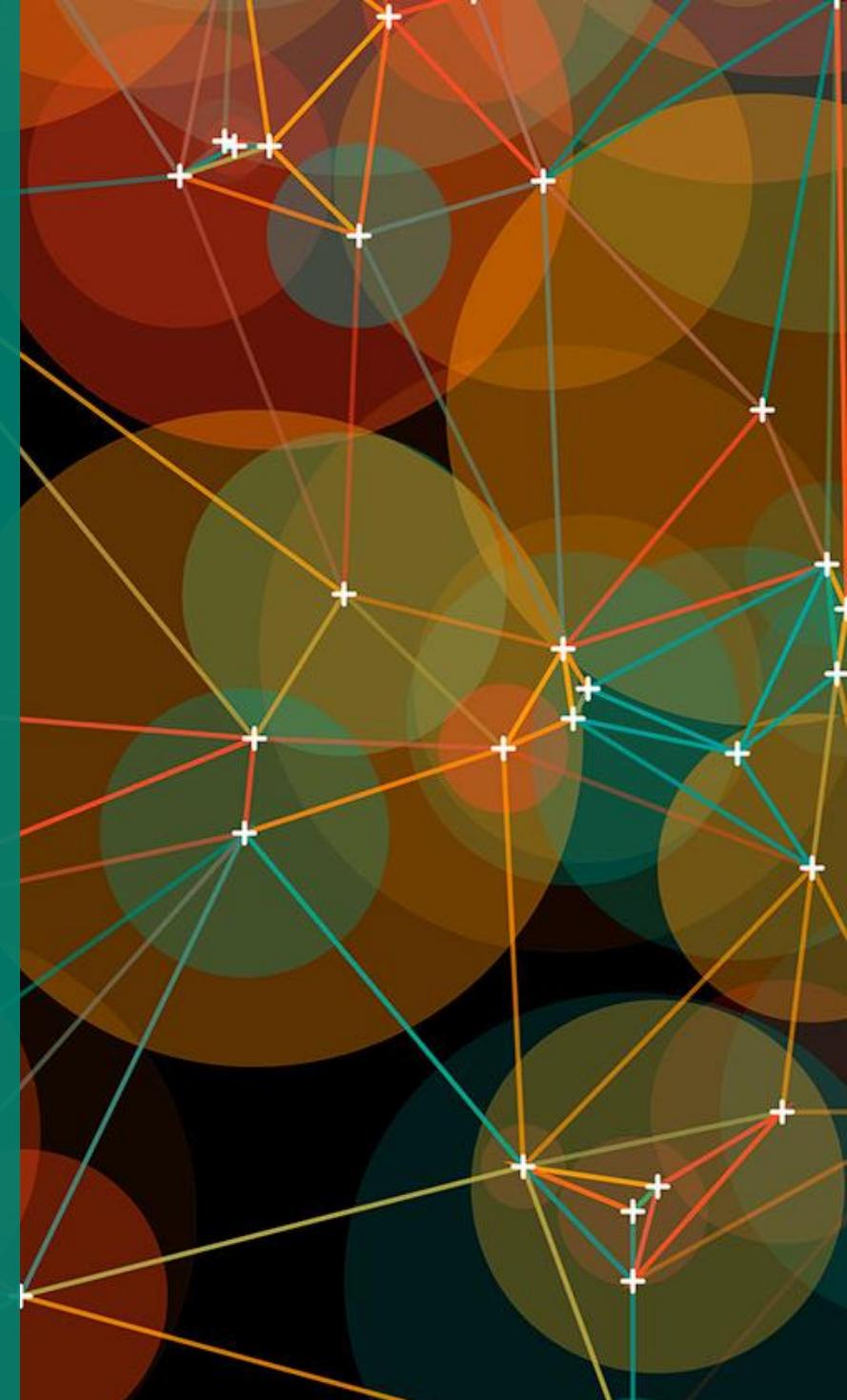


# *Discover your Data Science Super Power*

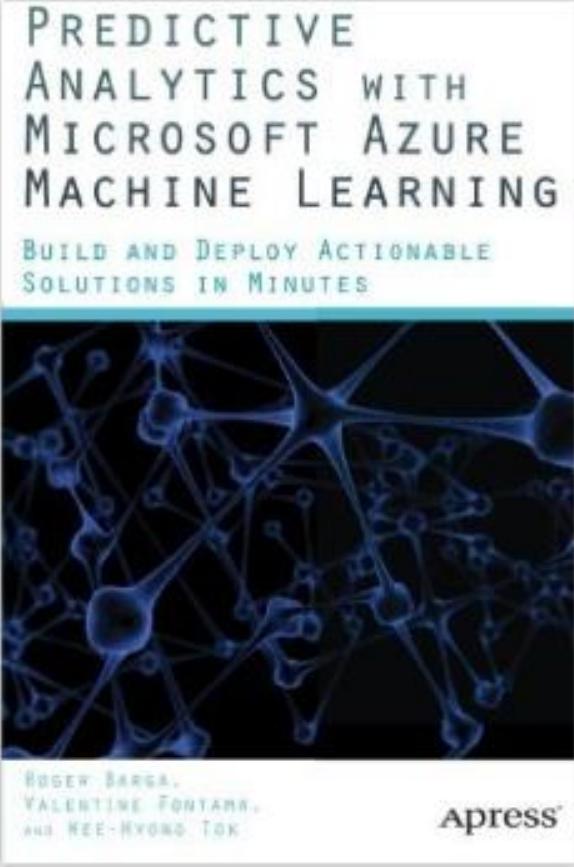
Tips and Tricks to a Successful Data Science Project

Wee Hyong Tok  
Algorithms and Data Science  
Microsoft

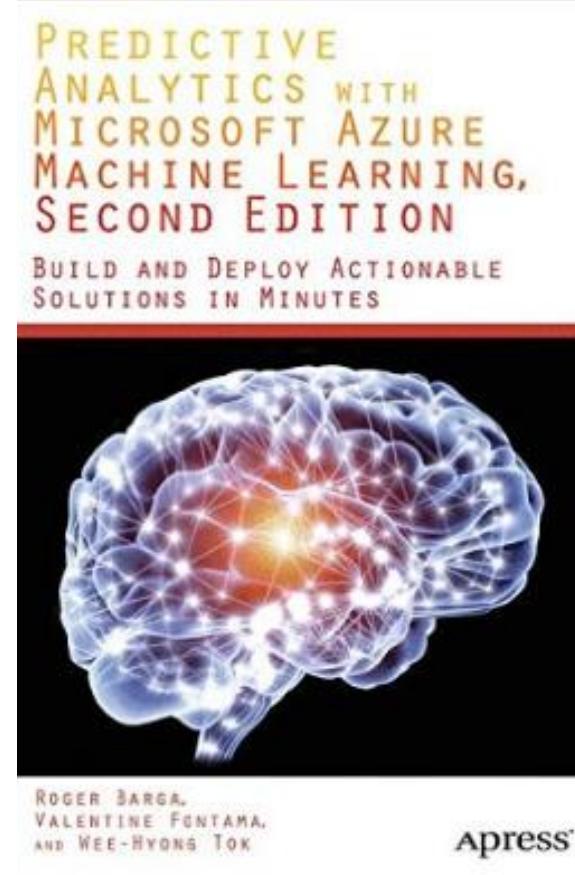
 @weehyong



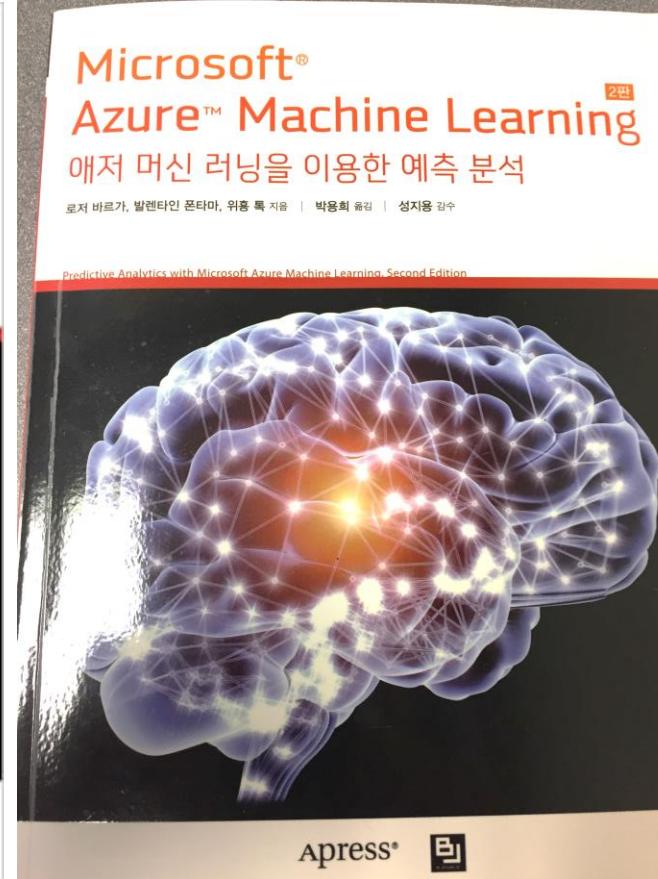
# My Data Science Journey



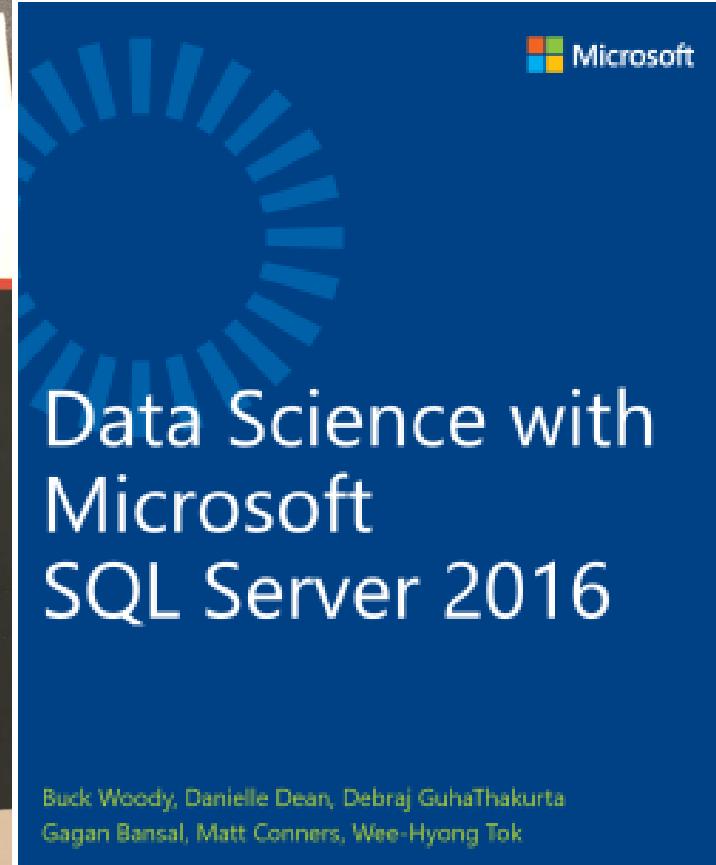
2014

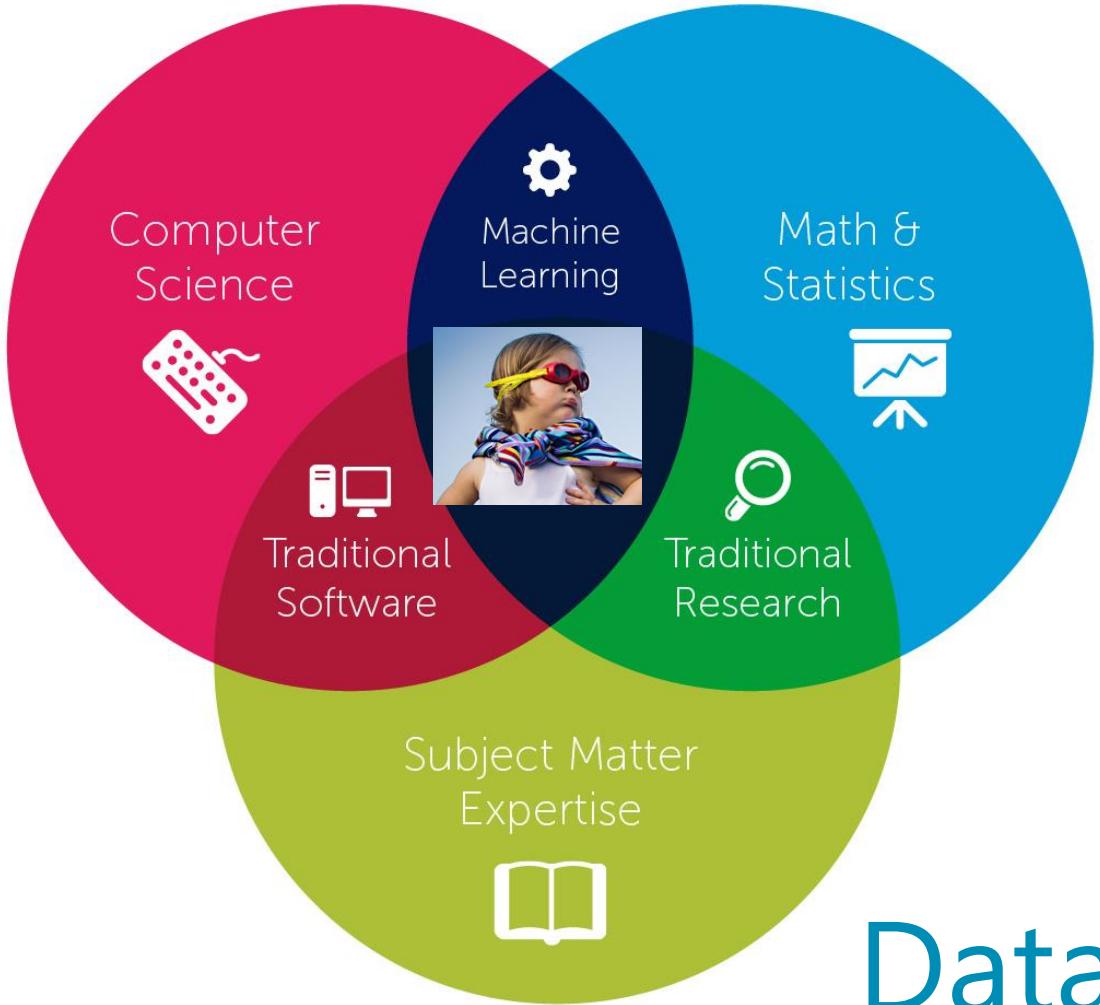


2015



2016





# Data Science

Copyright © 2014 by Steven Geringer Raleigh, NC.  
Permission is granted to use, distribute, or modify this image,  
provided that this copyright notice remains intact.

# 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...



I am a Machine  
Learning  
Researcher

I work on  
Deep  
Learning

I am a Data  
Scientist

We work on  
Text Analytics

# What are your Super Powers?



# Super Power #1

## Be Obsessed with Data



The screenshot shows a user interface for classifying galaxies. At the top, there is a navigation bar with links: Explore, Guided Tours, Search, Upload, Classification (which is highlighted in blue), View, Settings, and a user profile for 'Wee Hyong Tok'. Below the navigation bar, the title 'Unclassified Galaxies' is displayed. Underneath the title, there is a grid of nine small square images, each showing a different galaxy or celestial object. To the right of the grid, a dropdown menu is open, listing three machine learning models: Random Forest, Gradient Boosted Machine, and Generalized Linear Model. A button labeled 'Show Classes' is also visible next to the dropdown.

# Helping Astronomers Classify Galaxies

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

Rolls-Royce Demonstrator

rolls-royce.azurewebsites.net/#/enginedetails

Wee Hyong  
Engineering Supervisor

### Engine Details

|                           |             |
|---------------------------|-------------|
| TAIL NUMBER               | 7INTG       |
| TYPE                      | Airbus A350 |
| ENGINE                    | TRENT XWB   |
| SERIAL#                   | 21001       |
| LAST SERVICE DATE         | 02/15/2016  |
| CYCLES SINCE LAST SERVICE | 3,100       |
| NEXT WASH DATE            | 08/03/2016  |

### Performance

92%

### Aircraft Service Notes

| DATE | TECHNICIAN    | LOC | SERVICE                |
|------|---------------|-----|------------------------|
| 2015 | Joe Healy     | FRA | Full Inspection        |
| 2015 | Chen Yang     | LHR | Engine Wash            |
| 2015 | Lola Jacobsen | FRA | Landing Gear Repair    |
| 2015 | Katie Jordan  | DTW | Door Seal Inspection   |
| 2015 | Hamish Hill   | LGA | Full Inspection        |
| 2015 | Jiri Karpeta  | FRA | Bleed Air Valve Repair |

### Engine Overview

TRENT XWB

### Engine Systems

| STATUS  | ATA CODE | COMPONENT         | RUL |
|---------|----------|-------------------|-----|
| WARNING | ATA 28   | PRIMARY FUEL PUMP |     |
| OK      | ATA 71   | POWER PLANT       |     |
| OK      | ATA 74   | IGNITION          |     |
| OK      | ATA 75   | BLEED AIR SYSTEM  |     |
| OK      | ATA 76   | ENGINE CONTROLS   |     |
| OK      | ATA 78   | EXHAUST           |     |

### Normalized Fuel Burn

Sum of Optimization(%)

Hours

Threshold

Fuel

### Engine Wash Optimization

Sum of Optimization(%)

Hours

### Engine Lead Indicators

TGT Margin

88%

RUL

52%

# Predicting Remaining Useful Lifetime (RUL) of Engine Systems



# New Generation of Smart Refrigerators



August 10, 2016

Is it possible to predict whether students are at risk of dropping out of school? The Tacoma Public School district thinks so. Using predictive analytics tools based on Microsoft cloud technologies, the district is providing comprehensive data snapshots of student success indicators and has already helped to improve graduation rates from 55 to 82.6 percent.

**Customer**  
Tacoma Public Schools

**Account Website**  
[www.tacomaschools.org/Pages/default.aspx](http://www.tacomaschools.org/Pages/default.aspx)

*"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?





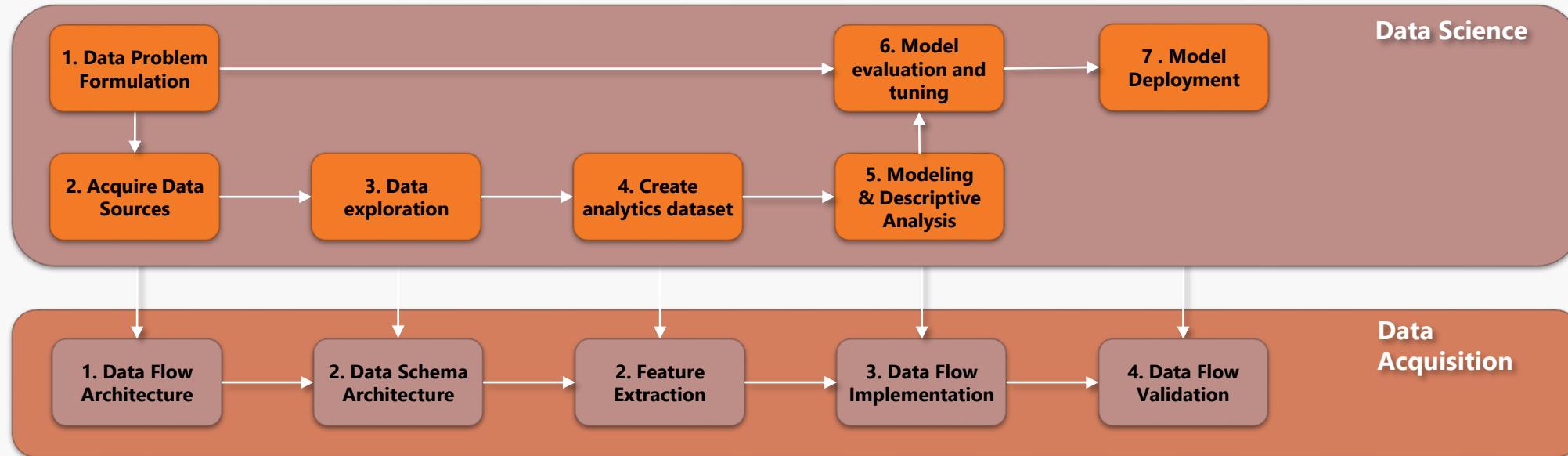
# Super Power #2

## A Linear Method for Non-Linear Work



# Data Science

**Goal:** Meet business goals through close collaboration with the business



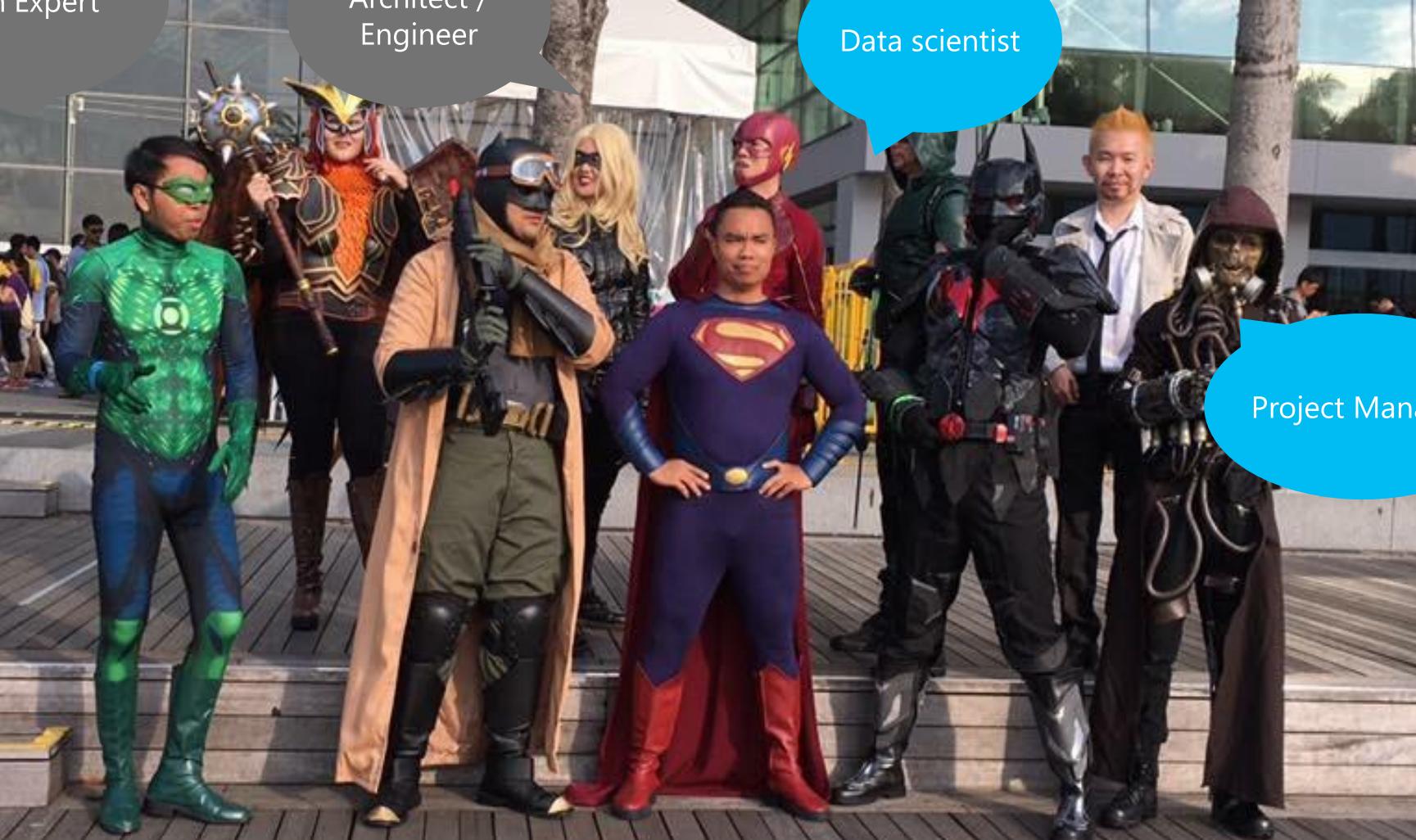
# Your process should include a team

Domain Expert

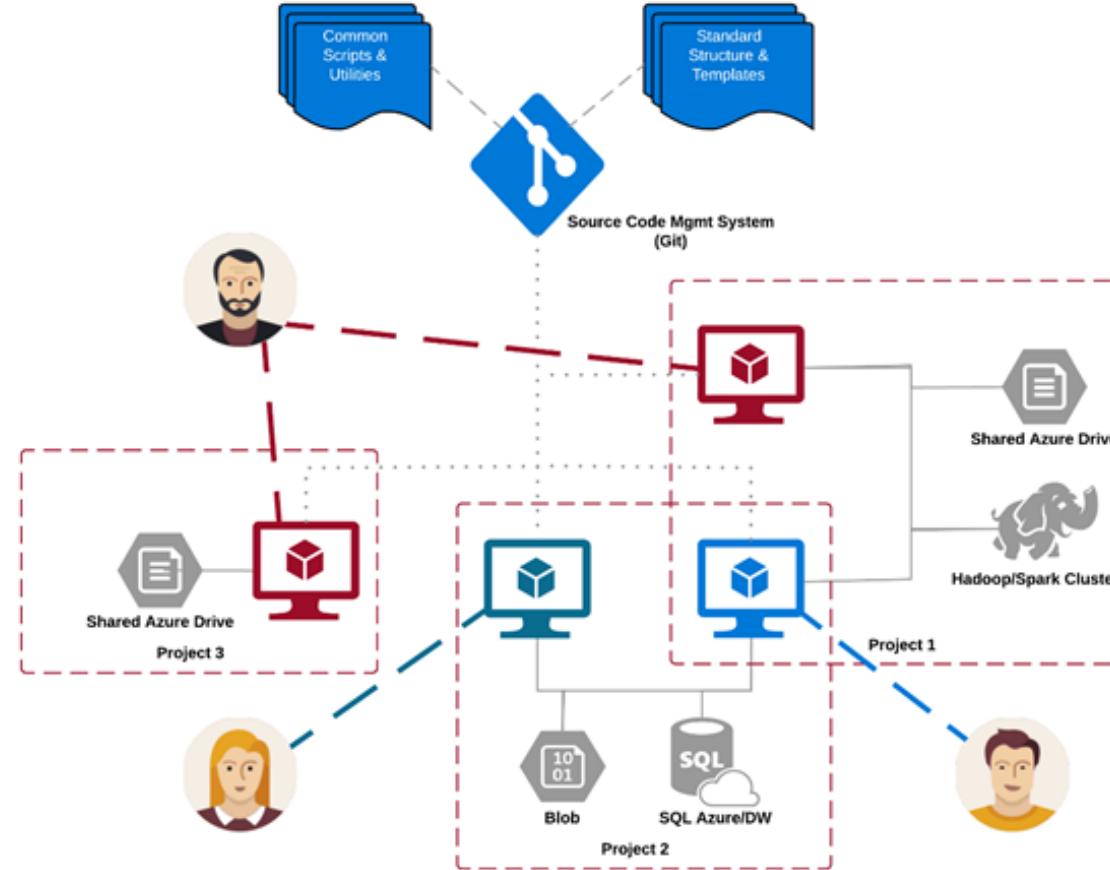
Architect /  
Engineer

Data scientist

Project Manager



# 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



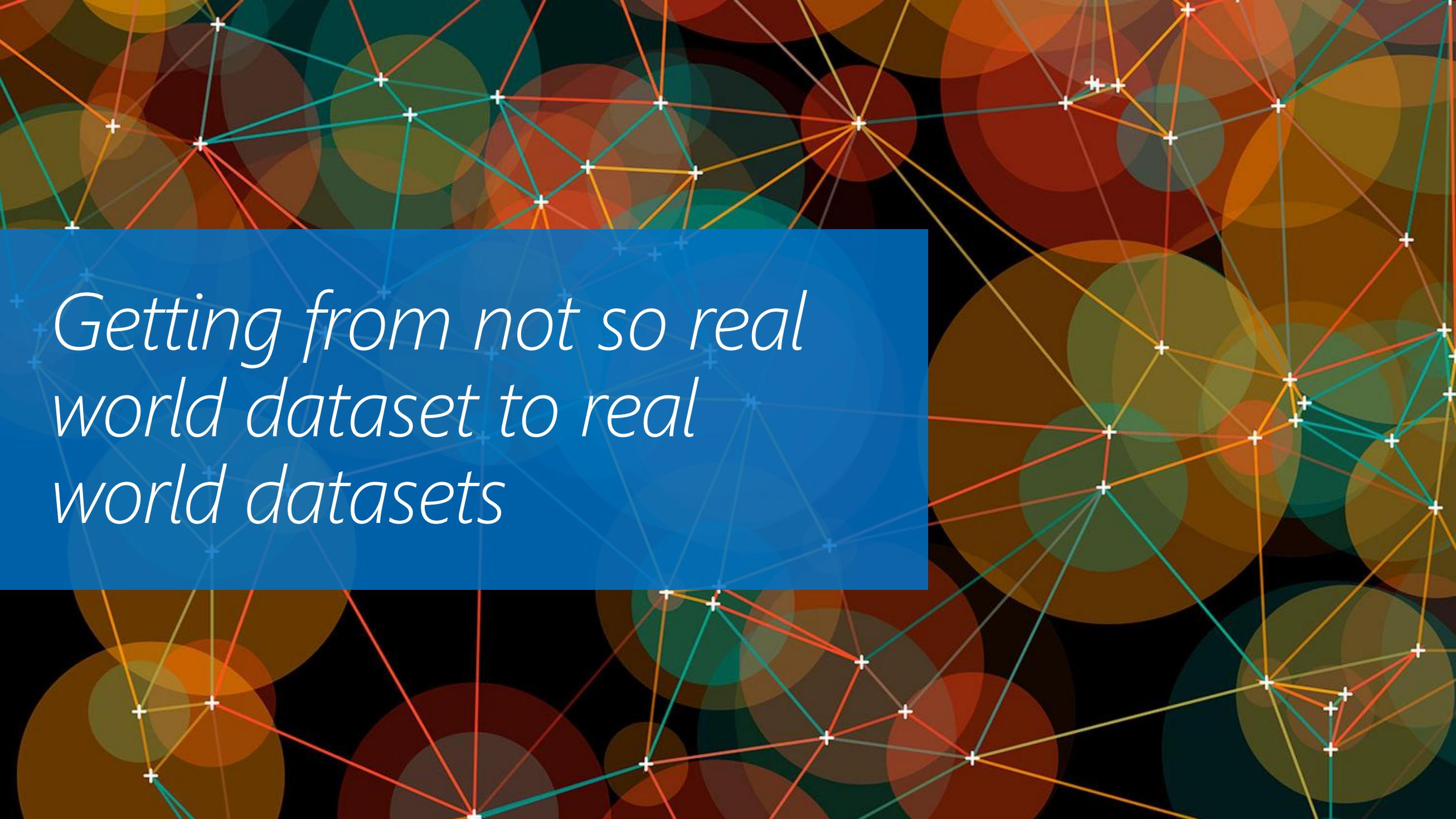
# Super Power #3

## 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.

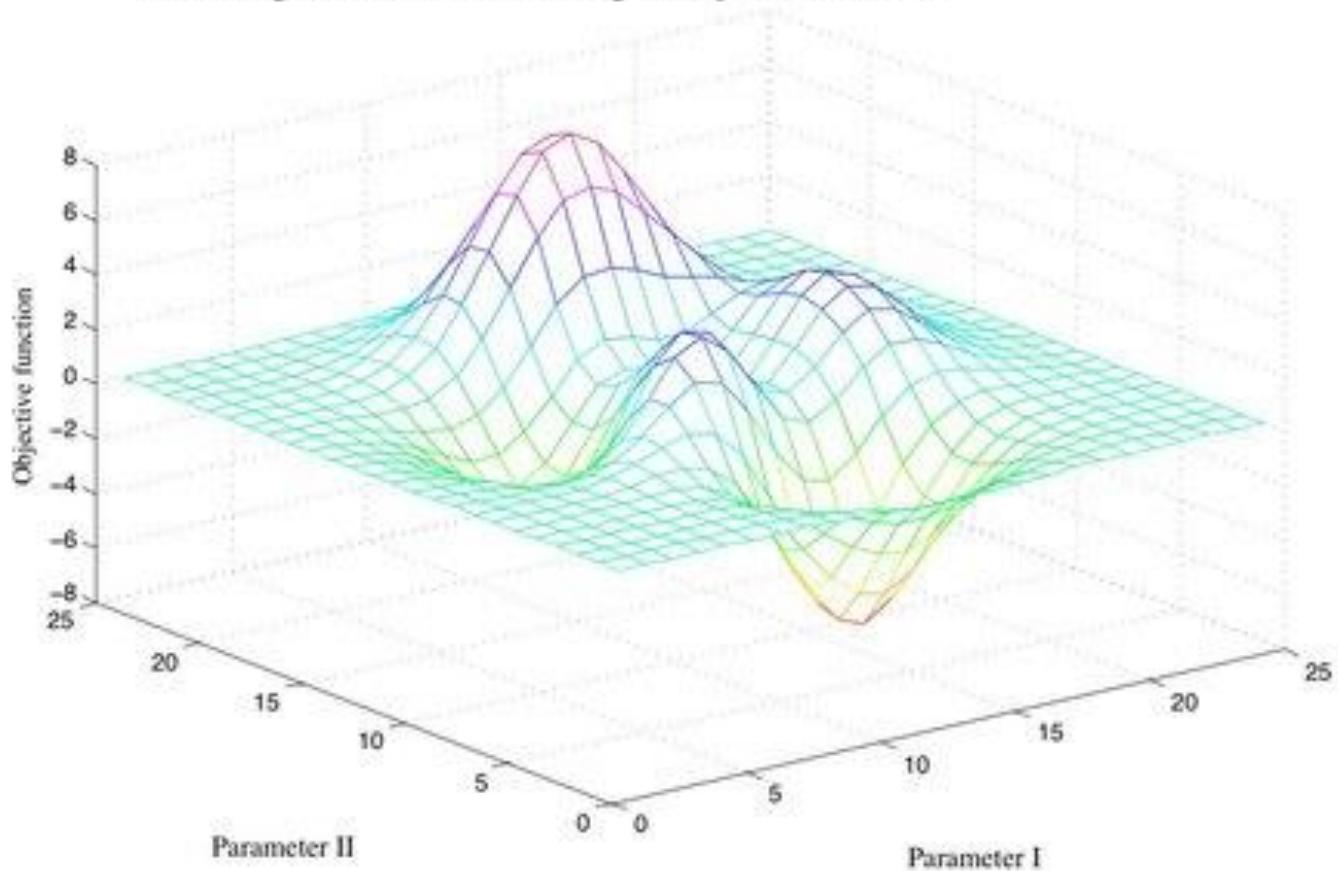
Minority class tend to be misclassified.

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

SMOTE - Synthetic Minority Oversampling Technique

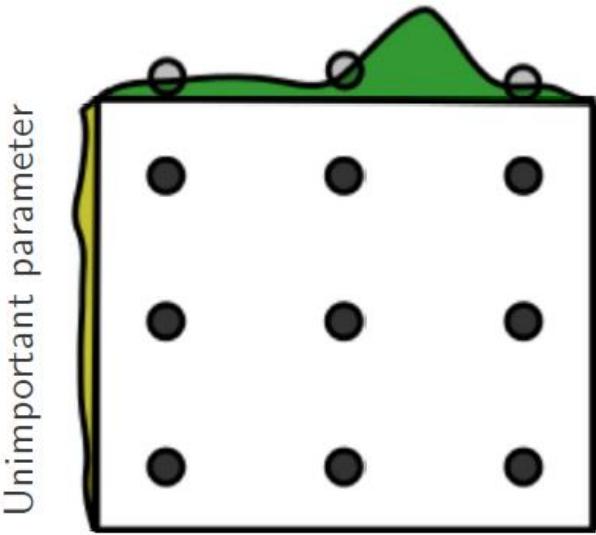
*Why am I doing so many iterations of trial and error?*

Random grid search in minimizing the objective function



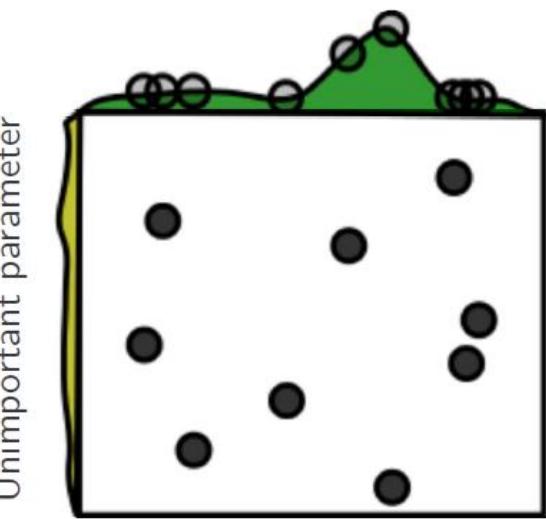
Simplest approach  
Create a grid over all parameters

## Grid Layout



Important parameter

## Random Layout

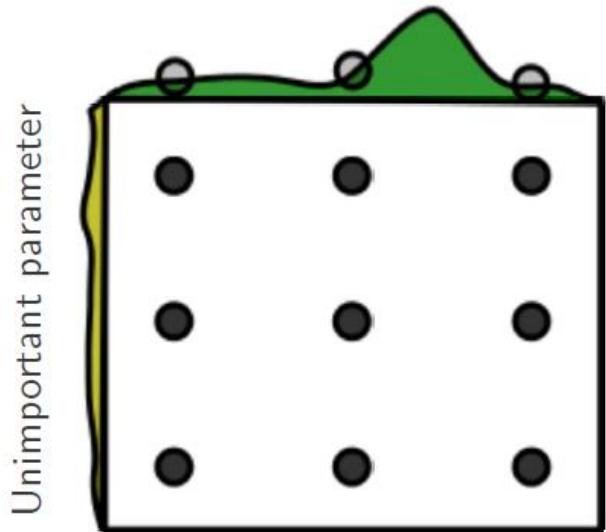


Important parameter

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

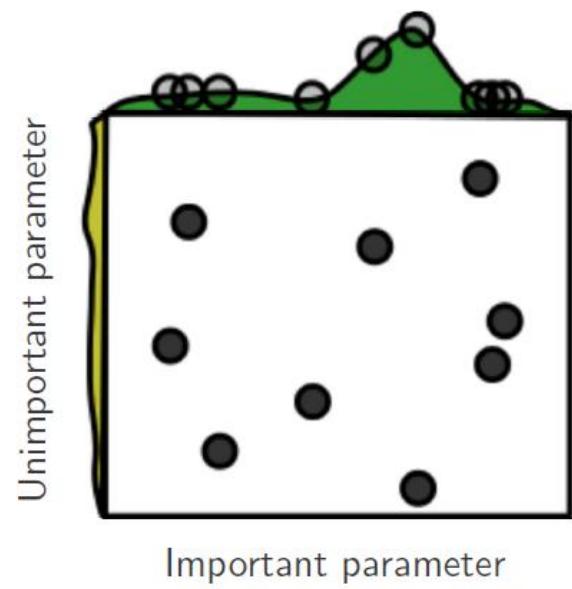
Credits: Random Search for Hyper-Parameter Optimization, Bergstra and Bengio JMLR2012

## Grid Layout

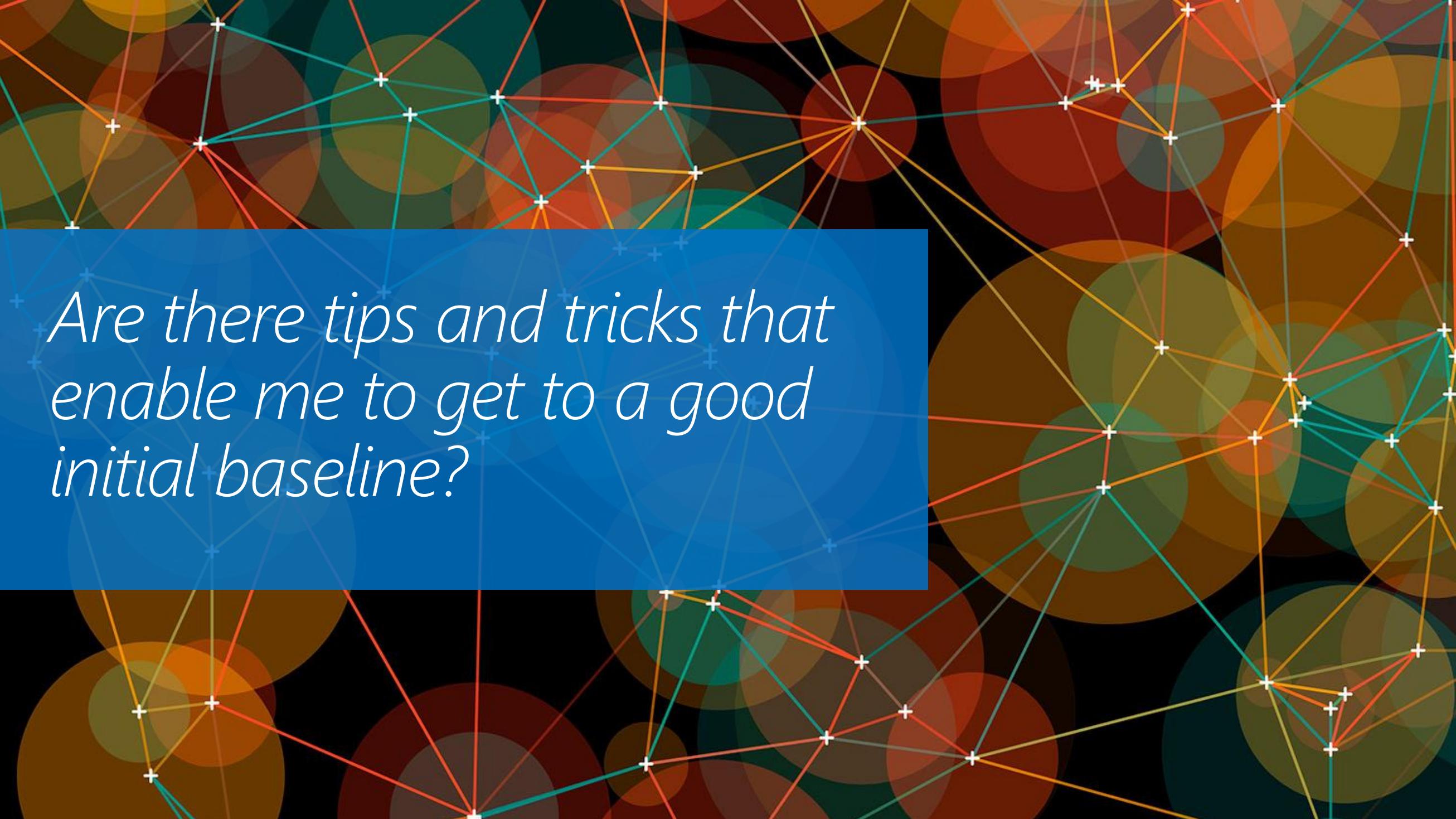


Choosing 59 random points from parameter grid will guarantee with 95% confidence that top 5% of the accuracy will be obtained.

## Random Layout



# 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             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:38:29 | server  | navigate |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:38:39 | server  | access   |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:38:39 | server  | access   |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:38:48 | server  | access   |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:41:49 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:41:50 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:42:28 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:42:30 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:43:20 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 2014-06-14T09:43:25 | browser | problem  |
| 1             | 9Uee7oEuuMmgPx2lzPfFk<br>WgkHZyPbWr0 | DPnLzkJJqOOPRJfBxIHbQE<br>ERiYHu5ila | 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

RFM



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
- 26<sup>th</sup> place

| #  | Rank | 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                    | 0.907793997455148  | 272     | 12 Jul 2015 19:58:52 |
| 7  | —    | ttlibb                            | 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                         | 0.9070665595706881 | 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 |
| 25 | —    | ...                               | 0.9005770000000001 | 76      | 11 Jul 2015 12:27:41 |
| 26 | —    | NoSmartsJustTools                 | 0.9006523107823934 | 25      | 30 Jun 2015 02:00:34 |
| 27 | —    | Yata                              | 0.8996550369900479 | 4       | 11 Jul 2015 16:08:20 |

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

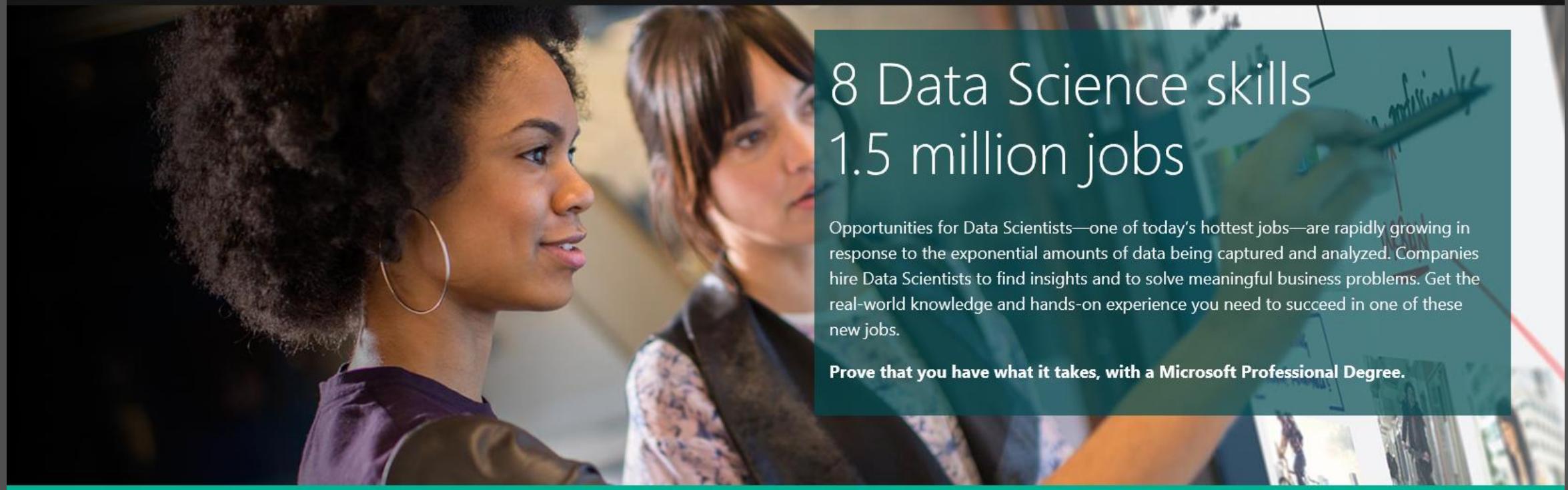


# Community

The screenshot shows the Cortana Intelligence Gallery homepage. At the top, there's a navigation bar with a menu icon, the title "Cortana Intelligence Gallery", a search bar, and user icons for sign-in and notifications. Below the navigation is a horizontal menu with links: "Browse all", "Industries", "Solutions", "Experiments", and "More". A main message states: "Cortana Intelligence Gallery enables our growing community of developers and data scientists to share their analytics solutions. [Learn how to contribute.](#)" Below this, there are six cards representing different solutions:

- EXPERIMENT**: Winning Solution Decoding Brain Signals challenge by AL B. It features an illustration of a brain and a house, with a network of nodes connecting them.
- TUTORIAL**: Create an End-to-End (E2E) Deployment-ready Data by Microsoft. It includes icons for an elephant, a SQL database, and a lab flask.
- EXPERIMENT**: Building a Regression Model to Predict Real Estate Sales Price by Data Science Dojo. It shows a house, a tree, and a city skyline.
- COMPETITION**: Womens Health Risk Assessment by Microsoft. It features a silhouette of a woman and medical icons.
- COLLECTION**: Text Analytics Modules in Azure Machine Learning by Microsoft. It displays a collage of words related to machine learning and text analysis.

<https://gallery.cortanaintelligence.com/>



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.

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.

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<br>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!

PublishCode  
CollabDashboard  
BlogService  
StoredProcedure

A photograph of several cosplayers in superhero costumes standing outdoors. From left to right: a person in a Green Lantern suit, a person in a Starfire costume, a person in a Cyborg costume holding a large gun-like prop, a person in a Hawkeye costume, a person in a Supergirl costume with hands on hips, a person in a Black Canary costume, a person in a Green Arrow costume, and a man in a white shirt and tie standing next to a person in a Star Wars stormtrooper costume.

I am obsessed  
with Data

I use a non-  
linear  
process

I have a  
toolbox of  
tricks

I make  
insights  
actionable

I learn from  
Experts

*What is your Super Power?*

