Lecture 1: Introduction to Biomedical Data Science

Instructor:

Parisa Rashidi

COURSE:
BIOMEDICAL DATA SCIENCE
FALL 2019

References

The Content, Graphics And Images In The Lecture Notes Are Partially Based On:

- Vijay Pande, Patrick Walters, Peter Eastman, Bharath Ramsundar. Deep Learning For The Life Sciences, 2019.
- David Sontag, Machine Learning For Healthcare 6.S897, Hst.S53, Mit, 2017
- Azizi, Palla, Belgrave, ICML Tutorial: Machine Learning For Personalised Health, 2018
- Yan Liu, Jimeng Sun, Deep Learning Models For Health Care - Challenges And Solutions, 2017
- O'neil, Cathy; Schutt, Rachel. Doing Data Science: Straight Talk From The Frontline, 2016
- Shortliffe, Edward H.; Cimino, James J. (2013). Biomedical Informatics. Springer London.
- Mark Musen, Introduction To Big Data And The Data Lifecycle, The Big Data To Knowledge (Bd2k), 2017
- Guide To The Fundamentals Of Data Science Computing Overview, Patricia Kovatch, 2017

Agenda

- Paper Presentation Logistics (First Paper: Next Thursday)
- Introduction to biomedical data science
- Introduction to Python Programming

Papers

- First paper is available on Canvas.
- Each time one graduate student <u>group</u> will be presenting.
  - For now, groups of 4 students.
  - Upload your presentation before noon on Thursday.
- 15 minutes presentation (background, methods, results, discussions):
  - Incorporate your own discussions, insights, criticism.
  - 2 minutes for Q/A.
- Followed by discussions, both graduate and undergraduate students.

Volume 29, Issue 2
February 2018

Article Contents

EDITOR'S CHOICE

Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board FREE

S P Somashekhar, M -J Sepúlveda, S Puglielli, A D Norden, E H Shortliffe, C Rohit Kumar, A Rauthan, N Arun Kumar, P Patil, K Rhee, ... Show more

Annals of Oncology, Volume 29, Issue 2, 1 February 2018, Pages 418–423, https://doi.org/10.1093/annonc/mdx781
Published: 09 January 2018

How to criticize scientific articles?

- Read this short guide: Link

- Please do not be afraid to criticize papers! That is part of our goal, to teach you critical thinking.
- Take a look at the rubrics on Canvas

Plan Your Presentations!

- Plan your presentations
  - Introduction
    - Problem (brief)
    - Importance (give background)
  - Problem (detail)
  - Solution (detail)
  - Your criticism (very important!)
    - Your Suggestions
  - Possible future directions
  - Conclusion
  - QA

Inverted Pyramid

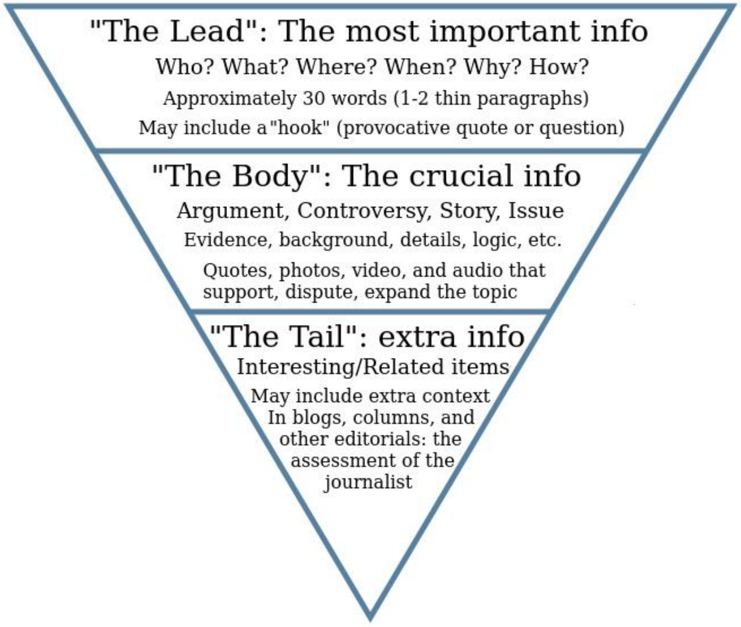


*Image from Wikipedia

Some Presentation Tips

- Timing and pace is key.
  - Almost ~1 minute per slide
  - Finish on time
- Be clear and concise
  - Avoid self-talking!
- Make a point
  - Why we need to know it, why we would care about it...
- Engage your audience with illustrations
  - Light on text and heavy on figures

More Tips here

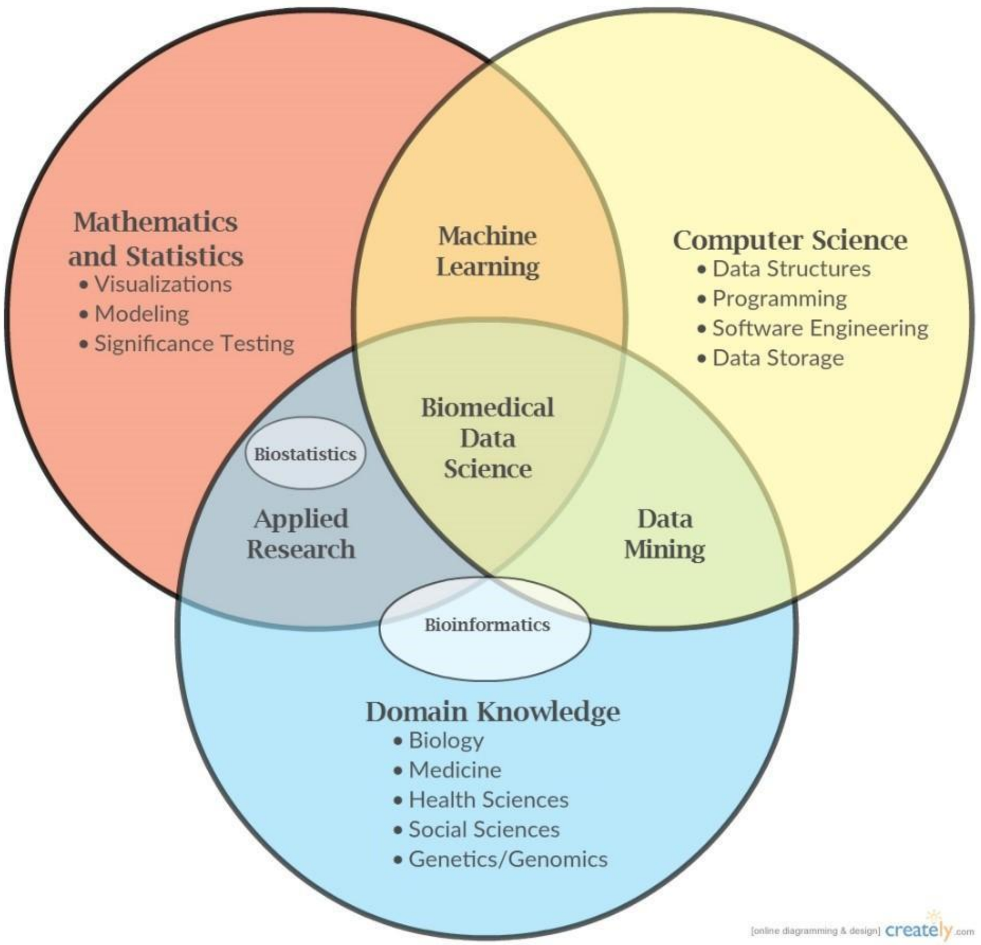
Some More Presentation Tips

- Giving credit to others
  - Figures, citations, …
- Talk, don’t read!
  - Use notes only sporadically
  - Non-verbal communication
- Rehearse
  - Don’t memorize! (plan how to present complex ideas)

HISTORICAL BACKGROUND

Biomedical Data Science

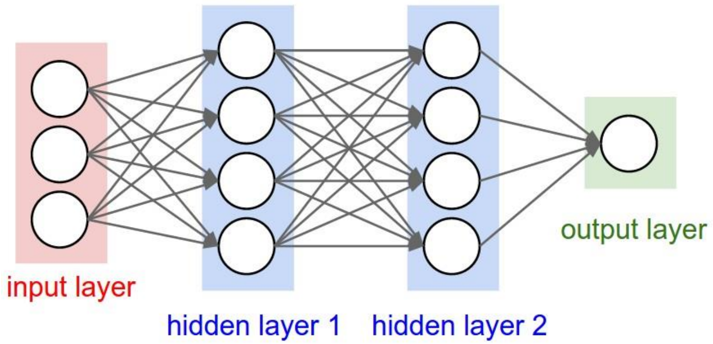
- Data science techniques applied to biomedical science problems



Question

Take a guess: the first artificial neural network models were developed in:

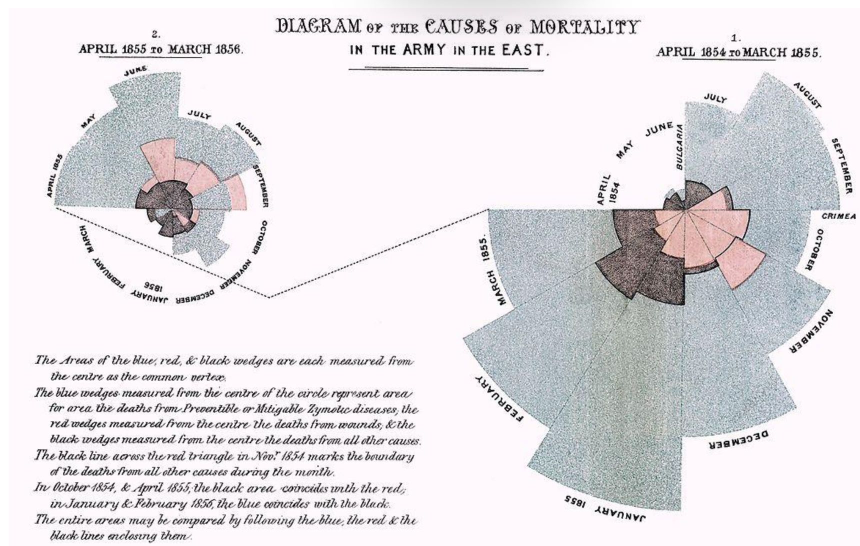
- ❑ 1950s
- ❑ 1980s
- ❑ 1990s
- ❑ 2000s



Dawn of Data-Driven Health:1850s

- Studying the causes of mortality in the army
- 16,000 to 18,000 army death due to preventable conditions

Florence Nightingale

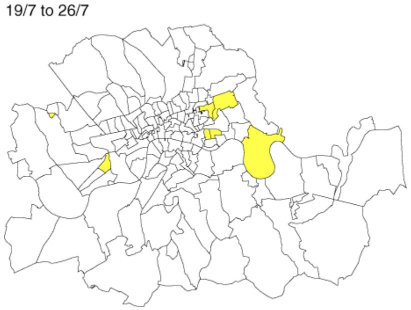


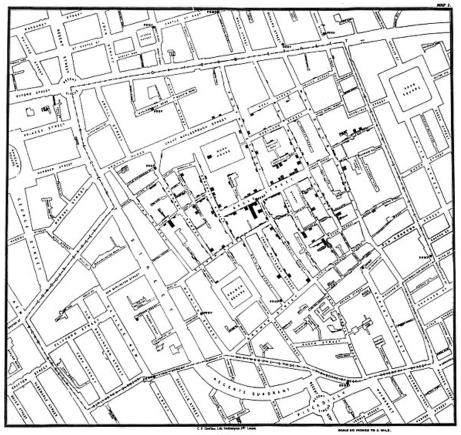
Polar area diagram (circular histogram)

Dawn of Data-Driven Health:1850s

- Tracing the outbreak of Cholera in London in 1854
- Father of modern epidemiology

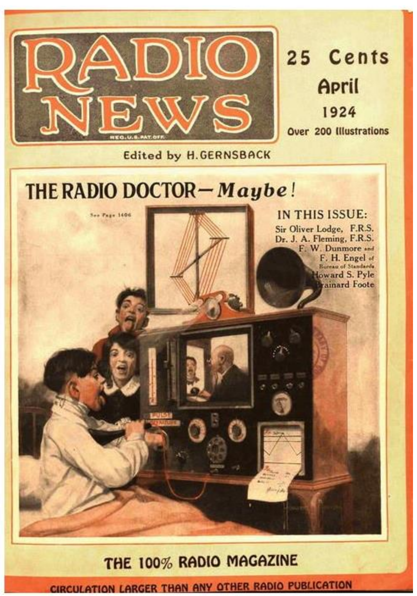
John Snow

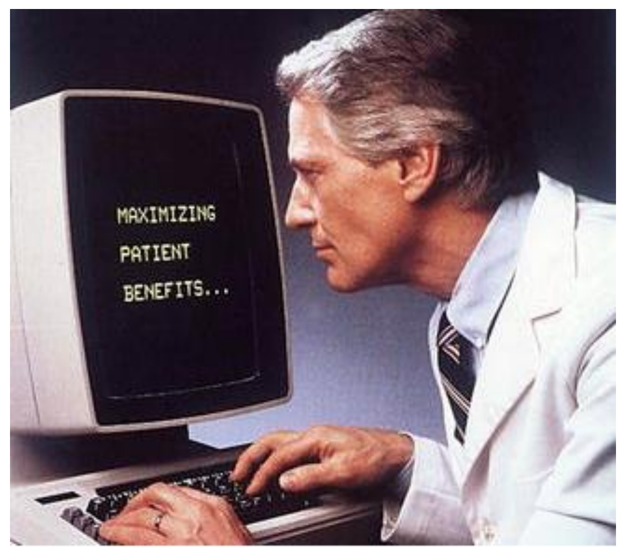




Early Clinical Information Systems: 1960s

- Hospital information systems or HIS (1960s)
  - Mostly a single, large, time-shared computer
  - Distributed HIS started to appear in 1980s





Doctor of the future (Early 1980s).

Machine Learning History 1

- 1950 — Alan Turing creates the “Turing Test”
- 1952 — Arthur Samuel wrote the first computer learning program to play checkers.
- 1957 — Frank Rosenblatt designed the first neural network for computers (the perceptron)
- 1967 — The “nearest neighbor” algorithm was written

Expert Systems: 1970

- MYCIN Expert System developed at Stanford in 1972
  - For identifying blood infections based on reported symptoms and medical test results
  - Using about 500 production rules
  - Roughly the same level of competence as blood infection specialists and rather better than general practitioners
  - Brittle, lots of manual entry, not scalable

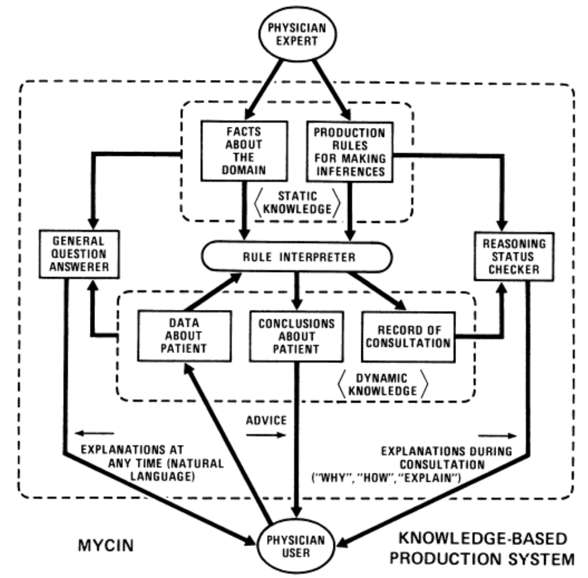
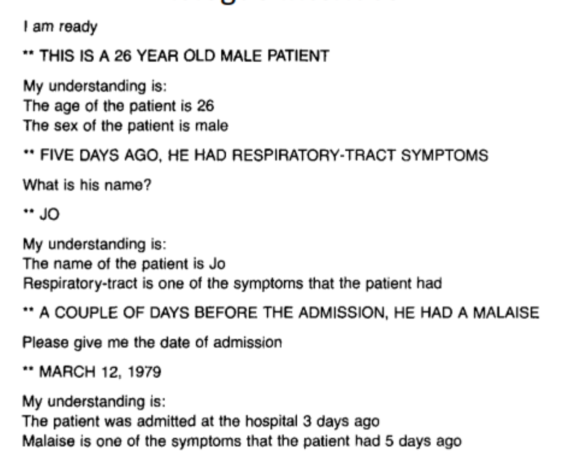


Figure 2 - Diagram summarizing the organization and flow of information within MYCIN. The correlation between this design and the human consultation process depicted in Fig. 1 is discussed in the text. (Figure reproduced from reference 10).

Dialogue interface


I am ready

** THIS IS A 26 YEAR OLD MALE PATIENT

My understanding is:
The age of the patient is 26
The sex of the patient is male

** FIVE DAYS AGO, HE HAD RESPIRATORY-TRACT SYMPTOMS

What is his name?

** JO

My understanding is:
The name of the patient is Jo
Respiratory-tract is one of the symptoms that the patient had

** A COUPLE OF DAYS BEFORE THE ADMISSION, HE HAD A MALAISE

Please give me the date of admission

** MARCH 12, 1979

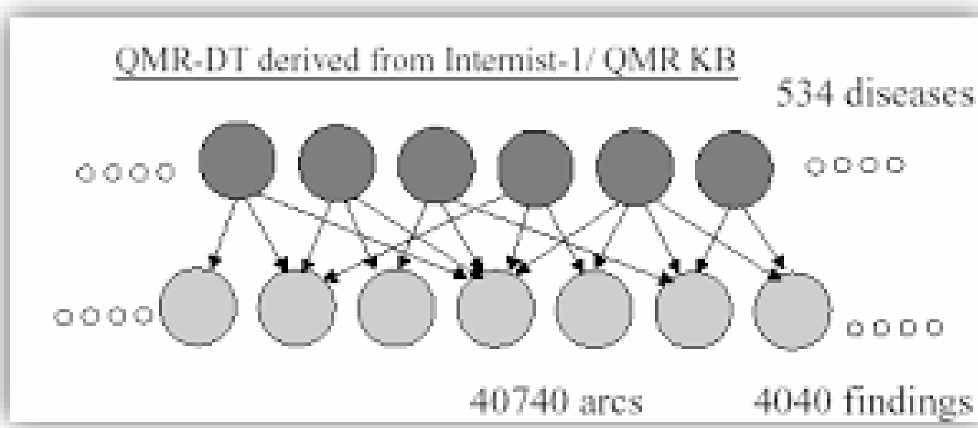
My understanding is:
The patient was admitted at the hospital 3 days ago
Malaise is one of the symptoms that the patient had 5 days ago

FIGURE 33-1 Short sample dialogue. The physician's inputs appear in capital letters after the double asterisks.

Shortliffe EH. Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. Proceedings of the Annual Symposium on Computer Application in Medical Care. October 1977:66-69.

Probabilistic Models: 1980

- INTERNIST/QMR was developed at University of Pittsburgh, 15 person-years of work
- A broad-based computer-assisted diagnostic tool
- Probabilistic model with 534 binary disease variables 4,040 binary symptom variables, 45,470 edges

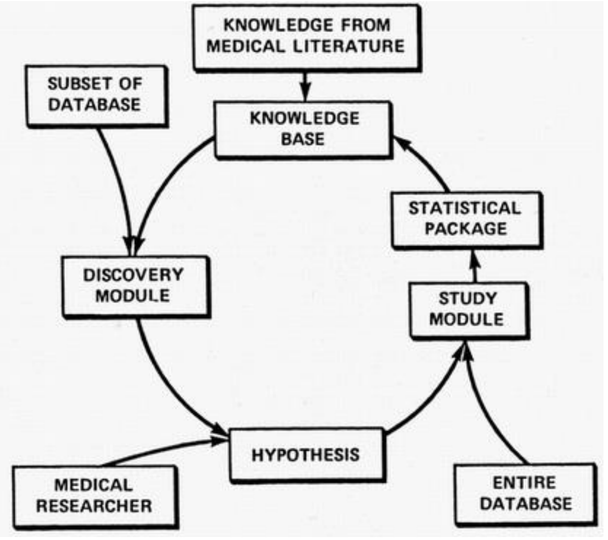


Issues
- Manual Symptom entry by physicians
- Difficulty in maintenance and generalization

Data Mining: 1980

- The RX Project: discovering medical facts
- An early example of data mining under AI control
- Data from 50 severe Lupus patients
  - 52 attributes

Blum, R. L. (1982). Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical data base: the RX project. Computers and Biomedical Research, 15(2), 164-187.



Neural Networks: 1990

- Neural networks in clinical applications started to appear in 1990

Small networks, poor generalization,

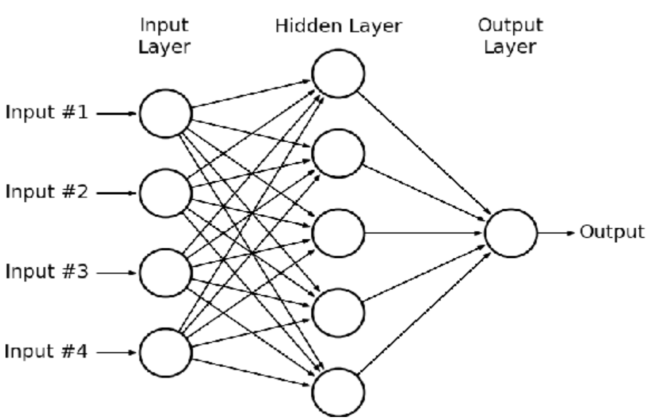


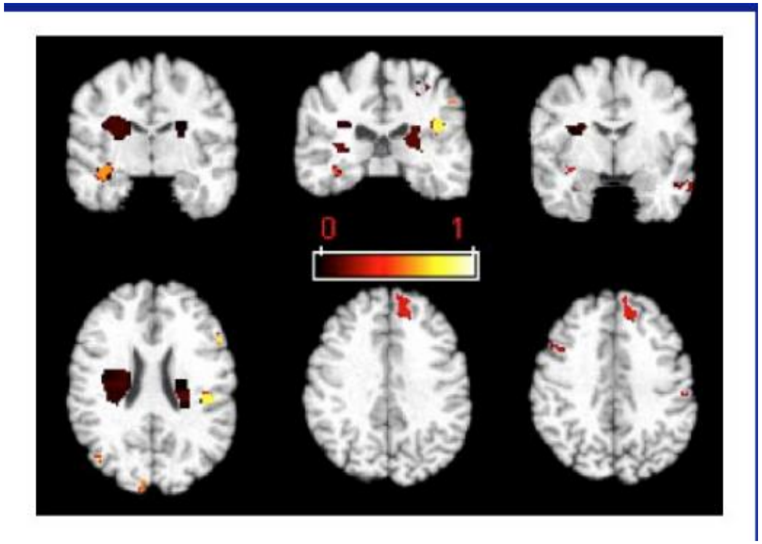
Table 1 • 25 Neural Network Studies in Medical Decision Making*

Subject	No. of Examples: Training	No. of Examples: Test	P†	Network	D‡	Accuracy§: Neural	Accuracy§: Other
Breast cancer[4]	57	20	60	9–15–2	0.6	80	75
Vasculitis[2]	404	403	73	8–5–1	8.0	94	—
Myocardial infarction[6]	351	331	89	20–10–10–1	1.1	97	**84**
Myocardial infarction[8]	356	350	87	20–10–10–1	1.1	97	**94**
Low back pain[11]	100	100	25	50–48–2	0.2	90	90
Cancer outcome[13]	5,169	3,102	—	54–40–1	1.4	0.779	0.776
Psychiatric length of stay[17]	957	106	73	48–400–4	0.2	74	**76**
Intensive care outcome[23]	284	138	91	27–18–1	0.5	0.82	**0.82**
Skin tumor[21]	150	100	80	18	—	80	**90**
Evoked potentials[35]	100	67	52	14–4–3	3.8	77	**77**
Head injury[47]	500	500	50	6–3–3	20	66	**77**
Psychiatric outcome[54]	289	**92**	**60**	41–10–1	0.7	79	—
Tumor classification[55]	53	**6**	**38**	8–9–3	1.4	99	**88**
Dementia[57]	75	18	19	80–10–7–7	0.6	61	—
Pulmonary embolism[59]	607	606	69	50–4–1	2.9	0.82	**0.83**
Heart disease[62]	460	230	54	35–16–8–2	3	**83**	**84**
Thyroid function[62]	3,600	1,800	93	21–16–8–3	**22**	**98**	**93**
Breast cancer[62]	350	175	66	9–4–4–2	**10**	**97**	**96**
Diabetes[62]	384	192	65	8–4–4–2	**12**	**77**	**75**
Mycardial infarction[63]	2,856	1,429	56	291–1	9.8	85	—
Hepatitis[65]	39	42	38	4–4–3	3.3	74	**79**
Psychiatric admission[76]	319	339	85	53–1–1	6.0	91	—
Cardiac length of stay[83]	713	696	73	15–12–1	3.5	0.70	—
Anti-cancer agents[89]	127	141	25	60–7–6	1.5	91	**86**
Ovarian cancer[91]	75	98	—	6–6–2	2.6	**84**	**81**
MEDIAN VALUE	350	175	71	20	2.8		

Penny, Will, and David Frost. "Neural networks in clinical medicine." Medical Decision Making 16.4 (1996): 386-398.

Support Vector Machines: 2000

- Support vector machines became very popular, especially in neuroimaging.



(Fan et al, Lect. Notes in CompSci, 2005)

Machine Learning History 2

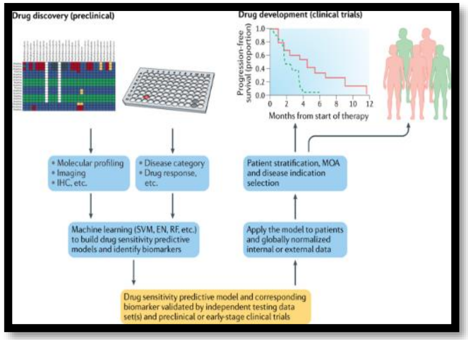
- 1986 — back-propagation by Rumelhart
- 1992 — SVMs close to their current form introduced by Vapnik
- 1997 — LSTM introduced.
- 1997 — IBM's Deep Blue beats the world champion at chess.
- 2006 — Geoffrey Hinton coins the term "deep learning"

Bernard Marr, A Short History of Machine Learning -- Every Manager Should Read

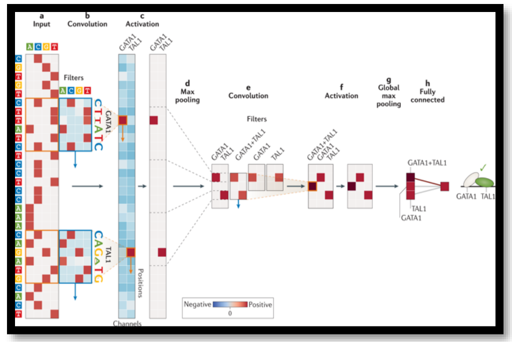
ML History 3

- 2011 — IBM’s Watson beats its human competitors at Jeopardy.
- 2014 – Facebook develops DeepFace
- 2016 – Google’s algorithm beats a professional player at the board game Go
- 2018 - 2019 moving beyond ImageNet

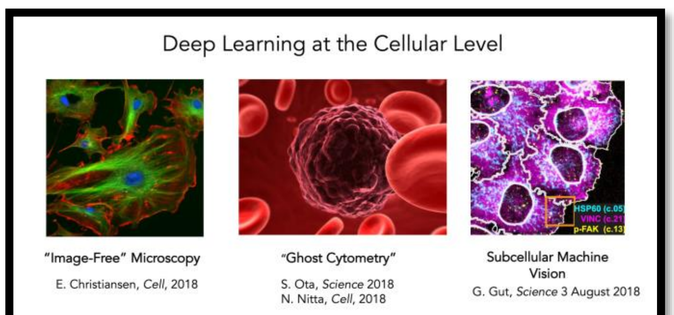
Deep Networks: 2019

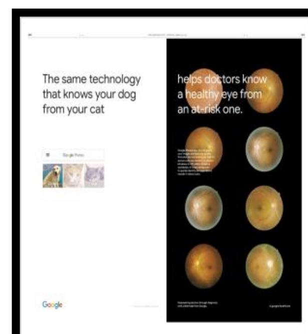


ML FOR DRUG DISCOVERY
(NATURE REVIEWS-DRUG DISCOVERY, 2019)

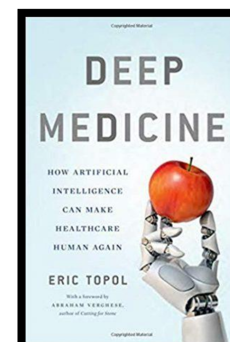


MODELLING TRANSCRIPTION FACTOR BINDING SITES
(NATURE REVIEW -GENTICS, 2019)





AI.GOOGLE/HEALTHCARE



What makes healthcare different?

- Life or death decisions
  - Need **robust** algorithms
  - Checks and balances built into ML deployment
  - (Also arises in other applications of AI such as autonomous driving)
  - Need **fair** and **accountable** algorithms
- Many questions are about unsupervised learning
  - Discovering disease subtypes, or answering question such as "characterize the types of people that are highly likely to be readmitted to the hospital"?
- Many of the questions we want to answer are *causal*
  - Naïve use of supervised machine learning is insufficient

What makes healthcare different?

- Often very little labeled data (e.g., for clinical NLP)
  - Motivates semi-supervised learning algorithms
- Sometimes small numbers of samples (e.g., a rare disease)
  - Learn as much as possible from other data (e.g. healthy patients)
  - Model the problem carefully
- Lots of missing data, varying time intervals, censored labels

What makes healthcare different?

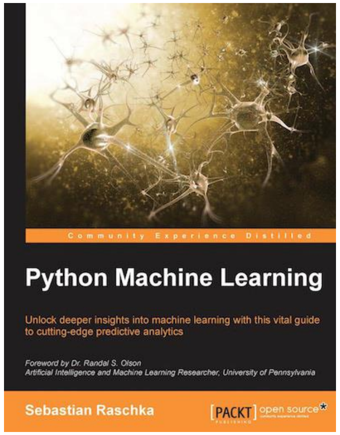
- Difficulty of de-identifying data
  - Need for data sharing agreements and sensitivity
- Difficulty of deploying ML
  - Commercial electronic health record software is difficult to modify
  - Data is often in silos; everyone recognizes need for interoperability, but slow progress
  - Careful testing and iteration is needed

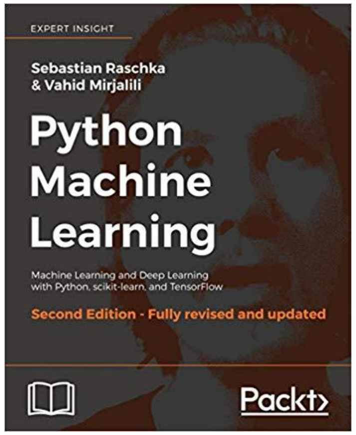
Institutional Review Board (IRB)

- Raise your hand if you know about IRB!
- A committee that reviews research studies to ensure they are complying with ethical guidelines.
- Link to [UF IRB]

Disclaimer

The following slides are partially based on:



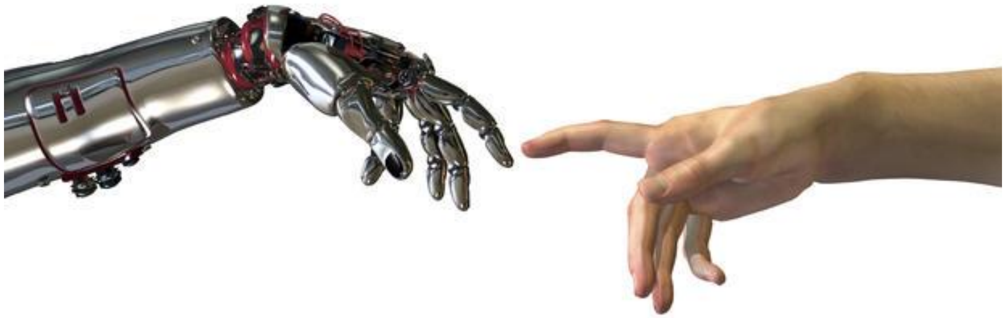


Agenda

- Machine learning introduction
- Simple classification models
  - KNN, decision trees
- More advanced models
  - XGBoost
  - SVM
- Deep learning

Artificial Intelligence

- Artificial Intelligence (AI) has many subfields
  - Machine Learning (ML)
  - Natural Language Processing (NLP)
  - Vision
  - Robotics
  - …



What is “Learning” ?

- Machine learning is programming computers to optimize a performance criterion in a certain task using example data (i.e. past experience).

- Example task: predicting if there will be any complication 30 days after surgery
  - Performance Criteria: Number of cases correctly predicted
  - Example data: patients’ medical history + outcome after 30 days

Capturing Informal Knowledge: Early Days

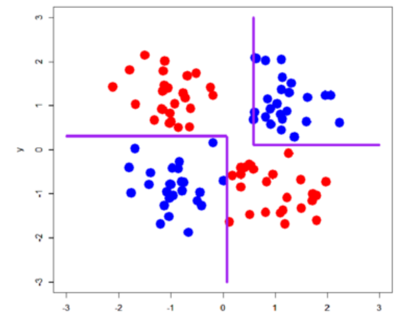
- We need to get informal knowledge to computers
  - Several systems tried to hard-code this knowledge
    - Knowledge-base approach
    - Example: Cyc, the world's longest-lived AI project (1984)
      - A knowledgebase of the basic concepts and "rules of thumb" about how the world works

Capturing Informal Knowledge: Modern Approach

- Machine learning
  - Instead of dictating rules, let's provide data to the machine and let it learn from data.
  - Even simple algorithms might work: deciding if C-section is needed using logistic regression (Mor-Yosef et al., 1990)

Key Terms

- Instance = example = data point
- Feature = independent variable
- Class label = dependent variable
- Decision boundary = separates examples in different classes

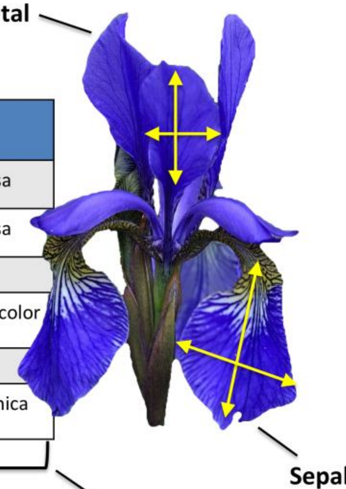


Samples
(instances, observations)

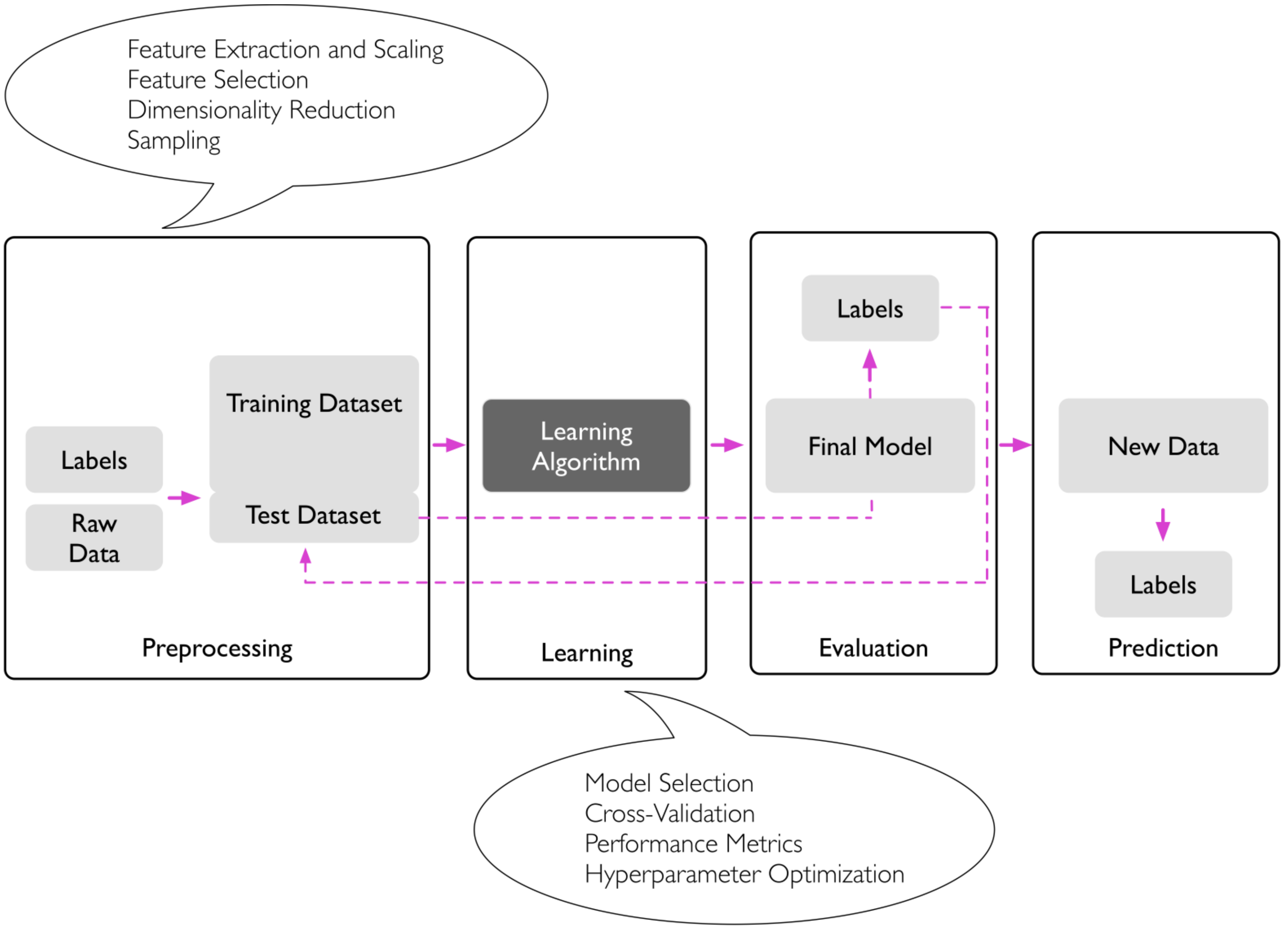
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



Roadmap



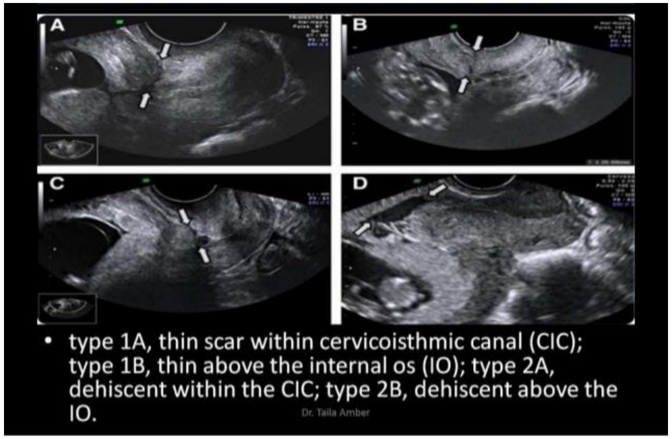
Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Preprocessing
Learning Algorithm
Learning
Labels
Final Model
Evaluation
New Data
Labels
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Representation

- Each piece of information included in the representation of the patient is known as a feature.
- The algorithm will learn how the features are correlated with the outcome.

	age	Previous pregnancies	Scar	C-section
P1	21	0	0	0
P2	39	2	1	1
P3	36	1	1	?

or



Tabular Representation

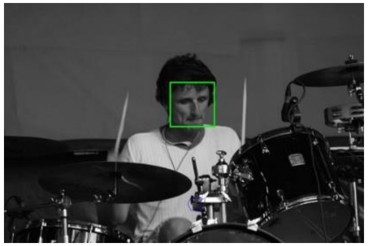
- The most common type
  - Simple records in Tables
  - Can be analyzed using regular machine learning techniques.
  - Most other data types are converted to this type (not always, we will later talk about deep learning).

ID	WGT	HGT	Cholesterol	**Risk (Class)**
1	high	short	260	high
2	high	med	254	high
3	high	tall	142	med

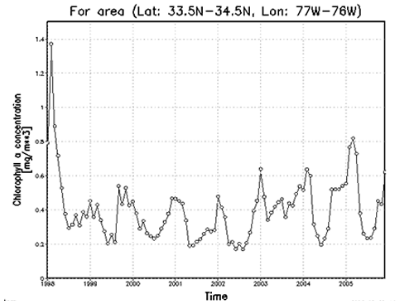
A Simple Table

Other Input Representations

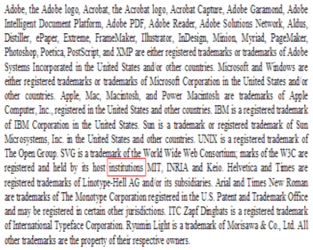
- Image, video
  - is preprocessed using Vision techniques or
  - using deep learning techniques such as deep convolutional neural networks (CNN)
- Text
  - is preprocessed using NLP techniques or
  - using deep learning techniques such as Long Short Term Memory Networks (LSTM)
- Continuous measures along time (Time series)
  - is preprocessed using Time Series analysis or
  - using deep learning techniques such as LSTMs



Image

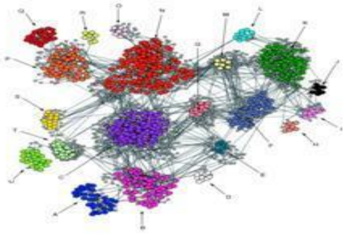


Time series



Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Aldus, Distiller, ePaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Microsoft and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Apple, Mac, Macintosh, and Power Macintosh are trademarks of Apple Computer, Inc., registered in the United States and other countries. IBM is a registered trademark of IBM Corporation in the United States. Sun is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries. UNIX is a registered trademark of The Open Group. SVG is a trademark of the World Wide Web Consortium; marks of the W3C are registered and held by its host $\boxed{\text{institutions}}$ MIT, INRIA and Keio. Helvetica and Times are registered trademarks of Linotype-Hell AG and/or its subsidiaries. Arial and Times New Roman are trademarks of The Monotype Corporation registered in the U.S. Patent and Trademark Office and may be registered in certain other jurisdictions. ITC Zapf Dingbats is a registered trademark of International Typeface Corporation. Ryumin Light is a trademark of Morisawa & Co., Ltd. All other trademarks are the property of their respective owners.

Text



Graph

Data Representation

- We usually represent data in a matrix

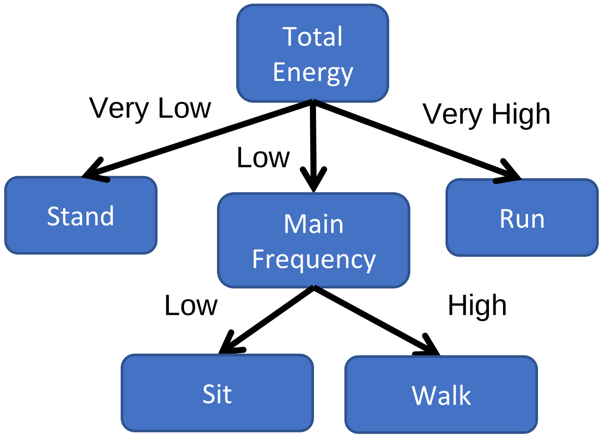
$$X = \begin{array}{cc} & \text{Features} \\ \text{Instances} & \begin{bmatrix} 2 & 5 & 1 & 1 & 1 & 2 & 1 & 3 \\ 2 & 5 & 4 & 4 & 5 & 7 & 10 & 3 \\ 3 & 2 & 1 & 1 & 1 & 2 & 5 & 4 \end{bmatrix} \end{array}$$

$$Y = \begin{array}{cc} & \text{Label} \\ \text{Instances} & \begin{bmatrix} -1 \\ +1 \\ ? \end{bmatrix} \end{array}$$

Note: We can also assign a probability to each label (we'll discuss it later)

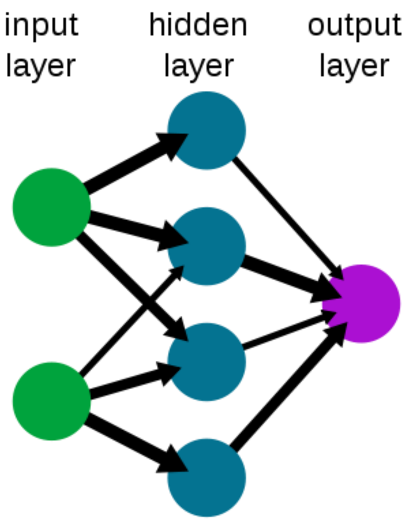
Example ML Algorithms

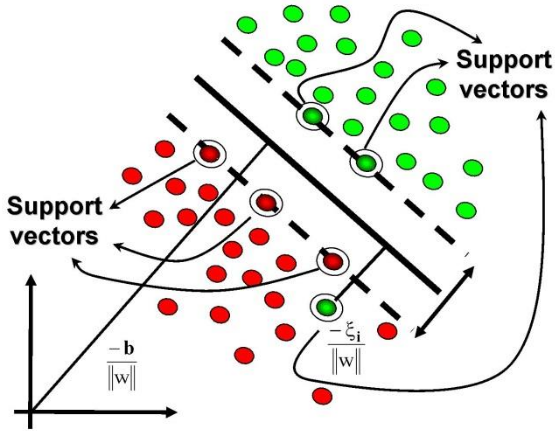
- Linear Regression
- Decision trees, neural network, support vector machine, …



A simple decision tree

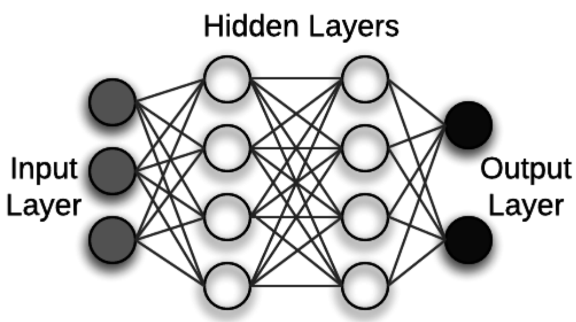
A simple neural network





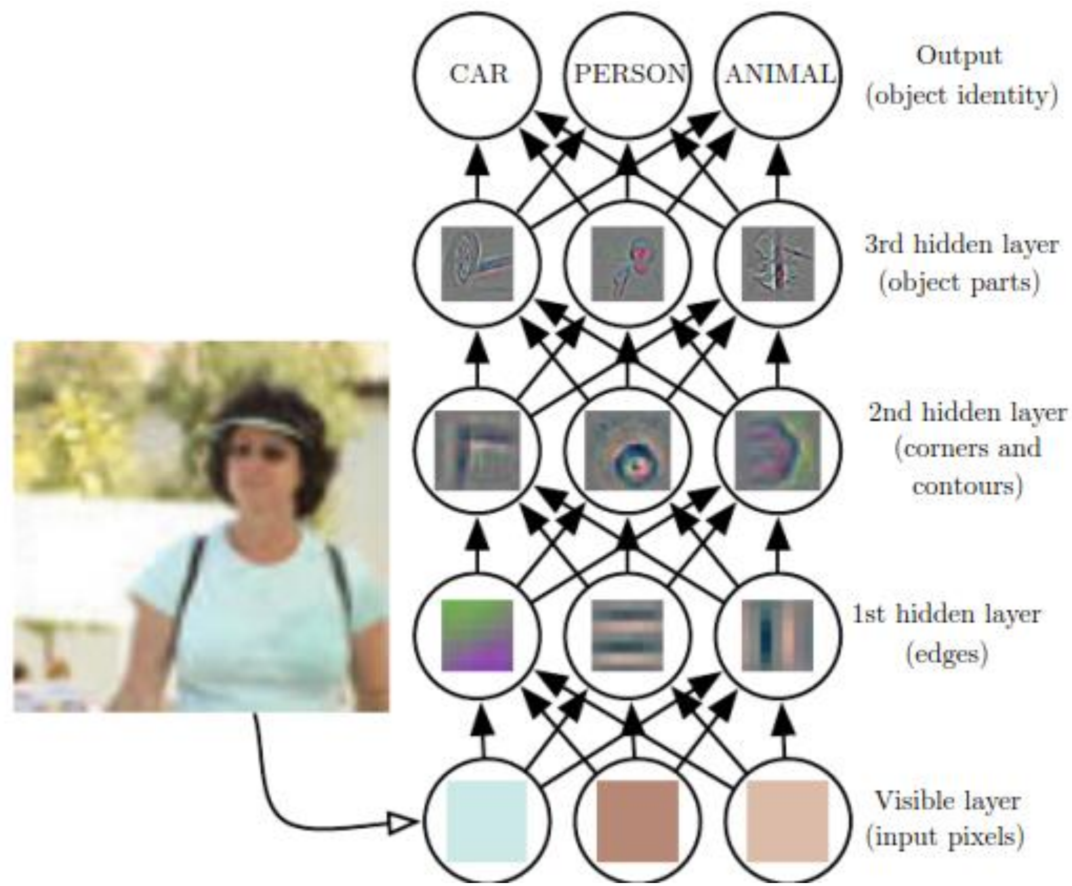
Support Vector Machines

Deep Learning



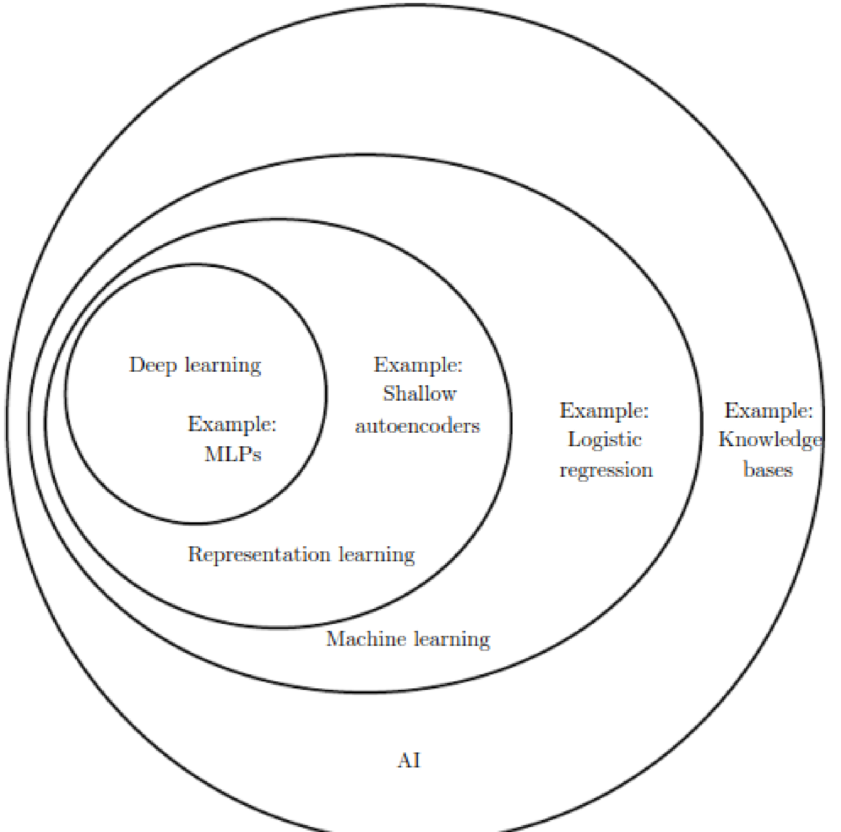
A simple Neural network

Deep Neural network



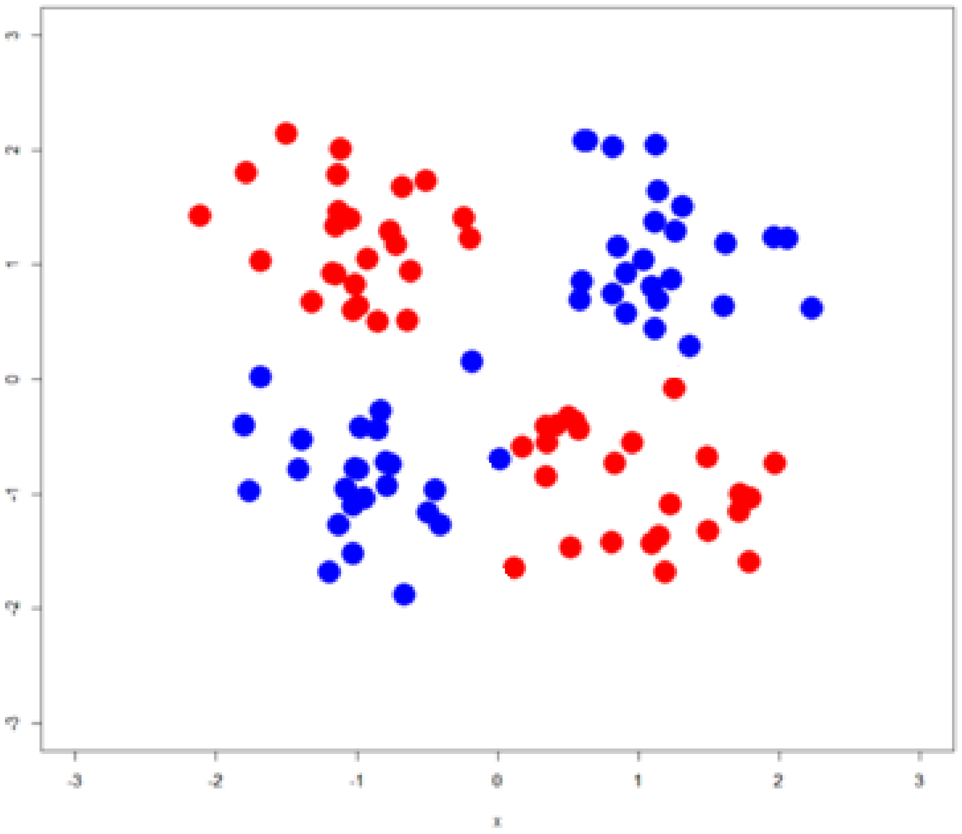
Relation with Sub-areas

- Deep learning is not equal to machine learning

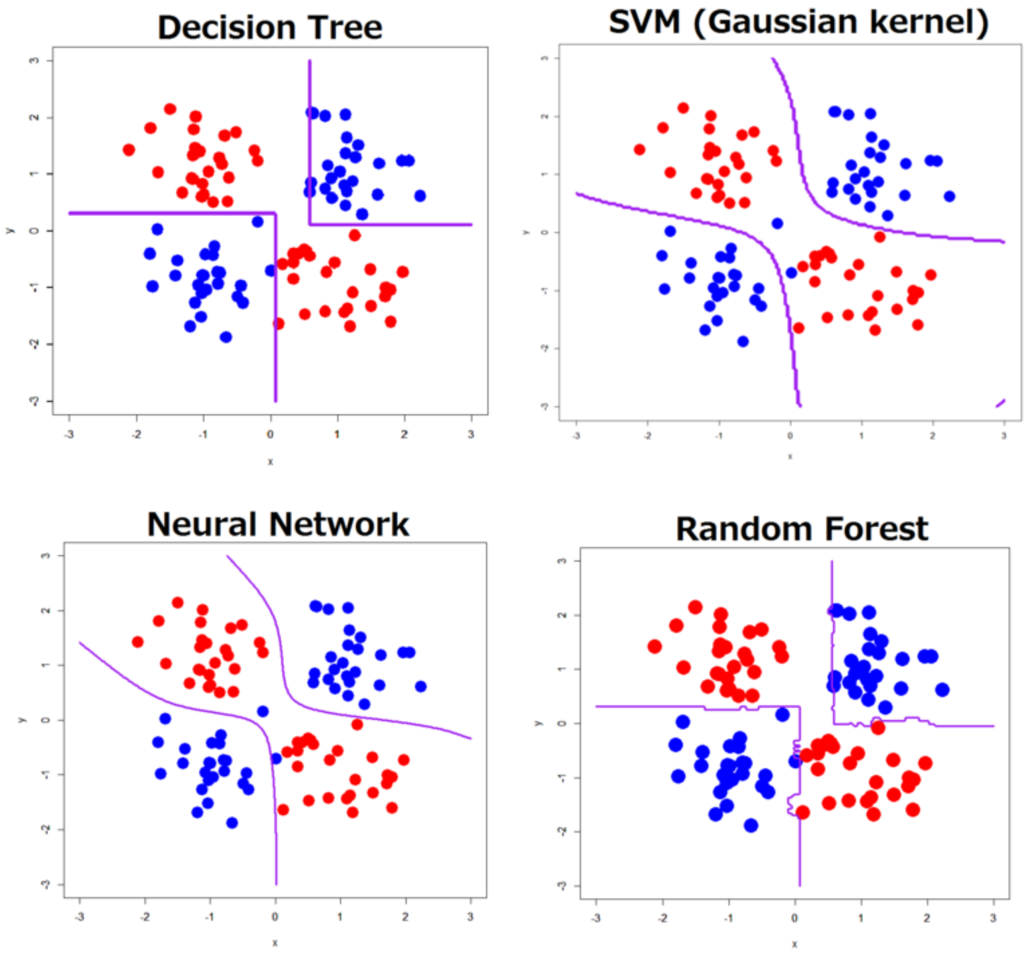


Decision Boundary

- Example dataset



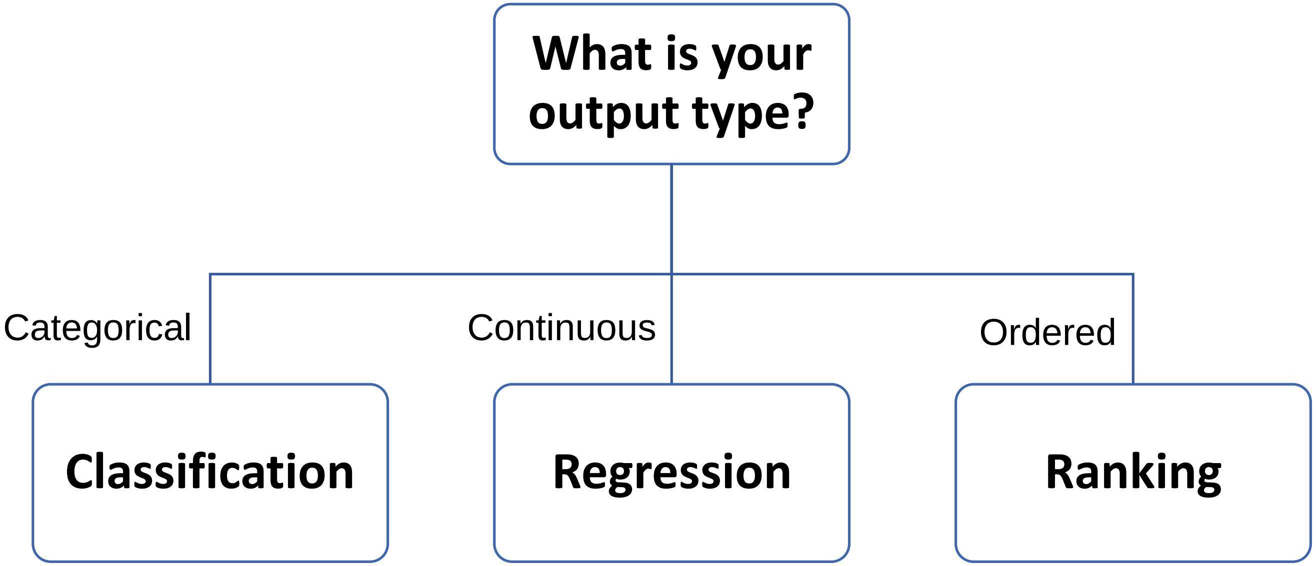
Example Decision Boundaries



http://www.kdnuggets.com/2015/06/decision-boundaries-deep-learning-machine-learning-classifiers.html

Task Type

- Categorical: Classification task
  - Classifier
- Continuous: Regression task
- Ordered: Ranking task



Supervised vs. Unsupervised Learning

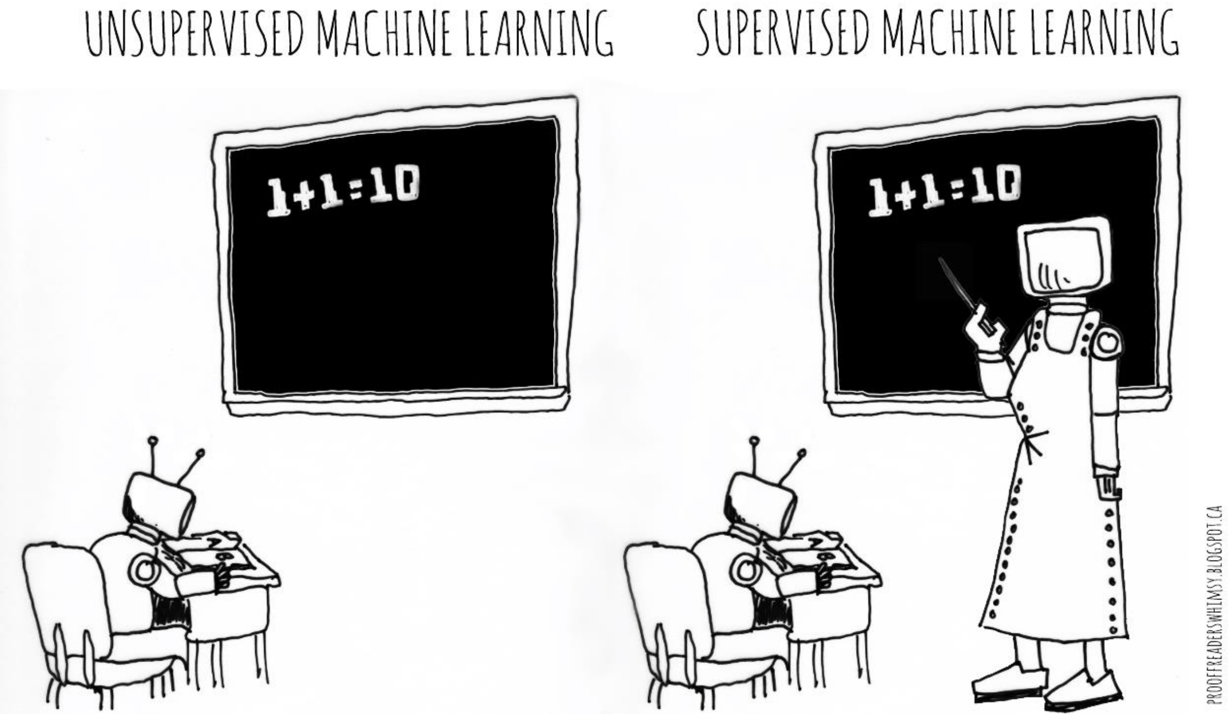
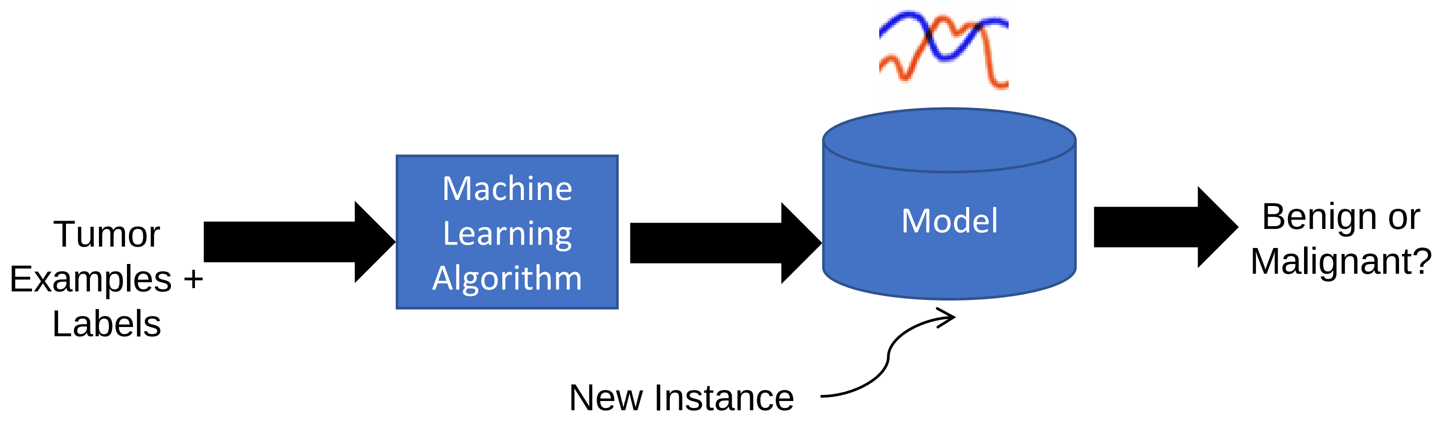


Image from *http://prooffreaderswhimsy.blogspot.com/2014/11/machine-learning.html

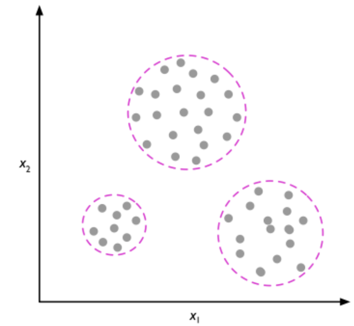
Supervised Machine Learning

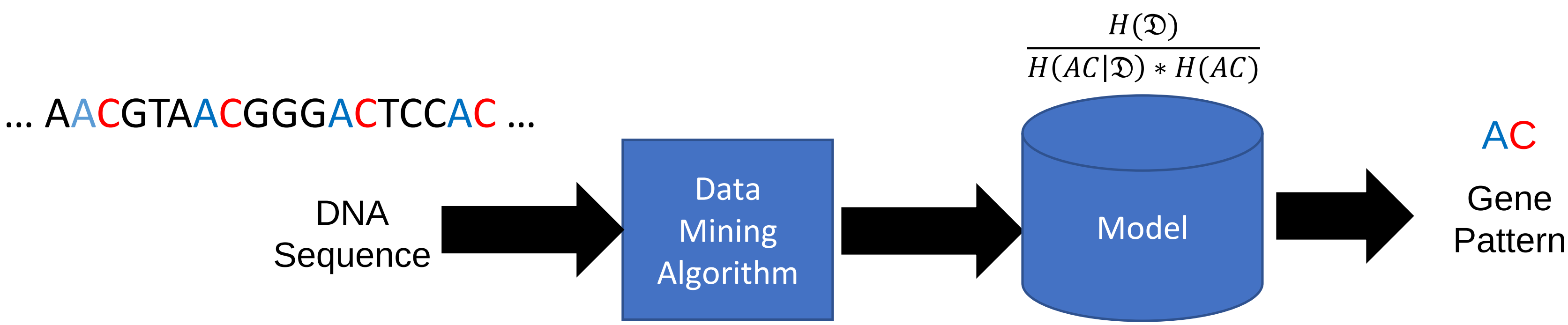
- Goal is Prediction (classification or regression)
- Example:
  - **Input:** examples of benign (-) and malignant (+) tumors defined in terms of tumor shape, radius, ..
  - **Output:** predict whether a previously unseen example is benign or malignant



Unsupervised Machine Learning

- Also known as data mining
- Goal is knowledge discovery
- Example:
  - **Input:** DNA Sequence as a long string of {A,C,G,T}
  - **Output:** frequent subsequences (gene patterns)





Supervised vs. Unsupervised Learning

- **Supervised Learning** (“learn from my example”)
  - Goal: A program that performs a task as good as humans.
  - TASK – well defined (the target function)
  - EXPERIENCE – training data provided by a human
  - PERFORMANCE Metric – error/accuracy on the task
- **Unsupervised Learning** (“see what you can find”)
  - Goal: To find some kind of structure in the data.
  - TASK – vaguely defined
  - No EXPERIENCE: no labeled data
  - No PERFORMANCE Metric (but, there are some evaluations metrics)

*TAKIS METAXAS, CS 315 – Web Search and Data Mining

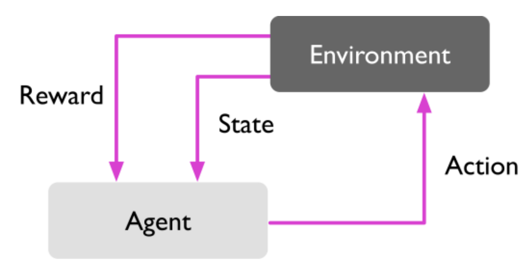
Beyond Supervised/Unsupervised

- Also
  - Semi-supervised learning => when a small amount of data is labeled
  - Transfer Learning => when labeled data is available in another domain

Supervised Learning
- Labeled data
- Direct feedback
- Predict outcome/future

Unsupervised Learning
- No labels
- No feedback
- Find hidden structure in data

Reinforcement Learning
- Decision process
- Reward system
- Learn series of actions



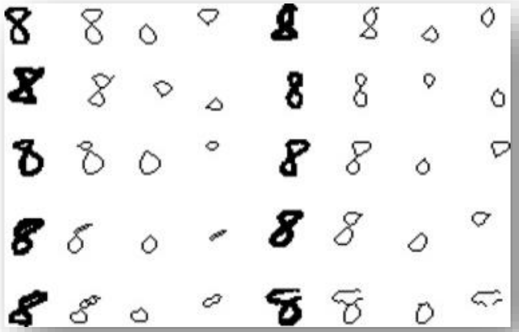
You don't Always need Machine Learning!

- Machine Learning definition (supervised):
  - The ability to learn and to improve with experience instead of using pre-determined rules.
- Consider the following two tasks:

Problem: Is **m** a prime number?

Solution: test up to $\sqrt{m}$ to see if m can be factored into two values.

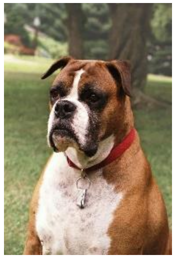
Testing for
Prime Numbers



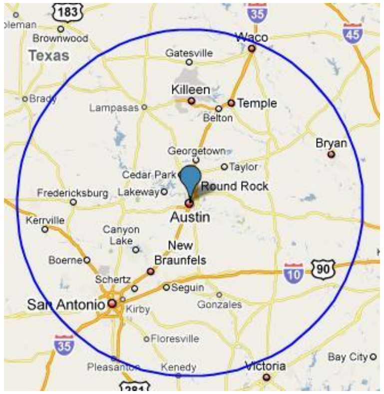
Recognizing
Handwritten Digits

Which Task Requires ML?

- Dog Recognition



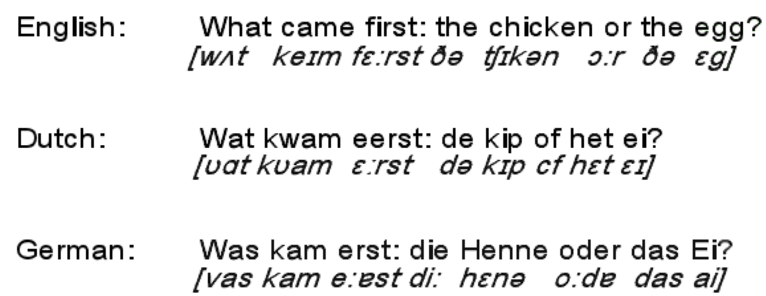
- Location Proximity Detection from GPS Signal



Which Task Requires ML?

- Speech Recognition

- Detecting if a given sentence is in English or German



en.wikipedia.org and-bmod.com

“When” Learning is needed?

- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition, image analysis)
  - Solution changes in time (decision support during surgery)
  - Solution needs to be adapted to particular cases (personalized medicine)