

# Course Overview

Dr. Nimisha Roy

Lecturer, College of Computing

# Your Instructor

## Instructor: Dr. Nimisha Roy

- 1<sup>st</sup> time teaching this course
- Lecturer in the College of Computing (SCI/OMSA)
- PhD in Computational Science & Engineering from Georgia Tech
- Research Interests: Computing Education, Gen AI and Sustainability Integrations, Software Development
- Hobbies: singing (jamming), interior designing, admiring my plants



# Your TAs



We have an army  
of ~45 TAs who  
help make this  
class successful



Check out the  
class website to  
know about them

Refer to:

# Class Website

For anything (updates, lectures, logistics, and so on) related to this class

# Some important notes:

- All communication must be via Ed (Chat and Threads). No Email please.
- Please read the website carefully. Website will be the first thing that we update if there any changes.
- Add HWs dues and Quizzes to your desired calendar.
- Please come to the class (required); it will help a lot ☺
  - Attendance is mandatory. Attendance will be taken few times in the semester. This genuinely comes from a place of care.

# Some important notes:

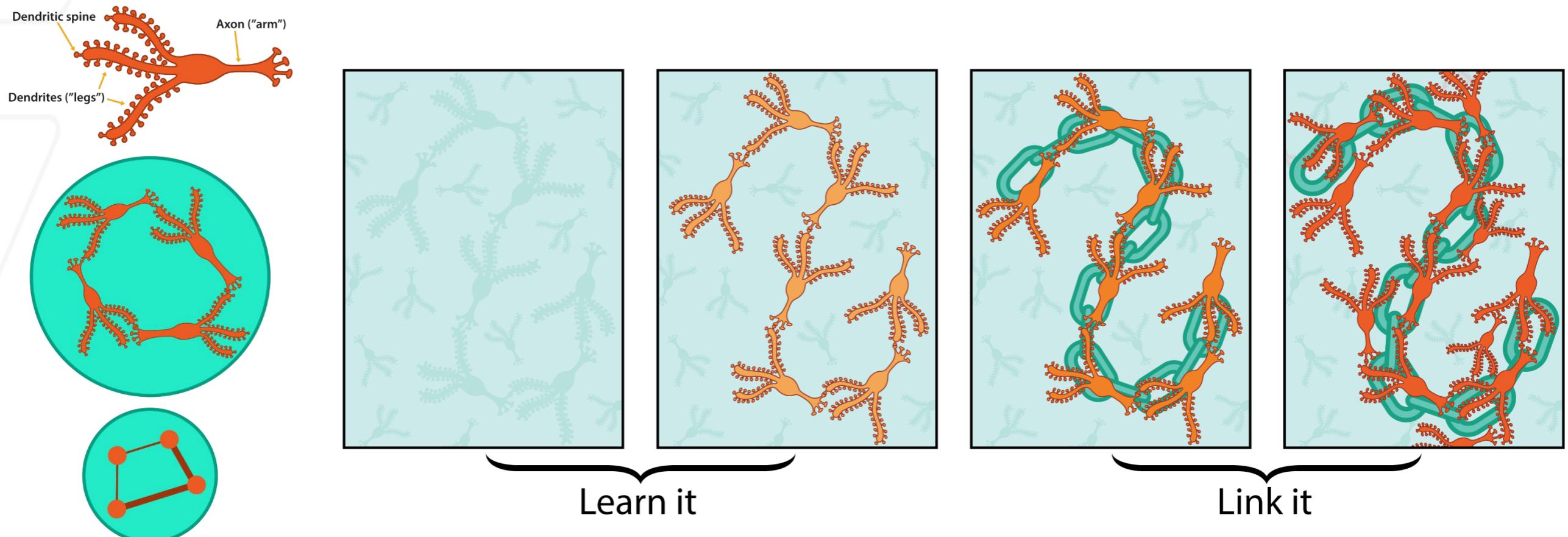
- ML is **NOT** a programming class.
- Lectures will be Math Heavy and HWs will be mostly programming. ML is all about Linear Algebra, Probability, Statistics and Optimization. You need to have both mathematical and programming skills to be successful in this area.
  - Please come to class and follow along.
  - Ask questions!
- HWs are substantial. Please start as early as possible. Don't procrastinate.

# Course Objectives

- Introduce to you the **pipeline of Machine Learning**
- Help you understand **major machine learning algorithms**
- Help you learn to **apply tools for real data analysis problems**
- Encourage you to **do research** in data science and machine learning

Before we talk about machine learning,  
let's talk about human learning...

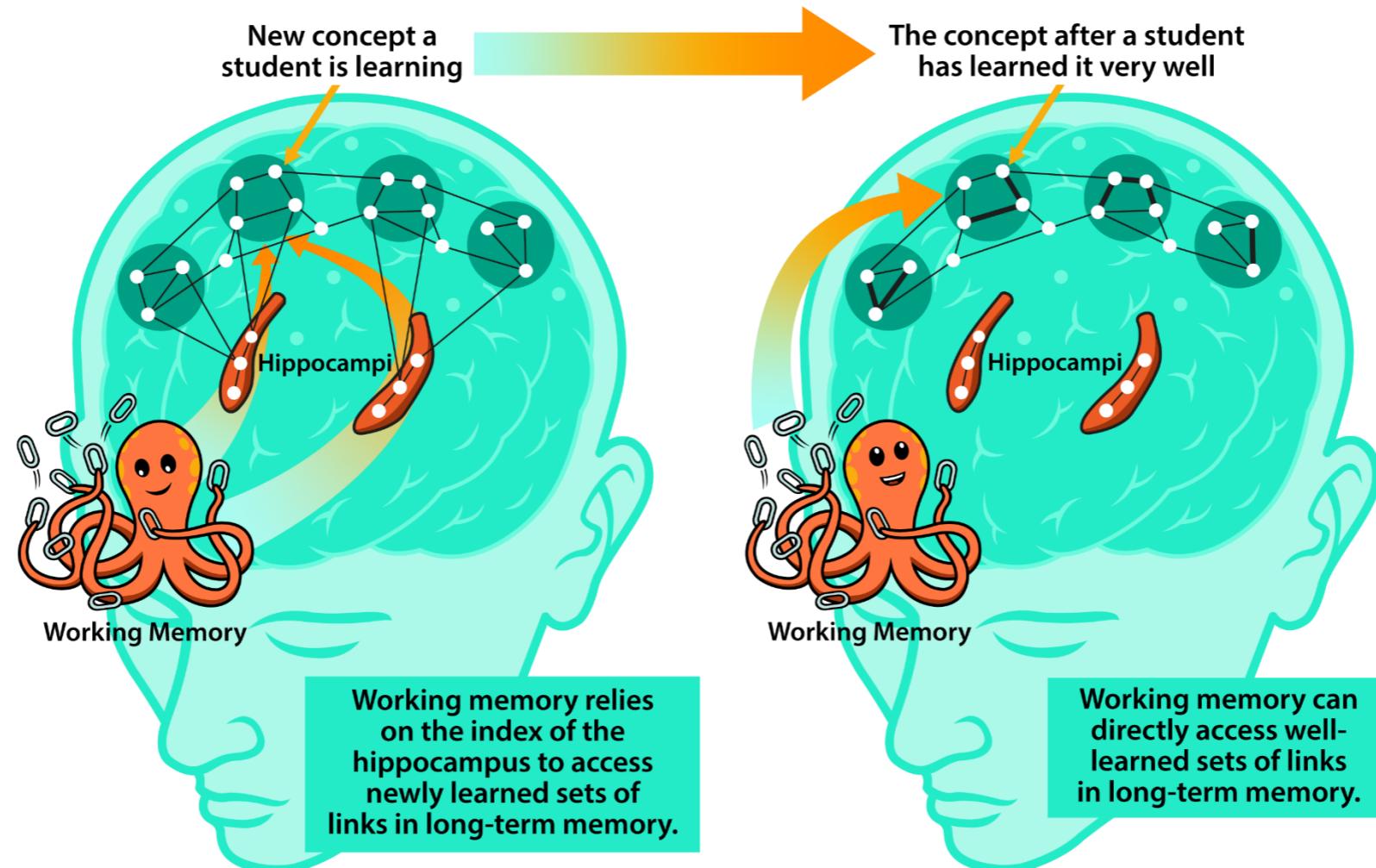
# Learning involves changing your brain!



Source: Oakley, Rogowsky & Sejnowski (2021)

9

# Declarative learning system



Source: Oakley, Rogowsky & Sejnowski (2021)

10

# Our approach to this course

- **Lectures** will introduce key theoretical concepts to enable you to understand different machine learning algorithms, maths behind them and their application
- **HWs** will help you deepen your understanding of topics covered in class and the implementation of machine learning algorithms
- **Project** will give you an opportunity to gain hands-on-experience in machine learning application to a real-world problem and expand beyond topics covered in class
- **Quizzes** will assess your understanding of the materials and motivate you to stay up to speed on topics covered in class

# Syllabus

## Part I: Basic math for computational data analysis

- Probability, statistics, linear algebra

## Part II: Unsupervised learning for data exploration

- Clustering analysis, dimensionality reduction, kernel density estimation

## Part III: Supervised learning for predictive analysis

- Tree-based models, linear classification/regression, neural networks

# Prerequisites

Basic knowledge in probability, statistics, and linear algebra

Basic programming skills in Python (Jupyter Notebook)

*No background in machine learning is required*

# Office hours

## Notes:

- 1) Each student may fill in their name and create a BlueJeans Meeting for the office hour
- 2) The signup policy is first come, first serve. **Please, DO NOT override others' sign-ups without their permissions.**
- 3) Because of the time limit, each student will have a maximum time of 10 minutes to work with the TA
- 4) Please come prepared with specific questions so we can help you more effectively

The slots were cleared on Nov 15th. Please add your name again if you had filled it for this week. Apologies for the inconvenience.

Time Slots	Student Name	Question of Interest (Please be more specific about the question you want to ask)	BlueJeans Address	Label
Jayanta's OH				
12:00--12:10				
12:10--12:20				
12:20--12:30				
12:30--12:40				
12:40--12:50				
12:50--13:00				
Waitlist				
	1			
	2			
	3			
Huili's OH				
14:30--14:40				
14:40--14:50				
14:50--15:00				
15:00--15:10				
15:10--15:20				
15:20--15:30				
Waitlist				
	1			
	2			

- You have just 10 minutes for each slot.
- Please be mindful of other students (you can't block multiple office hours, just one and we have the waitlist!)
- Make sure you pinpoint your problem exactly before joining the office hour.
- Office hour is not for general code debugging. You need to be specific about your question.

# Assignments

Four assignments (Submitted via GradeScope)  
Each can include written analysis or programming

**Start Early as soon as they are out**

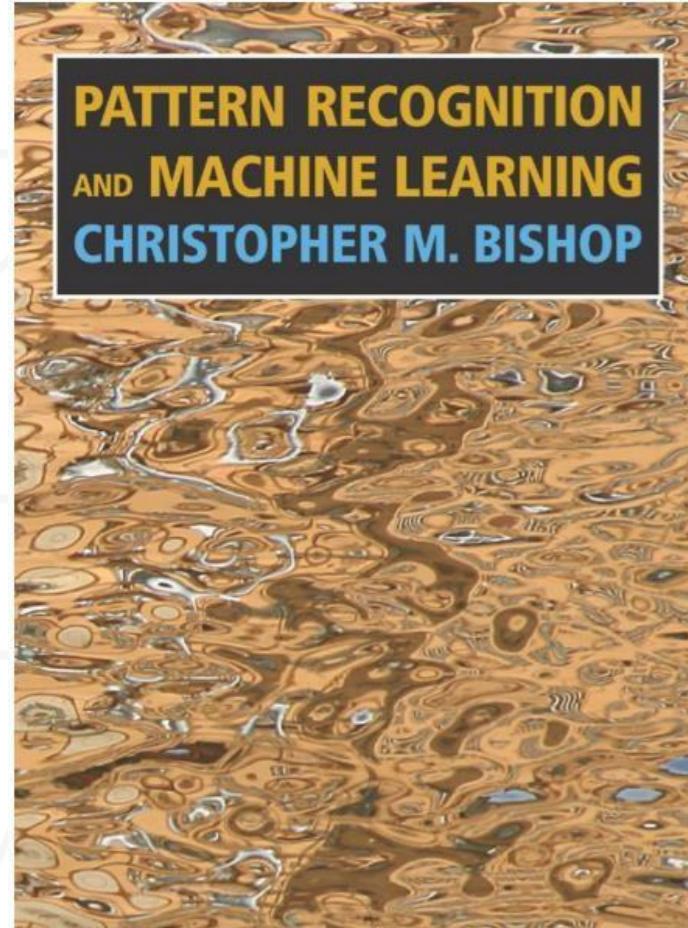
**Read late policy on the class website**

**Don't copy**

Because of the large size of our class, if we observe any (even small) similarity\plagiarisms detected by GradeScope or our TAs, WE WILL DIRECTLY REPORT ALL CASES TO OSI, which may unfortunately lead to a very harsh outcome.

# Projects

- Work on a real-life Machine Learning problem
  - What is the problem? What is your method? How do you evaluate it?
- Exactly 5 people in a team (Grad and undergrad can't be mixed in a group)
- Deploy via GitHub Pages (index.html)
- Start your projects early
- Ask for comments and feedbacks from the teaching staff



## Text Books

[Pattern Recognition and Machine Learning, by Chris Bishop](#)

Other recommended books:

[Learning from data](#), by Yaser S. Abu-Mostafa

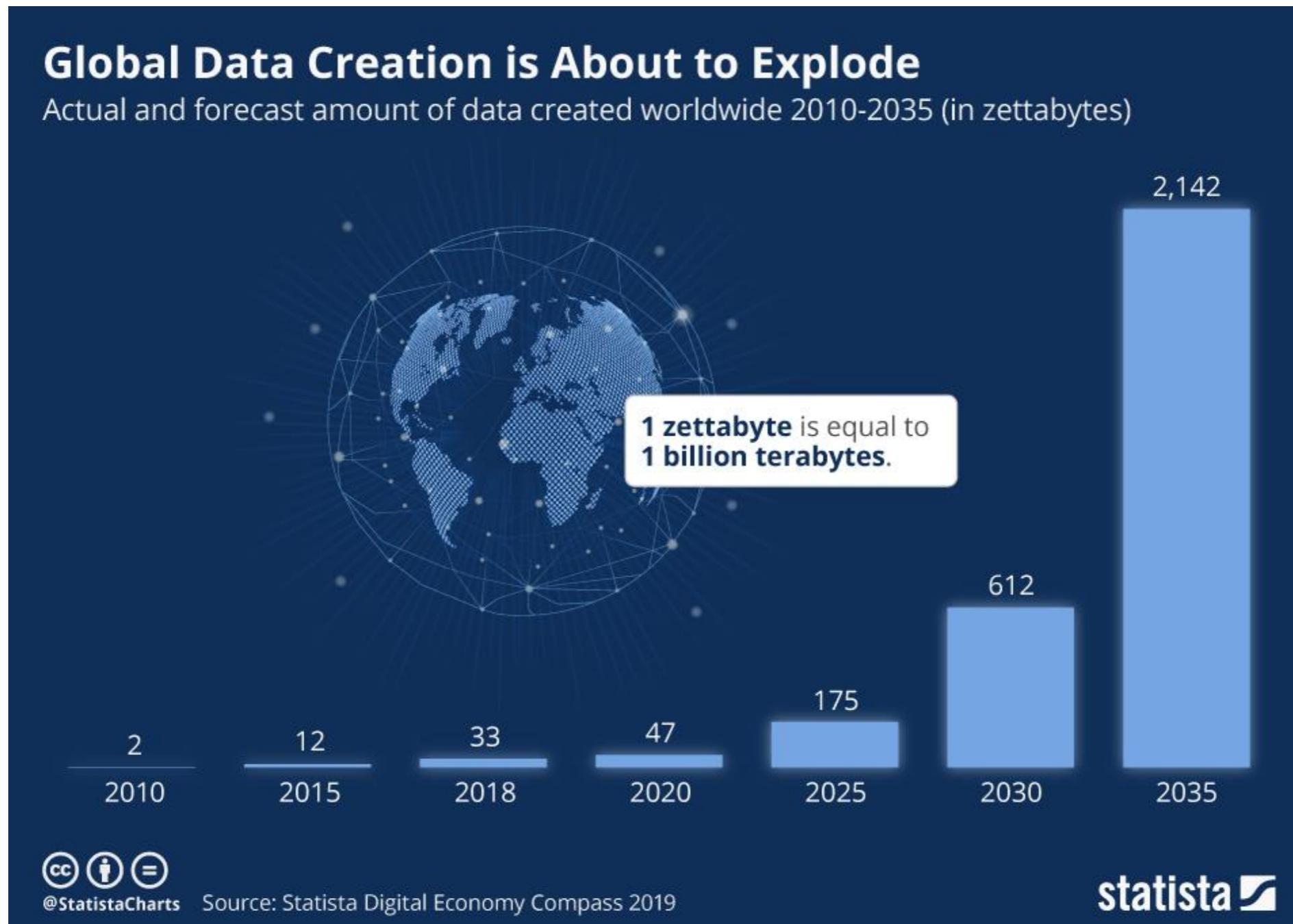
[Machine learning](#), by Tom Mitchell

[Deep Learning](#), by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

# Questions?

# Why Machine Learning?

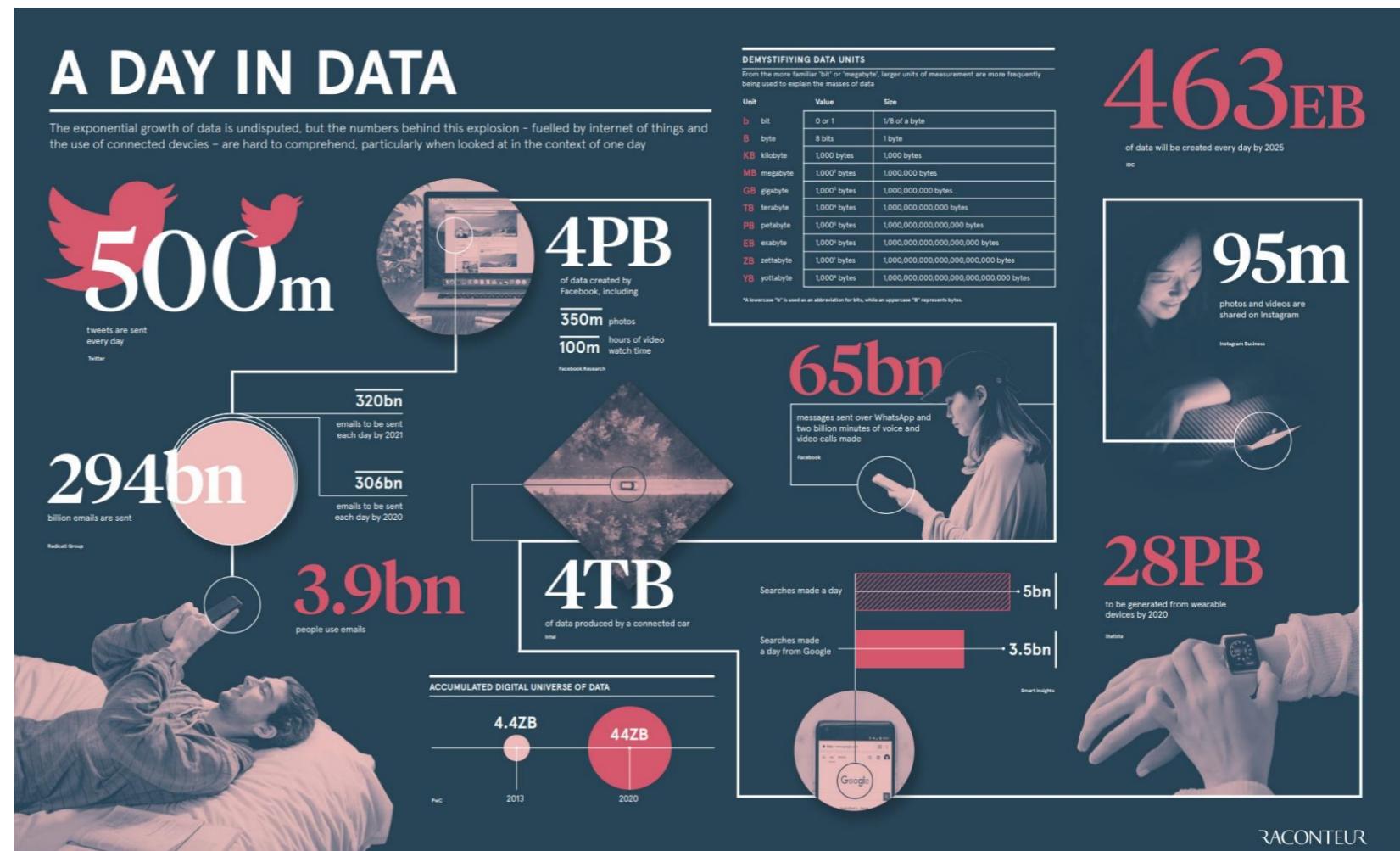
***"We are drowning in information but starved for knowledge."*** — John Naisbitt



33 zettabytes is equal to the storage capacity of 33 million human brains altogether

# The Booming Age of Data

- 30 trillion web pages
- 500 million tweets per day
- 2.7 billion monthly active users on Facebook
- 1.8 billion images uploaded per day
- 2.9 billion base pairs in human genome



## Levels of AGI

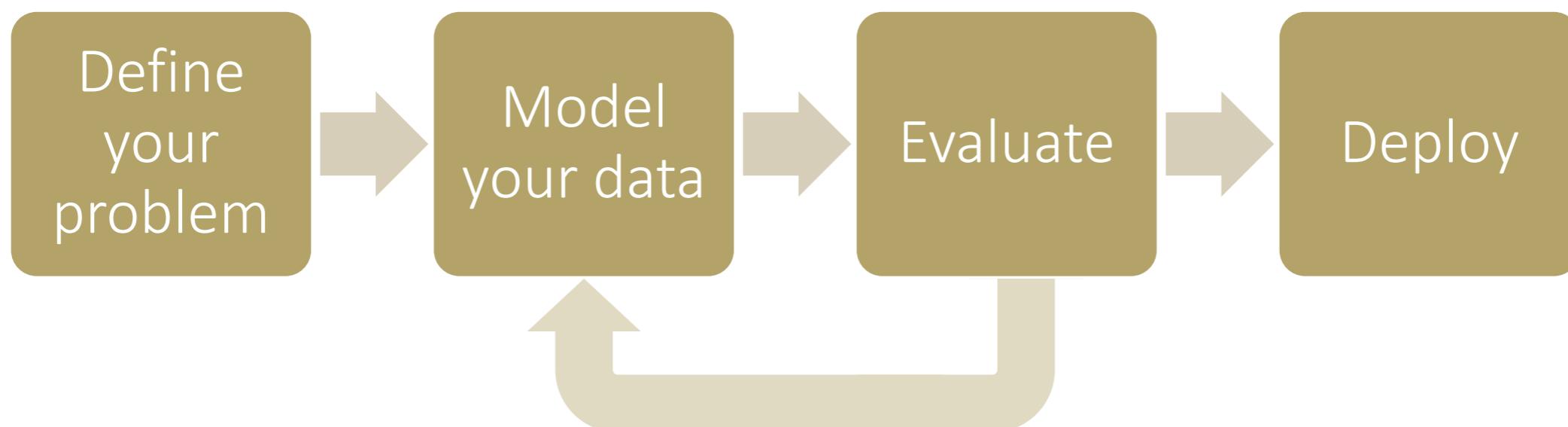
Performance (rows) x Generality (columns)	Narrow <i>clearly scoped task or set of tasks</i>	General <i>wide range of non-physical tasks, including metacognitive abilities like learning new skills</i>
<b>Level 0: No AI</b>	<b>Narrow Non-AI</b> calculator software; compiler	<b>General Non-AI</b> human-in-the-loop computing, e.g., Amazon Mechanical Turk
<b>Level 1: Emerging</b> <i>equal to or somewhat better than an unskilled human</i>	<b>Emerging Narrow AI</b> GOFAI <sup>4</sup> ; simple rule-based systems, e.g., SHRDLU ( <a href="#">Winograd, 1971</a> )	<b>Emerging AGI</b> ChatGPT ( <a href="#">OpenAI, 2023</a> ), Bard ( <a href="#">Anil et al., 2023</a> ), Llama 2 ( <a href="#">Touvron et al., 2023</a> )
<b>Level 2: Competent</b> <i>at least 50th percentile of skilled adults</i>	<b>Competent Narrow AI</b> toxicity detectors such as Jigsaw ( <a href="#">Das et al., 2022</a> ); Smart Speakers such as Siri ( <a href="#">Apple</a> ), Alexa ( <a href="#">Amazon</a> ), or Google Assistant ( <a href="#">Google</a> ); VQA systems such as PaLI ( <a href="#">Chen et al., 2023</a> ); Watson ( <a href="#">IBM</a> ); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	<b>Competent AGI</b> not yet achieved
<b>Level 3: Expert</b> <i>at least 90th percentile of skilled adults</i>	<b>Expert Narrow AI</b> spelling & grammar checkers such as Grammarly ( <a href="#">Grammarly, 2023</a> ); generative image models such as Imagen ( <a href="#">Saharia et al., 2022</a> ) or Dall-E 2 ( <a href="#">Ramesh et al., 2022</a> )	<b>Expert AGI</b> not yet achieved
<b>Level 4: Virtuoso</b> <i>at least 99th percentile of skilled adults</i>	<b>Virtuoso Narrow AI</b> Deep Blue ( <a href="#">Campbell et al., 2002</a> ), AlphaGo ( <a href="#">Silver et al., 2016, 2017</a> )	<b>Virtuoso AGI</b> not yet achieved
<b>Level 5: Superhuman</b> <i>outperforms 100% of humans</i>	<b>Superhuman Narrow AI</b> AlphaFold ( <a href="#">Jumper et al., 2021</a> ; <a href="#">Varadi et al., 2021</a> ), AlphaZero ( <a href="#">Silver et al., 2018</a> ), StockFish ( <a href="#">Stockfish, 2023</a> )	<b>Artificial Superintelligence (ASI)</b> not yet achieved

# Has machine learning touched your life today?

- The toothbrush you used this morning
  - What you ate for breakfast
  - The device you used to track your physical activity
  - When you asked your phone to set a timer for your food in the oven
  - The message you received from a friend on social media
- 
- **...machine learning was involved in all these activities directly or indirectly**

# Machine learning in practice

Machine learning is the process of turning data into actionable knowledge for task support and decision making.



# Brief History of Machine Learning

## 1950s

Samuel's checker player  
Selfridge's Pandemonium

## 1960s

Neural networks: Perceptron  
Pattern recognition  
Learning in the limit theory  
Minsky and Papert prove limitations of perceptron

## 1970s

Symbolic concept induction  
Winston's arch learner  
Expert systems and the knowledge acquisition bottleneck  
Quinlan's ID3  
Michalski's AQ and soybean diagnosis  
Scientific discovery with BACON  
**Mathematical discovery with AM (Automated Mathematician)**

# Brief History of Machine Learning

## 1980s

Advanced decision tree and rule learning  
Explanation-based Learning (EBL)  
Learning and planning and problem solving  
Utility problem  
Analogy  
Cognitive architectures  
Resurgence of neural networks  
(connectionism, backpropagation)  
Valiant's PAC Learning Theory  
Focus on experimental methodology

## 1990s

Data mining  
Adaptive software agents and web applications  
Text learning  
Reinforcement learning (RL)  
Inductive Logic Programming (ILP)  
**Ensembles: Bagging, Boosting, and Stacking**  
Bayes Net learning

# Brief History of Machine Learning

## 2000s

Support vector machines  
Kernel methods  
Graphical models  
Statistical relational learning  
Transfer learning  
Sequence labeling  
Collective classification and structured outputs  
Computer Systems Applications  
Learning in robotics and vision

## 2010s

Deep learning  
Reinforcement learning  
Generative models  
Adversarial learning  
Multi-task learning  
Transfer learning  
Learning in NLP, CV, Robotics, ...

## 2020s

What will your contribution be?

# Unsupervised and Supervised learning

dimension or attribute or feature

datapoint

Point 1

Point 2

Point 3

Point 4

Point 5

	Weight(lb)	Height(cm)	Fur color	Eye color	Label
Point 1	10	20	w	g	<u>cat</u>
Point 2	50	100	br	bl	dog
Point 3	8	15	bl	bl	dog
Point 4	12	25	w	bl	cat
Point 5	14	10	bl	g	dog

$X_{n \times d}$

$Y_{n \times 1}$

training data

EDA  $\rightarrow$

Unsupervised just focuses on  $X_{n \times d}$

testing data  $\rightarrow X [ ]$   
 ~~$n \times 1 \times d$~~

Supervised focus on  $X_{n \times d}$  and  $Y_{n \times 1}$

# We can do better than Cat and Dog

	Weight(lb)	Height(cm)	Fur color	Eye color	Label
Point 1	10	20	w	g	Blob Fish
Point 2	50	100	br	bl	opossum
Point 3	8	15	bl	bl	opossum
Point 4	12	25	w	bl	Blob Fish
Point 5	14	10	bl	g	opossum

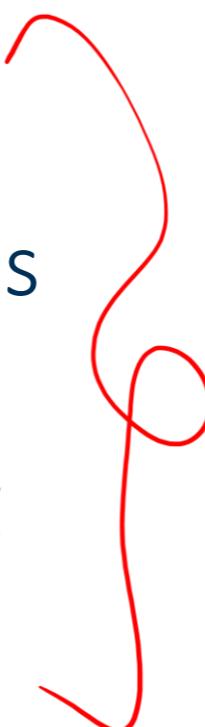
$X_{n \times d}$        $Y_{n \times 1}$



# Syllabus: Unsupervised Learning

## Clustering Analysis

- K-means
- Gaussian Mixture Models
- Hierarchical clustering
- Density-based clustering
- Clustering evaluation



## Probability distributions

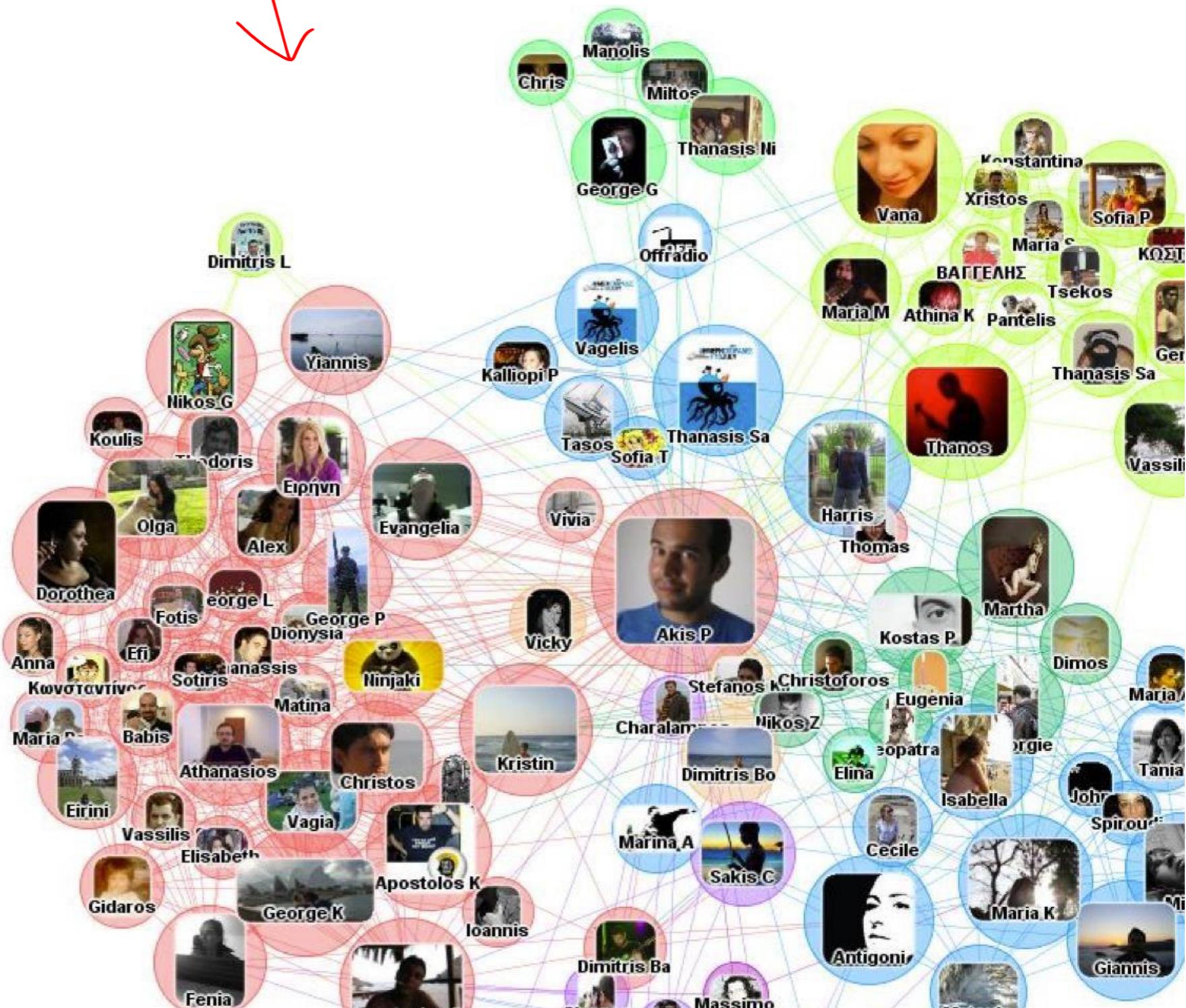
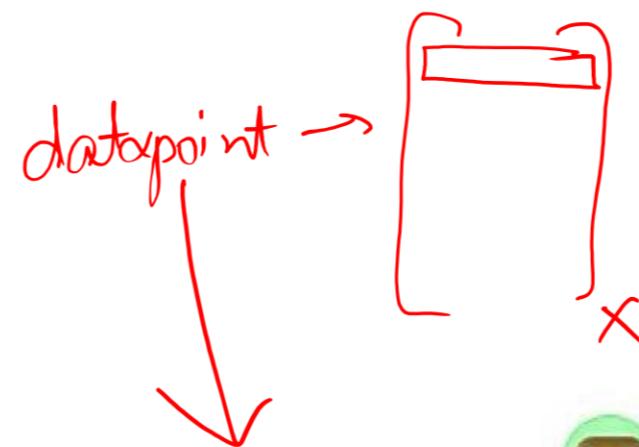
- Kernel density estimation
- Parametric density estimation
- Non-parametric density estimation

## Dimensionality reduction

- Principal component analysis

# Community Detection in Social Networks

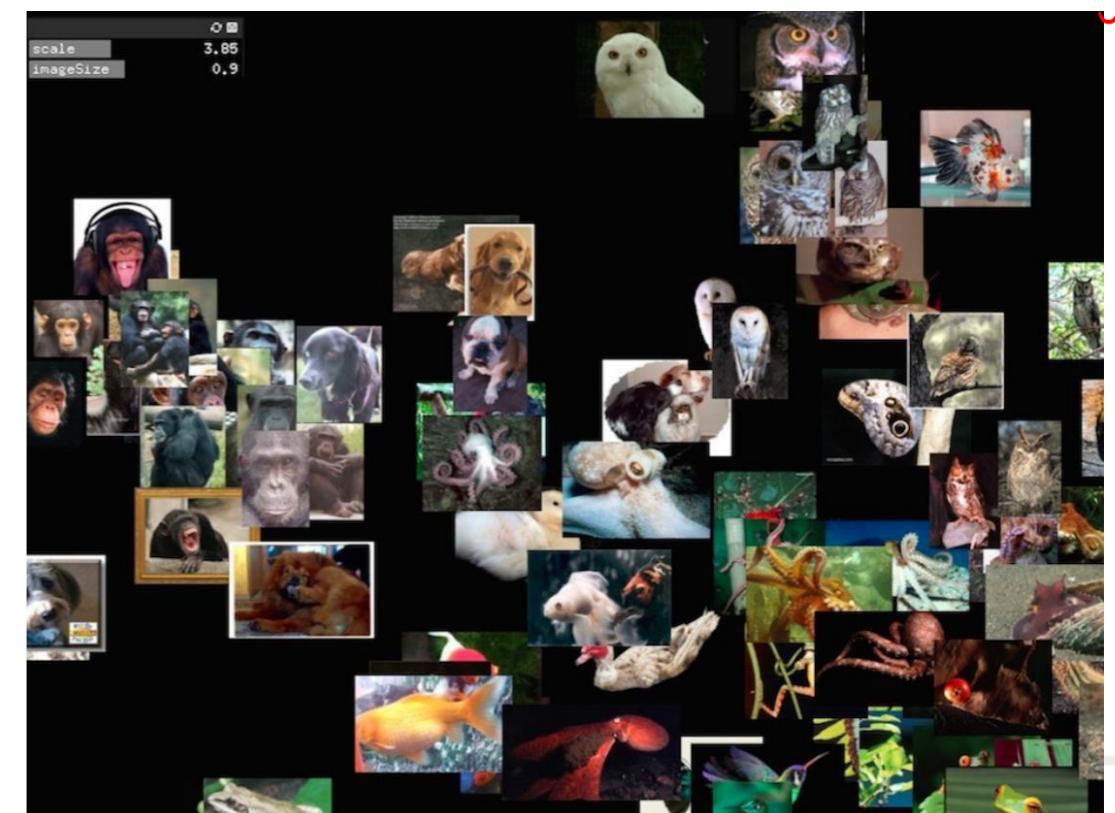
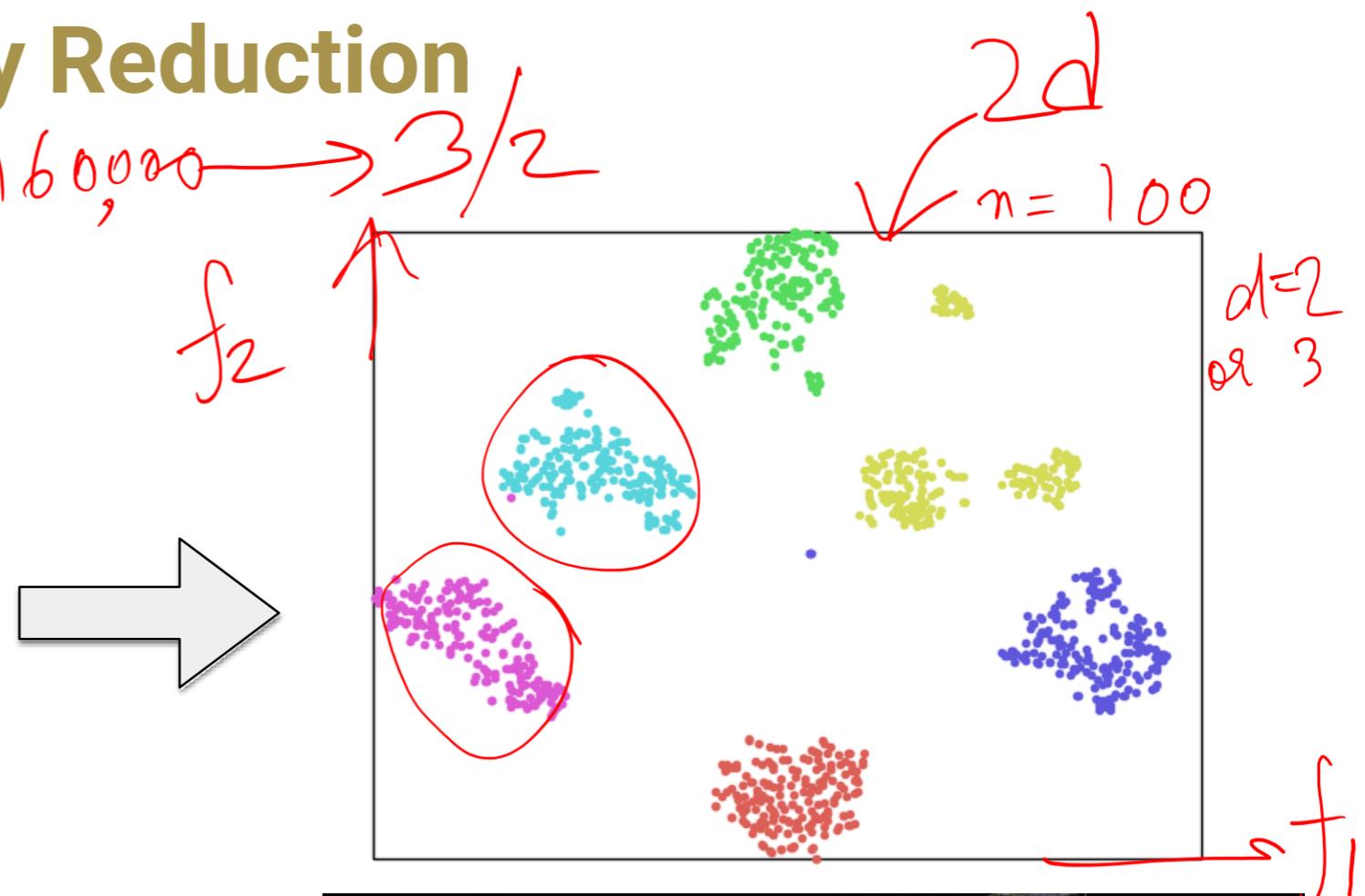
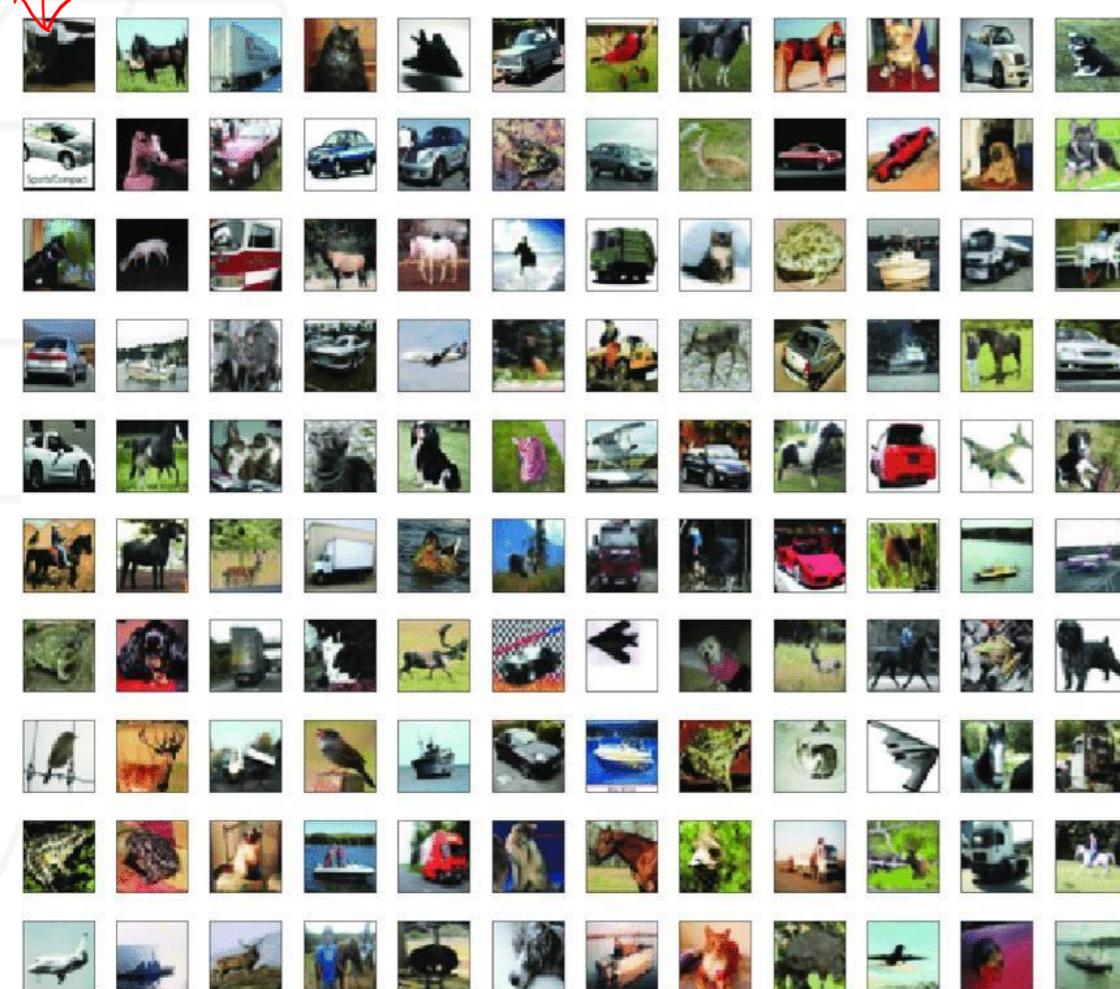
- What are the inputs and how to represent them?
- What are the desired outputs?
- What learning algorithms to choose?



$400 \times 400$  pixels

# Dimensionality Reduction

datapoint  $\rightarrow X \cdot [d]$   $d = 160,000 \rightarrow 3/2$



- What are the inputs and how to represent them? ~~✓~~
- What are the desired outputs? ~~✓~~ reduce  $d$  to 2 or 3
- What learning algorithms to choose?

# Syllabus: Supervised Learning

## Tree-based models

- Decision tree
- Ensemble learning/Random forest

## Linear classification/regression models

- Linear regression
- Naive Bayes
- Logistic regression

## Neural networks

- Feedforward neural networks and backpropagation analysis
- CNN

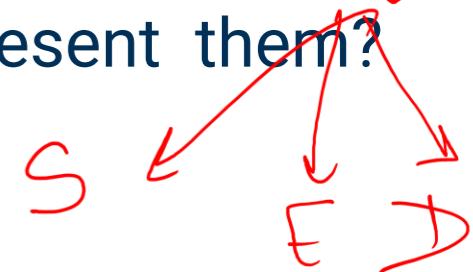
## Support vector machine

# News Classification

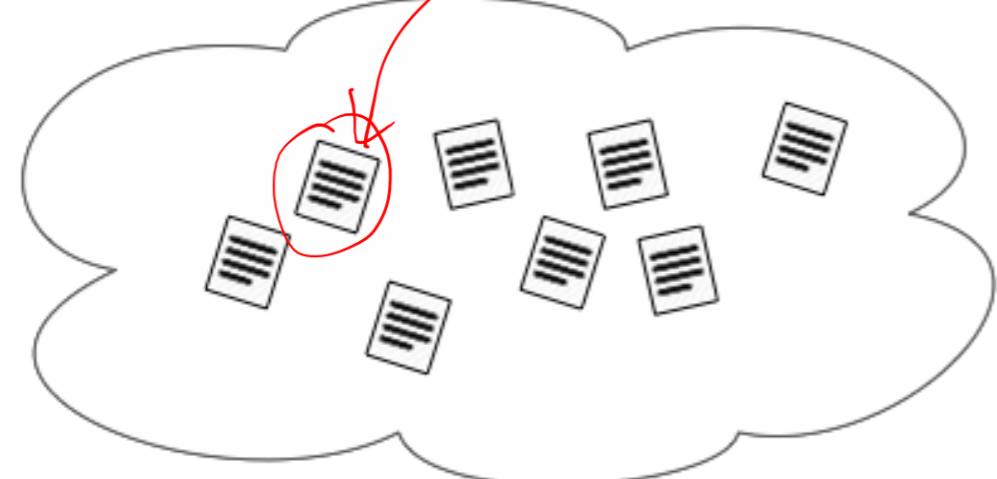


test set → a new article → trained model

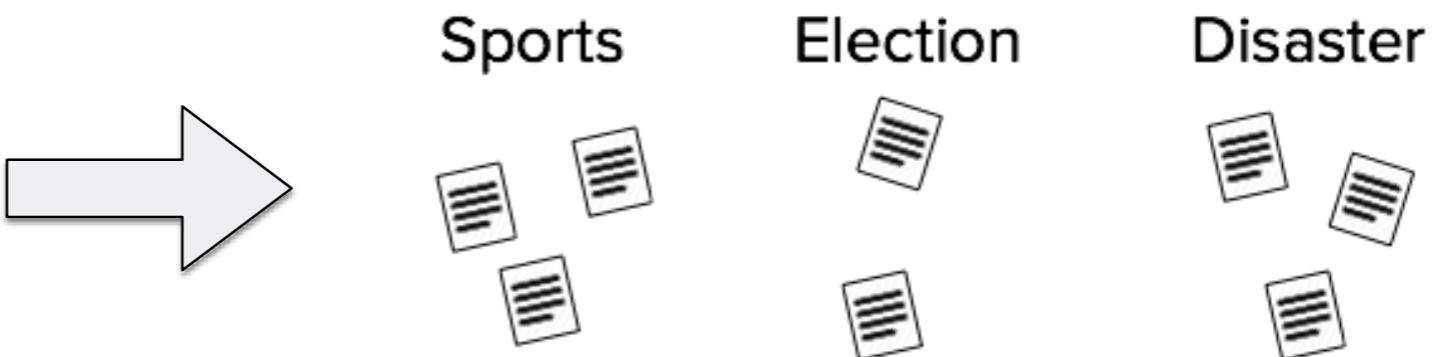
- What are the inputs and how to represent them?
- What are the desired outputs?
- What learning algorithms to choose?



1 datapoint



$$X = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} = \begin{bmatrix} S \\ E \\ D \end{bmatrix} \xrightarrow{\text{train}}$$



# Spam Detection

Google in:spam

Gmail • More

Compose

Inbox (994)

Starred

Sent Mail

Drafts

Less

Important

Customer Service

Sherley Rhoda

Customer Service

Healthy Living

Shiningltd Team

wikiHow Community Team (2)

FreeLotto

Good Fella's

Jason Squires

Bunnings

You still have product(s) in your basket - Healthy Living Lifestyle Pre  
From Sherley Rhoda

Activate your favorite videotraining service - Your activation code is re

We have added your shopping credits today - Healthy Living & Co. I

15 inch wifi Android OS tablet pc - SHININGLTD Our Alibaba Shop C

Congratulations on your article's first Helpful Vote! - Congratulations! A I

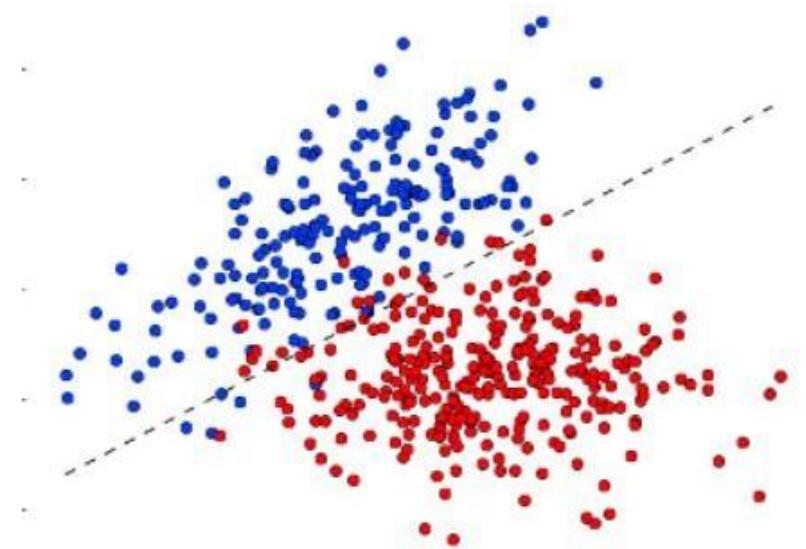
Jesse, NOTICE of FORFEITURE - Do not ignore! - NEVER miss an i

Our team assigned you to receive our new phone - Good Fella's Au

Make 2018 your best year yet - Hi there, Hope you're well, and have h

January arrivals - Image Congratulations Jesse Eaton! We have a very

NOT SPAM



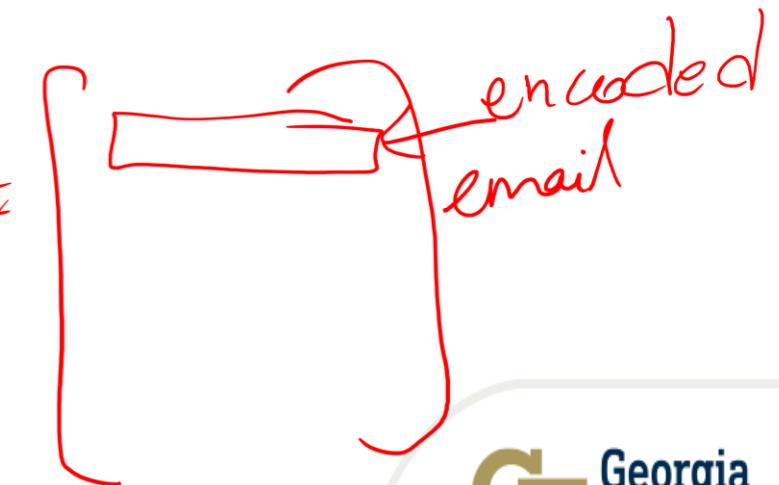
SPAM

- What are the inputs and how to represent them?

- What are the desired outputs?

$S_{\theta\theta} \text{ NS}$

- What learning algorithms to choose?



# Questions?

# Projects

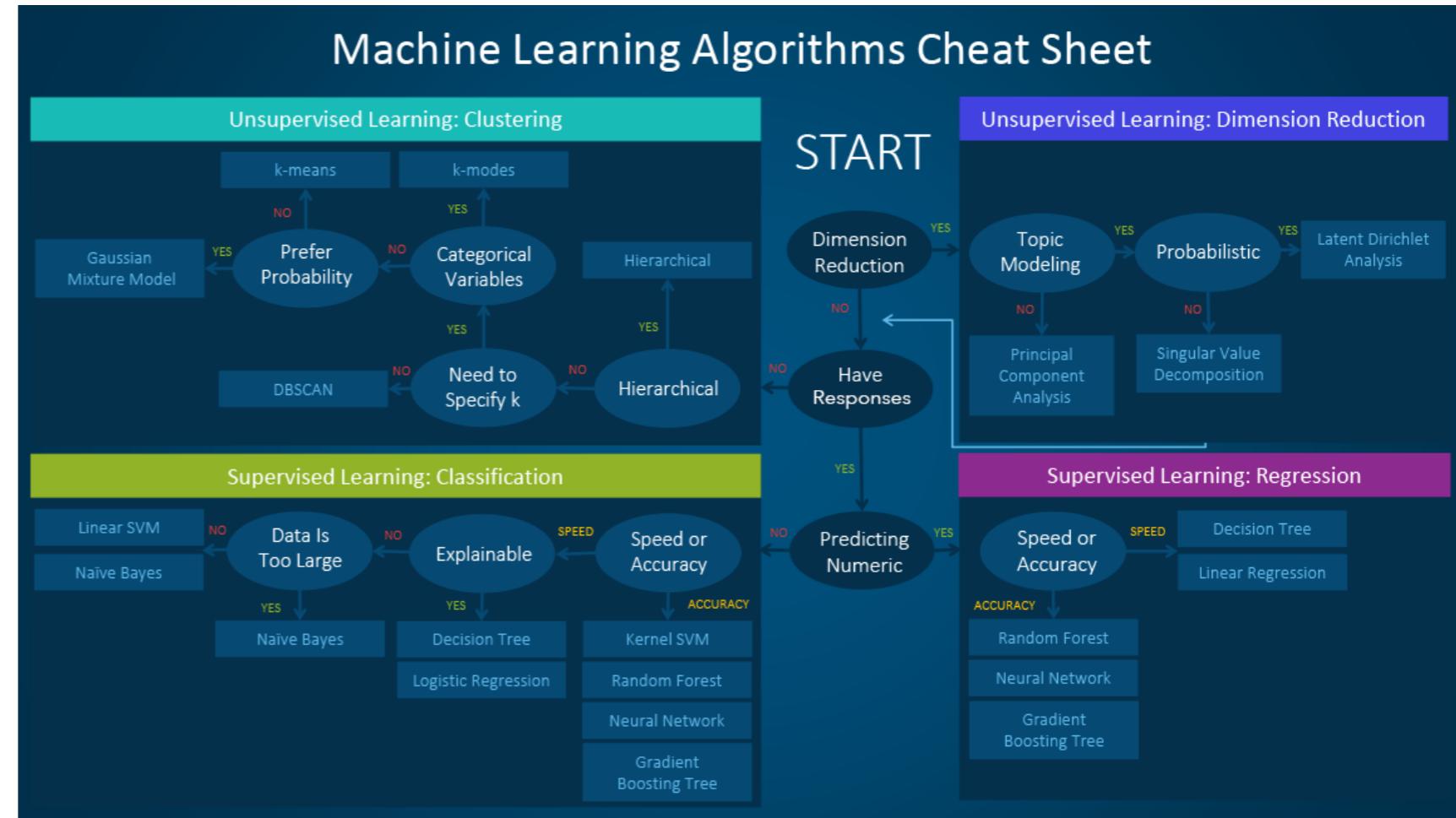
# Before you start a machine learning project

- **Defining the problem. Is this a machine learning task?**
- **What kind of machine learning task is it?**
  - Clustering, distribution estimation, classification, regression, other?
- **Do I have the data and resources to support it?**
- **What kind of data am I working with?**
  - Spatial (map, trajectory), visual (images), text (documents, tweets, customer reviews), behavioral data (smoking habits), time-series data (stock prices)
- **How am I evaluating success?**

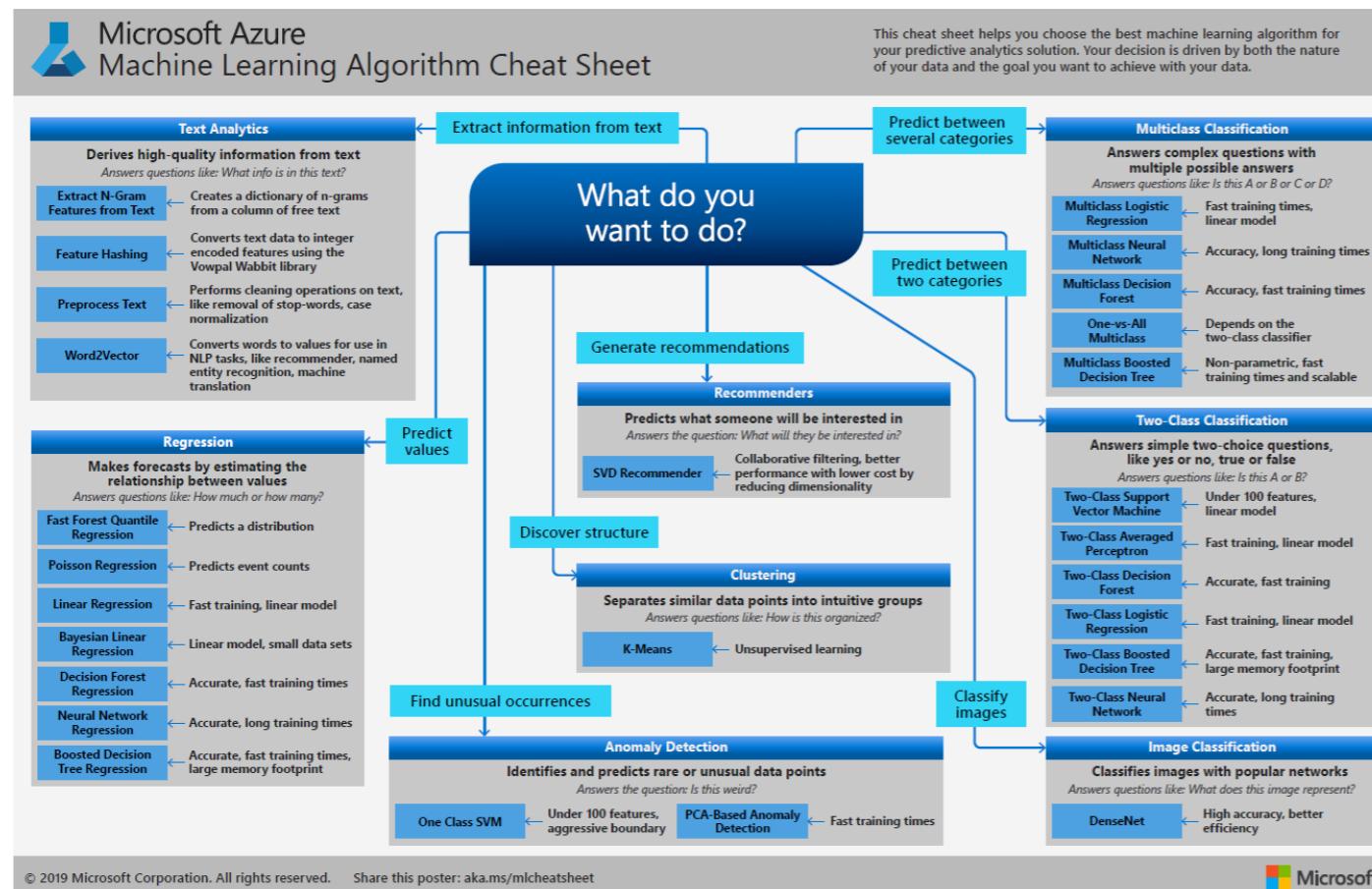
# Machine learning workflow process

- **Defining the problem:** Clearly define what you want to achieve with your model.
- **Data collection:** Gather and preprocess the data that you will use to train your model.
- **Data exploration:** Explore the data to gain insights and detect patterns, outliers, and anomalies.
- **Feature engineering:** Create features that capture relevant information from the data.
- **Model selection:** Choose a model that suits your problem, considering factors such as computational requirements and prediction accuracy.
- **Model training:** Train the model using the preprocessed data.
- **Model evaluation:** Evaluate the model's performance on a held-out test dataset.
- **Model tuning:** Fine-tune the model's hyperparameters to improve its performance.

# What algorithm should I use? (Not a comprehensive list)



Infographic by SAS Data Science Blog. Click on the figure to access the source.



## General Project Guidance on Class Website

# Logistics and teamwork advice

- Define clear means of communication (I strongly recommend you create a channel for your group on MS Teams)
- Listen to your teammates and be upfront about your availability
- Play to your individual strengths when tackling an activity
- Create an inclusive environment in your team and make sure all voices are heard
- Resolving conflicts within the team is part of the job
- The people you work with are part of your professional network
- Peer reviews will be used to assess your participation in the project

# Important aspects to consider

- Complex tasks demand large datasets in order to achieve satisfactory generalization. The course webpage has an extensive list of databases from which you can obtain datasets to work on your project
- The training phase of techniques involving large datasets and/or deep learning architectures are computationally intensive and require GPUs to be performed in a reasonable amount of time.
- Make sure you have the appropriate resources when dealing with such techniques. Here are some options of free GPU resources:
  - [Colab](#)
  - [Kaggle](#)
  - [AWS Educate](#)

# Example Projects applying these concepts is on our class website

- **Sample Projects**
  - Sample Project from previous semester [[Undergrad Canvas Access for previous ML projects](#)]; [[Grad Canvas Access for previous ML projects](#)]; [Stanford Project Examples](#);
- **General project guidance**
  - Your project will be graded based on the following criteria:  
**Was the motivation clear?**
    - What is the problem?
    - Why is it important and why we should care?  
**Were the dataset and approach used effectively?**