Course Overview

DSA 6000: Data Science and Analytics, Fall 2018

Wayne State University

Course Plan

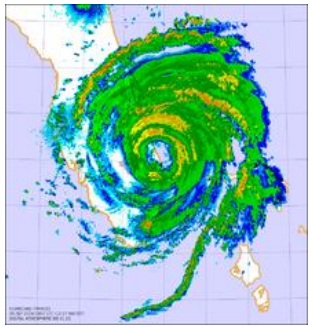
- In this course, we examine general principles of data science.
- Learn the structure and principles to systematically analyze business problems from the data perspective.
- Discuss use cases, business scenarios and success stories.
- Learn the statistical computing language R, and complete a marketable data science project with it.
- Learn data visualization and dashboard tools.

Data Science and Analytics

- Extract useful information from large volumes of data to improve business decision-making
- Discover patterns that are not obvious
- Generate insights
- Make predictions
- Make optimal decisions
- Automate, scale up the impact

Prominent applications of DSA

- Targeted marketing
- Online advertising
- Recommendations for cross-selling
- CRM - analyze customer behavior, manage attrition, value
- Finance: Credit scoring, trading, fraud detection, workforce management
- Supply chain management in retail and manufacturing
- Healthcare: EMR analysis, mining, process improvement



Targeted marketing

- Focusing marketing efforts on specific groups who are more likely to respond.
- When marketing is relevant, people are more likely to spend money on that service or product.
- Glaceau's **vitamin-enriched water** was marketed to people of age 18 - 49 who indicated that they were interested in **health and fitness**. Because of this targeted marketing, the Glaceau Smartwater brand grew approximately 28% in less than a year.
- Data science can help identify the most valuable targets.

Online advertising

- Programmatic advertising
- Various forms: email, search engine ads, mobile, social media
- Data science can help optimize the targeting, placement, and bidding strategy.

Typical Data Science Tasks

1. Classification: predict the class membership of an individual
2. Regression: predict the numerical value of some attribute for an individual
3. Similarity matching: identify similar individuals based on data known about them
4. Clustering: group individuals together by their similarity, but not driven by any specific purpose
5. Association rule mining: find associations between entities based on transactions involving them
6. Profiling: characterize the typical behavior of an individual, group, or population
7. Link prediction: predict connections between data items
8. Data reduction: reduce a large set of data to a smaller set of data that contains much of the important information
9. Causal modeling: understand what events or actions actually influence the quality/quantity of interest

What is Machine

- Machine Learning
  - The 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 P, if its performance on T, as measured by P, improves with experience E. -- Tom Mitchell, 1997
    - If you downloaded a copy of Wikipedia, your computer has a lot more data, but it is not suddenly better at any task. Thus, it is not machine learning.

How humans know

Spam or Ham?

Major Rolland Jude,

Am Major Rolland Jude, and am US army and very happy to meet you, during our mission in 2015 here in Damascus i came across a vault box which contain precious stones such as Diamond and Gold with huge amount of Millions of dollars in a cave which the rebels use as their hid-out to exchange arms from nearby country to fight against the country and the Innocent civilian. i took the trunk box and relocated it far from here and later move it out from this country by the help of my red cross team to deposited safe and secure under a Currier company, am looking for a benefactor who can contact the Currier company on my behalf to do everything and any solve any cost that is involved to receive the package safe and keep with faith and trust until i return and come over so we can share on agreement, 35% will be giving to you respectively as your share from each and every content which is the Diamond, Gold and the (Twelve million dollars six hundred and seventy five thousand USD) $12.675,000,000 million dollars, i have invest my money into this by moving it from here safely to to another country and depositing it safe in the company for this long, so i will be happy any cost that will be involved which you will solve to make the delivery get to your destination will be refund to you by agreement of 15% from each content making it 50% \ 50% which will be understanding. contact me back for evidence i make from my packages content before depositing in the delivery company.

Major Rolland Jude,

Good Day,

I am Mr.Heng Srey,I have an interesting business proposal that will yield something good to both of as at the end as long as you can handle the project from your side.If you are interested to partner with me you reply back for more details.

Best Regards,

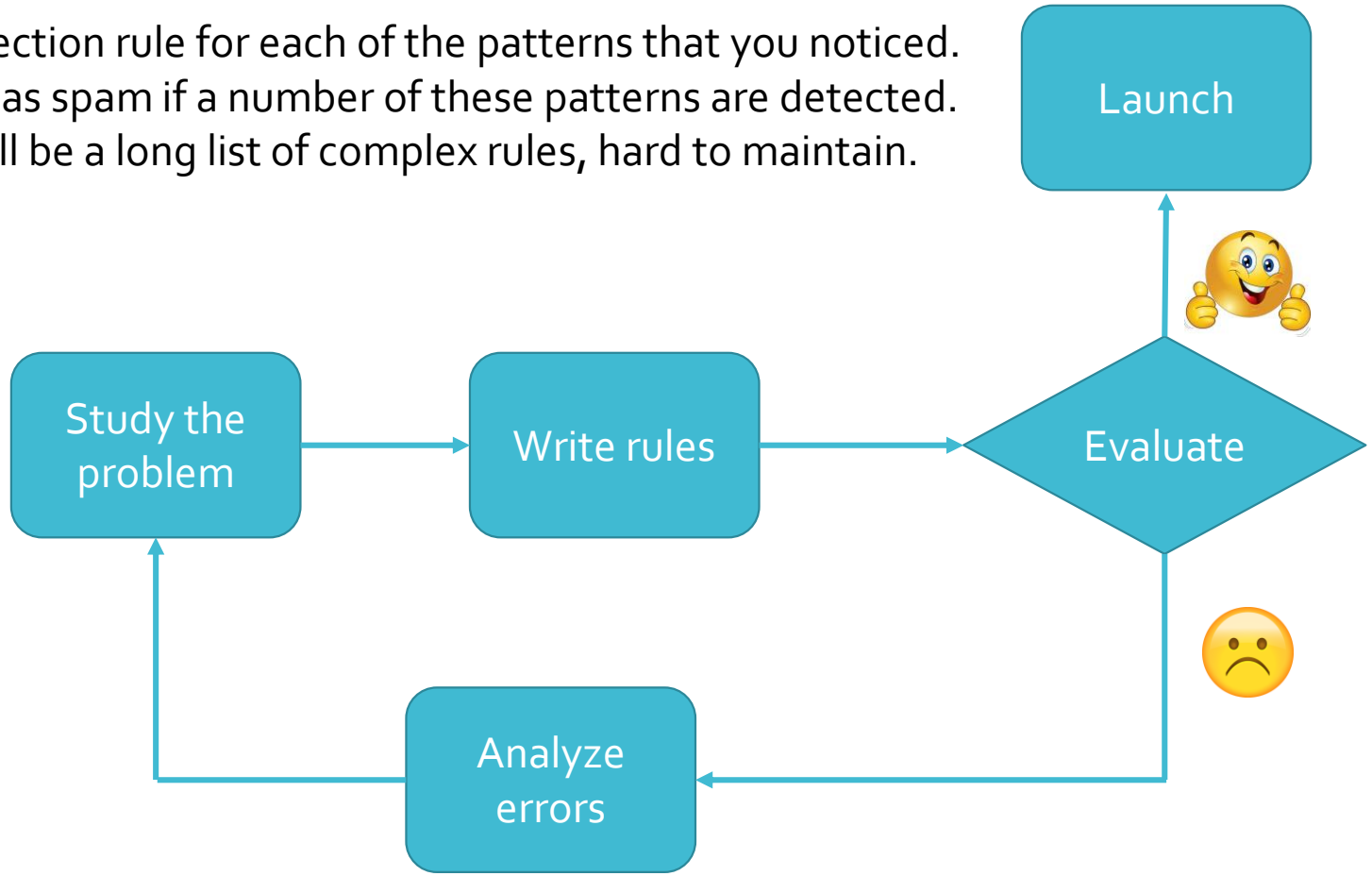
Mr.Heng Srey

Polanyi's Paradox

- Human knowledge and capability relied on skills and rulesets that are often beneath our conscious appreciation. (http://hplusmagazine.com/2015/10/07/polanyis-paradox-will-humans-maintain-any-advantage-over-machines/)
- We know more than we can tell.
- We cannot codify many tasks easily.

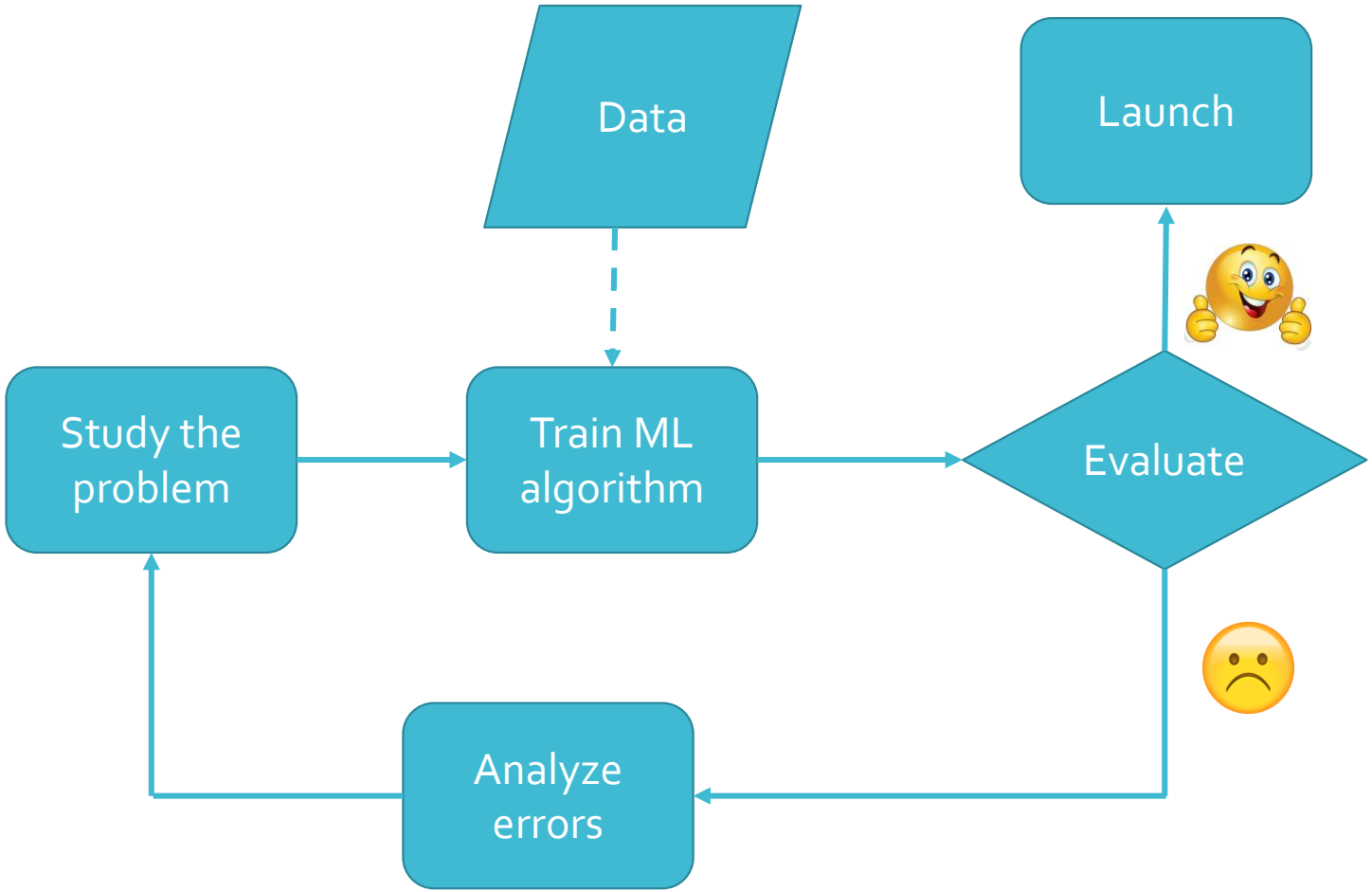
The traditional approach

Write a detection rule for each of the patterns that you noticed.
Flag emails as spam if a number of these patterns are detected.
Program will be a long list of complex rules, hard to maintain.



What if the spammer changed his style or storyline?
What about more complex tasks, like speech recognition?

Machine learning approach



Automatically learns which words/phrases are good predictors of spam.
Program is much shorter, easier to maintain, likely more accurate.

Statistical Learning

Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam *and* email.

Why using machine learning?

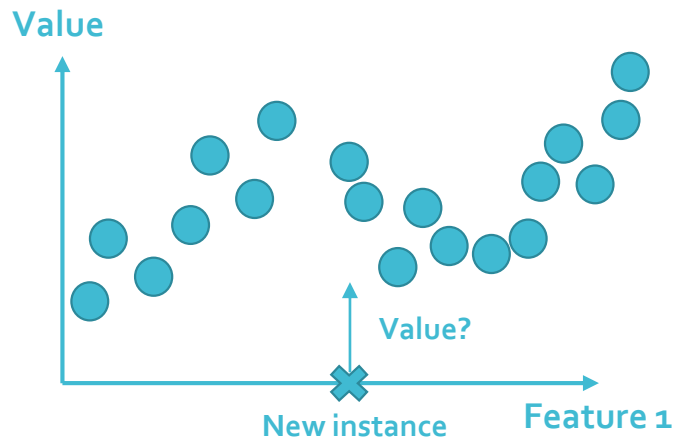
- Machine learning is great for:
  - Problems for which existing solutions require a lot of hand-tuning or long lists of rules
  - Complex problems for which there is no good solution at all using a traditional approach
  - Fluctuating environments: a machine learning system can adapt to new data
  - Getting insights about complex problems and large amounts of data

Types of machine learning systems

- Machine learning systems are usually classified based on:
  - Whether or not they are trained with human supervision (supervised, unsupervised, semi-supervised, reinforcement learning)
  - Whether or not they can learn incrementally on the fly (online v.s. batch learning)
  - Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model (instance-based v.s. model-based learning)

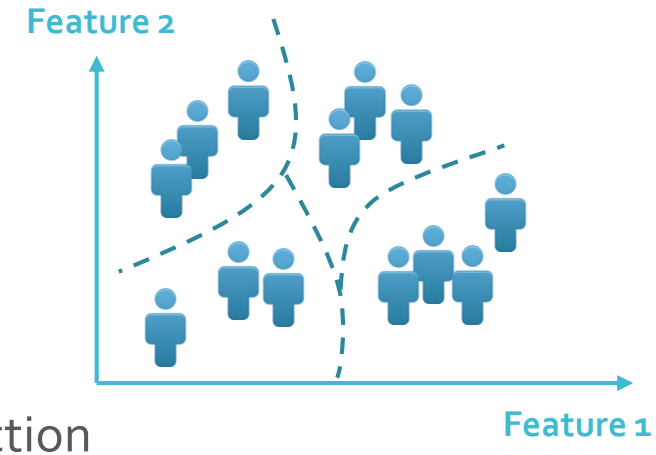
Supervised learning

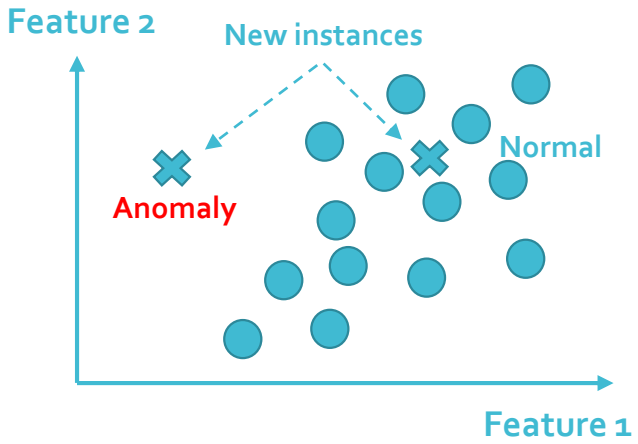
- The training data that you feed to the algorithm includes the desired solutions, called *labels*.
- Classification: learn to classify new instances (e.g., spam or ham)
- Regression: learn to predict a target numeric value (e.g., the price of a house), given a set of *features* (sqft, # of rooms, lot size, etc.) called *predictors*.
- Important algorithms:
  - K-Nearest Neighbors
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVMs)
  - Decision Trees and Random Forests
  - Neural Networks



Unsupervised learning

- The training data is unlabeled. The system tries to learn without a teacher.
- Tasks and algorithms:
- Clustering and Anomaly Detection
  - K-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization
- Visualization and Dimensionality Reduction
  - Principle Component Analysis (PCA)
  - Locally-Linear Embedding (LLE)
- Association Rule Learning
  - Apriori
  - Eclat





Semi-supervised learning

- Deals with partially labeled data, usually a lot of unlabeled data and a little bit of labeled data.

Figure: Classification with only labeled samples

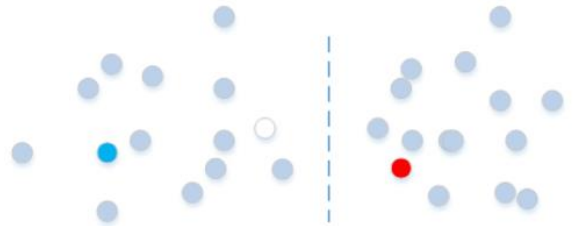


Figure: Classification with a complete set of samples

Unlabeled data can help reveal structures that are otherwise unseen from (the small amount of) labeled data, thus increase prediction accuracy

Reinforcement learning

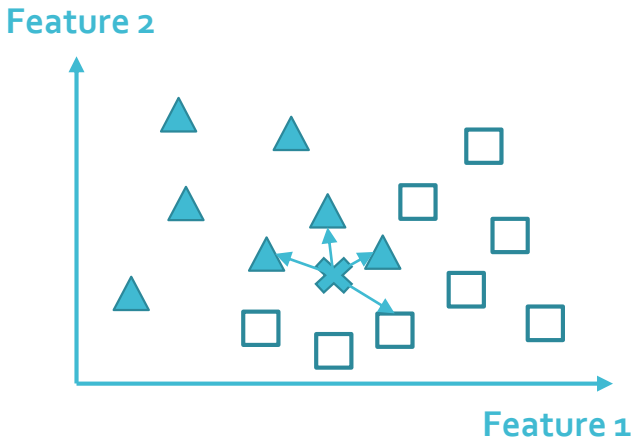
- The learning system is an *agent*. The agent observes the environment, select and perform actions, and get rewards in return. It must learn by itself what is the best strategy, called a policy, to get the most reward over time.
  - Walking robots by Boston Dynamics (https://www.youtube.com/watch?v=vjSohj-Iclc)
  - AlphaGo by Google (https://www.tastehit.com/blog/google-deepmind-alphago-how-it-works/)
  - Deep Q-learning plays Atari game (https://www.youtube.com/watch?v=V1eYniJoRnk)

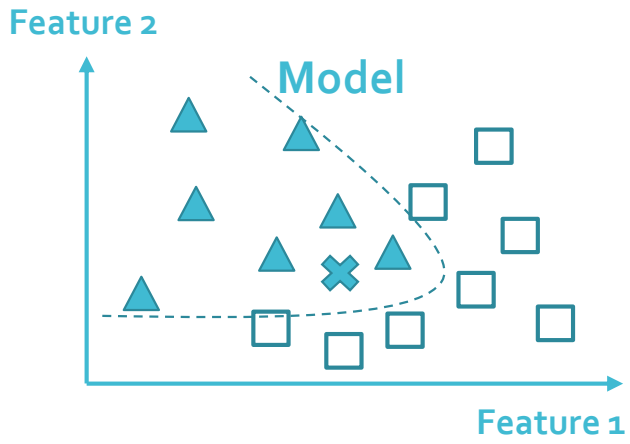
Batch v.s. Online Learning

- Batch learning
  - Trained with all available data, then launched into production without learning any more – it just applies what has learned.
  - New data? Train the model again and launch again, periodically. Can be automated.
- Online learning
  - Train the system incrementally by feeding it data instances sequentially
  - Great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly and autonomously
  - Also useful for training systems on huge datasets that cannot fit in one computer's memory, called *out-of-core* learning.

Instance-based v.s. Model-based learning

- Instance-based learning
  - Flag as spam emails that are very similar to known spam emails
  - Need a *measure of similarity* between two emails, e.g., the number of words they have in common
    - K-nearest neighbors algorithm is an instance-based learning algorithm
- Model-based learning
  - Generalize from a set of examples by building a model of these examples, then use the model to make predictions
  - E.g., LDA, Linear regression, etc.





Main challenges of machine learning

- Two things can cause problems: “bad data” and “bad algorithm”.
  - Insufficient quantity of training data
  - Non-representative training data
  - Poor-quality data
  - Irrelevant features
  - Overfitting the training data
  - Underfitting the training data

Insufficient quantity of data

- Data matters more than algorithms for complex problems.
- "We may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development." – Banko and Brill

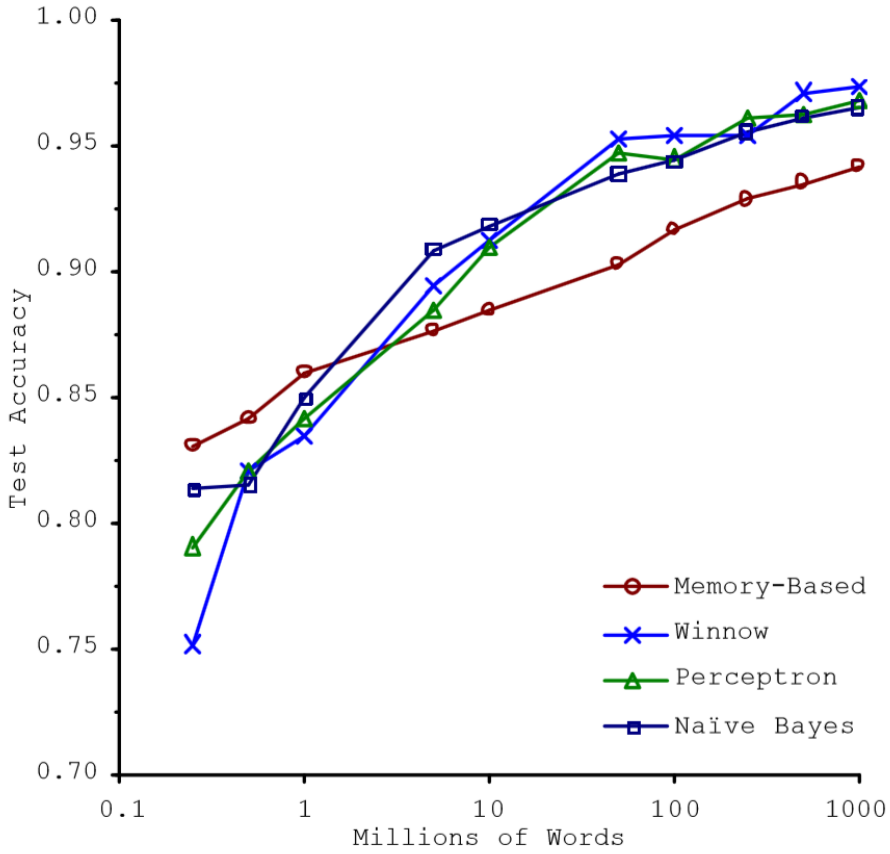
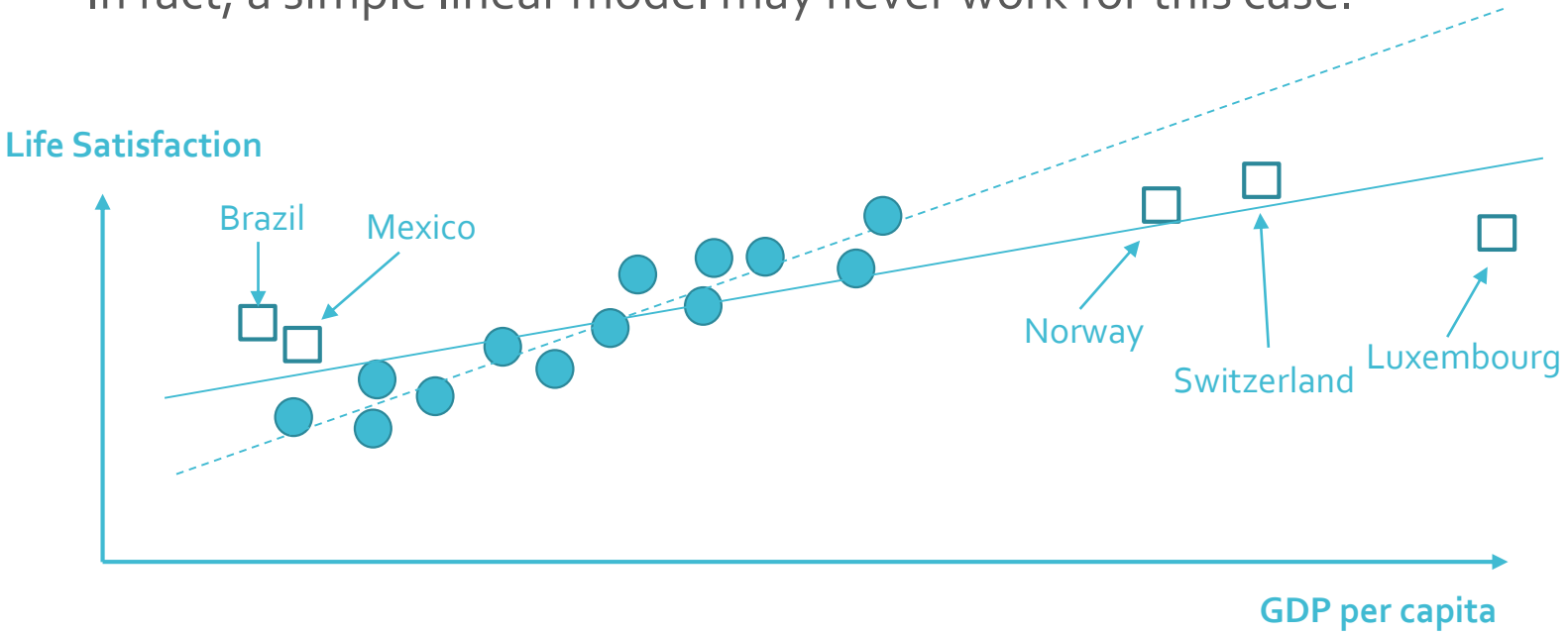


Figure source: Michele Banko and Eric Brill (2009). Scaling to Very Very Large Corpora for Natural Language Disambiguation.

Non-representative training data

- When a few data points are missing from the training data, a linear model (dashed line) seems to work well.
- Adding the data points significantly altered the linear fit (solid line).
- In fact, a simple linear model may never work for this case.



- It is crucial to use a training set that is representative of the cases you want to generalize to.
- Making extrapolations can be dangerous.

Poor-quality data

- If your training data is full of errors, outliers and noise, it will make it harder for the system to detect the underlying patterns.
- It is worth the effort to spend time cleaning up your training data.
  - That's what data scientists spend a significant part of their time doing.
- Discard instances that are clearly outliers, and diagnose the possible issue in your data collection process that produced the bad data in the first place.
- If some instances are missing a few features, decide whether to ignore the feature altogether, ignore these instances, or fill in (impute) the missing values.

Irrelevant Features

- Garbage in, garbage out.
- Provide the algorithm with enough relevant features and not too many irrelevant features.
- Feature engineering: produce a good set of features to train on.
  - Feature selection
  - Feature extraction
  - Creating new features by gathering new data

Overfitting

- Say you are visiting a foreign country and the taxi driver rips you off. You might be tempted to say *all* taxi drivers in that country are thieves.
- That is overgeneralization.
- In machine learning it is called *overfitting*.
  - The model performs well on the training data, but it does not generalize well on new data.
- Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. Possible cures:
  - Simplify the model (use fewer parameters, or add constraints)
  - Gather more training data
  - Reduce the noise in the training data, e.g., errors and outliers.

Underfitting

- Underfitting happens when your model is too simple to learn the underlying structure of the data.
- Reality is more complex than the model, so prediction are bound to be inaccurate, even on the training examples.
- Possible cures:
  - Select a more powerful model, with more parameters
  - Feed better features to the learning algorithm
  - Reduce the constraints on the model

No free lunch theorem

- A model is a simplified version of the observations. The simplifications are meant to discard the superfluous details that are unlikely to generalize to new instances.
- To decide what data to discard and what data to keep, you must make *assumptions*.
- No Free Lunch theorem: **There is no model that is guaranteed to work better than all others over all possible data sets.**
- It is impossible to evaluate all models and find out the best.
- In practice you make some reasonable assumptions about the data and evaluate only a few reasonable models.

D. Wolperts (1996). The Lack of A Priori Distinctions Between Learning Algorithms.

Machine Learning v.s. Statistical Learning

- *Machine learning* arose as a subfield of Artificial Intelligence.
  - Emphasizes large-scale applications and prediction accuracy
- *Statistical learning* arose as a subfield of Statistics.
  - Emphasizes models and their interpretability, precision and uncertainty
- Both focus on supervised and unsupervised problems.

References

- *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, by Aurelien Geron, O'Reilly.
  - Most content in this presentation is adopted from Chapter 1 of the above book.
- The Unreasonable Effectiveness of Data (http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf)
- ISLR
- Extended Reading:
  - *What can machine learning do? Workforce implications. (http://science.sciencemag.org/content/358/6370/1530.full)*