

# Data Science Methodology

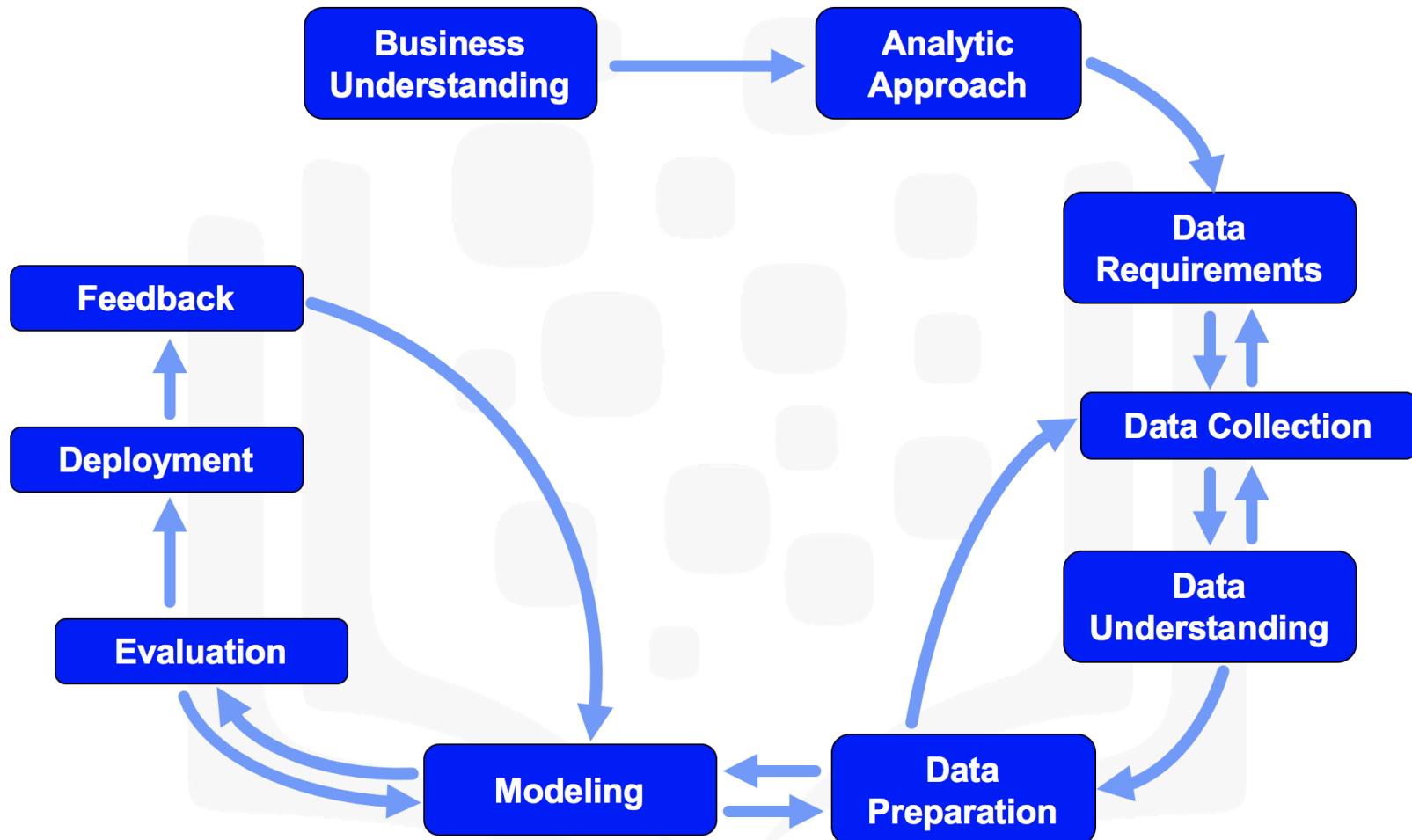
Data Science for Mathematicians

# Learning Objectives

In this course you will learn about:

- The major steps involved in tackling a data science problem.
- The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment.
- How data scientists think through tackling interesting real-world examples.

# Data Science Methodology



# Module 1

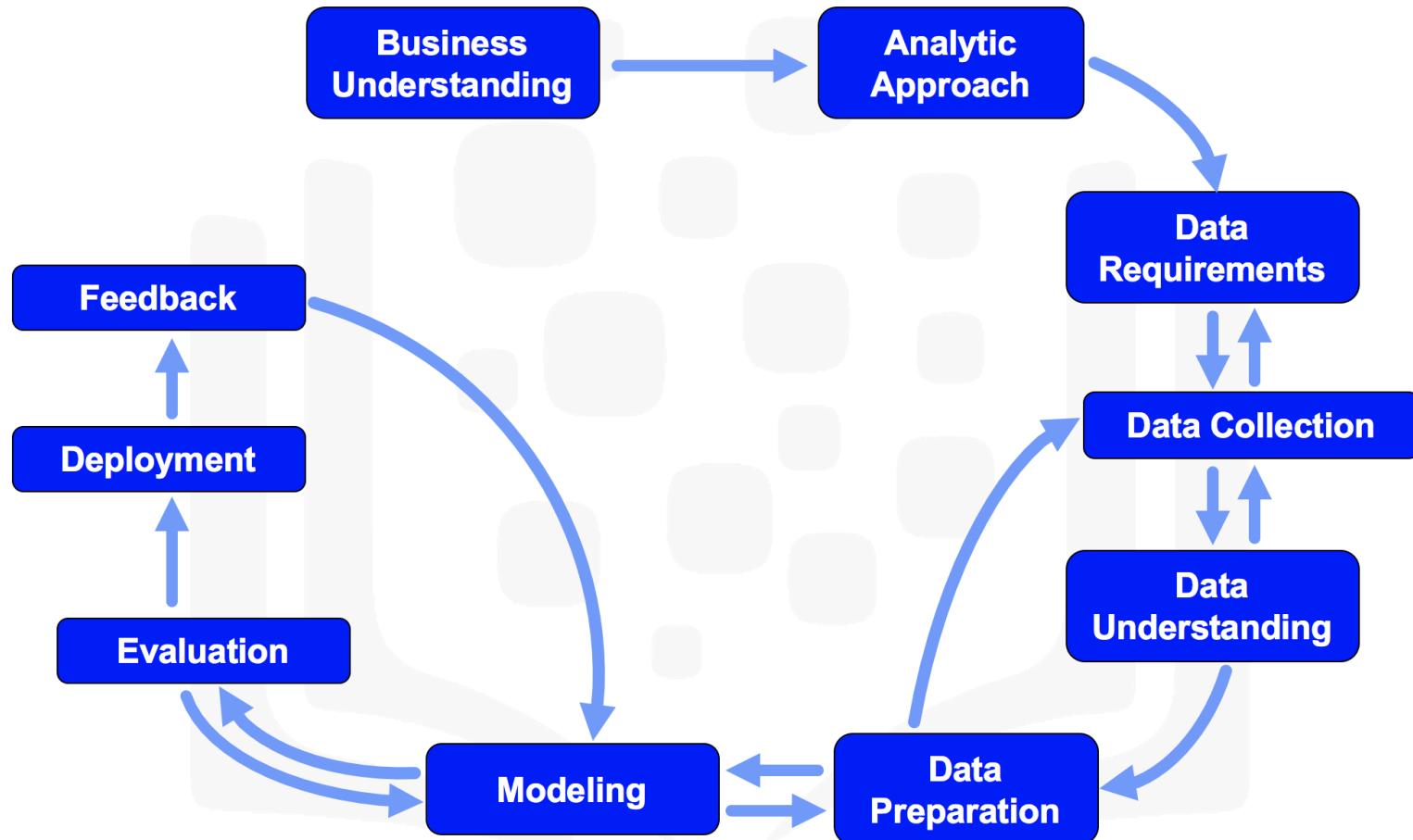
From Problem to Approach

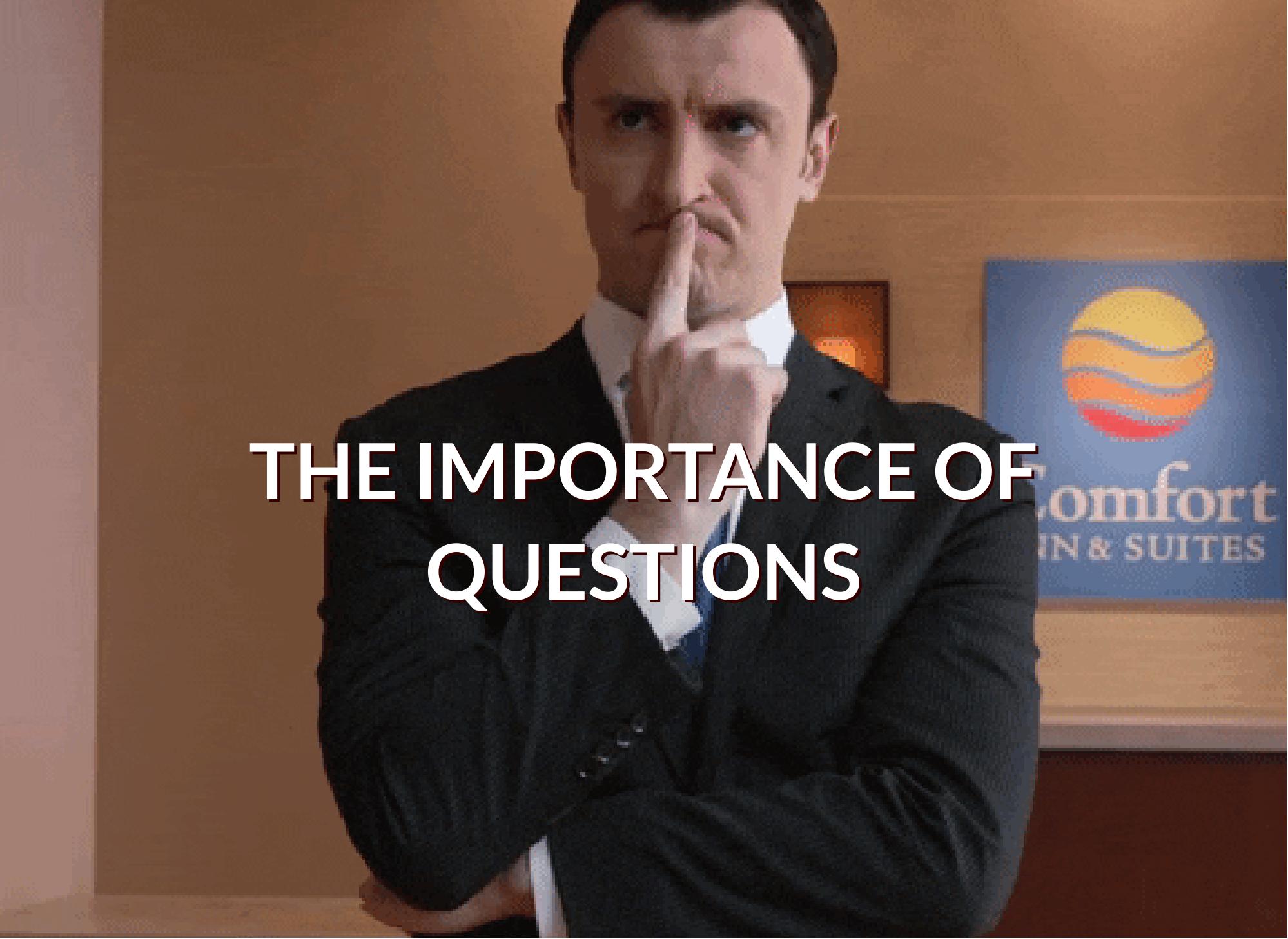
# Learning Objectives

In this lesson you will learn about:

- Why we are interested in data science.
- What a methodology is, and why data scientists need a methodology.
- The data science methodology and its flowchart.
- How to apply business understanding and the analytic approach to any data science problem.

# Business Understanding



A man in a dark suit and white shirt is standing in front of a brown wall. He is holding a blue and silver microphone in his right hand and has his left hand raised to his mouth, with his index finger touching his lips, indicating silence or a secret. To his right, there is a blue sign for "Comfort Inn & Suites" featuring a stylized orange and yellow swirl logo.

# THE IMPORTANCE OF QUESTIONS

# From understanding to approach

## Business Understanding

- What is the problem that you are trying to solve?



## Analytic Approach

- **How can you use data to answer the question?**





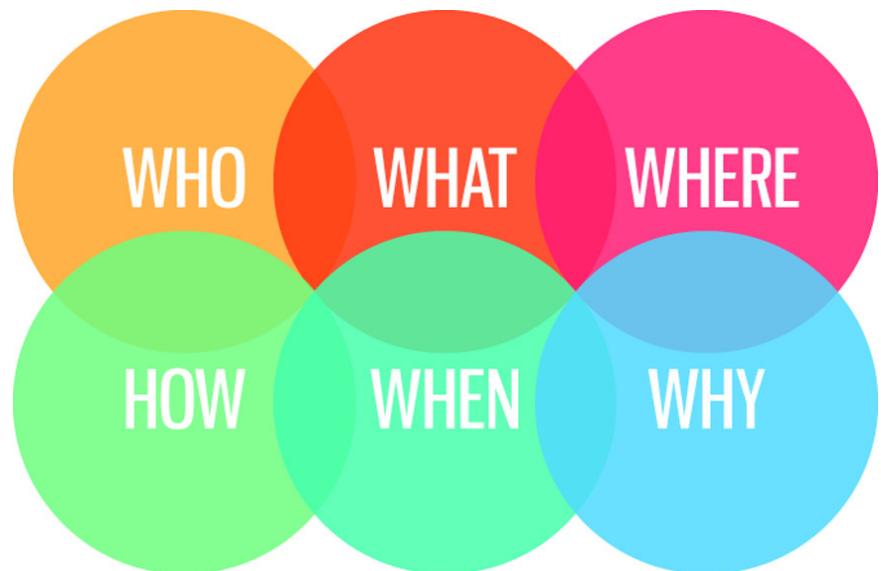
WHAT'S THE GOAL?

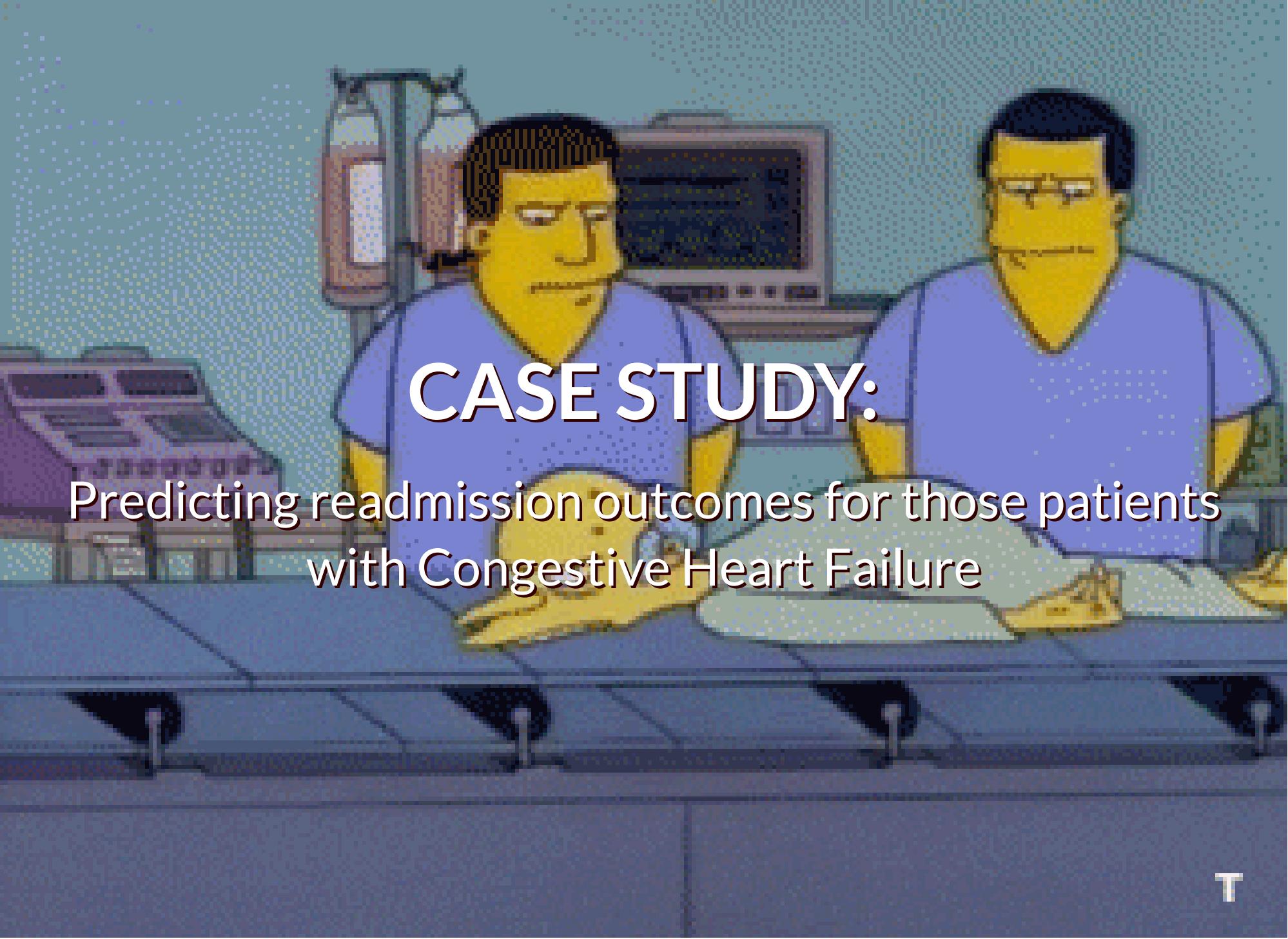


**SUPPORTING THE GOAL**



# Getting Stakeholders buy-in and support

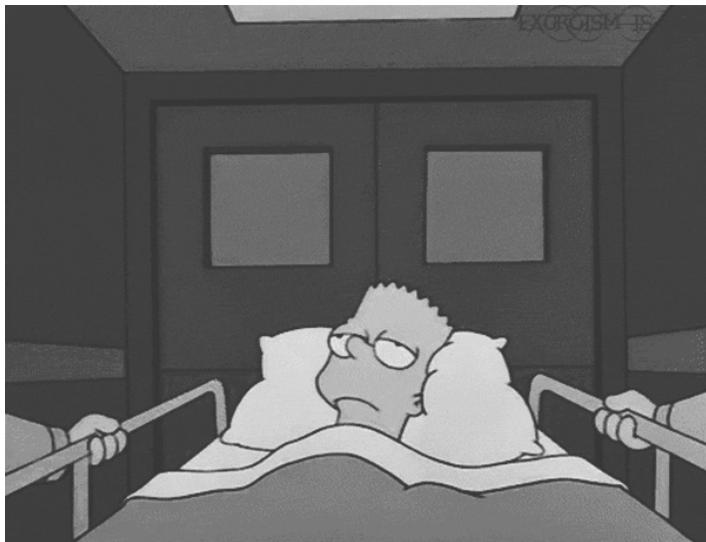


A photograph of a medical office or hospital room. Two male doctors in blue scrubs are standing behind a desk. The doctor on the left is looking down at a computer monitor, while the doctor on the right is looking towards the camera. There are medical charts and a computer keyboard on the desk. In the background, there are shelves with medical supplies and a window.

# CASE STUDY:

Predicting readmission outcomes for those patients  
with Congestive Heart Failure

# Case Study: CHF Re-admission



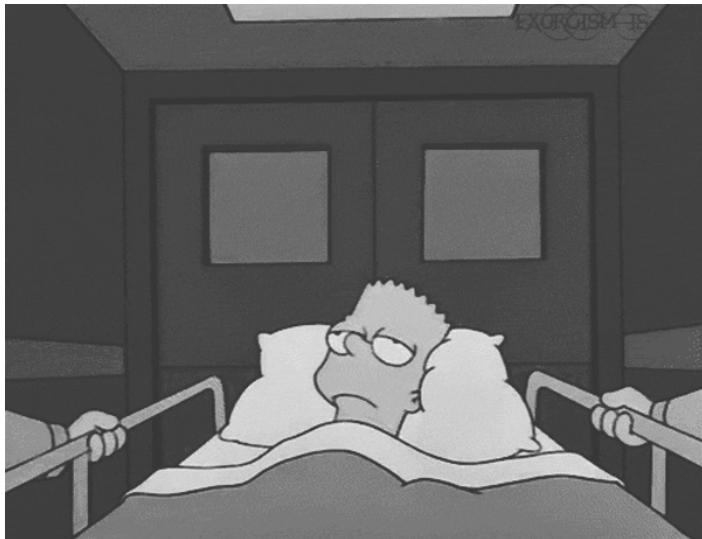
Define the **GOAL**

- To provide the care without increasing cost

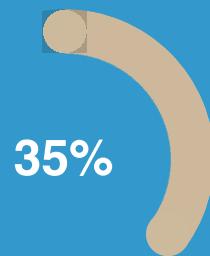
Define the **OBJECTIVE**

- To review the process to identify inefficiencies

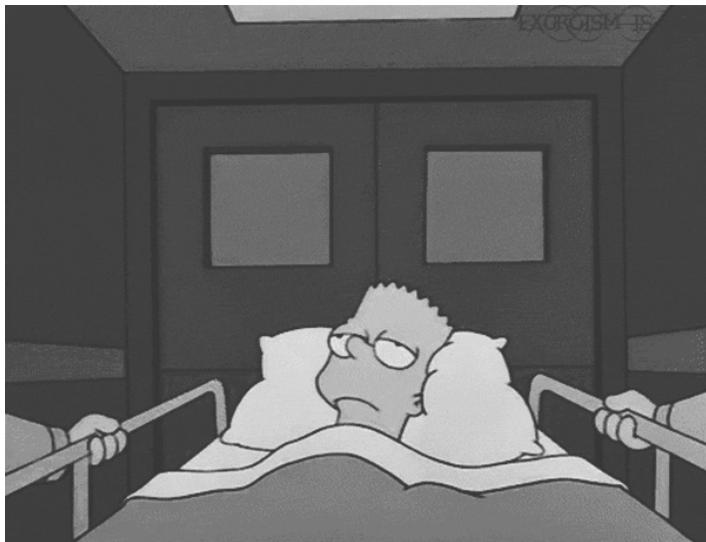
# Case Study: CHF Re-admission



Approximately **25-35%** of individuals who finish rehab treatment would be readmitted to a rehab center within **one year**; and that **50%** would be readmitted within **five years**.

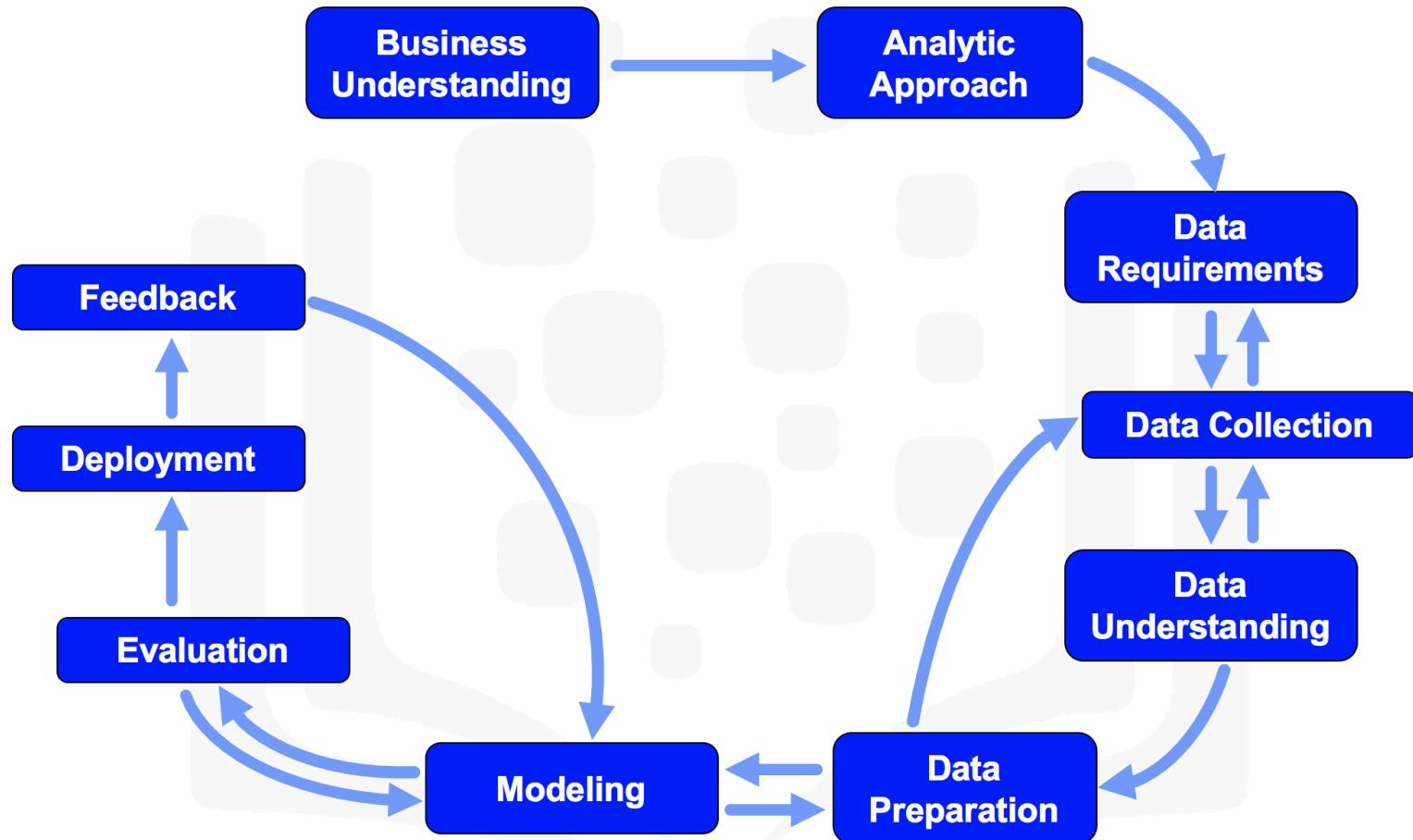


# Case Study: CHF Re-admission



1. Predicting CHF readmission outcomes (Y or N) for each patient.
2. Predicting readmission risk.
3. Understanding the combination of events that led to the predicted outcome
4. Applying an easy-to-understand process to new patients, regarding their readmission risk.

# Analytic Approach



# From understanding to approach

## Business Understanding

- What is the problem that you are trying to solve?



## Analytic Approach

- How can you use data to answer the question?



# Pick analytic approach based on type of question



<b>DESCRIPTIVE</b>
Current status
<b>DIAGNOSTIC (STATISTICAL ANALYSIS)</b>
What happened? Why is this happening?
<b>PREDICTIVE (FORECASTING)</b>
What if this trends continue? What will happen next?
<b>PRESCRIPTIVE</b>
How do we solve it?

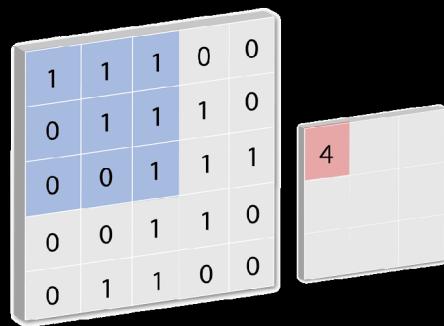
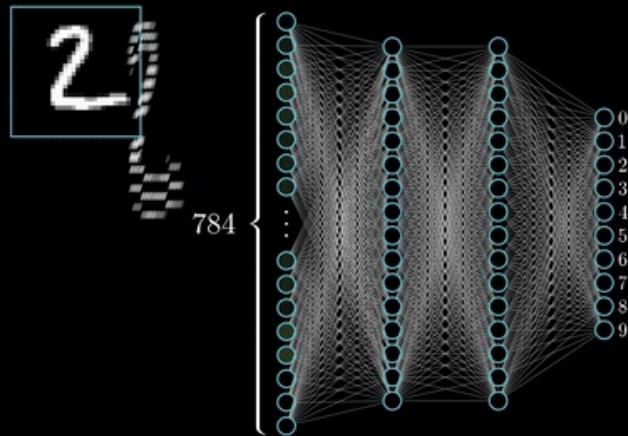
# What are the types of questions?

- If the question is to determine **probabilities** of an action
  - Use a **predictive model**
- If the question is to show **relationships**
  - Use a **descriptive model**
- If the question requires a yes/no answer
  - Use a **classification model**



The correct approach depends on business requirements for the model

# Will machine learning be utilized?

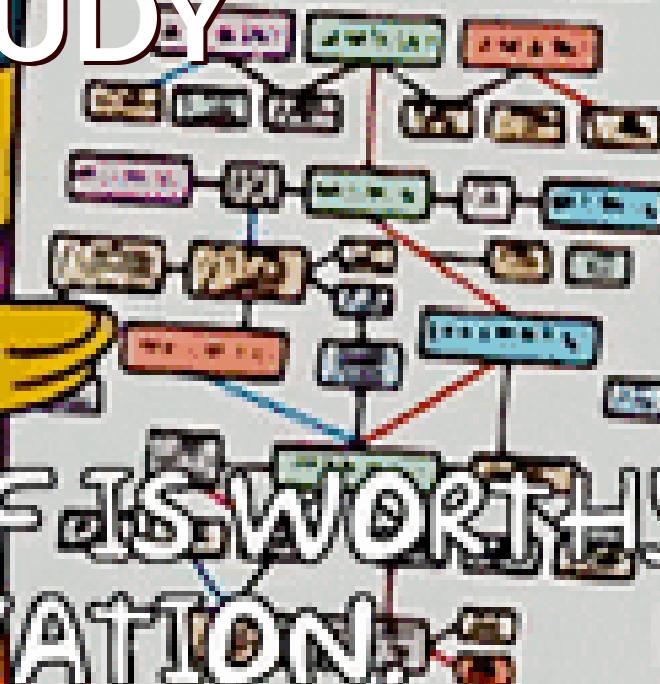


- Machine Learning
  - Learning without being explicitly programmed
  - Identifies relationships and trends in data that might otherwise not be accessible or identified
  - Use clustering association approaches

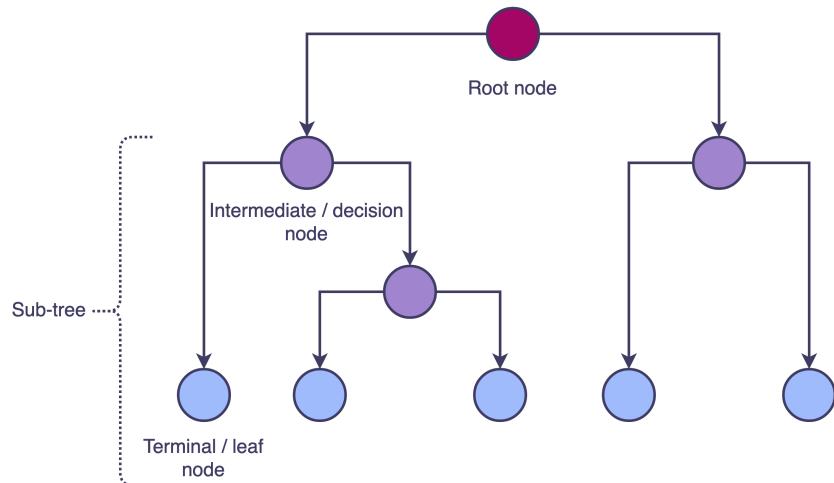
WHICH IN ITSELF IS WORTHY  
OF PRESERVATION.

## CASE STUDY

### DECISION TREE



# Case study: Decision tree classification selected!



- Predictive model
  - To predict an outcome
- Decision tree classification
  - Catalogical outcome
  - Explicit decision path showing conditions leading to high risk
  - Likelihood of classified outcome
  - Easy to understand and apply

# Module 2

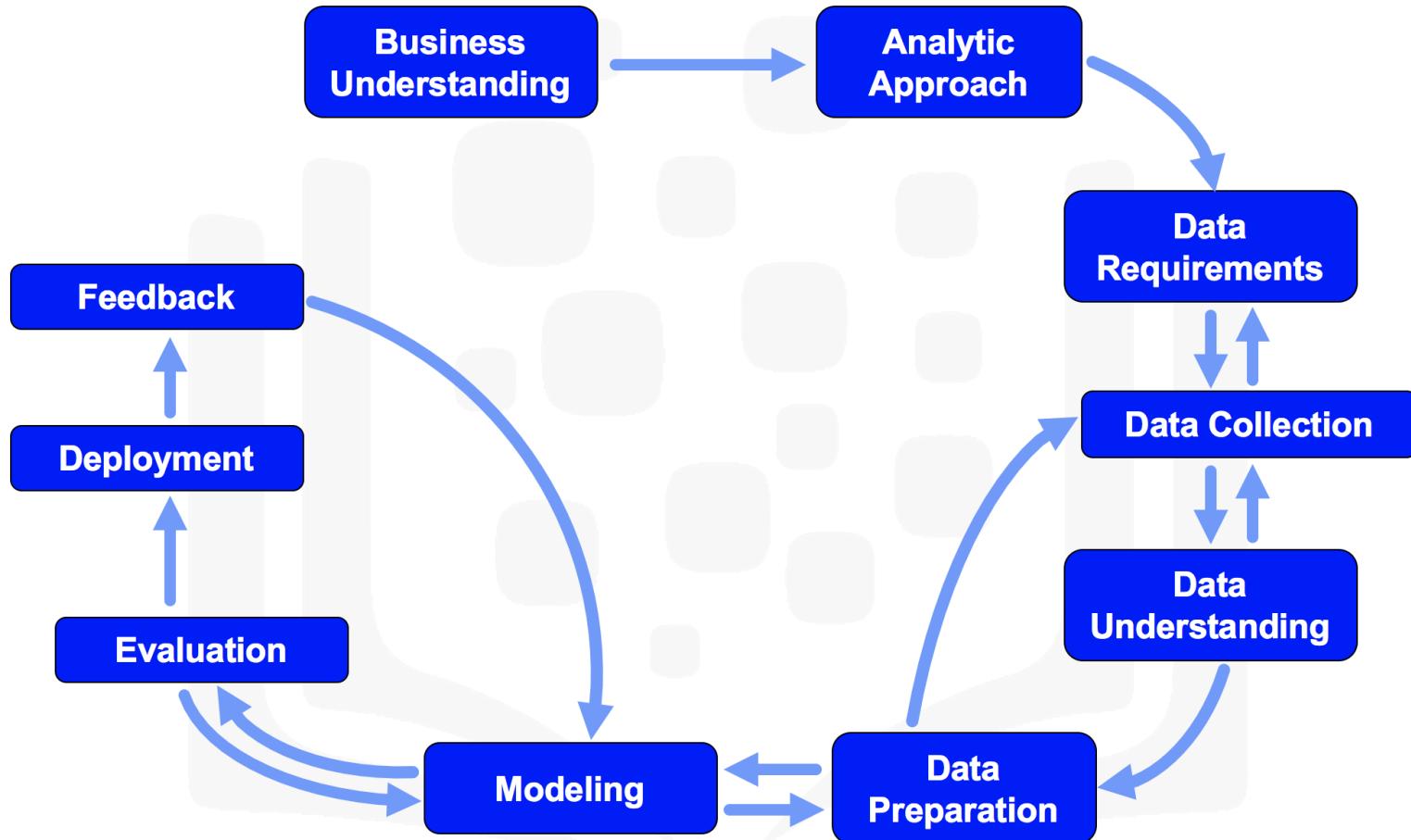
From Requirements to Collection

# Learning Objectives

In this lesson you will learn about:

- Data requirements and data understanding.
- What occurs during data collection.
- How to apply data requirements and data collection to any data science problem.

# Data Requirements



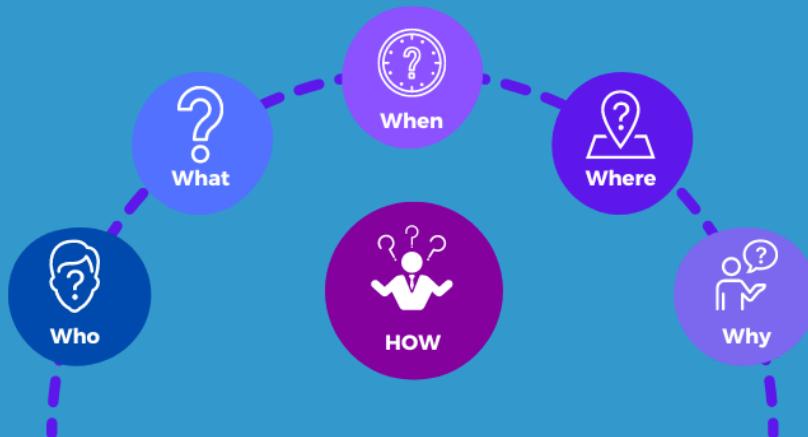
A professional chef in a white uniform and tall white toque is stirring a large, bubbling pot of what appears to be a creamy, yellowish soup or sauce over a gas burner. The kitchen background is filled with stainless steel equipment, including ovens and a large walk-in refrigerator.

# DATA REQUIREMENTS >>> COOKING WITH DATA <<<

# From requirements to collection

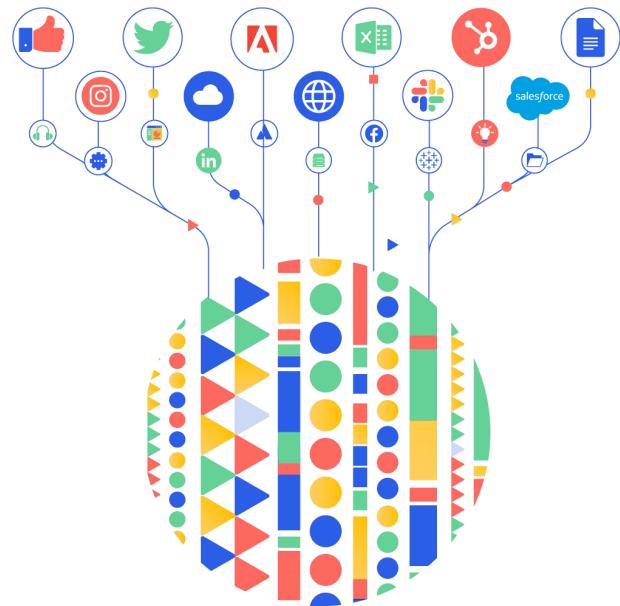
## Data Requirements

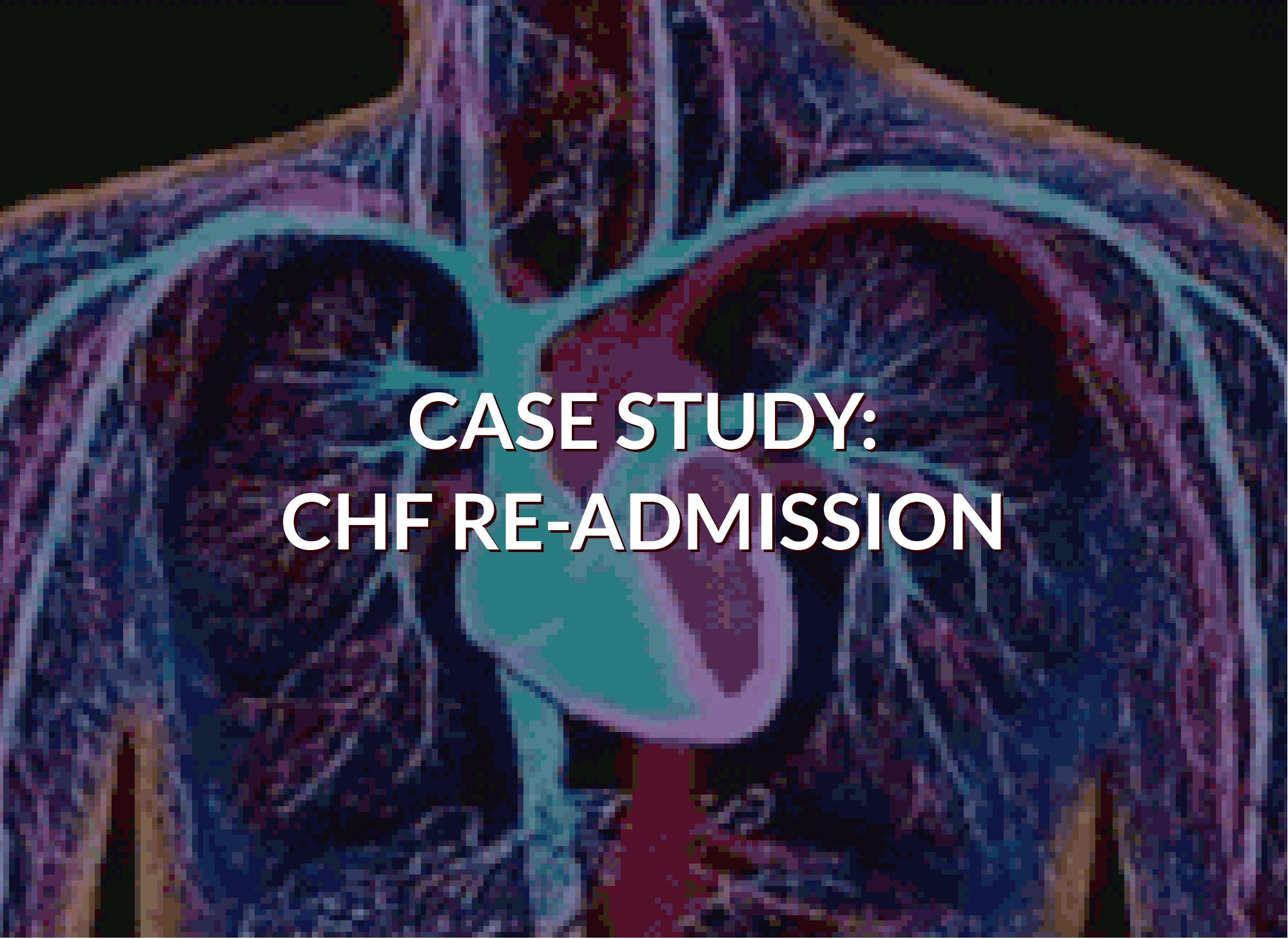
- What are data requirements?



## Data Collection

- What occurs during data collection?

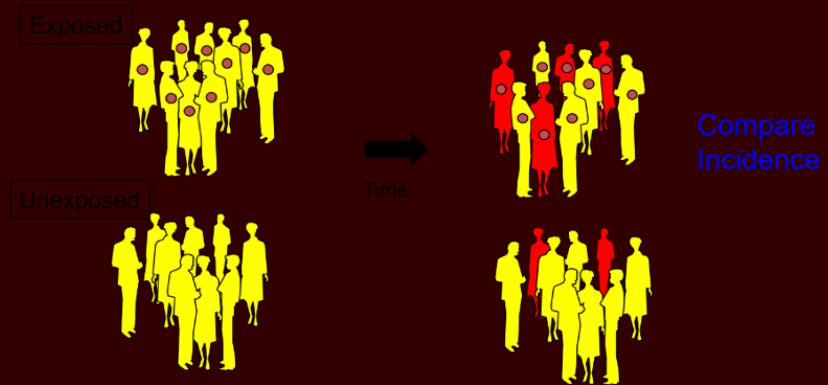


A photograph of a person sitting on a bench in a park at night. The person is wearing a dark jacket and light-colored pants, and is looking down at a smartphone held in their hands. The background is dark with some blurred lights from nearby trees and buildings.

# CASE STUDY: CHF RE-ADMISSION

# Case Study: Selecting the cohort

- Define and select cohort
  - In-patient within health insurance provider's service area
  - Primary diagnosis of CHF in one year
  - Continuous enrollment for at least 6 months prior to primary CHF admission
  - Disqualifying conditions

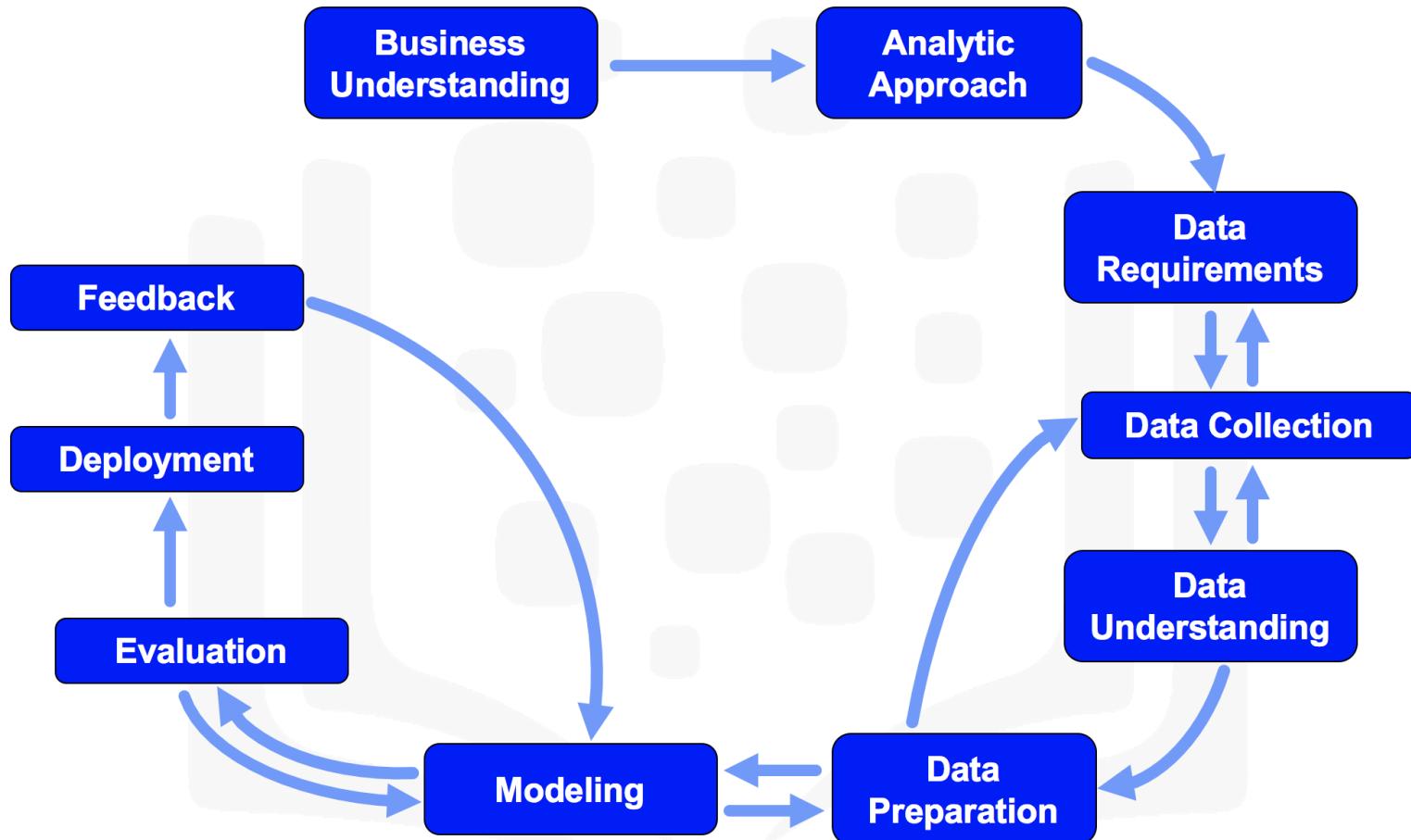


# Case Study: Defining the data

- Contents, formats, representations suitable for decision tree classifier
  - One record per patient with columns representing variables (*dependent variable and predictors*)
  - Content covering all aspects of each patient's clinical history
    - Transaction format
    - Transformations required

age	sex	amn.	diab.	bldprs.	creat.	CHF
...	...	...	...	...	...	...
55	F	1	0	60/144	176	1
78	F	0	0	64/121	106	0
66	M	1	0	79/125	112	1
42	F	1	1	68/133	141	1
60	M	0	1	73/145	133	0
...	...	...	...	...	...	...

# Data Collection



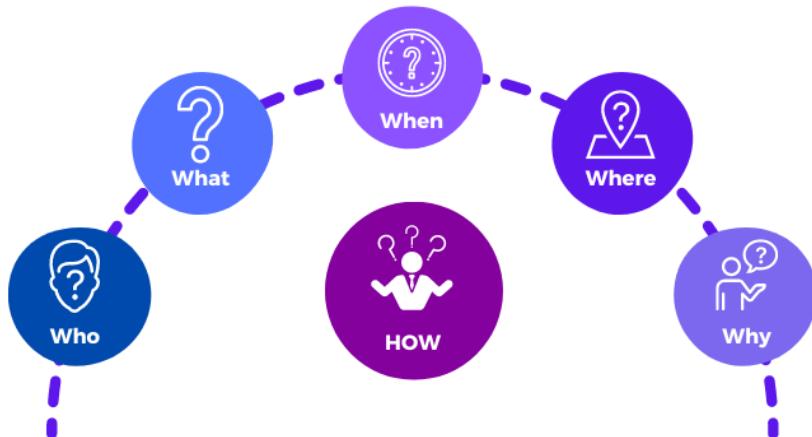
A photograph of a kitchen counter featuring a variety of fresh ingredients. In the foreground, there are several small white bowls containing different spices and seasonings. Behind them, there's a large whole lemon, a red onion, some green onions, and a small bowl of what looks like dried herbs or flowers. In the background, there are more ingredients like avocados and possibly some mushrooms or other vegetables. The lighting is warm and focused on the ingredients.

DATA COLLECTION  
->>> INGREDIENTS? <<<

# From requirements to collection

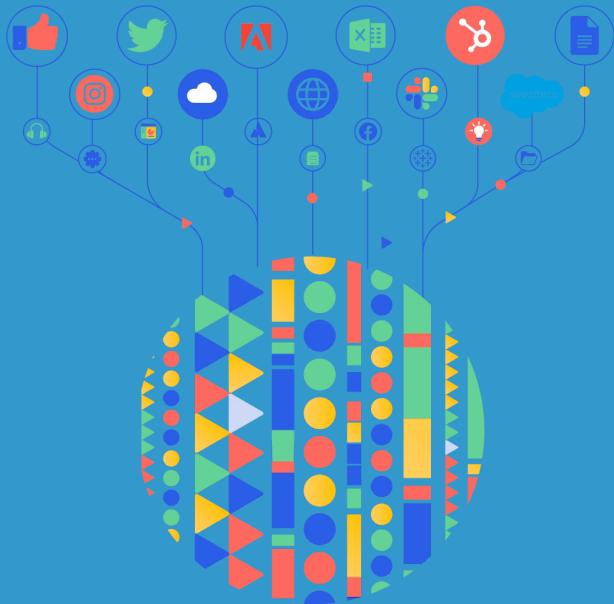
## Data Requirements

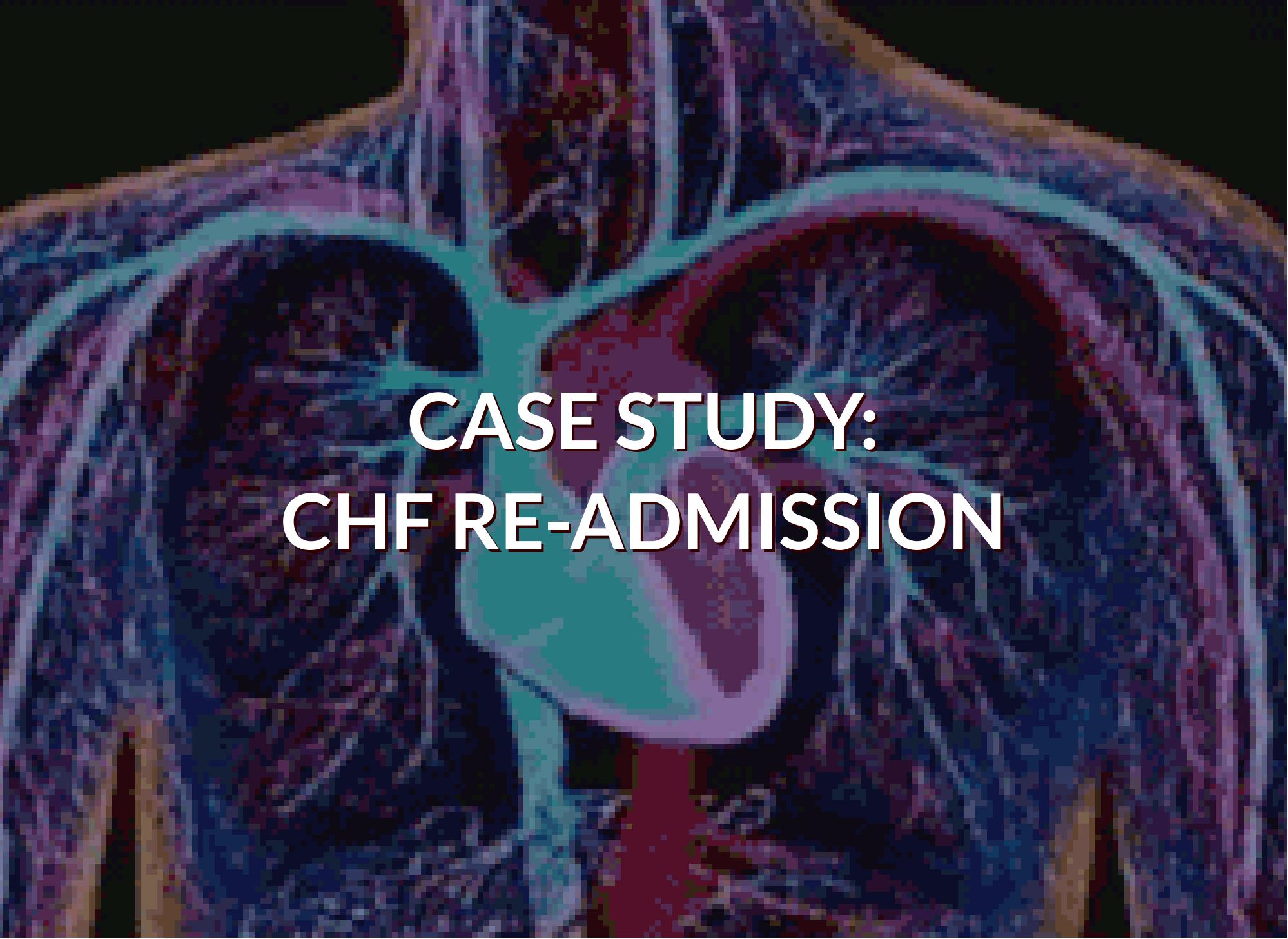
- What are data requirements?



## Data Collection

- What occurs during data collection?

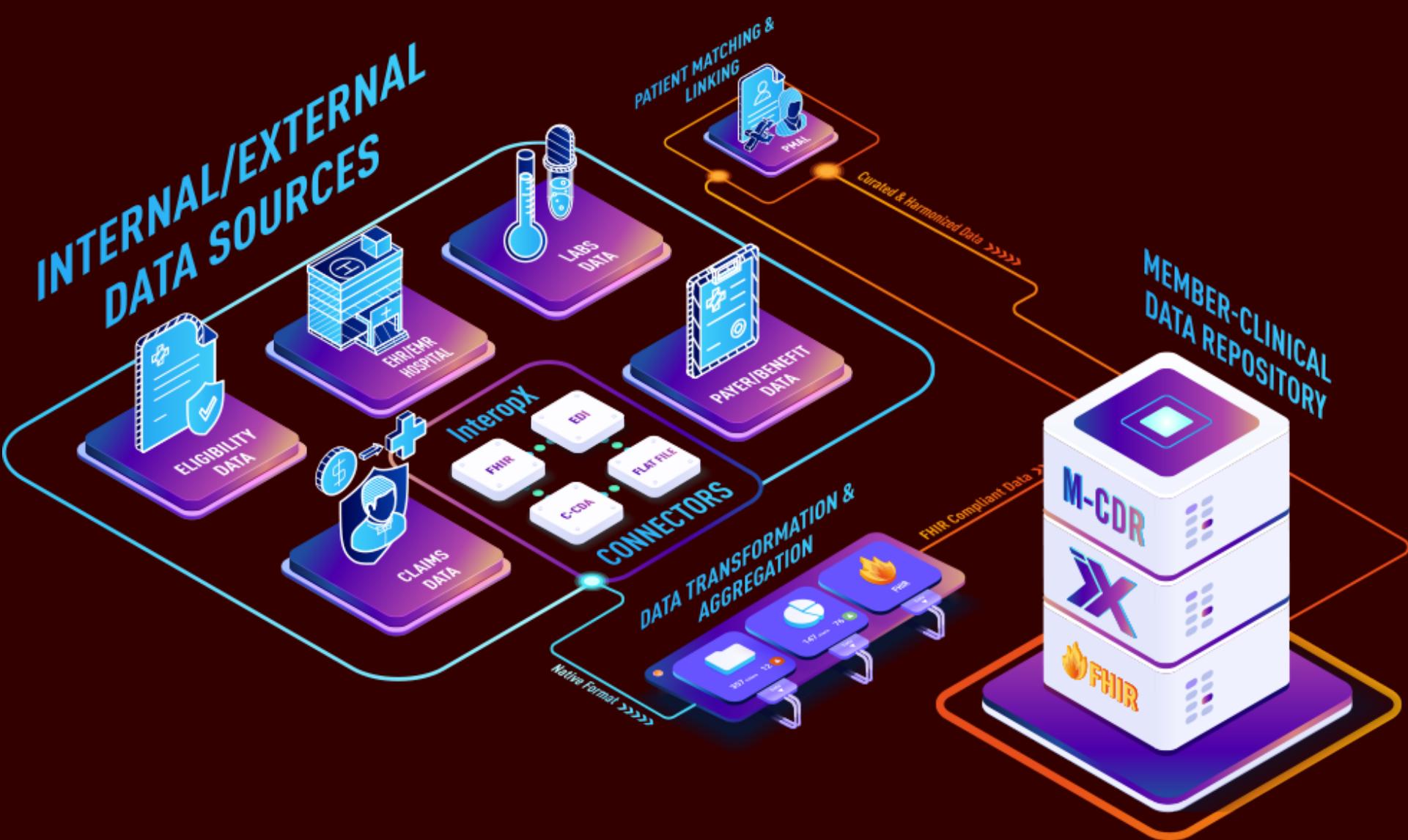




# CASE STUDY: CHF RE-ADMISSION

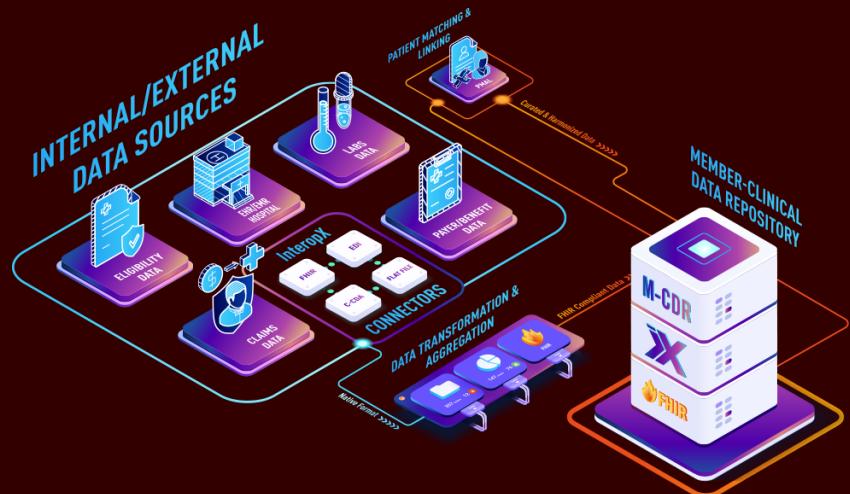
# Case Study: Gathering available data

# INTERNAL/EXTERNAL DATA SOURCES



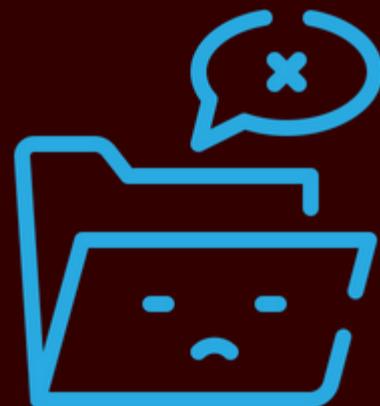
# Case Study: Gathering available data

- Available data sources
  - Cooperate data warehouse (single source of medial & claim, eligibility, provider and member information)
  - In-patient record system
  - Claim payment system
  - Disease management program information



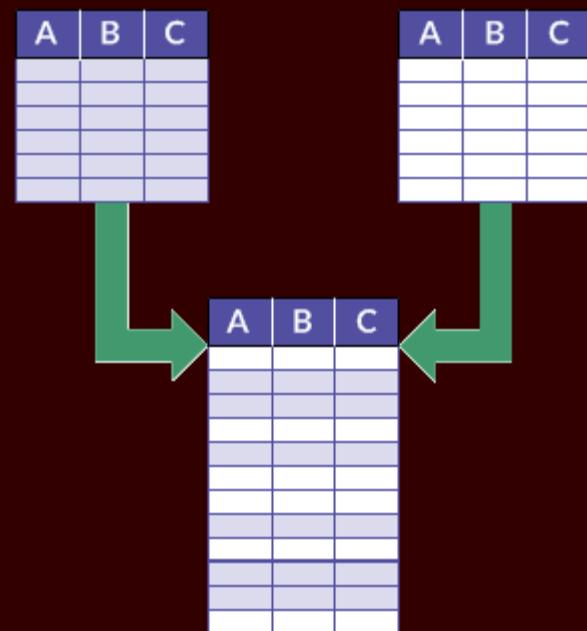
# Case Study: Deferring inaccessible data

- Data wanted but not available
  - Pharmaceutical records
  - Decide to defer



# Case Study: Merging data

Eliminate redundant data



# Module 3

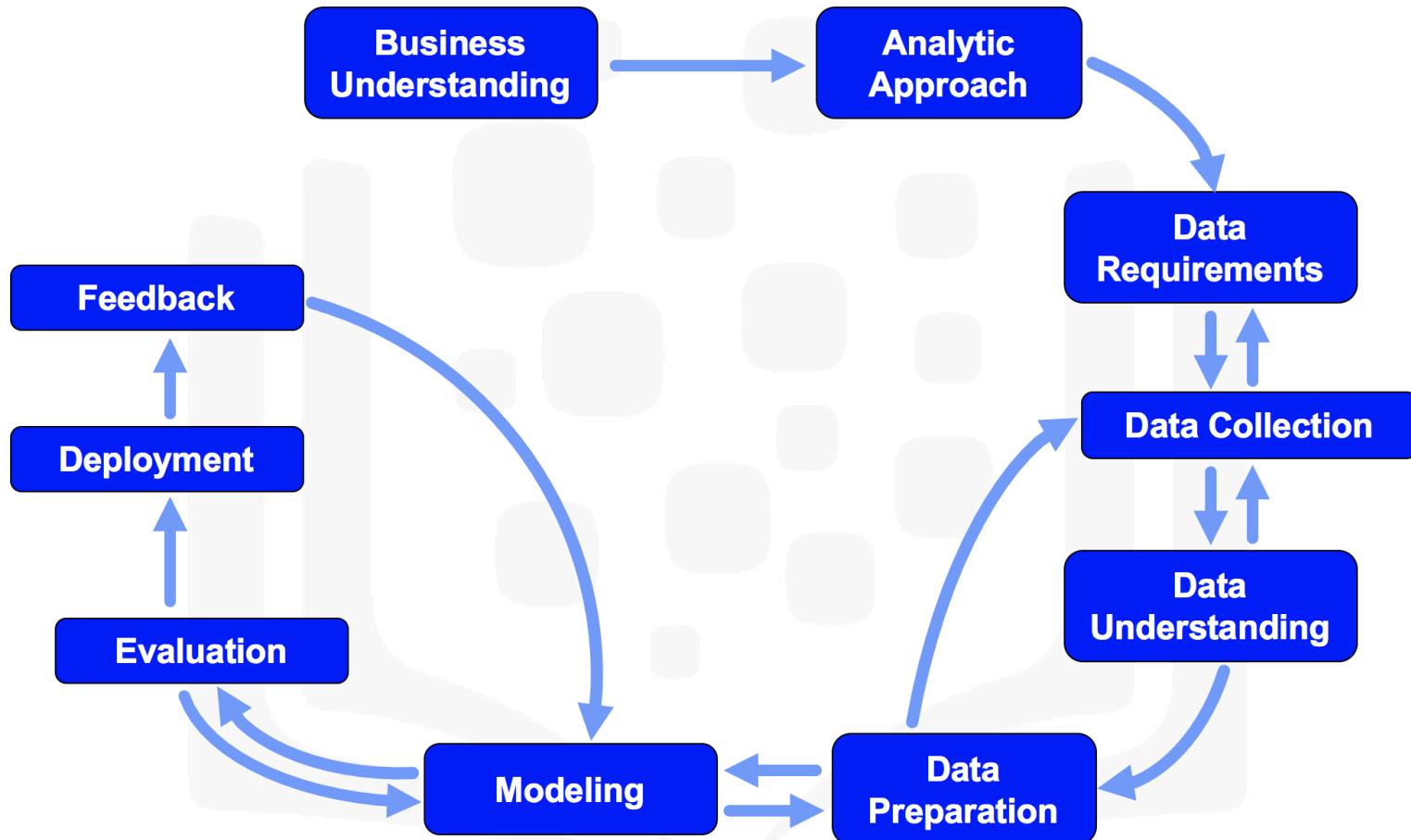
From Understanding to Preparation

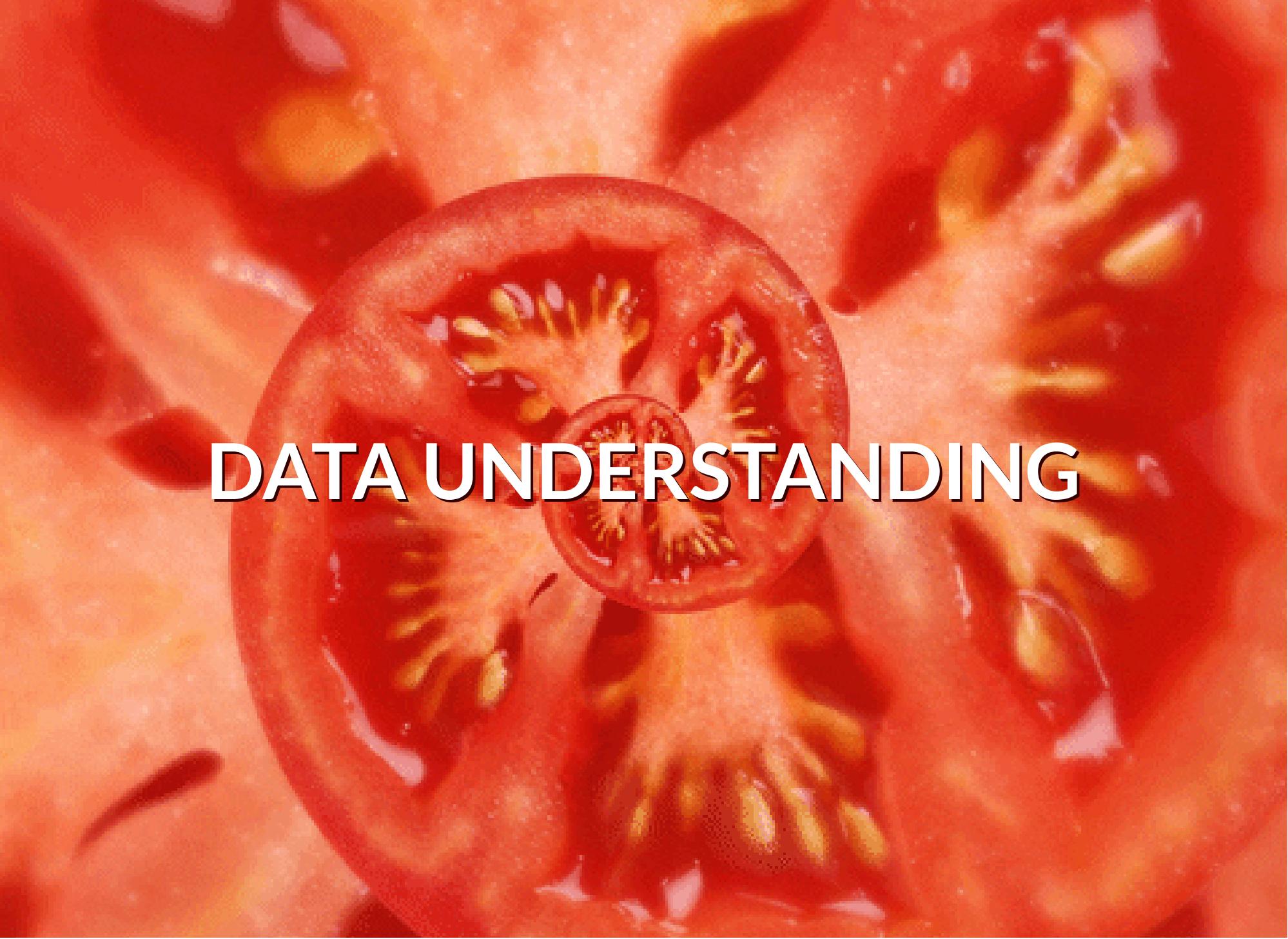
# Learning Objectives

In this lesson you will learn about:

- What it means to understand data.
- What it means to prepare or clean data.
- Ways in which data is prepared.
- How to apply data understanding and data preparation to any data science problem.

# Data Understanding



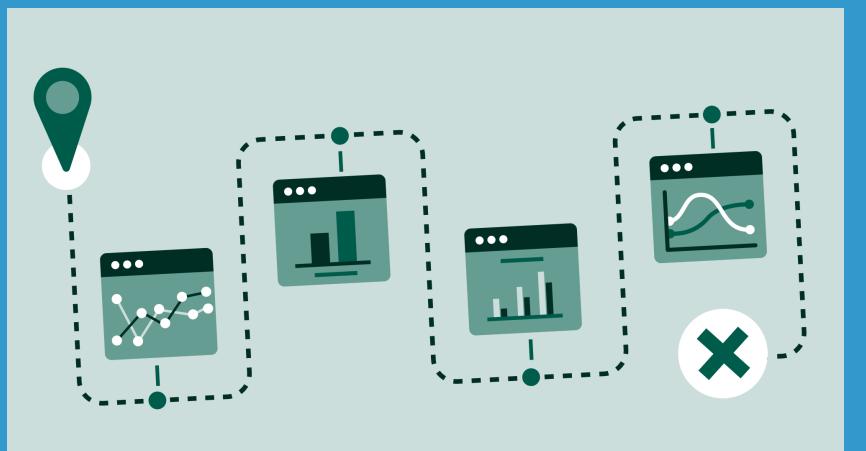


# DATA UNDERSTANDING

# From understanding to preparation

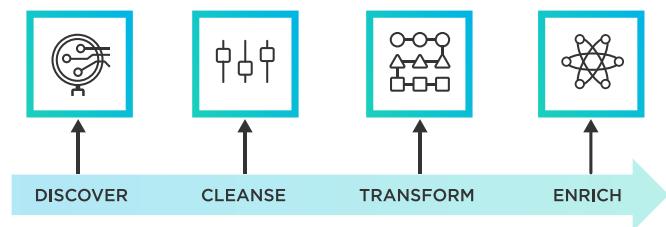
## Data Understanding

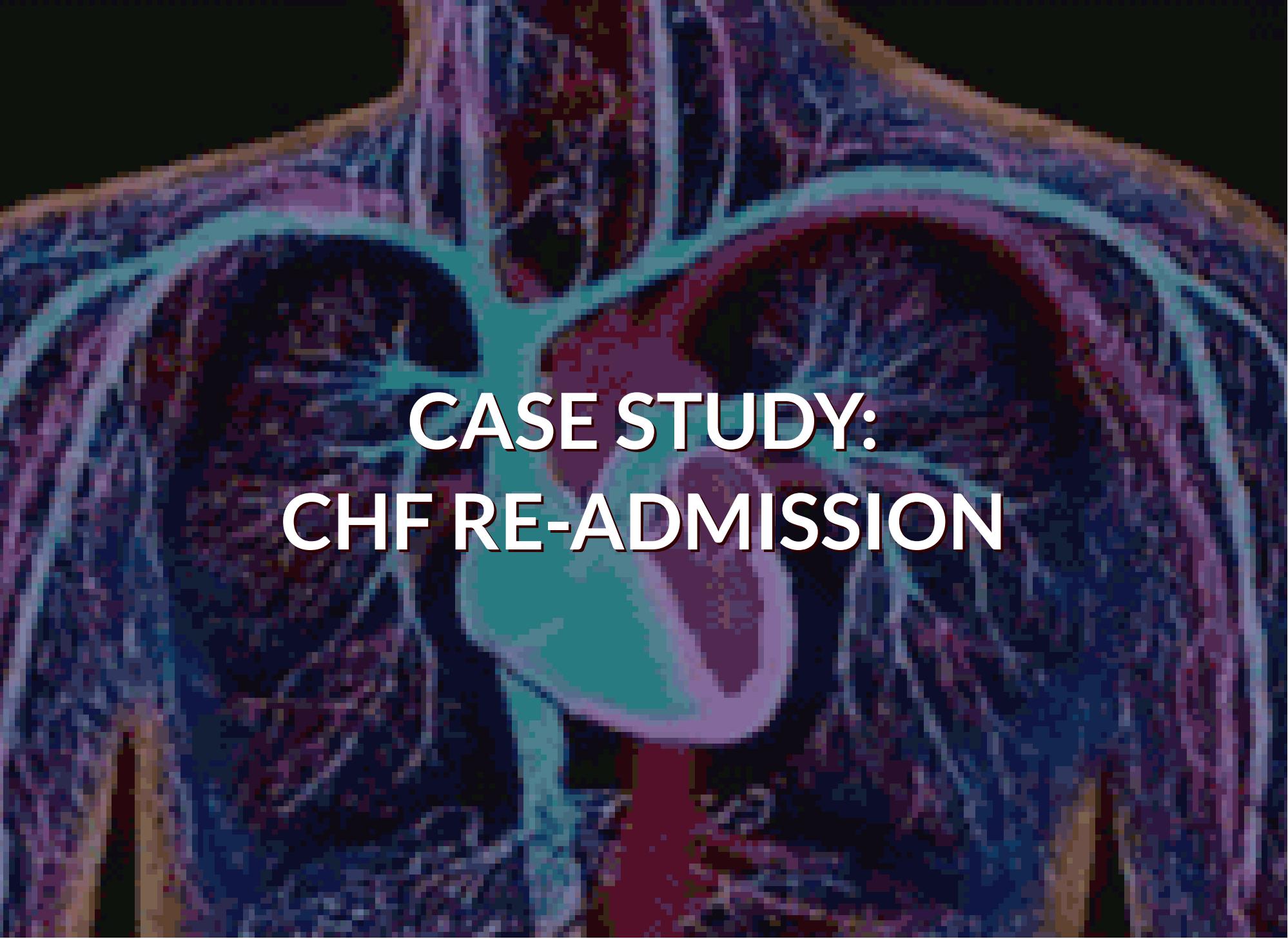
- What does it mean to "prepare" or "clean" data?



## Data Preparation

- What are ways in which data is prepared?



A photograph of a person sitting on a bench in a park at night. The person is wearing a dark jacket and light-colored pants, and is looking down at a smartphone held in their hands. The background is dark with some blurred lights from nearby trees and a building.

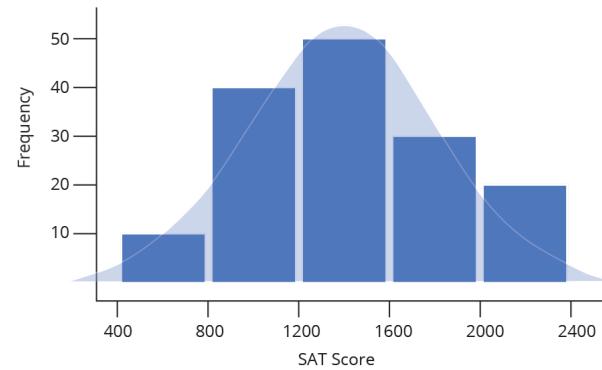
# CASE STUDY: CHF RE-ADMISSION

# Case Study: Understanding the data

- Descriptive statistics
  - **Univariate statistics**
  - **Pairwise correlations**
  - **Histogram**

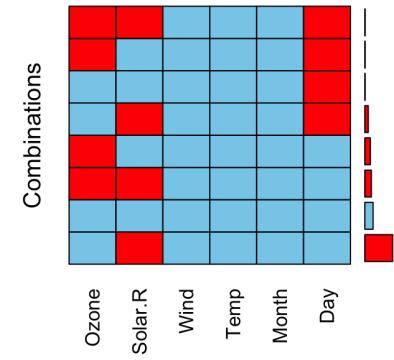
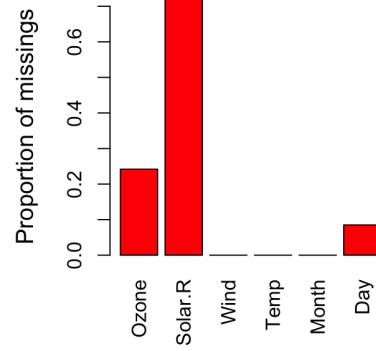
$$f(a) + \sum_{k=1}^n \frac{d^k}{dt^k} f(u(t)) + \int_0^1 \frac{(1-t)^n}{n!} \frac{d^{n+1}}{dt^{n+1}} f(u(t)) dt$$

Histograms are a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.



# Case Study: Looking at data quality

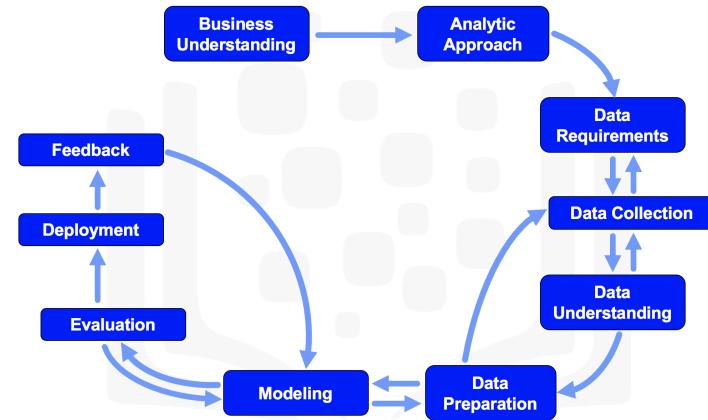
- Data quality
  - Missing values
  - Invalid or misleading values



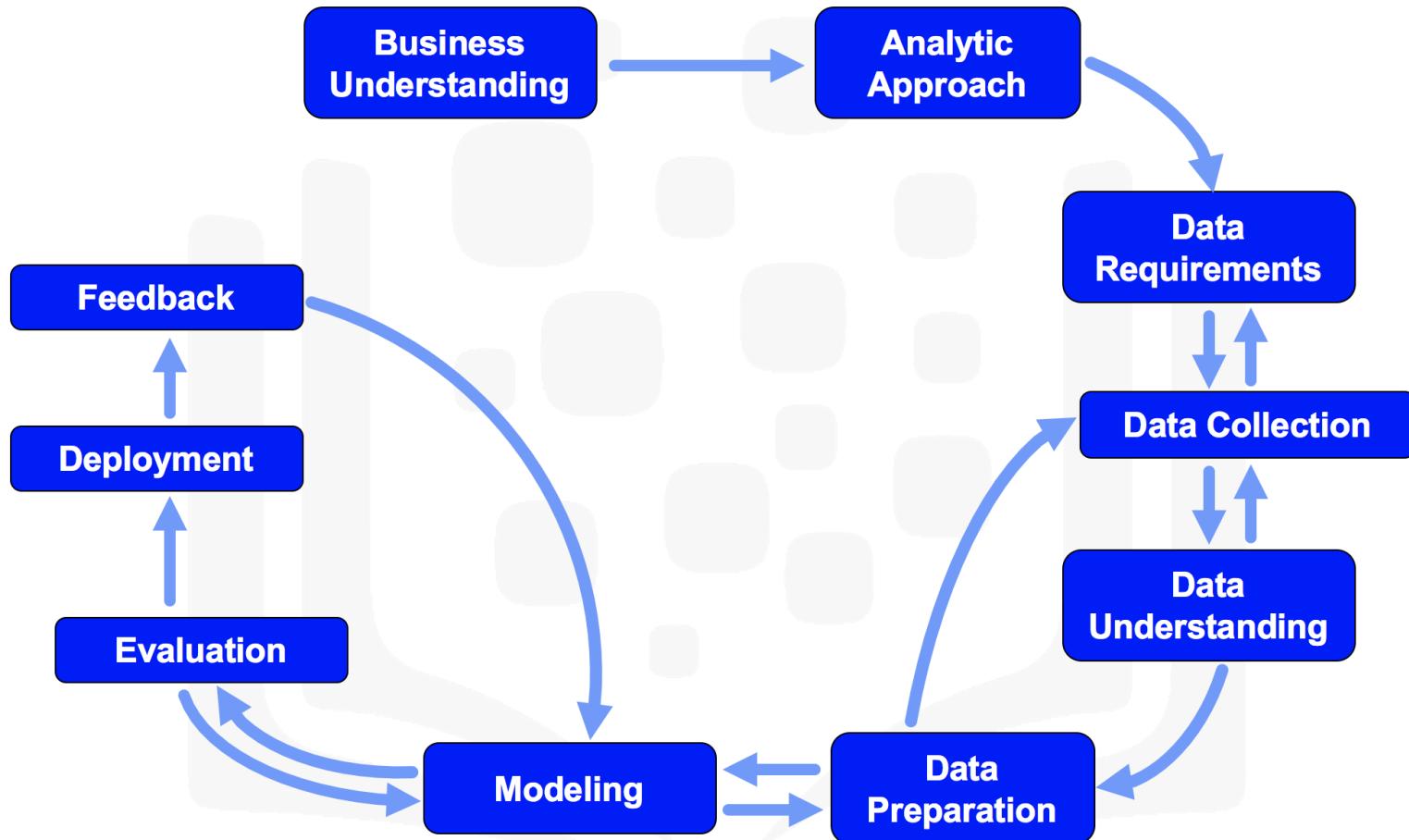
	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

# Case Study: This is an iterative process

- Iterative data collection and understanding
  - Refined definition of "CHE Re-admission"



# Data Preparation



The background of the image is a vibrant orange color. A single, ripe orange is suspended in the center, surrounded by numerous small, translucent bubbles of varying sizes. Some bubbles are clustered around the orange, while others are scattered throughout the frame.

# DATA PREPARATION

>>> CLEANING <<<

A close-up photograph of a person's hand holding a white smartphone. The phone's screen displays a presentation slide with a dark background and large white text. The text reads "DATA PREPARATION" on the first line and ">>> TRANSFORM <<<" on the second line. The person's hand is visible on the right side of the phone, and the background is blurred.

DATA PREPARATION  
>>> TRANSFORM <<<

# From understanding to preparation

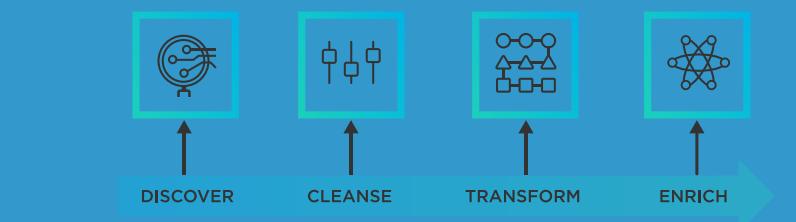
## Data Understanding

- What does it mean to "prepare" or "clean" data?



## Data Preparation

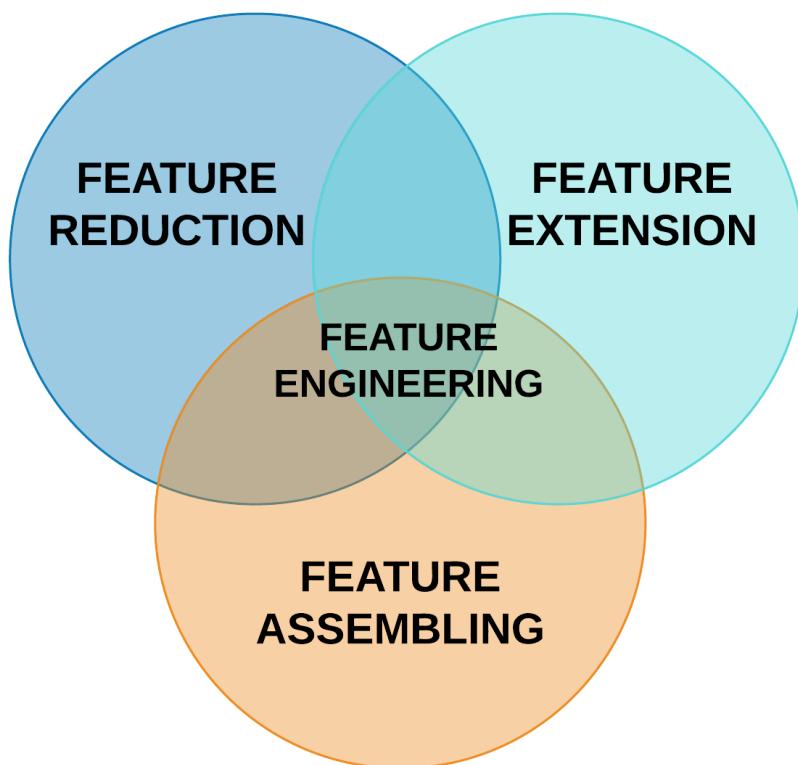
- What are ways in which data is prepared?



# Example of data cleaning

Name	Date	EXP	Location	Country
John Doe	20022012	22	BKK	THA
Marry Jane	2013-02-03	2	BKK	TH
Henny Ozbourne	30-Sep-12	15	Bangkok	Thailand
Kelly, Tom	2015 02 20		B	Thai
Marry Jane	2013-02-03	2	BKK	TH

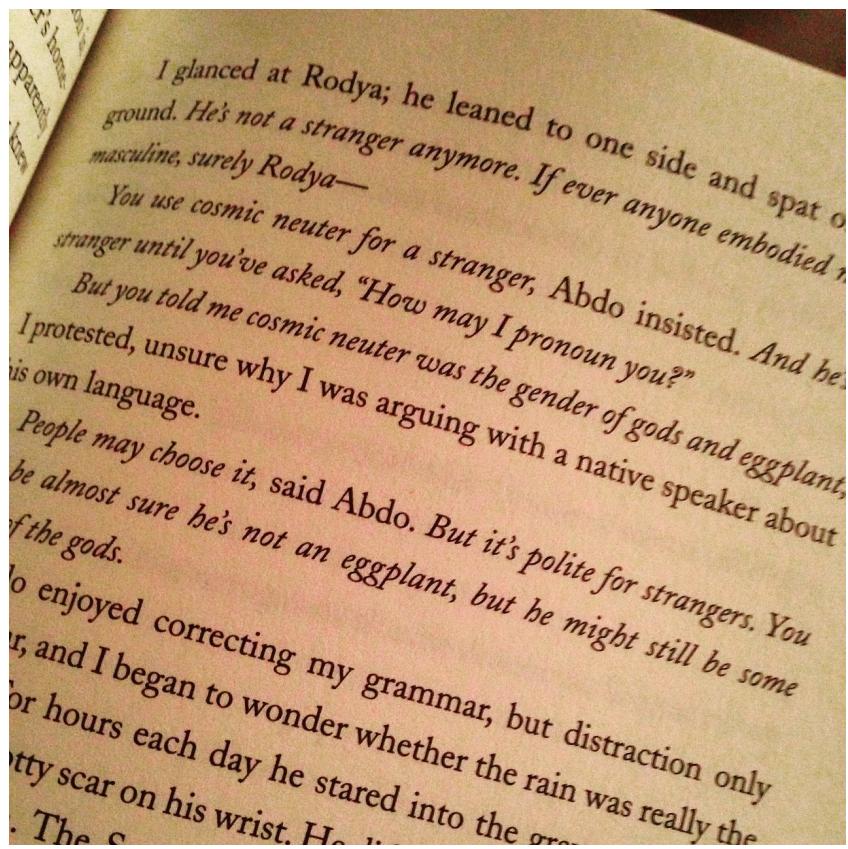
# Using domain knowledge



Feature engineering is the process of using domain knowledge if the data to create features that make the machine learning algorithms work.

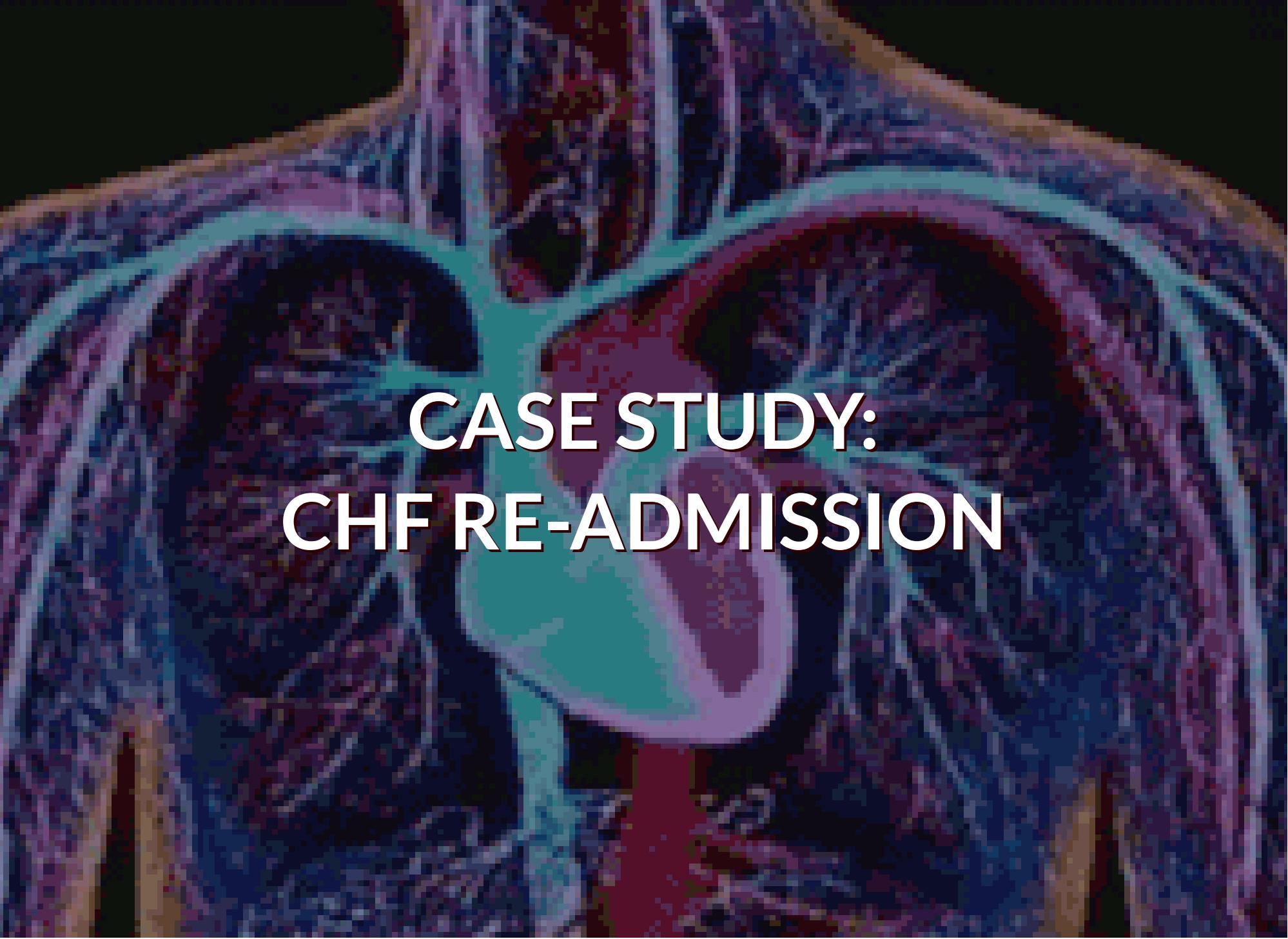
Feature engineering is critical when machine learning tools are being applied to analyze the data.

# Working with text analysis



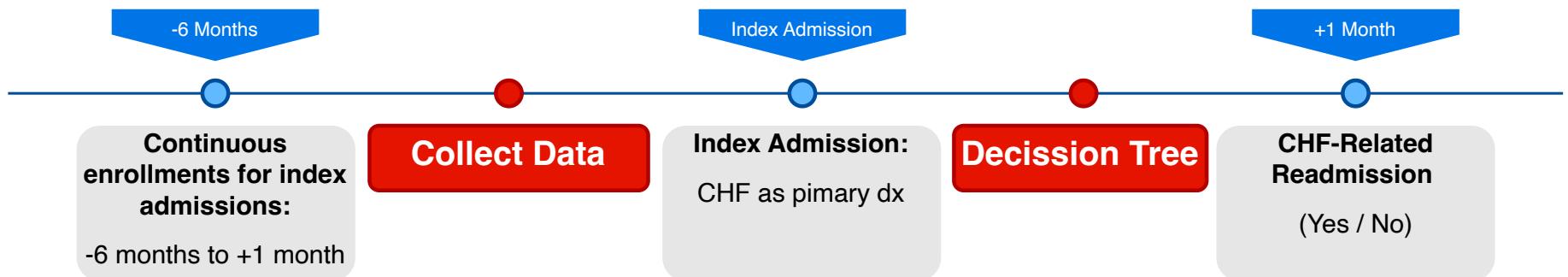
When working with text, **text analysis** steps for coding the data are required to be able to manipulate the data.

The text analysis is critical to ensure that the proper groupings are set, and that the programming is not overlooking what is hidden within.

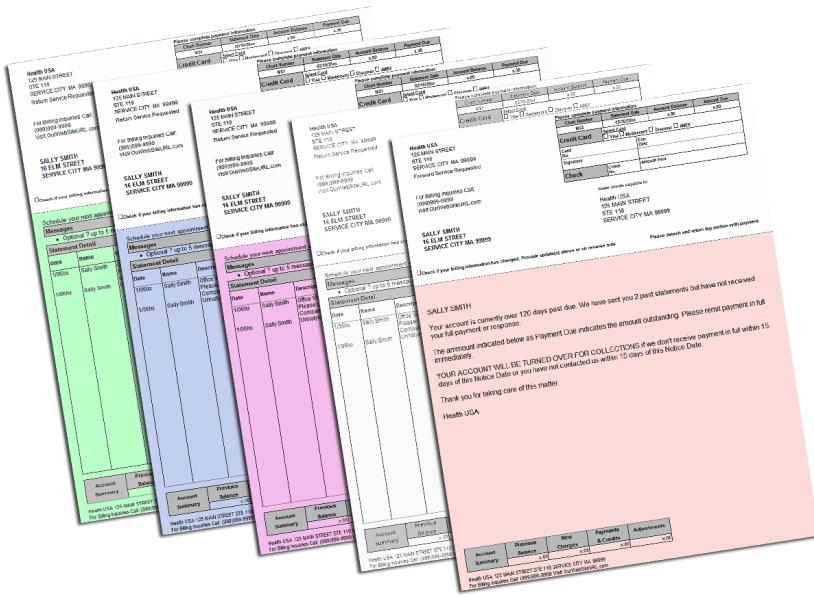


# CASE STUDY: CHF RE-ADMISSION

# Case Study: Defining CHF admission/readmission

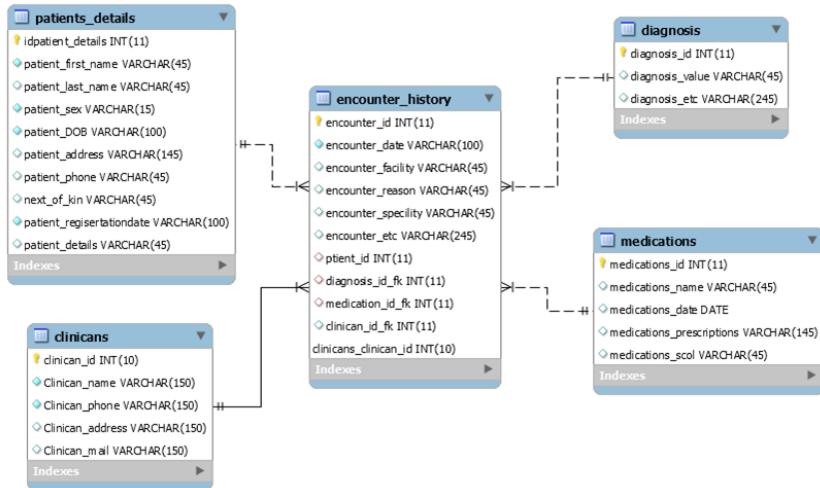


# Case Study: Aggregating records



- Transaction records
  - Claims: professional provider, facility, pharmaceutical
  - Inpatient & outpatient records: diagnoses, procedures, prescriptions, etc.
  - Possibly thousands per patient, depending on clinical history

# Case Study: Aggregating to patient level



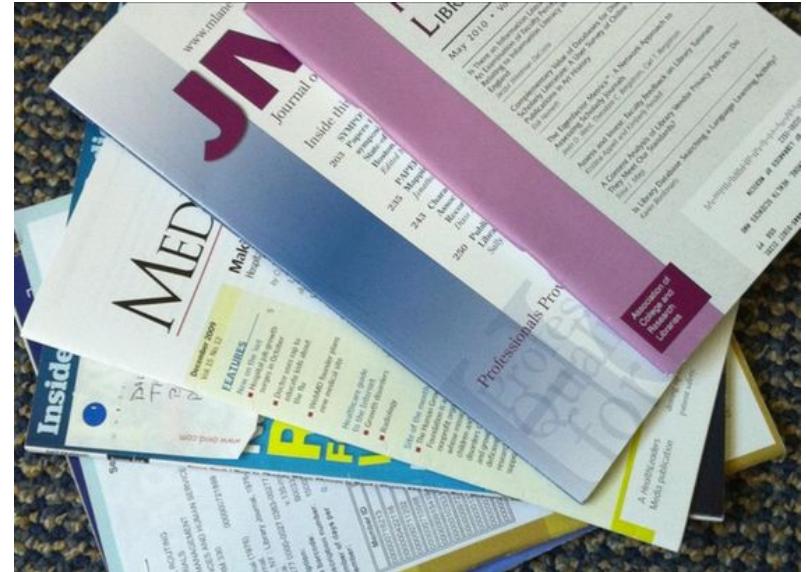
- Roll up to 1 record per patient
- Create new columns representing the transaction
  - Outpatients visits/ Inpatients episodes: frequency, recency, diagnoses, length of stay, procedures, prescriptions
  - Comorbidities with CHF



# Case Study: More or less data needed?



- Literature reviews of important factors for CHF readmission



- Loop back to data collection stage and add additional data, if needed

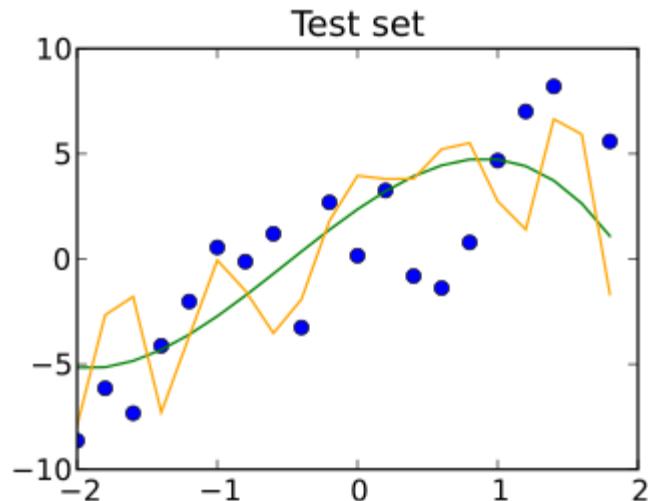
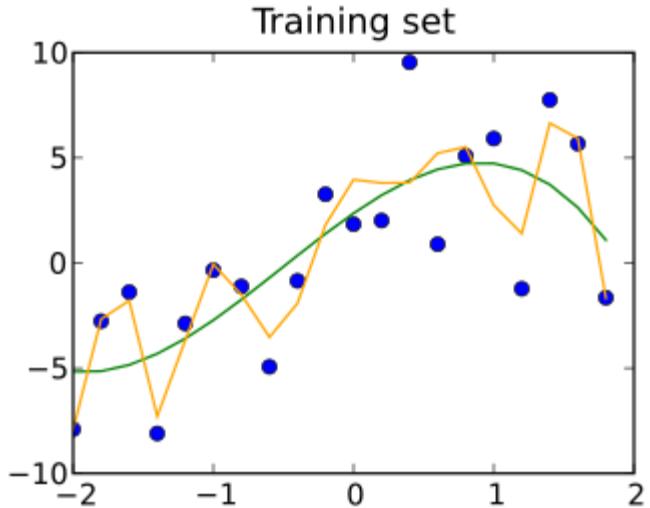
# Case Study: Completing the dataset

age	sex	diab.	bld.	...	CHFR
...	...	...	...	...	...
55	F	0	144	...	N
78	F	0	121	...	N
66	M	0	125	...	Y
42	F	1	133	...	Y
60	M	1	145	...	N
43	F	0	132	...	Y
33	F	0	162	...	N
...	...	...	...	...	...

Merge all data into one table

- One record per patient
- Create new variables
- List of variables used in modeling
  - Target: CHF readmission with 30 days (Yes/No), following discharge from CHF hospitalization
  - Measures: gender, age, primary drug, ...
  - Diagnosis flags (Y/N): CHF, pneumonia, ...

# Case Study: Using training and testing sets



- Cohort: 2,343 patients
  - Randomly divided into training and testing sets: 70% / 30% split
- Training: 1,640 patients
- Testing: 703 patients

# Module 4

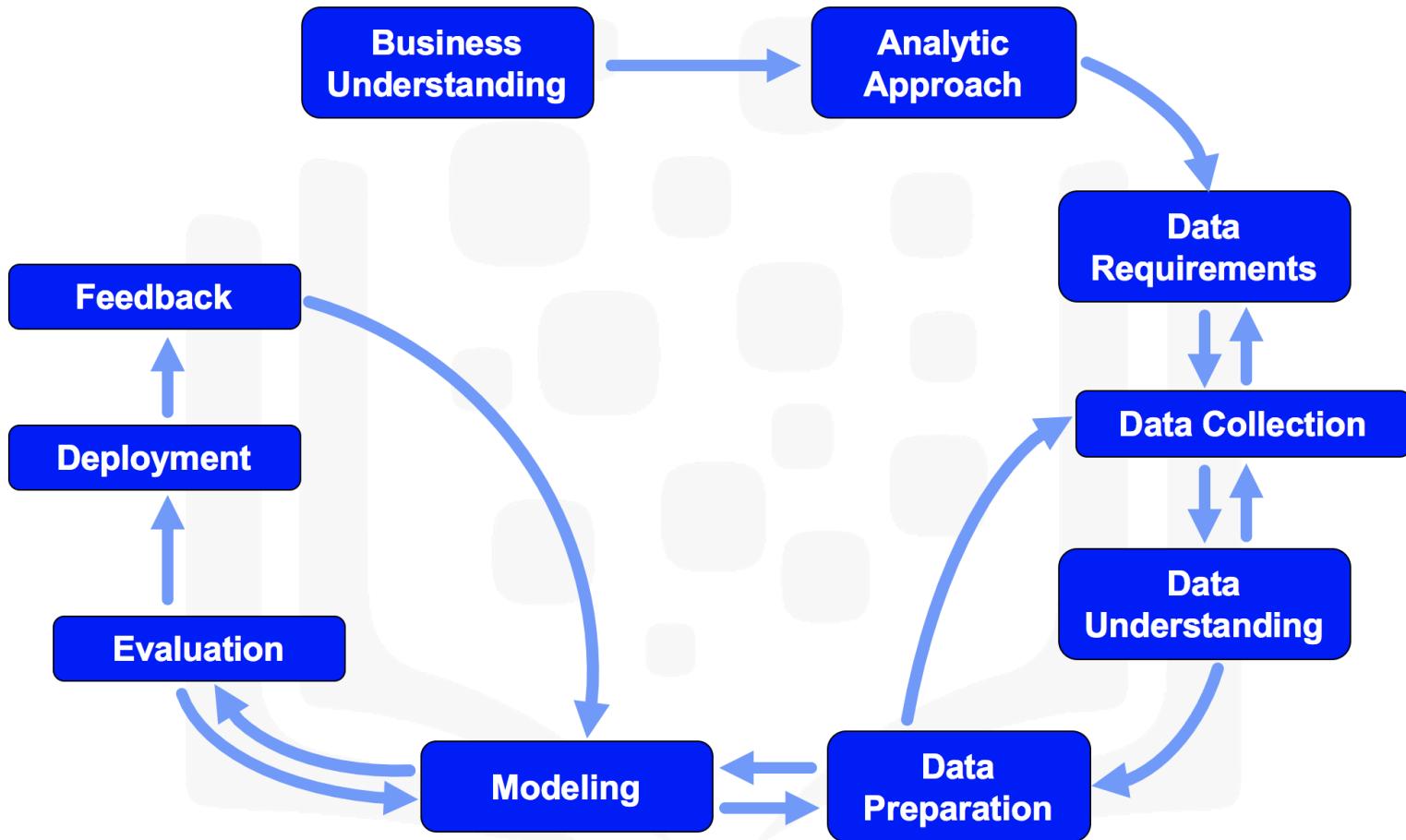
From Modeling to Evaluation

# Learning Objectives

In this lesson you will learn about:

- What the purpose of data modeling is.
- Some characteristics of the modeling process.
- What it means to evaluate a model.
- Ways in which a model is evaluated.
- How to apply modeling and model evaluation to any data science problem.

# Modeling



A professional chef in a white chef's coat is shown from the waist up, leaning over a stainless steel counter in a commercial kitchen. He is holding a large piece of raw meat, likely a chicken or fish fillet, with both hands, examining it closely. On the counter in front of him is a large metal mixing bowl containing some ingredients. To his left is a sink area with a faucet. In the background, there are stainless steel surfaces and equipment, including what looks like a large oven or grill unit.

**MODELING**  
**>>> SAMPLING THE FOOD <<<**

# From modeling to evaluation

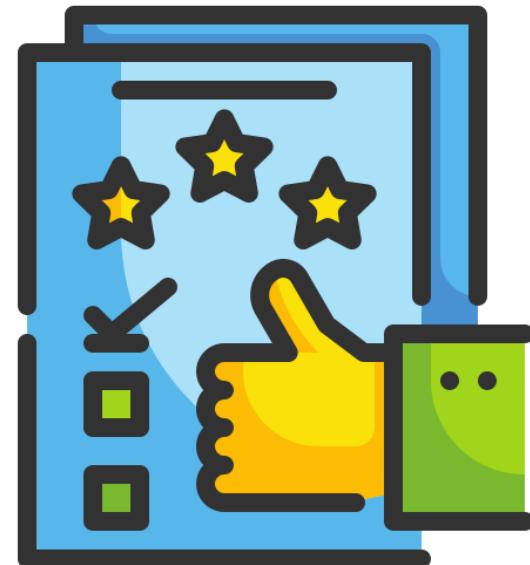
## Modeling

- In what way can the data be visualized to get the answer that is required?

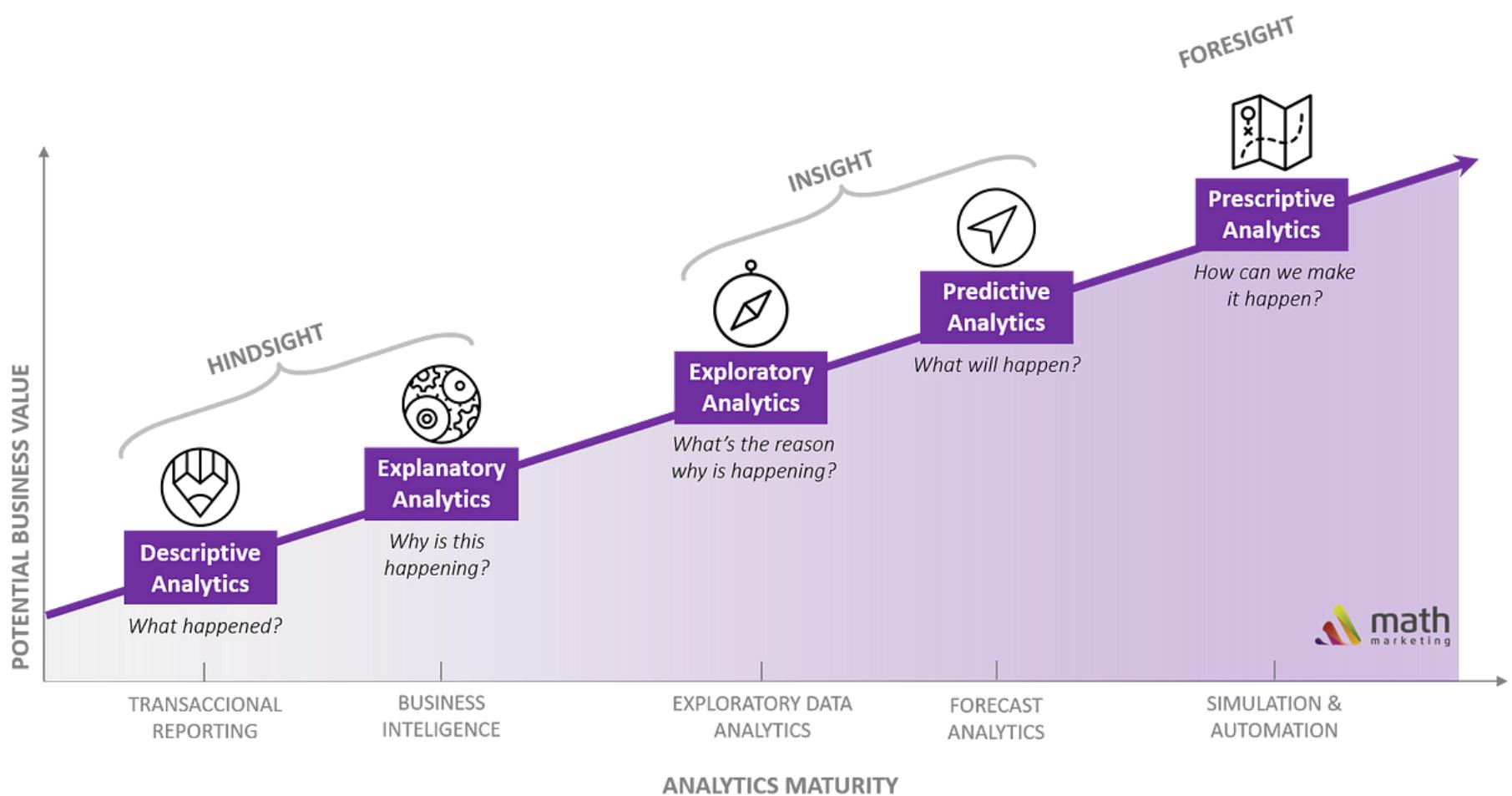


## Evaluation

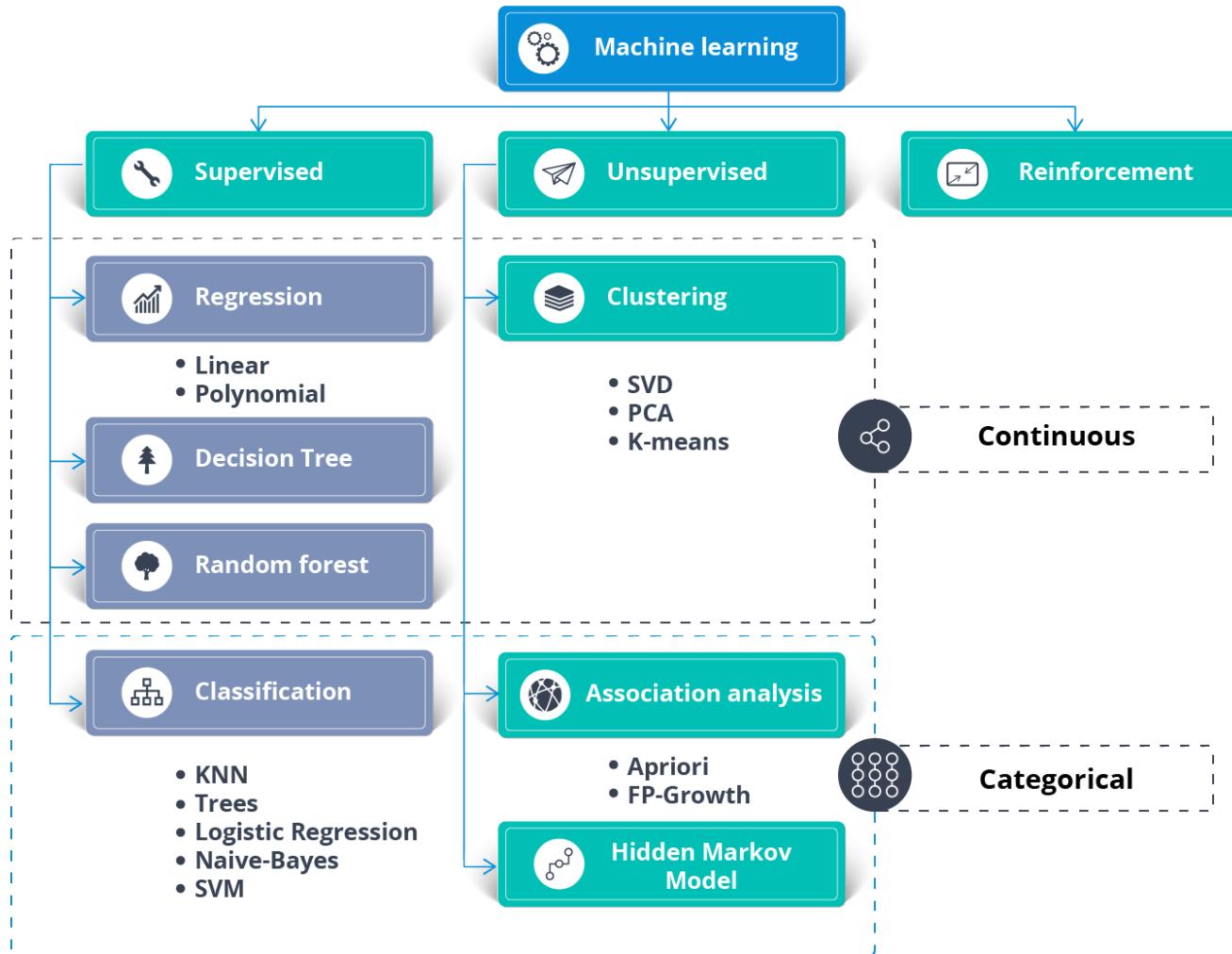
- Does the model used really answer the initial question?
- Does it need to be adjusted?

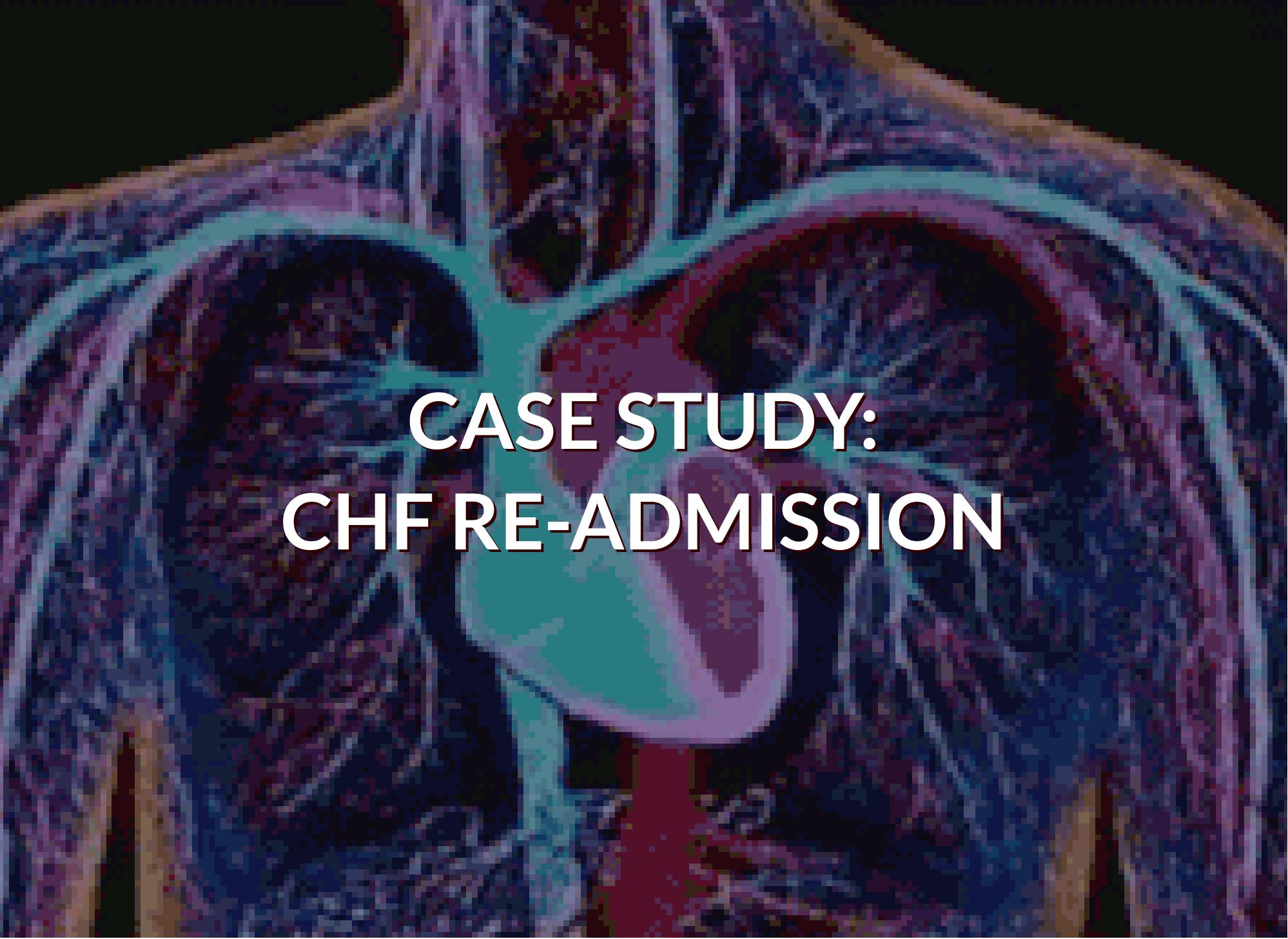


# Analytics Approaches



# Machine Learning Models



A photograph of a person sitting on a bench in a park at night. The person is wearing a dark jacket and light-colored pants, and is looking down at a smartphone held in their hands. The background is dark with some blurred lights from nearby trees and a building.

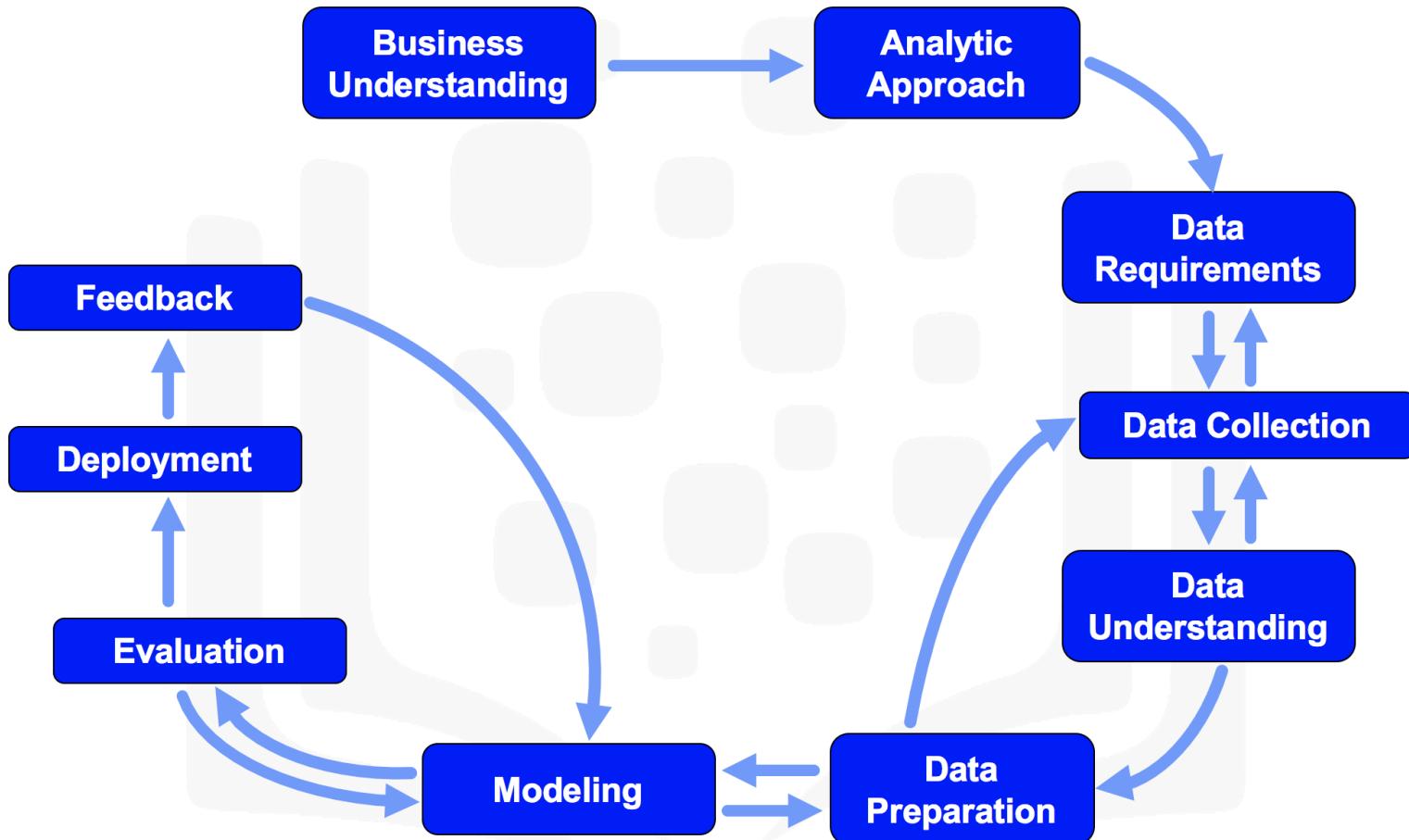
# CASE STUDY: CHF RE-ADMISSION

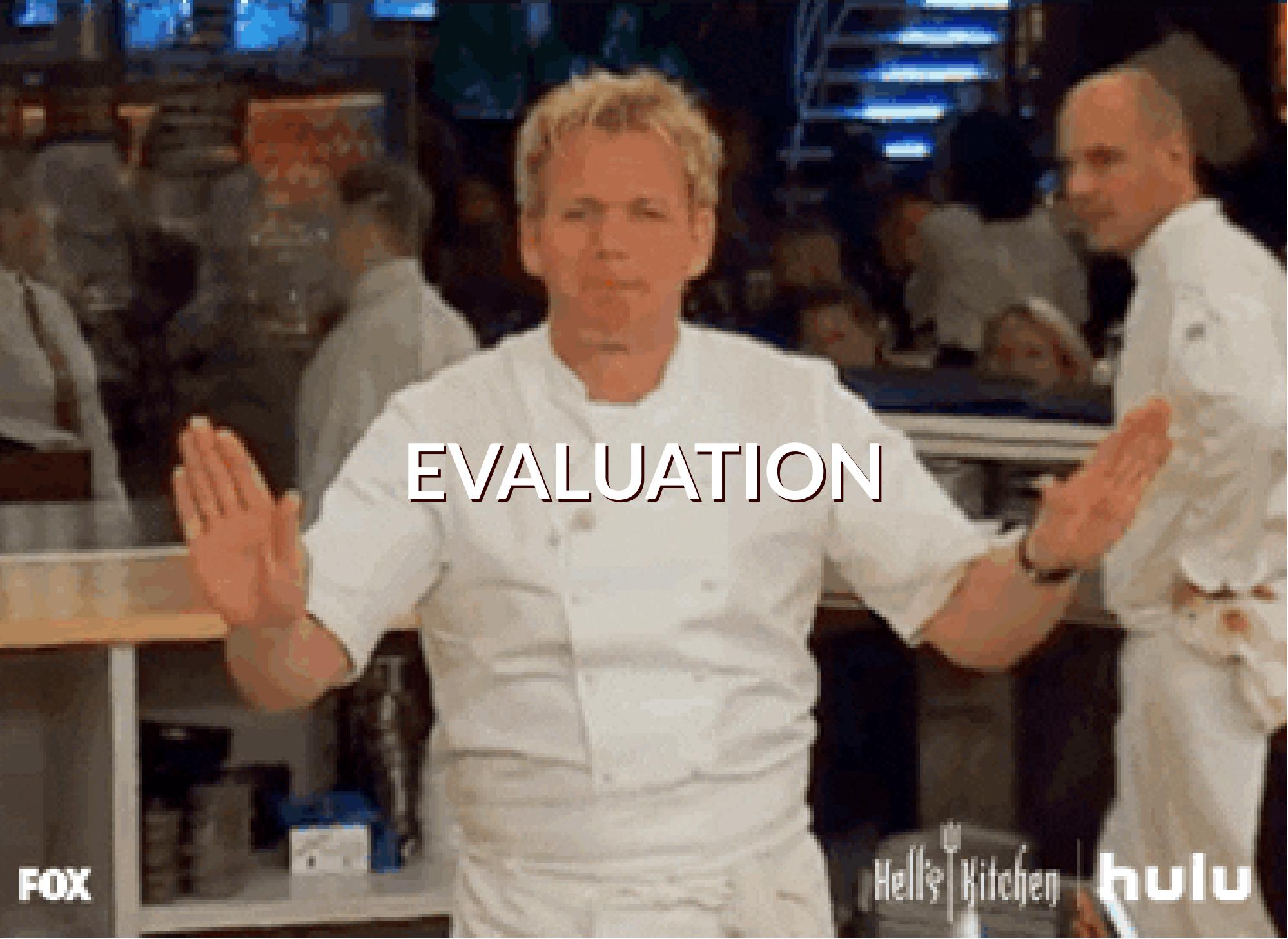
# Case Study: Analyzing the models

## Decission tree classification models

Model	Relative Cost (Y:N)	Overall Accuracy (% correct of Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

# Evaluation



A dramatic scene from the TV show Hell's Kitchen. Chef Gordon Ramsay, in the center, is shown from the waist up, wearing a white chef's coat. He has a serious, intense expression and is gesturing with his hands raised near his shoulders. In the background, another chef in a white coat is visible, looking towards the camera. The setting appears to be a professional kitchen with stainless steel equipment.

# EVALUATION

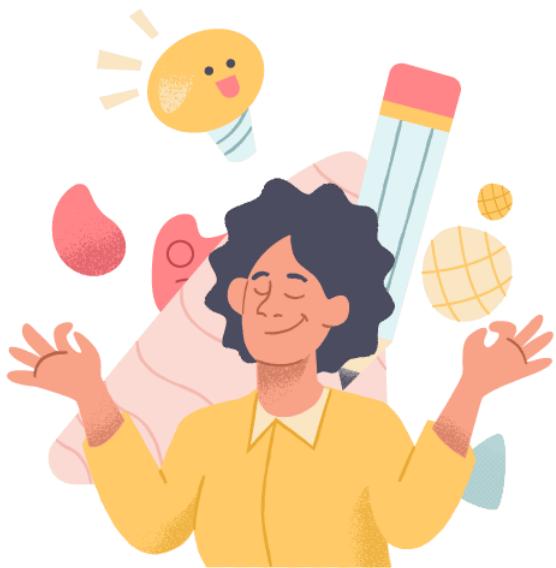
FOX

Hell's Kitchen hulu

# From modeling to evaluation

## Modeling

- In what way can the data be visualized to get the answer that is required?



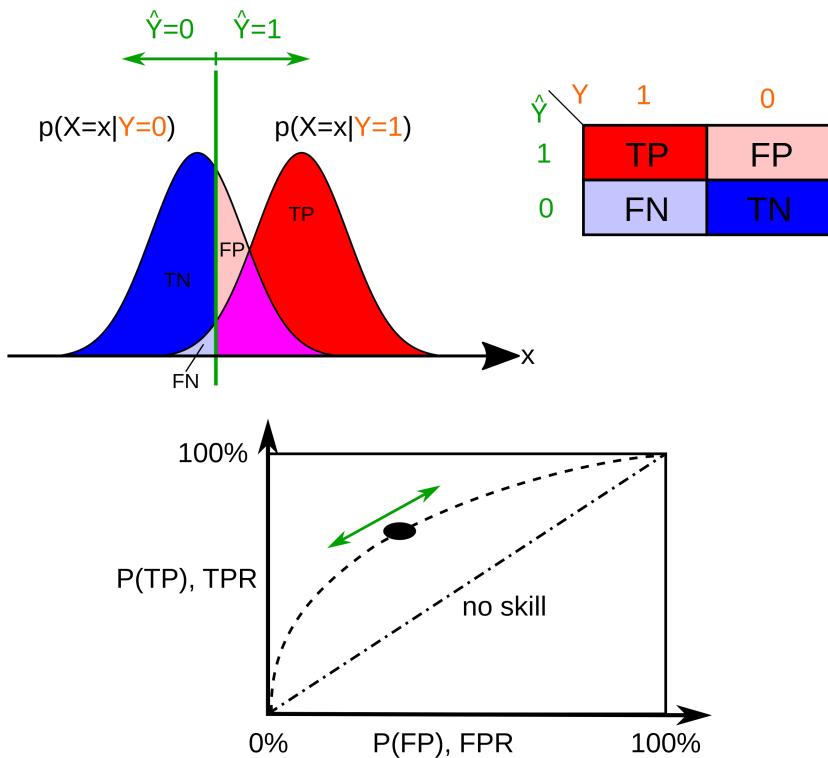
## Evaluation

- Does the model used really answer the initial question?
- Does it need to be adjusted?

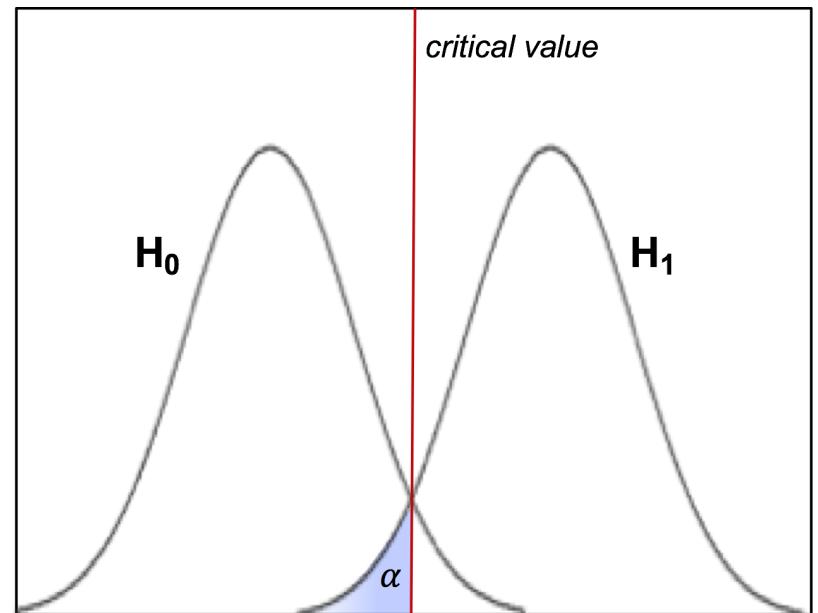


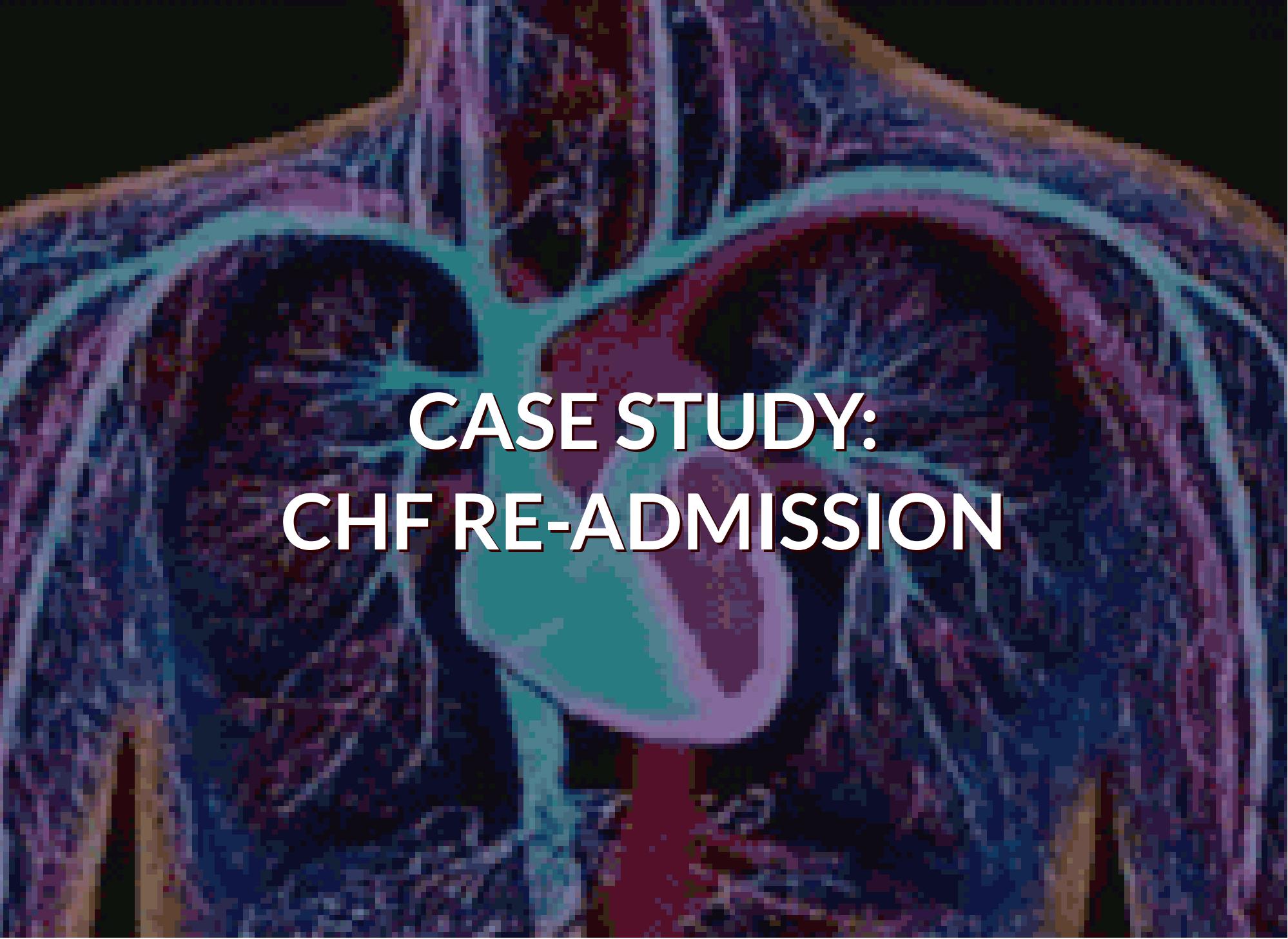
# Model evaluation

## Phase 1: Diagnostic measures



## Phase 2: Statistical significance





# CASE STUDY: CHF RE-ADMISSION

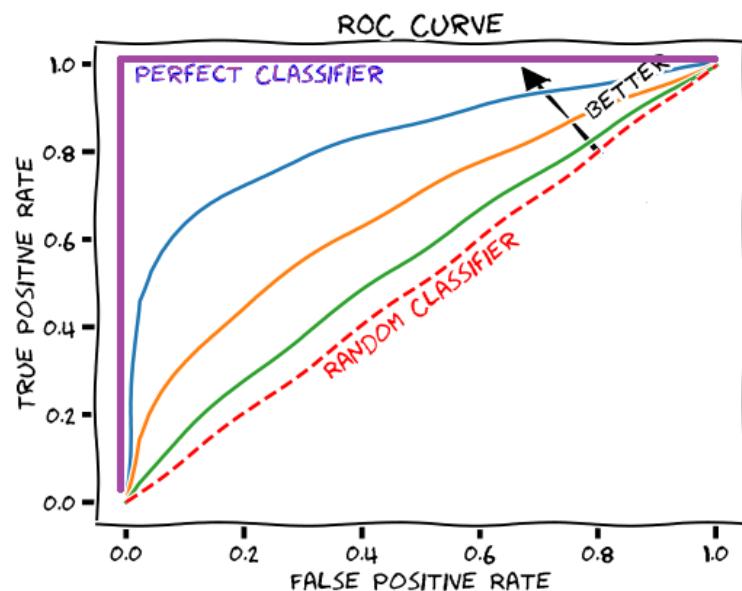
# Case Study: Misclassification costs

Model	Relative Cost (Y:N)	TPR (Sensitivity)	Specificity (N accuracy)	FPR (1-Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

# Case Study: Using the ROC curve

Receiver Operating Characteristic  
curve

- Diagnostic tool for classification model evaluation
- Classification model performance
- True-Positive Rate (TPR) vs False-Positive Rate (FPR)
- Optimal model at maximum separation



# Module 5

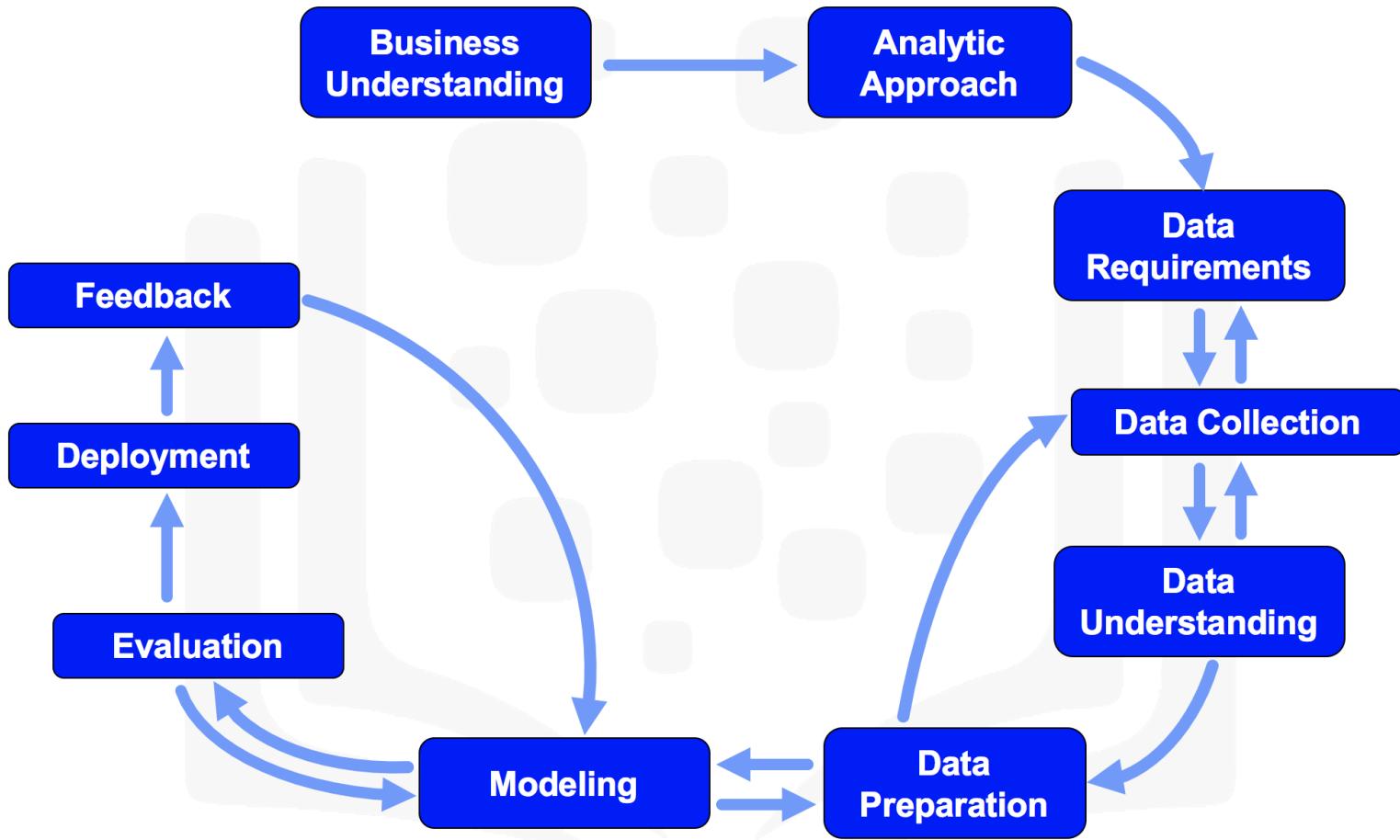
From Deployment to Feedback

# Learning objectives

In this lesson you will learn about:

- What happens when a model is deployed.
- Why model feedback is important.

# Deployment





**DEPLOYMENT**

# From deployment to feedback

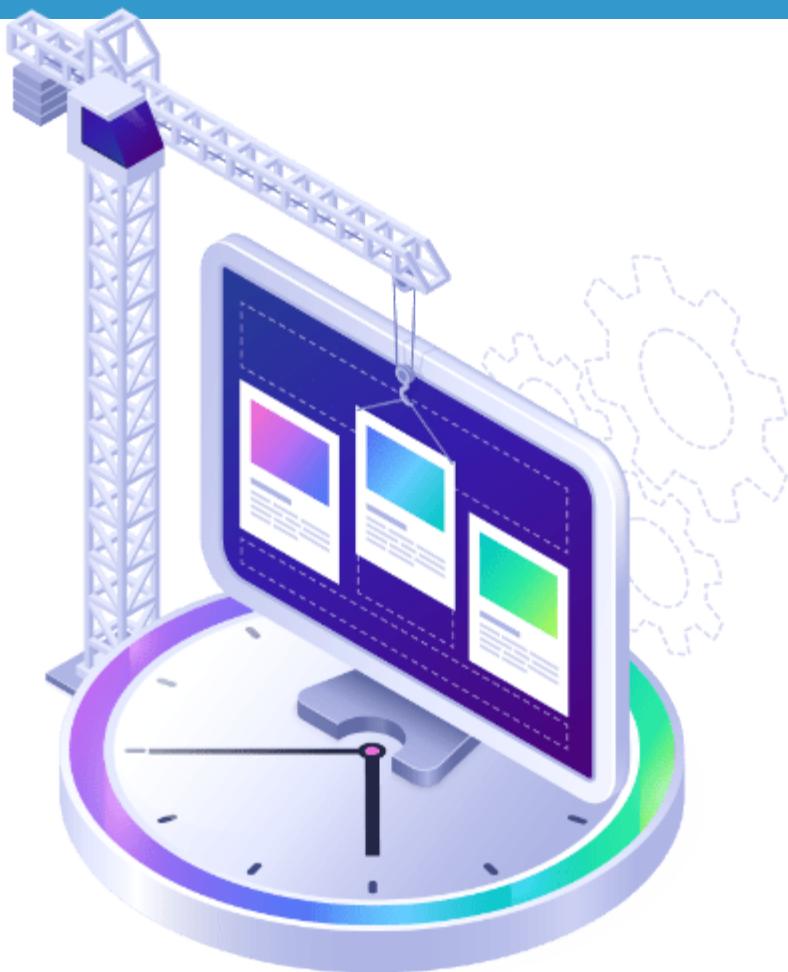
Deployment

Feedback

- Can you get constructive feedback into answering the question?



- Can you put the model into practice?



# Case Study: Understand the result

Assimilate knowledge for business

- Practical understanding of the meaning of model results
- Implications of model results for designing intervention actions



# Case Study: Assembling requirements

- Application requirements
  - Automated, near-real-time risk assessments of CHF inpatients
  - Easy to use
  - Automated data preparation and scoring
  - Up-to-date risk assessment to help clinicians target high-risk patients
- Additional requirements
  - Training for clinical staffs
  - Tracking/ monitoring processes



# Example: Random number simulation of the Dunning and Kruger experiments

# Random number simulation of the Dunning and Kruger experiments

source code e10v blog post

The Dunning–Kruger effect suggests that people with low competence in a domain often overestimate their abilities. But do the foundational experiments truly confirm this effect? This app offers a random number simulation challenging those original findings. For a deep dive into the topic, check out my blog post [Debunking the Dunning–Kruger effect with random number simulation](#).

## Parameters

Number of participants

100

50 150

Correlation

0.50

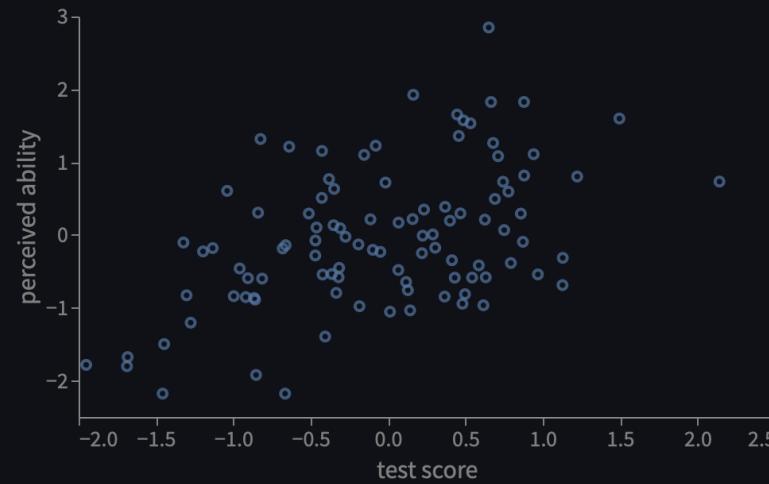
0.00 1.00

Random seed

42

- +

## Test score vs. perceived ability



<https://dunning-kruger.streamlit.app>

# Example: State Movement

# State Movement

Choose a state

OVERALL US

Choose a direction

Incoming

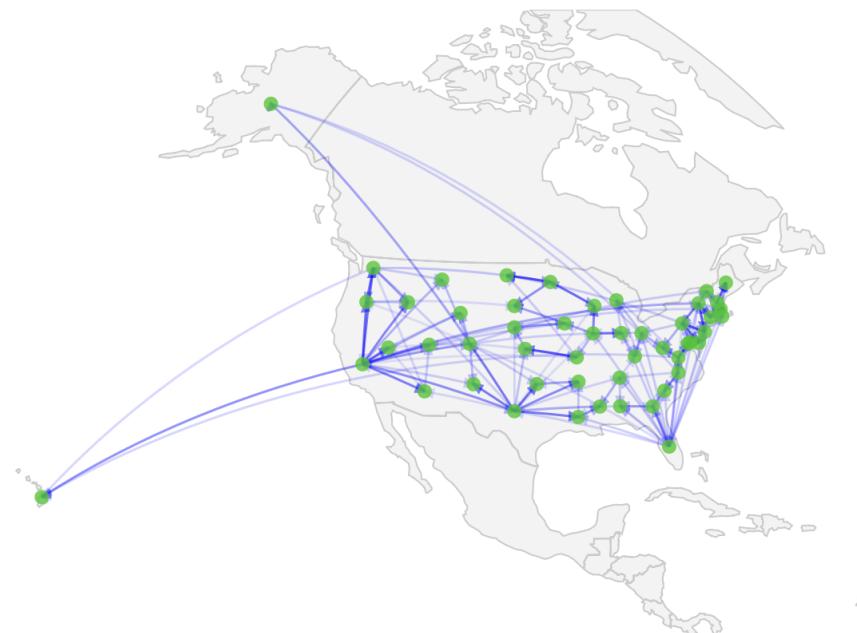
Set top N Migration per state

3 - +

Results Limited to top 5 per State in overall US

**Build Migration Map**

What is this?



Hope you like the map!

## Migration Table

Source State	Destination State	Total People	% of Incoming Migration from Source
NEW YORK	NEW JERSEY	58,959	40.61%

# Example: Euro 2024 Simulation

# Euro 2024 Torba Simülasyonu

Galler

Türkiye

0	-	+	0	-	+
---	---	---	---	---	---

Hırvatistan

Ermenistan

0	-	+	0	-	+
---	---	---	---	---	---

Romanya

İsviçre

0	-	+	0	-	+
---	---	---	---	---	---

Cebelitarık

Hollanda

0	-	+	0	-	+
---	---	---	---	---	---

Grp	Team	P	Av.	Siralama
J	Portekiz	24	28	1
B	Fransa	21	26	1
A	İspanya	21	20	1
F	Belçika	20	18	1
C	İngiltere	20	18	1
G	Macaristan	18	9	1
D	Türkiye	17	7	1
H	Danimarka	16	4	1
E	Arnavutluk	15	8	1
I	Romanya	14	4	1
F	Avusturya	19	10	2
A	İskoçya	17	9	2
J	Slovenya	16	5	2
B	Slovakya	16	5	2
D	Hollanda	16	4	2
G	Çekya	15	6	2
E	Hırvatistan	14	8	2

**Pot 1**  
Takımları

Almanya

Portekiz

Fransa

İspanya

Macaristan

Danimarka

İngiltere

Türkiye

Belçika

Romanya

Avusturya

**Pot 2**  
Takımları

Macaristan

Fransa

Slovenya

İspanya

Danimarka

Arnavutluk

Belçika

Romanya

Çekya

Avusturya

**Pot 3**  
Takımları

İskoçya

Slovenya

Slovakya

Çekya

Hollanda

Avusturya

Hırvatistan

**Pot 4**  
Takımları

İtalya

Sırbistan

İsviçre

/ /

/ /

/ /

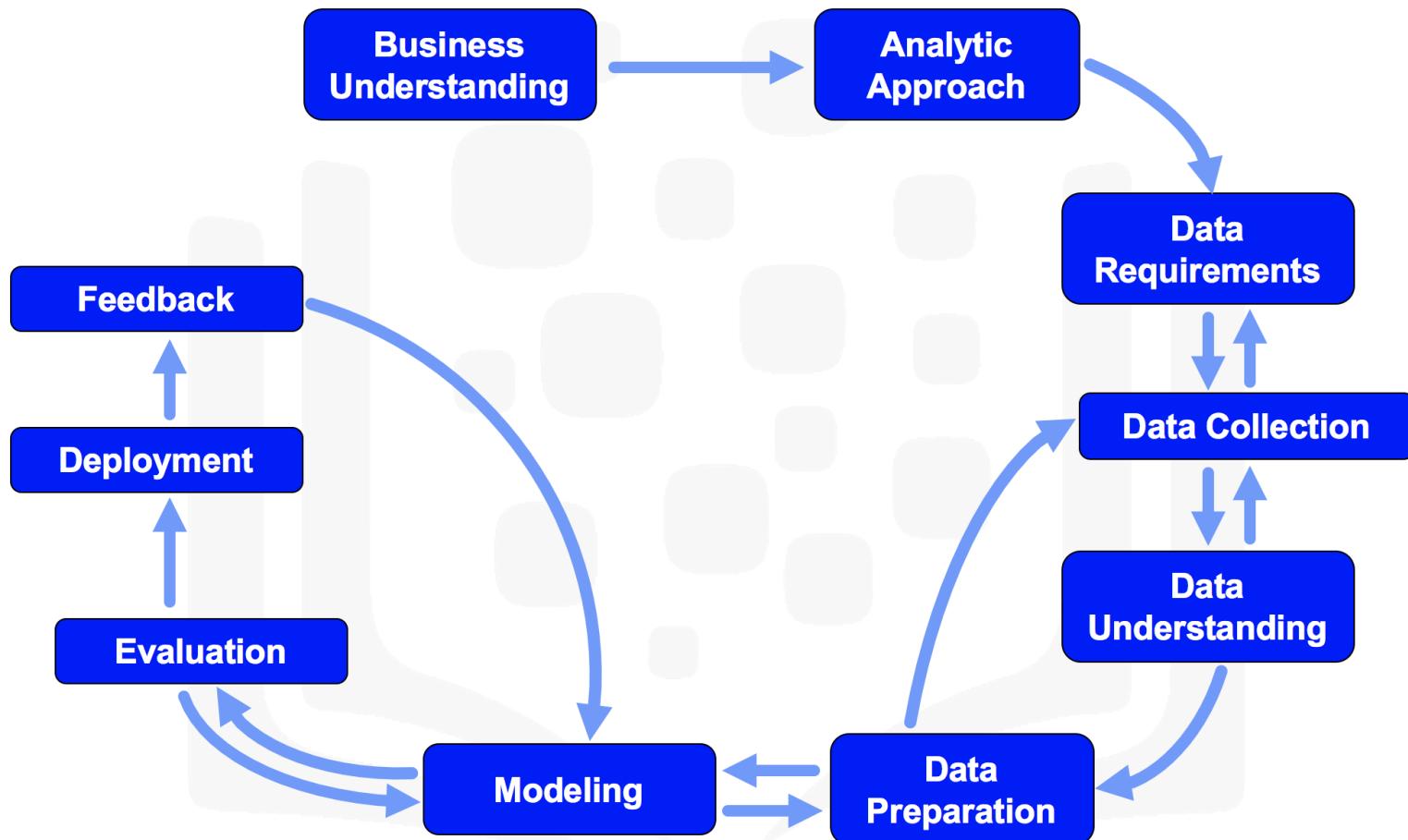
/ /

/ /

/ /

<https://euro2024pots.streamlit.app>

# Feedback



# From deployment to feedback

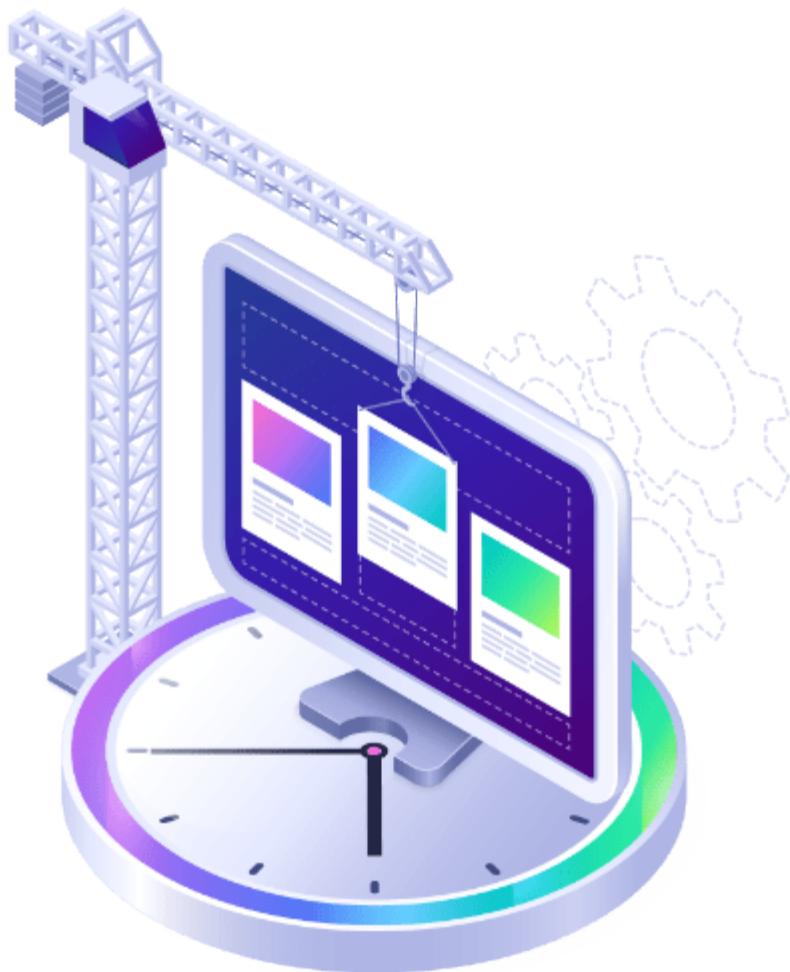
## Deployment

## Feedback

- Can you get constructive feedback into answering the question?



- Can you put the model into practice?



# From deployment to feedback



Once the model is evaluated, and the data scientist is confident it will work, it is deployed and put to the ultimate test

- Actual real-time use in the field

# Case study: Assessing model performance

Define review process



# Case study: Refinement



## Refine model

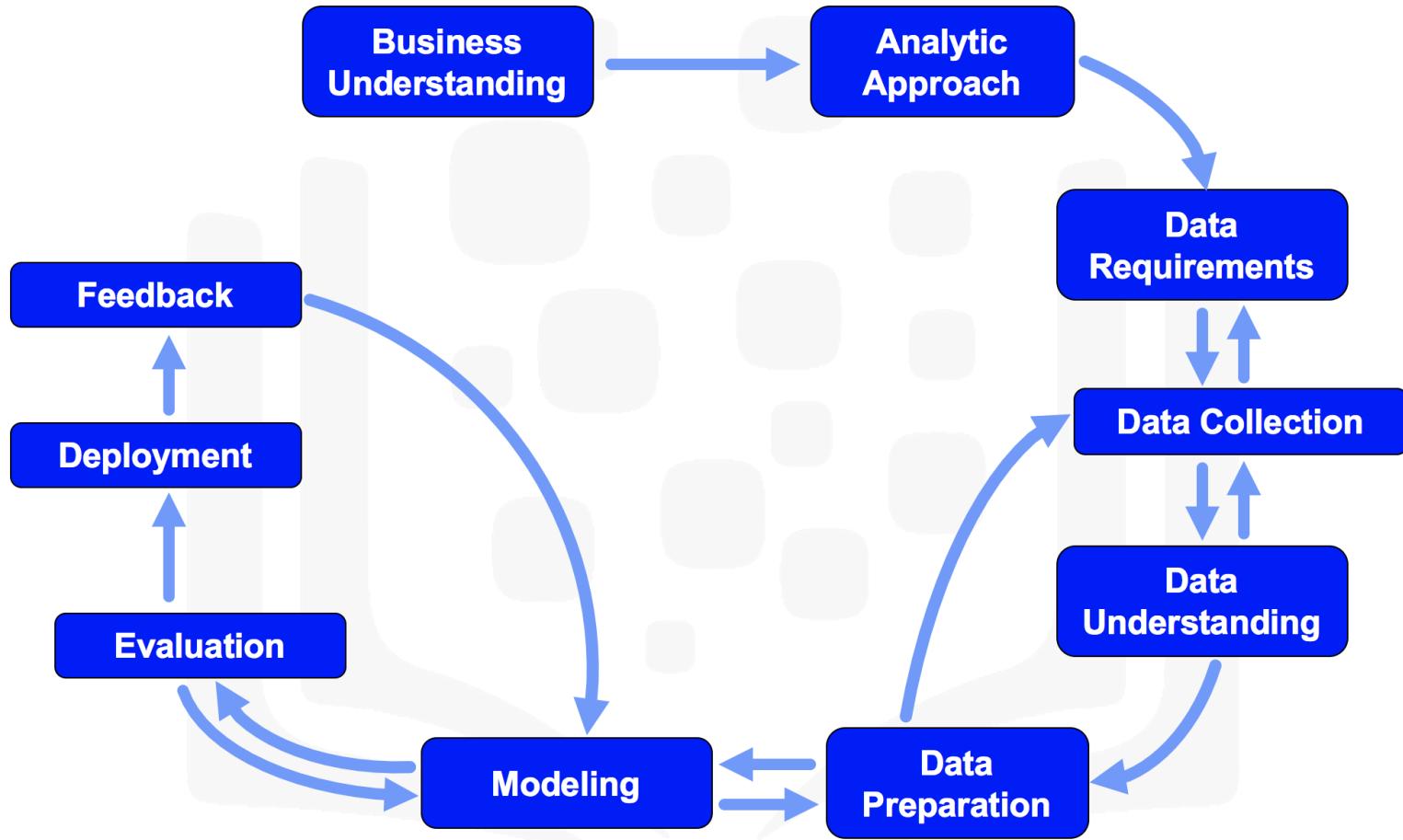
- Initial review after the first year of implementation
- Based on feedback data and knowledge gained
- Participation in interception program
- Possibly incorporate detailed pharmaceutical data originally deferred
- Other possible refinements as yet unknown

# Case study: Redeployment

- Review and refine intervention actions
- Redeploy
  - Continue modeling, deployment, feedback and refinement throughout the life of the intervention program

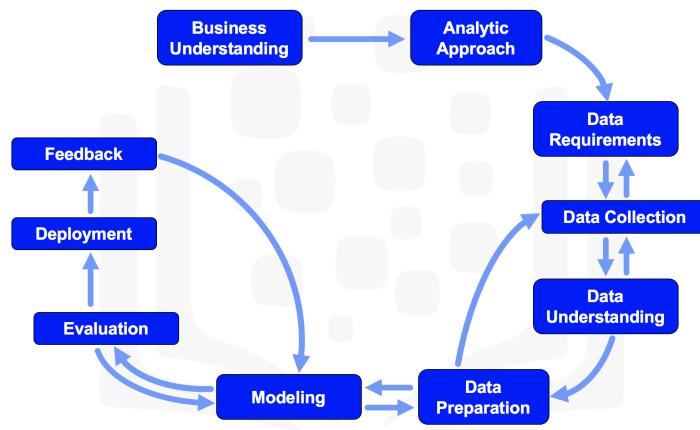


# Summary



# Data Science Methodology

Aims to answer the following questions in this prescribed sequence:

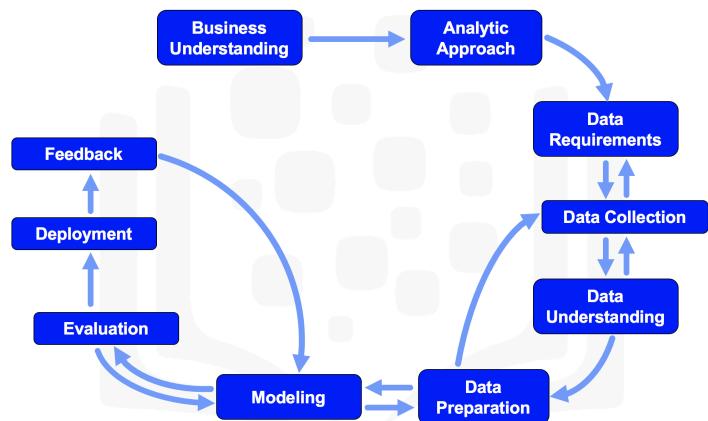


## From problem to approach:

1. What is the problem that you are trying to solve?
2. How can you use the data to answer the question?

# Data Science Methodology

Aims to answer the following questions in this prescribed sequence:

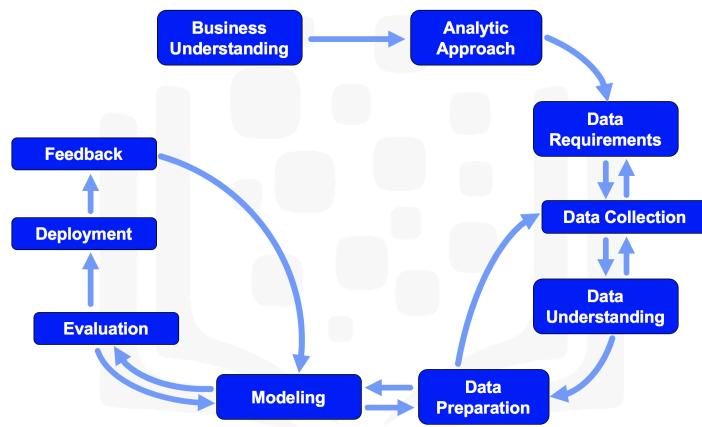


## Working with the data:

3. What data do you need to answer the question?
4. Where is the data coming from (identify sources) and how will you get it?
5. Is the data that you collected representative of the problem to be solved?
6. What additional work is required and work with the data?

# Data Science Methodology

Aims to answer the following questions in this prescribed sequence:



## Deriving the answer:

7. In what way can the data be visualized to get to the answer that is required?
8. Does the model used really answer the initial question or does it need to be adjusted?
9. Can you put the model into practice?
10. Can you get constructive feedback into answering the question?

A portrait of a man with short brown hair, wearing a dark tuxedo jacket over a white dress shirt with a black bow tie. He is looking slightly to his left with a neutral expression. The background is a dark, out-of-focus space filled with numerous small, glowing, colorful particles in shades of blue, green, yellow, and red, resembling a star field or a futuristic digital environment.

DONE