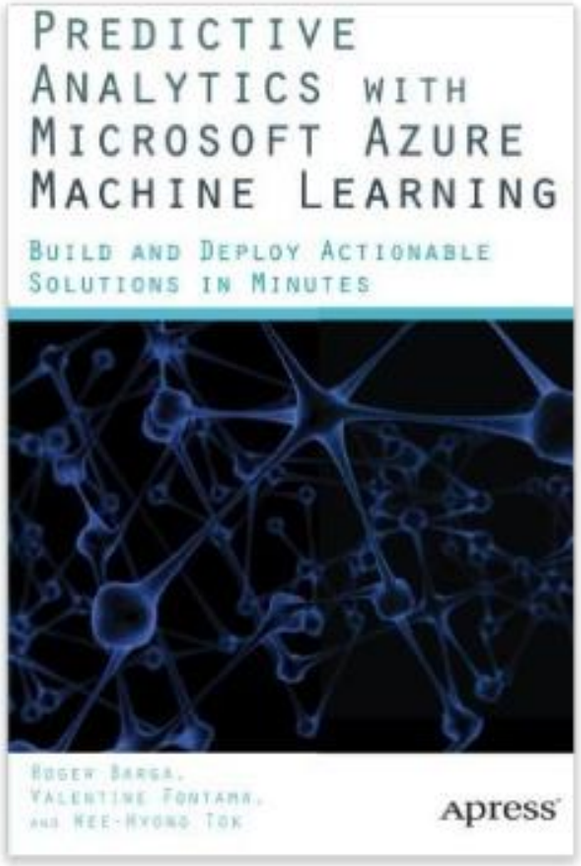
Discover your Data Science Super Power

Tips and Tricks to a Successful Data Science Project

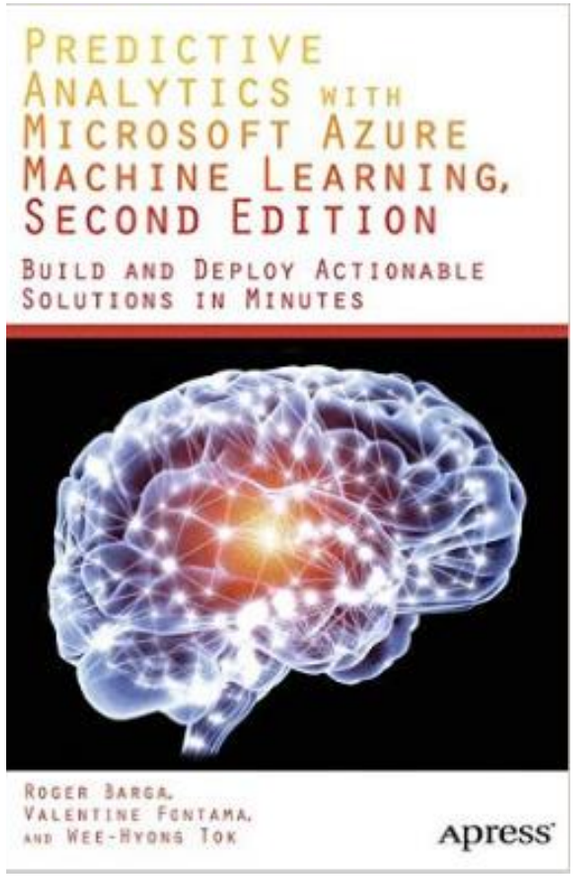
Wee Hyong Tok
Algorithms and Data Science
Microsoft

@weehyong

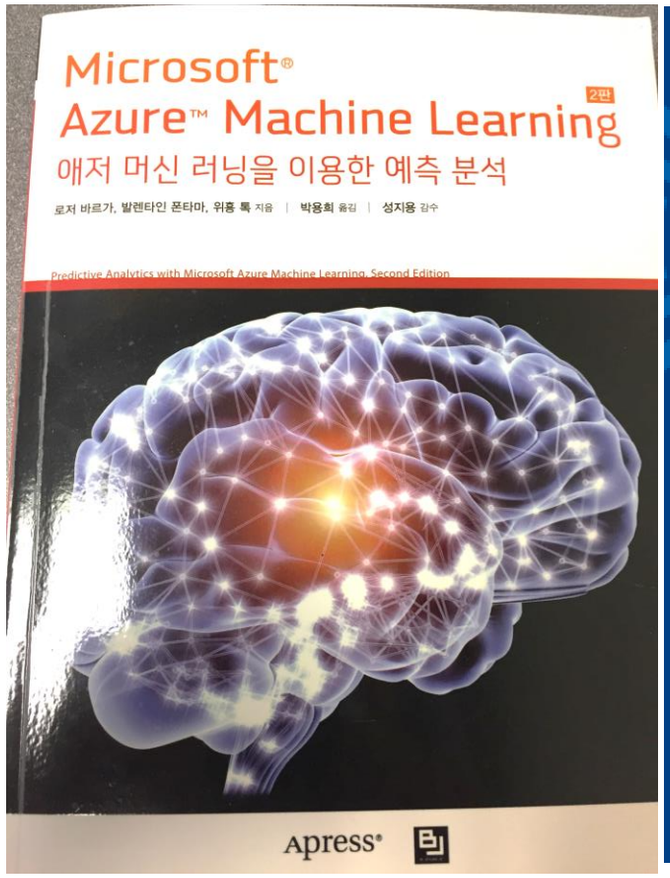
My Data Science Journey

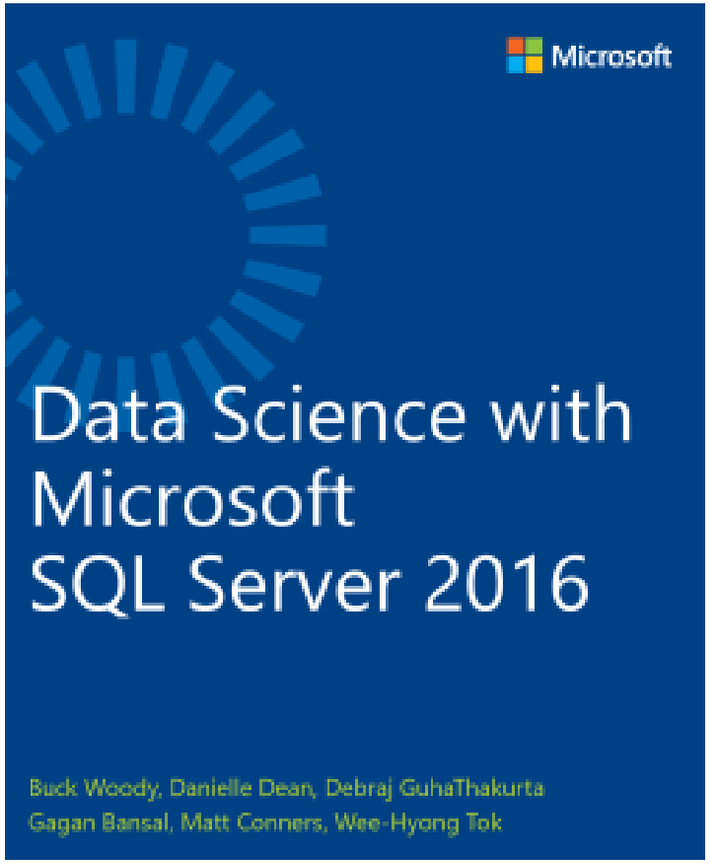


2014

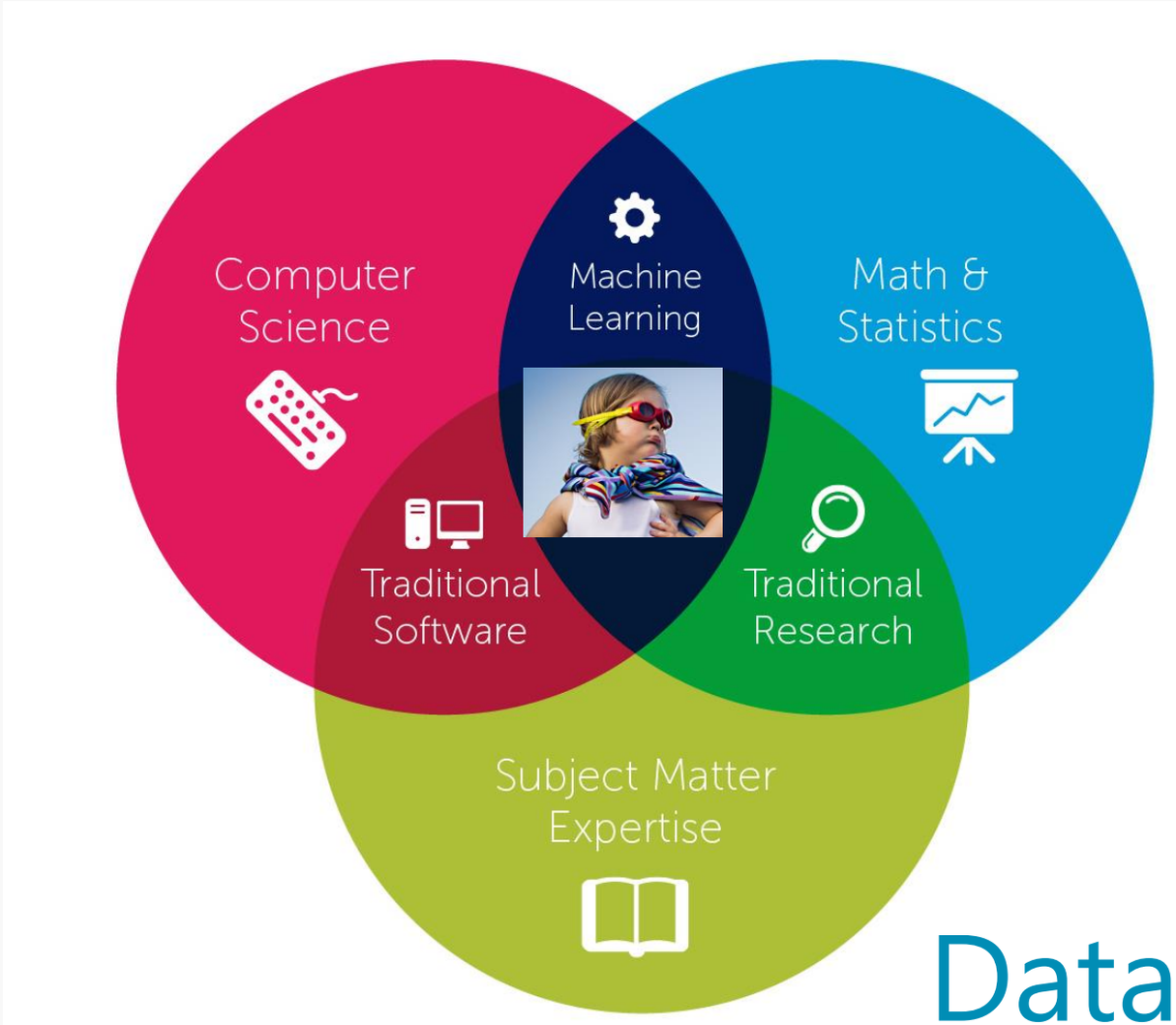


2015





2016



Data Science

Terminology

Training Data : A set of samples (table of data)

Features: Individual columns in our data set

Label / target: Historical outcome or result related to a set of samples

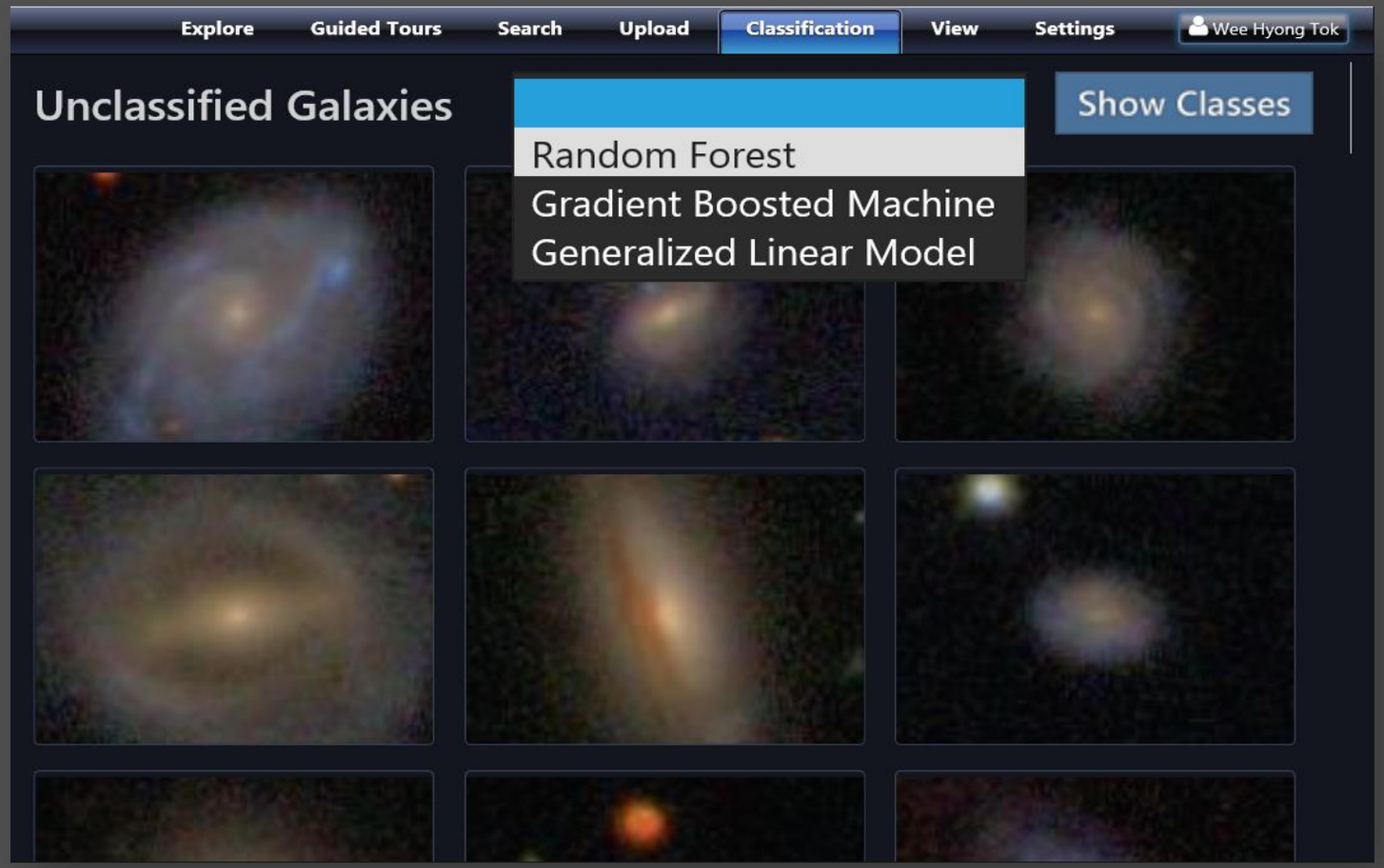
Learner: Machine learning algorithm

Feature Engineering / Munging: Manipulating the data to come to the training dataset


Let’s discover your super powers…

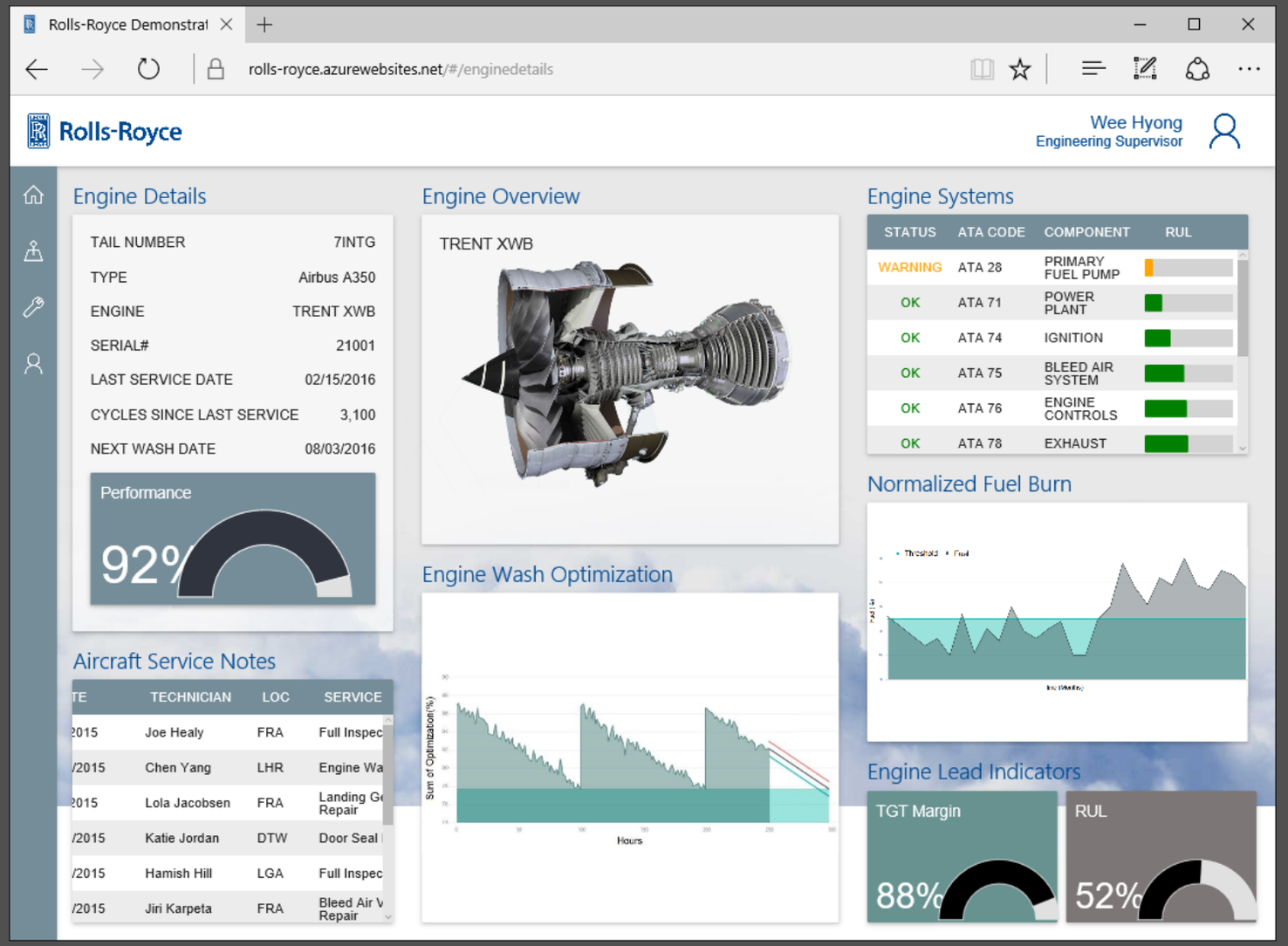
What are your Super Powers?

Super Power #1
Be Obsessed with Data



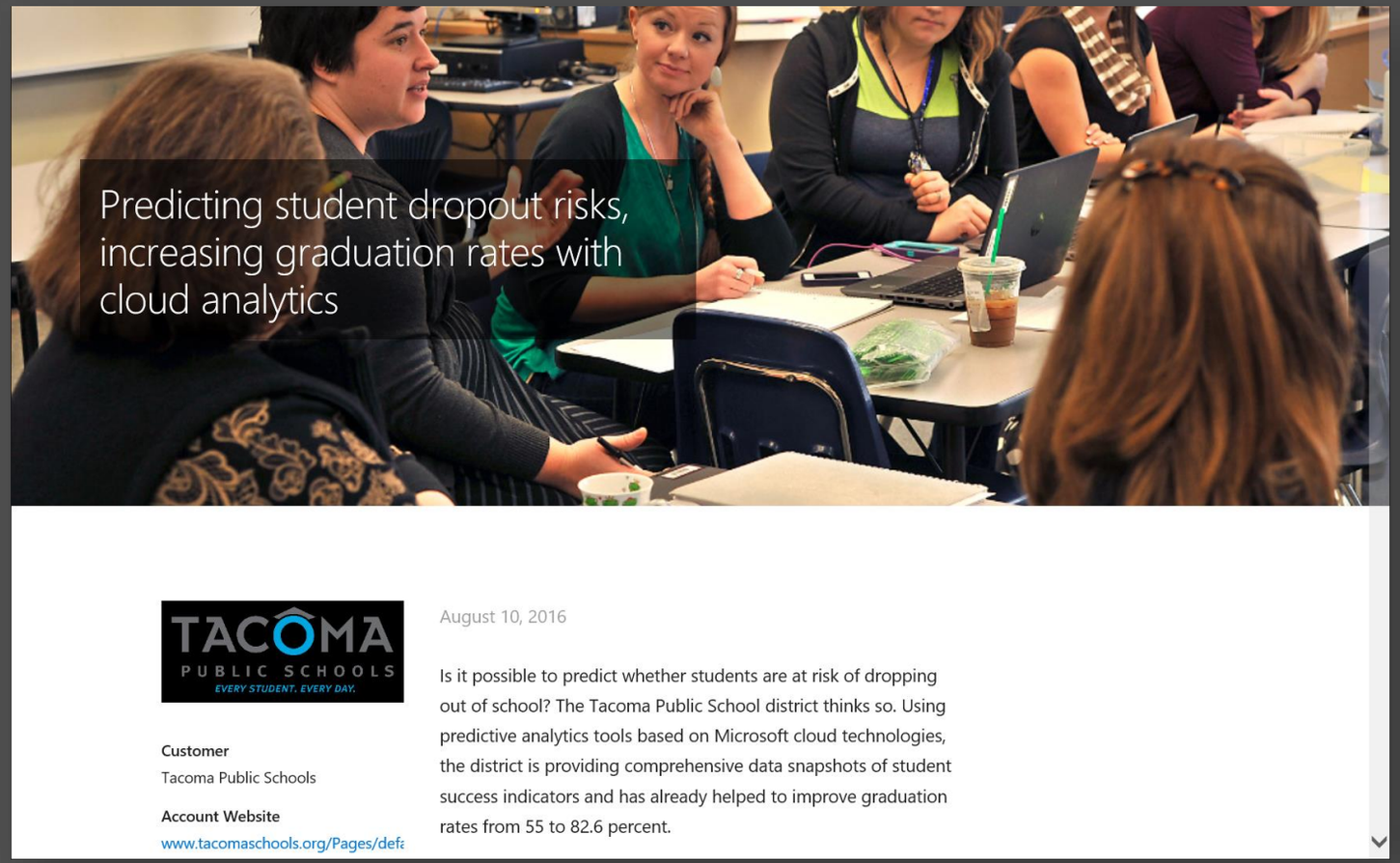
Helping Astronomers Classify Galaxies

Data Volume:12TB
SQL Server Database
(Sloan Digital Sky Survey DR12)



Predicting Remaining Useful Lifetime (RUL) of Engine Systems

New Generation of Smart Refrigerators



"With Azure Machine Learning, we proved that we have the right tool to get us where we want to go in terms of predicting student success. It's a tool our educators will be able to use to start tackling the problem of student disengagement.
"

Shaun Taylor: CIO
Tacoma Public Schools

Getting from Vague to Precise Questions

Vague

I want to do advanced analytics

I want to use the data to improve my business. How do I do that?

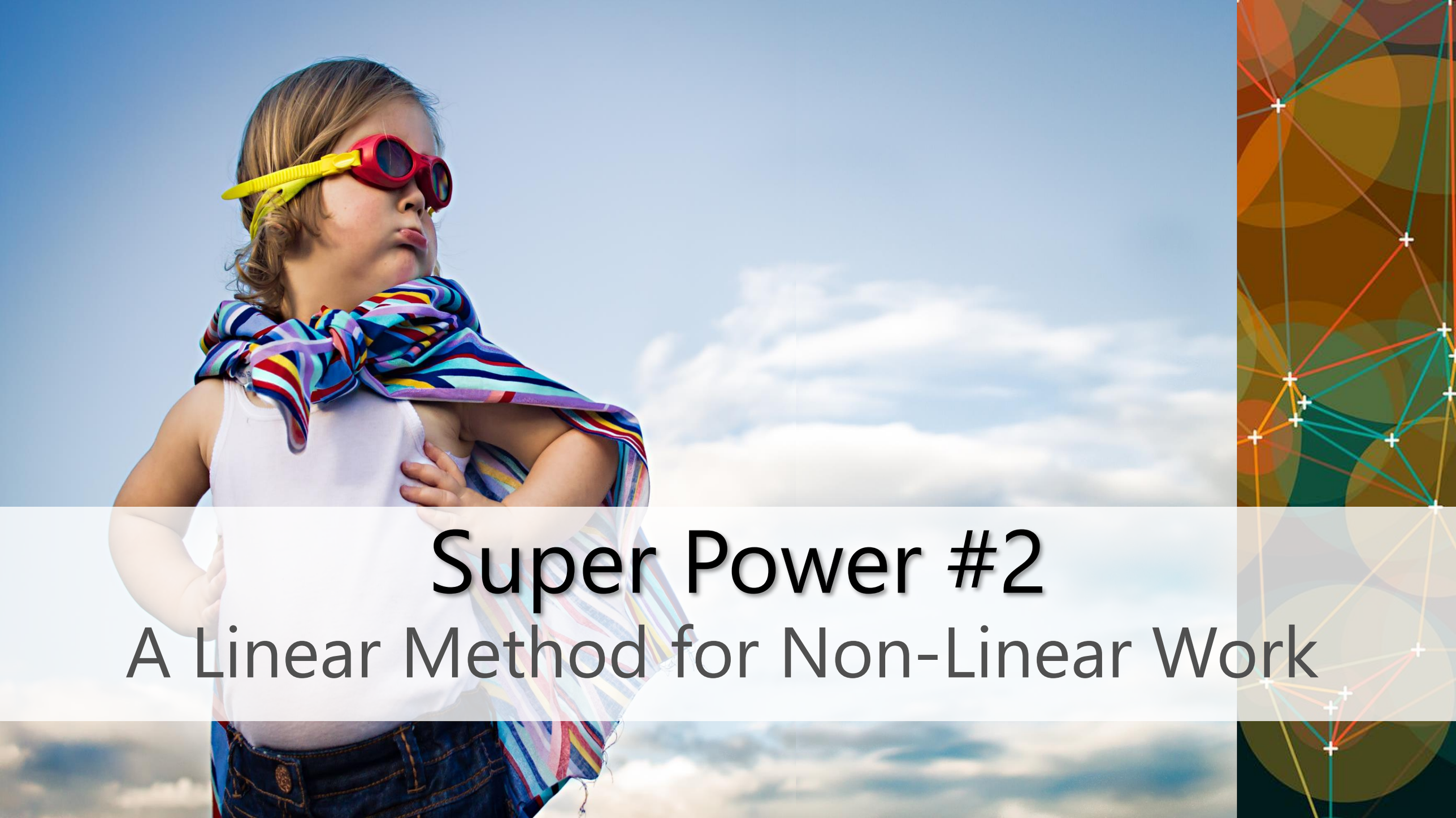
I want to build a model to increase customer retention

Precise

I want to understand how many Model Z cars will I sell in Atlanta this month

I want to know which customers will churn in the next month

What products should I recommend to my customers when they choose Product X?

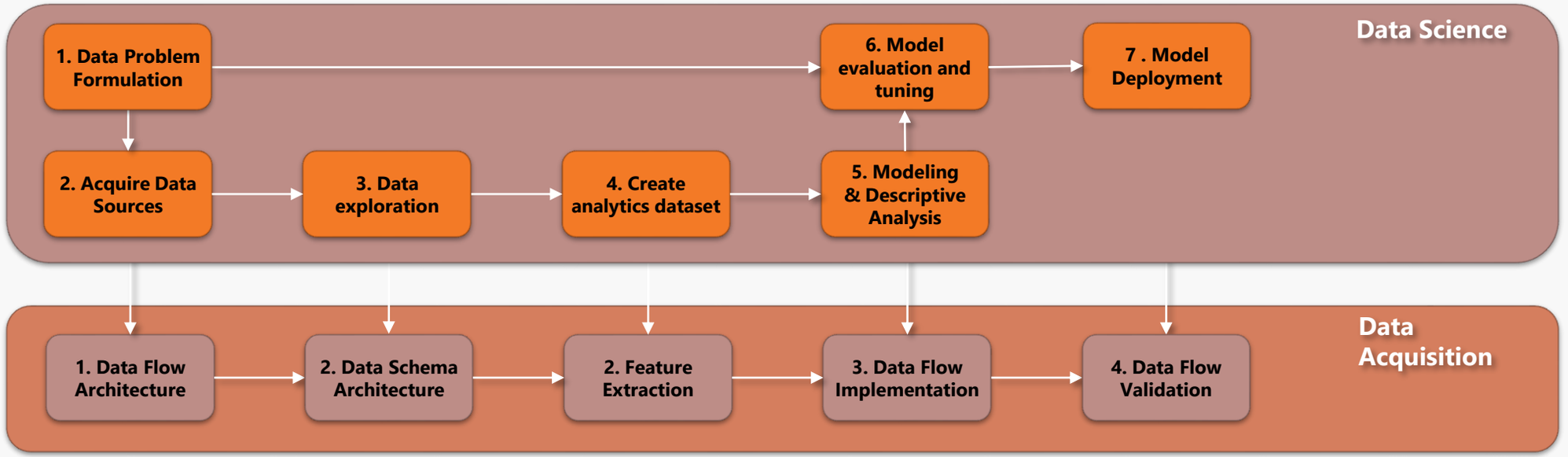


Super Power #2

A Linear Method for Non-Linear Work

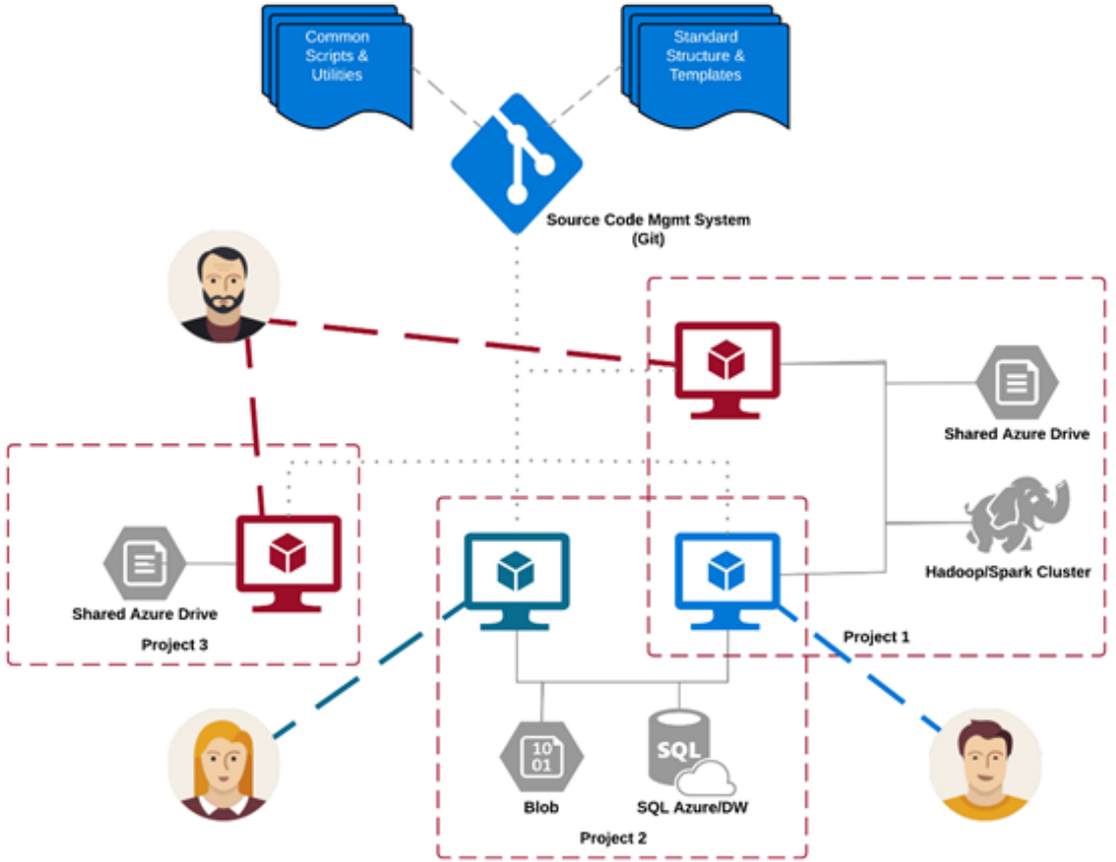
Data Science

Goal: Meet business goals through close collaboration with the business



Your process should include a team

Microsoft's **Team Data Science Process**



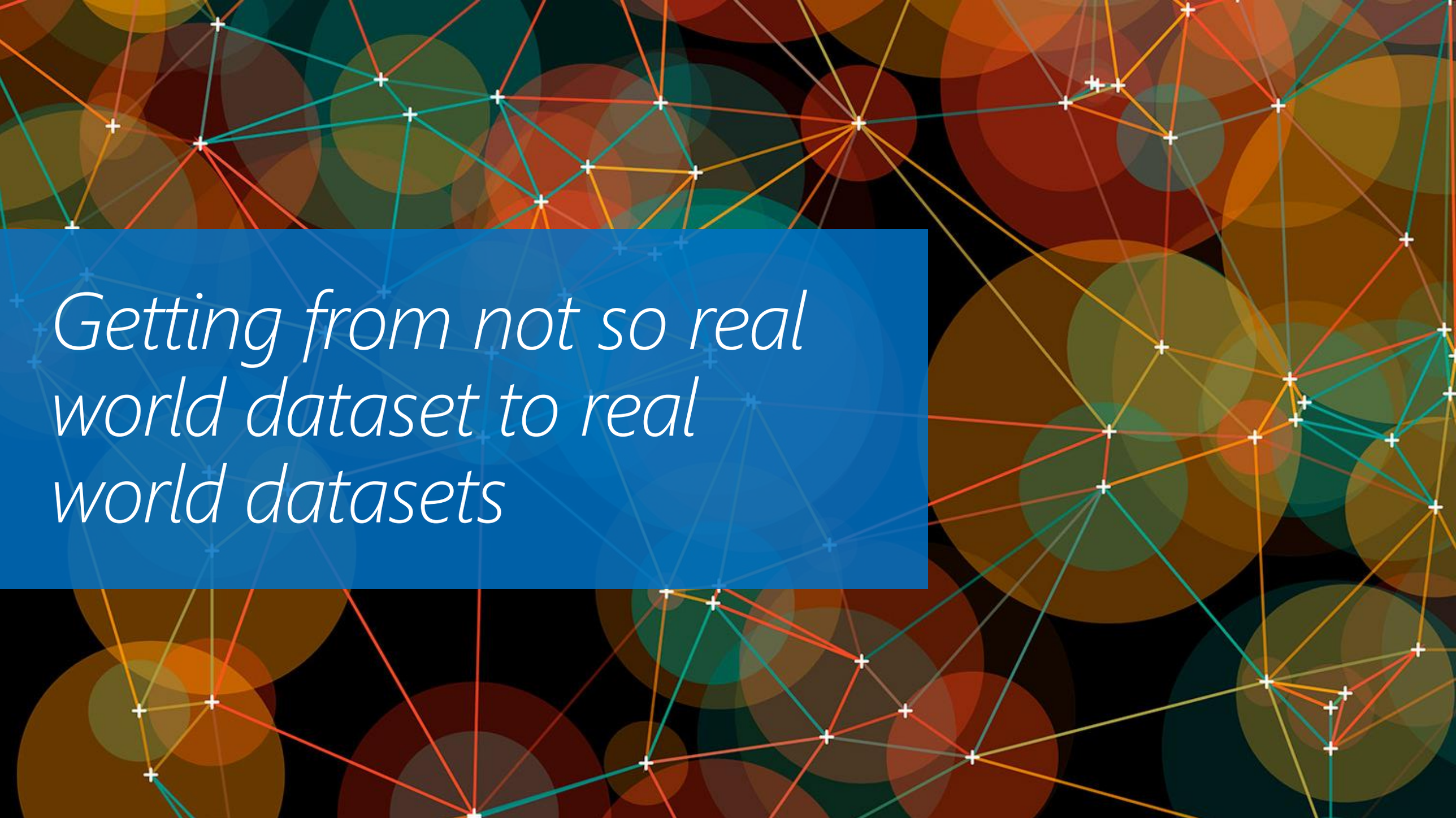
"Data Science Doesn't Just Happen, It Takes a Process. Learn about Ours…" session from 3:00-3:50pm

Super Power #3

Maintain and grow a toolbox of tricks

" Feature engineering is the most important but underrated step of machine learning."

Better features are better than better algorithms…


Getting from not so real world dataset to real world datasets

Aircraft Engine Failure

Helicopter Gearbox Faults

Fraudulent Telephone Calls

Customer Churn

What's common in all these dataset?

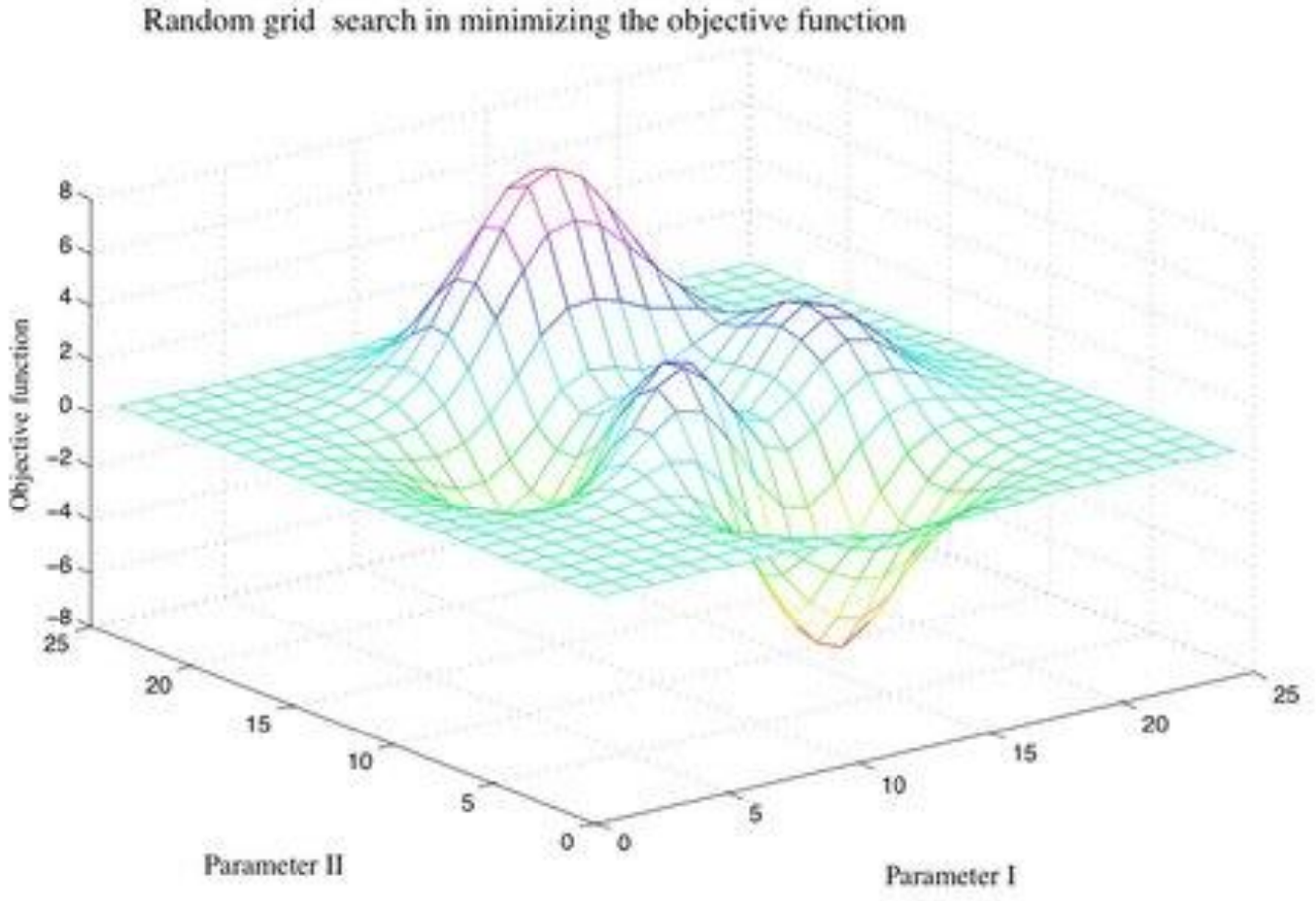
Learners are biased towards majority class.

Minority class tend to be misclassified.

Informed oversampling of the minority class with random undersampling of the majority class.

SMOTE - Synthetic Minority Oversampling Technique

Why am I doing so many iterations of trial and error?

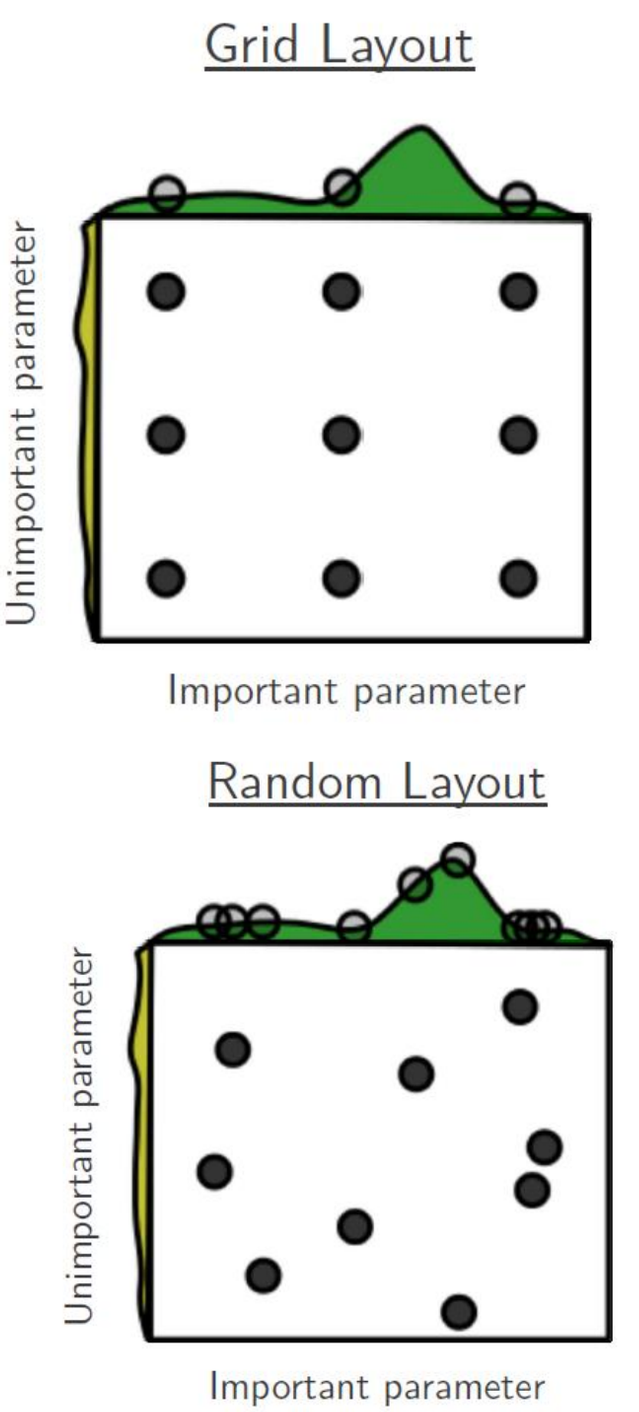


Simplest approach

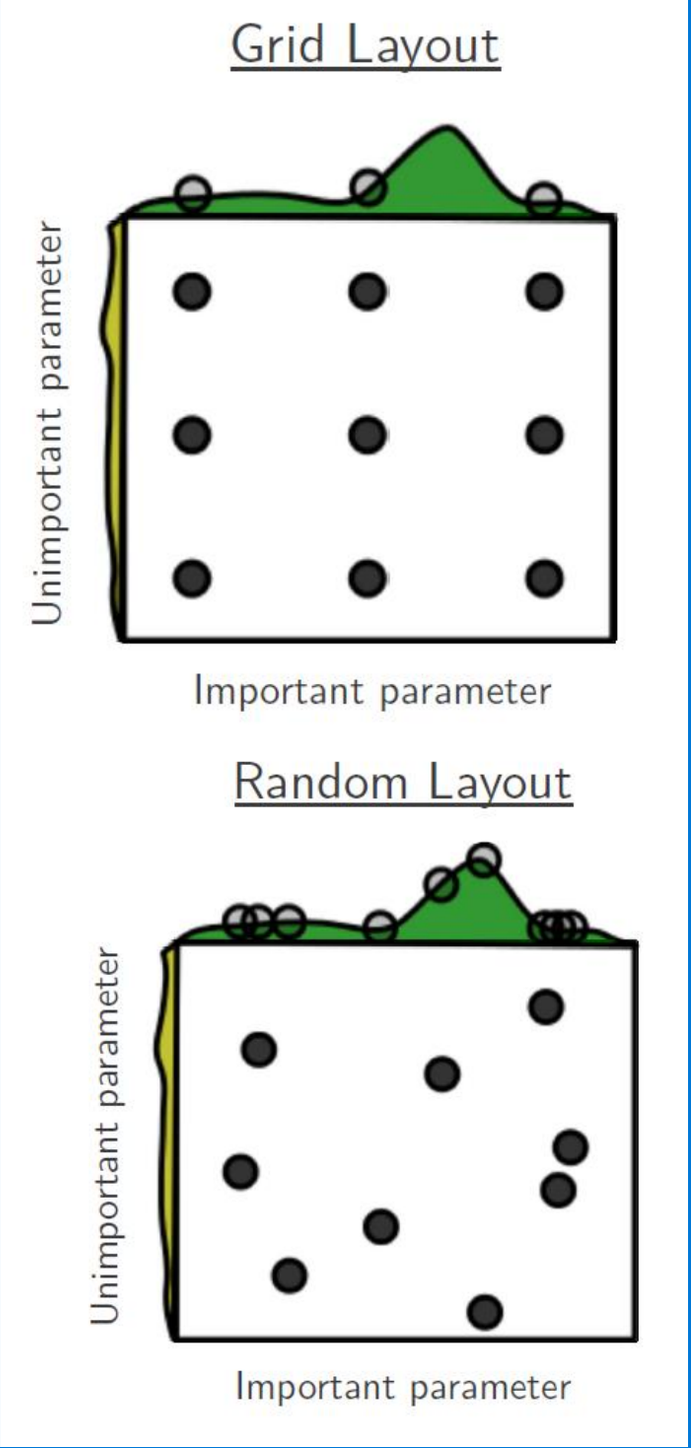
Create a grid over all parameters

At the same computational cost, selecting parameters randomly is better than specific grid

Credits: Random Search for Hyper-Parameter Optimization, Bergstra and Bengio JMLR2012



Choosing 59 random points from parameter grid will guarantee with 95% confidence that top 5% of the accuracy will be obtained.



Use Hyper Parameter Tuning


Are there tips and tricks that enable me to get to a good initial baseline?

enrollment_id	username	course_id	time	source	event
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:38:29	server	nagivate
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:38:39	server	access
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:38:39	server	access
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:38:48	server	access
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:41:49	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:41:50	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:42:28	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:42:30	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:43:20	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:43:25	browser	problem
1	9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	2014-06-14T09:43:25	server	problem

KDD Cup 2015

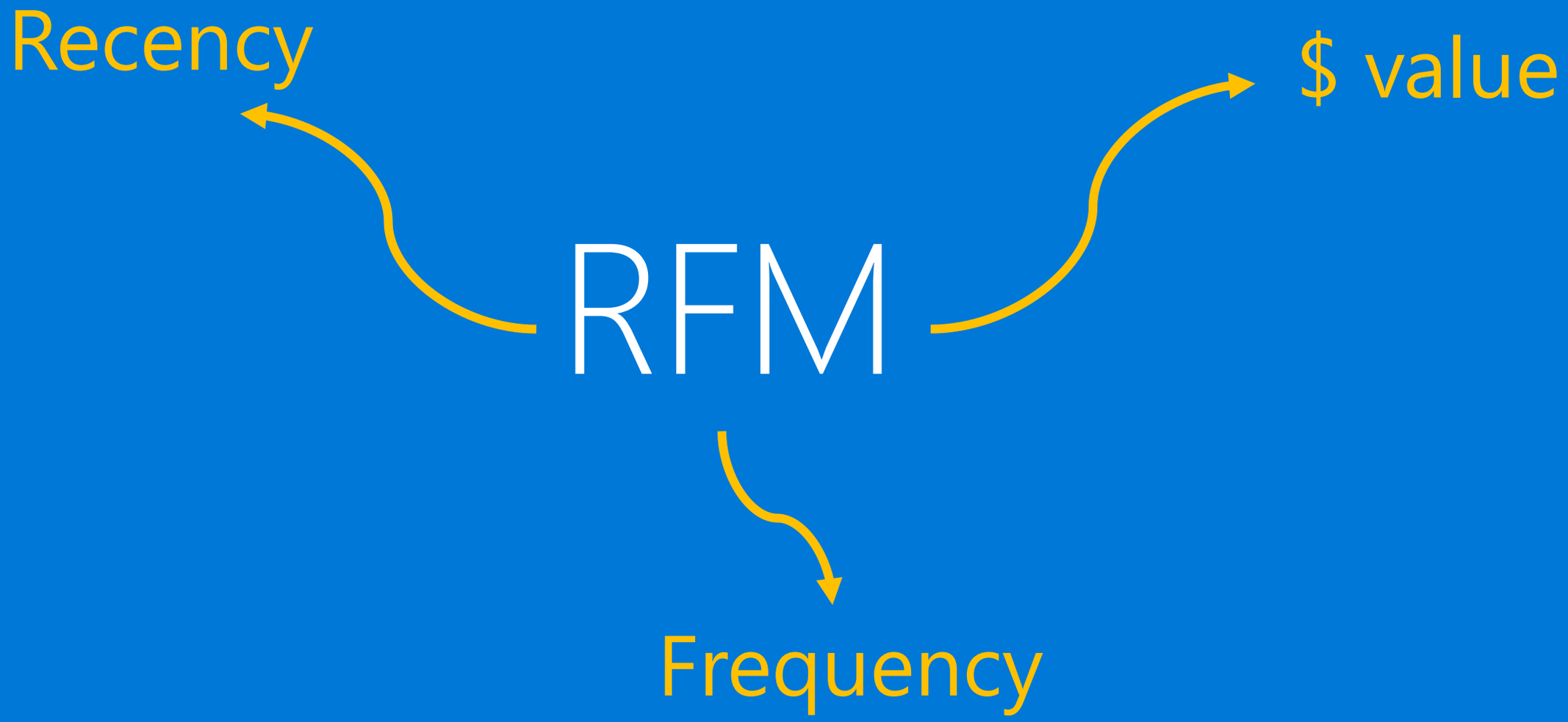
Student logs in online university courses

8M rows
7 columns

Goal
Predict student churn

Challenge 1 – How do we build an initial experiment that will be within the 1% accuracy of the winning solution?

Challenge 2 – How do we build an initial experiment that is within the top 30 (on the leaderboard?)


Recency
$ value
RFM
Frequency

Have the student
Attended class recently?

Enrollment ID's Last timestamp.

How many hours have the student spent on the course?

Number of unique hours on which the enrollment ID has had an eve

Count (events) Group By Enrollment ID.

How many problems have the student solved?
How many videos have the student watched?

800+ entries

Winner – AUC 90.9

- Merger of 9 teams
- ~250 iterations

RFM solution – AUC 90.1

- Quick configuration
- 26th place

#	Δrank	team name	score	Entries	Last Submission UTC
1	—	Intercontinental Ensemble	0.9091817339587759	284	12 Jul 2015 20:27:17
2	—	FEG&NSSOL@DataVeraci	0.9088631699682458	286	12 Jul 2015 22:51:36
3	—	CLMS	0.9085724159294324	272	12 Jul 2015 12:56:45
4	—	Data Sapiens	0.9079240957270056	95	12 Jul 2015 16:23:12
5	—	FirstTimeEver	0.907797205557665	123	12 Jul 2015 21:45:51
6	—	KDDILABS&Keiku	0.9077939997455148	272	12 Jul 2015 19:58:52
7	—	ttllbb	0.9077348384507644	131	11 Jul 2015 03:11:06
8	—	xiaochuan	0.907345041689206	38	11 Jul 2015 19:08:13
9	—	Donquote	0.9071736582217601	2	11 Jul 2015 19:04:25
10	—	NLP Logix	0.9070665859706881	33	12 Jul 2015 21:22:10
11	—	NCCU	0.9066275471846055	143	12 Jul 2015 19:16:38
12	—	kyazuki&DT@Keio univ. Ohmori Lab	0.9060999005005896	216	12 Jul 2015 23:13:24
13	—	Core	0.9060593629122135	204	12 Jul 2015 17:28:00
14	—	PLiu	0.9054985257613515	40	12 Jul 2015 23:13:06
15	—	^_^	0.9052381856935979	69	11 Jul 2015 13:00:49
16	—	marugari	0.9044325201064038	12	12 Jul 2015 21:37:43
17	—	xja31415	0.9044135851925345	77	12 Jul 2015 21:44:20
18	—	orange	0.904112231351442	192	04 Jul 2015 01:19:02
19	—	Opera.LCCL	0.9036351896308003	62	12 Jul 2015 22:33:11
20	—	beader	0.9035907863204043	29	12 Jul 2015 05:22:16
21	—	floydsoft	0.9035833604596284	31	12 Jul 2015 22:03:28
22	—	Carl.Hwang	0.9025204837523291	34	12 Jul 2015 19:43:10
23	—	Shi Xiaolin	0.9017338280687623	12	11 Jul 2015 17:05:14
24	—	Miner	0.9015676226440993	133	09 Jul 2015 01:09:21
26	—	NoSmartsJustTools	0.9006523107823934	25	30 Jun 2015 02:00:34

Try it - https://gallery.cortanaintelligence.com/Experiment/Auto-featurization-Churn-Prediction-on-KDDCup2015-Dataset-1

RFM

- Churn Prediction
- Segmentation
- Targeted Advertisements & Personalization

Building My Toolbox

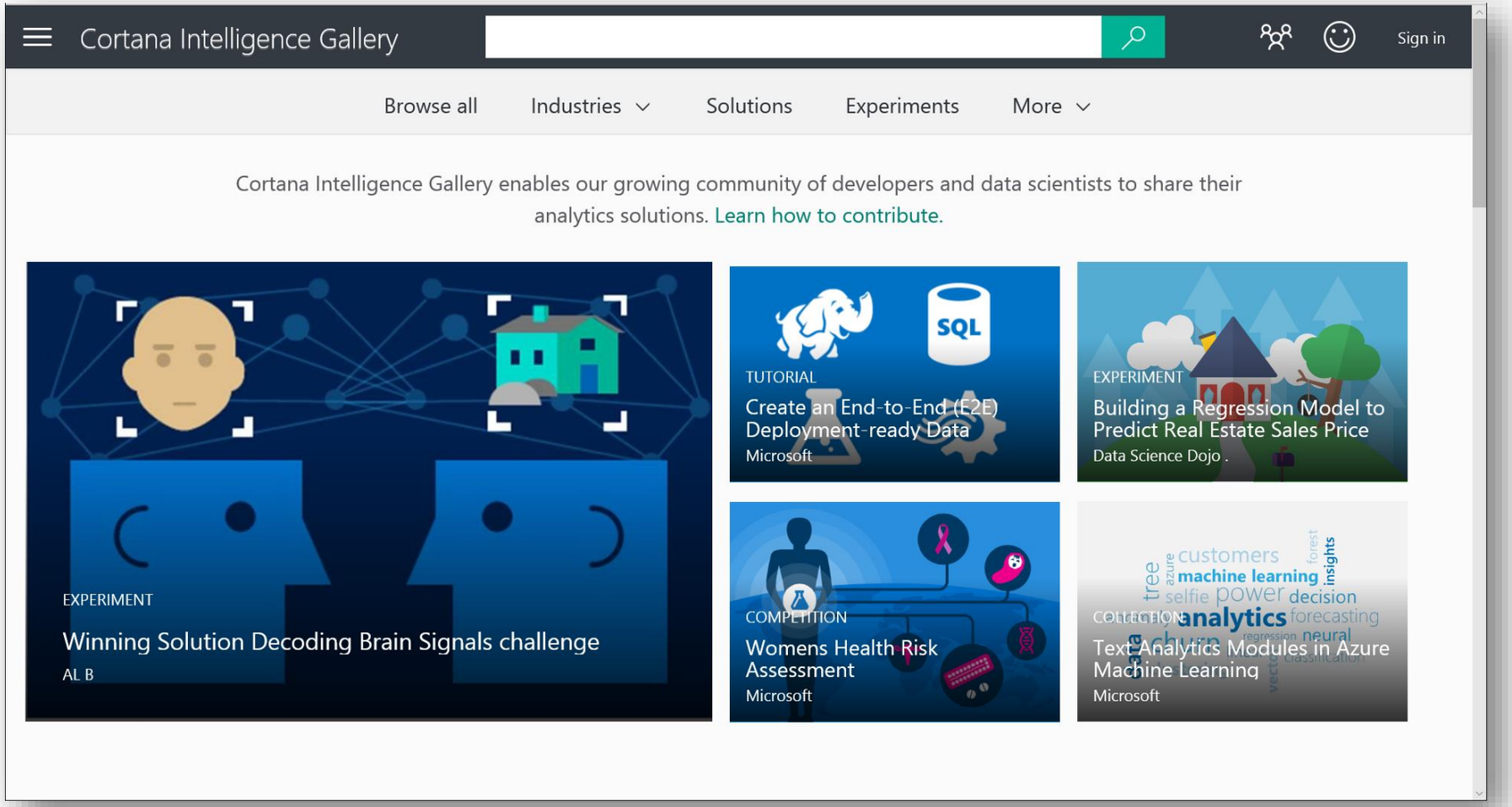
- RFM – User Behavior Modeling
- Handling Minority classes
- Hyper parameter tuning
- Auto Featurization
  Note: Domain expertise is still helpful

Can I auto-generate a machine learning model?

Super Power #4

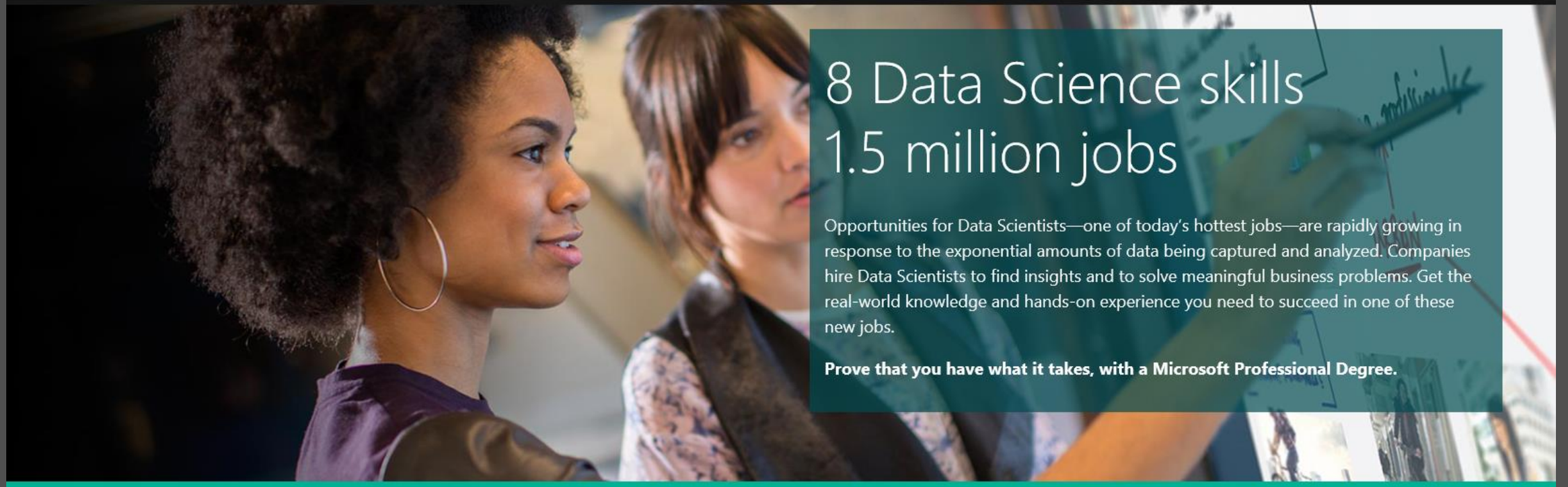
Learn from the Experts, and Internalize

Community



https://gallery.cortanaintelligence.com/

Microsoft



8 Data Science skills 1.5 million jobs

Opportunities for Data Scientists—one of today's hottest jobs—are rapidly growing in response to the exponential amounts of data being captured and analyzed. Companies hire Data Scientists to find insights and to solve meaningful business problems. Get the real-world knowledge and hands-on experience you need to succeed in one of these new jobs.

Prove that you have what it takes, with a Microsoft Professional Degree.

BETA RELEASE

Microsoft Professional Degree in Data Science

Microsoft consulted Data Scientists and the companies that employ them to identify the requisite core skills. We then developed a curriculum to teach these functional and technical skills, combining highly rated online courses with hands-on labs, concluding in a final capstone project. Graduates earn a Microsoft Professional Degree in Data Science—a digitally sharable, résumé-worthy credential.

The program opens soon! To be notified, submit your email address.

Enter your email address

Submit

https://academy.microsoft.com/en-US/professional-degree/data-science/

Data Science and Machine Learning Essentials

Learn key concepts of data science and machine learning with examples on how to build a cloud data science solution with R, Python and Azure Machine Learning from the Cortana Analytics Suite.

Microsoft

Self-Paced

Enroll Now

I would like to receive email from Microsoft and learn about its other programs.

About this course

45 Reviews 4/5

Demand for Data science talent is exploding. Learn these essentials with experts from MIT and the industry, partnering with Microsoft to help develop your career as a data scientist. By the end of this course, you will know how to build and derive insights from data science and machine learning models. You will learn key concepts in data acquisition, preparation, exploration and visualization

See more

What you'll learn

- The data science process
- Overview of data science theory
- Data acquisition, ingestion, sampling, quantization, cleaning and transformation
- Building data science workflows with Azure ML

Length:	5 weeks
Effort:	3 - 4 hours/week
Price:	FREE Add a Verified Certificate for $49
Institution:	Microsoft
Subject:	Computer Science
Level:	Intermediate
Languages:	English
Video Transcripts:	English

Super Power #5

Make Insights Actionable

But there are more ways to make insights actionable!

From Data to Intelligent Action!

What is your Super Power?

Microsoft