FOUNDATIONS OF MACHINE LEARNING

ANUSH SANKARAN

Most of the slide content are borrowed from multiple source online. Thanks to the original authors

WHO AM I?

- PhD in Computer Science, IIIT Delhi (2010 - 2017)
  - Thesis: “Learning Representations from Fingerprint Variants”
  - TCS Research Grant

- IBM Research AI, India (Nov. 2015 - Now)
  - Advisory Research Scientist
  - 25+ Publications
  - 12 Patents
  - IBM Research recognition award for Outstanding Technical Leadership for serving as a global technical lead in shaping the agenda for “Machine Learning for Creativity”
  - Technical Lead in end-to-end delivery of the product “Neural Network Modeller” as a part of Watson Studio

- Visiting Faculty, IIIT Bangalore (Jan 2019 – May 2019)
  - Grad Level Course on “Visual Recognition”

OVERVIEW OF THE COURSE

Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research

COURSE OUTCOMES

- Understand the fundamental concepts of different machine learning models
  - Supervised learning
  - Unsupervised learning
- Ability to formulate a business problem as machine learning task. Identify machine learning opportunities in businesses.
- Appreciate the challenging involved in data driven machine learning problems
- Ability to manage the building of tools and products that involves different aspects of machine learning

EASY LOGISTICS: GITHUB

- Github Repo: https://github.com/goodboyanush/isme-bangalore-Oct-Nov-2019
- Lectures slides, Hands-on code, Assignment solutions
- Have any doubt in my lectures or assignments?
  - Go ahead and create an issue in the repo!
  - I will try to answer them asap!
  - Everyone will be benefitted by the questions asked by one

EASY LOGISTICS: EXPECTED TASKS FROM YOU

- Task 1: Clone this Repo: https://github.com/goodboyanush/isme-bangalore-Oct-Nov-2019
  - All your assignment codes, blogs, hands-on exercises will be available here
  - Easy for me to follow
- Task 2: Use one of your existing blogs. And if you don't have one, create one at Github Pages: https://pages.github.com/
- Task 3: Add the link of your blog in the your forked Github project's README

EASY LOGISTICS: EVALUATION OUTLINE

- Weekly hands-on exercises + 1 blog post
- 2 project assignment from Kaggle
- 1 mid-semester exam
- 1 end-semester exam

BACKGROUND MATERIAL

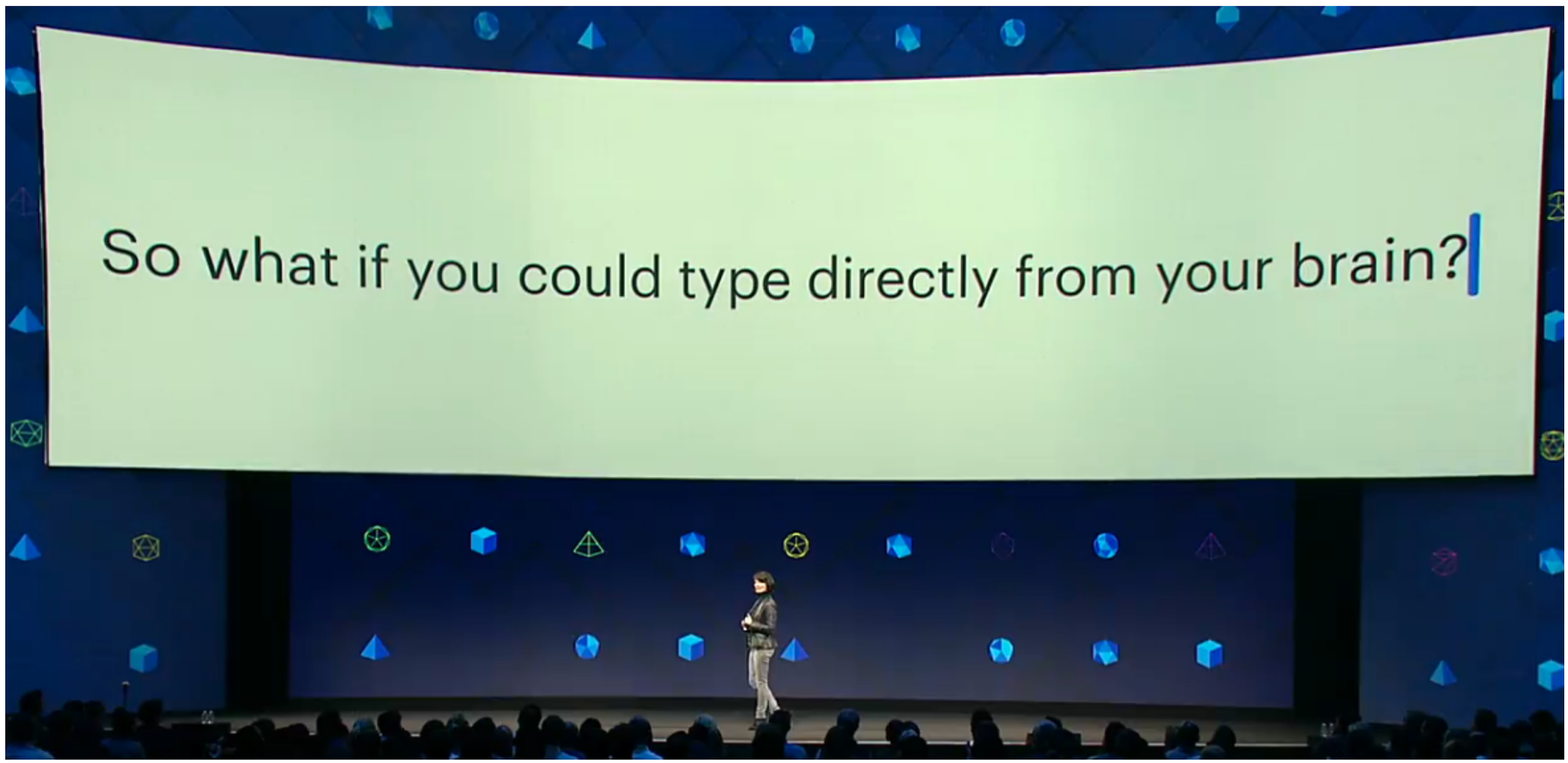
- Suggested Book: “Machine Learning in Action” by Peter Harrington
- Kaggle Competitions: https://www.kaggle.com/
- Coursera Course by Stanford: https://www.coursera.org/learn/machine-learning
- UCI Machine Learning Repository: https://archive.ics.uci.edu/

WEEK 1:

INTRODUCTION TO MACHINE LEARNING

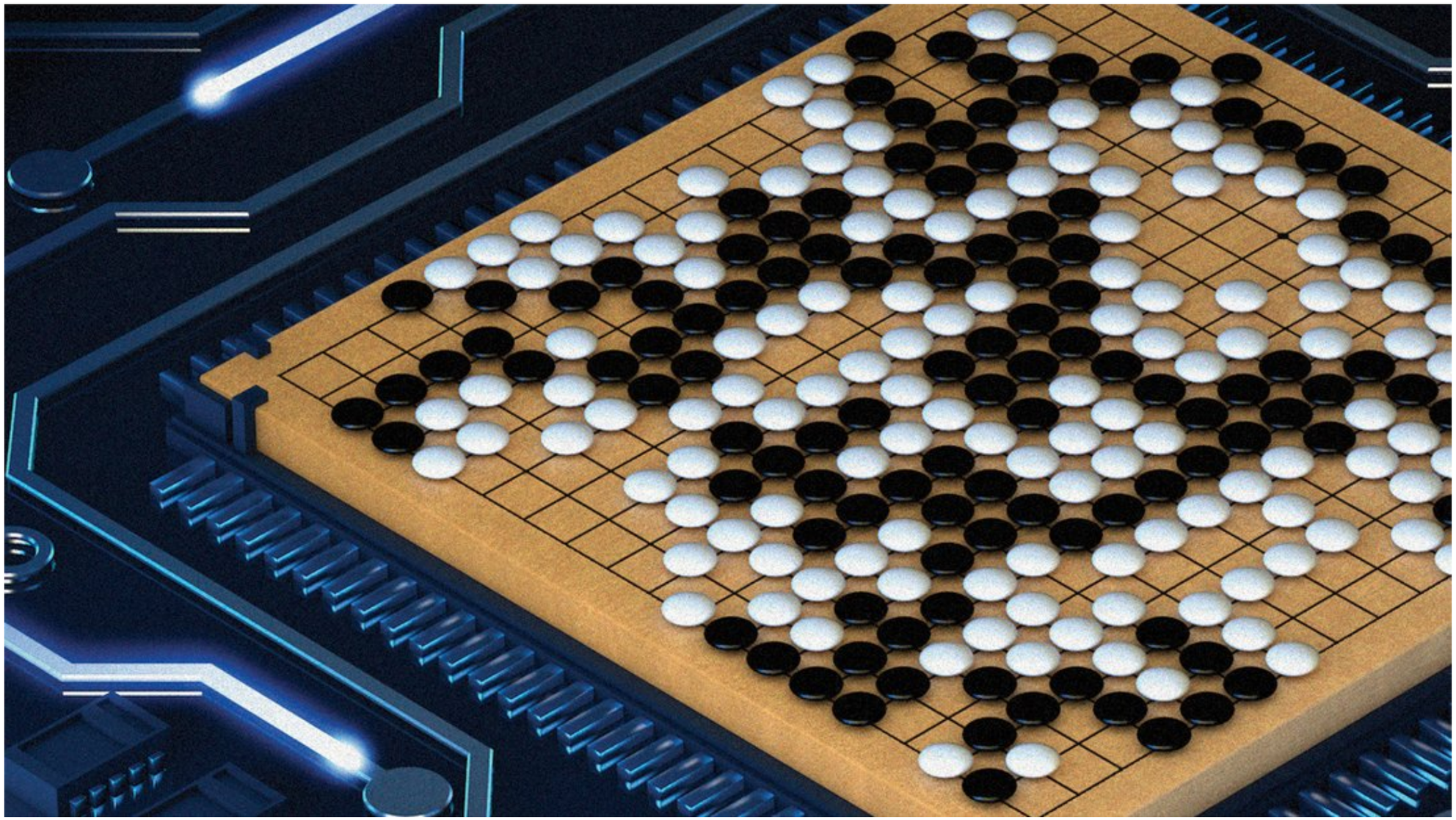
BUILDING THE ML MINDSET IN BUSINESS

NEURAL MAPPING OF THOUGHTS

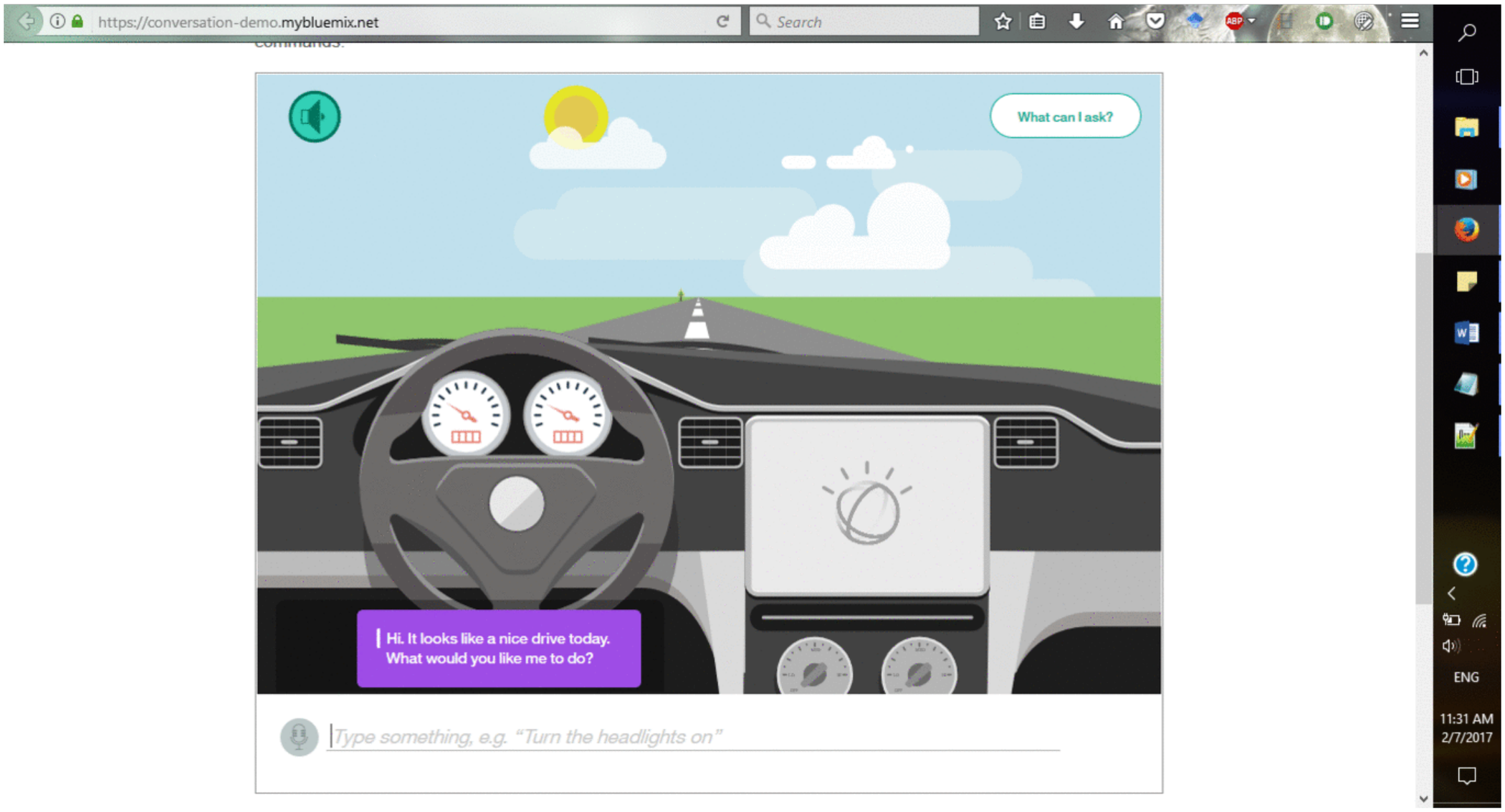


(credit: Facebook)

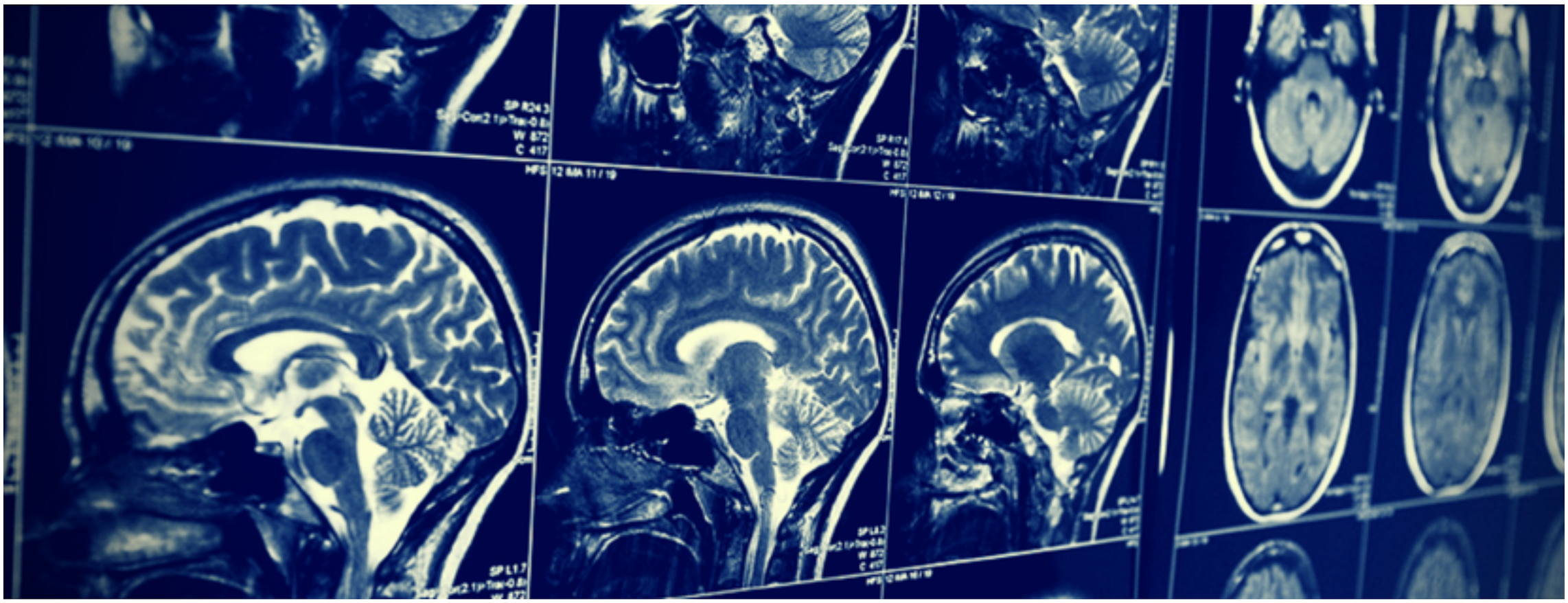
AI DEFEATS WORLD'S TOP GO PLAYERS



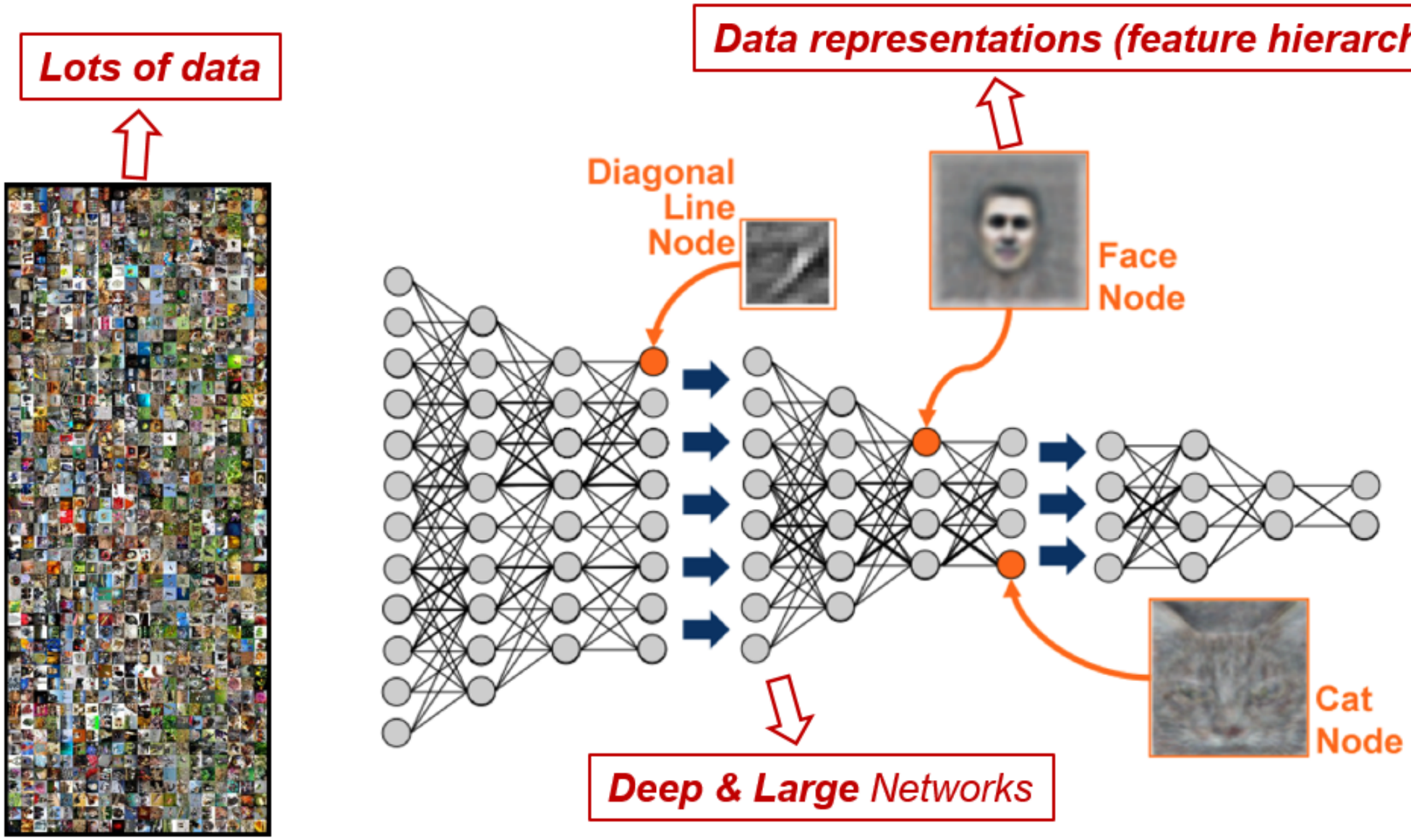
HUMAN LIKE CHATBOTS



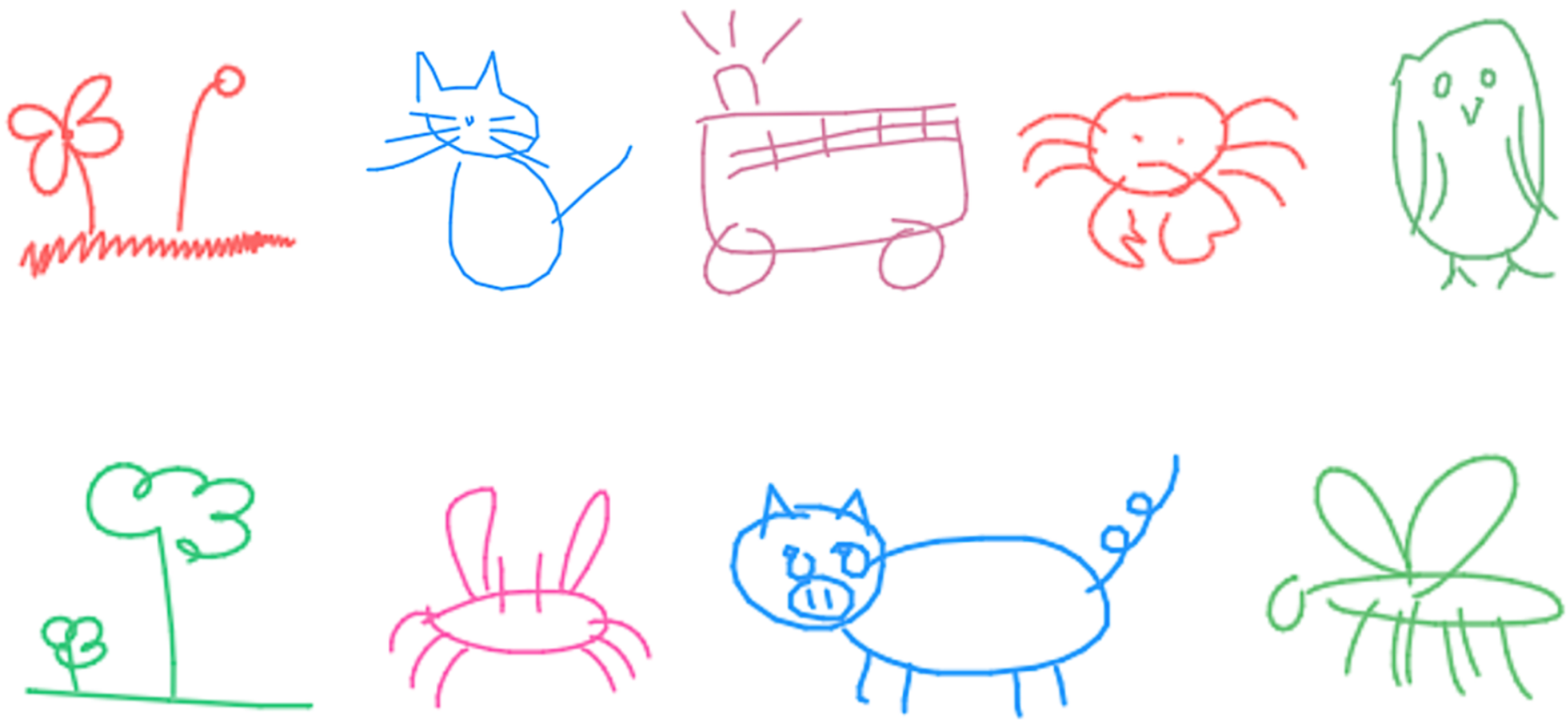
DETECTING BRAIN CANCER



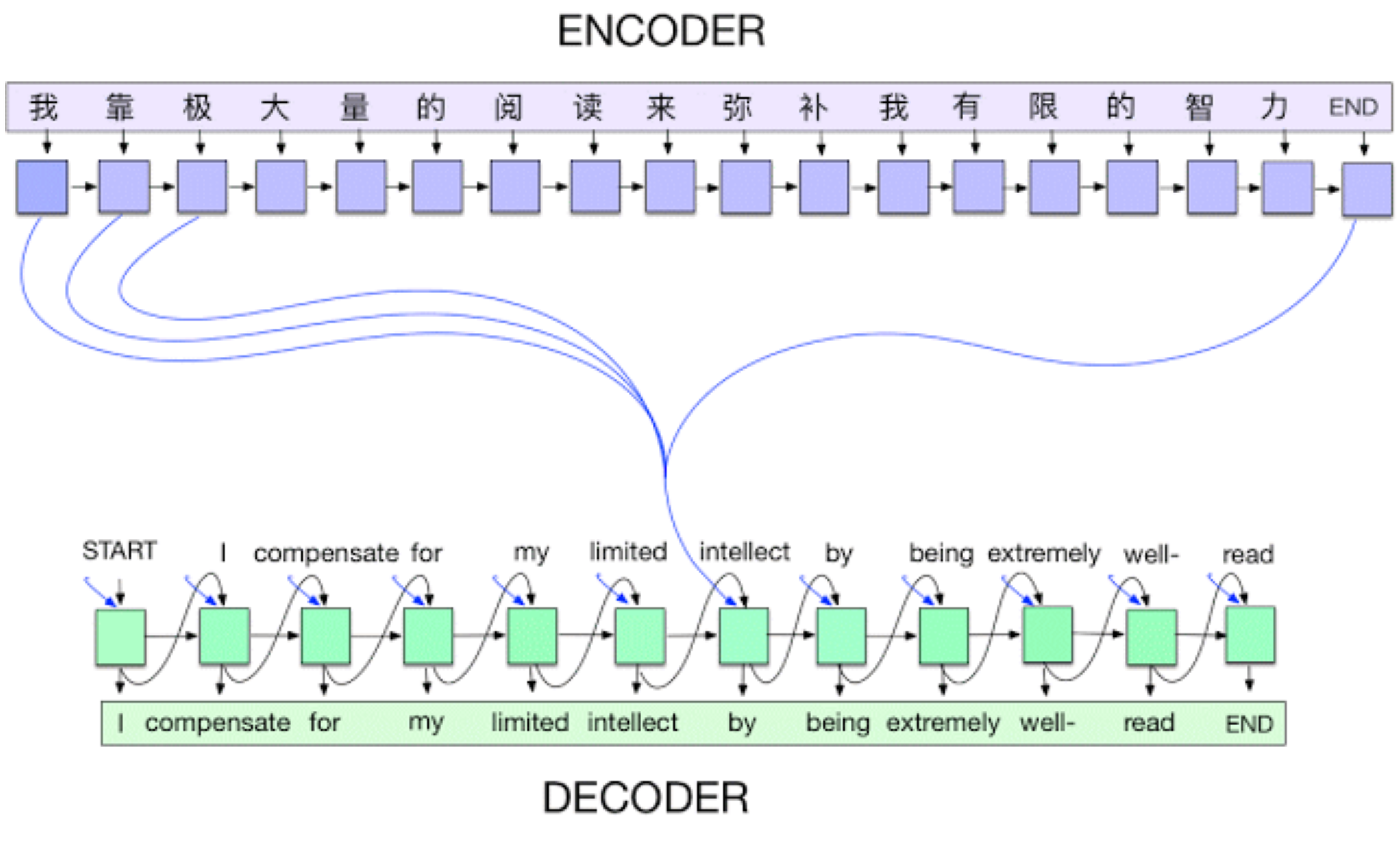
UNCONSTRAINED HUMAN FACE RECOGNITION



MACHINE LEARNING TO DRAW

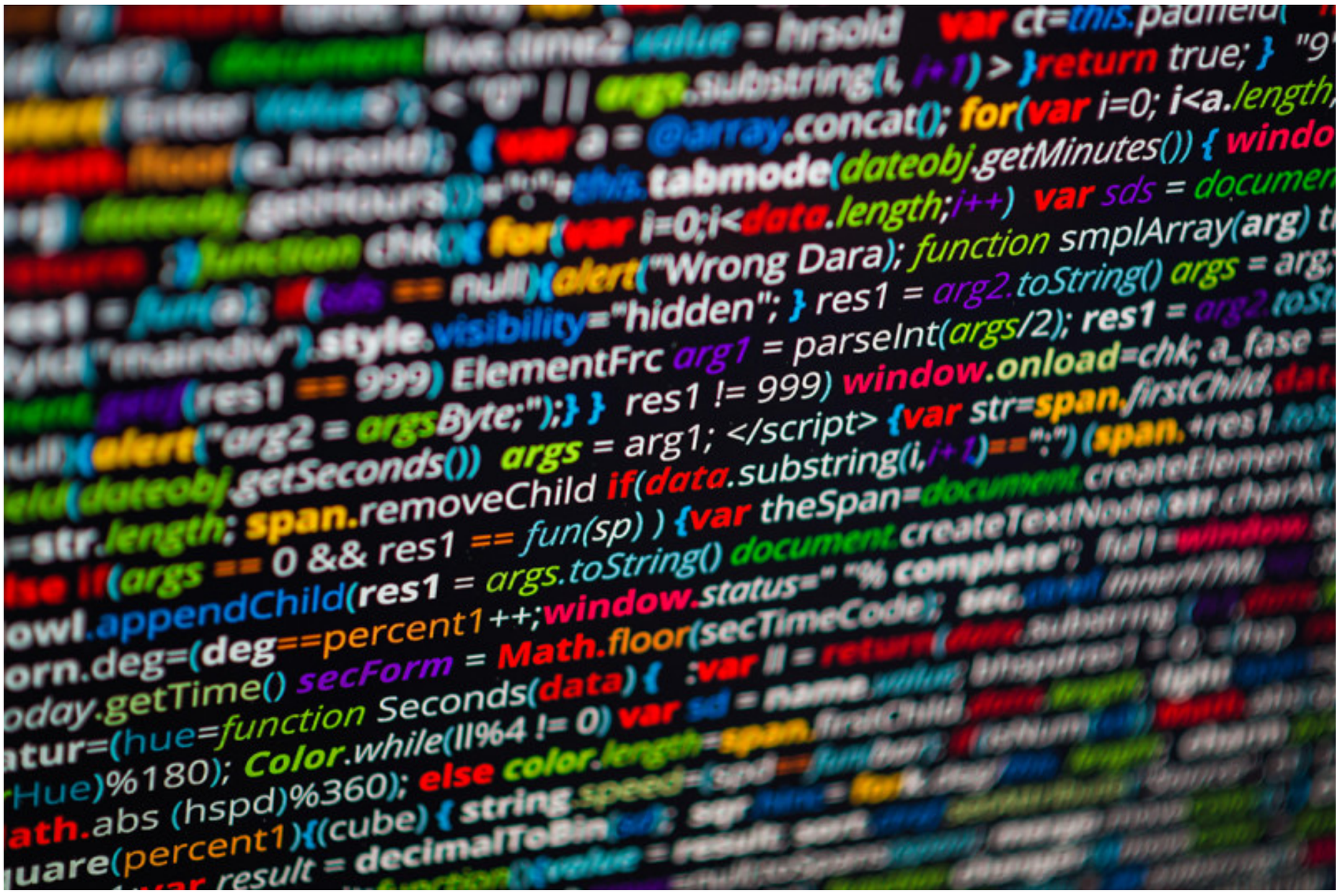


REAL TIME LANGUAGE TRANSLATION



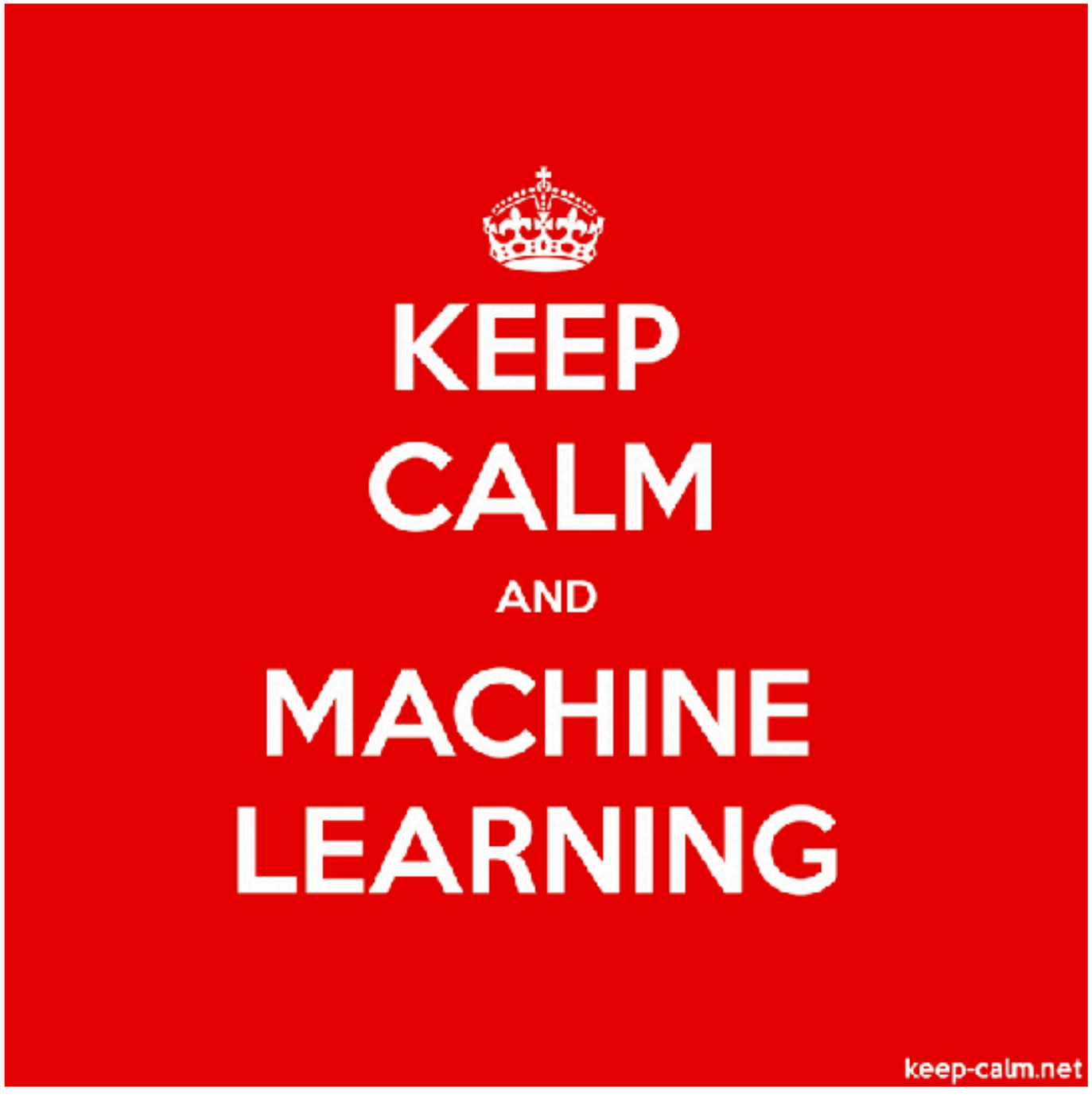
SELF DRIVING CARS

LEARNING TO WRITE PROGRAMS

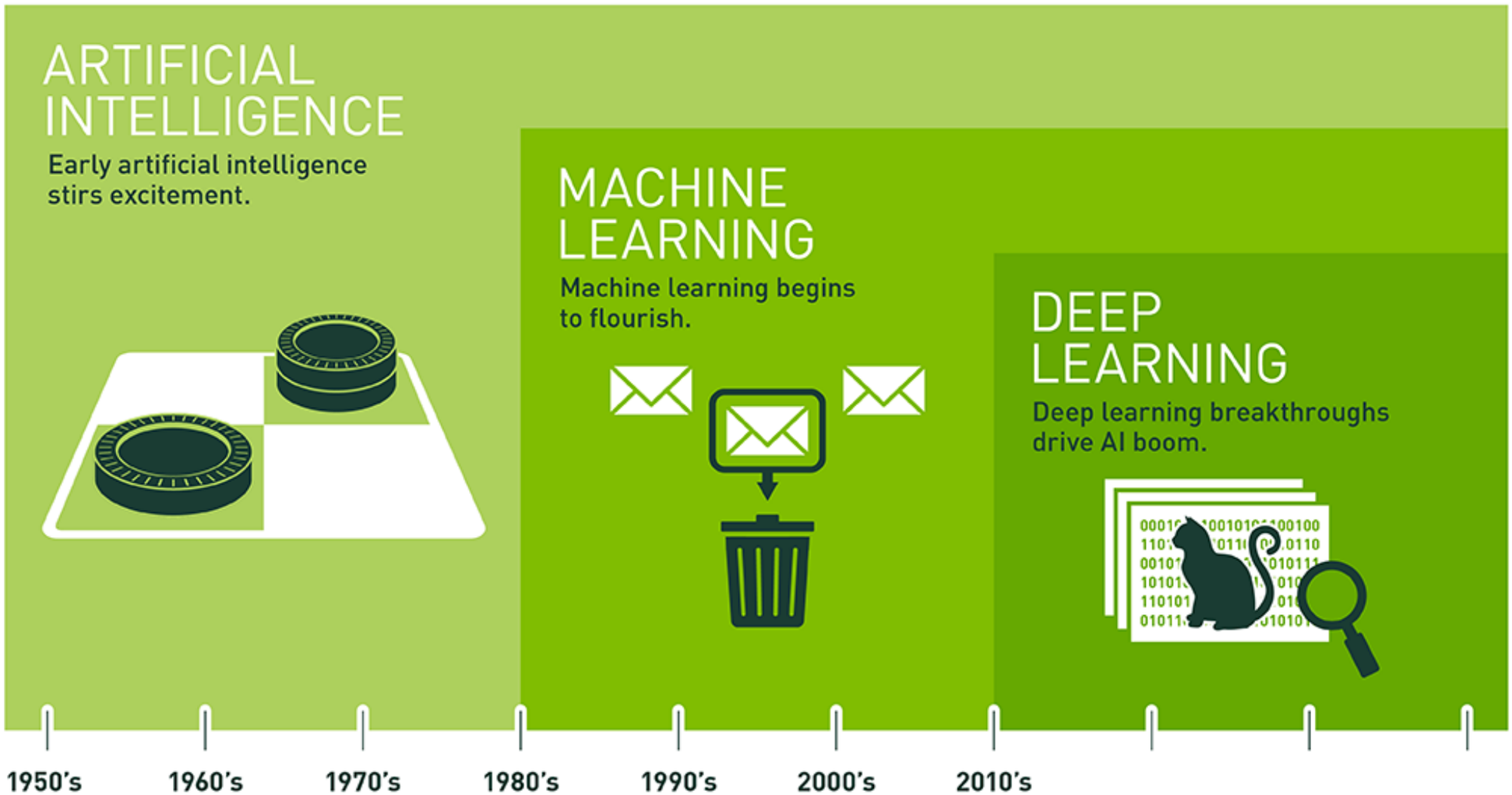


AI GETTING MORE CREATIVE

ALL POSSIBLE BECAUSE OF MACHINE LEARNING



HISTORY OF MACHINE LEARNING



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

WHAT IS MACHINE LEARNING?

- Classifying a person as MALE or FEMALE

Instances/ Input Data Points

1. Height
2. Weight

Features

1. Male:
   1. Weight > 75kg
   2. Height > 5'9"
2. Female:
   1. Weight < 70kg
   2. Height < 5'7"

Classification

1. **Feature extraction**: How to represent the data points ?
2. **Classification**: How to use the features to distinguish the classes ?

LOSS OF GENERALIZATION !!!

WHAT IS MACHINE LEARNING?

- Learn a classifier: learn a mapping function

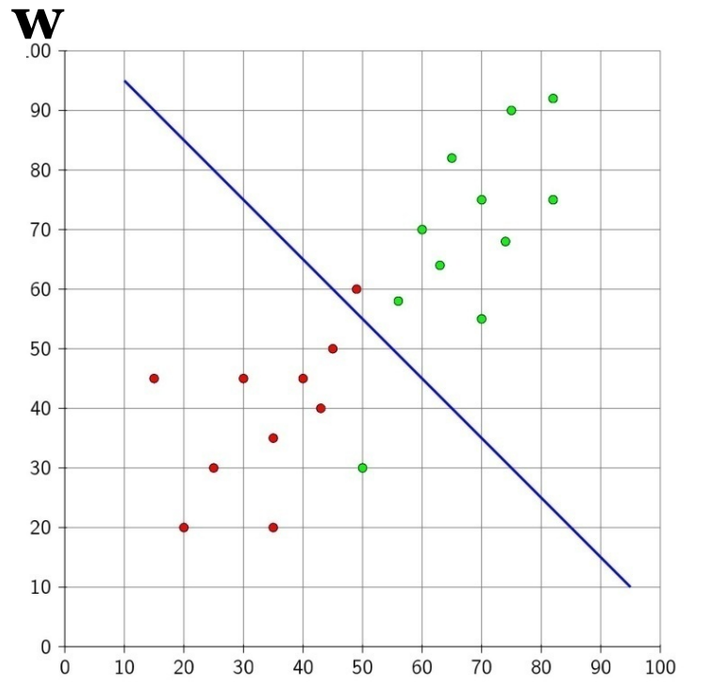
Instances/ Input Data Points

1. M: <h1, w1>
2. F: <h2, w2>
3. …

…

N. M: <hn, wn>

Labelled Features

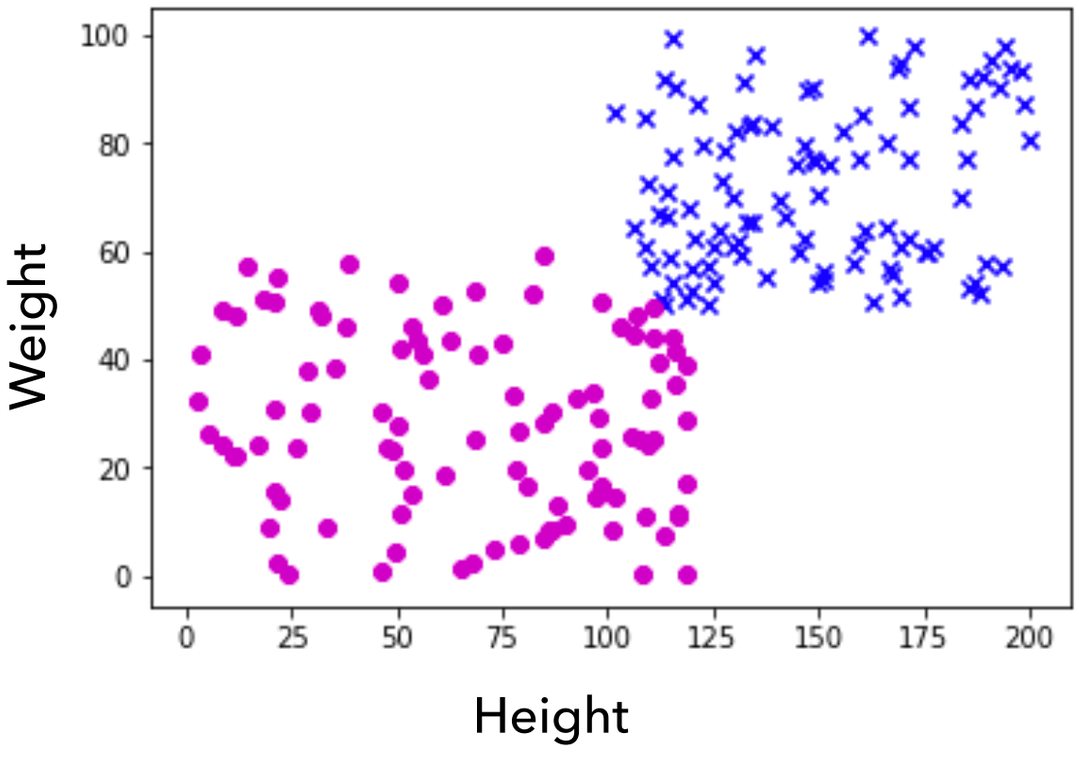


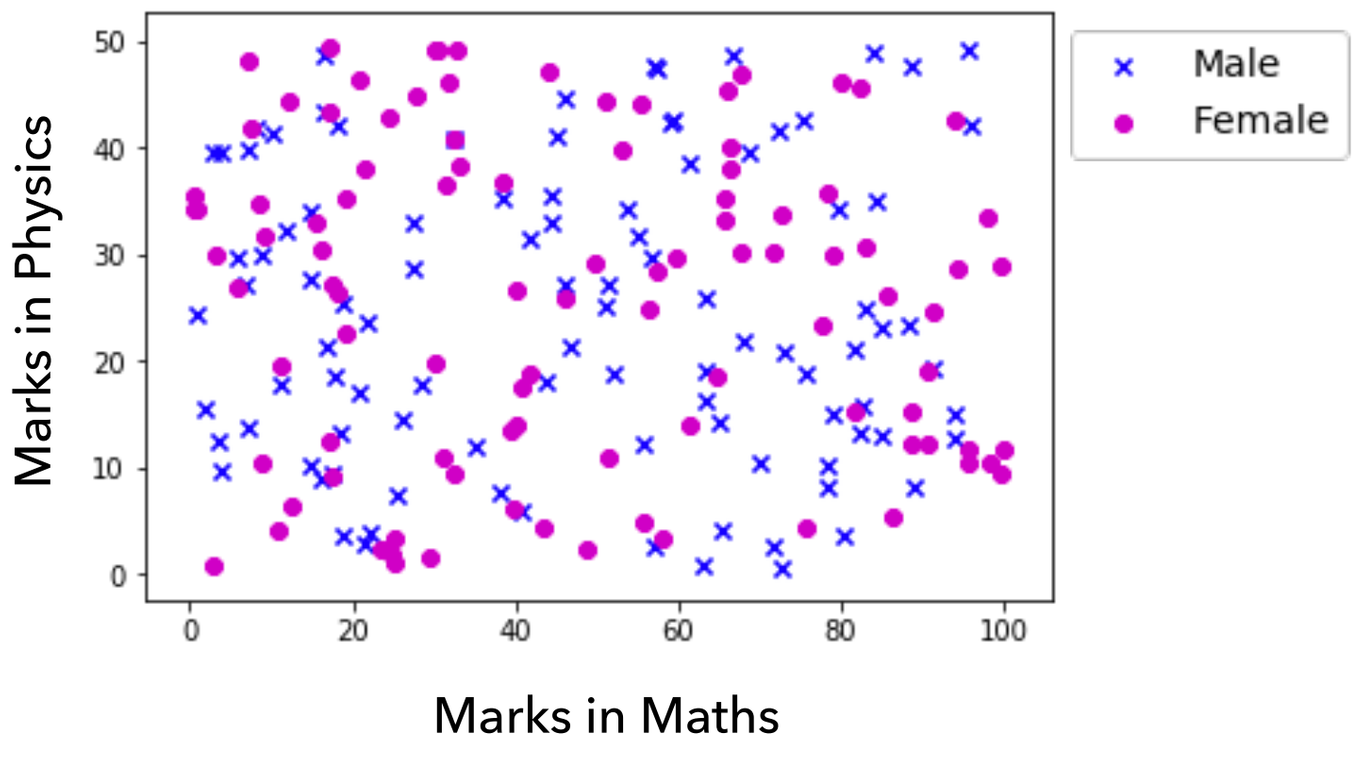
Classification

1. **Boundary**: Can be linear or non-linear boundary
2. **Method**: Can be a generative classifier or discriminative classifier

Examples: Naïve Bayes, Decision trees, Neural Network, Support Vector Machines etc.

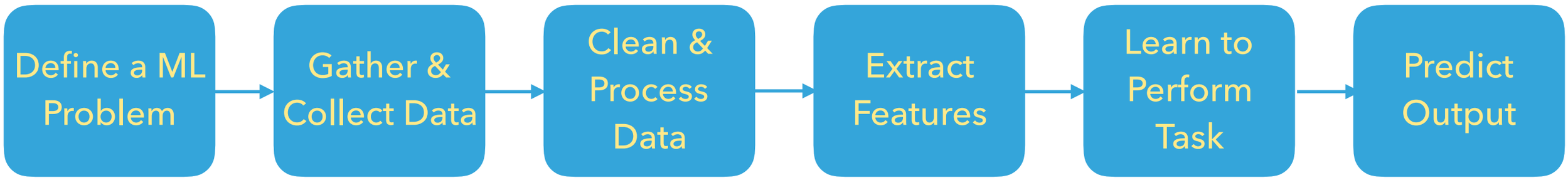
WHAT IS MACHINE LEARNING?



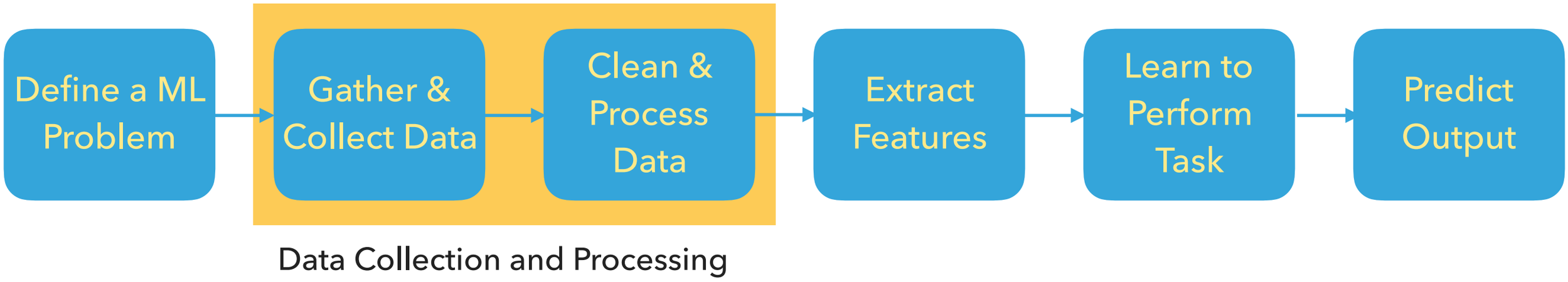


Classification is only as good as feature extraction !!!

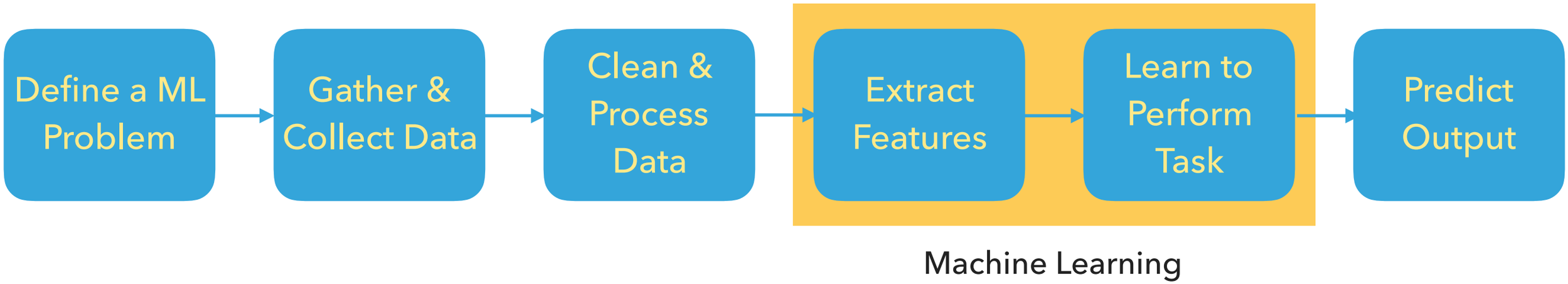
MACHINE LEARNING PIPELINE



MACHINE LEARNING PIPELINE

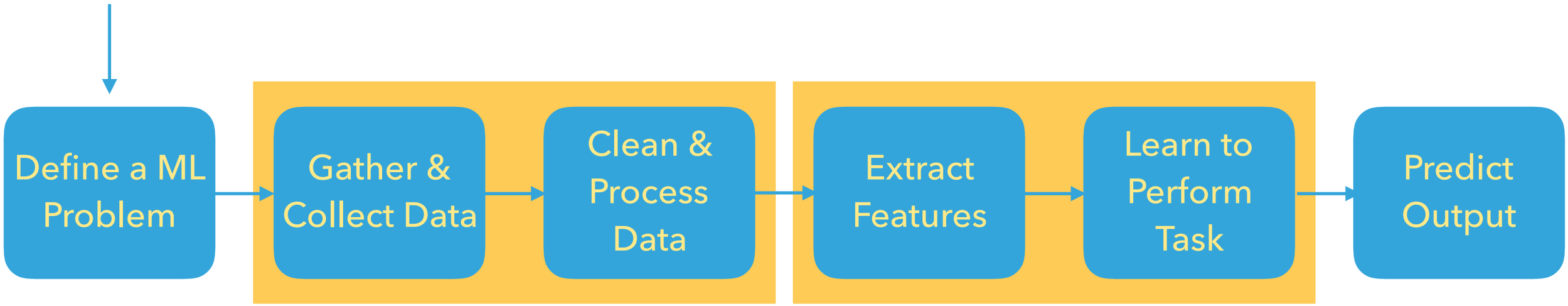


MACHINE LEARNING PIPELINE



MACHINE LEARNING PIPELINE

What are we focussing on today's lecture?



1. Articulate the problem (task)
2. Data Drive Strategy: Look for labelled data
3. Design your data for the task
4. Determine easily obtained inputs
5. Determine easily quantifiable outputs

COMMON LINGO

Input data/ Features

Output
Task Label

Instances

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N

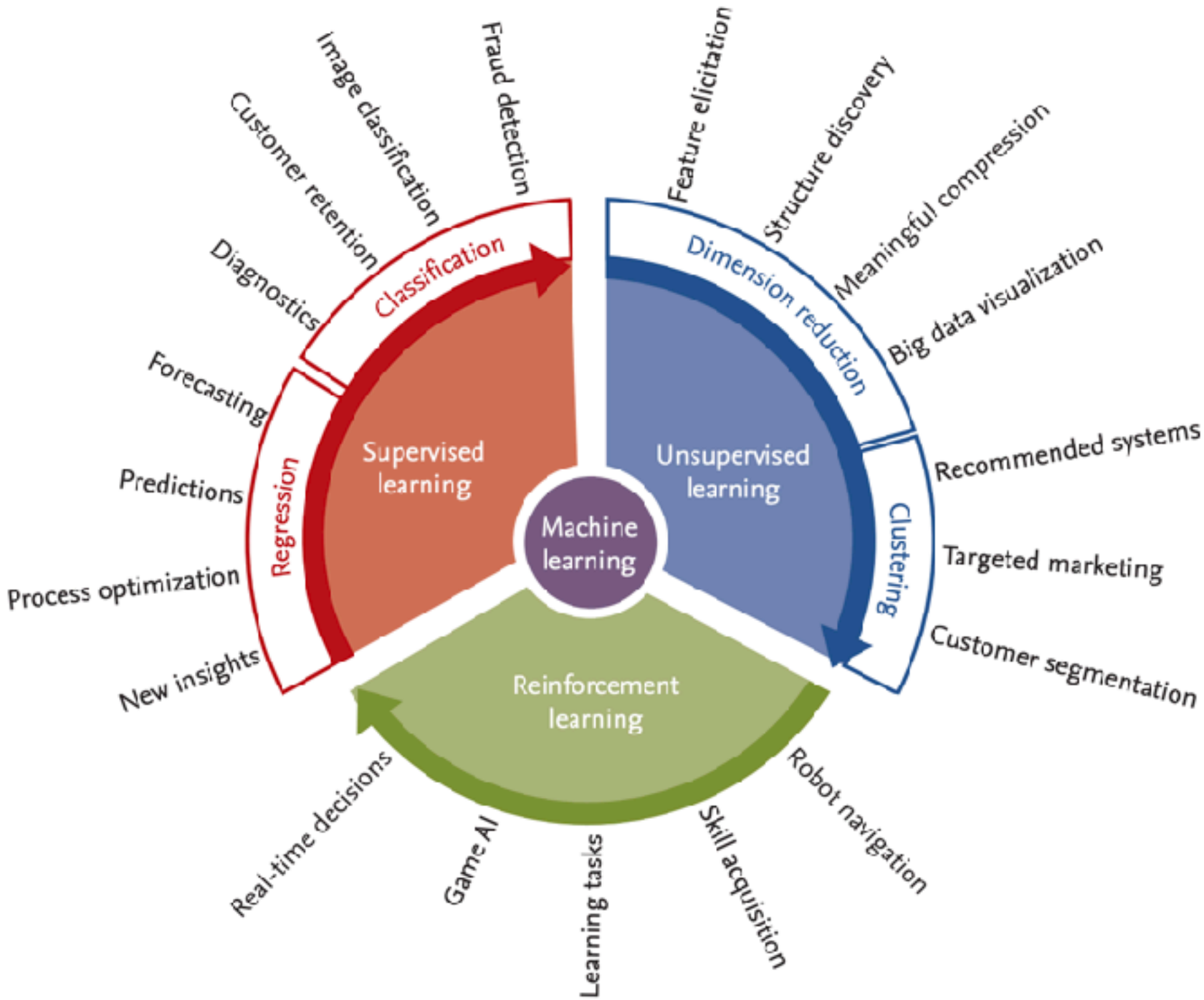
Seen data / Training data

Unseen data/ Test data

LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	
LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	
LP001047	Male	Yes	0	Not Graduate	No	2600	1911	116	360	0	Semiurban	

Machine Learning Glossary: https://developers.google.com/machine-learning/glossary

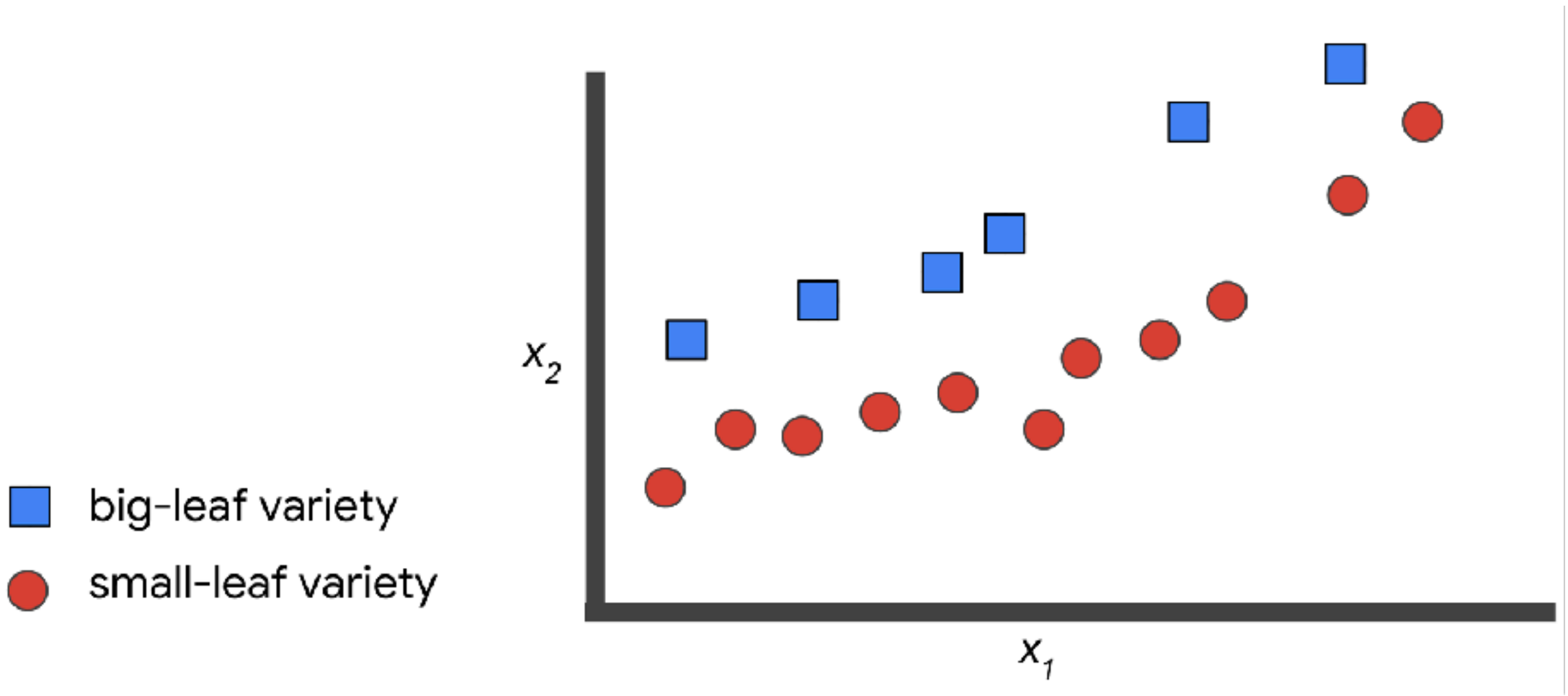
DIFFERENT TYPES OF ML ALGORITHMS



SUPERVISED LEARNING: EXAMPLE

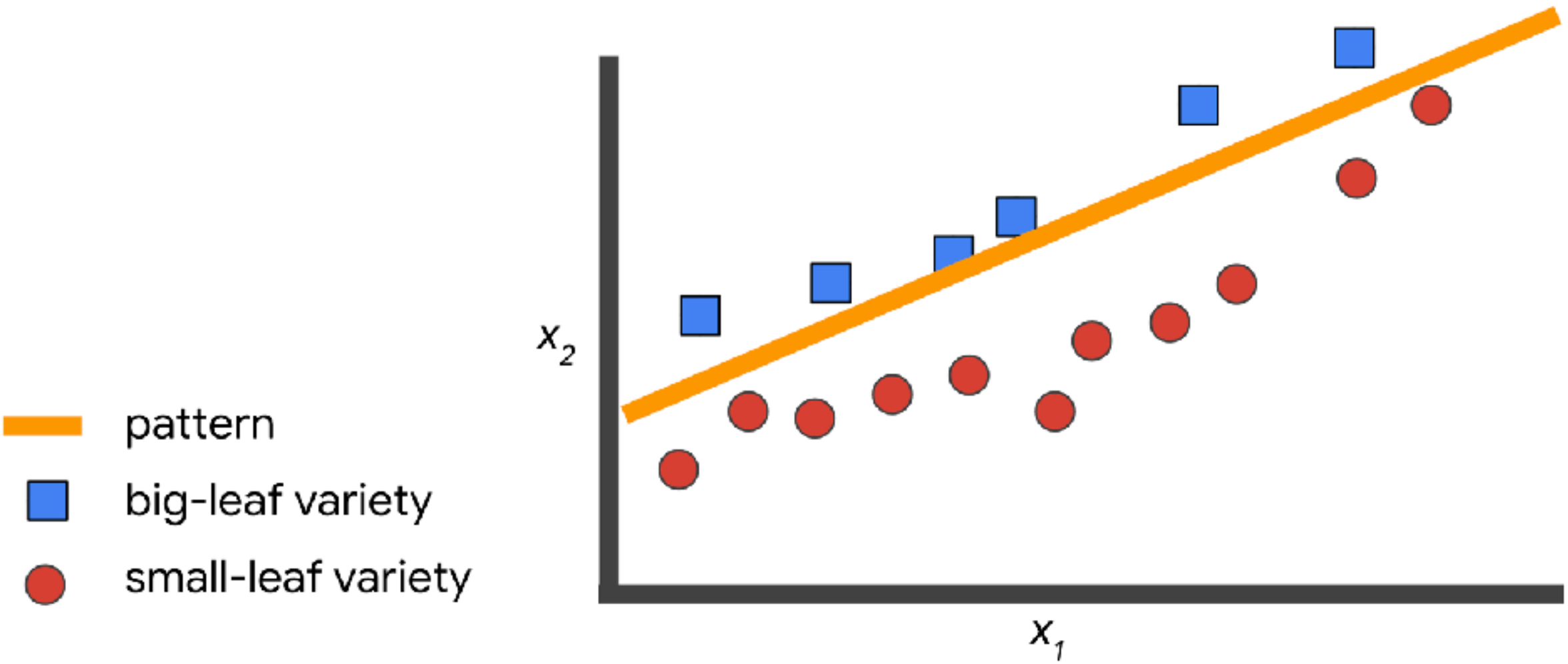
Leaf Width	Leaf Length	Species
2.7	4.9	small-leaf
3.2	5.5	big-leaf
2.9	5.1	small-leaf
3.4	6.8	big-leaf

SUPERVISED LEARNING: EXAMPLE



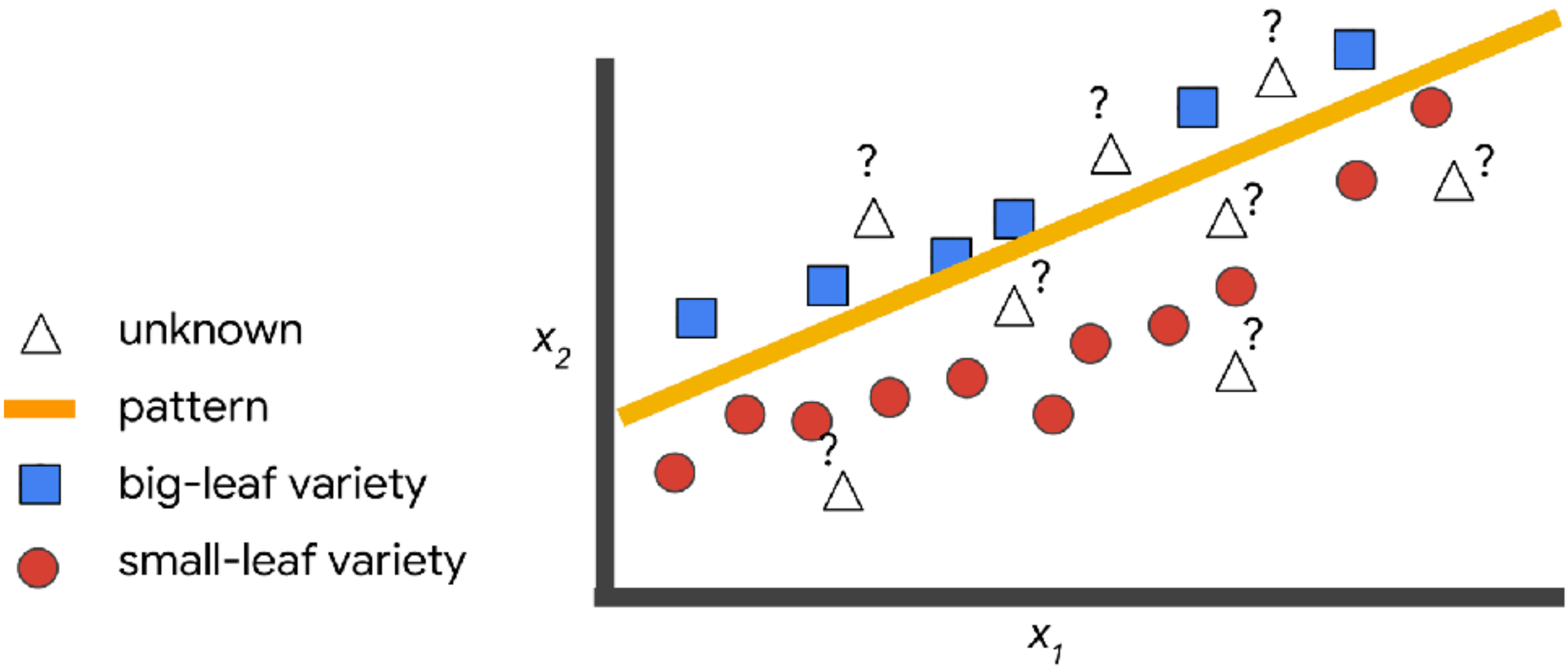
Source: https://developers.google.com/machine-learning/problem-framing/cases

SUPERVISED LEARNING: EXAMPLE



Source: https://developers.google.com/machine-learning/problem-framing/cases

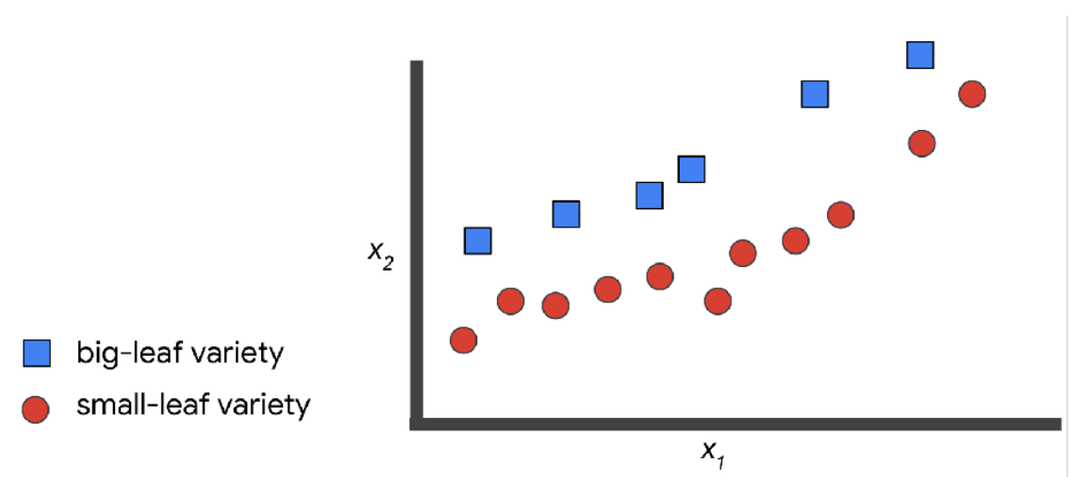
SUPERVISED LEARNING: EXAMPLE



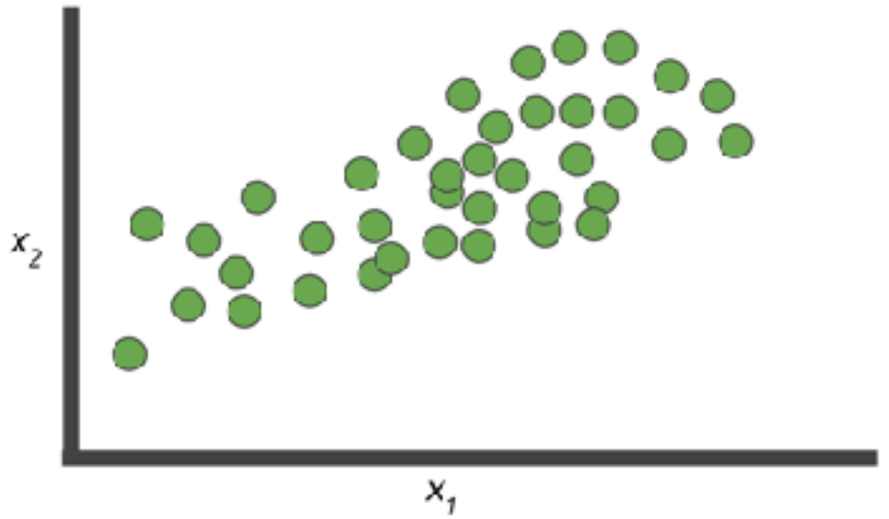
Source: https://developers.google.com/machine-learning/problem-framing/cases

SUPERVISED VS. UNSUPERVISED

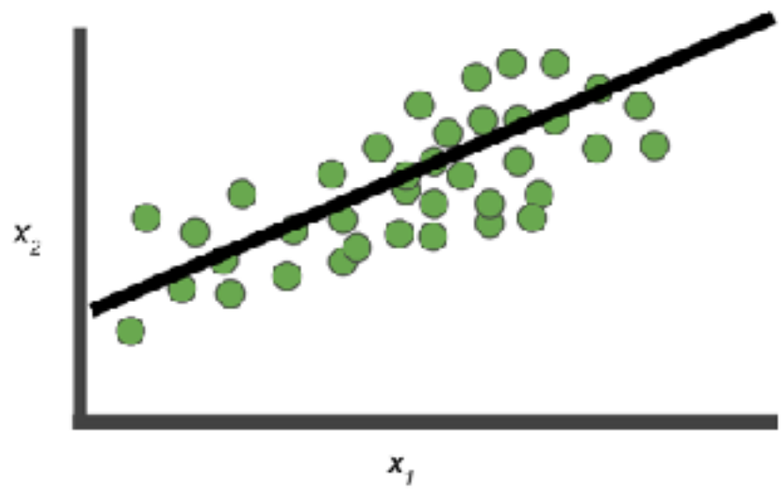
Supervised learning

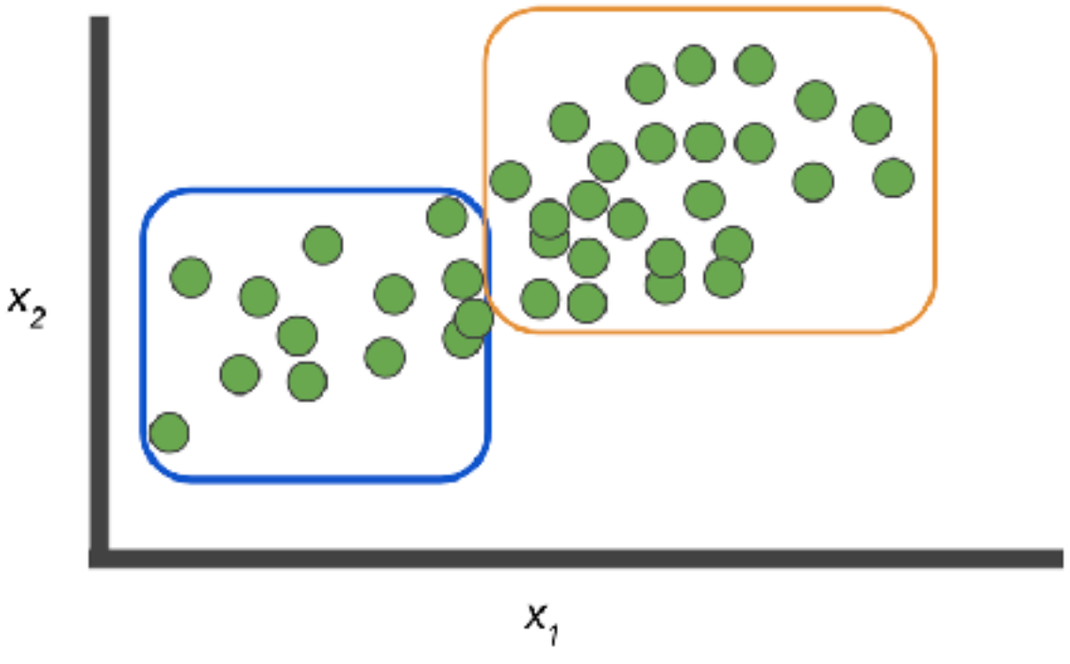


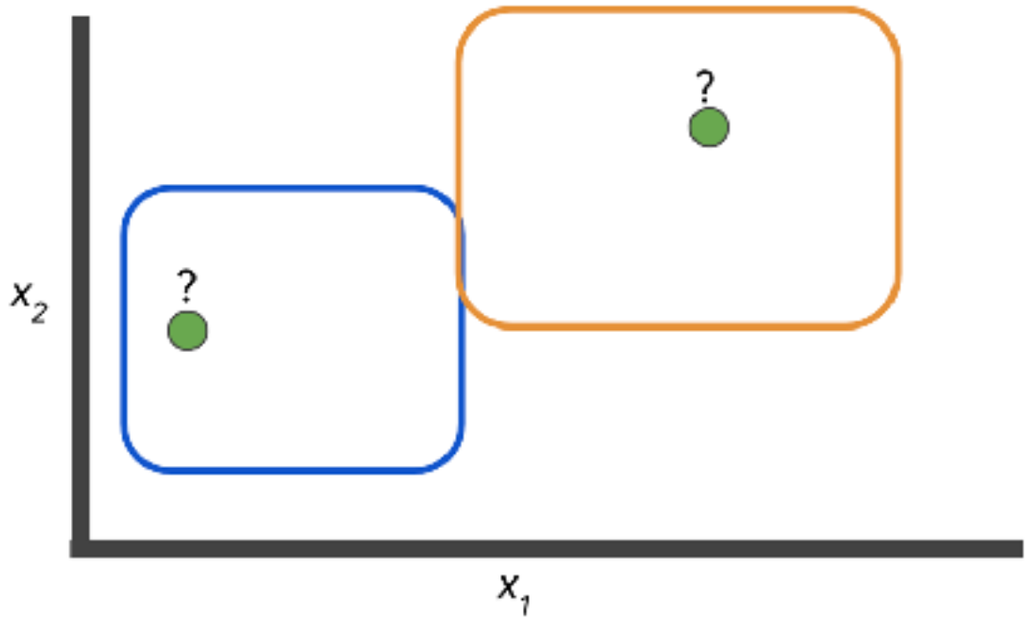
Unsupervised learning



UNSUPERVISED LEARNING: EXAMPLE







DIFFERENT TYPES OF ML LEARNING ALGORITHMS

Type of ML Problem	Description	Example
Classification	Pick one of N labels	Cat, dog, horse, or bear
Regression	Predict numerical values	Click-through rate
Clustering	Group similar examples	Most relevant documents (unsupervised)
Association rule learning	Infer likely association patterns in data	If you buy hamburger buns, you're likely to buy hamburgers (unsupervised)
Structured output	Create complex output	Natural language parse trees, image recognition bounding boxes
Ranking	Identify position on a scale or status	Search result ranking

THE ML MINDSET

"Machine Learning changes the way you think about a problem.

The focus shifts from a mathematical science to a natural science, running experiments and using statistics, not logic, to analyse its results."

- **Peter Norvig**

THE ML MINDSET

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems?
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

Get Comfortable with Some Uncertainty !

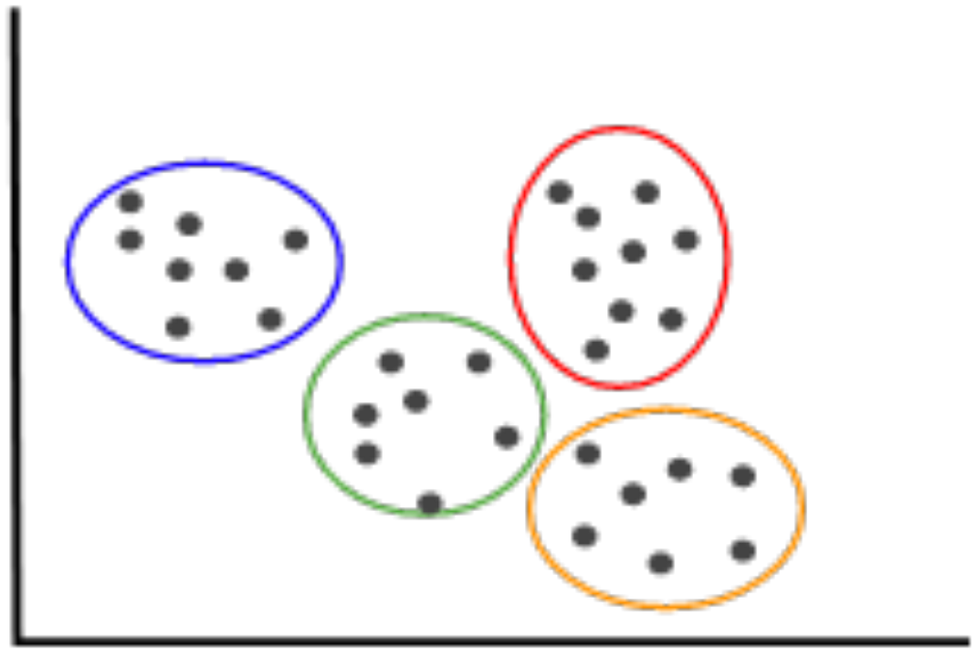
IDENTIFY GOOD PROBLEMS FOR ML

My product is facing a problem. Is it a good problem for ML?

- Understand the problem thoroughly. If you were to solve manually, how would you?
- Understand the different between data-driven patterns and problem-driven patterns
  - Could result in stereotypes or bias
- Typically, understand the data from 1000's of data points or hundreds of 1000's of data points
- Use your problem/domain knowledge - don't throw in all the features/data
- Deriving insights vs actionable decisions

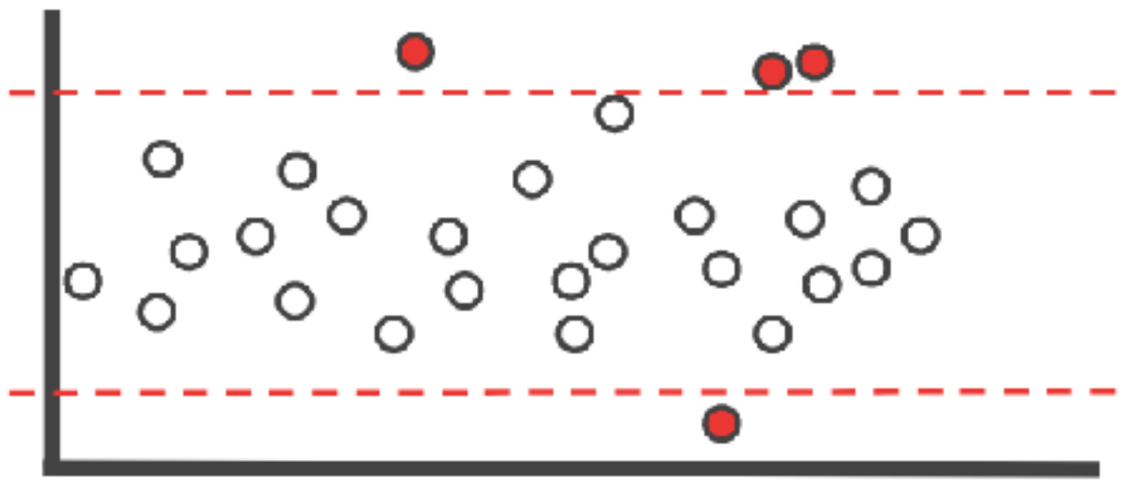
WHAT ARE SOME HARD ML PROBLEMS?

- Clustering
  - What does each cluster mean?
  - How do we make actionable insights from them?
  - Example: Crime scene fingerprint clustering



WHAT ARE SOME HARD ML PROBLEMS?

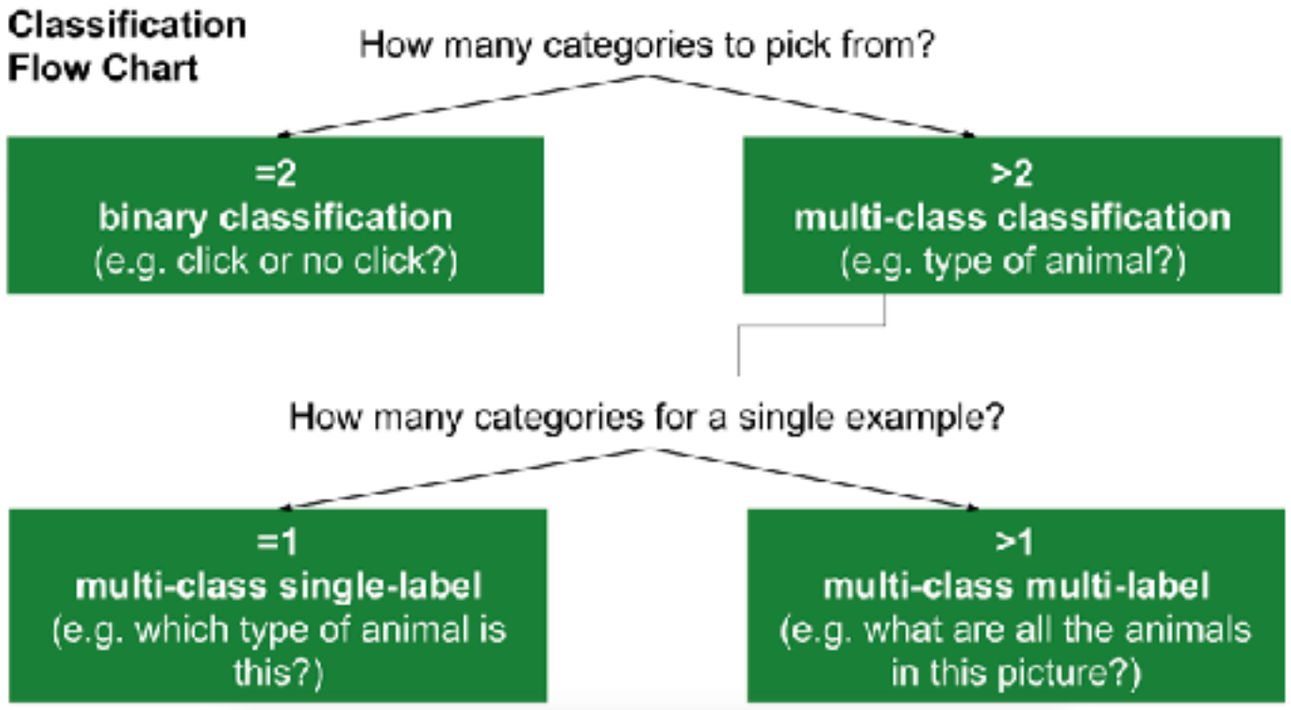
- Anomaly Detection
  - What constitute anomaly?
  - Can we use heuristics/rules? Will they be too easy?
  - Use a ML model to find out anomalous data points for another ML model

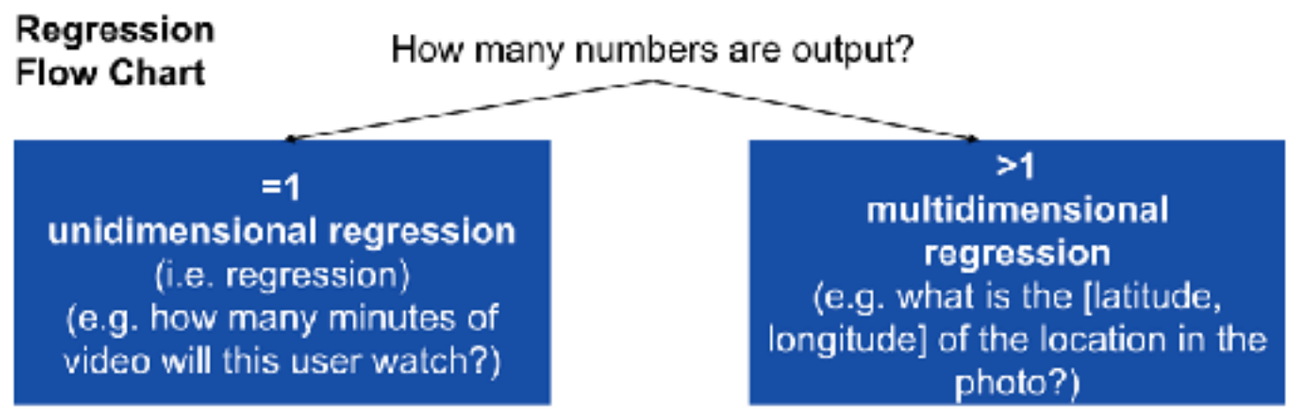


WHAT ARE SOME HARD ML PROBLEMS?

- Causation
  - Correlation does not mean causation!
  - Location and gender maybe correlated but may not be the cause!
  - How to identify the cause for the output, from observational data?

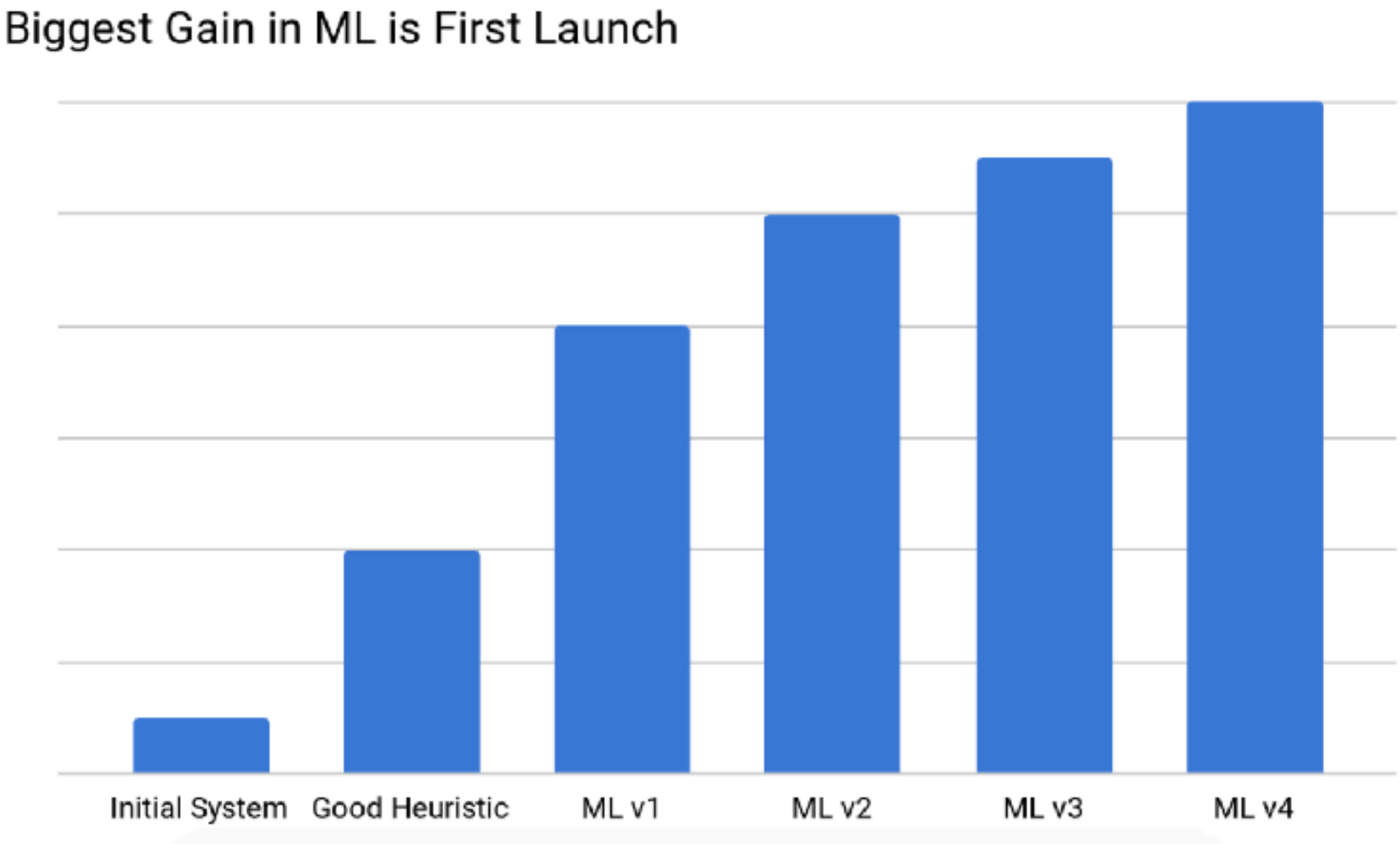
ML FORMULATION: 1. ARTICULATE YOUR PROBLEM





Example: Taste of the food

ML FORMULATION: 2. START SIMPLE



ML FORMULATION: 3. IDENTIFY YOUR DATA SOURCE AND DESIGN THE DATA

- How much of labelled data do we have? 100's, 1000's, tens of 1000's
- What is the source of your label? Manually labeled, crowd sourced, rule based
- Is the label closely related to the action we want to perform?
  - Label - how popular the video is (Very popular, popular, not popular)
  - Action - Show the video as recommendation
- Design the data for the model - <input, output>

Title	Channel	Upload Time	Uploader's Recent Videos	Output (label)
My silly cat	Alice	2018-03-21 08:00	Another cat video, yet another cat	Very popular
A snake video	Bob	2018-04-03 12:00	None	Not popular

ML FORMULATION: 4. DEFINE METRICS TO EVALUATE PERFORMANCE

- What are your metrics to evaluate the model's performance?
  - Accuracy, precision, time, happiness, #FB likes, (world peace?), (reply from ET?)
- Are the metrics measurable?
- Different different metrics for success and failure
- What does the ideal outcome look like?
- Can we collect some sample outputs to evaluate our model?

TASKS FOR THIS WEEK

- Write a blog post on:
  - Three different example applications from your day-to-day life
  - How can ML help in solving them?
  - How do you think you can use ML to solve them?
  - This is surely a non-technical blog, to see if you can formulate problems around you as ML problems!
  - The blog can be as small as one paragraph - Quantity doesn’t matter, quality does!

THANK YOU – NEXT WEEK

Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research