Linear Regression!

Class 2: BDM3014

Course Logistics

Good question!!

Course Plan*

Week	Date	Topic	Activity
1	May 13, 2023	ML Intro, Types. Regression	Group Exercise (Risk and Applications)
2	May 20, 2023	Linear Regression, OLS	Practice Exercise, Group Activity (Excel OLS)
3	May 27, 2023	Gradient Descent	Indiv Assignment 1, Indiv Exercise (Compute GD)
4	Jun 3, 2023	Regularization	Group Exercise (Model building)
5	Jun 10, 2023	Logistic Regression	Group Assignment 2
6	Jun 17, 2023	Logistic Regression	Group Assignment 3
7	Jun 24, 2023	Goodness of Fit	Mid-Term

*Course Plan is tentative and subject to change. All assessments are in-class assessments. No extensions will be provided.

Course Plan*

Week	Date	Topic	Activity
9	Jul 8, 2023	Sensitivity Analysis, SVM	Group Exercise (GoF and Sensitivty Analysis)
10	Jul 15, 2023	SVM Kernels	Indiv Assignment 4, Group Exercise (SVM)
11	Jul 22, 2023	Decision Trees	Indiv Assignment 5, Group Exercise (DT)
12	Jul 29, 2023	Ensembling, Random Forests	Group Exercise (RF)
13	Aug 5, 2023	~~Boosting, Adaboost, XGBoost~~	Group Assignment 6
14	Aug 12, 2023	~~Naive Bayes, Nearest Neighbors~~	Final
15	Aug 19, 2023	~~Neural Networks~~	Group Presentation

*Course Plan is tentative and subject to change. All assessments are in-class assessments. No extensions will be provided.

Grading & Class Structure

Grades will be absolute

Any integrity violations will get you 0 on the assessment

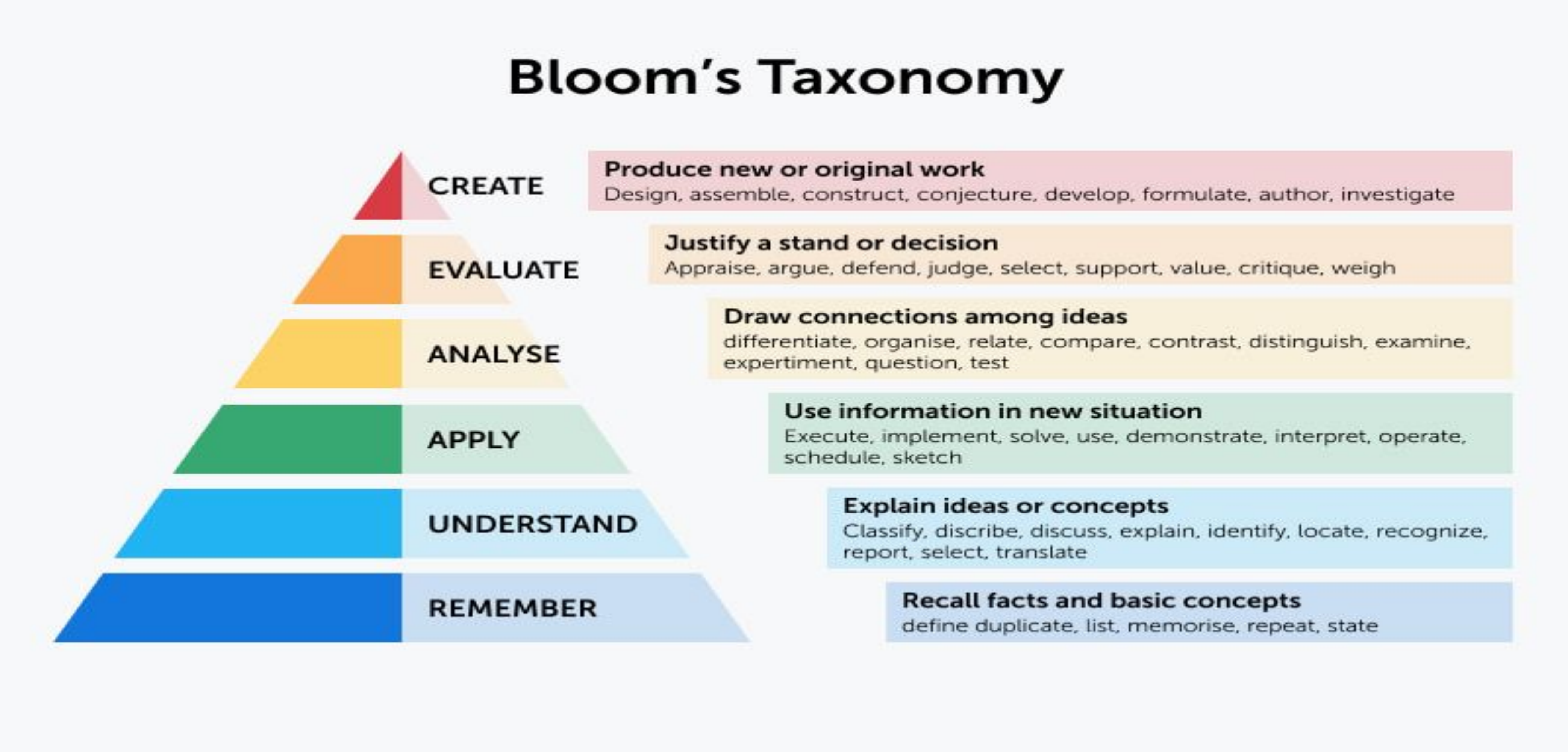
Bonus points can help you improve your grade

But you will still have to work! 😅

The class will usually be 1 - 1.5 hrs on instruction

The remaining 2-2.5 hrs will be on application and analysis of the material and in-class activities

Bloom’s Taxonomy



Slides, Learning Resources etc

Slides used here are only for reference, further asynchronous slides will be posted. However assessments will only be on the taught material (understanding, application, analysis, evaluation etc. - check's Bloom's taxonomy)

Various concepts will be explained using whiteboard, so feel free to take notes

Additional resources will be posted in Moodle.

Rubrics will be used to evaluate assignments. The assignments will not have bonus pts but upto 10% bonus pts may be awarded for extraordinary work as per the instructor's discretion.

Bonus pt assessments (usually a short quiz) may be held in the last weeks if a make-up test is needed for all the students

Class Decorum

No talking in the class even to yourself even if it is related to the subject matter except during breaks and group assignments/exercises.

Please step out for at least 15 minutes if you need to talk. If you are found talking you will be asked to step out and instruction to the class will be paused until you step out.

No language other than English in class even during breaks or group work. You need to get English and AI-ready. If found talking in another language you will be asked to leave the class until the next break and instruction to the class will be paused until you do.

No mobile phone ringing/vibrating in class. No recording in the class. If either of these happen you will be asked to leave the class until the next break and instruction will be paused until you do.

You can join in late in class, just come in and close the door. However attendance may be missed, you will miss on some content/instruction and you will not be allowed to join in late if an assessment has begun. Also, you will not get any make-up opportunities so be on time.

Academic Integrity

Academic Integrity is extremely important

- Cannot accept any submissions over email
- If you face any issues submitting the assignment, follow the below steps
  - Take screenshots
  - Send an email to bhavikg@quernscollege.ca with your complete assignment attached
  - Email itsupport@queenscollege.ca or moodlesupport@queenscollege.ca with these screenshots and get it resolved asap
  - Late submissions will incur a penalty so budget your time accordingly, genuine issues as verified by the it team or moodle team will be granted penalty-free extension
- Read the instructions and do not indulge in plagiarism. Anti-plagiarism softwares are very smart
- Please do not make me file ADRs 🙏 The risk is not worth the few extra pts you may get
- Not following instructions during test, even if you are not explicitly cheating, will still result in an ADR

Bonus Points!

- Total Bonus Points for the course will be capped at 20 (20%), total points capped at 100
- Attendance, Participation/Contribution: Upto 6 bonus points at the instructor’s discretion (includes participation in class, exercises, forums etc.). Will call it out in forums, exercise feedback or by email etc.
- Slide Corrections/Suggestions: Upto 4 bonus points at the instructor’s discretion, kindly email suggestions/corrections and save the response emails.
- Upto 10 bonus points distributed across assignments, exams and project.
- Bonus Point Assignment: Within 2 days of the last/W15 class each student will be required to submit how many bonus points they have and how they’d like their bonus points to be assigned. Failure to do so may lead to forfeiture of your bonus points. Set reminders now!

Group Exercises and Assignments!

- Points will be a combination of group performance and individual performance
- Individual responsibilities should be clearly outlined by the group in a MECE format in the assignment/exercise report
- The best submitted solutions at the instructor’s discretion may be shared with the class. You may get some bonus pts for knowledge sharing is your submission in a graded assessment is selected for sharing. Please write to me if you have any concerns regarding this.

Success Factors

No one can teach you all of Machine Learning

There's just too much to teach, so the idea of this course is to give you a primer into these fields so that you can explore them correctly.

To succeed in this course, apart from following instructions and having a strong sense of integrity, you will need

1. Basic Mathematical Skills
2. Problem Solving
3. Time Management
4. Group Working Skills
5. Practice

Things to Do!

This Week

Discussion Group

May 16, 12-2 pm, Rec Room

Get Your SIN/Service Canada on Campus

May 18, 9:30 - 3:30, Book Appt

Devant Career Cafe

May 17, 7-8 pm

Tangerine Tuesdays

May 16, 2-4 pm

Next Week

Outdoor Meditation

May 23, 12-2 pm, Backyard

Immigration Consultation

May 24, 10:30-1:30 pm

Devant Immigration Q&A

May 24, 7-8 pm

Tangerine Tuesdays

May 23, 2-4 pm

Today's Outline!

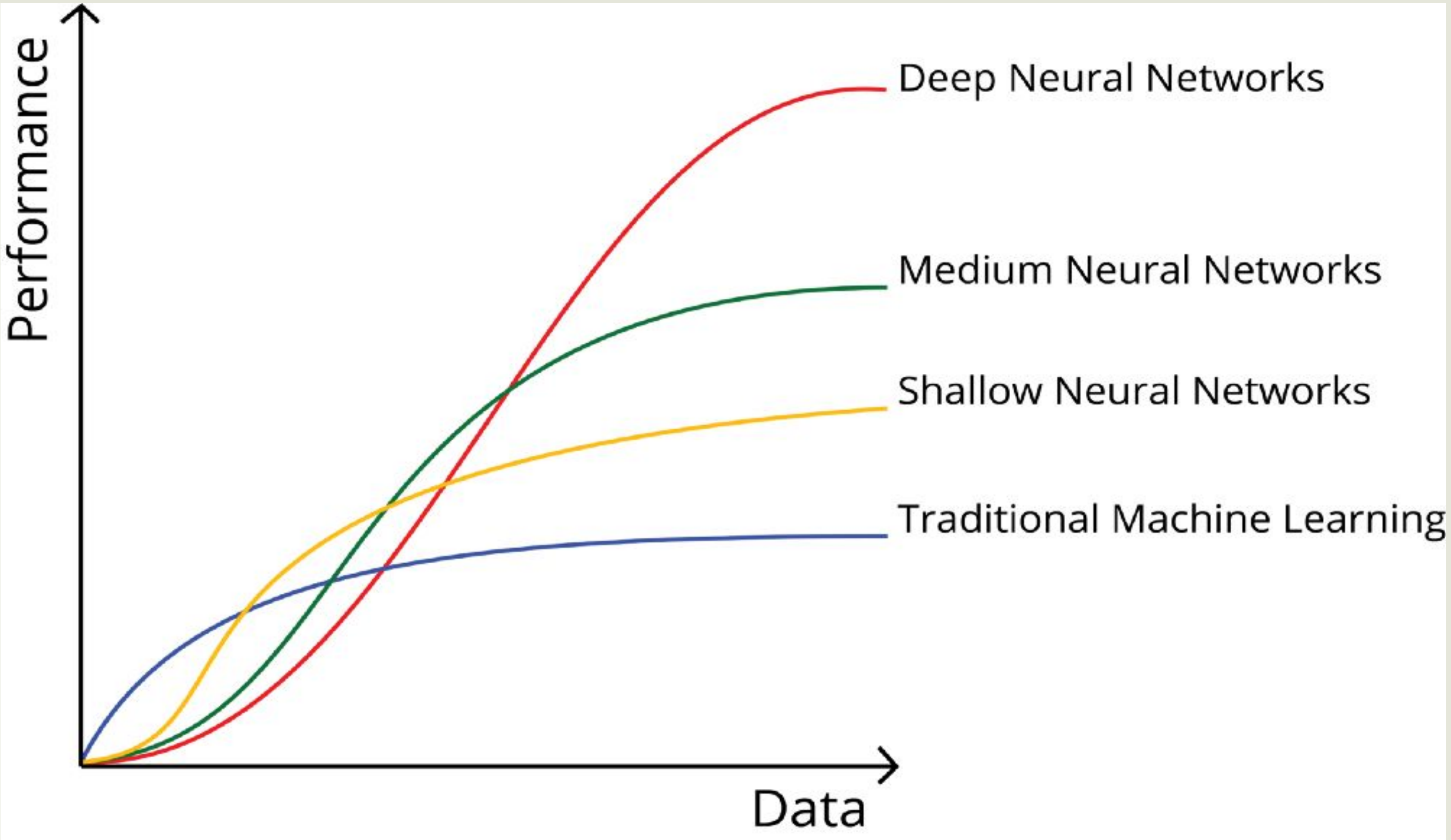
- You will be able to differentiate between supervised and unsupervised learning
- You will gain an appreciation for different types of problems such as classification and regression
- You will gain an introduction to classic Linear Regression
- You will understand where OLS-LR comes from

Review

What did we talk about last time?

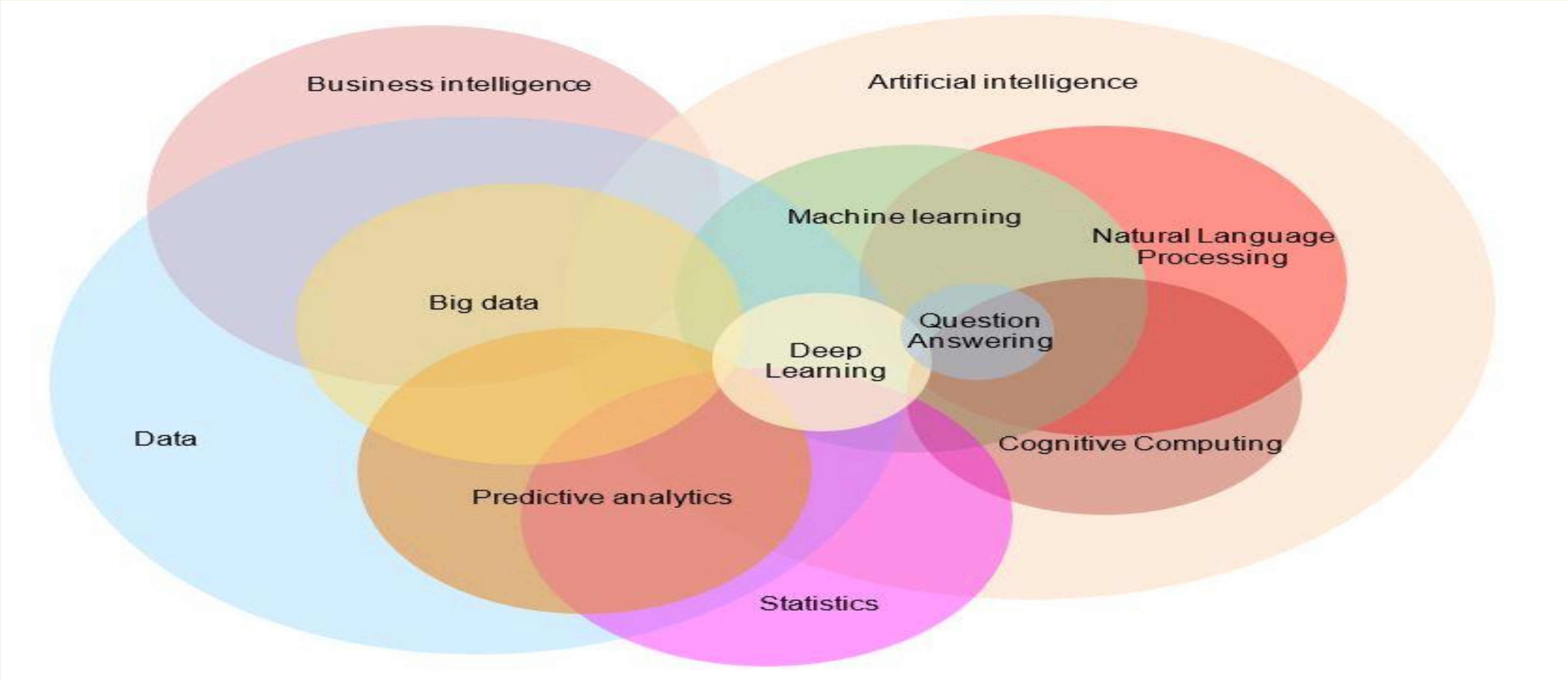
Intelligence, Learning & Challenges

- Intelligence
- Data Science
- Machine Learning
- Applications



"Machine learning is the next Internet" -- Tony Tether, former director, DARPA

AI vs ML vs DS


Business intelligence
Artificial intelligence
Machine learning
Natural Language Processing
Big data
Question Answering
Deep Learning
Data
Cognitive Computing
Predictive analytics
Statistics

Name	AI	Expert Sys?	ML?	DL?	NLP?	CV?	Supervised Learning?	Unsupervised Learning?	Reinforcement Learning?	Reactive Learning?	Limited Memory Learning?	Theory of Mind AI?	Self Aware AI?
Helmut Simonis's Sudoku Solver	Yes	Yes	No	No	No	No	No	No	No	Yes	No	No	No
Roomba j7+	Yes	No	Yes	No	No	No	No	No	Yes	No	Yes	No	No
Bitcoin PoW Algorithm	Yes	Yes	No	No	No	No	No	No	No	Yes	No	No	No
Siri	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No
Copilot	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No
DeepBlue	Yes	Yes	No	No	No	No	No	No	No	Yes	No	No	No
GauGAN2	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No	No
Pairwise Kidney Exchange	No	No	No	No	No	No	No	No	No	No	No	No	No
IBMProjectDebater	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	No
AlphaGO	Yes	No	Yes	Yes	No	No	No	No	Yes	No	Yes	No	No
ChatGPT	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No
AmazonGo	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No	No
Issac Asimov's Cosmic AC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes
Baxter's Xeelee	No	No	No	No	No	No	No	No	No	No	No	No	No

Types of AI

Reactive AI (Has no memory and predicts outputs based on the input that it receives)

eg. DeepBlue or Collaborative filtering

Limited Memory AI (comprised of supervised AI systems that derive knowledge from experimental data or real-life events) eg. Self-driving cars, ChatGPT

Theory of Mind AI (possess the understanding that people, creatures and objects in the world can have thoughts and emotions that affect their own behavior)

Self Aware AI (systems that can form representations about themselves, possess consciousness)

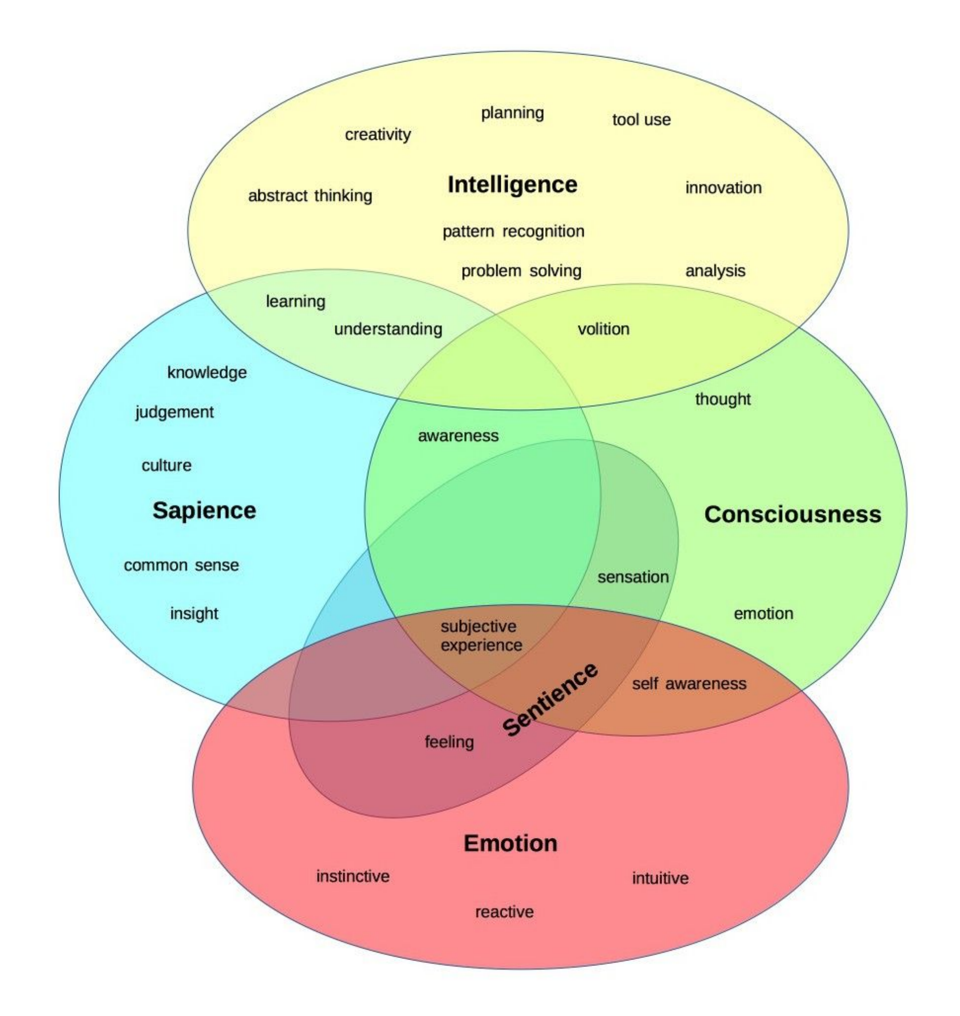
On a different spectrum

On similar lines there exist narrow/weak AI, general AI and super/strong AI

ANI or Artificial Narrow Intelligence is the ability to be intelligent in specific scenarios (eg. Image Classification, Sentiment Analysis)

AGI or Artificial General Intelligence is the ability of be intelligent in many different scenarios (eg. General Problem Solver)

ASI or Artificial Super Intelligence is the ability to be intelligent beyond humans



AI Agent Types

AI Agents can be specified using the PEAS (Performance Measure, Environment, Actuator and Sensor) framework

Task Environments can be fully or partially observable, episodic or sequential, static or dynamic, discrete or continuous, stochastic or deterministic and allowing for single or multi-agents

There are 4 categories in which agents can be classified. These are

Simple Reflex agents (Condition-Action Setup - If the car in front is braking, initiate braking)

Model based Reflex agent (Model based agent keeps track of the current complete state of a partially observable environment using a model and then applies condition-action setup)

Goal based agents (Goal helps defines a desirable situation, like getting closer to the destination)

Utility based agents (Utility allows for handling tradeoffs - for eg speed vs safety while driving)

Expert Systems

In artificial intelligence, an expert system is a computer system emulating the decision-making ability of a human expert.

Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if–then rules rather than through conventional procedural code.

The first expert systems were created in the 1970s and then proliferated in the 1980s. Expert systems were among the first truly successful forms of artificial intelligence (AI) software.

Applications include MYCIN, an early backward chaining expert system written in 1970s that used artificial intelligence to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight

Conventional System vs Expert System

Conventional System	Expert System
Knowledge and processing are combined in one unit.	Knowledge database and the processing mechanism are two separate components.
The programme does not make errors (Unless error in programming).	The Expert System may make a mistake.
The system is operational only when fully developed.	The expert system is optimized on an ongoing basis and can be launched with a small number of rules.
Step by step execution according to fixed algorithms is required.	Execution is done logically & heuristically.
It needs full information.	It can be functional with sufficient or insufficient information.

expert systems with applications

Turing Test

The Turing test, originally called the imitation game by Alan Turing in 1950,is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human. Turing proposed that a human evaluator would judge natural language conversations between a human and a machine designed to generate human-like responses. The evaluator would be aware that one of the two partners in conversation is a machine, and all participants would be separated from one another.

History of AI - Timeline

Decade	Summary
<1950s	Statistical methods are discovered and refined.
1950s	Pioneering machine learning research is conducted using simple algorithms.
1960s	Ray Solomonoff lays the foundations of a mathematical theory of AI, introducing universal Bayesian methods for inductive inference and prediction.
1970s	Marvin Minsky and Seymour Papert publish Perceptrons, demonstrating previously unrecognized limits of this feed-forward two-layered structure. This book is considered by some to mark the beginning of the AI winter of the 1970s, a failure of confidence and funding for AI due to pessimism in the promise of AI
1980s	First expert system shells and commercial applications. Rediscovery of backpropagation causes a resurgence in ML research. Lack of generality in expert systems and expenses required to maintain them lead to a second AI Winter (88-93)
1990s	Work on AI shifts from a knowledge-driven approach to a data-driven approach. Scientists begin creating programs for computers to analyze large amounts of data and draw conclusions – or "learn" – from the results. Support vector machines (SVMs) and recurrent neural networks (RNNs) become popular.
2000s	Support Vector Clustering and other Kernel methods and unsupervised machine learning methods become widespread.
2010s	Deep learning becomes feasible, leading to AI becoming integral to many widely used software services and applications.
2020s	Large Language Models including ChatGPT, Bard and AutoGPT lead to massive shifts in how questions are answered

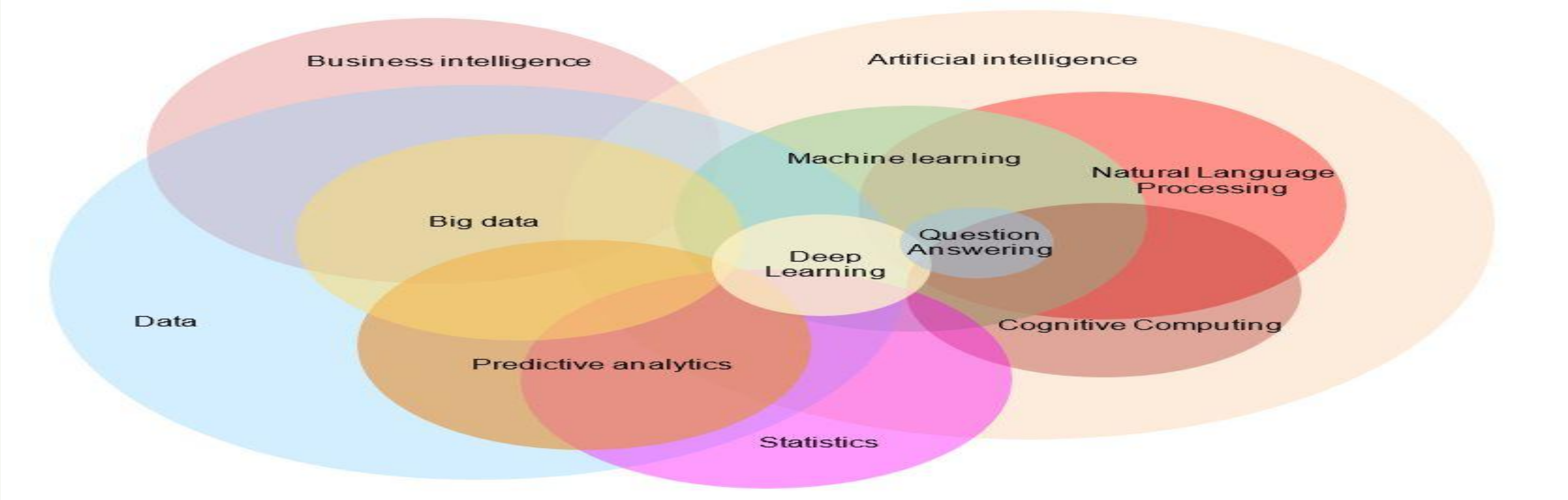
Based on Wikipedia

Natural Language Processing

GOAL OF NLP IS TO WRITE COMPUTER PROGRAMS THAT CAN INTERPRET NATURAL (HUMAN) LANGUAGES

NATURAL LANGUAGE PROCESSING

- **Natural language processing (NLP)** is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.



PROJECT DEBATER IN A DEBATE

https://www.youtube.com/watch?v=PkSzmnA1CQQ

MOSTLY SOLVED!! ☺

Spam detection

- Let's go to Agra! ✓
- Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON		ORG		LOC
Einstein	met with	UN	officials in	Princeton

SOMEWHAT SOLVED!! ☺

Sentiment analysis

- Best roast chicken in San Francisco!
- The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May
27
add

STILL VERY HARD! ☹

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

Housing prices rose

The S&P500 jumped

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

CHALLENGES WITH NLP

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤️

ambiguity

Red tape holds up new bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug’s Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

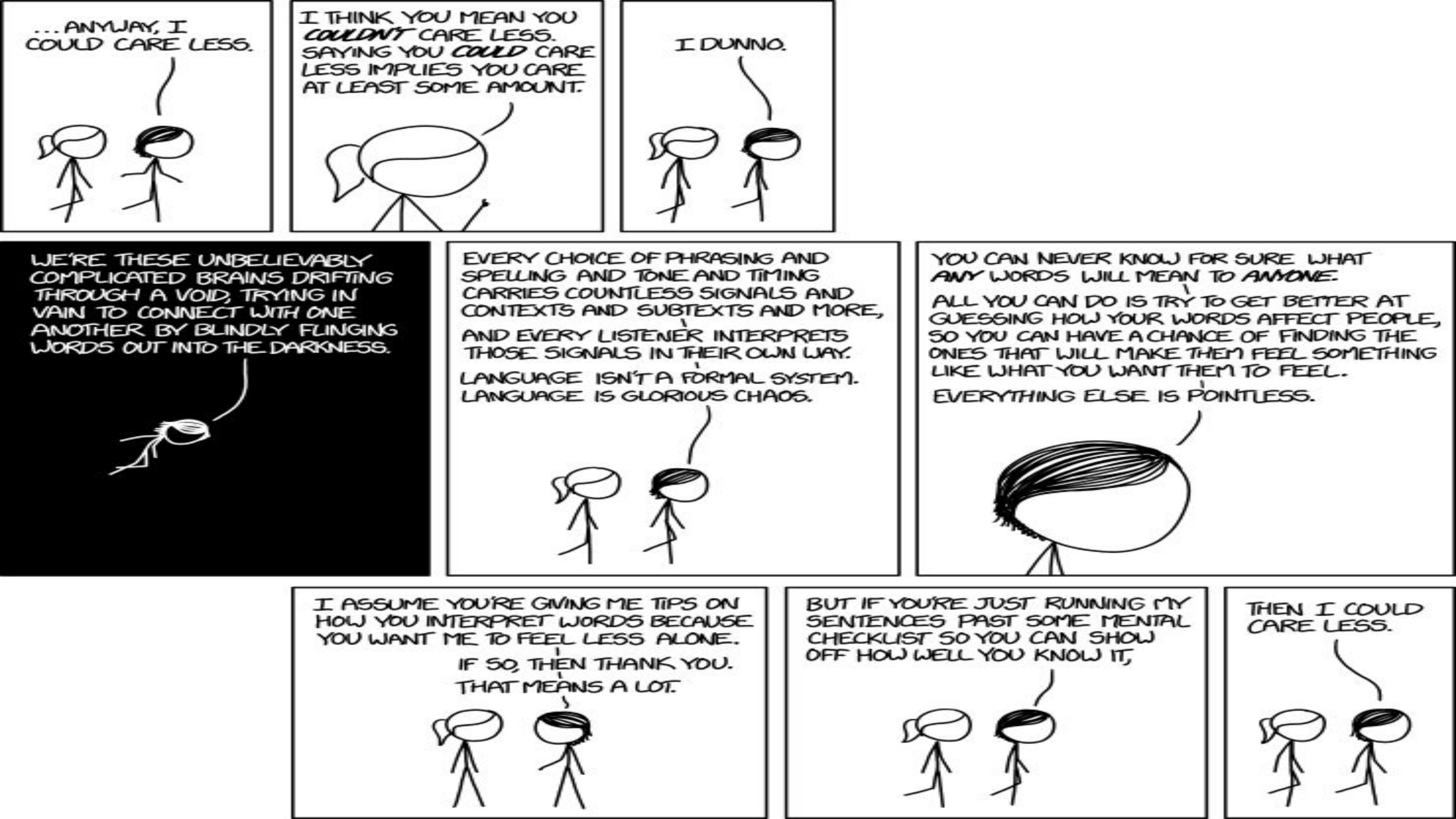
But that’s what makes it fun!

“80% of the data in the enterprise is unstructured, a lot of it is text

AutoGPT and Duplex Innovations

https://www.youtube.com/watch?v=ixtAXlKhmSo

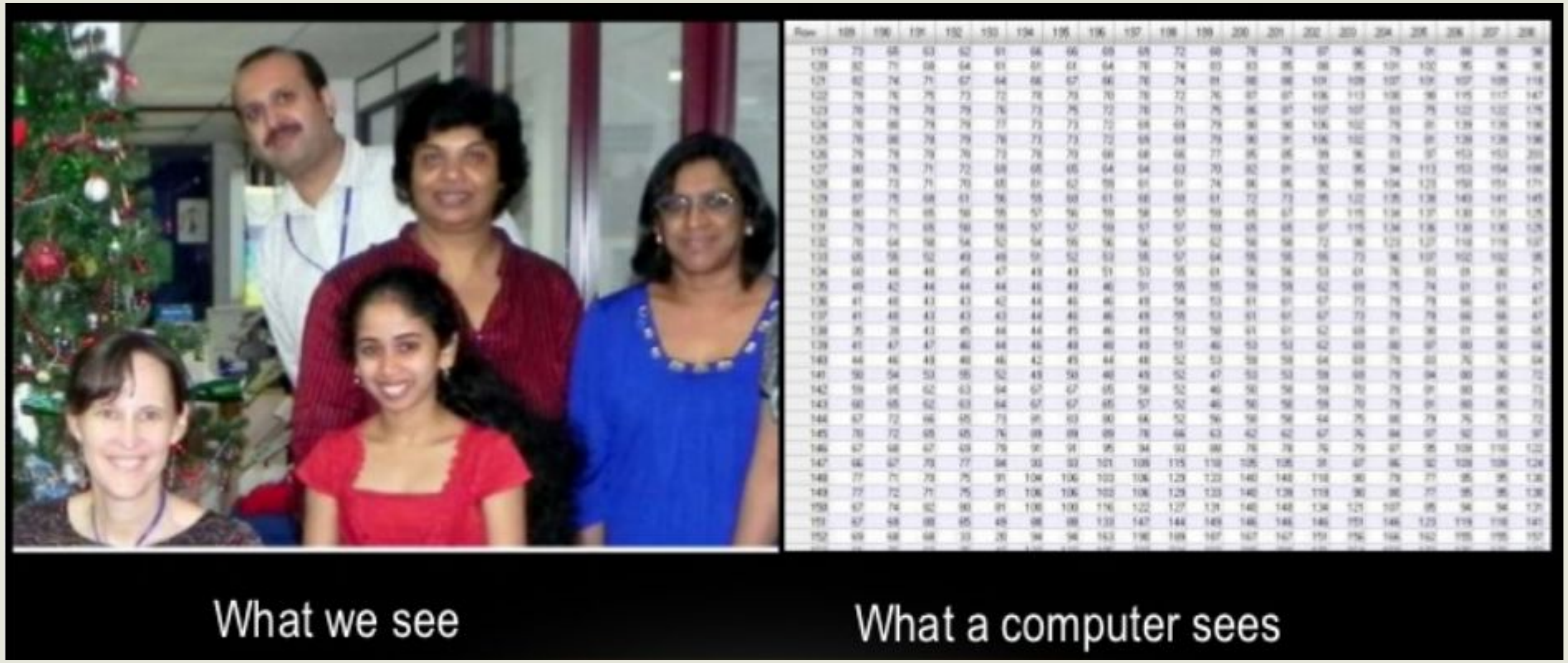
Google Video: https://www.youtube.com/watch?v=nTt8EEa53fA&t=488s


...ANYWAY, I COULD CARE LESS.
I THINK YOU MEAN YOU *COULDN'T* CARE LESS. SAYING YOU *COULD* CARE LESS IMPLIES YOU CARE AT LEAST SOME AMOUNT.
I DUNNO.
WE'RE THESE UNBELIEVABLY COMPLICATED BRAINS DRIFTING THROUGH A VOID, TRYING IN VAIN TO CONNECT WITH ONE ANOTHER BY BLINDLY FLINGING WORDS OUT INTO THE DARKNESS.
EVERY CHOICE OF PHRASING AND SPELLING AND TONE AND TIMING CARRIES COUNTLESS SIGNALS AND CONTEXTS AND SUBTEXTS AND MORE,
AND EVERY LISTENER INTERPRETS THOSE SIGNALS IN THEIR OWN WAY.
LANGUAGE ISN'T A FORMAL SYSTEM. LANGUAGE IS GLORIOUS CHAOS.
YOU CAN NEVER KNOW FOR SURE WHAT *ANY* WORDS WILL MEAN TO *ANYONE*.
ALL YOU CAN DO IS TRY TO GET BETTER AT GUESSING HOW YOUR WORDS AFFECT PEOPLE, SO YOU CAN HAVE A CHANCE OF FINDING THE ONES THAT WILL MAKE THEM FEEL SOMETHING LIKE WHAT YOU WANT THEM TO FEEL.
EVERYTHING ELSE IS POINTLESS.
I ASSUME YOU'RE GIVING ME TIPS ON HOW YOU INTERPRET WORDS BECAUSE YOU WANT ME TO FEEL LESS ALONE.
IF SO, THEN THANK YOU.
THAT MEANS A LOT.
BUT IF YOU'RE JUST RUNNING MY SENTENCES PAST SOME MENTAL CHECKLIST SO YOU CAN SHOW OFF HOW WELL YOU KNOW IT,
THEN I COULD CARE LESS.

COMPUTER VISION

GOAL OF COMPUTER VISION IS TO WRITE COMPUTER PROGRAMS THAT CAN INTERPRET IMAGES

CHALLENGES WITH COMPUTER VISION



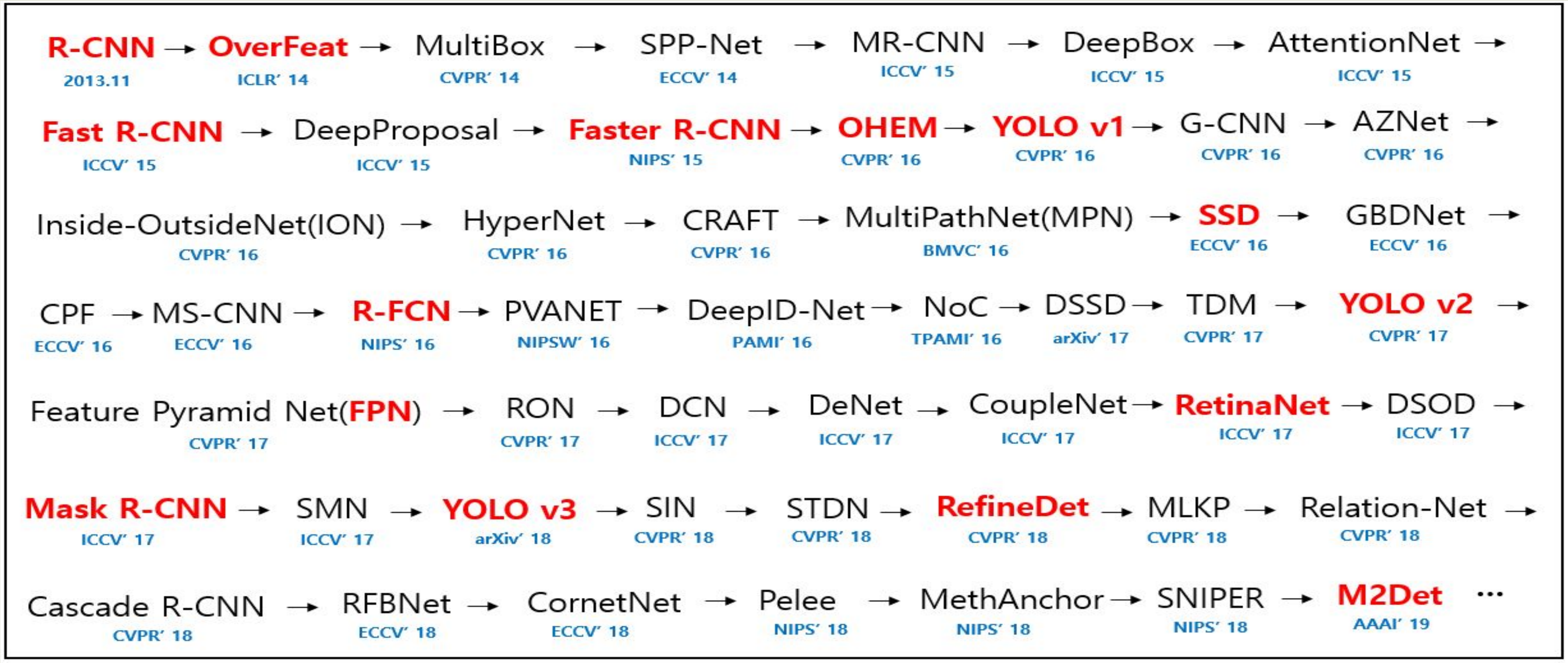
HUMAN VISION

- Vision is an amazing feat of natural intelligence
  - Visual cortex occupies about 50% of Macaque brain
  - More human brain devoted to vision than anything else
  - About 30,000 visual categories, learnt by 6 years (~13.5/day)
- Machines can’t replicate Human Image Recognition yet, they do not possess our ability to recognize distorted images
- Drawing to Images

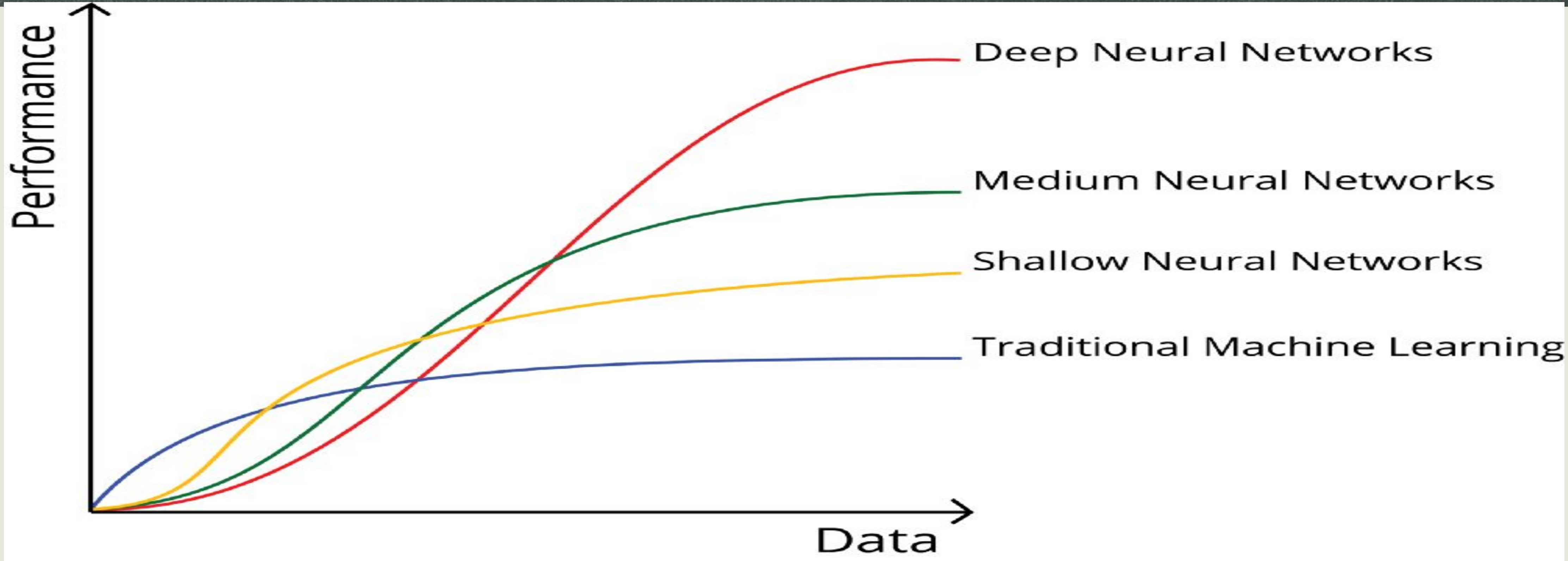
APPLICATIONS

- OCR
- Face Detection
- Smile Detection
- Face Recognition
- Medical Imaging
- Self-driving cars
- Sports Technologies
- Gaming Technologies

ADVANCES IN OBJECT DETECTION



DEEP LEARNING



HUMAN WINS IBM DEBATER

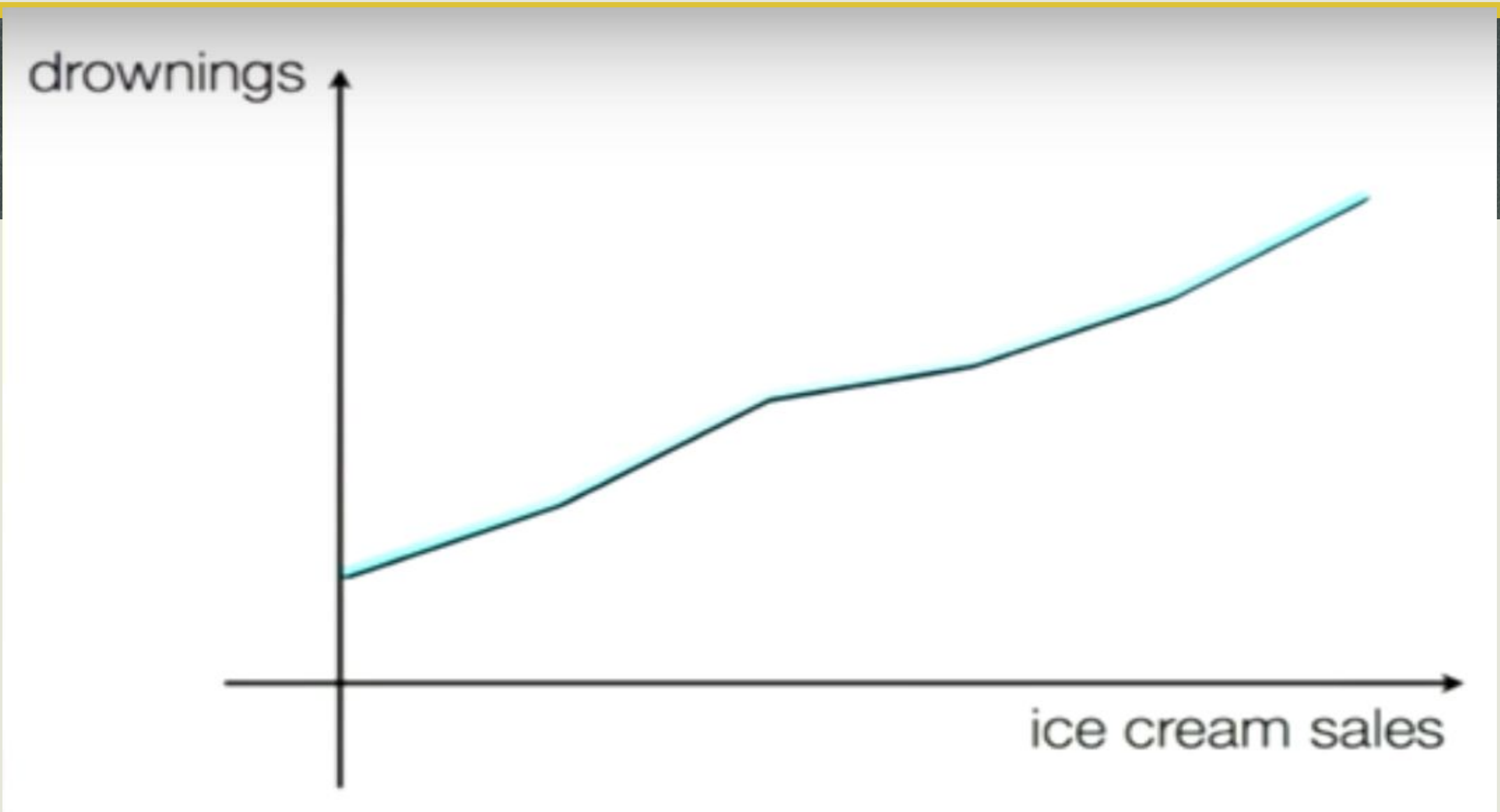
RECENT HEADLINES IN BLOOMBERG

INTERESTING DEMOS

- Language Translation
- Self driving Car
- Retail Store
- Duplex

DATA

- Data is insightful
- Data is revealing
- Data is difficult to manage at scale
- Data is tricky to interpret
  - HHT and HTH


drownings
ice cream sales

MARRIED MEN HAVE HIGHER LIFE EXPECTANCY THAN SINGLE MEN

KIDS WHO HAVE HIGH SELF ESTEEM GET GOOD GRADES

STATISTICAL THINKING WILL ONE DAY BE AS NECESSARY FOR EFFICIENT CITIZENSHIP AS THE ABILITY TO READ AND WRITE.

– H G WELLS

LURKING VARIABLE PROBLEM

- Also known as Simpson’s Paradox

- Smokers have higher life expectancy than non-smokers

CROOKS ALREADY KNOW HOW TO LIE WITH STATISTICS, HONEST MEN MUST LEARN THEM FOR SELF DEFENCE
- D HUFF

AI IS POWERFUL

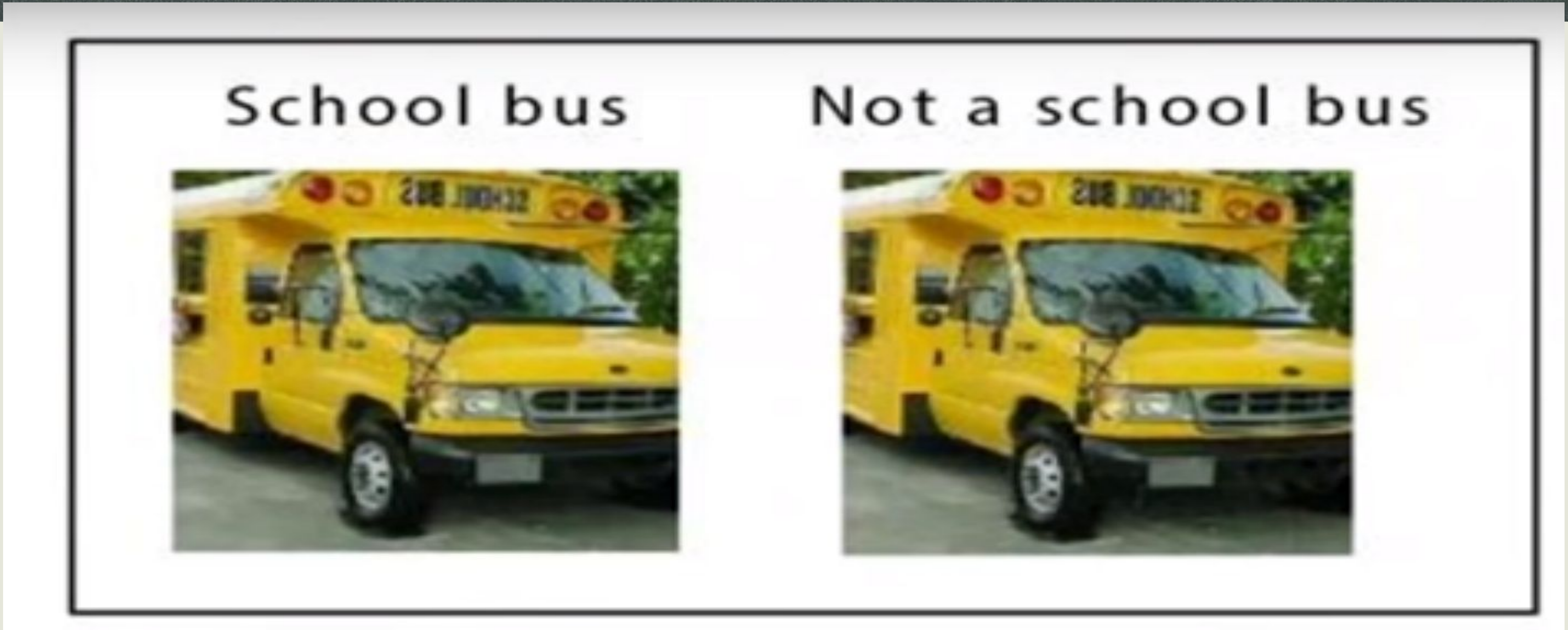
AI IS COMPLEX

DEEP NEURAL NETWORKS ARE BASED ON THE HUMAN BRAIN

THE AVERAGE HUMAN BRAIN MANAGES ABOUT 30 QUADRILLION COMPUTATIONS PER SECOND

A MODEST OBJECT DETECTION ALGORITHM DOES 3 BILLION COMPUTATIONS PER IMAGE

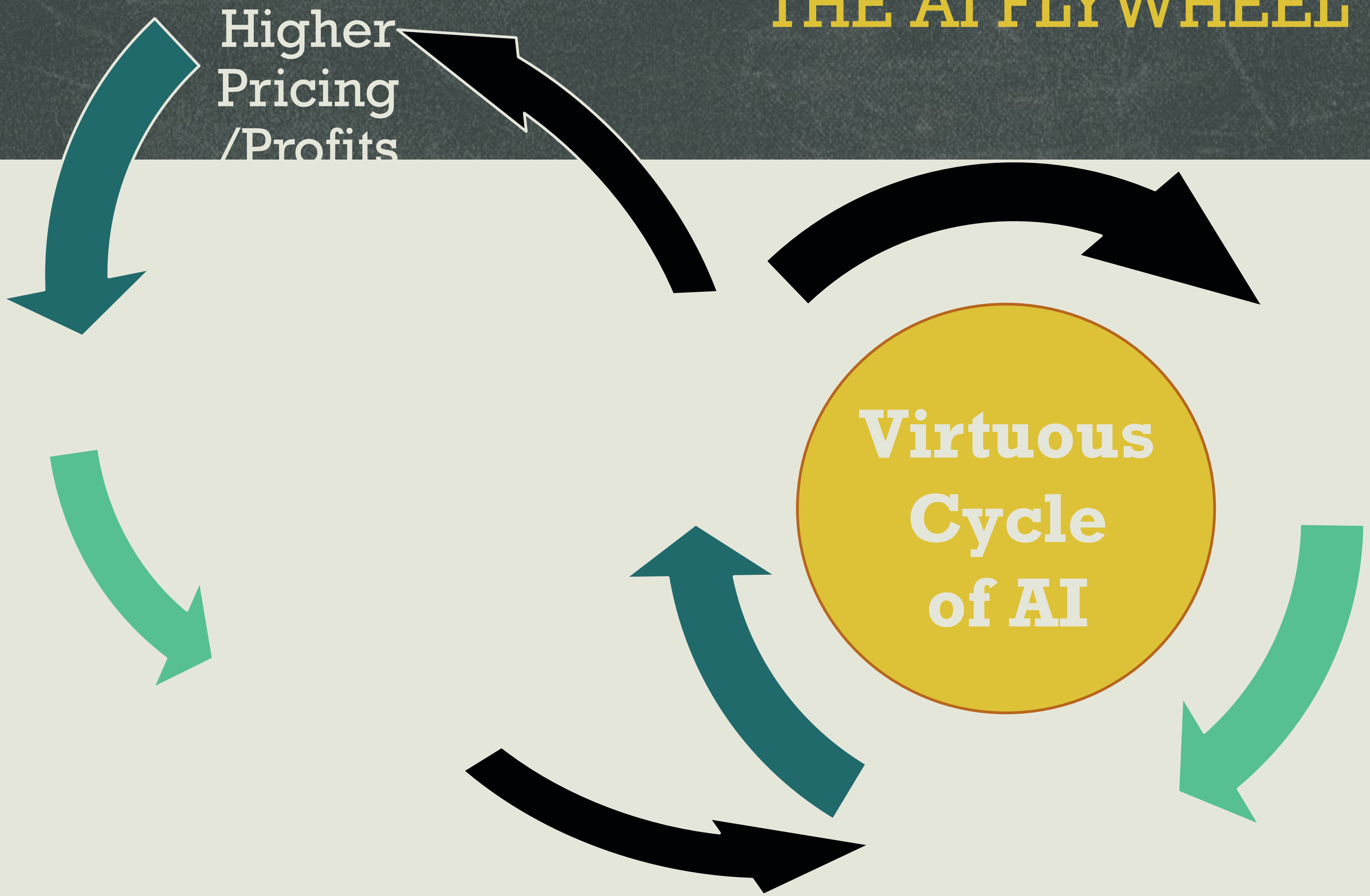
AI IS STUPID



AI IS EVOLVING LIKE CRAZY

ULMFiT (Jan-18) → ELMo (Mar-18) → BERT (Oct-18) → Transformer – XL (Jan-19) → MT-DNN (Feb-19)

THE AI FLYWHEEL



HOW TO LEARN AI?

- AI is extremely detailed and complex
- AI is strongly rooted in Mathematics, Statistics and Computing
- Work on your problem solving and analytical skills
- Learning AI is almost a life-long journey
- Read, Implement, Question, Explain as a framework works well with AI
- Lots of self-study and self-work will help you understand how and what to apply where and when

HOW TO BE A GOOD AI Engineer?

Comparison to logistic regression [edit]

Discriminant function analysis is very similar to logistic regression, and both can be used to answer the same research questions.[9] Logistic regression does not have as many assumptions and restrictions as discriminant analysis. However, when discriminant analysis' assumptions are met, it is more powerful than logistic regression.[25] Unlike logistic regression, discriminant analysis can be used with small sample sizes. It has been shown that when sample sizes are equal, and homogeneity of variance/covariance holds, discriminant analysis is more accurate.[7] With all this being considered, logistic regression has become the common choice, since the assumptions of discriminant analysis are rarely met.[8][7]

HOW TO GET A JOB AS IN AI?

- What kind of people apply?
- What I look for in Resumes?
- What I check for in Interviews?
- How to acquire the required skillsets?
- What does a day look like for a data scientist?
- How to be successful in a job?

Let’s brush up a few things

Sure, it’s been a while!

Some Maths Concepts

- Log and e
- Linear and Non-linear
- Lines and Planes
- Slopes and Derivatives
- Vectors and Matrices

Some Statistics Concepts

- Moments, Central Tendency, Means
- Variance, Standard Deviation
- Central Limit Theorem & Normal Distribution
- Uniform Distribution, Binomial, Multinomial, Poisson Distribution
- Hypothesis Testing & Statistical Significance
- Z-test vs T-Test

Some Computing Concepts

- Lists, Sets, Dictionaries, Trees
- Loops and Conditions
- Functions and Recursion
- Time and Space Complexity
- Version Control and Deployment Process

Data Types

Hmm, what are these types?

Types of Data – There are four types

- **Nominal (categorical)**.

  The values that are often represented as strings or it could be numbers also.
  Example: customer's name is a nominal and Student id is also nominal
  In general, a few statistics can be computed on nominal features. Examples are the chi-square statistic (2) and the mode (most common feature value). For example, one can find the most common first name among customers. The only possible transformation on the data is comparison. For example, we can check whether our customer's name is John or not. Nominal feature values are often presented in a set format.

- **Ordinal**. Ordinal features lay data on an ordinal scale.

The feature values have an intrinsic order to them.

In our example, Money Spent is an ordinal feature because a High value for Money Spent is more than a Low one.

- **Interval**. In interval features, in addition to their intrinsic ordering, differences are meaningful whereas ratios are meaningless. For interval features, addition and subtraction are allowed, whereas multiplications and division are not. Consider two time readings: 6:16 PM and 3:08 PM. The difference between these two time readings is meaningful (3 hours and 8 minutes); however, there is no meaning to 6:16 PM = 3:08 PM * 2.

- **Ratio**. Ratio features, as the name suggests, add the additional properties
- of multiplication and division. An individual’s income is an
- example of a ratio feature where not only diferences and additions
- are meaningful but ratios also have meaning (e.g., an individual’s
- income can be twice as much as John’s income).

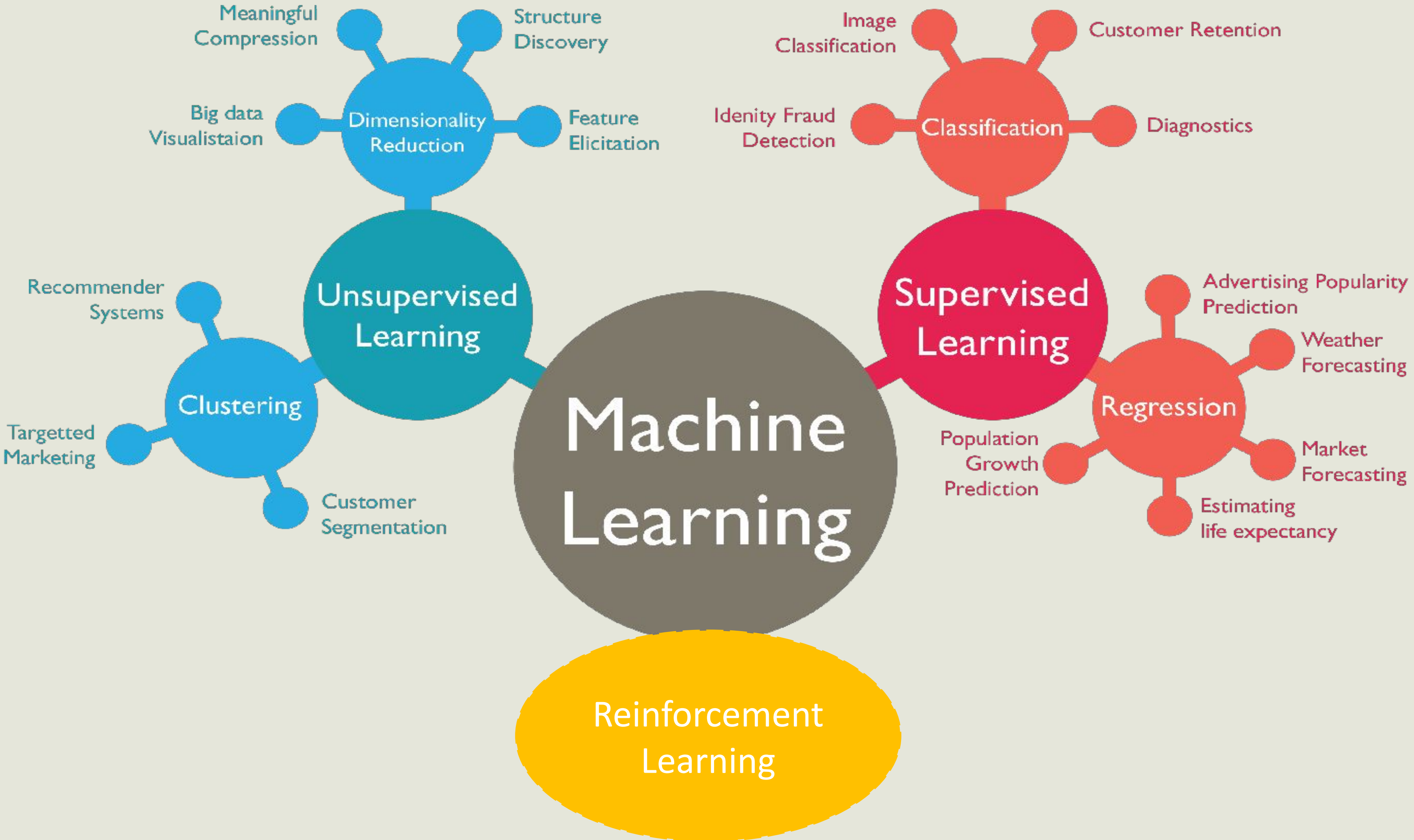
Sample Dataset Twitter Users

Activity	*Date Joined*	*Number of Followers*	*Verified Account?*	*Has Profile Picture?*
High	2015	50	FALSE	no
High	2013	300	TRUE	no
Average	2011	860000	FALSE	yes
Low	2012	96	FALSE	yes
High	2008	8,000	FALSE	yes
Average	2009	5	TRUE	no
Very High	2010	650,000	TRUE	yes
Low	2010	95	FALSE	no
Average	2011	70	FALSE	yes
Very High	2013	80,000	FALSE	yes
Low	2014	70	TRUE	yes
Average	2013	900	TRUE	yes
High	2011	7500	FALSE	yes
Low	2010	910	TRUE	no
Ordinal	**Interval**	**Ratio**	**Nominal**	**Nominal**

Supervised Learning

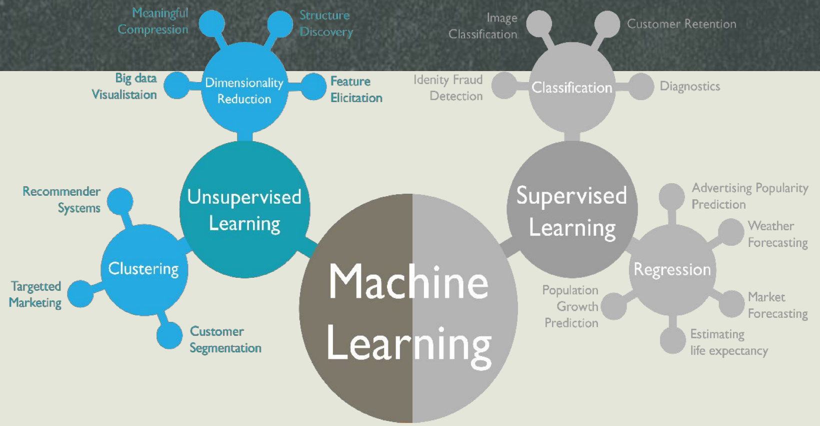
How do we supervise it?

Supervised Learning!



All of these distinctions have some arbitrariness. They change back and forth over time. The methods all sit on a spectrum and can be adapted. If you can imagine it, it exists! In the end, it’s all 1s & 0s!

Machine Learning – Supervised vs **Unsupervised learning**



The machine is learning without any supervision

The algorithm requires less from the data;

It doesn’t have to be preprocessed (labelled)

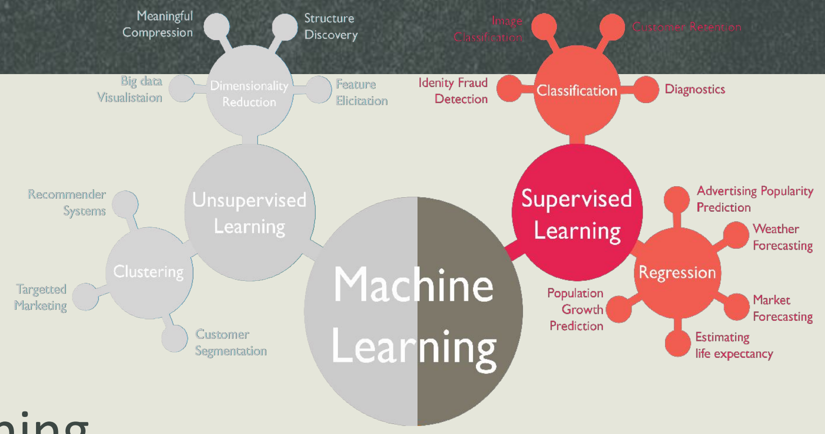
The machine handles the data on its own by sorting and filtering it to tease out underlying structures/patterns in the data

Typically we divide the unsupervised learning into two (main) areas Clustering/Association & Dimensional Reduction

<u>Unsupervised methods</u> include looking for similarities/differences in data

We typically don’t know what outputs to expect

Machine Learning – **Supervised** vs Unsupervised learning



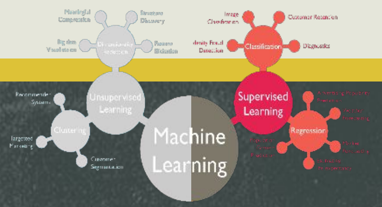
The machine is learning but requires supervision

That supervision is in the form of “labelled” data

These labelled data sets are used to train (supervise) the learning

In other words, we start by telling the machine **<u>what</u>** it is seeing using these labels, and as it learns, it will hopefully develop to the point where it no longer needs our supervision

Typically we divide the supervised learning into two areas Classification & Regression



Machine Learning – **Supervised** vs Unsupervised learning

Typically we divide the supervised learning into two areas Classification & Regression properties:

-less complicated computationally but upfront labelling work is higher

-the outputs are usually expected class members

-usually much easier to explain the results (less “blackbox”)

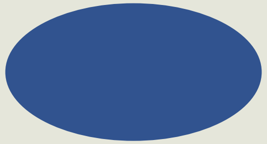
In either case it’s just 1s & 0s!

So where is the ‘magic’, how does it work?

Machine Learning – Supervised learning

Data has an implicit governing relationship that maps inputs (features), x_i, to outputs, y_i. Our job is to guess the relationship and remember it using trial and error.

e.g. what colour is this:



Red? ⟶ **NO!**

Green? ⟶ **NO!** (+not Red)

Blue? ⟶ **YES!** (+not Green & not Red)

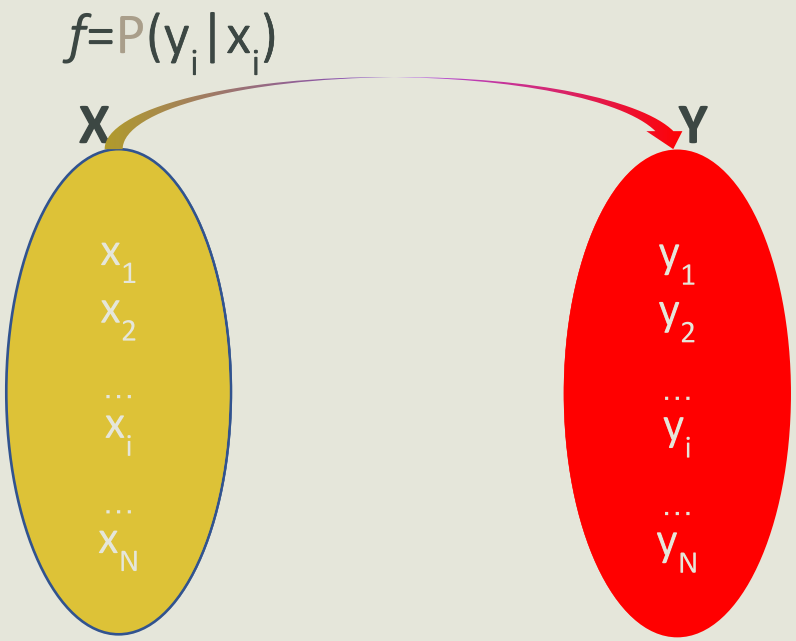
We do this many times, and build up a memory <u>usually</u> based on probability (joint/conditional).

This is no different than what we do.

Machine Learning – Supervised learning

In effect we are mapping inputs to outputs via some function, *f*:

D,R={x_iϵ**X** & y_iϵ**Y**| **X** x **Y** = $\mathbb{R}^2$}



But how do we know *f* is any good, or that this f is the best?

Machine Learning – Supervised learning

But how do we know *f* is any good, or that this f is the best?

Typically we define a cost/loss/objective function and measure a 'price' paid in doing the fitting/predicting. We usually want to optimize that function.

There are two main ways to deal with this price.

1. Empirically: we assume the data is good, and we want to fit it nicely

But is this the best approach?

Machine Learning – Supervised learning

Goal: minimize loss

1. Empirically: we assume the data is good, and we want to fit it nicely

-but variance is high (model is complex)

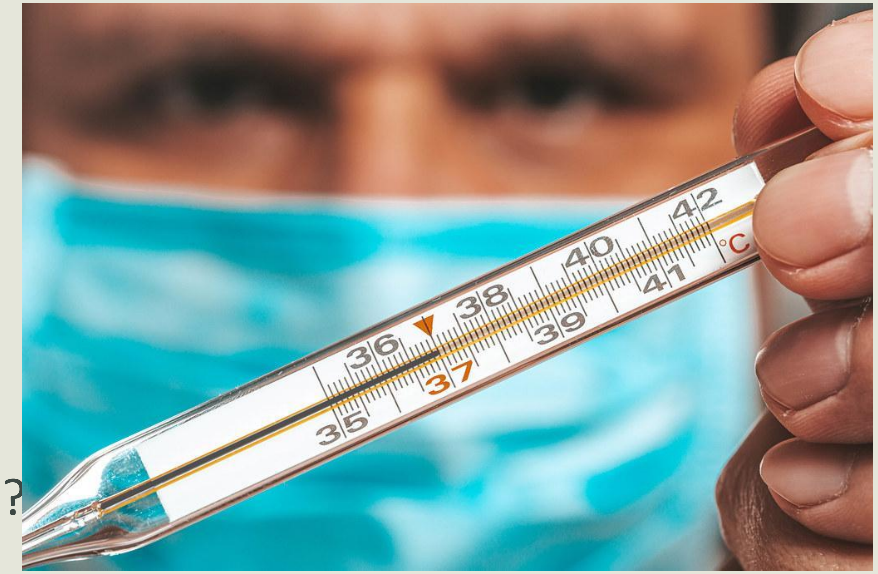
What if the data is wrong due to:

Precision? 37? 37.1? 36.9? 36.899?...

Accuracy? 36.85? 36.85? 36.85?

Is real life continuous/
discrete?

What if we are out of our domain and want our model to predict?



Machine Learning – Supervised learning

Goal: minimize loss

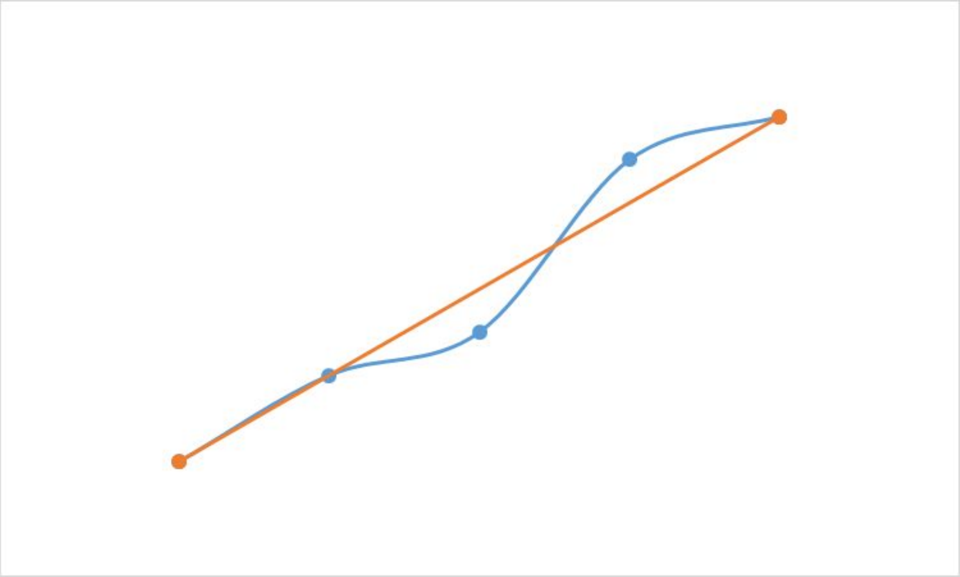
2. Model-wise: we assume the model is good (can predict, smooth, cts.)

But bias is high (systematically shifts outputs/does not fit the data well)

What if the data is great?!

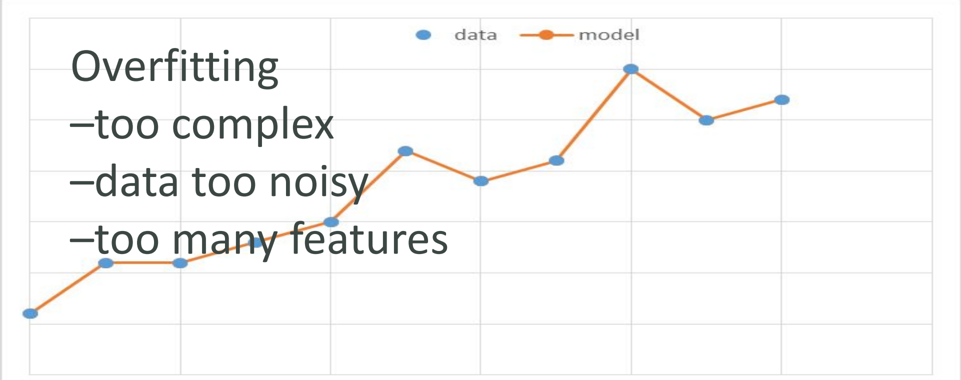
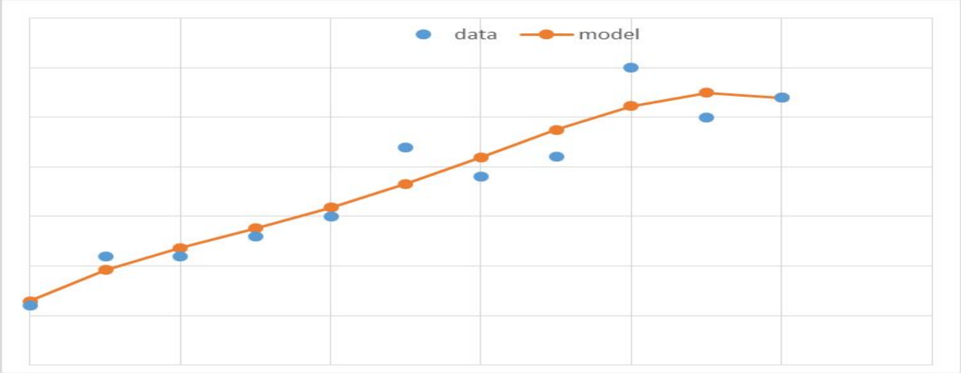
Why did we get it in the first place?

What if the model is too simple?



Machine Learning – Supervised learning

Goal: minimize loss

<table>
<tr><td colspan="2" rowspan="2">DATA FITTING</td><td colspan="2">BIAS (shift)</td></tr>
<tr><td>High -doesn’t respect data, oversimplifies model</td><td>Low</td></tr>
<tr><td rowspan="2">VARIANCE (spread)</td><td>High –too much emphasis on data, poor prediction</td><td></td><td>
Overfitting
–too complex
–data too noisy
–too many features</td></tr>
<tr><td>Low</td><td>
Underfitting
–didn’t learn
–not enough data
–under sophisticated model</td><td></td></tr>
</table>

Machine Learning – Supervised learning

Goal: minimize loss

We want a compromise!

3. Structurally:

We want to design a loss function that attaches importance to low bias and low variance.

In other words we want to control the trade off:

Good Loss $\equiv$ variance + bias

(we want an overall smaller total)

Machine Learning – Supervised learning

Goal: minimize loss by compromising (balance fitting)

3. Structurally:

Loss ~ Exact Interpolation vs. Roughness

So we create a penalty function, *J*,

$$J = RMSE + \lambda \, Regularization(weights)$$

$\lambda = 0$? We penalize data infidelity ⟶ end up loyal to data (overfit)

$\lambda >> 0$? We penalize roughness ⟶ smooth fit (underfit)

So how do we pick the right λ?

Machine Learning – Supervised learning

Goal: minimize loss by compromising (balance fitting)

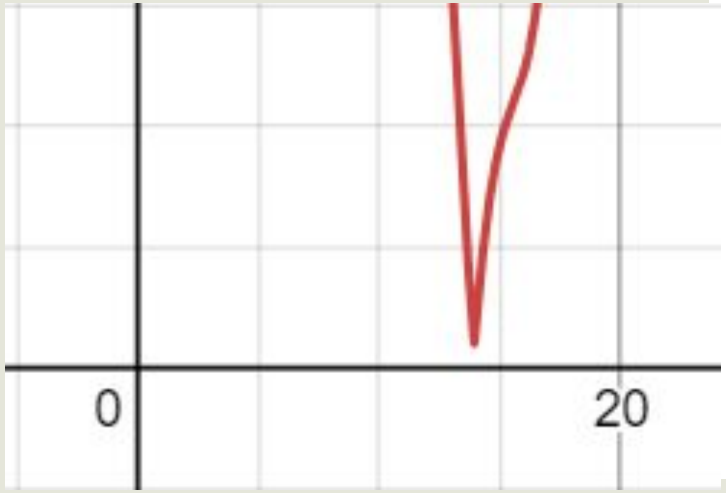
$$J(\lambda) = \text{RMSE} + \lambda\text{Regularization(weights)}$$

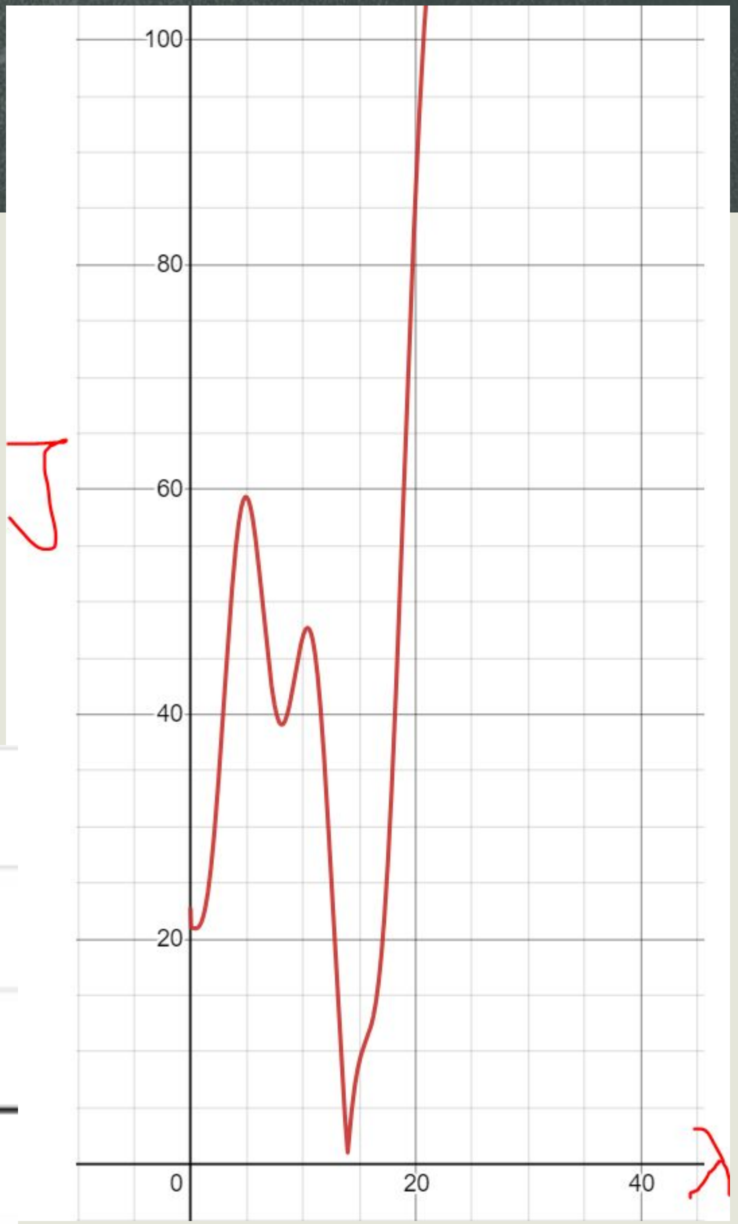
$\lambda = 0 \longrightarrow J = \text{RMSE}$

$\lambda >> 0 \longrightarrow J = R_o$

We call this GCV:

Generalized Cross Validation





Classification & Regression

Classify and Regress, eh?

Machine Learning –Supervised learning –Regression vs. Classification

Mini Recap:

Unsupervised learning occurs with simpler data requirements and goes by way of sorting/filtering data on its own → underlying patterns

Supervised learning occurs with intervention in the form of labelled data and goes by way of comparing predictions against known data (training) using a balanced loss function in an iterative manner → underlying patterns

Each of these has their uses –though there is much overlap.

Two such applications in Supervised learning are:

Regression & classification

Machine Learning –Supervised learning –Regression vs. **Classification**

Regression & Classification are methods used to achieve a common goal:

-to discover the underlying structure that separates data in order to use the same to predict where new data fits

The difference is in the interpretation of the results:

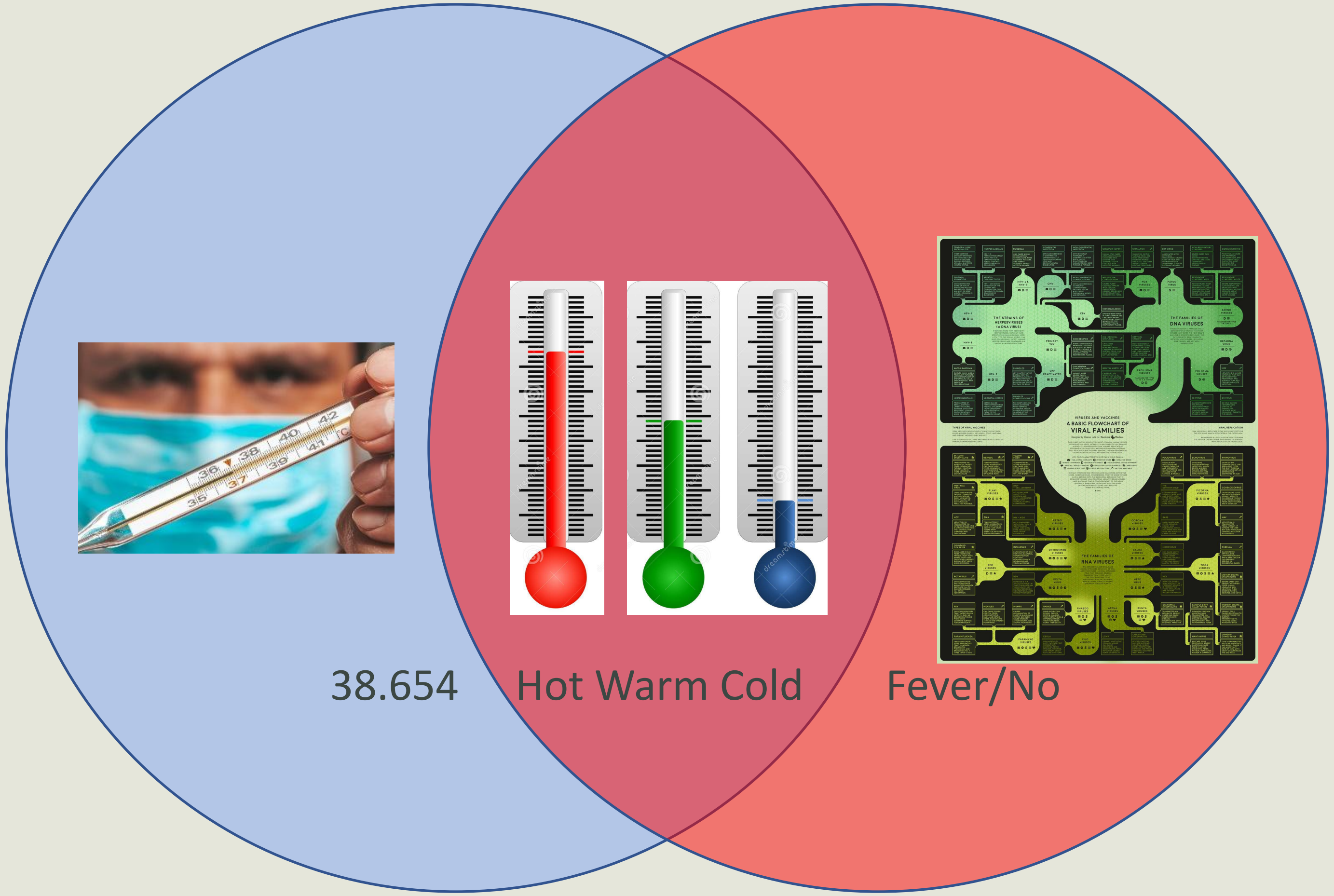
Regression deals with input data that is continuous in nature

Classification deals with input data that is discrete

But there can be overlap, and there is an arbitrariness in labelling.

Machine Learning –Supervised learning –Regression vs. **Classification**

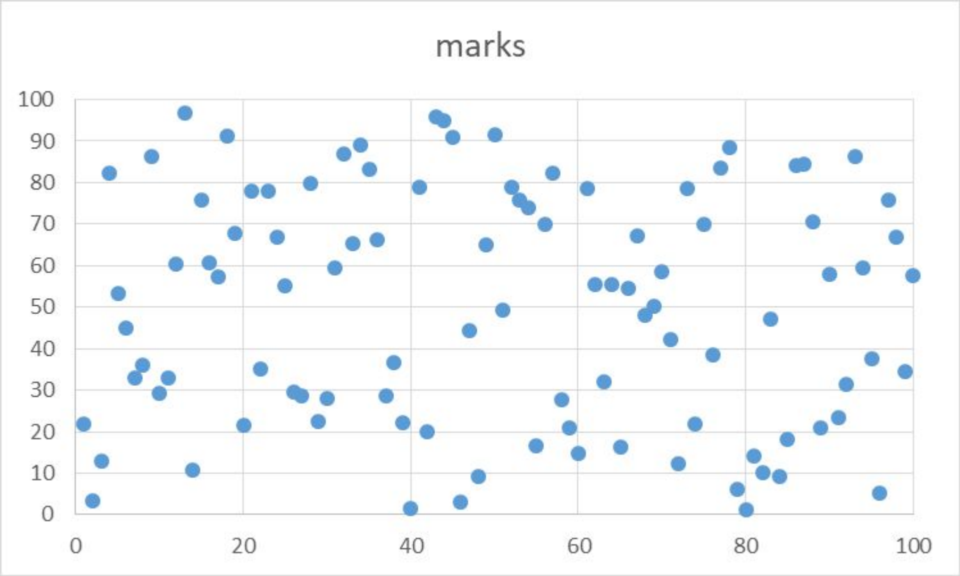
Regression & Classification –mostly a function of granularity

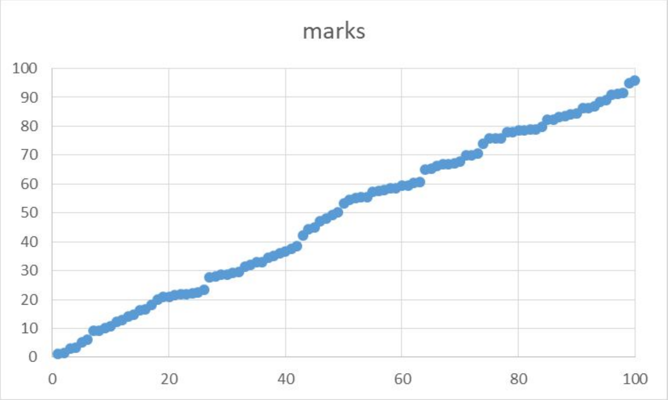


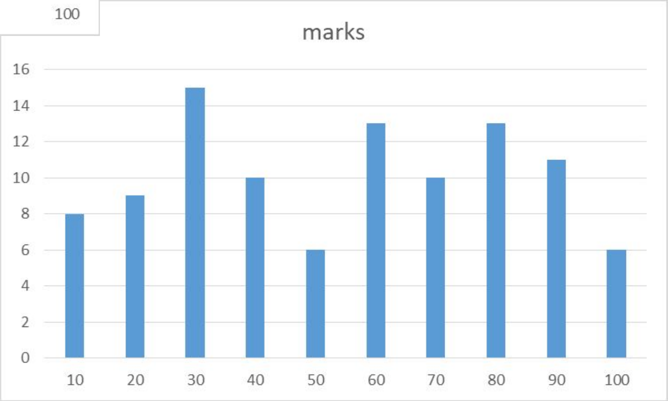
Machine Learning –Supervised learning –Regression vs. **Classification**

Regression & Classification –mostly a function of granularity

Marks distribution:



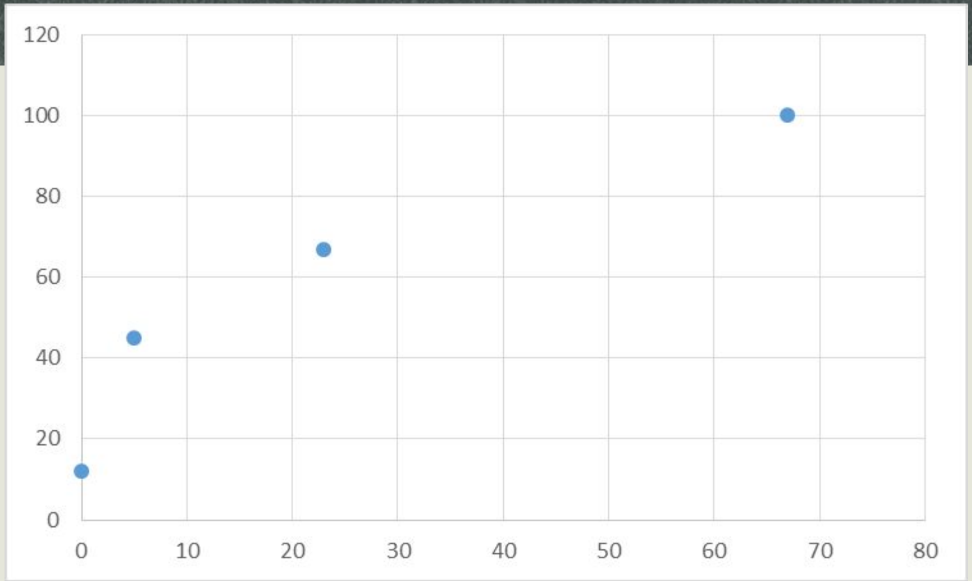


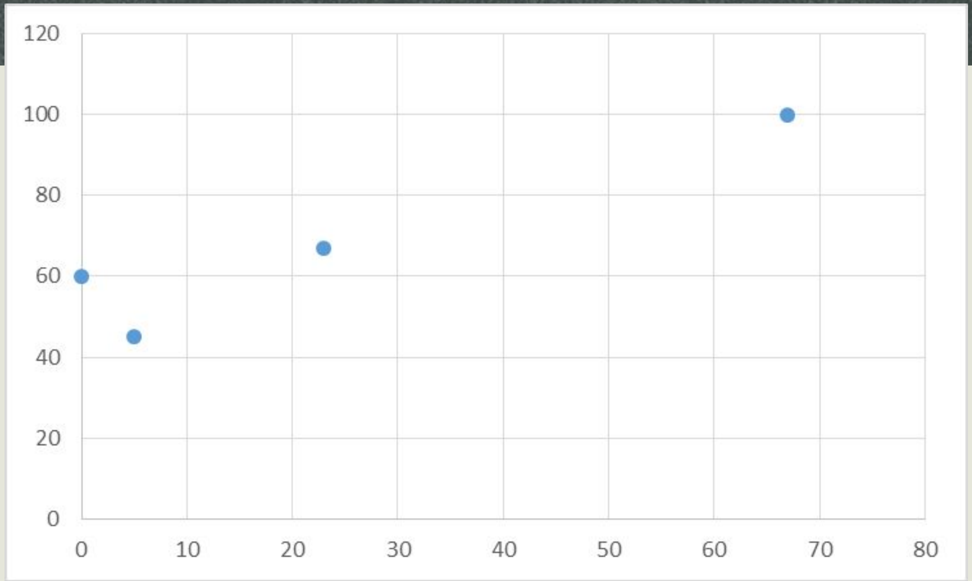


Linear Regression

Why linear?

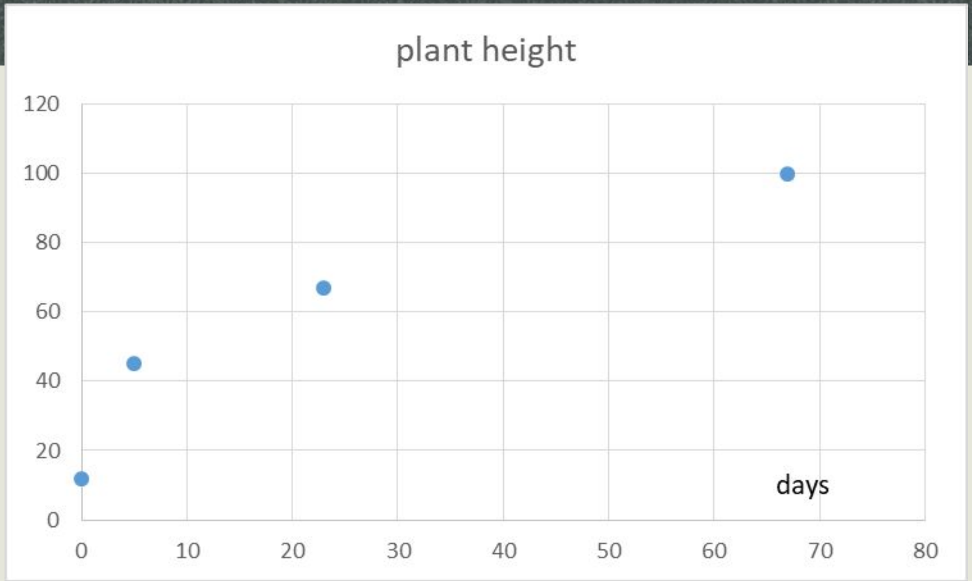
ML – Regression I –Start with Data

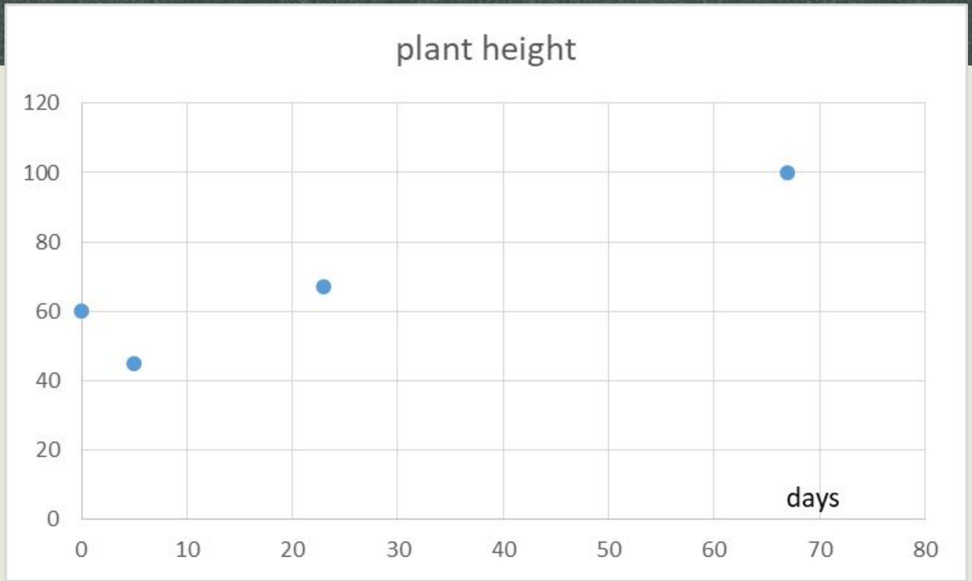




- The datasets tell us different things.
- There appears to be a pattern on the left, less so on the right.

ML – Regression I –Start with Data

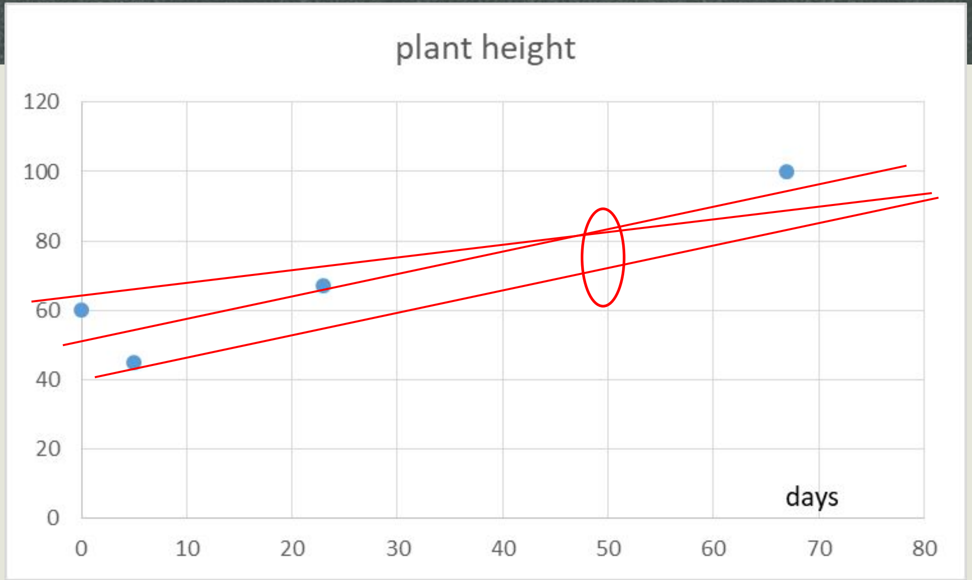




More info let’s us ask questions

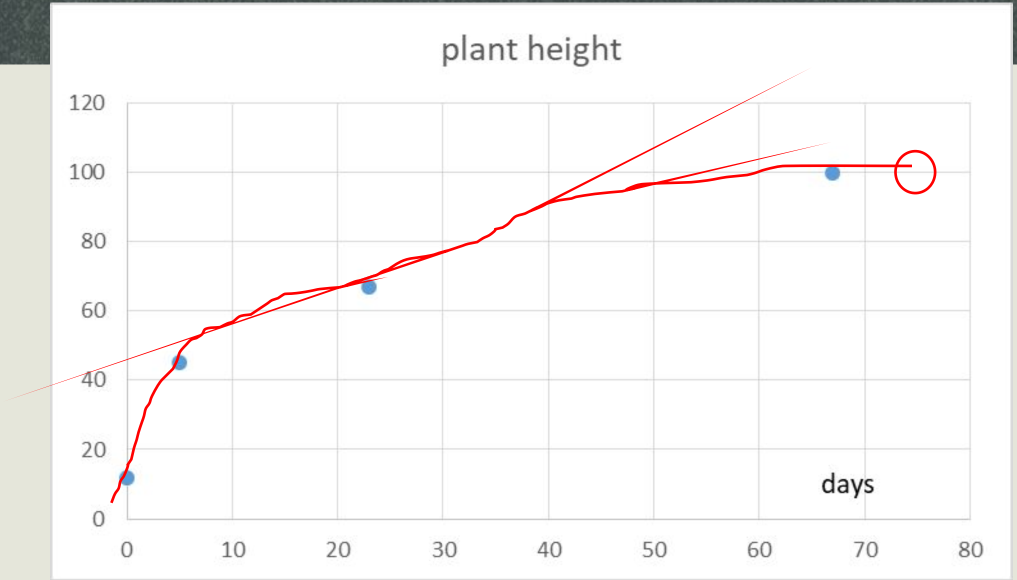
- E.g. students were asked to measure everyday, but some forgot
- How do we determine what happened on day 50?

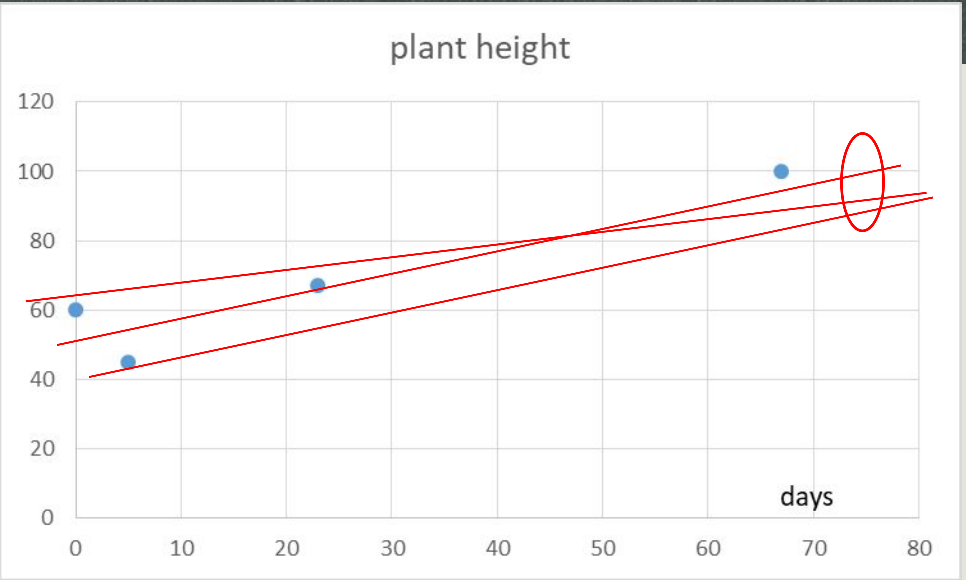
ML – Regression I –Start with Data



- How do we determine what happened on day 50?
- We call this process, INTERPOLATION
- Interpolation is affected by *noise in the data (bad data, missing data) and the model used (which red line)

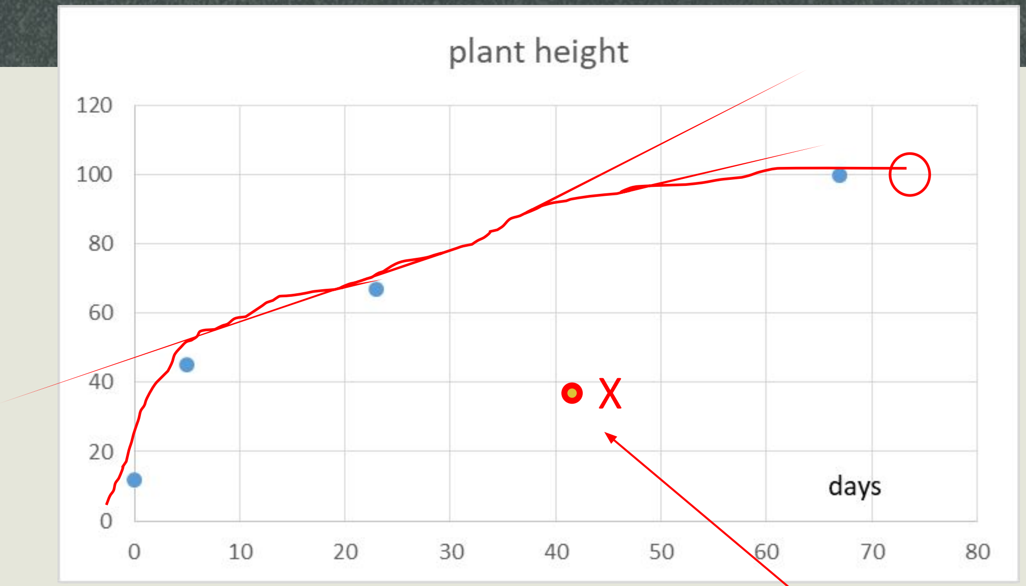
ML – Regression I –Start with Data





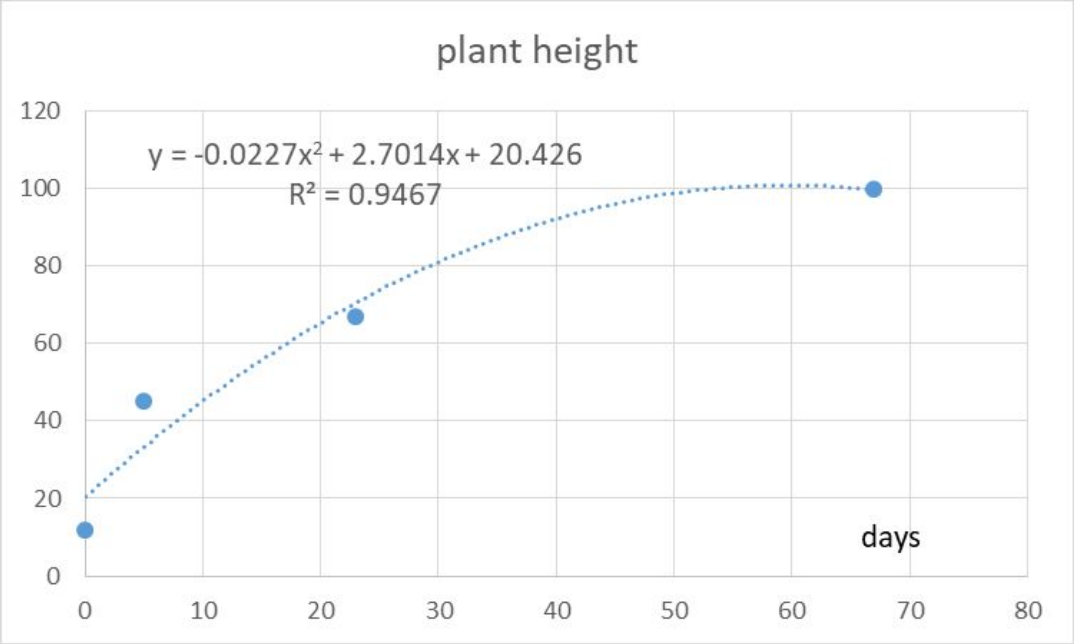
- How do we determine what will happen on day 75?
- We call this process, EXTRAPOLATION (outside the data)
- Extrapolation is affected by *noise in the data (bad data, missing data) and the model used (which red line)

ML – Regression I –Start with Data



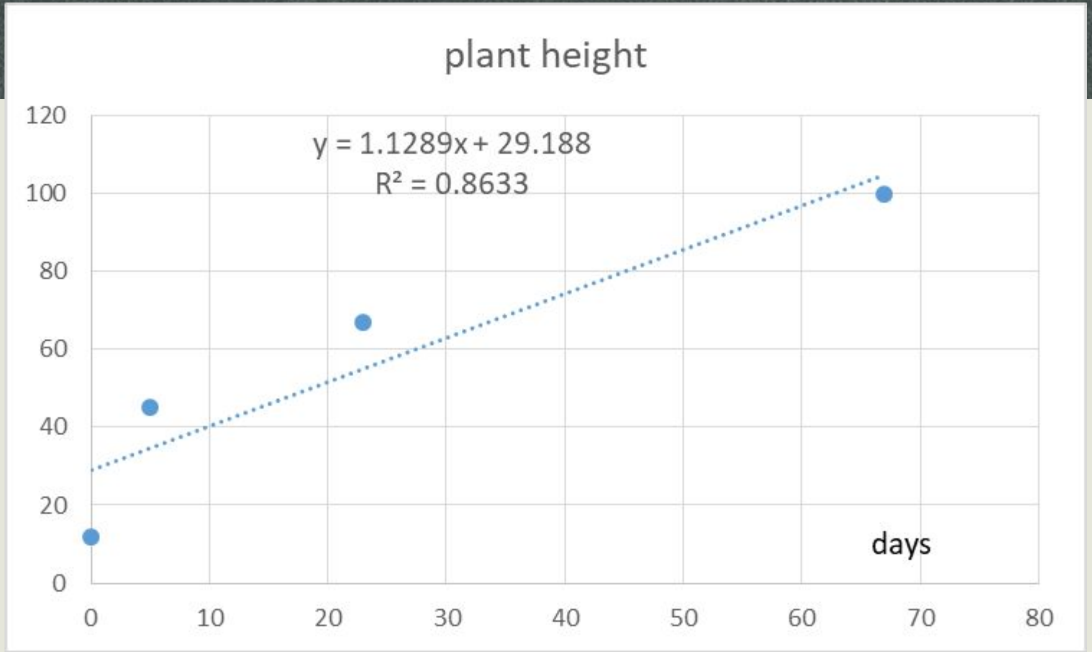
- Data modelling is as much art as it is science
- But we will look at it from a strictly mathematical viewpoint
- E.g. we will NOT ignore OUTLIERS/ANOMALIES

ML - Regression I –Start with Data



- On day 50: height ~100
- On day 75: height = $-0.0227(75^2) + 2.7014\ (75) + 20.426 \sim 95$

ML - Regression I –Start with Data



OR maybe

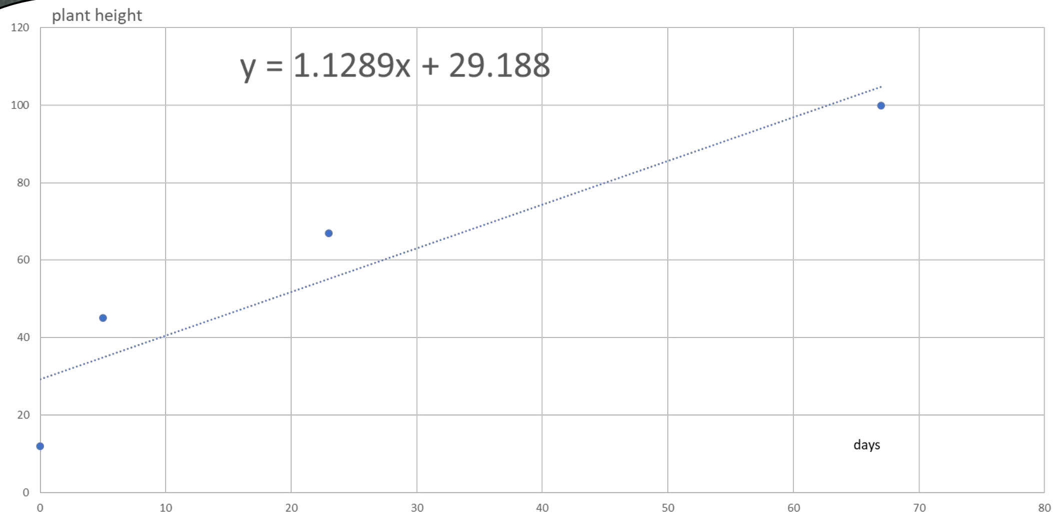
- On day 50: height ~85 (vs. 100)
- On day 75: height ~115 (vs. 95)
- So we need a way to figure out the best solution, with the highest confidence

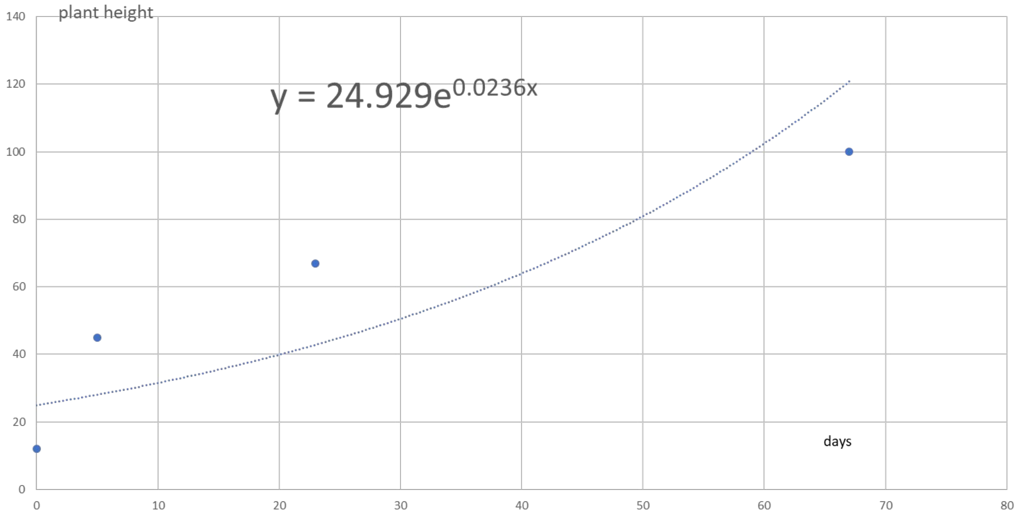
ML - Exact Interpolation –Fitting the Data

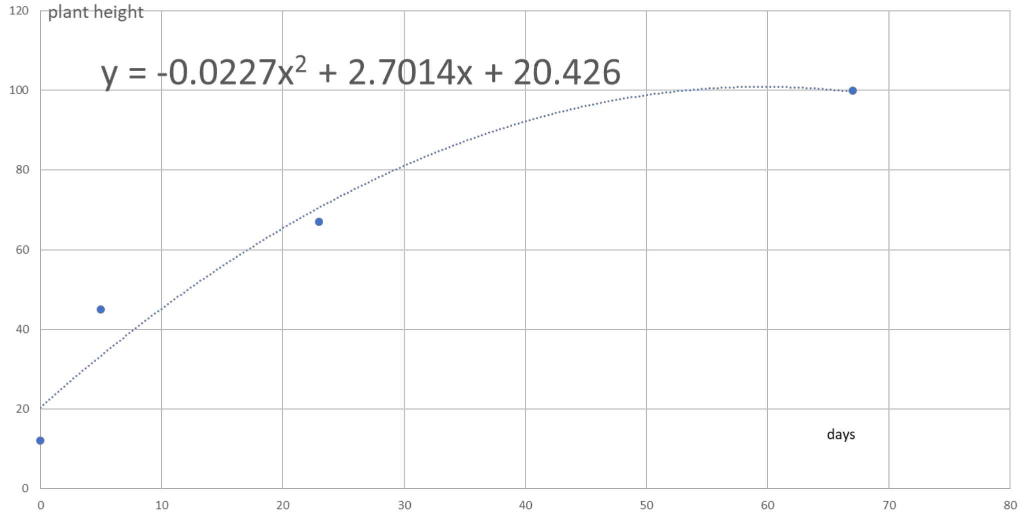
We can connect the dots under the assumption that they are LINEARLY connected (segmentwise)

- Pros: faithful to the data (*is this good?), easy
- Cons: fails completely outside the data during extrapolation, appears "jerky" non-realistic

ML-Regression I –Quality



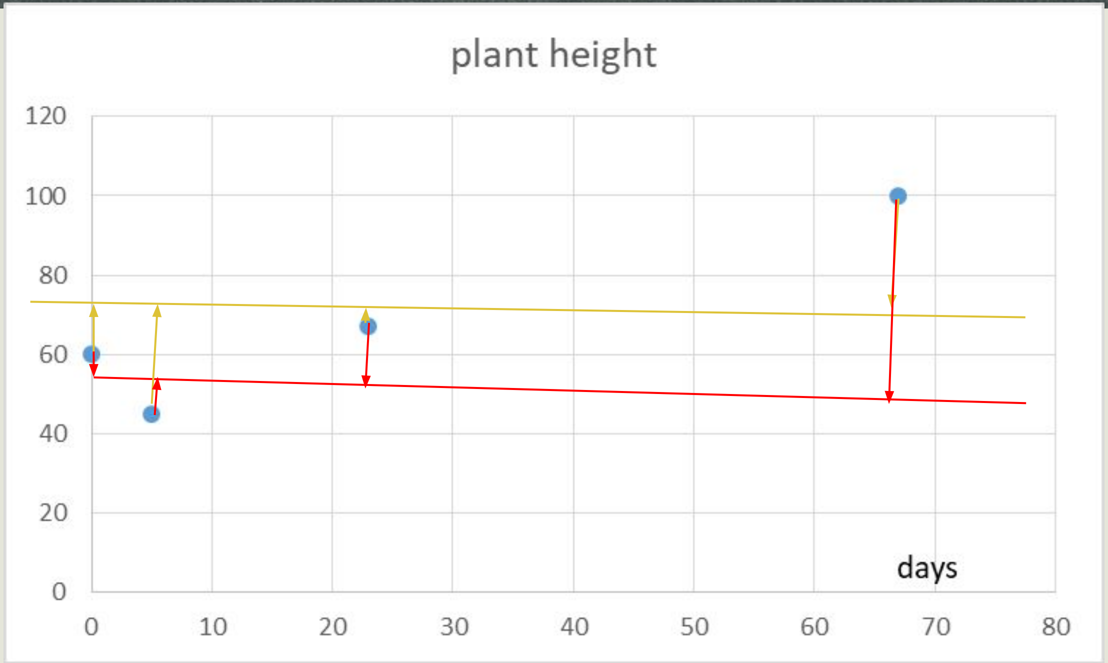




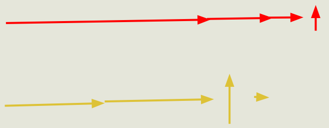
Which one? How do we know what is good?

ML - Regression I –choosing the model

Let's start simple:

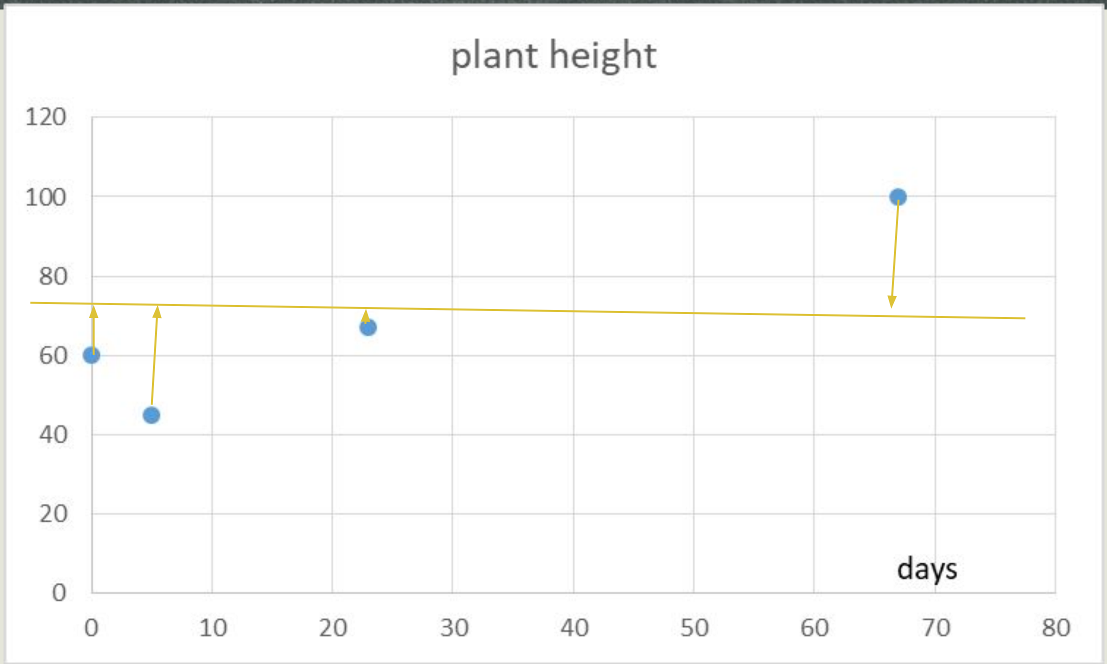


We want to be as close to the data as possible, so overall which are further?



ML - Regression I –choosing the model

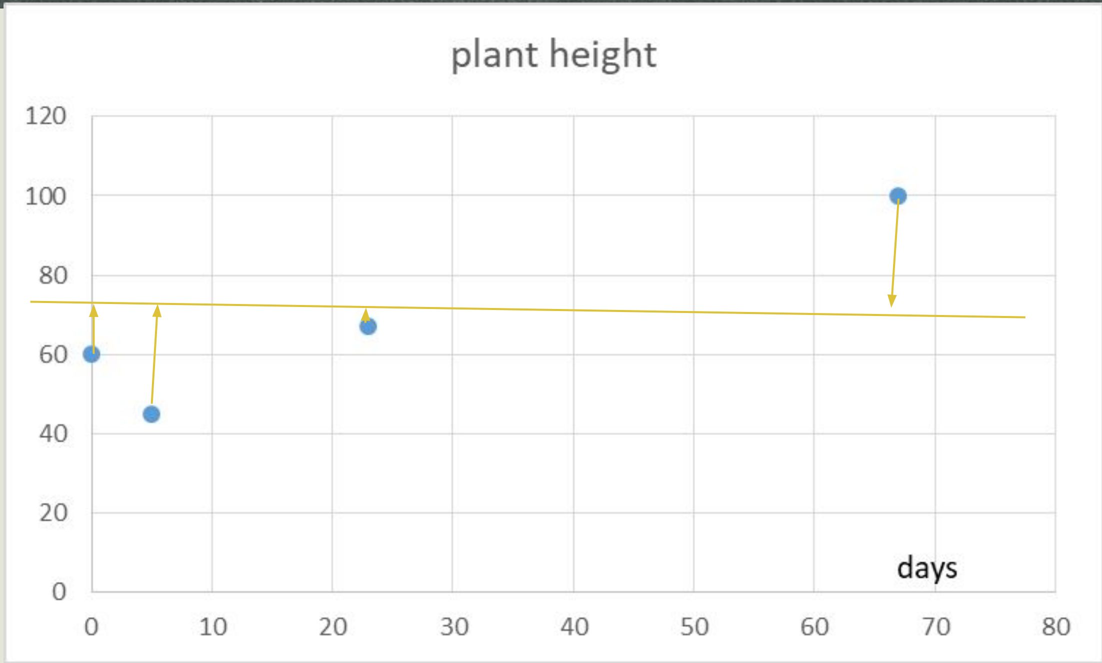
Let's start simple:



Is there a way to make this more mathematical?
How do we find the distance?
How do we ensure the model is linear (a line, not curved)?

ML - Regression I –formalizing linear model

Let's start simple:



The overall line must be of the form, y = m x + b (no bends, kinks)

How do we find the smallest set of distances?

We want the predicted heights to be as close to the line as possible (i.e. the vertical distance).

ML - Regression I –formalizing linear model

Let's start simple:

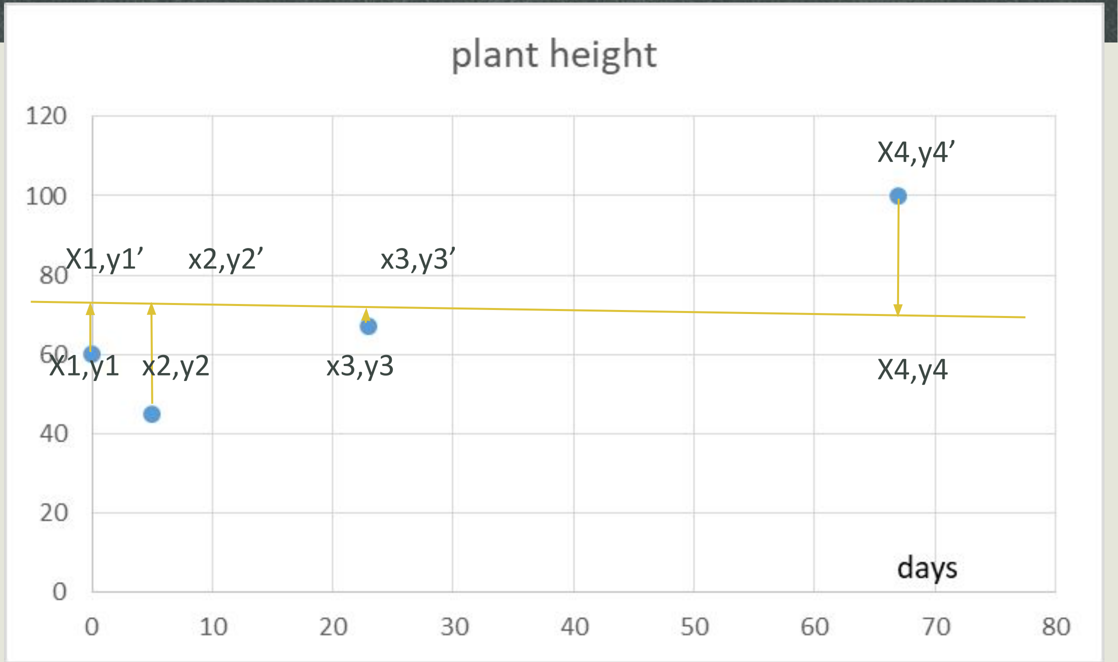
How do we find the smallest set of distances?

$\overrightarrow{(Y_1'-Y_1)} + \overrightarrow{(Y_2'-Y_2)} + \overrightarrow{(Y_3'-Y_3)} + \overrightarrow{(Y_4'-Y_4)}$

But some of these are > 0 (x_1,x_2,x_3),
and some are a lot less (x4), so they cancel

So take the squares:

$(Y_1'-Y_1)^2 + \ldots + (Y_4'-Y_4)^2$



But this grossly inflates the values, so we can take the square root of the sum: √

To be fair, we average them (<u>R</u>oot <u>M</u>ean <u>S</u>quare <u>E</u>rror): $\sqrt{}$**[$(Y_1'-Y_1)^2$ + … + $(Y_4'-Y_4)^2$] /4**

ML – Regression I –formalizing linear model

Mathematically:

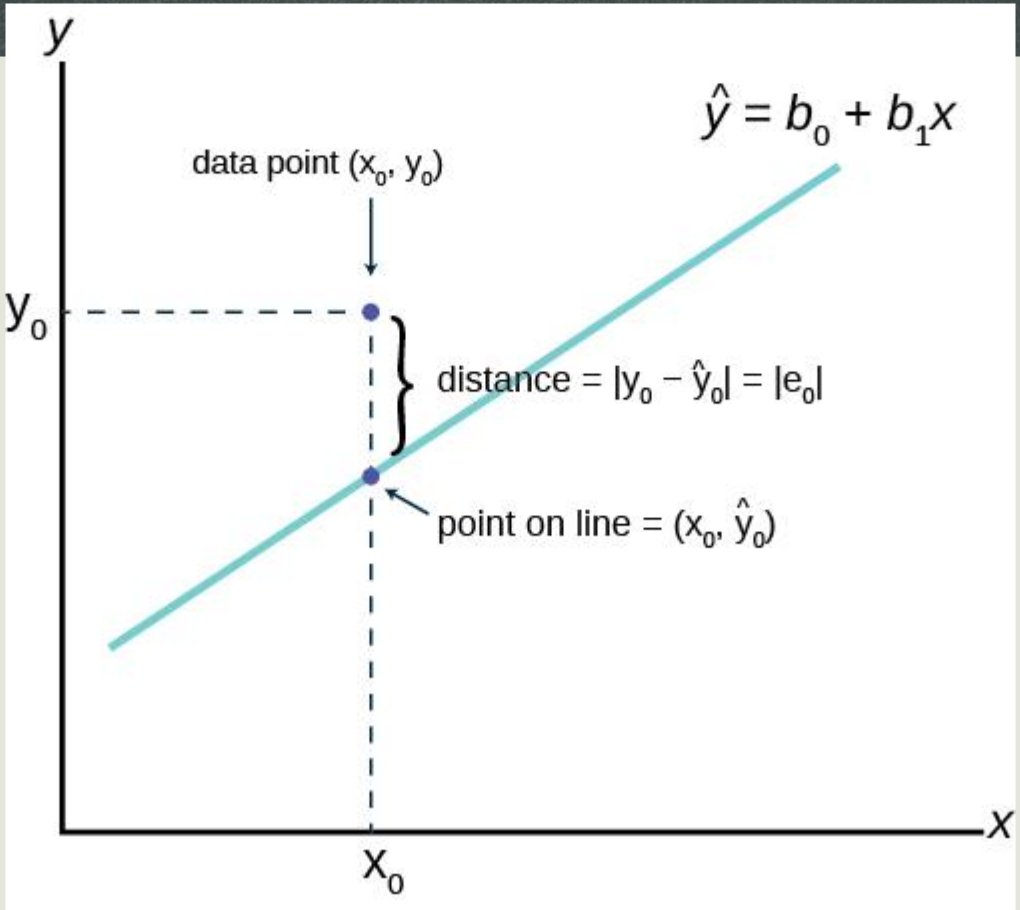
We call the distance (difference between the estimate and the data point) the RESIDUAL $|e_0|$.

But there are a whole set of residuals:

$|e_0|, |e_1| \ldots |e_N|$

We can sum these and minimize the total using partial derivatives (calculus).

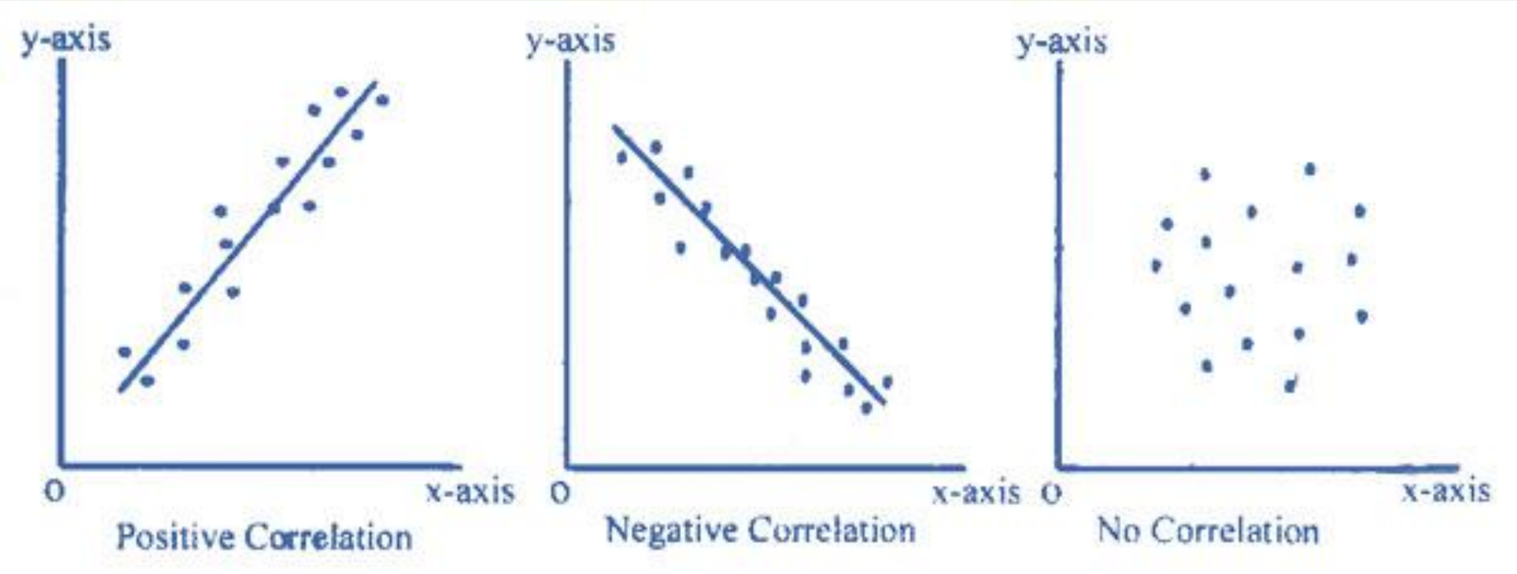
This becomes another OPTIMIZATION problem.



ML-Regression I –Fit? R

The correlation coefficient, R, tells us how 2 variables co-relate, i.e. how they are related to each other.

If one variable goes up, does the other one go up as well? Or does it go down? Or does it do neither?:



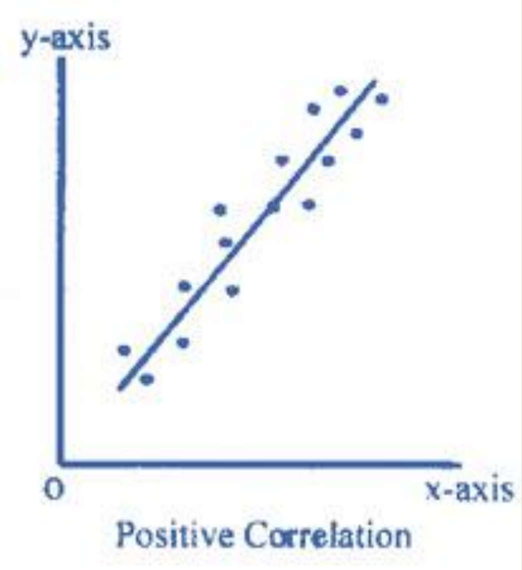
ML-Regression I –R^2

The **Determination Coefficient** (square of the correlation), **R^2**, shows how much variance is explained by the other variable

Say there is 80% correlation, R=0.8, then R^2=0.64. We say the two variables are positively co-related, and that 64% of the variability in one is explained by the other.

Variance is how much the data is spread (indirectly measuring the residuals)

Correlation explains how strongly they link



ML-Regression I –R^2

What is 100% correlation?

When 2 variables are absolutely correlated (100%), it means every time one is observed, the other has also been observ**ed**. **This is NOT to be confused with causality!**

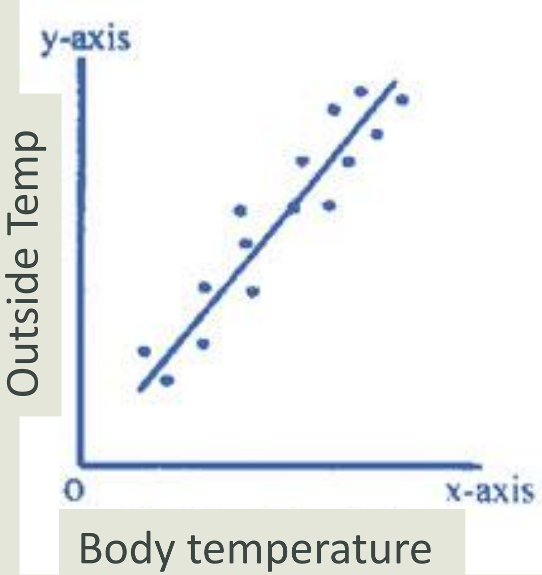
In data science, a great many errors can be made by doing this:

“It is cold outside, <u>implies</u> I am cold.”

Is it true that the temperature outside is making me cold?

ML-Regression I – R^2

"It is cold outside implies I am cold."



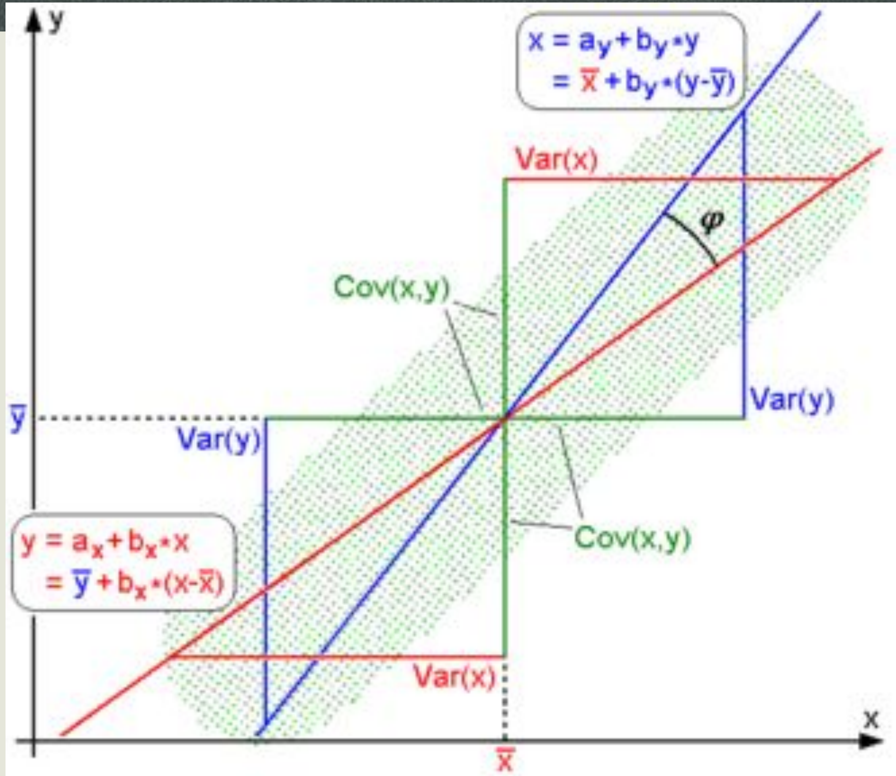
this ***causal***? Is the outside causing
or vice versa?

be <u>partially true</u>, but there are
explanations:

-maybe I have no clothing/shelter?

ML-Regression I –R^2–calculating

How do we calculate the R?



- R(raw scores, means)
- **Standardized covariance**
- Standardized slope of the regression line
- Geometric mean of the two regression slopes
- **Square root of the ratio of two variances**
- **Function of the angle between two standardized regression lines**
- Rescaled variance of the difference between standardized scores
- Ratio of two means

ML-Regression I – R^2 – calculating

How do we calculate the R?

Recall the difference between the fitted (predicted) variables and the data points can be summed:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2,$$

here the $\bar{Y}$ refers to the vector average of the Y values.

Think about why this is true….

ML-Regression I – R^2 – calculating

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2,$$

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}.$$

We can apply some principles such as individual sample correlation with their difference from the mean, as well as some algebra:

$$R = \sqrt{\frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}}.$$

ML-Regression I – R^2 – calculating

$$r(Y,\hat{Y})^2 = \frac{\sum_i(\hat{Y}_i - \bar{Y})^2}{\sum_i(Y_i - \bar{Y})^2}$$

The square of the correlation coefficient is the amount of the variance (variations or spread) in the fit, ***Y***, given the linear function, $\hat{Y}$(x).

If R^2 is high (1), then the model is "good".

Read:https://rpubs.com/beane/n3_1b

Group Exercise

Yass! Let’s exercise :)

Questions

Do the following tasks in your groups (very important to delegate well)

1. Settle down with your groups. Make sure each machine in your row is working. If not, write to itsupport@queenscollege.ca for help. Make sure each machine has Lockdown Browser and it is working. All graded in-class assessments will happen using the in-class terminal (no personal laptops) and quizzes will use the lockdown browser.
2. Work with your groups on the questions of AI provided in Moodle

And we learnt our first ML model!! Yay!!

Ok, that was a lot of Maths! Let’s make sure we practice it!