Introduction to Tools for Data Science

Data Science Fundamental

Ratthaprom PROMKAM, Dr.rer.nat.

Department of Mathematics and Computer Science,
RMUTT

Learning Objectives

- Explore programming languages, tools, and data that data scientists use.
- Use open source tools to perform data science tasks.
- Discover IBM tools focused on data science.

List of modules

1. Language of Data Science
2. Data Science Tools
3. Packages, APIs, Datasets and Models
4. GitHub

Module 1

Language of Data Science

Module 1.1: Programming languages

Which language should I learn?

So many languages!

Your verdict ...

Python!


matplotlib


SciPy


scikit
learn


pandas


python™

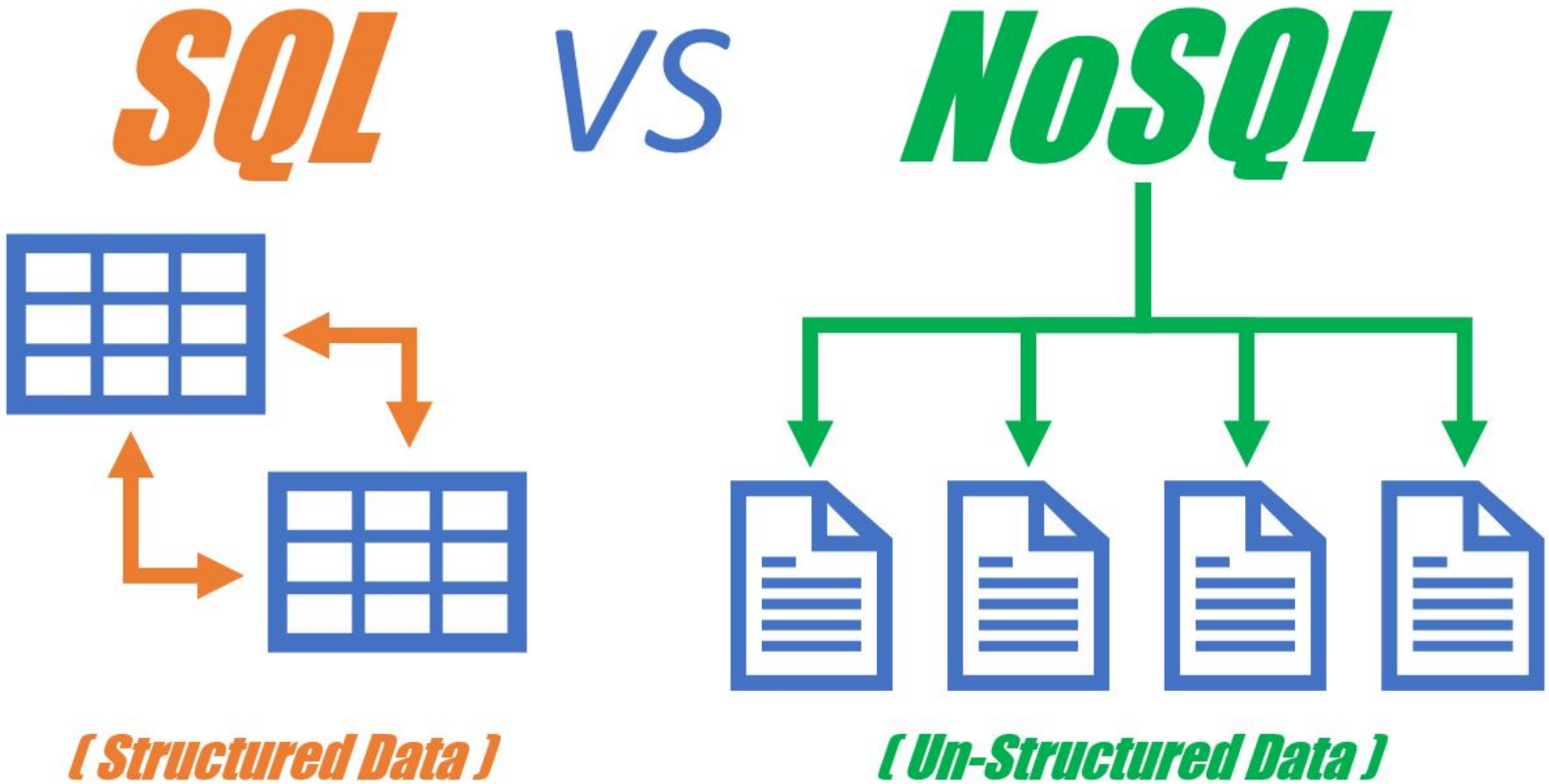

NumPy


K
Keras


GENSIM
topic modelling for humans


TensorFlow

What databases?



SQL Databases

NoSQL Databases

Software Types

	Free software	Open-source software	Freeware	Public-domain software
Definition	"FREE" is a matter of liberty, not price	"OPEN" doesn't just mean access to the source code	"FREE" refers to price, while freedom of the use is restricted by creator	"PUBLIC DOMAIN" belongs to the public as a whole
Ground philosophy	Social movement	Development methodology	Marketing goals	Copyright disclamation
Ground rules	Four Freedoms https://www.gnu.org/philosophy/free-sw.html	Open Software initiative https://opensource.org/osd		Creative Common Organization https://creativecommons.org
Free of charge	Not necessary	Not necessary	✓YES	✓YES
Covered by copyright law	✓YES	✓YES	✓YES	✗ NO
Examples	Linux, ubuntu	MySQL, Apache	Skype, Adobe Acrobat	SQLite

Software Types

	Copyright	Copyleft	Permissive	Creative Commons
What is a user allowed to do with the code?	What creator dictates	What user wants under certain rules	What user wants with a few restrictions	What user wants without restrictions
Clause of the use	As creator dictates	Derivative work must be attributed to creator, open-source and copyleft	Derivative work must be attributed to a creator	Derivative work must be attributed to a creator
Source code	As creator dictates	Must be open	Don't have to be open	No specific terms about the distribution of source code
Is creator liable for bugs?	✓ YES	✓ YES	✗ NO	✗ NO
Re-licensing	As creator dictates	Derivative work cannot be released as proprietary software	Derivative work can be released under another licen se or as proprietary software	Derivative work can be released under another licens e or as proprietary software
Commercial restrictions	As creator dictates	Permitted	Permitted	Permitted

Software Types

							
Type	**Permissive**	**Permissive**	**Permissive**	**Copyleft**	**Copyleft**	**Copyleft**
Provides copyright protection	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Can be used in commercial applications	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Provides an explicit patent license	✓ TRUE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE
Can be used in proprietary (closed source) projects	✓ TRUE	✓ TRUE	✓ TRUE	✗ FALSE	✗ FALSE partially	✗ FALSE for web
Popular open-source and free projects	Kubernetes Swift Firebase	Django React Flutter	Angular.js JQuery, .NET Core Laravel	Joomla Notepad++ MySQL	Qt SharpDevelop	SugarCRM Launchpad

Module 1.2: Review questions

Review questions

Which of the following statements is true?

- ○ 80% of data scientists worldwide use Python.
- ○ Python is the most popular language in data science.
- ○ Keras, Scikit-learn, Matplotlib, Pandas, and TensorFlow are all built with Python.
- ○ Python is useful for AI, machine learning, web development, and IoT.
- ○ All of the above

Review questions

Which of the following are SQL databases? (Select all that apply.)

- ○ MongoDB
- ○ MariaDB
- ○ MySQL
- ○ PostgreSQL
- ○ CouchDB
- ○ Oracle

Review questions

Which statements are true about Open Source and Free Software? (Select all that apply.)

- ○ Free Software and Open Source can be used interchangeably.
- ○ Free Software can always be run, studied, modified and redistributed with or without changes.
- ○ Most of Free Software licenses also qualify for Open Source.
- ○ Open Source Software can be modified without sharing the modified source code depending on the Open Source license.

Review questions

Is the following statement true or false:
"R integrates well with other computer languages like C++, Java, C, .Net and Python."

- ○ True
- ○ False

Review questions

Which of the following languages can be used for data science?

- ○ R
- ○ SQL
- ○ Java
- ○ Scala
- ○ Julia
- ○ Javascript
- ○ All of the above

Review questions

Which of the following is used to make Artificial intelligence and Machine Learning possible? (Select all that apply.)

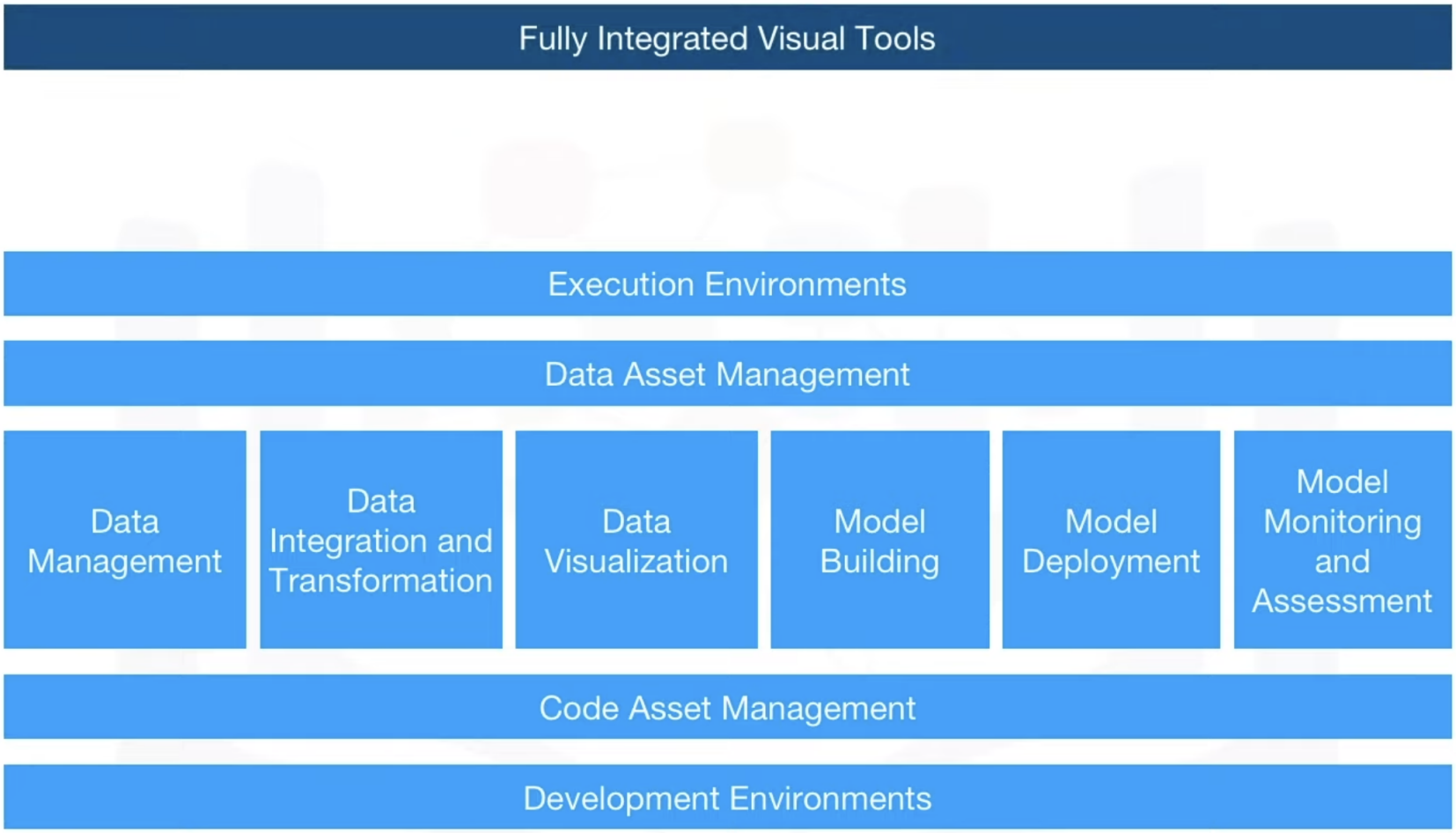
- ○ Oracle
- ○ PyTorch
- ○ TensorFlow.js
- ○ Apache Spark
- ○ GNU
- ○ Caffe

Module 2

Data Science Tools

Module 2.1: Categories of Data Science Tools

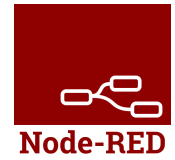
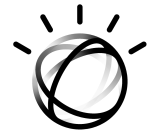
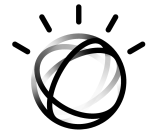
Categories of Data Science Tools



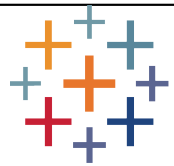
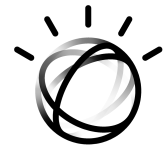
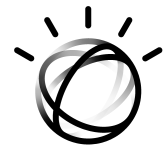
Data management

Open Source	        
Commercial	   
Cloud	    

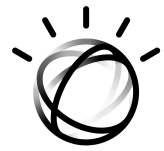
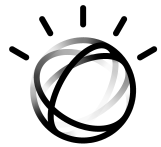
Data integration and transformation

Open Source	 Apache Airflow     Kubeflow 
Commercial	 Informatica  IBM Watson  
Cloud	 Informatica  IBM Watson

Data visualization

Open Source	 kibana  Apache Superset™ 
Commercial	 +ableau®  Power BI  IBM Cognos Analytics  IBM Watson®
Cloud	 Datameer  IBM Cognos Analytics  IBM Watson®

Model building

Open Source	          
Commercial	    
Cloud	  

Model deployment

Open Source	 kubernetes  OPENSHIFT  mleap  PredictionIO  SELDON  TensorFlow Serving
Commercial	SPSS®
Cloud	 Hugging Face amazon web services™ EC2

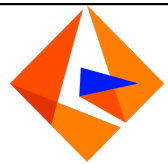
Model monitoring and assessment

Open Source	Prometheus VertaAI/**modeldb** Open Source ML Model Versioning, Metadata, and Experiment Management  Adversarial Robustness Toolbox AI Fairness 360 AI Explainability 360
Commercial	
Cloud	IBM **Watson**® **OpenScale** Amazon SageMaker

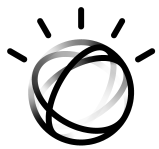
Code asset management

Open Source	
Commercial	
Cloud	  

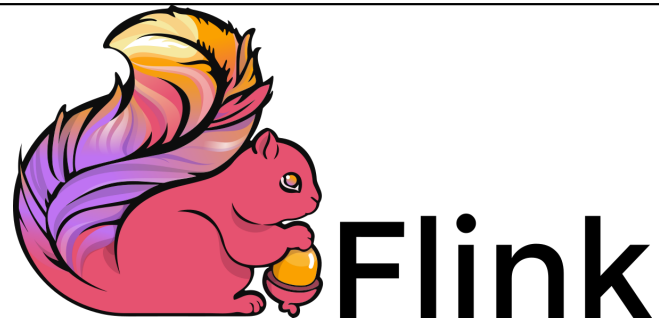
Data asset management

Open Source	 Apache Atlas kylo  ODPi EGERIA
Commercial	
Cloud	 Informatica IBM Infosphere

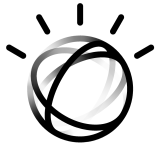
Development environment tools

Open Source	   
Commercial	
Cloud	 IBM **Watson**®

Execution environments

Open Source	  
Commercial	
Cloud	 **Hugging Face**

Fully integrated visual tools

Open Source	 Open for Innovation KNIME
Commercial	
Cloud	  IBM **Watson**®

Module 2.2: Review questions

Review questions

Which of the following are common tasks in data science?

- ○ Model Building
- ○ Data Integration and Transformation
- ○ Model Deployment
- ○ Data Visualization
- ○ Data Management
- ○ Model Monitoring and Assessment
- ○ All of the above

Review questions

Which of the following are data management tools? (Select all that apply.)

- ○ GitHub
- ○ MySQL
- ○ PostgreSQL
- ○ KubeFlow
- ○ PixieDust

Review questions

Which of the following are data management tools? (Select all that apply.)

- ○ GitHub
- ○ MySQL
- ○ PostgreSQL
- ○ KubeFlow
- ○ PixieDust

Review questions

Which of the following are Data Integration and Transformation tools? (Select all that apply.)

- ○ Cassandra
- ○ Apache Kafka
- ○ Apache Nifi
- ○ Apache AirFlow
- ○ Ceph

Review questions

Which statement about JupyterLab is correct?

- ○ JuypterLab can run Python code only.
- ○ JuypterLab can run R code only.
- ○ JuypterLab can run R and Python code only.
- ○ JuypterLab can run R and Python code in addition to other programming languages.

Review questions

Which statement about RStudio is correct?

- ○ RStudio is the primary choice for development in the Python programming language.
- ○ RStudio is the primary choice for development in the R programming language.
- ○ RStudio is the primary choice for web development.

Review questions

Which statements about IBM Watson Studio and OpenScale are correct? (Select all that apply.)

- ○ Watson Studio together with Watson OpenScale is a database management system.
- ○ Watson Studio together with Watson OpenScale covers the complete development life cycle for all data science, machine learning and AI tasks.
- ○ Watson Studio together with Watson OpenScale is available as a Cloud offering as well as a package running on top of Kubernetes/RedHat OpenShift in a local data center called IBM Cloud Pak for Data.

Module 3

Packages, APIs, Datasets and Models

Module 3.1: Libraries for Data Science

Scientific computing libraries in Python

(*n*-dimensional arrays)

(Data structures and tools)

Visualization libraries in Python

(Plots & graphs, most poplular)

(Plots: heat maps, time series, violin plots)

Machine learning and deep learning libraries in Python

(Machine learning:
regression, classification, ...)

(Deep learning:
neural networks, ...)

Deep learning libraries in Python

TensorFlow

(Deep learning:
production & deployment)

PyTorch

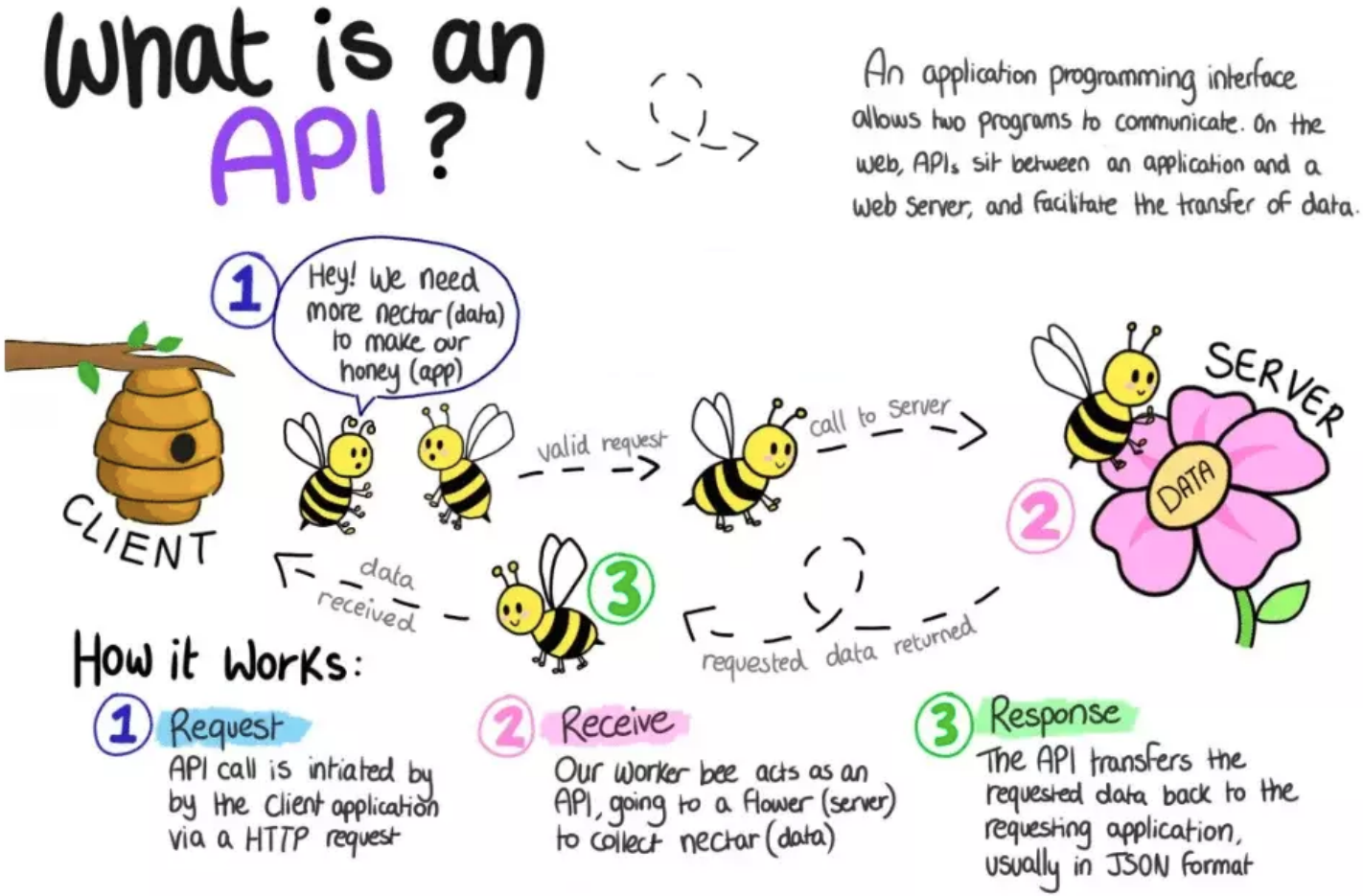
(Deep learning:
regression, classification, ...)

Computing Clusters

Apache Spark is a general-purpose cluster-computing framework that enables processing data using clusters.

Module 3.2: Application Programming Interfaces (API)

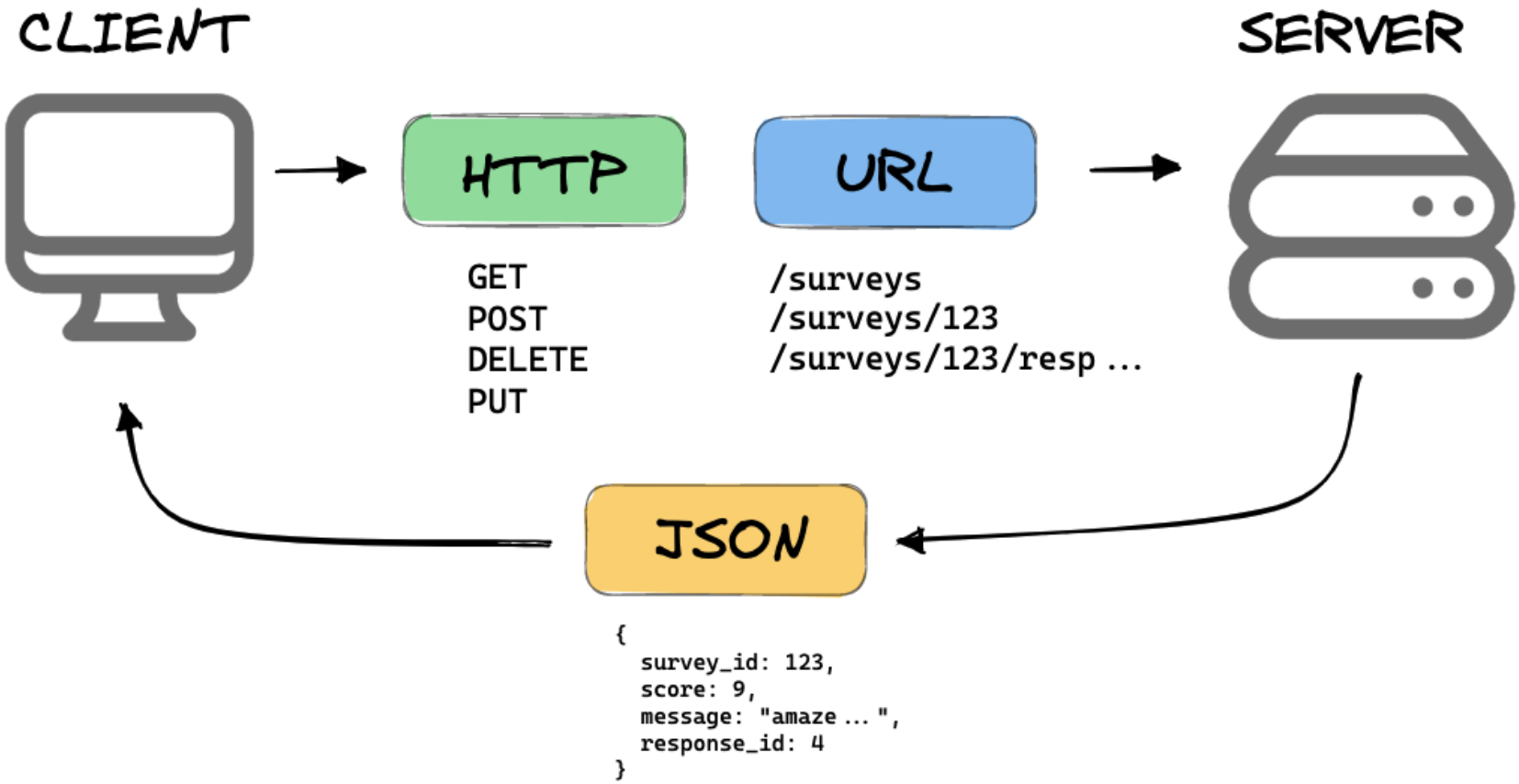
Application Programming Interface: API



An API lets two pieces of software talk to each other.

REST API

REpresentation – State – Transfer



A REST API lets two pieces of software communicate using the **HTTP** protocol.

Module 3.3: Datasets – Powering Data Science

What is a dataset?

- Collection of data
- Data structures

What is a dataset?

- Collection of data
- Data structures
  - Tabular data

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0

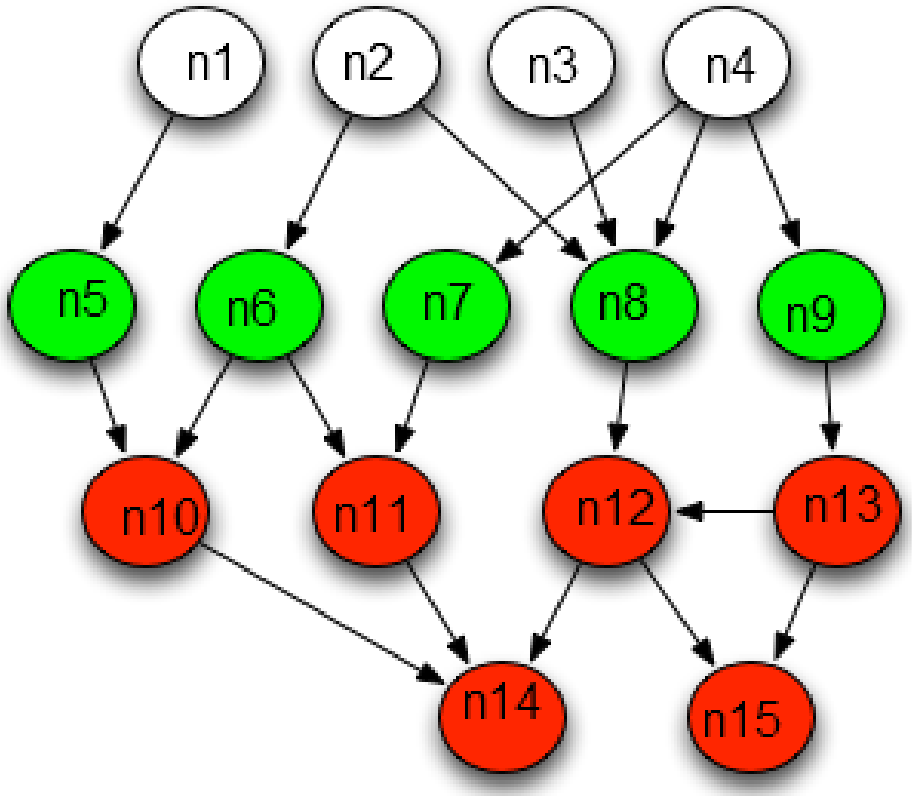
What is a dataset?

- Collection of data
- Data structures
  - Tabular data

```
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
```

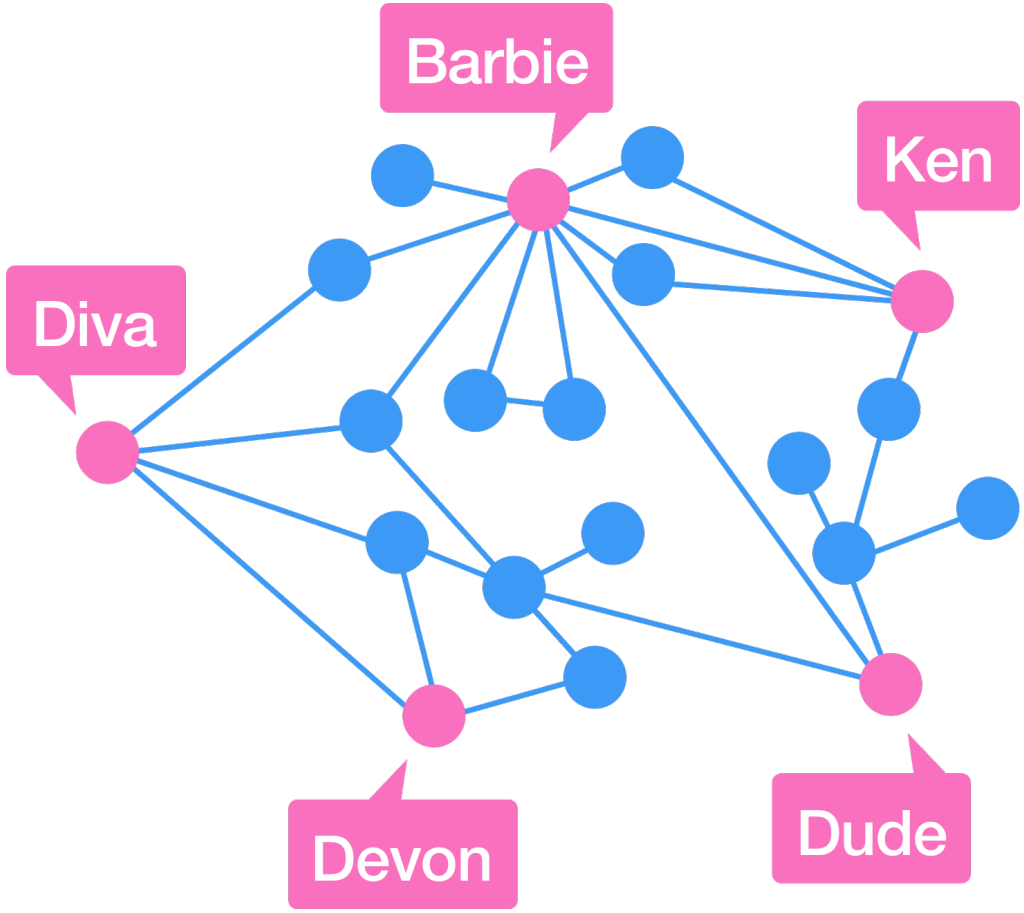
What is a dataset?

- Collection of data
- Data structures
  - Tabular data
  - Hierarchical data, network data



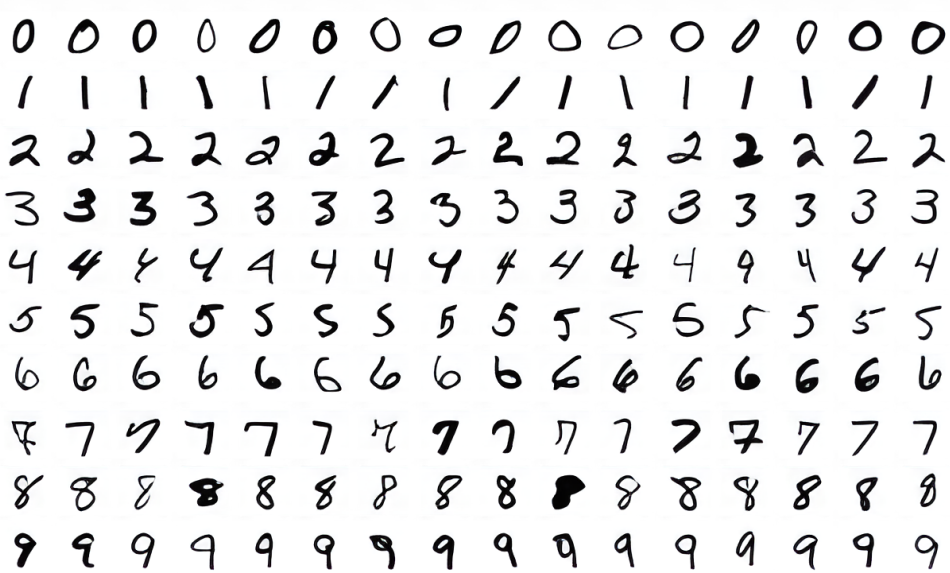
What is a dataset?

- Collection of data
- Data structures
  - Tabular data
  - Hierarchical data, network data



What is a dataset?

- Collection of data
- Data structures
  - Tabular data
  - Hierarchical data, network data
  - Raw files



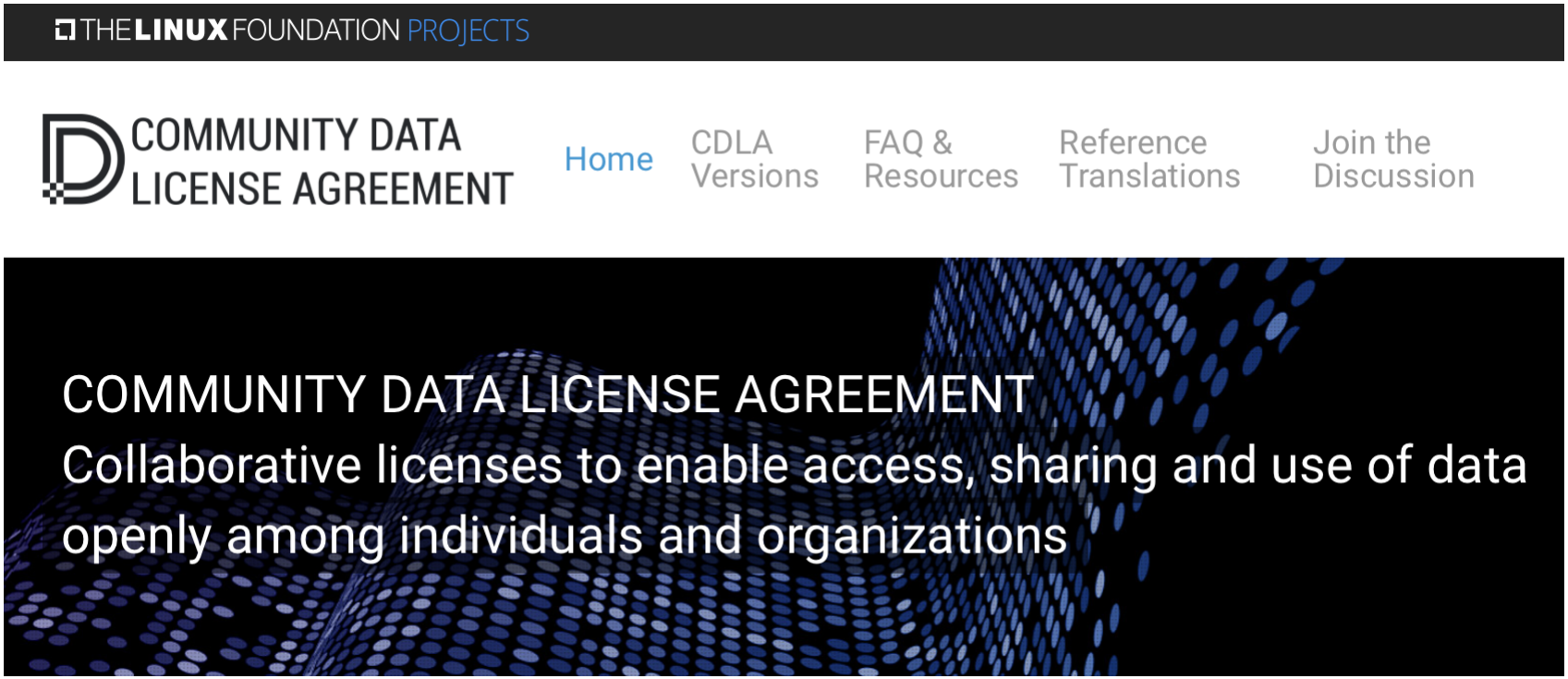
Data ownership

- Private data
  - Confidential
  - Personal information
  - Commercially sensitive
- Open data
  - Scientific institutions
  - Governments
  - Organizations
  - Companies
  - Publicly available

Where to find open data

- Open data portal list from around the world:
  - `https://dataportals.org`
- (Inter) Governmental & organization websites:
  - `https://data.un.org`
  - `https://data.gov`
  - `https://datacatalog.worldbank.org`
  - `https://gdcatalog.go.th`
  - `https://data.go.th`
- Kaggle:
  - `https://www.kaggle.com/datasets`
- GitHub:
  - `https://github.com/datasets`
  - `https://github.com/awesomedata/awesome-public-datasets`
- Google dataset search:
  - `https://datasetsearch.research.google.com`

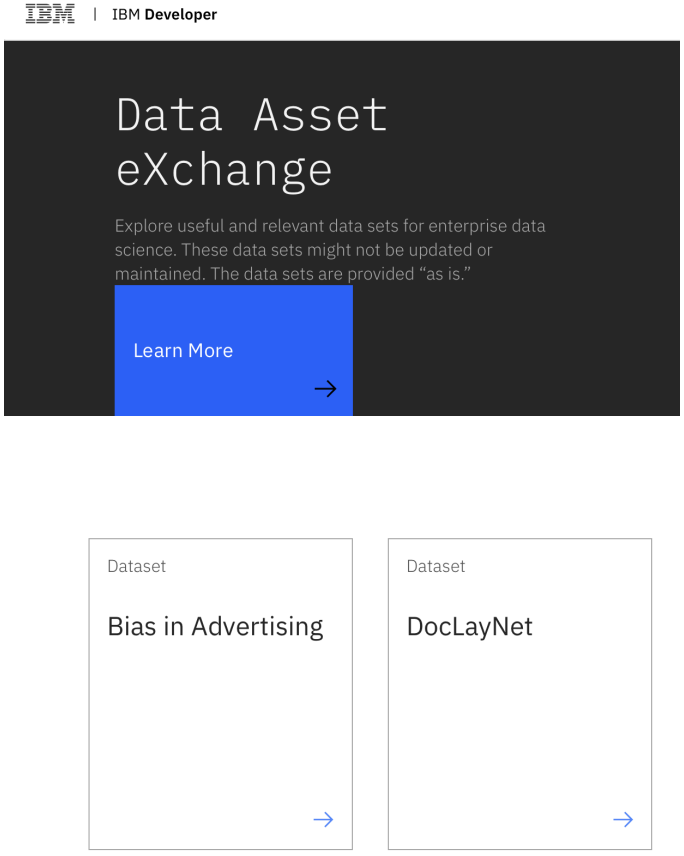
Community data license agreement



- The Linux foundation projects: `https://cdla.io`
- CDLA-Sharing: Permission to use and modify data; Publication only under same terms.
- CDLA-Permissive: Permission to use and modify data; No obligations.

Module 3.4: Data Asset eXchange

Data Asset eXchange: DAX



- Curated collection of datasets:
  - From IBM Research and 3rd parties
  - Multiple application domains
- Data science friendly liceses

`https://developer.ibm.com/exchanges/data/`

Module 3.5: Machine Learning Models

What's a model?

- Data can contain a wealth of information
- Machine Learning (ML) models identify patterns in data
- A model must be trained before it can be used to make predictions
- Supervised, unsupervised & reinforcement learnings are types of ML

Supervised learnings

- Data is labeled, & model trained to make predictions
- Regression
  - Predict real numeric values
  - e.g., home sales prices, stock market prices
- Classification
  - Classify things to categories
  - e.g., span email filters, fraud detection, image classification

Unsupervised learnings

- Data is not labeled
- Model tries to identify pattern without external helps
- Common learning problems:
  - Clustering: models are used to divide each record of a data set into one of a small number of similar groups, e.g., recommendation systems.
  - Anomaly detections: models identify outliers in a data set, such as fraudulent credit card transactions or suspicious online log-in attempts.

Reinforcement learnings

- Conceptually similar to human learning process
  - Models learn the best set of actions to take, given its current environment, in order to get the most **reward** over time.

- e.g., robot learning to walk; Chess, Go and other game of skill

Deep learnings

- Tries to loosely emulate how human brains solve a wide range of problems
- Applications:
  - Natural language processing
  - Image, audio, video analysis
  - Time series forecasting
  - Much more!
- Required typically big datasets of labeled data and is compute intensive

Deep learnings

The Impact of Fruit and Vegetable Consumption and Physical Activity on Diabetes Risk among Adults

Data to Paper

June 23, 2023

Abstract

Diabetes is a global health concern, and identifying modifiable risk factors is essential for prevention. We investigated the association between fruit and vegetable consumption, physical activity, and the risk of diabetes among adults. Using data from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey, logistic regression analysis was conducted, controlling for age, sex, BMI, education, and income. Our results show that higher fruit and vegetable consumption is associated with a reduced risk of diabetes. Moreover, engaging in regular physical activity strengthens this association. This study addresses a gap in the literature by providing evidence on the protective effects of fruit and vegetable consumption and physical activity in relation to diabetes risk. However, limitations, such as self-reported data and potential confounders, should be considered. Our findings highlight the importance of promoting healthy lifestyle behaviors and have implications for diabetes prevention interventions among adults.

Introduction

Diabetes is a major global health concern, affecting nearly half a billion people worldwide, with projections estimating an increase of 25% in 2030 and 51% in 2045 [1]. The increasing prevalence of diabetes poses both an economic and a public health burden [2]. Identification of modifiable risk factors, such as dietary habits and physical activity, is crucial for the prevention and management of diabetes [3].

Previous research has demonstrated the beneficial impact of fruit and vegetable consumption and regular physical activity on diabetes risk [4, 5],

Deep learnings

Deep learnings

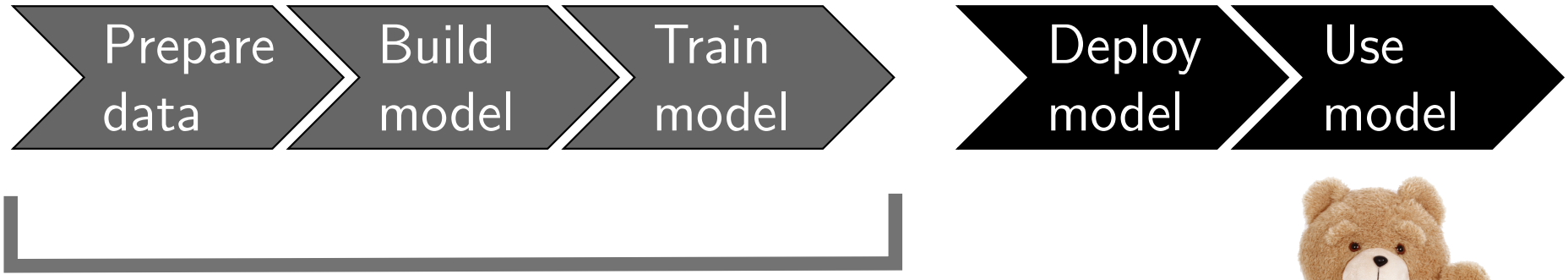
Deep learnings

Deep learning models

- Build from scratch or download from public model repositories
- Built using frameworks such as:
  - TensorFlow
  - PyTorch
  - Keras
- Popular model repositories:
  - Most frameworks provide a "model zoo"
  - ONNX model zoo
  - Hugging Face

Using models to solve a problem

What is this?



Iterative process:
Require data, expertises, time and resources

This is ted!

Module 3.6: Review questions

Review questions

Which scientific computing library provides data structures and data analysis tools for Python?

- ○ YumPies
- ○ Seahorse
- ○ Pandas
- ○ TensorFlow

Review questions

What does the acronym API stand for?

- ○ Application Programming Interface
- ○ Abstract Python Interface
- ○ Abstract Programming Interface
- ○ Algorithmic Programming Interface

Review questions

True or False: Open data is always distributed under a Community Data License Agreement.

- ○ True
- ○ False

Review questions

Which of the following is not a type of Machine Learning?

- ○ Unsupervised learning
- ○ Supervised teaching
- ○ Supervised learning
- ○ Reinforcement learning

Review questions

Which of the following is NOT a deep learning framework?

- ○ PyTorch
- ○ Tommy
- ○ TensorFlow
- ○ Keras

Module 4

GitHub

Module 4.1: Overview Git/GitHub

Version control

Commits on May 5, 2020

Fix references to defineq and ratint

NeilStrickland committed on May 5, 2020

`dd4c6ce`

Commits on Feb 9, 2020

Edit index.html

NeilStrickland committed on Feb 9, 2020

`39fe558`

Add image to index file

NeilStrickland committed on Feb 9, 2020

`25ea174`

Add docs folder

NeilStrickland committed on Feb 9, 2020

`4777cc1`

Add docs folder

NeilStrickland committed on Feb 9, 2020

`c7bcc45`

Minor changes

NeilStrickland committed on Feb 9, 2020

`b9eba21`

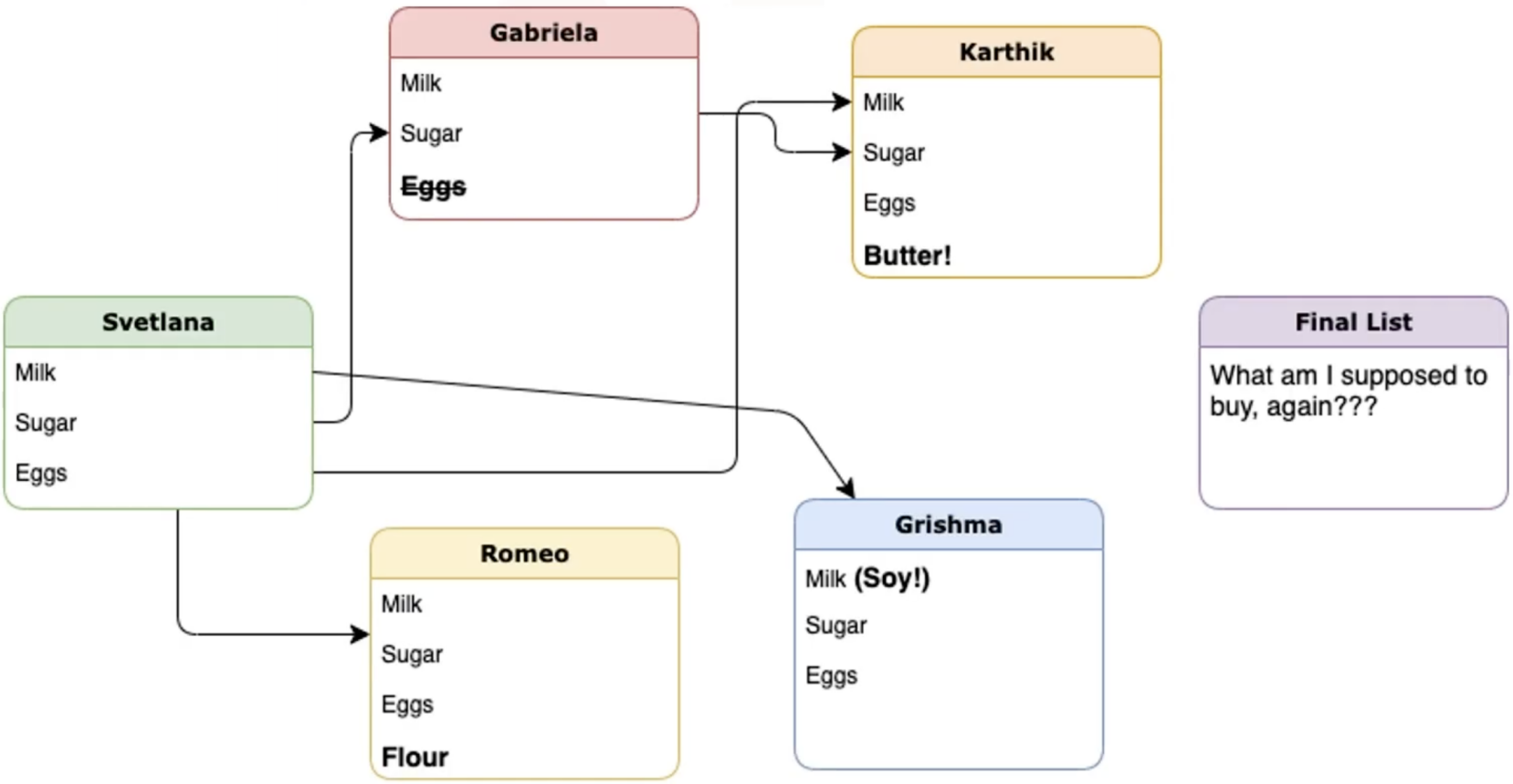
Commits on Feb 8, 2020

Initial commit

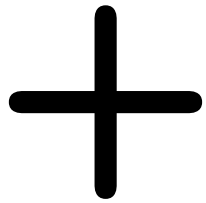
NeilStrickland committed on Feb 8, 2020

`4f2e835`

Working without version control?



- Free and open source
- Distributed version control system
- Accessible anywhere in the world
- The most common version control system
- Can version control images, documents, etc.



(

GitLab

Bitbucket

)

Glossary

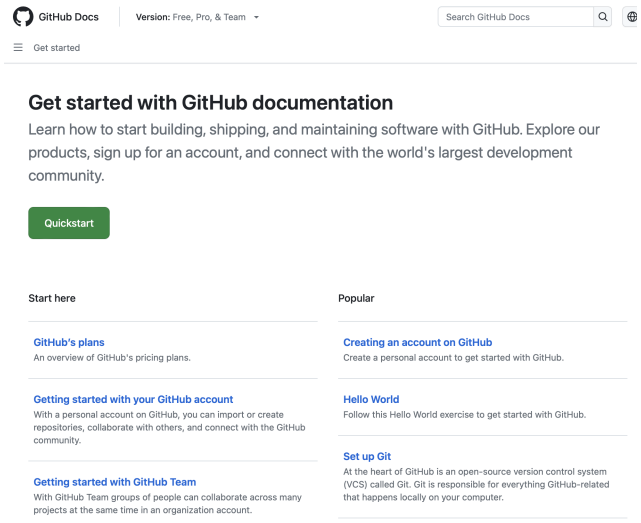
- **Secure SHell Protocol: SSH** – A method for secure remote login from one computer to another.
- **Repository** – The folders contains your project that are set up for version control.
- **Fork** – A copy of a repository.
- **Pull request** – The process you request that someone reviews and approves your changes before they become final.
- **Working directory** – The folder contains the files and subdirectories *on your computer* that are *associated with a Git repository*.

Basic Git commands

- `init`
- `add`
- `status`
- `commit`
- `log`
- `branch`
- `checkout`
- `merge`
- `revert`
- `reset`

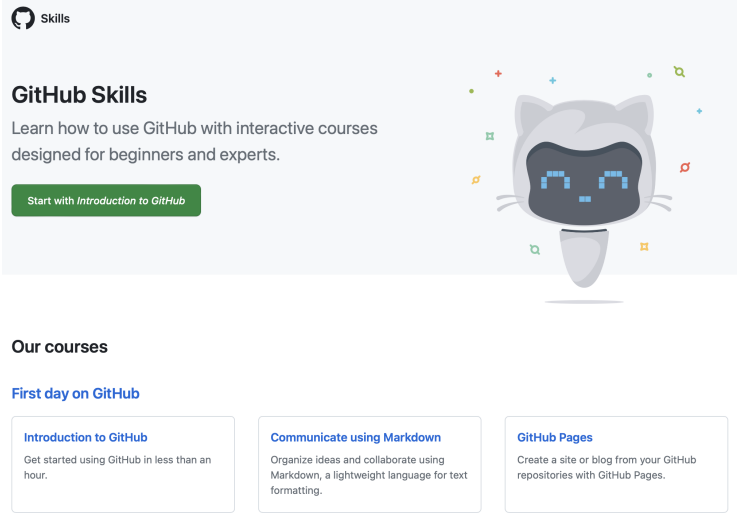
Learning resources

GitHub Documentation



https://docs.github.com/en/get-started

GitHub Skills



https://docs.github.com/en/get-started

Module 4.2: Review Question

Review questions

Which of the following statements are true? (Select all the apply)

- ○ Git is an integrated development environment for data science.
- ○ Git is a system for version control of source code.
- ○ Git is very useful for data science as well since data science often involves a lot of source code to be written and managed.

Review questions

Which of the following statements about repositories are correct? (Select all that apply.)

- ○ The remote repository is only accessible by myself.
- ○ The local repository is only accessible by myself.
- ○ The staging is only accessible by myself.
- ○ The remote repository is accessible by all contributors.
- ○ The local repository is accessible by all contributors.

Review questions

What is the best process contributing a bugfix to a foreign repository?

- ○ Send the fix via email to the author.
- ○ Fork the repository, update the fork and create a pull request.
- ○ Ask the repository owner for write access to the repository.