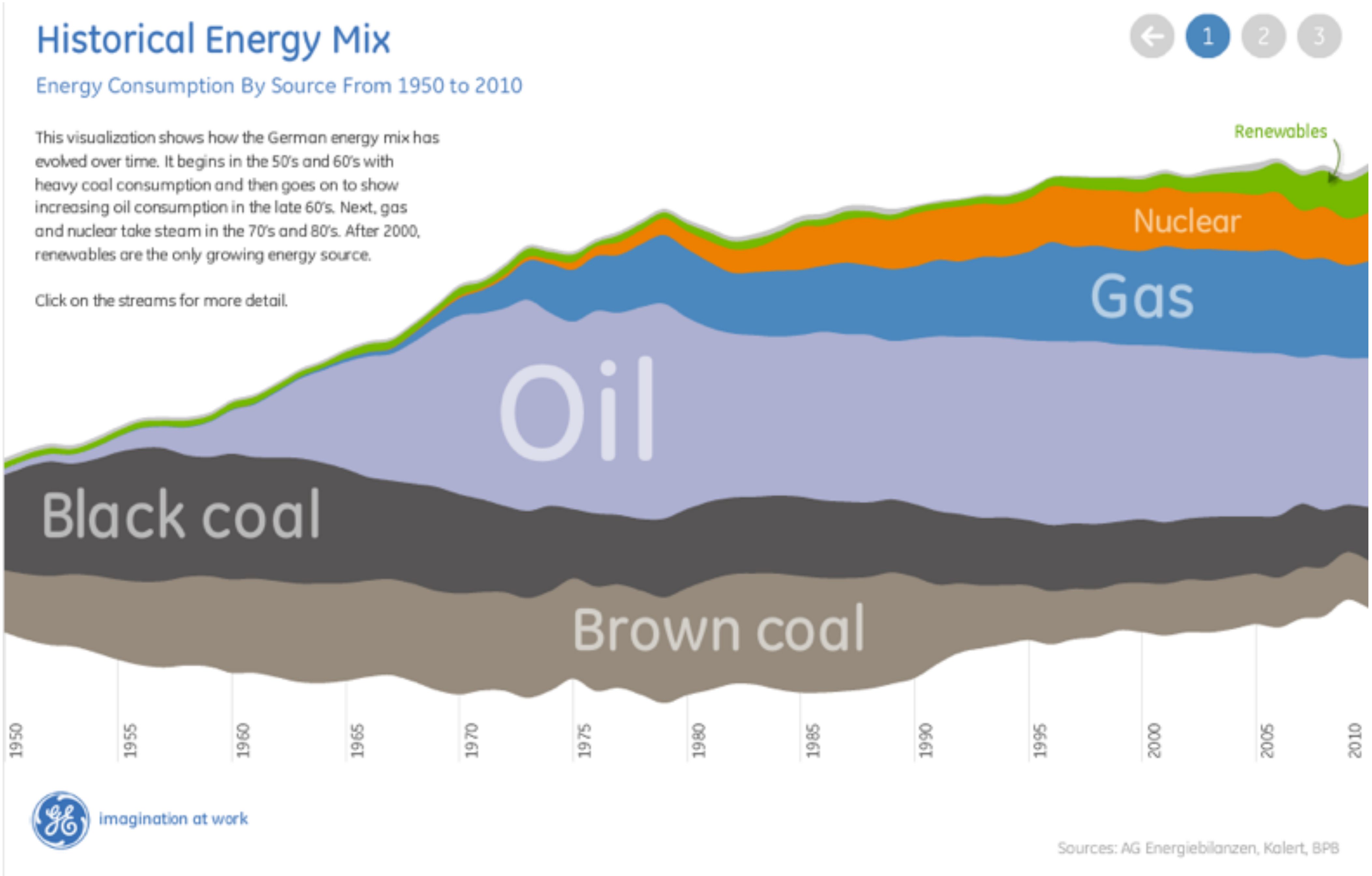
Visualization and GGPLOT



Historical Energy Mix

Energy Consumption By Source From 1950 to 2010

This visualization shows how the German energy mix has evolved over time. It begins in the 50's and 60's with heavy coal consumption and then goes on to show increasing oil consumption in the late 60's. Next, gas and nuclear take steam in the 70's and 80's. After 2000, renewables are the only growing energy source.

Click on the streams for more detail.

Sources: AG Energiebilanzen, Kolert, BPB

Strategic use of well designed graphics is essential for communication

http://www.theusrus.de/Blog-files/GermanEnergy.png

WHY!

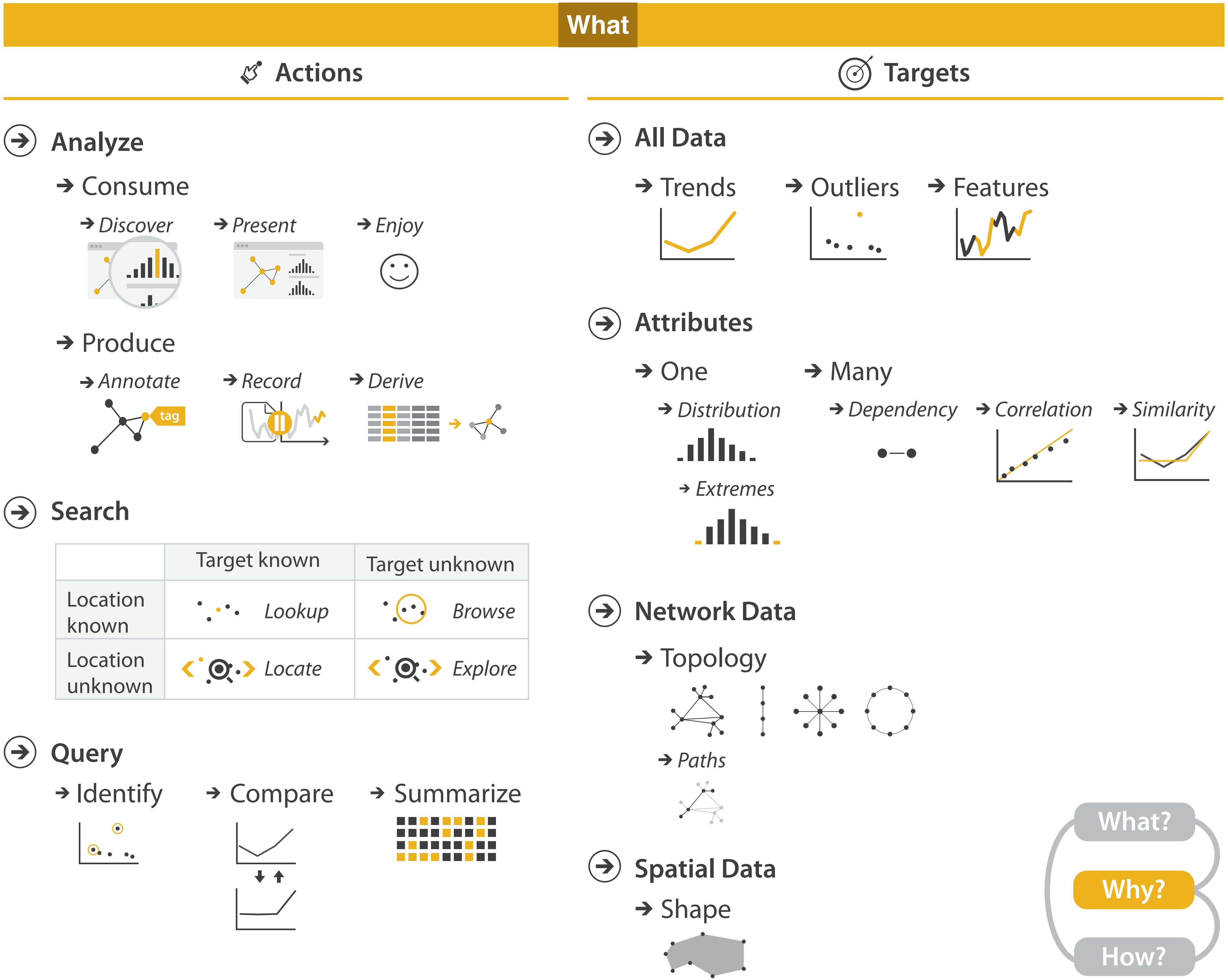
- Before you embark on visualization - know what your goal is!
  - communicate X to an audience
  - help you-your group understand something

Communicate to an audience

- Write down what you want to communicate
- Think through what your audience needs to know to understand (background, terminology…)

Understanding (for you!)

- Exploratory data analysis - PLAY - try different things
- Documentation - think of yourself as an audience



What

Actions

Analyze

- Consume
  - Discover
  - Present
  - Enjoy
- Produce
  - Annotate
  - Record
  - Derive

Search

	Target known	Target unknown
Location known	Lookup	Browse
Location unknown	Locate	Explore

Query

- Identify
- Compare
- Summarize

Targets

All Data

- Trends
- Outliers
- Features

Attributes

- One
  - Distribution
    - Extremes
- Many
  - Dependency
  - Correlation
  - Similarity

Network Data

- Topology
  - Paths

Spatial Data

- Shape

What?

Why?

How?

- {action, target} pairs
  - – *discover distribution*
  - – *compare trends*
  - – *locate outliers*
  - – *browse topology*

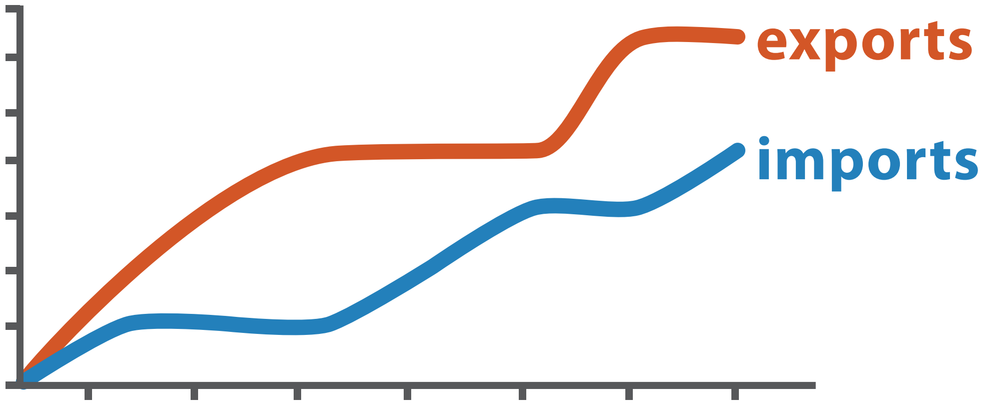
Munzer, T. (University of British Columbia, Canada;
IEEE VIS 2015 TutorialOctober 2015, Chicago IL

Don’t just use the data ‘as is’

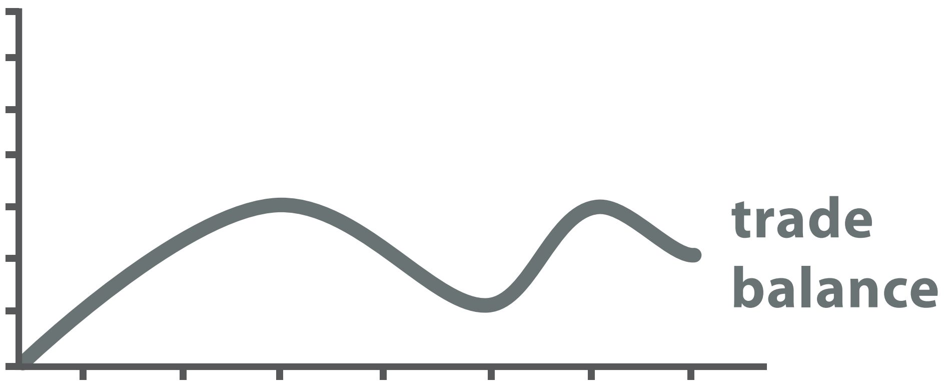
– think about how to achieve your goal - what makes it easiest to see

– create it with a series of transformations from the original dataset

– draw that



Original Data

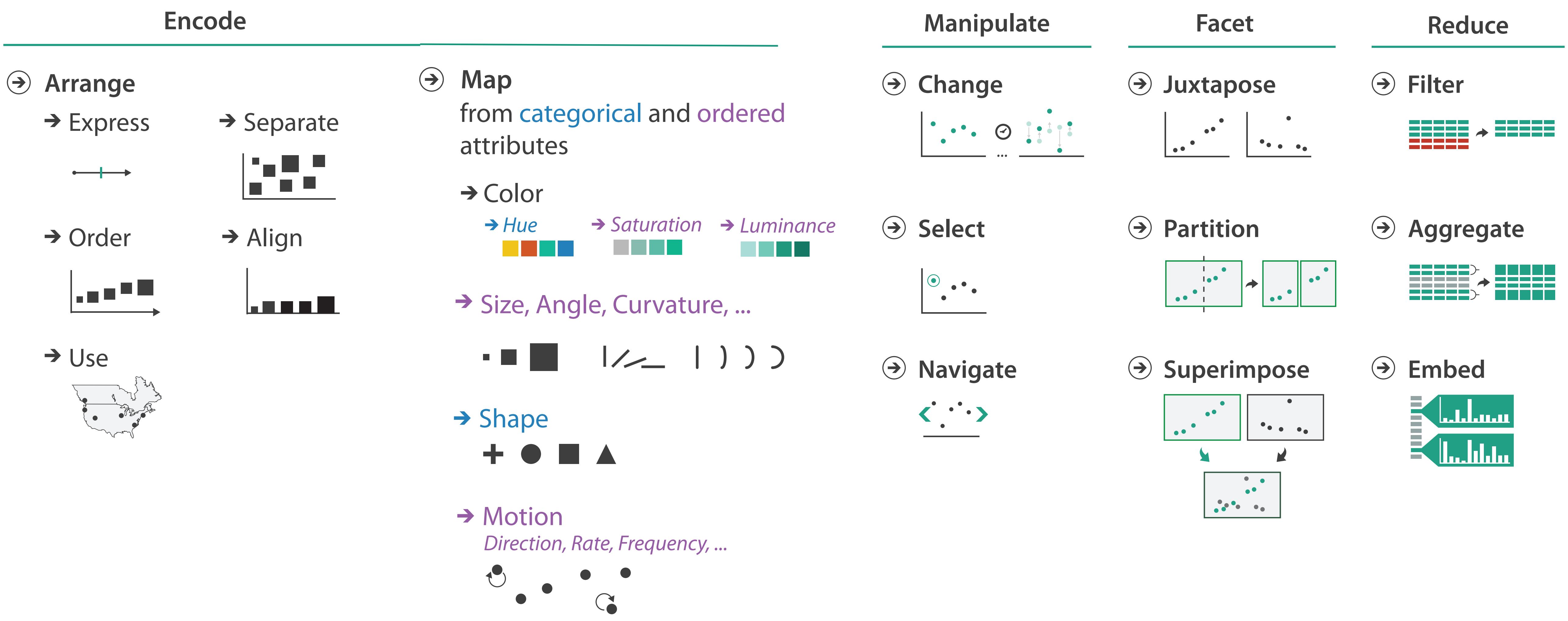


trade balance = exports – imports

Derived Data

Munzer, T. (University of British Columbia, Canada; IEEE VIS 2015 TutorialOctober 2015, Chicago IL

How?



Encode

Arrange

- → Express
- → Separate
- → Order
- → Align
- → Use

Map

from categorical and ordered attributes

- → Color
  - → *Hue*
  - → *Saturation*
  - → *Luminance*
- → Size, Angle, Curvature, ...
- → Shape
- → Motion

Direction, Rate, Frequency, ...

Manipulate

Change

Select

Navigate

Facet

Juxtapose

Partition

Superimpose

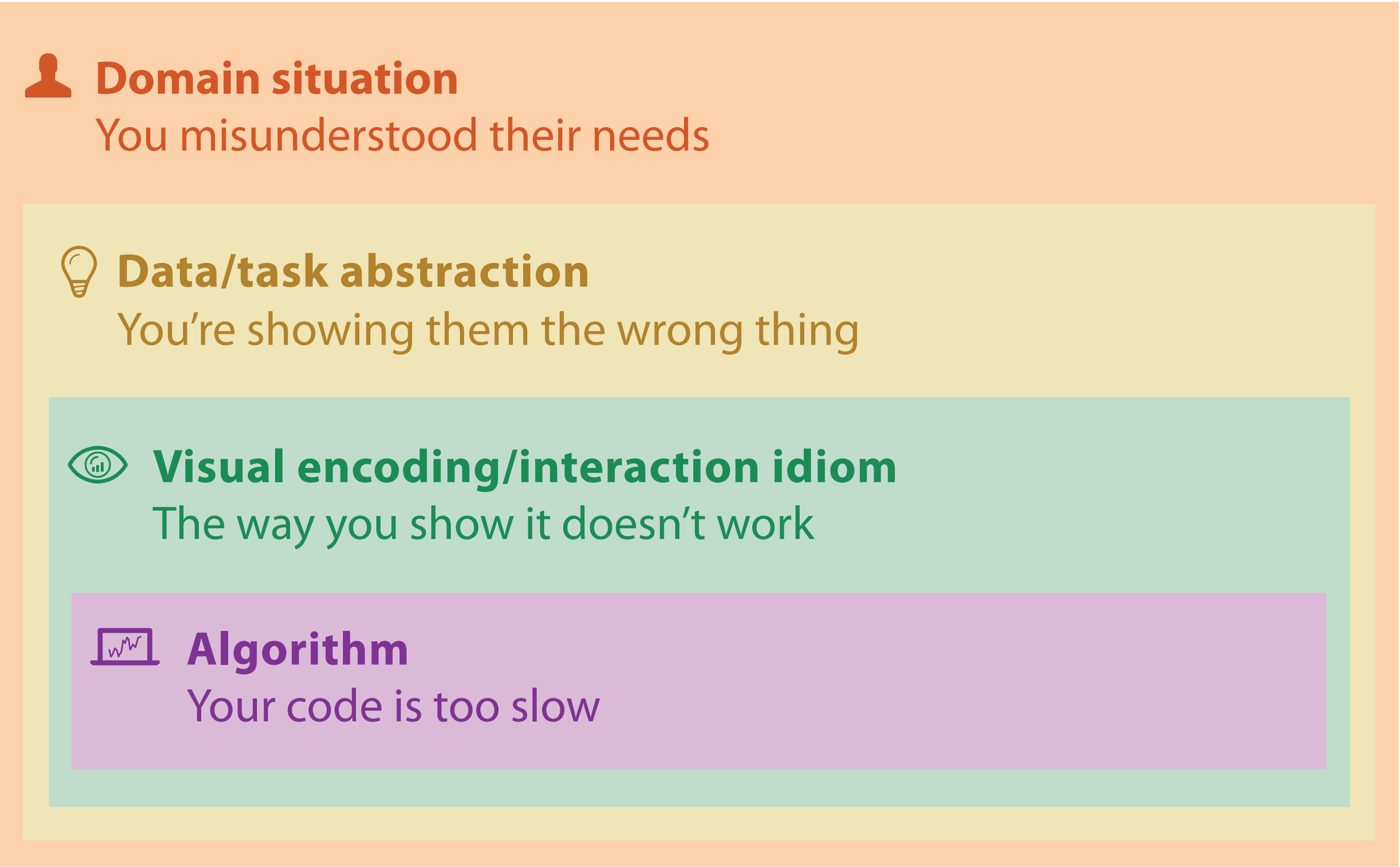
Reduce

Filter

Aggregate

Embed

- Common Problems in Visualization



Domain situation
You misunderstood their needs

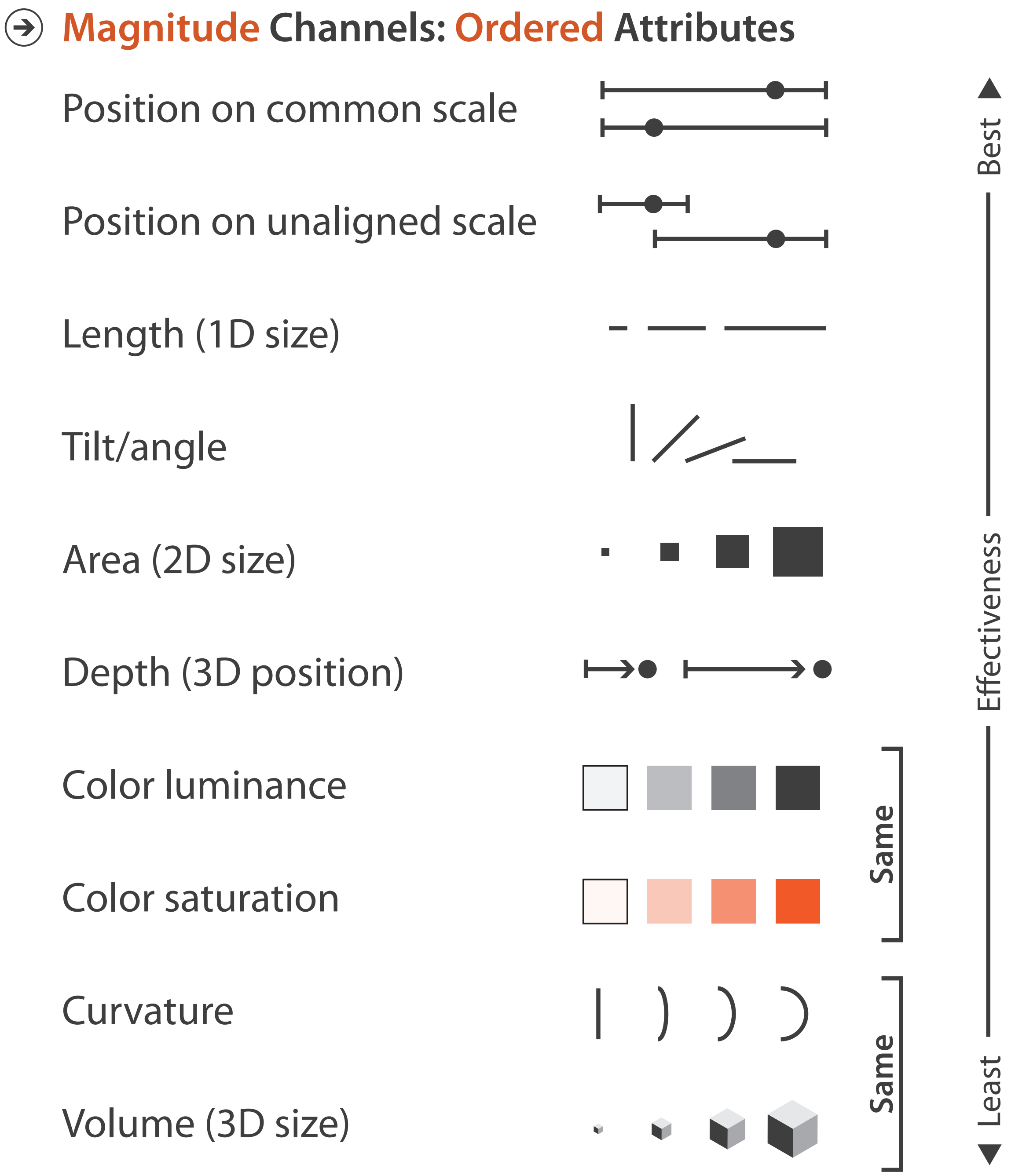
Data/task abstraction
You're showing them the wrong thing

Visual encoding/interaction idiom
The way you show it doesn't work

Algorithm
Your code is too slow

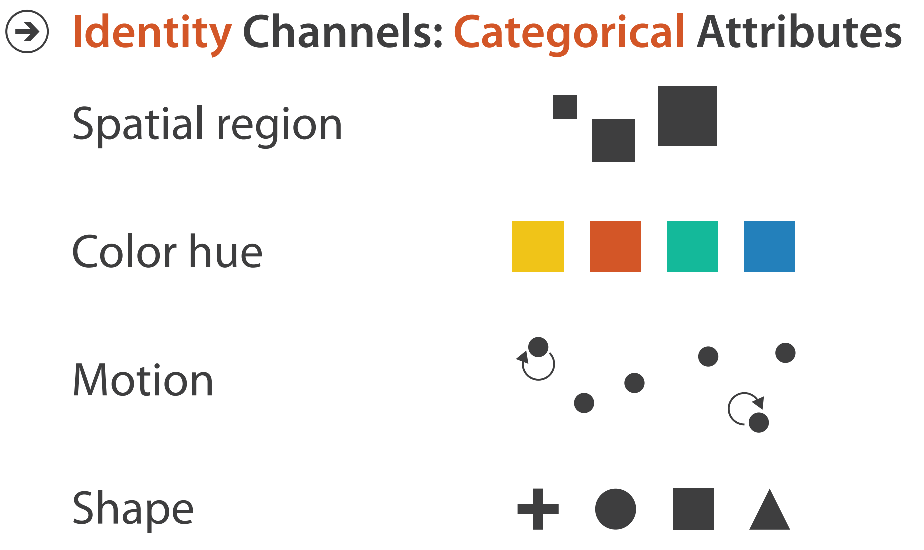
Munzer, T. (University of British Columbia, Canada;
IEEE VIS 2015 TutorialOctober 2015, Chicago IL

How - channels…



Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)



Identity Channels: Categorical Attributes

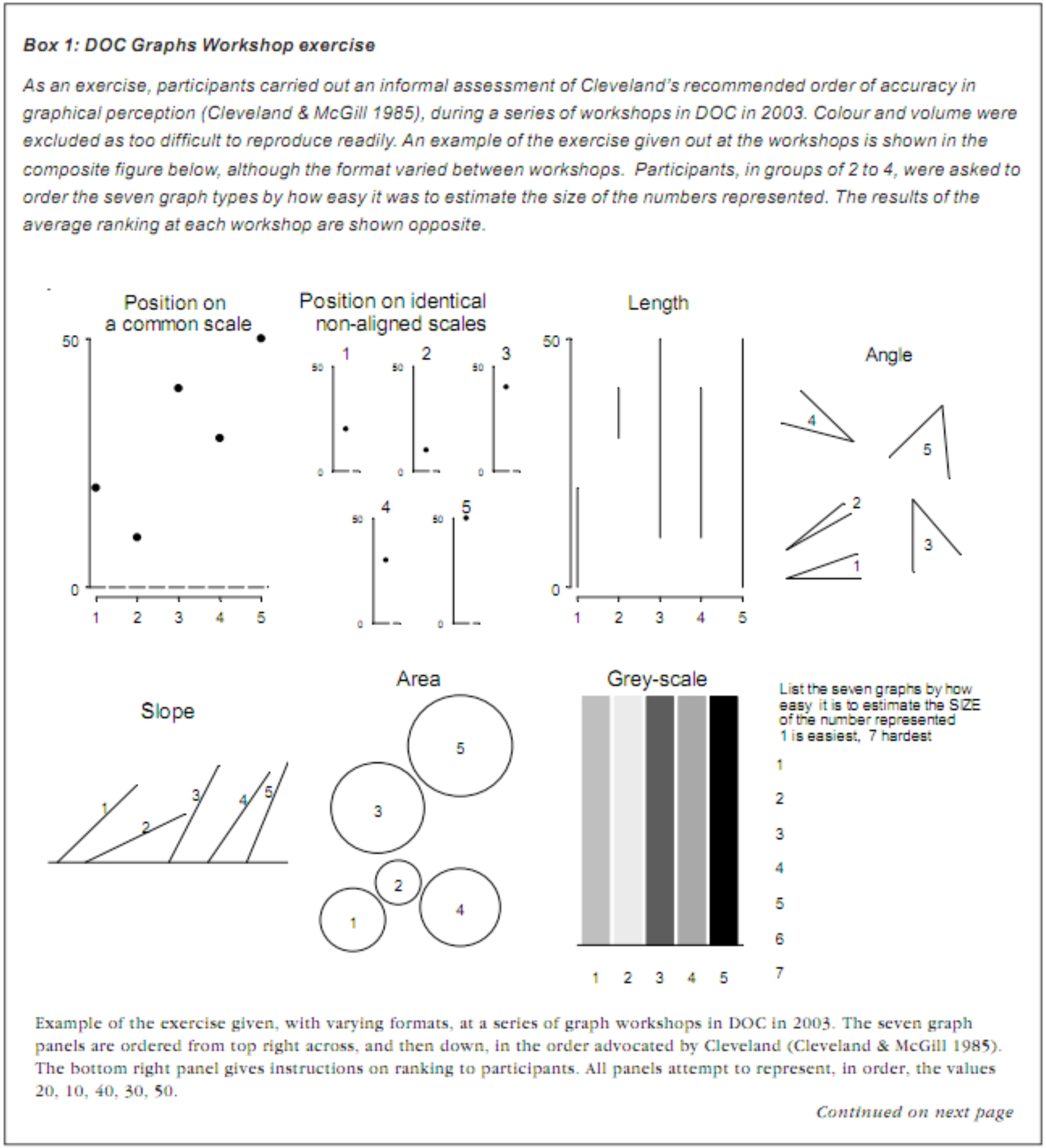
- Spatial region
- Color hue
- Motion
- Shape

- expressiveness principle
  - match channel and data characteristics
- effectiveness principle
  - encode most important attributes with highest ranked channels

Munzer, T. (University of British Columbia, Canada;
IEEE VIS 2015 TutorialOctober 2015, Chicago IL

Accuracy of decoding
Cleveland and McGill (1985)

Common scale is the easiest..note color is least informative but useful to add a dimension
Why we don't use pie graphs if small differences need to be seen



Box 1: DOC Graphs Workshop exercise

As an exercise, participants carried out an informal assessment of Cleveland's recommended order of accuracy in graphical perception (Cleveland & McGill 1985), during a series of workshops in DOC in 2003. Colour and volume were excluded as too difficult to reproduce readily. An example of the exercise given out at the workshops is shown in the composite figure below, although the format varied between workshops. Participants, in groups of 2 to 4, were asked to order the seven graph types by how easy it was to estimate the size of the numbers represented. The results of the average ranking at each workshop are shown opposite.

List the seven graphs by how easy it is to estimate the SIZE of the number represented
1 is easiest, 7 hardest
1
2
3
4
5
6
7

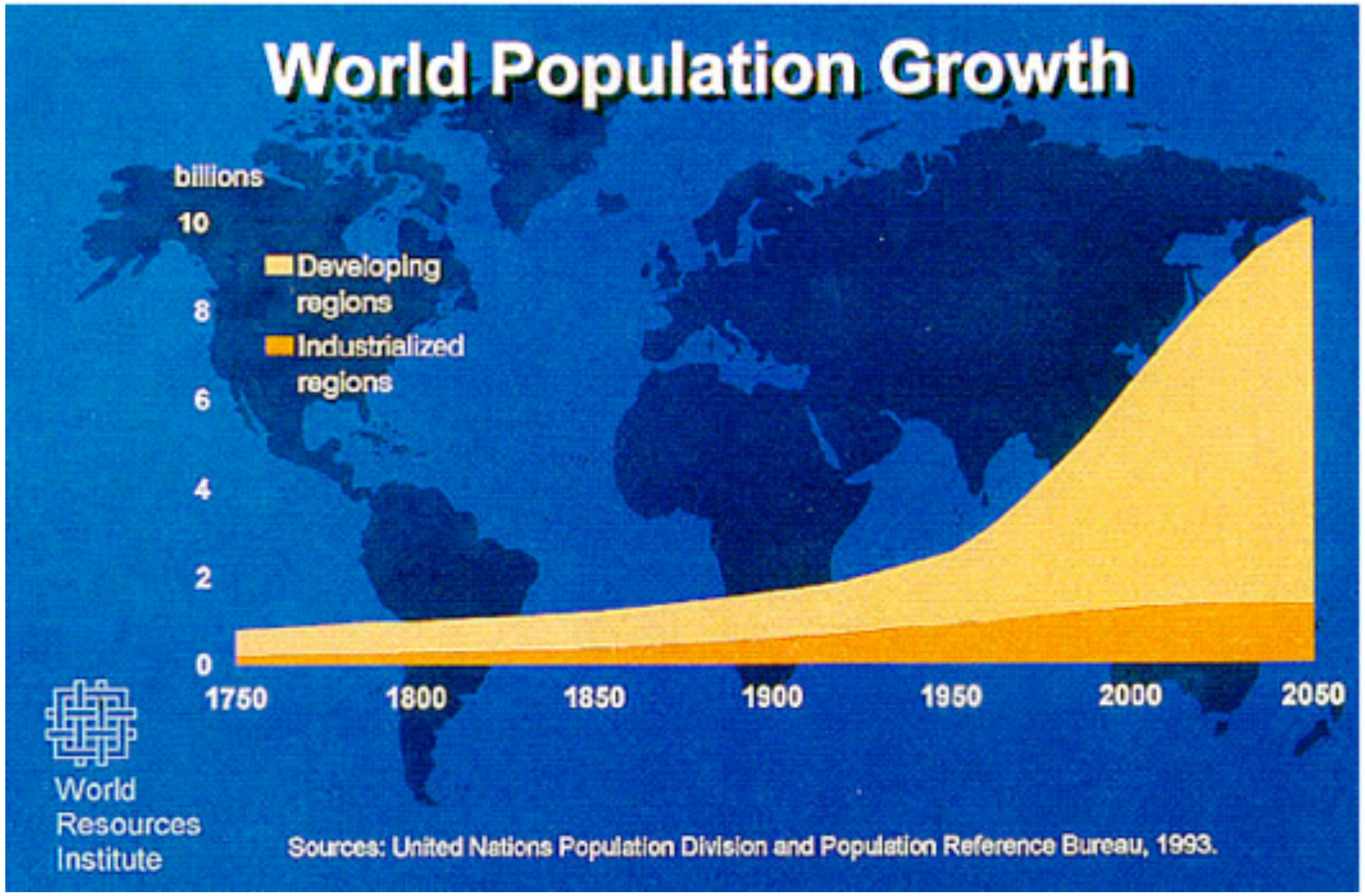
Example of the exercise given, with varying formats, at a series of graph workshops in DOC in 2003. The seven graph panels are ordered from top right across, and then down, in the order advocated by Cleveland (Cleveland & McGill 1985). The bottom right panel gives instructions on ranking to participants. All panels attempt to represent, in order, the values 20, 10, 40, 30, 50.

Continued on next page

How to excel in visualization: Notice! Deconstruct visualizations that you see

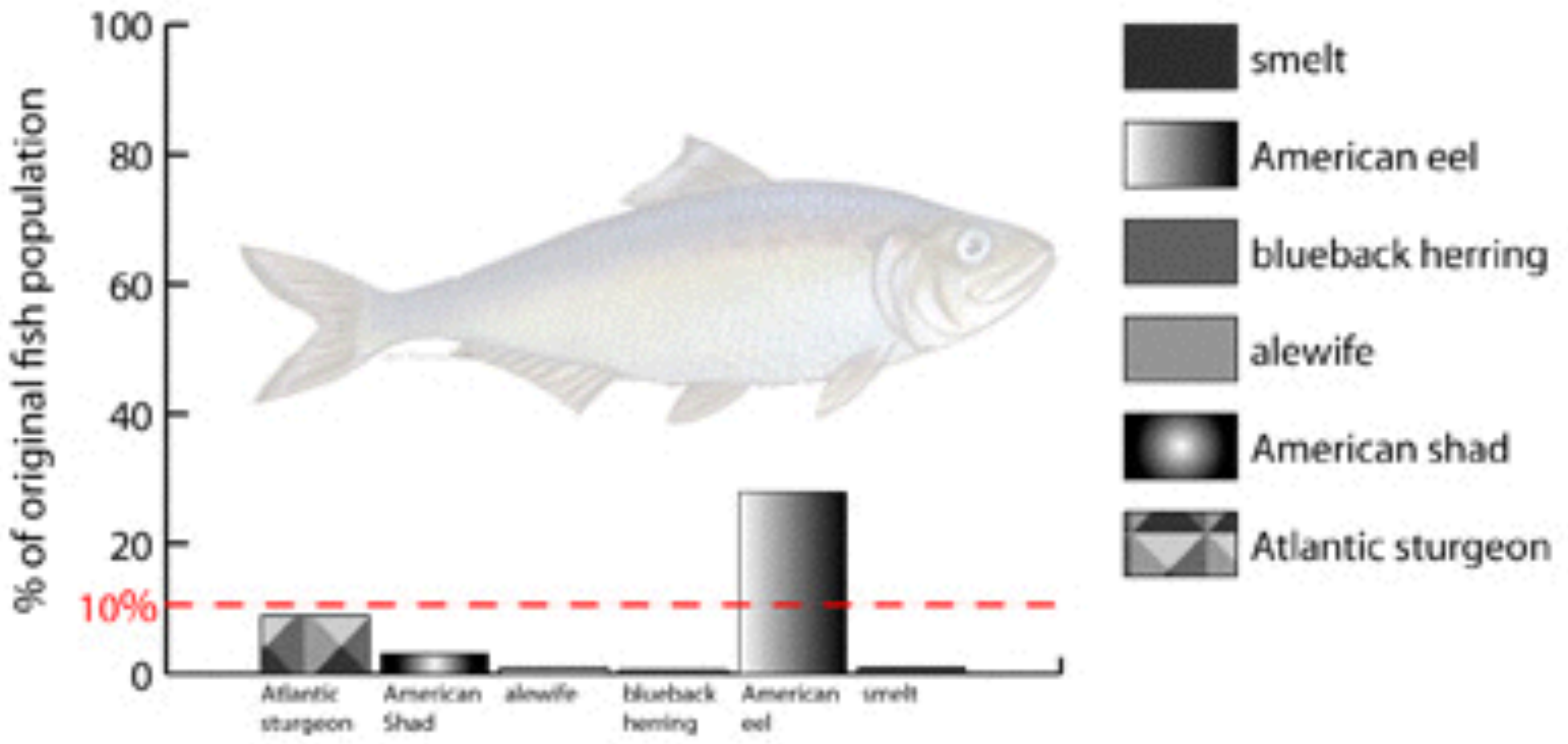
http://blog.udacity.com/2015/01/15-data-visualizati
blow-mind.html

http://spatial.ly/2014/11/r-visualisations-desig
(this one is useful because it points you to the ggp
other R tools used)



Sources: United Nations Population Division and Population Reference Bureau, 1993.

Simplicity

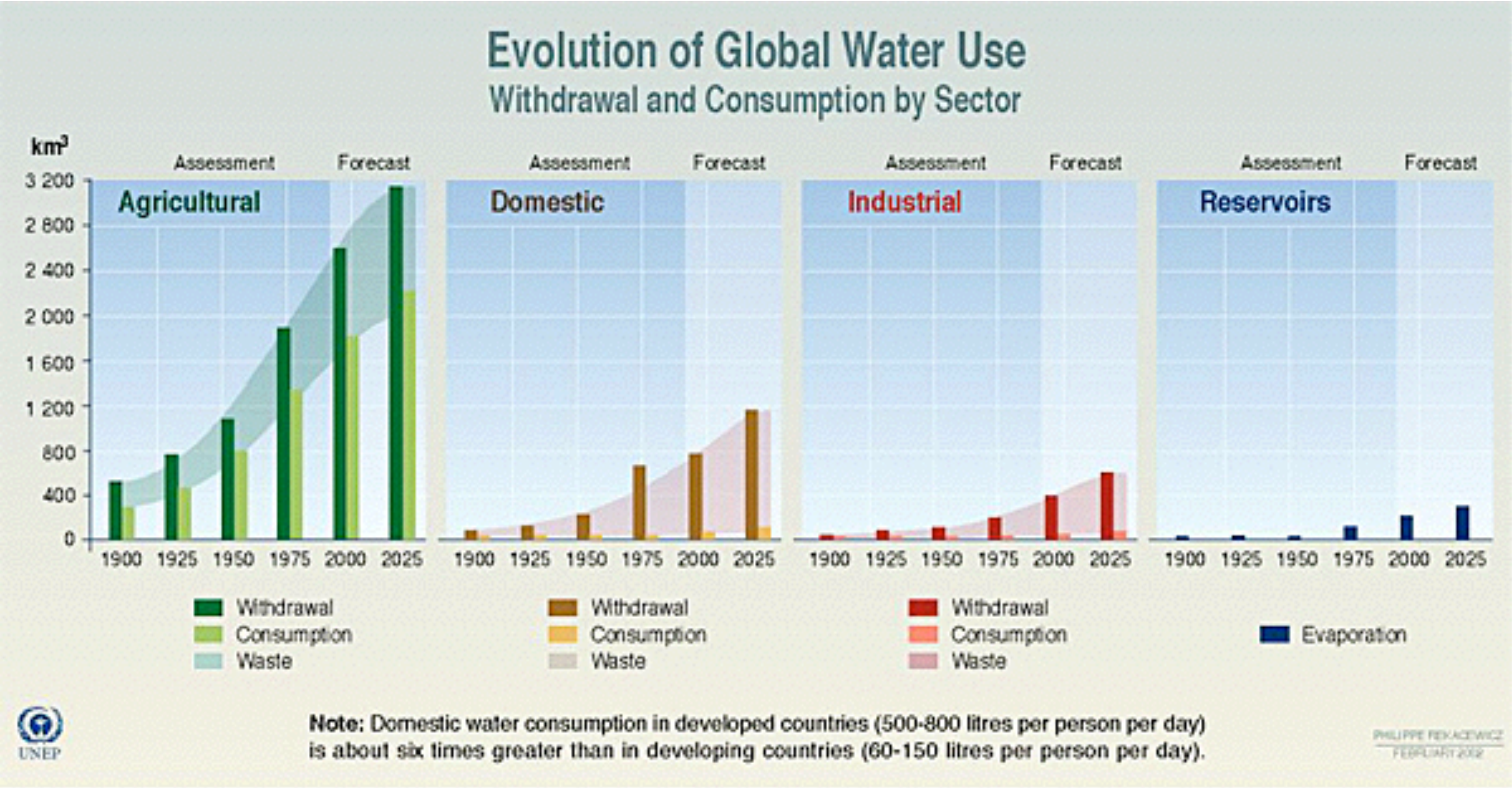


EcoFocus

A biweekly column in the Poughkeepsie Journal

It's Time to Remember to Protect Fish Population

Use of threshold line



Evolution of Global Water Use

Withdrawal and Consumption by Sector

Note: Domestic water consumption in developed countries (500-800 litres per person per day) is about six times greater than in developing countries (60-150 litres per person per day).

Source: Igor A. Shiklomanov, State Hydrological Institute (SHI, St. Petersburg) and United Nations Educational,Scientific and Cultural Organisation (UNESCO, Paris), 1999.

Strategic use of aggregation (by years) and disaggregation (Agricultural, Domestic, Industrial, Reservoirs)

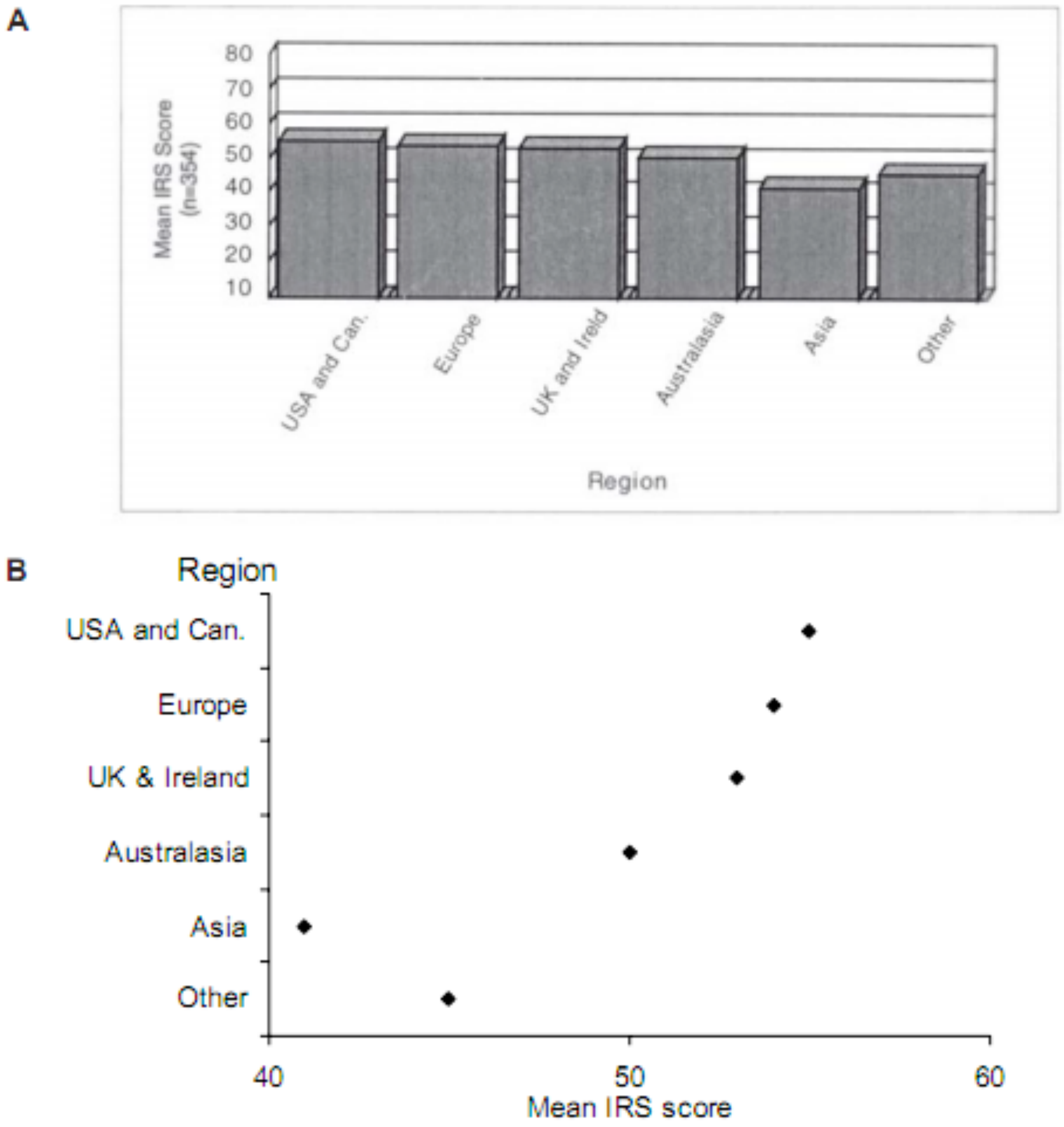
Use of shading to separate Assessment, Forecast, colors and shading for withdrawl and consumption

Good graphing practices: easy to see the distinctions you want to make (but be careful to notice when axis does not go to zero)

Figure 6. Mean scores for individual responsibility by region, from a survey regarding hazard warning signs of visitors to Franz Josef and Fox Glaciers: original (A) and regraphed (B). The differences in bar lengths in the original are difficult to distinguish, made all the more difficult by false 3-dimensional representation. Values in the original dataset were between 11 and 77, so the axes, and the length of the bars, are slightly misleading. The new version below highlights the relative values for the different groups and gives a much tidier appearance by using horizontal, not oblique type.

See Box 2, section 3.6.1 for guidelines on how to change the graph.

Original caption for A: Mean scores for individual responsibility by region.



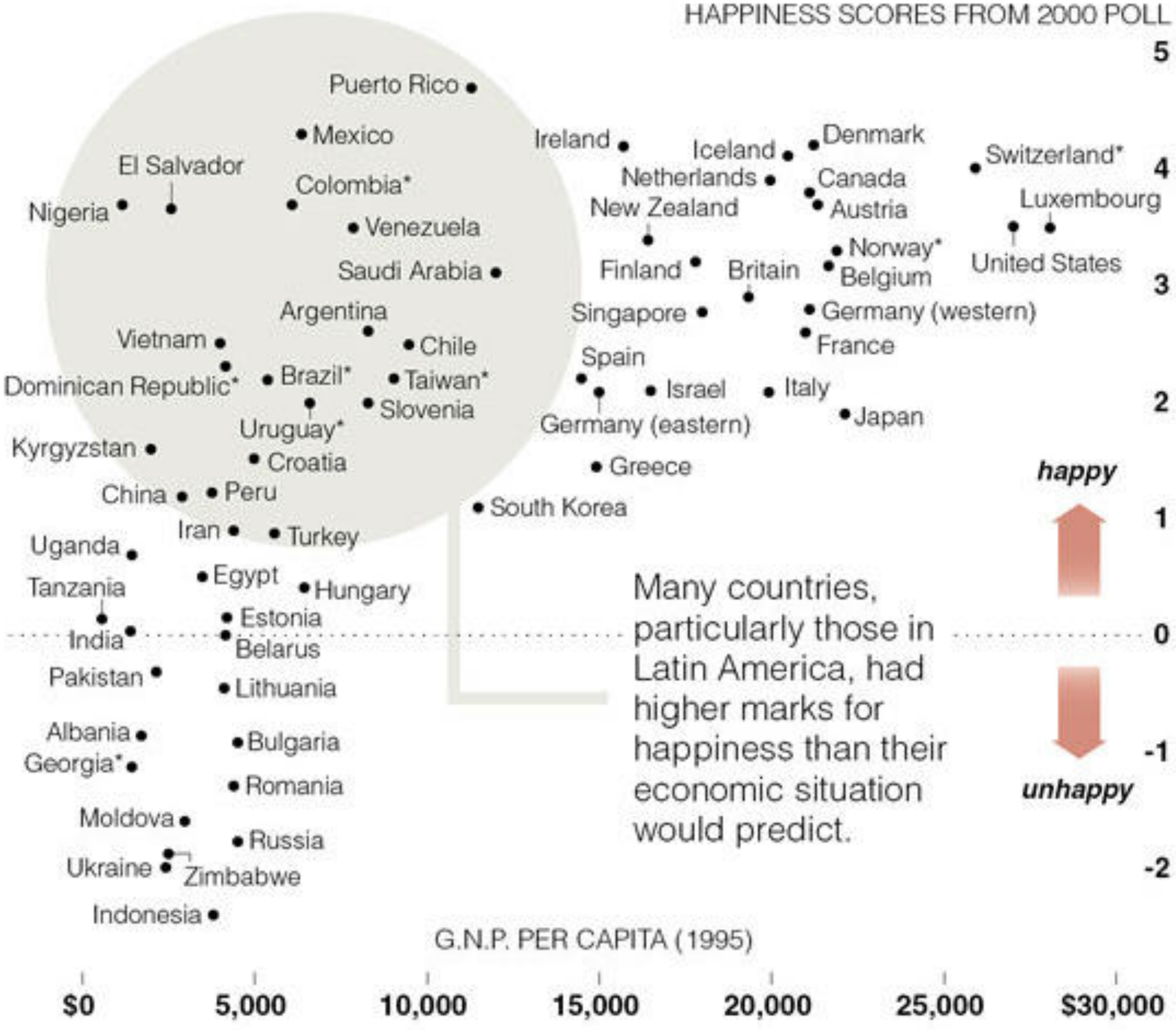
http://www.doc.govt.nz/upload/documents/science-and-technical/docts32.pdf

- what’s wrong with this graphic?

A Plateau of Happiness

A country’s wealth may not always dictate the happiness of its people.

As part of the World Values Survey project, inhabitants of different countries and territories were asked how happy or satisfied they were. Below is a sampling of happiness rankings, along with economic status.



Many countries, particularly those in Latin America, had higher marks for happiness than their economic situation would predict.

*Poll results for these countries were from 1995.

Source: Ronald Inglehart, “Human Beliefs and Values : A Cross-Cultural Sourcebook Based on the 1999-2002 Values Surveys”

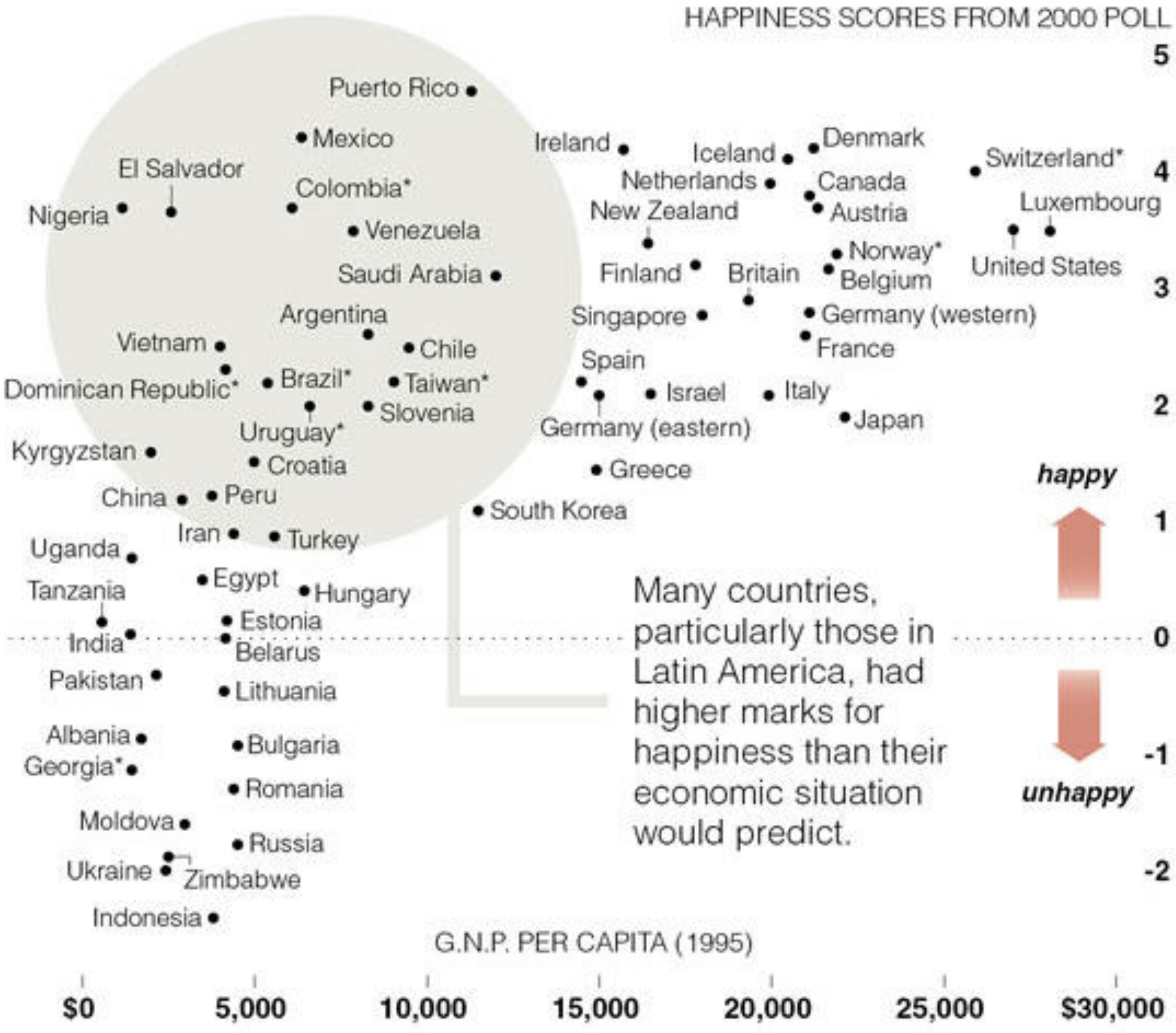
What axis to use? Transforming data

- Axis scale matters (http://www.datavis.ca/gallery/missed.php)
- Conclusion come from not using a log(GPP) axis - which is more appropriate given the data

A Plateau of Happiness

A country's wealth may not always dictate the happiness of its people.

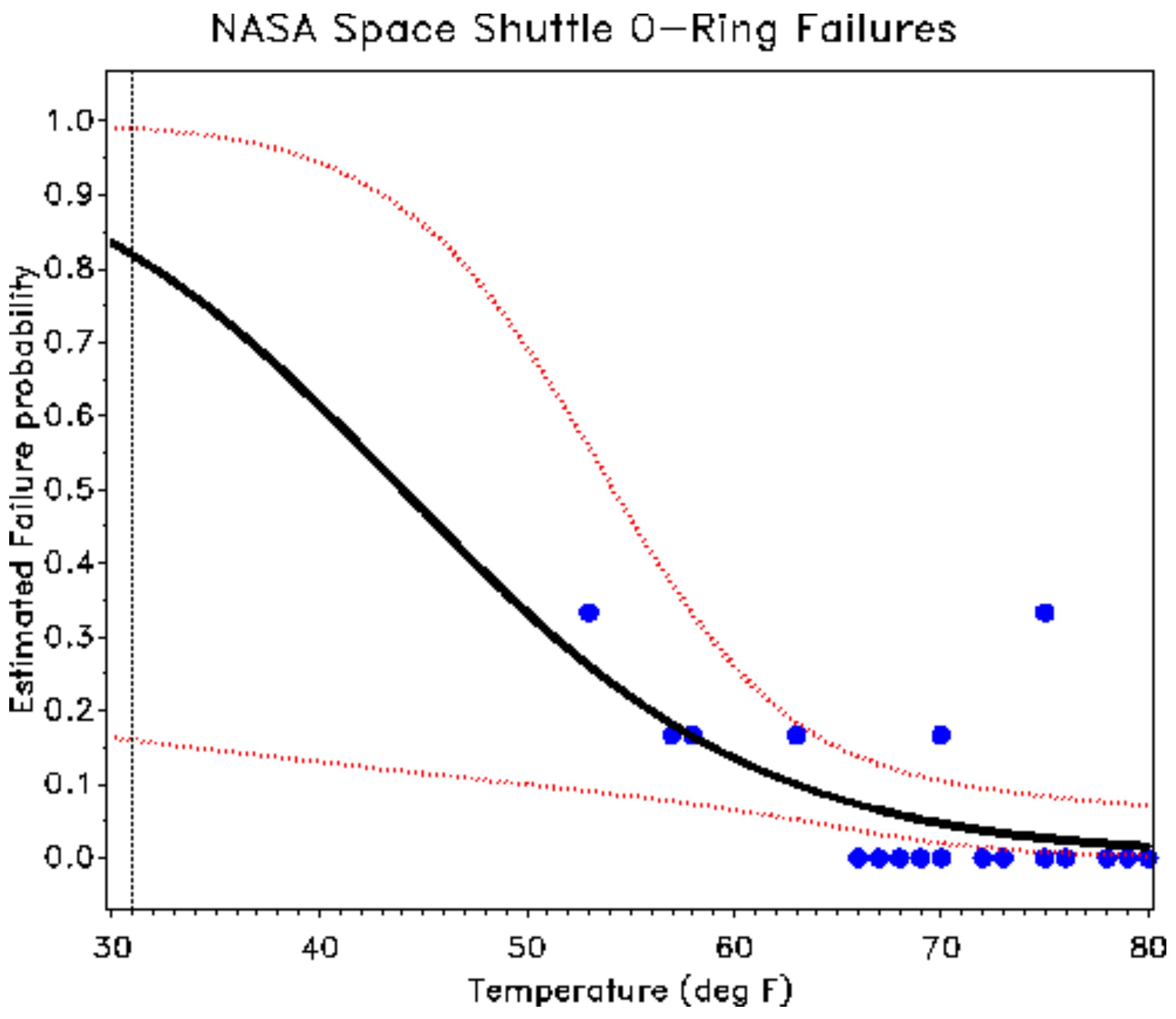
As part of the World Values Survey project, inhabitants of different countries and territories were asked how happy or satisfied they were. Below is a sampling of happiness rankings, along with economic status.



Many countries, particularly those in Latin America, had higher marks for happiness than their economic situation would predict.

*Poll results for these countries were from 1995.

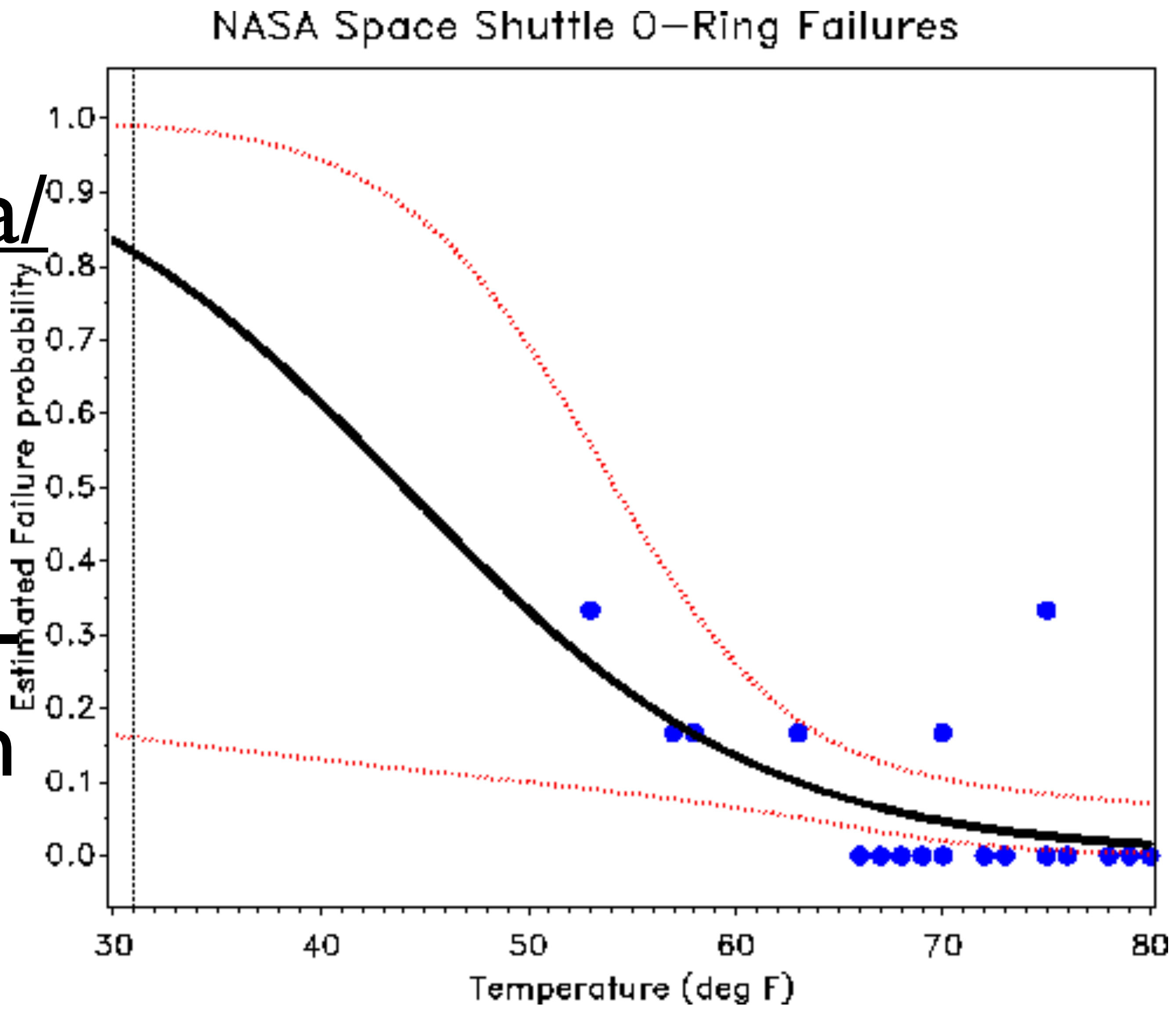
Source: Ronald Inglehart, "Human Beliefs and Values : A Cross-Cultural Sourcebook Based on the 1999-2002 Values Surveys"

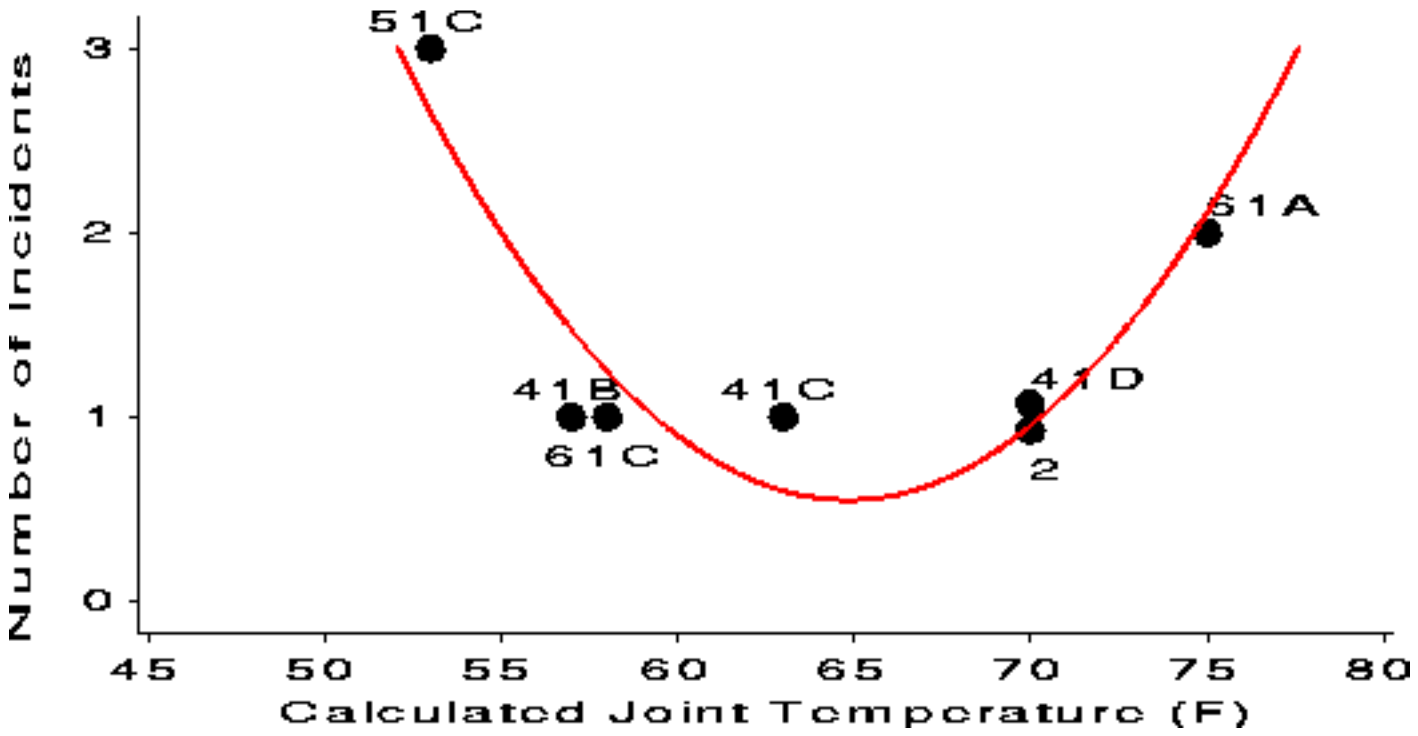


What about this one?

http://www.datavis.ca/gallery/missed.php

Showing relationships - use of trend lines from fitted models can obscure patterns if wrong model is used



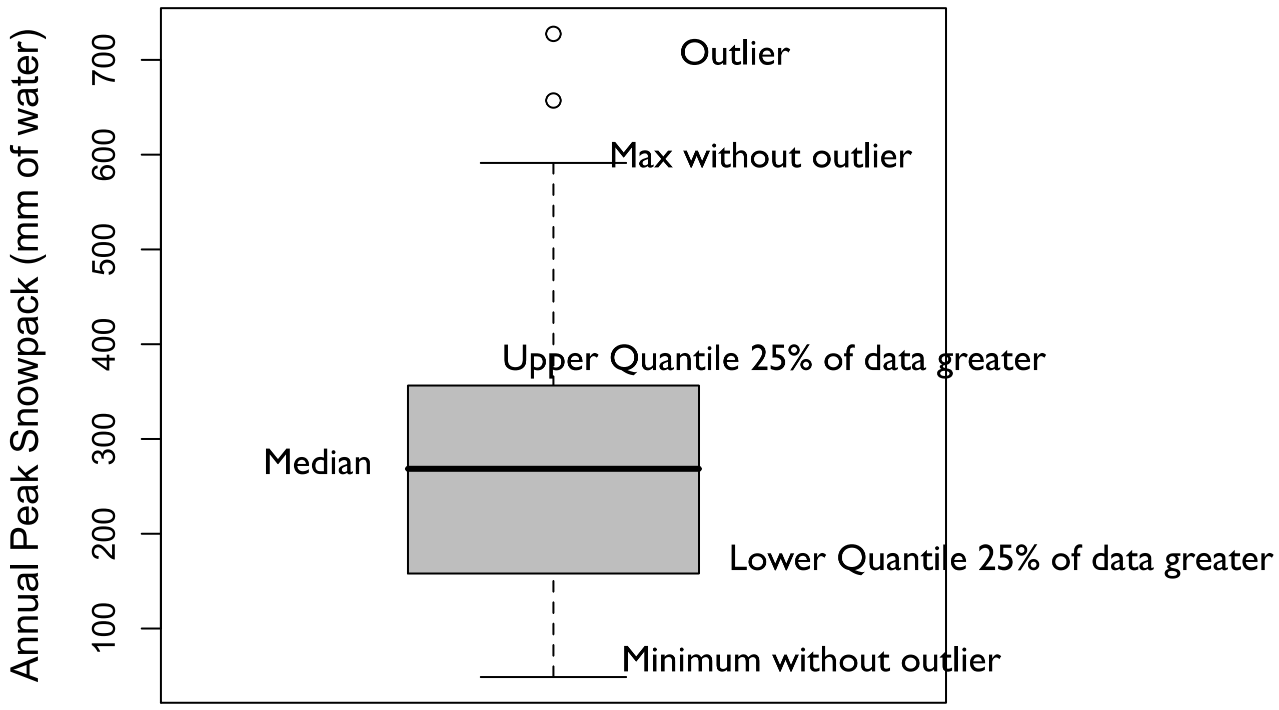


To get a sense of just how sophisticated visualization can be

http://www.datavis.ca/milestones/

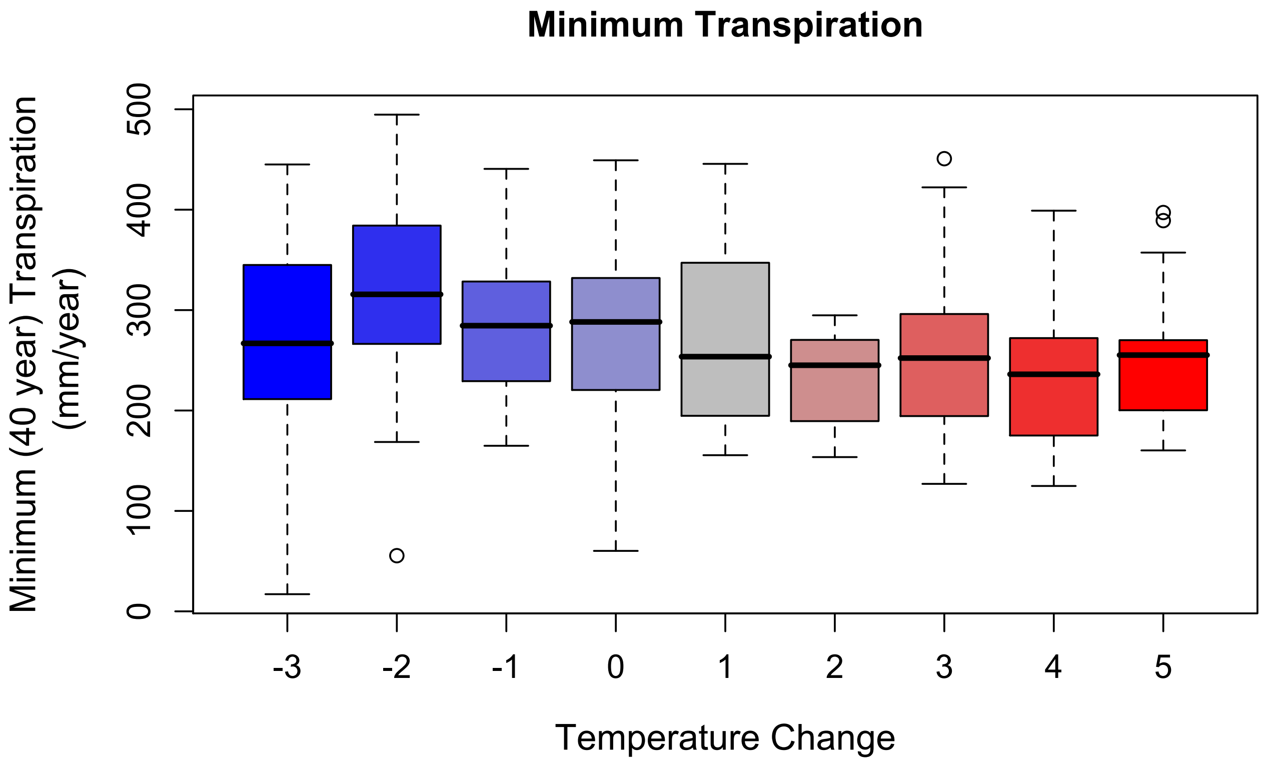
Box-plots

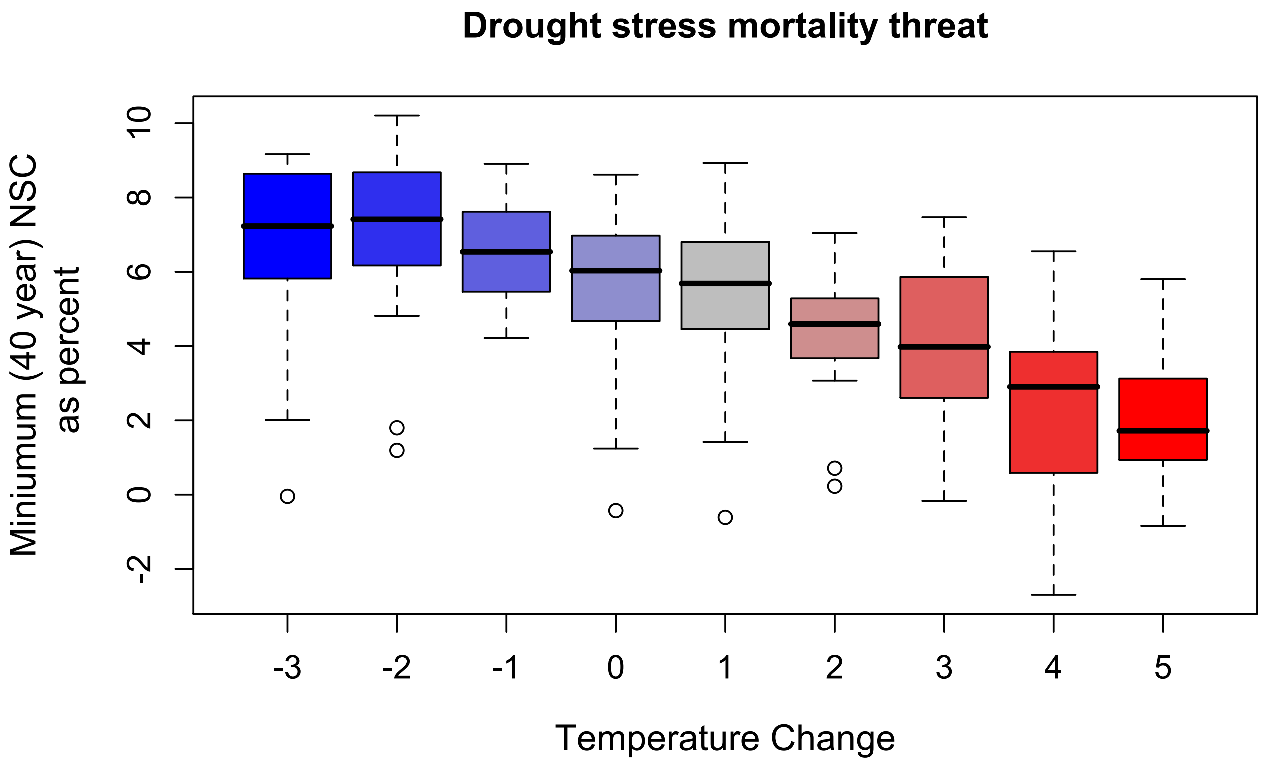
- concise, easy to get a visual sense of the range of data
- obscure the finer features of the distribution (think highly skewed data versus normal, or a multi-modal distribution)



Multiple Box-plots

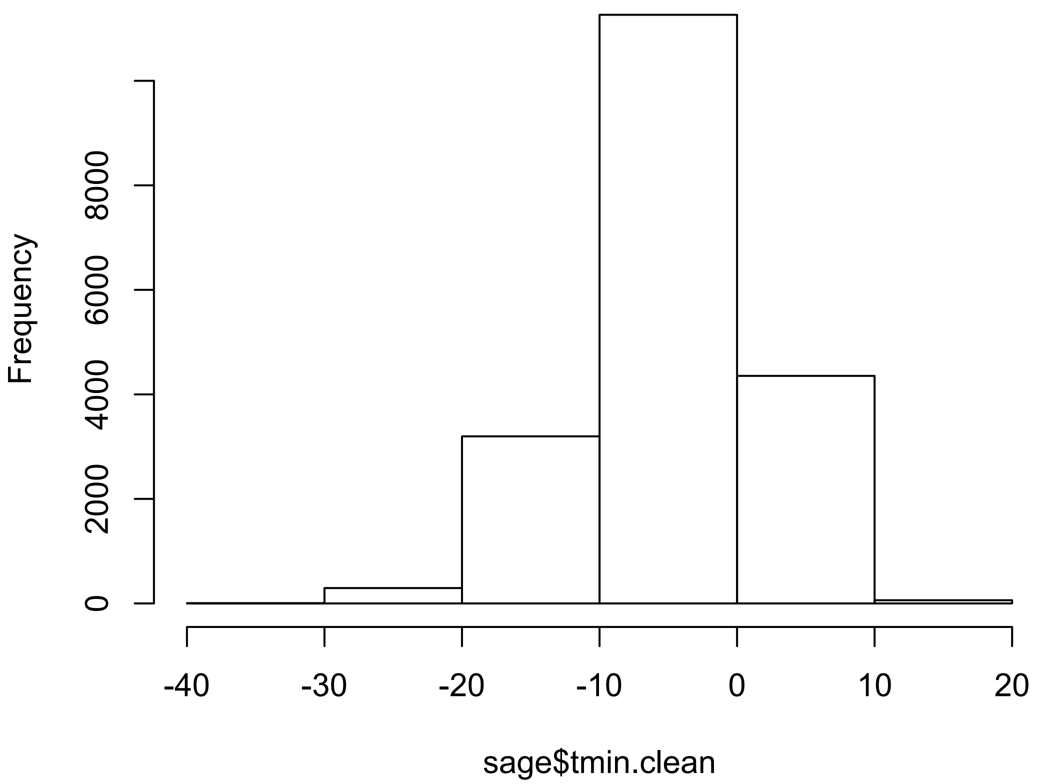
Estimates of Ponderosa Pine ecophysiologial responses to warming/cooling


Minimum Transpiration

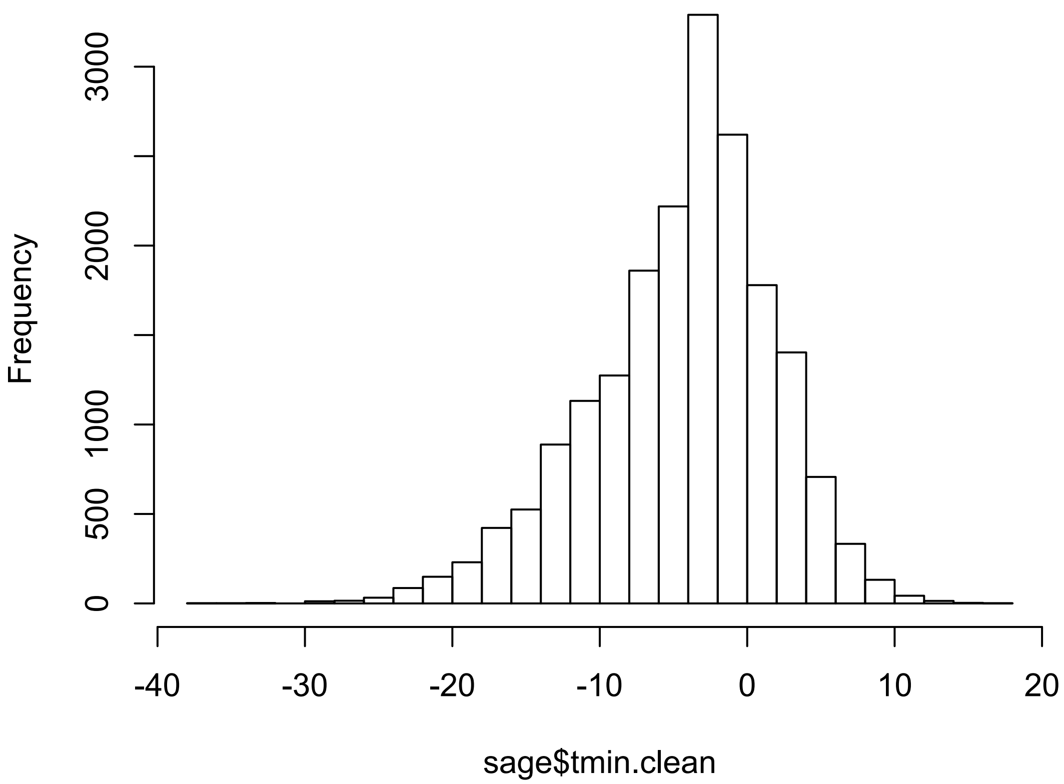

Drought stress mortality threat

Histograms

- Advantage: simple, commonly used
- Disadvantage: no information about distribution within the bins; can lead to mis-interpretation - by changing the width of bins you can dramatically alter the histogram

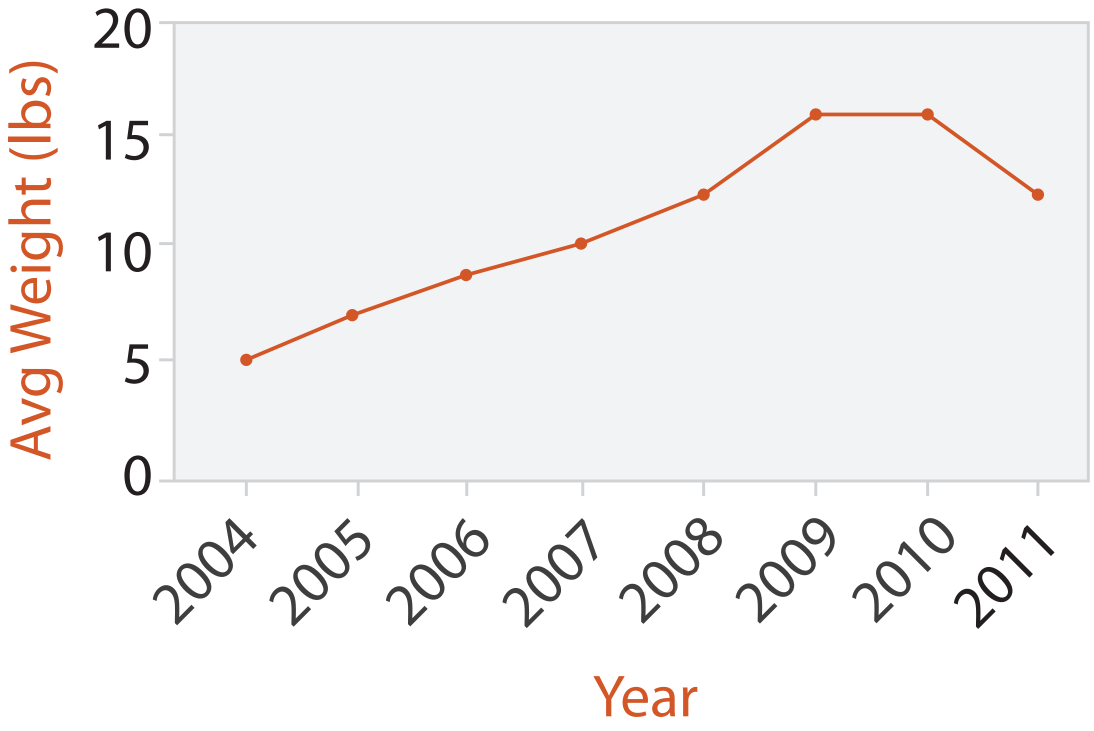


Histogram Tmin (Breaks = 4)



Histogram Tmin (Breaks = 30)

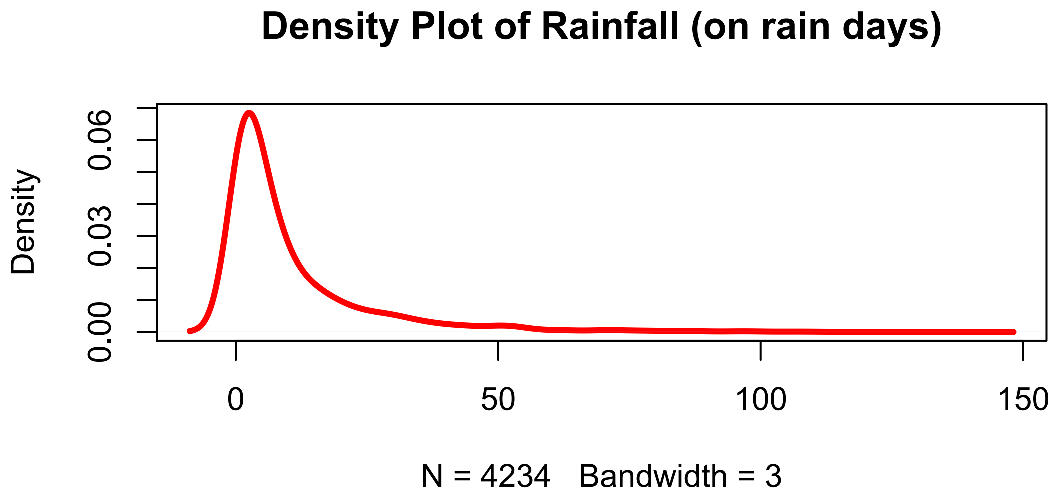
Line Graphs

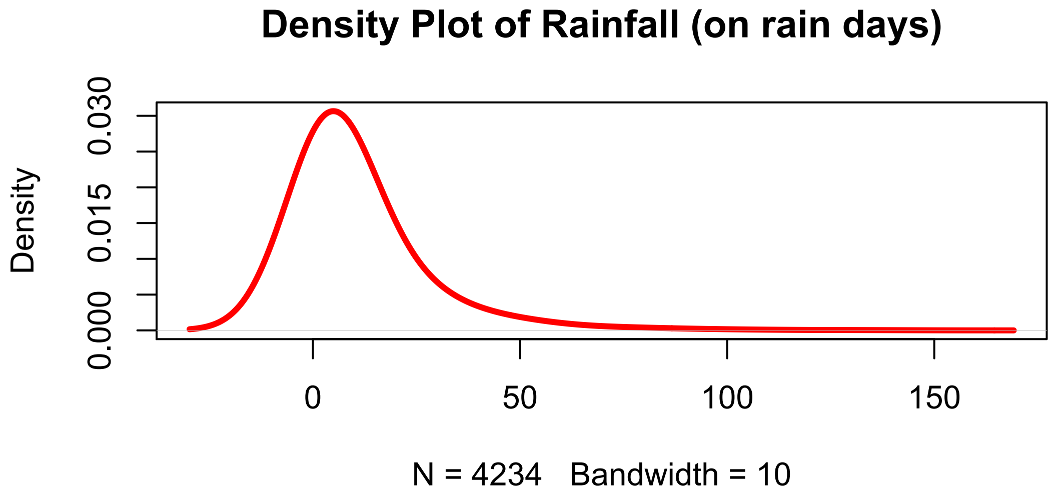


- Useful when looking for trends
- Understanding the relationship between consecutive items

Density

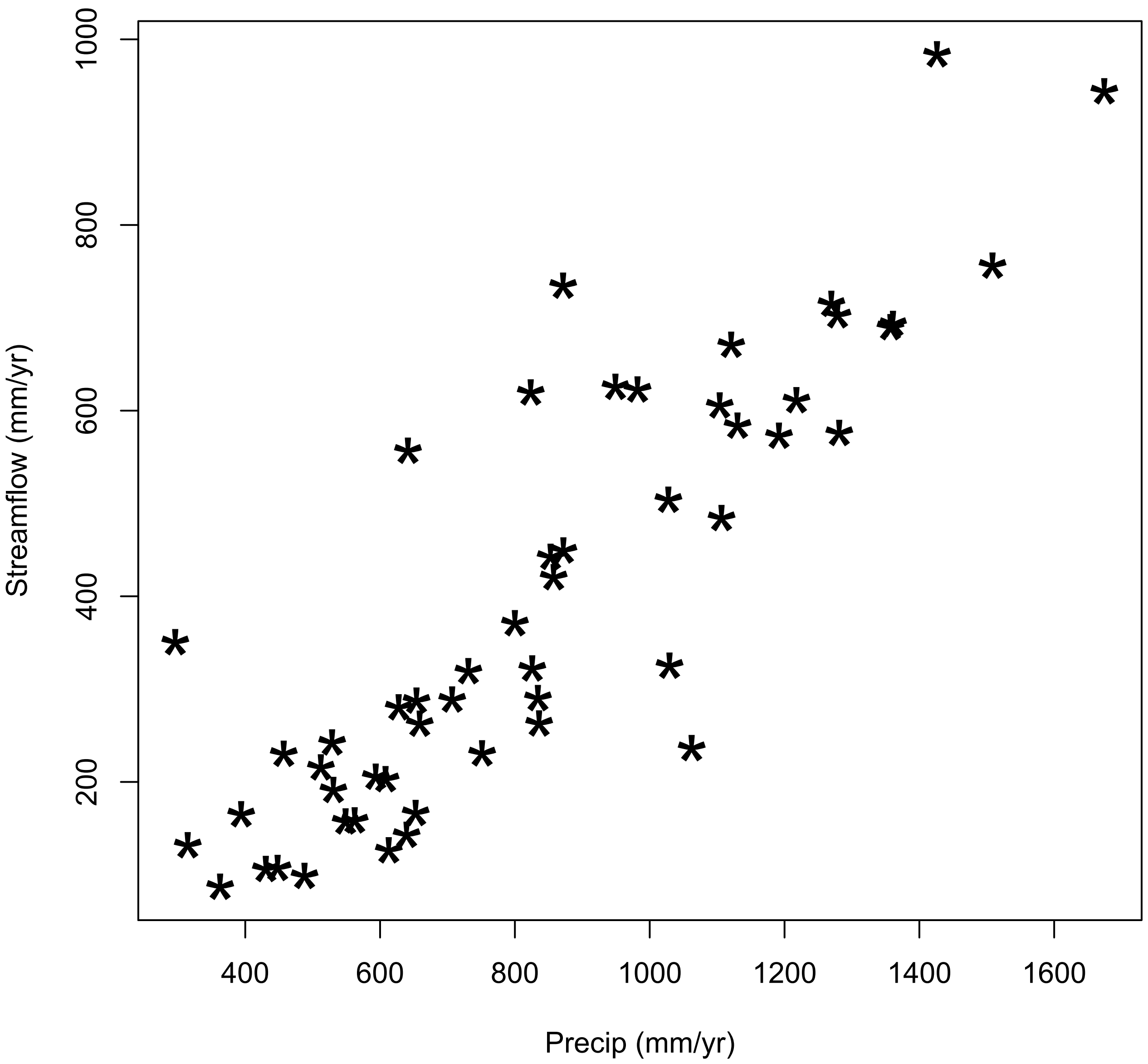
- useful for getting a sense of the shape of the distribution of the data
- does not have histograms problems of sensitivity to bin width
- smoothing (always occurs) can blur (sometimes important) patterns in the distribution
- bandwidth (or smoothing kernal) controls the degree of smoothing (think of as a smoothing window)
- local density =
  - number of x $x-h/2 < x_i < x+h/2$
  - where h is width
- most use more complex formula that weight observations relative to distance from window center





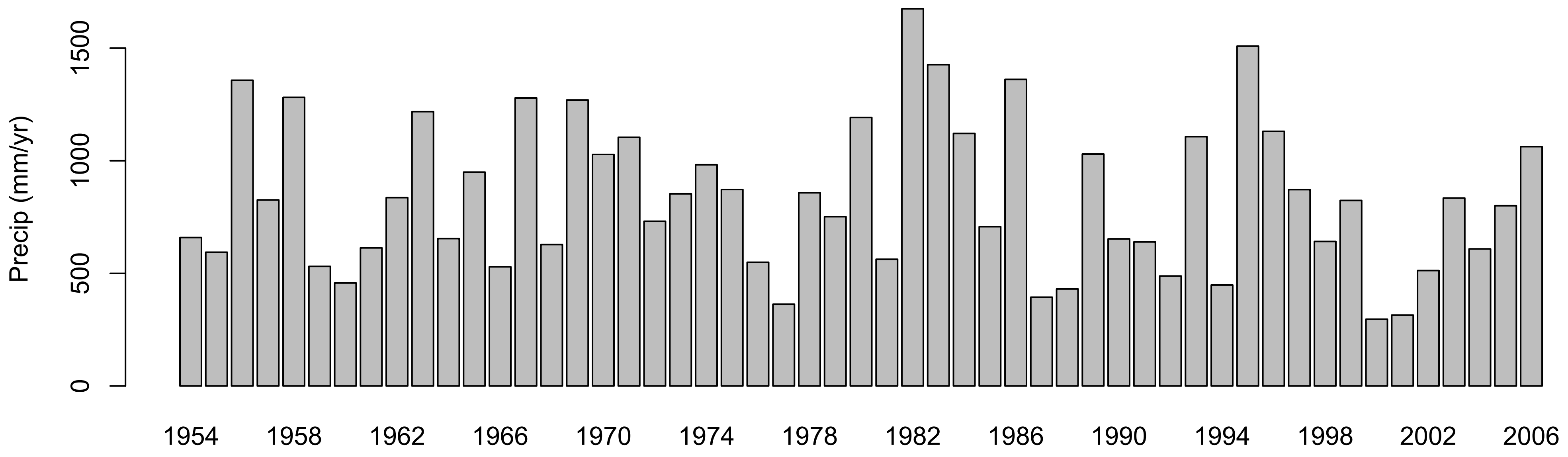
Scatterplots

- Useful for comparing data sets or looking at relationships between variables



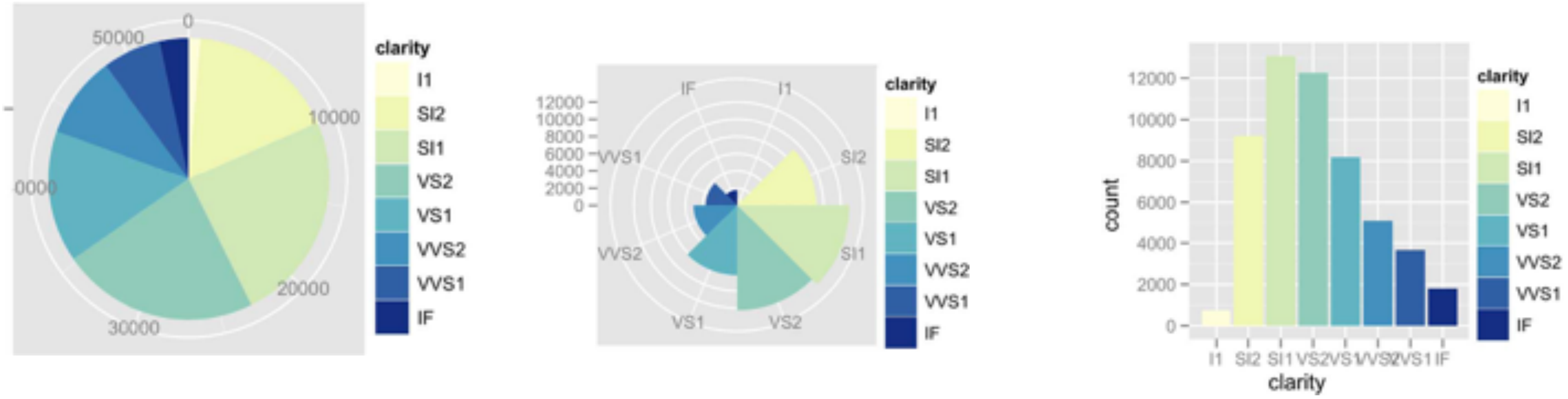
Barplots

- Useful for categorical data
- Implied binning/aggregation of data (sum of precip)



pie chart, polar area chart

- relationship between part and whole



[A layered grammar of graphics. Wickham. Journ. Computational and Graphical Statistics 19:1 (2010), 3–28.]

Some important choices

Accuracy, honesty, simplicity, ease of interpretation in graphing depend on choices about

Type of graph (boxplot, line, scatter plot, multiple plots versus 1 plot..3-d)

Axis (linear, log, other, more than one)

Symbols and colors used (line implies continuous or interpolation, box discrete)

Legends, labels, units (always add, even for yourself!)

If and how to communication uncertainty/variability

What to graph (aggregation and disaggregation, normalize/deviations, combinations of variables)

GGPLOT Resources

http://ggplot2.org/

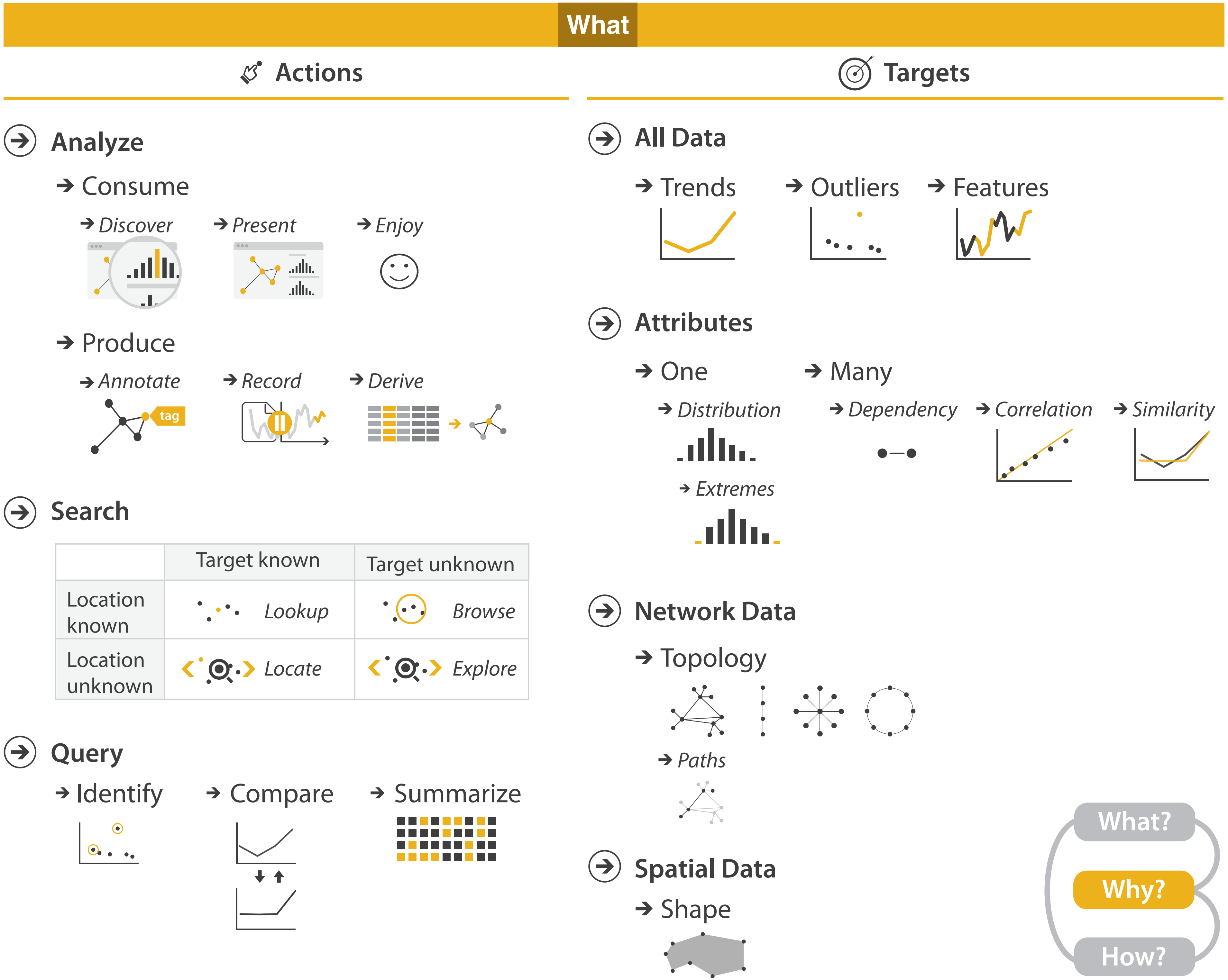
http://ggplot2.org/book/ (best book worth it if you are goi
do this a lot)

With your group

- Brainstorm about example of both visualization for communication; and visualization for understanding

- For communication - describe a possible data set, state your goal and a possible displays you would want to create to achieve that goal

- For understanding - describe a possible data set and come up with an example of a goal where you might meet it using a) all data b) attributes of the data c) networks d) spatial (from slide 6/33)



What

Actions

Analyze

- Consume
  - *Discover*
  - *Present*
  - *Enjoy*
- Produce
  - *Annotate*
  - *Record*
  - *Derive*

Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

- Identify
- Compare
- Summarize

Targets

All Data

- Trends
- Outliers
- Features

Attributes

- One
  - *Distribution*
    - *Extremes*
- Many
  - *Dependency*
  - *Correlation*
  - *Similarity*

Network Data

- Topology
  - *Paths*

Spatial Data

- Shape

What?

Why?

How?

- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Munzer, T. (University of British Columbia, Canada;
IEEE VIS 2015 TutorialOctober 2015, Chicago IL

Example