ENTERPRISE ML SOLUTIONS: STRATEGIC KEY CONSIDERATIONS

Thomas G. Dietterich, Chief Scientist, BigML and Professor (Emeritus), Oregon State University
Guillem Vidal, Machine Learning Engineer, BigML

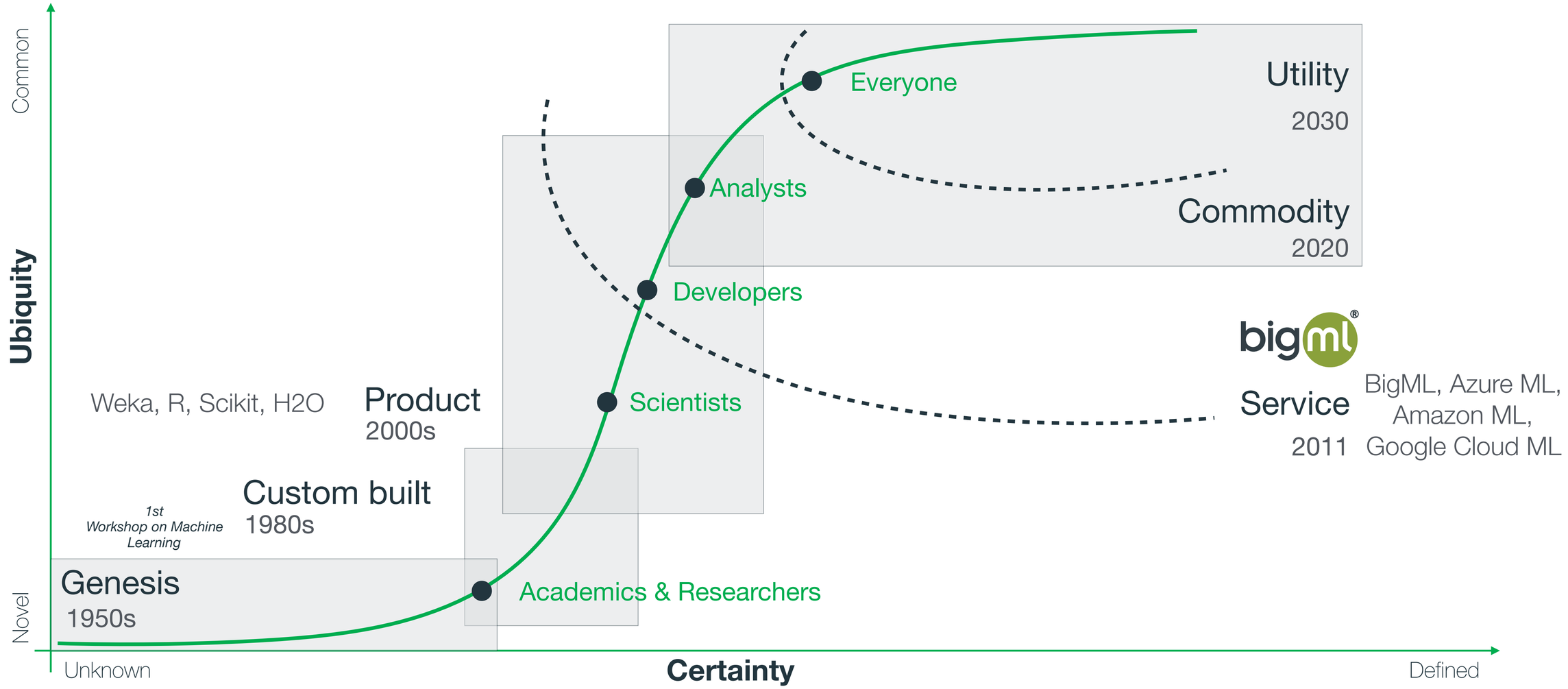
Agenda

1	Machine Learning Business Trends
2	What is Machine Learning?
3	What are the Engineering Challenges in Machine Learning?
4	Example Application Stories from BigML
5	Strategic Considerations

Agenda

1	Machine Learning Business Trends
2	What is Machine Learning?
3	What are the Engineering Challenges in Machine Learning?
4	Example Application Stories from BigML
5	Strategic Considerations

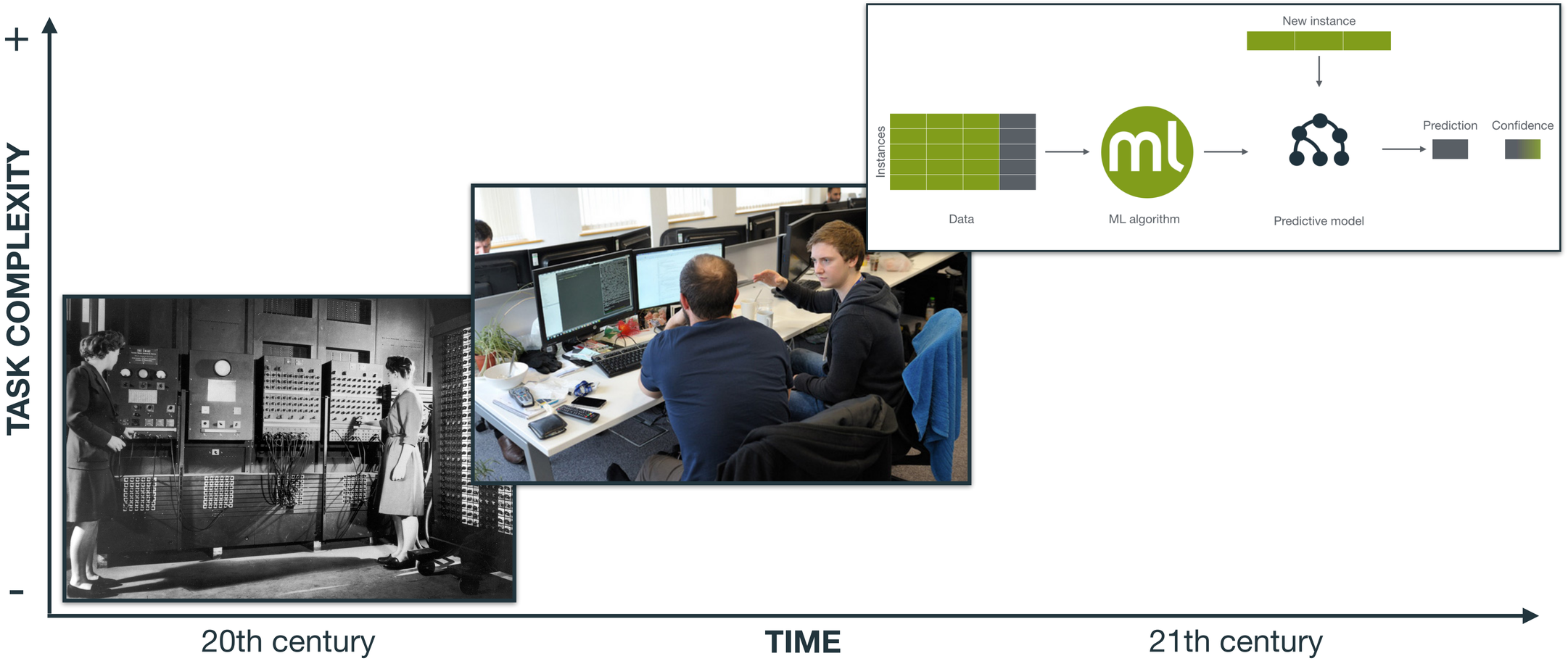
Machine Learning Evolution



Agenda

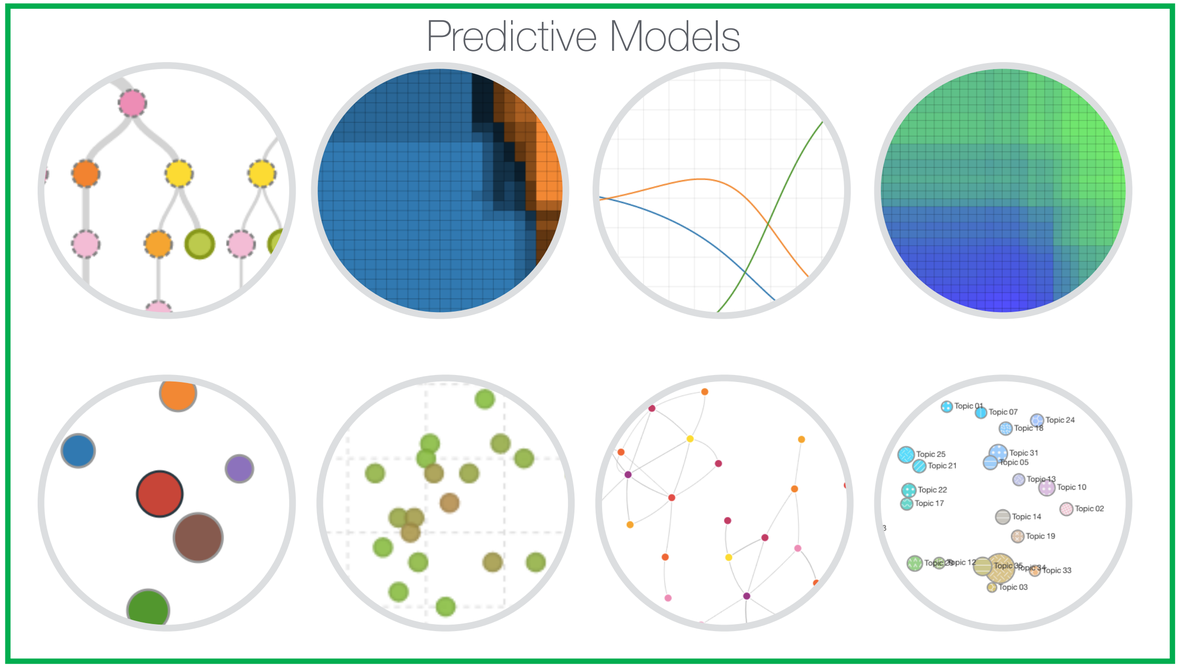
1	Machine Learning Business Trends
2	**What is Machine Learning?**
3	What are the Engineering Challenges in Machine Learning?
4	Example Application Stories from BigML
5	Strategic Considerations

A New Programming Paradigm



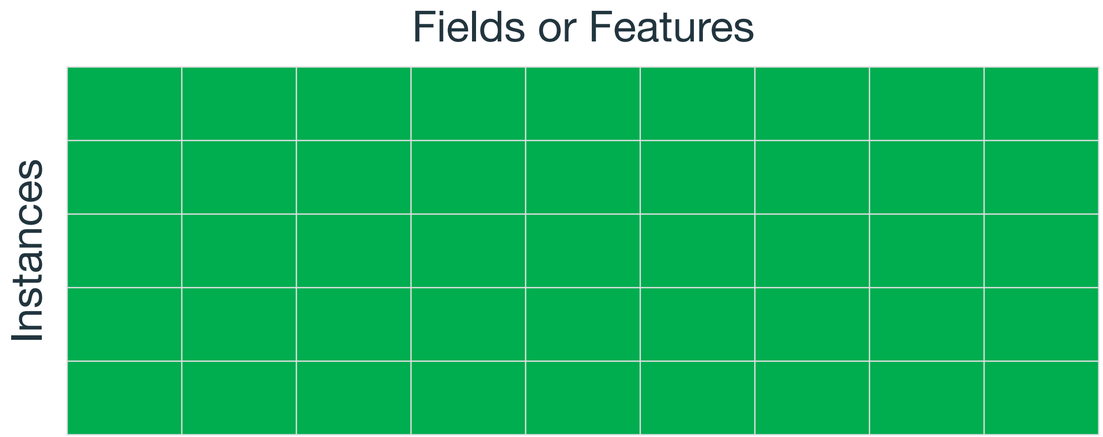
What's Machine Learning?

AIRLINE	ORIGIN	DESTINATION	DEPARTURE DELAY	DISTANCE	ARRIVAL DELAY
AS	ANC	SEA	-11	1448,0	-22
AA	LAX	PBI	-8	2330,0	-9
US	SFO	CLT	-2	2296,0	5
AA	LAX	MIA	-5	2342,0	-9
AS	SEA	ANC	-1	1448,0	-21
DL	SFO	MSP	-5	1589	8
NK	LAS	MSP	-6	1299	-17
US	LAX	CLT	14	2125,0	-10
AA	SFO	DFW	-11	1464,0	-13
DL	LAS	ATL	3	1747,0	-15



Finding **patterns** in data that can be used to make inferences

Machine Learning-Ready Data



Field types: Numeric, categorical, text, date-time, items

Machine Learning-Ready Data

A DATASET ABOUT FLIGHT DELAYS

AIRLINE	FLIGHT NUMBER	ORIGIN	DESTINATION	DEPARTURE DELAY	DISTANCE	ARRIVAL DELAY
AS	98	ANC	SEA	11	1448,0	22
AA	2336	LAX	PBI	8	2330,0	9
US	840	SFO	CLT	-2	2296,0	5
AA	258	LAX	MIA	-1	2342,0	0
AS	135	SEA	ANC	45	1448,0	21
DL	806	SFO	MSP	-5	1589	8
NK	612	LAS	MSP	32	1299	17
US	2013	LAX	CLT	14	2125,0	10
AA	1112	SFO	DFW	3	1464,0	0
DL	1173	LAS	ATL	-3	1747,0	0

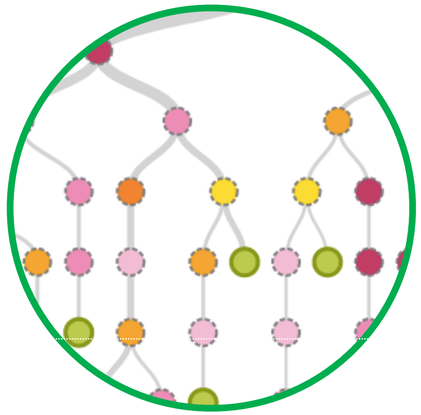
Machine Learning Tasks

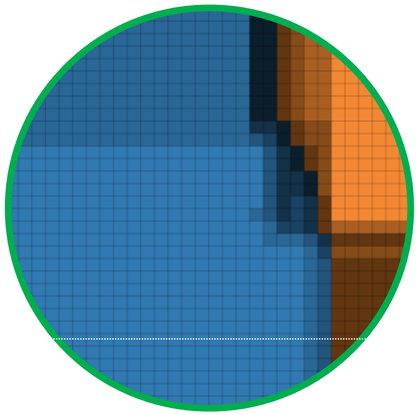
SUPERVISED

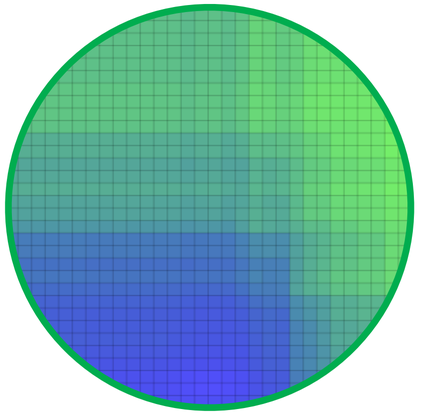
UNSUPERVISED

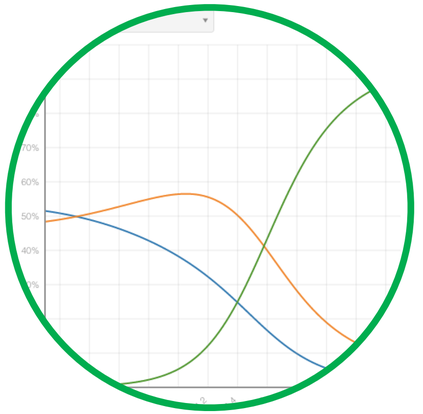
CLASSIFICATION AND REGRESSION

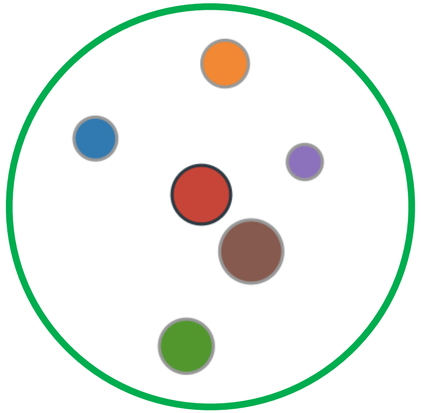
TIME SERIES



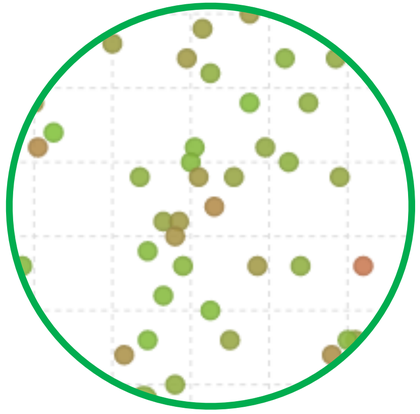




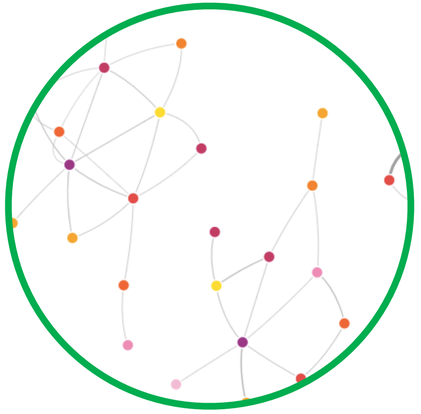




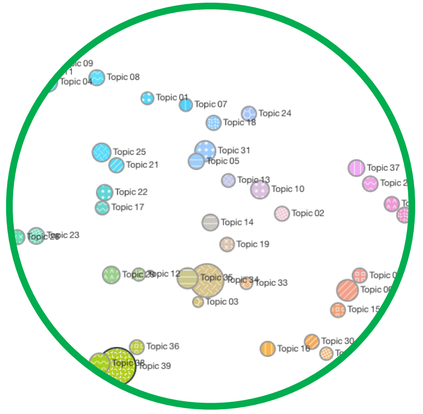
CLUSTER ANALYSIS



ANOMALY DETECTION

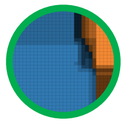
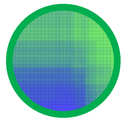


ASSOCIATION DISCOVERY



TOPIC MODELING

Airline Tasks

 CLASSIFICATION	Predict flight arrival delay (On Time / Delay)
 REGRESSION	Predict arrival delay time in minutes
 TIME SERIES FORECASTING	Forecast future delays over the next 3 months
 CLUSTER ANALYSIS	Discover groups of similar flights (By departure delay, origin, etc.)
 ANOMALY DETECTION	Detect unusual flights (Errors in data, diverted flights, etc.)
 ASSOCIATION DISCOVERY	Discover relationships (United + SFO + dep delay > 20 mins)

Agenda

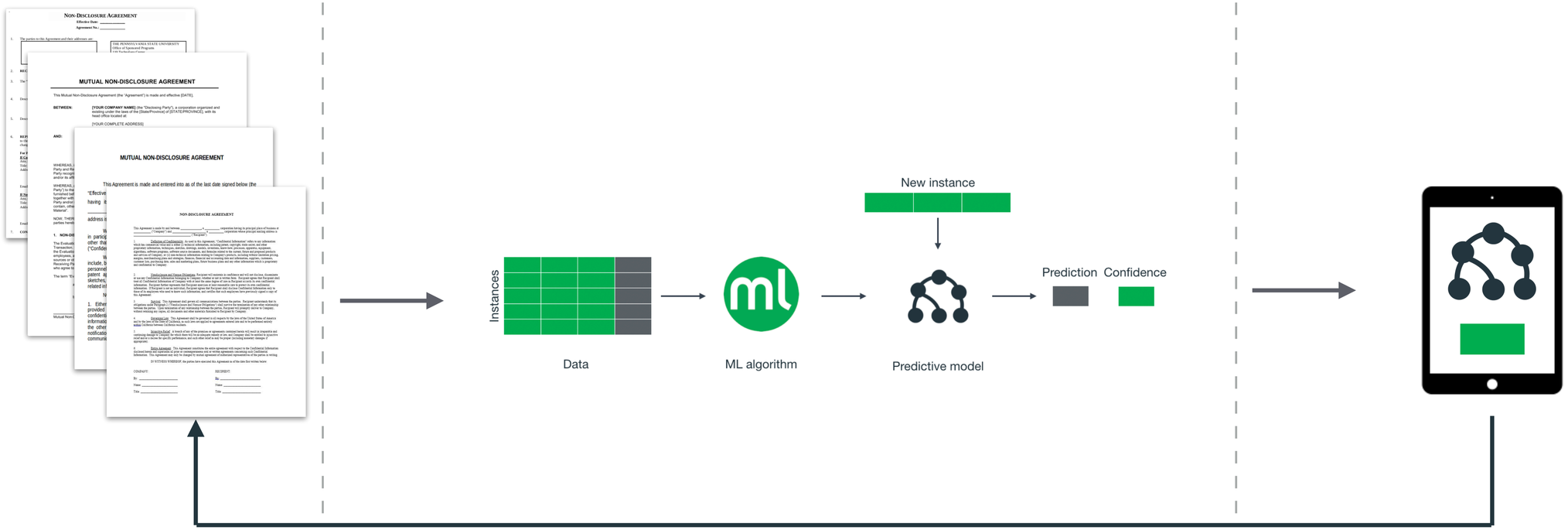
1	Machine Learning Business Trends
2	What is Machine Learning?
3	**What are the Engineering Challenges in Machine Learning?**
4	Example Application Stories from BigML
5	Strategic Considerations

Real World ML is Iterative

PREPARING AND TRANSFORMING DATA

MODELING

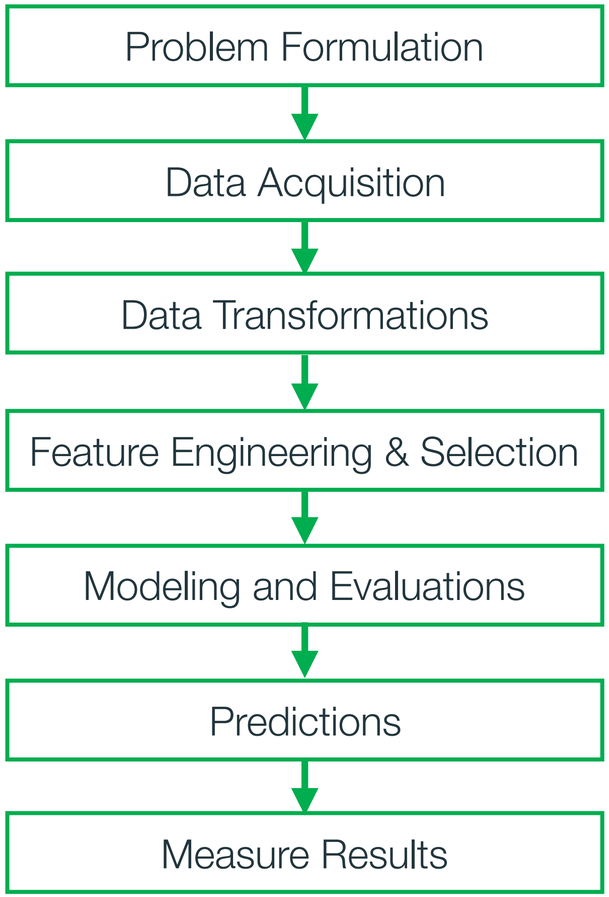
OPERATING



ITERATIVE

Real World ML is Iterative

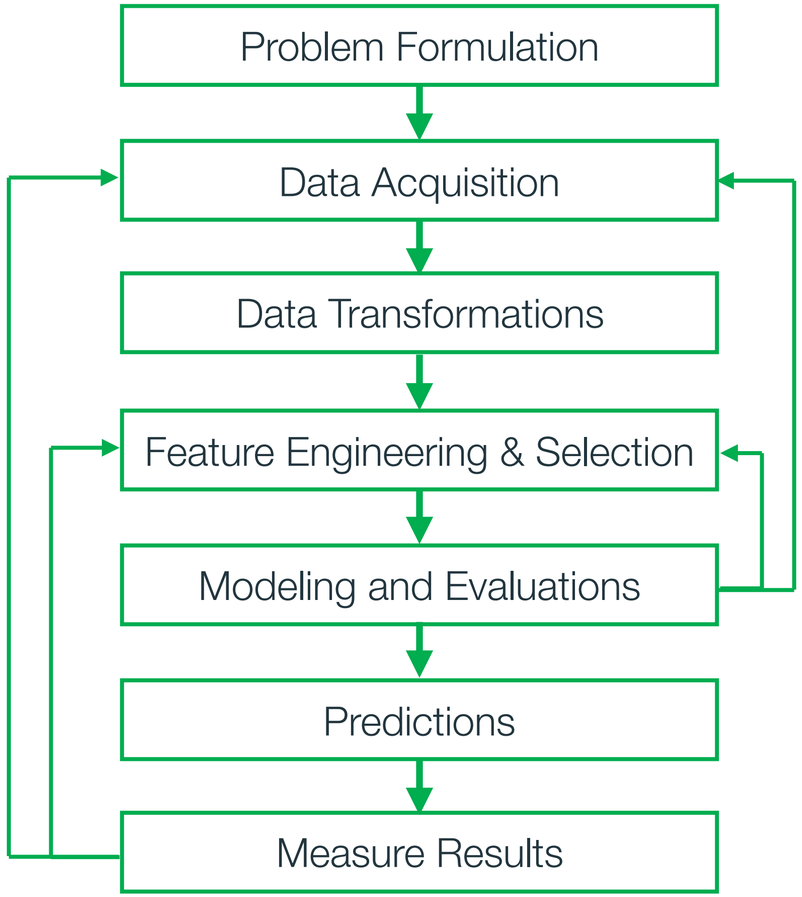
SEQUENTIAL STEPS



- Define a precise goal. What exact questions do we need to answer?
- Get the data: multiple locations and formats are possible
- Clean, join and transform the data into a ML ready format
- Define and calculate appropriate features
- Find the best model
- Make predictions
- Measure the prediction results on new data

Real World ML is Iterative

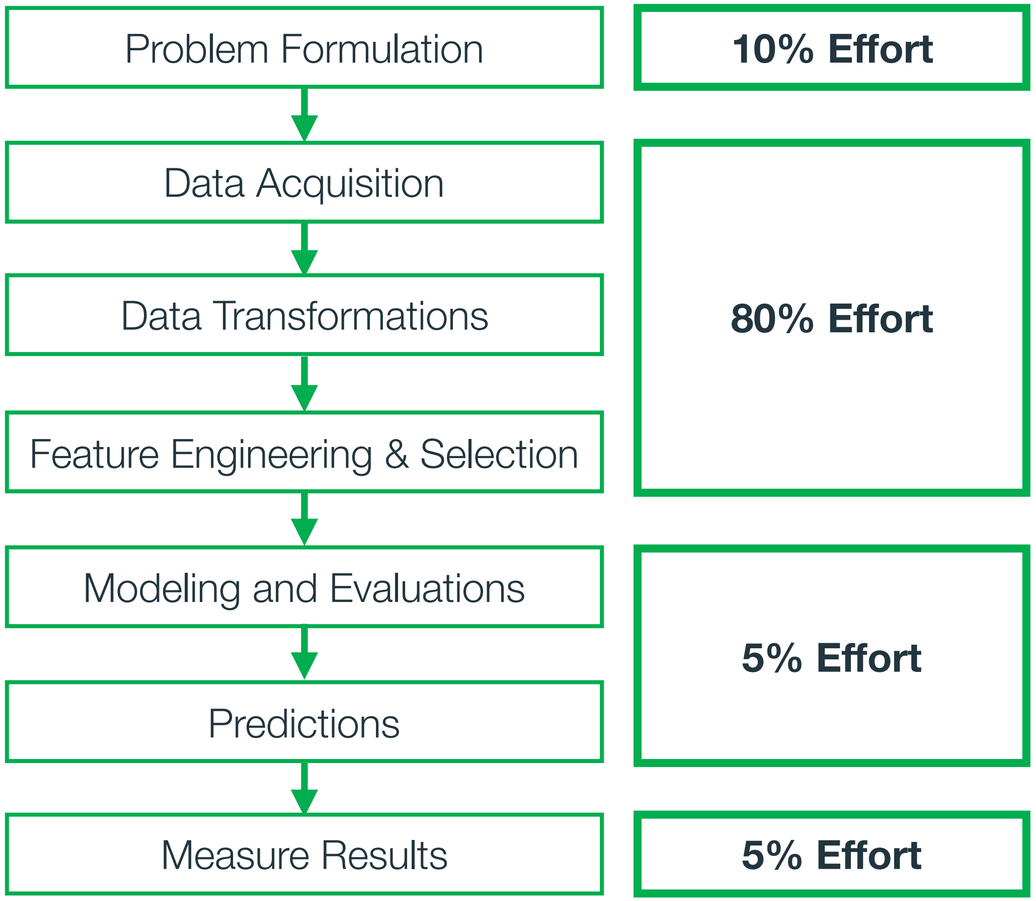
DATA ITERATIONS



- Data quality is the most important factor, especially for the target variable
- If there will be measurement errors in the run-time data, then the training time data should contain similar errors
- The exact choice of ML model is not as important as improving the quality and quantity of data and designing good features
- Model training and parameter tuning can be automated' Feature engineering mostly cannot

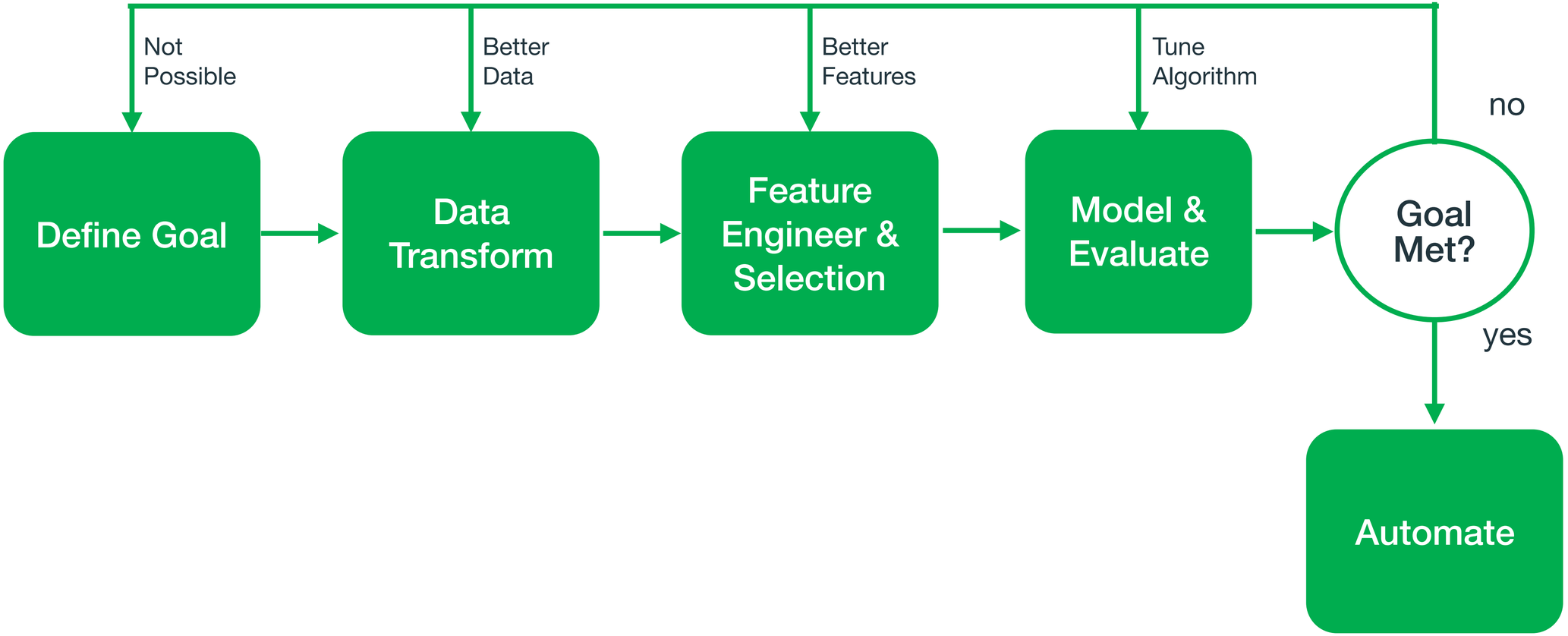
Data Preparation Takes Most of the Time

EFFORT ESTIMATES



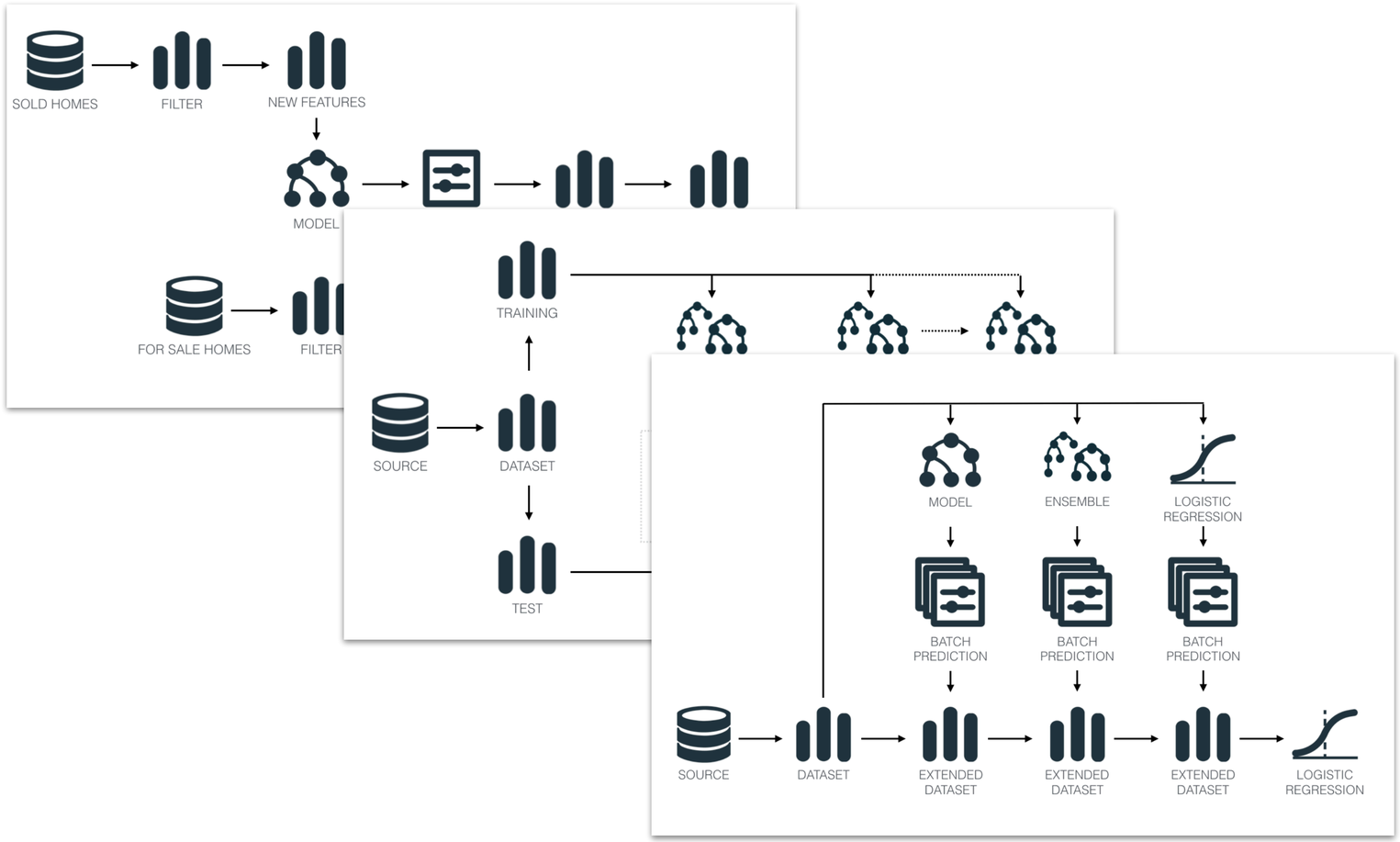
- Goal definition takes time
- Data tasks are the most consuming
- With tools like BigML, modeling is fast

Real World ML is Iterative



End to End ML is Compositional

BETTER PERFORMANCE WITH SIMPLER MODELS



Training Data vs Reality

KEY ML ASSUMPTION:

"Training data is a representative sample of what we will see at runtime"

- Features will be measured in the same way
- Customer behavior will not change
- Customer behavior will not change in response to deploying the model

REALITY:

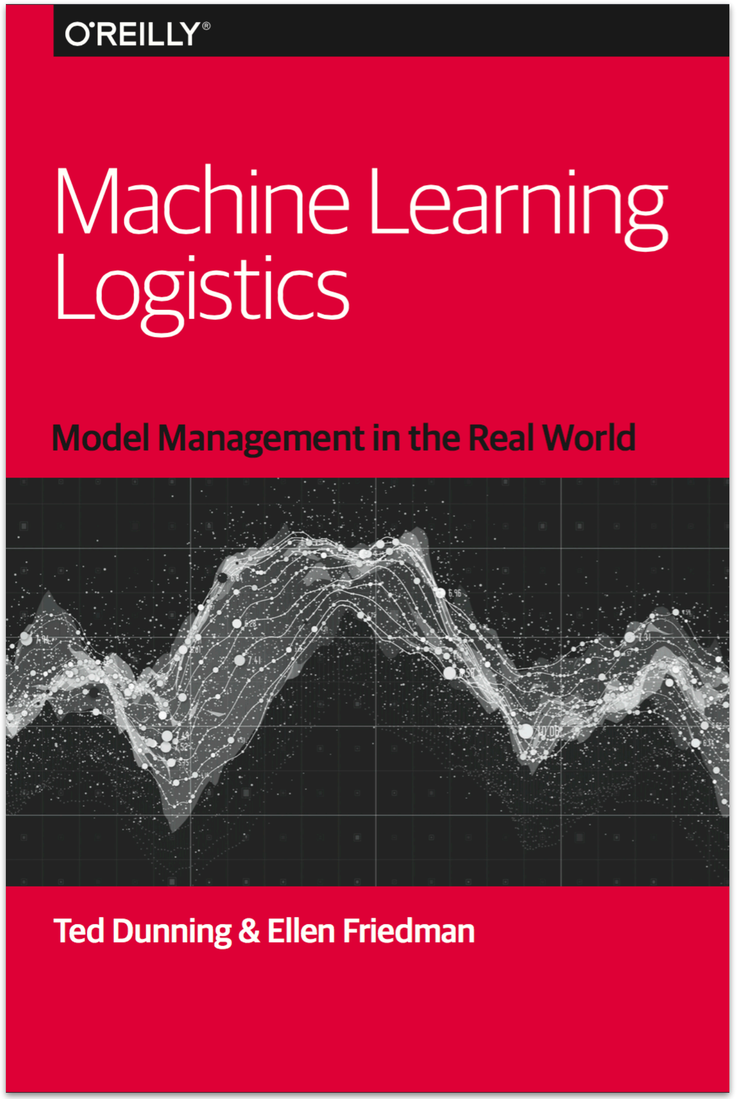
Run time data distribution is always different from training time data

Dealing With Changing Data

STRATEGY 1: RETRAIN FREQUENTLY

- Your application must allow you to observe the target variable (e.g., actual flight delay)

- This is the primary strategy at Google. They retrain many models every day

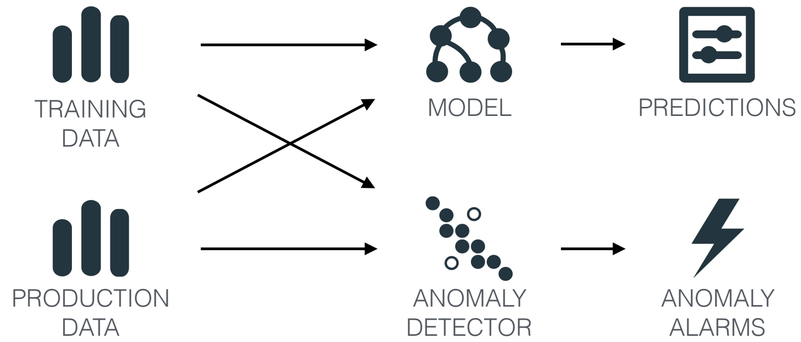


Dealing With Changing Data

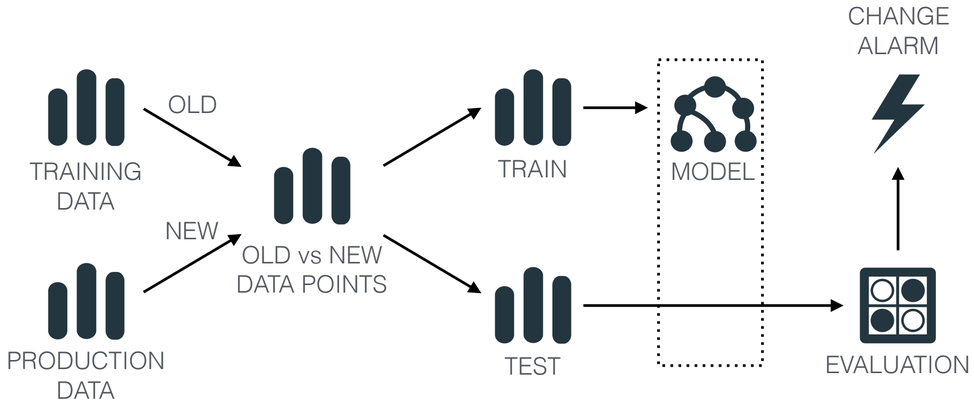
STRATEGY 2: MONITOR TO DETECT CHANGE

- Monitor changes in the distribution of predictions
- Monitor changes in the distribution of input data
  - Anomaly detection
  - Change detection: Train a classifier to see how well it can tell the difference between training data and production data
  - Monitor changes in model accuracy (if you can measure it)
- Retrain if change is detected
- Google also does automated model rollback to the best recent model

ANOMALY DETECTION



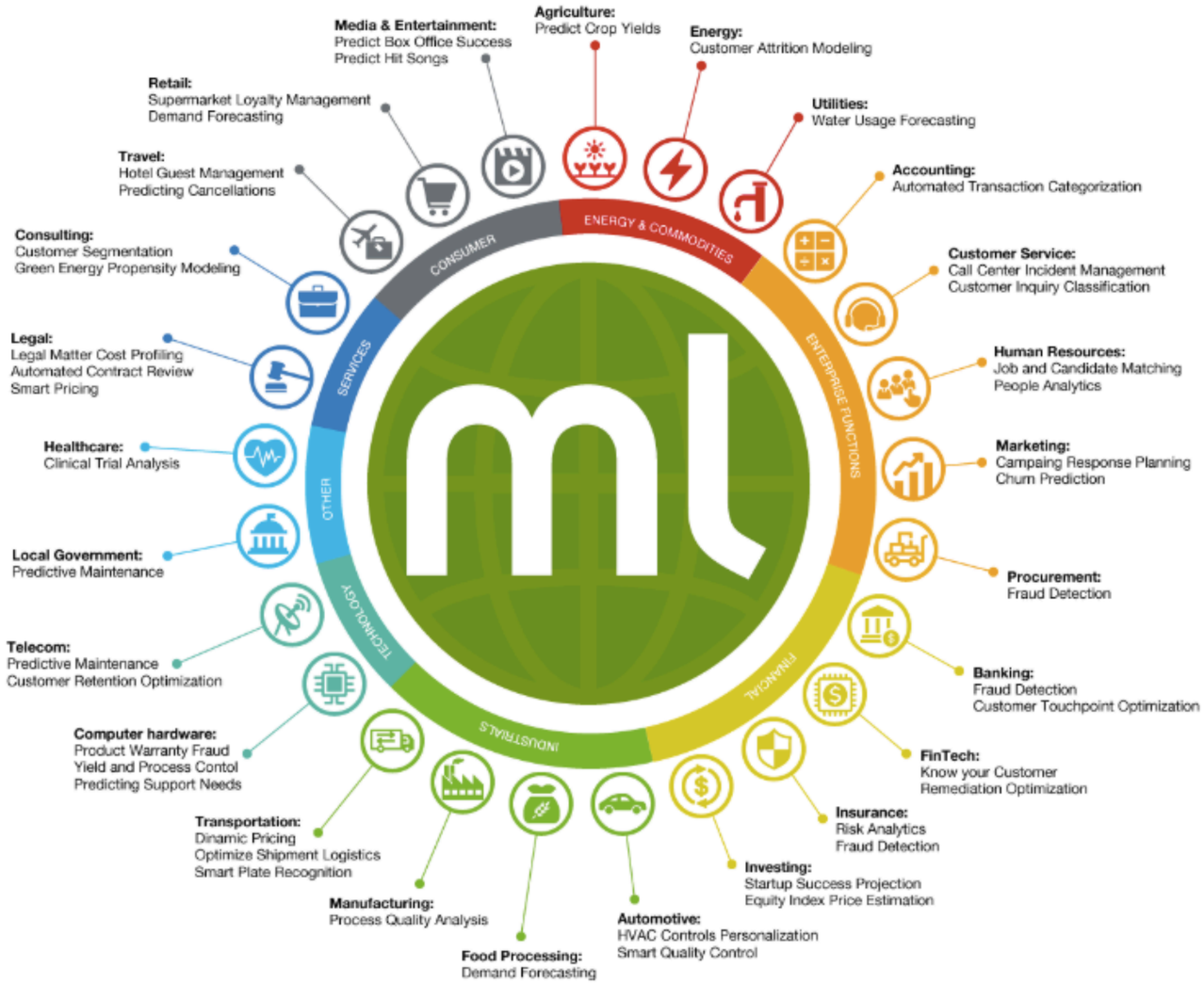
CHANGE DETECTION



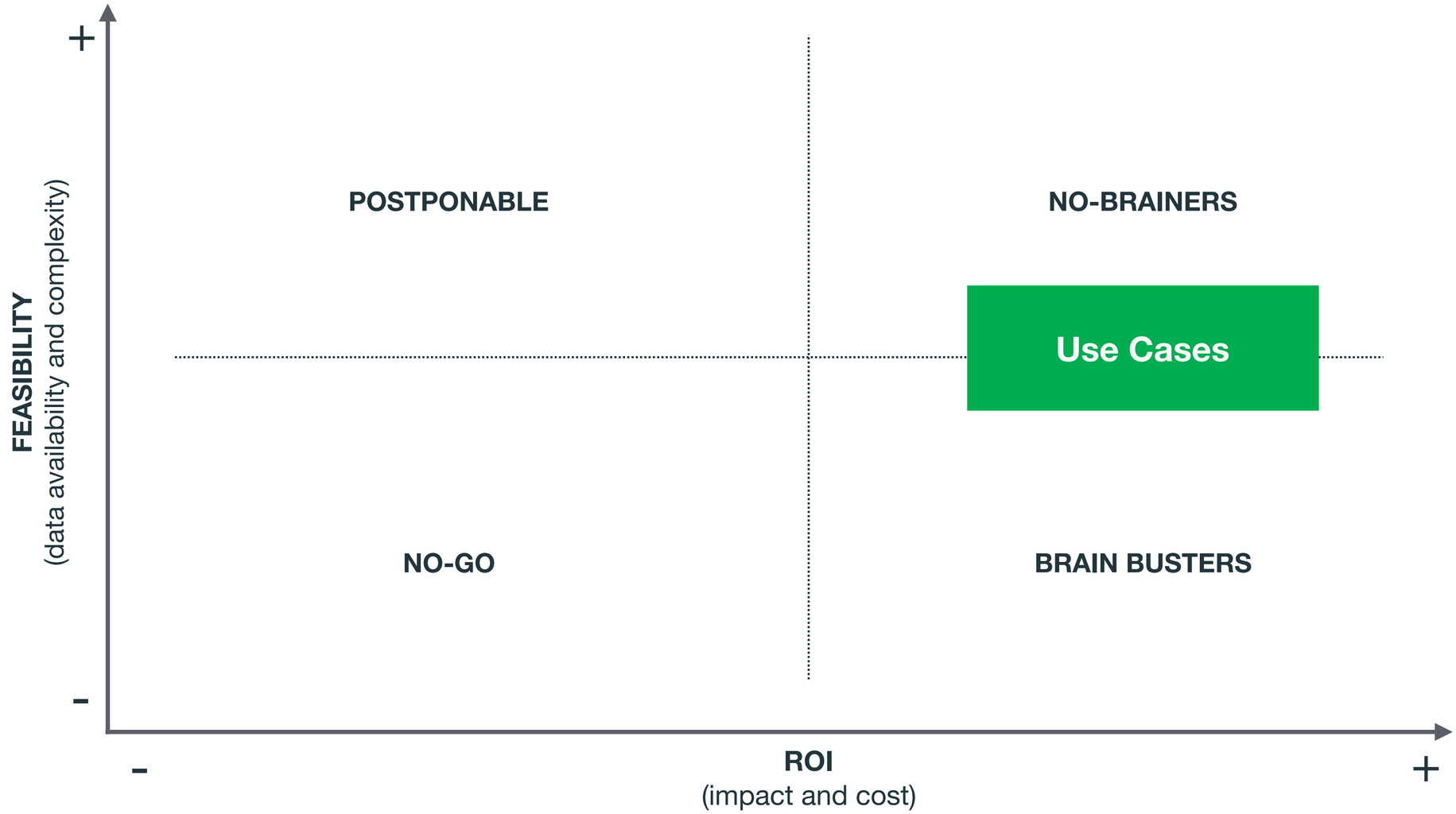
Agenda

1	Machine Learning Business Trends
2	What is Machine Learning?
3	What are the Engineering Challenges in Machine Learning?
4	**Example Application Stories from BigML**
5	Strategic Considerations

ML Use Cases


Media & Entertainment:
Predict Box Office Success
Predict Hit Songs
Agriculture:
Predict Crop Yields
Energy:
Customer Attrition Modeling
Retail:
Supermarket Loyalty Management
Demand Forecasting
Utilities:
Water Usage Forecasting
Travel:
Hotel Guest Management
Predicting Cancellations
Accounting:
Automated Transaction Categorization
Consulting:
Customer Segmentation
Green Energy Propensity Modeling
Customer Service:
Call Center Incident Management
Customer Inquiry Classification
Legal:
Legal Matter Cost Profiling
Automated Contract Review
Smart Pricing
Human Resources:
Job and Candidate Matching
People Analytics
Healthcare:
Clinical Trial Analysis
Marketing:
Campaing Response Planning
Churn Prediction
Local Government:
Predictive Maintenance
Procurement:
Fraud Detection
Telecom:
Predictive Maintenance
Customer Retention Optimization
Banking:
Fraud Detection
Customer Touchpoint Optimization
Computer hardware:
Product Warranty Fraud
Yield and Process Contol
Predicting Support Needs
FinTech:
Know your Customer
Remediation Optimization
Transportation:
Dinamic Pricing
Optimize Shipment Logistics
Smart Plate Recognition
Insurance:
Risk Analytics
Fraud Detection
Investing:
Startup Success Projection
Equity Index Price Estimation
Manufacturing:
Process Quality Analysis
Automotive:
HVAC Controls Personalization
Smart Quality Control
Food Processing:
Demand Forecasting
CONSUMER
ENERGY & COMMODITIES
ENTERPRISE FUNCTIONS
FINANCIAL
INDUSTRIALS
TECHNOLOGY
OTHER
SERVICES

Picking Use Cases



Predict Customer Churn

PROBLEM

- In the solar panel leasing sector in Africa, both service usage and payments are erratic.
- It is difficult to anticipate cancelations.
- The cost of losing a customer is about 100 times higher than paying to retain one.

DATA

Historical Customer Data

- Customer profiles.
- Usage patterns.
- Payment behavior.

Source: OFF·GRID ELECTRIC 2ML18

ALGORITHMIC MODELING PROCESS

bigml

Classification
To classify if a customer is going to churn in the next month.

Associations Discovery
Used in combination with feature importance and prediction explanations to understand factors and conditions which lead to churn.

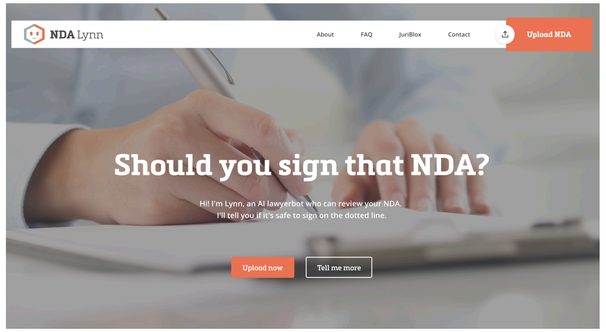
Cluster Analysis
To determine which customers have similar behavior.

APPLICATION / REPORTS

- Predict if a customer will churn next month along with a confidence level.
- Prediction models help provide personalized support.
- Timely payers may get more advantageous rates on new products.
- Identifying customers or prospects that are more likely to pay on time.

Predict NDA Signature

PROBLEM



- In the legal sector there is a need to automatically identify if an NDA should be signed or not as driven by the existence of potential deal breaker clauses.

DATA

Historical NDA Data

- NDA metadata and text prepared to train the algorithms, from a large corpus of original NDAs:

Source: juriblox 2ML18

ALGORITHMIC MODELING PROCESS

bigml

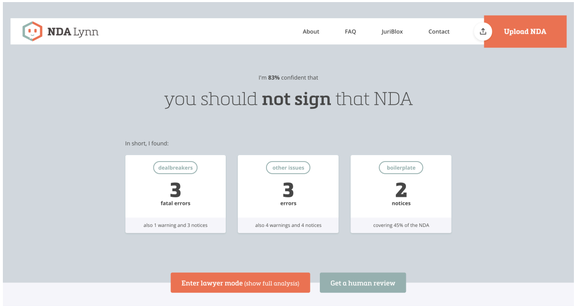
Classification
Ensembles to classify the sentences in different clauses.

Classification
Ensembles to classify different clauses.

Classification
Boosted trees to predict if the NDA should be signed or not.

APPLICATION / REPORTS

- Machine Learning-based approach can forecast whether the related party should sign an NDA given an automatically calculated confidence level:



Predict Manufacturing Results

PROBLEM

- In film metallization manufacturing, when an execution is launched many components inside the machine can affect the outcome. Monitoring the machine in real time is crucial to optimize the final result and prevent costly mistakes.
- High variety of sensor measurements and parameters turn monitoring into a very complex task.

DATA

Historical Manufacturing Data

- Historic executions monitoring measures.
- Sensor and internal measures data.
- Process parameters.
- Quality assessment results.

ALGORITHMIC MODELING PROCESS

bigml

Classification
To predict in real time if current execution result will be acceptable.

Anomaly Detection
To raise alerts about components out of normal behavior

Associations Discovery
To find relevant relationships in data

APPLICATION / REPORTS

- Machine Learning-based approach can forecast the outcome of a manufacturing process in order to increase quality rates and prevent machine downtime .
- A monitoring application including real time graphical analytics and alert predictions is provided to monitor the process.
- During maintenance windows, an alert summary is displayed to focus attention on specific components that may be operating outside ideal boundaries.
- System is capable of anticipating manufacturing problems in real time.

Predict Account Transactions

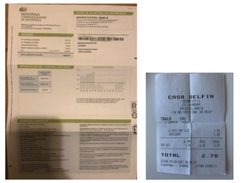
PROBLEM

- Accounting has been mainly dependent on manual human activities. More automation is now possible for tasks such as classifying transactions.

DATA

Historical Transaction Data

- Accounting transactions.



Source: anfix 2ML18

ALGORITHMIC MODELING PROCESS

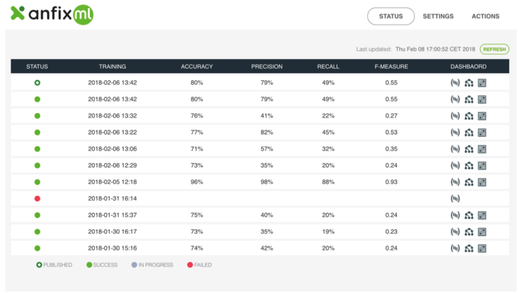
bigml

Classification
Boosted trees to classify different types of transactions.

Classification
To predict the detailed category of transactions.

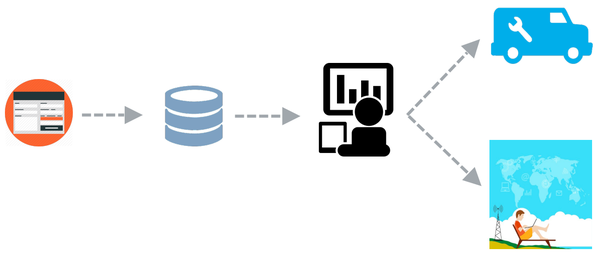
APPLICATION / REPORTS

- Machine Learning-based approach can recognize different types of transactions e.g, income or expense and applicable sub-types or tags.



Predict Incident Category

PROBLEM



- Printing company's call center receive thousands of requests to resolve customer's incidents.
- Tp make Call Center operations more efficient, we automate the classification of the type of request for better and faster issue resolution.

DATA

Historical Incident Data

- Printer characteristics.
- Incident description.
- Contract characteristics.
- How solved? Onsite vs. remote etc.

Source: CleverData (a CleverTask Company) BigData Prediction VSSML18 School

ALGORITHMIC MODELING PROCESS

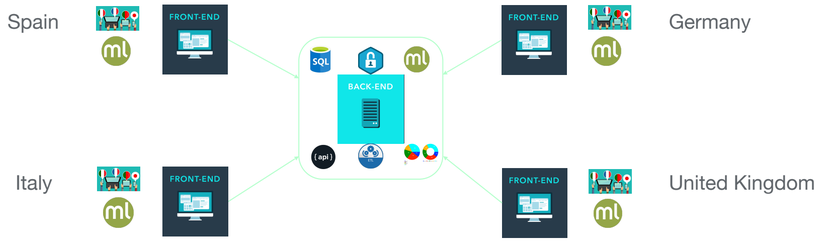
bigml

Classification
To predict best channel to solve customer tickets (Remote or OnSIte).

Topic Model
To produce features from incident descriptions (free text).

APPLICATION / REPORTS

- Machine Learning-based approach can predict proper actions for incidents.
- The business objective is to make operations in Call Center more efficient.
- The technical objective is to make an automatic incident dispatching bot.



Predict Startup Success

PROBLEM

- Investment in early stage companies is still dominated by human inuition. These investors are often influenced by who the investment seeking startups know in their network or how well they present as opposed to many unseen factors hidden in the data

DATA

Historical Company Data

- Crunchbase data
- Twitter
- Patents

Source: preseries 2ML18

ALGORITHMIC MODELING PROCESS

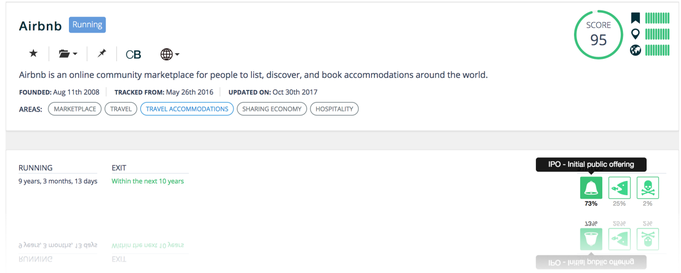
Classification
To predict if a company will have an IPO in the next 10 years

Cluster Analysis
To determine which companies display similar behavior

Associations Discovery
Used in combination with feature importance and prediction explanations to understand factors and conditions which lead to startup success

APPLICATION / REPORTS

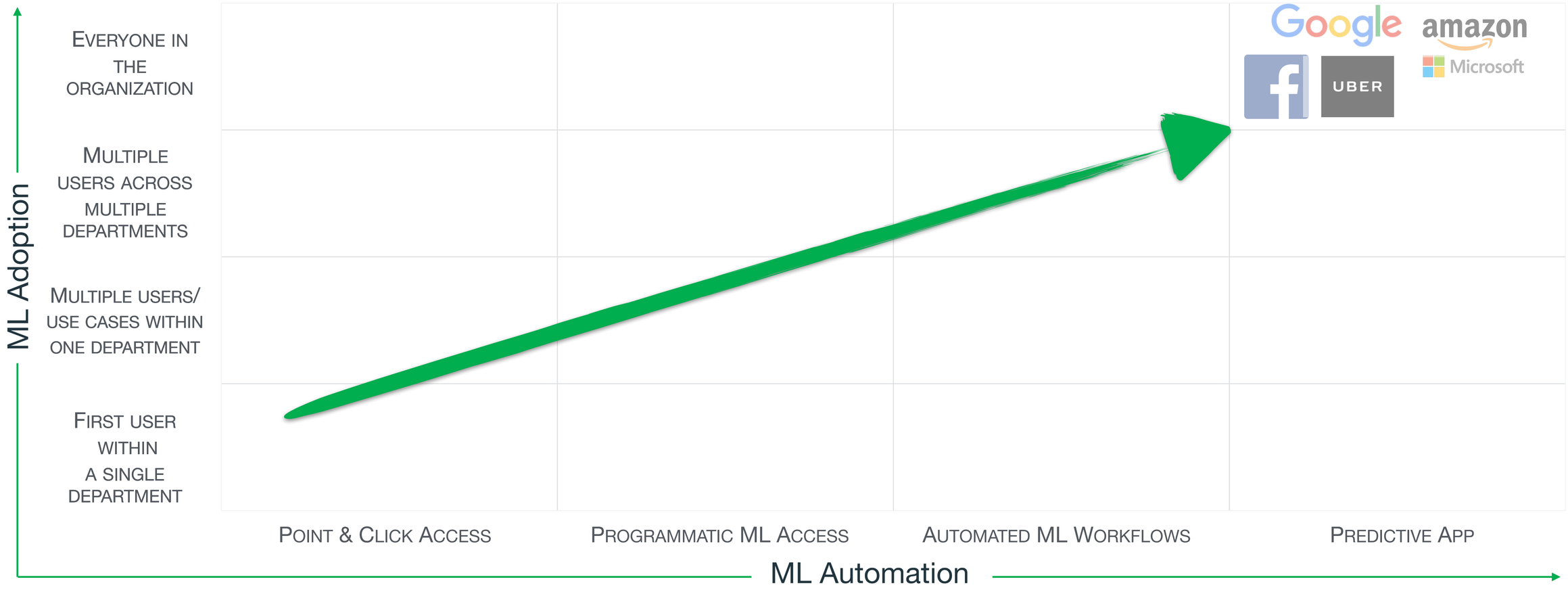
- Machine Learning-based approach can predict the future evolution of a startup e.g., whether the company will have an IPO in the next 10 years with the associated probability level or just continue operating or shut down



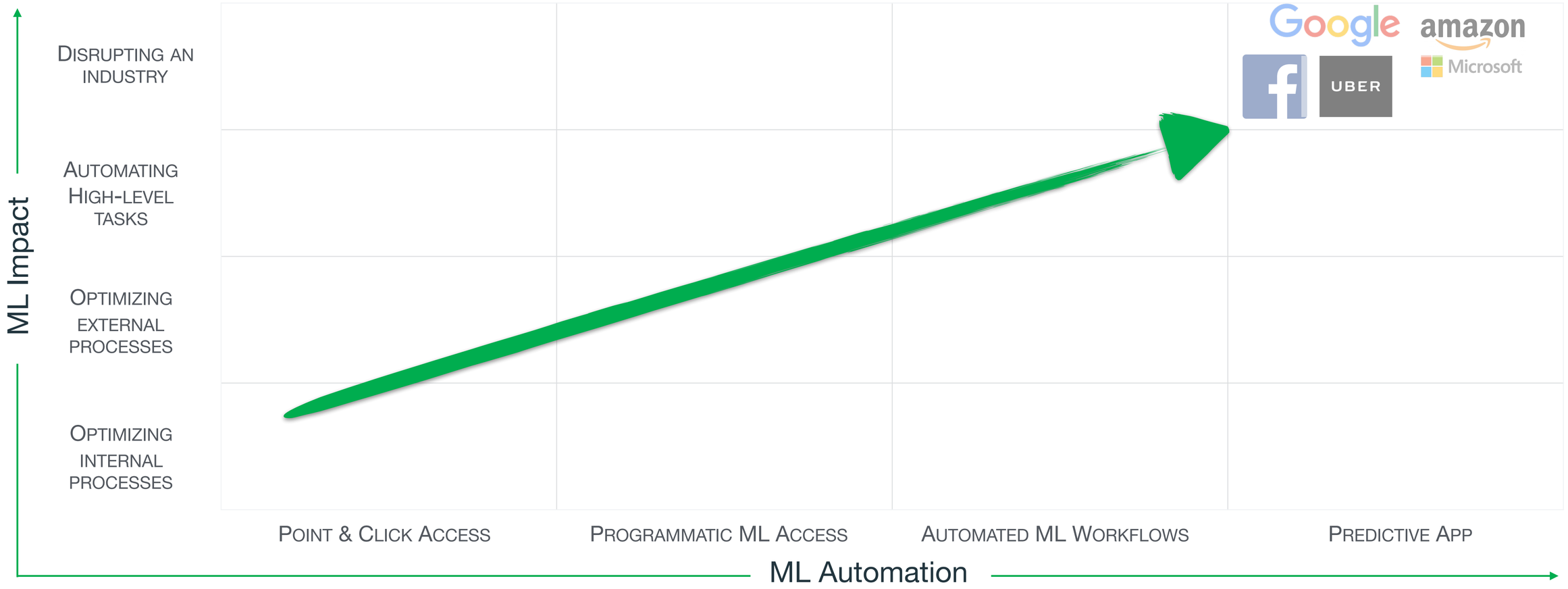
Agenda

1	Machine Learning Business Trends
2	What is Machine Learning?
3	What are the Engineering Challenges in Machine Learning?
4	Example Application Stories from BigML
5	Strategic Considerations

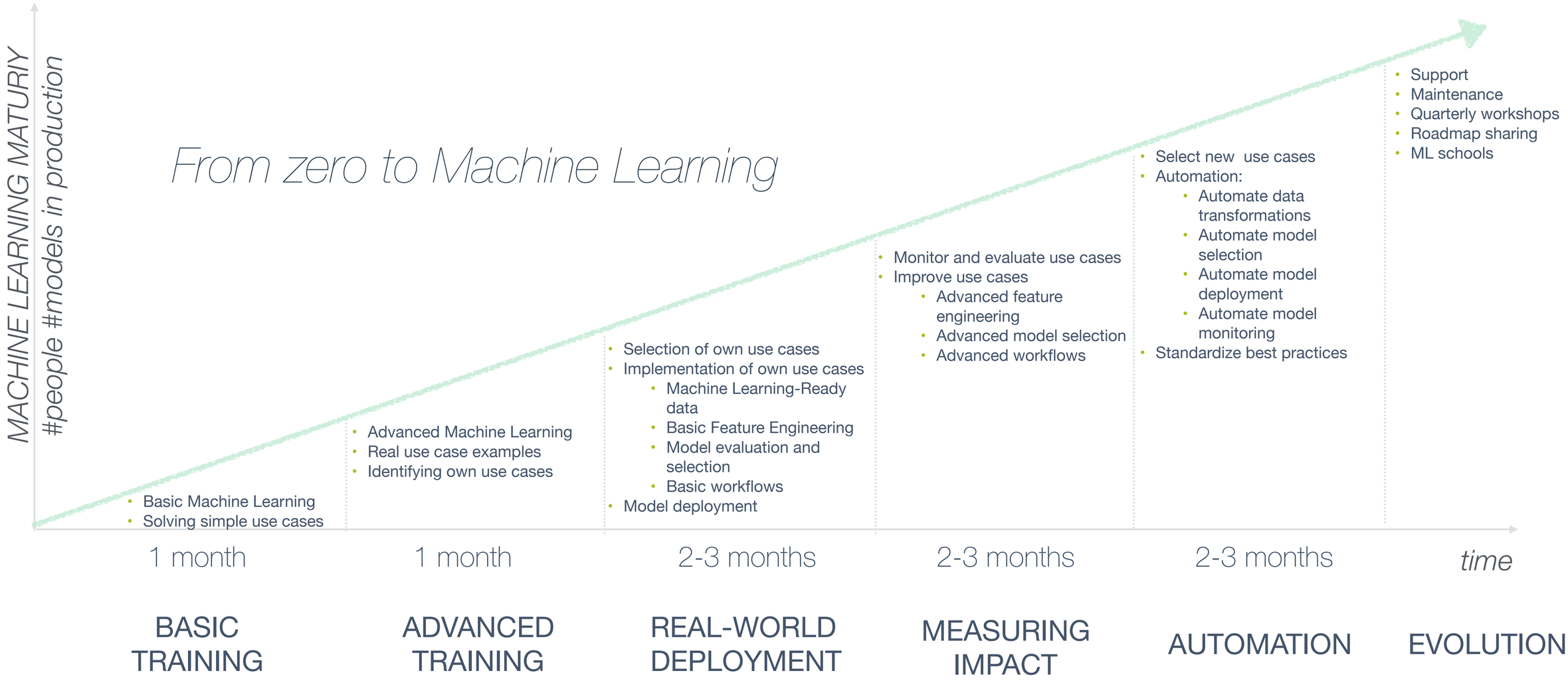
Machine Learning Adoption



Machine Learning Impact



Roadmap to Adapt ML at Scale



Q&A


DSN
Data Science Nigeria


1868
Oregon State
University


bigml®