Machine learning and AI meetings introduction

James D Pearce • University of Victoria • jdpearce@uvic.ca


WHAT ARE THESE MEETINGS ABOUT?
learning
kernels
models
problems
protein
splicing
structured
methods
prediction
Learning
individual
collaboration
machine
classification
sequence
alignment
alternative
work
organisms
sequences
multiple DNA
biology
approach
problem
solving
Systems
model
output
ML
data
also
developed
dynamical
graphical
number
gene
challenging
micro-exons
train
learn
Machine
PALMA
derived
research
predicting
sequencing
optimization
task
information
experts
tissue
several fields
important
genome
Alternative
localization
events
optimizing
partly
accurate
results
detection
splicegraphs
subjects
training
tasks
genomic
corresponding
explicit
experiments
outputs
image
interested
parts
space
Imaging
Consequently
predictions
intensive
features
shown
components
scenario
causal
activity
automated
investigate
Biology
discovery
scientific
fMRI
higher
different
obtain
highly
Protein
novel
success
positive
done
motifs
classifier
knowledge
subcellular
SVM
nuclei
analysis
similarity
mechanisms
biological
margin
idea
projects
definiteness
mGene
well
View
new
computational
University
consider
one
dynamic
domain
objects
intractable
alignments
levels
are
project

Official doctrine: “Provide an opportunity for students to learn about machine learning and AI algorithms through discussions, tutorials, and competitions.”

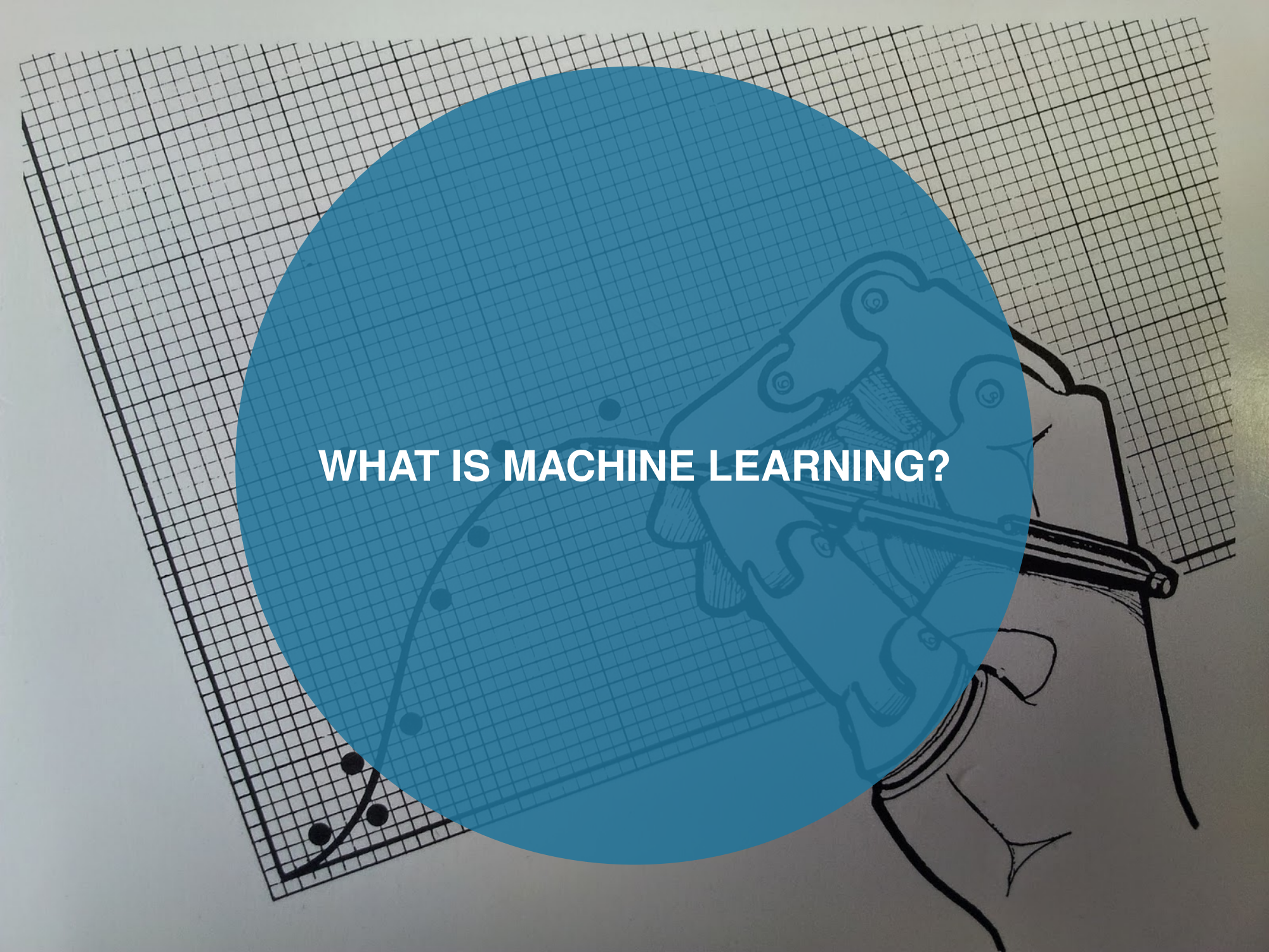
- **Talks:** Any and everyone is encouraged to give a talk. They don’t have to be long and you don’t need to be an expert!
- **Competitions:** For now we are focusing on Kaggle competitions. They tend to run for months and some have very large cash prizes.
- **Tutorials:** We’ll show you how to get started in Kaggle competitions.

Resources: People and resources will be shared between these Tuesday meetings and the regular Programming club Wednesday meetings. However, these two meetings will run independently, i.e. you shouldn't have to show up to both.

- **Programming club Website:** `http://uvic-programming-club.github.io/`
- **Email:** `uvic.programming.club@gmail.com`
- **Communications:** `https://uvic.slack.com`
- **Competitions:** `kaggle.com`
- **ML python package:** `http://scikit-learn.org/stable/`

Getting terms straight:

- **AI:** “The study and design of intelligent agents.”
- **Machine Learning:** “A branch of artificial intelligence; concerns the construction and study of systems that can learn from data.”
- **Data Mining:** Focuses on the discovery of previously unknown properties in the data.
- **Big Data:** Buzzword to describe the exponential increase of data and the inadequacy of database systems to scale at the same rate.
- **Data Science:** Catch-all phrase to describe the study of data analysis.


WHAT IS MACHINE LEARNING?

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*”

- Mitchell, T. (1997). Machine Learning, McGraw Hill

Example:

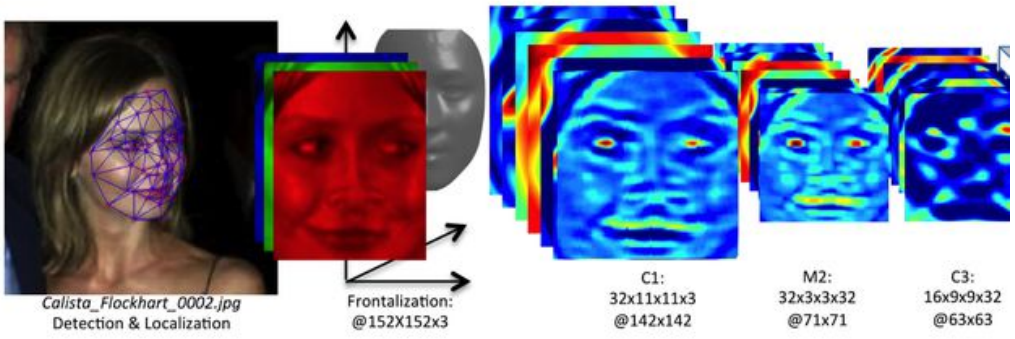
- *T*: classify handwritten digits 0-9
- *E*: Data set of handwritten digit images with associated human generated labels 0-9
- *P*: Accuracy of classification with respect to associated labels

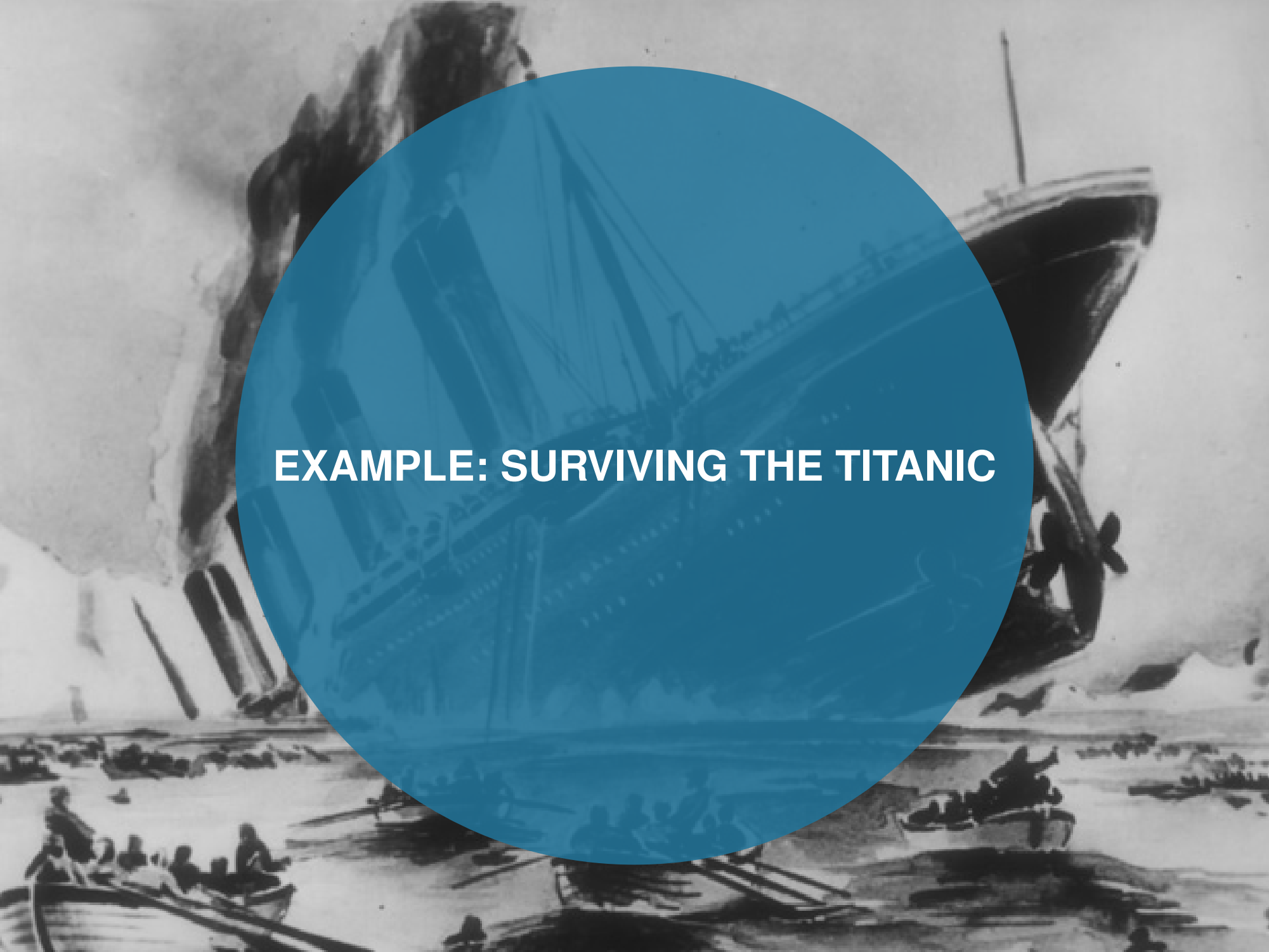
Types of machine learning algorithms:

- **Supervised:** algorithms are trained on labelled examples, i.e., input where the desired output is known. Examples: classification and regression problems.
- **Unsupervised:** algorithms operate on unlabelled examples, i.e., input where the desired output is unknown. Examples: clustering, association and feature extraction.
- **Hybrid algorithms:** Use some combination of both. Such algorithms include: Semi-supervised, transductive, reinforcement and self-taught learning.

Applications of machine learning:

- Facial recognition (Facebook)
- Email spam filters (Gmail)
- Recommender systems (Amazon/Netflix)
- Search engines (Google)
- credit card fraud detection (Visa, MasterCard, etc)
- Voice recognition systems (Google voice search, Siri)
- and many, many more . . .



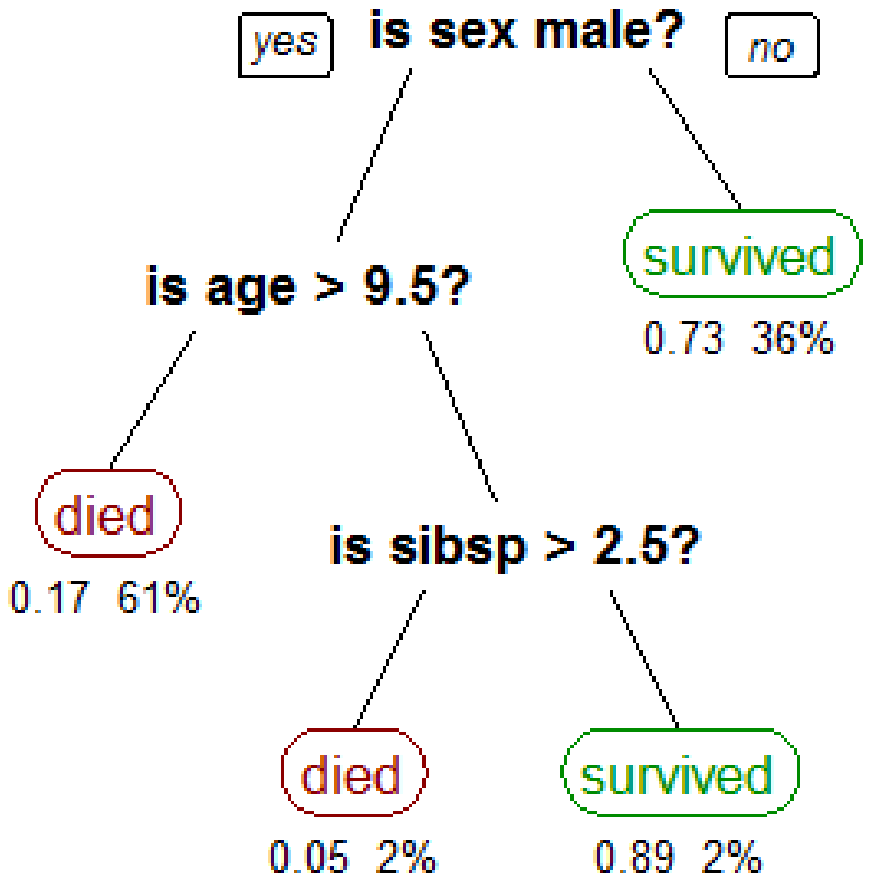


EXAMPLE: SURVIVING THE TITANIC

A simple example: classification with decision trees

Splits are chosen in such a way to minimize the overall entropy:

$$S = -\sum_i p(survival|split_i) \log p(survival|split_i)$$



A better algorithm: Random forests for classification

As the name suggests many different decision trees are used to vote on the class output (survived/died).

Algorithm:

For *tree* in *forest*

1. Sample, with replacement, a subset of training examples with class labels.
2. Sample, with replacement, a subset of the input variables.
3. Train *tree* with subset of training examples and input variables.

Take the mode (majority vote) of the *forest* to determine class.

Popular classification/regression algorithms:

- kth nearest neighbour (kNN)
- Support vector Machines (SVM)
- Linear discriminant analysis (LDA)
- Artificial neural networks (ANN)
- Bayesian networks
- Decision trees

Pretty much all classification/regression models derive from the above set.

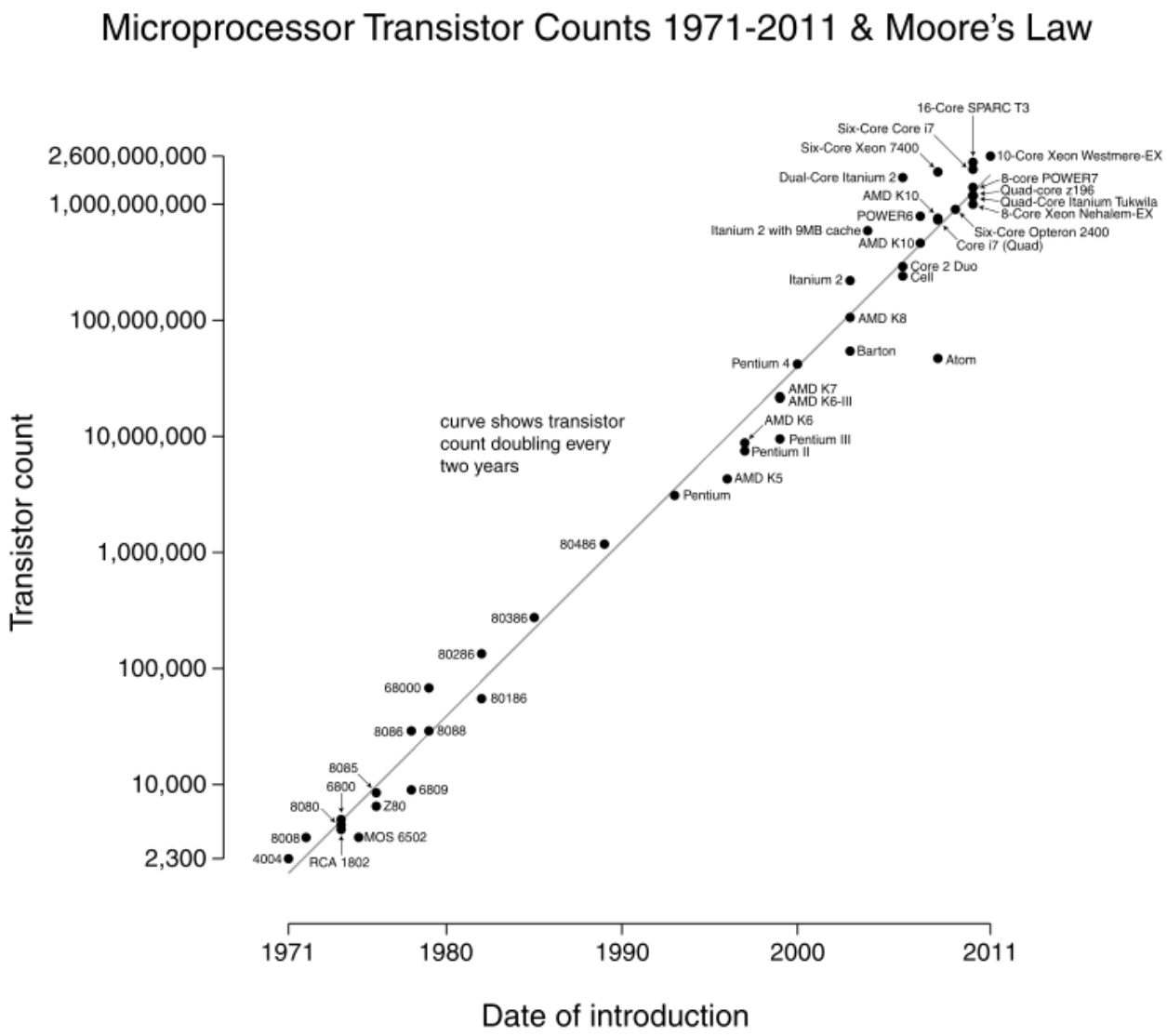

MOORE DATA

What's all the buzz about? There's a lot of excitement about machine learning right now. However, as a field of study it's been around since the 1960's! So what's driving this new found interest?

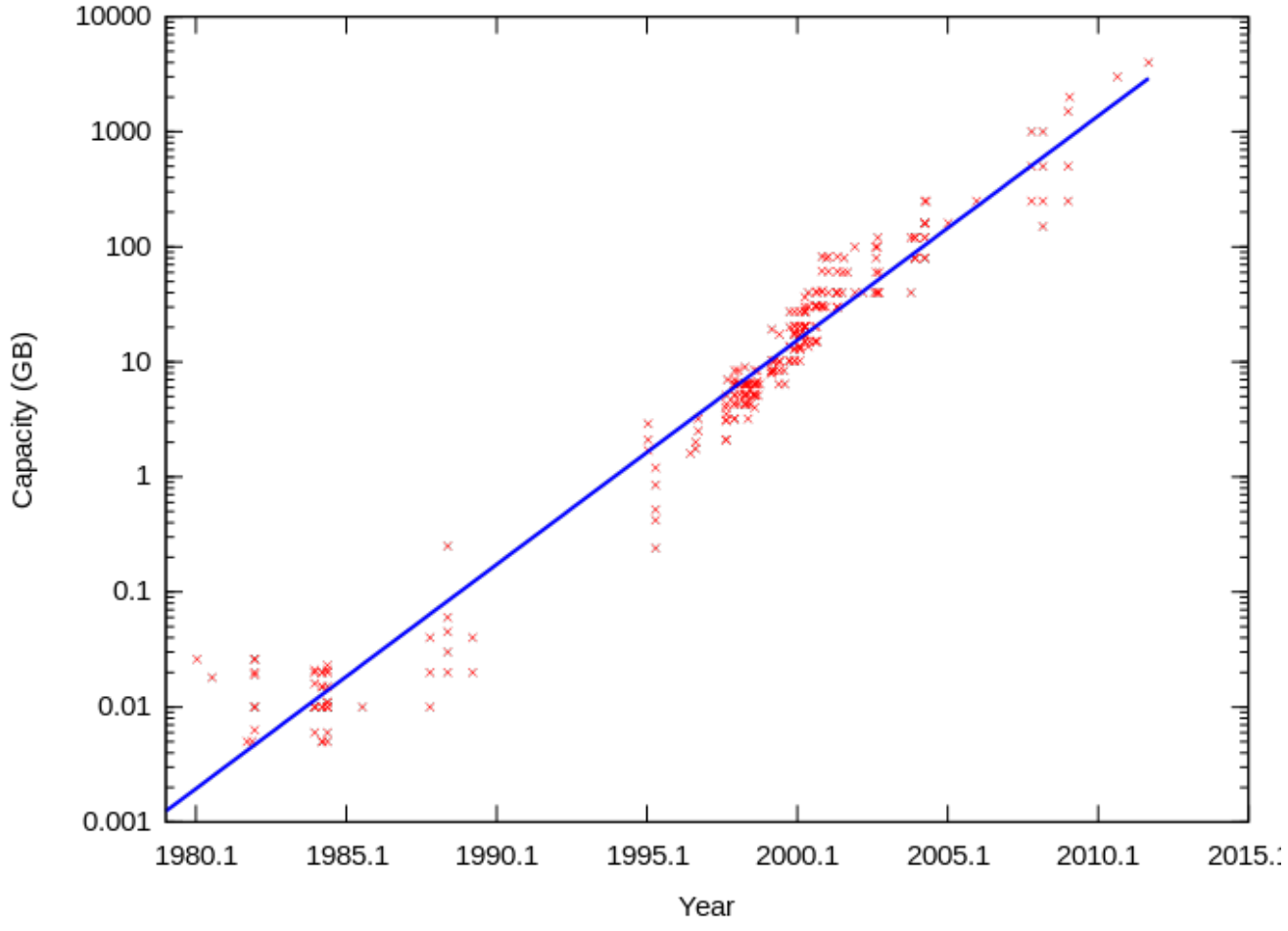
Answer:

- Moore's Law
- Big data
- Parallel architectures (GPUs)
- "Deep learning"

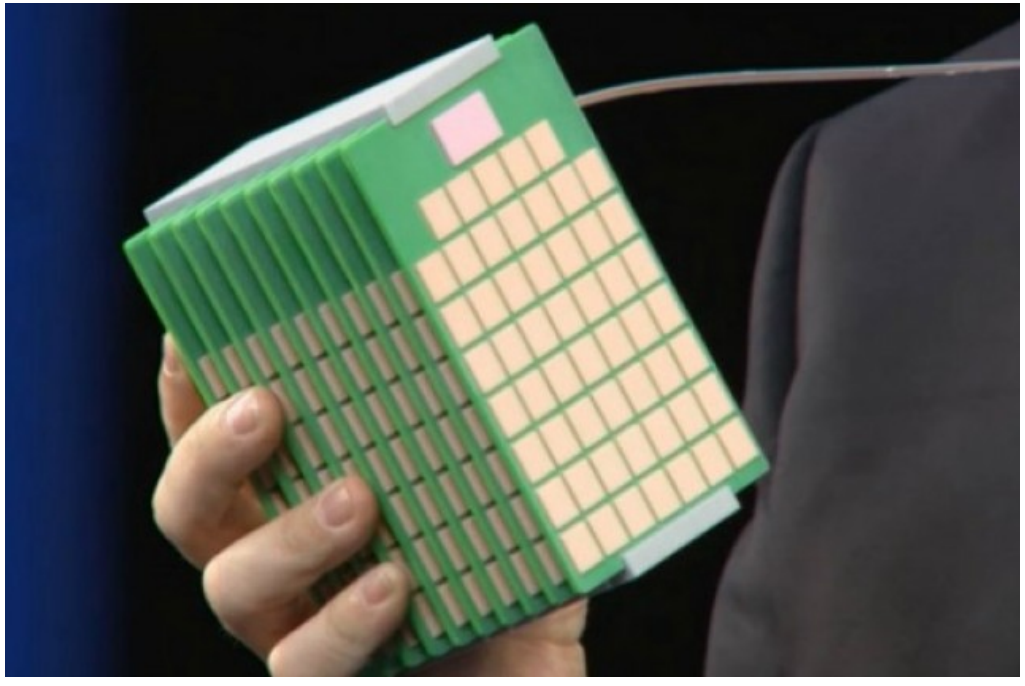
The number of transistors on a chip doubles every **18 months**



The amount of data collected doubles every **13 months!**

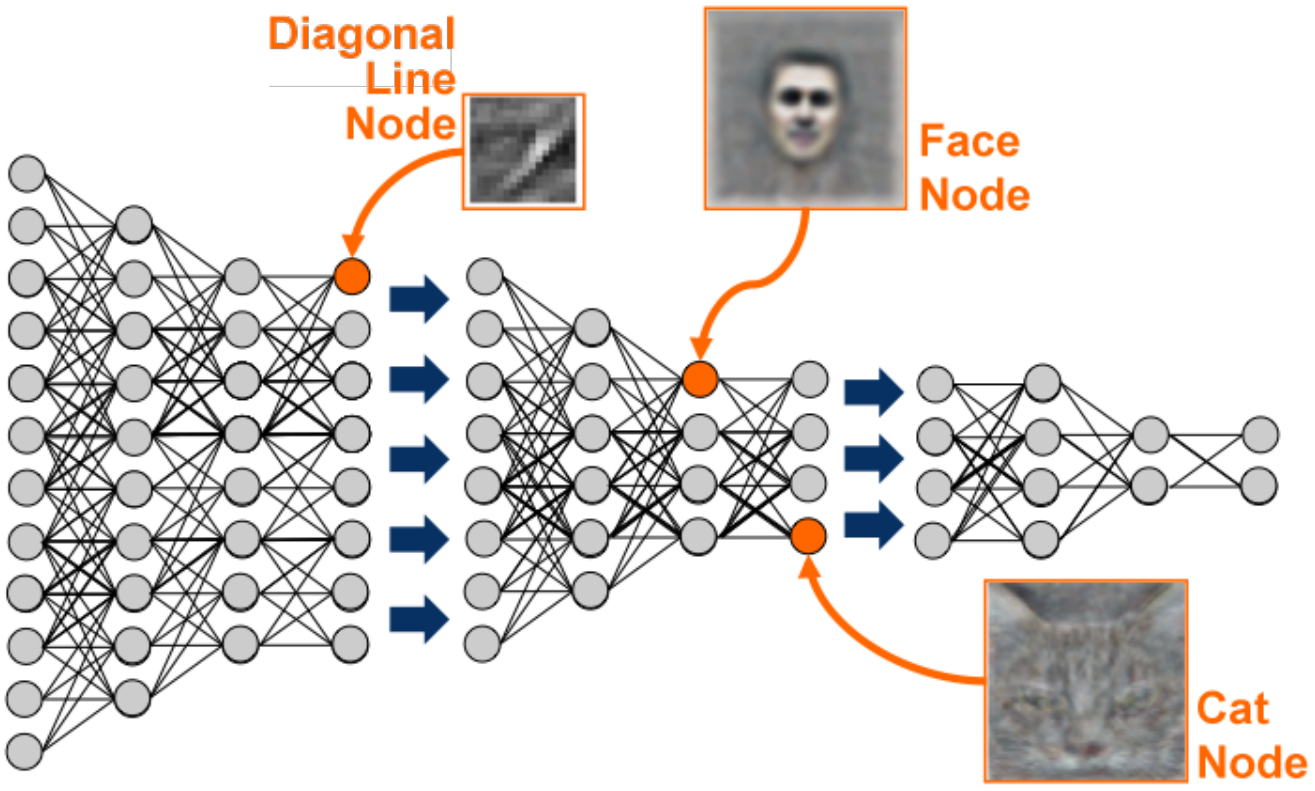


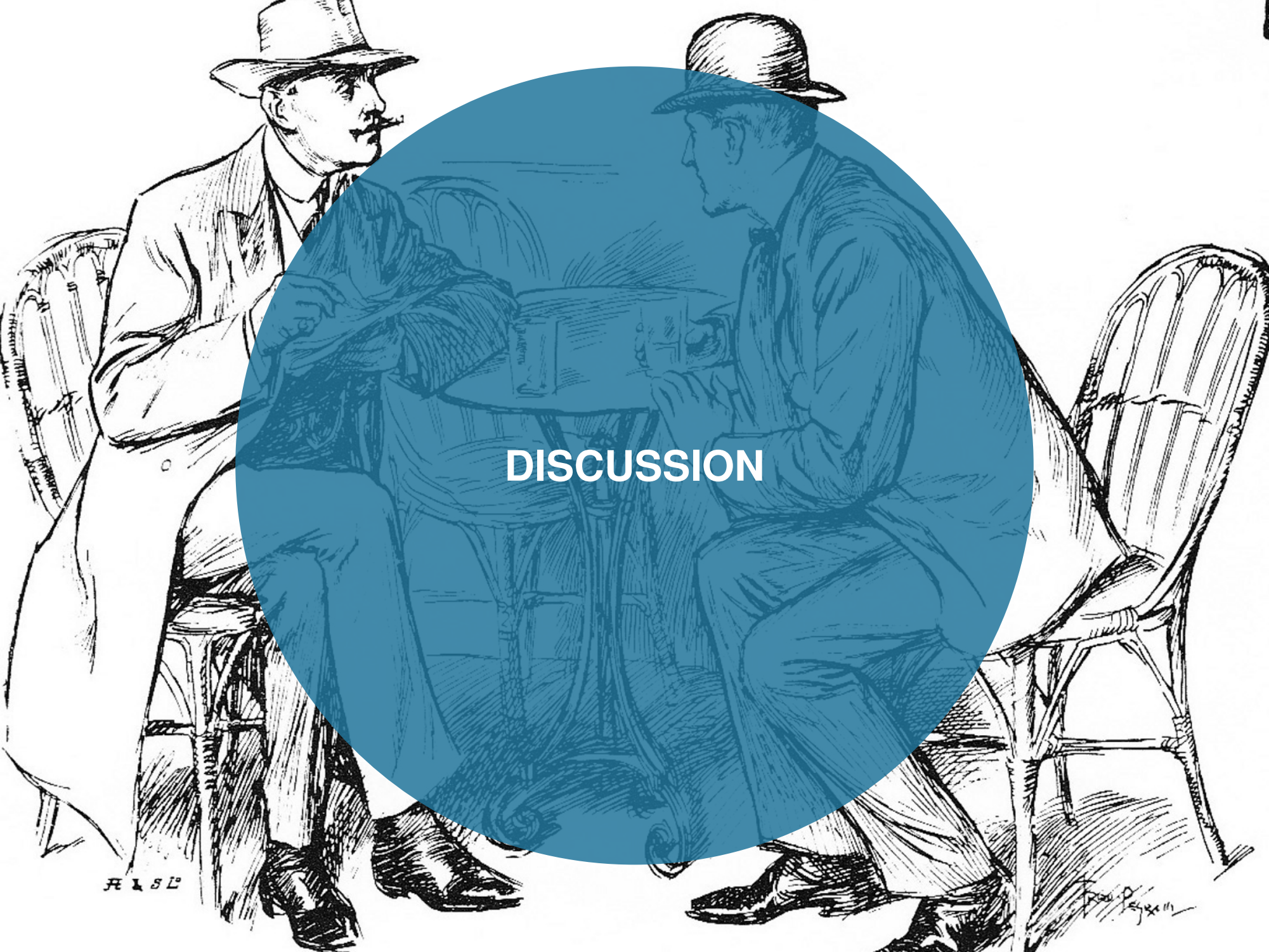
Neuromorphic hardware: Hardware architectures with a high level of parallelism can process data and train ML algorithms much more efficiently. Companies such as Qualcomm, IBM and HP are currently developing these next generation computing architectures



According to HP, The Machine can manage 160 PB of data in a mere 250 ns.

Deep learning: Deep neural networks are able to automatically extract features of data using a hierarchical series of representations, i.e. multiple levels of abstraction.




DISCUSSION