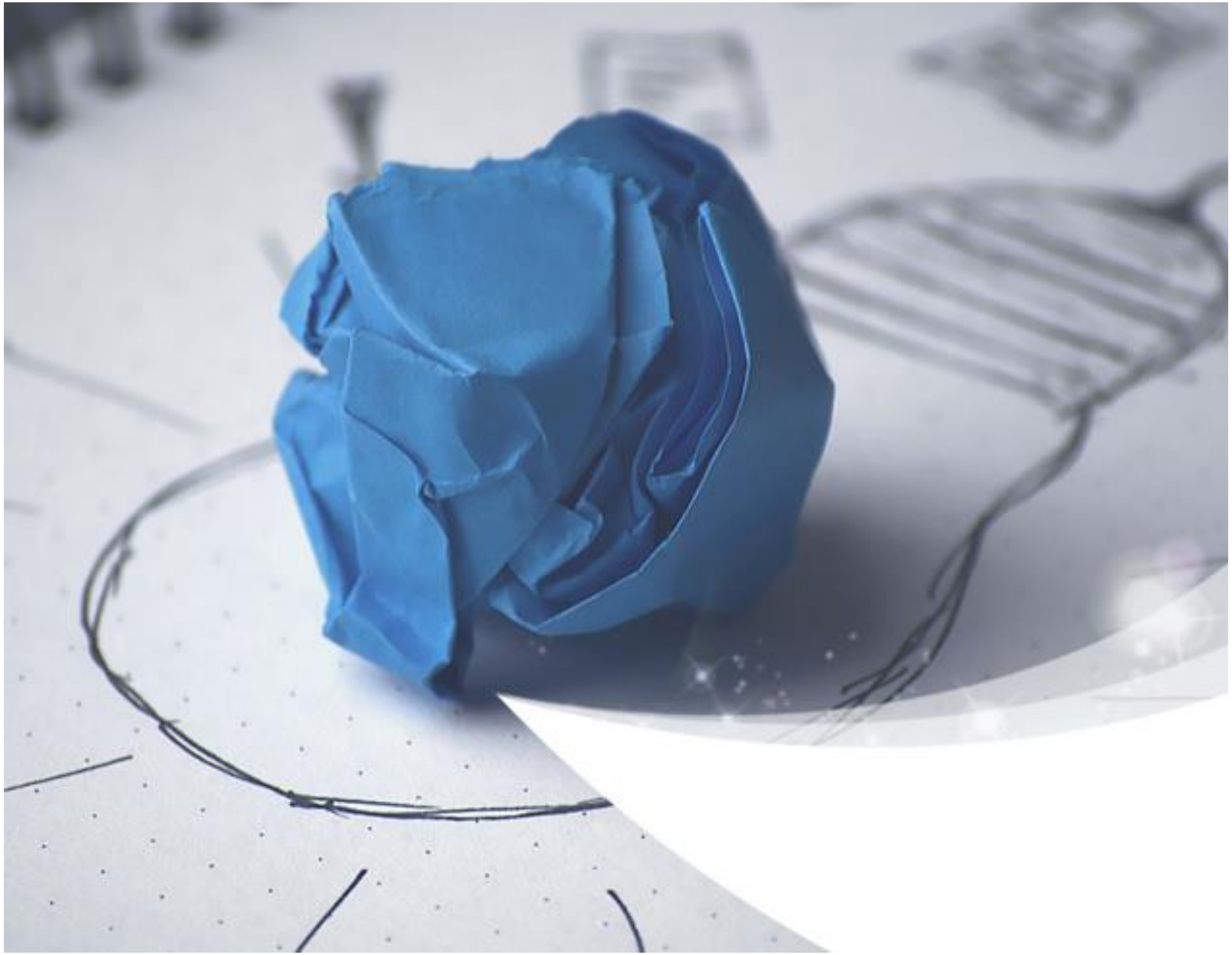


ADVANCED MACHINE LEARNING & TEXT MINING

Session 1

OUTLINE

Data Pre-processing

TEXT MINING

- Definitions
  - Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text." –Wikipedia
  - "Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known." -Hearst, 1999
  - "Text mining applies techniques such as categorization, entity extraction, sentiment analysis and natural language processing to transform text into data that can be used for further analysis." – expert system

TEXT MINING

Why Text Mining?

- 1.8 Zettabytes unstructured data
- Unlike search engines, which surface documents based on keywords, text mining tools analyze documents to identify entities and extract relationships between them, unlocking hidden information to help:
  - identify and develop new hypotheses attain knowledge
  - attain knowledge
  - improve understanding.

TEXT MINING

Why Text Mining?

- Text mining ensures the use of all of available information to make better informed decisions, automate information-intensive processes, gather business critical insights and mitigate operational risks.
- Text mining can solve high-value knowledge discovery problems in many different areas of application.

TEXT MINING

Types of text mining

- **Search and information retrieval**
  - Storage and retrieval of text documents, including search engines, and keyword search
- **Document Clustering**
  - Grouping and categorizing terms snippets, paragraph or documents using clustering algorithms
- **Document classification**
  - Grouping and categorizing terms snippets, paragraph or documents using classification methods, based on models trained on labels example

TEXT MINING

- Types of text mining
  - **Web mining**
    - Data and text mining on the Internet with specific focus on scale and interconnectedness
  - **Information Extraction**
    - Identification and extraction relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured data

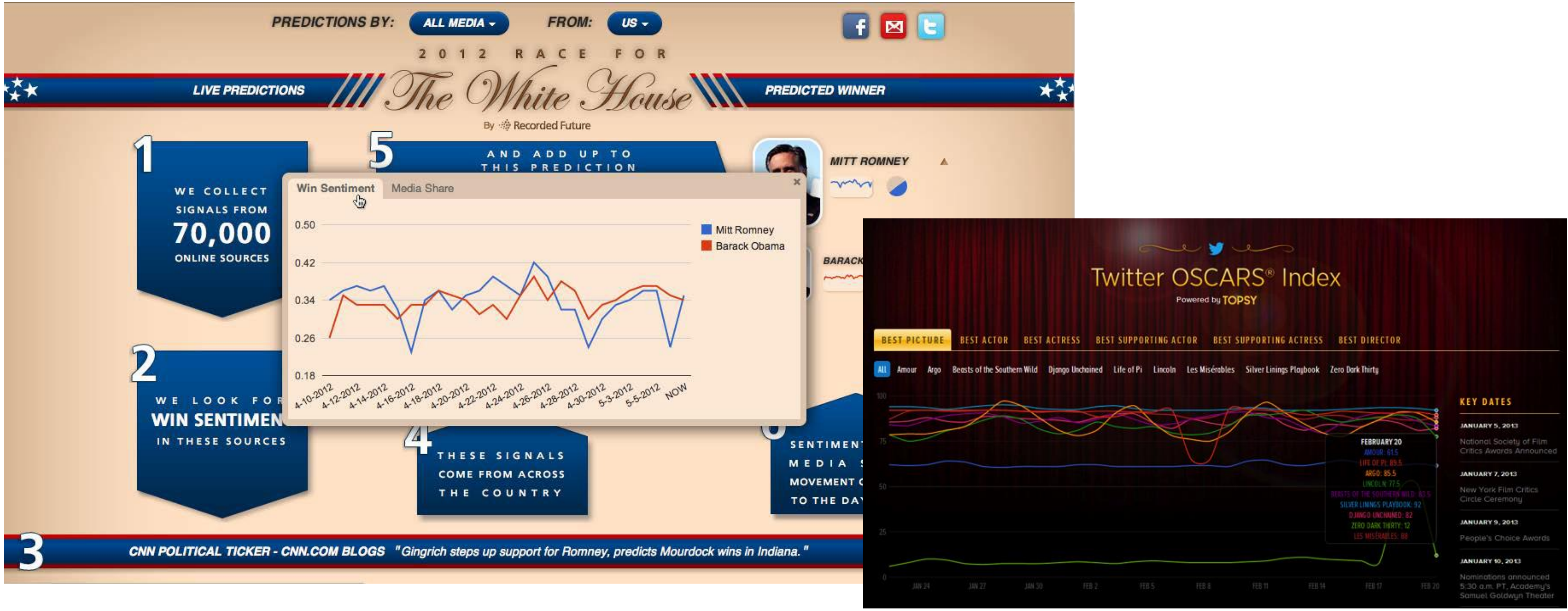
TEXT MINING

- Types of text mining
  - **NLP**
    - Low level language processing and understanding tasks often used synonymously with computational linguistics
  - **Concept extraction**
    - Grouping of words and phrases into semantically similar groups

TEXT MINING

- Text mining around us

Sentiment Analysis



1 WE COLLECT SIGNALS FROM 70,000 ONLINE SOURCES

2 WE LOOK FOR WIN SENTIMEN IN THESE SOURCES

3 CNN POLITICAL TICKER - CNN.COM BLOGS "Gingrich steps up support for Romney, predicts Mourdock wins in Indiana."

4 THESE SIGNALS COME FROM ACROSS THE COUNTRY

5 AND ADD UP TO THIS PREDICTION

6 SENTIMENT MEDIA S MOVEMENT C TO THE DAY

Twitter OSCARS® Index

Powered by TOPSY

FEBRUARY 20
AMOUR: 61.5
LIFE OF PI: 89.5
ARGO: 85.5
LINCOLN: 77.5
BEASTS OF THE SOUTHERN WILD: 83.5
SILVER LININGS PLAYBOOK: 92
DJANGO UNCHAINED: 82
ZERO DARK THIRTY: 12
LES MISÉRABLES: 88

KEY DATES

JANUARY 5, 2013
National Society of Film Critics Awards Announced

JANUARY 7, 2013
New York Film Critics Circle Ceremony

JANUARY 9, 2013
People's Choice Awards

JANUARY 10, 2013
Nominations announced 5:30 a.m. PT, Academy's Samuel Goldwyn Theater

TEXT MINING

- Text mining around us

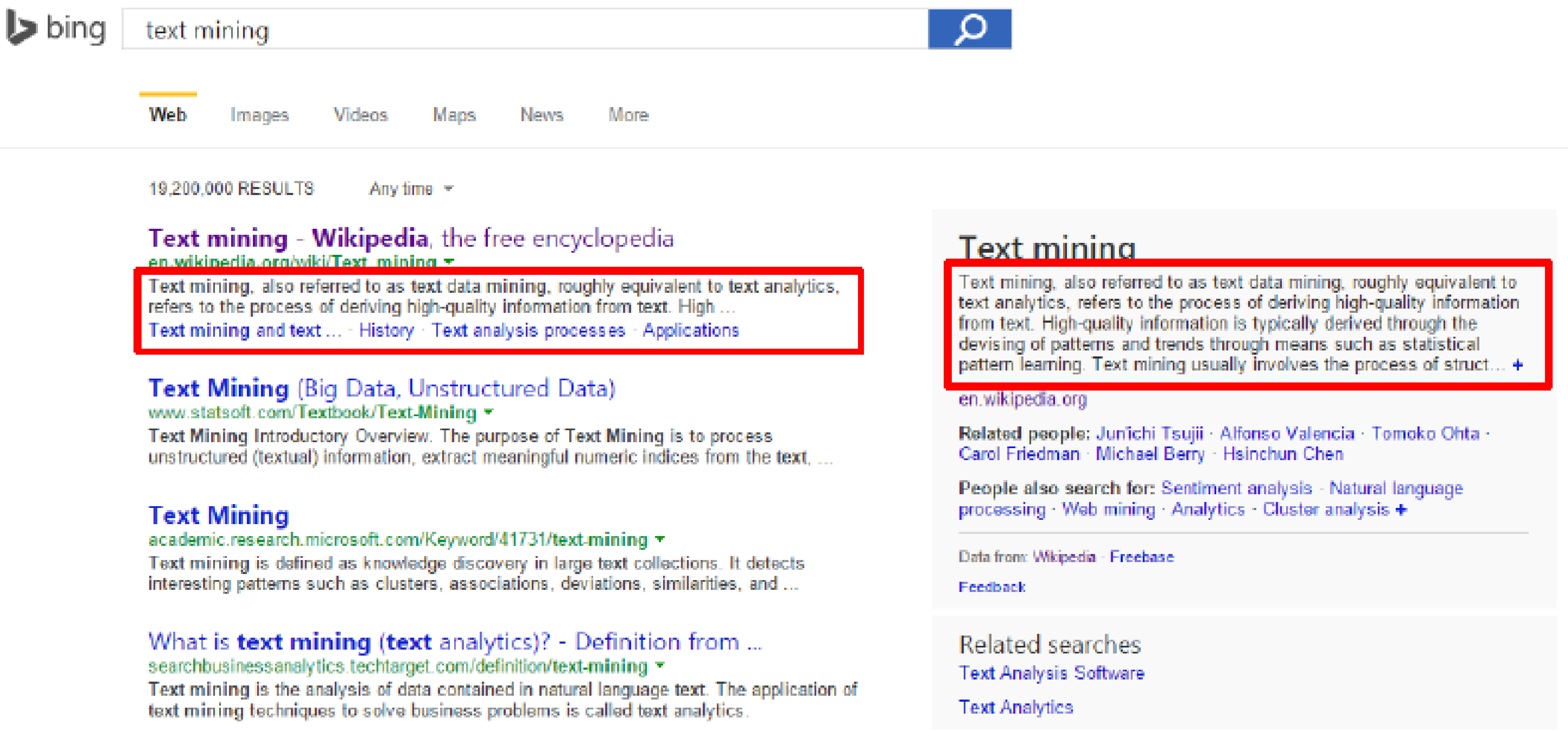
Document Summarization

identity
desired
allows
aspects
operations
strategy
used
body
Change
governance
increase
ensuring
together
manage
transformation
coherent
terms
approach
using
infrastructure
support
operating
software
changes
design
defining
effective
efficiently
set
supporting
informed
ensure
environment
best
organization
evolution
meet
decision
enables
get
tools
value
Strategic
execution
framework
reduce
stakeholders
processes
organizations
future
needs
technical
transform
practice
realize
knowledge
technologies
structure
people
customer
investment
architecture
EA
capabilities
help
change
management
enable
discipline
technology
align
delivery
effectively
planning
view
state
integration
mind
mission
resources
information
execute
system
time
data
art
way
development
cost
principles
context
complexity
services
blueprint
efficient
means
Planning
new
process
across
model
strategic
systems
organisation
aligning
alignment
objectives
requirements
building
corporate
solution
maintain
organization's
holistic
better
helps
efficiency
models
evolving
improve
achieve
vision
goals
decisions
right
structural
one
agility
opportunities
practices
organizational
Architecture
make
Agility
growth
strategies
plan
company
current
standards
functional
Efficiency
components
create
enabling
making
order
analysis
assets
provides

TEXT MINING

- Text mining around us

Document Summarization



19,200,000 RESULTS Any time

Text mining - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Text_mining
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High …
Text mining and text … · History · Text analysis processes · Applications

Text Mining (Big Data, Unstructured Data)
www.statsoft.com/Textbook/Text-Mining
Text Mining Introductory Overview. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, …

Text Mining
academic.research.microsoft.com/Keyword/41731/text-mining
Text mining is defined as knowledge discovery in large text collections. It detects interesting patterns such as clusters, associations, deviations, similarities, and …

What is **text mining** (**text** analytics)? - Definition from …
searchbusinessanalytics.techtarget.com/definition/text-mining
Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics.

Text mining

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of struct… +

en.wikipedia.org

Related people: Jun'ichi Tsujii · Alfonso Valencia · Tomoko Ohta · Carol Friedman · Michael Berry · Hsinchun Chen

People also search for: Sentiment analysis · Natural language processing · Web mining · Analytics · Cluster analysis +

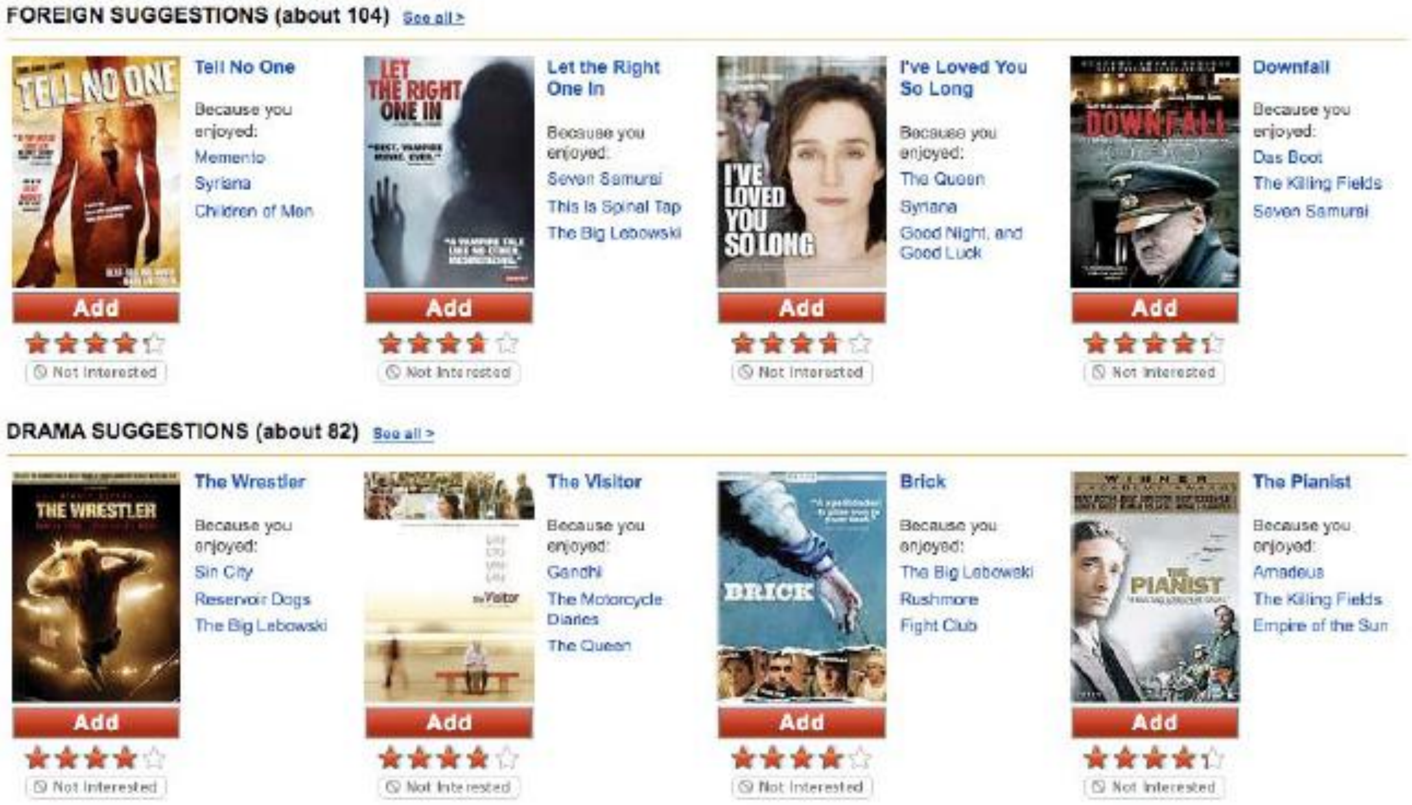
Data from: Wikipedia · Freebase

Feedback

Related searches
Text Analysis Software
Text Analytics

- Text mining around us

Movie Recommendation



FOREIGN SUGGESTIONS (about 104) See all >

Tell No One
Because you enjoyed:
Memento
Syriana
Children of Men
Add
Not Interested

Let the Right One In
Because you enjoyed:
Seven Samurai
This Is Spinal Tap
The Big Lebowski
Add
Not Interested

I've Loved You So Long
Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck
Add
Not Interested

Downfall
Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai
Add
Not Interested

DRAMA SUGGESTIONS (about 82) See all >

The Wrestler
Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski
Add
Not Interested

The Visitor
Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen
Add
Not Interested

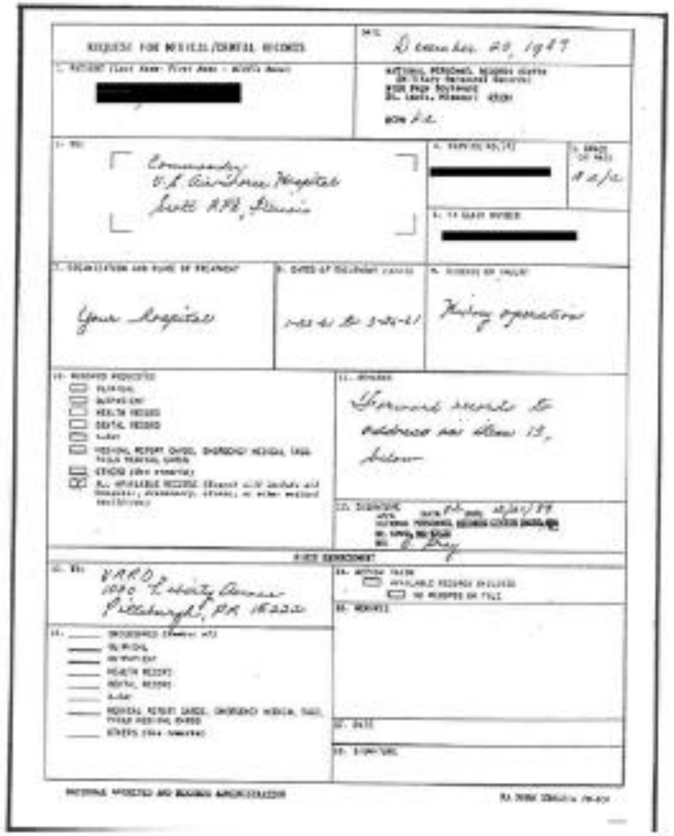
Brick
Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club
Add
Not Interested

The Pianist
Because you enjoyed:
Amadeus
The Killing Fields
Empire of the Sun
Add
Not Interested

TEXT MINING

- Text mining around us

Text mining in healthcare



REQUEST FOR MEDICAL/DENTAL RECORDS

DATE December 20, 1977

1. PATIENT (Last Name - First Name - Middle Name)
[redacted]

NATIONAL PERSONNEL RECORDS CENTER
(Military Personnel Records)
9700 Page Boulevard
St. Louis, Missouri 63132

3. TO:
Commander
U.S. Air Force Hospital
Scott AFB, Illinois

4. SERVICE NO.(S)
[redacted]

5. GRADE OR RATE
N/A

6. VA CLAIM NUMBER
[redacted]

7. ORGANIZATION AND PLACE OF TREATMENT
Your hospital

8. DATES OF TREATMENT
1-23-61 to 3-26-61

9. NATURE OF ILLNESS
Kidney operation

10. RECORDS REQUESTED
□ CLINICAL
□ OUTPATIENT
□ HEALTH RECORD
□ DENTAL RECORD
□ X-RAY
□ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
□ OTHERS (See remarks)
☒ ALL AVAILABLE RECORDS

11. REMARKS
Forward records to address in item 13, below

12. SIGNATURE

FIRST ENDORSEMENT

13. TO:
VARO
1000 Liberty Avenue
Pittsburgh, PA 15222

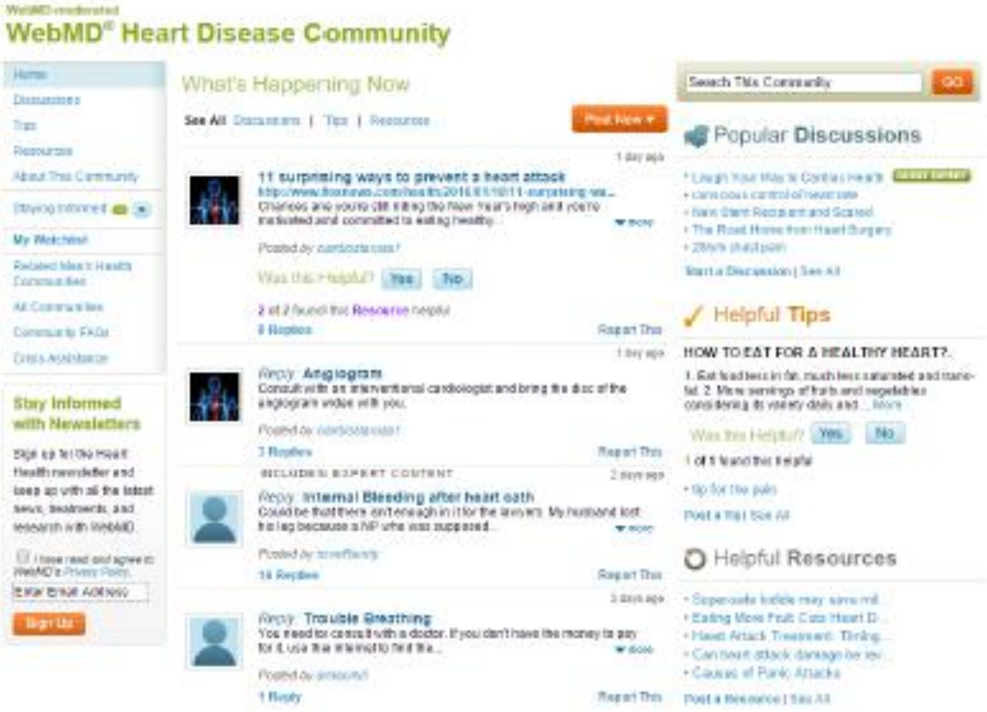
14. ACTION TAKEN
□ AVAILABLE RECORDS ENCLOSED
□ NO RECORDS ON FILE

15. REMARKS

17. DATE

18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION



WebMD-moderated

WebMD® Heart Disease Community

Home
Discussions
Tips
Resources
About This Community
Staying Informed
My WebMD
Related Men's Health Communities
All Communities
Community FAQs
Crisis Assistance

Stay Informed with Newsletters

Sign up for the Heart Health newsletter and keep up with all the latest news, treatments, and research with WebMD.

□ I have read and agree to WebMD's Privacy Policy.

Enter Email Address

Sign Up

What's Happening Now

See All Discussions | Tips | Resources

Post New

1 day ago

11 surprising ways to prevent a heart attack
http://www.foxnews.com/health/2016/01/01/11-surprising-wa...
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy...
more

Posted by cardiostaci

Was this Helpful? Yes No

2 of 2 found this Resource helpful

8 Replies

Report This

1 day ago

Reply: Angiogram
Consult with an interventional cardiologist and bring the disc of the angiogram video with you.

Posted by cardiostaci

3 Replies

Report This

INCLUDES EXPERT CONTENT

2 days ago

Reply: Internal Bleeding after heart cath
Could be that there isn't enough in it for the lawyers. My husband lost his leg because a NP who was supposed...
more

Posted by [illegible]

16 Replies

Report This

3 days ago

Reply: Trouble Breathing
You need to consult with a doctor. If you don't have the money to pay for it, use the internet to find the...
more

Posted by [illegible]

1 Reply

Report This

Search This Community GO

Popular Discussions

- Laugh Your Way to Cardiac Health
- Conscious control of heart rate
- New Stent Recipient and Scared
- The Road Home from Heart Surgery
- 28M chest pain

Start a Discussion | See All

Helpful Tips

HOW TO EAT FOR A HEALTHY HEART?

1. Eat food less in fat, much less saturated and trans-fat. 2. More servings of fruits and vegetables considering its variety daily and ... More

Was this Helpful? Yes No

1 of 1 found this helpful

- tip for the pain

Post a Tip | See All

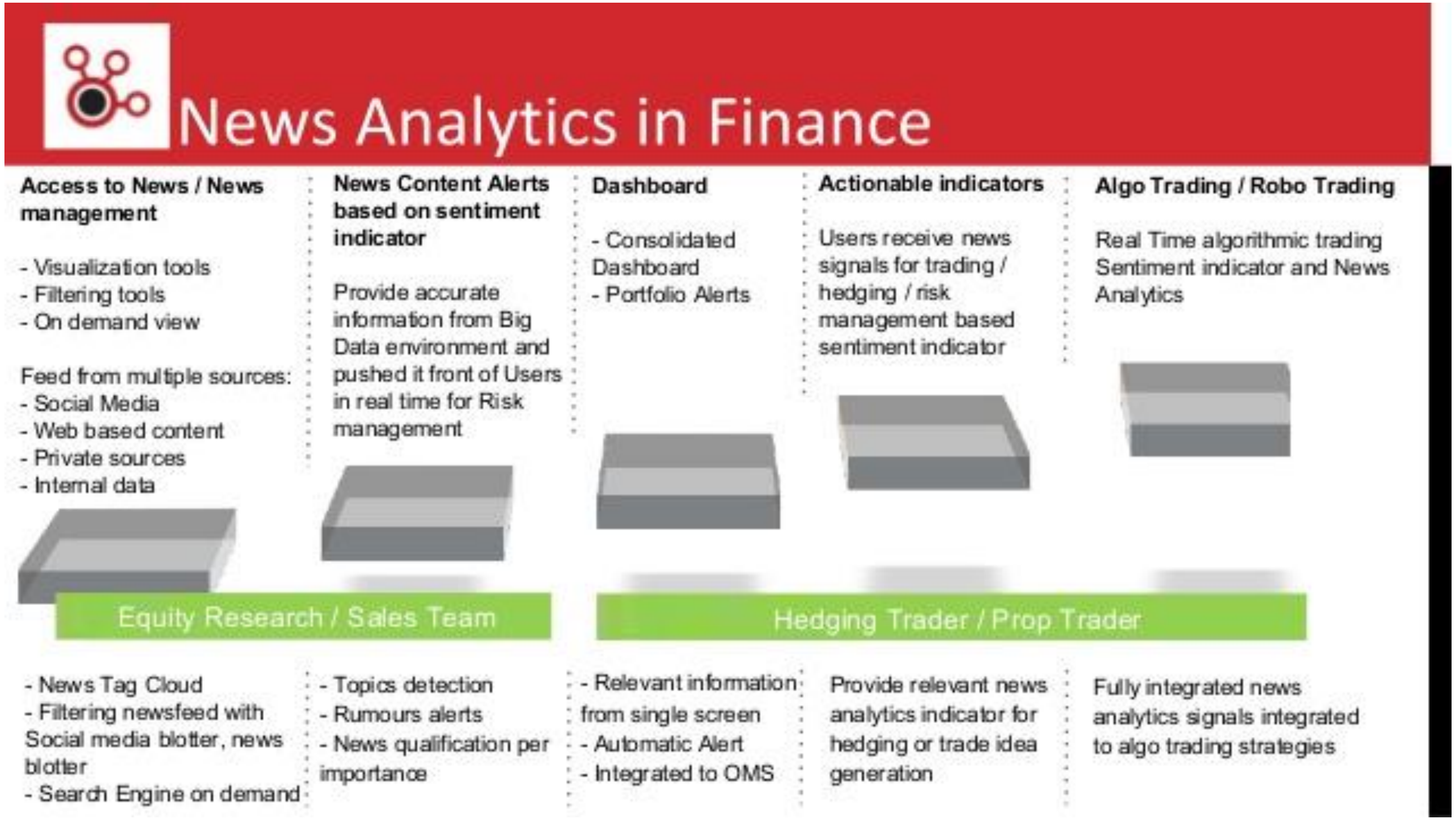
Helpful Resources

- Supersafe bubble may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be rev...
- Causes of Panic Attacks

Post a Resource | See All

- Text mining around us

Text mining in Finance



News Analytics in Finance

Access to News / News management

- Visualization tools
- Filtering tools
- On demand view

Feed from multiple sources:
- Social Media
- Web based content
- Private sources
- Internal data

News Content Alerts based on sentiment indicator

Provide accurate information from Big Data environment and pushed it front of Users in real time for Risk management

Dashboard

- Consolidated Dashboard
- Portfolio Alerts

Actionable indicators

Users receive news signals for trading / hedging / risk management based sentiment indicator

Algo Trading / Robo Trading

Real Time algorithmic trading Sentiment indicator and News Analytics

Equity Research / Sales Team

Hedging Trader / Prop Trader

- News Tag Cloud
- Filtering newsfeed with Social media blotter, news blotter
- Search Engine on demand

- Topics detection
- Rumours alerts
- News qualification per importance

- Relevant information from single screen
- Automatic Alert
- Integrated to OMS

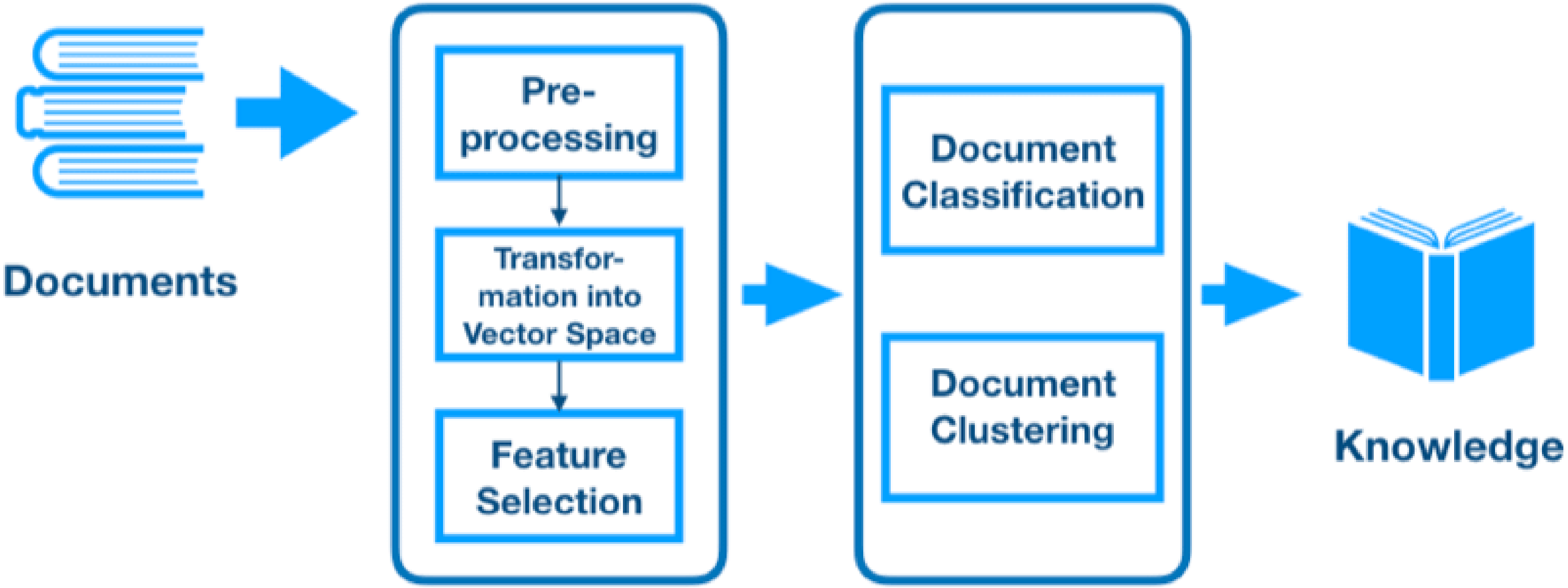
Provide relevant news analytics indicator for hedging or trade idea generation

Fully integrated news analytics signals integrated to algo trading strategies

TEXT MINING

Text Mining Process Flow

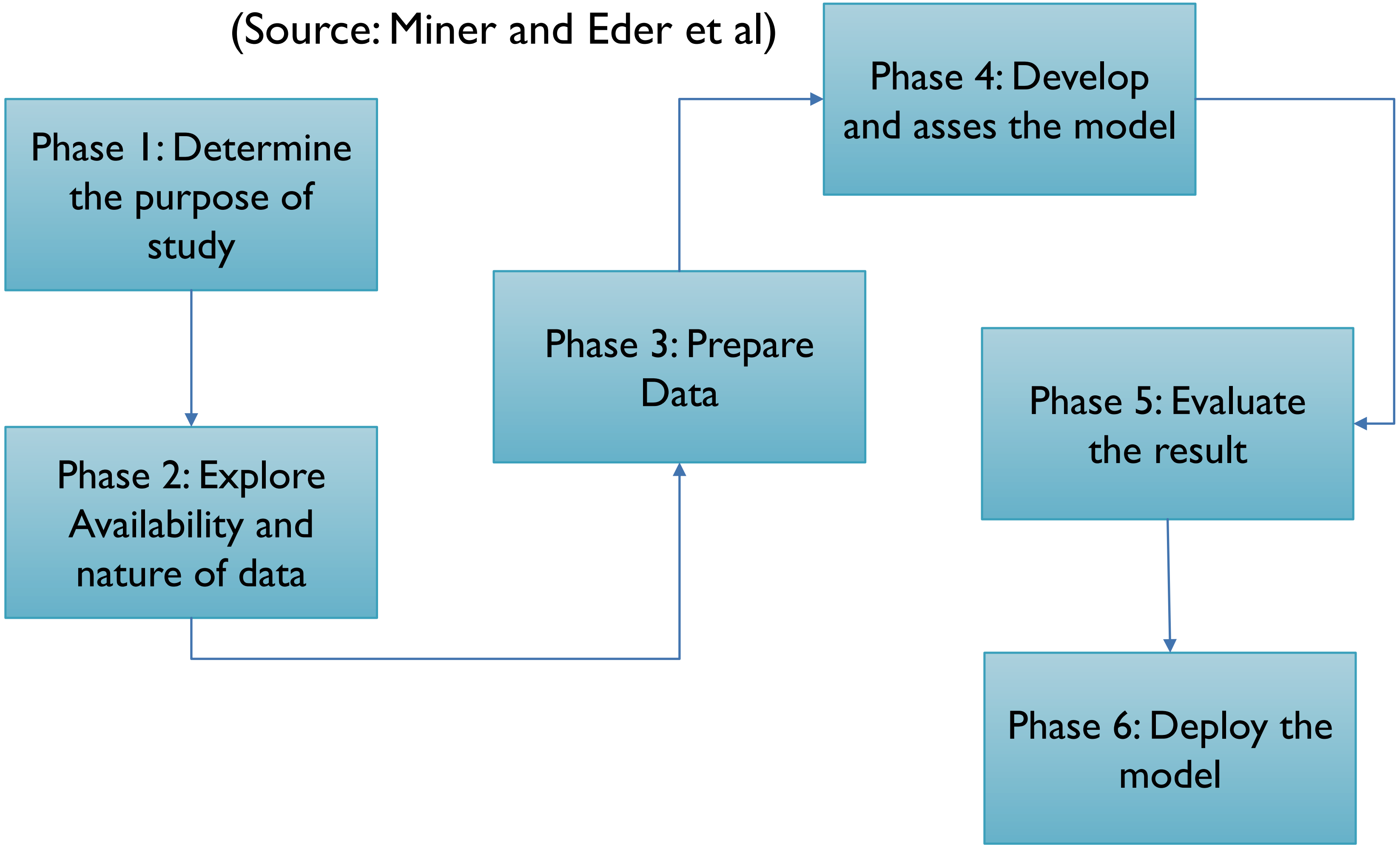
(Source: Claudia Peersman)



TEXT MINING

Text Mining Process Flow

(Source: Miner and Eder et al)


Phase 1: Determine the purpose of study

Phase 2: Explore Availability and nature of data

Phase 3: Prepare Data

Phase 4: Develop and asses the model

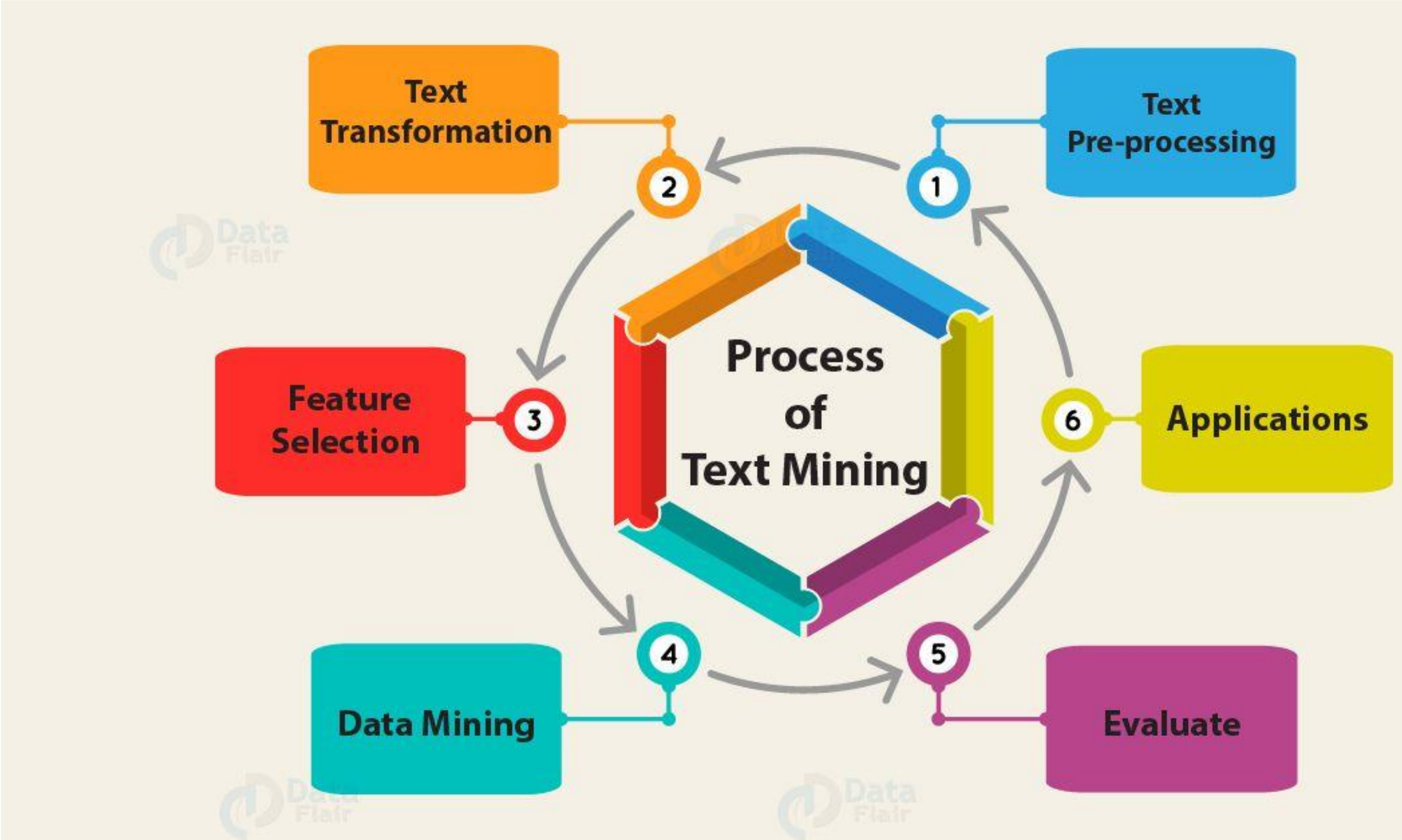
Phase 5: Evaluate the result

Phase 6: Deploy the model

TEXT MINING

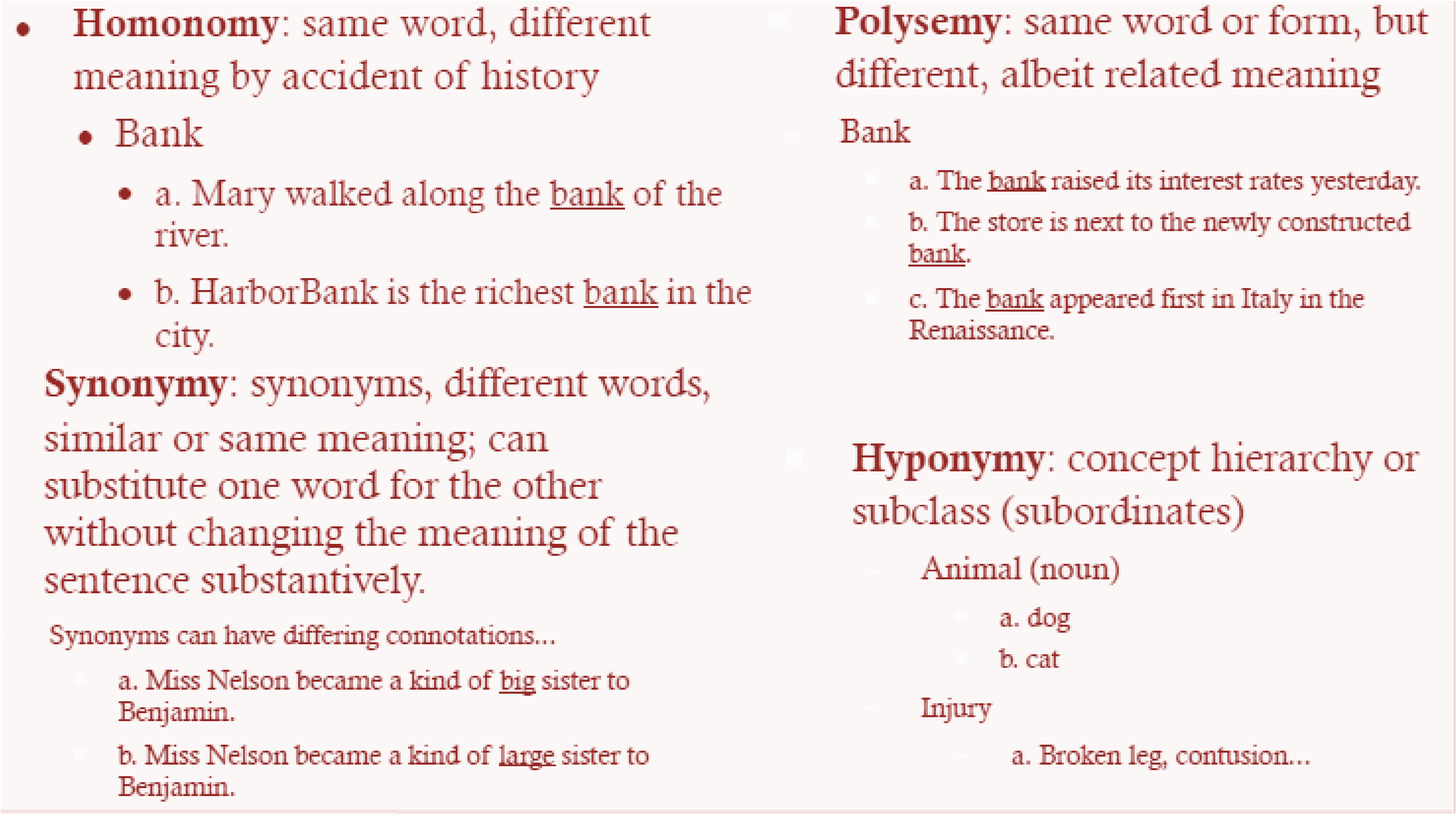
- Text Mining Process Flow

(Source: Data Flair)



TEXT MINING

- Text Mining is not easy
  - Language is ambiguous



- **Homonomy**: same word, different meaning by accident of history
  - Bank
    - a. Mary walked along the bank of the river.
    - b. HarborBank is the richest bank in the city.

Synonymy: synonyms, different words, similar or same meaning; can substitute one word for the other without changing the meaning of the sentence substantively.

Synonyms can have differing connotations…

- a. Miss Nelson became a kind of big sister to Benjamin.
- b. Miss Nelson became a kind of large sister to Benjamin.

Polysemy: same word or form, but different, albeit related meaning

Bank

- a. The bank raised its interest rates yesterday.
- b. The store is next to the newly constructed bank.
- c. The bank appeared first in Italy in the Renaissance.

Hyponymy: concept hierarchy or subclass (subordinates)

Animal (noun)

- a. dog
- b. cat

Injury

- a. Broken leg, contusion…

TEXT MINING

- Concepts and word extraction usually results in a huge dimension
  - Thousands of new fields
  - Each field typically has low information
- Misspelling abbreviations, spelling variants

TEXT MINING

Label	Challenges
Words and morphemes	Word segmentation: dividing text into words. Fairly easy for English and other languages that use whitespace; much harder for languages like Chinese and Japanese. – Assigning part of speech. – Identifying synonyms; synonyms are useful for searching. – Stemming: the process of shortening a word to its base or root form. For example, a simple stemming of *words* is *word*. – Abbreviations, acronyms, and spelling also play important roles in understanding words.

TEXT MINING

Multiword and sentence	Phrase detection: *quick red fox*, *hockey legend Bobby Orr*, and *big brown shoe* are all examples of phrases. – Parsing: breaking sentences down into subject-verb and other relationships often yields useful information about words and their relationships to each other. – Sentence boundary detection is a well-understood problem in English, but is still not perfect. – Coreference resolution: "Jason likes dogs, but he would never buy one." In this example, *he* is a coreference to Jason. The need for coreference resolution can also span sentences.

TEXT MINING

Multiword and sentence	– Words often have multiple meanings; using the context of a sentence or more may help choose the correct word. This process is called *word sense disambiguation* and is difficult to do well. – Combining the definitions of words and their relationships to each other to determine the meaning of a sentence
Multisentence and paragraph	At this level, processing becomes more difficult in an effort to find deeper understanding of an author’s intent. Algorithms for summarization often require being able to identify which sentences are more important than others.

OUTLINE

Introduction to Machine Learning

LEARNING – HOW IT IS DONE?

- Mr. Micky found unknown food
- He takes the first bite

Scenario 1: Learning

- Mr. Micky experience poisonous features in food

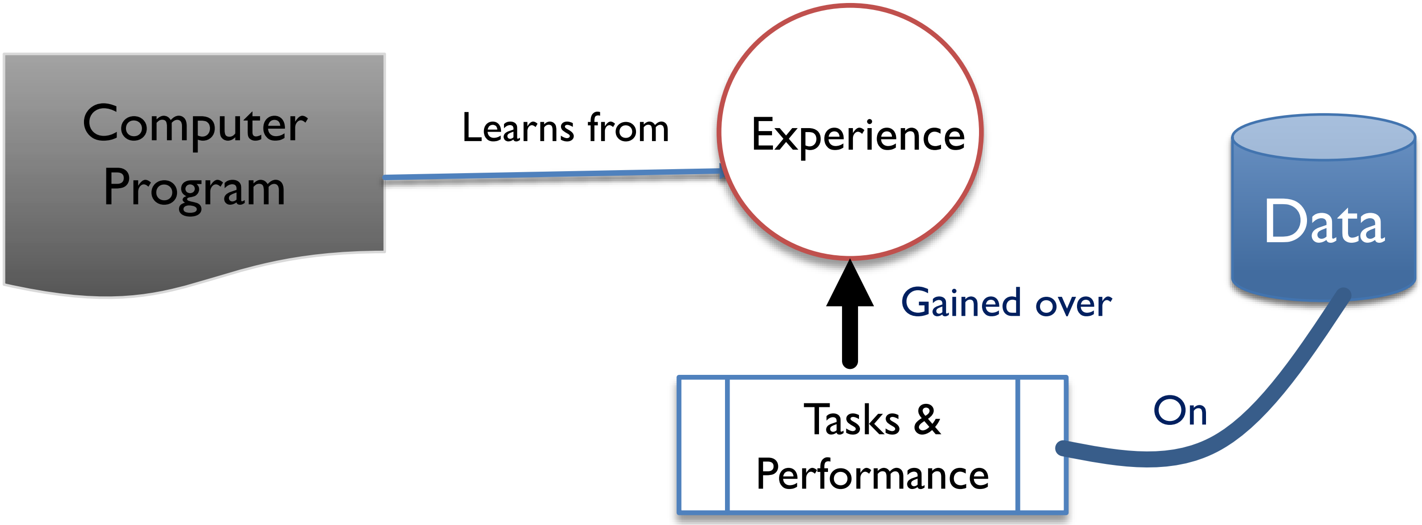
- Mr. Micky experience no poisonous features in food

- Mr. Micky eats known food without second thought
- He has knowledge (through experience) of the features of the food

Scenario 2: How to

HOW MACHINE LEARNS

- Machine learns through **computer programs** (the language of machine) which learns of from **experience**.

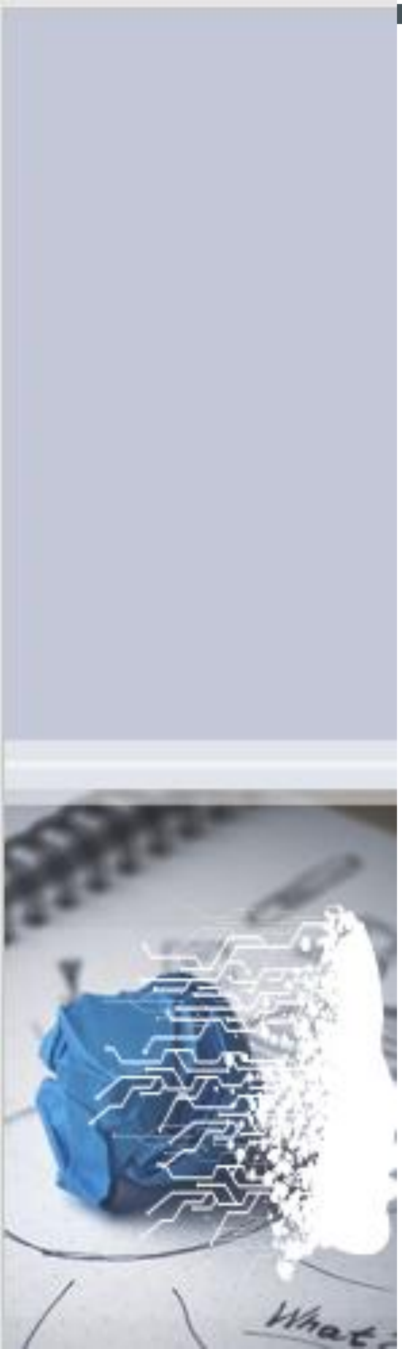


- Machine learns from **changes** in computer program or **data** (input); e.g., samples of speech from different person improve speech recognition.

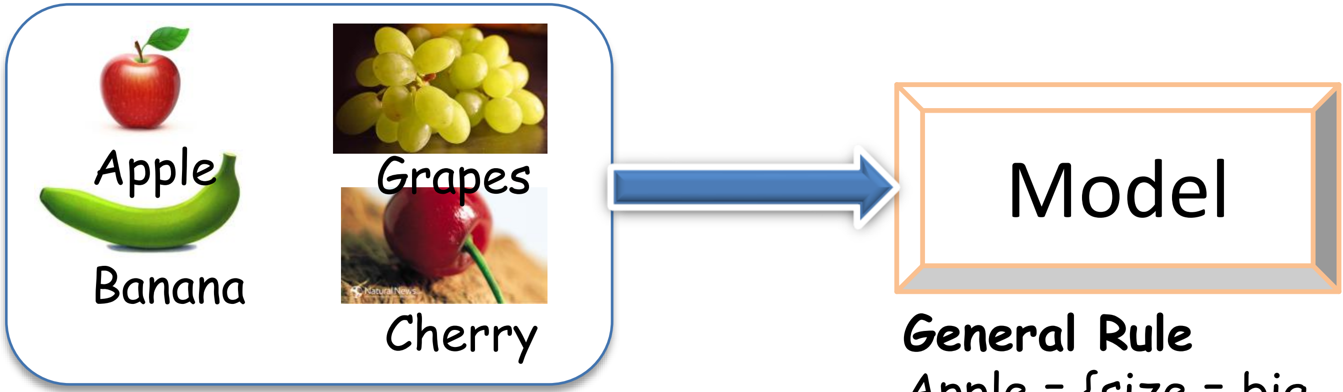
MACHINE LEARNING APPROACHES

- Definitions of Machine Learning
  - Machine learning is a method that gives computer the ability to learn **without** being explicitly programmed (Arthur, S.)
  - Machine learning is a set of **tools** that, broadly speaking, allow to "**teach**" computers how to perform tasks by providing examples of how they should be done (Mitchel, T.)
  - Machine learning is a method of data analysis that **automates** analytical model building (SAS)

LEARNING METHODS



- Learning methods of machines
  - Inductive learning: It involves the process of learning by example – where a system tries to induce a model (general rule/function) from training data.



Apple

Grapes

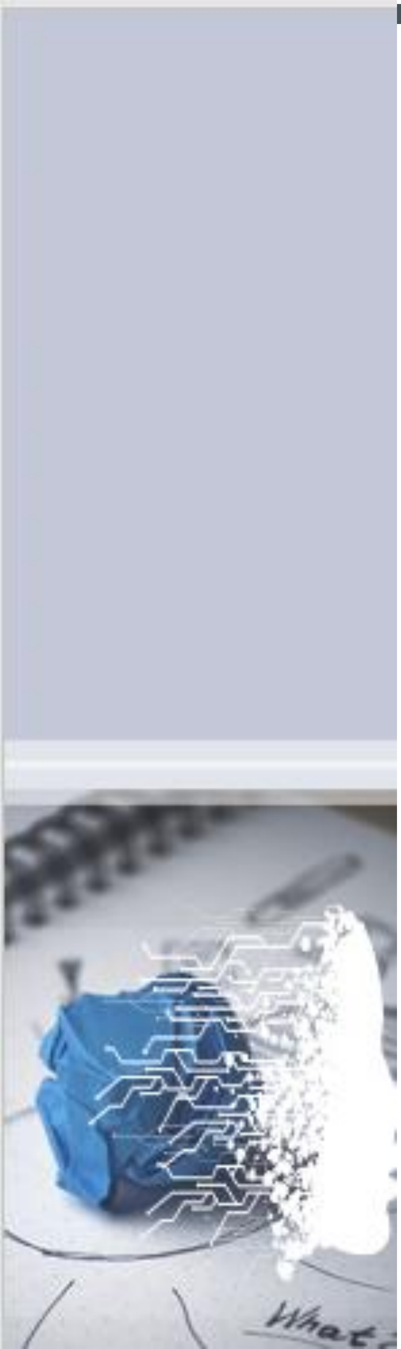
Banana

Cherry

Model

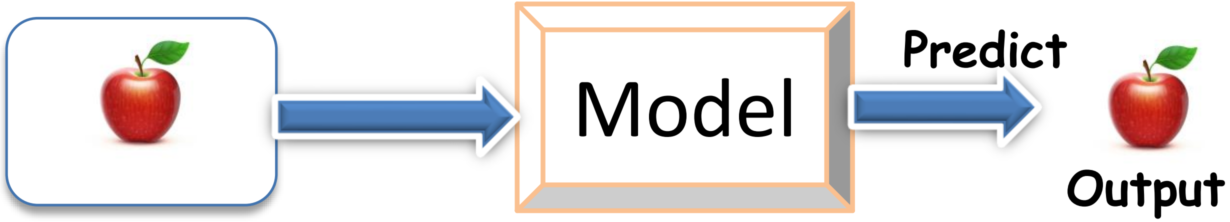
Training Data

General Rule
Apple = {size = big, shape rounded, color = red}

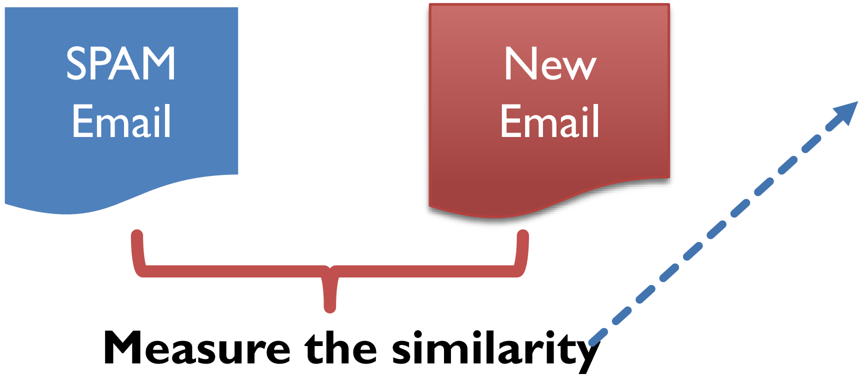


LEARNING METHODS

- Deductive learning: It applies the learning model to predict the outcome.

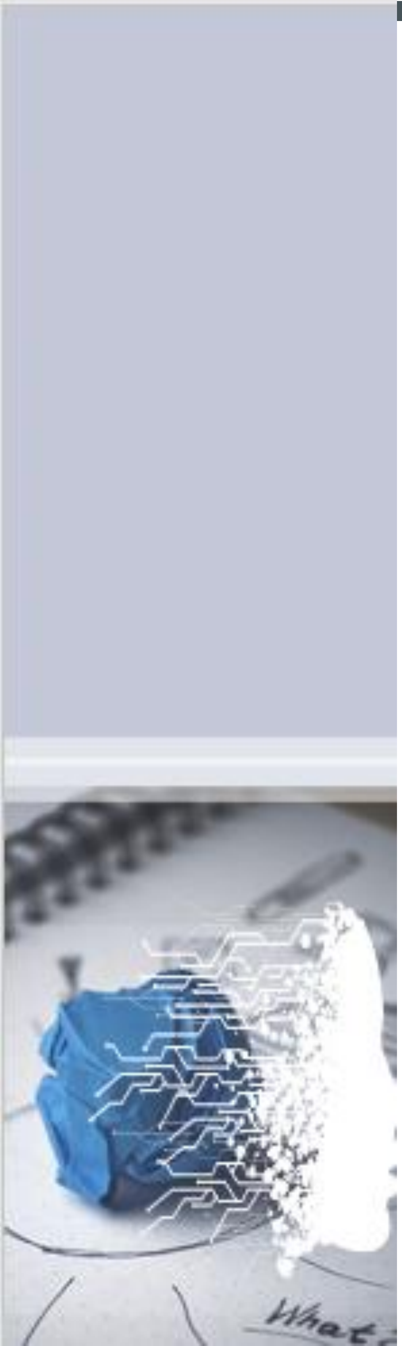


- Instance-based learning
  - The most trivial form of learning
  - The system learns the examples by heart, then generalizes to new cases by using a similarity measure to compare them to the learned examples (or a subset of them)



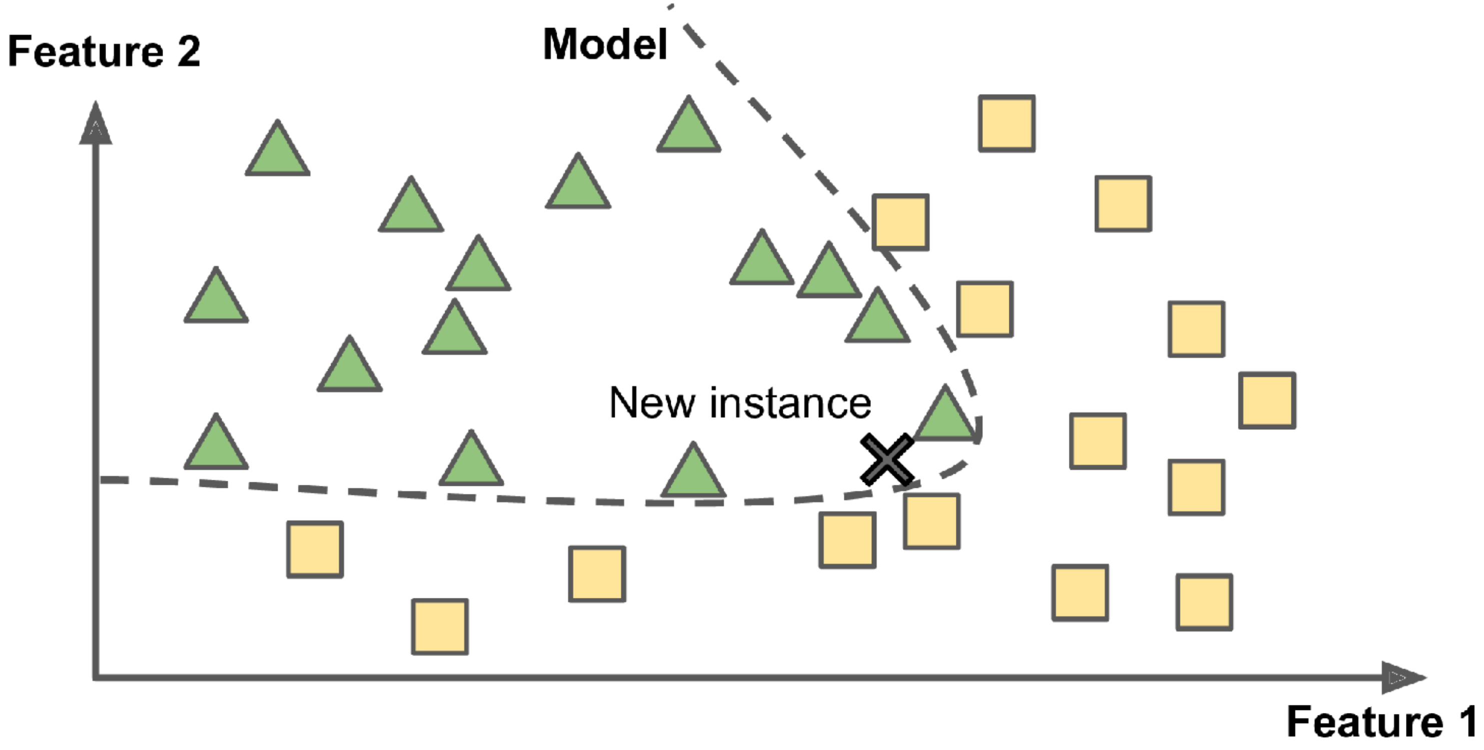
- Find common words between two mails
- If number of common words is significant then flag the new email as Spam

LEARNING METHODS

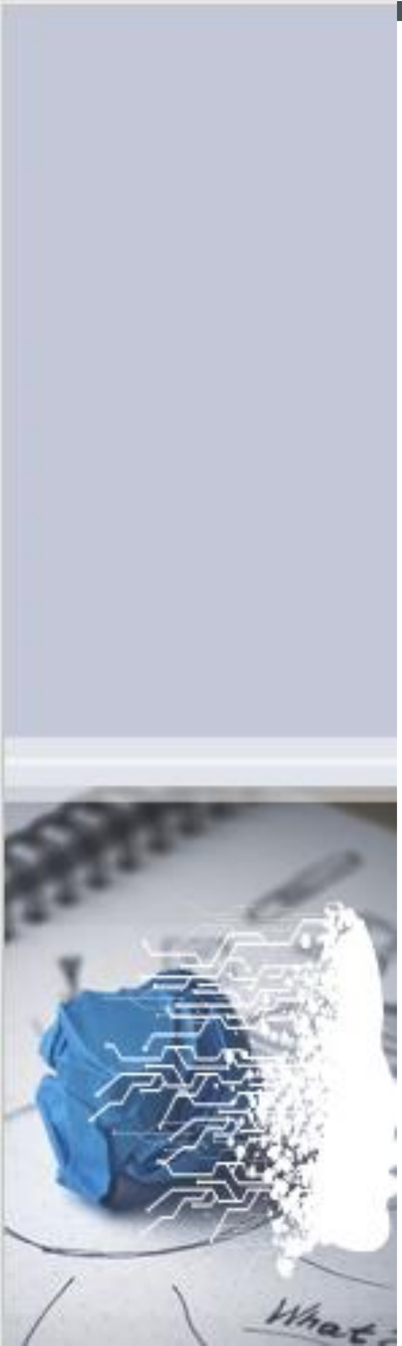


Model-based learning

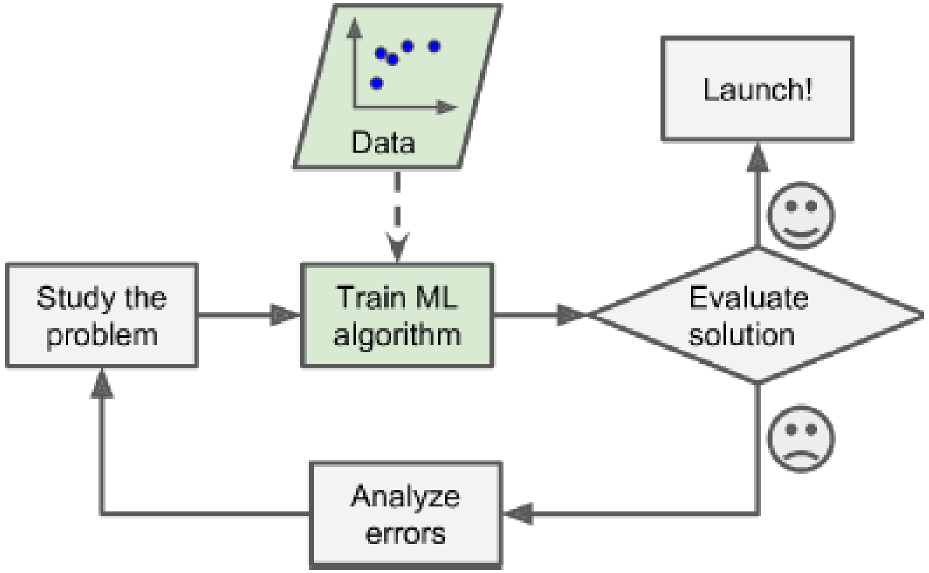
- Generalize from a set of examples is to build a model of these examples and then use that model to make predictions



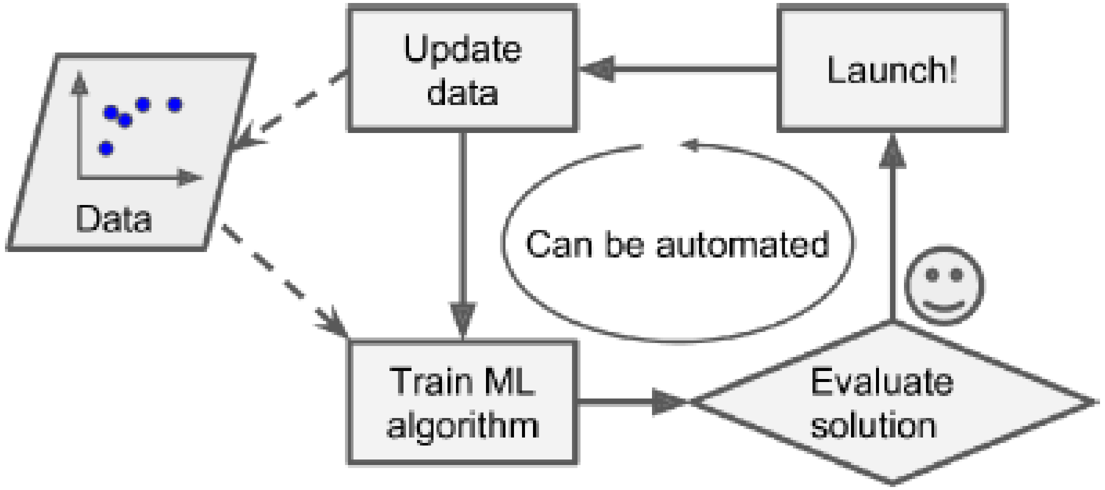
MACHINE LEARNING APPROACHES



- There are different approaches for machine learning systems
  - **Semi Automatic**
  - **Fully Automatic**



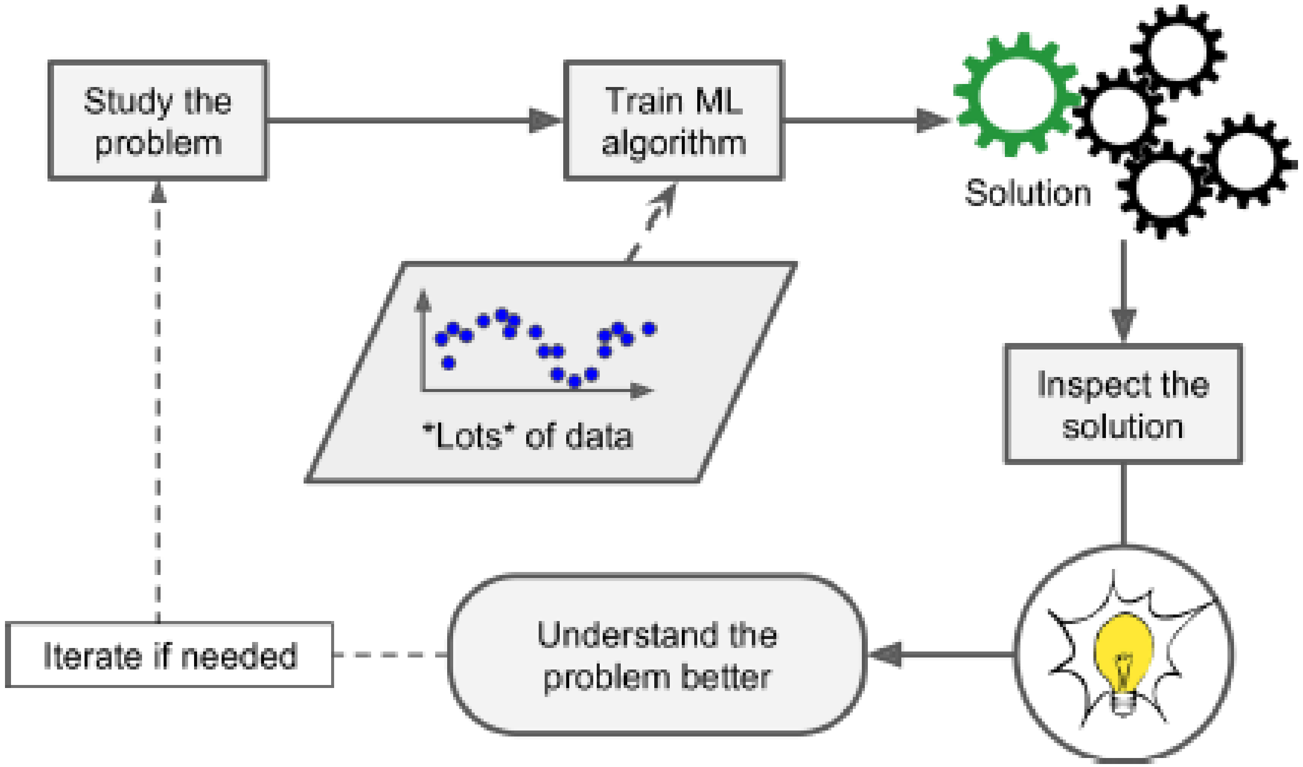
Semi-automatic system approach



Automatic system approach

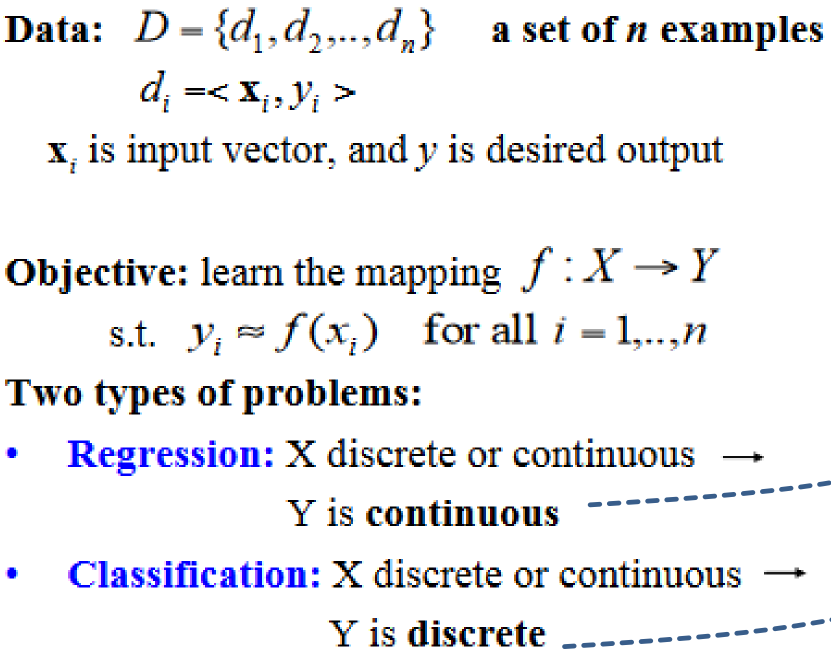
MACHINE LEARNING APPROACHES

- Machine learning helps human to learn



TYPES OF MACHINE LEARNING SYSTEM

- **Supervised Learning**: In supervised learning, the training set you feed to the algorithm includes the desired solutions, called **labels**



Data: $D = \{d_1, d_2, \ldots, d_n\}$ **a set of** n **examples**

$d_i = < \mathbf{x}_i, y_i >$

$\mathbf{x}_i$ is input vector, and y is desired output

Objective: learn the mapping $f : X \rightarrow Y$

s.t. $y_i \approx f(x_i)$ for all $i = 1,..,n$

Two types of problems:

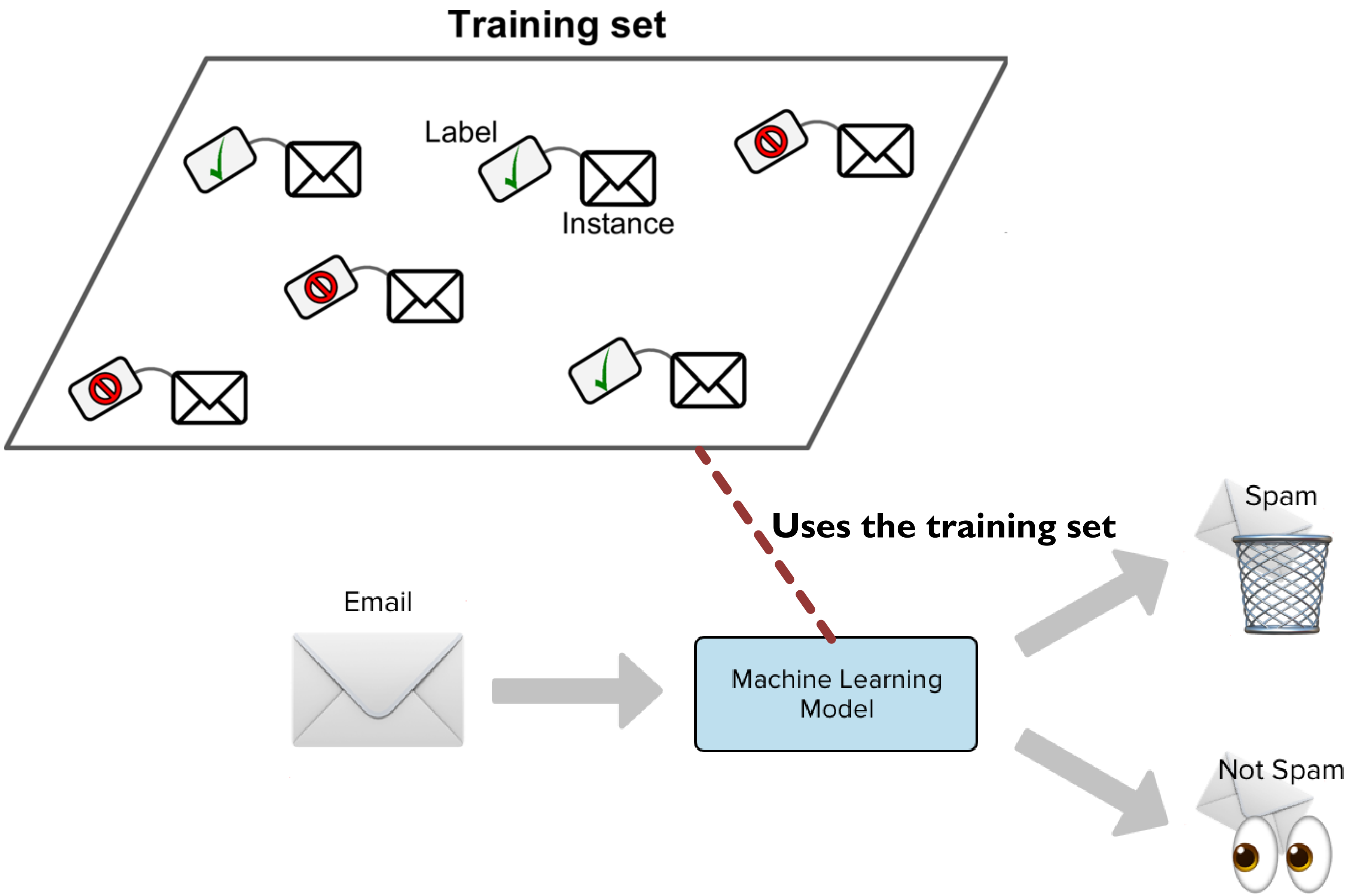
- **Regression:** X discrete or continuous → Y is **continuous**
- **Classification:** X discrete or continuous → Y is **discrete**

Ex: Price of a car, given a set of features

Ex: Classifying emails along with their class spam or Not Spam

- **Note:** Some regression algorithms can be used for classification as well, and vice versa.
  - For example, Logistic Regression is commonly used for classification, as it can output a value that corresponds to the probability of belonging to a given class

TYPES OF MACHINE LEARNING SYSTEM



TYPES OF MACHINE LEARNING SYSTEM

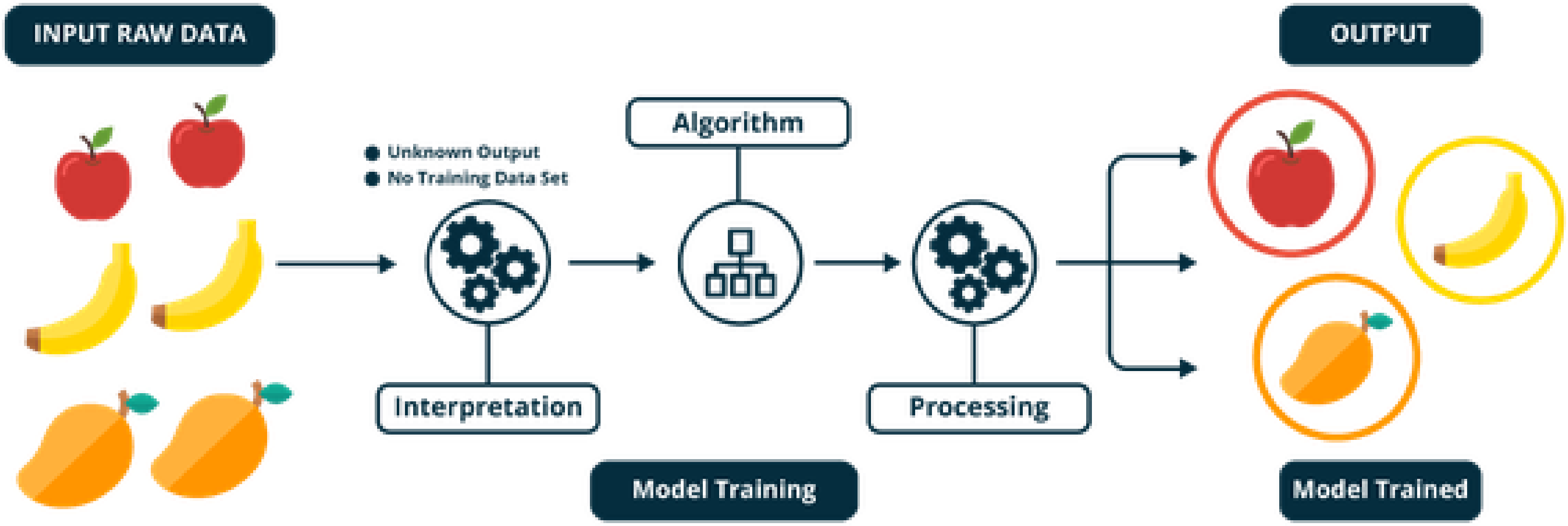
- The most widely used Supervised Learning Algorithms
  - k-Nearest Neighbors
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVMs)
  - Decision Trees and Random Forests
  - Neural networks

TYPES OF MACHINE LEARNING SYSTEM

- **Unsupervised Learning**
  - It is a type of learning where algorithm used to draw inferences from datasets consisting of input data (x) without labeled responses.
  - Goal:
    - In some pattern recognition problems, the training data consists of a set of input vectors x without any corresponding target values.
    - The goal in such unsupervised learning problems may be to discover groups of similar examples within the data.

TYPES OF MACHINE LEARNING SYSTEM

- **Unsupervised Learning**
  - Training data provides “example”, but we have no specific outcomes.
  - In simple word **there is no label associated with this learning.**
  - In unsupervised learning the machine tries to find interesting patterns in the data.

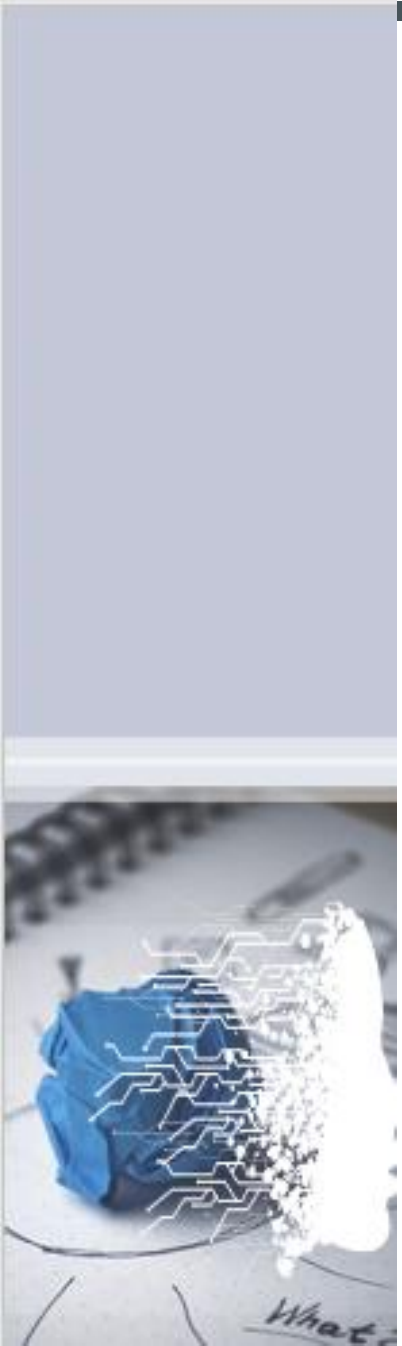


TYPES OF MACHINE LEARNING SYSTEM

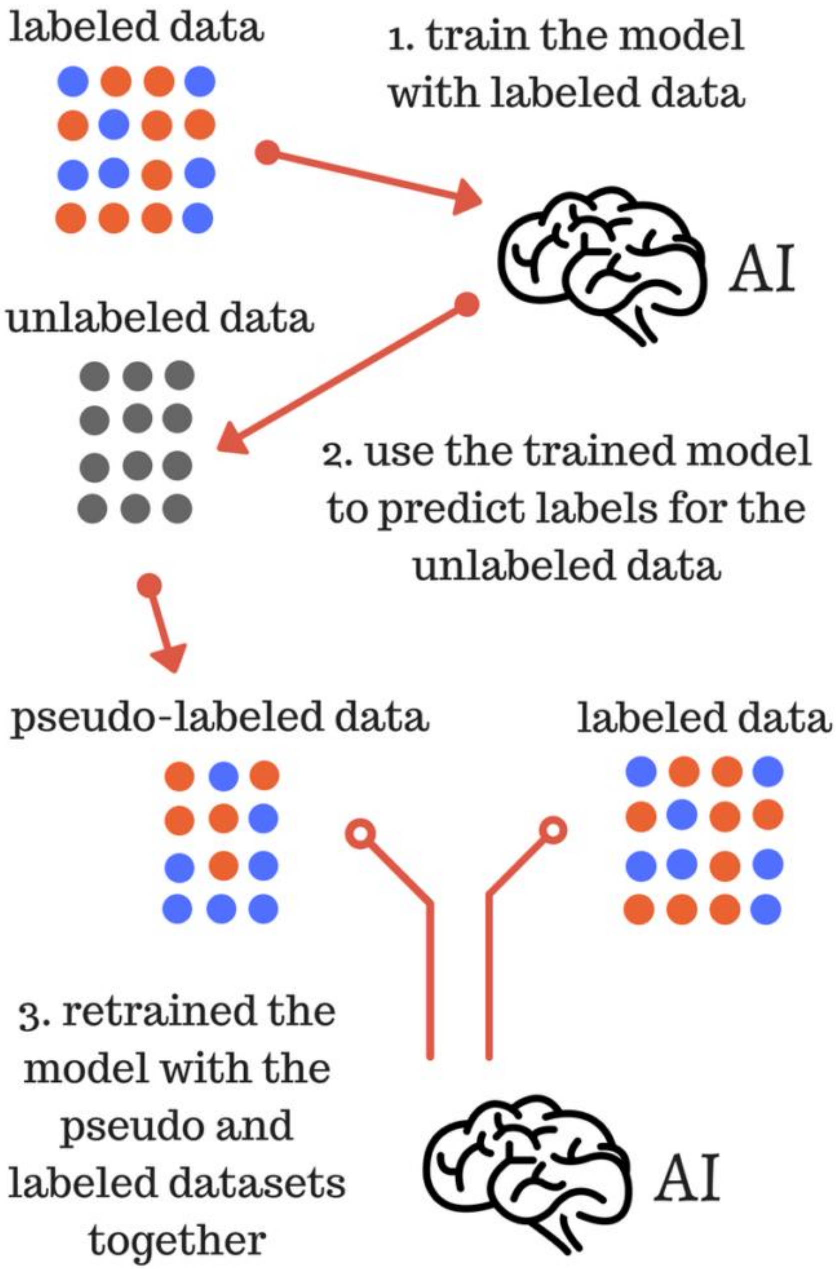
- **Unsupervised Algorithms**

Unsupervised Algorithms	
Clustering	—K-Means —DBSCAN —Hierarchical Cluster Analysis (HCA)
Anomaly detection and novelty detection	—One-class SVM —Isolation Forest
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 SYSTEM



- **Semi-supervised Learning**
  - Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning
  - In addition to unlabelled data, the algorithm is provided with some super-vision information – but not necessarily for all examples



labeled data

1. train the model with labeled data

unlabeled data

2. use the trained model to predict labels for the unlabeled data

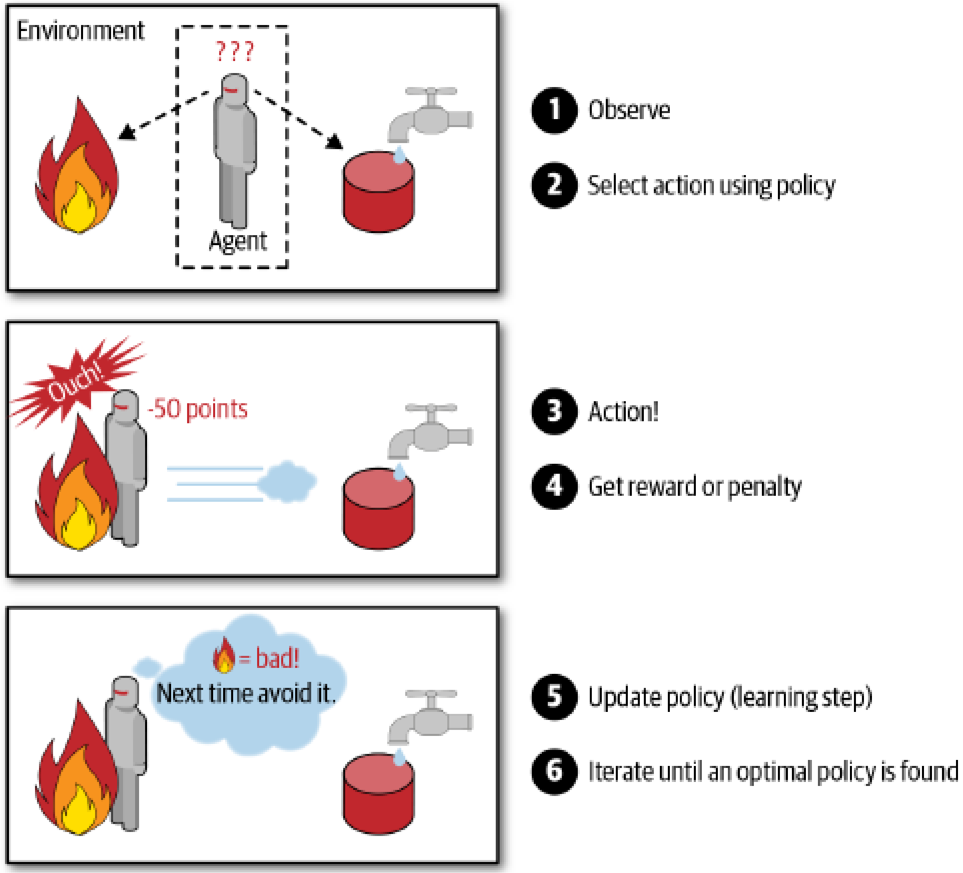
pseudo-labeled data

labeled data

3. retrained the model with the pseudo and labeled datasets together

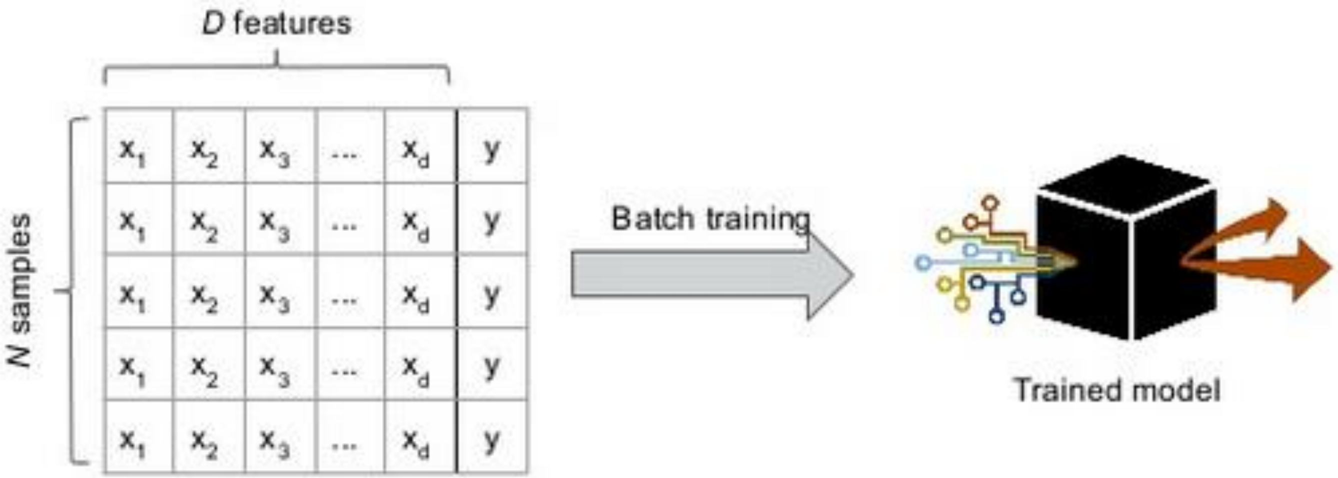
TYPES OF MACHINE LEARNING SYSTEM

- **Reinforcement Learning**
  - 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.


Environment
???
Agent
1 Observe
2 Select action using policy
Ouch!
-50 points
3 Action!
4 Get reward or penalty
= bad!
Next time avoid it.
5 Update policy (learning step)
6 Iterate until an optimal policy is found

DEPLOYMENT MODELS

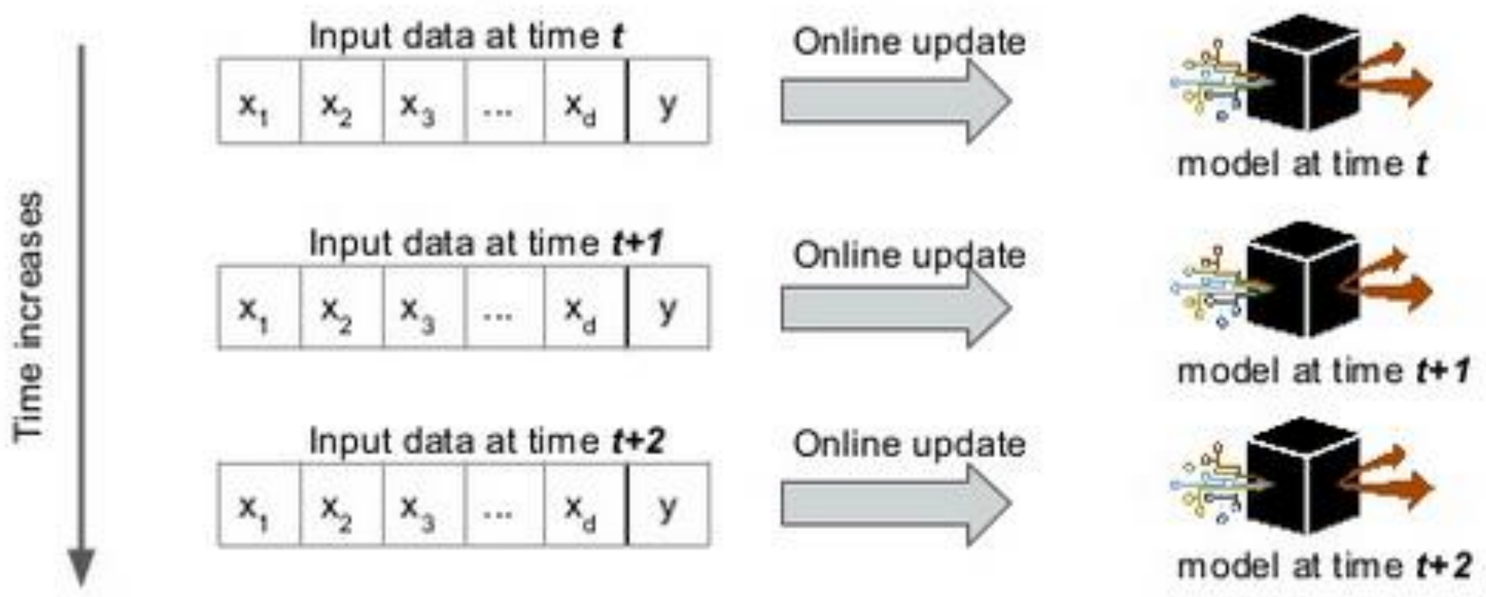
- **Batch Learning**
  - The system must be trained using all available data
  - The system is incapable of learning incrementally
  - For new data (such as a new type of spam),
    - Train a new version of the system from scratch on the full dataset (not just the new data, but also the old data)
    - Then stop the old system and replace it with the new one.



x_1	x_2	x_3	...	x_d	y
x_1	x_2	x_3	...	x_d	y
x_1	x_2	x_3	...	x_d	y
x_1	x_2	x_3	...	x_d	y
x_1	x_2	x_3	...	x_d	y

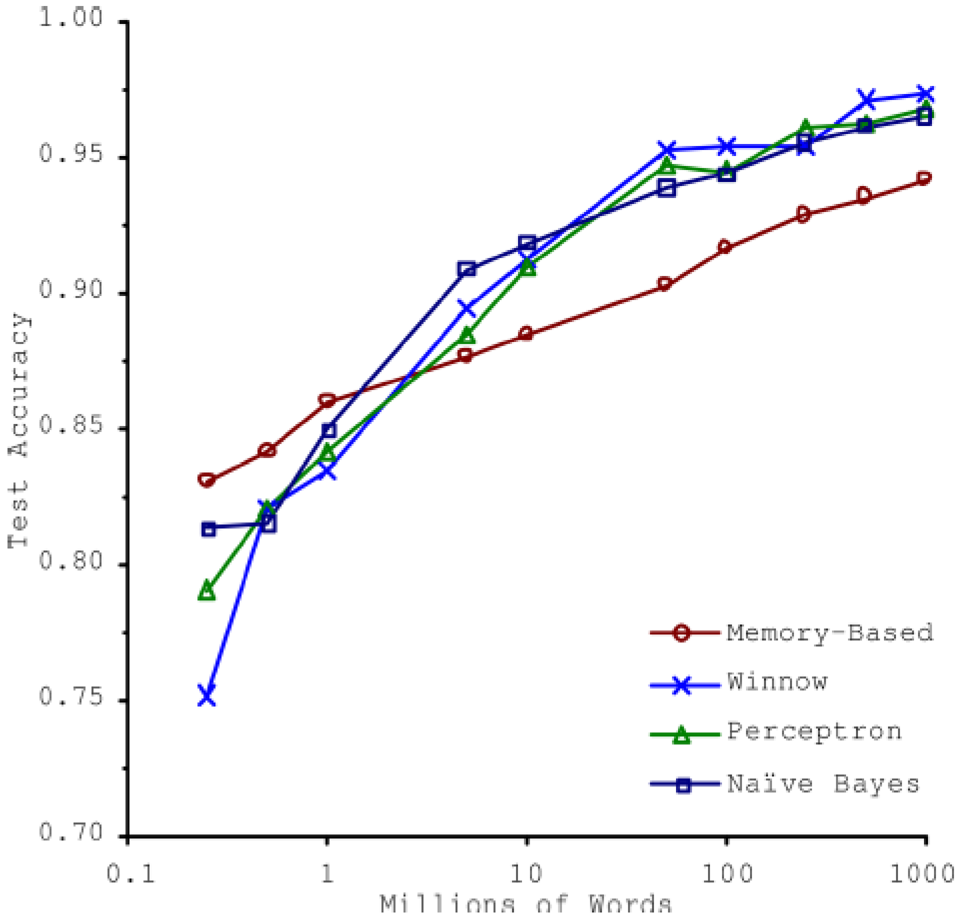
DEPLOYMENT MODELS

- **Online Learning**
  - Train the system incrementally by feeding it data instances sequentially, either individually or in 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



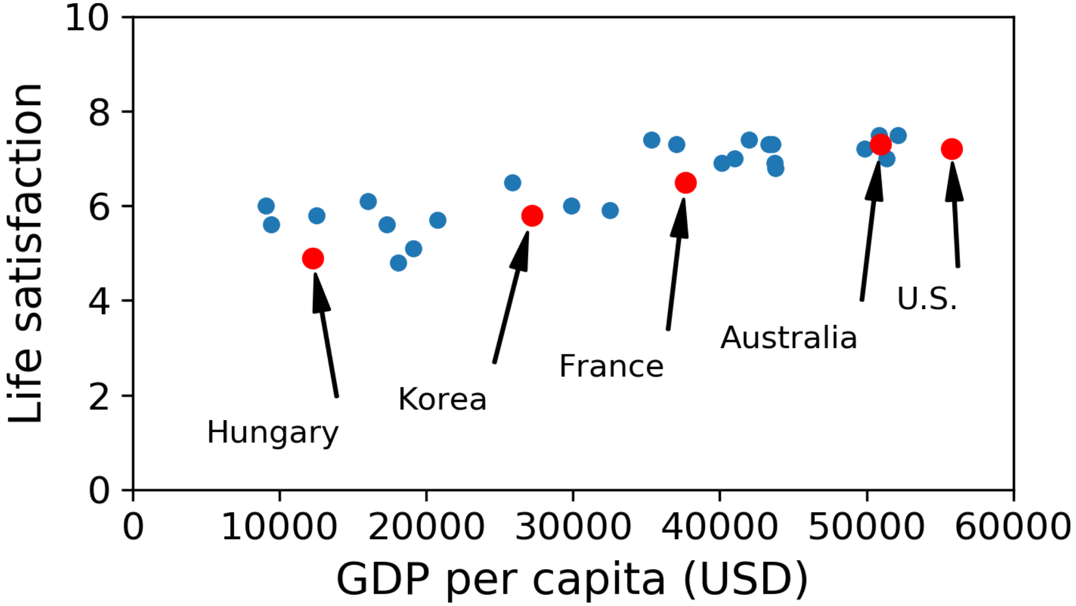
CHALLENGES OF MACHINE LEARNING

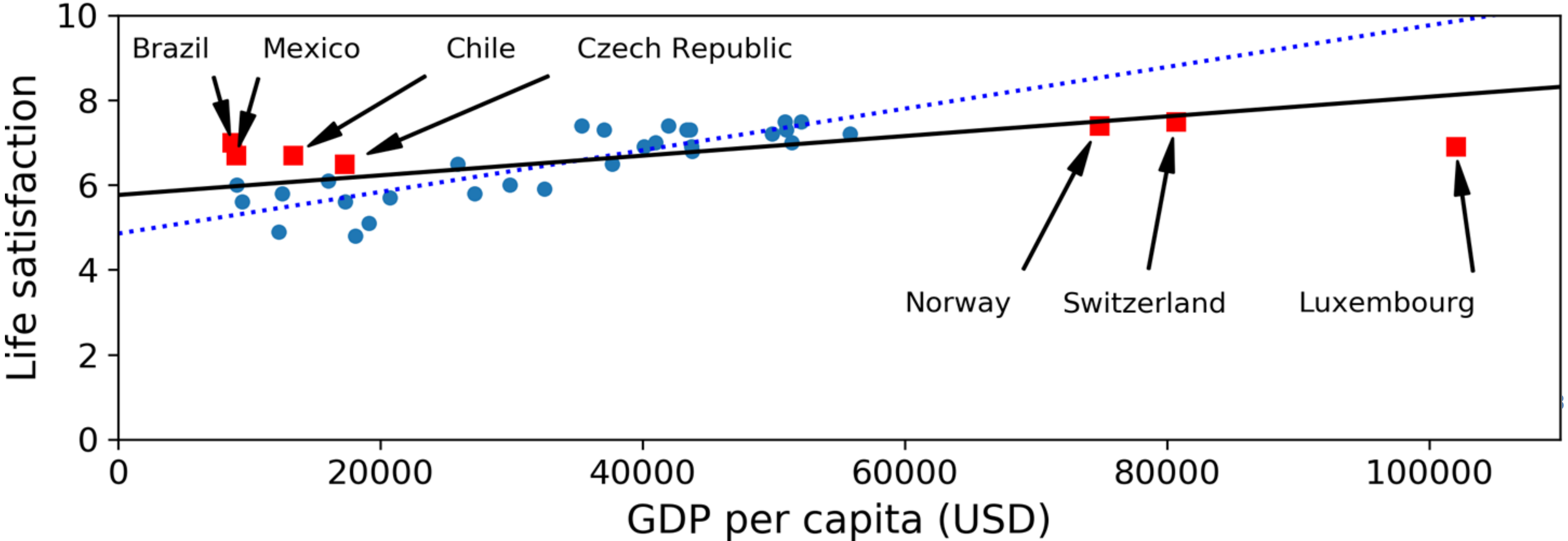
- **Insufficient Quantity of Data**
  - For very simple problems you typically need thousands of examples
  - For complex problems such as image or speech recognition you may need millions of examples



CHALLENGES OF MACHINE LEARNING

- **Non-Representative Training Data**
  - It is crucial that your training data be representative of the new cases you want to generalize to

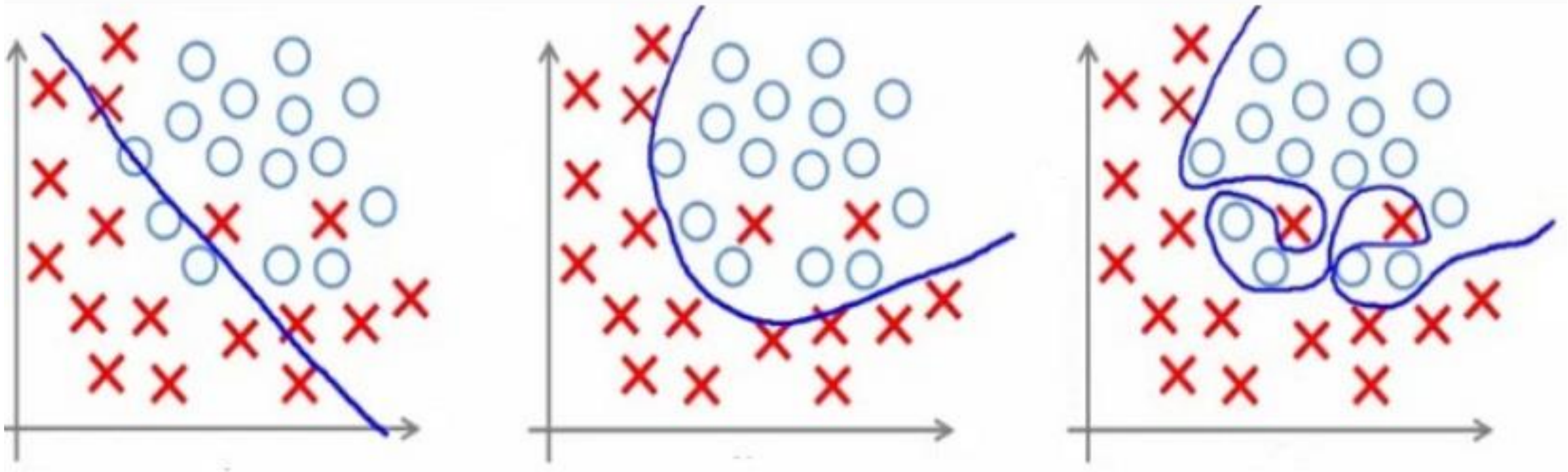




CHALLENGES OF MACHINE LEARNING

- **Poor Quality Data**
  - If training data is full of errors, outliers, and noise it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well
- **Overfitting training data**
  - The model performs well on the training data, but it does not generalize well.
- **Underfitting**
  - It occurs when your model is too simple to learn the underlying structure of the data

CHALLENGES OF MACHINE LEARNING



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

WHAT PROBLEM WE WILL SOLVE

- **Data Pre-processing**

  Feature extraction

  - Tokenization
  - Removing unnecessary punctuation, tags
  - Removing stop words
  - Stemming
  - Lemmatization/POS tagging

  Feature selection

  - Term frequency
  - Chi-square
  - Expected cross entropy
  - Odds Ratio

WHAT PROBLEM WE WILL SOLVE

- Algorithms
  - Support Vector Machines
  - Naive Bayes
  - Maximum Entropy
  - K-nearest neighbour
  - Decision Tree
  - TF-IDF
  - Conditional Random Field (CRF)
  - Latent Dirichlet Allocation (LDA)
  - Artificial Neural Networks
  - LSTM
  - Bi-directional LSTM
  - Recurrent neural networks
  - Models of word embedding:
    - ·Word2Vec
    - ·Glove

WHAT PROBLEM WE WILL SOLVE

- Common NLP techniques
  - Named Entity Recognition
  - Sentiment Analysis
  - Language detection
  - Topic Modeling
  - Text Summarization
  - Machine Translation