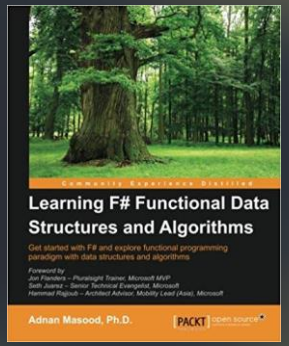
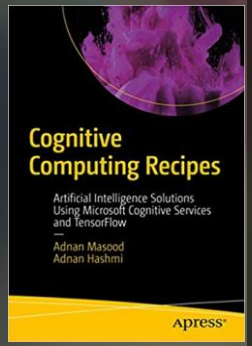
Beyond Hello World - Common Machine Learning Use Cases with ML.NET

Adnan Masood, PhD.

Microsoft Azure
+AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection

About the Speaker



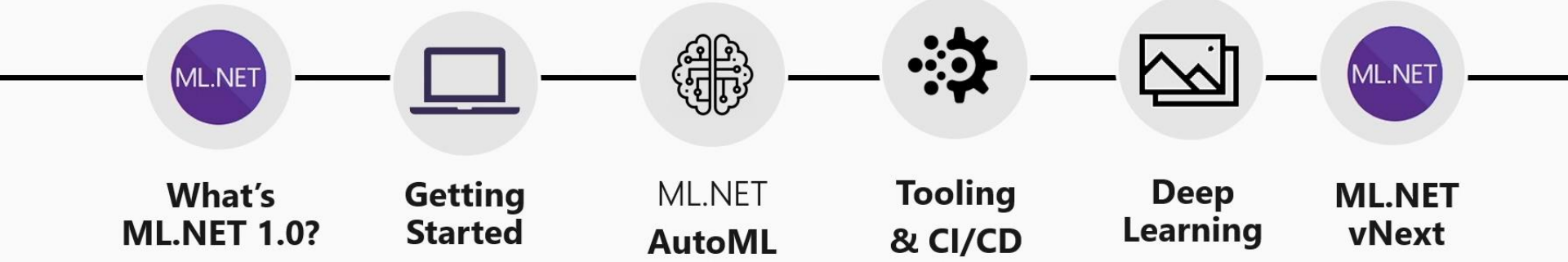


Adnan Masood, Ph.D. is an Artificial Intelligence and Machine Learning researcher, software architect, and Microsoft MVP (Most Valuable Professional) for Artificial Intelligence. As Chief Architect of AI and Machine Learning, at UST Global, he collaborates with Stanford Artificial Intelligence Lab, and MIT AI Lab for building enterprise solutions

Author of Amazon bestseller in programming languages, **"Functional Programming with F#",** Dr. Masood teaches Data Science at Park University, and has taught Windows Communication Foundation (WCF) courses at the University of California, San Diego. He is a regular speaker to various academic and technology conferences (WICT, DevIntersection, IEEE-HST, IASA, and DevConnections), local code camps, and user groups. He also volunteers as STEM (Science Technology, Engineering and Math) robotics coach for elementary and middle school students

A strong believer in giving back to the community, Dr. Masood is a co-founder and president of the Pasadena .NET Developers group, co-organizer of Tampa Bay Data Science Group, and Irvine Programmer meetup. His recent talk at Women in Technology Conference (WICT) Denver highlighted the importance of diversity in STEM and technology areas, and was featured by variety of news outlets.

What are we going to talk about today?



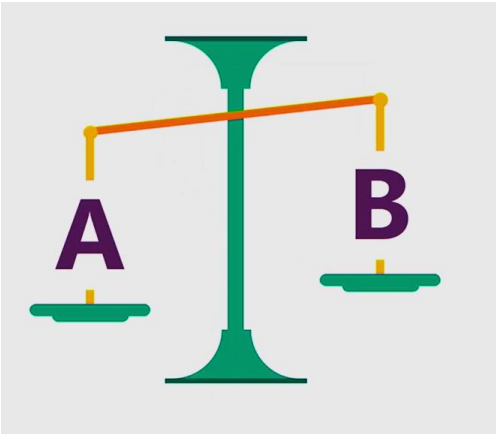
Abstract

- **ML.NET is an open source and cross-platform machine learning framework built for .NET developers. You can use your .NET and C# or F# skills to easily integrate custom machine learning into your applications without any prior expertise in developing or tuning machine learning models.**
- **In this talk we will write code in ML .NET to build the use cases around sentiment analysis, price prediction, Product recommendation, Customer segmentation, Spam detection, and sales forecasting to showcase the power and ease of framework.**

© Microsoft Azure + AI Conference All rights reserved.

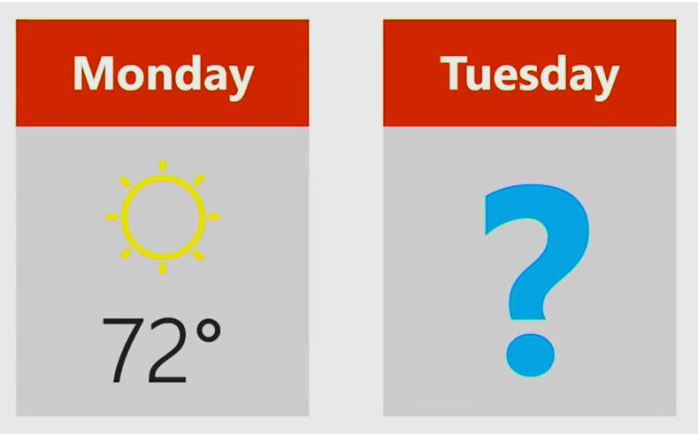
Many ML Tasks

Is this A or B?



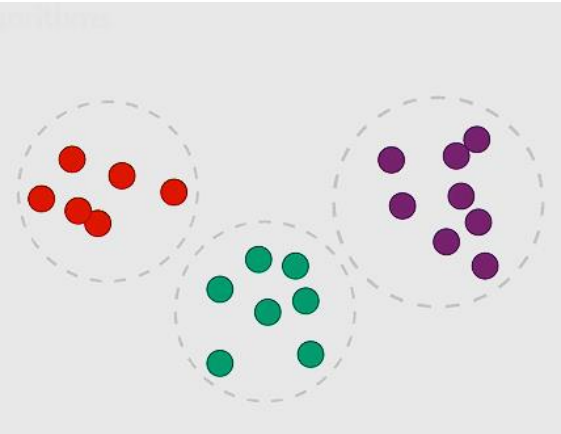
Classification

How much? How many?



Regression

How is this organized?



Clustering

And many more…

ML.NET 1.0 @Build 2019

ANNOUNCED @//BUILD 2019/

ML.NET 1.0

Machine Learning framework for building custom ML Models

Reuse .NET Skills
C# and F# for custom models

Proven at scale
Azure, Office, Windows

Custom ML made easy
Automated ML and Tools (Model Builder and CLI)

Extensible
TensorFlow, ONNX and Infer.NET

Free, Cross-platform and Open Source

A few things you can do with ML.NET 1.0 ...

Sentiment analysis

Analyze the sentiment of customer reviews using a binary classification algorithm.

Sentiment analysis sample >

Product recommendation

Recommend products based on purchase history using a matrix factorization algorithm.

Product recommendation sample >

Price prediction

Predict taxi fares based on distance traveled etc. using a regression algorithm.

Price prediction sample >

Customer segmentation

Identify groups of customers with similar profiles using a clustering algorithm.

Customer segmentation sample >

GitHub labeler

Suggest the GitHub label for new issues using a multi-class classification algorithm.

GitHub labeler sample >

Fraud detection

Detect fraudulent credit card transactions using a binary classification algorithm.

Fraud detection sample >

Spam detection

Flag text messages as spam using a binary classification algorithm.

Spam detection sample >

Image classification

Classify images (e.g. broccoli vs pizza) using a TensorFlow deep learning algorithm.

Image classification sample >

Sales forecasting

Forecast future sales for products using a regression algorithm.

Sales forecasting sample >

You can find more ML.NET samples on GitHub, or take a look at the ML.NET tutorials.

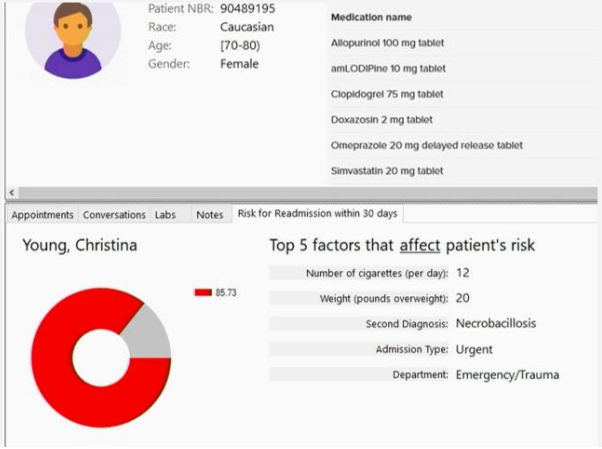
ML.NET enables interpretable machine learning

Explain & debug model predictions to build trust

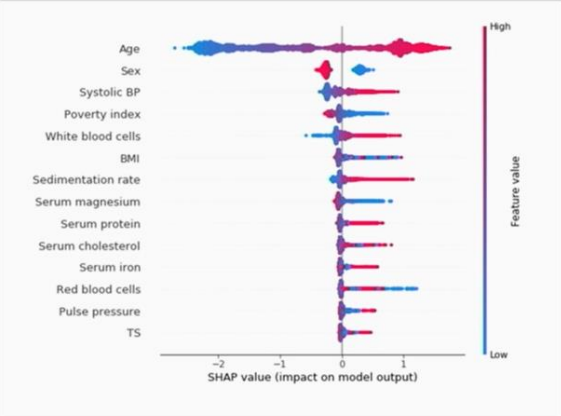
Meet regulatory requirements (health care, finance, more)

Ensure fairness & remove biases (address issues in data)

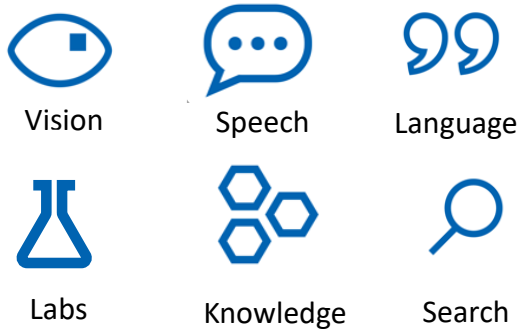
How do features impact specific prediction?



Which features are important in your data?



Pre-built machine learning models (i.e. Azure Cognitive Services)



Consume (C#, VB, F#)

e.g. Sentiment Analysis using Azure Cognitive Services

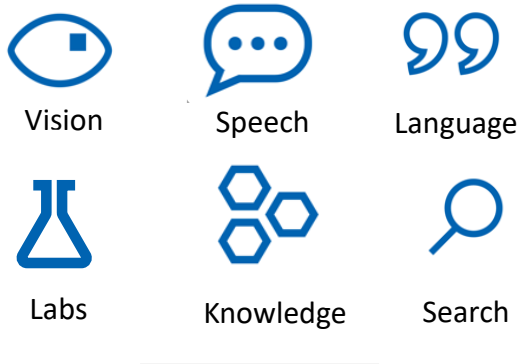
```
TextAnalyticsAPI client = new TextAnalyticsAPI();
client.AzureRegion = AzureRegions.Westus;
client.SubscriptionKey = "1bf33391DeadFish";

client.Sentiment(
new MultiLanguageBatchInput(
new List<MultiLanguageInput>()
{
 new MultiLanguageInput("en","0",
     "This is a great vacuum cleaner")
}));
```

96% positive

© Microsoft Azure + AI Conference All rights reserved.

Limitations with pre-built machine learning models



e.g. Sentiment Analysis using Azure Cognitive Services

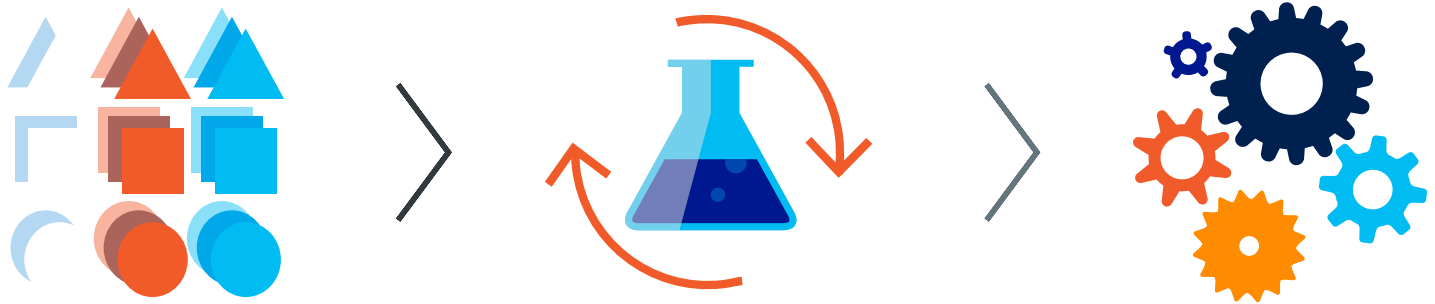
```
TextAnalyticsAPI client = new TextAnalyticsAPI();
client.AzureRegion = AzureRegions.Westus;
client.SubscriptionKey = "1bf33391DeadFish";

client.Sentiment(
new MultiLanguageBatchInput(
new List<MultiLanguageInput>()
{
new MultiLanguageInput("en","0",
    "This vacuum cleaner sucks so much dirt")
}));
```

☹ 9% positive

© Microsoft Azure + AI Conference All rights reserved.

Build your own custom machine learning models



Prepare Your Data

Build & Train

Run

© Microsoft Azure + AI Conference All rights reserved.

ML.NET is a framework for building custom ML Models

Developer friendly APIs for Machine Learning

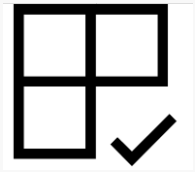
Training & Consumption

Transforms	Learners	Misc.	Extensions
Text	Linear	ML Data framework	
Schema	Boosted Trees	Evaluators	
Missing values	Svm	Calibrators	
Categorical	K-Means	Data loaders	
Normalization			
Feature Selection			

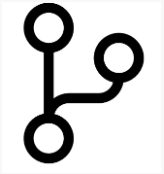
ML.NET is a machine learning framework made for .NET developers

Supported on Windows, Linux, and macOS

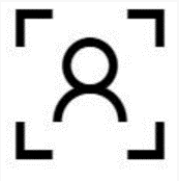
Build your own



Proven & Extensible



Open Source



Developer Focused

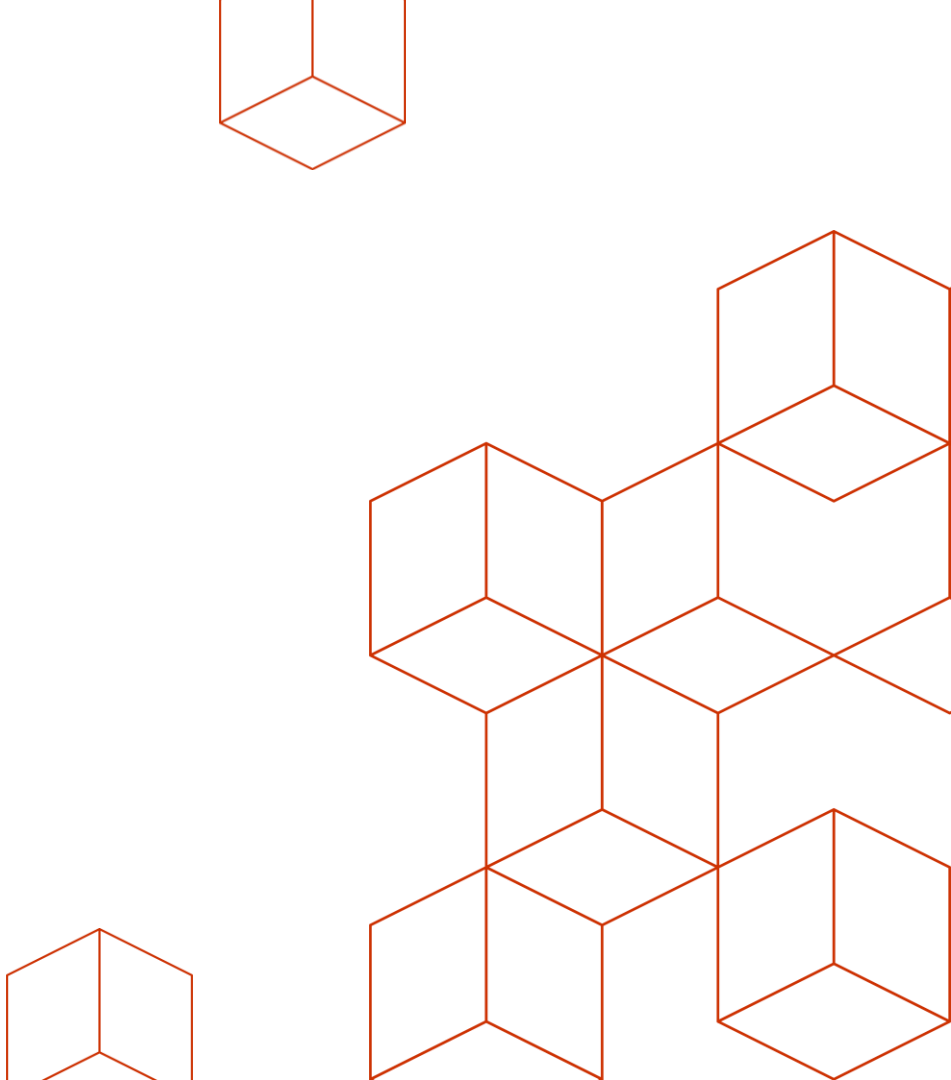
ML.NET 1.1 is available @ https://github.com/dotnet/machinelearning

A few things you can do with ML.NET …

- Sentiment Analysis
- Forecasting
- Issue Classification
- Predictive maintenance
- Image classification
- Recommendations
- Object detection
- Customer segmentation

And more! Samples @ https://github.com/dotnet/machinelearning-samples

Getting Started with Sentiment Analysis



Sentiment Analysis

Comment	Toxic? (Sentiment)
==RUDE== Dude, you are rude …	1
== OK! == IM GOING TO VANDALIZE …	1
I also found use of the word "humanists" confusing …	0
Oooooh thank you Mr. DietLime …	0

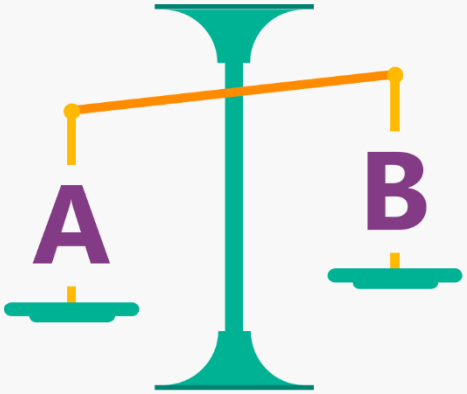
Features (input)

Label (output)

Wikipdia detox data at https://figshare.com/articles/Wikipedia_Talk_Labels_Personal_Attacks/4054689

Sentiment analysis explained

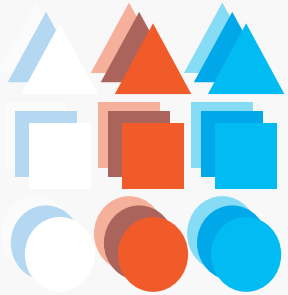
Is this A or B?

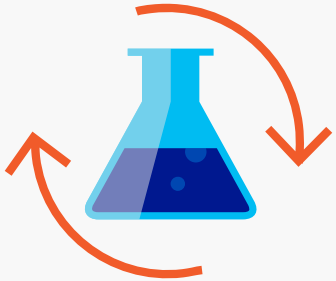


Is this a toxic comment?

Yes or no

Machine learning workflow





Prepare Your Data

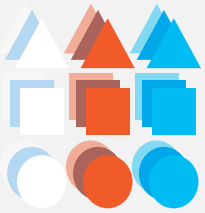
Build & Train

Run

Load Data → Extract Features → Train Model → Evaluate Model → Model Consumption

Important concepts: Data

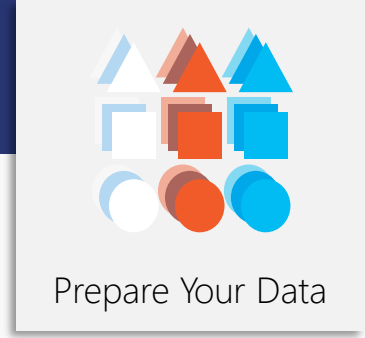
Data

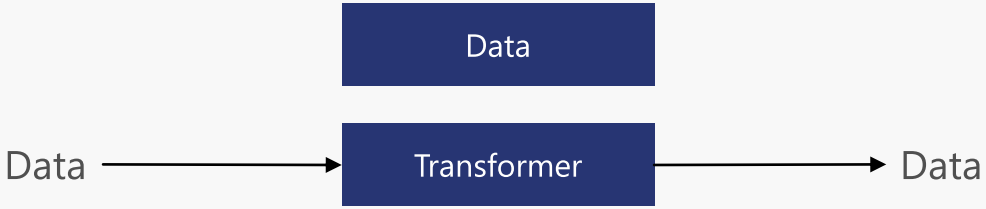


Example

Comment	Toxic? (Sentiment)
==RUDE== Dude, you are rude …	1
== OK! == IM GOING TO VANDALIZE …	1
I also found use of the word "humanists" confusing …	0
Oooooh thank you Mr. DietLime …	0

Important concepts: Transformer





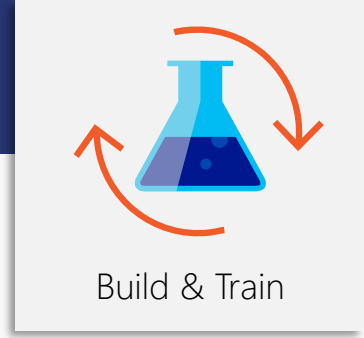
Example

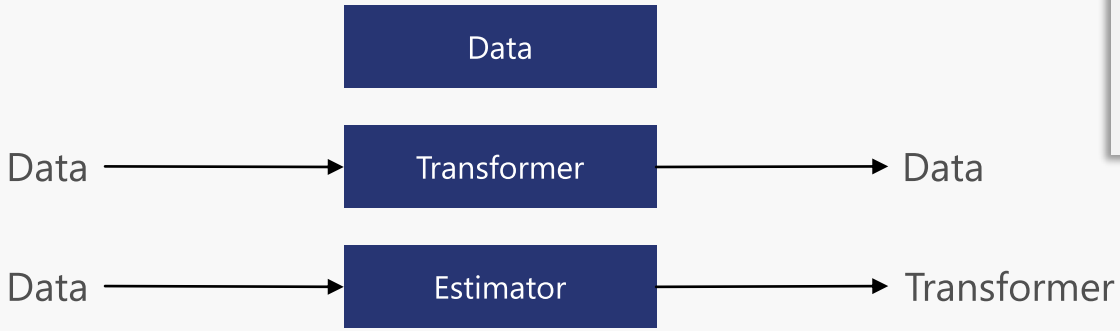
Text
==RUDE== Dude, you are rude …
== OK! == IM GOING TO VANDALIZE …
I also found use of the word "humanists" …
Oooooh thank you Mr. DietLime …

Text Featurizer

Featurized Text
[0.76, 0.65, 0.44, …]
[0.98, 0.43, 0.54, …]
[0.35, 0.73, 0.46, …]
[0.39, 0, 0.75, …]

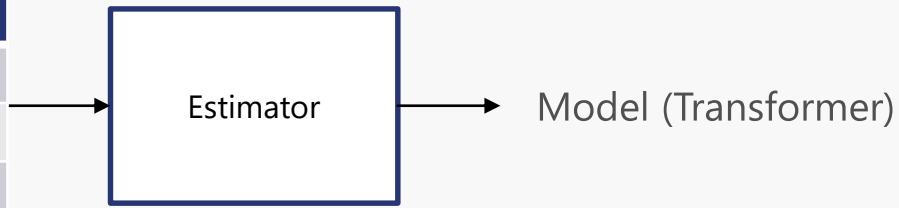
Important concepts: Estimator



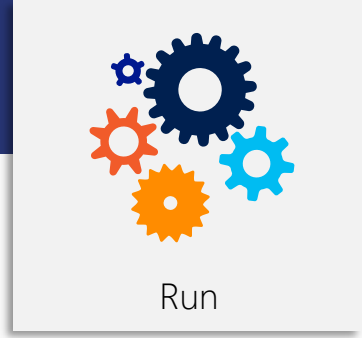


Example

Comment	Toxic? (Sentiment)
==RUDE== Dude, you …	1
== OK! == IM GOING …	1
I also found use of the …	0
Oooooh thank you Mr. …	0



Important concepts: Prediction Function



Data

Data → Transformer → Data

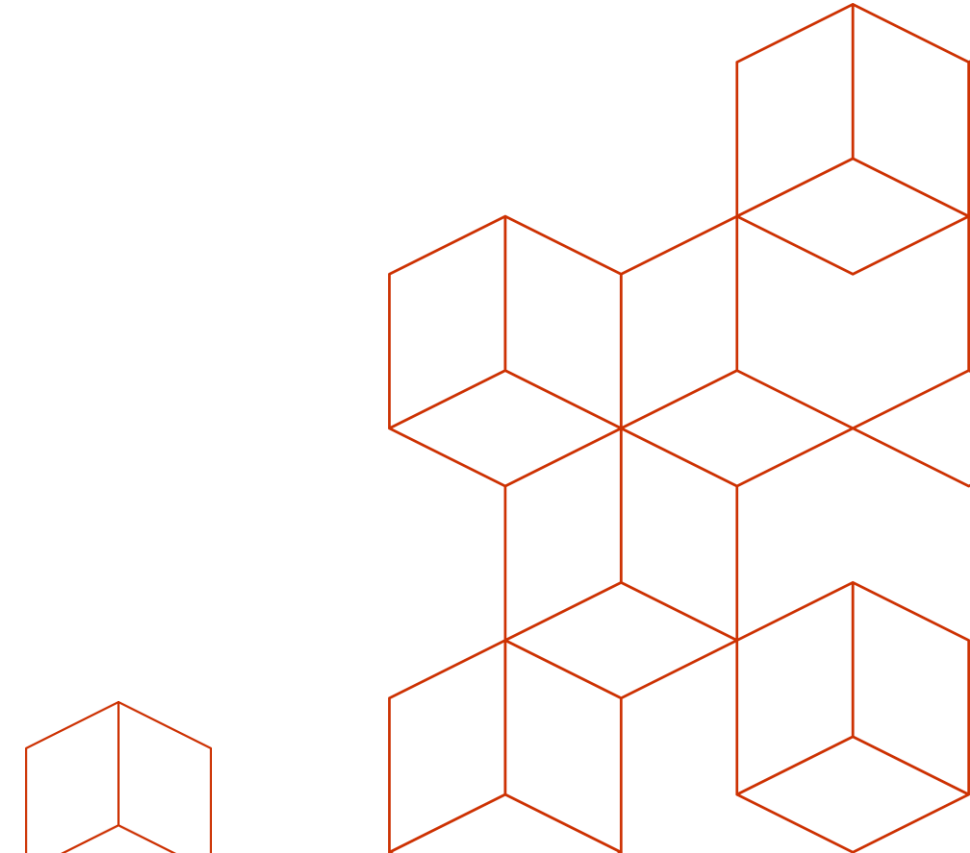
Data → Estimator → Transformer (e.g. model)

Data (single row) → Prediction Function → Data (single row)

Example

Comment	Prediction Function	Predicted Label – Toxic? (Sentiment)
==RUDE== Dude, you …		1

Demo: Sentiment Analysis



GitHub issue classification

dotnet / corefx

Watch 1,567 | Star 13,864 | Fork 3,983

Code | Issues 2,209 | Pull requests 58 | Projects 0 | Wiki | Insights

IsolatedStorage fails on Mac if ~/.local/share directory does not exist #29354

New issue

Open **jasongin** opened this issue 5 days ago · 13 comments

jasongin commented 5 days ago

Repro steps:

1. Start with a clean Mac OS machine, or at least a clean user account on Mac that does not have any `~/.local/share` directory created yet.
2. From a .NET Core app, call `IsolatedStorageFile.GetUserStoreForAssembly()`

Result

Assignees

maryamariyan

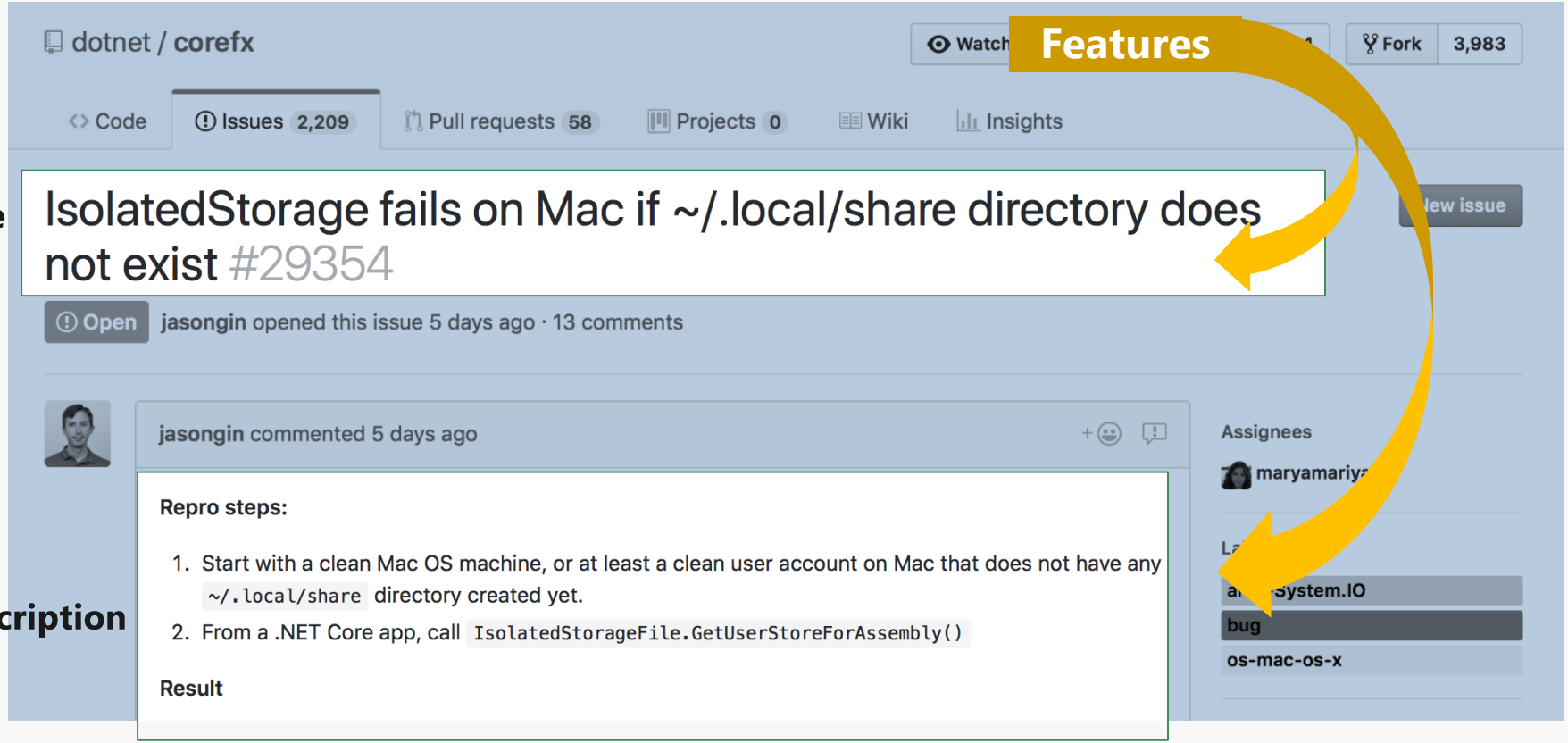
Labels

area-System.IO

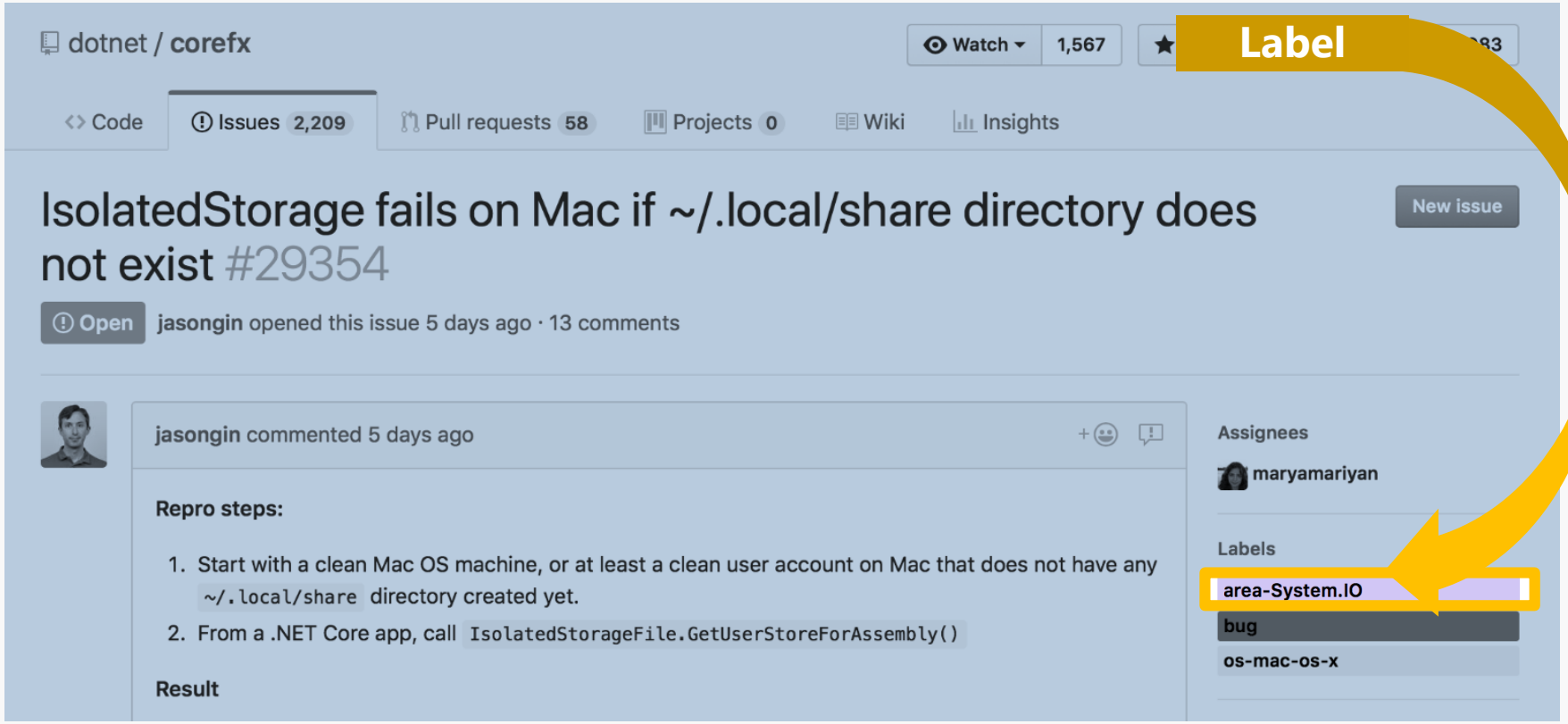
bug

os-mac-os-x

Features (input)

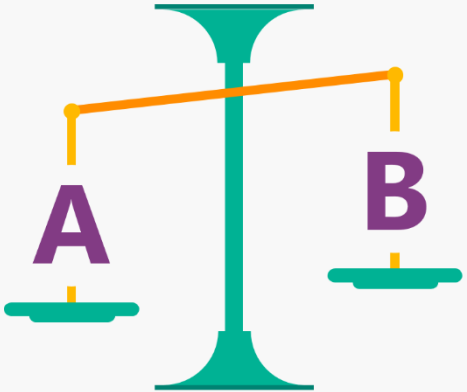


Label (output)



GitHub issue classification explained

Is this A or B?



Which label should this issue be assigned?

area-Microsoft.CSharp

area-Microsoft.VisualBasic

area-Microsoft.Win32

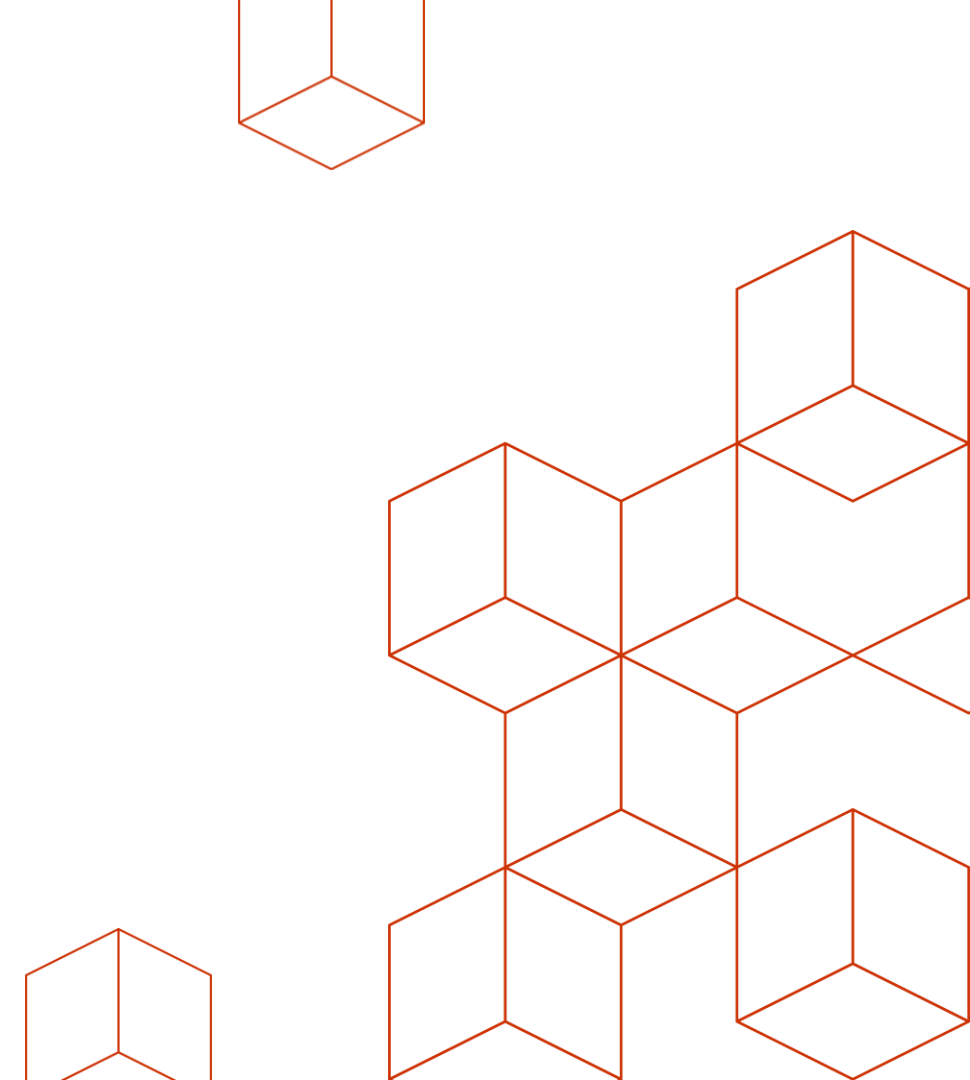
area-Serialization

area-System.AppContext

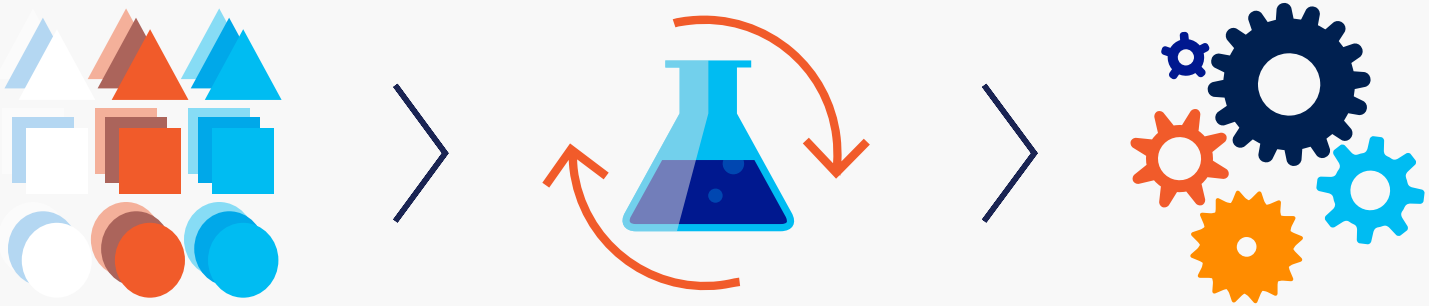
area-System.Buffers

area-System.CodeDom

Demo: Issue Classification (GitHub)



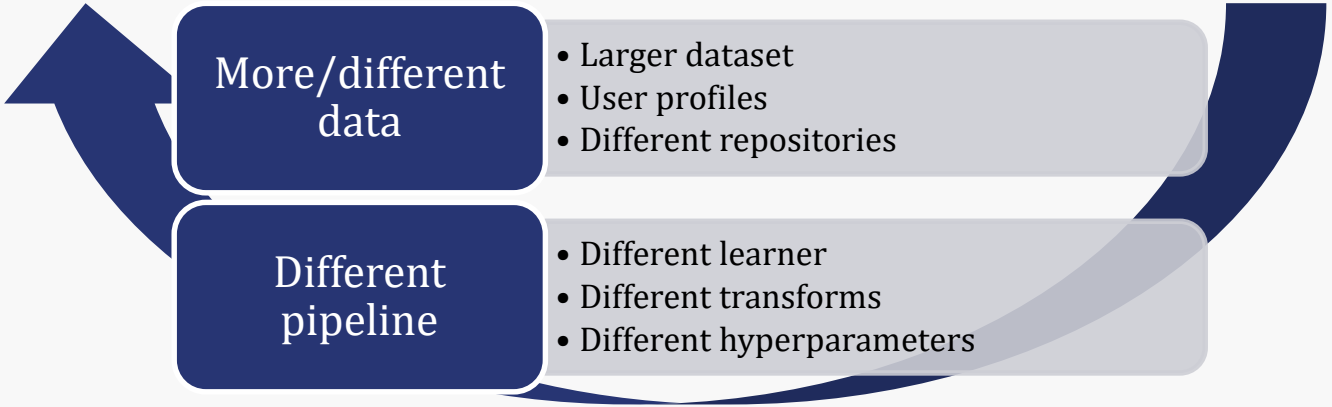
Machine learning is iterative



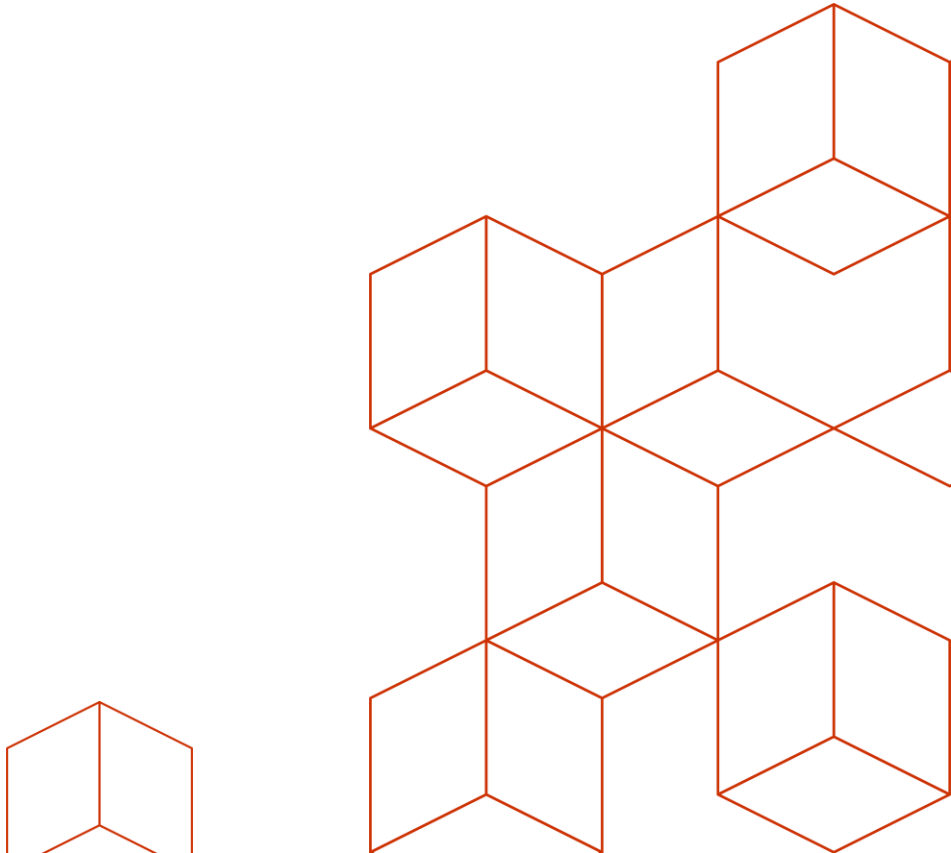
Prepare Your Data

Build & Train

Run



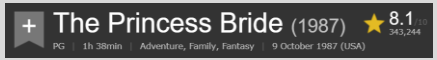
Movie Recommendations



Approaches to Movie Recommendations

1. Population Averages

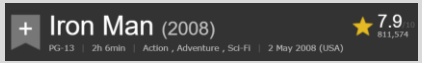
e.g. IMDB, Yelp



IMDB Rating > 8.0, Recommend Movie!

2. Content Based Filtering

e.g. Movie genre



Genre = Superhero movies, Recommend!

3. Collaborative Filtering

Approaches to Movie Recommendations

3. Collaborative filtering

If a person A (e.g. Gal) has the same opinion as a person B (e.g. Cesar) on an issue,
A (Gal) is more likely to have B's (Cesar) opinion on a different issue than that of a random person

	Heat	Mission Impossible	Home Alone	Terminator 2	Casino Royale
Ankit	×	×	✓	✓	×
Gal	✓	✓	**×**	✓	**?**
Cesar	✓	✓	×	✓	✓

What's the probability of Gal liking Casino Royale?

ML.NET 0.3 provides support for Collaborative Filtering (Factorization Machines)

Approaches to Movie Recommendation

Which Dataset to use?

Movie Lens Dataset

20 million ratings, 27,000 movies across 138,000 users.

https://grouplens.org/datasets/movielens/

Ratings.csv

UserId	movieId	Rating	TimeStamp
1	1	2	1094785740
1	2	4.0	1094785734
1	6	4.0	1112485573
1	10	4.0	1112484703

…

Movies.csv

movieID	Title	Genre
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
6	Heat (1995)	Action|Crime|Thriller
10	GoldenEye (1995)	Action|Adventure|Thriller

Features (input)

Ratings

UserId	movieId	Rating	TimeStamp
1	1	0	1094785740
1	2	1	1094785734
1	6	1	1112485573
1	10	1	1112484703

Movies

movieID	Title	Genre
1	Toy Story (1995)	Adventure|Animation| Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children| Fantasy
6	Heat (1995)	Action|Crime|Thriller
10	GoldenEye (1995)	Action|Adventure|Thriller

Label (output)

Ratings

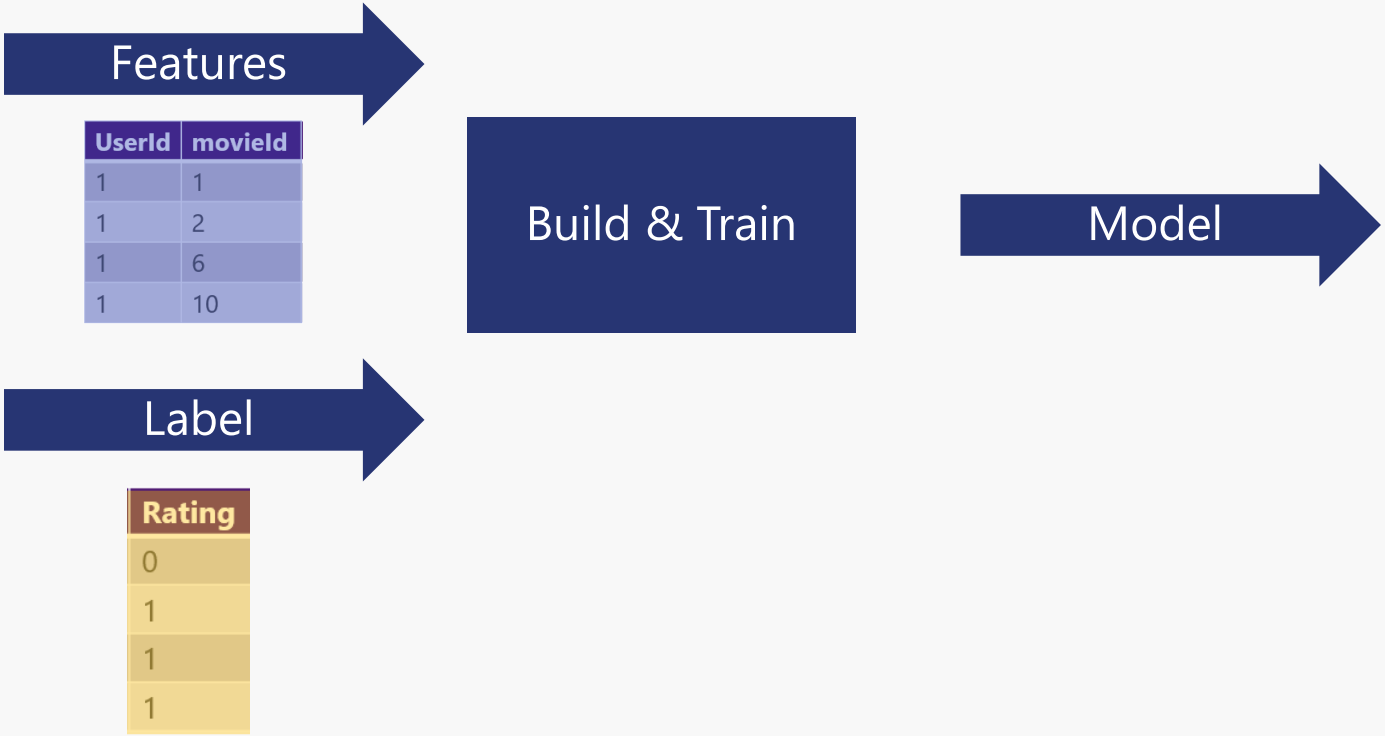
UserId	movieId	Rating	TimeStamp
1	1	0	1094785740
1	2	1	1094785734
1	6	1	1112485573
1	10	1	1112484703

Movies

movieID	Title	Genre
1	Toy Story (1995)	Adventure|Animation| Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children| Fantasy
6	Heat (1995)	Action|Crime|Thriller
10	GoldenEye (1995)	Action|Adventure|Thriller

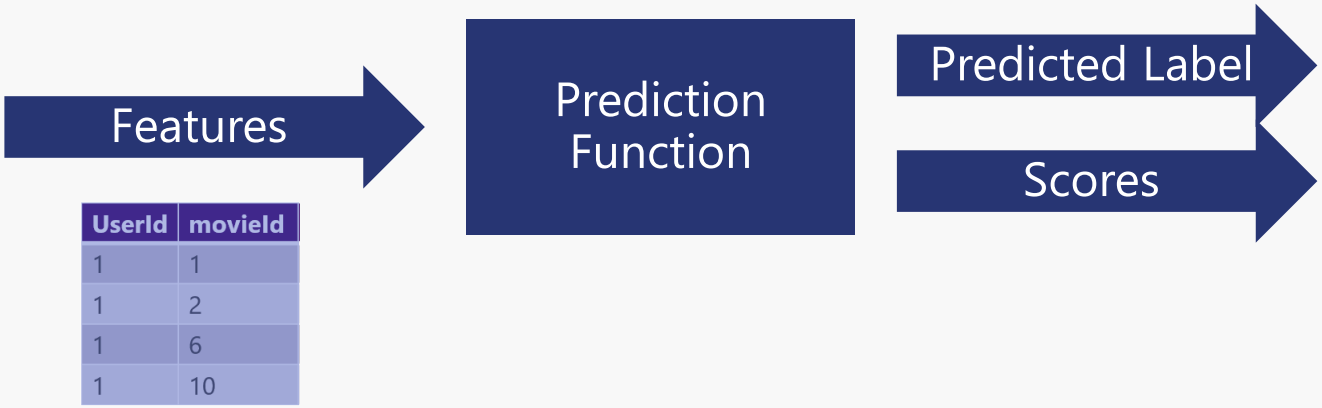
Movie Recommendation with ML.NET

Training



Movie Recommendation with ML.NET

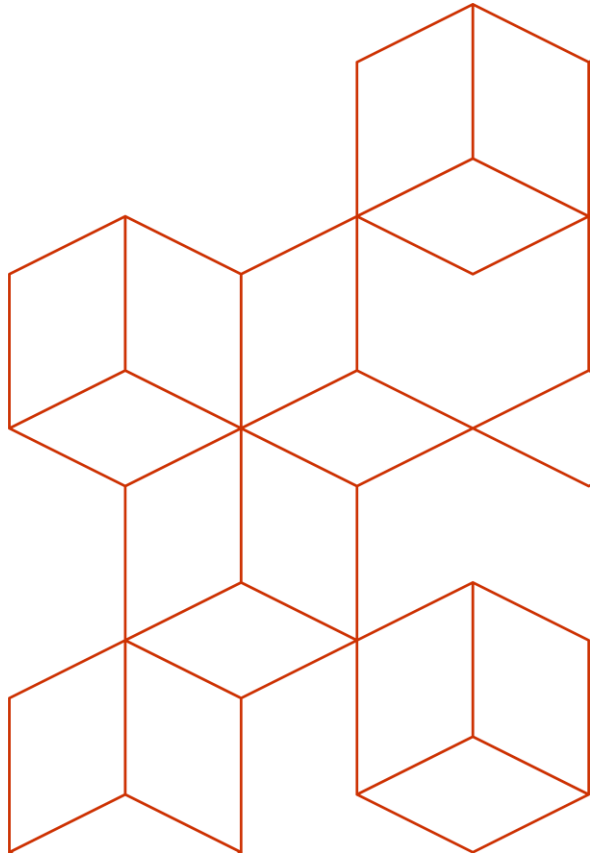
Prediction



UserId	movieId
1	1
1	2
1	6
1	10

Movie Recommendations

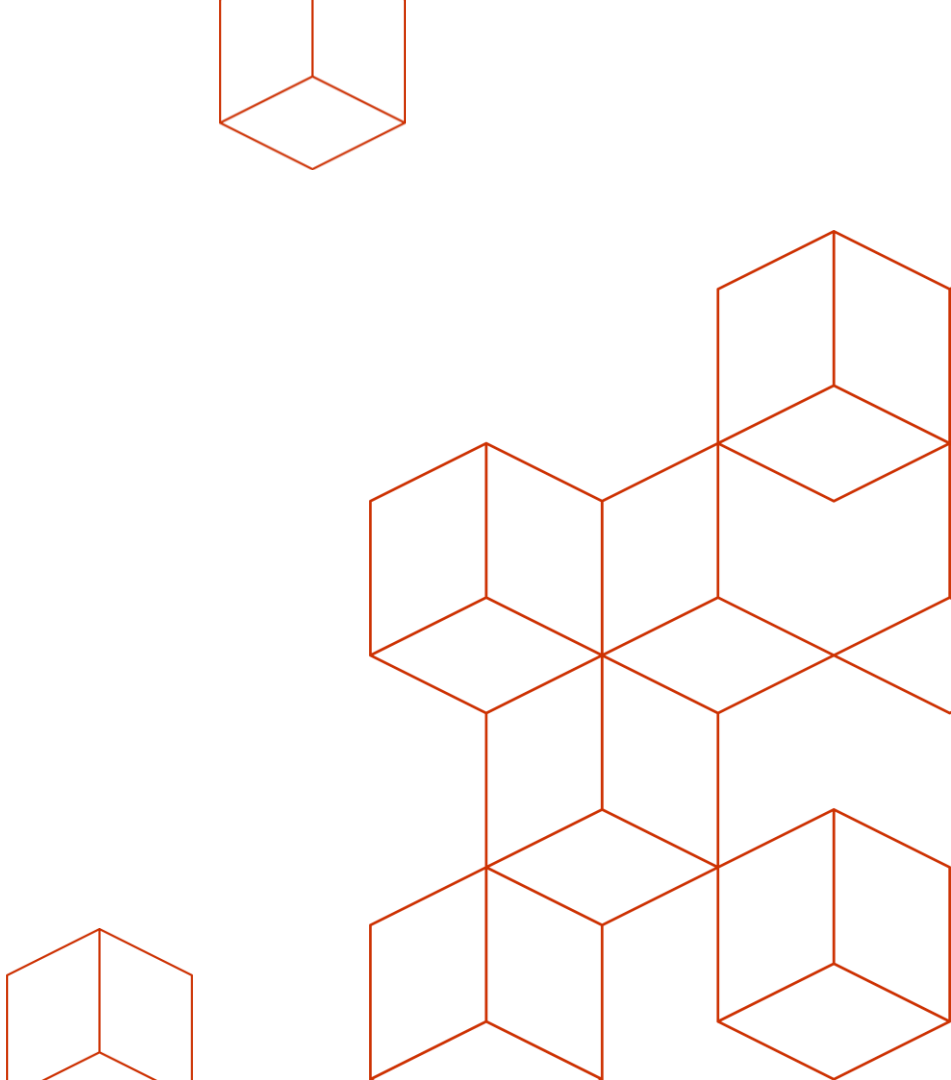
Demo: ML.NET + Factorization Machines



Field Aware Factorization Machines

- Popular in Click Prediction and Recommender Systems competitions
- Combines advantages from Support Vector Machines & Factorization Methods

Deep Learning

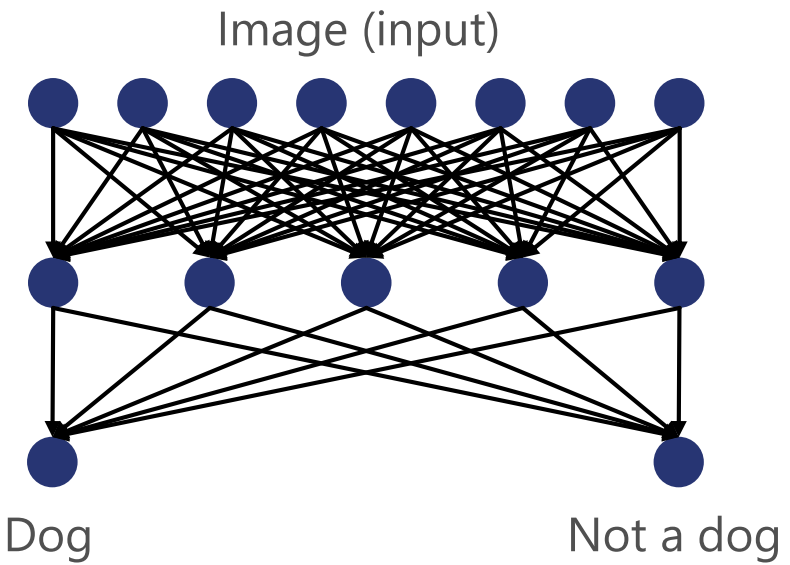


Deep learning

- Revolutionizing areas like computer vision and speech recognition
- Takes advantage of large amounts of data and compute

Dog

Not a dog



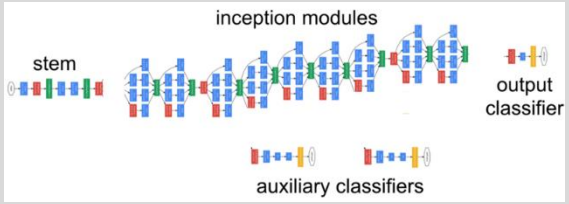
Deep learning with ML.NET

+

TensorFlow

Predicting Image classes with ML.NET + TensorFlow

- **TensorFlow: popular deep learning framework**
- **Pretrained model: Inception**
  - 1000 classes e.g. cats, cars, breeds of dogs
- **Implemented as a transformer in ML.NET**



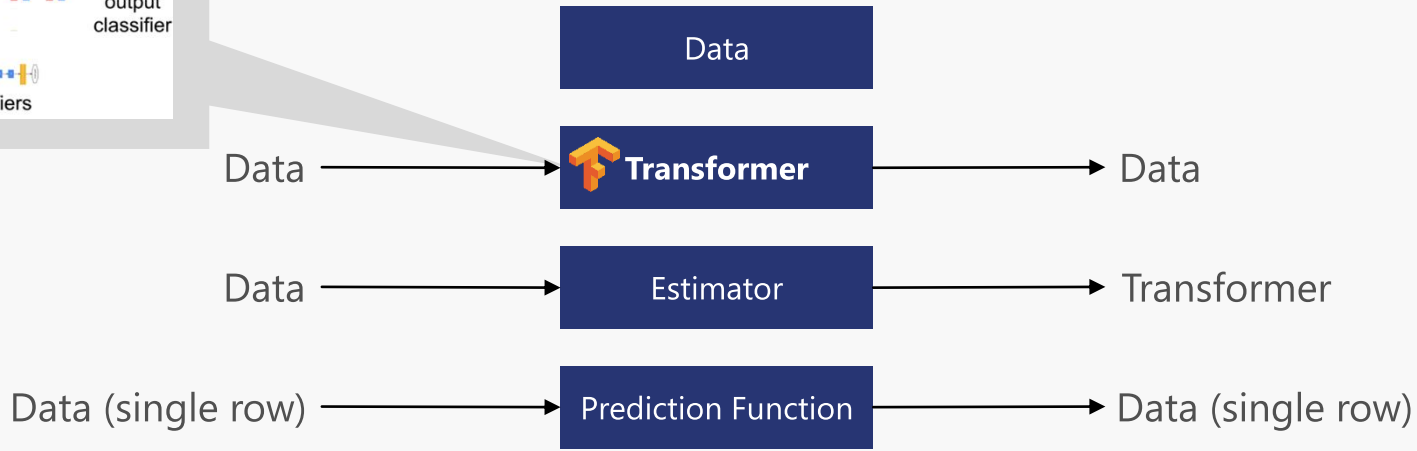
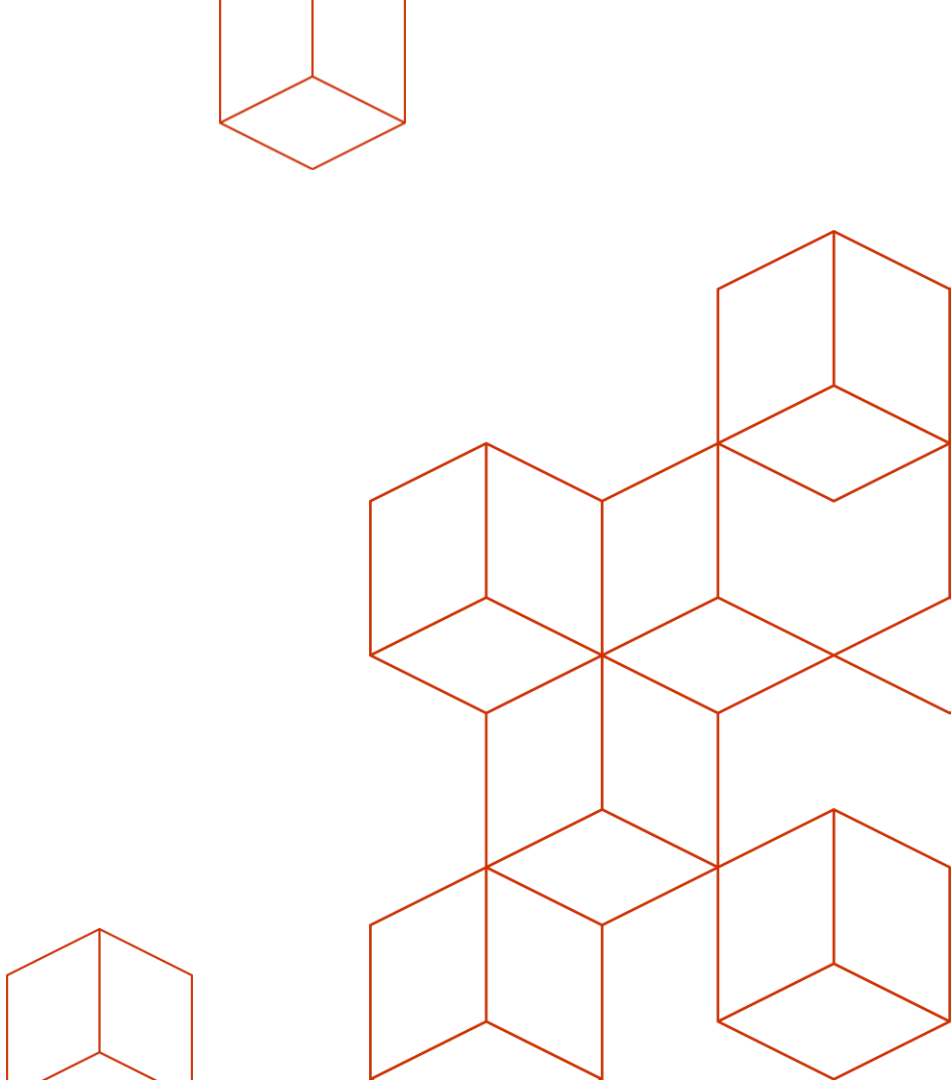


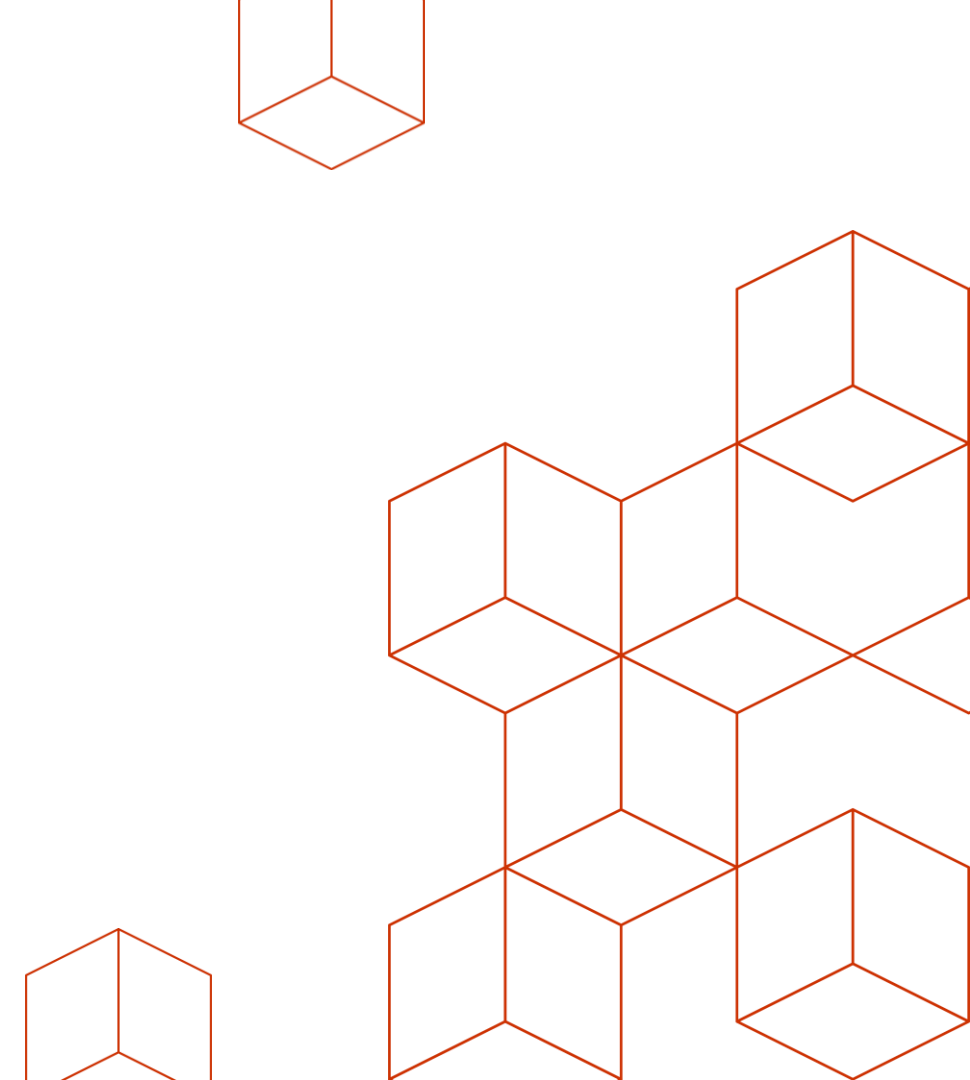
Image Classification

Demo: TensorFlow + ML.NET



Style Transfer

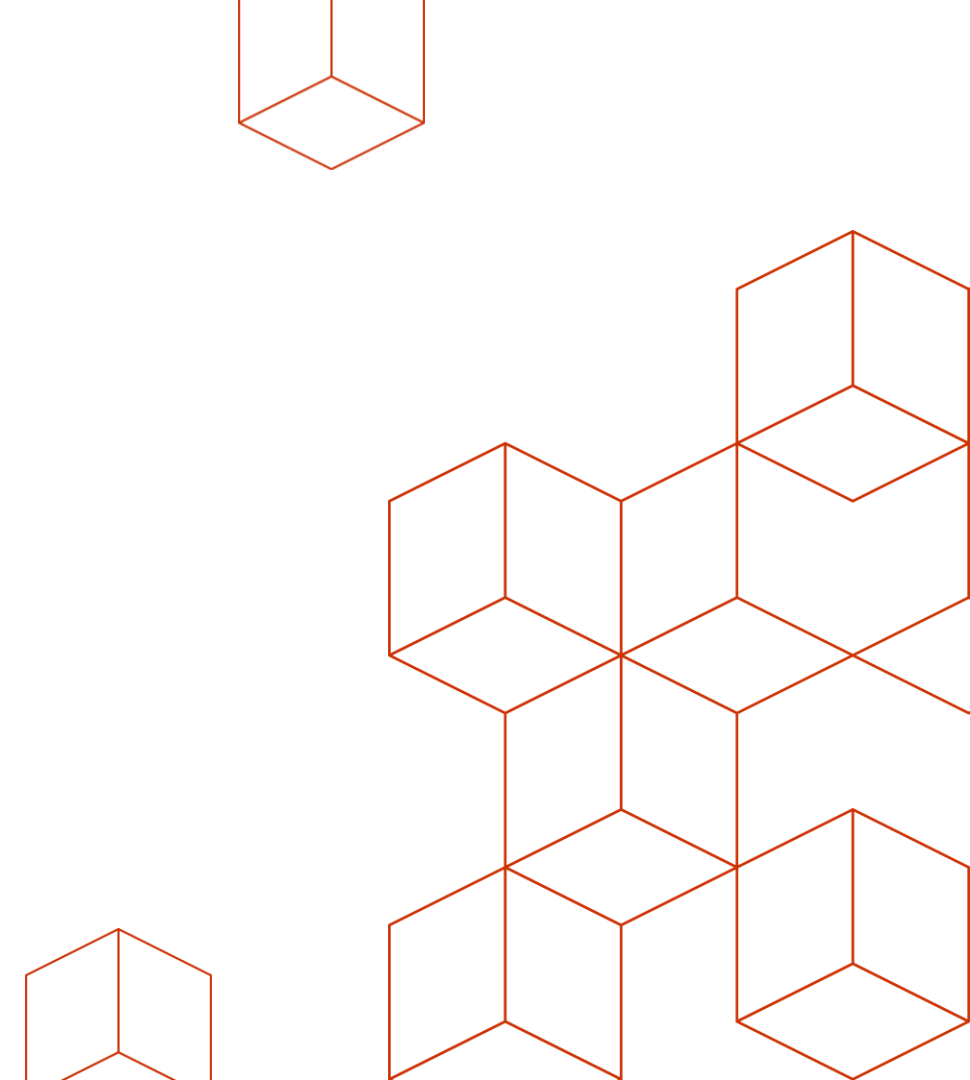
Demo: TensorFlow + ML.NET



ONNX: Open and interoperable AI

Emotion Detection

Demo: ONNX + ML.NET



Deep learning with ML.NET

+

`ApplyTensorFlowGraph`

`ApplyONNXGraph`

A few things you can do with ML.NET …

	Issue Classification ✓		Forecasting
	Sentiment Analysis ✓		Predictive maintenance
	Image classification ✓		Recommendations ✓
	Object detection		Customer segmentation

And more! Samples @ https://github.com/dotnet/machinelearning-samples

What's next with ML.NET?

- API improvements
- Additional ML Tasks and Scenarios
- Improved Deep Learning with TensorFlow
- Scale-out on Azure
- Better GUI to simplify ML tasks
- Improved tooling in Visual Studio
- Improvements for F#
- Language Innovation for .NET

Try ML.NET today!

Get started at http://dot.net/ml

Try the samples at http://aka.ms/mlnetsamples

Read the docs at http://aka.ms/mlnetdocs

Request features or contribute at http://aka.ms/mlnet

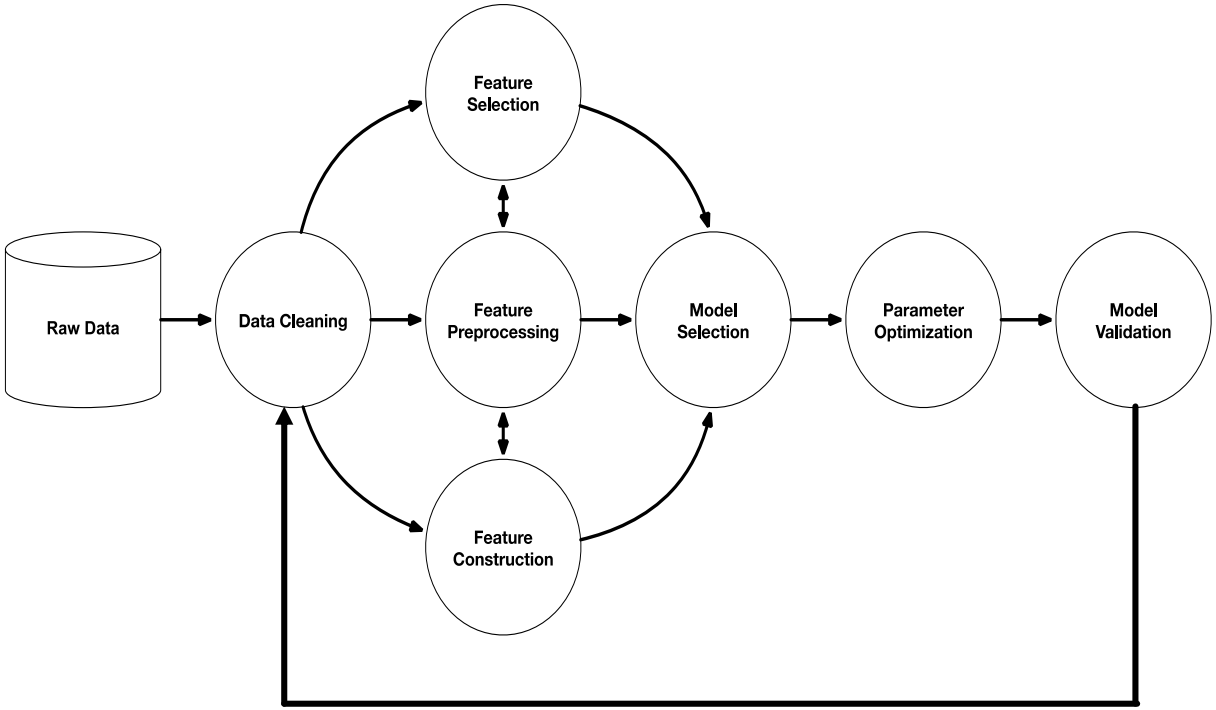
“AutoML is a quiet revolution in AI…”

Automated Machine Learning—A Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb

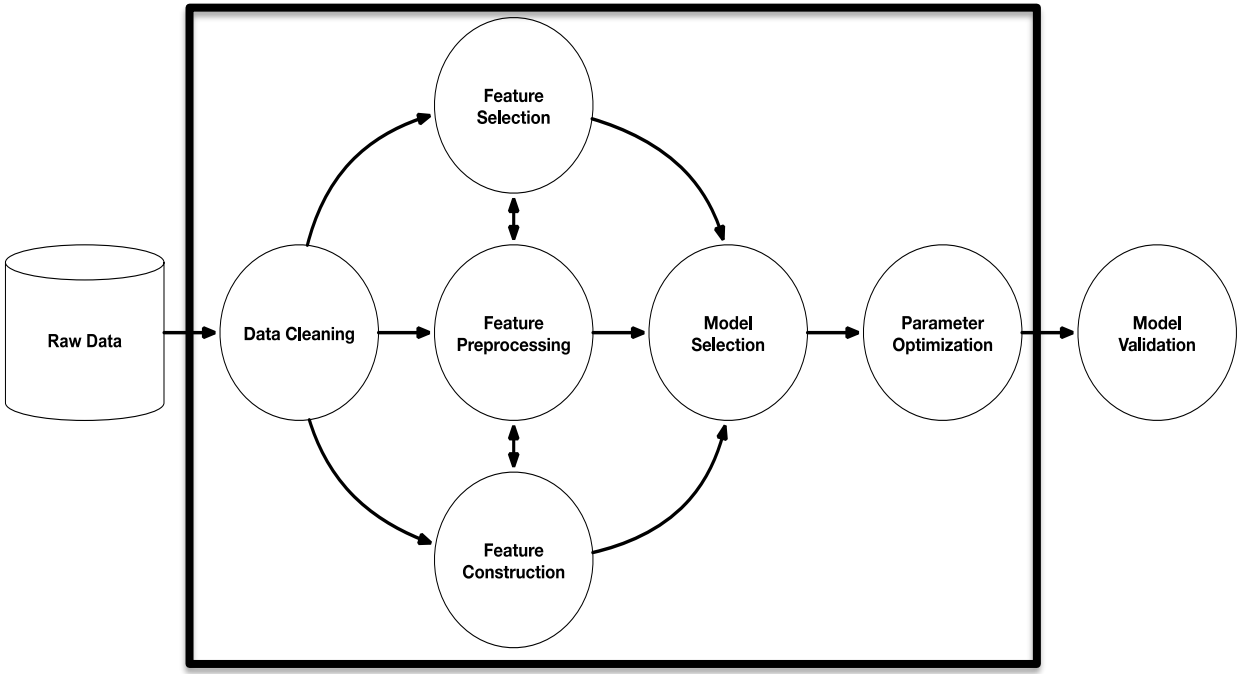
Building A.I. That Can Build A.I.

Google and others, fighting for a small pool of researchers, are looking for automated ways to deal with a shortage of artificial intelligence experts.

ML still requires a lot of manual programming

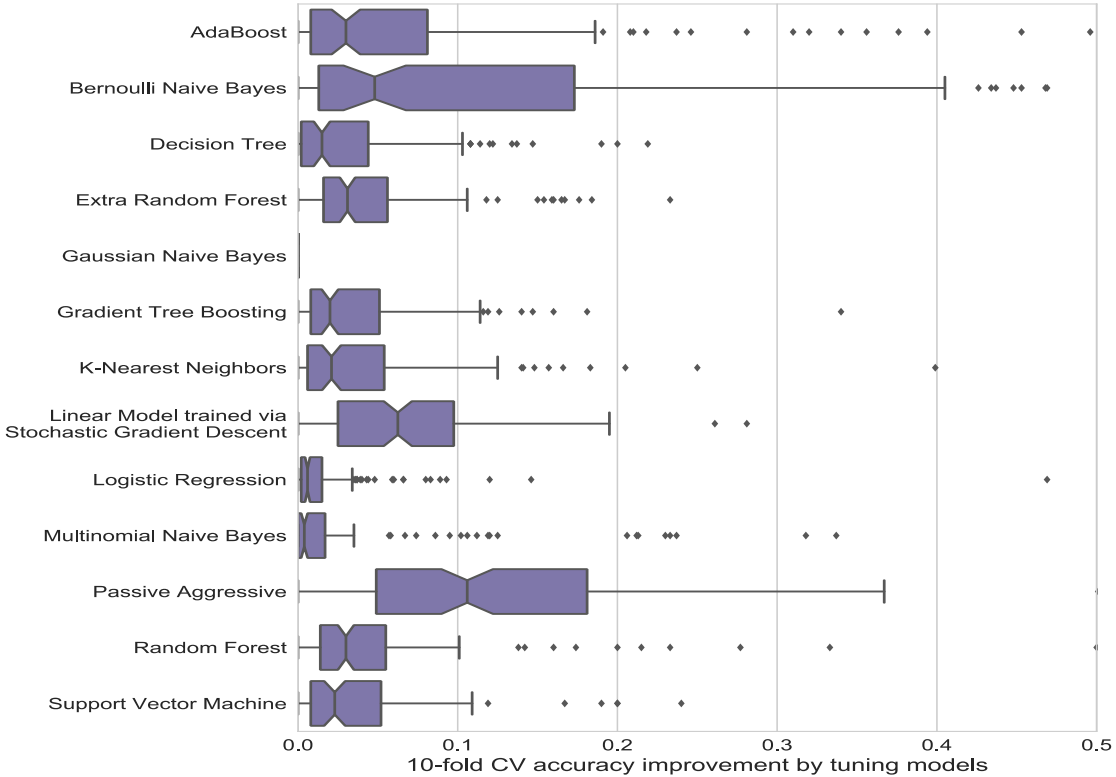


AutoML aims to automate the entire ML workflow



Default parameters are almost always bad

AutoML handles this for you!



AutoML is a huge time-saver

AutoML handles (some of) this for you!

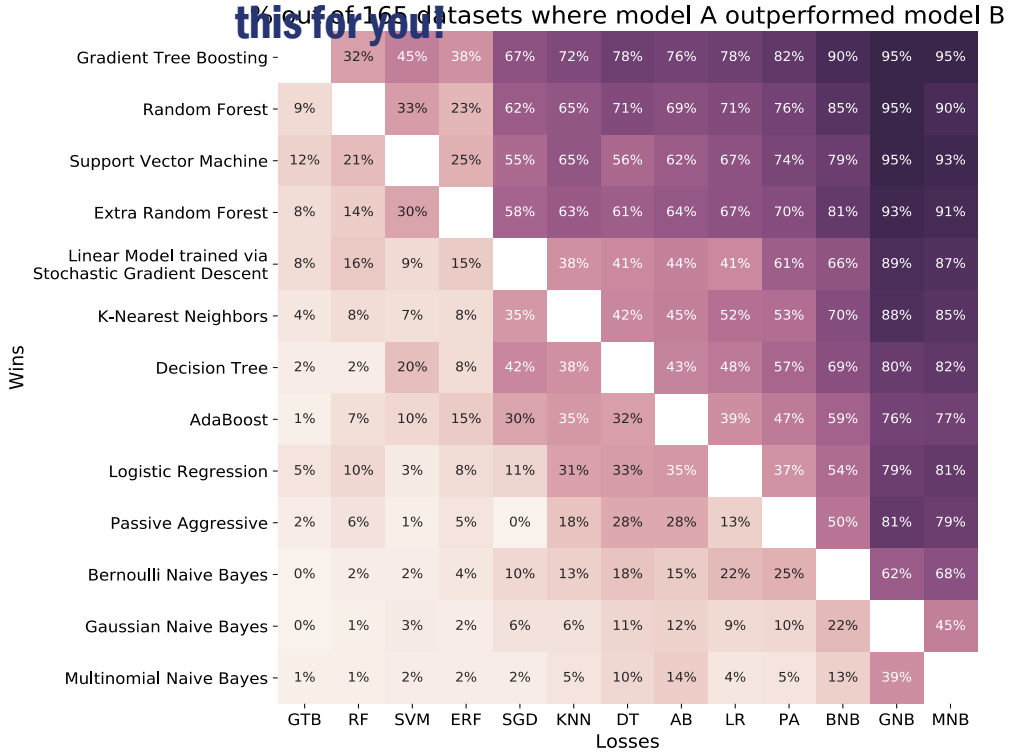


Image source: R. Olson & W. La Cava *et. al.* (2017) “Data-driven advice for applying machine learning to bioinformatics problems.”

AutoML is a huge time-saver

AutoML handles (some of) this for you!

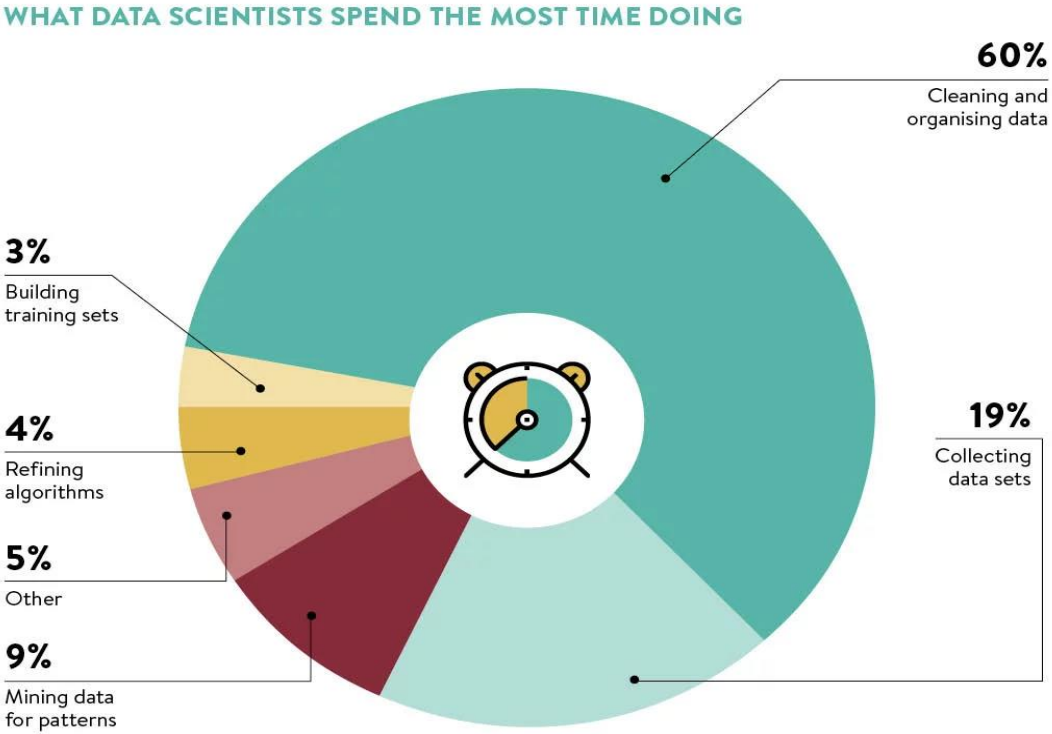


Image source: visit.crowdflower.com/data-science-report.html

The business case for AutoML

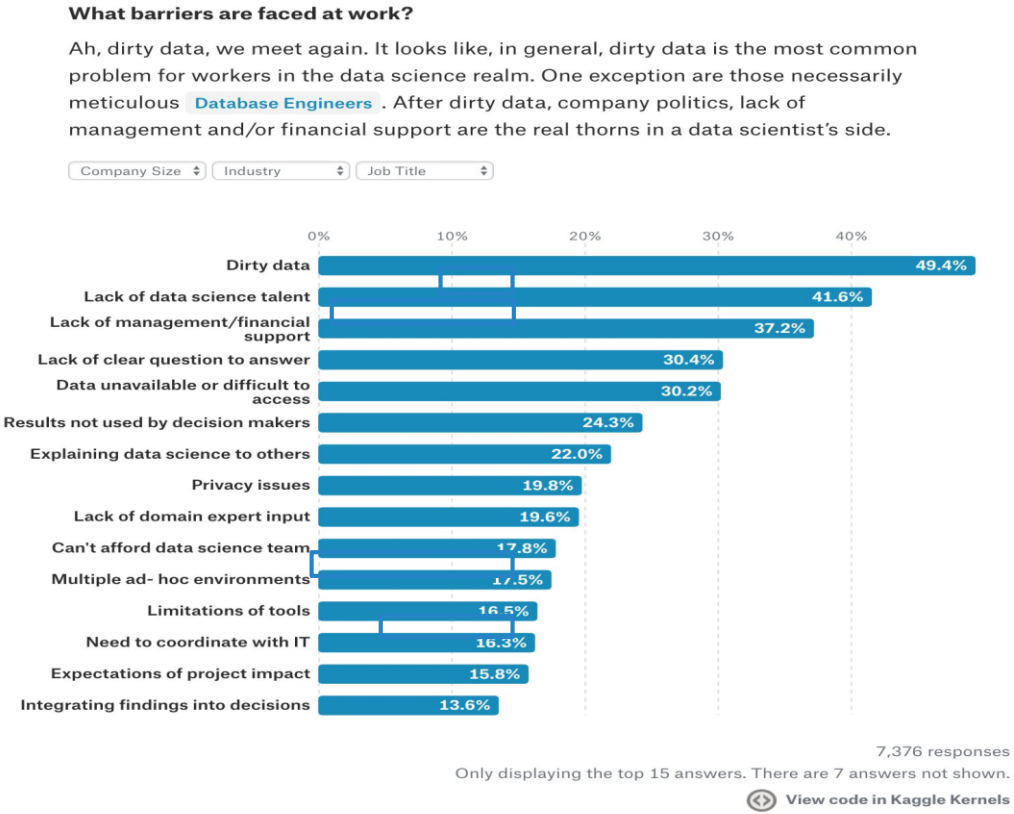


Image source: kaggle.com/surveys/2017

Early AutoML focused on only parameter tuning

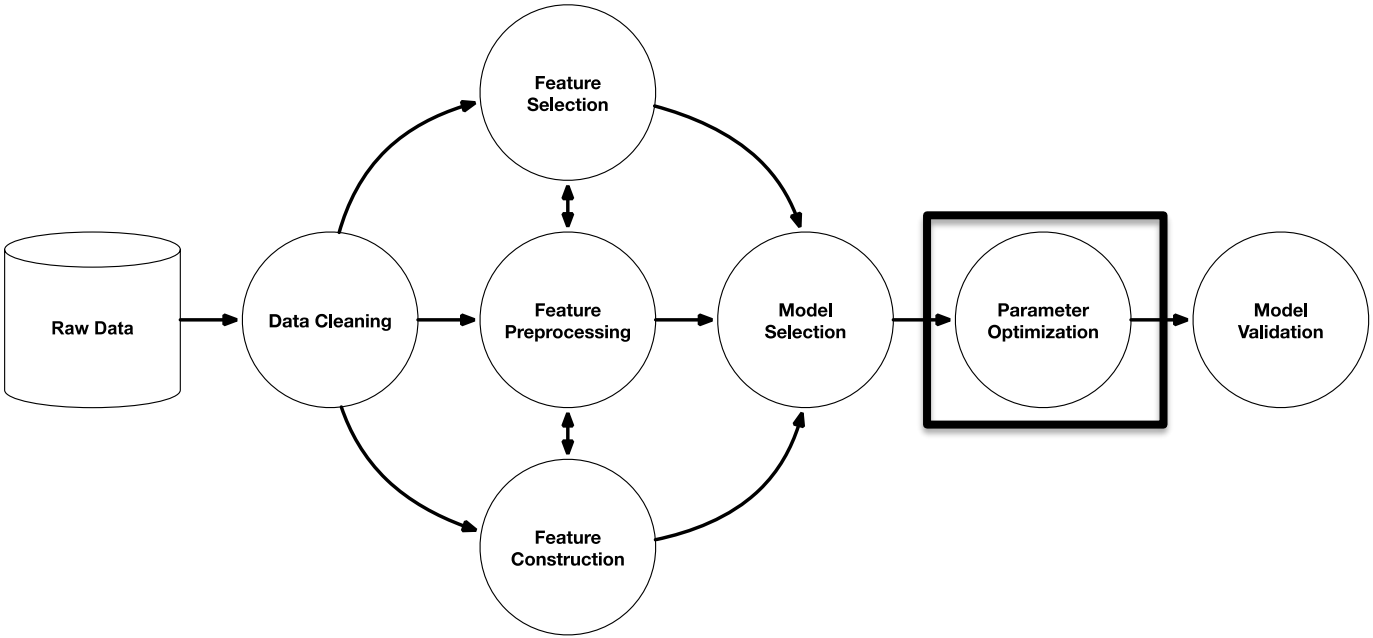
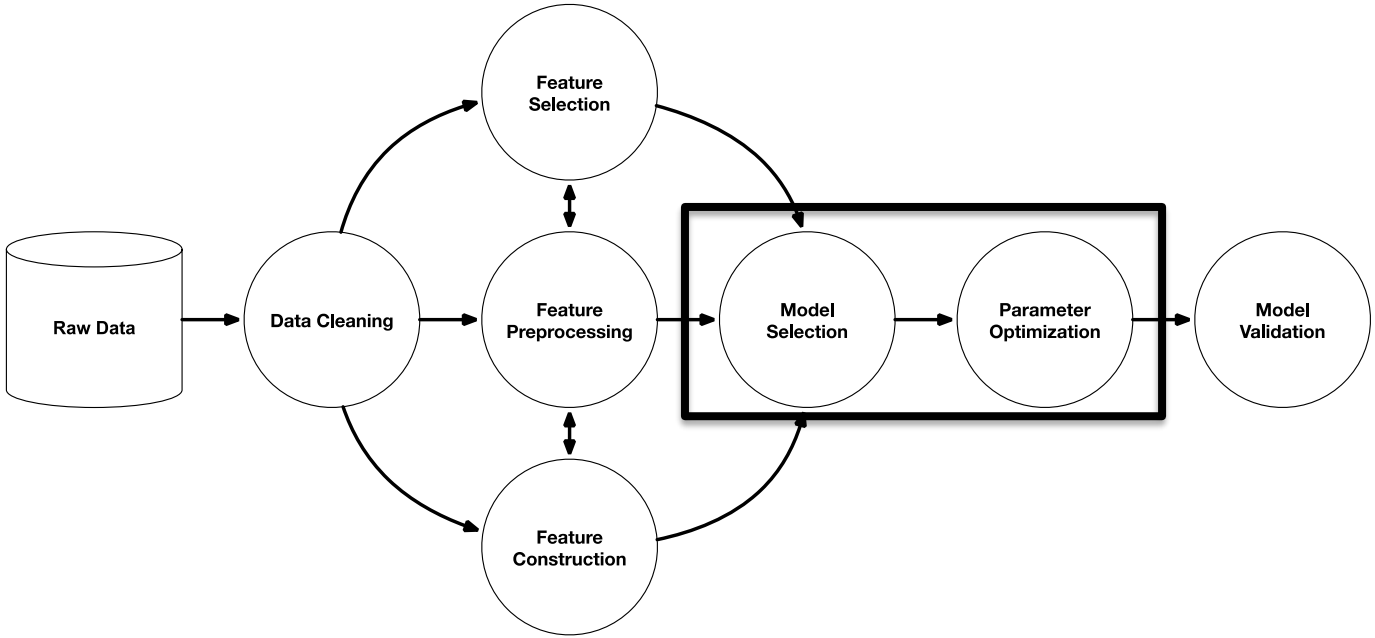


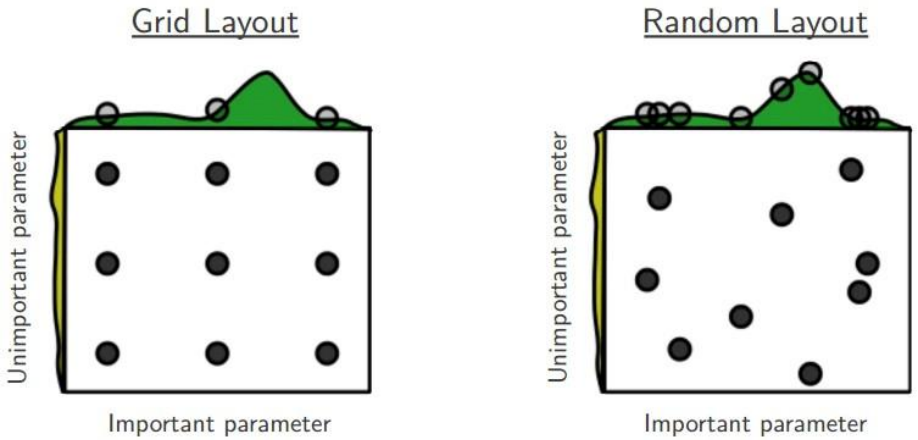
Image source: R. Olson e*t. al.* (2016) “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.”

Early AutoML focused on only parameter tuning



... and maybe (limited) model selection

We mostly used grid search and random search



Nowadays, we wouldn’t really call this AutoML

Modern AutoML optimizes the entire ML workflow

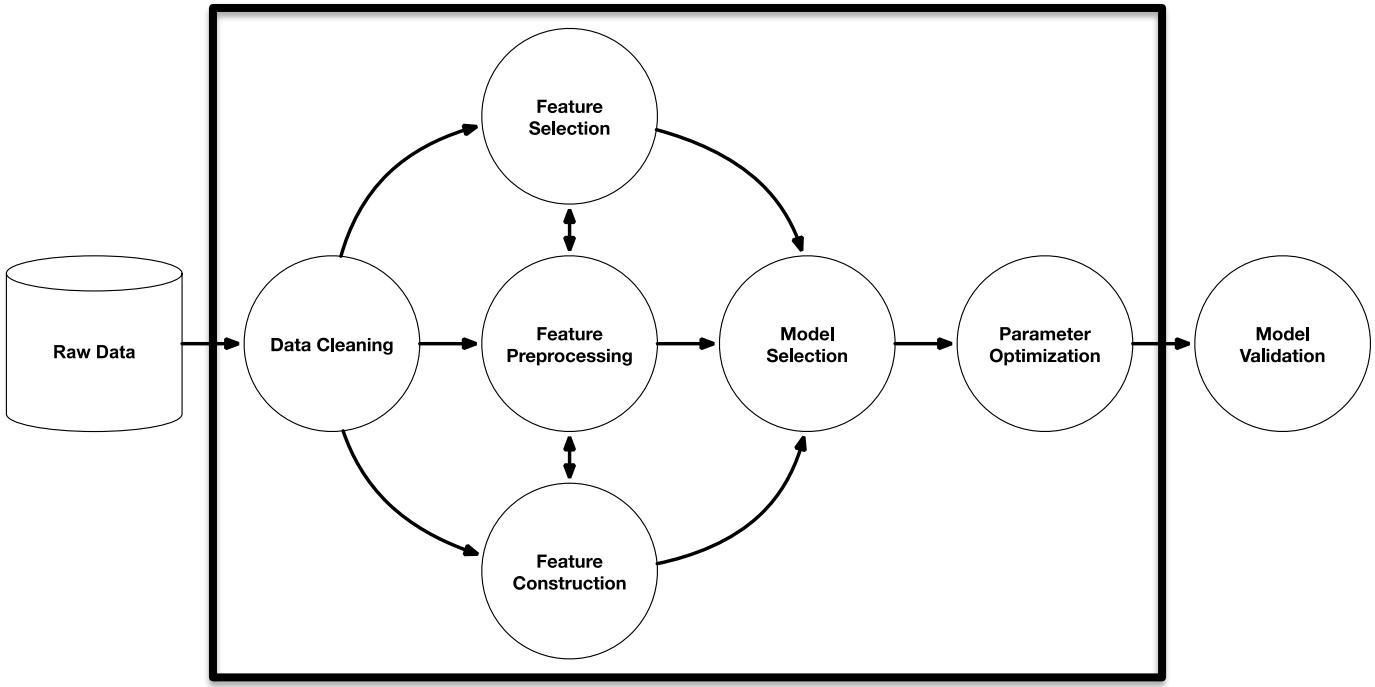
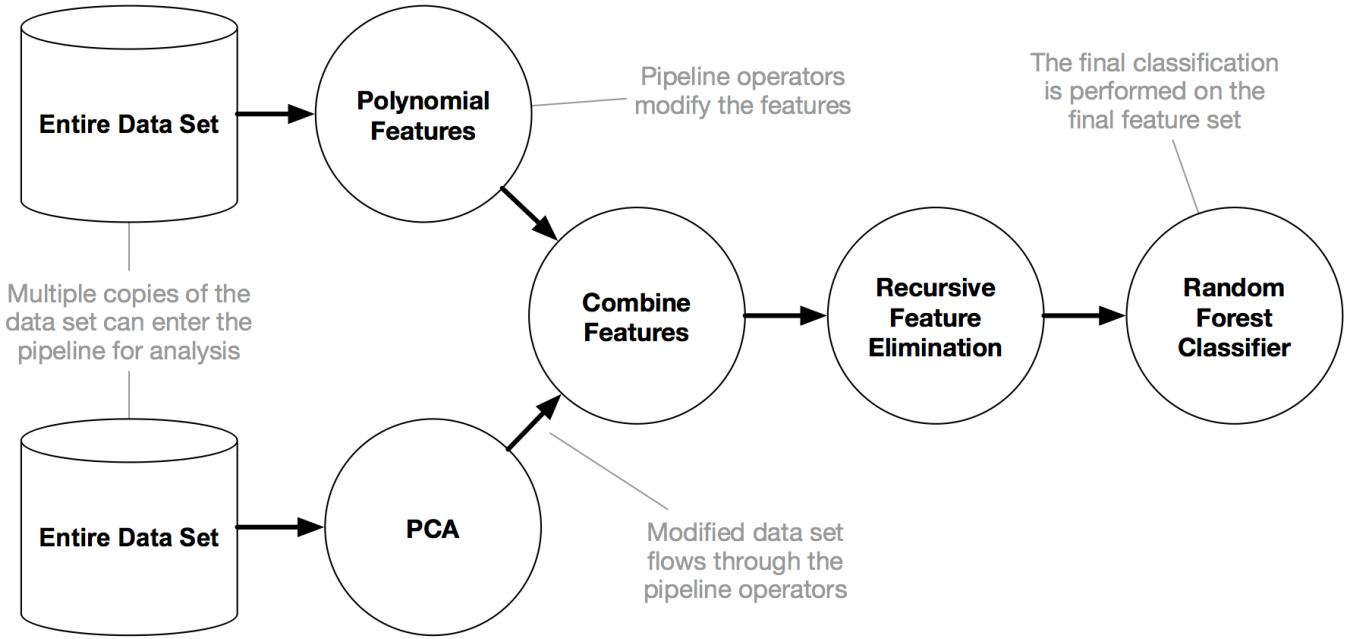


Image source: R. Olson et. *al.* (2016) “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.”

Modern AutoML optimizes the entire ML workflow



Open source AutoML tools

- **auto-sklearn [Python]**
  - Bayesian optimization over a fixed 3-step ML pipeline
  - github.com/automl/auto-sklearn
- **auto-Weka [Java]**
  - Similar to auto-sklearn, but built on top of Weka
  - github.com/automl/autoweka
- **TPOT [Python]**
  - Genetic Programming over a configurable ML pipeline
  - github.com/rhiever/tpot
- **H2O.ai AutoML [Java w/ Python, Scala, & R APIs and web GUI]**
  - Basic data prep w/ mix of grid and random search over ML algorithms
  - github.com/h2oai/h2o-3
- **devol [Python]**
  - Deep Learning architecture search via Genetic Programming
  - github.com/joeddav/devol

© Microsoft Azure + AI Conference All rights reserved.

AutoMLaaS: Commercial AutoML tools

- **DataRobot**
  - Web-based interface
  - Fixed search over thousands of ML pipelines
- **H2O.ai Driverless AI**
  - Web-based interface
  - H2O.ai AutoML + better feature construction
- **Google AutoML**
  - Integrated in the Google Cloud Compute platform
  - DNN architecture search
- **SAS Factory Miner**
  - Fixed search over a handful of ML methods
- **IBM SPSS Modeler**
  - Basic automated data preparation and ML modeling

© Microsoft Azure + AI Conference All rights reserved.

AutoML in the near future

- **AutoML will also handle most of the data cleaning process**
  - Unstructured data → tabular data ready for analysis
  - Capture & automate human approaches to data cleaning
- **AutoML will vastly improve Deep Learning**
  - Automated DNN architecture design
  - Automated preprocessing of data prior to modeling
- **AutoML will scale to large datasets**
  - AutoML is very slow right now on “Big Data”
  - Spark, dask, TensorFlow, etc. will help bring AutoML to scale
- **AutoML will become human-competitive**
  - Already human-competitive on several Kaggle challenges
  - Already human-competitive in DNN architecture design (Google AutoML)

© Microsoft Azure + AI Conference All rights reserved.

AutoML in the future

- **AutoML will transform the practice of data science as we know it**
  - "Data Science Assistant" → Junior Data Scientist level
  - Less focus on choosing the right ML workflow
  - More focus on posing the right questions, collecting & curating the right data, and "thinking like a data scientist"

- **AutoML will become productized**
  - Not AutoMLaaS!
  - "Siri, set an alarm for 6am" → "Siri, set an alarm for the best time for me to wake up"
  - "Siri, [given my personal medical history] should I worry about this rash on my face?"

- **AutoML is only a small part of a greater meta-learning movement**
  - Computer programming is focused on automating rote tasks
  - Machine learning is focused on automating the automation of rote tasks
  - Meta-learning is focused on *automating the automation of automation*
    - i.e., enabling the machine to learn *how* to learn in the best way possible

© Microsoft Azure + AI Conference All rights reserved.

Questions

Microsoft Azure
+ AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection

© Microsoft Azure + AI Conference All rights reserved.

Please use EventsXD to fill out a session evaluation.

Thank you!

Microsoft Azure
+ AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection

© Microsoft Azure + AI Conference All rights reserved.