



Data Science Lifecycle DSC8201

Week 1: Lecture 01 (MSDS_1:1)

Topic: *The Data Science Lifecycle*

Dr. Daphne Nyachaki Bitalo
Department of Computing & Technology
Faculty of Engineering, Design & Technology

A Complete Education for A Complete Person



Online class rules

1. Keep your microphone off when a participant is speaking
2. Use the “raise hand” icon to ask a question or make a remark
3. Video is optional
4. Sessions may be recorded and posted on Moodle
5. Feedback is invited on Moodle

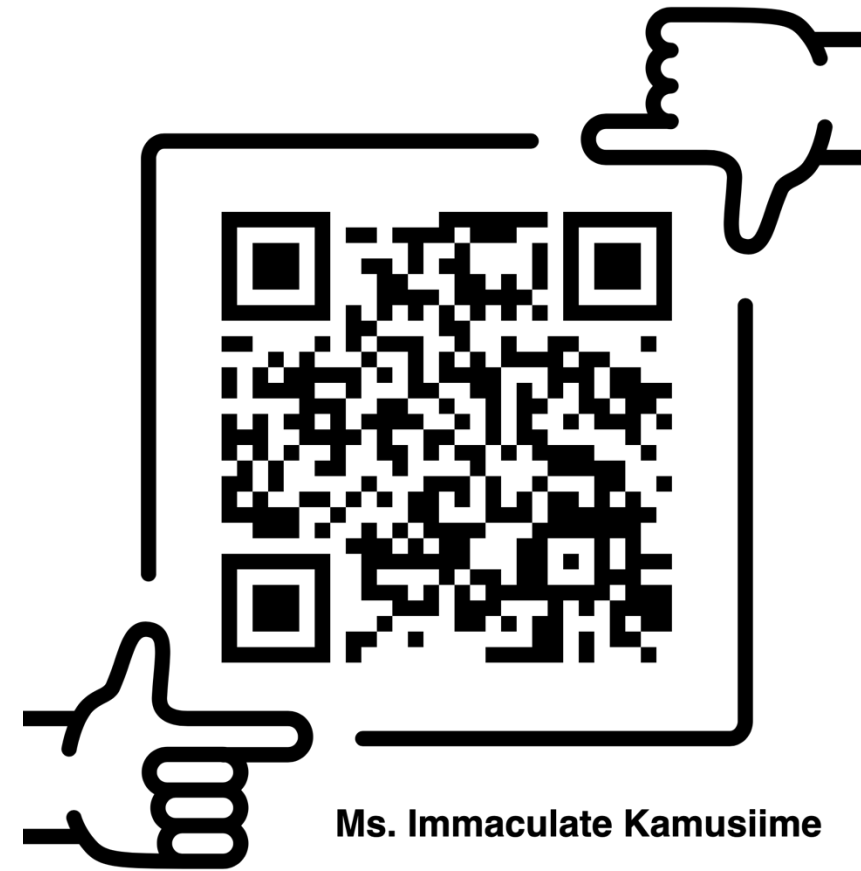




About the lecturers



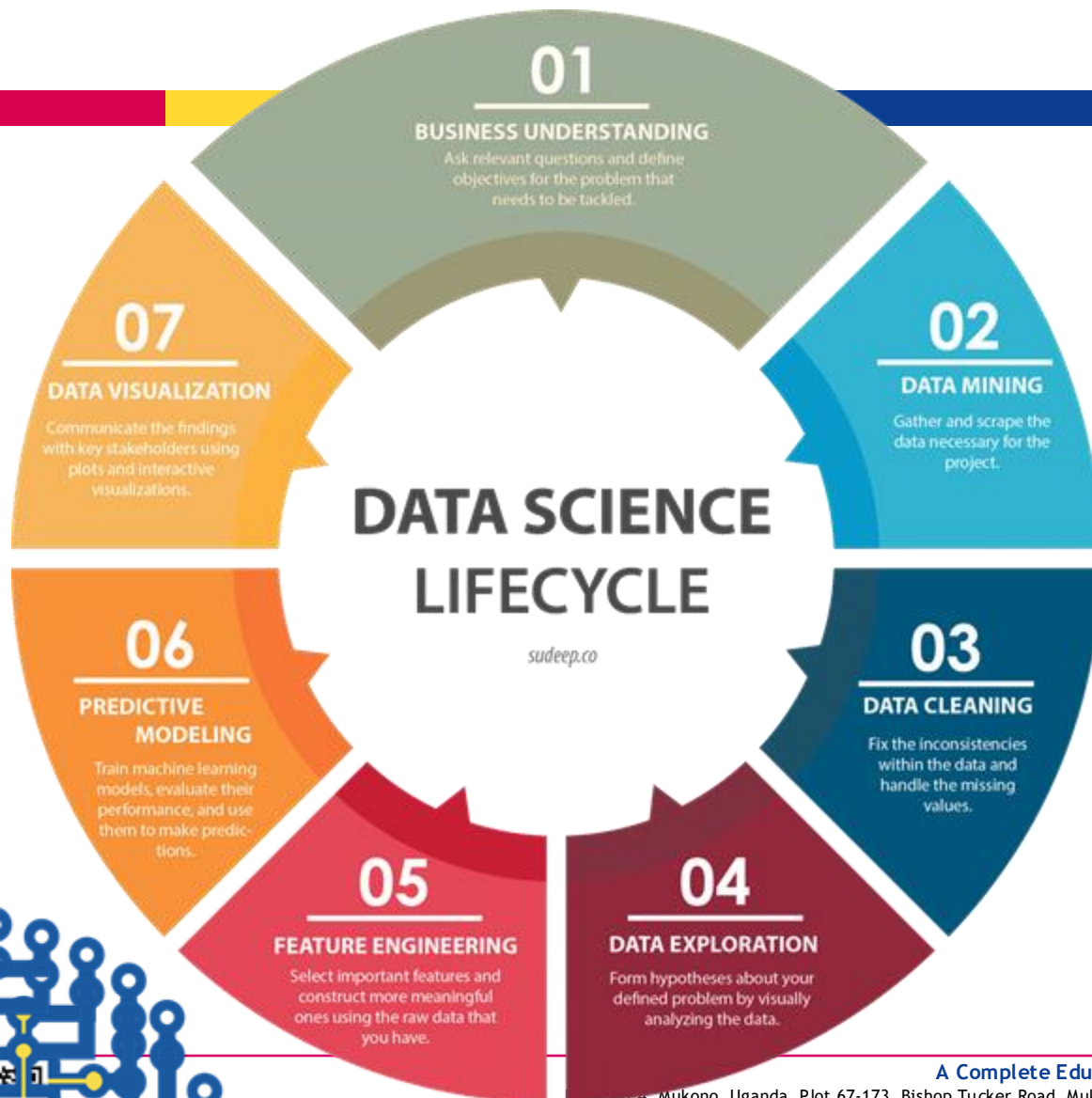
Dr. Daphne Nyachaki Bitalo



Ms. Immaculate Kamusiime



Course Outline



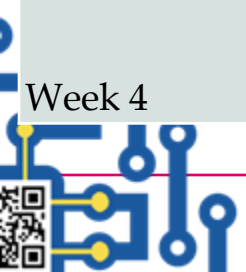
1. Interactive online classes (Thursday 5pm-9pm; Friday 5pm-9pm)
2. Practical classes (Group assignments)
3. Course Work; 70% (End of each lecture Week)
4. Exam (Project); 30% (Week 6-Week 7)

Breakdown of Course Outline

Weeks	Course Goals	Focus Area	Topic Break Down	Practical Skills/ Tools
Week 1	Data Science Fundamentals & Ethics	Foundation & Project Setup	The Data Science Lifecycle (CRISP-DM, Cross-Industry Standard Process for Data Mining, or a similar framework), problem framing, project scoping, Data Privacy	Python, pandas, numpy setup, documentation
Week 1	Data Acquisition, Wrangling, and Storage	Data Engineering	Data ingestion from various sources (APIs, SQL, NoSQL, web scraping), data cleaning techniques (missing values, outliers), feature engineering basics,	ETL/ELT pipeline design, Pandas for data manipulation, Basic use of an object storage system (e.g., S3).
Week 2	Exploratory Data Analysis & Visualization	Analysis & Storytelling	Hypothesis generation and testing, descriptive statistics, data visualization principles, statistical inference (t-tests, ANOVA), communicating data stories.	Matplotlib, Seaborn, Plotly for visualization, statistical software (e.g., StatsModels, SciPy), Storytelling with data

Breakdown of Course Outline

Weeks	Course Goals	Focus Area	Topic Break Down	Practical Skills/ Tools
Week 3	Machine Learning Engineering (MLOps I)	Modeling & Deployment	Supervised (Regression, Classification), Unsupervised (Clustering, PCA), Model selection,	Scikit-learn, MLflow/DVC for experiment tracking, FastAPI/Flask for basic deployment, Docker basics.
week 3	Deep Learning and Advanced Architectures	Advanced Techniques	Introduction to Neural Networks (NNs), CNNs (Convolutional NNs)	TensorFlow/PyTorch, leveraging pre-trained models.
Week 4	Causal Inference and Experimentation (A/B Testing)	Business Impact	The logic of A/B Testing and experimentation, design and analysis of Randomized Controlled Trials (RCTs),)	Practical A/B test setup and analysis (statistical significance),



Breakdown of Course Outline

Weeks	Course Goals	Focus Area	Topic Break Down	Practical Skills/ Tools
Week 4	Big Data and Cloud Computing for Data Science	Scalability & Infrastructure	Distributed computing frameworks (Spark/Dask), Cloud service providers	Hands-on with Spark/PySpark,
Week 5	Capstone Project/Dissertation	Synthesis & Real-World Application	Students work on a real-world problem from industry or research, applying the full lifecycle.	All tools from the program, intensive project management, technical report writing.
Week 5	Professional Practices and Technical Communication	Soft Skills & Career Prep	Project management for data science (Agile, Scrum), effective presentation of technical results to non-technical audiences, preparing technical documentation, Data Governance, interview skills and portfolio development.	Presentation software, Markdown for professional reports.

Academic Journeys are Complex



Essay

1771

The importance of stupidity in scientific research

Martin A. Schwartz

Department of Microbiology, UVA Health System, University of Virginia, Charlottesville, VA 22908, USA
e-mail: maschwartz@virginia.edu

Accepted 9 April 2008
Journal of Cell Science 121, 1771 Published by The Company of Biologists 2008
doi:10.1242/jcs.033340

I recently saw an old friend for the first time in many years. We had been Ph.D. students at the same time, both studying science, although in different areas. She later dropped out of graduate school, went to Harvard Law School and is now a senior lawyer for a major environmental organization. At some point, the conversation turned to why she had left graduate school. To my utter astonishment, she said it was because it made her feel stupid. After a couple of years of feeling stupid every day, she was ready to do something else.

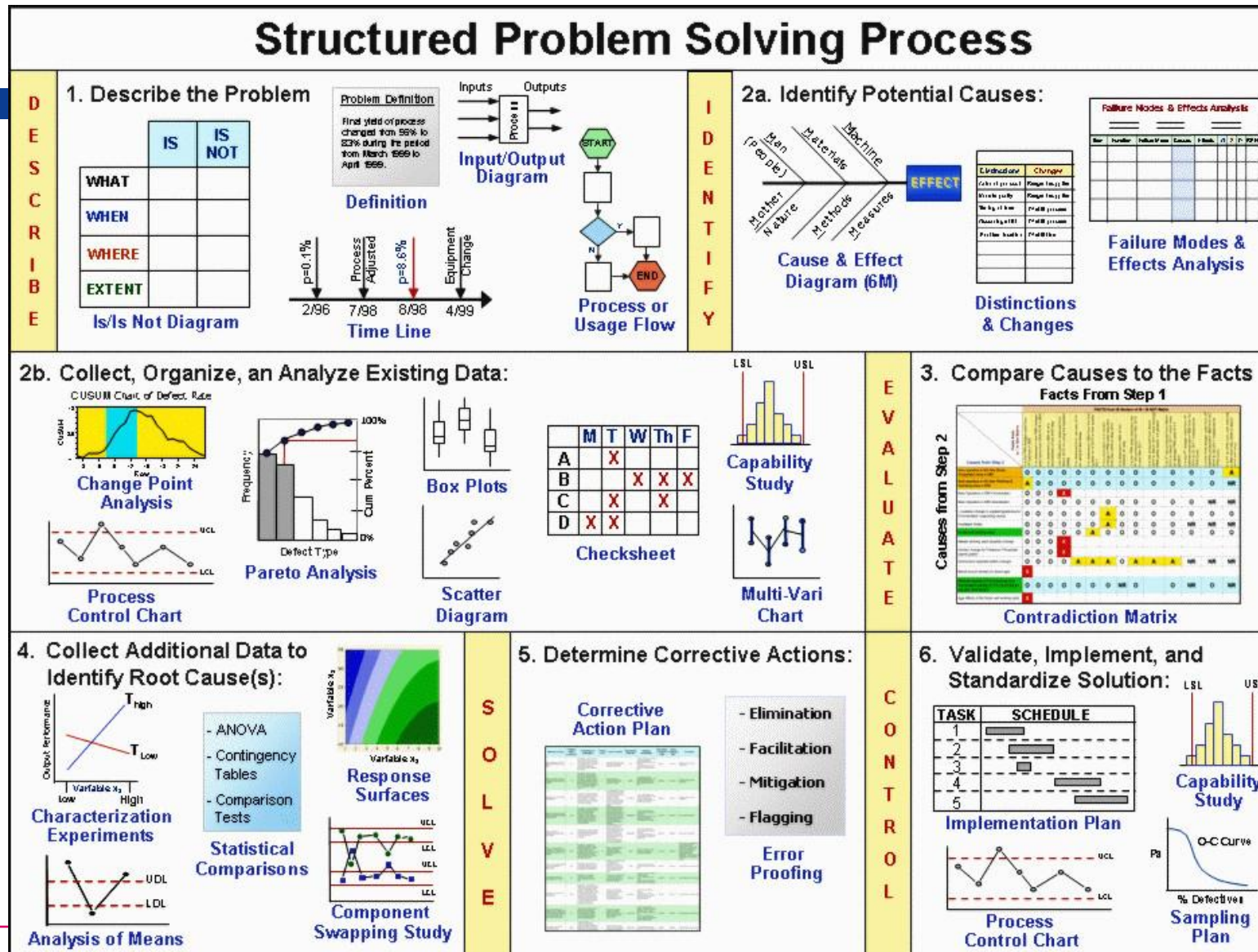
I had thought of her as one of the brightest people I knew and her subsequent career supports that view. What she said bothered me. I kept thinking about it; sometime the next day, it hit me. Science makes me feel stupid too. It's just that I've gotten used to it. So used to it, in fact, that I actively seek out new opportunities to feel stupid. I wouldn't know what to do without that feeling. I even think it's supposed to be this way. Let me explain.

For almost all of us, one of the reasons that we liked science in high school and college is that we were good at it. That can't be the only reason – fascination with understanding the physical world and an emotional need to discover new things has to enter into it too. But high-school and college science means taking courses, and doing well in courses means getting the right answers on tests. If

I'd like to suggest that our Ph.D. programs often do students a disservice in two ways. First, I don't think students are made to understand how hard it is to do research. And how very, very hard it is to do important research. It's a lot harder than taking even very demanding courses. What makes it difficult is that research is immersion in the unknown. We just don't know what we're doing. We can't be sure whether we're asking the right question or doing the right experiment until we get the answer or the result. Admittedly, science is made harder by competition for grants and space in top journals. But apart from all of that, doing significant research is intrinsically hard and changing departmental, institutional or national policies will not succeed in lessening its intrinsic difficulty.

Second, we don't do a good enough job of teaching our students how to be productively stupid – that is, if we don't feel stupid it means we're not really trying. I'm not talking about 'relative stupidity', in which the other students in the class actually read the material, think about it and ace the exam, whereas you don't. I'm also not talking about bright people who might be working in areas that don't match their talents. Science involves confronting our 'absolute stupidity'. That kind of stupidity is an existential

Academic Journeys require structured problem-solving



SOLVE

4. Collect Additional Data to Identify Root Cause(s):

Characterization Experiments

Output Performance

Variable x₁

T_{High} T_{Low}

Statistical Comparisons

- ANOVA
- Contingency Tables
- Comparison Tests

Response Surfaces

Variable x₂

Variable x₁

Component Swapping Study

UCL LCL

CONTROL

5. Determine Corrective Actions:

Corrective Action Plan

Task	Schedule
1	
2	
3	
4	
5	

- Elimination
- Facilitation
- Mitigation
- Flagging

Error Proofing

CONTROL

6. Validate, Implement, and Standardize Solution:

Implementation Plan

TASK SCHEDULE

Task	Schedule
1	
2	
3	
4	
5	

Capability Study

LSL USL

Process Control Chart

UCL LCL

Sampling Plan

O-C Curve

% Defectives

Academic Journeys bring new insights



Computer Science > Computation and Language

[Submitted on 15 Oct 2025]

LLMs Can Get "Brain Rot"!

Shuo Xing, Junyuan Hong, Yifan Wang, Runjin Chen, Zhenyu Zhang, Ananth Grama, Zhengzhong Tu, Zhangyang Wang

We propose and test the LLM Brain Rot Hypothesis: continual exposure to junk web text induces lasting cognitive decline in large language models (LLMs). To causally isolate data quality, we run controlled experiments on real Twitter/X corpora, constructing junk and reversely controlled datasets via two orthogonal operationalizations: M1 (engagement degree) and M2 (semantic quality), with matched token scale and training operations across conditions. Contrary to the control group, continual pre-training of 4 LLMs on the junk dataset causes non-trivial declines (Hedges' $g > 0.3$) on reasoning, long-context understanding, safety, and inflating "dark traits" (e.g., psychopathy, narcissism). The gradual mixtures of junk and control datasets also yield dose-response cognition decay: for example, under M1, ARC-Challenge with Chain Of Thoughts drops $74.9 \rightarrow 57.2$ and RULER-CWE $84.4 \rightarrow 52.3$ as junk ratio rises from 0% to 100%.

Error forensics reveal several key insights. First, we identify thought-skipping as the primary lesion: models increasingly truncate or skip reasoning chains, explaining most of the error growth. Second, partial but incomplete healing is observed: scaling instruction tuning and clean data pre-training improve the declined cognition yet cannot restore baseline capability, suggesting persistent representational drift rather than format mismatch. Finally, we discover that the popularity, a non-semantic metric, of a tweet is a better indicator of the Brain Rot effect than the length in M1. Together, the results provide significant, multi-perspective evidence that data quality is a causal driver of LLM capability decay, reframing curation for continual pretraining as a \textit{training-time safety} problem and motivating routine "cognitive health checks" for deployed LLMs.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2510.13928](https://arxiv.org/abs/2510.13928) [cs.CL]

(or [arXiv:2510.13928v1](https://arxiv.org/abs/2510.13928v1) [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.2510.13928> 

A Complete Education for A Complete Person

P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: info@ucu.ac.ug Web: <https://ucu.ac.ug>
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda

Academic Journeys breed independent Data Scientists



ARTICLE

Contextualizing AI Ethics in Uganda Through Adaptive Sensitive Reweighting (ASR) for Equitable Microcredit

Emmanuel Isabirye, Daphne Nyachaki Bitalo

<https://doi.org/10.1093/9780198945215.003.0179>

Published: 15 October 2025

 PDF  Split View  Annotate  Cite
 Permissions  Share ▼

Abstract

This research tackles the pressing ethical concerns of using AI in Uganda's microcredit sector, namely to develop an adaptive sensitive reweighting (ASR) model to mitigate algorithmic bias and promote equitable access to credit. Traditional credit scoring models— and fairness-aware machine learning algorithms trained on Western-biased data—discriminate against marginalized groups because they are based on formal financial records, reinforcing structural disadvantages. By iterative engagement with Ugandan policymakers, lenders, borrowers, and AI experts, the most significant ethical concerns and context-specific fairness metrics were identified. The ASR approach



Oxford Intersections: AI in Society

(In Progress)

Philipp Hacker (editor in chief)

Search in this collection 

Contents

- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Acknowledgments
- References

Lecture Objectives and Learning outcomes

The Objectives of this lecture are to learn:

- ☐ Data Science Fundamentals & Ethics
- ☐ Problem-framing
- ☐ CRISP-DM (Cross Industry Standard Process for Data Mining)

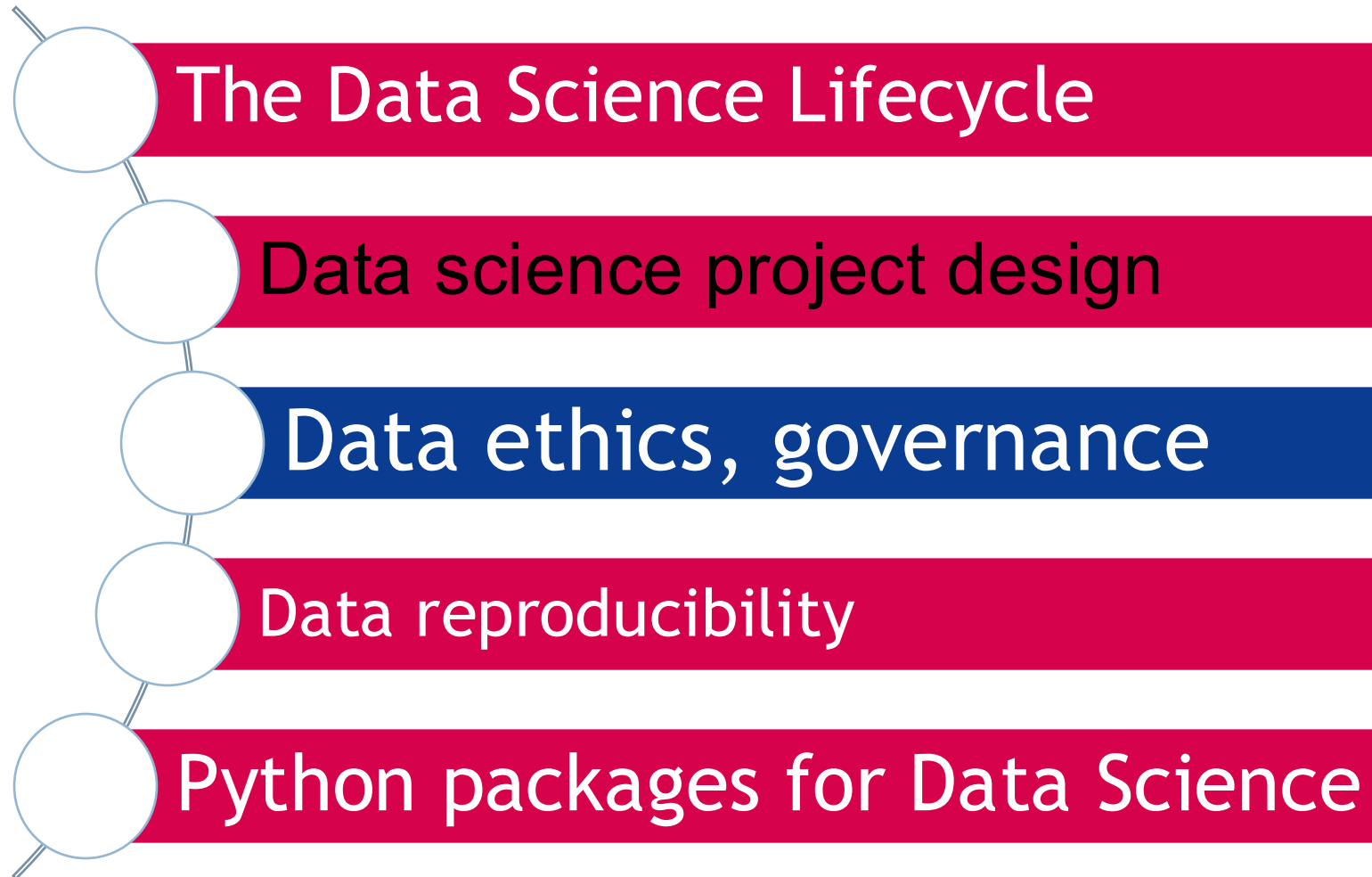
By the end of this lecture week, students should be able to:

- ☐ Design and scope a data science project from a vague business question to a deployable model.

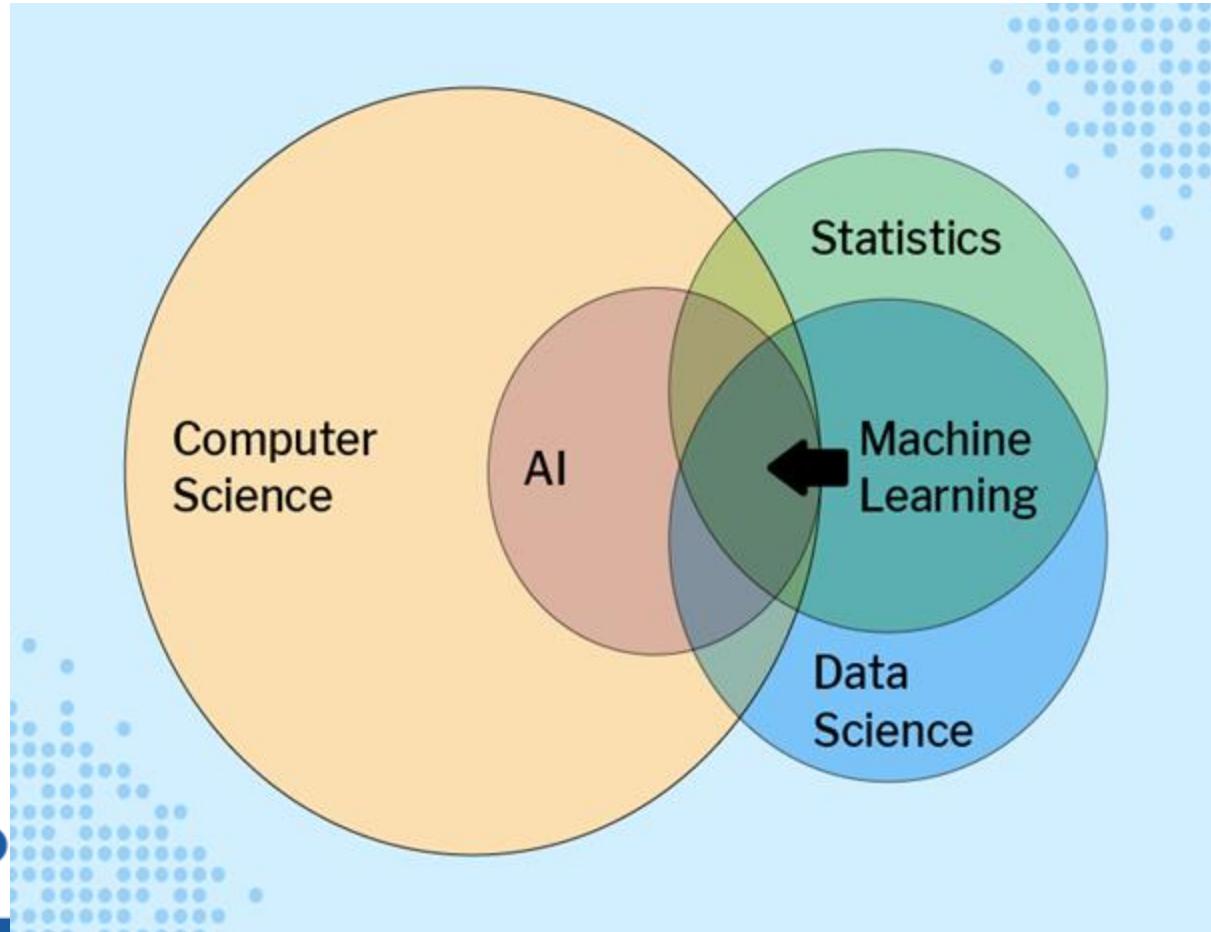




Lecture Overview



What is Data Science?



Interdisciplinary =
Multidisciplinary
Stats: Probability and
regression

Data Science: Data
preparation and exploration

ML: Trains data to generate
AI algorithm

CRISP-DM Method

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used methodology that provides a framework for conducting data mining/science projects. It outlines a series of phases that guide the process from business understanding to deployment.

Documentation: Maintain clear documentation throughout the project.





CRISP-DM Method

Phases of CRISP-DM:

1. Business Understanding:

- ☐ Define the business objectives and goals.
- ☐ Identify the relevant data sources.
- ☐ Create a project plan.

2. Data Understanding:

- ☐ Collect and gather the necessary data.
- ☐ Explore the data to understand its characteristics, quality, and completeness.
- ☐ Identify potential data quality issues.



CRISP-DM Method

3. Data Preparation:

- ☐ Clean and preprocess the data to address any quality issues.
- ☐ Transform the data into a suitable format for analysis.
- ☐ Create features or attributes that are relevant to the problem.

4. Modeling:

- ☐ Select appropriate data mining techniques based on the business objectives.
- ☐ Build and train models using the prepared data.
- ☐ Evaluate the performance of the models.



CRISP-DM Method

5. Evaluation:

- ☐ Assess the quality and reliability of the models.
- ☐ Compare the performance of different models.
- ☐ Validate the models using unseen data.

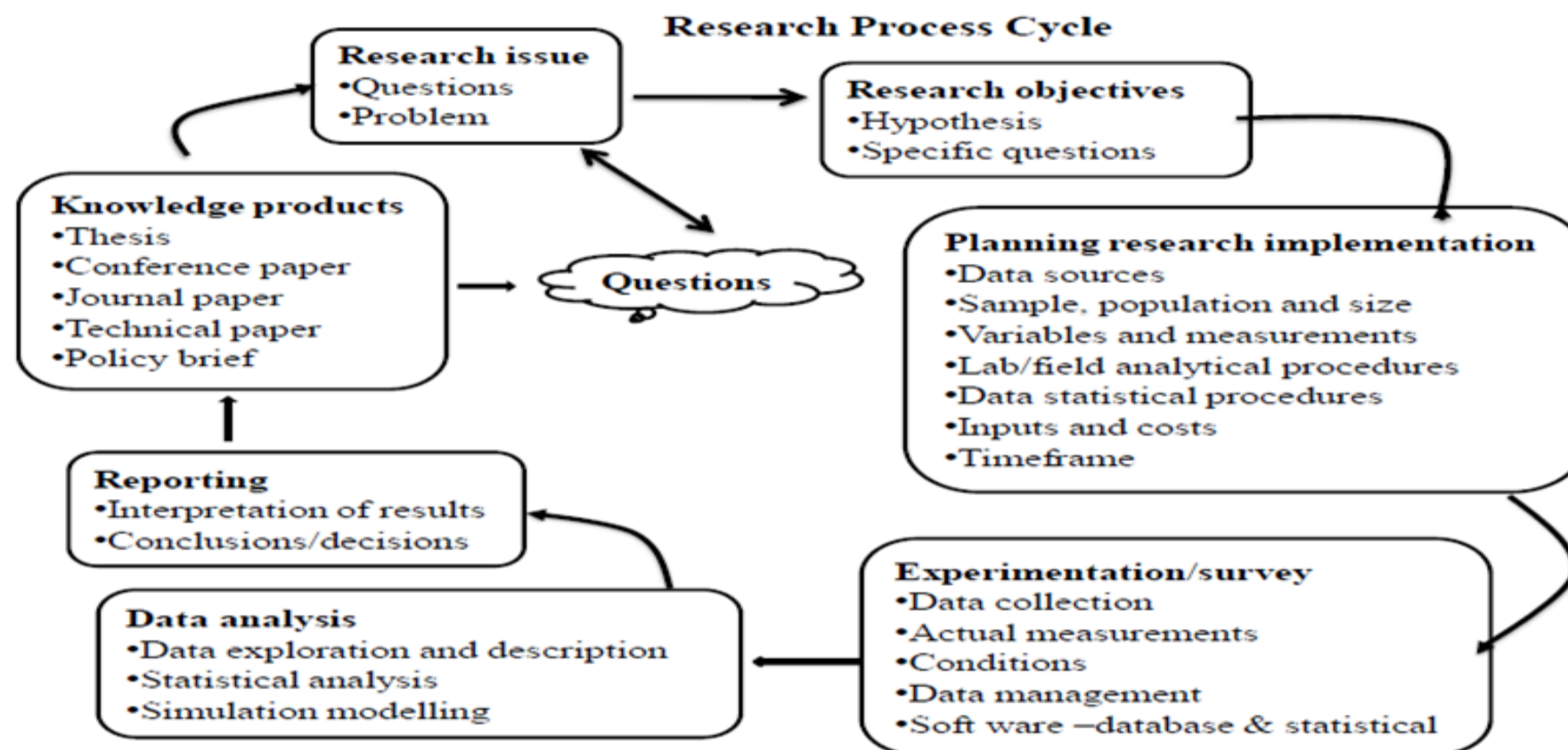
6. Deployment:

- ☐ Integrate the chosen model into the production environment.
- ☐ Monitor the model's performance and update it as needed.

Although CRISP-DM is the most widely adopted data science methodology, its original form has been adapted and challenged by several variants and alternatives to better suit modern data, tools, and practices

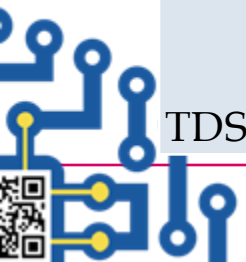
CRISP-DM = Designing a research cycle

Research Cycle



Core variants of CRISP-DM

Methodology	Acronym Meaning	Core Focus	Key Difference from CRISP-DM
KDD	Knowledge Discovery in Databases	The process of turning raw data into useful knowledge.	More research-oriented; the Data Mining step is only one phase within the larger KDD process. It lacks the initial Business Understanding phase of CRISP-DM.
SEMMA	Sample, Explore, Modify, Model, Assess	A framework heavily focused on the technical steps of model building and assessment.	Tool-specific (developed by SAS); it omits the critical Business Understanding and Deployment phases of CRISP-DM, focusing almost entirely on the data manipulation and modeling.
TDSP	Team Data Science Process	A methodology for collaborative, team-based data science projects, integrating cloud services and Agile principles.	Cloud-native (developed by Microsoft); it provides richer guidance on project management, artifacts, and MLOps/Deployment, aligning the lifecycle with an Agile approach.



More recent variants of CRISP-DM

1. Integration with Agile:

- ❑ CRISP-DM Agile/Scrum: allows for quicker feedback and value delivery, rather than completing all Business Understanding, then all Data Preparation, etc.
- ❑ Data Driven Scrum: framework specifically designed to address the unique challenges of data science teams (e.g., data exploration having uncertain outcomes)
- ❑ Kanban: visualizes workflow and managing the amount of work in progress, can identify shortcomings

More recent variants of CRISP-DM

2. Machine Learning Operations (MLOPs) focus:

- ❑ Continuous Integration: Automating code and environment testing.
- ❑ Continuous Delivery: Automating model deployment to production
- ❑ Continuous Training: Automating model retraining on new data
- ❑ Monitoring: Tracking model performance (e.g., data drift, model drift) and system health in production.



More recent variants of CRISP-DM

3. Domain-specific adaptations:

- ❑ CRISP-MED-DM : tailored for data mining in the medical domain, addressing specific challenges like privacy, ethics etc.
- ❑ CRISP-DM for Agriculture/Finance: Methodologies that add steps for handling non-traditional data (like satellite imagery or high-frequency trade data) and ensuring regulatory compliance specific to those industries

“Contextualizing AI Ethics in Uganda Through Adaptive Sensitive Reweighting (ASR) for Equitable Microcredit,” Emmanuel Isabirye & Daphne Nyachaki Bitalo, Journal: Oxford Intersections, 2025.





Data Science Problem Framing

1. Defining a hypothesis:

- ☐ A statement of expectation or prediction that will be tested by research.
- ☐ A hypothesis can be used to predict the relationship between variables for instance.

Two types of hypotheses:

- ☐ Null hypothesis: Predicts that the results will show no or little effect. The null hypothesis is a predictive statement that researchers use when it is thought that the Independent variable (IV) will not influence the Dependent Variable (DV).
- ☐ Alternative hypothesis: Predicts the reverse. Expecting that IV will significantly influence the DV





Key Terminology

Samples- Individual plants

Observations- Collected data

Populations- Defined groupings that are being compared

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Sesn	locn	block	rep	illage	ferTco	Plants_harvested	No_bigtubers	Weigh_bigtubers	No_mediumtubers	Weight_mediumtubers	No_smalltubers	Weight_smalltubers	Tc
2	2	1	1	1	onv	F2150	28	0	0	61	2.5	319	4.7	
3	2	1	1	1	onv	F1100	28	0	0	110	4.6	260	4	
4	2	1	1	1	onv	F3200	28	2	0.2	115	5.2	319	4.4	
5	2	1	1	1	onv	F5300	28	6	0.7	60	2.7	303	4.8	
6	2	1	1	1	onv	F4250	28	3	0.3	82	3.4	332	4.7	
7	2	1	2	2	onv	F5300	28	6	0.5	65	2.7	299	4.5	
8	2	1	2	2	onv	F3200	28	0	0	91	3.9	289	4.8	
9	2	1	2	2	onv	F4250	28	0	0	72	3.5	246	4.6	
10	2	1	2	2	onv	F1100	28	1	0.2	64	2.6	305	5.4	
11	2	1	2	2	onv	F2150	28	1	0.1	56	2.5	308	4.5	
12	2	1	3	3	onv	F4250	28	4	0.5	63	3	376	5	
13	2	1	3	3	onv	F5300	28	3	0.3	97	4.5	290	4.8	
14	2	1	3	3	onv	F2150	28	3	0.3	59	2.7	286	4.7	
15	2	1	3	3	onv	F3200	28	0	0	81	3.3	138	3.9	
16	2	1	3	3	onv	F1100	28	0	0	68	2.9	276	5.2	
17	2	1	1	1	minimum	F2150	28	1	0.2	59	2.5	316	4.2	
18	2	1	1	1	minimum	F1100	28	0	0	77	3	266	3.5	
19	2	1	1	1	minimum	F3200	28	2	0.3	54	2.4	243	4	
20	2	1	1	1	minimum	F5300	28	0	0	60	3	278	4.7	
21	2	1	1	1	minimum	F4250	28	2	0.3	50	2.6	296	5.1	
22	2	1	2	2	minimum	F5300	28	2	0.2	37	1.7	285	4.1	
23	2	1	2	2	minimum	F3200	28	0	0	49	2.1	272	4.9	

Key Terminology

Quantitative vs Qualitative

- quantitative data is measurable
- qualitative: data is described

Non-experimental vs Experimental

- non-experimental: contribute to background conditions of experiment
- experimental: purposely chosen to be studied in defined conditions

Data can be primary or secondary

	Qualitative	Quantitative
Conceptual	<p>Concerned with understanding human behaviour from the informant's perspective</p> <p>Assumes a dynamic and negotiated reality</p>	<p>Concerned with discovering facts about social phenomena</p> <p>Assumes a fixed and measurable reality</p>
Methodological	<p>Data are collected through participant observation and interviews</p> <p>Data are analysed by themes from descriptions by informants</p> <p>Data are reported in the language of the informant</p>	<p>Data are collected through measuring things</p> <p>Data are analysed through numerical comparisons and statistical inferences</p> <p>Data are reported through statistical analyses</p>
Source: Adapted from Minichiello <i>et al.</i> (1990, p. 5)		



Key Terminology: Data Model

Data = Pattern+ Residual

- Response (observed data) is influenced by factors on the right hand side of the model
- Pattern is the part of data that can be explained or can be attributed to some known sources, e.g. experimental treatments, experimental design
- Residual is that part of data whose source cannot be explained (assumed random)

Variation in data = explained variation + unexplained variation

A good data process cycle will maximise explained variation and minimize unexplained variation



Hypotheses

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

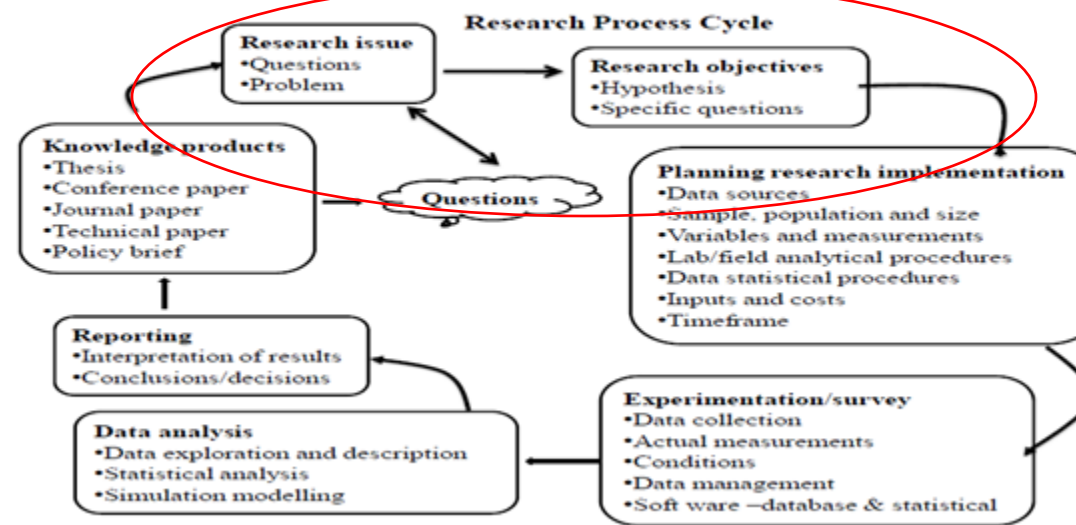
Null hypothesis: Personal appearance does not affect success

OR There is no significant difference between appearance and success

Alternative hypothesis: Personal appearance affects success

OR There's a significant difference between appearance and success

Research Cycle



Formulating a hypothesis



Hypothesis Requirement	Description
Write as predictive statements regarding the relationship between the IV and DV.	The researcher should be able to predict what they expect to find from the study results. The researcher could state that they expect to see a difference.
It should be formulated based on background research	Hypotheses should not be based on guesswork. Instead, researchers should use previously published research to predict the study's expected outcome Anecdotal (citizen Science)
Identify the IV.	IV is what the experimenter manipulates to see if it affects the DV.
Identify the DV.	DV is the variable being measured after the IV has been manipulated or after it changes during the experiment.

Formulating a hypothesis

Hypothesis Requirement	Description
The variables should be operationalised.	The researchers must define how each variable (IV and DV) will be measured. When a hypothesis is operationalised, it is testable.
The hypotheses needs to be falsifiable.	Other researchers need to be able to replicate the research using the same variables to see whether they can verify the results.
The hypotheses should be clear.	Hypotheses are usually only a sentence long and should only include the details summarised above.

Structuring an experimental/analysis design



2. Ask important questions

- a) Why is this problem important? (Hypothesis formulation, objectives)
- b) Who does this problem affect? (samples, populations)
- c) What if we don't have the right data? (data sources, defined variables and how to measure them, experimental survey)
- d) When is the project over? (Aligning project expectations to achieved results)
- e) What if we don't like the results? (Risks and mitigation)



Defining Variables

Continuous variables- take an infinite set of values

Discrete- Finite set of values (Can be categorical)

Dependent variable (DV)- being measured in the experiment

Independent variable (IV)- Is not measured (manipulated variable)

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Sesn	locn	block	rep	tillage	ferTco	Plants harvested	No_bigtubers	Weigh_bigtubers	No_mediumtubers	Weight_mediumtubers	No_smalltubers	Weight_smalltubers	Tc
2	2	1	1	1	1 conv	F2150	28	0	0	61	2.5	319	4.7	
3	2	1	1	1	1 conv	F1100	28	0	0	110	4.6	260	4	
4	2	1	1	1	1 conv	F3200	28	2	0.2	115	5.2	319	4.4	
5	2	1	1	1	1 conv	F5300	28	6	0.7	60	2.7	303	4.8	
6	2	1	1	1	1 conv	F4250	28	3	0.3	82	3.4	332	4.7	
7	2	1	2	2	2 conv	F5300	28	6	0.5	65	2.7	299	4.5	
8	2	1	2	2	2 conv	F3200	28	0	0	91	3.9	289	4.8	
9	2	1	2	2	2 conv	F4250	28	0	0	72	3.5	246	4.6	
10	2	1	2	2	2 conv	F1100	28	1	0.2	64	2.6	305	5.4	
11	2	1	2	2	2 conv	F2150	28	1	0.1	56	2.5	308	4.5	
12	2	1	3	3	3 conv	F4250	28	4	0.5	63	3	376	5	
13	2	1	3	3	3 conv	F5300	28	3	0.3	97	4.5	290	4.8	
14	2	1	3	3	3 conv	F2150	28	3	0.3	59	2.7	286	4.7	
15	2	1	3	3	3 conv	F3200	28	0	0	81	3.3	138	3.9	
16	2	1	3	3	3 conv	F1100	28	0	0	68	2.9	276	5.2	
17	2	1	1	1	1 minimum	F2150	28	1	0.2	59	2.5	316	4.2	
18	2	1	1	1	1 minimum	F1100	28	0	0	77	3	266	3.5	
19	2	1	1	1	1 minimum	F3200	28	2	0.3	54	2.4	243	4	
20	2	1	1	1	1 minimum	F5300	28	0	0	60	3	278	4.7	
21	2	1	1	1	1 minimum	F4250	28	2	0.3	50	2.6	296	5.1	
22	2	1	2	2	2 minimum	F5300	28	2	0.2	37	1.7	285	4.1	
23	2	1	2	2	2 minimum	F3200	28	0	0	49	2.1	272	4.9	

In-class Assignment

You will be assigned to breakout rooms for this exercise:

- Create a hypothetical problem pertinent to your research e.g.

“The department of computer science at UCU is investigating whether student performance is improved by continuous assessment.”

1. Formulate two research hypotheses that can be used to address your research problem.
 - a) A null hypothesis
 - b) An alternative hypothesis
2. Expand your research cycle by generating measurable objectives (be sure to define your variables)



Data Ethics and Governance

Data privacy and Compliance:

The practice of safeguarding sensitive information from unauthorized access and ensuring compliance with legal standards.

Regulation/Concept	Jurisdiction & Focus	Core Impact on Data Science	Techniques
GDPR (General Data Protection Regulation)	European Union (EU)	Requires explicit consent, mandates the "Right to be Forgotten" (erasure), and requires Data Protection Impact Assessments (DPIAs) for high-risk processing.	Data Minimization: Only collecting data essential to the project. Pseudonymization: Replacing direct identifiers with artificial ones (reversible).
Data Protection and Privacy Act	NITA (Uganda)	A data subject's prior consent is generally required for data collection and processing	De-identification: Removing all 18 identifiers (e.g., names, dates, geo-codes) to make the data legally usable for research.
Differential Privacy	Algorithmic concept	A mathematically rigorous way to provide aggregate data insights while guaranteeing that an individual's presence or absence in the dataset does not significantly change the outcome of a query.	Adding Noise: Injecting controlled noise into the data or query results to obscure individual records, protecting privacy during analytical operations.

A Complete Education for A Complete Person

Data Ethics: Reproducibility



- Reproducibility ensures duplication of work generating the exact same results. It is essential for quality control, collaboration and auditing.
- Code Versioning: Use public repositories for code/data such as Github, Kaggle, NCBI, etc
- Environment management: State all software/library versions used or use Conda or Docker for consistency
- Experiment tracking: Use platforms like MLflow





Python packages for Data Science

Category	Package Name	Primary Function in Data Science	Key Use Cases
Core Computing & Arrays	NumPy	Fundamental package for scientific computing; provides powerful N-dimensional array objects.	Linear algebra, Fourier transforms, working with matrices and vectors, underlying many other libraries.
Data Manipulation & Analysis	Pandas	Provides fast, flexible, and expressive DataFrames for working with structured data.	Data cleaning (handling missing values, outliers), data transformation, merging, filtering, and time-series analysis.
Machine Learning (General)	Scikit-learn	A comprehensive library for classic ML algorithms and model utilities.	Classification, regression, clustering, dimensionality reduction (PCA), cross-validation, and performance metrics.
Deep Learning	TensorFlow / PyTorch	Frameworks optimized for building and training complex neural networks on GPUs.	Computer Vision (CNNs), Natural Language Processing (NLP), sequential modeling (RNNs/LSTMs), and Generative AI.



Python packages for Data Science

Category	Package Name	Primary Function in Data Science	Key Use Cases
Statistical Modeling	StatsModels	Focused on statistical models, testing, and exploration.	Regression models (OLS, GLMs), time-series analysis, and statistical hypothesis testing.
Visualization (Static)	Matplotlib	The foundational library for creating static, publication-quality 2D plots.	Basic line charts, scatter plots, histograms, and customizing plot elements.
Visualization (Statistical)	Seaborn	Provides a high-level interface for drawing attractive and informative statistical graphics.	Visualizing distributions (violin plots), relationships between variables (pair plots), and heatmap visualization.
Visualization (Interactive)	Plotly / Dash	Tools for creating interactive plots, dashboards, and web applications without JavaScript.	Interactive reports, geospatial visualizations, and exploratory data analysis applications.
Big Data & Distributed Computing	PySpark / Dask	Python API for Apache Spark (PySpark) and Dask for parallel and distributed computing.	Processing massive datasets that exceed the memory capacity of a single machine.
ML Operations (MLOps)	MLflow	Manages the end-to-end machine learning lifecycle.	Experiment tracking, model packaging, and model registry for deployment.
Explainable AI (XAI)	SHAP / LIME	Libraries for interpreting the predictions of black-box machine learning models.	Providing feature importance and local explanations for individual predictions.

Python Libraries

LIBRARIES

Numpy

np. Multidimensional data arrays

Pandas

pd. Generates data frames. Used in Data Science

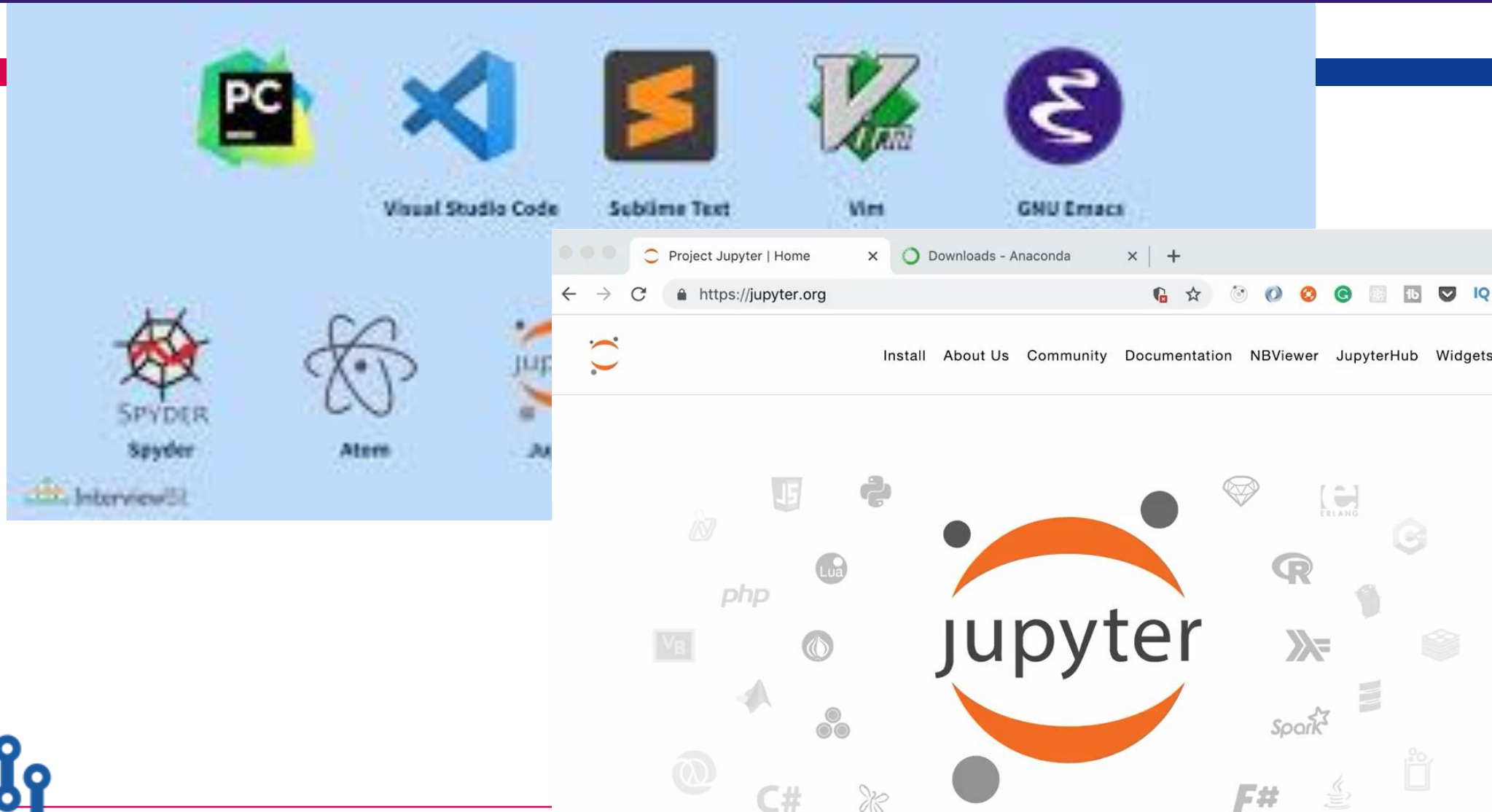
Matplotlib

mat. Creates plots and graphs in 2D

Scikit-Learn

sklearn. Has ML algorithms for SL and USL

Python script editors



A Complete Education for A Complete Person

P.O. Box 4, Mukono, Uganda, Plot 67-173, Bishop Tucker Road, Mukono Hill | Tel: +256 (0) 312 350 800 Email: info@ucu.ac.ug Web: <https://ucu.ac.ug>
Founded by the Province of the Church of Uganda. Chartered by the Government of Uganda



Group Exercise

1. Install python
2. Install Visual Studio Code and Jupyter Notebook
3. Install common python packages/libraries (pandas, matplotlib etc)
4. Upload the libraries of the packages
5. Start practicing importing a dataset in to VS Code.

