DataScience for Development and Social Change, 2015

Science

Getting information from your datasets. Truthfully.

Machine Learning

- Build a model that extract patterns from data
- Use those patterns to predict missing or future data
- Do it automatically

(part of artificial intelligence)

Example Uses

- Search engines: producing results tailored to you
- Spam filters: learning what should go into your spam folder (and not)
- Handwritten character recognition: recognizing “2”
- Gap-filling: estimating missing data values
- Prediction

Why Use Machine Learning?

- You don’t have all the features you want in your data.
  - Some features don’t exist
  - Some features have missing data

- You want to predict an outcome (eg. loan default) based on previous examples

Machine Learning Areas

- Classification: predict classes
- Regression: predict numerical values
- Clustering: predict group membership

Classification

Classification is the allocation of a piece of information to a category (e.g. male, female, other).

- Often need to automate this:
  - e.g. huge dataset, specialist knowledge (e.g. Tagalog, non-obvious connections), regularly-updated data etc.

Classification: Practice

- Which of these people are male or female?
  1. She previously worked at Kings College London.
  2. Bayani advises the UN on the use of drones.
  3. Kim is a leading scholar on text classification.
  4. His work on sand eels is renowned.
  5. Diwata Jones.
  6. He works closely with Sandra Smith and her work on fish.

Classification: Practice

1. Female: “She” at start of description.
2. Male: male first name (look in babynames.ch)
3. Unknown: “Kim” is male or female first name.
4. Male: “His” at start of description.
5. Female: female first name (look in babynames.ch)
6. Male: “He” at start of description, but note the “her” later in the text.

- Alternative gender classification methods include:
  - GenderAnalyzer: guesses gender of writer (has API)
  - Name endings: Names ending in a, e and i are more likely to be female, while names ending in k, o, r, s and t are more likely to be male

The Scikit.learn Library

- Python library
- Contains most common machine learning algorithms
- import “sklearn”

Scikit.learn: classification

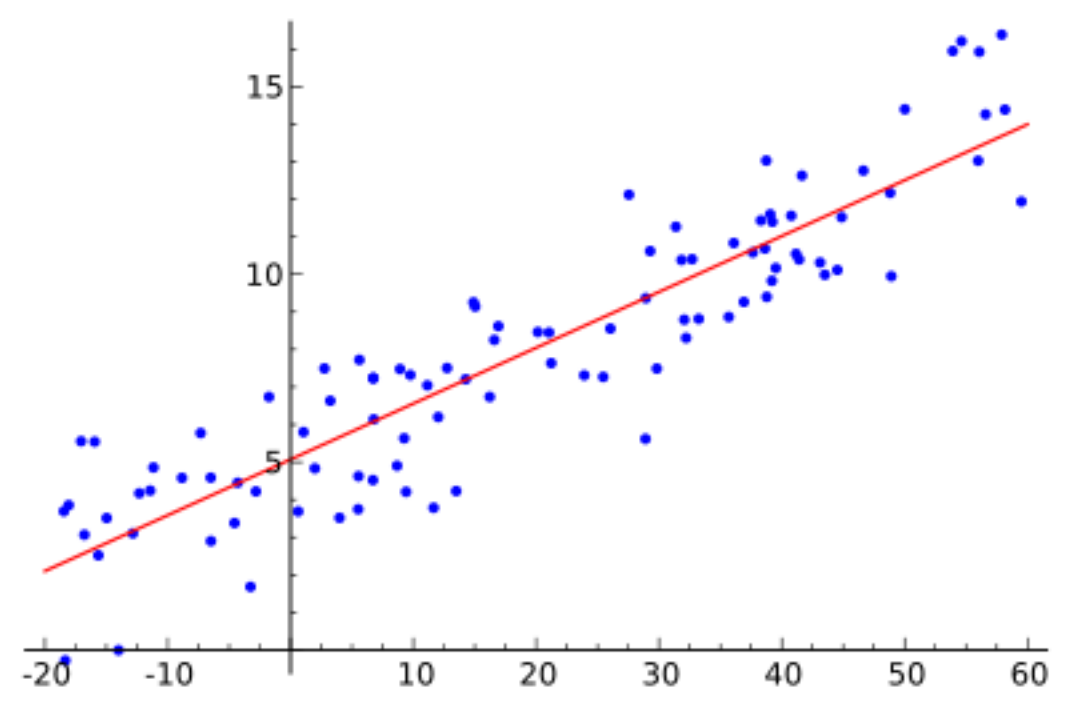
Feature Selection

- For classification to work, you need a feature set that's useful, e.g.
  - First names
  - First name endings
  - Personal pronouns in text
  - Not: department (e.g. Engineering)

- A "training set": a set of pre-classified examples, for the classifier to learn from
- A "test set": a smaller set of pre-classified examples, for the classifier to check its predictions

Regression

- Finds the best-fit line between points
- Because: you need to know the relationship between 2 or more features
- Needs:
    - 1) data points, in 2 or more dimensions
    - 2) “best-fit” equation, e.g. city-block metric, least-squares, edit distance etc



Logistic Regression

- Because: you want to predict the class of a datapoint

Unsupervised Learning

- Supervised learning:
    - We ‘teach’ the computer how to do a task
    - E.g. we tag male/female names until the machine can do this reliably (i.e. low number of misclassifications) for itself
- Unsupervised learning:
    - The computer learns for itself, without teaching
    - E.g. the computer separates a dataset into classes of closely-related data points, without being told what or where these are

Clustering

- Divide a dataset up into n related “clusters” or classes
    - **Because**: you want to know if groups exist in your data (so you can investigate further / use characteristics of those groups to advantage)
    - **Issue**: knowing “n”
    - **Algorithms**: k-nearest neighbours
    - **Python module**: sklearn.cluster

k-Means

Hypothesis Testing

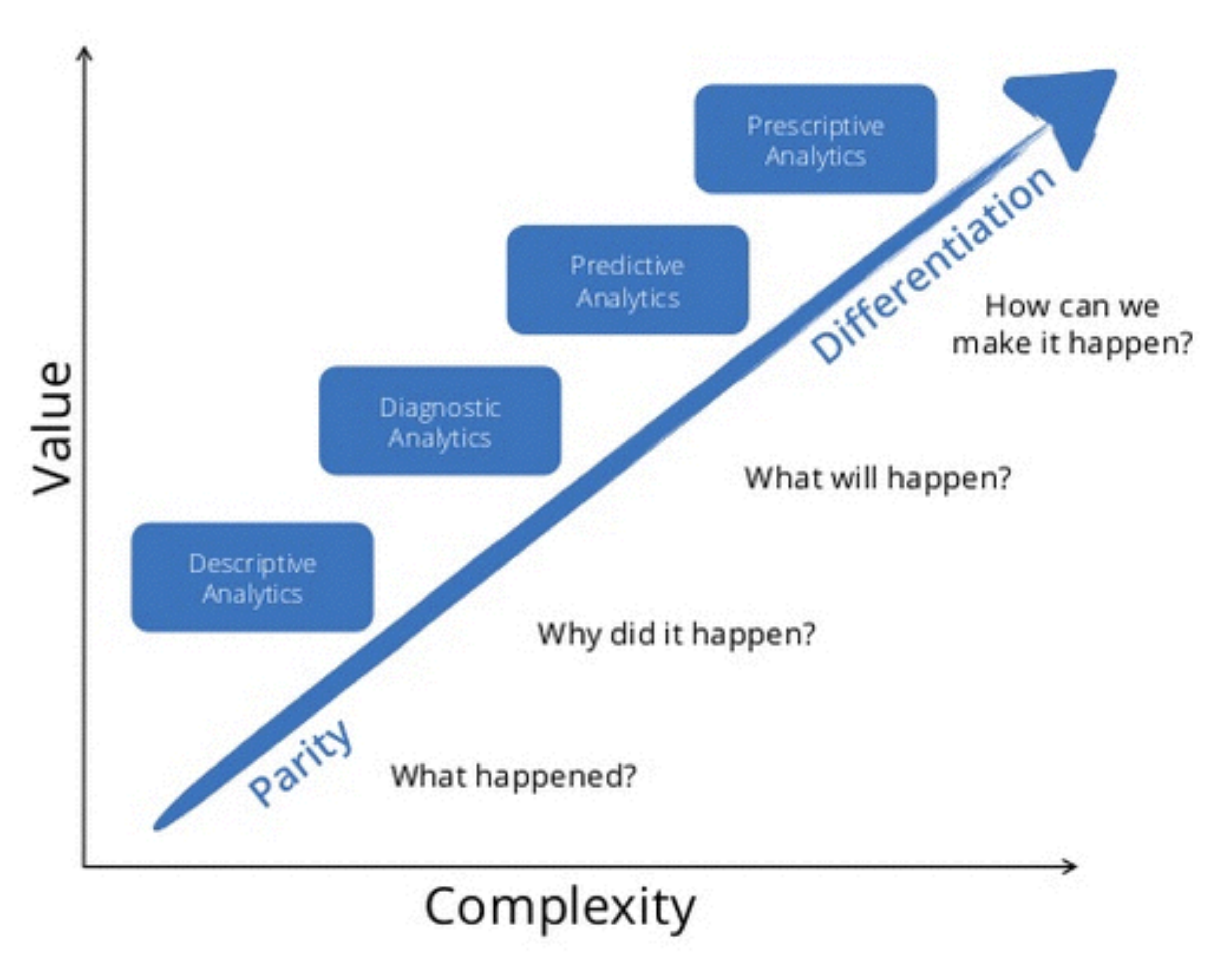
- Data scientists often mention "p-values"
    - Part of hypothesis testing:
        - Null hypothesis = no relationship between two items
        - significance level = lowest value we can reject null hypothesis at, e.g. 1%, 5% etc. (below this = sampling error)
        - p-value: calculated strength of relationship. If p-value < significance level, assume no relationship

Improving learning: Bagging and Boosting

- Sometimes, you don’t have to choose between models…
  - Bagging: combine different types of model
  - Boosting: combine the same type of model

Network Analysis

What's next?



(Image: Ken Collier, Thoughtworks)

Continuing your Science journey

- http://scikit-learn.org/stable/index.html