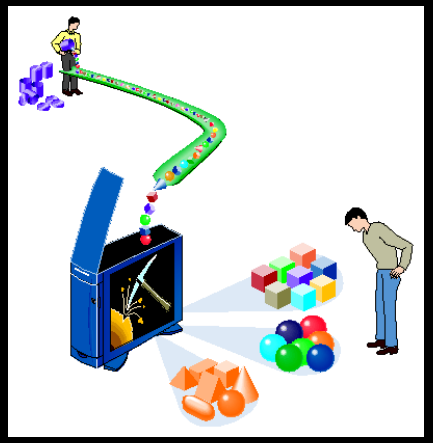
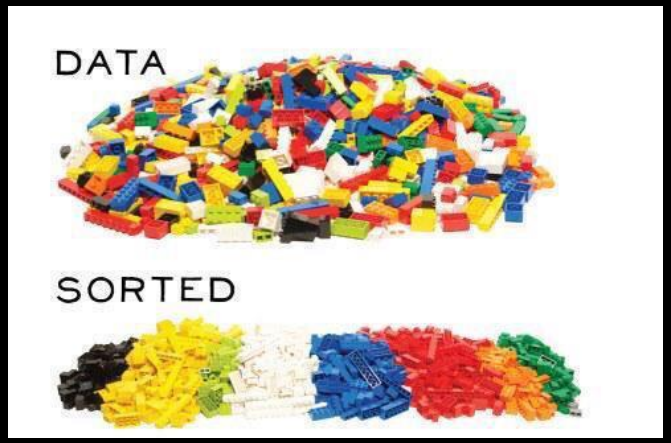
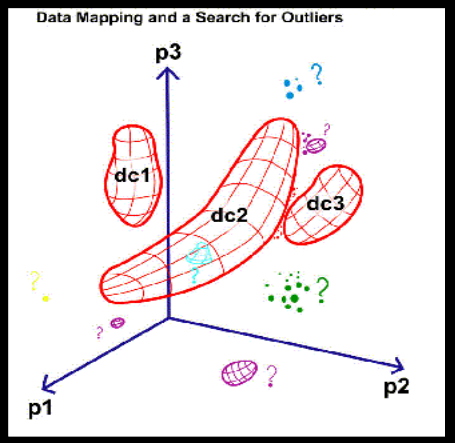
Booz Allen







Statistical and Data Literacy in the Era of Big Data

Kirk Borne

@KirkDBorne

Principal Data Scientist, Booz Allen Hamilton

http://www.boozallen.com/datascience

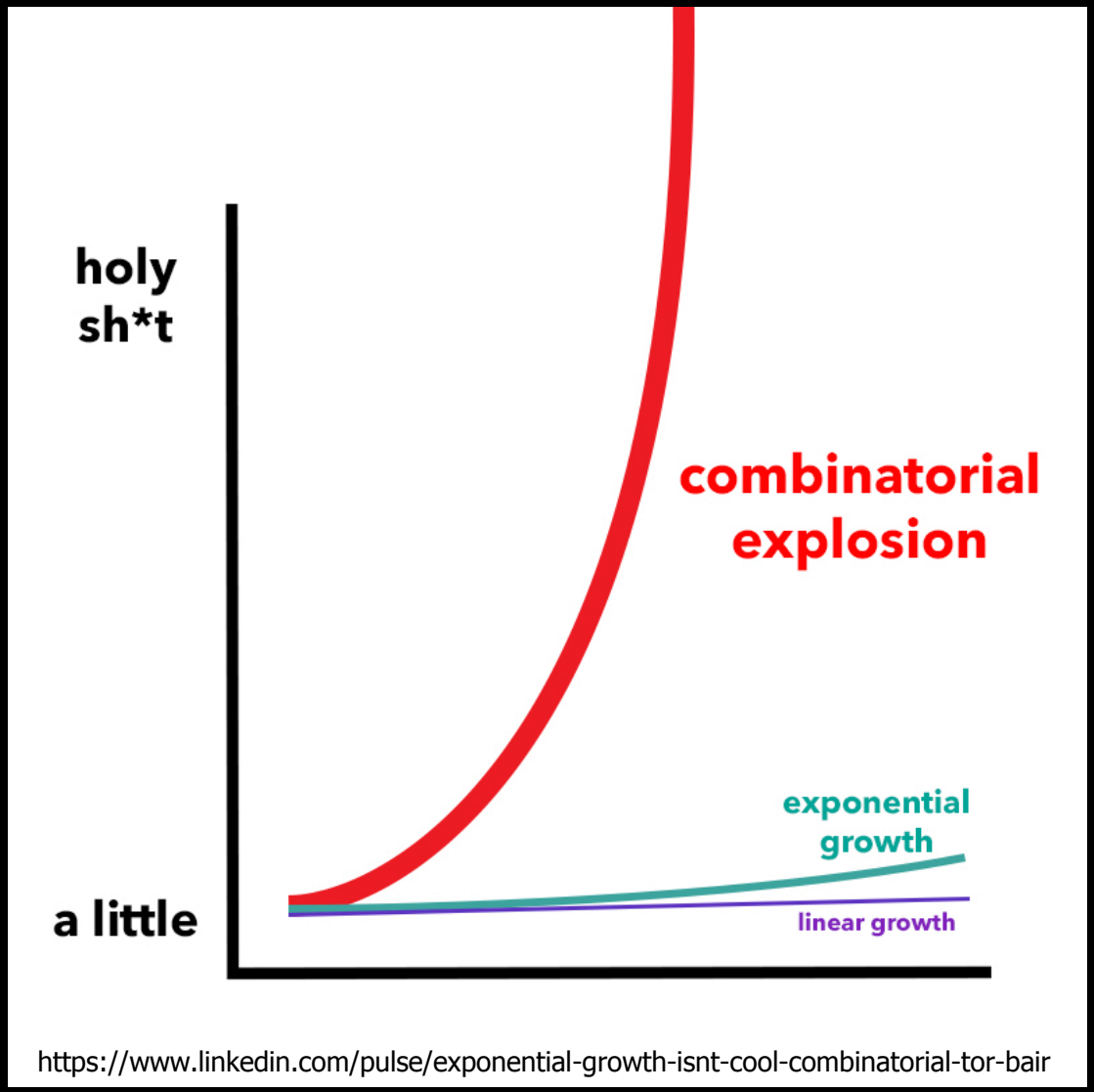
Ever since we first explored our world…

http://www.livescience.com/27663-seven-seas.html

…We have asked questions about everything around us.

https://jefflynchdev.wordpress.com/tag/adobe-photoshop-lightroom-3/page/5/

So, we have collected evidence (data) to answer our questions, which leads to more questions, which leads to more data collection, which leads to more questions, which leads to **BIG DATA!**



https://www.linkedin.com/pulse/exponential-growth-isnt-cool-combinatorial-tor-bair

$y \sim x! \approx x \wedge x$

$y \sim 2 \wedge x$

$y \sim 2 * x$

Huge quantities of data are now being **used** everywhere!

https://datafloq.com/read/dont-let-big-data-hr-program-backfire/1445

Huge quantities of data are now being **used** everywhere!

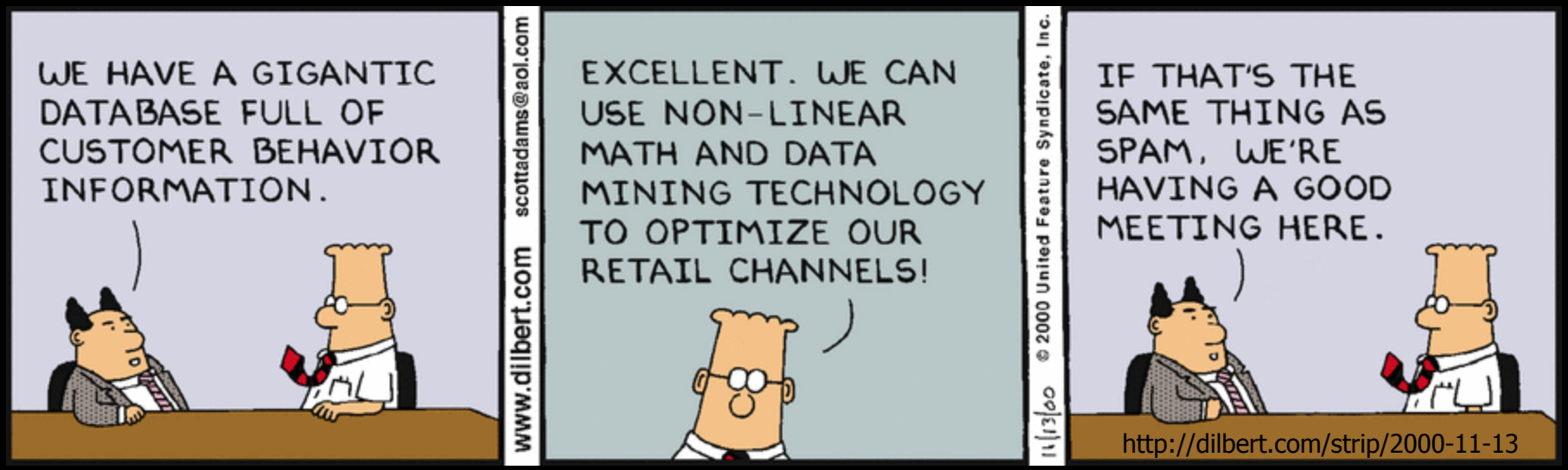
"With great power comes great responsibility."

– Spiderman's uncle (…or… Voltaire)

http://verix.com/6-new-articles-on-harnessing-the-power-of-big-data/

Data Ethics in Data Science in 2 parts: unbiased data & unbiased models

http://www.kirkborne.net/cds151/



The value of evaluation

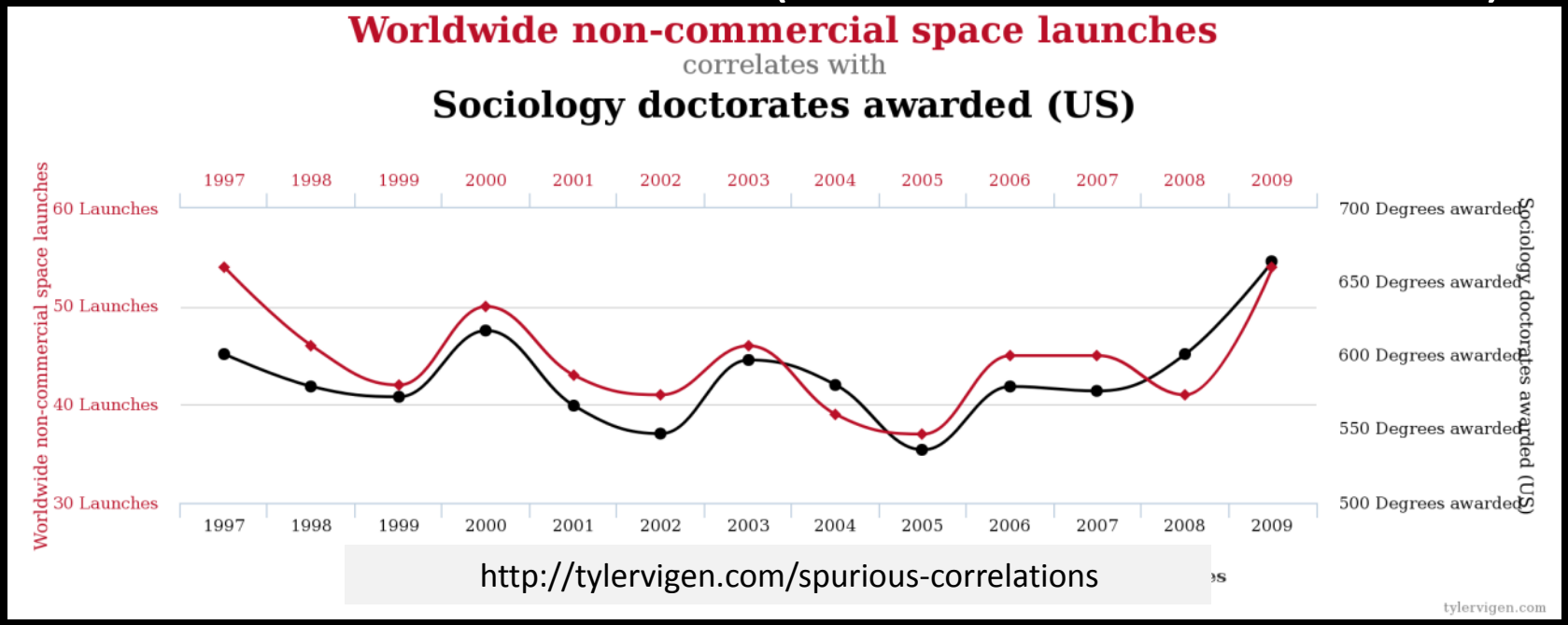
Data analysis can be fun and exploratory, BUT:

"If you torture the data long enough, it will confess to anything."

-Ronald Coase, economist

Statistical Fallacies can still appear in Data Science in the era of Big Data

- Correlation ≠ Causation (beware hidden variables)



- Biased sampling or biased models (underfitting)
- Ignoring natural variance in the data (overfitting)
- Absence of Evidence ≠ Evidence of Absence

Quote from H.G. Wells (1903; writer) …

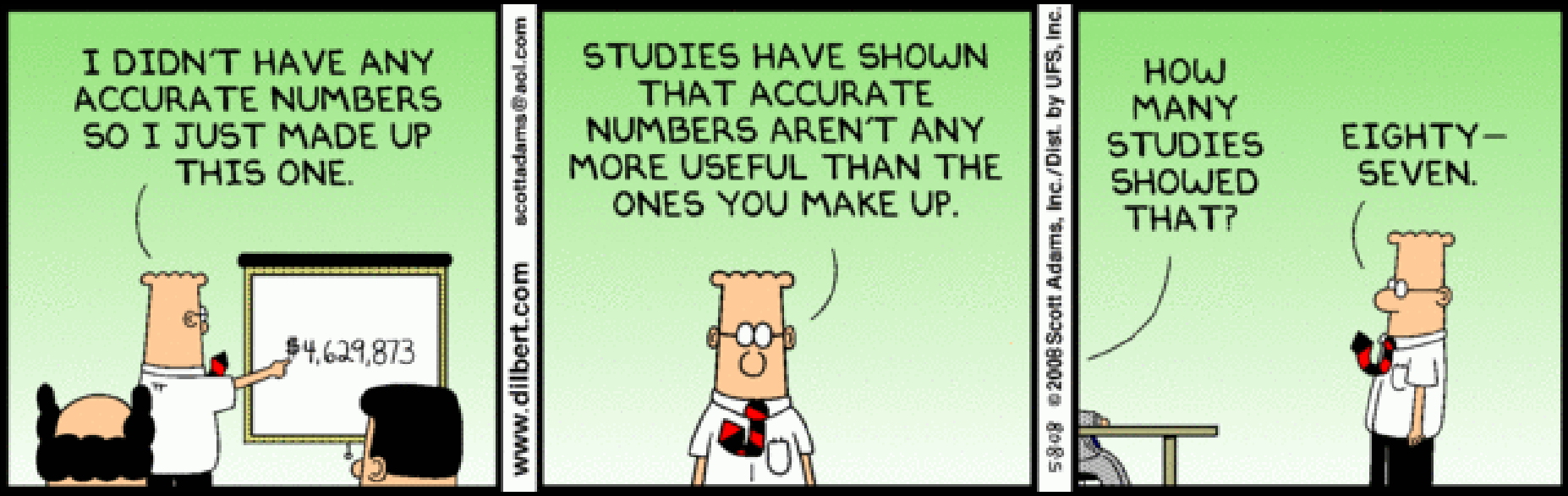
"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

Well, that day is here now!

Statistical & Data Literacy Matters!

Quote from Steven Wright (comedian) …

"42.7% of all statistics are made up on the spot."



http://dilbert.com/strip/2008-05-08

Quote from somebody (?) …

"It is now beyond any doubt that cigarettes are the biggest cause of statistics"

Let us start with a Statistics Quiz …

Which of these statements was made seriously and publicly:

a) "All models are wrong but some are useful"

b) "There are 3 types of lies – lies, damned lies, and statistics!"

c) "I am shocked that half the students in this country score below average on their standardized test scores"

d) "The best outcome for our education system is for more than half of our students to score below average on their standardized test scores!"

e) All of the above

Let us start with a Statistics Quiz …

Which of these statements was made seriously and publicly:

a) "All models are wrong but some are useful"

b) "There are 3 types of lies – lies, damned lies, and statistics!"

c) "I am shocked that half the students in this country score below average on their standardized test scores"

d) "The best outcome for our education system is for more than half of our students to score below average on their standardized test scores!"

✓ **e)All of the above**

Let us start with a Statistics Quiz …

Which of these statements was made seriously and publicly:

a) "All models are wrong b George Box, famous statistician

b) "There are 3 types of lies – lies, damned lies, and statistics!" Benjamin Disraeli, British Prime Minister

c) "I am shocked that half the students in this country score below average on their standardized test scores" (famous American politician from 1990's)

d) "The best outcome for our education system is for more than half of our students to score below average on their standardized test scores!" … me!

✓ **e)All of the above**

Let us start with a Statistics Quiz …

Which of these statements was made seriously and publicly:

Let us examine these last two statements…

c) "I am shocked that half the students in this country score below average on their standardized test scores" (famous American politician from 1990's)

d) "The best outcome for our education system is for more than half of our students to score below average on their standardized test scores!" … me!

✓ **e)All of the above**

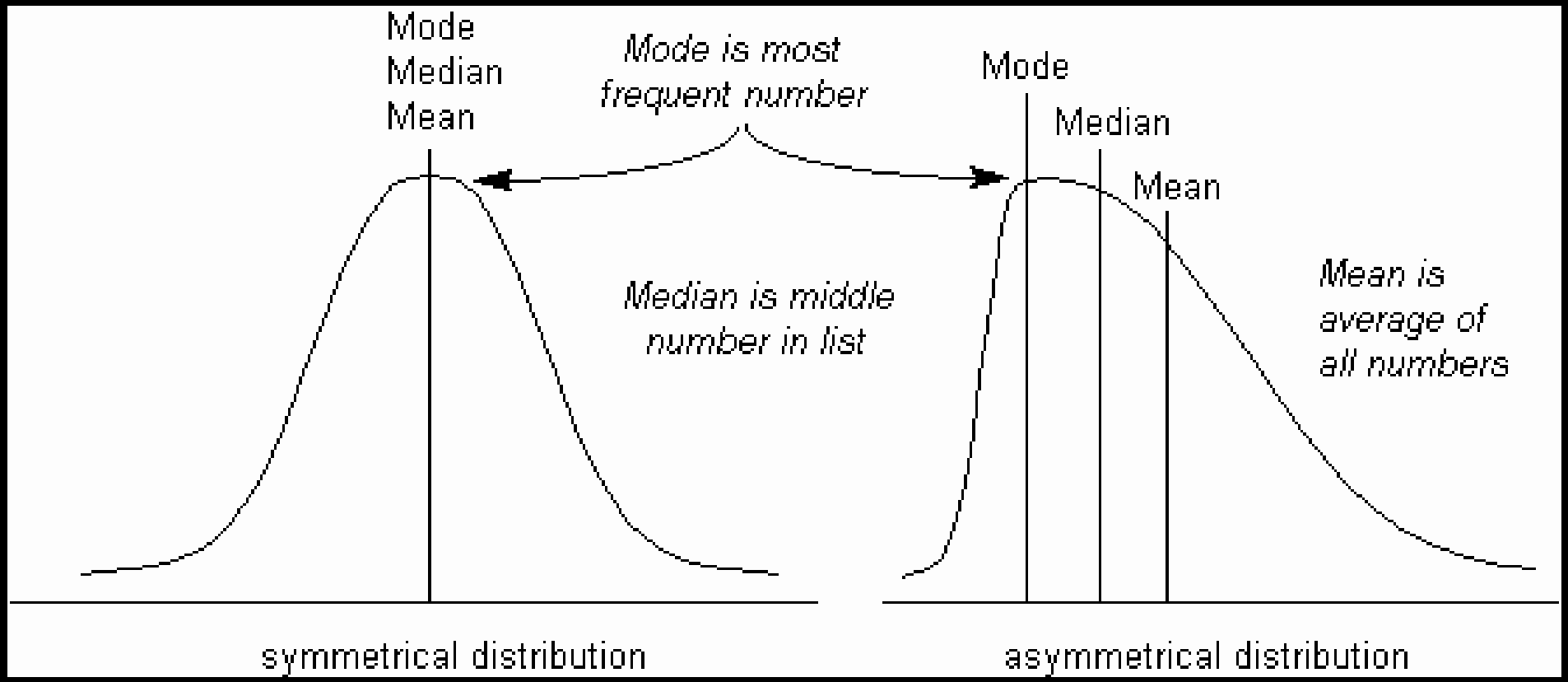
First question to ask ourselves:

What do we mean by *average?*

…

Is it the Mean or Median or Mode?

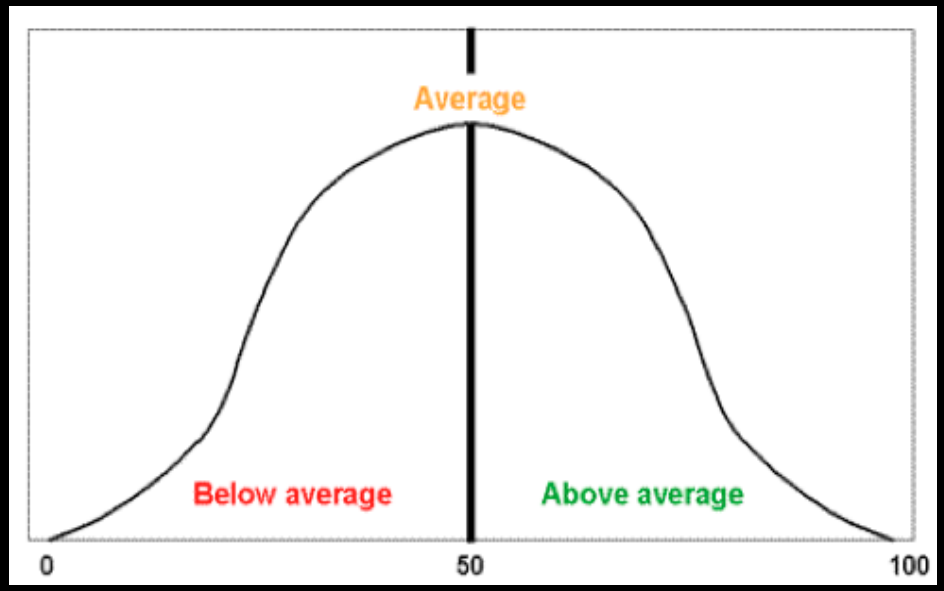
Means, Medians, and Modes, oh my



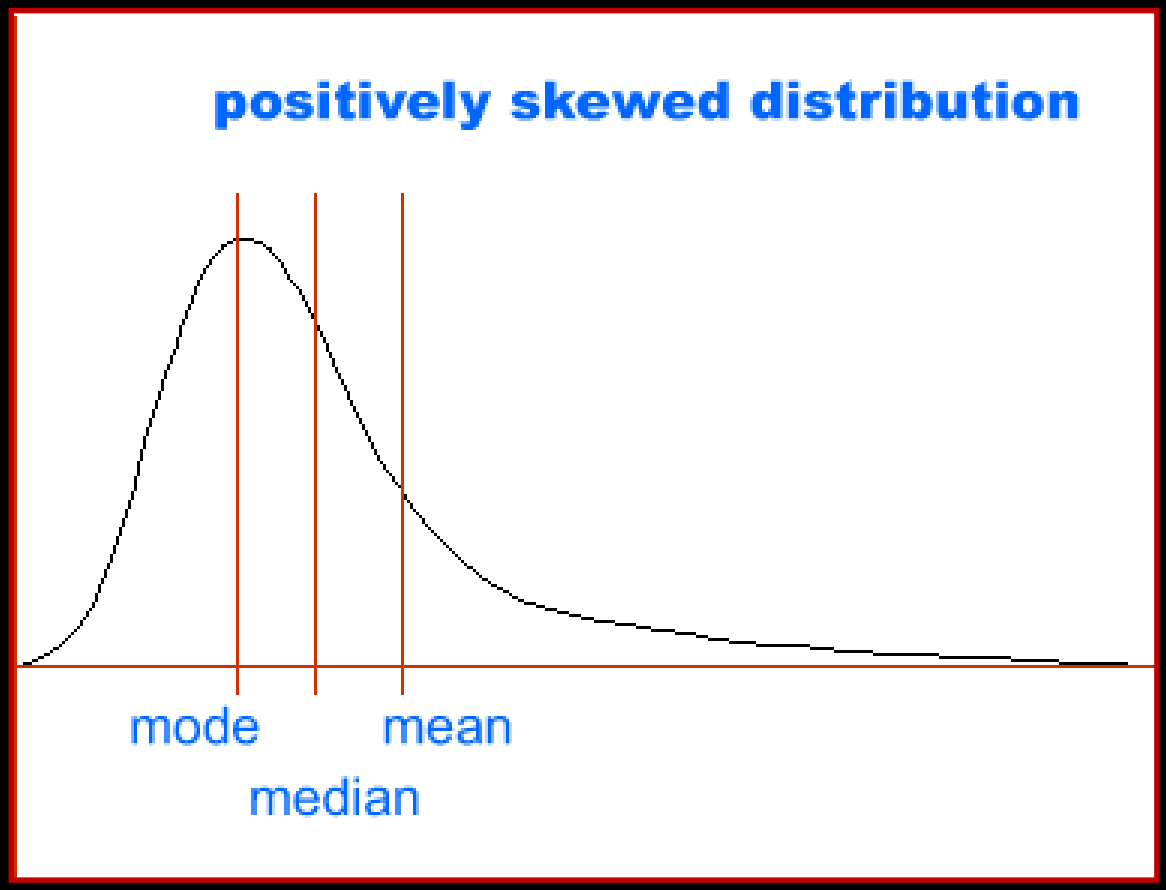
Here is the analysis …

"…half of the students in this country score below average on their standardized tests."

For a bell curve (normal) distribution of test scores, this statement is **neither** a good thing nor a bad thing, since half of the students will always score below average (by any definition of the word "average").

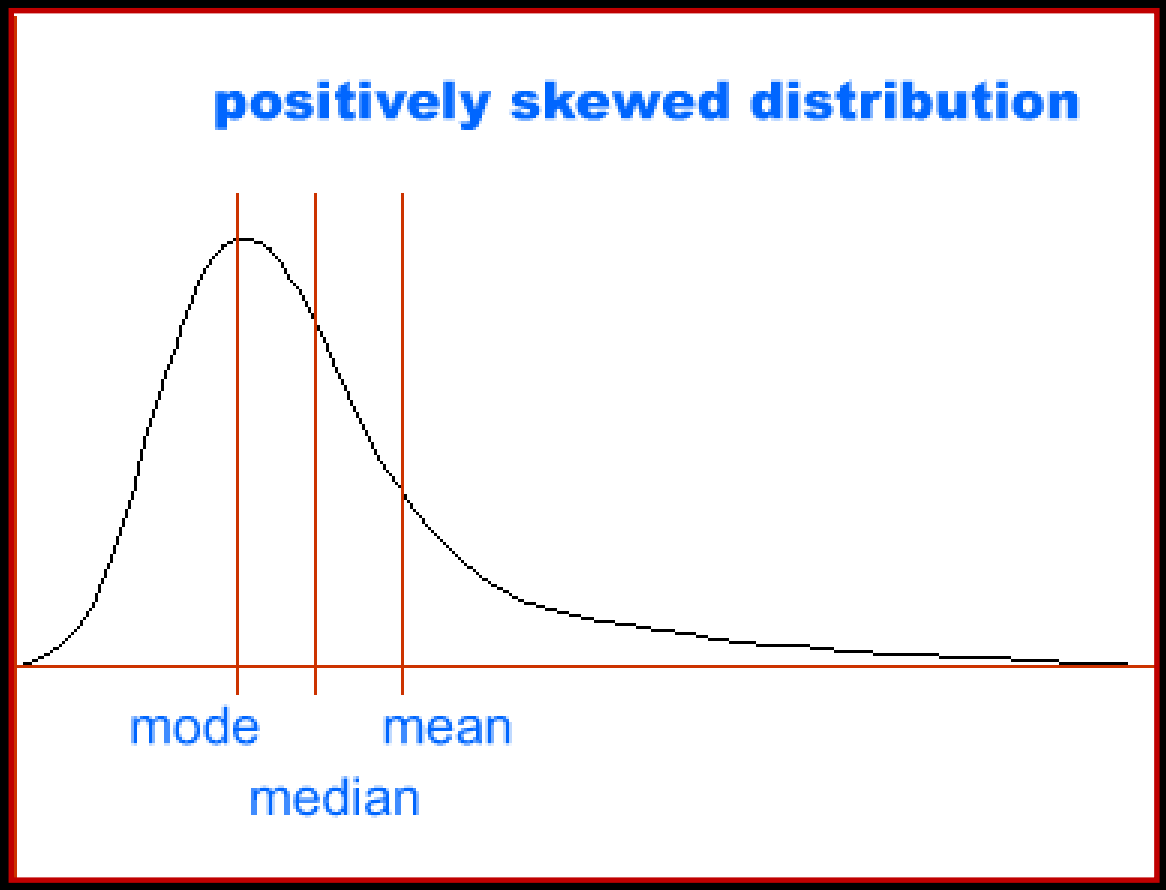


The Positively Skewed Distribution



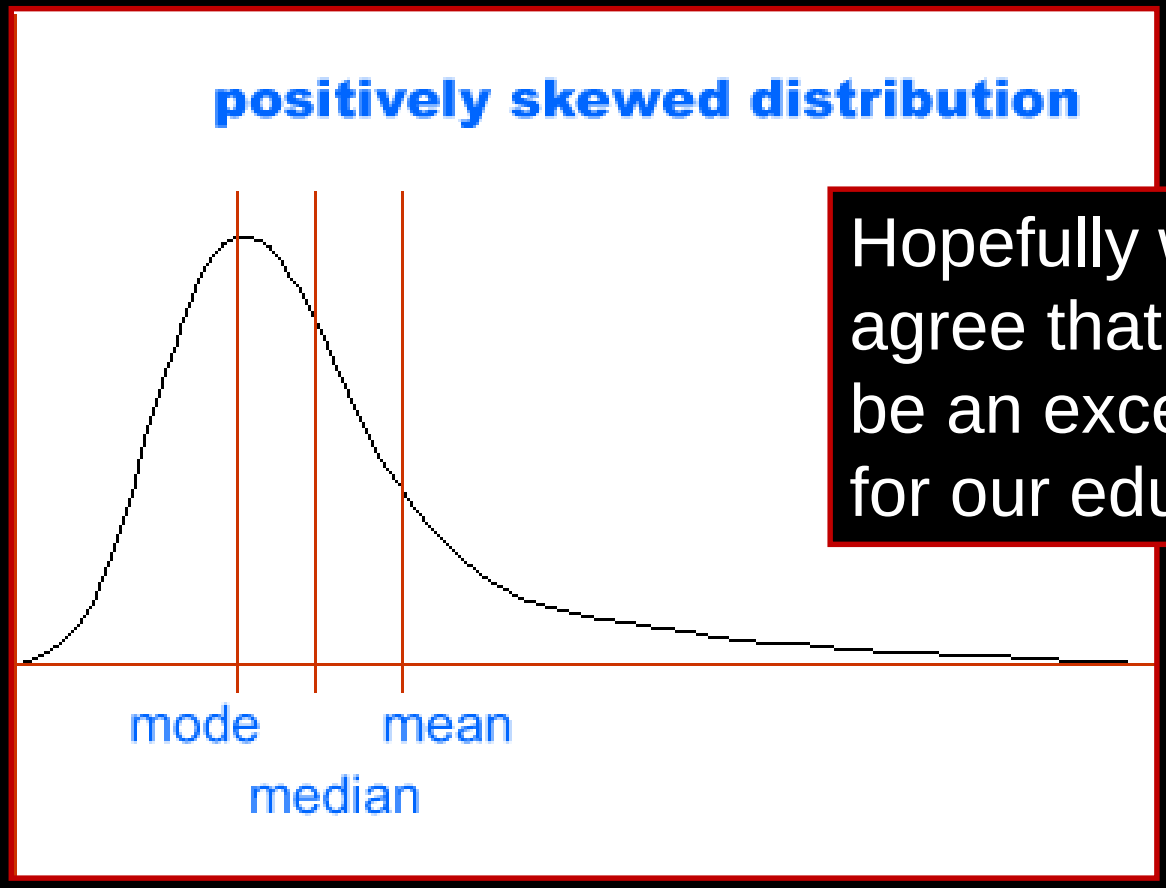
The Positively Skewed Distribution

- **This case corresponds to the statement:**
  "More than half of the values are below average" (i.e., below the mean)



The Positively Skewed Distribution

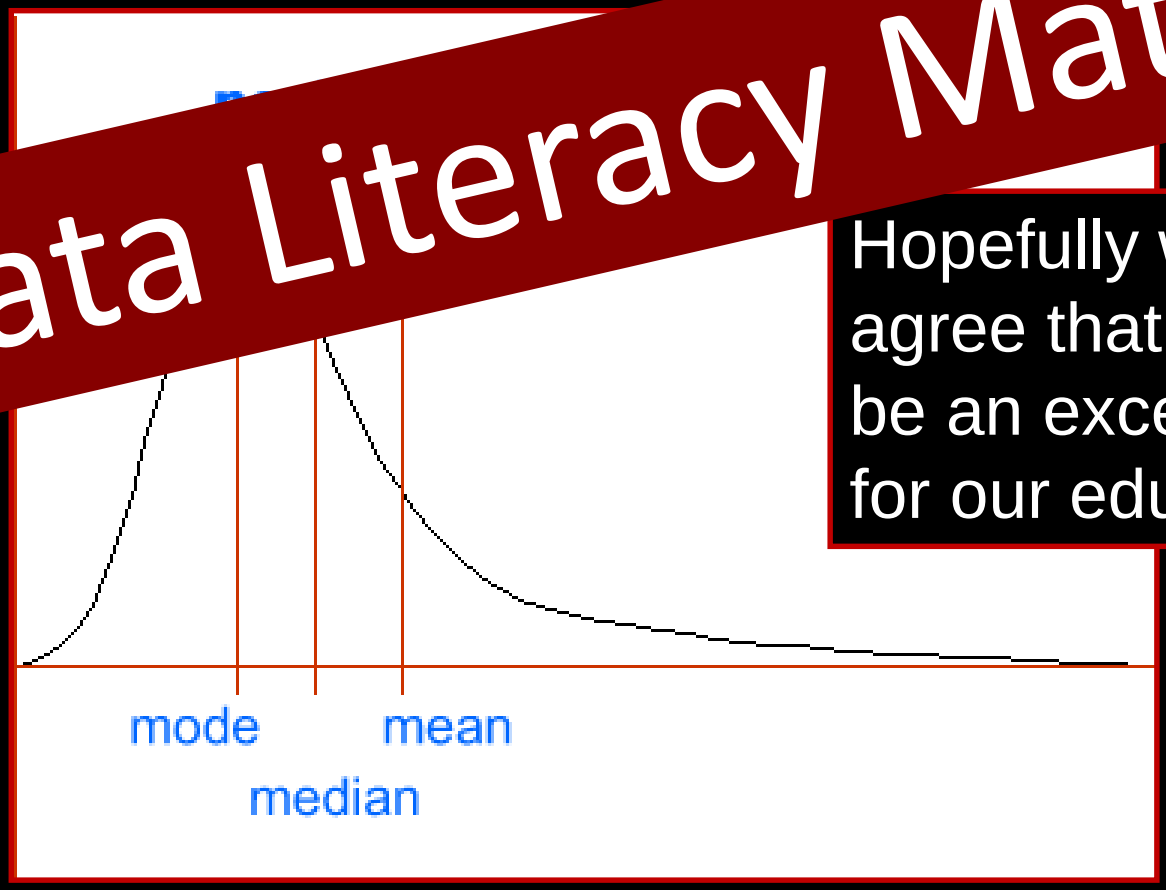
- **This case corresponds to the statement:**
  "More than half of the values are below average" (i.e., below the mean)



Hopefully we can all agree that this would be an excellent outcome for our education system!

The Positively Skewed Distribution

- **This case corresponds to the statement:**
  "More than half of the values are belo
  average" (i.e., below the

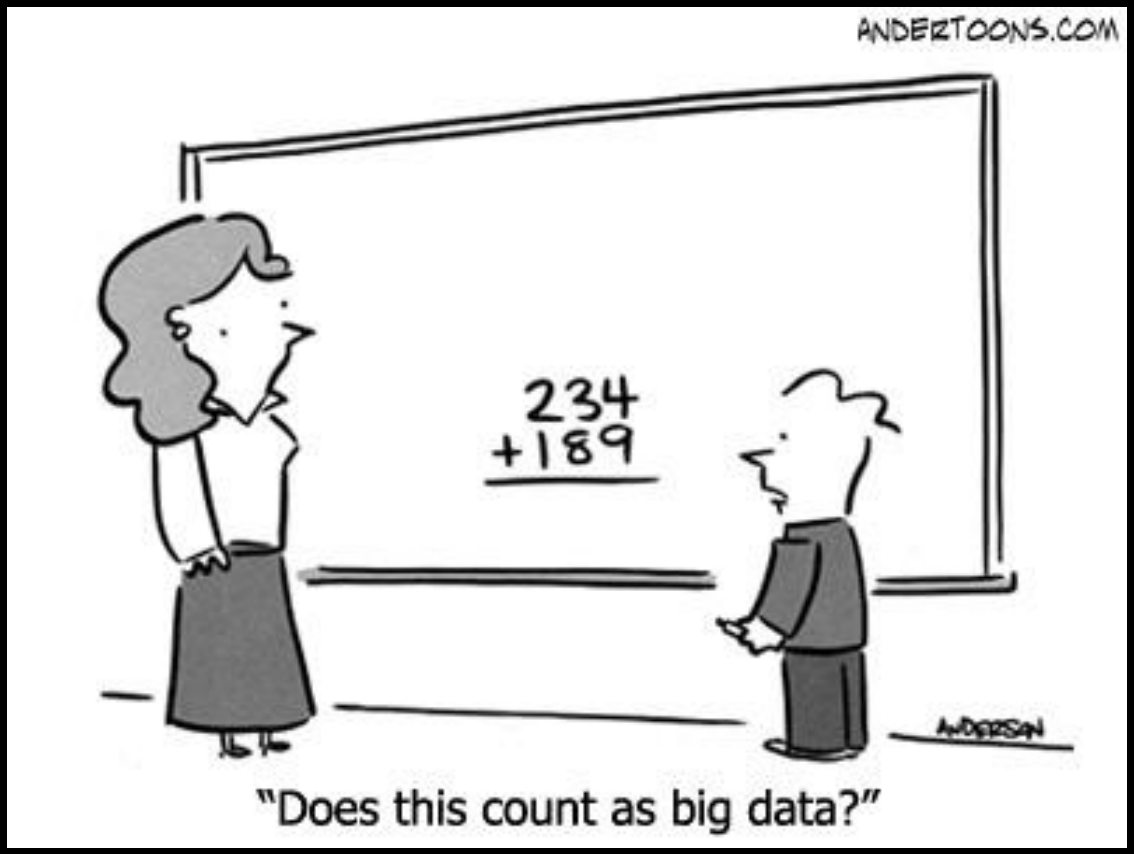


Data Literacy Matters!

Hopefully we can all agree that this would be an excellent outcome for our education system!

Enough with the quiz!

- These were examples of statistical literacy
- **Data Literacy** includes statistical literacy … and more…



The ~~5~~ 6 Commandments of Data Science

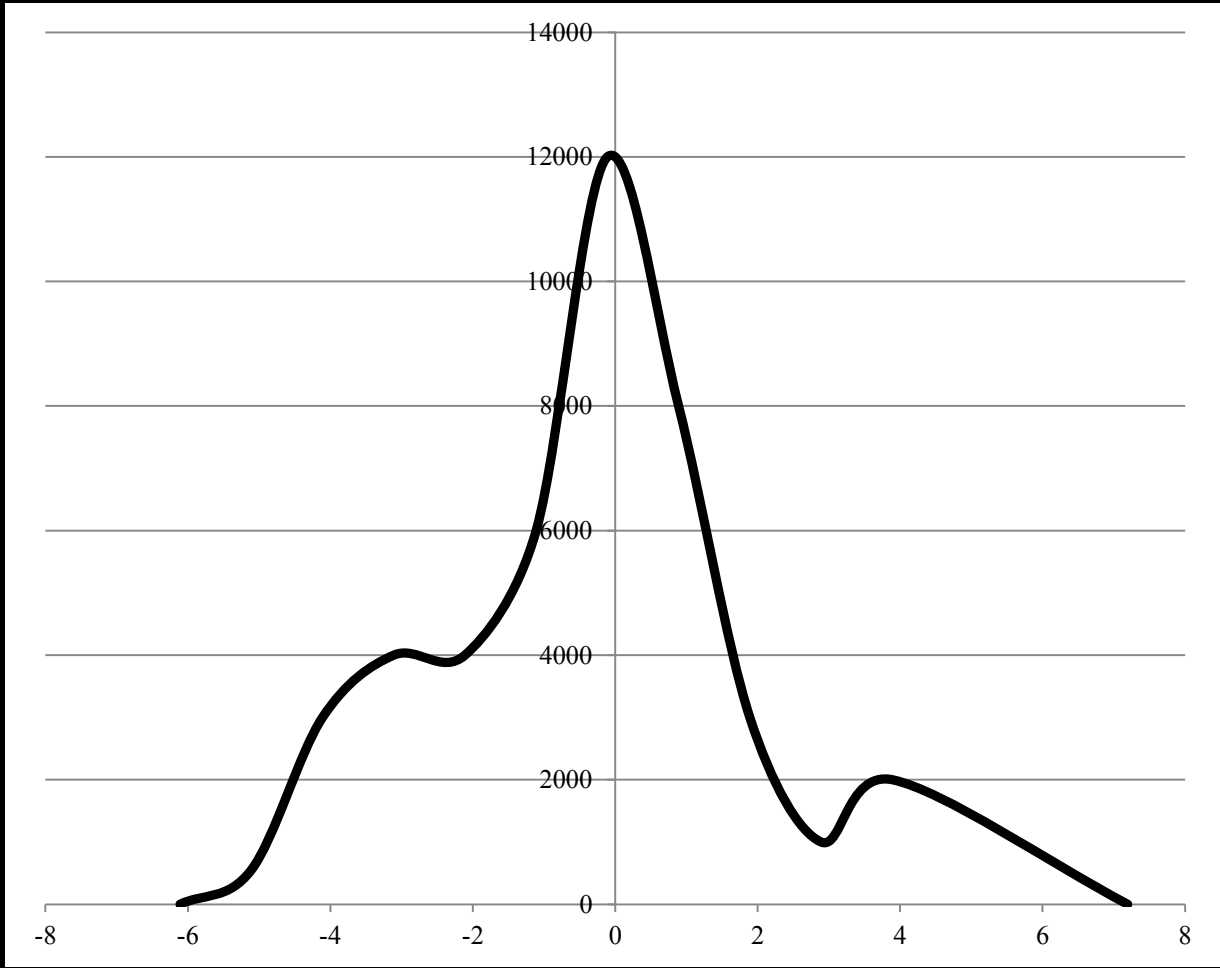
http://www.statisticsviews.com/details/feature/5459931/Five-Fundamental-Concepts-of-Data-Science.html

1. **Begin with the end in mind**
2. **Data Science is Science**
3. **Know thy data**
4. **Love thy data**
5. **Overfitting is a sin**
6. **Honor thy data's first mile and last mile**

(a) The First Mile is the hardest :
…integrating ubiquitous heterogeneous data.

(b) The Last Mile is the hardest :
…extracting actionable intelligence.

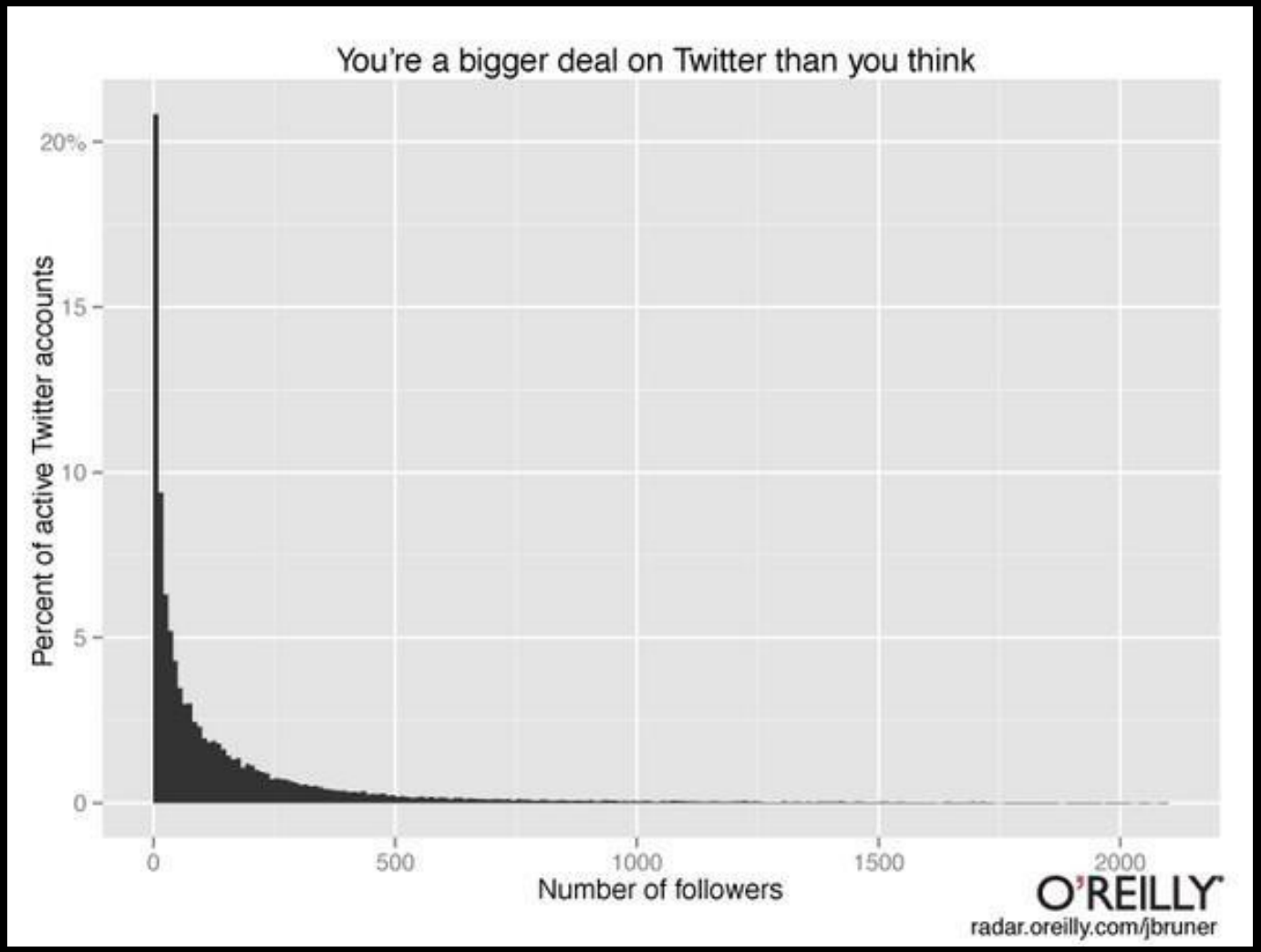
All of the features in the data histogram convey valuable (actionable) information (the long tail, outliers, multi-modal peaks, …)



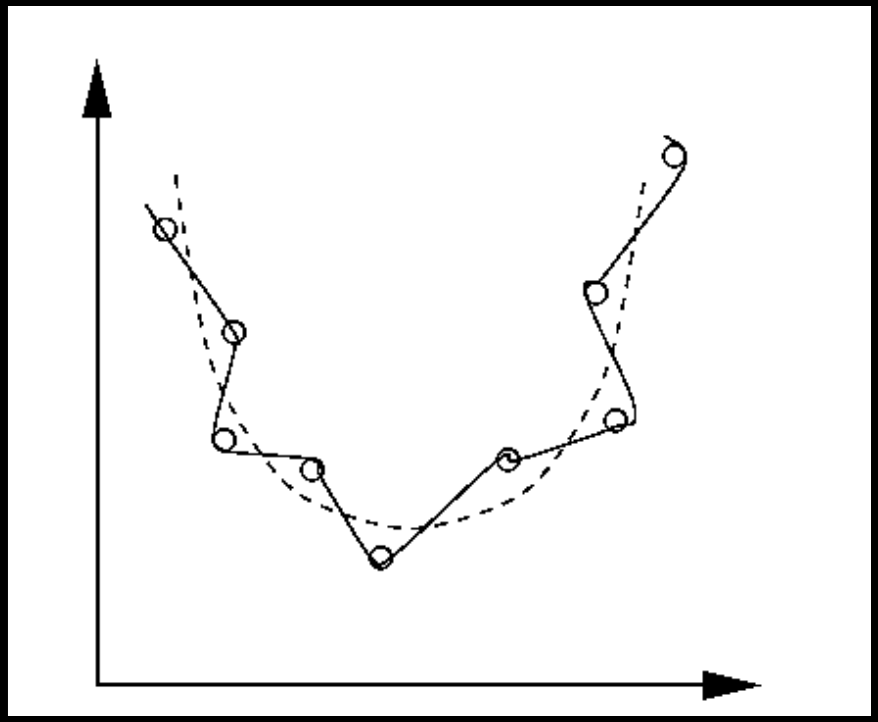
You can discover real insights in the long tail

e.g., the median Twitter account has only 1 follower!

http://radar.oreilly.com/2013/12/tweets-loud-and-quiet.html

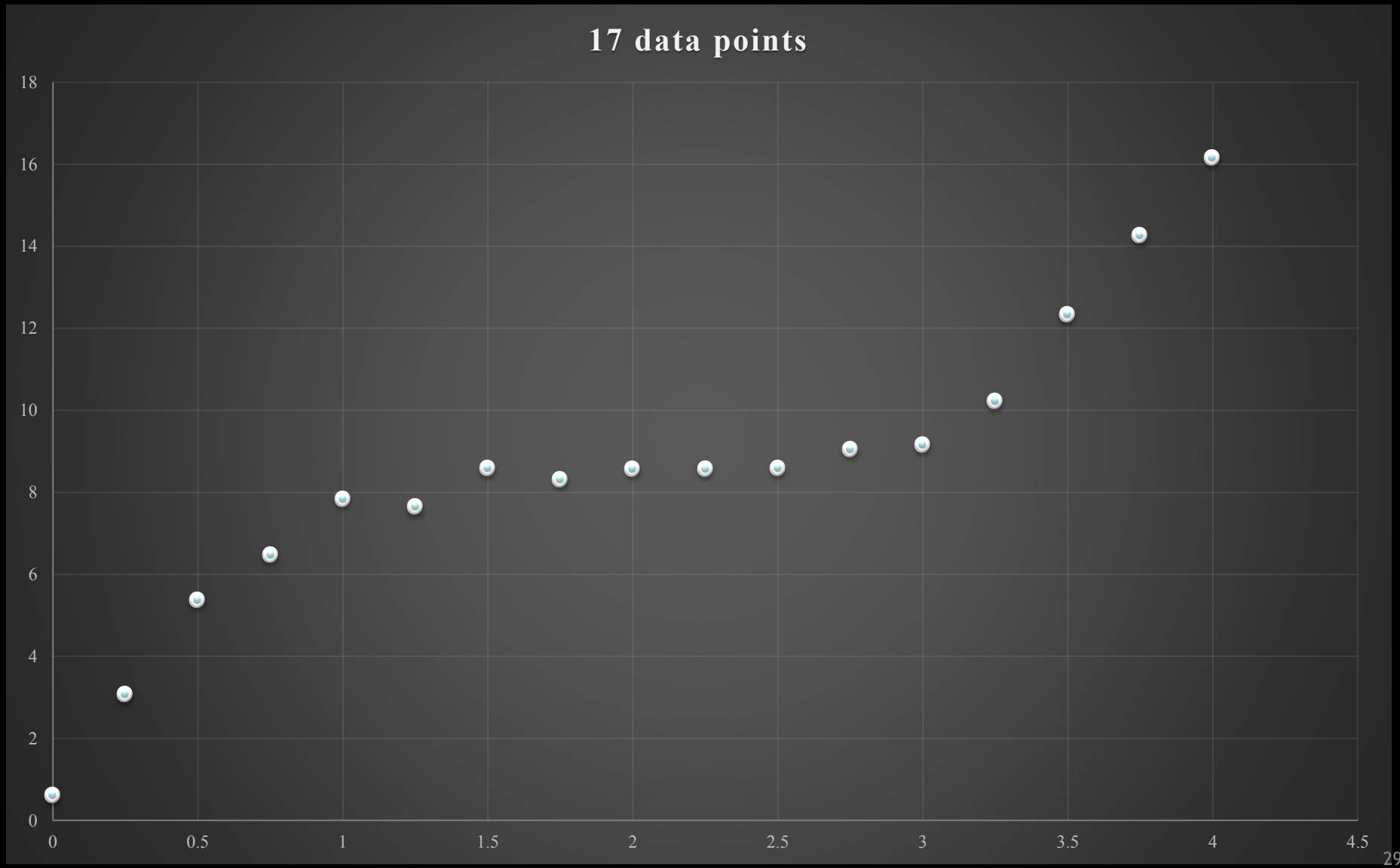


Overfitting is a sin

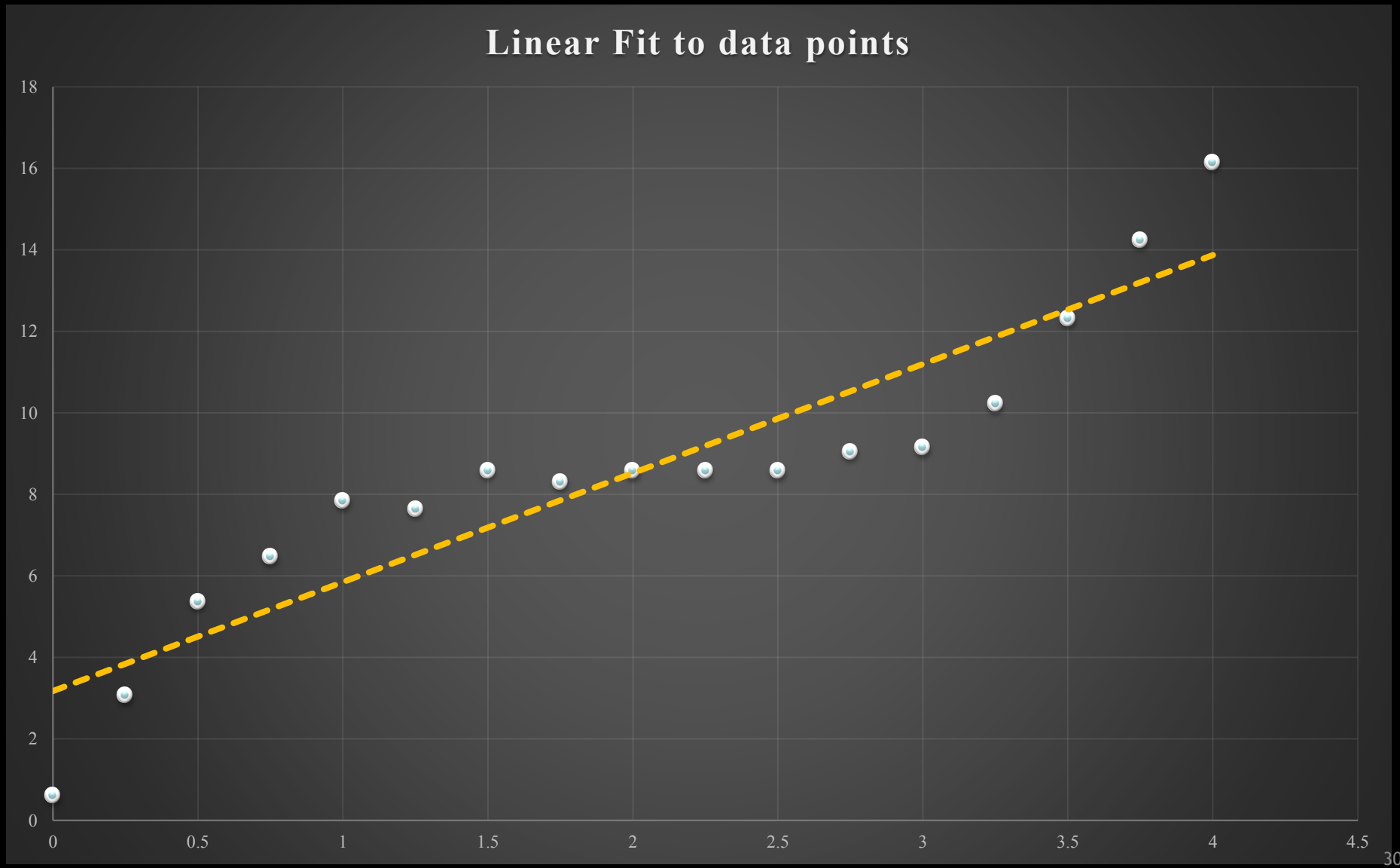


Data Literacy Lesson:

Build a Predictive Model of your data points

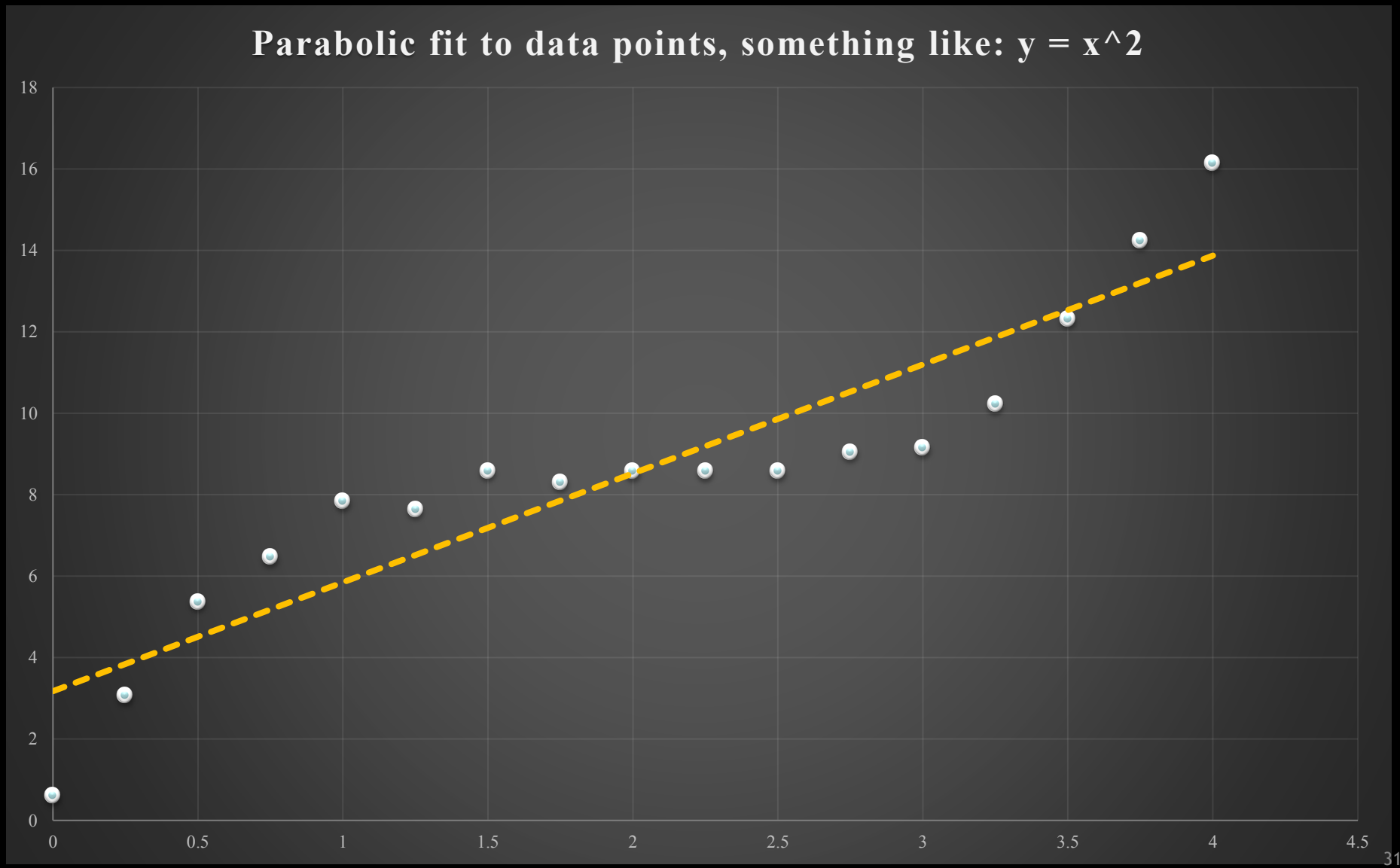


Data Literacy Lesson:
1st attempt – fit a line through the points

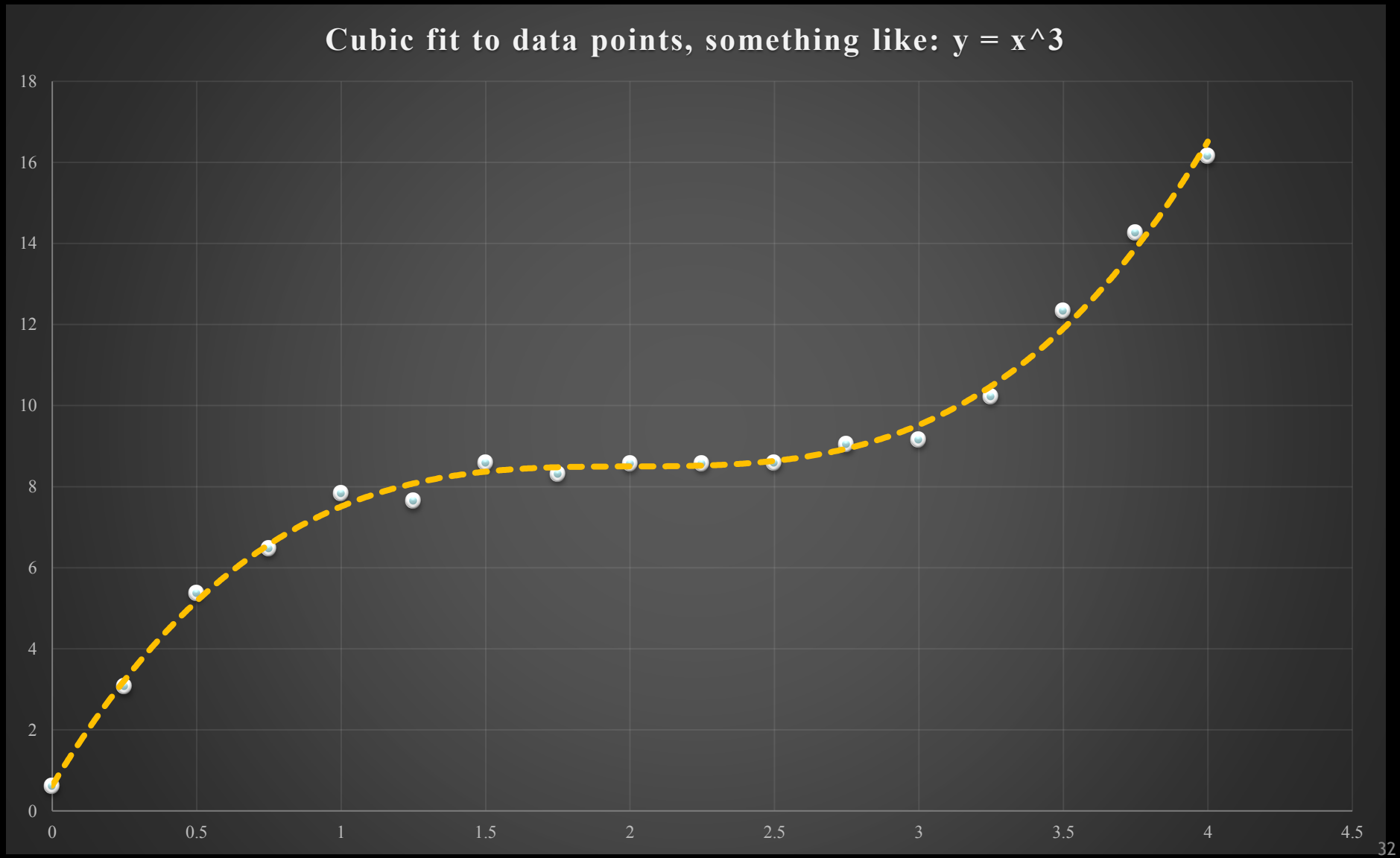


Data Literacy Lesson:

2nd attempt – use a quadratic polynomial

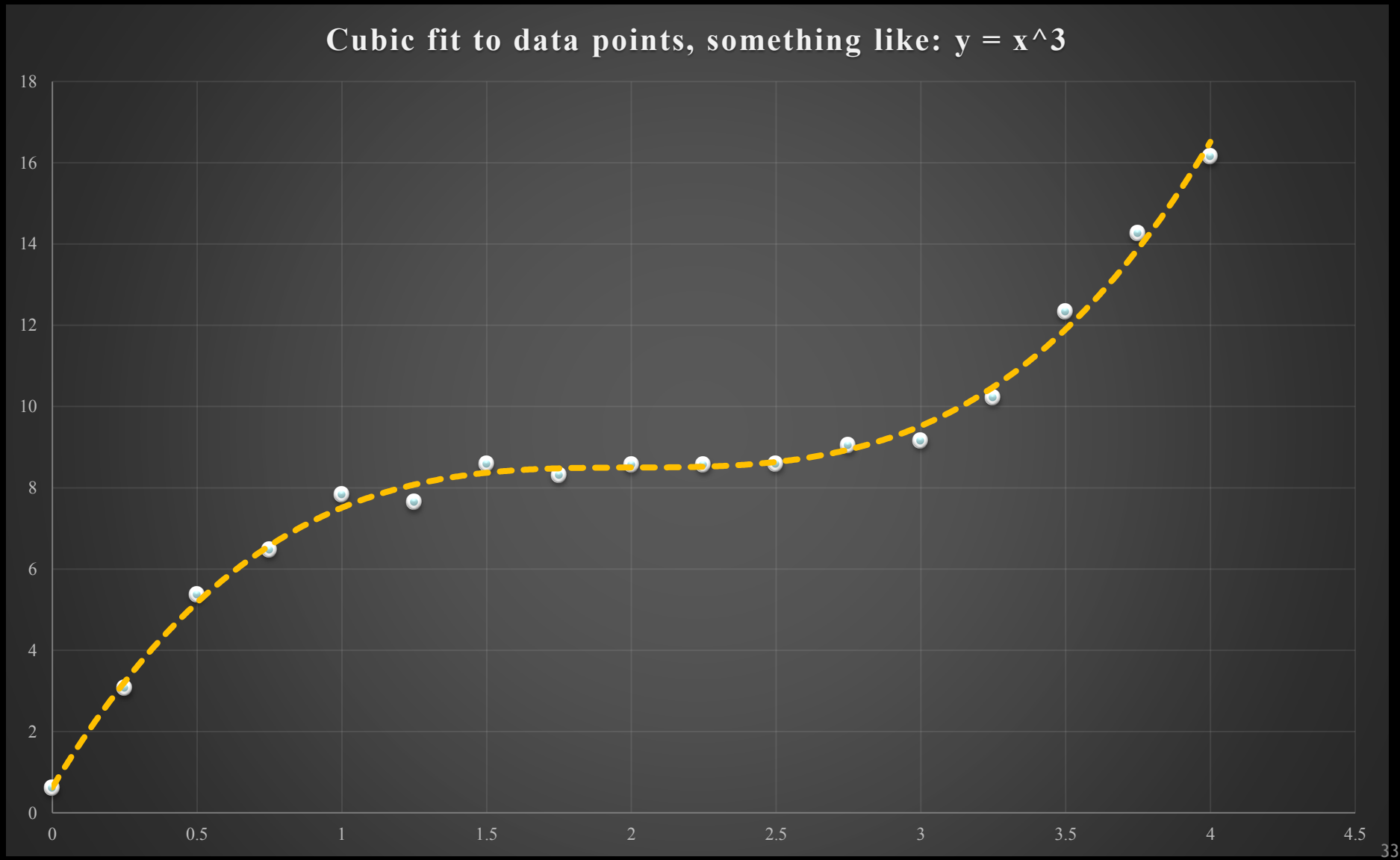


Data Literacy Lesson:
3rd attempt – use a cubic polynomial



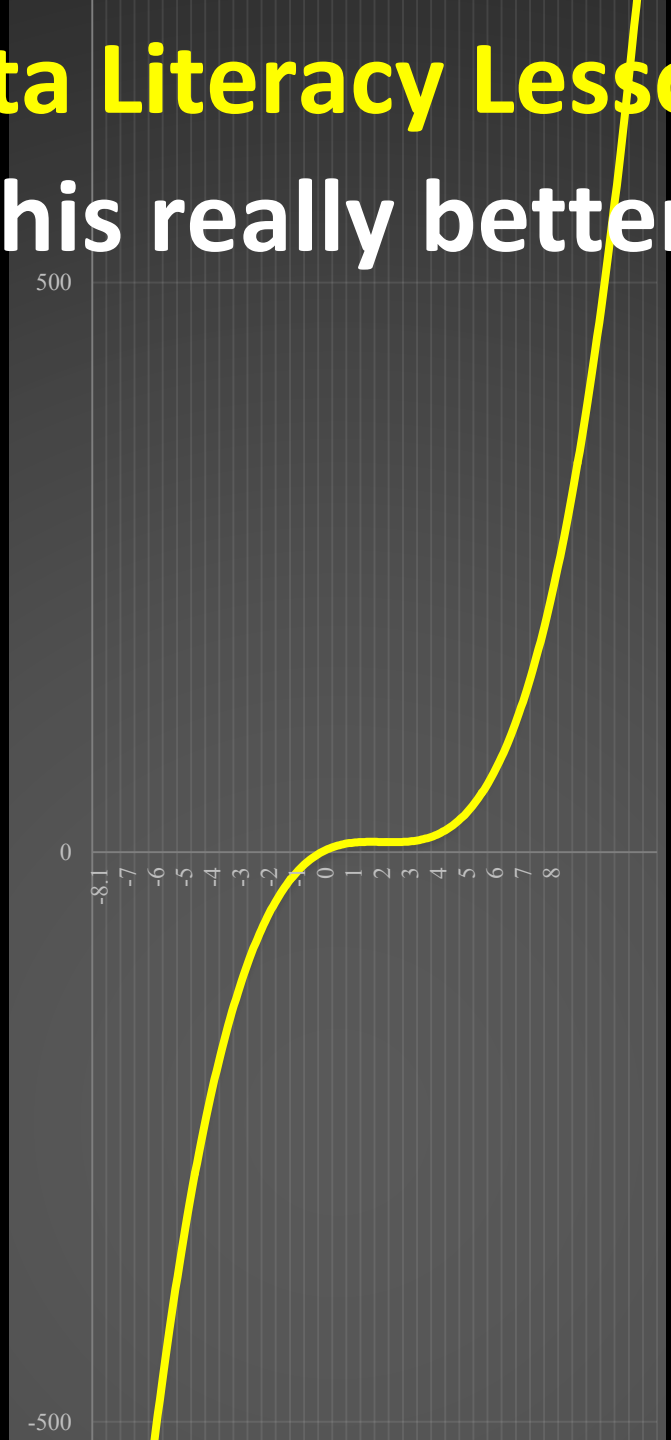
Data Literacy Lesson:

***** Is this really better?? *****



Data Literacy Lesson:

***** Is this really better?? *****

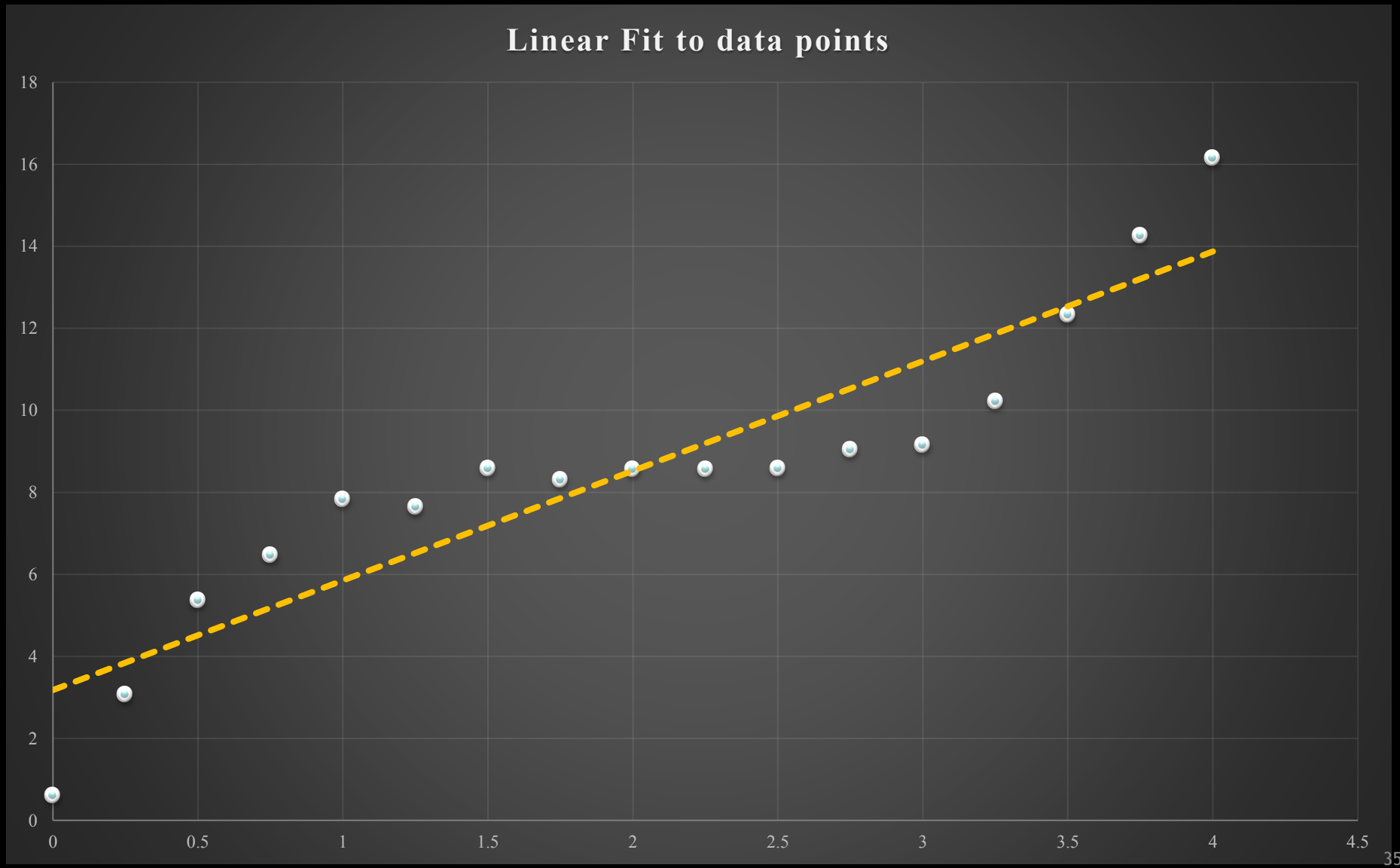


This is example of Overfitting the data!

Data Science rule #1: Don't Overfit Your Model

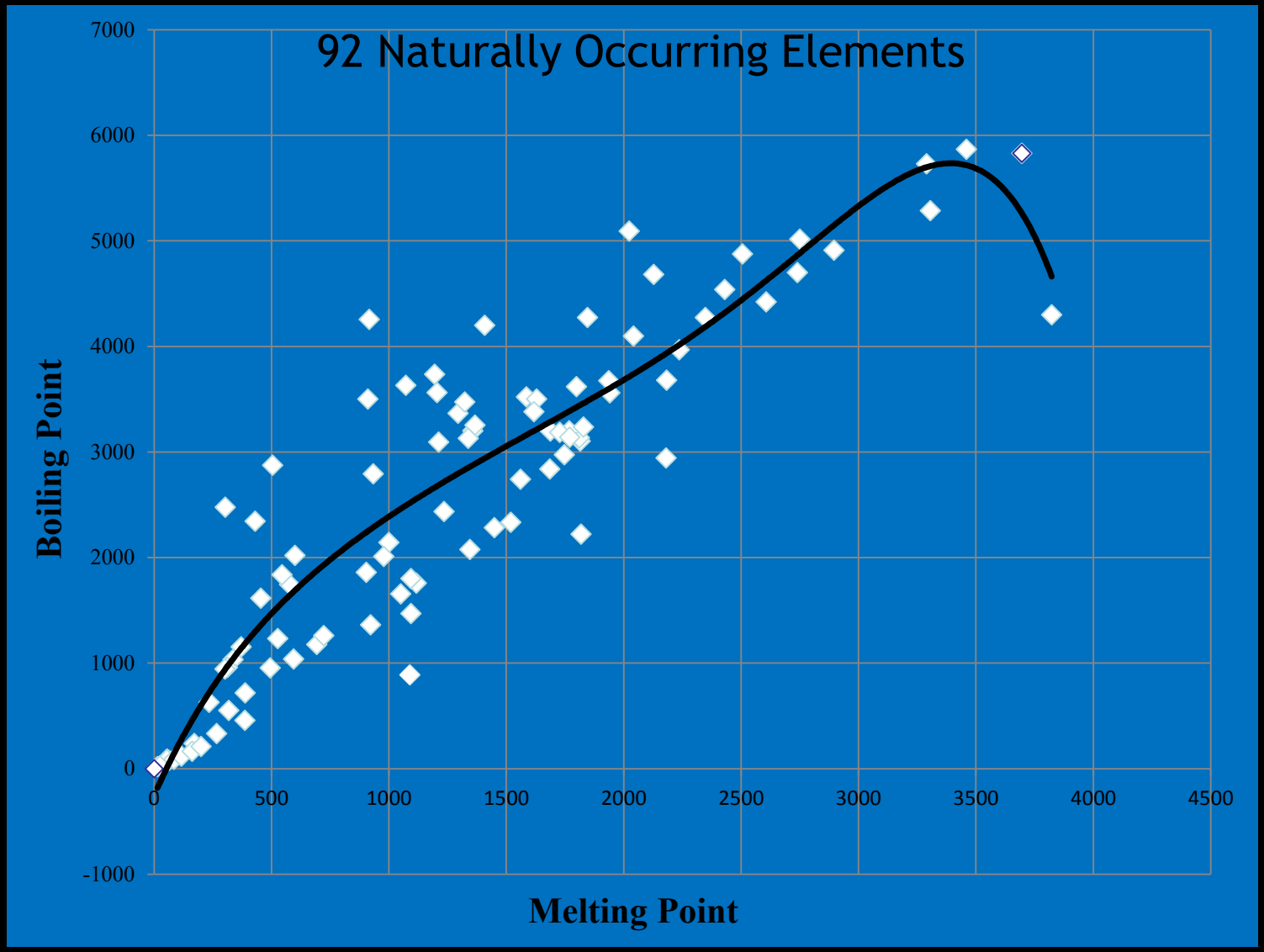
Data Literacy Lesson

1st attempt was a Good Fit to data points



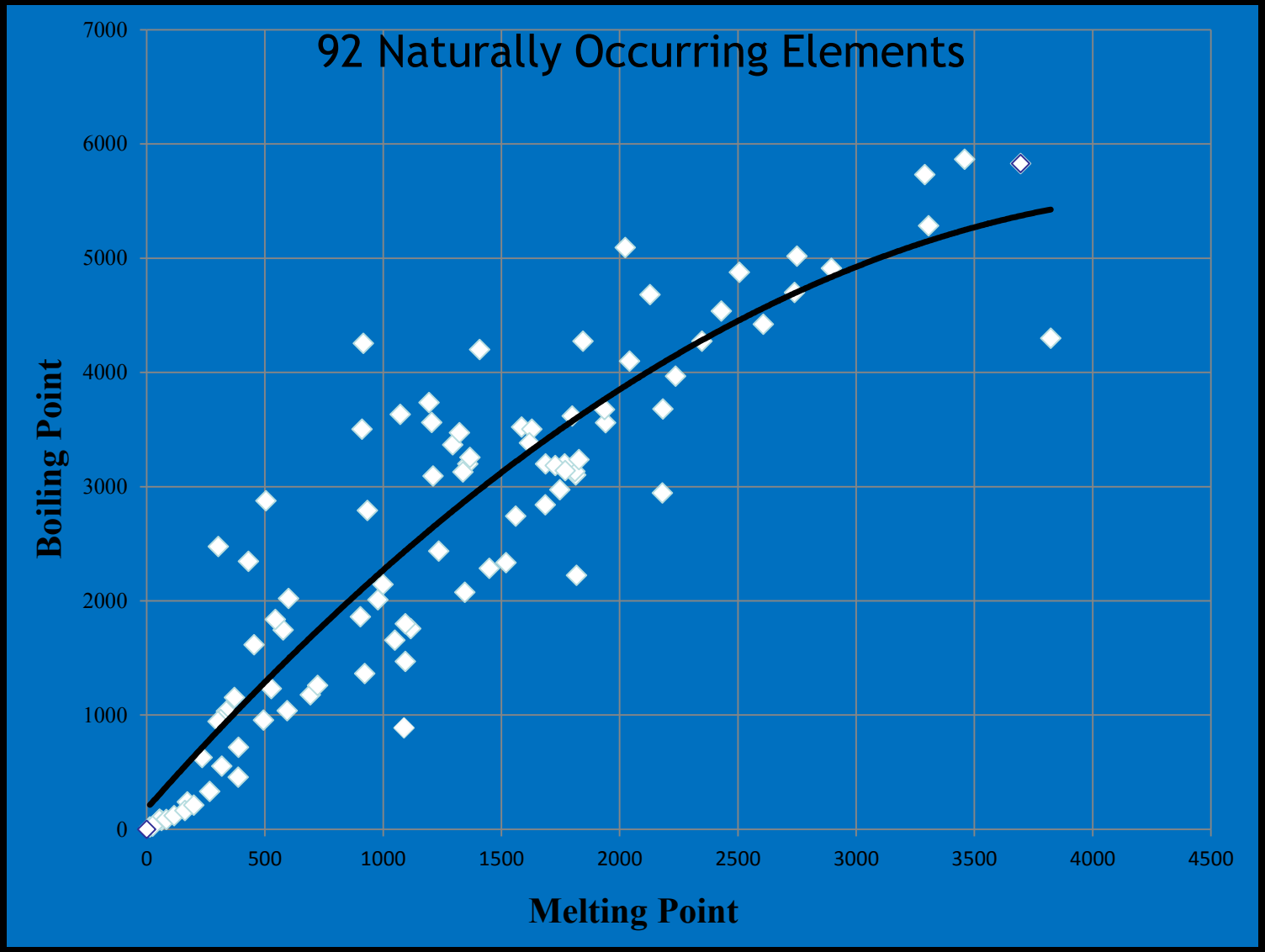
Trend Lines in big data sets: Descriptive Analytics!

It is tempting to over-fit every wiggle in the data.

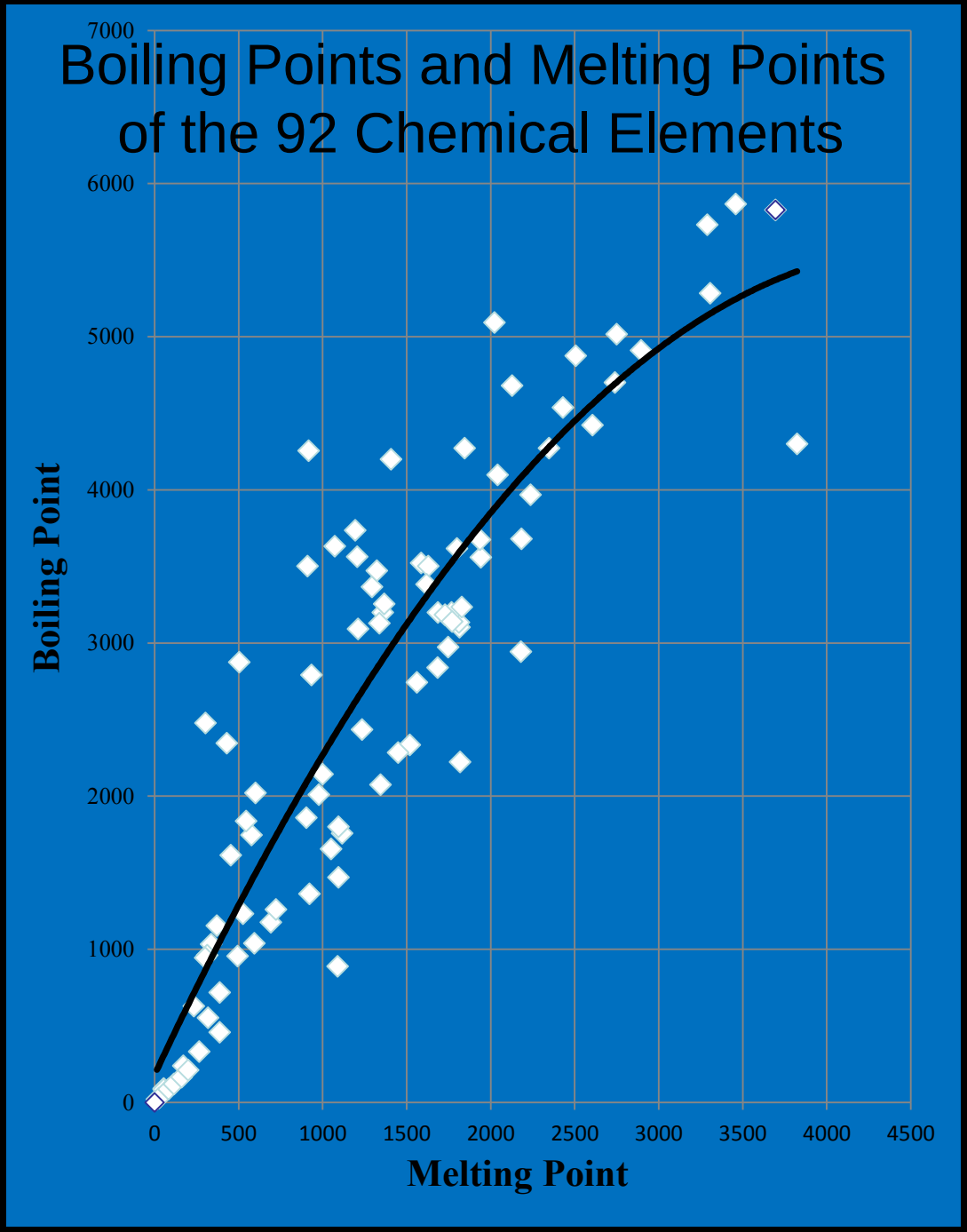


This is a better fit to the trend line...

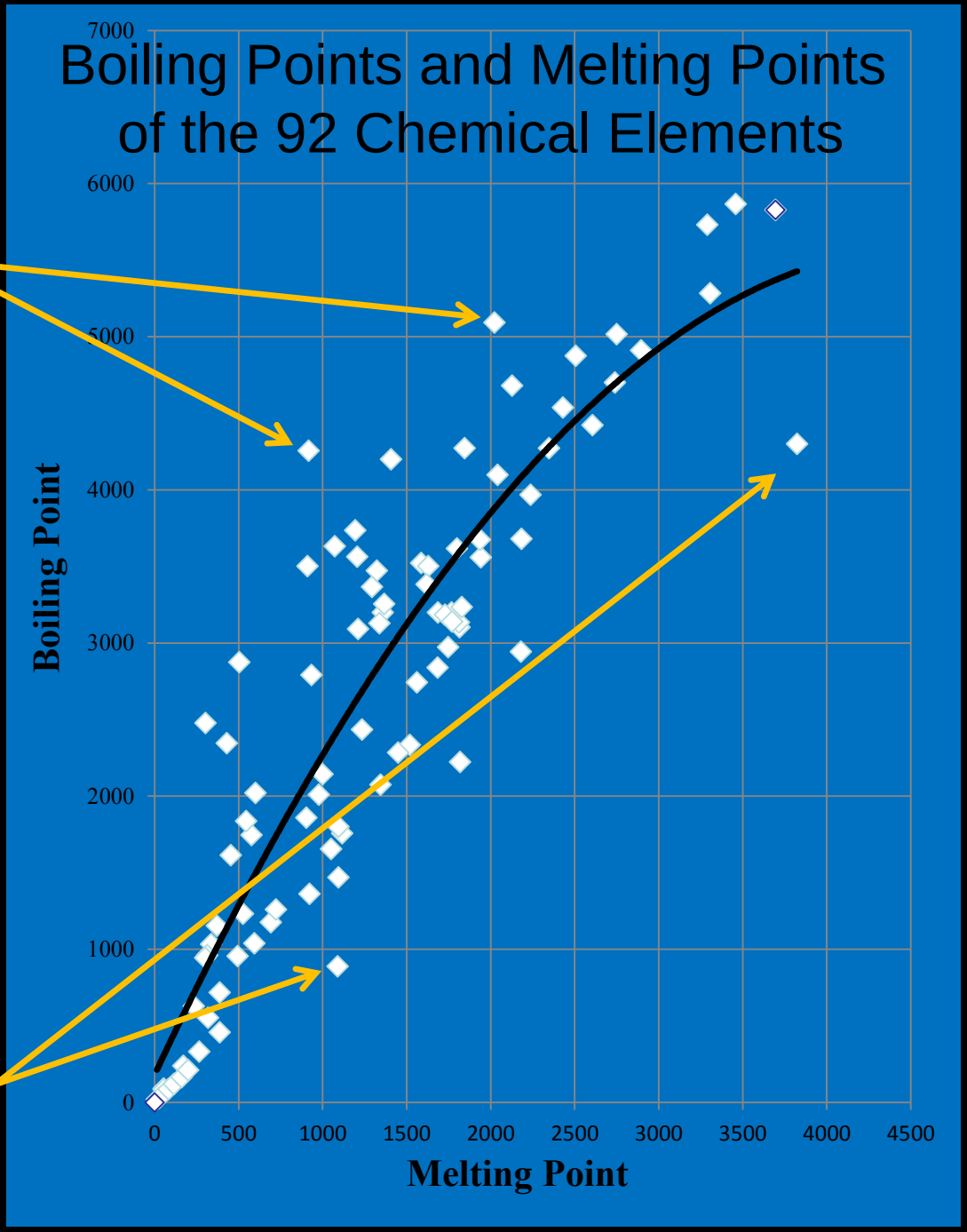
for use in Predictive Analytics!



Trend Line


Boiling Points and Melting Points of the 92 Chemical Elements
Boiling Point
Melting Point
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500

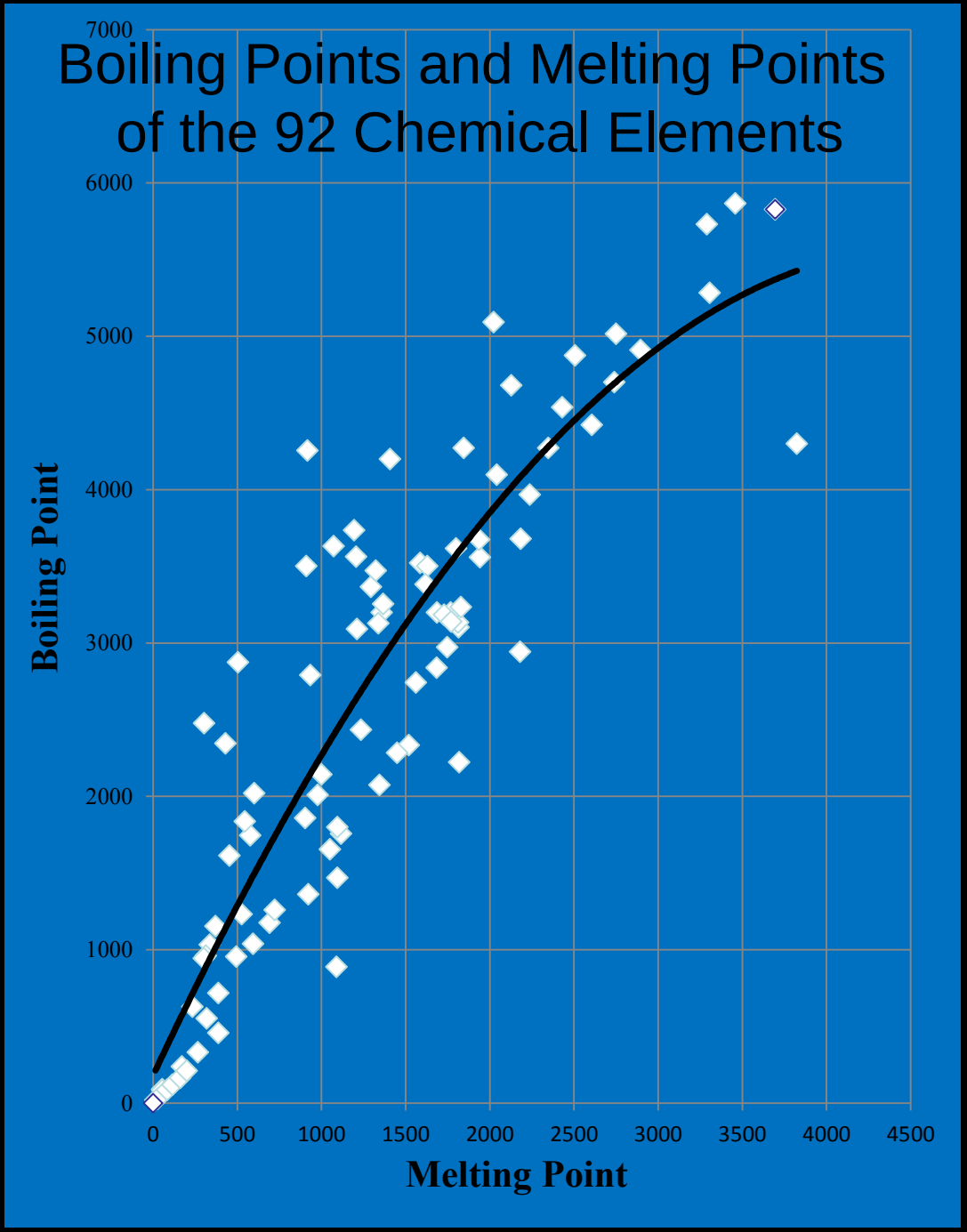
Trend Line and Outliers:



Sometimes we are tempted to think that outliers are just noise or natural variance.

Trend Line and Outliers: where is the real discovery?

Sometimes we are tempted to think that outliers are just noise or natural variance.

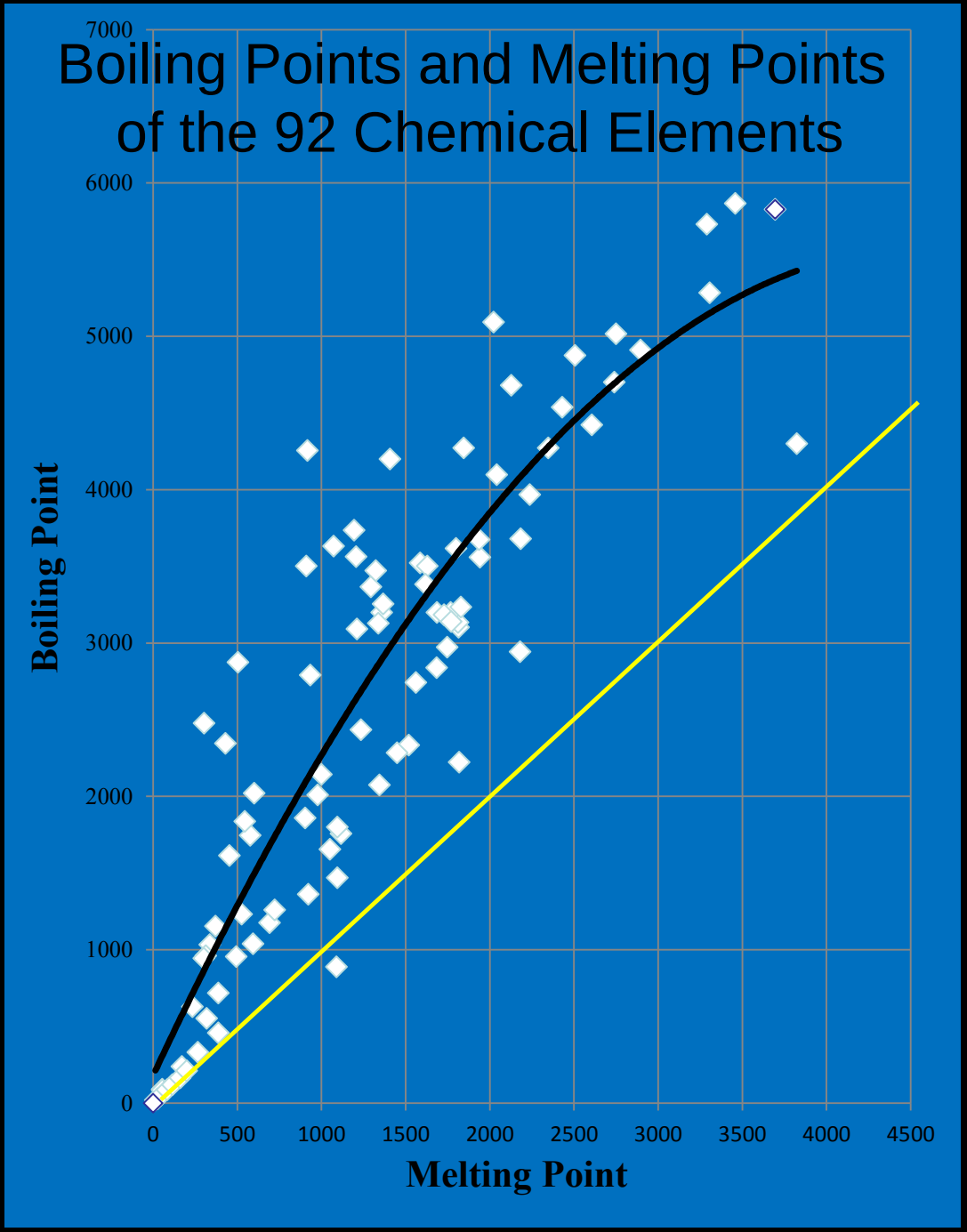


Trend Line and Outliers:

Add some context to the data!

…that diagonal line in the plot (where melting point = boiling point)

… this provides some context (related to your prior knowledge)!

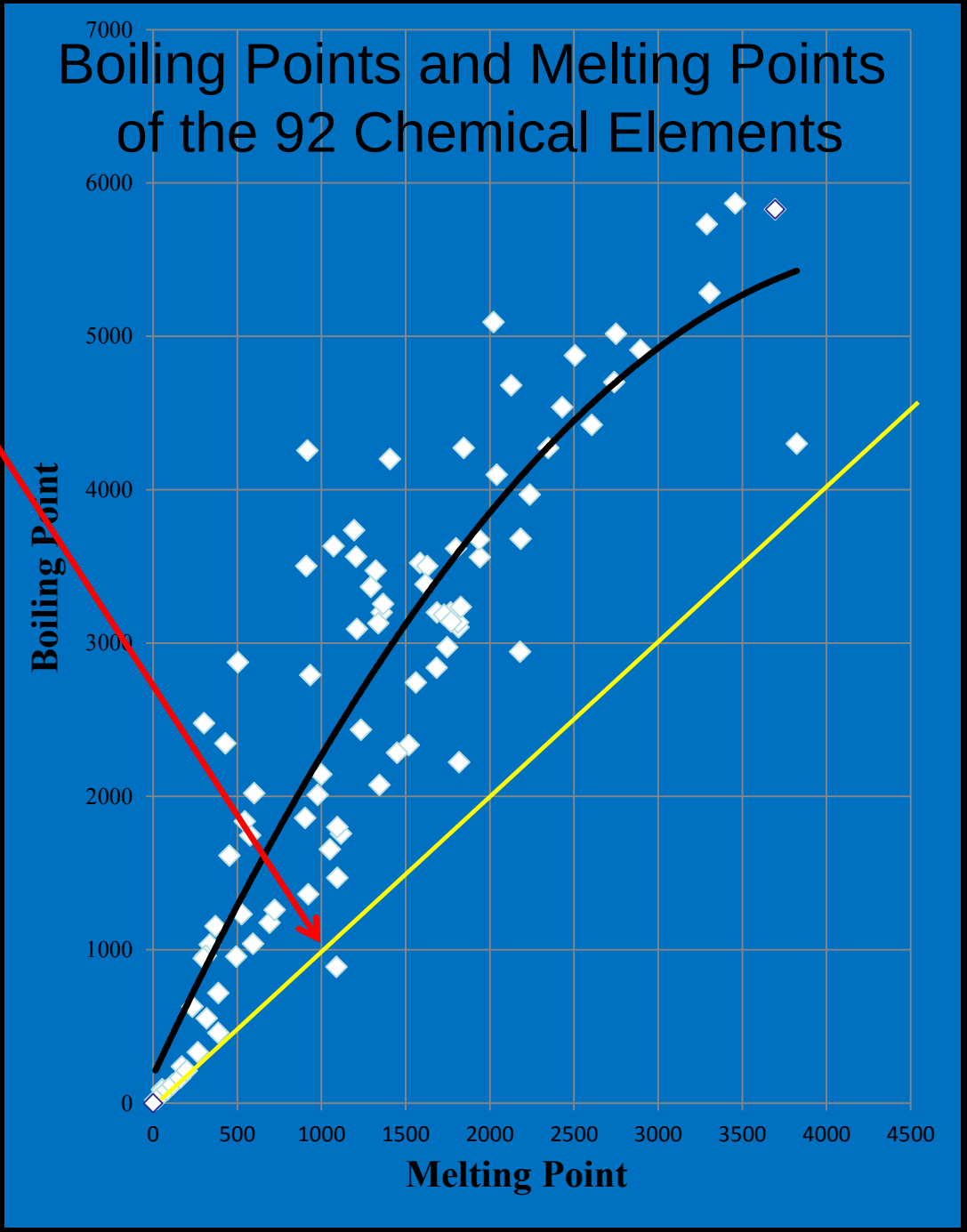


Trend Line and Outliers:

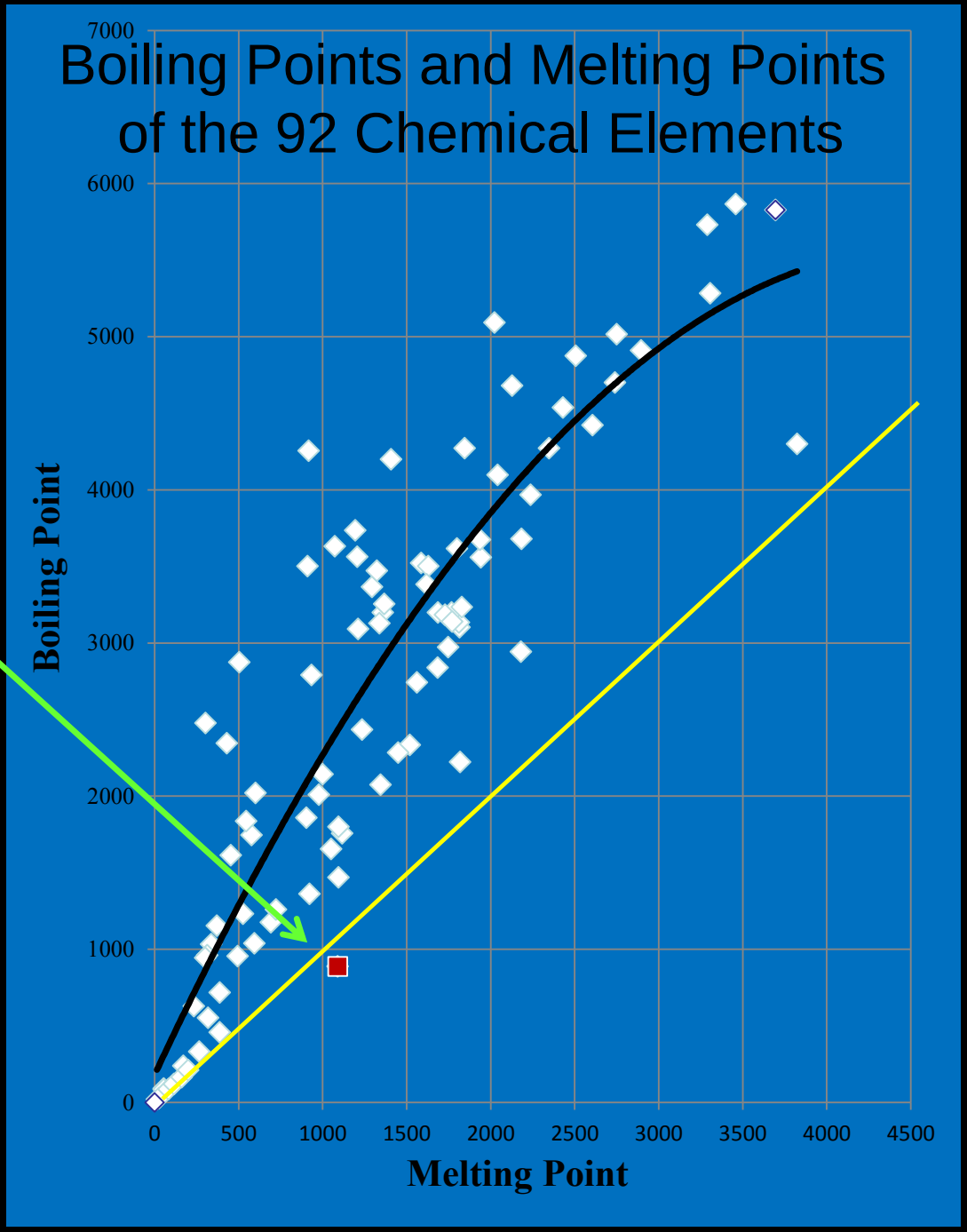
What is that point below the line?

…that diagonal line in the plot (where melting point = boiling point)

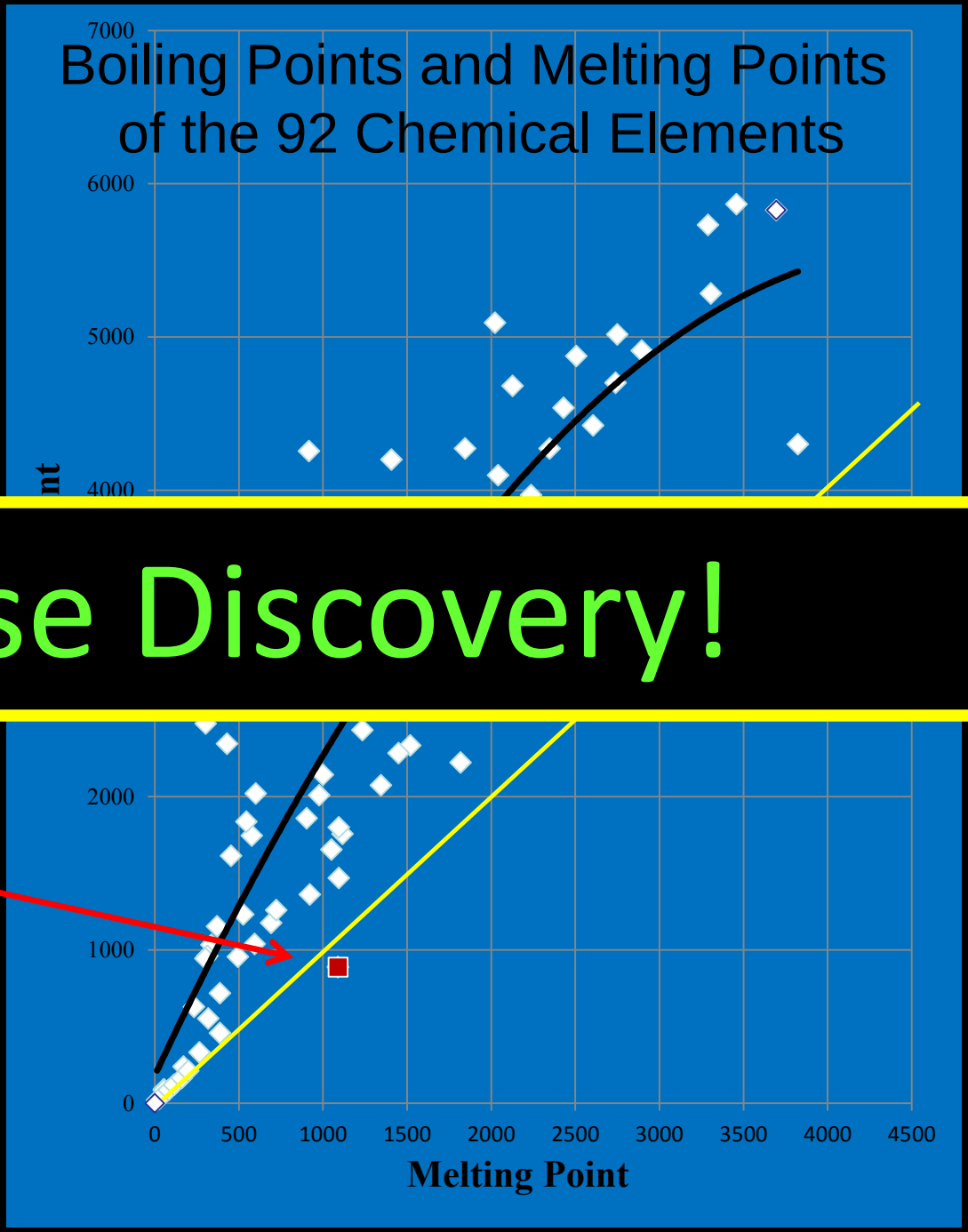
… this provides some context (related to your prior knowledge)!



Trend Line and Outliers: there's the real discovery!



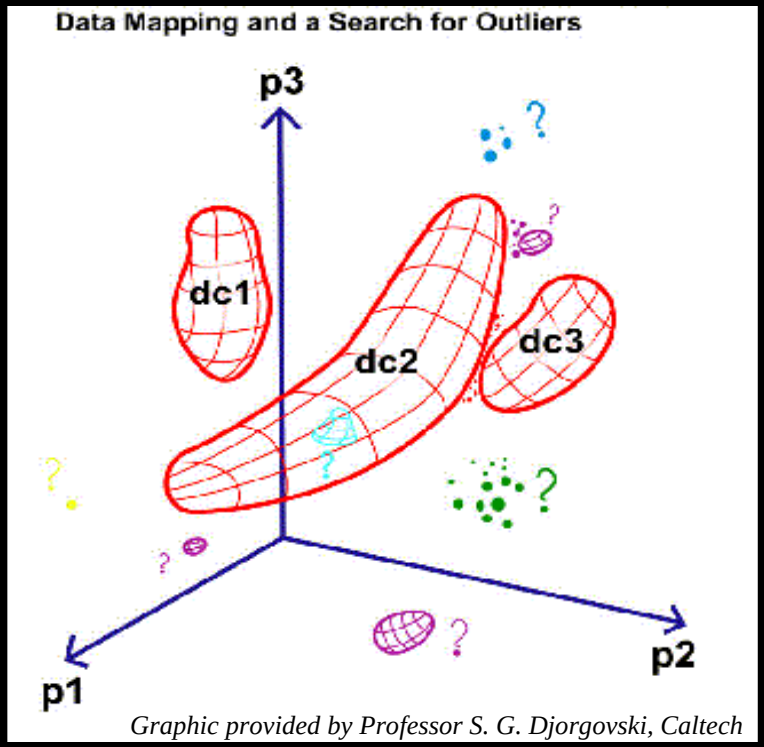
Trend Line and Outliers: there's the real discovery!



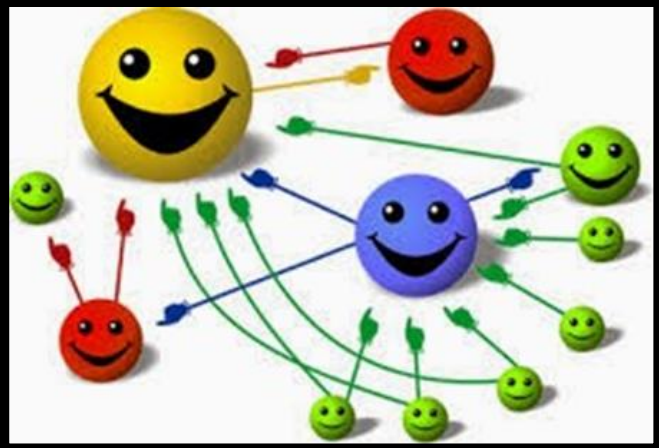
Surprise Discovery!

Melts @ 1089°K
Boils @ 889°K
Arsenic!

Data Literacy Matters:
Using Data for Discovery in 4 flavors



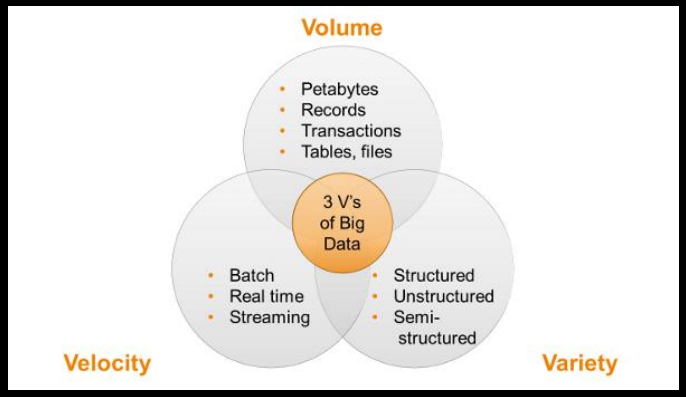
Graphic provided by Professor S. G. Djorgovski, Caltech



1) **Class Discovery:** Finding new classes of objects (population segments), events, and behaviors. This includes: learning the rules that constrain the class boundaries.
2) **Correlation (Predictive and Prescriptive Power) Discovery:** Finding patterns and dependencies, which reveal new governing principles or behavioral patterns (the "DNA").
3) **Novelty (Surprise!) Discovery:** Finding new, rare, one-in-a-[million / billion / trillion] objects and events.
4) **Association (or Link) Discovery:** Finding unusual (improbable) co-occurring associations.

Big Data Challenges

- The 3 V's of Big Data are not just hype – they represent really big challenges:
  1. **Volume**
  2. **Variety**
  3. **Velocity**



- But… **Volume** is not the problem! Storage is manageable.
- Data Science & Analytics (integrating and combining disparate data sources to achieve Data-to-Discovery, Data-to-Decisions, and Data-to-Dividends) are hard…
- … especially on complex (diverse, **high-Variety**) and fast-moving (real-time, **high-Velocity**) data!
- Enabling Advanced Analytics / Data Science capabilities is therefore the key to conquering these challenges.

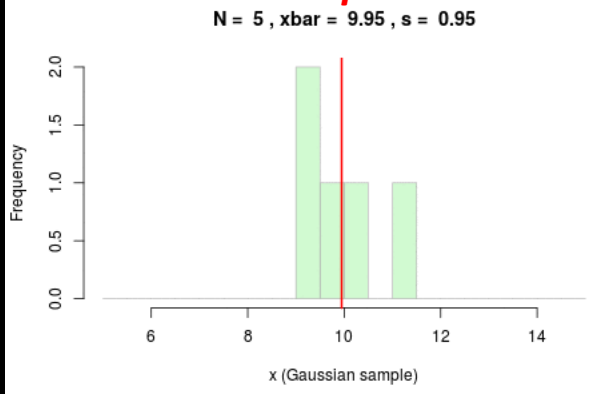
Source for graphic: http://www.vitria.com/blog/Big-Data-Analytics-Challenges-Facing-All-Communications-Service-Providers/

46

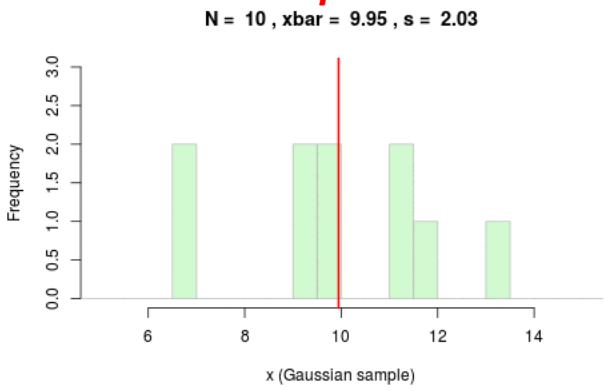
Big Data **Volume** is great news…

… more data means less uncertainty, and more laser-focused insights & intelligence!

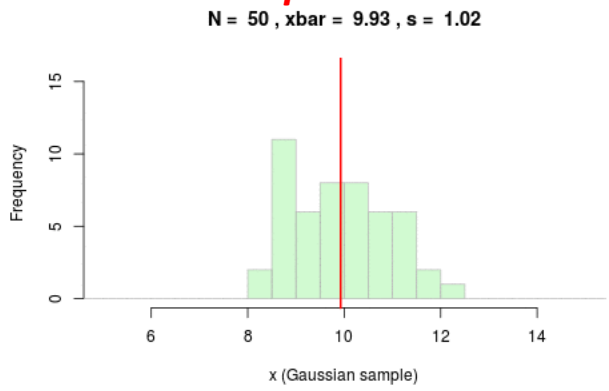
5 data points



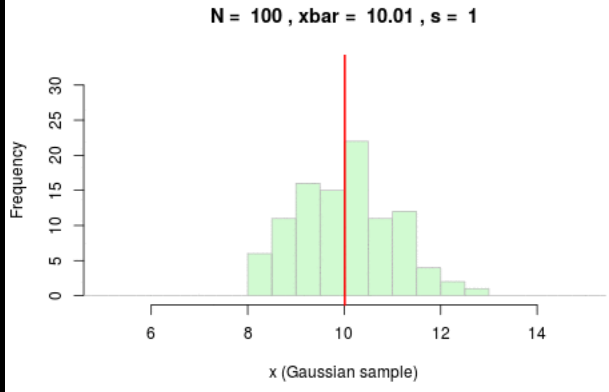
10 data points



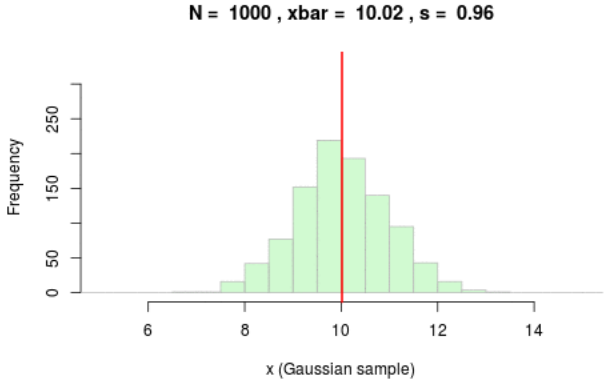
50 data points



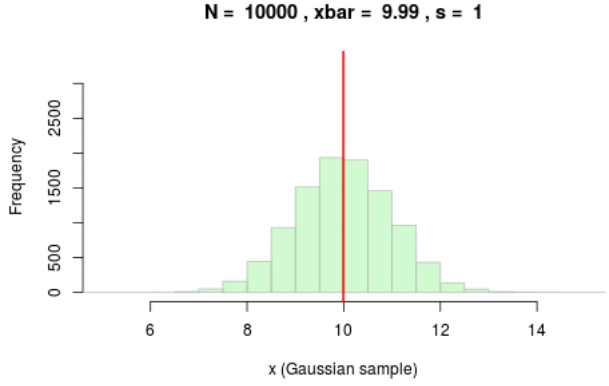
100 data points



1000 data points



10000 data points



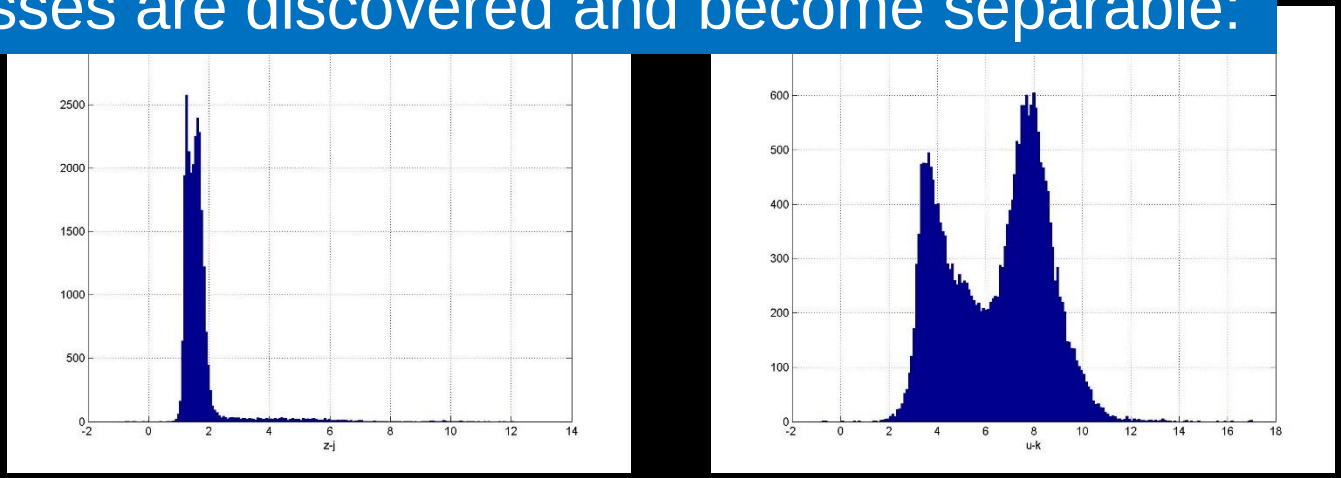
Source for graphics: https://rexplorations.wordpress.com/2015/09/05/animated-mean-and-sample-size/

47

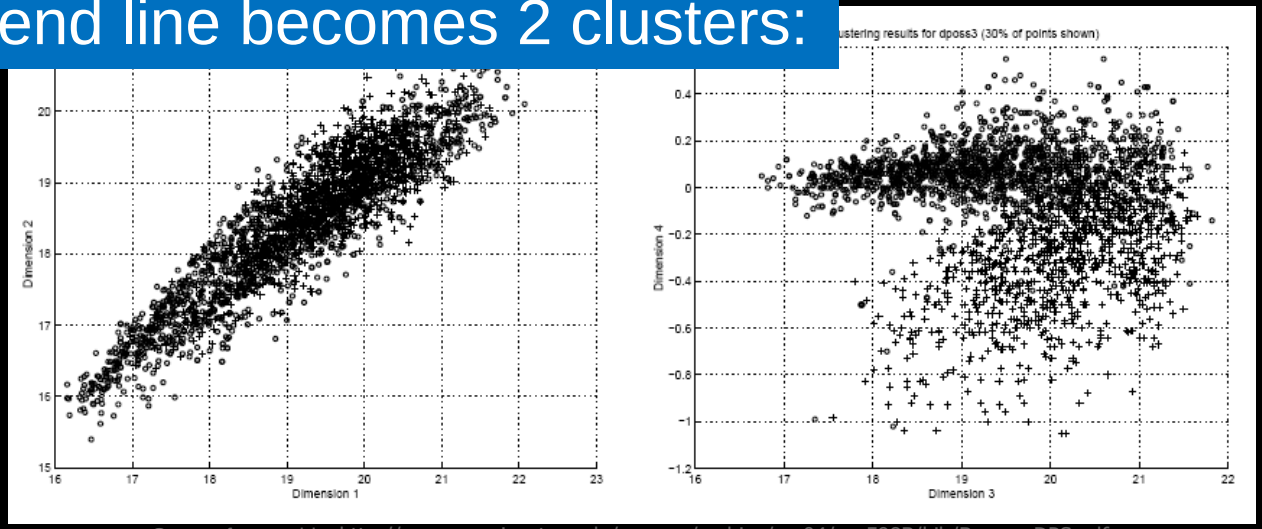
…but the greatest of V's is Variety

The discovery and separation of classes improves when a sufficient number of "correct" features are available for exploration:

(a) 2 classes are discovered and become separable:



(b) One trend line becomes 2 clusters:



Source for graphic: http://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/BrunnerDPS.pdf

48

Clustering

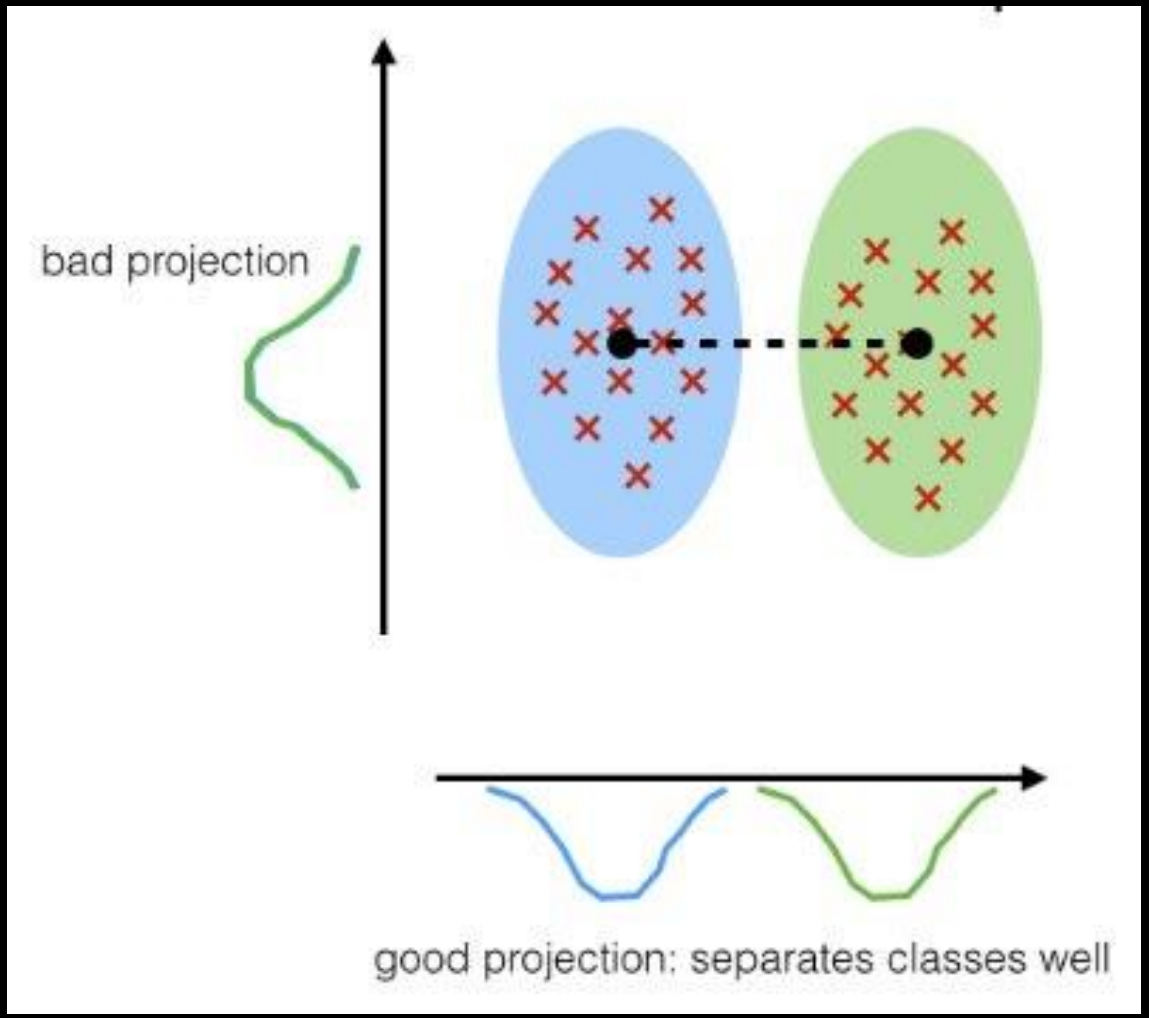
(for class discovery, segmentation, personalization, SegOne marketing,…)

Exploiting the 3rd V of Big Data
(Data Exploration and Data Exploitation)

1. Volume
2. Velocity
3. Variety

49

Feature Selection and Projection



Feature Selection is important to disambiguate different classes. More importantly, **Class Discovery** depends on selecting the right features!

50

Feature Selection and Model Bias: choosing features in the dark

I picked out two socks from my sock drawer this morning!

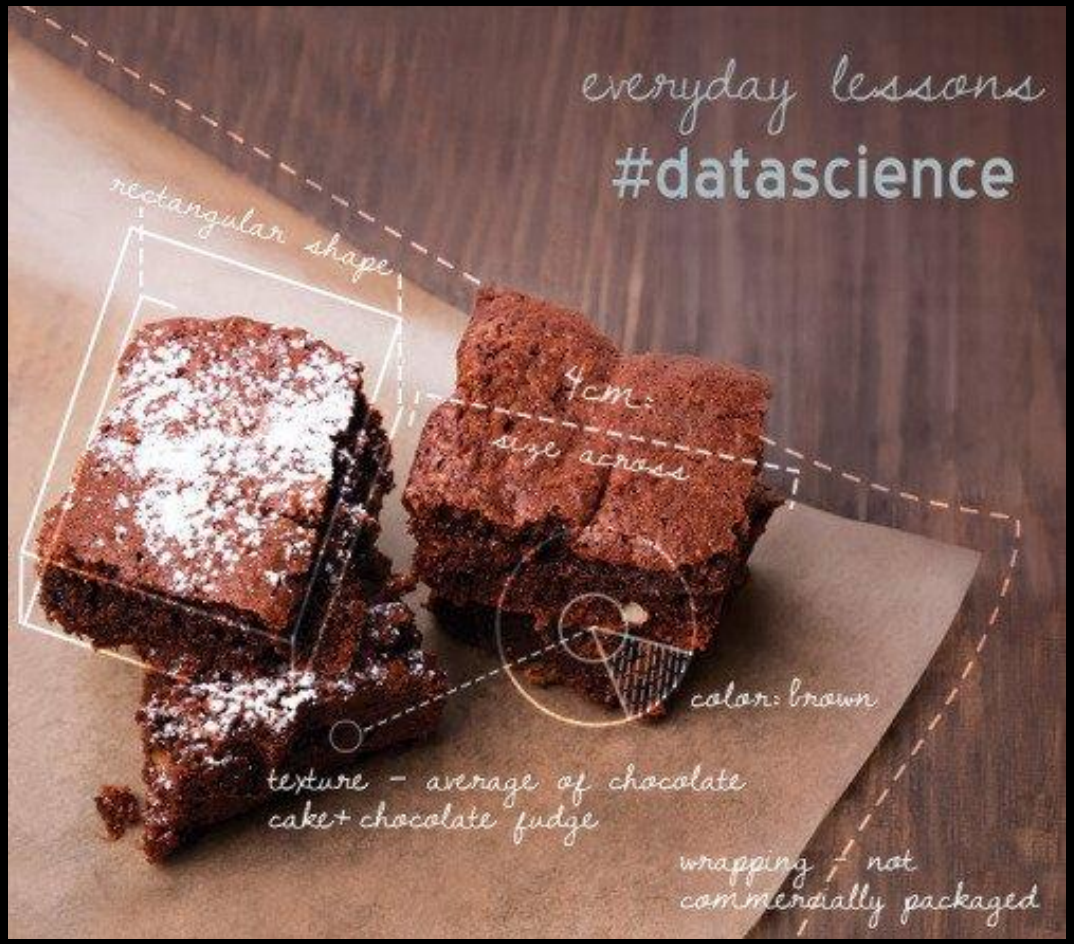
It was still dark, but that shouldn't matter, right? After all, they are the **same size** … **THE SAME ?!?**

The Era of Big Data represents the END OF DEMOGRAPHICS (*i.e.,* our models should no longer be based on and biased by a limited selection of attributes and features)

51

High-Variety Data enables better (and tastier) analytics models

Variety is the spice of discovery!



The Data Science of Feature-rich Chocolate Brownies

http://www.datasciencebowl.com/data-science-of-chocolate-brownies/

52

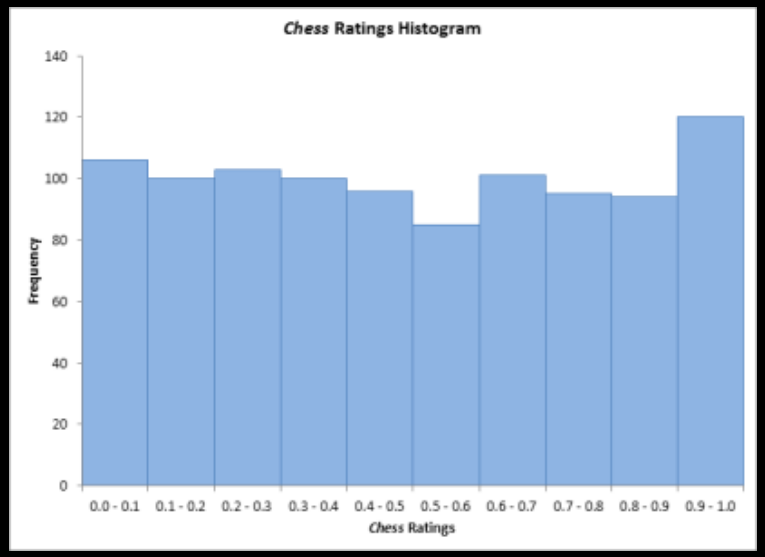
A Statistical Data Puzzle

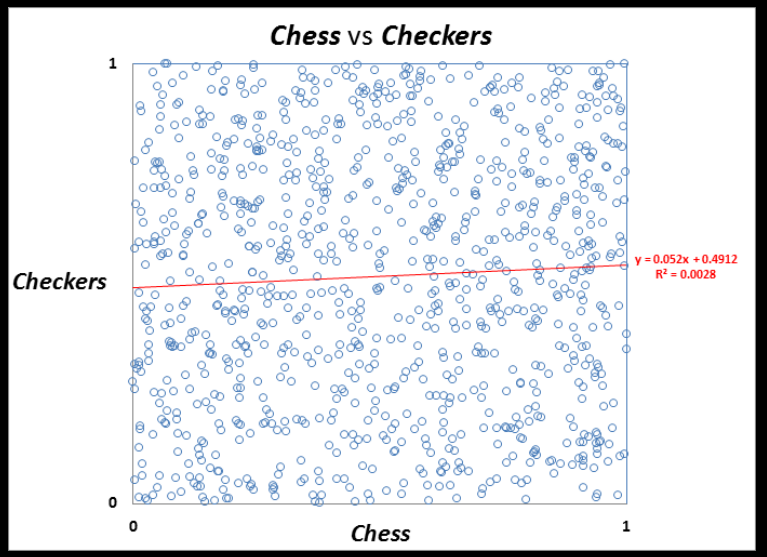
53

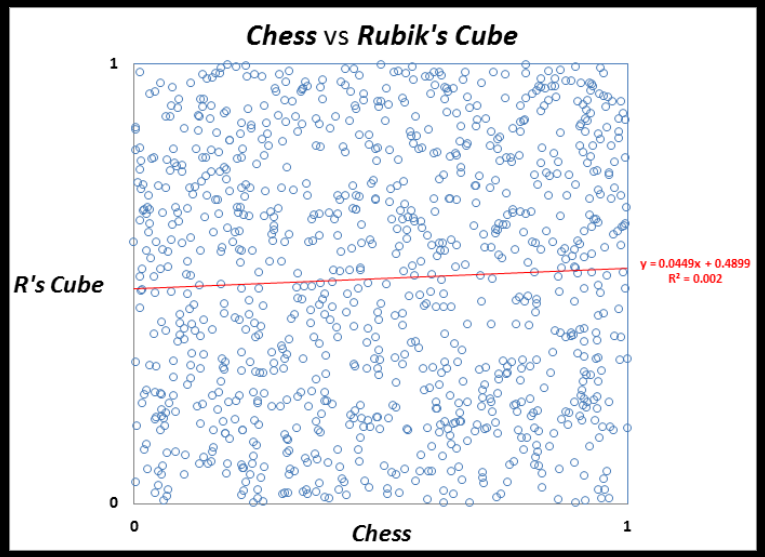
The Island of Games Puzzle:

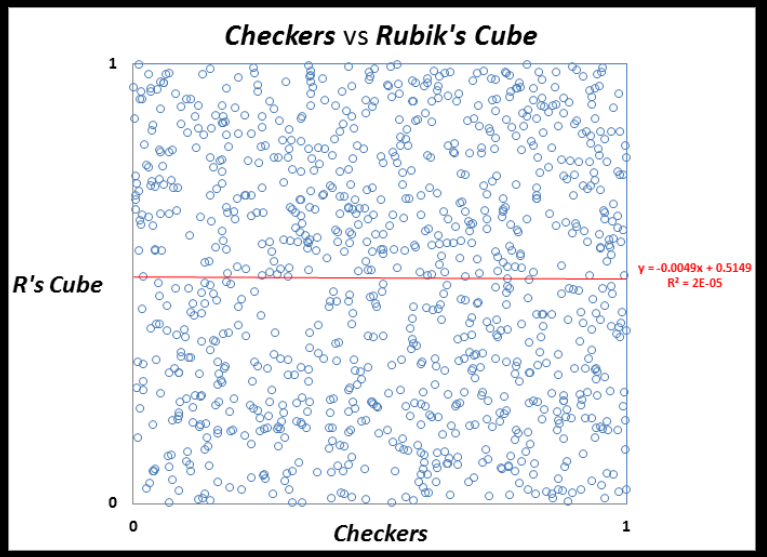
Can you find a pattern in the player ratings data?

http://www.datasciencecentral.com/profiles/blogs/island-of-games-puzzle-problem-statement









54

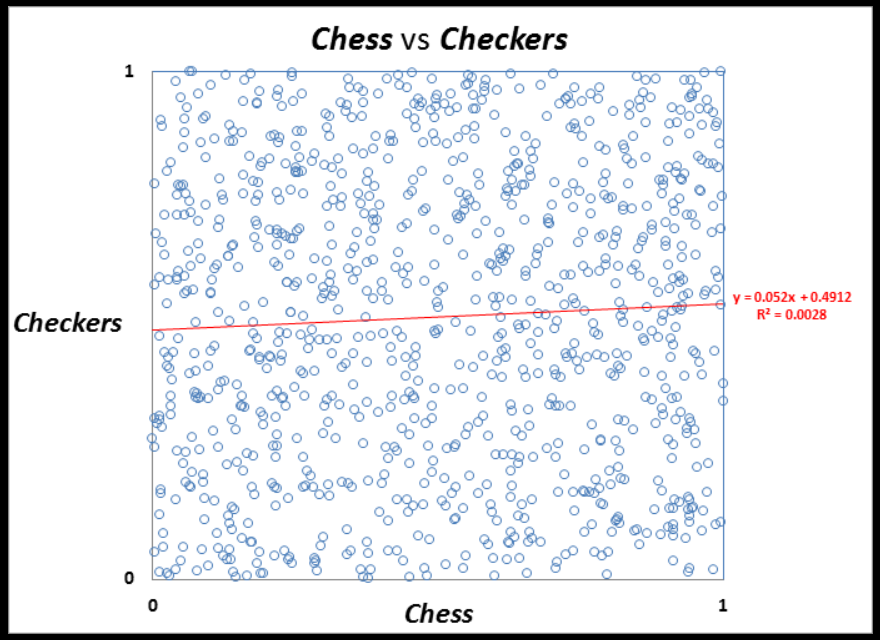
Solution to the Island of Games Puzzle

55

Island of Games

Color-coding the player ratings data distribution

Green and Red = High Rubik's Cube ranking (> 0.5) ; Blue and Yellow = Low Rubik's Cube ranking (< 0.5)

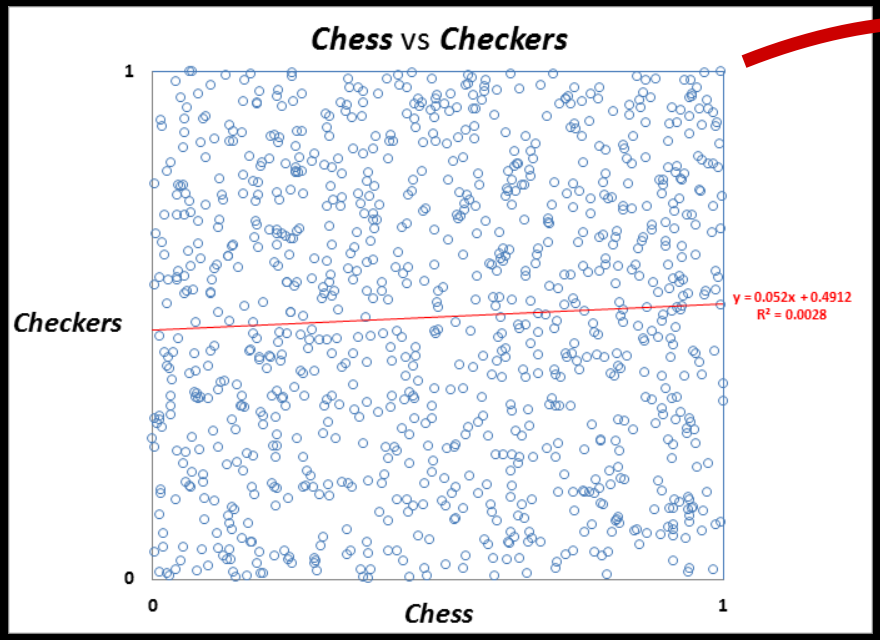


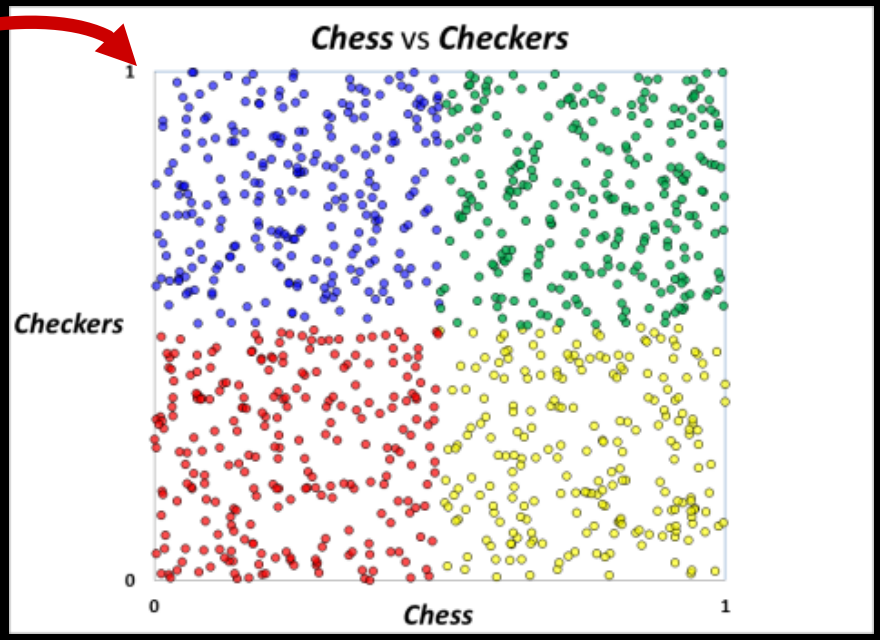
The intrinsic **patterns** in the player ratings data are not revealed in 2-D scatter plots or by using traditional statistical methods.

Island of Games

Color-coding the player ratings data distribution

Green and Red = High Rubik's Cube ranking (> 0.5) ; Blue and Yellow = Low Rubik's Cube ranking (< 0.5)

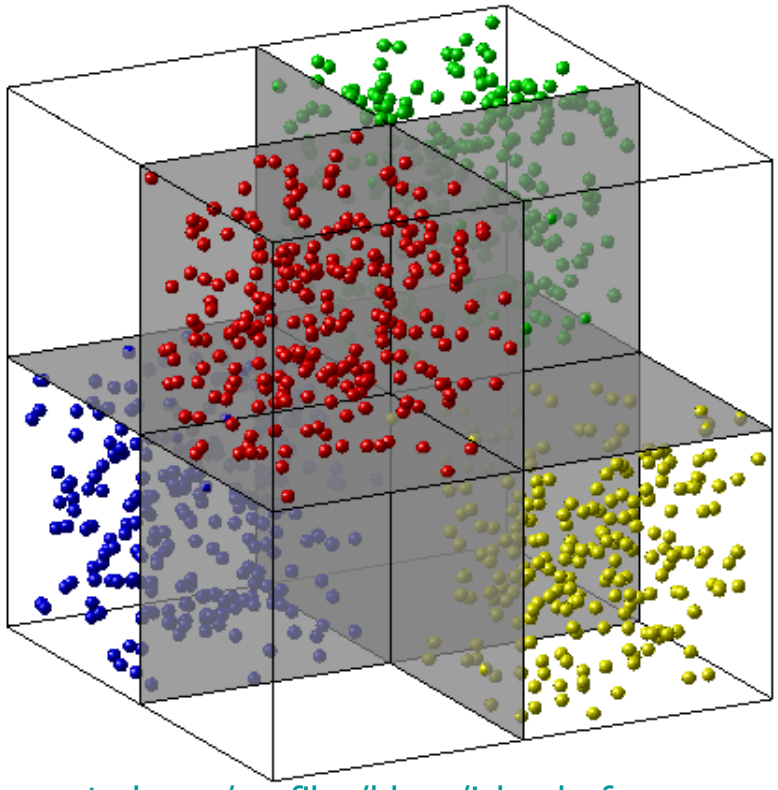




The intrinsic **patterns** in the player ratings data are not revealed in 2-D scatter plots or by using traditional statistical methods. However, exploration in the 3-D input parameter space (of player ratings for 3 games) reveals the actual player groupings…

3-D view of the Player Ratings Data

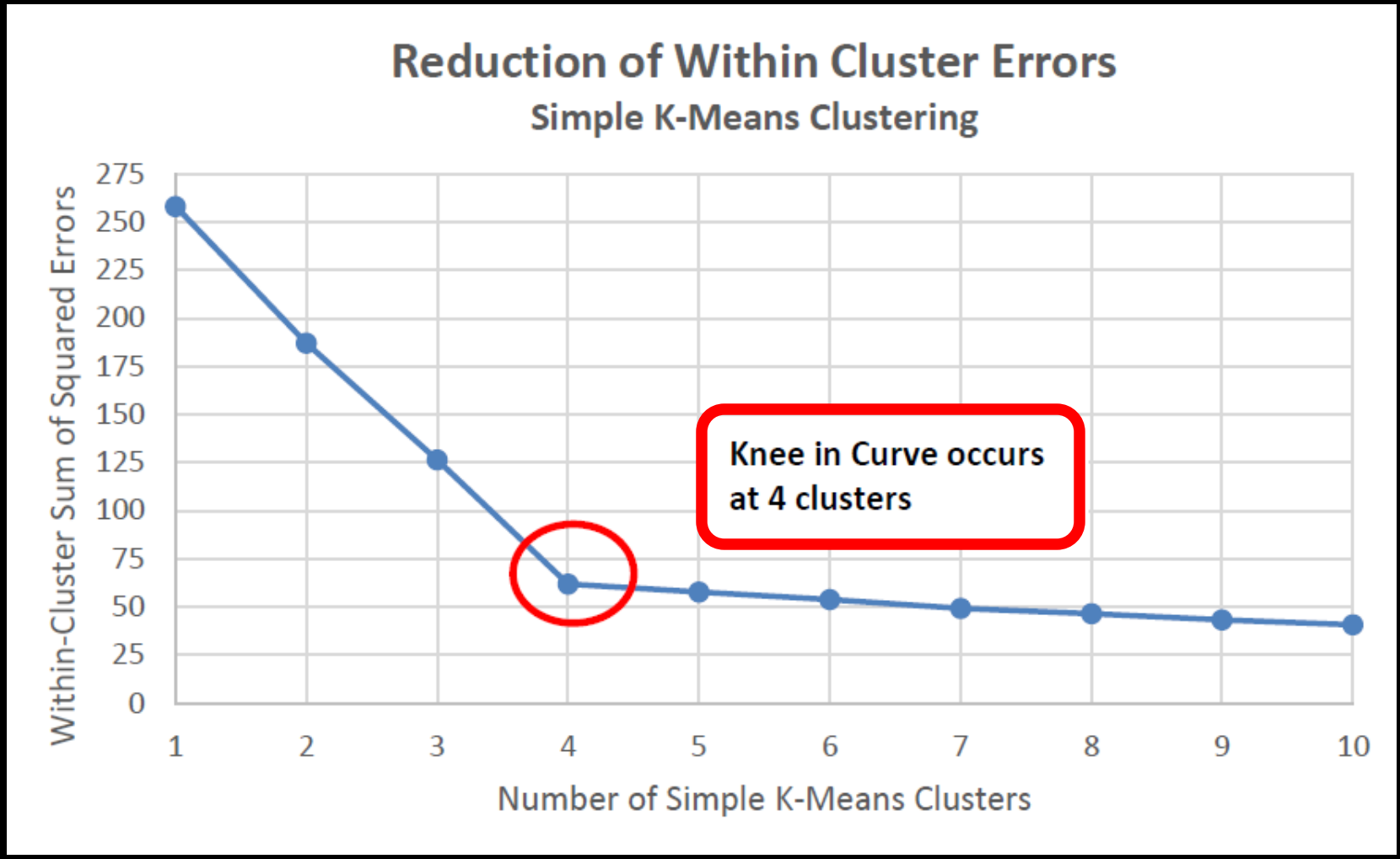
The true 3-D data distribution = 4 separable groups!



http://www.datasciencecentral.com/profiles/blogs/island-of-games-puzzle-problem-statement

Data Visualization Revelations

Solving the Island of Games Puzzle with a sequence of cluster models

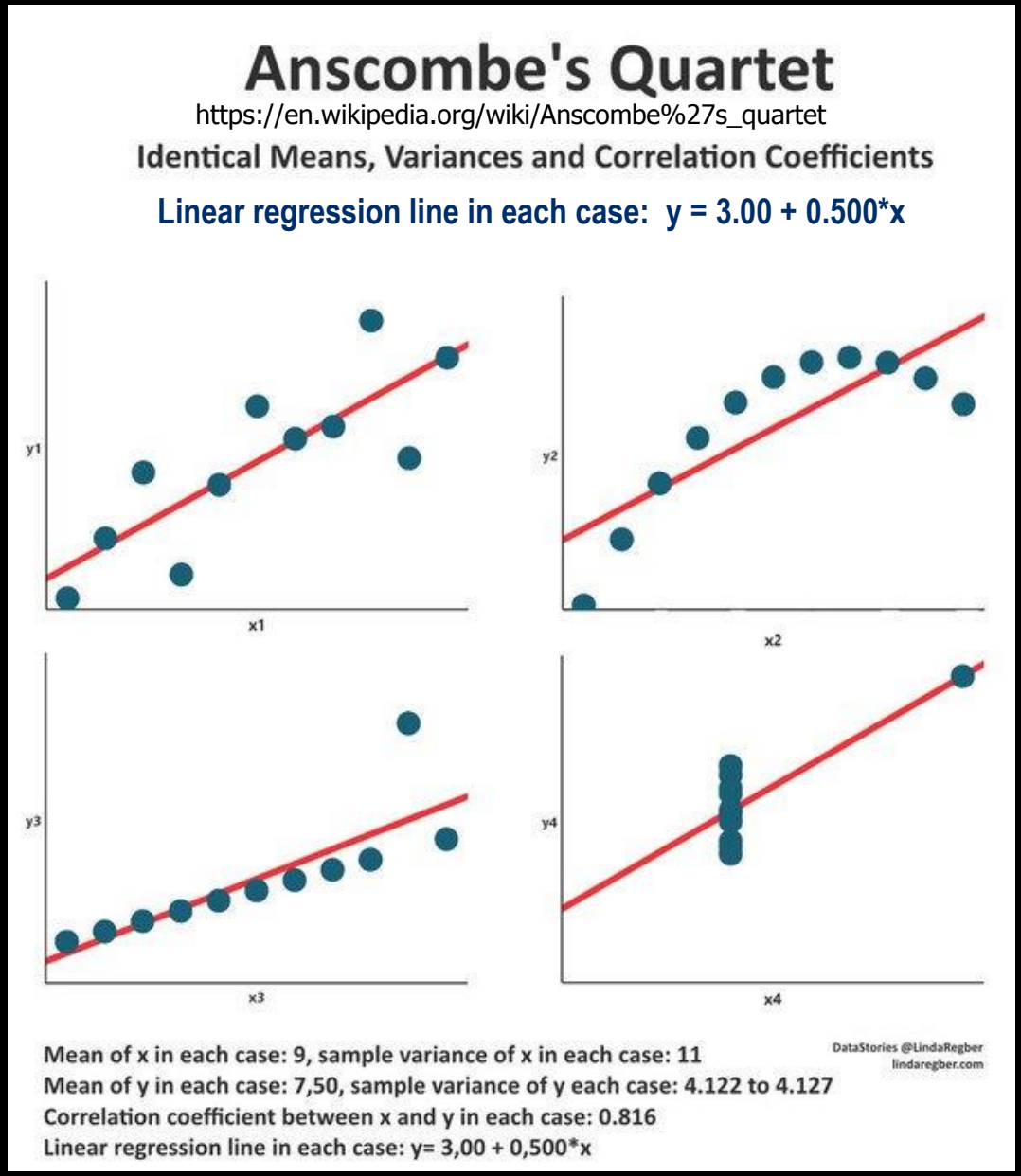


Reference: Dr. Joseph Marr, GMU

59

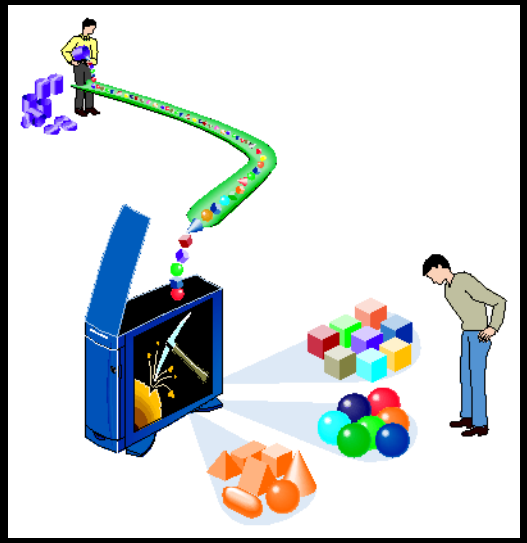
Further Data Visualization Revelations

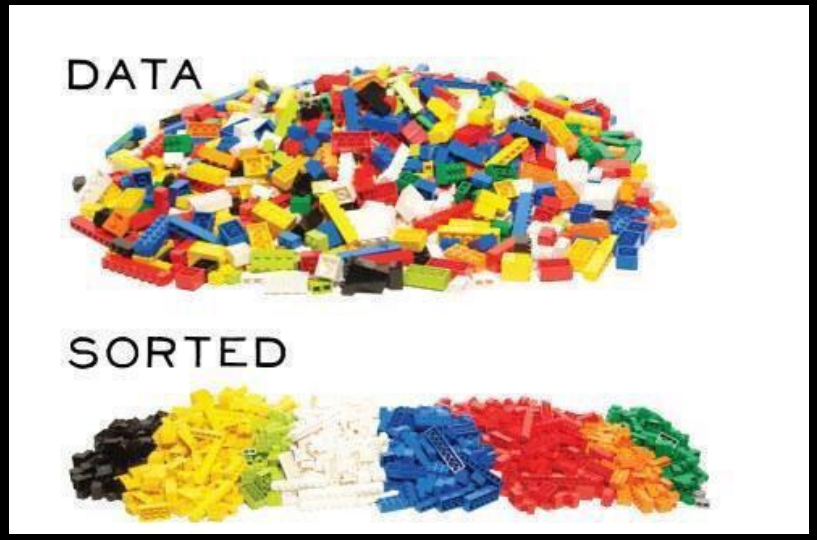
Anscombe's Quartet demonstrates classic case where statistical parameters are truly unrevealing of the different distributions of these four data sets, but the data structure is finally revealed through data visualization!



Start Young with Statistical and Data Literacy!

- Incorporate data & statistical literacies in every grade, in every type of course (not just math / STEM classes).
- Teach the 4 R's: Reading, wRiting, aRithmetic, and "R"
- Humans (even little ones) naturally characterize, sort, cluster, and classify by different observable attributes:





- Sign the Global Data Literacy petition:
  http://oceansofdata.org/call-action-promote-data-literacy

61

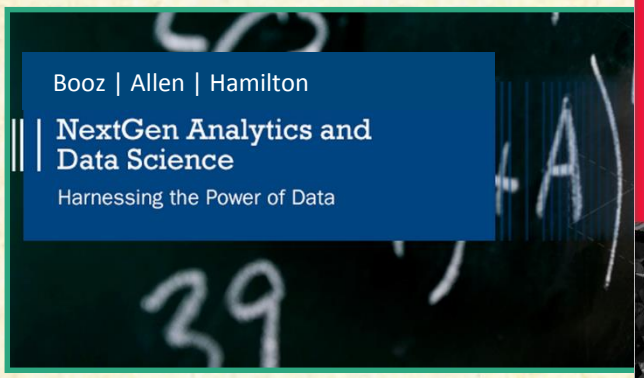
Data Literacy with Data Science will conquer Big Data's big challenges:

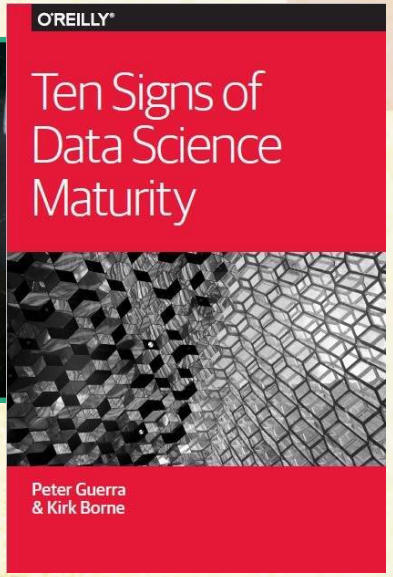
a) **V**olume
b) **V**ariety
c) **V**olume

The 3 new V's of Big Data will become:

1) **Veni** (I came)
2) **Vidi** (I saw)
3) **Vici** (I conquered)

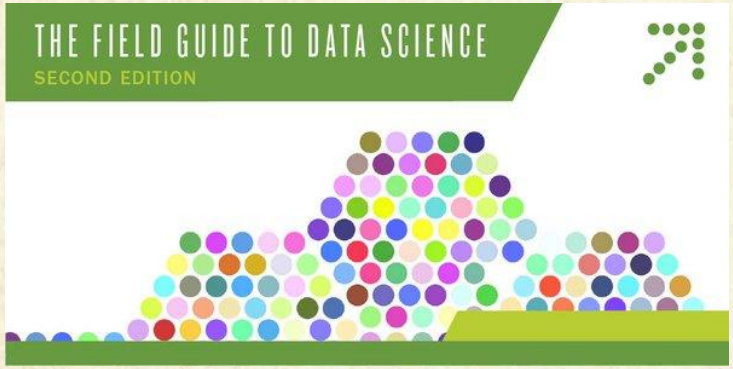
Let's conquer the world's problems together with Data & Statistical Literacy!

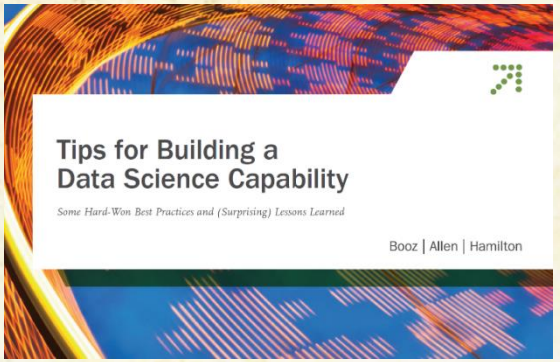




LISTEN

@KirkDBorne
@DataSci4Good
@BoozAllen





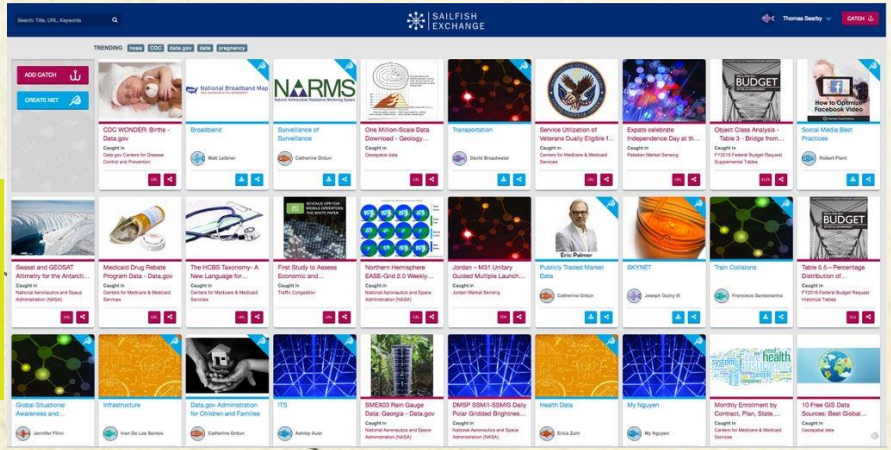
READ

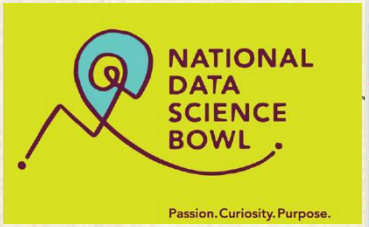
www.boozallen.com/datascience

- ❑ The Field Guide to Data Science
- ❑ Building a Data Science Capability
- ❑ 10 Signs of Data Science Maturity

PARTICIPATE

datasciencebowl.com
sailfish.boozallen.com
careers.boozallen.com





Booz | Allen | Hamilton

63