CDC Data Science Certificate Program

April 5th, 2021

Agenda

- About BeVera Solutions
- About your team of trainers
- Program Details
- Session 1: Intro to Data Science
- Preparing for the upcoming months

About BeVera

- Founded in 2013
- HQ in Metro Atlanta
- Veteran-Owned Small Business
- Data Science Training and Consulting
- Technical Staffing
- Dedicated to making Data Science understandable and usable at every level of the organization
- Helping our customers become self-sufficient

BK Ashford, Director of Data Science

- Over 22 years of Data Science solution delivery experience
- Analytical Consultant for SAS Institute for over 11 years
- Successful Data Science solution delivery across multiple industries
  - Insurance Carriers
  - Major Airlines
  - Credit Bureaus
  - Retail
- Education:
  - M.Sc. Applied Mathematics, Lehigh University
  - B.S. Mathematics, Morehouse College

BeVera

Yvonne Phillips, Instructor

- 19 years of advanced analytics work
- Currently a consultant and speaker in Data Science. Also, Adjunct Professor, Morehouse College – Computer Science/Data Science
- Experienced analytic professional with demonstrated history in:
  - Insurance: Underwriting Auto & Property, Motor Vehicle Reporting Analytics
  - Government: Federal and State/Local government customer solutions needs
  - Financial/Credit Cards
  - Retail
- Education:
  - M.Sc., Decision Science, Georgia State University, Robinson College of Business
  - B.S., Mathematics, Spelman College

BeVera

Mo Medwani, Instructor

- Analytics professional, consultant and entrepreneur in Data Science
- Experienced IT and analytic professional with demonstrated history specifically in:
  - Banking
  - Computer Technology
- Education:
  - Ph.D. Candidate in AI
  - Postgraduate Diploma in Data Science, Emeritus Institute of Management,
  - M.Sc., Data Science, University of Maryland University College
  - MBA., Southern Polytechnic State University
  - B.S., Computer Science, Jackson State Community College

Program Details

Training	Full	Group A	Group B	Group C
Introduction to Data Science	5-Apr			
Basics of R Programming		7-Apr	9-Apr	14-Apr
Data Wrangling and Visualization Basics I		21-Apr	27-Apr	28-Apr
Data Wrangling and Visualization Basics II		23-Apr	29-Apr	30-Apr
Exploratory Data Analysis		11-May	12-May	13-May
Statistical Analysis in R		25-May	27-May	2-Jun
Introduction to Machine Learning	14-Jun			
Clustering Analysis	30-Jun			
Regression Analysis	12-Jul			
Decision Trees/Random Forest	16-Aug			
Ensemble Modeling	8-Sep			

Additional Learning Opportunities

- Assigned Homework
- After-Class Group Sessions
- Applied Learning Project

Applied Learning Project

You will present your finding in a 20-minute presentation:
Including the following:

- Topic Selection
  - How did you select your topic?
  - What stakeholders did you engage?
  - What are the business needs you are trying to address?
  - What are the objectives of the project?

Applied Learning Project

Extracting, Cleaning and Exploring the Data

- What data was available to you?
- From where did you source the data and in what format was it in?
- Did you find and merge additional data sets?
- Did you utilize unstructured data and how did you manage it?
- What types of EDA methods did you employ?
  - What issues did you find with the data and how did you resolve them?
  - What visualization techniques did you employ and what was learned from those visualizations?

Applied Learning Project

- Modifying and Preparing Data for Modeling
  - Describe the feature engineering methods you employed and why
    - Transformations
    - Modifications
    - Missing value imputations
- Gaining Deeper Insights (Modeling Techniques)
  - Describe your findings from Clustering (cluster analysis)
    - What did you learn from Profiling your data?
  - Which algorithm did you gain the most insight?
  - How can you leverage the learnings from this project to add value to your organization?

Applied Learning Project Topic

14 June

Participants should not finalize a project topic until they've completed the Introduction to Machine Learning class on June 14th.

Date for finalizing the project topic is June 30th.

30 June

Module 1: Why should you become a Data Scientist?

- Data Explosion
- Why Data Science?
- What is Data Science?
- Type of Analytics
- Data Science Portfolio
- Data Science Process
- Career in Data Science

BeVera

Data Science is the Profession of the Future

According to Tim Berners Lee, the inventor of the World Wide Web–

Data is a Precious Thing and will Last Longer than the Systems themselves.

Discussion: Why Data Science For You?

Article Read:

https://www.dexlabanalytics.com/blog/application-of-data-science-in-healthcare

- Google gets over 3.5 billion searches daily.

- WhatsApp users exchange up to 65 billion messages daily.
- 80-90% of the data we generate today is unstructured.
- The market of Big Data analytics in healthcare could be worth $67.82 billion by 2025.

The Why in Data Science

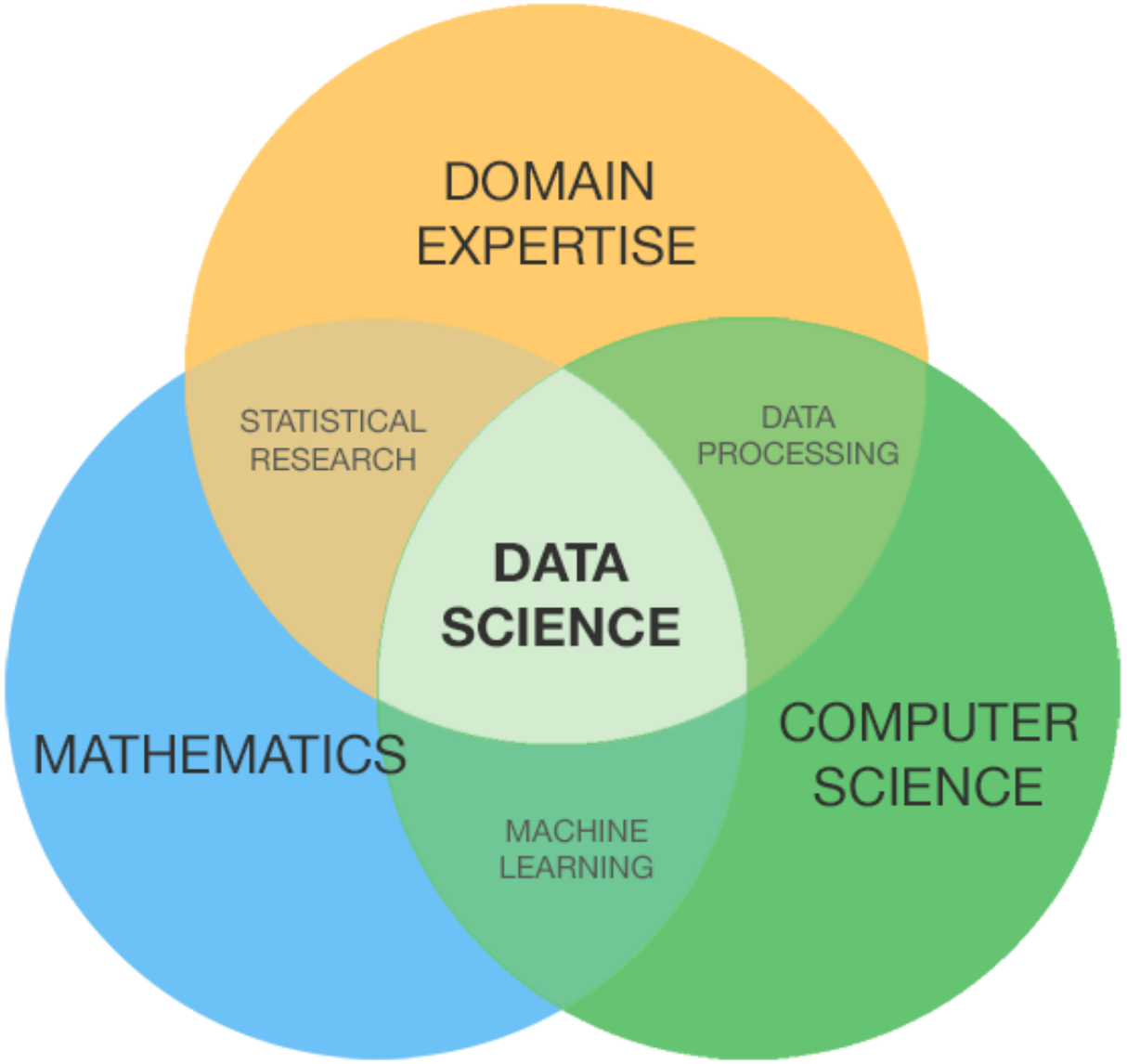
In data science, the algorithmic principles are applied to greater areas of uncertainty, often producing probabilistic answers to interdisciplinary questions about business.

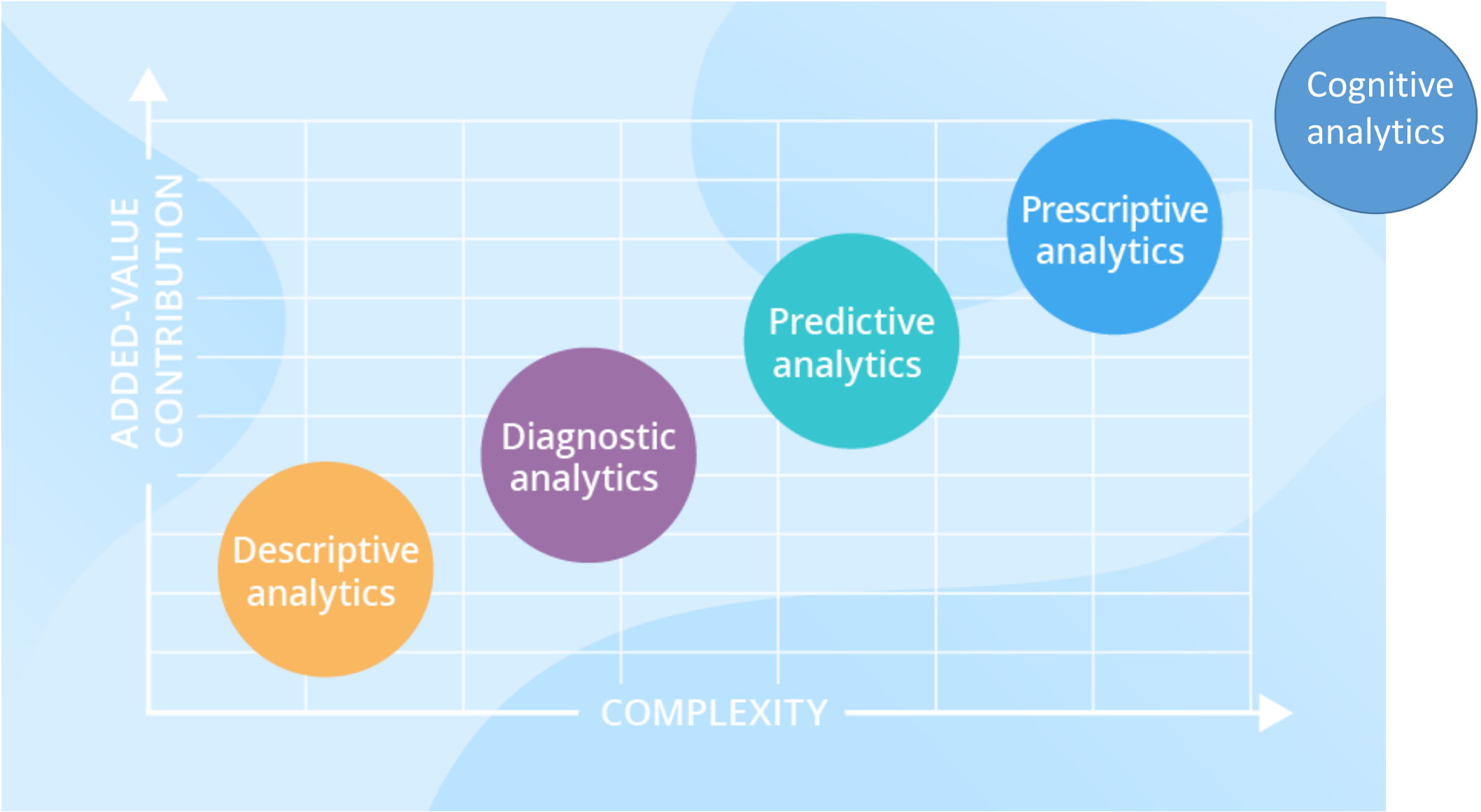
Modern data scientists typically have a proficiency in computer science, but they can come from mathematical, statistical, or even business backgrounds.

Working on top of what computer science has built, data scientists design unique ways to filter through the massive amounts of data that flow through network systems and then extract actionable insights.

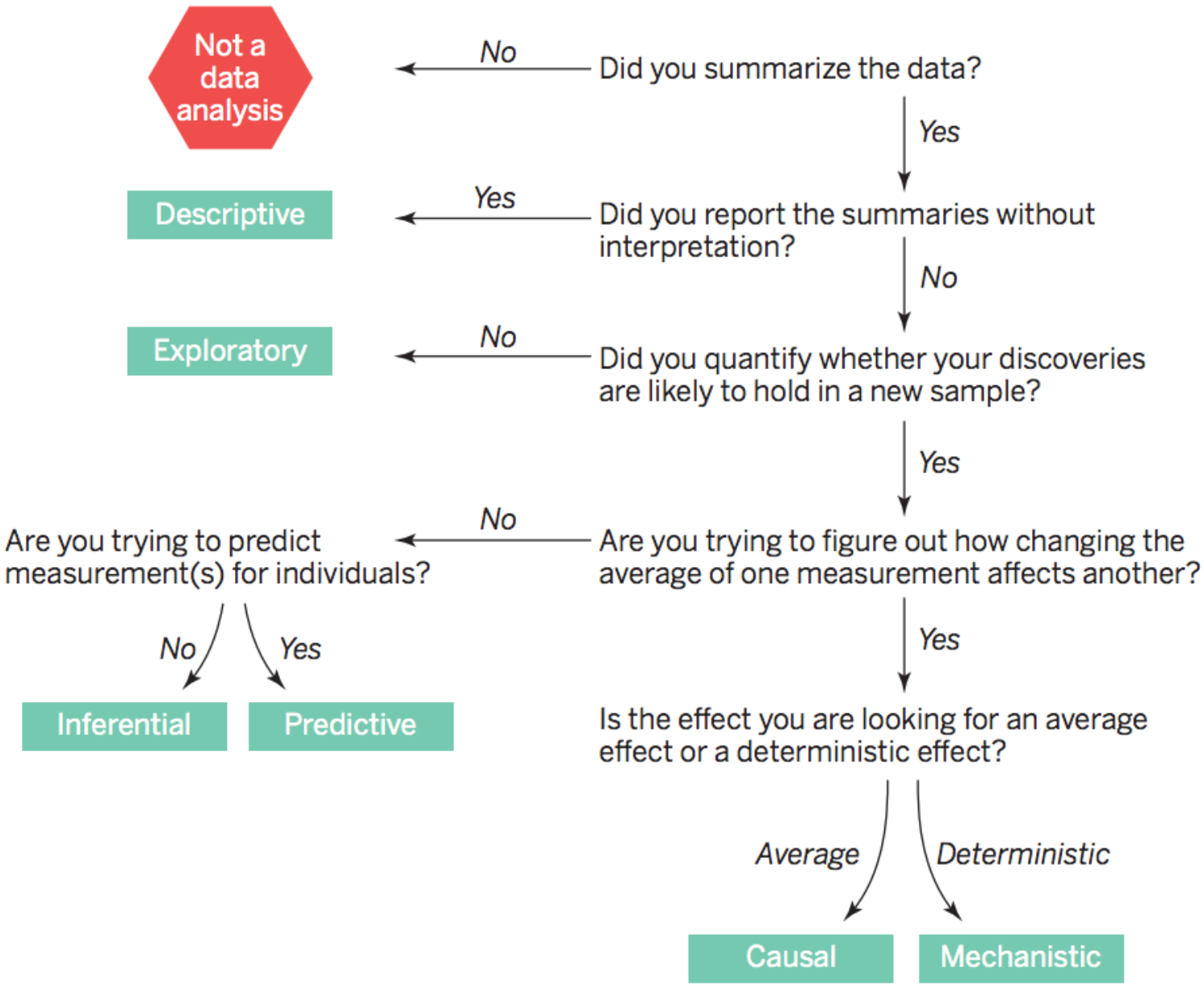
Those actionable insights can make a business more efficient and effective; they can widen the world's understanding of health sciences; and they can even filter back into computer science to create better datasets and customer experiences.

What is Data Science?

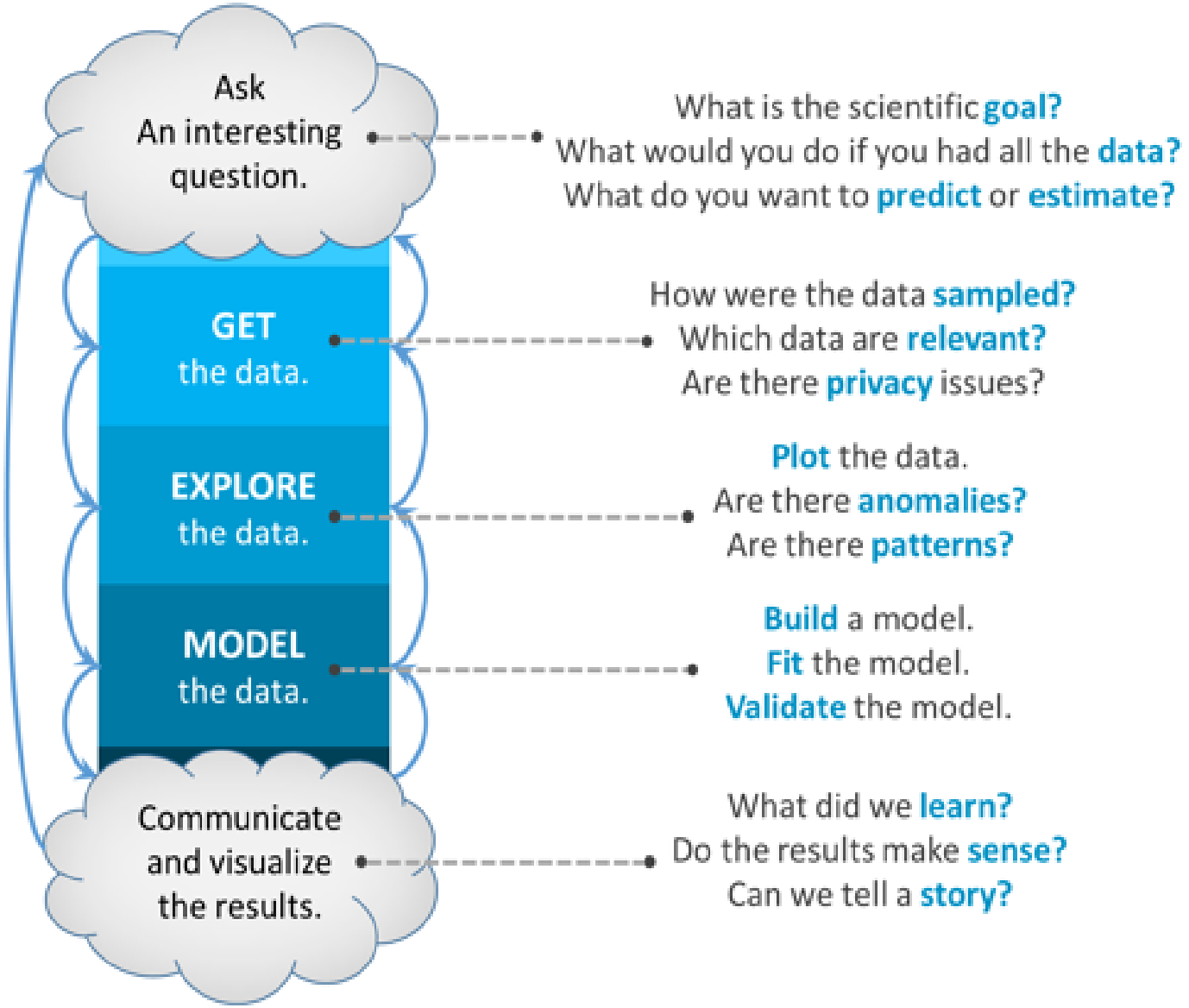



Cognitive analytics
Prescriptive analytics
Predictive analytics
Diagnostic analytics
Descriptive analytics
ADDED-VALUE CONTRIBUTION
COMPLEXITY

Data analysis flowchart



The
Data Science Process



https://www.kpipartners.com/services/data-science-and-ai

What do Data Scientists do?

- Using various statistical tools, a Data Scientist must develop models.
- Analytic objectives and approaches are planned and defined by the Data Scientists who collaborate with the internal consulting team.
- Data Scientists also formulate work plans to provide support – programming as well as analytical to internal consulting.
- Possess the most sought non-technical skill - communication skills
- In addition, based on the domain of expertise of the company, the specific requirements for the job will vary accordingly.

Data Scientist Responsibilities

- Ask the right questions to begin the discovery process
- Acquire data
- Process and clean the data
- Integrate and store data
- Initial data investigation and exploratory data analysis
- Choose one or more potential models and algorithms
- Apply data science techniques, such as ML, statistical modeling, and AI
- Measure and improve results
- Present result to stakeholders
- Adjust based on feedback
- Repeat the process to solve a new problem

The Typical Data Scientist 2020

Predominantly Male
(71%)

8.5 years
in the workforce overall

Bilingual

Python/R
(90%)

3.5 years
as a Data Scientist

Master/PhD
(80%)

365 √ DataScience

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Common Data Scientist Job Titles

- **Data scientists:** Design data modeling processes to create algorithms and predictive models and perform custom analysis
- **Data analysts**: Manipulate large data sets and use them to identify trends and reach meaningful conclusions to inform strategic business decisions
- **Data engineers:** Clean, aggregate, and organize data from disparate sources and transfer it to data warehouses.
- **Business intelligence specialists:** Identify trends in data sets
- **Data architects:** Design, create, and manage an organization’s data architecture

The Jobs Landscape in 2022

WORLD ECONOMIC FORUM

emerging roles, global change by 2022

133 Million

Top 10 Emerging

1. Data Analysts and Scientists
2. AI and Machine Learning Specialists
3. General and Operations Managers
4. Software and Applications Developers and Analysts
5. Sales and Marketing Professionals
6. Big Data Specialists
7. Digital Transformation Specialists
8. New Technology Specialists
9. Organisational Development Specialists
10. Information Technology Services

declining roles, global change by 2022

75 Million

Top 10 Declining

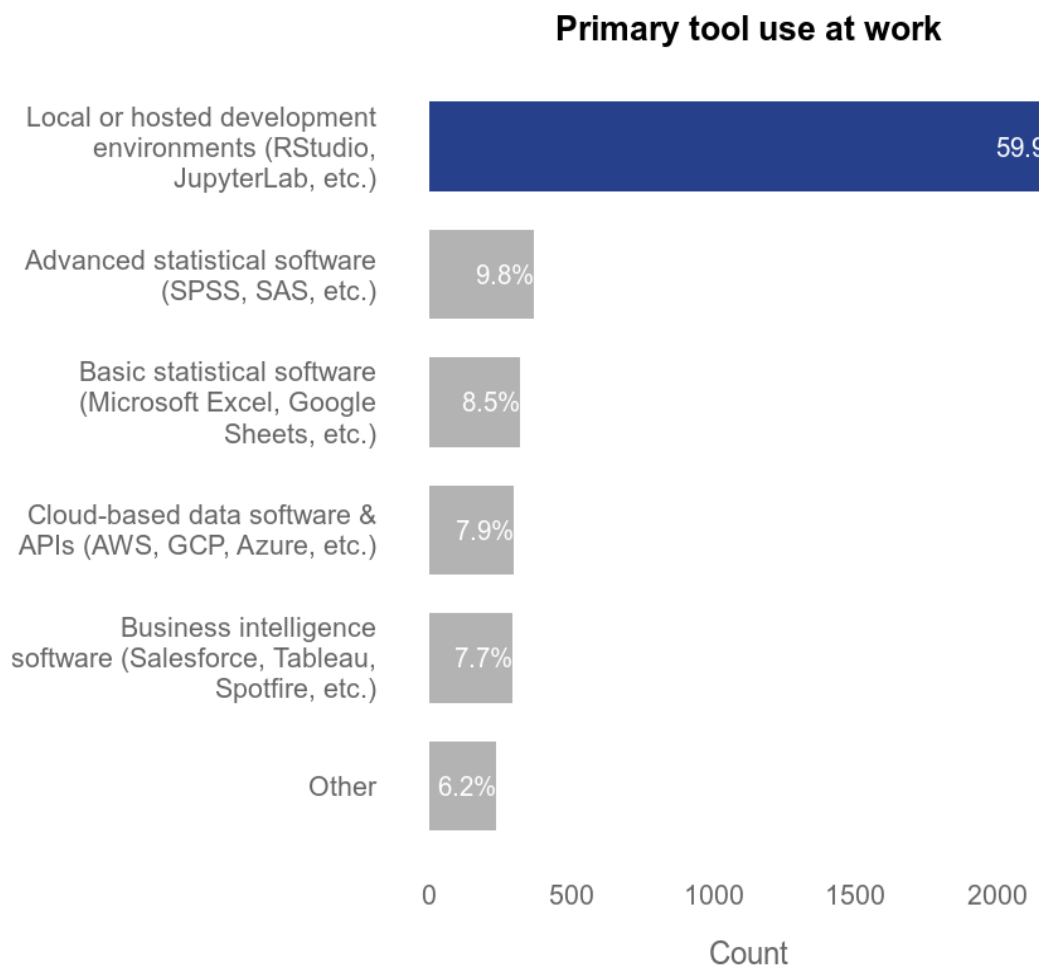
1. Data Entry Clerks
2. Accounting, Bookkeeping and Payroll Clerks
3. Administrative and Executive Secretaries
4. Assembly and Factory Workers
5. Client Information and Customer Service Workers
6. Business Services and Administration Managers
7. Accountants and Auditors
8. Material-Recording and Stock-Keeping Clerks
9. General and Operations Managers
10. Postal Service Clerks

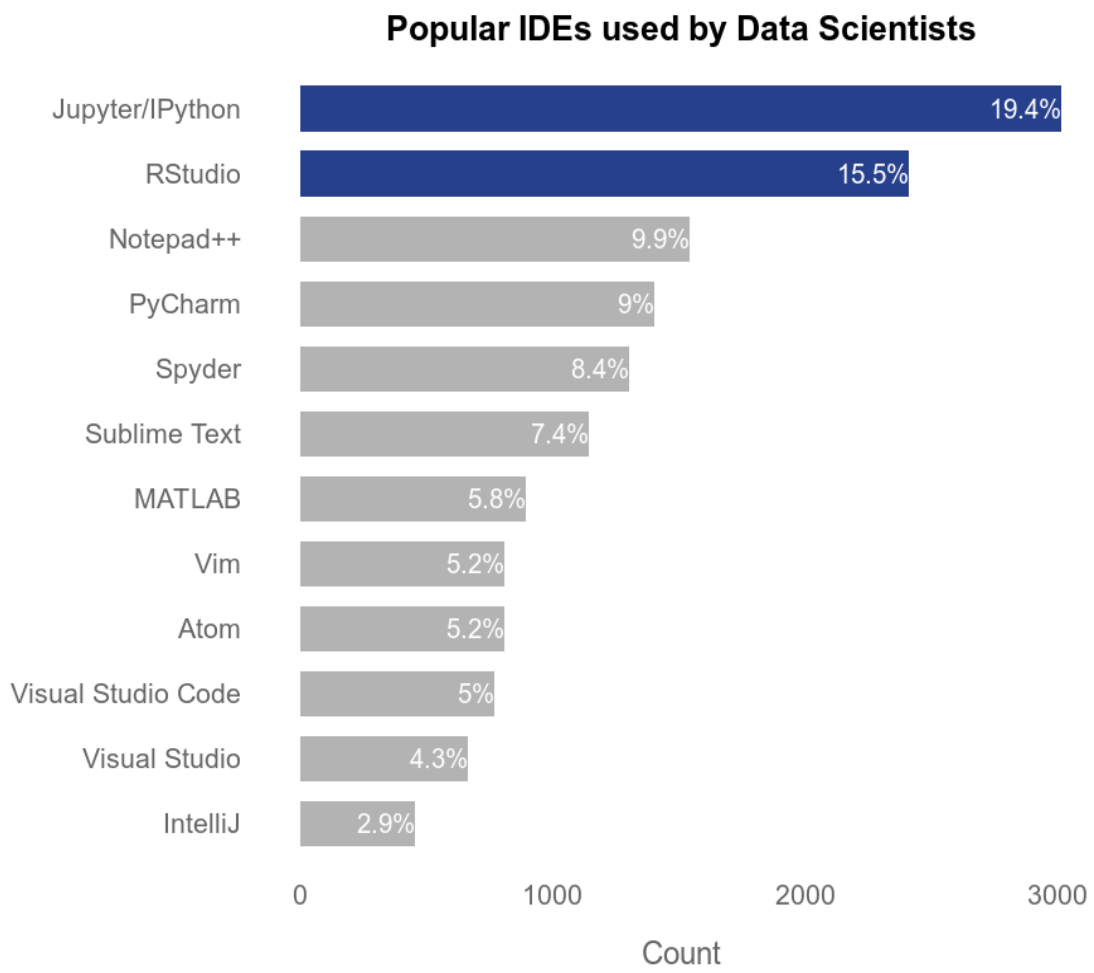
Source: Future of Jobs Report 2018, World Economic Forum

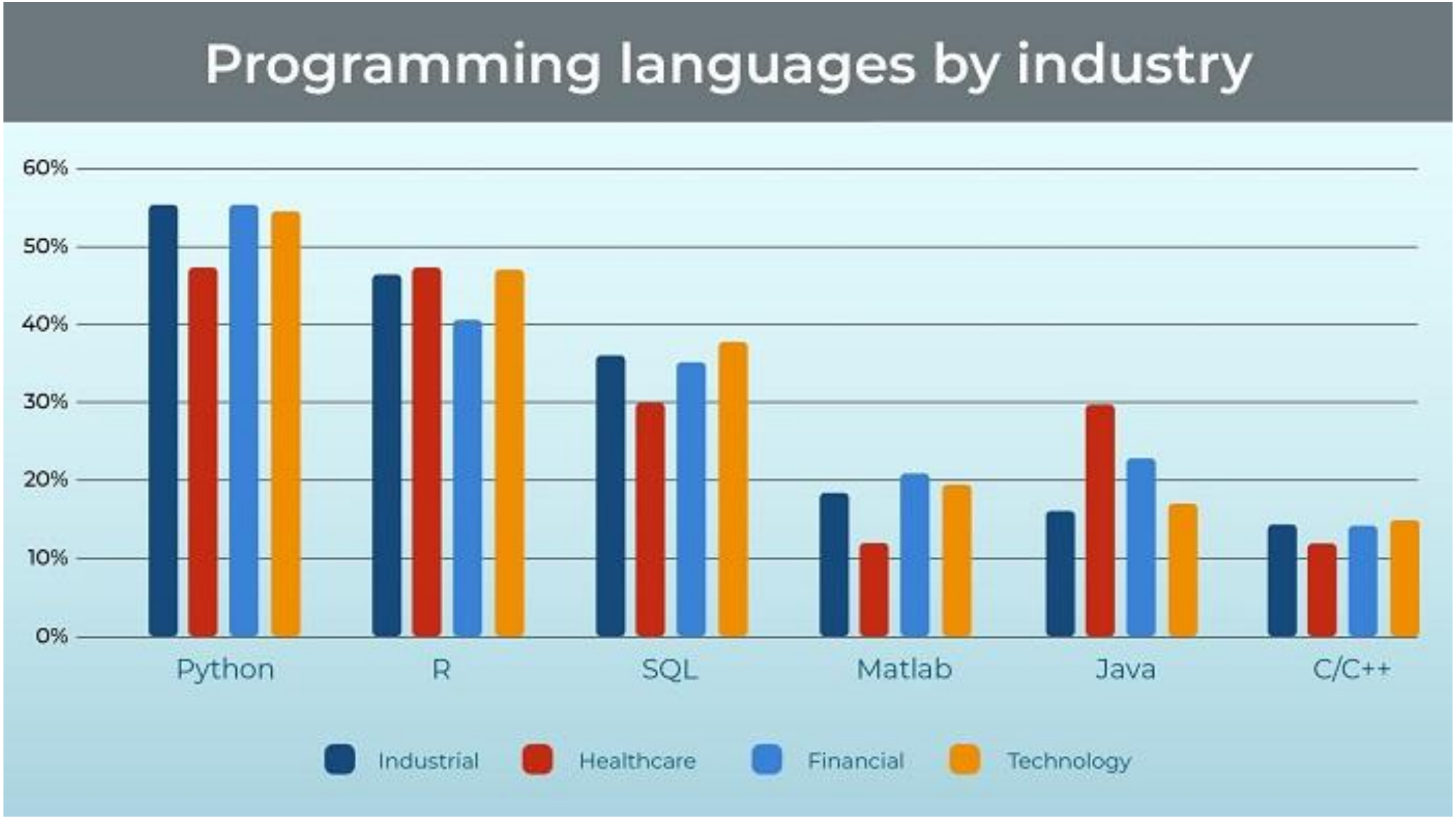
DataScience.US
A Data Professionals Community

According to Robert Half Technology's 2020 Salary Guide,

- Data Scientists earn an average annual salary between $105,750 and $180,250 per year.
- However, compensation can vary depending on location.
- Additionally, as data scientists gain experience, they often move into more senior positions with higher pay.
- These include:
  - Senior Data Scientist: $125,925
  - Data Science Manager: $135,401
  - Data Science Director: $157,273


Primary tool use at work
Local or hosted development environments (RStudio, JupyterLab, etc.)
59.9%
Advanced statistical software (SPSS, SAS, etc.)
9.8%
Basic statistical software (Microsoft Excel, Google Sheets, etc.)
8.5%
Cloud-based data software & APIs (AWS, GCP, Azure, etc.)
7.9%
Business intelligence software (Salesforce, Tableau, Spotfire, etc.)
7.7%
Other
6.2%
0
500
1000
1500
2000
Count


Popular IDEs used by Data Scientists
Jupyter/IPython
19.4%
RStudio
15.5%
Notepad++
9.9%
PyCharm
9%
Spyder
8.4%
Sublime Text
7.4%
MATLAB
5.8%
Vim
5.2%
Atom
5.2%
Visual Studio Code
5%
Visual Studio
4.3%
IntelliJ
2.9%
0
1000
2000
3000
Count



https://www.kdnuggets.com/2019/03/typical-data-scientist-2019.html

Tools

python

hadoop

tableau

Microsoft SQL Server

sas THE POWER TO KNOW.

TERADATA

hadapt

Apache Zeppelin

splunk>

R

Microsoft PowerPoint

KNIME

Excel

Java

Spotfire

APACHE Spark

rapidminer

Steps to a Data Scientist

- Develop skills in Algebra, Statistics, and Machine Learning
- Learn to love data
- Gain a thorough knowledge of databases
- Learn to code
- Master data munging, visualization, and reporting
- Work on real projects
- Look for knowledge everywhere
- Communication skills
- Compete
- Stay up-to-date with the data science community

Module 1
Learning Outcomes

- Definition of data science and what data scientists do
- Tools and algorithms used daily within the field
- Skills needed to be a successful data scientist
- The role of data science within a business

Module 2: What is Data Science, Machine Learning and Deep Learning?

- Overview of:
  - Data science
  - Machine Learning
  - Deep Learning
- Describe the data science process and how its components interact
- Define big data and know the challenges

BeVera

What is Data Science?

In its purest form, data science is the fourth paradigm of science, following experiment, theory, and computational sciences. The fourth paradigm is a term coined by Dr. Jim Gray in 2007.

Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.

How Enterprises Use Data Science

DATA SCIENCE INFORMS BUSINESS DECISIONS

IT HELPS DEFINE GOALS

IT HIGHLIGHTS UNSEEN OPPORTUNITIES

IT POWERS EVIDENCE-BASED DECISION MAKING

IT HELPS TESTING AND IMPROVING ON DECISIONS

Data Mining? Statistical Modeling? Predictive Analytics?

- **Predictive analytics** is the branch of advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses techniques from data mining, statistics, modeling, machine learning, and artificial intelligence.

- **Data mining** is an analytics process designed to explore data (usually large amounts of data – also known as "big data"). Big data is defined by the volume, velocity, variety, variability and veracity. Data mining is a part of data cleaning that prepares data from multiple sources for analysis. This provides a complete view of the customer interactions.

- Data analysis is the process of inspecting, cleaning, transforming and modeling data with the objective of discovering useful information. **Statistical Analysis** enables to validate assumptions, hypothesis, and test them using standard statistical models.

- **Predictive modeling** provides the ability to automatically create accurate predictive models about future.

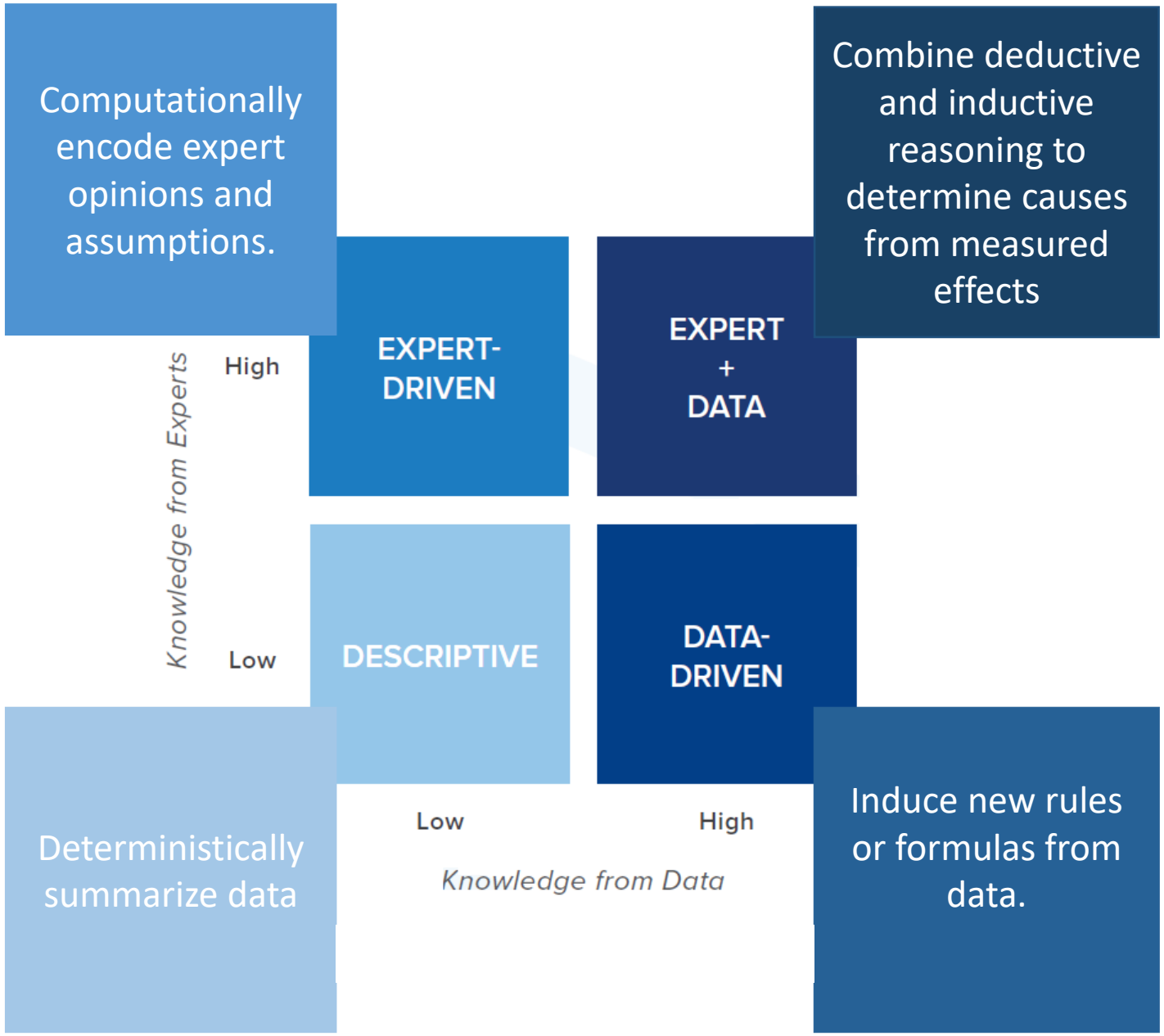
4 Levels of Knowledge Model

https://www.elderresearch.com/consulting-services/ten-levels-of-analytics

Sources of knowledge whether data or expert are independent. Combine data- and expert-driven approaches to maximize insight and value.

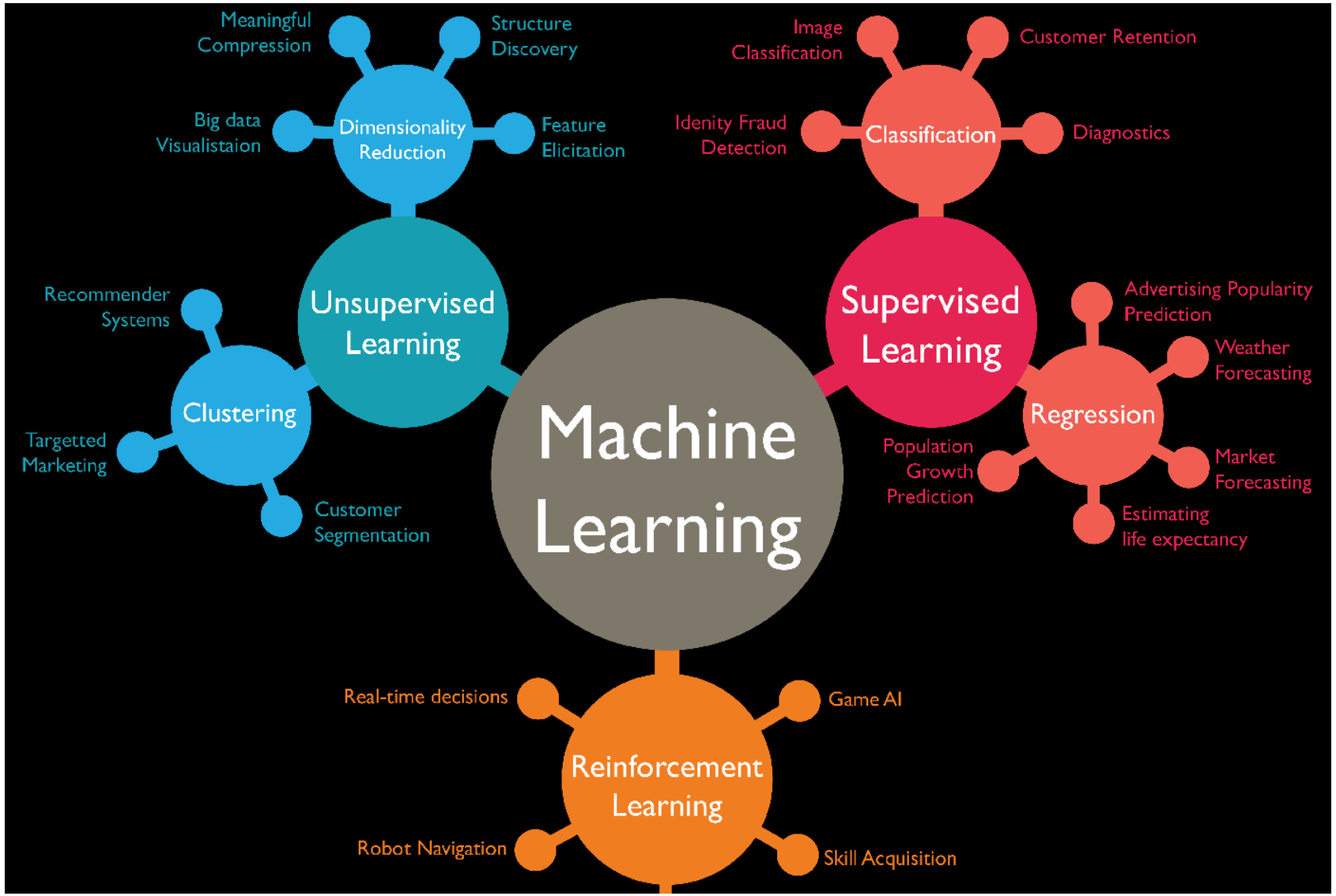
Expert-driven approaches is preferred if:

- an accepted premise
- the data is filtered
- poorly represents the full situation



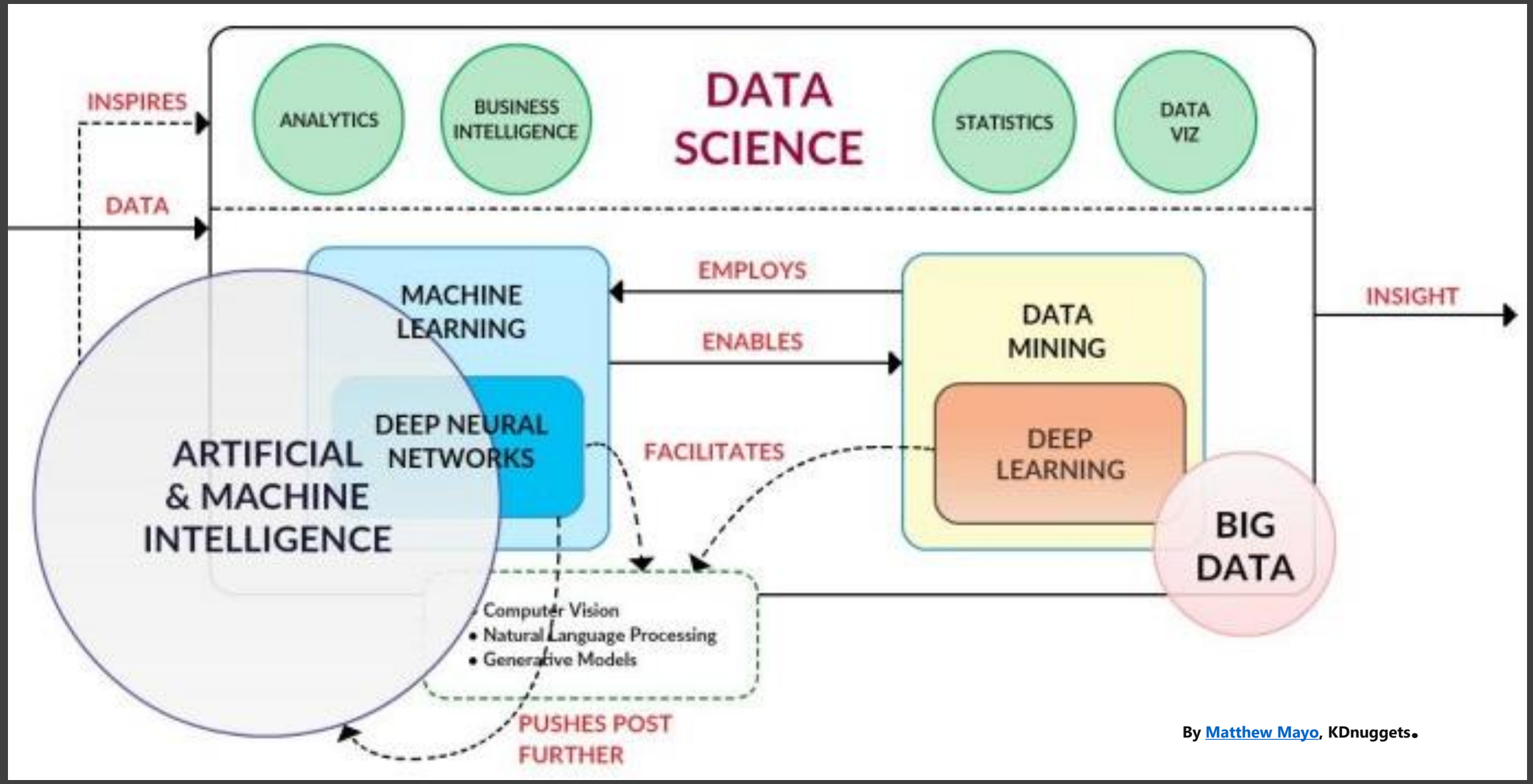
Data –driven modeling:

- is inductive
- graded superior to expert-driven ones.
- allow unknown rules or relationships to be discovered from the data
- are less susceptible to the biases and misconceptions common to human reasoning.



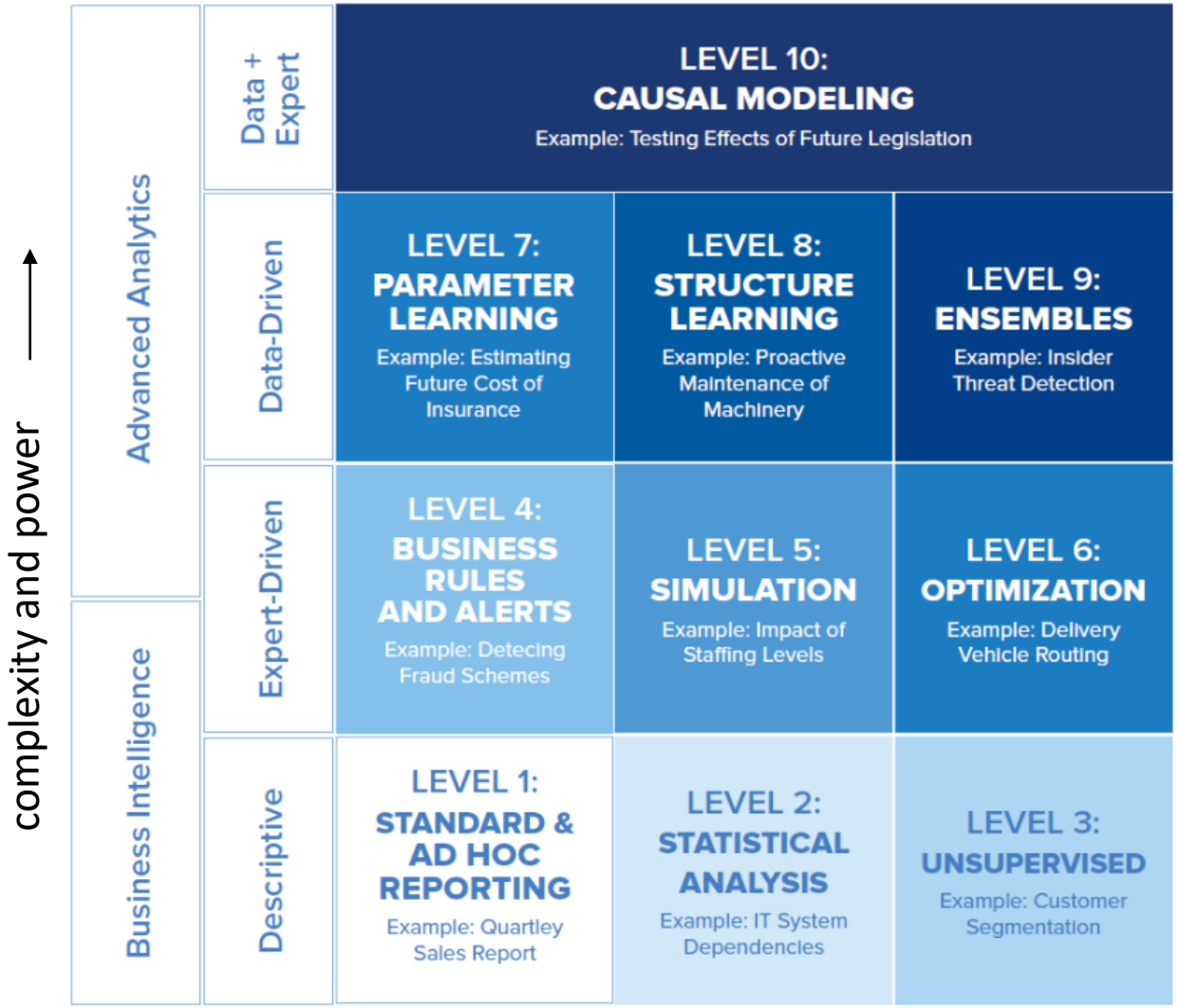
What is Machine Learning (ML)?

It is a class of algorithms that allow computers to perform pattern matching extremely efficiently.



By Matthew Mayo, KDnuggets.

10 Levels of Analytics

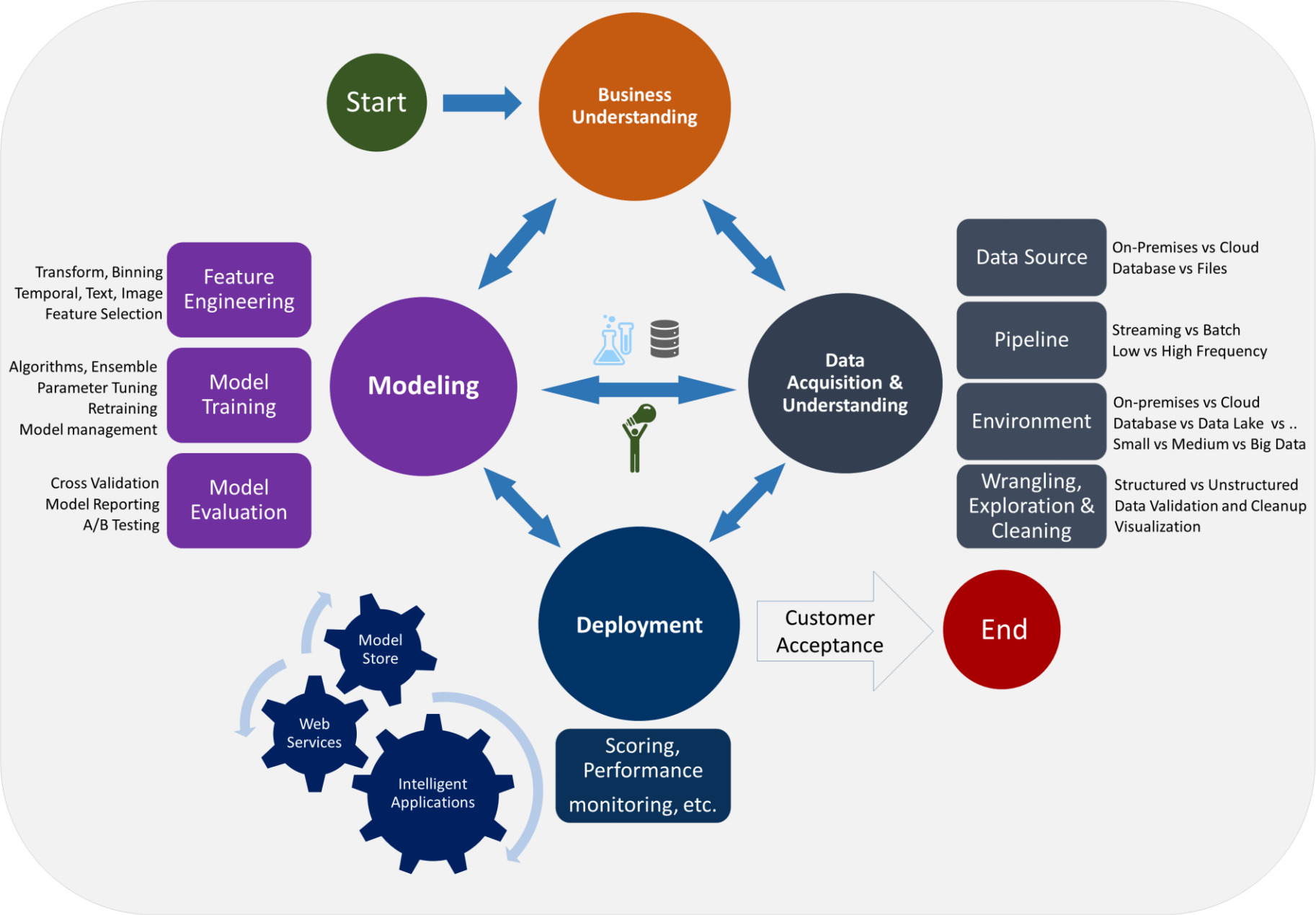


Ten Levels of Analytics

complexity and power →

https://www.elderresearch.com/consulting-services/ten-levels-of-analytics

Data Science Lifecycle



Big Data and Analytics

- Surprisingly, **99.5% of collected data never gets used or analyzed**. So much potential wasted!
- **Less than 50% of the structured data** collected from IoT is used in decision making.
- Predictive analytics are becoming more and more crucial for success. **79% of executives believe that failing to embrace big data will lead to bankruptcy**. This explains why **83% of companies** invest in big data projects.
- Fortune 1000 companies can gain more than **$65 million** additional net income, only by increasing their data accessibility by **10%**.
- Healthcare could also vastly benefit from big data analytics adoption. As much as **$300 billion** can be saved yearly!
- Companies that harness big data's full power could **increase** their operating margins by **up to 60%**!

“In the coming decade, data will be at the heart of transformative technologies that will affect every sector, creating new possibilities for healthcare, financial services, education and industrial manufacturing.”

— Alok Ohrie, President and MD of Dell Technologies, India

How is Machine Learning the Driving Force behind the Future of Data Science?

Data Science is expanding due to the immense contributions made by machine learning. It has improved the data science scenario in the following ways –

- 1. Advanced Personalizations
- 2. Giving Advanced Search Engine Results to the User
- 3. Code Free Environments
- 4. Quantum Computing

Article Read:
”Strengths of Employing Data Science in Healthcare”

https://www.iotforall.com/data-science-healthcare

BIG DATA GLOBAL MARKET SIZE FORECAST

by 2025 **$250* billion**

According to MarketsandMarkets, Adroit Market Research

by 2021 **$70* billion**

According to Wikibon, Frost & Sullivan, MarketsandMarkets, Statista

** - average assumption*

WHAT HAPPENS EVERY MINUTE

via Internet Live Stats

6,123 TB
TRAFFIC PRODUCED BY USERS

84,000
INSTAGRAM PHOTOS UPLOADED

5,200,000
GOOGLE SEARCHES

305,000
SKYPE CALLS

185,000,000
E-MAILS SENT

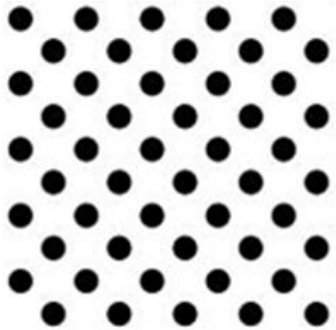
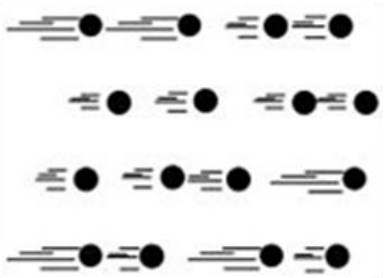
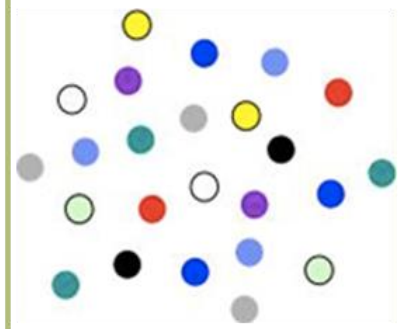
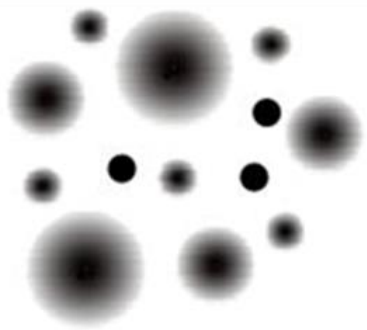
Big Time Big Data Statistics

- The big data analytics market is set to reach **$103 billion by 2023.**
- Poor data quality costs the US economy **up to $3.1 trillion yearly**.
- In 2020, **every person will generate 1.7 megabytes** in just a second.
- Internet users generate about **2.5 quintillion bytes of data each day.**
- **95% of businesses** cite the need to manage unstructured data as a problem for their business.
- **97.2% of organizations** are investing in big data and AI.
- Using big data, **Netflix saves $1 billion per year** on customer retention.

https://techjury.net/blog/big-data-statistics, March 2021

“Big data is a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis.” (Lisa Arthur, CMO Network, 8/15/2013). “Big Data” is data whose scale, diversity, and complexity require new architecture, new tools, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

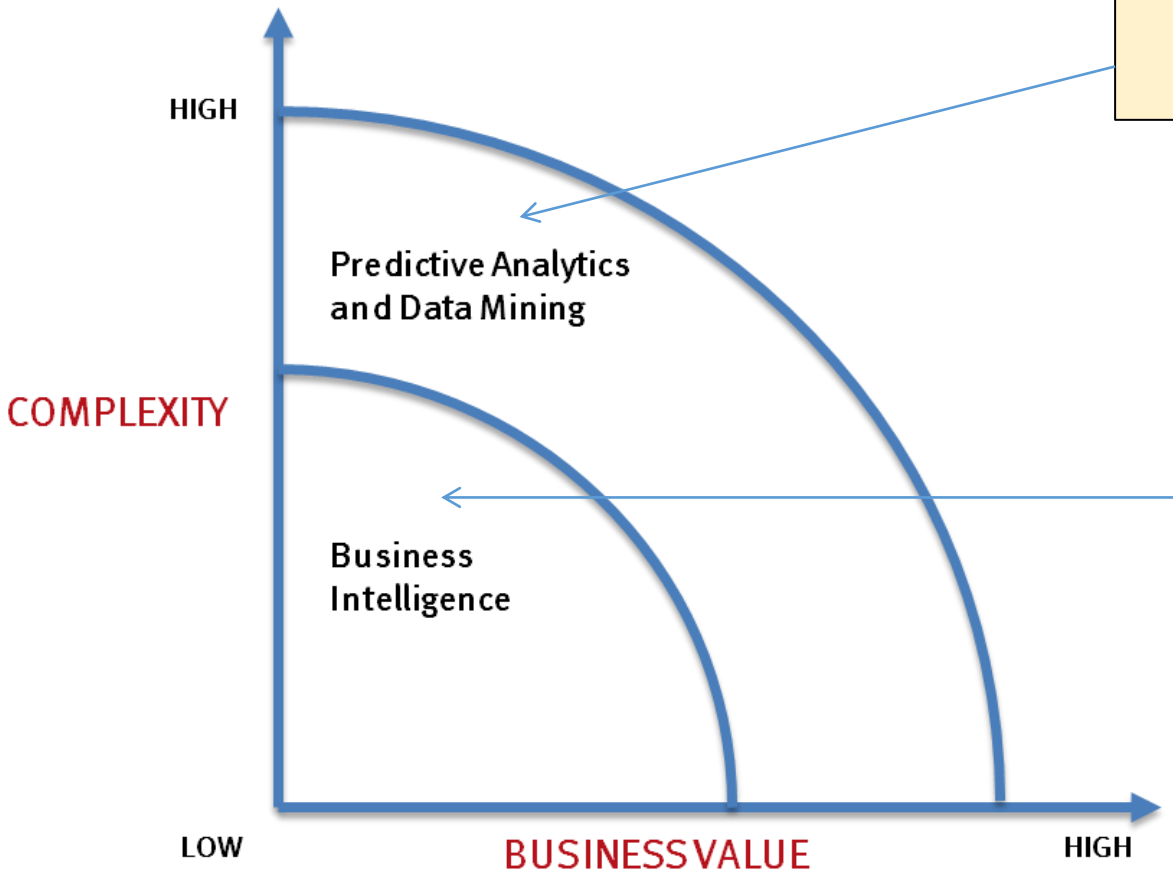
“Vs”of Big Data

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	**Streaming data, requiring milliseconds to seconds to respond**	**Structured, unstructured, text, multimedia,…**	**Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations**	**Business models can be associated to the data**

Adapted by a post of Michael Walker on 28 November 2012

Big data characteristics: Data structures

What's driving Big Data



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

Common questions

- What if…
- What's the optimal scenario for our business?
- What will happen next? What if these trends continue? Why is this happening?

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Common questions

- What happened last quarter?
- How many did we sell?
- Where is the problem? In which situation?

Article Read: “The 10 Vs of Big Data”

https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Traditional vs. big data

- Challenges
- Advantages

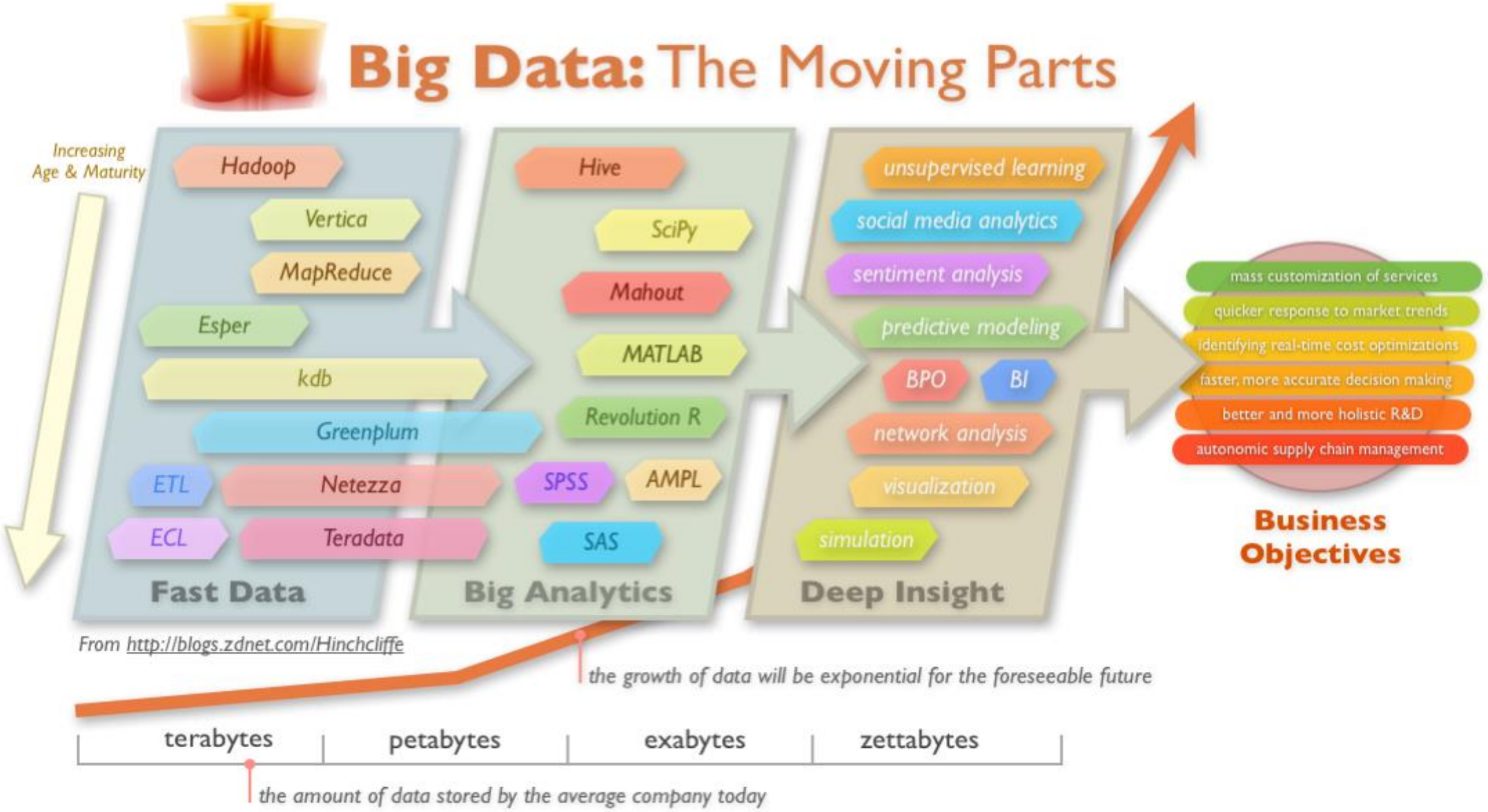
The biggest data silo at hand, the stronger the need for the organization and tidiness

To be effective in …healthcare industrial research, big data in the life science domain needs to be organized, well annotated, catalogued, correlated and integrated.

The heterogeneity of biological data is certainly a challenge. Possibility to also use data to refine data.

Big Data Challenges
Data Integration
Scalability
Security & Privacy of Data
Data Mining Techniques
Data processing
Data Visualization
Data Transmission
Data Quality
Lack of Standardization
Storage of data

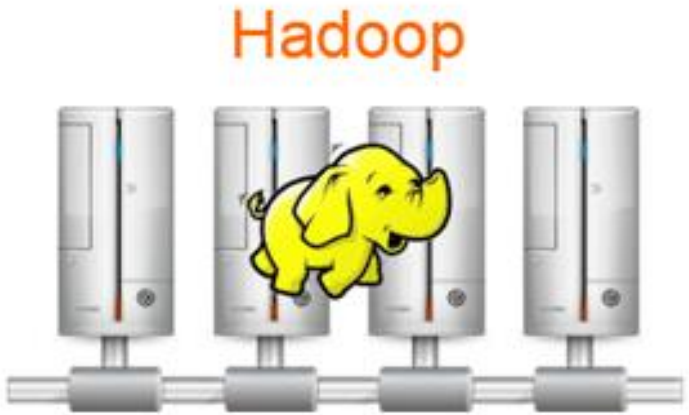
Big Data Technology

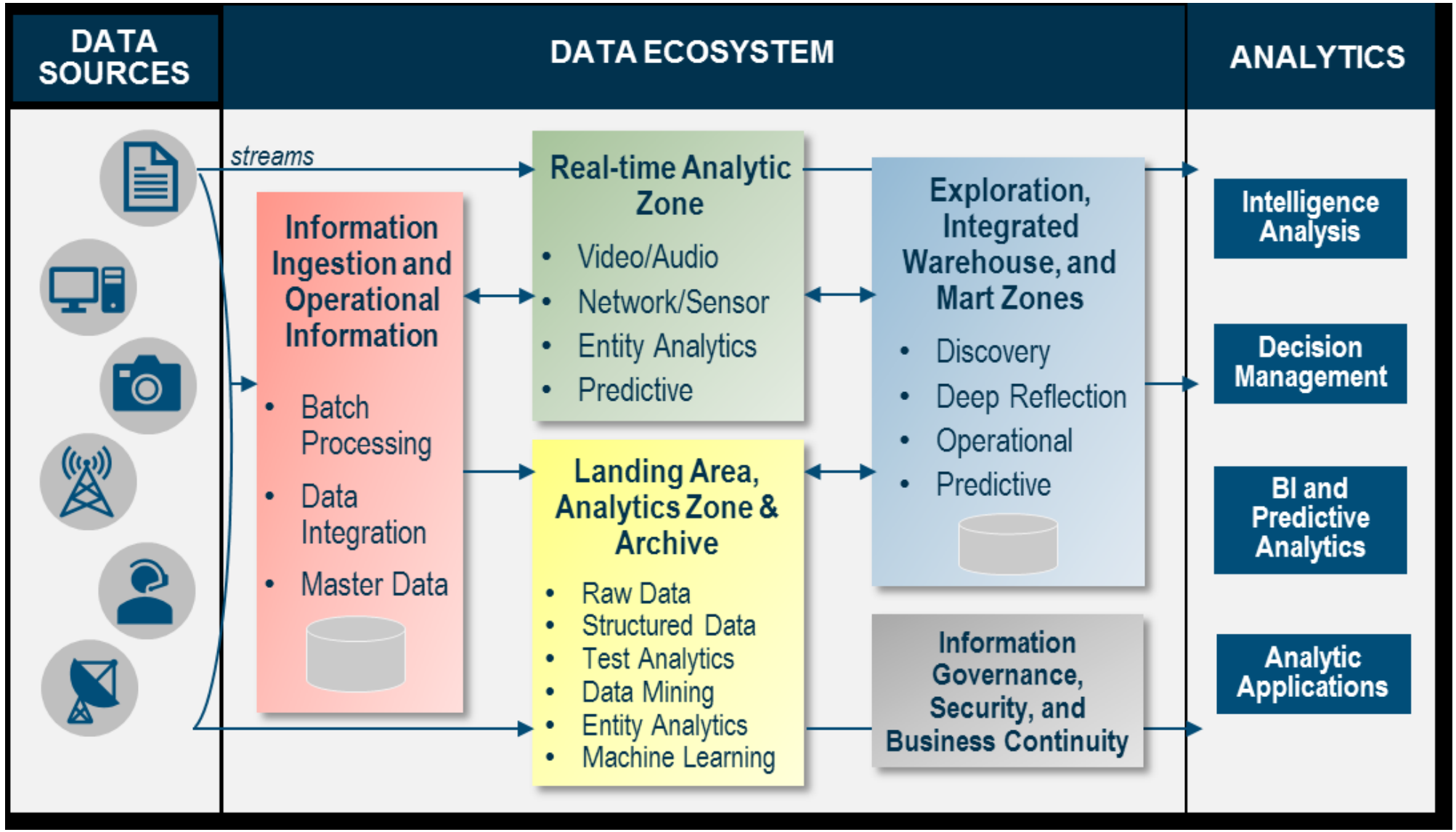


Types of tools used in Big-Data

- Where processing is **hosted**?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is **stored**?
  - Distributed Storage (e.g. Amazon S3)
- What is the **programming model**?
  - Distributed Processing (e.g. MapReduce)
- How data is **stored & indexed**?
  - High-performance schema-free databases (e.g. MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

SQL


Hadoop



https://www.analytics8.com/blog/7-elements-of-a-data-strategy/

Module 3: Introduction to Data Science Process – CRISP-DM

- Solutions Methodologies
  - Scientific Research Method
  - Operations Research Method
  - Waterfall Method
  - Agile Method
  - CRISP-DM
- Exploration and Discovery

- CRISP - DM process
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

BeVera

In descriptive Data Science Methodology, the framework is geared to do three things:

Understand the question at hand.

Select an analytic approach or method to solve the problem.

Obtain, understand, prepare, and model the data.

“Before doing your research I want to share my dogmas. Methodology is there to make it credible that you have planned and carried through your study as well as analysed and drawn conclusions in a way that we can rely on what you write.”

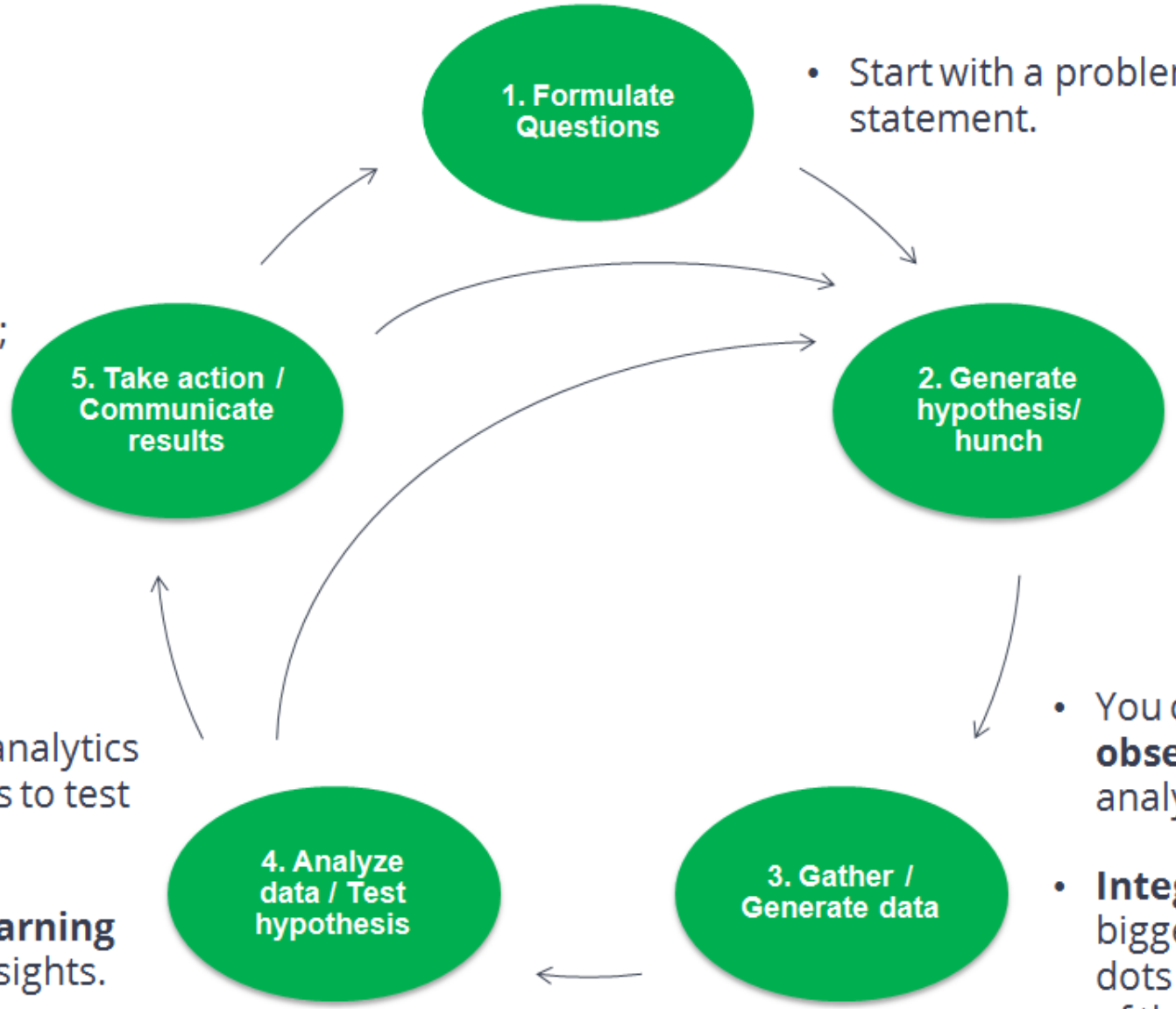
Professor Christer Karlsson,

Professor of Innovation and Operations Management and Academic Director for the Competitiveness platform at Copenhagen Business School, Denmark

General macro-methodology categories are covered:

1. The scientific research methodology;
2. The operations research project methodology;
3. The waterfall methodology;
4. The agile method;
5. And the cross-industry standard process for data mining (CRISP-DM) methodology

Getting Insight from Data: The Scientific Method



- Start with a problem statement.

- What are your hunches / hypotheses?
- Be sure your hypotheses are testable.

- You can use **experimental** or **observational** approach to analyzing data.
- **Integrate** your data silos to ask bigger questions; connect the dots and get a 360 degree view of the phenomenon you're studying.

- Employ **Predictive** analytics / Inferential statistics to test hypotheses.
- Employ **machine learning** to quickly surface insights.

- Implement your findings; inform decision-makers; optimize algorithms
- Use **Prescriptive** analytics to guide course of action.

BUSINESS BROADWAY

The Scientific Method helps you put together experiments, use data to find conclusions and interpret them. In short, the Scientific Method is a step-by-step process:

- Observe. Use your senses and take notes about the situation.
- Ask a question about the problem at hand.
- After that, form a hypothesis. A hypothesis is a potential explanation to your question.
- Predict what the hypothesis may lead to and conduct an experiment to test it out.
- Analyze the data to draw a conclusion from your findings.
- Share your results.

Commonly, two types of inferences: Deductive and Reasoning.

Deductive reasoning moves from a general premise to a more specific conclusion (what you predict you'll see in your data).

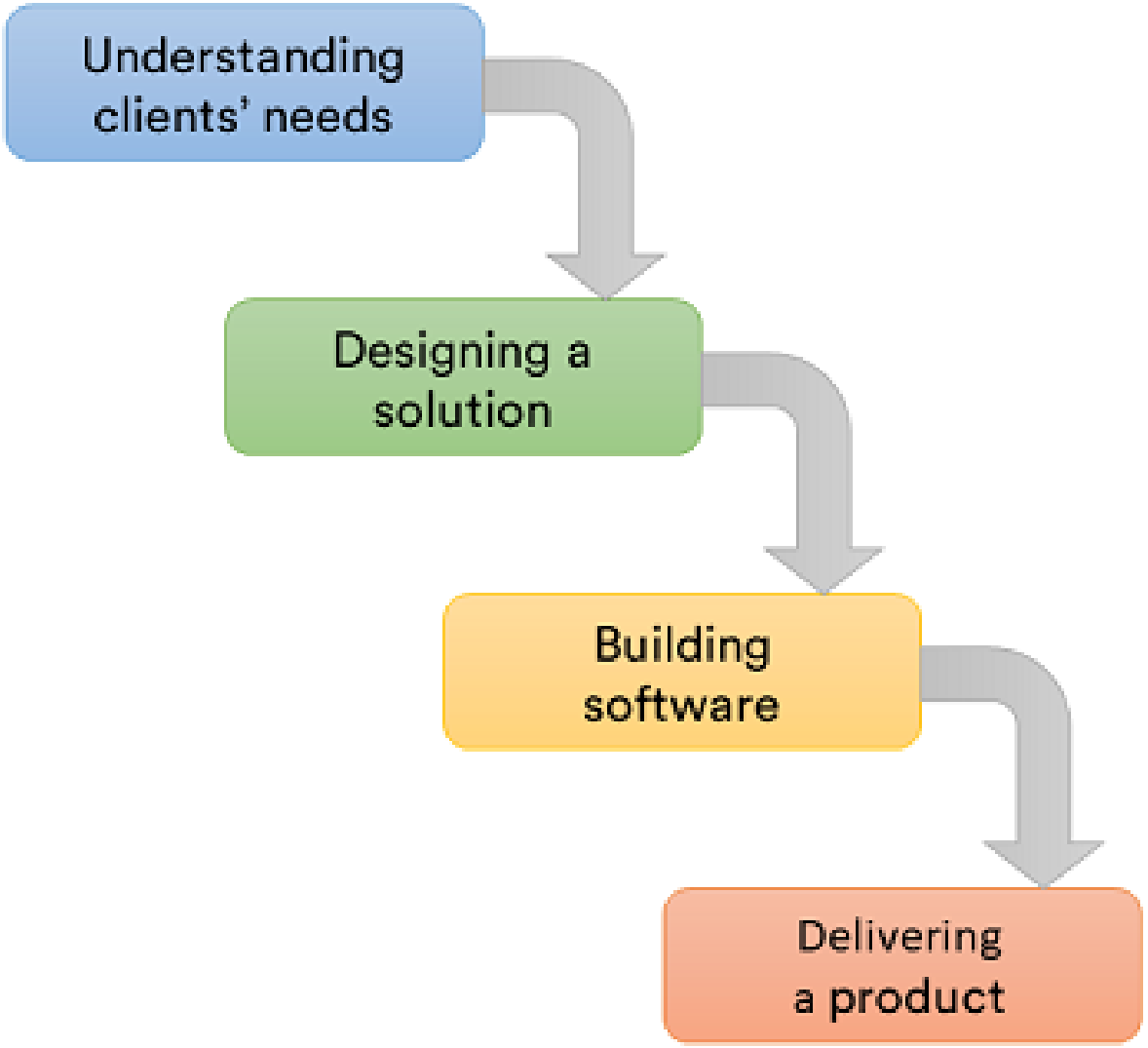
Inductive reasoning moves from data to a general conclusion.

1. Business knowledge	A. Formulate the right questions B. Generate hypothesis C. Gather data D. Communicate results
2. Technology/Programming skills	C. Gather/generate data E. Analyze data or test hypotheses
3. Statistics/Math skills	C. Gather data E. Analyze data or test hypotheses D. Communicate results.

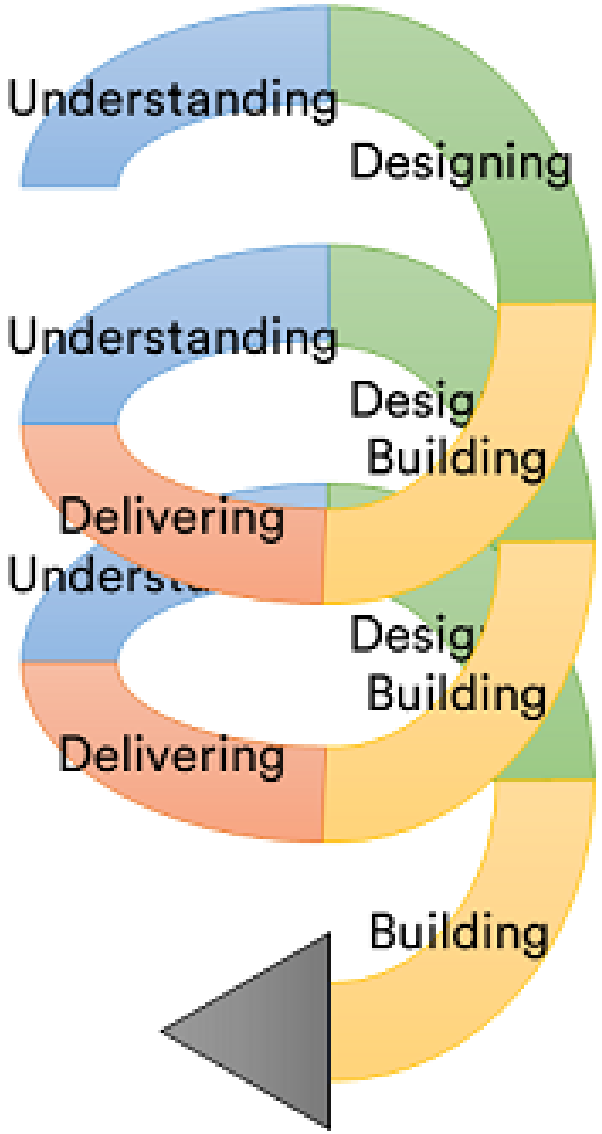
Operations Research Method

1. Business Problem
2. Theory
3. Model
4. Hypothesis
5. Methodology
6. Conclusion
7. Business Impact

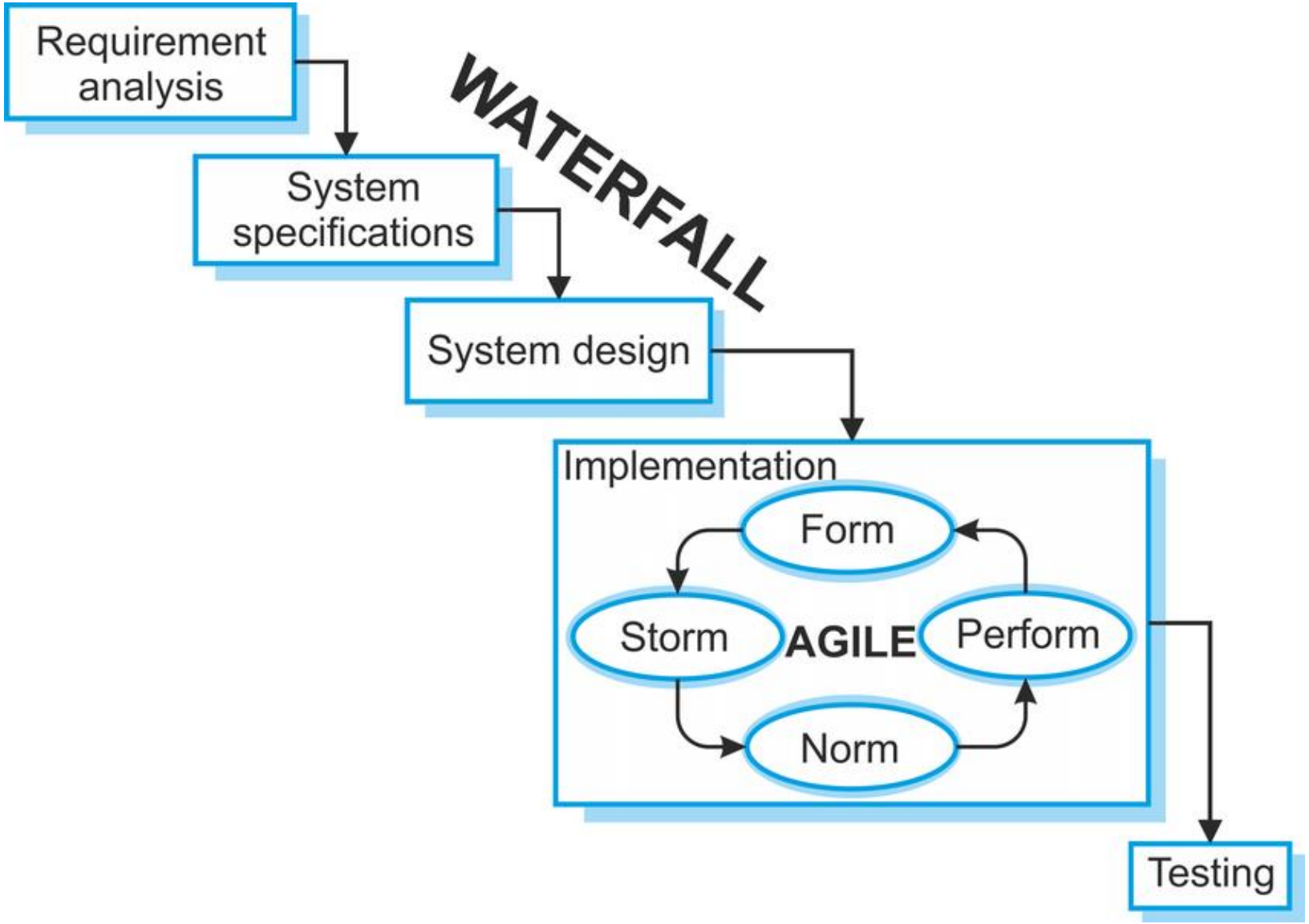
‘Waterfall’ process



‘Agile’ process

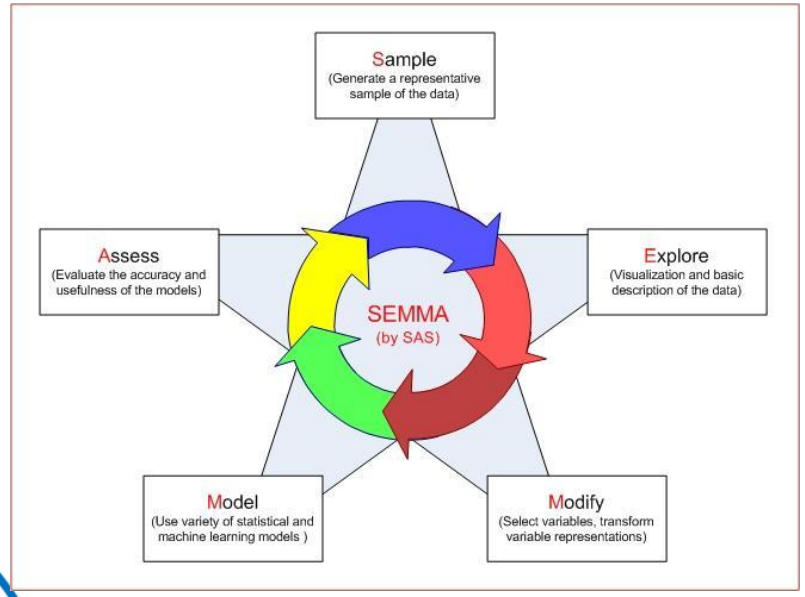


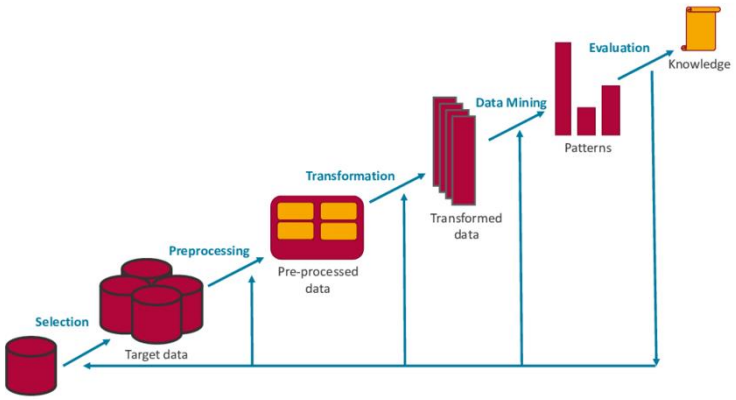
Hybrid approach



Motivation:

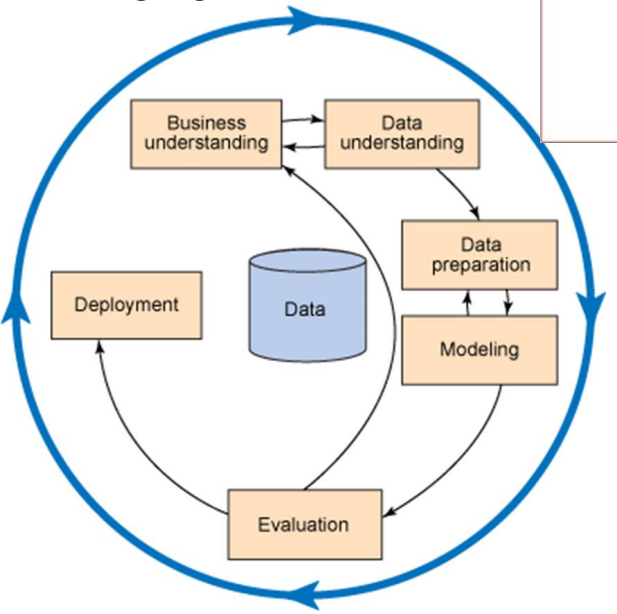
It’s all about the different methods used in data science.

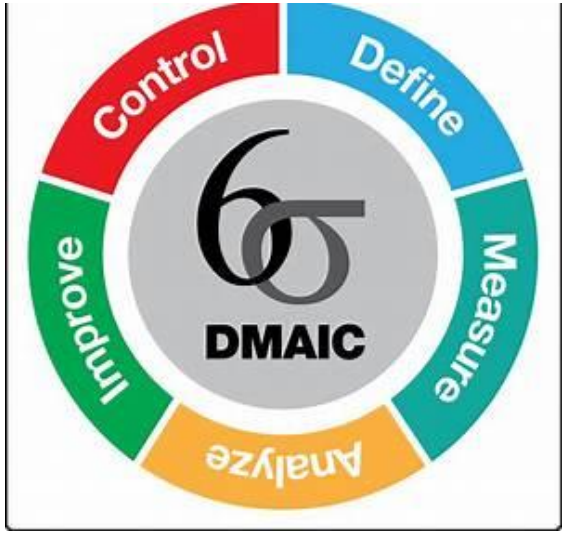




KDD

CRISP-DM





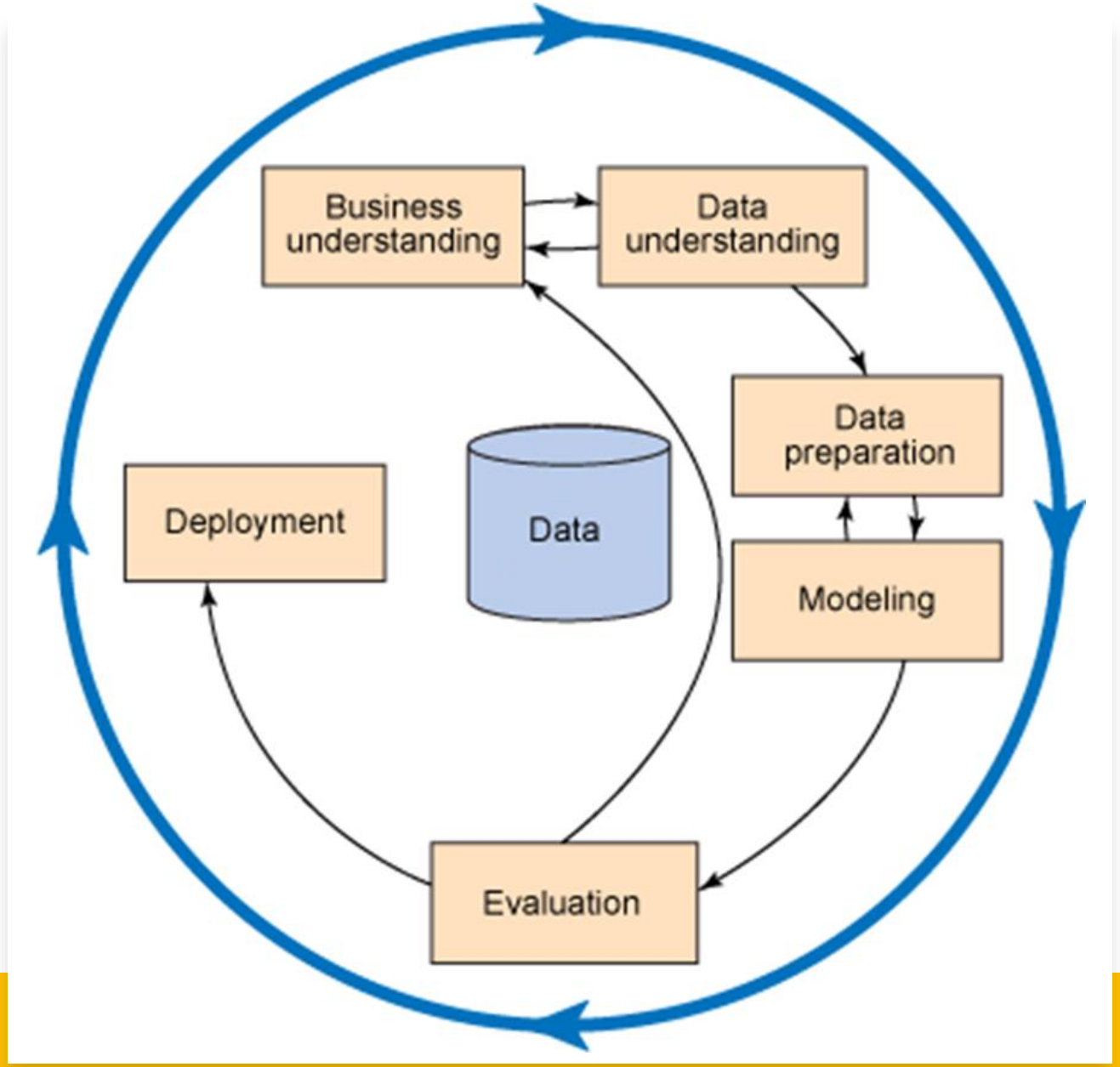
Standards

Different Life Cycles of Data Science

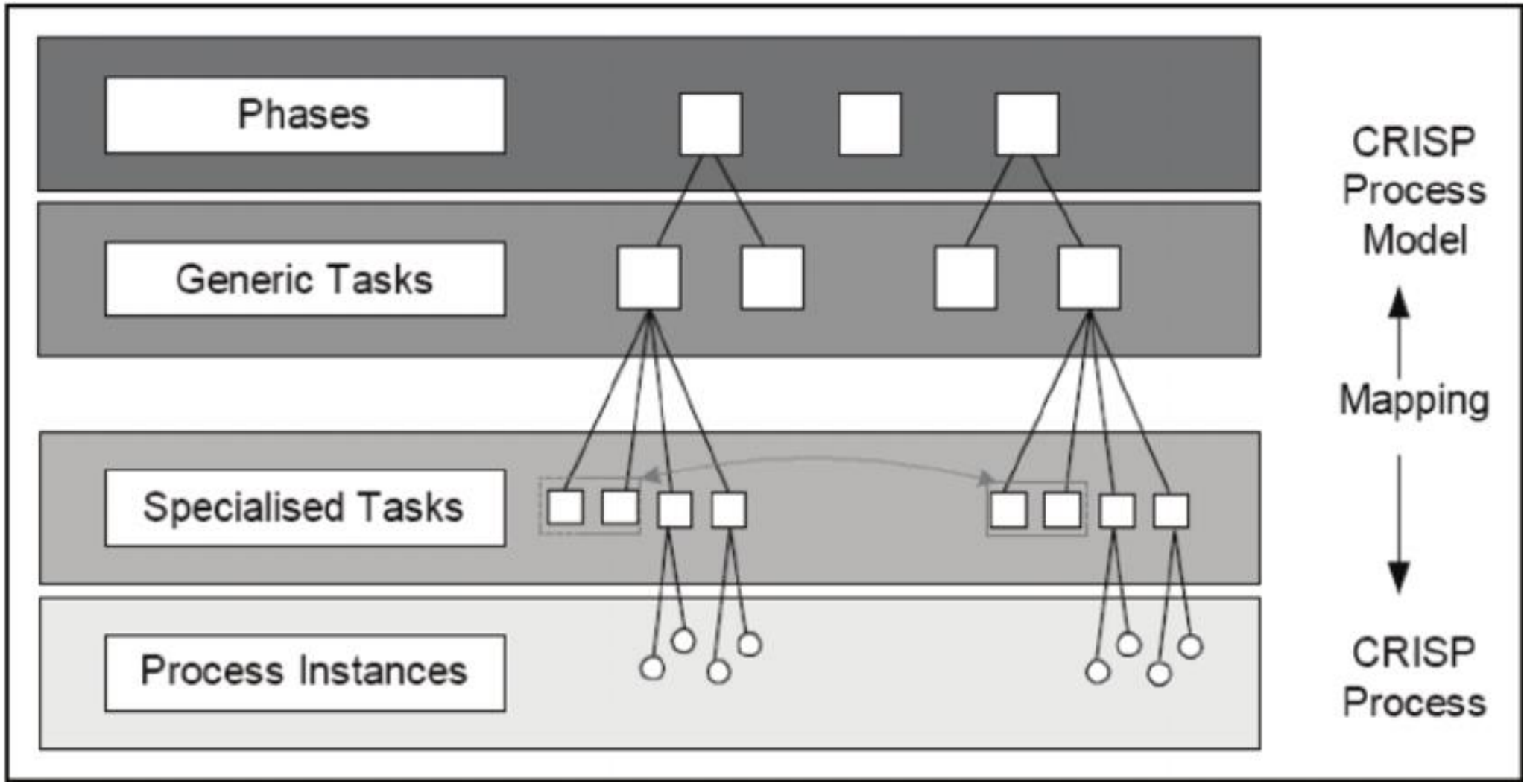
Cross Industry Standard Process for Data Mining Methodology (CRISP-DM)

Basic reasons why the CRISP-DM process is important:

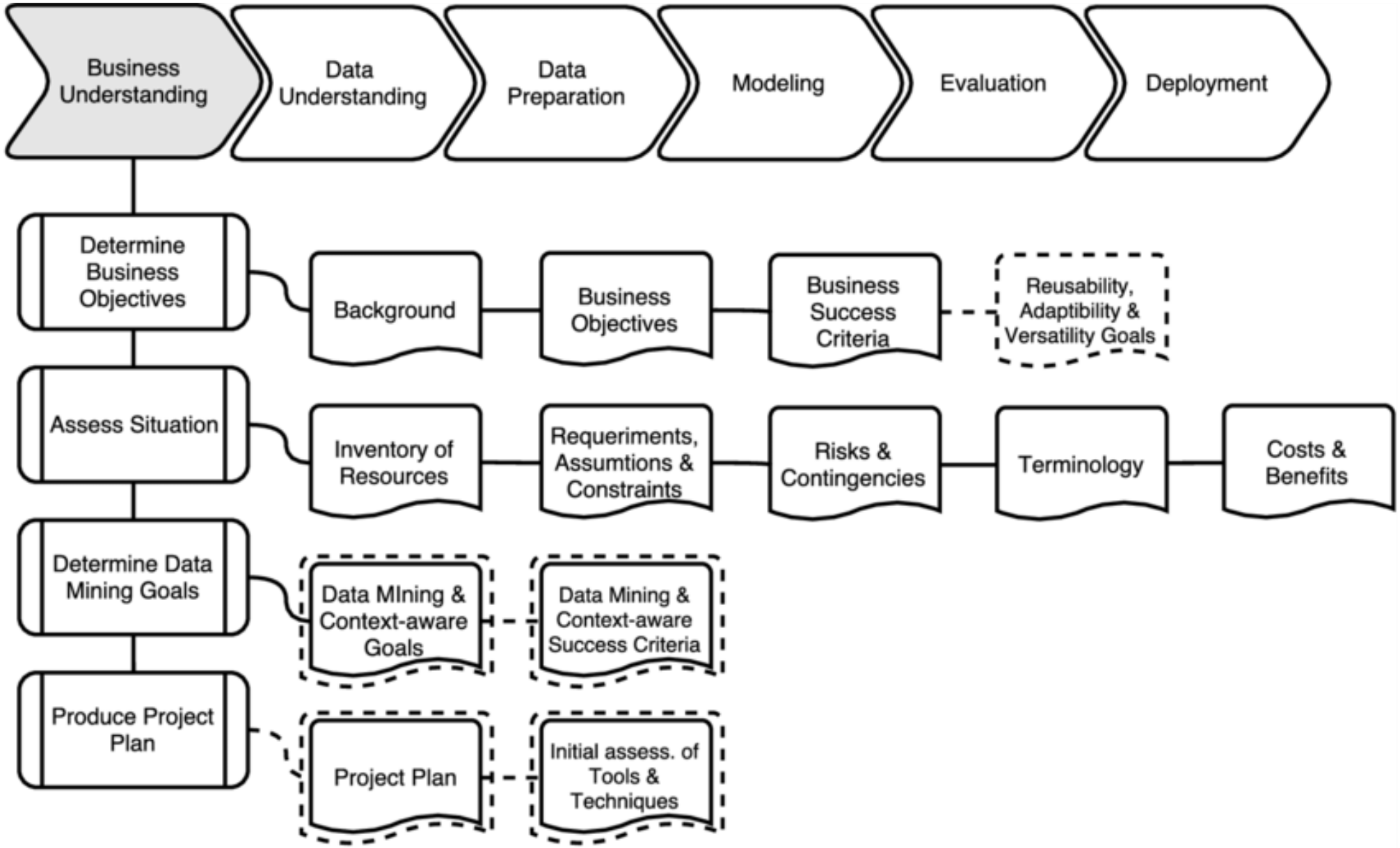
1. Flexibility
2. Functionality
3. Strategy



CRISP-DM Methodology



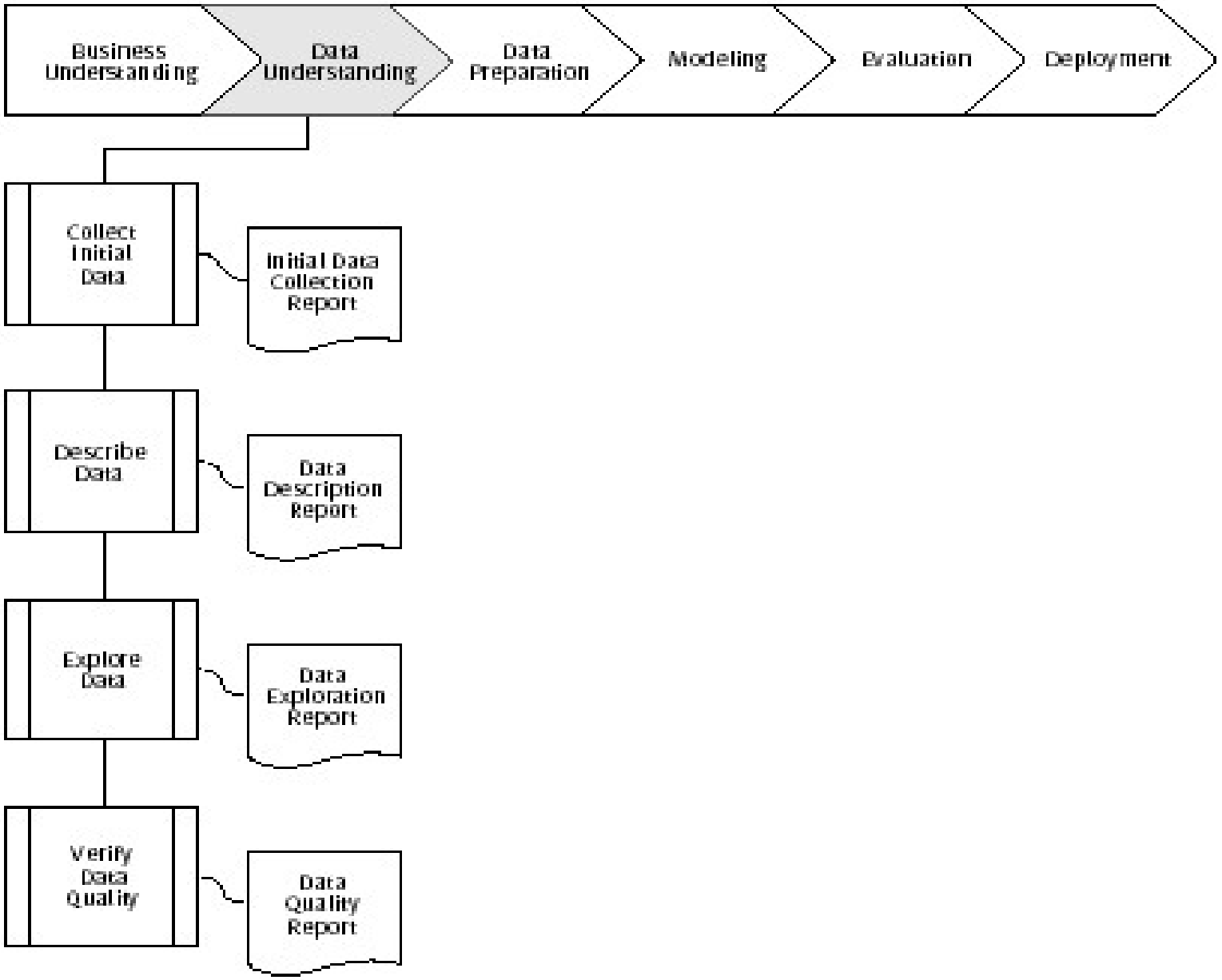
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Reusability, Adaptibility & Versatility Goals
Assess Situation
Inventory of Resources
Requeriments, Assumtions & Constraints
Risks & Contingencies
Terminology
Costs & Benefits
Determine Data Mining Goals
Data MIning & Context-aware Goals
Data Mining & Context-aware Success Criteria
Produce Project Plan
Project Plan
Initial assess. of Tools & Techniques

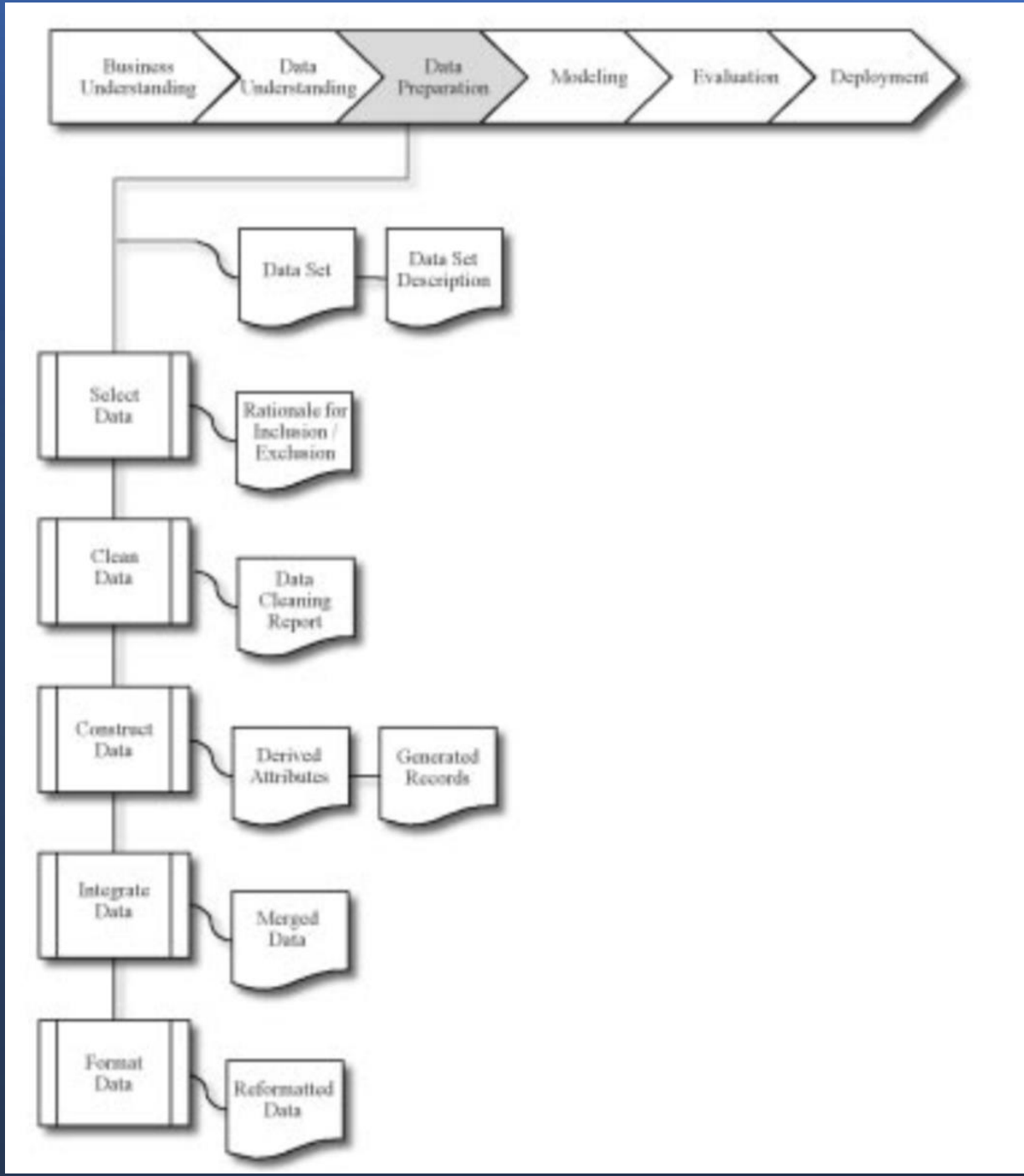
Determining Business Objectives

1

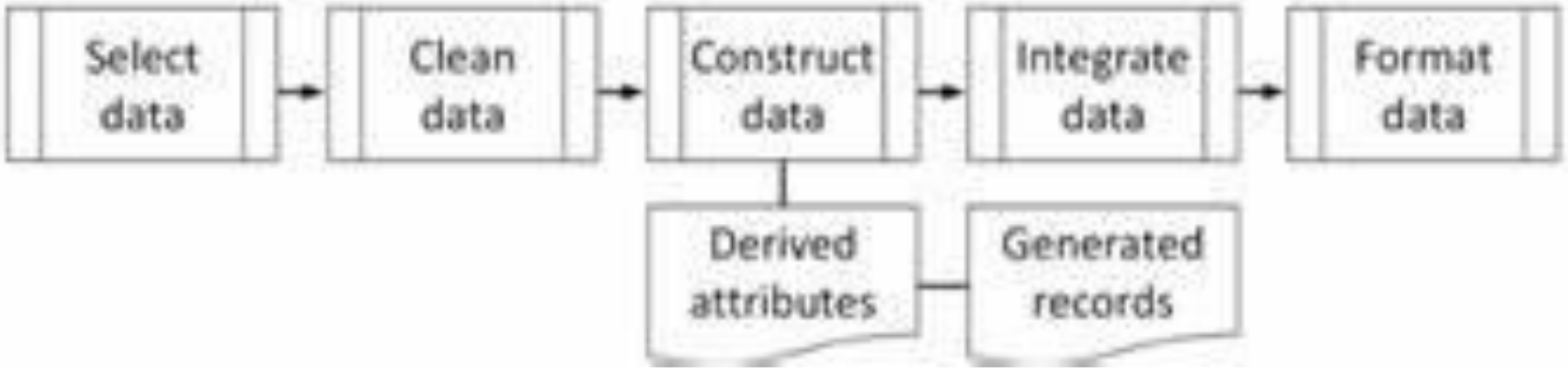
- Receive and refine the business problem
- Determine whether there is an analytics solutions
- Identify stakeholders in the problem and solution
- Define the initial set of business benefits
- Obtain stakeholder agreement on the problem statement


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report

- The data understanding step is often further detailed in the models reviewed and referred to using different names. It tends to include data acquisition, which is functional to data quality analysis.
- Data discovery, data prospecting, data requirements,, select target data set/ sample data or data acquisition, data validation/data quality analysis can all be considered part of the data understanding phase.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data Set
Data Set Description
Select Data
Rationale for Inclusion / Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

Data Preparation



Data Gathering

- Also known as data mining, collects the necessary data for the project. Some data scientists write their programs or work with data engineers and design applications that mine the required data.
- Many tools are available for scraping data off websites, collecting it from mobile apps, or sourcing it from third-party data warehouses. Data scientists need to identify where to source their data, and the collection methods depend on the location and approachability.
- In this step, the data science team also needs to establish databases to store their information for the project's duration. This task usually falls to a data engineer who designs the database to meet the data scientists' needs.

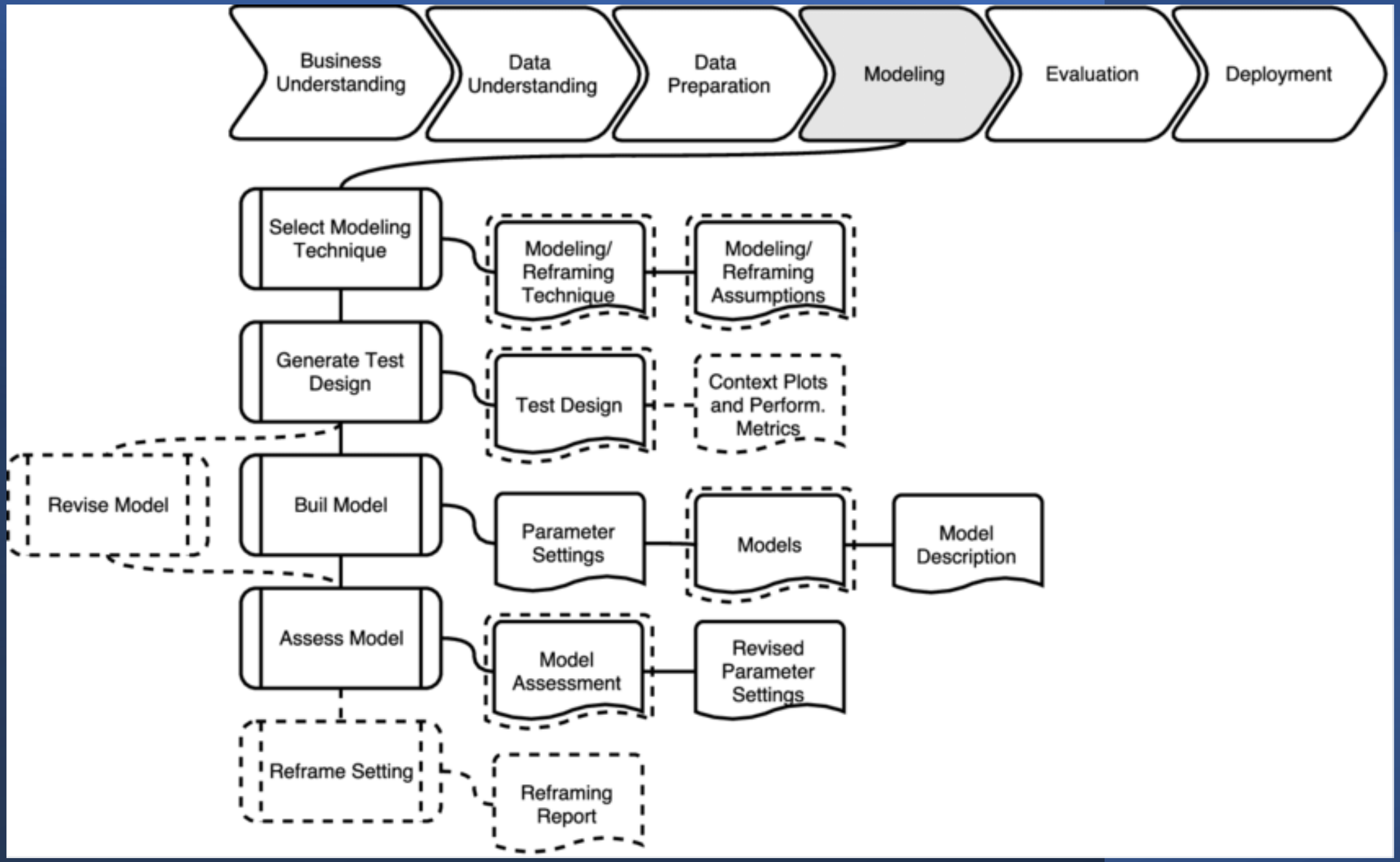
- Data preparation as a generic step comprising all tasks required to make the available data ready to be analyzed is reported in many models. It is also referred to as data pre-processing, data cleaning and data transformation. Once the sources are completely identified, proper selection, cleansing and construction and formatting is done .explanation of information may be executed noticing the patterns considering the business understandings.

Choice of a Representative Sample

- Picking the correct representative sample for your data science project is imperative, especially for big data plans. This step entails figuring out which data sets are necessary to answer the questions and solve the problems at your project's essence.
- Select a representative sample by determining which variables you need to answer the question or solve the problem posed in your project.

Data Cleaning

- It involves transforming the data you collect into a convenient form and ensuring that it applies within the representative sample. Make certain they are using quality data, then finding and tossing out data that does not meet their standards.
- Ensure the complete data is in the same format. This step may necessitate making calculations or administering the data through algorithmic functions, converting it, and matching the other variables you have collected.
- Data that isn’t clean may lead to unsatisfactory results. You will not get the correct results if the variables are incorrect, even if the algorithms for analyzing the data are flawless.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Modeling Technique
Modeling/ Reframing Technique
Modeling/ Reframing Assumptions
Generate Test Design
Test Design
Context Plots and Perform. Metrics
Revise Model
Buil Model
Parameter Settings
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Settings
Reframe Setting
Reframing Report

Development of a Data Model

- A data model selects the data and organizes it according to the project's needs and parameters
- The development of the data model will depend on what the data scientist wants to achieve.
- Despite the different data model choices, the intention is to generate results you can analyze and utilize to solve the initial problem and meet the project goal.
- Developing a data model is the step most people associate with data science.

Modeling

- First step is to select modeling technique
- Create model
- Generate test scenarios for validating model's quality
- Option to repeat previous task
- Do model assessments

Data Analysis

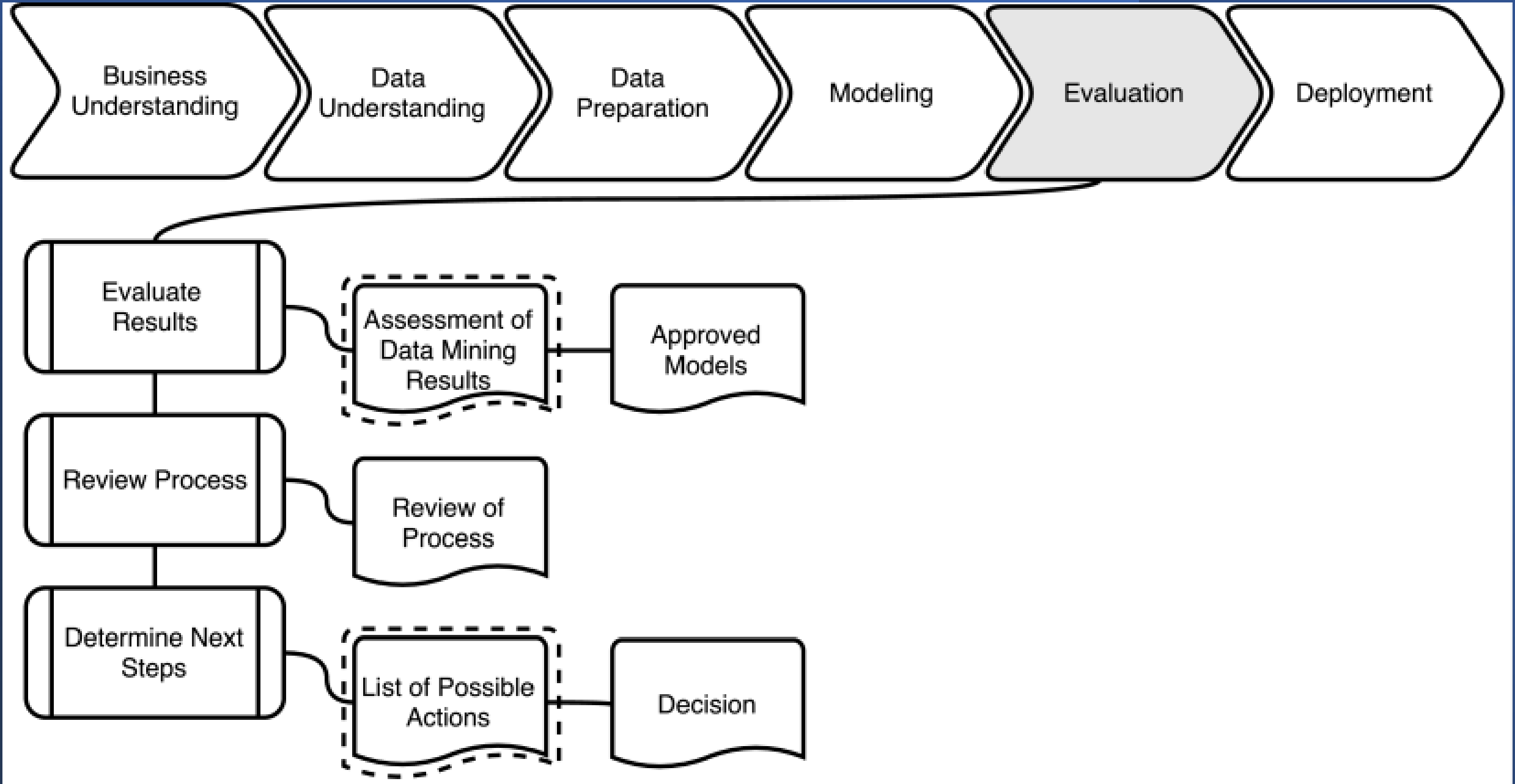
- It aims to discover patterns, trends, and connections that can help create hypotheses and gain insights. It may help you find answers to the questions and problems you identified at the beginning of the project. You can also find a ground for outliers or anomalies.
- The central aspect of data analysis is feature engineering. Features can refer to measurable attributes — they provide quantifiable results scientists can use for predictive modeling or machine learning. Data scientists often use multiple features to arrive at the most accurate predictions possible.

Data Visualization

- it is helpful on a couple of levels; data visualization involves creating graphs, tables, or charts explaining the results of your data study. It can assist data scientists and analysts as they scan for patterns related to their data. It can also help them interpret complex results to people who do not have the same academic background.
- Data scientists need knowledge of the software that can produce graphs and tables. However, the most significant data visualization aspect is to ensure that the visual elements accurately represent the data analysis results.

Setting out

- Final information gathered has to be prepared in a usable manner to the stakeholders. The process may be simple or complex depending n numerable factors.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment of Data Mining Results
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision

Evaluation

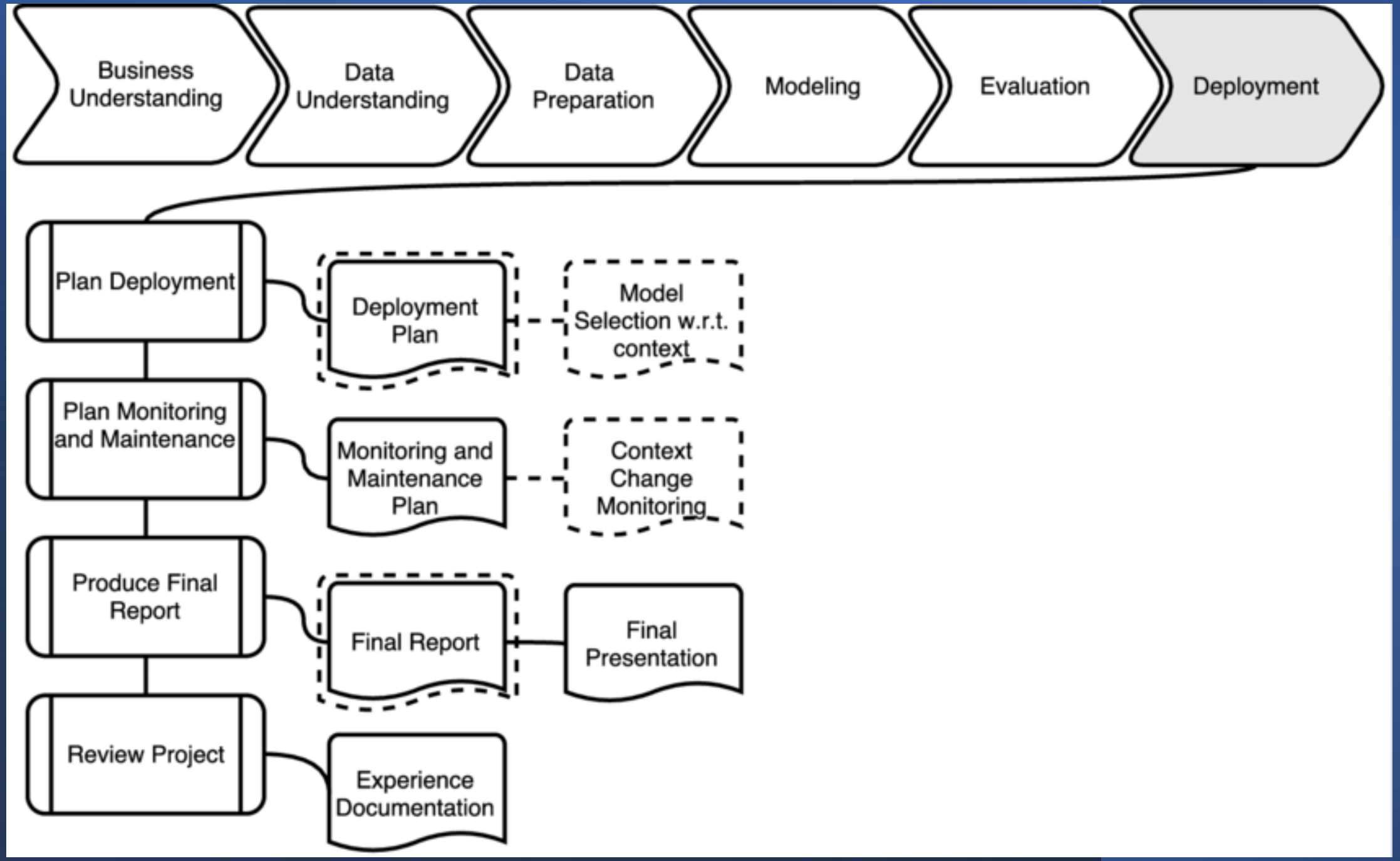
- The evaluation step is also included in all models. Beside
- evaluation, typical denominations for this step include
- interpretation, determine knowledge relevance, reflection,
- knowledge assessment, verification and validation.

Re-evaluation

- Re-evaluation may occur in several forms. During a project, data scientists may evaluate each step to decide if corrections can be made to improve the project or the deployed models or algorithms before stepping on to the next phase of the lifecycle.

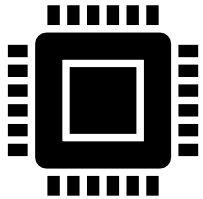
Long-Term Monitoring

- Some data science projects are continuous. Variables may change over time and require extended data mining, data cleaning, and analysis. Data scientists and engineers need to create systems that regularly mine and produce new data sets in these instances.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Model Selection w.r.t. context
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Context Change Monitoring
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

Deployment

- Data scientists need to test the system to ensure it provides accurate results or performs the required tasks accurately because deployment involves using the results of the analysis to perform predictive interpretation, launch a program for machine learning, or contribute insights to decision-makers in the organization using tools that were a result of the study.



Module 3 Learning Outcomes:

- The major steps involved in tackling a data science problem.
- Why data scientists need a methodology and an approach.
- What it means to understand data, and prepare or clean data

Module 4: Decisions Management and Problem Framing

- How to forming a concrete business question or research.
- Converting Business Problems into Analytics Solutions
- Stakeholders & analytical team variable

BeVera

Business Understanding

Ask, what are the desired project outputs?

1

1. Set objectives
    1. Describe the primary objective from a business perspective.
    2. Address relate questions
2. Business success criteria
    1. Write down the criteria the business will evaluate for project success?
    2. Criteria should be specific and measurable
3. Produce project plan
    1. Describe the plan for achieving the data mining and business goals.
    2. Specify the steps to be performed during the rest of the project.
    3. Include the initial selection of tools and techniques.

Business Understanding

Benefits:

- Assists in defining roles and responsibilities
- Reduces duplication of efforts
- Identifies accountability
- Improves team efficiency and effectiveness
- Identifies accountabilities
- Clarifies individual roles and responsibilities
- Can identify problems with existing process
- Removes ambiguities regarding activities and decision points

R — Responsible
The person assigned to do the work

A — Accountable
The person who makes the final decision and has the ultimate ownership

C — Consulted
The person who must be consulted before a decision or action is taken

I — Informed
The person who must be informed that a decision or action has been taken

TIPS

- Consider the roles and responsibilities from the To-Be perspective
- Assign the accountability (A) and responsibility to persons with lowest ranks.
- Assign only one accountable person (A) for each action.
- The accountable person (A) should hold approval authority.
- The number of people to be consulted (C) and informed (I) should be kept at a minimum.

TRAPS

- Trying to apply RACI charts to everything taking place.
- 100% accuracy is not always required.
- Filling in the chart without due diligence.
- Double check that the operation can be completed with the new role assignment.
- Only one accountability per activity.
- All roles & responsibilities must be documented and communicated.

The expertise of businesspeople allows you to answer the following questions:

- Is the data mining efforts necessary?
- Should we focus on a particular segment or subgroup?
- What are the relevant business rules?
- What do they (the businesspeople) know about the data?
  - Domain experts know where data resides and how it is stored.
  - They may know whether some sources are invalid.
- What do their intuition and experience say is important? Can be a source of insight.

Business Problem Framing

- Frame the problem…Framing is the process of stating the analytics problem to be solved. *Name* the issue as a problem.
  - State the problem – Why it is important, and to whom?
  - Identify key stakeholders and their interests
  - Articulate the current situation and pain points

  - Objectives – identify what needs to be achieved in business terms and what needs to be done to meet the needs
    - What is the goal? What are the criteria for success? What is “good enough”?
    - What is the failure criterion?
- Two common ways of framing are 1) framing an issue broadly as "common ground," and 2) framing an issue from a "single angle."

Frame the Problem

<u>Framing</u> is the process of stating the analytics problem to be solved.

- *Name* the issue as a problem.
- *State* the problem:
  - Why it is important, and to whom?
- *Identify* key stakeholders and their interests
- *Articulate* the current situation and pain points
- *Identify* what needs to be achieved in business terms and what needs to be done to meet the needs
  - What is the goal?
  - What are the criteria for success?
  - What is “good enough”?
  - What is the failure criterion?

Business Problem Framing

Who: are the stakeholders who satisfy one or more of the following with respect to the project: funding, using, creating, or are affected by the project's outcome?

What: problem/function is the project meant to solve/perform?

Where: does the problem occur? Or where does the function need to be performed? Are the physical and spatial characteristics articulated?

When: does the problem occur, or function need to be performed? When does the project need to be completed?

Why: does the problem occur, or function need to occur?

Re-framing business problems

Rethink the question

Brainstorm bad ideas

Unpack your assumptions

Business Understanding

- Statement of Business Objective
  - State goal in business terminology
- Statement of Data Mining objective
  - State objectives in technical terms
- Statement of Success Criteria
  - Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives
  - What the client really wants to accomplish?
  - Uncover important factors (constraints, competing objectives).