outrageously
AMBITIOUS

Module 2: Organizing ML Projects

Duke
PRATT SCHOOL of ENGINEERING

87%

of ML projects fail*

*VentureBeat, 2019

Module 2 Objectives:

At the conclusion of this module, you should be able to:

1) Organize projects using the CRISP-DM data science process
2) Structure a ML project team and define roles
3) Organize project team work using best practices and track progress

outrageously
AMBITIOUS

ML Projects vs. Software Projects

Duke
PRATT SCHOOL of ENGINEERING

ML vs. software projects

- Relative to normal software projects, ML projects:
  - Require a broader set of skills / team
  - Have higher technical risk
  - Are more challenging to plan and estimate
  - Are harder to show progress
  - Require more ongoing support

Challenges of ML projects

- Probabilistic rather than deterministic
  - How to define “good enough”
  - Art of model building
  - Variance of model outputs
- Higher technical risk
  - Data needs and quality
  - Model limitations

Challenges of ML projects

- Much more up-front work required
  - Correct data issues
  - Identify features

- Often require change management
  - Not just another tool - changes the user’s workflow
  - Build model trust

outrageously AMBITIOUS

CRISP-DM Data Science Process

Duke

PRATT SCHOOL *of* ENGINEERING

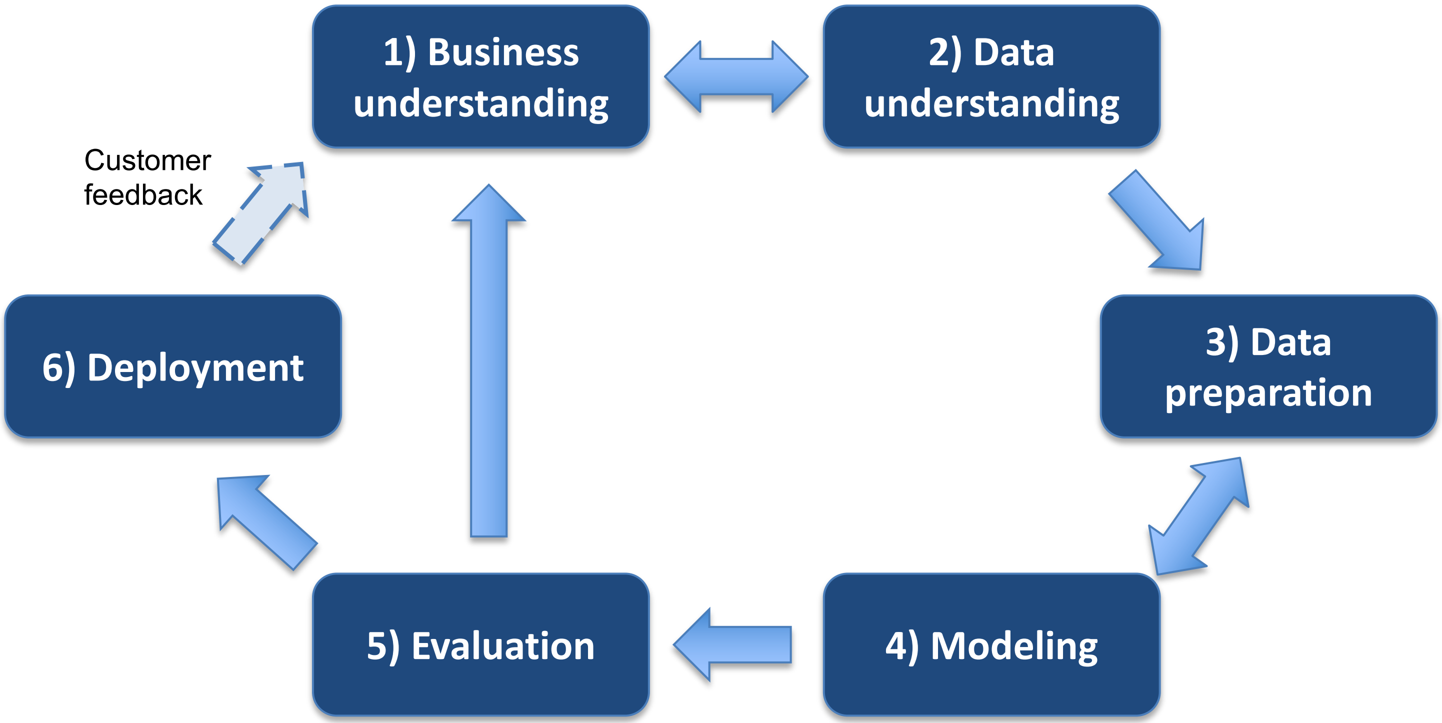
Why have a process?

- Prevent the tendency to jump right to solutions / modeling
- Avoid wasting time/money by working on a poorly defined problem
- Ensure discipline in doing the right things, in order
- Organize the work and the team responsibilities

CRISP-DM

- Developed in 1996 by a European consortium of companies
- Developed as a flexible, industry- agnostic approach to data mining projects
- Still the most widely used data science project methodology
- Major corporate champions include IBM

CRISP-DM Process



1) Business understanding

1.1 Define the problem	1.2 Define success	1.3 Identify factors
• Target user • Write the problem statement • Why it matters • How is it solved today? • Gaps in current state	• Quantify the expected business impact • Identify constraints • Translate impact into metrics – outcome & output metrics • Define success targets for metrics	• Gather domain expertise • Identify potentially relevant factors

2) Data understanding

2.1 Gather data	2.2 Validate data	2.3 Explore the data
• Identify data sources for each factor • Label data • Create features	• Quality control data • Resolve data issues – missing, erroneous, outliers	• Statistical analysis and visualization • Dimensionality reduction • Identify relationships & patterns

3) Data preparation

3.1 Split data	3.2 Determine feature set	3.3 Prepare for modeling
• Split data for training and test	• Feature engineering • Feature selection	• Encoding categorical features • Scale/standardize data • Resolve class imbalance

4) Modeling

4.1 Model selection	4.2 Model tuning
• Evaluate algorithms via cross-validation • Documentation and versioning	• Hyperparameter optimization • Documentation and versioning • Model re-training

5) Evaluation

5.1 Evaluate results	5.2 Test solution
• Model scoring on test set • Interpretation of model outputs and performance	• Software unit & integration tests • Model testing – unit tests, directional expectation • User tests

6) Deployment

6.1 Deploy	6.2 Monitor
• API framework • Product integration • Scaling infrastructure • Security • Software deployment process	• Model performance monitoring • Model retraining

CRISP-DM: Final thoughts

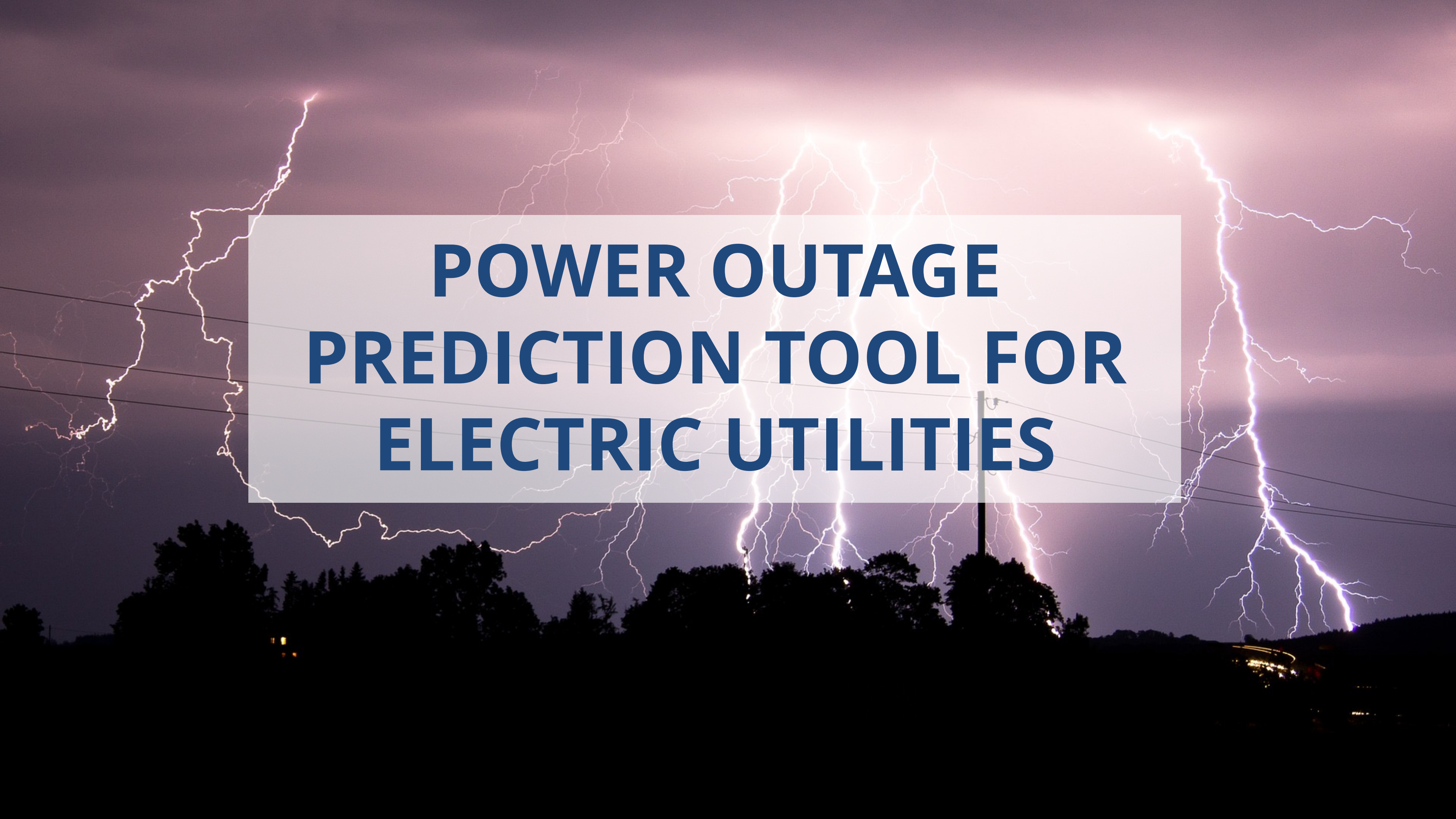
- Data science work is iterative, not linear
- Each step itself is iterative, as is the whole process
- You may want to adjust steps based on your project
- Skipping a step can be very dangerous!

outrageously
AMBITIOUS

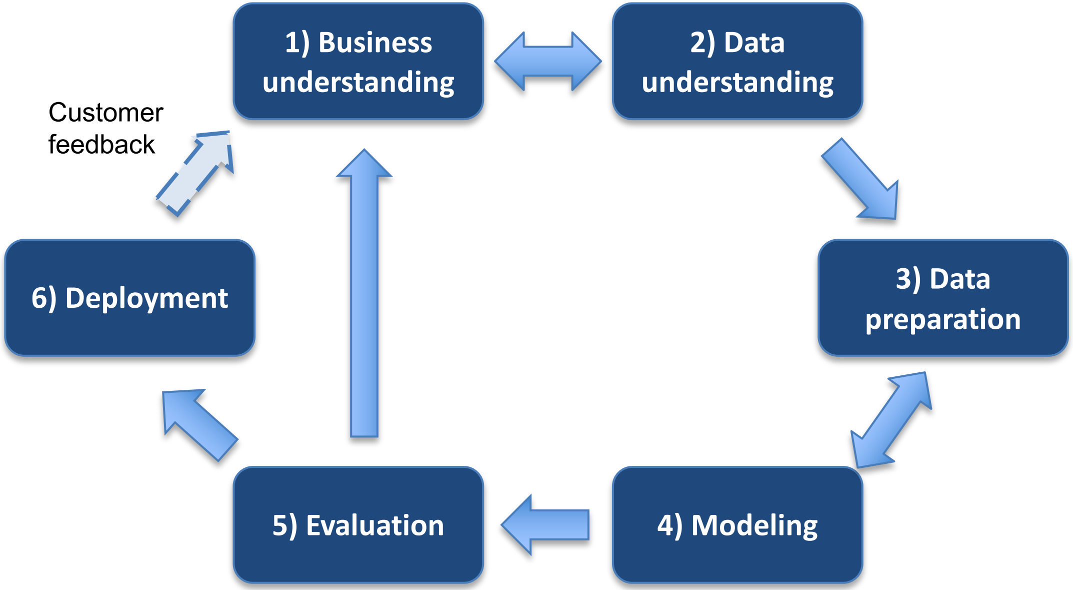
CRISP-DM Case Study

Duke
PRATT SCHOOL *of*
ENGINEERING

POWER OUTAGE PREDICTION TOOL FOR ELECTRIC UTILITIES



CRISP-DM Process



1) Business Understanding

Define the problem

Target user	Electric utility Director of Operations
Problem	Need to decide 2-3 days in advance how many crews to call in to repair expected storm damage
Why it matters	If they call in too many, they waste significant money. If they call in too few, customers are upset
Current state	They use weather forecasts and their own intuition to make an educated guess

1) Business Understanding

Define success

Expected impact	Improve restoration times and minimize wasted cost
Metrics	Outcome: Reduction of average restoration time Output: MSE of aggregate predictions
Targets	Outcome: Reduction of average restoration time by X minutes Output: MSE < XX
Constraints	Predictions must be delivered >48hrs in advance of storm start

1) Business Understanding

Identify factors

- Weather
  - Wind, gusts, precipitation, ice etc
- Density
  - Location/concentration of assets
- Trees
  - Proximity to power lines
  - Seasonality

2) Data Understanding

Source data

- Sources:
  - Weather: Weather providers
  - Trees: Satellite imagery vendors
  - Density: Utility customers
  - Historical outages (target): Utility customers
- Considerations:
  - How much data?
  - Sensitivity
  - Cost

2) Data Understanding

Validate data

- Significant missing data
- Map disparate sources to common geospatial resolution
- Outlier storms – major outages

3) Data Preparation

Define Features

- Many possible features
  - Weather parameters, time scales
- Interactions between features
- Possible missing features

4) Modeling

Model Selection

- Balance of performance & interpretability
- Single model or tailored models

5) Evaluation

Evaluate results / testing

- Performance on test set(s)
- Customer testing – live data
- Debugging – data issues

6) Deployment

Deploy

- Visualization product integration
- Customer change management

Monitor

- Model performance & outcomes
- Re-training plan

outrageously
AMBITIOUS

Team Organization

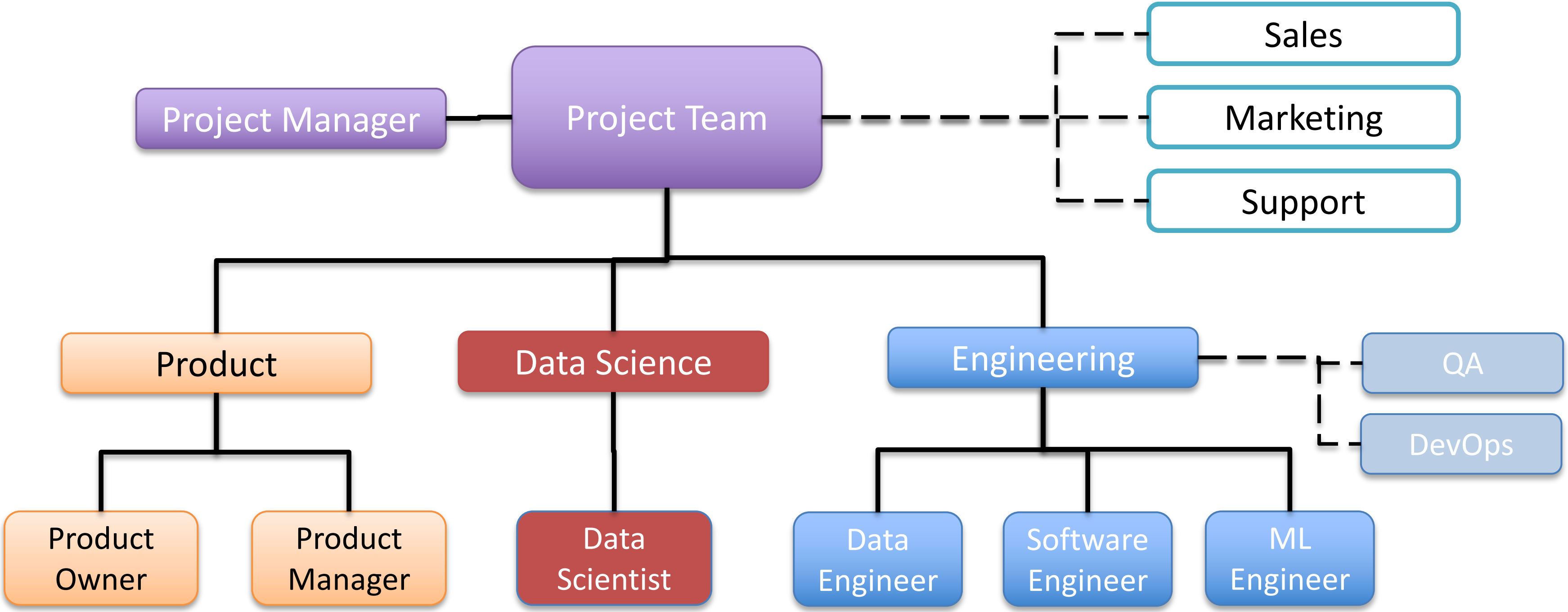
Duke
PRATT SCHOOL of ENGINEERING

Project team

- There is no “right” or “wrong” way to structure a team
  - Some teams are larger, some are smaller
  - Some are directly aligned, some are matrix
  - Different organizations use different titles
- What is important is defining responsibilities

Typical team roles

Some roles may have more than one person, or some people may have more than one role



Data Scientist vs. ML Engineer

Data Scientist

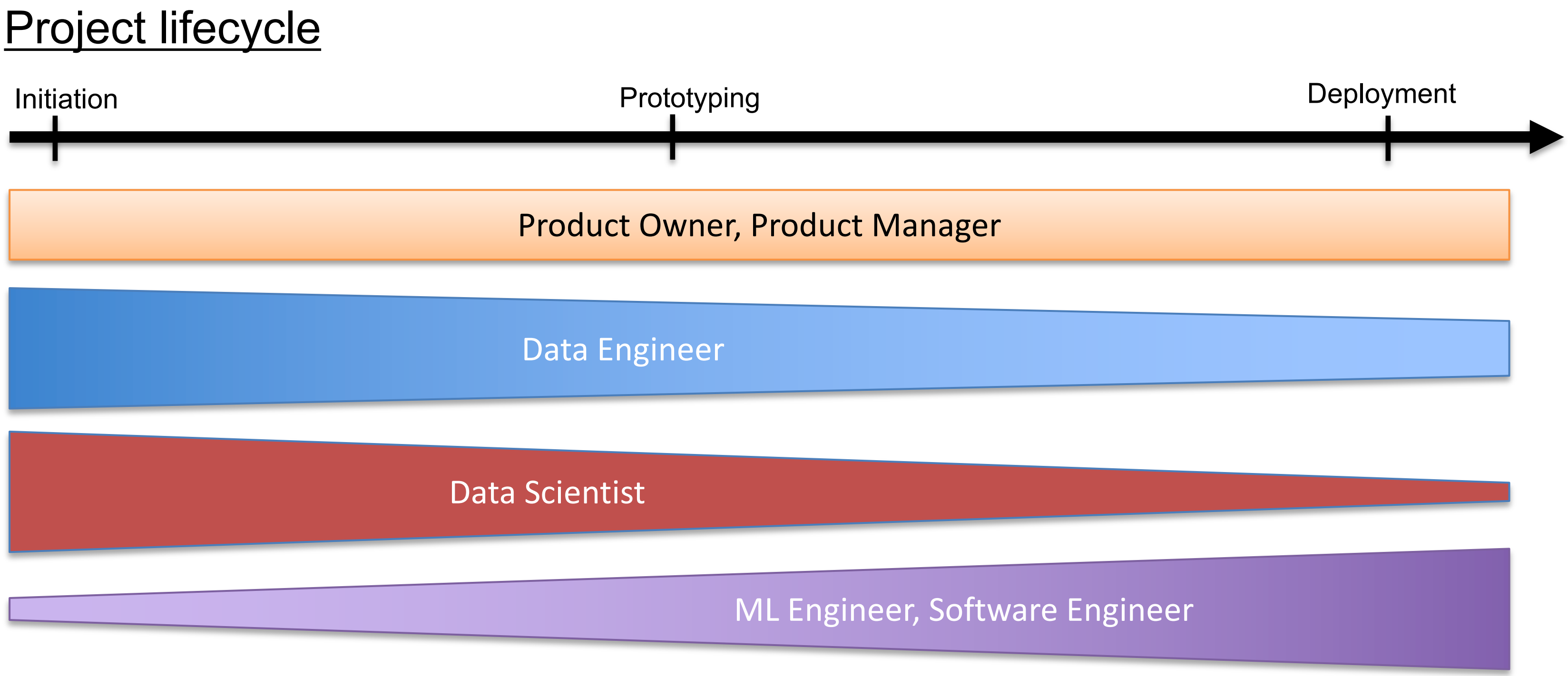
- Statistical / data science background, plus programming skills and domain expertise
- Gather, process & derive insights from data
- Determination of ML approach and prototyping

Data Scientist vs. ML Engineer

ML Engineer / MLOps

- Computer science or engineering background plus ML training
- Develop production data pipelines and ML system
- Work with software engineering & DevOps on model integration & deployment

Involvement over project cycle



Project Business Sponsor

- Having a business champion is a key success factor for AI projects
- Business champion secures resources and ensures alignment of project with company strategy
- Particularly important due to higher uncertainty & technical risk – protects team from business pressures

Organizing the Project

Agile approach to ML

- Sequence of iterative experiments
  - Explore a hypothesis
  - Build it, using more of CRISP-DM each time
  - Observe it in action, get feedback
  - Analyze results and repeat

Agile approach to ML

Iteration	What	CRISP-DM Steps Involved
1	Mockup of potential solution	Business Understanding

Agile approach to ML

Iteration	What	CRISP-DM Steps Involved
1	Mockup of potential solution	Business Understanding
2	Small subset of historical data and mocked up model	Business Understanding, Data Understanding

Agile approach to ML

Iteration	What	CRISP-DM Steps Involved
1	Mockup of potential solution	Business Understanding
2	Small subset of historical data and mocked up model	Business Understanding, Data Understanding
3	Real data, heuristic as model	Business Understanding, Data Understanding, Data Processing

Agile approach to ML

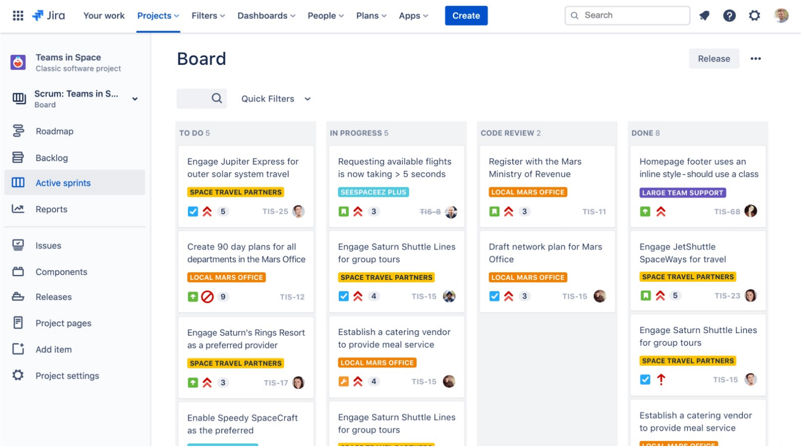
Iteration	What	CRISP-DM Steps Involved
1	Mockup of potential solution	Business Understanding
2	Small subset of historical data and mocked up model	Business Understanding, Data Understanding
3	Real data, heuristic as model	Business Understanding, Data Understanding, Data Processing
4	Real data, simple ML model	Business Understanding, Data Understanding, Data Processing, Modeling
…	…	…

Collaboration - cadence

- Monthly/quarterly roadmap sessions
  - Align on priorities
- Sprint planning & sprint reviews
  - Bi-weekly work planning
- Daily stand-ups
  - Not just for software dev – DoD, NWS
- Regular demo sessions
  - Visualize progress, get input

Collaboration - tools

- Roadmap & requirements
  - Confluence, Google Docs
- Project tracking
  - User stories, sprint planning, tracking
  - Jira, Trello
- Collaboration / version control
  - Git/GitHub



Atlassian, 2019. https://www.atlassian.com/software/jira outrageously AMBITIOUS

Measuring Performance

Duke

PRATT SCHOOL *of* ENGINEERING

Metrics

Outcome Metrics

- Refers to the desired business impact on your organization or for your customer
- Stated in terms of the expected impact (which is often $)
- Does NOT contain model performance metrics or other technical metrics

Metrics

Output Metrics

- Refers to the desired output from the model
- Measured in terms of a model performance metric
- Typically not communicated to the customer
- Set this AFTER setting the desired outcome

Tracking progress on metrics

Output Metrics

- Model validation and testing
- Can require customer input data

Outcome Metrics

- Hindsight scenario testing
- A/B testing
- Beta testing

Non-performance considerations

- Explainability / interpretability
  - Easier to debug issues and identify bias
  - Fault tolerant vs. fault intolerant

- Data and computational cost
  - Cost of sourcing & storing data
  - Compute requirements for training & inference

outrageously
AMBITIOUS

Wrap-up

Duke
PRATT SCHOOL of ENGINEERING

Wrap Up

- ML projects differ substantially from software projects
- Process is critical to ensure doing the right things in the right order
- Process does NOT imply linear working – ML is highly iterative