NODES 22

(g:graphData)-[:REQUIRES]->(e:graphEDA)

Daniel Bukowski, Graph Data Science CSA, Neo4j

About Me

Certified Public Accountant who worked as a financial intelligence analyst and forensic accountant in the public sector.

Anti-money laundering and fraud investigations at Ernst & Young and a Fortune 500 travel company.

M.S. Data Science from Northwestern University.

Security engineering and threat intelligence roles at startup and large tech companies.

Agenda

Background

EDA Conceptual Foundation

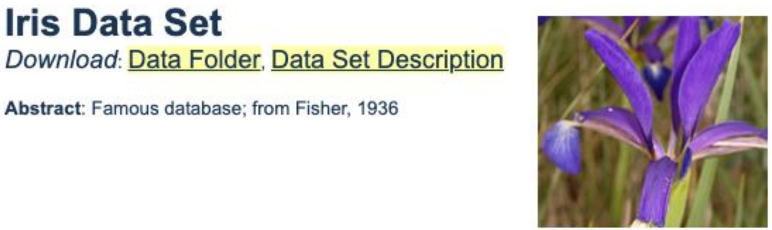
Notebook-Based Demos:

- Traditional (Tabular) EDA
- Graph Data Model and Data Loading
- Graph EDA *(with Neo4j Browser and Bloom)*

Notebooks are available at https://github.com/danb-neo4j/NODES2022_GraphEDA

Why are we here?

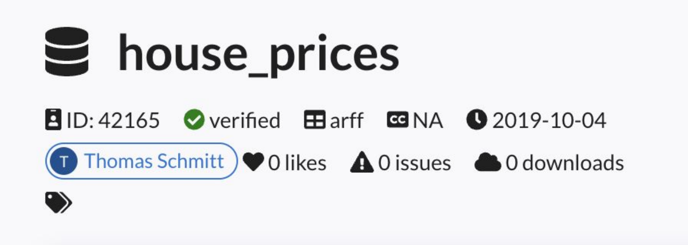
Learning on Tabular Data



Iris Data Set

Download: Data Folder, Data Set Description

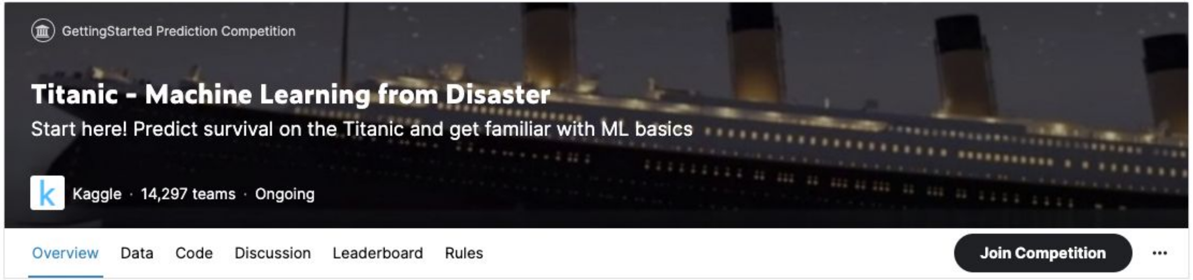
Abstract: Famous database; from Fisher, 1936



house_prices

ID: 42165 ✓ verified arff NA 2019-10-04

Thomas Schmitt 0 likes 0 issues 0 downloads

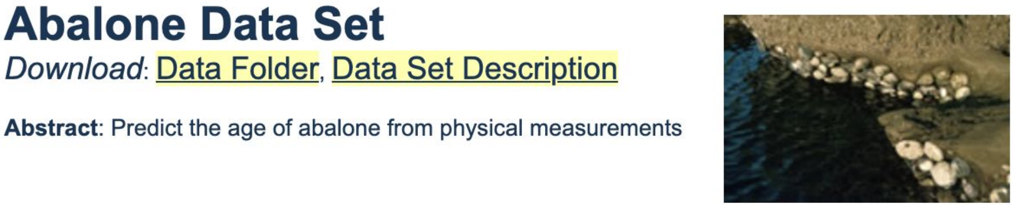


GettingStarted Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 14,297 teams · Ongoing

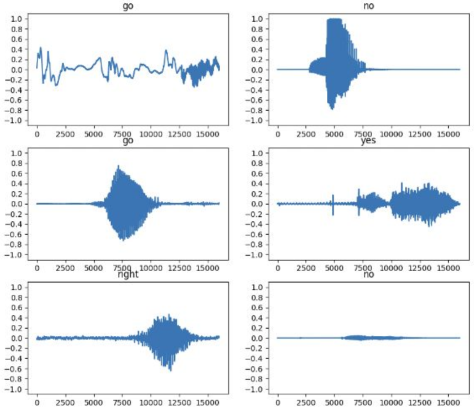


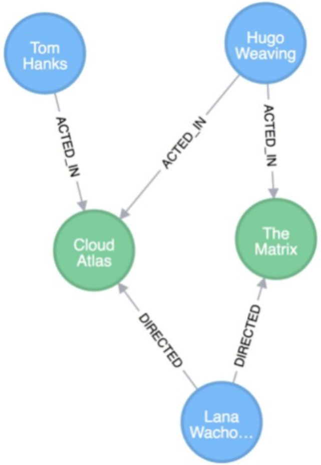
Abalone Data Set

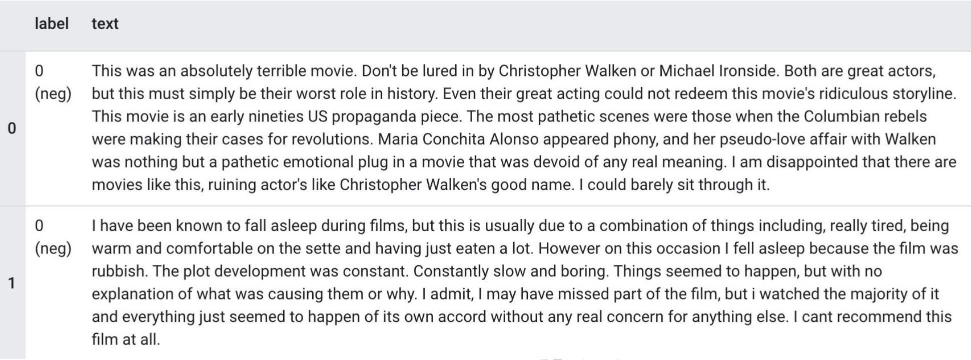
Download: Data Folder, Data Set Description

Abstract: Predict the age of abalone from physical measurements

But…Not All Data is Tabular







	label	text
0	0 (neg)	This was an absolutely terrible movie. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their worst role in history. Even their great acting could not redeem this movie's ridiculous storyline. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it.
1	0 (neg)	I have been known to fall asleep during films, but this is usually due to a combination of things including, really tired, being warm and comfortable on the sette and having just eaten a lot. However on this occasion I fell asleep because the film was rubbish. The plot development was constant. Constantly slow and boring. Things seemed to happen, but with no explanation of what was causing them or why. I admit, I may have missed part of the film, but i watched the majority of it and everything just seemed to happen of its own accord without any real concern for anything else. I cant recommend this film at all.

(t:TraditionalEDA)-[:ADAPT]->(g:GraphEDA)

Build Upon What Works

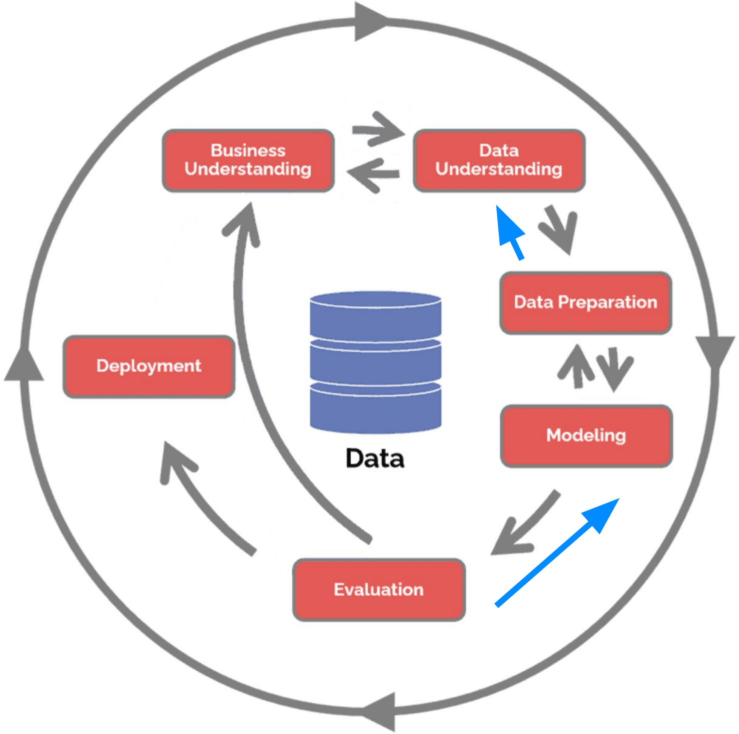
- Focus on business question
- Data completeness
- Data quality
- Data visualization
- Preprocessing and feature engineering
- Common tools and libraries

Adapt where Necessary

- Graph-shaped problem
- Informed graph data model
- Pre-graph feature engineering
- Nodes and relationships
- Visualize graph patterns
- Graph-based features

Exploratory Data Analysis

CRISP-DM

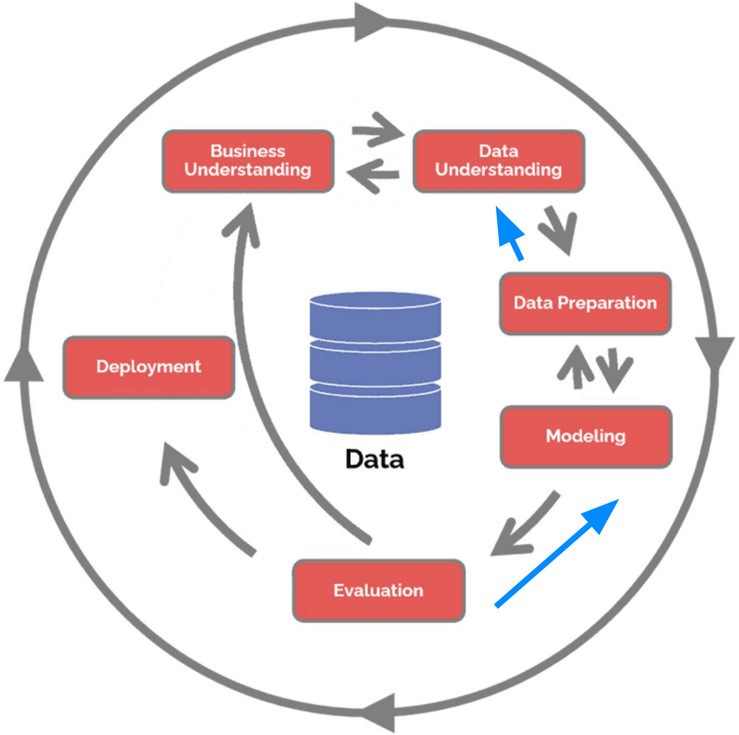


- CRISP-DM is a common framework many data scientists learn as part of their education.
- With small adjustments we can use it when working with graph data, especially if converting structured data to graph.
- *Many of the steps will be iterative when you change the form of the data.*

Adapting CRISP-DM to Graph

(Graph) Business Understanding:

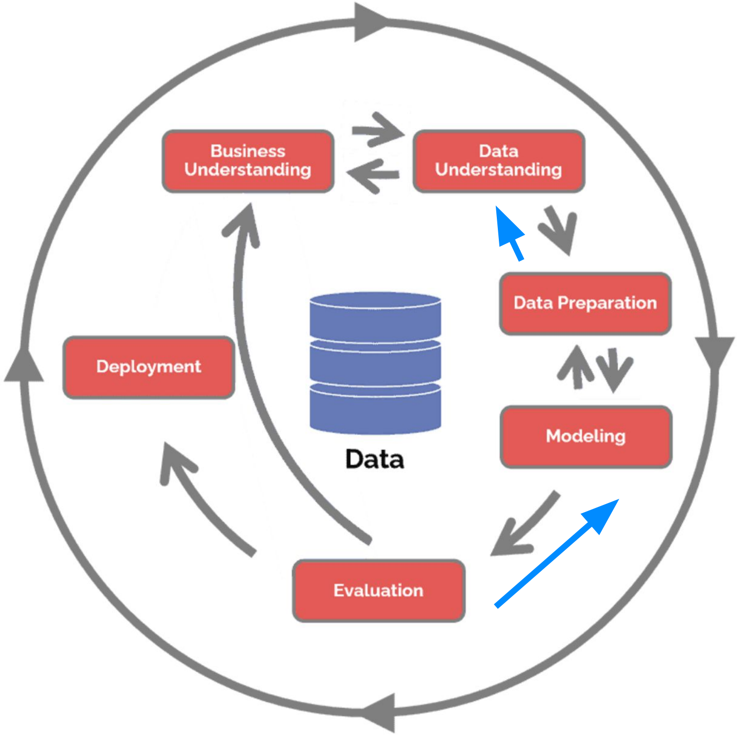
- Is the primary business question a ‘graph shaped’ problem?
- Does the problem focus on relationships?
- Will we be trying to identify relationships? Or patterns of specific relationships?
- How does graph fit into the overall data science workflow?



Adapting CRISP-DM to Graph

Data Understanding:

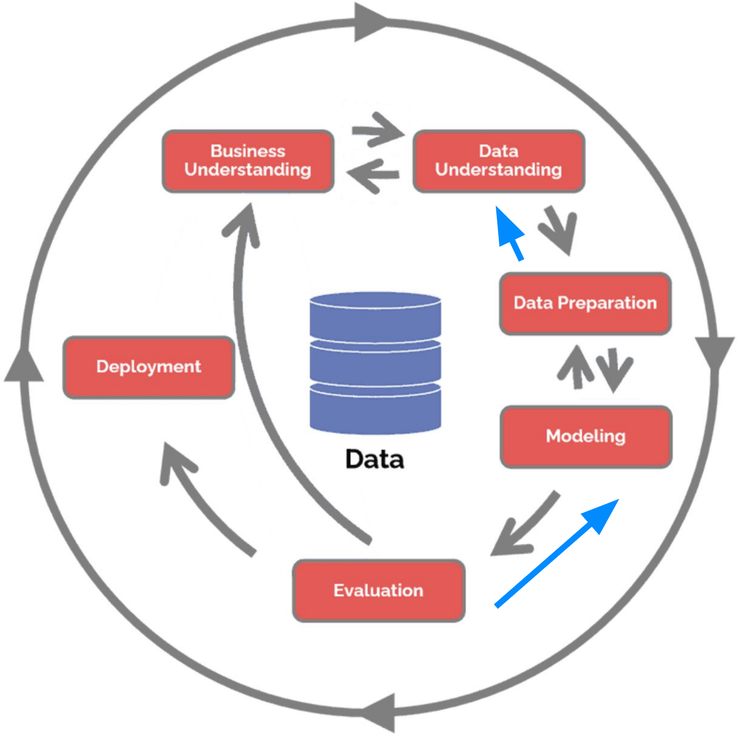
- Perform on data as-is to understand the starting point:
    - Inform the graph data model
    - Identify necessary preprocessing
- Perform again after the data is converted into graph format:
    - Graph-centric understanding
    - Validate the graph data model
    - Make adjustments if necessary



Adapting CRISP-DM to Graph

Data Preparation:

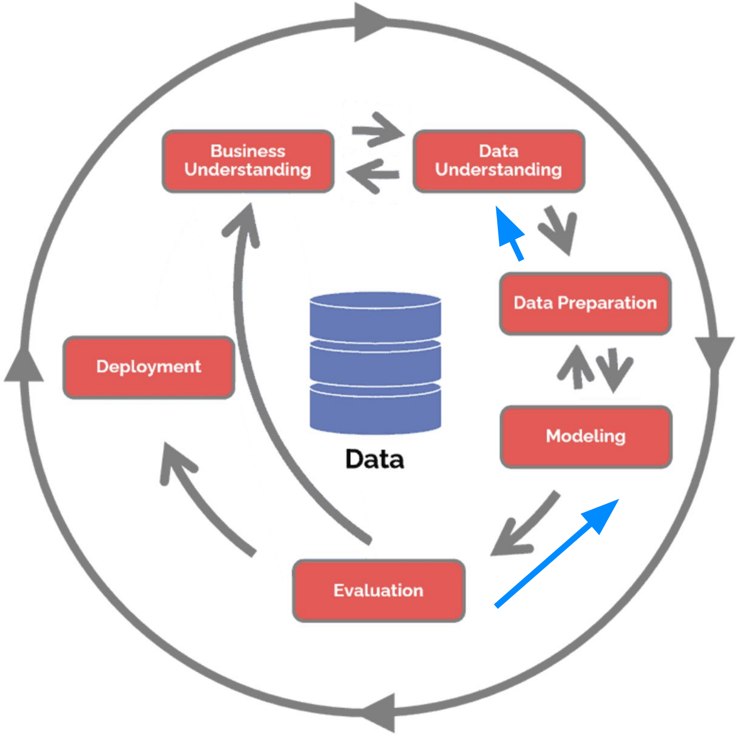
- Address any quality or formatting issues identified during EDA.
- Perform feature engineering to prepare for conversion to graph.
- Repeat once the data is in graph form to:
  - Address graph-based quality or formatting issues
  - Generate graph-based features



Adapting CRISP-DM to Graph

Modeling:

- Consider generating a pre-graph model such as random forest to:
    - Serve as a baseline
    - Identify feature importance
- After converting the data to graph form, leverage GDS algorithms for graph exploration and feature generation.



Notebook and Tool Demos

Notebooks are available at https://github.com/danb-neo4j/NODES2022_GraphEDA

Spaceship Titanic

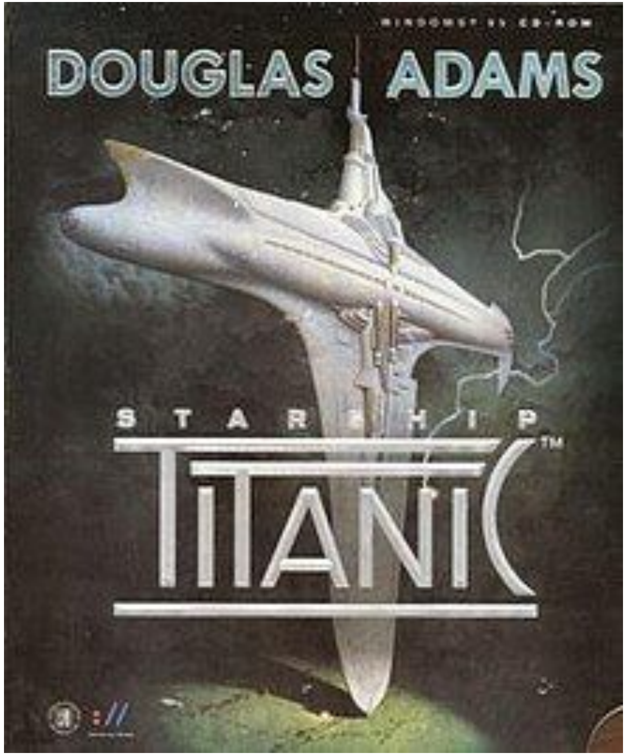


Image Source: Wikipedia

- Updated version of the Kaggle Titanic data set with more features and observations.
- Requires data cleaning and feature engineering.
- Not inherently graphy data or graphy use case.

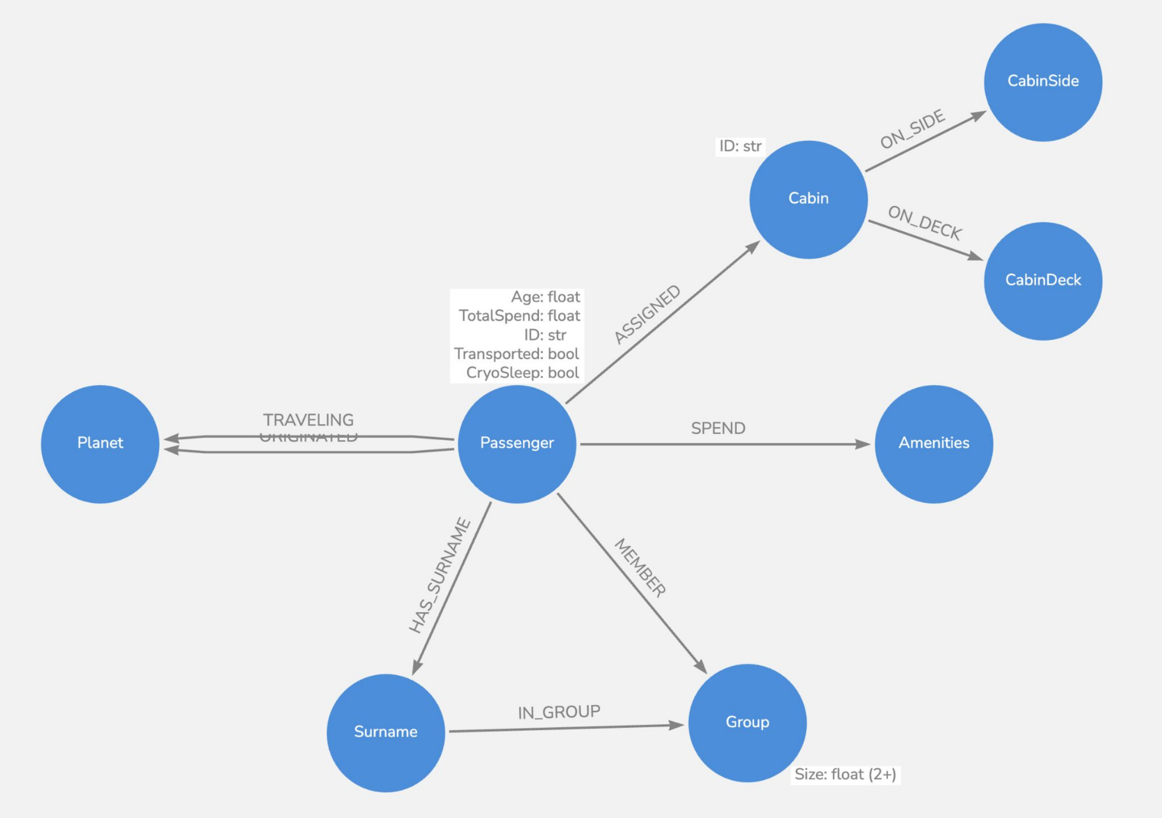
Tabular EDA Notebook:

(g:GraphEDA)-[:STARTS_WITH]->(t:TabularEDA)

Notebook to Load Graph Data

(td:TabularData)-[:BECOMES]->(gd:GraphData)

Graph Data Model



GraphEDA Notebook

(g:GraphEDA)-[:PRODUCES]->(u:GraphUnderstanding)

Conclusion

(t:TraditionalEDA)-[:ADAPT]->(g:GraphEDA)

Build Upon What Works

- Focus on business question
- Data completeness
- Data quality
- Data visualization
- Preprocessing and feature engineering
- Common tools and libraries

Adapt where Necessary

- Graph-shaped problem
- Informed graph data model
- Pre-graph feature engineering
- Nodes and relationships
- Visualize graph patterns
- Graph-based features

Now Your Turn…

How can you adapt your workflow to incorporate GraphEDA?

Review the code, build on the code, and please share any feedback or improvements with me and the community so we can all learn from each other.

Please let me know if you have any feedback or would like to see more content related to this workflow.

Check out ***GraphAcademy*** and the rest of the ***NODES 22*** lineup for additional learning resources.

Thank You for Attending!