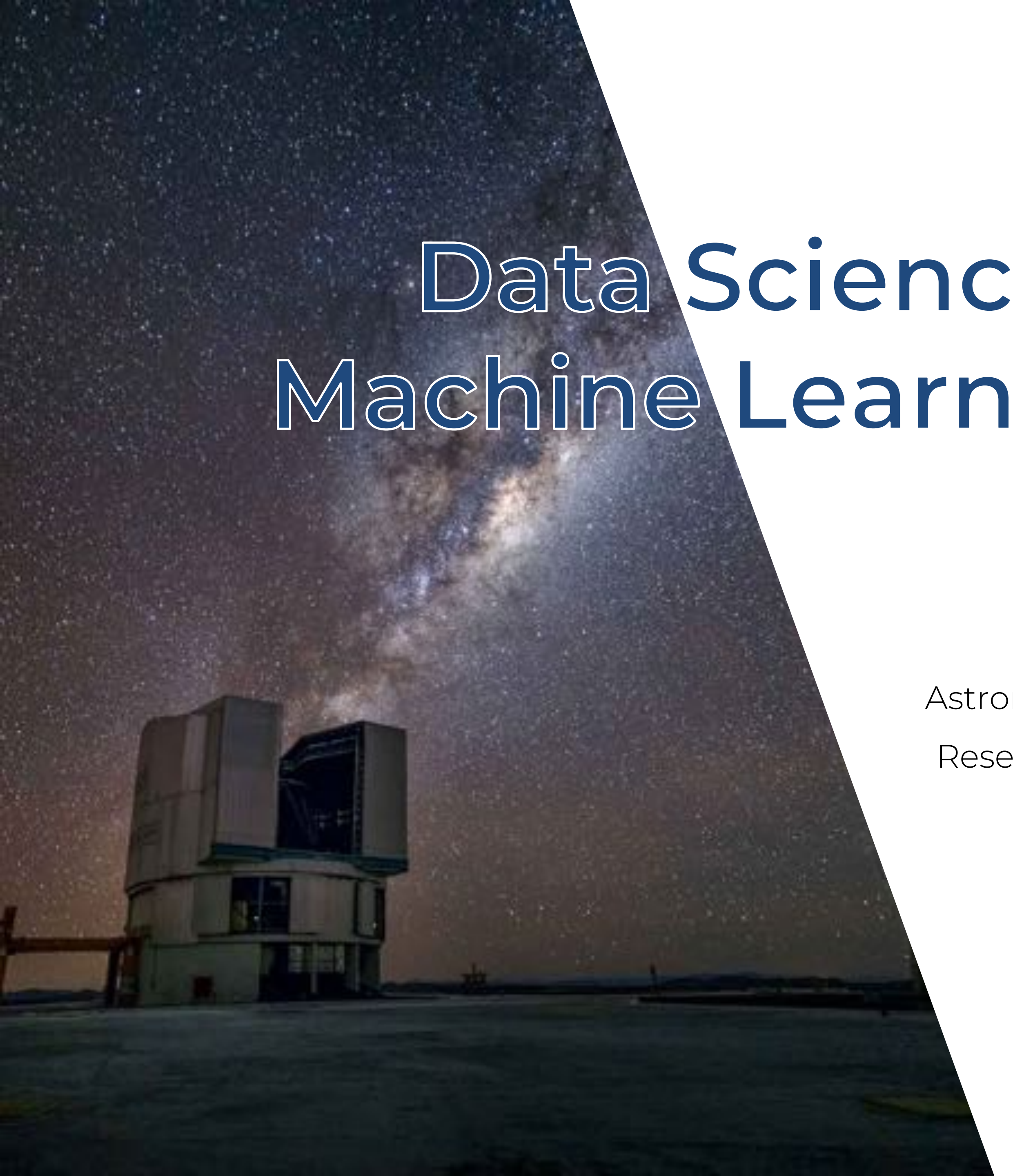


Data Science & Machine Learning

Roberto Muñoz

Astronomer and Data Scientist

Research officer at MetricArts

@RobertoKPax

Microsoft

METRICARTS

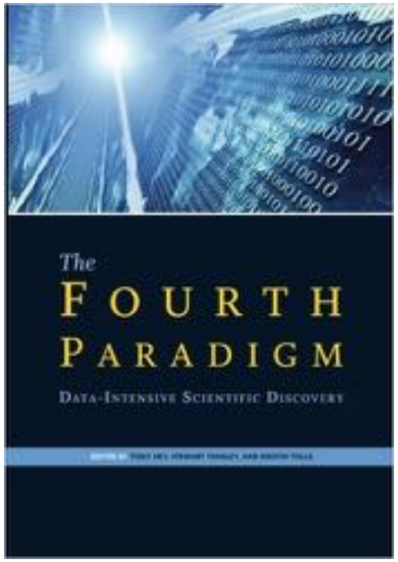
DATA: THE NEW OIL

The world's most valuable resource

Data and the new rules of competition

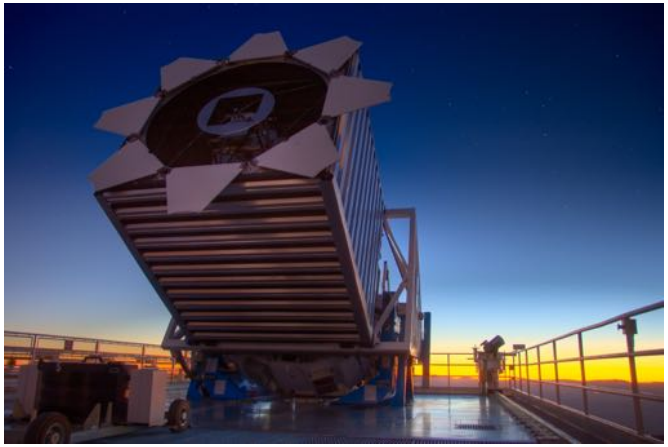
NEW PARADIGM

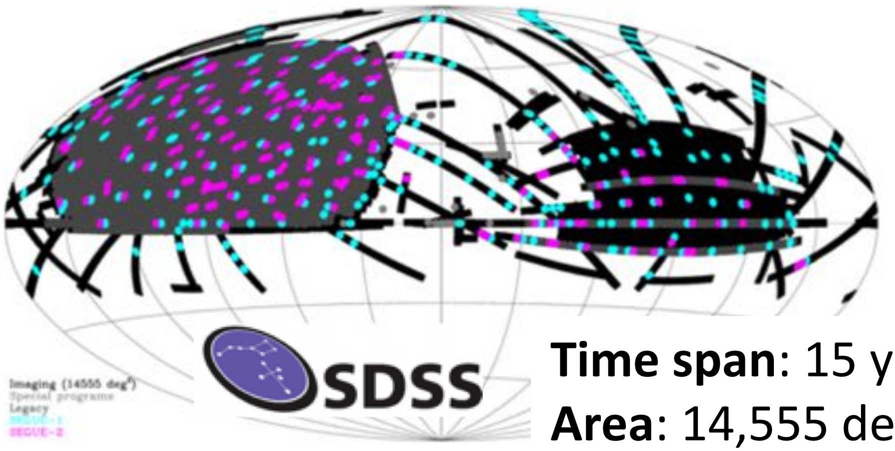
- Digital information and technology have changed the way we live and understand the world
- Jim Gray, researcher at Microsoft and pioneer in database coined the term The Fourth Paradigm
- Experimental age, Theoretical, Computational and lately Data-driven age



DATA IN ASTRONOMY

- Raw data is collected by instruments at telescopes, stored in data servers, processed and published
- Big data in Astronomy: SDSS (1998)
- Filled 8GB HDD every 25 minutes





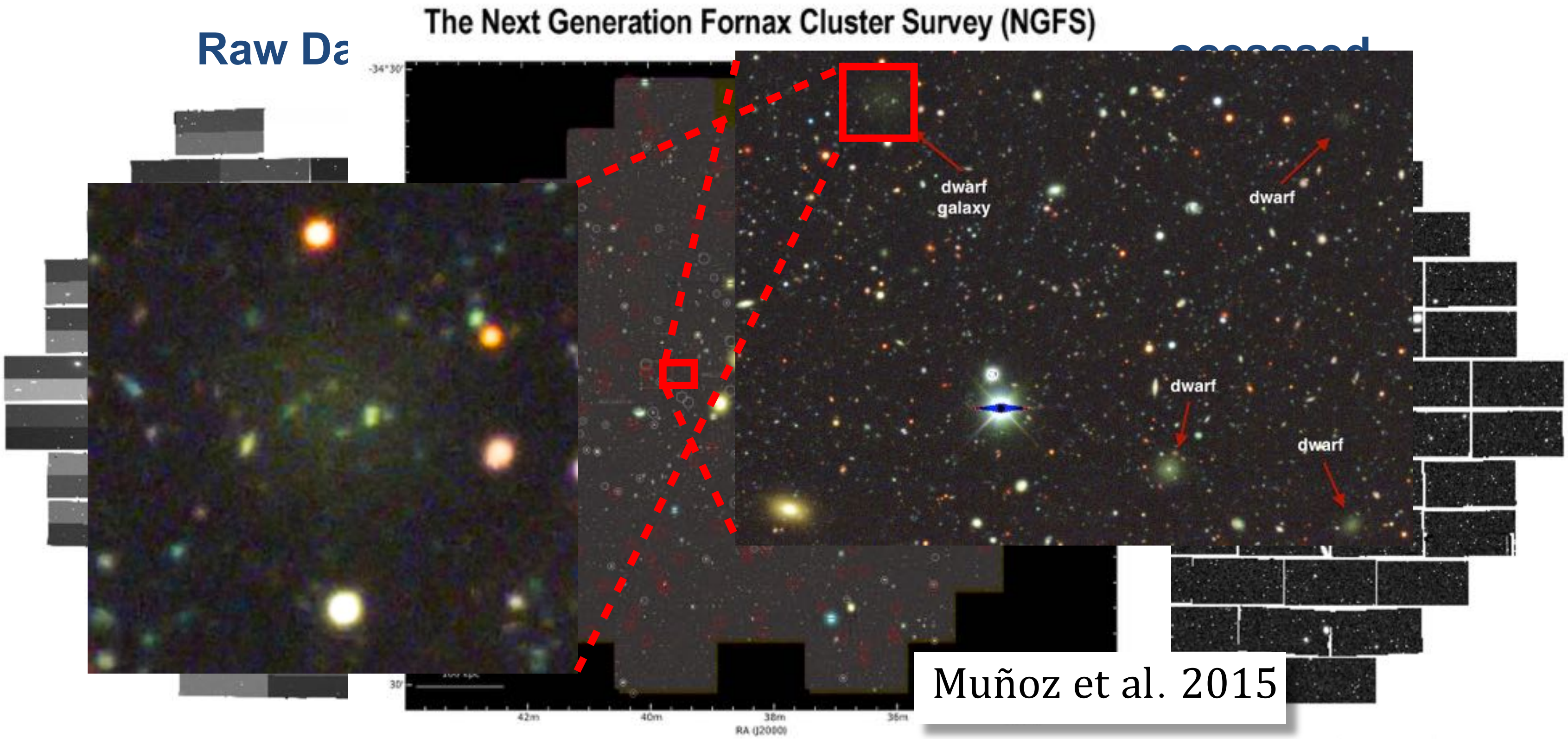
Time span: 15 years
Area: 14,555 deg^2
# sources: 469,053,874

Data Analytics

PROCESSING AND ANALYTICS

IMPORTANCE OF PROCESSING

- Telescopes in Chile produce 1.5 PB/year
- By 2023 telescopes will produce 17 PB/year (EY 2018)



Muñoz et al. 2015

DATA PREPARATION

- **Preprocessing**
  - Preparation of data directly after accessing it from a data source
  - Initial transformations, aggregations and data cleansing
- **Wrangling**
  - Preparation of data during the interactive data analysis and model building
  - Cleaning, structuring and enriching dataset until it works well for finding insights

DATA ANALYTICS

- Analytics is the process of examining datasets in order to draw conclusions about the information they contain
- Data-intensive, statistics, quantitative, descriptive and predictive analysis
- Decision making, competitive advantage, generate value

ANALYTICAL PIPELINE

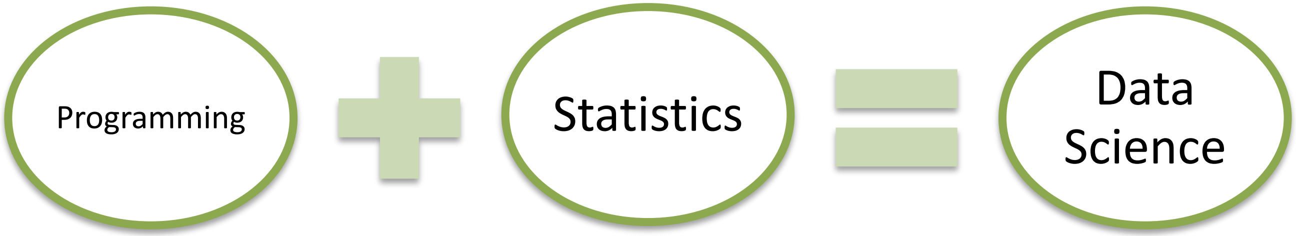
- Data Access
- Data Preprocessing
- Exploratory Data Analysis
- Model building
- Model validation
- Model execution
- Deployment

DS

DATA SCIENCE

WHAT IS DATA SCIENCE?

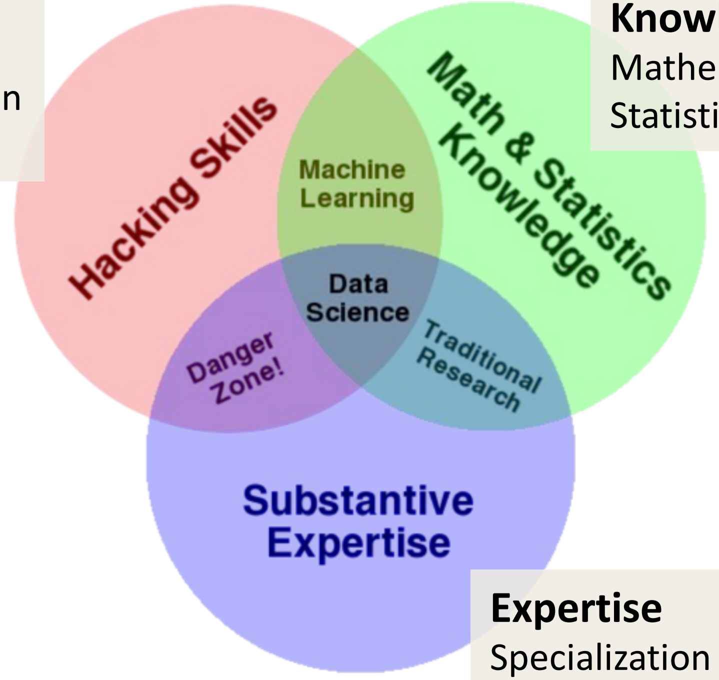
- Data Science is an interdisciplinary field that use processes and systems to extract information and knowledge from data.
- It requires knowledge from multiple fields: Computer Science, Programming, Mathematics and Statistics.



INTERDISCIPLINARY

Skills
Programming
Data exploration
Creativity

Knowledge
Mathematics
Statistics



Expertise
Specialization
Domain knowledge

Venn diagram for Data Science
Drew Conway (2010)

PROGRAMMING LANGUAGES

- The most used languages in Data Science are Python and R. Python has more than 30 million (M) users and R more than 16M.
- Julia has emerged as an all-around and efficient language. Around 2M users.

ML

MACHINE LEARNING

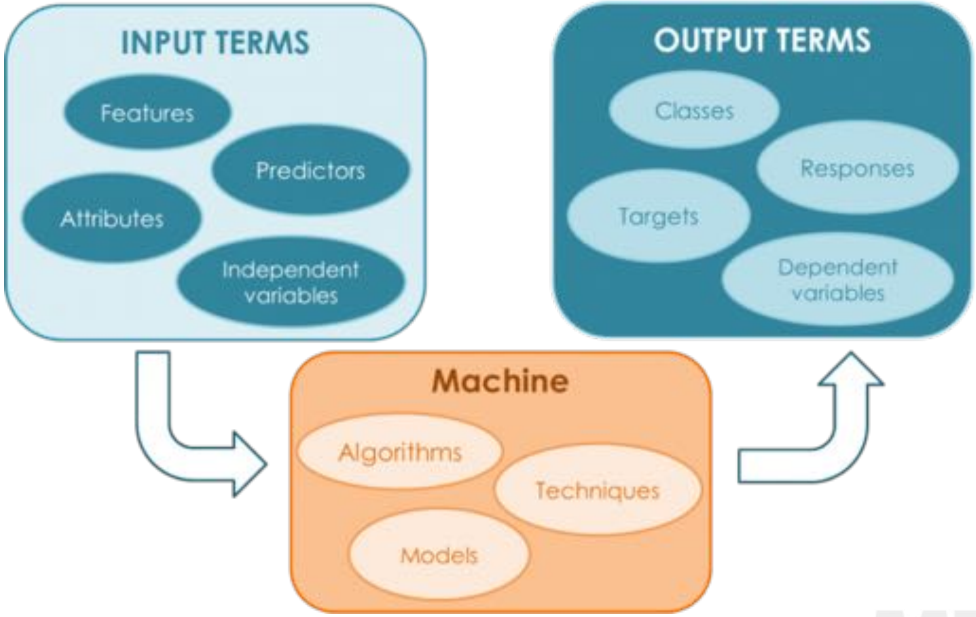
USUAL TASKS

- **Regression**
  “The company Entel wants to learn how many GB their clients will consume next month”
- **Classification**
  “Falabella wants to promote a new product. What gender and age they should focus their ads?”
- **Association**

  “User 1 from Netflix has watched movies A and B, while user 2 has watched movies B and C. Next time Netflix will offer movie C to user 1”

MACHINE LEARNING

- M*achine learning* is a subfield from Computer Science. Heavily based in Statistics
- The goal is to develop techniques that allow to computers to learn or imitate human cognitive skills

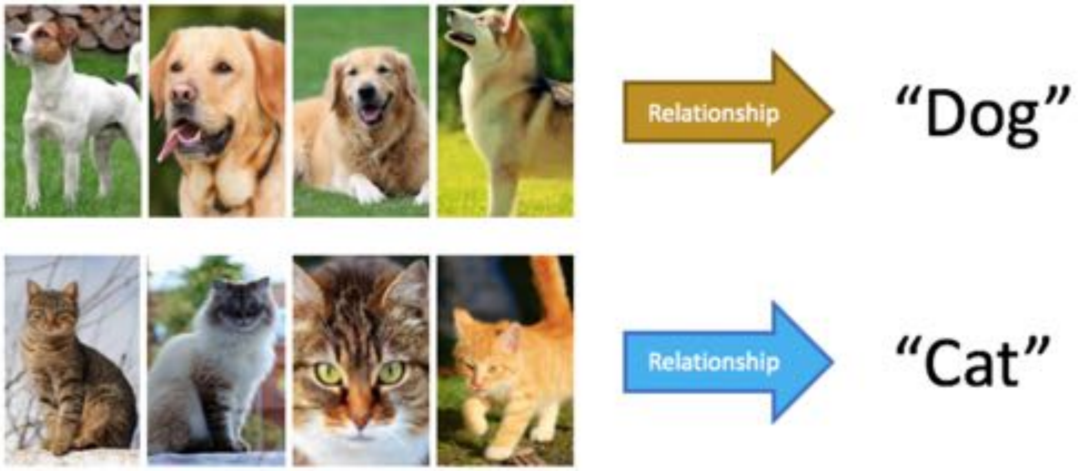


TYPES OF ML

- **Supervised learning**

  The system learns using previously classified data. The data can be structured or unstructured.

  Algorithm generates a model that establish the correspondence between the input data and expected output of the system.

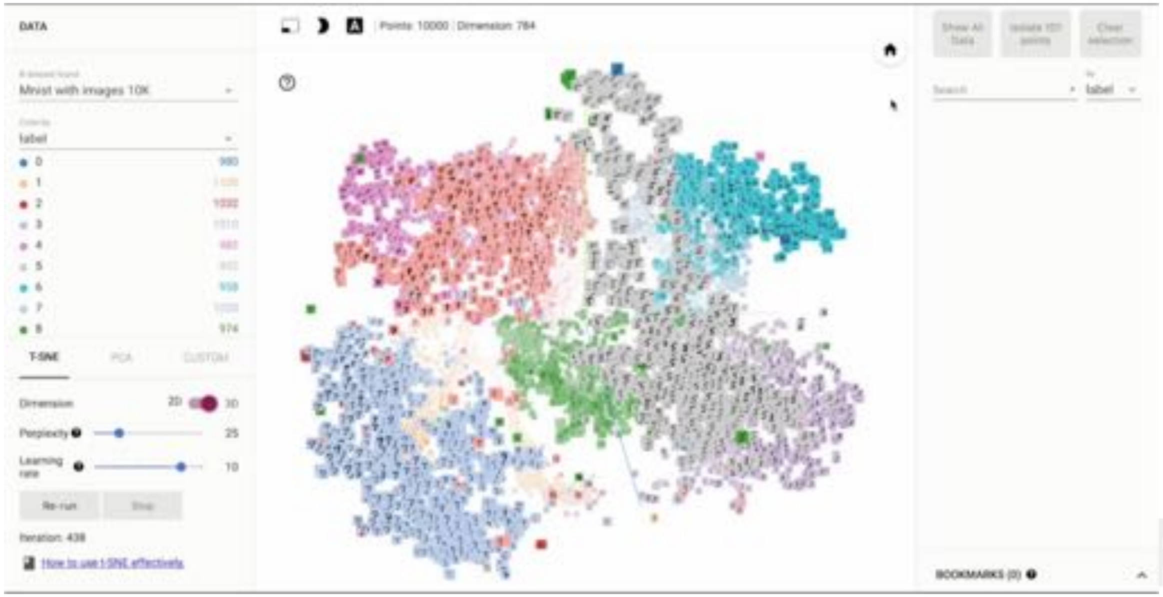


TYPES OF ML

- **Unsupervised learning**

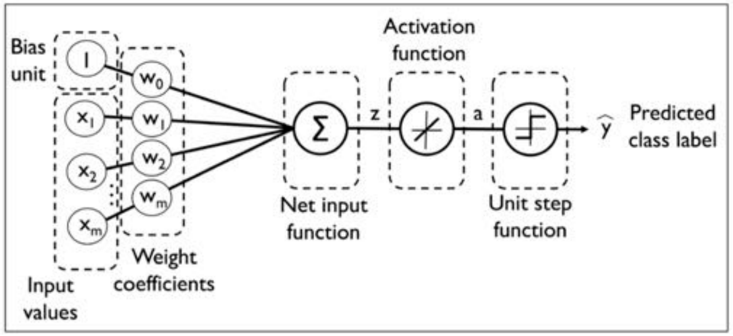
  The data has not been classified previously.

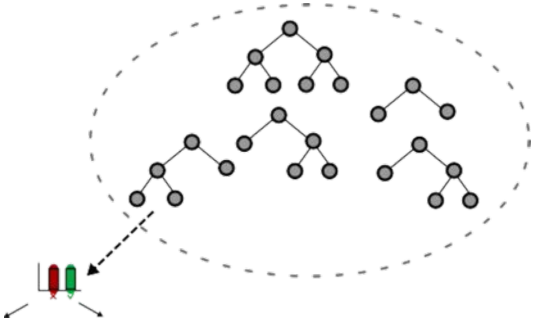
  The system should be able to recognize patterns and generate their own labels. Model should classify new input data.



CLASSIFICATION ALGORITHM

- The most used methods to solve Classification problems using ML are
  - Support Vector Machine (SVM)
  - Decision trees
  - Random forest
  - Deep Neural Networks





CONFUSION MATRIX

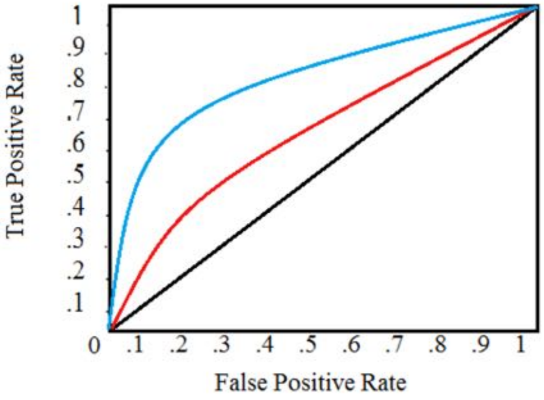
		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

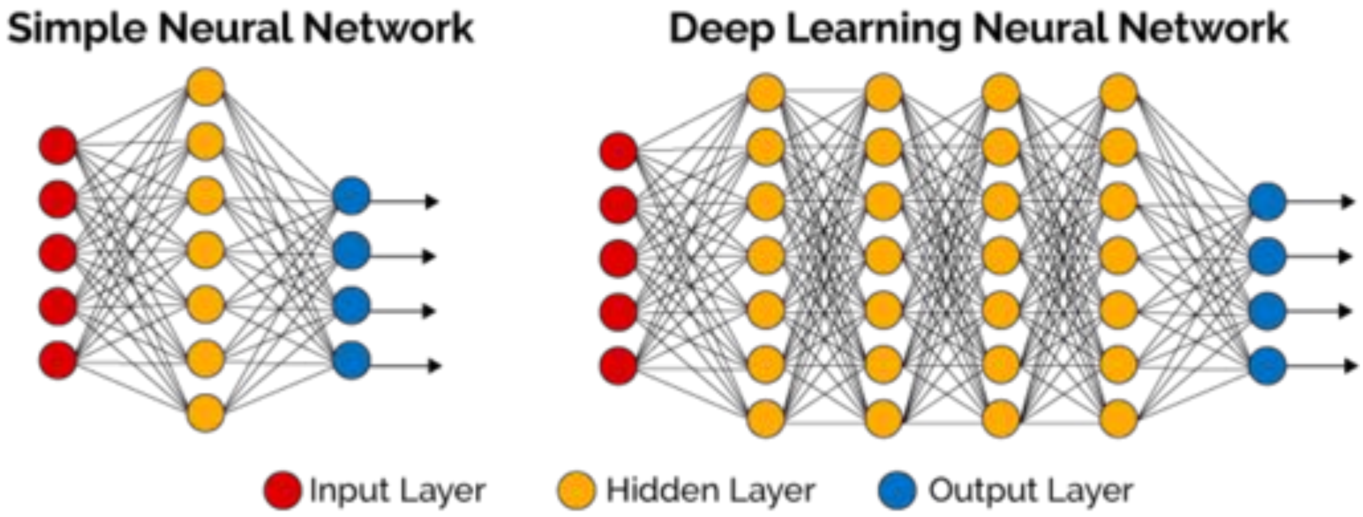
$$true\ positive\ rate = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$false\ positive\ rate = \frac{false\ positives}{false\ positives + true\ negatives}$$

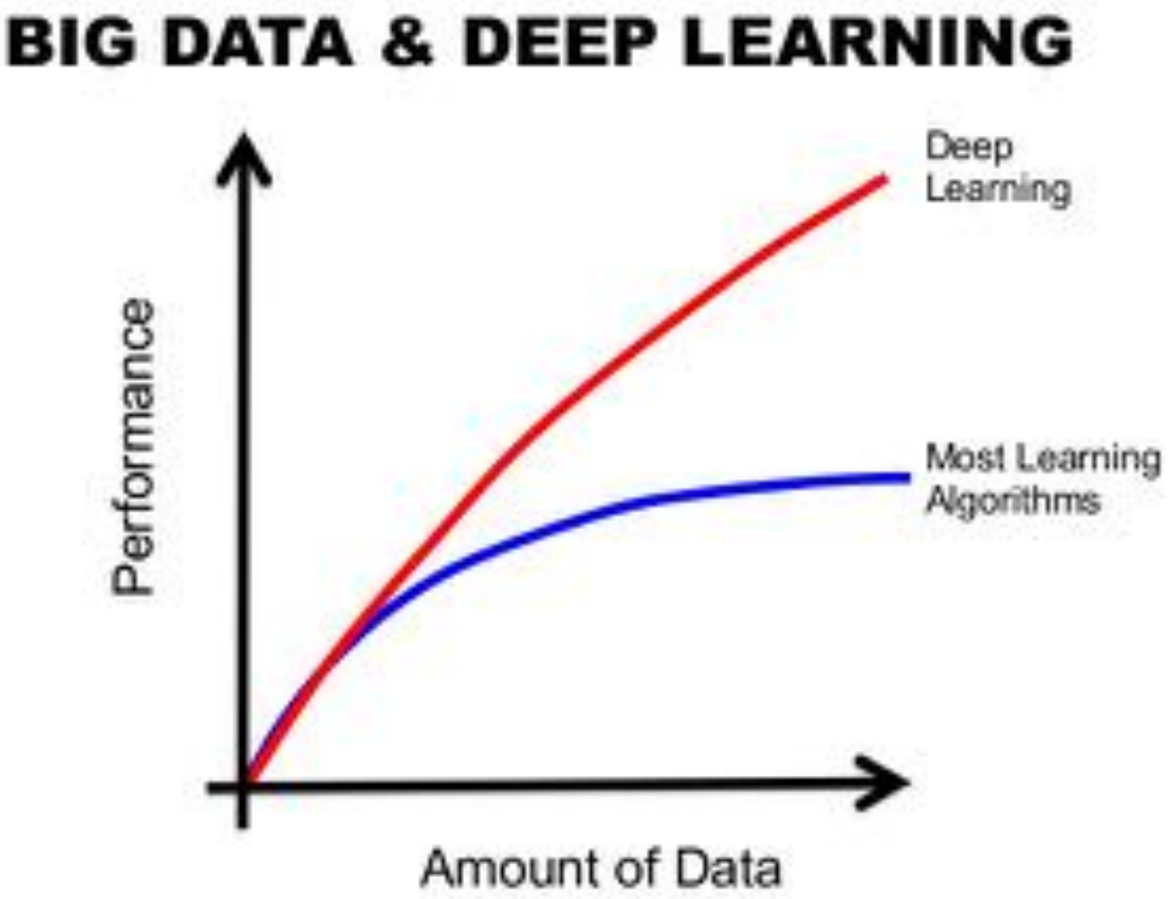


DEEP LEARNING

- Deep learning is a collection of machine learning methods based in feature or representation learning.
- The most famous are Deep Neural Networks (DNN). They are based on Artificial Neural Networks and multiple layer architectures.



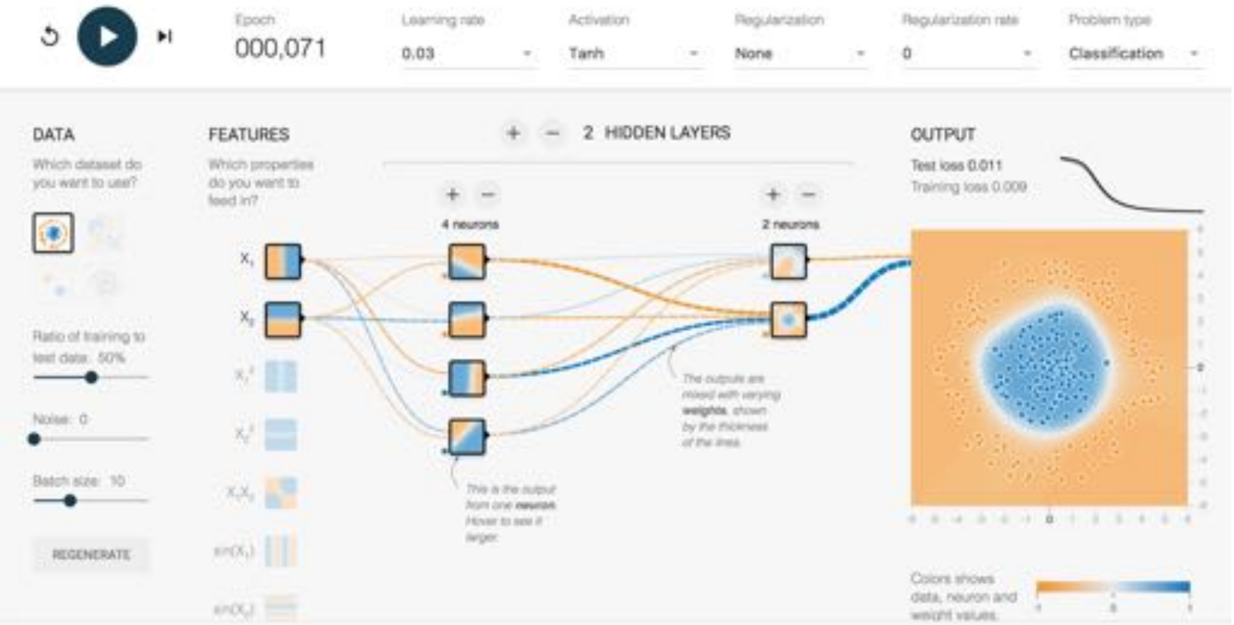
HOW MUCH DATA?



NEURAL NETWORKS

Visit the website

http://playground.tensorflow.org



Epoch 000,071

Learning rate 0.03

Activation Tanh

Regularization None

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES

Which properties do you want to feed in?

2 HIDDEN LAYERS

4 neurons

2 neurons

This is the output from one **neuron**. Hover to see it larger.

The outputs are mixed with varying **weights**, shown by the thickness of the lines.

OUTPUT

Test loss 0.011

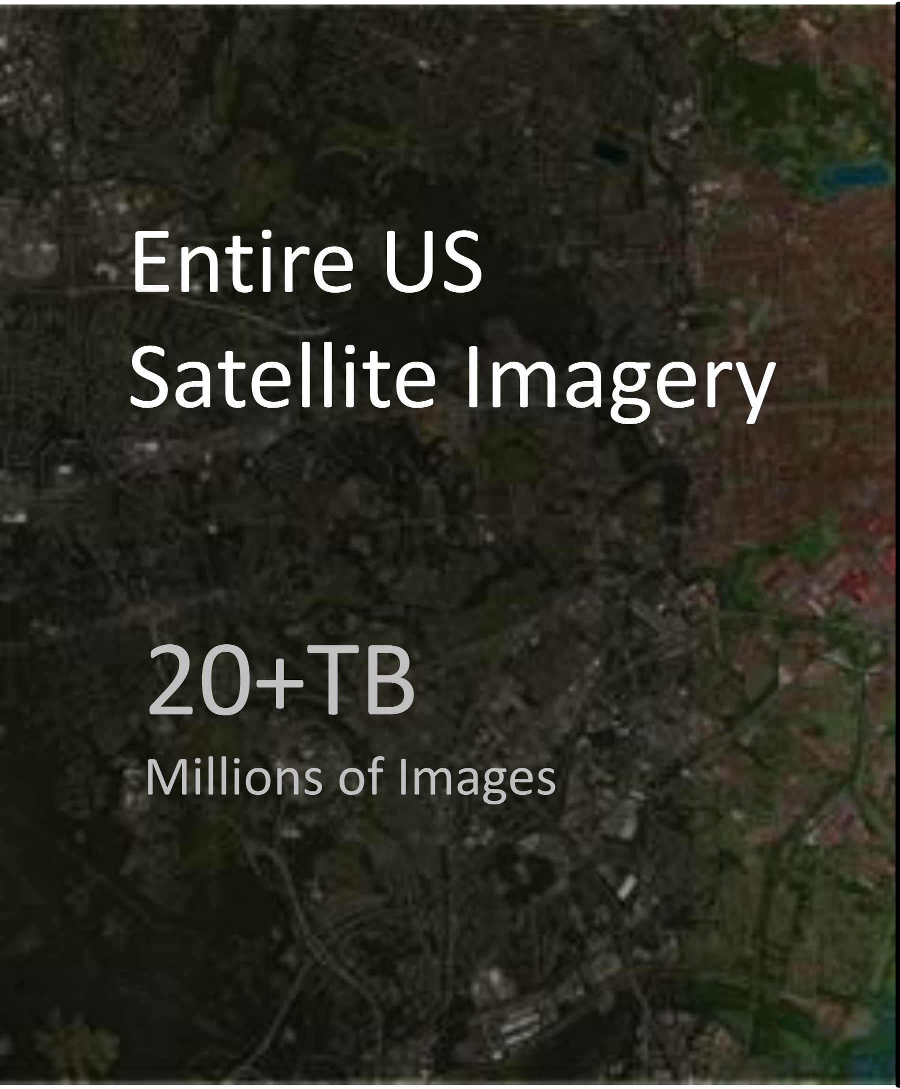
Training loss 0.009

Colors shows data, neuron and weight values.

Trends

TRENDS IN MACHINE LEARNING

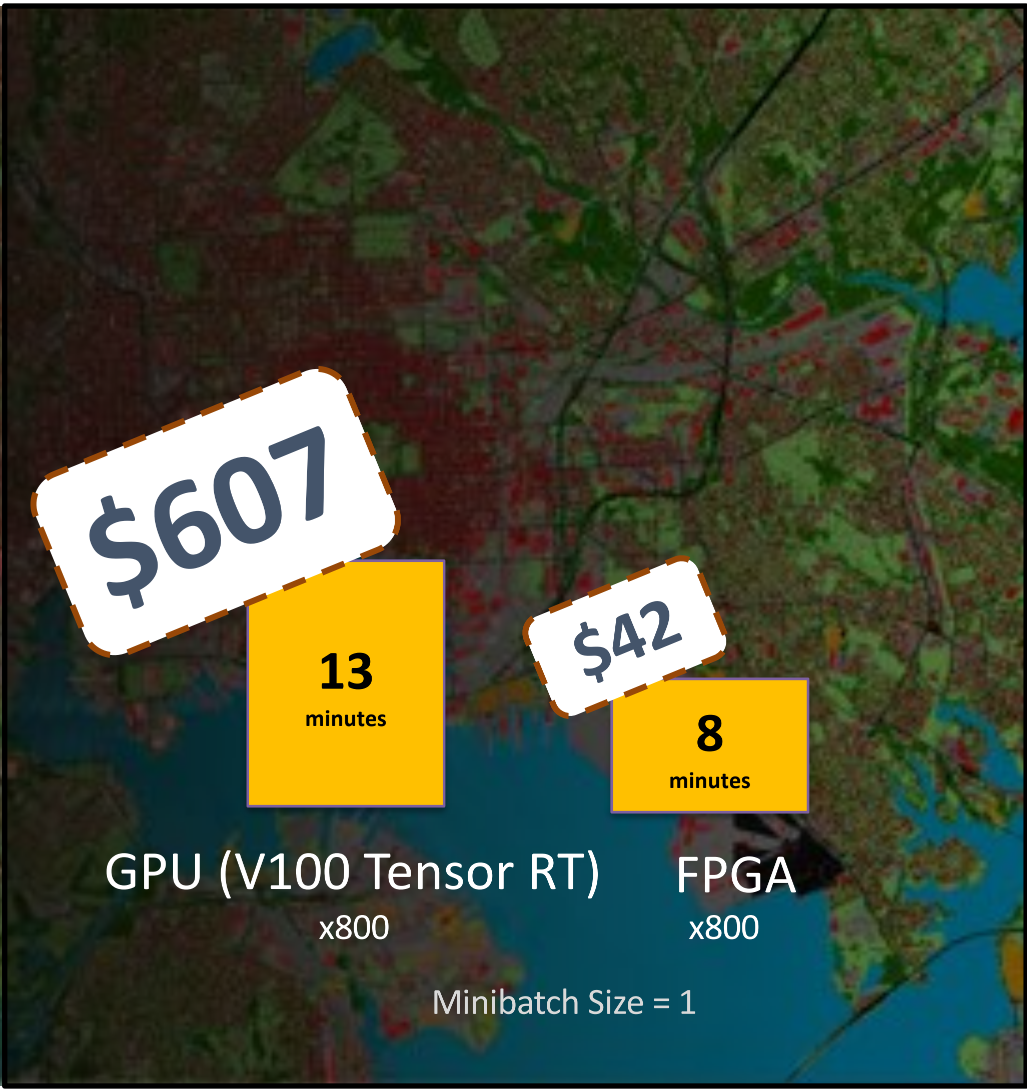
CPU VS GPU



Entire US Satellite Imagery

20+TB

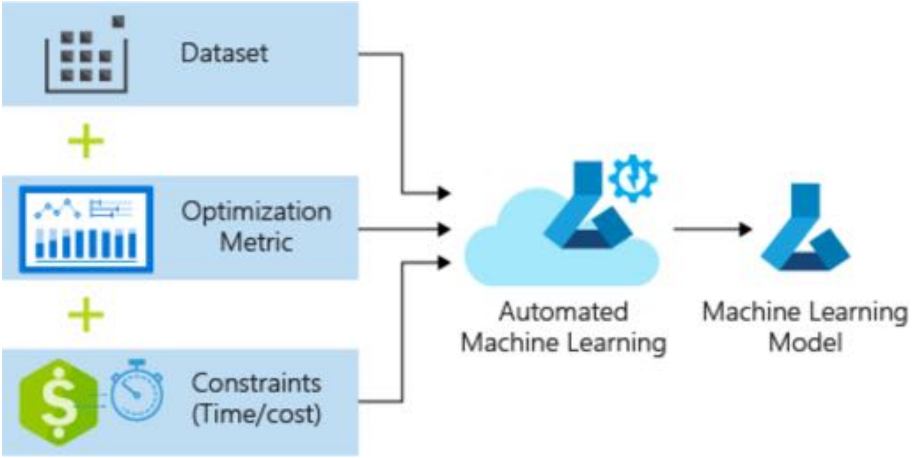
Millions of Images



METRICARTS

AUTOMATED MACHINE LEARNING

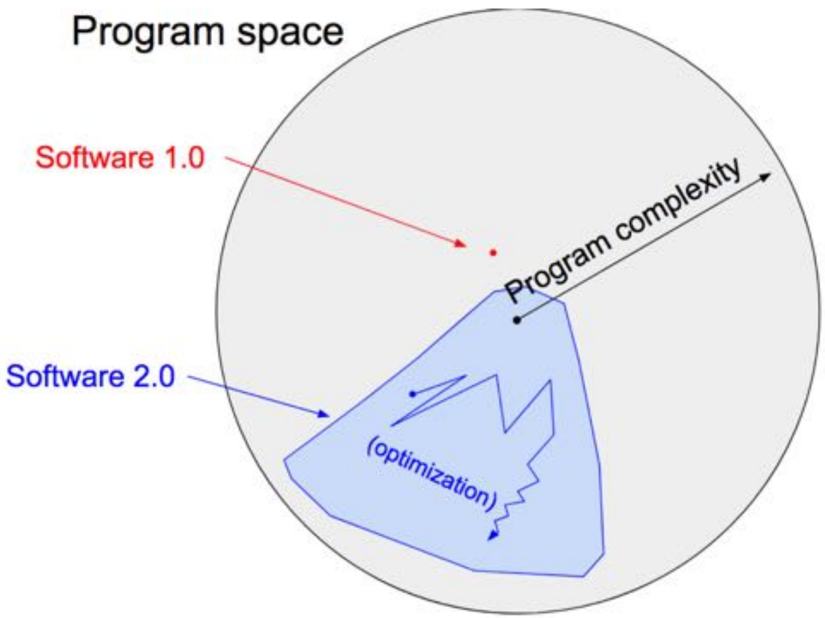
- Automatically search for algorithms, architectures and hyperparameters to find best scoring machine learning model



AUTO KERAS

METRICARTS

SOFTWARE 2.0



Software 1.0 is what we're all familiar with — it is written in languages such as Python, C++, etc. It consists of explicit instructions to the computer written by a programmer

Software 2.0 can be written in much more abstract, human unfriendly language, such as the weights of a neural network. No human is involved in writing this code because there are a lot of weights and coding directly in weights is kind of hard

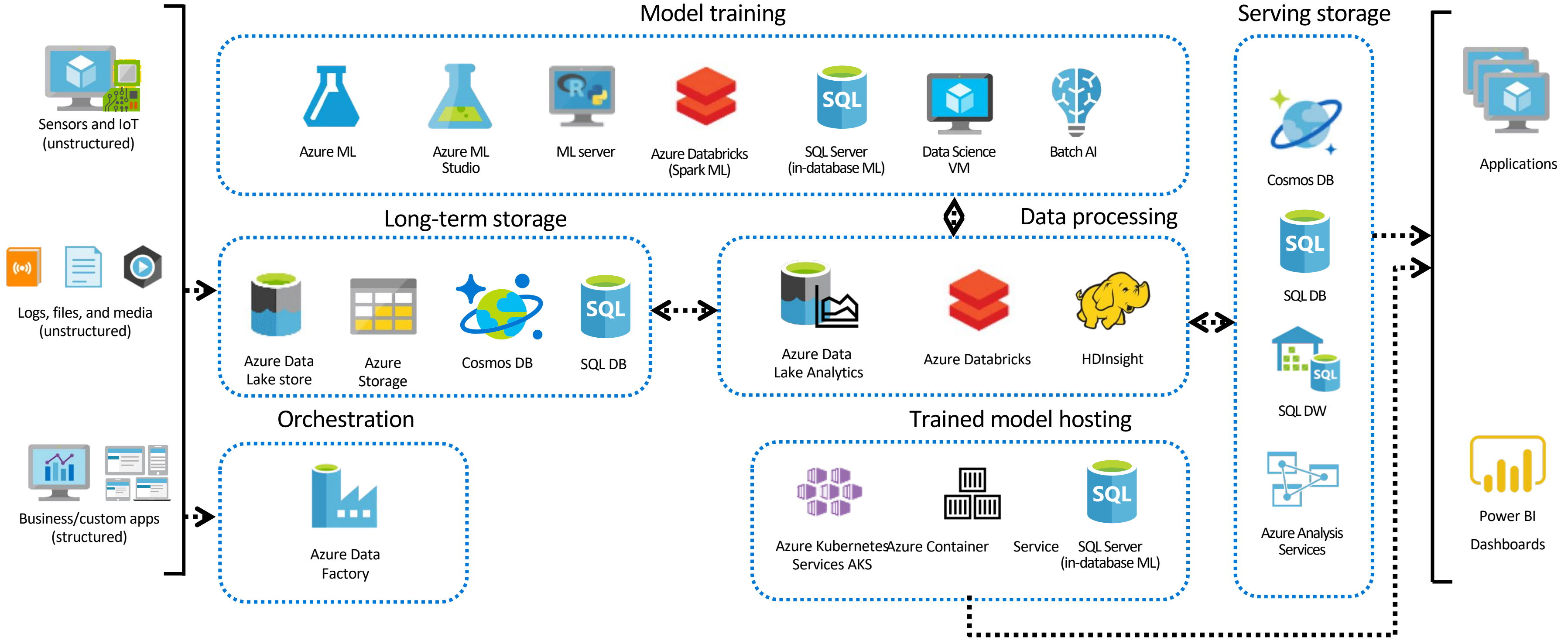
Andrej Karpathy
Director of AI at Tesla

Azure

AZURE MACHINE LEARNING

Advanced analytics pattern in Azure

Data collection and understanding, modeling, and deployment



Leverage out-of-the-box AI tools and services

Cognitive services

Use pre-built AI services
to solve business problems

Map complex
information and data

Allow your apps to
process natural language

Azure search

Get up and
running quickly

Reduce complexity with
a fully-managed service

Use artificial intelligence
to extract insights

Bot services

Speed development with a purpose-built
environment for bot creation

Infuse intelligence into your bot using
cognitive services

Integrate across multiple
channels to reach more customers

Create a seamless developer experience across desktop, cloud, or at the edge using Visual Studio AI Tools

Azure Databricks for deep learning modeling

Fast, easy, and collaborative Apache Spark-based analytics platform

Tools

Use HorovodEstimator via a native runtime to enable build deep learning models with a few lines of code

Load images natively in Spark DataFrames to automatically decode them for manipulation at scale

Simultaneously collaborate within notebooks environments to streamline model development

Frameworks

Full Python and Scala support for transfer learning on images

Seamlessly use TensorFlow, Microsoft Cognitive Toolkit, Caffe2, Keras, and more

Use built-in hyperparameter tuning via Spark MLLib to quickly drive model progress

Infrastructure

Leverage powerful GPU-enabled VMs pre-configured for deep neural network training

Automatically store metadata in Azure Database with geo-replication for fault tolerance

Improve performance 10x-100x over traditional Spark deployments with an optimized environment

H2O.ai in Microsoft Azure Cloud for the Enterprise

- Enterprise customers chose Azure for security, ease of deployment, enterprise ready capabilities

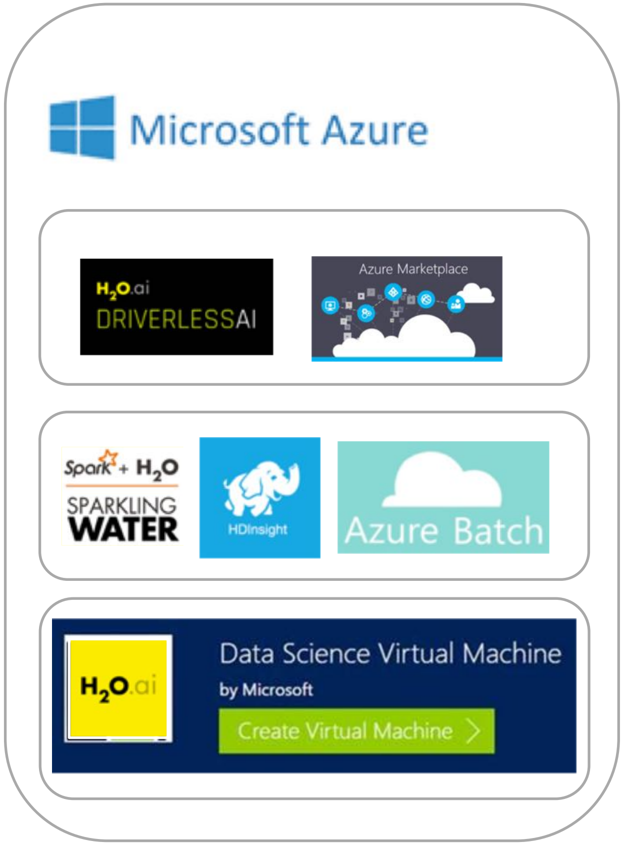
- H2O Driverless AI is available on the Marketplace

- H2O Core and Sparkling Water - open source - also available in Azure
  - Integrated with HDInsights, Batch, Databricks

Customers

KAISER PERMANENTE.

Travelport



Azure

TUTORIALS USING AZURE

AZURE NOTEBOOKS

Visit the website and Sign in

https://notebooks.azure.com/

Clone the Github repo

https://github.com/rpmunoz/workshop_eso_2019

jupyter

Notebooks hosted on
Microsoft Azure