01.01.2019

Statistical Methods in AI (CSE/ECE 471)

Lecture-1: Intro and Administrivia

Ravi Kiran (ravi.kiran@iiit.ac.in)

Center for Visual Information Technology (CVIT), IIIT Hyderabad

SMAI (Statistical Methods in AI)

- SMAI ~ Introduction to Machine Learning

- Good news: One half is already familiar to you [Machine !]
- Other half = What this course is about !

Machine Learning

Study of Algorithmic methods that use data to improve their knowledge of a task

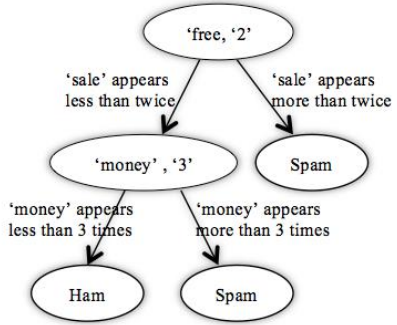
Machine Learning: Examples

Algorithmic methods that use data to improve their knowledge of a task

Task: Detect spam email

Data: Labelled emails (in inboxes of other users as well !)

Knowledge:



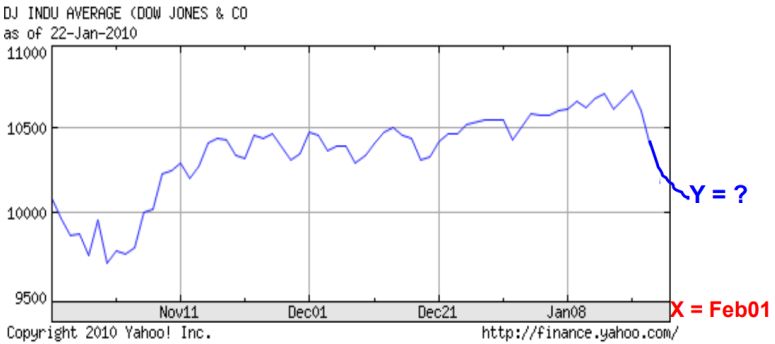
Improve → 85% reduction of spam emails in Inbox over 3 months

Algorithmic method: Decision Tree

Machine Learning: Examples

Algorithmic methods that use data to improve their knowledge of a task

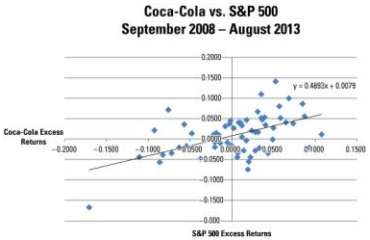
Task: Predict value of a stock (GOOG)



Data: Historical stock value (time, price/share)

Knowledge: Model coefficients

Improve → Predict stock to 95% of its value



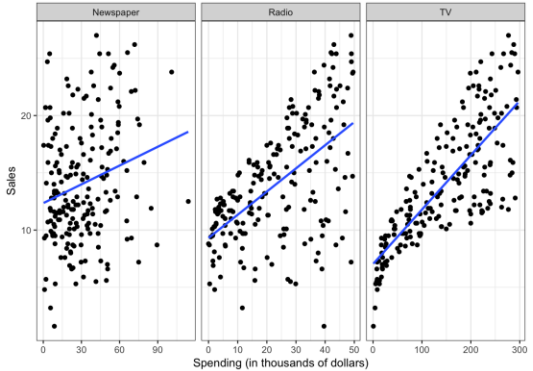
Algorithmic method: Linear Regression

Machine Learning: Examples

Algorithmic methods that use data to improve their knowledge of a task

Task: Predict effect of advertising on 'furniture' sales

Algorithmic method: Linear Regression



Data: Amount spent on ad spots in TV, radio, newspaper

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Knowledge: For a given amount of TV, newspaper advertising spending additional 10,000 rupees on FM radio leads to an additional sale of 150 units

Machine Learning: Examples

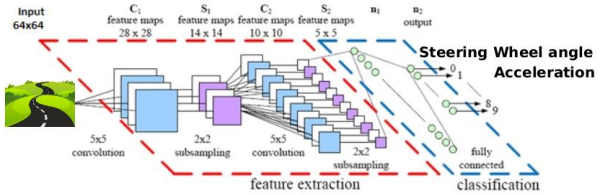
Algorithmic methods that use data to improve their knowledge of a task

Task: Drive car 'safely' without human intervention

Data: Camera, Laser , GPS data ; Synthetic data

Knowledge: Model coefficients

Improve → Drive 160,000 miles without accident/human intervention

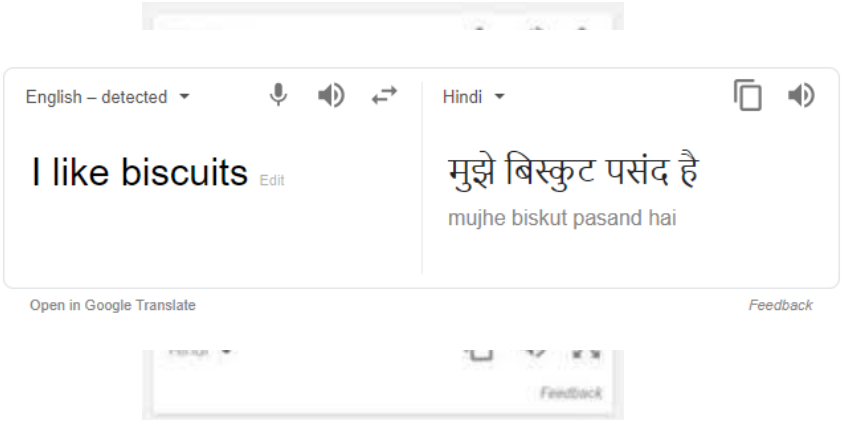


Algorithmic method: Deep Reinforcement Learning

Machine Learning: Examples

Algorithmic methods that use data to improve their knowledge of a task

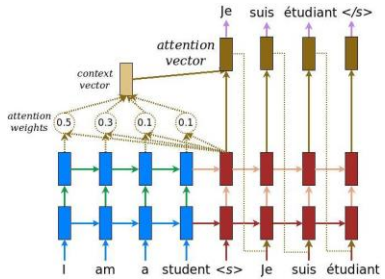
Task: Translate text from one language to another



Data: Paired sentences from source and target languages

Knowledge: Model coefficients

Improve → Reduce number of mistakes by 78%



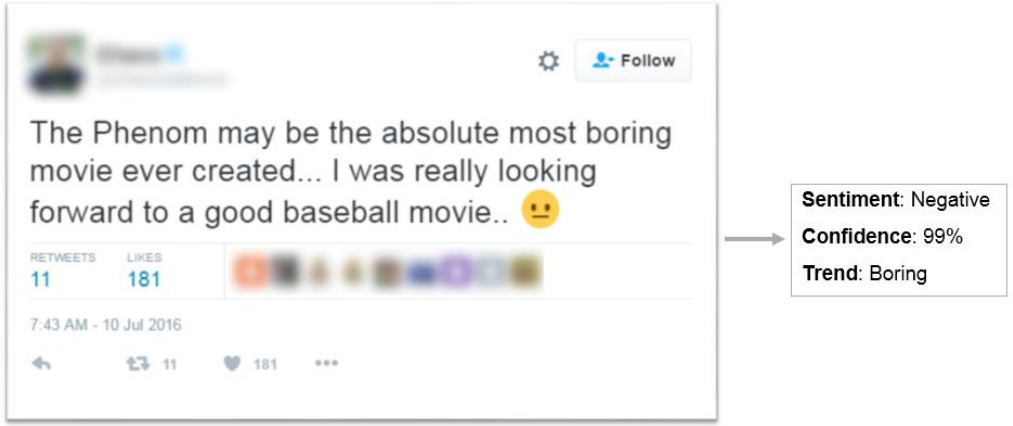
Algorithmic method: Deep Recurrent Neural Networks

Machine Learning: Examples

Algorithmic methods that use data to improve their knowledge of a task

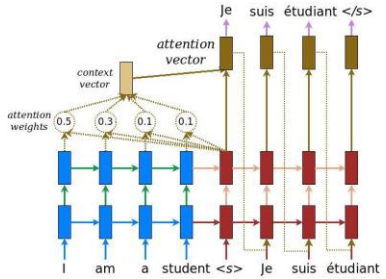
Task: Sentiment Analysis

Data: Text and 'Sentiment' label



Knowledge: Model coefficients

Improve → Reduce number of sentiment mislabelings by 80%



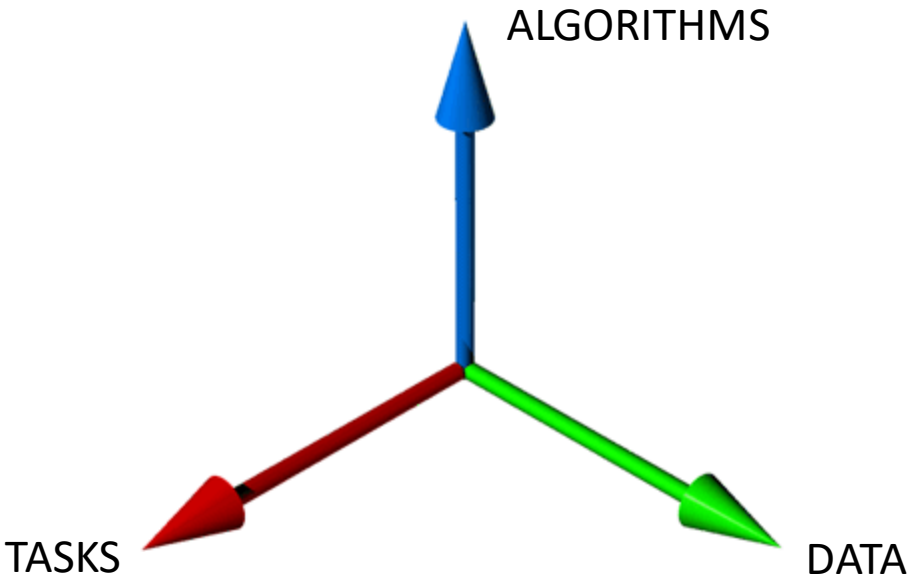
Algorithmic method: Deep Recurrent Neural Networks

What is ML ?

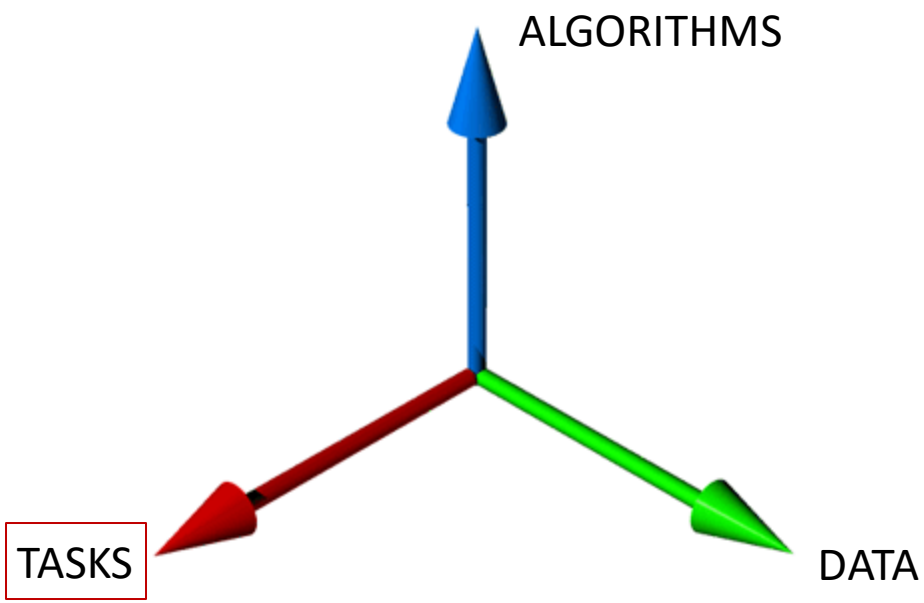
- Computer program whose behavior evolve based on empirical data (Wikipedia)
- Computer program that learns from experience E in order to improve its performance P on a task T (Tom Mitchell)

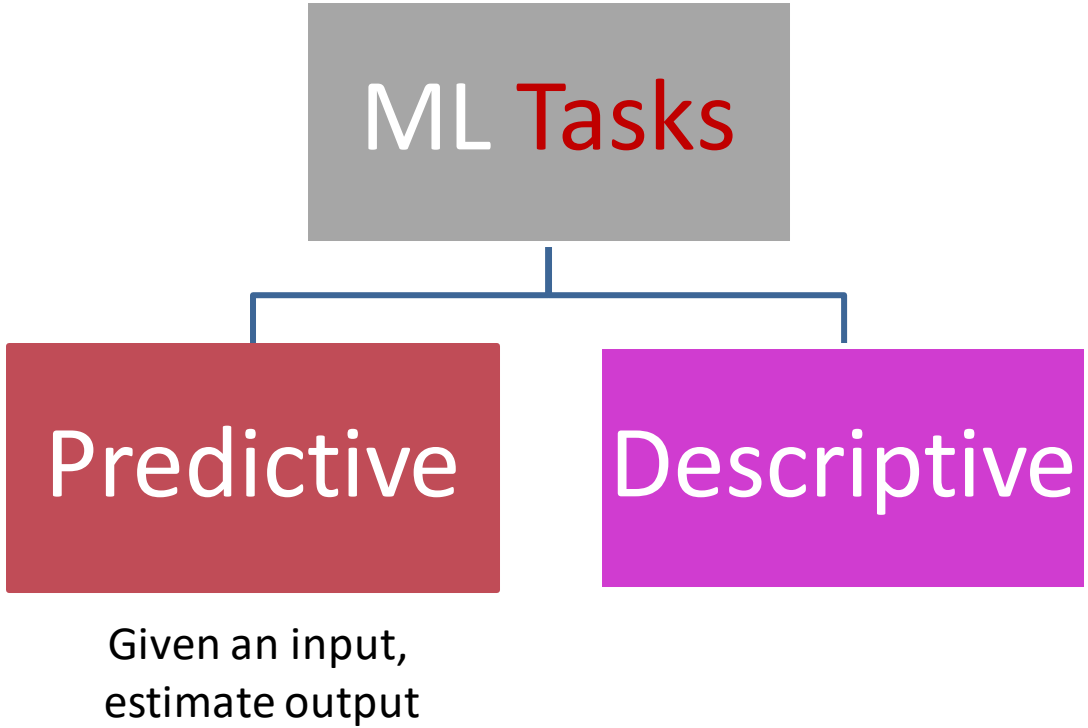
experience E : images, text, sensor measurements, biological data
task T : estimating probabilities, predicting object label, dimensionality reduction, clustering
performance P : probability of success, money/time saved,

3 axes of ML



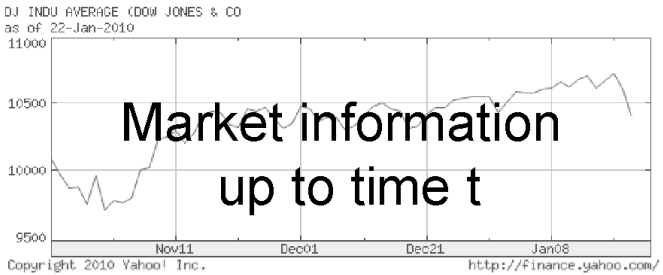
3 axes of ML




ML Tasks
Predictive
Descriptive
Given an input,
estimate output

ML::Tasks → Predictive

Feature Space $\mathcal{X}$



Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
…

Share Price
"$ 24.50"

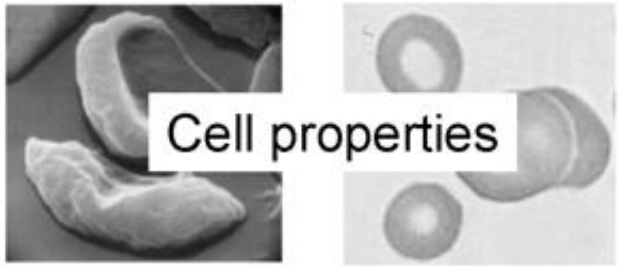
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

ML::Tasks → Predictive → Classification

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
...



"Anemic cell"
"Healthy cell"

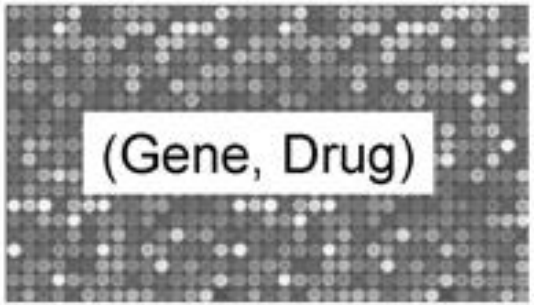
Discrete Labels

ML::Tasks → Predictive → Regression

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

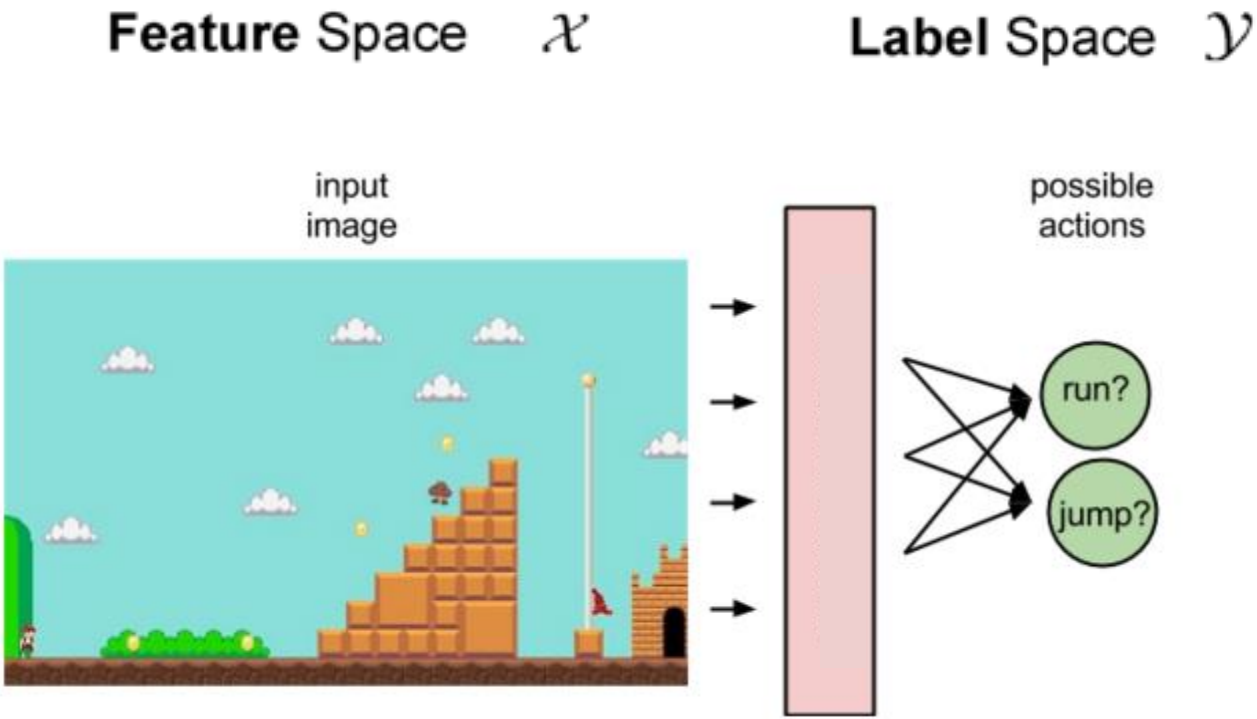
→ Share Price
"$ 24.577"



→ Expression level
"6.88"

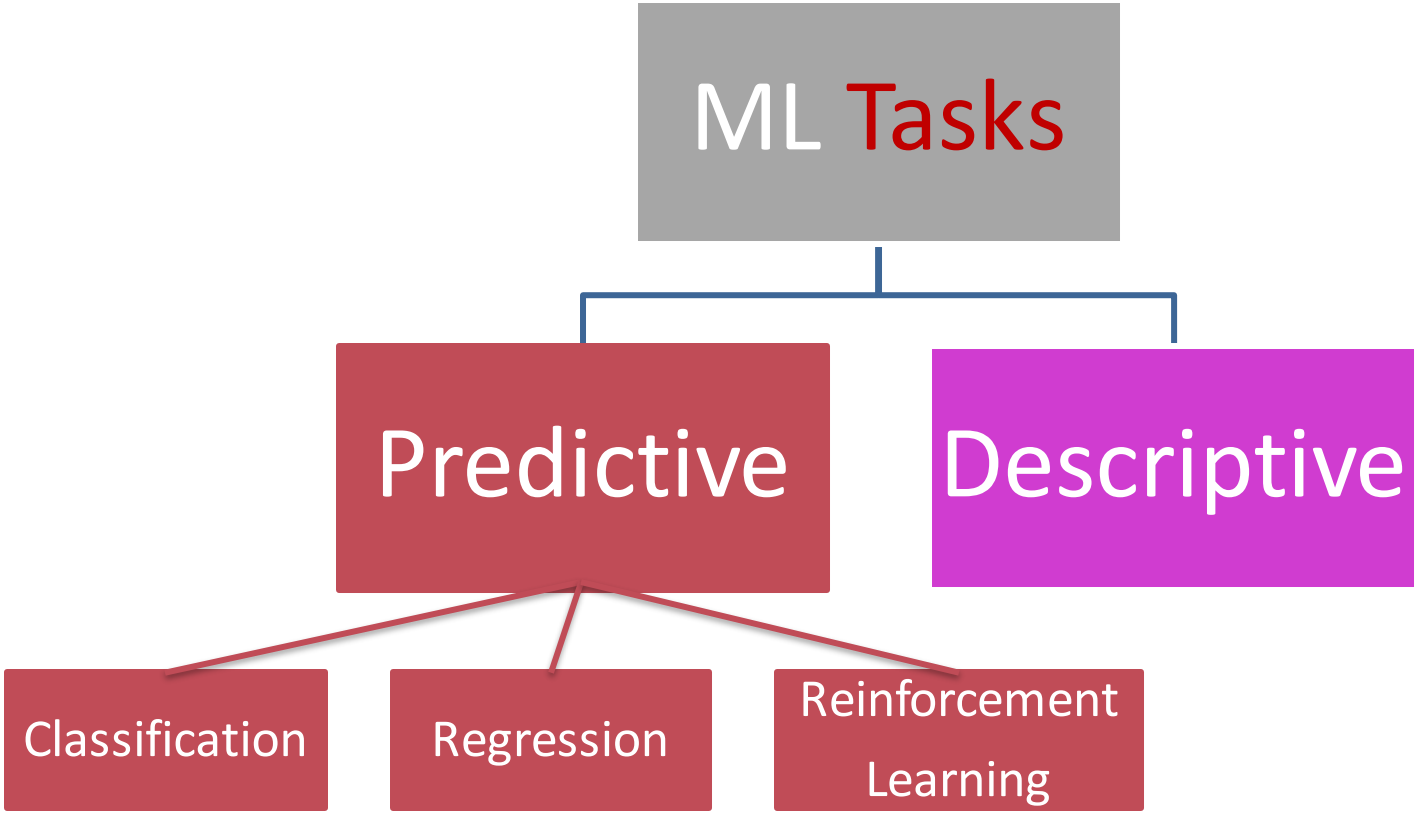
Continuous Labels

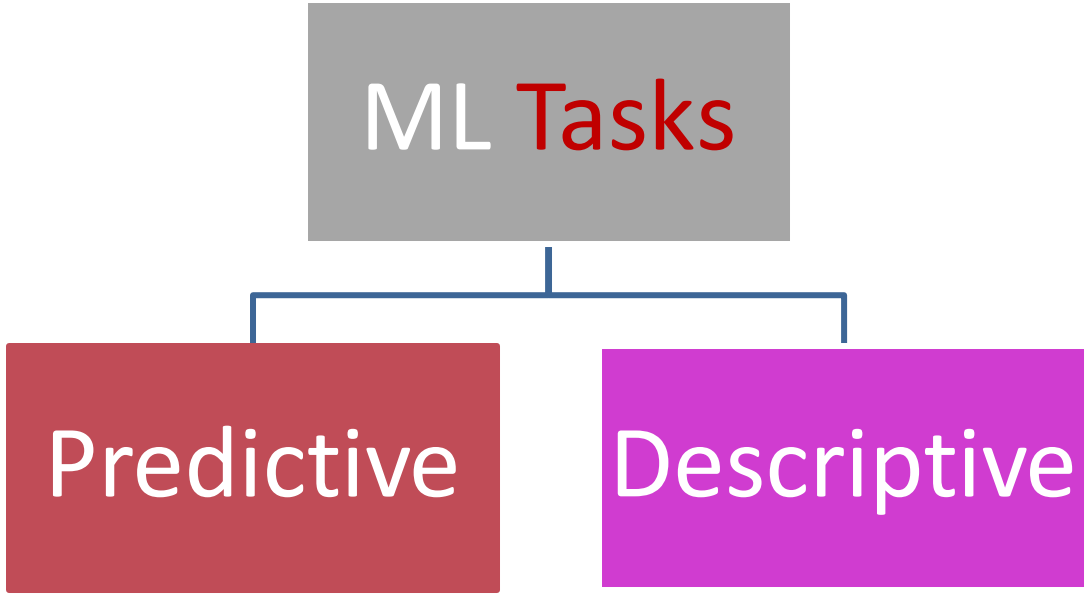
ML::Tasks → Predictive → Reinforcement Learning



ML::Tasks → Predictive → Reinforcement Learning

Feature Space $\mathcal{X}$ **Label** Space $\mathcal{Y}$


ML Tasks
Predictive
Descriptive
Classification
Regression
Reinforcement Learning


ML Tasks
Predictive
Descriptive

ML::Tasks → Descriptive

- Study/Exploit the ‘structure’ of data
  - Density Estimation
  - Clustering
  - Dimensionality Reduction
- Also studied as ‘Unsupervised Learning’
  - ‘Input’ data without paired ‘Output’

Unsupervised Learning → Density Estimation

Aka "learning without a teacher"

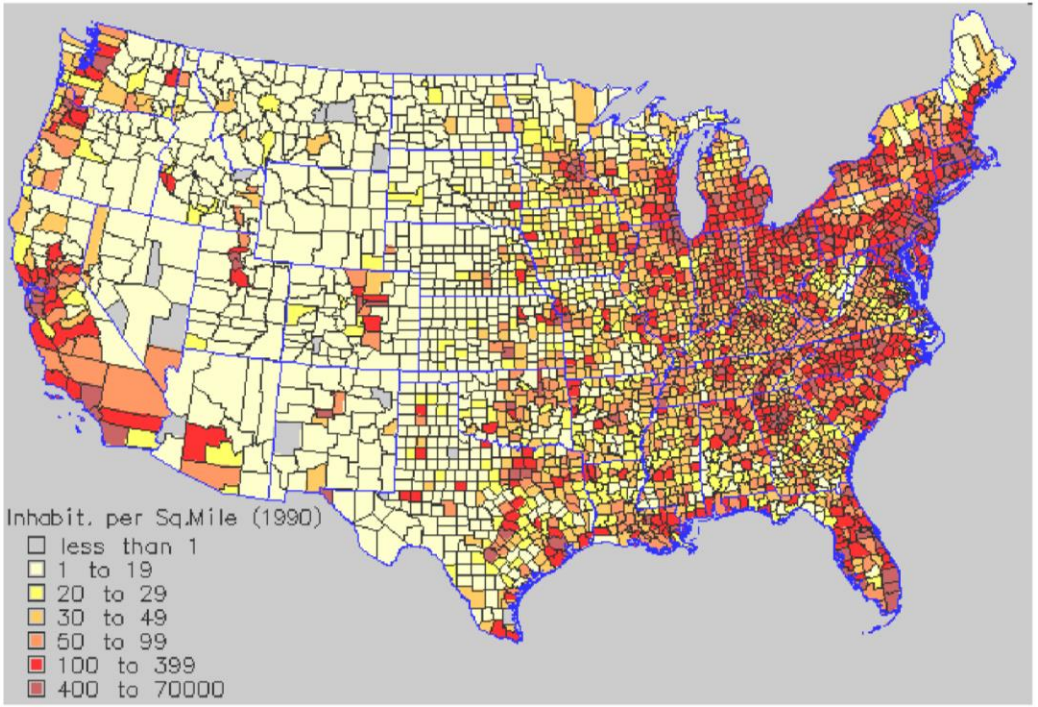
Feature Space $\mathcal{X}$

Word distribution
(Probability of a word)

Task: Given $X \in \mathcal{X}$, learn $f(X)$.

Unsupervised Learning → Density Estimation

Population density

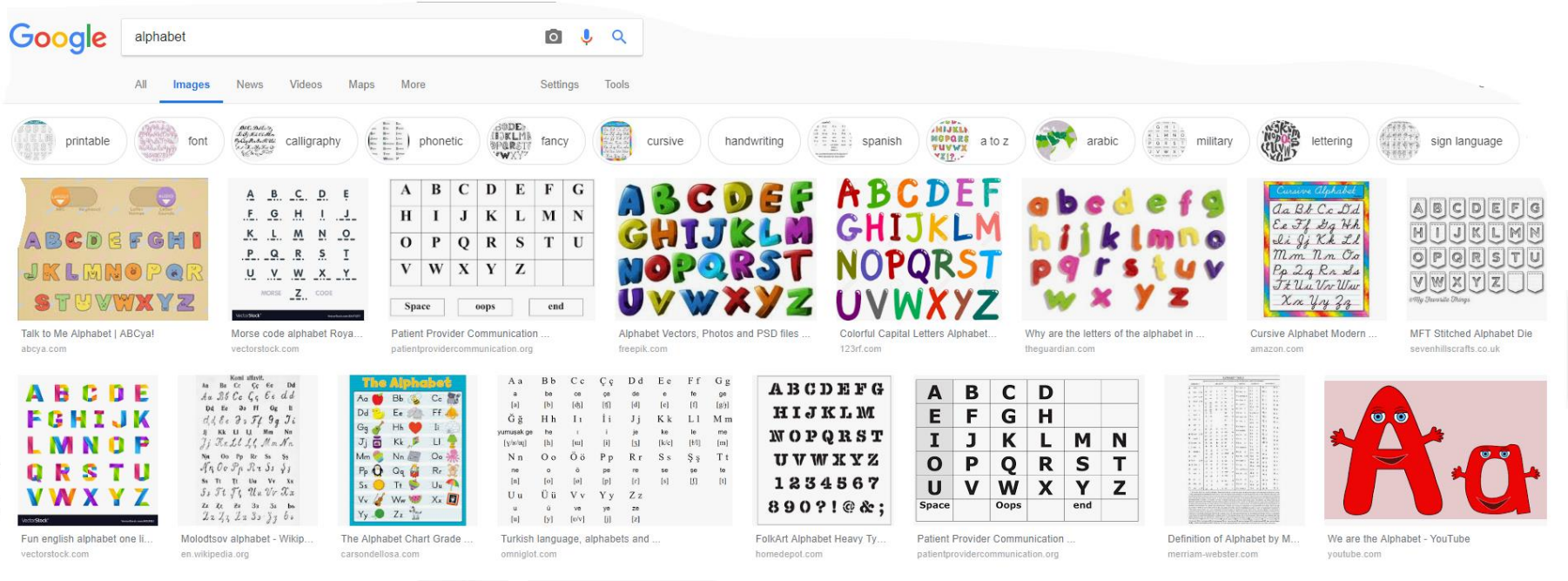


Unsupervised Learning → Clustering

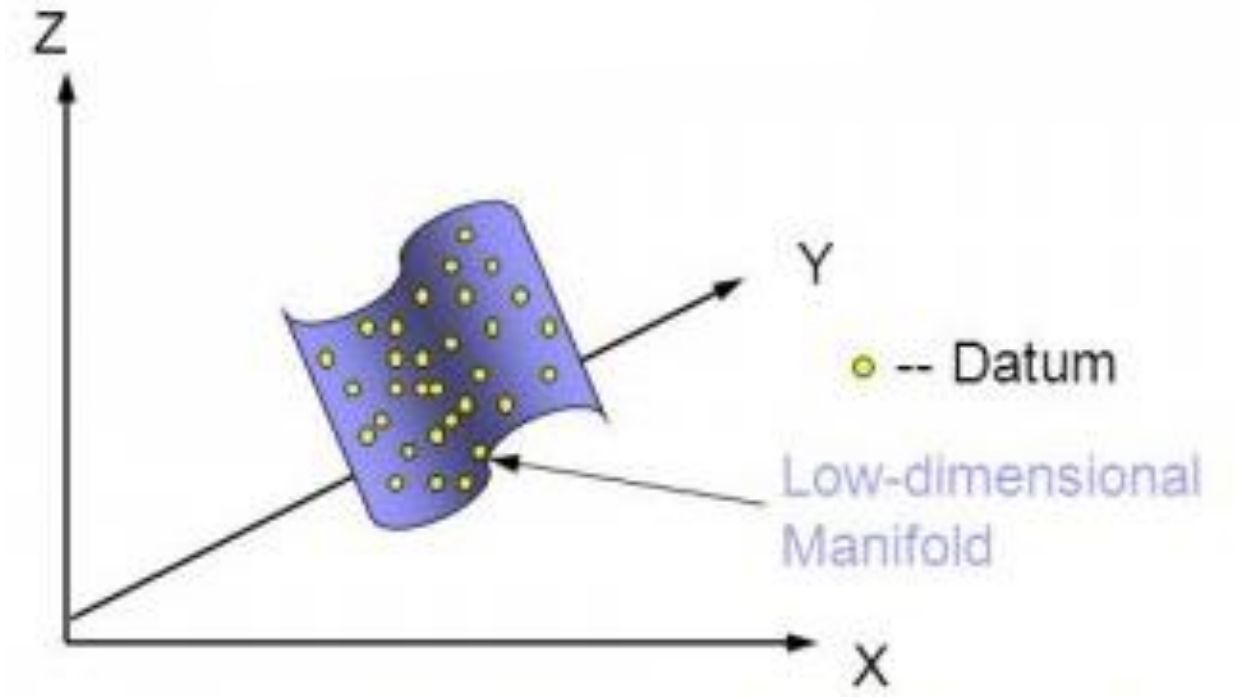
Group similar things e.g. images

[Goldberger et al.]

Unsupervised Learning → Web Search



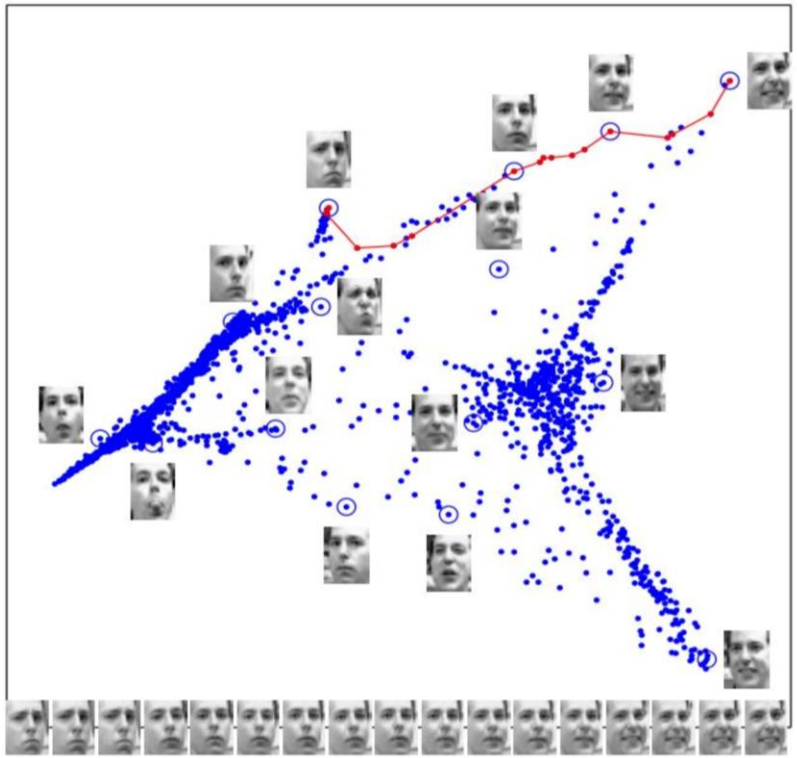
Unsupervised Learning → Dimensionality Reduction



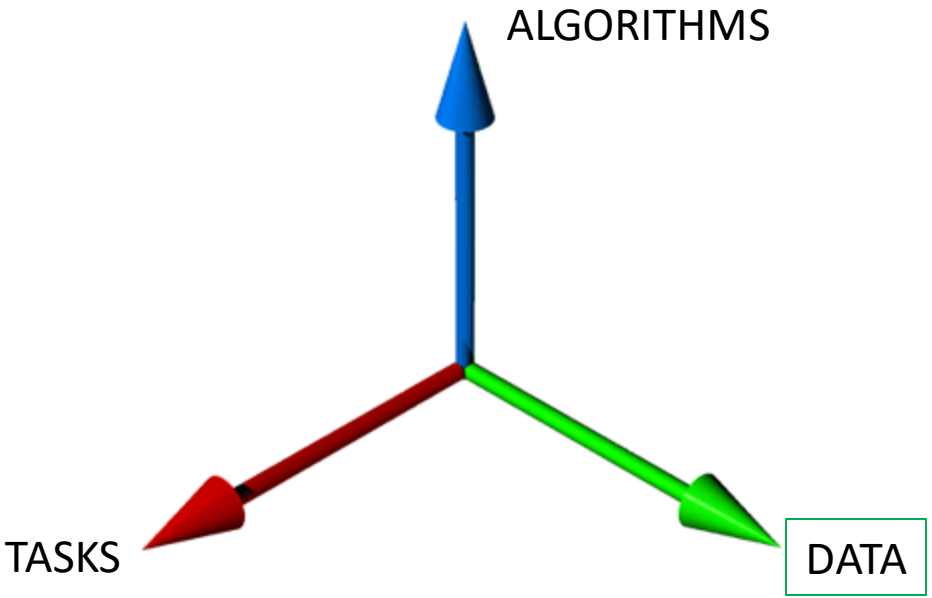
Unsupervised Learning → Dimensionality Reduction + Visualization

Images have thousands or millions of pixels.

Can we give each image a coordinate,
such that similar images are near each other?

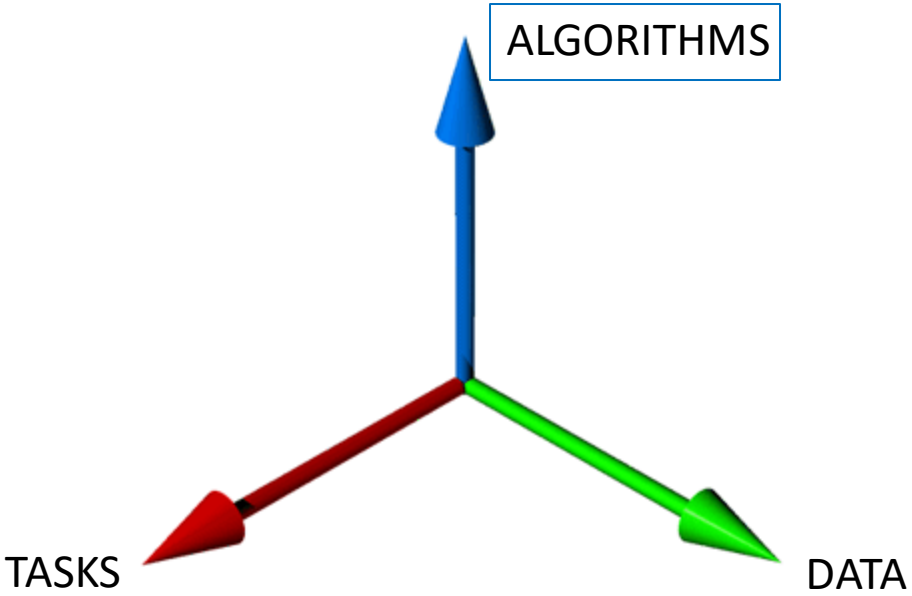


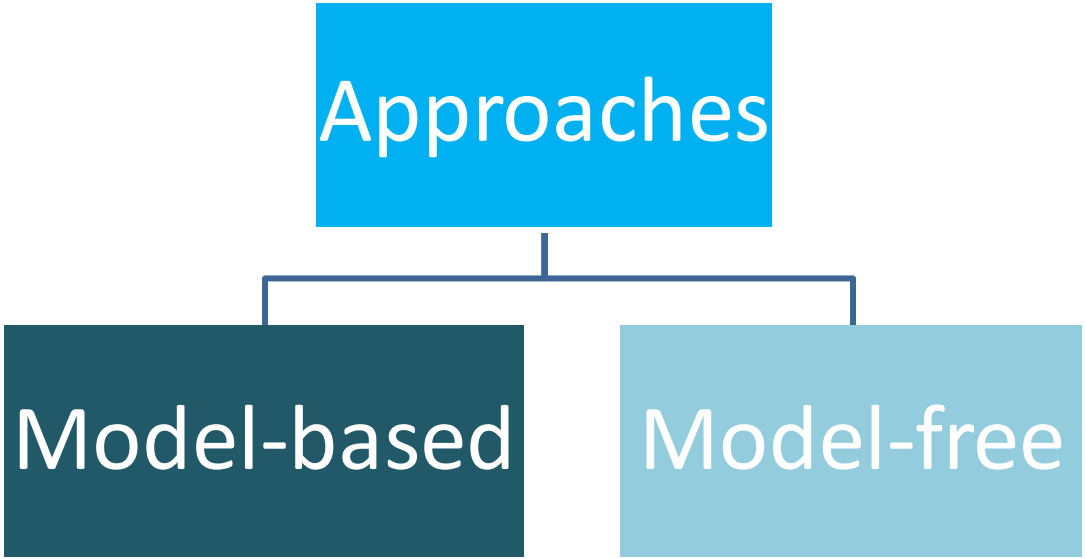
3 axes of ML



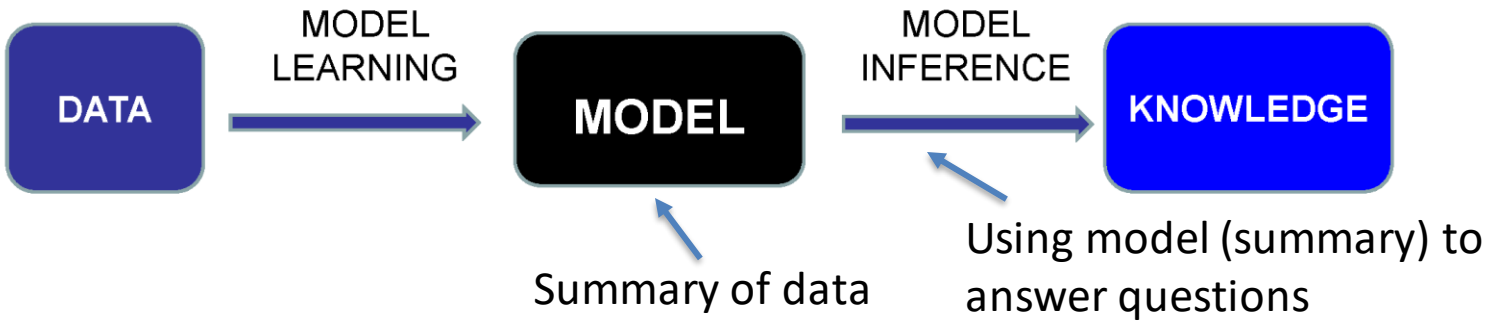
- Fully Observed
- Partially Observed
  - Some variables systematically not observed (e.g. ‘topic’ of a document)
  - Some variables missing some of the time (e.g. ‘faulty sensor’ readings)
- Actively collect / sense data (e.g. exploration robots)

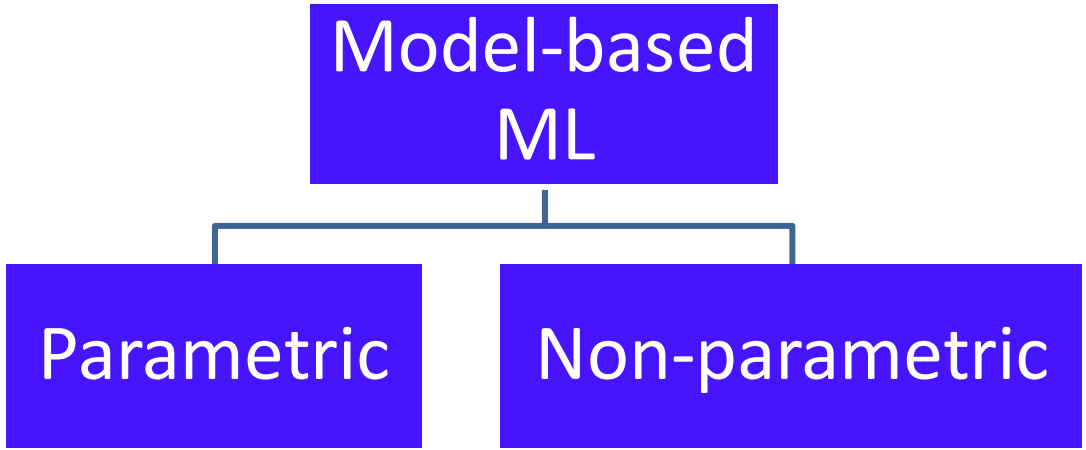
3 axes of ML




Approaches
Model-based
Model-free

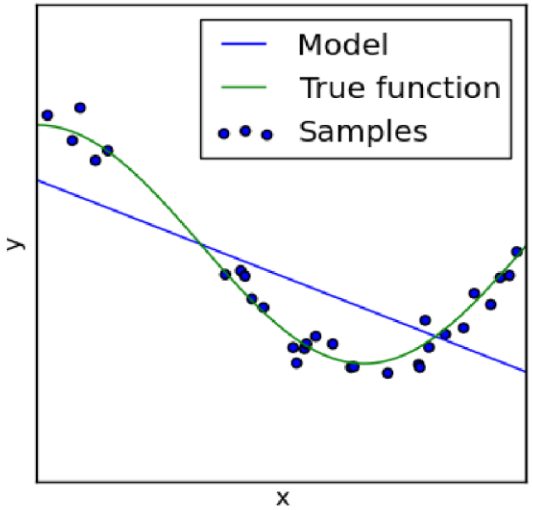
Model-based ML




Model-based ML
Parametric
Non-parametric

Parametric Models

- "Fixed-size" models that do not "grow" with the data
- More data just means you learn/fit the model better

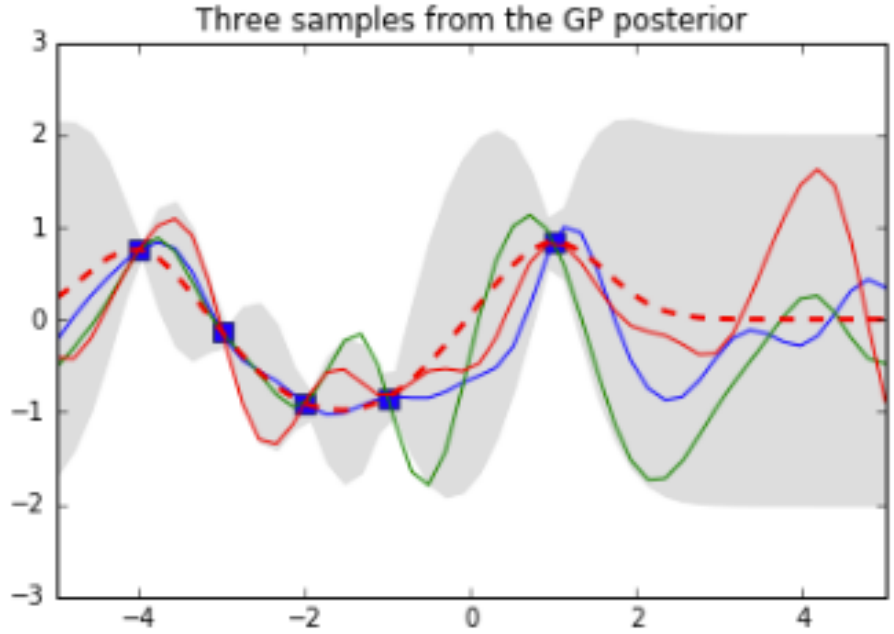


Fitting a simple line (2 params) to a bunch of one-dim. samples

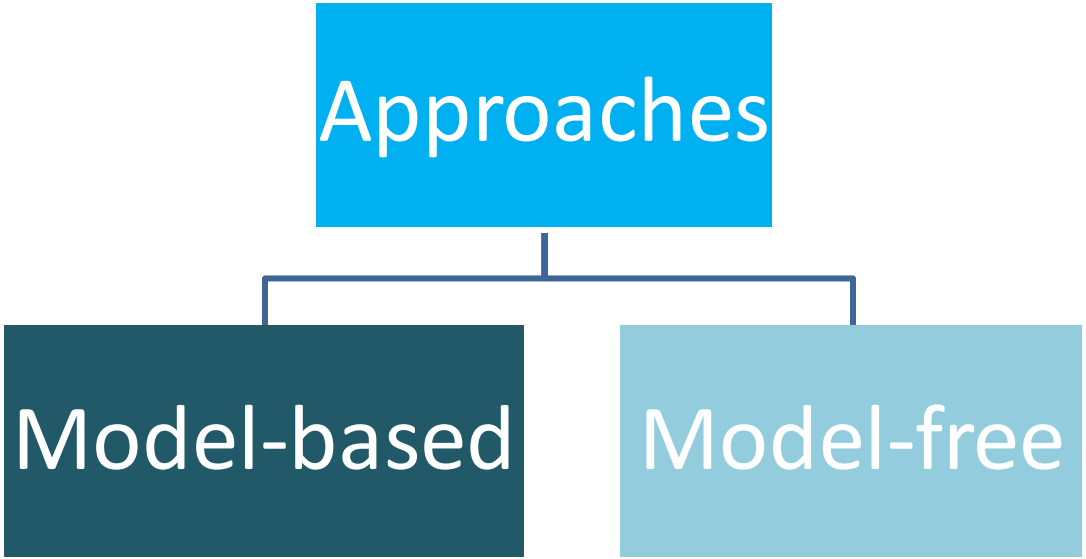
Model: data = point on line + noise

Nonparametric Models

- Models that grow with the data
- More data means a more complex model

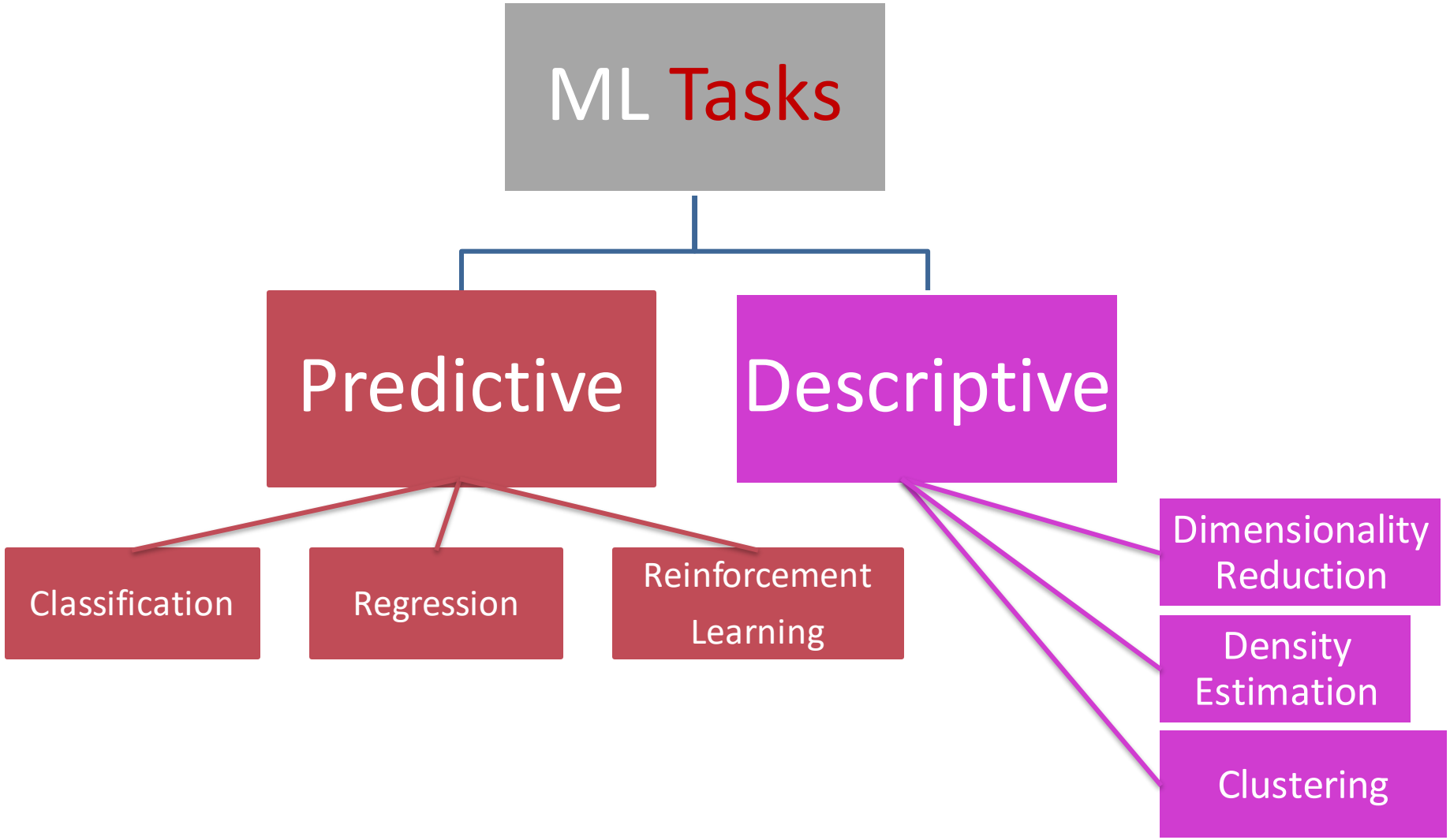


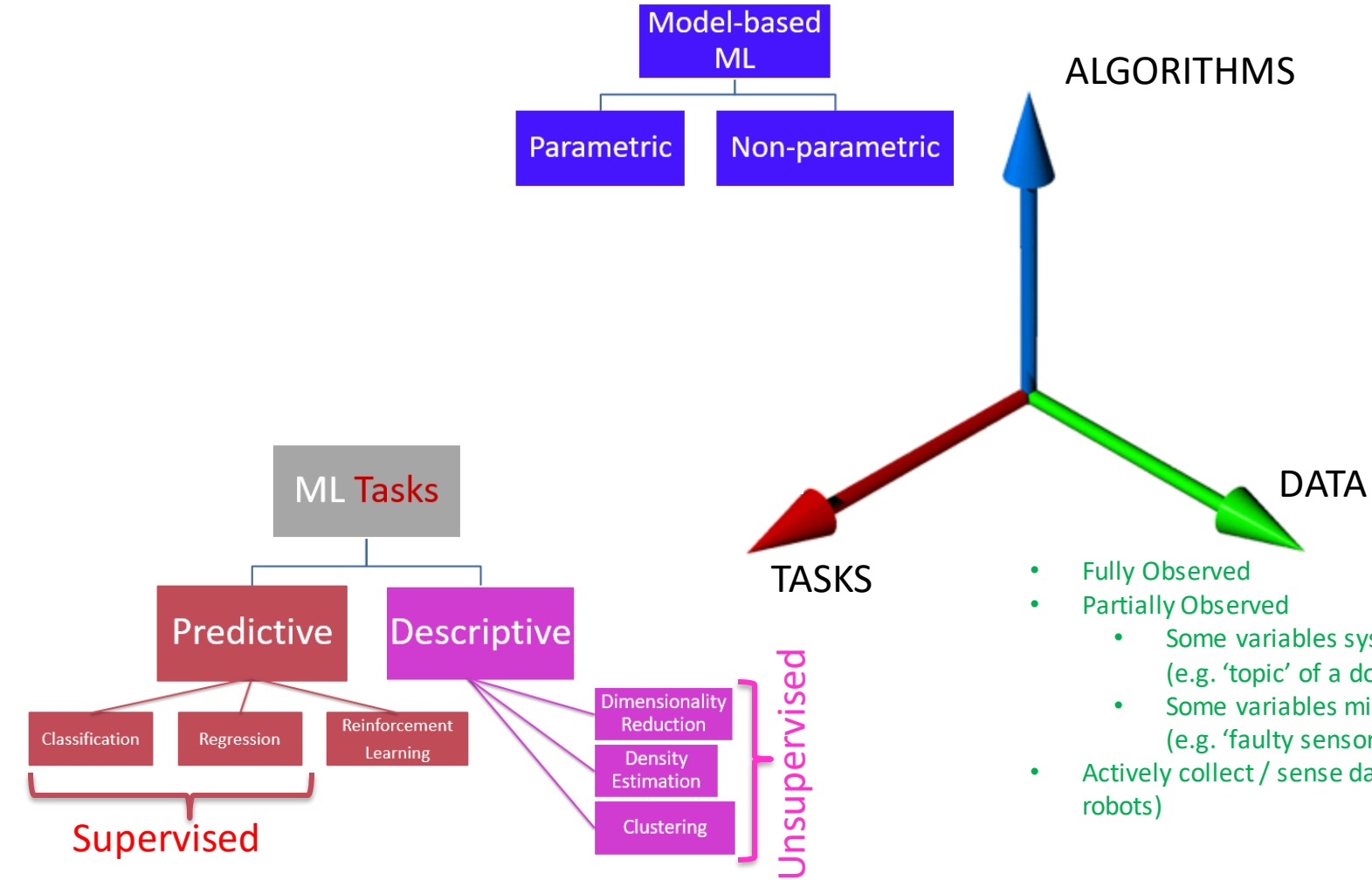
Gaussian Process


Approaches
Model-based
Model-free

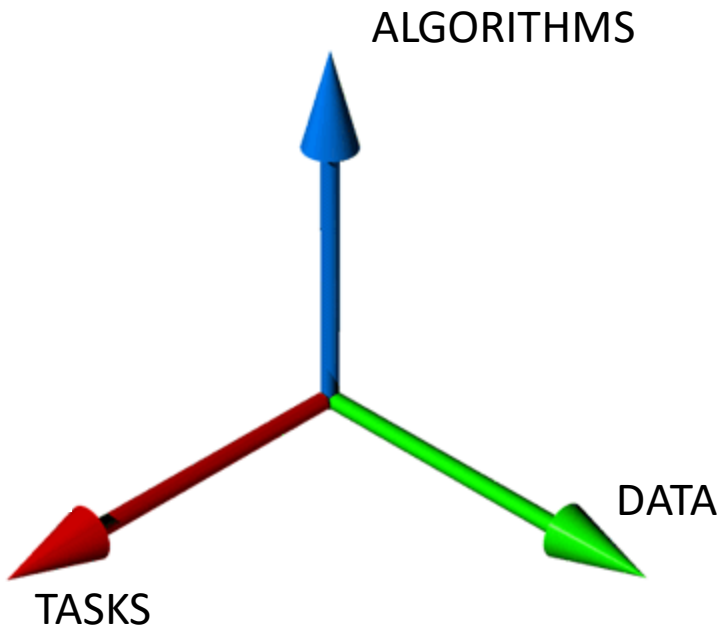
Model-Free

- Occurs in context of Reinforcement Learning
- …

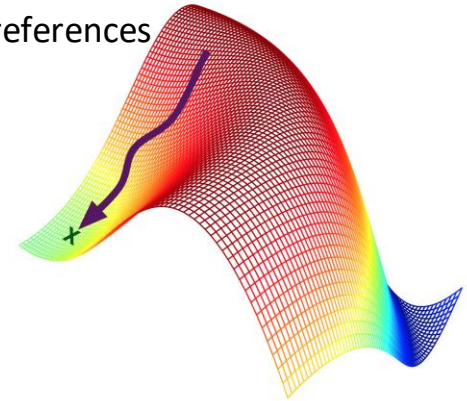

ML Tasks
Predictive
Descriptive
Classification
Regression
Reinforcement Learning
Dimensionality Reduction
Density Estimation
Clustering



- Fully Observed
- Partially Observed
  - Some variables systematically not observed (e.g. 'topic' of a document)
  - Some variables missing some of the time (e.g. 'faulty sensor' readings)
- Actively collect / sense data (e.g. exploration robots)


ALGORITHMS
DATA
TASKS

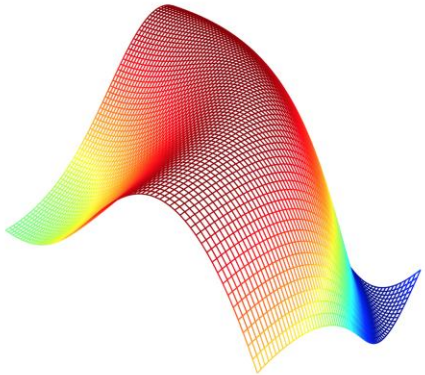
Strategy for fulfilling preferences



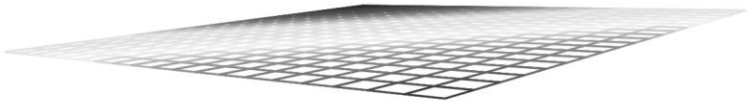
Optimization

Evaluation

Representation



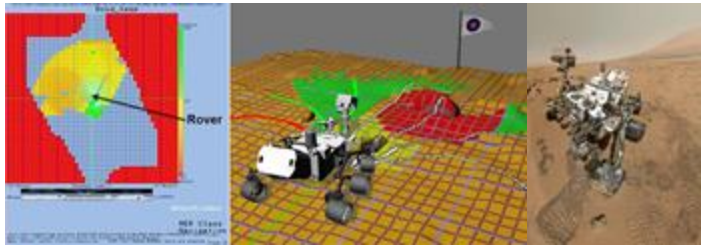
Preferences over the landscape



The landscape of allowed models

https://medium.com/@devnag/machine-learning-representation-evaluation-optimization-fc7b26b38fdb

When to “Learn”

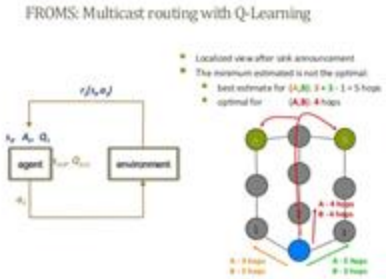


Human expertise does not exist

(‘learning’ to navigate on Mars)

Humans unable to explain their expertise

(‘learning’ to understand speech)



Solution changes over time

(‘learning’ to route network packet traffic)

Solution needs to be adapted to particular cases

(user-specific ‘learning’)

ML is everywhere !

- Wide applicability
- Very large-scale complex systems
  - Internet (billions of nodes)
  - Sensor network (new multi-modal sensing devices)
  - Genetics (human genome)
    - 20,000 genes x 10,000 drugs x 100 species x …
- Improved machine learning algorithms
- Improved H/W
  - data capture (Terabytes, Petabytes of data)
  - Networking
  - faster computers (GPUs)

Pattern Recognition

Machine Learning

Data Science

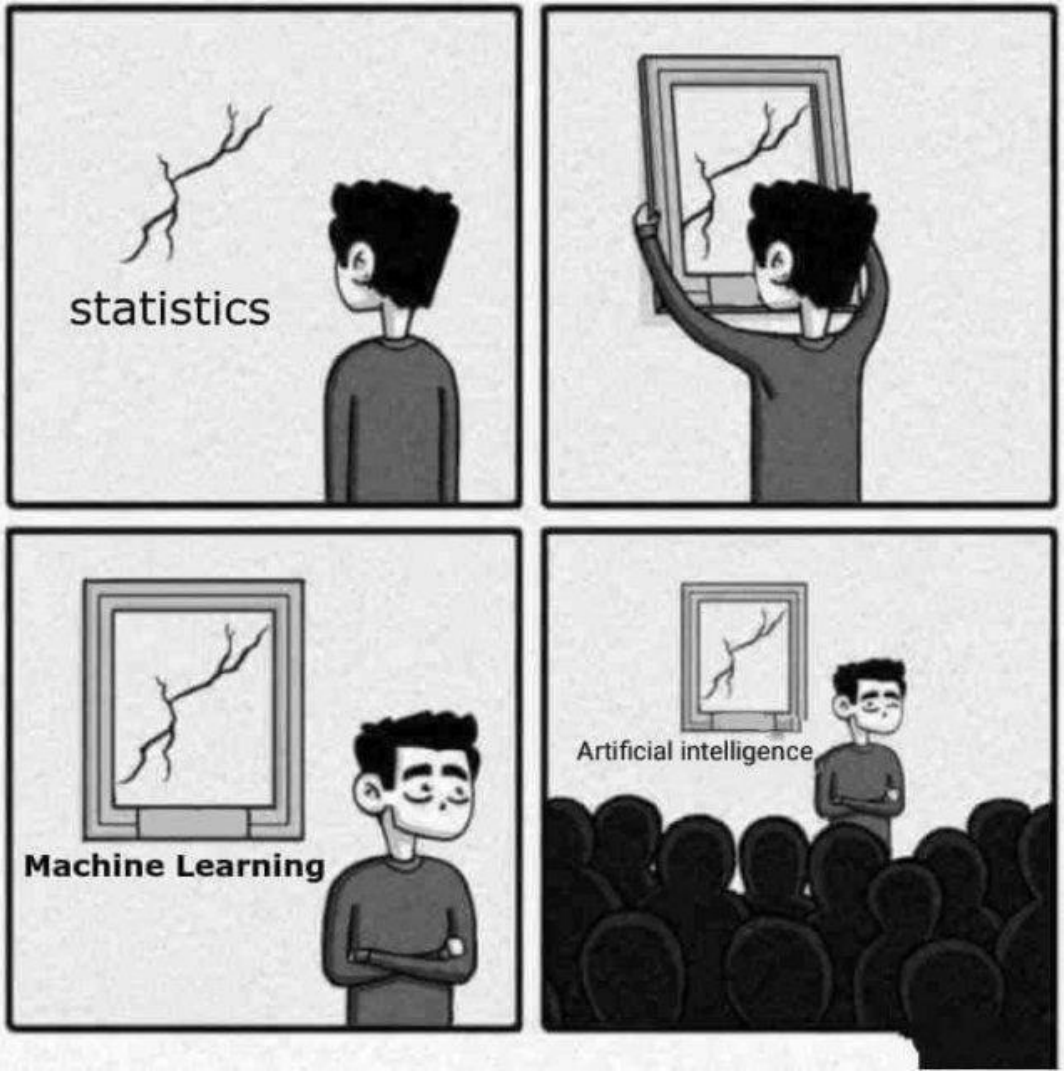
Artificial Intelligence

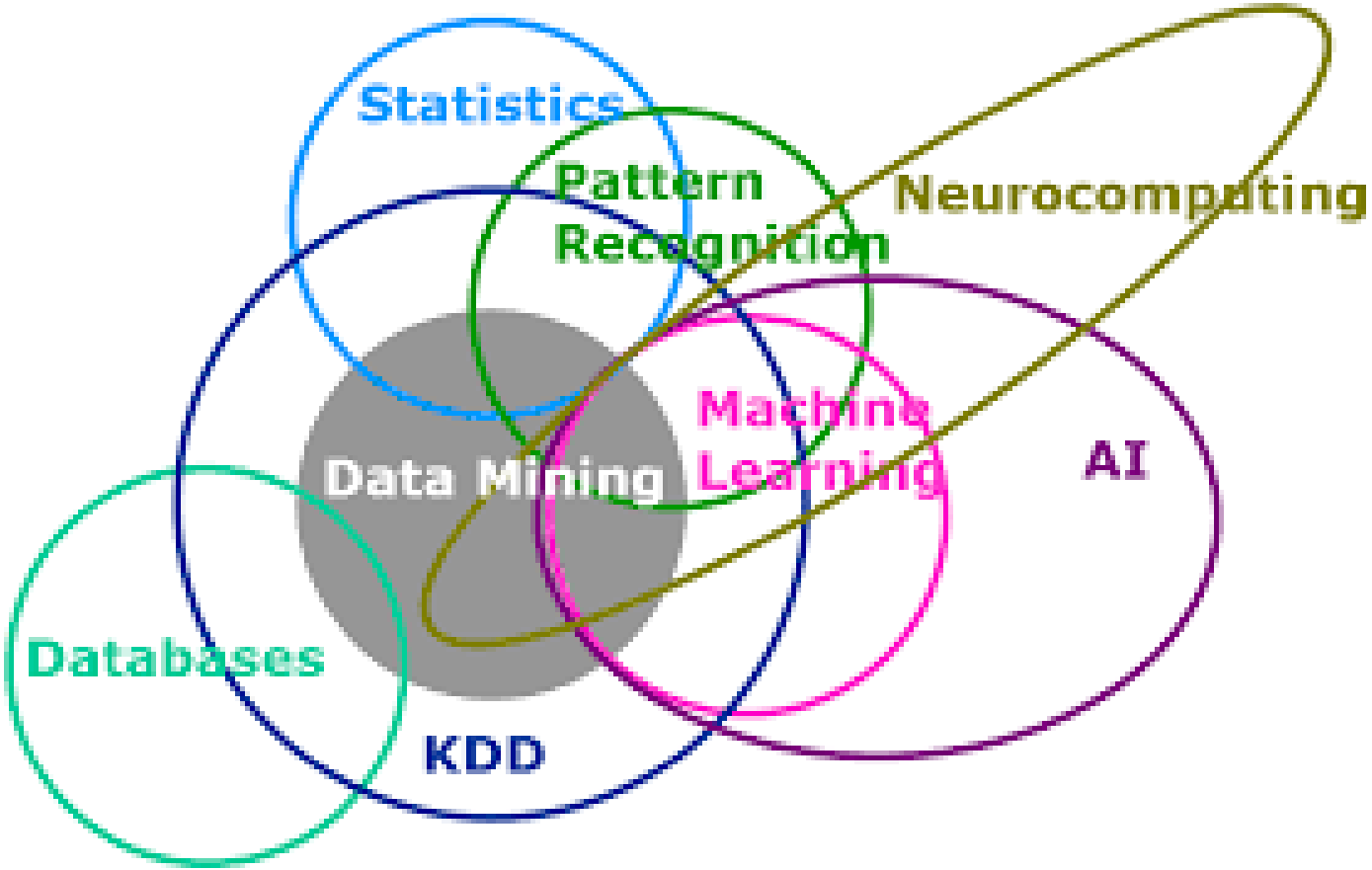
Statistical Learning

Information Retrieval

Data Mining

Statistics

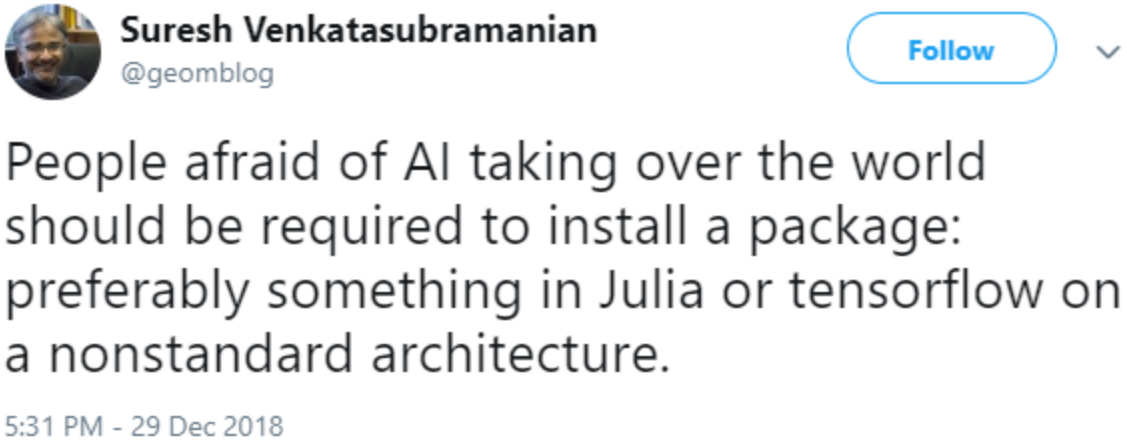

statistics
Machine Learning
Artificial intelligence



https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

Concerns about ML/AI

- Ethics and Bias
- Will AI take over the world ?



About the course (471)

- Timings: Tue, Fri (Himalaya 205, 2.00p – 3.30p)
- Tutorial: Sat, Himalaya 205, 3.30p – 4.30p
- Website/Moodle : https://moodle.iiit.ac.in/course/view.php?id=1472

Course Overview

- Part-1: Supervised Learning
- Part-2: Unsupervised Learning
- Part-3: Neural Networks
- Part-4: Model Selection and Statistical Estimation
- Part-5: ML for sequential data
- Part-6: Case studies [Vision, Robotics, NLP, Speech, Bioinformatics]

Common Themes

- Mathematical framework

  - Well defined concepts based on explicit assumptions

- Representation

  - How do we encode text? Images?

- Model selection

  - Which model should we use? How complex should it be?

- Use of prior knowledge

  - How do we encode our beliefs? How much can we assume?

Pre-requisites

- [MUST]
  - Programming
  - Data Structures
  - Algorithms
- [RECOMMENDED REVISION]
  - Linear Algebra
  - Statistics
  - Probability
  - Calculus

Course Objectives

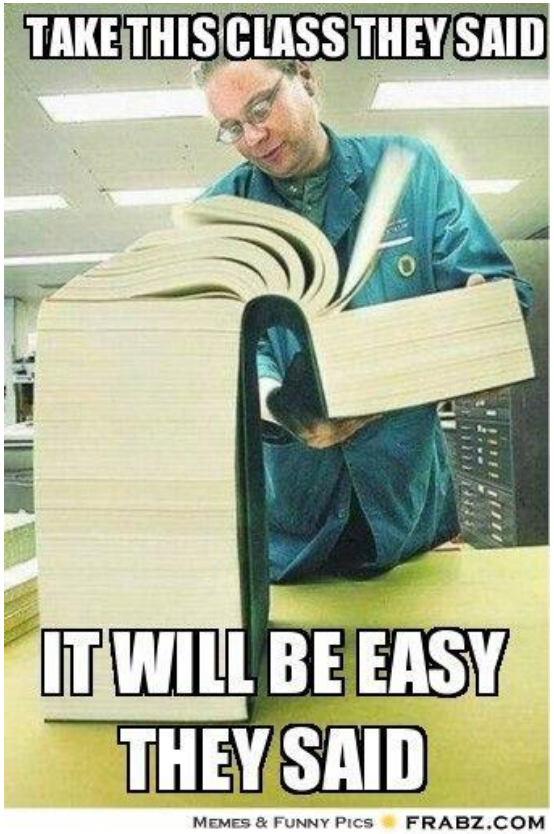
- Determine whether ML is suitable for a problem
- Formulate a problem as a ML problem (data ,representations, tasks, algorithms)
- **Understand** and apply ML method(s)
- Be aware of ML pitfalls, follow best practices
- Be ready to dive deeper (into ML theory or applied areas)

About the course - TAs

- “Sarthak Sharma” <sarthak.sharma@research.iiit.ac.in>
- "Nikhil Gogate" <nikhil.gogate@students.iiit.ac.in>
- "Maheshwari Shubh Jagmohan" <shubh.maheshwari@students.iiit.ac.in>
- "Aditya Aggarwal" <aditya.aggarwal@students.iiit.ac.in>
- "Vachaspati R" <vachaspathi.r@students.iiit.ac.in>
- "Himanshi Sharma" <himanshi.sharma@students.iiit.ac.in>
- "Ranajit Saha" <ranajit.saha@students.iiit.ac.in>
- "Sanjoy Chowdhury" <sanjoy.chowdhury@students.iiit.ac.in>
- "Murtuza Bohra" <murtuza.bohra@research.iiit.ac.in>
- "Avinash Kumar" <avinash.kumar@students.iiit.ac.in>
- "Satyam Mittal" <satyam.mittal@students.iiit.ac.in>
- "Kanay Gupta" <kanay.gupta@students.iiit.ac.in>

About the course – Grading Policy

- Assessment
  - 2 mid semester exams (2x15% = 30%)
  - 1 Final Exam (30%)
  - 13 Assignments (13x2% = 26%)
  - 1 Project (14%)



About the course - assignments

- Code
  - **MATLAB**
  - * Python (scikit-learn + jupyter notebook)
  - Neural Networks: TF, Pytorch, Keras
- This Saturday: Tutorials on MATLAB, Python (tentative)

About the course – collaboration policy

- OK to discuss assignment questions and approaches
- But work must be your own (no copying – partially or fully)
- If you worked with someone, mention their name(s)
- We will be checking for copying/plagiarism
- Better to own up than be caught !

https://twitter.com/terago_networks/status/707283921323429889

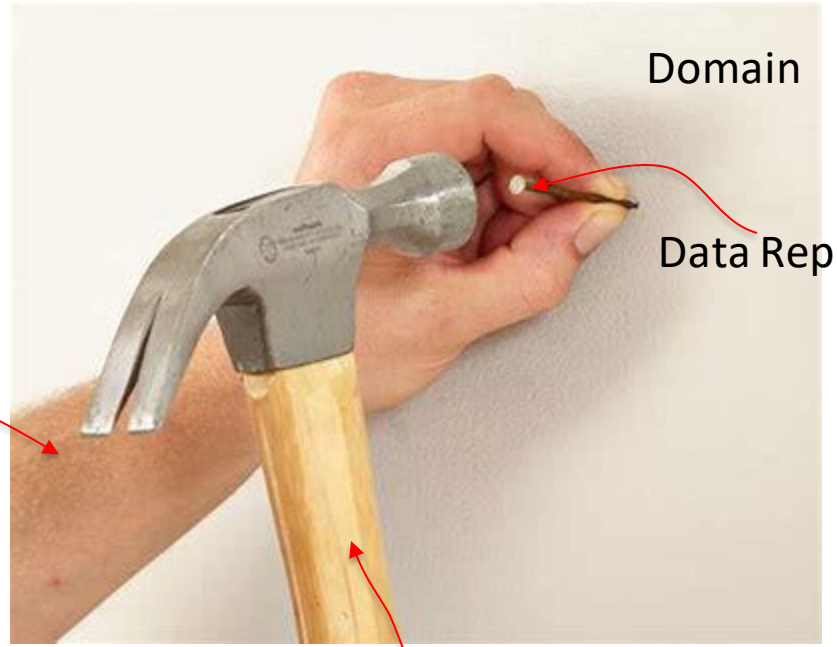
About the course – Grading Policy

- **Homework Late Policy:** 50% if one day late; zero percent if more than one day late
- **A one time late submission bonus:** only applicable to HW (with maximum of three days delay). You must adhere to standard late submission policy after using your bonus. No exceptions will be made. You'll need to inform TAs beforehand if you wish to use the late submission bonus.

About the course - Projects

- Groups (max 4 students/group)

Projects


Domain
Data Representation
SMAI student group
ML approach

Projects

- Problem/Task-first thinking
- Come up with a top-3 list and why the problem is interesting / novel to study
- Be creative / original
- Project-related guidelines will be shared later

…

Some ideas

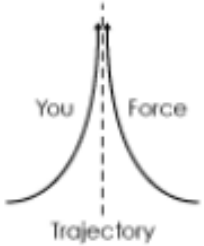
- For each subject exam (including SMAI)
  - collect anonymous data of
    - 1: # of hours studied before an exam
    - 2: # of hours slept before an exam
    - 3: Location in the classroom (# of rows away from front of the room)
    - marks obtained
  - Analyze above data
  - How well can marks be predicted from {1,2,3} above ?

Some ideas

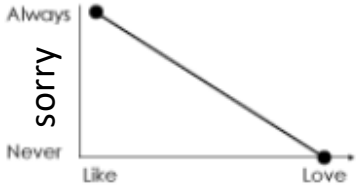
- Predict someone's mother tongue from the way they speak English/sing English songs ☺
- Solve a civic problem
  - E.g. Detect garbage from satellite imagery of Hyderabad
  - Predict traffic density at IIIT/DLF junction
- Predict if photo of biryani is from Paradise ☺

Some ideas

- Generate charts for famous movie quotes

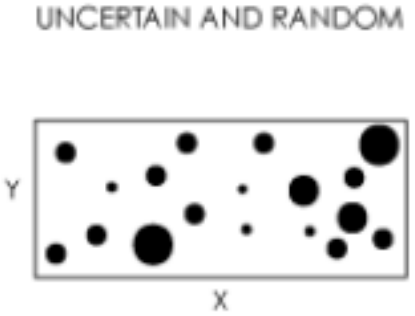


– *Star Wars Episode IV*, 1977



– *Love Story*, 1970

– Dr. No, 1962

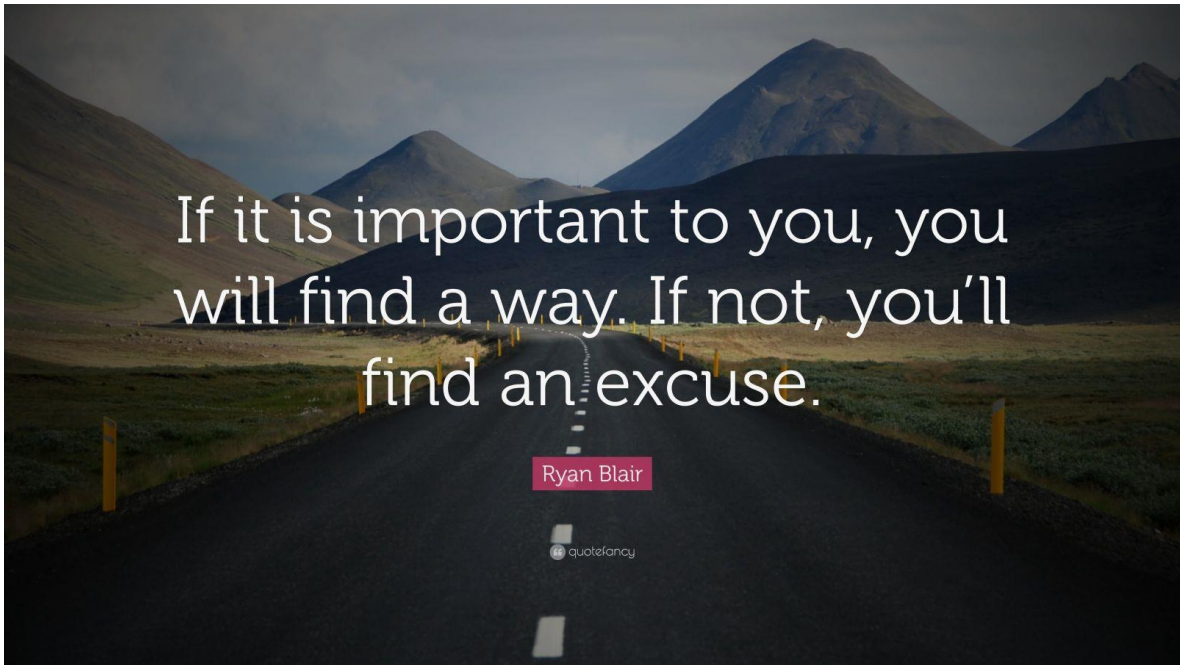


– *Forrest Gump*, 1994

- Indian Language movies ☺

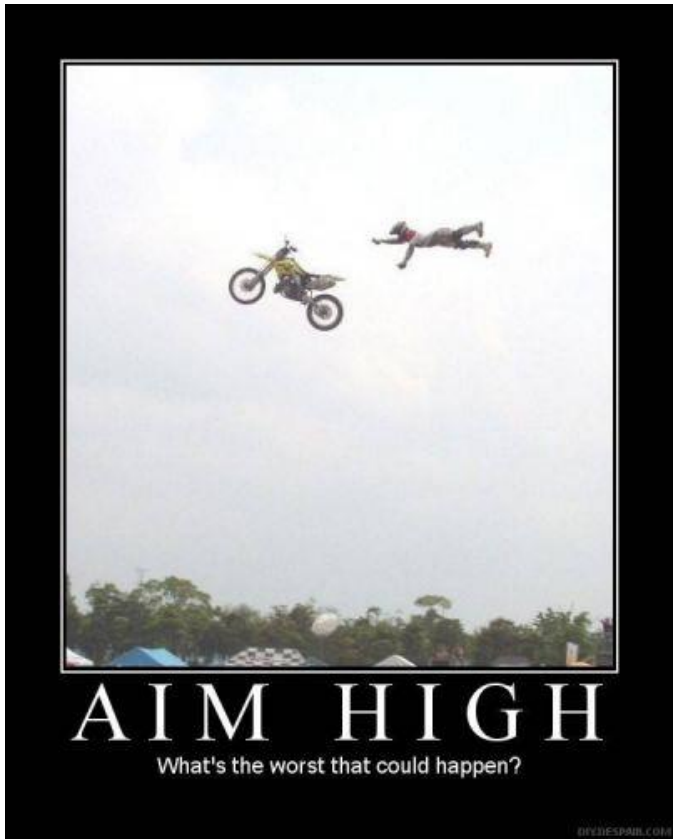
Some ideas

- Given a quote, find an accompanying picture (or pictures) which best `fit' the quote (and vice versa)



Some ideas

- Given a quote, find an accompanying picture (or pictures) which best `fit' the quote



Some ideas

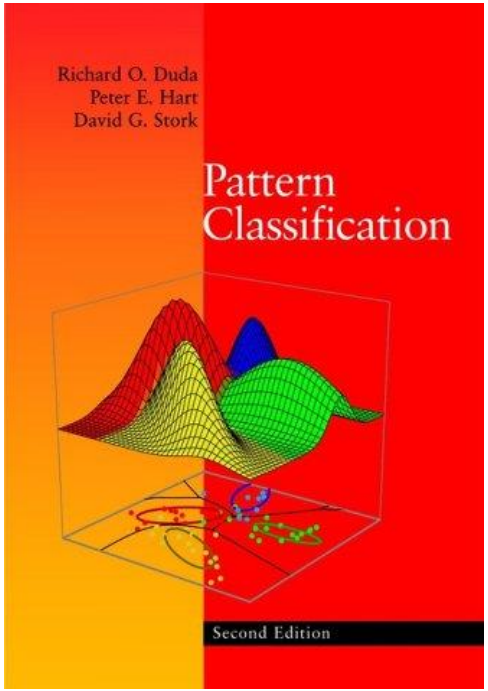
- Generating text whose rhyme matches song lyrics

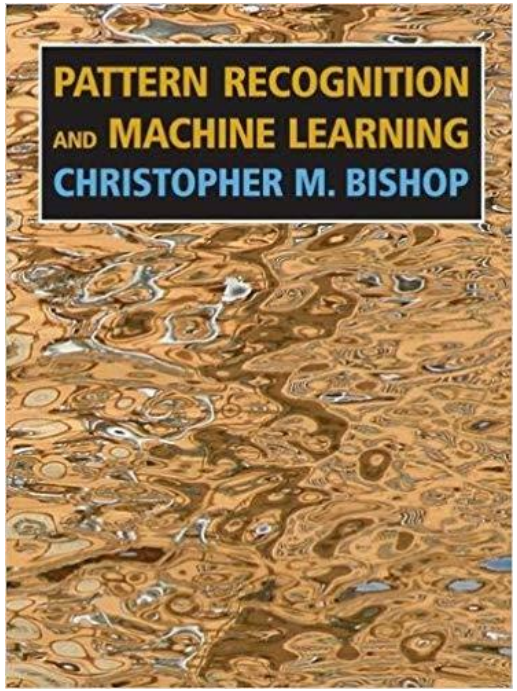
To the tune of Skyfall (by Adele)

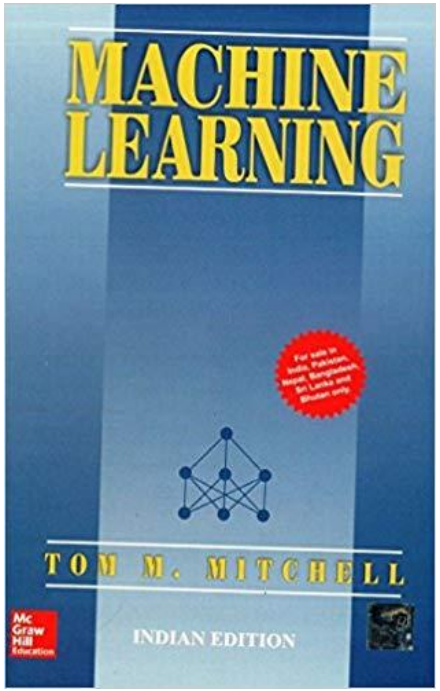
School hell is where we start,
A thousand grades and marks apart,
Where hell collides, and Sundays are dark
You can have my control, You can take my brain
But you'll never have my heart,

Let the grades fall, when they fail,
I will stand tall, and face them all together
Let the grades fall, when they fail
I will stand tall, and face it all together
At School Hell

About the course - Material







About the course - Material

- Will be provided on a per lecture basis
- Scattered Resources across Internet

- SMAI = Introduction to Machine Learning
- Companion courses this semester
  - Artificial Intelligence (371) [Mon, Thu] (UG)
  - Optimization (481) [Mon, Thu]

  - Computer Vision (578) [Tue, Fri]
  - NLP Applications (573) [Mon, Thu]

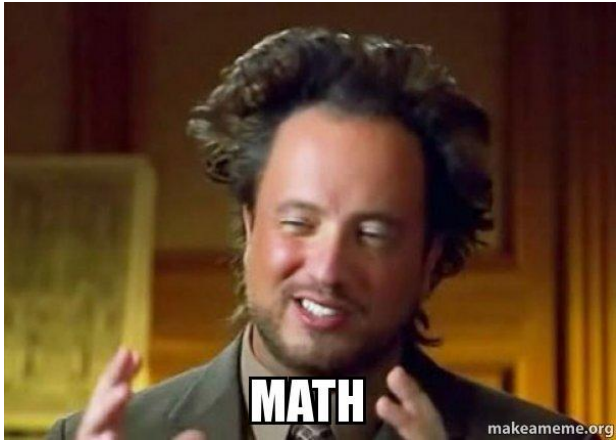
Survey

- For those **seriously** planning to take the course …

- Take the anonymous survey: https://bit.ly/2QOTlMr
- Deadline to submit survey: **Fri 4, 12pm**

- … Understand your background
- … Will help tailor the course content

Additionally …



- **Understand**, don’t just memorize
- Love the math, not the toolbox !
- Capture the broad ideas and insights (useful years down the line)
- Implement ! No substitute for experience.
- Just the beginning ….

A tale of two airplanes

[“The Gimli Glider – 30 years later”](https://www.youtube.com/watch?v=3ffryZAd4Nw)
https://www.youtube.com/watch?v=3ffryZAd4Nw

[“Fatal Flight 447:Chaos in the Cockpit”](https://www.youtube.com/watch?v=YJzg6W2f7Ng)
https://www.youtube.com/watch?v=YJzg6W2f7Ng

OK, so what !

- So many resources (and courses) online
- Why bother taking this course ?
- Answer: Look around you (and me)