Data Science

Chapter 1: Introduction to Big Data

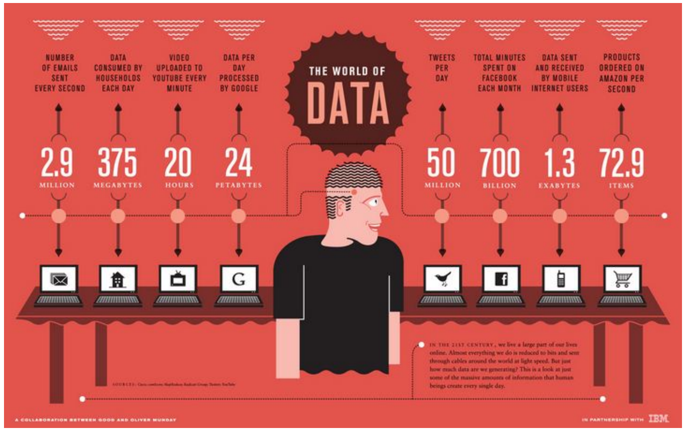
Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

How Much Data Do We have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs
  (100 MB/s)


THE WORLD OF DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
SOURCES: Cisco, comScore, MapReduce, Radicati Group, Twitter, YouTube
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

Big Data Definition

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

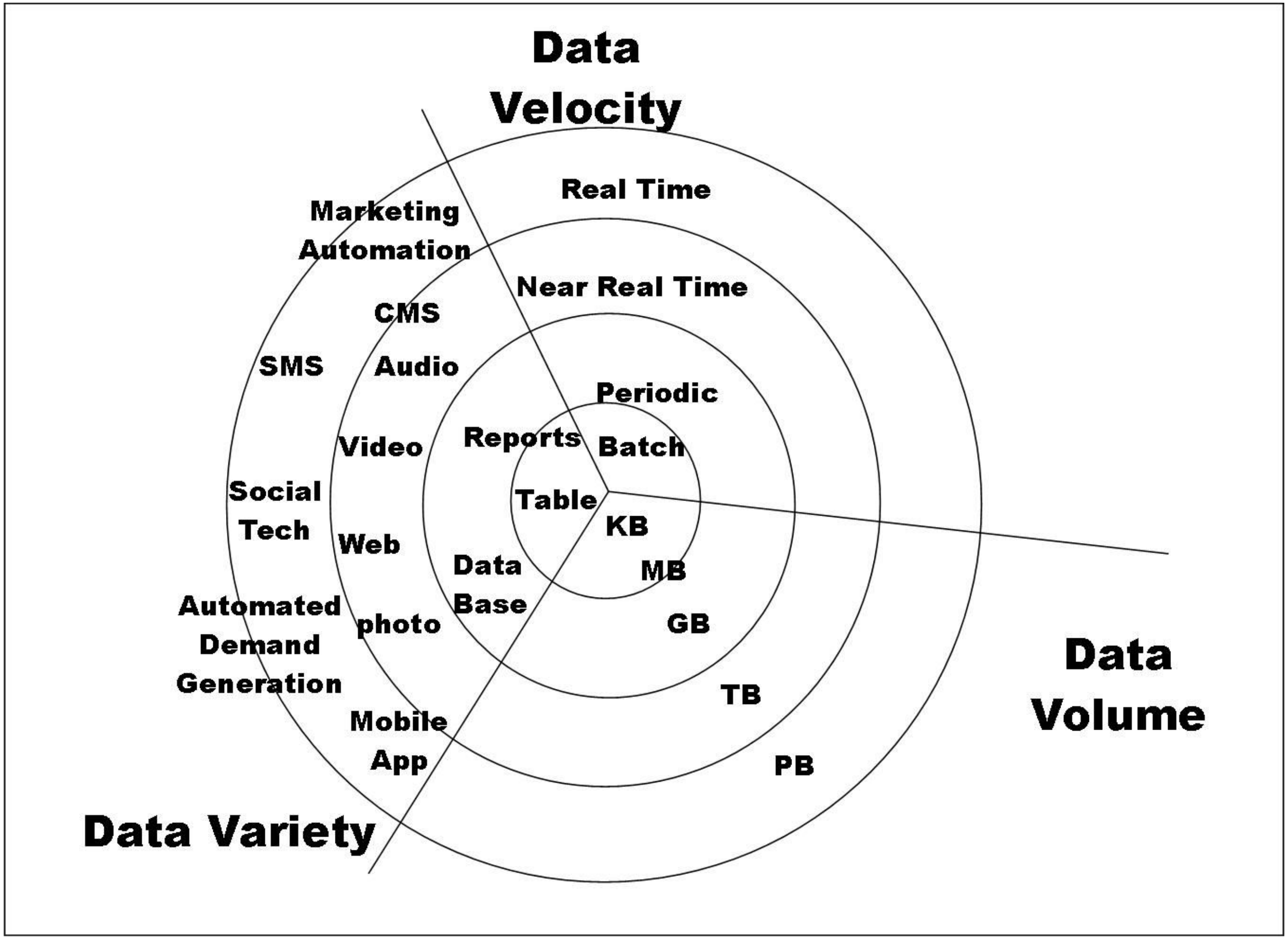
Big Data Overview

- Industries that gather and exploit data
  - Credit card companies monitor purchase
    - Good at identifying fraudulent purchases
  - Mobile phone companies analyze calling patterns – e.g., even on rival networks
    - Look for customers might switch providers
  - For social networks data is primary product
    - Intrinsic value increases as data grows

Attributes Defining Big Data Characteristics

- Huge volume of data
  - Not just thousands/millions, but billions of items
- Complexity of data types and structures
  - Varity of sources, formats, structures
- Speed of new data creation and grow
  - High velocity, rapid ingestion, fast analysis

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
Data
Base
KB
MB
GB
TB
PB
Automated
Demand
Generation
photo
Mobile
App
Data
Volume
Data Variety

Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

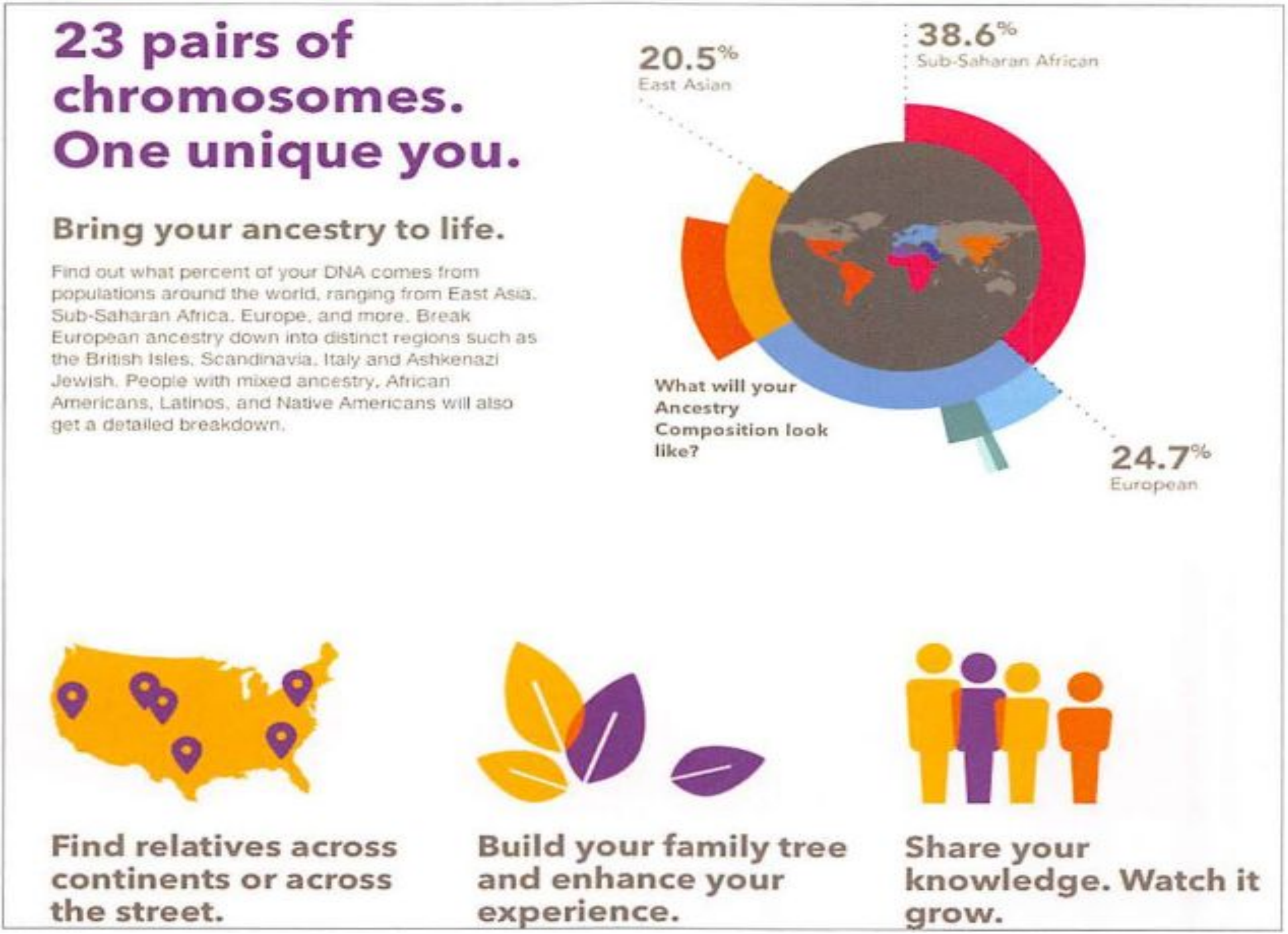
Sources of Big Data Deluge

- Mobile sensors – GPS, accelerometer, etc.
- Social media – 700 Facebook updates/sec in2012
- Video surveillance – street cameras, stores, etc.
- Video rendering – processing video for display
- Smart grids – gather and act on information
- Geophysical exploration – oil, gas, etc.
- Medical imaging – reveals internal body structures
- Gene sequencing – more prevalent, less expensive, healthcare would like to predict personal illnesses

What's Driving Data Deluge?

Mobile Sensors
Social Media
Video Surveillance
Video Rendering
Smart Grids
Geophysical Exploration
Medical Imaging
Gene Sequencing

Example:
Genotyping from 23andme.com



23 pairs of chromosomes. One unique you.

Bring your ancestry to life.

Find out what percent of your DNA comes from populations around the world, ranging from East Asia, Sub-Saharan Africa, Europe, and more. Break European ancestry down into distinct regions such as the British Isles, Scandinavia, Italy and Ashkenazi Jewish. People with mixed ancestry, African Americans, Latinos, and Native Americans will also get a detailed breakdown.

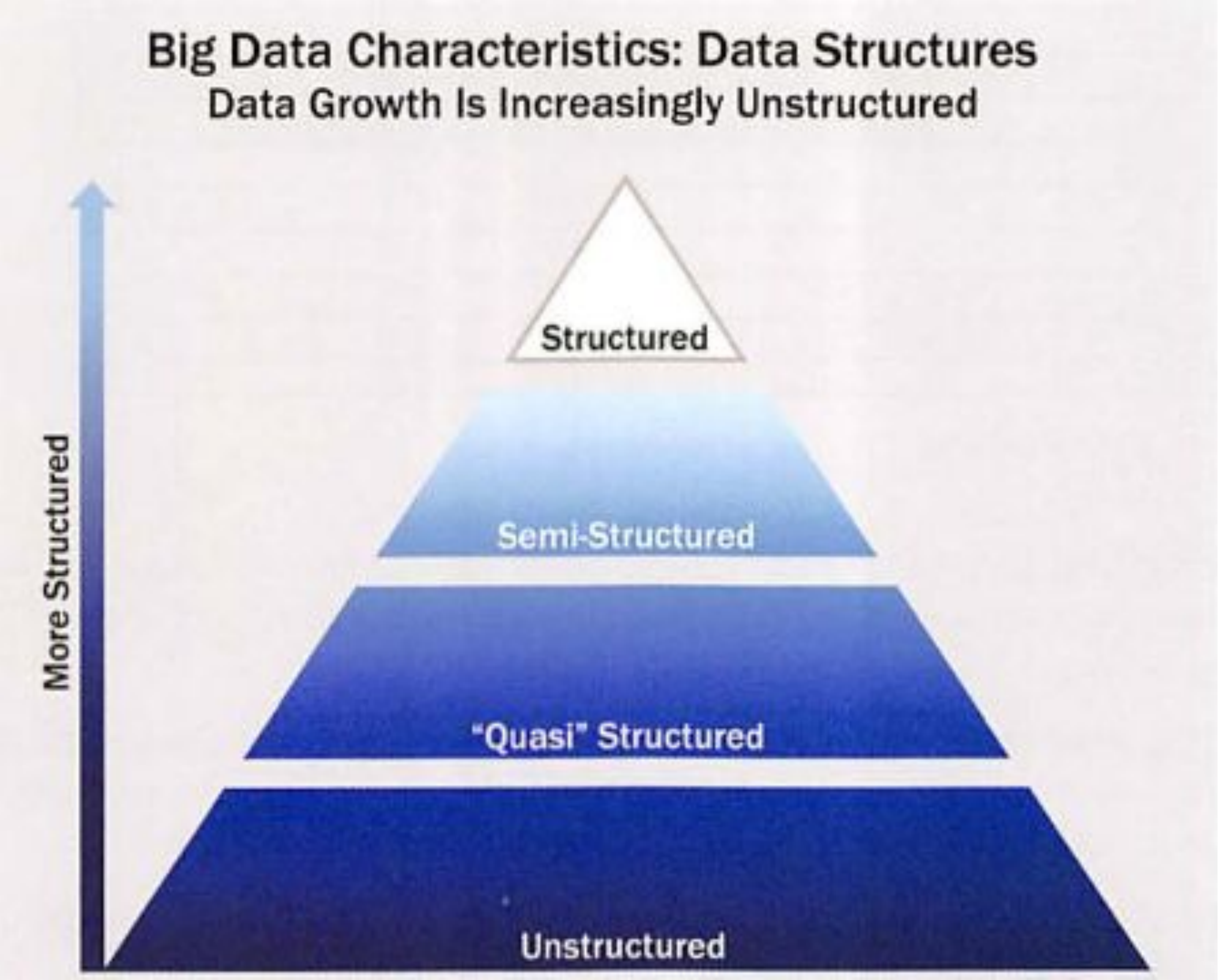
What will your Ancestry Composition look like?

Find relatives across continents or across the street.

Build your family tree and enhance your experience.

Share your knowledge. Watch it grow.

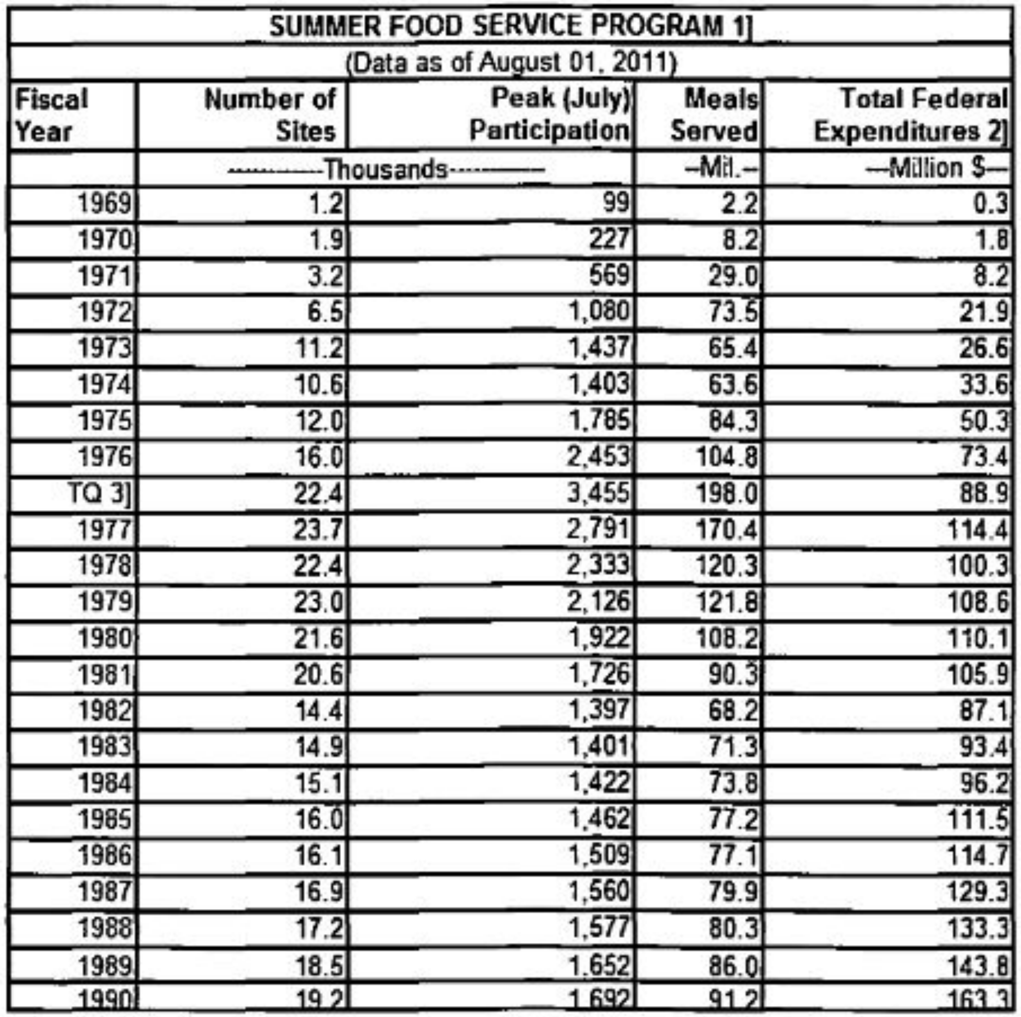
Characteristics of Big Data



Big Data Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

- Structured – defined data type, format, structure
  - Transactional data, OLAP cubes, RDBMS, CSV files, spreadsheets
- Semi-structured
  - Text data with discernable patterns – e.g., XML data
- Quasi-structured
  - Text data with erratic data formats – e.g., clickstream data
- Unstructured
  - Data with no inherent structure – text docs, PDF’s, images, video

Example of Structured Data

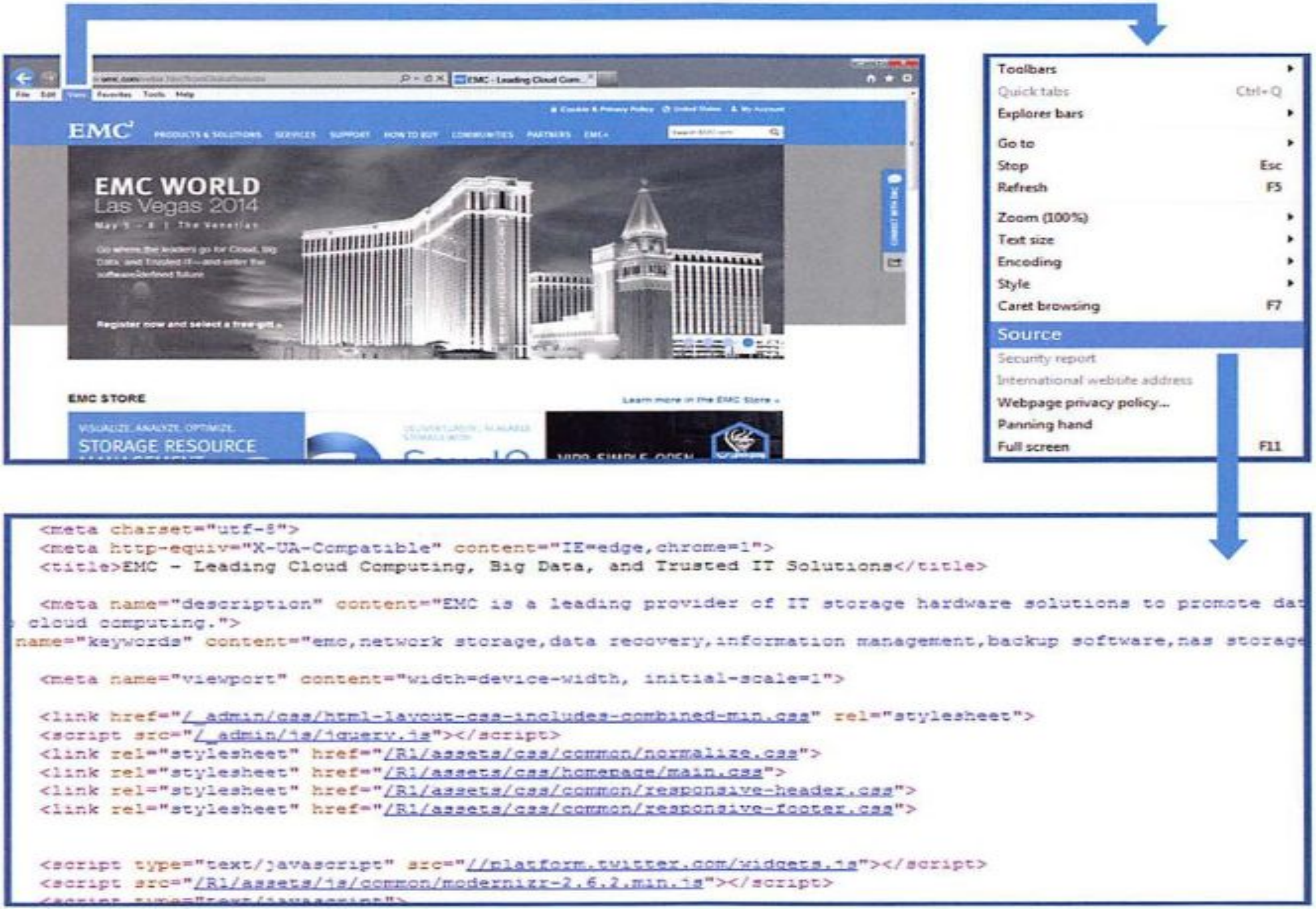


SUMMER FOOD SERVICE PROGRAM 1]

(Data as of August 01, 2011)

Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	--Million $--
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

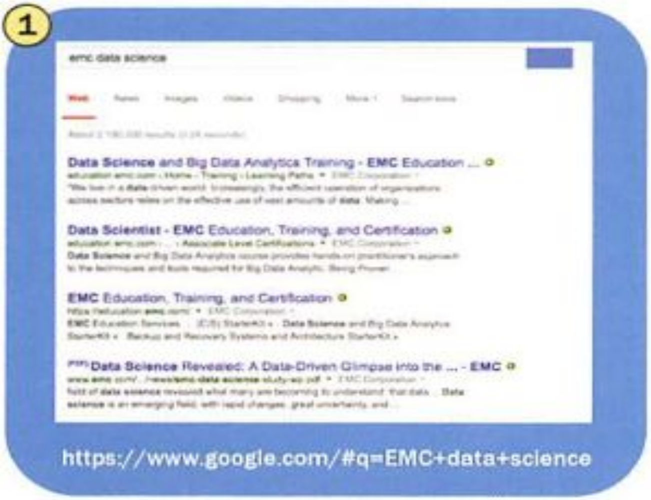
Example of Semi-Structured Data



```
  <meta charset="utf-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
  <title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

  <meta name="description" content="EMC is a leading provider of IT storage hardware solutions to promote dat
 cloud computing.">
name="keywords" content="emc,network storage,data recovery,information management,backup software,nas storage

  <meta name="viewport" content="width=device-width, initial-scale=1">

  <link href="/_admin/css/html-layout-css-includes-combined-min.css" rel="stylesheet">
  <script src="/_admin/js/jquery.js"></script>
  <link rel="stylesheet" href="/R1/assets/css/common/normalize.css">
  <link rel="stylesheet" href="/R1/assets/css/homepage/main.css">
  <link rel="stylesheet" href="/R1/assets/css/common/responsive-header.css">
  <link rel="stylesheet" href="/R1/assets/css/common/responsive-footer.css">


  <script type="text/javascript" src="//platform.twitter.com/widgets.js"></script>
  <script src="/R1/assets/js/common/modernizr-2.6.2.min.js"></script>
  <script type="text/javascript">
```

Example of Quasi-Structured Data

visiting 3 websites adds 3 URLs to user's log files



emc data science

Data Science and Big Data Analytics Training - EMC Education ...

"We live in a data driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Making ...

Data Scientist - EMC Education, Training, and Certification

Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytic. Being Proven ...

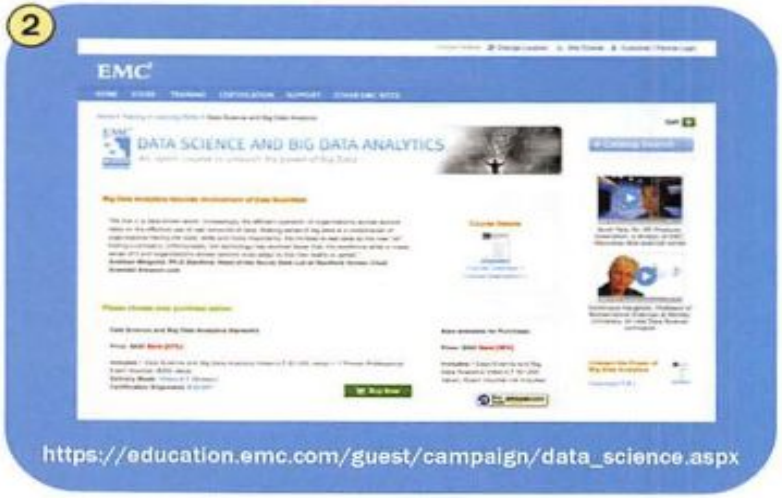
EMC Education, Training, and Certification

EMC Education Services ... Data Science and Big Data Analytics StarterKit » Backup and Recovery Systems and Architecture StarterKit »

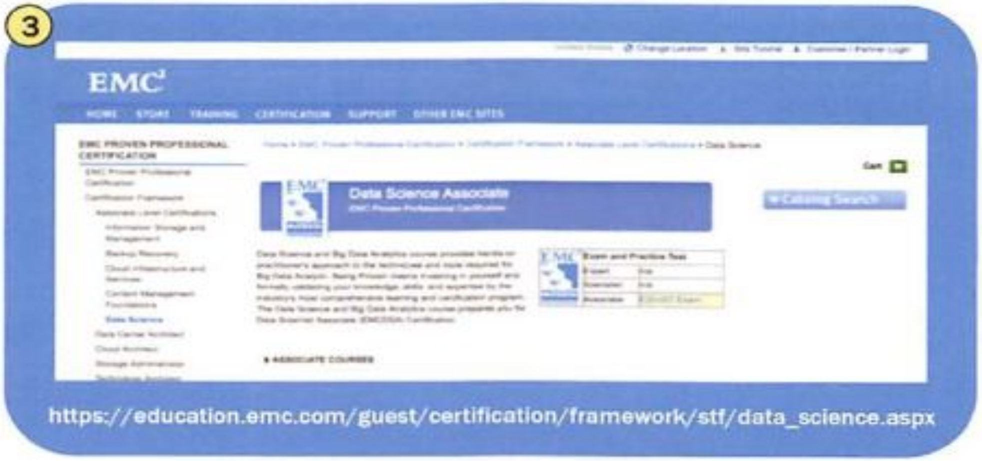
[PDF] Data Science Revealed: A Data-Driven Glimpse into the ... - EMC

field of data science revealed what many are becoming to understand: that data ... Data science is an emerging field, with rapid changes, great uncertainty, and ...

https://www.google.com/#q=EMC+data+science



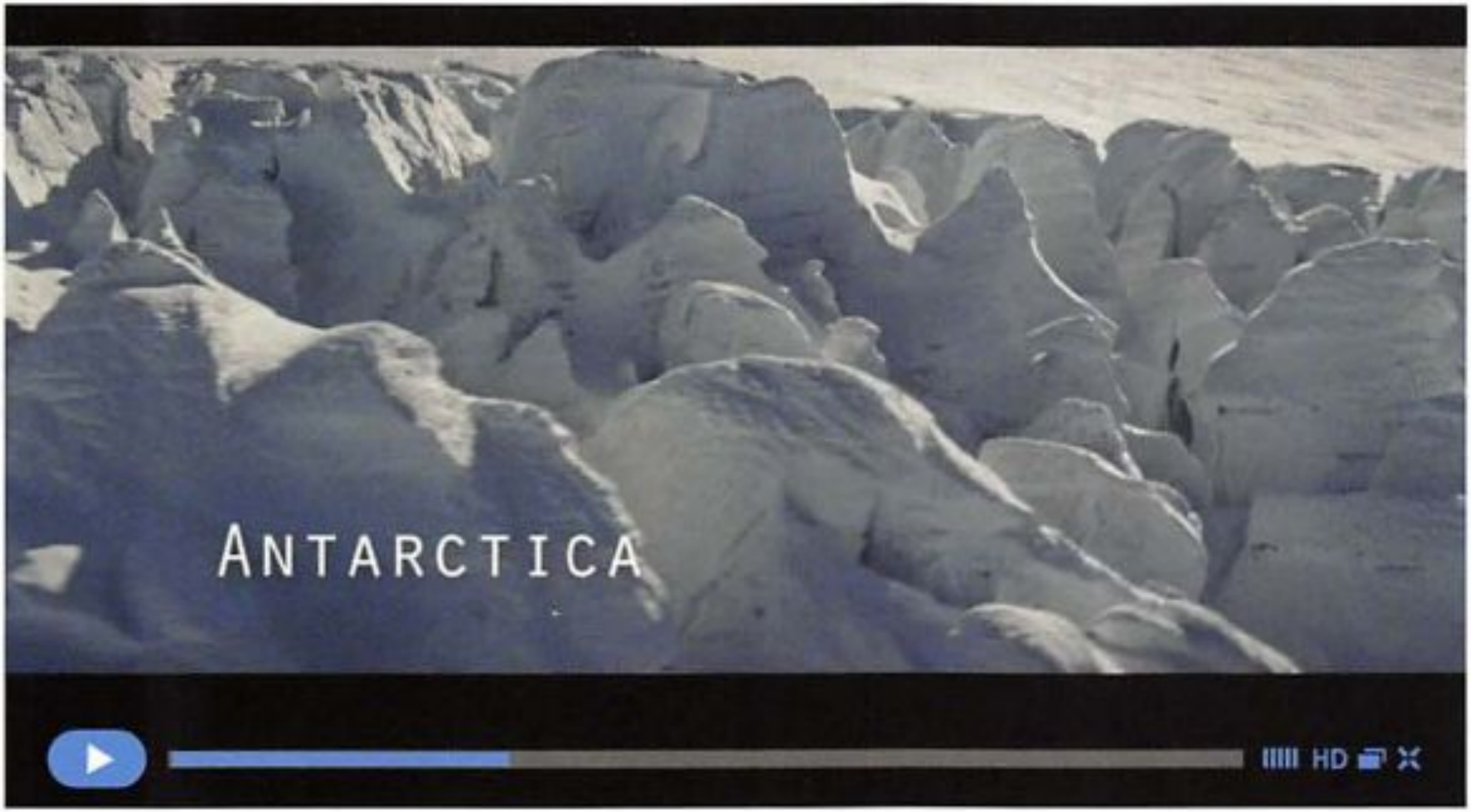
https://education.emc.com/guest/campaign/data_science.aspx



https://education.emc.com/guest/certification/framework/stf/data_science.aspx

Example of Unstructured Data

Video about Antarctica Expedition



Datafication

- Datafication is a process of “taking all aspects of life and turning them into data”
- Examples:

  1. Google’s augmented-reality glasses datafy the gaze.
  2. Twitter datafies stray thoughts.
  3. LinkedIn datafies professional networks
- Once we datafy things, we can transform their purpose and turn the information into new forms of value.

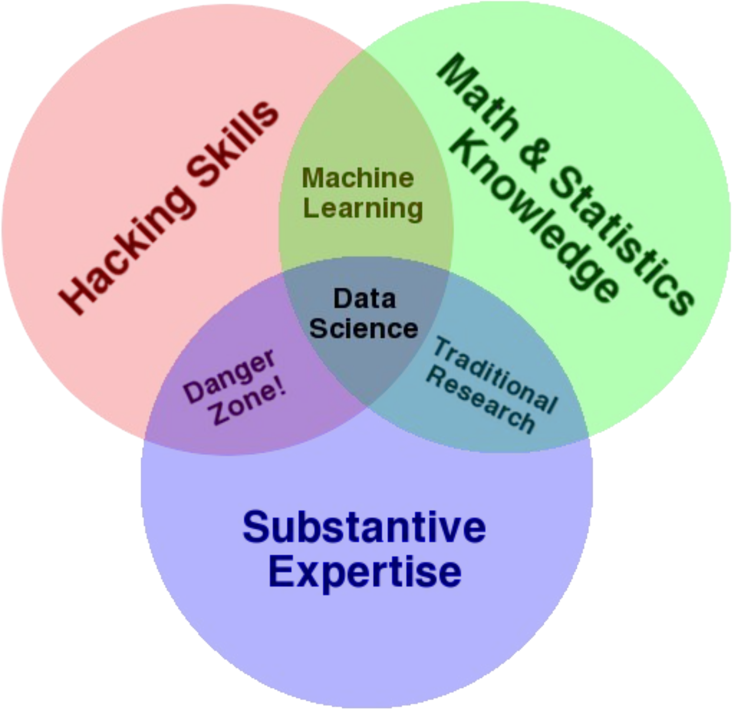
Introduction to Data Science

- Applying advanced statistical tools to existing data to generate new insights is known as “Data Science”.
- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

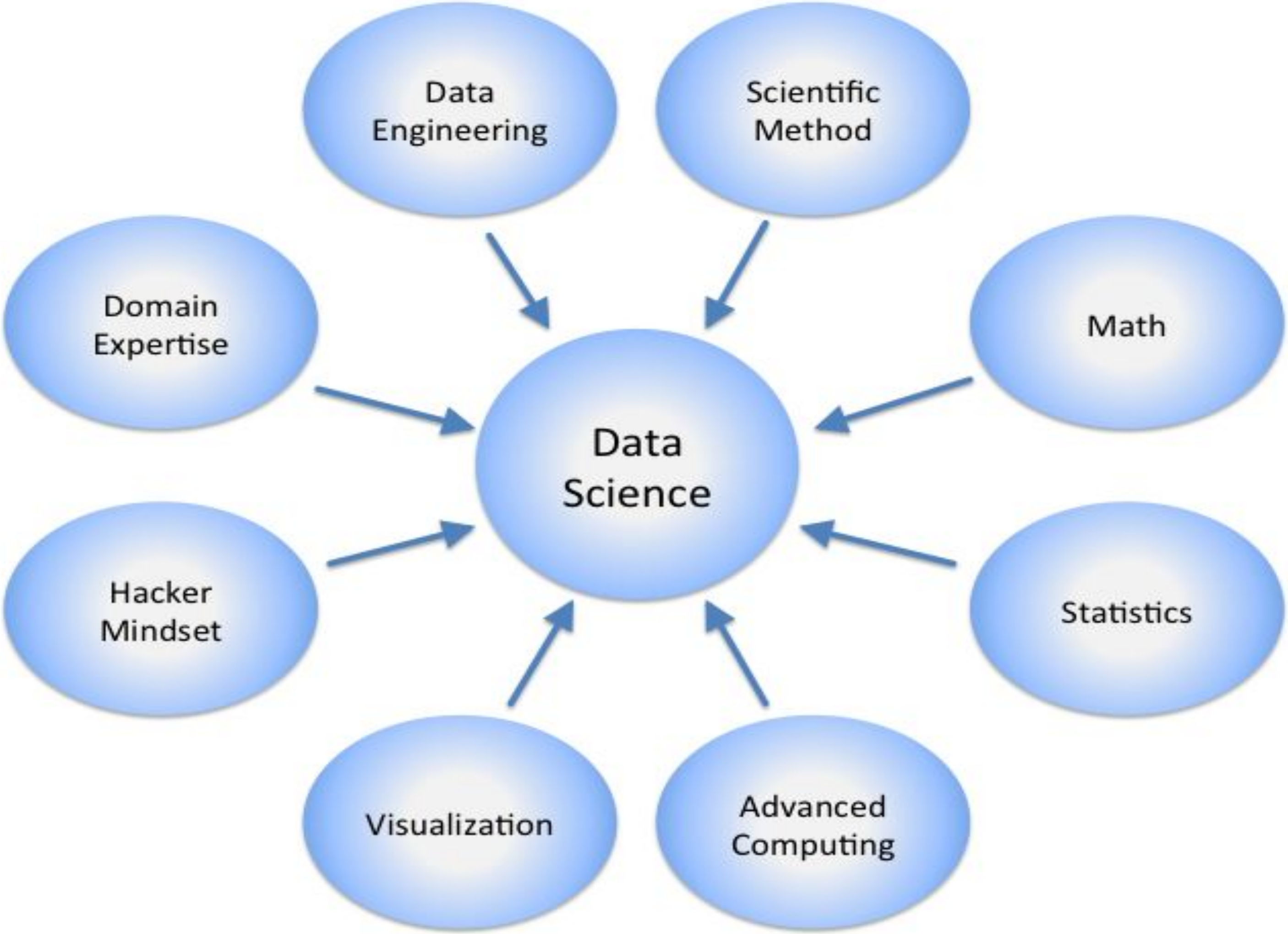
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

Goal of Data Science

- Turn data into data products.
- Visual Definition



- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.



Data Engineering

Scientific Method

Domain Expertise

Math

Data Science

Hacker Mindset

Statistics

Visualization

Advanced Computing

Real Life Examples

- Companies learn your secrets, shopping patterns, and preferences
  - For example, can we know if a woman is pregnant, even if she doesn't want us to know? Target case study
- Data Science and election (2008, 2012)
  - 1 million people installed the Obama Facebook app that gave access to info on "friends"

Data Scientists

- They find stories, extract knowledge. They are not reporters
- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

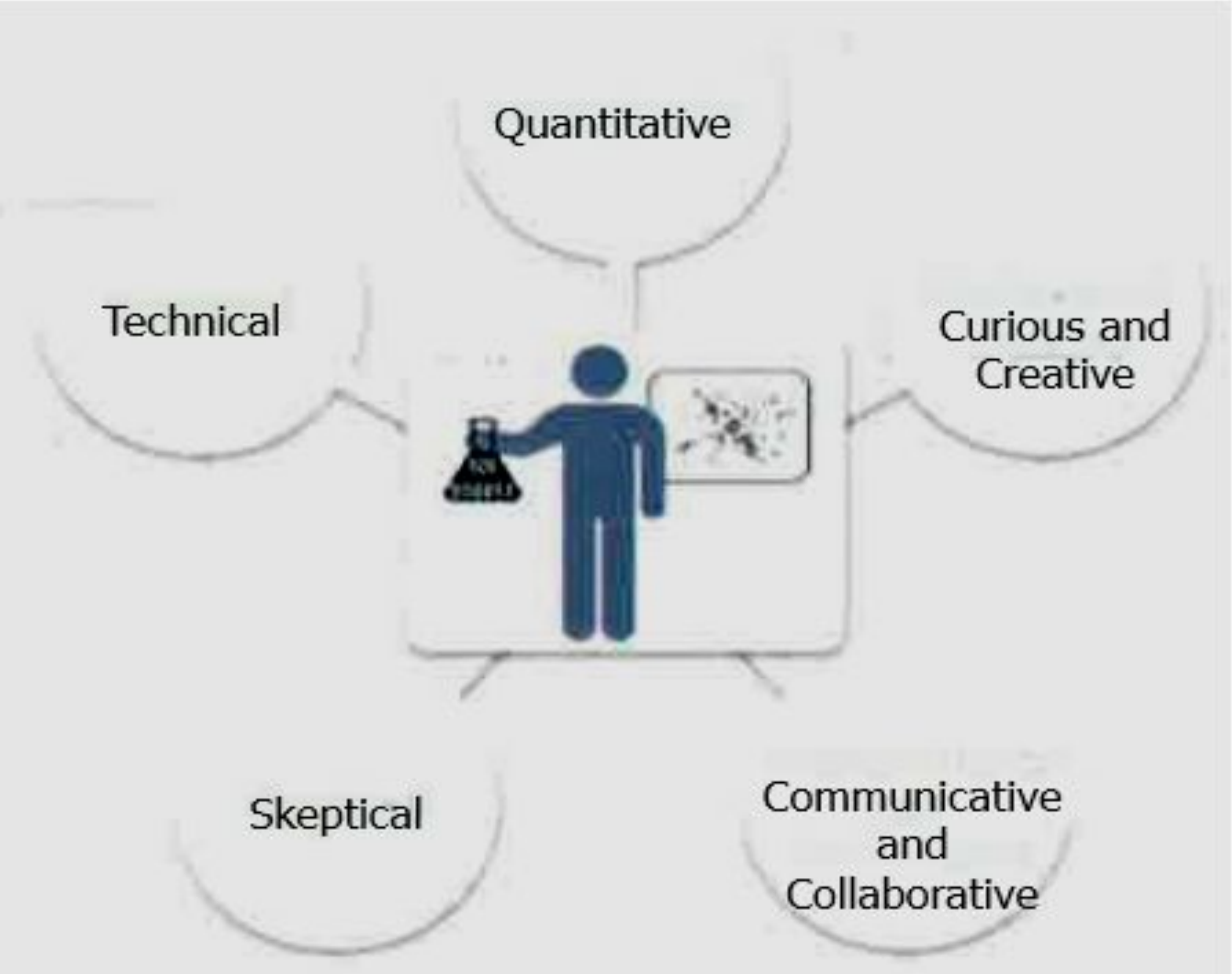
What do they do?

- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

Three Recurring Data Scientist Activities

1. Reframe business challenges as analytics challenges
2. Design, implement, and deploy statistical models and data mining techniques on Big Data
3. Develop insights that lead to actionable recommendations

Profile of Data Scientist
Five Main Sets of Skills



Profile of Data Scientist
Five Main Sets of Skills

- Quantitative skill – e.g., math, statistics
- Technical aptitude – e.g., software engineering, programming
- Skeptical mindset and critical thinking – ability to examine work critically
- Curious and creative – passionate about data and finding creative solutions
- Communicative and collaborative – can articulate ideas, can work with others

- Data science projects differ from BI projects
  - More exploratory in nature
  - Critical to have a project process
  - Participants should be thorough and rigorous
- Break large projects into smaller pieces
- Spend time to plan and scope the work
- Documenting adds rigor and credibility

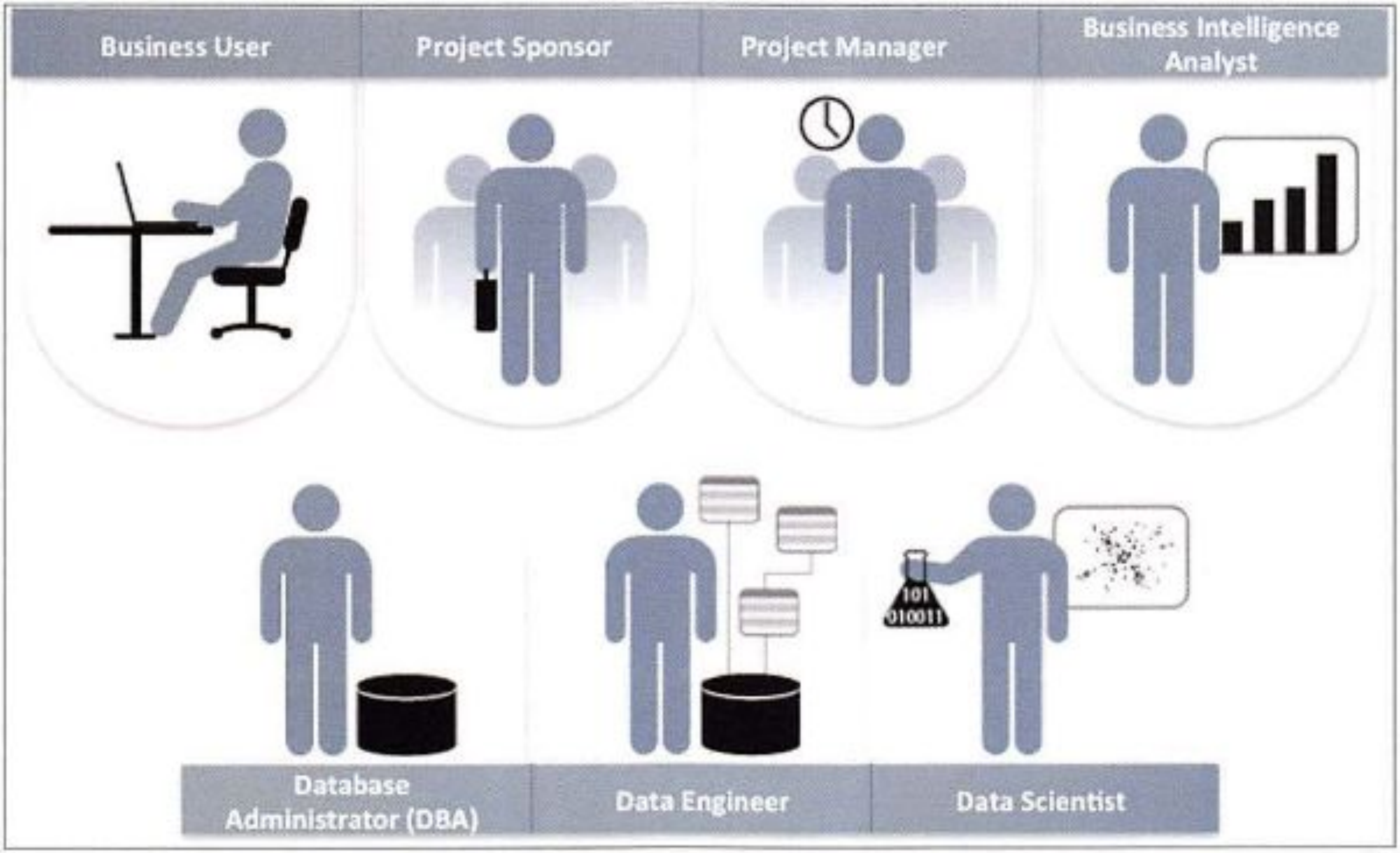
Data Analytics Lifecycle

- Data Analytics Lifecycle Overview
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Communicate Results
- Phase 6: Operationalize

Data Analytics Lifecycle Overview

- The data analytic lifecycle is designed for Big Data problems and data science projects
- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered

Key Roles for a Successful Analytics Project



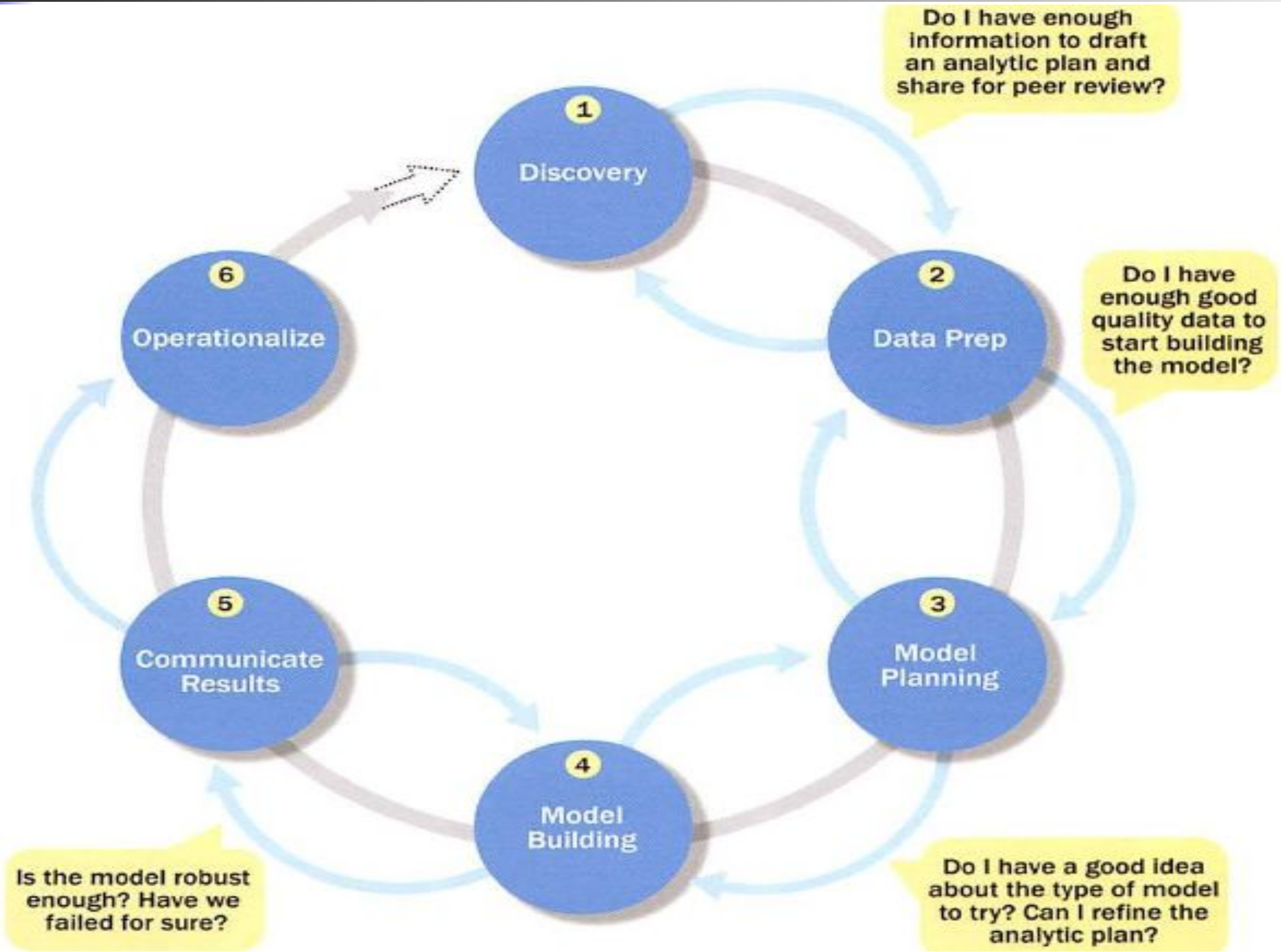
Key Roles for a Successful Analytics Project

- Business User – understands the domain area
- Project Sponsor – provides requirements
- Project Manager – ensures meeting objectives
- Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data
- Database Administrator (DBA) – creates DB environment
- Data Engineer – provides technical skills, assists data management and extraction, supports analytic sandbox
- Data Scientist – provides analytic techniques and modeling

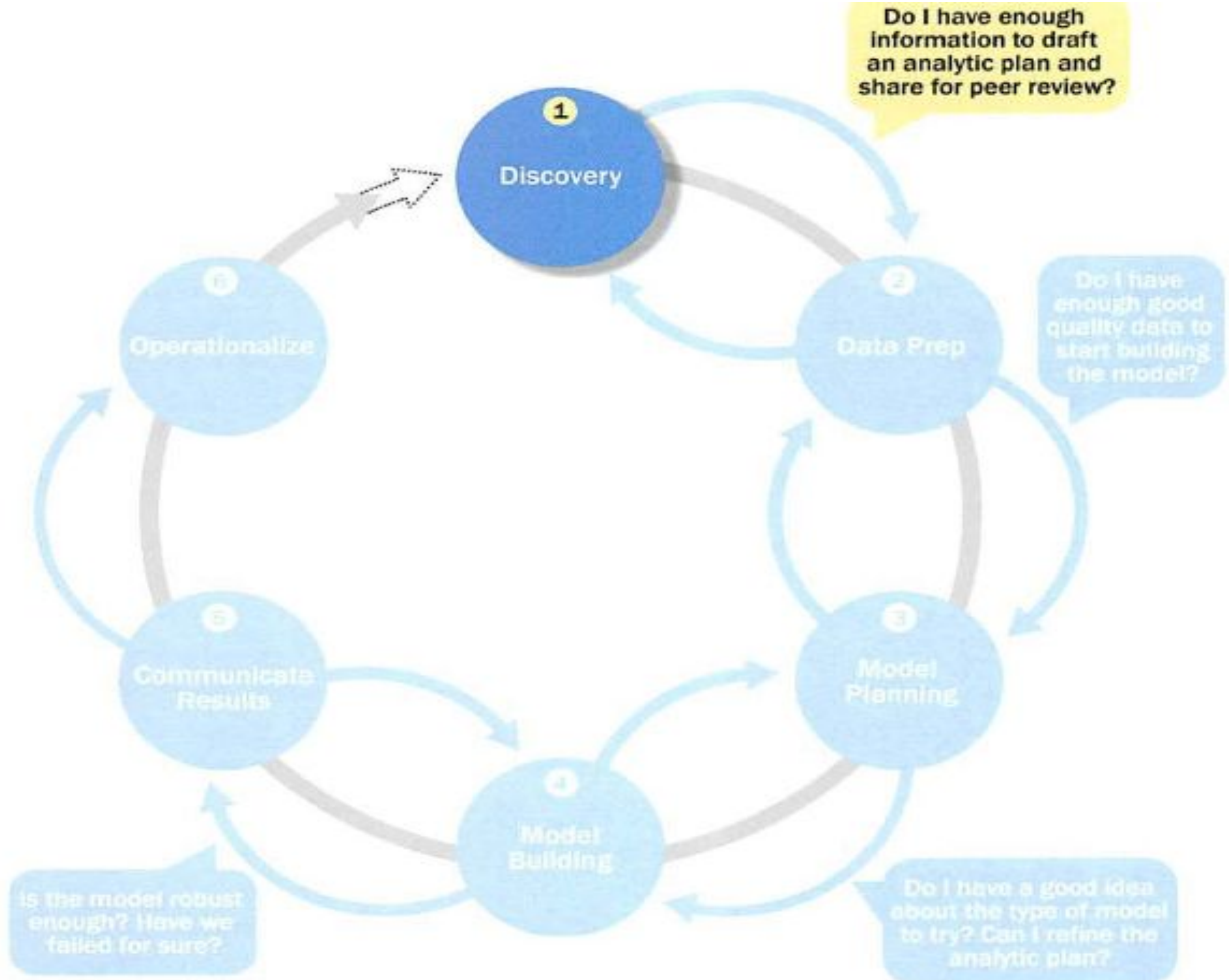
Background and Overview of Data Analytics Lifecycle

- Data Analytics Lifecycle defines the analytics process and best practices from discovery to project completion
- The Lifecycle employs aspects of
  - Scientific method
  - Cross Industry Standard Process for Data Mining (CRISP-DM)
    - Process model for data mining
  - Davenport’s DELTA framework
  - Hubbard’s Applied Information Economics (AIE) approach
  - MAD Skills: New Analysis Practices for Big Data by Cohen et al.

Overview of Data Analytics Lifecycle


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

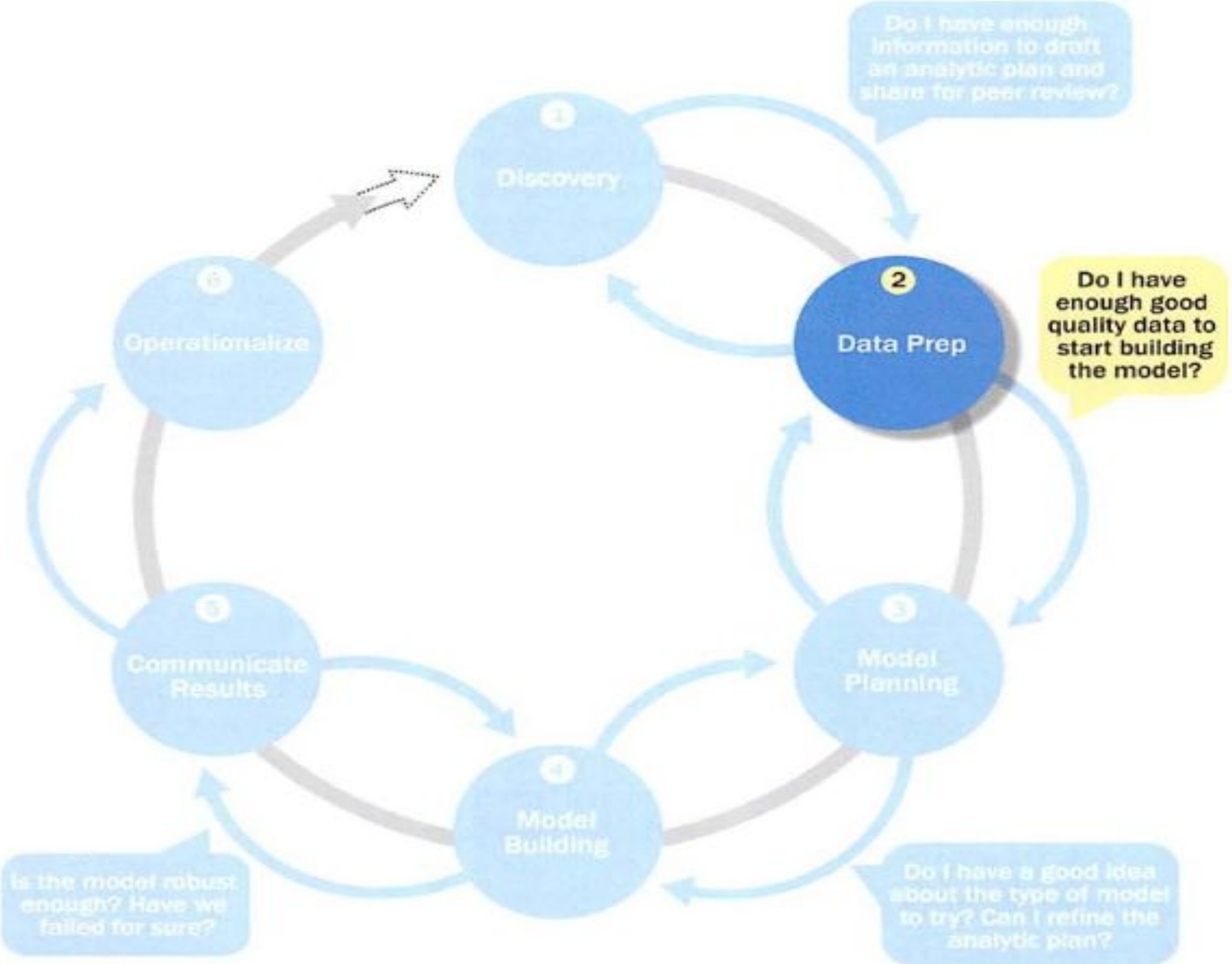
Phase 1: Discovery


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

Phase 1: Discovery

1. Learning the Business Domain
2. Resources
3. Framing the Problem
4. Identifying Key Stakeholders
5. Interviewing the Analytics Sponsor
6. Developing Initial Hypotheses
7. Identifying Potential Data Sources

Phase 2: Data Preparation


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

Phase 2: Data Preparation

- Includes steps to explore, preprocess, and condition data
- Create robust environment – analytics sandbox
- Data preparation tends to be the most labor-intensive step in the analytics lifecycle
  - Often at least 50% of the data science project's time
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often

Preparing the Analytic Sandbox

- Create the analytic sandbox (also called workspace)
- Allows team to explore data without interfering with live production data
- Sandbox collects all kinds of data (expansive approach)
- The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics
- Although the concept of an analytics sandbox is relatively new, this concept has become acceptable to data science teams and IT groups

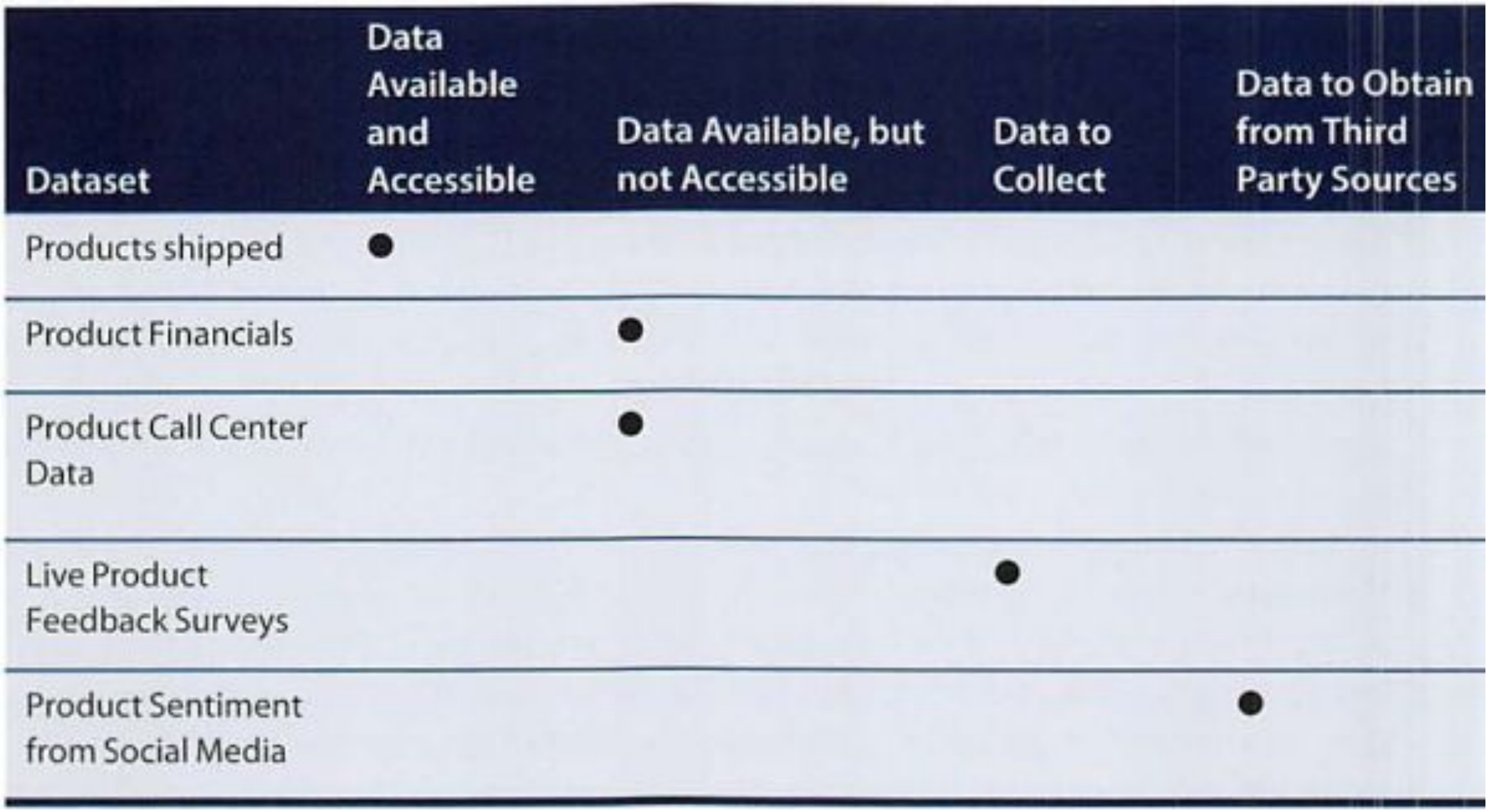
Performing ETLT (Extract, Transform, Load, Transform)

- In ETL users perform extract, transform, load
- In the sandbox the process is often ELT – early load preserves the raw data which can be useful to examine
- Example – in credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database
- Hadoop is often used here

Learning about the Data

- Becoming familiar with the data is critical
- This activity accomplishes several goals:
  - Determines the data available to the team early in the project
  - Highlights gaps – identifies data not currently available
  - Identifies data outside the organization that might be useful

Learning about the Data Sample Dataset Inventory



Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Data Conditioning

- Data conditioning includes cleaning data, normalizing datasets, and performing transformations
  - Often viewed as a preprocessing step prior to data analysis, it might be performed by data owner, IT department, DBA, etc.
  - Best to have data scientists involved
  - Data science teams prefer more data than too little

Data Conditioning

- Additional questions and considerations
  - What are the data sources? Target fields?
  - How clean is the data?
  - How consistent are the contents and files? Missing or inconsistent values?
  - Assess the consistence of the data types – numeric, alphanumeric?
  - Review the contents to ensure the data makes sense
  - Look for evidence of systematic error

Survey and Visualize

- Leverage data visualization tools to gain an overview of the data
- Shneiderman’s mantra:
  - “Overview first, zoom and filter, then details-on-demand”
  - This enables the user to find areas of interest, zoom and filter to find more detailed information about a particular area, then find the detailed data in that area

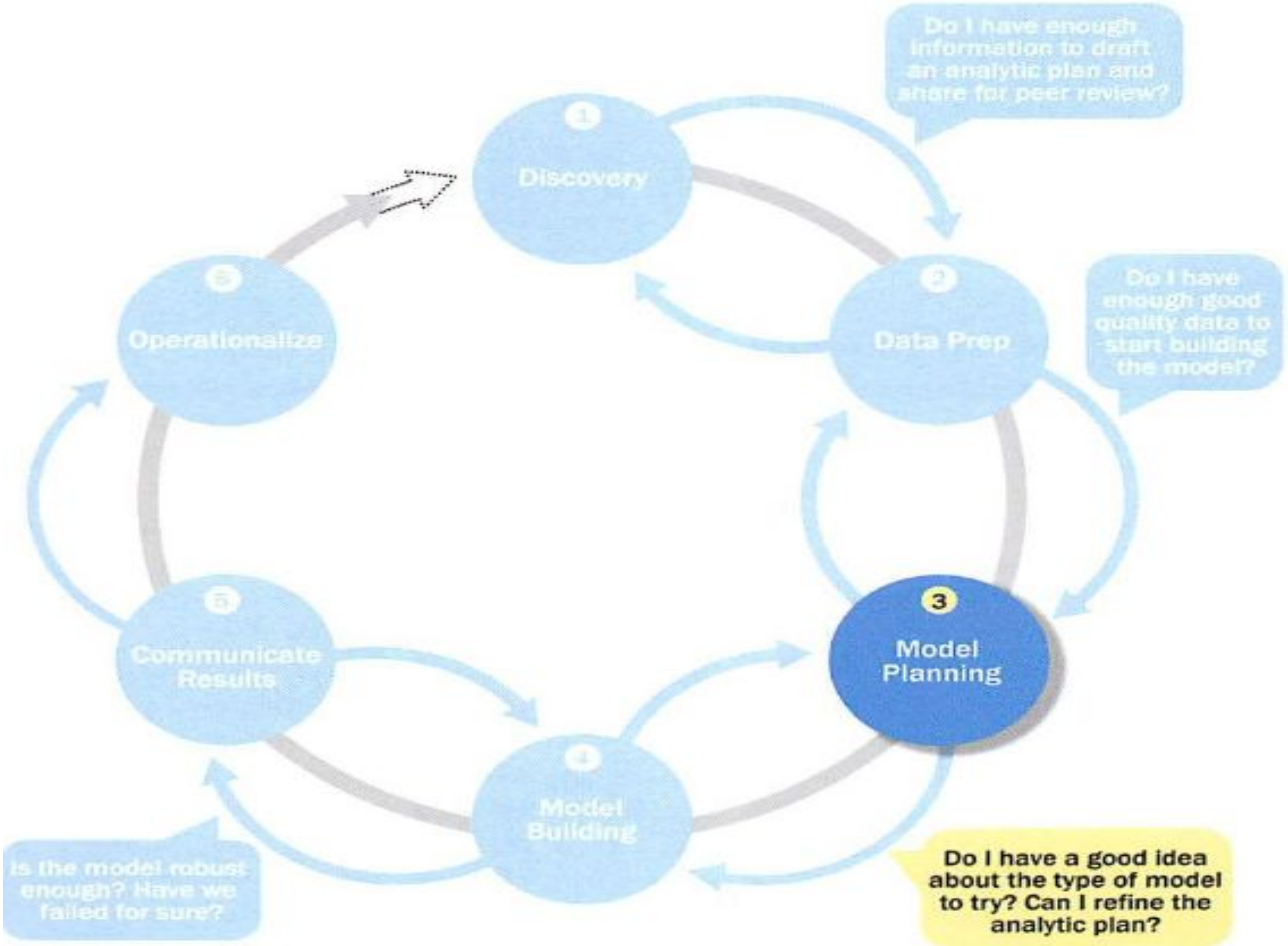
Survey and Visualize Guidelines and Considerations

- Review data to ensure calculations are consistent
- Does the data distribution stay consistent?
- Assess the granularity of the data, the range of values, and the level of aggregation of the data
- Does the data represent the population of interest?
- Check time-related variables – daily, weekly, monthly? Is this good enough?
- Is the data standardized/normalized? Scales consistent?
- For geospatial datasets, are state/country abbreviations consistent

Common Tools for Data Preparation

- **Hadoop** can perform parallel ingest and analysis
- **Alpine Miner** provides a graphical user interface for creating analytic workflows
- **OpenRefine** (formerly Google Refine) is a free, open source tool for working with messy data
- Similar to OpenRefine, **Data Wrangler** is an interactive tool for data cleansing and transformation

Phase 3: Model Planning

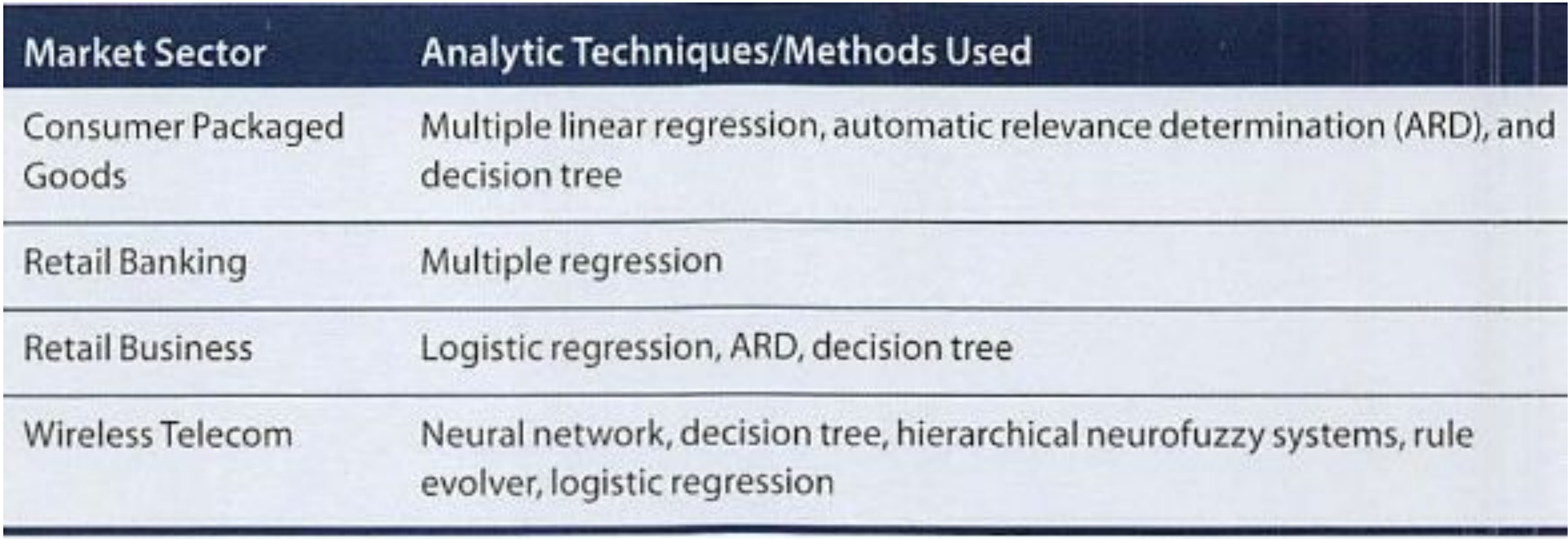

Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

Phase 3: Model Planning

- Activities to consider
  - Assess the structure of the data – this dictates the tools and analytic techniques for the next phase
  - Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses
  - Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow
  - Research and understand how other analysts have approached this kind or similar kind of problem

Phase 3: Model Planning
Model Planning in Industry Verticals

- Example of other analysts approaching a similar problem



Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

Data Exploration and Variable Selection

- Explore the data to understand the relationships among the variables to inform selection of the variables and methods
- A common way to do this is to use data visualization tools
- Often, stakeholders and subject matter experts may have ideas
  - For example, some hypothesis that led to the project
- Aim for capturing the most essential predictors and variables
  - This often requires iterations and testing to identify key variables
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model

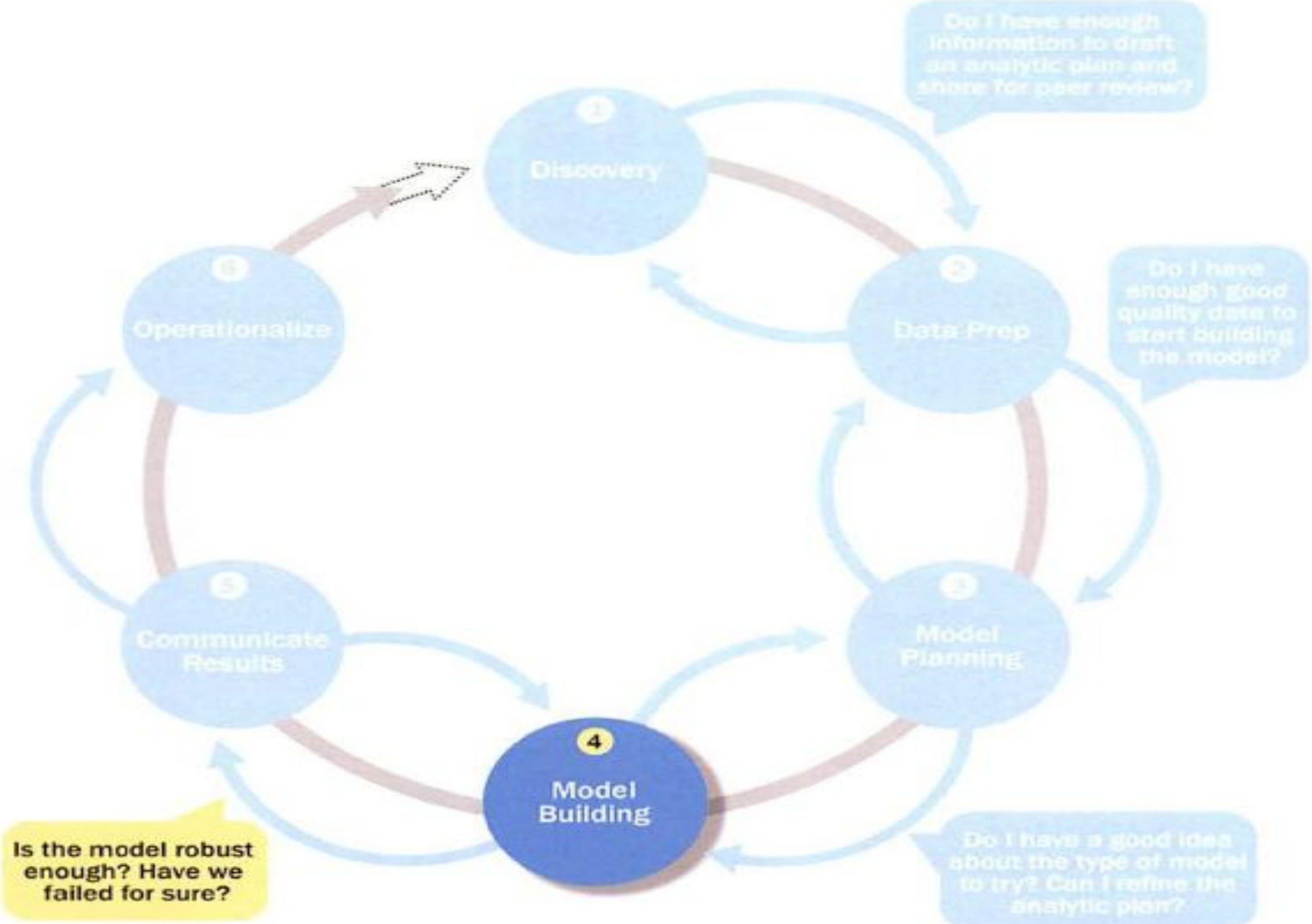
Model Selection

- The main goal is to choose an analytical technique, or several candidates, based on the end goal of the project
- We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions
    - A model is simply an abstraction from reality
- Determine whether to use techniques best suited for structured data, unstructured data, or a hybrid approach
- Teams often create initial models using statistical software packages such as R, SAS, or Matlab
    - Which may have limitations when applied to very large datasets
- The team moves to the model building phase once it has a good idea about the type of model to try

Common Tools for the Model Planning Phase

- **R** has a complete set of modeling capabilities
  - R contains about 5000 packages for data analysis and graphical presentation
- **SQL Analysis services** can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models
- **SAS/ACCESS** provides integration between SAS and the analytics sandbox via multiple data connections

Phase 4: Model Building


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

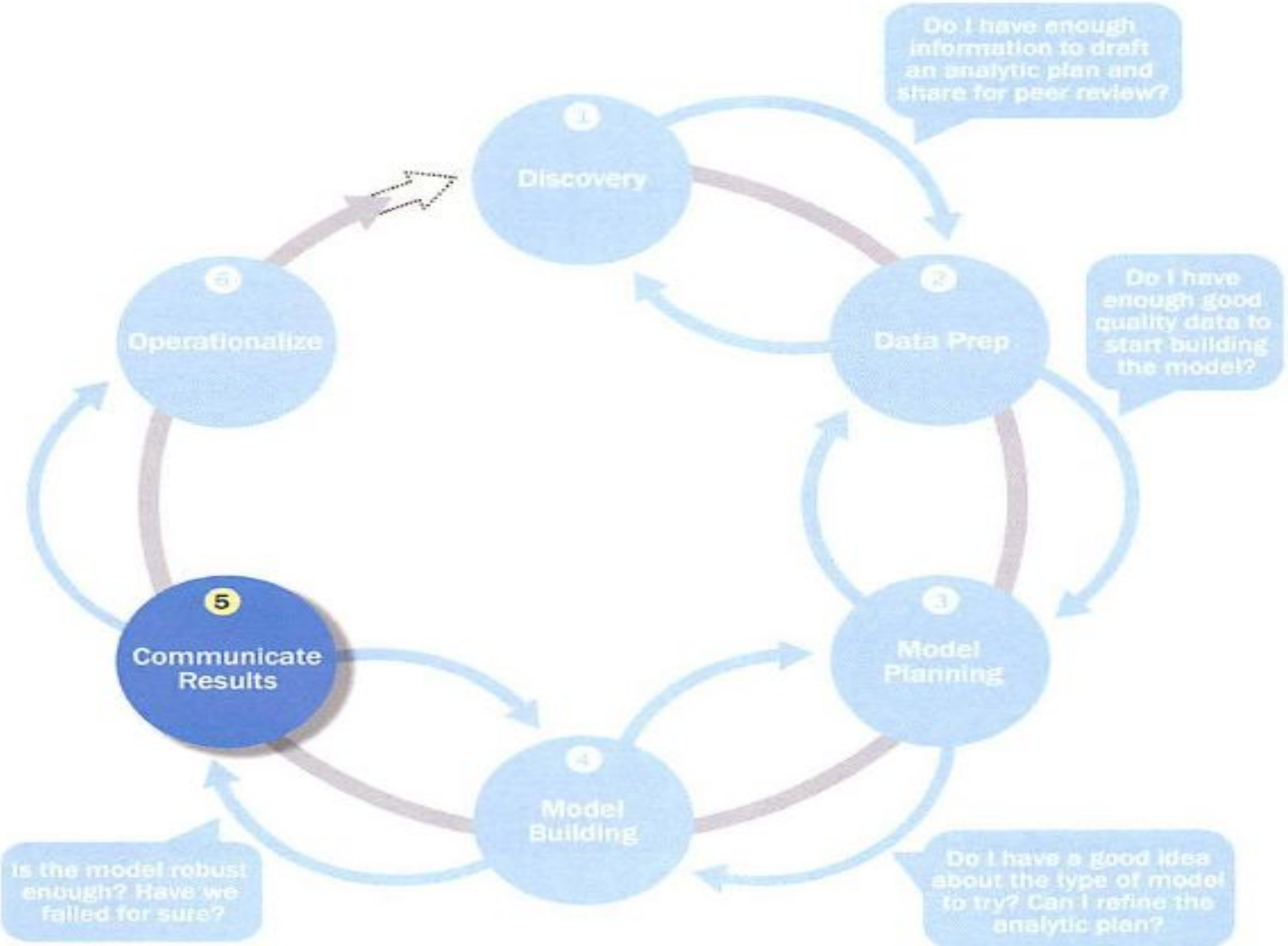
Phase 4: Model Building

- Execute the models defined in Phase 3
- Develop datasets for training, testing, and production
- Develop analytic model on training data, test on test data
- Question to consider
  - Does the model appear valid and accurate on the test data?
  - Does the model output/behavior make sense to the domain experts?
  - Do the parameter values make sense in the context of the domain?
  - Is the model sufficiently accurate to meet the goal?
  - Does the model avoid intolerable mistakes?
  - Are more data or inputs needed?
  - Will the kind of model chosen support the runtime environment?
  - Is a different form of the model required to address the business problem?

Common Tools for the Model Building Phase

- Commercial Tools
  - SAS Enterprise Miner – built for enterprise-level computing and analytics
  - SPSS Modeler (IBM) – provides enterprise-level computing and analytics
  - Matlab – high-level language for data analytics, algorithms, data exploration
  - Alpine Miner – provides GUI frontend for backend analytics tools
  - STATISTICA and MATHEMATICA – popular data mining and analytics tools
- Free or Open Source Tools
  - R and PL/R - PL/R is a procedural language for PostgreSQL with R
  - Octave – language for computational modeling
  - WEKA – data mining software package with analytic workbench
  - Python – language providing toolkits for machine learning and analysis
  - SQL – in-database implementations provide an alternative tool

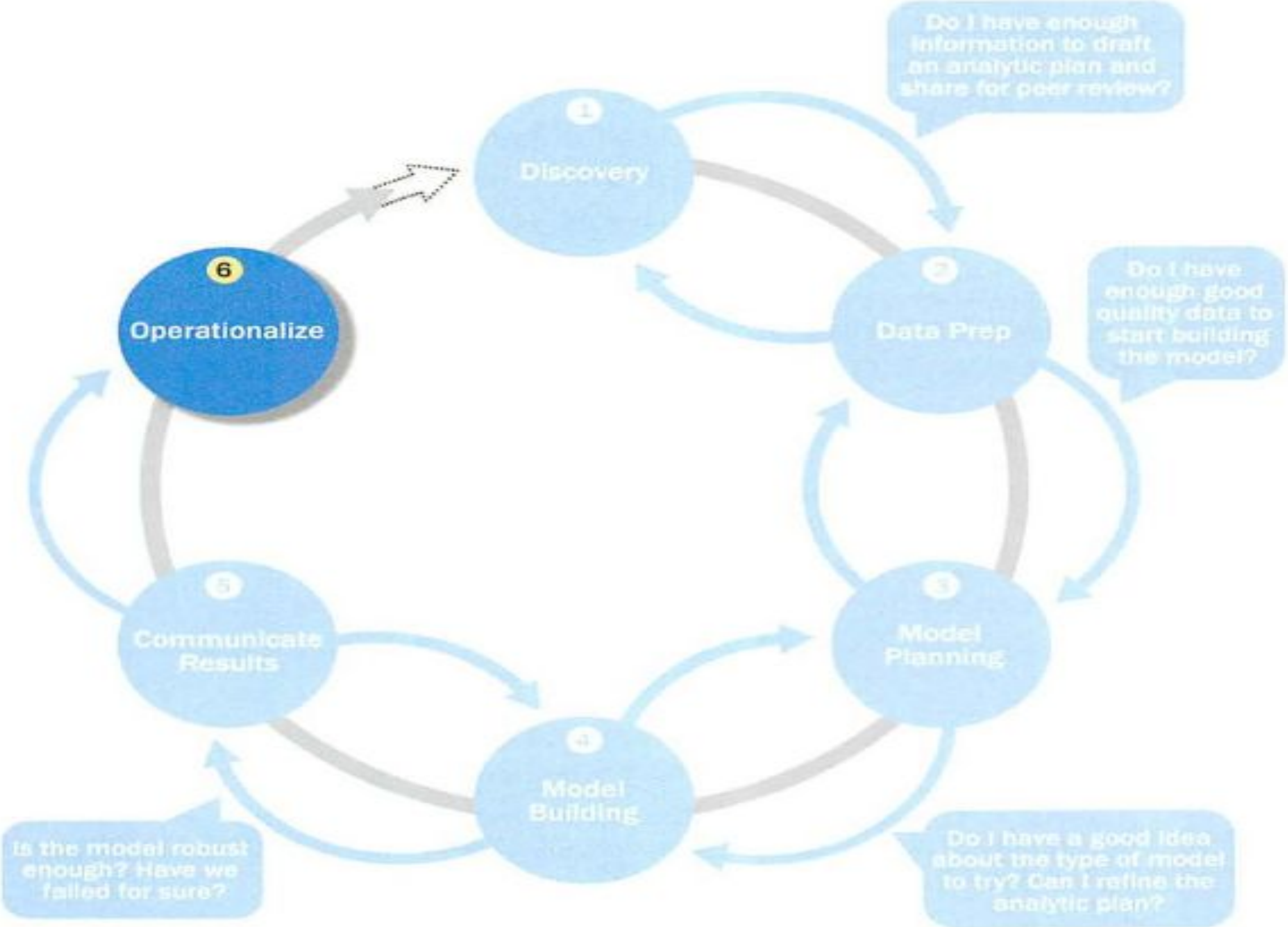
Phase 5: Communicate Results


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
6
Operationalize
5
Communicate Results
3
Model Planning
4
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Phase 5: Communicate Results

- Determine if the team succeeded or failed in its objectives
- Assess if the results are statistically significant and valid
  - If so, identify aspects of the results that present salient findings
  - Identify surprising results and those in line with the hypotheses
- Communicate and document the key findings and major insights derived from the analysis
  - This is the most visible portion of the process to the outside stakeholders and sponsors

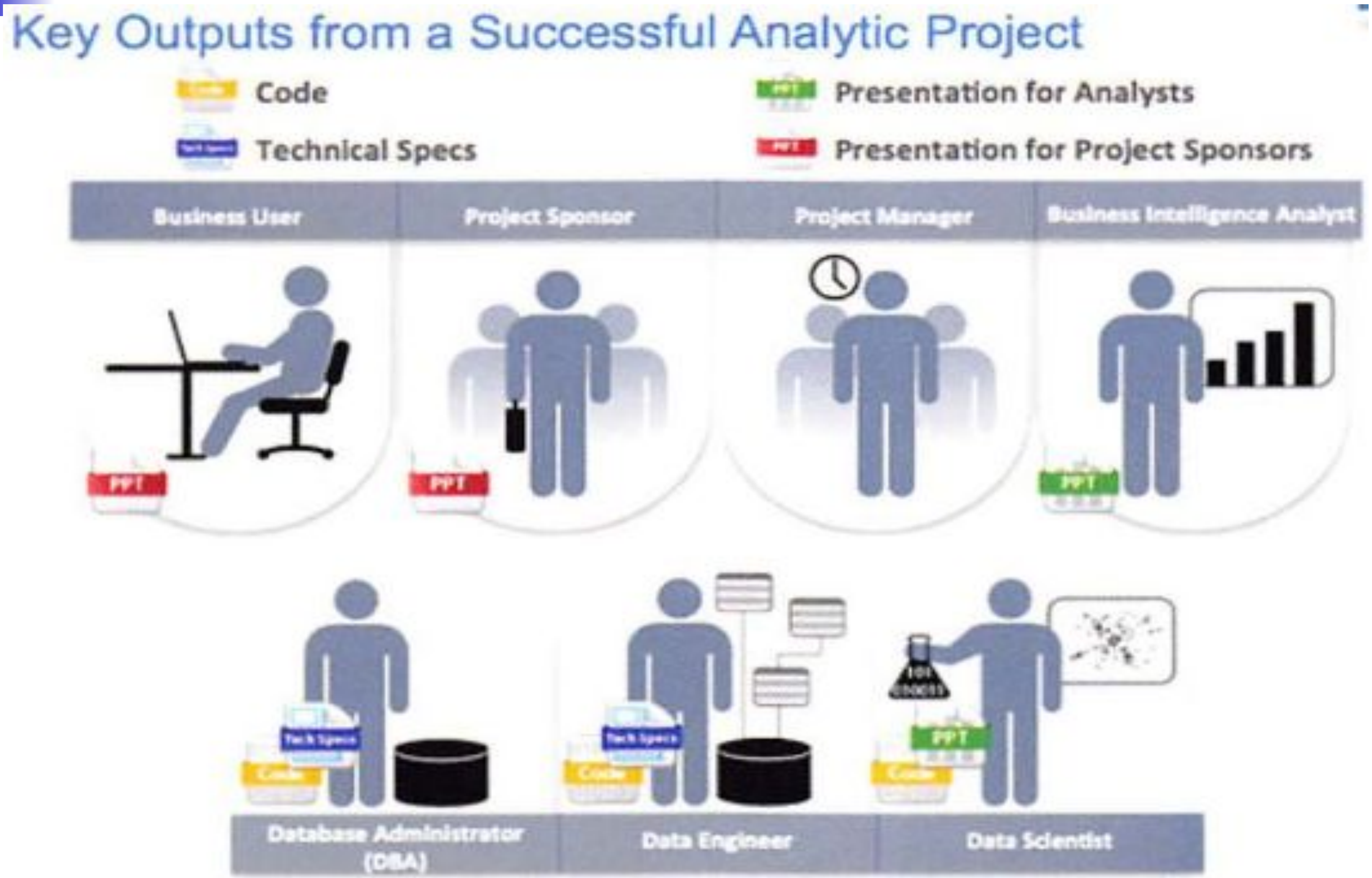
Phase 6: Operationalize


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
6
Operationalize
3
Model Planning
5
Communicate Results
4
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Phase 6: Operationalize

- In this last phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
- Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
- During the pilot project, the team may need to execute the algorithm more efficiently in the database rather than with in-memory tools like R, especially with larger datasets
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business
- Monitor model accuracy and retrain the model if necessary

Phase 6: Operationalize
Key outputs from successful analytics project


Key Outputs from a Successful Analytic Project
Code
Technical Specs
Presentation for Analysts
Presentation for Project Sponsors
Business User
Project Sponsor
Project Manager
Business Intelligence Analyst
PPT
PPT
PPT
Tech Specs
Code
Tech Specs
Code
PPT
Code
Database Administrator (DBA)
Data Engineer
Data Scientist

Phase 6: Operationalize
Key outputs from successful analytics project

- Business user – tries to determine business benefits and implications
- Project sponsor – wants business impact, risks, ROI
- Project manager – needs to determine if project completed on time, within budget, goals met
- Business intelligence analyst – needs to know if reports and dashboards will be impacted and need to change
- Data engineer and DBA – must share code and document
- Data scientist – must share code and explain model to peers, managers, stakeholders

Phase 6: Operationalize
Four main deliverables

- Although the seven roles represent many interests, the interests overlap and can be met with four main deliverables
  1. Presentation for project sponsors – high-level takeaways for executive level stakeholders
  2. Presentation for analysts – describes business process changes and reporting changes, includes details and technical graphs
  3. Code for technical people
  4. Technical specifications of implementing the code

Case Study: Global Innovation Network and Analysis (GINA)

- In 2012 EMC's new director wanted to improve the company's engagement of employees across the global centers of excellence (GCE) to drive innovation, research, and university partnerships
- This project was created to accomplish
  - Store formal and informal data
  - Track research from global technologists
  - Mine the data for patterns and insights to improve the team's operations and strategy

Phase 1: Discovery

- Team members and roles
  - Business user, project sponsor, project manager – Vice President from Office of CTO
  - BI analyst – person from IT
  - Data engineer and DBA – people from IT
  - Data scientist – distinguished engineer

Phase 1: Discovery

- The data fell into two categories
  - Five years of idea submissions from internal innovation contests
  - Minutes and notes representing innovation and research activity from around the world
- Hypotheses grouped into two categories
  - Descriptive analytics of what is happening to spark further creativity, collaboration, and asset generation
  - Predictive analytics to advise executive management of where it should be investing in the future

Phase 2: Data Preparation

- Set up an analytics sandbox
- Discovered that certain data needed conditioning and normalization and that missing datasets were critical
- Team recognized that poor quality data could impact subsequent steps
- They discovered many names were misspelled and problems with extra spaces
- These seemingly small problems had to be addressed

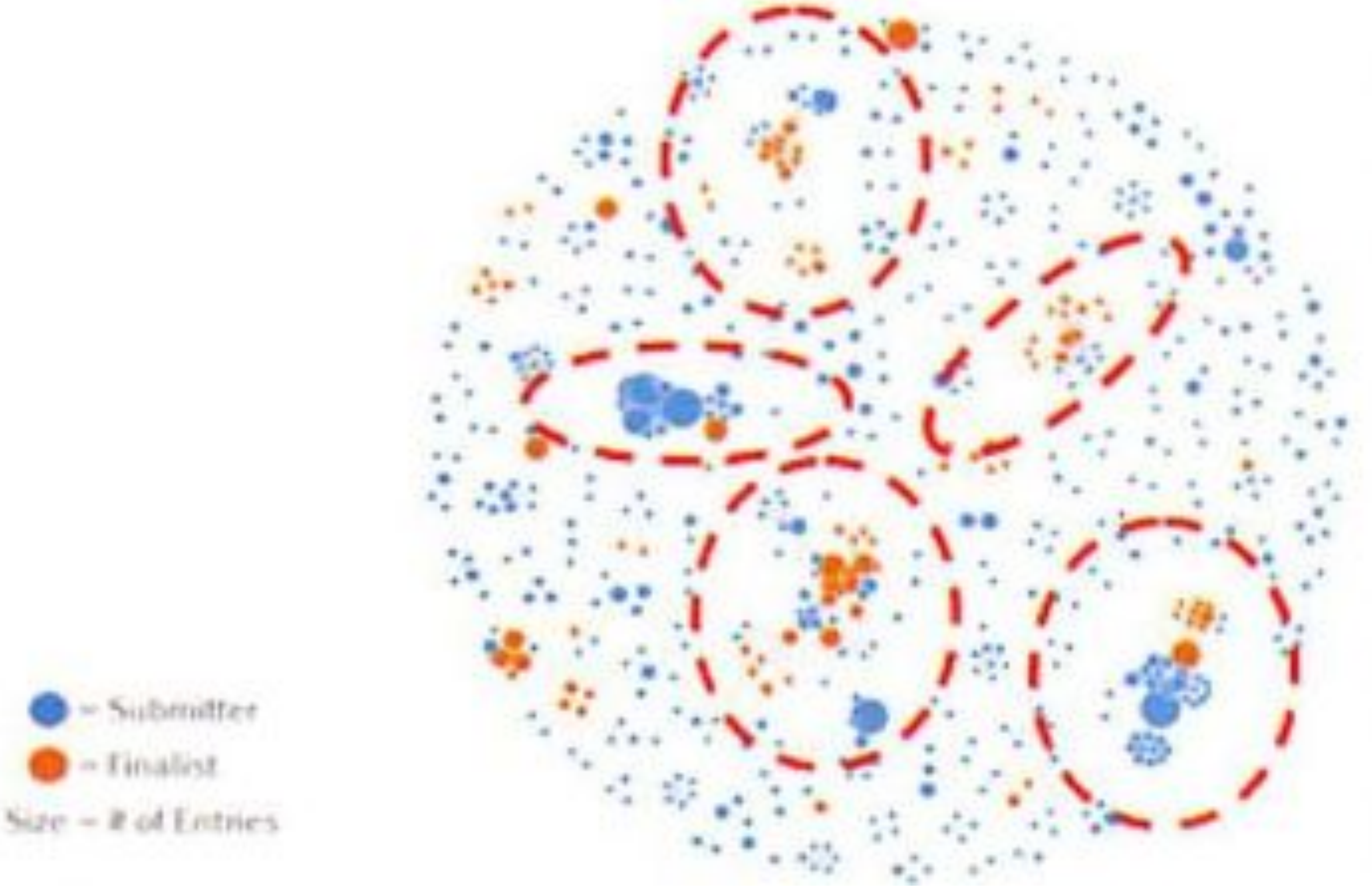
Phase 3: Model Planning

- The study included the following considerations
  - Identify the right milestones to achieve the goals
  - Trace how people move ideas from each milestone toward the goal
  - Tract ideas that die and others that reach the goal
  - Compare times and outcomes using a few different methods

Phase 4: Model Building

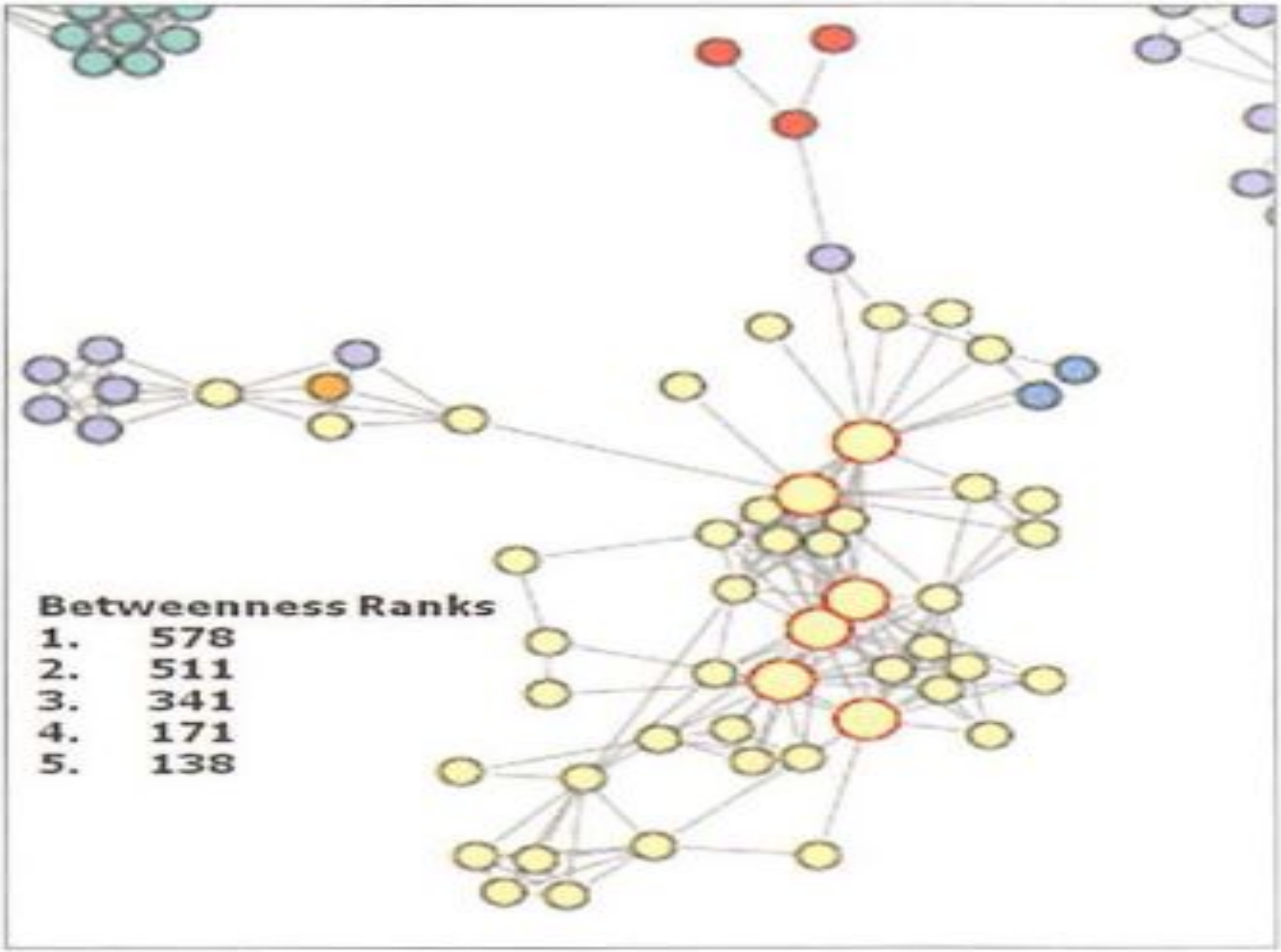
- Several analytic method were employed
  - NLP on textual descriptions
  - Social network analysis using R and Rstudio
  - Developed social graphs and visualizations

Phase 4: Model Building
Social graph of data submitters and finalists



Phase 4: Model Building

Social graph of top innovation influencers



Betweenness Ranks

1.	578
2.	511
3.	341
4.	171
5.	138

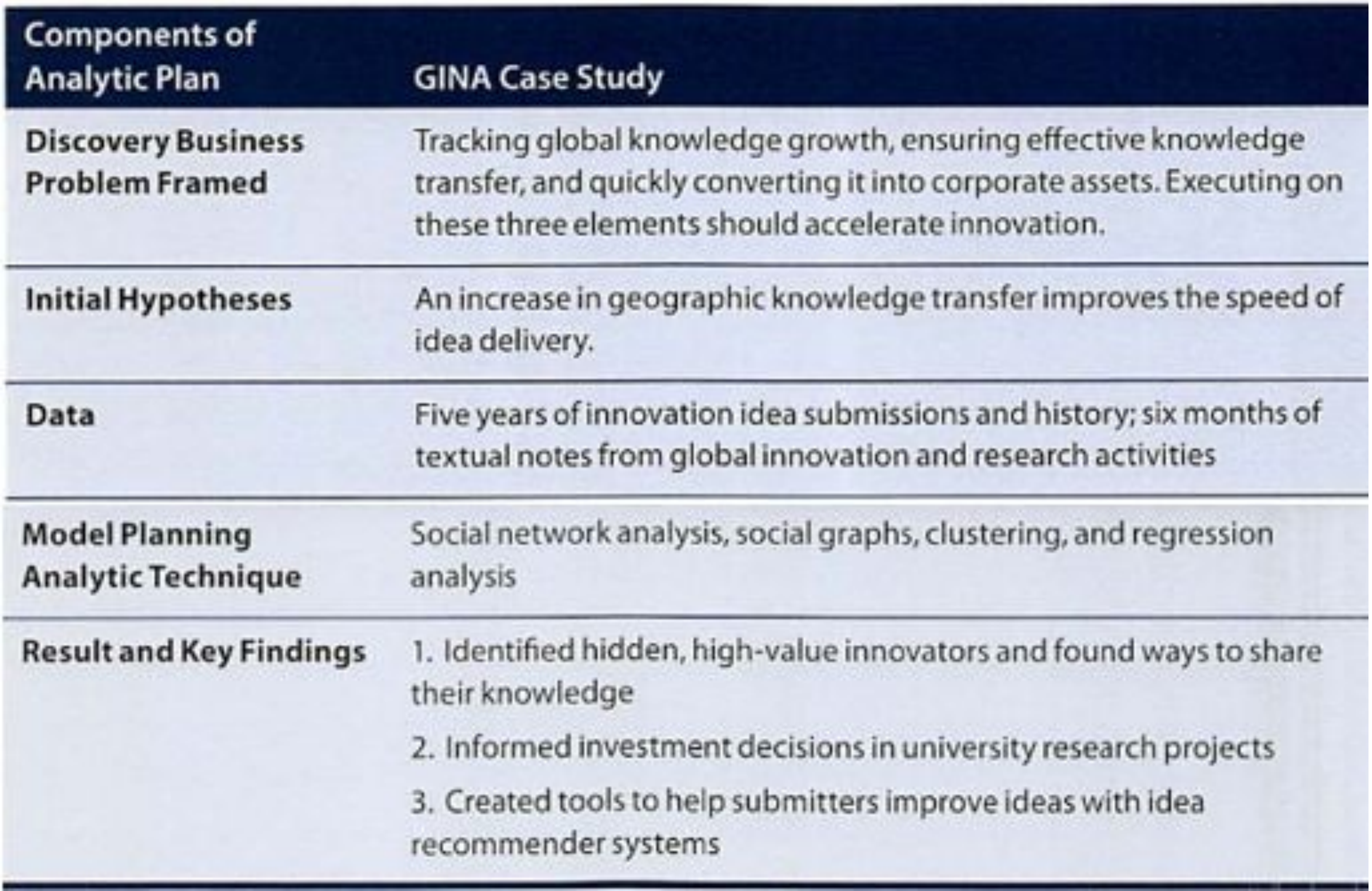
Phase 5: Communicate Results

- Study was successful in in identifying hidden innovators
    - Found high density of innovators in Cork, Ireland
- The CTO office launched longitudinal studies

Phase 6: Operationalize

- Deployment was not really discussed
- Key findings
  - Need more data in future
  - Some data were sensitive
  - A parallel initiative needs to be created to improve basic BI activities
  - A mechanism is needed to continually reevaluate the model after deployment

Phase 6: Operationalize



Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems