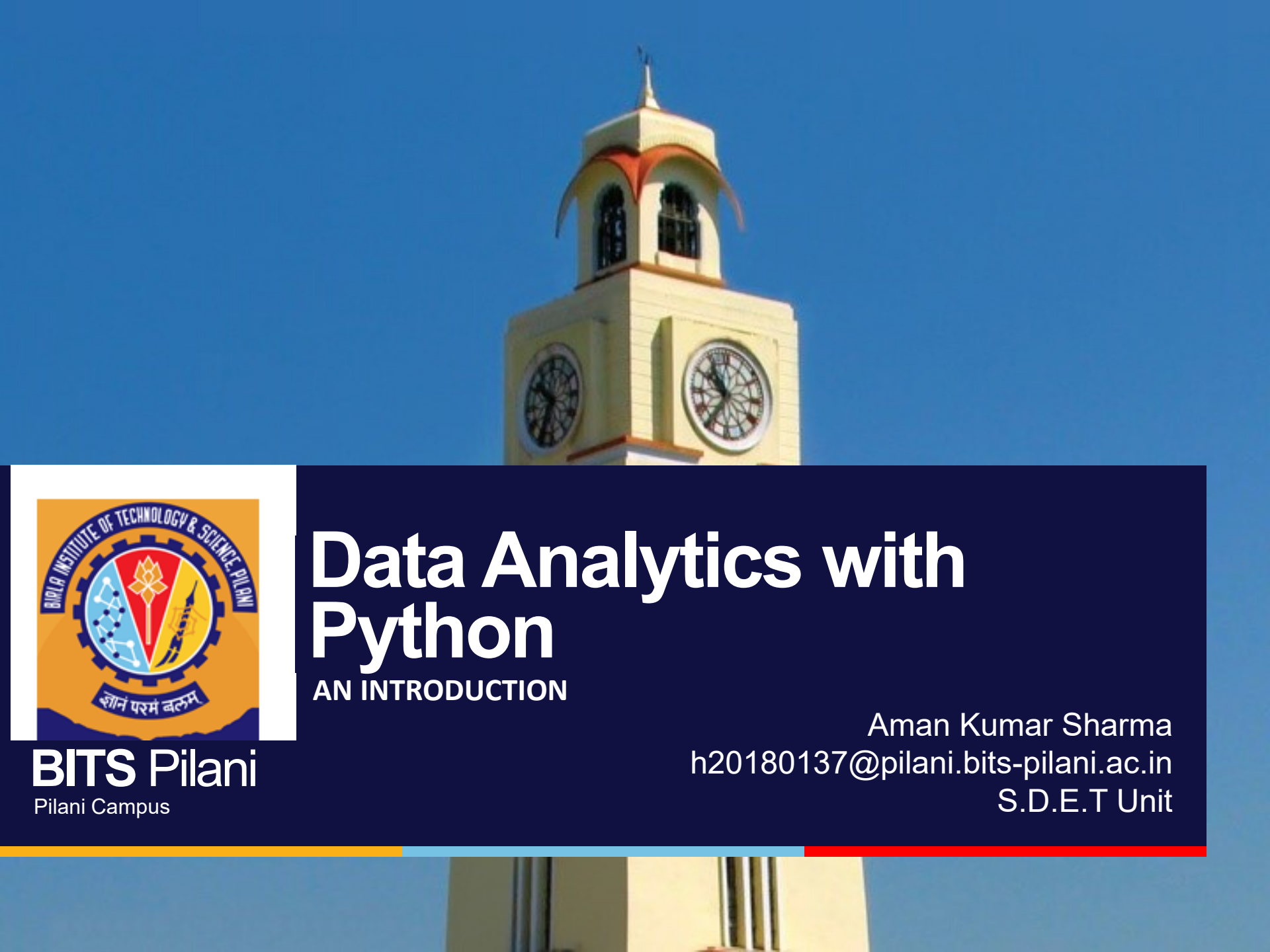


Data Analytics with Python

AN INTRODUCTION

Aman Kumar Sharma
h20180137@pilani.bits-pilani.ac.in
S.D.E.T Unit

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

• Instructors

- **Aman Kumar Sharma** (M.E Computer Science)
- **Rahul Bothra** (B.E Computer Science)
  - ➢ GSoC’18 Sugar Labs

innovate achieve lead

BITS Pilani
Pilani Campus

- Tell us about yourself

About yourself ?

- Your name, branch, year
- Do you know Programming ? If yes, which language(s) ?
- Do you know what is Machine Learning ?
- Do you know basics of Machine Learning ?
- Why did you opt for this course ?

COURSE LOGISTICS

1. **Timing and Venue:** WF 5:30 – 7:00 (±00:30), NAB 6108
2. **Course Website:** nalanda.bits-pilani.ac.in
3. **Discussion Site:** Nalanda + WhatsApp
4. **Background assumed:** Basics of Linear Algebra , probability and statistics. (Do not worry)
5. **Grading :**

Component	Duration	Weightage(%)	CB/OB
2 quizzes	TBA	20	TBA
Mid Sem Exam	60 mins	25	CB
Project	-	20	OB
Comprehensive Exam	90 mins	35	OB + CB

Make-up Policy: Make-up shall be granted in genuine cases with prior notification to the Instructor

Course Structure

- The course will teach you Python and Applications of Machine Learning using Python.
- The first half (almost) will cover Python. After the first half, you will be able to do **ANYTHING** with Python.
- The second half will introduce you to Machine Learning and its applications.

Course Study Material

- Text Books:
  - *Guido Van Rossum. "Python 3.7.2 Documentation"*
  - *A. Muller, "Intro to Machine Learning with Python: A Guide for Data Scientists"*
- Reference Books:
  - *Allen B. Downey, "Think Python: How to Think Like a Computer Scientist"*
  - *R2. Ian Goodfellow, Yoshua B., Aaron C., "Deep Learning"*
- All slides will be shared (Yay!)

innovate achieve lead

BITS Pilani

Pilani Campus

Intro to Machine Learning

Brief Overview

INTRODUCTION TO MACHINE LEARNING

- JARGON / TERMINOLOGIES
- MATHEMATICS FOR ML
- TOOLS

SUPERVISED LEARNING

- CLASSIFICATION VS REGRESSION
- SUPERVISED ML ALGO : KNN, DECISION TREE, SVM, KERNELIZED SVM, ARTIFICIAL NEURAL NETWORKS

UNSUPERVISED LEARNING AND PREPROCESSING

- TYPES OF UNSUPERVISED LEARNING – TRANSFORMATION OF DATASET AND CLUSTERING

Continued…

MODEL EVALUATION AND IMPROVEMENT

- CROSS – VALIDATION
- GRID SEARCH
- EVALUATION METRICS AND SCORING

MISCELLANEOUS / SELECTED TOPICS

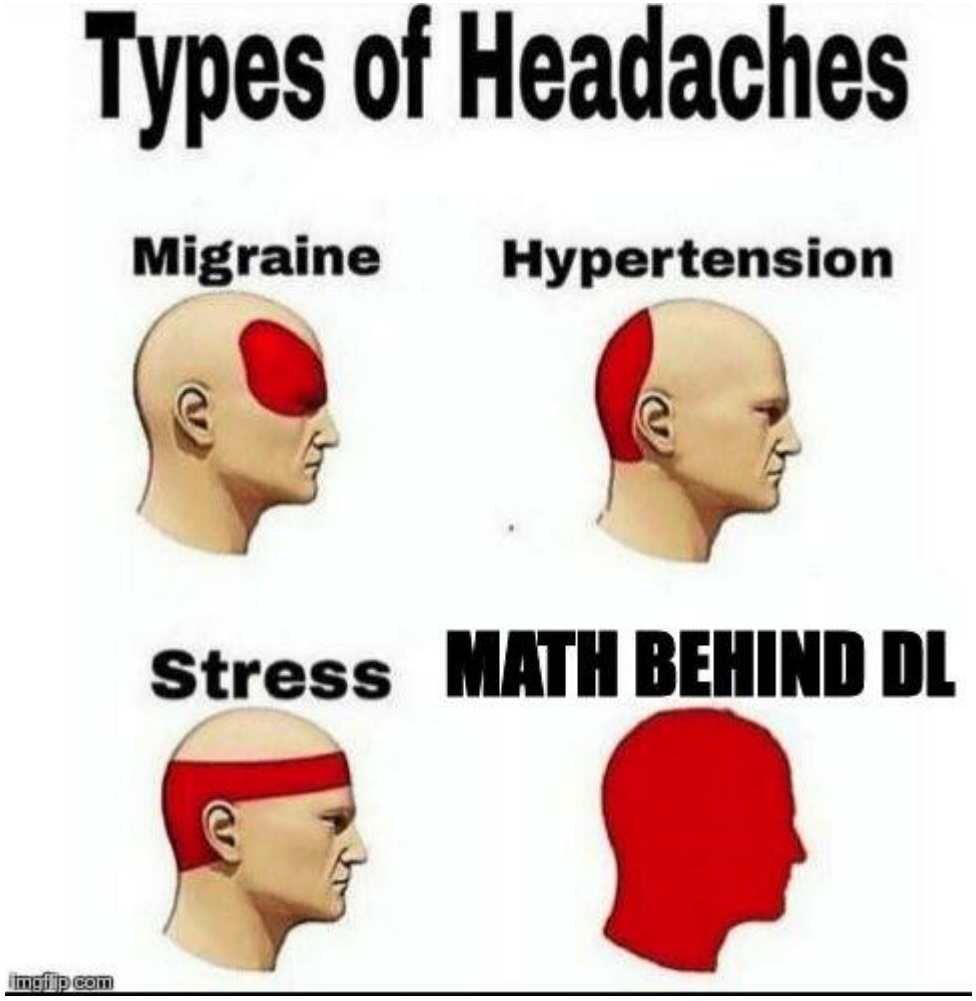
- SEMI-SUPERVISED LEARNING
- TBA

CASE STUDIES

- DEEP LEARNING BASED EXAMPLES

Course Goals

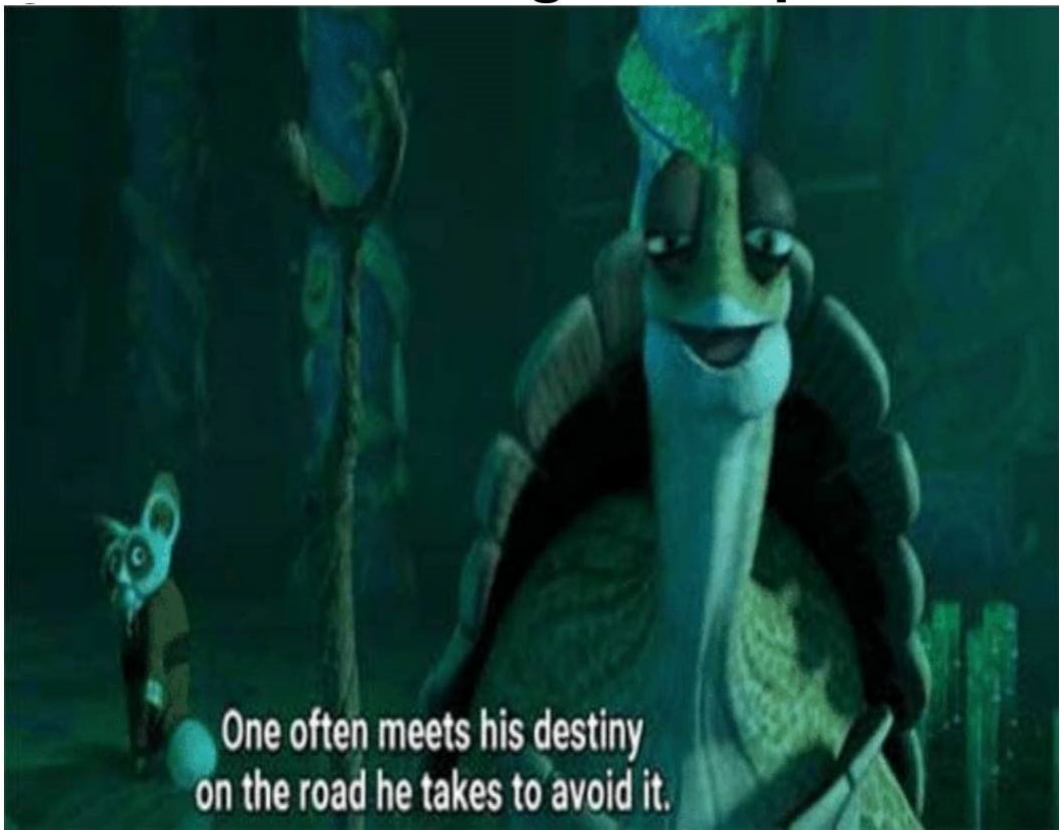
- By the end of the semester, you should be able to:
  - Understand how various machine learning algorithms work
  - Implement them (and, hopefully, their variants/improvements) on your own
  - Look at a real-world problem and identify if ML is an appropriate solution
  - If so, identify what types of algorithms might be applicable
  - **Feel inspired to work on and learn more about Machine Learning :-)**
- This class is not about:
  - Mathematics behind Machine Learning
  - Deep Learning (Advanced ML Topics)


Types of Headaches
Migraine
Hypertension
Stress
MATH BEHIND DL
imgflip.com

Course Goals (Extra)

- Decipher papers on Machine Learning.
- Build cool applications to impress your friends.
- Debunk a lot of garbage that is in the media today about ML.
- **But most importantly, you'll be able to understand Machine Learning Memes.**

When you accidentally flip the sign of your gradient update but still land on the global optimum



Introduction

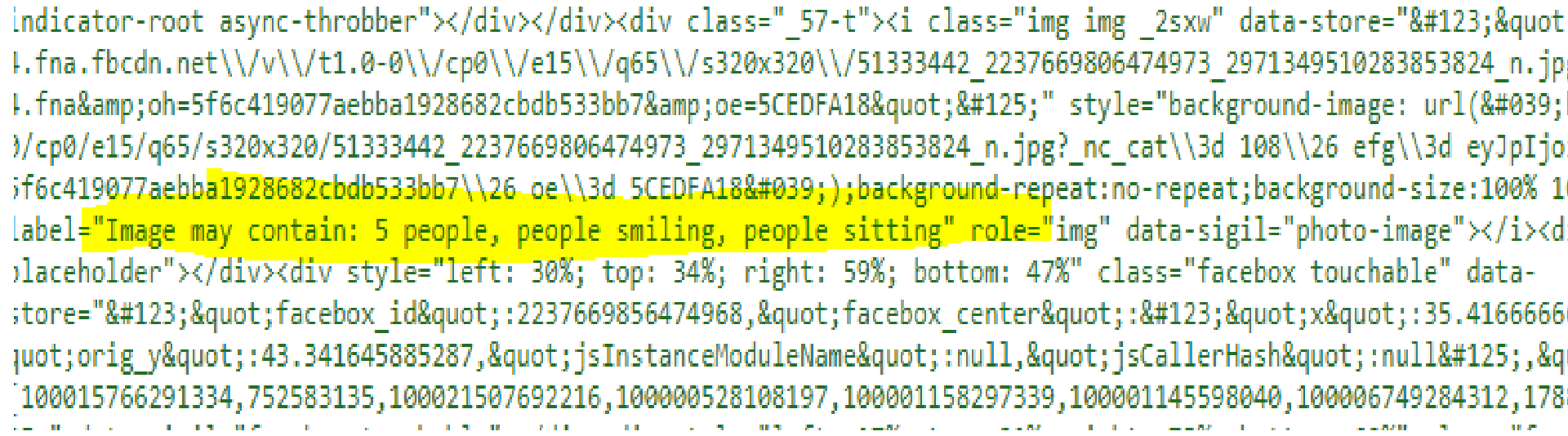
1. What is Machine Learning ?
2. Why is everyone talking about it ?
3. Why should it matter to me ?

IT'S UBIQUITIOUS

ML IN EVERYDAY LIFE

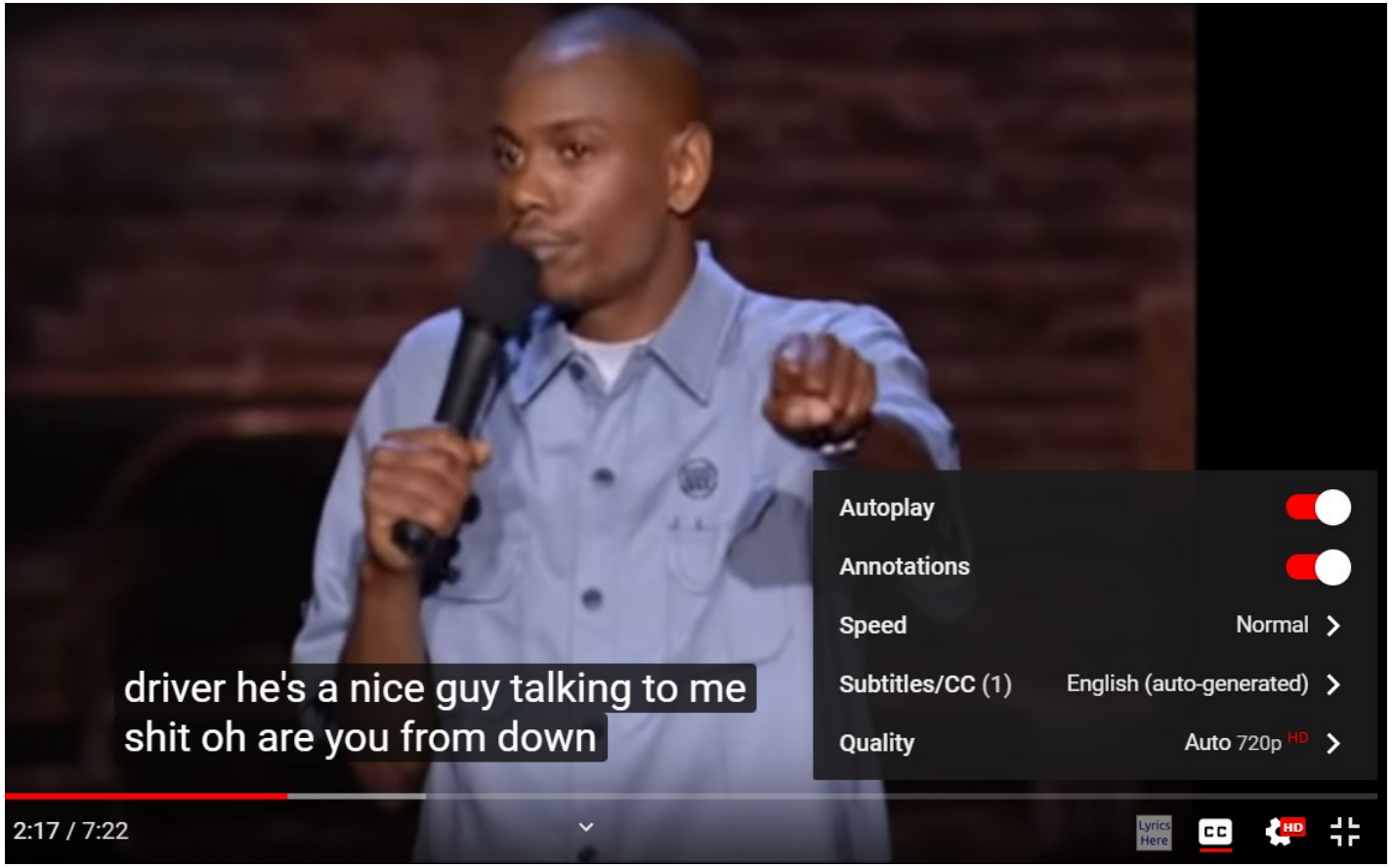
- Voice Assistants [Duplex] : Signal Processing + NLP + Deep Neural Nets (RNNs)

- Facebook : Image Captioning (CNN + RNN)

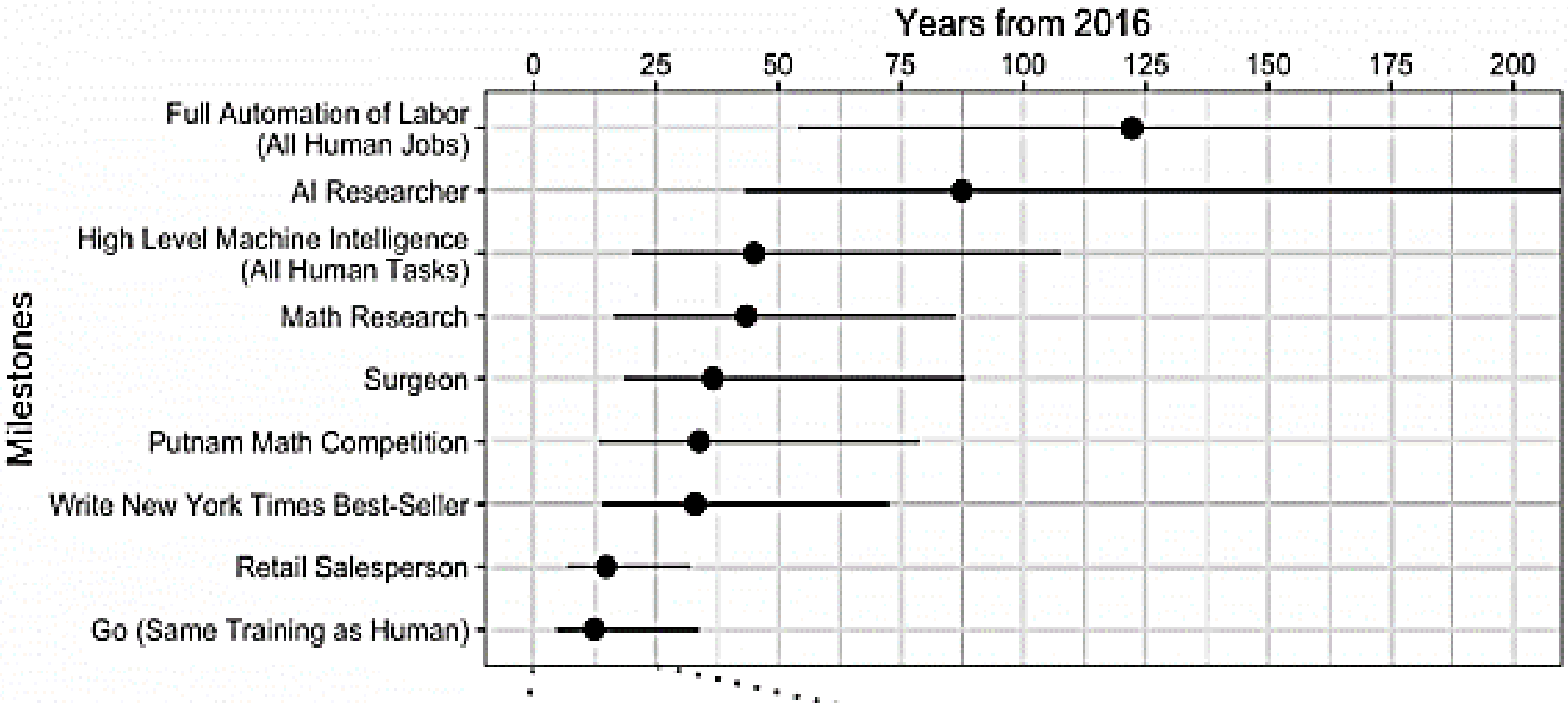

indicator-root async-throbber"> <i class="img img _2sxw" data-store="{&quot
4.fna.fbcdn.net\\/v\\/t1.0-0\\/cp0\\/e15\\/q65\\/s320x320\\/51333442_2237669806474973_2971349510283853824_n.jp
4.fna&oh=5f6c419077aebba1928682cbdb533bb7&oe=5CEDFA18"}" style="background-image: url('
)/cp0/e15/q65/s320x320/51333442_2237669806474973_2971349510283853824_n.jpg?_nc_cat\\3d 108\\26 efg\\3d eyJpIjo
5f6c419077aebba1928682cbdb533bb7\\26 oe\\3d 5CEDFA18');background-repeat:no-repeat;background-size:100% 1
label="Image may contain: 5 people, people smiling, people sitting" role="img" data-sigil="photo-image"></i><d
)laceholder"> <div style="left: 30%; top: 34%; right: 59%; bottom: 47%" class="facebox touchable" data-
store="{"facebox_id":2237669856474968,"facebox_center":{"x":35.4166666
quot;orig_y":43.341645885287,"jsInstanceModuleName":null,"jsCallerHash":null},&q
100015766291334,752583135,100021507692216,100000528108197,100001158297339,100001145598040,100006749284312,178

- Amazon : Recommender System (Collaborative Filtering)

- YouTube : Video Auto-Caption (NLP + RNNs)

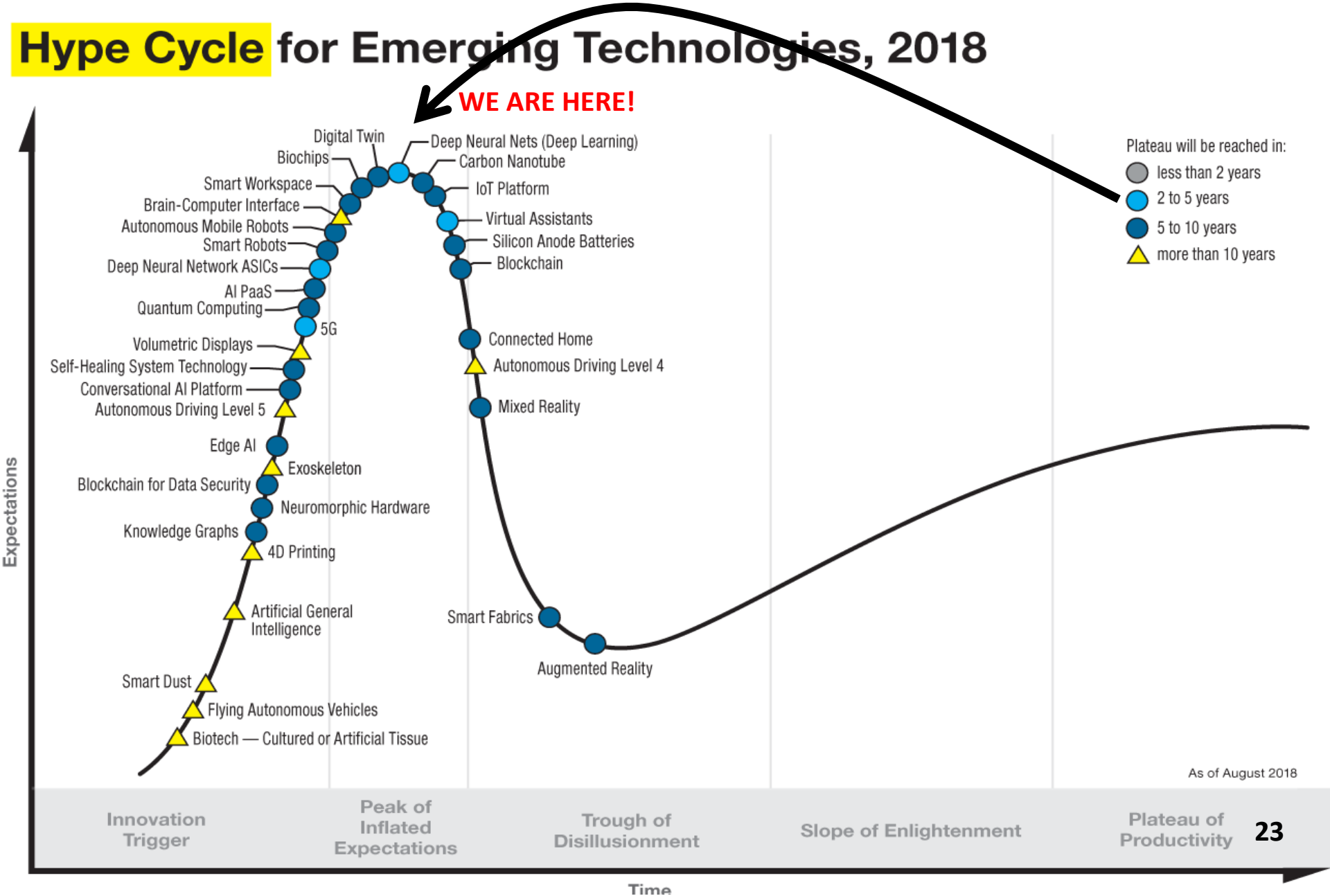


WHERE IS THIS HEADED ?



Ref: arxiv.org/abs/1705.08807 : When Will AI Exceed Human Performance? Evidence from AI Experts

Hype Cycle for Emerging Technologies, 2018



INFACT,

- No sector in the commercial industry would be left untouched by Machine Learning in 3-5 years.
- Data Analysis would become a fundamental prerequisite in most of the jobs.
- Above all, knowledge is power and ignorance is not bliss.

Machine Learning : An Interdisciplinary Field

It draws on results from

- Artificial Intelligence
- Probability & Statistics
- Computational Complexity Theory
- Control Theory
- Information Theory
- Philosophy
- Neurobiology

Machine Learning :

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

{ Tom Mitchell, Prof at CMU}

Example: **A handwriting recognition learning problem.**

- **Task T:** recognizing and classifying handwritten words within images
- **Performance measure P:** percent of words correctly classified
- **Training experience E:** a database of handwritten words with given classifications.

Example: **A robot driving learning problem**

- **Task T:** driving on public four-lane highways using vision sensors.
- **Performance measure P:** average distance traveled before an error (as judged by human overseer)
- **Training experience E:** a sequence of images and steering commands recorded while observing a human driver

Machine Learning Systems - Taxonomy

- Based on :
  - Whether or not they are trained with human supervision **(Supervised, Unsupervised, Semi-supervised, Reinforcement Learning)**
  - Whether or not they can learn incrementally on the fly **(Online Learning vs Batch Learning)**
  - Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do **(instance-based vs model-based learning)**

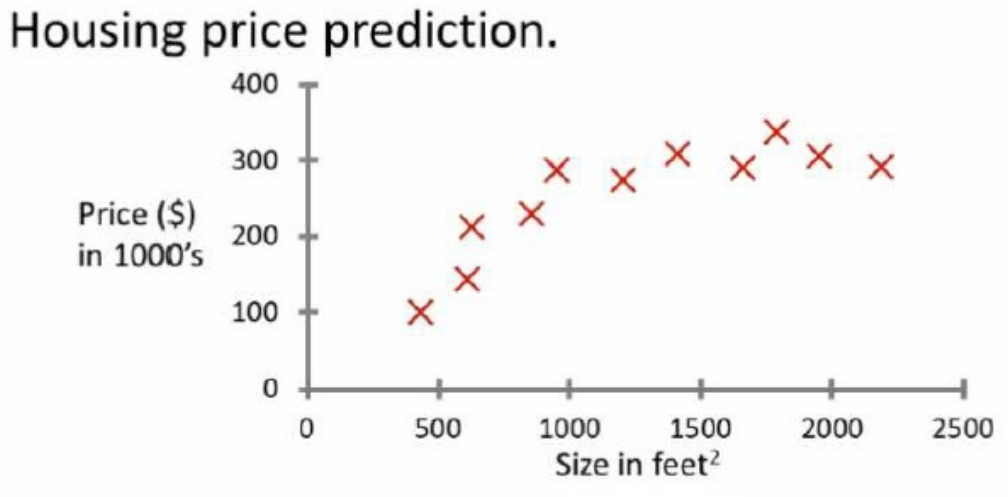
TYPES OF MACHINE LEARNING

based on the amount and type of supervision they get during training

innovate achieve lead

• SUPERVISED LEARNING

- Given: Training data as labeled instances $\{(x_1, y_1), \ldots, (x_N, y_N)\}$
- Goal: Learn a rule $(f: x \rightarrow y)$to predict outputs y for new inputs x.
- Real-valued outputs (e.g., price of a house): **Regression**

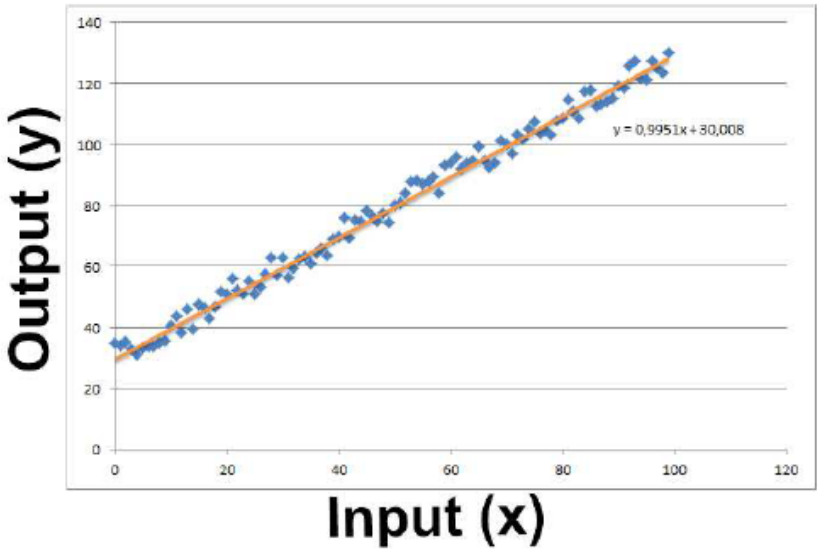


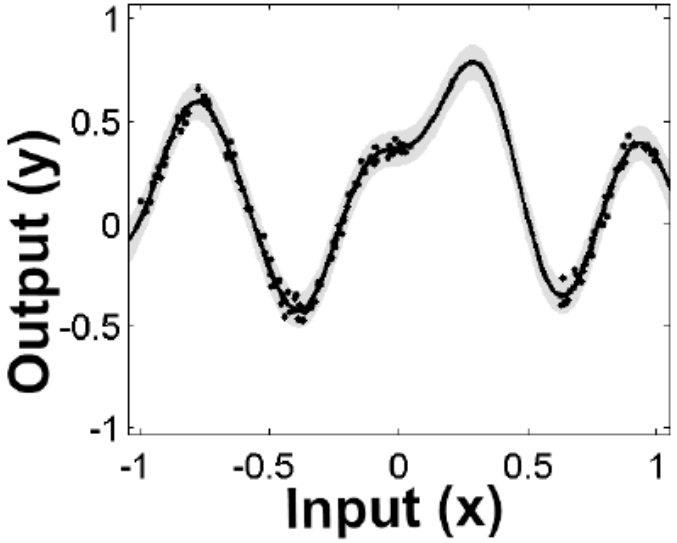
Supervised Learning (Cont....)

- Discrete-valued outputs (e.g., label of a hand-written digit): **Classification**

Supervised Learning (Pictorially)

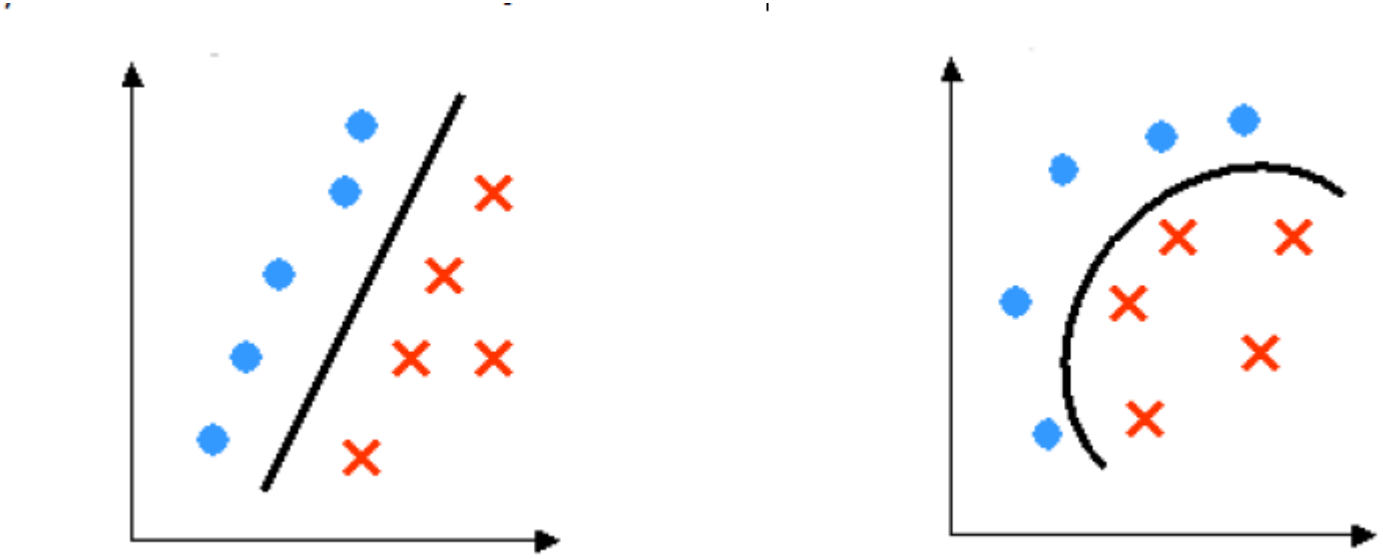
- Regression: fitting a line/non-linear curve





Supervised Learning (Pictorially)

- Classification: finding a linear/nonlinear separator

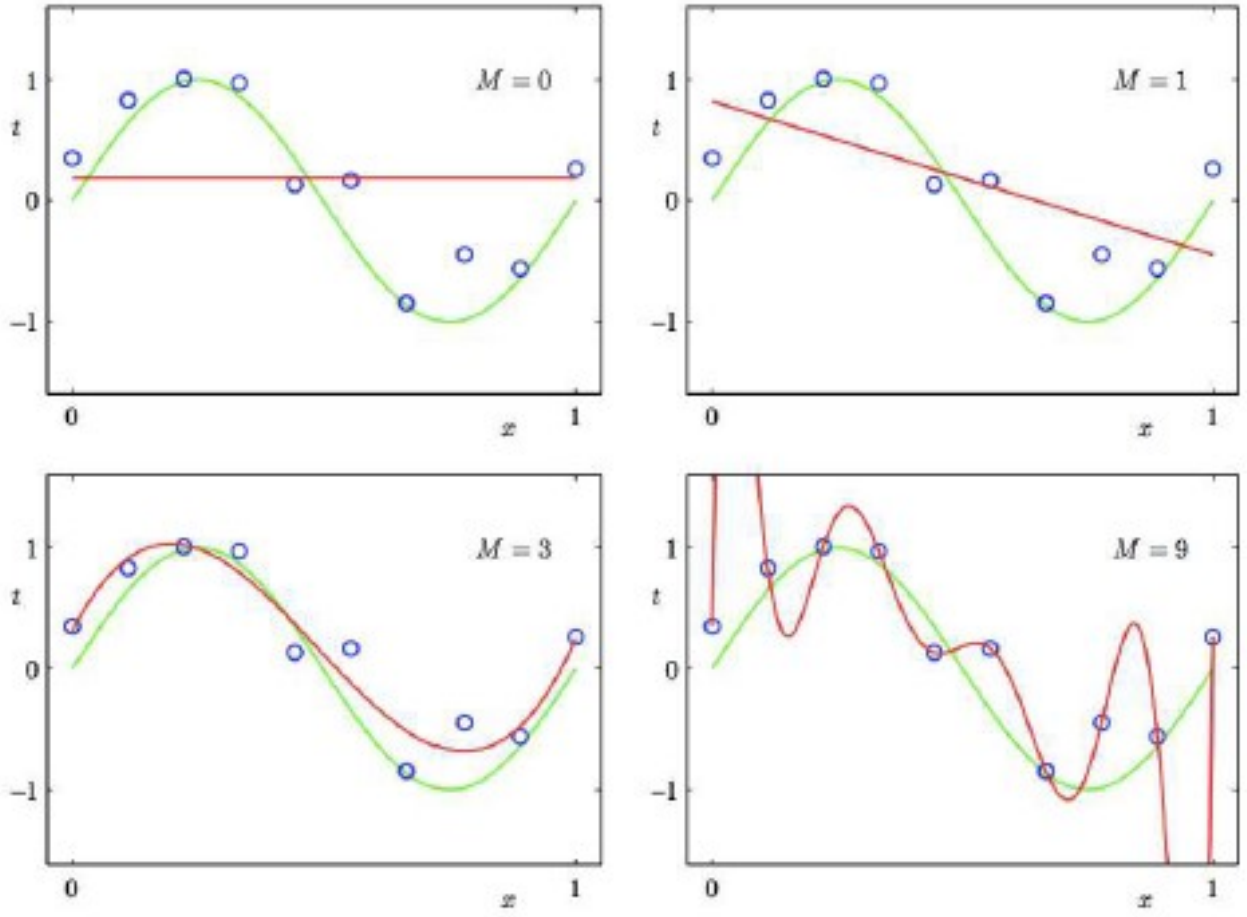


- Generalization is crucial (must do well on test data)

Some Supervised Algorithms

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks

Generalization

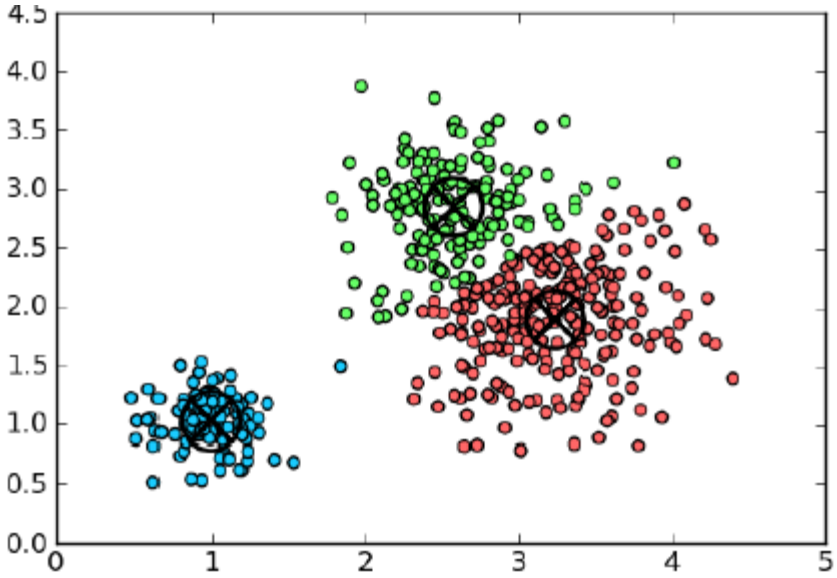


Desired: hypotheses that are not too simple, not too complex (to avoid **overfitting** on training data)

TYPES OF MACHINE LEARNING

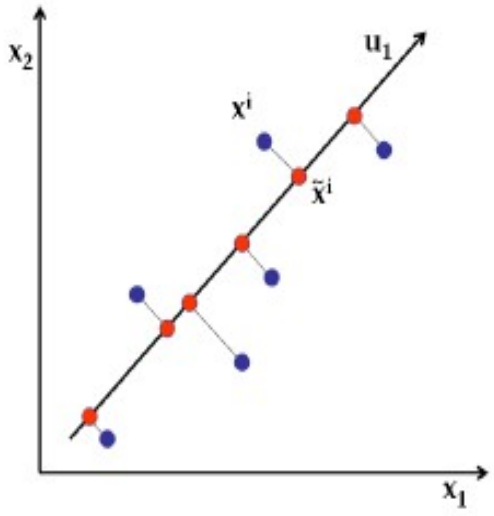
• UNSUPERVISED LEARNING

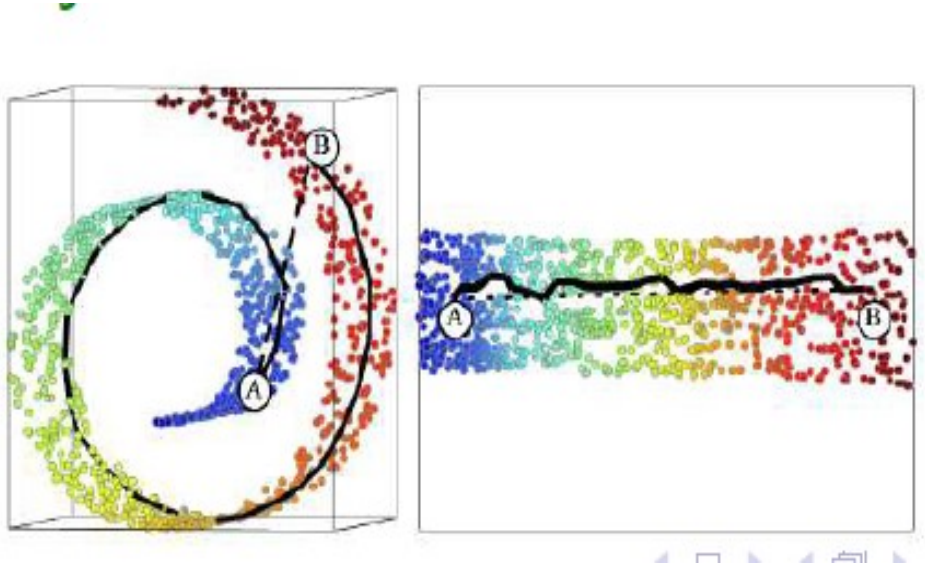
- Given: Training data in form of unlabeled instances $\{ x_1, \ldots, x_N \}$
- Goal: Learn the intrinsic latent structure that summarizes/explains data
- Homogeneous groups as latent structure: **Clustering**



Unsupervised Learning (Cont.…)

- Low-dimensional latent structure: **Dimensionality Reduction**





Unsupervised Learning (Examples)

- Clustering large collections of images

- Also used as a preprocessing step for many supervised learning algorithms (e.g., to learn/extract) good features, to speed up the algorithms, etc.)
- CURSE OF DIMENSIONALITY

Some Unsupervised Learning Algorithms

- Clustering
  - k-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization
- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
  - Apriori
  - Eclat

TYPES OF MACHINE LEARNING

innovate achieve lead

• SEMI-SUPERVISED LEARNING

- Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. This is called **semi-supervised learning**
- **Ex : Google Photos**
- Most semi-supervised learning algorithms are combinations of unsupervised and supervised algorithms.

TYPES OF MACHINE LEARNING

• REINFORCEMENT LEARNING

- Reinforcement Learning is a very different beast.
- The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards.
- It must then learn by itself what is the best strategy, called a policy, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.

- There are an astonishing 10 to the power of 170 possible board configurations - more than the number of atoms in the known universe - making Go a **googol** times more complex than Chess.
- AlphaGo Zero learnt to play the game of Go simply by playing games against itself, starting from completely random play.

TYPES OF MACHINE LEARNING
Based on if they learn on the fly

innovate achieve lead

- BATCH LEARNING

- In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data.
- This will generally take a lot of time and computing resources, so it is typically done offline.
- First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called offline learning.

TYPES OF MACHINE LEARNING

• ONLINE LEARNING

- In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.
- Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.
- **Online learning is great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously.**
- It is also a good option if you have limited computing resources: once an online learning system has learned about new data instances, it does not need them anymore,

TYPES OF MACHINE LEARNING
Based on how they generalize

innovate achieve lead

- INSTANCE-BASED LEARNING

- Most trivial form of learning is simply to learn by heart.
- If you were to create a spam filter this way, it would just flag all emails that are identical to emails that have already been flagged by users— not the worst solution, but certainly not the best.
- Instead of just flagging emails that are identical to known spam emails, your spam filter could be programmed to also flag emails that are very similar to known spam emails. This requires a measure of similarity between two emails.

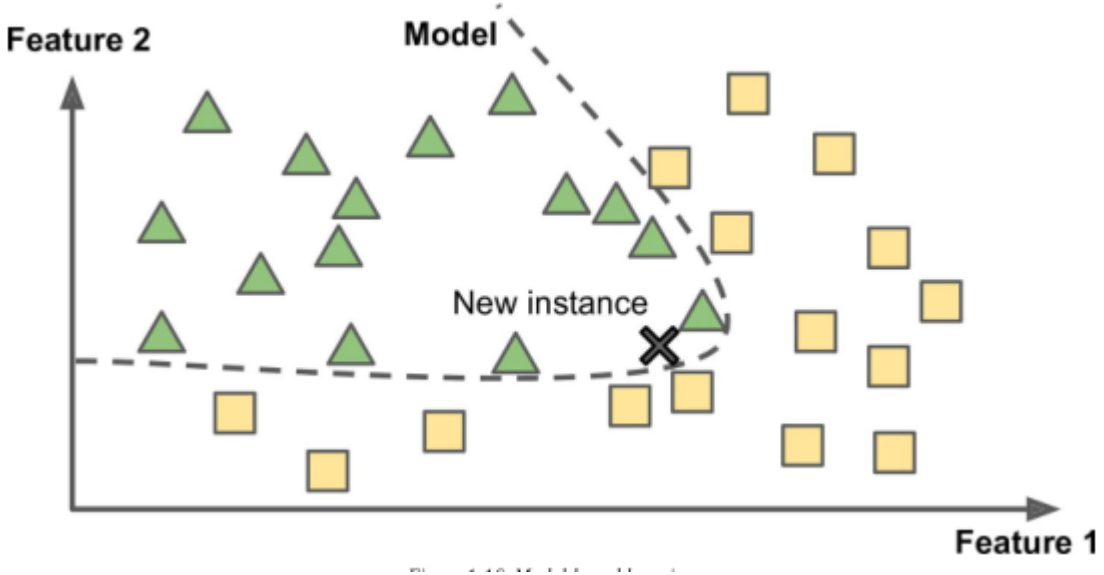
Contd..

- A (very basic) similarity measure between two emails could be to count the number of words they have in common.
- The system would flag an email as spam if it has many words in common with a known spam email.
- This is called instance-based learning: the **system learns the examples by heart, then generalizes to new cases using a similarity measure**

TYPES OF MACHINE LEARNING

• MODEL-BASED LEARNING

- Another way to generalize from a set of examples is to build a model of these examples, then use that model to make predictions. This is called model-based learning.



To Do

- **DOWNLOAD AND INSTALL ANACONDA.**
- **DON'T FORGET TO BRING YOUR MACHINES NEXT TIME ITS TIME TO TEACH THEM.**

Tentative List of Topics

- Supervised Learning
  - nearest-neighbors methods, decision trees, naïve Bayes
  - linear/non-linear regression and classification
    - SVM, Kernelized SVM, Neural Networks ,
- Unsupervised Learning
  - Clustering and density estimation
    - K-Means Clustering
    - Agglomerative Clustering
    - DBSCAN
  - Dimensionality reduction and manifold learning
    - PCA
    - Non Negative Matrix Factorization
    - Manifold Learning with t-SNE
  - Latent factor models and matrix factorization
- Ensemble Methods
- Deep Learning


That's all Folks!