Overview

DATA 201: Thinking With Data
Winter 2021

Jonathan Hudson, Ph.D
Instructor
Department of Computer Science
University of Calgary

Wednesday, January 6, 2021

UNIVERSITY OF CALGARY

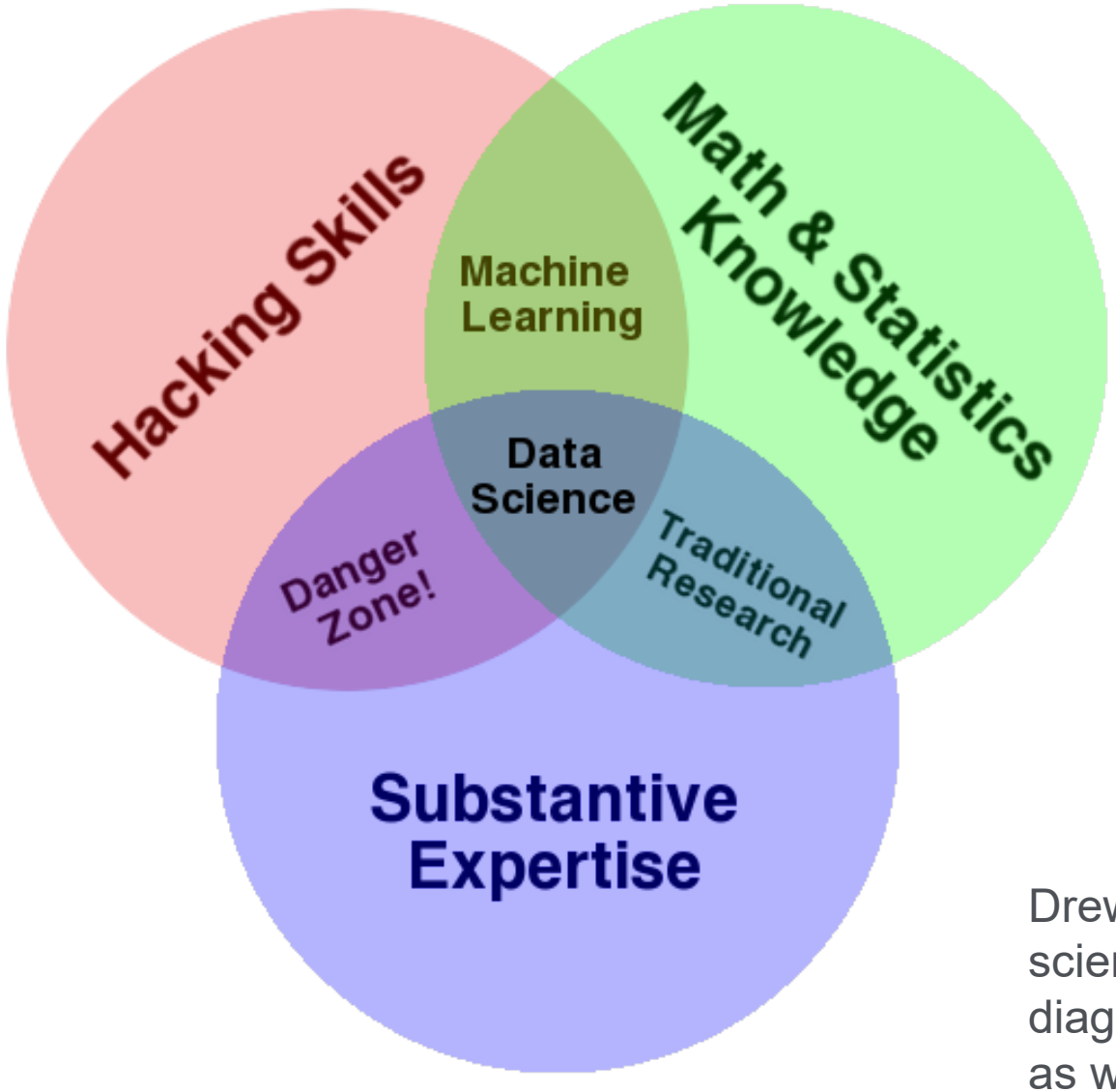
What is Data Science?

What (do you think) is Data Science?

Data science is a ‘concept to unify statistics, data analysis, machine learning and their related methods’ in order to ‘understand and analyze actual phenomena’ with data.

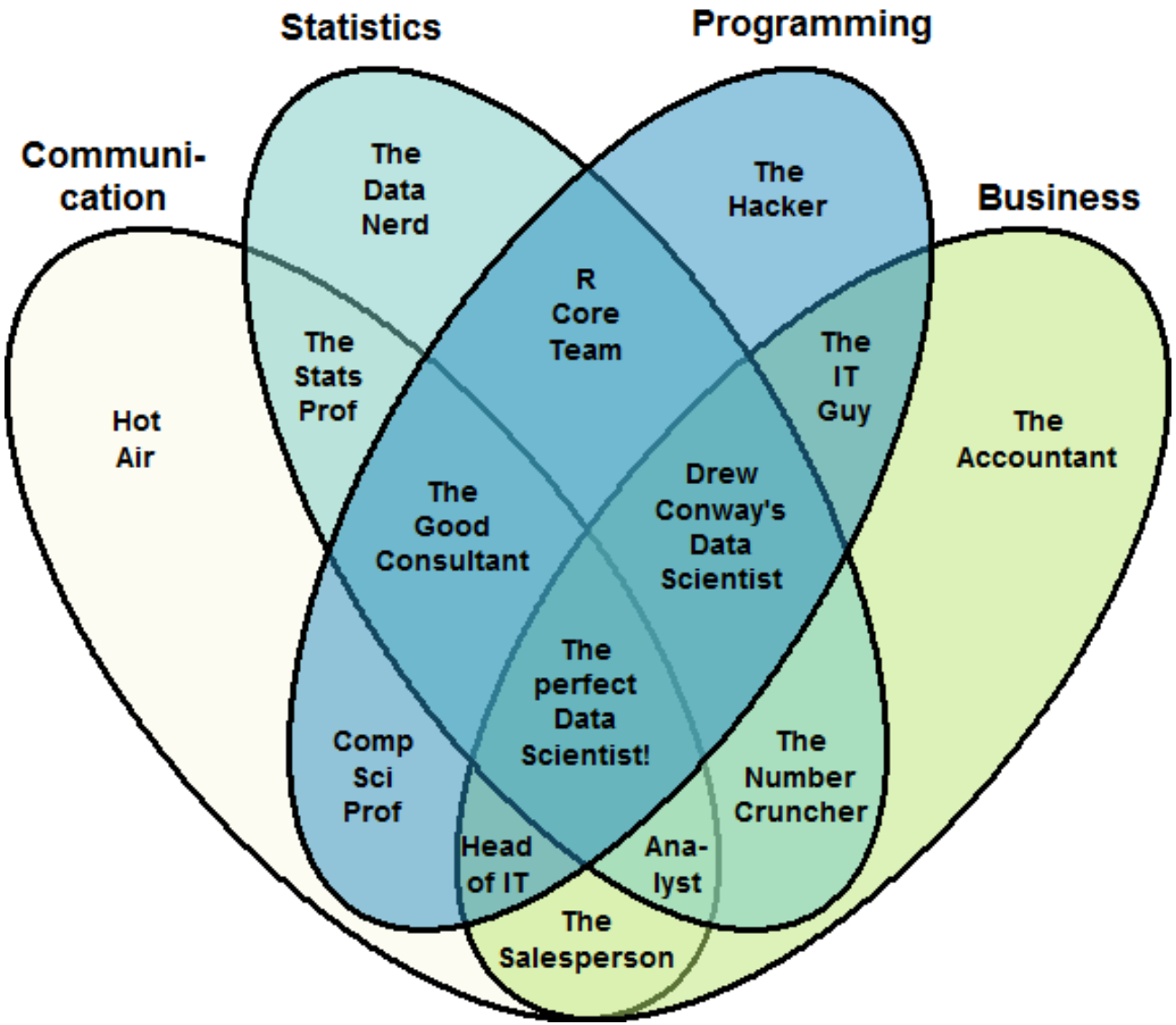
“Data science is the discipline of making data useful.”

Drew Conway

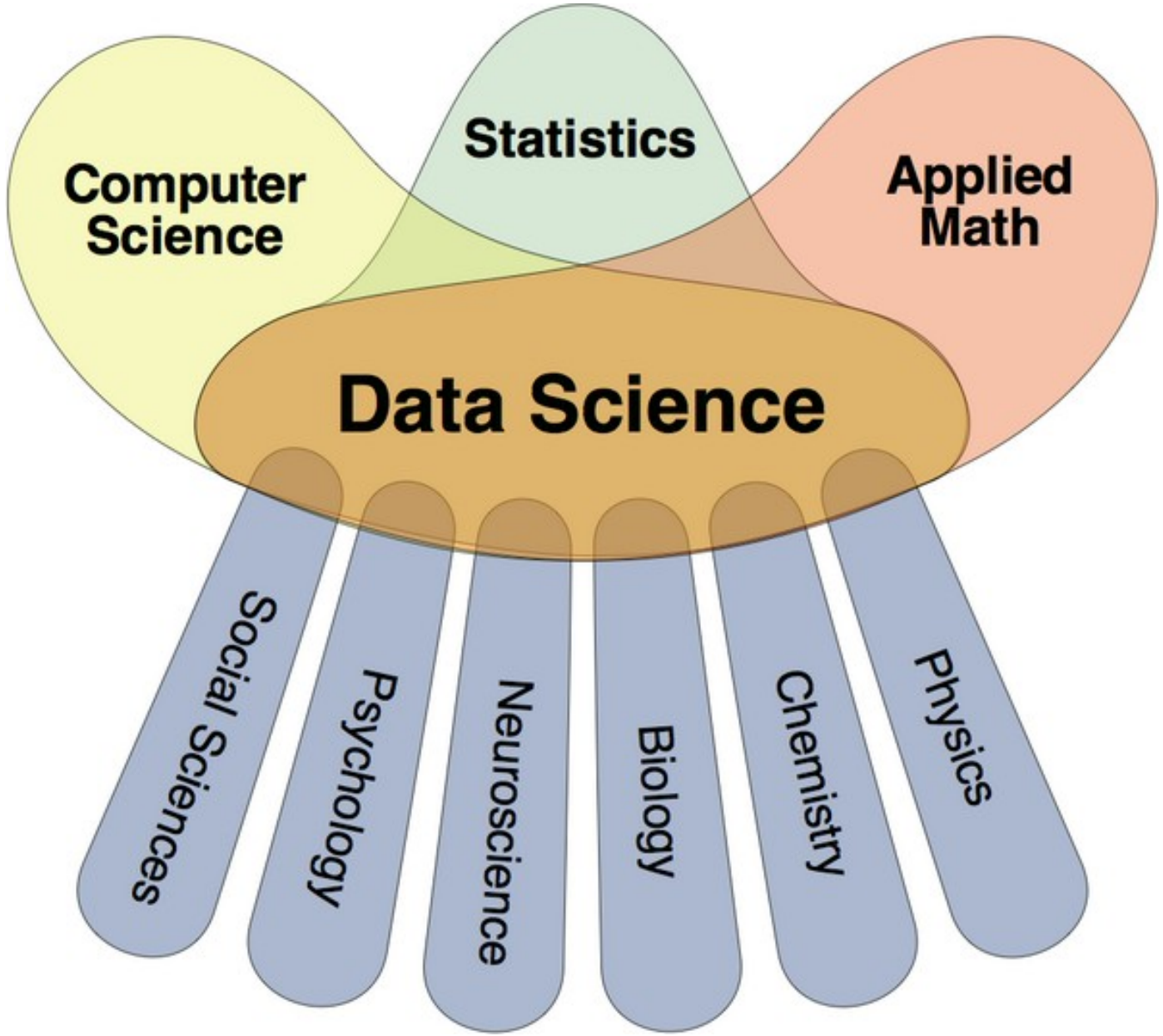


Drew Conway is an American data scientist known for his venn diagram definition of data science as well as applying data science to study terrorism.

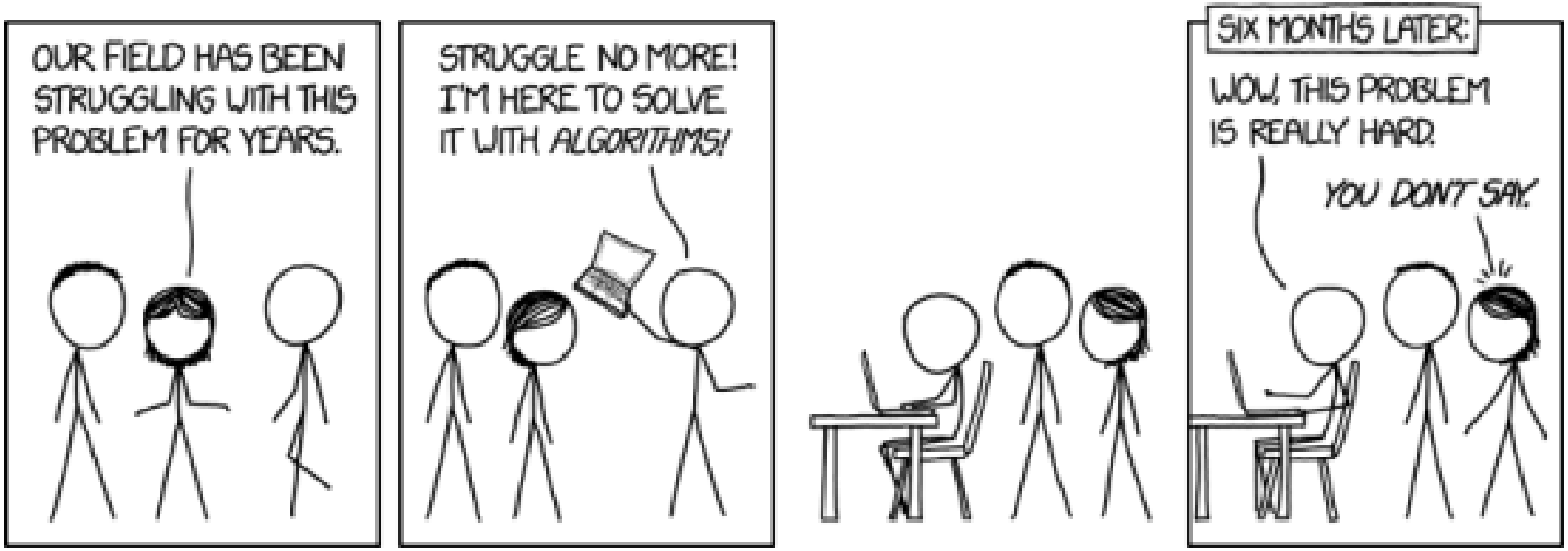
The Data Scientist Venn Diagram



UNIVERSITY OF CALGARY



Area Knowledge

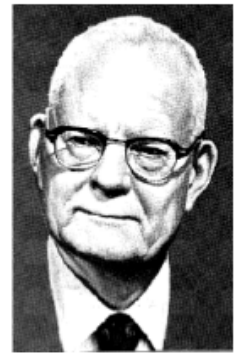


1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

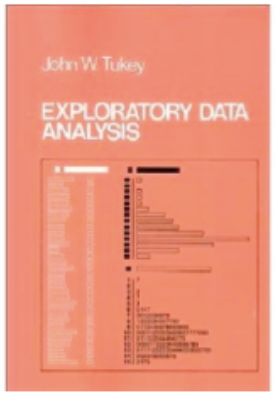
W.E. Demming



1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

Howard Dresner

1997: "Machine Learning"

2010:"The Data Deluge"

1996: Google

Google™

2007:"The Fourth Paradigm"

2009:"The Unreasonable Effectiveness of Data"

Abridged Version of Jeff Hammerbacher's timeline for Berkeley CS 194, 2012

Short History – How We Got Here

- The Design of Experiments is a 1935 book by the English statistician Ronald Fisher about the design of experiments
- Deming edited a series of lectures delivered by Shewhart at USDA, Statistical Method from the Viewpoint of Quality Control, into a book published in 1939
- IBM Peter Luhn paper 1958 “An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. “
- 1977 This book serves as an introductory text for exploratory data analysis.
- In 1989 Gartner analyst Howard Dresner again brought the phrase “business intelligence” into the common vernacular. He employed it as a general term to cover the cumbersome-sounding names for data storage and data analysis.
- Tom Michell 1997 Machine Learning is the study of computer algorithms that improve automatically through experience.

Short History – How We Got Here

- The first version of Google was released in August 1996 on the Stanford website. It used nearly half of Stanford's entire network bandwidth. Algorithm to rank documents (webpages).
- 2007 Data is fourth paradigm. Outlines a two-part plea for the funding of tools for data capture, curation, and analysis, and for a communication and publication infrastructure. He argued for the establishment of modern stores for data and documents that are on par with traditional libraries.

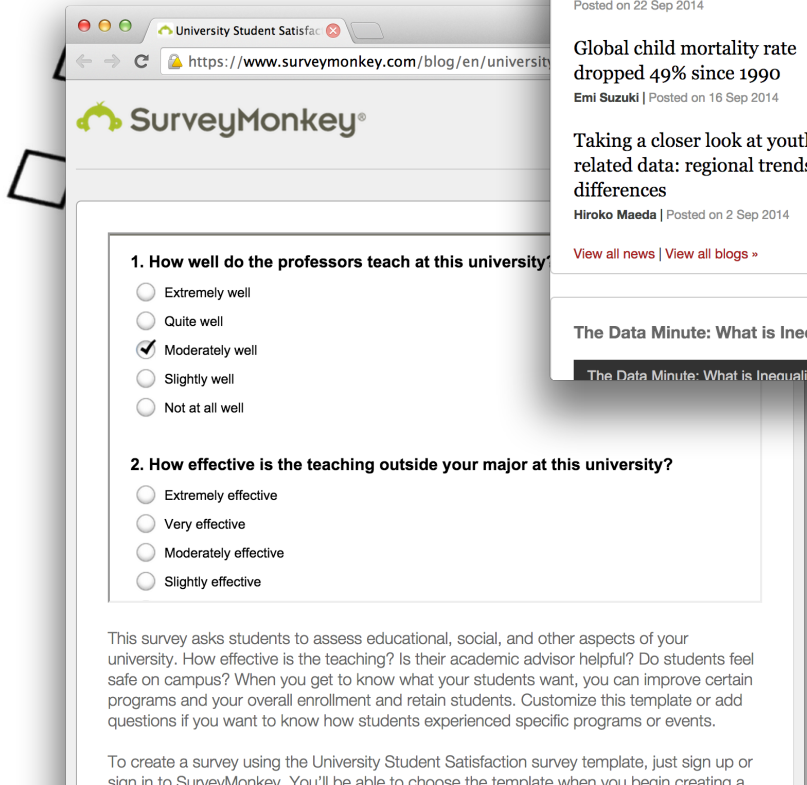
Short History – How We Got Here

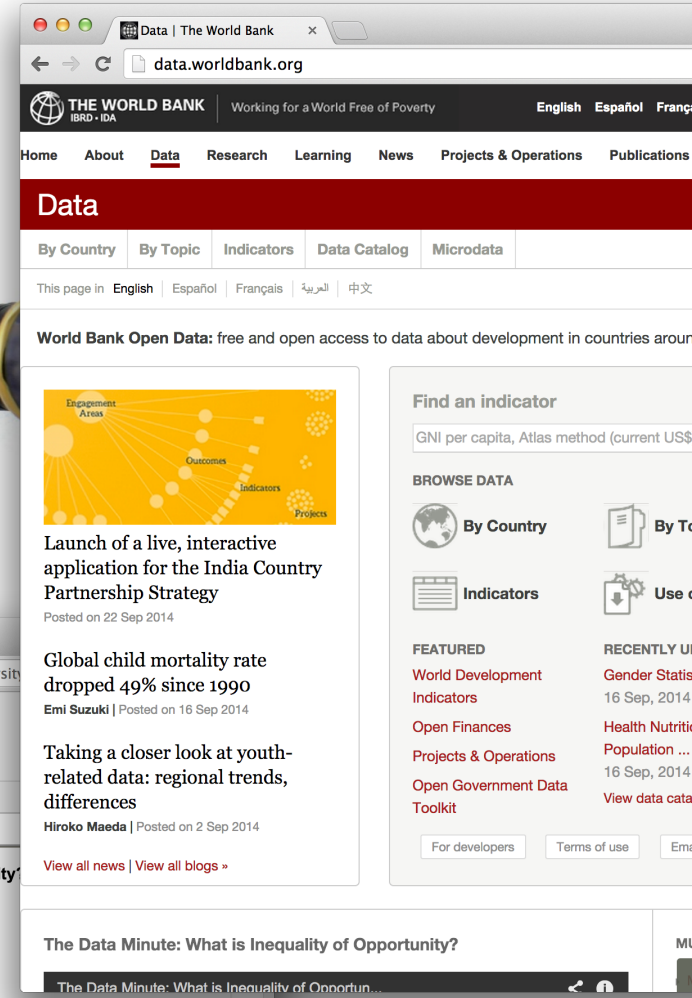
- 2009 Problems that involve interacting with humans, such as natural language understanding, have not proven to be solvable by concise, neat formulas like F = ma. Instead, the best approach appears to be to embrace the complexity of the domain and address it by harnessing the power of data.
- 2010 Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life. It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.

Why Data Science?

Data Science Pipeline - Collection

Collecting → Cleaning → Analyzing → Presenting

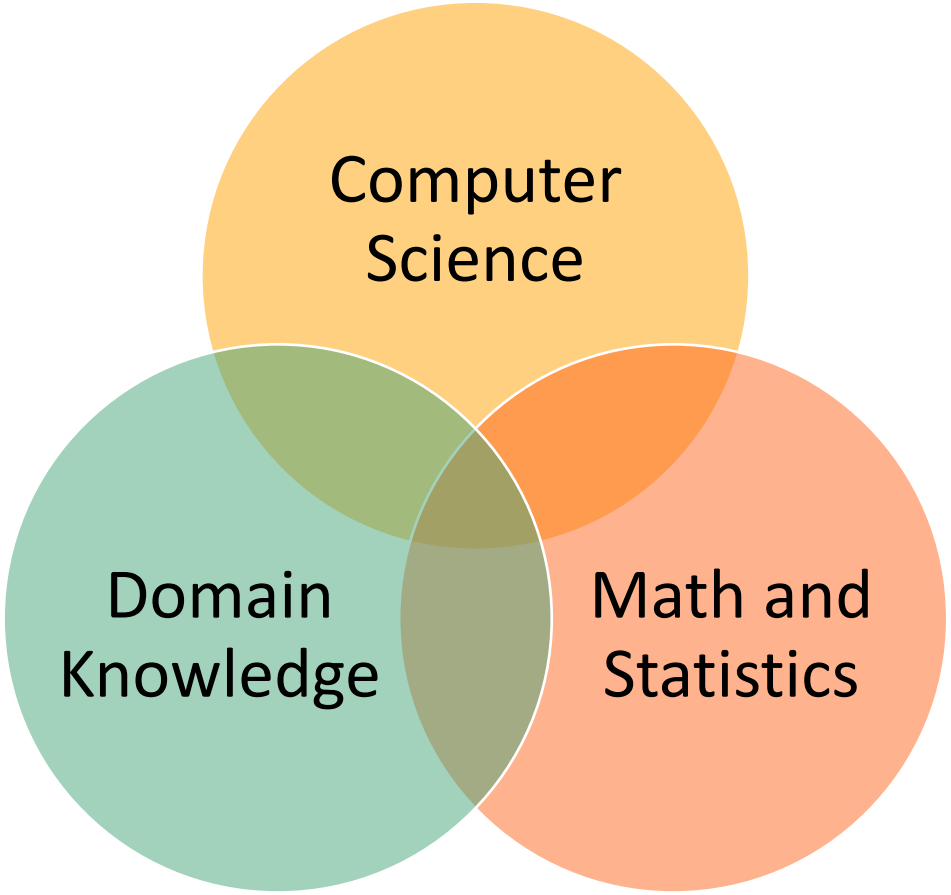


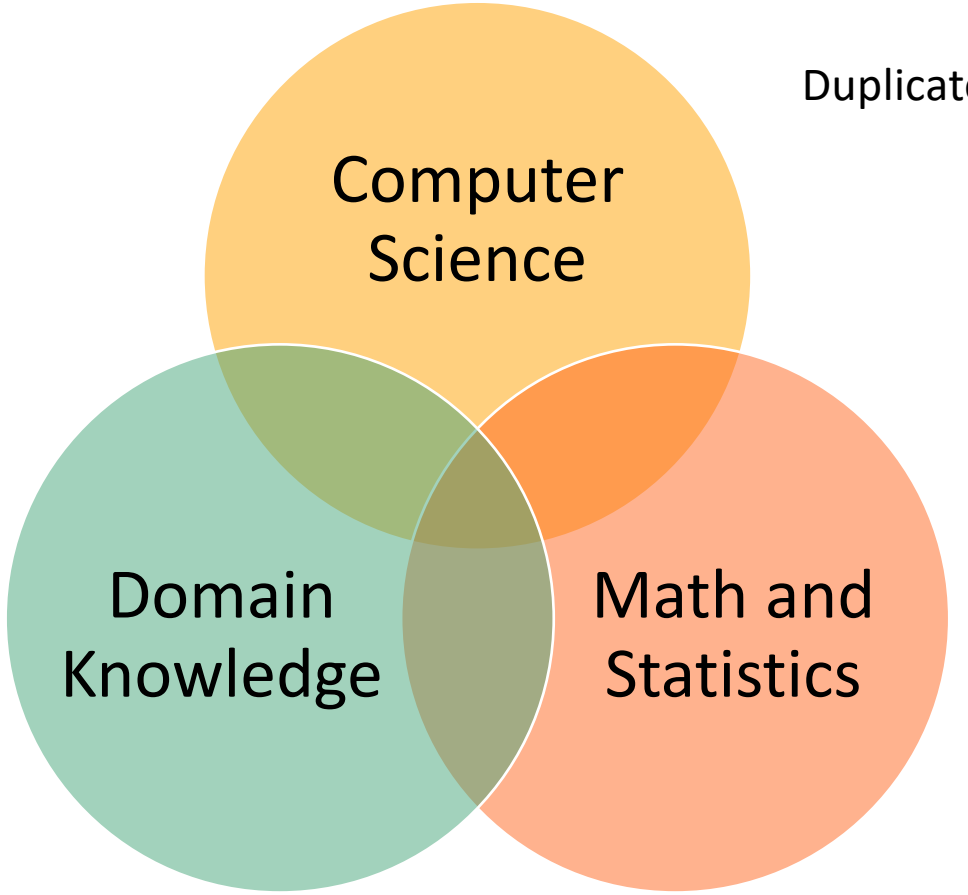


UNIVERSITY OF CALGARY

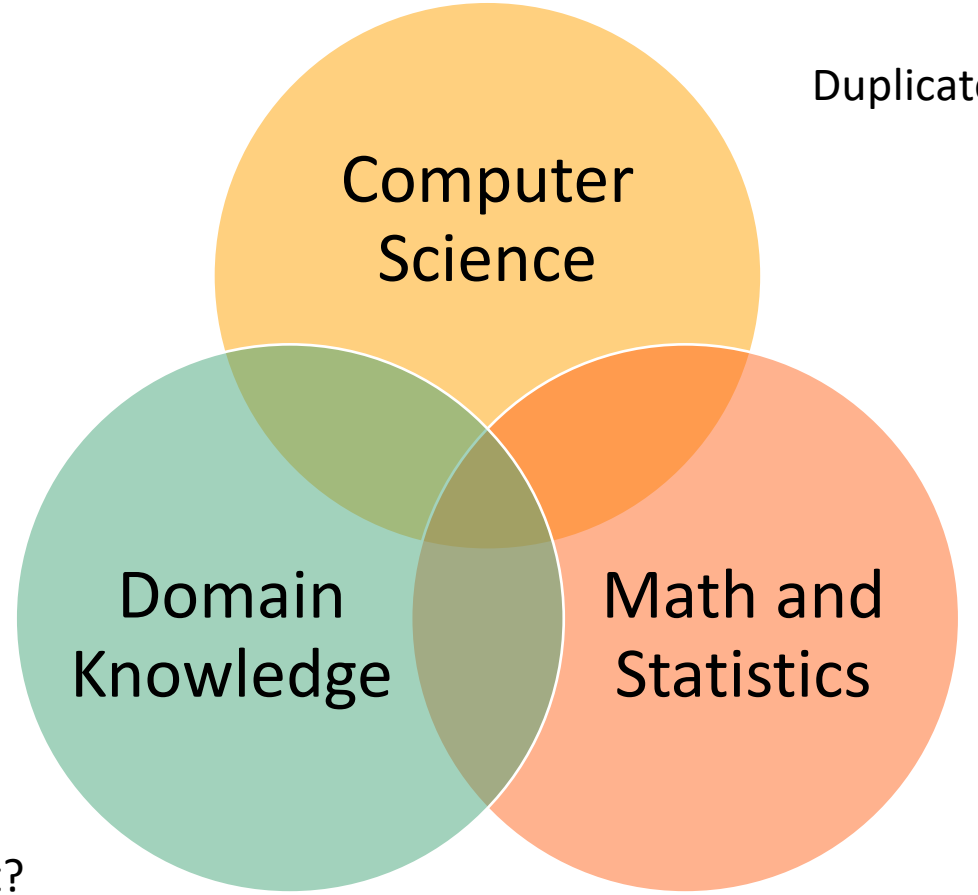
Data Science Pipeline - Cleaning

Collection → **Cleaning** → Analyzing → Presenting



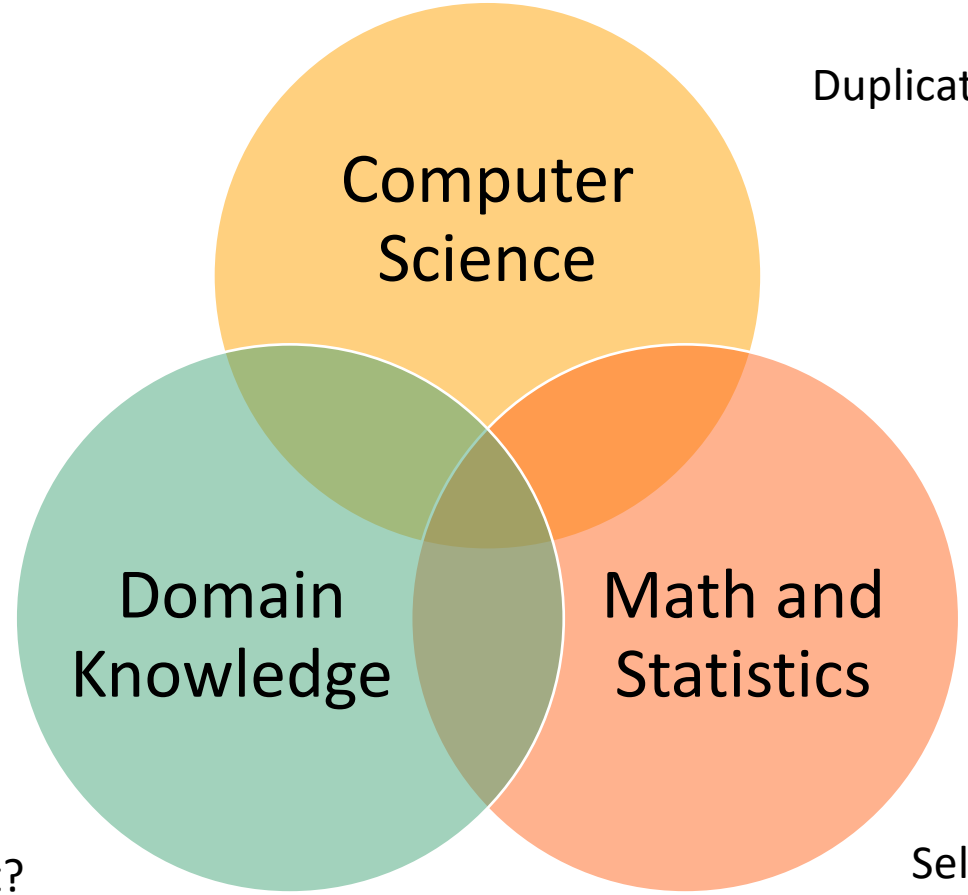


Duplicate values, incorrect formatting, etc.



Duplicate values, incorrect formatting, etc.

Data doesn’t look right?
Answer doesn’t look right?



Duplicate values, incorrect formatting, etc.

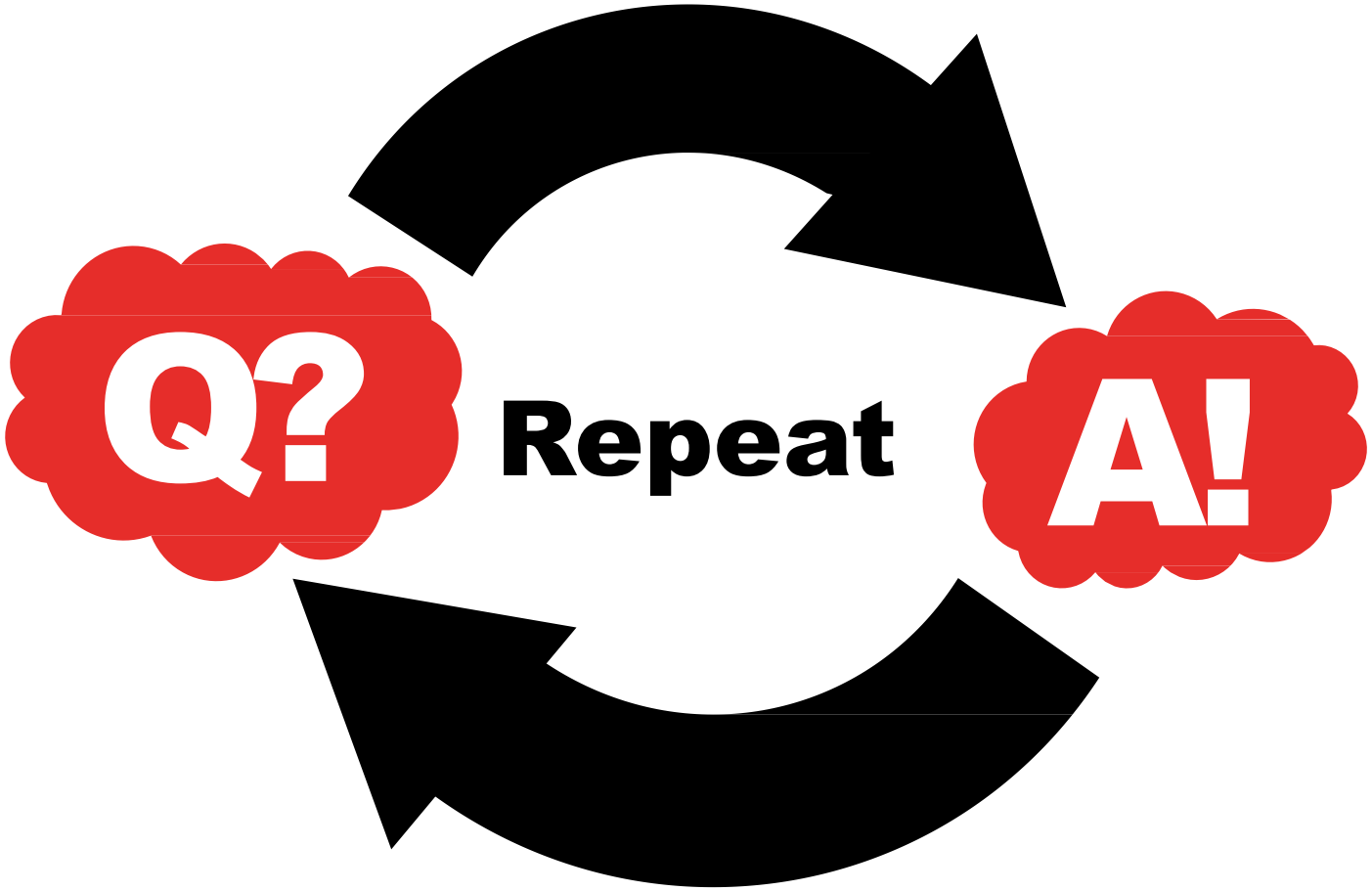
Data doesn't look right?
Answer doesn't look right?

Selection bias, dependence, etc.

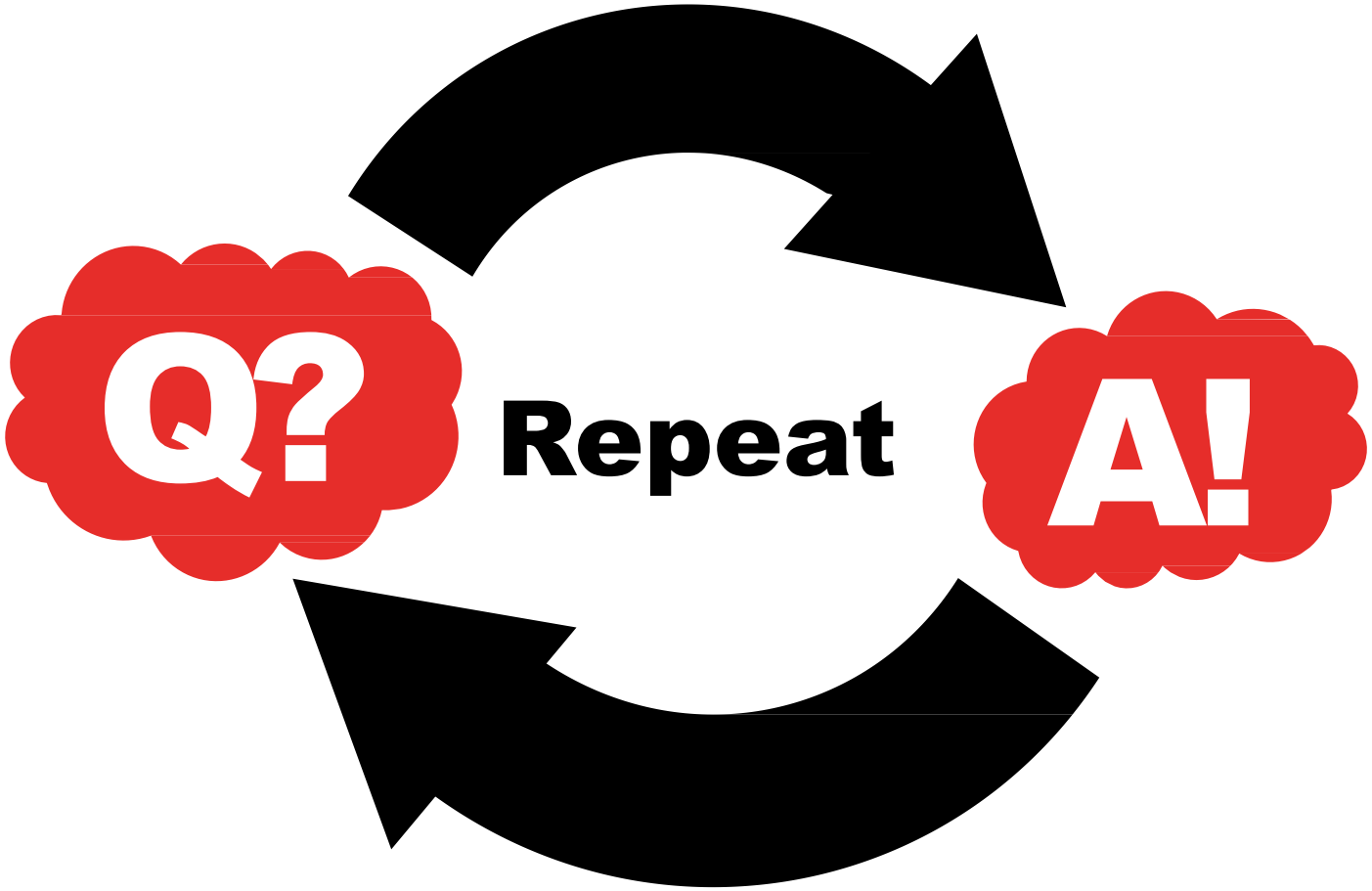
OpenRefine

Data Science Pipeline - Analysis

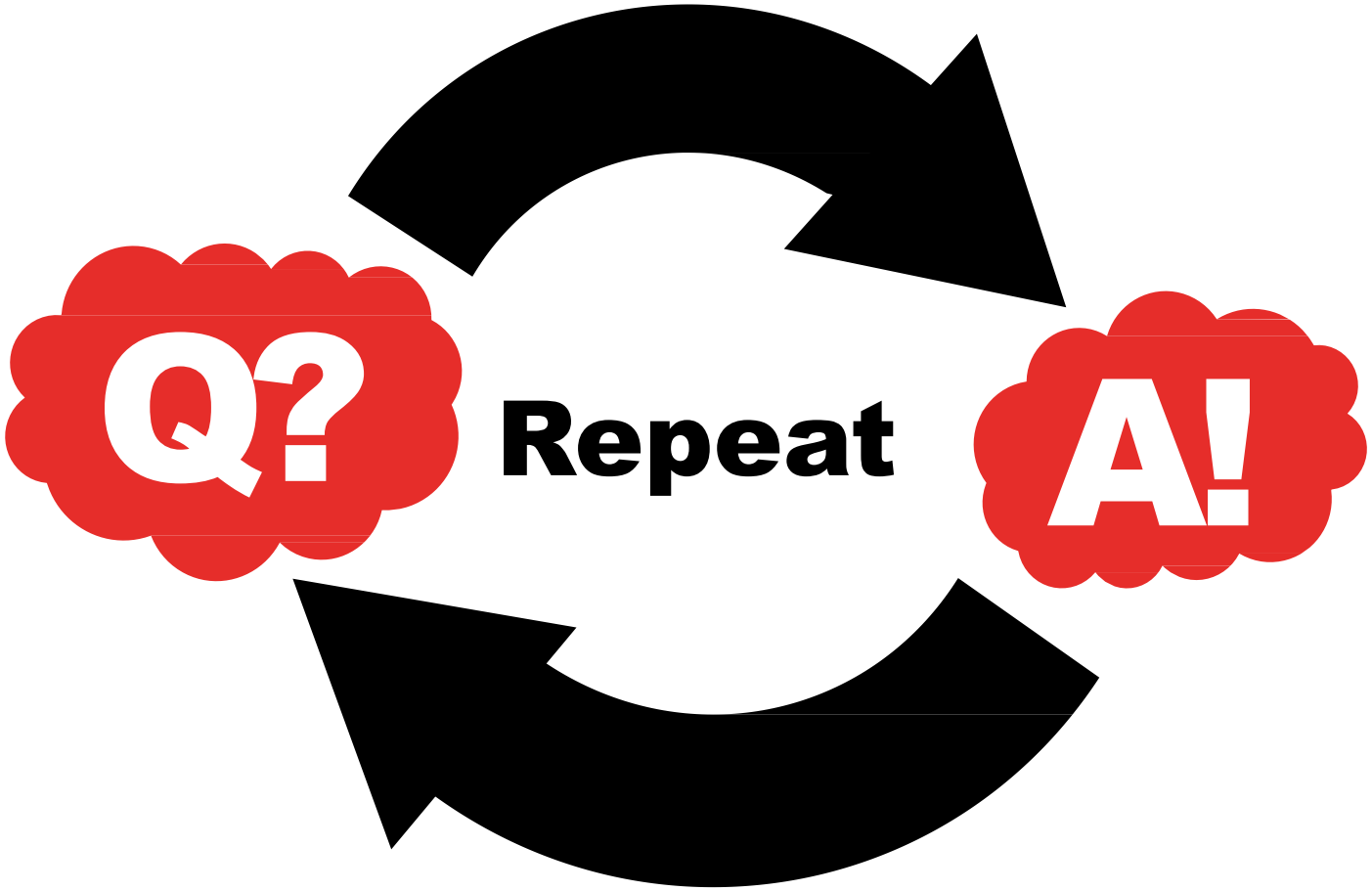
Collection → Cleaning → **Analyzing** → Presenting


Q?
Repeat
A!

gather data, clean data, apply statistical tools, visualize data to address questions



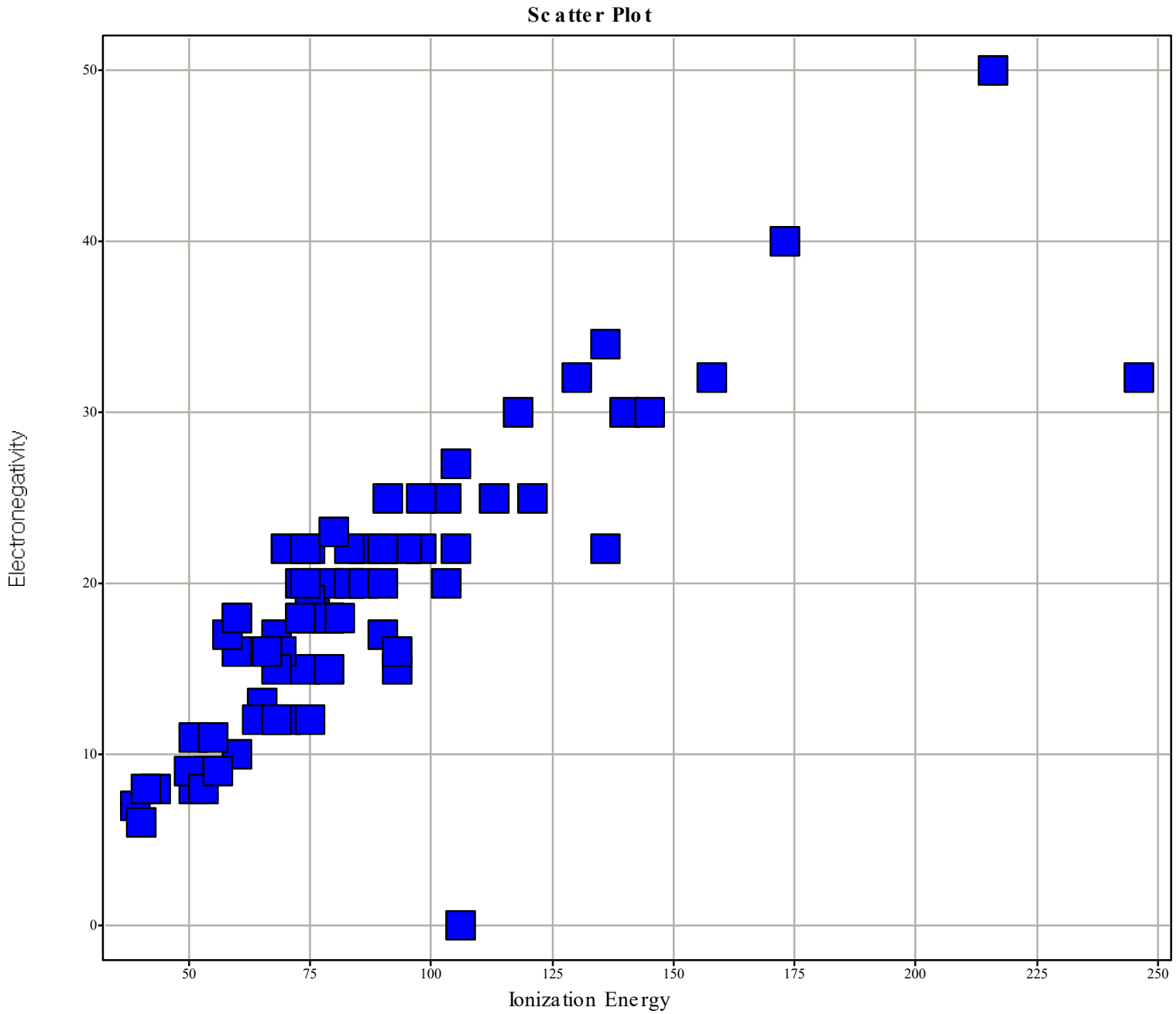
gather data, clean data, apply statistical tools, visualize data to address questions

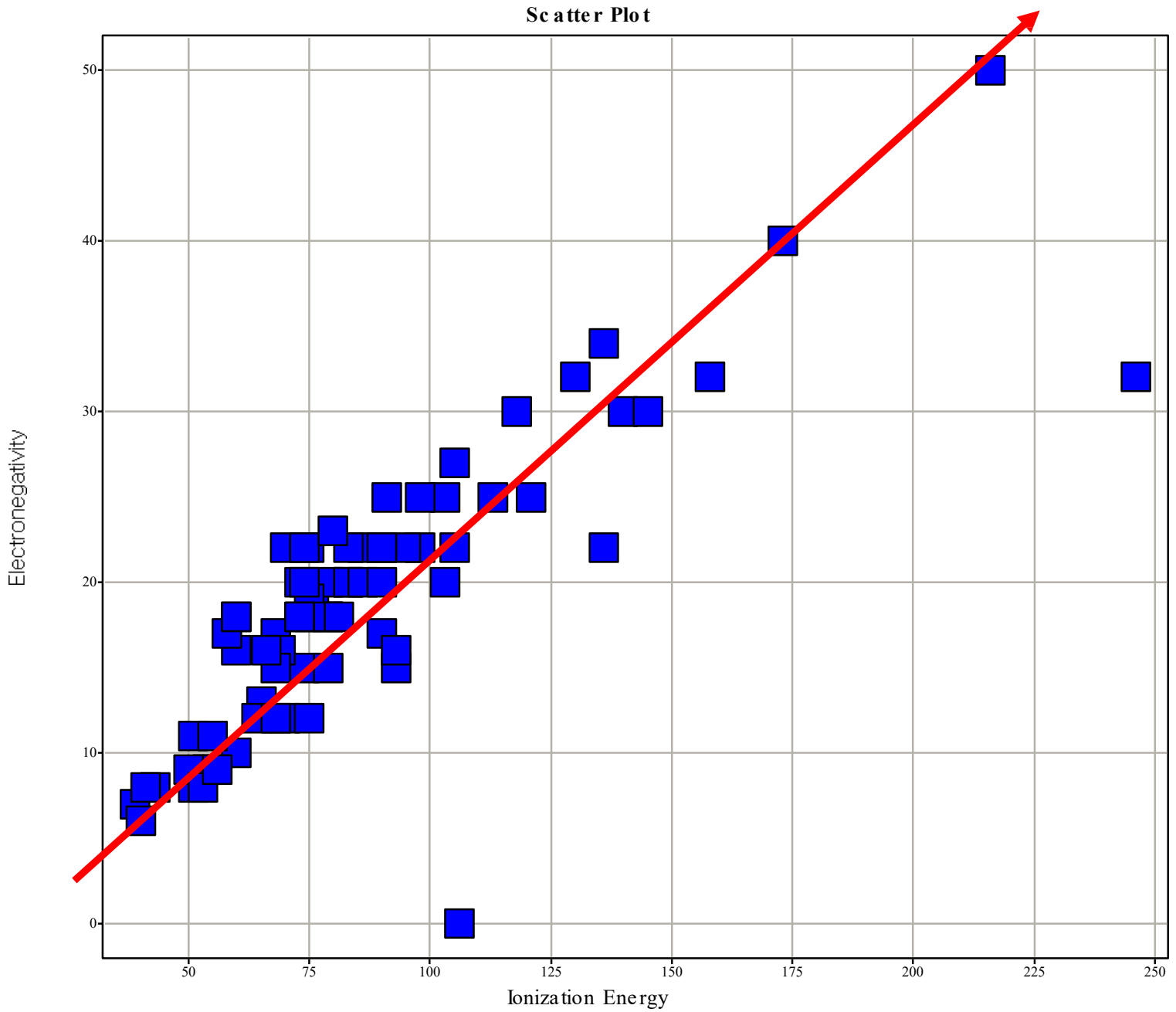


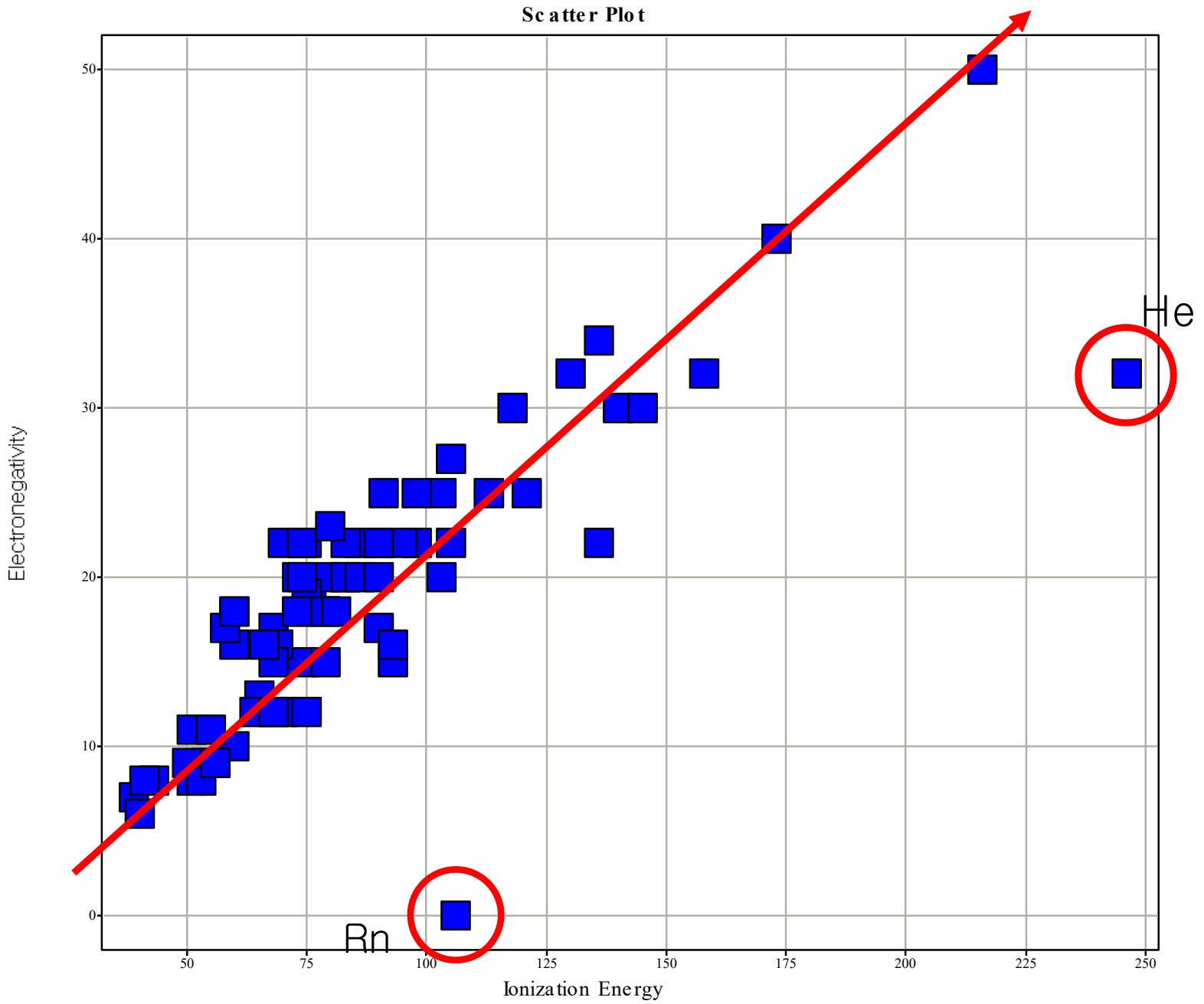
inspect “answers” and ask new questions

	A	B	C	D	E	F	G	H	I	J	K
1	Element	*P1	*P2	Atomic Num	Atomic Mas	Atomic Radi	Ionic Radius	Ionization E	Electronega	*C1	*C2
2	Ac	140	0	89	227	200	126	51	11	62	56
3	Ag	630	80	47	107	144	129	75	18	124	40
4	Al	750	160	13	27	143	67	60	16	28	25
5	Ar	1050	160	18	39	98	154	158	32	176	51
6	As	870	120	33	75	120	72	98	22	115	33
7	At	990	40	85	210	140	76	95	22	119	22
8	Au	630	40	79	197	144	99	91	25	131	22
9	B	750	200	5	10	85	41	83	20	101	8
10	Ba	80	40	56	137	222	149	51	8	46	56
11	Be	80	200	4	9	112	59	93	15	82	15
12	Bi	870	40	83	209	150	117	73	20	140	27
13	Br	990	120	35	79	114	182	118	30	161	44
14	C	810	200	6	12	77	30	113	25	82	1
15	Ca	80	120	20	40	197	114	60	10	70	51
16	Cd	690	80	48	112	151	109	90	17	113	43
17	Cl	990	160	17	35	100	167	130	32	173	47
18	Co	500	120	27	59	125	83	79	18	120	30
19	Cr	320	120	24	52	128	75	68	17	91	28
20	Cs	20	40	55	132	265	181	39	7	7	56
21	Cu	630	120	29	63	128	87	76	19	118	32
22	F	990	200	9	19	72	119	173	40	39	1
23	Fe	440	120	26	55	126	83	79	18	115	32
24	Fr	20	0	87	223	269	194	40	6	1	56
25	Ga	750	120	31	69	135	76	60	18	89	31
26	Ge	810	120	32	72	122	87	79	20	118	33
27	H	20	240	1	1	32	0	136	22	40	1
28	He	1050	240	2	4	31	93	246	32	1	1
29	Hf	200	40	72	178	159	85	70	12	95	44
30	Hg	690	40	80	200	151	116	103	20	147	27
31	I	990	80	53	126	133	206	105	27	153	44
32	In	750	80	49	114	167	94	58	17	93	42
33	Ir	500	40	77	192	136	82	90	22	116	25
34	K	20	120	19	39	227	152	43	8	37	56
35	Kr	1050	120	36	83	112	169	140	30	163	47

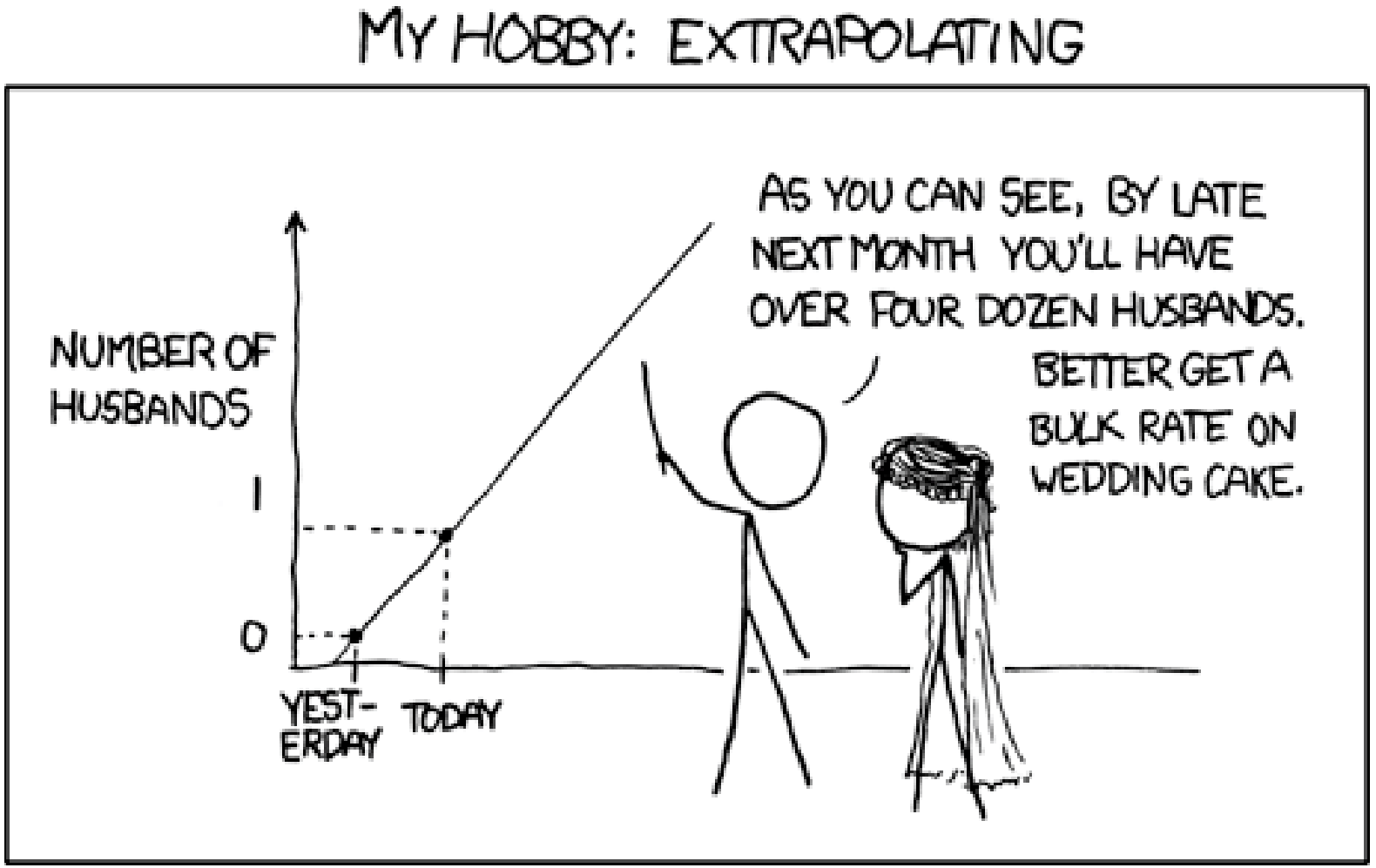
periodic



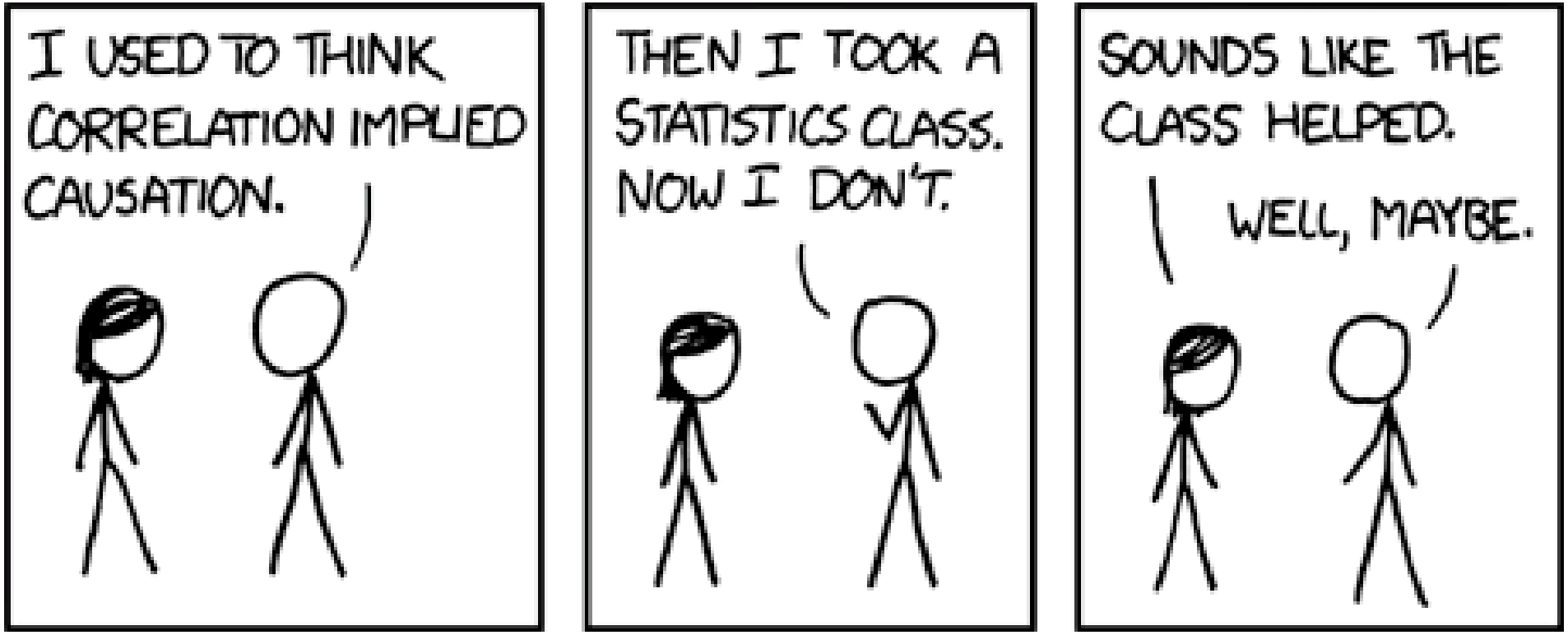




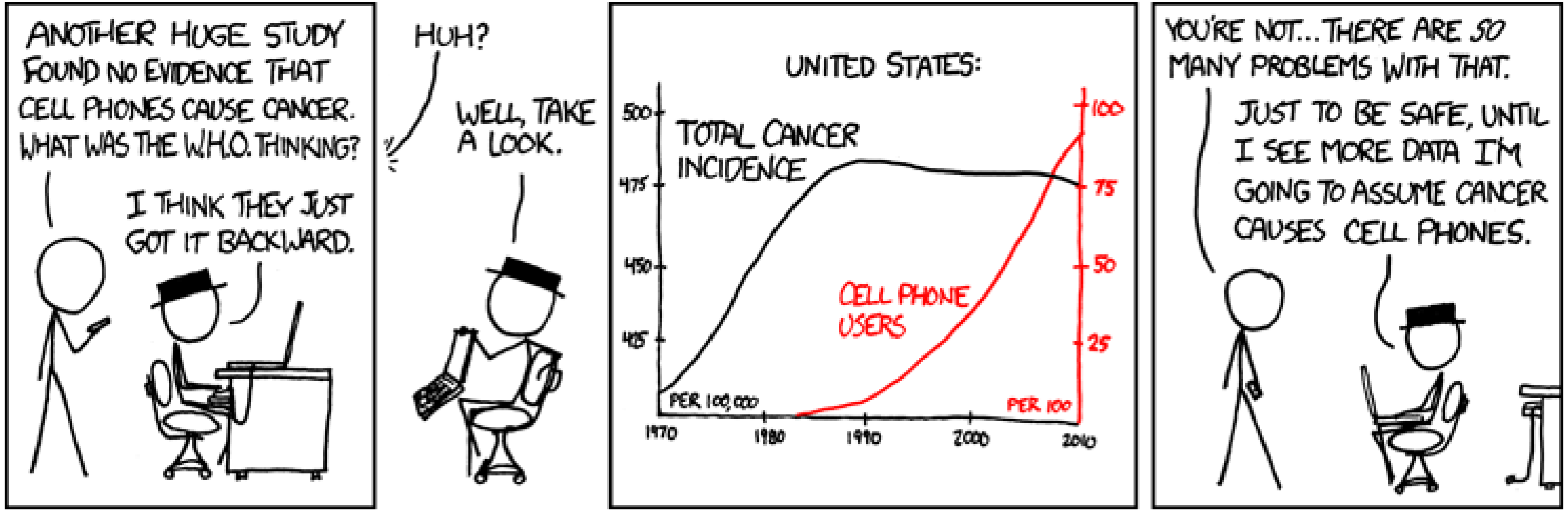
Extrapolating?

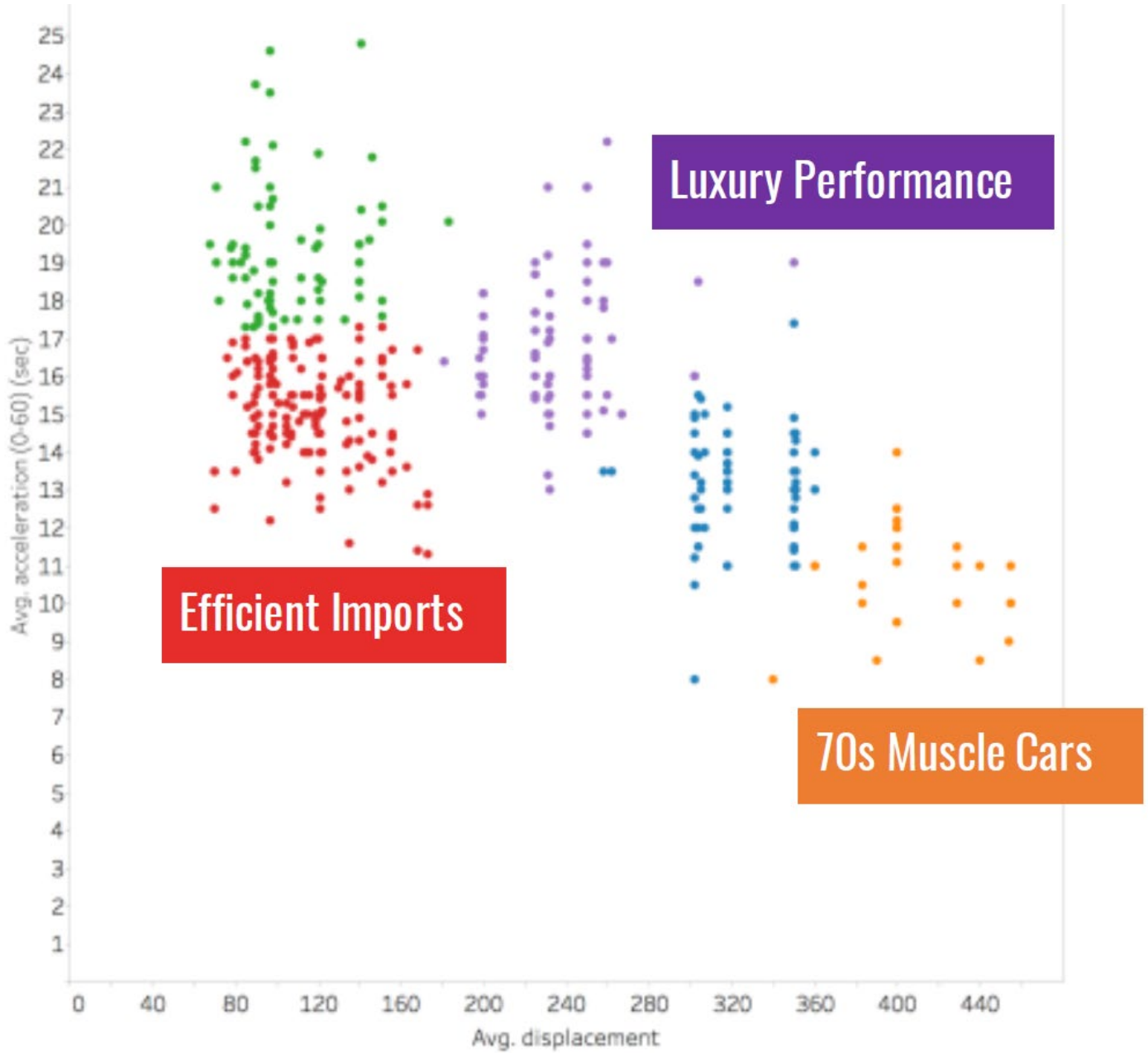


Correlation?

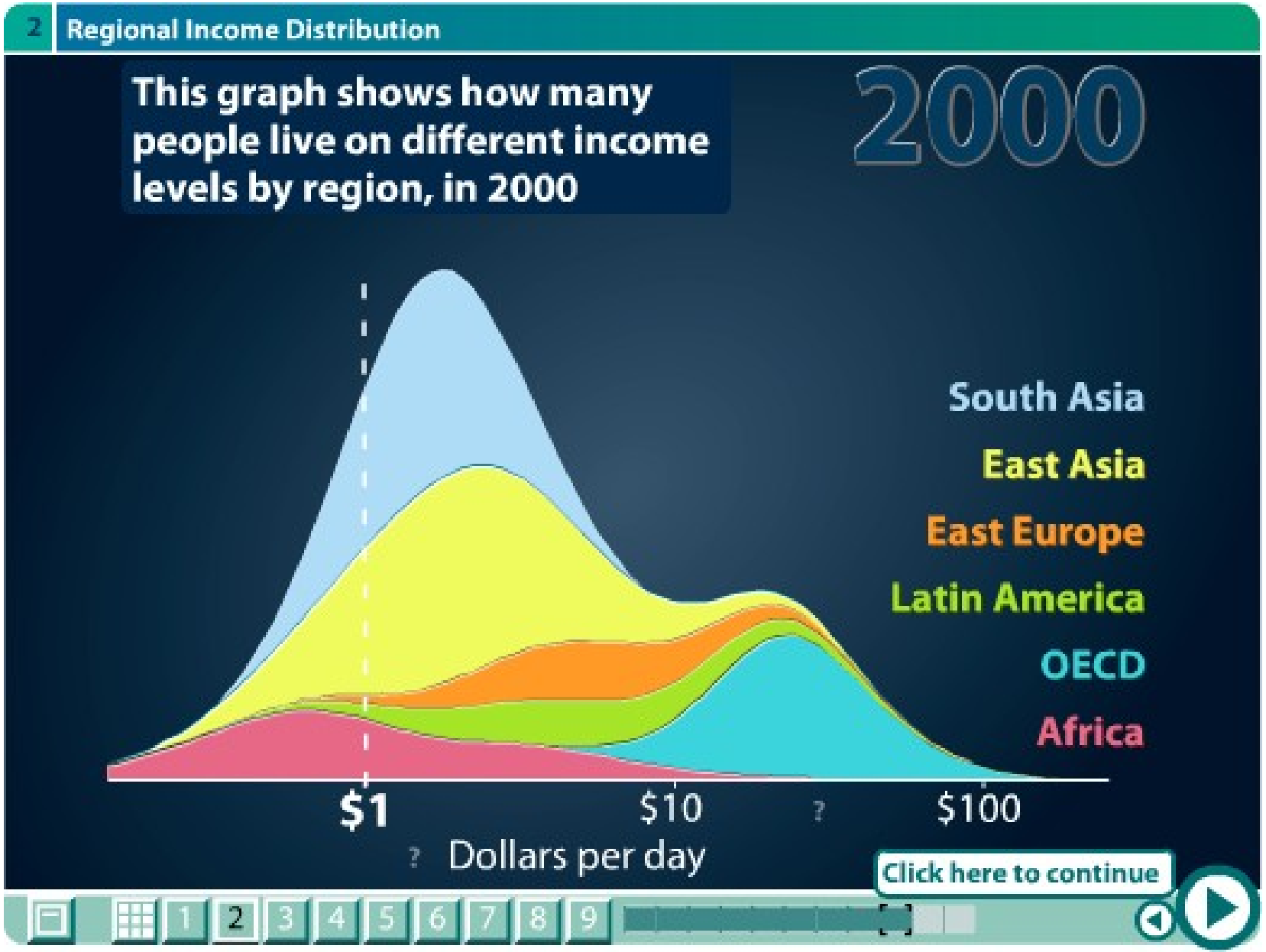


Causation?



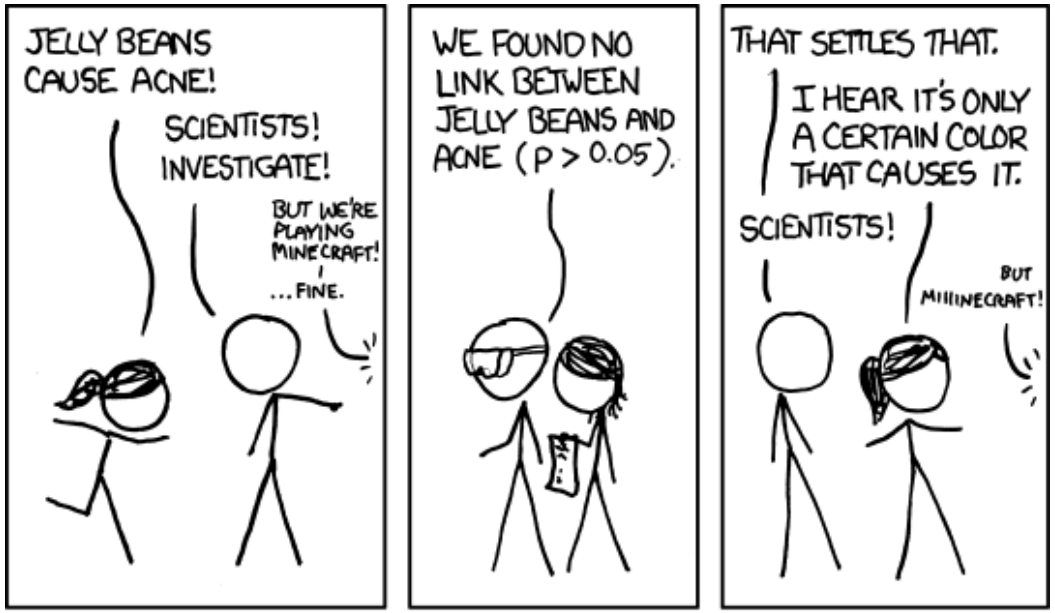


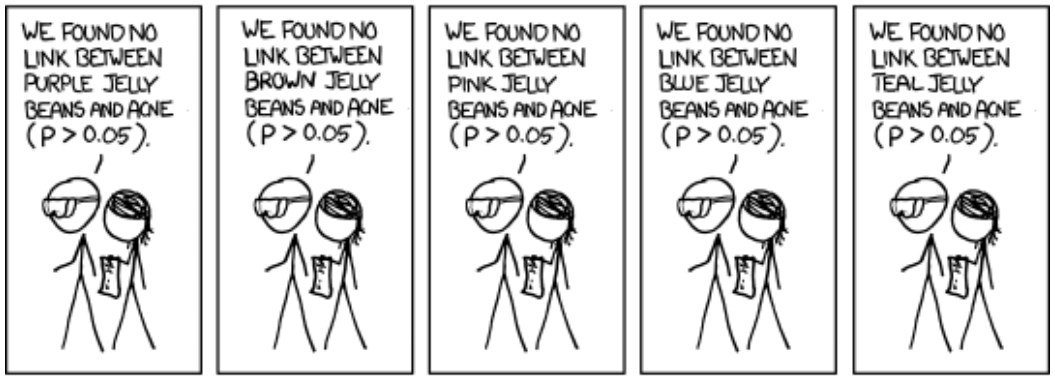
UNIVERSITY OF CALGARY

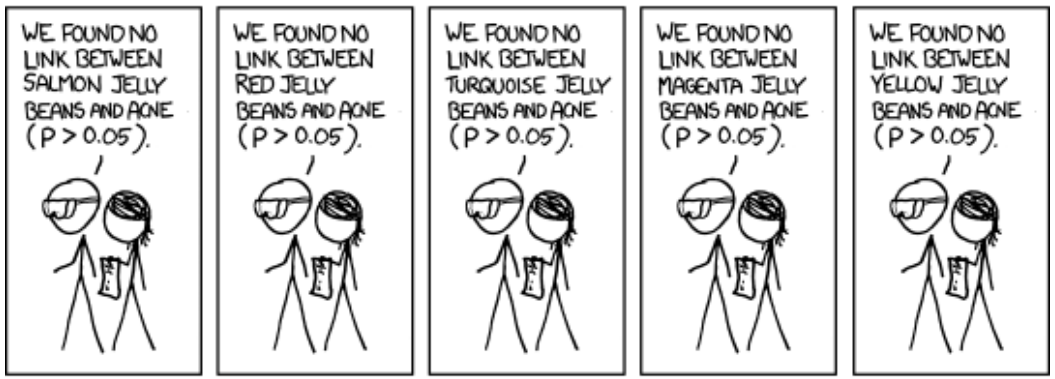


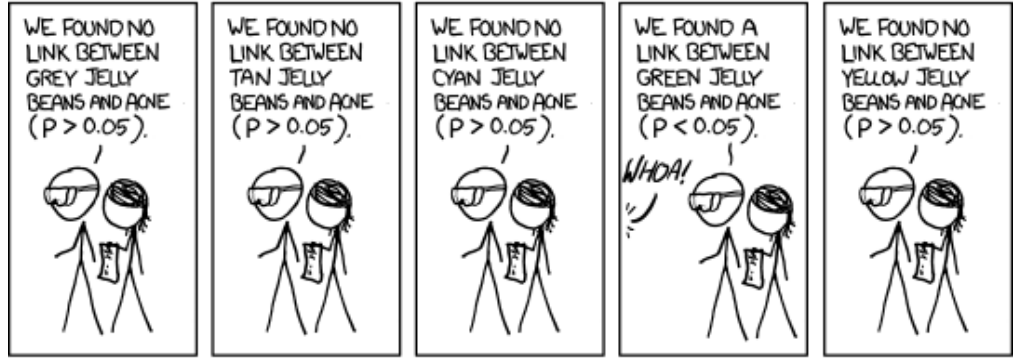
UNIVERSITY OF CALGARY

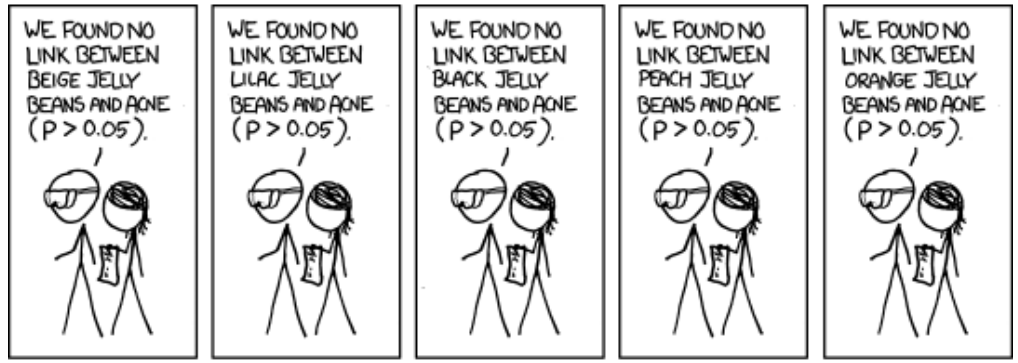

+ableau

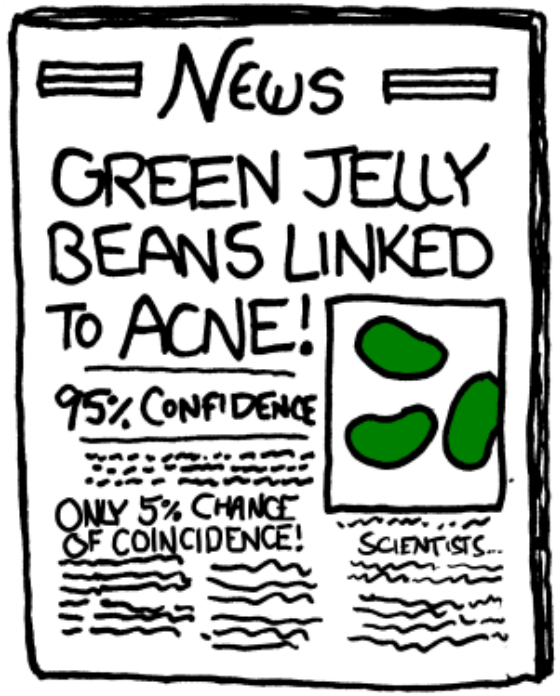




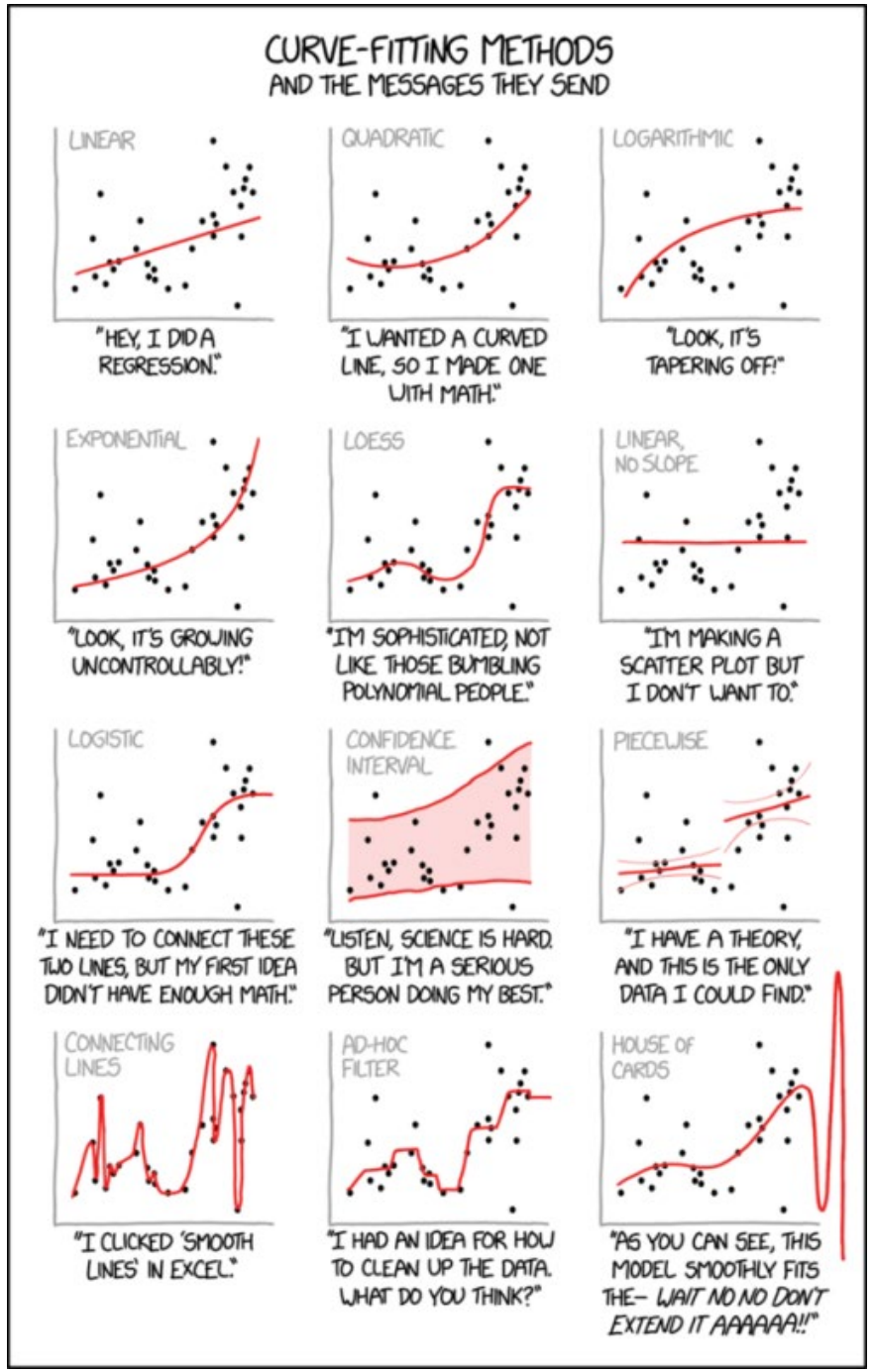








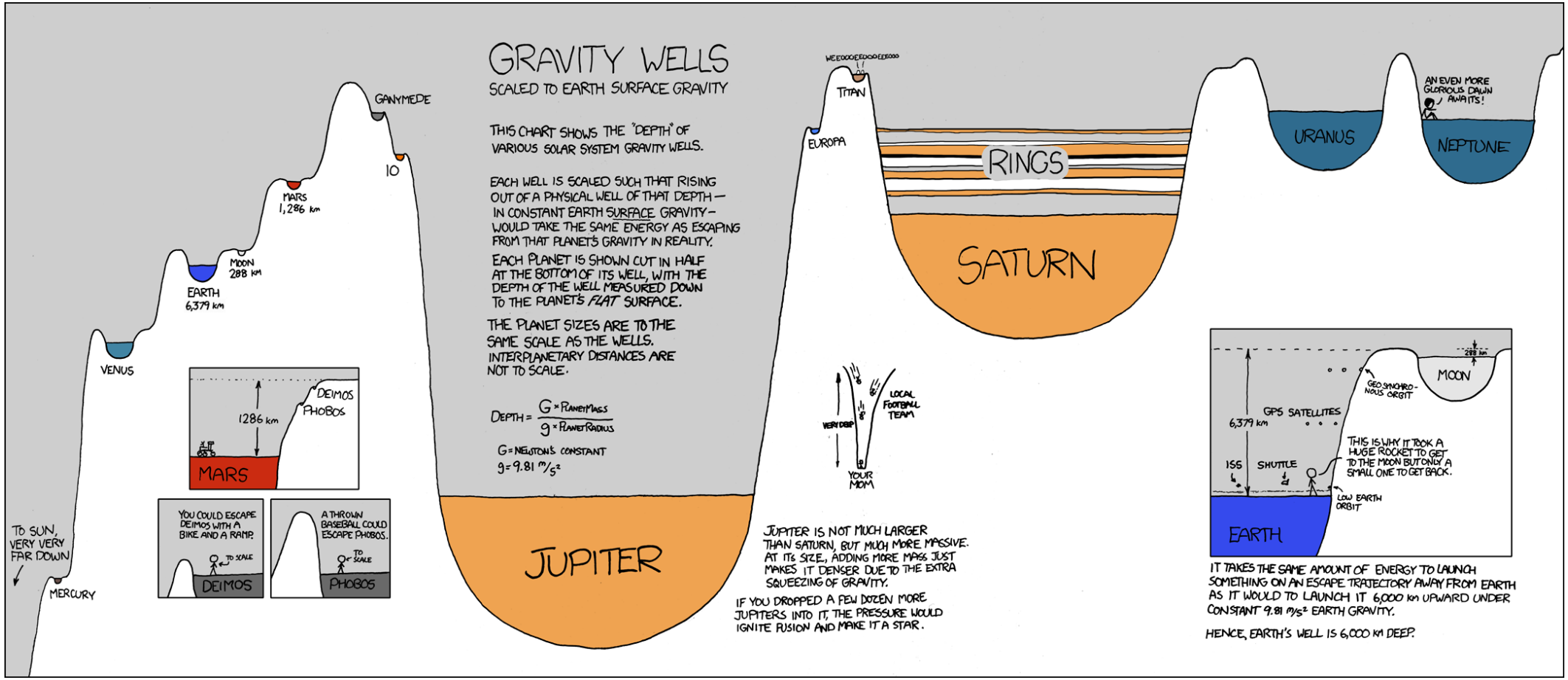
Curve-Fitting



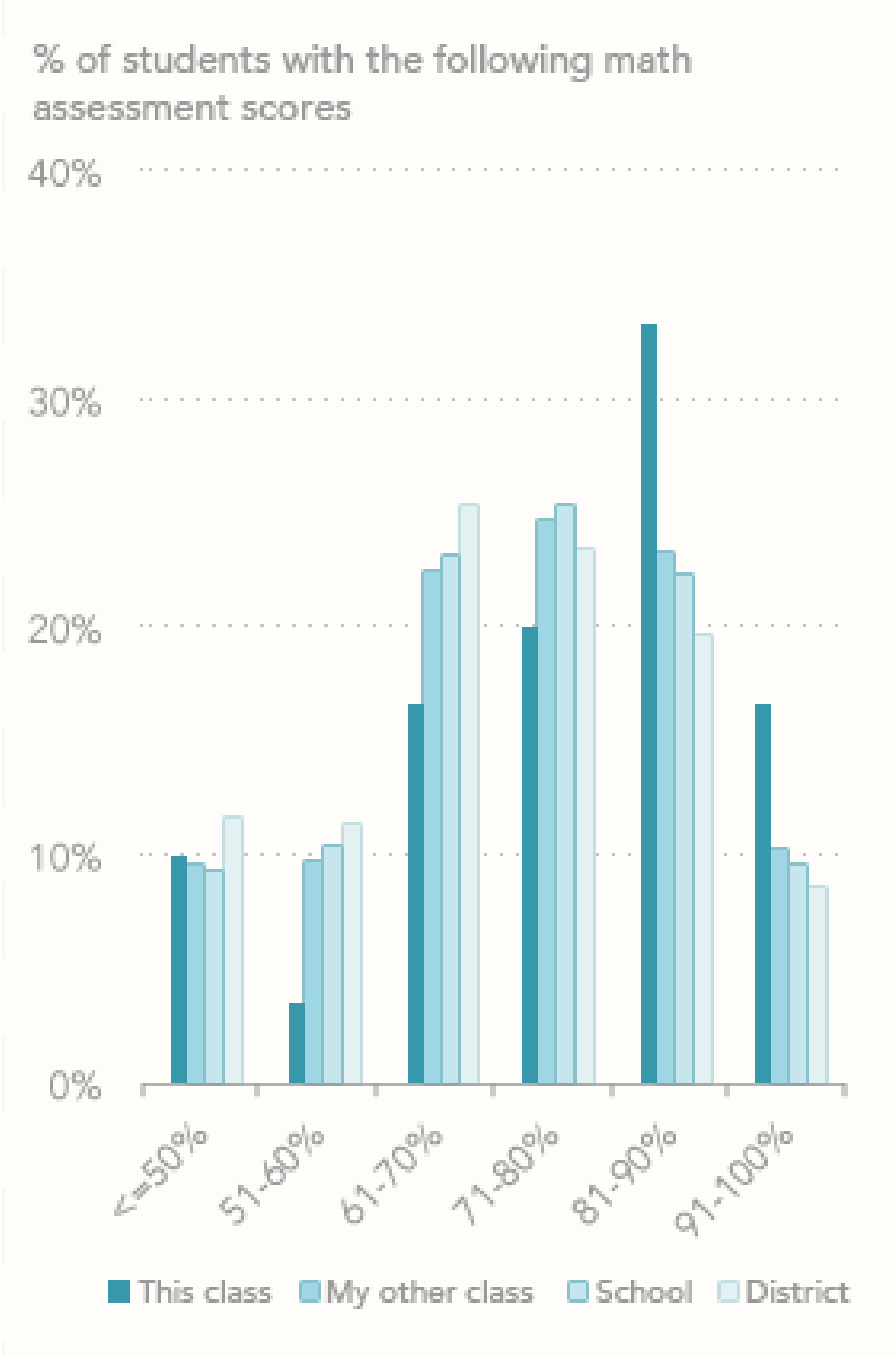
Data Science Pipeline - Presentation

Collection → Cleaning → Analyzing → **Presenting**

Visualization via Drawing - XKCD

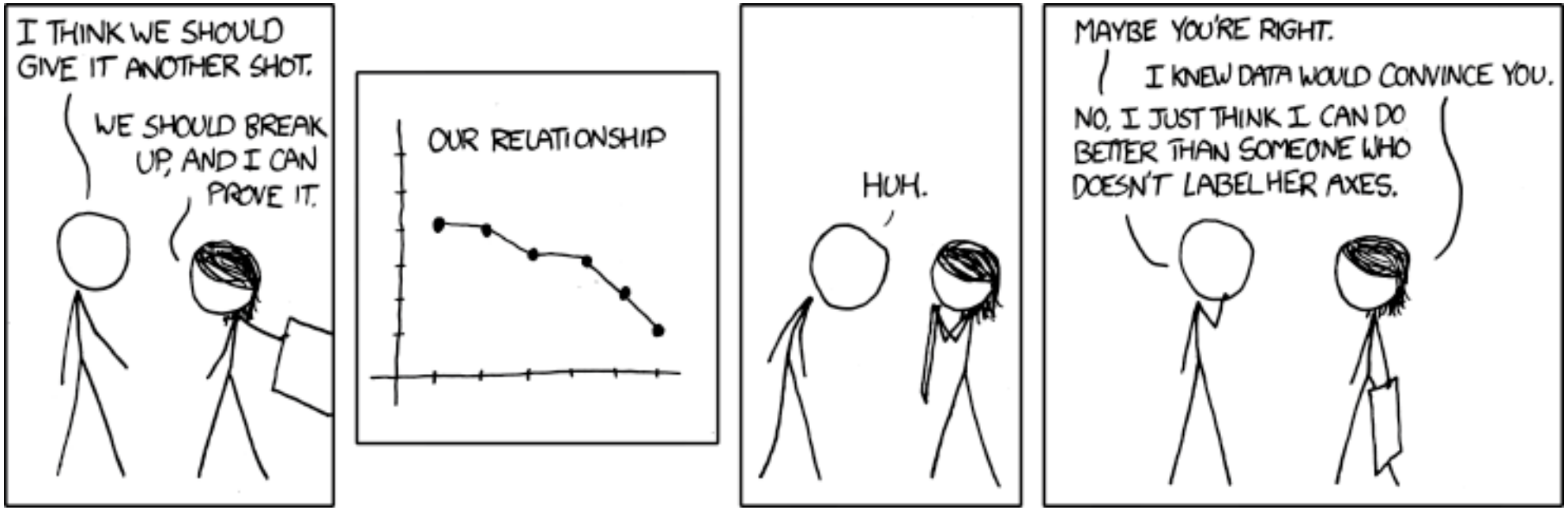


UNIVERSITY OF CALGARY



UNIVERSITY OF CALGARY

Presentation



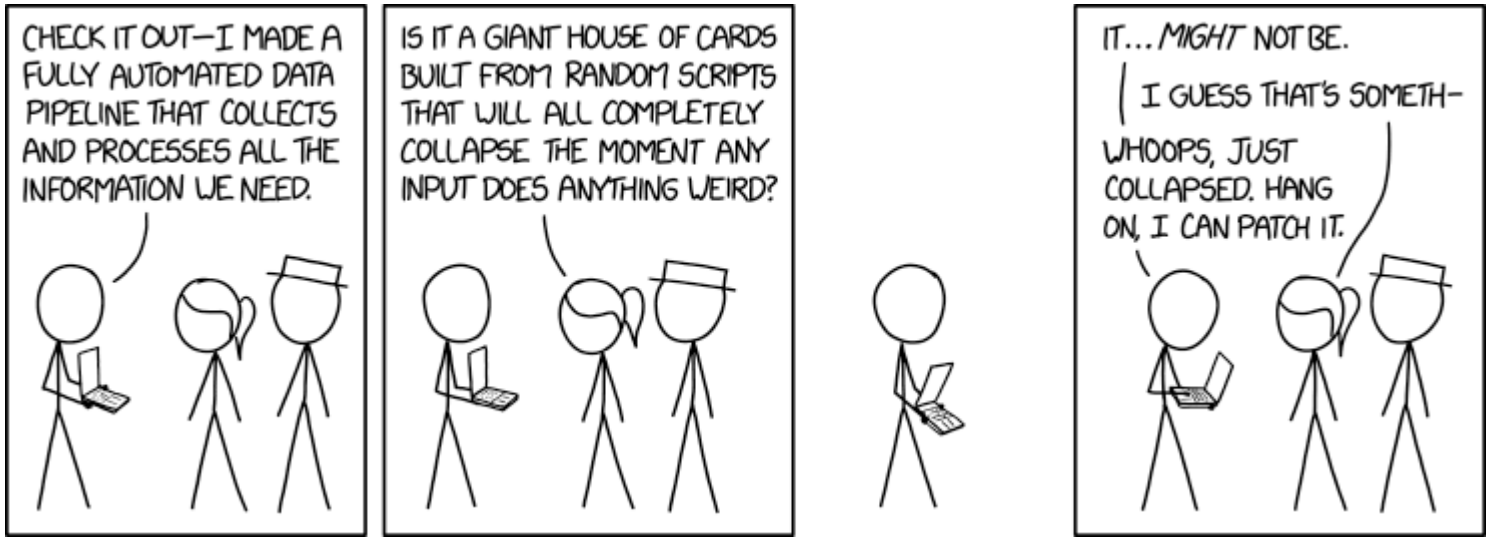
Written Reports

Written reports and responses

- How well can you convey ideas?
- How much details have you provided?
- Are you effectively using the combination of text and images? Is your report well structured?
- Is it easy for people to follow?

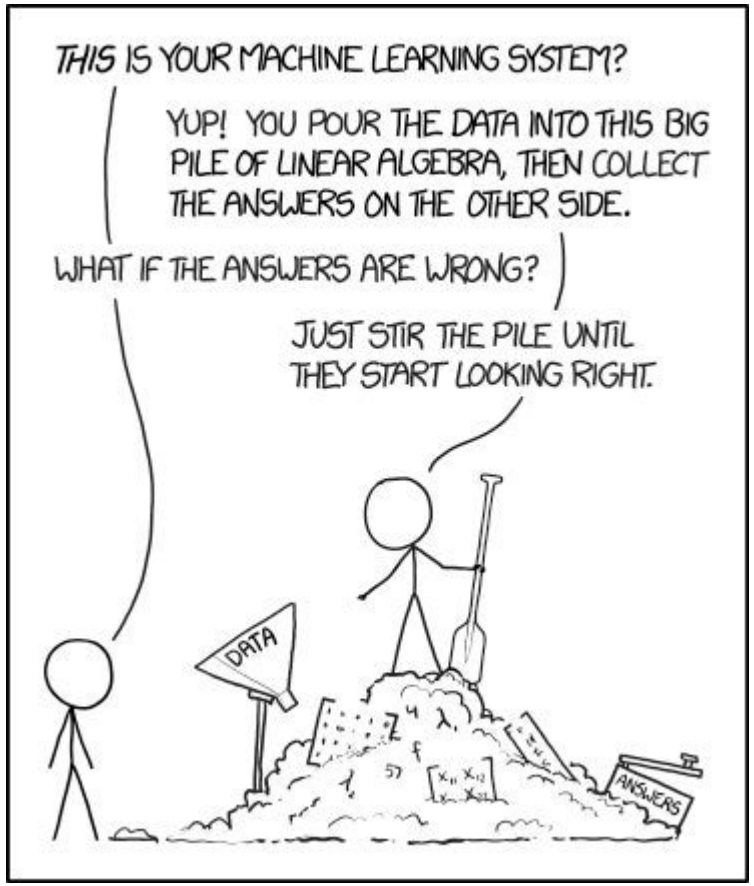
Data Pipeline – Automation?

Collection → Cleaning → Analyzing → Presenting



UNIVERSITY OF CALGARY

Machine Learning



UNIVERSITY OF CALGARY

Onward to …
Visualization

Jonathan Hudson
jwhudson@ucalgary.ca
https://pages.cpsc.ucalgary.ca/~jwhudson/

UNIVERSITY OF CALGARY