COMP47360: Data Analytics for Research Practicum (Conv)

Dr. Georgiana Ifrim
georgiana.ifrim@ucd.ie

Insight Centre for Data Analytics
School of Computer Science
University College Dublin

2016/17

Dynamic Bus Schedule Estimation

When presented with any bus route, an origin stop and a destination stop, a time, a day of the week, current weather, the system should produce and display via the interface an accurate estimate of travel time for the selected journey.

Dynamic Bus Schedule Estimation

Mapping to DA formulation:

- **Input:** Historical data describing a route
  - Descriptive features: route (origin, destination), time, date, weather, traffic/GPS, events, school holidays, …
  - Target feature: Actual Travel Time

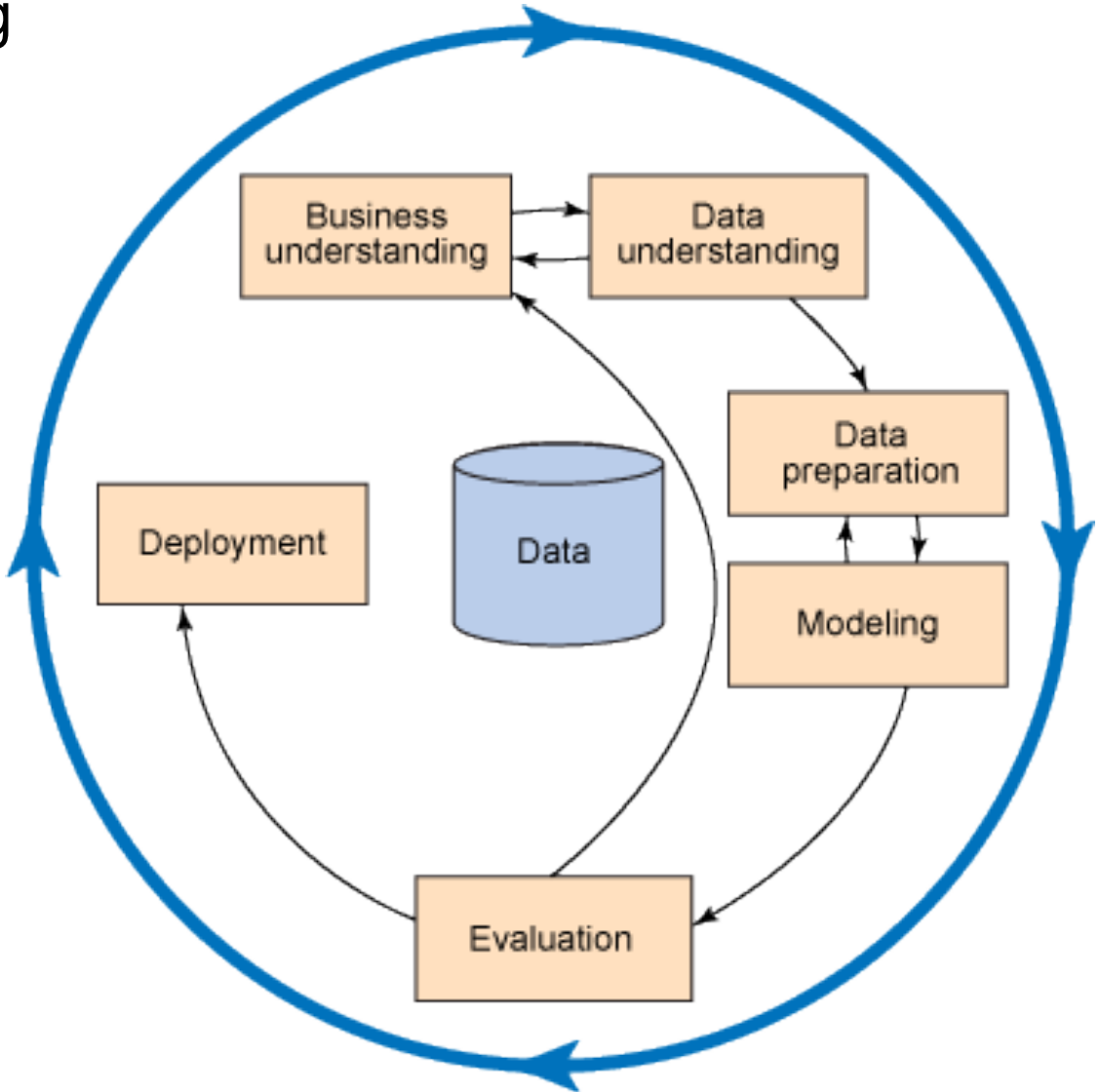
- **Output:** Regression model that can take descriptive features for a route and predict travel time

DA Module Topics

- **Python Environment** (Anaconda, Jupyter Notebook, PyCharm)
- **Getting Data** (Web scrapping, APIs, DBs)
- **Understanding Data** (slicing, visualisation)
- **Preparing Data** (cleaning, transformation)
- **Modeling & Evaluation** (machine learning)

Data Analytics Project Lifecycle: CRISP-DM

CRISP-DM: CRoss-Industry Standard Process for Data Mining



Data Analytics Project Lifecycle:
CRISP-DM

1. **Business Understanding:** What is the business problem?
2. **Data Understanding:** What is the data required to solve the business problem?
3. **Data Preparation:** Where is the data, how should it be collected, transformed and stored?
4. **Modeling:** What analytics algorithms should be used?
5. **Evaluation:** How well do the algorithms work?
6. **Deployment:** How can the analytics results/model be integrated into the work process (specific to the organisation)?

CRISP-DM Summary (FMLPDA Book)

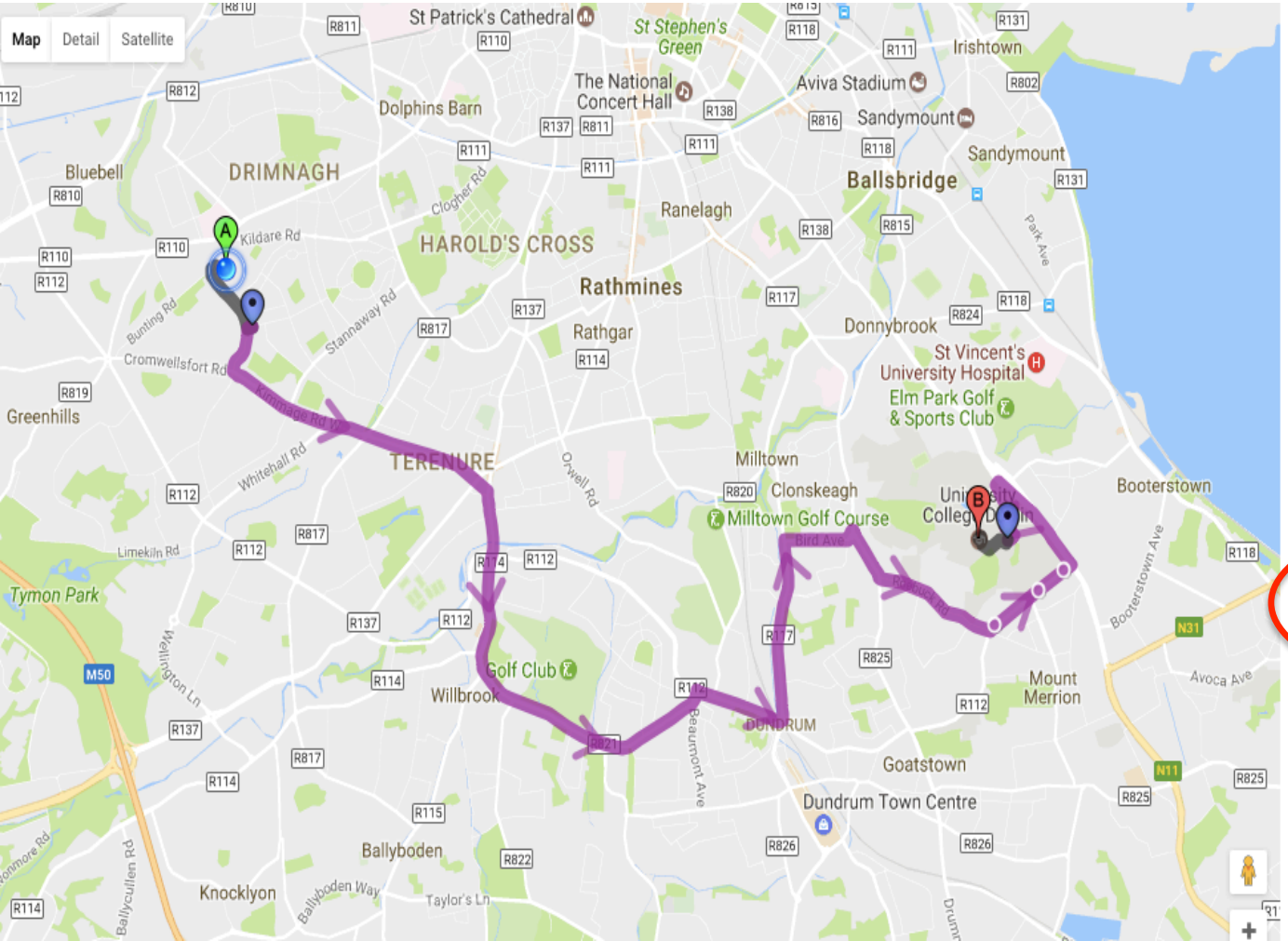
CRISP-DM	Open Questions	Chapter
Business Understanding	*What is the organizational problem being addressed? In what ways could a prediction model address the organizational problem? Do we have situational fluency? What is the capacity of the organization to utilize the output of a prediction model? What data is available?*	Chapter 2
Data Understanding	*What is the prediction subject? What are the domain concepts? What is the target feature? What descriptive features will be used?*	Chapter 2

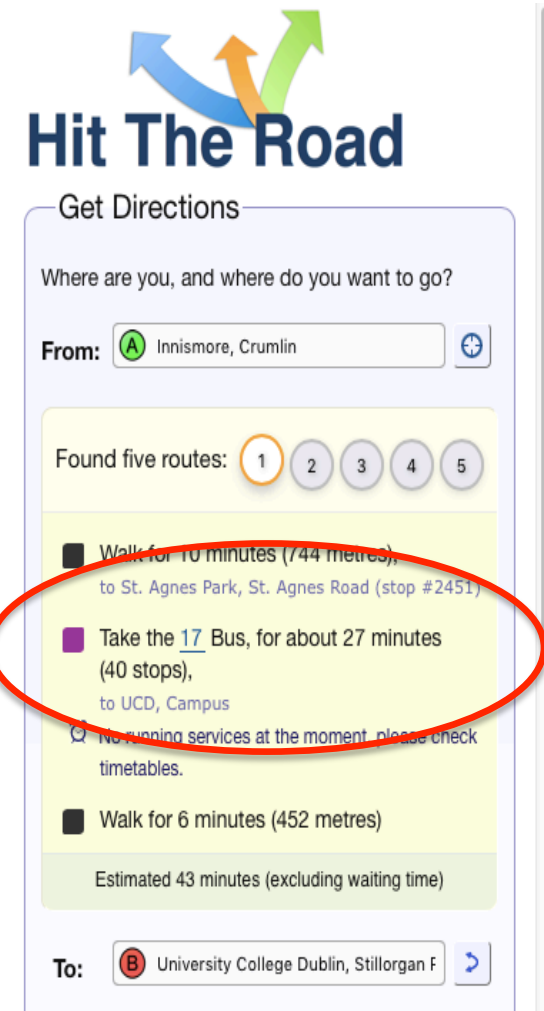
Business Understanding

- **What is the business need?**
  - What are the existing solutions for this problem?
  - Customers unhappy with current estimates?
  - Introduce new routes?
- **How will the prediction model be used?**
  - Sell service (i.e., better estimate) to third-party?
  - Dynamic estimate requires low prediction time (Can we predict in real-time? Under 5 mins?)
- **What is an acceptable error margin?**
  - Is an error of +- 5 mins acceptable?
  - Is a 10% improvement on current estimate a good goal?

Business Understanding

How will the prediction model be used?





Business Understanding

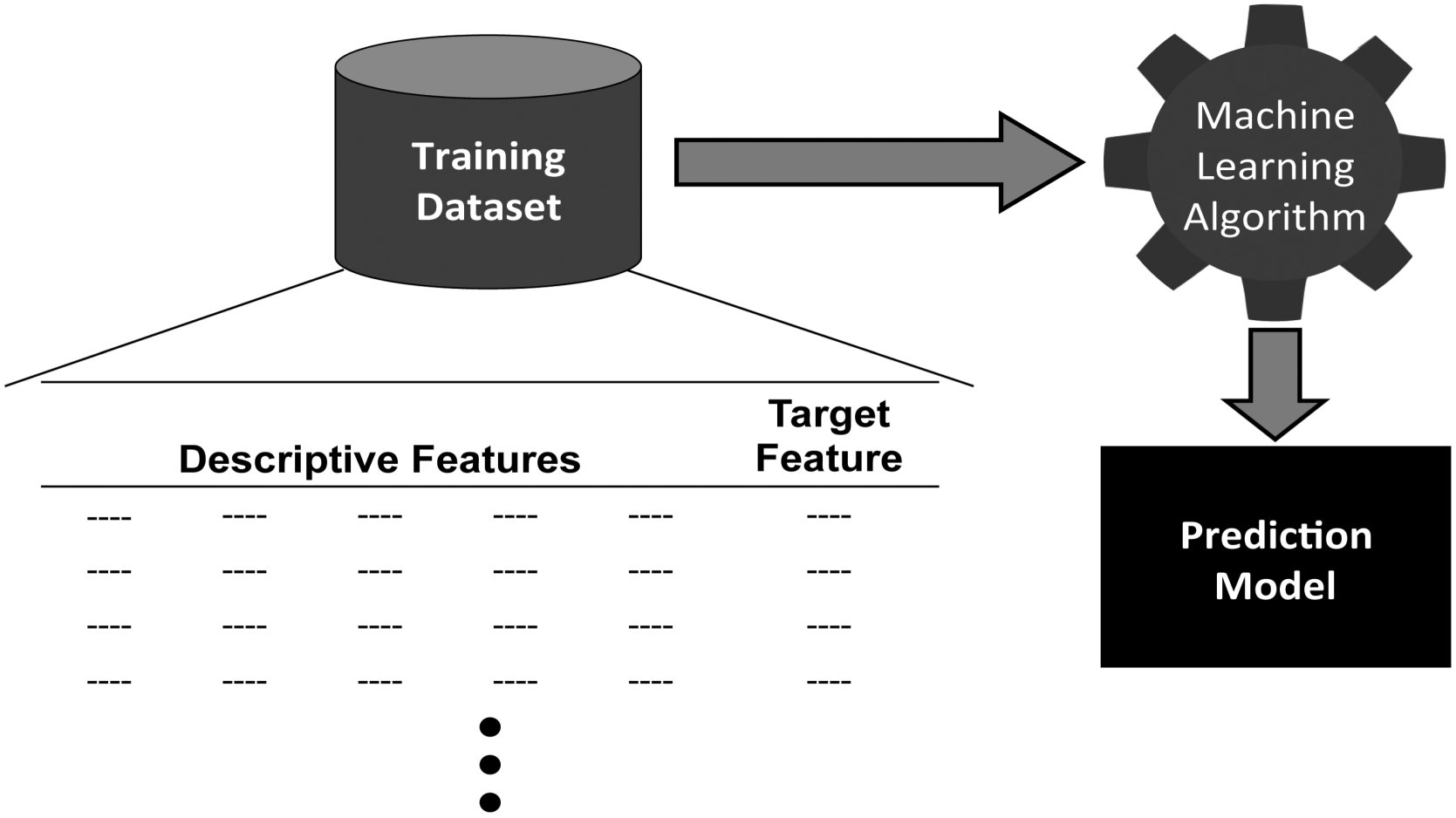
- **What data is available?**
  - Dublin Bus GPS
  - Weather
- **What data is required?**
  - Calendar (sporting events, concerts, school holidays)
  - Other??? (how many buses service a route)

CRISP-DM Summary (FMLPDA Book)

CRISP-DM	Open Questions	Chapter
Business Understanding	*What is the organizational problem being addressed? In what ways could a prediction model address the organizational problem? Do we have situational fluency? What is the capacity of the organization to utilize the output of a prediction model? What data is available?*	Chapter 2
Data Understanding	*What is the prediction subject? What are the domain concepts? What is the target feature? What descriptive features will be used?*	Chapter 2

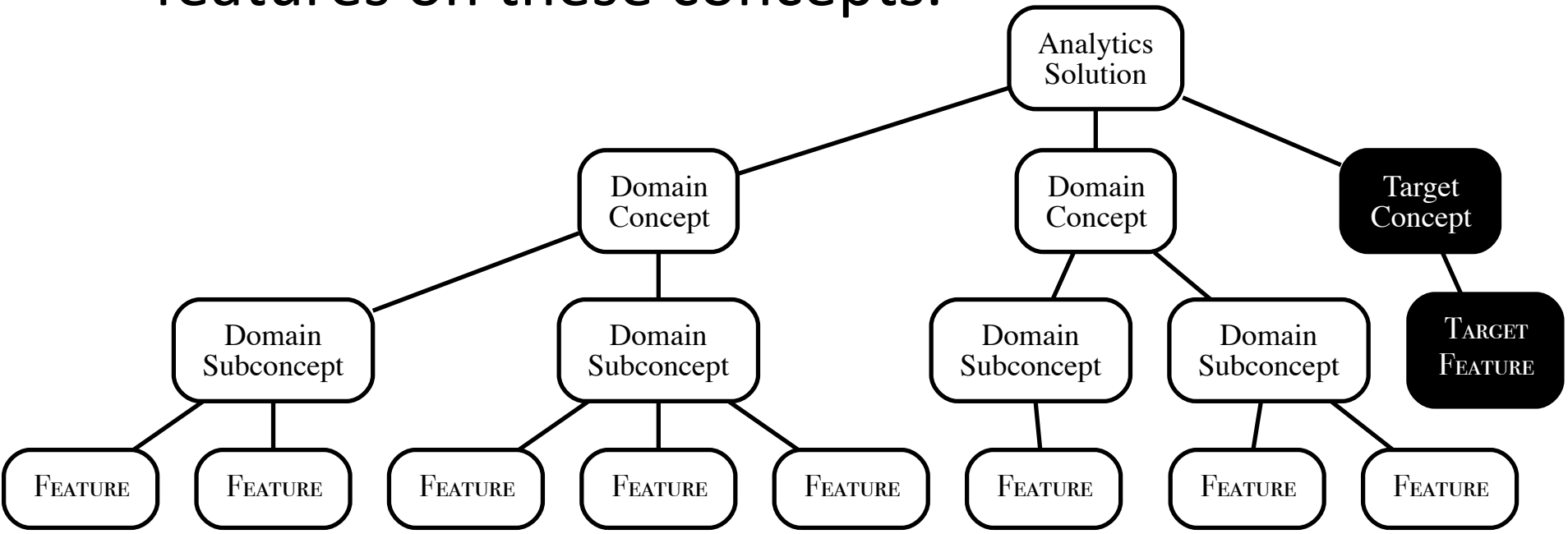
Analytics Base Table

The basic structure in which we capture historical datasets is the **analytics base table** (looks like a spreadsheet or a CSV file)



Analytics Base Table

- **Defining features can be difficult**
- A good way to define features is to identify the key **domain concepts** and then to base the features on these concepts.



Data Understanding

- **What are the domain concepts?**
  - Bus route (features: each pair origin_destination; separate features for origin, destination?)
  - Weather (features: rainfall, temperature, wind)
  - Time (features: hour of day, rush_hours)
  - Date (features: month, season)

- **What is the target feature?**
  - Do we have access to actual travel time from GPS traces?

Data Understanding

Data Quality Report (tables, plots)

- Getting To Know The Data
- Identifying Data Quality Issues

Data Quality Plan

- Handling Data Quality Issues (missing values, irregular cardinality, outliers, incorrect data)

Data Quality Report

First round:

- **Tables:** descriptive statistics for each feature (e.g., min, max, average, median value)
  - Decide which features make more sense as categorical vs numeric

Second round:

- Look at tables with descriptive stats again

Data Quality Report

Visualizations/plots:

- Plot each feature
  - Continuous feature: histogram, boxplot
  - Categorical feature: barplot
- Plot pairs of features
  - Continuous-continuous: scatter plot
  - Continuous-categorical: boxplot (or histogram) of continuous feature, for each level of the categorical feature
  - Categorical-categorical: barplot of one categorical feature for each level of the other categorical feature (or stacked barplot)

Data Understanding: Summary

1. Describe the samples/rows and features/columns within the ABT: how many rows/columns, central tendencies, variations, and distributions

2. Identify data quality issues within the ABT: duplicate samples or features, missing values, irregular cardinality (constant columns), outliers (care with handling outliers!!!)

3. Correct data quality issues due to invalid data

4. Record data quality issues due to valid data in a data quality plan along with potential handling strategies

5. Does enough good quality data exists to continue with a project?

CRISP-DM Summary (FMLPDA Book)

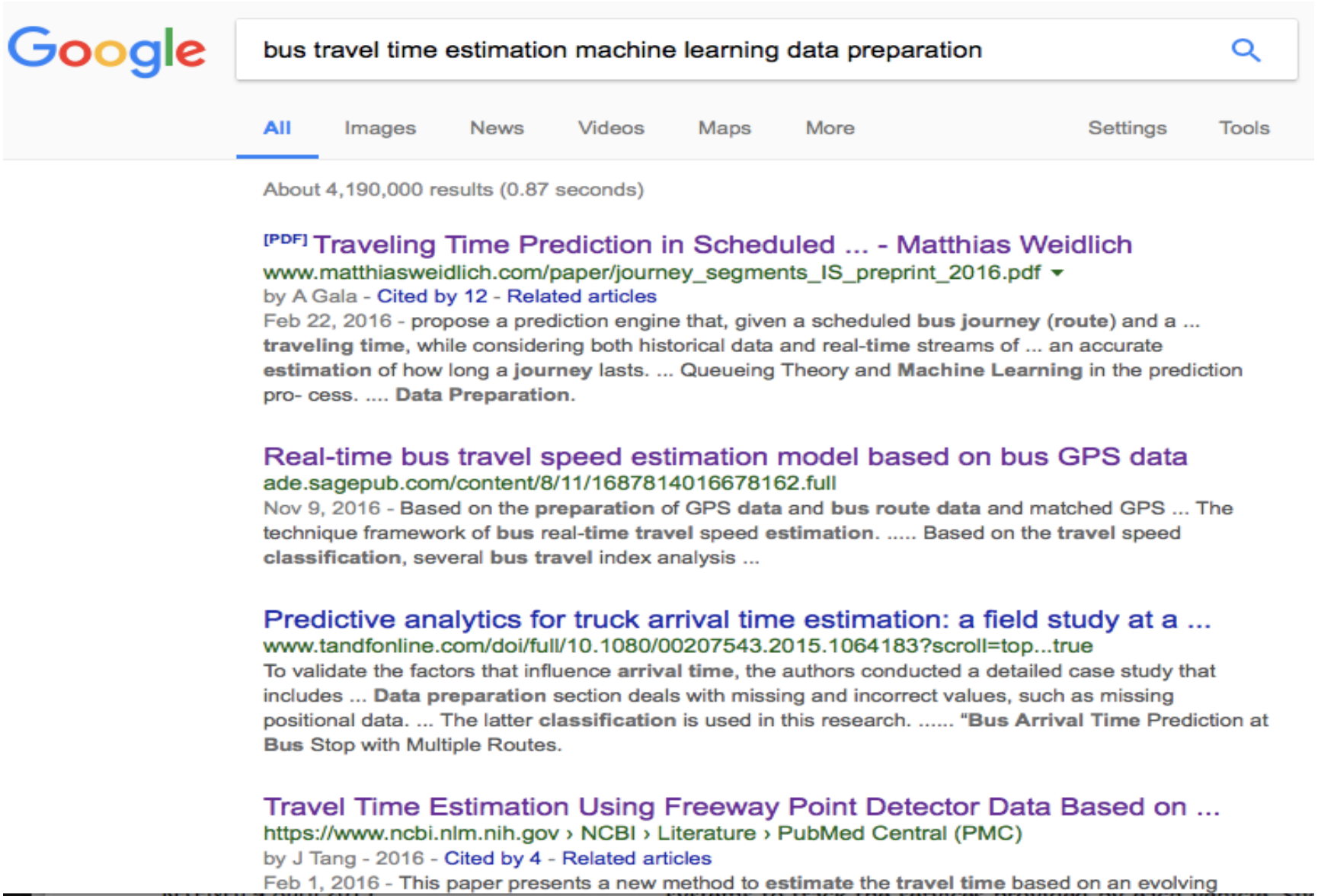
CRISP-DM	Open Questions	Chapter
Data Preparation	*Are there data quality issues? How will we handle missing values? How will we normalize our features? What features will we include?*	Chapter 3
Modelling	*What types of models will we use? How will we set the parameters of the machine learning algorithms? Have underfitting or overfitting occurred?*	Chapters 4, 5, 6 and 7

Data Preparation

Changing the original data to make it more compatible with some machine learning algorithms and improve predictive accuracy

- Converting categorical features
- Normalization
- Binning
- Sampling

Task Specific Data Preparation



Data Preparation Tips

- Start with simple and quick data prep first (e.g., duplicate/constant columns, missing values)
- Iterate data prep and modeling
- What is a good baseline model that you aim to improve upon? (e.g., average time on this route at this hour, last week?)

CRISP-DM Summary (FMLPDA Book)

CRISP-DM	Open Questions	Chapter
Data Preparation	*Are there data quality issues? How will we handle missing values? How will we normalize our features? What features will we include?*	Chapter 3
Modelling	*What types of models will we use? How will we set the parameters of the machine learning algorithms? Have underfitting or overfitting occurred?*	Chapters 4, 5, 6 and 7

Modeling

Supervised Learning

- Learn a model of the relationship between a set of **descriptive features** and a **target feature**
- **Key steps for modeling:**
  - **Feature engineering** (problem/data understanding)
  - **Feature selection** (feature importance, grid search)

Modeling

How to build prediction models

- **Regression:** predicting a numeric target feature (**linear regression, random forests**)
- **Classification:** predicting a categorical target feature (**logistic regression, random forests**)

Modeling

- **Linear regression:**
  - Simple model to begin with
  - Need to convert categorical features to numeric
  - Strongly affected by feature correlation
  - Linearity assumption (may need non-linear features)

Modeling

- **Random Forests** (ensemble via bagging):
  - Powerful non-linear model
  - No need to convert categorical features (but don't be lazy in data prep, data has to make sense)
  - RF feature importance also affected by feature correlation
  - Good for building a quick strong baseline

Modeling

- **New contenders: Gradient Boosting** (ensemble via boosting):
  - Powerful non-linear model
  - No need to convert categorical features
  - Xgboost tool wins a lot of Kaggle competitions
  - Good for building a quick strong baseline

CRISP-DM Summary (FMLPDA Book)

CRISP-DM	Open Questions	Chapter
Evaluation	*What evaluation process will we follow? What performance measures will we use? Is the model fit for purpose?*	Chapter 8
Deployment	*How will we continue to evaluate the model after deployment? How will the model be integrated into the organization?*	Section 8.4.6 and Chapters 9 and 10

Model Evaluation

Experiment Design

- Underfitting/Overfitting
- Out-of-sample Testing
- Cross-Validation
- Offline/Online

Evaluation Metrics

- Regression: What are appropriate error metrics?

Model Evaluation

Very important for the design of an evaluation experiment:

- Make sure that the data used to evaluate the model is not the same as the data used to train the model!

Model Evaluation

The purpose of evaluation is threefold:

1. To determine which model is the most suitable for a task (problem modeling)
2. To estimate how the model will perform when deployed (generalization)
3. To convince users that the model will meet their needs (accuracy, efficiency, retraining)

Evaluation

How to evaluate prediction models

To avoid overfitting:

- **Training/test split:** fit model on the training set, predict on the test (use the test error)
- **Cross-validation:** repeated train/test splits and averaging the test error
- Need care with feature selection (need a test set that is never seen for model selection)
- Need care with time split: train on past data, predict on future data

Regression: Evaluation Metrics

RMSE (Root Mean Squared Error)

n = number of examples

y_i = true value of the target feature

ŷ_i = predicted value of the target feature

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Regression

RMSE (Root Mean Squared Error)

- Used for regression problems
- Square root of the MSE
- Easily interpretable (in the “y” units)
- “Punishes” larger errors
- RMSE recommended over MAE as it is more pessimistic (slightly overestimates prediction error; it is more conservative)

Regression: Evaluation Metrics

- Many other metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R^2, etc.
- RMSE and R^2 most popular, depending on the community
- RMSE is domain dependent (need to have an understanding of what the units mean)
- R^2 is domain-independent (in [0,1) range)

Evaluation/Deployment

Offline vs online evaluation

Offline: Data in place, features pre-computed, target feature known, cross-validation

Online:

- Feature computation can be a bottleneck (how long does it take, do we need all features?)
- Evaluate error of online predictions vs actual travel time (evaluation simulates real use case)

Evaluation/Deployment

Retraining the prediction models

- Is a once-off trained model still working well after a year? If not, how often do we need to retrain?
- Can we adapt the model by clever use of features? (i.e., features capture global and local information about traffic, time)
- How do we deal with route updates or new routes? Do we deploy only for subset of stable routes?

Modeling: Unsupervised Learning

- **Clustering:** finding structure in data by grouping samples into similar groups
- Might make sense to first cluster similar routes, learn a model for each cluster
- This might also help with system scalability

References

- **FMLPDA Book: Fundamentals of Machine Learning for Predictive Data Analytics**, by J. Kelleher, B. Mac Namee and A. D'Arcy, MIT Press, 2015 (machinelearningbook.com)

- Feature selection approaches: http://machinelearningmastery.com/an-introduction-to-feature-selection/