UNSW SYDNEY | Australia's Global University

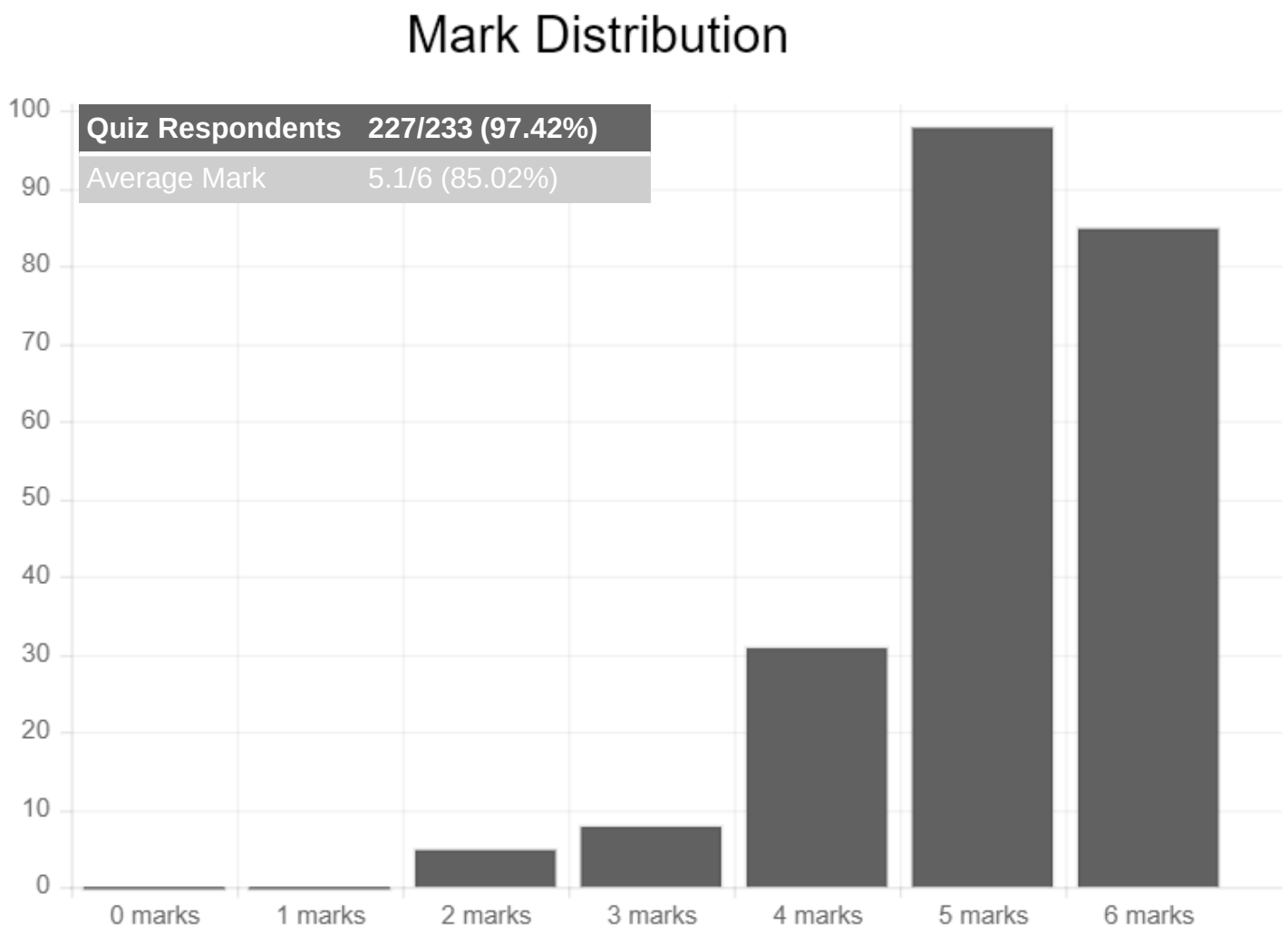
COMP9321

Data Services Engineering

Term 1, 2019

Week 3 Lecture 1

Statistics for Quiz 1 (marks released)



1. Which of the following is NOT correct on data extraction?

- Many HTML documents are auto-generated these days, making it relatively easier to precisely target relevant patterns in a document
- XML and JSON are already structured and can convey hierarchical structure of the data well
- CSV data is always normalised, making it an ideal candidate for database import tasks
- Processing text from a PDF document requires understanding its layout as well as the content

2. Select the NOT correct statements about Impedance Mismatch?

- It is a translation layer between the objects in the application code and the database model of tables/row/columns
- It is one of the problems mostly related to relational databases
- It is also called "Object-Relational Mismatch”
- Refers to the problem of a mismatch between application data model

3. Which statement is NOT true about NoSQL databases

- NoSQL databases can have no schema
- NoSQL databases are suitable for embedded document-like data model
- MongoDB is an example of NoSQL database
- Performance-wise, NoSQL databases can only scale vertically

4. Which of the following can be considered as metadata?

- The description of an object and its type
- The time an object is created or modified
- Information about the origin of the object
- All of the above mentioned options

5. Generate the same table

```
class Contact(db.Model):
    __tablename__ = 'contacts'
    id = db.Column(db.Integer, primary_key=True)
    first_name = db.Column(db.String(100))
    last_name = db.Column(db.String(100))
    phone_number = db.Column(db.String(32))
    address = db.Column(db.String(100))
    post_code = db.Column(db.Integer)

    def __repr__(self):
        return '<Contact {0} {1}: {2}>'.format(self.first_name,
                                               self.last_name,
                                               self.phone_number,
                                               self.address,
                                               self.post_code)
```

```
CREATE TABLE CONTACTS(
    ID INT PRIMARY KEY          NOT NULL,
    FIRST_NAME      CHAR(100)   NOT NULL,
    LAST_NAME       CHAR(100)   NOT NULL,
    PHONE_NUMBER    CHAR(32)    NOT NULL,
    ADRESS          CHAR(100)   NOT NULL,
    POST_CODE       INT,
);
```

```
CREATE TABLE CONTACTS(
    ID INT PRIMARY KEY          NOT NULL,
    FIRST_NAME      CHAR(100),
    LAST_NAME       CHAR(100),
    PHONE_NUMBER    CHAR(32),
    ADRESS          CHAR(100),
    POST_CODE       INT NOT NULL,
);
```

```
CREATE TABLE CONTACTS(
    ID INT PRIMARY KEY          NOT NULL,
    FIRST_NAME      CHAR(100),
    LAST_NAME       CHAR(100),
    PHONE_NUMBER    CHAR(32),
    ADRESS          CHAR(100),
    POST_CODE       INT,
);
```

```
CREATE TABLE CONTACTS(
    ID INT PRIMARY KEY          NOT NULL,
    FIRST_NAME      CHAR(100),
    LAST_NAME       CHAR(100),
    PHONE_NUMBER    CHAR(32),
    ADRESS          CHAR(100),
    POST_CODE       INT(32),
);
```

```
CREATE TABLE CONTACTS(
    ID INT PRIMARY KEY,
    FIRST_NAME      CHAR(100),
    LAST_NAME       CHAR(100),
    PHONE_NUMBER    CHAR(32),
    ADRESS          CHAR(100),
    POST_CODE       INT,
);
```

What's the relational pattern for following code:

```
from sqlalchemy import Table, Column, Integer, ForeignKey
from sqlalchemy.orm import relationship
from sqlalchemy.ext.declarative import declarative_base

Base = declarative_base()

association_table = Table('association', Base.metadata,
    Column('left_id', Integer, ForeignKey('left.id')),
    Column('right_id', Integer, ForeignKey('right.id'))
)

class Parent(Base):
    __tablename__ = 'left'
    id = Column(Integer, primary_key=True)
    children = relationship("Child",
                            secondary=association_table,
                            backref="parents")

class Child(Base):
    __tablename__ = 'right'
    id = Column(Integer, primary_key=True)
```

- Many-To-Many
- One-To-One
- One-To-Many
- Many-To-One

Assignment 1

Due 23:59:59, 9TH March

This Saturday

128/233 submissions so far

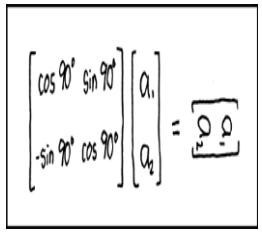
Data Visualization(1)

COMP9321 2019T1

Data Scientist's Workflow

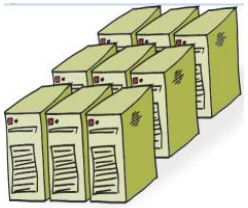
Sandbox

Digging Around in Data

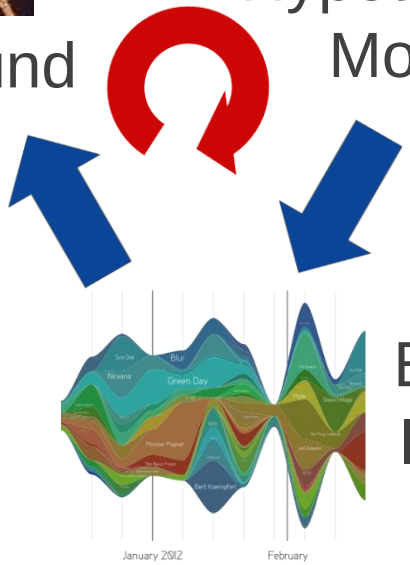


Hypothesize Model

Production



Large Scale Exploitation

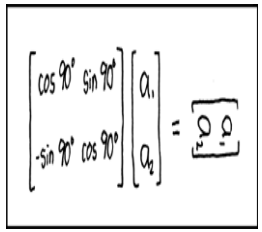


Evaluate Interpret

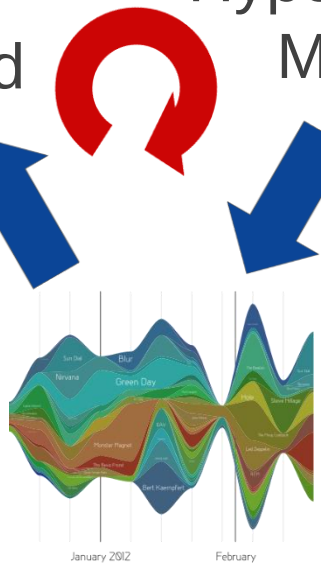
Data Scientist's Workflow

Sandbox

Digging Around in Data

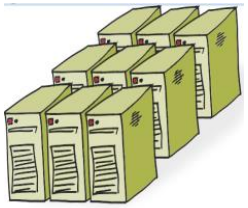


Hypothesize Model



Evaluate Interpret

Publish Information

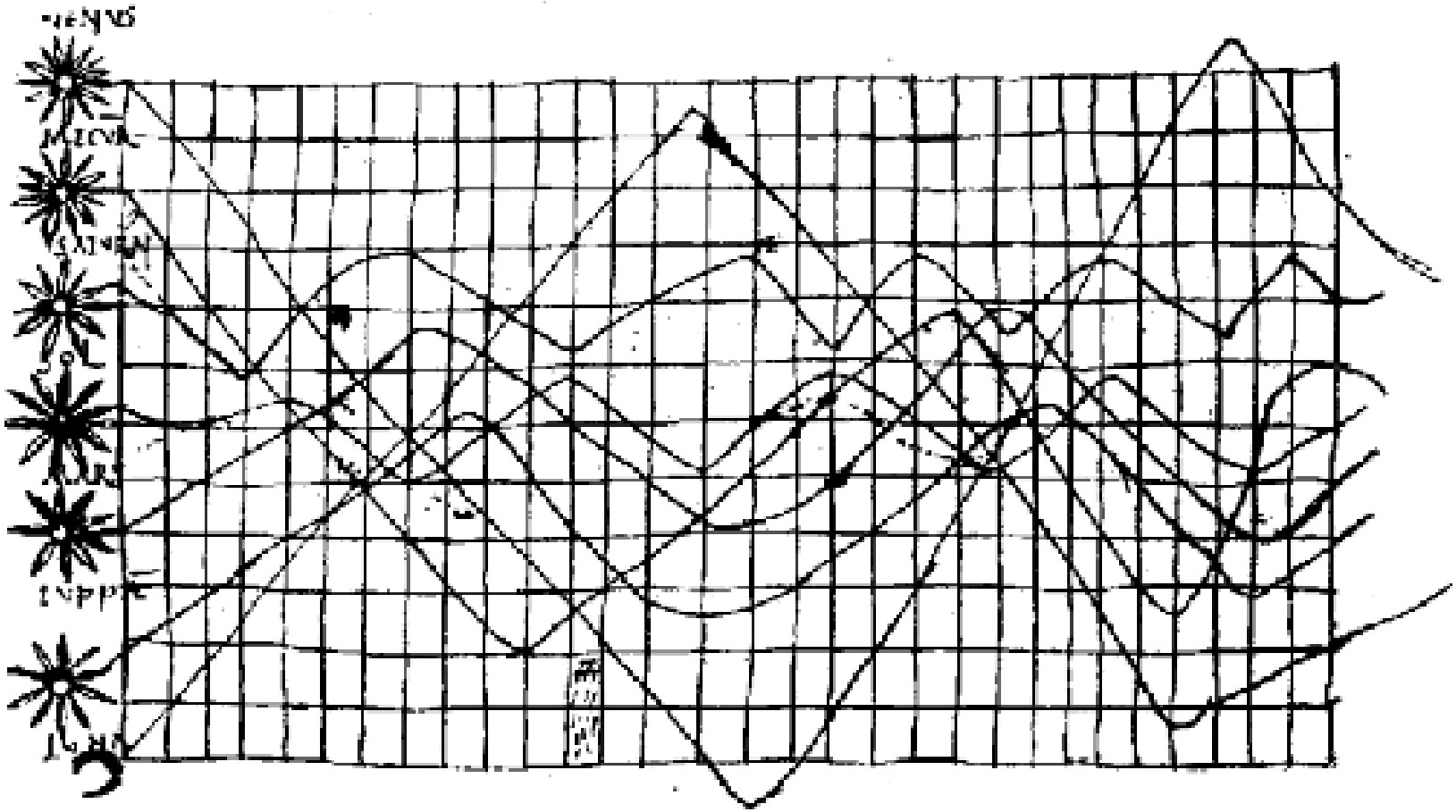


Early Babylonian world map (600 BC)

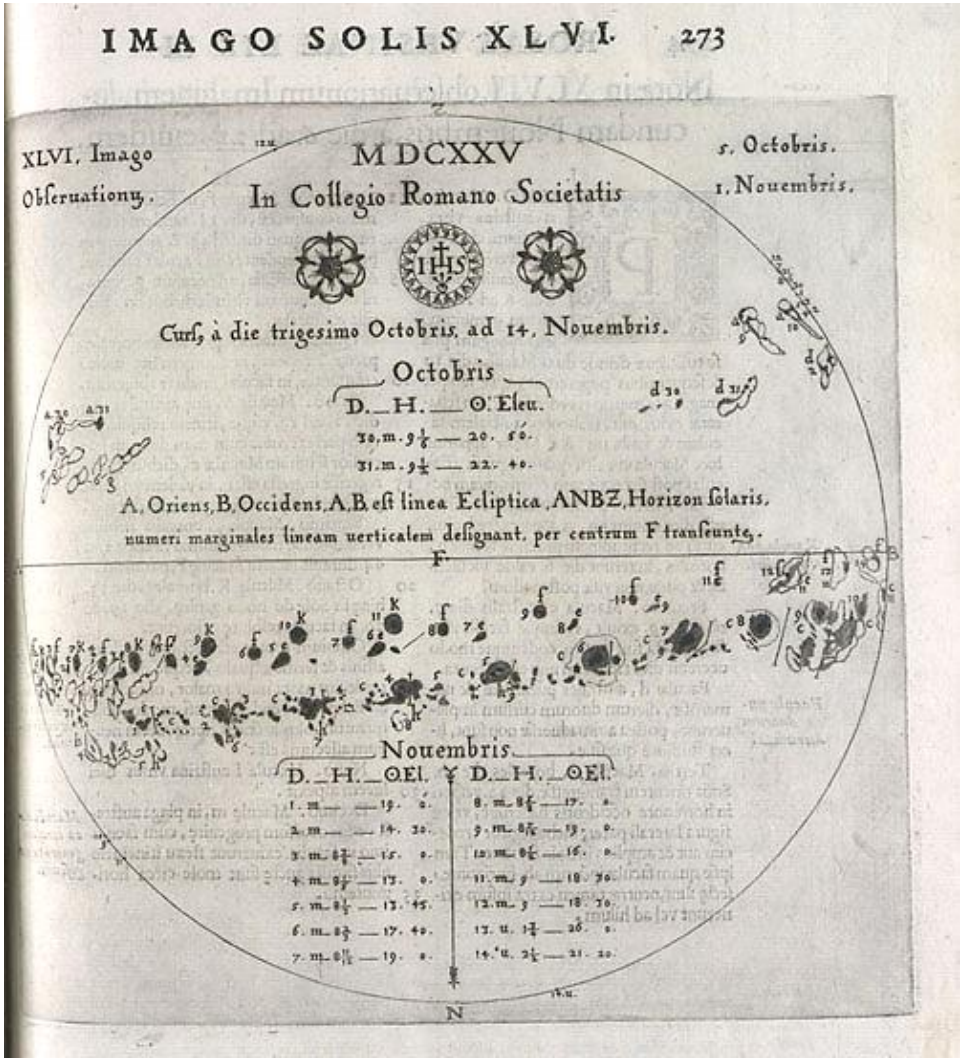
These concentric circles that represent the ocean, named “bitter water” or the “salt sea.”

an early interpretation of the layout of the world

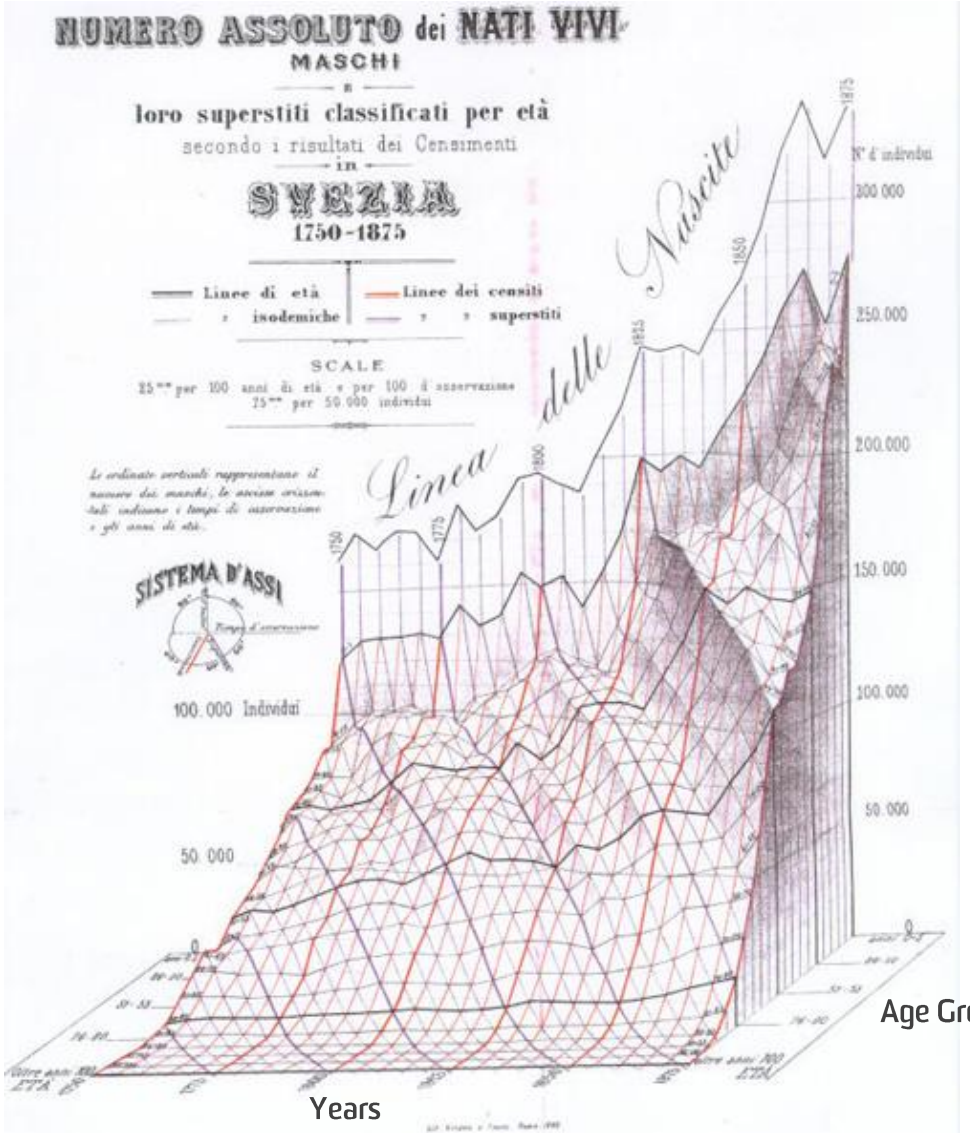
Positions of the Sun, Moon, and Planets Throughout the Year (Europe, 950 AD)



Christoph Scheiner, Images of Sunspots 1626

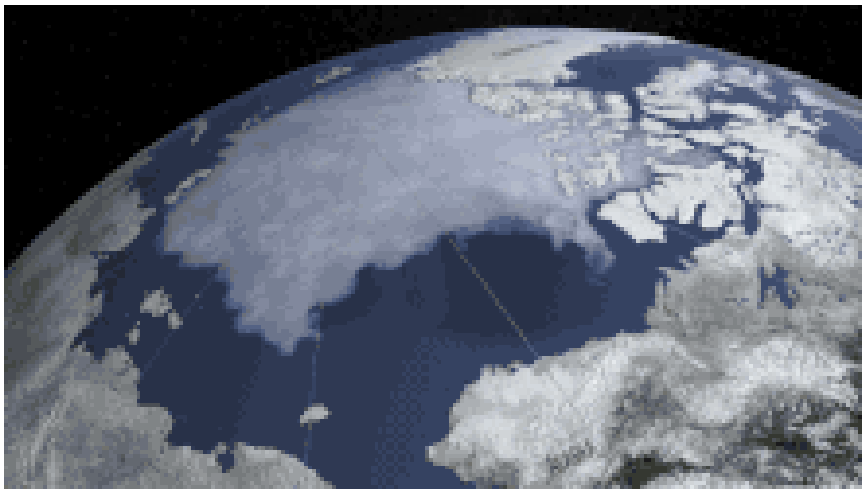


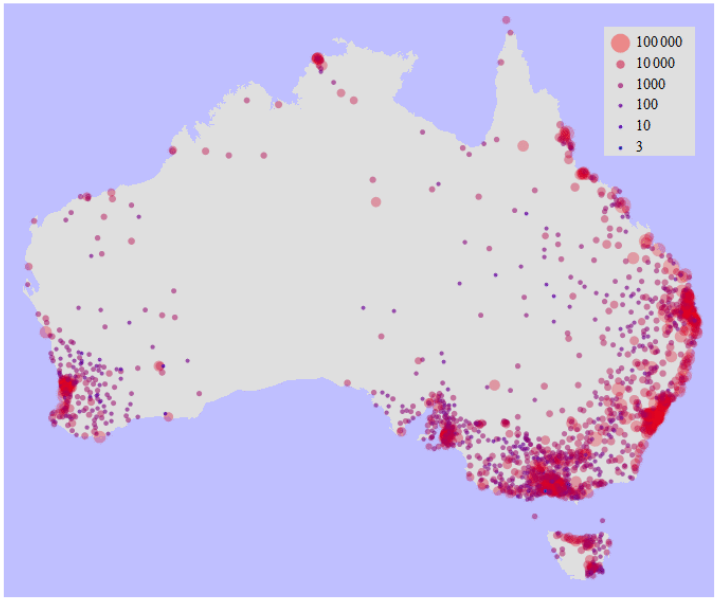
Luigi Perozzo, 3D Model of the Swedish Census 1879

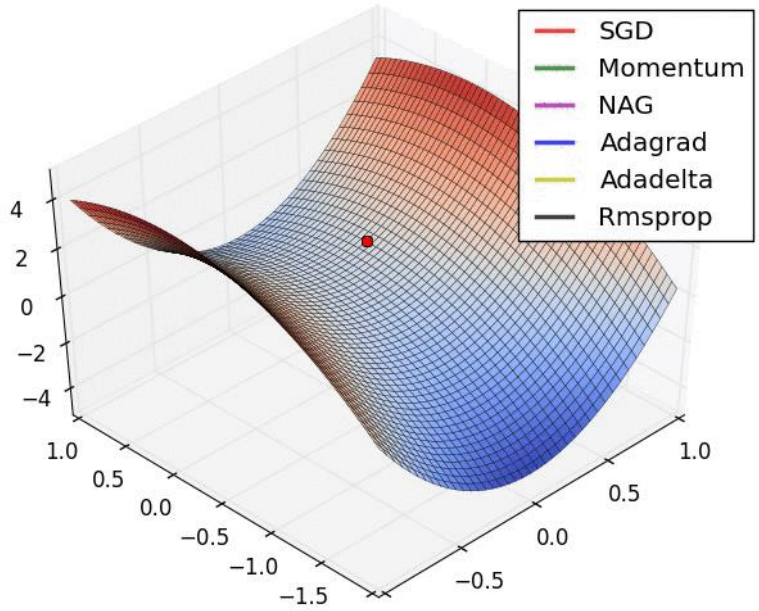


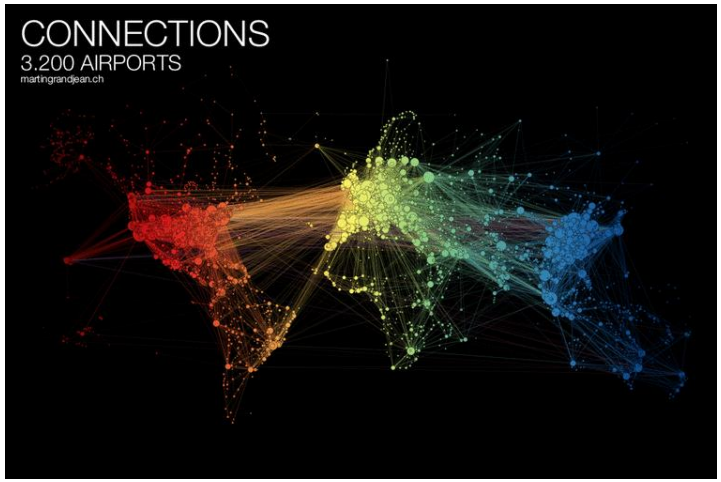
Pretty old, yet famous Gapminder Video, Hans Rosling:
200 Countries, 200 Years in 4 Minutes

Nowadays…









Network Attacks

Scene Reconstruction

Question

What is the purpose of a data visualization?

A. Find relationships in data

B. Find patterns in data

C. Discover meaning in data

D. All of the above

Clarity

DATA VISUALIZATION IS COMMUNICATION

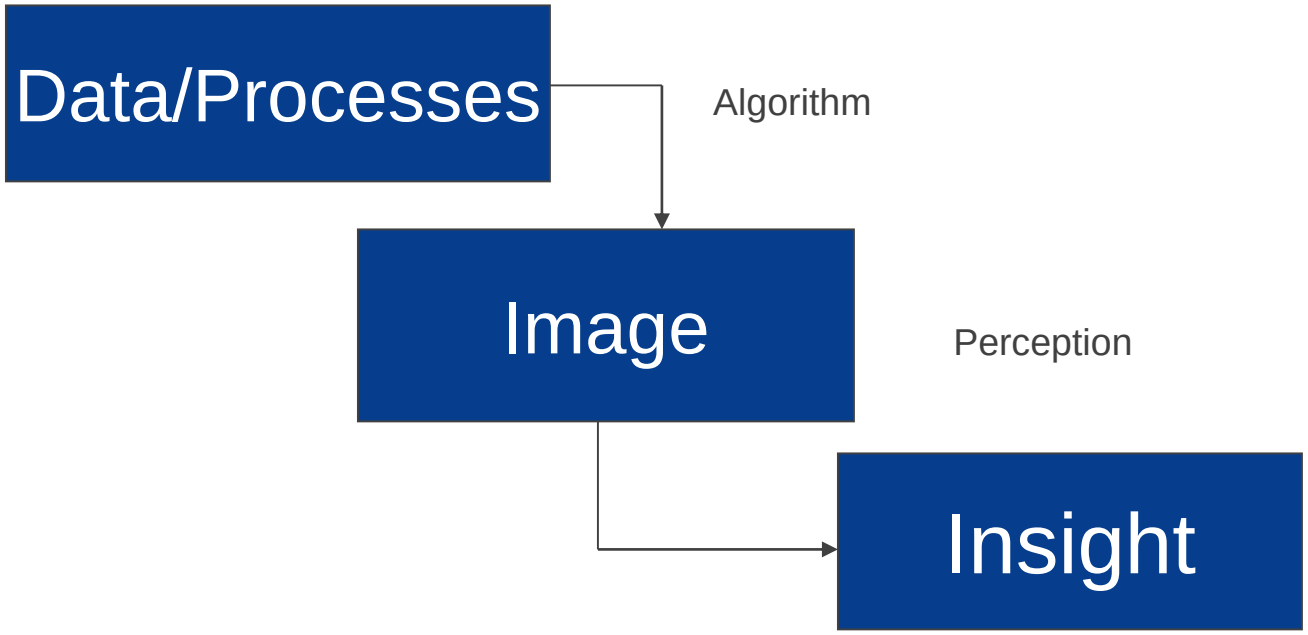
Precision

Efficiency

Maximize ideas, minimize ink

Data -> Easily Understood Pictures

Jacques Bertin who wrote the classic works of graphical visualization "Semiology of Graphics" states that the "transformation from numbers to insight requires two stages."



Bertin's 7 Visual Variables

Seven Visual Variables

- position
- form
- orientation
- color
- texture
- value
- size

combined with a visual semantics for linking data attributes to visual elements

Image Theory

Visual Processing occurs in 3 steps.

- 1) formation of the retinal image,
- 2) decomposition of the retinal image information into an array of specialized representations and
- 3) reassembly of the information into object perception.

Image

Bertin's key concept is the *image*, from which the theory derives its name.

Roughly speaking, an *image* is the fundamental perceptual unit of a visualization.

- An ideal visualizations will contain only a single *image* in order to optimize "efficiency," the speed with which observer can extract the information

Uses Today

Data-driven actions are increasingly made without access to information provided by traditional information presentation

Information visualization is emerging as an important fusion of graphics, scientific visualization, database, and human-computer interaction.

- In Military, Commercial Industries use Data Visualization to convey complex results as understandable images.

What is Data Visualization

Data visualization is the process of converting raw data into easily understood pictures of information that enable fast and effective decisions.

The study of visual representations of data to reinforce human cognition.
"Help people understand the, structure, relationships meaning in data."

Techniques: Charts, Graphs, Maps

What is Data Visualization

Data visualization is used in software applications to provide an intuitive graphical interface.

It is applied to many areas to enable users to glean useful information from their data for faster, more informed decision making.

These areas include: Military, private business sectors and scientific research.

What are the benefits of Data Visualization?

Data visualization allows users see several different perspectives of the data.

Data visualization makes it possible to interpret vast amounts of data

Data visualization offers the ability to note exceptions in the data.

Data visualization allows the user to analyze visual patterns in the data.

Exploring trends within a database through visualization by letting analysts navigate through data and visually orient themselves to the patterns in the data.

Benefits

Data visualization can help translate data patterns into insights, making it a highly effective decision-making tool.

Data visualization equips users with the ability to see influences that would otherwise be difficult to find.

With all the data available, it is difficult to find the nuances that can make a difference.

By simplifying the presentation, Data Visualization can reduce the time and difficulty it takes to move from data to decision making.

Anatomy of A Visualization

1. Title
2. X-Axis
3. Y-Axis
4. Series
5. Data Points

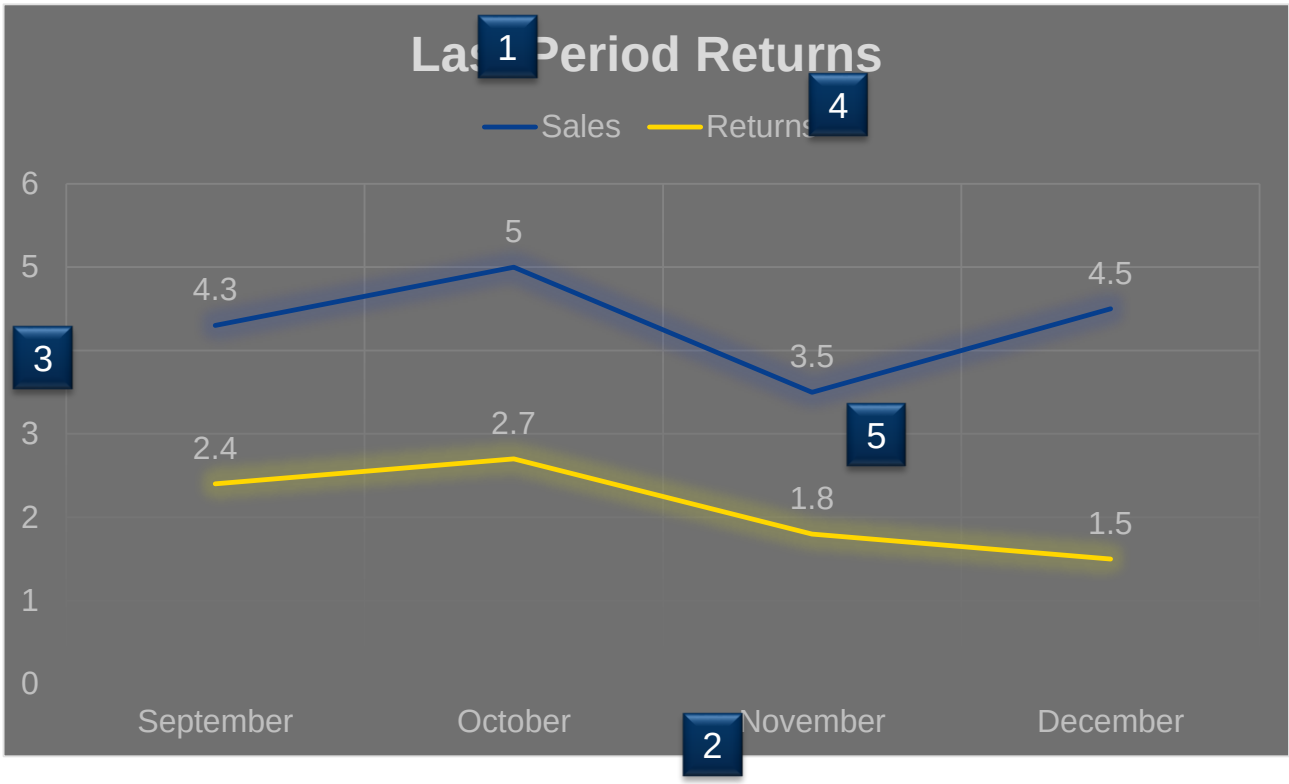


Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

Among Items

- Two Variables per Item
  - Variable Width Column Chart
- One Variable per Item
  - Many Categories
    - Table or Table with Embedded Charts
  - Few Categories
    - Many Items
      - Bar Chart
    - Few Items
      - Column Chart

Over Time

- Many Periods
  - Cyclical Data
    - Circular Area Chart
  - Non-Cyclical Data
    - Line Chart
- Few Periods
  - Single or Few Categories
    - Column Chart
  - Many Categories
    - Line Chart

Relationship

- Two Variables
  - Scatter Chart
- Three Variables
  - Bubble Chart

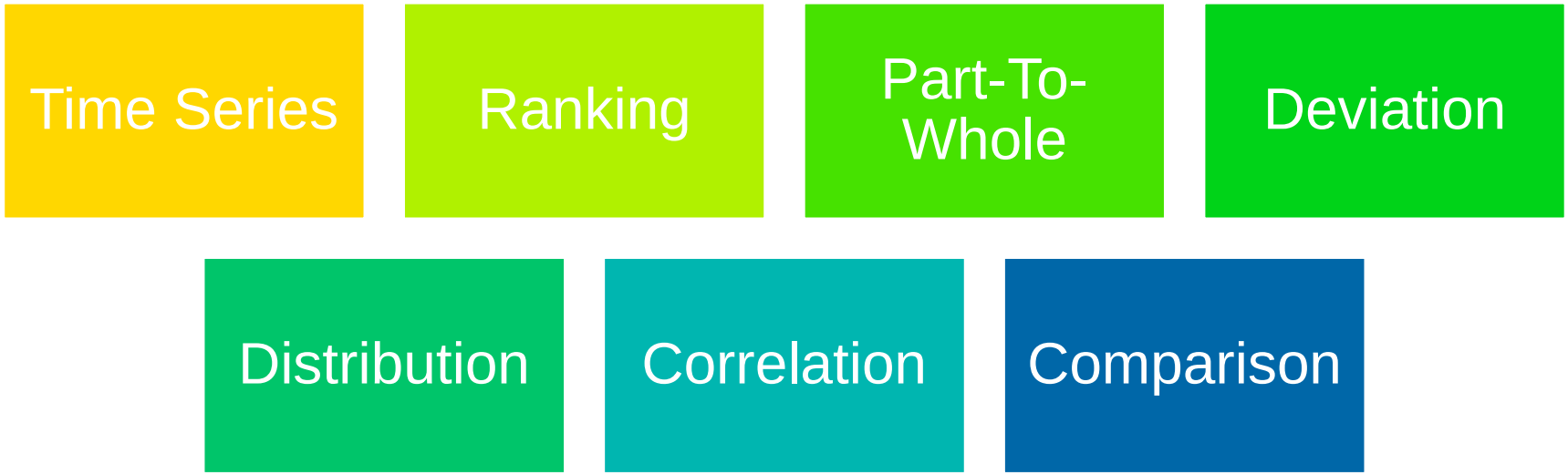
Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

What do you want to show with your data?



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

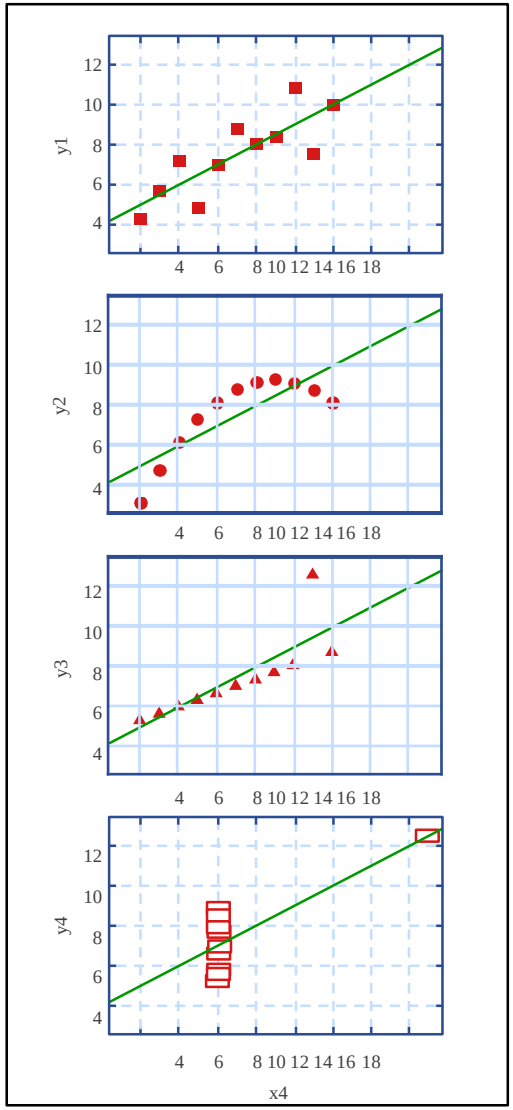
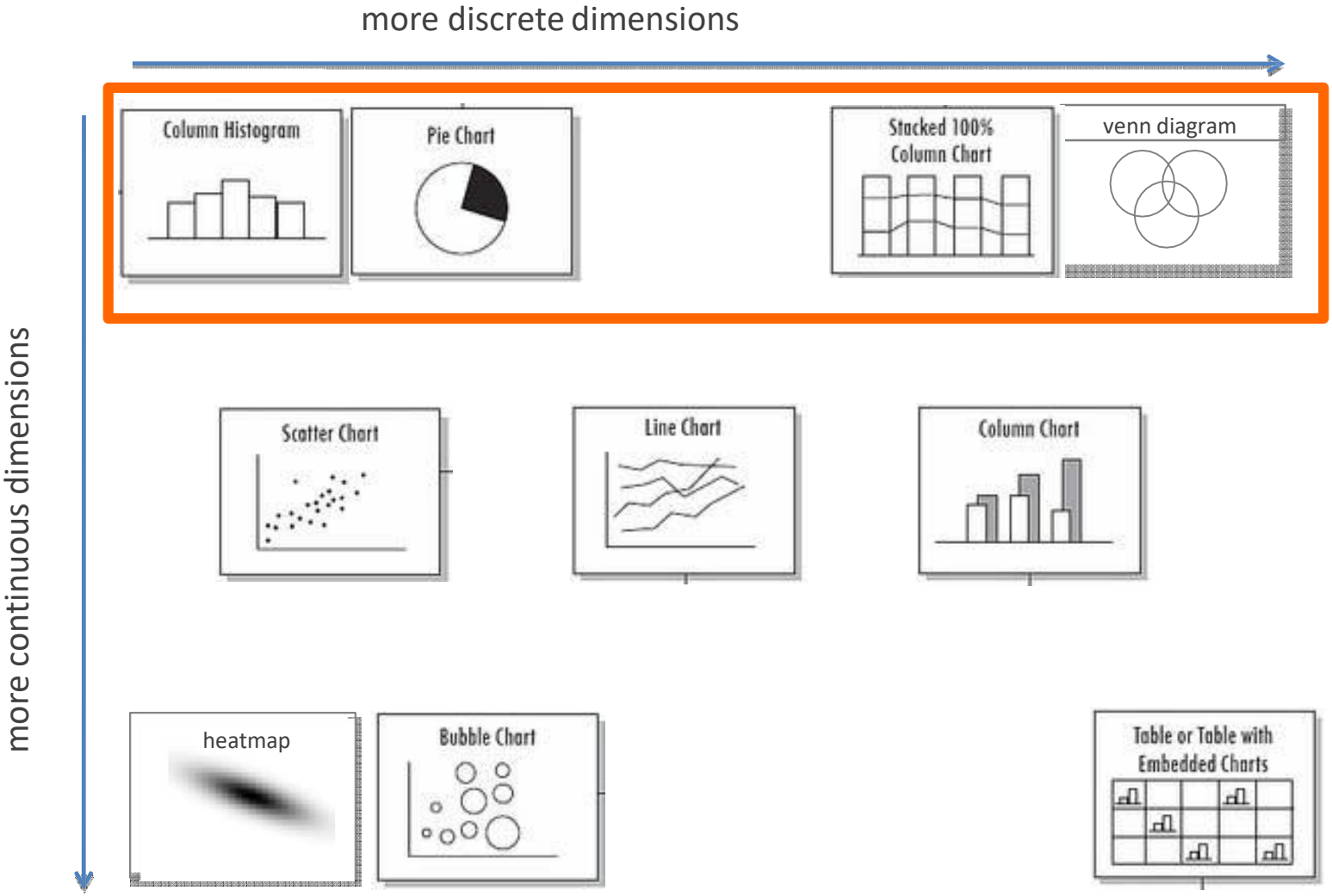


Figure by MIT OpenCourseWare.

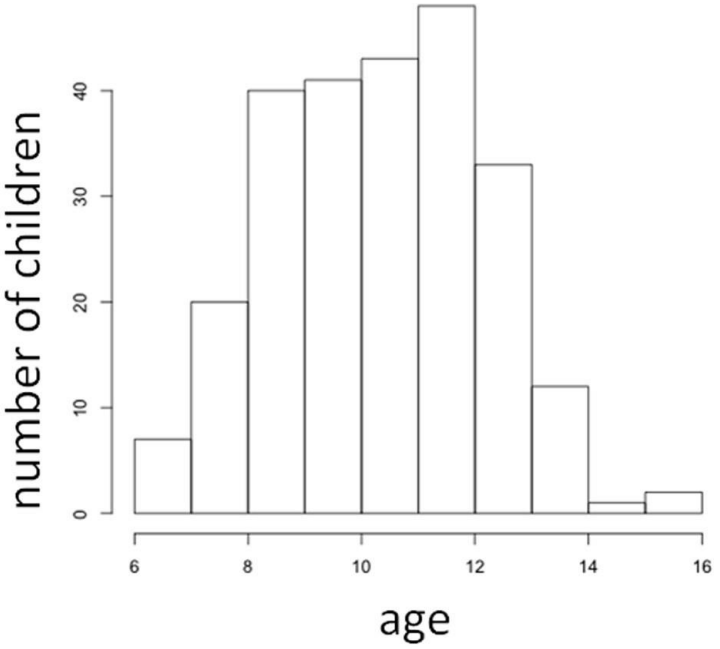
F. J. Anscombe, 1973

Conventional visualizations

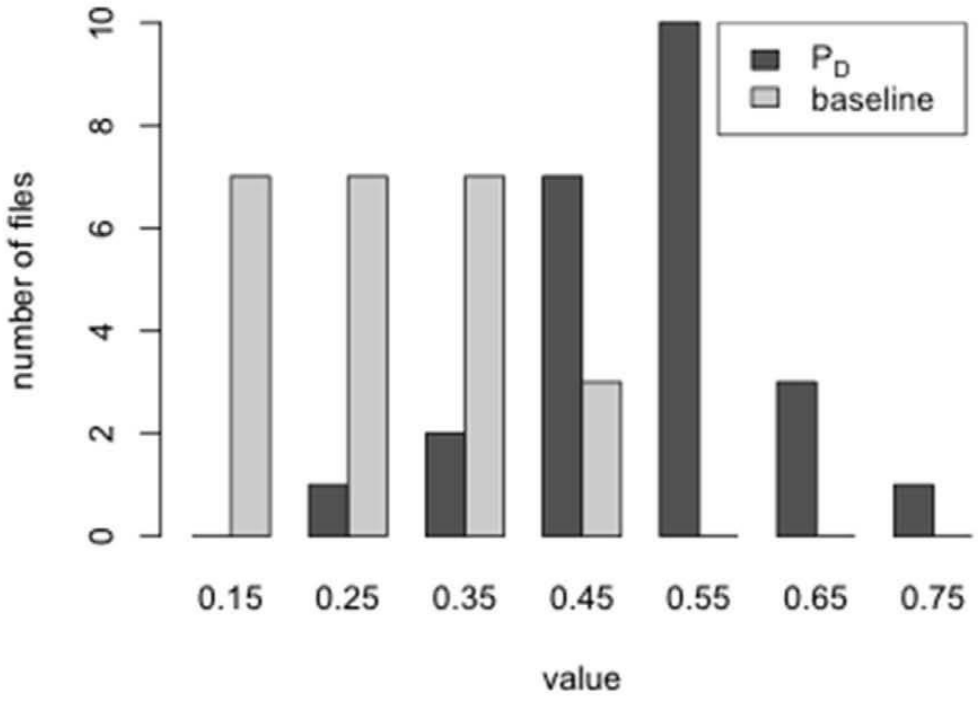


Histogram

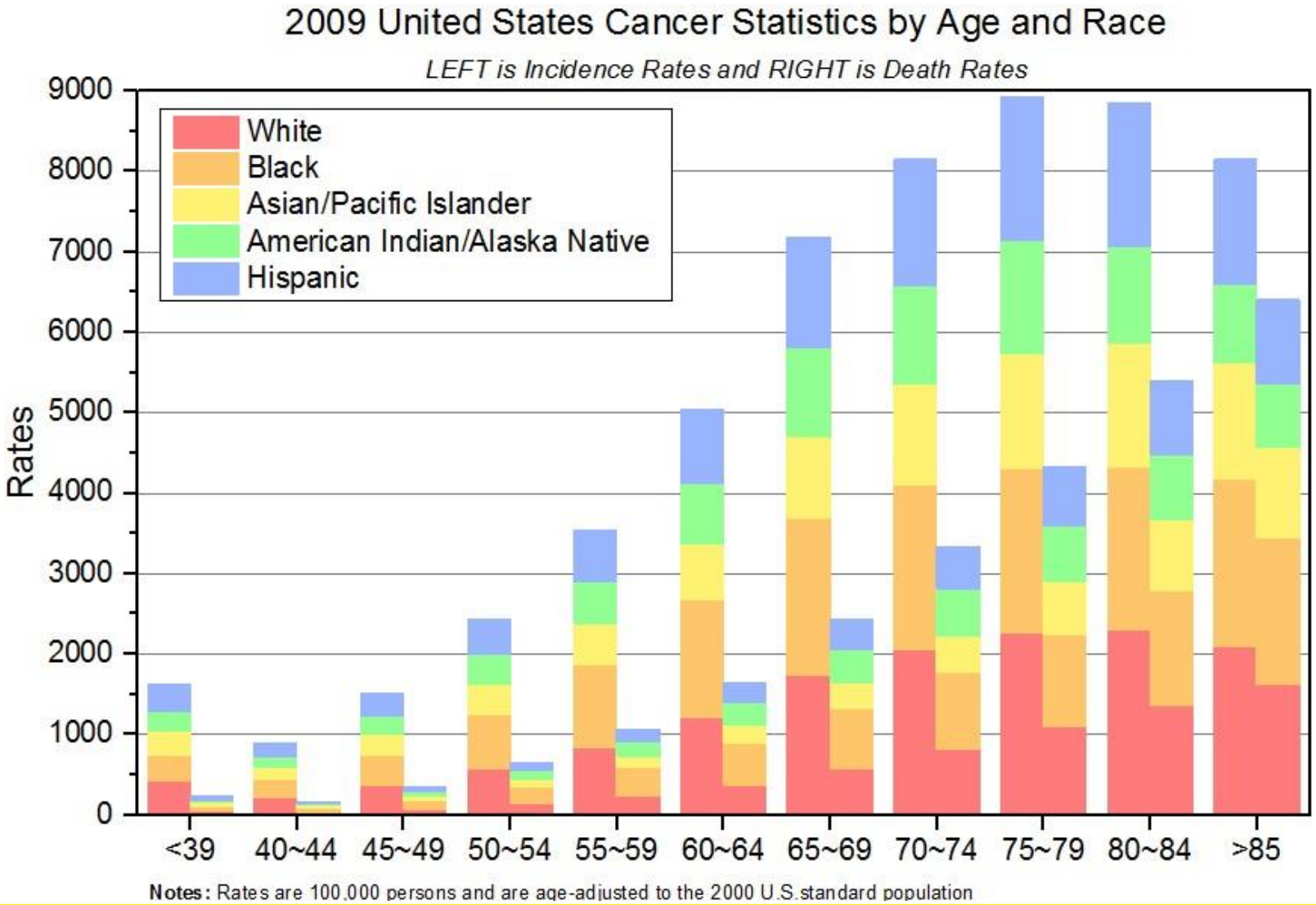
- Important first way of looking at your data
- One dimensional
- Shows shape by binning a continuous distribution



Grouped histogram



Grouped histogram



Pie chart

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

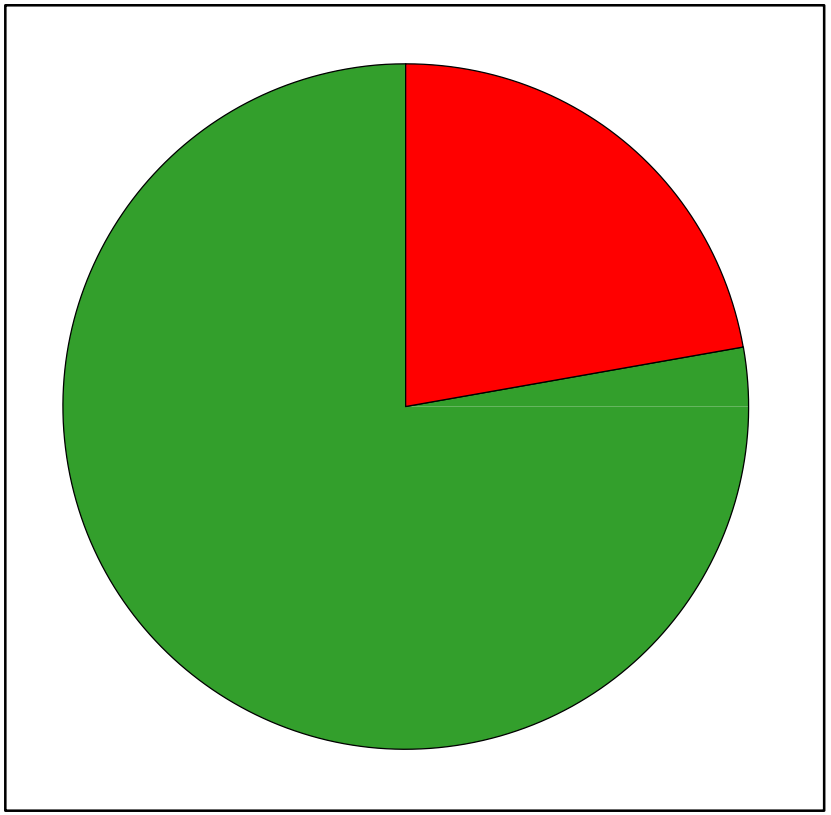
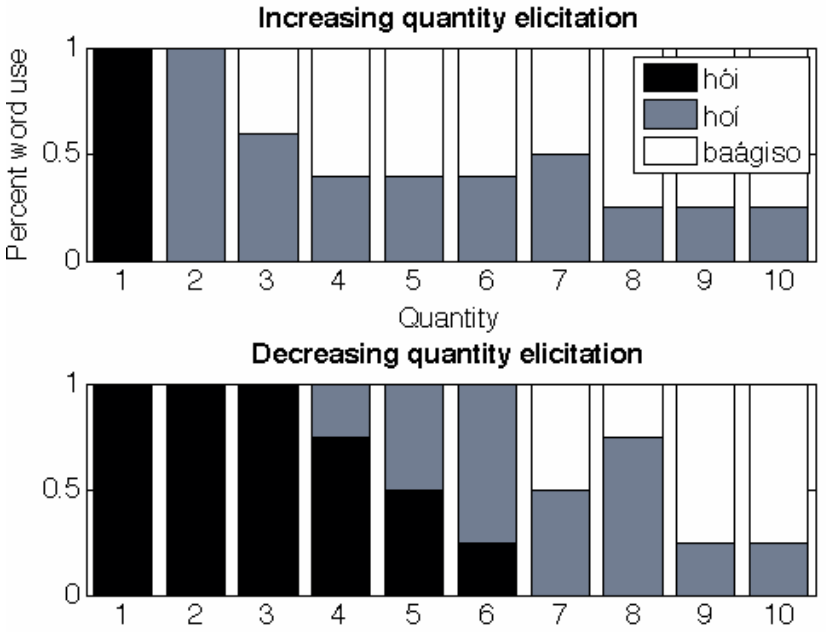


Figure by MIT OpenCourseWare.

Stacked bar graph

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



Courtesy Elsevier, Inc., http://www.sciencedirect.com. Used with permission.

Venn diagram

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

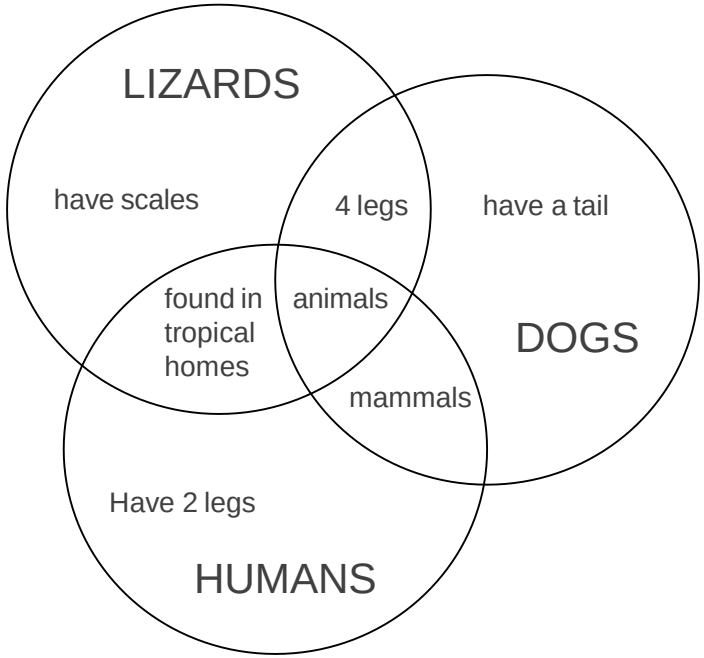
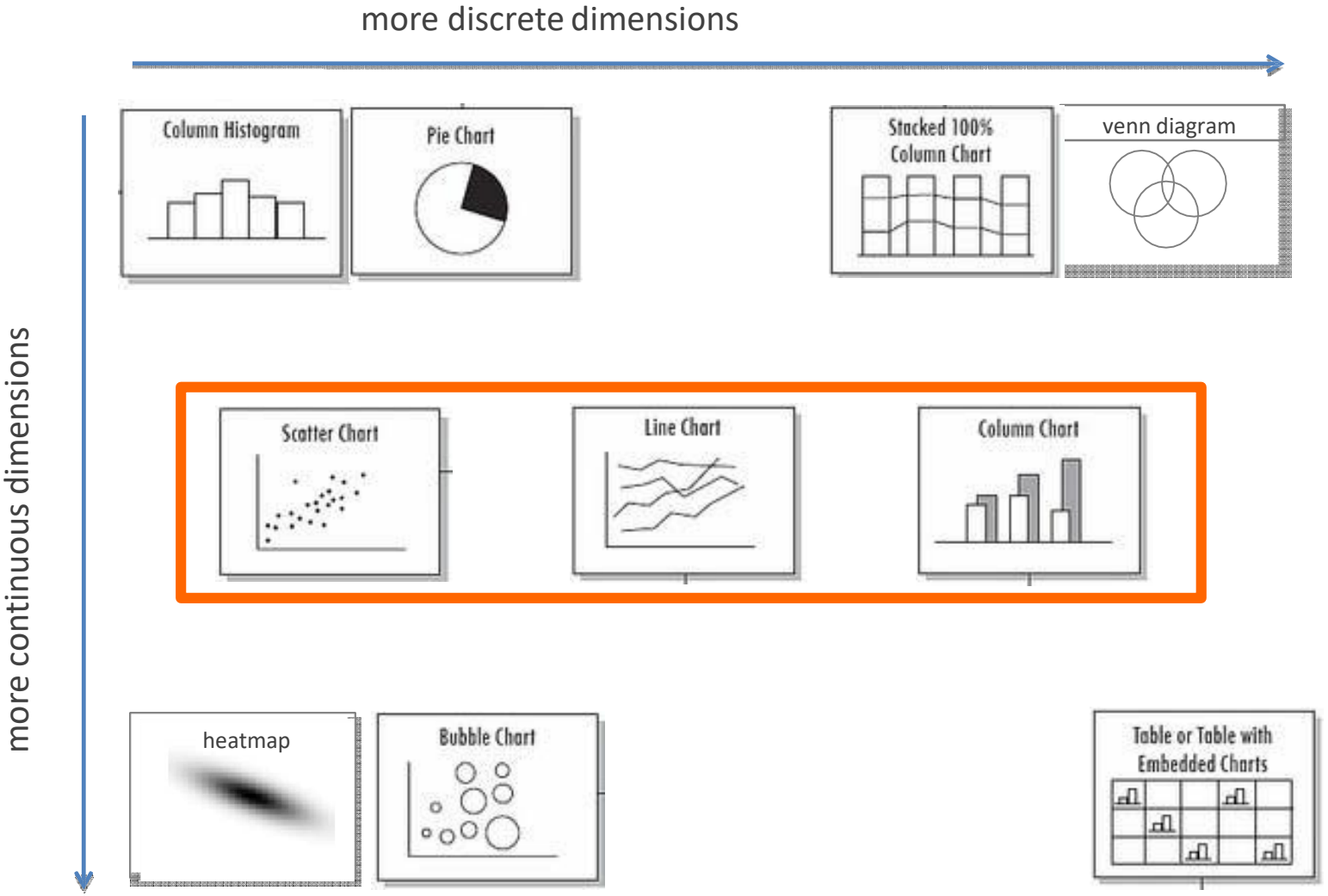


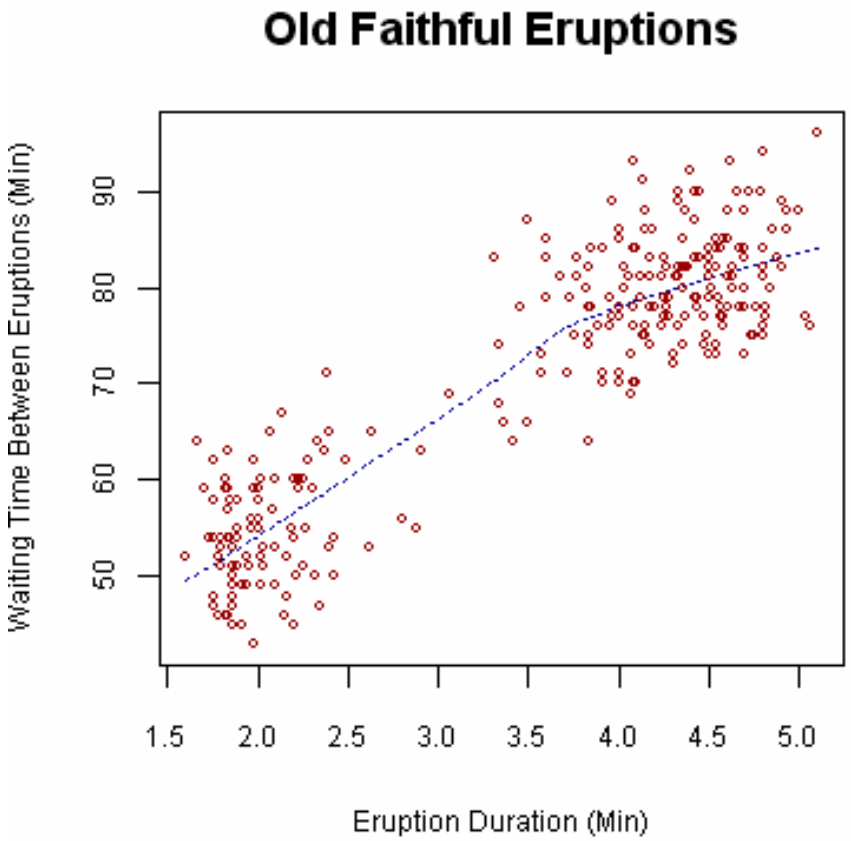
Figure by MIT OpenCourseWare.

Conventional visualizations



Scatter plot

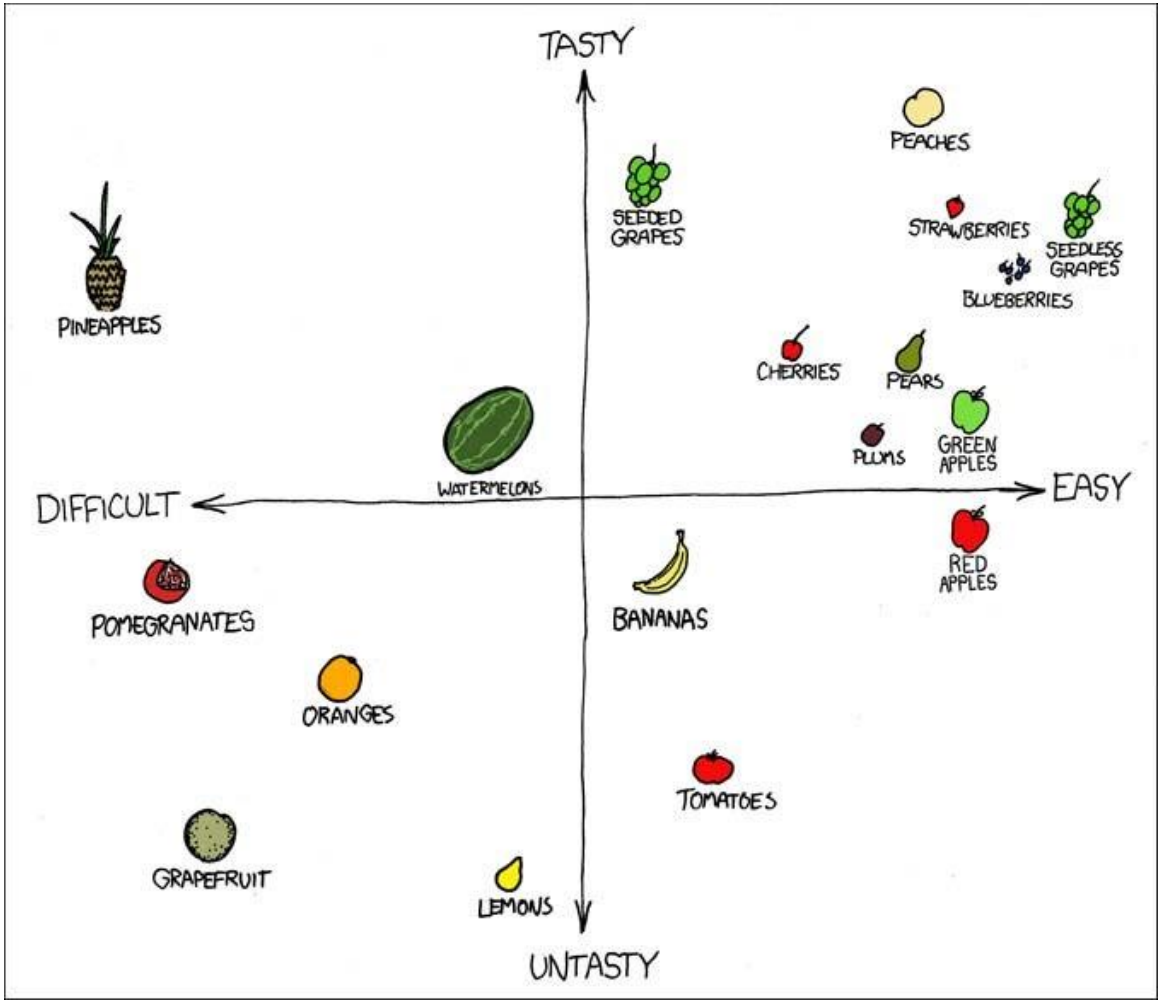
- Relationship between observations on two continuous dimensions
- Can show multiple groups
- Can show trend lines etc.
- Uninformative with too much data



Courtesy of xkcd.org.

Scatter plot

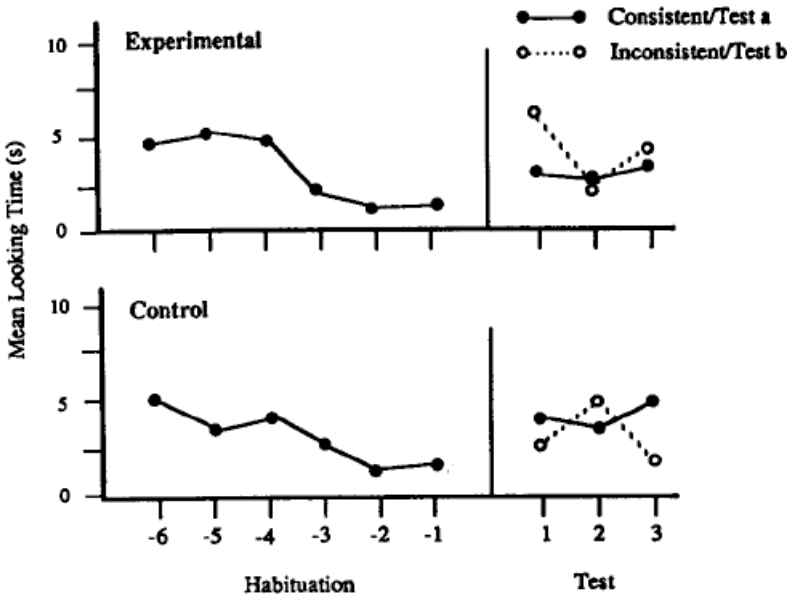
… with many discrete items (identity as a dimension)



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



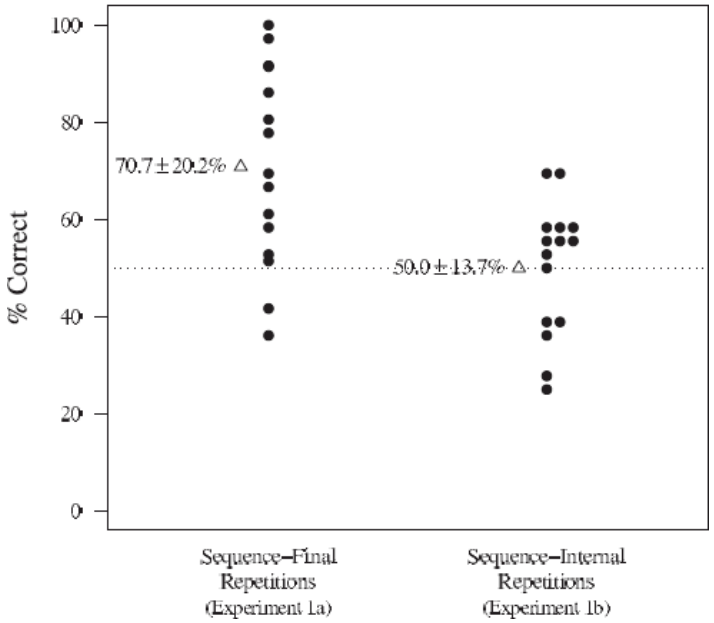
Courtesy of American Psychological Association. Used with permission.

Bar graph

- Can be used for lots of discrete grouping factors
- Natural semantics of grouping
- Conceals data

More bar graphs

Strip chart

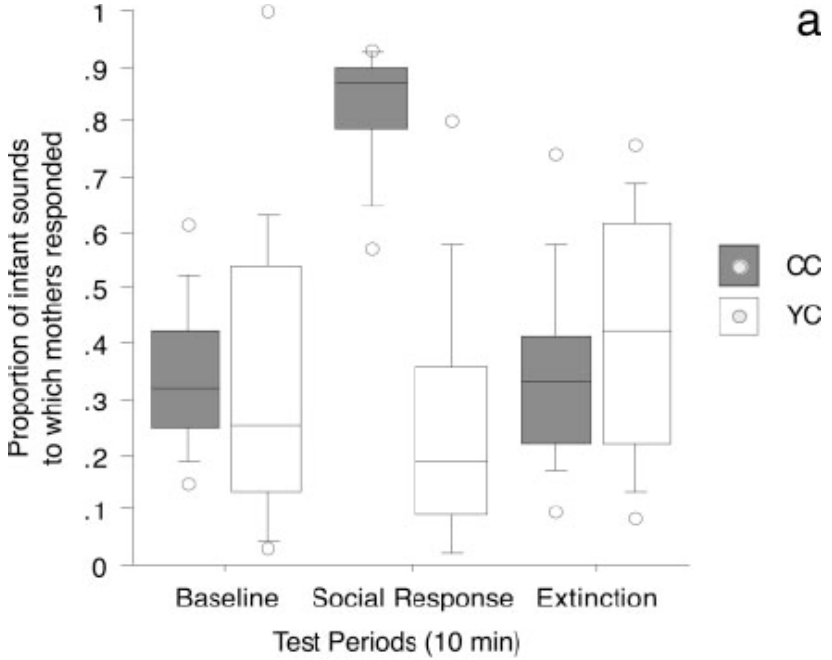


Very useful for showing individual subject means

Courtesy of American Psychological Association

Endress, Ansgar D. et. Al. "The Role of Salience in the Extraction of Algebraic Rules."
Journal of Experimental Psychology: General, Vol. 134, No. 3 (2005): 406-419.

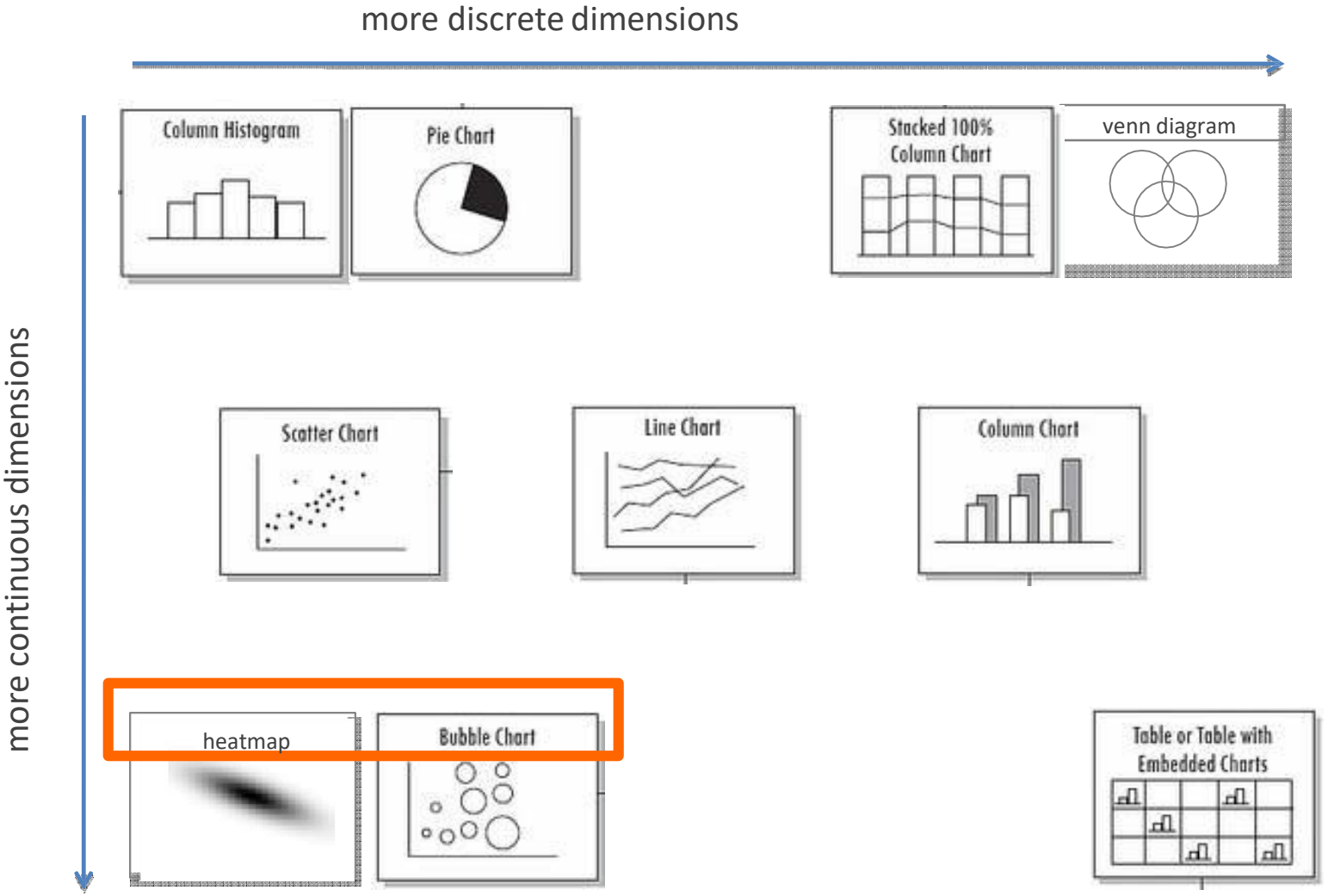
Box plot



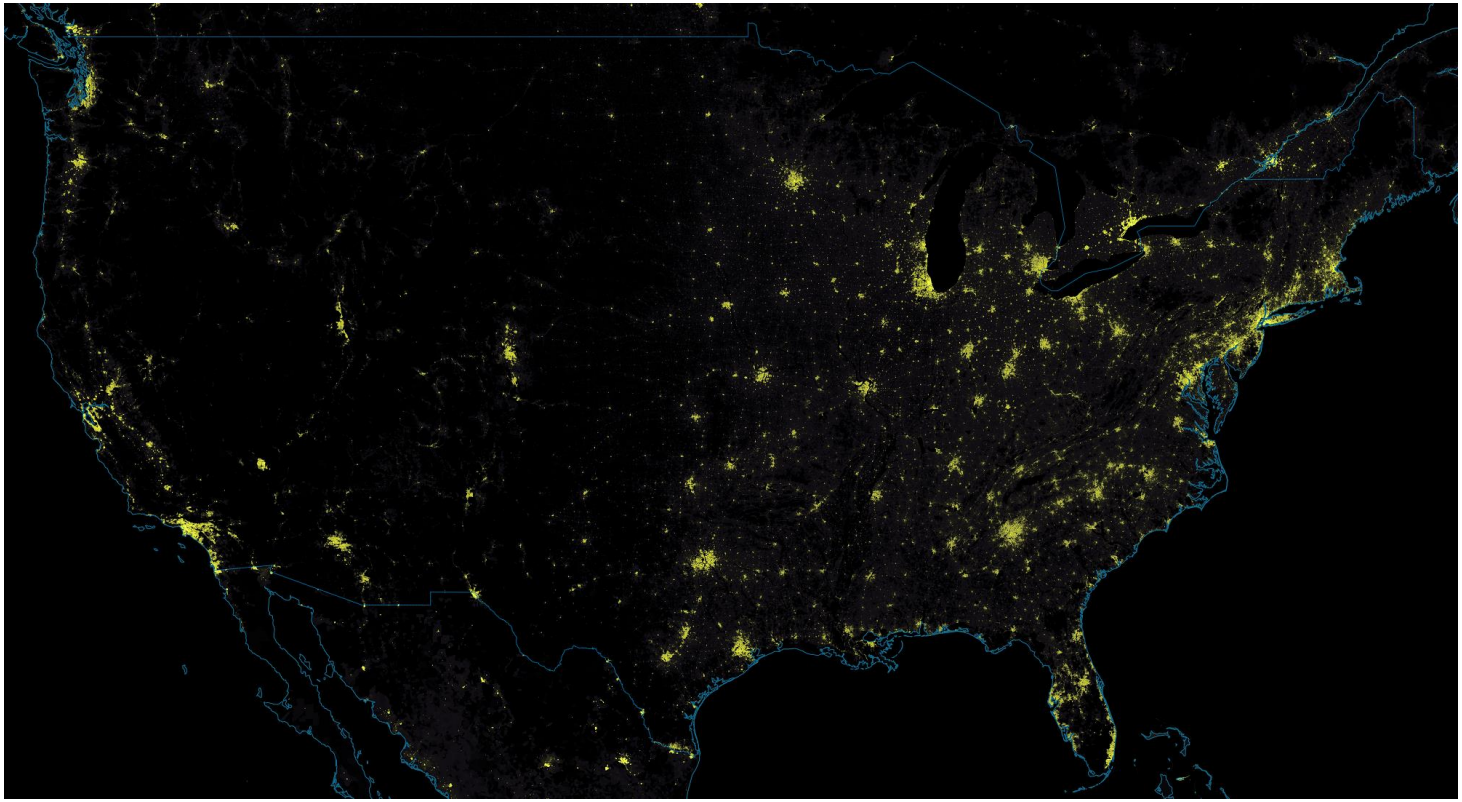
Shows the shape of distribution but not focused on individual subjects

Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Goldstein et. al. "Social Interaction Shapes Babbling: Testing Parallels Between Birdsong and Speech." *PNAS* 100, no. 13 (2003): 8030-8035.
Copyright © 2003, National Academy of Sciences, U.S.A.

Conventional visualizations

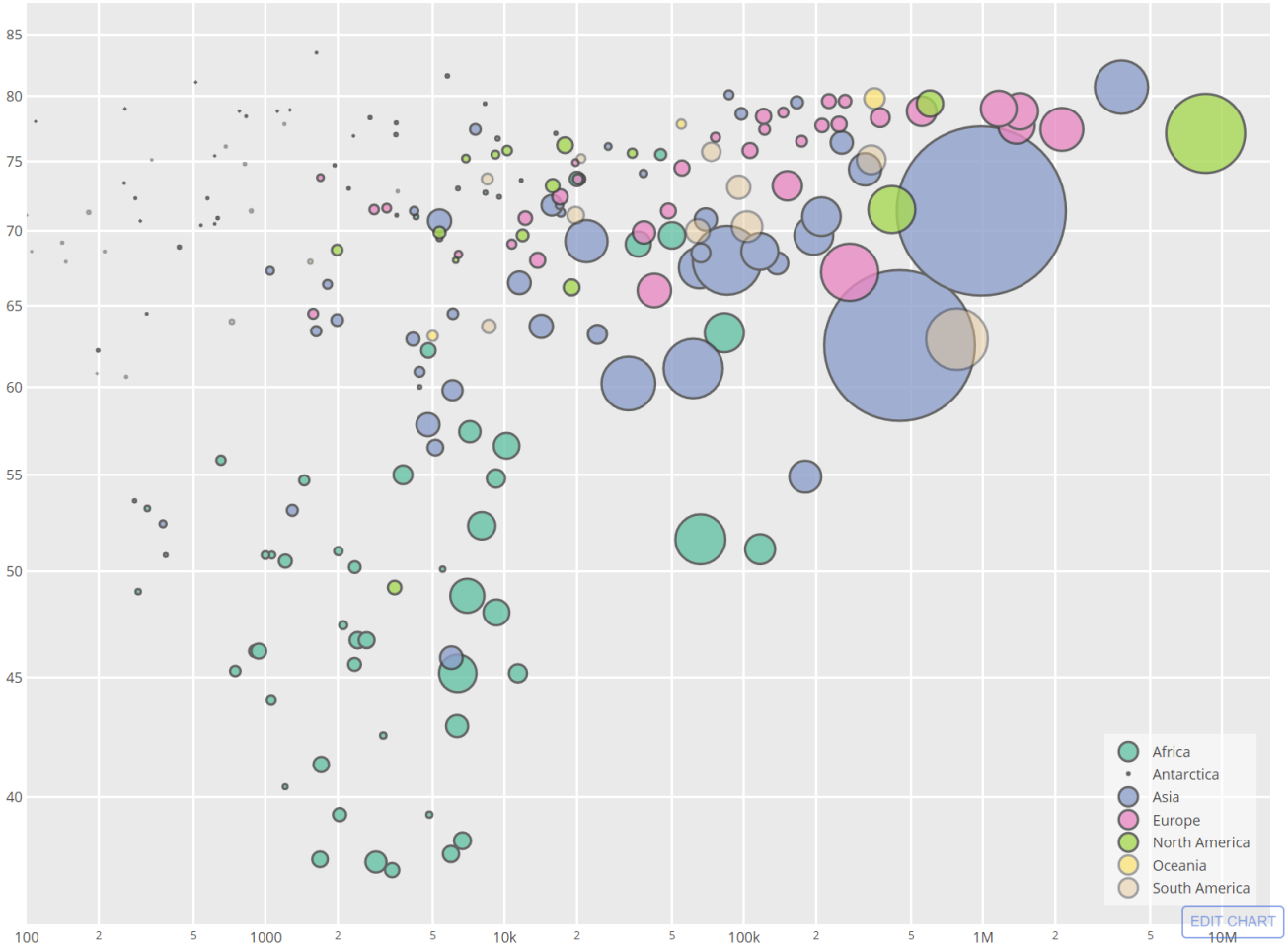


Heat map



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

Bubble plot

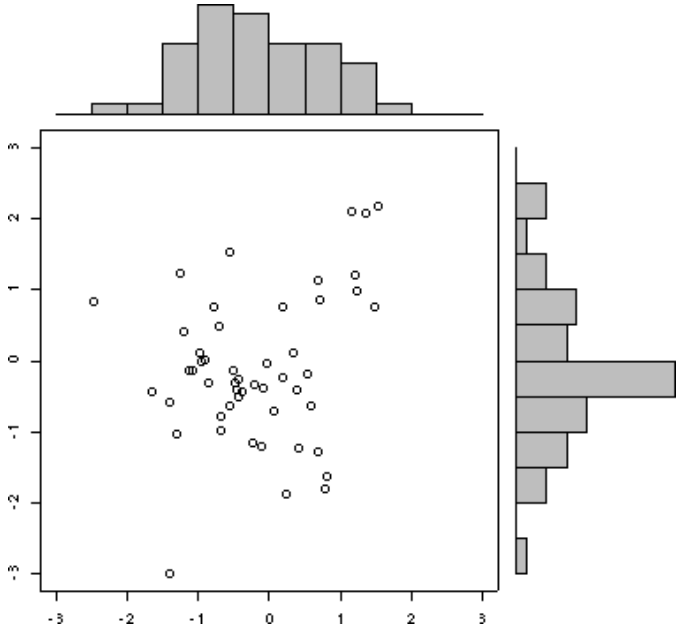


- A bubble chart is a type of chart that displays three dimensions of data. Each entity with its triplet (v1, v2, v3)
- Can be very intuitive
- Size is not perfectly quantitative

Three tricks for doing more with less

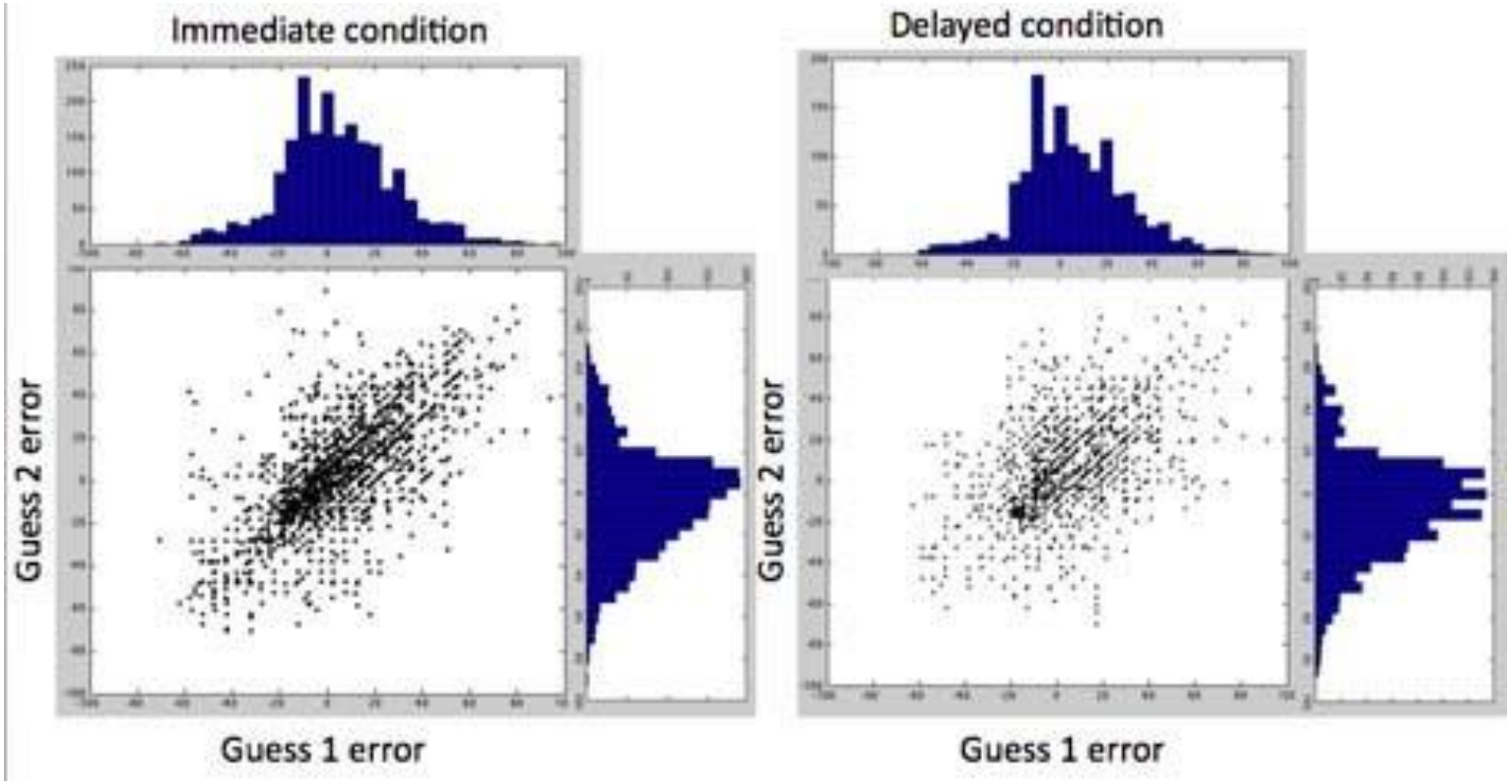
- Multiple plots
  - simple, easily interpretable subplots
  - can be beautiful but overwhelming
- Hybrid plots
  - a scatter plot of histograms
  - or a venn-diagram of histograms, etc.
- Multiple axes
  - plot two (or more) different things on one graph

Hybrid plots

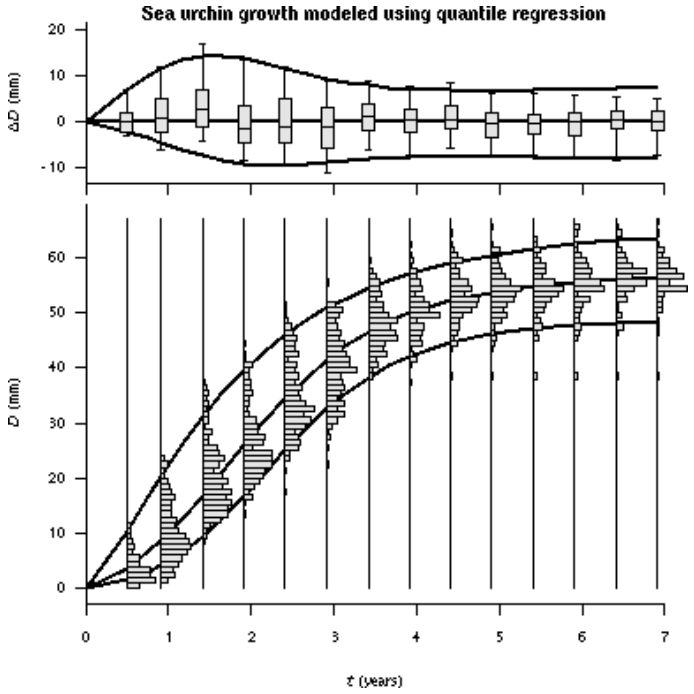


Courtesy of http://addictedtor.free.fr/graphiques/addNote.php?graph=78

Hybrid plots



Hybrid plots



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

Multiple plots

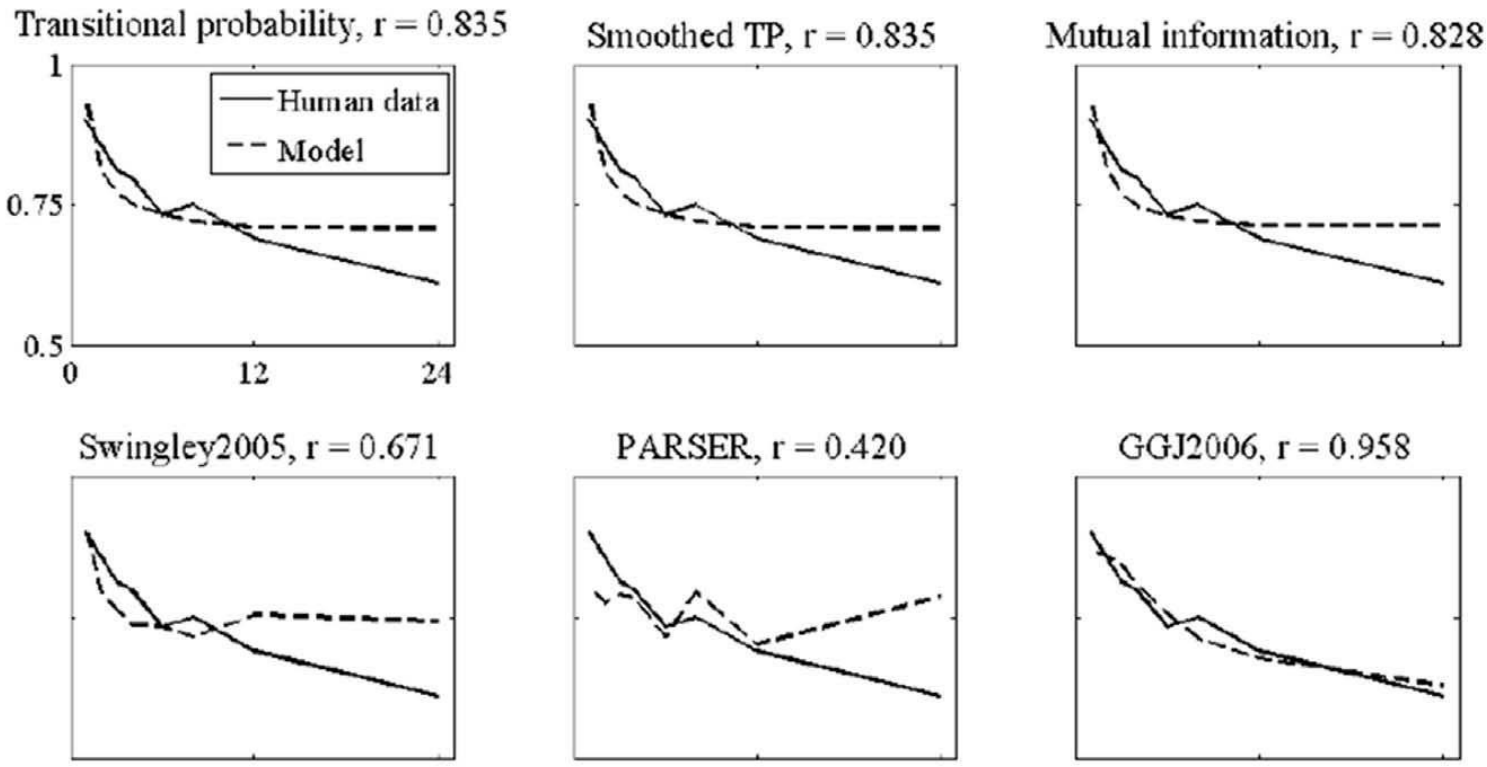


Figure 2. Best linear fit of each model's performance to human data, graphed by sentence length. The vertical axis represents decision probabilities for models and percentage correct for human data; the horizontal, sentence length.

Multiple plots

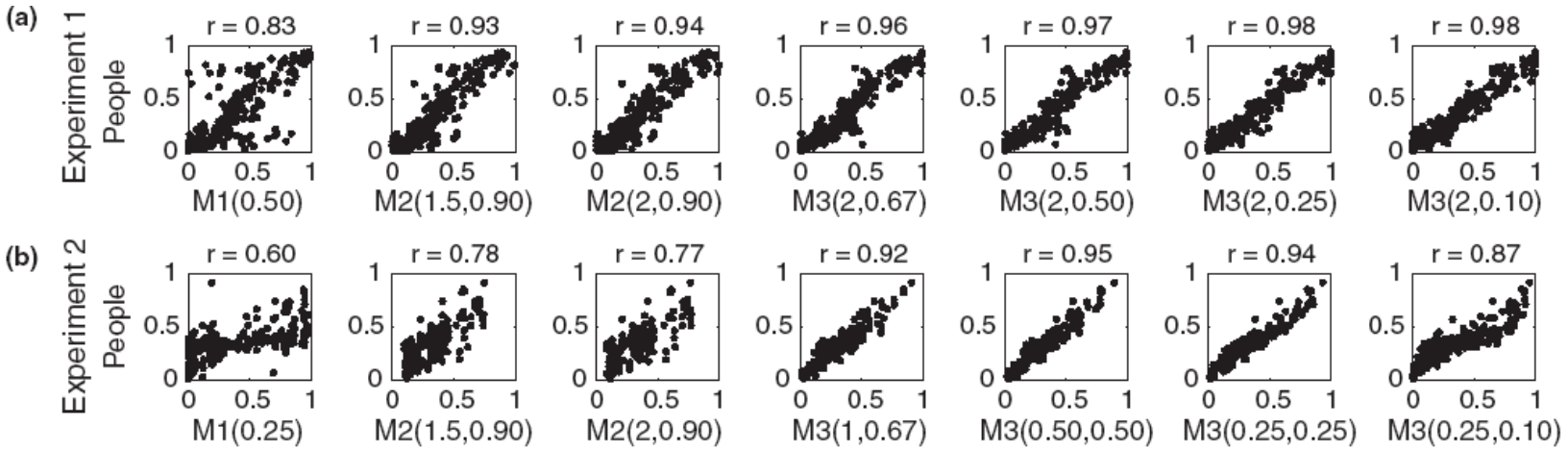
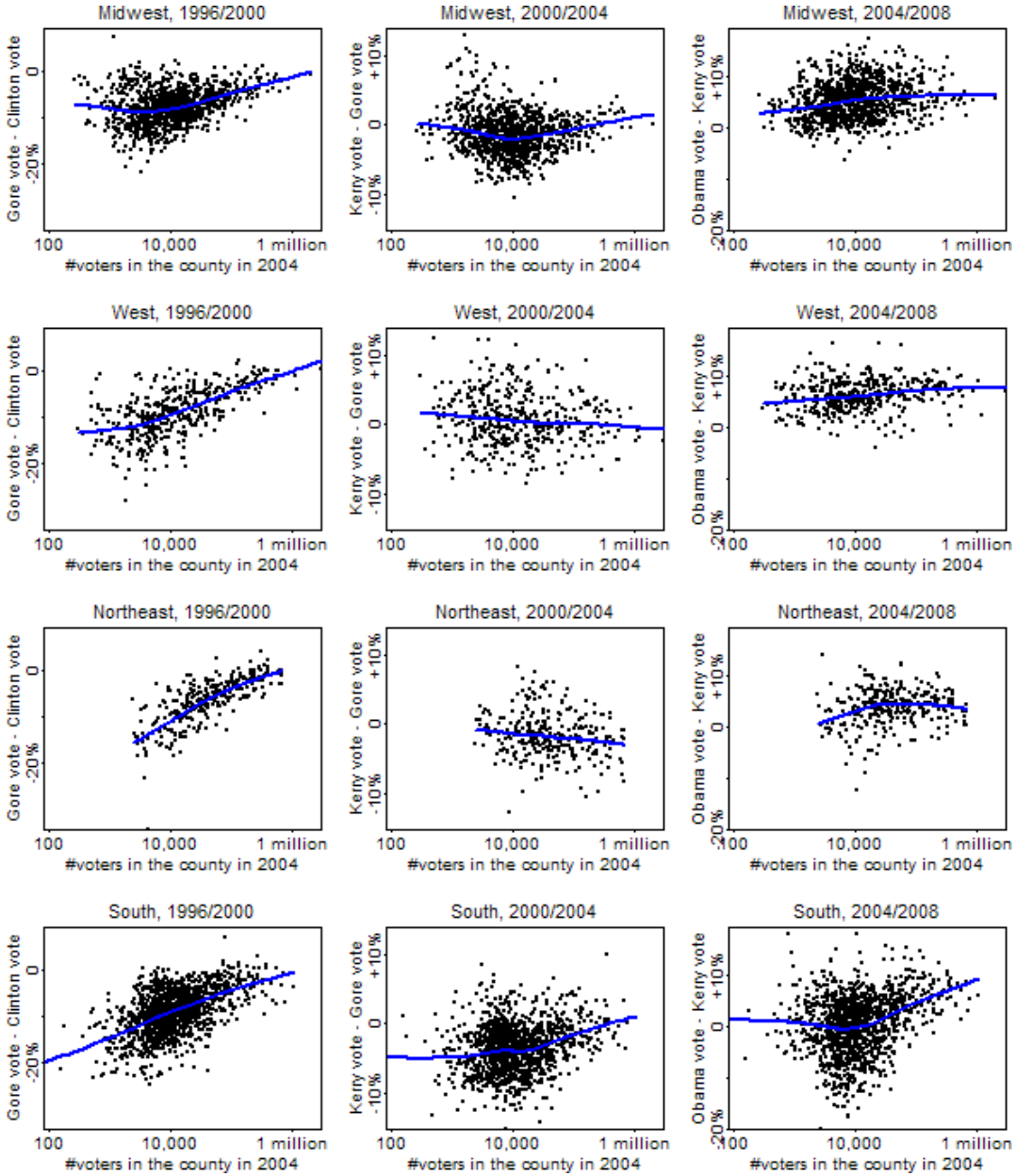


Figure 3: Example scatter plots of model predictions against subject ratings. Plots of model predictions use the parameter settings with the highest correlation from each model column of Tables 1 and 2. (a) Experiment 1 results. (b) Experiment 2 results.

Baker, Tenenbaum, & Saxe (2007)

Multiple plots



Multiple axes

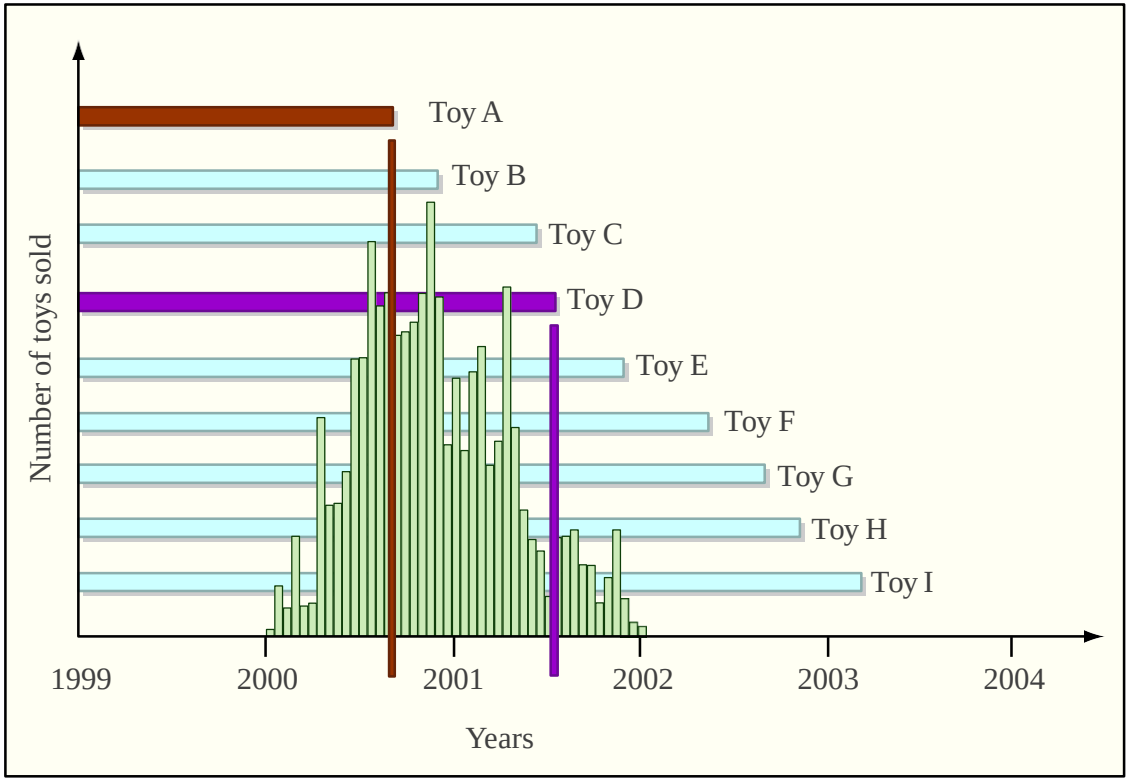
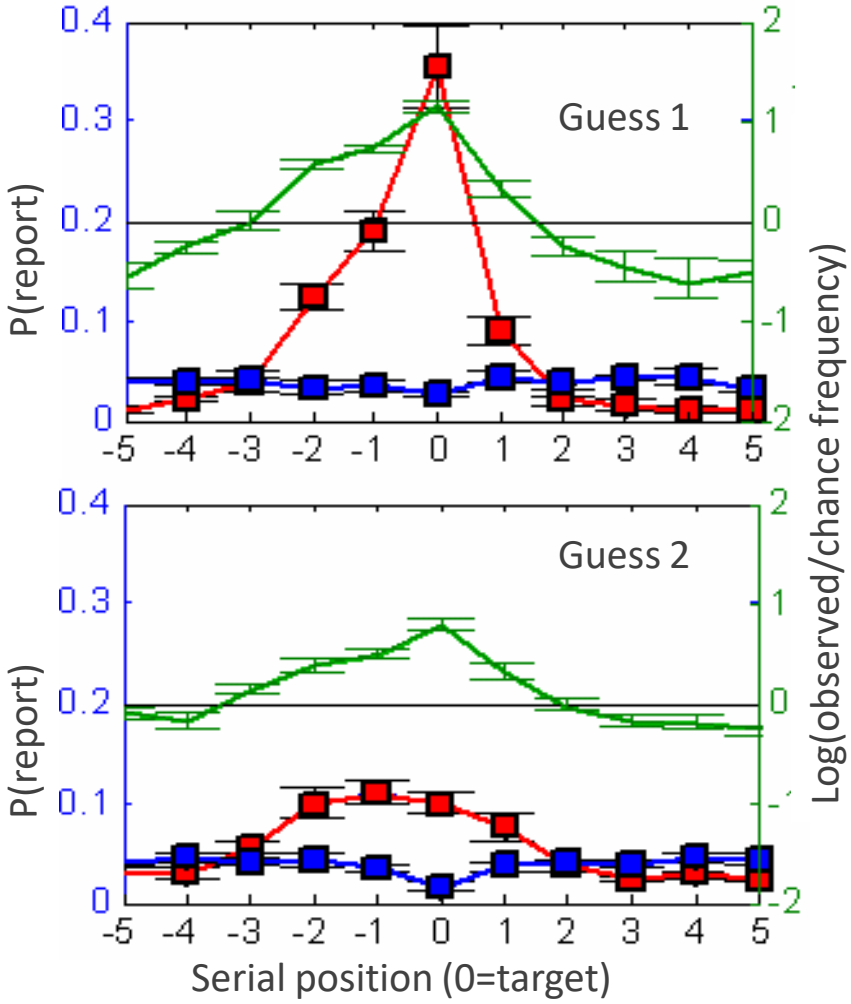


Figure by MIT OpenCourseWare.

Multiple axes



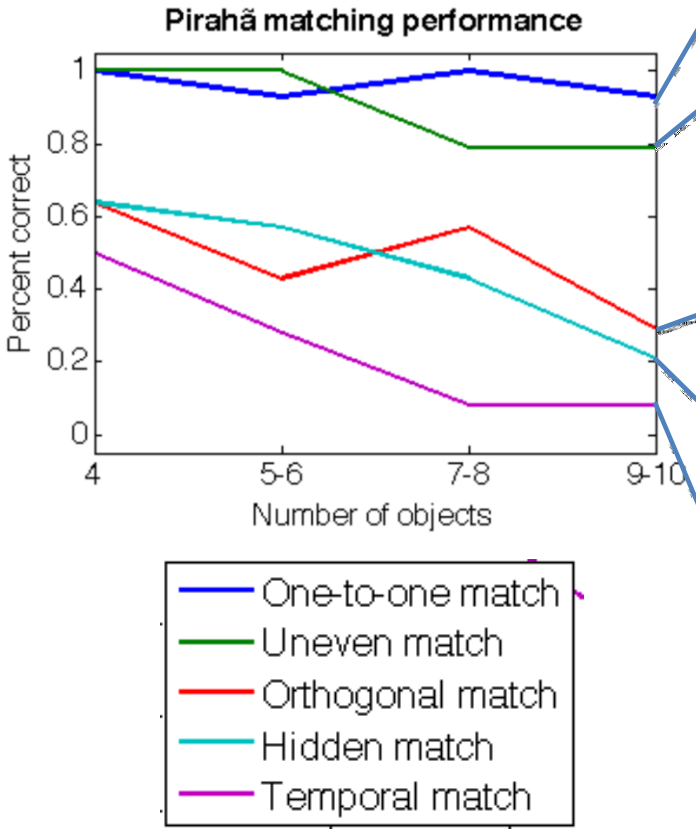
Two tradeoffs

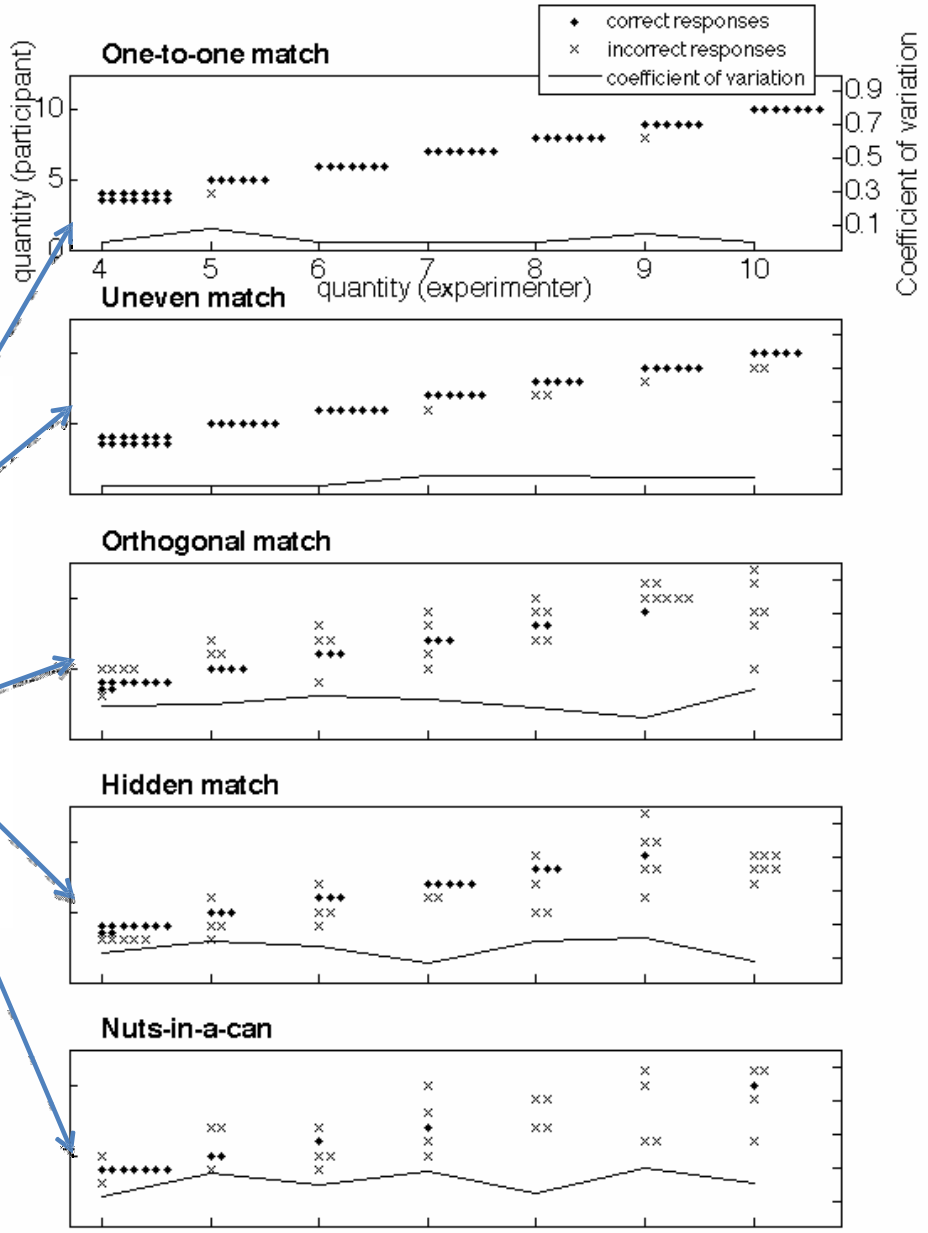
- Informativeness vs. readability
  - Too little information can conceal data
  - But too much information can be overwhelming
  - Possible solution: hierarchical organization?
- Data-centric vs. viewer-centric
  - Viewers are accustomed to certain types of visualization
  - But novel visualizations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

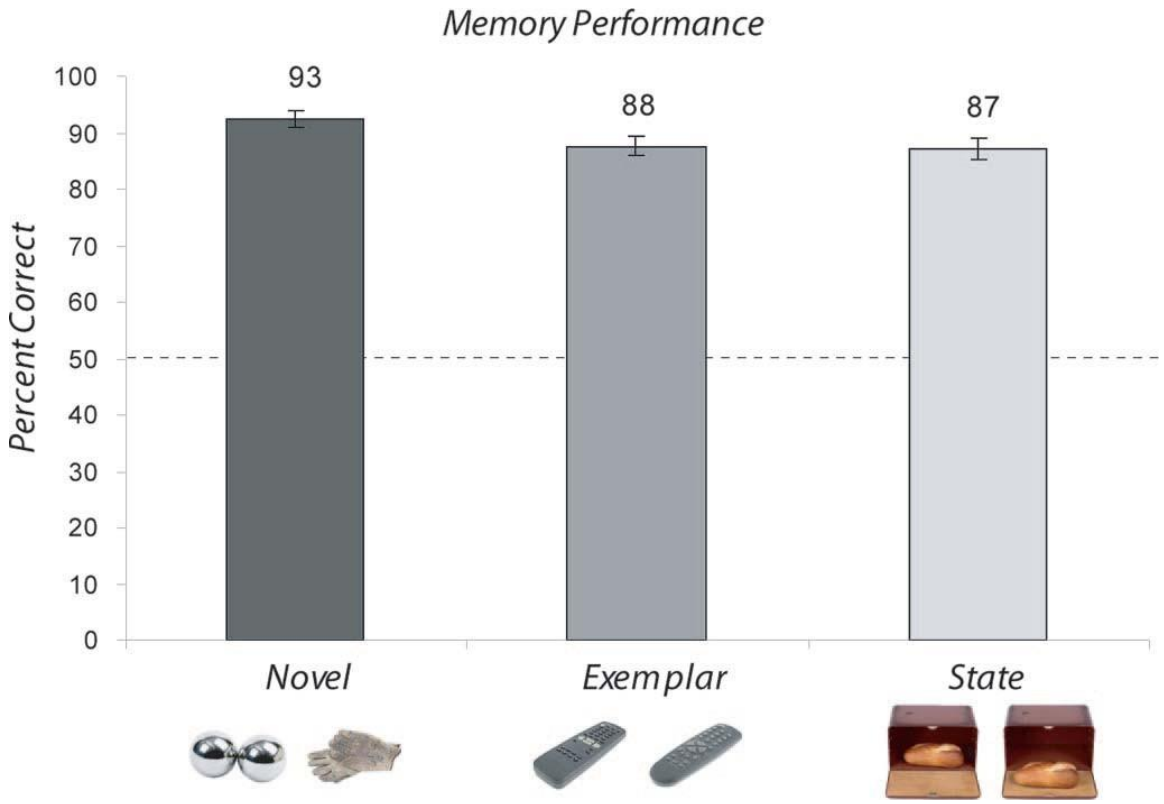




Information vs. readability

Brady, Konkle, Alvarez, Oliva (2008)

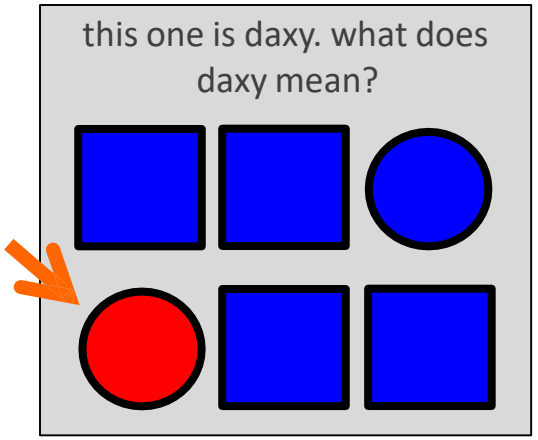
Information vs. readability



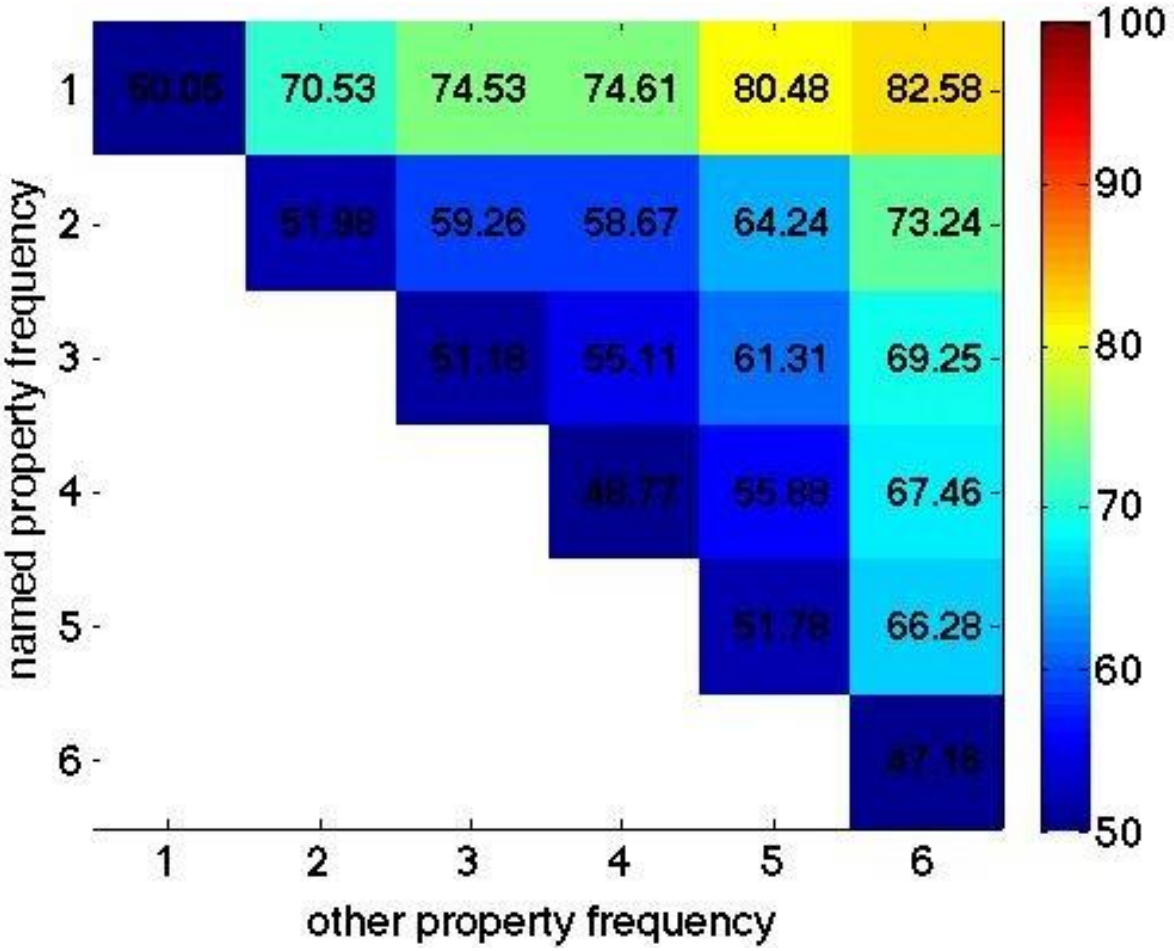
Brady, Konkle, Alvarez, Oliva (2008)

Data-centric vs. viewer-centric

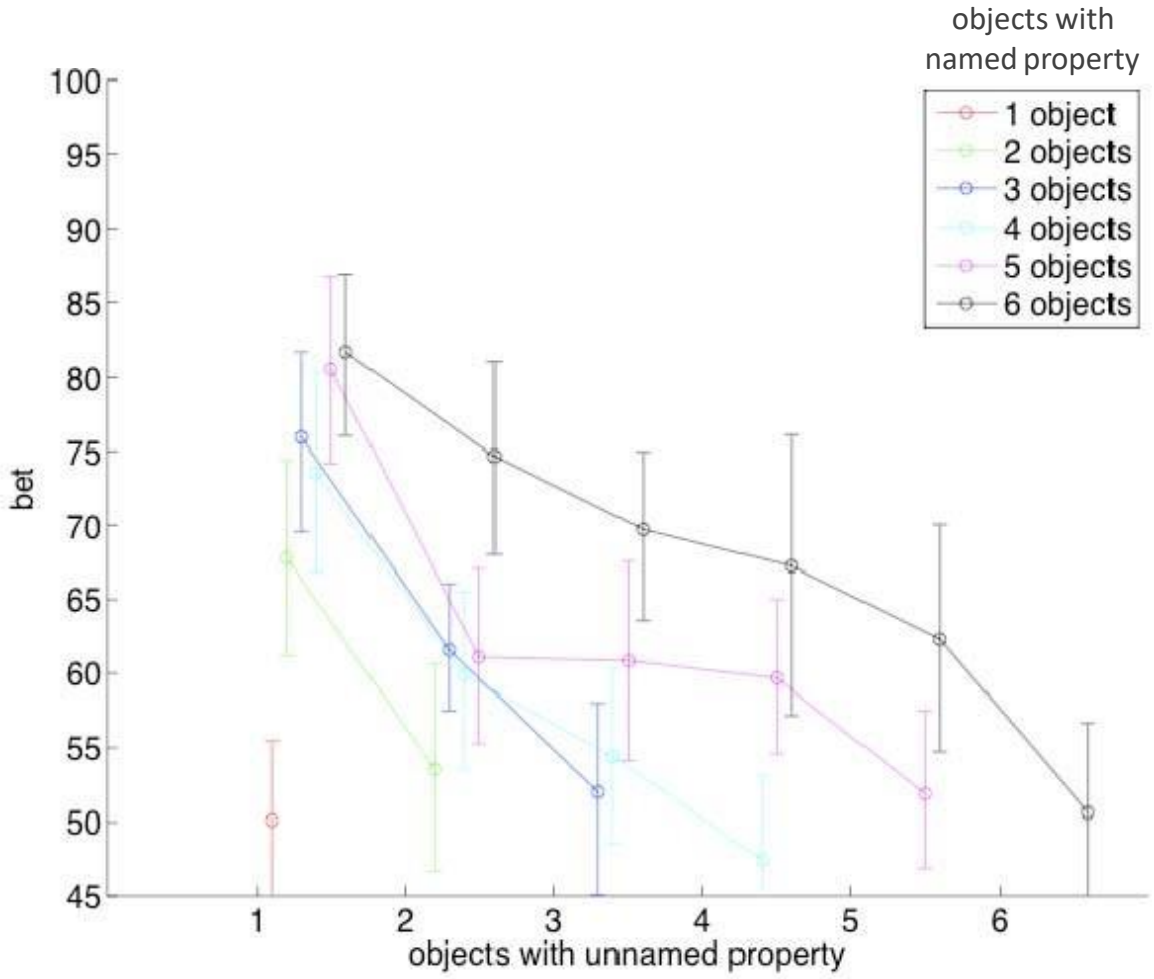
- Web study of word learning
  - n=700
  - lots of noise
- varied number of objects with different properties
  - asked for bets
- had a model that predicted performance



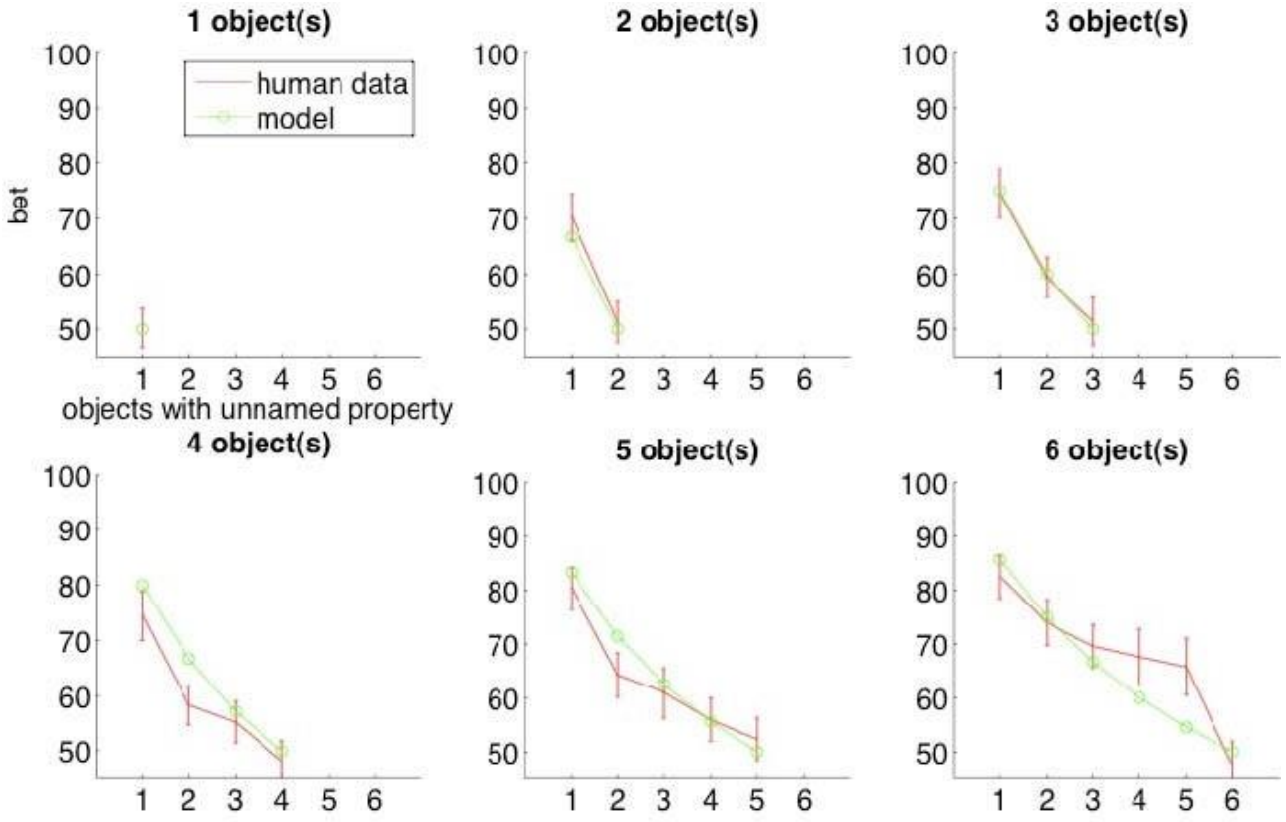
Data-centric vs. viewer-centric

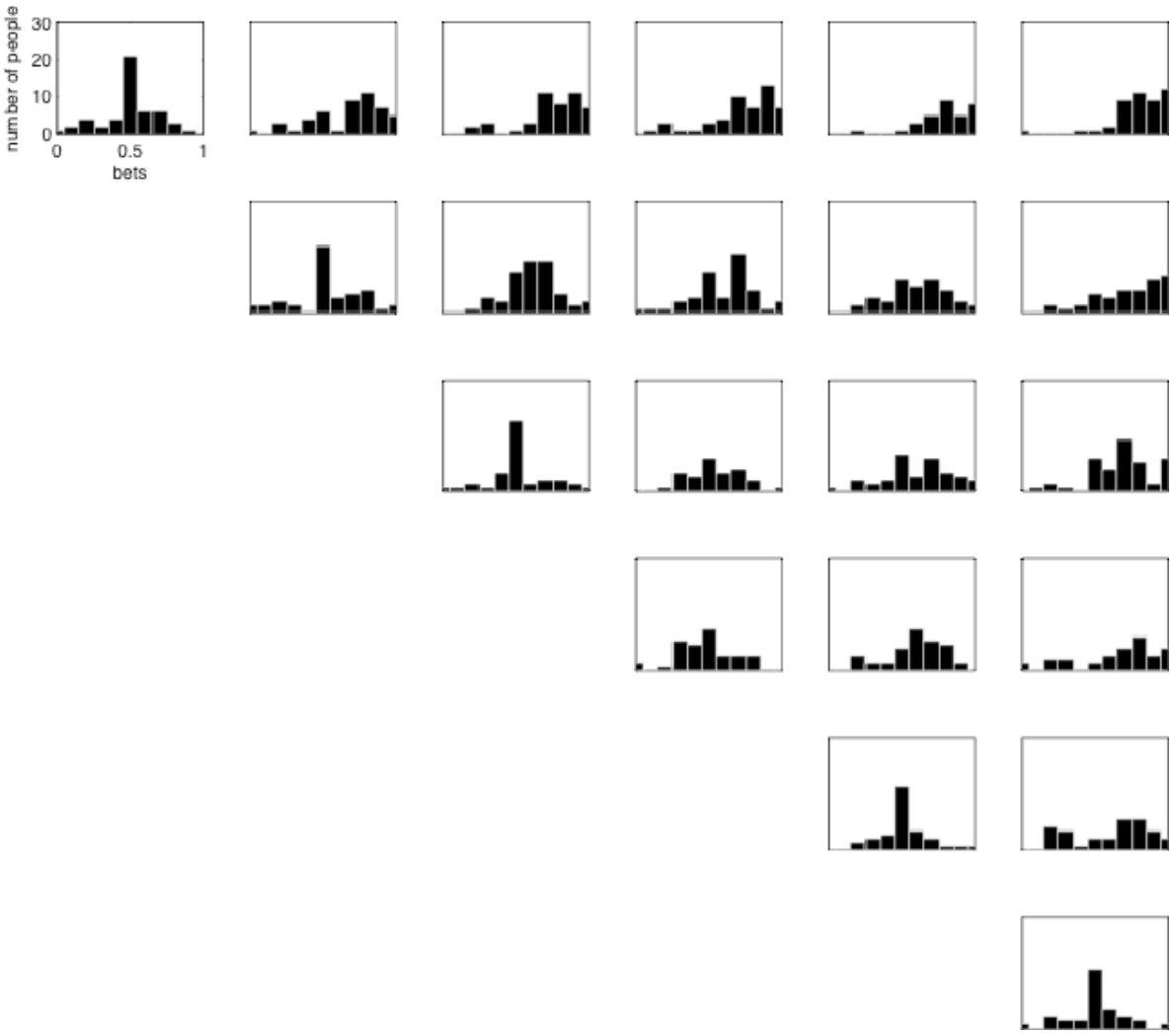


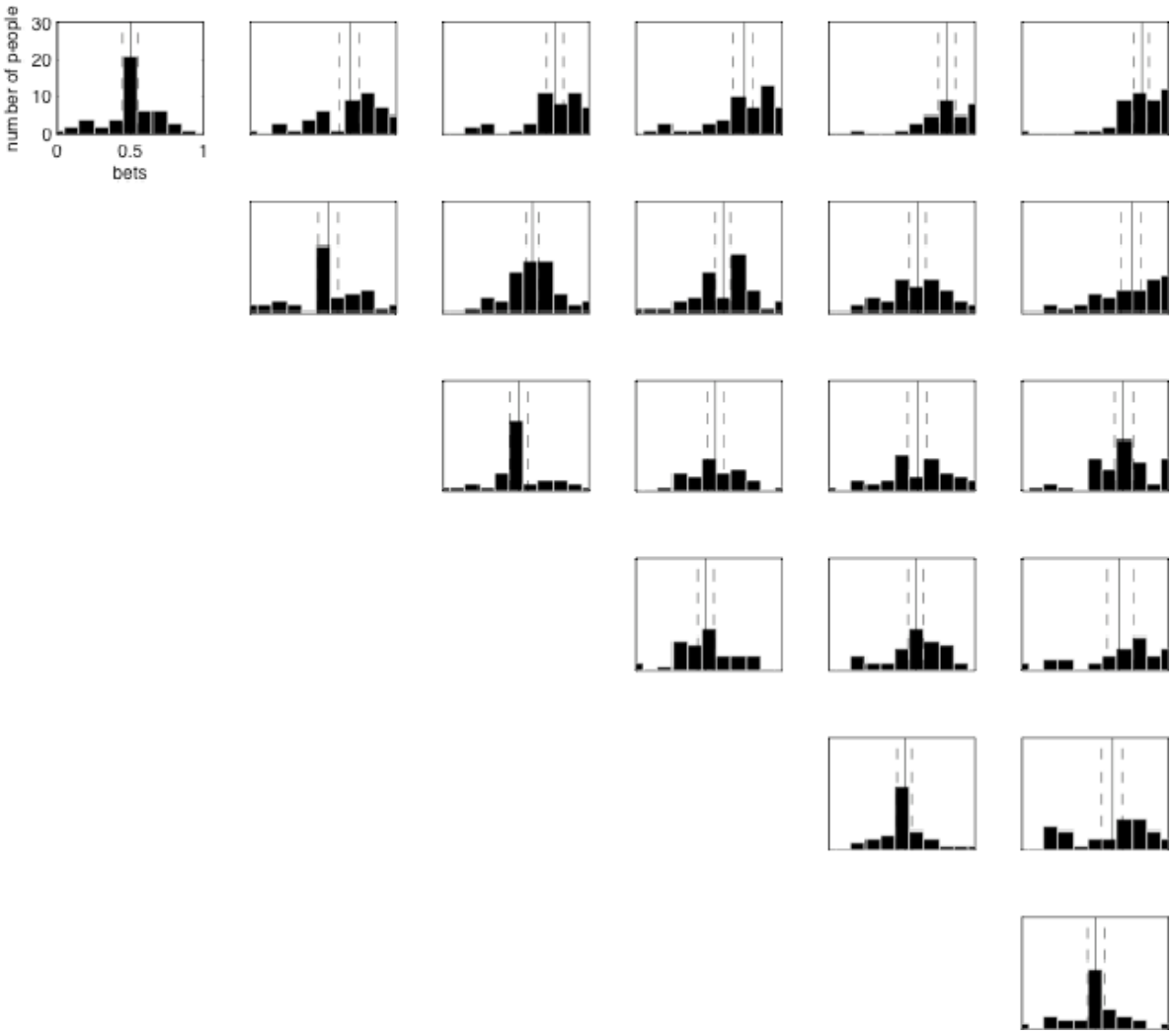
Data-centric vs. viewer-centric

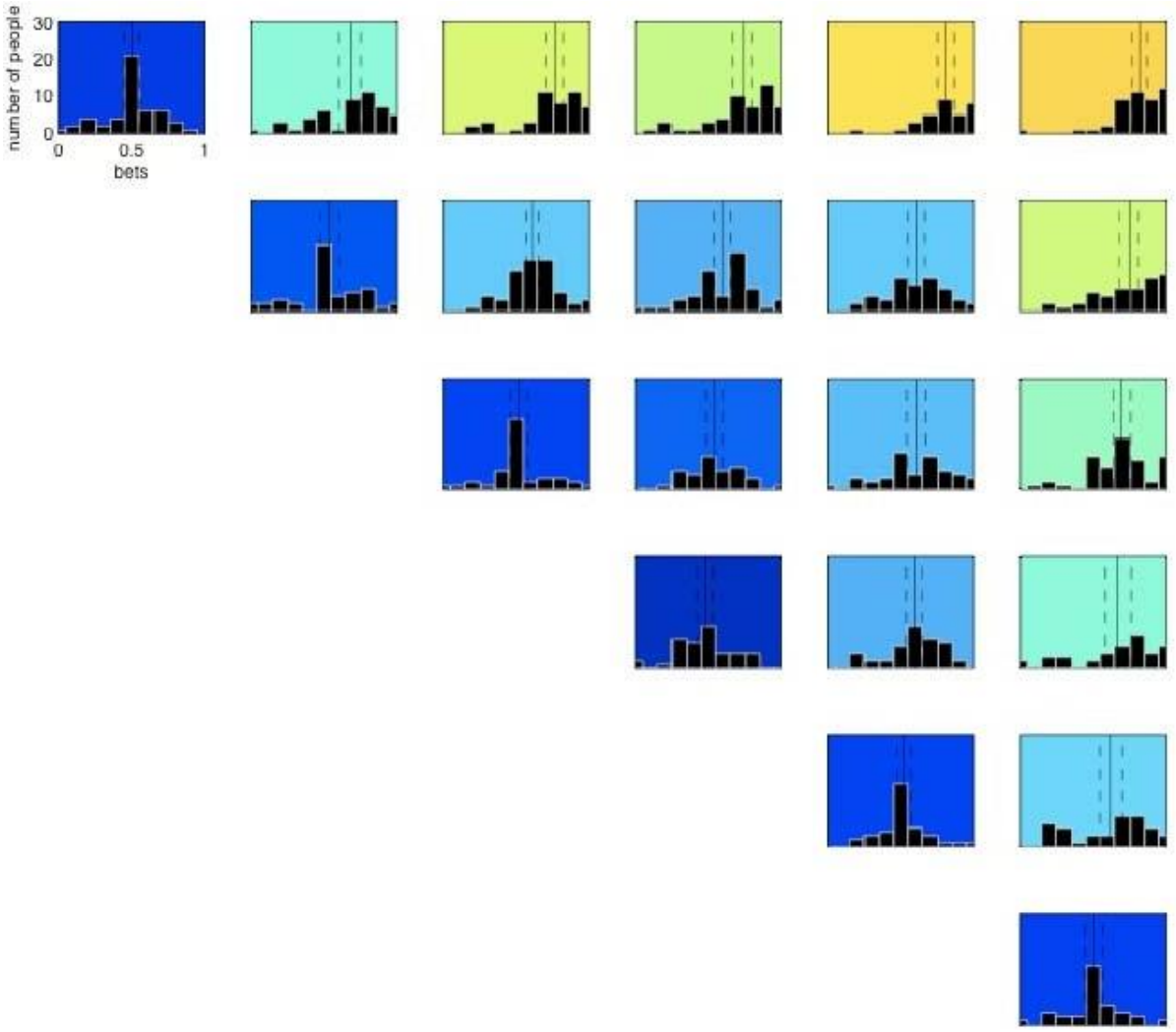


Data-centric vs. viewer-centric









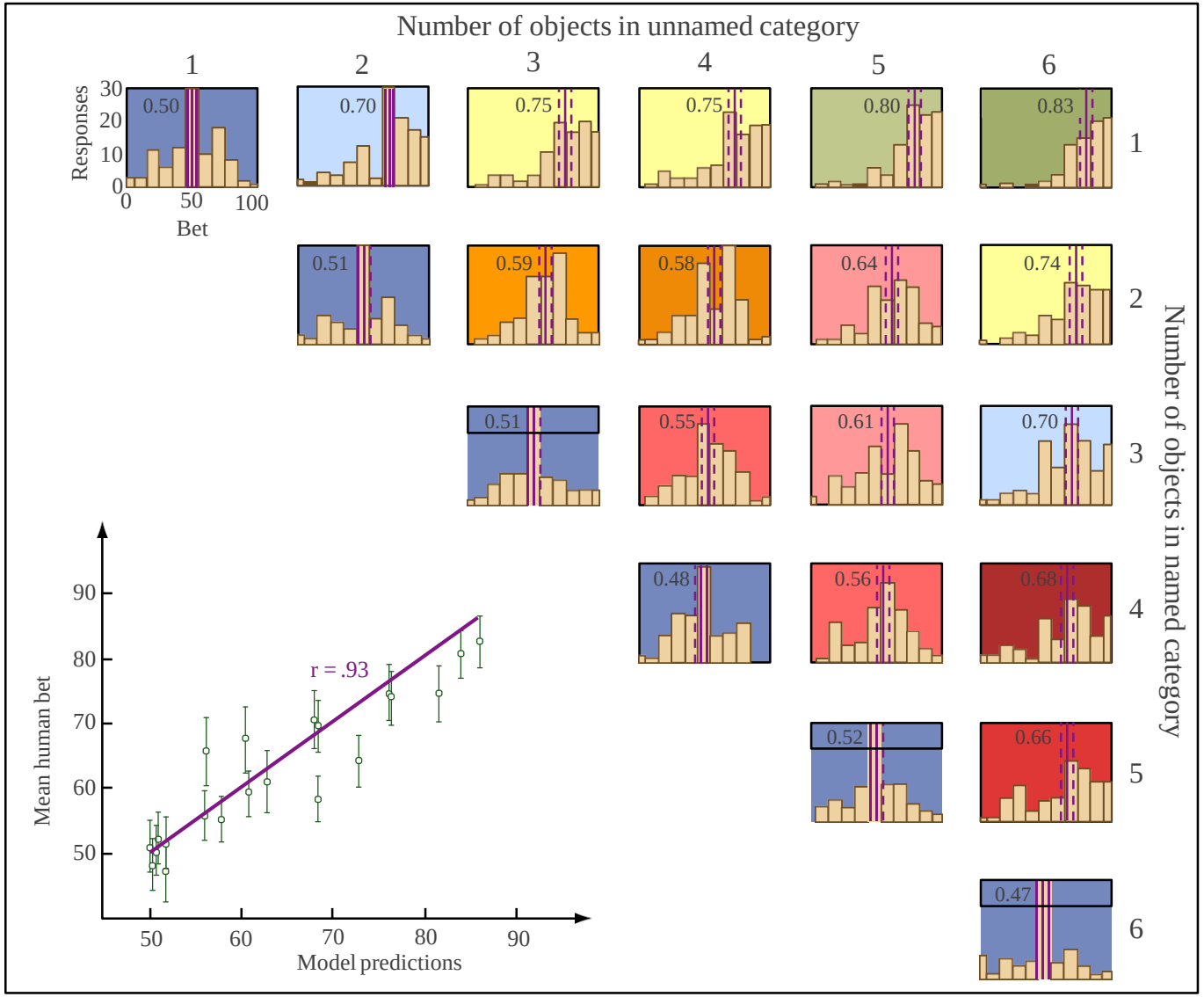
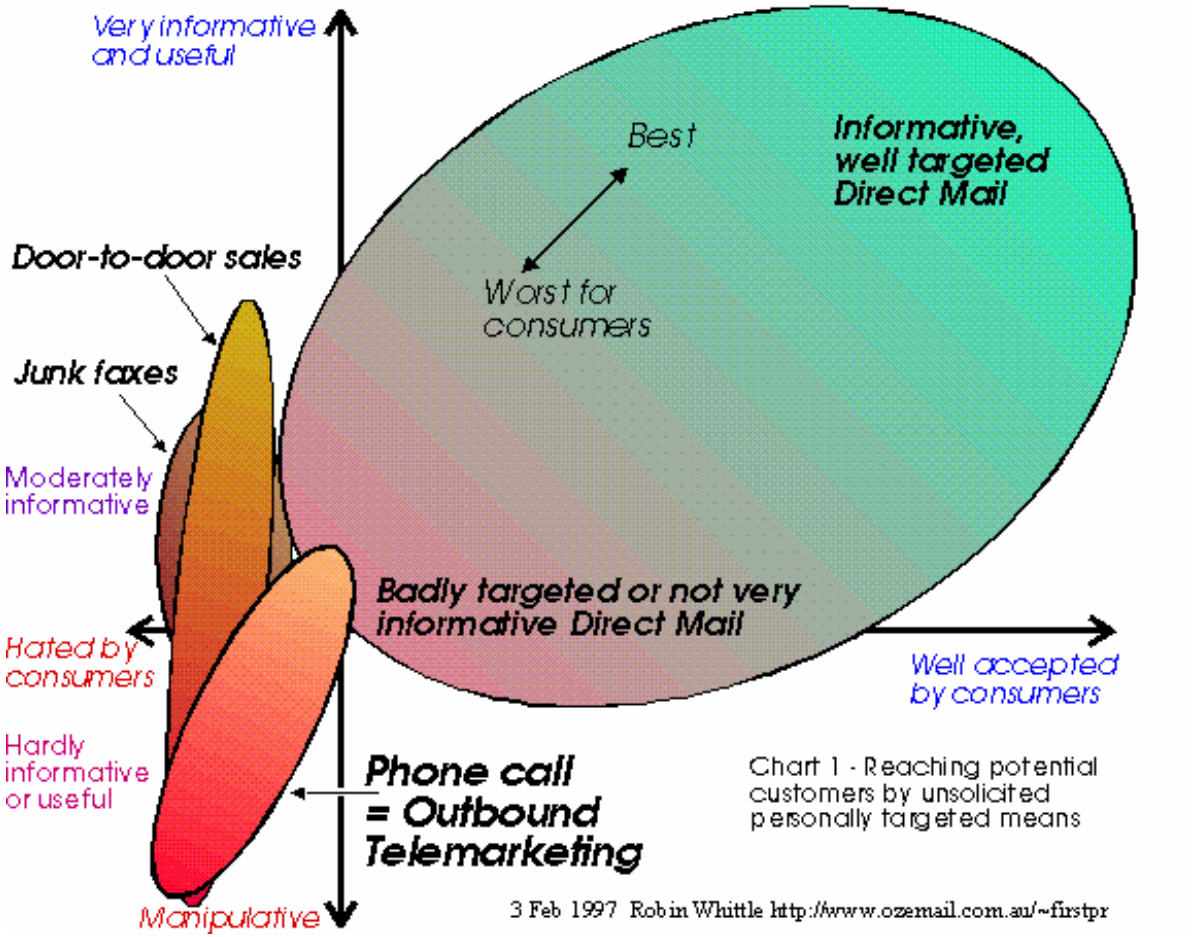
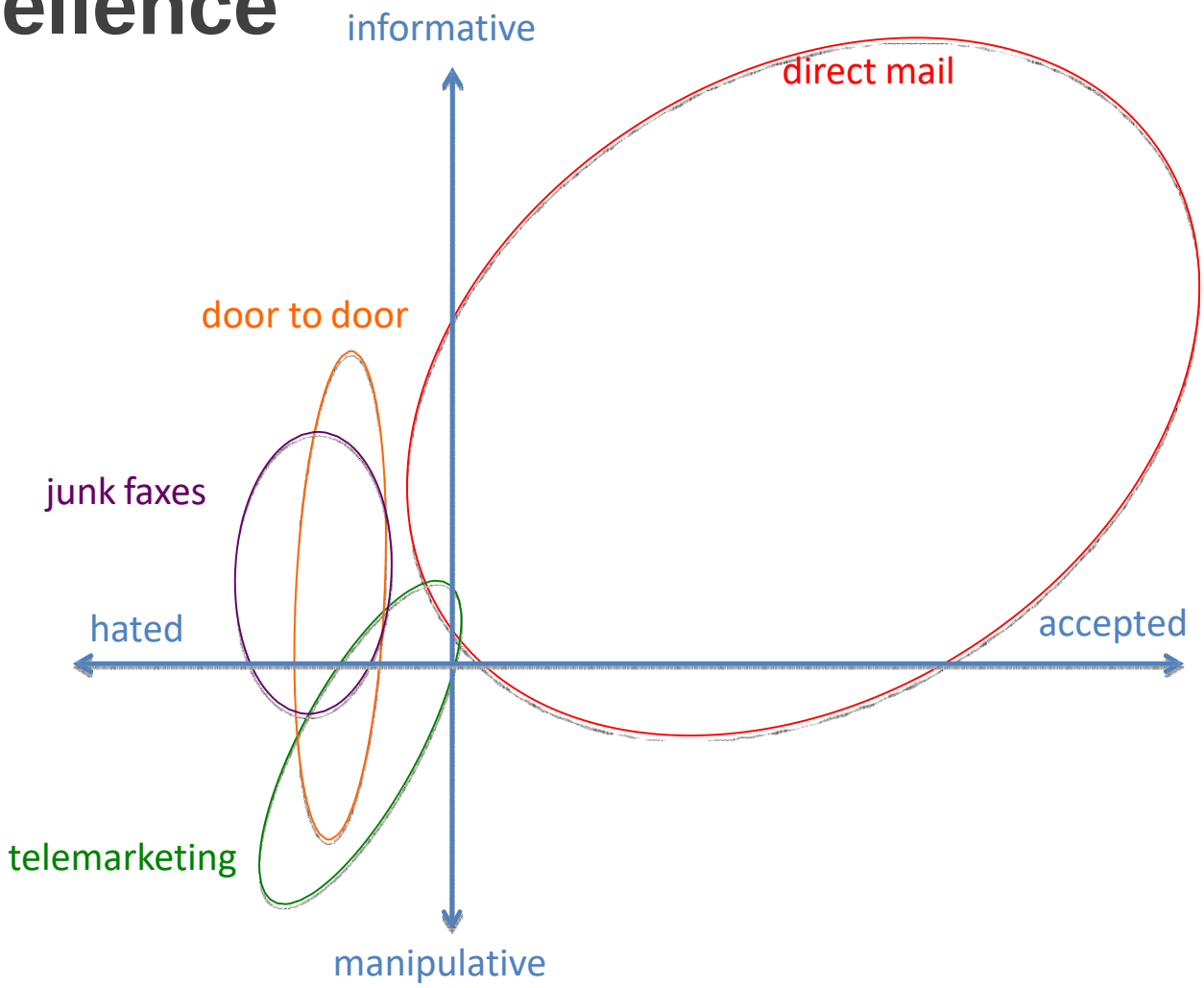


Figure by MIT OpenCourseWare.

High Dimensionality Doesn't Guarantee Excellence



High Dimensionality Doesn't Guarantee Excellence



Messy bar graphs

- Sometimes you can discretize way too many variables

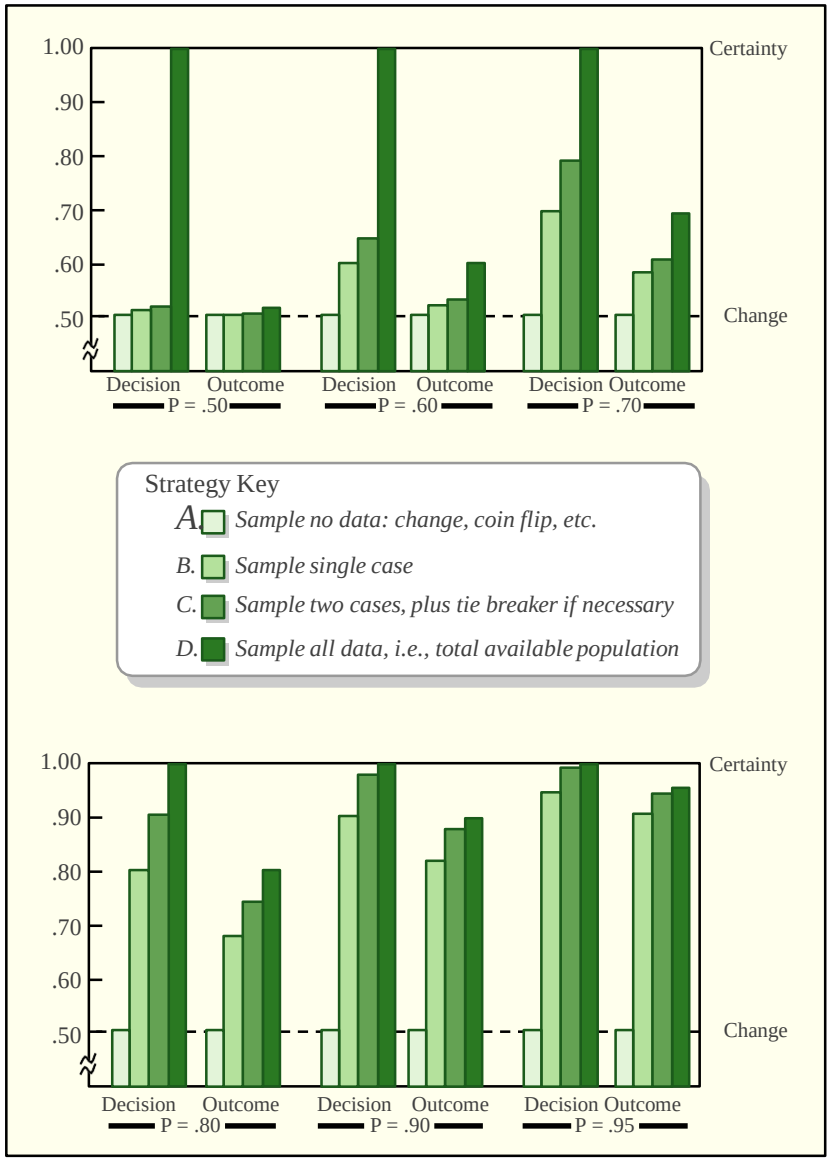
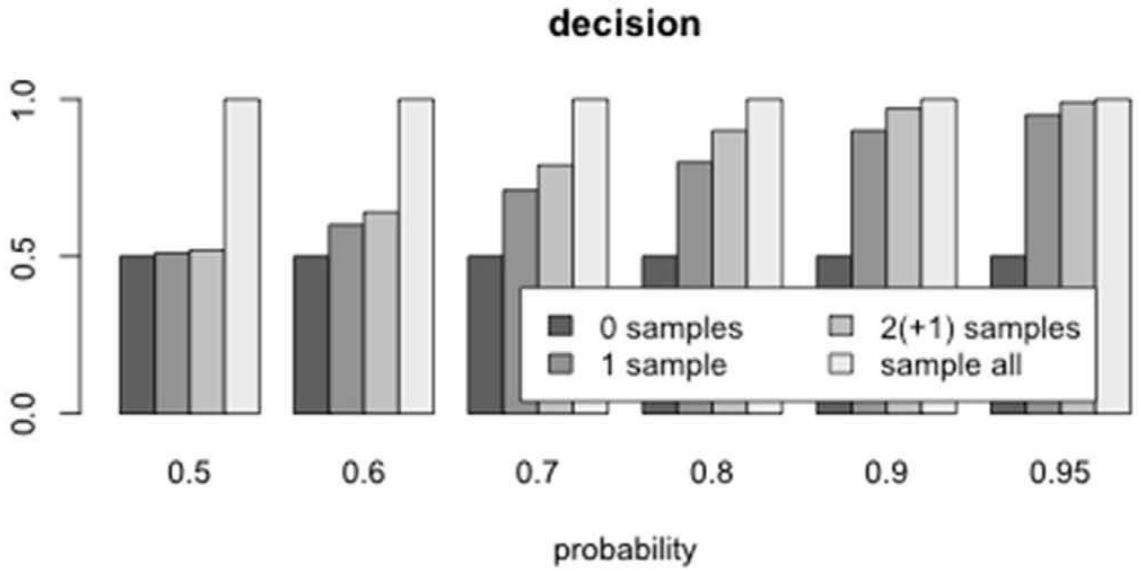
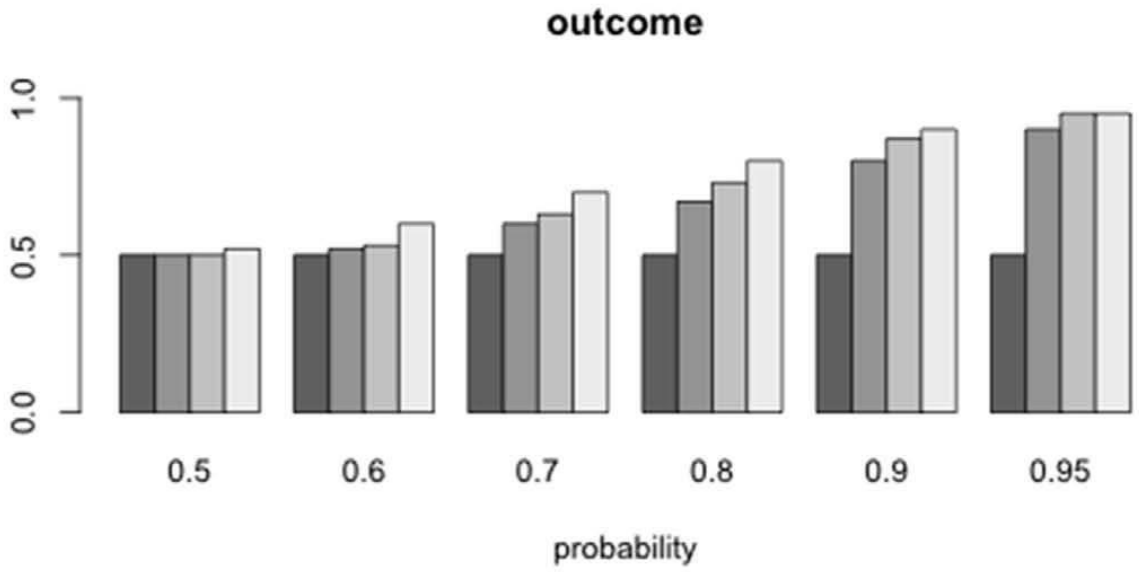
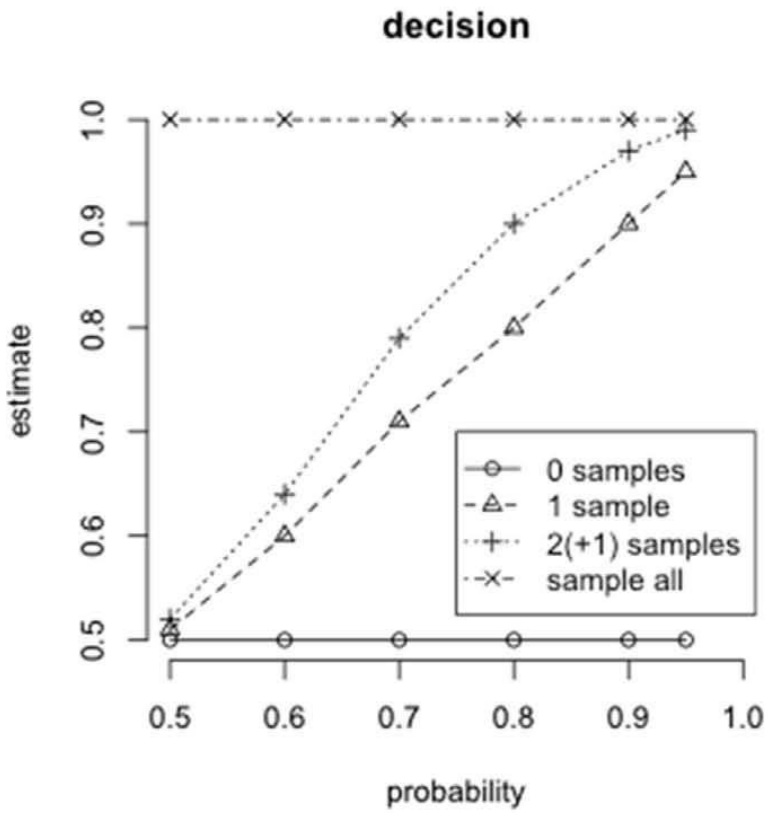


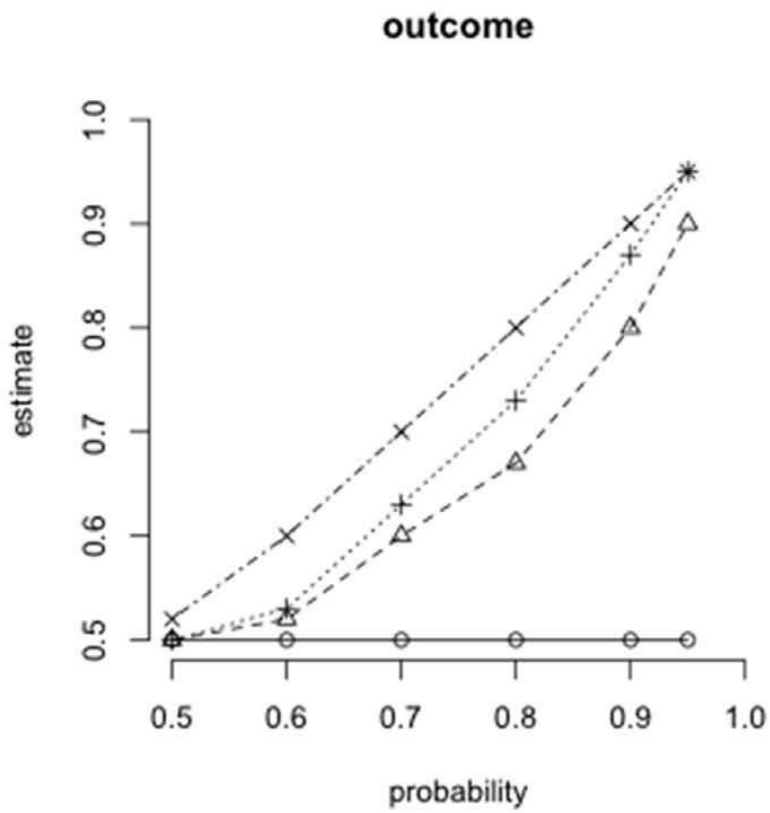
Figure by MIT OpenCourseWare.

Nisbett& Ross (1980) - ??

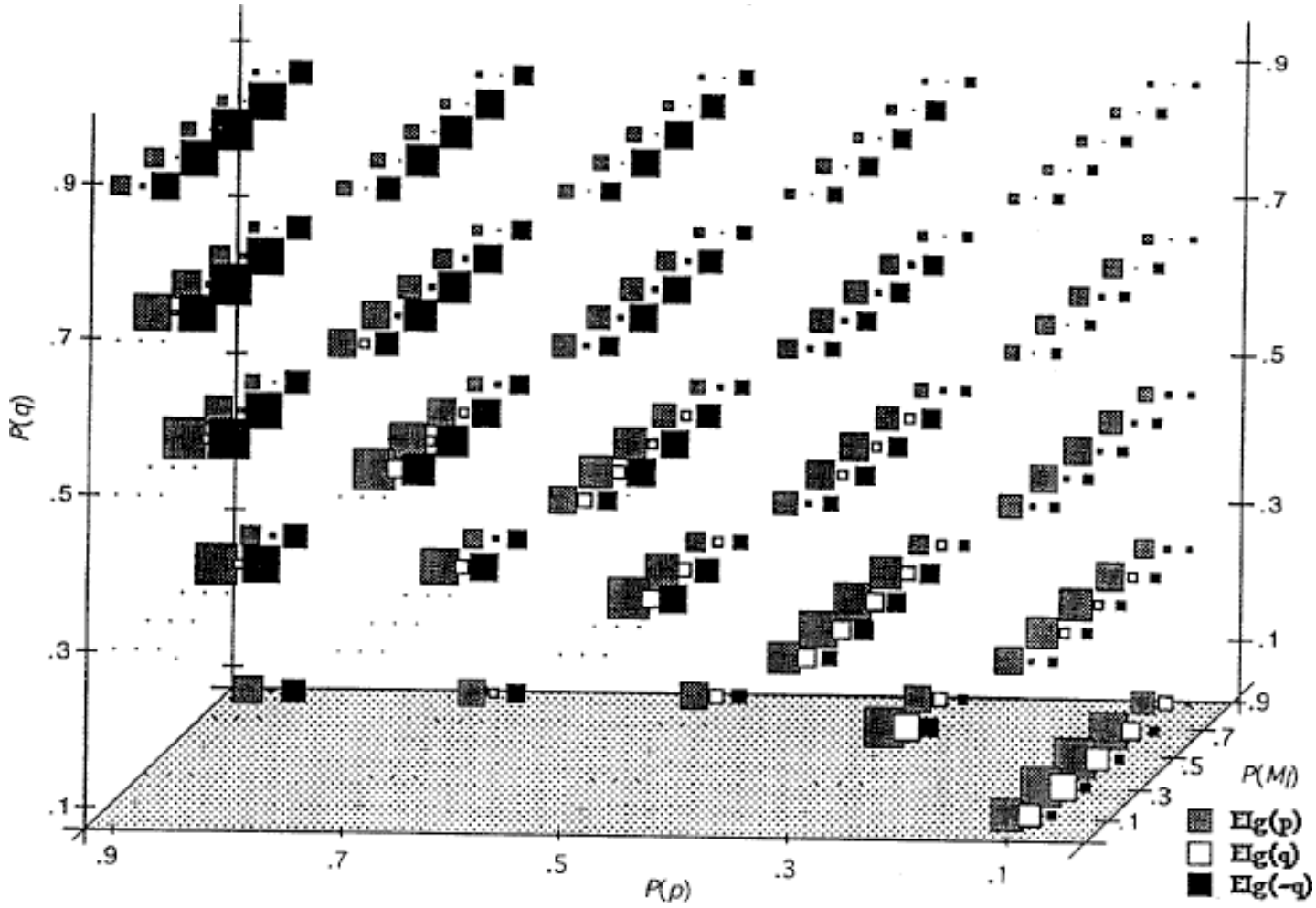

decision
1.0
0.5
0.0
0 samples
1 sample
2(+1) samples
sample all
0.5
0.6
0.7
0.8
0.9
0.95
probability


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

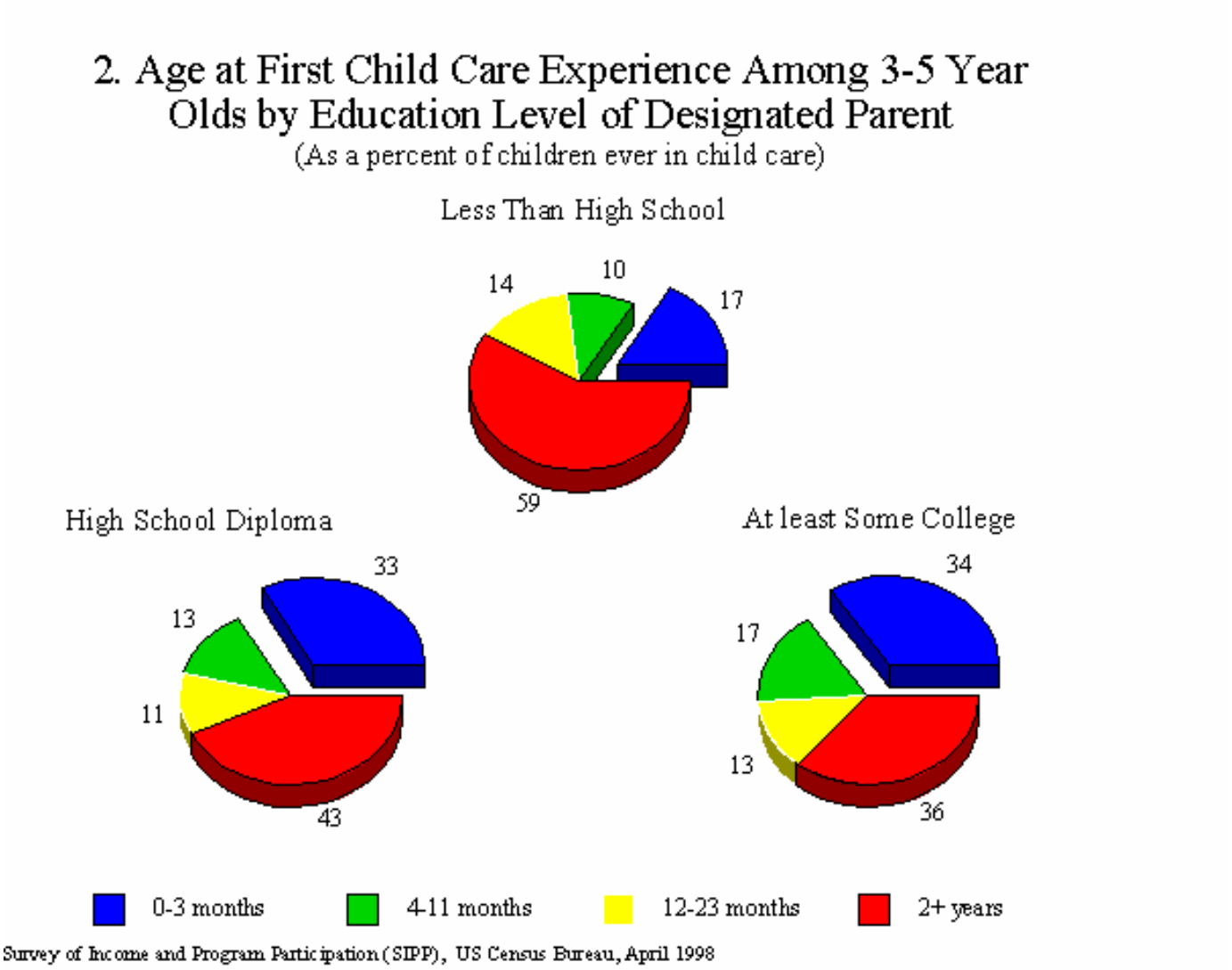

outcome
estimate
probability
0.5
0.6
0.7
0.8
0.9
1.0

Too much data for one plot



Oaksford & Chater, 1994

Difficulty of comparison



Which packages/functions

Standard charts (e.g. line chart, bar chart, scatter plot):

- Matplotlib, Pandas, Seaborn, ggplot, Altair, ...

Thematic maps

- Folium, Basemap, Cartopy, Iris, …

Other visualisations

- Bokeh (interactive plots), plotly, …

ggplot

Based on one of the most popular R package (ggplot2)

Based on the Grammar of Graphics (Wilkinson, 2005)

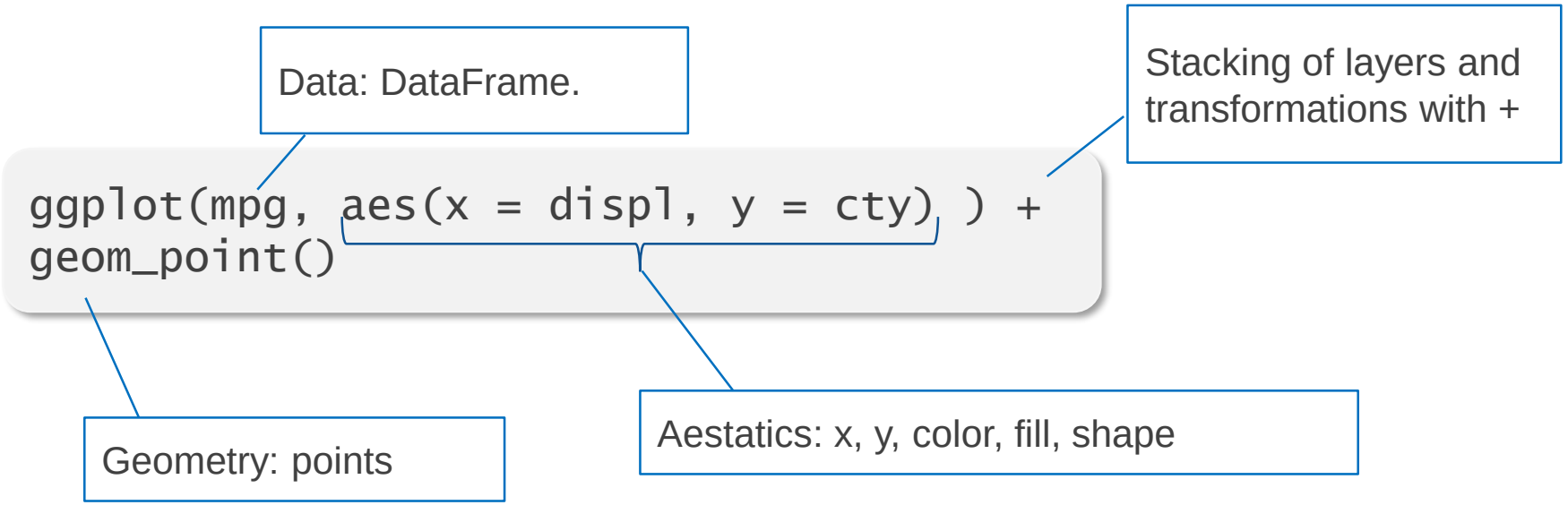
Charts are build up according to this grammar:

- data
- mapping / aestetics
- geoms
- stats
- scales
- coord
- Facets

Pandas DataFrames are used natively in ggplot.

ggplot and qplot

Shortcut function: qplot (quick plot):

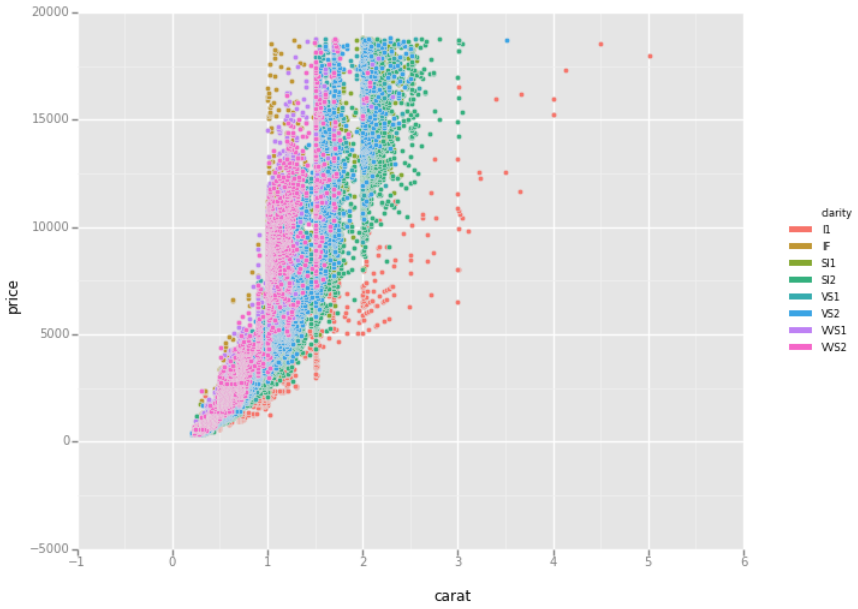



```
qplot(diamonds.carat, diamonds.price)
```

Aesthetics

Mapping of data to visual attributes of geometric objects:

- Position: x, y
- Color: color
- Shape: shape

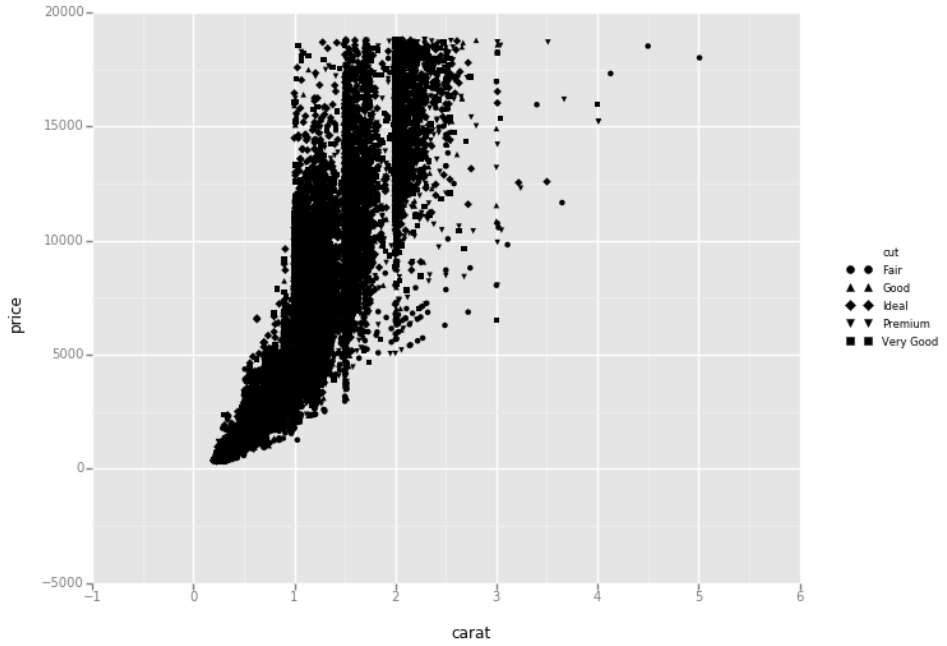



```
ggplot(aes(x='carat', y='price', color='clarity'), diamonds) +
geom_point()
```

Aesthetics

Mapping of data to visual attributes of geometric objects:

- Position: x,y
- Color: color
- Shape: shape




```
ggplot(aes(x='carat', y='price', shape="cut"), diamonds) +
geom_point()
```

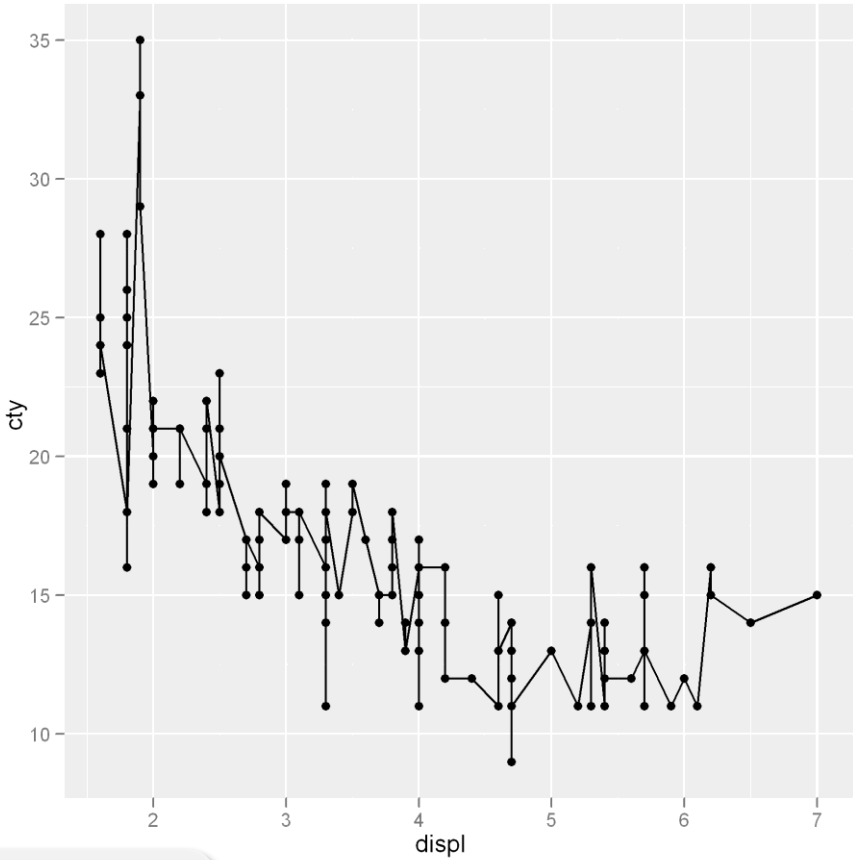
Geom

Geometric objects:

- Points, lines, polygons, …
- Functions start with “geom_”

Also margins:

- geom_errorbar(), geom_pointrange(), geo
- Note: they require the aesthetics ymin anc




```
ggplot(mpg, aes(x = displ, y = cty)) +
geom_point() + geom_line()
```

Stat

stat_smooth() and stat_density() enable statistical transformation

Most geoms have default stat (and the other way round)

geom and stat form a layer

One or more layers form a plot

stat_smooth

```
ggplot(aes(x='date', y='beef'), data=meat) + geom_point() + \
    stat_smooth(method='loess')
```

stat_density

```
ggplot(aes(x='price', color='clarity'), data=diamonds) + stat_density()
```

Scales (and axes)

A scale indicates how the value of a variable scales with an aesthetic

Therefore:

- A scale belongs to one aesthetic (x, y, color, fill, etc.)
- The axis is an essential part of a scale
- With scale_XXX, the scales and axes can be adjusted (XXX stands for the a combination of aesthetic and type of scale, e.g. scale_fill_gradient)

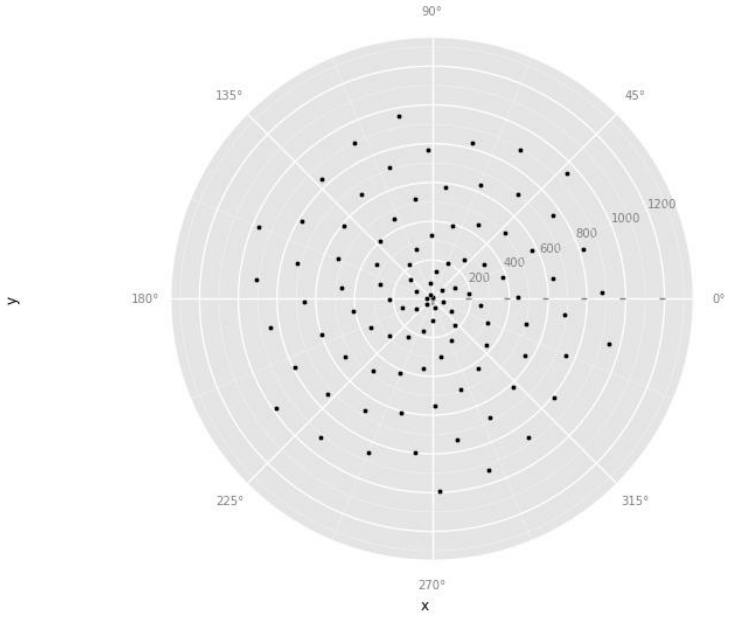
scale_x_log

```
ggplot(diamonds, aes(x='price')) + geom_histogram() + scale_x_log(base=100)
```

Coord

A chart is drawn in a coordinate system. This can be transformed.

A pie chart has a polar coordinate system.

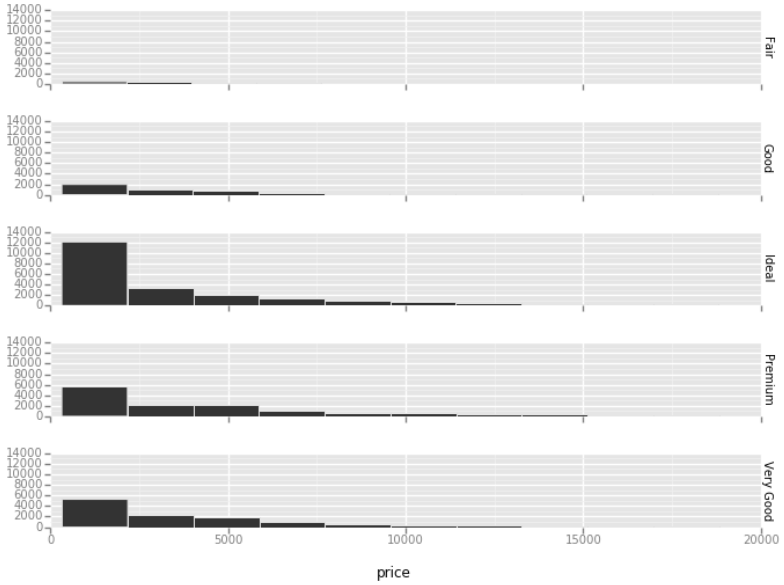



```
df = pd.DataFrame({"x": np.arange(100)})
df['y'] = df.x * 10  # polar coords
p = ggplot(df, aes(x='x', y='y')) + geom_point() + coord_polar()
print(p)
```

Facets

With facets, small multiples are created.

Each facet shows a subset of the data.

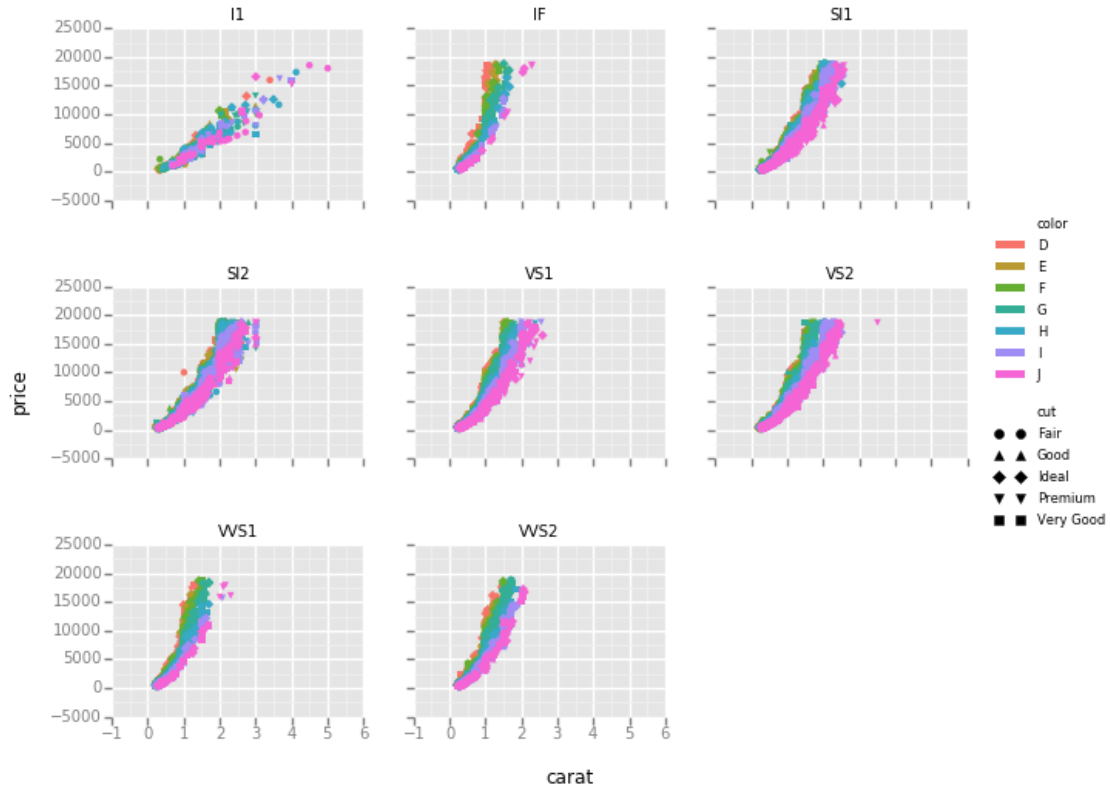



```
ggplot(diamonds,
aes(x='price')) + \
geom_histogram() + \
facet_grid("cut")
```

Facets example

```
ggplot(chopsticks, aes(x='chopstick_length',
y='food_pinching_effeciency')) + \
geom_point() + \
geom_line() + \
scale_x_continuous(breaks=[150, 250, 350]) + \
facet_wrap("individual")
```

Facets example 2




```
ggplot(diamonds, aes(x="carat", y="price", color="color",
shape="cut")) + geom_point() + facet_wrap("clarity")
```

ggplot tips

- You can annotate plots

```
ggplot(mtcars, aes(x='mpg')) + geom_histogram() + \
      xlab("Miles per Gallon") + ylab("# of Cars")
```

- Assign a plot to a variable, for instance g:

```
g = ggplot(mpg, aes(x = displ, y = cty)) +
geom_point()
```

- The function save saves the plot to the desired format:

```
g.save(“myimage.png”)
```

Folium: Thematic maps

A thematic map is a visualization where statistical information with a spatial component is shown.

Other libraries are: Basemap, Cartopy, Iris

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library.

Manipulate your data in Python, then visualize it in on a Leaflet map via Folium.

Folium features

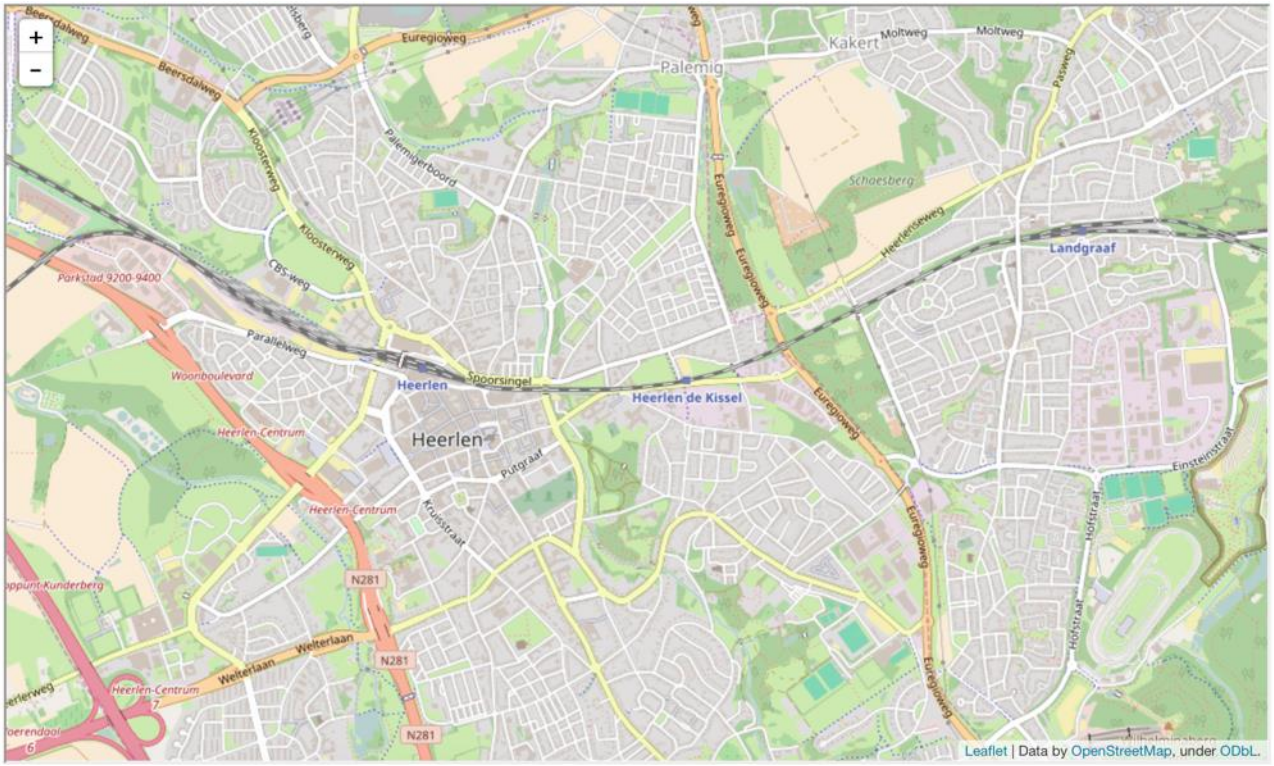
Built-in tilesets from OpenStreetMap, MapQuest Open, MapQuest Open Aerial, Mapbox, and Stamen

- Supports custom tilesets with Mapbox or Cloudmade API keys.

Supports GeoJSON and TopoJSON overlays,

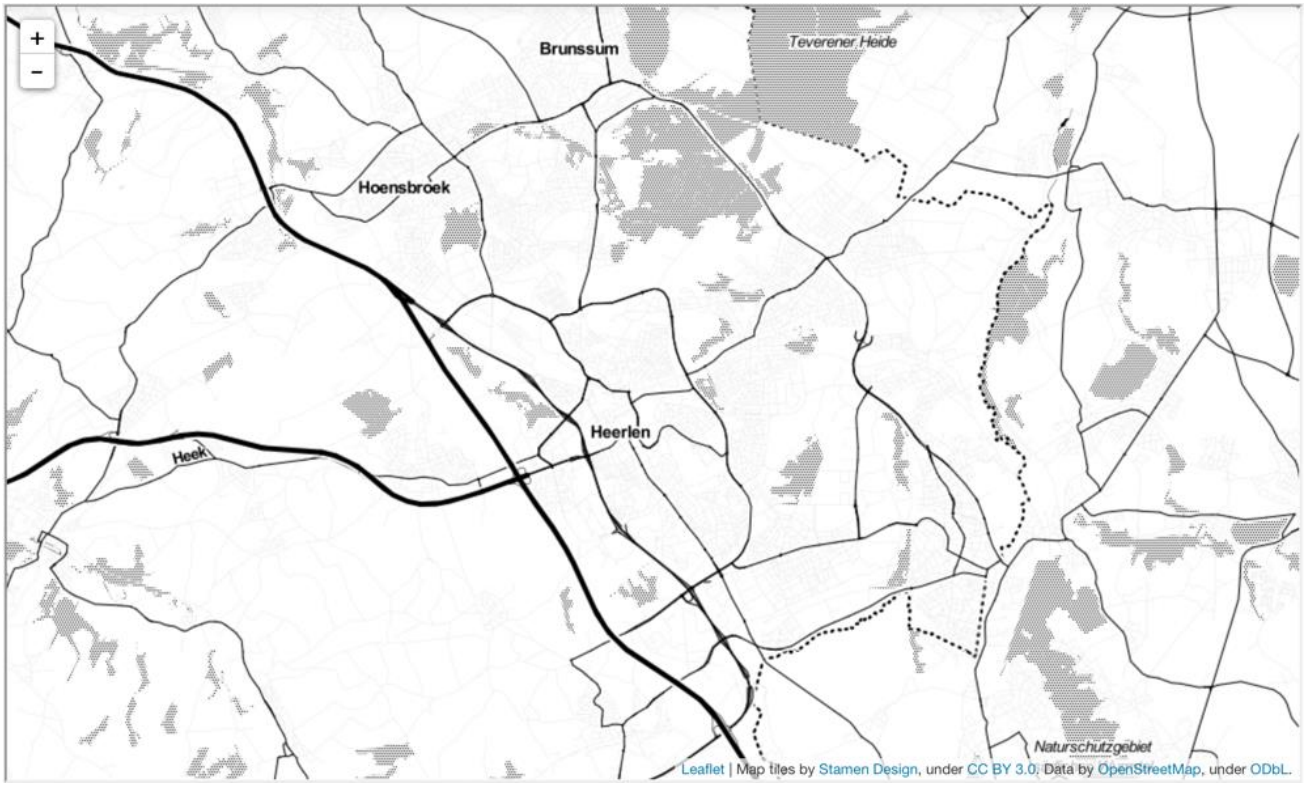
- as well as the binding of data to those overlays to create choropleth maps with color-brewer color schemes.

Basic Maps




```
folium.Map(location=[50.89, 5.99], zoom_start=14)
```

Basic maps




```
folium.Map(location=[50.89, 5.99], zoom_start=14, tiles='Stamen Toner')
```

GeoJSON/TopoJSON Overlays

```
ice_map = folium.Map(location=[-59, -11], tiles='Mapbox Bright', zoom_start=2)
ice_map.geo_json(geo_path=geo_path)
ice_map.geo_json(geo_path=topo_path, topojson='objects.antarctic_ice_shelf')
ice_map.create_map(path='ice_map.html')
```

Choropleth maps

```
map = folium.Map(location=[48, -102], zoom_start=3)
map.choropleth(geo_path=state_geo, data=state_data,
               columns=['State', 'Unemployment'], key_on='feature.id',
               fill_color='YlGn', fill_opacity=0.7, line_opacity=0.2,
               legend_name='Unemployment Rate (%)')
```

Summary

Python has many options for data visualization

Each visualisation library has a particular audience

Javascript backend is mostly used to extend power of the visualisation

Python's extensive data processing tools integrates well with visualisation requirements

References

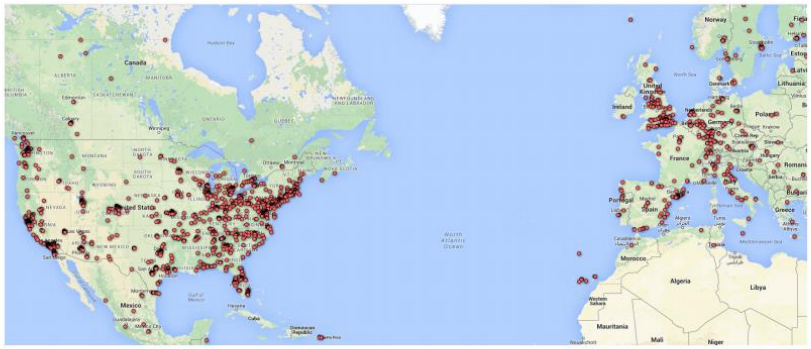
http://yhat.github.io/ggplot/

https://folium.readthedocs.io/en/latest/

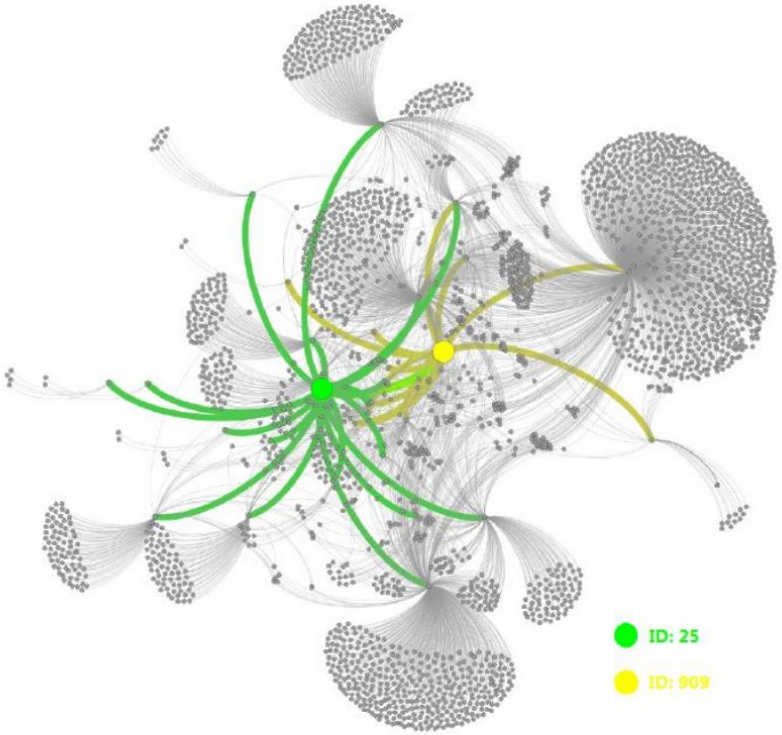
Use Cases

Research data visualization from my group

Yao, L., Sheng, Q.Z., Wang, X., Zhang, W.E. and Qin, Y., 2018. Collaborative location recommendation by integrating multi-dimensional contextual information. *ACM Transactions on Internet Technology (TOIT)*, *18*(3), p.32.



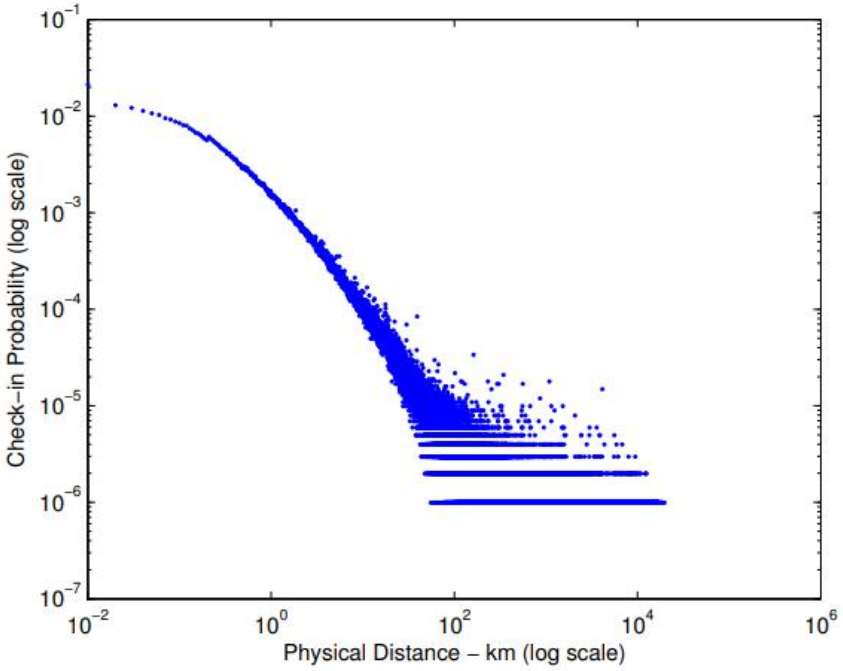
(a)



(b)

Fig. 3. (a) Check-in distribution over the processed dataset of BrightKite; (b) User friendships over the processed BrightKite dataset, each node represents a user. For example, the links of user id = 25 and id = 909 are highlighted in green and yellow respectively

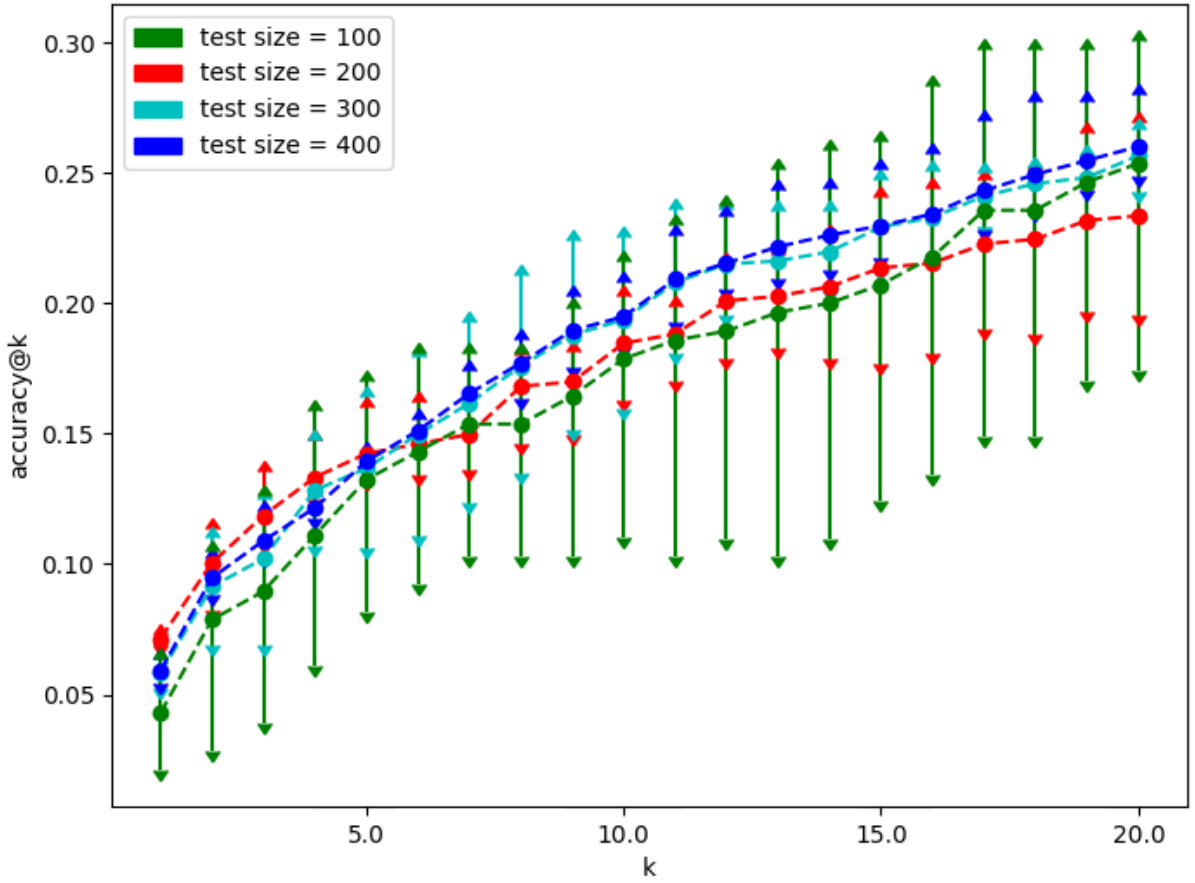
Yao, L., Sheng, Q.Z., Wang, X., Zhang, W.E. and Qin, Y., 2018. Collaborative location recommendation by integrating multi-dimensional contextual information. *ACM Transactions on Internet Technology (TOIT)*, *18*(3), p.32.



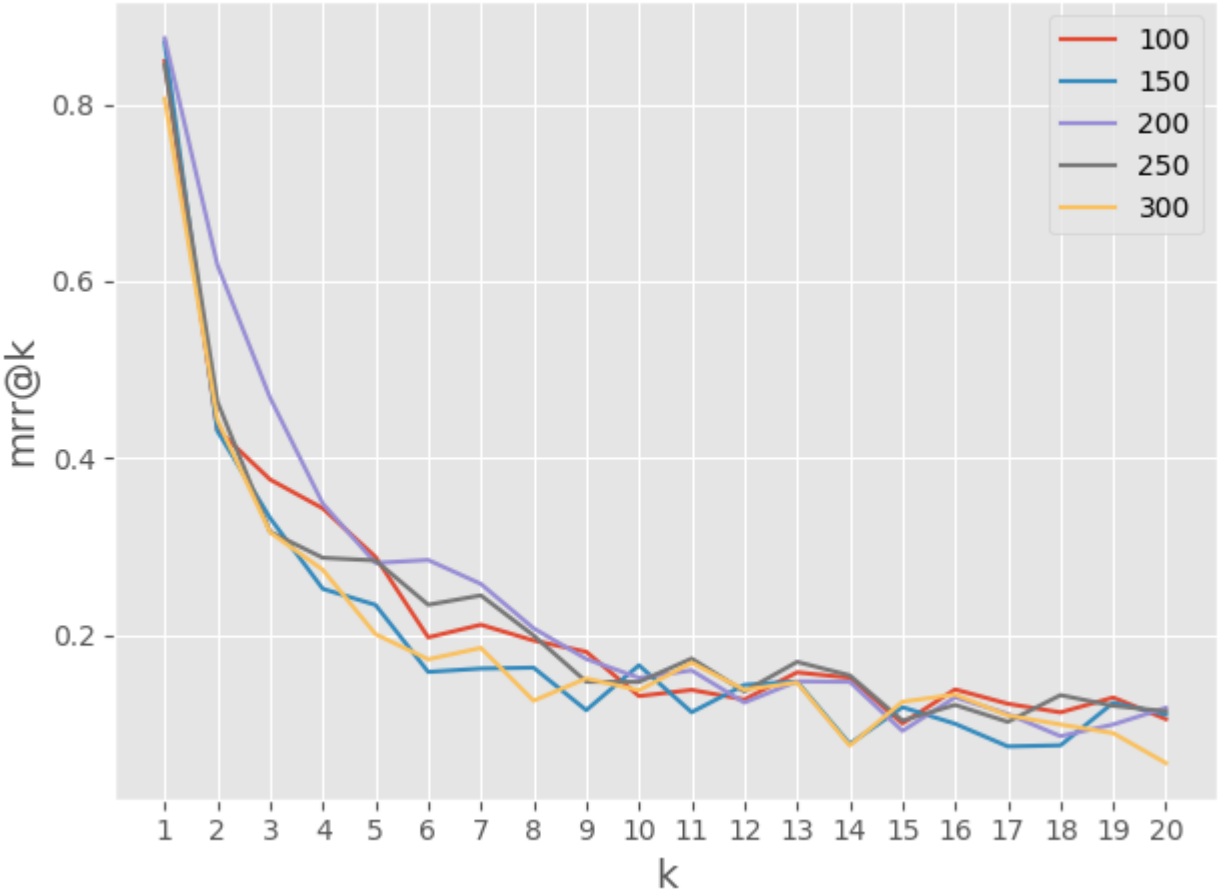
(a)

(b)

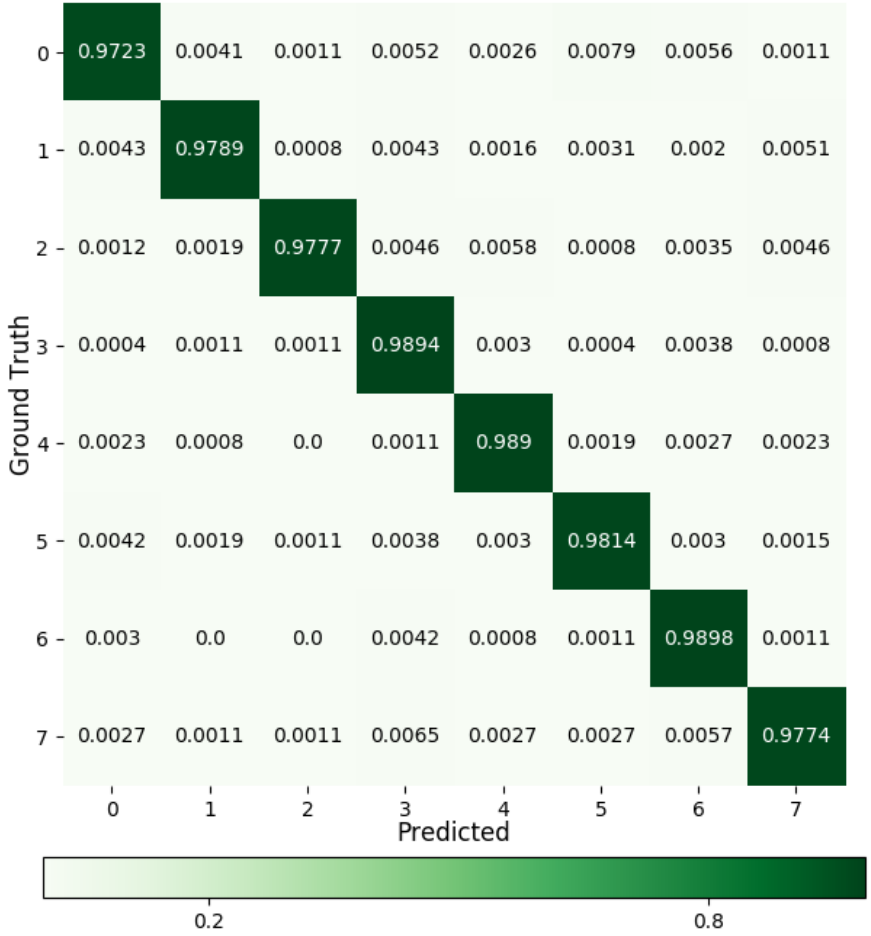
Huang, C., Yao, L., Wang, X., Benatallah, B., & Sheng, Q. Z. (2017, June). **Expert as a service: Software expert recommendation via knowledge domain embeddings in stack overflow**. In *2017 IEEE International Conference on Web*

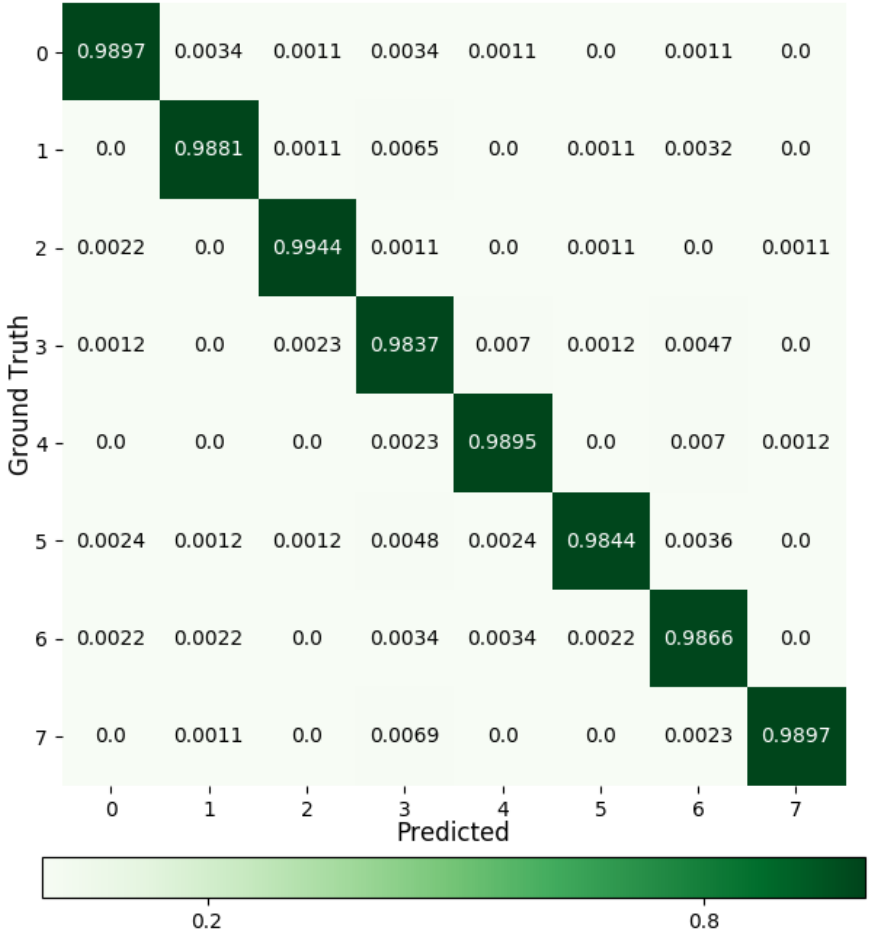


Huang, C., Yao, L., Wang, X., Benatallah, B., Zhang, S., & Dong, M. (2018, November). **Expert recommendation via tensor factorization with regularizing hierarchical topical relationships**. In *International Conference on Service-Oriented Computing* (pp. 373-387). Springer, Cham.

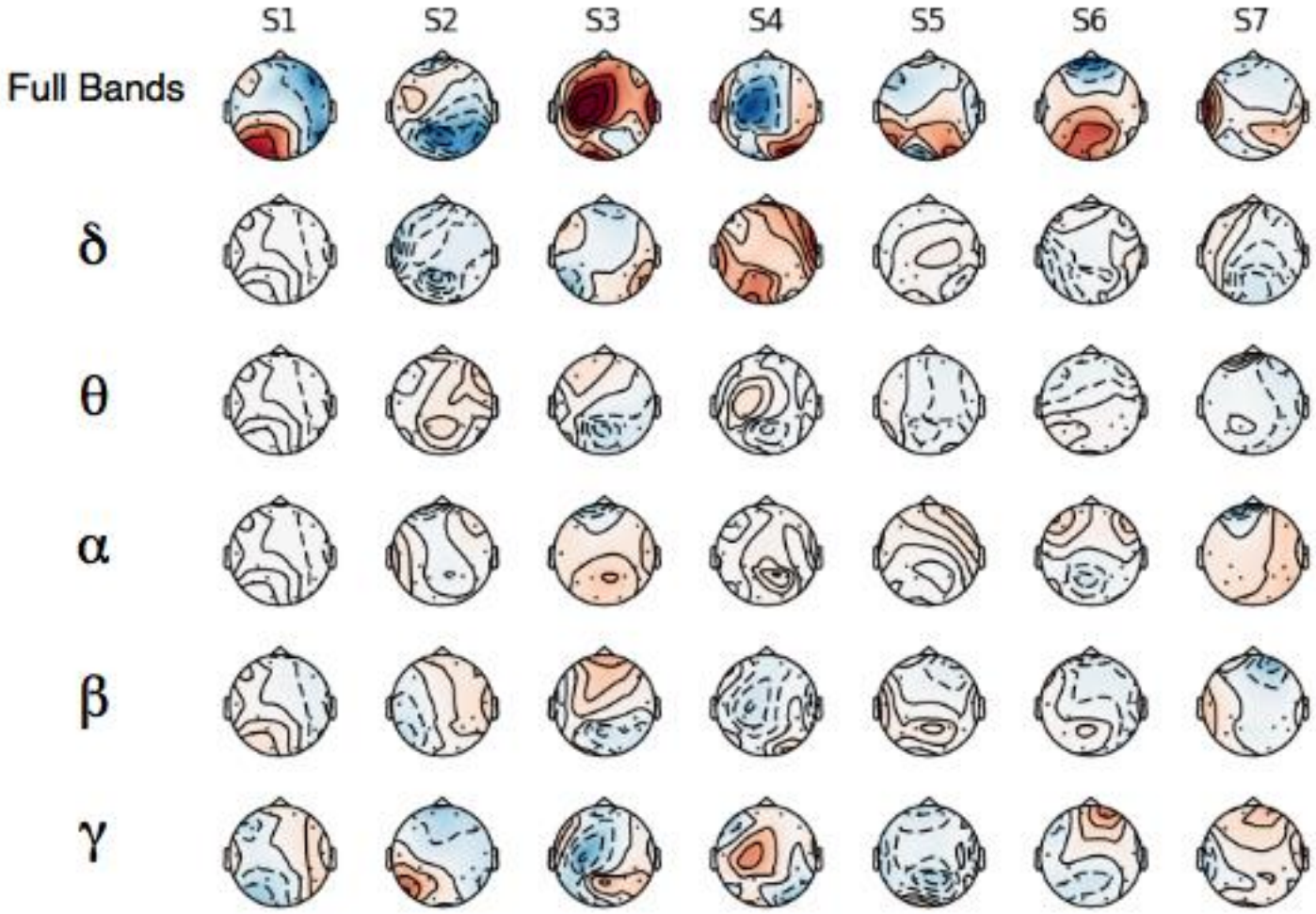


Zhang, X., Yao, L., Kanhere, S. S., Liu, Y., Gu, T., & Chen, K. (2018). MindID: Person Identification from Brain Waves through Attention-based Recurrent Neural Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 149.

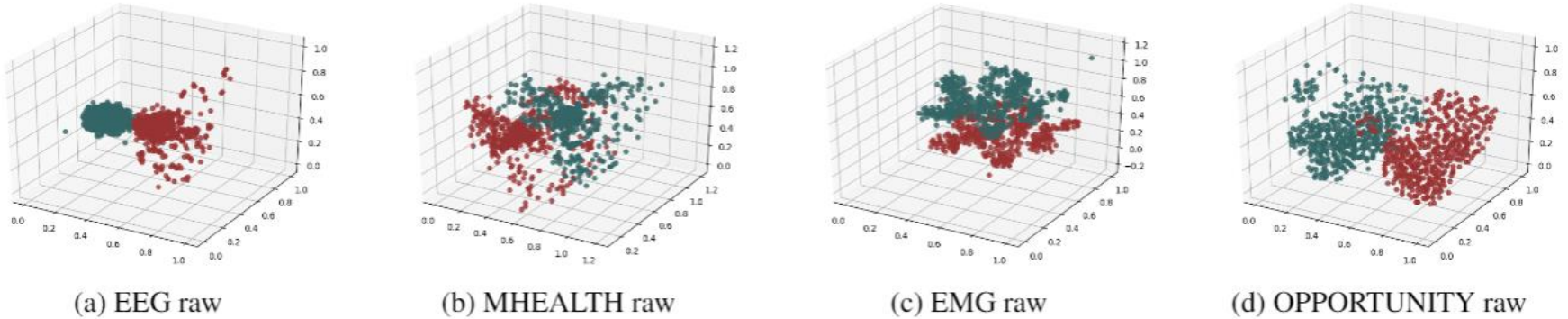




Zhang, X., Yao, L., Chen, K., Wang, X., Sheng, Q., & Gu, T. (2017**). DeepKey: An EEG and Gait Based Dual-Authentication System.** *arXiv preprint arXiv:1706.01606*. (Submitted to Ubicomp 2019)



Chen, K., Yao, L., Zhang, D., Chang, X., Long, G., & Wang, S. (2018). **Distributionally Robust Semi-Supervised Learning for People-Centric Sensing**.
Accepted. *arXiv preprint arXiv:1811.05299*.

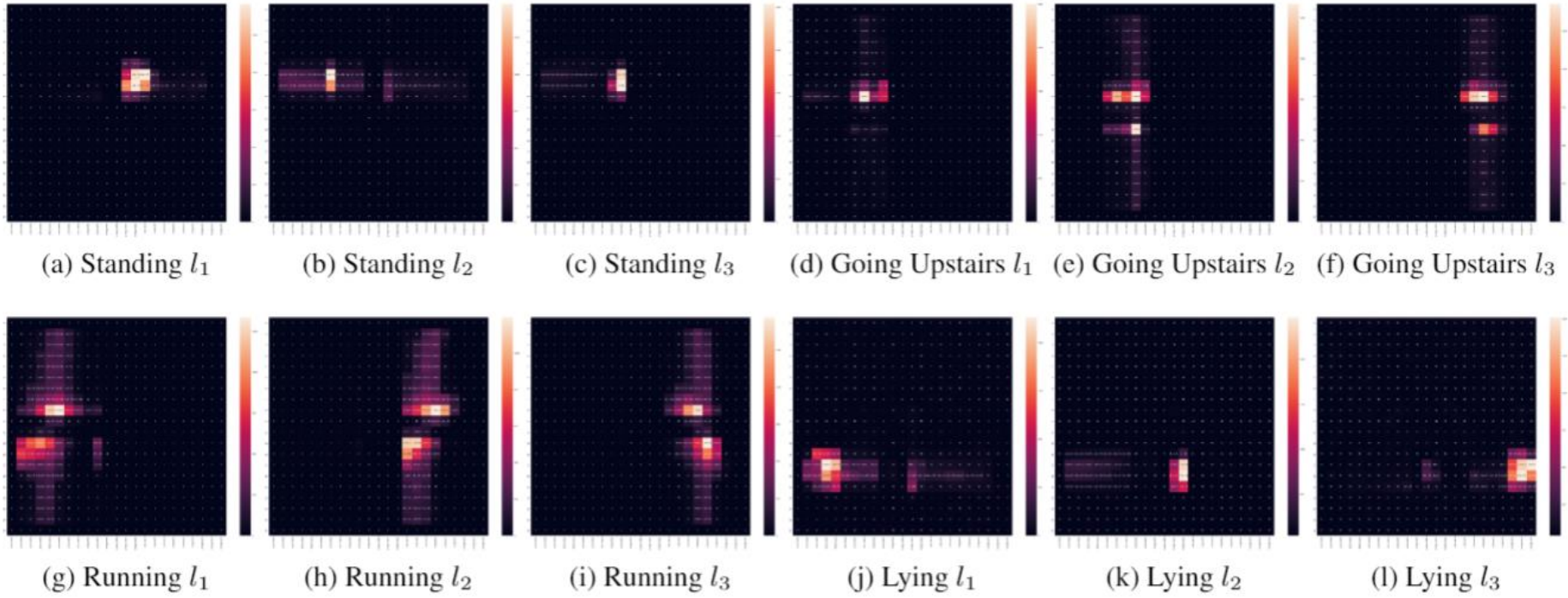


(a) EEG raw (b) MHEALTH raw (c) EMG raw (d) OPPORTUNITY raw

(e) EEG features (f) MHEALTH features (g) EMG features (h) OPPORTUNITY features

Figure 3: Visualization of Latent Features. Green points correspond to the labeled data or features, while the red points correspond to the unlabeled data or features. In all cases, our model is effective in reducing distribution discrepancy.

Kaixuan Chen and Lina Yao et al, **Multi-agent Attentional Activity Recognition** Submitted IJCAI 2019

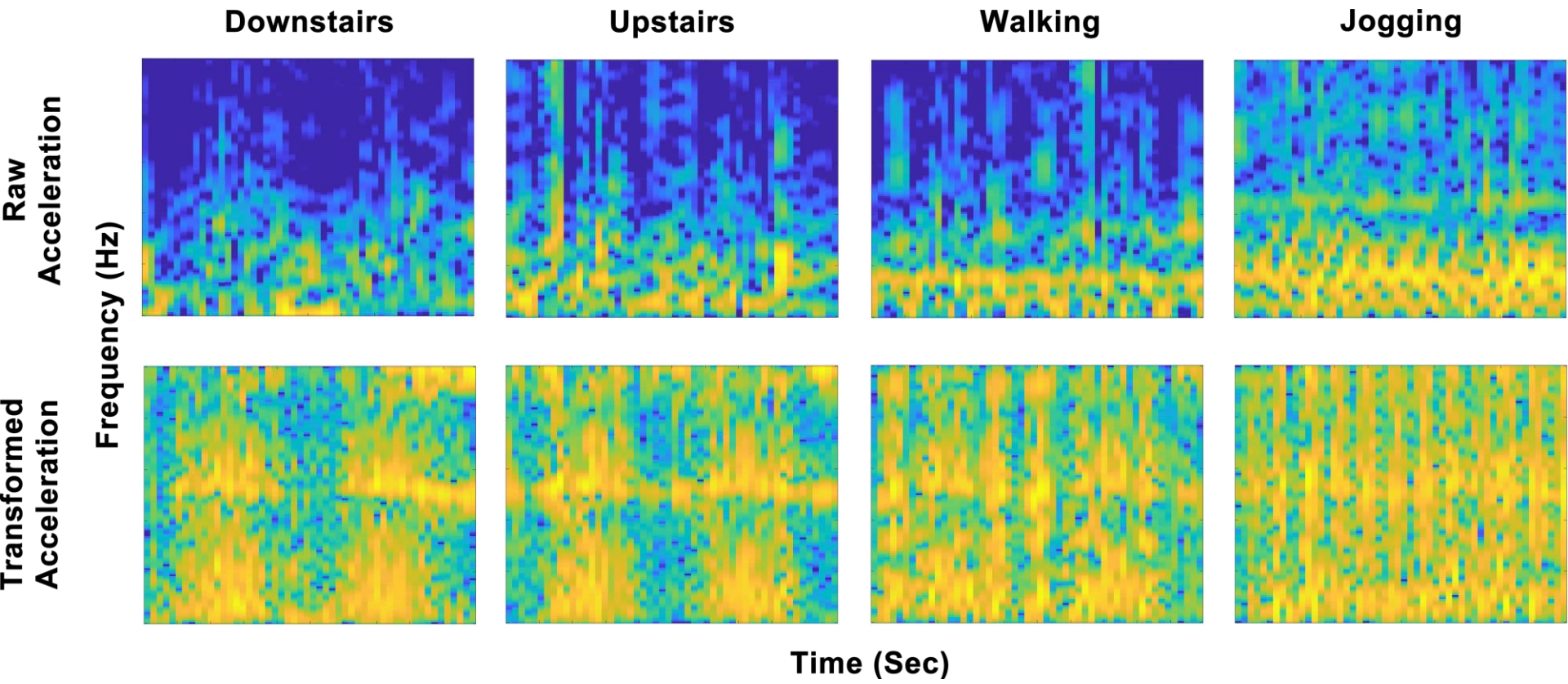


(a) Standing l_1 (b) Standing l_2 (c) Standing l_3 (d) Going Upstairs l_1 (e) Going Upstairs l_2 (f) Going Upstairs l_3

(g) Running l_1 (h) Running l_2 (i) Running l_3 (j) Lying l_1 (k) Lying l_2 (l) Lying l_3

Figure 2: Visualization of the selected modalities and time on MHEALTH. The input matrices' size is 20×23, where 20 is the length of the time window and 23 is the number of modalities. Thus each grid denotes an input feature, and the values in the grids represent the frequency with which this feature is selected. Lighter colors denote higher frequency. To be clear, detailed illustration is provided in Table 3.

Dalin Zhang and Lina Yao et al. **Collective Privacy Protection: Preventing Sensitive Inferences via Integrative Transformation**
submitted ***IJCAI 2019***



“I didn't have time to write a short letter, so I wrote a long one instead.”

—Mark Twain