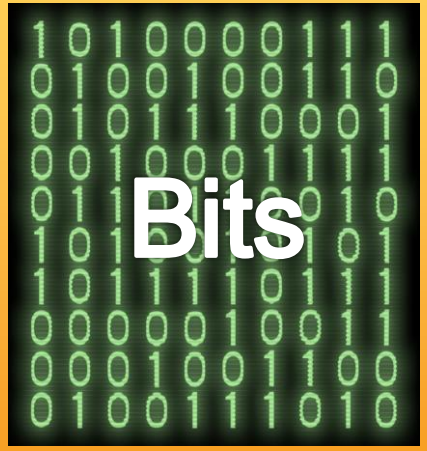
HelioCAMPUS
Better Data. Brilliant Insights.

What is Data Science?

Renée Teate | Director of Data Science, HelioCampus

What is Data?


Bits


Numbers
+740.21 -6.87
+122.56 -9.45
+140.04 -3.36
+180.98 -0.21


Text


Images

What is Data?

Generated & Collected

HelioCAMPUS

What is Data?

Analyzed and Visualized

Data-driven systems you likely interact with daily

Social Media Apps

Google searches and resources like Wikipedia

Streaming music and movies

Medical records and digital scans

Online Stores and Ratings

Bank databases that store transactions

Maps and navigational software

Any time you log into or even just visit any web page, it's tracked

Who builds these systems that:

Decide which social media posts, ads, songs, and movies to present to you? Determine your credit card interest rate? Predict what you want to type next? Highlight concerning areas on medical scans? Unlock your phone with your face? Navigate your self-driving car to your destination?

Data Scientists

Data Science is a combination of

Computer Science

- Summarizing data into the form required for model input
- Writing code to train, evaluate, and apply model
- Designing databases and interfaces
- Production deployment

Math and Statistics

- Exploring and analyzing data
- Understanding mathematical algorithms and selecting parameters
- Evaluating model results

Domain/Business Expertise

- Knowing what questions to ask
- Interpreting results
- Considering real-world implications

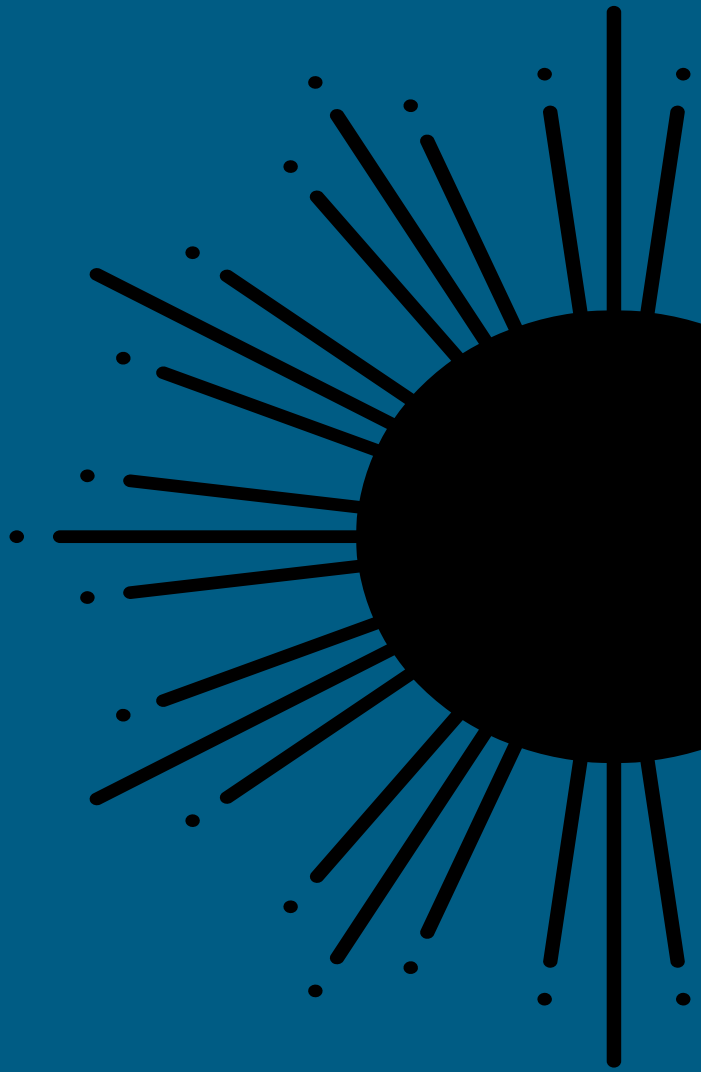
…and is increasingly seen as a "team sport", since it's rare for one person to have advanced skills in all of these areas

Data Scientists are sometimes also called

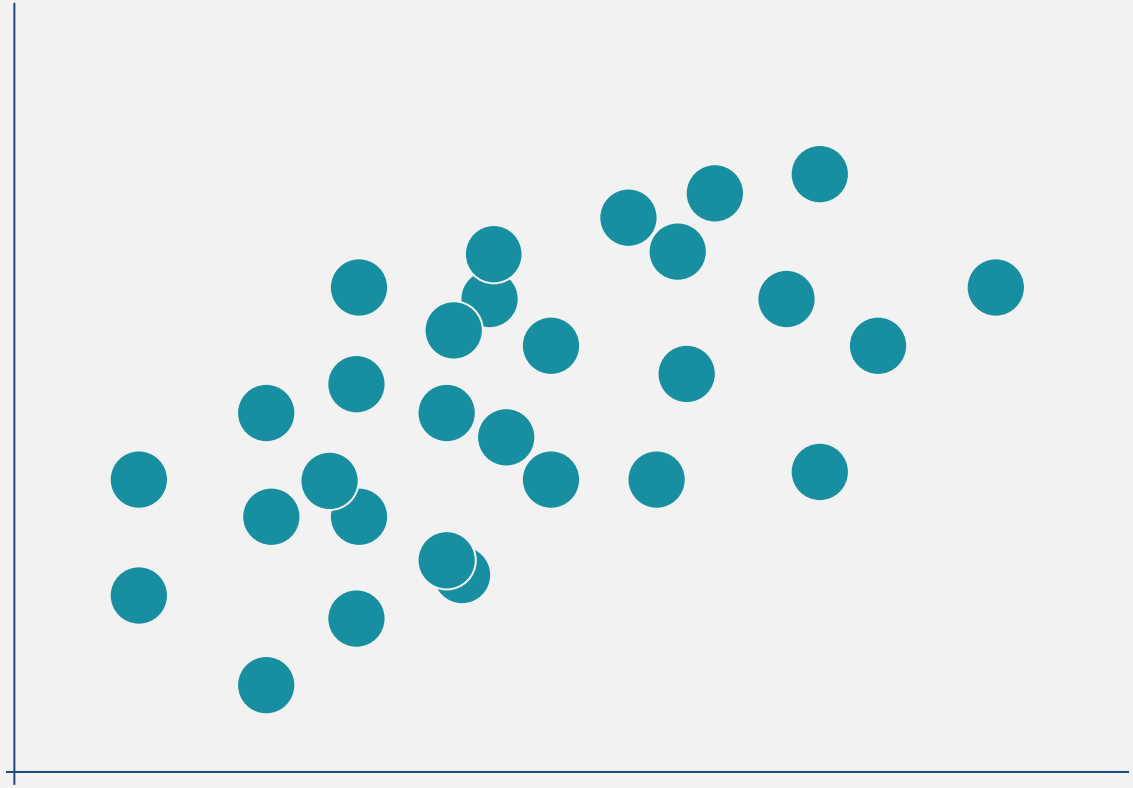
- Biostatistician
- GIS Analyst
- Natural Language Processing Researcher
- Neuroscientist
- Computer Vision Engineer
- Computational Physicist
- Financial Data Analyst
- Machine Learning Engineer
- Recommendation System Developer
- AI Researcher

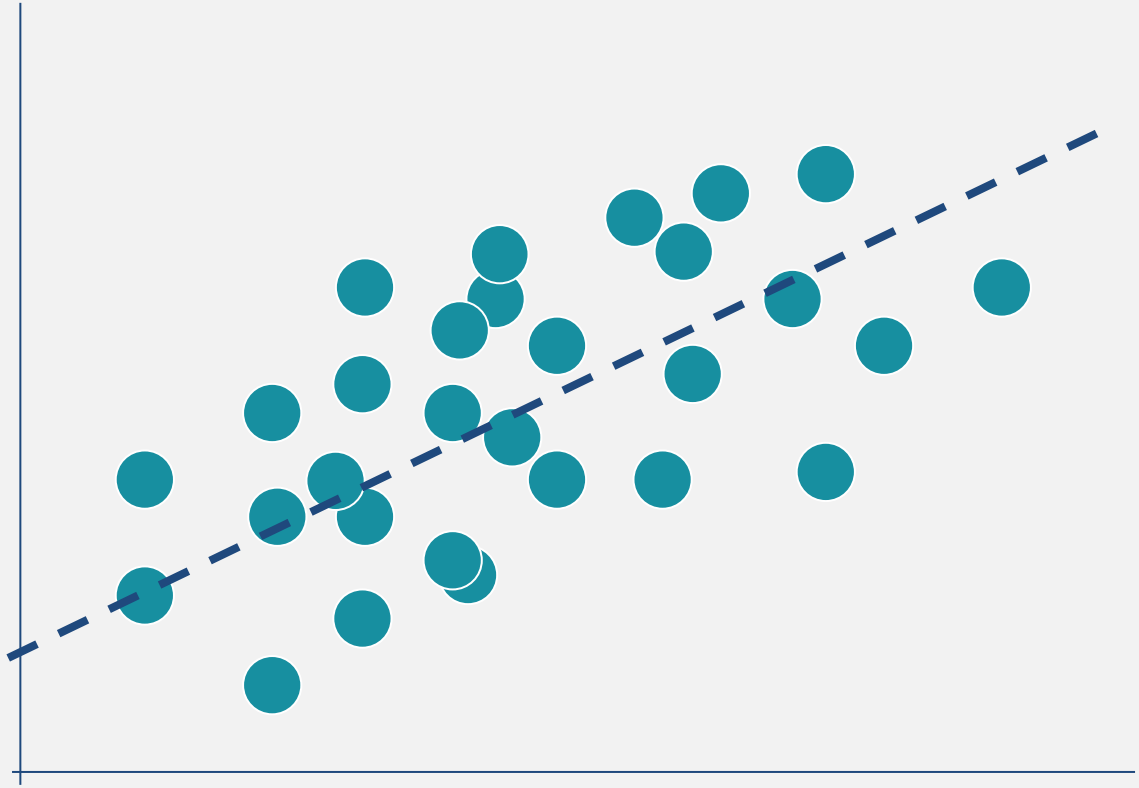
…there are data scientists working in just about every industry now

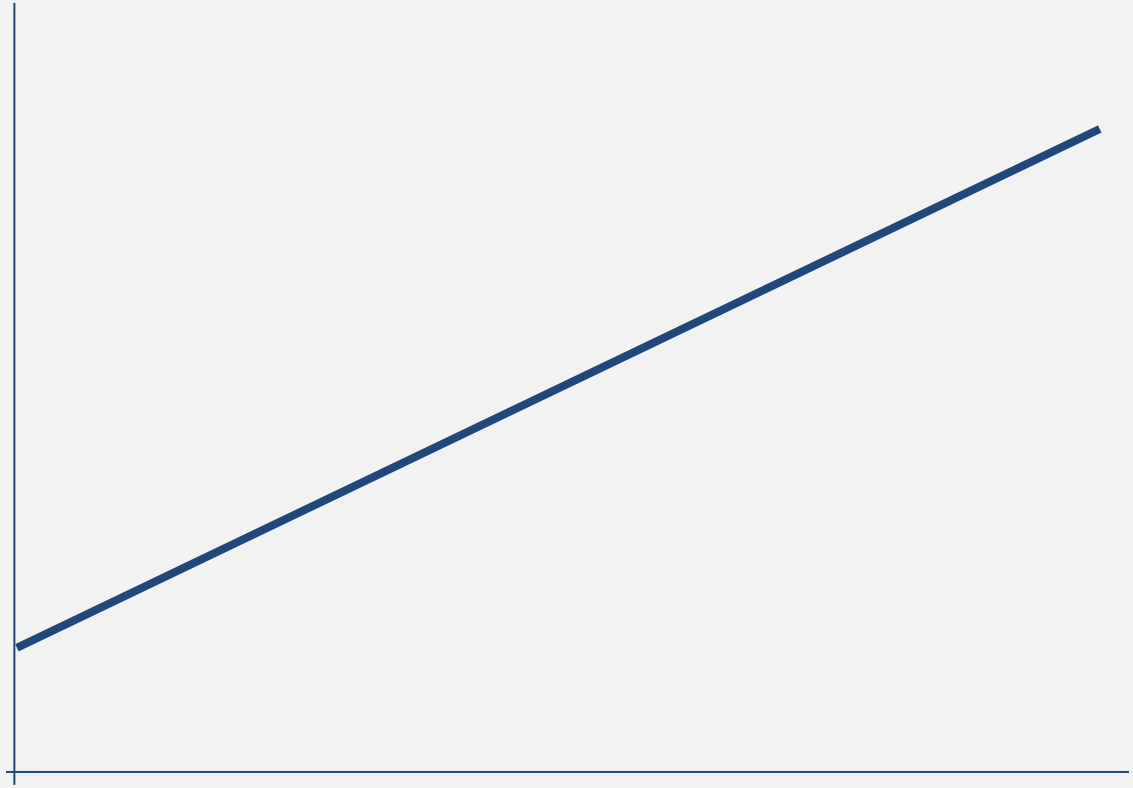
What is a predictive model?

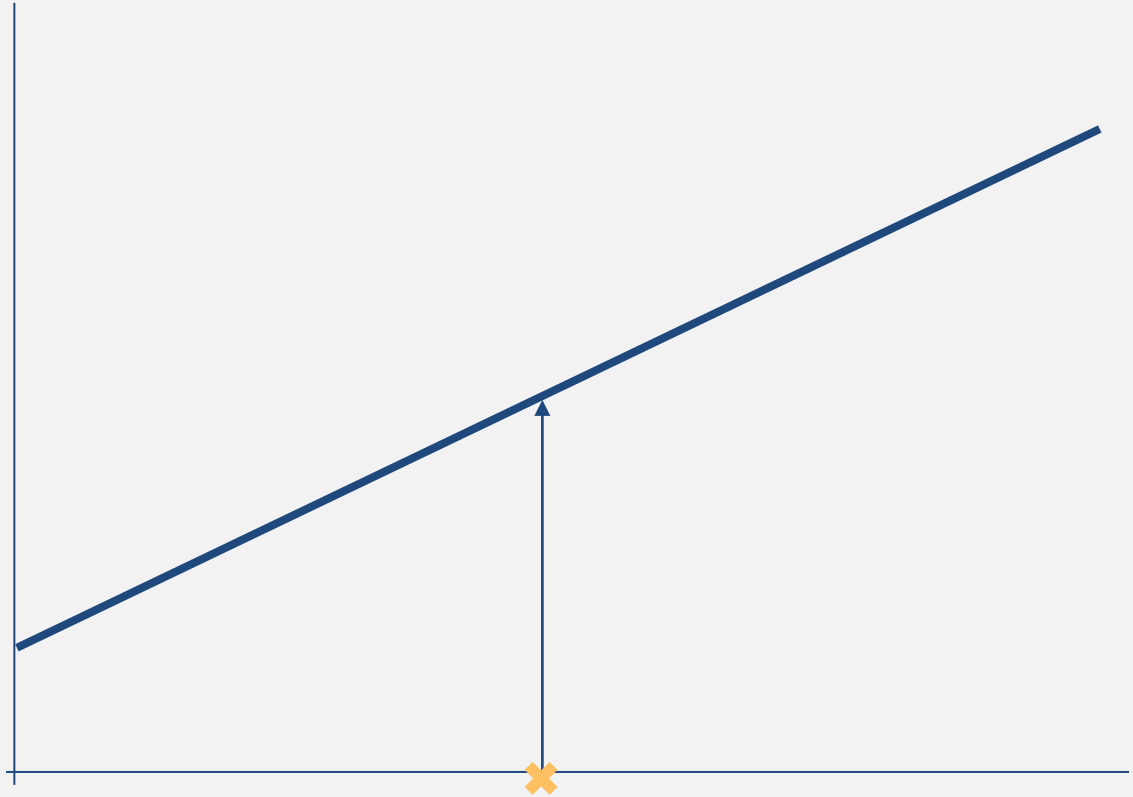


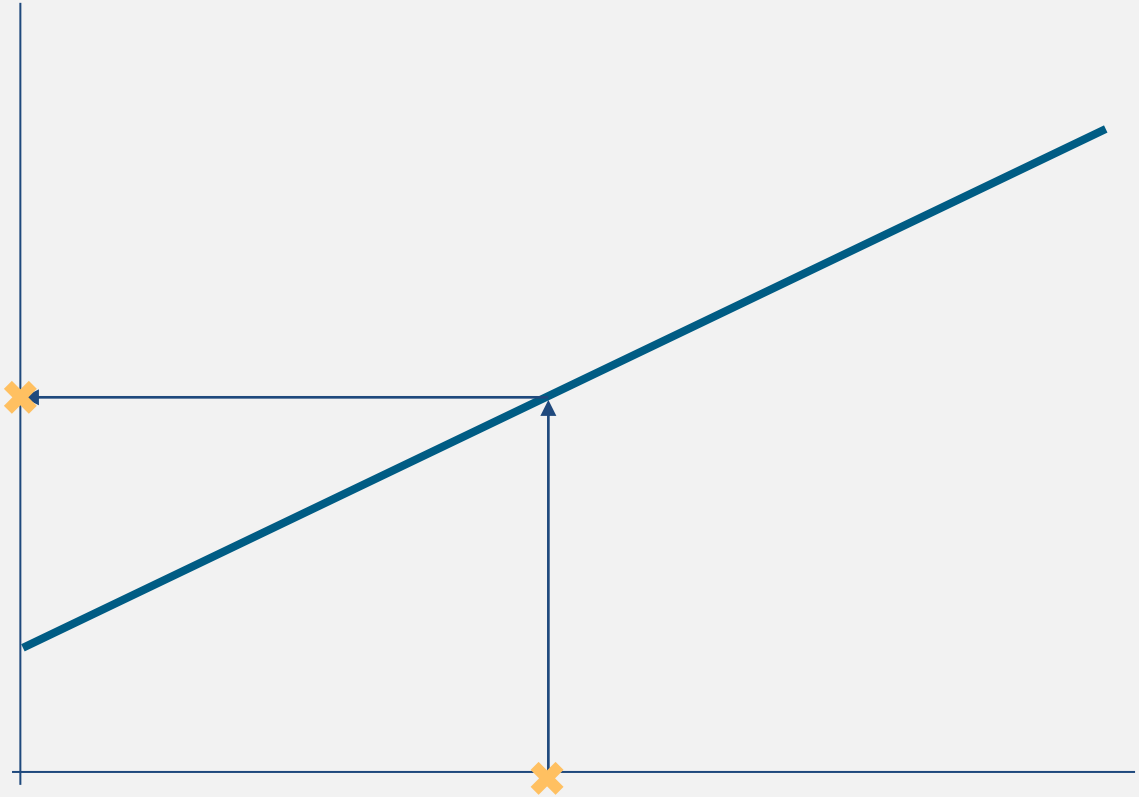
Data Science involves “training” machine learning algorithms to find patterns in existing data, and use those to predict what might happen in the future, or classify something into a category, or cluster similar things into groups.

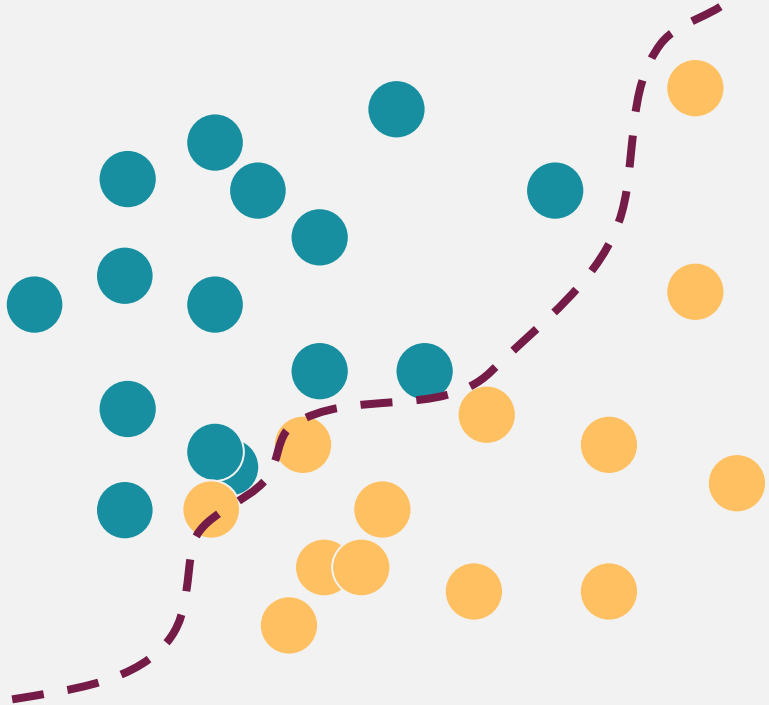


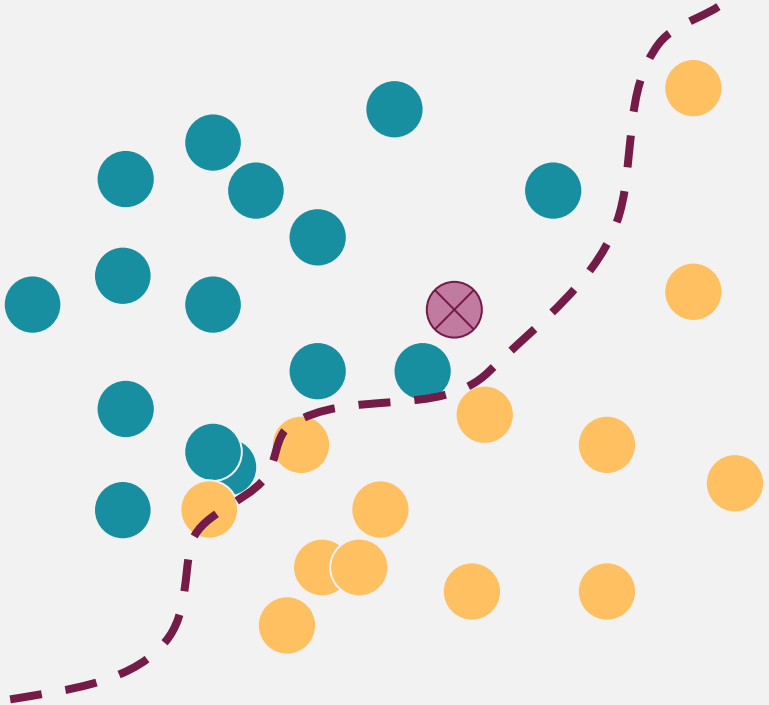


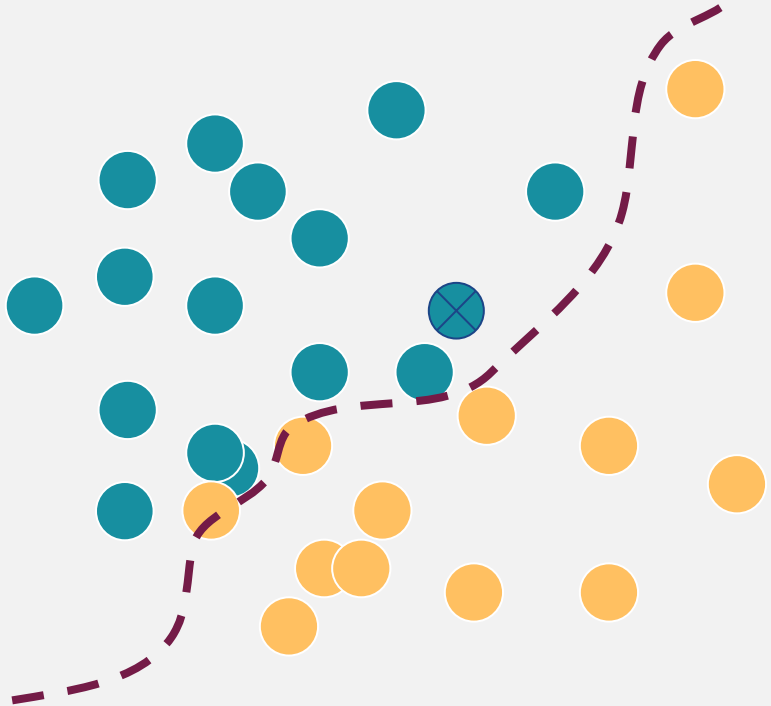












Model Evaluation: Confusion Matrix & Cost

Accuracy
What percentage of the predictions are correct?

If 99 people out of 100 don't have cancer, and there is a test that just always comes back negative (predicts that no one has cancer), that test is still 99% accurate, but has no "predictive power"

Other types of evaluation metrics can be used to optimize the model, minimize "cost" of error

	Has Cancer	Does Not Have Cancer
Tests Positive for Cancer	TRUE POSITIVE	FALSE POSITIVE
Tests Negative for Cancer	FALSE NEGATIVE	TRUE NEGATIVE

Model Evaluation: Confusion Matrix & Cost

Accuracy
What percentage of the predictions are correct?

If 99 people out of 100 don't have cancer, and there is a test that just always comes back negative (predicts that no one has cancer), that test is still 99% accurate, but has no "predictive power"

Other types of evaluation metrics can be used to optimize the model, minimize "cost" of error

	Has Cancer	Does Not Have Cancer
Tests Positive for Cancer	0	0
Tests Negative for Cancer	10	990

Some examples of predictive models you might see in the Higher Ed industry:

Time-Series models to predict how many students will matriculate next semester, based on counts of applications, deposits, and enrollments per week

Clustering models to identify peer institutions, or to group student populations or sets of courses into segments by similarity

Classification models to determine likelihood of an applicant to enroll, a first-time student to retain for 1 year, persist to the next term, or graduate within 4 years

Regression models to predict success metrics such as GPA

DS-OPS at HelioCampus

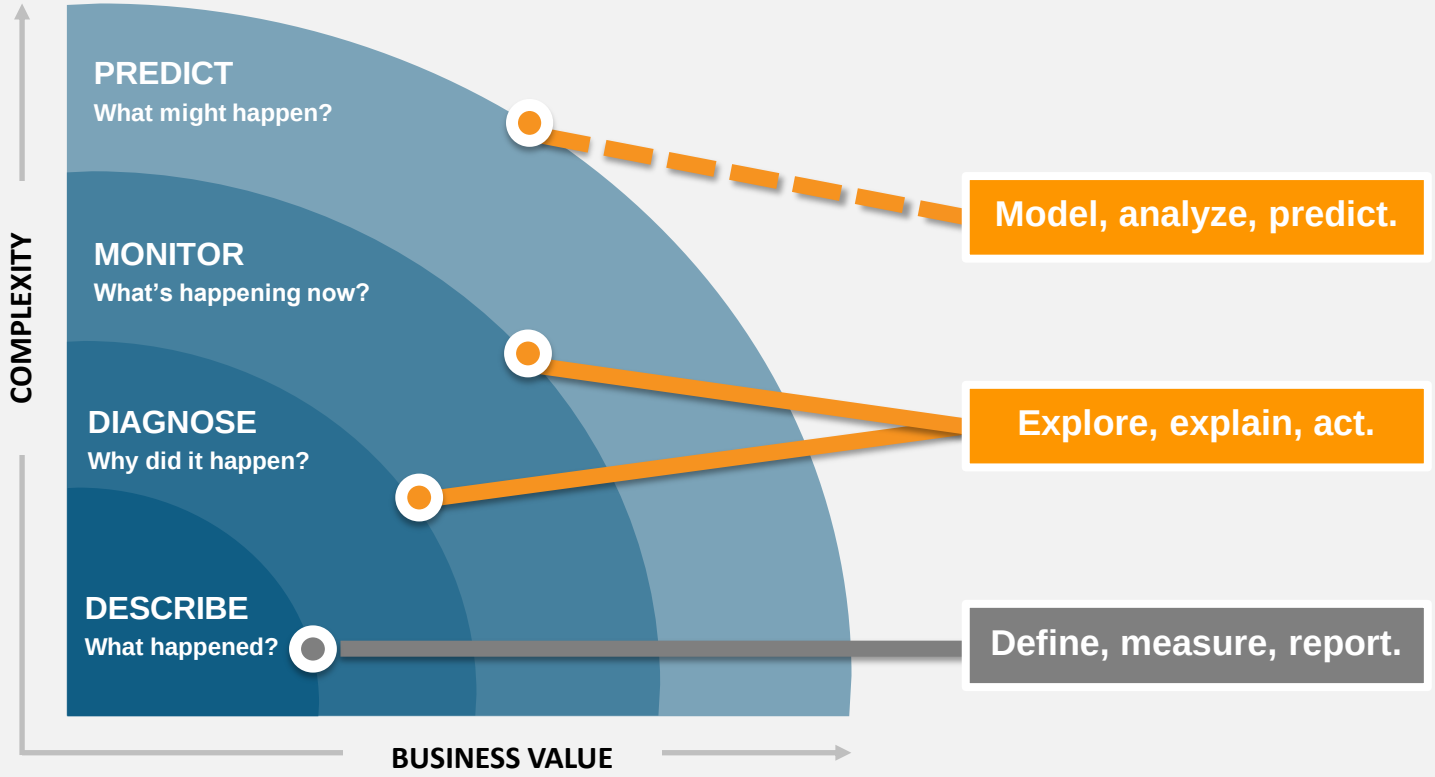
Renée
Teate

Rick
Ruiz

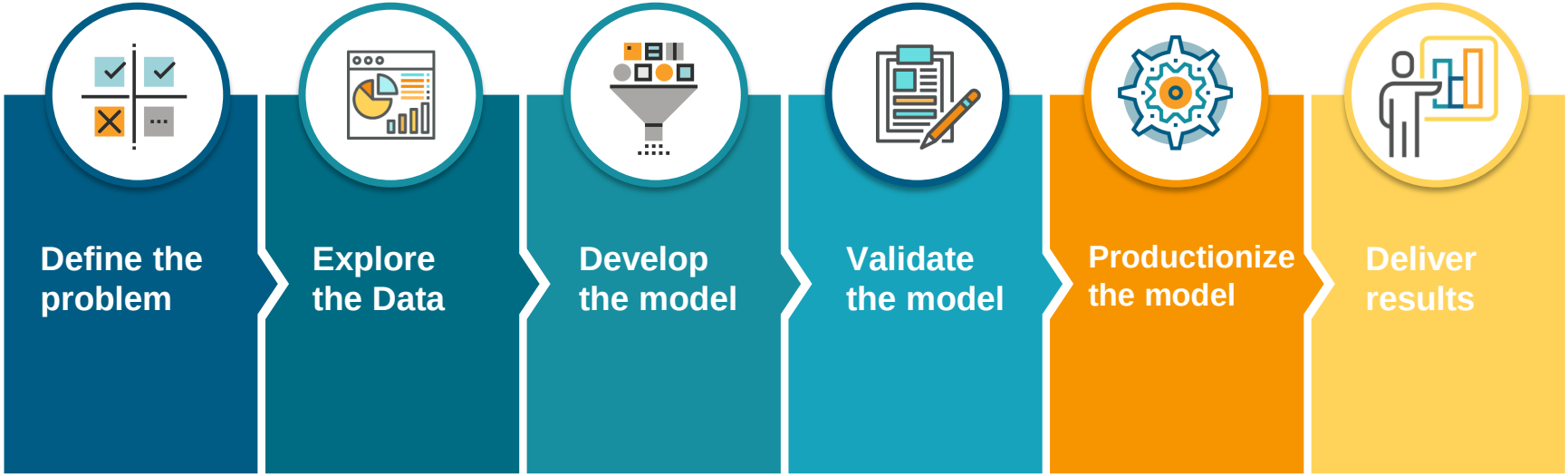
Karen
Heil

Brian
Richards

Finding Value in Analytics



Data science project flow



Step 1:
Define the question

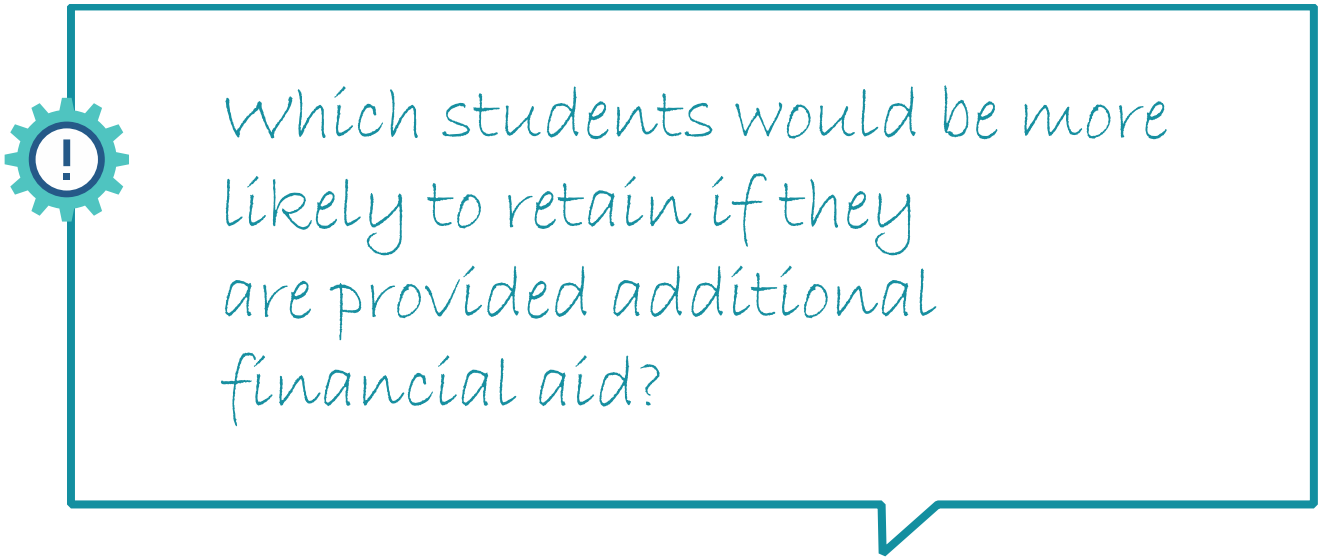
Step 1: Define the question

Data science needs a specific question to answer – the **outcome**

Which students are at risk of not retaining for 1 year?

Step 1: Define the question

The answer to the question should be *actionable*

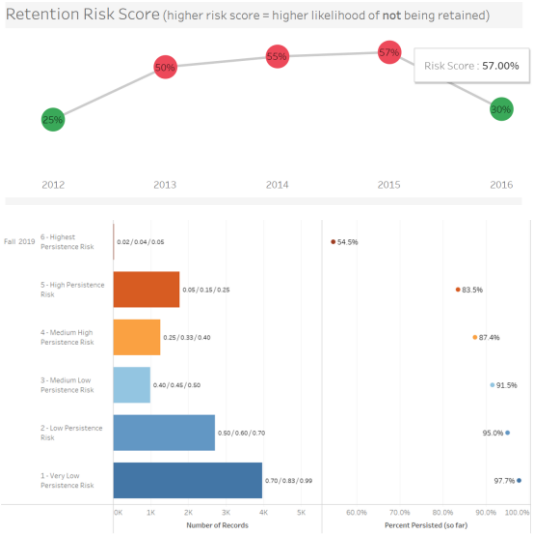


Step 1: Define the question

The **deliverable** of the model depends on the type of question

Model Scores

(such as predicted likelihood of each student retaining to next fall)

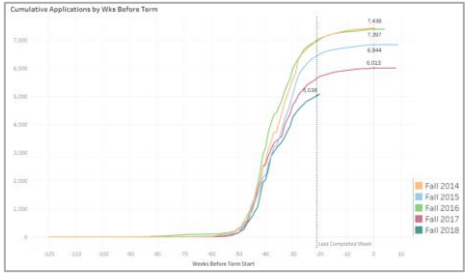


Top variables that impact the outcome

- Cumulative Institutional GPA
- Percent of Credits Completed
- Percent of Need Remaining Unmet
- Cumulative Credits Earned
- Cumulative Credits Attempted
- Term Type (Fall or Spring)
- Admissions HS GPA
- Grant/Scholarship Offer Amount (binned)
- Student Age

Forecasts

(such as number of full-time undergraduates expected to enroll next year)



HelioCAMPUS

Step 1: Define the question

The target group, whose past data is used to answer the question, will define how the model can be interpreted and applied

Which population are we most likely to have relevant data for, and be able to impact with the results?

Step 2:
Explore the Data

Step 2: Explore the data

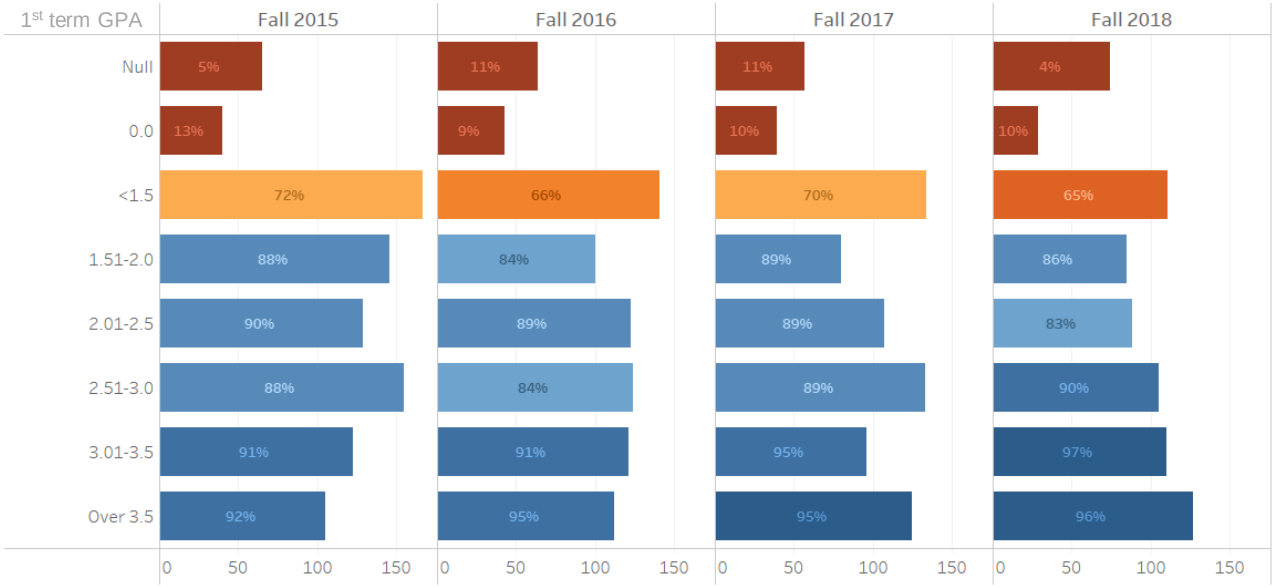
- Can the question be answered with the data available?

> If we only have data from 2016 to present, can we predict 4-year graduation likelihood of incoming students?

- Confirm that our initial descriptive statistics align with the institution's subject matter experts' experience and expectations on persistence and graduation rates
- Look for correlations between outcome variable and potential inputs

Step 2: Explore the data

Exploratory Data Analysis in Partnership with Subject Matter Experts at the Institution



HelioCAMPUS

Step 3:
Develop the model

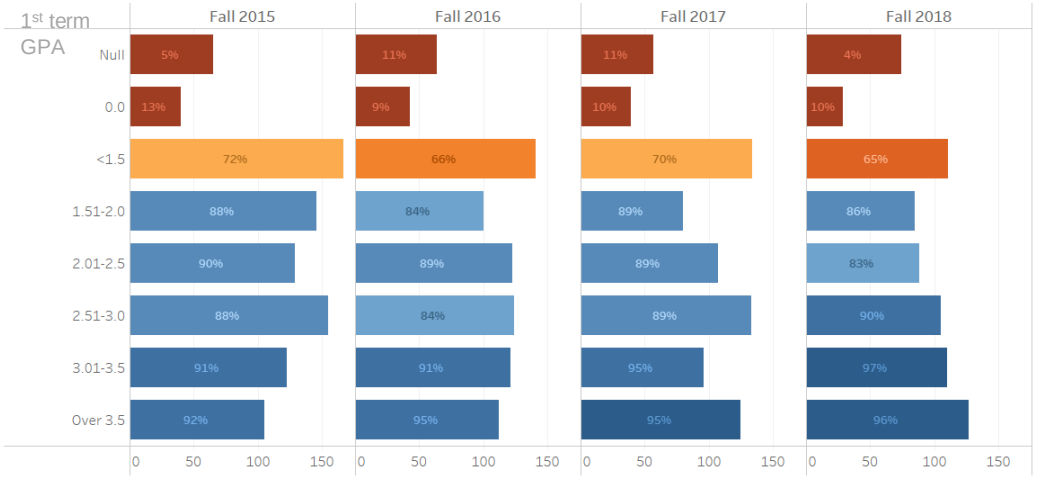
Step 3: Develop the model

Start building the “baseline” predictive model

- Feature Engineering
- Training/Testing/Evaluation on past year known outcomes
- Revisit “Explore” step as-needed to get additional inputs

Check with stakeholders to ensure that the variables chosen make sense

- Cumulative Institutional GPA
- Percent of Credits Completed
- Percent of Need Remaining Unmet
- Cumulative Credits Earned
- Cumulative Credits Attempted
- Term Type (Fall or Spring)
- Admissions HS GPA
- Grant/Scholarship Offer Amount (binned)
- Student Age



HelioCAMPUS

Step 4:
Validate the model

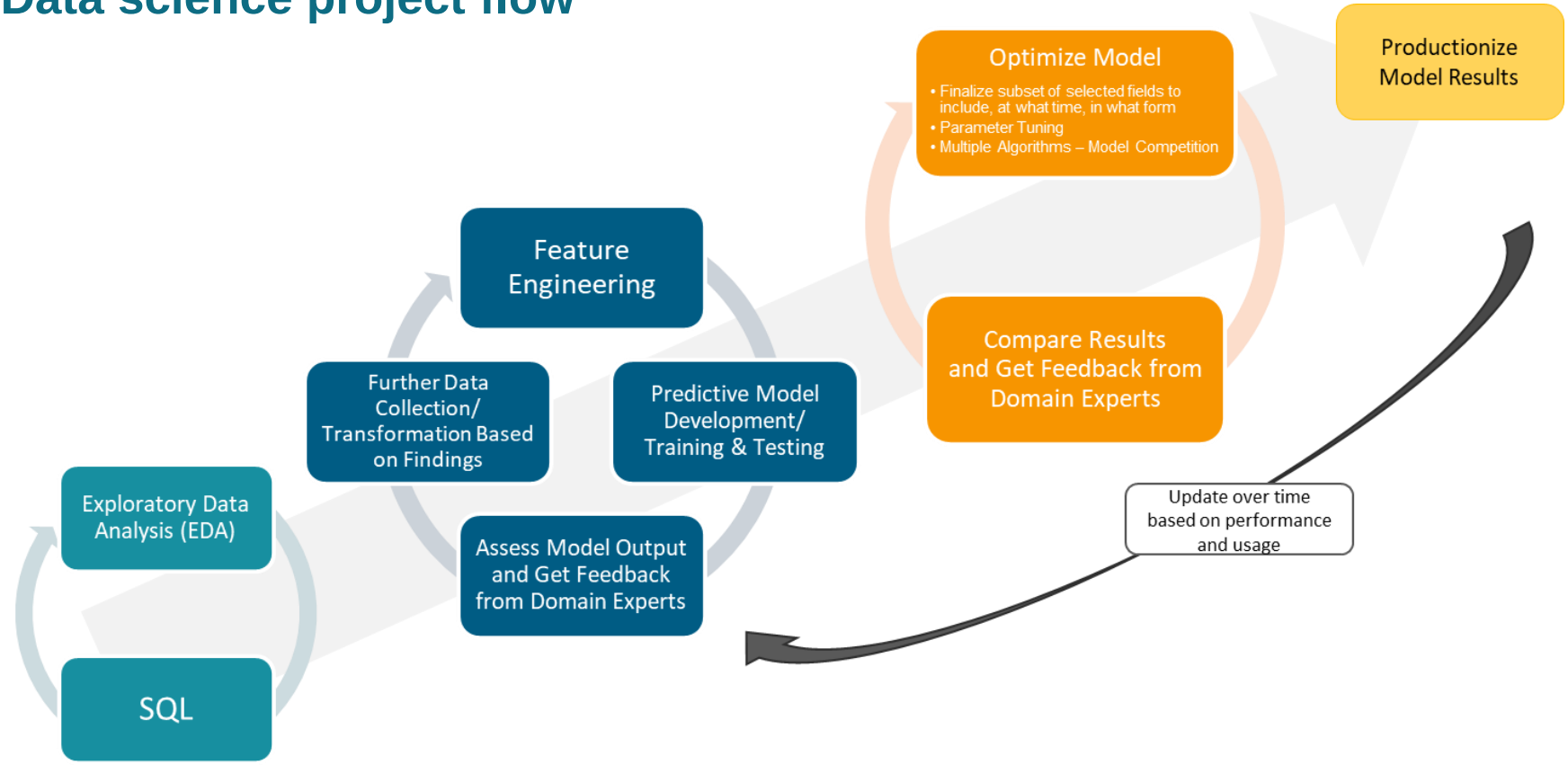
Step 4: Validate the model

- Using a hold-out dataset with known outcomes, check to see how well each model actually predicts – how well it “generalizes” to score previously-unseen records

  For example, a 1-year freshmen retention model may be trained/tested on a split dataset of combined of Fall 2016 and 2017 freshmen records, validated by predicting retention outcomes for Fall 2018 freshmen

- Tune the model parameters as-needed to improve results
- Understand where each model performs well and where it doesn’t

Data science project flow

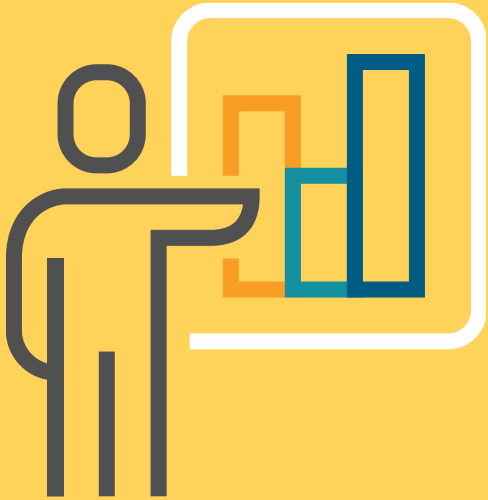


Step 5:
Productionize the model

Step 5: Productionize the model

- HelioCampus data scientists and data engineers collaborate to:
  - Add the score to the Redshift database (SQL)
  - Establish Tableau connection to the model table in database
- Perform ongoing validation of model and incorporate user feedback into subsequent versions

Step 5: Deliver the model



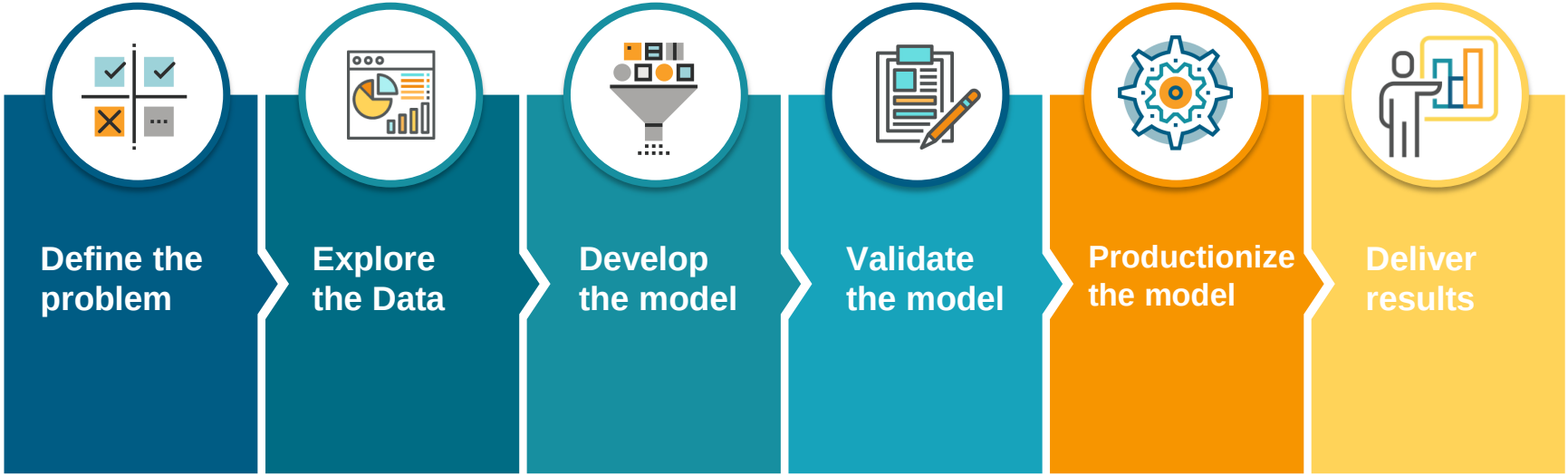
Step 5: Deliver the model

- Tableau dashboard for end-users to access the scores or forecasts
- Executive-level results and strategic recommendations based on the findings from the modeling process
- Additional training and materials as needed

What happens when the patterns change dramatically?

Is a model trained to predict student retention during “normal” years still useful during a pandemic?

We partner with you at every step



Questions?

renee.teate@heliocampus.com
@becomingdatasci on twitter