Machine Learning CS 4641 C

Course Overview

Nimisha Roy

Lecturer, College of Computing

Slides adapted from Max Roozbahani and Rod Borela

Your Instructor

Instructor: Dr. Nimisha Roy

- 1st 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

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 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.

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:

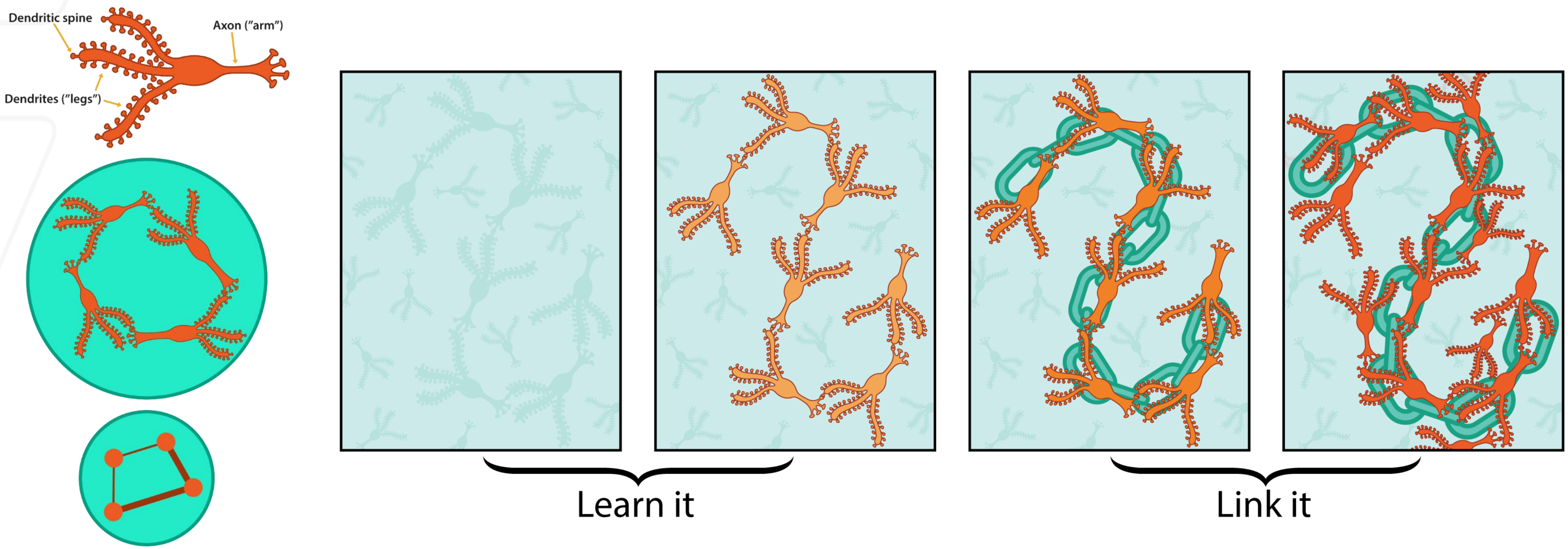
- ML is **NOT** a programming class.
- Lectures will be Math Heavy and HWs will be mostly programming. ML is all about Linear Algebra, Probability, Statistics and Optimization. You need to have both mathematical and programming skills to be successful in this area.
  - Please come to class and follow 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

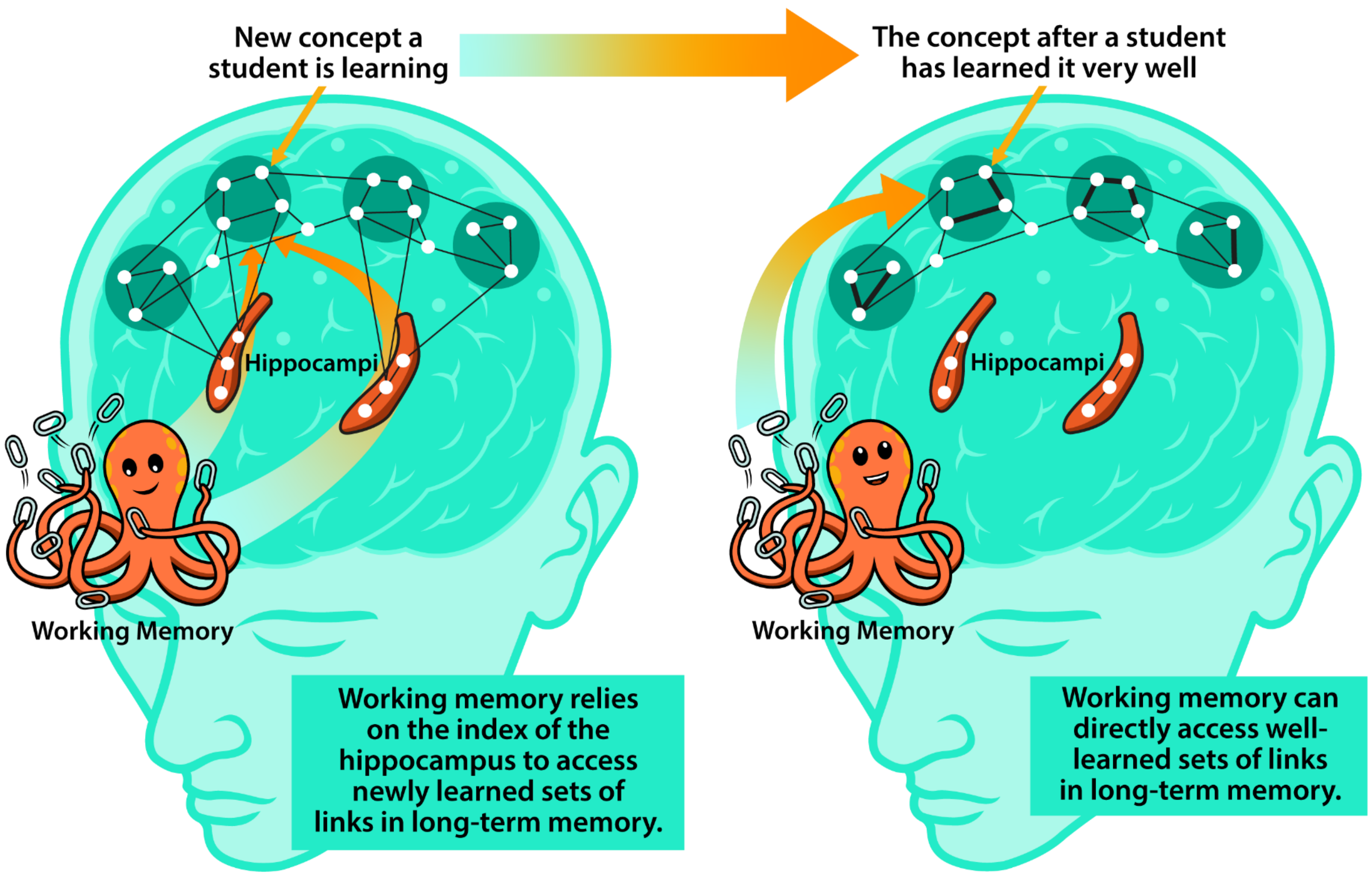
Before we talk about machine learning,
let’s talk about human learning…

Learning involves changing your brain!



Source: Oakley, Rogowsky & Sejnowski (2021)

Declarative learning system



Source: Oakley, Rogowsky & Sejnowski (2021)

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
- 1 in class assessment in Vocareum

Georgia Tech

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

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

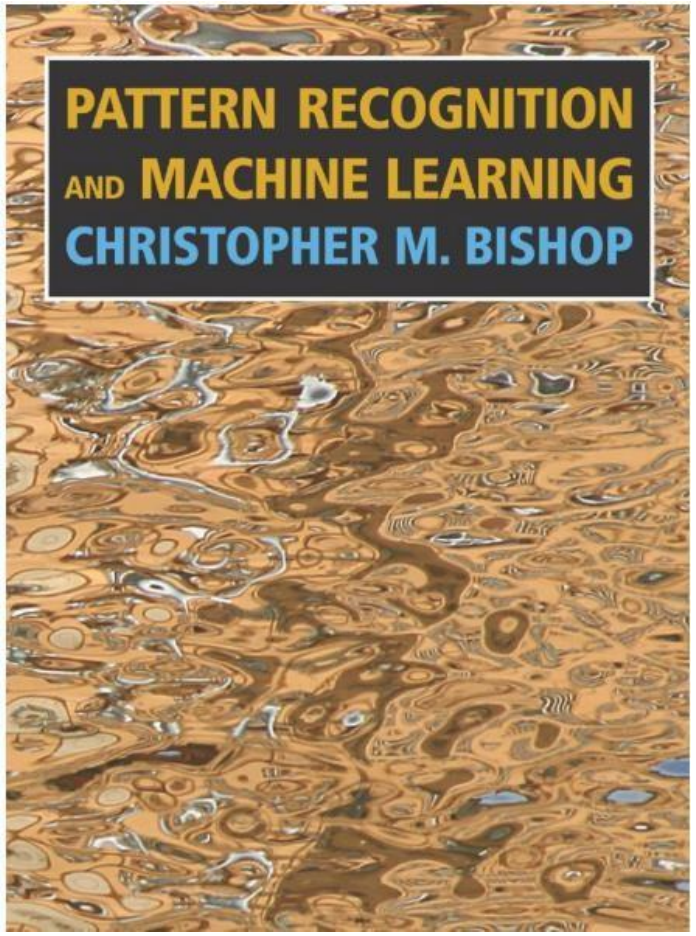
Because of the large size of our class, if we observe any (even small) similarity\plagiarisms detected by GradeScope or our TAs, WE WILL DIRECTLY REPORT ALL CASES TO OSI, which may unfortunately lead to a very harsh outcome.

Quizzes

- 14% of your grade
- Quiz 0 is 1% of your grade. Covers information about class website, course structure etc.
- Quiz 1 onwards covers lecture materials covered in class
- Honorlock based, open notes from canvas
- Can take from anywhere when quiz is open.

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

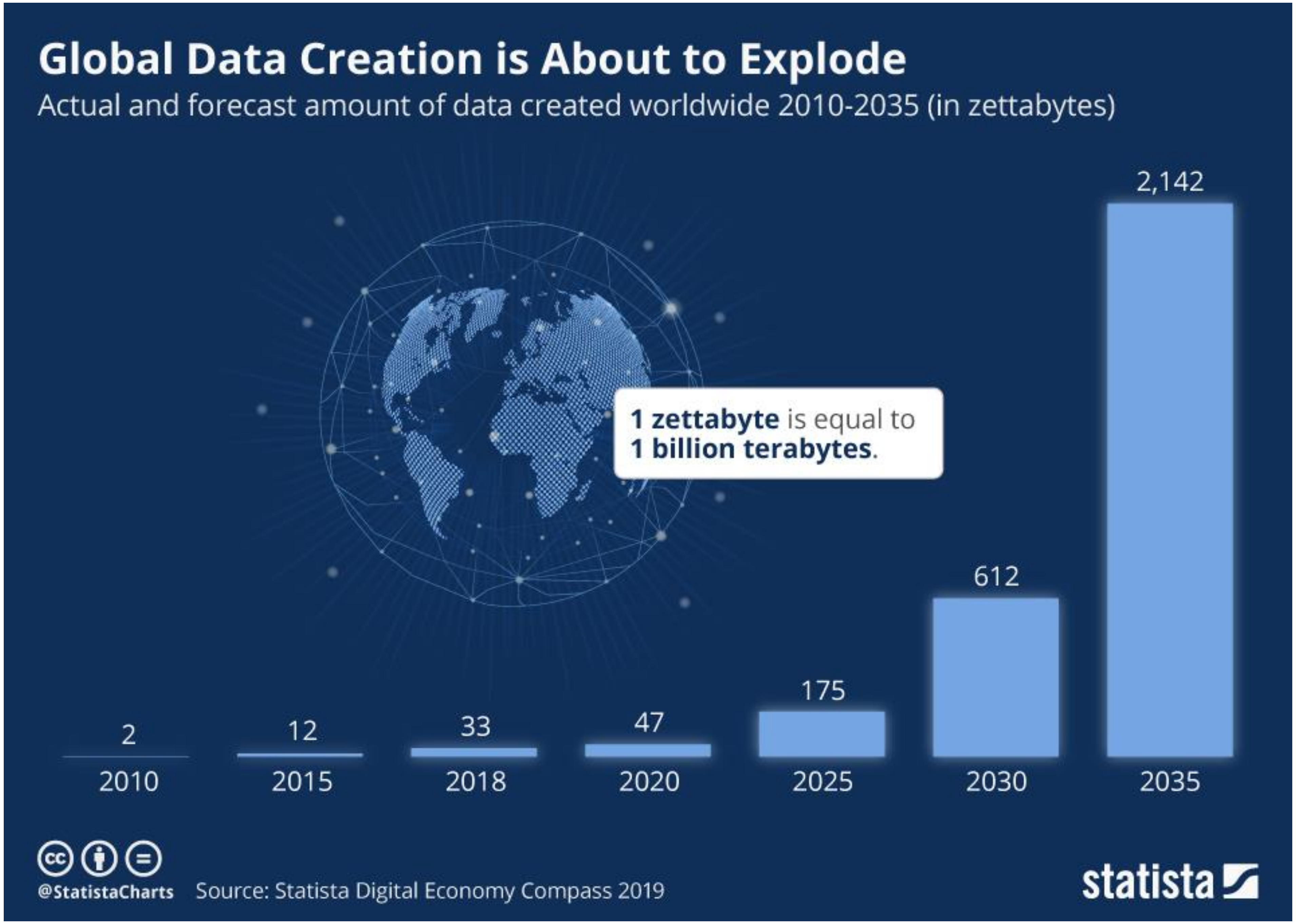
Machine learning, by Tom Mitchell

Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Questions?

Why Machine Learning?

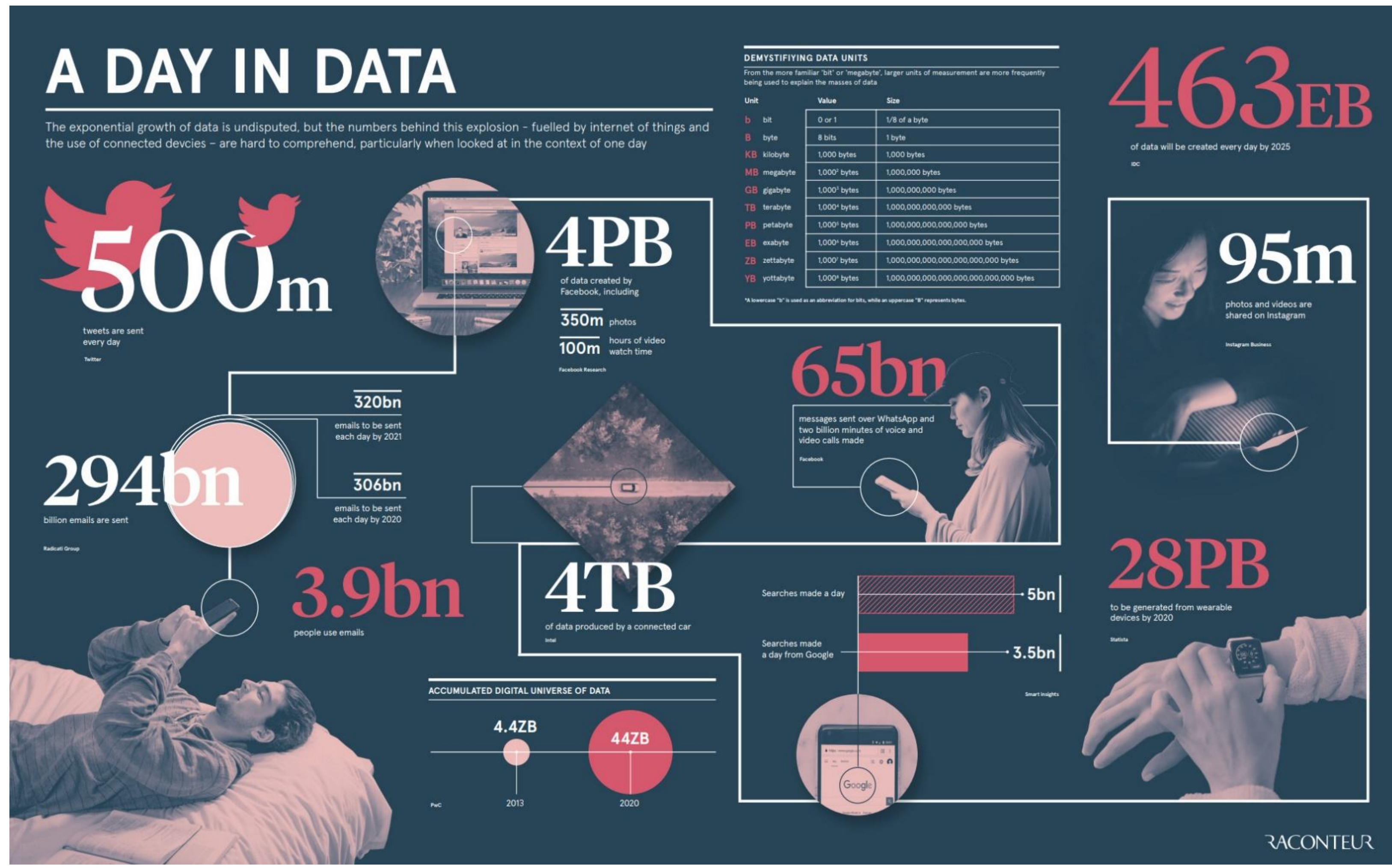
"***We are drowning in information but starved for knowledge.***" — John Naisbitt



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



Infographic by Raconteur. Click on the figure to access full size pdf

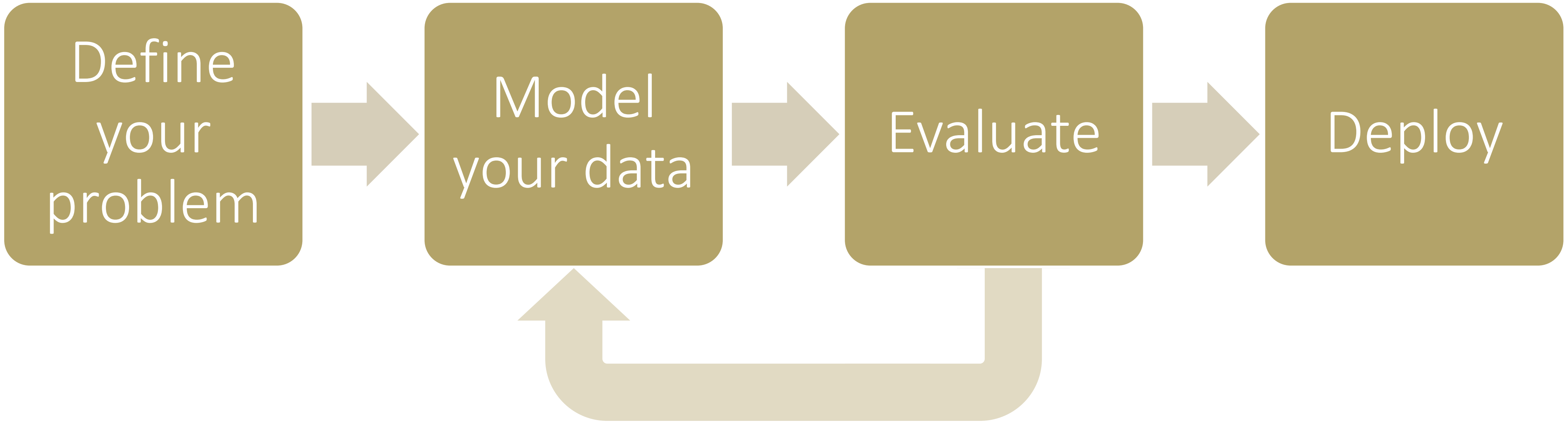
Levels of AGI

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

https://arxiv.org/abs/2311.02462

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

$$
\begin{array}{r|cccc|c}
 & \text{Weight(lb)} & \text{Height(cm)} & \text{Fur color} & \text{Eye color} & \text{Label} \\
\hline
\end{array}
$$

$$
\begin{array}{r}
Point\ 1 \\ Point\ 2 \\ Point\ 3 \\ Point\ 4 \\ Point\ 5
\end{array}
\begin{bmatrix}
10 & 20 & w & g \\
50 & 100 & br & bl \\
8 & 15 & bl & bl \\
12 & 25 & w & bl \\
14 & 10 & bl & g
\end{bmatrix}_{X_{n\times d}}
=
\begin{bmatrix}
cat \\ dog \\ dog \\ cat \\ dog
\end{bmatrix}_{Y_{n\times 1}}
$$

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

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

We can do better than Cat and Dog

$$
\begin{array}{r|ccccc|}
 & \text{Weight(lb)} & \text{Height(cm)} & \text{Fur color} & \text{Eye color} \\
\end{array}
\qquad \text{Label}
$$

$$
\begin{array}{r}
\textit{Point 1} \\ \textit{Point 2} \\ \textit{Point 3} \\ \textit{Point 4} \\ \textit{Point 5}
\end{array}
\begin{bmatrix}
10 & 20 & w & g \\
50 & 100 & br & bl \\
8 & 15 & bl & bl \\
12 & 25 & w & bl \\
14 & 10 & bl & g
\end{bmatrix}
=
\begin{bmatrix}
\text{Blob Fish} \\
\text{opossum} \\
\text{opossum} \\
\text{Blob Fish} \\
\text{opossum}
\end{bmatrix}
$$

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

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 (Clustering)

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

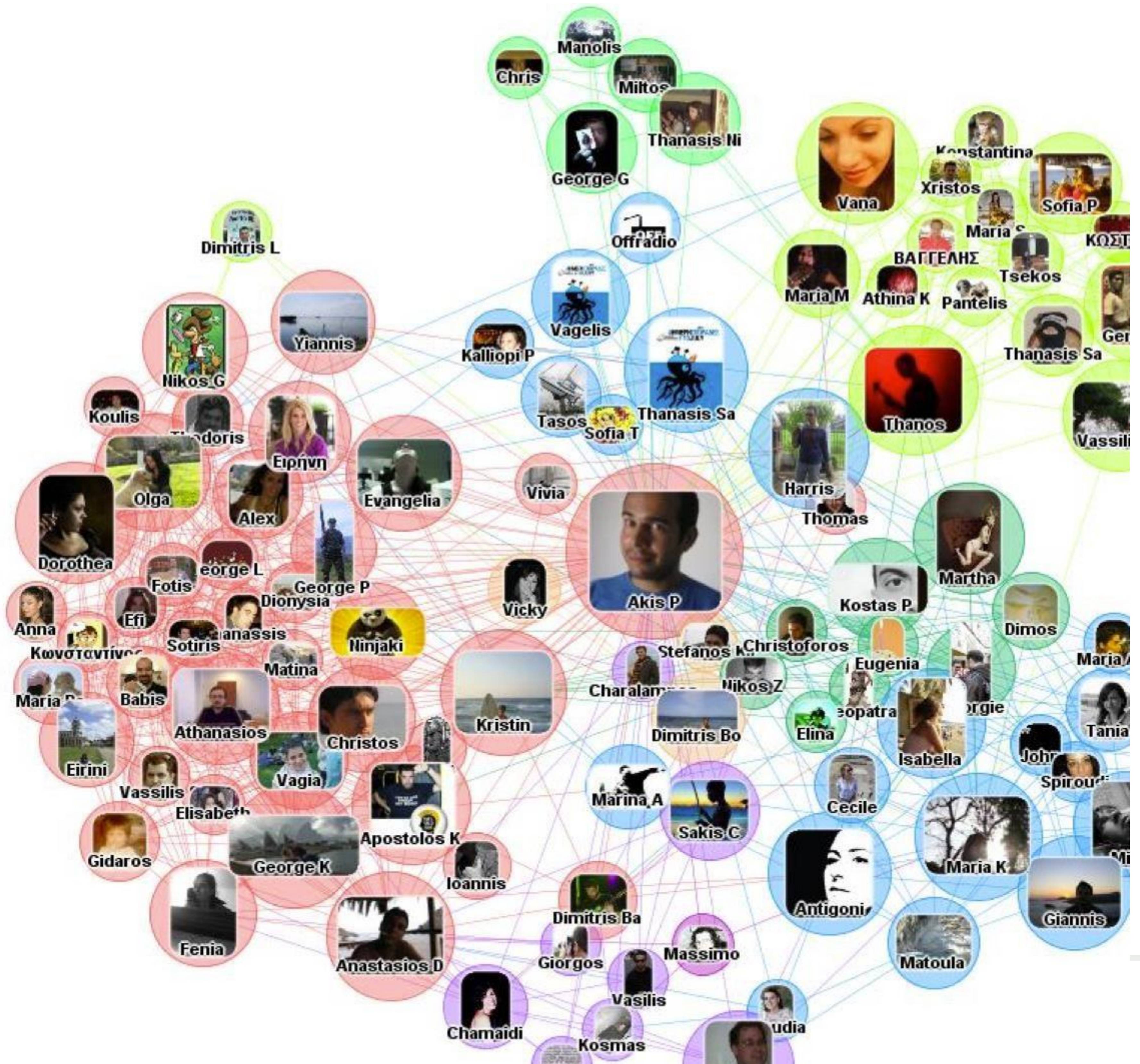
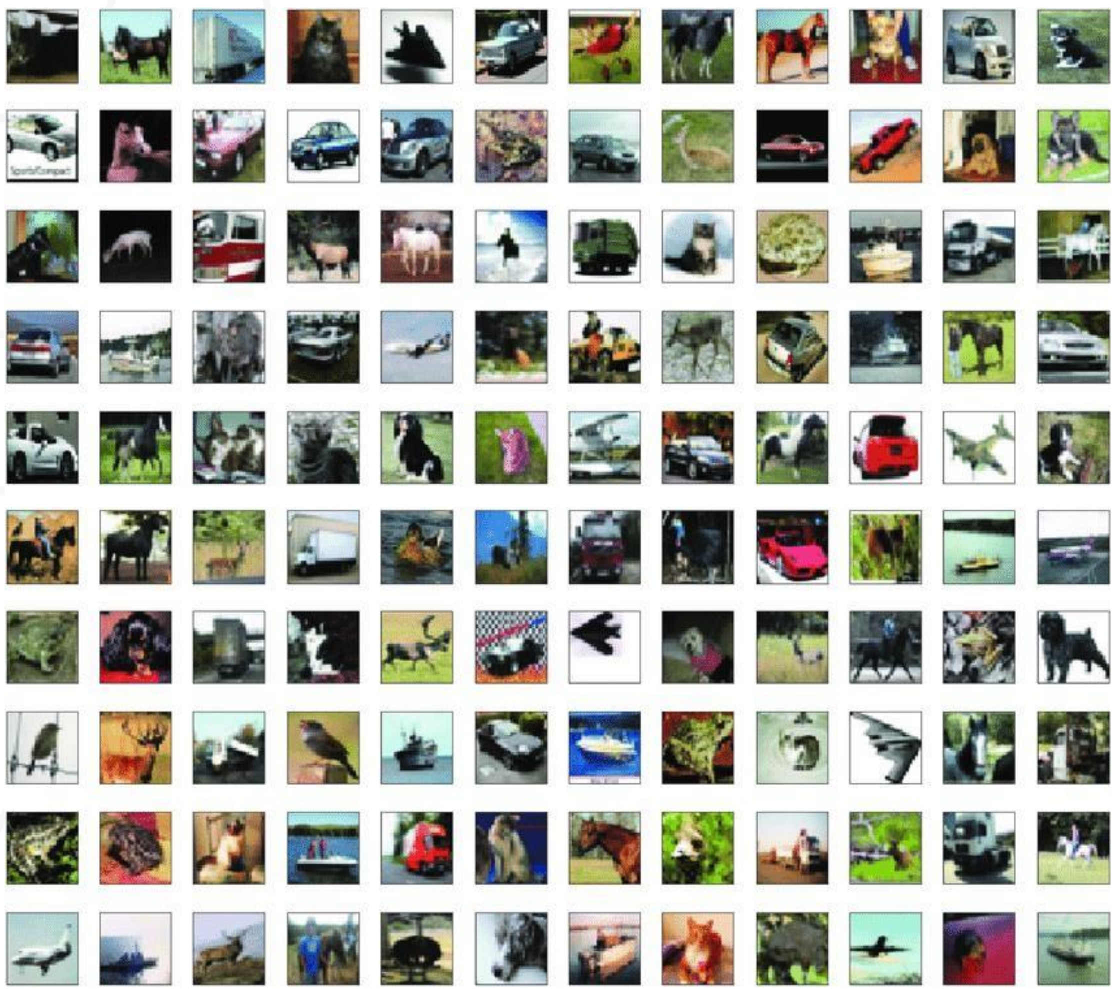
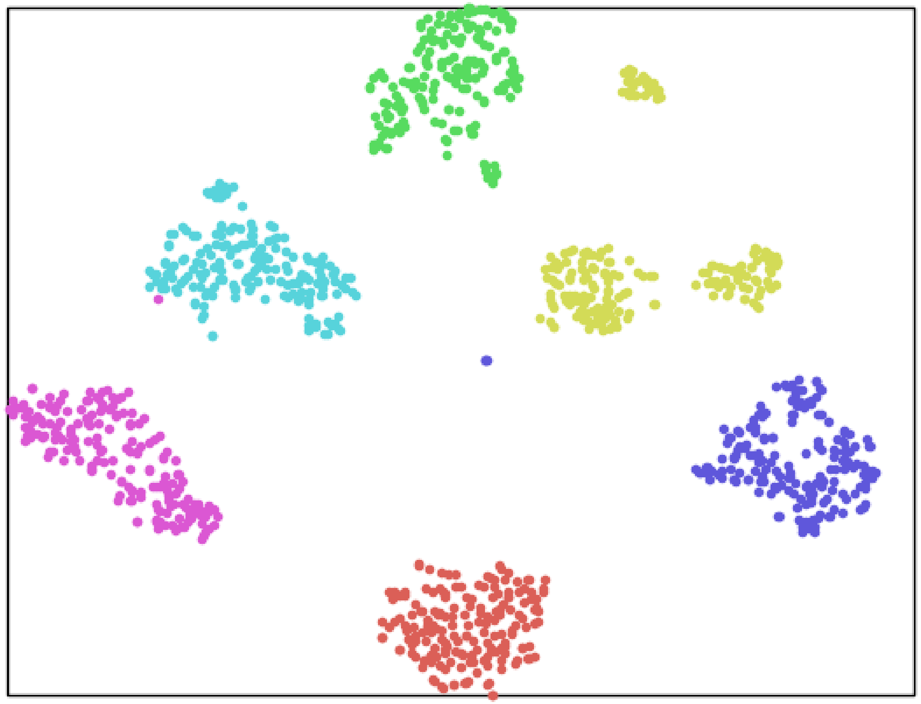
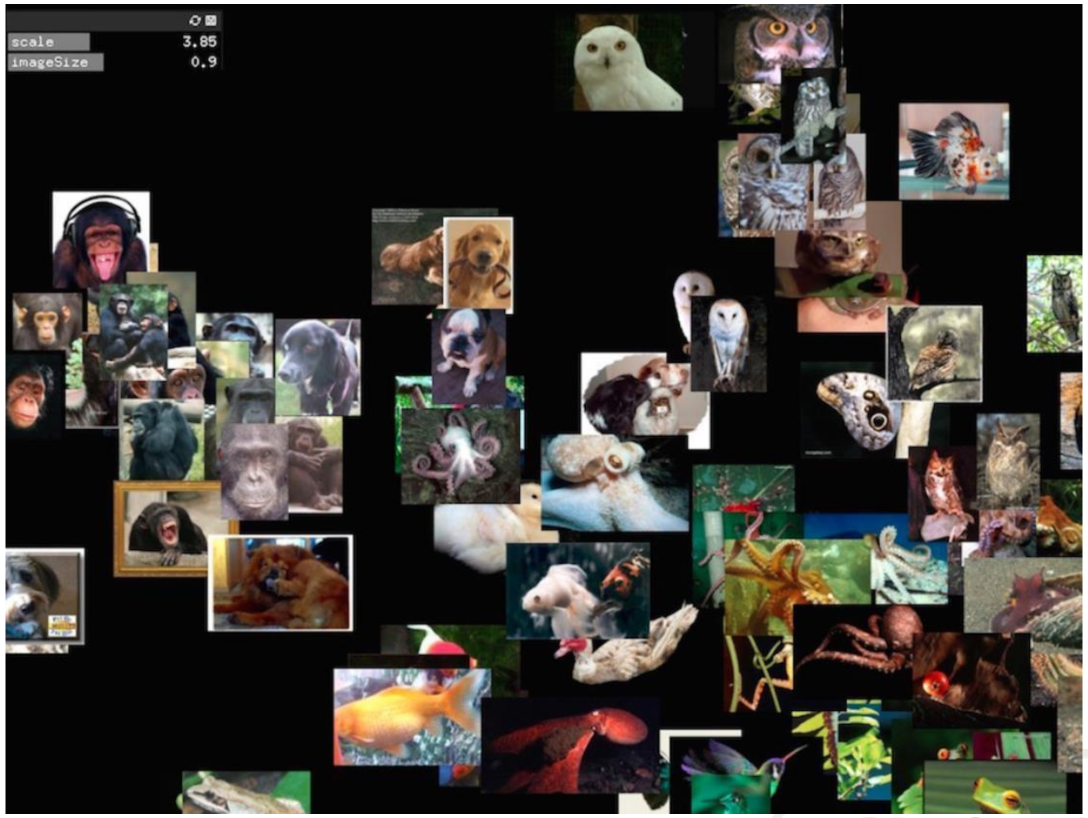


Image credit: Symeon Papadopoulos

Dimensionality Reduction







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

Image credit: Gene Kogan

Syllabus: Supervised Learning

Tree-based models

- Decision tree
- Ensemble learning/Random forest

Linear classification/regression models

- Linear regression
- Naive Bayes
- Logistic regression

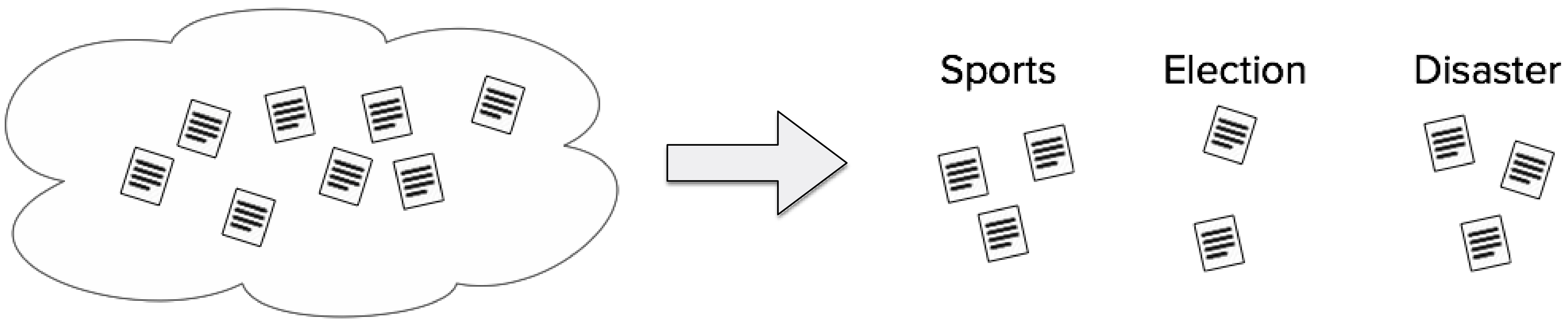
Neural networks

- Feedforward neural networks and backpropagation analysis
- CNN

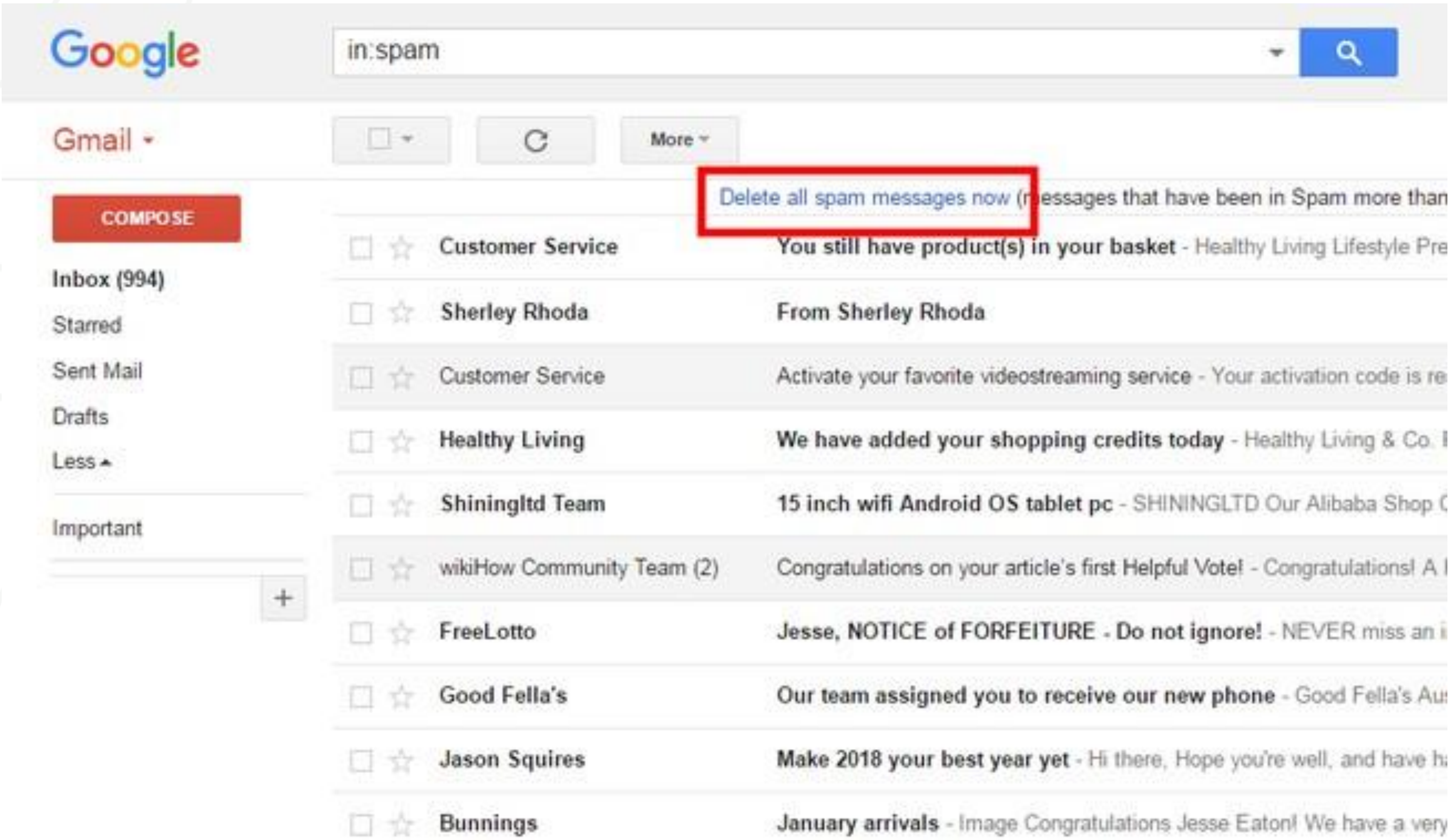
Support vector machine

News Classification

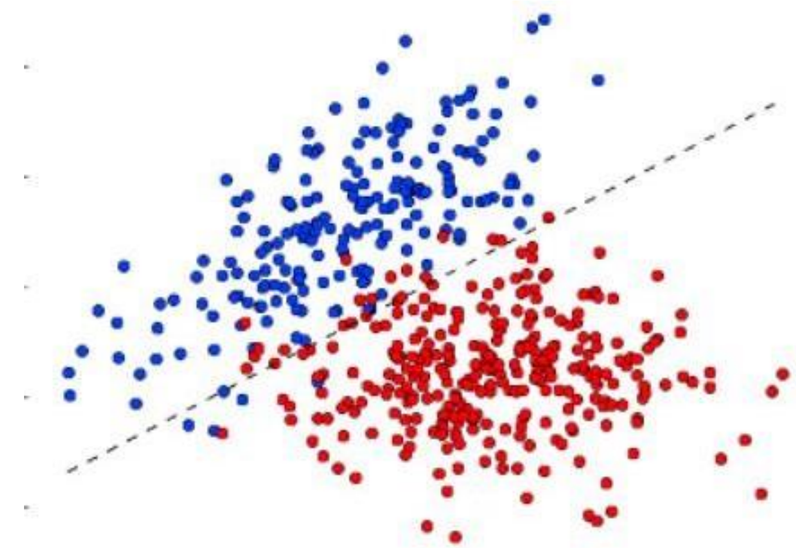
- What are the inputs and how to represent them?
- What are the desired outputs?
- What learning algorithms to choose?



Spam Detection



NOT SPAM



SPAM

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

Image credit: WikiHow

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?**

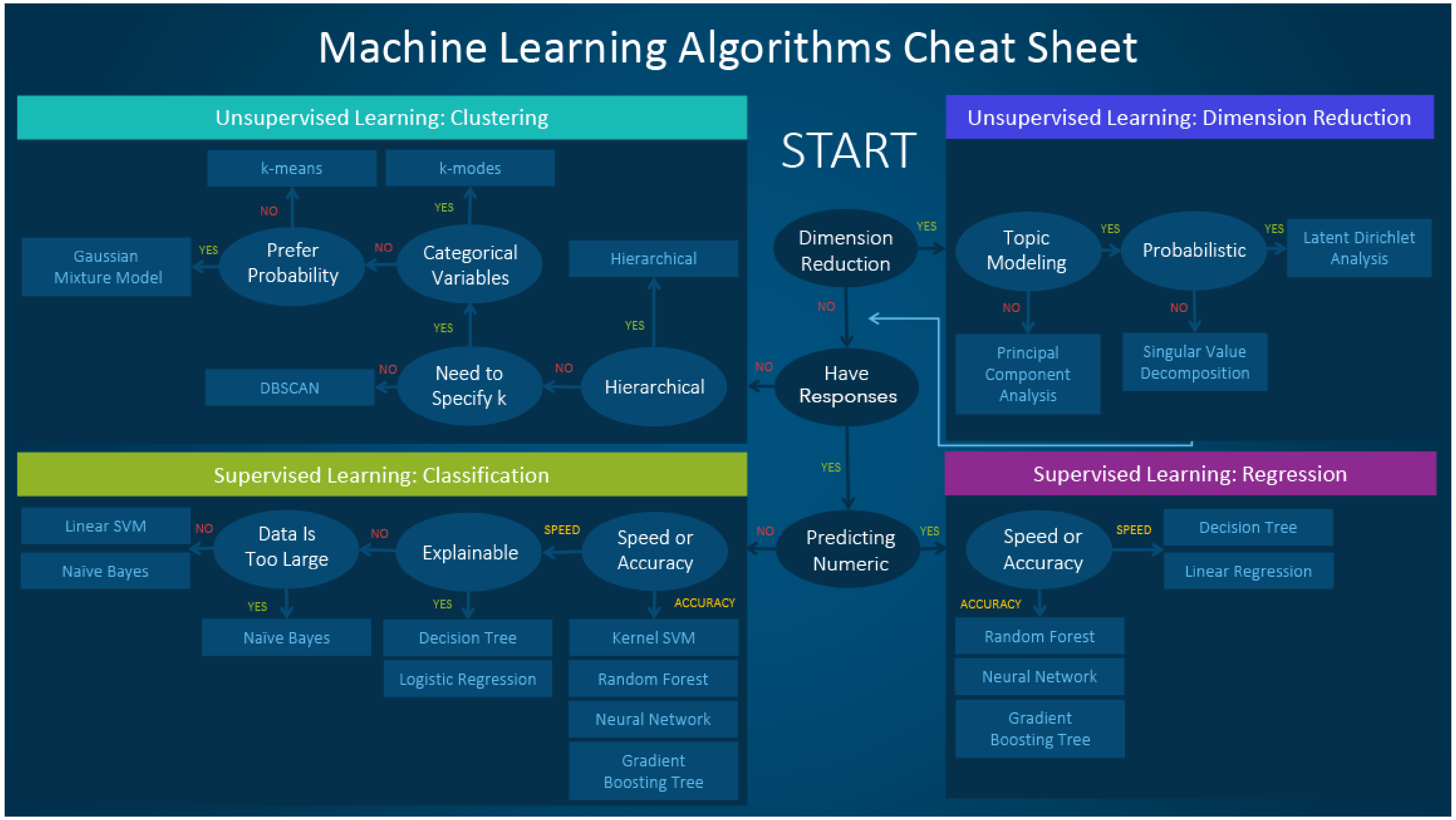
Machine learning workflow process

- **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.

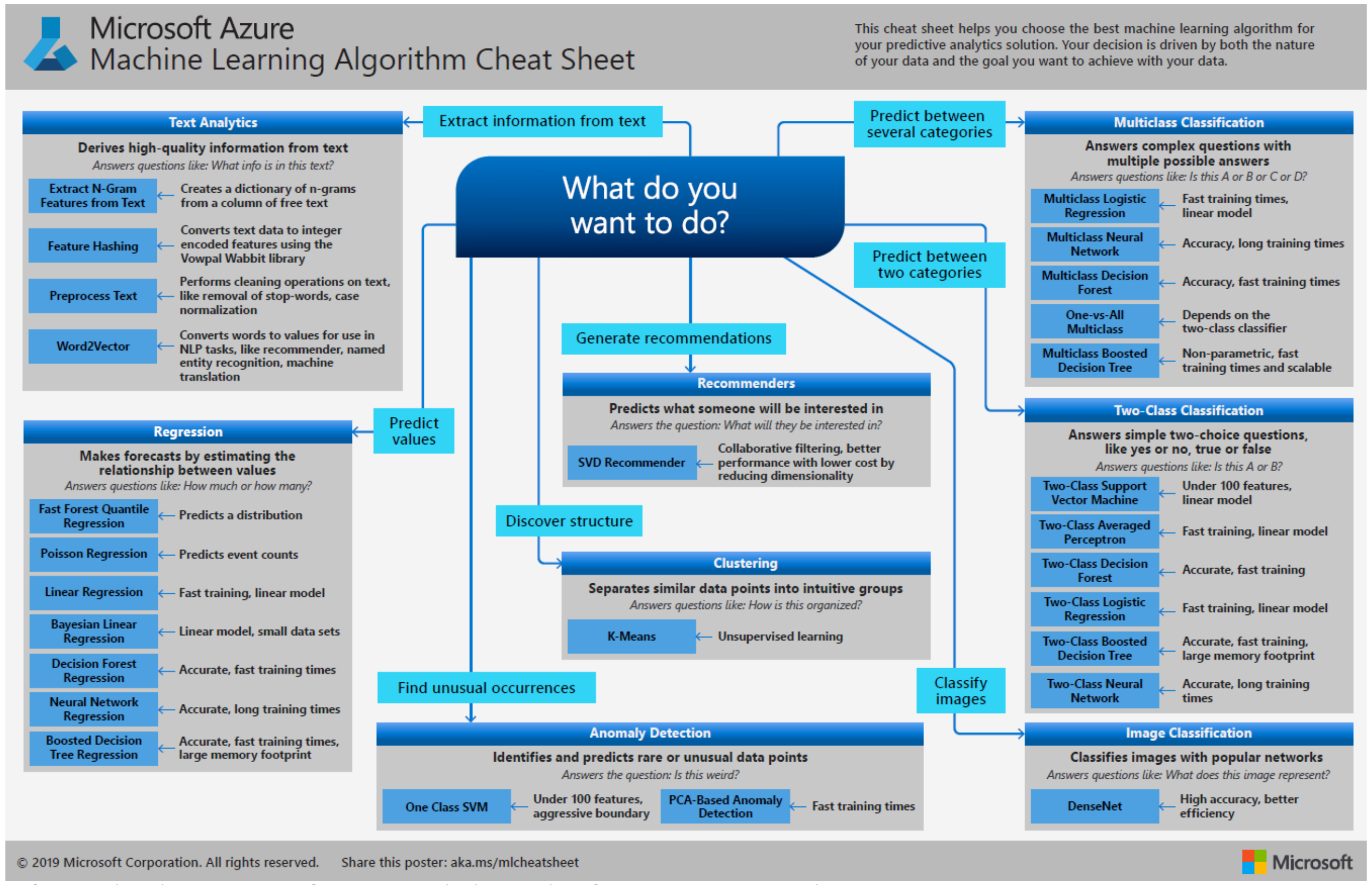
- **Data exploration:** Explore the data to gain insights and detect patterns, outliers, and anomalies.
- **Feature engineering:** Create features that capture relevant information from the data.

- **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.



Infographic by Microsoft Azure. Click on the figure to access the source.

[General Project Guidance on Class Website]

Georgia Tech

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

Georgia Tech

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?