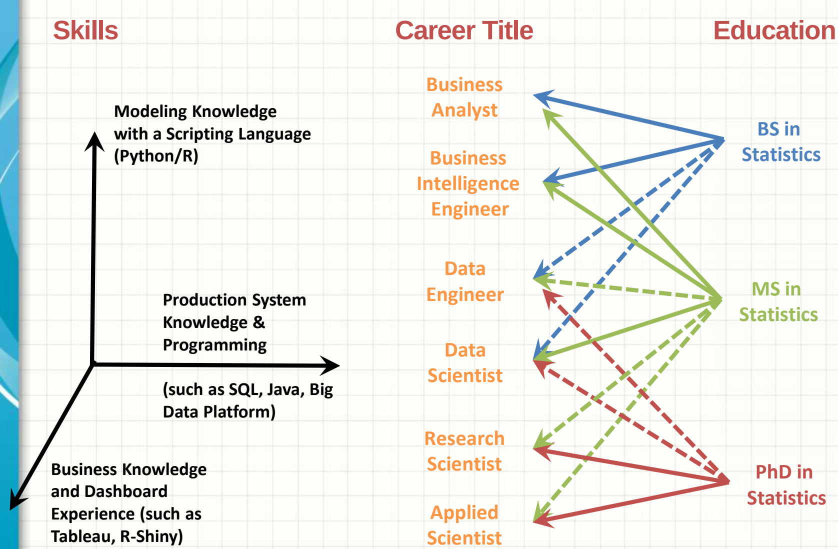
SOFT SKILLS & PROJECT CYCLES IN DATA SCIENCE

DIFFERENCE BETWEEN STATISTICIAN AND DATA SCIENTIST

Science	Engineering
Physics	Electrical Engineering
Statistics	Statistical Engineering + Big Data & Software = Data Science

Collection of materials for statistical engineering: *http://asq.org/divisions-forums/statistics/quality-information/statistical-engineering*

Statis-tician	Data Scientist
Relatively focus on modeling (i.e. science)	Mainly focus on business problem & result (i.e. engineering)
Bring data to model	Bring models to data
Data is relatively small in size and clean in text file formats	Need to work with messy and large amount data in various formats
Usually structured data	Both structured & unstructured data
Usually isolated from production system	Usually embedded in production system



If you are good at all three areas, you become a "*full-stack*" data scientist!

Each career path has its own promotion cycle.

Generalist

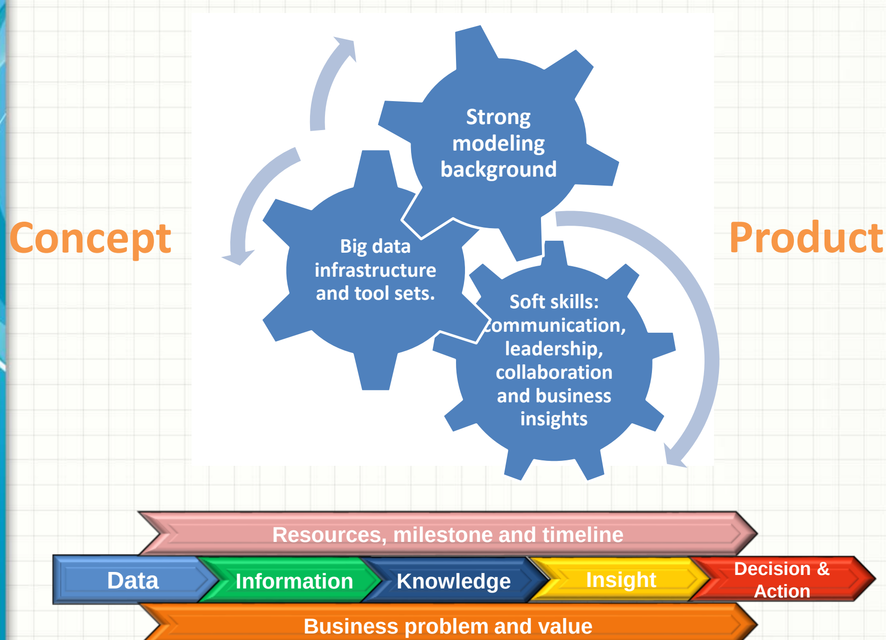
- Business problem formulation skills
- Data preprocessing skills
- Statistical and machine learning methods
- Deep learning applications knowledge
- Model building process experience

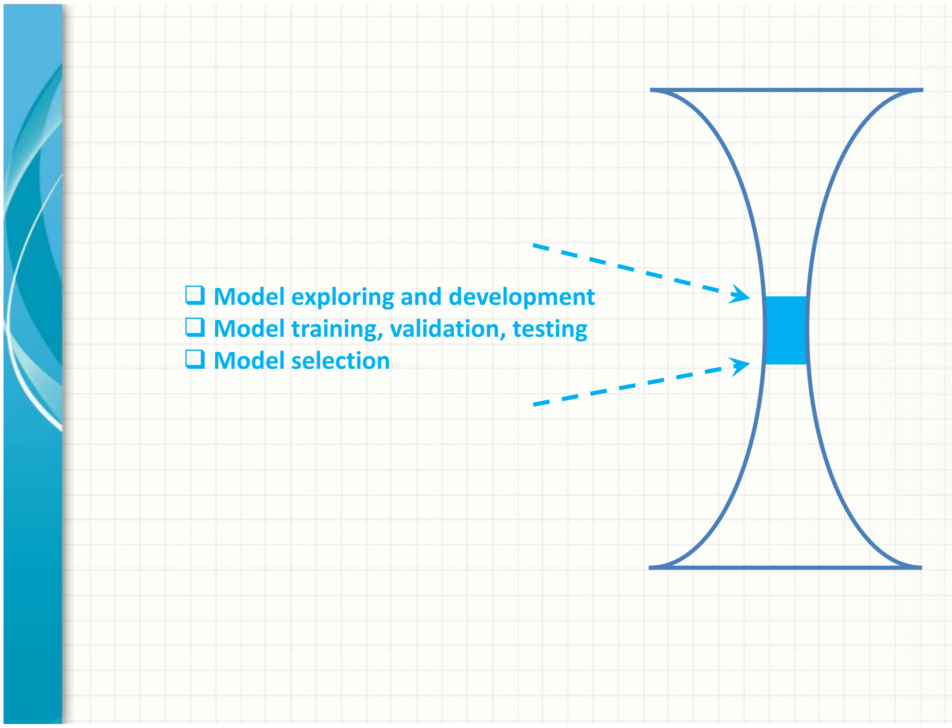
Specialist

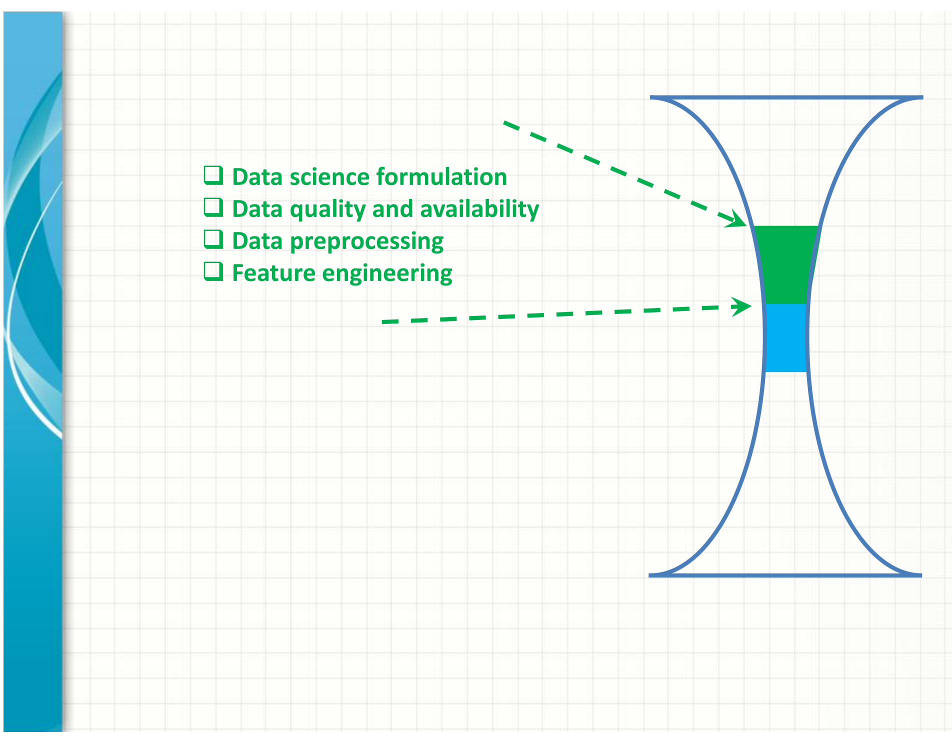
- Natural language understanding
- Image and video analysis
- Voice recognition
- Language translation

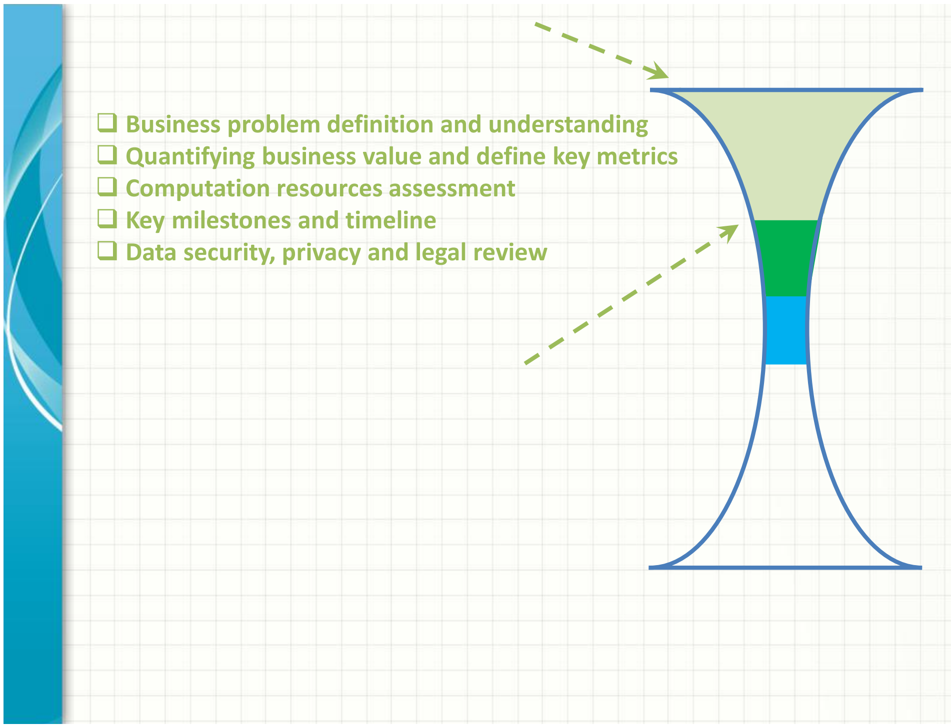
DATA SCIENCE PROJECT CYCLES

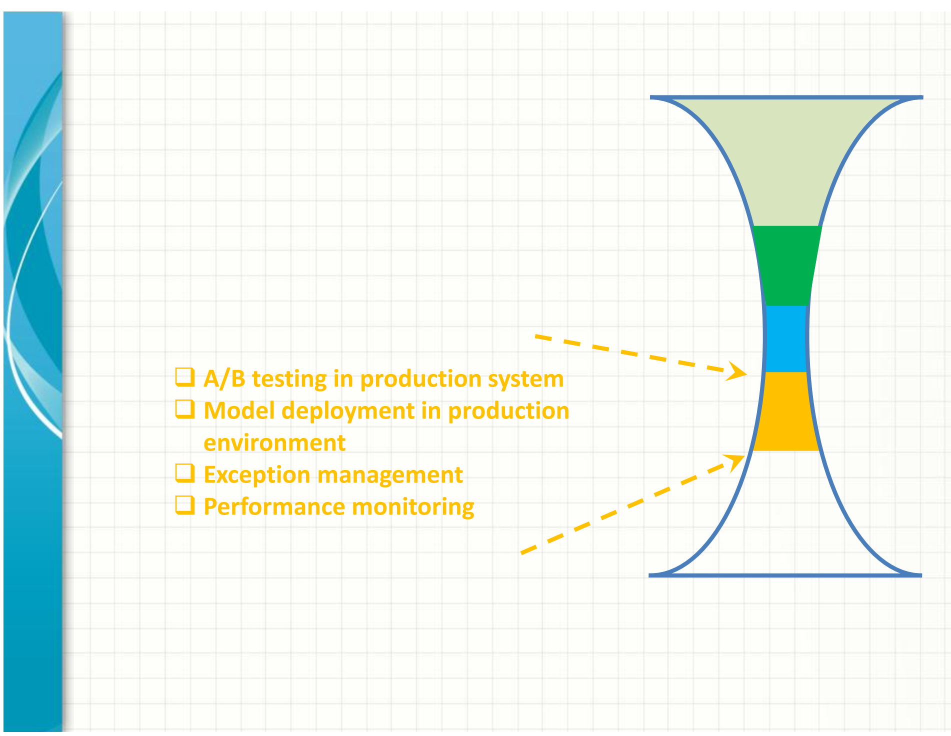
Data Science Project Cycle (Overview)



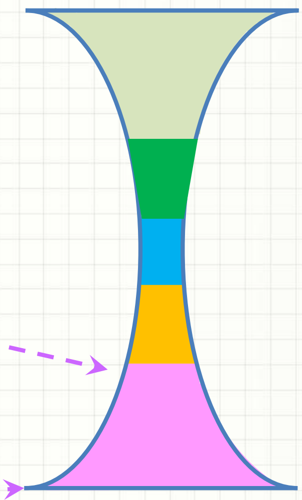

❑ Model exploring and development
❑ Model training, validation, testing
❑ Model selection


❑ Data science formulation
❑ Data quality and availability
❑ Data preprocessing
❑ Feature engineering

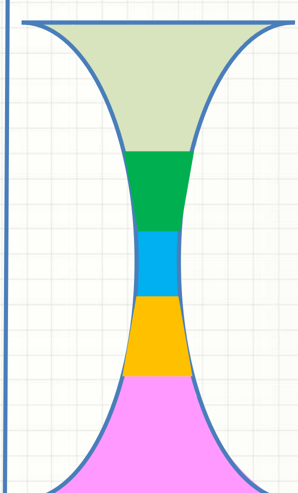

❑ Business problem definition and understanding
❑ Quantifying business value and define key metrics
❑ Computation resources assessment
❑ Key milestones and timeline
❑ Data security, privacy and legal review


❑ A/B testing in production system
❑ Model deployment in production environment
❑ Exception management
❑ Performance monitoring

- ❑ Model tuning and re-training
- ❑ Model update and add-on
- ❑ Model failure and retirement
- ❑ …



Project Cycle



Planning
- ❑ Business problem definition and understanding
- ❑ Quantifying business value and define key metrics
- ❑ Computation resources assessment
- ❑ Key milestones and timeline
- ❑ Data security, privacy and legal review

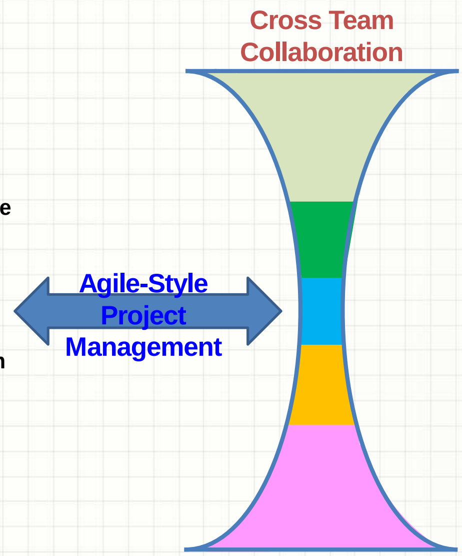
Formulation
- ❑ Data science formulation
- ❑ Data quality and availability
- ❑ Data preprocessing
- ❑ Feature engineering

Modeling
- ❑ Model exploring and development
- ❑ Model training, validation, testing
- ❑ Model selection

Production
- ❑ A/B testing in production system
- ❑ Model deployment in production environment
- ❑ Exception management
- ❑ Performance monitoring

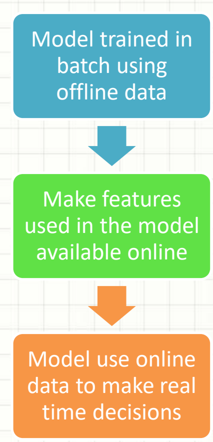
Post-Production
- ❑ Model tuning and re-training
- ❑ Model update and add-on
- ❑ Model failure and retirement

- Business teams
  - Operation team
  - Business analyst team
  - Insight and reporting team
- Technology team
  - Database and data warehouse team
  - Data engineering team
  - Infrastructure team
  - Core machine learning team
  - Software development team
  - Visualization dashboard team
  - Production implementation
- Project management team
- Program management team
- Product management team
- Senior leadership team
- Leaders across organizations



Online vs Offline Training

- Concept of offline data
  - Historical data in the data warehouse system
  - Slow to retrieve (i.e. hours to get needed data)
  - Very rich in general (i.e. click stream, detailed shopping history etc.)
  - Low cost to maintenance in data warehouse (i.e. hard disk)
  - Can be extracted in batch as raw material for features
- Offline data are typical used for
  - Data exploratory
  - Feature engineering
  - Statistical or machine learning model development and selection
- Concept of online data
  - Data available in real time to a model for real time decision making
  - High cost to maintenance for low latency
  - Limited selection of data usually available
  - Additional data can be added with effort



There are also **offline-only models** with regular batch process; and **models training using online real time data** and deploy in real time depending on different applications.

Common Pitfalls of Data Science Projects

Project Planning Stage

❑Solving the wrong problem

- Vague description of business needs
- Misalignment across many teams (Scientist, Developer, Operation, Project Managers etc.)
- Scientist team are not actively participating in the problem formulation process

❑Too optimistic about the timeline

- Project managers may not have past experience for ML and data science projects
- Many ML method-specific uncertainties are not accounted for at planning stage
- ML and data science projects are fundamentally different from each other and from software development projects (such as online vs. offline model, batch model, real time training, re-training etc.)

❑Over promise on business value

- Unrealistic high expectation (i.e. advertisement vs actual product)
- Many assumptions about the project are usually not true
- Similar projects from other teams/companies are not evaluated thoroughly to set realistic expectation of time line and outcome

Problem Formulation Stage

- ❑ Too optimistic about standard statistical and ML methods
  - Extra efforts are needed to abstract business problem into a set of analytics problems
  - Standard methods are usually not enough to solve the business problems
- ❑ Too optimistic about data availability and quality
  - “Big data” is not a guarantee of good and relevant data, usually big and messy
  - Ideal data for the business problem is almost always not available
  - Unexpected efforts to bring the right data
  - Under estimate effort to evaluate quality of data
- ❑ Too optimistic about needed effort on data preprocessing
  - Table or column descriptions are not detailed enough
  - Lack in-depth understanding of the dataset
  - Under estimate of date preprocessing (such as dealing missing data)
  - Under estimate the effort for feature engineering
  - Mismatch between different data sources (such as online vs offline, different tables etc.)

Modeling Stage

- ❑ Un-representative data (such as lack of future outlook of what will happen in production or biased data)
- ❑ Too optimistic about model selection and hyper-parameter tuning to reach desired performance
- ❑ Overfitting and obsession for complicated models (heavy models may leads poor production performance)
- ❑ Take too long to fail

Productionization Stage

❑ Bad production performance

- Lack shadow mode dry run
- Lack needed A/B testing
- Data availability and stability issue in real time
- Lack exception management on issues such as timeout and missing data

❑ Fail to scale in real time applications

- Computation capacity limitation
- Real time data storage and processing limitation
- Latency constrains
- Not enough engineering resources (i.e. SDE, DE) during implementation

Post-Production Stage

❑ Missing necessary checkup

- Lack model monitoring for key metrics
- Lack exception notification
- Lack model failures/timeout notification
- Online feature not stored for future analysis

❑ Production performance degradation

- Not aware of dynamic nature of the business problem
- Not aware of changing input data quality and availability
- Lack model tuning and re-training plan
- Lack model retirement or replacement plan

Soft Skills

Leading With Statistics

- ❑ Strong modeling background should guide the project from the beginning of the cycle
- ❑ Keep a high standard with data-driven and model-backed decision making process
- ❑ Clearly communicate potential issues for the project as well as providing proactive suggestions

Communication: Speaking the Same Language

- ❑ Interact with multiple teams across the entire project cycle
  - ✓ Easy to understand language that everyone understand
  - ✓ Be clean on deliverables, timeline and resource allocation
- ❑ Technical modeling part requires communication skills too
  - ✓ Statistician, Operation Researcher, Economist, Computer Scientist, Market Researcher, …
- ❑ Need to be familiar with different terminology, for example:
  - ✓ Label = Target = Outcome = Class = Response = Dependent Variables (i.e. Y)
  - ✓ Features = Attribute = Independent Variables = Predictors = Covariates (i.e. X)
  - ✓ Weights = Parameters
  - ✓ Learning = Fitting
  - ✓ Generalization = Applying to population or test data
  - ✓ Sensitivity = recall = hit rate = true positive rate

Communication: Different Styles

Statistics	Data Science
All kind of errors ◦ Type-I error ◦ Type-II error ◦ Mean square error	Accuracy Precision
Dummy variables Lack of fit Loss function	One-hot encoding Faithfulness Information gain
Failure Rate Hazard Model Penalty Discrimination Function …	Golden Standard Smart Algorithm Intelligent Procedure Knowledge Discovery …

* Partially Adopted from Dennis Lin's FTC Talk

Business Domain Knowledge

- Many technical skills and soft skills are easily transferable from one business sector to another such as
  - ✓ Statistical and ML methods, SQL, Spark
  - ✓ Procedures and best practices in problem formulation and modeling
  - ✓ Communication, leadership and collaboration

- How to quickly obtain business domain knowledge?
  - ✓ Very similar to statistical consulting projects
    - Understand the current decision making process
    - Get familiar with current data acquisition procedures
    - Understand current modeling process and data flow
    - Outline business problems to solve
  - ✓ Job shadowing with office and field agents
    - Ask questions to understand business operation procedures
    - Identify current pain points and known-unknowns
    - Outbox thinking to identify unknown-unknowns
  - ✓ Current best practice across the industry
    - Read research/white paper, attend conference, meetup and talks
    - Reach out to domain specific experts

Keep on Track for Data Science Career

- Learning New Methods
  - Deep Learning
  - Reinforced Learning
- Keep up with New Tools
  - TensorFlow, MxNet etc.
  - Spark
  - R/Python
  - Dynamic Dashboard
- Explore New Applications
  - Internet of Things (IoT)
  - Robotics
  - Automatic Driving Cars
- Apply New Methods to Existing Applications
  - Identify problems at daily work
  - Apply novel ways for existing solutions
  - It could be much faster / more accurate / more efficient etc.
- Brand Yourself
  - LinkedIn
  - GitHub
  - Blogs and Posts
  - Personal Professional website

Fun Video: THE EXPERT
https://youtu.be/BKorP55Aqvg

Hilarious but sadly true for many data science projects!
Probably you are the only data scientist in the room next time, be prepared to fight back!

Learning outcomes:

After taking the CE course, participants will:

1. Get familiar with deep learning methods such as feedforward neural network, CNN, and RNN with hands-on how to apply these deep learning methods through R keras package with TensorFlow backend
2. Understand data science in general and the end-to-end data science project cycles.
3. Get familiar with cloud-based big data platforms (i.e., Databrick's Spark) for data preprocessing and model development that are widely used in the development and production setting for industry and know how to transit from academia environment to enterprise environment quickly.
4. Learn soft skills to ensure the successful delivery of data science projects and get familiar with typical data science project pitfalls.

THANK YOU!