Scaling Predictive Analytics with Data Science Automation

主讲人：Max Kanter，Feature Labs CEO

Who am I?

- Passionate about **making data science more accessible**
- Former student and machine learning researcher at the **Massachusetts Institute of Technology** (MIT)
- Previously, an engineer at Twitter, Hewlett Packard, The New York Times, and Fitbit
- Currently, the **CEO of Feature Labs (Boston, USA)**

Computer Science and Artificial Intelligence Lab @ MIT

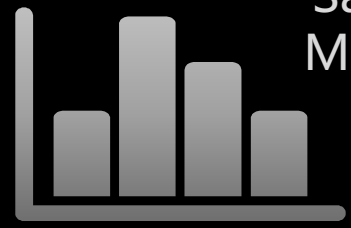
Why is Data Science Important?

Customer Behavior

Economic

Energy

Health



Sales and Marketing

Finance

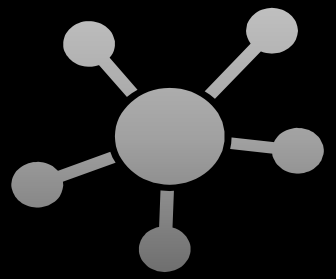
Data problems are everywhere

Agriculture

Supply chain

Cyber

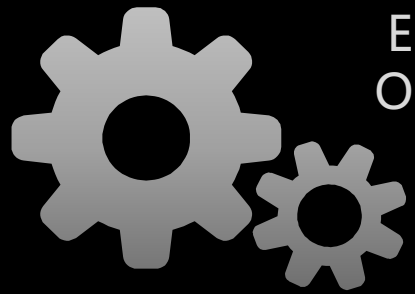
Device and IOT



Sports

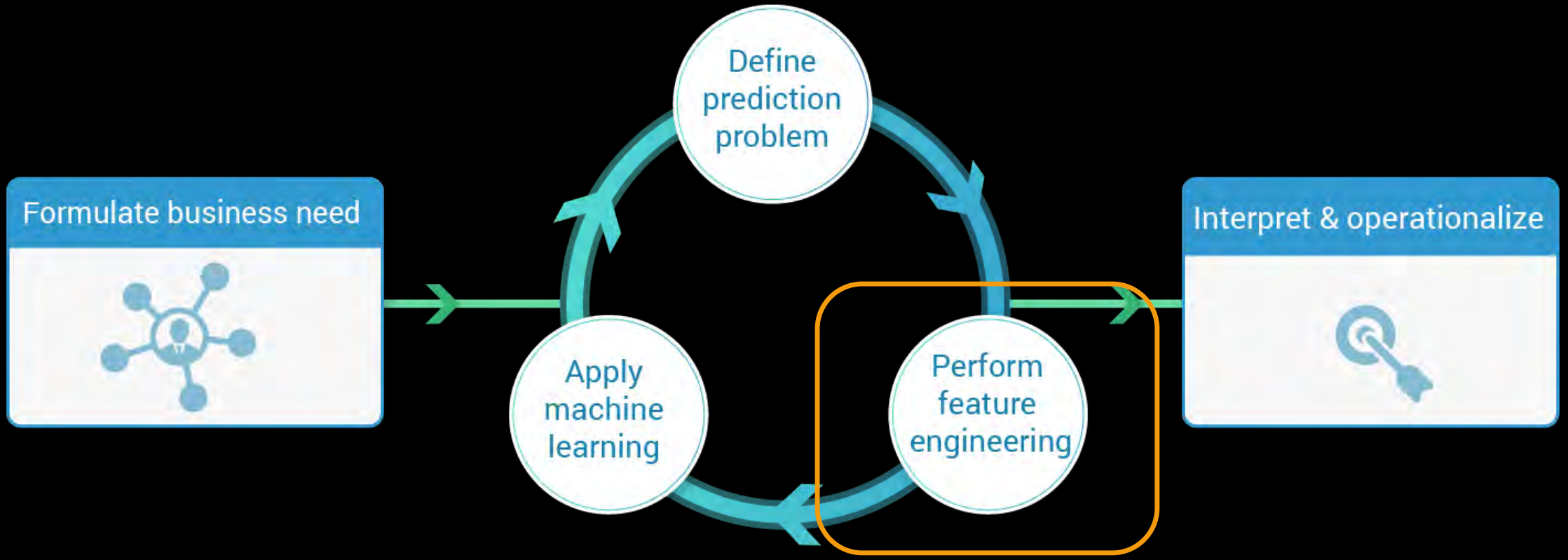
Education

Manufacturing



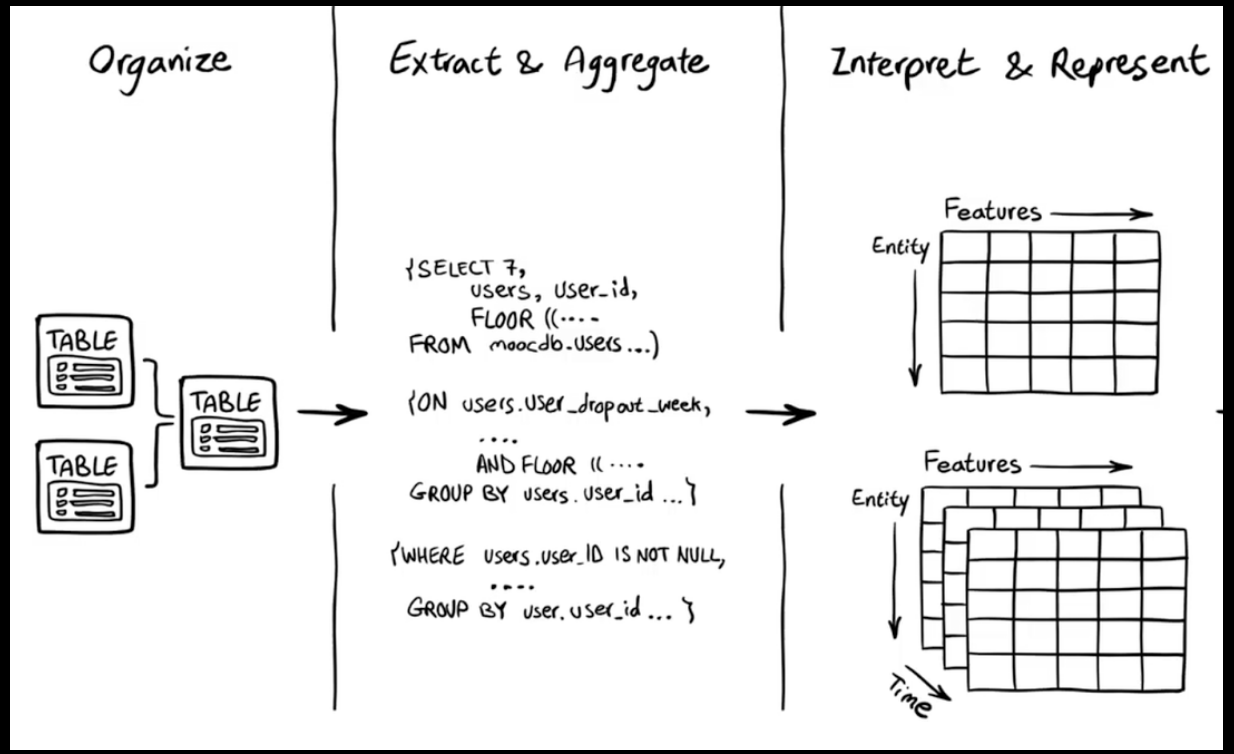
Enterprise Operations

The Data Science Process



Feature Engineering

- Human-driven
- Iterative
- Most time consuming part of data science process
- **BUT, it is key to developing high performing models**



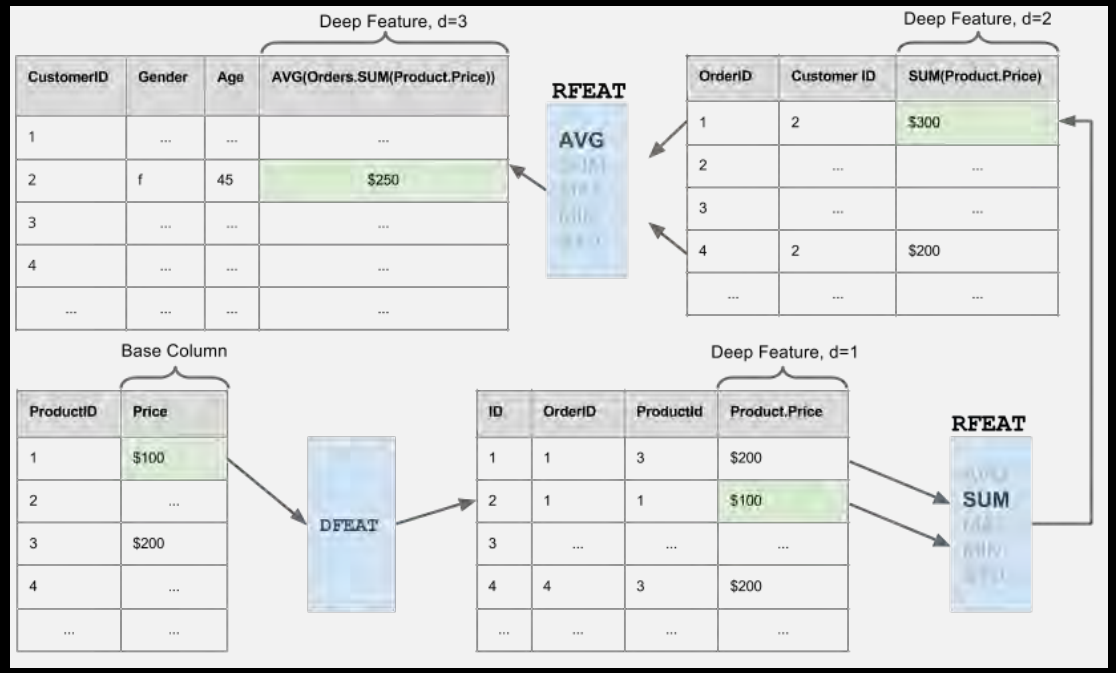
Typical feature engineering process

Automated Feature Engineering

Invented at MIT in 2015

How DFS works

1) Define repository of "primitive" functions
   - Primitives each define simple transformation
2) Apply primitives to columns or across relationships
3) "Stack" primitives to form complex calculations



Applying DFS to example database

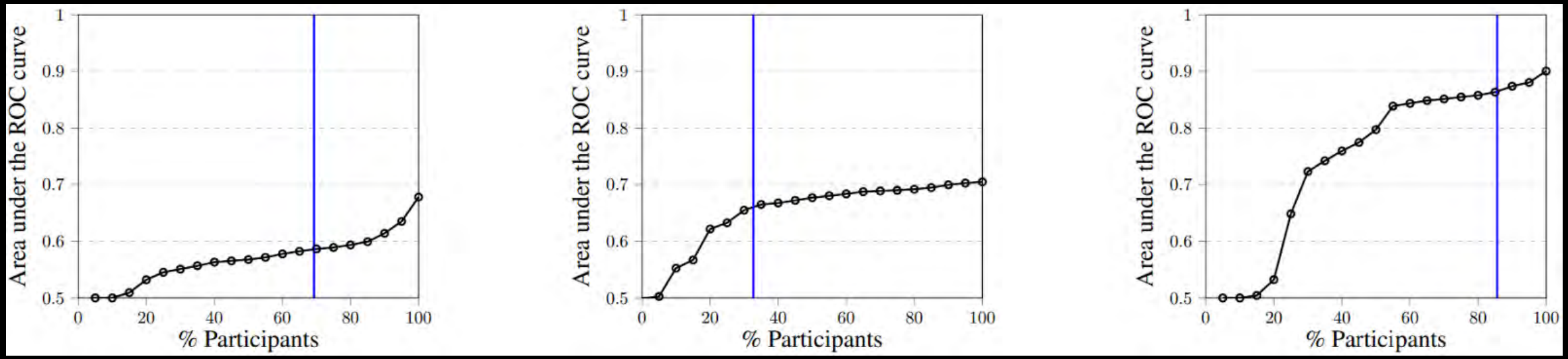
Validating Deep Feature Synthesis

tested against
1000
data scientists on
kaggle

on average
92%
of top score

over
1200
days saved

Validating Deep Feature Synthesis



1. Project Excitement

2. Repeat Buyer

3. Dropout Prediction

Open Source Library for Deep Feature Synthesis

Featuretools

Install for free at **www.featuretools.com**

Relational data -> feature matrix with one function

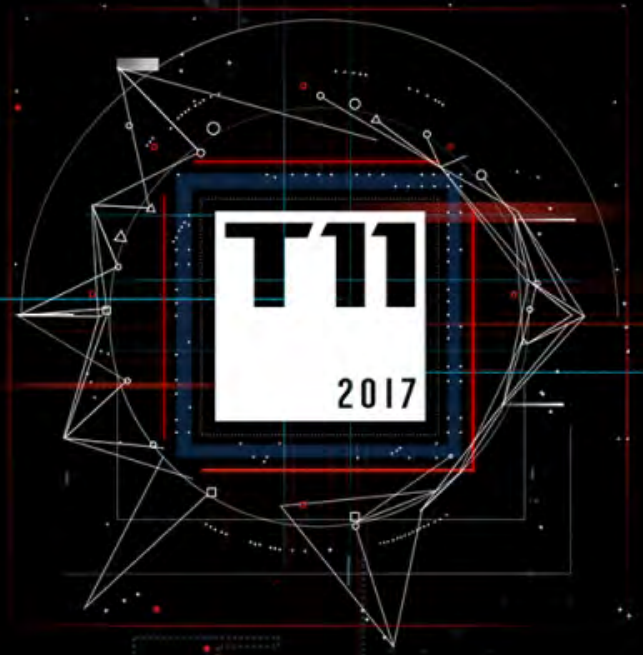
```
In [11]: feature_matrix_customers, features_defs = ft.dfs(entities=entities,
   ....:                                                  relationships=relationships,
   ....:                                                  target_entity="customers")
   ....:

In [12]: feature_matrix_customers
Out[12]:
             zip_code  COUNT(transactions)  COUNT(sessions)  SUM(transactions.amount) MODE(sessi
customer_id
1               60091                  131               10                  10236.77
2               02139                  122                8                   9118.81
3               02139                   78                5                   5758.24
4               60091                  111                8                   8205.28
5               02139                   58                4                   4571.37

[5 rows x 69 columns]
```

Featuretools

Conclusions

What is Data Science Automation?

1. Technology to help those new to data science
2. Tools for to improve existing data scientists
3. Ready to use **TODAY**! (featuretools.com)



THANKS

Email: max@featurelabs.com

Twitter: @maxk