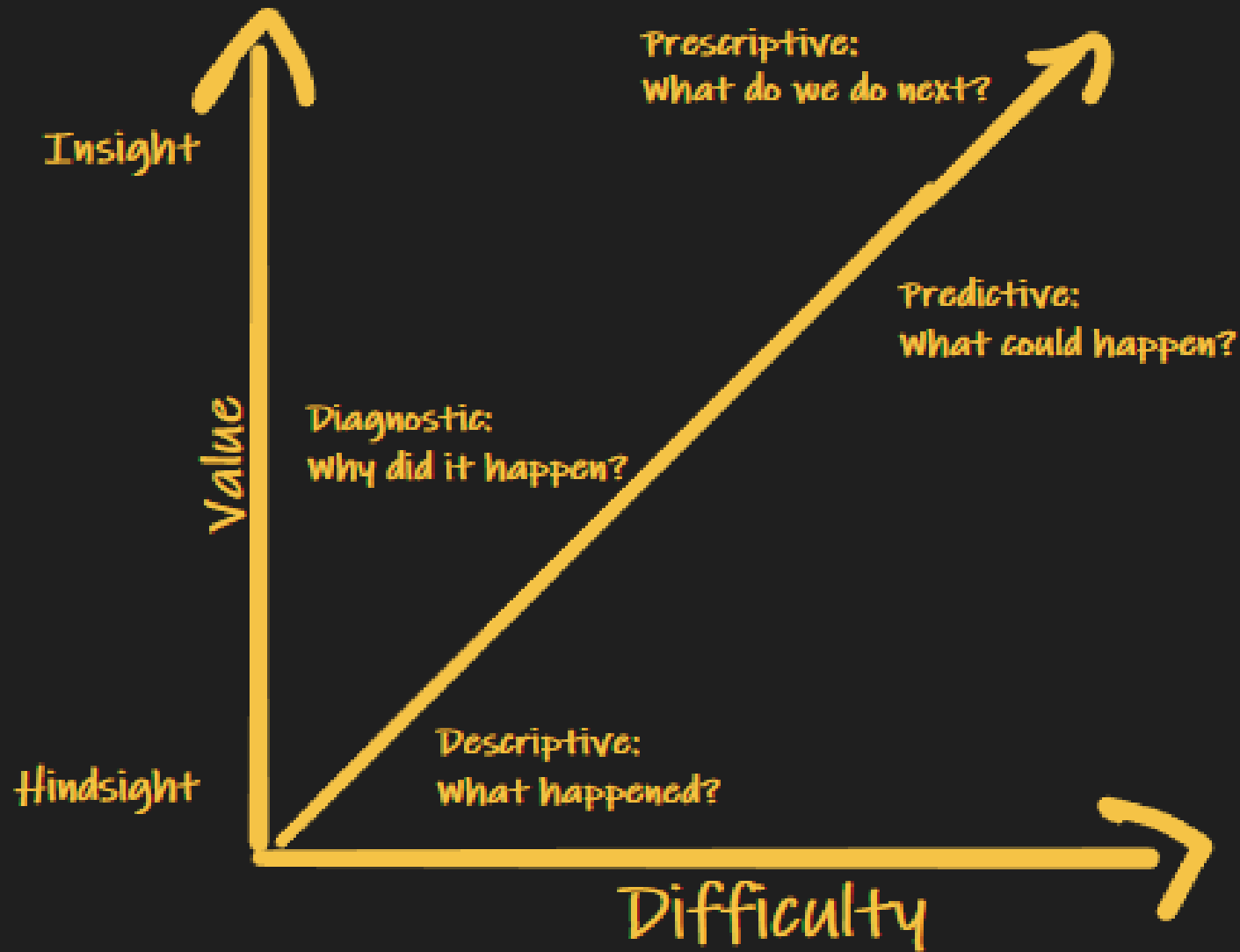


# Why is Data Literacy Important?

- Understanding technology is less important than understanding data
  - Pick the right tool for the user and use case
- Self-service analytics (aka “data democratization”, “data literacy”, or “data fluency”) initiatives at most companies are “underwhelming”
- Your users’ level of data literacy (the ability to find, work with, analyze, and “discuss” data is critical to building a self-service, *insights-driven* culture

*It is my ambition to help you better integrate business analytics into the decision-making process, and brandish it for competitive advantage.*

# Analytics Maturity Models



# How The Industry Has Traditionally Done Data Projects...

# Business Problem

“We spend \$10M every month on Facebook ad campaigns? Should we spend more or less”

Process:

- “If I could provide you with an answer, what would the UX look like”
- What is the current ROI?
- What data will I need?
- How do I attribute sales to a given campaign?

---

# How to draw an Owl.

*"A fun and creative guide for beginners"*

---

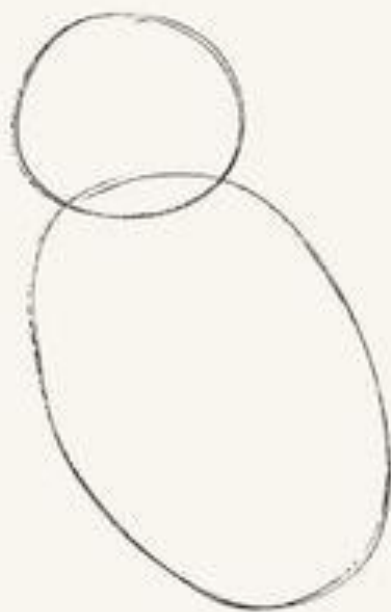


Fig 1. Draw two circles

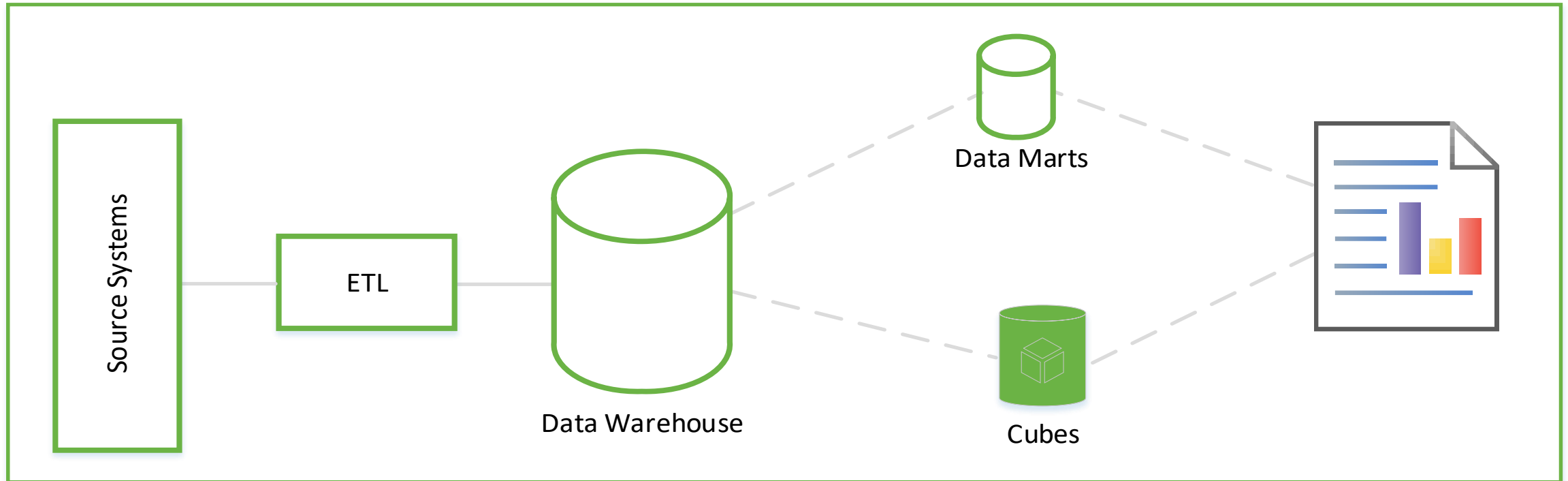


Fig 2. Draw the rest of the damn Owl

---

# Path Dependency Thinking (how we did it in the past...)

**The Philosophy:** Model data » Transform data » Load data » *Understand* data



# Data Projects have a high fail rate

---

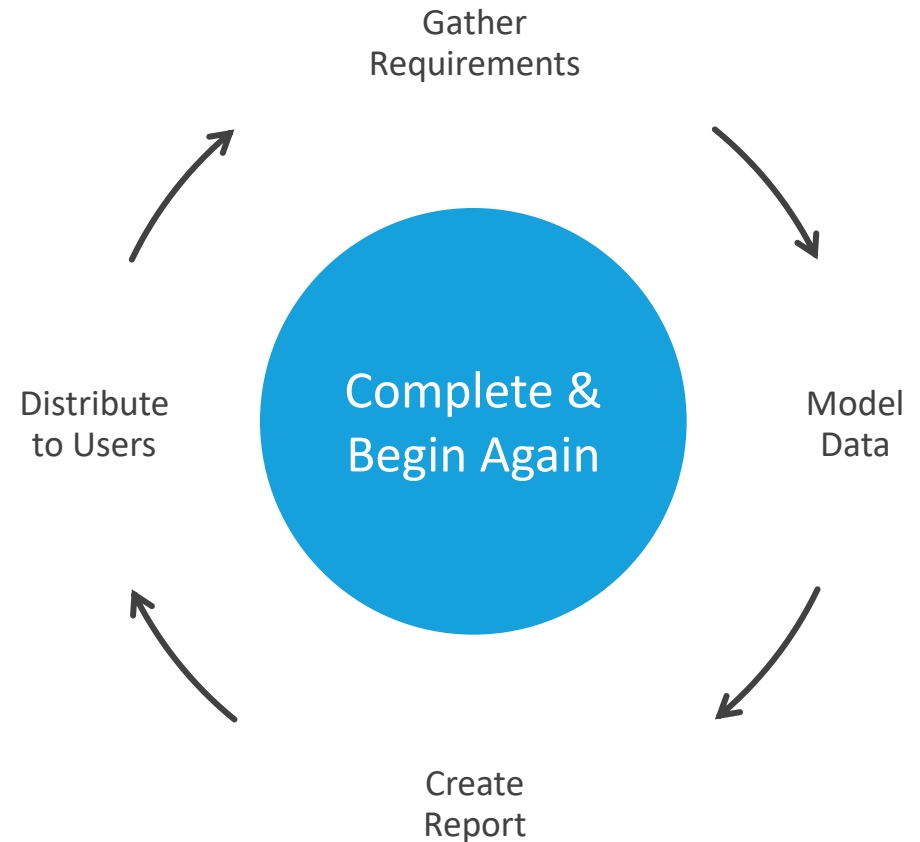
Too much time is spent in:

- Requirements gathering
- Data modeling
- ETL

Users only see the fruits of the endeavor after the reports are created

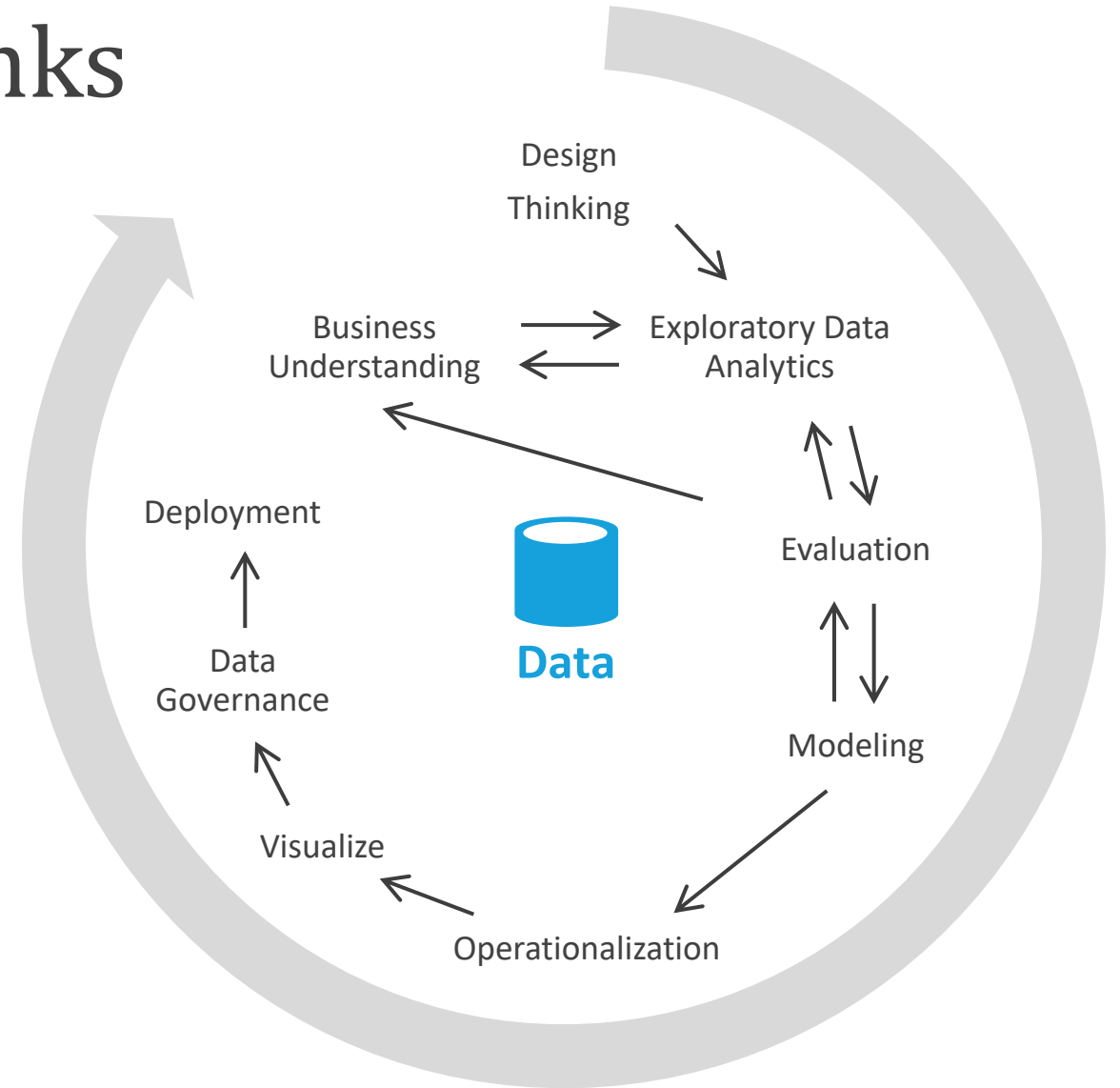
In 2014, the Project Management Institute (PMI) released its Pulse of the Profession report. PMI found that “37 percent of all organizations reported inaccurate requirements as the primary reason for project failure.”

<https://www.pmi.org/learning/library/poor-requirements-management-source-failed-projects-9341>



# How a Data Scientist Thinks

- A robust and well-proven methodology.
- Data science-like.
- Iterative.
- Stresses up-front understanding of data.
- Modeling is done later in the process (schema-on-read).
- ETL might not be needed





science

question



research



hypothesis

test



analyze



report



# AI, Machine Learning and Deep Learning

Artificial  
Intelligence



1950

1960

1970

1980

# Understanding the “Shape” of Data

## Thinking Like a Data Scientist

Will this loan be charged off?

	A	B	C	D	E	F	G	H	I	J	K
1	Loan ID	Customer ID	Loan Status	Current Loan	Term	Credit Score	Years in current job	Home Ownership	Annual Incon	Purpose	Monthly Debt
2	000025bb-	5ebc8bb1-5eb9	Fully Paid	11520	Short Term	741	10+ years	Home Mortgage	33694	Debt Cons	\$584.03
3	00002c49-	927b388d-2e01	Fully Paid	3441	Short Term	734	4 years	Home Mortgage	42269	other	\$1,106.04
4	00002d89-	defce609-c631-	Fully Paid	21029	Short Term	747	10+ years	Home Mortgage	90126	Debt Cons	\$1,321.85
5	00005222-	070bcecb-aae7	Fully Paid	18743	Short Term	747	10+ years	Own Home	38072	Debt Cons	\$751.92
6	0000757f-	dde79588-12f0	Fully Paid	11731	Short Term	746	4 years	Rent	50025	Debt Cons	\$355.18
7	0000a149-	62ddc017-7023	Fully Paid	10208	Short Term	716	10+ years	Rent	41853	Business L	\$561.52
8	0000afa6-	e49c1a82-a0f7-	Charged Off	24613	Long Term	6640	6 years	Rent	49225	Business L	\$542.29
9	0000afa6-	e49c1a82-a0f7-	Charged Off	24613	Long Term		6 years	Rent		Business L	\$542.29
0	00011dfc-	ef6e098c-6c83-	Fully Paid	10036	Short Term		5 years	Rent		Debt Cons	\$386.36

# Terminology

**Training Data** : A set of samples (table of data)

**Testing Data**: A set of samples (training data) set aside to test your model

**Features**: Individual columns in our data set. These might be used to help make our prediction, or not.

**Factors**: aka features

**Categorical Features**: features with a known domain of values

**independent variables**: aka features

**Feature Engineering**: manipulating existing data to make it more meaningful  
very similar to ETL

**Data Wrangling/Munging**: ETL

**Data Dredging**: make the data fit the hypothesis (don't do this)

**Label**: Historical outcome or result related to a set of samples.

What you are trying to predict.

aka, “the target” or “the dependent variable”, or “response”

# But, how do I get started on my first use case?

- One way, set aside a bit of budget to test something a bit weird. Maybe the opposite of what you think is actually true?
- Don't worry if an experiment doesn't replicate. Most experiments will not replicate every time.
- You don't have to be right all the time. That's what an academic does, not a business person.
- You don't always need robust data. It isn't physics. You just want to try something you wouldn't otherwise try. If the cost of failure is low, then why not try something different?
- If you test counterintuitive things, it's much more valuable when they pay off because your competitors aren't already doing those things.

# How to put a pear inside a bottle?

If the Question  
is wrong...



... the solution  
will never be  
right

How to get a bottle with a pear inside?









# How to keep a watermelon cold?







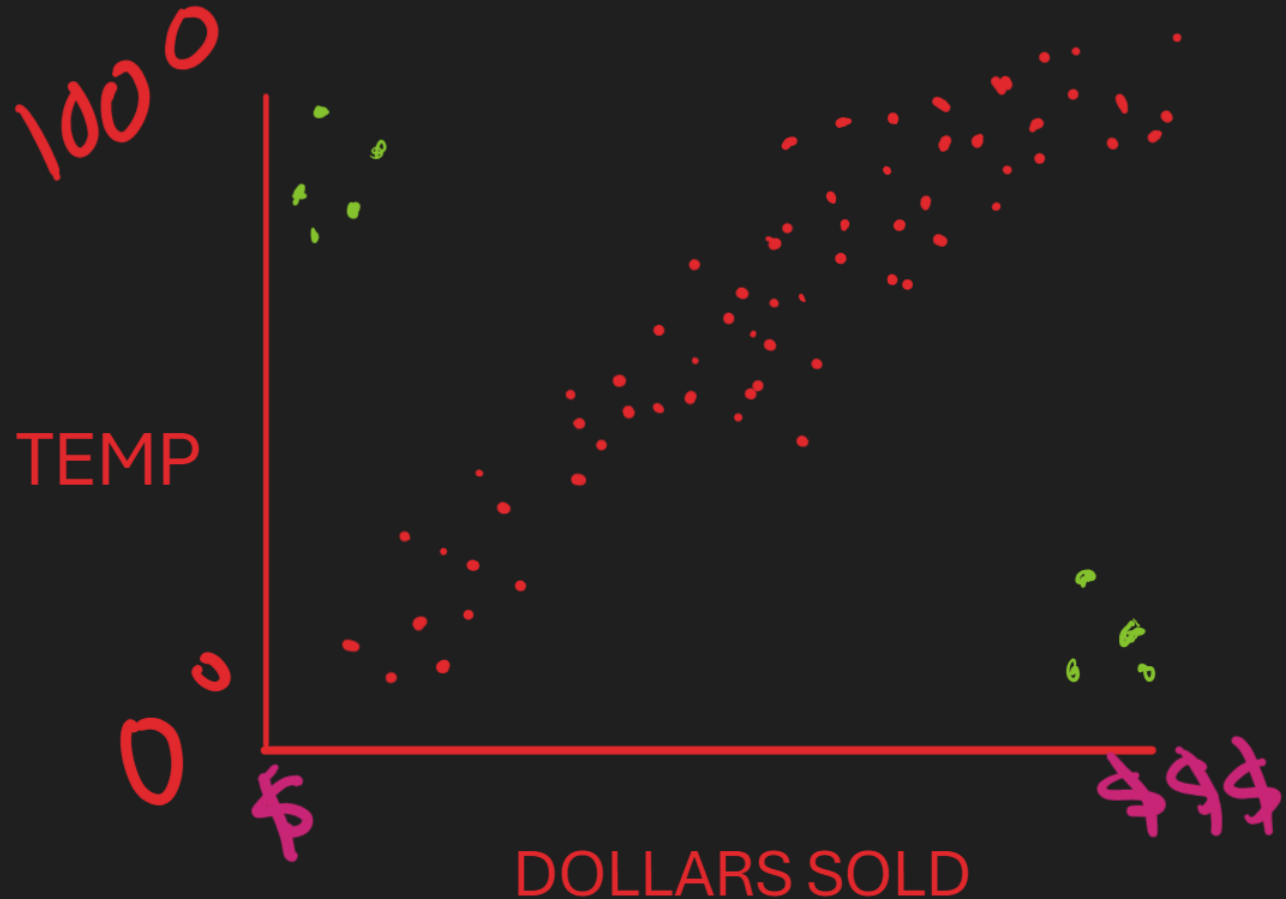


# Avoiding Cognitive Mistakes With Data

## DON'T MAKE THIS MISTAKE!

- Many data scientists will make this mistake.
- ...don't rush to build a predictive model without understanding the data.
- ...an example
- Business Problem: We are a convenience store. We believe that as temperatures rise we can sell more ice cream. Therefore, if we know what the temperature is NEXT WEEK, we can reconfigure our freezers to carry more ice cream.

each dot represents one day's sales



# Why did this fail?

- The convenience store didn't take the outliers into consideration.
- These are known as “confounders”
- In data science, a confounding variable is a variable that can affect the relationship between other variables (in this case, “temperature” and “dollars sold”), leading to distorted or unreliable results.

As I like to say, “**Humans are confounding!!**”

So, what happened? What is the “confounder”?



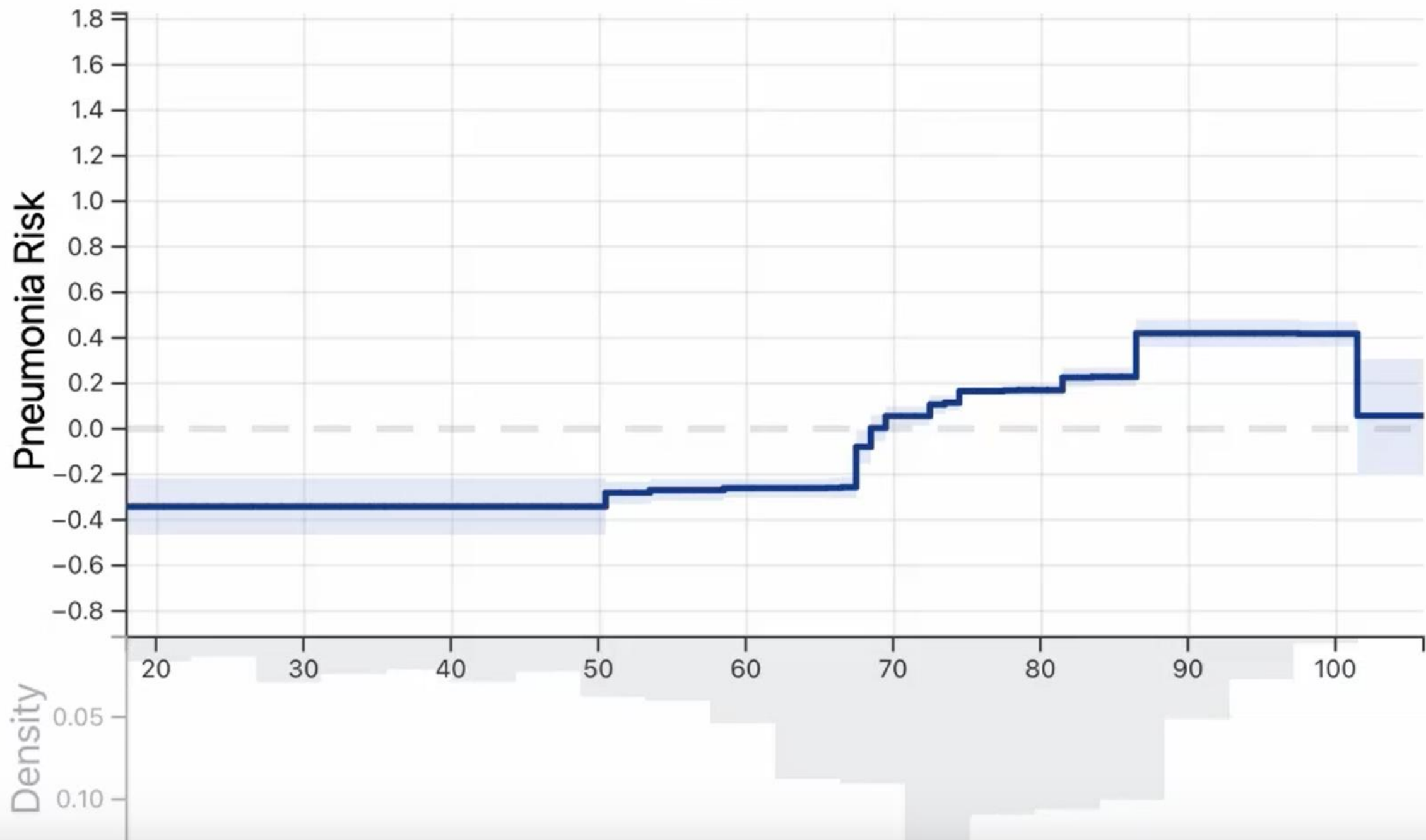
Always remember: Data Scientists understand math, statistics, building predictive models, etc.

But *rarely* do data scientists understand the business domain (ie, the problem).

That's where YOU can likely do it better.

Here's an example:

# Age





Geoffrey Hinton

“I think that if you work as a radiologist, you are like Wile E. Coyote in the cartoon. You’re already over the edge of the cliff, but you haven’t yet looked down. There’s no ground underneath. People should stop training radiologists now. It’s just completely obvious that in five years deep learning is going to do better than radiologists.”

Nov 24, 2016

@硅谷陈源-海外杂谈



# Model Evaluation: Accuracy may not be the most important thing





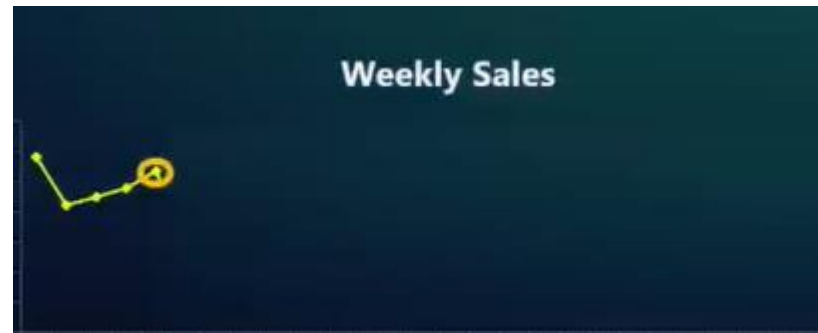
“We have a problem with customer loyalty”

“Our best customers (top 10%) in 2023 bought 30% less in 2022”

This might be a “regression to the mean” problem

“Sales have increased for the past 4 weeks. We’re on an upswing!!”

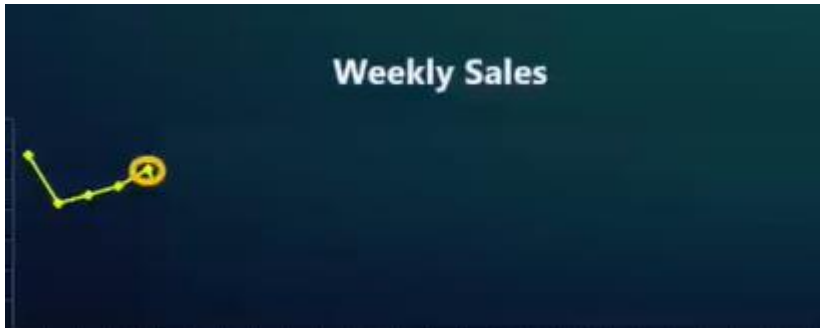
Is this a valid conclusion?



“Sales have increased for the past 4 weeks. We’re on an upswing!!”

Is this a valid conclusion?

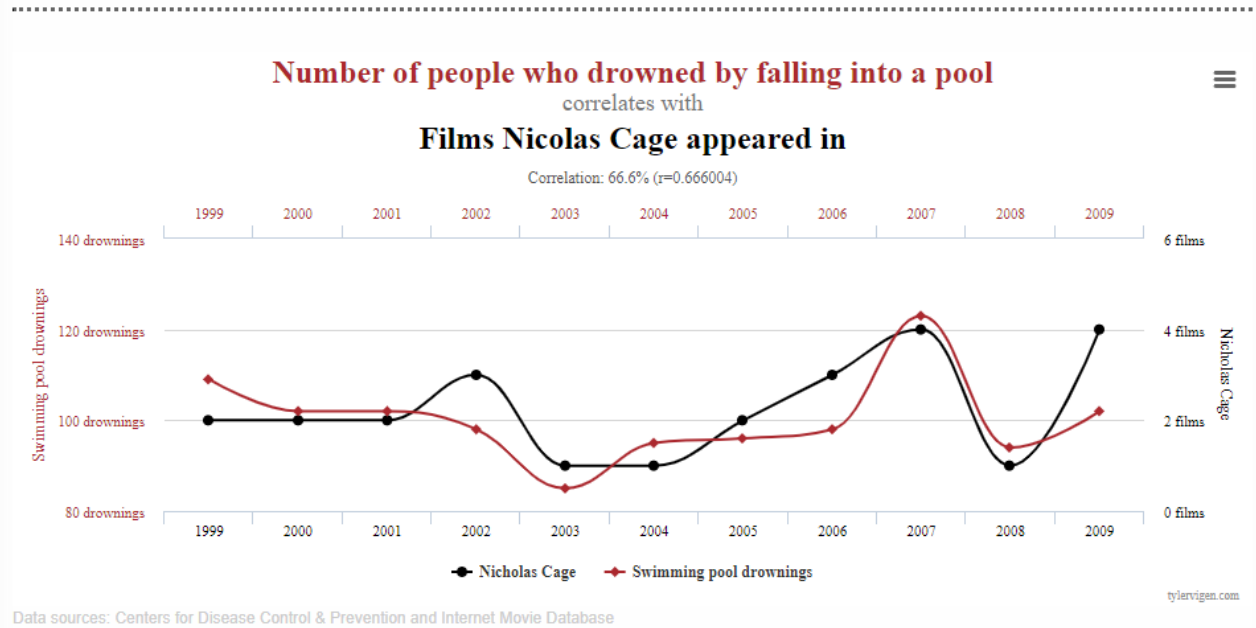
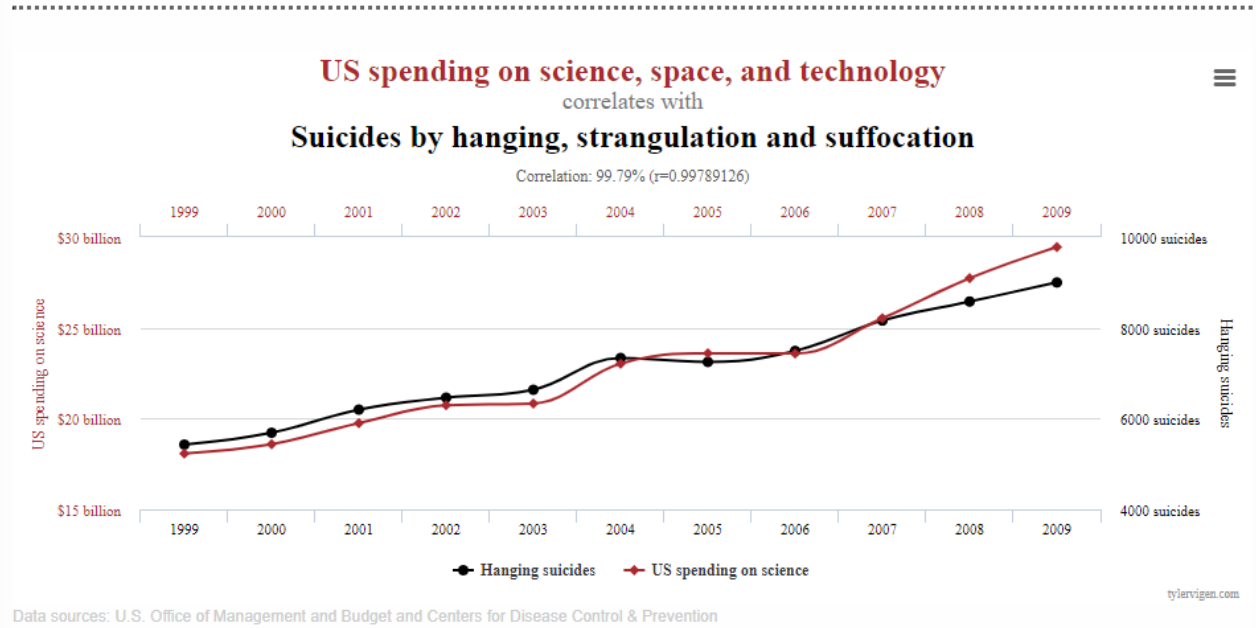
this is an issue of “spotting trends too early”



# Learning Something That Isn't True

- “Superstitious learning occurs when the connection between the cause of an action and the outcomes experienced aren't clear or misattributed.”
- This can be due to:
  - regression to the mean issues (Issue 1)
  - trends that are really random (Issue 2)
  - causation inferred from correlation
  - faulty case studies (“one-off occurrences”)





# A statistically significant correlation between two variables may be due to:

- chance
  - the usual statistical significance burden of proof is 5%. If there is no relationship between 2 variables then we would be concluding there IS a statistical significance 1 in 20 times.
  - If you look at relationships among 15 variables (by looking at pairs), 5 correlations will be statistically significant simply by chance.
- underlying (hidden) factor
- a true cause-effect relationship (but which causes which)



# Most observational studies tend to be wrong

In business, we rarely expect academic excellence in our studies...but we should at least be aware of our possible cognitive shortcomings.

## PROCEEDINGS OF THE ROYAL SOCIETY B

### BIOLOGICAL SCIENCES

 Restricted access

 Check for updates

 View Full Text

 View PDF


 Tools  Share

Cite this article 

Section

Research article

## You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans

Fiona Mathews , Paul J Johnson and Andrew Neil

Published: 22 April 2008 | <https://doi.org/10.1098/rspb.2008.0105>

### Abstract

Facultative adjustment of sex ratios by mothers occurs in some animals, and has been linked to resource availability. In mammals, the search for consistent patterns is

# Standard Deviation Helps Confirm Your Intuition

Your CMO presents a new ad campaign and claims it will increase sales to \$570/mo. That seems aggressive. What is the likelihood of this happening?

Year	Avg Monthly Sales	Min	Max
2023	500	480	520
2022	...	...	...
2021	...	...	...

StdDev "Shortcut" =  $(MAX-MIN)/4 = 10$

Now you can use stddev to calculate the "Z-Score".

z-score = how far you are from the mean =  $(x - \text{mean})/\text{stddev}$

Example: Assume Jan 2024 = 510. z-score =  $(510-500)/10 = 1$  stddev

What is the CMO forecast?  $(570-500)/10 = 7$  stddev

Generally accepted principle: zscore > 3 is an outlier and not likely

# Let's change the Historical Data and Look at it Again

Your CMO presents a new ad campaign and claims it will increase sales to \$570/mo. That seems aggressive. What is the likelihood of this happening?

Year	Avg Monthly Sales	Min	Max
2018	500	<del>480</del> 400	<del>520</del> 600
2017	...	...	...
2016	...	...	...

$$\text{StdDev} = (\text{MAX}-\text{MIN})/4 = (600-400)/4 = 50$$

$$\text{z-score} = (x - \text{mean})/\text{stddev}$$

NOW what is the CMO forecast?

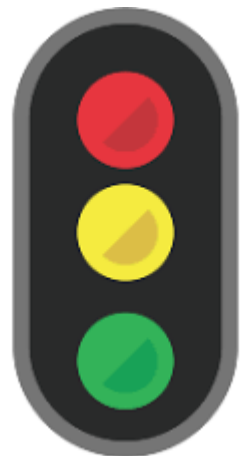
$$(570-500)/50 = 70/50 = 1.4 \text{ std dev}$$

**Z-scores**

**>3**

**1.65-3**

**< 1.65**



# How Business Leaders Can Avoid Analytics Mistakes

- Data Scientists tend to be well-versed in math, probabilities, and statistics.
- Even so, humans in general are susceptible to cognitive biases, especially when interpreting data
  - Especially “conditional probabilities” (the likelihood of an event happening)
- Trust your intuition
- Don't be fooled by randomness
- Always get feedback from other people (design thinking)
- Quantitative methods can tell you WHAT or HOW, but never WHY

**I CAN'T DO THIS. I'm scared. This is too much math!**

# Github Repo

- Presentation, examples using Jupyter notebooks, add'l material
- <https://github.com/davew-msft/datacamp-dl>
- [linkedin.com/in/dwentzel](https://www.linkedin.com/in/dwentzel)