Advanced Data Analysis

DATA 71200

Class 1

Course Description

- This course will provide you with skills necessary to a**pply machine learning techniques to data**, and i**nterpret and communicate their results**.

- You will also begin to develop **intuitions** about when machine learning is an appropriate tool versus other statistical methods.

- This course will cover both **supervised methods** (e.g., k-nearest neighbors, naïve Bayes classifiers, decision trees, and support vector machines) and **unsupervised methods** (e.g., principal component analysis and k-means clustering).

    - The supervised methods will focus primarily on **“classic” machine learning techniques** where features are designed rather than learned, although we will briefly look at recent deep learning models with neural networks.

- This is an **applied machine learning class** that emphasizes the intuitions and know-how needed to get learning algorithms to work in practice, rather than mathematical derivations.

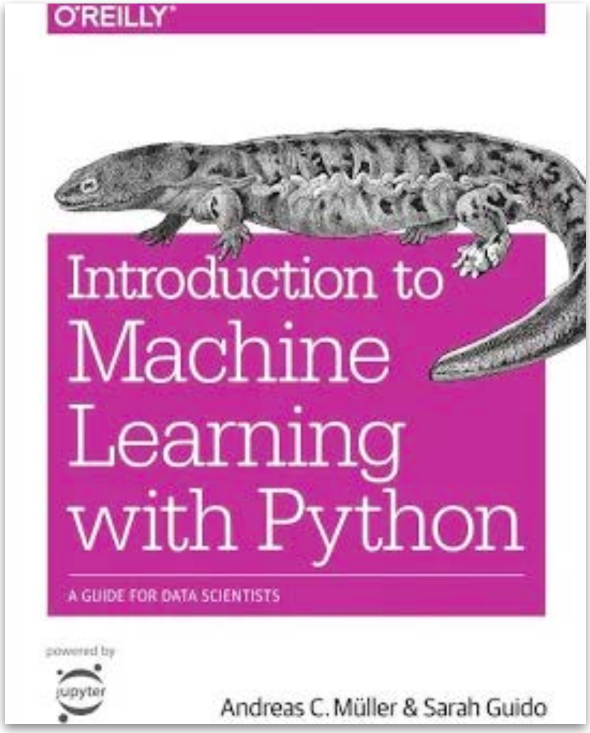
- The course will be taught in **Python**, primarily using the **scikit-learn** library.

Course Objectives

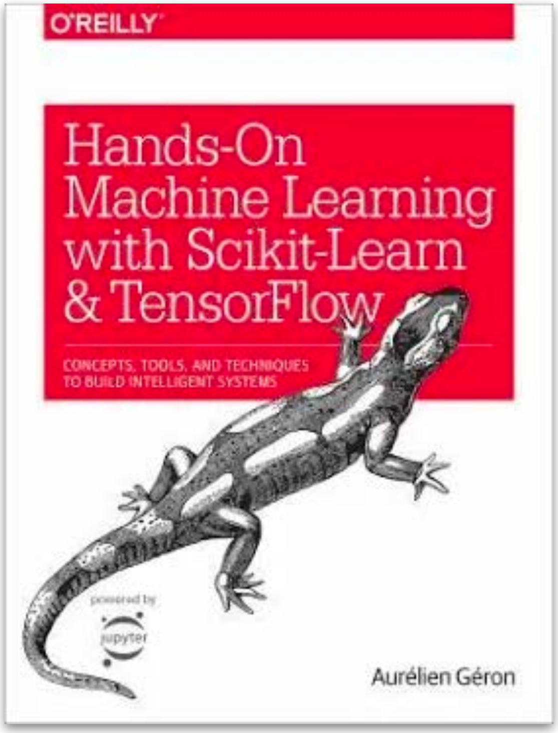
- By the end of the course, you will be able to
  - articulate the main assumptions underlying machine learning approaches
  - demonstrate the basic principles of dataset creation
  - articulate the importance of data representations
  - evaluate machine learning algorithms
  - articulate the difference between supervised and unsupervised learning
  - apply a range of supervised and unsupervised learning techniques

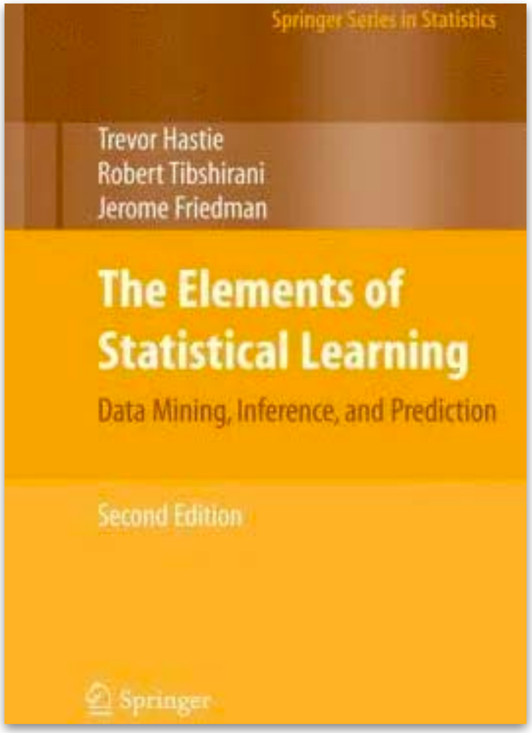
Textbooks

Required


O'REILLY
Introduction to Machine Learning with Python
A GUIDE FOR DATA SCIENTISTS
powered by jupyter
Andreas C. Müller & Sarah Guido

Recommended


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by jupyter
Aurélien Géron


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Grade Breakdown

Class Participation	10%
Datacamp Assignments	25%
Project 1: Dataset creation	15%
Project 2: Supervised learning	15%
Project 3: Unsupervised learning	15%
Final Paper	20%

Grade Breakdown Details

- **Class Participation: 10%**
    - The participation grade is a combination of attendance (including arriving on time); attentiveness, engagement, and participation during class; and general preparedness for class discussions.
- **Datacamp Assignments: 25%**
    - These projects are hands-on activities designed to both provide coding background and reinforce the concepts covered in class.

Grade Breakdown Details

- **Project 1 (Dataset creation): 15%**
  - Curation and cleaning of a labeled data set that you will use for the supervised and unsupervised learning tasks in project 2 and 3. The dataset can built from existing data and should be stored in your GitHub repostiory.
- **Project 2 (Supervised learning): 15%**
  - Application of two supervised learning techniques on the dataset you created in Project 1. This assignment should be completed as a Jupyter notebook your GitHub repository.

Grade Breakdown Details

- **Project 3 (Unsupervised learning): 15%**
  - Application of two unsupervised learning techniques on the dataset you created in Project 1. This assignment should be completed as a Jupyter notebook your GitHub repository.
- **Final Paper: 20%**
  - A 5--8 page paper describing the work you did in projects 1--3 (your dataset and your supervised and unsupervised experiments). The paper should describe both what you did technically and what you learned from the relative performance of the machine learning approaches you applied to your dataset. This assignment should be posted as a PDF in your GitHub repository.

Course Schedule

31-May	Introduction/What is Machine Learning?
1-Jun	Getting Started with Machine Learning
2-Jun	Inspecting Data
6-Jun	Representing Data
7-Jun	*Async: DataCamp Modules*
8-Jun	Evaluation Methods

Course Schedule

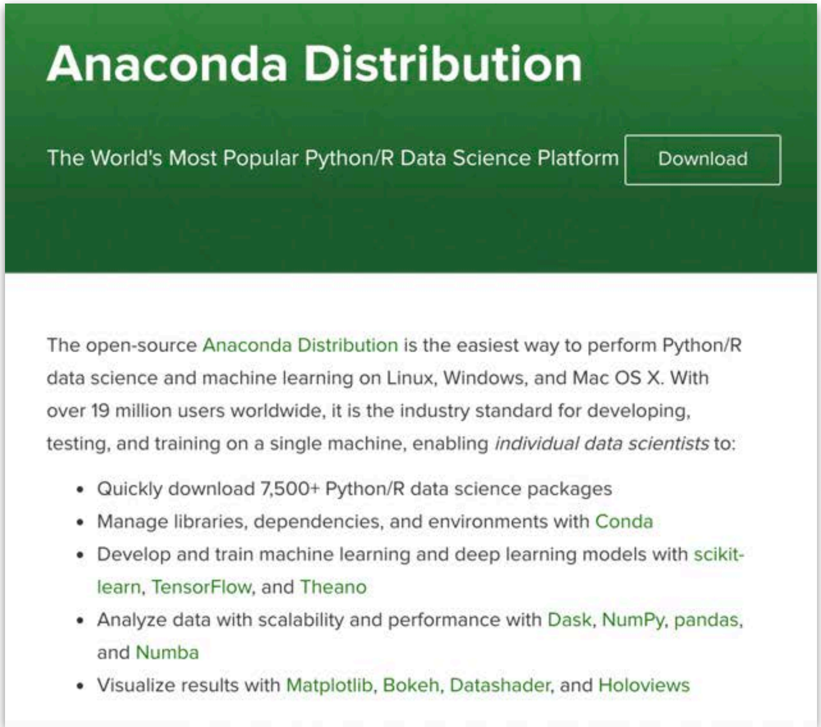
9-Jun	*Async: DataCamp Modules*
13-Jun	Supervised Learning (k-Nearest Neighbors, Linear Models, and Naive Bayes Classifiers) *Project 1 Due*
14-Jun	*Async: DataCamp Modules*
15-Jun	Supervised Learning (k-Nearest Neighbors, Linear Models, and Naive Bayes Classifiers)
16-Jun	*Async: DataCamp Modules*

Course Schedule

21-Jun	Supervised Learning (Decision Trees, Support Vector Machines and Uncertainty estimates from Classifiers) *Project 2 Due*
22-Jun	Unsupervised Learning (Dimensionality Reduction & Feature Extraction, and Manifold Learning)
23-Jun	*Async: DataCamp Modules*
1-Jul	*Project 3 Due* *DataCamp Assignments Due*
8-Jul	*Final Project Due*

Coding Environment

- **Python 3**
  - matplotlib, NumPy, Pandas, SciPy, scikit learn (+ mglearn)
- **Jupytr notebooks (Anaconda or Google Colab)**

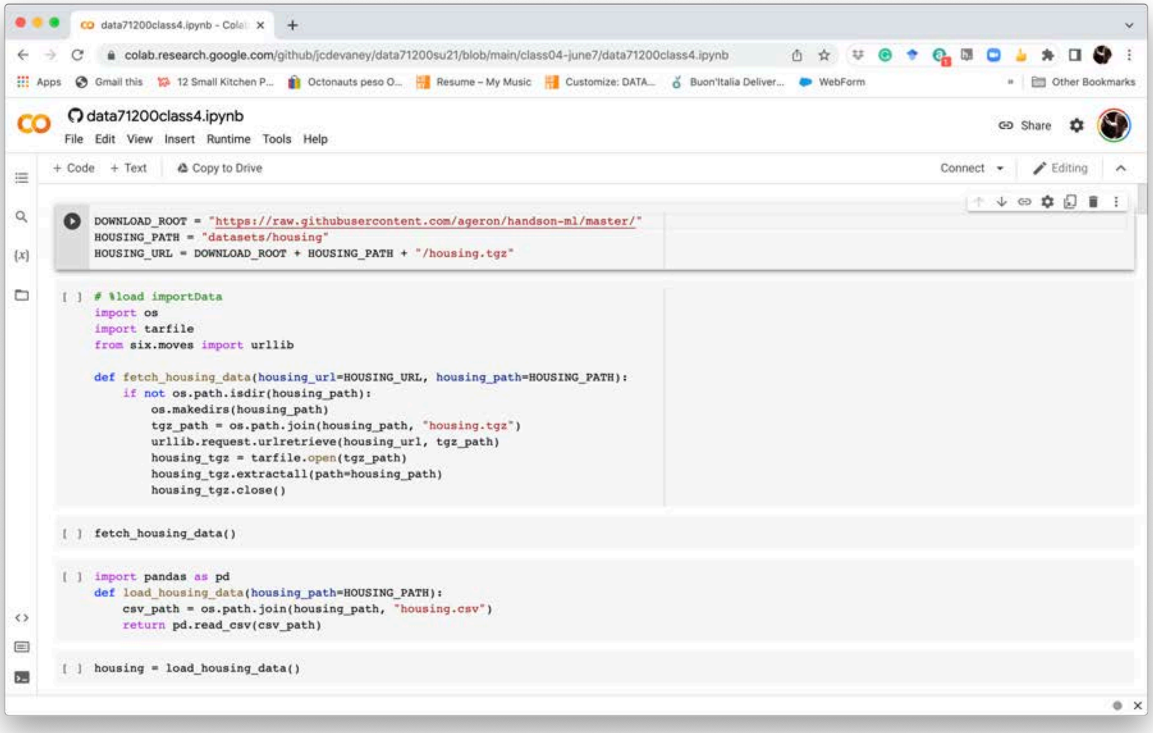


Anaconda Distribution

The World's Most Popular Python/R Data Science Platform

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 19 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling *individual data scientists* to:

- Quickly download 7,500+ Python/R data science packages
- Manage libraries, dependencies, and environments with Conda
- Develop and train machine learning and deep learning models with scikit-learn, TensorFlow, and Theano
- Analyze data with scalability and performance with Dask, NumPy, pandas, and Numba
- Visualize results with Matplotlib, Bokeh, Datashader, and Holoviews



```
DOWNLOAD_ROOT = "https://raw.githubusercontent.com/ageron/handson-ml/master/"
HOUSING_PATH = "datasets/housing"
HOUSING_URL = DOWNLOAD_ROOT + HOUSING_PATH + "/housing.tgz"
```

```
[ ] # %load importData
    import os
    import tarfile
    from six.moves import urllib

    def fetch_housing_data(housing_url=HOUSING_URL, housing_path=HOUSING_PATH):
        if not os.path.isdir(housing_path):
            os.makedirs(housing_path)
            tgz_path = os.path.join(housing_path, "housing.tgz")
            urllib.request.urlretrieve(housing_url, tgz_path)
            housing_tgz = tarfile.open(tgz_path)
            housing_tgz.extractall(path=housing_path)
            housing_tgz.close()
```

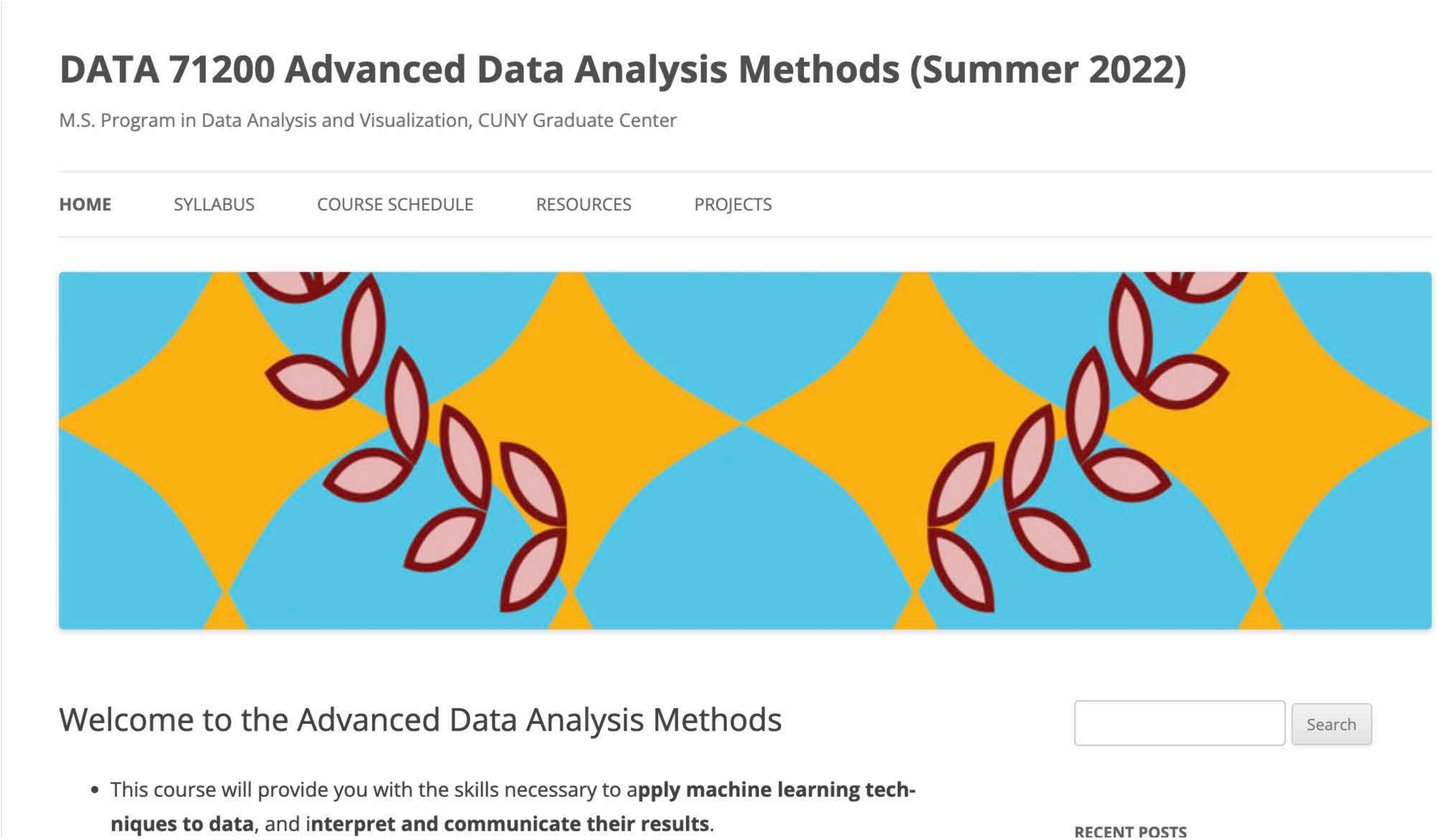
```
[ ] fetch_housing_data()
```

```
[ ] import pandas as pd
    def load_housing_data(housing_path=HOUSING_PATH):
        csv_path = os.path.join(housing_path, "housing.csv")
        return pd.read_csv(csv_path)
```

```
[ ] housing = load_housing_data()
```

Tutorial: https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/

Class Website



DATA 71200 Advanced Data Analysis Methods (Summer 2022)

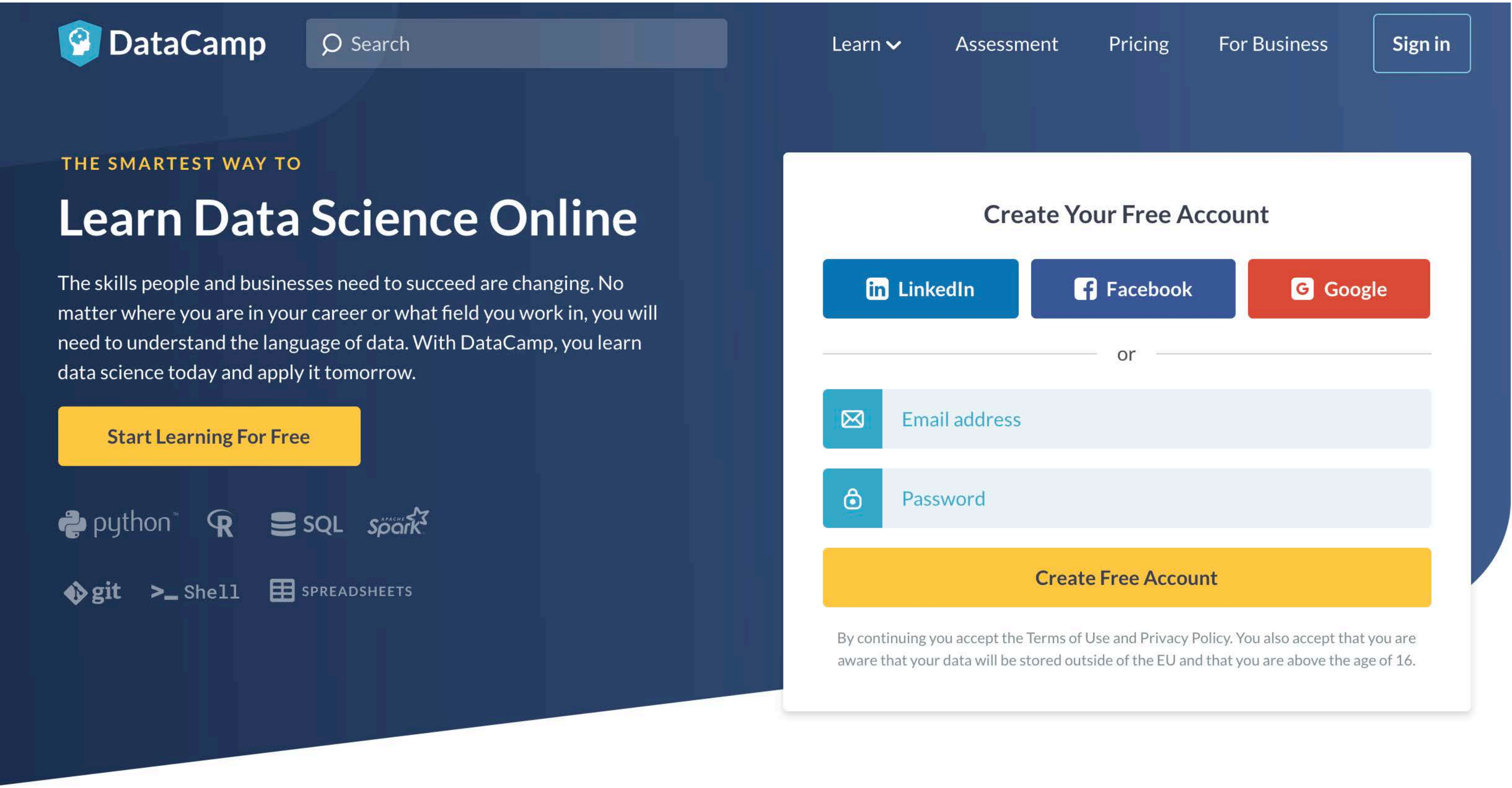
M.S. Program in Data Analysis and Visualization, CUNY Graduate Center

Welcome to the Advanced Data Analysis Methods

- This course will provide you with the skills necessary to a**pply machine learning techniques to data**, and i**nterpret and communicate their results**.

https://data71200su22.commons.gc.cuny.edu/

Data Camp



THE SMARTEST WAY TO

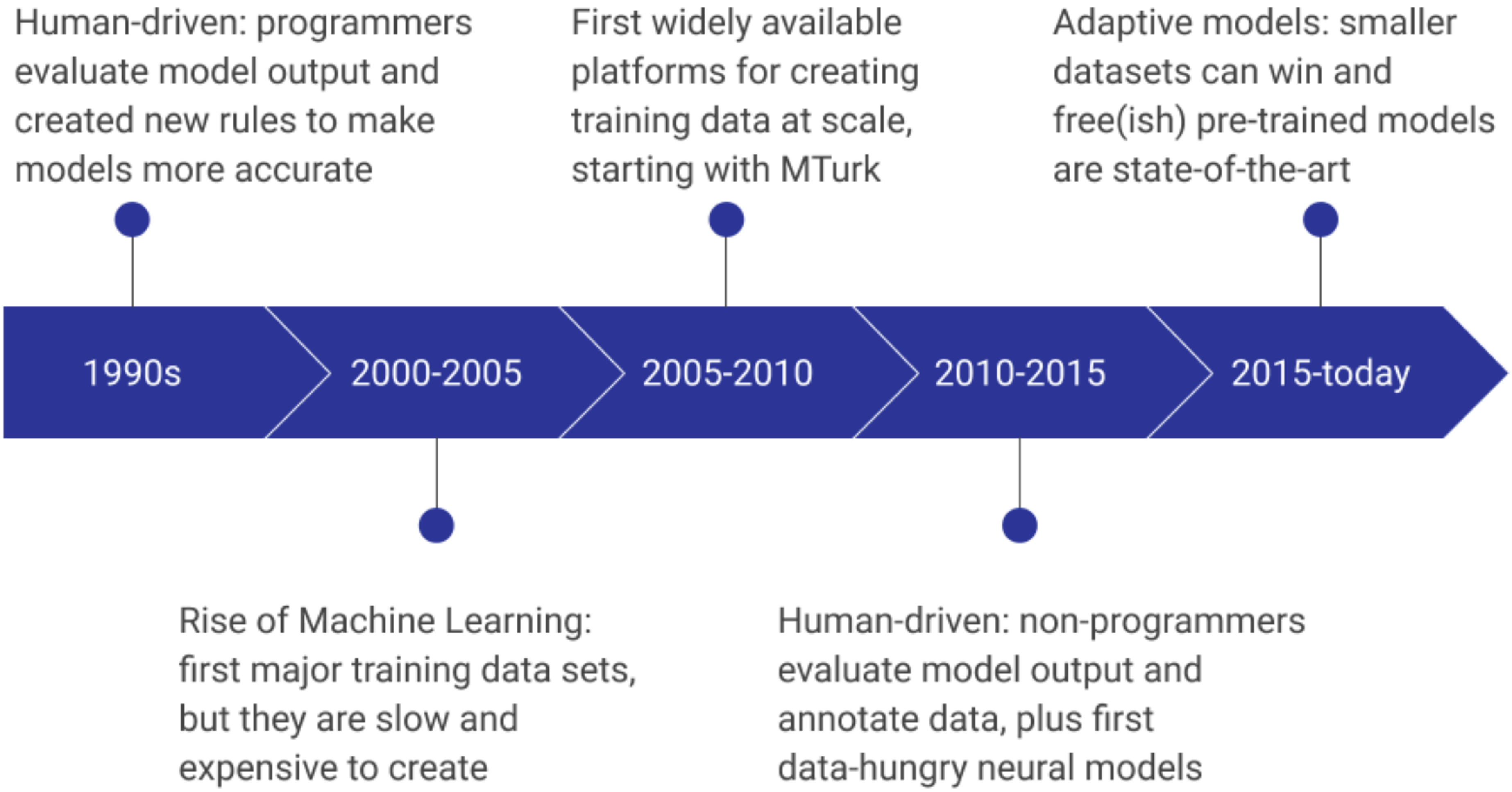
Learn Data Science Online

The skills people and businesses need to succeed are changing. No matter where you are in your career or what field you work in, you will need to understand the language of data. With DataCamp, you learn data science today and apply it tomorrow.

Create Your Free Account

By continuing you accept the Terms of Use and Privacy Policy. You also accept that you are aware that your data will be stored outside of the EU and that you are above the age of 16.

Machine Learning



1990s

Human-driven: programmers evaluate model output and created new rules to make models more accurate

2000-2005

Rise of Machine Learning: first major training data sets, but they are slow and expensive to create

2005-2010

First widely available platforms for creating training data at scale, starting with MTurk

2010-2015

Human-driven: non-programmers evaluate model output and annotate data, plus first data-hungry neural models

2015-today

Adaptive models: smaller datasets can win and free(ish) pre-trained models are state-of-the-art

https://towardsdatascience.com/a-brief-history-of-training-data-9c513fc95b3e

Key Questions

- “How can one construct computer systems that automatically improve through experience?”
- “What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?”
- “How accurately can the algorithm learn from a particular type and volume of training data?”
- “How robust is the algorithm to errors in its modeling assumptions or to errors in the training data”

Jordan, Michael I. and Tom M. Mitchell. (2015). “Machine Learning: Trends, perspectives, and prospects” *Science*.

Machine Learning vs Traditional Programming

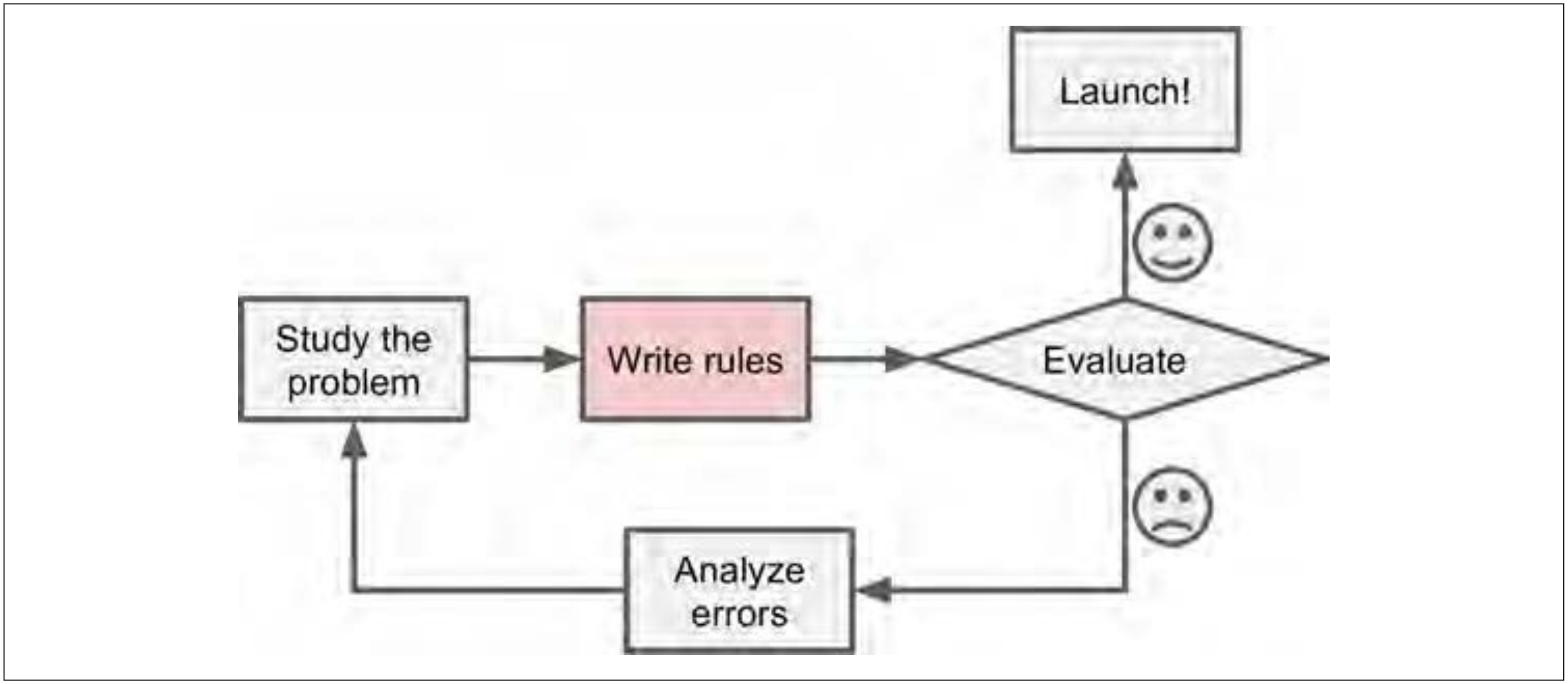


Figure 1-1. The traditional approach

Image from: Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O'Reilly Media, Inc.

Machine Learning vs Traditional Programming

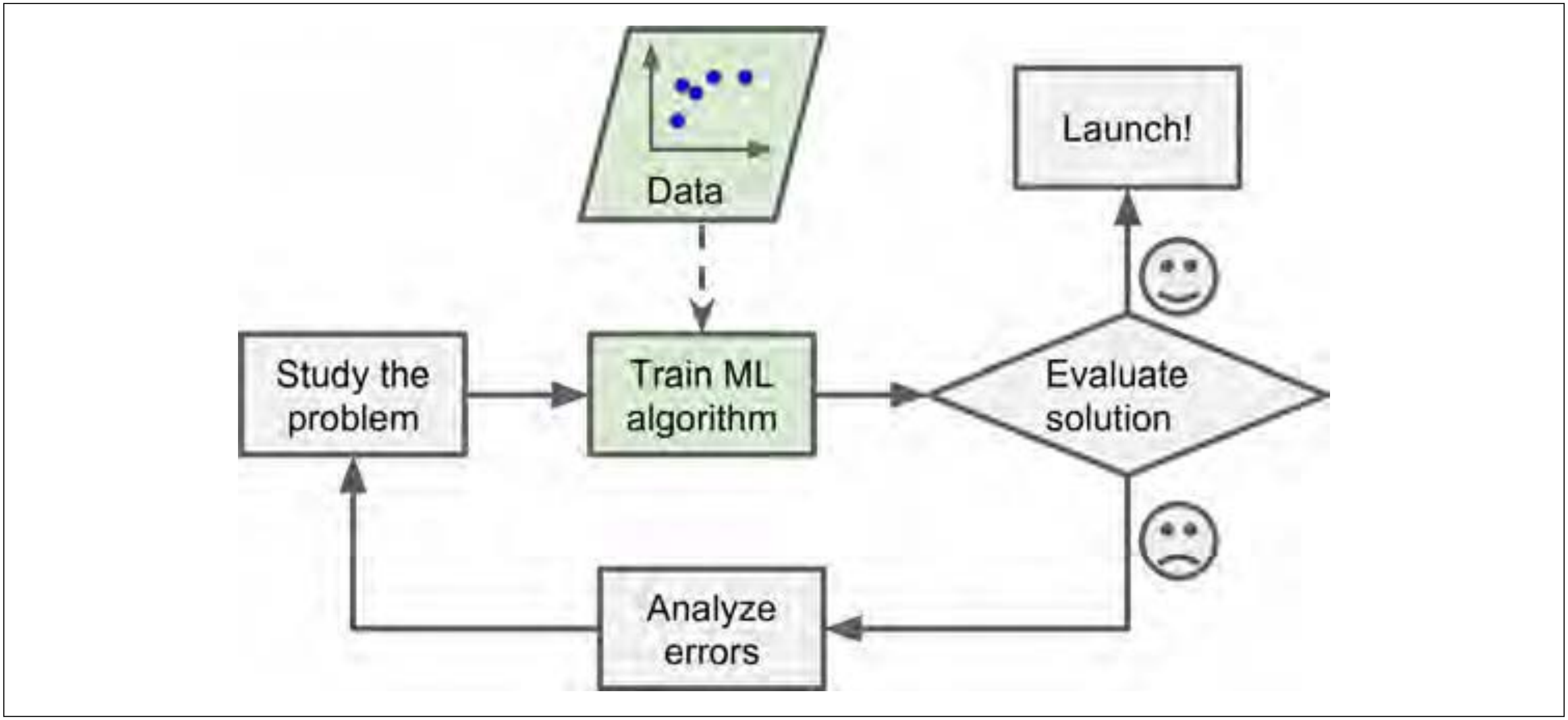


Figure 1-2. Machine Learning approach

Image from: Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O'Reilly Media, Inc.

Challenges

- “huge data sets require computationally tractable algorithms”
- “highly personal data raise the need for algorithms that minimize privacy effects”
- “the availability of huge quantities of unlabeled data raises the challenge of designing learning algorithms to take advantage of it”

Jordan, Michael I. and Tom M. Mitchell. (2015). “Machine Learning: Trends, perspectives, and prospects” *Science*.

Supervised Learning

- **Function approximation problem**
  - “the training data take the form of a collection of (x, y) pairs and the goal is to produce a prediction y* in response to a query x*”
  - Task is to learn a mapping, f(x), which outputs a y value for each inputted x value

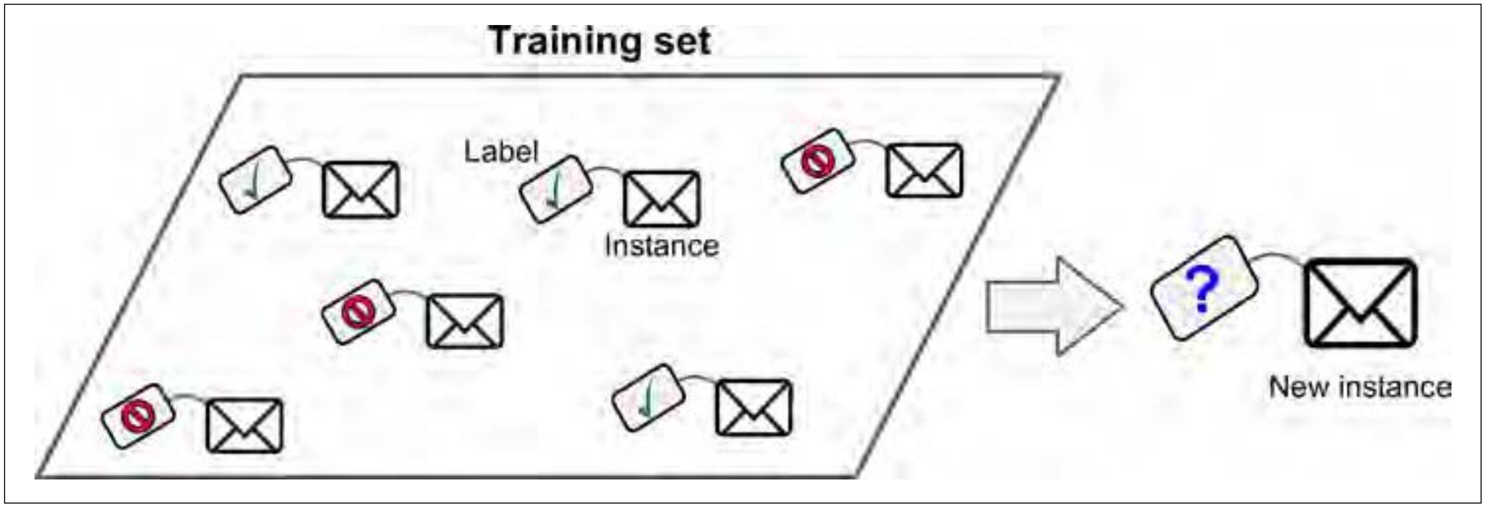


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

Jordan, Michael I. and Tom M. Mitchell. (2015). “Machine Learning: Trends, perspectives, and prospects” *Science*.
Image from: Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O’Reilly Media, Inc.

Supervised Learning

- ***k*-Nearest Neighbors**
- **Linear Regression**
- **Logistic Regression**
- **Support Vector Machines (SVMs)**
- **Decision Trees and Random Forests**
- **Naive Bayes Classifiers**
- **Neural networks**

Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O'Reilly Media, Inc.

Supervised Learning

- **“diversity of learning architectures and algorithms reflects the diverse needs of applications”**
  - “with different architectures capturing different kinds of mathematical structures, offering different levels of amenability to post-hoc visualization and explanation, and providing varying trade-offs between computational complexity, the amount of data, and performance.”

Jordan, Michael I. and Tom M. Mitchell. (2015). “Machine Learning: Trends, perspectives, and prospects” *Science*.

Unsupervised Learning

- **"the analysis of unlabeled data under assumptions about structural properties of the data (e.g., algebraic, combinatorial, or probabilistic)"**

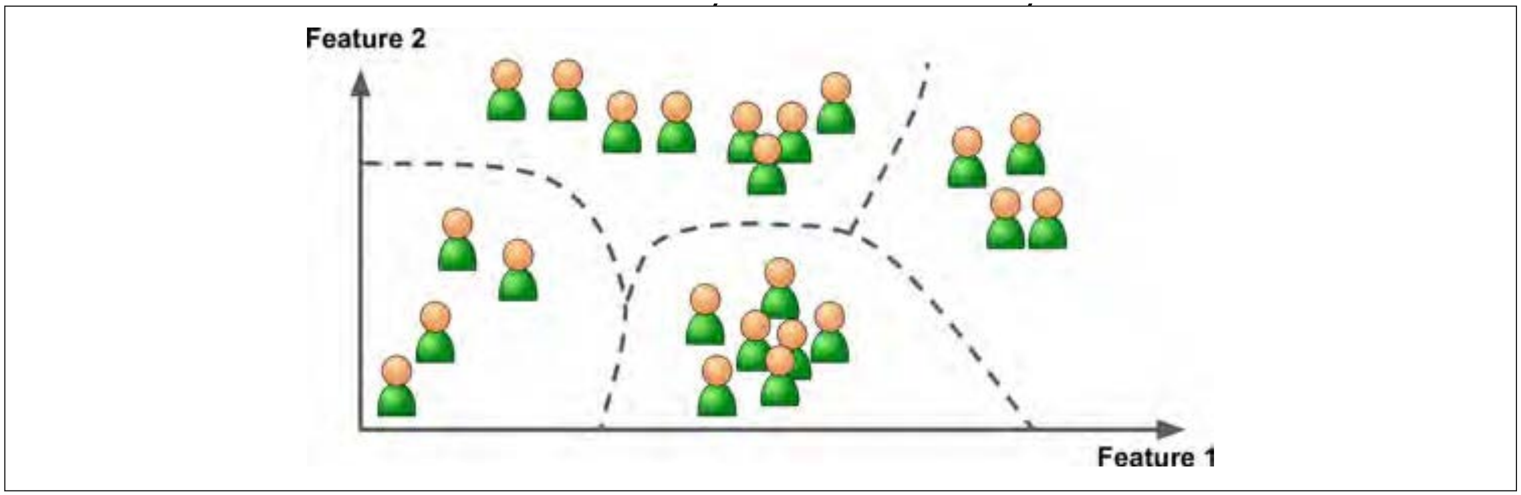


Figure 1-8. Clustering

Jordan, Michael I. and Tom M. Mitchell. (2015). "Machine Learning: Trends, perspectives, and prospects" *Science*.
Image from: Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O'Reilly Media, Inc.

Unsupervised Learning

- **The models make the assumption "that data lie on a low-dimensional manifold and aim to identify that manifold explicitly from the data"**
  - Dimensionality reduction (e.g., PCA)
  - Clustering (e.g., *k*-means)

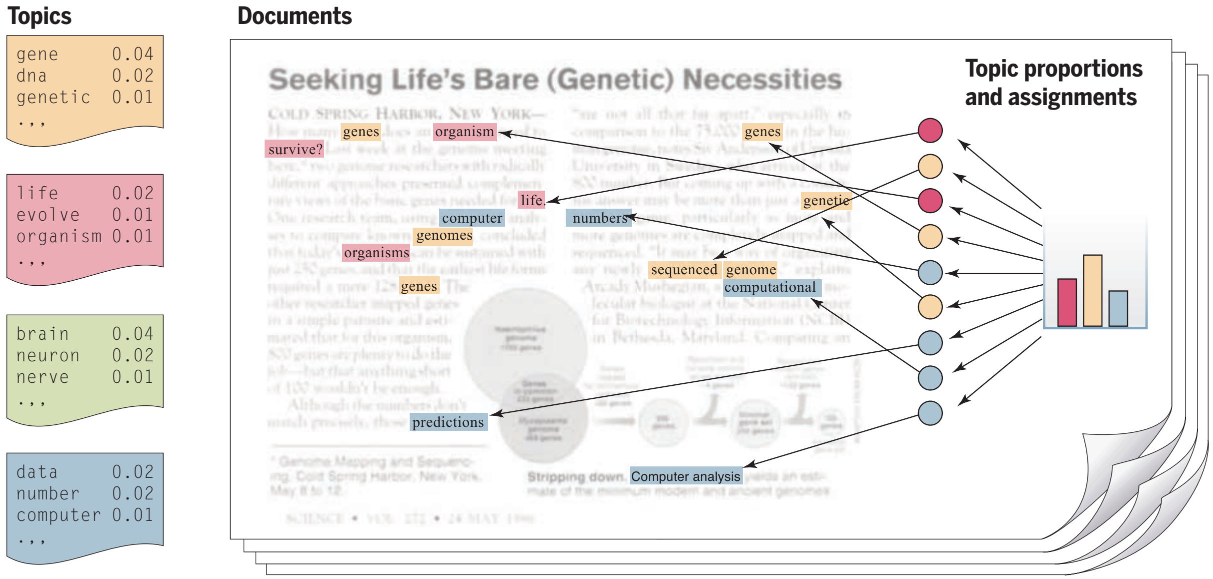


Fig. 3. Topic models. Topic modeling is a methodology for analyzing documents, where a document is viewed as a collection of words, and the words in the document are viewed as being generated by an underlying set of topics (denoted by the colors in the figure). Topics are probability distributions across words (leftmost column), and each document is characterized by a probability distribution across topics (histogram). These distributions are inferred based on the analysis of a collection of documents and can be viewed to classify, index, and summarize the content of documents. [From (*31*). Copyright 2012, Association for Computing Machinery, Inc. Reprinted with permission]

Text and image from Jordan, Michael I. and Tom M. Mitchell. (2015). "Machine Learning: Trends, perspectives, and prospects" *Science*.

Feature Engineering

- **Feature selection**
  - “selecting the most useful features to train on among existing features”
- **Feature extraction**
  - “combining existing features to produce a more useful one (as we saw earlier, dimensionality reduction algorithms can help)”
- **“Creating new features by gathering new data”**

Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O’Reilly Media, Inc.

Question Set 1

Géron (p. 4–9)

- **How would you define Machine Learning?**
- **Can you name four types of problems where it shines?**
- **What is a labeled training set?**

Question Set 1

- **How would you define Machine Learning?**
  - “Machine Learning is about building systems that can learn from data. Learning means getting better at some task, given some performance measure.”
- **Can you name four types of problems where it shines?**
  - “Machine Learning is great for complex problems for which we have no algorithmic solution, to replace long lists of hand-tuned rules, to build systems that adapt to fluctuating environments, and finally to help humans learn (e.g., data mining).”

Question Set 1

- **What is a labeled training set?**
  - “A labeled training set is a training set that contains the desired solution (a.k.a. a label) for each instance.”

Typical Machine Learning Project Steps

- **“You studied the data.”**
- **“You selected a model.”**
- **Feature Engineering**
- **“You trained it on the training data (i.e., the learning algorithm searched for the model parameter values that minimize a cost function).”**
- **“Finally, you applied the model to make predictions on new cases (this is called inference), hoping that this model will generalize well.”**

Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O’Reilly Media, Inc.

Main Challenges

- **Insufficient training data**
  - Quantity and/or quality and/or non-representative
- **Irrelevant features**
- **Overfitting training data**
- **Under-fitting training data**

Géron, Aurélien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* O'Reilly Media, Inc.