Microsoft

CodeMotion 2017 – Madrid, Spain

Overfitting:
Top 10 Machine Learning Gotchas

@DynamicWebPaige

Paige Bailey

Sr. Cloud Developer Advocate

Work Experience

- Focus at Microsoft is *machine learning* and *artificial intelligence*.
- Prior to joining Microsoft, was a *data scientist* and *geophysical application developer* in the energy industry for 5 years.
- *GIS Technician* (Esri products) for two years.

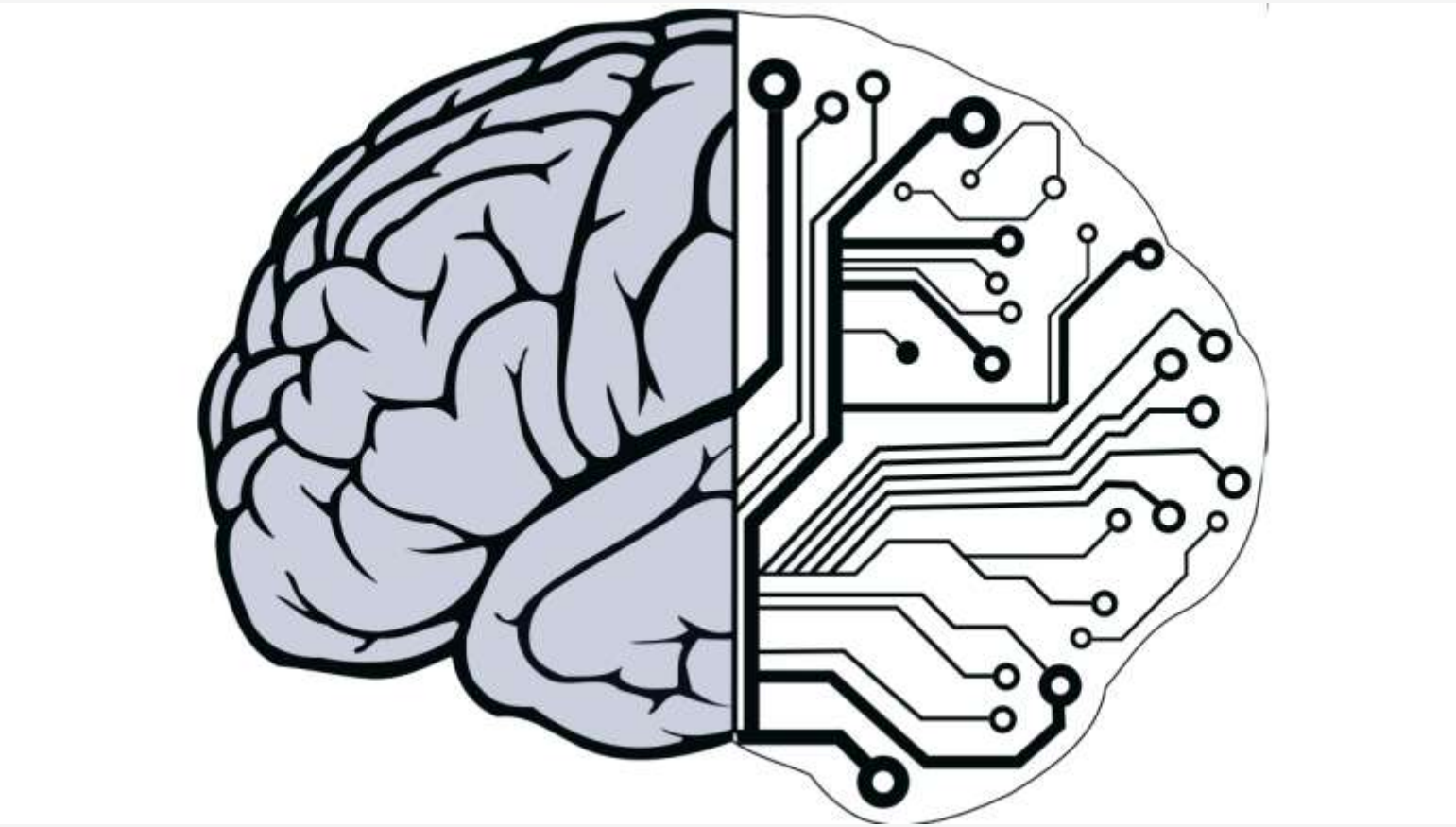
Toolkit

- Python *(10 years)*
- R *(4 years)*
- Spark, Kafka, Hive, HBase *(2 years)*

Location: *Austin, TX*

Twitter: *@DynamicWebPaige*

Microsoft | Data Platform



Microsoft

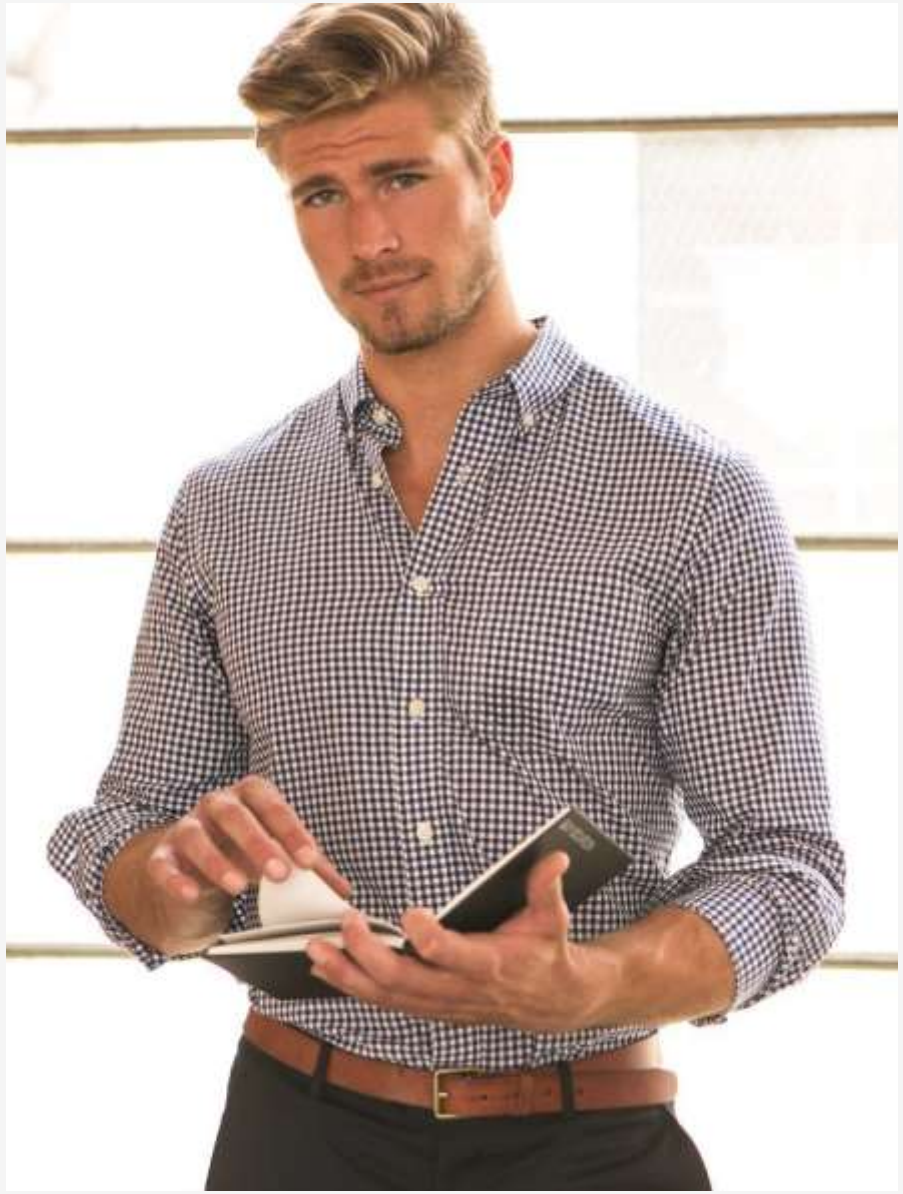
What is Machine Learning?

Microsoft

An example:

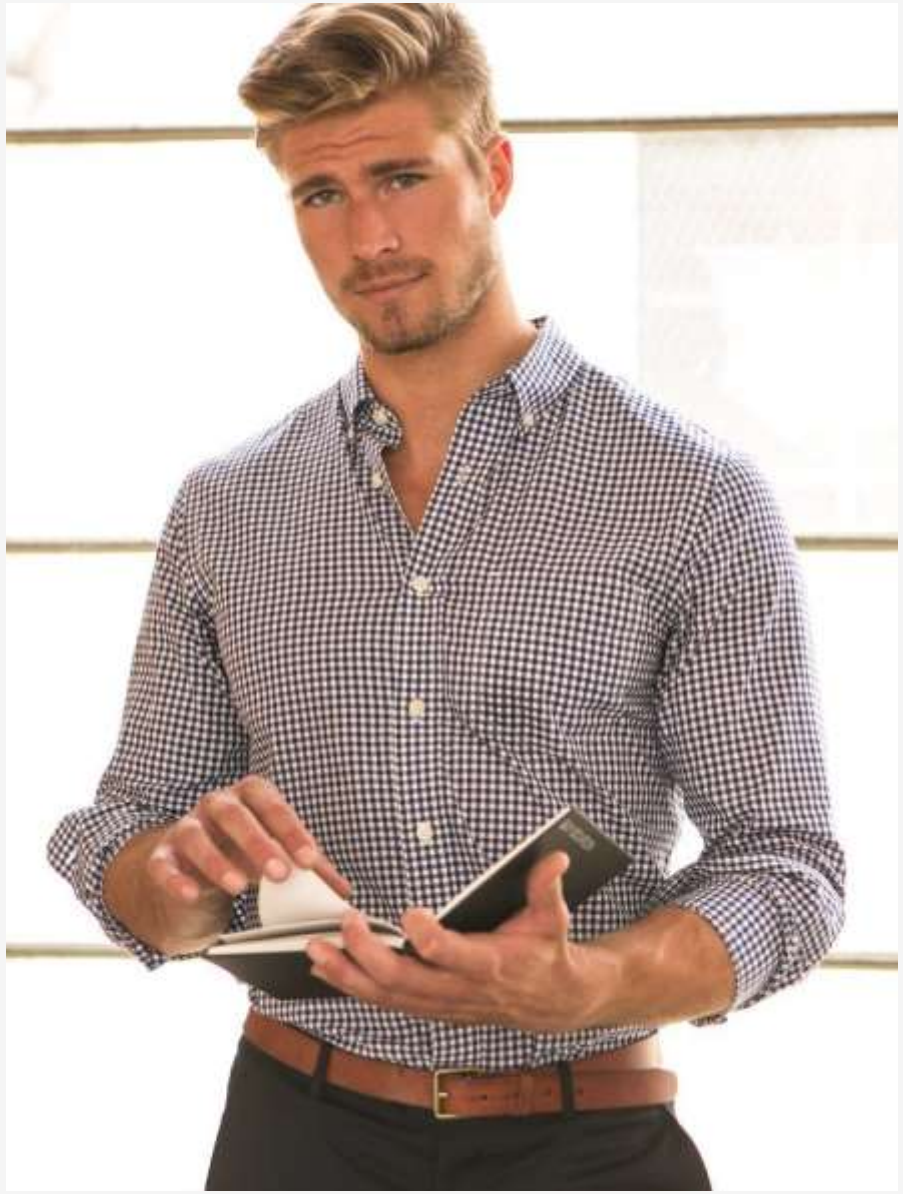
Credit Risk

Sven



We know some things about Sven:

- Age: 32
- Gender: M
- Credit Score: 1
- Homeowner: N
- Account balance: -150
- Employment status: Y
- Education: BS
- Current debt: 2000



Maria

Jeff

Maria

...and we have similar information about each of them.

Jeff

Name	Age	Gender	Credit Score	Homeowner	Checking Account	Employment Status	Education	Current Debt
Sven	32	M	1	N	-150	Y	BS	2000
Jeff	29	M	4	Y	700	Y	MS	250
Maria	42	F	5	Y	1500	Y	MBA	1500
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

RECORD

Name	Age	Gender	Credit Score	Homeowner	Checking Account	Employment Status	Education	Current Debt
Sven	32	M	1	N	-150	Y	BS	2000
Jeff	29	M	4	Y	700	Y	MS	250
Maria	42	F	5	Y	1500	Y	MBA	1500
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

RECORD

FEATURE

Name	Age	Gender	Credit Score	Homeowner	Checking Account	Employment Status	Education	Current Debt
Sven	32	M	1	N	-150	Y	BS	2000
Jeff	29	M	4	Y	700	Y	MS	250
Maria	42	F	5	Y	1500	Y	MBA	1500
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

RECORD

FEATURE

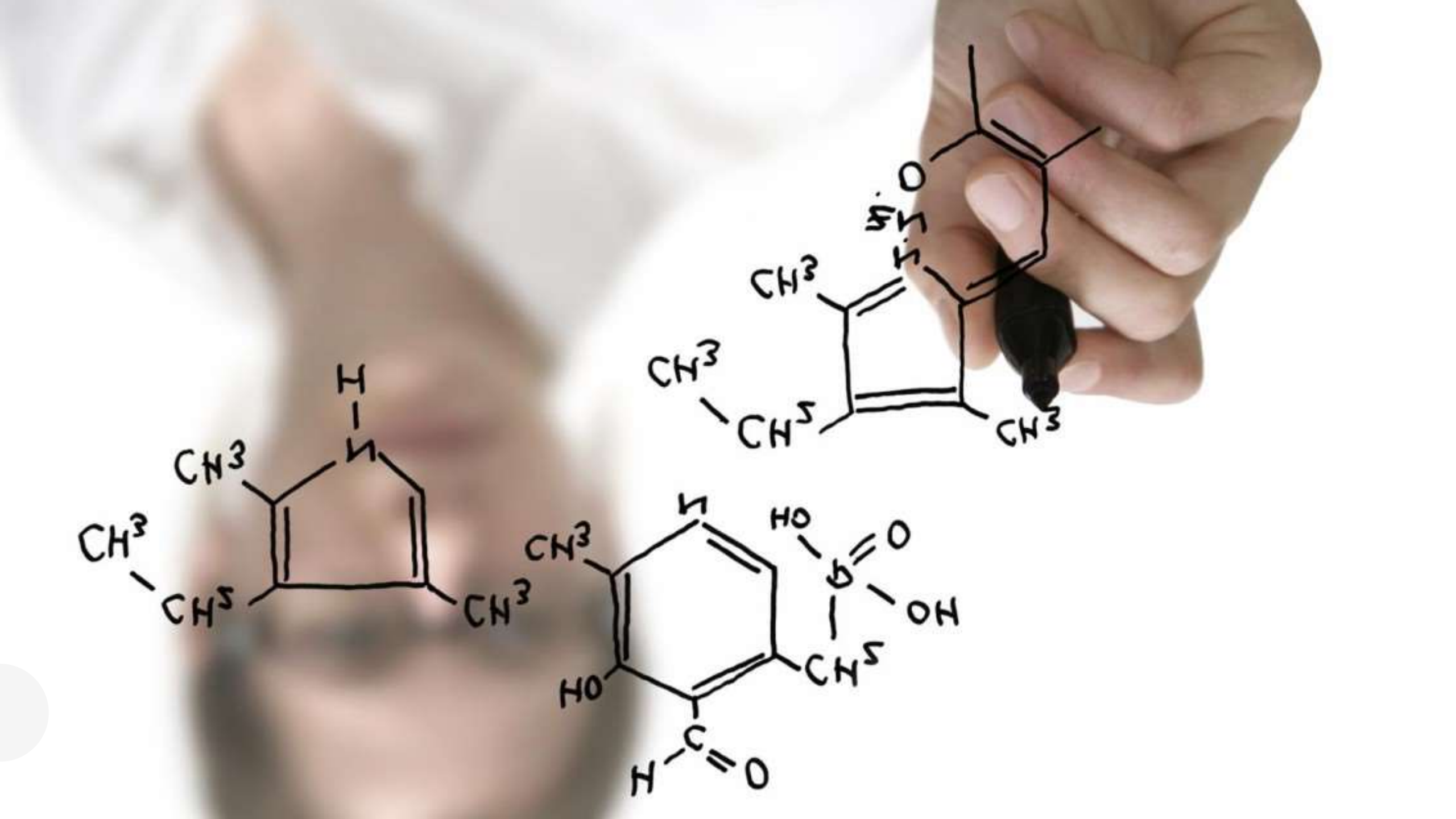
What does this look like in code?

Microsoft

Live Demo

…so what does that mean?

What is Overfitting?


O
H
CH3
CH3
CH5
CH3
H
CH3
CH3
CH5
CH3
H
CH3
HO
O
OH
CH5
HO
C
H
O

@DynamicWebPaige
@DynamicWebPaige

An analogy:

Remember those kids in school who were phenomenal at memorization, but could never apply those memorized concepts to a new problem? Every time they were shown a problem they hadn't seen before, their conceptual framework fell apart.

That's #overfitting.

2:28 AM - 22 Nov 2017

Microsoft

Live Demo

Microsoft

Top 10
Machine Learning Gotchas

(and to make it interesting…)

Microsoft

Top 10
Machine Learning Gotchas
(...let's talk about Real Madrid)

CodeMotion 2017 – Madrid

Microsoft

The owner of Real Madrid has heard about your legendary data science skills…

Microsoft

...and wants you to predict performance for the club in 2018.

Microsoft

Not just whether they'll win

Microsoft

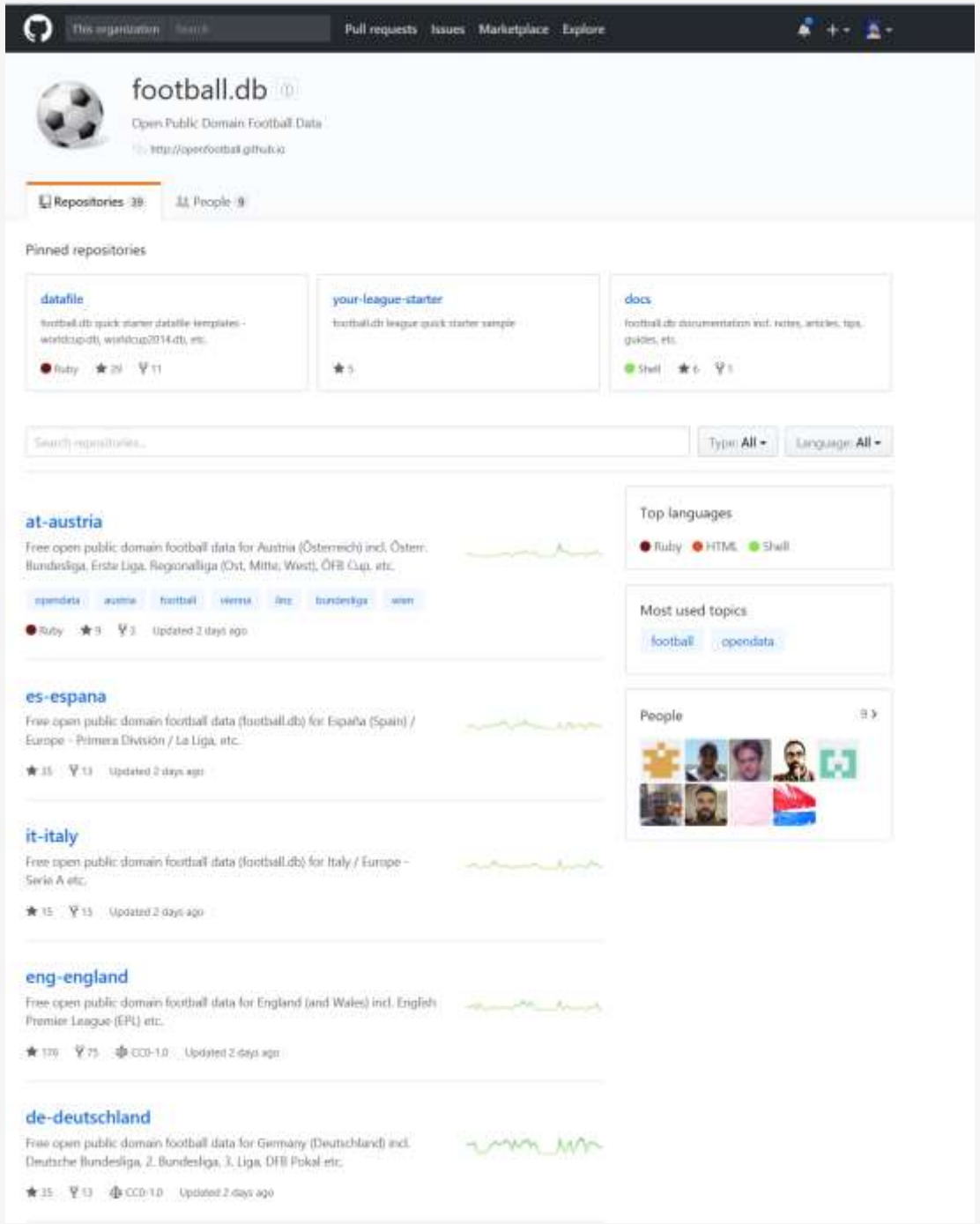
(obviously)

But…

- By how much?
- How will individual players perform?
- How many successful passes?
- How will players grow / deteriorate?
- Will there really be a home field advantage?

etc.

football.db

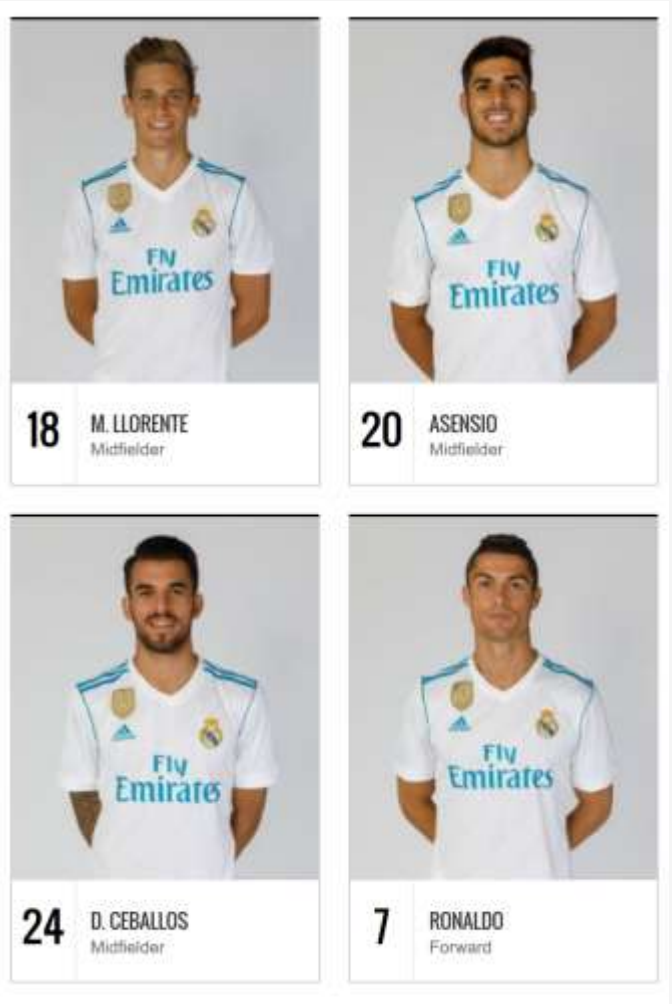

This organization
Pull requests
Issues
Marketplace
Explore
football.db
Open Public Domain Football Data
http://openfootball.github.io
Repositories 39
People 9
Pinned repositories
datafile
your-league-starter
docs
Type: All
Language: All
at-austria
Free open public domain football data for Austria (Österreich) incl. Österr. Bundesliga, Erste Liga, Regionalliga (Ost, Mitte, West), ÖFB Cup, etc.
Updated 2 days ago
es-espana
Free open public domain football data (football.db) for España (Spain) / Europe - Primera División / La Liga, etc.
Updated 2 days ago
it-italy
Free open public domain football data (football.db) for Italy / Europe - Serie A etc.
Updated 2 days ago
eng-england
Free open public domain football data for England (and Wales) incl. English Premier League (EPL) etc.
Updated 2 days ago
de-deutschland
Free open public domain football data for Germany (Deutschland) incl. Deutsche Bundesliga, 2. Bundesliga, 3. Liga, DFB Pokal etc.
Updated 2 days ago
Top languages
Ruby
HTML
Shell
Most used topics
football
opendata
People

Example 1:

Too Few Samples

only having data for a few games…

19	16 January 2011	Almería	1–1	Real Madrid	Almería
20	23 January 2011	Real Madrid	1–0	Mallorca	Madrid
21	30 January 2011	Osasuna	1–0	Real Madrid	Pamplona
22	6 February 2011	Real Madrid	4–1	Real Sociedad	Madrid
23	13 February 2011	Espanyol	0–1	Real Madrid	Cornellà de Llobregat
24	19 February 2011	Real Madrid	2–0	Levante	Madrid
25	26 February 2011	Deportivo La Coruña	0–0	Real Madrid	A Coruña
26	3 March 2011	Real Madrid	7–0	Málaga	Madrid

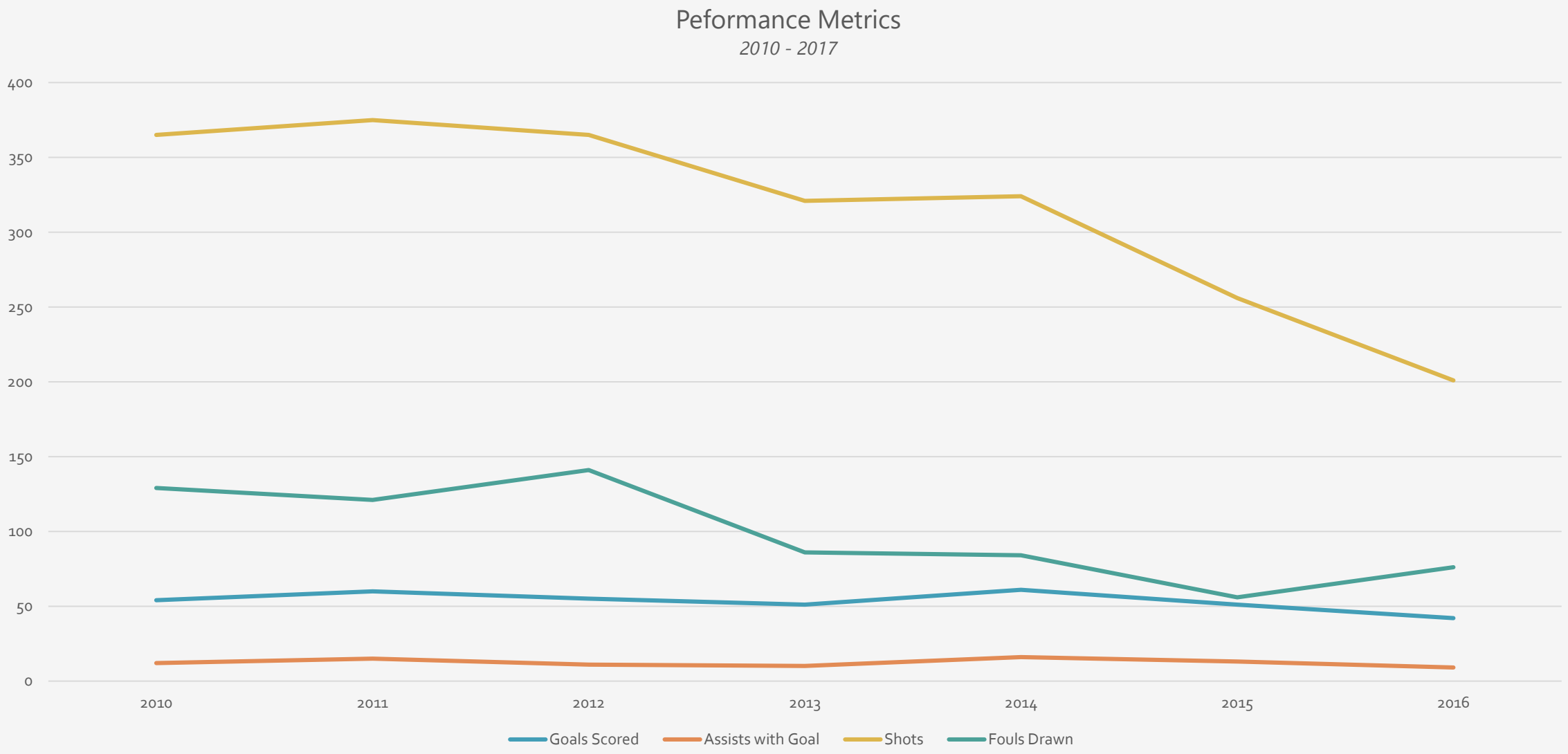


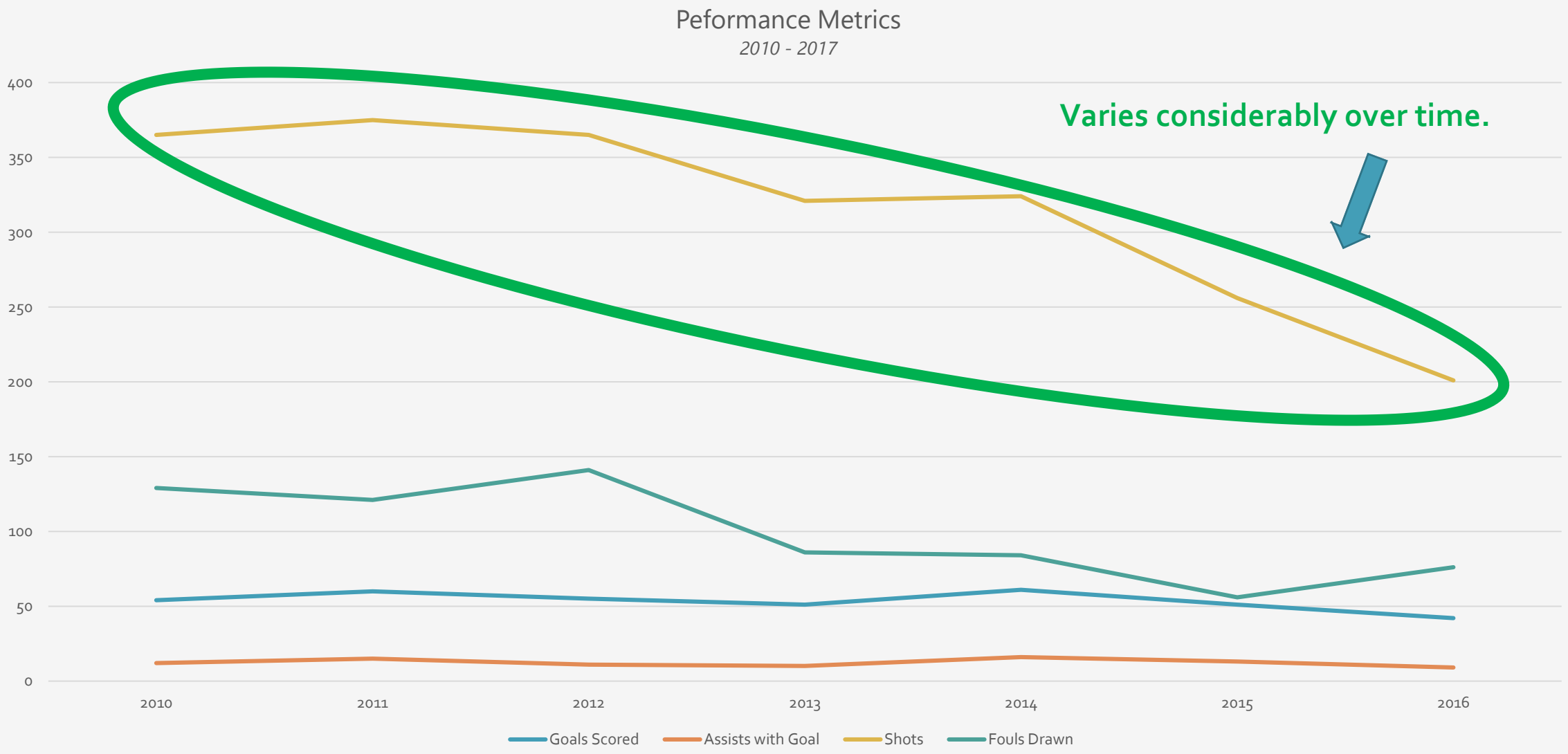
…or a few players.

Microsoft

Example 2:

Using Old Data


Peformance Metrics
2010 - 2017
400
350
300
250
200
150
100
50
0
2010
2011
2012
2013
2014
2015
2016
Goals Scored
Assists with Goal
Shots
Fouls Drawn


Peformance Metrics
2010 - 2017
Varies considerably over time.
400
350
300
250
200
150
100
50
0
2010
2011
2012
2013
2014
2015
2016
Goals Scored
Assists with Goal
Shots
Fouls Drawn

Microsoft

Example 3:

Choice of Measure

What does a “good” season look like?

Season 2012 - 2013

55	55	11	4861	365	141
Fixtures played	Goals Scored	Assists w/ Goal	Minutes Played	Shots	Fouls Drawn

	CHAMPIONS LEAGUE	COPA DEL REY	LIGA ESPAÑOLA 1ª DIVISIÓN	SUPERCOPA DE ESPAÑA
Fixtures played	12	7	34	2
Goals Scored	12	7	34	2
Assists w/ Goal	1	1	9	0
Minutes Played	1141	690	2841	189
Shots	80	42	235	8
Fouls Drawn	30	29	80	2
Matches Started	12	7	30	2
Matches Subbed In	0	0	4	0
Headers Scored	1	1	6	1
Goals w/ Right Foot	9	2	18	1
Goals w/ Left Foot	2	4	10	0
Penalty Kicks Scored	0	1	6	0
Free Kicks Scored	0	0	3	0
Fouls Committed	7	10	27	4
Steals	25	9	49	4
Penalties Received	0	1	1	0
Passes Completed	399	224	884	40

What does a “good” season look like?

Season 2012 - 2013

55	55	11	4861	365	141
Fixtures played	Goals Scored	Assists w/ Goal	Minutes Played	Shots	Fouls Drawn

	CHAMPIONS LEAGUE	COPA DEL REY	LIGA ESPAÑOLA 1ª DIVISIÓN	SUPERCOPA DE ESPAÑA
Fixtures played	12	7	34	2
Goals Scored	12	7	34	2
Assists w/ Goal	1	1	9	0
Minutes Played	1141	690	2841	189
Shots	80	42	235	8
Fouls Drawn	30	[illegible]	80	2
Matches Started	12	7	30	2
Matches Subbed In	0	0	4	0
Headers Scored	1	[illegible]	6	1
Goals w/ Right Foot	9	[illegible]	18	1
Goals w/ Left Foot	2	4	10	0
Penalty Kicks Scored	0	[illegible]	6	0
Free Kicks Scored	0	[illegible]	3	0
Fouls Committed	7	10	27	4
Steals	25	9	49	4
Penalties Received	0	1	1	0
Passes Completed	399	224	884	40

Microsoft

Example 4:

Cherry-picking Data

Microsoft

Example 5:

Reprobleming

Altering the problem so that your performance improves.

Microsoft

Example 6:

Parameter Tweaking

Only selecting parameters that you know will give you **good** results.

Season 2012 - 2013

55	55	11	4861	365	141
Fixtures played	Goals Scored	Assists w/ Goal	Minutes Played	Shots	Fouls Drawn

	CHAMPIONS LEAGUE	COPA DEL REY	LIGA ESPAÑOLA 1ª DIVISIÓN	SUPERCOPA DE ESPAÑA
Fixtures played	12	7	34	2
Goals Scored	12	7	34	2
Assists w/ Goal	1	1	9	0
Minutes Played	1141	690	2841	189
Shots	80	42	235	8
Fouls Drawn	30	29	80	2
Matches Started	12	7	30	2
Matches Subbed In	0	0	4	0
Headers Scored	1	1	6	1
Goals w/ Right Foot	9	2	18	1
Goals w/ Left Foot	2	4	10	0
Penalty Kicks Scored	0	1	6	0
Free Kicks Scored	0	0	3	0
Fouls Committed	7	10	27	4
Steals	25	9	49	4
Penalties Received	0	1	1	0
Passes Completed	399	224	884	40

Only selecting parameters that you know will give you **good** results.

Season 2012 - 2013

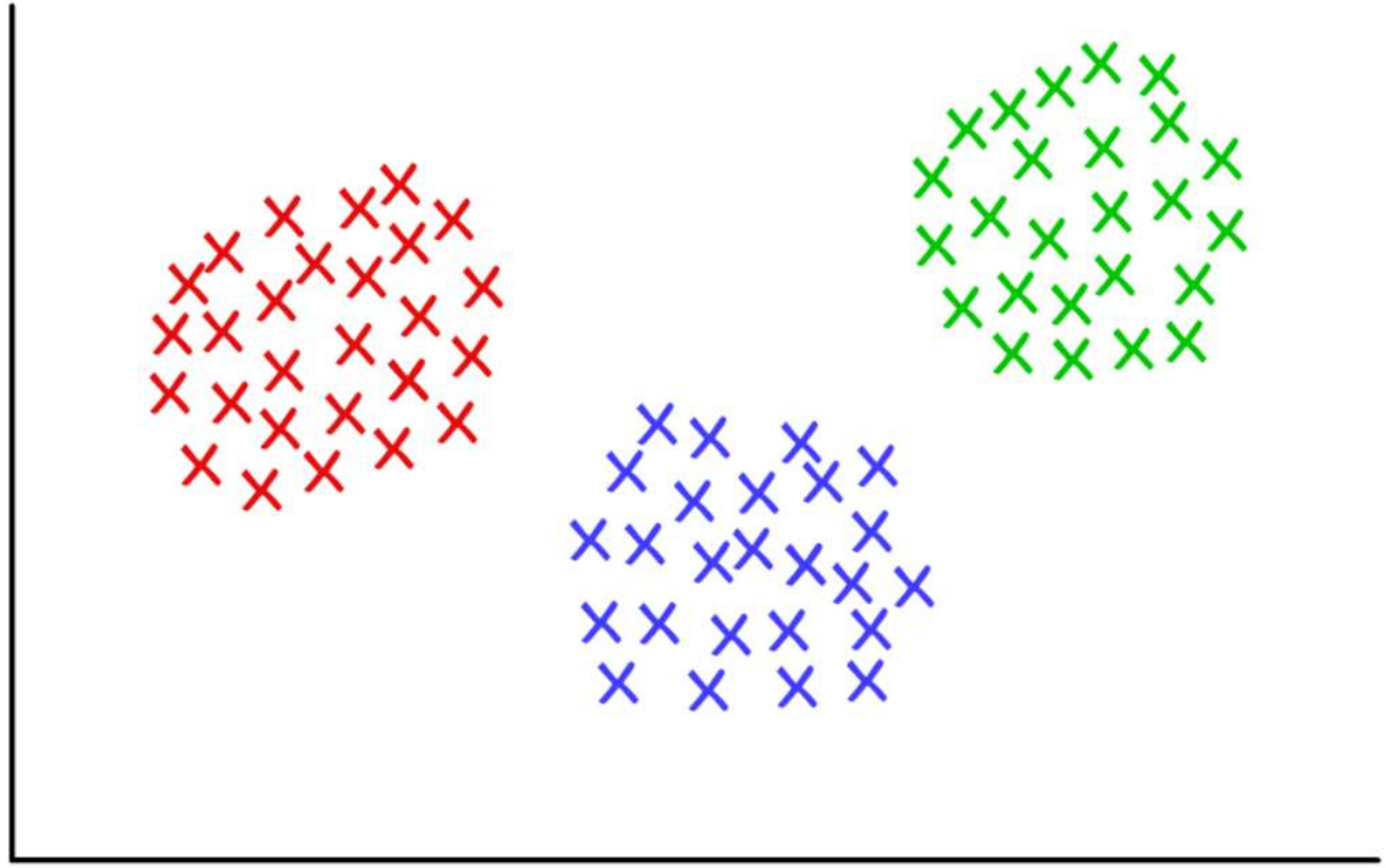
55	55	11	4861	365	141
Fixtures played	Goals Scored	Assists w/ Goal	Minutes Played	Shots	Fouls Drawn

	CHAMPIONS LEAGUE	COPA DEL REY	LIGA ESPAÑOLA 1ª DIVISIÓN	SUPERCOPA DE ESPAÑA
Fixtures played	12	7	34	2
Goals Scored	12	7	34	2
Assists w/ Goal	1	1	9	0
Minutes Played	1141	690	2841	189
Shots	80	42	235	8
[redacted]				
Matches Started	12	7	30	2
Matches Subbed In	0	0	4	0
Headers Scored	1	1	6	1
Goals w/ Right Foot	9	2	18	1
Goals w/ Left Foot	2	4	10	0
Penalty Kicks Scored	0	1	6	0
Free Kicks Scored	0	0	3	0
[redacted]				
Steals	25	9	49	4
[redacted]				
Passes Completed	399	224	884	40

Microsoft

Example 7:

Human-Loop Overfitting



Example:

Using a clustering algorithm (on training and test samples) to guide learning algorithm choice.

Example 8:

Collinearity

For example:

Player *age*, *date of birth*, and number of *years playing professionally* would probably give you similar information.

Microsoft

Example 9:

Overfitting by Review

Example:

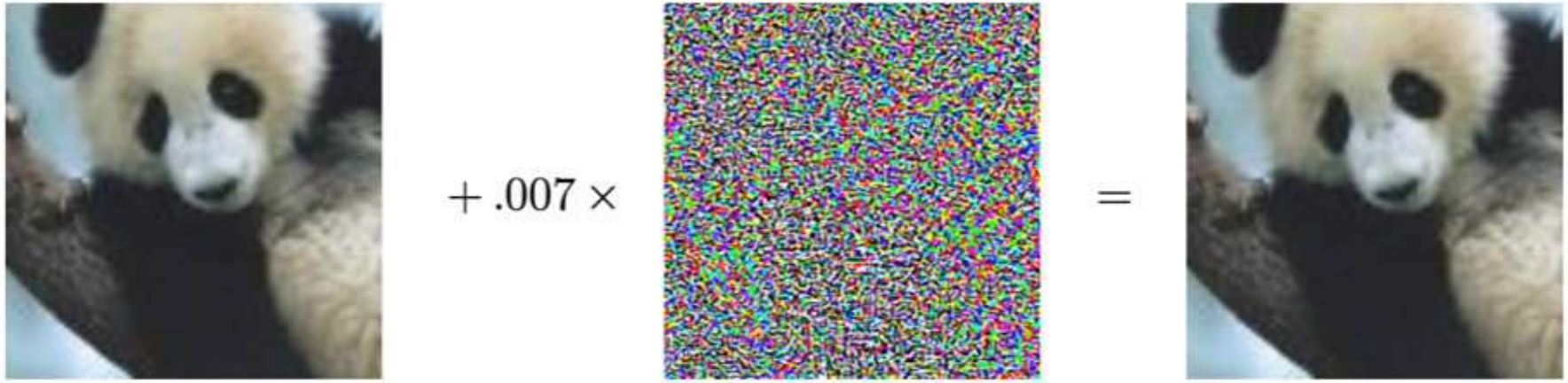
We **only** see data for the **public** games – **not** the data for **trial** matches.

Microsoft

Example 10:

Deep Learning in General

A common way to find perturbations forcing models to make wrong predictions is to compute **adversarial examples** [SZS13]. They yield perturbations that are very slight and often indistinguishable to humans, yet force machine learning models to produce wrong predictions. For instance, in the illustration below reproduced from [GSS14], the image on the left is correctly classified by a machine learning model as a panda, but adding the noise represented in the middle to that same image results in the image on the right, which is classified as a gibbon by the model.



$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Microsoft

You **will** overfit.

DENIAL

How could my performance on test data be so low? There must be a mistake!

ANGER

What did I do wrong?! This is the [data / algorithm / programming language]'s fault!

BARGAINING

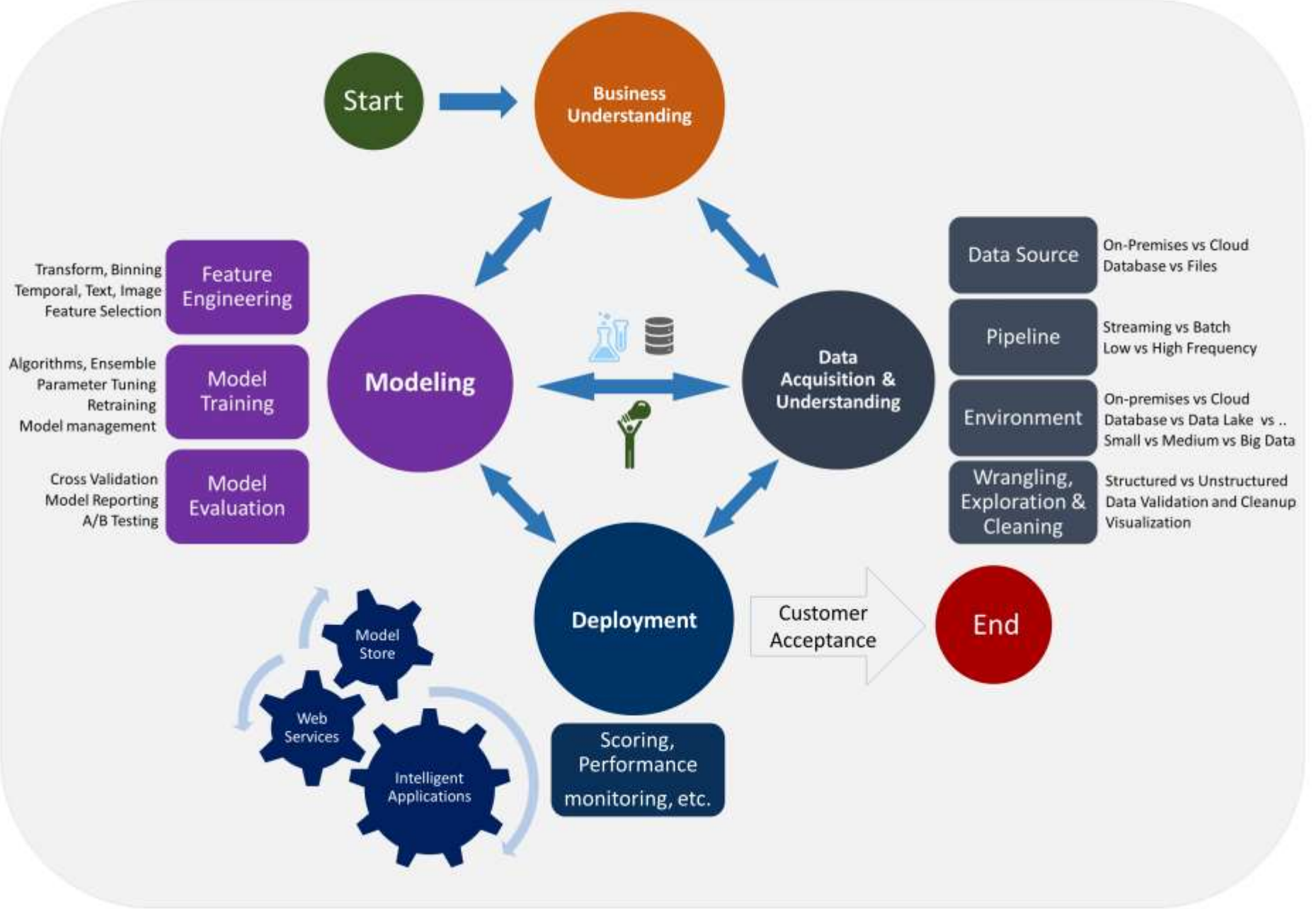
Maybe if I add more data? Or if I do more feature engineering?

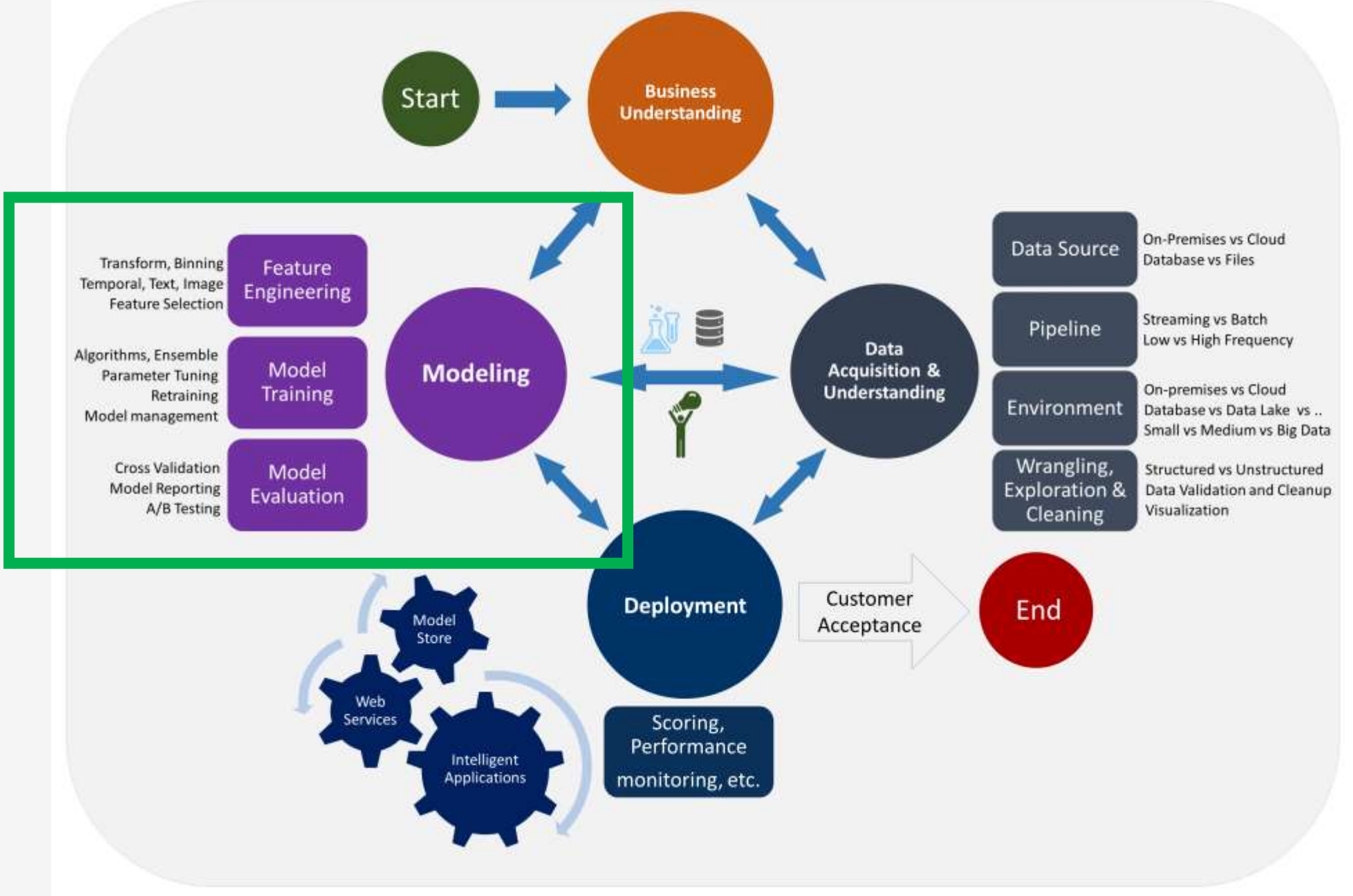
DEPRESSION

Why can't I fix this? Why does machine learning hate me? Maybe I should go be a coffee farmer

ACCEPTANCE

I will overfit. I am one with that reality. I must overfit as little as possible.


Start
Business Understanding
Transform, Binning
Temporal, Text, Image
Feature Selection
Feature Engineering
Algorithms, Ensemble
Parameter Tuning
Retraining
Model management
Model Training
Cross Validation
Model Reporting
A/B Testing
Model Evaluation
Modeling
Data Acquisition & Understanding
Data Source
On-Premises vs Cloud
Database vs Files
Pipeline
Streaming vs Batch
Low vs High Frequency
Environment
On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data
Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleanup
Visualization
Deployment
Customer Acceptance
End
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

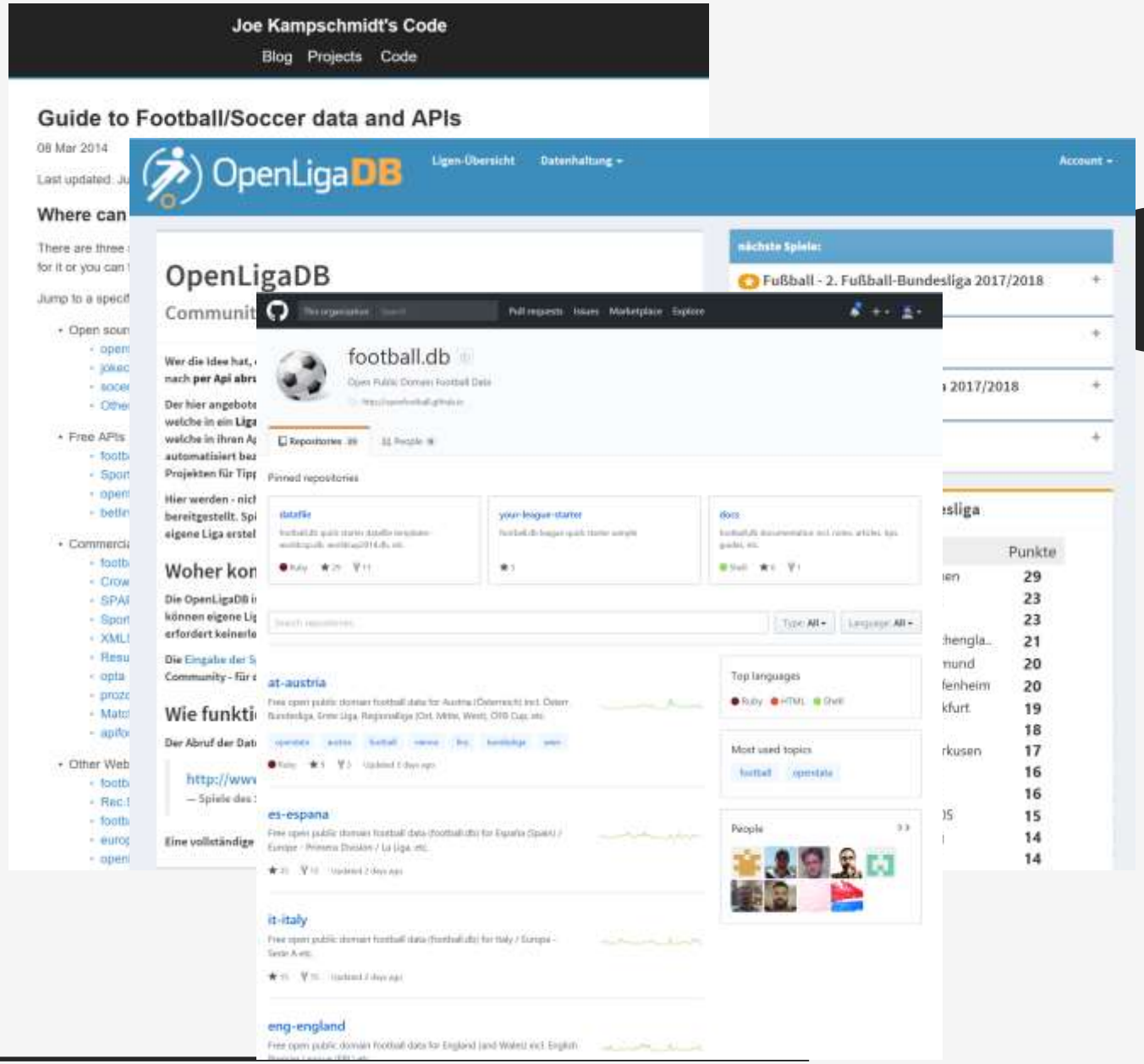

Start
Business Understanding
Transform, Binning
Temporal, Text, Image
Feature Selection
Feature Engineering
Algorithms, Ensemble
Parameter Tuning
Retraining
Model management
Model Training
Cross Validation
Model Reporting
A/B Testing
Model Evaluation
Modeling
Data Acquisition & Understanding
Data Source
On-Premises vs Cloud
Database vs Files
Pipeline
Streaming vs Batch
Low vs High Frequency
Environment
On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data
Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleanup
Visualization
Deployment
Customer Acceptance
End
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

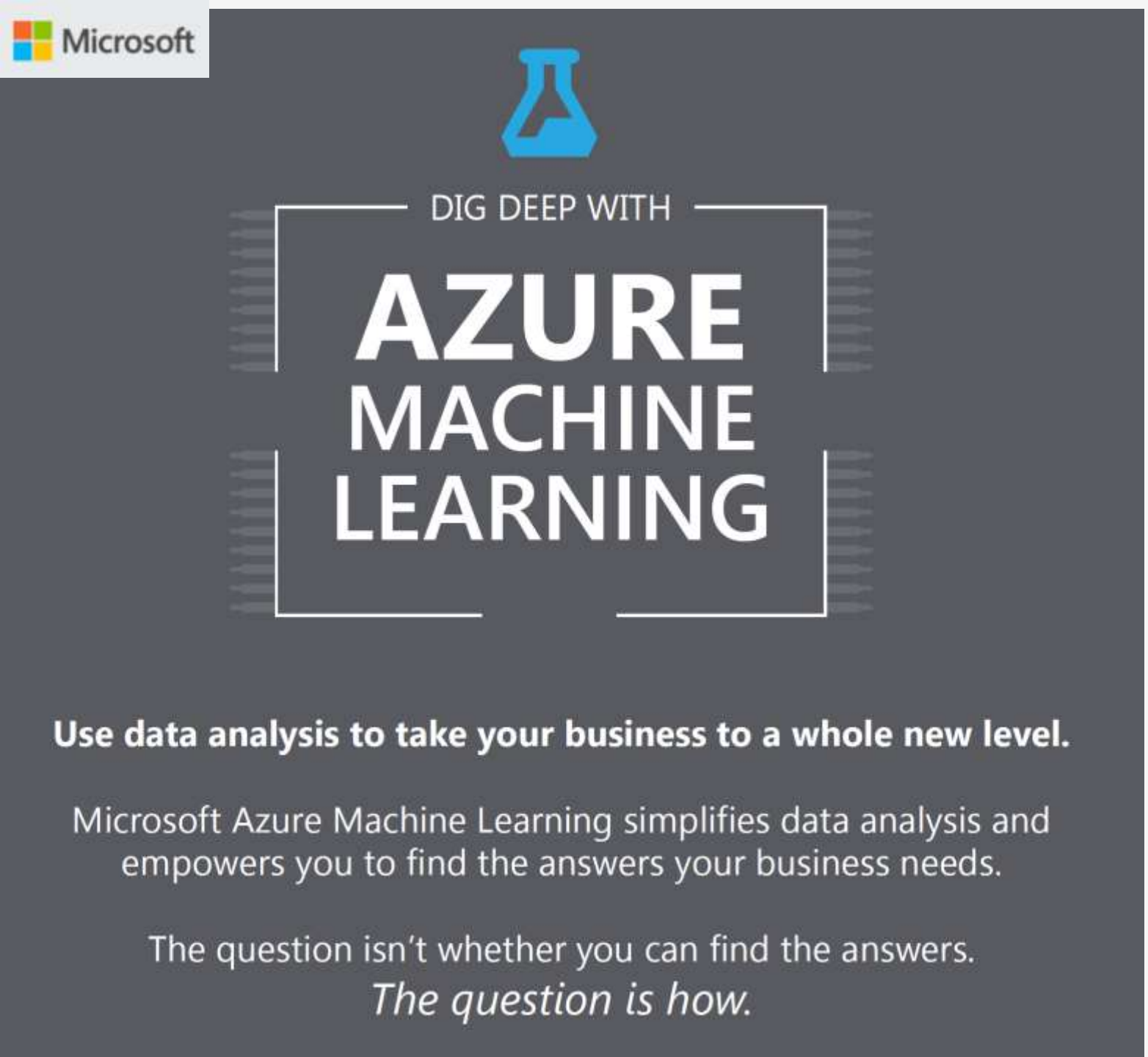
RESOURCES (football data):

- Football.db database
- Guide to Football and Soccer APIs
- Sports Open Data Statistics
- OpenLigaDB

RESOURCES (machine learning):

- Team Data Science Process
- Azure Machine Learning Workbench
- Learn Analytics @ Microsoft
- Microsoft AI School
- DataCamp
- AI Tools for Visual Studio
- Machine Learning (Coursera)
- Deep Learning (Coursera)





SELECTING A MODEL

Thank you!