Introduction to Data Science

dr Maciej Świtała
Ewa Weychert

Class 1: Data science and its economic context

Who am I? ;)

- mgr Ewa Weychert
- Research interests: demography, machine learning, NLP
- Collaboration with LabFam (Interdisciplinary Centre for Labour Market and Family Dynamics)
- Working at University of Florence
- `e.weychert@uw.edu.pl`

Who is Maciek? ;)

- Maciej Świtała
- Doctor of Social Sciences in the field of Economics and Finance
- Master of Laws
- Research interests: natural language processing, machine learning, empirical legal studies
- NLPath
- `ms.switala@uw.edu.pl`

Plan

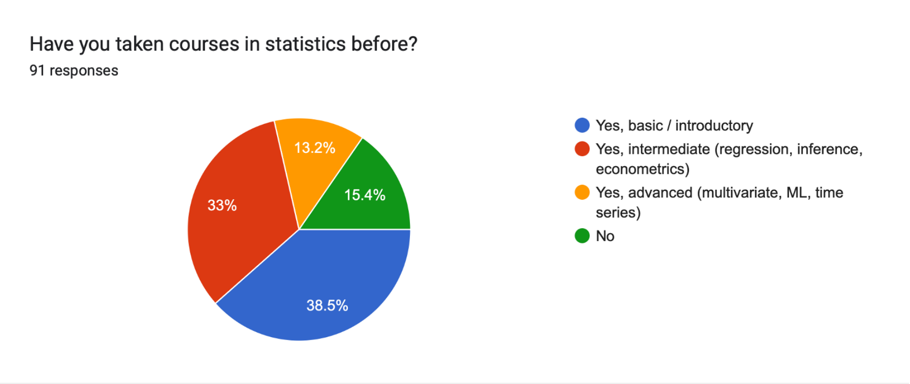
Date	Ewa	Maciek
Monday, October 06, 2025	Data science and its economic context	
Monday, October 13, 2025	Computer programming for data science	
Monday, October 20, 2025		Basics of statistics and econometrics
Monday, October 27, 2025	Introduction to machine learning	
Monday, November 03, 2025		Introduction to machine learning
Thursday, November 13, 2025		Natural language processing
Monday, November 17, 2025		Natural language processing
Monday, November 24, 2025	AI and prompt engineering	

Final Grade

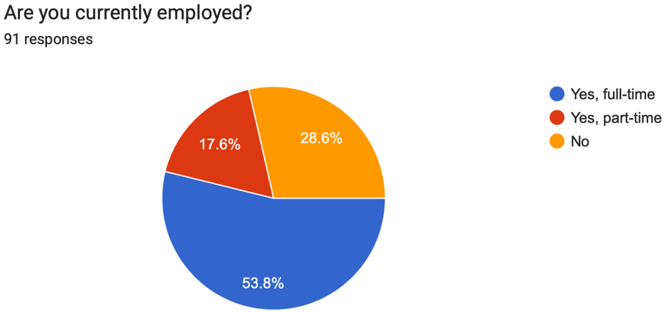
How to pass this class?

- Format: Multiple-choice questions (MCQs) (A multiple-choice question is a type of objective assessment in which a question has zero or more possible answers)
- Number of questions: 25
- Time limit: 90 minutes
- Date of exam – during the examination session: 26 January 2026 – 8 February 2026 (to be determined later by the administration office)

Who are you? 1/8



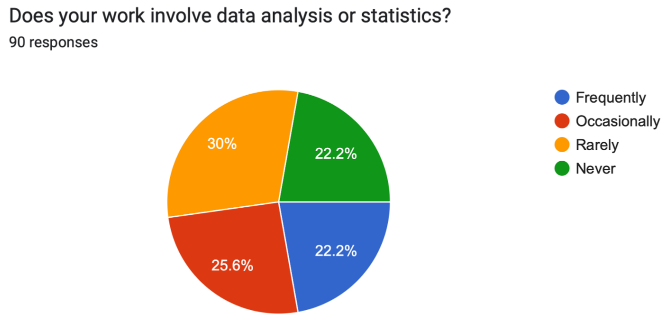
Main insight: Only 15.4% have had no prior exposure.
Interpretation: The group has a solid foundation in statistics, mostly at the introductory to intermediate level. The course can therefore move beyond the basics relatively quickly, but should still provide brief refreshers for those with limited background.



Main insight: 28.6% are not currently employed. **Interpretation:**

The majority of students balance full-time work with studies, indicating that flexibility and asynchronous learning options may be important for maintaining engagement and accessibility.

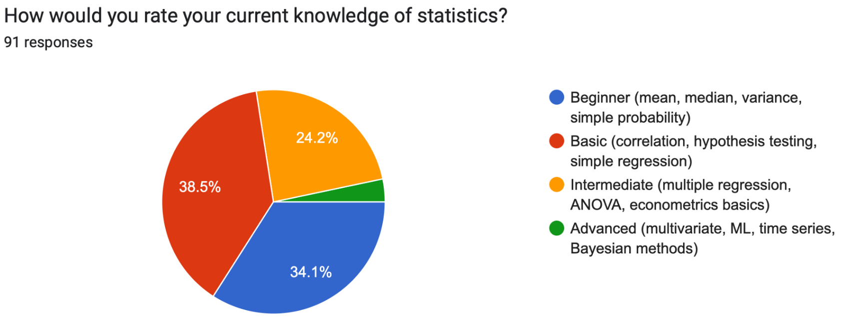
Who are you? 3/8



Main insight: Responses are evenly distributed - 30% rarely and 22.2% never do.

Interpretation: About half of participants have at least some practical engagement with data analysis, while the other half have limited or no direct experience. This mix suggests the need to balance

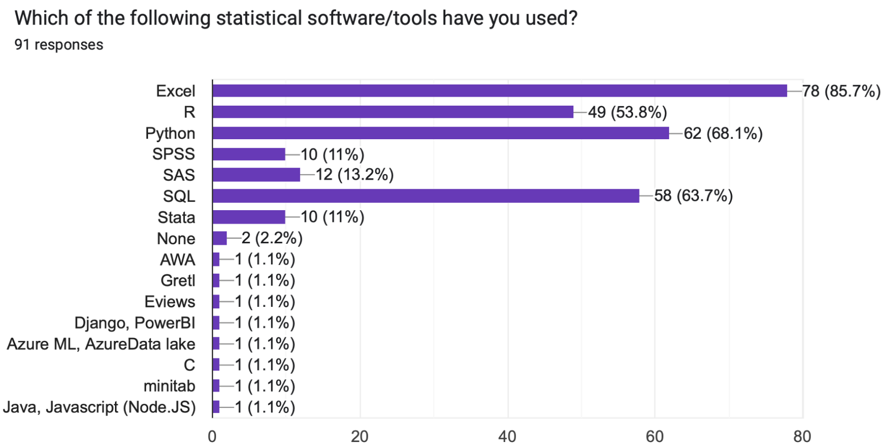
Who are you? 4/8



Main insight: The majority of respondents self-assess as either *basic* (38.5%) or *beginner* (34.1%) in statistical knowledge.

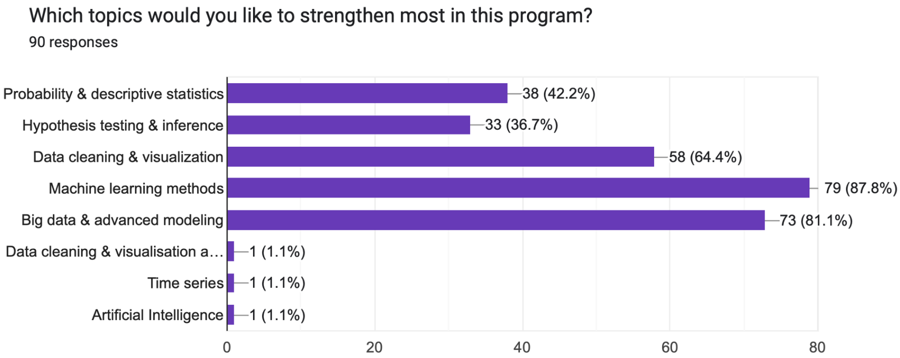
Interpretation: The cohort's confidence level aligns with early-career learners who have had limited exposure beyond fundamental methods. Course content should focus on consolidating foundational concepts

Who are you? 5/8



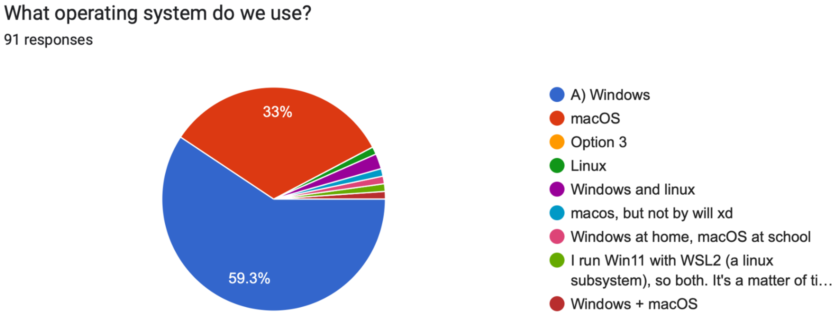
Main insight: The most widely used tools are *Excel* (85.7%), *Python* (68.1%). Only 2.2% reported having used no statistical software

Who are you? 6/8



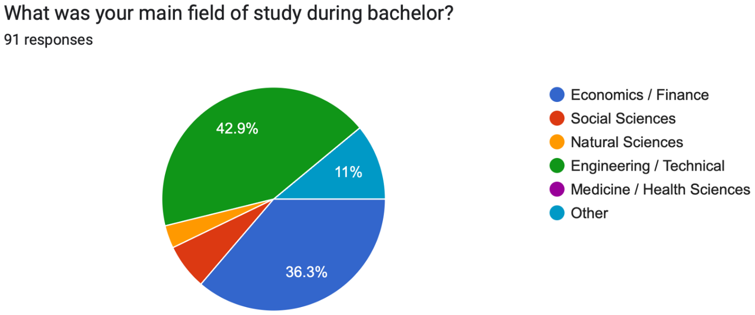
Main insight: The strongest interest areas are *machine learning methods* (87.8%) and *big data & advanced modeling* (81.1%)
Interpretation: Participants are eager to deepen their applied and computational skills, particularly in modern, data-driven methods. The

Who are you? 7/8



Main insight: A majority of respondents (59.3%) use *Windows*, while about one-third (33%) use *macOS*.

Interpretation: Given that most participants use Windows or macOS, course materials and software setup guides should focus on these systems, with optional notes for Linux users to ensure full



Main insight: The cohort is dominated by *Engineering/Technical* backgrounds (42.9%), with a large share from *Economics/Finance* (36.3%). About 11% report *Other* fields, while only small shares come from *Social* and *Natural Sciences.*

Interpretation: Expect strong quantitative aptitude mixed with

Plan of today's class

1. What is data science?
2. What is econometrics?
3. What is machine learning?
4. What is the difference between econometrics and machine learning?
5. What is the road-map of the subjects and what will you learn during this Master Program - Data Science and Business Analytics?
6. Why is machine learning useful in economics? Examples of machine learning in economics
7. Types of data in data science: numerical, text, images
8. Challenges and ethical concerns in data science
9. Useful advice especially for beginners

Trending of Data Science-Relevant Terms

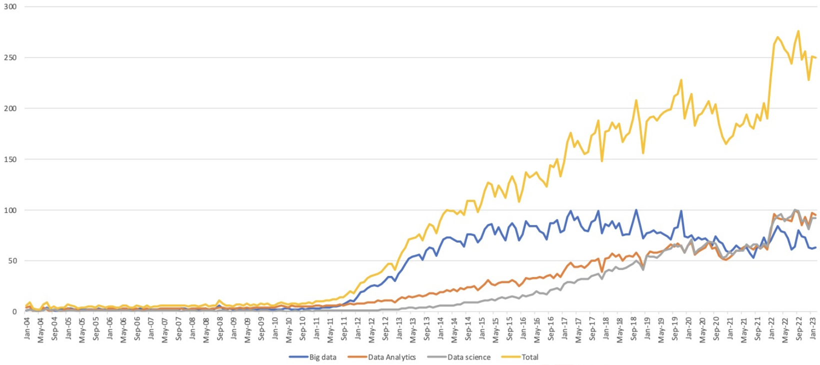


Figure 1: Trending of data science-relevant terms
Source: Google Trends on January 2023

What is Data ? - formal definition

- **Scientific Definition:** Data are **symbolic representations of empirical observations**, collected and recorded to measure or describe attributes of physical, social, or computational phenomena (Ackoff, 1989; Kitchin, 2014).

- **Data** = recorded facts, measurements, or observations about some phenomenon.
- Can be numbers, text, images, audio, signals—*representations* of reality.
- By themselves, data lack interpretation; with context they become **information**, then **knowledge**.
- **Data** → **Information** (add context, summarize) → **Knowledge** (explain/understand) → **Decisions** (act, evaluate impact).
- The same raw data can yield different insights depending on *questions*, *methods*, and *domain knowledge.*
- Good outcomes depend on both data properties and analytical practice.

What is Data ? - by structure

- **Structured:** tables with rows/columns (SQL tables, spreadsheets).
- **Semi-structured:** self-described but irregular (JSON, XML, logs).
- **Unstructured:** free text, images, audio, video.

Implication: structure affects storage, tooling, and analysis methods.

What is Data ? - by measurement scale (Stevens, 1946)

- **Nominal** — categories, no order (e.g., country, blood type).
- **Ordinal** — ordered categories, gaps not meaningful (e.g., Likert 1–5).
- **Interval** — numeric, equal steps, no true zero (e.g., °C temperature).
- **Ratio** — numeric, equal steps, true zero (e.g., income, weight).

Implication: scale determines valid operations (means, ratios, correlations).

Types of Data in Machine Learning — Big Picture

- **Core principle:** ML models operate on **numbers**. Every data type is **transformed to numeric** representations.
- **Common modalities** in practice:
  - **Tabular/Structured** (spreadsheets, SQL)
  - **Text** (documents, social media, logs)
  - **Images** (photos, medical scans)
  - **Geospatial** (coordinates, rasters/vectors)
- **Also frequent:** time series, audio, video, graphs/networks, sensor/IoT, events/logs, multimodal.
- *Pipeline:* Raw ⇒ Clean ⇒ **Encode to numeric** ⇒ Model ⇒ Evaluate.

Tabular / Structured Data

- Rows = observations; columns = features (types: **numeric**, **categorical**, **ordinal**, **datetime**, **boolean**).
- **Numeric encoding:**
  - Scale/normalize continuous features; log-transform skewed variables.
  - Encode categories: one-hot, ordinal (with care), target/impact encoding (with CV to avoid leakage).
  - Datetime $\rightarrow$ cyclic (sin/cos) or calendar features; interactions.
- **Missing values:** indicator flags; impute (median/knn/model-based).
- **Pitfalls:** data leakage (future info, target encoders without CV), inconsistent units, mixed granularities.

Text (NLP)

- Sources: documents, reviews, emails, support tickets, social media.
- **Numeric encoding:**
  - Classic: bag-of-words, **TF–IDF**, character n-grams.
  - Token embeddings: word2vec/GloVe; **subword** (BPE) for rare words.
  - **Contextual** embeddings (Transformers): sentence/CLS vectors.
- Preprocess: language detection, tokenization, lowercasing/stemming/lemmatization, stopwords (task-dependent).
- Labels: classification (topics, sentiment), sequence labeling (NER), ranking, QA.
- **Pitfalls:** target leakage via rare terms, class imbalance, domain shift (train vs deploy language/genre).

Text to numbers

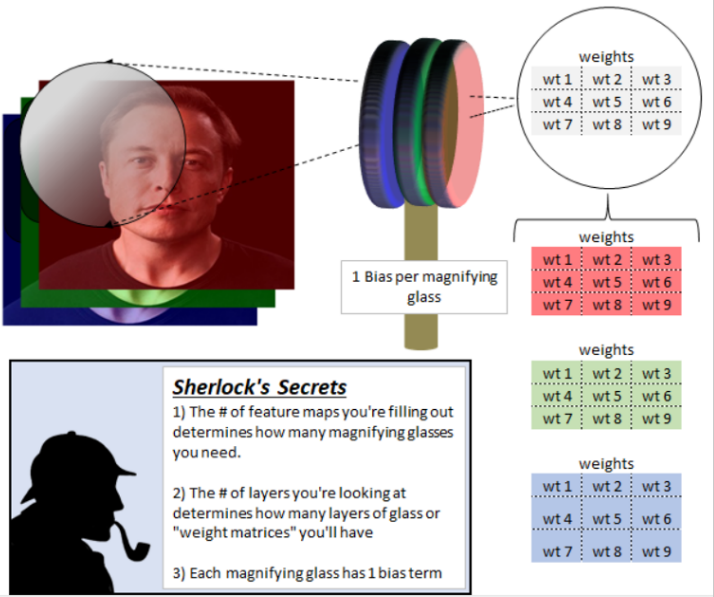
	food	is	tasty	not
$d_1 :=$	1	1	1	0

	food	is	tasty	not
$d_2 :=$	1	1	1	1

Images

- Representation: tensors ($H \times W \times C$); pixel intensities (0–255 or normalized).
- **Numeric encoding:**
  - Raw pixels for CNNs; handcrafted features (HOG/SIFT) for classical ML.
  - Pretrained backbones (CNN/ViT) $\rightarrow$ embeddings for downstream tasks.
- Preprocess: resize, center-crop, normalization per channel; **augmentation** (flip, rotate, color jitter, mixup/cutout).
- Tasks: classification, detection, segmentation, retrieval.
- **Pitfalls:** label noise, shortcut learning (spurious backgrounds), data imbalance, leakage via near-duplicates.

Images to numbers



Geospatial / Geographical Data

- Forms: **vector** (points/lines/polygons: roads, parcels) and **raster** (grids: satellite, elevation).
- **Numeric encoding:**
  - Coordinates (lon/lat) with proper **CRS** (e.g., WGS84); engineered features: distances, buffers, spatial joins.
  - Raster stacks $\rightarrow$ per-pixel feature vectors (e.g., spectral bands); tiling to fixed-size tensors for CNNs.
- **Spatio-temporal** aspects: add time, seasonality, lagged aggregates.

Other Common Data Types (All Become Numeric)

- **Time series:** sequences indexed by time; encode lags, rolling stats, Fourier terms; use sequence models.
- **Video:** frames + time; 3D CNNs or frame embeddings + temporal model.
- **Multimodal:** combine text+image+tabular

- **Categorical** → one-hot / ordinal / target encoding (with CV).
- **Text** → TF–IDF / static embeddings / contextual embeddings.
- **Images** → pixels / CNN features / ViT embeddings.
- **Geo** → CRS-aware coordinates, distances, raster bands, graph features.
- **Time** → lags, windows, seasonal/sinusoidal encodings.
- **Graphs** → adjacency, Laplacian features, node/edge embeddings.

Rule: choose encoders that **respect structure** (order, space, topology) and **avoid leakage**.

What is Data ? - by origin

- **Primary data:** collected directly for the study (surveys, experiments, sensors).
- **Secondary data:** obtained from others (administrative records, open datasets, APIs).

Implication: origin affects cost, control, documentation, and bias.

What is Data Science? 1/2

- **IBM (Industry Perspective):**
  "Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data."
  Source: IBM – What is Data Science?

- **Harvard Business Review (Academic & Professional View):**
  "Data scientists are professionals who combine the skills of software engineering, statistics, and storytelling to transform raw data into understanding."
  Source: Harvard Business Review, *"Data Scientist: The Sexiest Job of the 21st Century"* (2012)

What is Data Science? 2/2

- Data science is inherently **interdisciplinary**.
- It combines **mathematics, statistics, computing, and domain expertise**.
- Its goal: extract insights and support **evidence-based decision-making**.

Source: Özsu, M. T. (2023). *Foundations and Scoping of Data Science.* arXiv preprint arXiv:2301.13761 — esp. § Definition and framing of interdisciplinarity in data science.

What is Econometrics?

- **Econometrics** is the application of **statistical and mathematical methods** to analyze economic data.
- It aims to **test economic theories**, **estimate relationships**, and **forecast economic outcomes**.
- Econometrics combines **economic theory**, **data**, and **statistical inference** to quantify economic phenomena.
- It provides the empirical foundation for **evidence-based policy** and **decision-making in economics and finance**.

Source: Adapted from Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach.* 7th ed., Cengage Learning.

Econometrics vs Data Science: Core Goals

- **Econometrics:** Focuses on **causal inference**, **model-based estimation**, and testing economic theories.
- **Data Science:** Emphasizes **prediction**, **algorithmic performance**, and extracting patterns from complex data.
- Econometrics seeks **interpretation and validity**; data science seeks **accuracy and scalability**.
- Athey & Imbens (2019) describe this as a contrast between the **model-based culture** and the **algorithmic modeling culture** (after Breiman, 2001).

Source: Athey, S. & Imbens, G. (2019). *Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11:685–725.*

Econometrics vs Data Science: Methods and Evaluation

- **Econometrics:**
  - Relies on **statistical models** with explicit assumptions.
  - Prioritizes **consistency, unbiasedness, efficiency**, and valid **confidence intervals**.
- **Data Science / ML:**
  - Uses **algorithmic models** optimized for **out-of-sample prediction**.
  - Employs **cross-validation**, **regularization**, and **ensemble methods** (e.g., random forests, boosting).
  - Often trades formal inference for predictive performance.

Source: Athey & Imbens (2019)

Econometrics and Data Science: Integration and Outlook

- Both fields aim to **learn from data for decision-making**.
- **Econometrics** provides tools for **causal identification and inference**.
- **Data Science** provides tools for **high-dimensional prediction and scalability**.
- Modern research combines both:
  - **Causal machine learning** (e.g., double ML, causal forests).
  - **Hybrid methods** balancing interpretability and predictive power.
- Future econometrics must integrate **ML algorithms** without losing its focus on **causality and theory-driven modeling**.

Source: Athey & Imbens (2019); Mullainathan & Spiess (2017); Wager & Athey (2017).

Econometrics vs Data Science — A Comparative Overview

Dimension	Econometrics	Data Science / ML
Primary Goal	Causal inference, parameter estimation	Prediction, pattern recognition
Modeling Approach	Model-based (theory-driven)	Algorithmic / data-driven
Assumptions	Strong structural and statistical assumptions	Minimal assumptions, flexible models
Typical Methods	Regression, IV, GMM, panel models	Trees, random forests, neural networks, ensembles
Evaluation Criterion	Consistency, unbiasedness, valid inference	Out-of-sample predictive accuracy
Data Focus	Smaller, structured data sets	Large, high-dimensional, unstructured data
Validation	Theoretical fit, hypothesis testing	Cross-validation, regularization, tuning
Output	Interpretability and causal effects	Accuracy and scalability
Core Strength	Explanation and policy relevance	Prediction and automation
Emerging Integration	Causal ML, double machine learning, hybrid mo-	Incorporation of theory-based constraints

- **Goal:** program computers to *learn* from data (experience) to produce *expertise* usable for a task.
- Training data $\rightarrow$ learning algorithm $\rightarrow$ a model/program that performs the task.
- We seek formal answers to: **(i)** what is the data? **(ii)** how to automate learning? **(iii)** how to evaluate success?

Source: Shalev–Shwartz & Ben–David (2014), Ch. 1.

What is Learning?

- **Informal definition:** Learning is the process of transforming **experience (data or feedback)** into **expertise or knowledge** that improves performance on a task.
- **In nature:** Organisms adapt based on feedback from the environment. Example — *bait shyness in rats:* after associating a particular taste with nausea, the animal learns to avoid it in the future. $\Rightarrow$ Experience (stimulus + consequence) $\rightarrow$ Change in future behavior.
- **In machines:** A learning algorithm observes data and adjusts internal parameters (weights, rules, or trees) so that its predictions or decisions improve over time. *Goal:* perform well on **unseen data**, not just memorize the training examples.
- **Inductive inference:** Machines generalize from observed examples to unseen cases — the essence of learning is to infer general patterns rather than exact repetitions. Example — spam filtering: the system infers "spamness" patterns that apply to new emails it has never seen.
- **Key risk — Spurious associations:** Learners can mistake correlation

Spurious Patterns: The “Pigeon Superstition”

- **Origin:** B.F. Skinner (1948) observed pigeons developing “superstitious” behaviors. They were fed at random intervals, yet each bird repeated the action it happened to perform before food appeared — e.g., spinning or flapping — believing it caused the reward.

- **Lesson:** the pigeons **mistook correlation for causation.** They learned a pattern that seemed predictive but was actually meaningless.

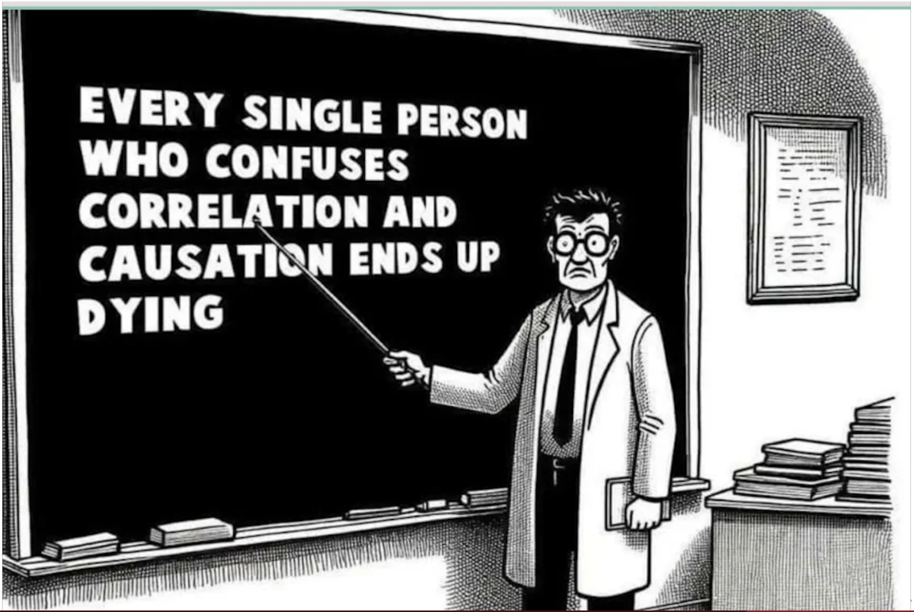
- **In Machine Learning:** Models can behave like “superstitious pigeons” — finding spurious correlations in data. Example: a model predicts “cows” when it sees grass, because cows in training data were always on grass.

- **Key takeaway:** Avoid overfitting and false generalization through proper validation, causal reasoning, and robustness checks.

Source: Skinner (1948), *“Superstition in the Pigeon.”* Journal of Experimental Psychology; Shalev–Shwartz & Ben–David (2014), Ch. 1.

Correlation vs Causation 1/2

Correlation vs Causation 2/2



Inductive Bias & Generalization

- Rats learn "food $\Rightarrow$ nausea" but not "food $\Rightarrow$ electric shock": **prior knowledge** shapes what is learnable.
- **Inductive bias:** assumptions that guide generalization beyond the data; essential for successful learning.
- **Trade-off:** stronger priors $\Rightarrow$ easier to learn with few samples, but less flexible; weaker priors $\Rightarrow$ more flexible, need more data.
- Foreshadowing: **No Free Lunch** (later) formalizes the necessity of bias.

Source: Shalev–Shwartz & Ben–David (2014), Sec. 1.1.

Correlation vs Causation — Why Both Matter

- **Correlation:** Two variables move together; signals association. *Useful for:* pattern detection, prediction, hypothesis generation.
- **Causation:** One variable directly affects another; explains mechanisms. *Useful for:* policy design, intervention, theory testing.
- **Why correlation still matters:**
  - Predictive ML models rely on stable correlations (e.g., spam filtering, credit scoring).
  - Correlations guide where to look for causal mechanisms.
  - In domains where experiments are impossible, correlation provides actionable insights.
- **Key principle:** You can have correlation without causation, but no causation without correlation. Correlation is the first step; causation is the explanation.

Sources: Pearl (2009); Shalev–Shwartz & Ben–David (2014); Angrist & Pischke (2009).

When Do We Need Machine Learning?

- **Complexity**
  - Human/animal skills hard to program explicitly (speech, vision, driving).
  - Ultra–complex data analysis beyond human capacity (astronomy, genomics, web-scale).
- **Adaptivity**
  - Environments/users change over time (handwriting, spam, speech).
  - ML systems update from data rather than fixed rules.

Source: Shalev–Shwartz & Ben–David (2014), Sec. 1.2.

Types of Learning (Taxonomy)

- **Supervised** — labels present; goal = prediction or classification.
  - *Examples:* predicting house prices, spam vs non-spam email classification, credit risk scoring.
- **Unsupervised** — no labels; goal = structure discovery or grouping.
  - *Examples:* customer segmentation in marketing, topic modeling of news articles, clustering countries by economic indicators.
- **Reinforcement** — learning to act based on feedback or rewards.
  - *Examples:* self-driving cars learning to navigate, algorithms optimizing bids in online auctions, robots learning to walk.

Source: Shalev–Shwartz & Ben–David (2014), Sec. 1.3.

Relation to Other Fields

- **Computer Science / AI:** algorithms + efficiency; leverage computation to complement human intelligence.
- **Statistics:** shared goals/techniques, but ML emphasizes **algorithms**, **finite-sample guarantees**, and often **distribution-free** settings.
- **Other links:** optimization, information theory, game theory.

Source: Shalev–Shwartz & Ben–David (2014), Sec. 1.4.

Data Science vs Machine Learning

- **Data Science:** An interdisciplinary field that employs statistical, computational, and domain knowledge methods to extract knowledge, insights, and actionable information from data (structured or unstructured).
- **Machine Learning:** A subfield of artificial intelligence focused on creating algorithms that improve automatically through experience — learning patterns from data without being explicitly programmed.

Note: Machine learning is commonly used *within* data science workflows for predictive modeling and pattern recognition.
Sources: Harvard SEAS (2023) "What is Data Science?"; Russell & Norvig (2021) *Artificial Intelligence: A Modern Approach.*

Econometrics vs Machine Learning: Conceptual Goals

- **Econometrics** focuses on **causal inference**, **parameter estimation**, and **testing economic hypotheses**.
- **Machine Learning (ML)** focuses on **prediction**, **pattern recognition**, and algorithmic optimization.
- Econometrics emphasizes **theoretical structure** and interpretability; ML emphasizes **data-driven learning** and predictive performance.
- Athey & Imbens (2019) describe this distinction as the difference between a **model-based** and an **algorithmic** approach to empirical analysis (cf. Breiman, 2001).

Source: Athey, S. & Imbens, G. (2019). *Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11:685–725.*

Econometrics vs Machine Learning: Methods and Evaluation

- **Econometrics:**
  - Relies on **parametric or semi-parametric models** grounded in economic theory.
  - Seeks **consistent**, **unbiased**, and **efficient** estimators.
  - Validation often uses **theoretical criteria** and statistical inference (e.g., hypothesis testing).
- **Machine Learning:**
  - Employs **non-parametric and algorithmic models** (e.g., trees, boosting, neural networks).
  - Optimizes for **out-of-sample prediction accuracy**.
  - Uses **cross-validation, regularization**, and **ensemble methods** to control overfitting.

Source: Athey & Imbens (2019), Sections 2–4; Mullainathan & Spiess (2017). *"Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives, 31(2):87–106.*

Integrating Econometrics and Machine Learning: Causal ML

- Modern research increasingly **integrates econometric inference with ML flexibility**.
- **Causal Machine Learning (CML):** Combines econometric identification strategies (e.g., instrumental variables, RCTs, difference-in-differences) with ML tools suited for high-dimensional or nonlinear data.
- **Key examples:**
  - **Double/Debiased Machine Learning (DML)** — Chernozhukov et al. (2018)
    - **Goal:** estimate causal parameters when many covariates exist, using ML for nuisance estimation.
    - **Idea:** use two ML models to predict the outcome and treatment; compute residuals (orthogonalization) and regress them to remove bias.
    - "Double" = two ML stages (for outcome and treatment).
    - "Debiased" = valid inference despite flexible ML.
    - **Advantages:** handles high-dimensional controls, nonlinearities, and provides valid confidence intervals.

Road-map of the Subjects – Data Science and Business Analytics

No.	Subject Name
1	Python and SQL: intro / SQL platforms
2	Algorithms for Data Science
3	Applied Finance
4	R: intro / data cleaning / basics of visualisation
5	Unsupervised Learning
6	Webscraping and Social Media Scraping
7	Statistics and Exploratory Data Analysis
8	Advanced Visualisation in R
9	Text Mining and Social Media Mining
10	Advanced Programming in R
11	Machine Learning 1: classification methods
12	Big Data Analytics
13	Machine Learning 2: predictive models, deep learning, neural networks
14	Reproducible Research
15	Communication and Autopresentation
16	Negotiations
17	Understanding Business

Tabela 1: List of Subjects – Data Science and Business Analytics (UW): https://www.wne.uw.edu.pl/application/files/4317/5075/9415/S2-DS.pdf

Data science vs. adjacent fields

- **Statistics / Econometrics:** focused on inference, causal relationships, and uncertainty quantification.
- **Machine Learning:** focused on predictive modeling and pattern discovery (goal often = high accuracy).
- **Data Engineering:** focused on reliable data collection, storage, and building pipelines at scale.
- **Data Science:** integrates all the above into an *end-to-end process*: from question → data → model → decision/impact.

Why data science entered economics

- **Explosion of new data sources:**
  - Administrative records, online platforms, transactions
  - Text (news, social media), satellite images, mobile phone data
- **Computational advances:** cheap storage, fast algorithms, cloud computing
- **Methodological shifts:**
  - Econometrics → causal inference
  - Data science → prediction, high-dimensional analysis
- **Applications:** targeting policies, market design, poverty mapping, recommender systems

Pioneers and early influence

- **Susan Athey** (Stanford): championed bringing ML to economics
  - Applications in digital platforms, auctions, causal ML
- **Guido Imbens** (Stanford): causal inference, integration with ML
- **Sendhil Mullainathan & Jann Spiess** (Harvard, Stanford): introduced ML methods to applied economists

Key contributions:

- Mullainathan & Spiess (2017) *Machine Learning: An Applied Econometric Approach*, JEP
- Athey (2018) *The Impact of Machine Learning on Economics*, NBER WP 24362
- Athey & Imbens (2019) *Machine Learning Methods that Economists Should Know About*, ARE

Applications in economics

- **Policy targeting:** predicting which households benefit most from welfare programs *(e.g., individualized subsidies, poverty mapping with satellite data).*
- **Labor markets:** resume screening, job matching, wage prediction.
- **Health economics:** personalized treatment effects, hospital resource allocation.
- **Finance:** credit scoring, fraud detection, risk assessment.
- **Market design:** auctions, ad placement, pricing algorithms on digital platforms.

Overview

- Machine Learning is a subfield of Artificial Intelligence focused on learning patterns from data.
- Two main paradigms:
  1. **Supervised Learning**
  2. **Unsupervised Learning**

Supervised Learning

Definition

Supervised learning is a type of machine learning where models are trained on labeled data — that is, each training example has an input and a corresponding correct output.

- The goal: learn a mapping from inputs X to outputs Y.
- Common algorithms:
  - Linear Regression
  - Logistic Regression
  - Decision Trees
  - Support Vector Machines (SVM)
  - Neural Networks
- Used for: classification, regression, prediction.

Unsupervised Learning

Definition

Unsupervised learning involves training models on unlabeled data — where the algorithm tries to find hidden structure or patterns without predefined outputs.

- The goal: discover structure, similarity, or relationships in data.
- Common algorithms:
  - K-Means Clustering
  - Hierarchical Clustering
  - Principal Component Analysis (PCA)
  - Association Rule Mining
- Used for: grouping, dimensionality reduction, pattern discovery.

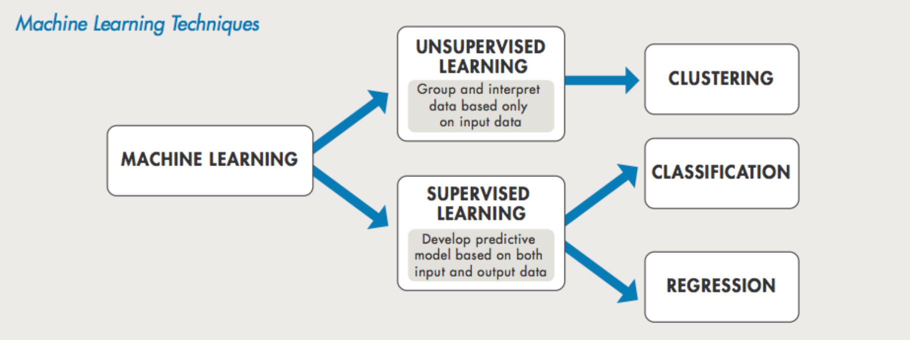
Key Differences

Aspect	**Supervised Learning**	**Unsupervised Learning**
Data Type	Labeled (input–output pairs)	Unlabeled (no predefined outputs)
Objective	Predict or classify new data	Find patterns or structure
Algorithms	Regression, SVM, Neural Nets	K-Means, PCA, Clustering
Example	Predict house prices from features	Group customers by buying behavior

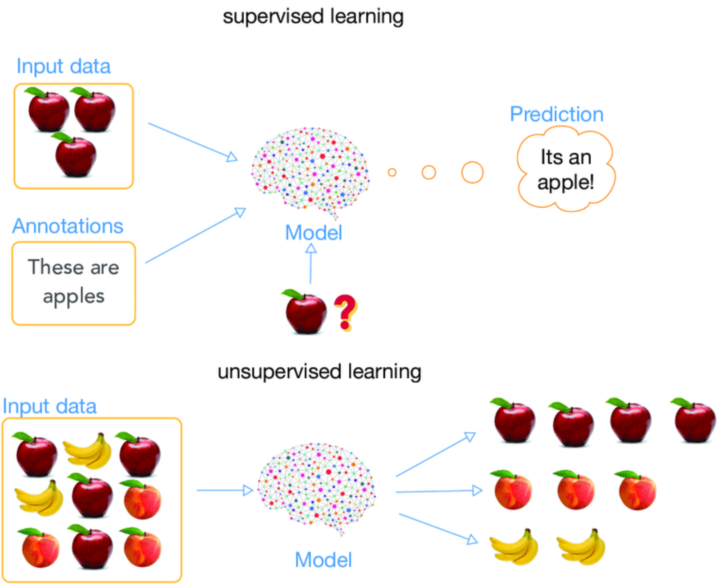
Examples in Practice

- **Supervised:** Email spam detection, medical diagnosis, stock price prediction.
- **Unsupervised:** Market segmentation, anomaly detection, image compression.

Supervised vs Unsupervised 1/2



Supervised vs Unsupervised 2/2



Summary

- Supervised learning uses labeled data to make predictions.
- Unsupervised learning uses unlabeled data to find patterns or structure.
- Both are essential for understanding and leveraging data.

Sources:

Alpaydin, E. (2020). *Introduction to Machine Learning* (4th ed.). MIT Press.
Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Data Leakage: A Critical Risk

- **Definition:** information from test set leaks into training
- Leads to overly optimistic performance estimates
- Common sources:
  - Temporal leakage (future info in training)
  - Unit leakage (same person/firm in train and test)
- Economists are vulnerable with panel/cross-sectional data

Reproducibility Crisis

- ML-based science faces reproducibility challenges
- Lack of transparent data processing and code sharing
- Risks: overfitting, cherry-picking, publication bias
- Economics must avoid importing these bad practices

Reproducibility vs Replicability in Research

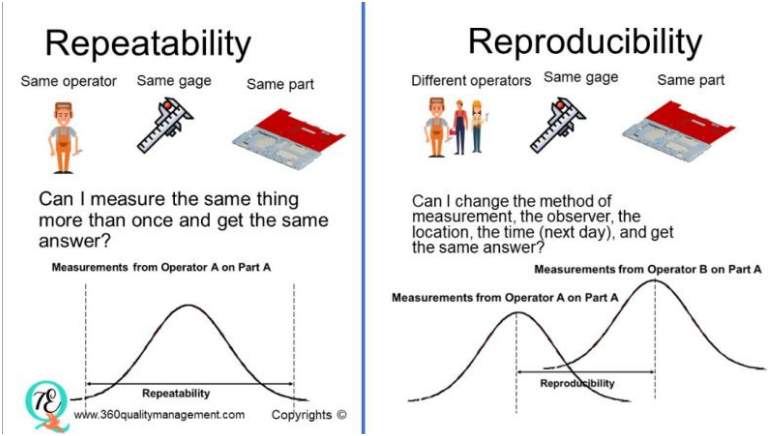
- **Reproducibility:** Re-conducting the *same study* using the *same data and methods* by a different researcher or team (Patil et al., 2016; Shokraneh 2022).
  - Confirms that results can be independently re-obtained.
  - Requires transparent reporting of data, code, and analytical procedures.
  - In systematic reviews: the ability to re-run the searches and obtain the same or very similar results.
  - Two forms: *Quantitative reproducibility* – same number of results. *Content reproducibility* – same records or studies retrieved.
- **Replicability:** Re-doing the same study to gather *new data* and check whether the same findings hold.
  - Tests robustness and generalizability of conclusions.
- **Why it matters:** Reproducibility is the hallmark of a *research study*. Without it, even a systematic review becomes merely a narrative review.

Sources: Patil et al. (2016); Shokraneh F. (2022) *Reproducibility and Replicability of Systematic Reviews*.

Reproducibility vs Replicability 1/2

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Reproducibility vs Replicability 2/2



Takeaways

- Economics embraced data science because:
  - new kinds of data became available,
  - new computational methods became feasible,
  - prediction tools complemented causal inference.
- Pioneers like **Athey, Imbens, Mullainathan, Spiess** shaped the integration.
- Today: ML methods are part of the econometrics toolbox, especially for policy targeting and heterogeneous effects.

ML in Econometrics: Famous Applications

Today: ML methods are part of the econometrics toolbox — especially for **policy targeting** and **heterogeneous effects.**

- **Targeting Poverty Alleviation (Jean et al., 2016):** Satellite imagery + convolutional neural networks used to predict regional wealth in Africa, enabling data-driven targeting of aid.
- **Tax Compliance and Audits (Kleinberg et al., 2018):** Governments use ML models to identify high-risk taxpayers and optimize audit allocation — improving efficiency and fairness.
- **Employment and Job Matching (Brynjolfsson et al., 2018):** Predictive algorithms used in labor markets to match job seekers with openings; econometric evaluation shows heterogeneous benefits across skill groups.
- **Heterogeneous Treatment Effects (Athey & Wager, 2019):** Causal forests applied to estimate individual-level (heterogeneous) treatment effects — a framework widely used in evaluating personalized policies such as health interventions, education programs, and social benefits.
- **Development Economics (Blumenstock et al., 2015):** Mobile

Responsible Data Science — Why It Matters

- **Individuals:** protect privacy, autonomy, and well-being via consent, minimization, and strong security. Build trust in data-driven systems.
- **Communities:** identify and mitigate bias; ensure transparency to foster inclusion and trust.
- **Society:** consider impacts on healthcare, education, justice, and policy; assess benefits/harms through social impact assessments and cross-disciplinary collaboration.
- **Takeaway:** technical excellence *and* ethical responsibility must go together.

Sources: Igual & Seguí (2024, Ch.12).

What Is Data Ethics?

- Ethical principles guiding the collection, storage, analysis, and sharing of data; focuses on effects on people, communities, and society.
- **Context matters** — guidance should adapt to sector-specific needs and legal duties.
- **Industry:** privacy, consent, responsible handling, transparency, accountability.
- **Academia:** research integrity, informed consent, data sharing, open science with privacy.
- **Government:** security, surveillance limits, citizens' rights, transparency and fairness.

Sources: Floridi 2016,Jobin 2019

Responsible Data Science — Core Components

1. **Principles** (lines you won't cross).
2. **Governance** (oversight).
3. **Transparency** (explainable + understandable).
4. **Fairness** (avoid unjust bias).
5. **Privacy** (lawful, respectful).
6. **Security** (defend against misuse).
7. **Robustness & Reliability** (consistent, trustworthy).
8. **Lawfulness, Accountability, Auditability** (who's responsible? keep audit trails).

Sources: Taylor 2019

Transparency & Explainability

- **Explainability:** clarify factors behind decisions (local vs global; stakeholder-tailored).
- **Approaches:** interpretable models (trees, linear, rules); feature importance; post-hoc (e.g., SHAP); visual explanations; interactive "what-if".
- **Algorithmic & Data Transparency:** document models and datasets (limits, biases, usage).

Sources: VonEschenbach 2021,Rudin 2019,Lundberg 2017,Ribeiro 2016,Mitchell 2019,Pushkarna 2022

- **Bad bias sources:** structural bias, biased collection/labels, measurement bias.
- **Individual fairness:** similar individuals ⇒ similar outcomes (incl. counterfactual fairness).
- **Group fairness (examples):**
  - *Demographic parity*
  - *Equal opportunity*
  - *Calibration:* predicted risk matches observed frequency in each group.
- **Mitigation:** pre-processing (reweighting, representation learning), in-processing (fair objectives/constraints), post-processing (threshold adjustment).

Sources: Mitchell 2021,Zemel 2013,Pleiss 2017,BarocasBook 2019,Carey 2022

Types of Bias in Research

Type of Bias	Description / Example
Selection Bias	Participants or data are not representative of the target population. *Example:* Using only urban hospitals for a national health study.
Measurement (Information) Bias	Errors in measuring exposure or outcome variables. *Example:* Self-reported income vs verified income data.
Recall Bias	Participants remember past events differently. *Example:* Patients recalling diet after diagnosis.
Observer Bias	Researcher expectations influence measurement or interpretation. *Example:* Interviewer infers positive outcomes from treated group.
Publication Bias	Studies with significant results are more likely to be published. *Example:* Journals prefer positive findings over null results.

Robustness & Reliability

- **Robustness:** withstand adversarial inputs, distributional shifts, noise/missingness.
- **Reliability:**
  - *Uncertainty awareness* — quantify epistemic/aleatoric uncertainty; "know when you don't know"
  - *Generalize under shift* — design/monitor for stability across environments.
- **Practice:** stress tests (to evaluate robustness under controlled but challenging scenarios), adversarial checks, confidence/interval reporting

Sources: Laskov 2010,Fort 2021,Mena 2021,Subbaswamy 2022

What is a Financial Stress Test?

- A financial stress test simulates how banks or economies would perform under severe but plausible adverse conditions.
- It evaluates the resilience of financial institutions to economic shocks such as:
  - Recessions
  - Market crashes
  - Interest rate spikes
  - Surges in unemployment

Purpose of Stress Testing

1. **Assess Resilience:** Test if a bank can absorb losses in extreme conditions.
2. **Protect the Economy:** Prevent failures that could trigger systemic crises.
3. **Guide Regulation:** Help regulators set capital requirements.
4. **Enhance Confidence:** Reassure investors, depositors, and the public.

How Stress Tests Work

1. **Scenario Design:**
   - Define "what-if" macroeconomic situations (e.g., GDP drop, housing crash).
2. **Modeling Impact:**
   - Estimate effects on loan losses, income, asset values, and capital.
3. **Evaluation:**
   - Compare capital ratios against regulatory thresholds.

Limitations

- Results depend on assumptions and models
- Cannot fully predict real-world crises
- May underestimate correlated global risks
- Can provide false confidence if scenarios are too mild

Quiz 1/3 — Zero or more answers may be correct

Q1. Foundations of Data Science
According to IBM's view, what best describes Data Science?

- A) The study of human–computer interaction in social networks
- **B) Combining math/statistics, programming, advanced analytics/AI and domain expertise to extract insights**
- C) The process of collecting data for statistical testing

Q2. Econometrics vs Data Science
Which statement correctly contrasts the two fields?

- A) Econometrics focuses on prediction; data science focuses on causality
- **B) Econometrics prioritizes interpretation/validity; data science prioritizes prediction/scalability**
- C) Both fields ignore statistical assumptions

Quiz 2/3 — Zero or more answers may be correct

Q3. Types of Learning
In supervised learning, models are trained on:

- A) Unlabeled data with unknown outputs
- **B) Data with known input–output pairs**
- C) Randomly generated data without structure

Q4. Structured vs unstructured data
Which of the following are *unstructured* data?

- A) Relational database
- **B) Tweets**
- **C) Photos/images**

Quiz 3/3 — Zero or more answers may be correct

Q5. Objective of ML
Main objective of machine learning:

- A) Testing economic theories
- B) Automating data collection
- **C) Learning patterns from data to improve task performance**

Q6. Data leakage
What is *data leakage*?

- A) Loss of records during preprocessing
- **B) Using test-set information during training**
- C) Incorrect encoding of categorical variables

Q7. What is NOT true about Data Science?
Which of the following statements correctly define Data Science?

- A) It focuses solely on collecting raw data without analysis
- B) It excludes programming and statistics entirely
- C) It avoids using any computational tools

Note: In this question, none of the options are correct.

Sample (Hypothetical) Exam Question

Question 1. Which of the following are *unstructured* data?

1. A) Photos/images
2. B) Tweets
3. C) Relational database

The correct answers are **(a) and (b)**. Therefore, the following provided answers yield these point values:

given anwser	a	b	c	ab	ac	bc	abc	–
results	0.4	0.4	0.0	2.0	0.2	0.2	0.4	0.2

given answer vs. **correct answer**

	a	b	c	ab	ac	bc	abc	–
a	2.0	0.2	0.2	0.4	0.4	0.0	0.2	0.4
b	0.2	2.0	0.2	0.4	0.0	0.4	0.2	0.4
c	0.2	0.2	2.0	0.0	0.4	0.4	0.2	0.4
ab	0.4	0.4	0.0	2.0	0.2	0.2	0.4	0.2
ac	0.4	0.0	0.4	0.2	2.0	0.2	0.4	0.2
bc	0.0	0.4	0.4	0.2	0.2	2.0	0.4	0.2
abc	0.2	0.2	0.2	0.4	0.4	0.4	2.0	0.0
–	0.4	0.4	0.4	0.2	0.2	0.2	0.0	2.0

Recommended References

- Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.
- VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.
- Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable* (2nd ed.). Random House Trade Paperbacks.

Online Learning Resources:

- Quant Psych — YouTube Channel
- JB Statistics — YouTube Channel
- Statistics Book in R — R Companion Handbook
- StatQuest with Josh Starmer — YouTube Channel

e.weychert@uw.edu.pl

UNIWERSYTET WARSZAWSKI | WYDZIAŁ NAUK EKONOMICZNYCH