Data Viz

April 2, 2020
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Josh Levin, Diane Mutako, Sol Zitter

Announcements

- Videos on if you can! Use raise-hand feature for questions.
- Any questions/concerns logistically?
- Extra Office Hours tomorrow

Today

- Questions from previous lectures? (Dimensionality Reduction, Classification, Regularization)

- Data Viz tips and best practices

When do I do data viz during a project?

When do I do data viz during a project?

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Run linear regression, control for various things, find large coefficient on whether student has two concentrations

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Viz #2: Quick histograms (or box-whiskers maybe) of hours of sleep vs. number of concentrations

Run linear regression, control for various things, find large coefficient on whether student has two concentrations

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Viz #2: Quick histograms (or box-whiskers maybe) of hours of sleep vs. number of concentrations

Run linear regression, control for various things, find large coefficient on whether student has two concentrations

Viz #3: Quick histogram of number of concentrations for CS vs. non-CS students

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Viz #2: Quick histograms (or box-whiskers maybe) of hours of sleep vs. number of concentrations

Run linear regression, control for various things, find large coefficient on whether student has two concentrations

Viz #3: Quick histogram of number of concentrations for CS vs. non-CS students

Viz #4: Final polished visualizations for poster/paper/ report

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

while not converged

Viz #1: Quick side-by-side histogram of CS students' sleep vs. the rest. Means + CIs

Viz #ia: Quick histograms (or box-whiskers maybe) of hours of sleep vs. number of concentrations

Run linear regression, control for various things, find large coefficient on whether student has two concentrations

Viz #ib: Quick histogram of number of concentrations for CS vs. non-CS students

Viz #N+1: Final polished visualizations for poster/paper/ report

Hypothesis: CS students sleep less than Brown students in general

When do I do data viz during a project?

- At the very start of analysis, to find out wth is going on in my data
- Periodically throughout, to vet the quantitative trends I am seeing
- At the very end of a project, to showcase the results

When do I do data viz during a project?

- At the very start of analysis, to find out wth is going on in my data
- Periodically throughout, to vet the quantitative trends I am seeing
- At the very end of a project, to showcase the results

More important

(matplotlib, excel, whatever is easy)

When do I do data viz during a project?

Most attention, cause its fun ;)

(D3, etc.)

- At the very start of analysis, to find out wtn is going on in my data
- Periodically throughout, to vet the quantitative trends I am seeing
- At the very end of a project, to showcase the results

When do I do data viz during a project?

- At the very start of analysis, to find out wth is going on in my data
- Periodically throughout, to vet the quantitative trends I am seeing
- At the very end of a project, to showcase the results

You are the main audience, goal is to make sure you understand what you are looking at

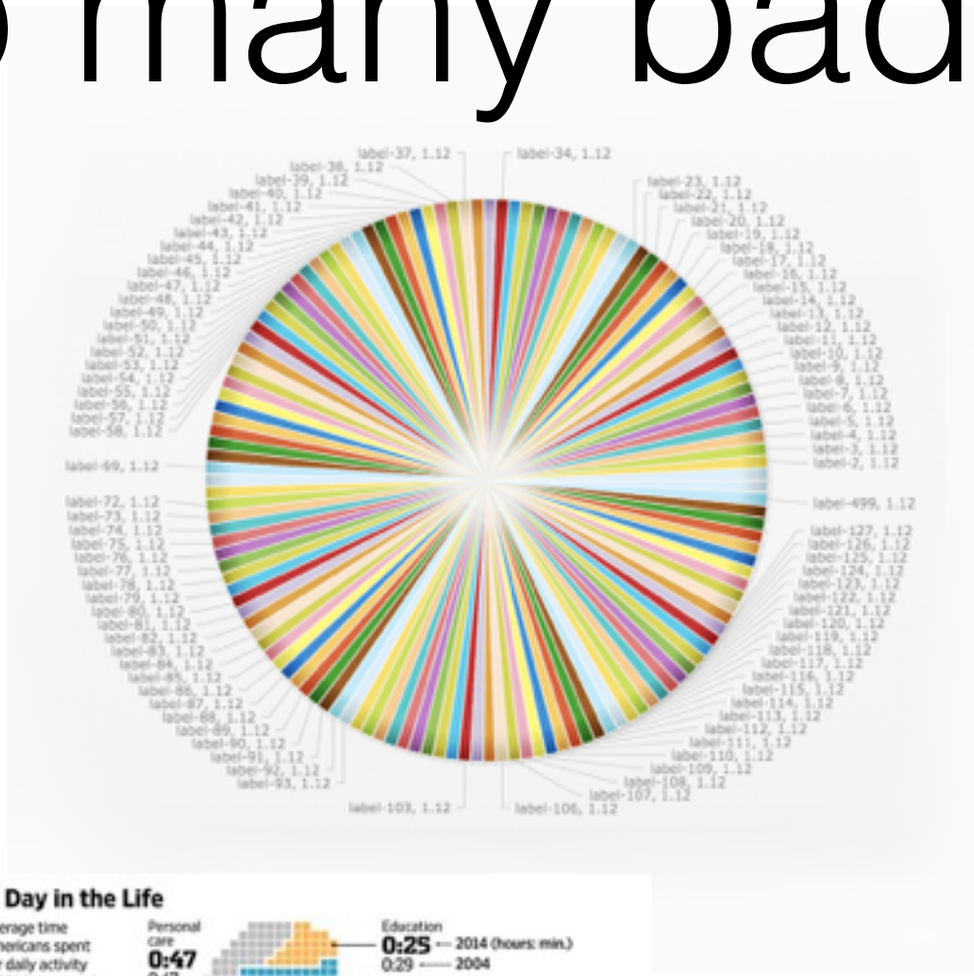
When do I do data viz during a project?

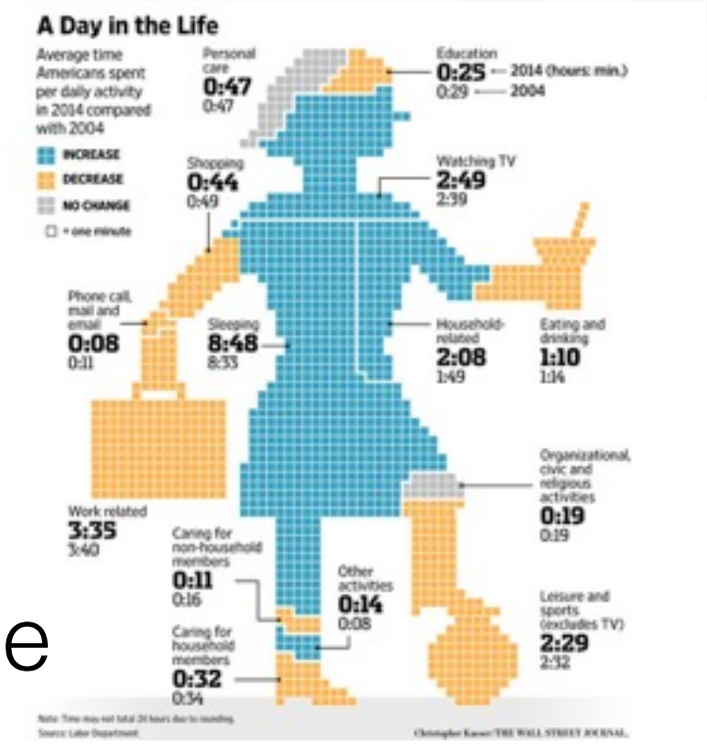
Everyone else is the main audience. Goal is to make point as clearly and concisely as possible.

- At the very start of analysis, to find out wtn is going on in my data
- Periodically throughout, to vet the quantitative trends I am seeing
- At the very end of a project, to showcase the results

So many bad figures…

Diane





Maggie

Neil

My “three pillars”* of Data Viz

*:)

My “three pillars” of Data Viz

clarity — Your figures should speak for themselves. The analysis should be understandable and your conclusions should be obviously supported, without too much effort

My “three pillars” of Data Viz

clarity — Your figures should speak for themselves. The analysis should be understandable and your conclusions should be obviously supported, without too much effort

Don’t obfuscate the data or **H**ide the pr**O**cess you used to come to your co**N**clusions. Giv**E** people enough data **S**o that **T**hey can disagree with **Y**ou if they want to.

My “three pillars” of Data Viz

clarity — Your figures should speak for themselves. The analysis should be understandable and your conclusions should be obviously supported, without too much effort

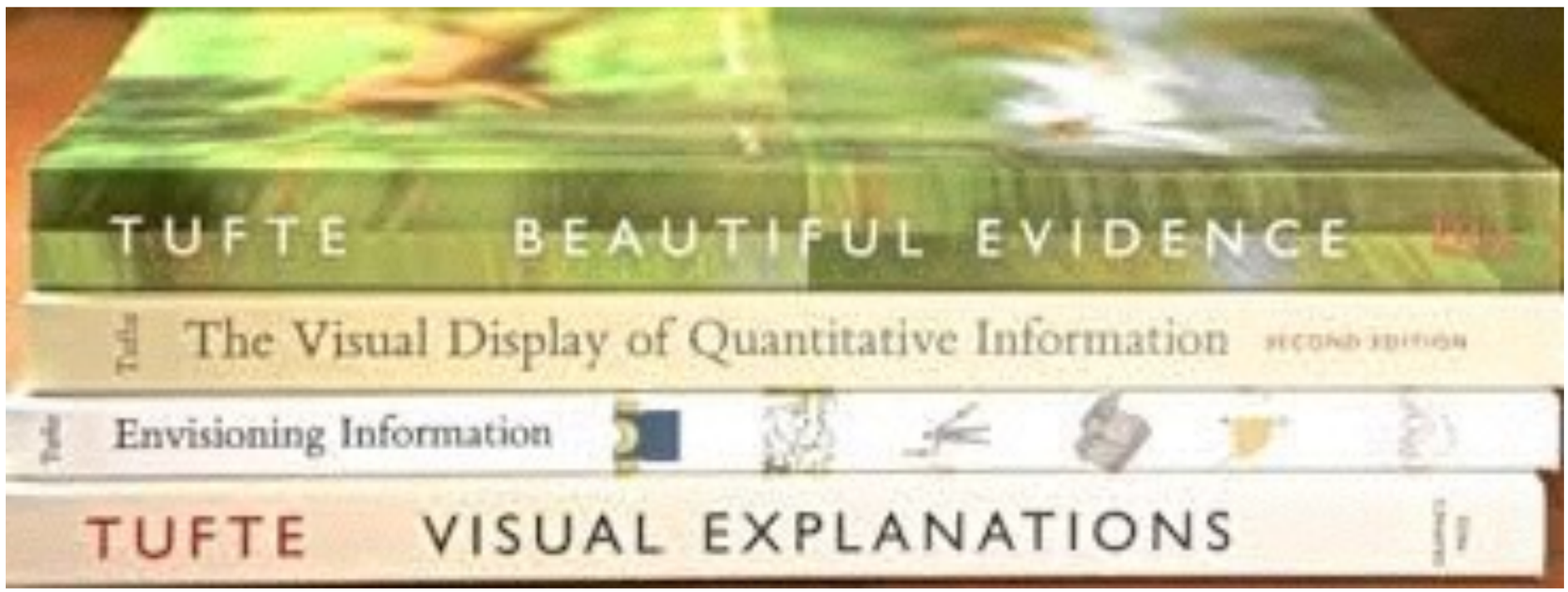
Don’t obfuscate the data or **H**ide the pr**O**cess you used to come to your co**N**clusions. Giv**E** people enough data **S**o that **T**hey can disagree with **Y**ou if they want to.

Minimalism — Substance over style. Make your point concisely, without redundant or distracting information or ornamentation.

Ellie rants about culture for 2 seconds. Indulge me….

“form follows function”

Great tangent to go on…

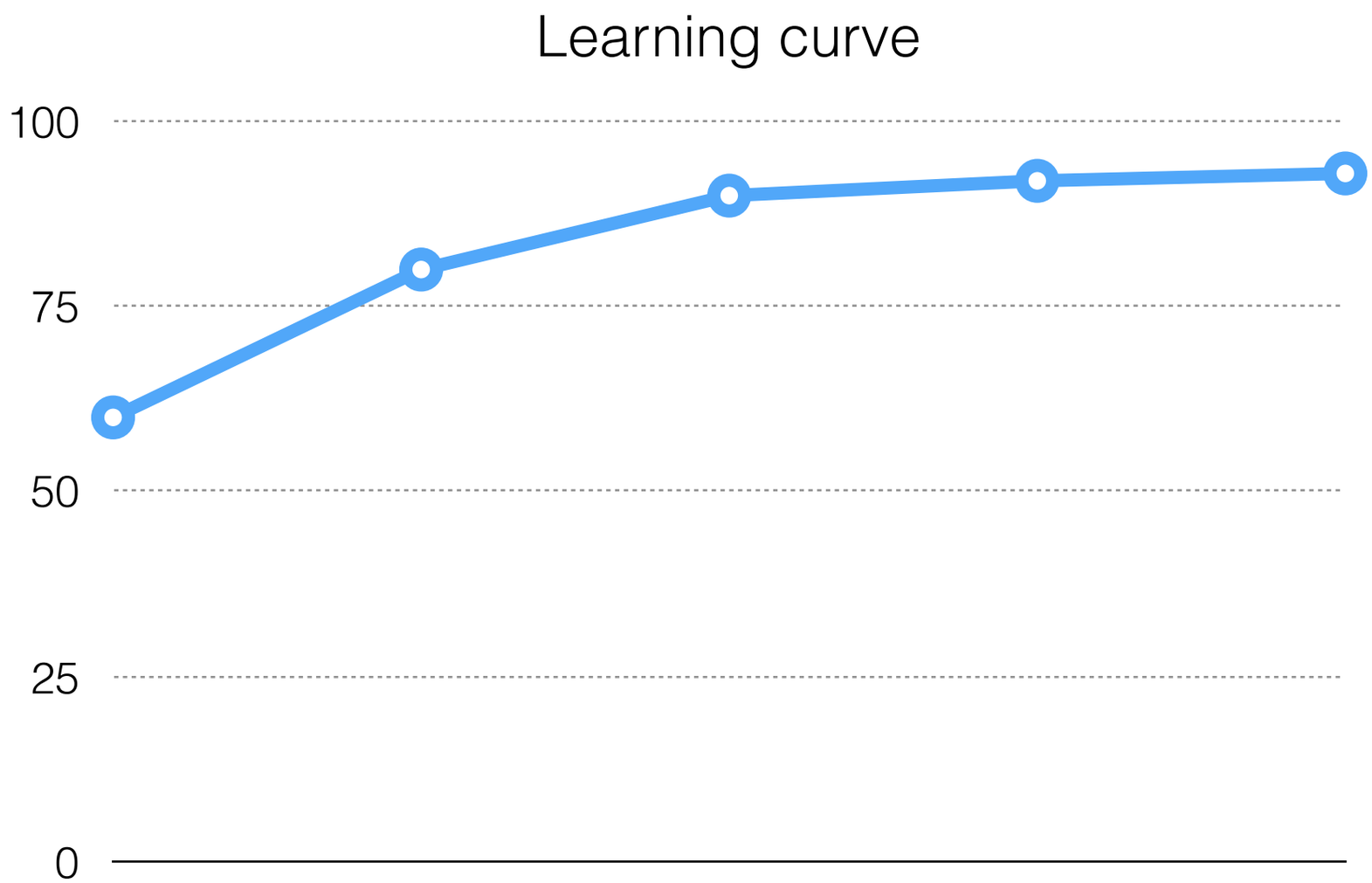


Edward Tufte—dogma of data viz

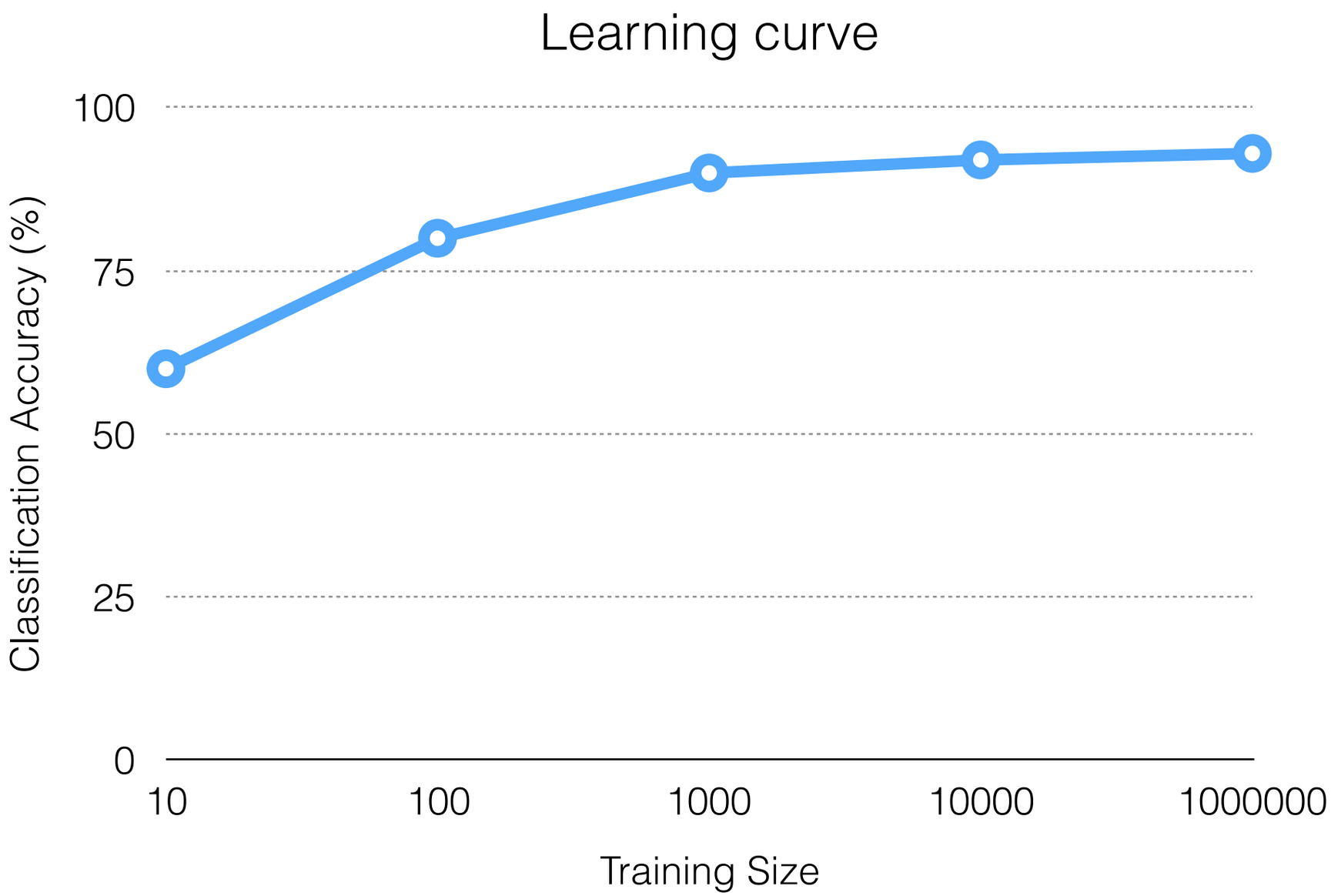
My “three pillars” of Data Viz

clarity — Your figures should speak for themselves. The analysis should be understandable and your conclusions should be obviously supported, without too much effort

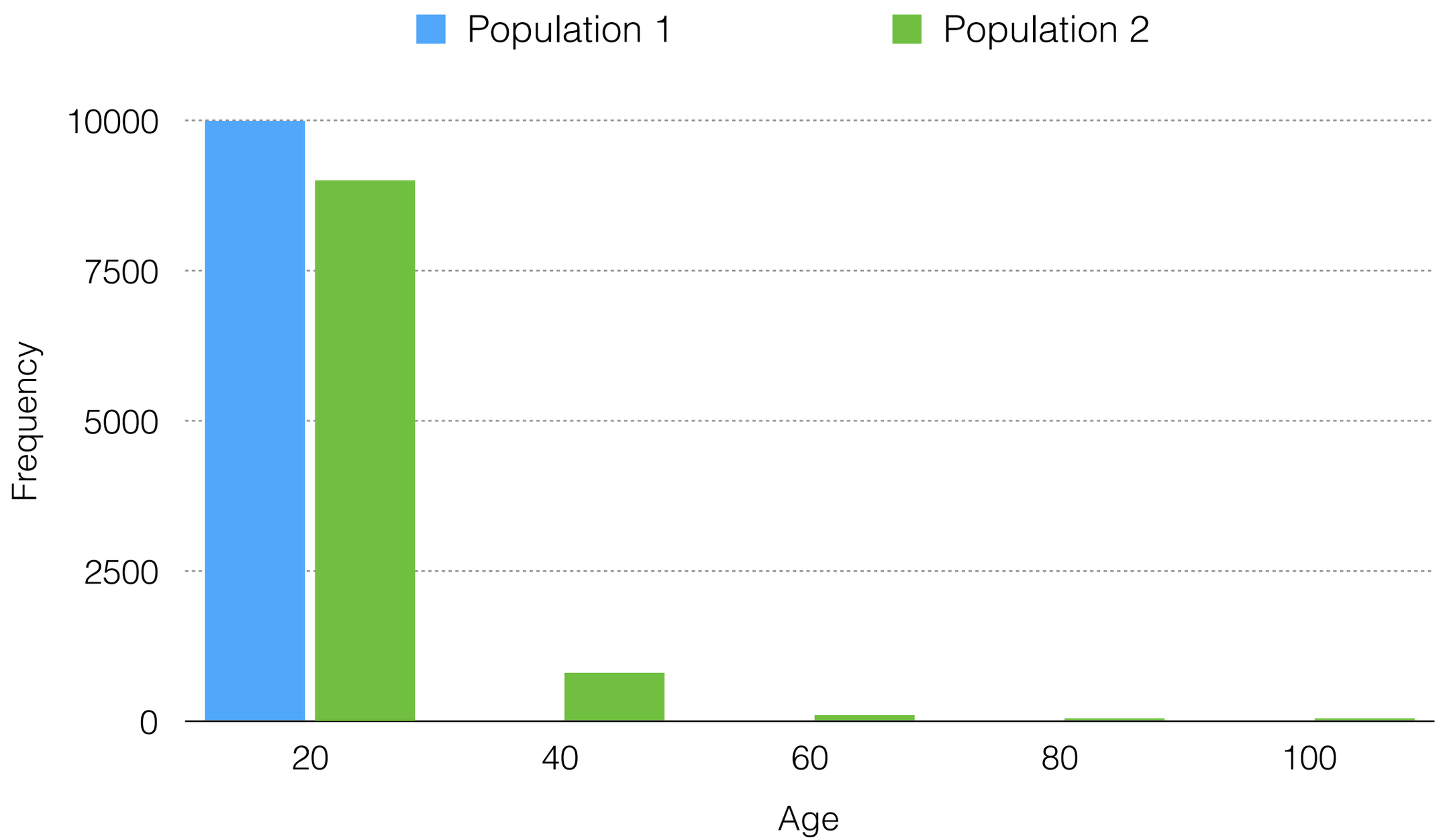
Missing or Cryptic Labels



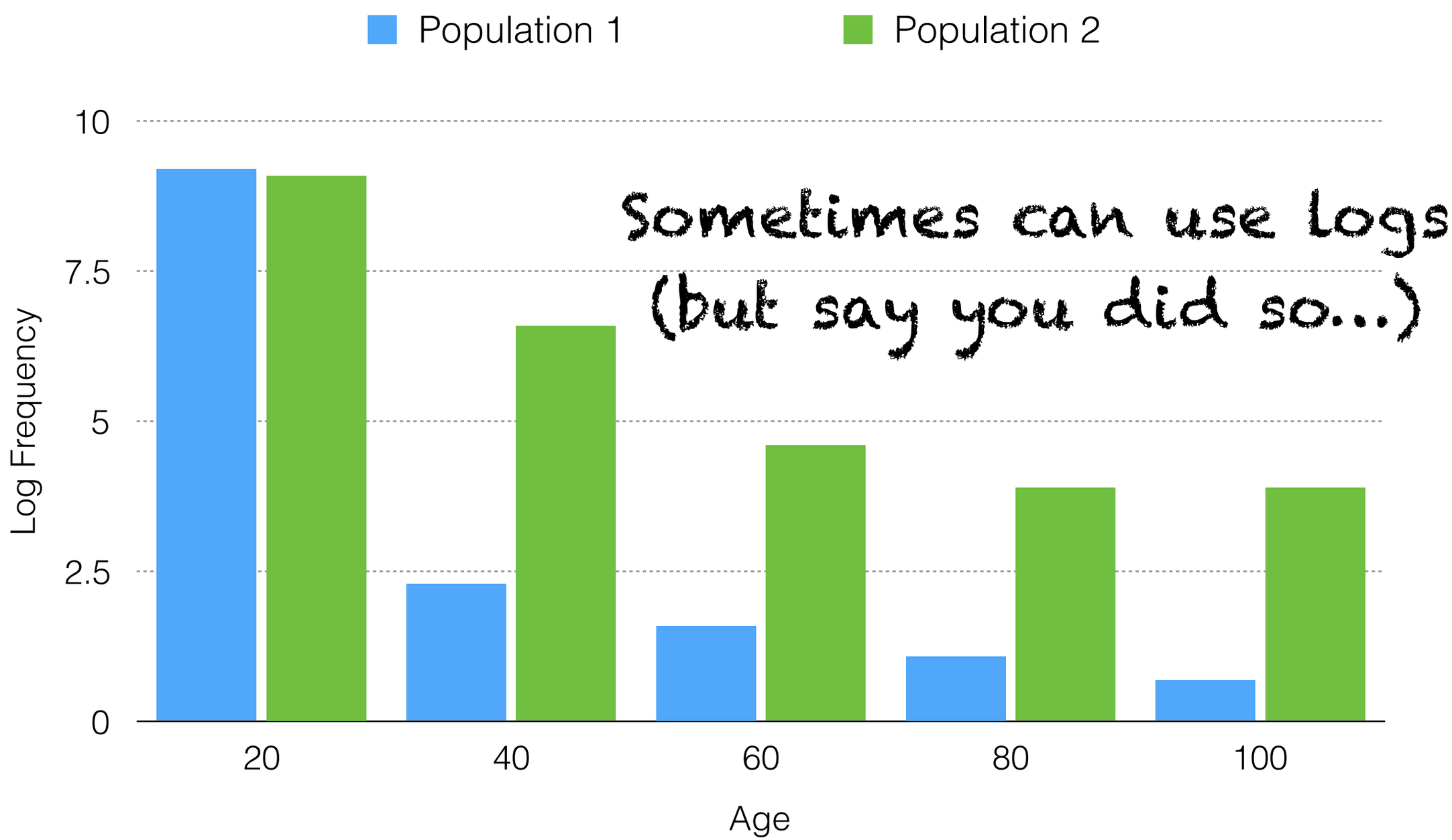
Missing or Cryptic Labels



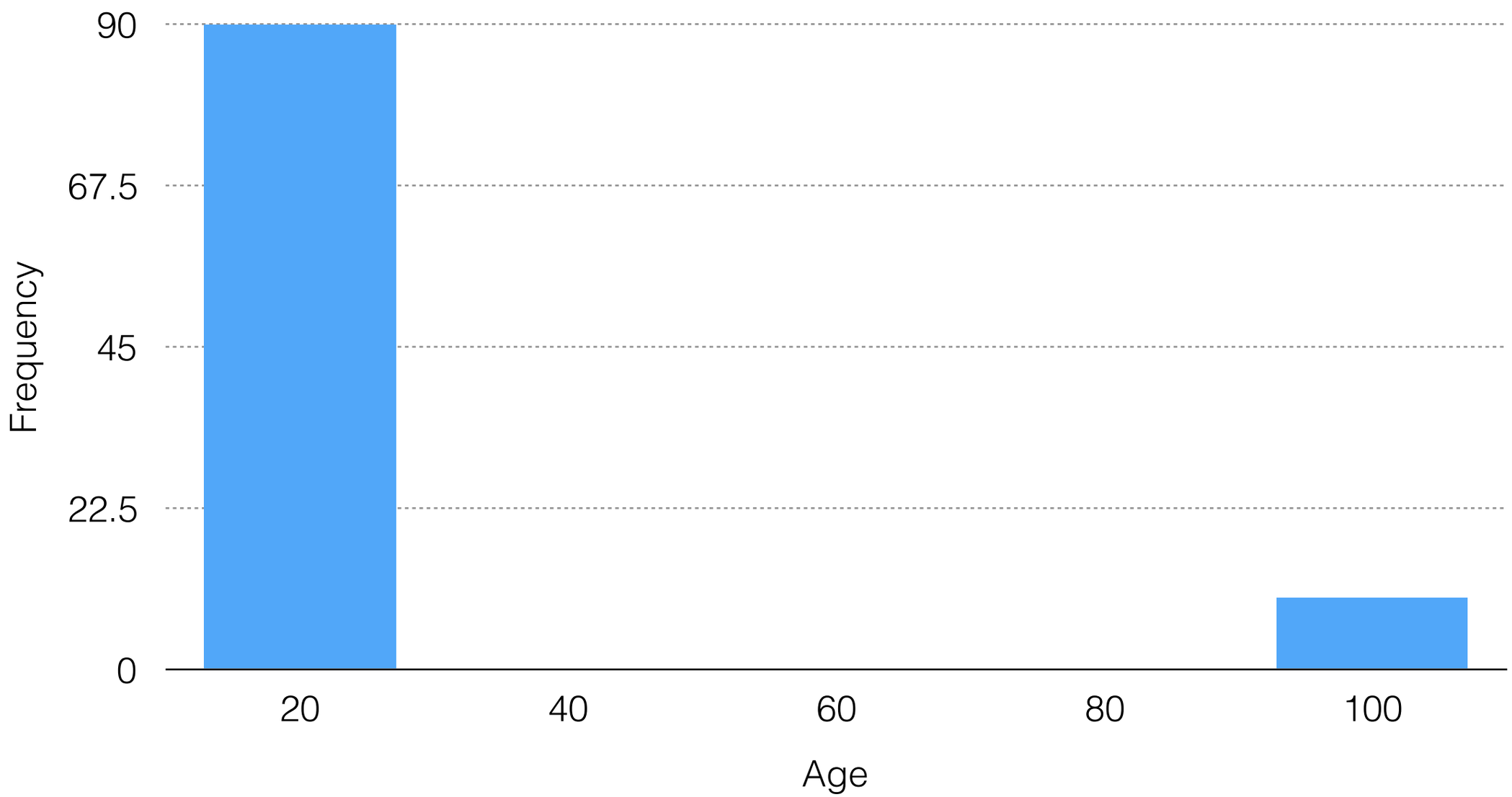
Skewed or Crunched Data



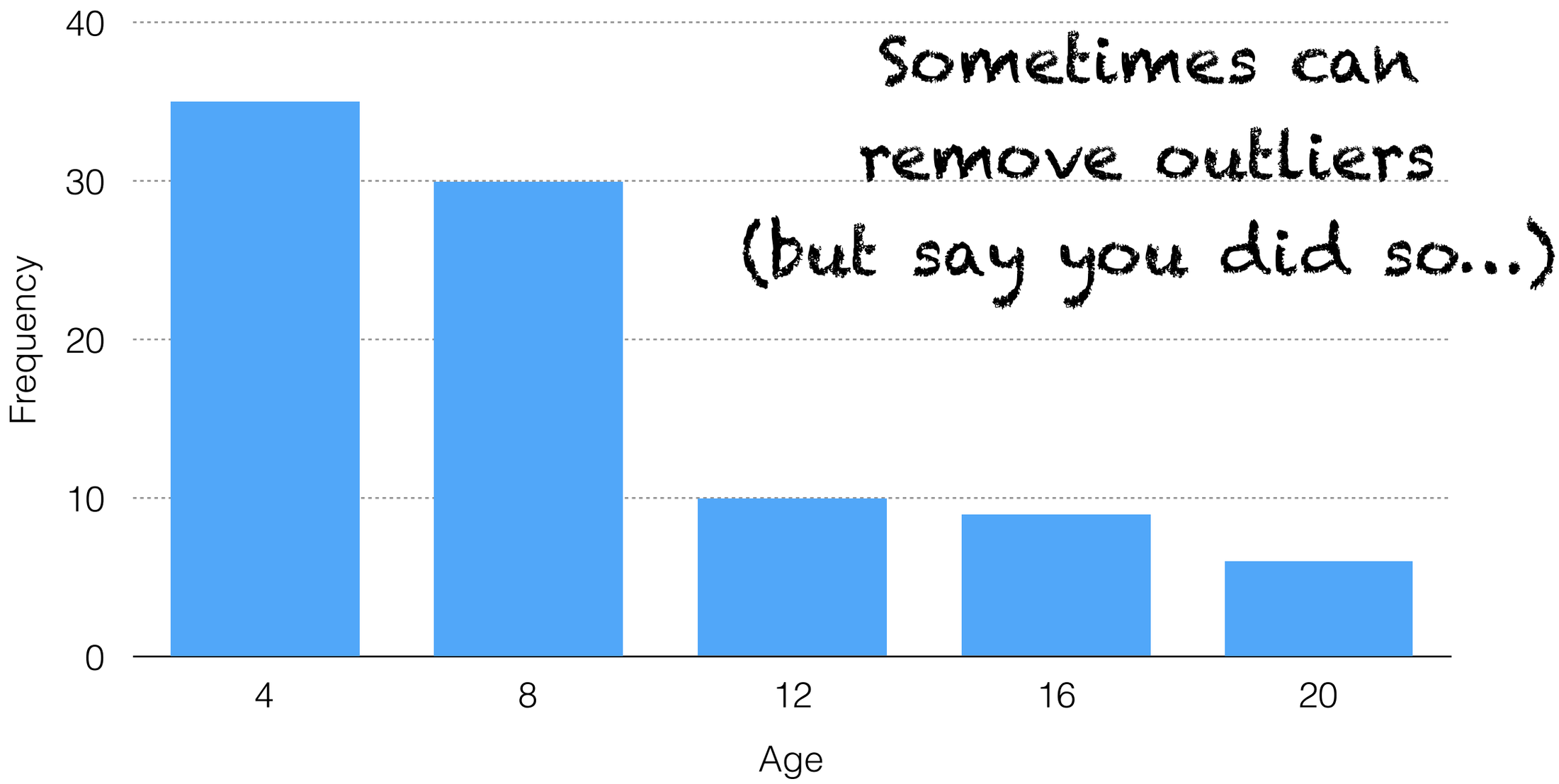
Skewed or Crunched Data



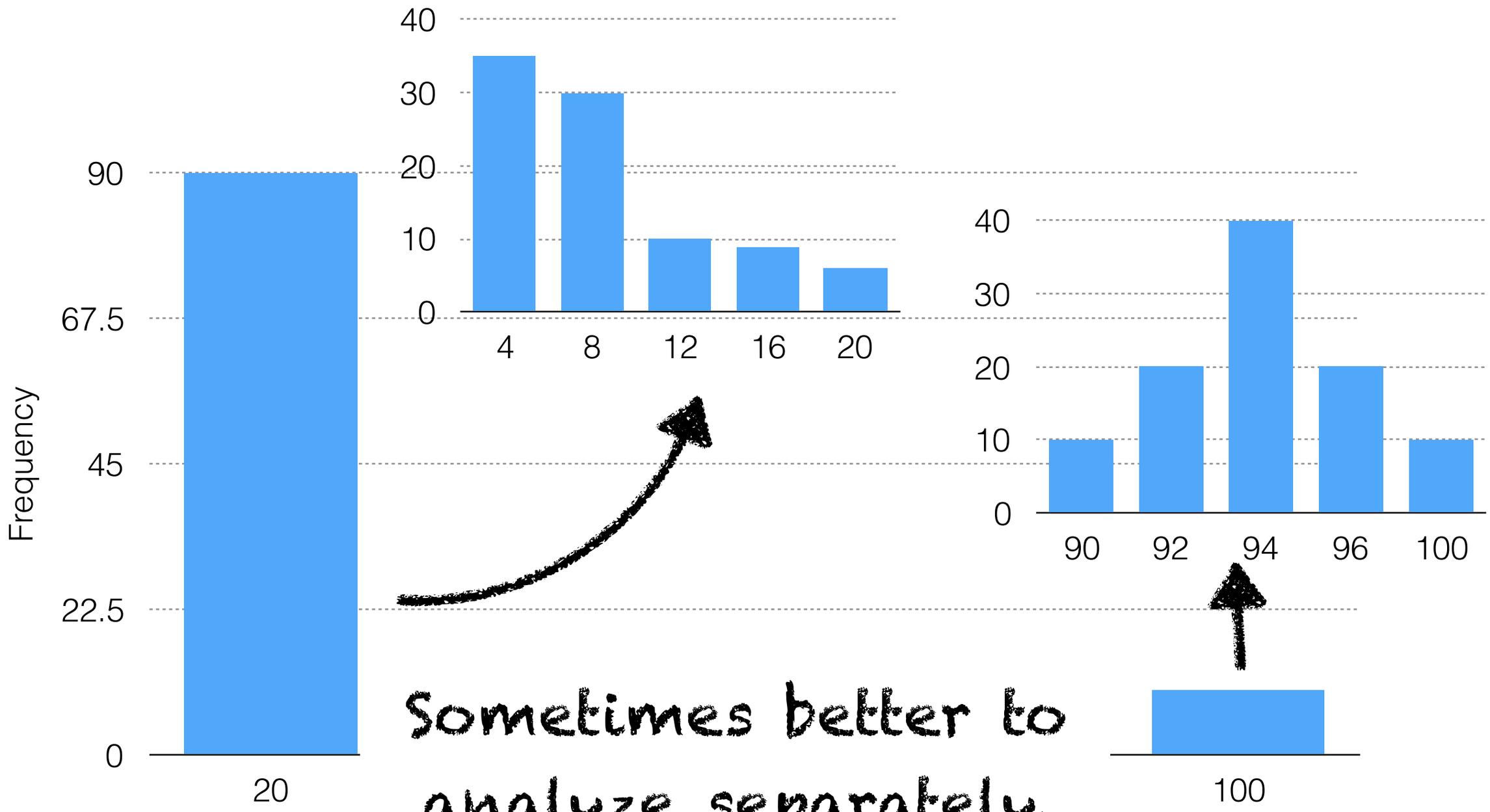
Skewed or Crunched Data



Skewed or Crunched Data

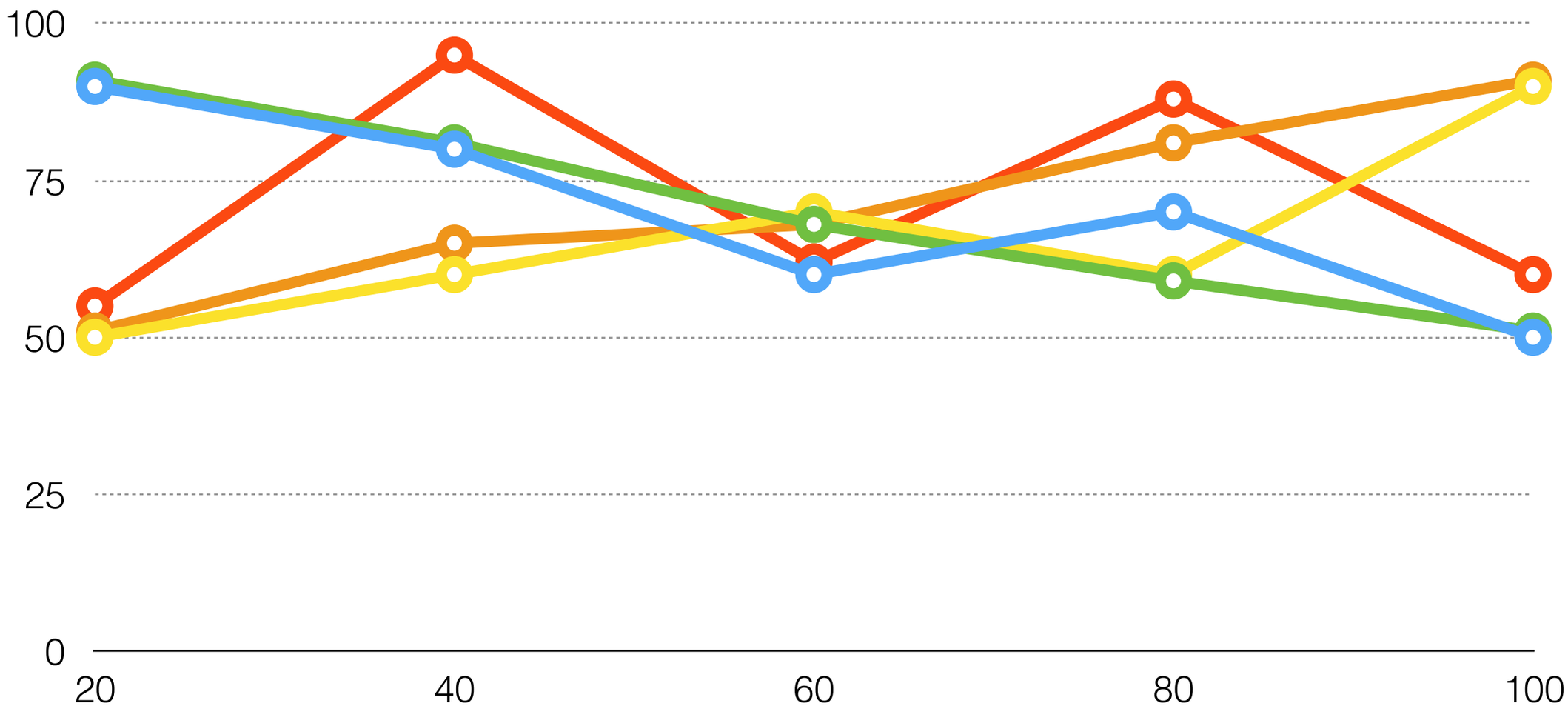


Skewed or Crunched Data

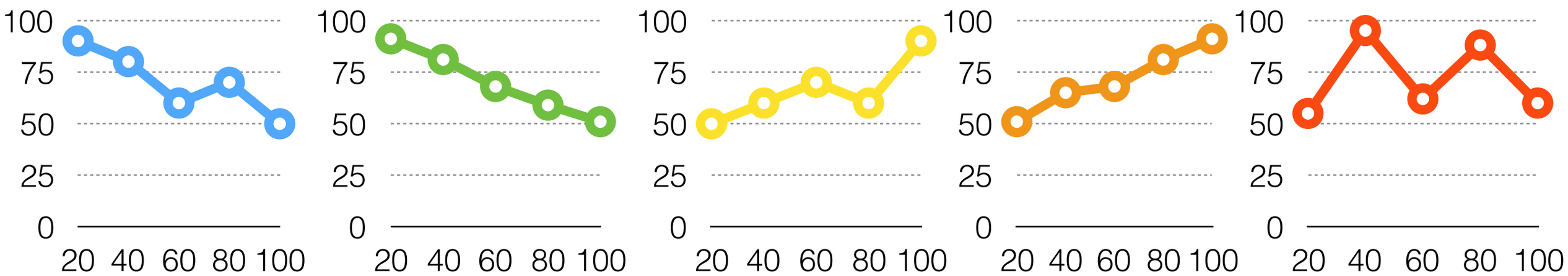


Sometimes better to analyze separately. (Look at your data!)

Skewed or Crunched Data

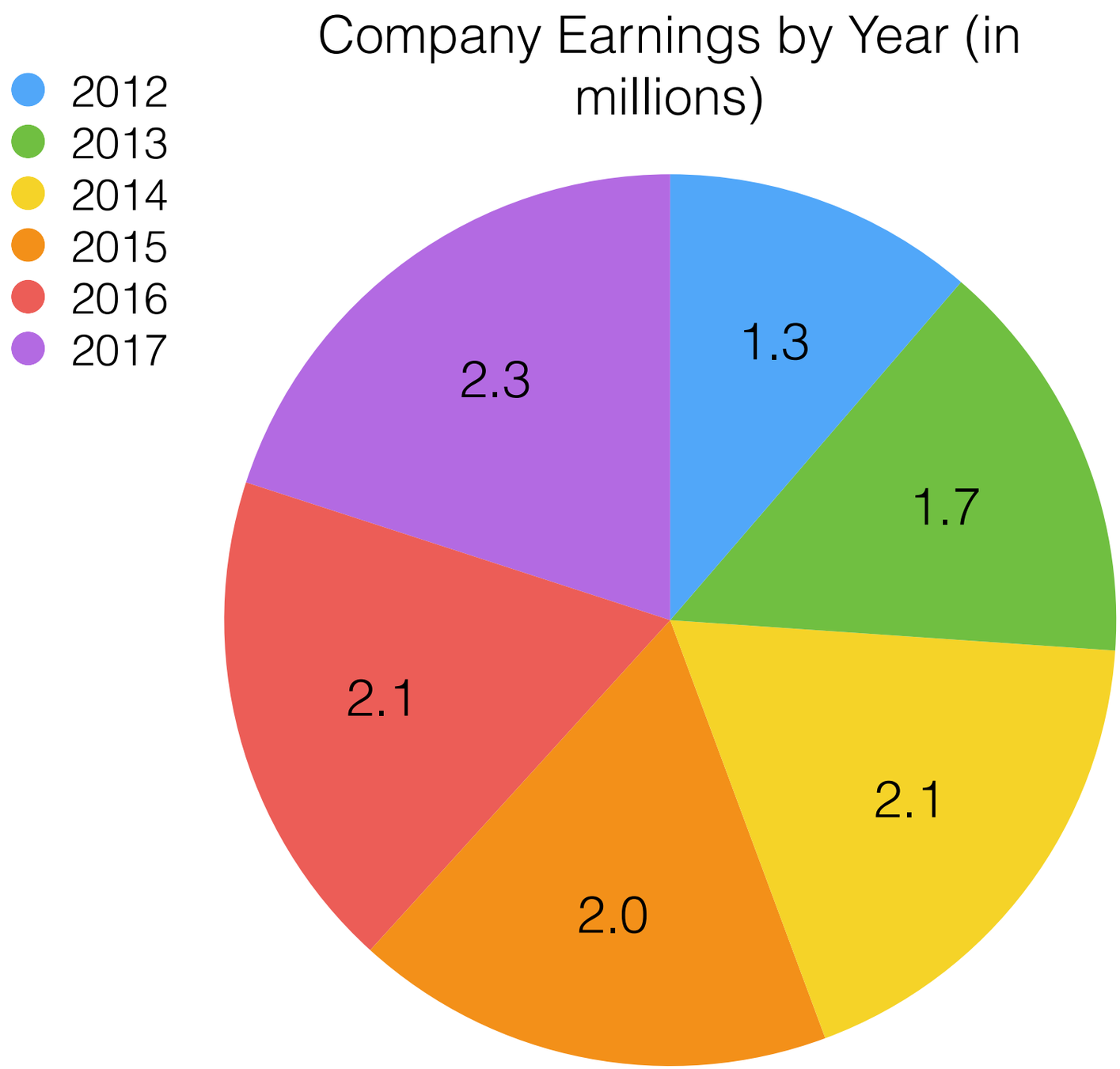


Skewed or Crunched Data

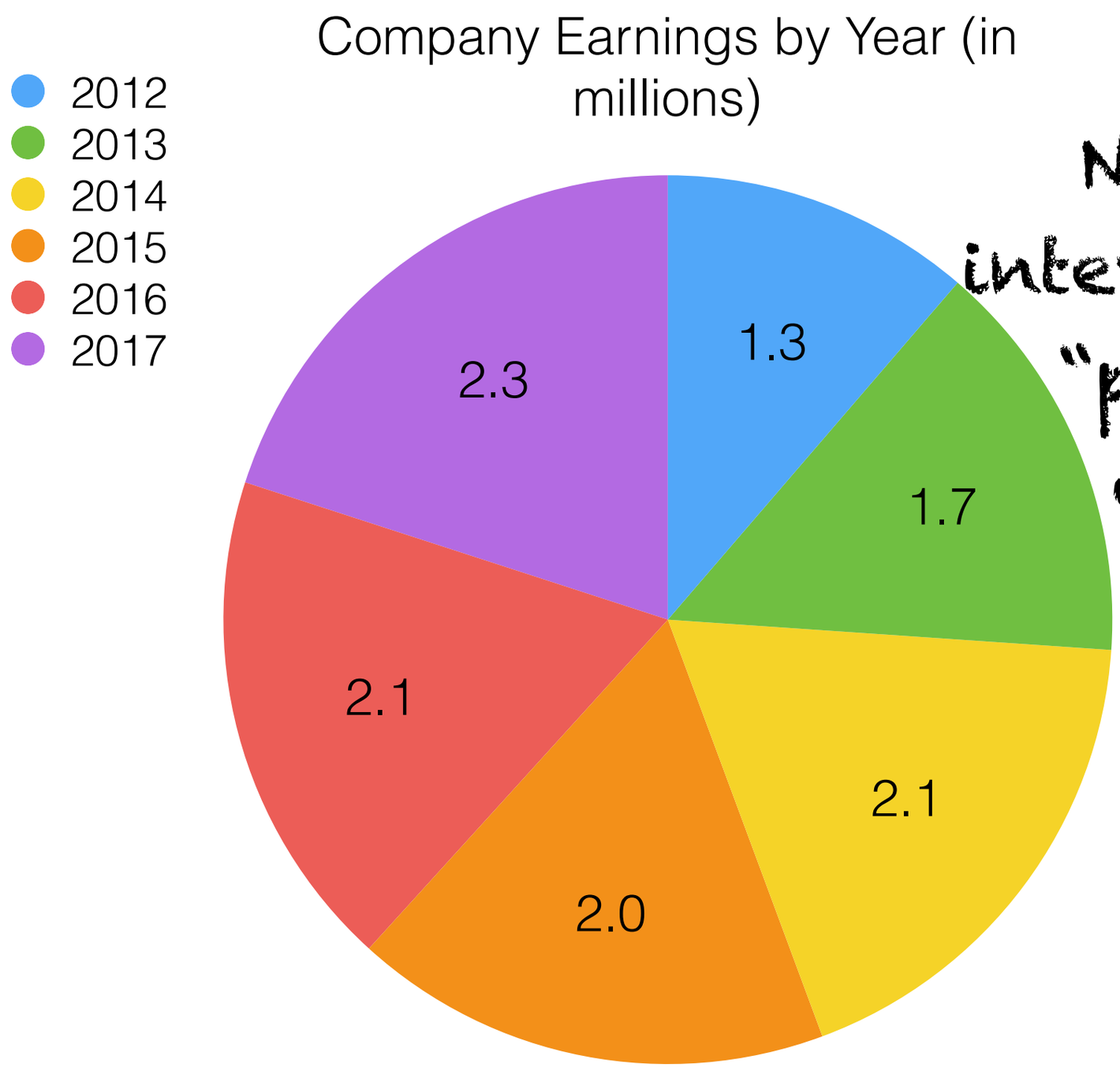


Sometimes better to split into multiple charts...

Chart/Data Type Mismatch

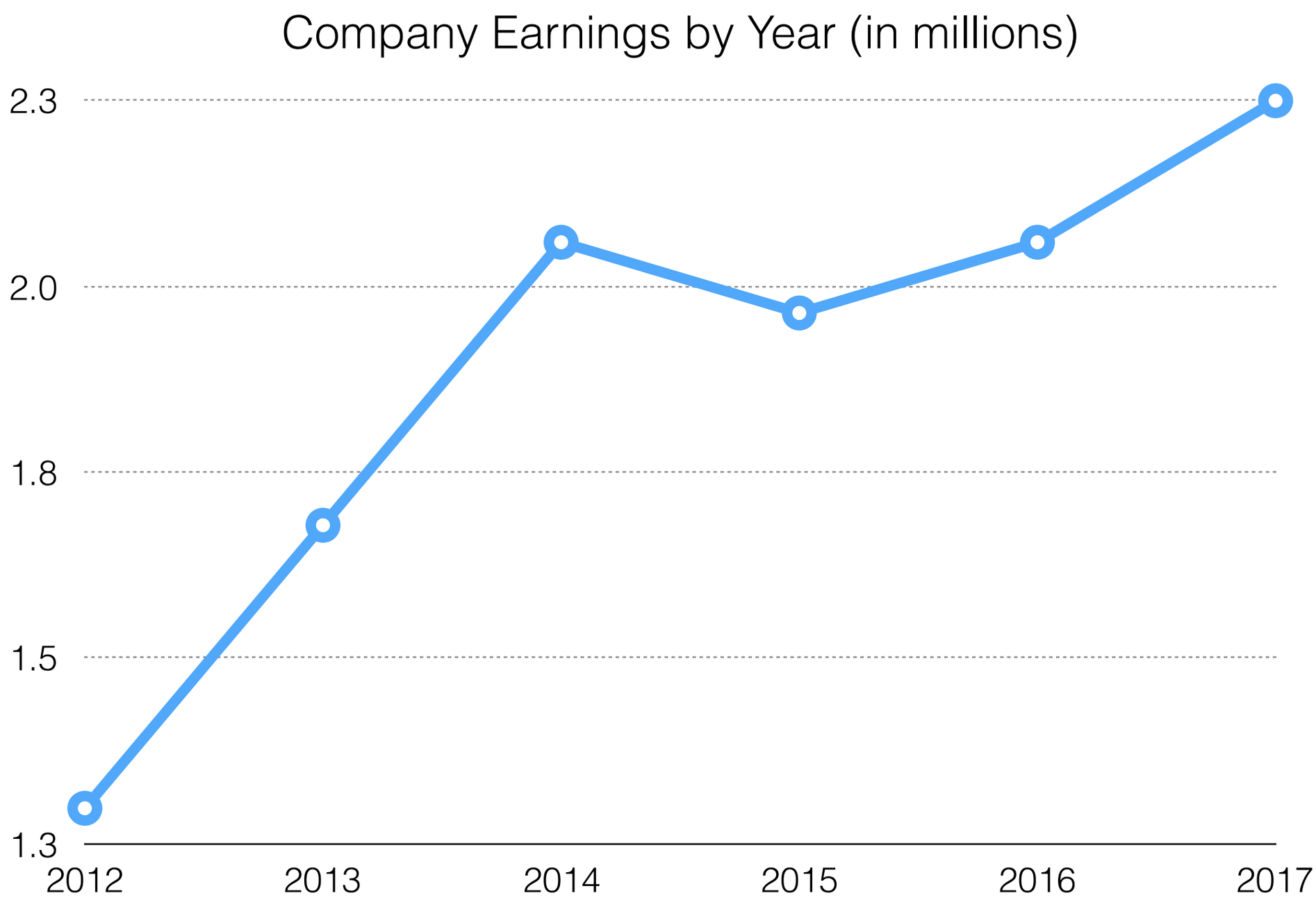


Chart/Data Type Mismatch

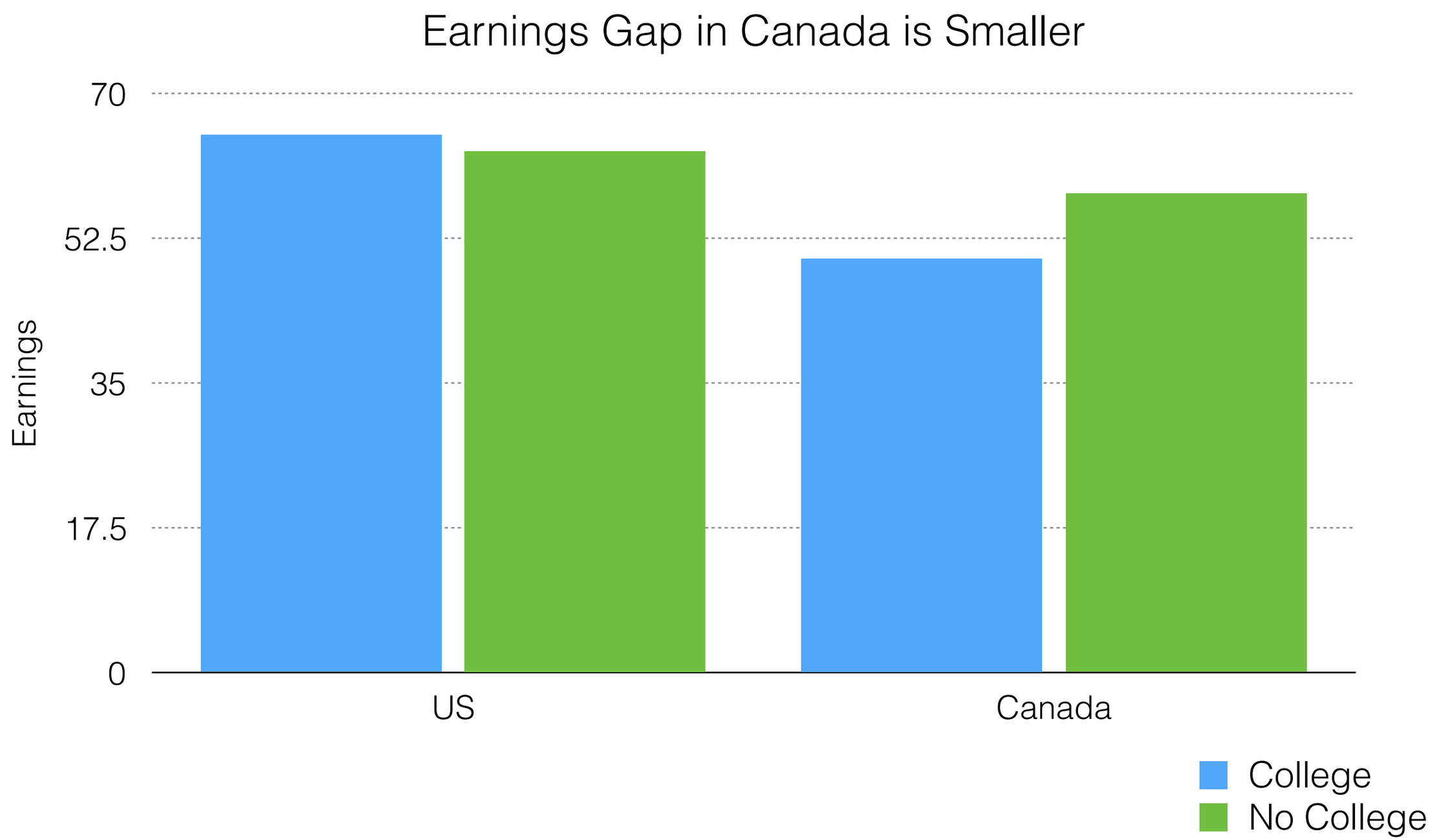


Not really interpretable as "parts of a whole"...

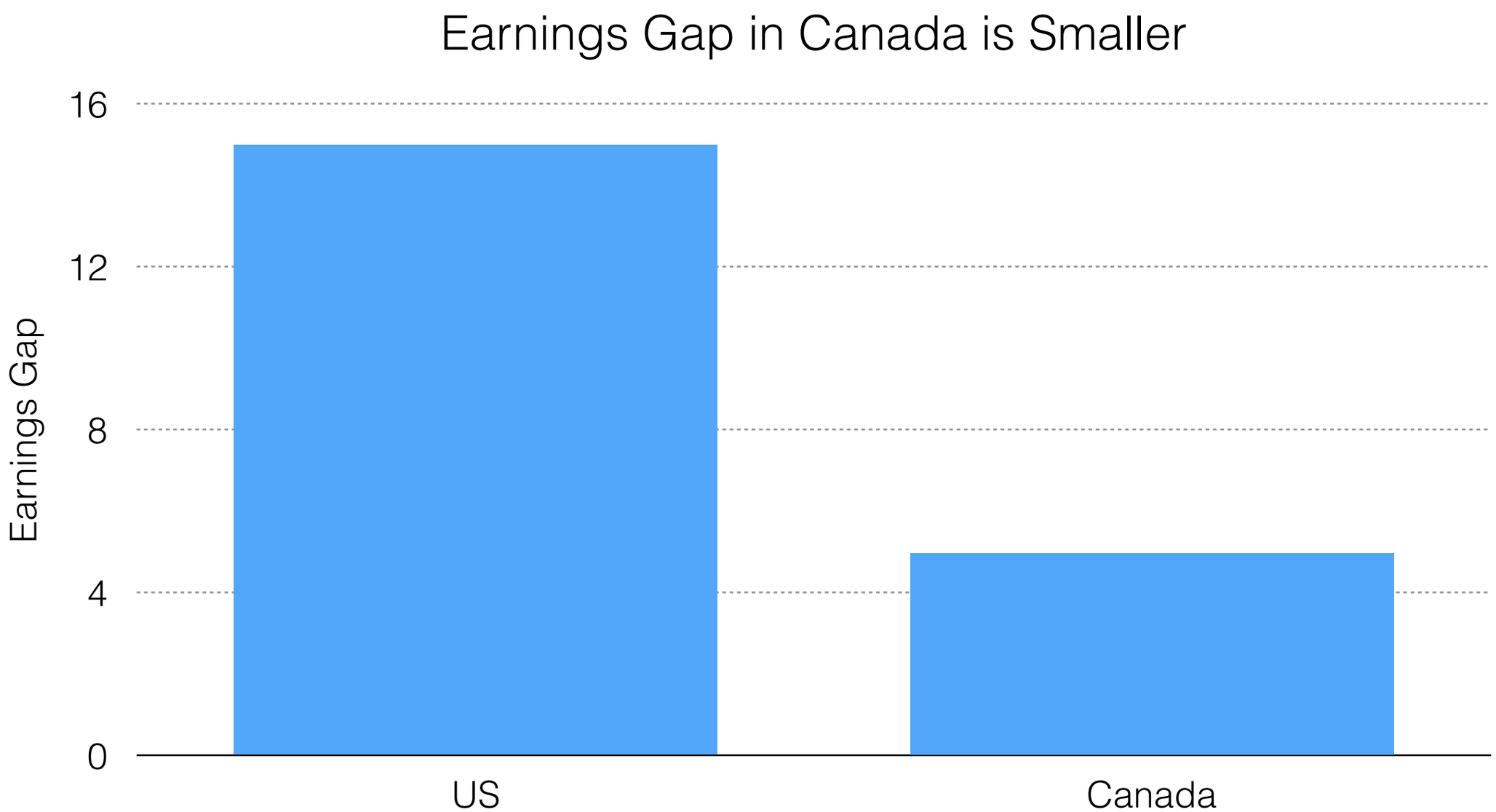
Chart/Data Type Mismatch



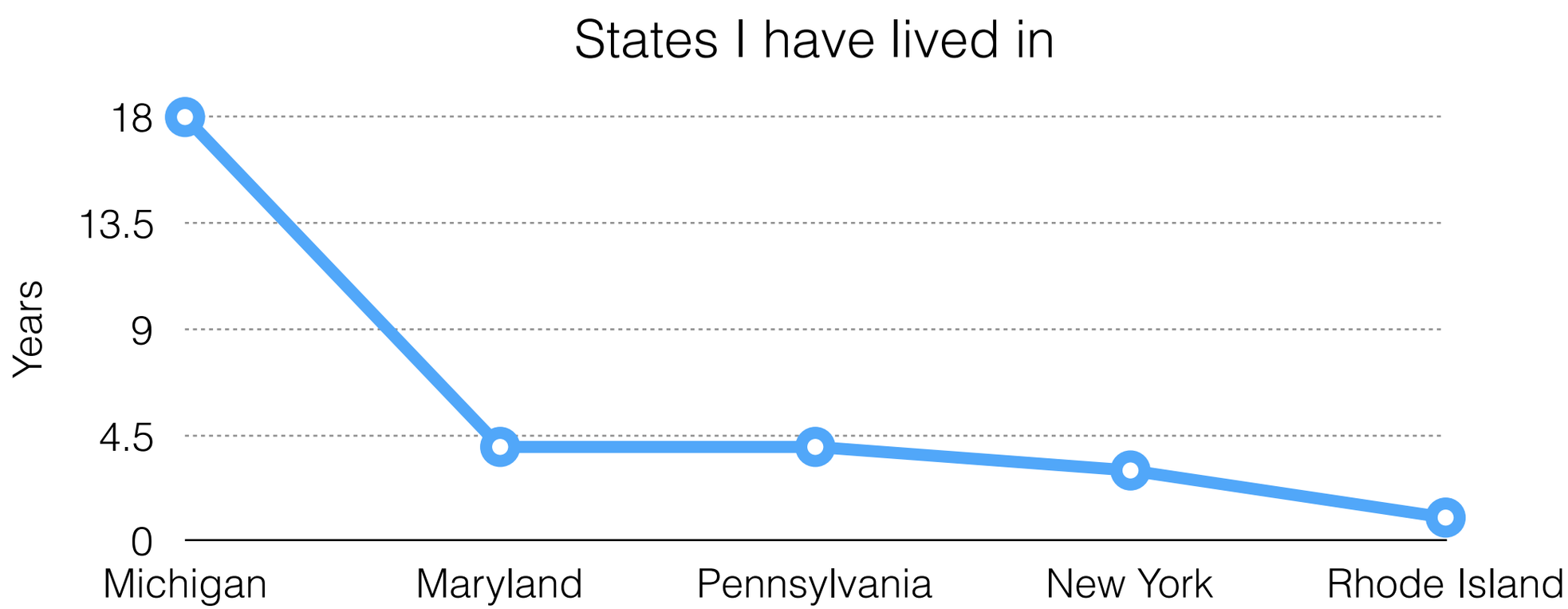
Chart/Data Type Mismatch



Chart/Data Type Mismatch



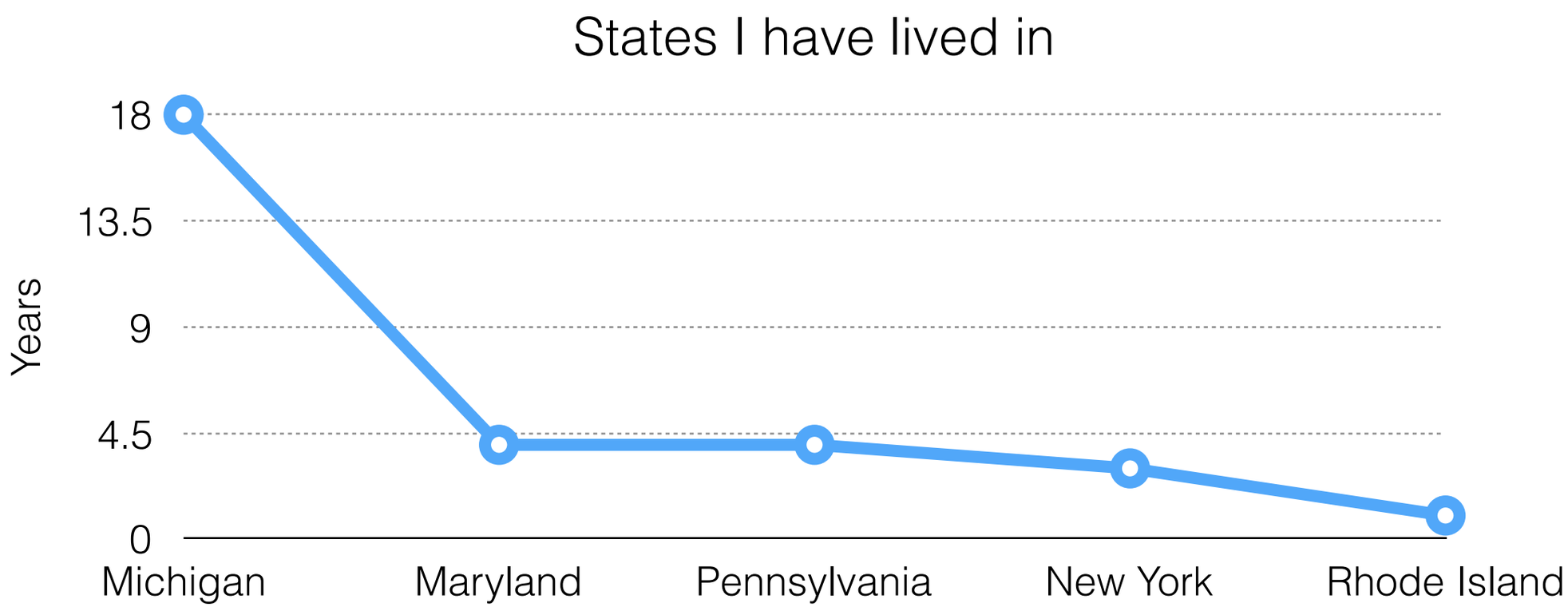
Clicker Question!



What is the biggest problem with this?

(a) Crunched/Skewed Data
(b) Missing/Cryptic Labels
(c) Chart/Data Type Mismatch
(d) Its just ugly

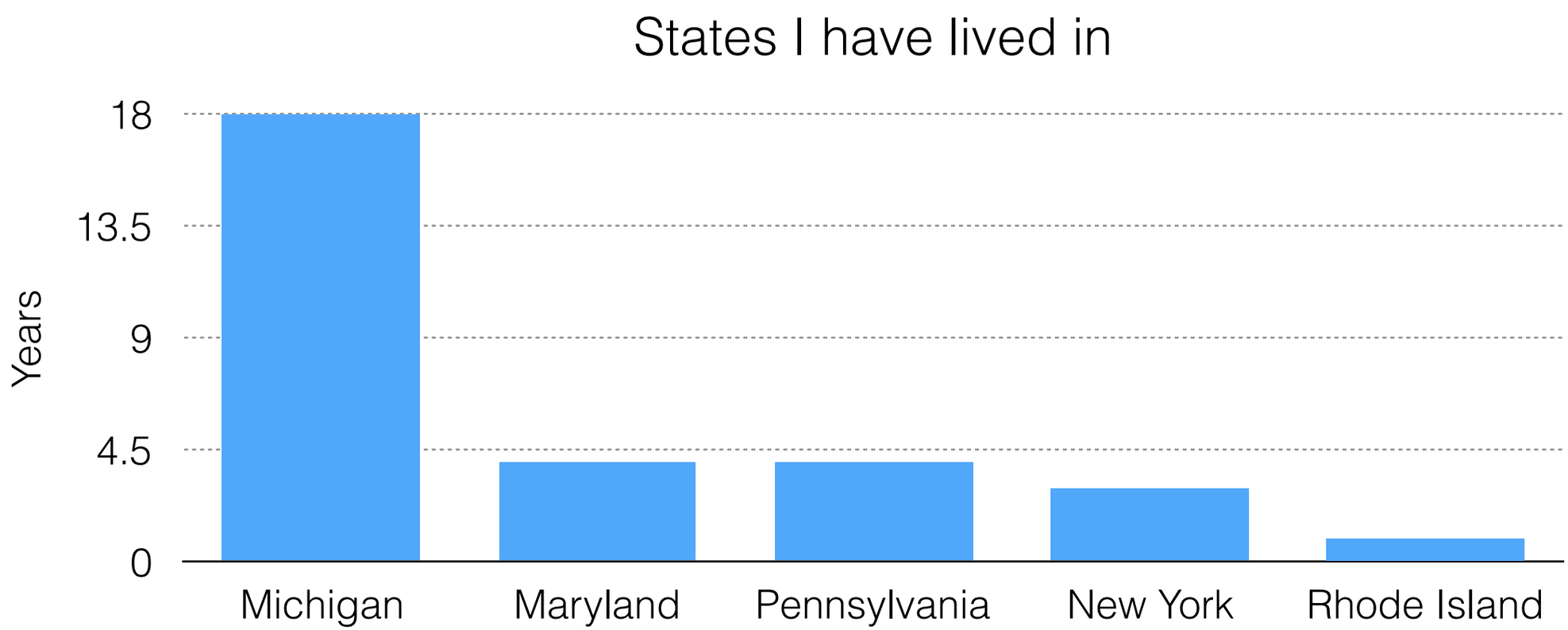
Clicker Question!



What is the biggest problem with this?

(a) Crunched/Skewed Data
(b) Missing/Cryptic Labels
(c) Chart/Data Type Mismatch
(d) Its just ugly

Clicker Question!



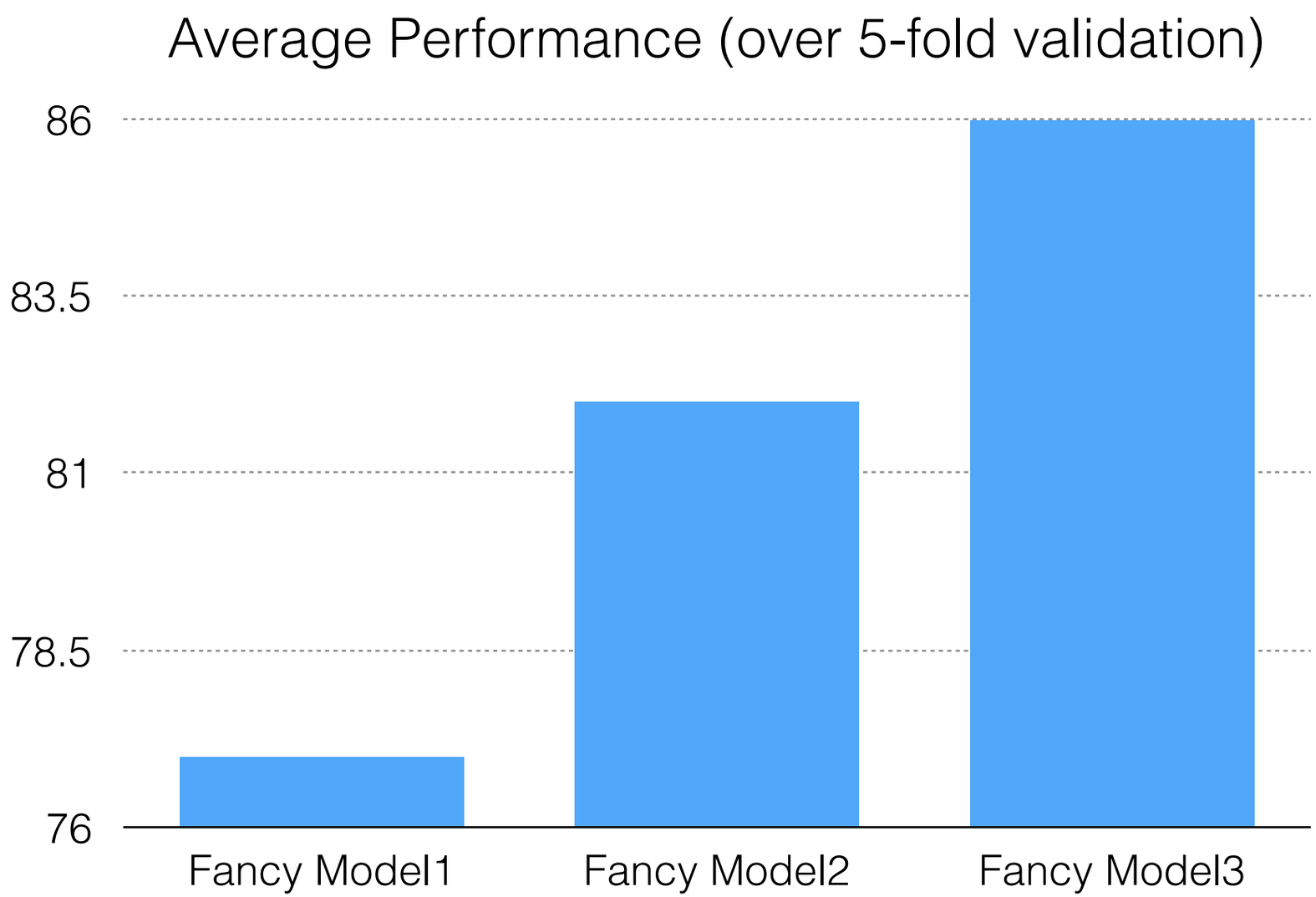
What is the biggest problem with this?

(a) Crunched/Skewed Data
(b) Missing/Cryptic Labels
(c) Chart/Data Type Mismatch
(d) Its just ugly

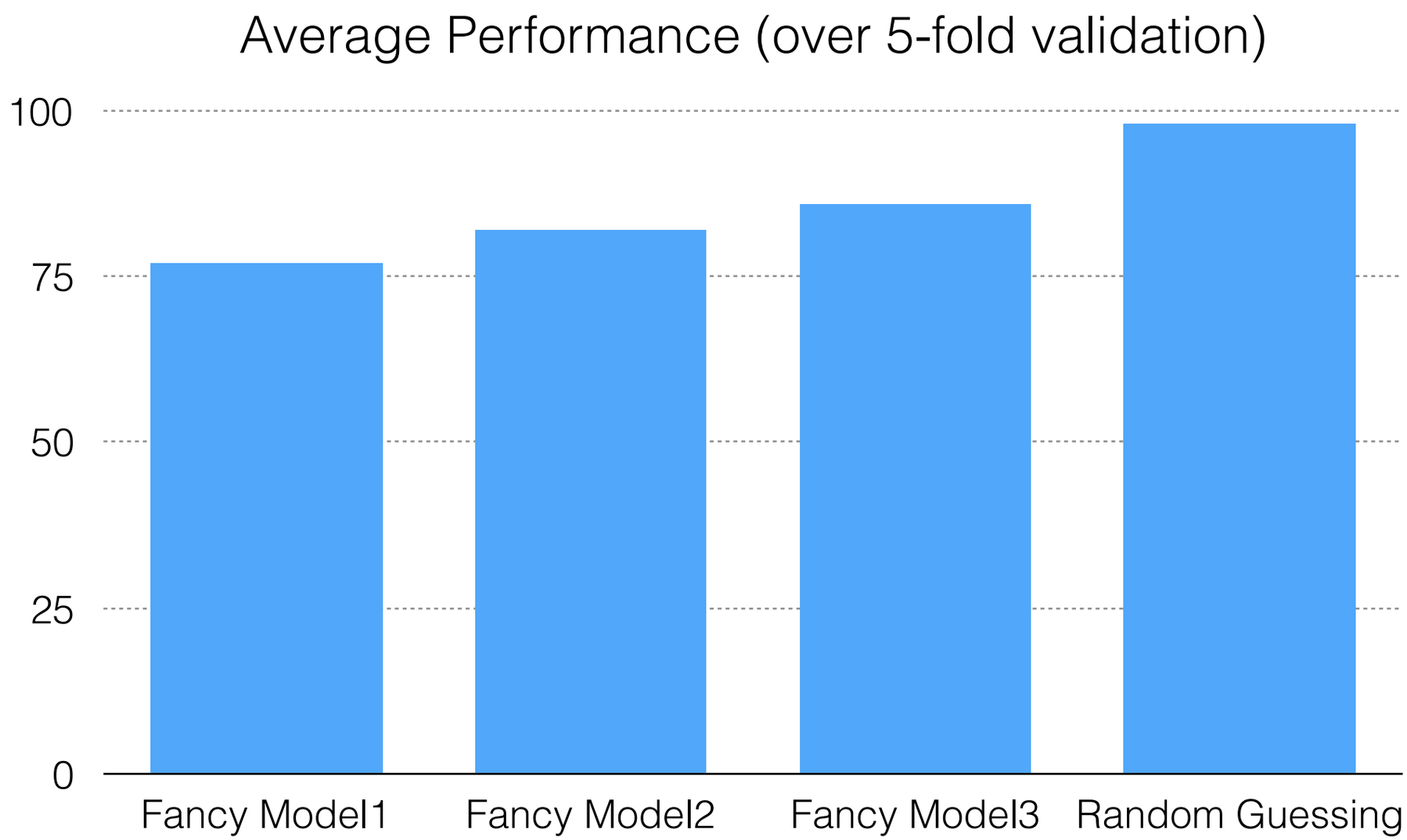
My “three pillars” of Data Viz

Don’t obfuscate the data or **H**ide the pr**O**cess you used to come to your co**N**clusions. Giv**E** people enough data **S**o that **T**hey can disagree with **Y**ou if they want to.

No Point of Comparison

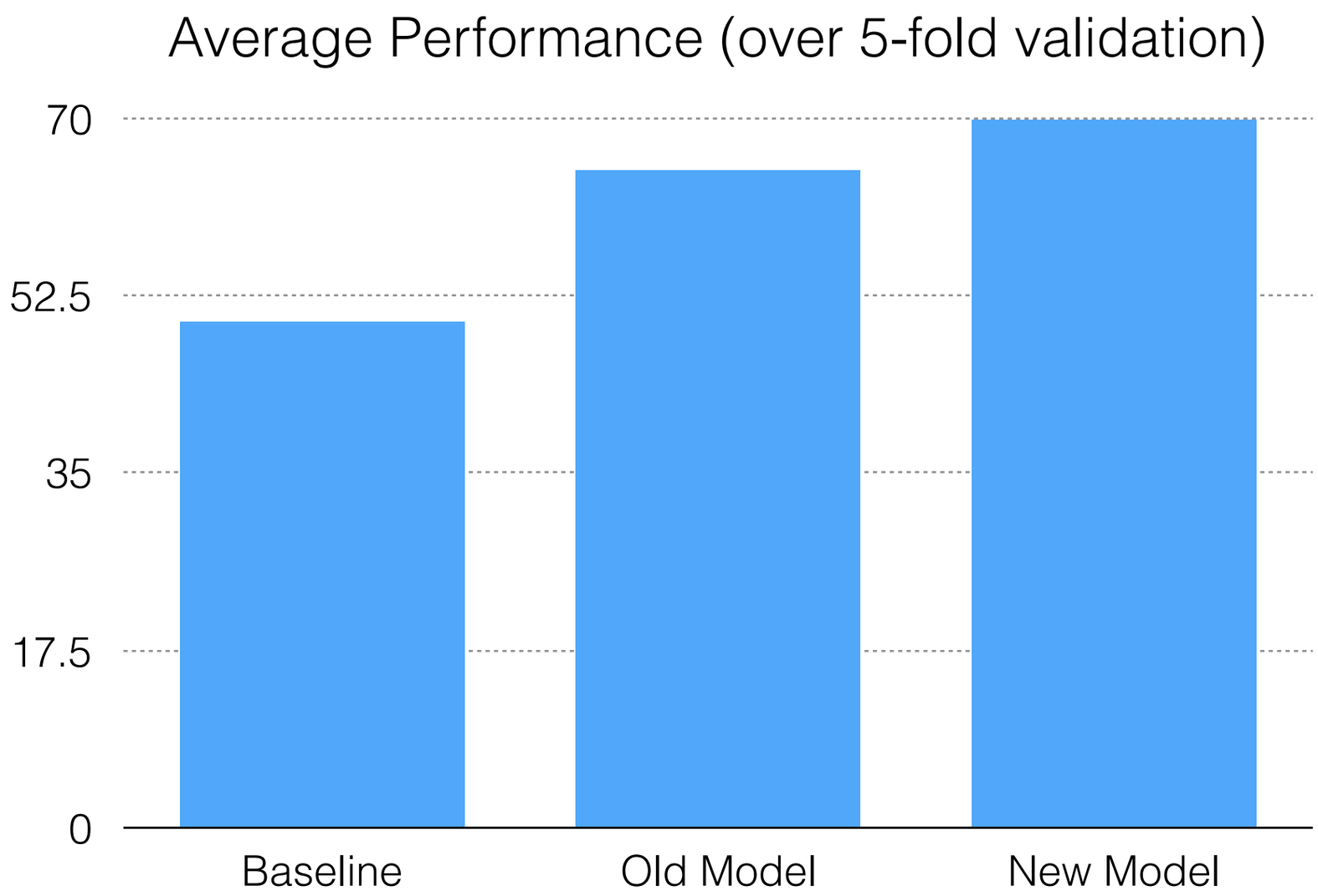


No Point of Comparison

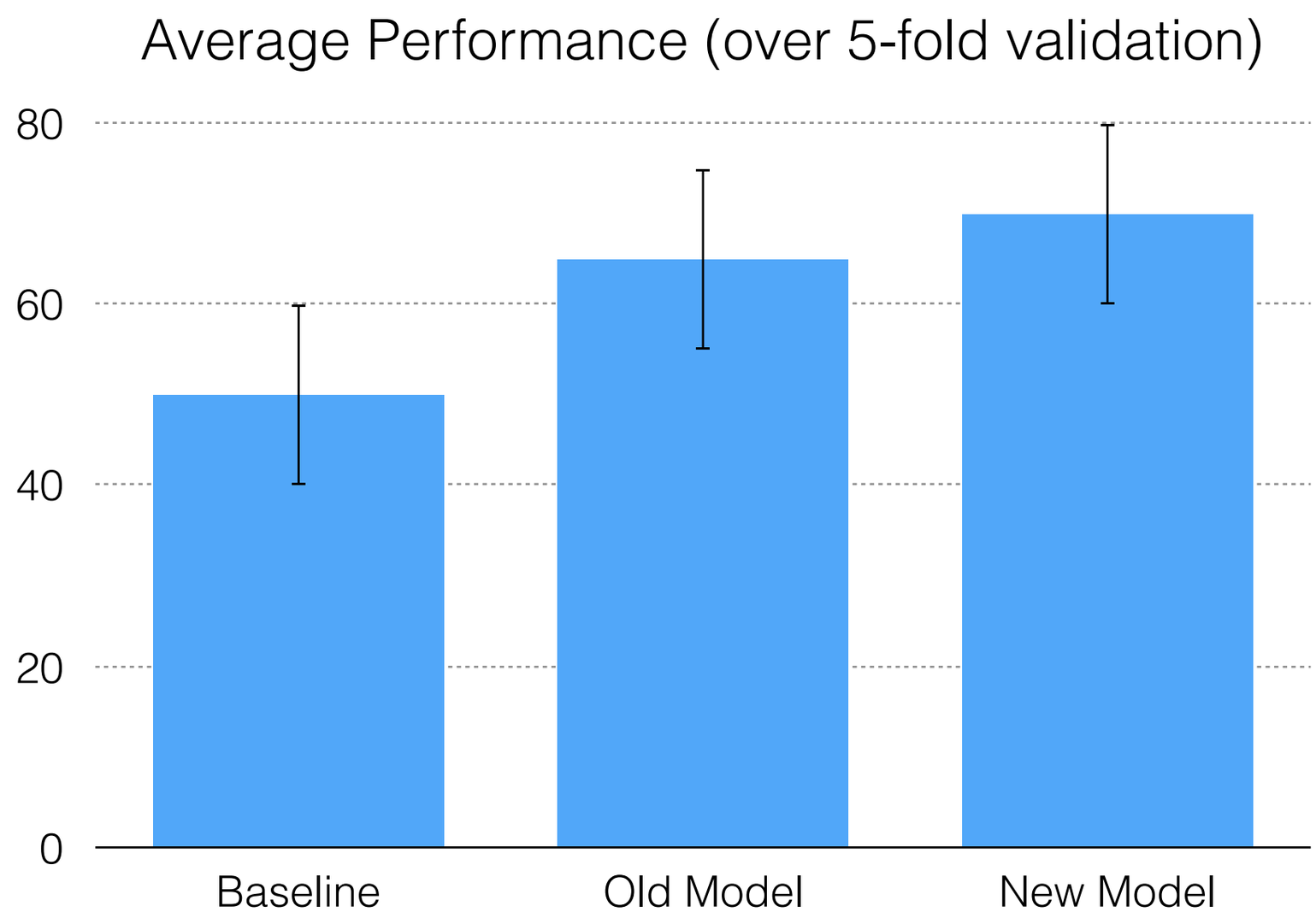


Help calibrate how easy/hard the problem is, what types of numbers to expect a priori...

Summary Stats Only

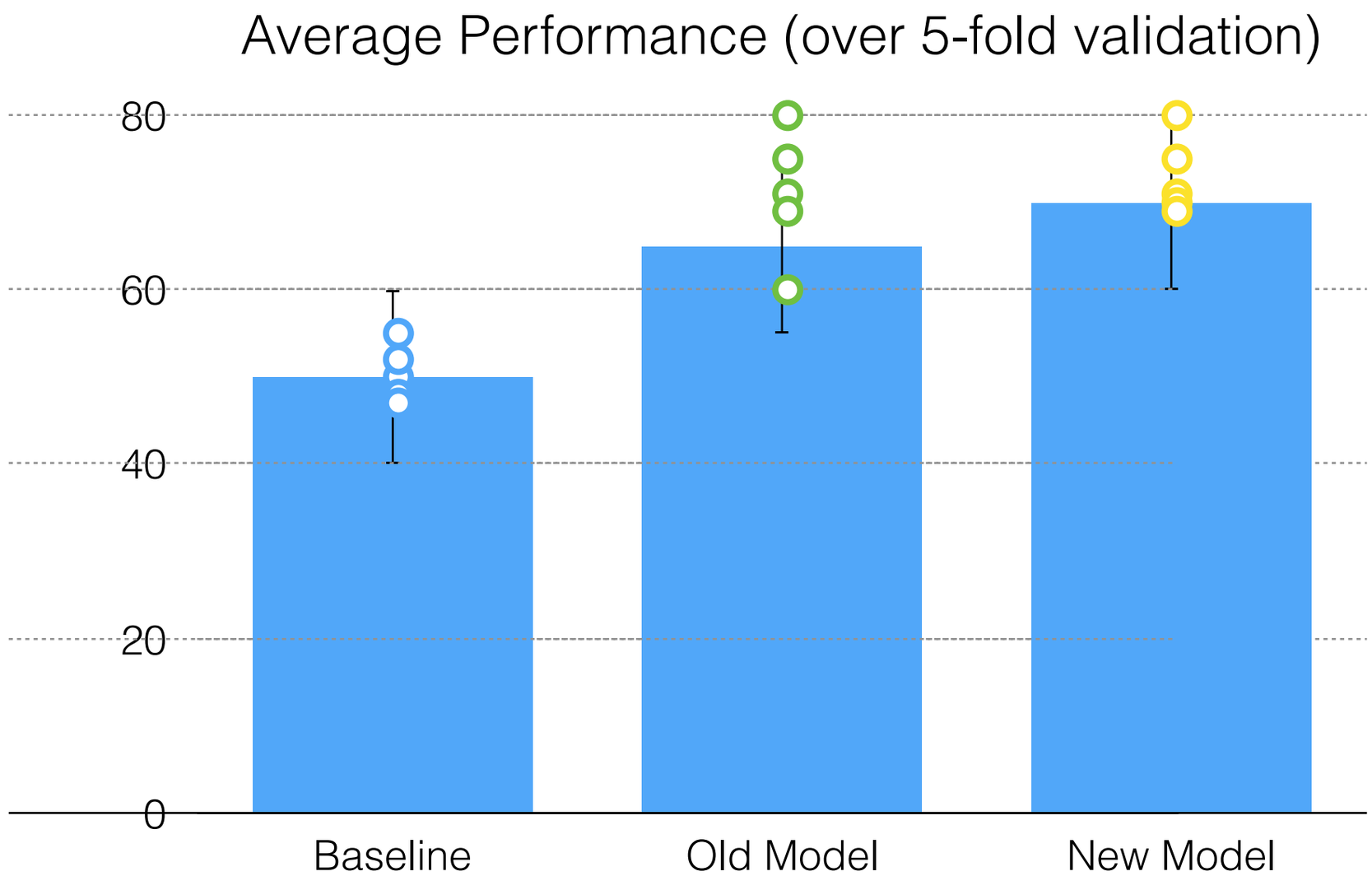


Summary Stats Only



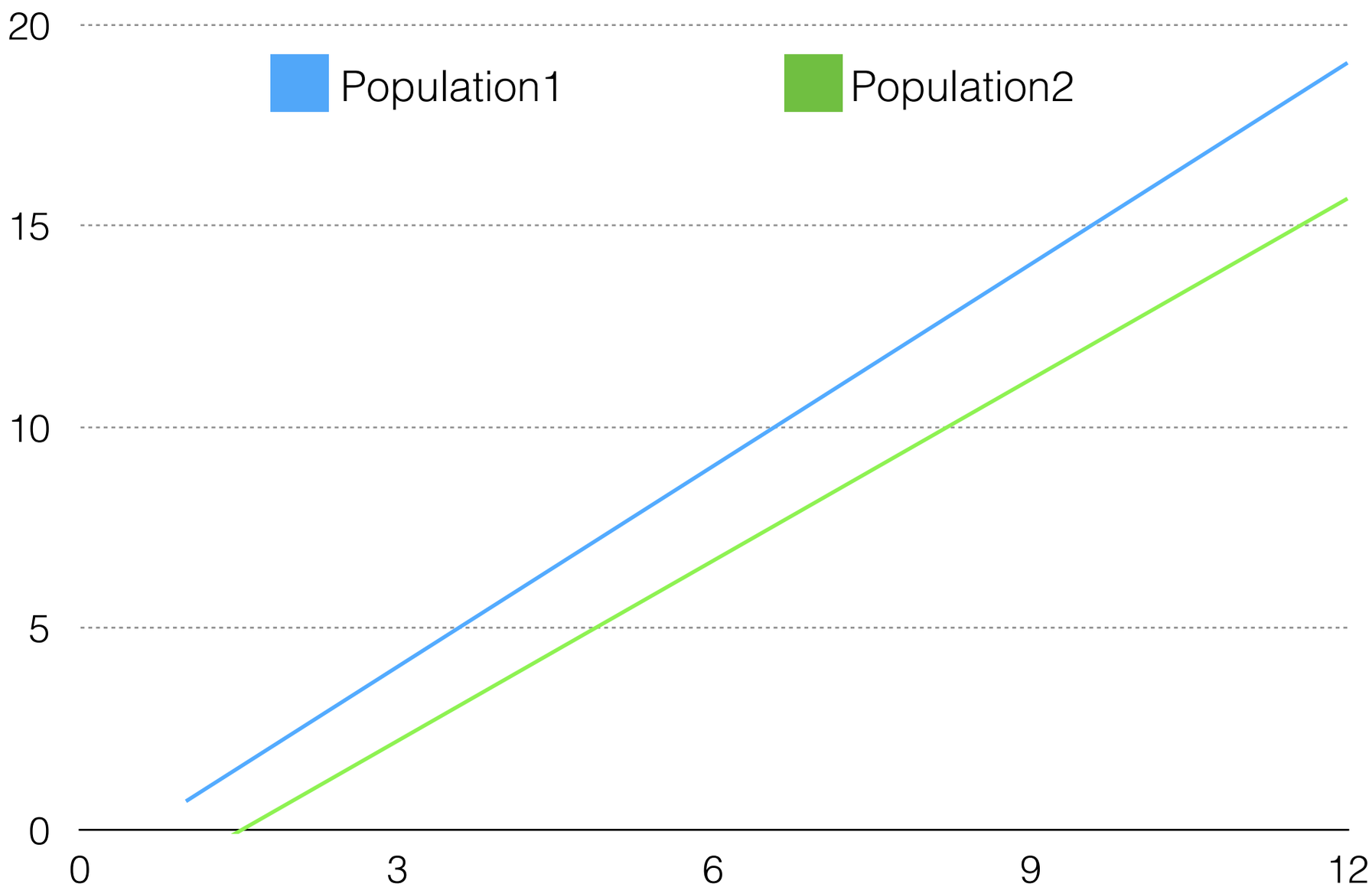
Sometimes can include error bars/ confidence intervals...

Summary Stats Only

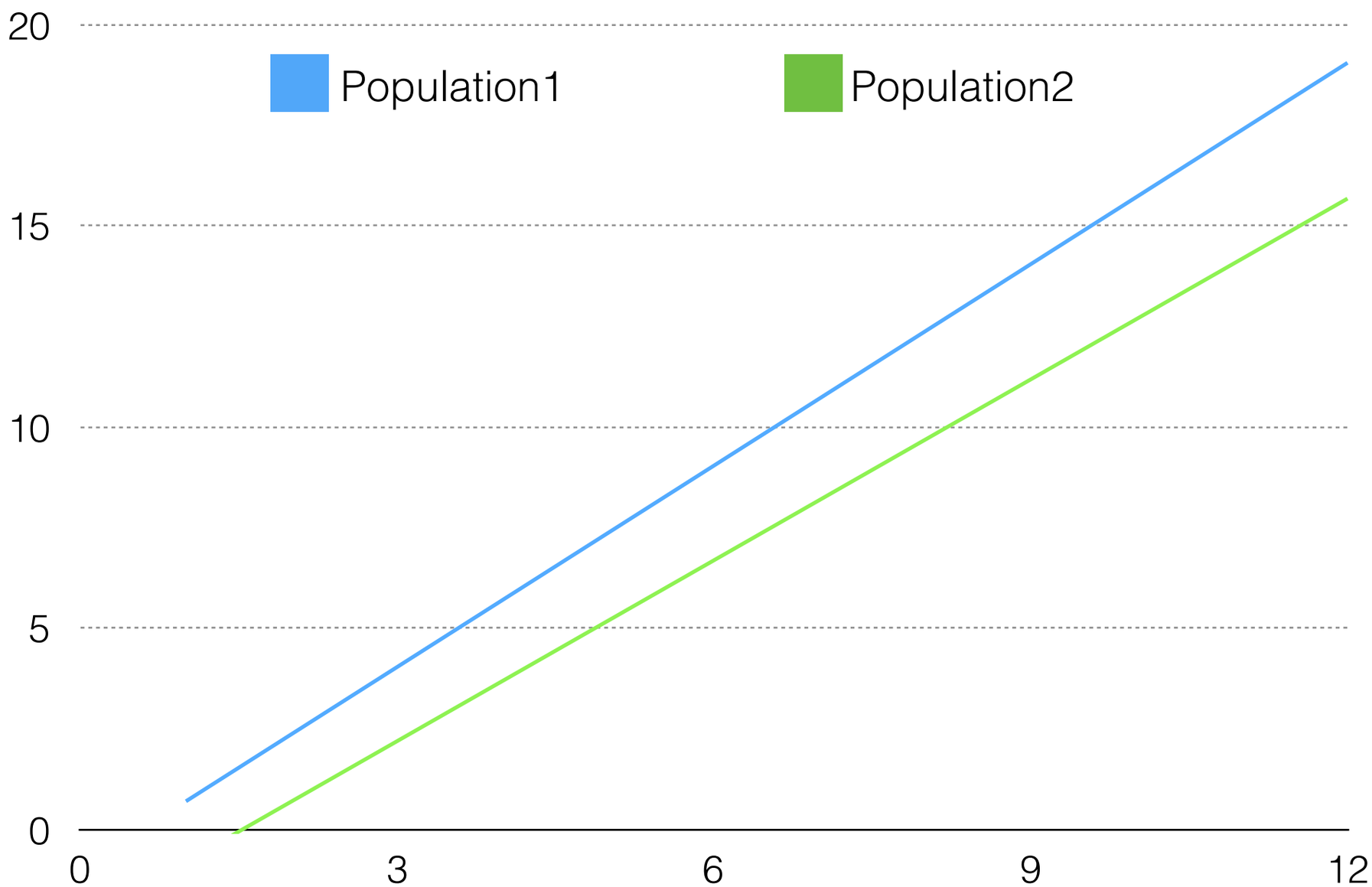


Even better, just show all the data alongside the summary stats.

Summary Stats Only

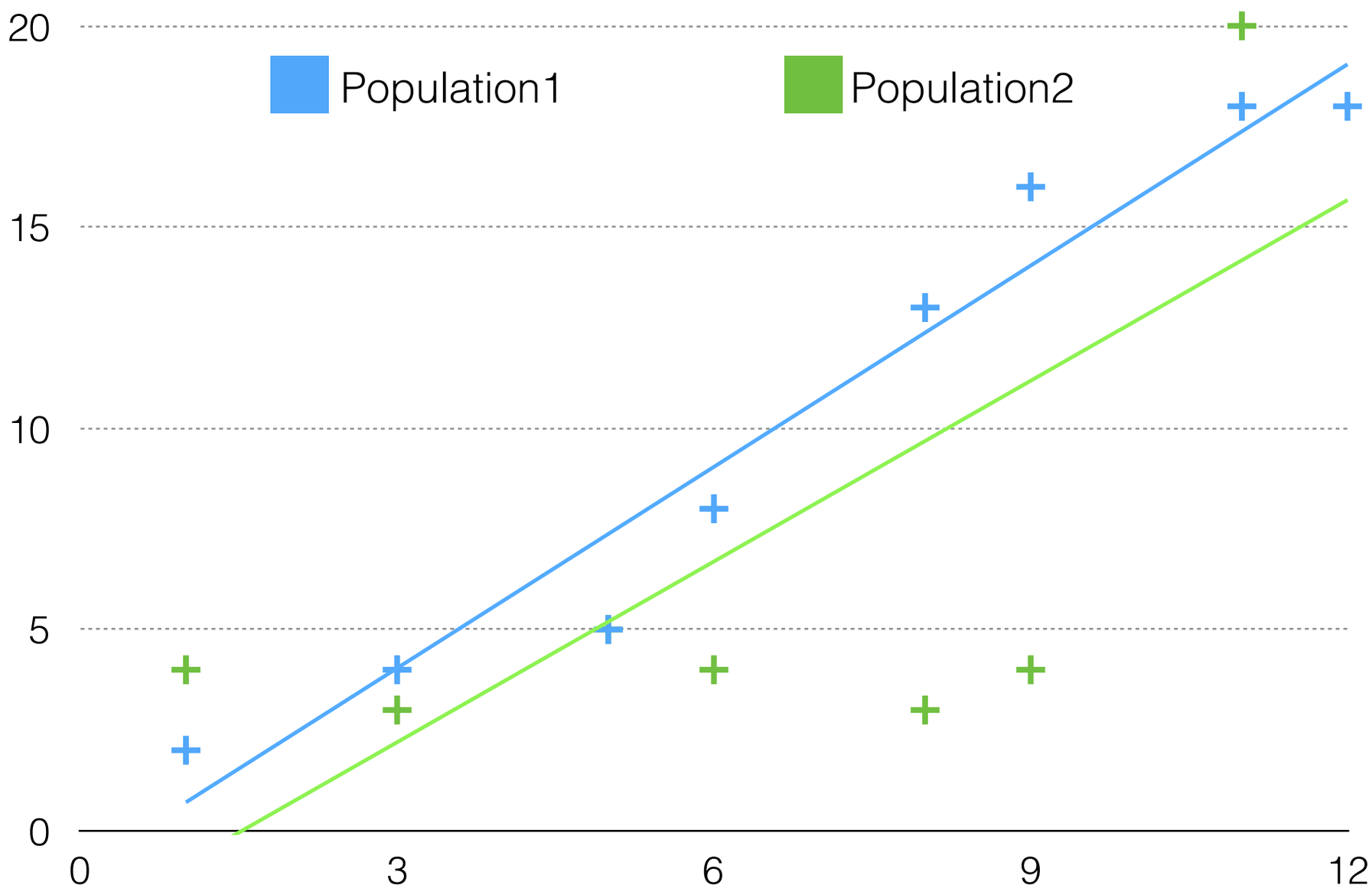


Summary Stats Only



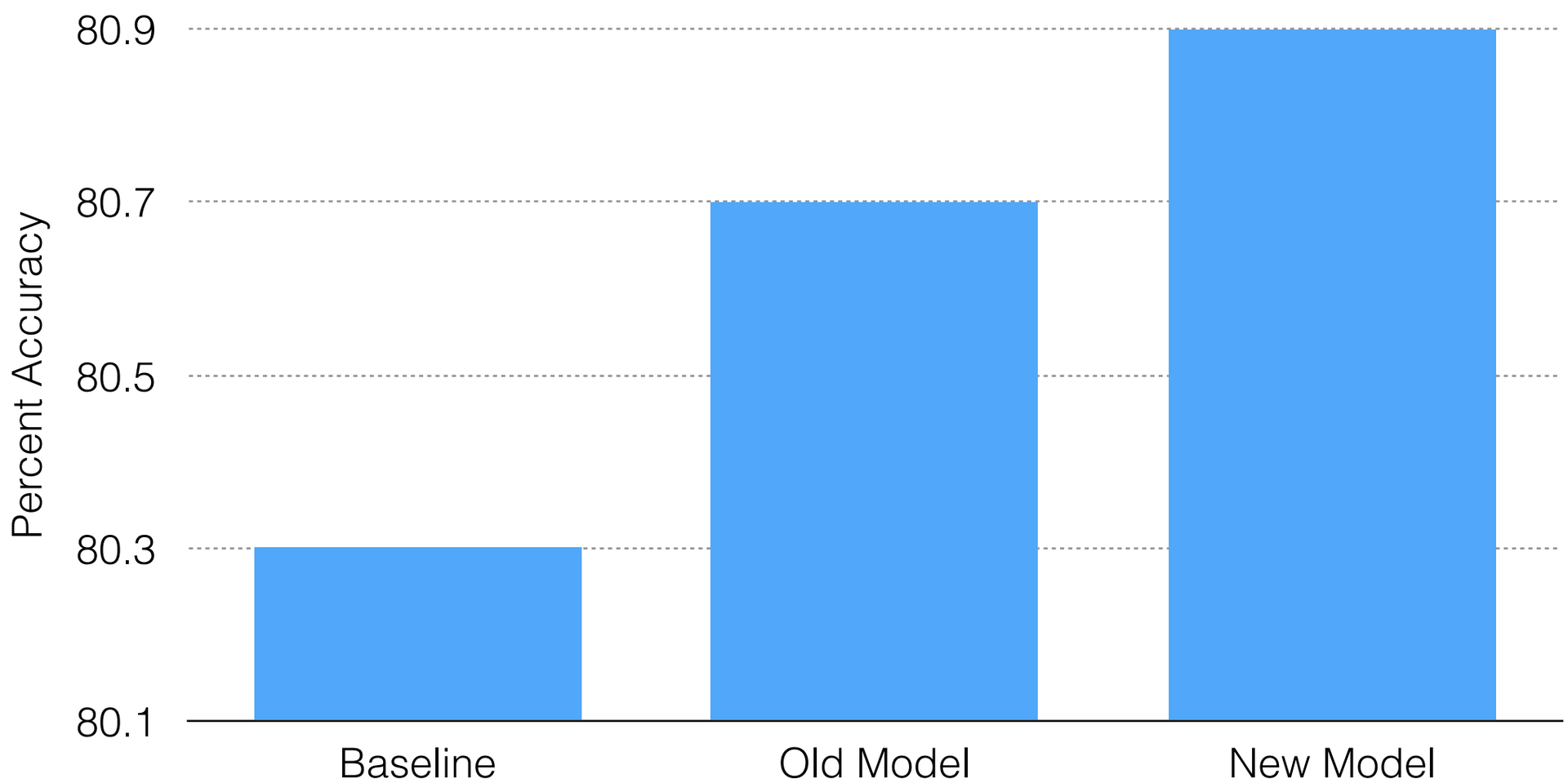
Be careful about showing smoothed/ estimated/aggregated trends only

Summary Stats Only

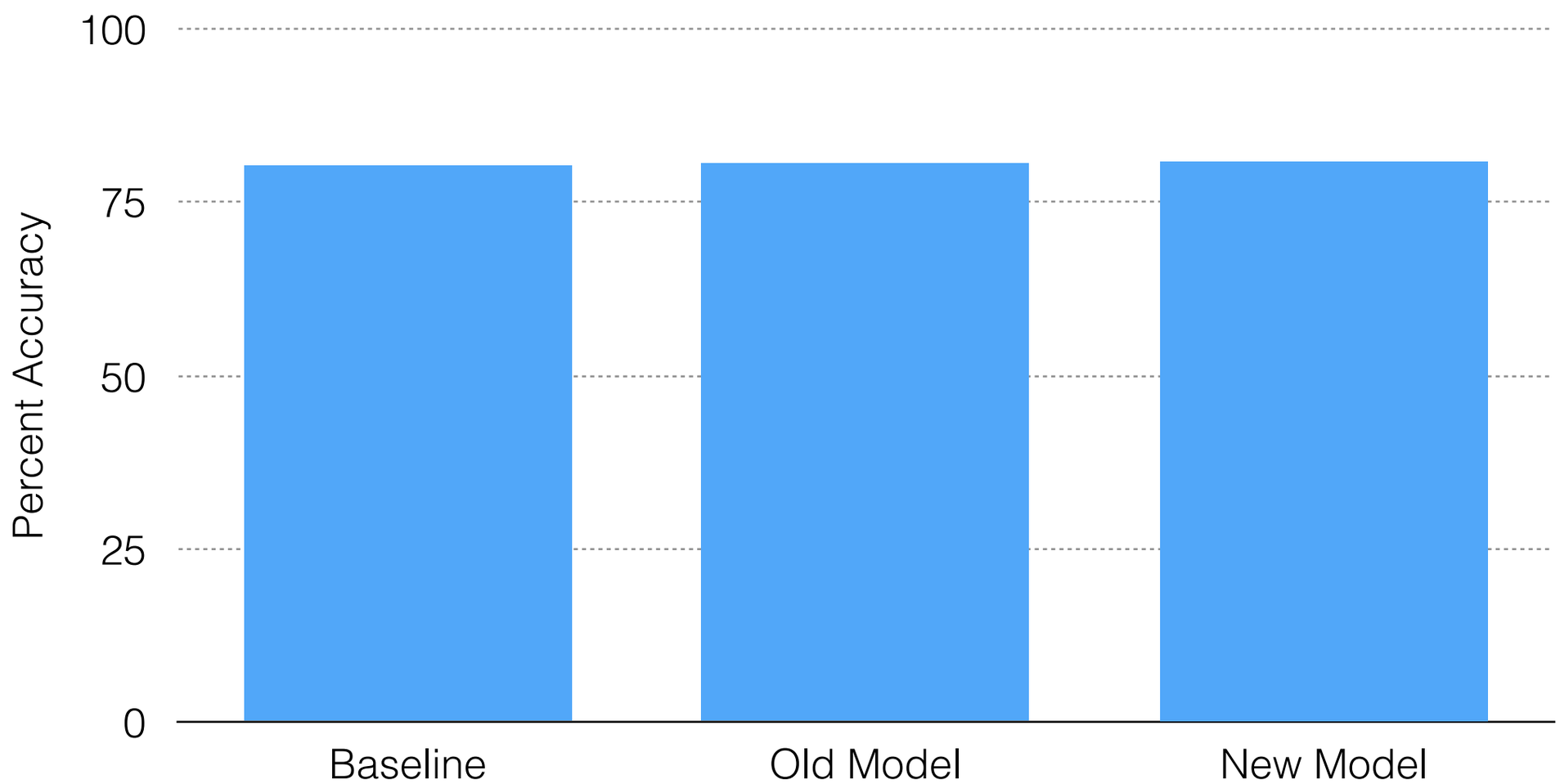


Whenever possible, show underlying data

Misleading/Badly Scaled Axes

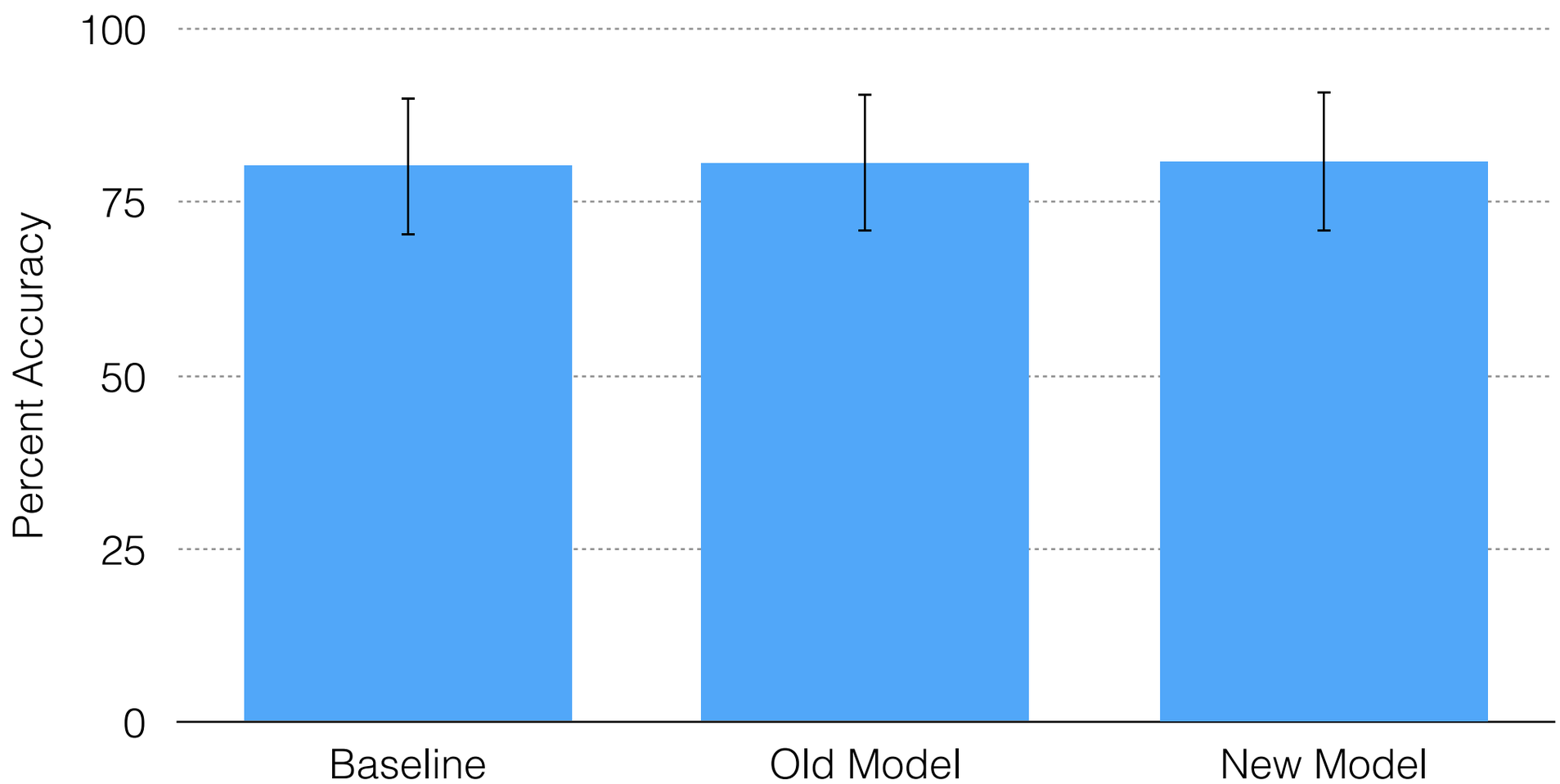


Misleading/Badly Scaled Axes



Rescale to a meaningful range (use full range of values that are interesting/expected)

Misleading/Badly Scaled Axes

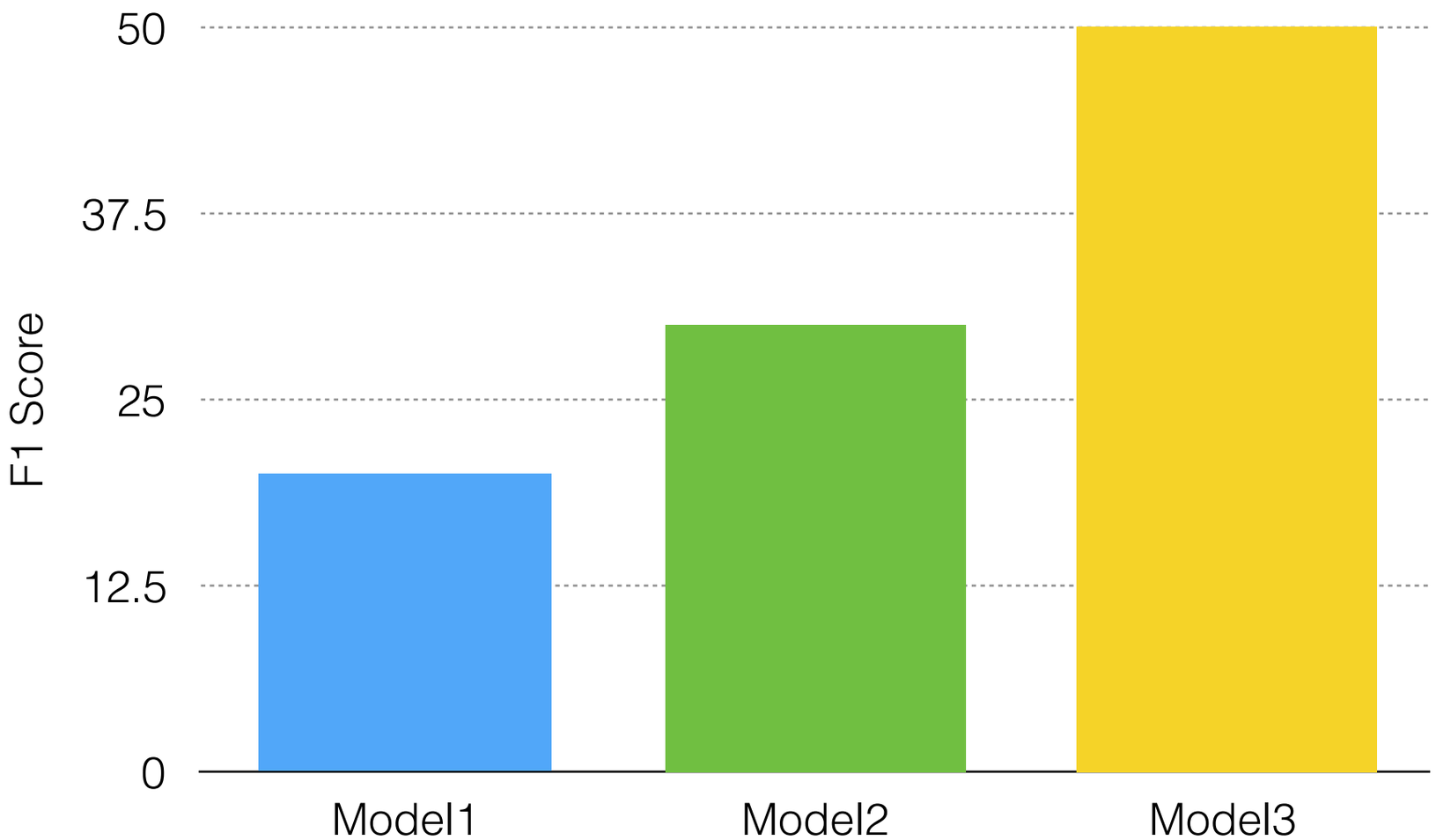


And/or include error bars

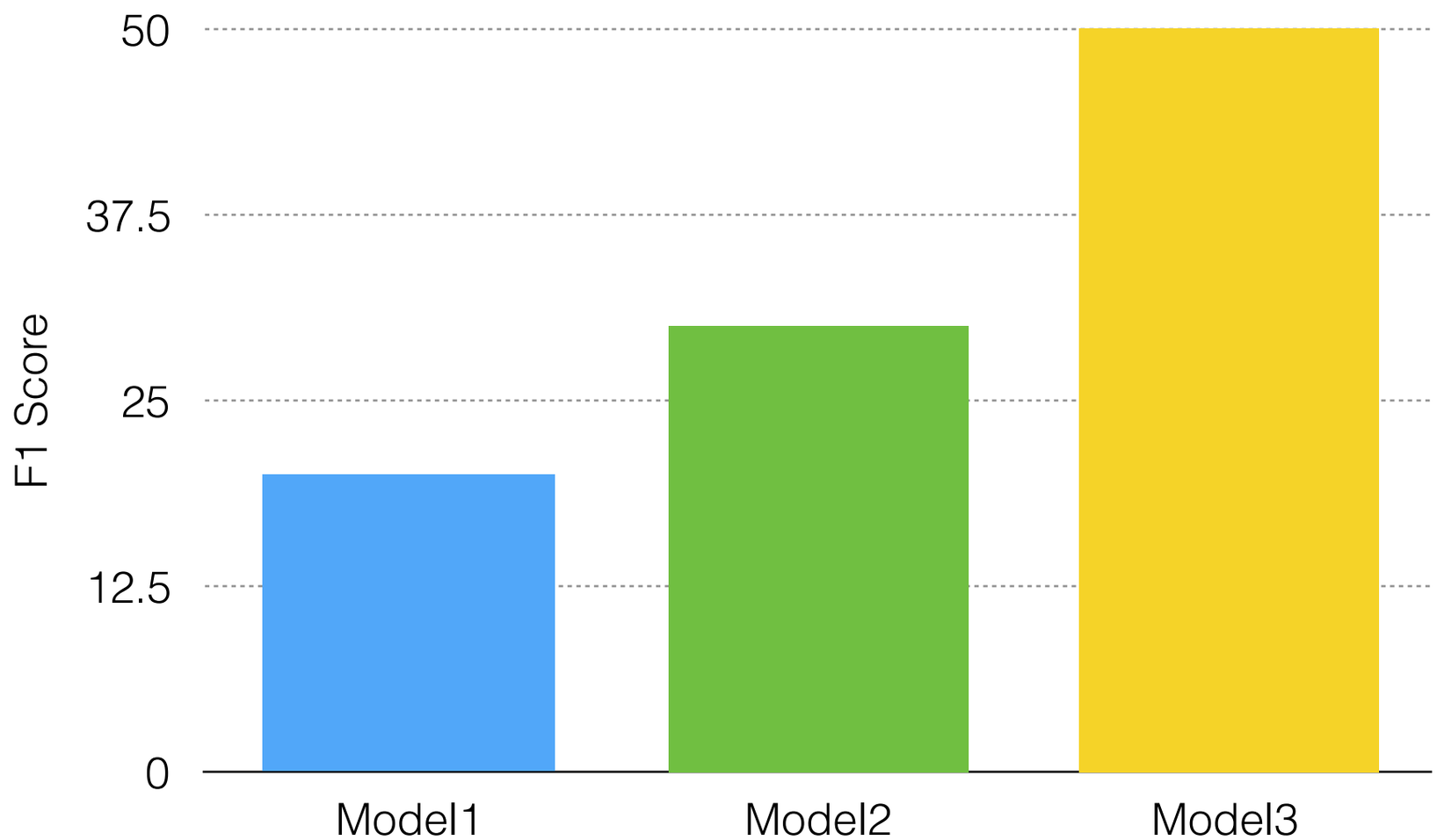
My “three pillars” of Data Viz

Minimalism — Substance over style. Make your point concisely, without redundant or distracting information or ornamentation.

Redundancy

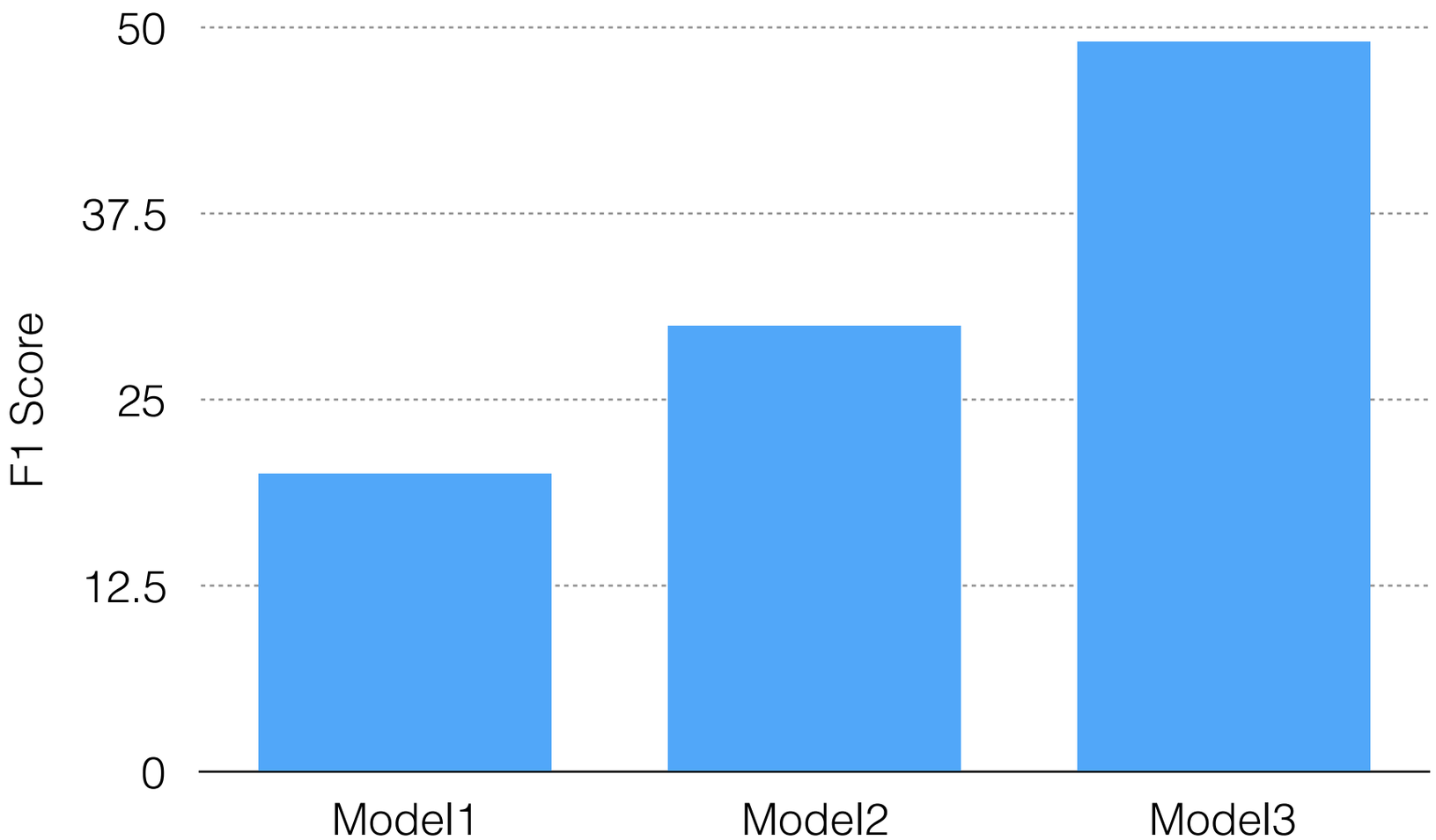


Redundancy



Don't use colors/decorations unless they add new information

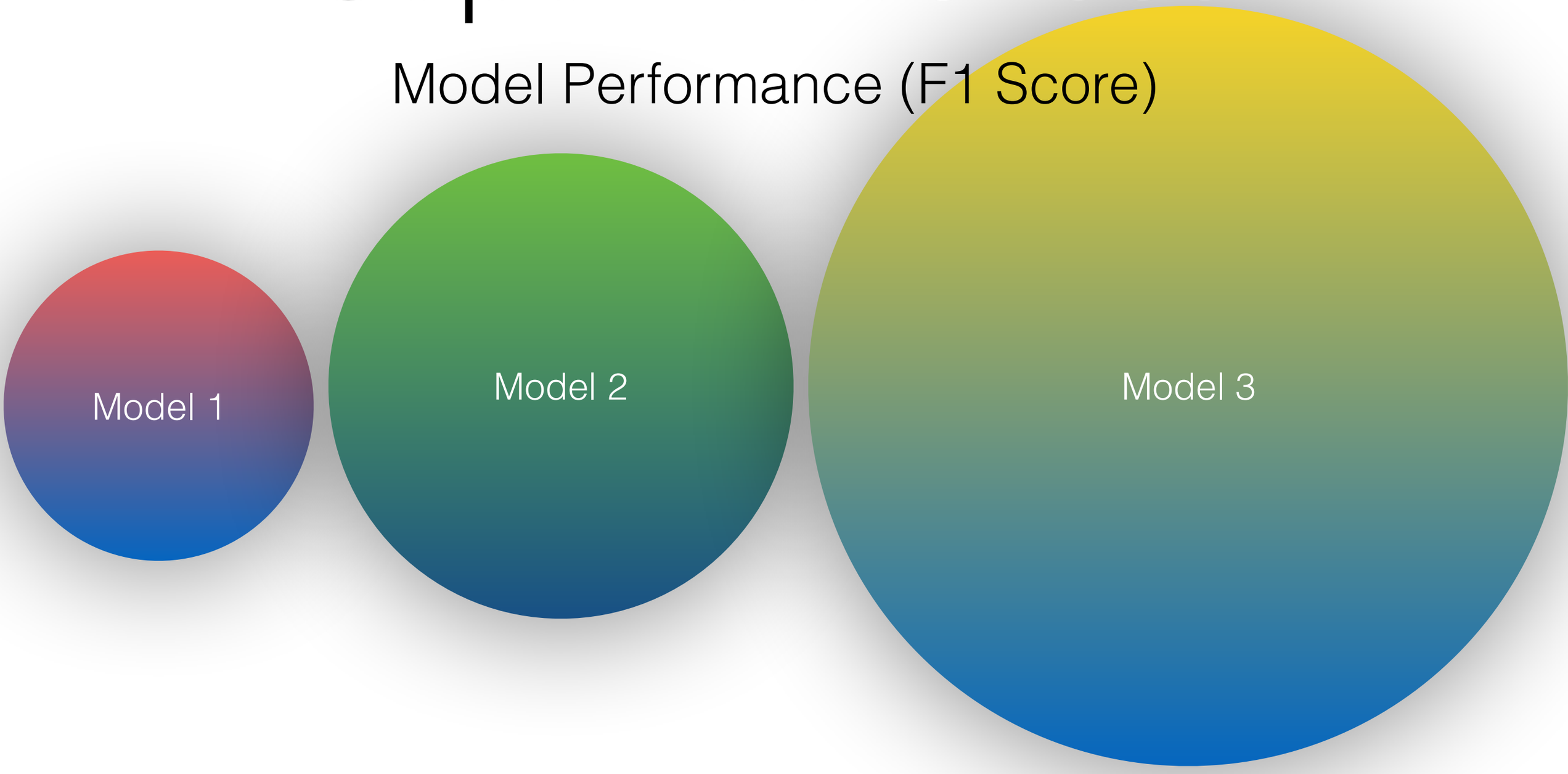
Redundancy



Just look how pretty that is

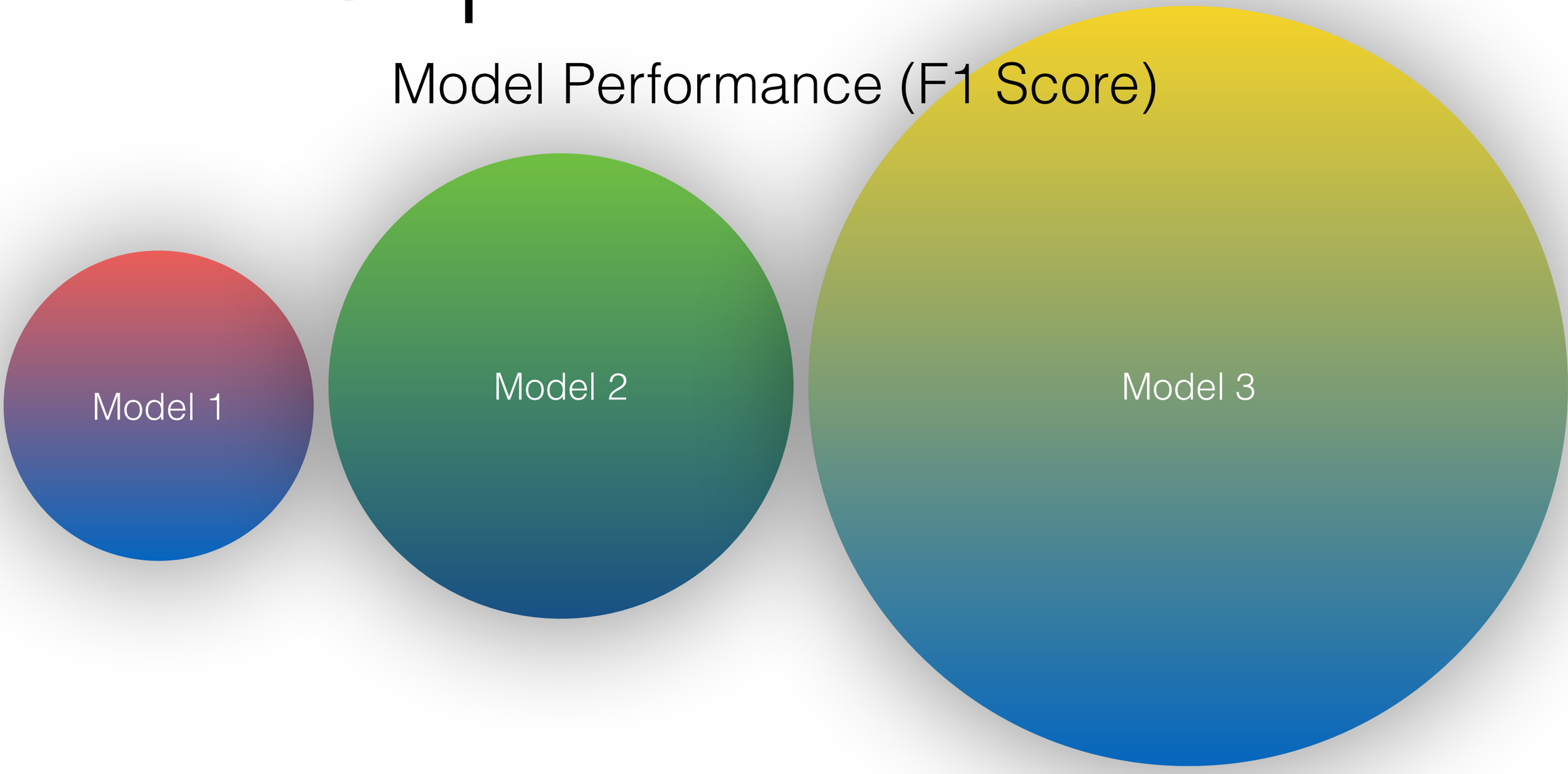
Superfluousness

Model Performance (F1 Score)



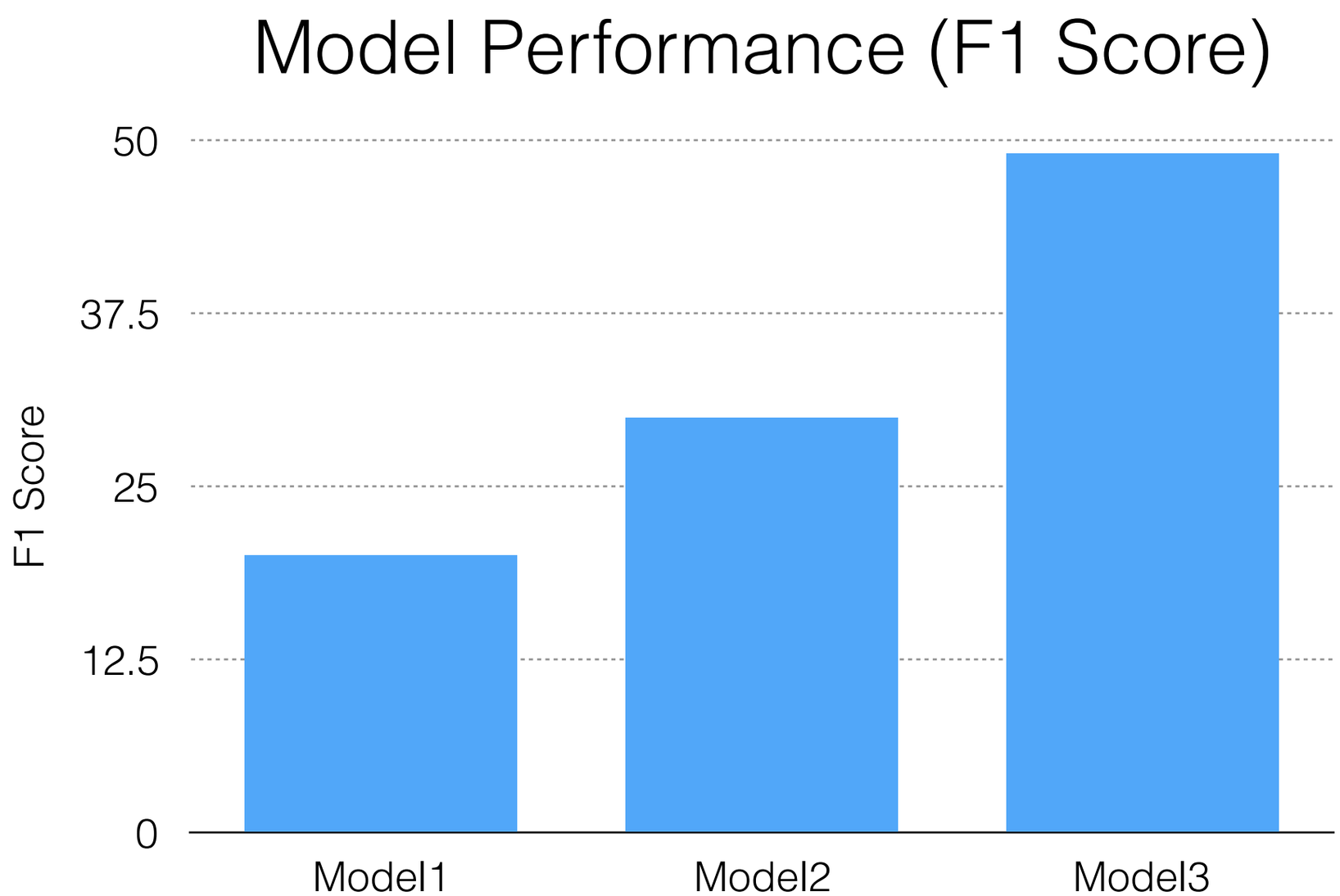
Superfluousness

Model Performance (F1 Score)



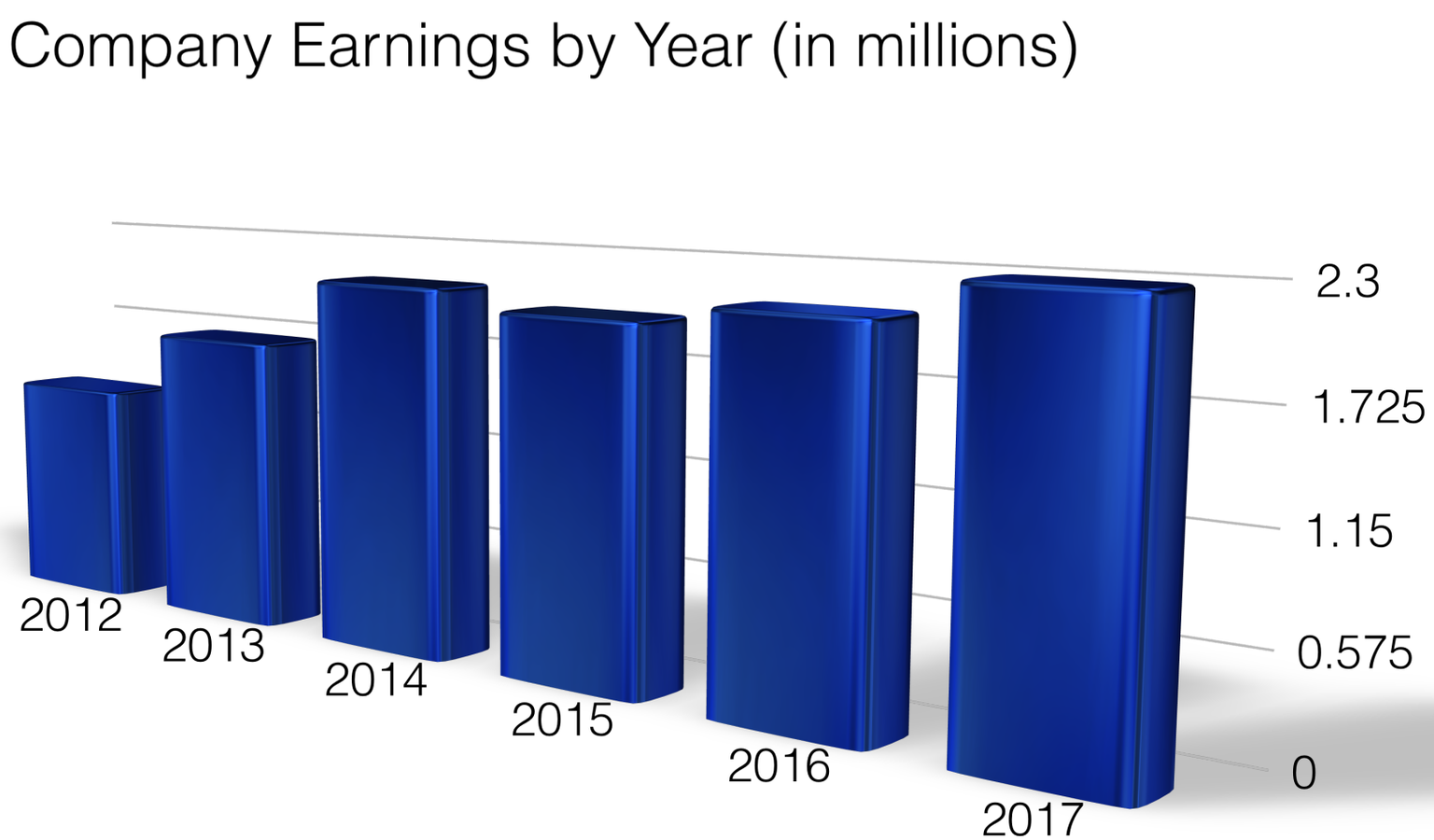
Don't use colors/decorations unless they add new information

Superfluousness

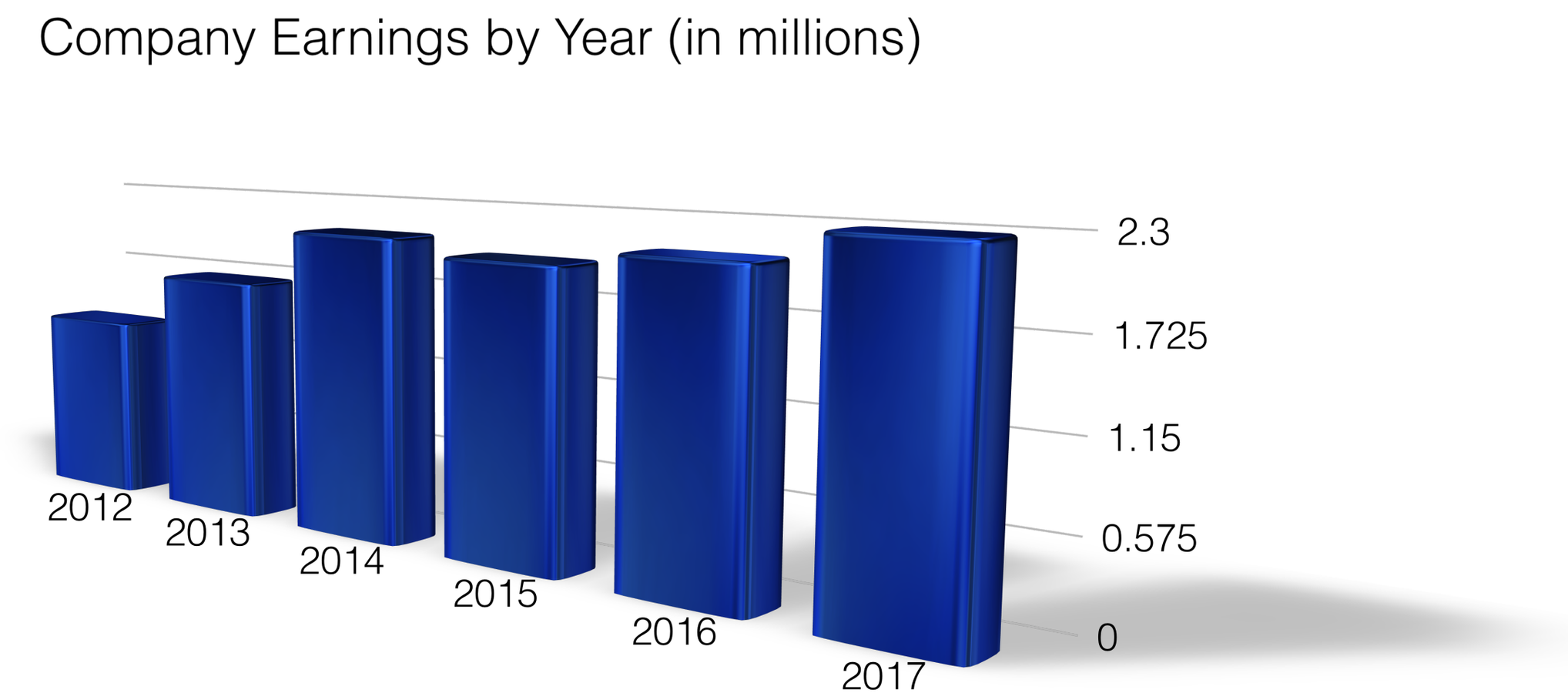


Just look how pretty that is

Why is that chart 3D?

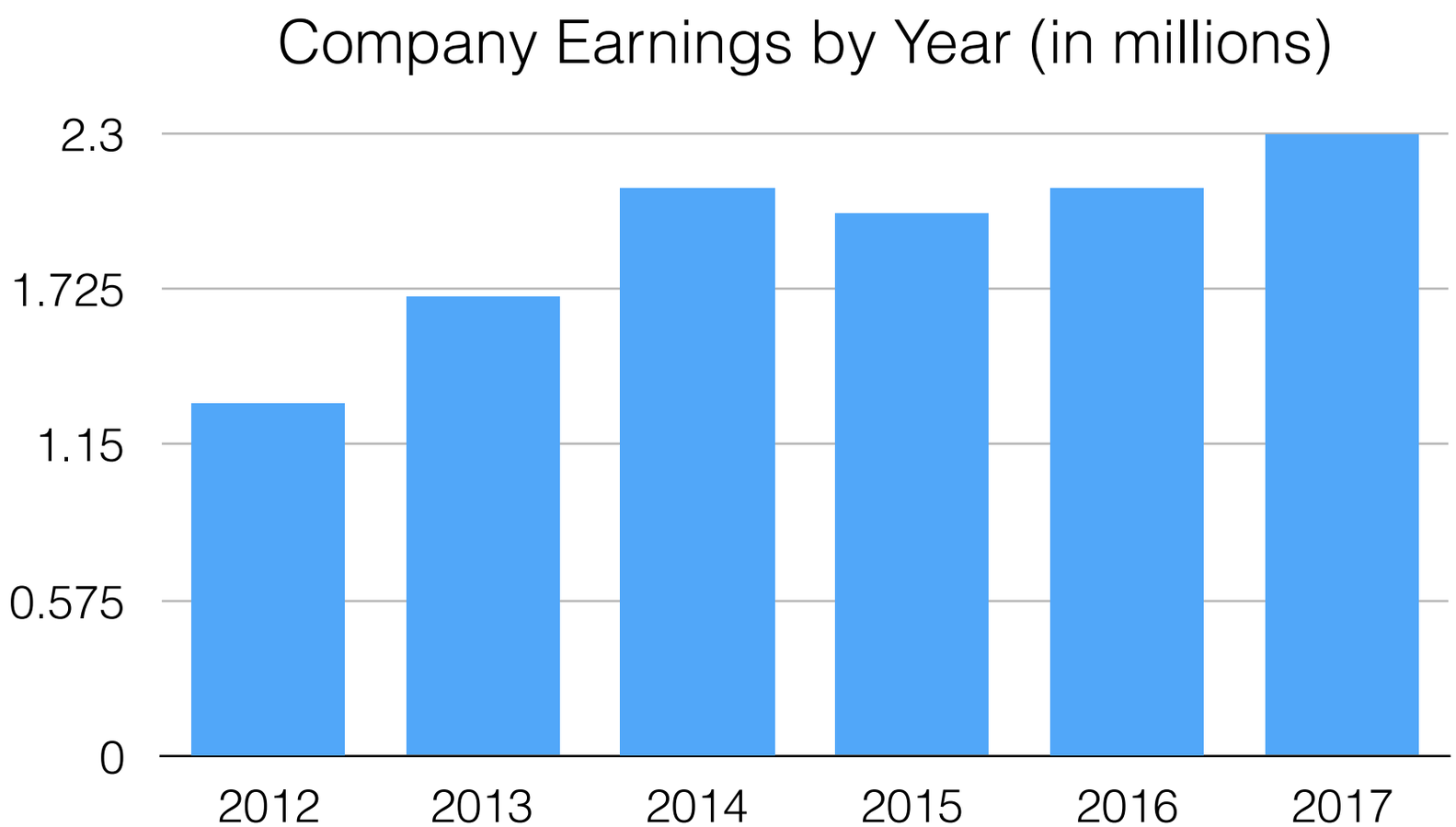


Why is that chart 3D?



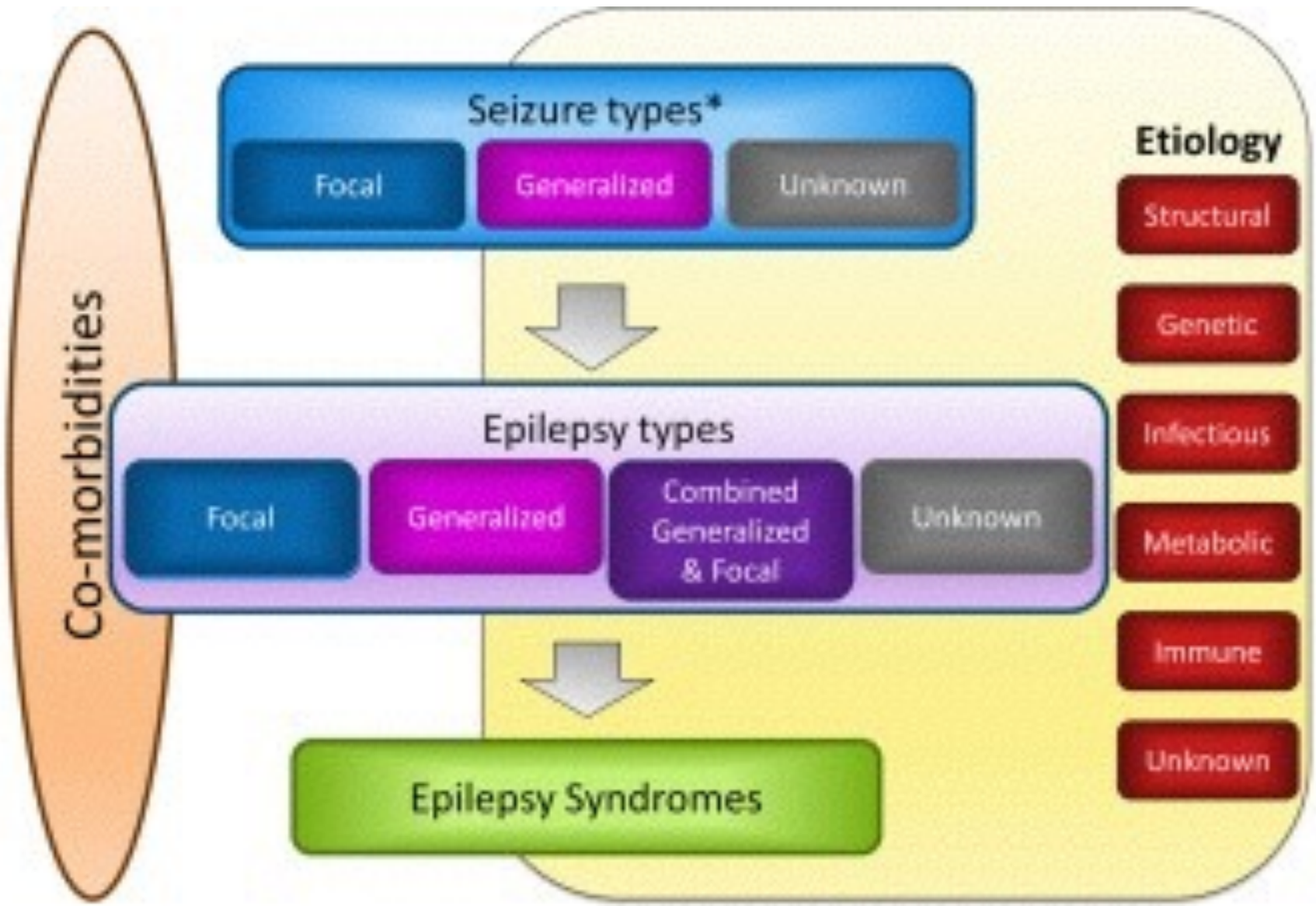
Just...don't

Why is that chart 3D?



Just look how pretty that is

. . . 😑


Seizure types*
Focal
Generalized
Unknown
Epilepsy types
Focal
Generalized
Combined Generalized & Focal
Unknown
Epilepsy Syndromes
Co-morbidities
Etiology
Structural
Genetic
Infectious
Metabolic
Immune
Unknown

Tips for quick-and-helpful data viz

Tips for quick-and-helpful data viz

Type of Hypothesis	First plot I'd make
Group A differs from Group B according to metric C	side-by-side histograms, with means and CIs
X effects Y	scatter plot, with correlation
Prediction Tasks, Recommendations	dim.reduction feature matrix to 2D, then scatter and color by label/group
Any of the above	correlation matrices between all features
Any of the above	counts of all features (broken down by groups/labels if relevant)

Tips for quick-and-helpful data viz

- Histograms:
  - Play with bin size/normalization until you can see clearly
  - Sometimes I use box plots if the variation is low (but always overlay the points themselves)
- Scatters:
  - Apply jitter or transparency to scatter plots so you can see overlapping points
  - Add labels (or use plotly for interaction) so you can see labels on points

Tips for quick-and-helpful data viz

- Matplotlib: https://matplotlib.org/ my <3, because I am old-school. Not super streamlined but does give you a lot of control

- Seaborn: https://seaborn.pydata.org/ plays well with numpy, streamlines process for making complex charts (e.g. large grids/side-by-sides) but harder to tweak little things

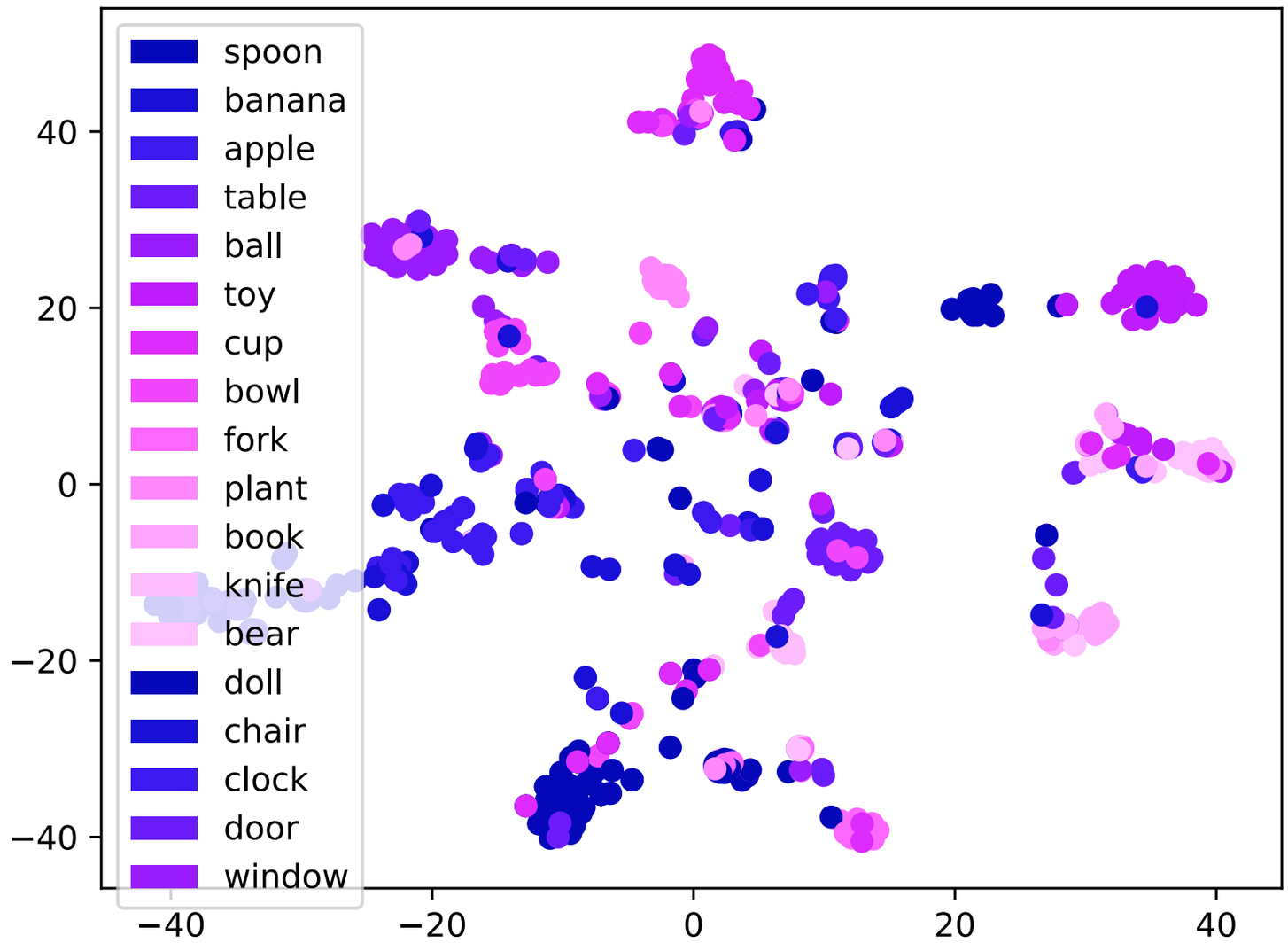
- Plotly: https://plotly.com/ good for quick interactive charts (I use this for messy scatter plots)

- D3: https://d3js.org/ good for making very flashy plots (and for doing your homeworks)

Clicker Question!

Look at this chart I made!!

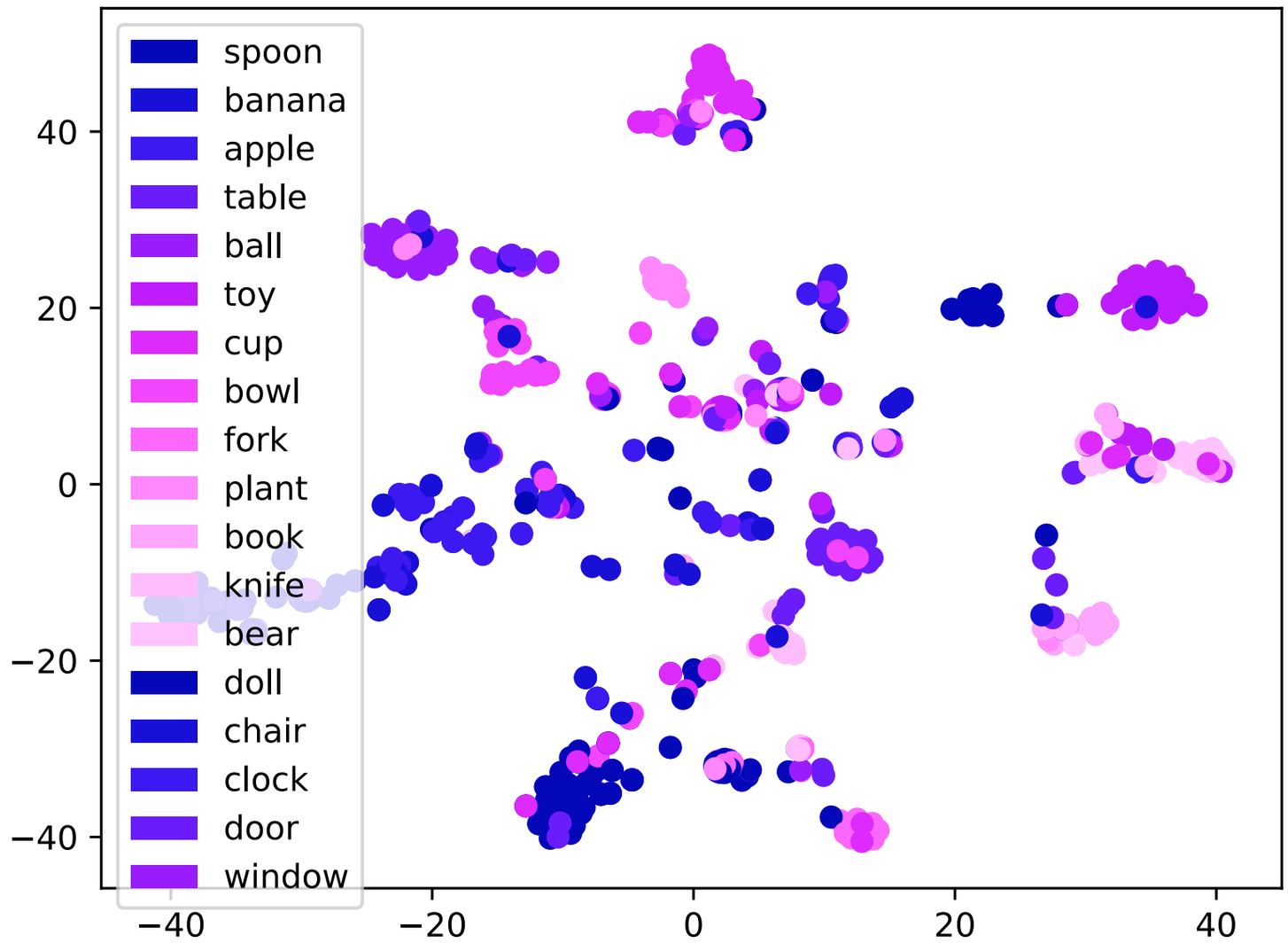
(a) A
(b) B
(c) C
(d) F



Clicker Question!

Look at this chart I made!!

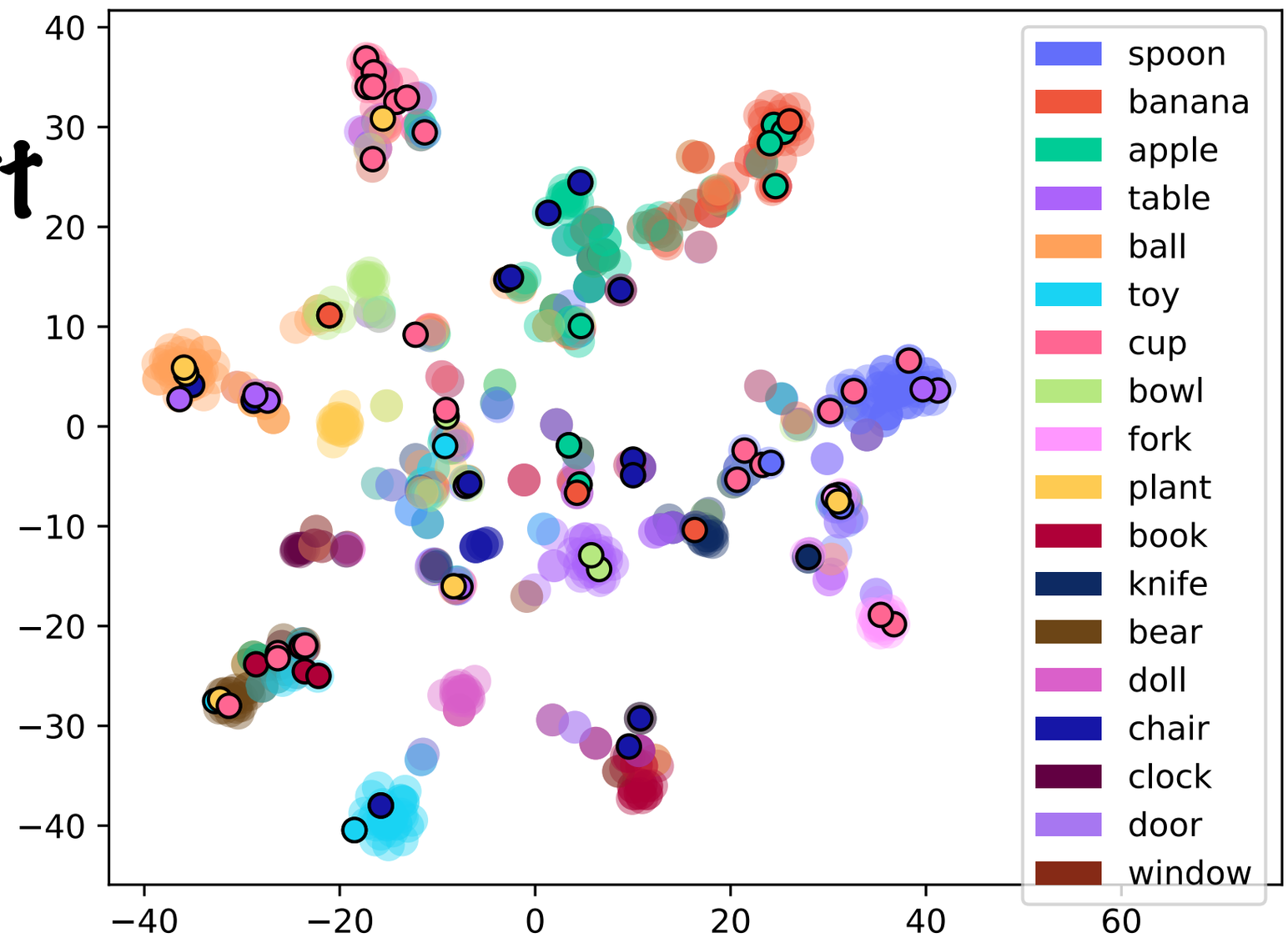
(a) A
(b) B
(c) C
(d) F



Clicker Question!

Look at this chart I made!!

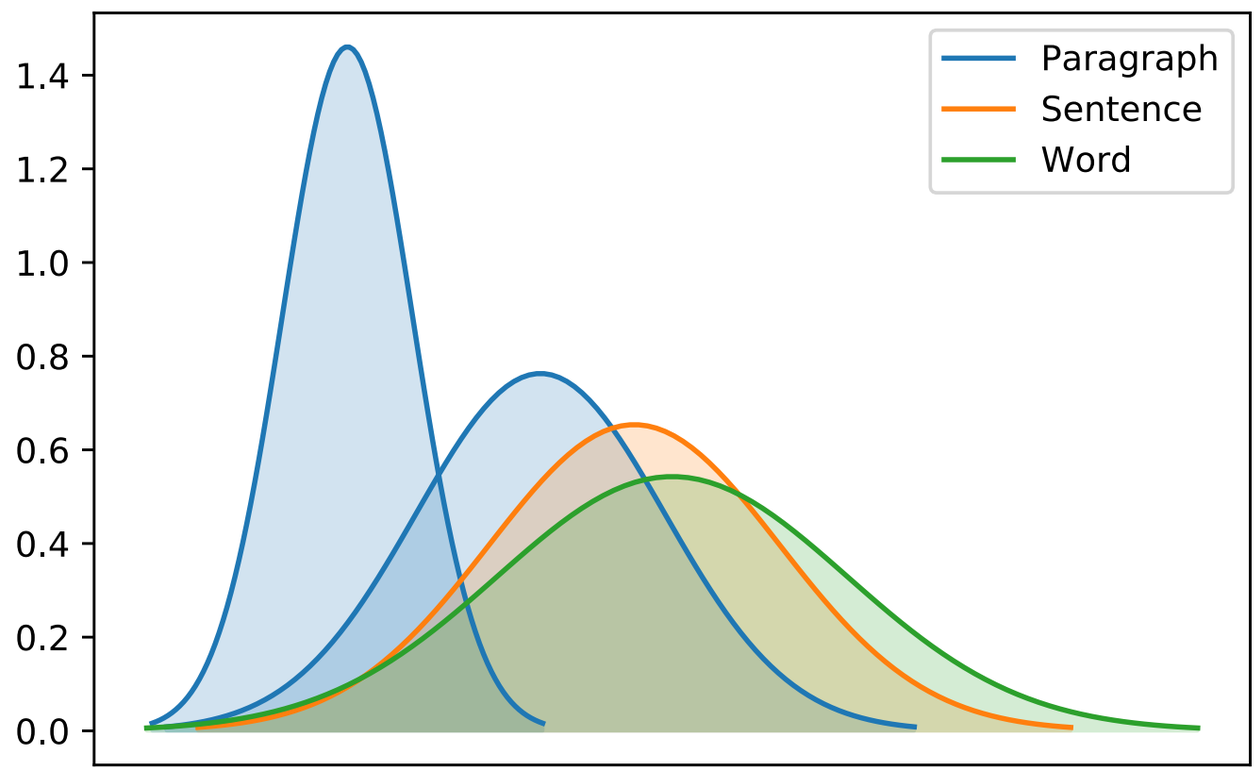
(a) A
(b) B
(c) C
(d) F



Clicker Question!

Look at this chart I made!!

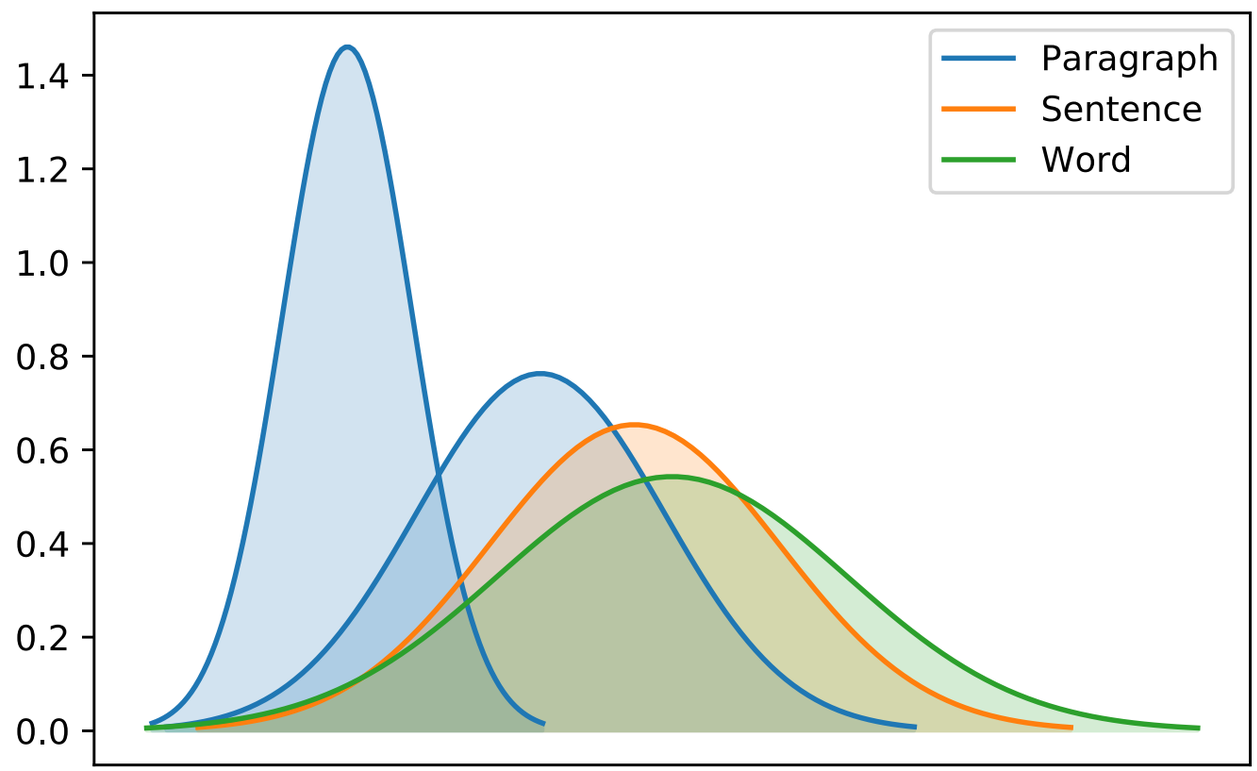
(a) A
(b) B
(c) C
(d) F



Clicker Question!

Look at this chart I made!!

(a) A
(b) B
(c) C
(d) F



Click

A watercolor painting celebrating that event hangs today in the Chenango Museum in Norwich. The canal itself was also utilized for recreation. In the summer months it supported swimming, **boating** and fishing. In the winter months, after the surface froze over, ice skating and even horse racing became favorite pastimes. Before the Chenango Canal was built, much of the Southern Tier and Central New York was still considered to be frontier.

Look at this chart I made!!

(a) A
(b) B
(c) C
(d) F

In the summer months it supported swimming , **picnicking** and fishing .