https://www.projectmates.eu/

MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Tutorial D01

# Research Methods in Data Science

Yuri Demchenko MATES Project

University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

Co-funded by the Erasmus+ Programme of the European Union

# Outline

- Research methods: Importance for Data Science
- Research methods and Research types
  - Research questions, Hypothesis and Hypothesis testing
- Business research

- CRISP-DM: Model, stages and tasks

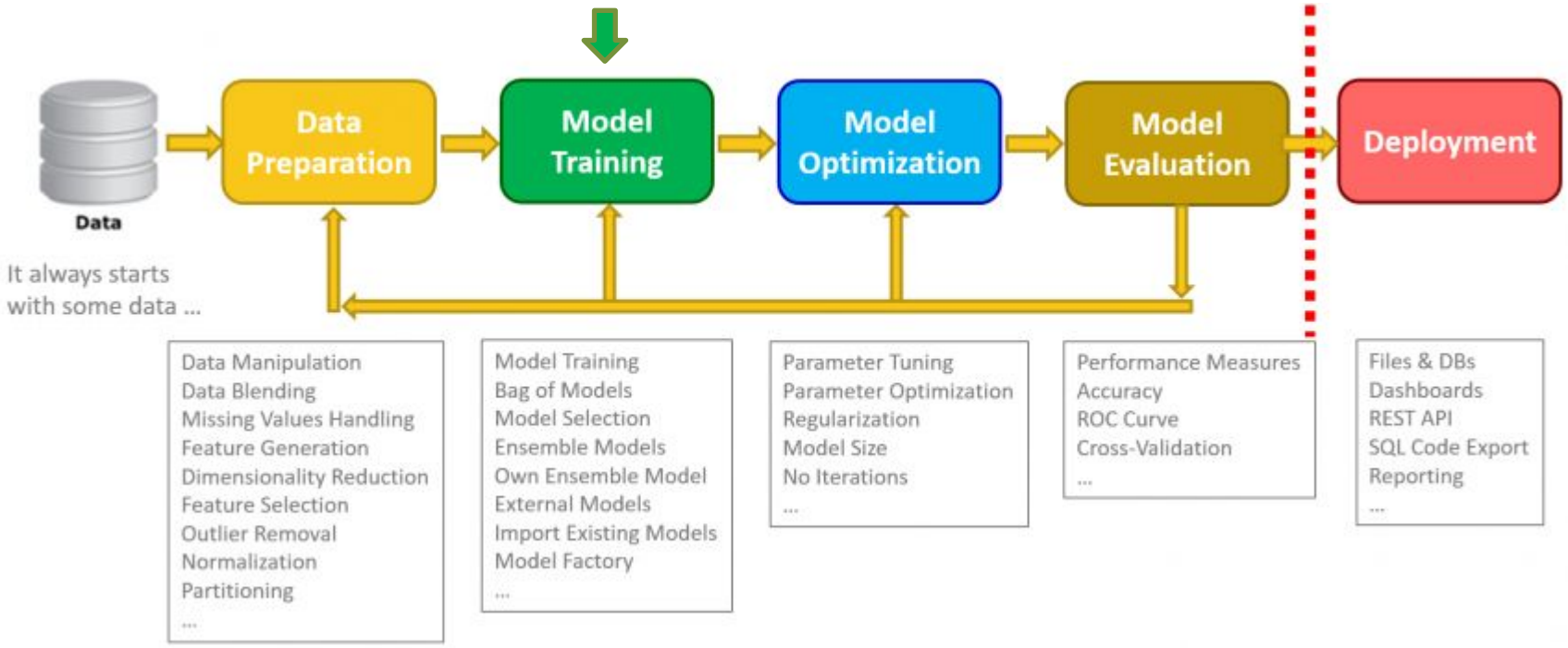


Co-funded by the
Erasmus+ Programme
of the European Union

# Every Data Science Project/Research is based on Data

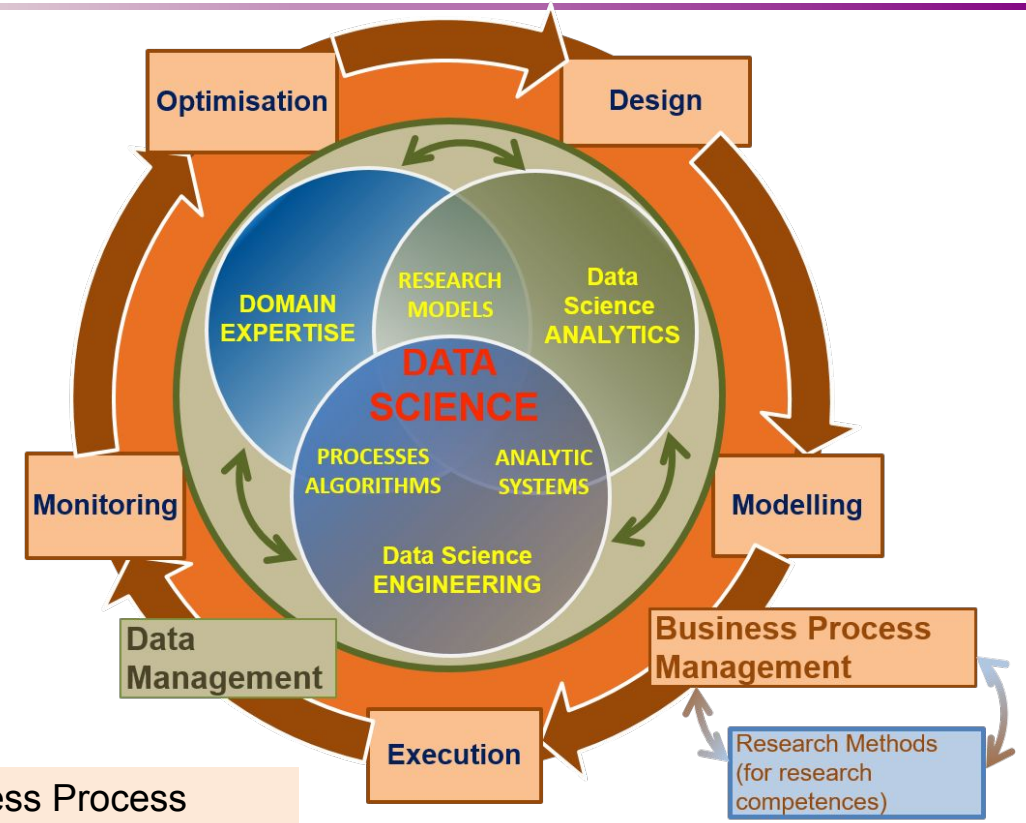
https://www.knime.com/blog/analytics-and-beyond

Model selection and Training




- Each phase – data preparation, model training and evaluation, and model deployment – operates on its own data set. All these data sets need to be completely isolated from each other. The pollution of data sets across the data science assembly line is one of the most frequent mistakes in model production.
- The data science journey always starts with some ***historical data*** *or* ***sample dataset*** lying around in a repository.
- They are our secret weapon in this ***data blending*** phase – connecting external sources

# EDISON Data Science Framework: Research Cycle and Business Process Cycle

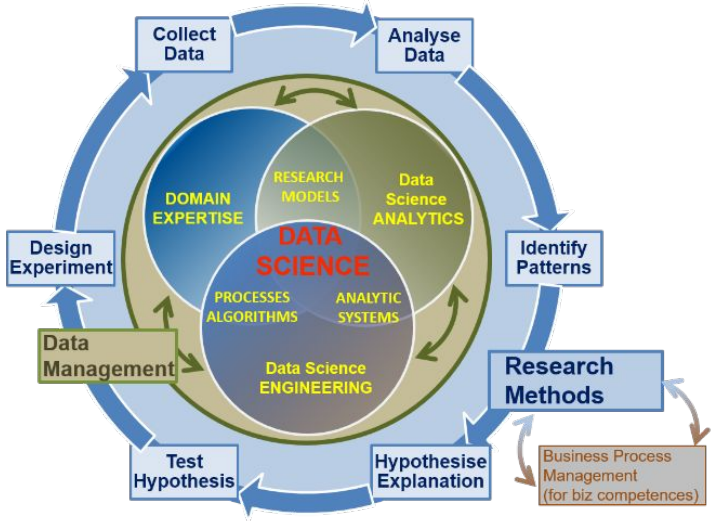



Business Process Operations/Stages

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Data Science Competences/Activities include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)




Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

# Data Science and Research Methods

- **Data Science needs methodology**
- Research is an organized and systematic way to find answers to questions
  - Data Science uses data analysis to answer questions
- Research is a creative process
  - So is Data Science
- To effectively work as a Data Scientist you need to know Research Methods
  - Sufficient to have your confident opinion (and ask questions)

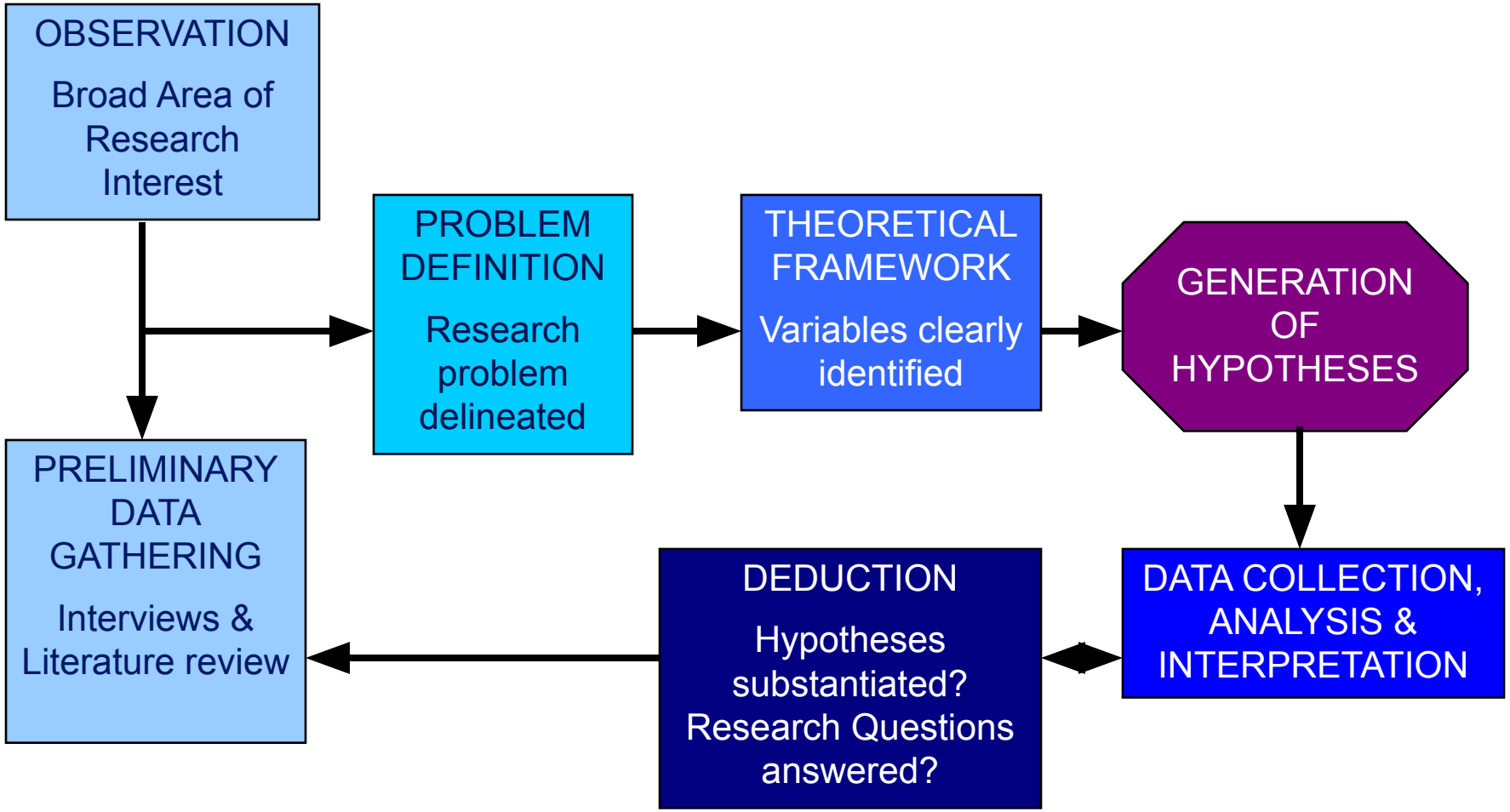
- **Data Science and Analytics projects need development and operationalization methodology and platform**

Research process cycle
- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

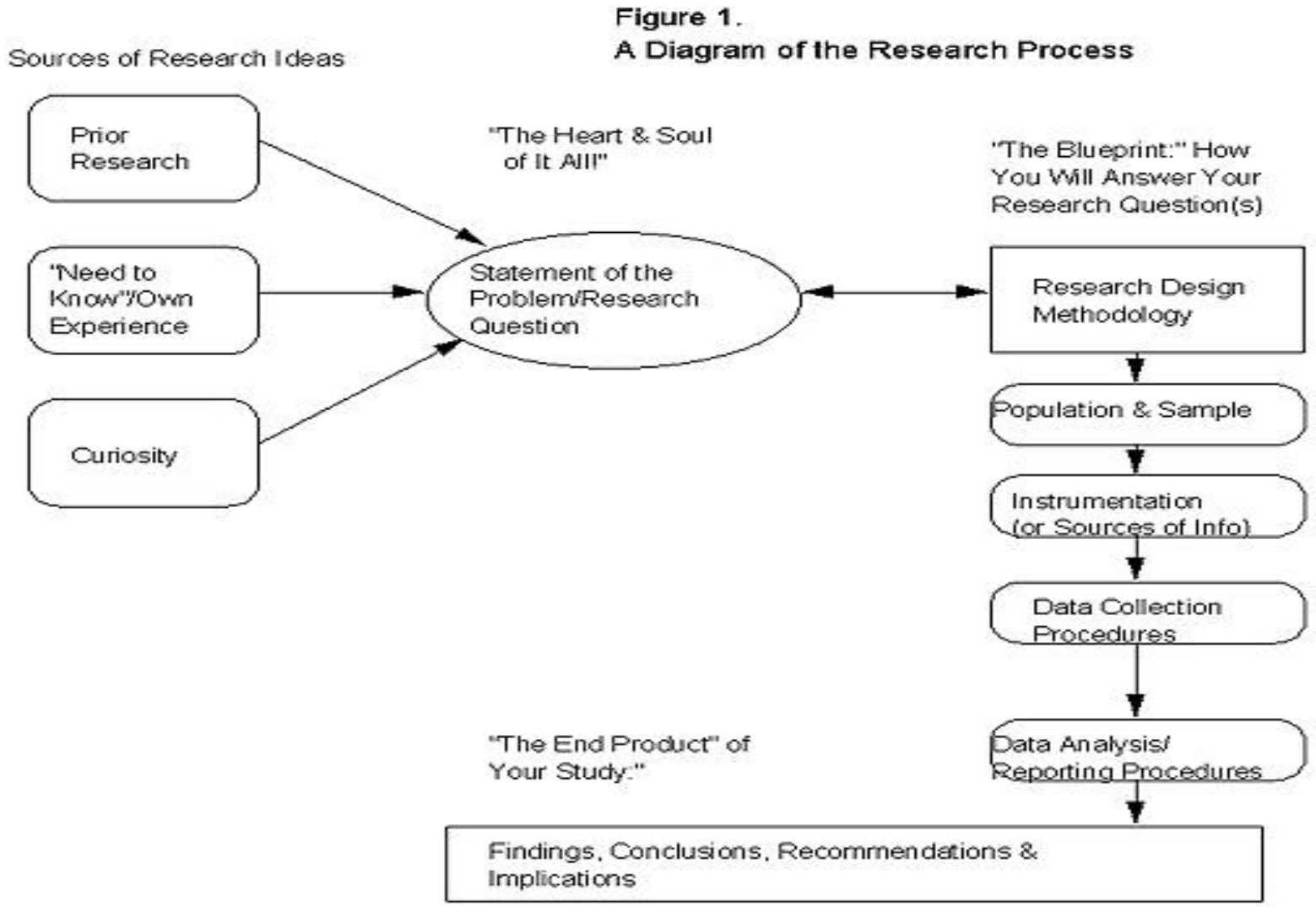
# Research Process – Overview

- Research Process
- Research Overview
- Research Method
- Research Design
- Conclusion

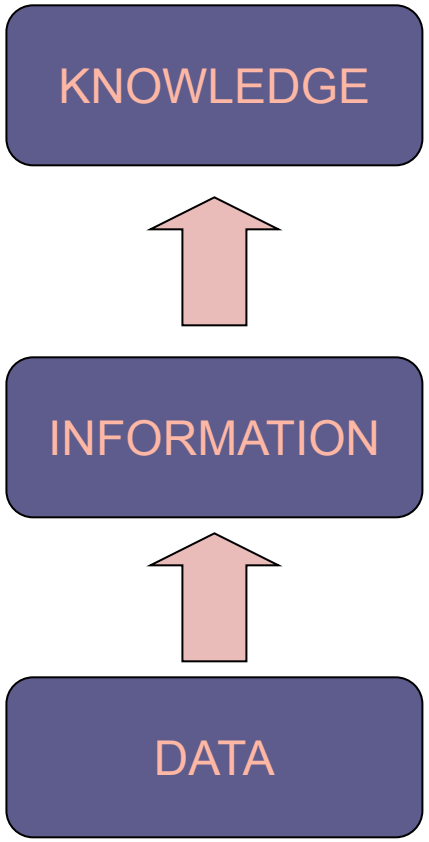
# The Research Process



# Research Process



# The Building Blocks of Research




Blend of information, experience and in-sights that provides a framework that can be thoughtfully evaluated when assessing new information or evaluating relevant situations

Determination of relationship amongst data with a view to facilitating understanding of the phenomena, their relationships and decision-making
(e.g. past and predicted future sales trends)

Measurements of phenomena
(e.g. sales statistics of a department store)

# The Research Idea

Where to get a research idea?

- Professional experience
- Burning questions
- Literature
- Professional meetings
- Discussions

# Criteria for developing a good research question

- Feasible
- Interesting
- Novel
- Ethical
- Relevant

- **Feasible**
  - Subjects
  - Resources
  - Manageable
  - Data available?
- **Interesting**
- **Novel**
  - In relation to previous findings
    - Confirm or refute?
  - New setting, new population

- **Ethical**
  - Social or scientific value
  - Safe
- **Relevant**
  - Advance scientific knowledge?
  - Influence clinical practice?
  - Impact health policy?
  - Guide future research?

[ref] Steven R. Cummings, Warren S. Browner, and Stephen B. Hulley, Conceiving the Research Question and Developing the Study Plan [online] https://www.topvelocity.net/wp-content/uploads/2017/10/Ch2-Hulley-Conceiving-the-Research-Question.pdf

# A Research Question Must Identify

1. The **variables** under study
2. The **population** (objects/subjects) being studied
3. The **testability** of the question

# Variables in Research

- Have 2 or more properties or qualities
  - Age, sex, weight, height
- Is one variable related to another?
  - " Is X related to Y? What is the effect of X on Y?" etc.
- Variables analysis is an initial stage of the data analysis
  - E.g. Data preparation by VRIS-DM
- Independent variable:
  - Has a presumed effect on the dependent variable (outcome)
  - May or may not be manipulated
- Dependent variable:
  - Something that varies with a change in the independent variable
  - *Outcome* variable

# Hypothesis

- Statement about the relationship between 2 or more variables
- Converts the question into a statement that predicts an expected outcome
- A unit or subset of the research problem

- Characteristics of hypotheses
  - Declarative statement that identifies the predicted relationship between 2 or more variables
  - Testability
  - Based on sound scientific theory/rationale

- Directional vs. Non-Directional Hypotheses
  - Directional hypothesis
    - Specifies the direction of the relationship between independent and dependent variables
  - Non-directional hypothesis
    - Shows the existence of a relationship between variables but no direction is specified

# Hypothesis Testing

- Explain the nature of relationships
- Establish differences among groups or the interdependence of two or more factor in a situation
- Explain the variance in the dependent variable or to predict organizational outcome
- Revisit statistical hypothesis testing
  - AB testing

# Examples of different types of research

- Basic/Fundamental Research
  - Correlation-prediction
  - Experiment
  - Survey-questionnaire
  - Theory construction
  - Trend analysis

- Applied Research
  - Case study
  - Comparison
- Business Research
  - Analysis
  - Evaluation
  - Design-demonstration
  - Status

# What is basic research

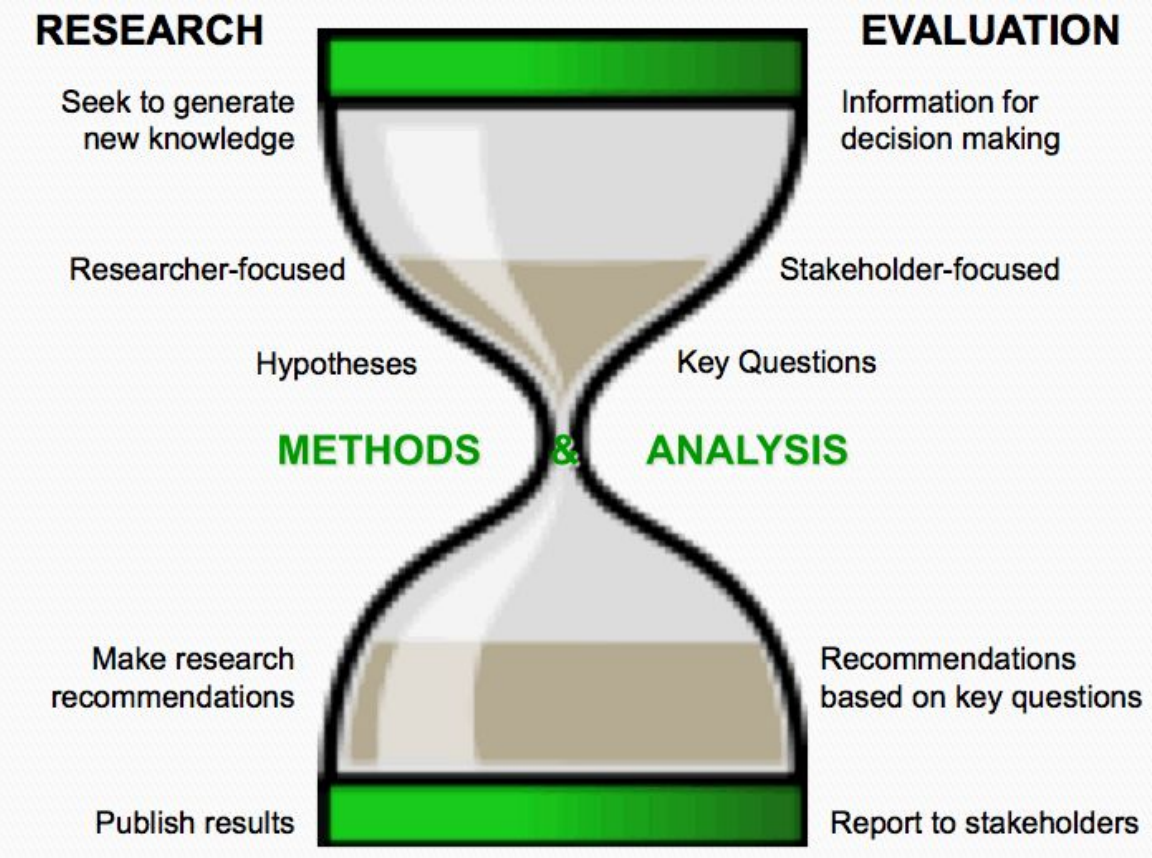
The goal is to produce new knowledge, which takes three main forms:

- Exploratory research, which structures and identifies new problems
- Constructive research, which develops solutions to a problem
- Empirical research, which tests the feasibility of a solution using empirical evidence

- Research can also fall into two distinct types:
  - Primary research
  - Secondary research

- Research is often illustrated using the hourglass model

# Hourglass Model




- Hypothesis checking based on key element of the suggested model
- Observable data must validate the hypothesis

[ref] https://aea365.org/blog/john-lavelle-on-describing-evaluation/

# Research Method

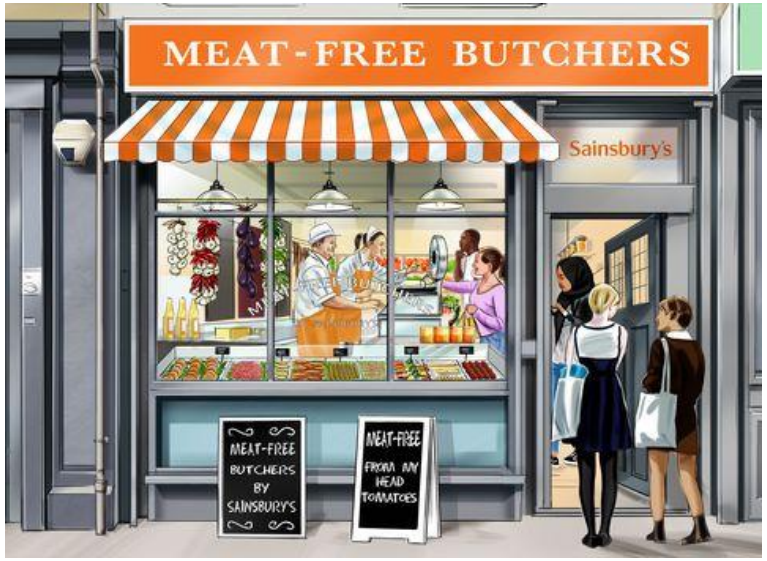
- The method section answers two main questions
  1. How was the data collected or generated?
  2. How was it analyzed?

***" It shows your reader how you obtained your results"***

- Why do you need to explain how you obtained your results?
  1. Method affects results
  2. Helps the reader evaluate the validity and reliability of your results and conclusions you draw from them

# Examples: Research Methods in Digital/Mobile Age

- Butcher shop in London
  - Center of London – very competitive space for rent
  - Family shop > 100 yrs
  - Sells dropped
- Goal of research
  - Research on local community to increase sell/buyers

# Examples: Research Methods in Digital/Mobile Age

- Butcher shop in London
  - Used mobile and wifi sensing for passing crowd
    - Move patterns: time and stopping (by shop)
  - Discovered
    - Crowds are passing by shop in the evening: after sport events time, after bar closes at 11pm
- Recommendation
  - Sell sandwiches in evening, after 9pm till midnight
- Results
  - Sell of sandwiches overcomes traditional butcher

# Examples: Research Methods in Digital/Mobile Age

- Butcher shop in London
  - Use mobile and wifi sensing for passing crowd
- Boutique windows advertisement in the mall
  - Investigate people passing by boutique and advertisement time
  - It will work for this case as well
  - Question: Is it legal to collect sensing data?
- Your example

# Method

- Your methodology should make **clear the reasons why** you choose a particular method or procedure.
- The data was collected or generated in a way **consistent with accepted practice in the field of study**
- The research method must be **appropriate to  the objectives** of the study
- The methodology should also **discuss the problems** that were anticipated and explain steps taken to prevent them from occurring

# Research Design vs Design Research

- **Design research** investigates the process of designing in all its many fields.
  - Widely used in UX (User eXperience) design research

- It is thus related to Design methods in general or for particular disciplines.
  - A primary interpretation of design research is that it is concerned with undertaking research *into* the design process.
  - Secondary interpretations would refer to undertaking research *within* the process of design. The overall intention is to better understand and to improve the design process.

# Quantitative Research

Data in numbers

- Investigates the why and how of decision making, as compared to what, where, and when of quantitative research.
- Quantitative research is the systematic scientific investigation of properties and phenomena and their relationships.
- The objective of quantitative research is to develop and employ mathematical models, theories and/or hypotheses pertaining to natural phenomena.
- The process of measurement is central to quantitative research because it provides the fundamental connection between empirical observation and mathematical expression of quantitative relationships.
- Quantitative research is widely used in both the natural sciences and social sciences, from physics and biology to sociology and journalism. It is also used as a way to research different aspects of education. The term quantitative research is most often used in the social sciences in contrast to qualitative research.

# Qualitative Research

## Data in words

- Involves an in-depth understanding of human behavior and the reasons that govern human behavior
- Qualitative research relies on reasons behind various aspects of behavior. Simply put, it investigates the **why** and **how** of decision making
- The need is for smaller but focused samples rather than large random samples, which qualitative research categorizes data into patterns as the primary basis for organizing and reporting results.
- Qualitative researchers typically rely on four methods for gathering information: (1) participation in the setting, (2) direct observation, (3) in depth interviews, and (4) analysis of documents and materials

# Exploratory Research

- Is a type of research conducted because a problem has not been clearly defined.
- Helps determine the best research design, data collection method and selection of subjects.
- Given its fundamental nature, it often concludes that a perceived problem does actually not exist.
  - Still an important activity: Negative results often tells more than positive

# Constructive Research

- Is perhaps the most common **computer science research method**.
- This type of approach demands a form of validation that doesn't need to be as empirically based as in other types of research like exploratory research.
- The conclusions have to be objectively argued and defined.
- This may involve evaluating the "construct" being developed analytically against some predefined criteria or performing some benchmark tests with the prototype.
- Working code is an answer and proof, and a way to production

# Empirical Research

- Any research that bases its findings on direct or indirect observation as its test of reality.
  - Such research may also be conducted according to hypothetico-deductive inference procedures
    - Developed in works by R. A. Fisher
  - The researcher attempts to describe accurately the interaction between the instrument (or the human senses) and the entity being observed.
  - If instrumentation is involved, the researcher is expected to calibrate her/his instrument by applying it to known standard objects and documenting the results before applying it to unknown objects.

# Primary Research

- **Primary research** (also called **field research**) involves the collection of data that don't already exist.

- ***Methods of collection primary data***
  - **Observation:** Looking at and recording what people do and how they behave. Today, store cameras can be used to observe consumer behaviour
  - **Experiments:** Market researchers can use experimental techniques. e.g. test marketing, blind taste tests
  - **Surveys:** Involves asking questionnaires to respondents
  - **Consumer panels:** Select a group of consumers that the company regularly surveys to identify changing attitudes

# Secondary Research

- **Secondary research** (also known as **desk research**)
  - Involves the summary, collation and/or synthesis of existing research rather than primary research, where data are collected from, for example, research subjects or experiments.
  - The term is widely used in market research and in medical research.
  - The principle methodology in medical secondary research is the systematic review, commonly using meta-analytic statistical techniques, although other methods of synthesis, like realist reviews and meta-narrative reviews, have been developed in recent years.
    - Current issues with executive papers and open data linked to academic papers

# Case Study

- Gathers in-depth information on a single entity
- Problem solving techniques
- Involve contextual analyses of similar situation in other organization

# Cohort Studies

- [Wikipedia] A **cohort** study is a particular form of longitudinal study that samples a **cohort** (a group of people who share a defining characteristic, typically those who experienced a common event in a selected period, such as birth or graduation), performing a cross-section at intervals through time.

- Often form of research in medical studies
  - Think about COVID-19

# Types of Investigation

- Clarification
  - Clear understanding of the concept
  - Related to exploratory and descriptive study
- Correlational
  - At least two concept or variables move simultaneously
- Causal relationship/experimental/group comparison
  - One concept or variables causes a movement in another concept or variables
  - True experiment vs quasi-experiment

# Business Research

- **What is Business Research?**
  - Business Research may be defined as the “systematic and objective process of gathering, recording and analyzing data for aid in making business decisions” (Zikmund, Business Research Methods, 2002, p. 6)
  - Systematics and Objectivity are its distinguishing features of Business Research, which is important tool for managers and decision-makers in corporate and non-corporate organizations

- **When is Business Research Used?**
  - Typically, business research methods are used in situations of uncertainty, that is, when decision-makers face two or more courses of action and seek to select the best possible alternative under the circumstances.
  - Business Research is hence aimed at improving the quality of decision-making which, in turn, benefits the organization and helps ensure its continuity and efficiency

# Who Uses the Business Research Methods

- Businesses and Corporations
- Public-Sector Agencies
- Consulting Firms
- Research Institutes
- Non-Governmental Organizations
- Non-Profit Organizations
- Independent Researchers and Consultants

# Common Business Research Methods & Techniques

- Surveys
- Interviews
- Observation
- Experiments
- Archival and Historical Data
- Qualitative Analysis
- Quantitative Analysis

# Common Business Research Methods & Techniques

- Surveys
- Interviews
- Observation
- Experiments
- Archival and Historical Data
- Qualitative Analysis
- Quantitative Analysis

What data do they produce?

# Fields where Business Research is often Used – (1)

**General Business Conditions and Corporate Research**

- Short- & Long-Range Forecasting,
- Business and Industry Trends
- Global Environments
- Inflation and Pricing
- Plant and Warehouse Location
- Acquisitions

**Management and Organizational Behaviour Research**

- Total Quality Management
- Morale and Job Satisfaction
- Leadership Style
- Employee Productivity
- Organizational Effectiveness
- Structural Issues
- Absenteeism and turnover
- Organizational Climate

**Financial and Accounting Research**

- Forecasts of financial interest rate trends,
- Stock, bond and commodity value predictions
- Capital formation alternatives
- Mergers and acquisitions
- Risk-return trade-offs
- Portfolio analysis
- Impact of taxes
- Research on financial institutions
- Expected rate of return
- Capital asset pricing models
- Credit risk
- Cost analysis

# Fields where Business Research is often Used – (2)

**Sales and Marketing Research**

- Market Potentials
- Market Share
- Market segmentation
- Market characteristics
- Sales Analysis
- Establishment of sales quotas
- Distribution channels
- New product concepts
- Test markets
- Advertising research
- Buyer behaviour
- Customer satisfaction
- Website visitation rates

**Information Systems Research**

- Knowledge and information needs assessment
- Computer information system use and evaluation
- Technical support satisfaction
- Database analysis
- Data mining
- Enterprise resource planning systems
- Customer relationship management systems

**Corporate Responsibility Research**

- Ecological Impact
- Legal Constraints on advertising and promotion
- Sex, age and racial discrimination / worker equity
- Social values and ethics

# Selected Examples of Real-Life Situations in which Business Research Methods are Used

- A firm wants to produce and market a new product but first wants to ascertain if there is a potential consumer demand for this product in markets x,y and z
- A multinational firm wants to establish a production facility in another country after determining its technical and economic feasibility
- A government agency wants to ascertain the satisfaction level of its employees, the causes for any possible discontent, and propose a scheme for enhancing this level
- A financial institution wants to invest in commodities and commissions a study to determine the past trends and forecast future returns in a portfolio of commodities
- The CEO of a firm wants to undertake a SWOT-Analysis as part of his plan to redefine his organization's priorities

# Basic and Applied Research

**Basic Research** aims to expand the frontiers of science and knowledge by verifying or disproving the acceptability of a given theory or attempting to discover more about a certain concept (non-specificity)

Example: How does motivation affect employee performance?

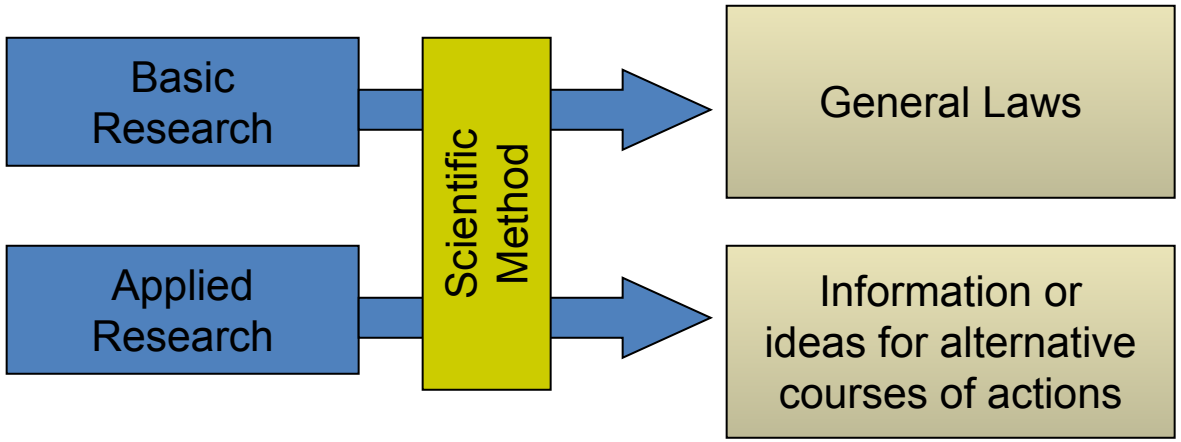
**Applied Research** focuses on a real-life problem or situation with a view to help reaching a decision how to deal with it (Specificity)

Example: Should corporation X adopt a paperless office environment?

# The Essence of the Scientific Method

**Characteristics of the Scientific Method**

Objectivity
Systematic Analysis
Logical Interpretation of Results




**Elements of the Scientific Method**

- Empirical Approach
- Observations
- Questions
- Hypotheses
- Experiments
- Analysis
- Conclusion
- Replication

# The Value of Business Research for Managers (1)

Reduction of uncertainty and improvement in the quality of decision-making with several consequent advantages (e.g. strategic, operational) and benefits for organizations

Business Research Methods can be employed in each of the following four stages:

(1) Identification of problems and/or opportunities

Useful for strategy planning, analysis of internal and external organizational environment

(2) Diagnosing and Assessment of problems and/or opportunities

Its purpose is to gain insight into the underlying reasons and causes for the situation. If there is a problem, it asks what happened and why? If there is an opportunity, it seeks to explore, clarify and refine the nature of the opportunity and, in the case of multiple opportunities, seeks to set priorities

(3) Selection and Implementation of Courses of Action

After alternative courses of action have been determined, selection of the best possible course.

(4) Evaluation the Course of actions

# The Value of Business Research for Managers – Evaluation Research

Evaluation Research – the formal objective measurement and evaluation of the extent to which an activity, project or programme has achieved its goal, and the factors which influence performance (e.g. audits). It is also the formal objective measurement and evaluation of the extent to which on-going activities, projects or programmes are meeting their goals (performance-monitoring research)

Examples of performance-monitoring research:

(1) Are railway passengers satisfied with the level of service the railway company is providing? If not, then research may need to be undertaken to ascertain the reasons for customer dissatisfaction and propose corrective measures

(2) What are the trends in retail and wholesale sector? Can research suggest new ways to improve efficiency in purchase transactions?

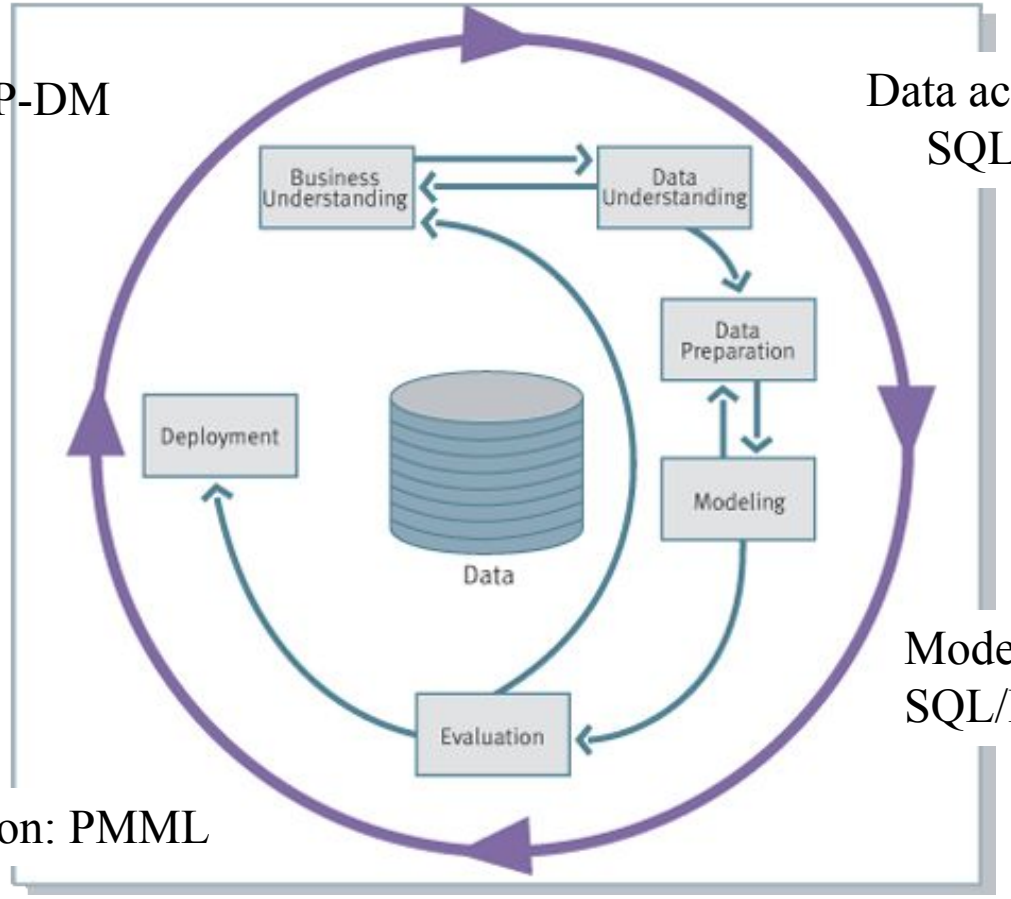
# Data Mining Models

- 5A – used by SPSS Clementine
  (Assess, Access, Analyze, Act and Automate)
- SEMMA – used by SAS Enterprise Miner
  (Sample, Explore, Modify, Model and Assess)
- CRISP–DM – tends to become a standard

# CRISP-DM

- CRISP-DM: A Standard Process Model for Data Mining - http://www.crisp-dm.org/

Global view: CRISP-DM




Data access:
SQL interfaces

Model generation:
SQL/MM, OLEDB DM

Model representation: PMML

The Predictive **Model** Markup Language (**PMML**)

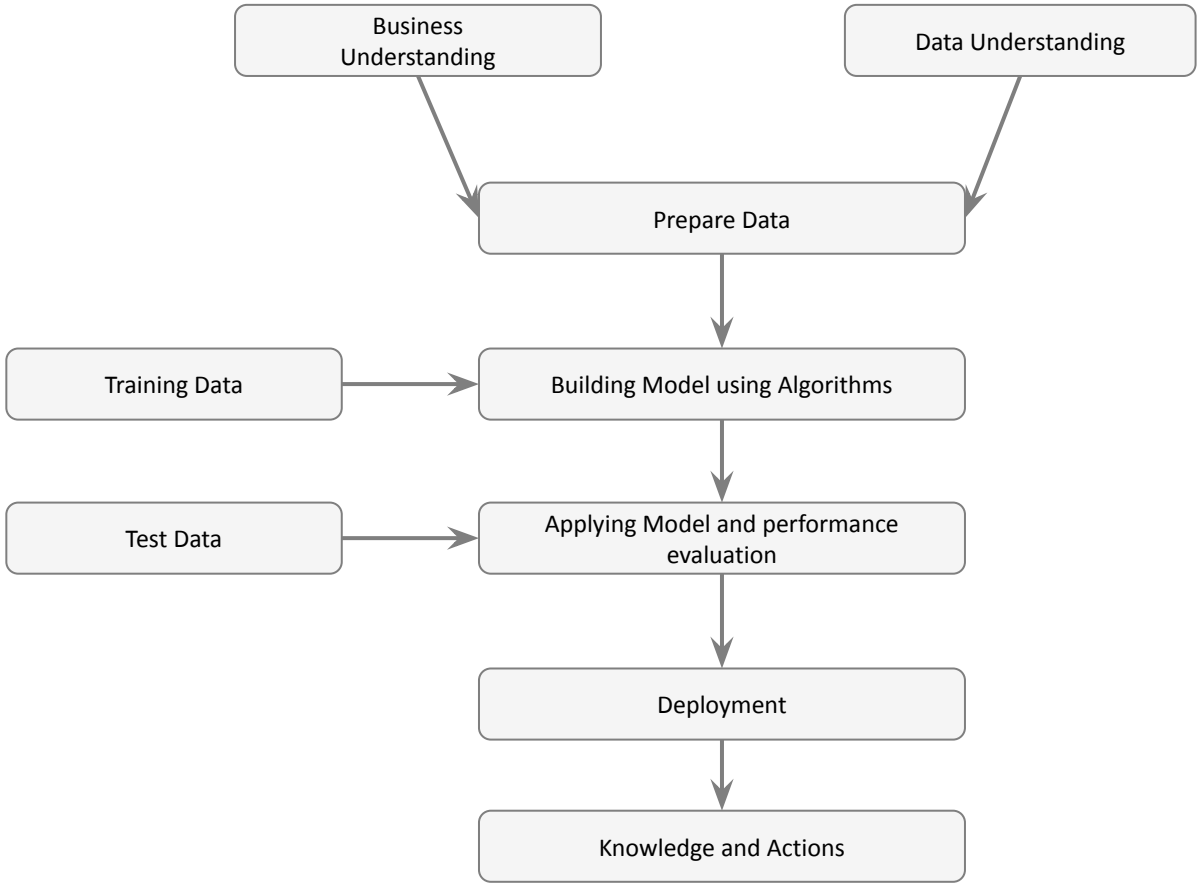
# What is CRISP-DM?

- Cross-Industry Standard Process for Data Mining
- Aim:
  - To develop an industry, tool and application neutral process for conducting Knowledge Discovery
  - Define tasks, outputs from these tasks, terminology and mining problem type characterization
- Founding Consortium Members: DaimlerChrysler, SPSS and NCR
- CRISP-DM Special Interest Group ~ 200 members
  - Management Consultants
  - Data Warehousing and Data Mining Practitioners

# Four Levels of Abstraction

- Phases
  - Example: Data Preparation
- Generic Tasks
  - A stable, general and complete set of tasks
  - Example: Data Cleaning
- Specialized Task
  - How is the generic task carried out
  - Example: Missing Value Handling
- Process Instance
  - Example: The mean value for numeric attributes and the most frequent for categorical attributes was used

# Process of Data Analysis (based on CRISP-DM)




**1. Prior Knowledge**

**2. Preparation**

**3. Modeling**

**4. Application**

**5. Knowledge**

- Note: the process is not linear, repeatedly backtracking

# Business Understanding Phase - Objectives

## Determining Business Objectives

- 1. Gather background information
  - Compiling the business background
  - Defining business objectives
  - Business success criteria
- 2. Assessing the situation
  - Resource Inventory
  - Requirements, Assumptions, and Constraints
  - Risks and Contingencies
  - Cost/Benefit Analysis
- 4. Determining data science goals
  - Data science goals
  - Data science success criteria
- 5. Producing a Project Plan

# Business Understanding Phase (1)

- Understand the business objectives
  - What is the status quo?
    - Understand business processes
    - Associated costs/pain
  - Define the success criteria
  - Develop a glossary of terms: speak the language
  - Cost/Benefit Analysis
- Current Systems Assessment
  - Identify the key actors
    - Minimum: The Sponsor and the Key User
  - What forms should the output take?
  - Integration of output with existing technology landscape
  - Understand market norms and standards

# Business Understanding Phase (2)

- Task Decomposition
  - Break down the objective into sub-tasks
  - Map sub-tasks to data mining problem definitions
- Identify Constraints
  - Resources
  - Law e.g. Data Protection
- Build a project plan
  - List assumptions and risk (technical/financial/business/ organisational) factors

# Example Project Plan

| Phase | Time | Resources | Risks |
|---|---|---|---|
| Business understanding | 1 week | All analysts | Economic change |
| Data understanding | 3 weeks | All analysts | Data problems, technology problems |
| Data preparation | 5 weeks | Data scientists, DB engineers | Data problems, technology problems |
| Modeling | 2 weeks | Data scientists | Technology problems, inability to build adequate model |
| Evaluation | 1 week | All analysts | Economic change, inability to implement results |
| Deployment | 1 week | Data scientist, DB engineers, implementation team | Economic change, inability to implement results |

# Are you ready for Data Understanding?

- From a business perspective:
  - What does your business hope to gain from this project?
  - How will you define the successful completion of our efforts?
  - Do you have the budget and resources needed to reach our goals?
  - Do you have access to all the data needed for this project?
  - Have you and your team discussed the risks and contingencies associated with this project?
  - Do the results of your cost/benefit analysis make this project worthwhile?

- From a data science perspective:
  - ✔ How specifically can data mining help you meet your business goals?
  - ✔ Do you have an idea about which data mining techniques might produce the best results?
  - ✔ How will you know when your results are accurate or effective enough? (Have we set a measurement of data mining success?)
  - ✔ How will the modeling results be deployed? Have you considered deployment in your project plan?
  - ✔ Does the project plan include all phases of CRISP-DM?
  - ✔ Are risks and dependencies called out in the plan?

# Data Understanding Phase (1)

- **Data Understanding**
  - Proceeds with activities aimed at:
    - Getting familiar with the data
    - Identifying data quality problems
    - Discovering first insights into the data
    - Detecting interesting subsets to form hypotheses for hidden information

- Collect Data
  - What are the data sources?
    - Internal and External Sources (e.g. Axiom, Experian)
    - Document reasons for inclusion/exclusions
    - Depend on a domain expert
    - Accessibility issues
      - Legal and technical
  - Are there issues regarding data distribution across different databases/legacy systems
    - Where are the disconnects?

# Data Understanding Phase (2)

- Data Description
  - Document data quality issues
    - requirements for data preparation
  - Compute basic statistics

- Data Exploration
  - Simple univariate data plots/distributions
  - Investigate attribute interactions
  - Data Quality Issues
    - Missing Values
      - Understand its source: Missing vs Null values
    - Strange Distributions

# Prepare for Data Preparation

- Are all data sources clearly identified and accessed? Are you aware of any problems or restrictions?
- Have you identified key attributes from the available data?
- Did these attributes help you to formulate hypotheses?
- Have you noted the size of all data sources?
- Are you able to use a subset of data where appropriate?
- Have you computed basic statistics for each attribute of interest? Did meaningful information emerge?
- Did you use exploratory graphics to gain further insight into key attributes? Did this insight reshape any of your hypotheses?
- What are the data quality issues for this project? Do you have a plan to address these issues?
- Are the data preparation steps clear? For instance, do you know which data sources to merge and which attributes to filter or select?

# Data Preparation Phase (1)

- Data Preparation
  - Covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data
  - Data preparation tasks are likely to be performed multiple times, and not in any prescribed order
  - Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools
- Integrate Data
  - Joining multiple data tables
  - Summarisation/aggregation of data
- Select Data
  - Attribute subset selection
    - Rationale for Inclusion/Exclusion
  - Data sampling
    - Training/Validation and Test sets

# Data Preparation Phase (2)

- Data Transformation
  - Using standard functions for data transformation
  - Factor/Principal Components analysis
  - Normalization/Discretisation/Binarisation
- Clean Data
  - Handling missing values/Outliers
- Data Construction
  - Derived Attributes

# Prepare for Data Modelling

- Based upon your initial exploration and understanding, were you able to select relevant subsets of data?
- Have you cleaned the data effectively or removed unsalvageable items? Document any decisions in the final report.
- Are multiple data sets integrated properly? Were there any merging problems that should be documented?
- Have you researched the requirements of the modeling tools that you plan to use?
- Are there any formatting issues you can address before modeling? This includes both required formatting concerns as well as tasks that may reduce modeling time.

# The Modelling Phase

- Selection of the appropriate modelling technique
  - Data pre-processing implications
    - Attribute independence
    - Data types/Normalisation/Distributions
  - Dependent on
    - Data mining problem type
    - Output requirements
- Develop a testing regime
  - Sampling
    - Verify samples have similar characteristics and are representative of the population

# The Modelling Phase

- Build model
  - Choose initial parameter settings
  - Study model behaviour
    - Sensitivity analysis
- Assess the model
  - Be beware of over-fitting
  - Investigate the error distribution
    - Identify segments of the state space where the model is less effective
  - Iteratively adjust parameter settings
    - Document reasons for these changes

# Prepare for Evaluation

- Are you able to understand the results of the models?
- Do the model results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration?
- From your initial glance, do the results seem to address your organization's business question?
- Have you used analysis nodes and lift or gains charts to compare and evaluate model accuracy?
- Have you explored more than one type of model and compared the results?
- Are the results of your model deployable?

# The Evaluation Phase

- Evaluation
  - At this stage, a model (or models) that appears to have high quality, from a data analysis perspective, has been built
  - Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives
  - A key objective is to determine if there is some important business issue that has not been sufficiently considered
  - At the end of this phase, a decision on the use of the data mining results should be reached

- Validate Model
  - Human evaluation of results by domain experts
  - Evaluate usefulness of results from business perspective
    - Define control groups
    - Calculate lift curves
    - Expected Return on Investment
- Review Process
- Determine next steps
  - Potential for deployment
  - Deployment architecture
  - Metrics for success of deployment

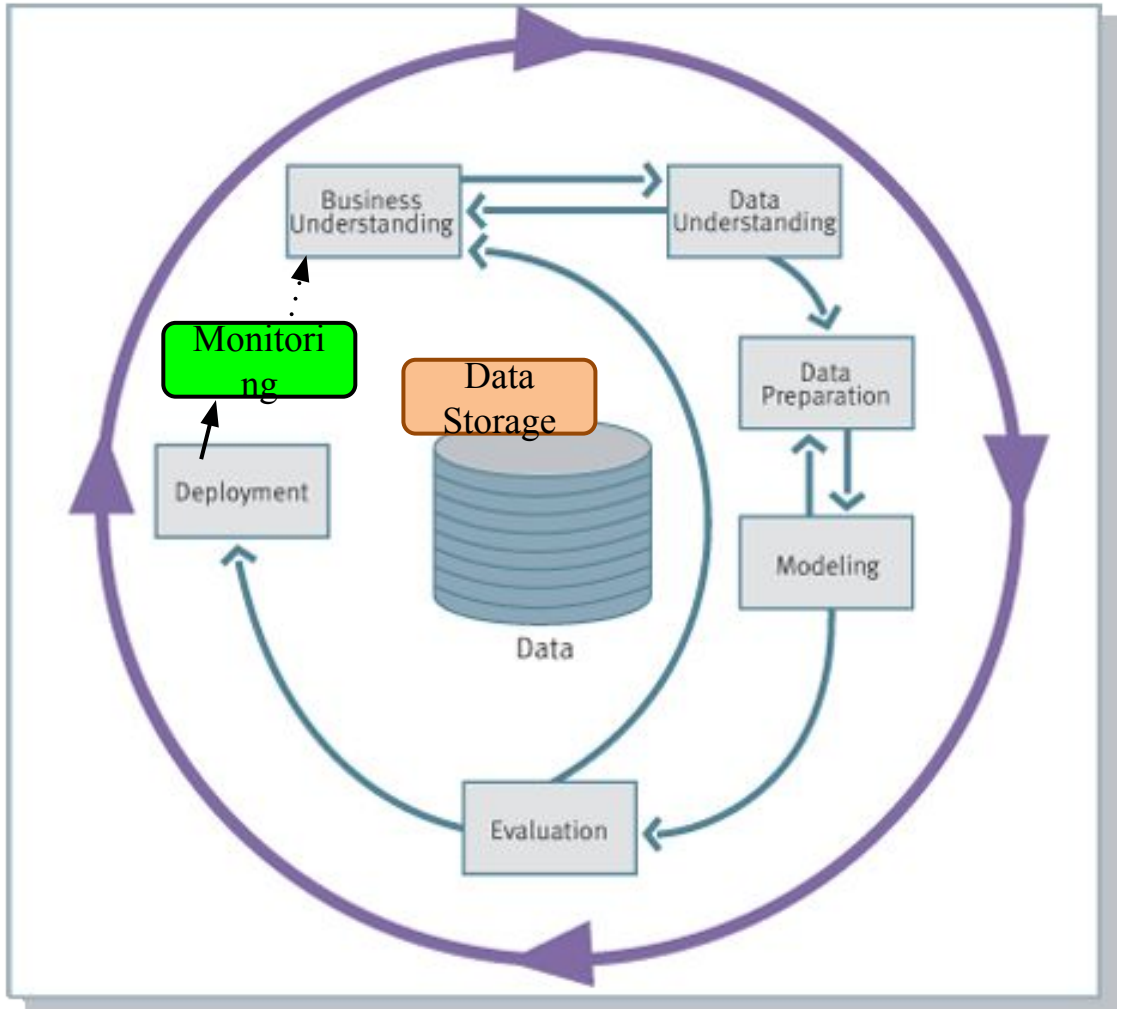
# The Deployment Phase

- Creation of the model is generally not the end of the project
- Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it
- Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process
- In many cases it will be the customer, not the data analyst, who will carry out the deployment steps
- However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models

# The Deployment Phase

- Knowledge Deployment is specific to objectives
  - Knowledge Presentation
  - Deployment within Scoring Engines and Integration with the current IT infrastructure
    - Automated pre-processing of live data feeds
    - XML interfaces to 3rd party tools
  - Generation of a report
    - Online/Offline
  - Monitoring and evaluation of effectiveness
- Process deployment/production
- Produce final project report
  - Document everything along the way

# Missing Steps – Model Monitoring and Data Management




## Closing the Loop: Model monitoring

- Changes in data
- Changes in environment

How do I know my model remains valid and applicable?
When should I update my model(s)?
How do I update my model(s)?

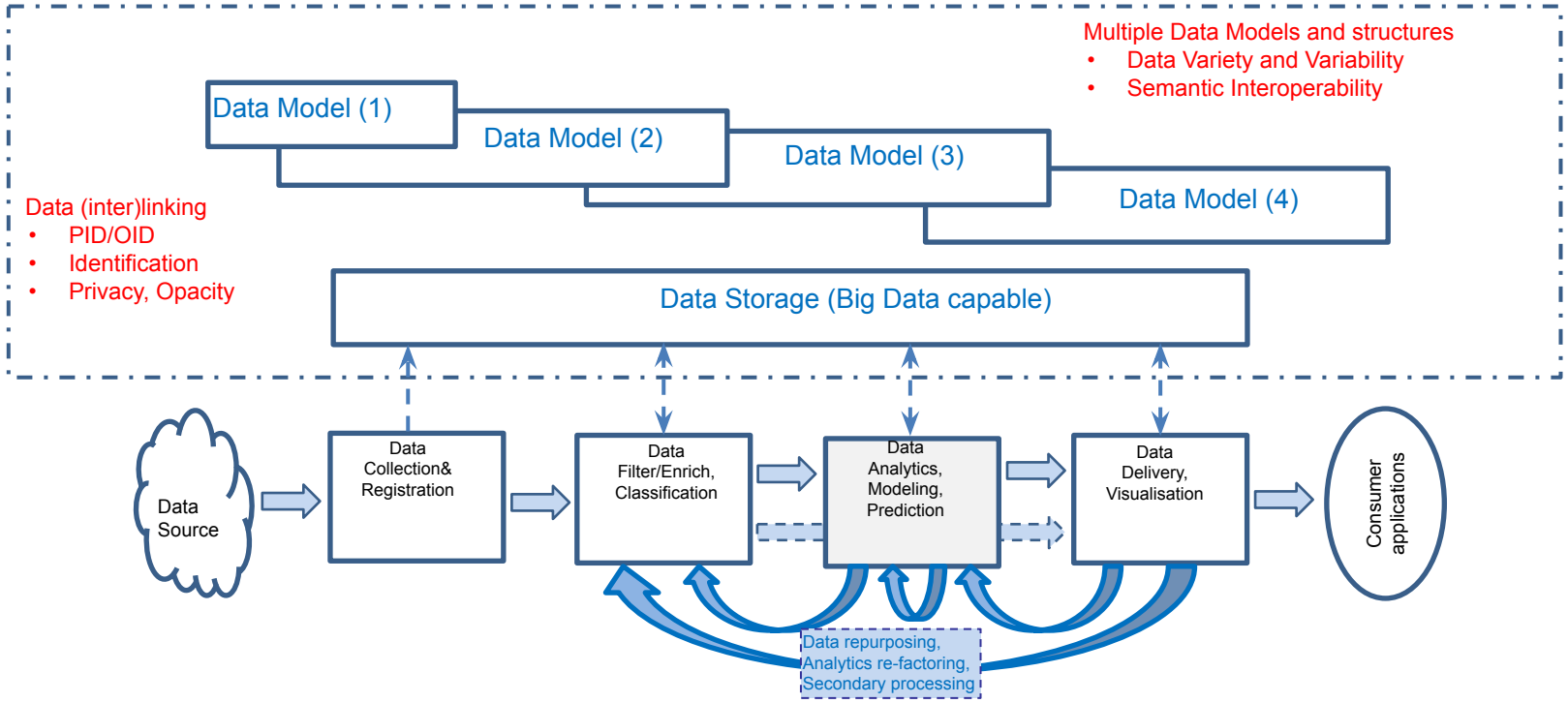
## Overall: Data Management

- Data Storage and retrieval
- Data Lineage and historical data
- Data reuse
- Digital Twins
- FAIR data principles
- Models explainability

# CRISP-DM and Data Lifecycle

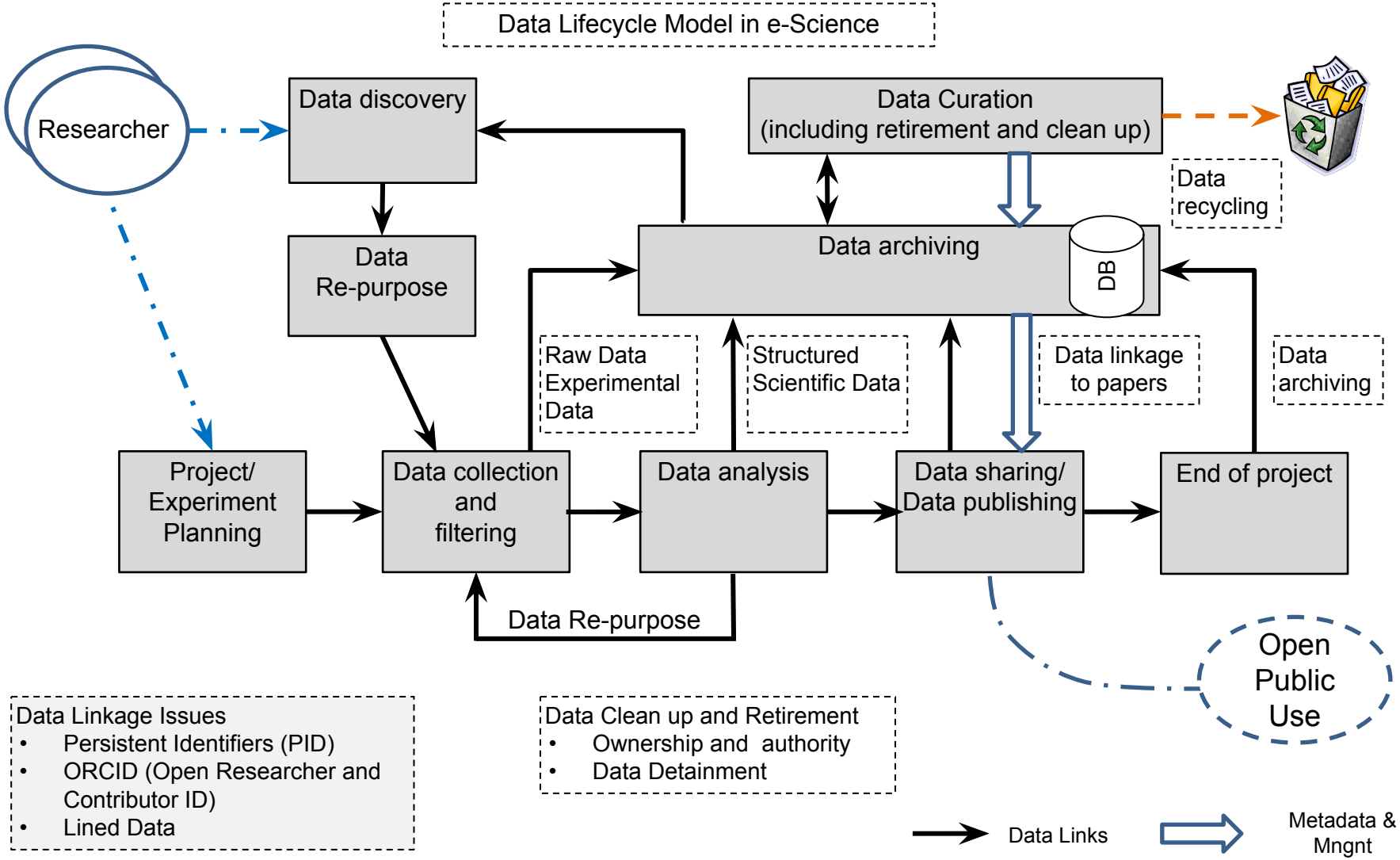
- Big (general) Data Lifecycle
- Research Data Lifecycle

# Data Lifecycle/Transformation Model




- Data Model changes along data lifecycle or evolution (Variability)
- Data provenance (lineage) is a discipline to track all data transformations along their lifecycle
- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity

# Scientific/Research Data Lifecycle Model



# Summary: CRIST-DM Phases & Tasks

## Business Understanding

**Determine Business Objectives**
*Background*
*Business Objectives*
*Business Success Criteria*

**Situation Assessment**
*Inventory of Resources*
*Requirements, Assumptions, and Constraints*
*Risks and Contingencies*
*Terminology*
*Costs and Benefits*

**Determine Data Mining Goal**
*Data Mining Goals*
*Data Mining Success Criteria*

**Produce Project Plan**
*Project Plan*
*Initial Assessment of Tools and Techniques*

## Data Understanding

**Collect Initial Data**
*Initial Data Collection Report*

**Describe Data**
*Data Description Report*

**Explore Data**
*Data Exploration Report*

**Verify Data Quality**
*Data Quality Report*

## Data Preparation

***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*

## Modeling

**Select Modeling Technique**
*Modeling Technique*
*Modeling Assumptions*

**Generate Test Design**
*Test Design*

**Build Model**
*Parameter Settings*
*Models*
*Model Description*

**Assess Model**
*Model Assessment*
*Revised Parameter Settings*

## Evaluation

**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*

## Deployment

**Plan Deployment**
*Deployment Plan*

**Plan Monitoring and Maintenance**
*Monitoring and Maintenance Plan*

**Produce Final Report**
*Final Report*
*Final Presentation*

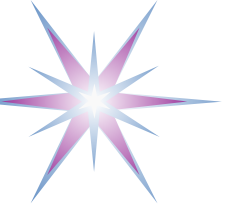
**Review Project**
*Experience Documentation*

# Summary and Takeaway

- Data Science is a creative process
- To effectively work as a data Scientist you need to know
  - Research methods
  - Statistics
  - Data Science and Data Analytics tools
- CRISP-DM is widely recognized model to Data Analytics that defines sequential step in performing data analytics projects from data and business understanding to model building and evaluation
- Practical Data Science and Analytics projects require consistent data management during the whole data lifecycle

# Practice Part – Investigating Selected Dataset

- Use self-study exercises provided for this course both in Python and RapidMiner
- Investigate and visualize dataset characteristics



# References