Информациски системи и големи податоци

Data Science, Machine Learning
- Introduction

Outline – Today

- **Data Science, Data Mining, ML Perspective**
- Machine Learning
- State of the Art ML Applications
- Illustrative Examples
- WEKA

Motivation

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data but starving for knowledge!
- Solution: Data warehousing, data mining, machine learning
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data mining, ML

- Process of semi-automatically analyzing large databases to find patterns that are:
  - **valid**: hold on new data with some certainity
  - **novel**: non-obvious to the system
  - **useful**: should be possible to act on the item
  - **understandable**: humans should be able to interpret the pattern
- Also known as Knowledge Discovery in Databases (KDD)

What Is Data Mining, ML?

- **What is not?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining / ML
  - Find all credit applicants who are poor credit risks. (Classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (Association rules)

Outline – Today

- Data Science, Data Mining, ML Perspective
- **Machine Learning**
- State of the Art ML Applications
- Illustrative Examples
- WEKA

Machine Learning

Machine Learning definition

“Learning is any process by which a system improves performance from experience.” Herbert Simon

Definition by Tom Mitchell (1998):

- Machine Learning is the study of algorithms that
  - improve their performance P
  - at some task T
  - with experience E.
- A well-defined learning task is given by <P, T, E>.

Samuel's Checkers-Player

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." -Arthur Samuel (1959)

"A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

- ◯ Classifying emails as spam or not spam.
- ◯ Watching you label emails as spam or not spam.
- ◯ The number (or fraction) of emails correctly classified as spam/not spam.
- ◯ None of the above—this is not a machine learning problem.

Defining the Learning Task

Improve on task T, with respect to

performance metric P, based on experience E

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself

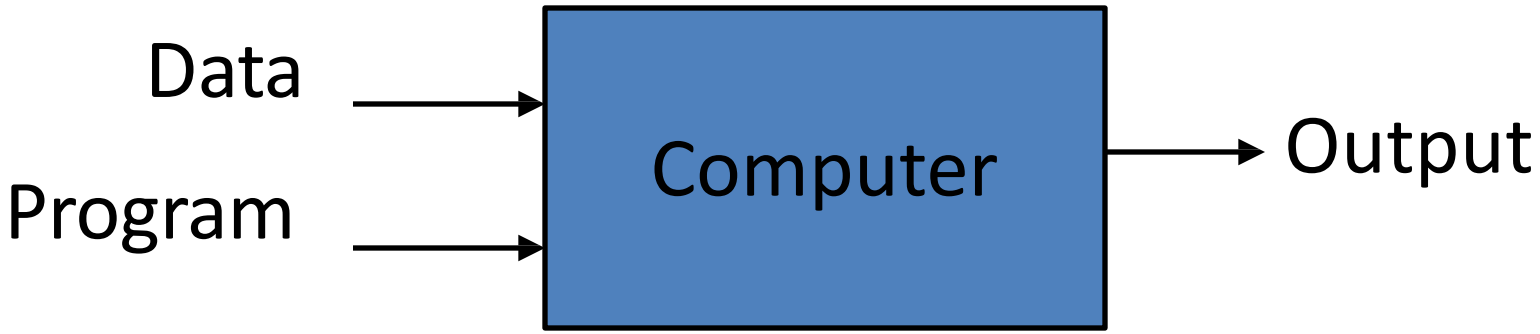
T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver.

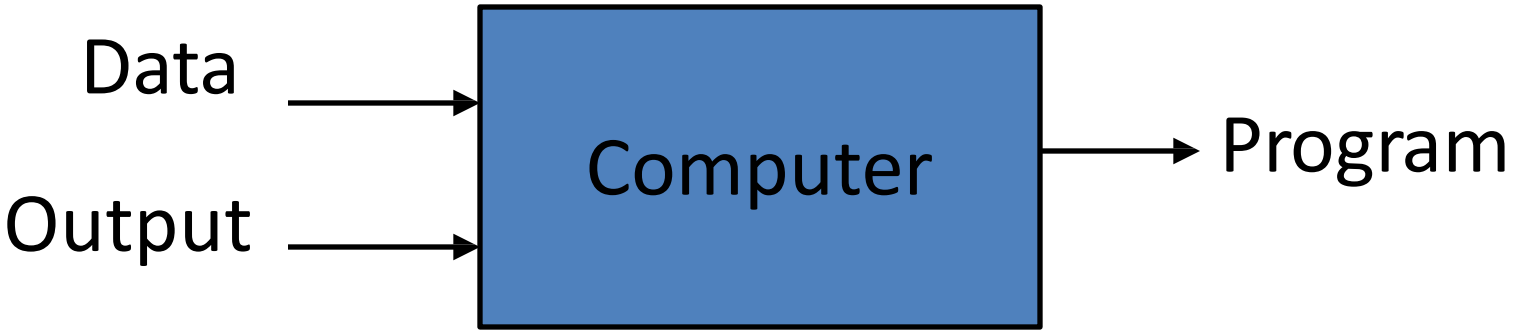
T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

Slide credit: Ray Mooney

Traditional Programming



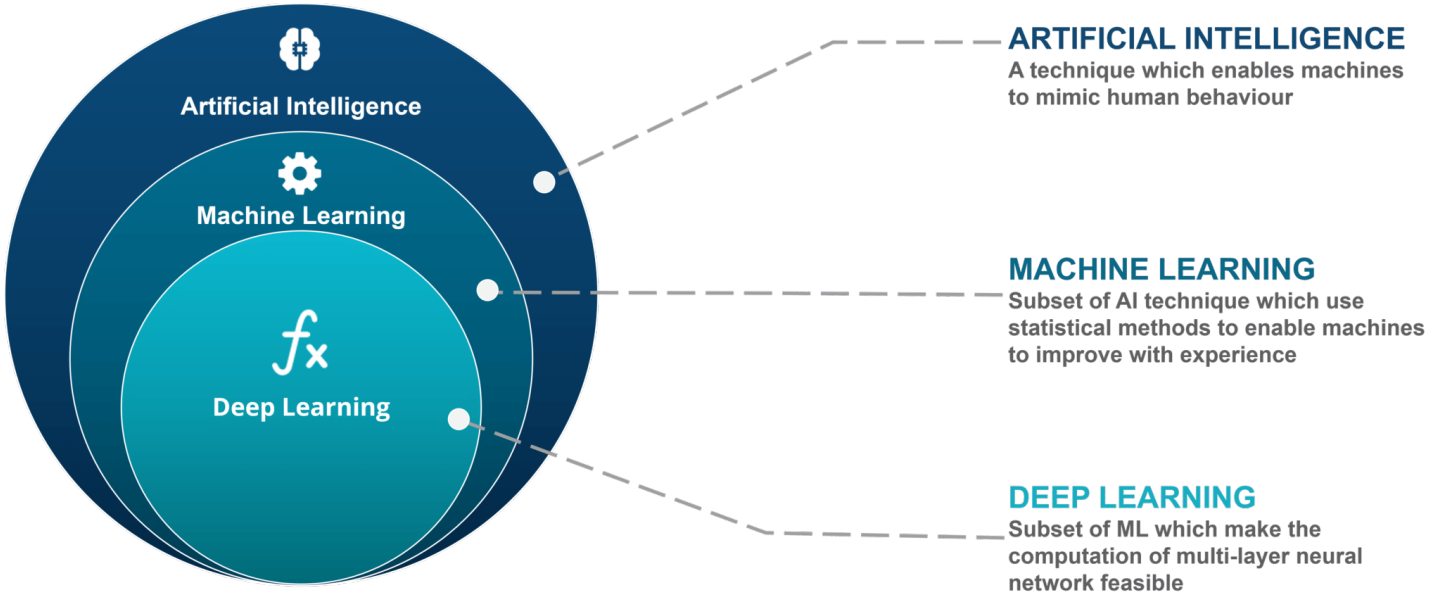
Machine Learning



Slide credit: Pedro Domingos

Machine Learning

- Scientific study of **algorithms** and **statistical models** that computer systems use to perform a **specific task** without using **explicit instructions**, relying on **patterns and inference instead**.



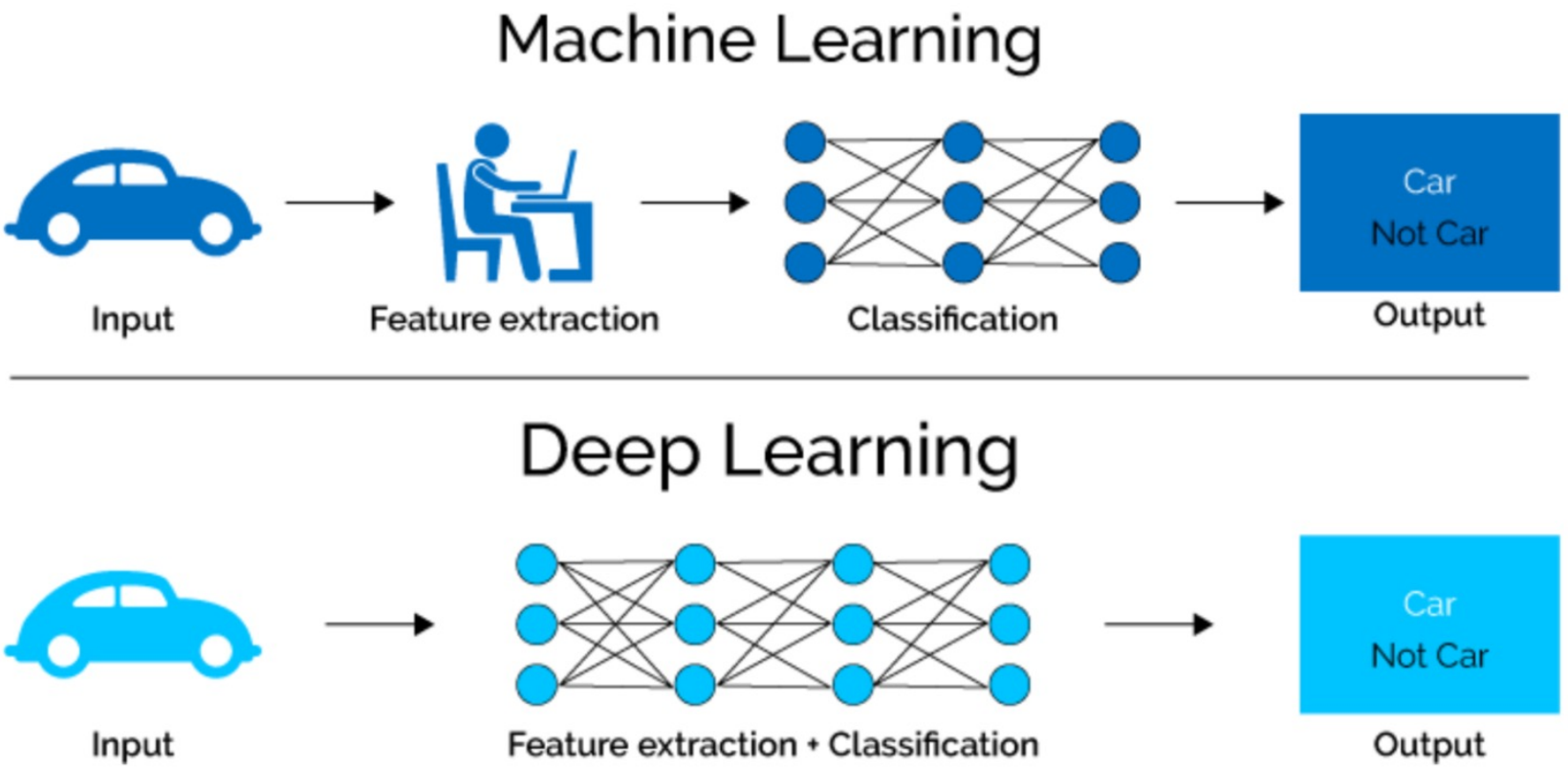


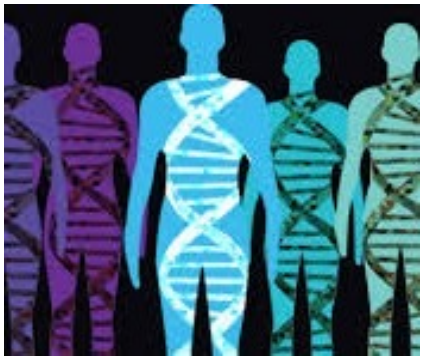
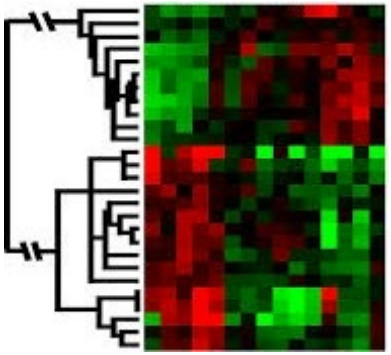
Figure 1: Machine Learning VS Deep Learning

When Do We Use Machine Learning?

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



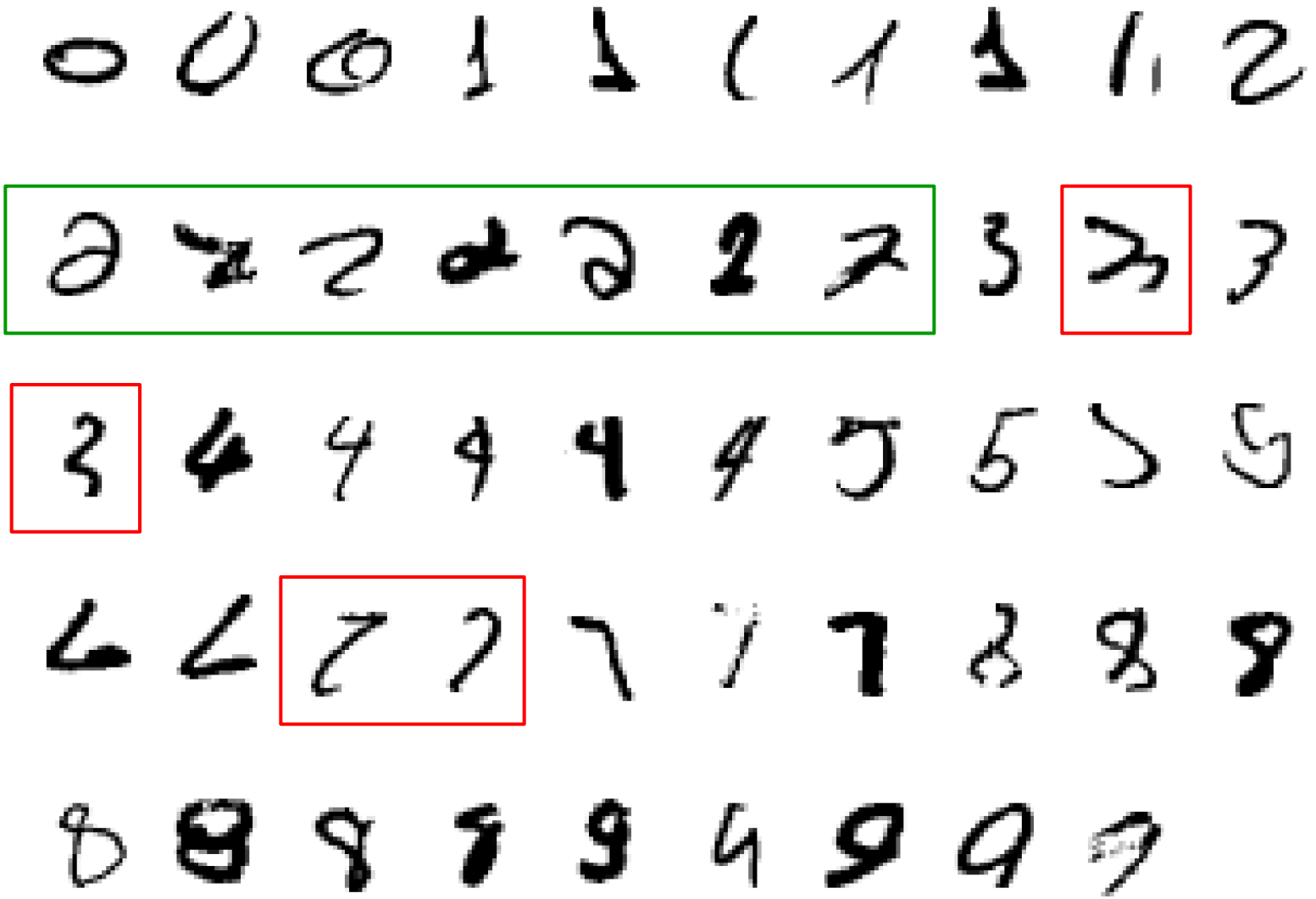



Learning isn't always useful:

- There is no need to "learn" to calculate payroll

Based on slide by E. Alpaydin

A classic example of a task that requires machine learning: It is very hard to say what makes a 2

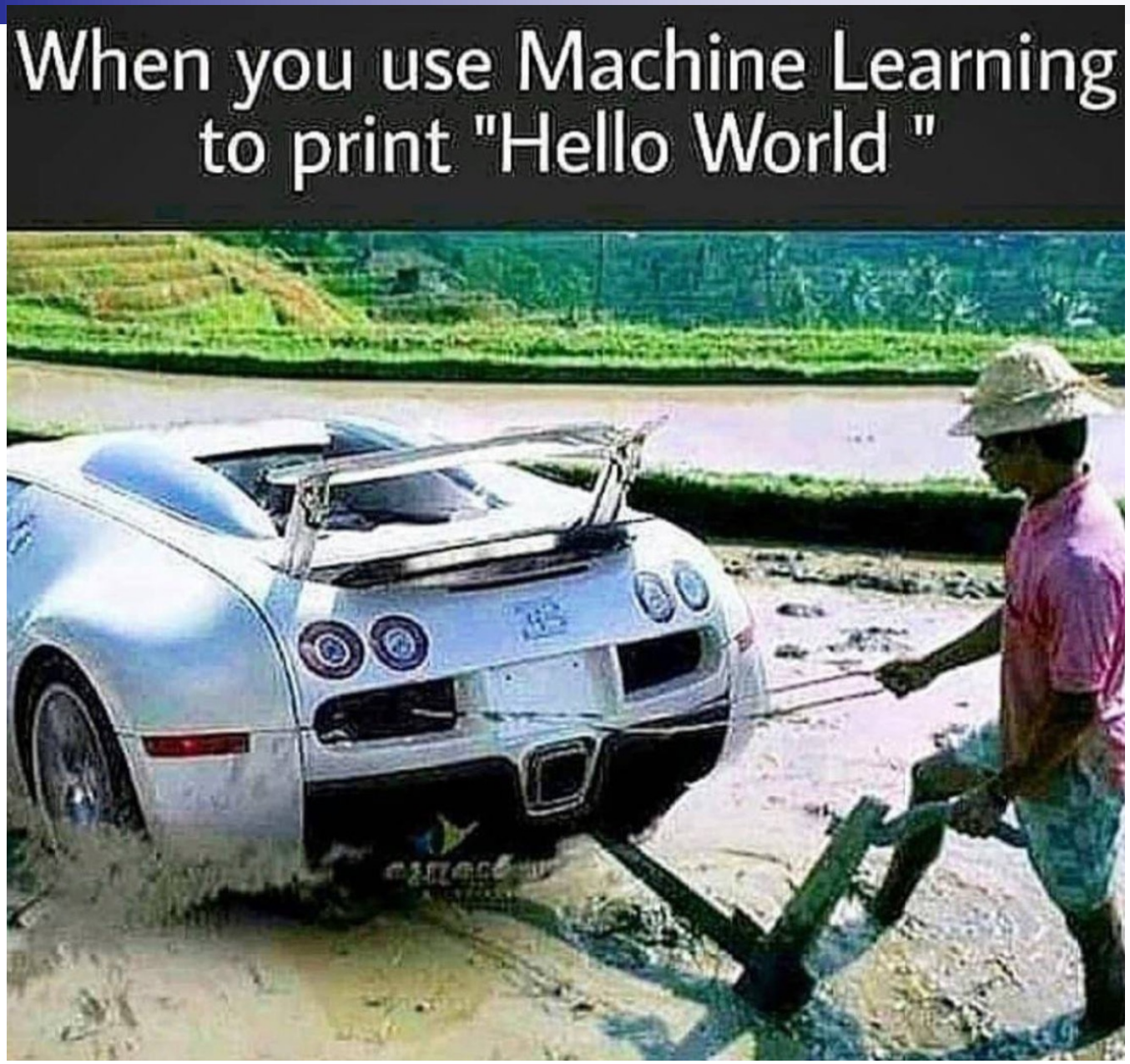


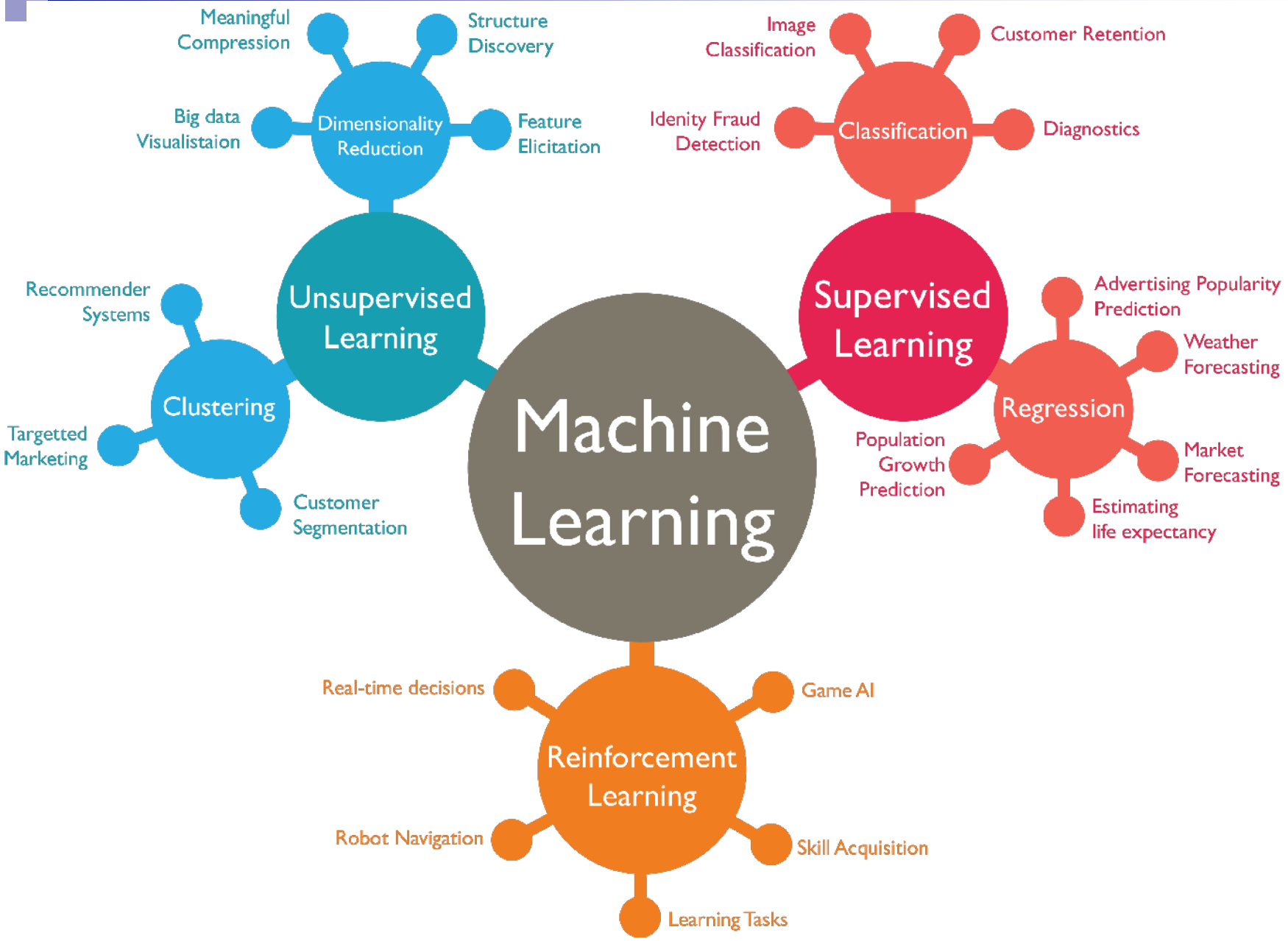
Slide credit: Geoffrey Hinton

Some more examples of tasks that are best solved by using a learning algorithm

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction:
  - Future stock prices or currency exchange rates

Slide credit: Geoffrey Hinton


When you use Machine Learning
to print "Hello World "


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Types of Learning

- **Supervised (inductive) learning**
  - Given: training data + desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Semi-supervised learning**
  - Given: training data + a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Based on slide by Pedro Domingos

Supervised Learning

- Like human learning from past experiences.
- A computer does not have “experiences”.
- A computer system learns from data, which represent some “past experiences” of an application domain.
- Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.
- The task is commonly called: Supervised learning, classification, or inductive learning.

Supervised Learning Example

- A credit card company receives thousands of applications for new cards. Each application contains information about an applicant,
  - age
  - Marital status
  - annual salary
  - outstanding debts
  - credit rating
  - etc.
- Problem: to decide whether an application should approved, or to classify applications into two categories, approved and not approved.

The data and the goal

- Data: A set of data records (also called examples, instances or cases) described by
  - *k* attributes/features: A_1, A_2, … A_k.
  - a class: Each example is labelled with a pre-defined class.
- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

An example: the learning task

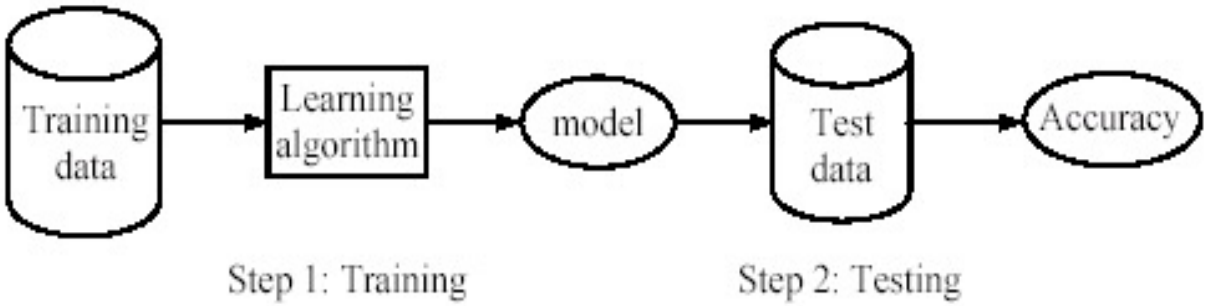
- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

Supervised learning process: two steps

- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



What do we mean by learning?

- Given
  - a data set *D*,
  - a task *T*, and
  - a performance measure *M*,

  a computer system is said to **learn** from *D* to perform the task *T* if after learning the system's performance on *T* improves as measured by *M*.
- In other words, the learned model helps the system to perform *T* better as compared to no learning.

An example

- Data: Loan application data
- Task: Predict whether a loan should be approved or not.
- Performance measure: accuracy.

No learning: classify all future applications (test data) to the majority class (i.e., Yes):

Accuracy = 9/15 = 60%.

- We can do better than 60% with learning.

Decision Tree Algorithm

- Decision tree learning is one of the most widely used techniques for classification.
  - Its classification accuracy is competitive with other methods, and
  - it is very efficient.
- The classification model is a tree, called decision tree.

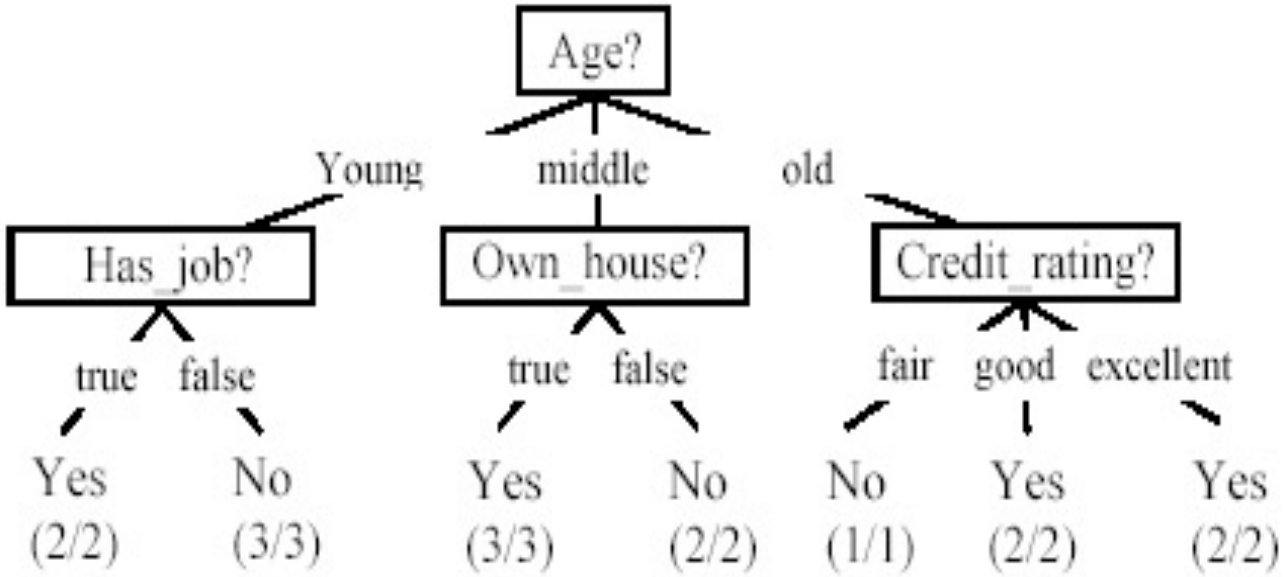
The loan data (reproduced)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

A decision tree from the loan data

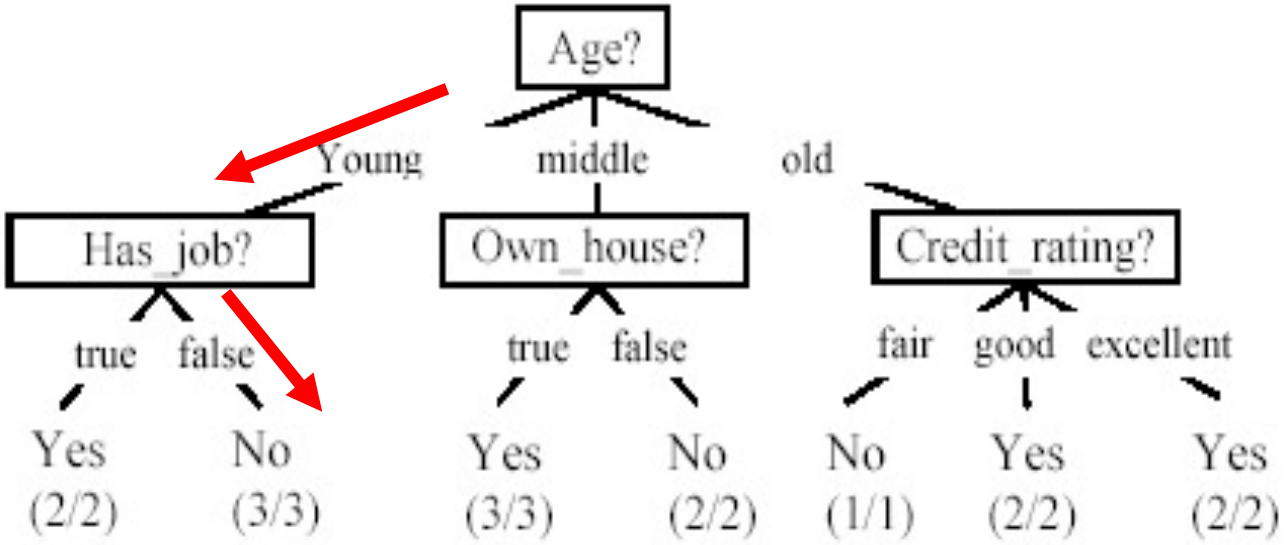
- Decision nodes and leaf nodes (classes)



Use the decision tree

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

No



Regression

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class	
1	young	false	false	fair	No	1.2
2	young	false	false	good	No	2.1
3	young	true	false	good	Yes	4
4	young	true	true	fair	Yes	5
5	young	false	false	fair	No	0.3
6	middle	false	false	fair	No	1.1
7	middle	false	false	good	No	0
8	middle	true	true	good	Yes	5
9	middle	false	true	excellent	Yes	4.4
10	middle	false	true	excellent	Yes	4.8
11	old	false	true	excellent	Yes	5
12	old	false	true	good	Yes	4.2
13	old	true	false	good	Yes	5
14	old	true	false	excellent	Yes	4.1
15	old	false	false	fair	No	1.2

Predict continuous valued output – credit score from 0 to 5

Quiz

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

- Treat both as classification problems.
- Treat problem 1 as a classification, problem 2 as a regression problem.
- Treat problem 1 as a regression, problem 2 as a classification problem.
- Treat both as regression problems.

Unsupervised Learning

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class	
1	young	false	false	fair	No	1.2
2	young	false	false	good	No	2.1
3	young	true	false	good	Yes	4
4	young	true	true	fair	Yes	5
5	young	false	false	fair	No	0.3
6	middle	false	false	fair	No	1.1
7	middle	false	false	good	No	0
8	middle	true	true	good	Yes	5
9	middle	false	true	excellent	Yes	4.4
10	middle	false	true	excellent	Yes	4.8
11	old	false	true	excellent	Yes	5
12	old	false	true	good	Yes	4.2
13	old	true	false	good	Yes	5
14	old	true	false	excellent	Yes	4.1
15	old	false	false	fair	No	1.2

Groups of similar customers, clusters

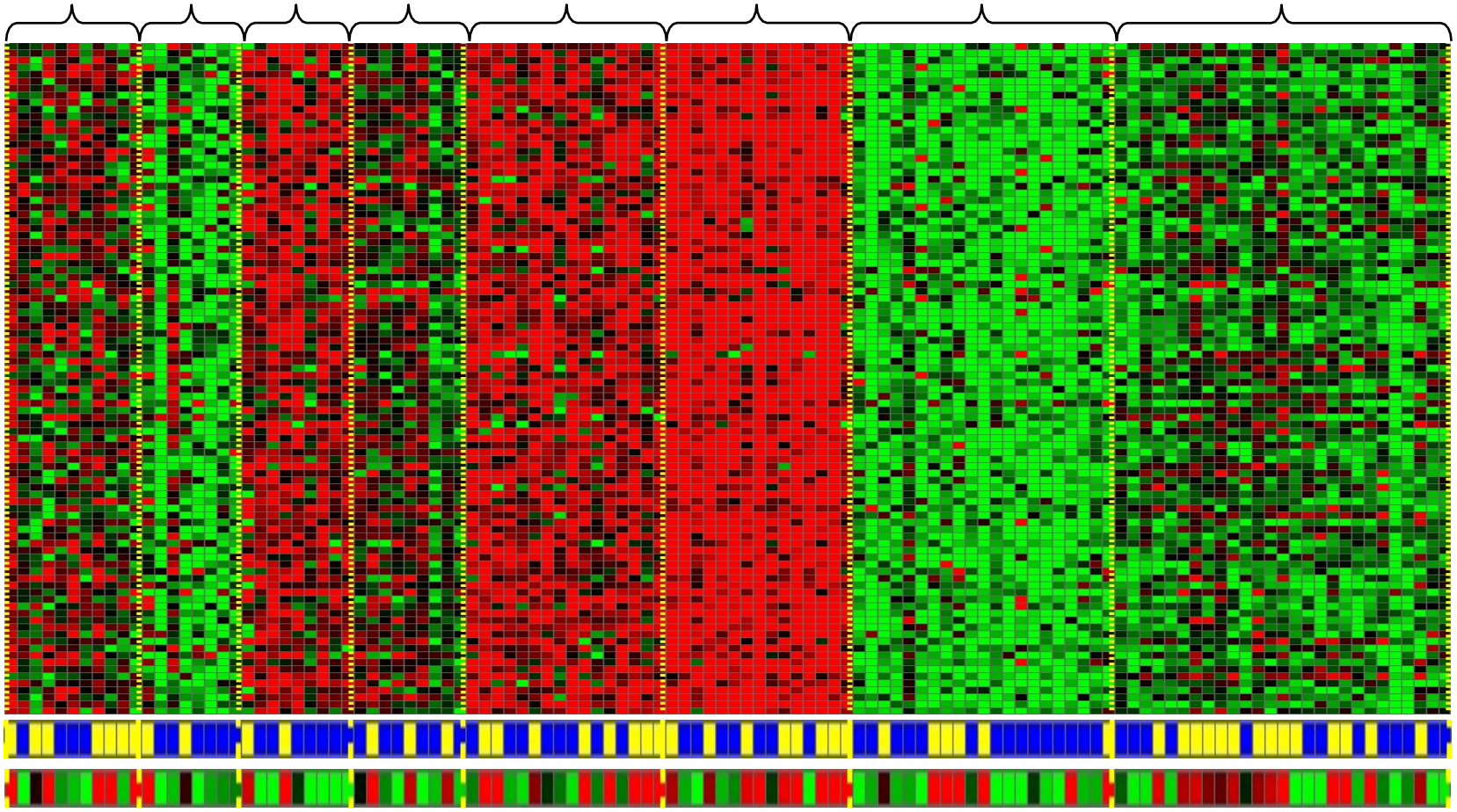
Supervised vs. unsupervised Learning

- Supervised learning: classification is seen as supervised learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a “teacher” gives the classes (supervision).
  - Test data are classified into these classes too.

- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Unsupervised Learning

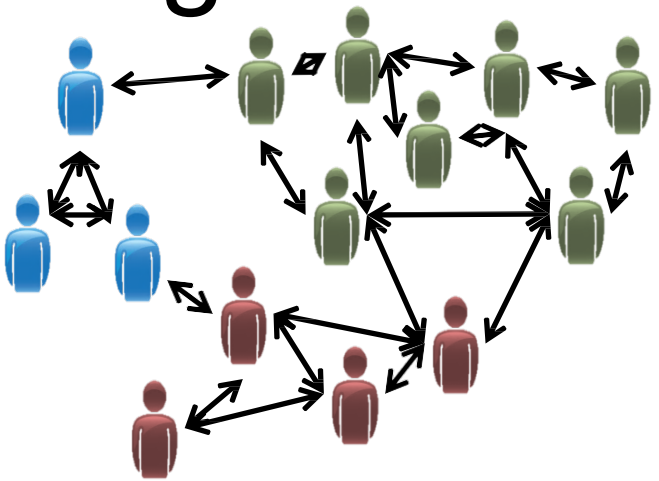
Genomics application: group individuals by genetic similarity



[Source: Daphne Koller]

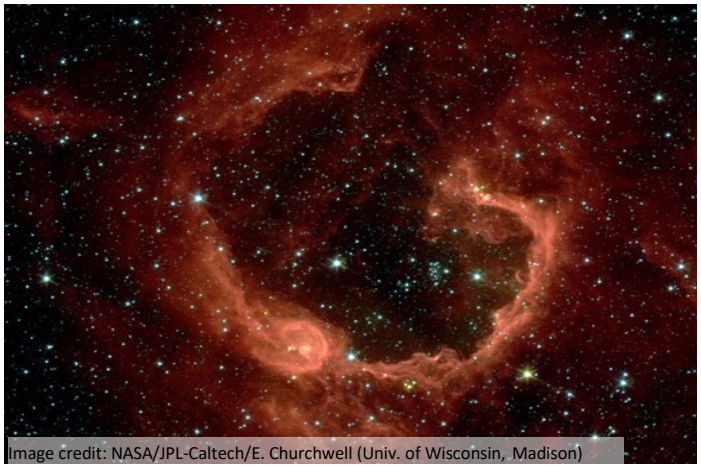
Unsupervised Learning

Organize computing clusters



Social network analysis

Market segmentation

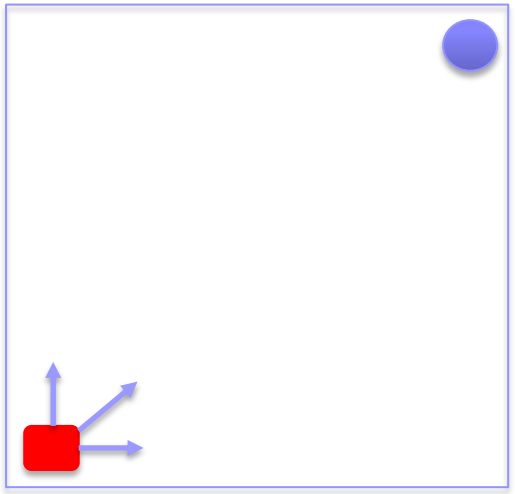


Astronomical data analysis

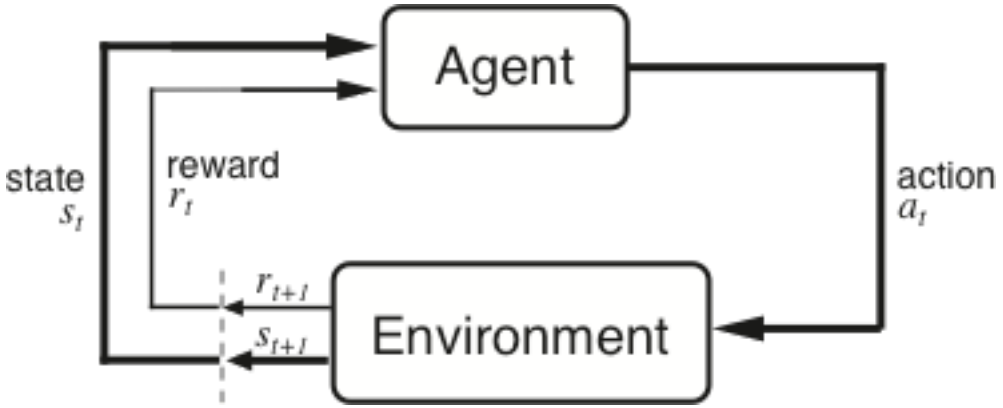
Slide credit: Andrew Ng

Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples:
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand



The Agent-Environment Interface



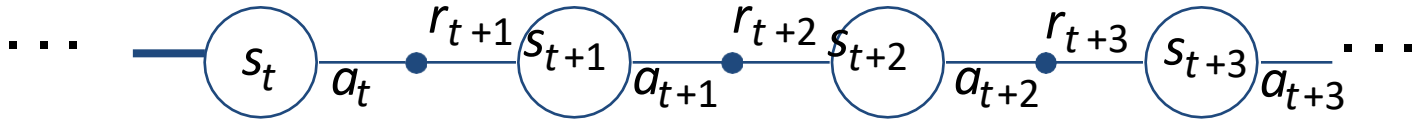
Agent and environment interact at discrete time steps : $t = 0, 1, 2, \mathbb{K}$

Agent observes state at step t : $s_t \in S$

produces action at step t : $a_t \in A(s_t)$

gets resulting reward : $r_{t+1} \in \Re$

and resulting next state : s_{t+1}



Slide credit: Sutton & Barto

Reinforcement Learning

https://www.youtube.com/watch?v=4cgWya-wjgY

Facebook AI Chief
Yann LeCun Cake

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**

Self-supervised learning

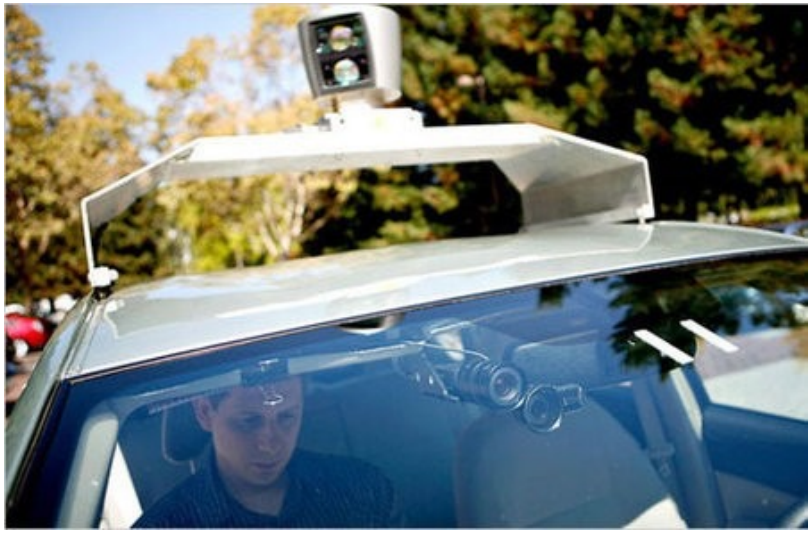
- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

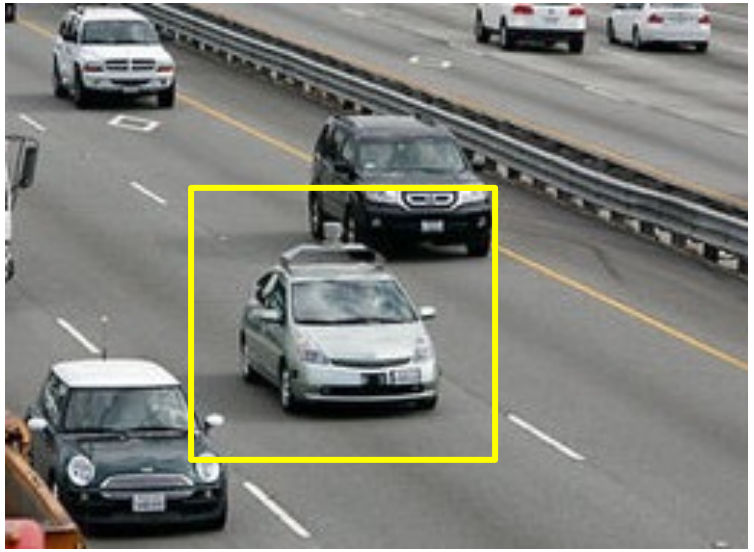
Self-supervised Learning

Outline – Today

- Data Science, Data Mining, ML Perspective
- Machine Learning
- **State of the Art ML Applications**
- Illustrative Examples
- WEKA

Autonomous Cars

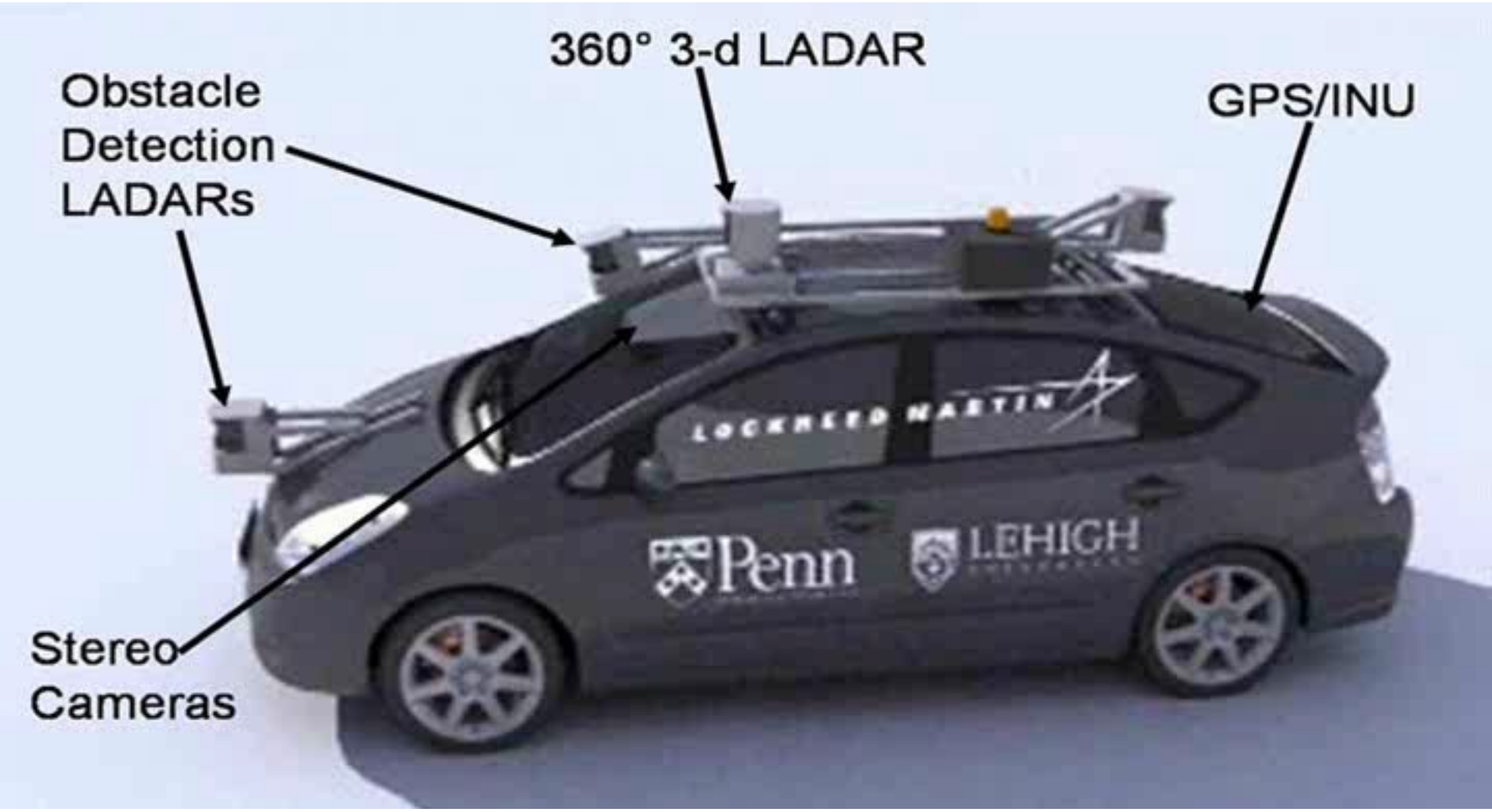




- Nevada made it legal for autonomous cars to drive on roads in June 2011
- As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars

Penn's Autonomous Car →
(Ben Franklin Racing Team)

Autonomous Car Sensors

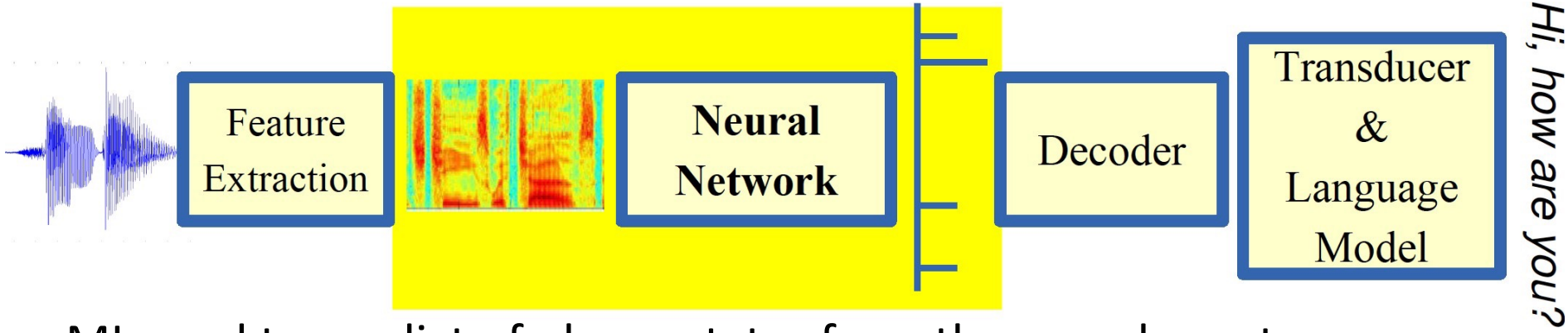


Scene Labeling via Deep Learning

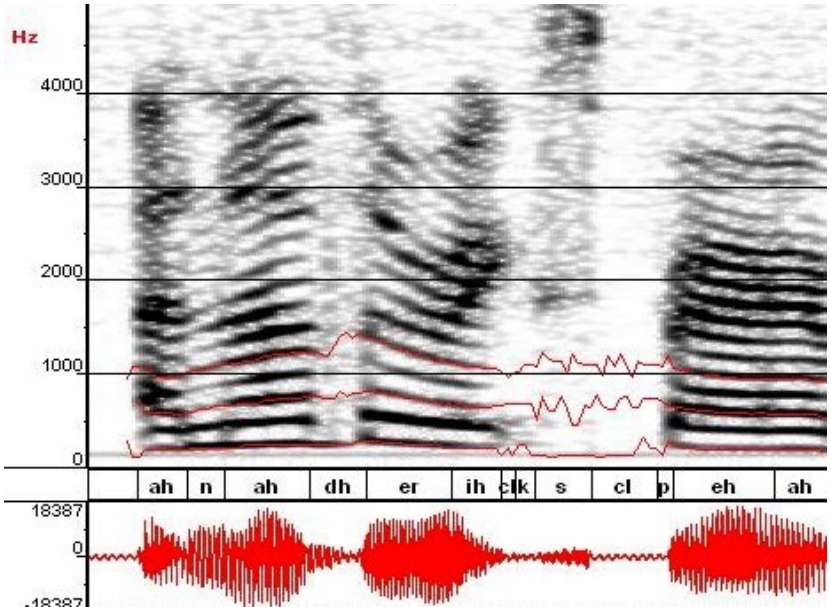
[Farabet et al. ICML 2012, PAMI 2013]

Machine Learning in Automatic Speech Recognition

A Typical Speech Recognition System



ML used to predict of phone states from the sound spectrogram



Deep learning has state-of-the-art results

# Hidden Layers	1	2	4	8	10	12
Word Error Rate %	16.0	12.8	11.4	10.9	11.0	11.1

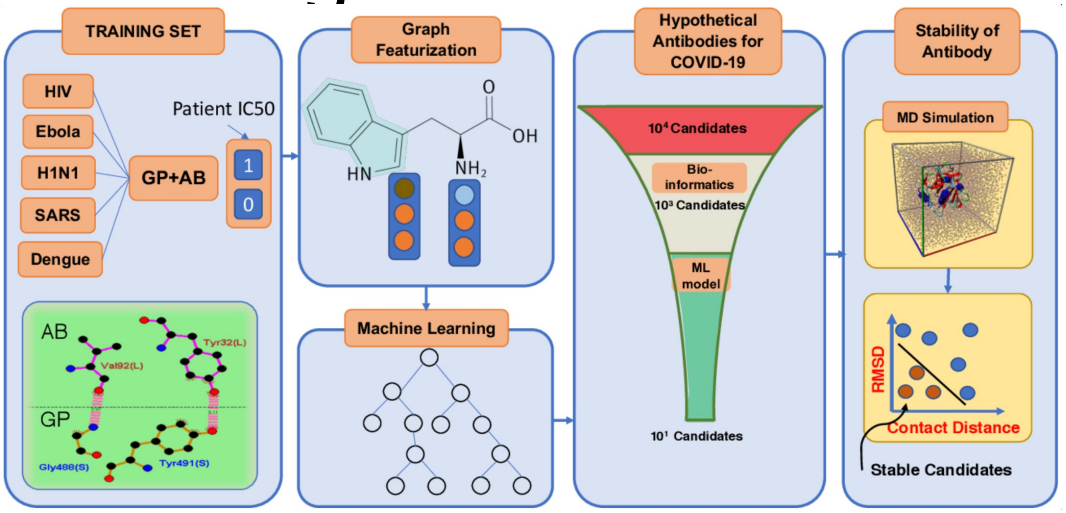
Baseline GMM performance = 15.4%

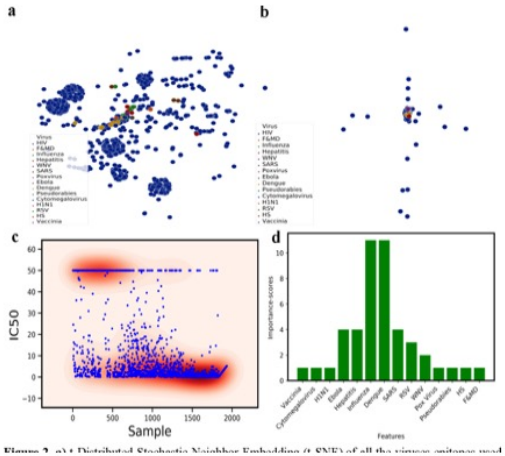
[Zeiler et al. "On rectified linear units for speech recognition" ICASSP 2013]

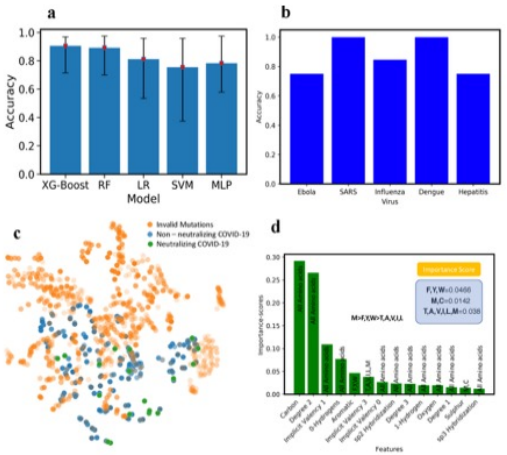
Impact of Deep Learning in Speech Technology

Slide credit: Li Deng, MS Research

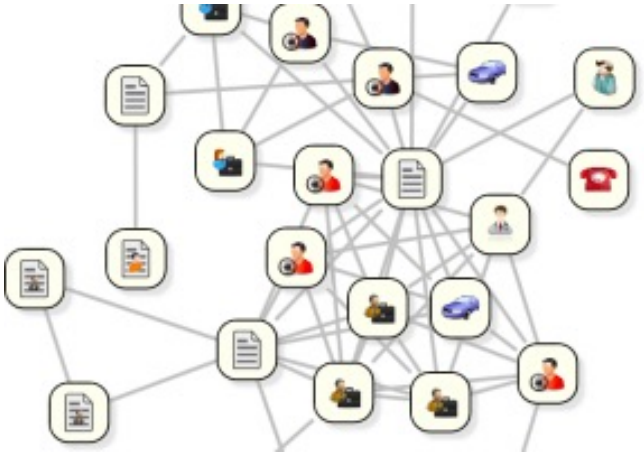
Potential Neutralizing Antibodies Discovered for Novel Corona Virus Using Machine Learning





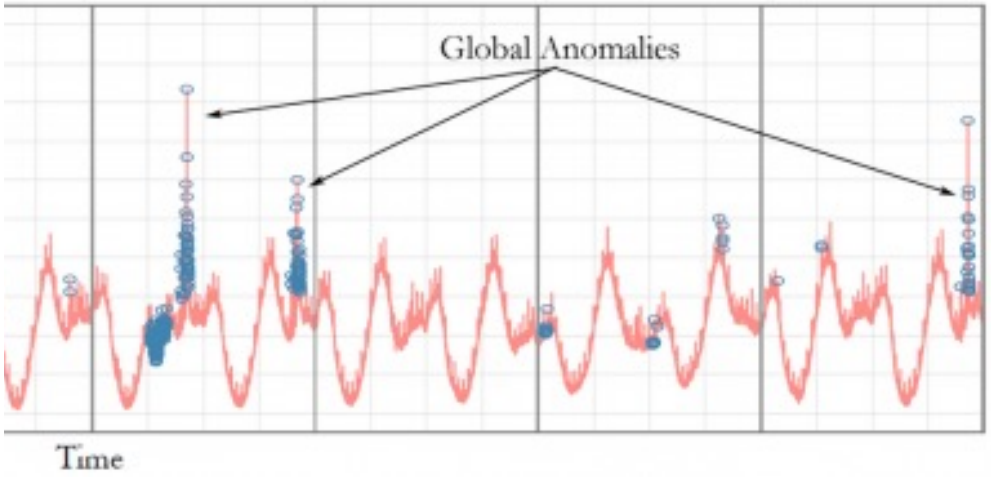


Examples of Machine Learning



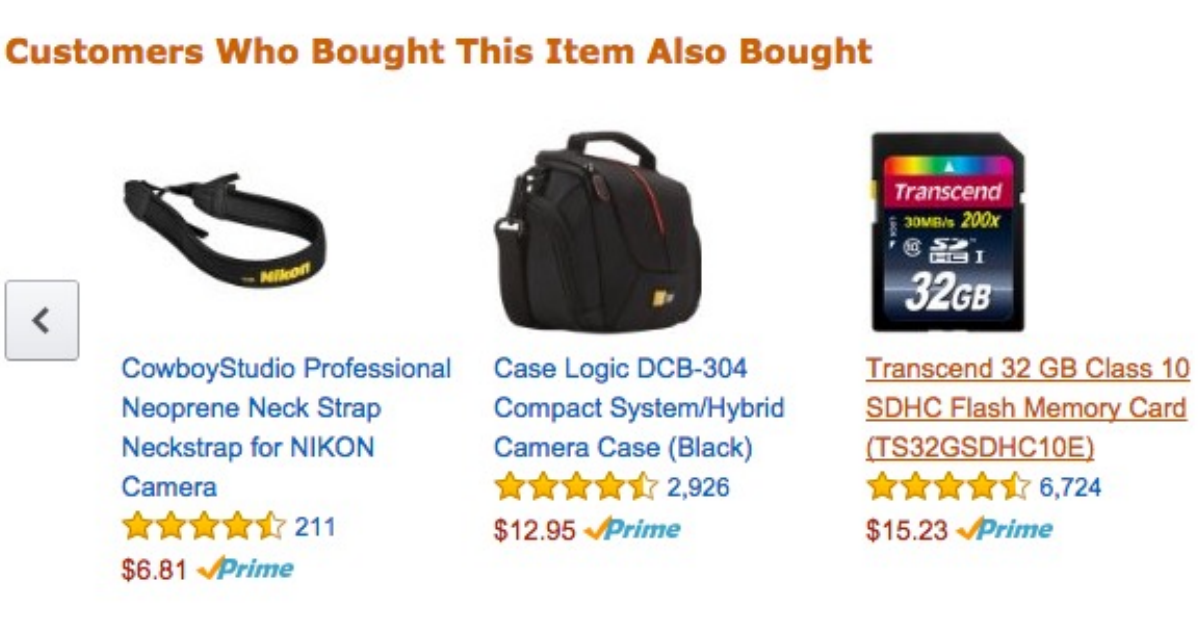
Fraud detection

Anomaly detection in signals



Examples of Machine Learning

Recommender systems



Case Study: Bank

- Business goal: Sell more home equity loans
- Current models:
  - Customers with college-age children use home equity loans to pay for tuition
  - Customers with variable income use home equity loans to even out stream of income
- Data:
  - Large data warehouse
  - Consolidates data from 42 operational data sources

Case Study: Bank (Contd.)

1. Select subset of customer records who have received home equity loan offer
   - Customers who declined
   - Customers who signed up

Income	Number of Children	Average Checking Account Balance	…	Reponse
$40,000	2	$1500		Yes
$75,000	0	$5000		No
$50,000	1	$3000		No
…	…	…	…	…

Case Study: Bank (Contd.)

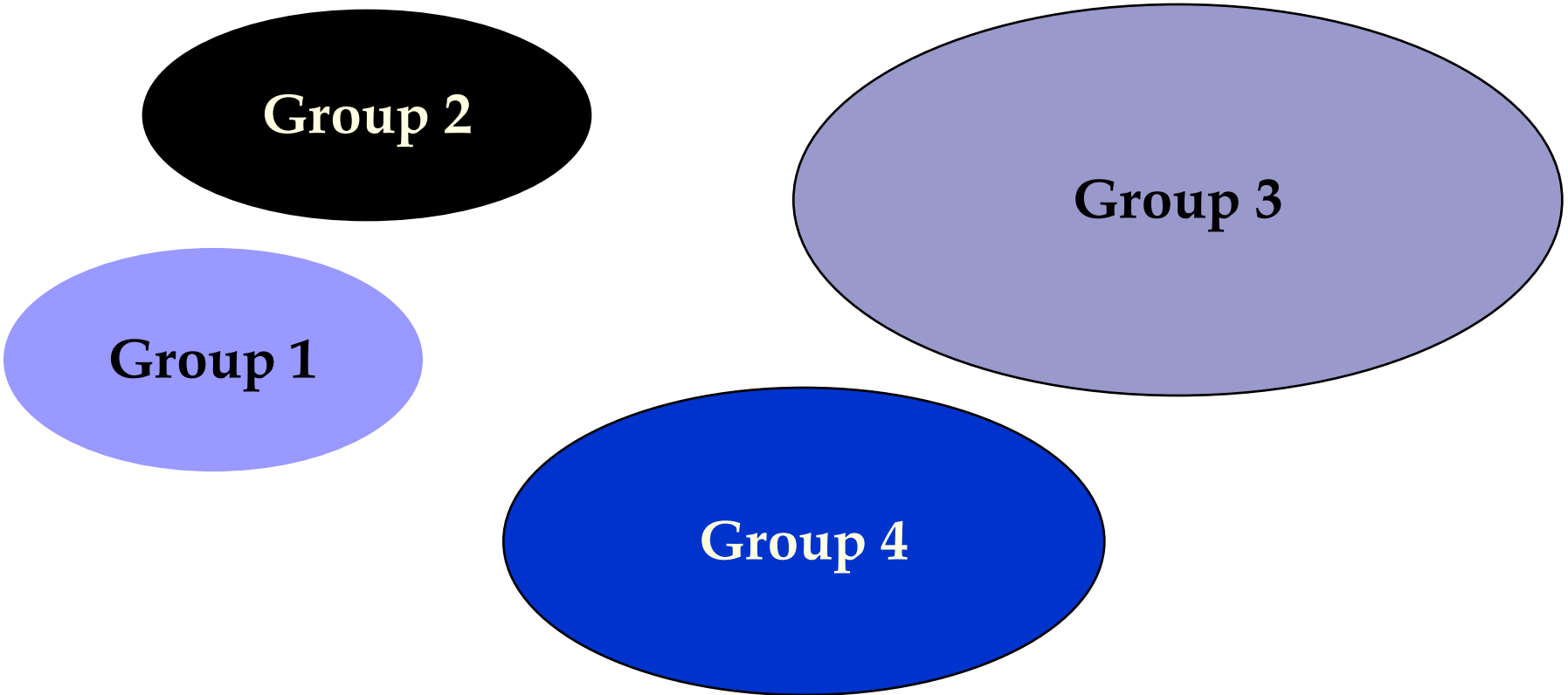
2. Find rules to predict whether a customer would respond to home equity loan offer

IF (Salary < 40k) and
(numChildren > 0) and
(ageChild1 > 18 and ageChild1 < 22)
THEN YES

…

Case Study: Bank (Contd.)

3. Group customers into clusters and investigate clusters



Case Study: Bank (Contd.)

4. Evaluate results:
   - Many “uninteresting” clusters
   - One interesting cluster! Customers with both business and personal accounts; unusually high percentage of likely respondents

Example: Bank (Contd.)

Action:

- New marketing campaign

Result:

- Acceptance rate for home equity offers more than doubled

Example Application: Fraud Detection

- Industries: Health care, retail, credit card services, telecom, B2B relationships
- Approach:
  - Use historical data to build models of fraudulent behavior
  - Deploy models to identify fraudulent instances

Fraud Detection (Contd.)

- Examples:
  - Auto insurance: Detect groups of people who stage accidents to collect insurance
  - Medical insurance: Fraudulent claims
  - Money laundering: Detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - Telecom industry: Find calling patterns that deviate from a norm (origin and destination of the call, duration, time of day, day of week).

Outline – Today

- Data Science, Data Mining, ML Perspective
- Machine Learning
- State of the Art ML Applications
- **Illustrative Examples**
- WEKA

Illustrative example 1:

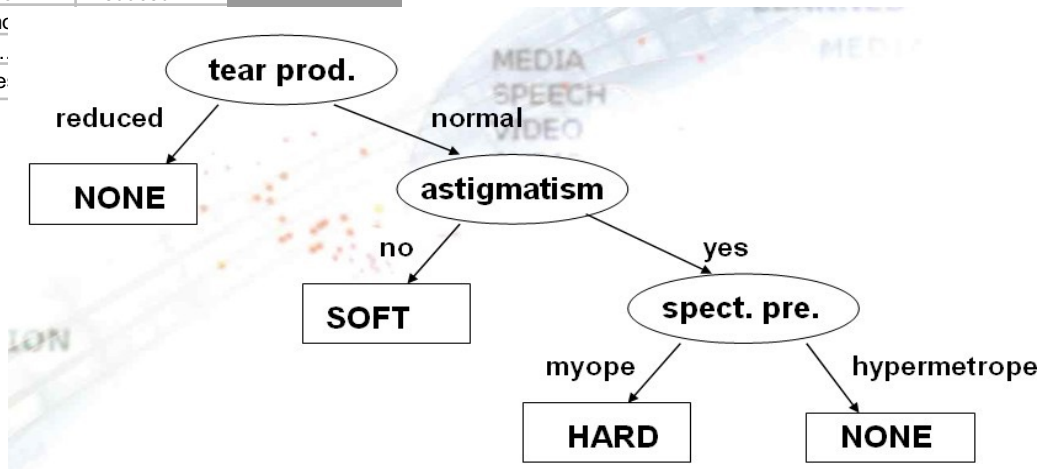
Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	young	myope	no	reduced	NONE
O2	young	myope	no	normal	SOFT
O3	young	myope	yes	reduced	NONE
O4	young	myope	yes	normal	HARD
O5	young	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	pre-presby	hypermetrope	no	normal	SOFT
O15	pre-presby	hypermetrope	yes	reduced	NONE
O16	pre-presby	hypermetrope	yes	normal	NONE
O17	presbyopic	myope	no	reduced	NONE
O18	presbyopic	myope	no		
O19-O23	...	...	...		
O24	presbyopic	hypermetrope	yes		

Data Mining



Learning from Numeric Class Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	LensPrice
O1	17	myope	no	reduced	0
O2	23	myope	no	normal	8
O3	22	myope	yes	reduced	0
O4	27	myope	yes	normal	5
O5	19	hypermetrope	no	reduced	0
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	5
O15	43	hypermetrope	yes	reduced	0
O16	39	hypermetrope	yes	normal	0
O17	54	myope	no	reduced	0
O18	62	myope	no	normal	0
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	0

Numeric class values – regression analysis

Learning from Unlabeled Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Unlabeled data - clustering: grouping of similar instances
- association rule learning

Illustrative Example 2: Customer Relationship Management

What is Customer Relationship Management (CRM)?

- CRM is used to learn more about your key customers needs in order to develop a stronger relationship with them
- CRM can be defined as "companies activities related to increasing the customer base by acquiring new customers and meeting the needs of the existing customers”
- Customer Oriented

ML for CRM

- "Extraction of hidden predictive information from large databases"
- Help companies focus on the most important information in their data warehouses
- ML tools **predict future trends** and behaviors, allowing businesses to make proactive, knowledge-driven decisions
- It is no longer possible to wait until the signs of customer dissatisfaction are obvious before action must be taken
- To succeed, companies must be proactive and anticipate what a customer desires
- More customers, more products, more competitors, and less time to react means that understanding customers is now much harder to do

ML Tasks

- Classification & Regression
  - Estimation
  - Prediction
- Affinity grouping or association rules
- Clustering
- Description and visualization

Use case 1: Offering new product

- Mailing directed at a given customer base
- Typically: 1% of contacted customers are responders who will purchase the offered product
- A mailing of 100,000 will result in about 1,000 sales
- Data mining: identify which customers are most likely to respond to the campaign (based on the past records)
- Response raised from 1% to 1.25%: the sales of 1,000 could be achieved with only 80,000 mailings, reducing the mailing cost by one-fifth

Use case 2: Car Insurance

- Sports car owners fall into a high-risk category
- By mining driver safety data in data warehouse: if sports car enthusiasts also own a second, conventional car, they may be safe-enough drivers to be attractive policy holders
- As a result of the discovered micro-niche among sports car owners, the company changed how they underwrite and price some sport car policies

Use case 3: Churn Prediction

- Churn – a customer of a mobile telephone company that is likely to leave in near future
- The cost of keeping customers around is significantly less than the cost of bringing them back after they leave
- Traditional approach: pick up good customers and persuade them (with a gift) to sign for another year of service
- ML: segment the customers, determine what is your value to them, give them what they need (reliability, latest features, better rate for evening calls)

Use case 3: Churn Prediction

- The traditional approach - pick out good customers and try to persuade them to sign up for another year of service … gift (possibly a new phone) or maybe a discount calling plan
- This solution is probably very wasteful - there are undoubtedly many "good" customers who would be willing to stick around without receiving an expensive gift
- The customers to concentrate on **are the ones that will be leaving** - don't worry about the ones who will stay
- Give your customers what they need - there are differences between your customers, and you need to understand those differences in order to optimize your relationships
- One big spending customer might value the relationship because of your high reliability, and thus wouldn't need a gift in order to continue with it

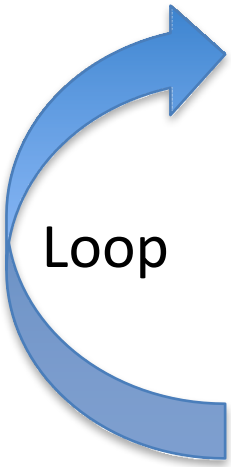
Use case 3: Churn Prediction – value for customer

- A customer who takes advantage of all of the latest features and special services might require a new phone or other gift in order to stick around for another year
- Or they might simply want a better rate for evening calls because their employer provides the phone and they have to pay for calls outside of business hours
- The key is determining which type of customer you're dealing with

Use case 3: Churn Prediction – value for customer and timing

- Consider timing in this process - do not wait until a week before a customer's contract and then pitch them an offer in order to prevent them from churning
- By then, they have likely decided what they are going to do and you are unlikely to affect their decision at such a late date
- Don't to start the process immediately - it might be months before they have an understanding of your company's value to them, so any efforts now would also be wasted

ML in Practice



- Understand domain, prior knowledge, and goals
- Data integration, selection, cleaning, pre-processing, etc.
- Learn models
- Interpret results
- Consolidate and deploy discovered knowledge

Based on a slide by Pedro Domingos

ML in a Nutshell

- Tens of thousands of machine learning algorithms
  - Hundreds new every year

- Every ML algorithm has three components:
  - **Representation**
  - **Optimization**
  - **Evaluation**

Slide credit: Pedro Domingos

10 Stages of an ML Project

1. Problem definition
2. Research
3. Data collection, aggregation, scrapping
4. Data preparation, preprocessing, augmentation
5. Model implementation
6. Training
7. Evaluation
8. Parameter tuning
9. Model conversion – mobile, cloud, …
10. Model deployment

https://towardsdatascience.com/10-stages-of-a-machine-learning-project-in-2020-and-where-you-fit-cb73ad4726cb

Outline – Today

- Data Science, Data Mining, ML Perspective
- Machine Learning
- State of the Art ML Applications
- Illustrative Examples
- **WEKA**

https://waikato.github.io/weka-wiki/downloading_weka/