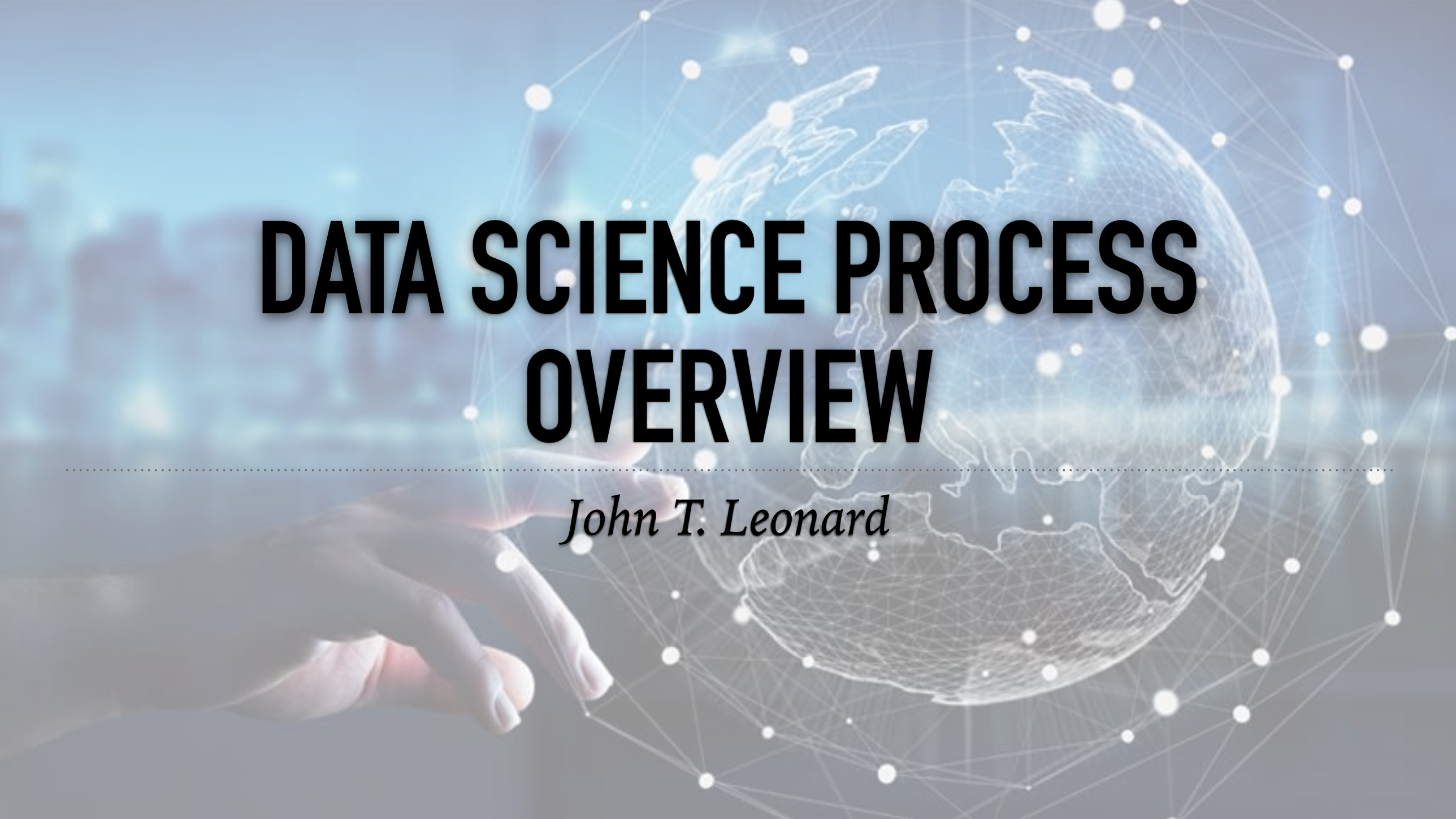


DATA SCIENCE PROCESS OVERVIEW

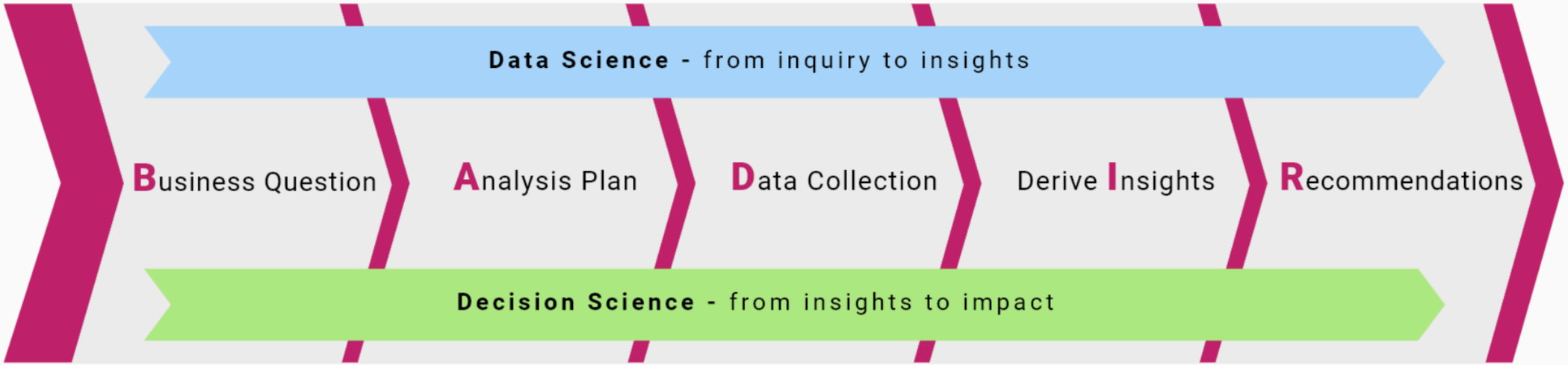
John T. Leonard

DATA SCIENCE PROCESS | GOALS

- Define data science process framework
- Outline analysis plan
- Project management structure/steps
- Potential risks in credit default data
- Preliminary insights

DATA SCIENCE PROCESS | BADIR FRAMEWORK

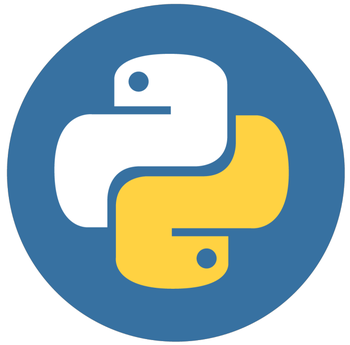
- BADIR: established framework used for data science process
- 5 key components: Business question, Analysis plan, Data collection, Insights, Recommendations



DATA SCIENCE PROCESS | BADIR FRAMEWORK: BUSINESS QUESTION & ANALYSIS PLAN

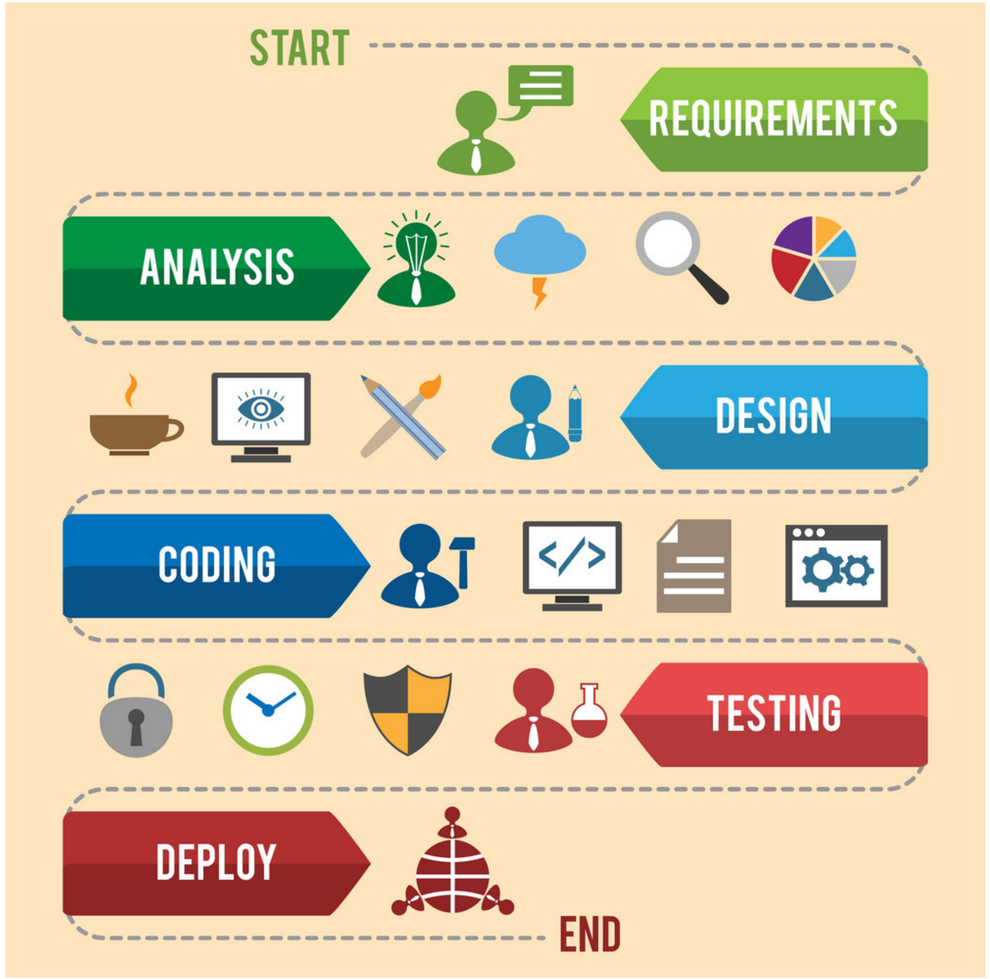
- Business Question:
  - Credit One experiencing increase in number of default loans. Risk of lost business if root cause / solution not identified.
- Analysis Plan:
  - Use python to inspect, clean, preprocess, validate ML models, and make predictions

Python + Analytics + Machine Learning



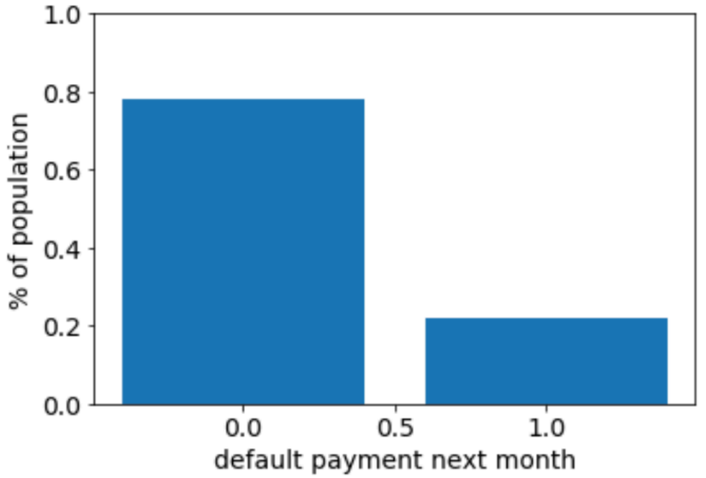
DATA SCIENCE PROCESS | MANAGEMENT

- Project stored as repository on GitHub:
  - LINK
- Jupyter Notebook w/ Python 3 kernel used for analytics + ML pipeline dev.
- Standard Python libraries used for enhance efficiency
  - pandas, numpy, matplotlib, sklearn, scipy
- Following v1.0.0 dev. in notebook, functions / operations transferred to python package for deployment.



DATA SCIENCE PROCESS | RISKS / ISSUES

- Dataset: ‘default of credit card clients.csv’
- Inspections performed:
  - date types
  - missing values
  - distributions
- No irregularities identified in preliminary inspection (data clean)
- Potential risks:
  - class imbalance: 80% pop: no default on loan
    - Can skew accuracy scores
    - Need to apply class balancing before model training



DATA SCIENCE PROCESS | PRELIMINARY INSIGHTS

- Defaults on payments next month strongly correlated to:
  - Pay_0
  - Pay_2
  - Limit bal
- Bill amounts have low correlation generally with defaults