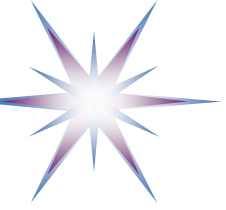


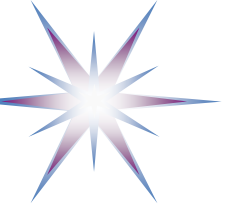
Research Methods
and
Research Process Models

Yuri Demchenko
SLICES Summer School
13-15 June 2023, Oulu, Finland



Outline

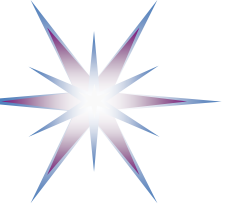
- Research methods: Importance for Scientific Research
- Experimental Research Methods
- Use of AI and ChatGPT in research
 - Scientific research in transition from 4th to 5th paradigm
- Research methods and Research types
 - Research questions, Hypothesis and Hypothesis testing
- Research process models
 - CRISP-DM: Model, stages and tasks
 - IBM ASUM
 - Microsoft TDSP



Workshop Materials

- https://drive.google.com/drive/folders/1mfoZs3OXOx_Klhy1r6-YVXIW_4MtadFh?usp=sharing





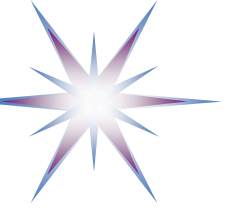
Research Methods (in Scientific Research)

- The scientific method is a body of **techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge** [ref4, ref5, ref6].
 - The general research methods include: observational methods, experimental and simulation methods, opinion based methods.
 - **Experimental data or simulation results were used to validate the hypothesis** formulated based on initial observation or domain knowledge study.
- Modern research are **data driven and computation powered** (Big Data and Data Science – AI/ML powered).
- The traditional research process may be concluded with the scientific publication and archiving of collected data.
- Data driven and data powered/driven research paradigm allows **research data re-use** and **combining them with other linked data sets** to reveal new relations between initially not linked processes and phenomena.

[ref 4] E. Bright Wilson Jr., An Introduction to Scientific Research, Dover Publications; Rev Sub edition, January 1, 1991

[ref 5] Scientific Methods, Wikipedia [online] https://en.wikipedia.org/wiki/Scientific_method

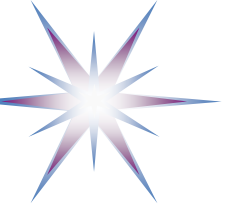
[ref 6] Research Methodology [online] <https://explorable.com/research-methodology>



Research in Computer Sciences – ICT (Information and Communication Technologies)

- Do we have any undiscovered properties?
- Do we expect to discover any unknown phenomena?
- Any other inhibiting/discouraging question?

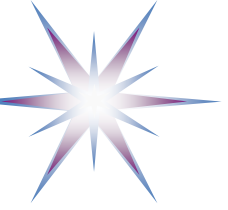
- Possible terra incognita:
 - Behavior of very large system of different levels of complexity and interaction
 - Behavior of large systems with limited resources
 - Behavior of AI powered/driven systems



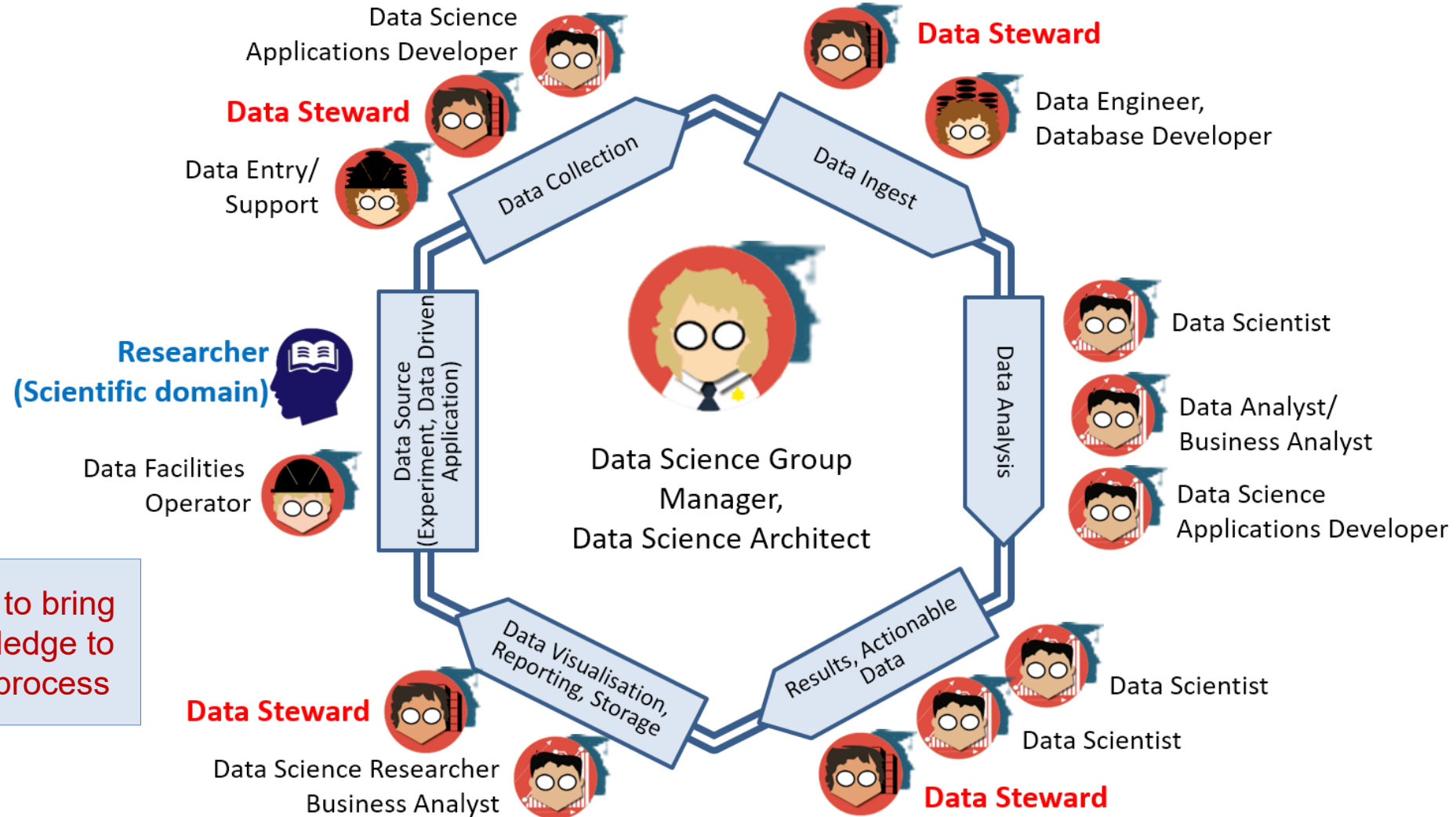
General steps in the Research process

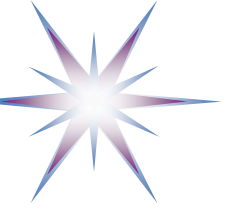
The research process is **iterative** by its nature and allows scientific model improvement by using a continuous research cycle that typically includes the following basic stages:

- Define research questions
- Design experiment representing an initial model of research object or phenomena
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis
- Refine model, publish if consistency achieved, and start new experiment cycle



Research and Data Science Team – Variety of Roles (Example)

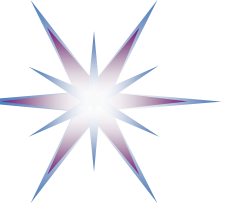




Experimental Research: General methods

Experimental research methods are characterized by the systematic manipulation of variables and control over the research environment to establish cause-and-effect relationships.

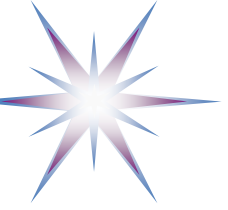
1. Research Hypotheses
2. Independent and Dependent Variables
3. Experimental Design
4. Control Group
5. Randomization
6. Experimental Manipulation
7. Measurement Instruments
8. Data Collection
9. Data Analysis
10. Internal and External Validity
11. Replication and Reliability
12. Ethical Considerations



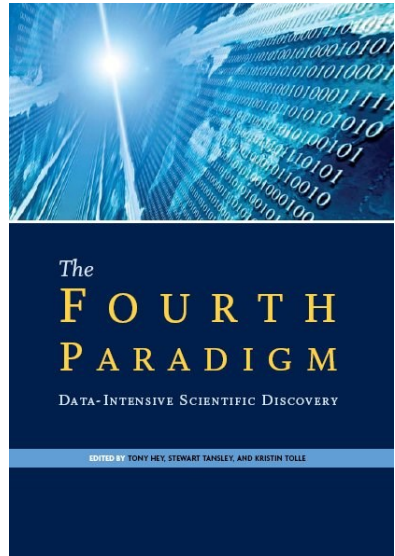
4 steps model in doing Academic Research

Approach and a model approved by British Computer Society

- Used as a template for a research proposal in Computer Science
- Hypothesis
 - Requires initial study of the domain to understand the context and initial conditions or data
- Research Methods
 - Requires initial knowledge of the different forms/types of research
- Artefact
 - Prototype or test implementation
- Validation
 - Testing working prototype and providing evidence



Big Data, Data Science and AI in Scientific Research



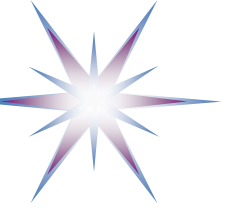
The Fourth Paradigm: Data-Intensive Scientific Discovery.

By Jim Gray, Microsoft, 2009. Edited by Tony Hey, Kristin Tolle, et al.

<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

Predicted advents of Big Data and Data Science

1. Theory, hypothesis and logical reasoning
2. Observation or Experiment
 - E.g. Newton observed apples falling to design his theory of mechanics
 - But Galileo Galilei made experiments with falling objects from the Pisa leaning tower
3. Simulation of theory or model
 - Digital simulation can prove theory or model
4. **Data-driven Scientific Discovery (aka Data Science)**
 - **More data beat hypnotized theory**
 - **e-Science as computing and Information Technologies empowered science**



Scientific Research in the transition from the 4th paradigm to 5th paradigms of AI powered research

0. Impersonation and Religion (theology and reasoning)

1. Theory, hypothesis and logical reasoning

2. Observation or Experiment

- E.g. Newton observed apples falling to design his theory of mechanics
- But Galileo Galilei made experiments with falling objects from the Pisa leaning tower

3. Simulation of theory or model

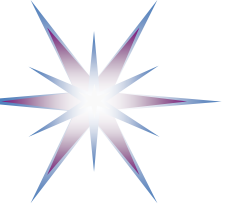
- Digital simulation can prove theory or model

4. Data-driven Scientific Discovery (aka Data Science)

- More data beat hypnotized theory
- e-Science as computing and Information Technologies empowered science

5. AI (Computer-human) driven science powered with **AI instruments**

- Based on large AI/ML models
- Machine discovers new patterns and formulates hypothesis in one or multiples knowledge spaces (synthesizing multiples knowledge domains)
- Scientist validates (asks questions) and designs additional tests or experiments
 - Danger that AI will intend to defend “own idea”



Using AI and ChatGPT in Research

- AI and ChatGPT is here, available and no reason to deny its using
 - Obviously, regulation and ethical norms are urgently needed
 - Many publications on using Chat GPT for education and research
- Many experimented with ChatGPT
 - What is your experience and examples?
 - You need to do your own critical evaluation and understand its limits
- ChatGPT represents a good snapshot of common knowledge
 - What level: literacy, informed, generalist, expert?
 - Can be defined as a Knowledge horizon, e.g. ChatGPT3 trained 2019, ChatGPT3.5 trained 2021, new ChatGPT4 released 14 March 2023

ChatGPT Cheat Sheet

By @hasantoxr

Define ChatGPT Role

- Act as a Linux Terminal
- Act as "position" Interviewer
- Act as a JavaScript Console
- Act as an Excel Sheet
- Act as an English Teacher
- Act as a Plagiarism Checker
- Act as an Advertiser
- Act as a Relationship Coach
- Act as a Recruiter

Learn from ChatGPT

- Explain clearly
- Explain uniquely
- Explain detailed
- Explain like I'm 5
- Explain with examples
- Explain to 5th grades
- Explain like Elon Musk
- Explain detailed with examples
- Explain to high school students

Chained Prompting

- Write an article about ChatGPT.
- First give me the outline, which consists of a headline, a teaser, and several subheadings.
[Output]
- Now write 5 different subheadings.
[Output]
- Add 5 keywords for each subheading.
[Output]

Prompts for Marketers

- Can you provide me with some ideas for blog posts about [topic].
- Write a product description for my [product or service or company]
- Suggest inexpensive ways I can promote my [company] without using social media.
- How can I obtain high-quality backlinks to raise the SEO of [website name]

Prompts for Designers

- Generate examples of UI design requirements for a [mobile app].
- How can I design a [law firm website] in a way that conveys [trust and authority].
- What are some micro-interactions to consider when designing fintech app.
- Create a text-based excel sheet to input your copy suggestions. Assume you have 3 members in your UX writing team

Prompts for Developers

- Develop an architecture and code for a <description> website with JavaScript.
- Help me find mistakes in the following code <paste code below>.
- I want to implement a sticky header on my website. Can you provide an example using CSS and JavaScript?
- Please continue writing this code for JavaScript <post code below>

Learn ChatGPT API and functions

- Tasks
 - NLP Tasks
 - Code
 - Structured Output Styles
 - Unstructured Output Styles
 - Media Types
 - Meta ChatGPT
 - Expert Prompting
- Writing style
 - Academic writing
 - Creative writing
 - Business writing
 - Social media
- https://bigdataanalyticsnews.com/wp-content/uploads/2023/02/ChatGPT_Cheatsheet_Costa.pdf
- <https://exposedata.com.au/openai-cheat-sheet-gpt-dall-e-codex-copilot-openai-api-playground-and-prompting/>

Writing

Creativity

Academic Life

Spreadsheet

Foreign Language

Content Creation

Data Science

Programming

Blog Posts/Essays/Articles:

Write a [adjective] [type of content] why: [topic1], topic2, ...]

- Write a friendly blog post on weight. Explain why obesity is a can lose weight, and why being
- Write a humorous essay on how exam.

Writing Assistant:

Can you [action] my text [adjective]?

- Can you make my paragraph less
- Can you make my sentence easier
- Can you complete my sentence?
- Can you continue my paragraph?

Copywriting:

Write a [type of content] for [subject] [benefit1], benefit2, ...].

- Write a social media ad for showcasing its simplicity and design
- Create a sales copy for a travel the price, destinations and comm

Email:

Reply to this email: [insert email]. B [ideas].

- Reply to this email. Be polite and offer if ...
- Reply to this email. Be friendly have time for the meeting

Social Media:

Create X [type of content] for [platform]

- Create 5 hashtags and a caption Instagram of a cute cat
- Create 2 bios for LinkedIn h achievements

Formula Generation in Excel:

Create a formula to [goal] in cell [range]

- Create a formula to calculate the average of A10
- Create a formula to calculate the sum of H10 to H20
- Create a formula that sums on have the status "No" in column

Explain formula

Explain the meaning of the formula

- Explain the meaning of the formula =SUM(A1:A10)
- Explain the meaning of the formula =AVERAGE(H10:H20)
- Explain the meaning of the formula =SUMIF(C1:C13, "No", B1:B13)

Advanced Formulas, Macros and VBA

- I have a table with phone numbers have the format: (xxx) xxx-xxx within parenthesis using Excel
- Create a macro that calculate the sum of B1 to B10 and inserts the result
- Create a macro that sorts the data by VBA.

Google Sheets

Create a formula to [goal] in cell [range]

- Create a formula in Google Sheets to sum in cells A1 to A10
- Create a formula in Google Sheets to move data from one sheet into the current sheet
- Explain this Google Sheet comment

Ask questions:

Coding Questions:

How to [action] in [programming language]?

- How to merge dictionaries in Python?
- How to join two datasets in Python?

What's the difference between [concept1] and [concept2] in [language]?

- What's the difference between lists and dictionaries in Python?
- What's the difference between OOP and functional programming?

Explain the [concept] in [language]

- Explain the while loop in Python
- Explain the try/except in Python to a child

Advice:

I want to learn [language] for [field]. Create a roadmap to learn [language]

- I want to learn Python for data science. Create a roadmap to learn Python
- I want to learn JavaScript for web development. Create a roadmap to learn it
- I want to automate [app]. Provide steps to [task] with [language]
- I want to automate Gmail. Provide steps to sending emails with Python
- I want to automate Tinder. Provide steps to swipe right with Python

Simulate a job interview for a [position] by asking and answering questions as if you are a potential employer and I'm the candidate. Consider that [details]

Web Development:

I want to build a [type of website]. Can you generate some HTML for my website?

- I want to build a job board website. Can you generate some HTML for my website?
- I want to build a price comparison website. Can you generate some HTML for my website?

Additional steps:

- Can you center the content in the HTML page, add a line break between the inputs and add an image on a logo?
- Add CSS to customize the size of the logo image

Automation:

- Send the message "[type message]" on Whatsapp using Python and pywhatkit
- Send an email from [email_1] to [email_2] with the subject [email subject] and the content [email content] using Python
- Use Python to rename all the [file extension] files in my directory. Add the word [word] to each name

Data Science Questions:

Where can I find data for my [describe project]?

- Where can I find data for my credit card fraud detection project?
- Where can I find data for my customer churn prediction project?

How to [action] my dataset?

- How to oversample/undersample my dataset?
- How to split my dataset into train and test sets with Python?

Best practices:

Can you [best practice] my code?

- Can you optimize the following Python code?
- Can you simplify the following R code?
- Can you document my code?
- Can you check for readability in my code?

Automate Data Science Tasks:

- Translate my Python code to R
- Turn this SQL code to Python pandas code
- Format my SQL code by converting the reserved keywords to uppercase
- Create a SQL query that [describe query]
- Write a regex in Python that [describe regex]

Data Science Workflow:

Web scraping (only working for a few websites)

- Web scrape [website] using [language] and [library]

Generate data

- Act as a fake data generator. Create a dataset that has X rows and Y columns: [column1], column2, ...]

Train Classification model

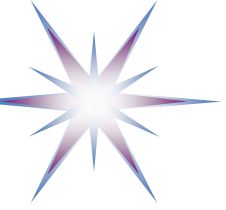
- I have a dataset of [describe dataset], build a machine learning model that predicts [target variable]

Tune Hyperparameter

- I've trained a [model name]. Write code to tune the hyperparameters

Feature Importance

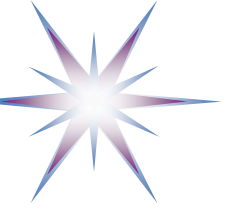
- I've trained a [model name]. Write code to find the most important features in the model



ChatGPT3.5: Experimenting and use examples

ChatGPT answer: **ChatGPT can be used in various ways for scientific research?**

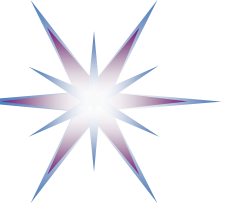
- Literature Review Assistance:** ChatGPT can help researchers in conducting literature reviews by summarizing scientific papers, extracting relevant information, and suggesting additional sources to explore. It can assist in identifying key concepts, methodologies, and findings across a wide range of research areas.
- Experimental Design:** Researchers can use ChatGPT to discuss and refine experimental designs. By describing the goals, variables, and constraints, researchers can receive feedback and suggestions from ChatGPT, which can provide insights, alternative approaches, and considerations that may have been overlooked.
- Hypothesis Generation:** ChatGPT can be a useful tool for brainstorming and generating research hypotheses. Researchers can engage in conversations with ChatGPT, describing their research question and background information. ChatGPT can offer insights, related concepts, and potential connections that may inspire new ideas or directions.
- Data Analysis and Interpretation:** ChatGPT can assist in data analysis tasks by providing guidance on statistical methods, visualizations, and interpretations. Researchers can discuss their datasets with ChatGPT, explore potential patterns or relationships, and receive suggestions for appropriate analysis techniques.
- Simulation and Modeling:** Researchers can collaborate with ChatGPT to develop and refine simulation models. By describing the system, variables, and parameters, researchers can explore different scenarios, discuss model assumptions, and receive feedback on the model's validity and potential improvements.
- Science Communication and Outreach:** ChatGPT can aid researchers in communicating their findings to a broader audience. It can help with writing summaries, abstracts, or press releases, providing plain language explanations, and addressing common questions or misconceptions related to the research.



How ChatGPT can be used for hypothesis generation?

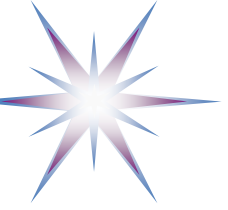
Researcher - ChatGPT

- 1. Background and Context:** Provide ChatGPT with the necessary background information and context of your research topic. Explain the problem you're investigating, any relevant existing theories or findings, and the specific area where you want to generate hypotheses.
- 2. Initial Question:** Pose a question or describe the research gap you're interested in exploring. Ask ChatGPT to help generate potential hypotheses or propose alternative explanations for the observed phenomenon.
- 3. Iterative Conversation:** Engage in a back-and-forth conversation with ChatGPT, elaborating on the question and discussing potential factors, variables, or mechanisms that might influence the phenomenon. Ask ChatGPT to brainstorm different ideas, consider alternative perspectives, and suggest relevant literature or related concepts.
- 4. Prompt Refinement:** As you receive responses from ChatGPT, refine your prompts or questions based on the generated hypotheses or insights. This iterative process helps to guide ChatGPT towards more relevant and specific hypotheses.
- 5. Evaluation and Validation:** Assess the generated hypotheses critically. Consider their plausibility, testability, and alignment with existing knowledge. Evaluate them in terms of feasibility, available resources, and potential impact.
- 6. Expert Input:** Once you have a set of generated hypotheses, it's crucial to involve domain experts or researchers with expertise in the relevant field. Discuss the generated hypotheses with them, seek their input, and incorporate their feedback to refine and prioritize the hypotheses.



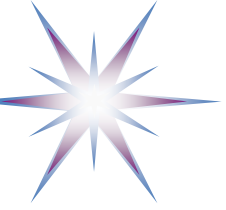
Rather common knowledge but useful for understanding the baseline and wider perspective

1. Hypothesis 1&2: **Optimizing computational workflows** in the digital research infrastructure will lead to a decrease in energy consumption, resulting in improved energy efficiency and carbon footprint.
2. Hypothesis 3: Implementing **efficient resource allocation** techniques and workload balancing algorithms in computational workflows will result in reduced power consumption and subsequently lower greenhouse gas emissions.
3. Hypothesis 4: By **optimizing software and algorithms** in the digital research infrastructure, computational tasks will be completed in a shorter time, leading to reduced energy consumption and a smaller climate impact.
4. Hypothesis 5: Integration of **advanced cooling and power management** techniques in the digital research infrastructure will optimize energy usage and reduce carbon emissions associated with excessive cooling and inefficient power distribution.
5. Hypothesis 6: Implementing data management strategies that prioritize **efficient data storage, retrieval, and transmission** in the digital research infrastructure will minimize energy requirements and, in turn, decrease the carbon footprint.
6. Hypothesis 7: **Adoption of cloud computing and virtualization technologies**, combined with intelligent workload scheduling, will enable better resource utilization, resulting in energy savings and reduced climate impact.
7. Hypothesis 8: Enhancing hardware efficiency through the **use of energy-efficient processors**, accelerators, and low-power components in the digital research infrastructure will lead to significant energy savings and a reduced environmental footprint.



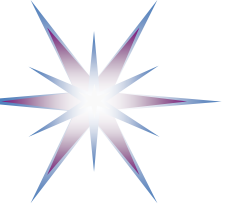
Formulate few hypotheses on the effect of wide 5G/6G technologies implementation to reduce climate impact of the digital research infrastructure

1. Hypothesis 1: The widespread implementation of 5G/6G technologies in the digital research infrastructure will lead to increased energy efficiency compared to previous generations, resulting in a reduced climate impact.
2. Hypothesis 2: **5G/6G technologies' advanced network management capabilities**, such as dynamic spectrum sharing and intelligent power control, will optimize energy consumption in the digital research infrastructure, leading to a decrease in greenhouse gas emissions.
3. Hypothesis 3: The **use of ultra-dense small cell deployments** in 5G/6G networks will enable more localized and efficient data transmission, reducing the energy required for long-distance data routing and resulting in a smaller carbon footprint.
4. Hypothesis 4: The implementation of **advanced communication protocols and coding techniques in 5G/6G networks** will enhance spectral efficiency, allowing for higher data throughput per unit of energy consumed, thereby reducing the climate impact of digital research infrastructure.
5. Hypothesis 5: The deployment of **edge computing infrastructure in 5G/6G networks** will reduce latency and enable more localized data processing, minimizing the need for long-distance data transfers and reducing energy consumption and carbon emissions.
6. Hypothesis 6: The integration of **energy harvesting techniques**, such as solar or kinetic energy, into 5G/6G infrastructure components, will contribute to renewable energy utilization and decrease reliance on fossil fuel-powered energy sources, thereby reducing the climate impact.
7. Hypothesis 7: The **utilization of network slicing and virtualization technologies** in 5G/6G networks will allow for more efficient resource allocation, leading to optimized energy usage and reduced environmental impact in the digital research infrastructure.
8. Hypothesis 8: The implementation of **advanced energy management systems and smart grid technologies** in 5G/6G networks will enable better monitoring and control of energy consumption, facilitating energy-efficient operations and minimizing the climate impact.



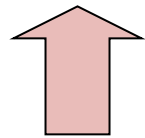
Research Process – Overview

- Research Process
- Research Method
- Research Design
- Conclusion

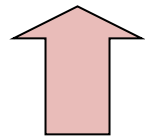


The Building Blocks of Research

KNOWLEDGE



INFORMATION

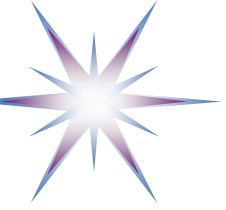


DATA

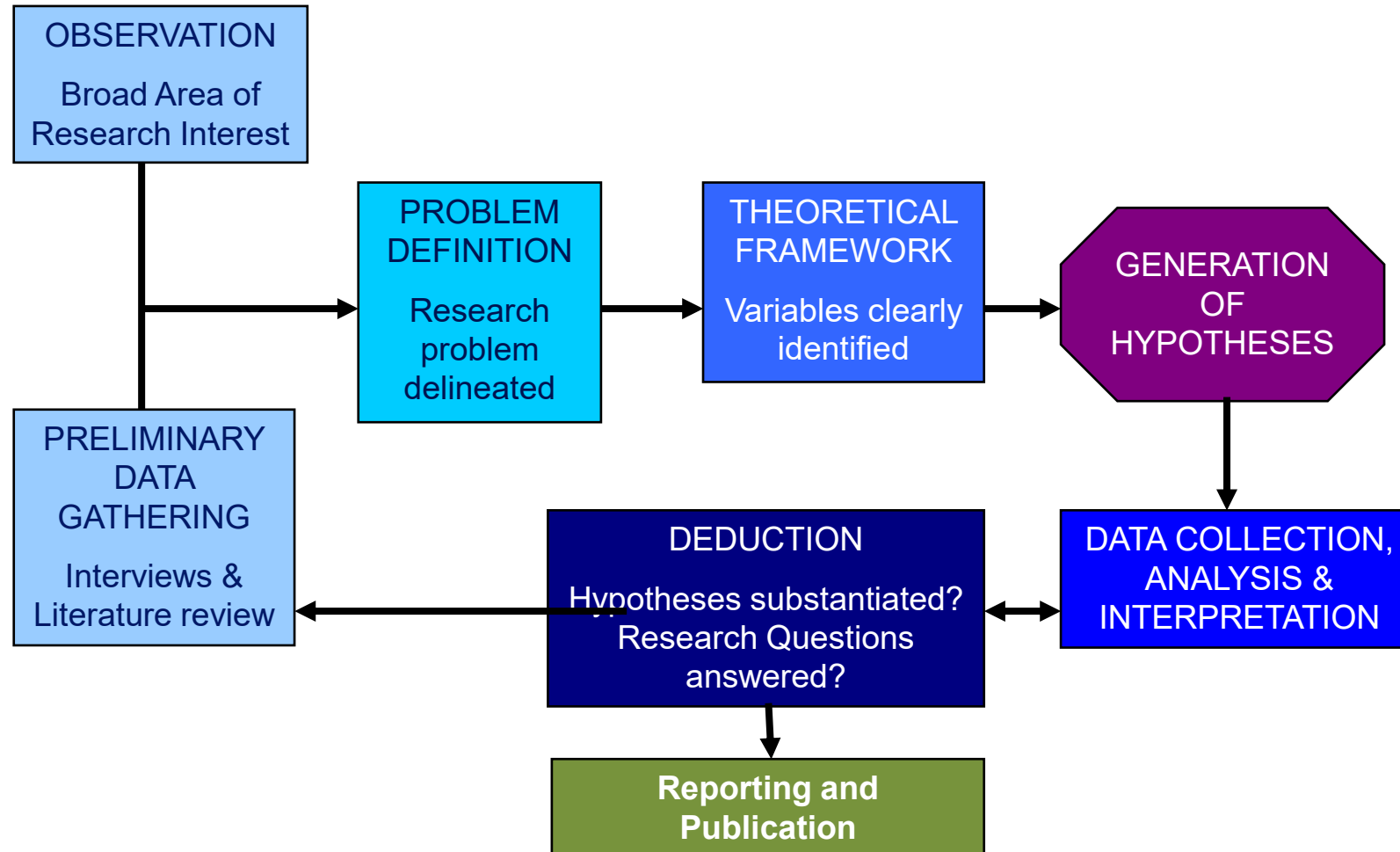
Blend of information, experience and in-sights that provides a framework that can be thoughtfully evaluated when assessing new information or evaluating relevant situations

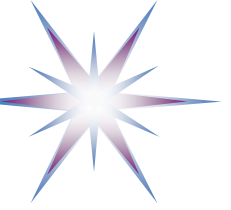
Determination of relationship amongst data with a view to facilitating understanding of the phenomena, their relationships and decision-making (e.g. past and predicted future sales trends)

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



The Research Process

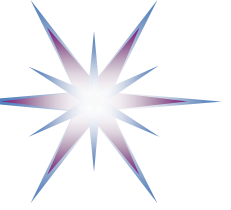




The Research Idea

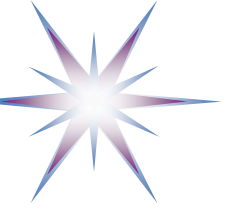
Where to get a research idea?

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



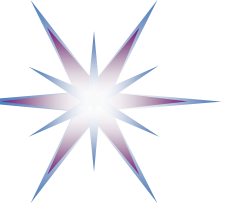
A Research Questions Must Identify

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 CRISP-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

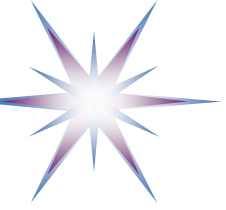


Hypothesis

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

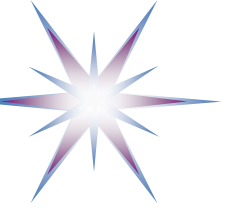
- Characteristics of hypotheses
 - Declarative statement that identifies the predicted relationship between 2 or more variables
 - Testability
 - Based on sound scientific theory/rationale

- 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 factors in a situation
- Explain the variance in the dependent variable or to predict experiment/test outcome.
- Revisit statistical hypothesis testing
 - AB testing (null hypothesis and alternative hypothesis), ANOVA and t-test



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**

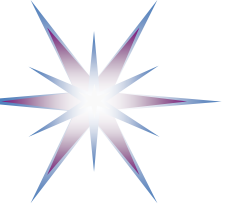
- Analysis
- Evaluation
- Design-demonstration
- Status

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 focusses on a real-life problem or situation with a view to helping reach a decision how to deal with it (Specificity)

Example: Should Corporation X adopt a paperless office environment?



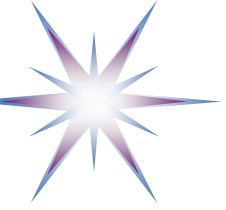
What is basic research

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

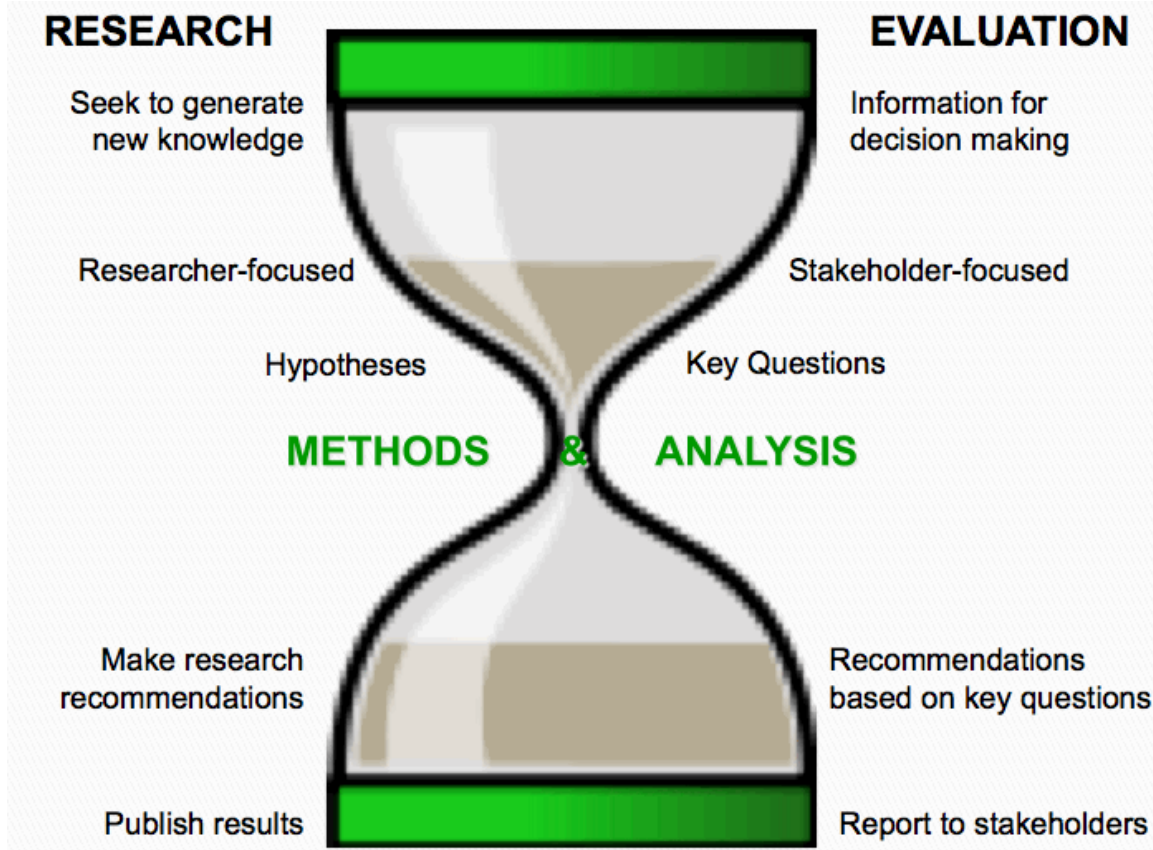
- **Exploratory research**, which structures and identifies new problems
- **Constructive research**, which develops solutions to a problem
- **Empirical research**, which tests the feasibility of a solution using empirical evidence

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

- Research is often illustrated using the hourglass model.

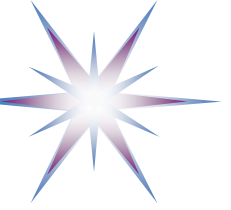


Hourglass Model



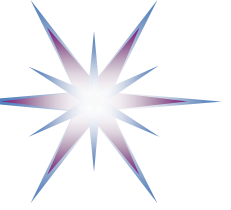
- Hypothesis checking based on key element of the suggested model
- Observable data must validate the hypothesis

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



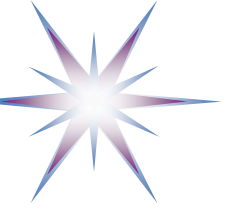
Different types of Research

- Quantitative (data in numbers)
- Qualitative (data in words)
- Exploratory research
- Constructive research
- Empirical research
- Primary research
- Secondary research
- Case studies
- Cohort research



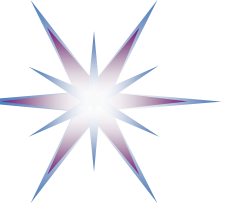
Data Driven Research Process models (after Data Science Analytics)

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- ASUM (Analytics Solutions Unified Method) by IBM
- Microsoft TDSP (Team Data Science Process)



What is CRISP-DM?

- Cross-Industry Standard Process for Data Mining
- Outcome of the European Union project under the ESPRIT funding initiative in 1997
 - Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation, and OHRA
 - Founding CRISP-DM Consortium Members: DaimlerChrysler, SPSS and NCR
- 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
- CRISP-DM Special Interest Group ~ 200 members
 - Management Consultants
 - Data Warehousing and Data Mining Practitioners

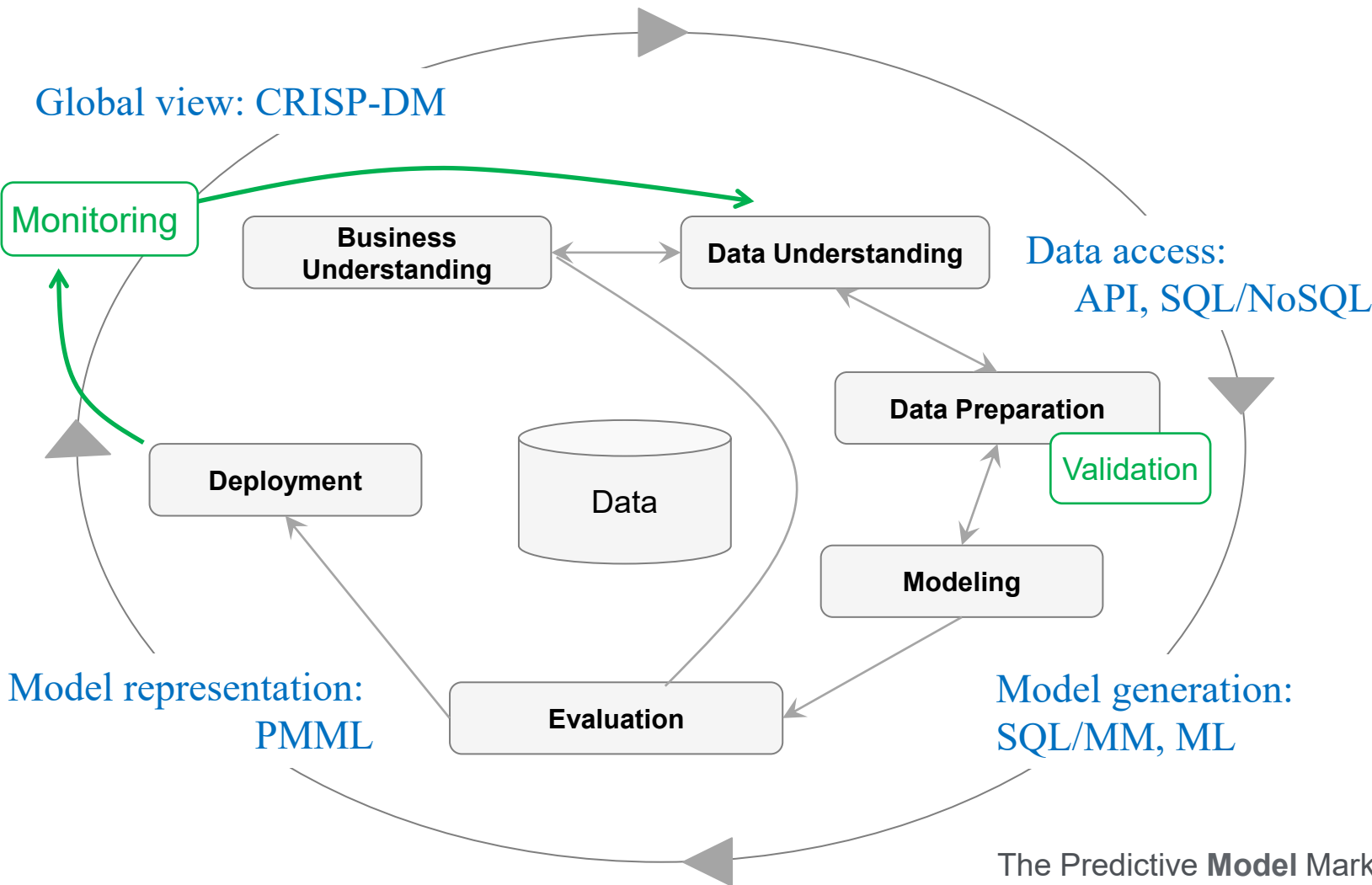


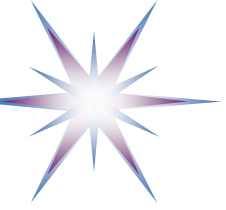
CRISP DM **Process**: Processes and Data Lifecycle (1)

Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
 - **Data Validation**
- Modelling
- Evaluation
- Deployment
 - **Process monitoring**

All stages are iterative with the goal to achieve effectiveness for business decision making



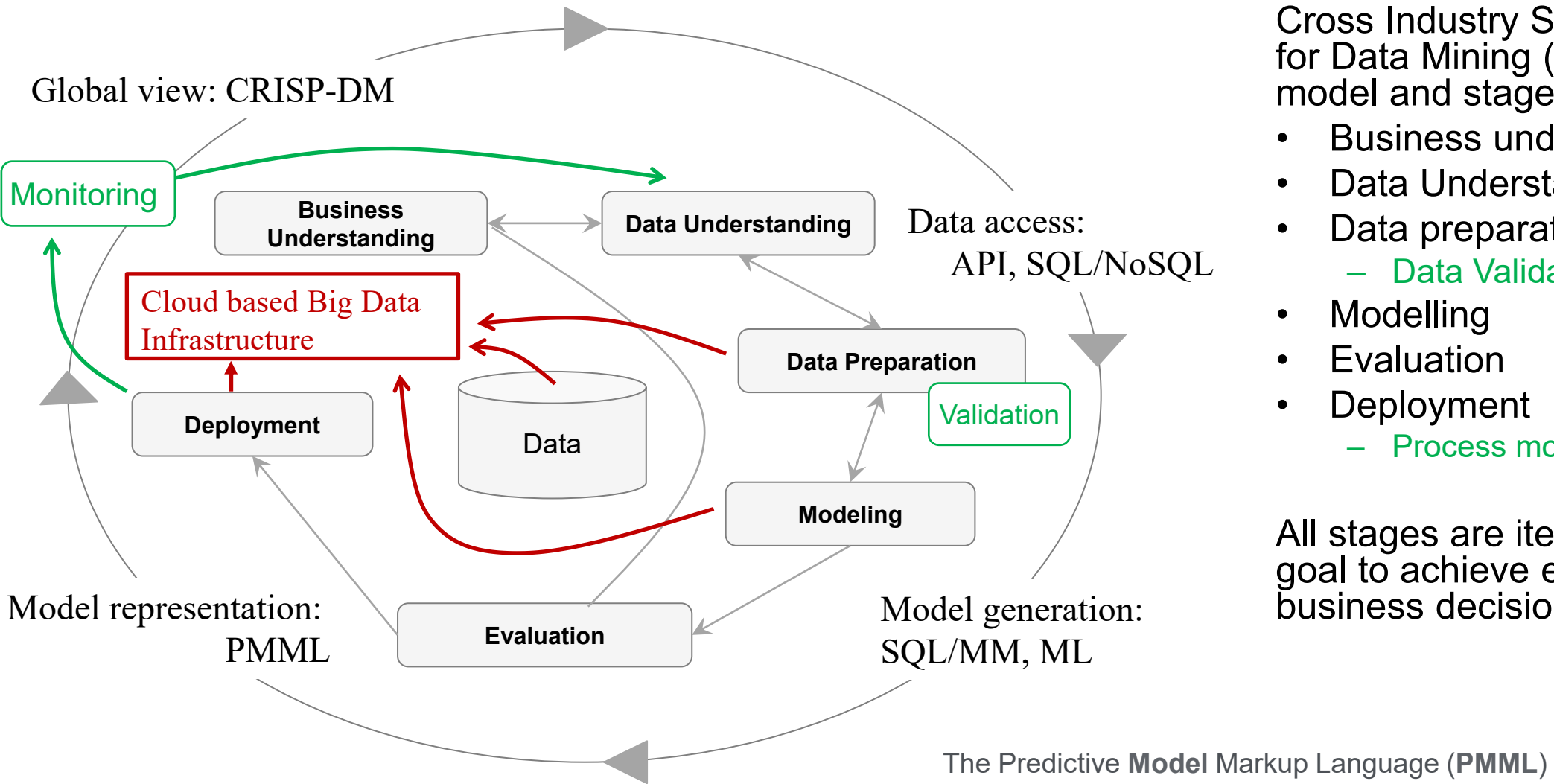


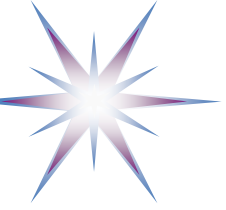
CRISP DM Process: Processes and Data Lifecycle (2)

Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
 - Data Validation
- Modelling
- Evaluation
- Deployment
 - Process monitoring

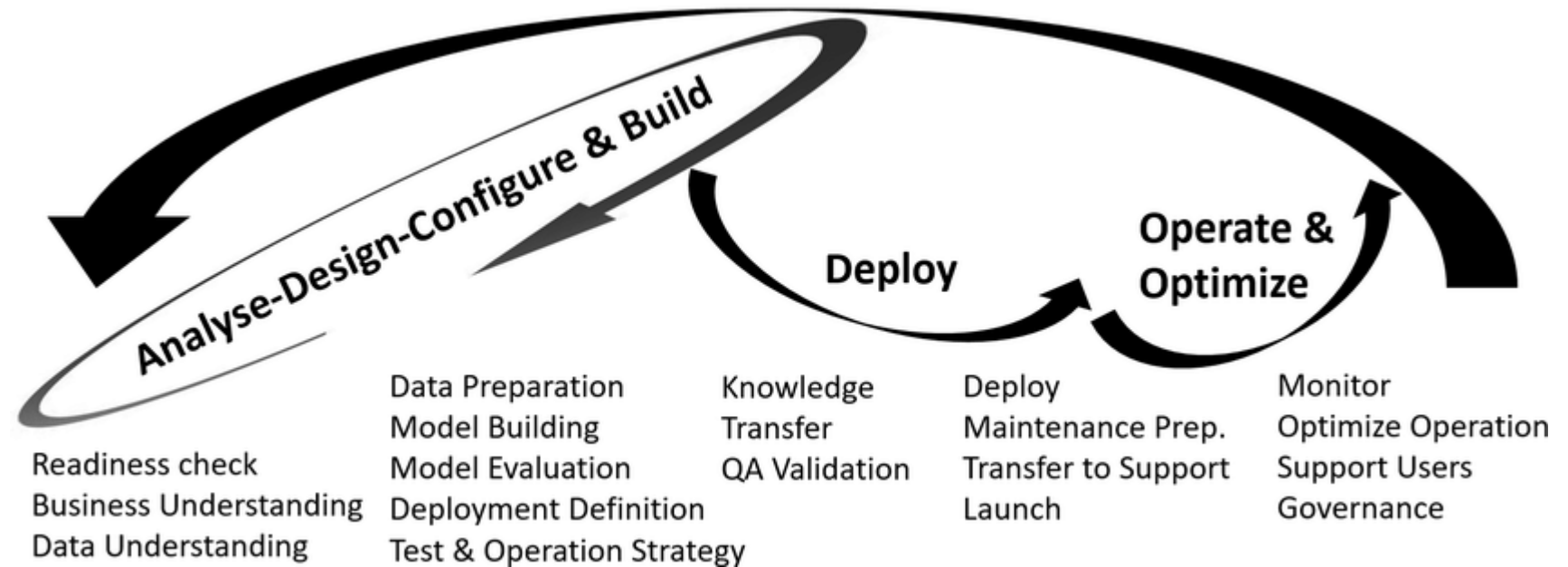
All stages are iterative with the goal to achieve effectiveness for business decision making





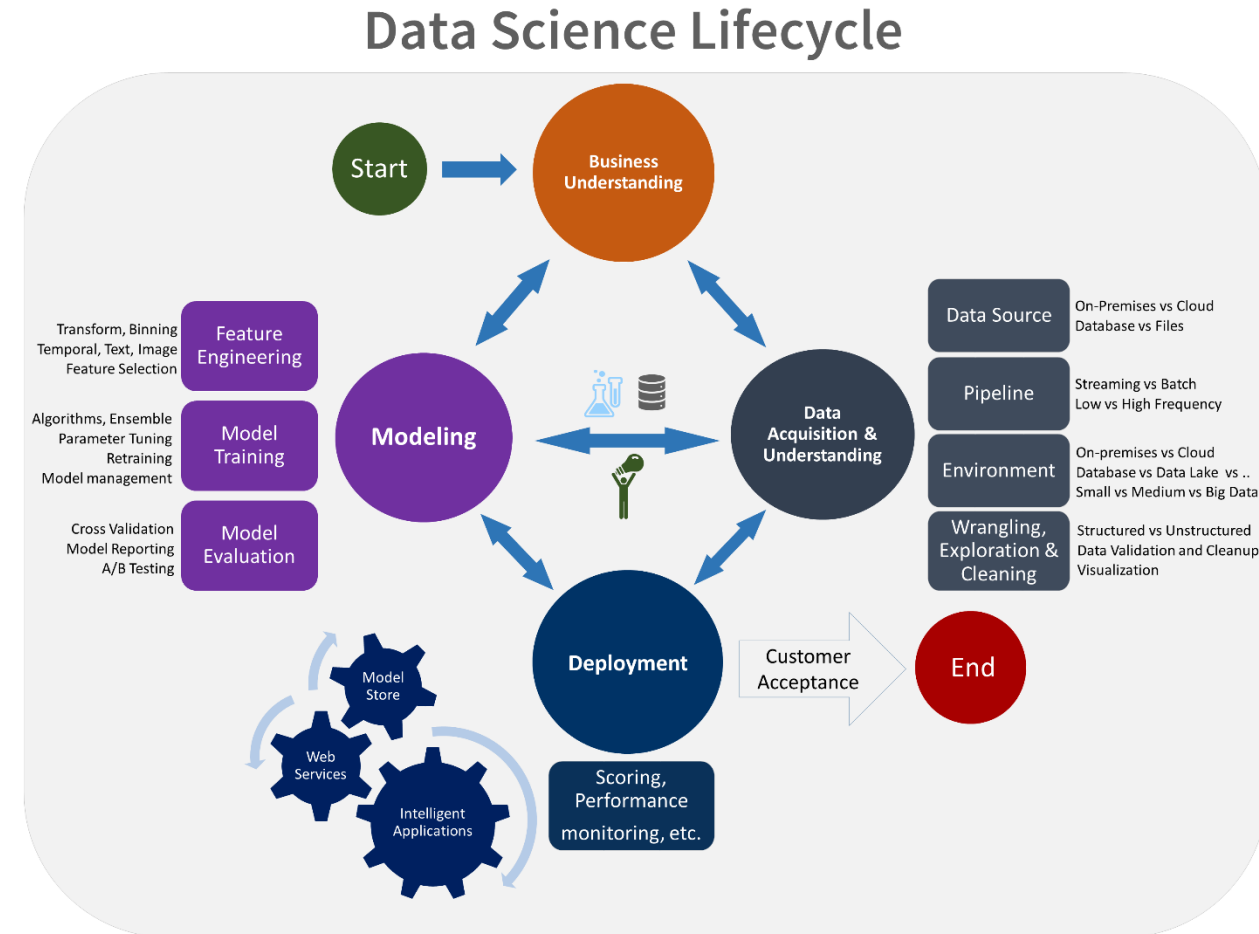
ASUM (Analytics Solutions Unified Method) by IBM

- **Analyze.** Requirements specified and agreed; contract or services agreement is signed.
- **Design.** Define all components of the solution and their relationships and dependencies, identify necessary resources.
- **Configure and Build.** The solution is developed, all components are integrated and configured.
- **Deploy.** Create a plan to run and maintain the developed solution, including configuration management and migration plan if necessary.
- **Operate and Optimize.** The solution is operational is monitoring data are collected and maintained.

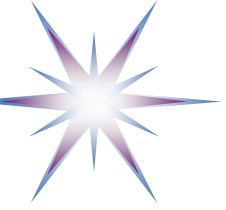


TDSP (Team Data Science Process) by Microsoft

- TDSP is an agile and iterative process model
 - Refactored into Azure DevOps in 2018
 - Currently supported by MLOps
- Includes components
 - A **data science lifecycle** definition
 - A **standardized project structure**
 - **Infrastructure and resources** recommended for data science projects
 - **Tools and utilities** recommended for project execution
- The lifecycle includes five sequential phases:
 - Business Understanding
 - Data acquisition and understanding
 - Modeling
 - Deployment
 - Customer acceptance



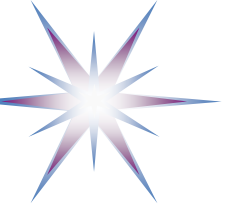
[ref] <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>



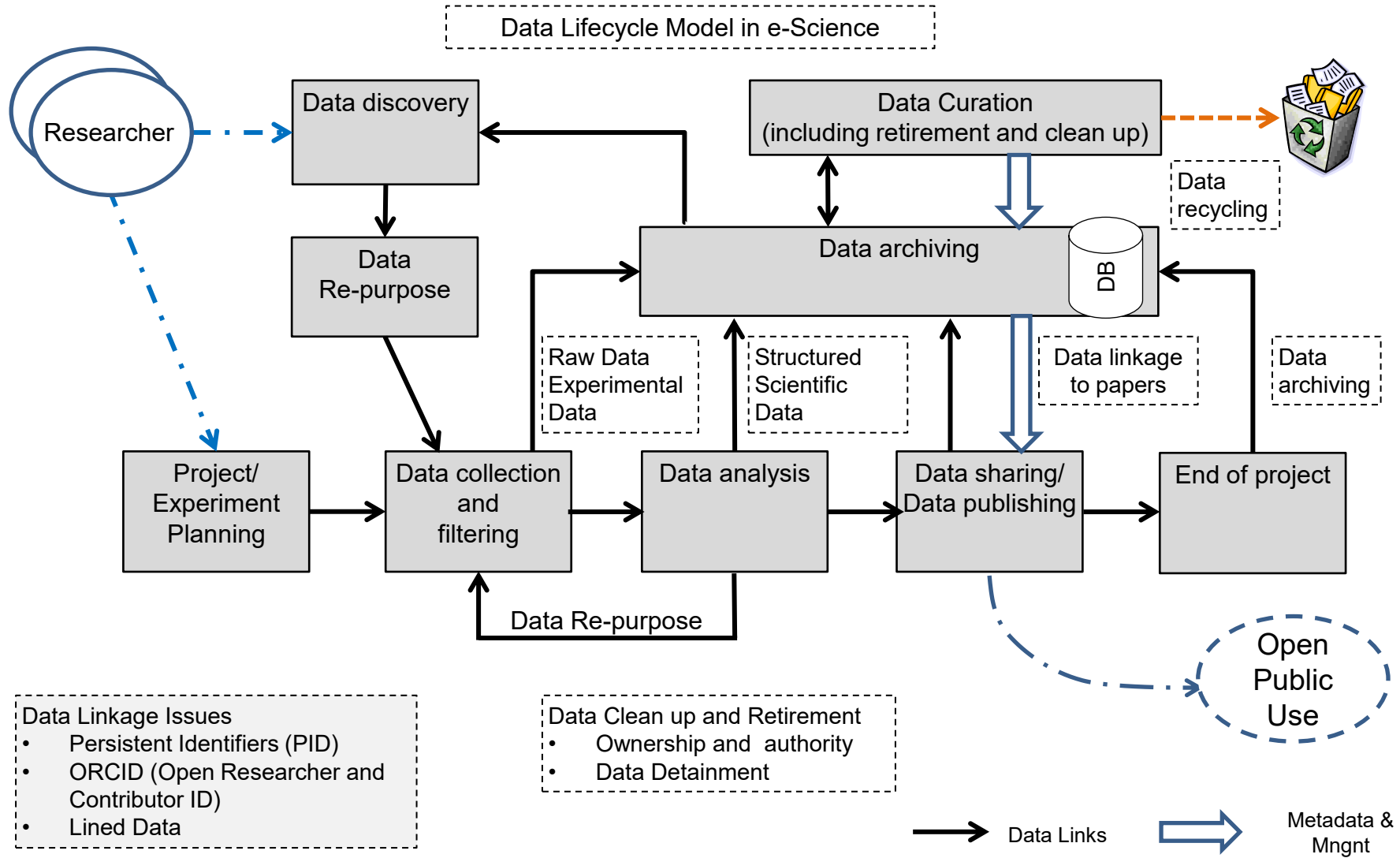
TDSP: Data Science Project Planning

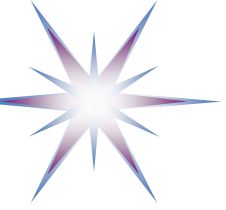
<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/team-data-science-process-project-templates>

- Microsoft Project template
- Excel template
<https://github.com/Azure/Azure-MachineLearning-DataScience/blob/master/Team-Data-Science-Process/Project-Planning-and-Governance/Advanced%20Analytics%20Microsoft%20Project%20Plan.xlsx>
- Repository template <https://github.com/Azure/Azure-TDSP-ProjectTemplate>
- Walkthroughs executing the Team Data Science Process
<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/walkthroughs>
 - [HDInsight Spark walkthroughs using PySpark and Scala](#) These walkthroughs use PySpark and Scala on an Azure Spark cluster to do predictive analytics.
 - [HDInsight Hadoop walkthroughs using Hive](#) These walkthroughs use Hive with an HDInsight Hadoop cluster to do predictive analytics.



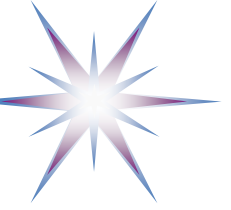
Scientific/Research Data Lifecycle Model



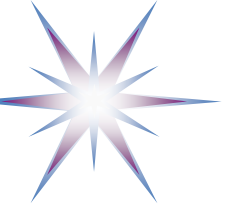


Summary and Takeaway

- Data Science is a creative process
- To effectively work as a research in data driven research you need to know
 - Research methods
 - Statistics
 - Data Science and Data Analytics tools (act as a Data Scientist)
 - General IT and Data Science projects management -> DevOps and MLOps
- 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

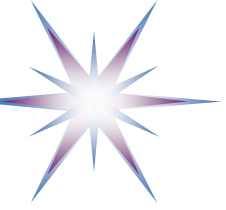


- Different types of research in details



Different types of Research

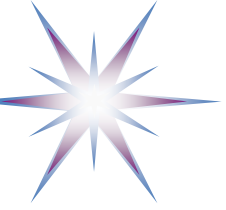
- Quantitative (data in numbers)
- Qualitative (data in words)
- Exploratory research
- Constructive research
- Empirical research
- Primary research
- Secondary research
- Case studies
- Cohort research



Quantitative Research

Data in numbers

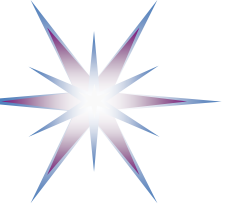
- Investigates the why and how of decision making, as compared to what, where, and when of quantitative research.
- Quantitative research is the systematic scientific investigation of properties and phenomena and their relationships.
- The objective of quantitative research is to develop and employ mathematical 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

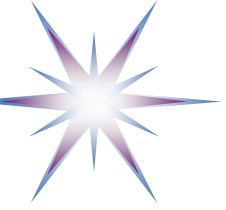
Data in words

- 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



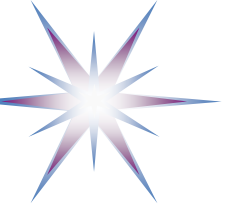
Exploratory Research

- 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's often concludes that a perceived problem does not actually exist.
 - Still an important activity: Negative results often tells more than positive



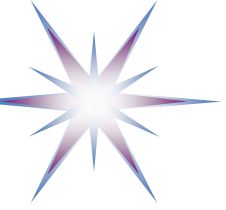
Constructive Research

- Is perhaps the most common **computer science research method**.
- This type of approach demands a form of validation that doesn't need to be quite 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



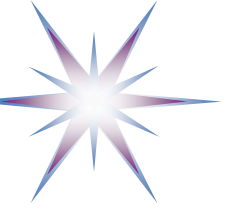
Empirical Research

- 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.



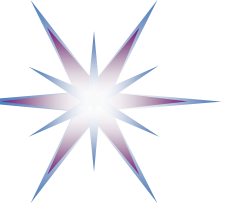
Primary Research

- **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



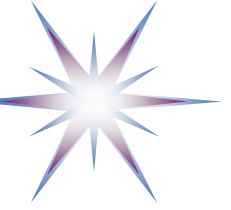
Secondary Research

- **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



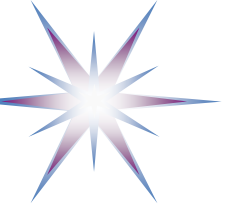
Case Study

- 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