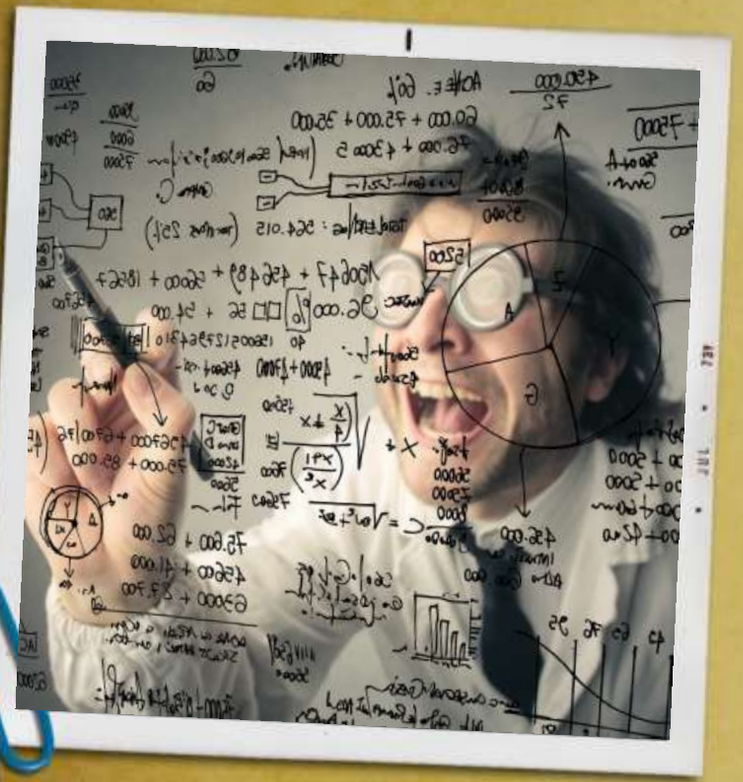


TOP SECRET

Microsoft

7 Secrets of Effective Data Science in a Business Context

Agent Name: Rui Carmo
Rank: Cloud Solution Architect
Code Name: @rcarmo

Disclaimer

THIS IS NOT A PRODUCT PITCH

IT'S A MINDSET PITCH

DATA SCIENCE & AI

KEY TRENDS

- Accelerating adoption of AI by developers (consuming models)
- Rise of hybrid training and scoring scenarios
- Push scoring/inference to the event (edge, cloud, on-prem)
- Moving high-end developers into deep learning as non-traditional path to DS / AI dev
- Growth of diverse hardware arms race across all form factors (CPU / GPU / FPGA / ASIC / device)
- Demonstrating success of transfer learning techniques while reducing dev complexity

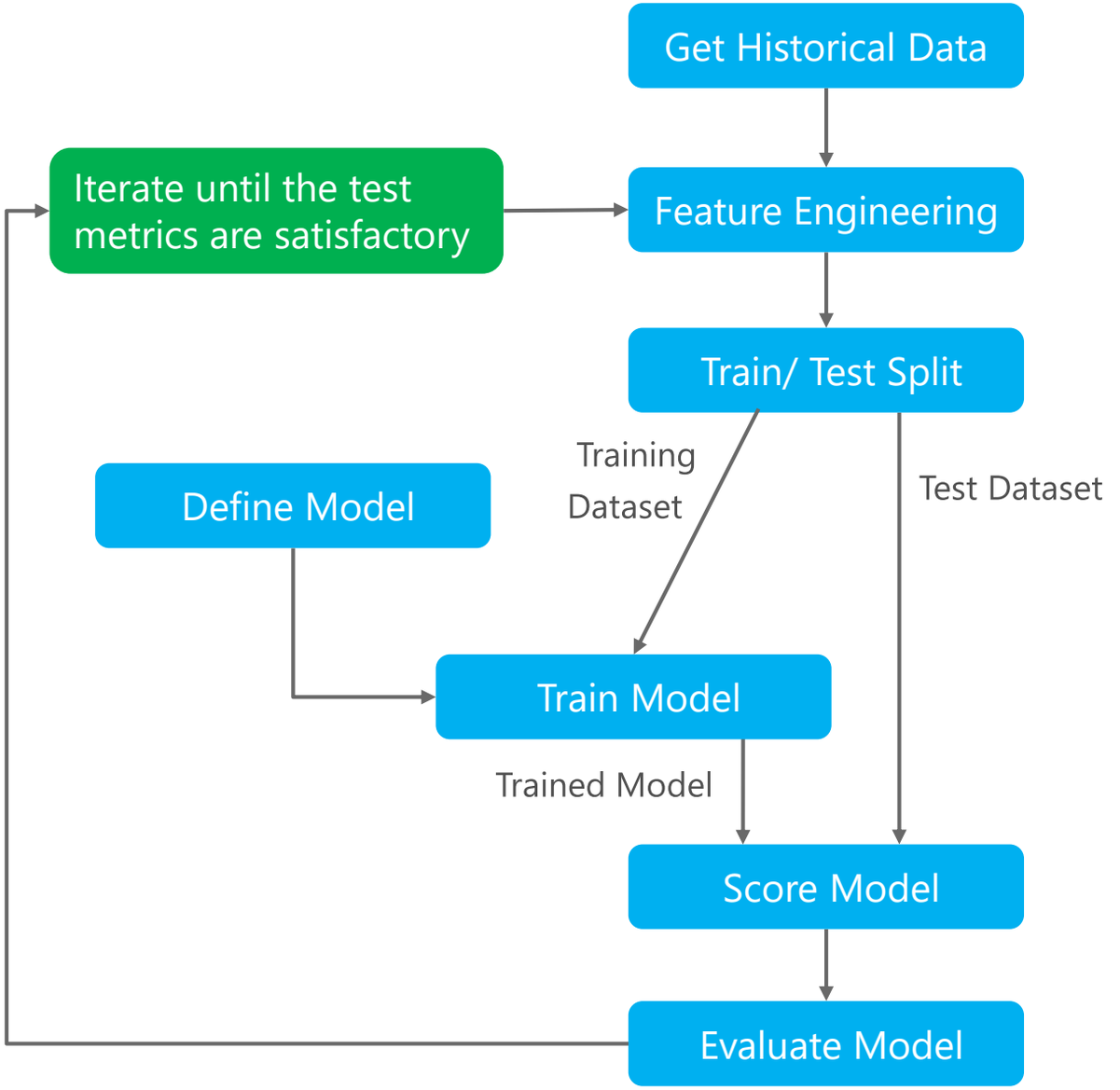
CHALLENGES

- Data prep
- Model deployment & management
- Model lineage & auditing
- Explain-ability

#1 DATA SCIENCE IS (PART OF) A PROCESS

The way we were taught

1. Define business problem
2. Understand data relationships
3. Prepare data
4. Construct models
5. Evaluate models
6. Improve models
7. (Cross) validate model
8. Publish model



BUT IN THE REAL WORLD…

trained model

real life!

~~business~~

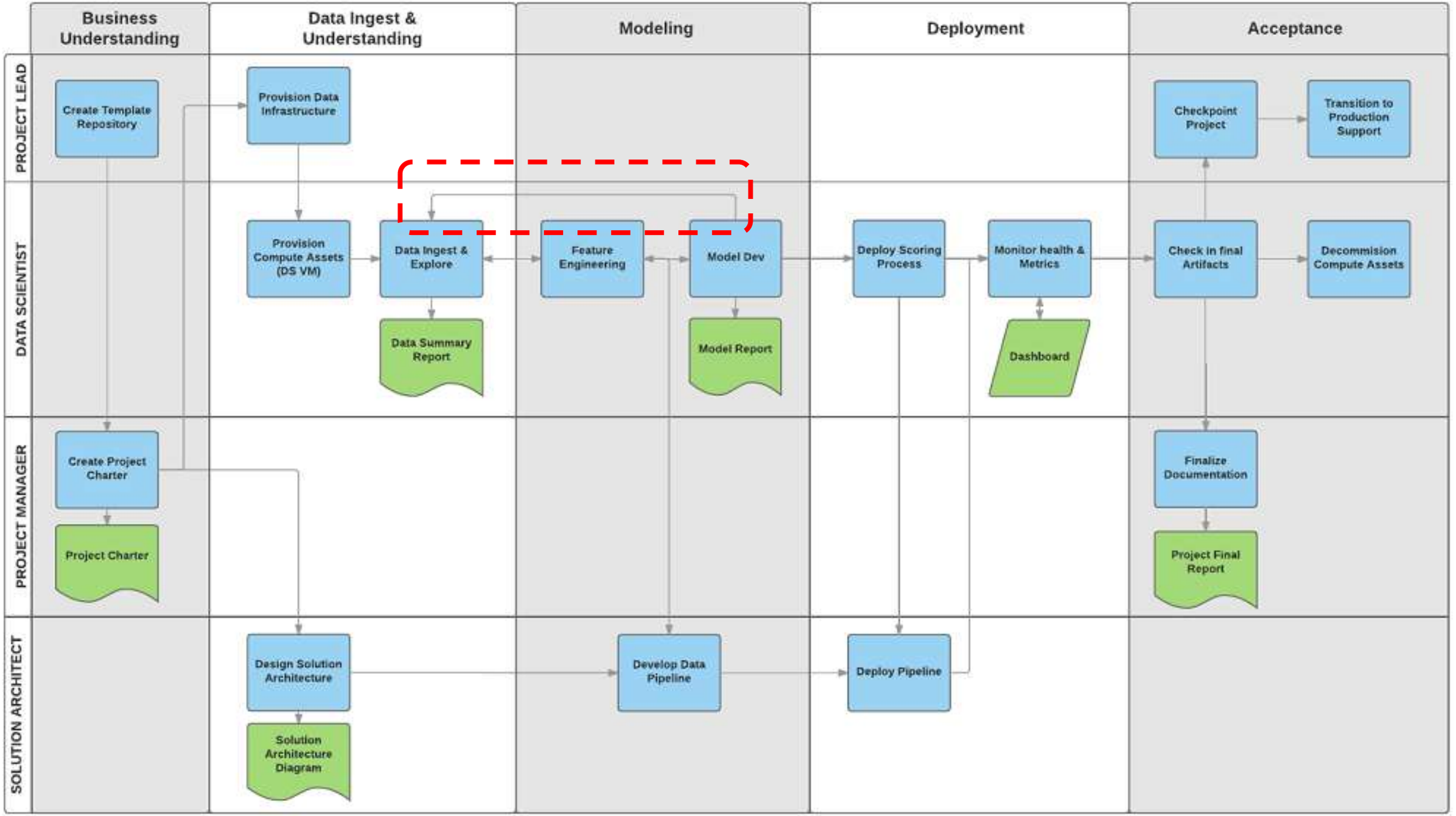
~~production~~

"No ~~battle plan~~ survives contact with ~~the enemy"~~

Helmuth von Moltke
Field Marshal, Prussian Army

"Strategy is a system of expedients"

Initiative and independent judgement along clearly defined roles



The Microsoft Team Data Science Process - https://github.com/Azure/Microsoft-TDSP

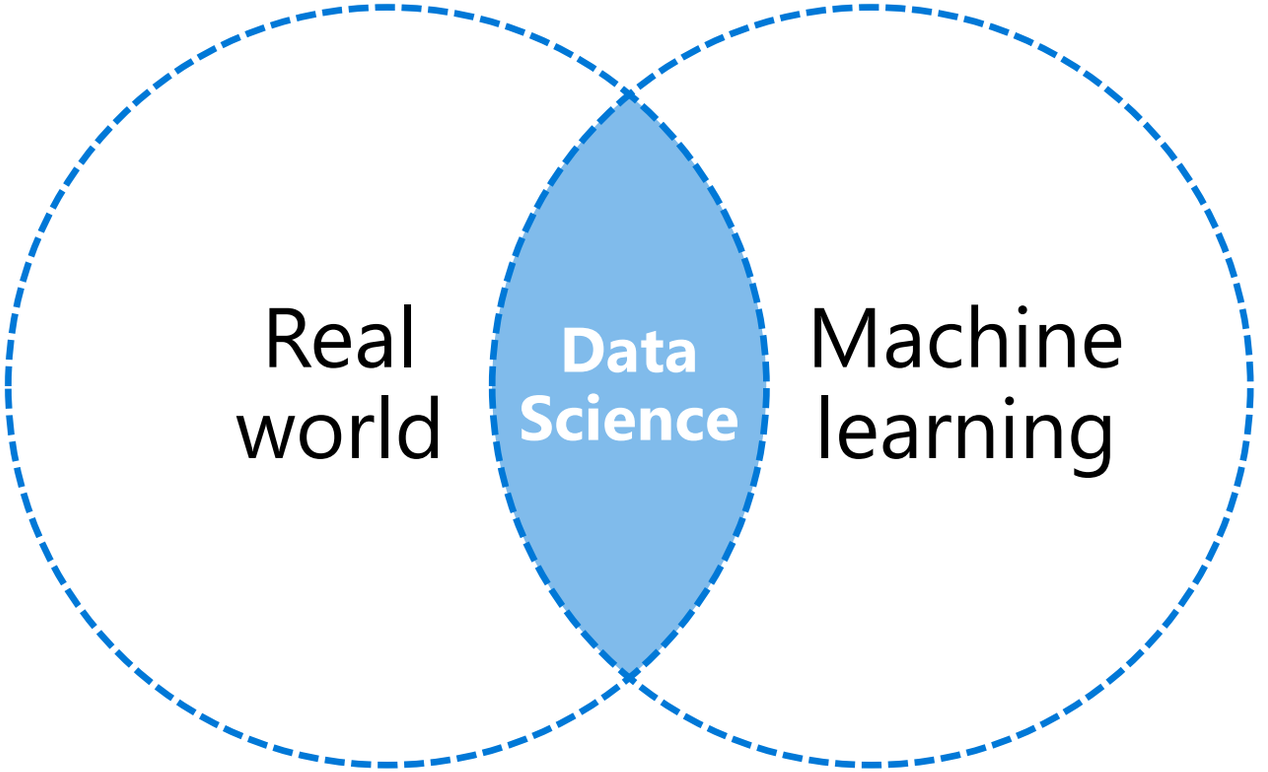
A DATA SCIENTIST NEVER WORKS ALONE

(UNLESS HE'S A MOVIE CHARACTER)

#2

YOUR SUCCESS IS THE BUSINESS' SUCCESS

DATA SCIENCE FOR DUMMIES



Data Science is the practice of deriving information and insight from real-world data to create business value.

ASKING THE RIGHT QUESTIONS

Business scenario	Key decision	Data Science question
Energy forecasting	Should I buy or sell energy contracts?	What will be the long/short-term demand for energy in a region?
Customer churn	Which customers should I prioritize to reduce churn?	What is probability of churn within X days for each customer?
Personalized marketing	What product should I offer first?	What is the probability that customer will purchase each product?
Product feedback	Which service/product needs attention?	What is social media sentiment for each service/product?

BUT CAN YOU GIVE **THE RIGHT KINDS** OF ANSWERS?

Are you looking at ~~all~~ **the right** data for the business outcome?

Can you **tell a story** with the data you have?

Are your answers **actionable** in business terms?

Can you help the business **prioritize**?

UNDERSTAND THE BUSINESS TO HELP IT PROSPER

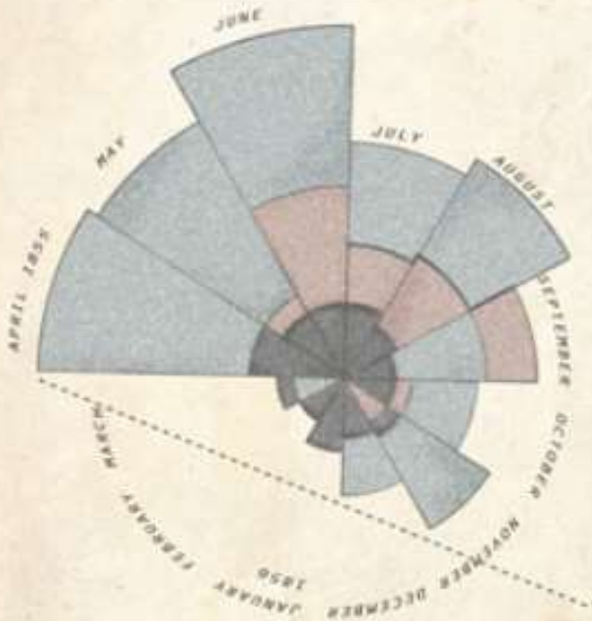
#3 COMMUNICATION IS KEY

Florence Nightingale (1820 – 1910)

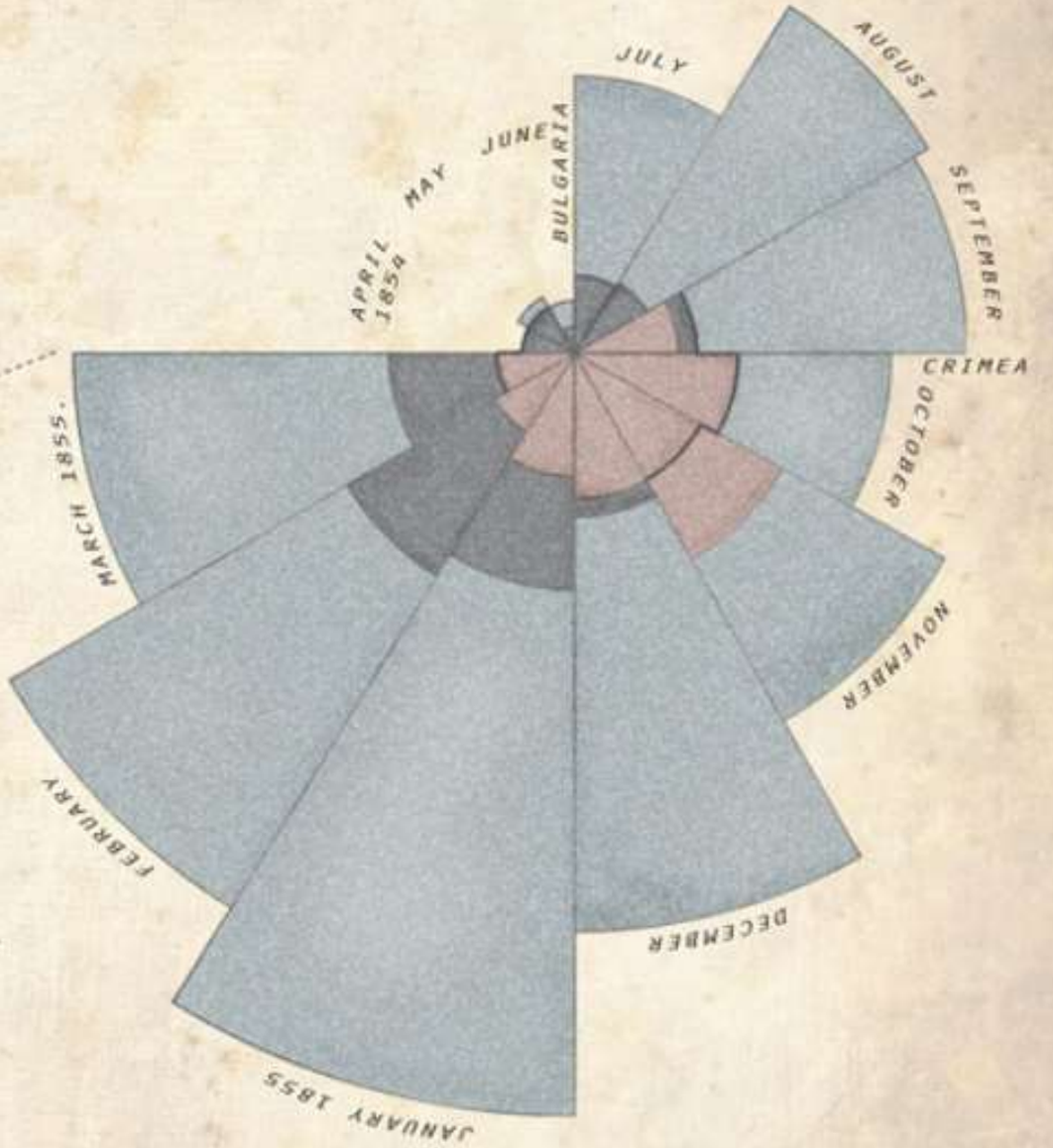
ONE OF THE EARLY DATA SCIENTISTS

DIAGRAM of the CAUSES of MORTALITY
IN THE ARMY IN THE EAST

2.
APRIL 1855 to MARCH 1856



1.
APRIL 1854 to MARCH 1855



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

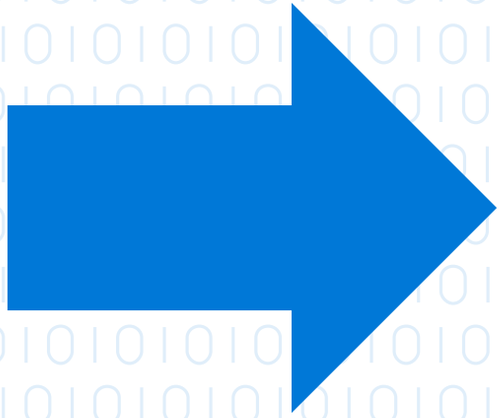
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases. the red wedges measured from the centre the deaths from wounds. & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov. 1854 marks the boundry of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red. in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

43%



2%

3 months after Nightingale arrived in Scutari

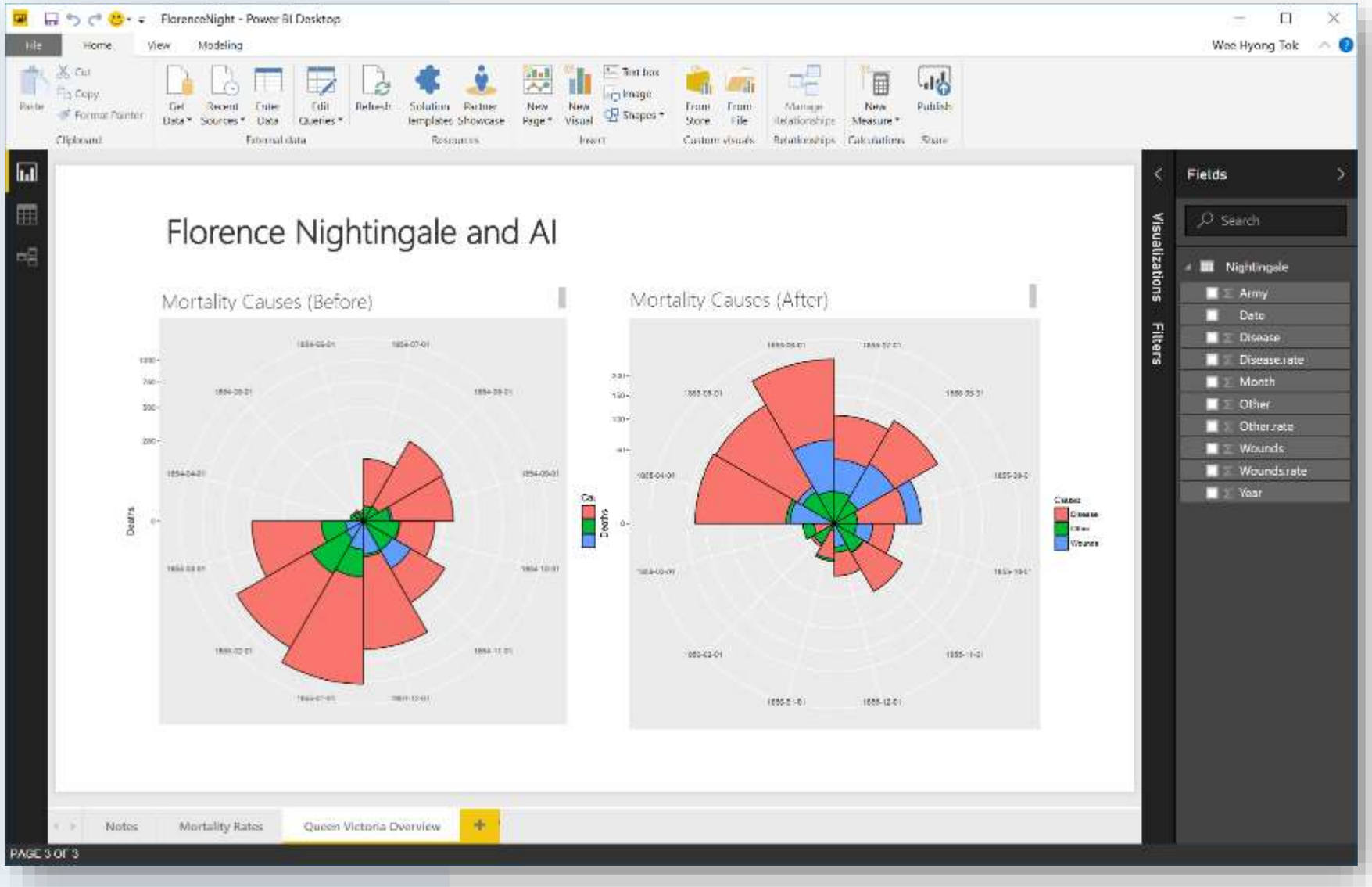
14 months later

Improved sanitary conditions in the hospitals

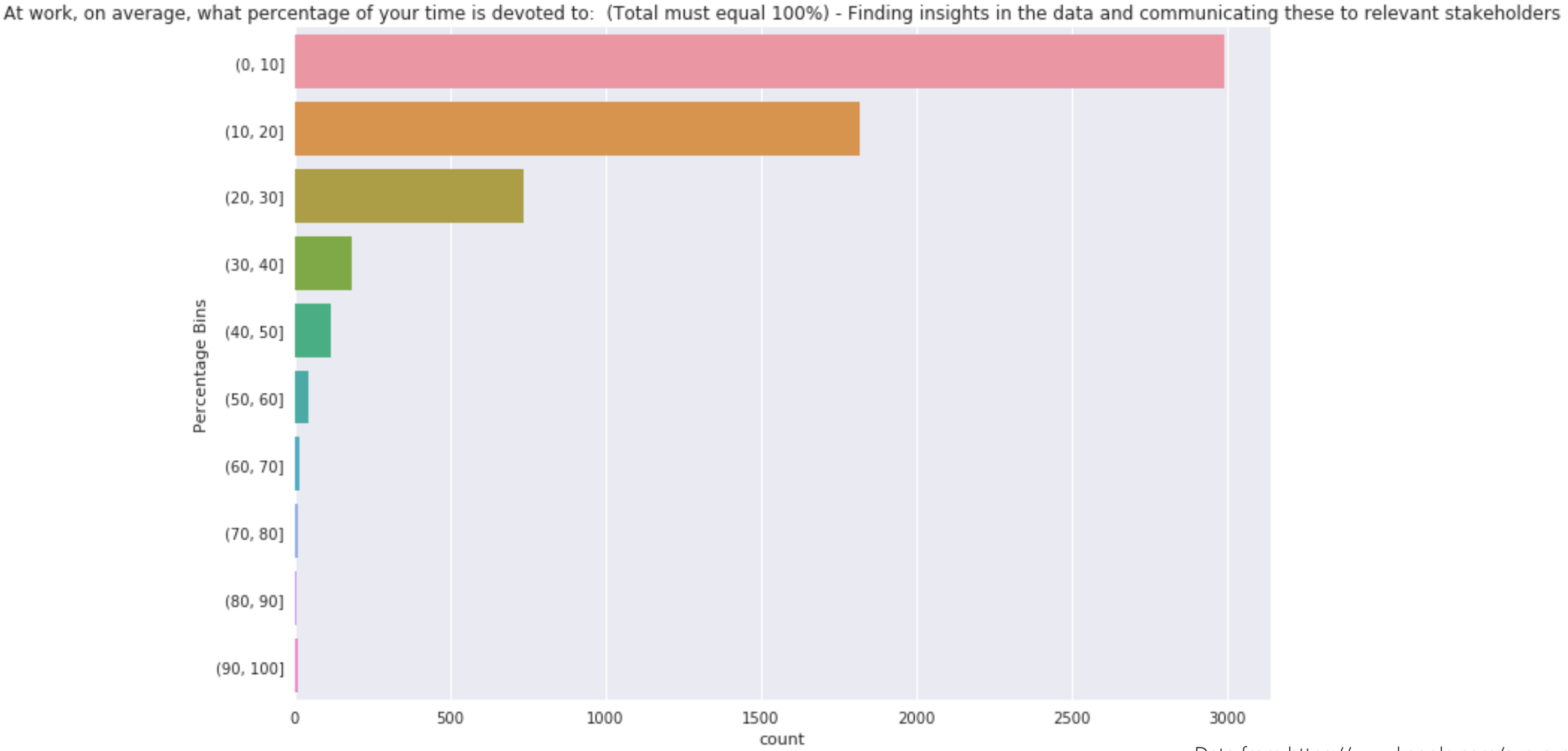
Laundry for washing linens

Today…

Nightingale Diagram?



AND YET...



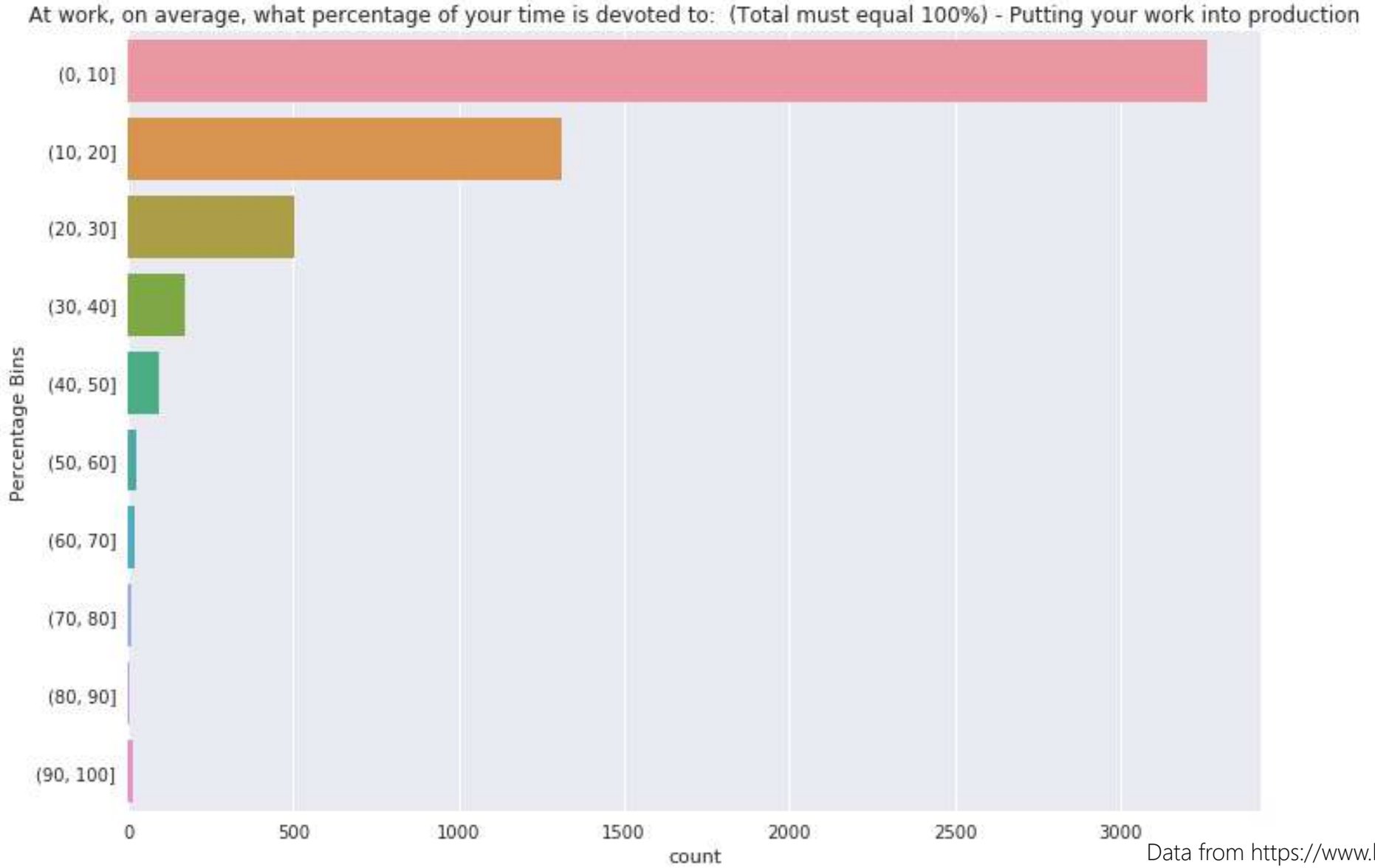
Data from https://www.kaggle.com/surveys/2017

MAKE YOUR IMPACT A MATTER OF RECORD

#4

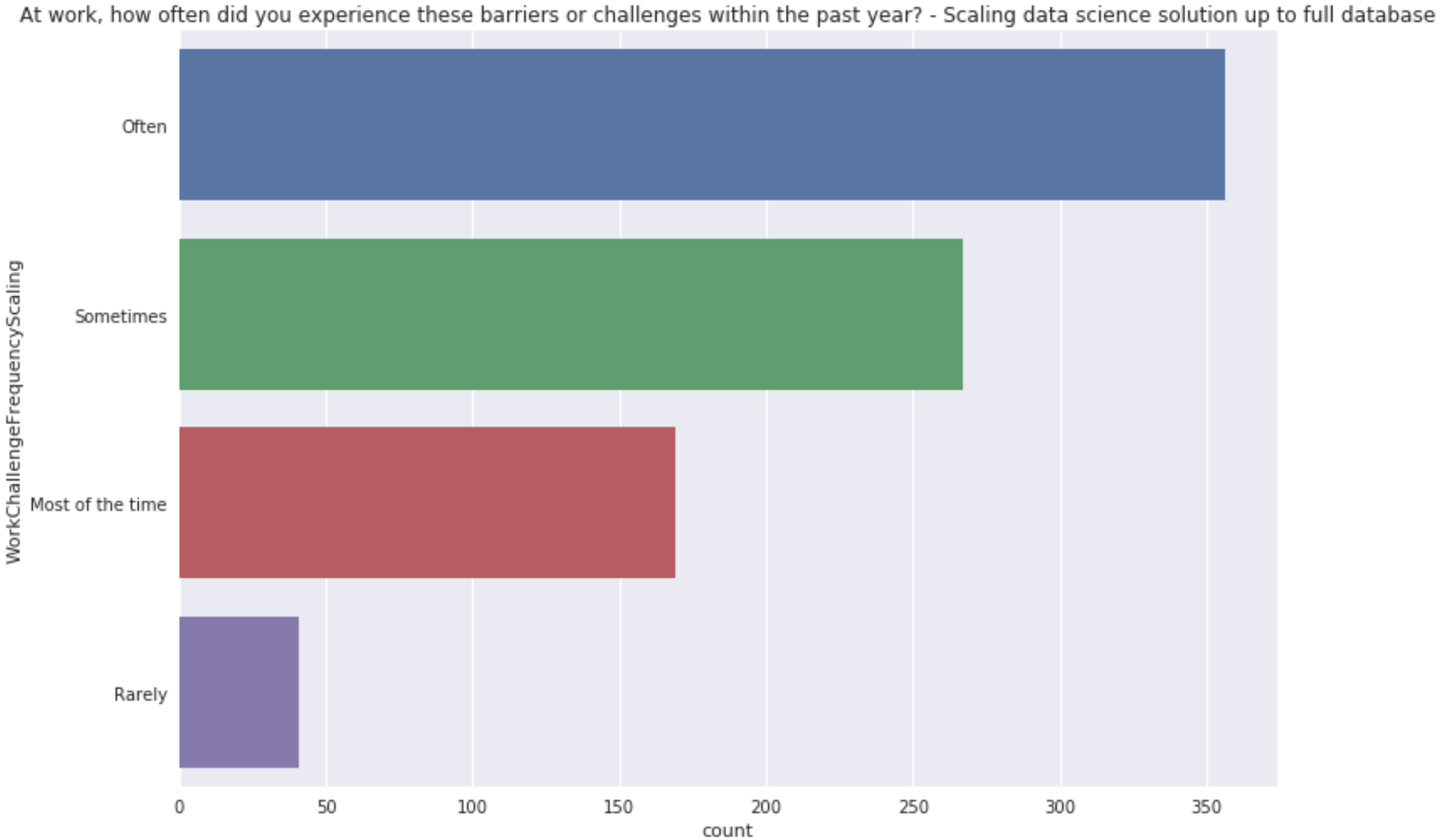
YOU **HAVE** TO WORRY ABOUT PRODUCTION

"IT WORKS ON MY MACHINE" (AND WHERE ELSE?)



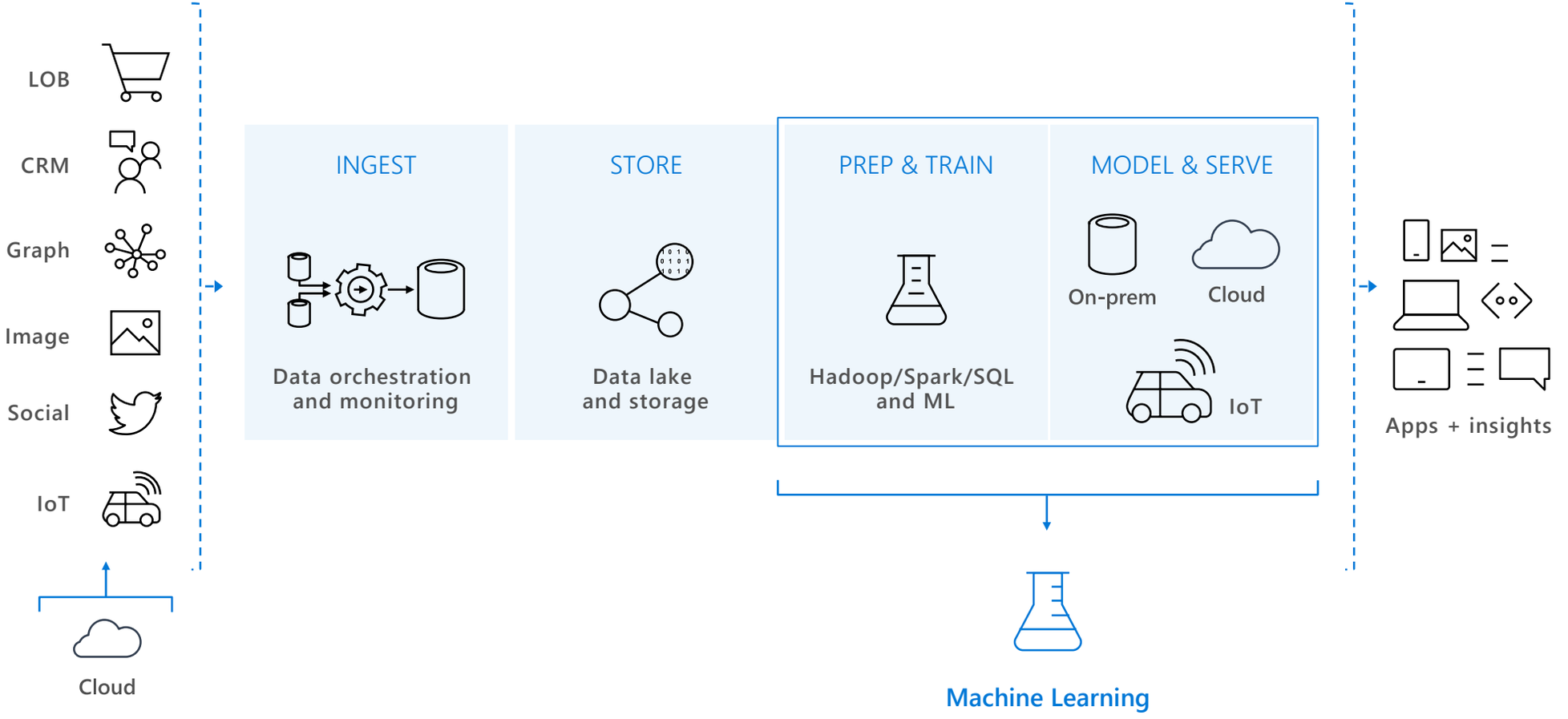
Data from https://www.kaggle.com/surveys/2017

SCALING IS A PROBLEM THAT MATTERS

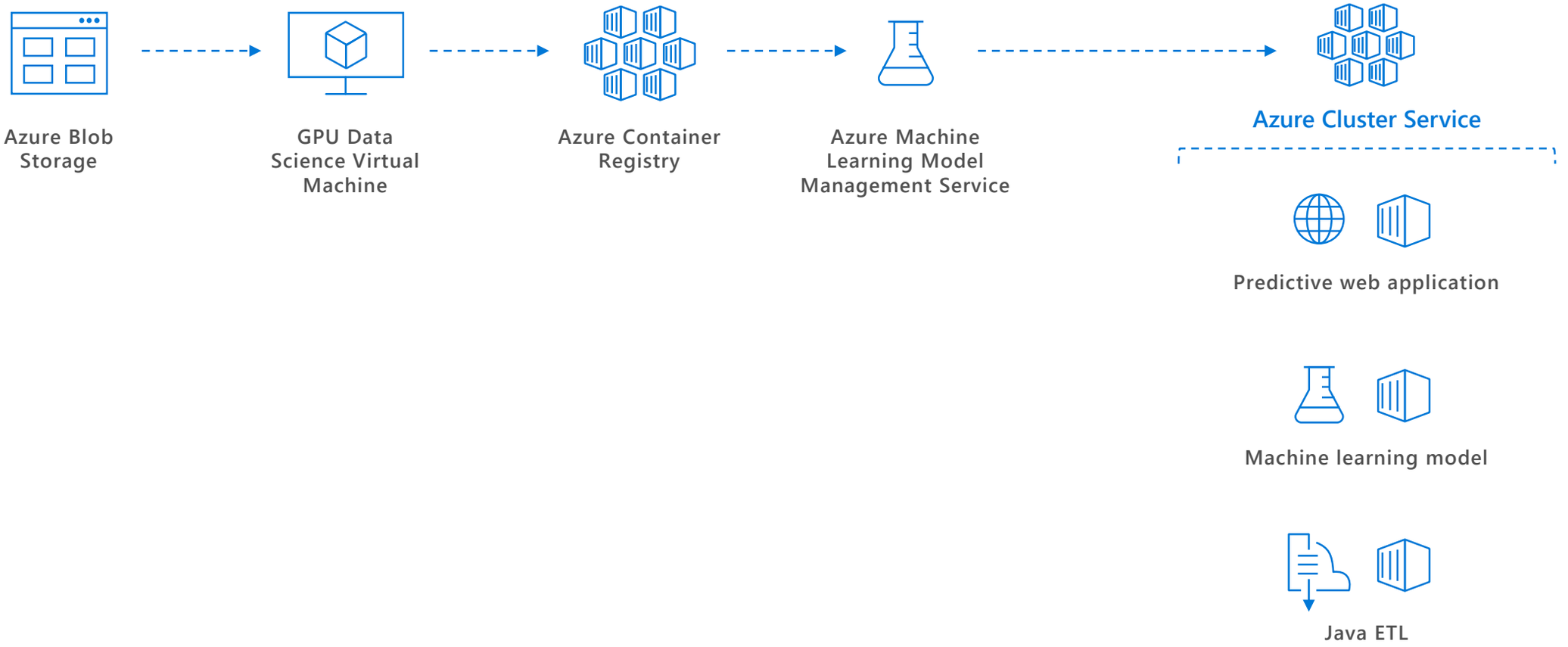


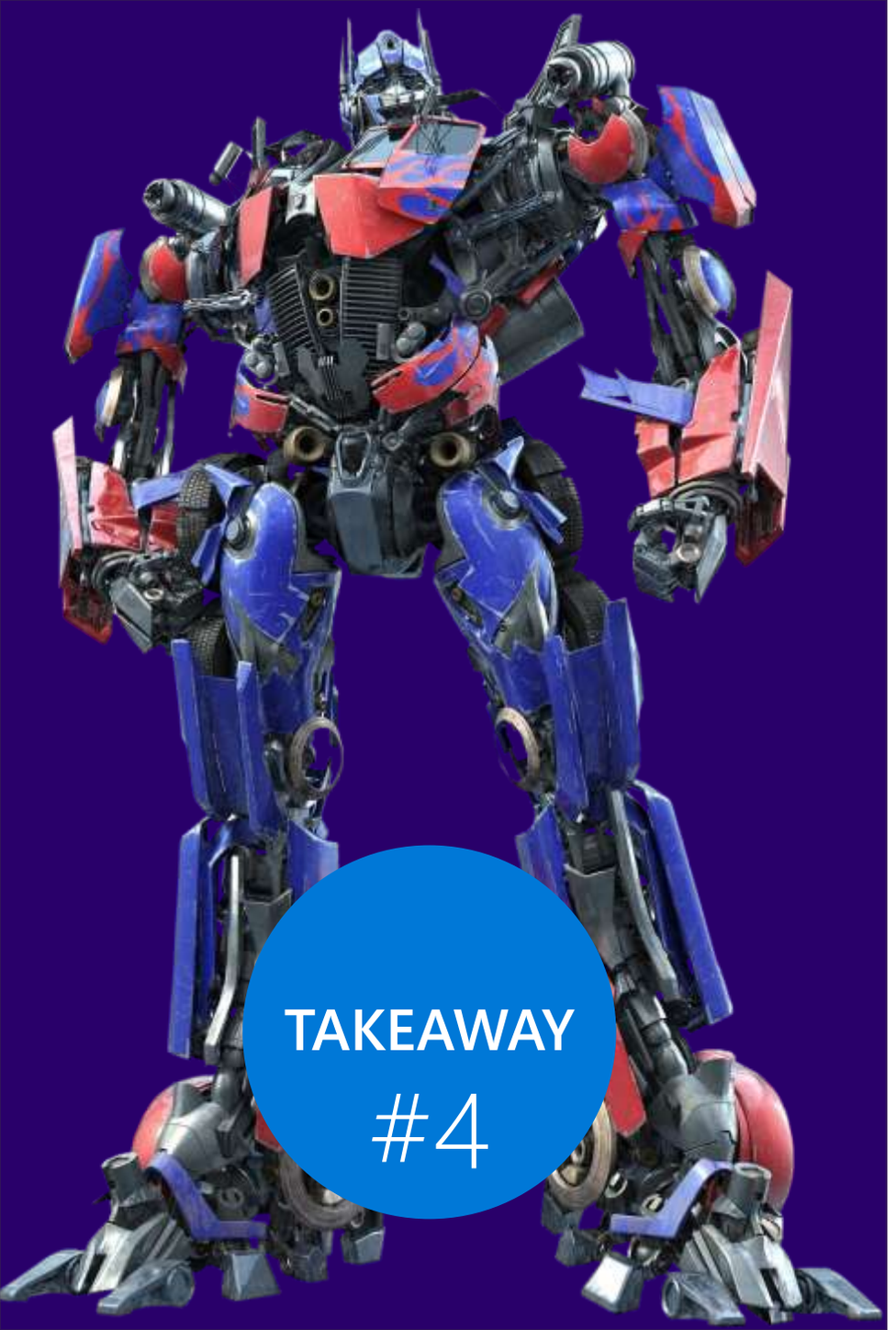
Data from https://www.kaggle.com/surveys/2017

EVOLUTION OF THE DATA ESTATE



PACKAGING AND DEPLOYNG MODELS IN THE CLOUD

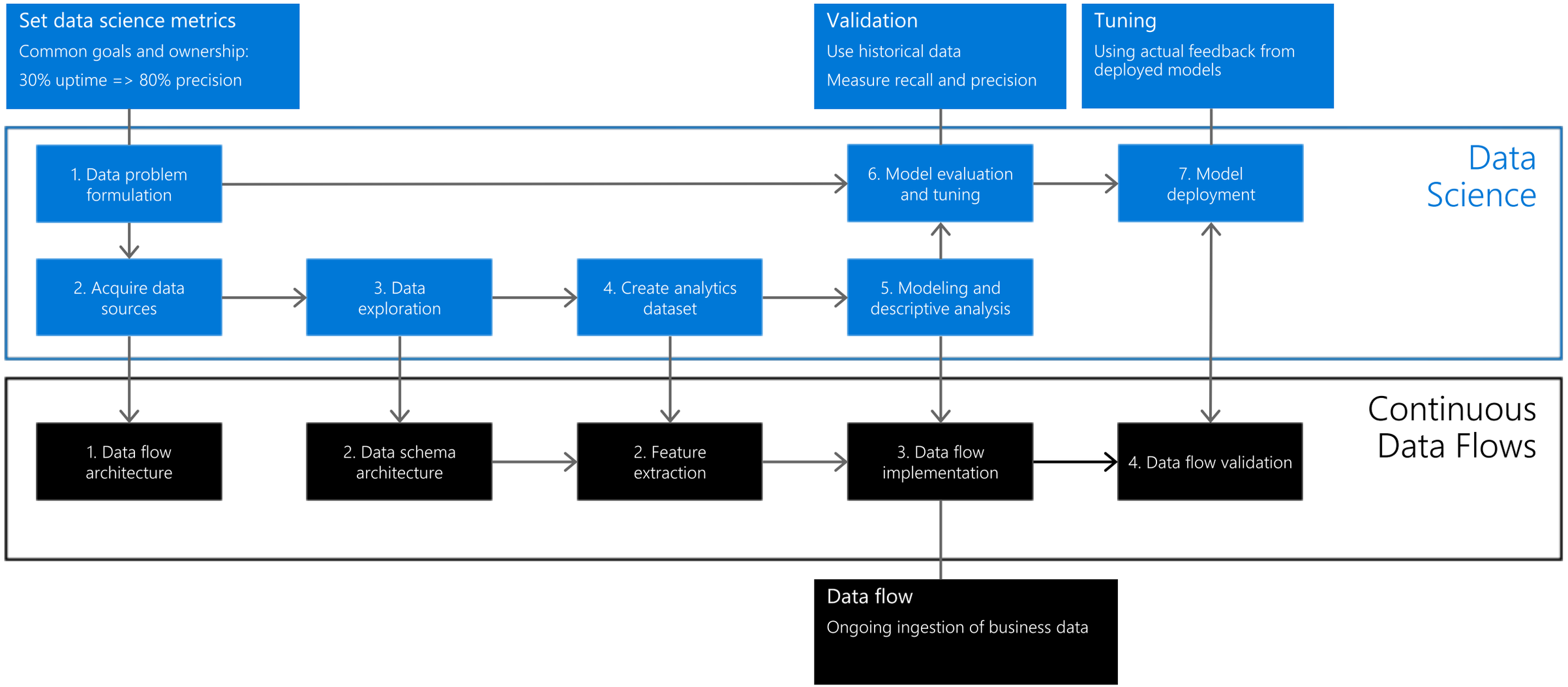




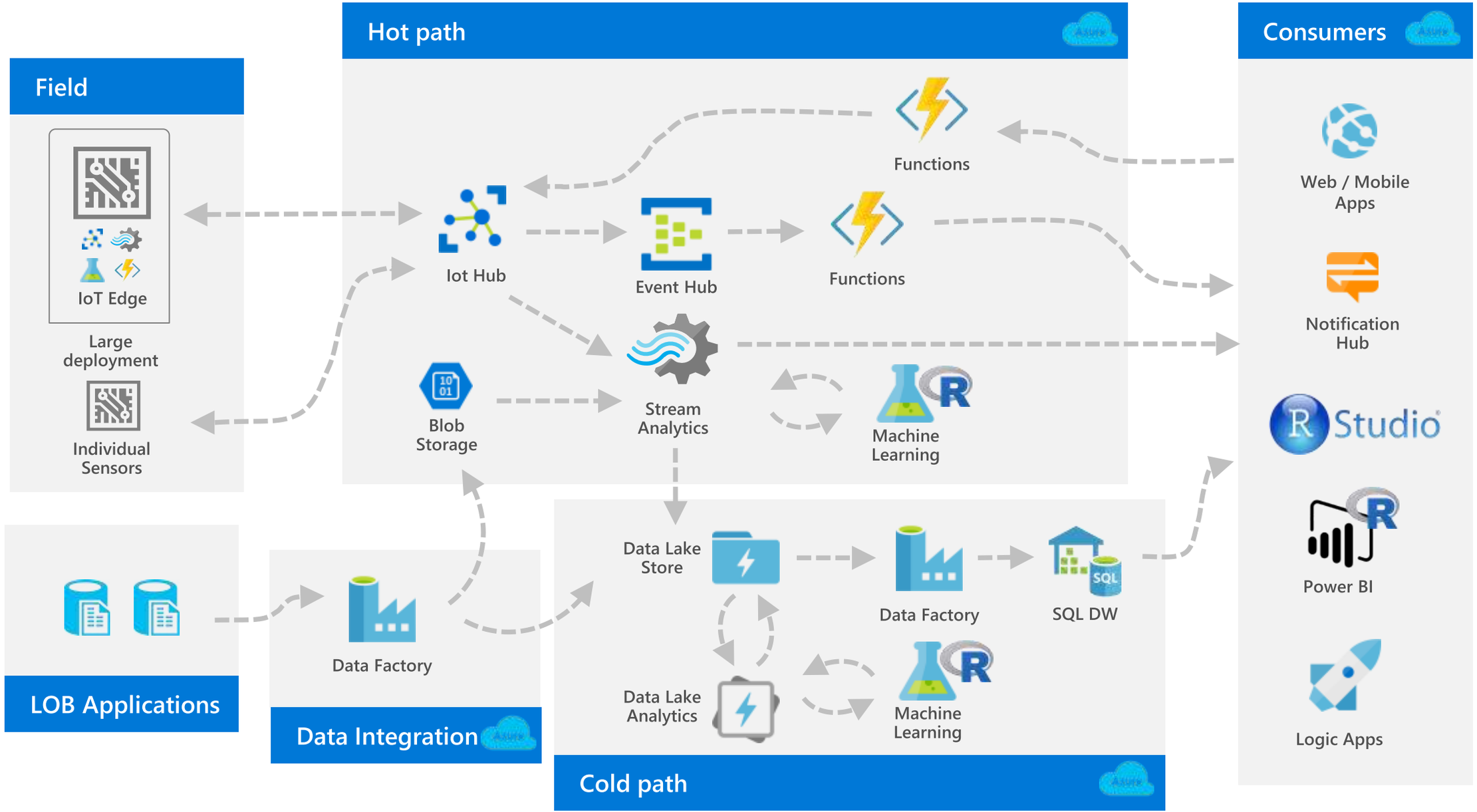
GET READY TO AUTOMATE IN A BIG WAY

#5 MASTERING THE DATA JUNGLE

DATA SCIENCE VS DATA ENGINEERING



REAL-LIFE DEPLOYMENTS

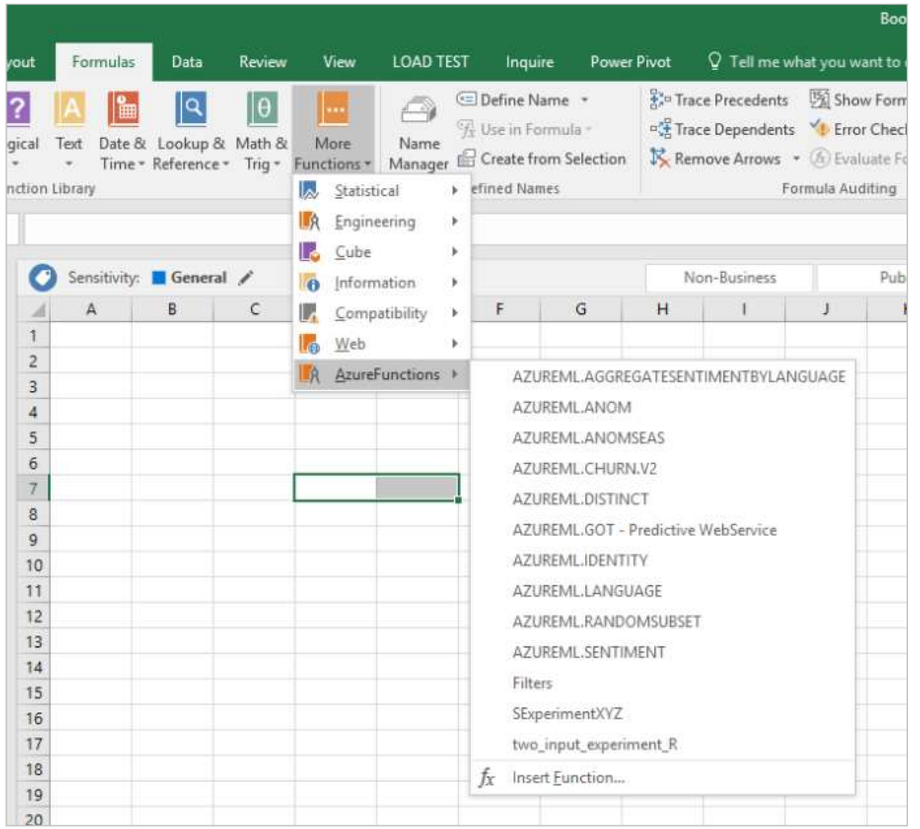


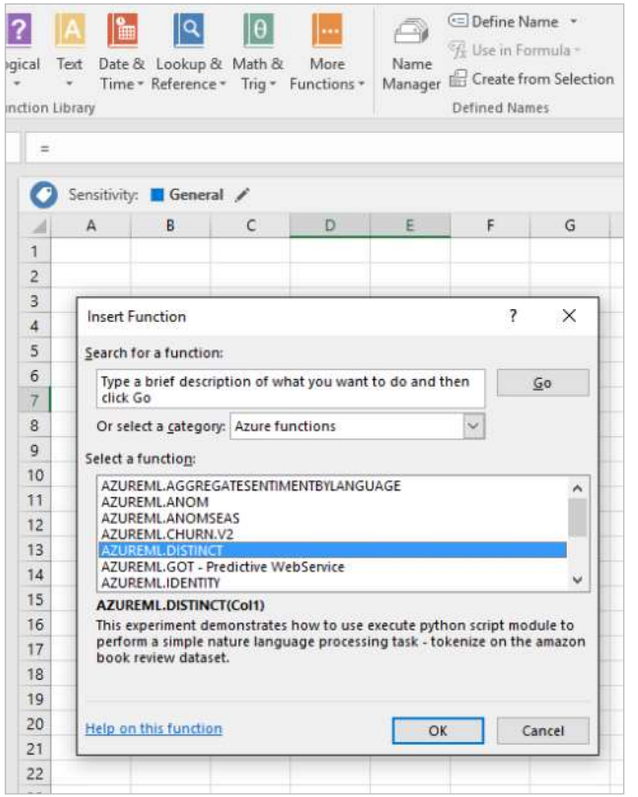
YOU MUST GO DEEPER
(AND GET YOUR FEET WET)

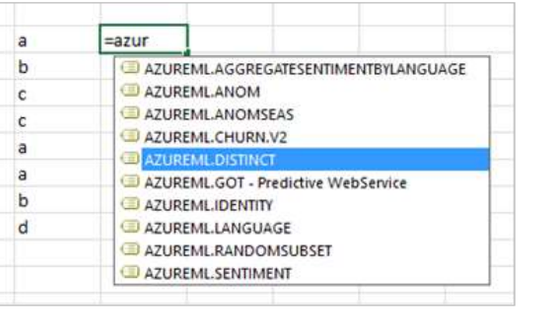
#6

TOOLING IS **NEVER** WHAT YOU EXPECT

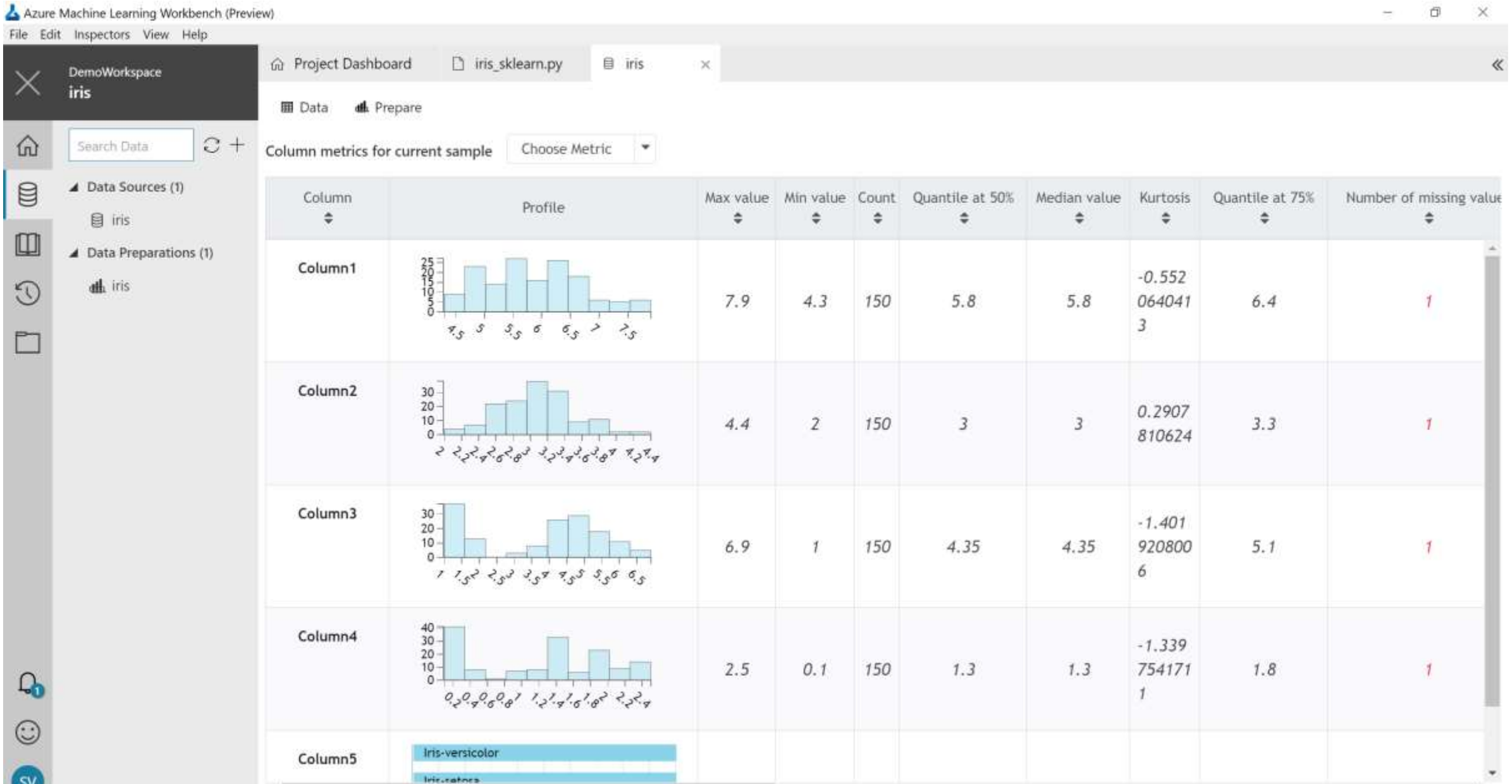
LET THEM USE EXCEL – IF THEY MUST...



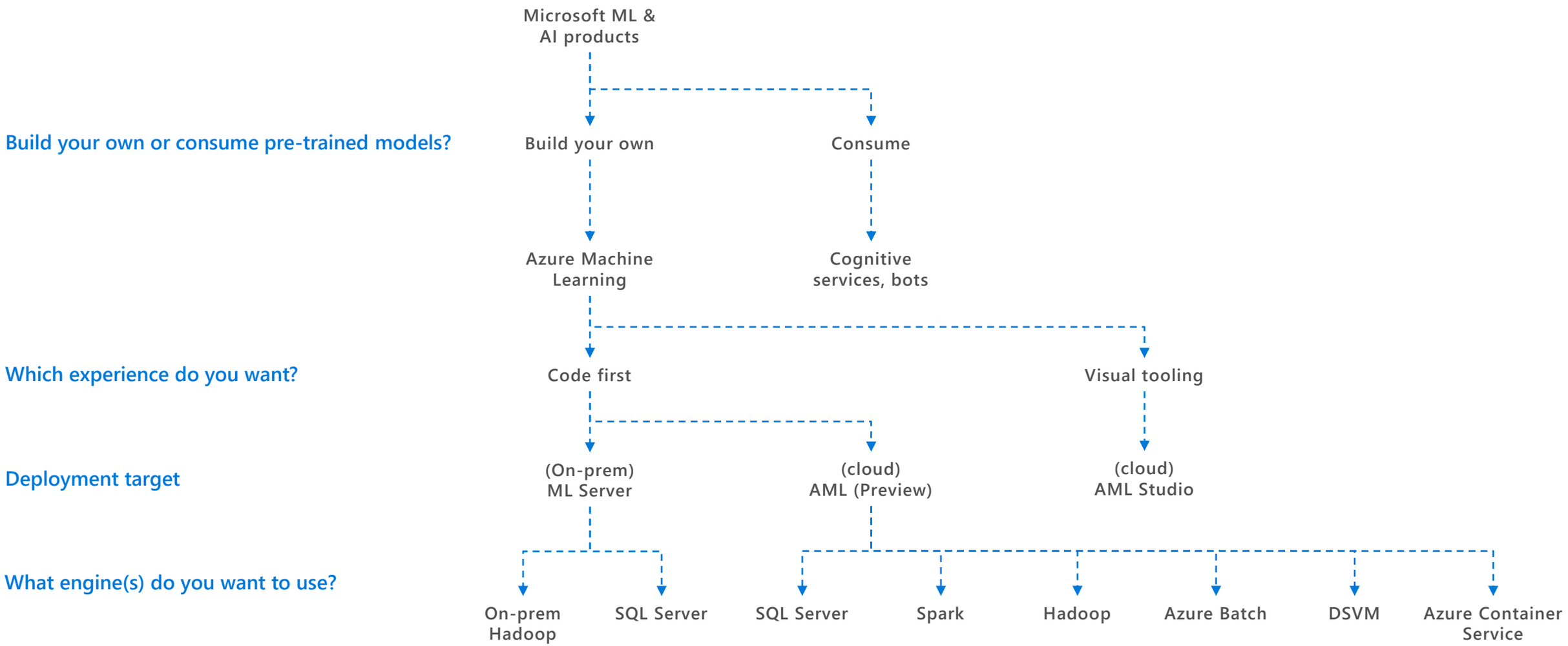




THINGS ARE IMPROVING FAST



NEW TOOLS MAKE IT EASIER TO DEPLOY



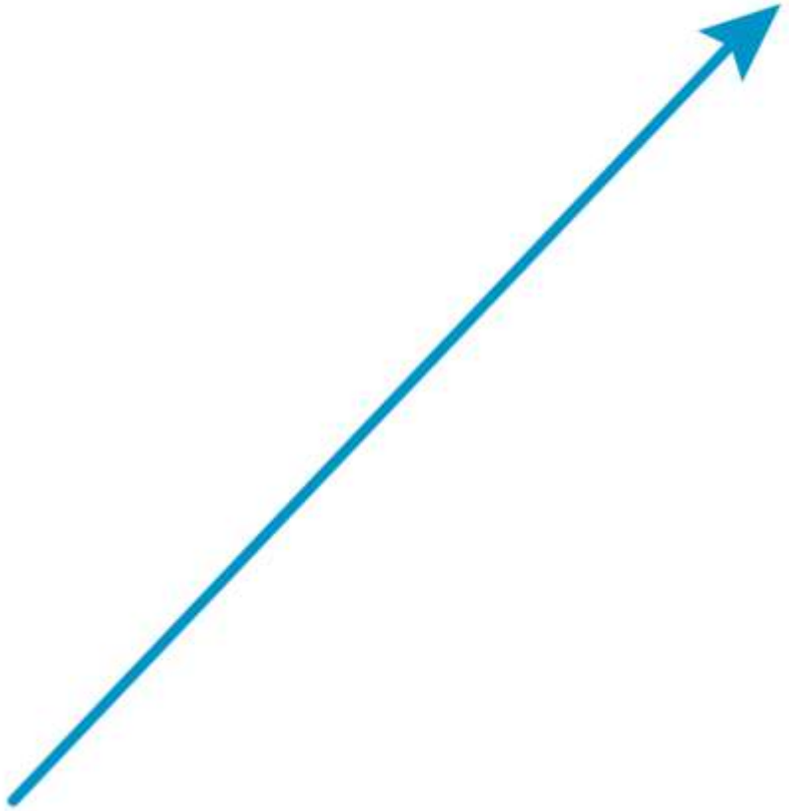
KEEP CALM AND CARRY ON

(YOU'LL FIGURE IT OUT)

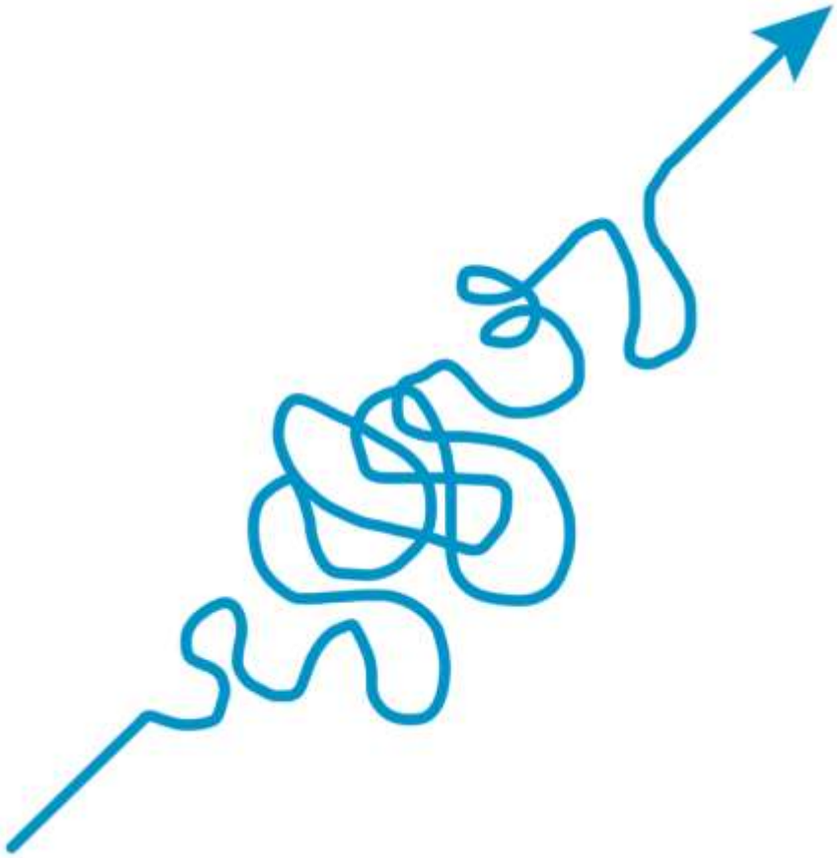
#7

MEASURING YOUR PROGRESS IN THE LONG RUN

SUCCESS



WHAT PEOPLE THINK IT LOOKS LIKE



WHAT IT ACTUALLY LOOKS LIKE

DEFINE YOUR OWN PERFORMANCE METRICS

Establish a **Qualitative Objective**

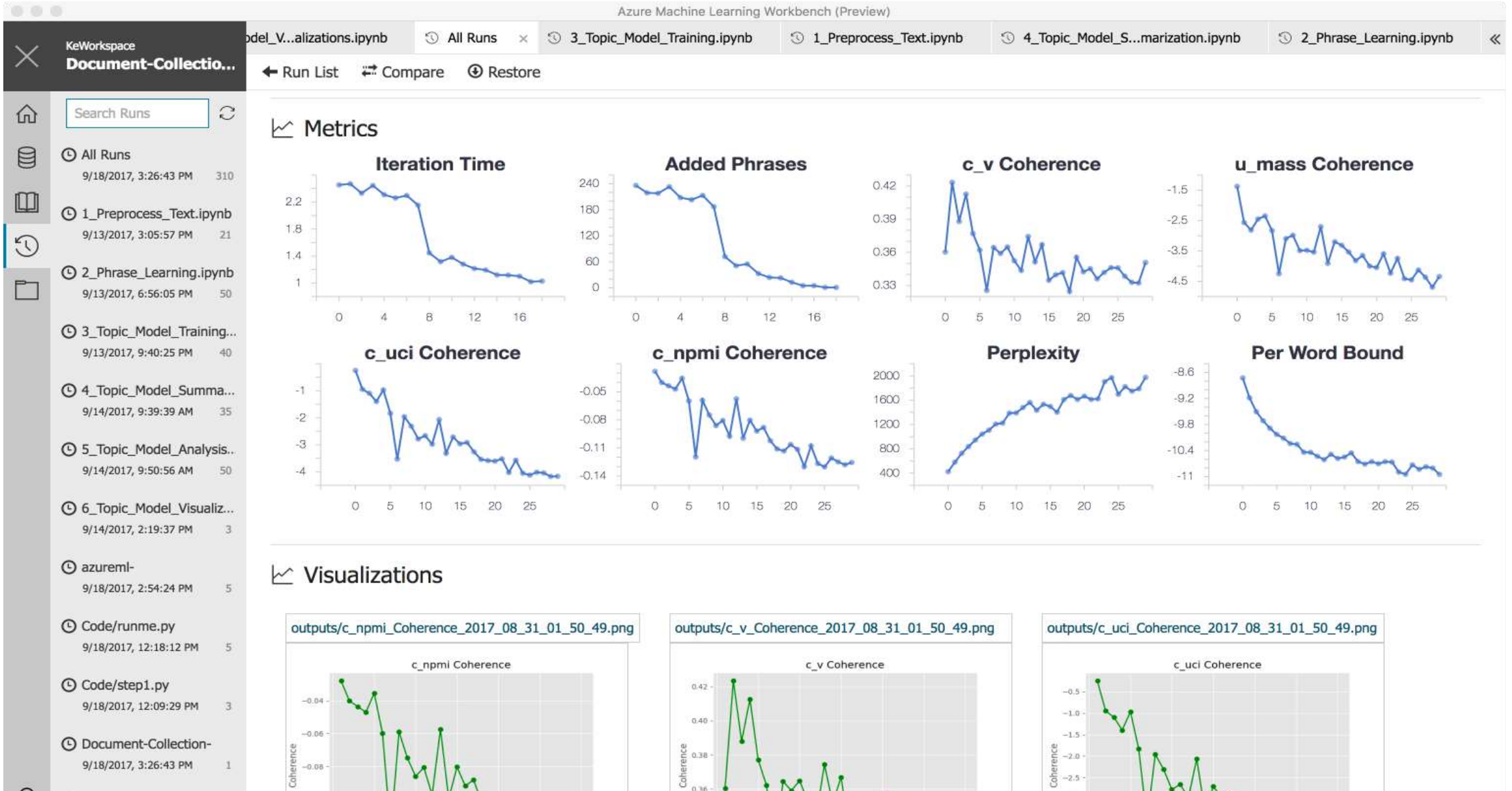
Translate into **Quantifiable Metric**

Quantify the **metric value improvement** for each iteration

Establish a **baseline**

Establish how to **measure the improvement** in the metric with the data science solution

PROFILING YOUR MODELS



MEASURED STEPS ARE THE WAY TO BIG LEAPS