Data Science Overview

LESSONS & BEST PRACTICES

JAY SWARTZ
JAYWSWARTZ@GMAIL.COM

My Background

- Multiple Domain Expertise Spanning 45+ Years
  - Led Seagate Technologies Global IT Function
- C-Level Roles
- Start-Up Experience
  - Founder
  - Advisor
- Consulting & Contracting
  - Virtual Environments
  - Machine Learning
  - Engineering
  - Marketing

Overview

- Strategic Factors
- Success Plan
- Data Science Scope
- Processes
- Applications
- Q & A

Strategic Intents

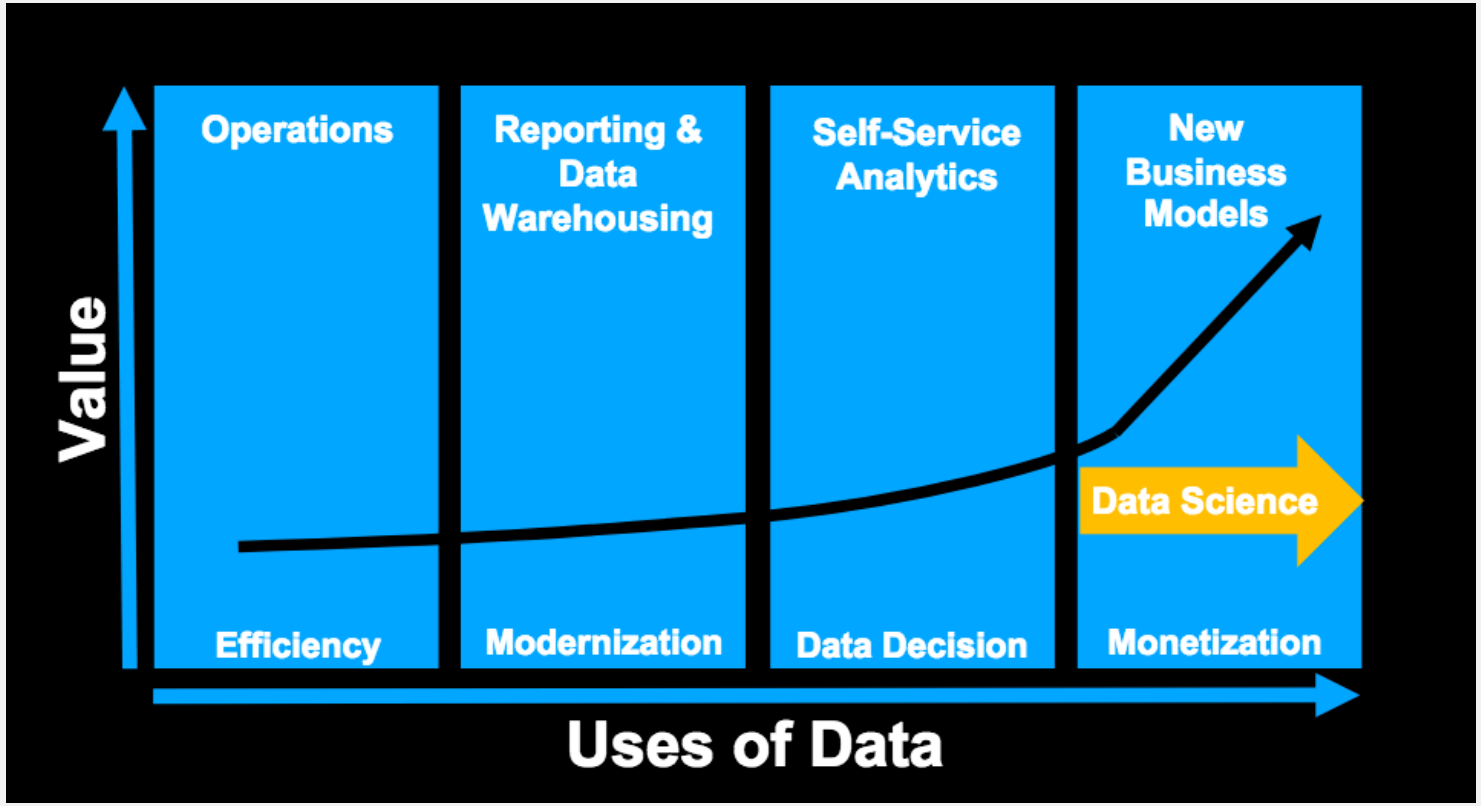
- ROI-Based Project Selection
  - Lifetime Value
  - Offense Over Defense
- Architect for Scale
- Concurrent Workflows
- Extensible Approaches
  - Reusable Patterns
- Adjacent Compatibility

Success Factors

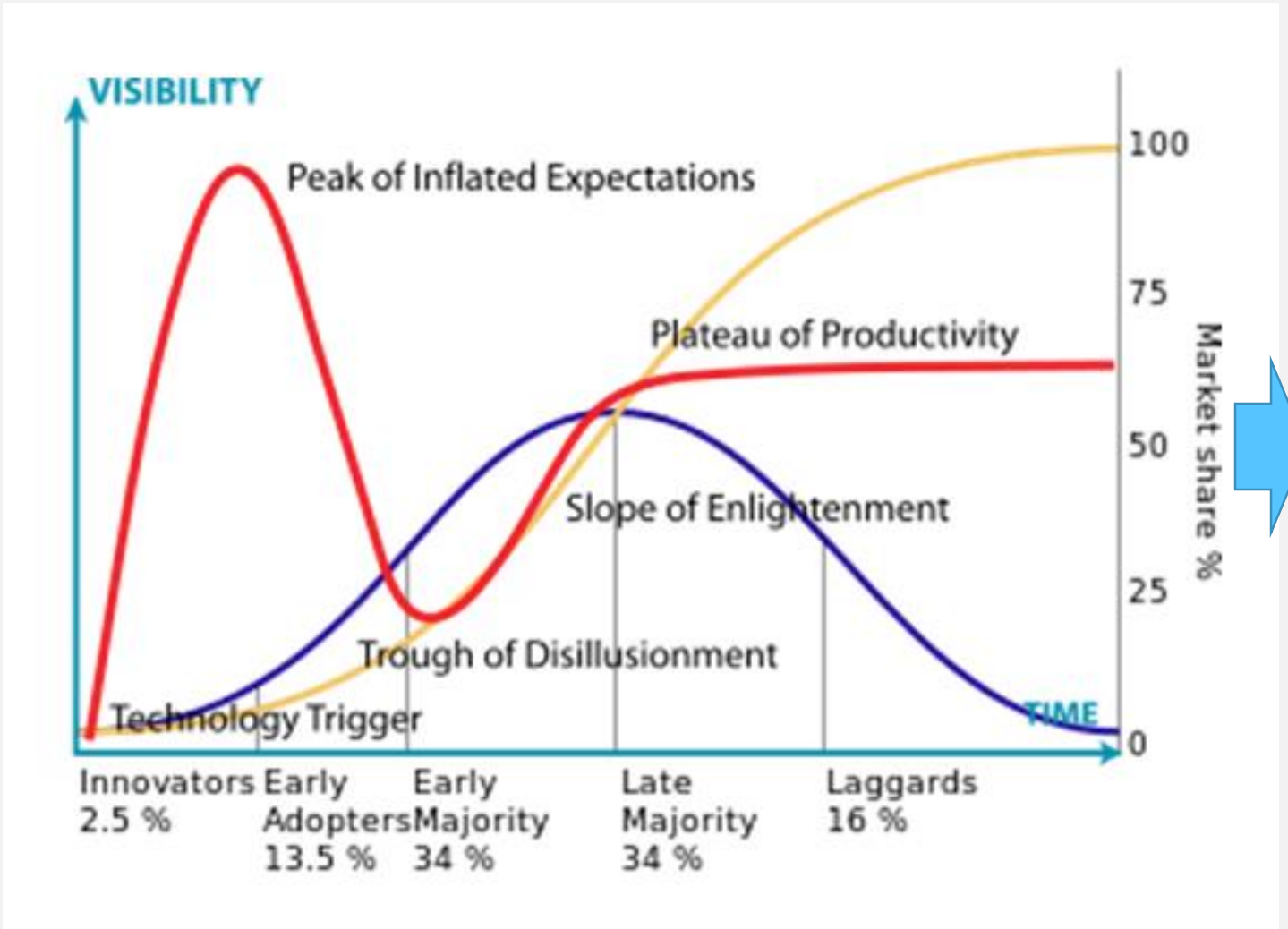
- Data Driven Decisions
- Tool Investment
  - Time & Error Reduction
- Shaping Techniques
- Reliable Methods
  - Incremental Process
  - Agile Kanban/Cynefin
  - Scope Bounding
- Expectation Management
  - Realistic Learning Abilities

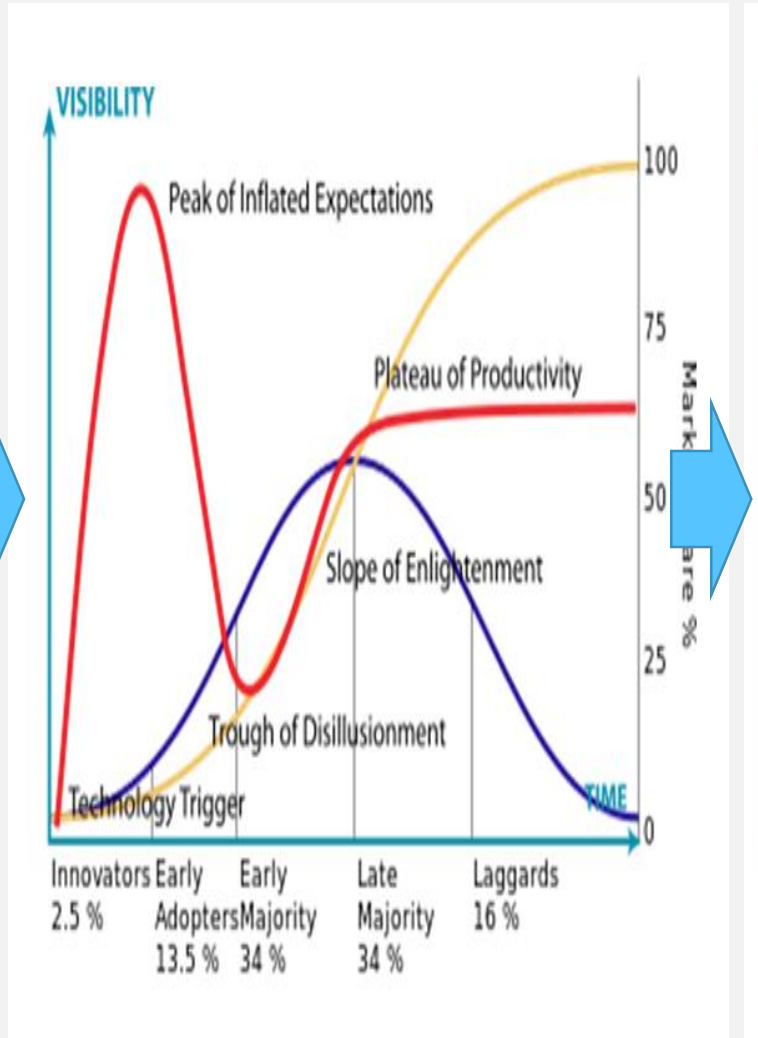
Data Process Maturity

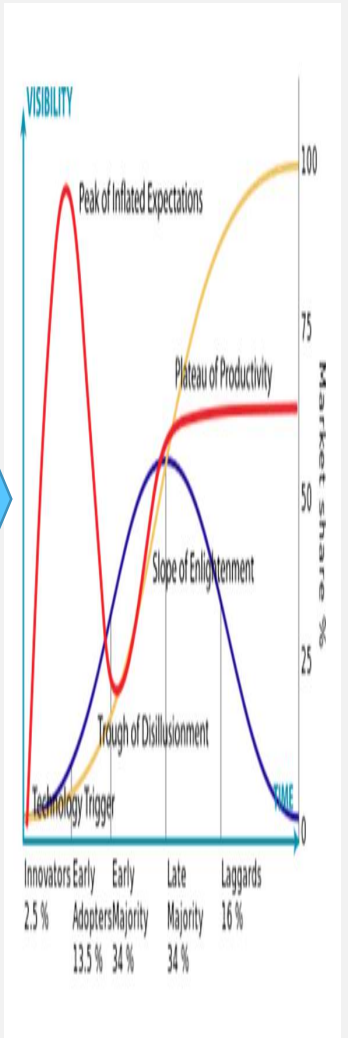
- Each Step Requires Previous Steps
- Bring All Data Streams to Same Phase Before Moving to Next Phase
- Decisions Based on Facts Remove Risk Associated with Instinctual Decisions



Intuition: Non-Obvious AI Advances







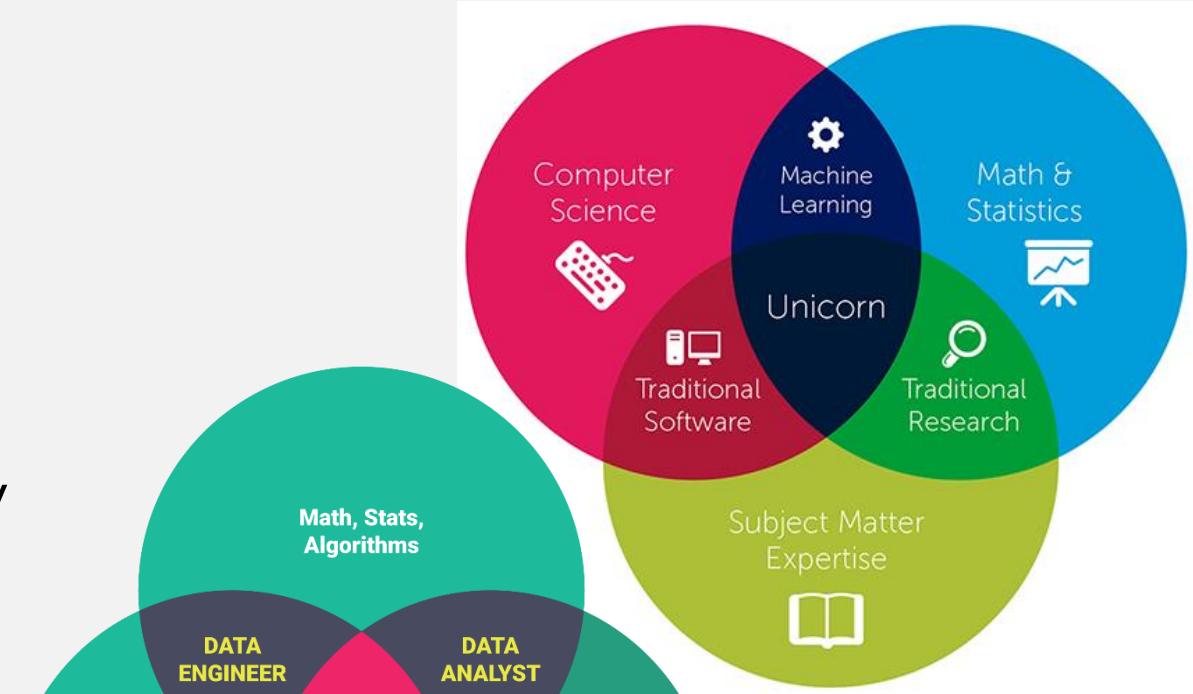
Worksheet for Data Science Success

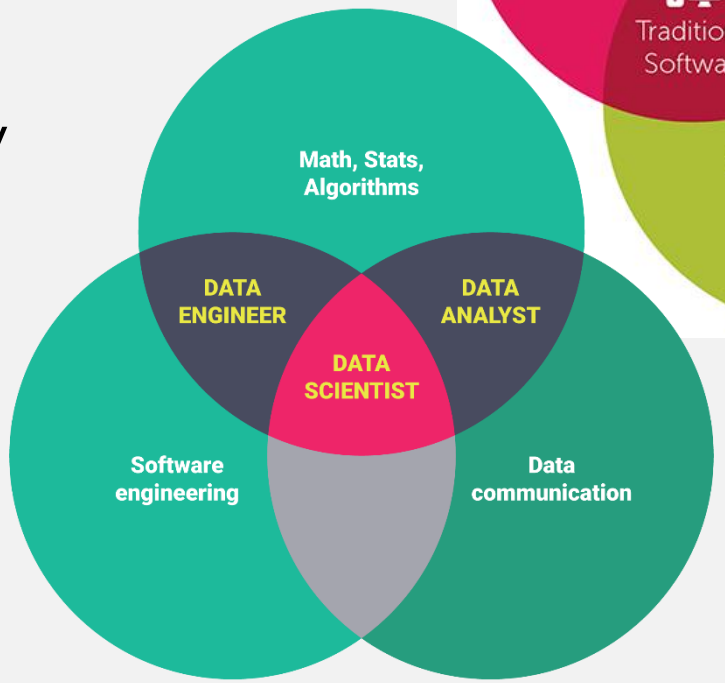
- Tenets
  - Clear Vision
  - Cross-Functional
  - Data Access
  - Data Hygiene
  - Recognition
- Factors
  - Culture
  - Data
  - Talent
  - Technology
- Artifacts
  - Current State
  - Future State
  - Targets and Goals
  - Action Plan
    - Immediate Next Steps
    - Major Milestones

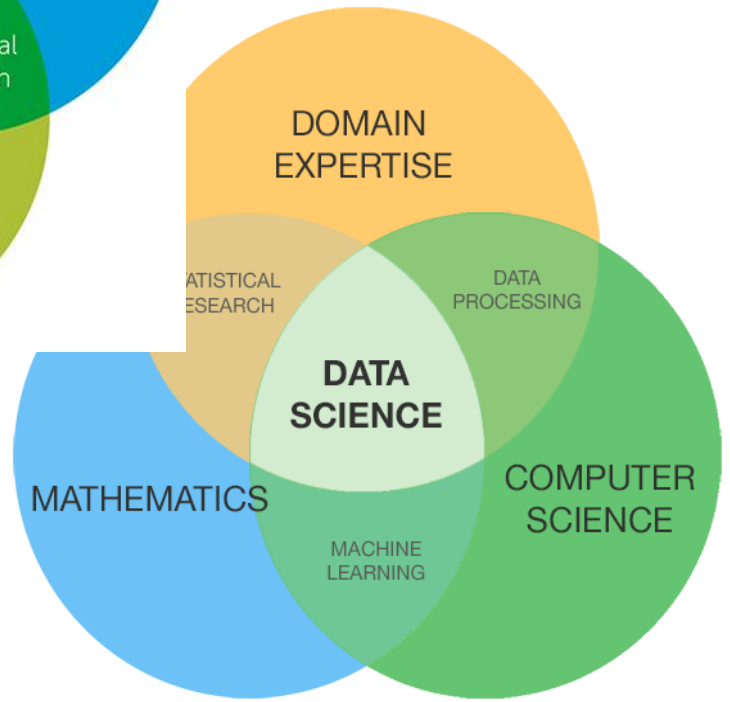
— Source: Nir Kaldero
Head of Data Science at Galvanize

Descriptions of Data Science

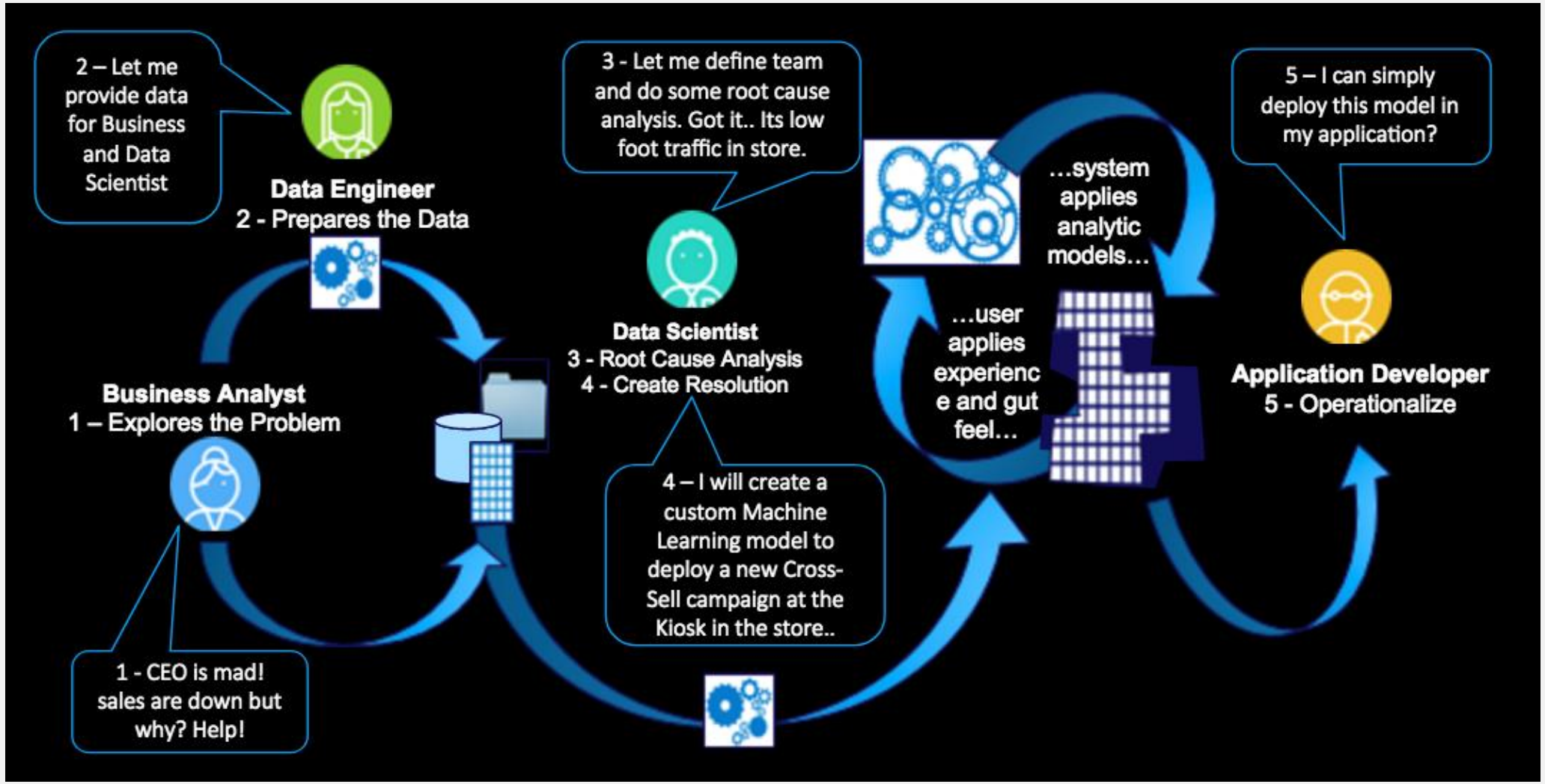
- Intersections of Three Domains
  - Each a Robust Discipline
  - Fundamentally Different
- Interdisciplinary Approach
- Many Variations Evolving
- DS Similar to Sports



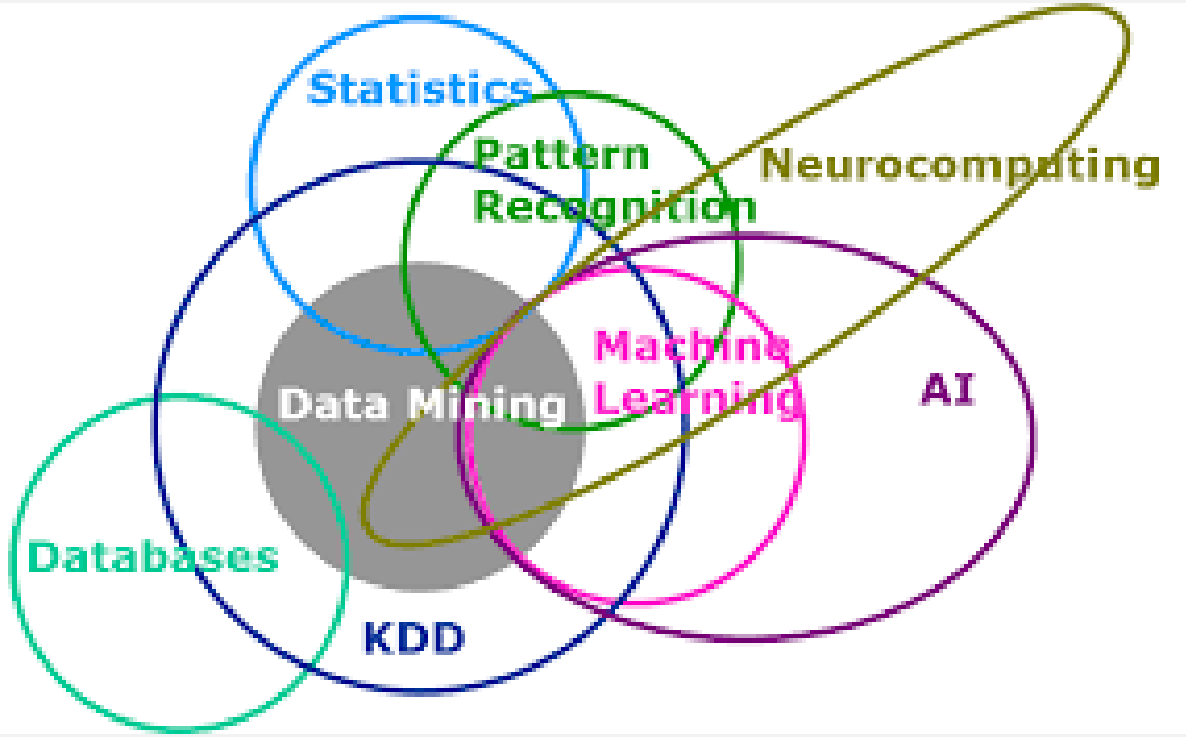


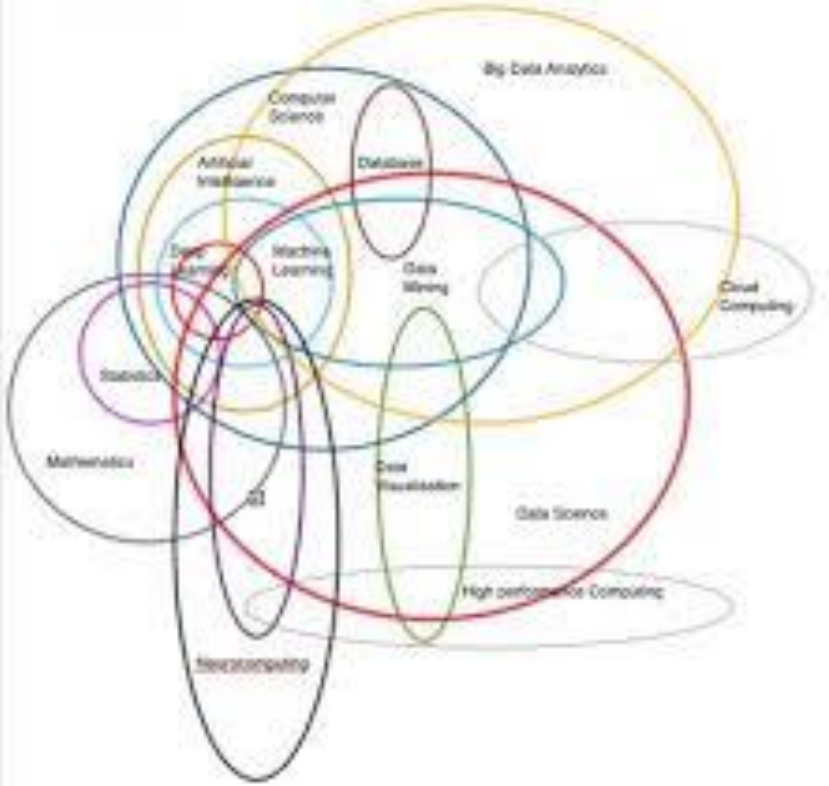


Example Team Flow



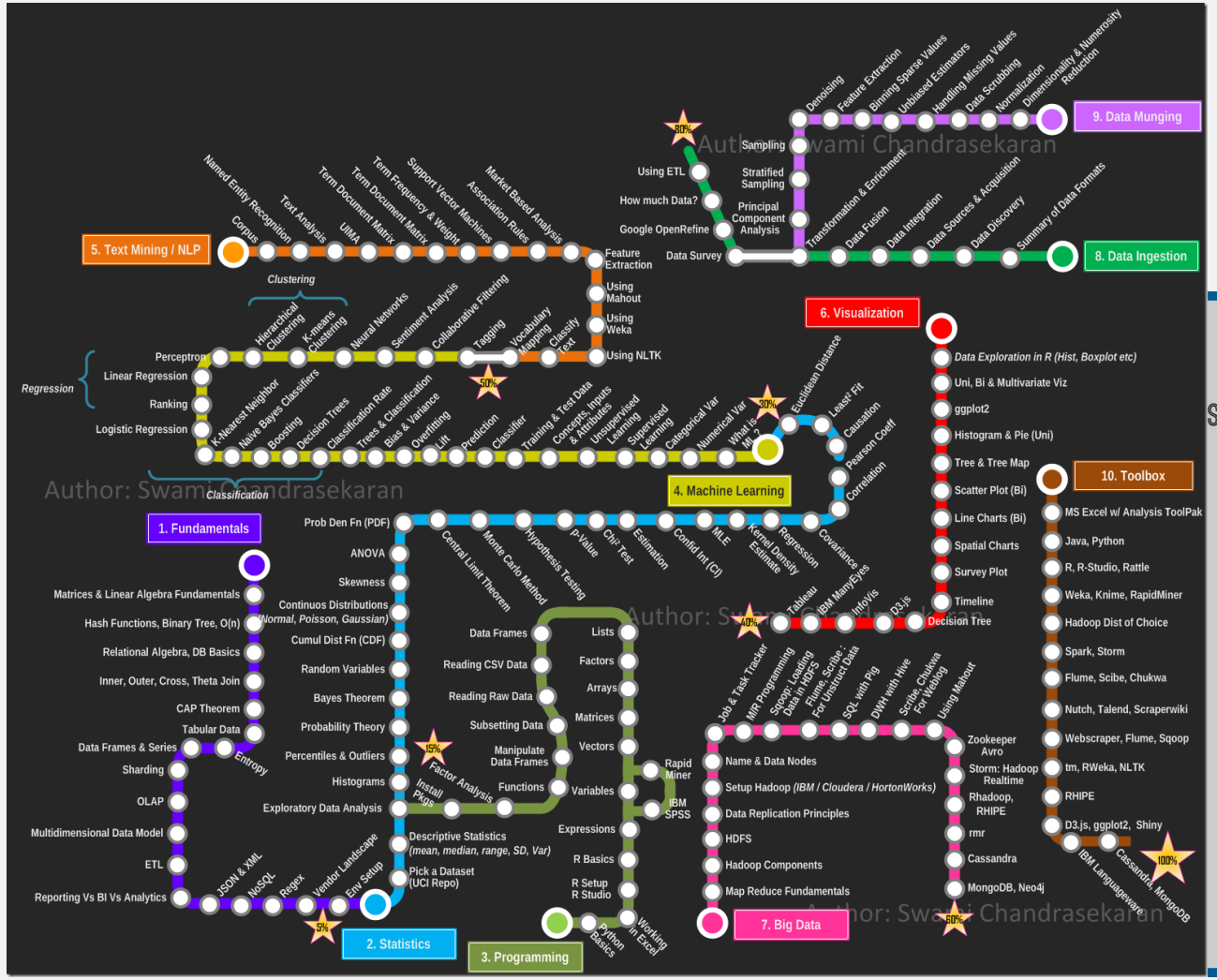
Data Science Complexity

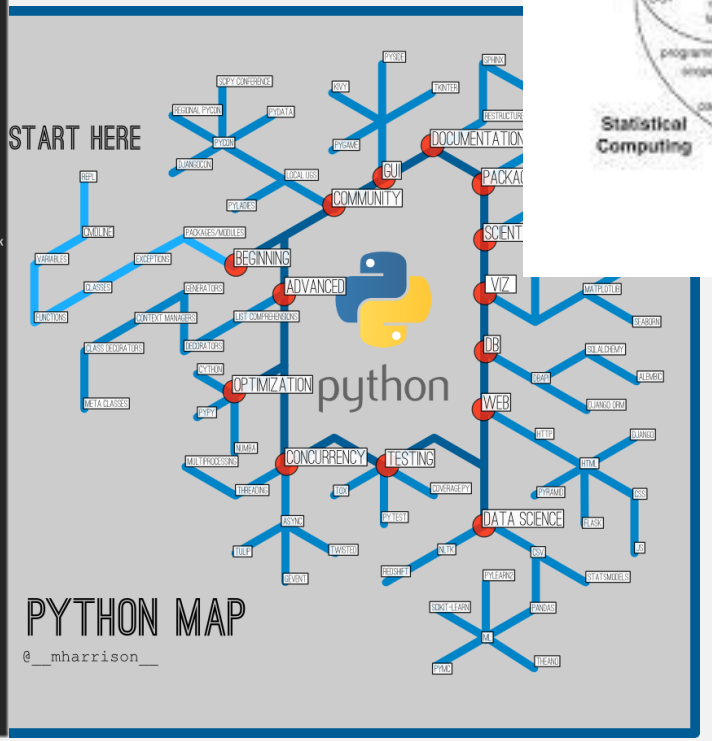


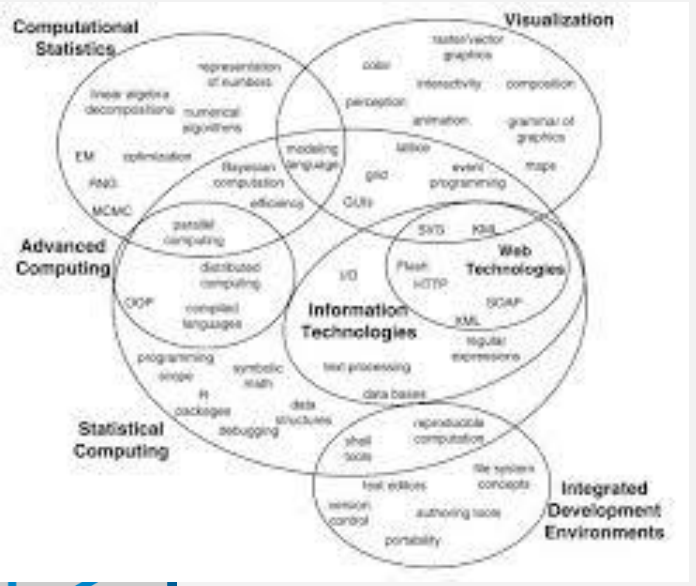


BI

Data Science Complexity







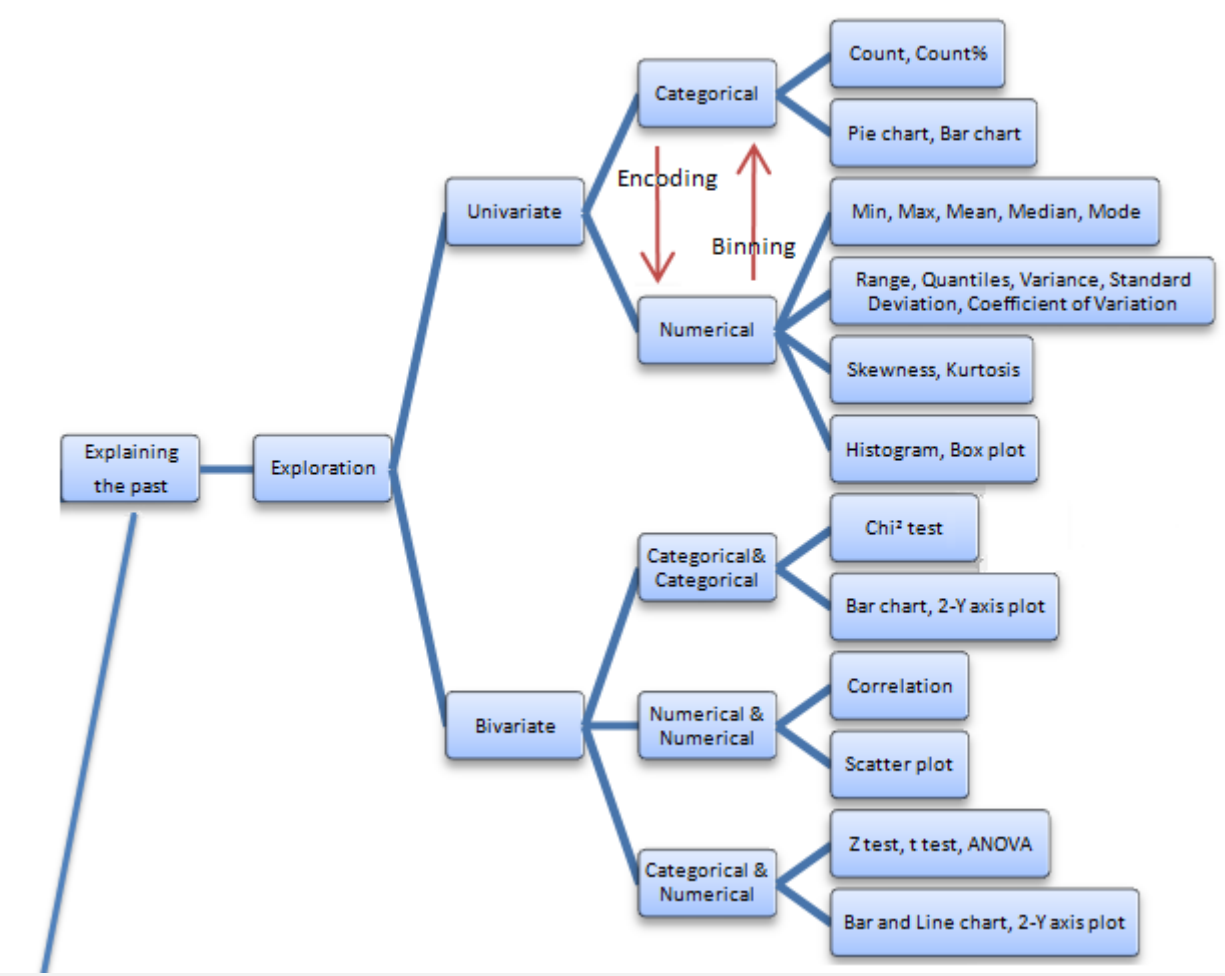
Statistical Methods Versus Machine Learning

- Statistical Modeling
  - Common Pre-ML Approach
  - Generalized Linear Models (GLM)
  - Coefficients Support Natural Interpretation
- Machine Learning Models
  - Flexibility to Handle Non-Linearity
  - No Assumptions to Test
- Machine Learning Usually Wins Against GLMs

Model Type	Average Rank
eXtreme Gradient Boosting	3.50
RuleFit Regularized Tree Ensembles	3.68
Random Forest Trees	5.06
Generalized Linear Models	5.11
Support Vector Machine	5.44
Extra Trees	6.39
Gradient Boosting Model	6.40
K-Nearest Neighbors	8.87
Vopal Wabbit	8.93
TensorFlow	9.71
Decision Tree	10.15

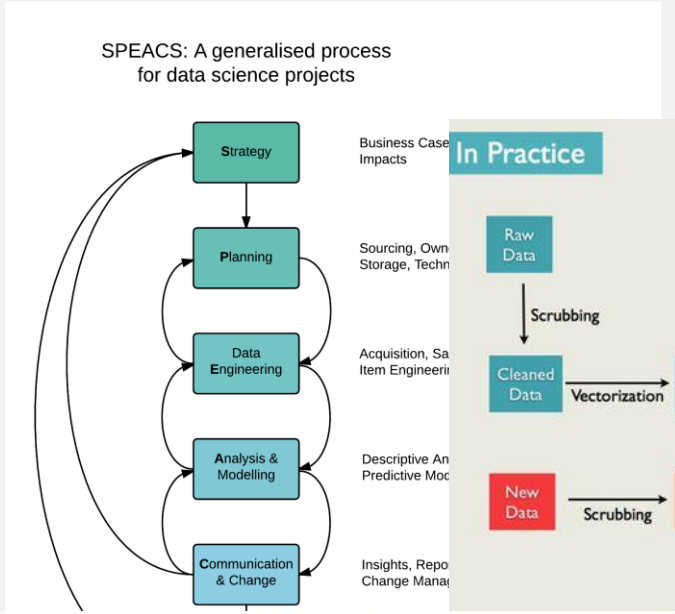
Classic Analytic Decision Tree Branch

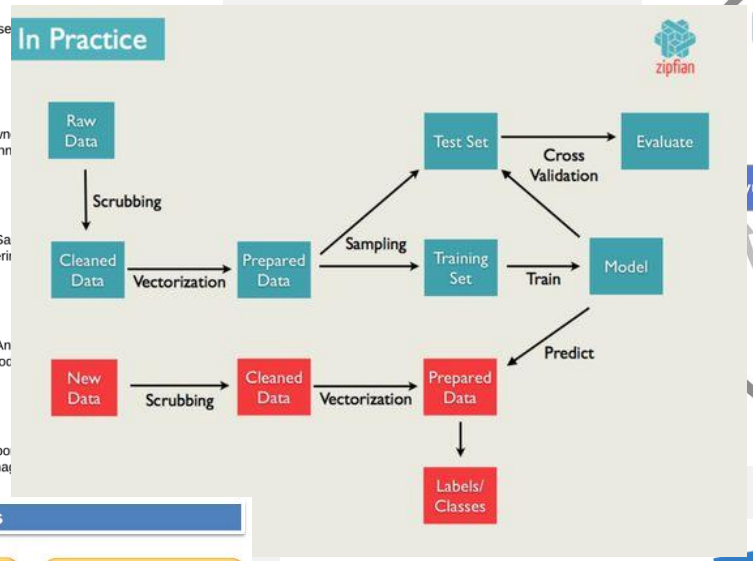
- Leaves Describe Visualizations
- Many Variations for Each Leaf
- Emphasis Varies with Data Set

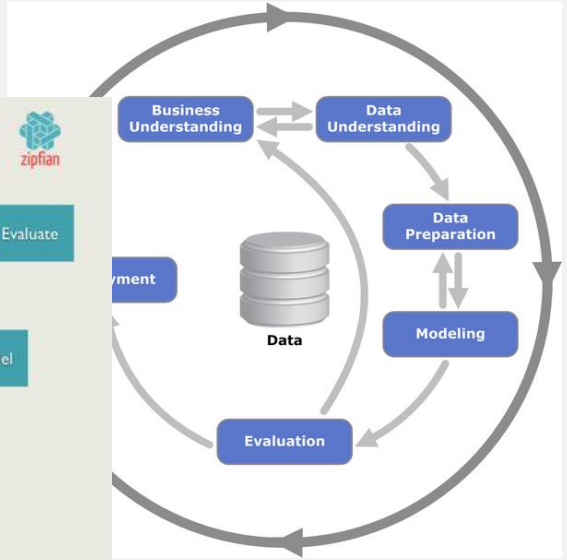


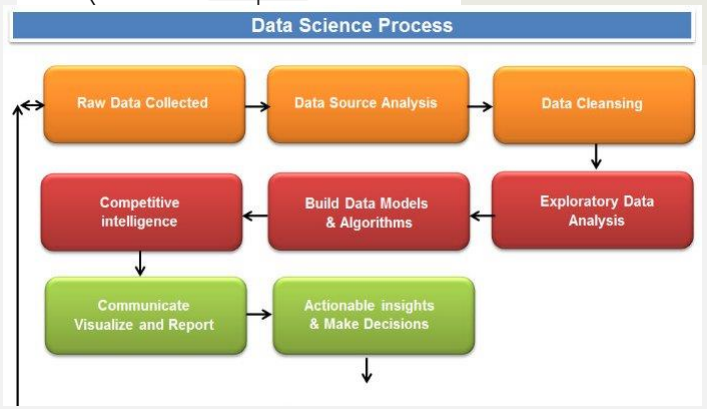
Data Science Processes

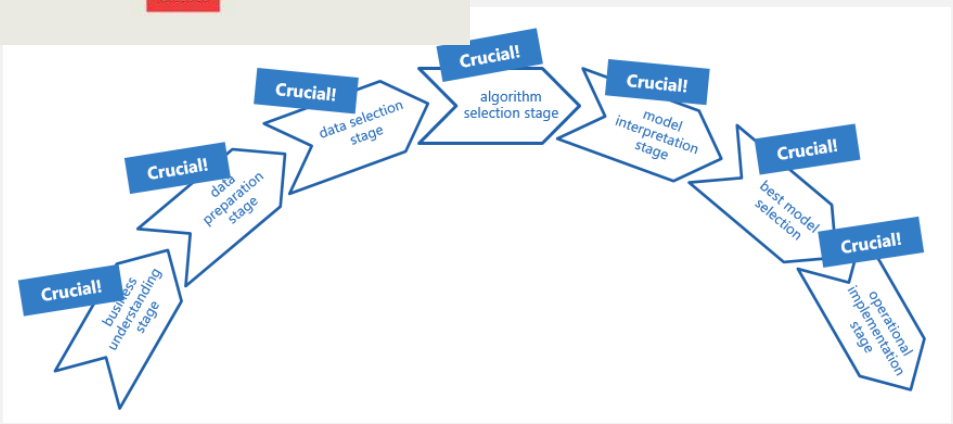
- Numerous Potential Workflows
- Driven By
  - Business Model
  - Skill Capabilities
  - Transparency Requirements
  - Technology Availability



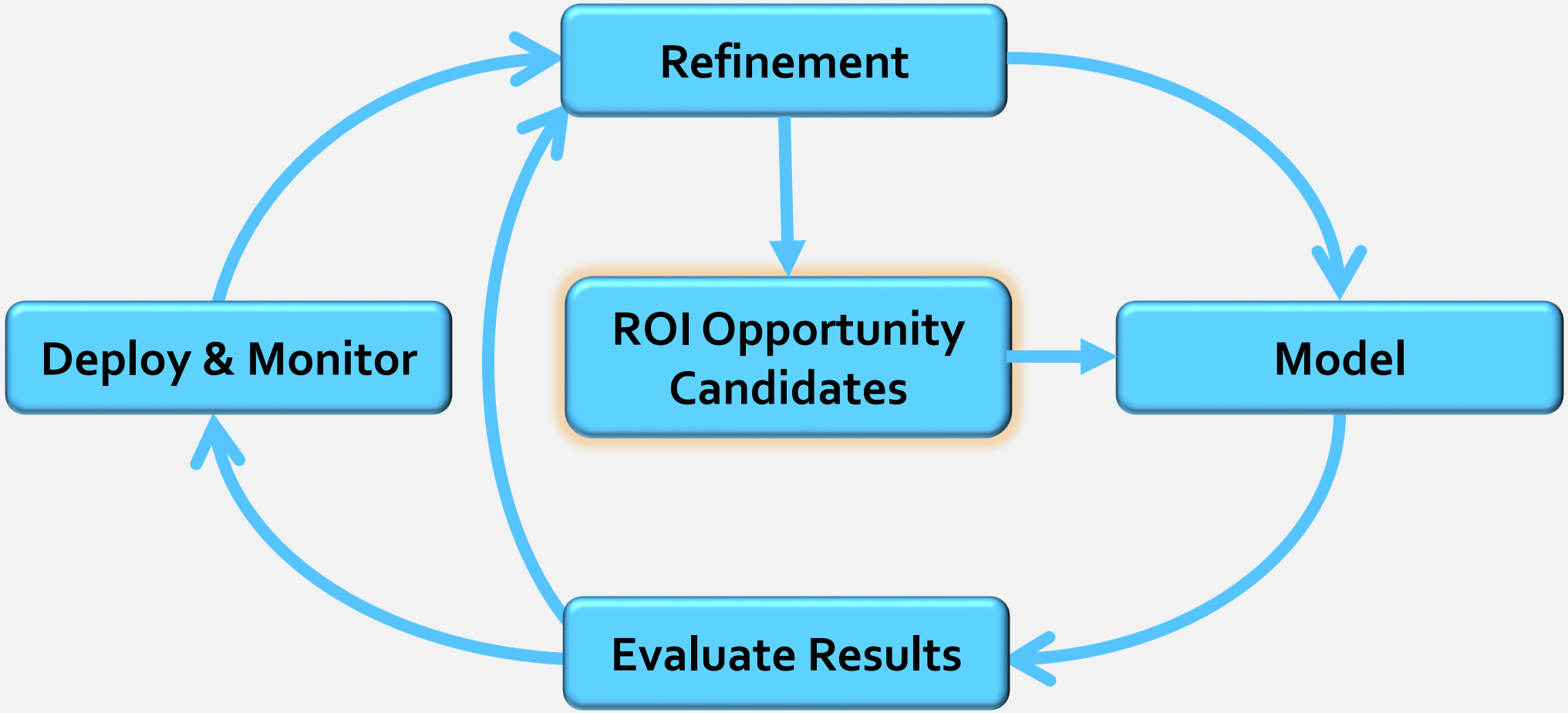








Macro Process



Model Development

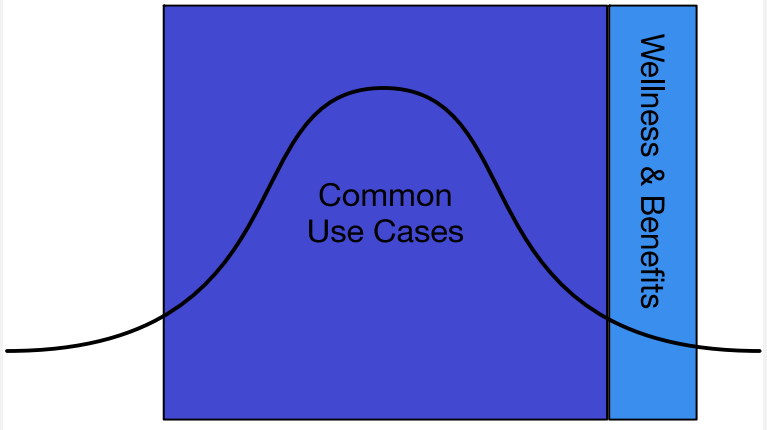
- Iterate Between Data Preparation & Model Creation
- Data Preparation
  - Source Extracts
  - Transformations
  - Data Frame Design
  - Record Creation
  - Feature Design
- Model Creation
  - Modeling Technique Selection & Calibration
  - Algorithm Selection

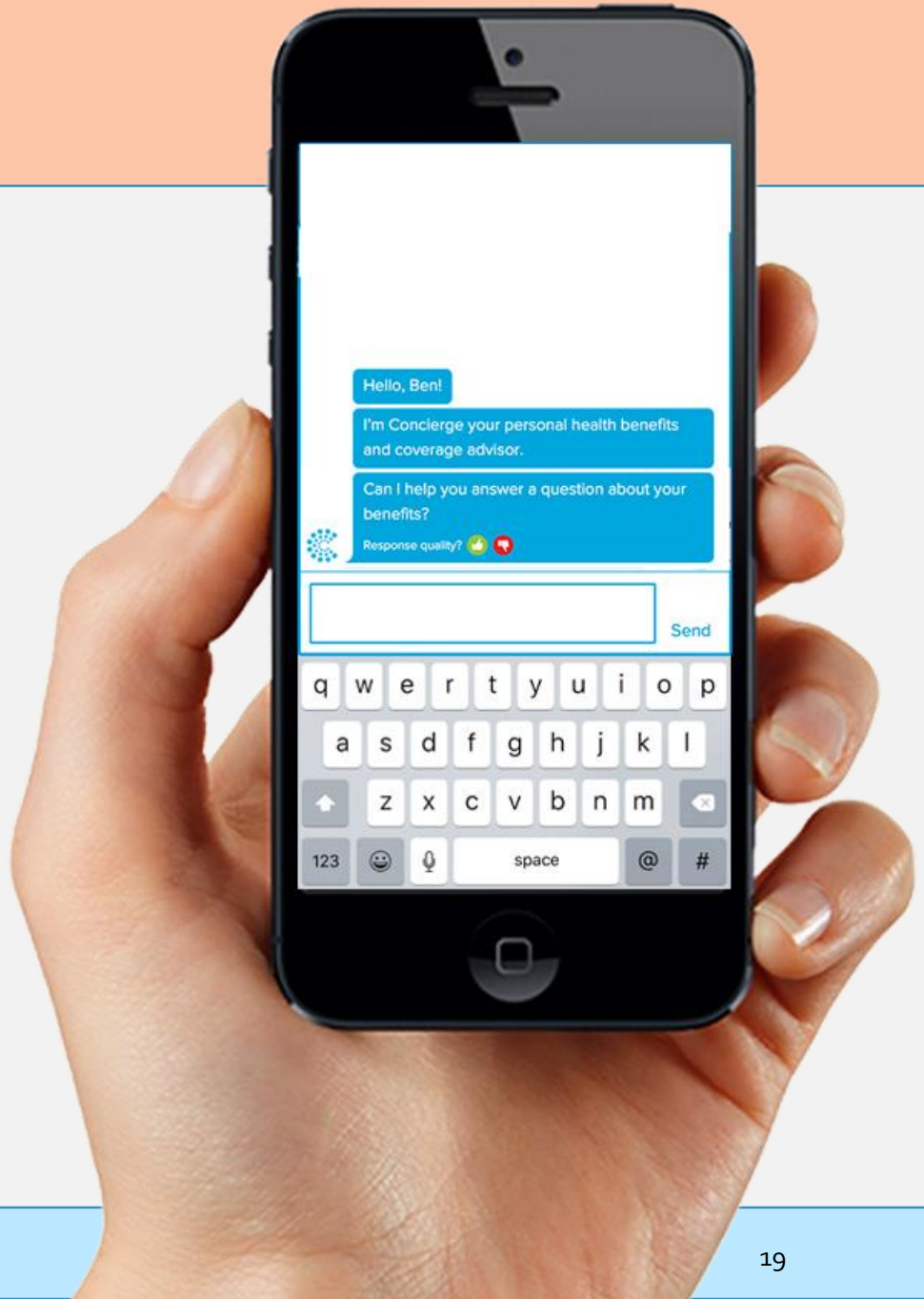
Completing the Process

- Evaluation
  - Confirm Hypotheses
  - Evaluate Confidence
  - Refine/Release Decision
- Deploy & Monitor
  - Produce Report
  - Implement Processes
  - Identify Anomalies
- Refinement
  - Correct for Drift
  - Adjust for Change
  - Incorporate New Data Streams
  - Resolve Edge Cases

Intelligent Mobile Agent

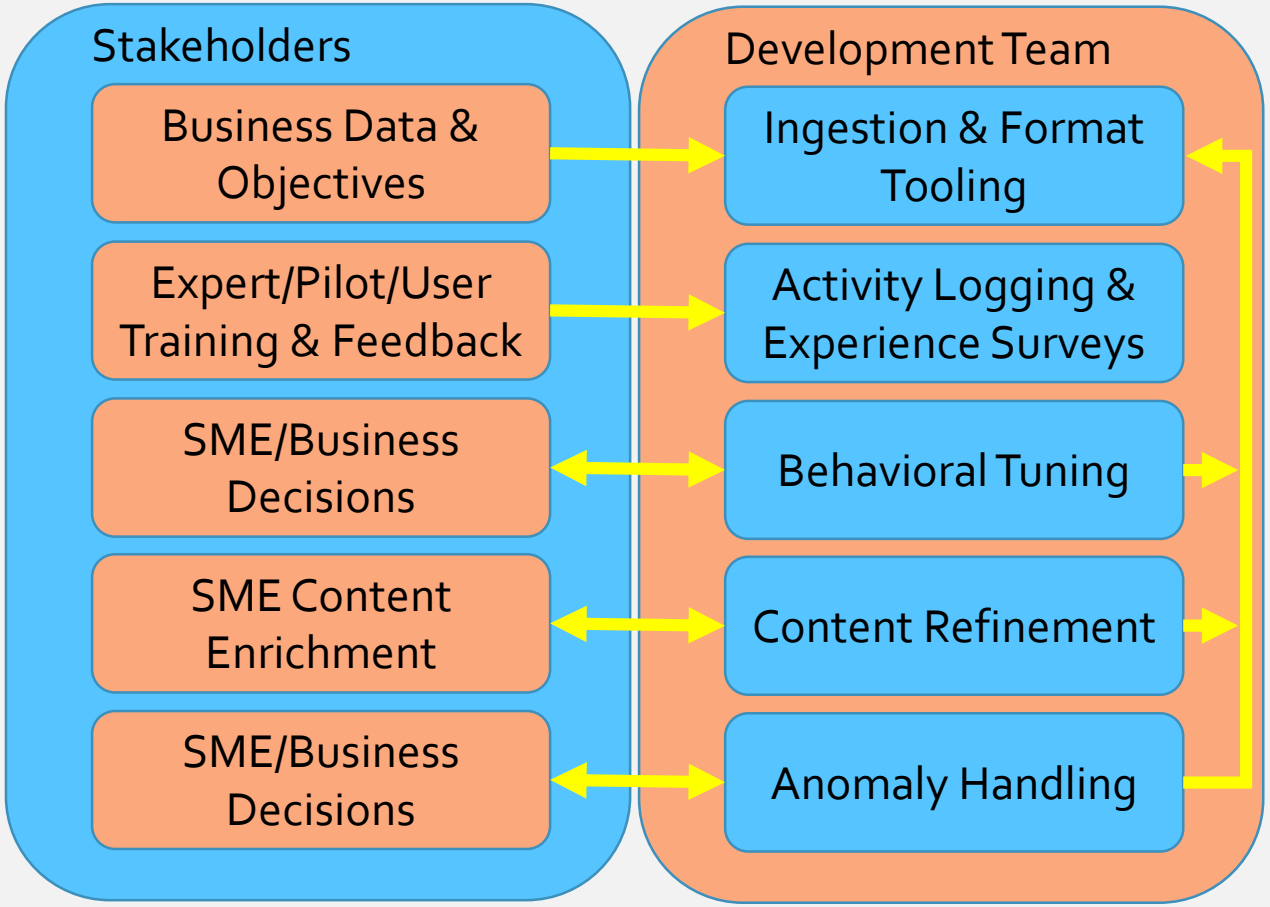
- Based on IBM Watson & TensorFlow
- Natural Language Interface
- Health and Wellness Topics
- Advanced Behaviors
- Varies by Use Case





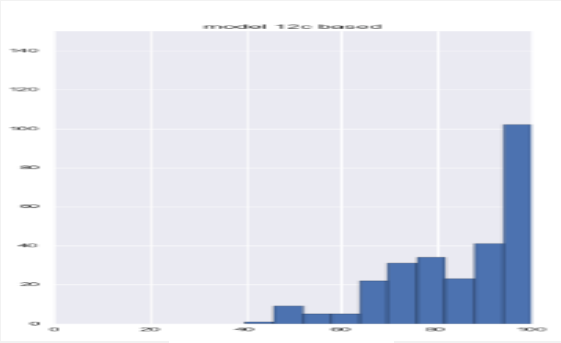
Development Process Flow

- Transform Content Sources
  - Ground Truth (GT)
  - Tone of Voice
- Distribute Activity Log Data to Appropriate Roles & Stakeholders
- Iterate Up Steps
- Continuous Improvement
  - Correct Drift
  - Diminish Edge Cases

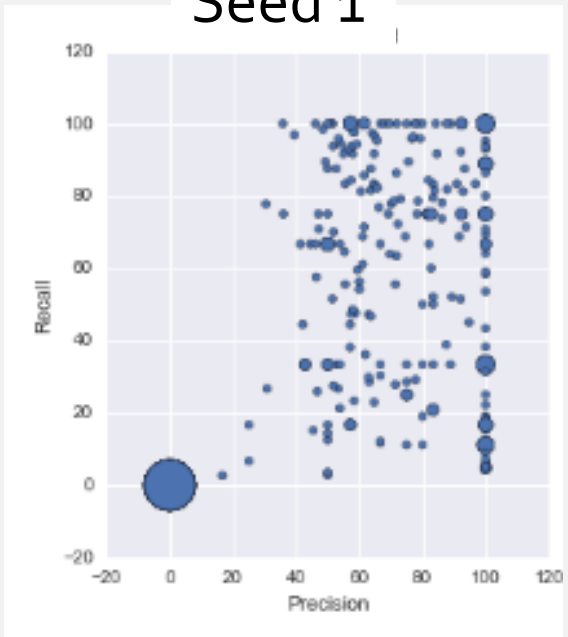


Ground Truth Improvements

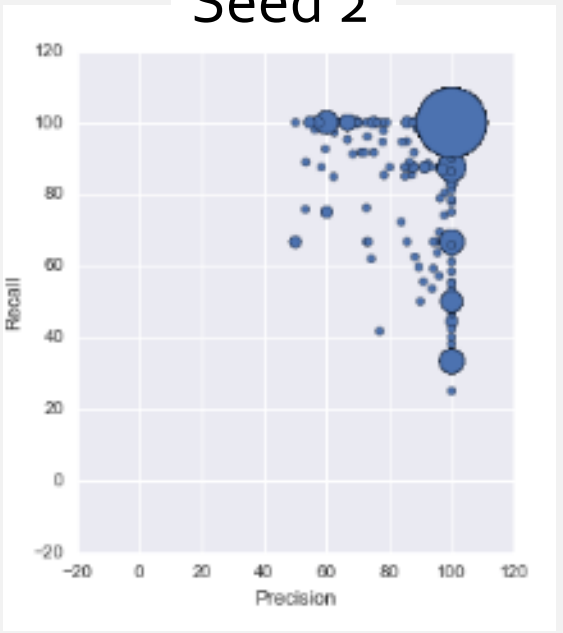
- Move to Synthetic Techniques
  - Seed 1: Manual GT Generation
  - Seed 2: Synthetic GT Generation
- Feature Testing
  - Pluralization
  - Synonyms
  - Spelling
  - 200+ Experiments



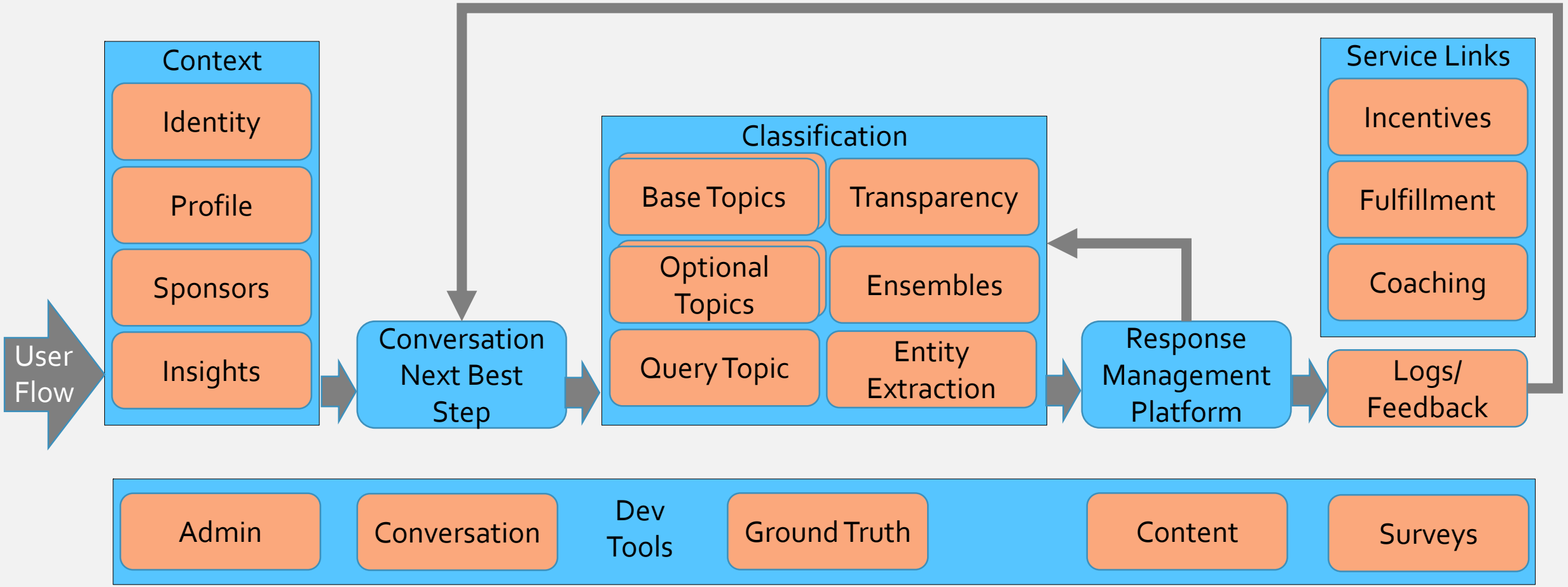
Seed 1



Seed 2



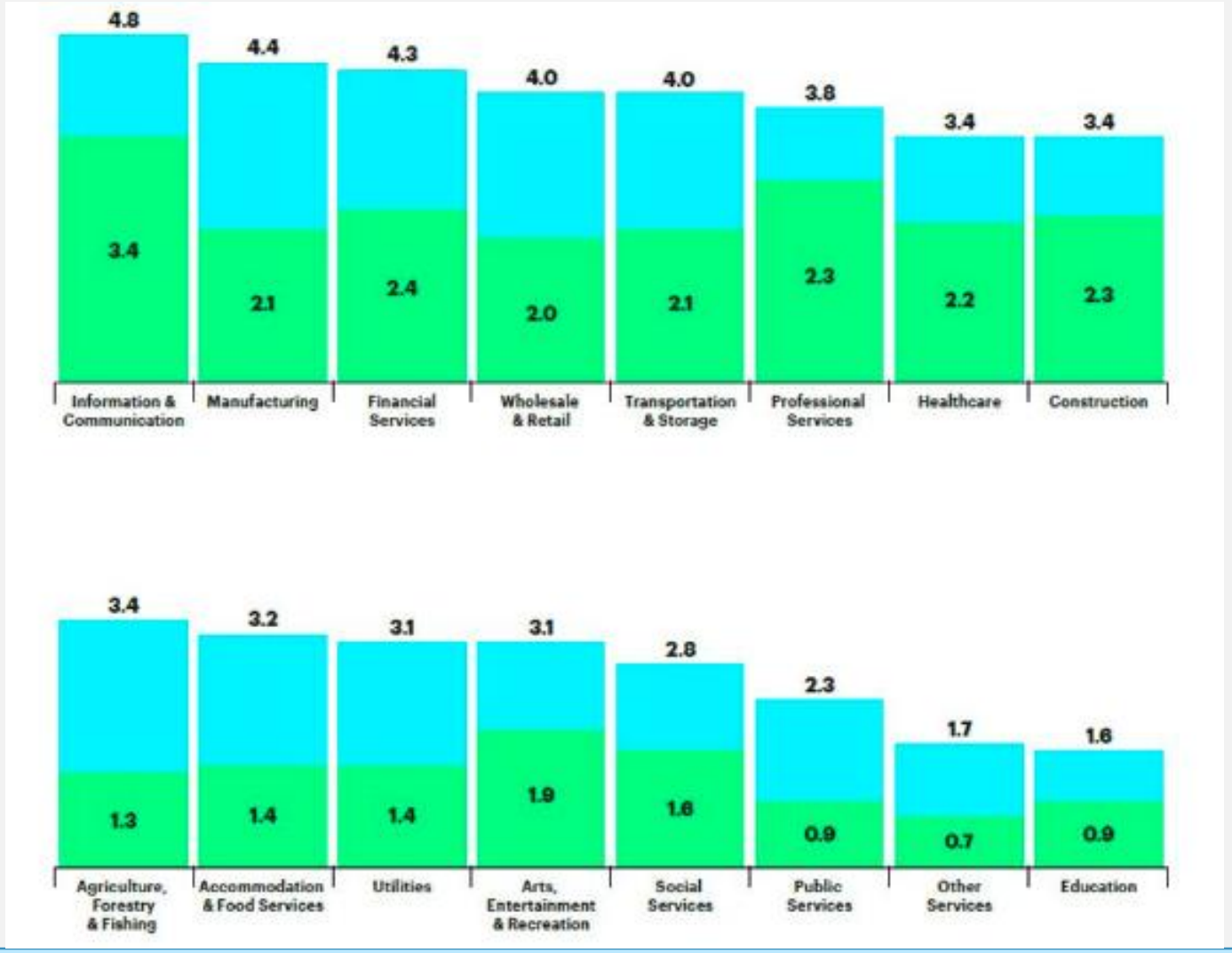
Intelligent Agent Architecture



Sampling of Machine Learning Implementations

- Image Recognition
  - GoogleMaps Street Digits
  - GoogleSearch Image Tab
- Language Recognition
  - Chatbots
  - Assistant Devices
- Detection/Prediction
  - User Behavior/Satisfaction
  - Creditworthiness
  - Churn Prevention
- Process Optimization
  - All Business Functions
  - Robotics
- Jeremy Howard TED Talk:
  The wonderful and terrifying implication of computers that can learn

Sampling of Machine Learning Implementations



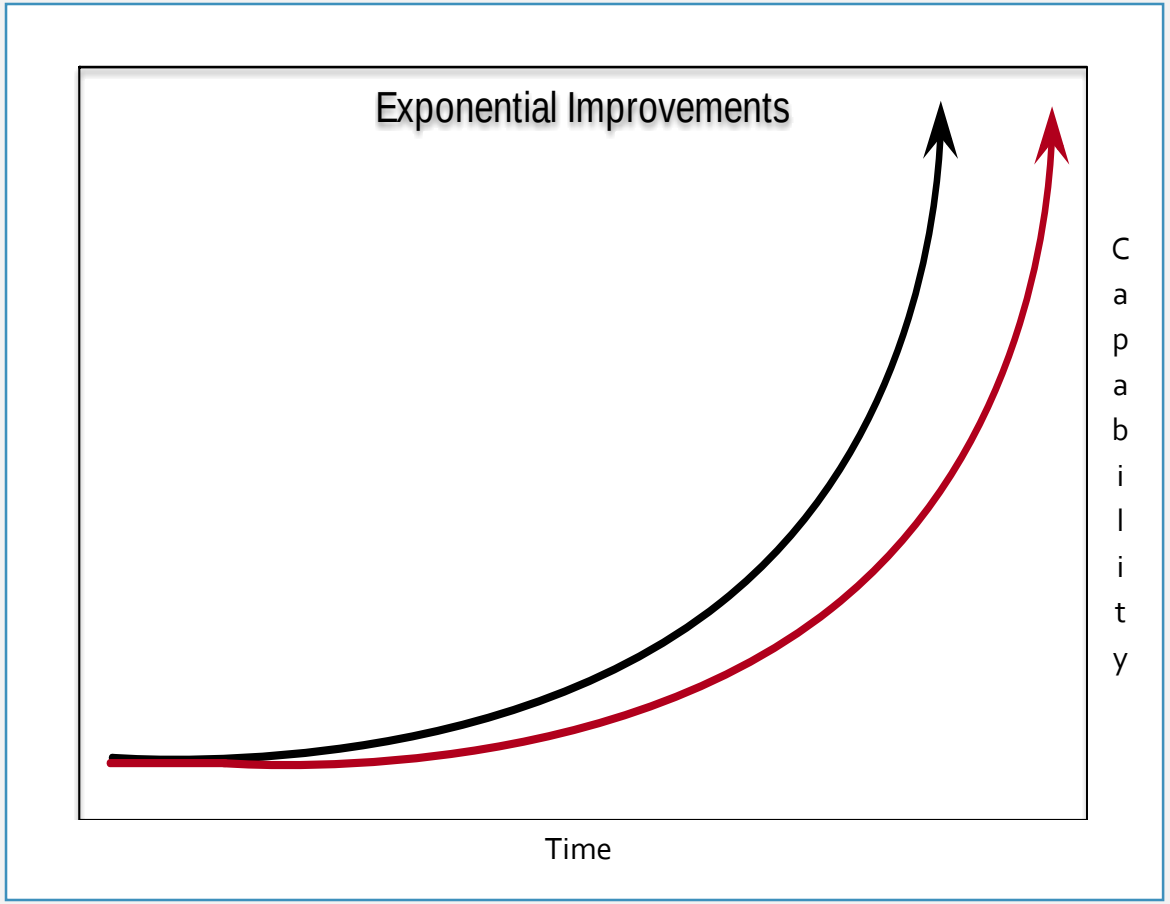
Q & A

Appendix

SUPPORTING INFORMATION

Machine Learning Rate of Change

- Early Start = Insurmountable Lead
- Multiple Accelerators
  - Open Source Community
  - New Effort Builds on Prior Effort
  - Refined Techniques
  - Magnitude Change In Compute
  - Growing Data Fidelity
- Adoption Decelerators
  - Talent Pool
  - Hype Obfuscation



Additional Factors

- Shaping Techniques
  - Corpora Structures
  - Ground Truth
  - Blackbox Validation
  - Classifier Balancing
- Proprietary Technology Tooling
  - Training @Scale
  - Corpus & Ground Truth Management
- Process Techniques
  - Context Boosting
  - Ethnomethodological Linguistic Expertise
- IBM Relationship
  - Broad & Deep Access
  - Infrastructure Tuning
  - Product Team Access & Influence
  - Early Technology Access
  - GitHub Contributions

Machine Learning Transparency

- Documentation
- Accuracy
- Sensitivity
- Input Impact on Outputs
- Use Case Specific Explanation