ML Foundations

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set For Modeling: Data Exploration & Munging

What you can expect in this session

01 Overview

02 Data Preparation for ML

03 Data Preparation Tools

04 Introduction to Data Preparation Demo

05 What's Next?

Overview

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set for Modeling: Data Exploration & Munging

Part: 1

AI & ML Foundations

H2O.ai

Session: X

AI Foundations

- Intro to Key AI Concepts
- No prior AI knowledge or background necessary
- No technical or coding experience necessary
- Exercises: Non-Technical and introductory

ML Foundations

- Applied AI Concepts
- Some experience with Python or R would be helpful to success
- Exercises: Technical and deeper

In both courses you get access to H2O.ai experts and community makers!

You can earn a badge for AI & ML Foundations by successfully completing the assessments at the end of each module (**not required**).

ML Foundations Overview

H2O.ai

Module 0: Start with Business Problem, Again

Module 1: Data Wrangling & Data Curation

- **Session 1: Getting the Right Data Set For Modeling** You Are Here
- Study Group
- Ask Me Anything

Module 2: Feature Engineering in Machine Learning

Module 3: Machine Learning Deep Dive

Interested in knowing the full schedule for the ML Foundations course? View the schedule on the community learning site

Data Preparation for Machine Learning

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set for Modeling: Data Exploration & Munging

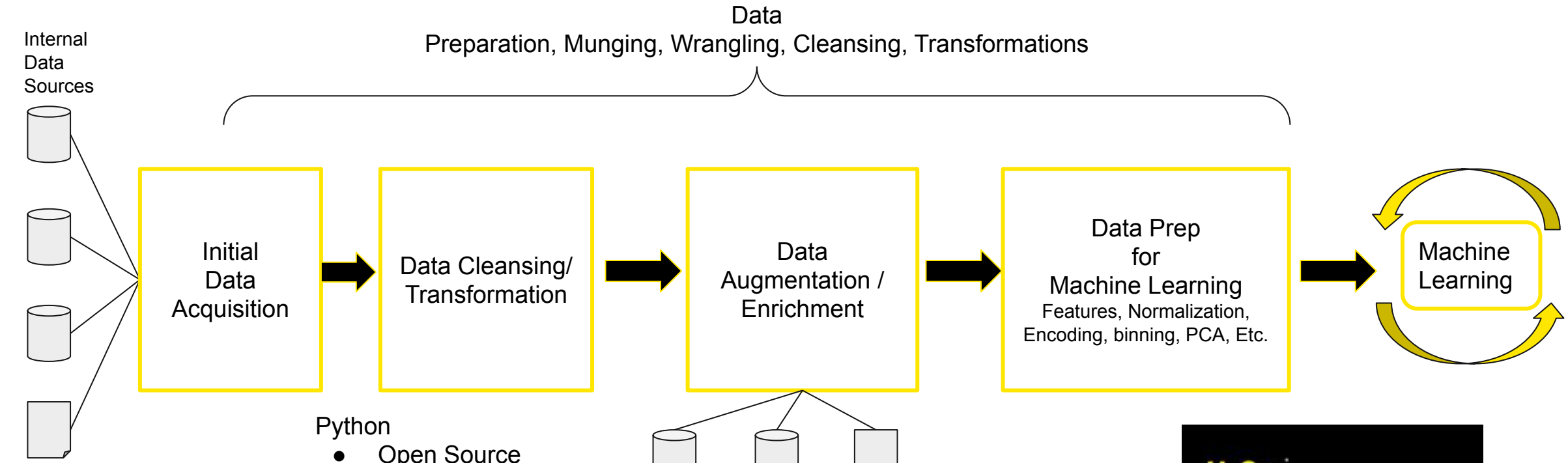
Part: 2

Data & The Role of Machine Learning

- Machine Learning happens when a computer can take lots of data (examples) and learn patterns from it to make predictions on new data based on those learned patterns.

- Constructing a **good** dataset is critical once your business problem has been identified and can be achieved using various data preparation techniques

Data Preparation for Machine Learning



Python
- Open Source
- Python DataTable

R
- Open Source
- R Data.Table

Spark

Commercial Offerings

alteryx TRIFACTA KNIME Informatica

and more….

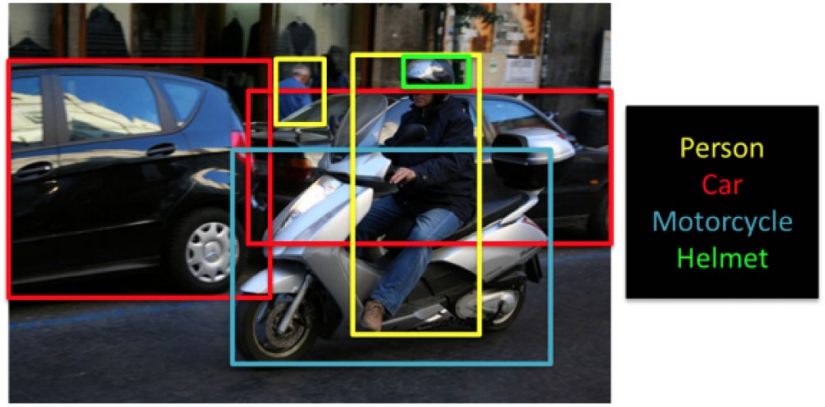
Free or Fee

Q Apps
- Data Insights
- Data Augmentation

H2O.ai DRIVERLESSAI

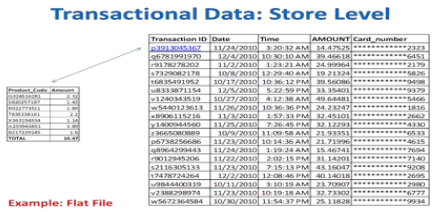
- Feature Engineering
- Normalization
- Basic Encoding
- Advanced Encoding
- Missing Data Imputation
- Binning
- Dimensionality Reduction
- Splitting train/test data

What kind of data can support an ML problem?

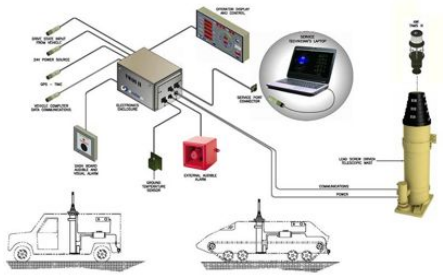


Image, Video, Audio

Transactional

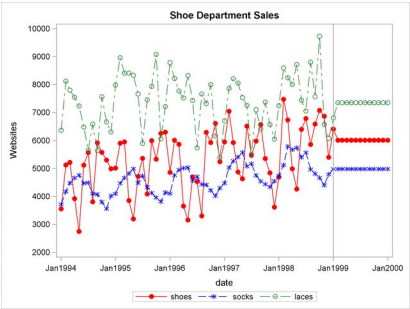


Sensor



Text, Log

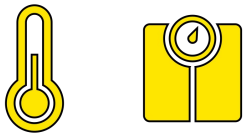
Time Series



Data Types Used in ML Problems

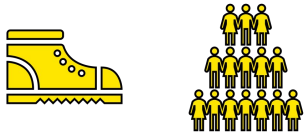
Numerical Data

Continuous



Any value within a range

Discrete



Data with distinct values

Quantitative Data

Categorical Data

Nominal

Categories with no particular order

Ordinal

Categories are ranked or ordered

Qualitative Data

Features

- A feature is an attribute or measurable characteristic that helps to explain a particular record
- The features can be pulled from multiple systems, sources, etc.
- Often a machine learning model can be improved using feature engineering

Module 2

Age	Historical Bill Amounts	Historical Payments Made	Account Balances
35	1000	500	3000
27	500	500	0

Features & Labels

Features

Labels

Observations

Age	Historical Bill Amounts	Historical Payments Made	Account Balances	Target: Did customer make a payment?
35	1000	500	3000	Yes
28	2000	2000	0	Yes
32	500	200	300	No

Some Other Considerations For Preparing Your Dataset

- Size of data available
  - Small amount of data available
  - Large data available (should you sample?)

- Quality
  - Is data reliable? Trustworthy?
  - Does data represent what you are trying to predict?
  - Is the data skewed?
  - Will the data used to train be available at the time of prediction?

- Accounting for rare event problems (Imbalanced Classification Problems)

- Obtaining more data when necessary

Data Preparation Tools

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set for Modeling: Data Exploration & Munging

Part: 3

Rich AI Ecosystem - Too Many Choices?

https://miro.medium.com/max/1400/0*j1pqHqSkGg-sS2x4.png

Rich AI Ecosystem - Too Many Choices?

Cloud Computing

Microsoft Azure | amazon web services | Google Cloud | vmware | IBM Cloud | ORACLE CLOUD

Databases

SQL, snowflake, SAP, ORACLE, neo4j

- Typical frontline store of data (relational, graph, etc)
- May be hosted in cloud if volume of data warrants it

Big Data/Distributed Computing

databricks, hadoop, Apache Spark, elasticsearch

- If data is too big to be useful for accessing it you can use big data platforms for distributed, parallel, high-performance computing

Module 5

Programming Languages

python, R Studio, SQL

- In terms of accessing, isolating, cleaning, transforming data, these are the big 3.
- Python + R are consistently used for DS & modeling

ML Foundations

Business Intelligence

tableau, TIBCO Spotfire, Power BI, Qlik

- Most common resources for descriptive statistics and dashboarding (specialize in descriptive stats)

Data Science/Analytics/AI

python, R, H2O.ai, Azure Machine Learning, Amazon SageMaker

- For predictive & advanced analytic insights use Data Science/AI platforms (and py+R) to apply the highest quality methods.
- Cloud computing may be needed to run heavy math for these models.

ML Foundations

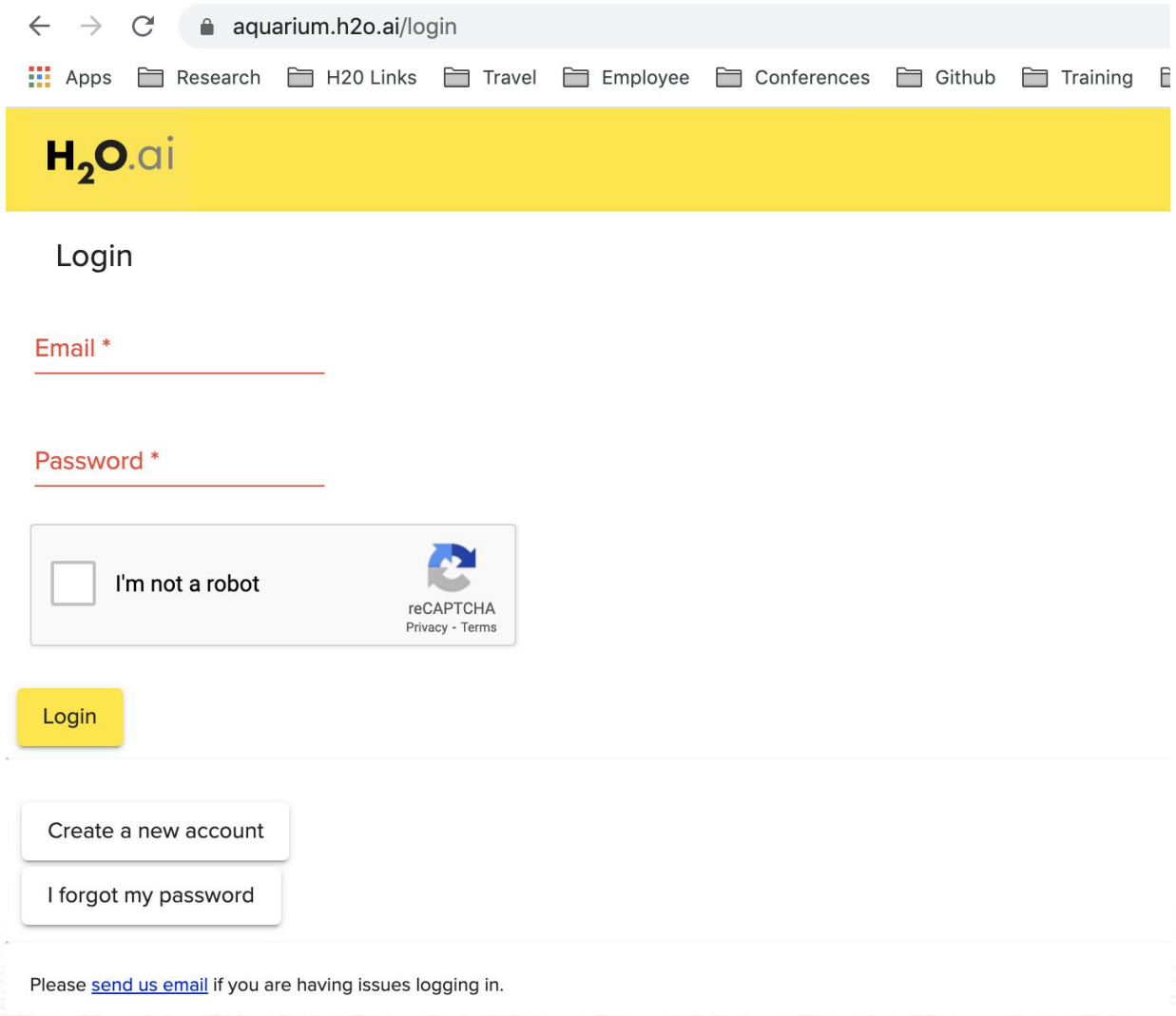
Introduction to Data Preparation Demo

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set for Modeling: Data Exploration & Munging

Part: x

Aquarium Account Creation



H2O.ai's Hands-On Learning System

- Contains 1 to 2 hours labs for H2O-3, Sparkling Water & Driverless AI
- Free account creation and use of labs
- Some labs are pre-packaged with robust examples

Demo

What’s Next?

Module 1: Data Wrangling & Data Curation

Session 1: Getting the Right Data Set for Modeling: Data Exploration & Munging

Part: x

Summary

- Data Preparation can be comprised of many different aspects of dealing with data from collecting it, to exploring it for areas to clean, to performing transformations
- Tools like Pandas, R are great for data preparation for small datasets
- H2O-3 can be used for large data to perform data preparation
- Python datatable & R data.table are great for big data prep
- Driverless AI automates many aspects of data pre-processing and feature engineering & can be extended with data recipes for other data tasks

Recording will be posted w/in 2 days

Upcoming Sessions

1. A recorded session for **Driverless AI Data Recipes** will be released by September 11 as part of this Module 1.
2. The next live session will begin Module 2: Feature Engineering For Machine Learning with **Session 1: Feature Engineering Techniques From an Expert Kaggler** will be held on **Tuesday September 15, 2020 @ 7:00AM PDT**

Additionally a recording for **Aquarium Account Setup & Site Exploration** will be released by September 11.

Quizzes & Study Groups

- Each session within a module will have a small quiz to complete and all quizzes for that module will be due before the next module starts.
- There are 2 options available for you to ask additional questions or get assistance on AI concepts covered in the sessions:
  - A **Study Group** for each Module will be held on **Saturdays @ 10:00AM PDT**
  - **Ask Me Anything** will be held on **Sundays @10:00AM PDT**

- **Reminder:** Don’t forget to complete Quiz 1: Getting the Right Data Set For Modeling (released by Friday September 11) to be on the path to earn your badge!

Resources

datatable vs Pandas

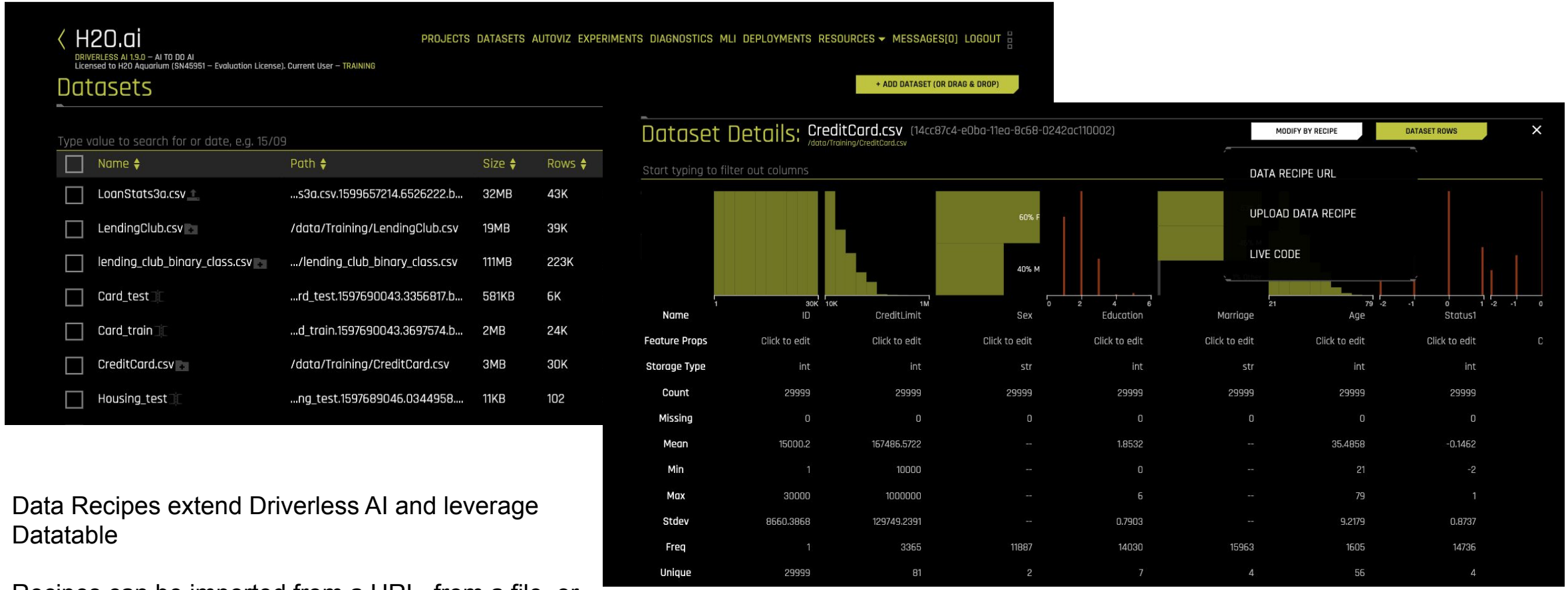
Python datatable

- Python Library
- 2D data frames
- Data munging, manipulation
- Great for **Big Data** (greater than 100GB)
- High performance
- In-memory and out-memory datasets
- Multi-threaded algorithms
- Powerful API similar to R data.table

Pandas

- Python Library
- 2D data frames
- Data munging and preparation
- Extensive data representation
- Great for **small data** (usually 100MB to 1GB)
- For **large data** (less than 100 GB)
  - Usually **low performance**
  - **Long runtime**
  - Insufficient memory usage

Data Preparation Using Driverless AI: Data Recipes



- Data Recipes extend Driverless AI and leverage Datatable
- Recipes can be imported from a URL, from a file, or can be implemented live

Additional Resources

H2O.ai's AI Glossary	Glossary
Public Datasets - H2O.ai	Datasets
Kaggle Datasets	Datasets
Google Dataset Search	Datasets
H2O-3 Documentation - Data Munging	H2O-3 Data Munging Documentation (Python & R)
Python Pandas	Python Pandas Documentation
R Tidyverse	R Data Manipulation Packages
Python datatable documentation	Python Datatable Documentation
R data.table documentation	R data.table Documentation

Additional Resources

Datatable Overview	Blog
Data Analysis With Datatable	Blog
Using A Data Recipe in Driverless AI	Documentation
Driverless AI Data Recipes (Github Repo)	Github
Driverless AI Data Recipes: Live Code (Github Repo)	Github
Introduction to Python Data Wrangling with Python Datatable	Meetup Replay