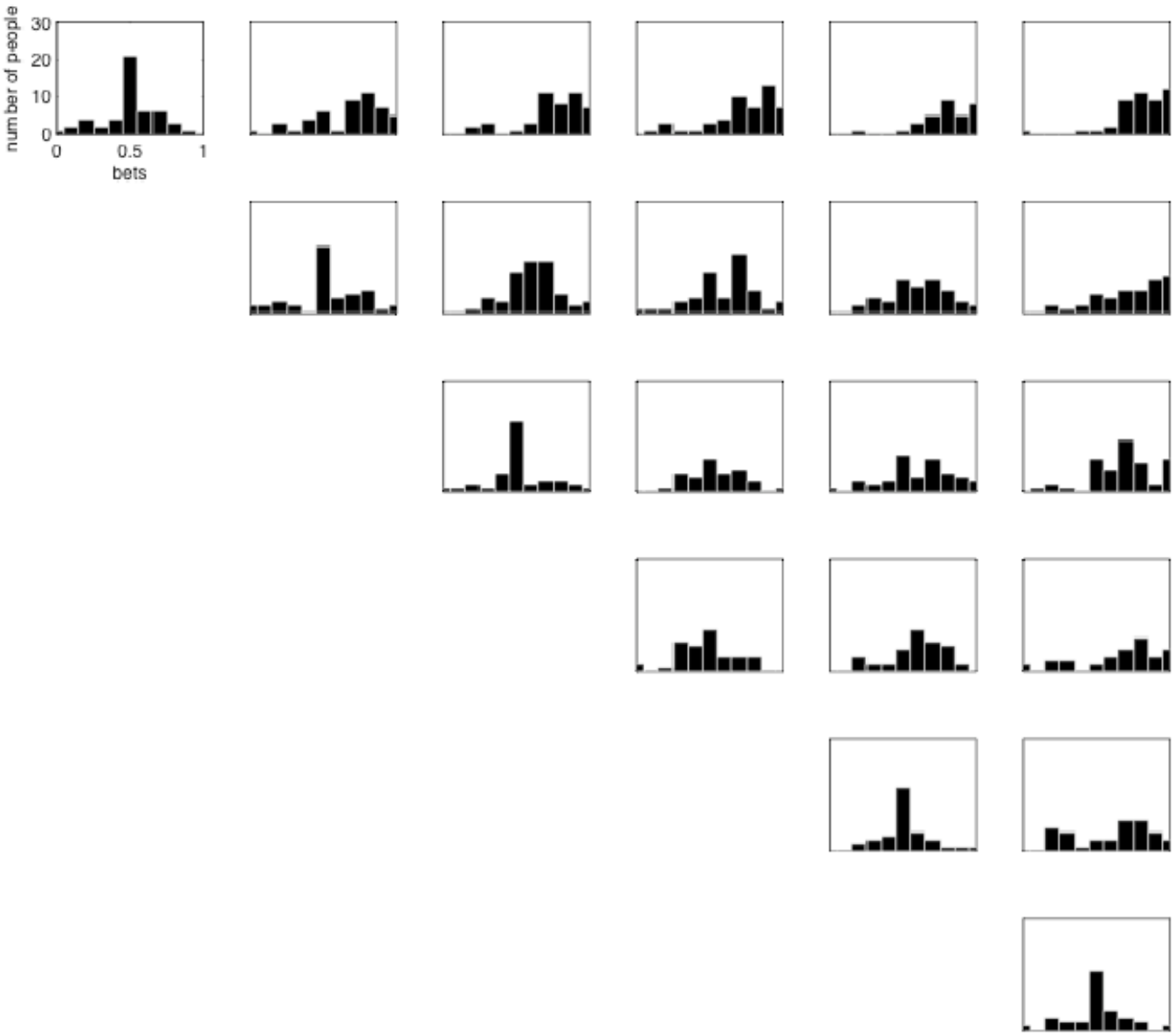


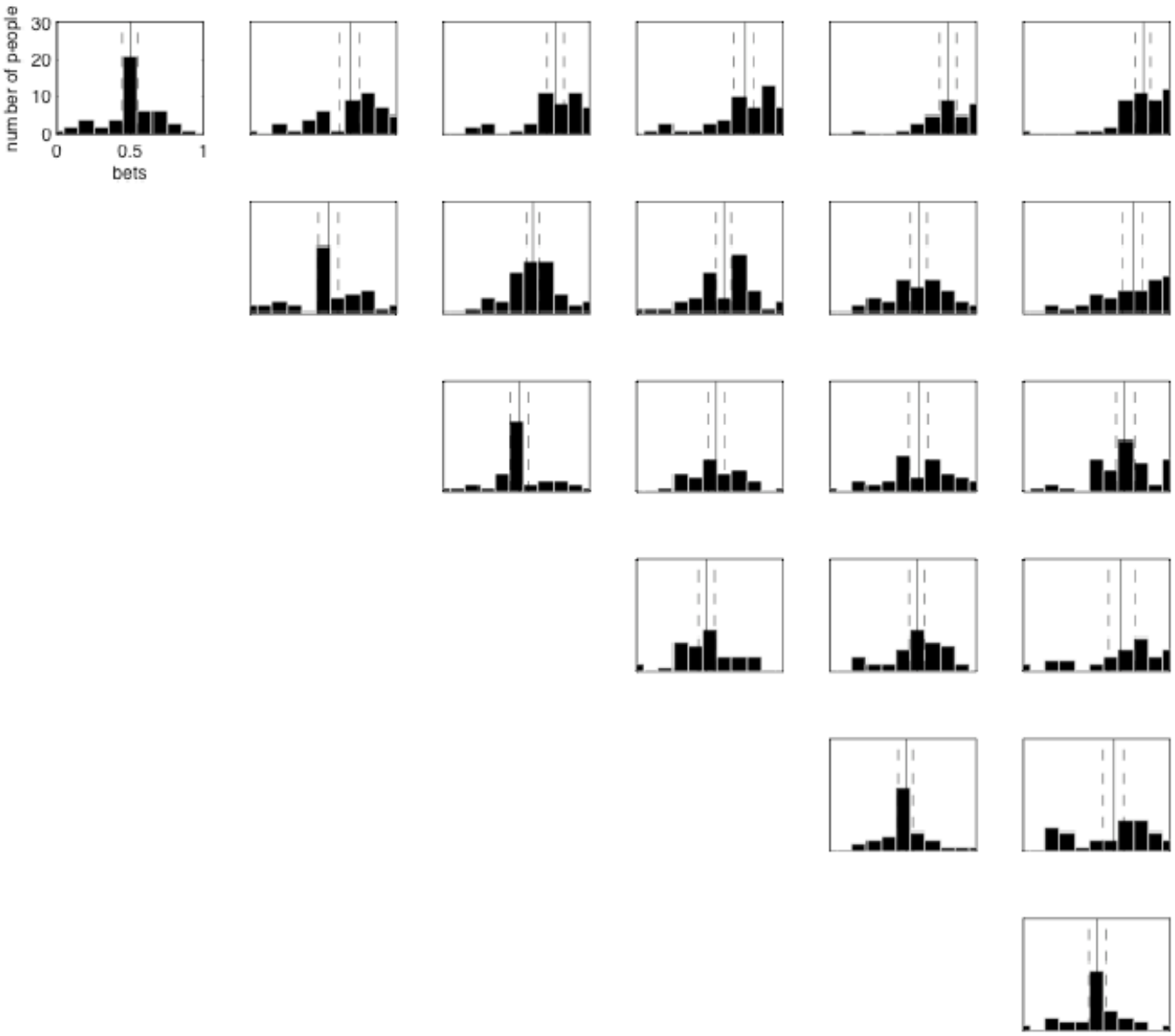
Data-centric vs. viewer-centric

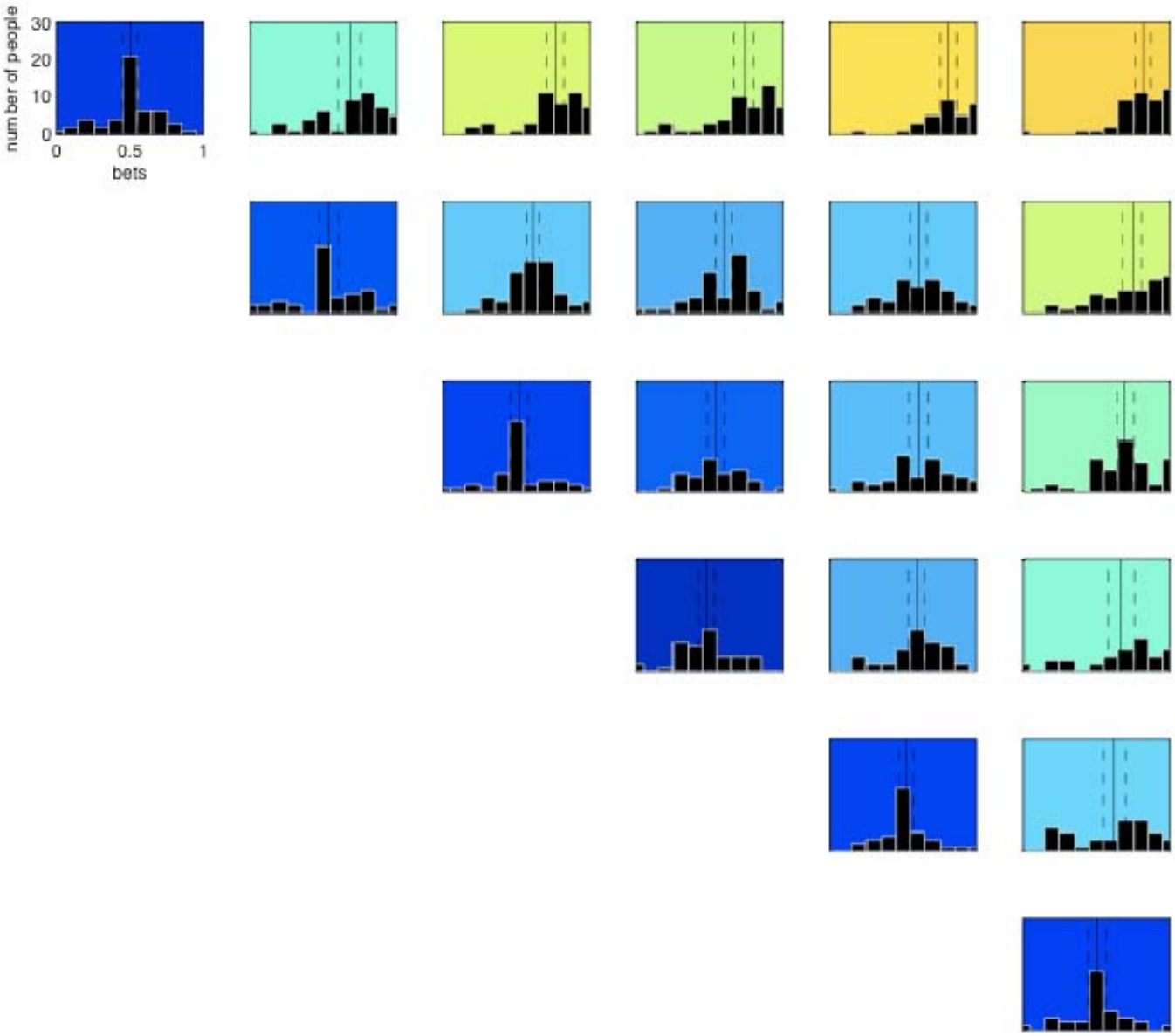


Data-centric vs. viewer-centric




number of people
30
20
10
0
0
0.5
1
bets


number of people
30
20
10
0
0
0.5
1
bets


number of people
30
20
10
0
0
0.5
1
bets

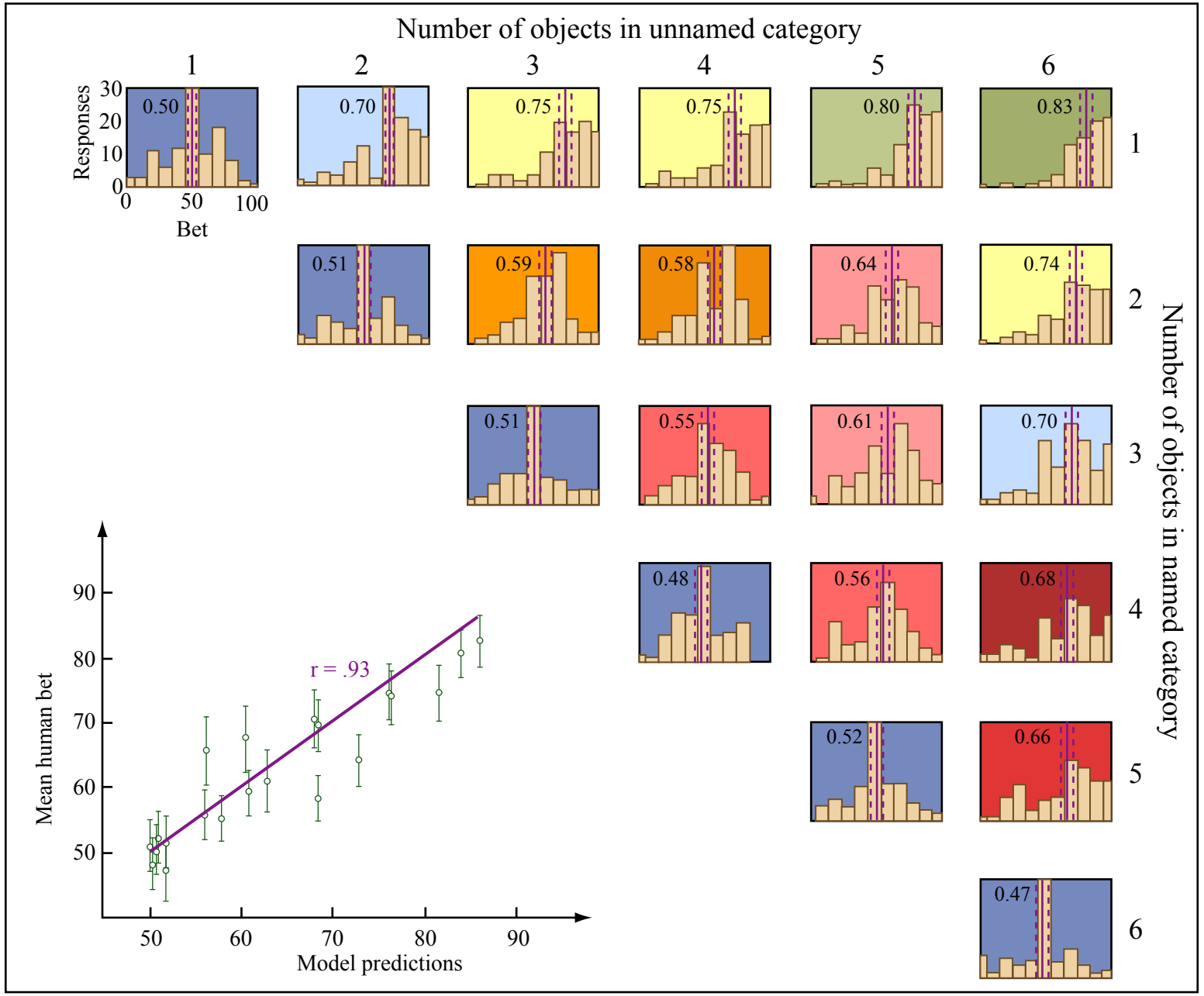
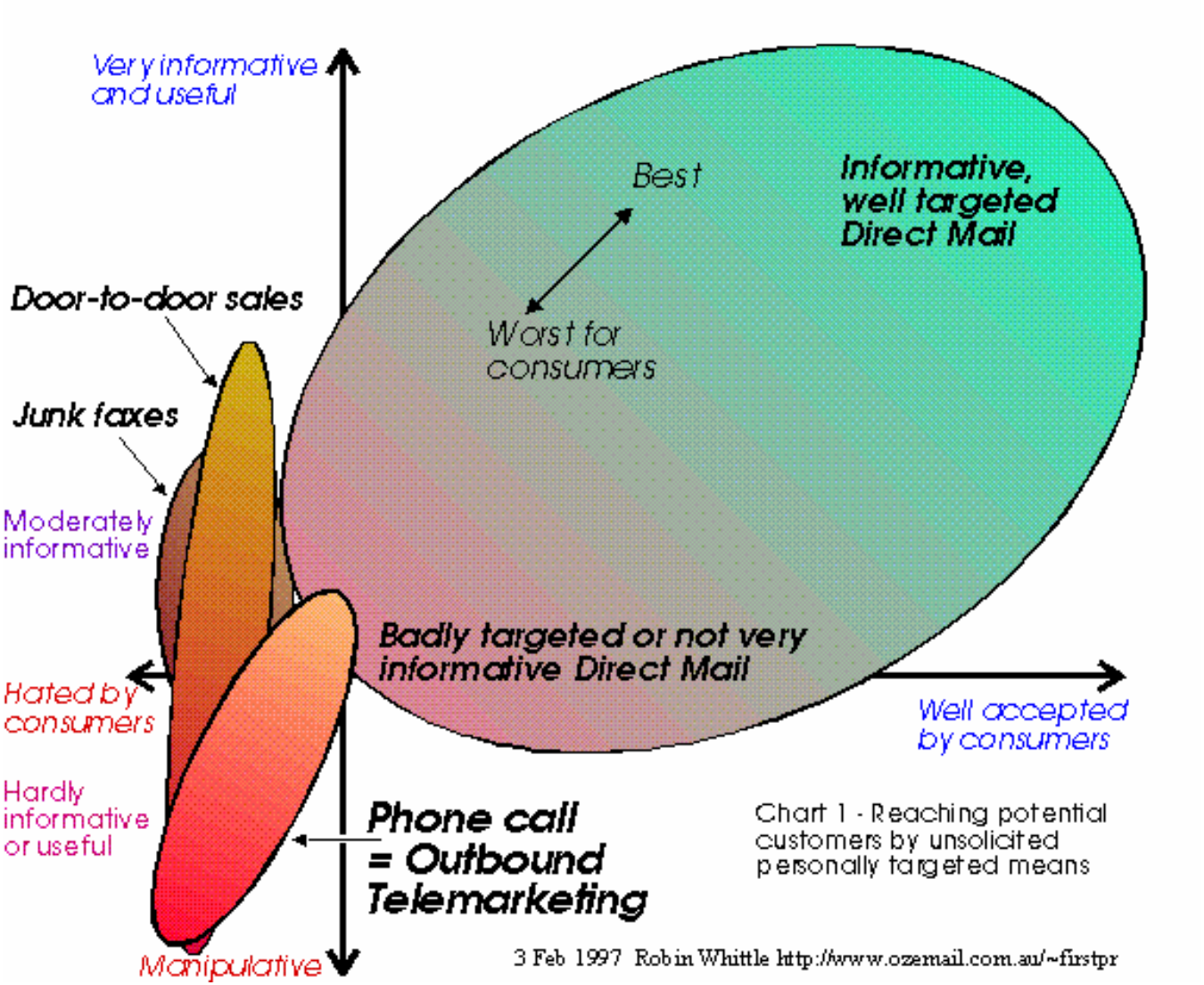


Figure by MIT OpenCourseWare.

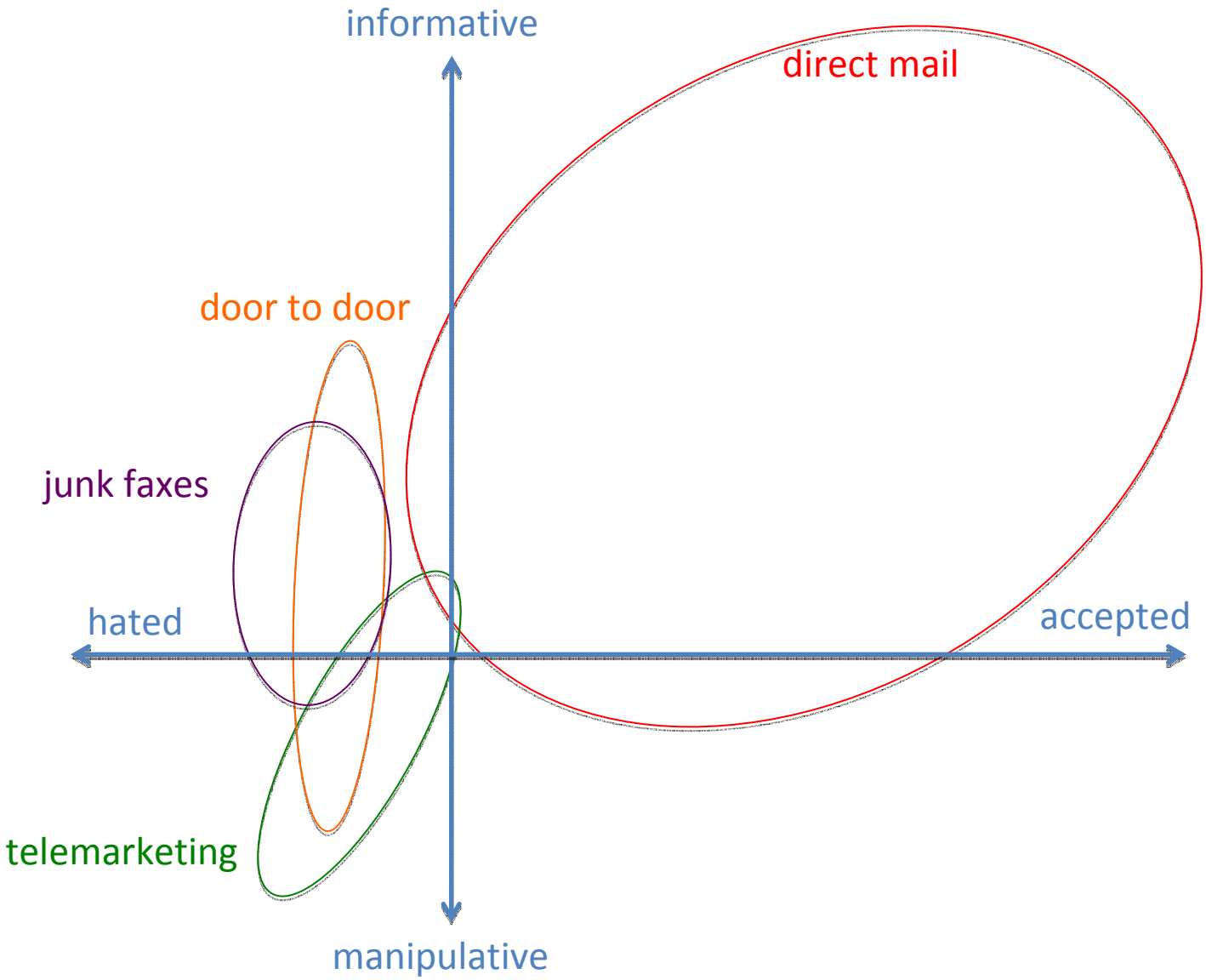
TROUBLE

High Dimensionality Doesn't Guarantee Excellence



Courtesy of Robin Whittle. Used with permission.

High Dimensionality Doesn't Guarantee Excellence



Messy bar graphs

- Sometimes you can discretize way too many variables

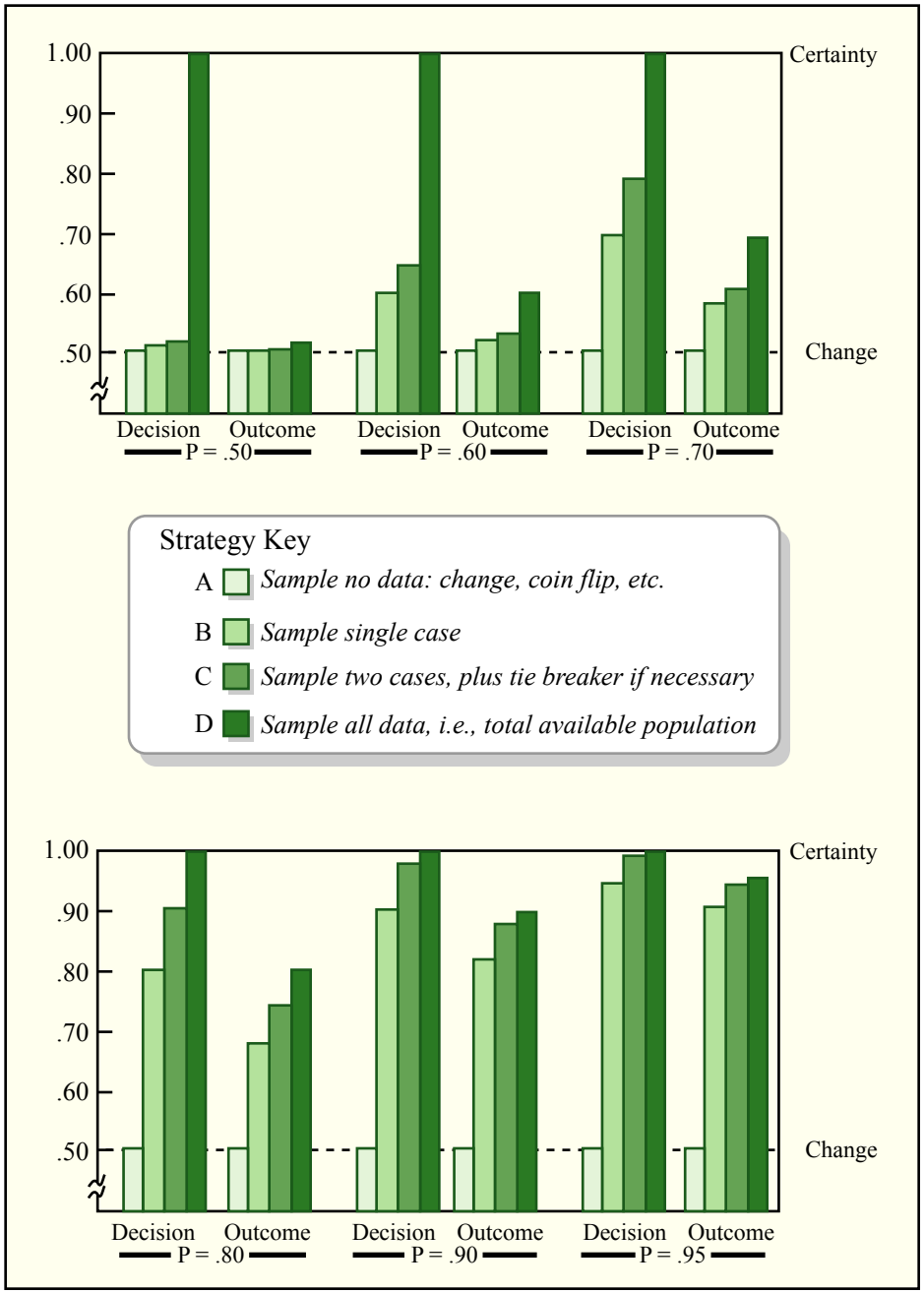
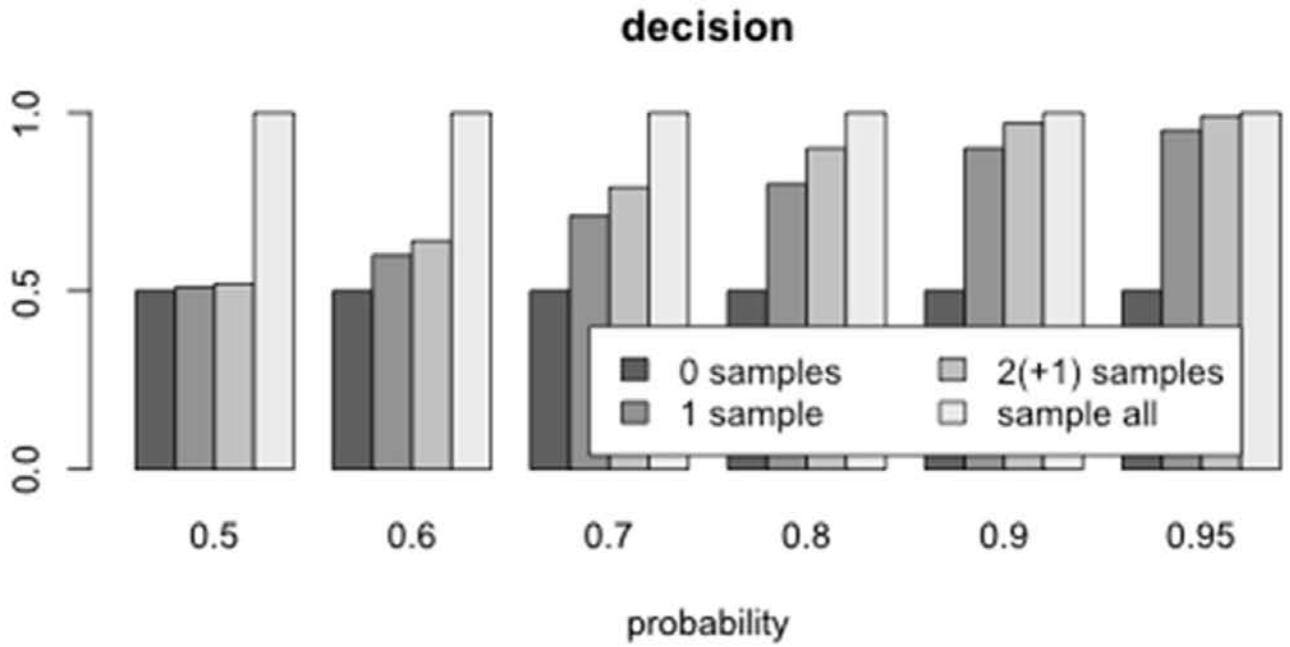
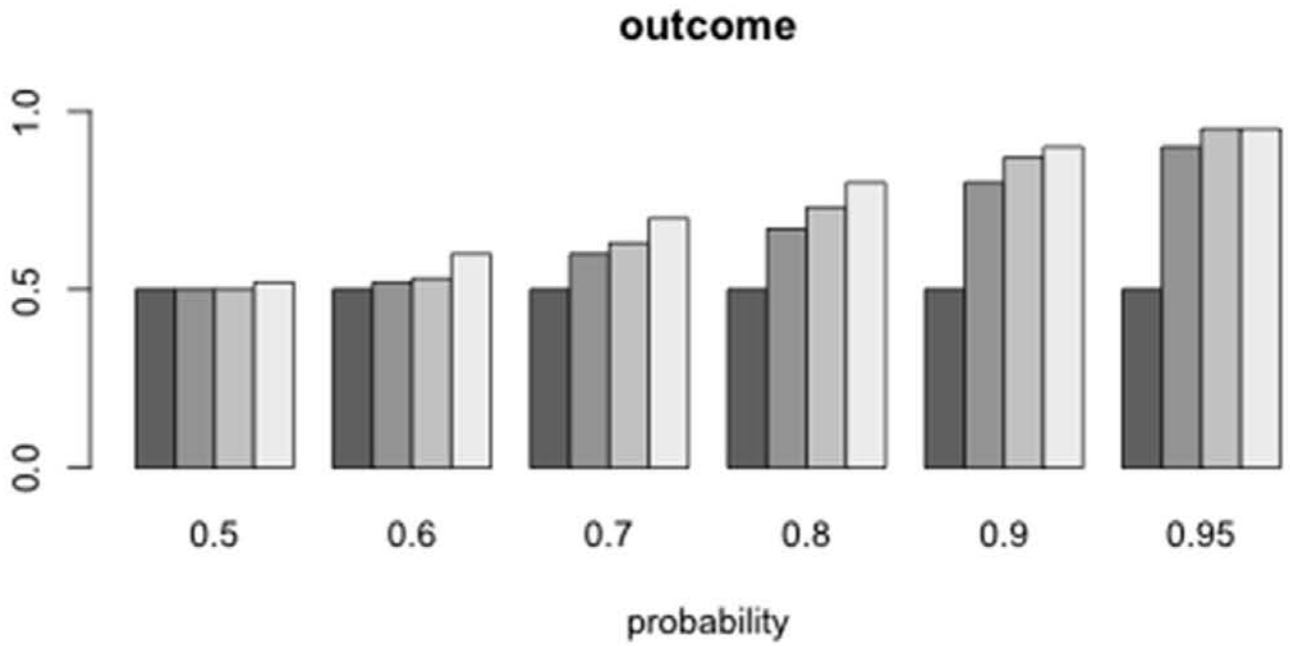
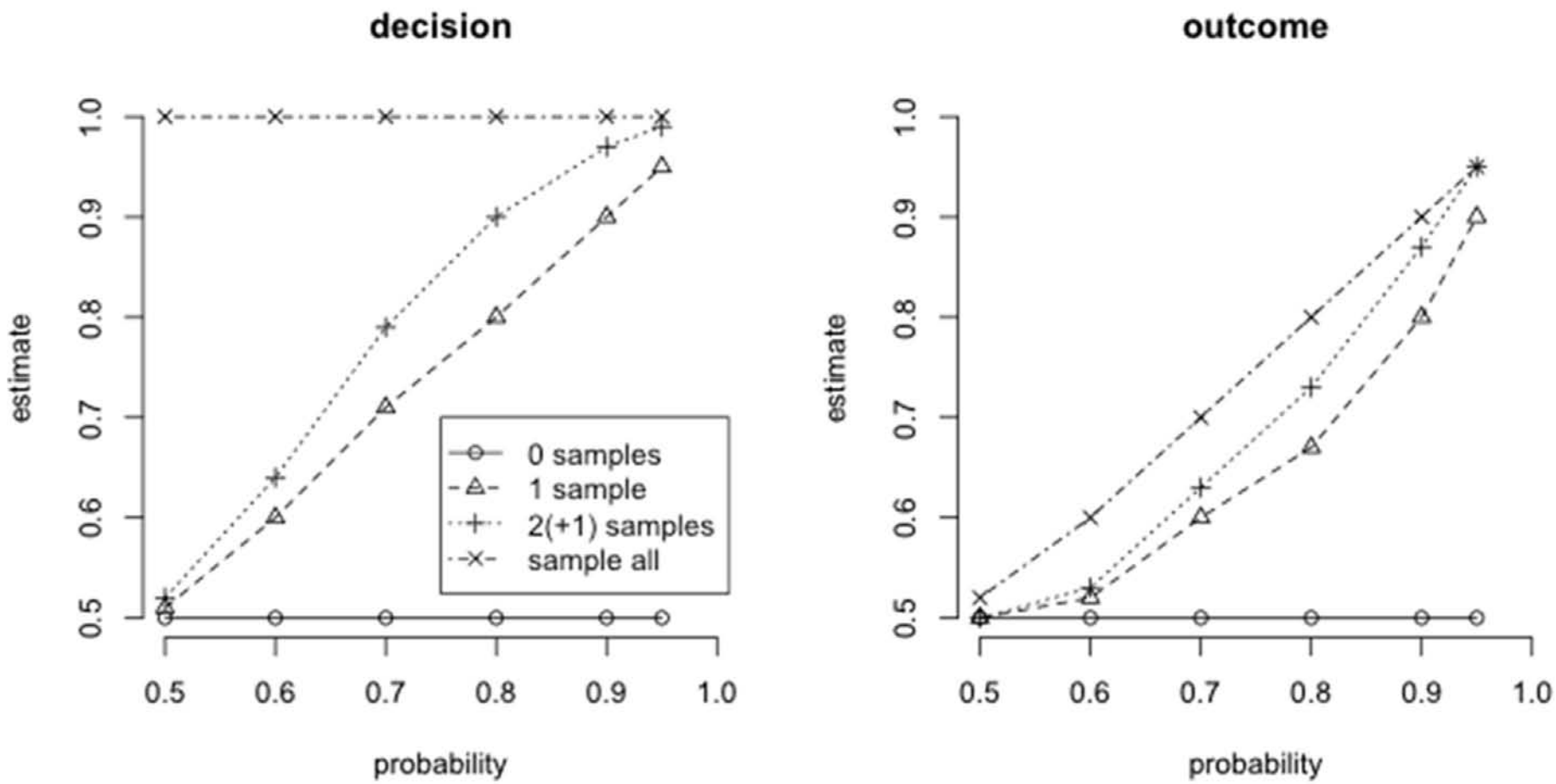


Figure by MIT OpenCourseWare.

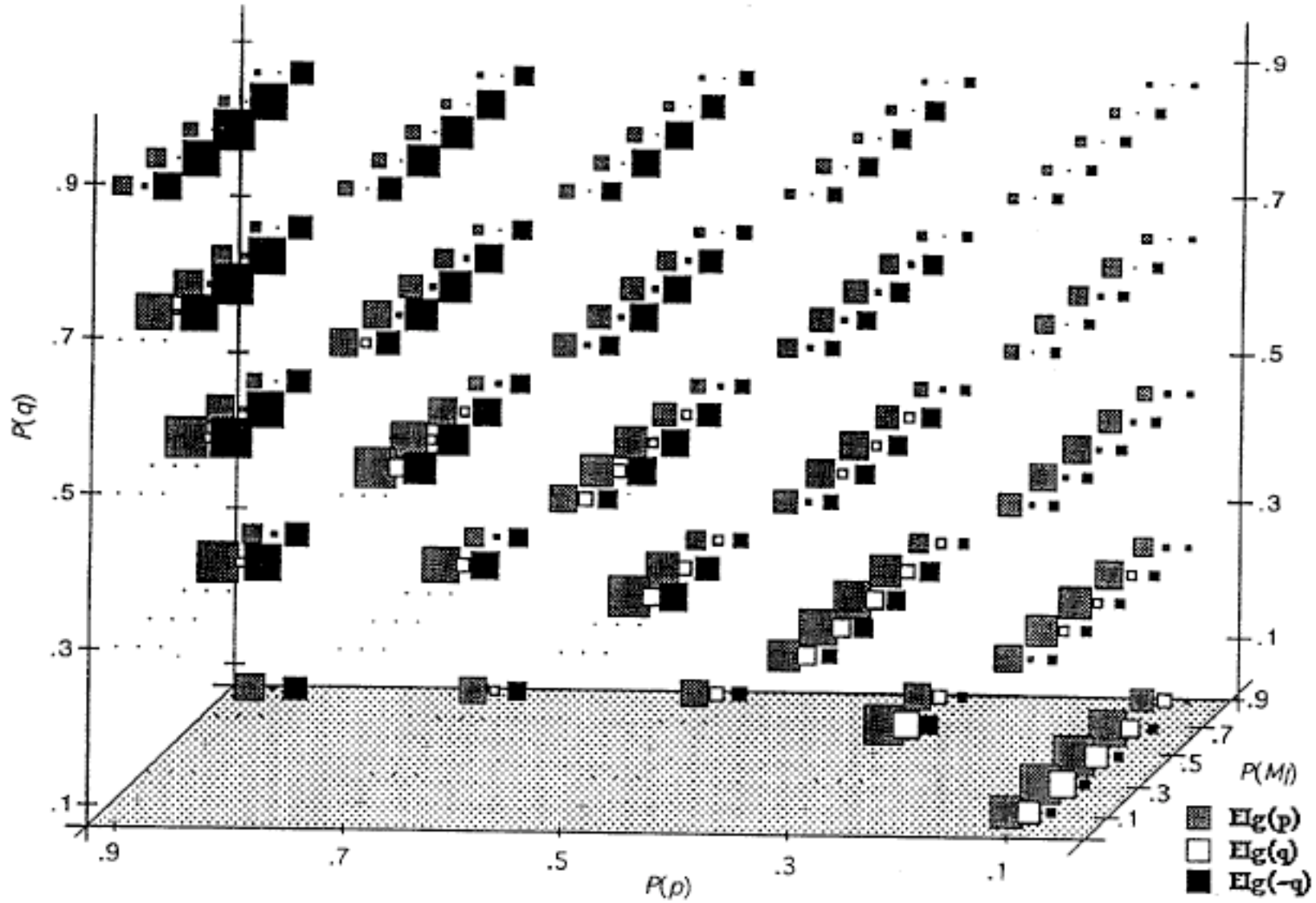
Nisbett& Ross (1980) - ??


decision
1.0
0.5
0.0
0 samples
1 sample
2(+1) samples
sample all
0.5
0.6
0.7
0.8
0.9
0.95
probability


outcome
1.0
0.5
0.0
0.5
0.6
0.7
0.8
0.9
0.95
probability

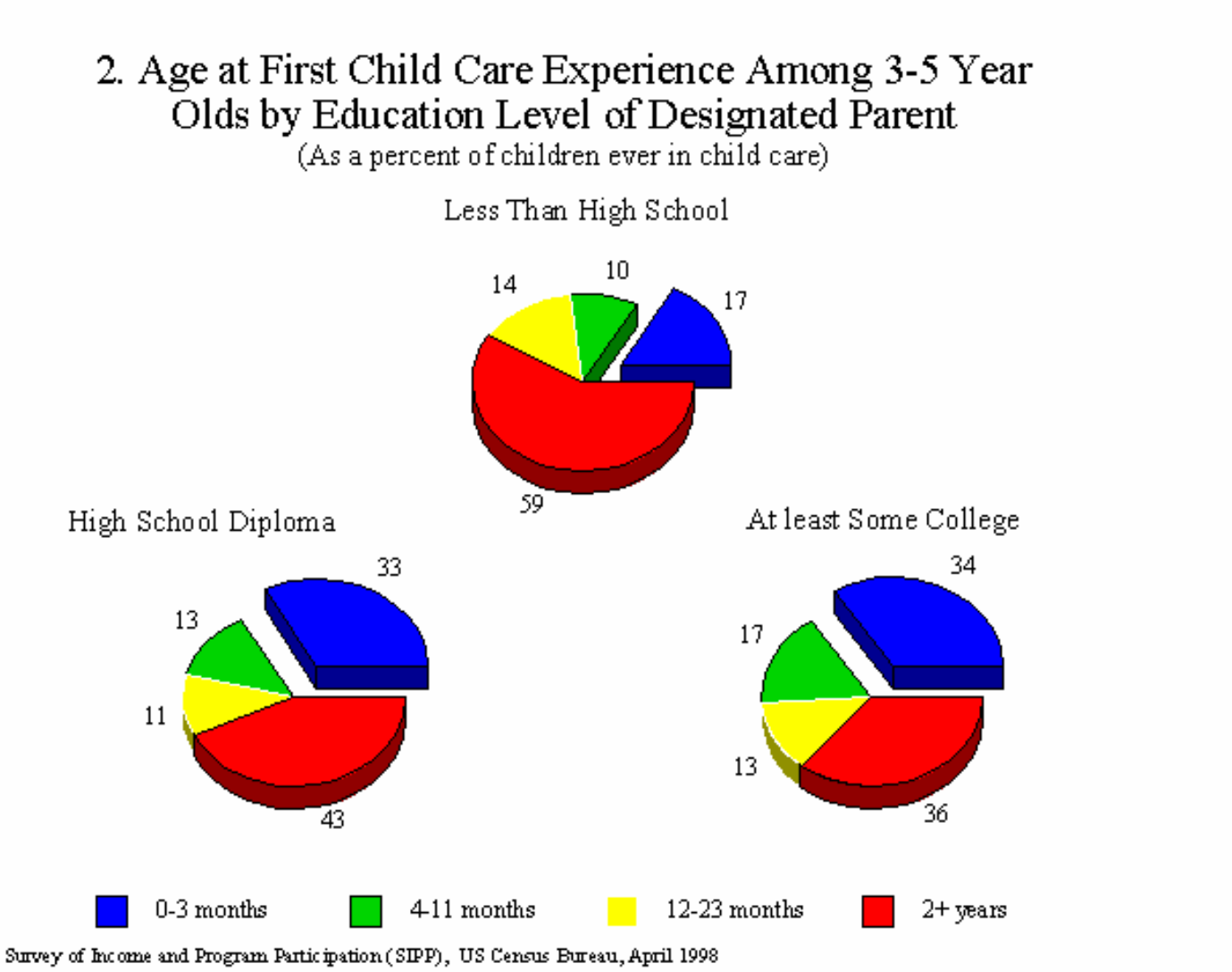

decision
estimate
probability
0.5
0.6
0.7
0.8
0.9
1.0
0 samples
1 sample
2(+1) samples
sample all
outcome
estimate
probability
0.5
0.6
0.7
0.8
0.9
1.0

Too much data for one plot



Oaksford & Chater, 1994

Difficulty of comparison



Bad semantics

Image removed due to copyright restriction.
(http://junkcharts.typepad.com/junk_charts/2008/02/ordering-and-gr.html)

Summary and conclusions

1. **Mapping data to a visual representation**
    1. What dimensions matter
    2. What vocabulary elements (color, shape, orientation) will map to those dimensions
2. **Three principles**
    1. be true to your research
    2. maximize information, minimize ink
    3. organize hierarchically

More worked examples

- Short email describing experiment
  - what is being measured, what is being manipulated
- .CSV (comma-separated value) file
  - header row with good variable names
    - GOOD: sub.name,trial.type,correct.ans
    - BAD: num,tt,g
  - each row is a single observation
    - e.g., one or two *ys* (dependent variable like answer correctness or RT), many *x*s (independent variables like subject, condition, etc.)