ML on structured data with GCP

Jérémie Gomez
Data consultant, Professional Services

Google Cloud

Topics

1. Machine learning landscape
2. AutoML Tables
3. BigQuery ML

Machine learning landscape

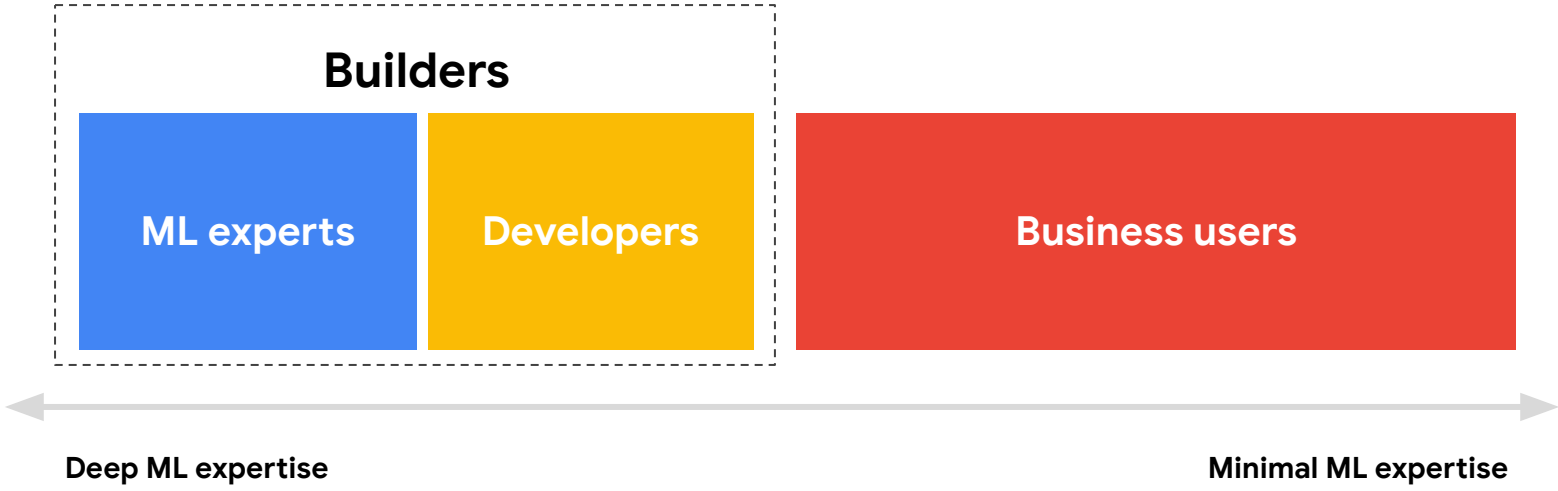
01

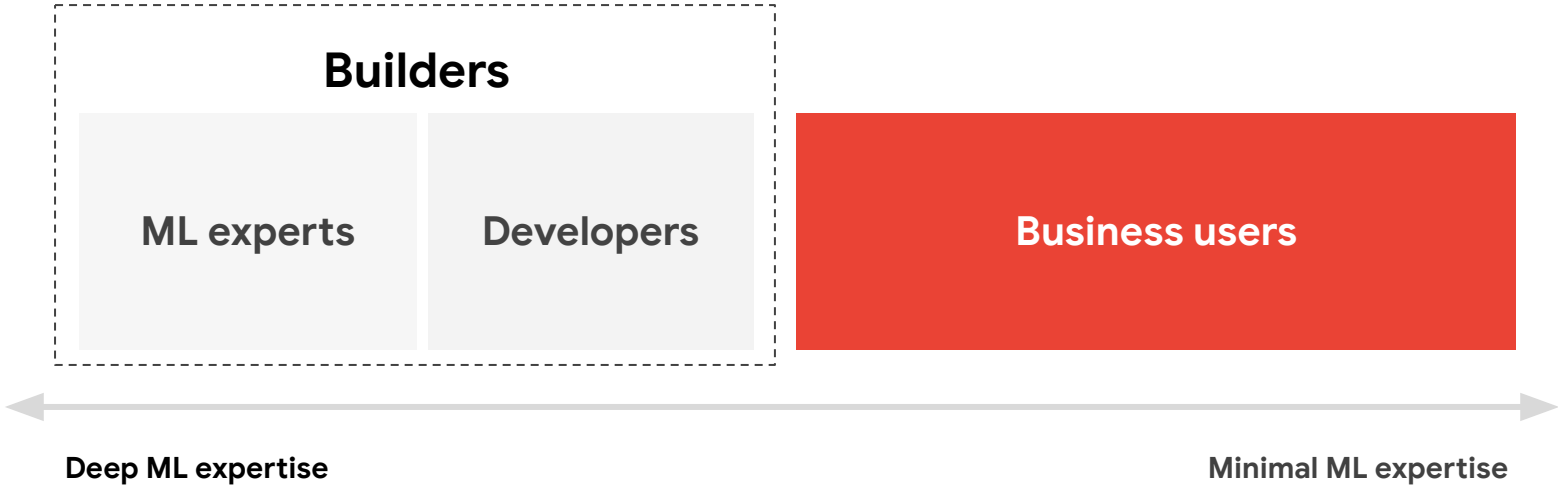
Google Cloud

Who can actually use AI today?

Very few people can create custom ML models today

Google Cloud


Builders
ML experts
Developers
Business users
Deep ML expertise
Minimal ML expertise


Builders
ML experts
Developers
Business users
Deep ML expertise
Minimal ML expertise

Ready-to-deploy AI solutions plug into your existing technology & workflows. No AI or ML expertise required.

Document Understanding AI

Easily, and cost-effectively, extract valuable insights from your documents.

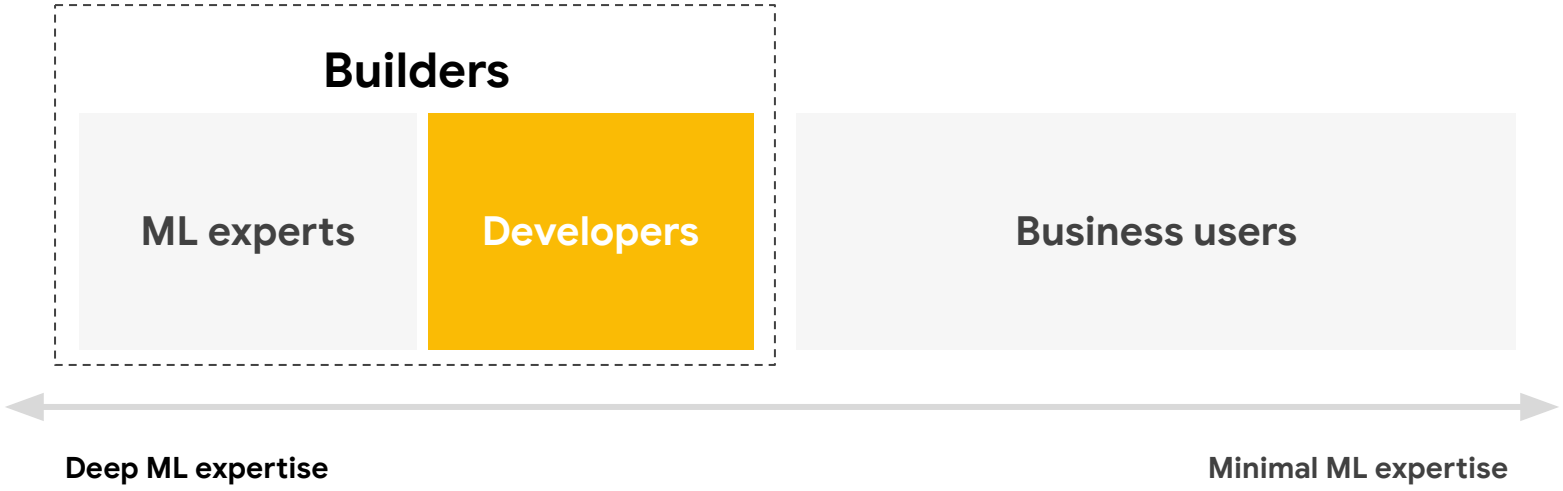
Contact Center AI

Efficiently provide world-class customer support.

Retail

Drive sales with hyper-personal recommendations and visual product search.

Google Cloud


Builders
ML experts
Developers
Business users
Deep ML expertise
Minimal ML expertise

Making AI easier for developers

Sight

- Vision
- Video Intelligence
- AutoML Vision
- AutoML Video Intelligence

Language

- Translation
- Natural Language
- AutoML Translation
- AutoML Natural Language

Conversation

- Dialogflow Enterprise Edition
- Text-to-Speech
- Speech-to-Text

Structured Data

- **AutoML Tables**
- **BigQuery ML**
- Recommendation AI


Builders
ML experts
Developers
Business users
Deep ML expertise
Minimal ML expertise

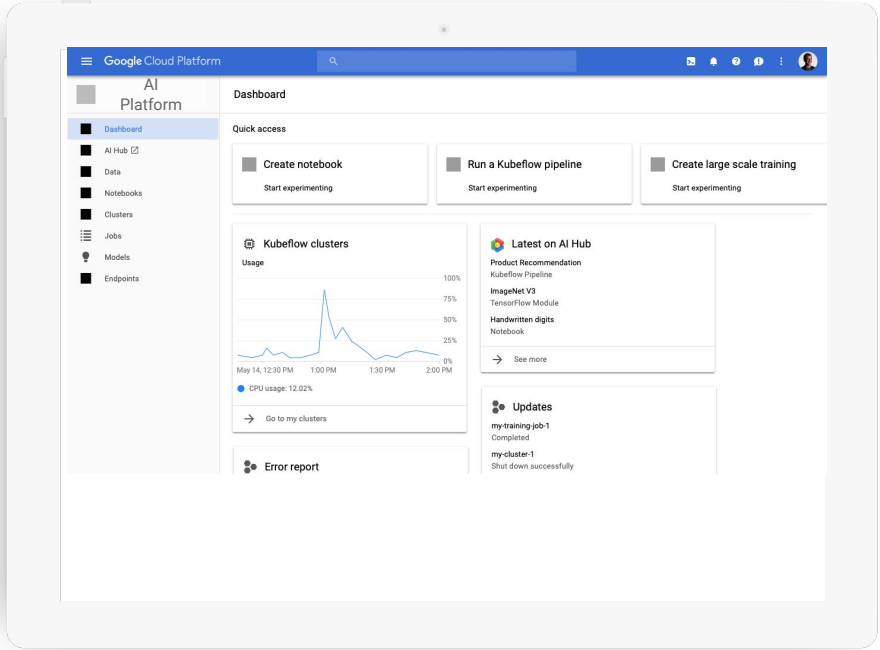
What is AI Platform?

End-to-end, code-based development environment for AI inside GCP console

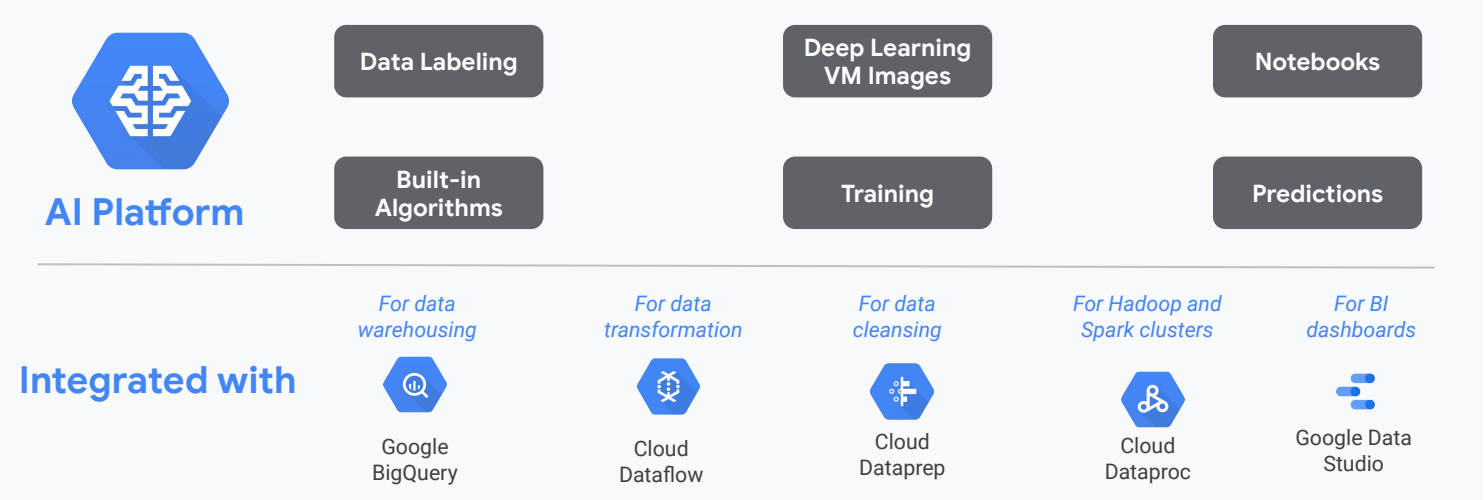
Built on Kubeflow, Google's open-source project, offers an integrated tool chain from data engineering to model deployment with "no lock-in"

Allows you to run on-premises or on Google Cloud without significant code changes.

Access to cutting-edge Google AI technology like TensorFlow, TPUs, and TFX tools as you deploy your AI applications to production.



What is included?



AI Hub

Kubeflow
(On premises)

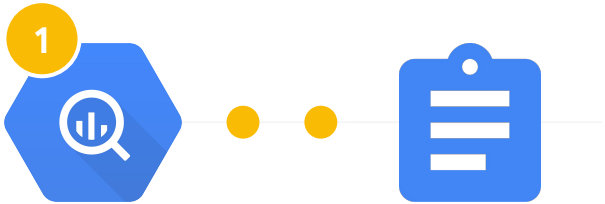
Google Cloud

AutoML Tables

02

Google Cloud

Months to create and deploy an ML model



Export data

2

Regression in Excel/Sheets

Export small amounts of data from BQ

Run linear regression

Get a model with **low accuracy** due to small data for training

Go back and get more data to create new features, and improve performance

Repeat. It's hard, so you **stop after a few iterations**

3

TensorFlow or scikit-learn

Only an expert data scientist can do this

Export small amounts of data from BQ

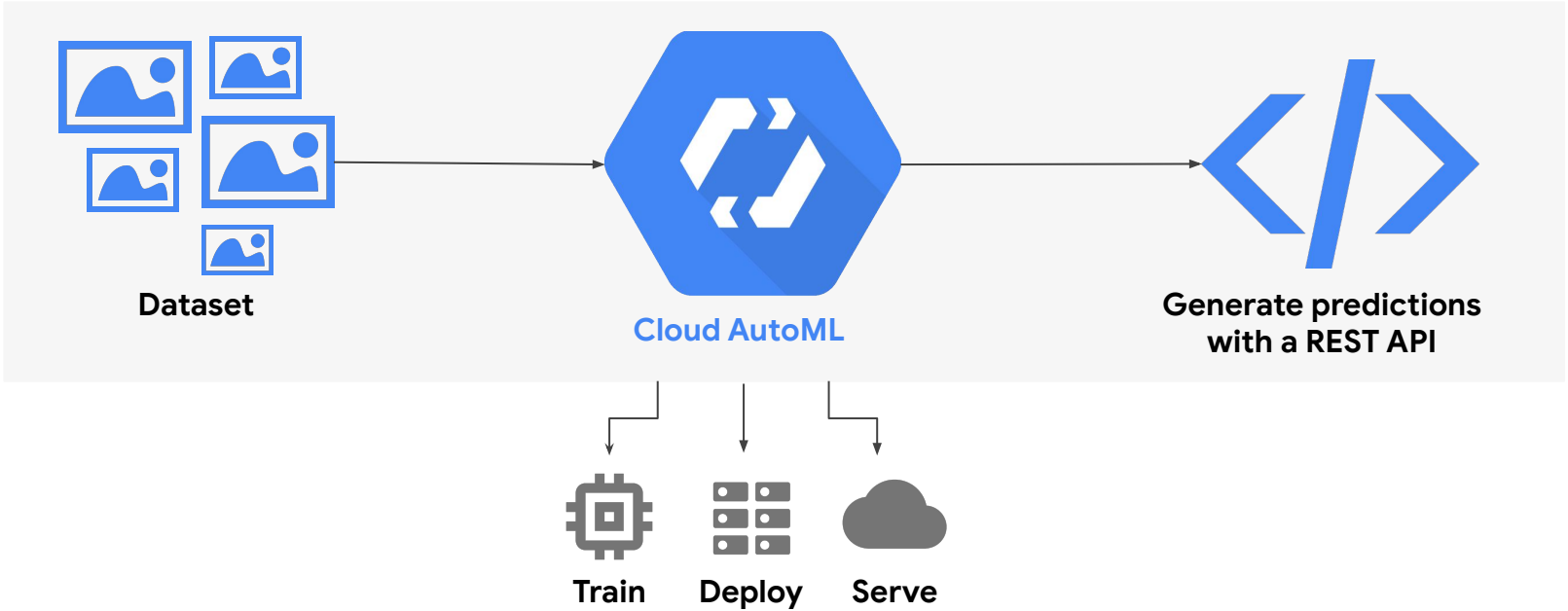
Create frames of data for use with TensorFlow Build model

Go back and get more data to create features, and improve performance

Repeat. It's hard, so you stop after a few iterations

Cloud AutoML to the rescue

ML that creates ML for your problem



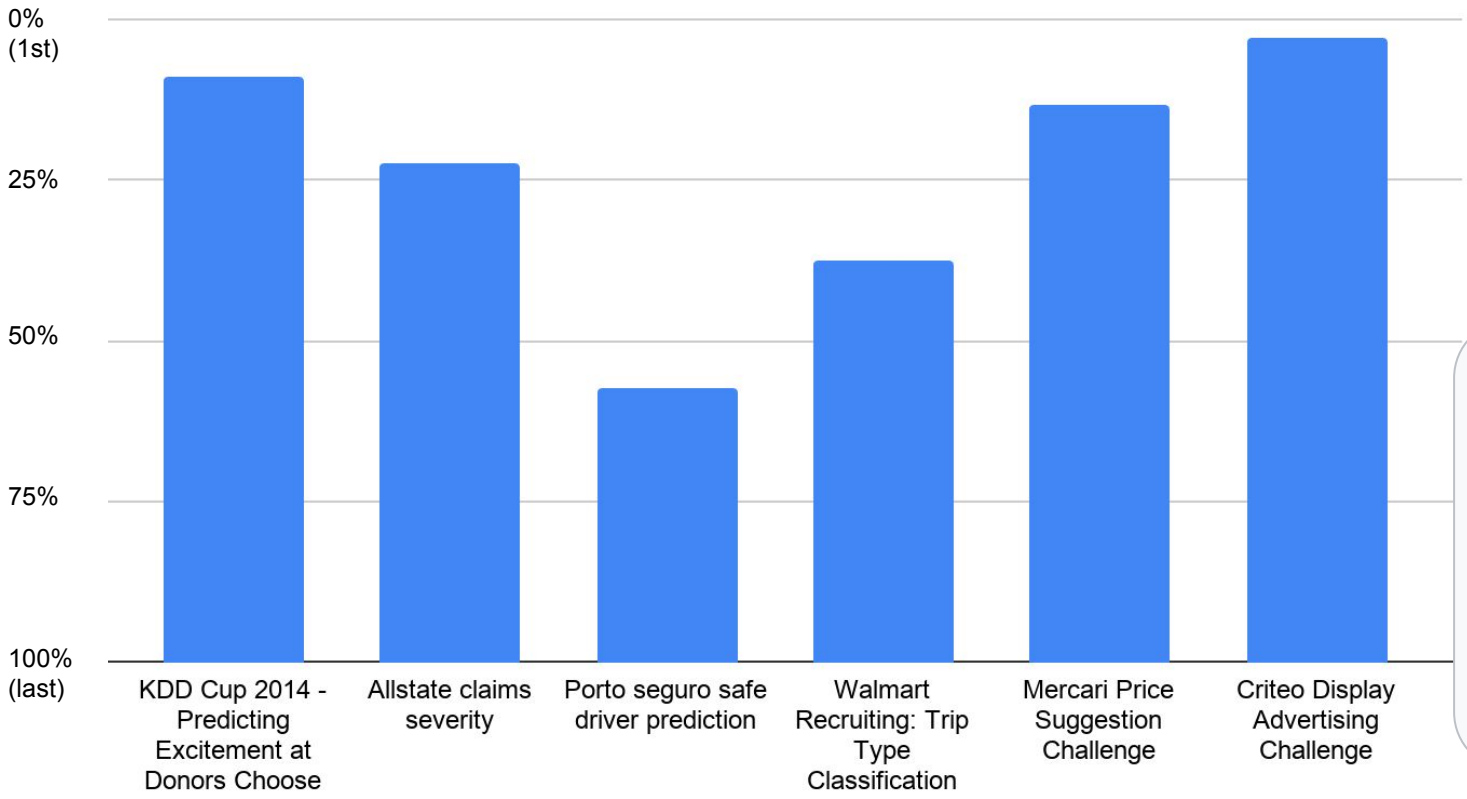
Cloud AutoML to the rescue

ML that creates ML for your problem

- Higher model accuracy and faster time to market
- Limited ML expertise needed to build custom models
- Data-first approach with simple graphical user interface

Leading to increased model quality

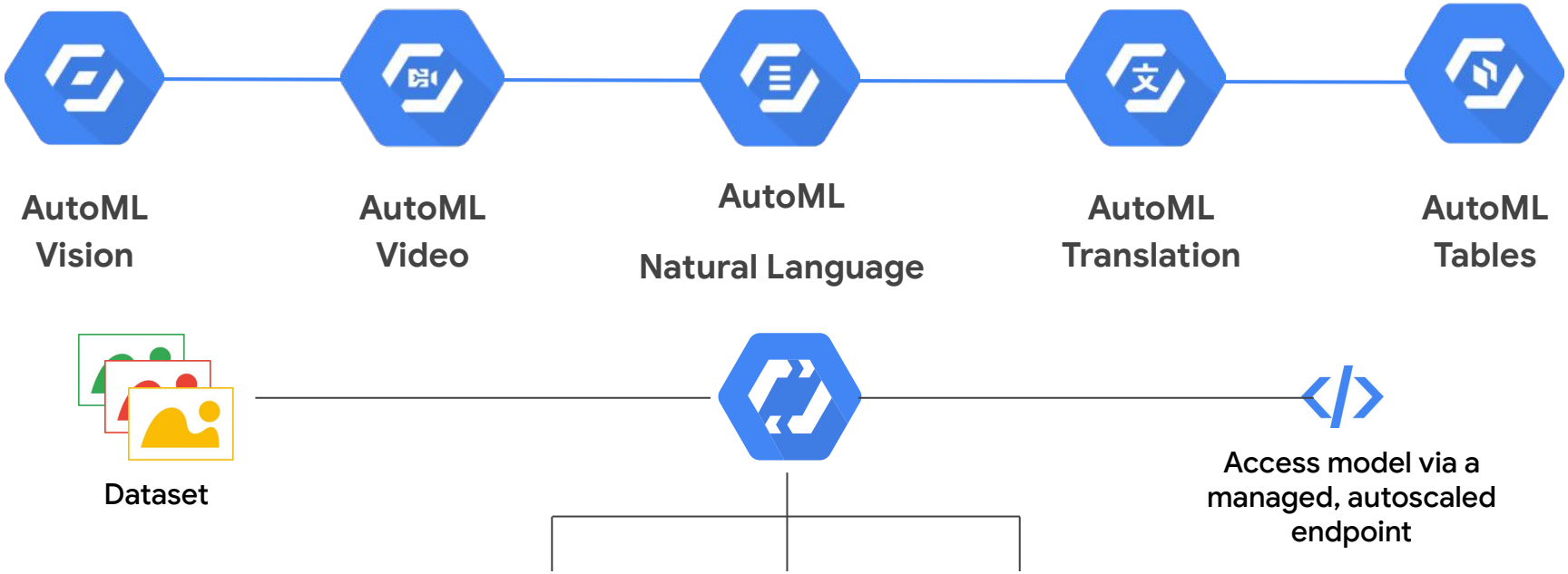
% ranking on Kaggle private leaderboard



For each product:

- Relevant tables joined by given IDs
- Some minimal preprocessing done to match input requirements
- Run until converge
- Benchmarks run between H2 2018 to today (as they became available)

AutoML products



ML on structured data: 101

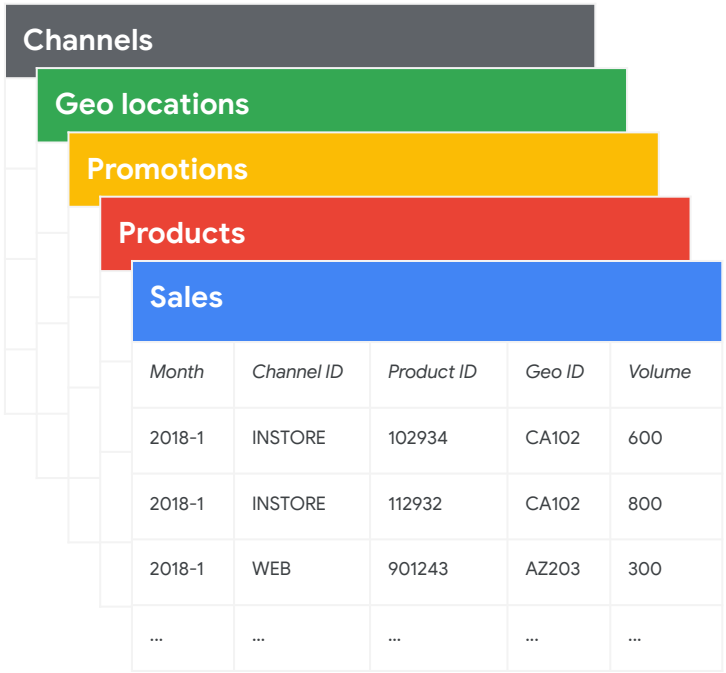
Historic offers from marketplace.xyz									
ID	**Geo**	**Domain**	**Posted on:**	**Title**	**Description**	**Category**	**Brand**	**...**	**Price sold:**
104	US	marketA	Feb 1, 2018	“Dark red...”	“Try this soft...”	[“A, B, ...”]	Nike	...	$92
204	US	marketB	Jan 20, 2018	“Women’s...”	“Medium-size...”	[“A, B, ...”]	Adidas	...	$58
302	US	marketA	Jan 12, 2018	“Running...”	“All-terrain...”	[“A, B, ...”]	Asics	...	$85
352	EU	marketB	Feb 13, 2018	“Running...”	“All-terrain...”	[“A, B, ...”]	Puma	...	?

Target column

Google Cloud

AutoML Tables

Start with **raw tabular data**



Month	*Channel ID*	*Product ID*	*Geo ID*	*Volume*
2018-1	INSTORE	102934	CA102	600
2018-1	INSTORE	112932	CA102	800
2018-1	WEB	901243	AZ203	300
...	...	...	...	...

- Build state-of-the-art models automatically
- Enriched treatment for a wide range of data primitives (#s, text, etc.)
- Gracefully handle datasets at BigQuery scale (currently up to 10TB)
- Code-less graphical UI for the full ML lifecycle

Demand forecasting

Stock-out prediction

Price optimization

Customer lifetime value

Predict customer conversion / churn

Fraud prevention

and more...

Google Cloud

Handle data as found in the wild

Automated feature engineering for:

- Numbers
- Timestamps
- Classes
- Lists
- Strings
- Nested fields

Resilient to + guardrails for:

- Imbalanced data
- Highly correlated features
- Missing values
- High cardinality features (like IDs)
- Outliers

Google Cloud

Automatically search through Google's whole model zoo...

Linear, logistic

Feedforward DNN

Wide and Deep NN

Gradient Boosted Decision Tree (GBDT)

DNN + GBDT Hybrid

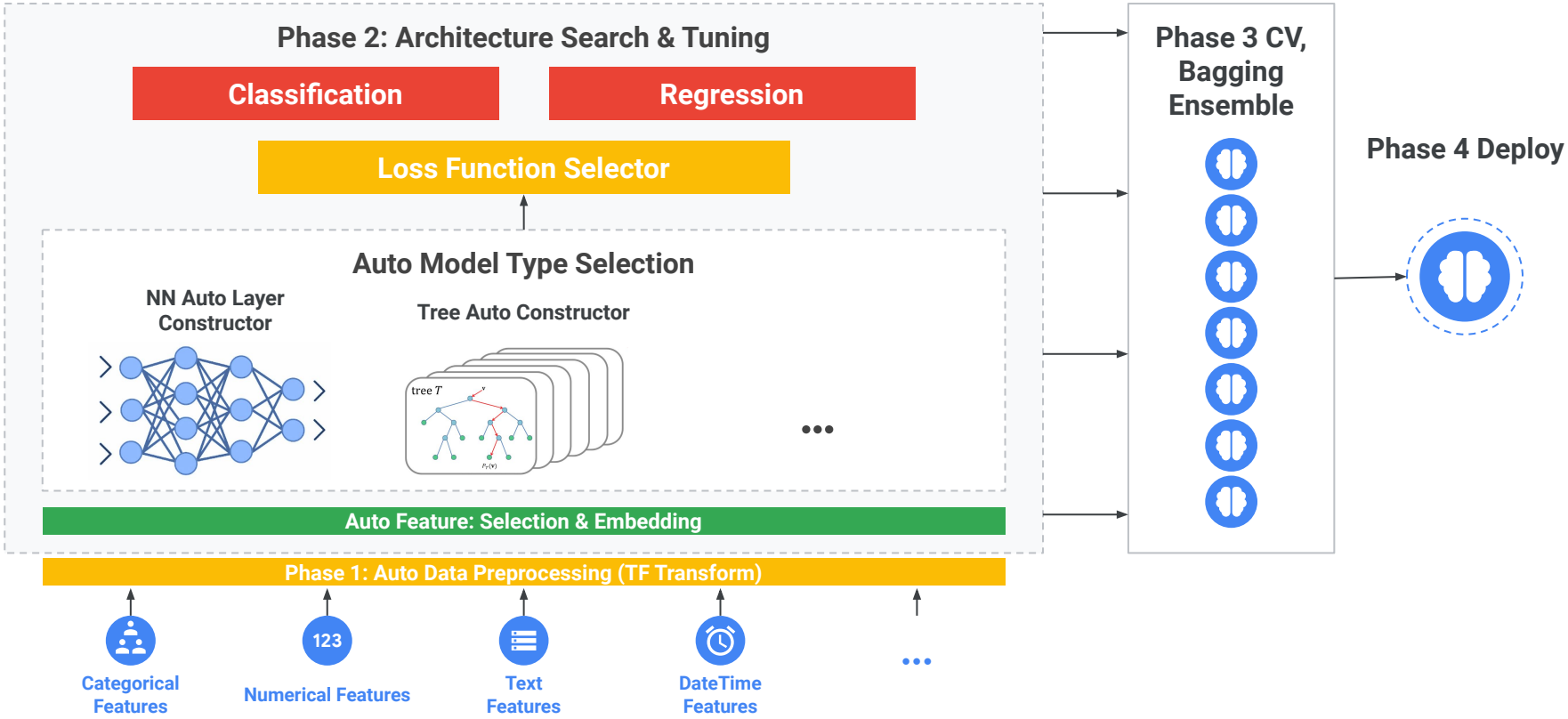
Adanet ensemble

Neural + Tree Architecture Search

...and more!

Google Cloud

...including the latest methods from Google Brain



Demo

Importing data into BigQuery, creating a dataset and a model with AutoML Tables

Google Cloud

BigQuery ML

03

Google Cloud

What is BigQuery?

- Fully Managed, Zero-Ops Data Warehouse
- Petabyte-Scale
- Industry-Standard SQL
- Automatically Encrypted, Durable, and Highly Available
- Virtually Unlimited Resources

Google Cloud

BigQuery ML

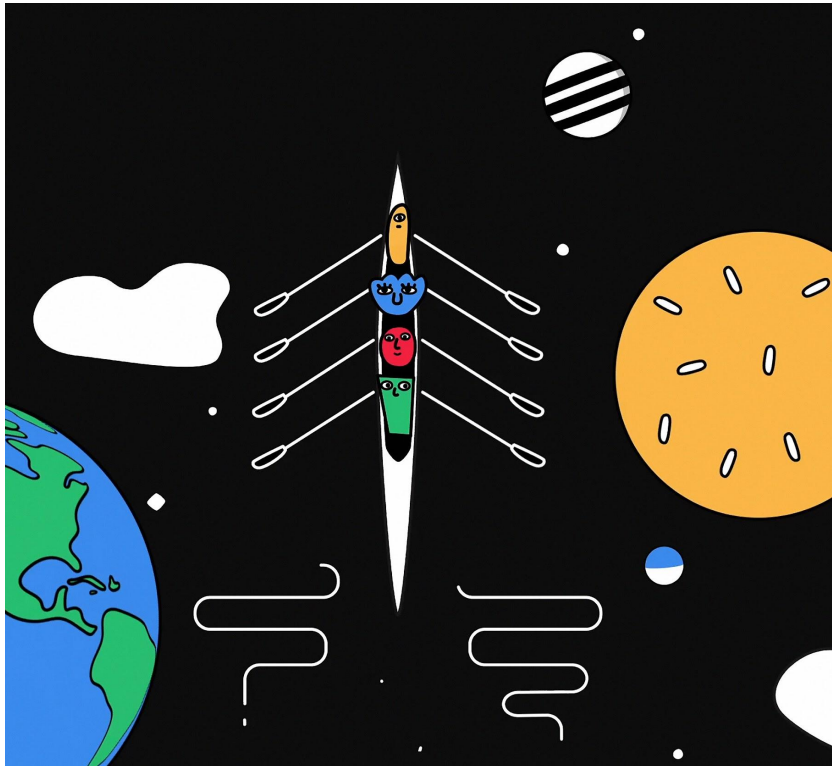
1. **Execute** ML initiatives without moving data from BigQuery
2. **Iterate** on models in SQL in BigQuery to increase development speed
3. **Automate** common ML tasks and hyperparameter tuning

Google Cloud

Behind the scenes

Through two lines of SQL

- Leverage BigQuery's processing power to build a model
- Auto-tuned learning rate
- Auto-split of data into training and test
- Null imputation
- Standardization of numeric features
- One-hot encoding of strings
- Class imbalance handling



Current BQML Model Types (with betas)

Classification

- Logistic regression

Other Models

- k-means clustering
- Matrix factorization

Regression

- Linear regression

Model Import/export

- TensorFlow model import
- Model export

SQL interface to Machine Learning

- Data scientists can:
  - Build 100 models in a week
  - Not spend time on data ETL and setup of multiple tools
  - Experiment in BigQuery, export models to CMLE for further tuning
- Data analysts can:
  - Build ML models with knowledge of basic ML concepts.
  - Easily utilize domain and data knowledge for ML models

Prediction

- BigQuery for batch prediction on structured data
- Export to Cloud AI Platform for online prediction
- Interactive prediction on BI dashboards through BigQuery

AutoML Tables vs. BigQuery ML

These are complementary, not competing products

AutoML Tables	BigQuery ML
For problems that require best-in-class accuracy that is fully automated	For problems that require fast experimentation and development time, and explainability (e.g., simpler models like logistic regression, trees)
Discovers the best model for the problem	Supports a variety of models
Code-less graphical UI	SQL interface
Consistent experience for users that used other AutoMLs	Will support AutoML Tables as a model_type in the future

Demo

BigQuery ML training, evaluation and prediction

Google Cloud

Model type	BigQuery ML	AutoML	Custom model
How	SQL in BigQuery for ML on structured data	AutoML uses neural architecture search and best-of-class model architectures for the specific problem	Machine learning libraries (scikit-learn, Tensorflow...), trained on Cloud ML Engine
Best if you are a	Data analyst who can wrangle data with SQL	Developer who can create the dataset in the required format	ML Engineer who knows Python and knows deep learning, NLP techniques
How long it takes an experienced practitioner	About an hour	About a day	A week to a month
Most of this time is spent in	Writing SQL	Waiting for job to finish	Coding Python and experimentation with ML
Cloud computing costs	Low	Medium	Medium to high depending on size of data, number of experiments, etc.
Accuracy	Moderate to high, mostly depending on the size of your dataset	High	Low if you don't know what you are doing; extremely high if you employ appropriate architectures and have a large-enough dataset

Thank you!