Data Science Methodology

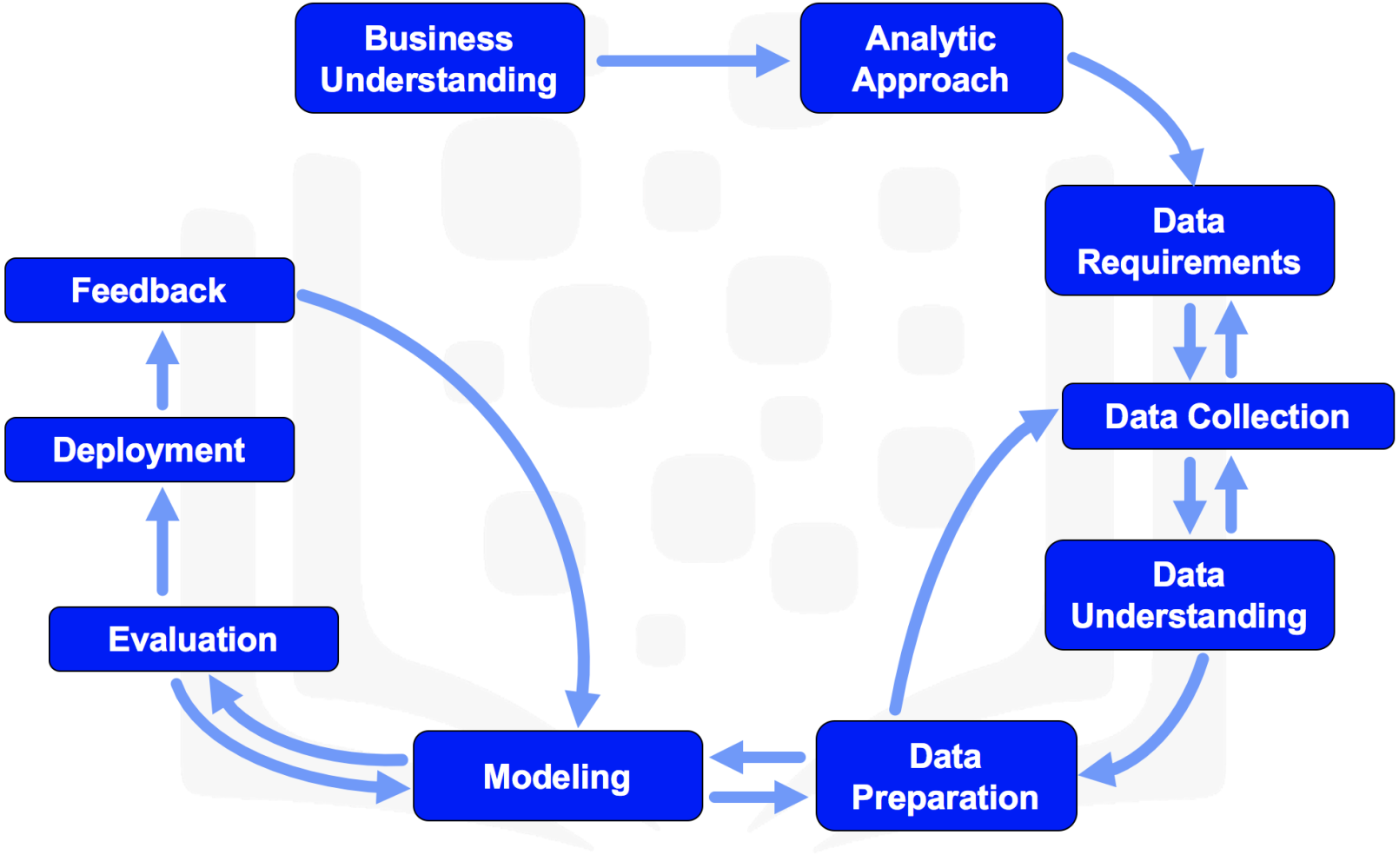
Data Science for Mathematicians

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

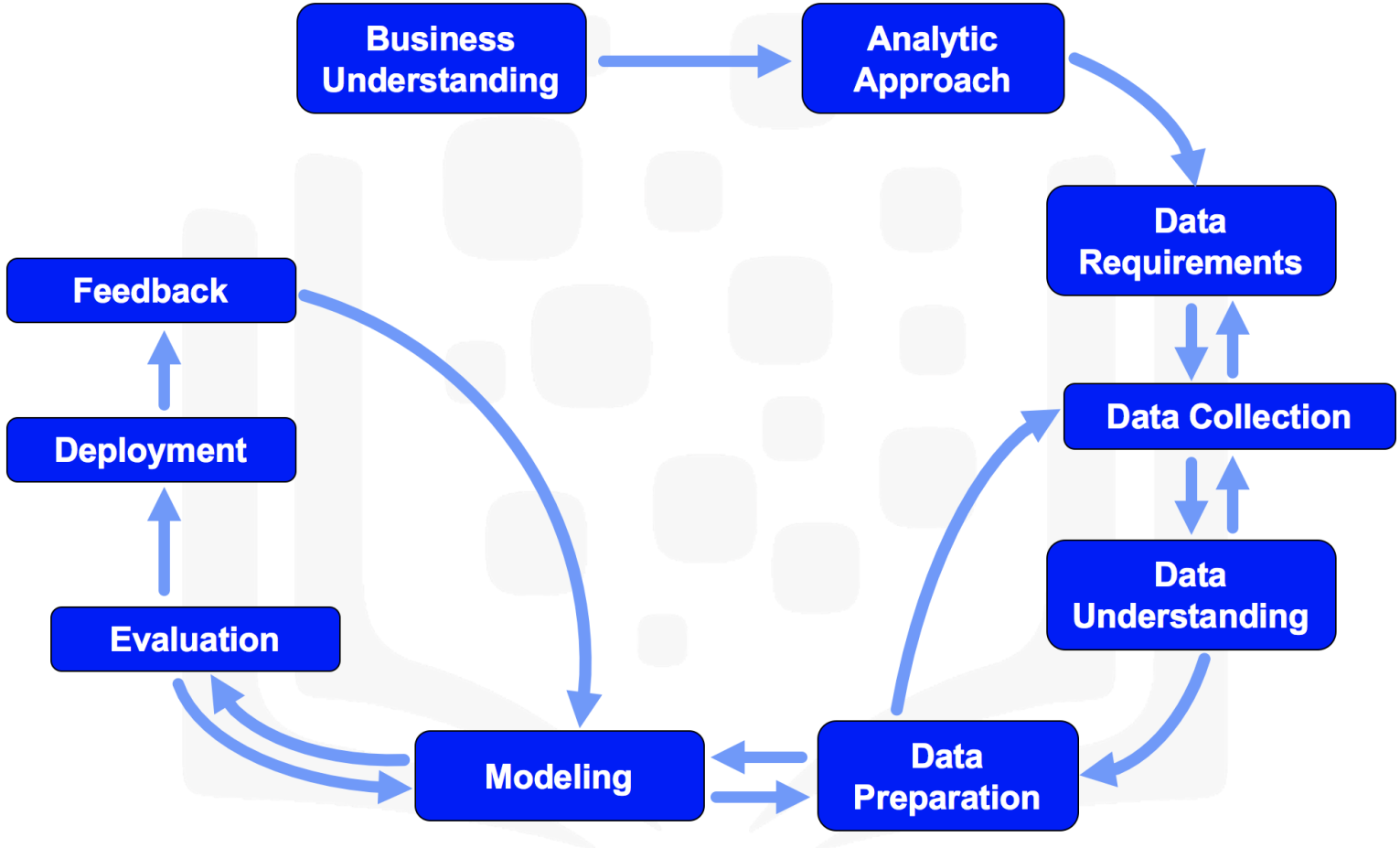
From Problem to Approach

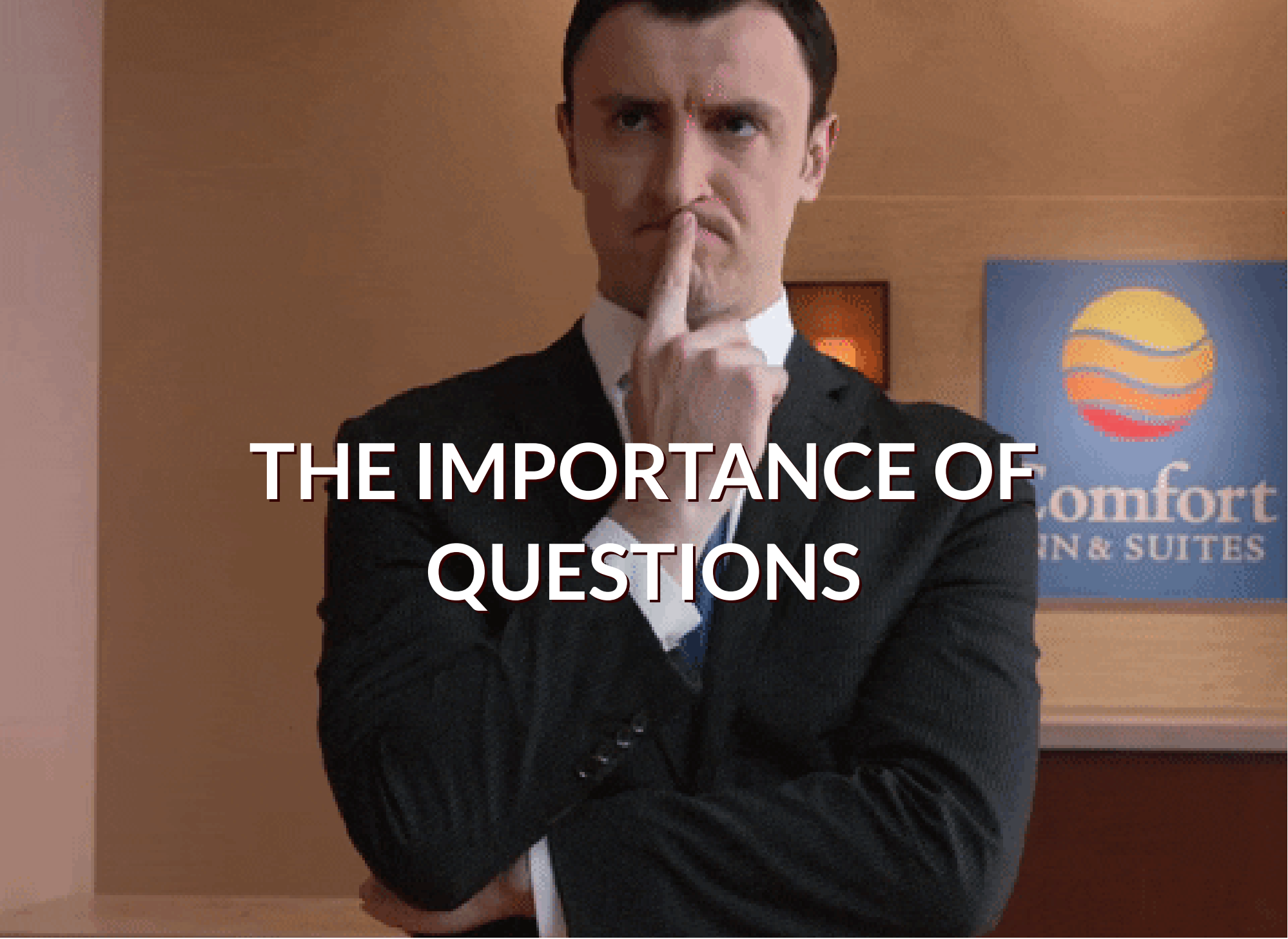
Learning Objectives

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




THE IMPORTANCE OF QUESTIONS

From understanding to approach

Business Understanding

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

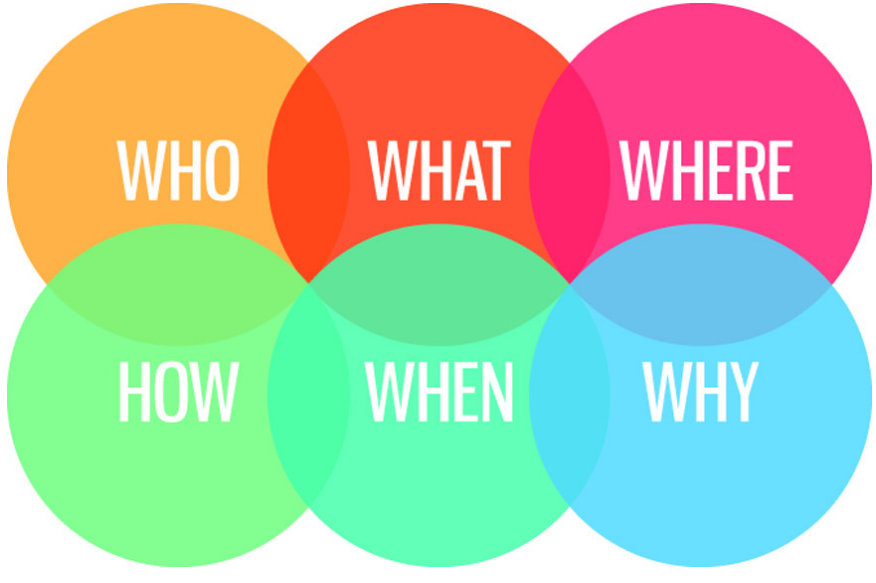
Analytic Approach

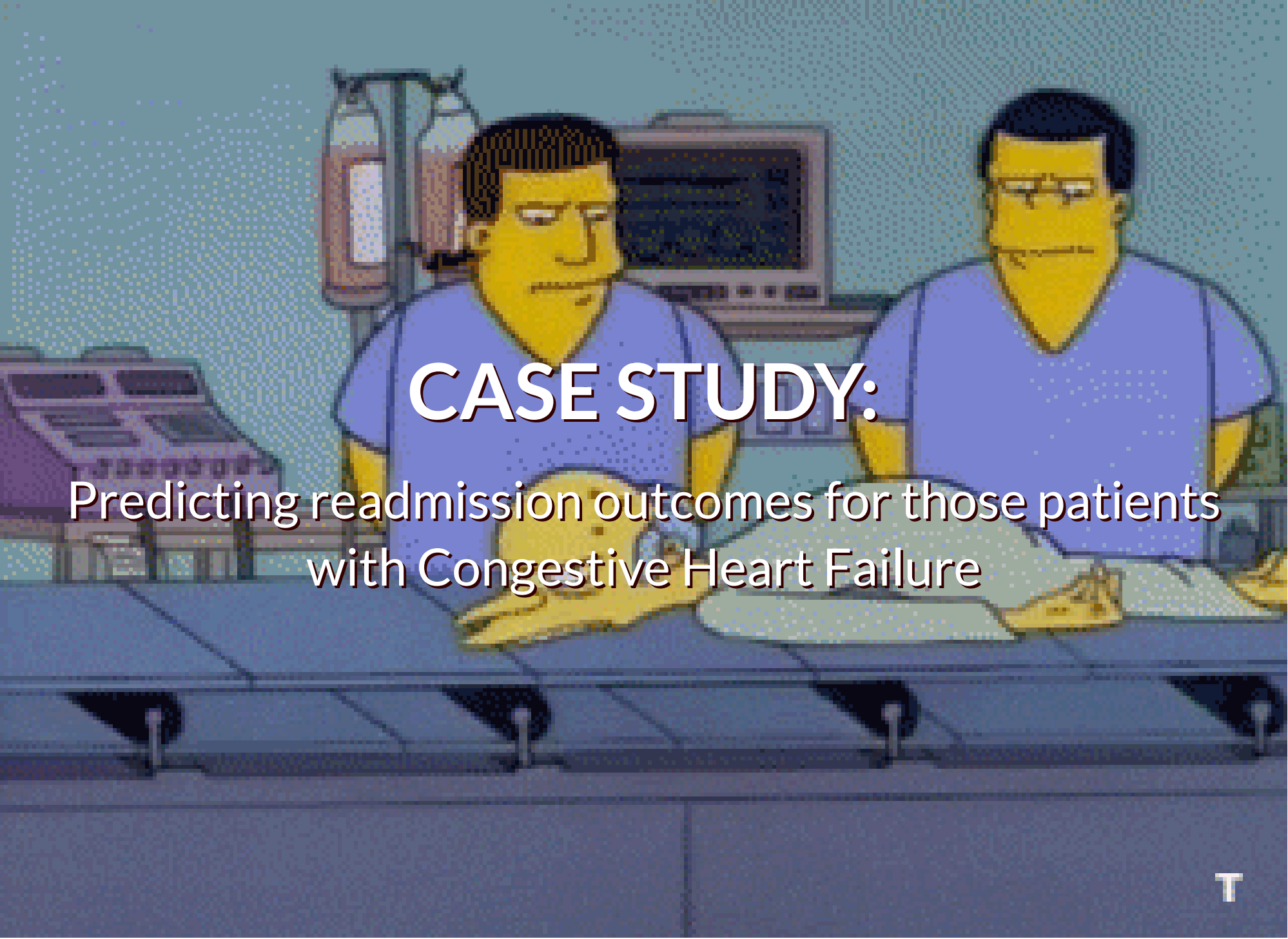
- **How can you use data to answer the question?**


WHAT'S THE GOAL?


SUPPORTING THE GOAL

Getting Stakeholders **buy-in** and **support**

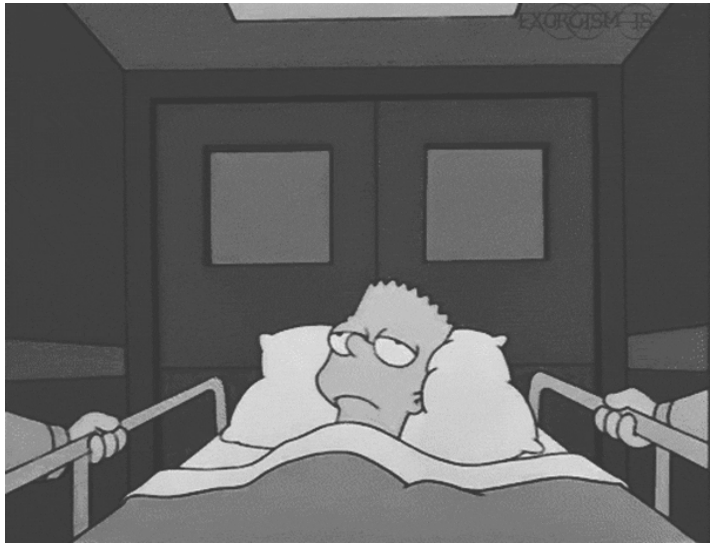




CASE STUDY:

Predicting readmission outcomes for those patients with Congestive Heart Failure

Case Study: CHF Re-admission



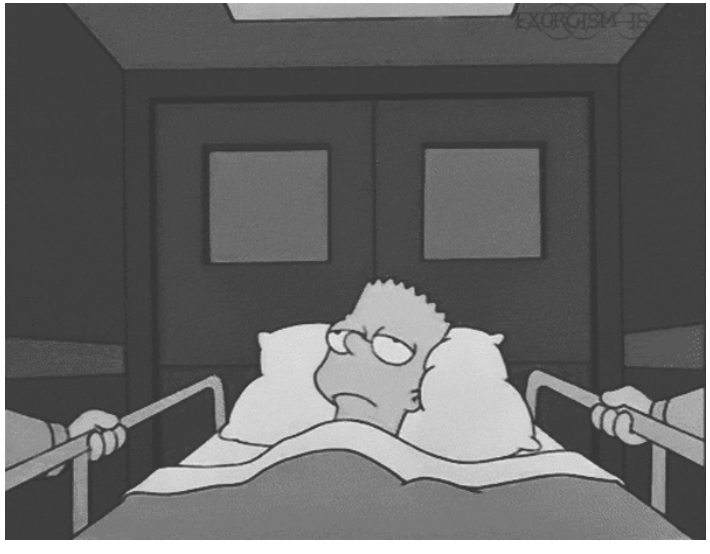
Define the **GOAL**

- To provide the care without increasing cost

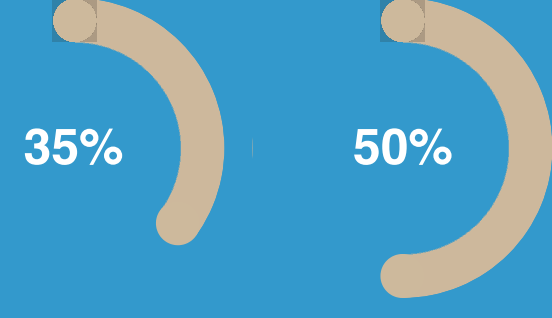
Define the **OBJECTIVE**

- To review the process to identify inefficiencies

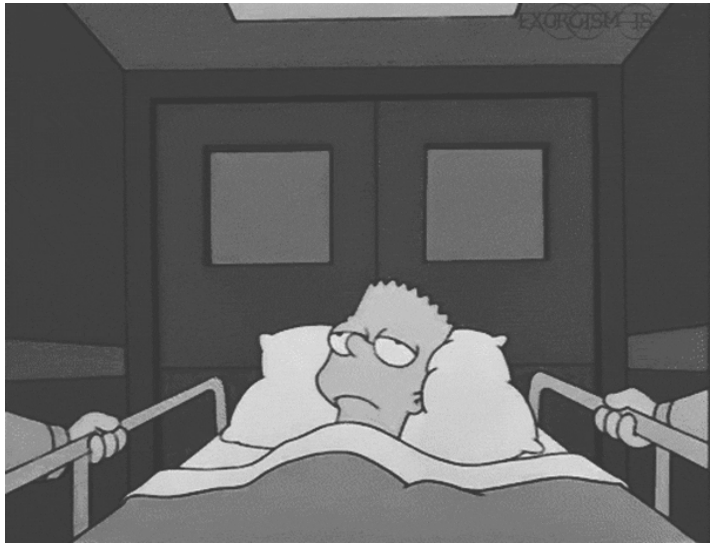
Case Study: CHF Re-admission



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**.

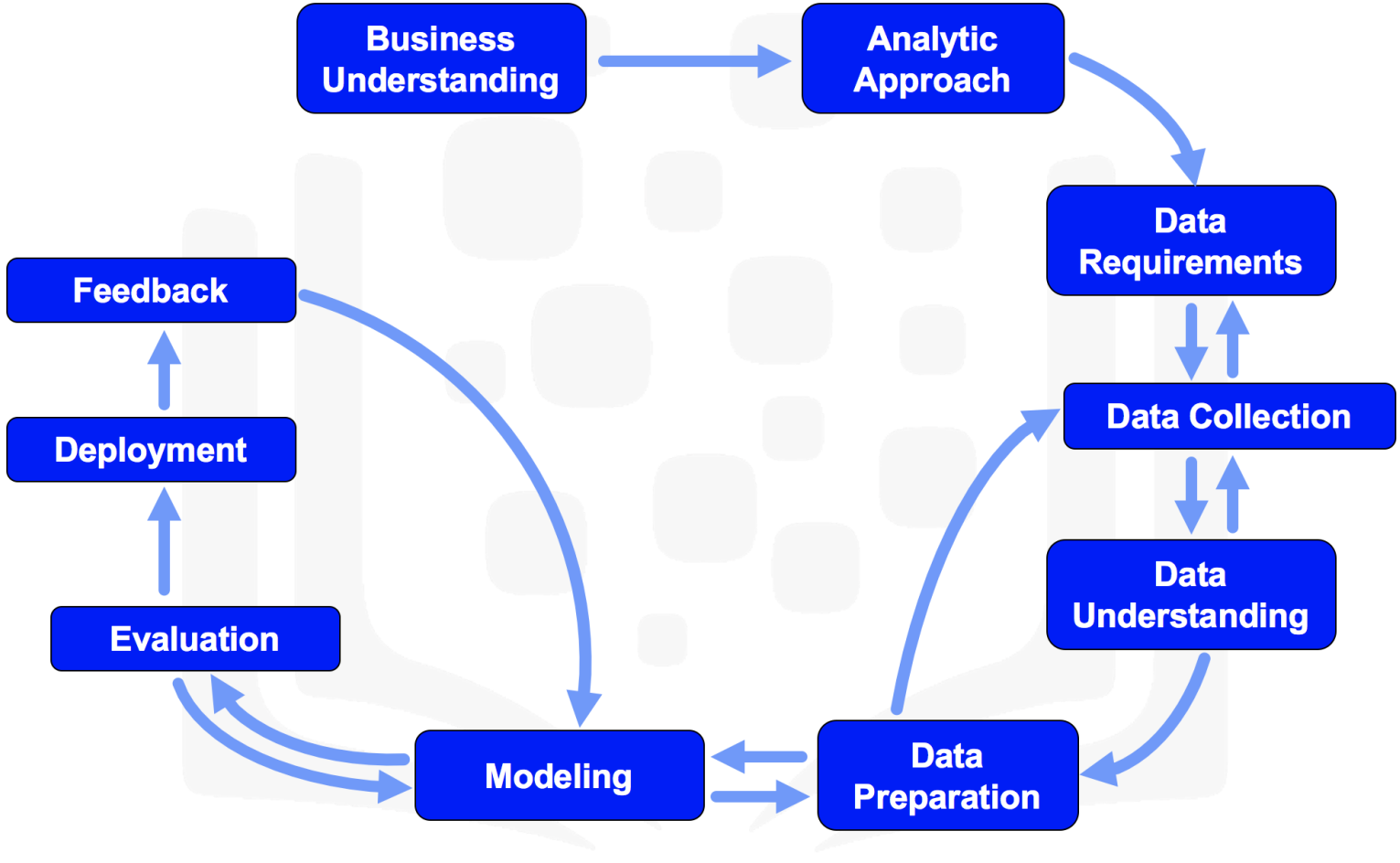


Case Study: CHF Re-admission



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

DESCRIPTIVE
Current status

DIAGNOSTIC (STATISTICAL ANALYSIS)
What happened? Why is this happening?

PREDICTIVE (FORECASTING)
What if this trends continue? What will happen next?

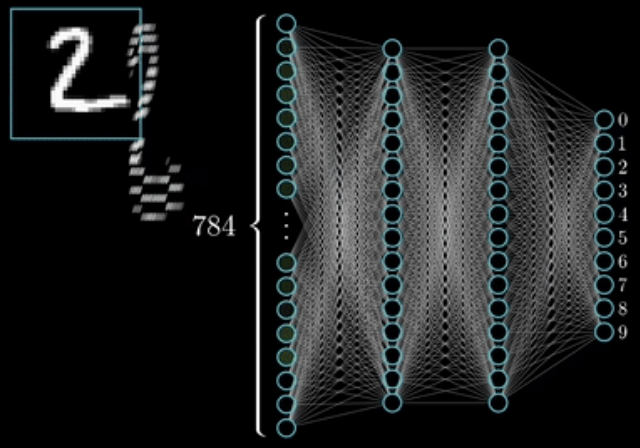
PRESCRIPTIVE
How do we solve it?

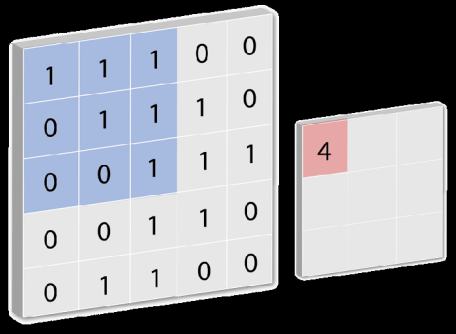
What are the types of questions?

- 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**

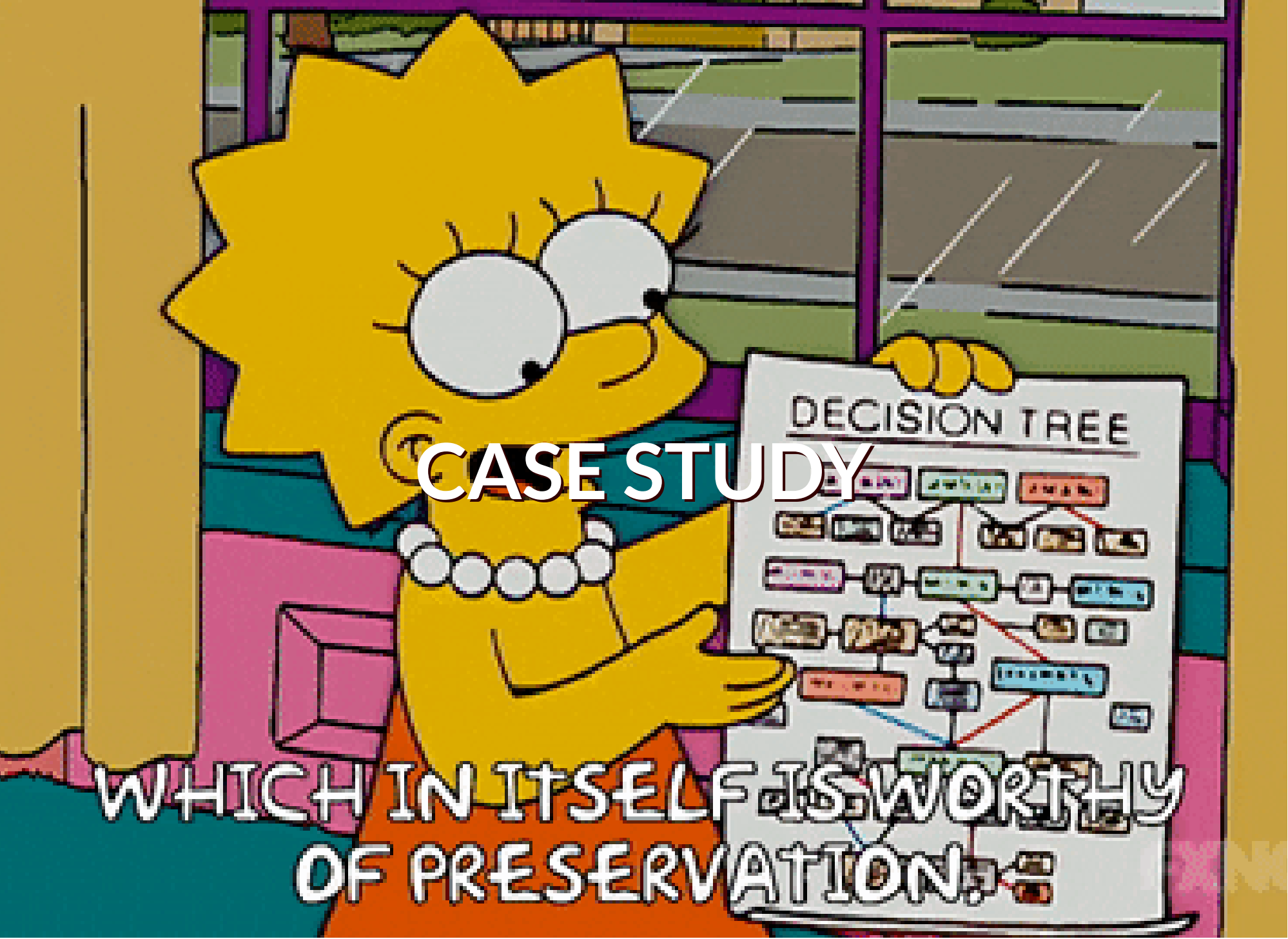
The correct approach depends on business requirements for the model

Will machine learning be utilized?

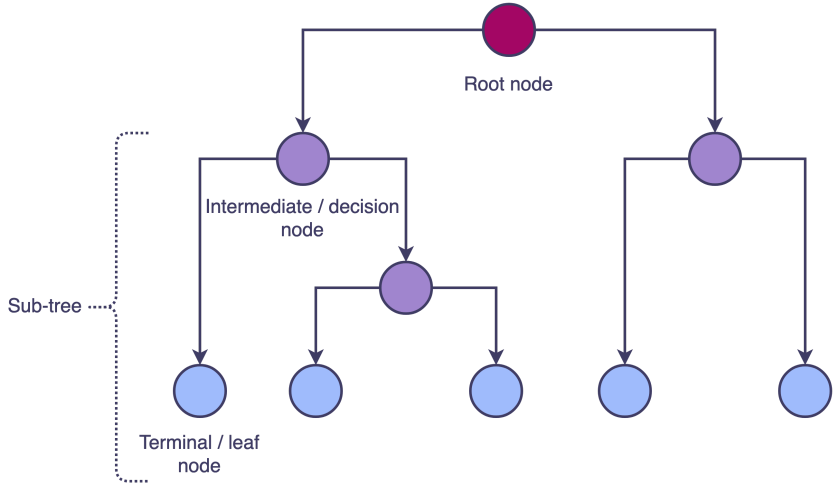




- Machine Learning
  - Learning without being explicitly programmed
  - Identifies relationships and trends in data that might otherwise not be accessible or identified
  - Use clustering ascociation approaches


CASE STUDY
DECISION TREE
WHICH IN ITSELF IS WORTHY
OF PRESERVATION.

Case study:
Decision tree classification selected!



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

Module 2

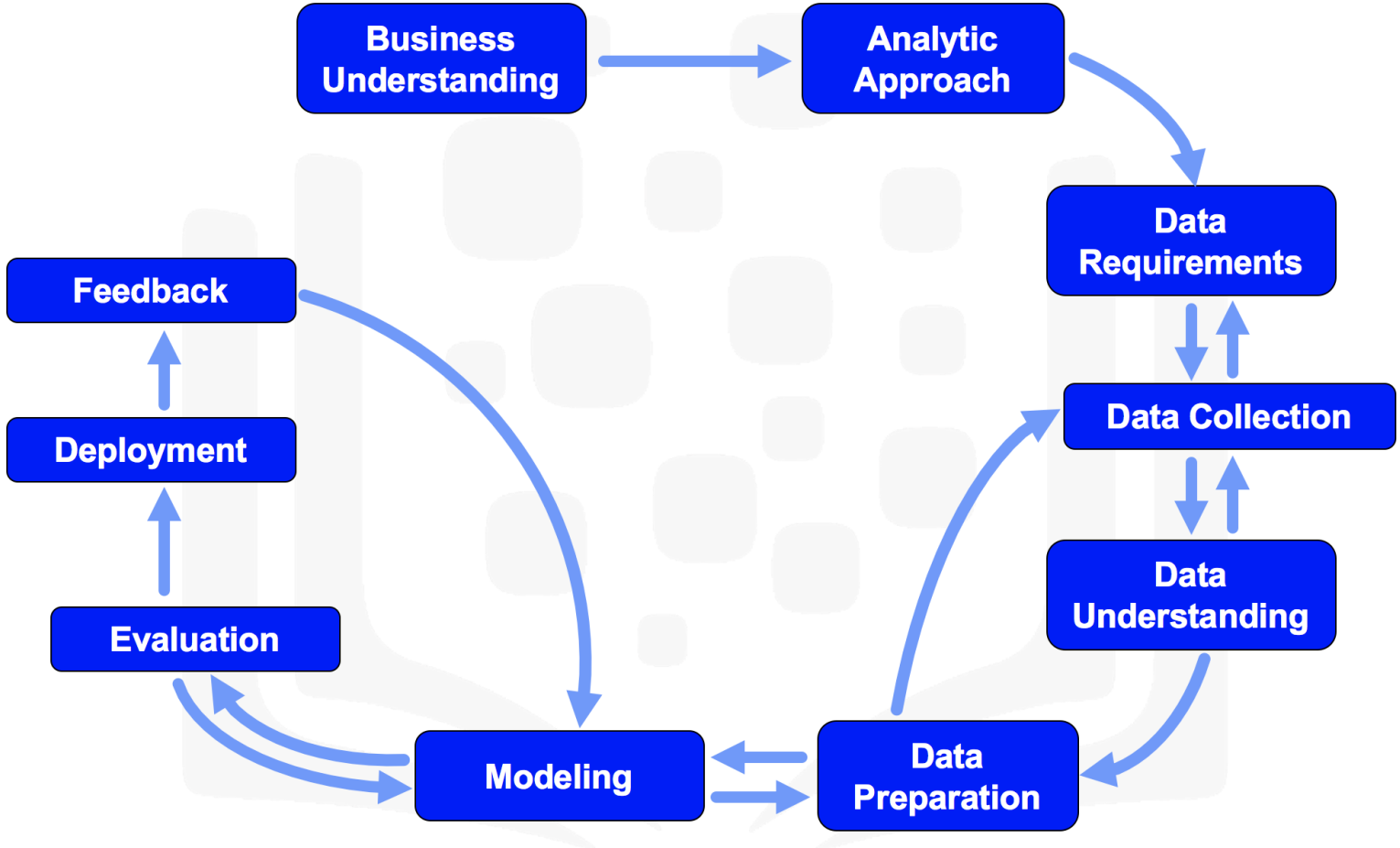
From Requirements to Collection

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

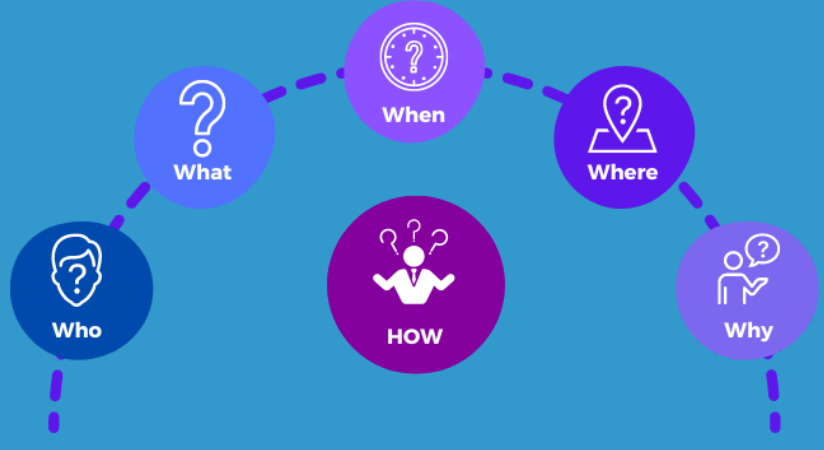



DATA REQUIREMENTS
>>> COOKING WITH DATA <<<

From requirements to collection

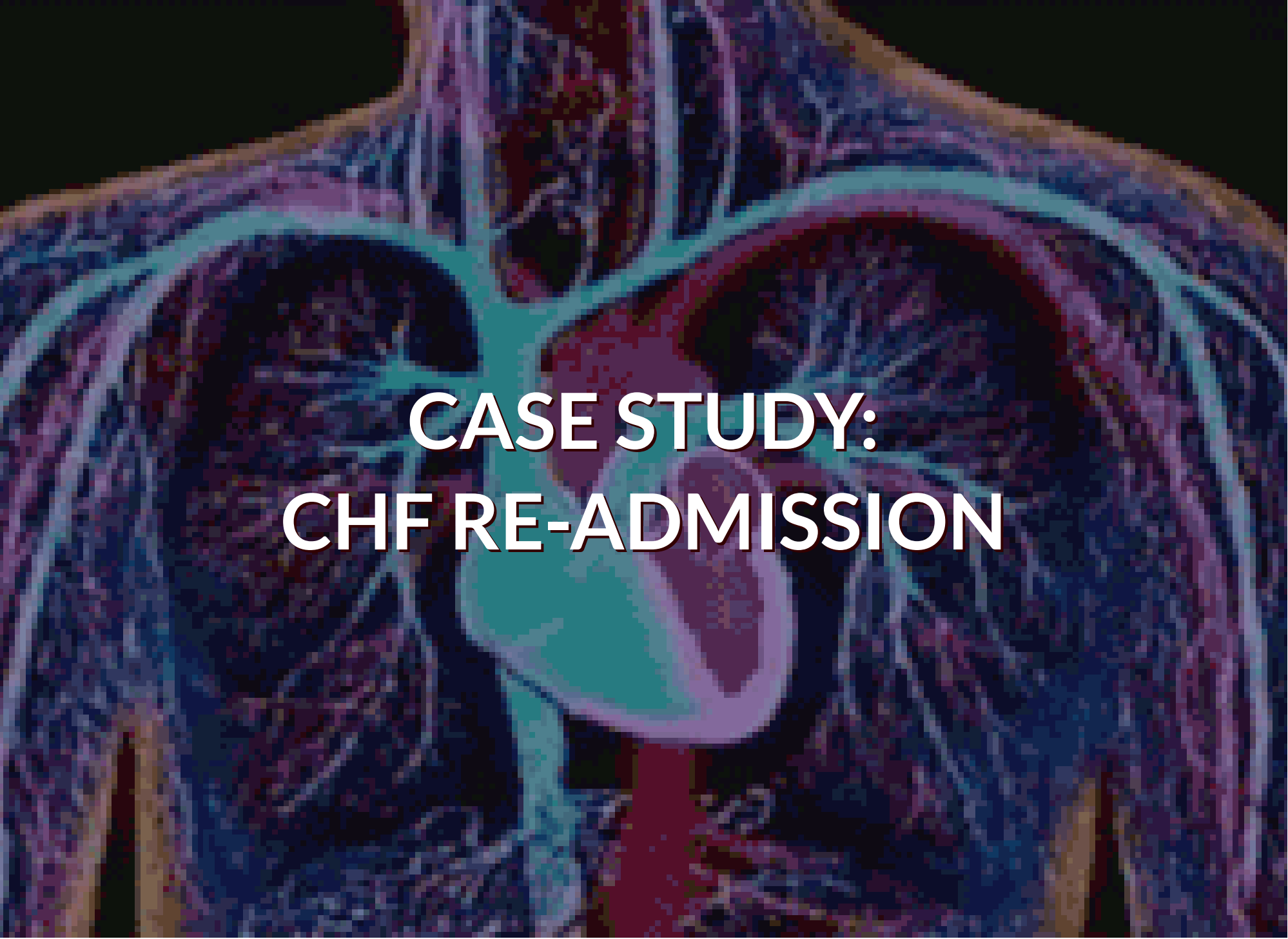
Data Requirements

- What are data requirements?



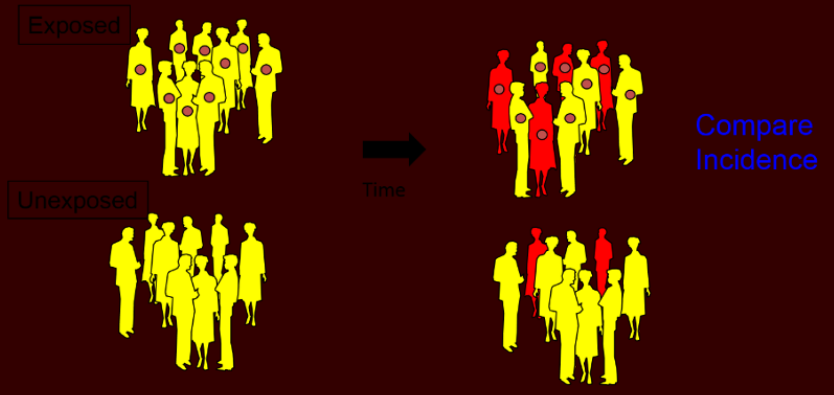
Data Collection

- What occurs during data collection?


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
  - Contiuous enrollment for at least 6 months prior to primary CHF admission
  - Disqualifying conditions

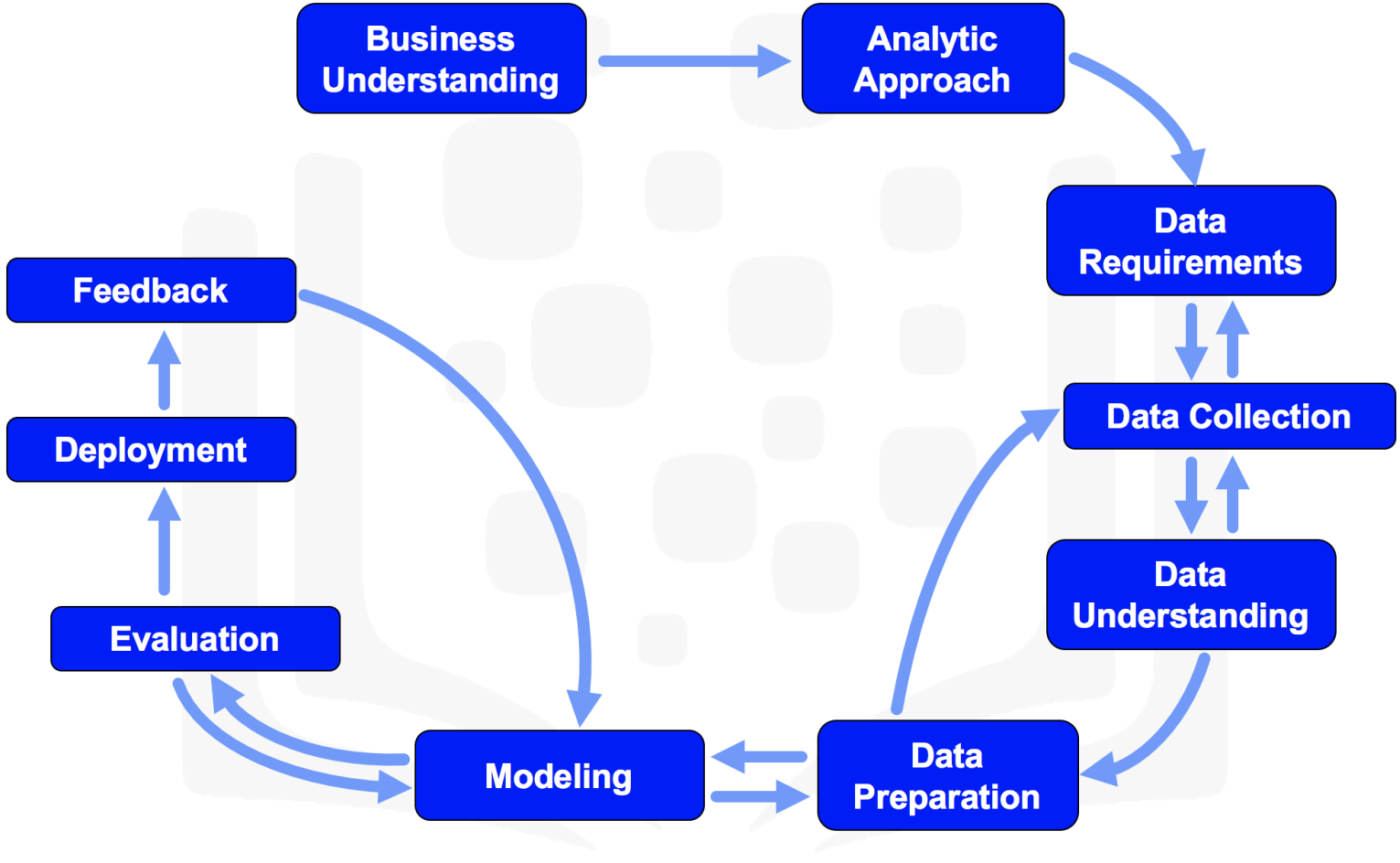


Case Study: Defining the data

- Contents, formats, representations suitable for decision tree classifier
  - One record per petient with columns represention variables (*dependent variable and predictors*)
  - Content covering all aspects of each petient'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

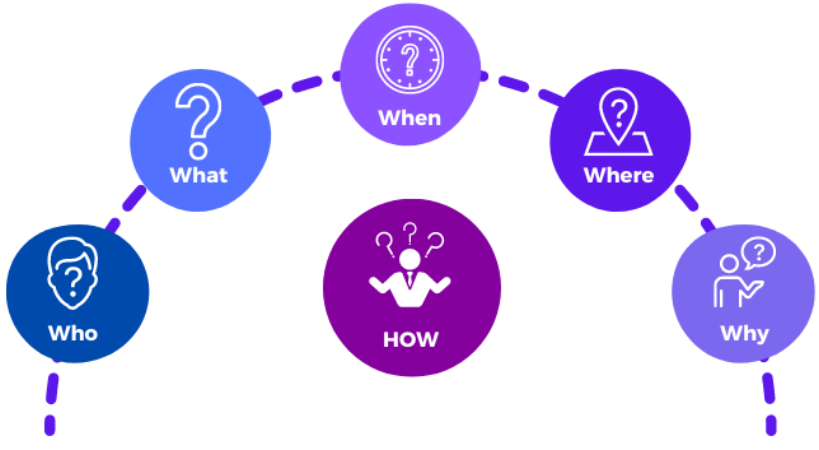



DATA COLLECTION
>>> INGREDIENTS? <<<

From requirements to collection

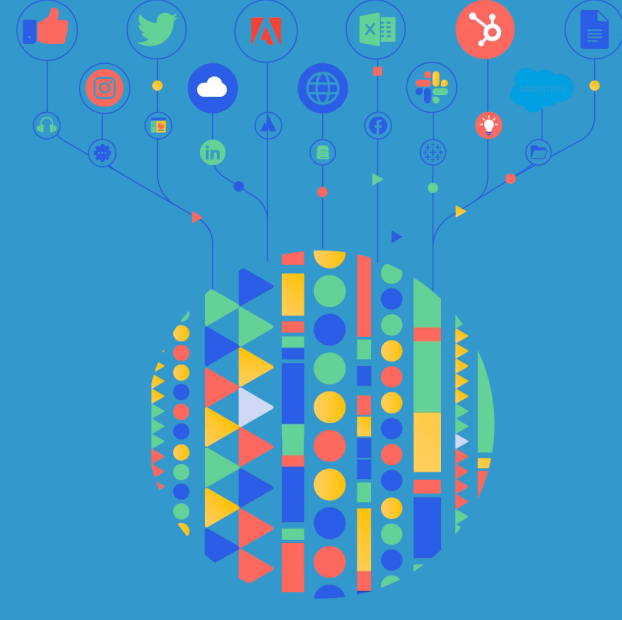
Data Requirements

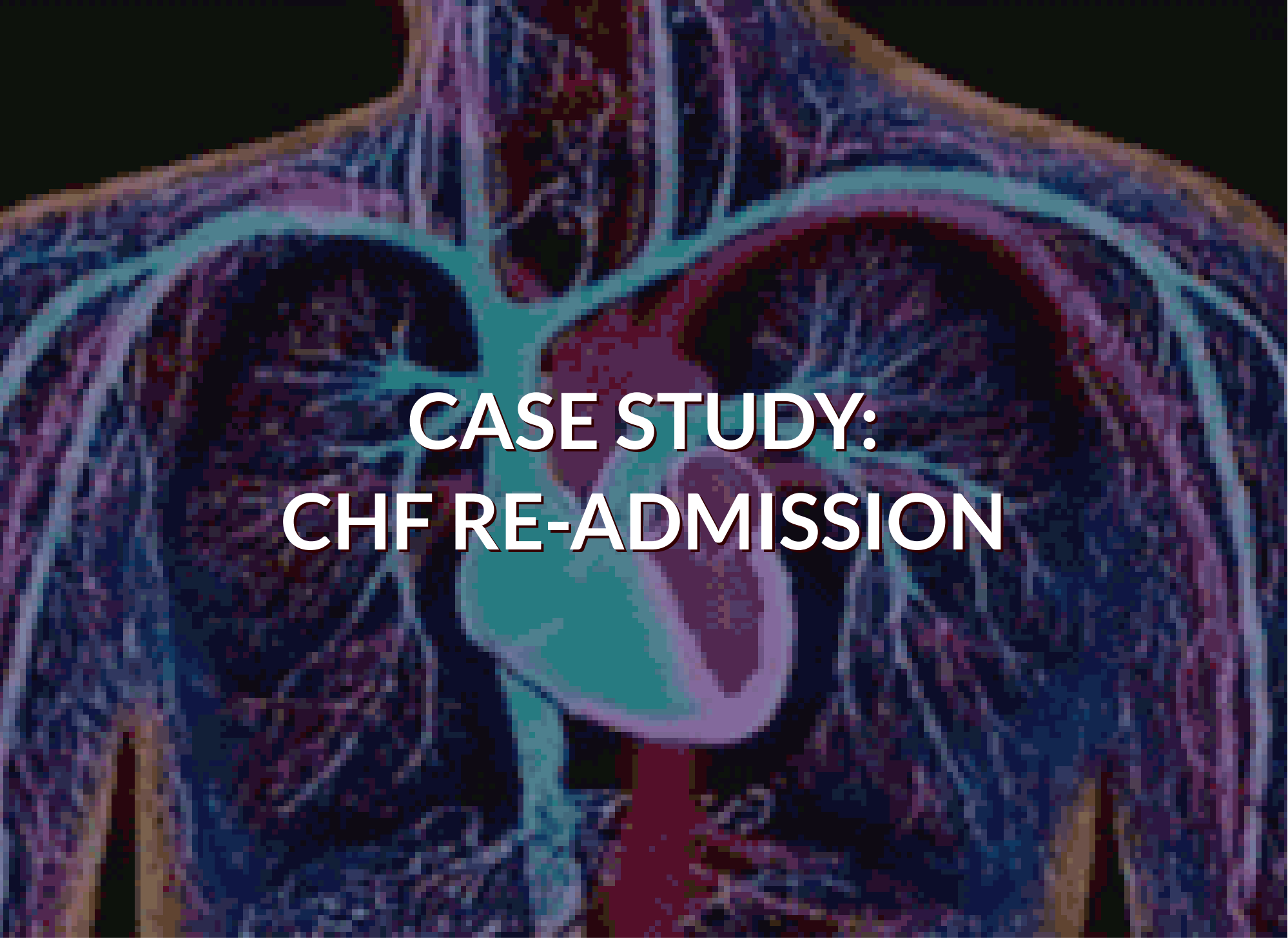
- What are data requirements?



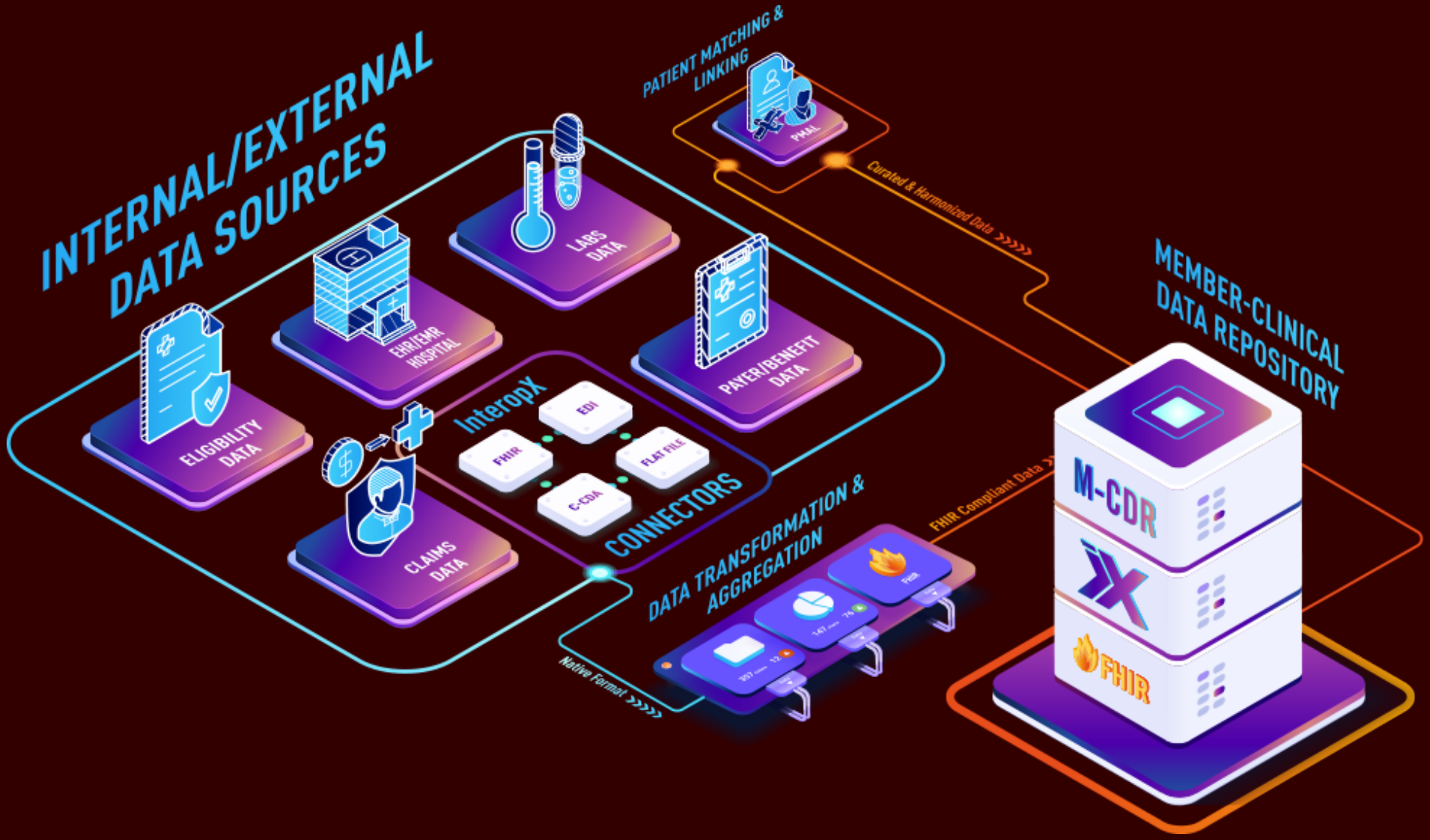
Data Collection

- What occurs during data collection?



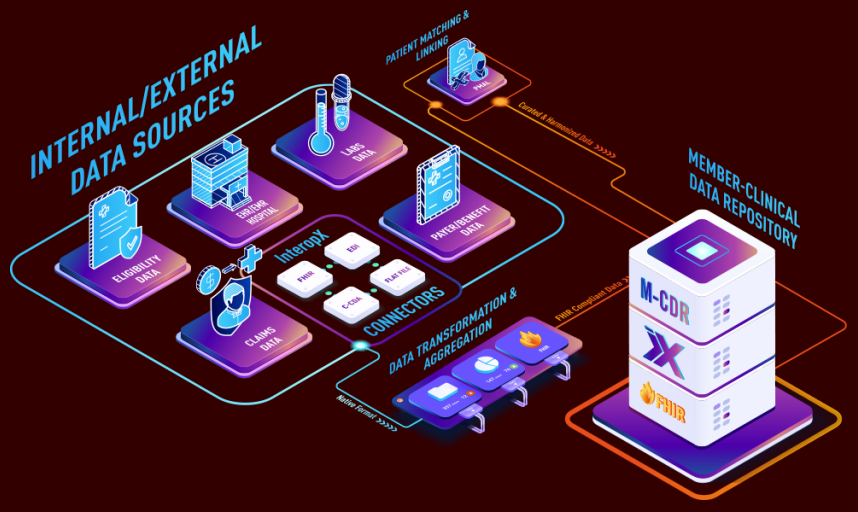

CASE STUDY:
CHF RE-ADMISSION

Case Study:
Gathering available data


INTERNAL/EXTERNAL DATA SOURCES
PATIENT MATCHING & LINKING
PMAL
Curated & Harmonized Data
MEMBER-CLINICAL DATA REPOSITORY
LABS DATA
EHR/EMR HOSPITAL
PAYER/BENEFIT DATA
ELIGIBILITY DATA
CLAIMS DATA
InteropX
EDI
FHIR
FLAT FILE
C-CDA
CONNECTORS
DATA TRANSFORMATION & AGGREGATION
FHIR Compliant Data
Native Format
M-CDR
FHIR

Case Study: Gathering available data

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

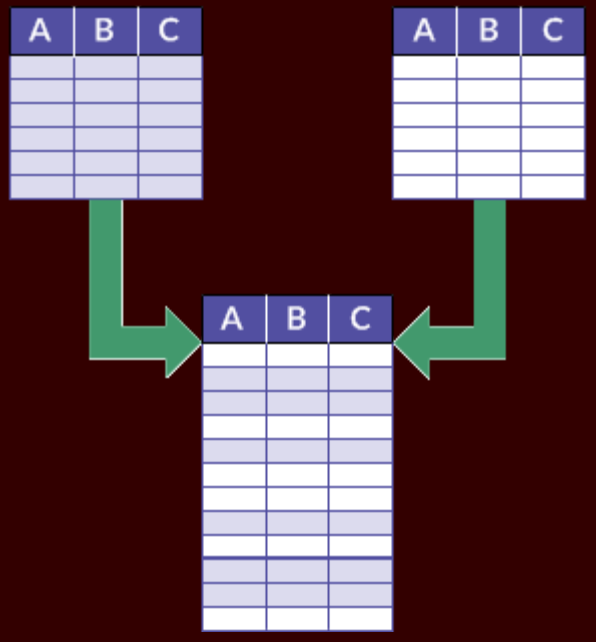


Case Study: Deferring inaccessible data

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

Case Study: Merging data

Eliminate redundant data



Module 3

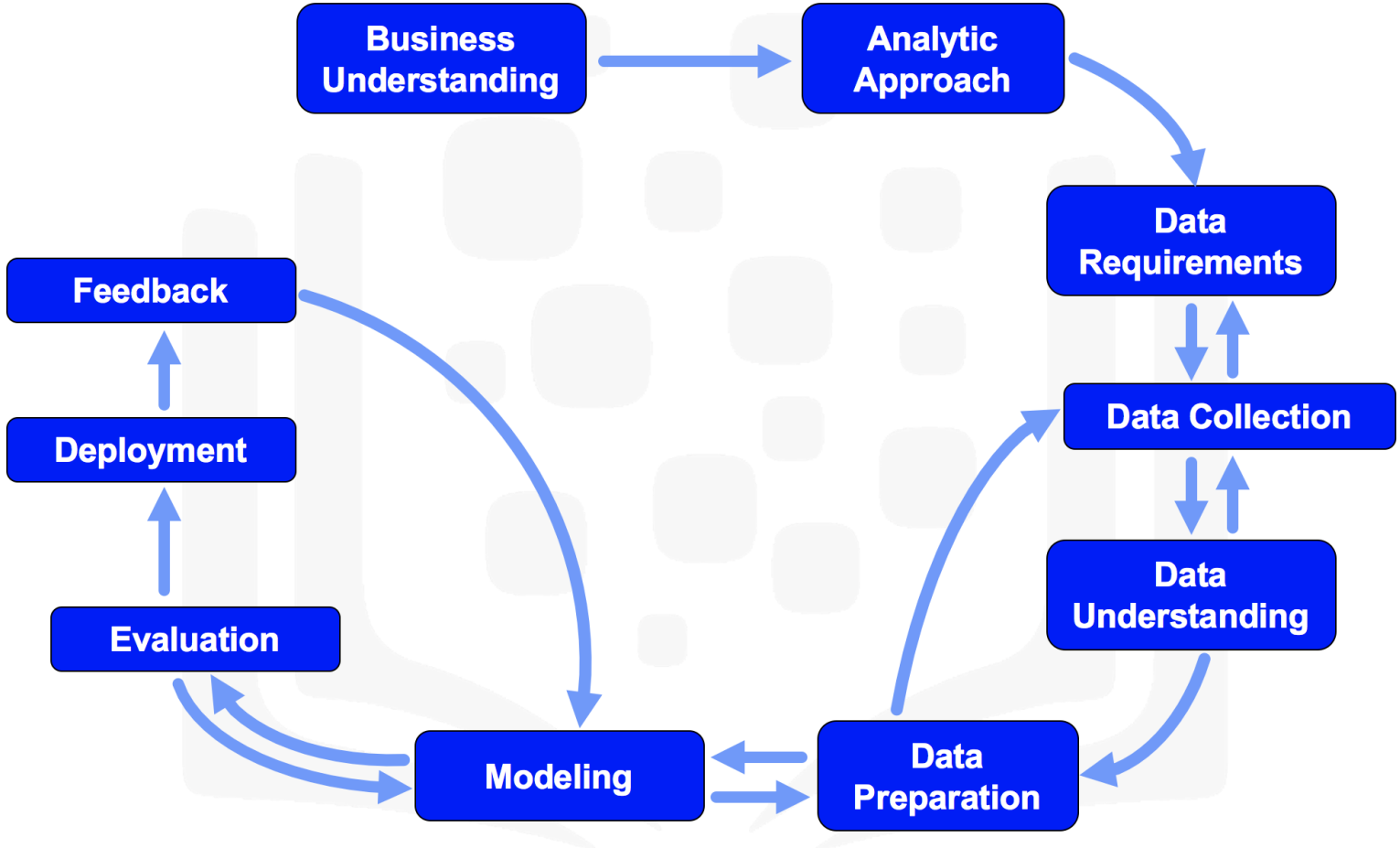
From Understanding to Preparation

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

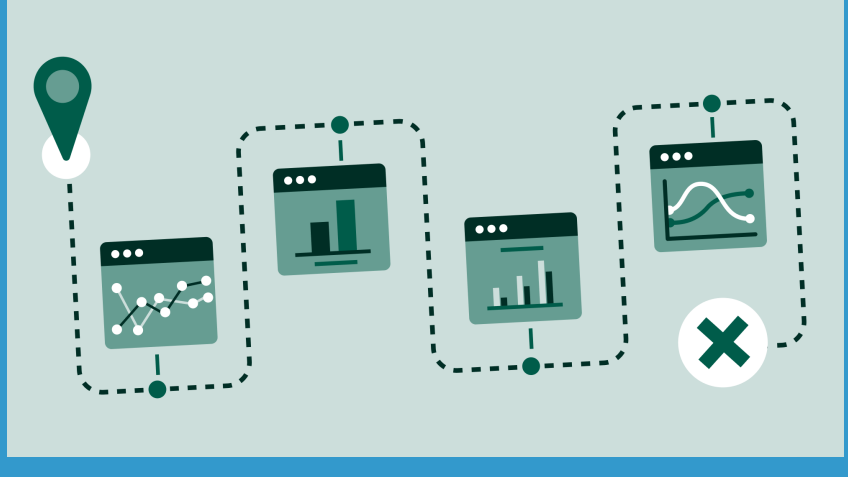



DATA UNDERSTANDING

From understanding to preparation

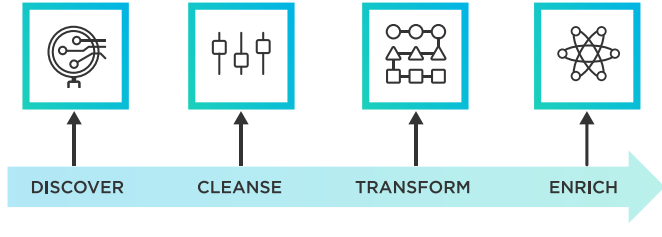
Data Understanding

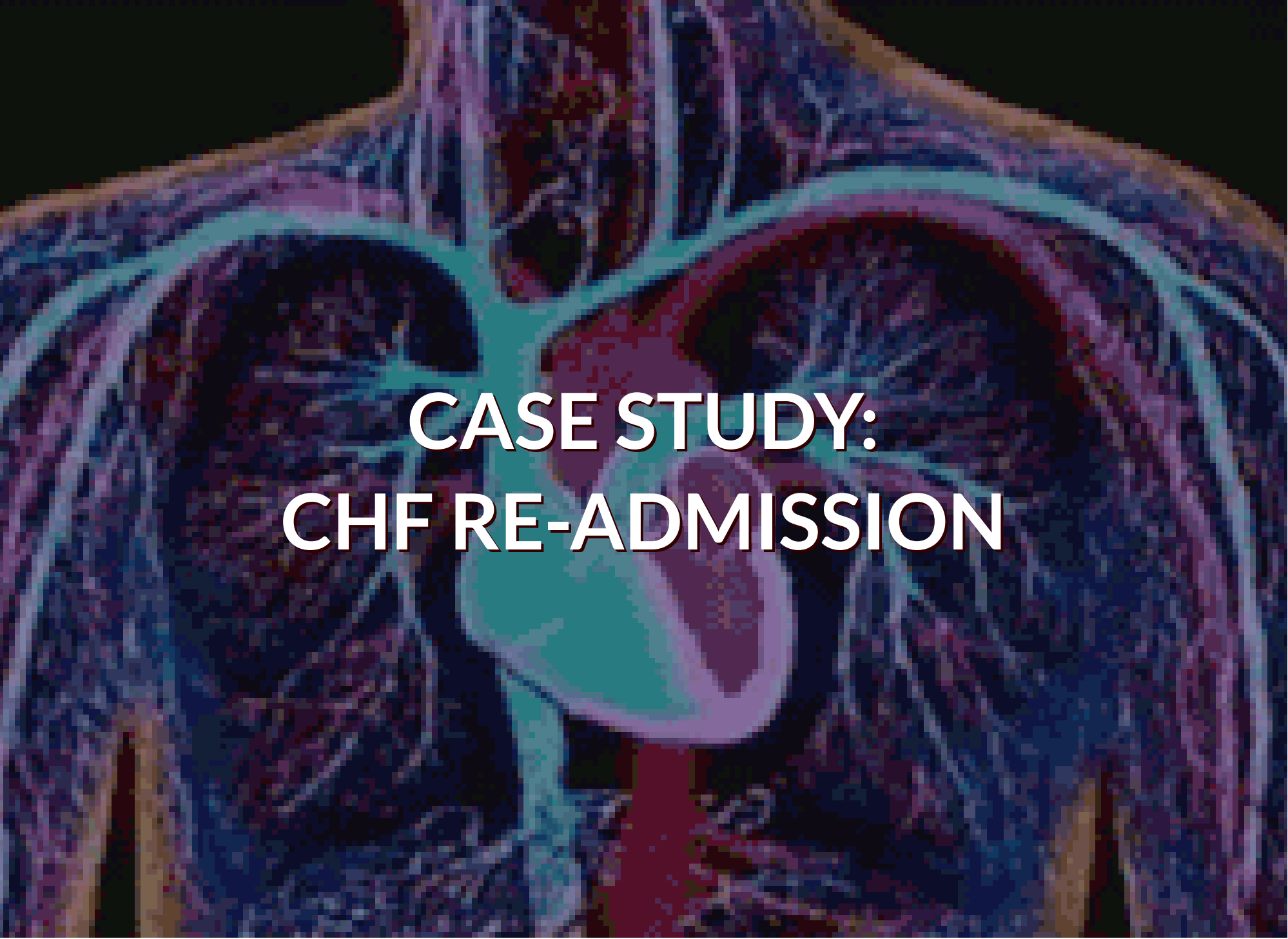
- What does it mean to "prepare" or "clean" data?



Data Preparation

- What are ways in which data is prepared?



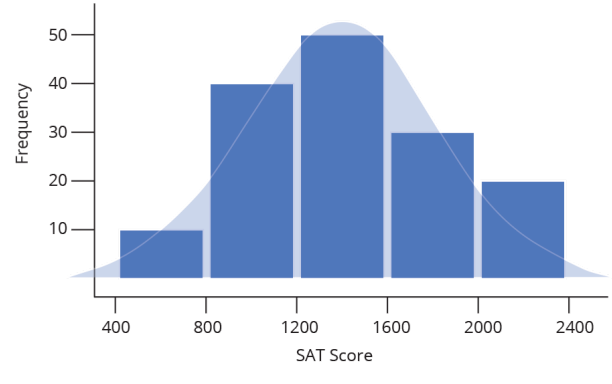

CASE STUDY:
CHF RE-ADMISSION

Case Study: Understanding the data

- Descriptive statistics
  - Univariate statistics
  - Pairwise correlations
  - Histogram

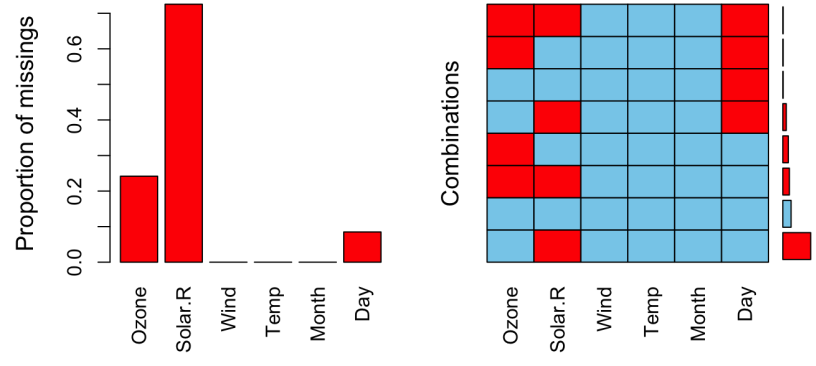
$$f(a) + \sum_{k=1}^{n} \frac{\mathrm{d}^k}{\mathrm{d}t^k} f(u(t)) + \int_0^1 \frac{(1-t)^n}{n!} \frac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}} f(u(t)) \mathrm{d}t$$

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

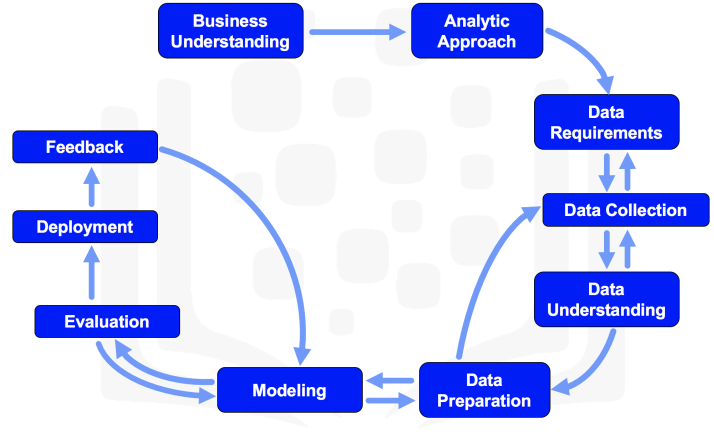
- Data quality
  - Missing values
  - Invalid or misleading values



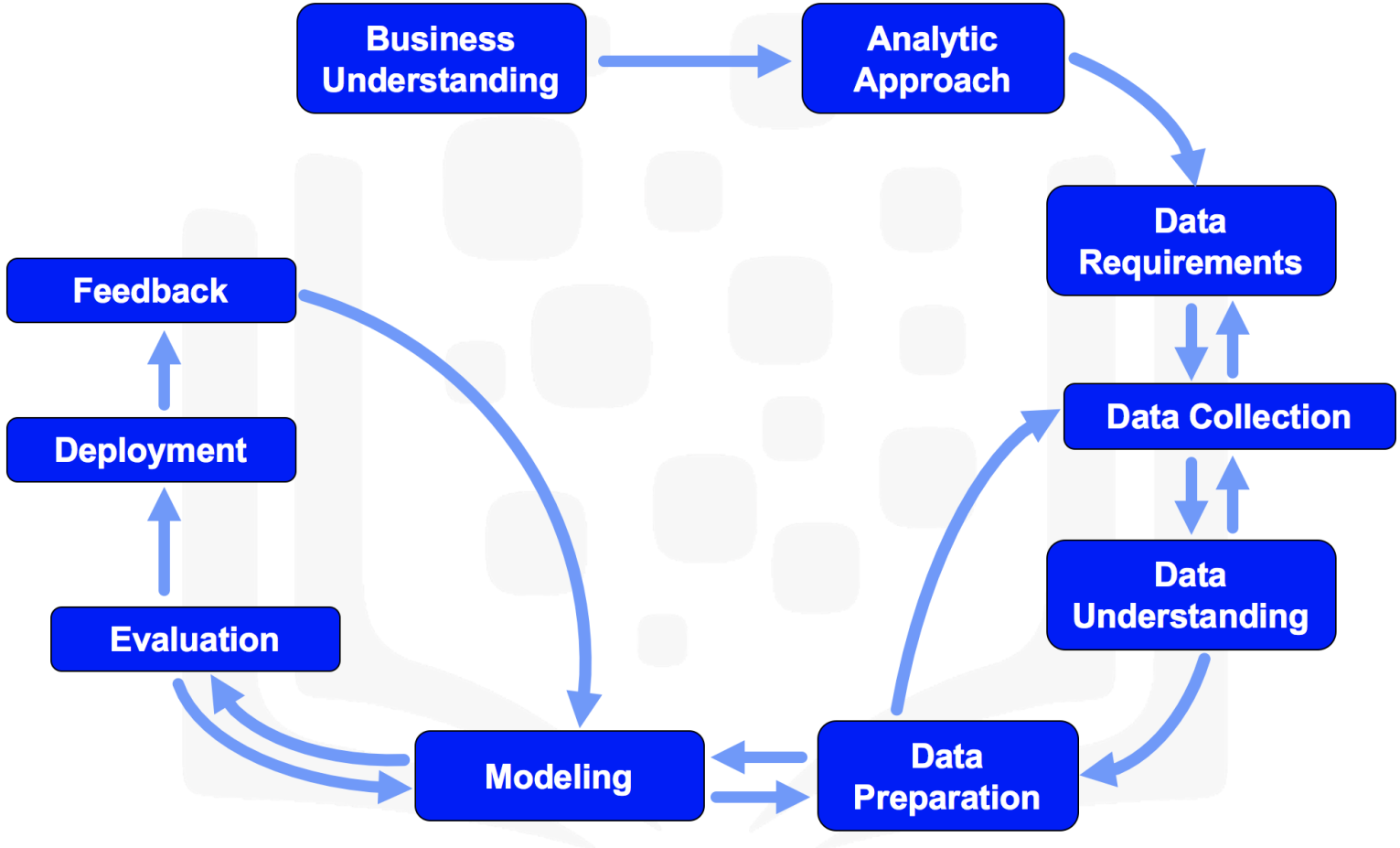
	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

Case Study:
This is an iterative process

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



Data Preparation




DATA PREPARATION
>>> CLEANING <<<


DATA PREPARATION
>>> TRANSFORM <<<

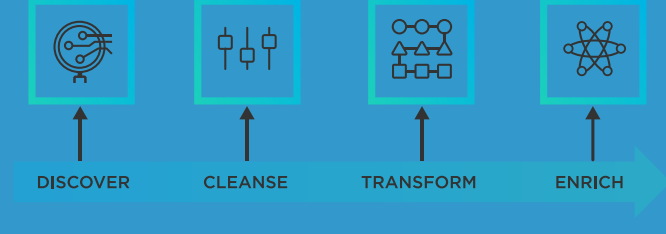
From understanding to preparation

Data Understanding

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

Data Preparation

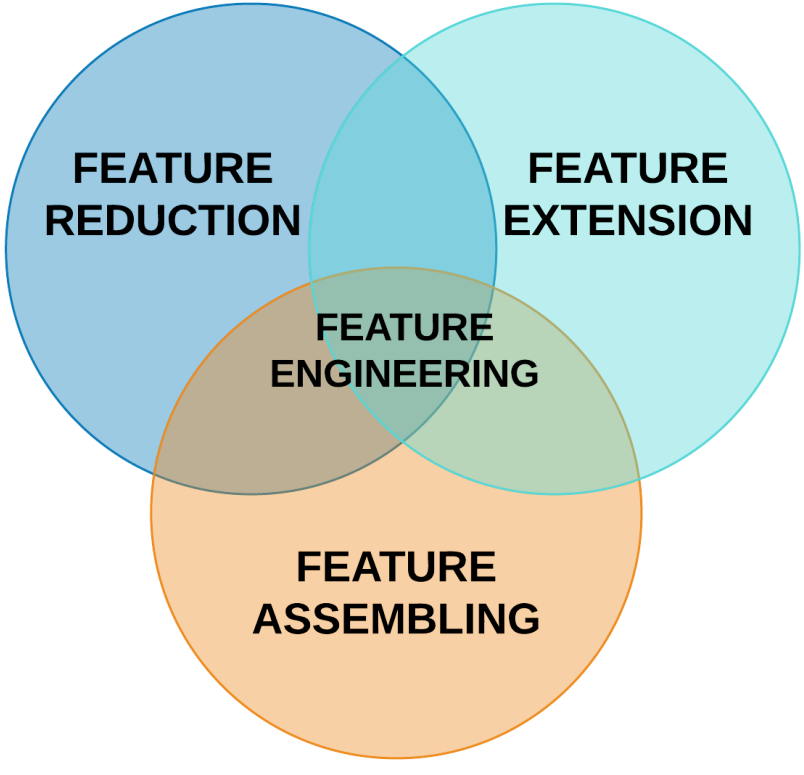
- What are ways in which data is prepared?



Example of data cleaning

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

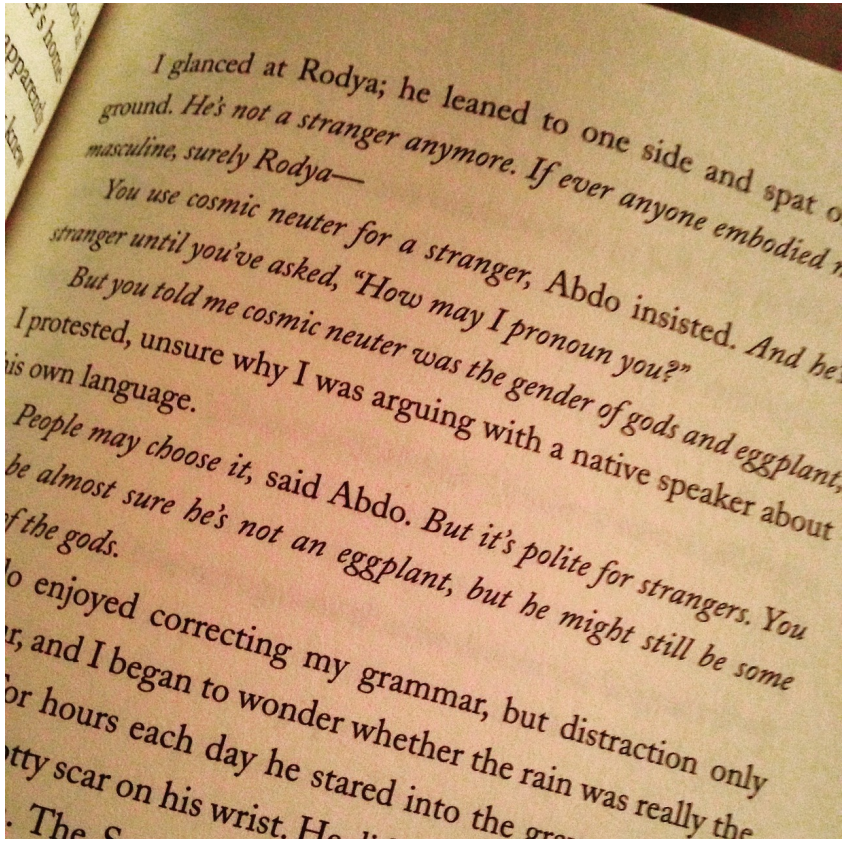
Using domian knowledge



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

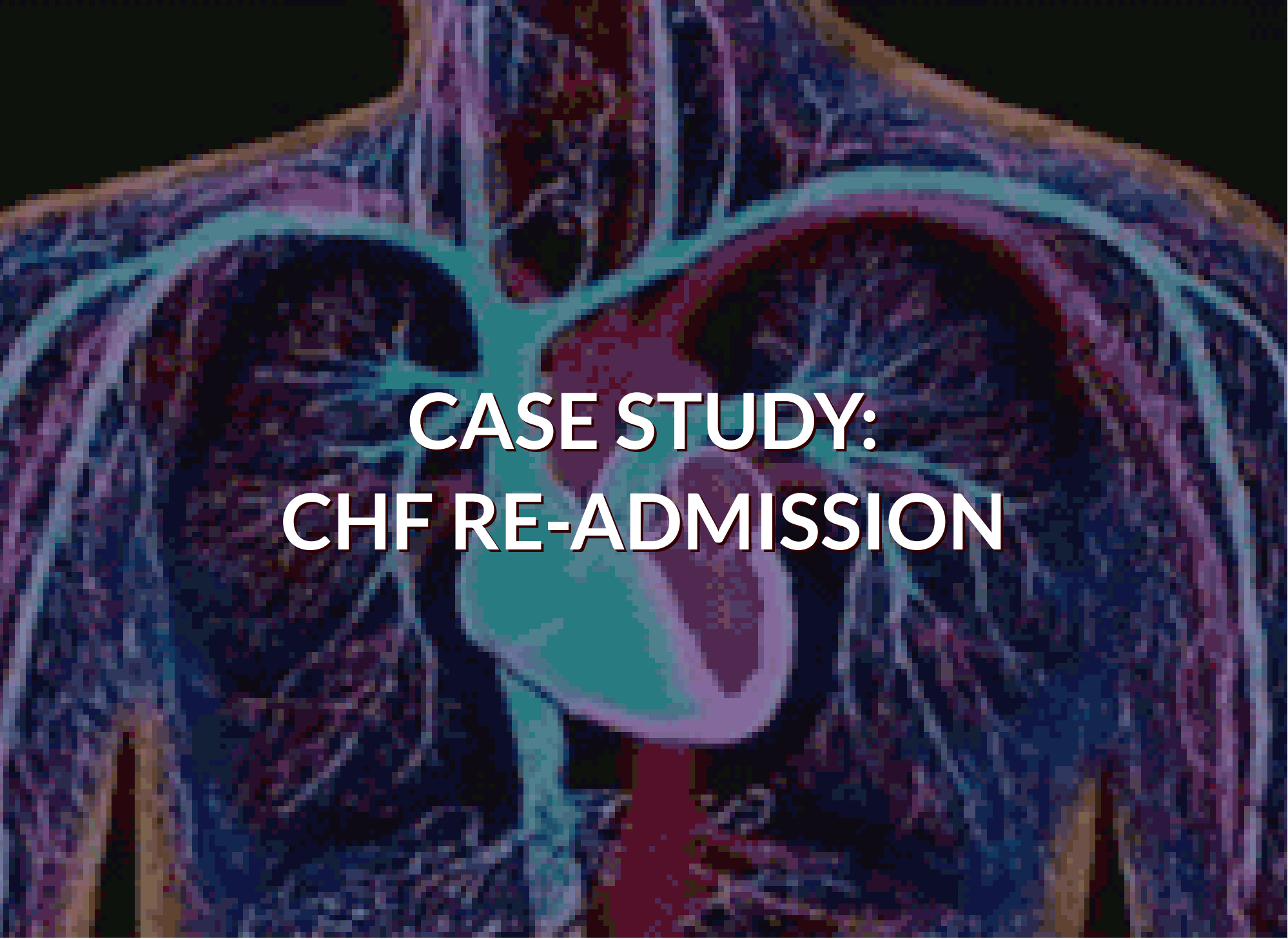
Feature engineering is critical when machine learning tools are being applied to analyze the data.

Working with text analysis

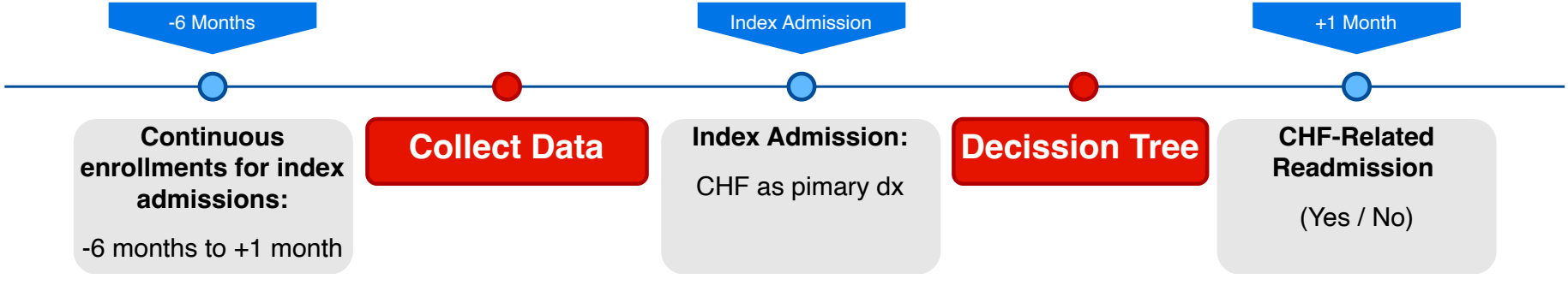


When working with text, **text analysis** steps for coding the data are required to be able to manipulate the data.

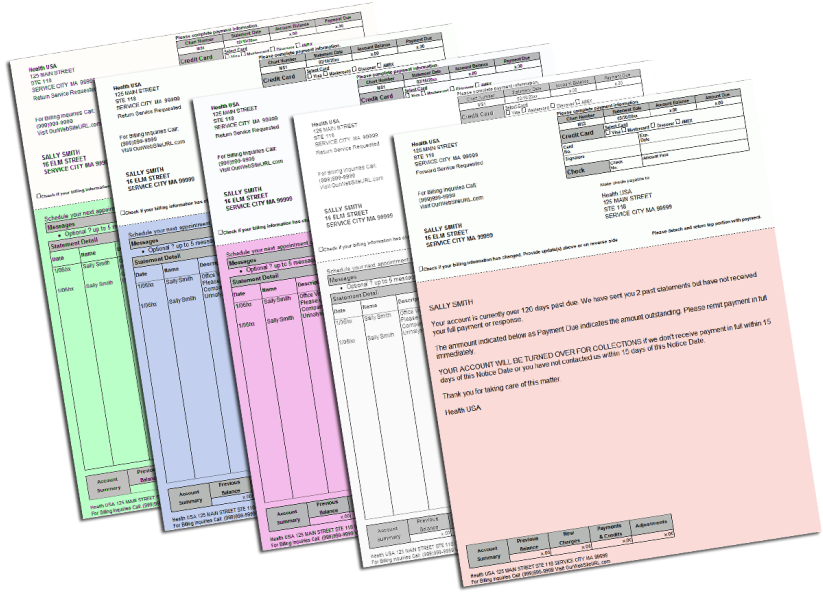
The text analysis is critical to ensure that the proper groupings are set, and that the programming is not overlooking what is hidden within.


CASE STUDY:
CHF RE-ADMISSION

Case Study:
Defining CHF admission/readmission

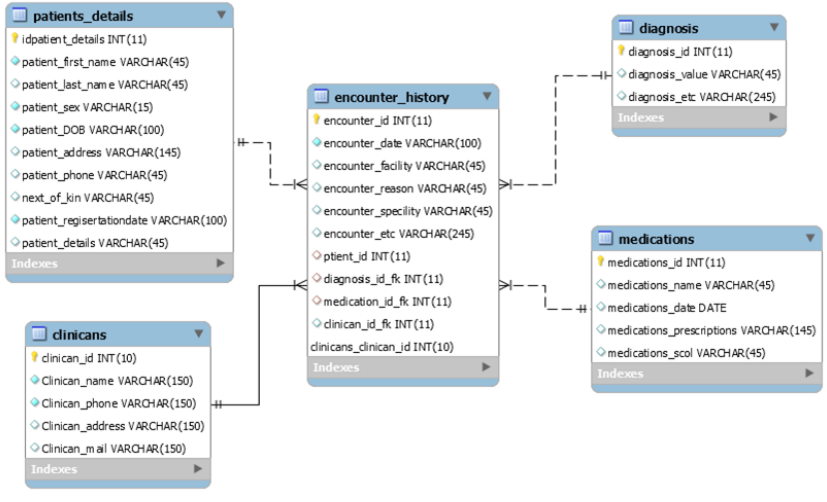


Case Study: Aggregating records

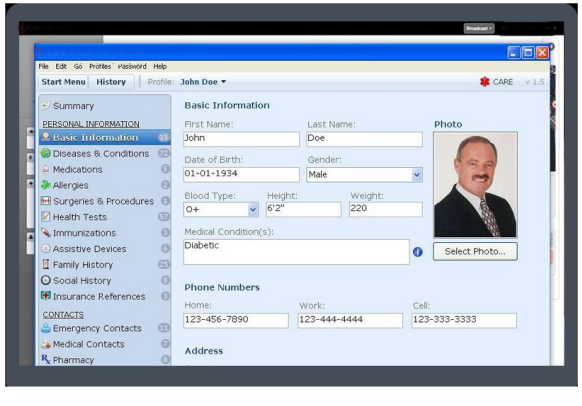


- Transaction records
  - Claims: professional provider, facility, pharmaceutical
  - Inpatien & outpetient records: diagnoses, procedures, prescriptions, etc.
  - Possibly thousands per petient, depending on clinical history

Case Study: Aggregating to patient level

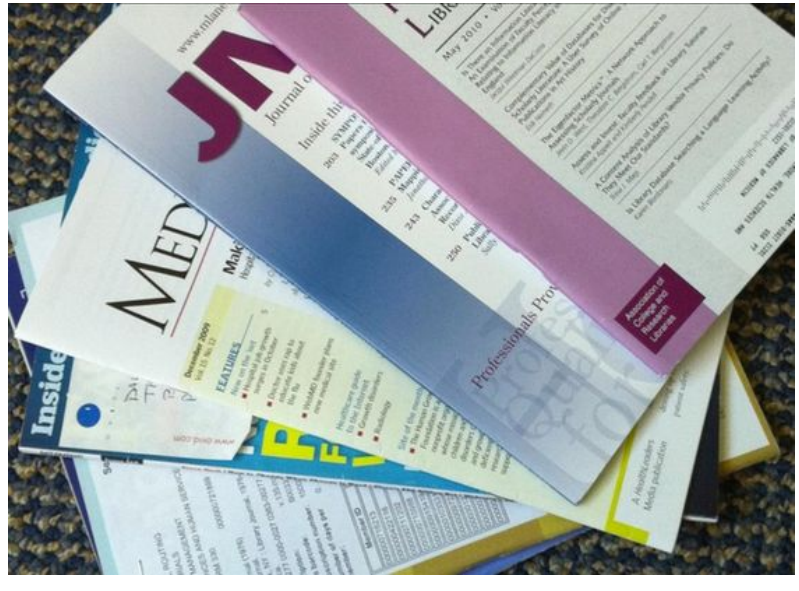


- Roll up to 1 record per patient
- Create new columns representing the transaction
  - Outpatients visits/ Inpatients episodes: frequency, recency, diagnoses, length of stay, procedures, prescriptions
  - Comorbidities with CHF



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

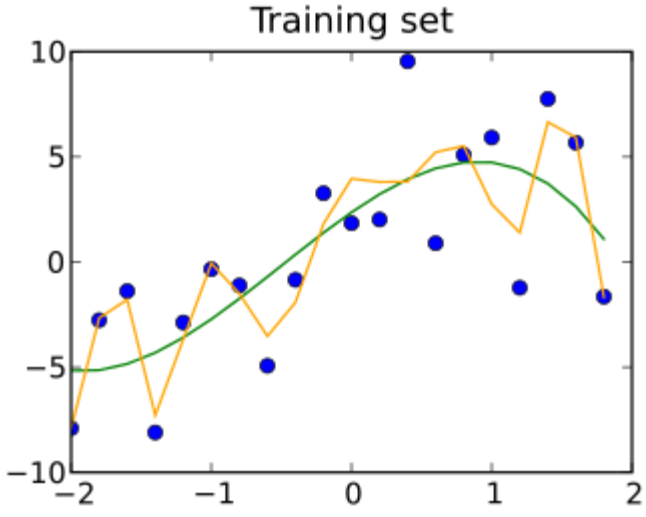
Case Study: Completing the dataset

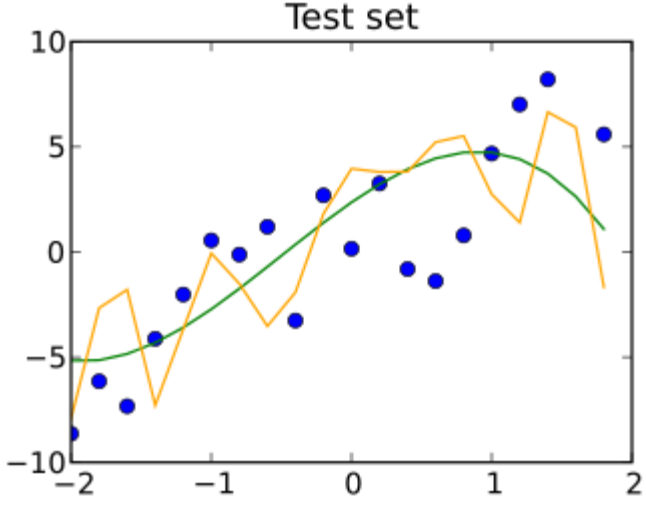
age	sex	diab.	bld.	...	CHFR
...	...	...	...	...	...
55	F	0	144	...	N
78	F	0	121	...	N
66	M	0	125	...	Y
42	F	1	133	...	Y
60	M	1	145	...	N
43	F	0	132	...	Y
33	F	0	162	...	N
...	...	...	...	...	...

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 divied into traning and tesing sets: 70% / 30% split
- Training: 1,640 patients
- Testing: 703 patients

Module 4

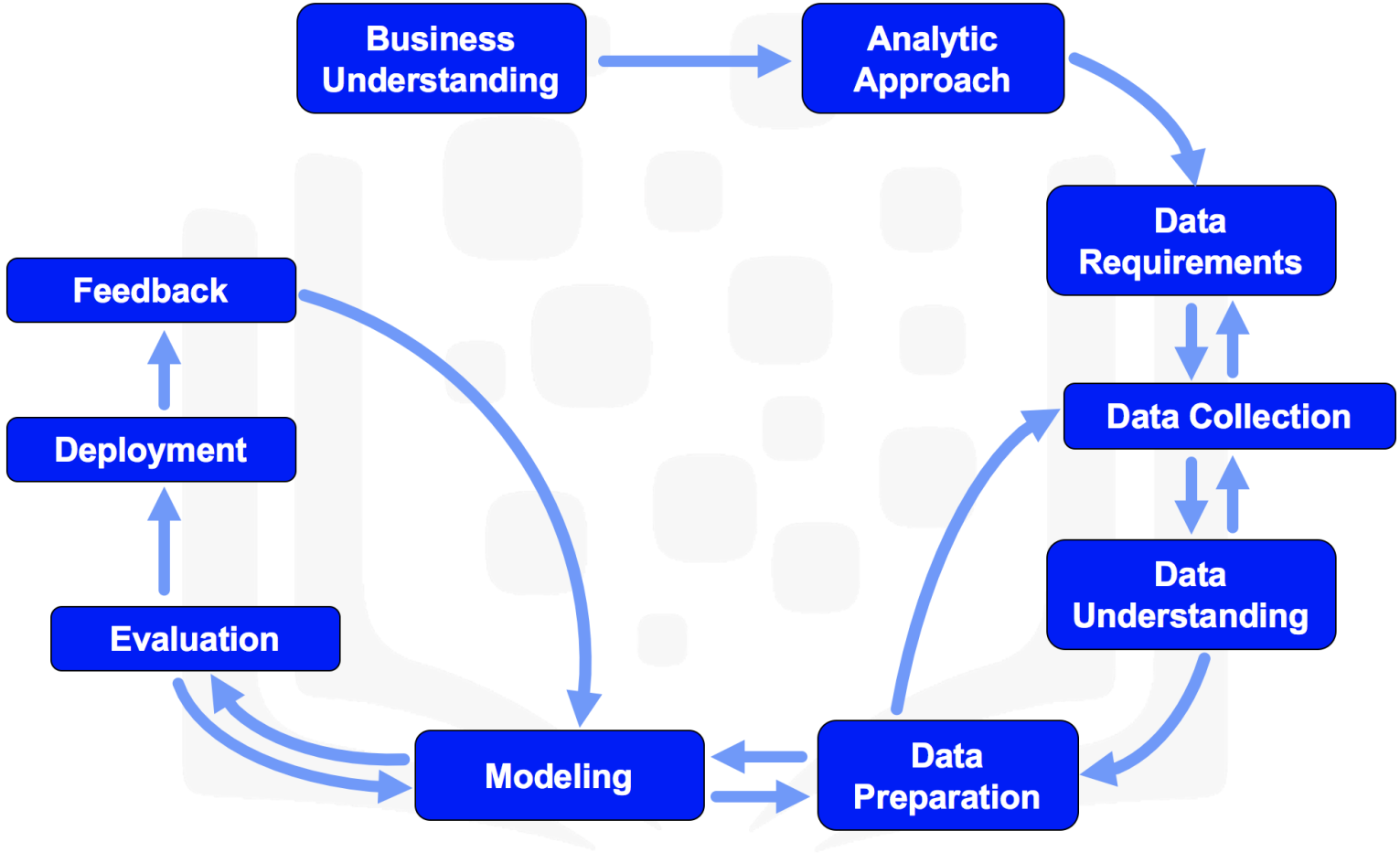
From Modeling to Evaluation

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


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback


MODELING
>>> SAMPLING THE FOOD <<<

From modeling to evaluation

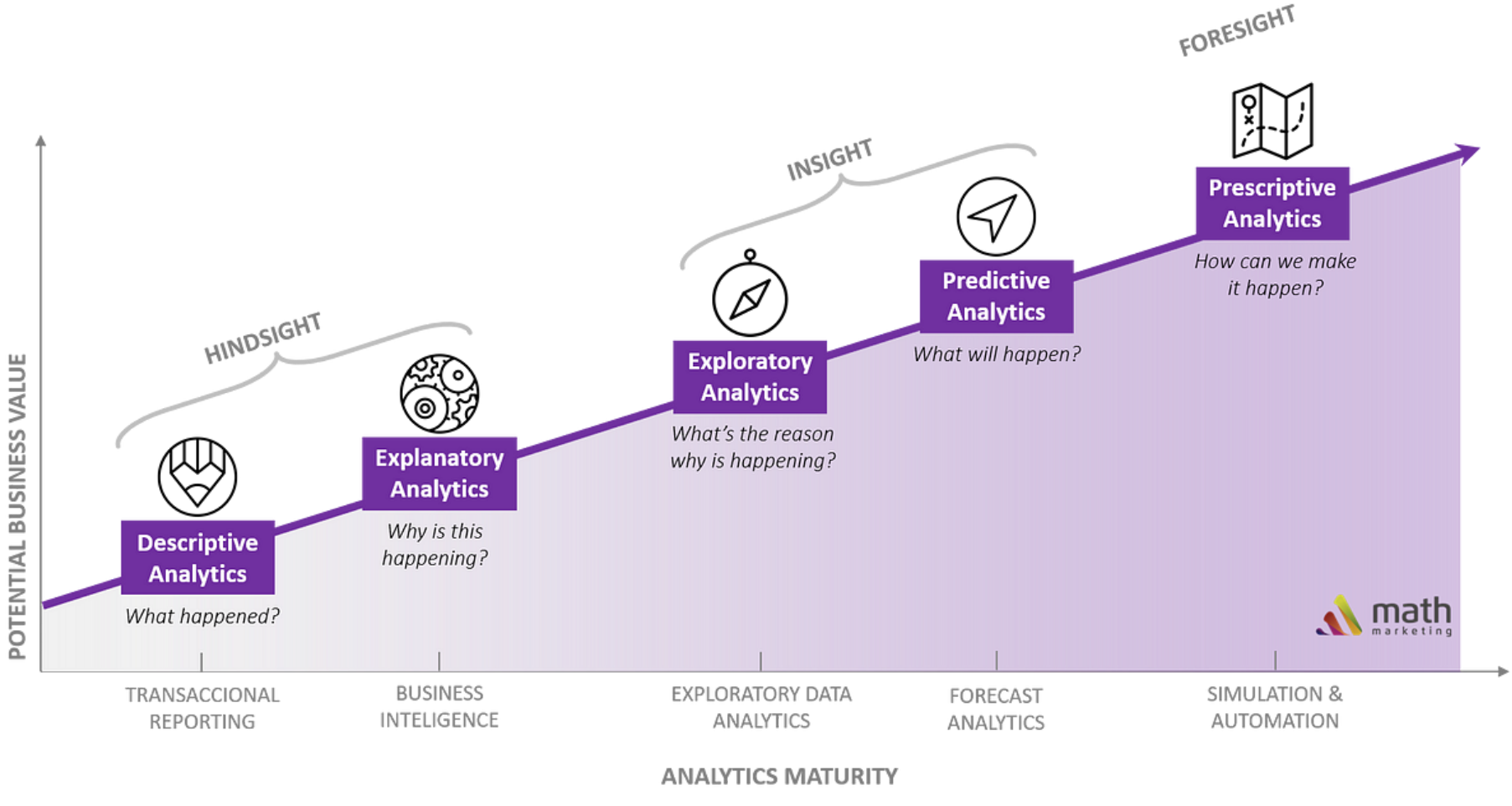
Modeling

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

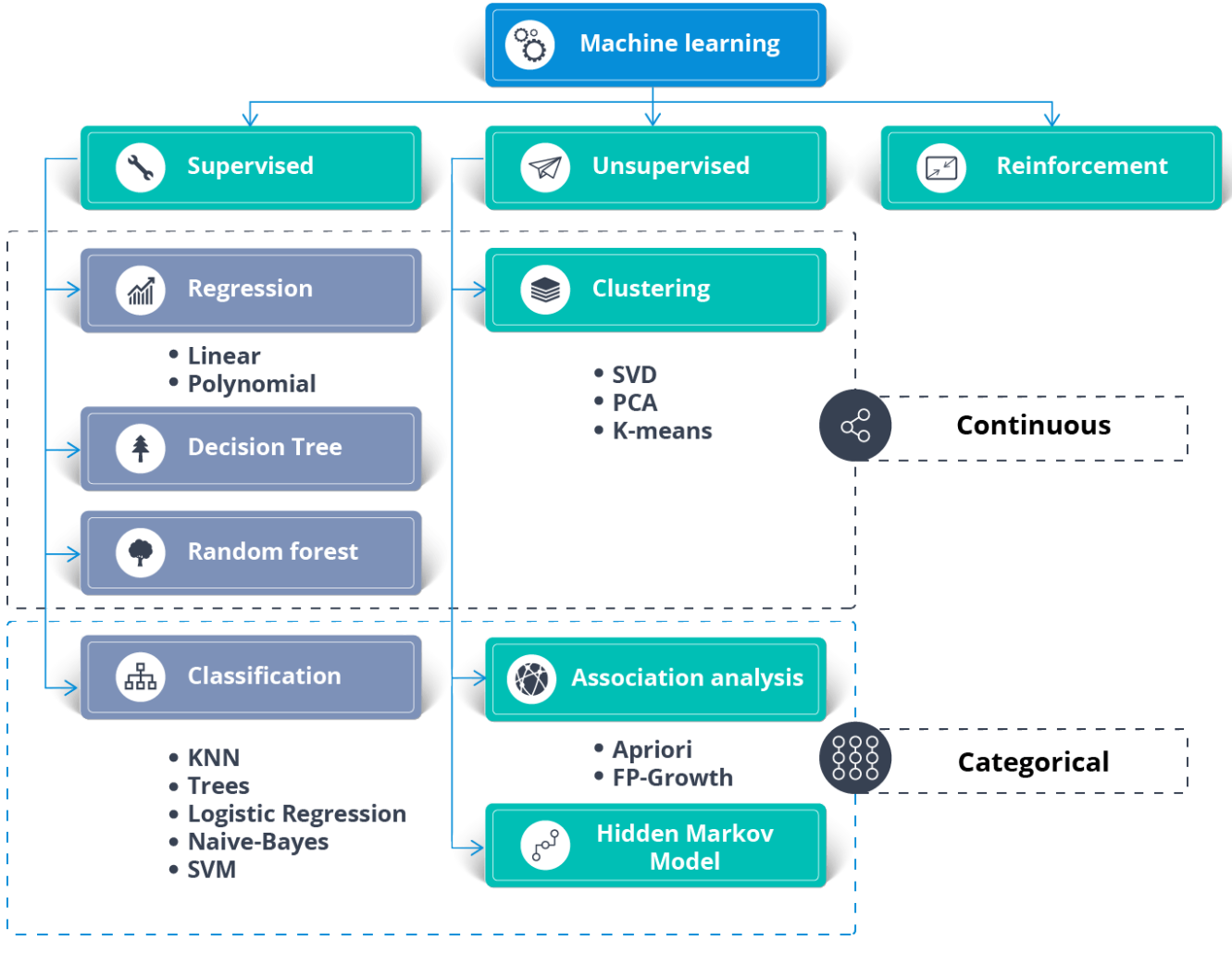
Evaluation

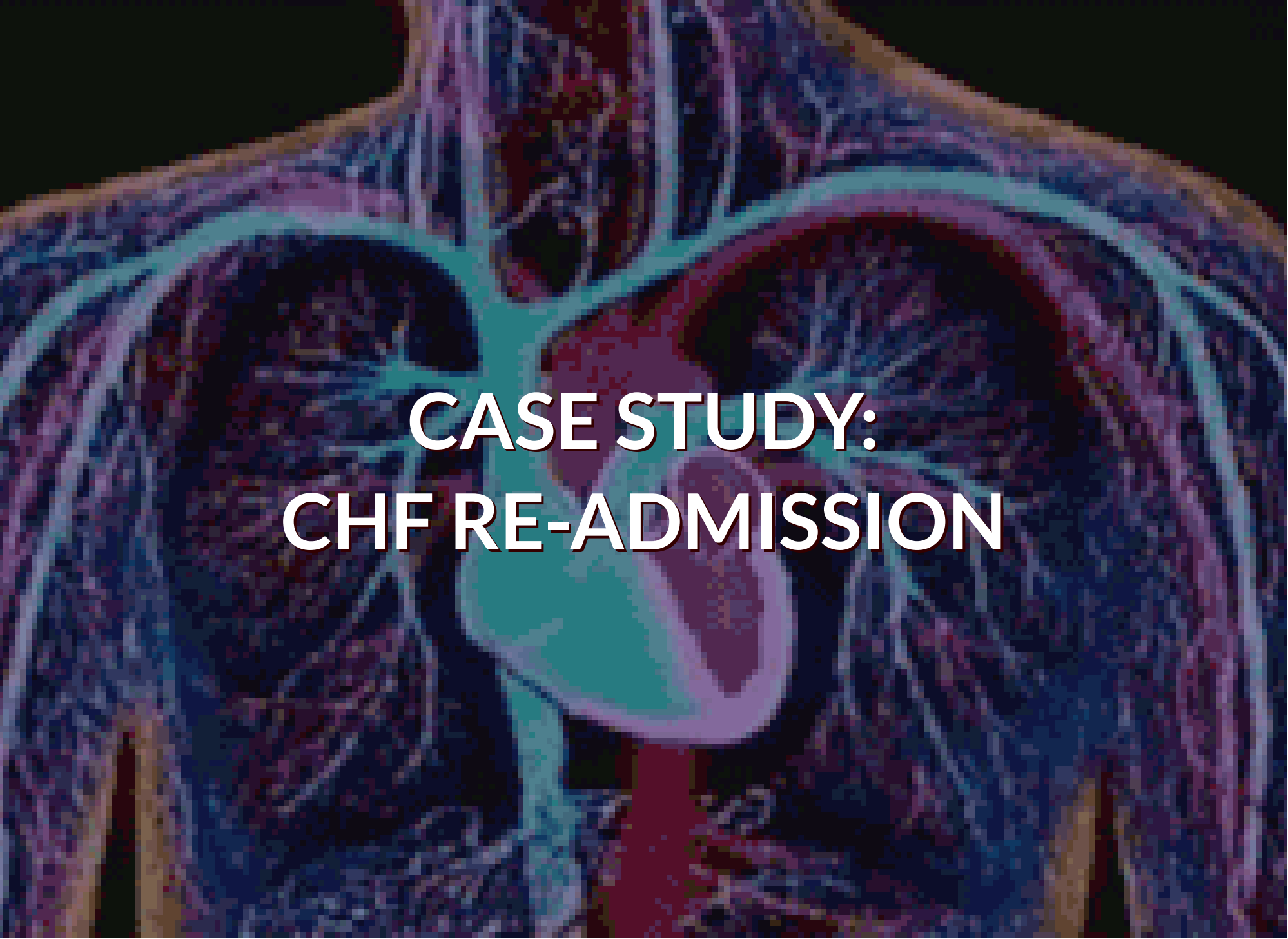
- Does the model used really answer the initial question?
- Does it need to be adjusted?

Analytics Approaches



Machine Learning Models



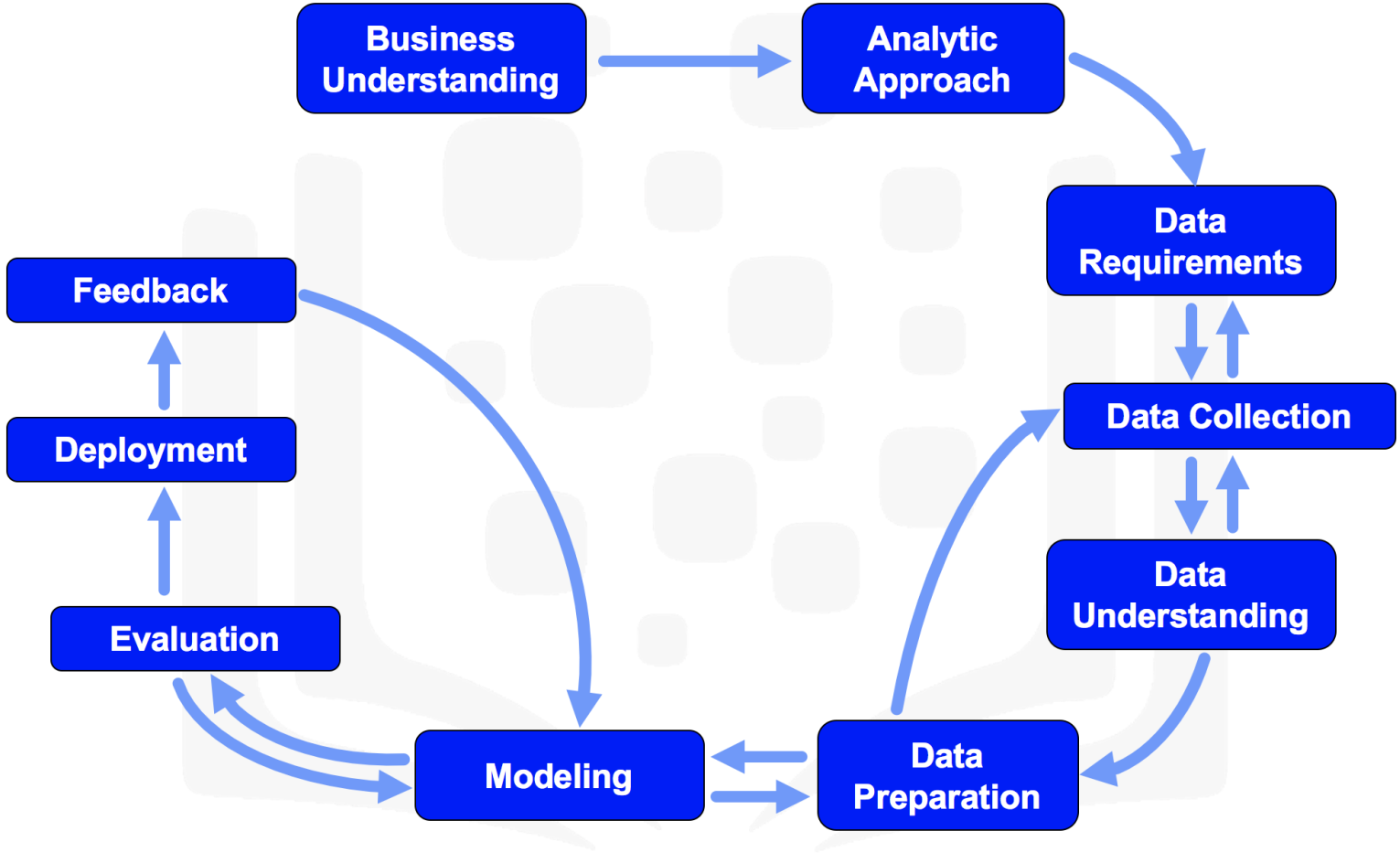

CASE STUDY:
CHF RE-ADMISSION

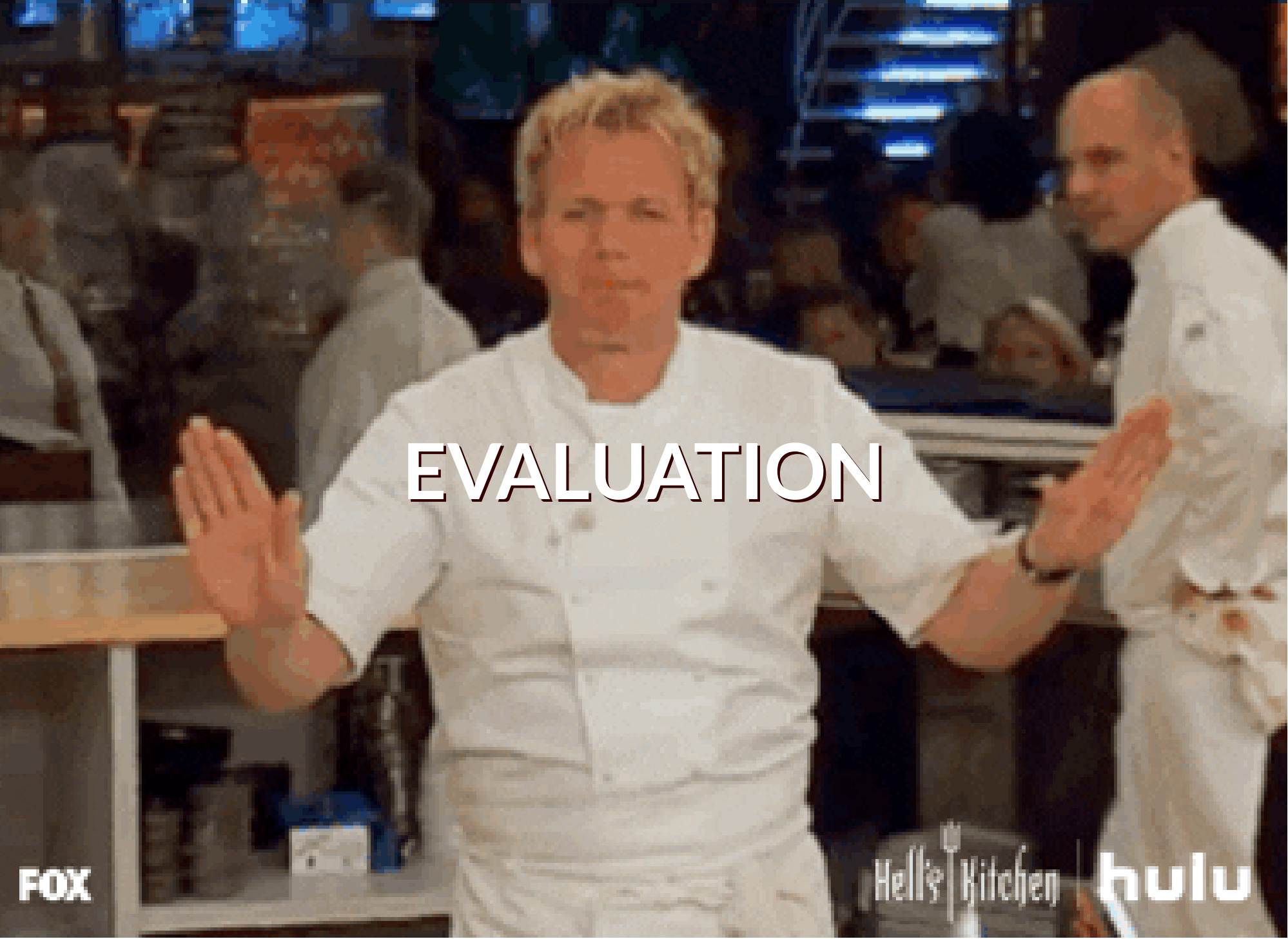
Case Study: Analyzing the models

Decission tree classification models

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

Evaluation


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback


EVALUATION
FOX
Hell's Kitchen
hulu

From modeling to evaluation

Modeling

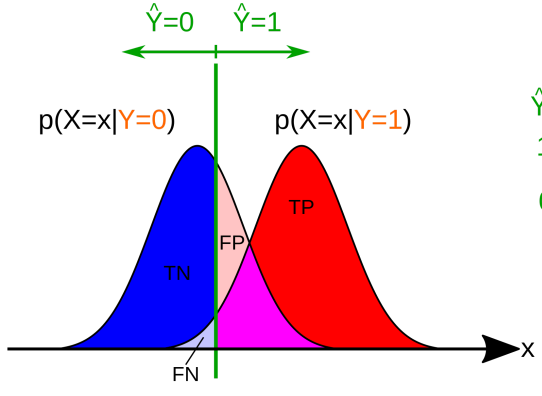
- In what way can the data be visualized to get the answer that is required?

Evaluation

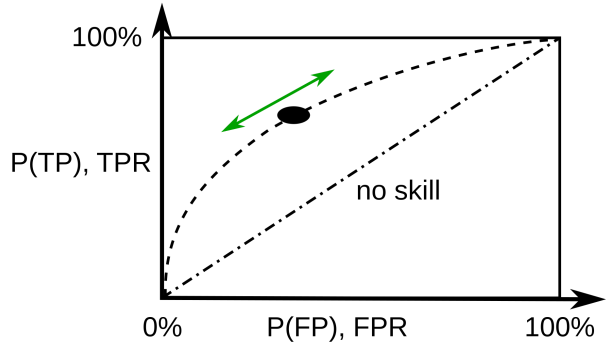
- Does the model used really answer the initial question?
- Does it need to be adjusted?

Model evaluation

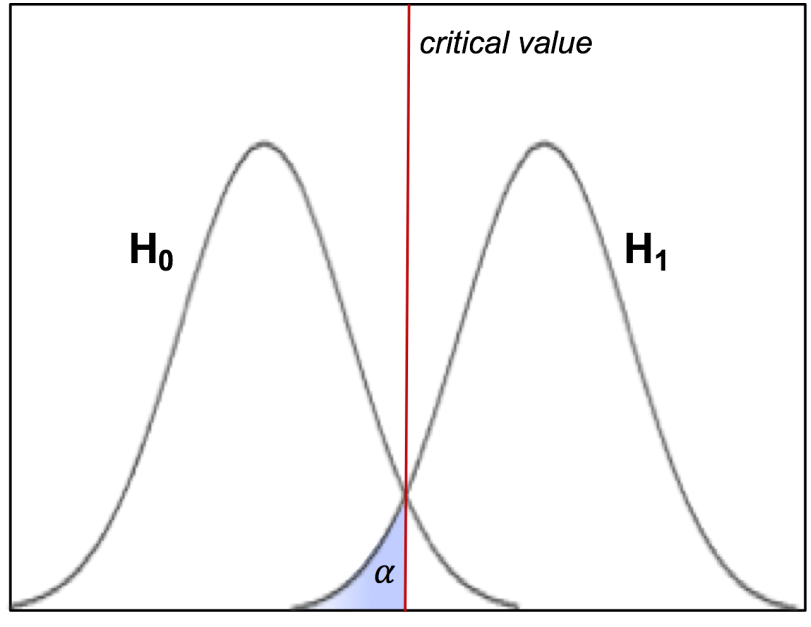
Phase 1: Diagnostic measures

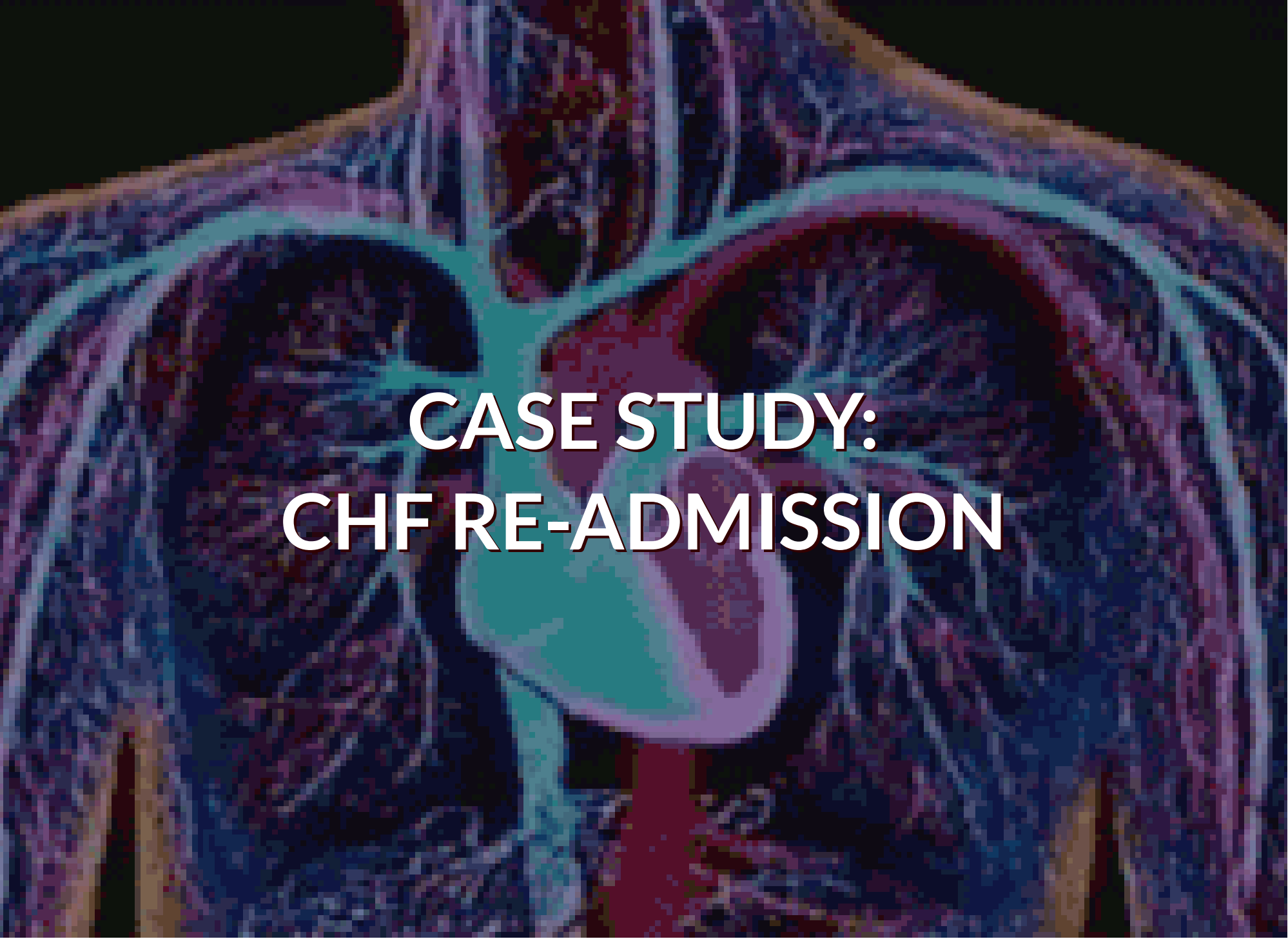


Ŷ \ Y	1	0
1	TP	FP
0	FN	TN



Phase 2: Statistical significance




CASE STUDY:
CHF RE-ADMISSION

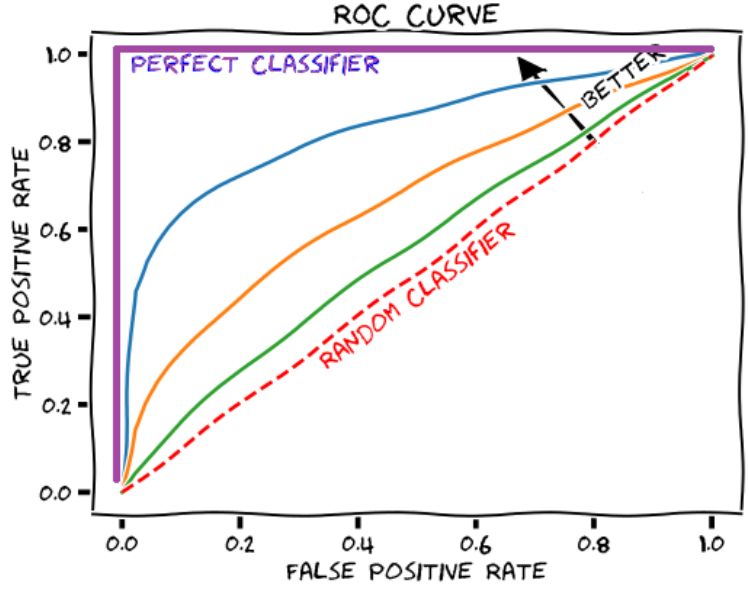
Case Study: Misclassification costs

Model	Relative Cost	TPR	Specificity	FPR
	(Y:N)	(Sensitivity)	(N accuracy)	(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-Poisve Rate (FPR)
- Optimal model at maximum seperation



Module 5

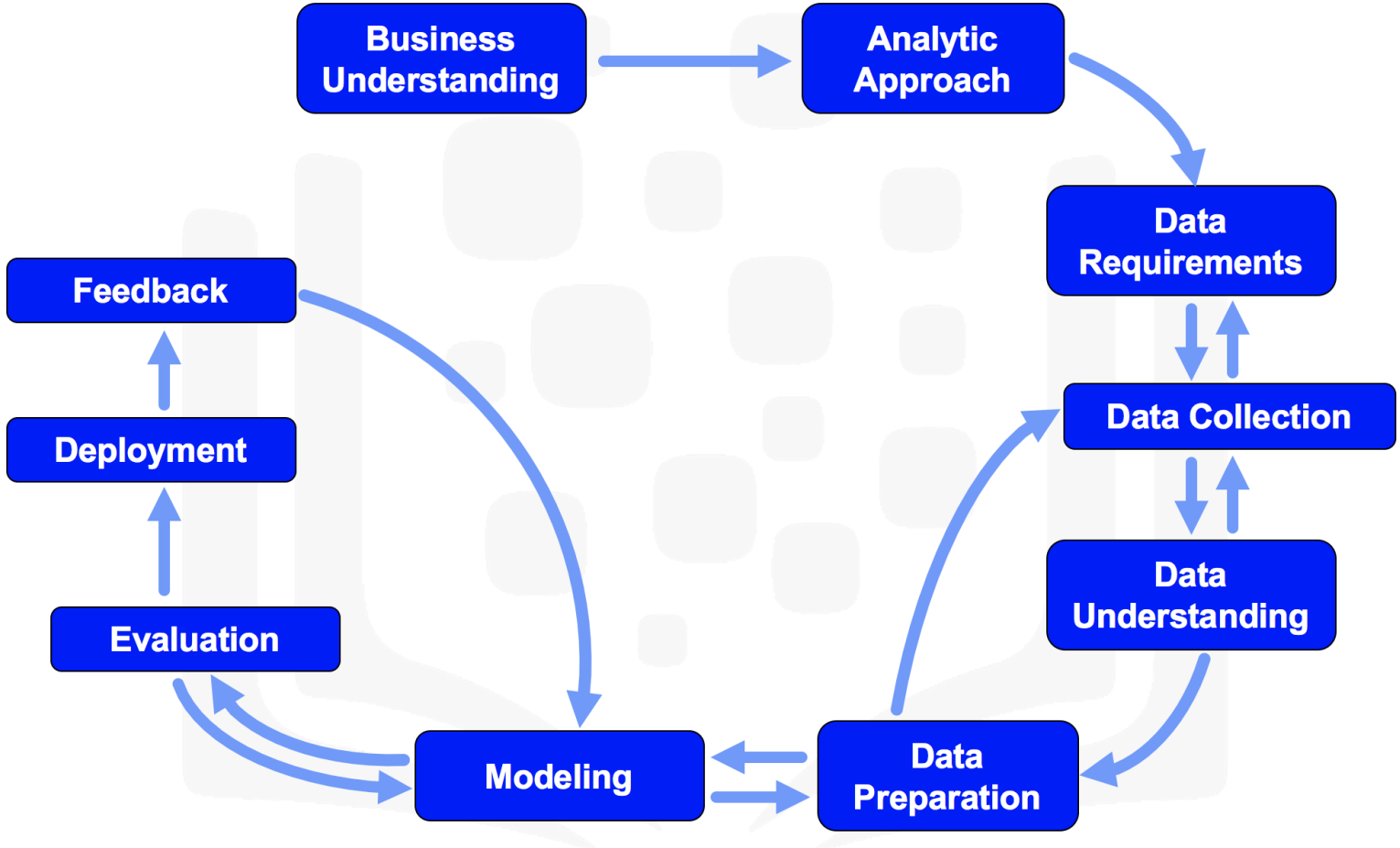
From Deployment to Feedback

Learning objectives

In this lesson you will learn about:

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

Deployment


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback


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 taget high-risk patients
- Additional requirements
  - Training for clinical staffs
  - Tracking/ monitoring processes

Example: Random number simulation of the Dunning and Kruger experiments

Fork this app

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.

Test score vs. perceived ability

Parameters

Number of participants

100

50 150

Correlation

0.50

0.00 1.00

Random seed

42 − +

https://dunning-kruger.streamlit.app

Example: State Movement

Fork this app

State Movement

Choose a state

OVERALL US

Choose a direction

Incoming

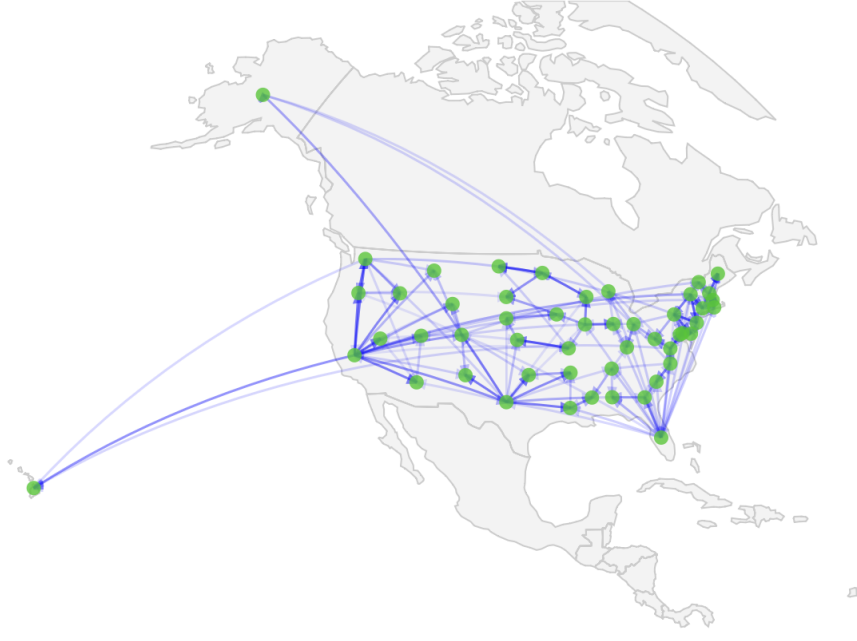
Set top N Migration per state

3

Results Limited to top 5 per State in overall US

Build Migration Map

What is this?



Hope you like the map!

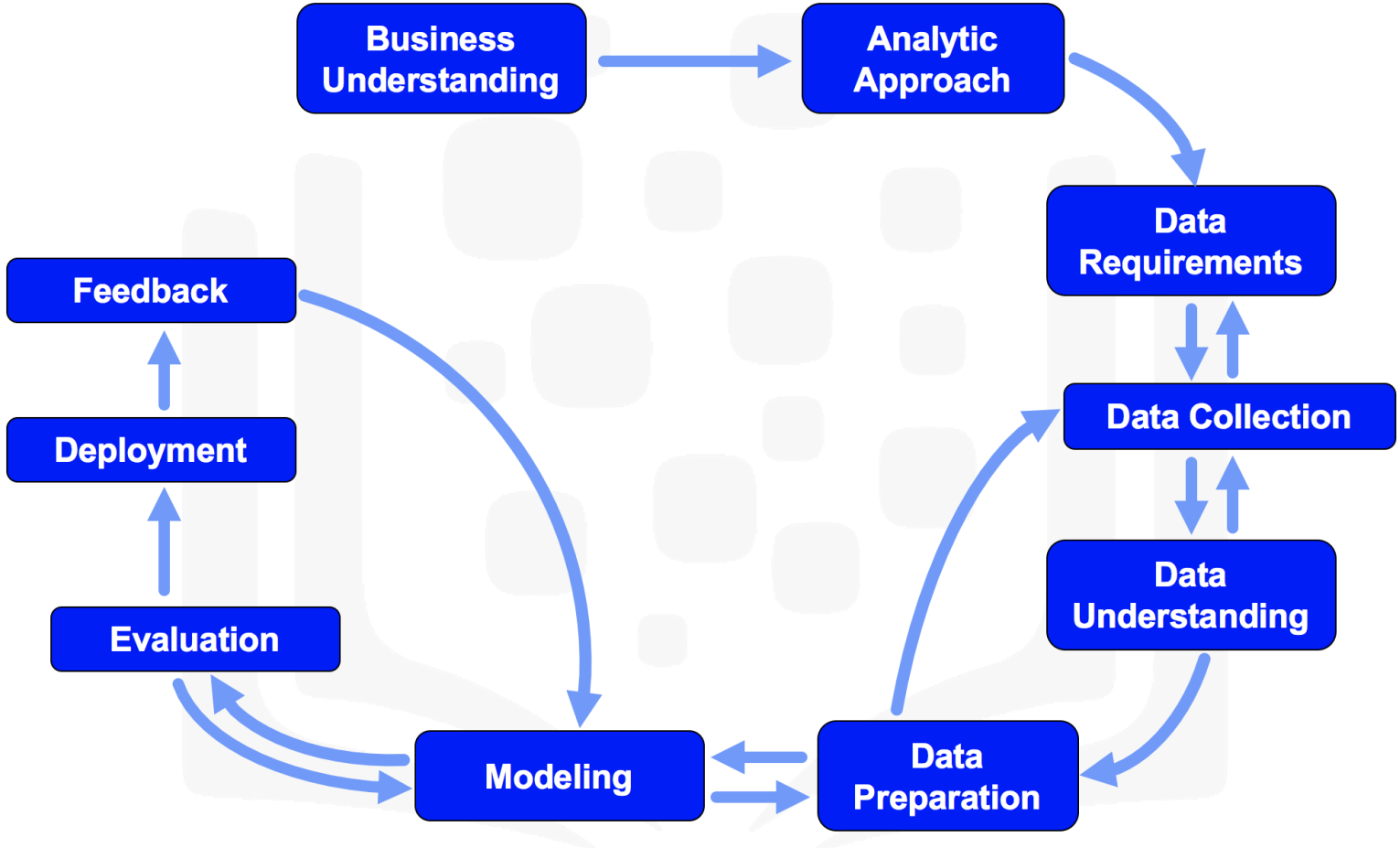
Migration Table

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

Example: Euro 2024 Simulation

https://euro2024pots.streamlit.app

Feedback


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
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

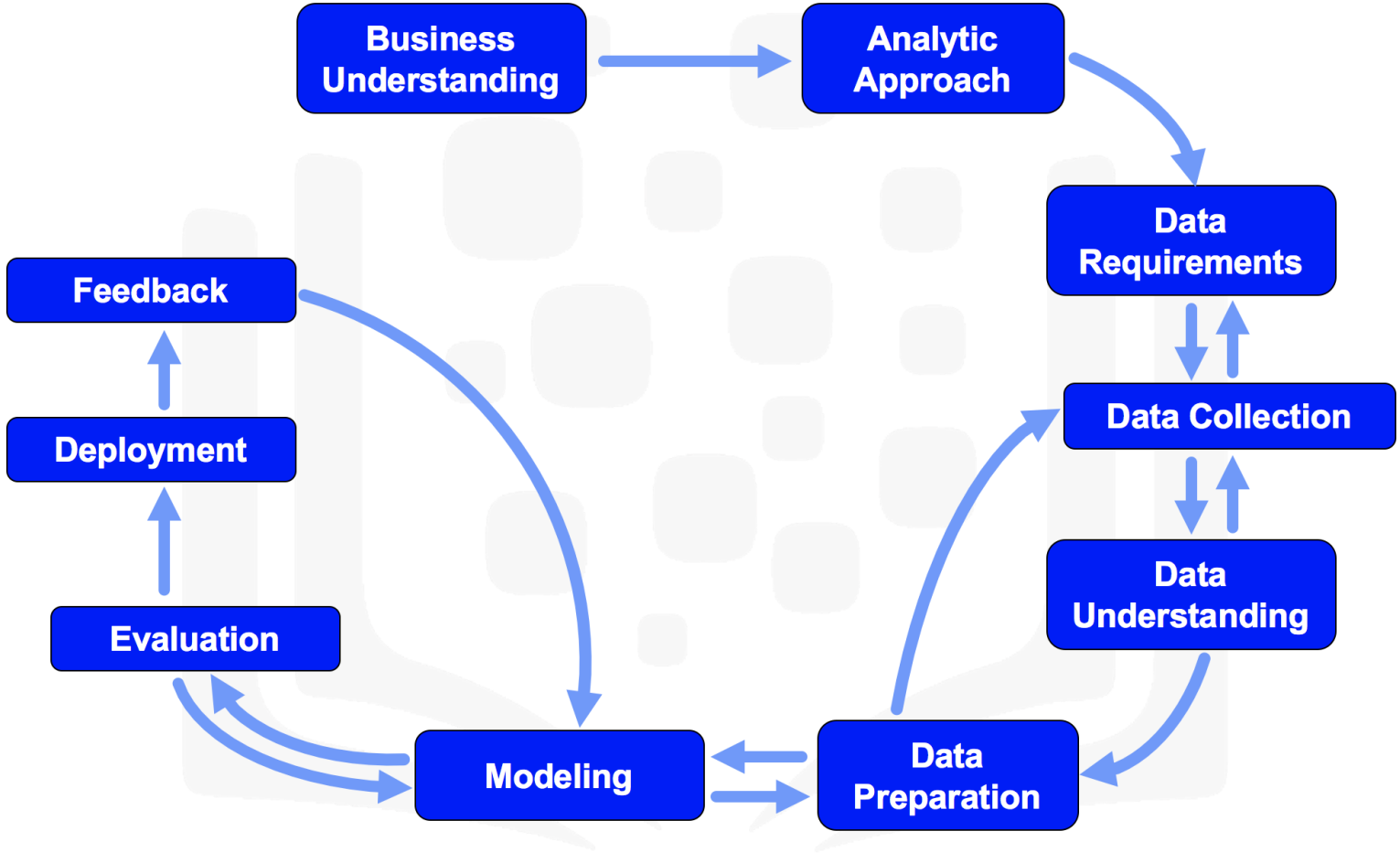
Refine model

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

Case study: Redeployment

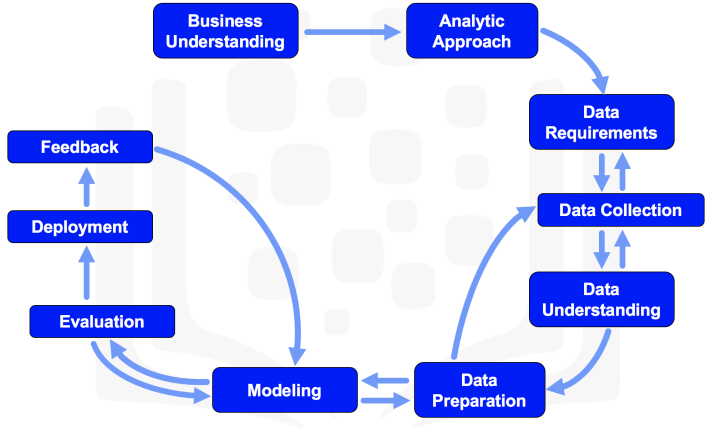
- Review and refine intervention actions
- Redeploy
  - Continue modeling, deployment, feedback and refinement throughout the life of the intervention program

Summary


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

Data Science Methodology

Aims to answer the following questions in this prescribed sequence:

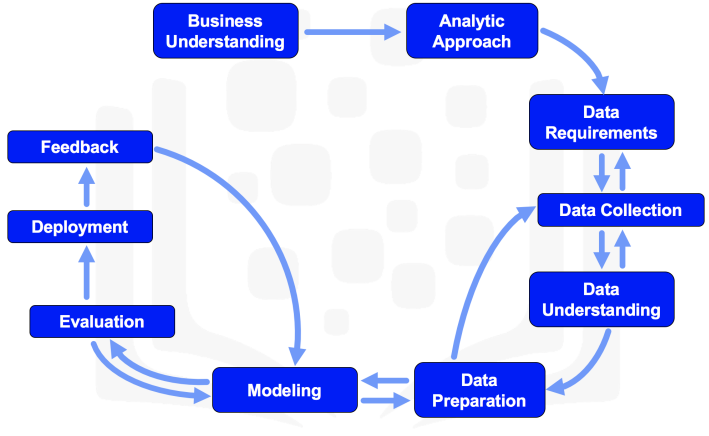


From problem to approach:

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:

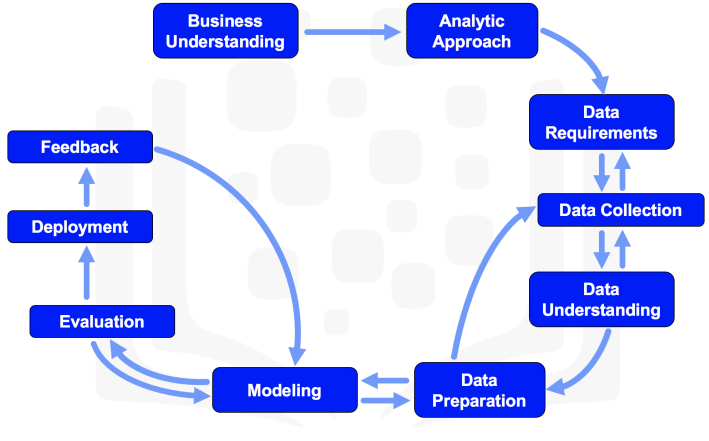


Working with the data:

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?


DONE