Data Science/AI Capabilities

Word cloud of this document

Professor Prem Saggar

Crypto

DARPA™

NATIONAL SECURITY AGENCY

UNITED STATES OF AMERICA

FINRA®

Microsoft

Twins, Triplets, and Crypto

- Twins & Triplets all have the same birthday 4 years apart
- All “6” boys
- 1/33*1/60,000*1/365 = 1/722 million <= chances < 1/33*1/200,000,000*1/365 = 1/2.4 Trillion
- Co-occurrence? → 3,805,024,570 → 3,805,024,570 /722,700,000 = 5 others are possible in the world → 1 of 5 on Earth ☺ or 1/1 in maybe the universe

MONEY
Product
IDEA
TEAM
INCOME
SUCCESS
1
2
3
400
500
ANALYSIS
MAIN
success
TEAMWORKING
Team
SUCEFSS
Business
IDEA
INCOME
Product
IDEA
money
IDEA

Data Science Defined

- **Data science** is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.
- The multiple disciplines are:
  - Mathematics
    - Calculus and Partial Differential Equations, …
  - Statistics
    - Distributions, missing data, regression techniques, predictions about a population based on a sample …
  - Computer Science
    - Programming, algorithms, complexity analysis, BIG DATA …
  - Information Science
    - Librarians, cataloging, information discovery, information retrieval, …

Basics of Statistics

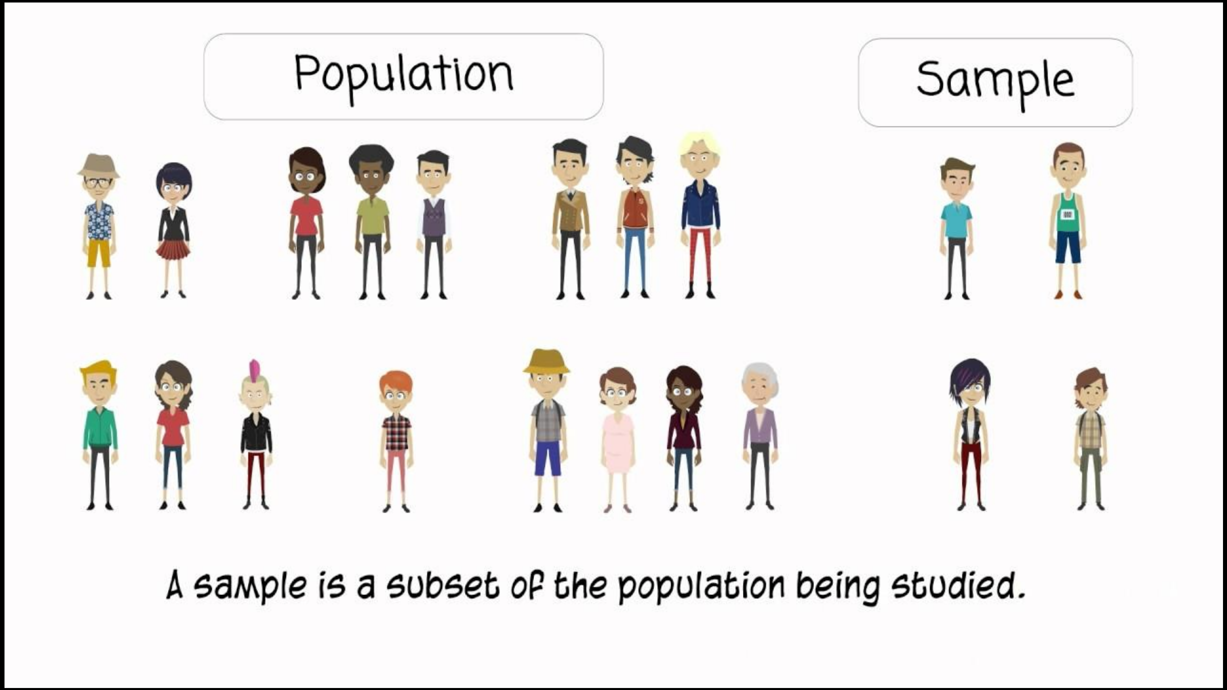
- Statistics is the science of making a prediction about an entire population based on a sample.
    - E.g.,
        - Predicting when the next attack will occur
        - Predicting who will win the next Presidential Election
    - How is this done? Through sampling, probabilities, and hypothesis testing.

Statistics = Use random samples to make confident statements about entire populations. Intelligent guesses/speculation.

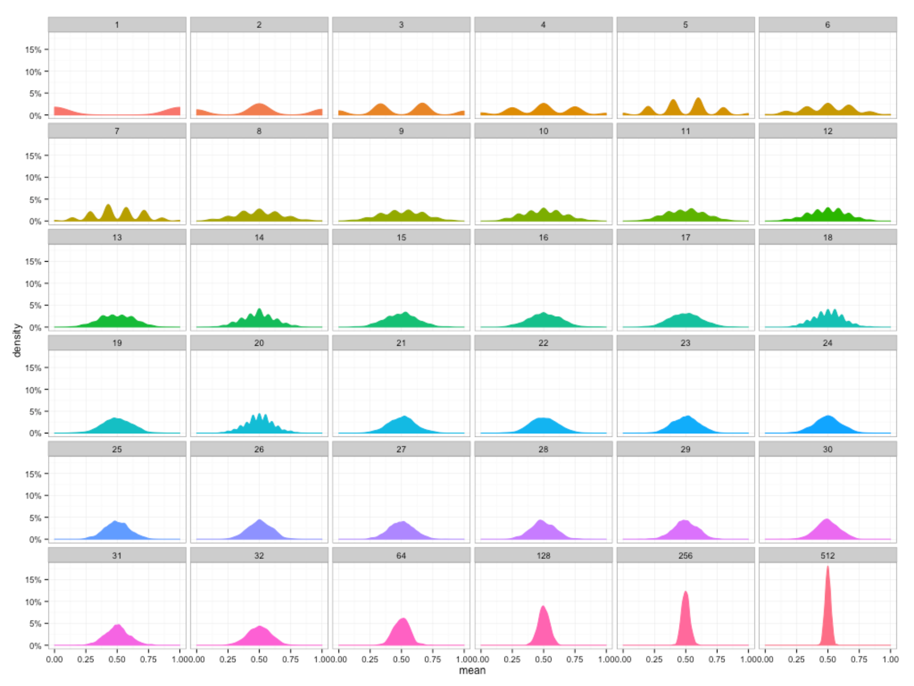
Sample, shape, location, and spread

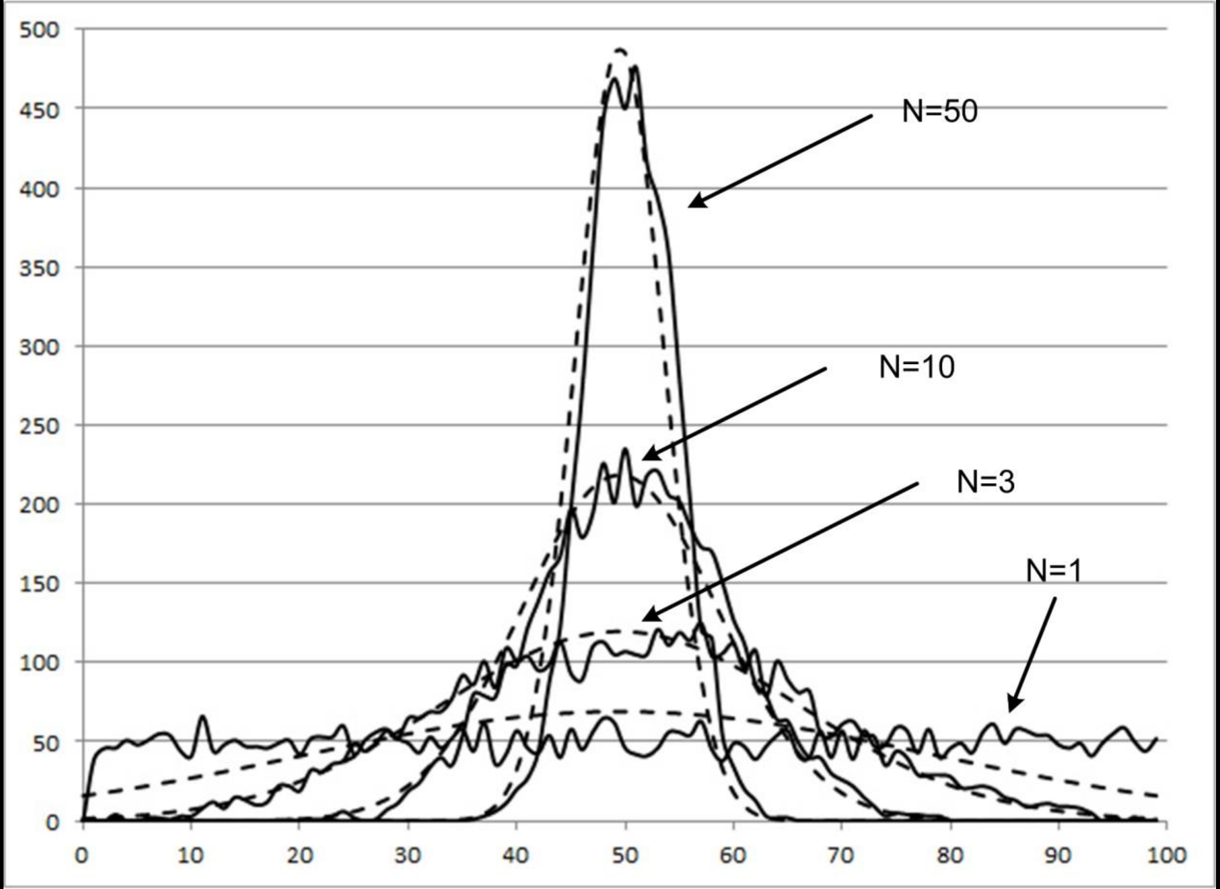
- Sample = make sure it's random, handle missing data (mcar, mar, nmar), imputation methods. NMAR!
- Shape = Is the data skewed, normal, or flat? If normal then we can use statistical analysis for normal distributions
- Location = Where does the data accumulate? Is it skewed, if so the median tells a better story.
- Spread = How much does the data differ? Standard deviations!

Samples give statistics, and populations (which we may never know give parameters)



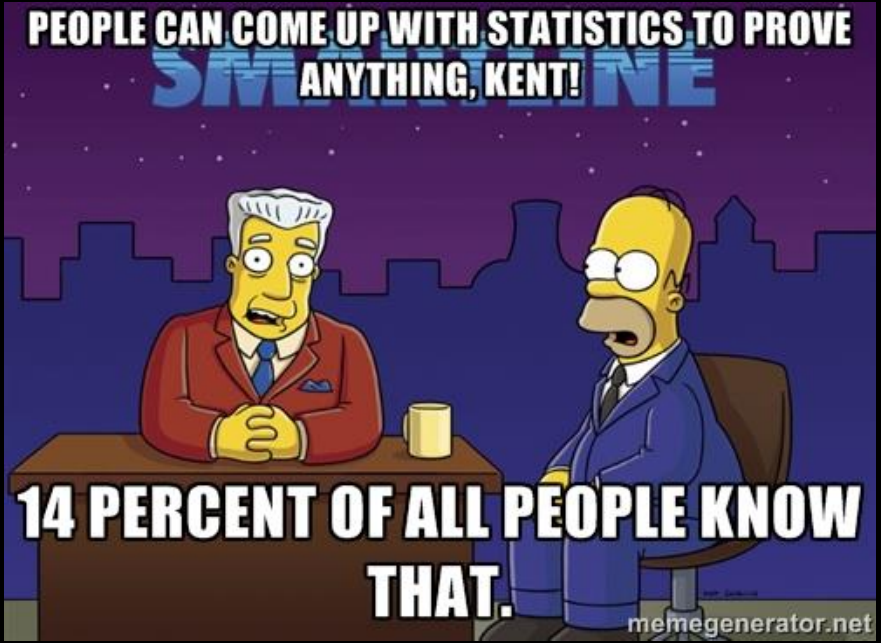
The Central Limit Theorem




500
450
400
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
N=50
N=10
N=3
N=1

We're all [redacted] until proven [redacted].

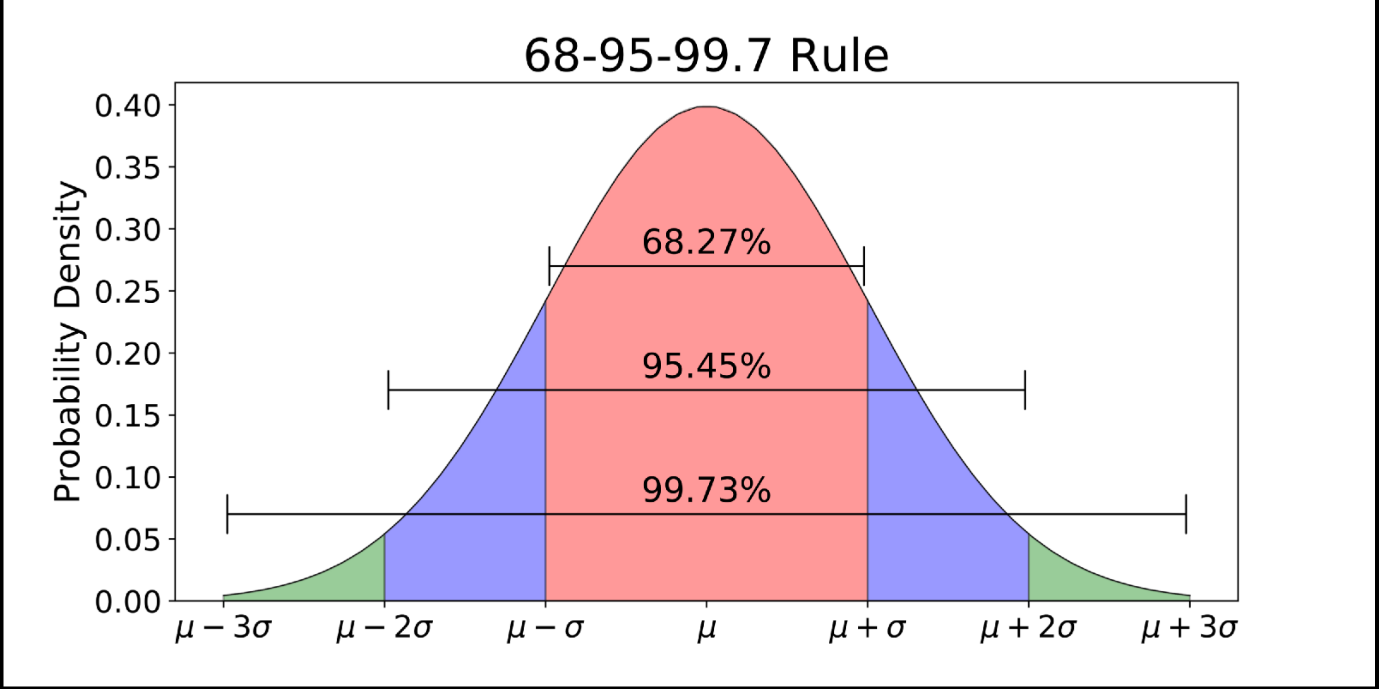
Hypothesis Testing


PEOPLE CAN COME UP WITH STATISTICS TO PROVE ANYTHING, KENT!
SMARTLINE
14 PERCENT OF ALL PEOPLE KNOW THAT.
memegenerator.net

How do we prove our statistical claims?

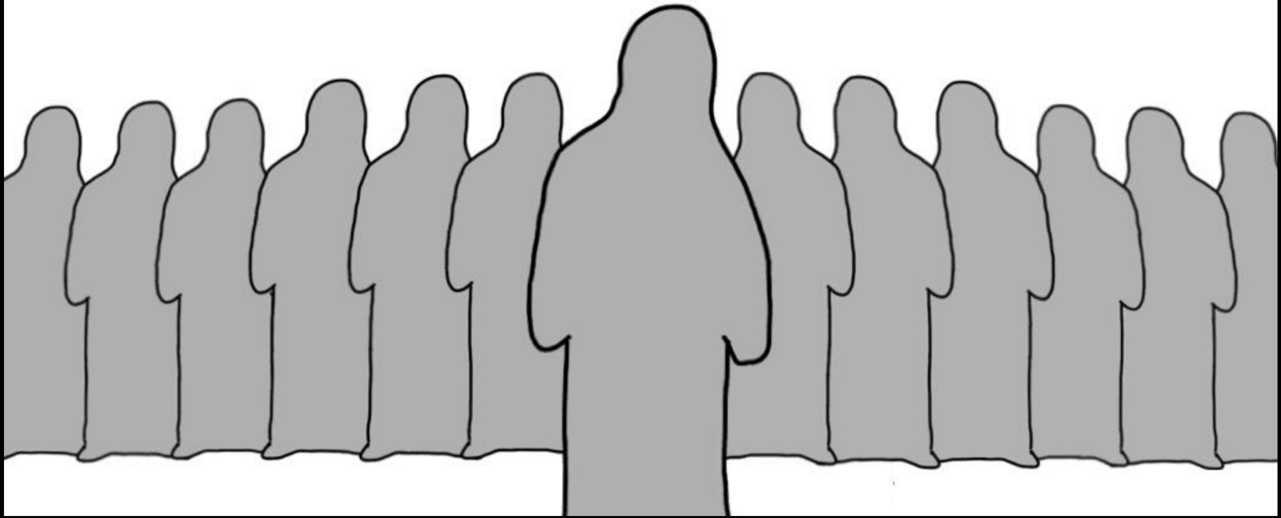
- Hypothesis Testing! The point of hypothesis testing is to make sure we don't jump to bad conclusions. Conclusions can be confusing (xylitol vs. fluoride). We are inherently speculating albeit rigorously. So we try to control the guessing by being conservative and use innocent until proven guilty.

Standard deviations and probability of the population mean



I am what is
The default, the status quo
I am already accepted, can only be rejected
The burden of proof is on the alternative

I am the null hypothesis



The alternative hypothesis tries to nullify the ghosts!

Warning

- It's important to note that statistics and inferences about "populations" are always approximations. We aren't using probabilities when we do have the entire population e.g., the entire Data Science class is our population. Here I can get the population mean etc... No guess work is necessary. Now what if we said all data science students in the world currently? That’s much harder if not impossible. Inference time!

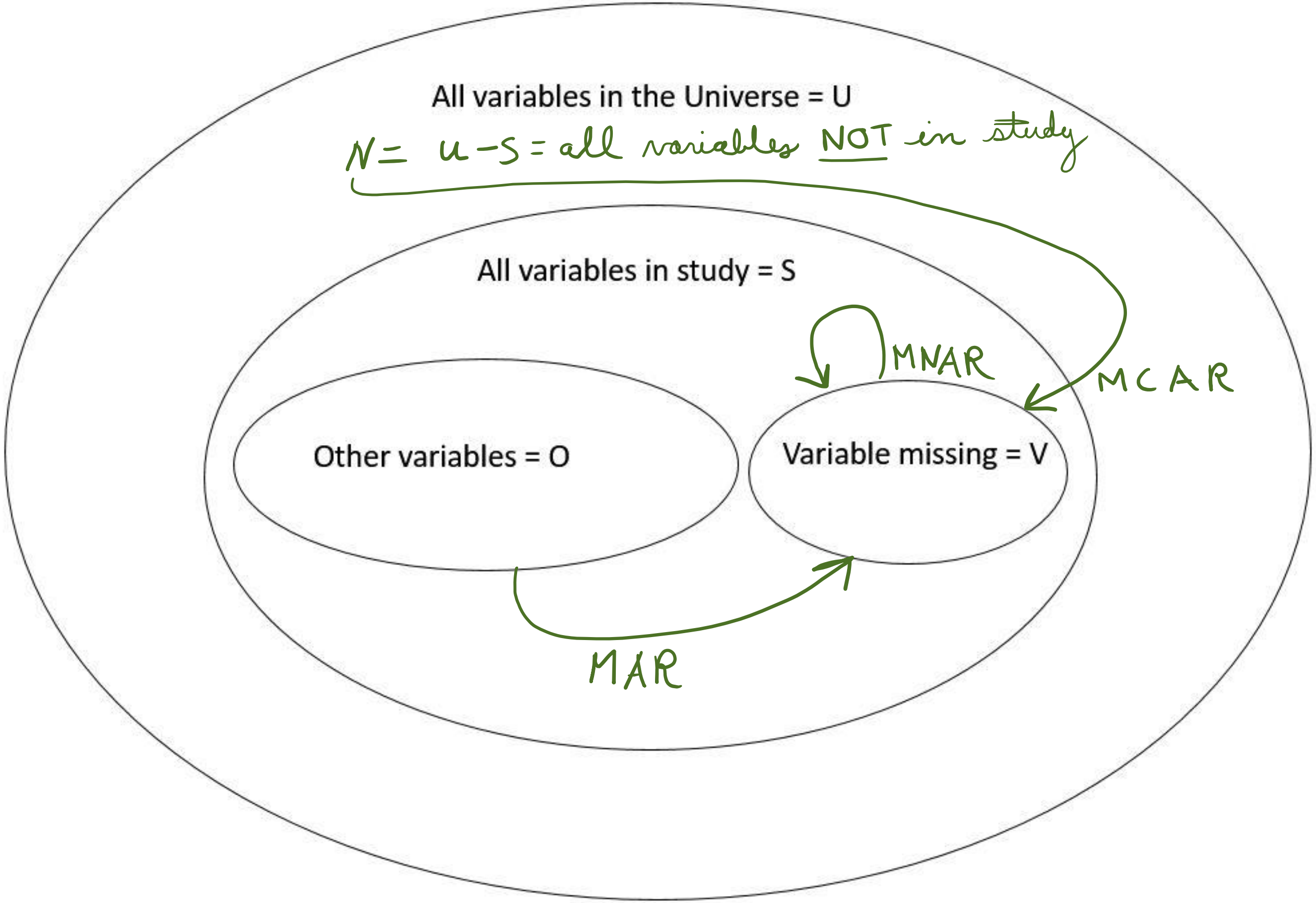
Survey

- Assumption: student will find Waldo given enough time.

Student	Time to find Waldo	Maximum Time
		10
		10
		10
		10
		10

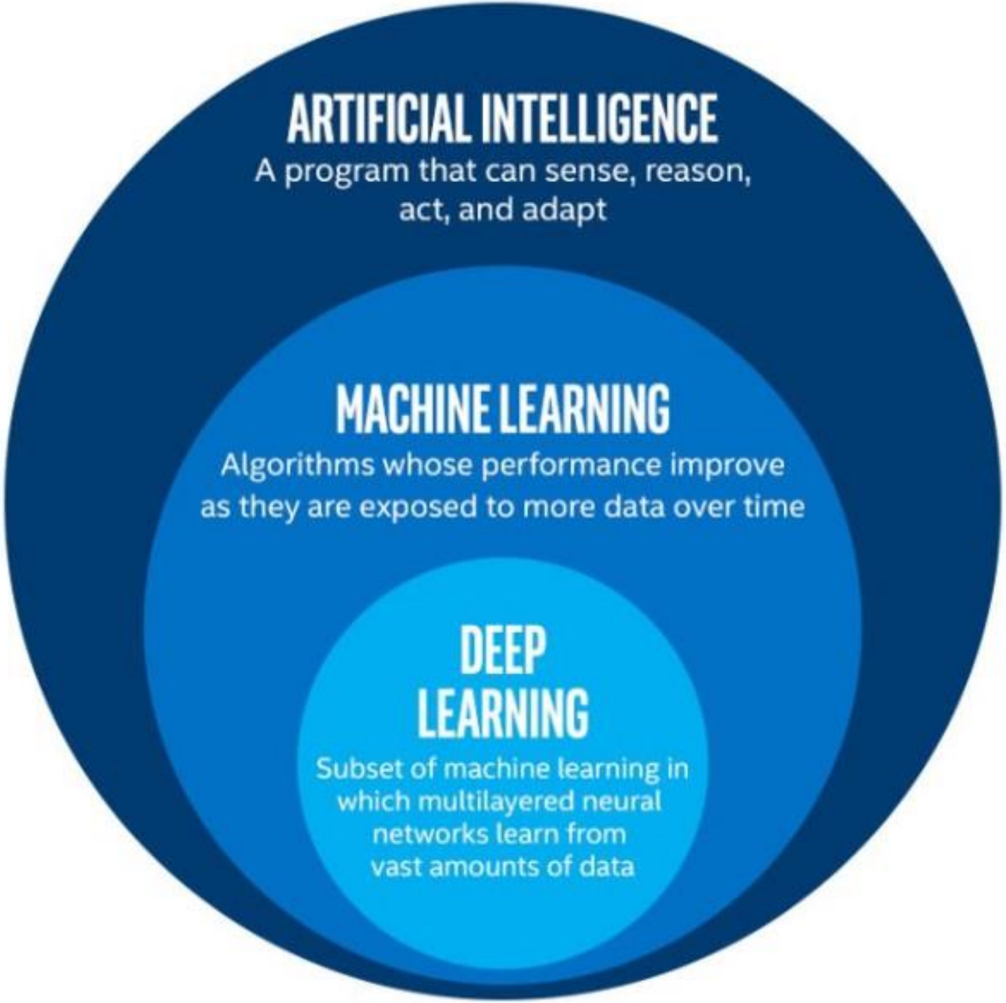
DOES MICHAEL BAY HATE your childhood?
CONFIRMED
Henching!
Magic show!
PLEASE COSPLAY US
this way
BITE OFF THEIR LIMBS
BEFORE
AFTER
Exit
FOOD
POKEMEAT
2DAYS SPECIAL PIKASTEW $10
Heart HEARTS $2
Got Hearts?
SELL UR USED GAMES HERE
WE'LL PAY $5!!!
EG GAMES
USED GAMES $99.99
HELP
iDeathNoteBook Air
KISSING BOOTH for CHARITY
Gamecon Tables
DISKINECT
PLANK HERO
Nice
PXL con COSPLAY COMPETITION
WORKSHOP HOW TO: Zombie COSPLAY
WORKSHOP HOW TO: KILL A Zombie
SPONSORED BY
FREAKSHOW COLLECTIBLES
LIFE-LIKE FIGURINES!
ZOMBIE-KILLING WORKSHOP AHEAD!!!
DEAD
AROD THE GAME
TINY SAILOR
DO NOT CROSS THE STREAMS!
tase us bro
CAUTION! MINE
TRY OUT REAL-LIFE GADGETS FROM MOVIES, GAMES & TV!
GADGET ZONE
Armor
Pink Unicorn Comics
Buff Vampire Slayer
FREE SHOULDER MASSAGE
FREE HUGS
Maid Cafe
I ♥ YAOI
SONY
Nogo
BLEACH
KILLS mold + germs + Evil Spirits
Closed Due to our booth being stolen
Wow!
Petting ZOO
GUN DAM
UR MOM
First Aid
PXL CON
jimmysomething
You Are Here

Missing data



All variables in the Universe = U
N = U − S = all variables NOT in study
All variables in study = S
MNAR
MCAR
Other variables = O
Variable missing = V
MAR

Missing Data	Cause of missing	Example	Type	Impute outcome	Imputable	Action
V	N	Variable missing because of freak event, or any variable you did not model	MCAR	freak event is random, so imputation is not biased	TRUE	Rerun if possible
V	O	Variable missing because of variable(s) you DID model, e.g., blood not drawn by clinician so verdict not possible or unable to find something based on time	MAR	50/50 chances for missing data is modelled so imputation is ok	TRUE	Rerun if possible
V	V	Variable missing because of the variable! I'm not telling you, e.g., what is your income	MNAR	skewed incomes everyone looks middle class	FALSE	Change survey



ARTIFICIAL INTELLIGENCE

A program that can sense, reason,
act, and adapt

MACHINE LEARNING

Algorithms whose performance improve
as they are exposed to more data over time

DEEP LEARNING

Subset of machine learning in
which multilayered neural
networks learn from
vast amounts of data

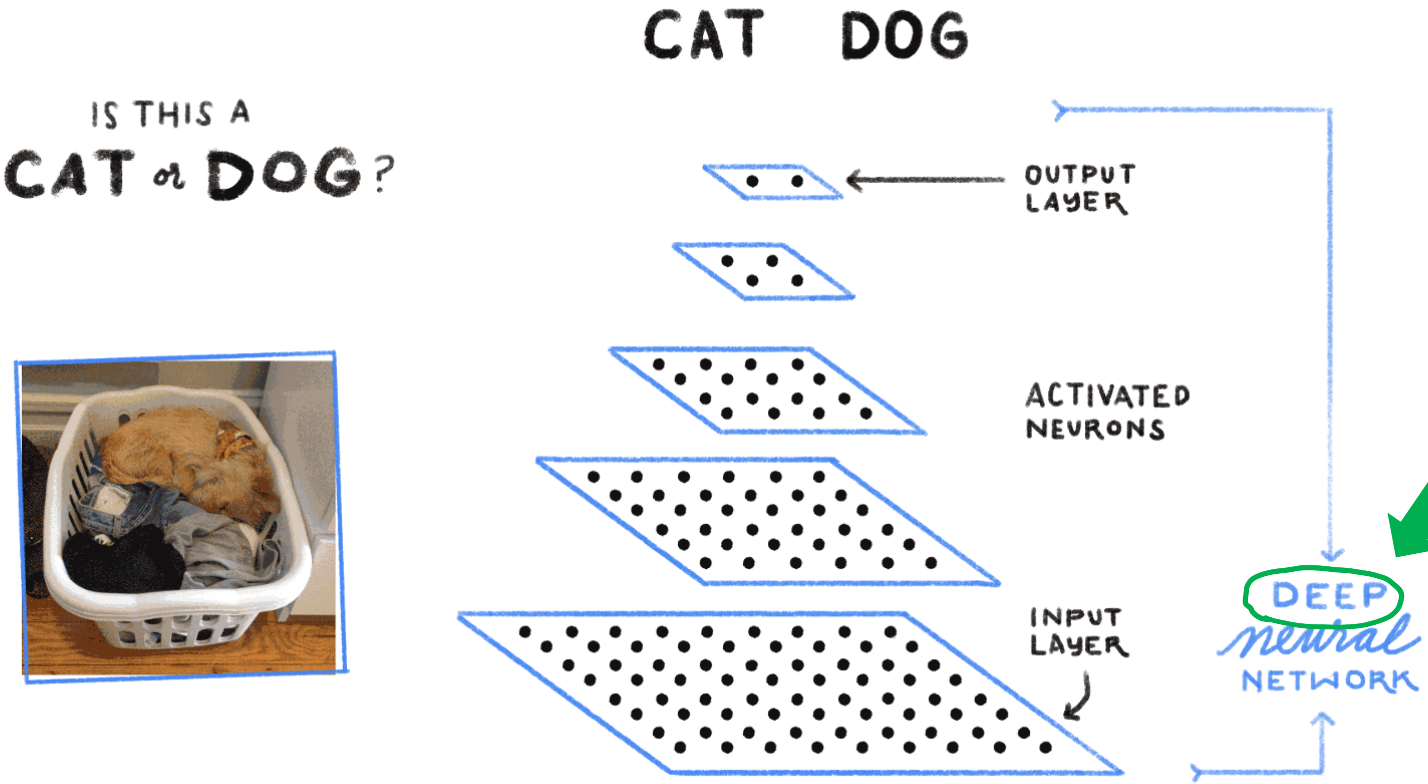
Unsupervised vs. Supervised Learning

Supervised learning is aided by training data and human correction.

Here’s some training data. Learn the patterns. Make your best guess at what the patterns are. We’ll feed you test data to figure out if you’ve understood it. If you stray off course, we’ll correct you and retrain.

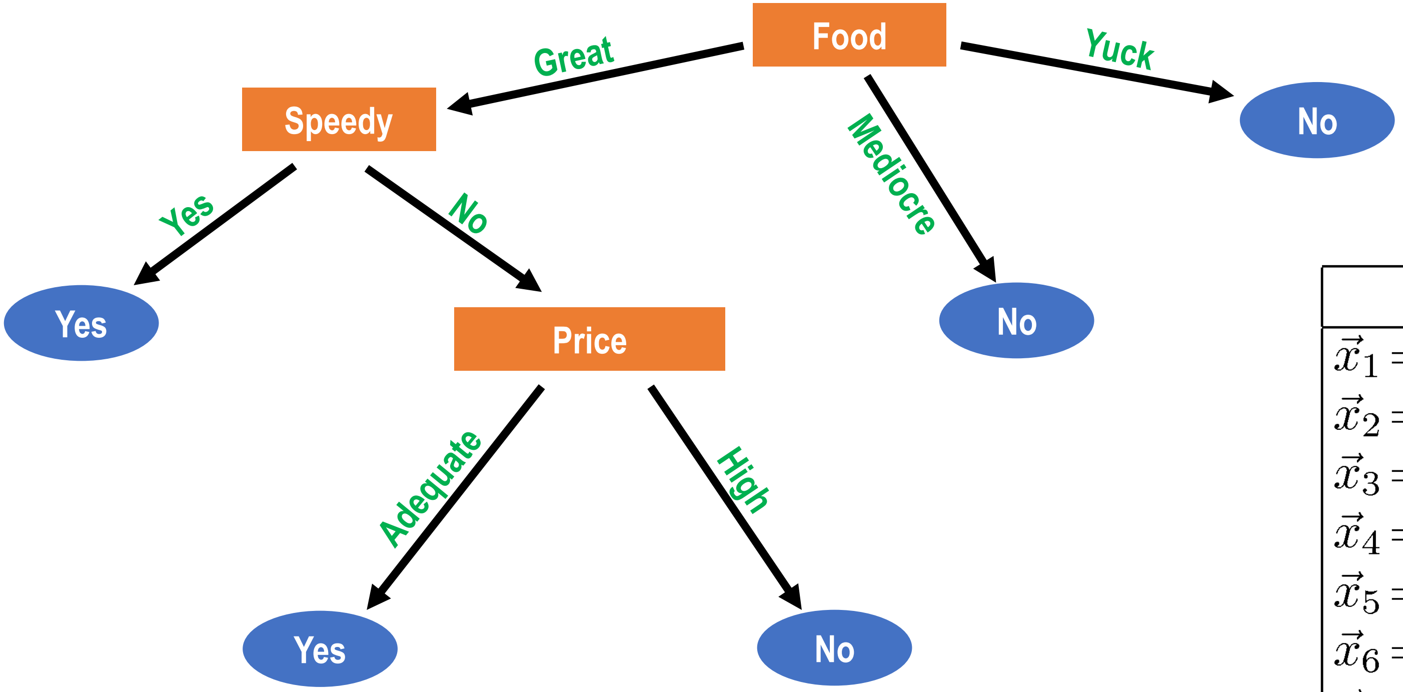
Examples include Decision Trees and Neural Networks.

Unsupervised learning is uncorrected and runs on data. It can’t classify things “yet”. But is very good at clustering and anomaly detection.



What's "deep" about it?

Decision Tree Example: “BigTip”



Our Data

F S P	BigTip
$\vec{x}_1 = (\ \mathrm{g}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_1) = 1$
$\vec{x}_2 = (\ \mathrm{g}, \mathrm{n}, \mathrm{h}\ )$	$f(\vec{x}_2) = 0$
$\vec{x}_3 = (\ \mathrm{g}, \mathrm{y}, \mathrm{h}\ )$	$f(\vec{x}_3) = 1$
$\vec{x}_4 = (\ \mathrm{g}, \mathrm{n}, \mathrm{a}\ )$	$f(\vec{x}_4) = 1$
$\vec{x}_5 = (\ \mathrm{m}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_5) = 0$
$\vec{x}_6 = (\ \mathrm{y}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_6) = 0$
$\vec{x}_7 = (\ \mathrm{g}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_7) = 1$
$\vec{x}_8 = (\ \mathrm{g}, \mathrm{y}, \mathrm{h}\ )$	$f(\vec{x}_8) = 1$
$\vec{x}_9 = (\ \mathrm{m}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_9) = 0$
$\vec{x}_{10} = (\ \mathrm{g}, \mathrm{y}, \mathrm{a}\ )$	$f(\vec{x}_{10}) = 1$

Association Rules!!!

Buying Pattern
Beer, Orange Juice, Diapers
Beer, Q-Tips, Diapers
Beer, Chips
Beer, Advil

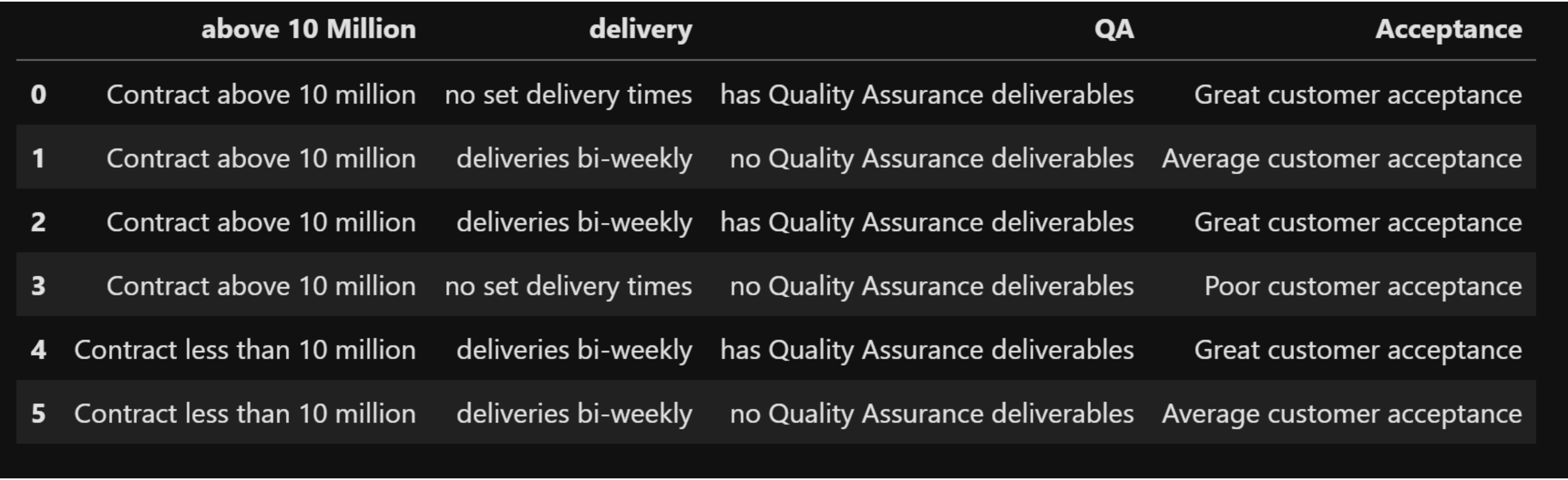
Derived rules

Beer → Diaper 50% confidence

Diapers → Beer 100% confidence

Association Rules example for financial transactions.

Please consider the following mock financial contract data. What patterns can we find for Great customer acceptance?



	above 10 Million	delivery	QA	Acceptance
0	Contract above 10 million	no set delivery times	has Quality Assurance deliverables	Great customer acceptance
1	Contract above 10 million	deliveries bi-weekly	no Quality Assurance deliverables	Average customer acceptance
2	Contract above 10 million	deliveries bi-weekly	has Quality Assurance deliverables	Great customer acceptance
3	Contract above 10 million	no set delivery times	no Quality Assurance deliverables	Poor customer acceptance
4	Contract less than 10 million	deliveries bi-weekly	has Quality Assurance deliverables	Great customer acceptance
5	Contract less than 10 million	deliveries bi-weekly	no Quality Assurance deliverables	Average customer acceptance

There are actually 190 rules we can derive and here are some key ones:
1) has Quality Assurance Deliverables → Great Customer acceptance @ **100%** confidence (happens every time)
2) no Quality Assurance Deliverables → Average customer acceptance @ **66.7%** confidence (happens 2 out of every 3 times)

What can we do with Association Rules?

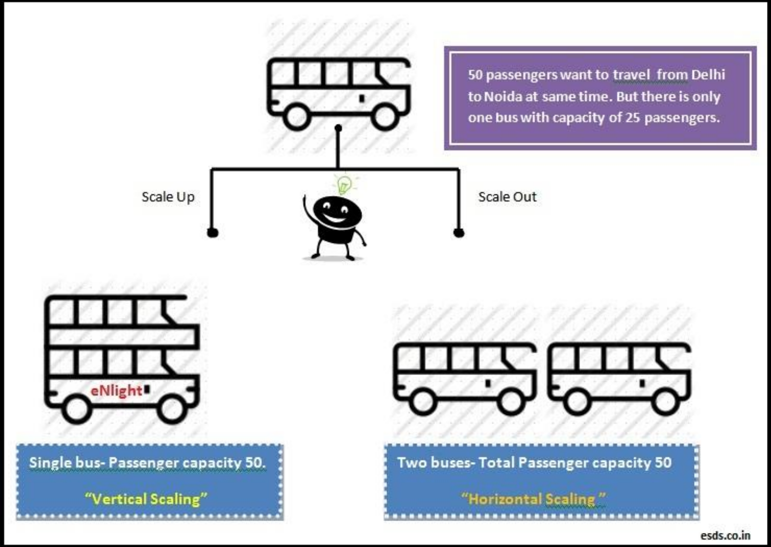
1. Allow us to objectively report these are the rules within our data
2. Discover new behaviors we weren’t aware of
3. Finds rules with confidences that are not 100% but are still important
4. After examining the rules found we can say these are the target outcomes we’re interested in e.g., Great Customer Acceptance and Contracts less than 10 million.
5. Once we have target outcomes, we can use Decision Trees to illustrate and predict rules for these target outcomes.

Interview Problem for data scientists

- Given the following:
  - 3,6,4 = 18, 108, 648, 3888
  - 2,-2,5 = -4,8,-16,32,-64
- What is
  - 2,3,4 = ?
- Is there a formula for the last term?
- Can you write a program to get the last term?
- Can you prove your program works?
- What is the time complexity?

Big Data

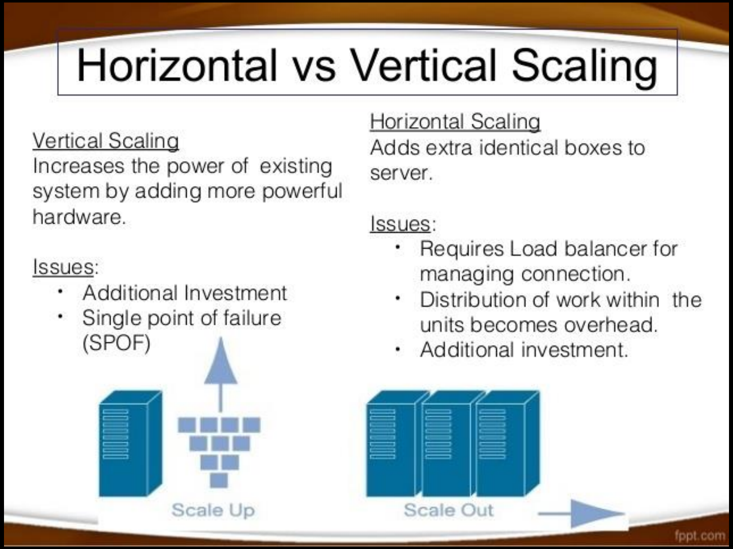
- What is Big Data? Is it Dangerous ☺?
- Well what is Big?
  - Big is a relative term here and is assumed to be big by “today’s” standards. However, it really means too “big” for you to handle via vertical scaling. By vertical we mean single node improvement vs multi node horizontal improvement.



50 passengers want to travel from Delhi to Noida at same time. But there is only one bus with capacity of 25 passengers.

Single bus- Passenger capacity 50. “Vertical Scaling”

Two buses- Total Passenger capacity 50 “Horizontal Scaling ”



Horizontal vs Vertical Scaling

Vertical Scaling
Increases the power of existing system by adding more powerful hardware.

Issues:
- Additional Investment
- Single point of failure (SPOF)

Horizontal Scaling
Adds extra identical boxes to server.

Issues:
- Requires Load balancer for managing connection.
- Distribution of work within the units becomes overhead.
- Additional investment.

The 4 Vs of Big Data

- Volume
  - How will your system deal with storage as data approaches ∞?
- Velocity
  - How will your system deal with fast responses as requests approach ∞?
- Variety
  - How will your system deal with different types of data as the types approach ∞?
- Veracity
  - How will your system ensure the accuracy of your Big Data is not crushed as the other 3 Vs approach ∞?

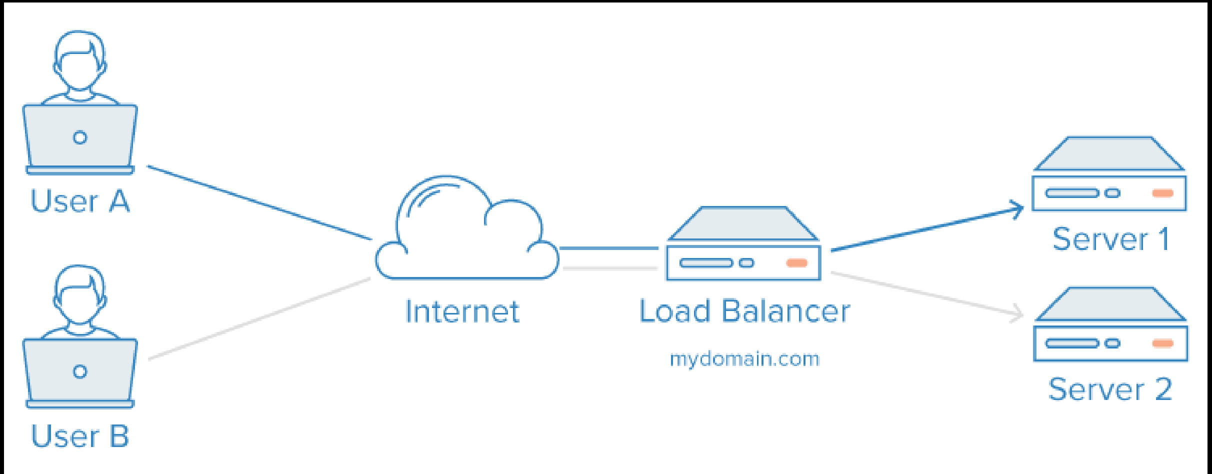
Volume

- Traditional approaches were to add more disks or (Storage area networks) SANs in data centers. This did not scale and the SANs were not geographically fault tolerant.

- Current approach is to use Cloud storage that is practically limitless and is geographically fault tolerant. Many cloud storage providers exist making it financially practical to move storage to the cloud.

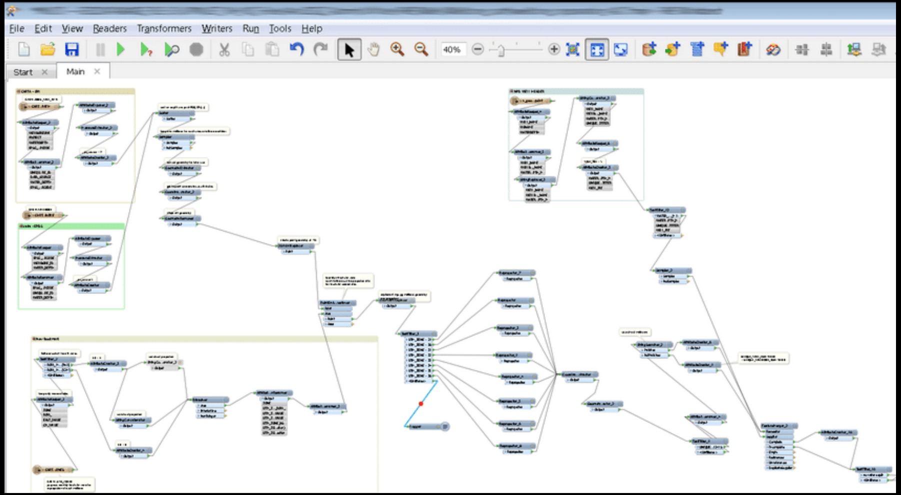
Velocity

- How do we deal with millions → infinite concurrent requests?
  - Traditionally we'd add load balancers and scale up the webserver.
  - Now cloud providers have geographic load balancers (that are themselves load balanced) and webservers that are horizontally scaled



Variety

- How do we deal with data changing and different data sets?
- Relational databases tend to crack with many changes (ultra structured so change is costly)
- Extraction, Transform, and Load (ETL) is cumbersome and expensive to run when data moves or is changed?
- Current solutions make data formatting easy and impose minimal structure (put this type of data in this folder).
- Data Lake → Folders ←→ Tables


File
Edit
View
Readers
Transformers
Writers
Run
Tools
Help
40%
Start
Main