Advanced Data Analysis

DATA 71200

Class 2

Course Schedule

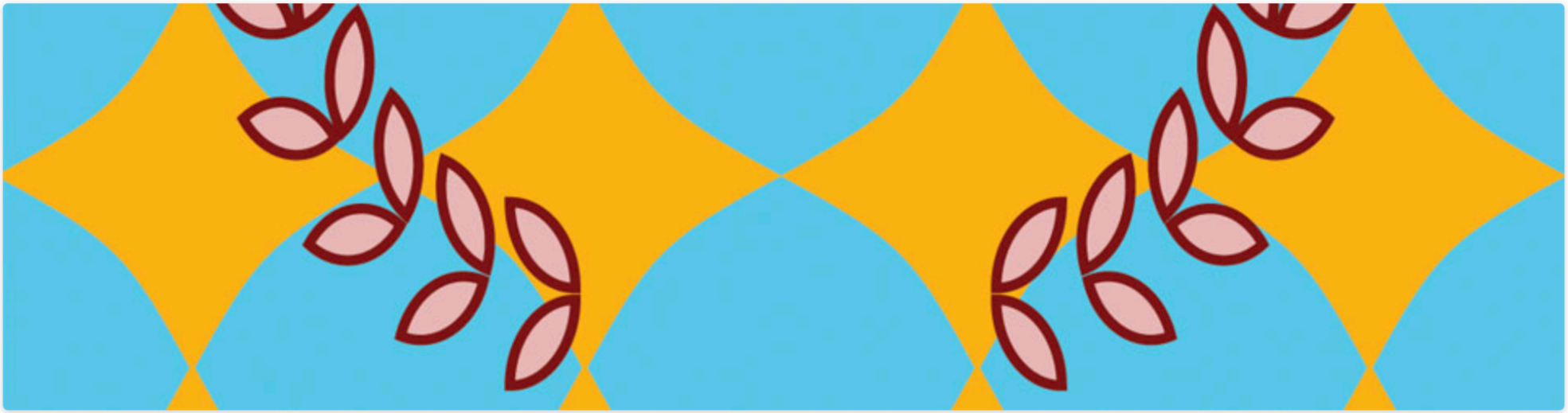
29-Jan	Introduction
5-Feb	**What is Machine Learning?**
12-Feb	No Class
19-Feb	Getting Started with Machine Learning
26-Feb	Inspecting Data
4-Mar	Representing Data

Class Website

DATA 71200 Advanced Data Analysis Methods

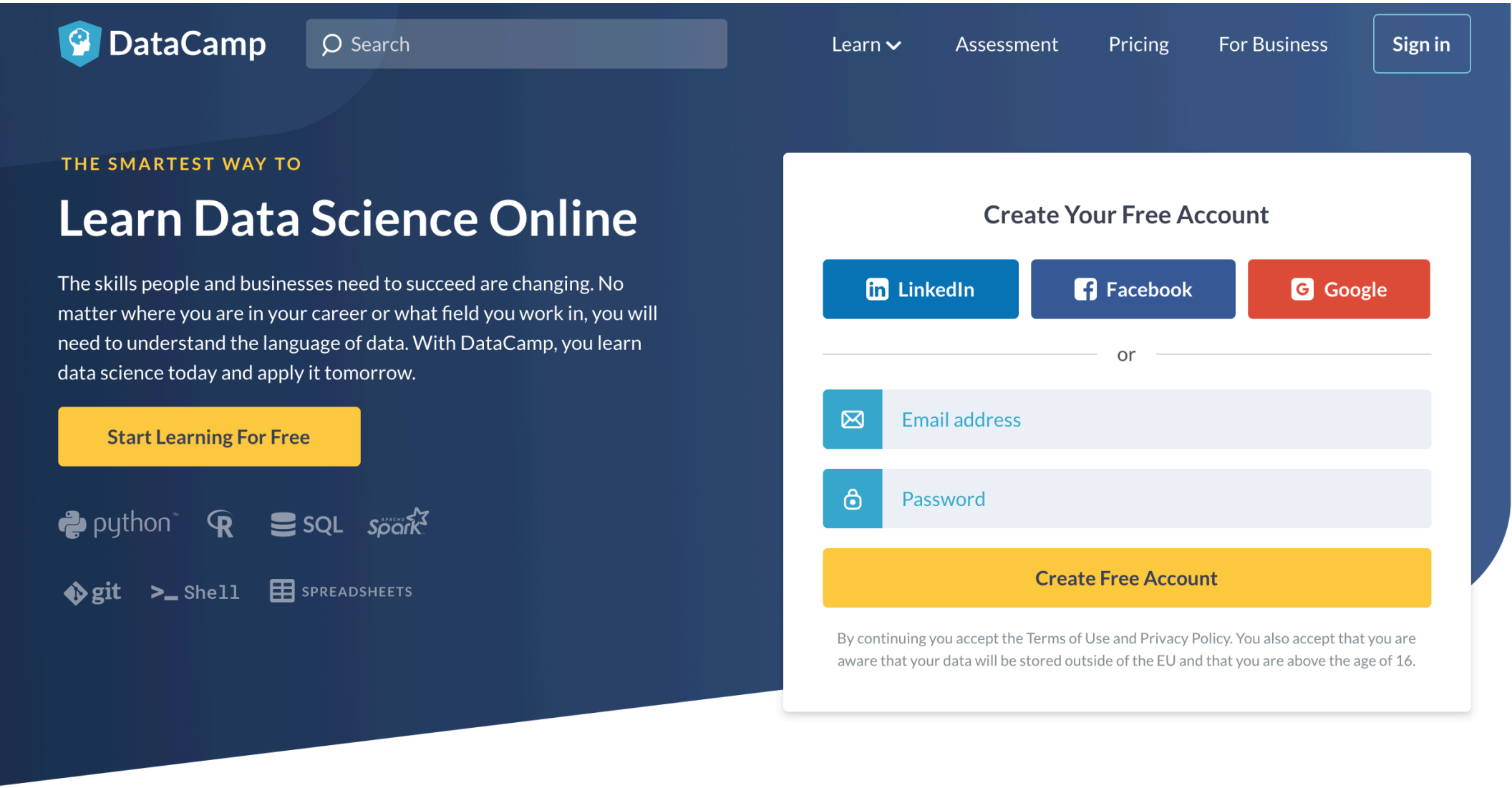
An introduction to supervised and unsupervised machine learning methods

HOME SYLLABUS COURSE SCHEDULE RESOURCES POSTS



https://data71200sp20.commons.gc.cuny.edu/

Data Camp


DataCamp
Search
Learn
Assessment
Pricing
For Business
Sign in
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.
Start Learning For Free
python
SQL
Spark
git
Shell
SPREADSHEETS
Create Your Free Account
LinkedIn
Facebook
Google
or
Email address
Password
Create 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.

GitHub

jcdevaney / **data71200sp20**

Unwatch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Actions | Projects 0 | Wiki | Security | Insights | Settings

DATA 71200 Advanced Data Analysis Methods

Edit

Manage topics

1 commit | **1** branch | **0** packages | **0** releases | **1** contributor

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

jcdevaney Initial commit — Latest commit `02f220c` 6 days ago

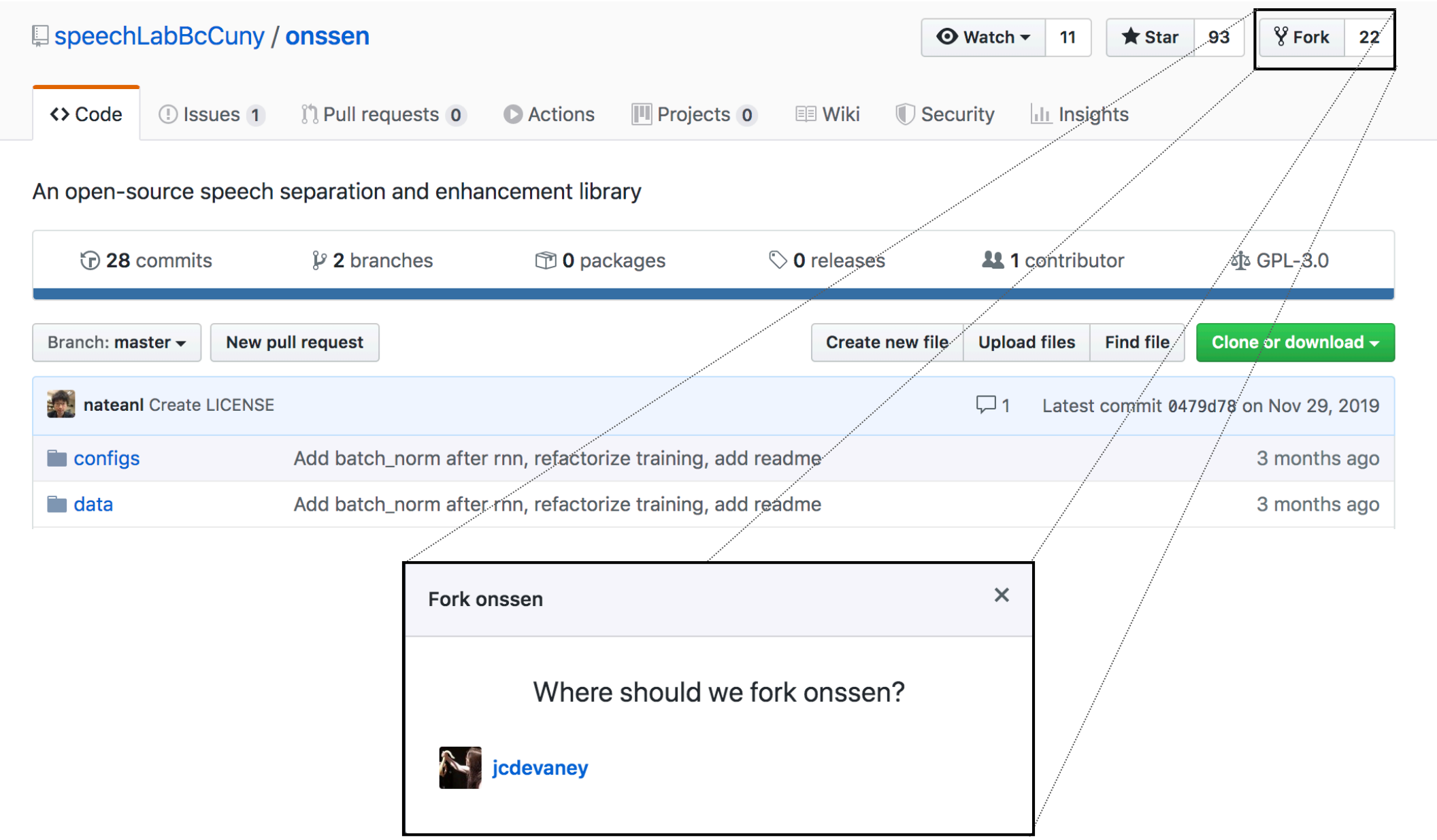
README.md — Initial commit — 6 days ago

README.md

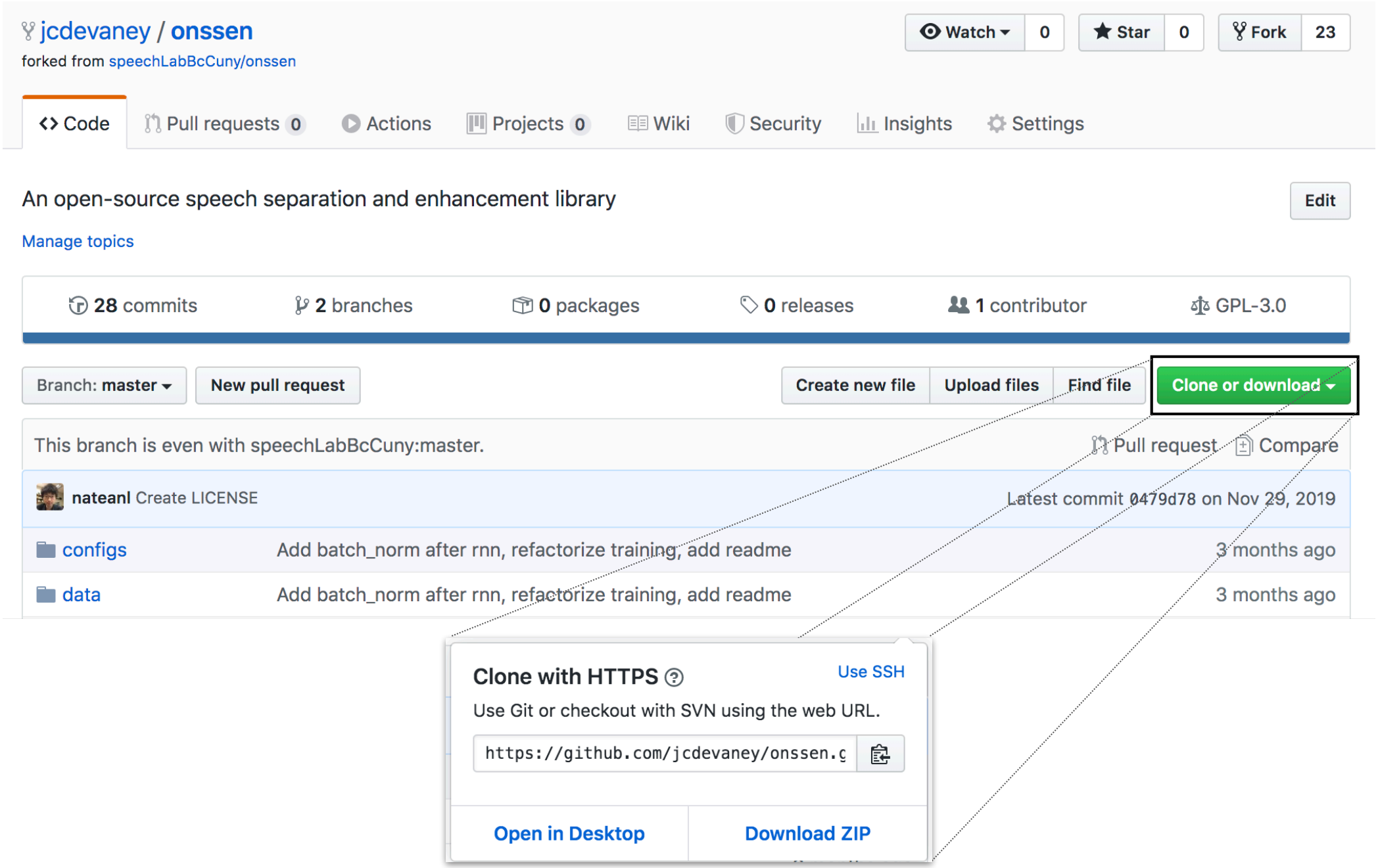
data71200sp20

DATA 71200 Advanced Data Analysis Methods

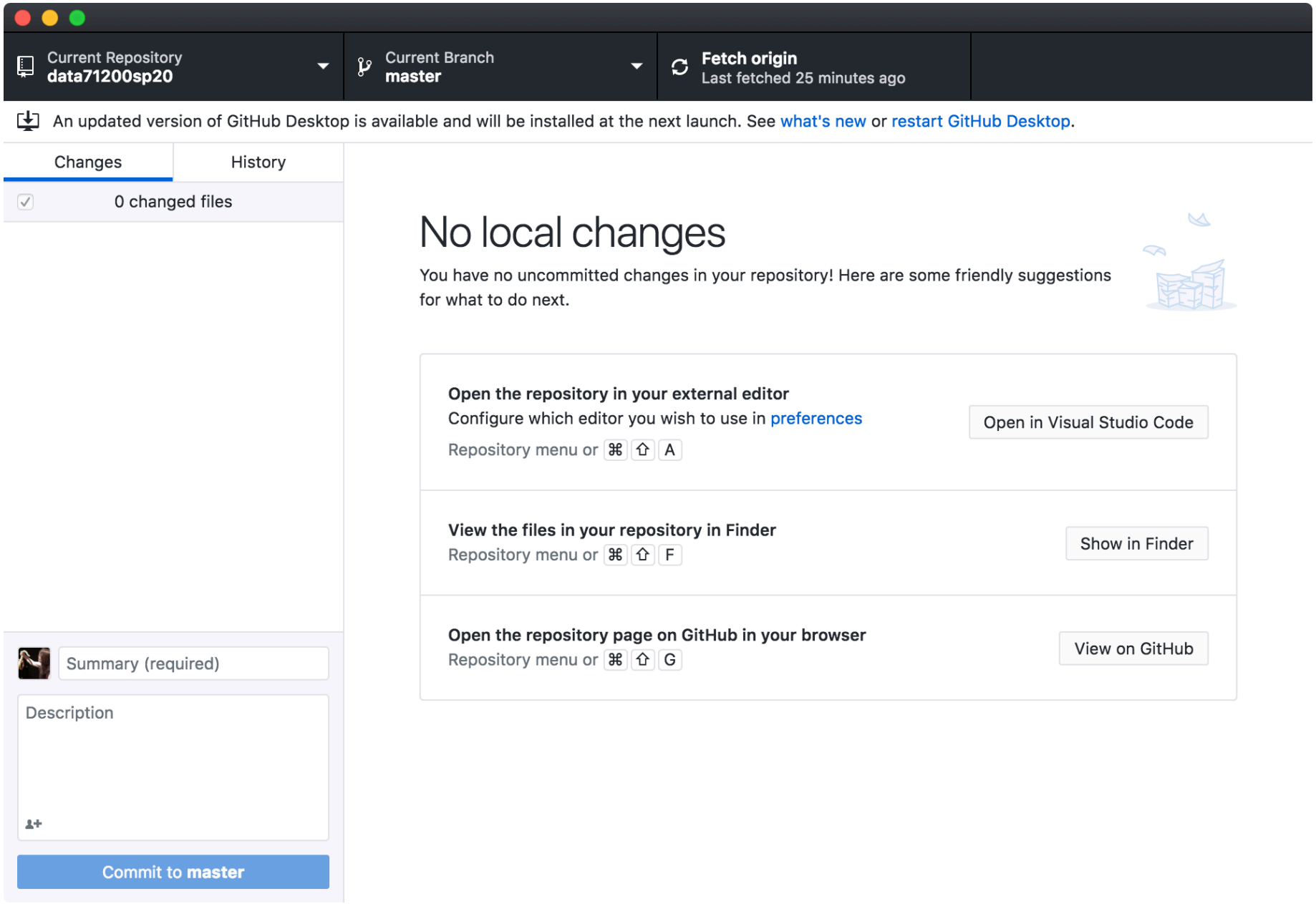
GitHub - Forking versus Cloning



GitHub - Forking versus Cloning

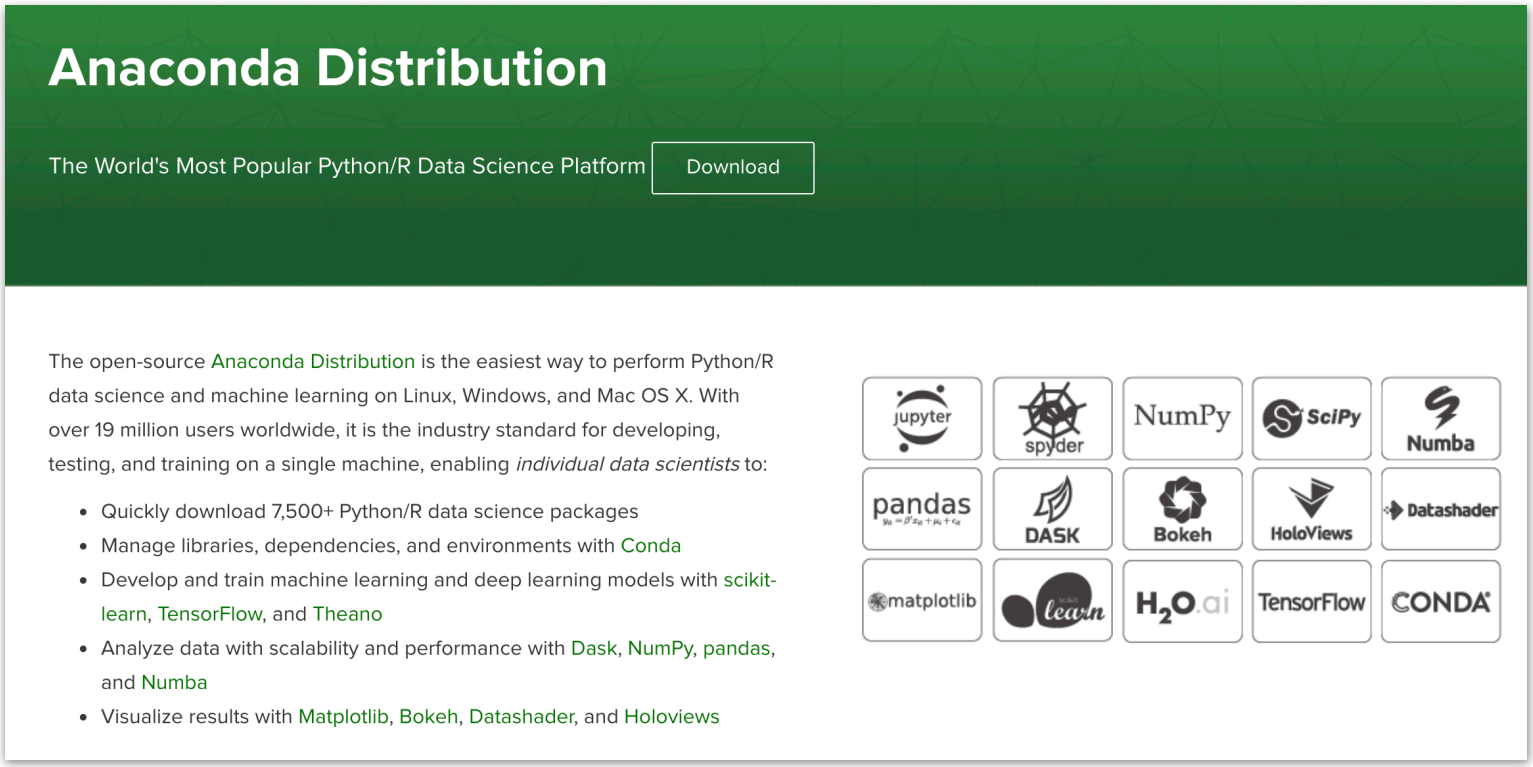


GitHub Desktop



Coding Environment

- **Python 3**
  - matplotlib, NumPy, Pandas, SciPy scikit learn
- **Jupytr notebooks**



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

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

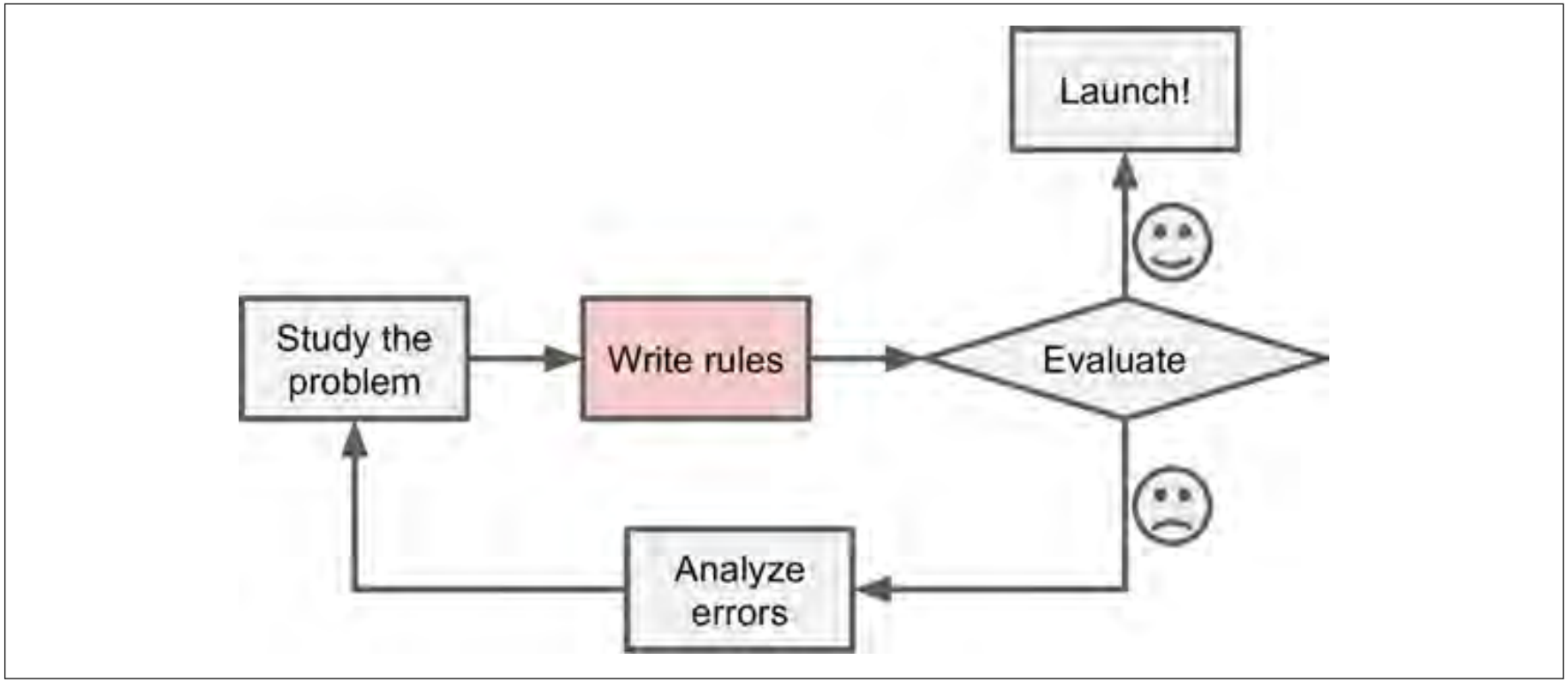


Figure 1-1. The traditional approach

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

Machine Learning vs Traditional Programming

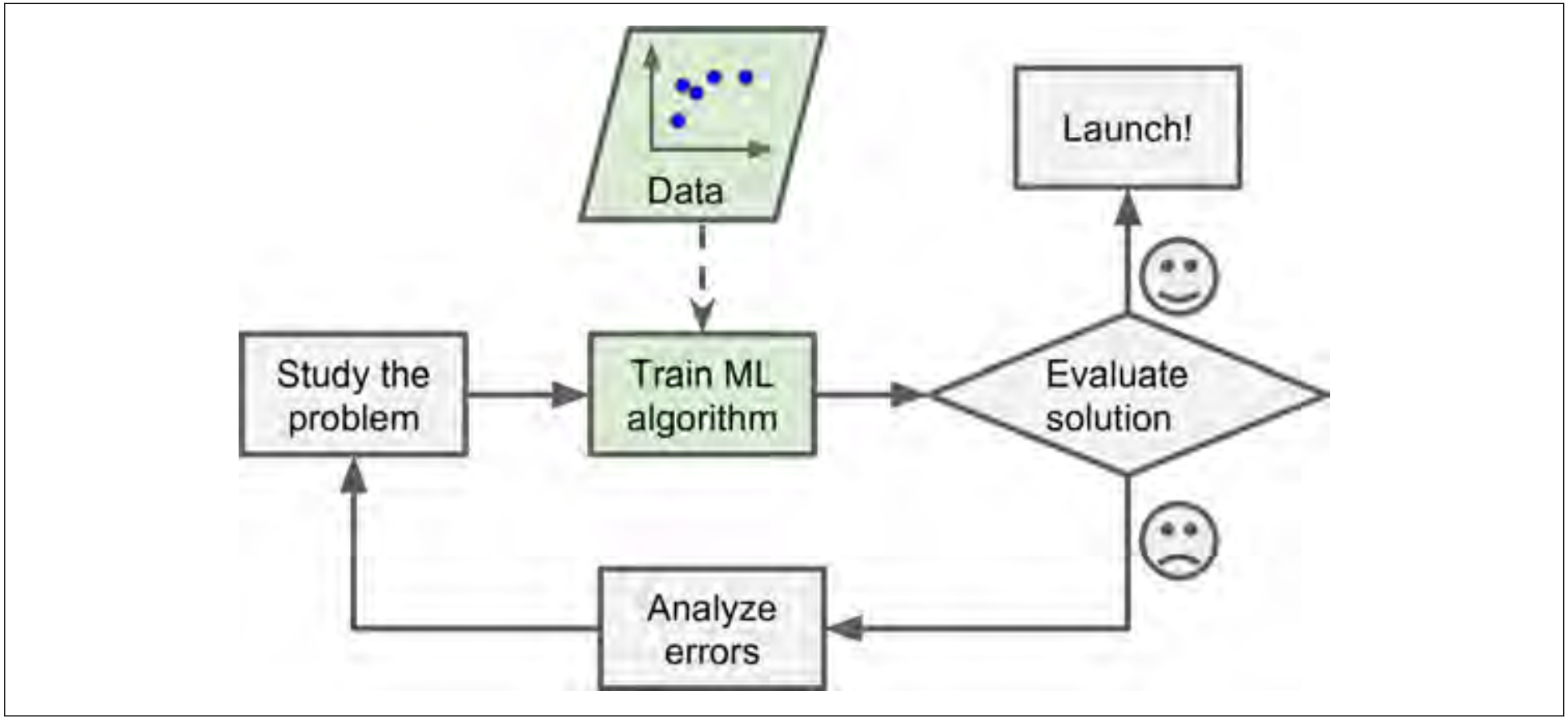


Figure 1-2. Machine Learning approach

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

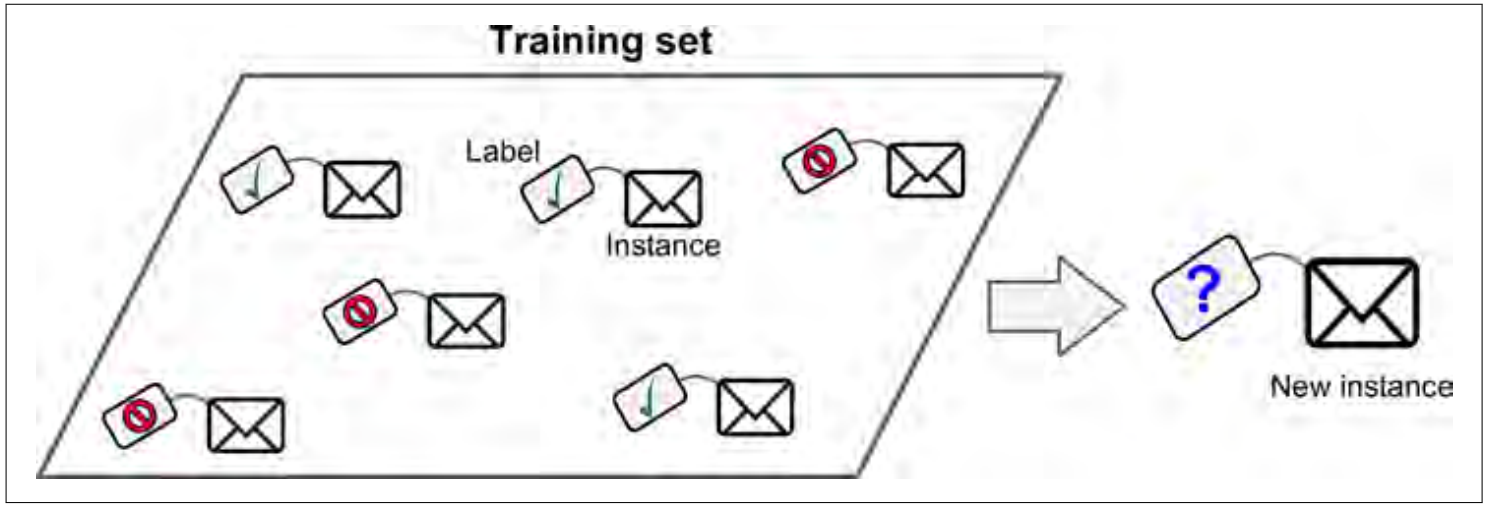


Figure 1-5. A labeled training set for supervised learning (e.g., spam classification)

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*.

Supervised Learning

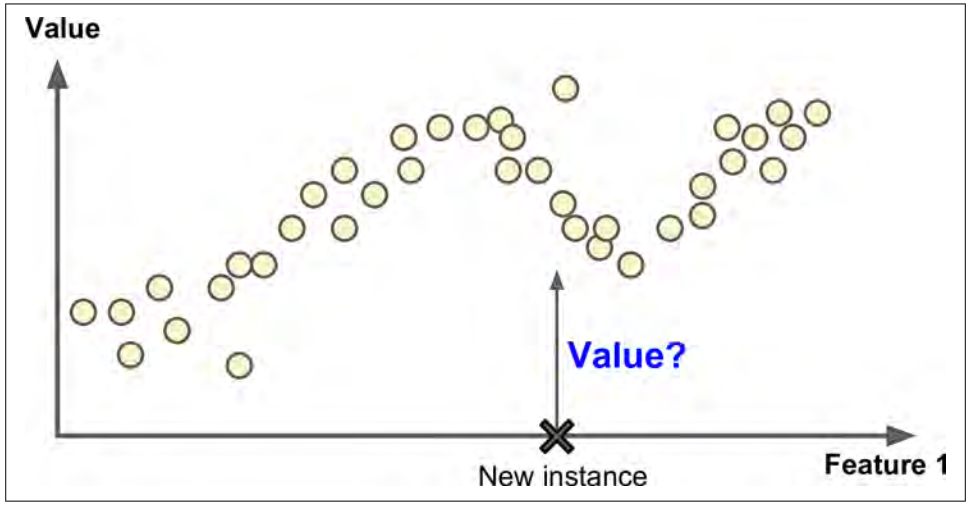


Figure 1-6. Regression

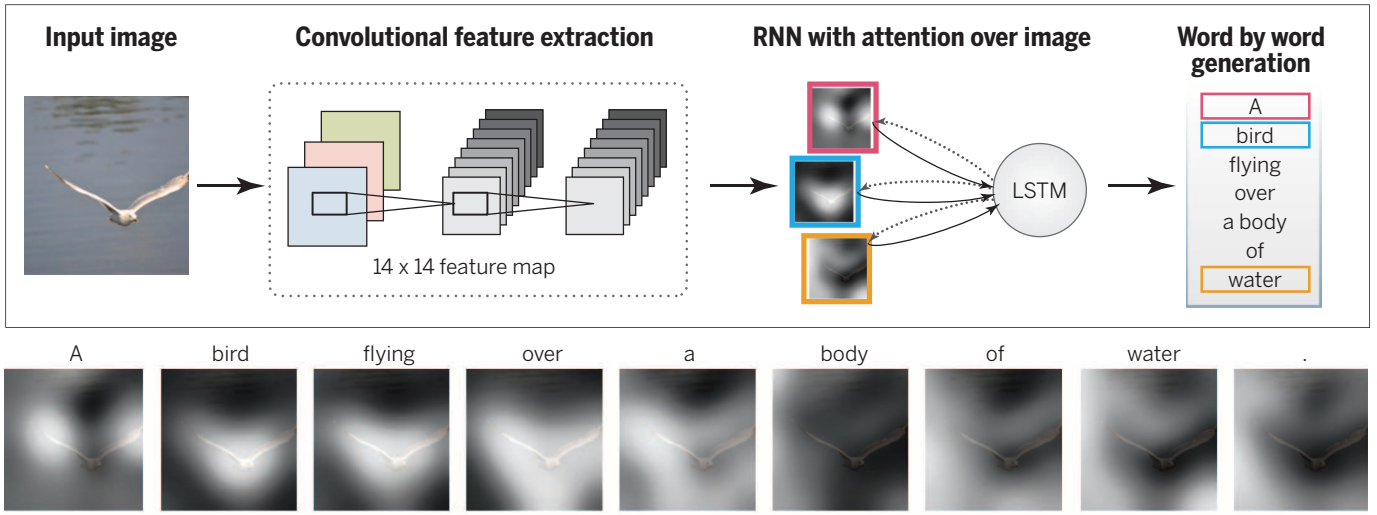


Fig. 2. Automatic generation of text captions for images with deep networks. A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top). The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word. [Adapted with permission from (*30*)]

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

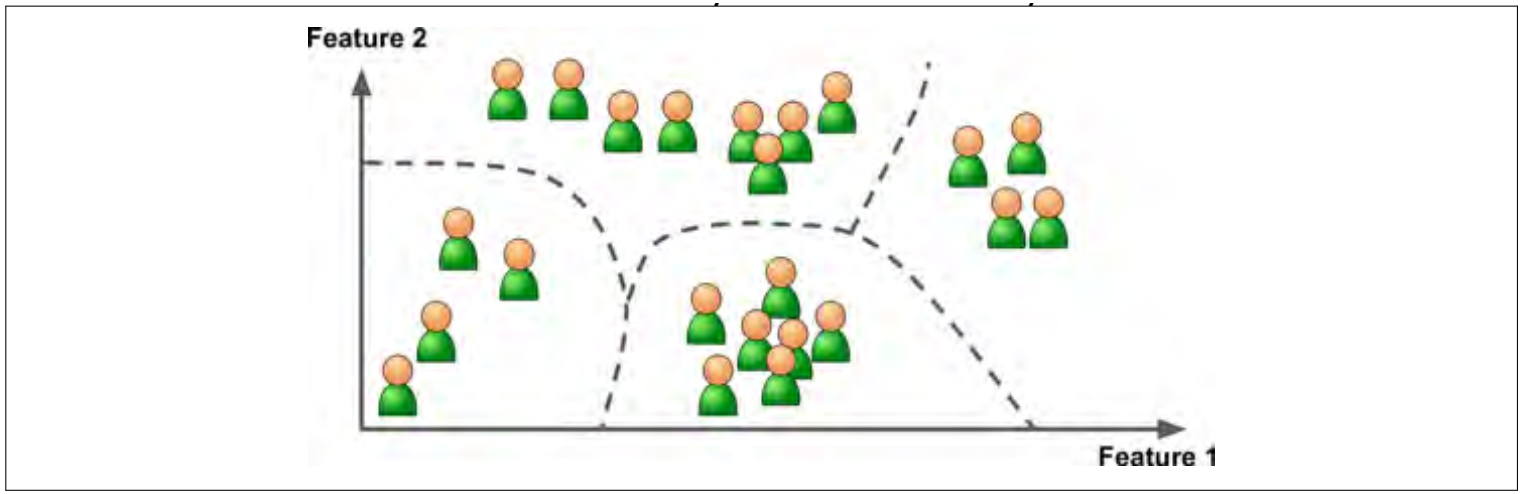


Figure 1-8. Clustering

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)

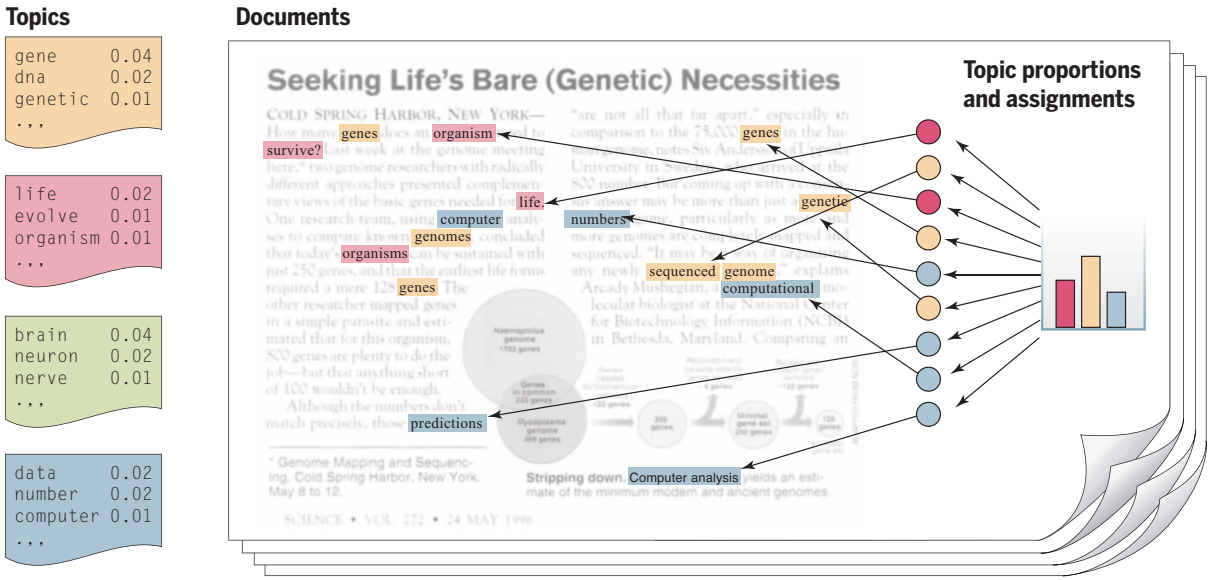


Fig. 3. Topic models. Topic modeling is a methodology for analyzing documents, where a document is viewed as a collection of words, and the words in the document are viewed as being generated by an underlying set of topics (denoted by the colors in the figure). Topics are probability distributions across 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*.

Semi-supervised Learning

- **“makes use of unlabeled data to augment labeled data in a supervised learning context, and discriminative training blends architectures developed for unsupervised learning with optimization formulations that make use of labels”**

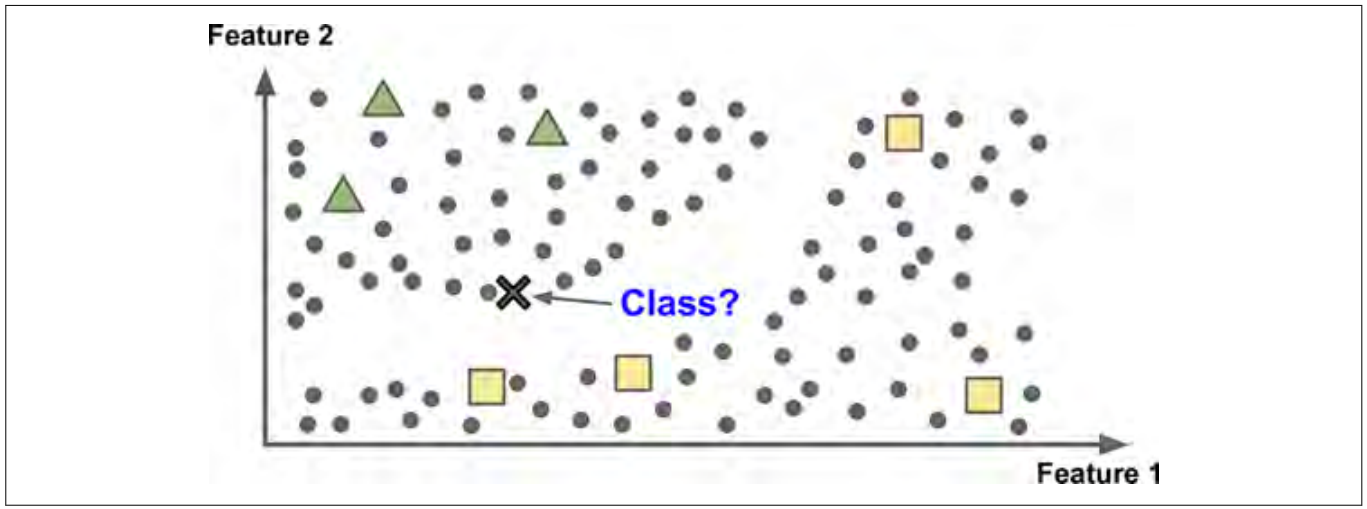


Figure 1-11. Semisupervised learning

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.

Reinforcement Learning

- **“Instead of training examples that indicate the correct output for a given input, the training data in reinforcement learning are assumed to provide only an indication as to whether an action is correct or not; if an action is incorrect, there remains the problem of finding the correct action.”**

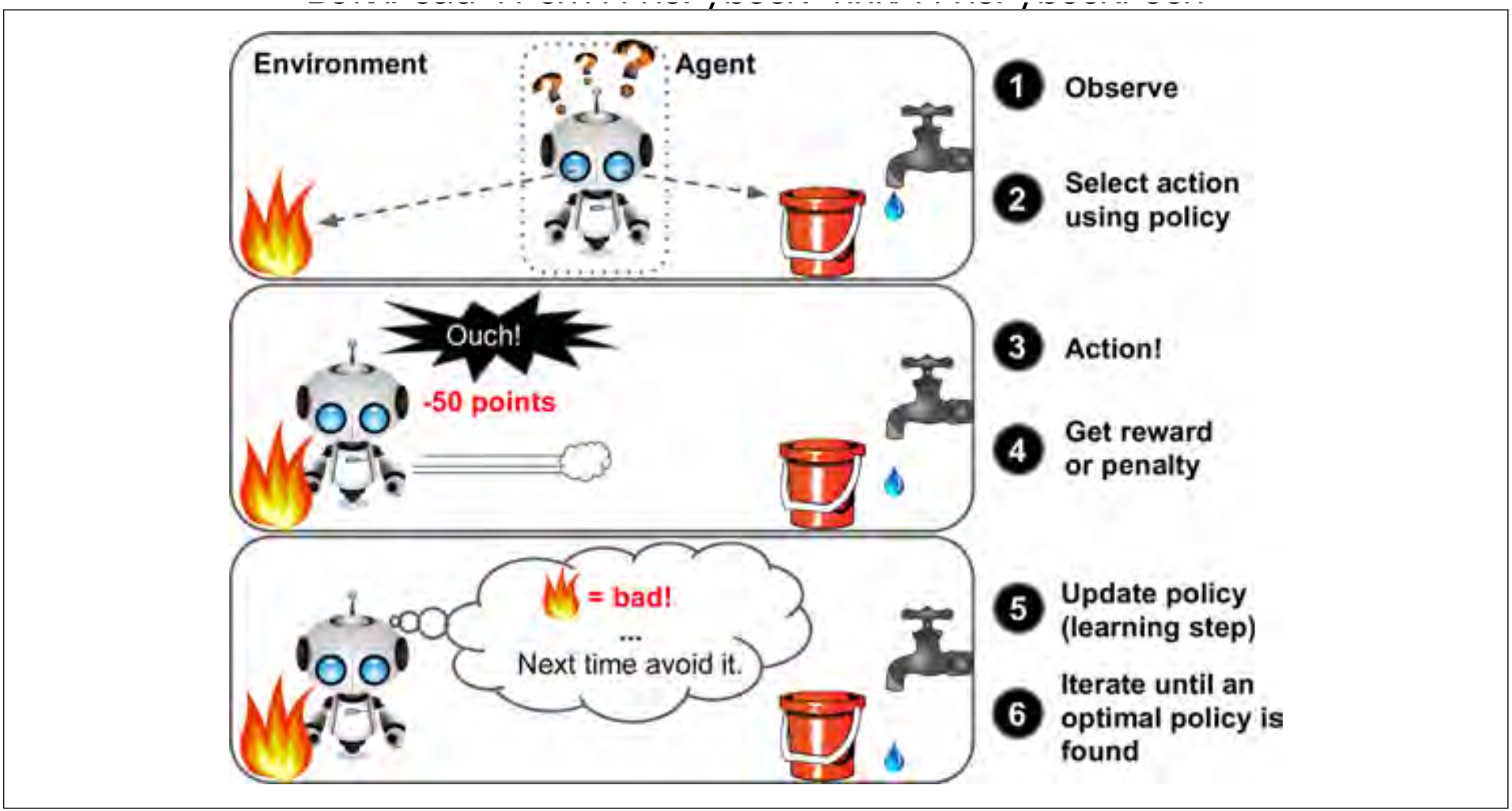
Jordan, Michael I. and Tom M. Mitchell. (2015). “Machine Learning: Trends, perspectives, and prospects” *Science*.

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.”**

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

Reinforcement Learning



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

Instance versus Model-Based Learning

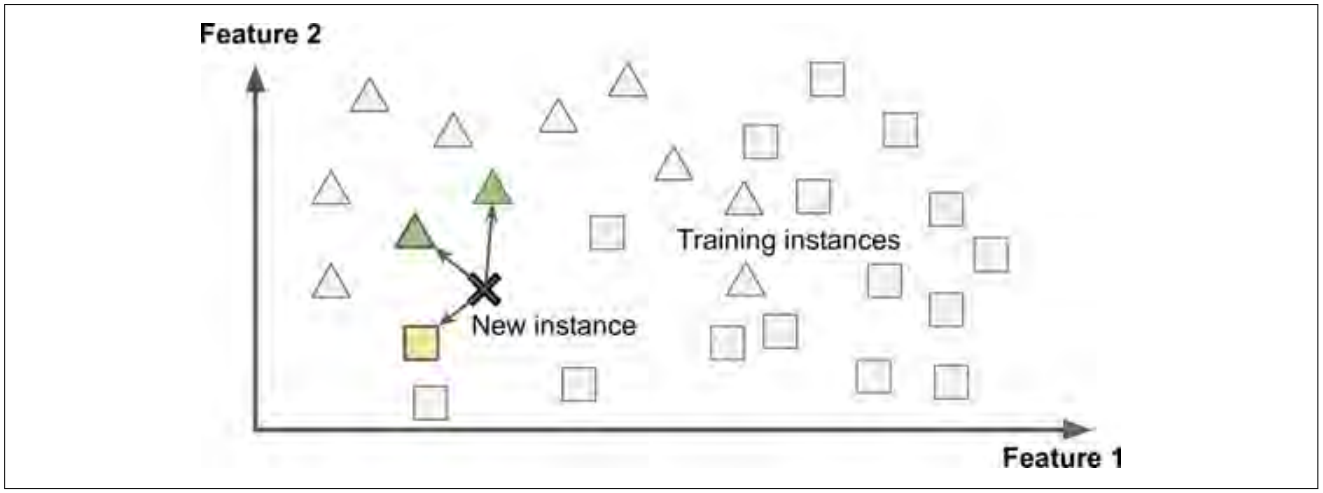


Figure 1-15. Instance-based learning

“the system learns the examples by heart, then generalizes to new cases using a similarity measure”

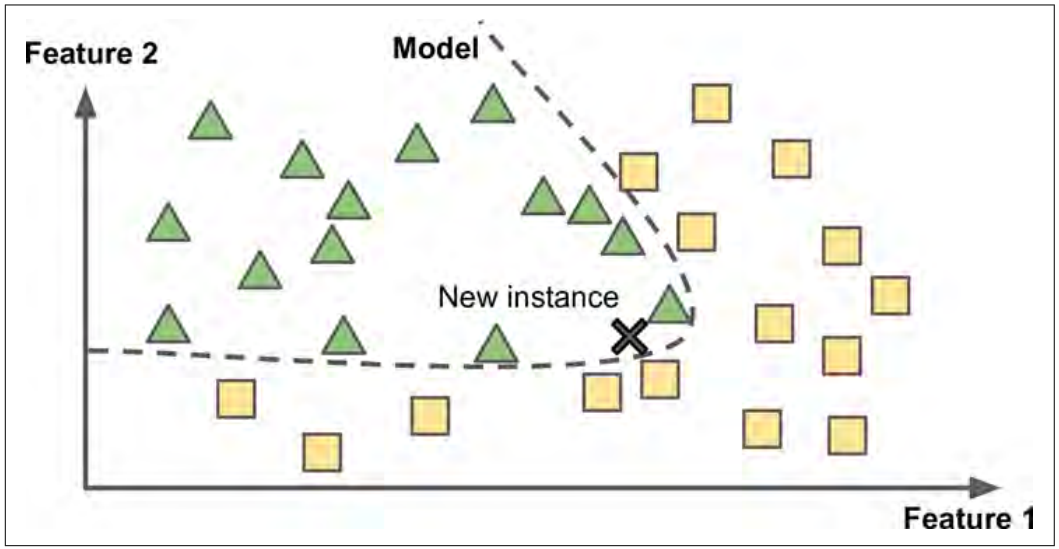


Figure 1-16. Model-based learning

"another way to generalize from a set of examples is to build a model of these examples, then use that model to make predictions”

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

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.

Review Questions 1

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

Review Questions 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)."

Review Questions 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.

Example: GDP and Happiness

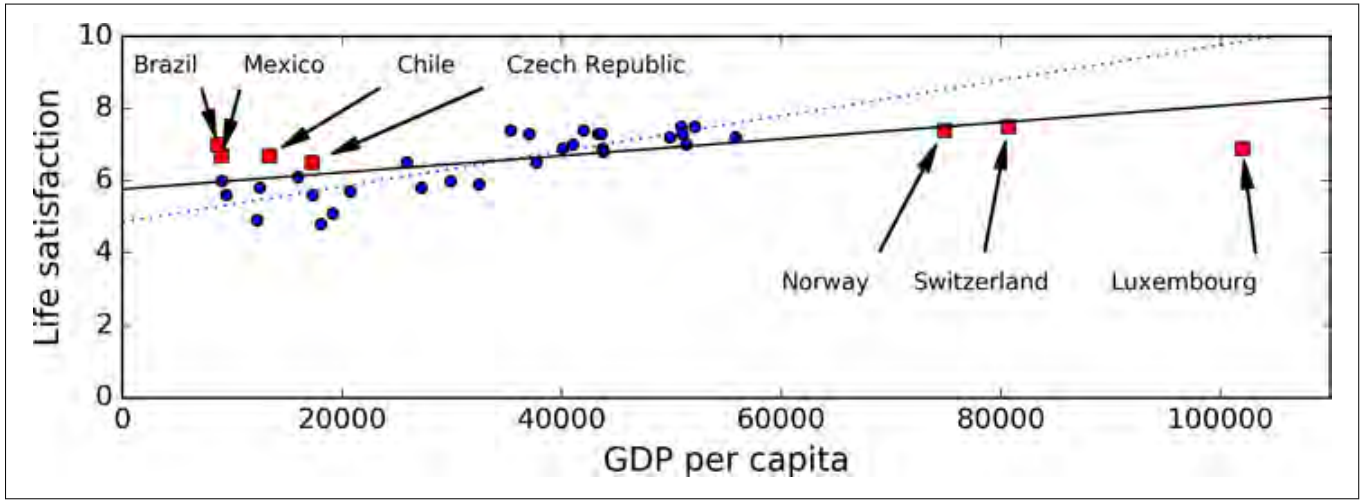


Figure 1-21. A more representative training sample

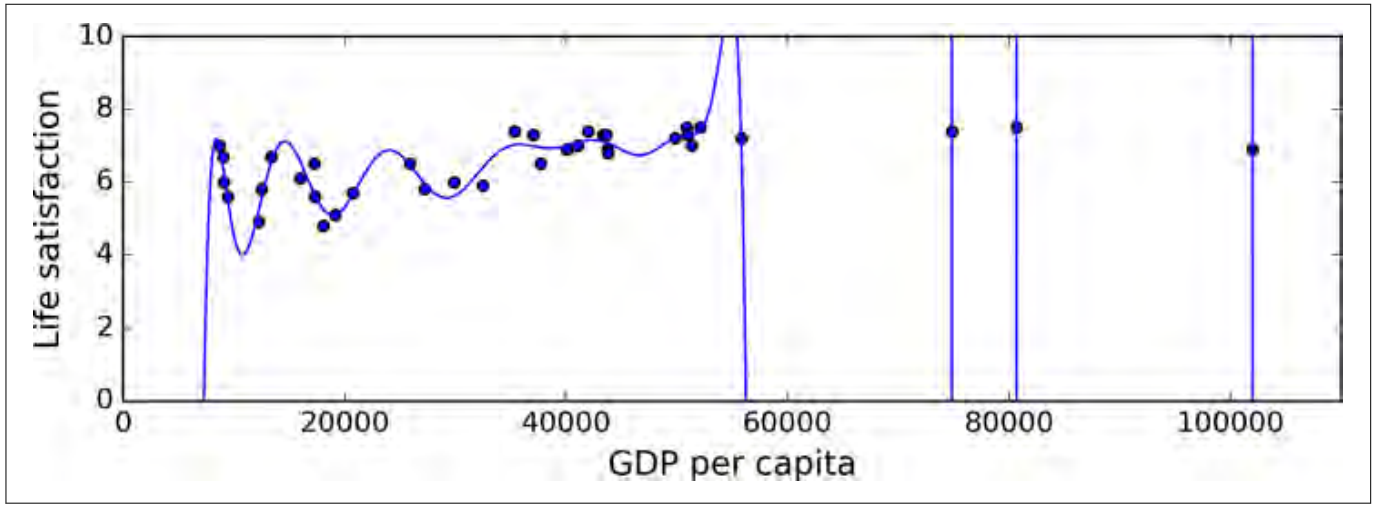


Figure 1-22. Overfitting the training data

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

Example: GDP and Happiness

- **regularization**
  - "constraining a model to make it simpler and reduce the risk of overfitting"
- **hyperparameter**
  - "amount of regularization to apply during learning"
  - "need too be set before training"

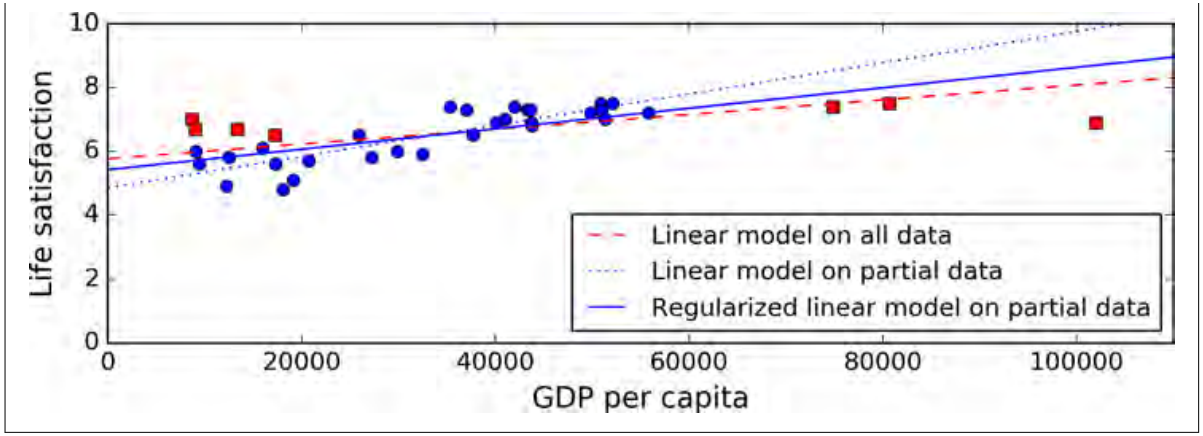


Figure 1-23. Regularization reduces the risk of overfitting

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

Training/Testing/Validation

- Training set - data used to train the model
- Testing set - hold out data used to estimate the generalization error on new data
- Validation set - used to compare models
- Cross-validation - iteratively holding out a subset of the data and testing on the rest (typically 80/20)

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

Data Analysis in Practice

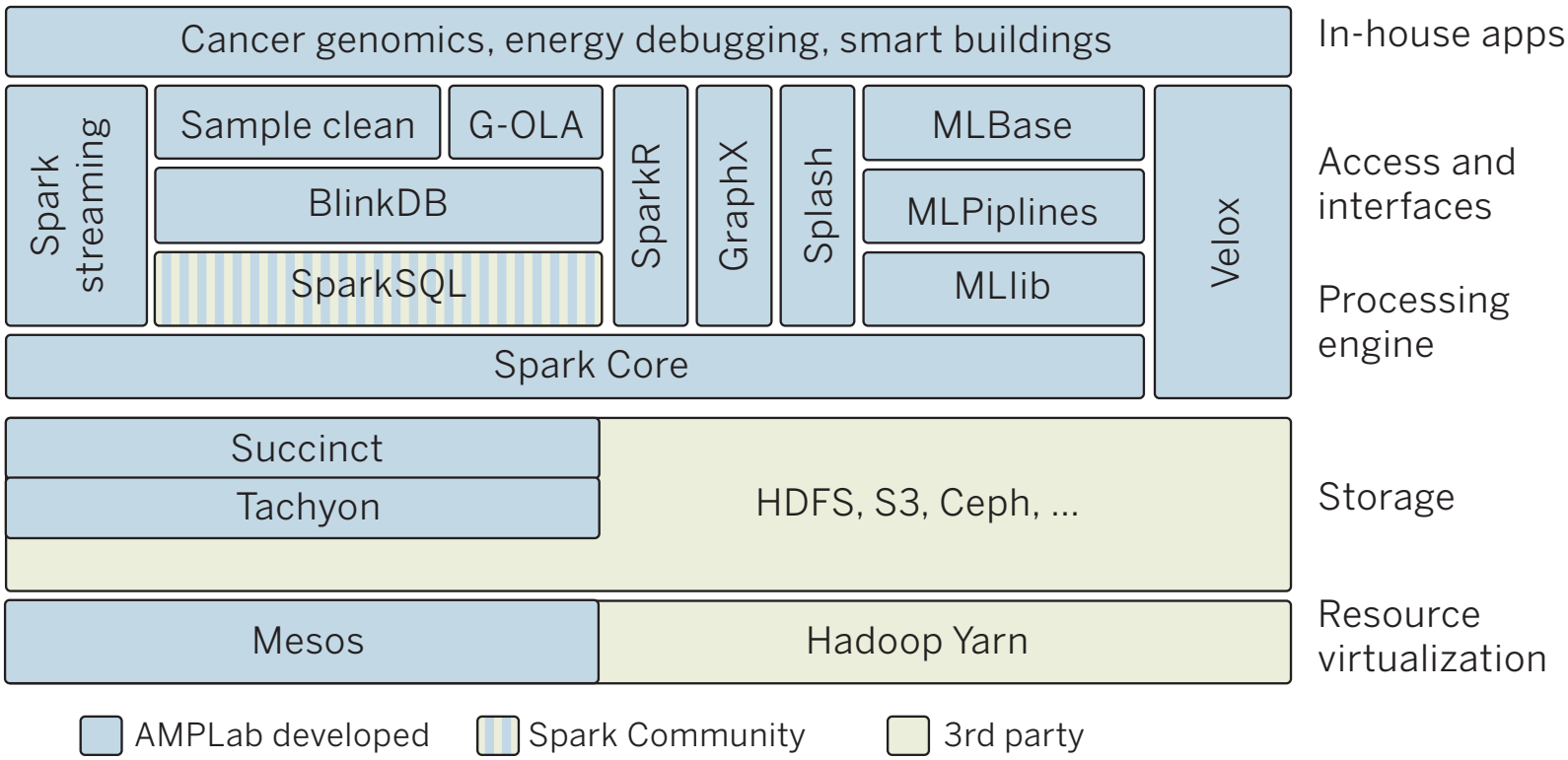


Fig. 5. Data analytics stack. Scalable machine-learning systems are layered architectures that are built on parallel and distributed computing platforms. The architecture depicted here—an open-source data analysis stack developed in the Algorithms, Machines and People (AMP) Laboratory at the University of California, Berkeley—includes layers that interface to underlying operating systems; layers that provide distributed storage, data management, and processing; and layers that provide core machine-learning competencies such as streaming, subsampling, pipelines, graph processing, and model serving.

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

Review Questions 2

- Can you name four of the main challenges in Machine Learning?
- If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?
- What is a test set and why would you want to use it?
- What is the purpose of a validation set?
- What can go wrong if you tune hyperparameters using the test set?
- What is cross-validation and why would you prefer it to a validation set?

Review Questions 2

- **Can you name four of the main challenges in Machine Learning?**
  - "Some of the main challenges in Machine Learning are the lack of data, poor data quality, non-representative data, uninformative features, excessively simple mod- els that underfit the training data, and excessively complex models that overfit the data."

Review Questions 2

- **If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?**
  - "If a model performs great on the training data but generalizes poorly to new instances, the model is likely overfitting the training data (or we got extremely lucky on the training data). Possible solutions to overfitting are getting more data, simplifying the model (selecting a simpler algorithm, reducing the number of parameters or features used, or regularizing the model), or reducing the noise in the training data."

Review Questions 2

- **What is a test set and why would you want to use it?**
  - "A test set is used to estimate the generalization error that a model will make on new instances, before the model is launched in production."
- **What is the purpose of a validation set?**
  - "A validation set is used to compare models. It makes it possible to select the best model and tune the hyperparameters."

Review Questions 2

- **What can go wrong if you tune hyperparameters using the test set?**
  - "If you tune hyperparameters using the test set, you risk overfitting the test set, and the generalization error you measure will be optimistic (you may launch a model that performs worse than you expect)."
- **What is cross-validation and why would you prefer it to a validation set?**
  - "Cross-validation is a technique that makes it possible to compare models (for model selection and hyperparameter tuning) without the need for a separate validation set. This saves precious training data."