Title: Data Collection

Course: Data Mining

Instructor: Claudio Sartori

Master: Data Science and Business Analytics

Master: Artificial Intelligence and Innovation Management

Academic Year: 2024/2025

BOLOGNA BUSINESS SCHOOL

Alma Mater Studiorum Università di Bologna

The prerequisites for a *data–driven* activity

- availability vs collection of data
- in both cases an *inventory* is needed
  - demand, customer preferences, customers, competitors, ...
- without inventory and/or acquisition plans the rate of failure is considered not less than 80%
- data collection is frequently necessary to continue for the entire project duration

What is Data Collection?

- collect, measure analyse different types of information
- standard and validated techniques
- cleaning
- transformation

Methods of Data Collection

- Primary Data Collection
  - directly from the source,
    - interviews, observations, surveys, focus groups, oral histories, ...
- Secondary Data Collection
  - data that has already been collected by someone else,
    - internet sources, government archives, libraries, ...

Primary Data Collection Methods I

- Interviews
  - the interviewer asks questions and records responses
    - flexibility in question adjustment
- Observations
  - observe and record findings of a situation
    - controlled or uncontrolled
    - straightforward
- Surveys and Questionnaires
  - broad perspective from large groups of people
    - can be conducted via various methods

Primary Data Collection Methods II

- Focus Groups
  - conducted with a group of people who share common characteristics
  - offers insights into group thinking but may lack privacy
- Oral Histories
  - opinions and personal experiences linked to a single phenomenon,
    - insights into historical events

Secondary Data Collection Methods

- Internet
  - a large pool of free and paid research resources available online
  - requires careful sourcing from authentic sites
- Government Archives
  - Offers authentic and verifiable data but may not always be readily available due to classification
- Libraries
  - Storehouse for various documents including academic research and business directories, providing valuable information for research

Use Case: Conducting Customer Surveys

- A research study was conducted by Rice University Professor Dr. Paul Dholakia and Dr. Vicki Morwitz to see whether a company could influence customers loyalty or buying habits.
- The research study was conducted over the course of a year.
- One group of customers were surveyed and the other set was not surveyed about customer satisfaction.
- In the next year, the group that took the survey were thrice as likely to renew their loyalty towards the organization than the other group

Outline

Data Collection Tools

- Word Association
- Sentence Completion
- Role-Playing
- In-Person Surveys
- Online/Web Surveys
- Mobile Surveys
- Phone Surveys
- Observation
- IOT
- Sensors
- Web Scraping

Issues and challenges

- Quality assurance and quality control
- Proactive prevention and detection of errors during and after the data collection process
- Data quality issues, inconsistent data, data downtime, ambiguous data, duplicate data, and dealing with big data

Key Steps in the Data Collection Process

The data collection process involves five key steps:

1. Decide What Data You Want to Gather
2. Establish a Deadline for Data Collection
3. Select a Data Collection Approach
4. Gather Information
5. Examine the Information and Apply Your Findings

Data Collection Considerations and Best Practices

- careful planning to collect richer, more accurate data
- evaluating the price of each data point,
- planning how to gather data,
- considering options for data collection using mobile devices, and ensuring relevance and accuracy of collected data

Outline

Quantitative

- gathering numerical data
- allow statistical analysis and objective measurement

Surveys

- asking predefined questions to a sample of respondents
- via paper-based questionnaires, online surveys, telephone interviews, or face-to-face interviews
- useful for collecting data on attitudes, opinions, behaviors, and demographic information

Experiments

- manipulating variables to observe their effects on other variables
- conducted in controlled settings to establish cause-and-effect relationships
- can be laboratory-based or conducted in real-world environments

Observational Studies

- systematically observing and recording behaviors, events, or phenomena
- researchers do not intervene or manipulate variables
- useful for studying behaviors, interactions, and patterns in natural settings

A/B Testing for Data Collection

Definition: A/B testing is a controlled experiment comparing two variants (A and B) to determine which performs better based on predefined metrics

Purpose in Data Collection:

- Collect data under **real-world conditions** to evaluate the effect of changes
- Ensure **data-driven decisions** before deploying changes widely
- Generate **clean, labeled data** for modeling or evaluation

Key Elements:

- **Control Group (A):** Existing system or process
- **Test Group (B):** Variant with a change (e.g., new feature, design)
- **Metrics:** Conversion rate, click-through rate, revenue per user, etc

Example: An e-commerce site tests two layouts:

A/B Testing: Use Cases Beyond Websites

Widely applicable across industries where decisions benefit from controlled experimentation

Domain	Use Case
Mobile Apps	Test UI layouts, onboarding flows, in-app notifications.
Email Campaigns	Compare subject lines, send times, or content styles.
Product Development	Evaluate feature sets, pricing models, or workflows.
Advertising	Test ad creatives, calls-to-action, targeting strategies.
Retail / E-commerce	Experiment with product placements, discounts, signage.
Healthcare	Compare treatment protocols, reminders, digital interventions.
Finance	Test risk models, credit scoring, or customer messaging.
Gaming	Evaluate game mechanics, rewards, level designs.
Education / EdTech	Test lesson formats, quiz styles, feedback mechanisms.

Example: Mobile App

- **A:** Current onboarding with 5 steps
- **B:** Streamlined 3-step onboarding + tutorial video
- **Goal:** Increase user retention after 1 week

Statistical Validation in A/B Testing

Purpose: Ensure observed differences between A and B are **statistically significant** and not due to random chance

Key Concepts:

- **Null Hypothesis (H_0):** No difference between A and B
- **Alternative Hypothesis (H_1):** A and B differ significantly
- **p-value:** Probability of observing the result (or more extreme) if H_0 is true
- **Significance Level (α):** Threshold for rejecting H_0 (commonly $\alpha = 0.05$)
- **Confidence Interval (CI):** Range within which the true effect likely falls (e.g., 95% CI)

Example: In an email A/B test, variant B shows a 3% higher click rate than A, with a **p-value = 0.03**. Since $p < 0.05$, the difference is statistically significant at 95% confidence

Note: Adequate sample size is crucial for reliable results

Outline

Qualitative data

- Qualitative data may not fit traditional numeric representations used in ML algorithms
- Converting qualitative data into a format suitable for ML can lead to information loss or distortion
- Choosing appropriate features and representations is crucial for preserving the richness of qualitative data

Subjectivity and Bias

- Qualitative data often contains subjective interpretations and biases
- ML models trained on biased data may perpetuate or amplify existing biases
- Addressing subjectivity and bias requires careful preprocessing, feature engineering, and model evaluation

Examples: Subjectivity and Bias in Data Collection

- **Qualitative Data Bias:** Customer feedback labeled as ?positive? or ?negative? may vary by annotator due to subjective interpretations
- **Historical Bias:** Loan approval datasets may reflect past discriminatory practices, leading to biased ML models that unfairly deny certain groups
- **Amplification of Bias:** A hiring algorithm trained on biased hiring data may favor resumes similar to historically hired profiles, reinforcing lack of diversity
- **Preprocessing to Address Bias:** Removing sensitive features (e.g., gender, race) or rebalancing class distributions during data preprocessing can mitigate bias
- **Model Evaluation for Fairness:** Use fairness metrics (e.g., equal opportunity, disparate impact) during model evaluation to detect and address bias

Lack of Ground Truth

- Unlike quantitative data, qualitative data may lack a clear ground truth or objective measure
- Evaluating the performance of ML models becomes challenging without a reliable benchmark
- Researchers must rely on alternative evaluation methods, such as expert judgment or consensus validation

Examples: Lack of Ground Truth in Data Collection for Qualitative Data

- **Subjective Labels:** In sentiment analysis, the same review may be labeled as *neutral* by one annotator and *positive* by another, with no objective ground truth
- **Ambiguity in Medical Imaging:** Diagnosing rare conditions from X-rays may lack a definitive ground truth, especially when expert opinions diverge
- **Creative Content Evaluation:** Assessing *quality* of art, music, or writing generated by ML models is subjective and lacks an objective benchmark
- **Impact on Model Evaluation:** Without ground truth, accuracy or precision may be unreliable; alternative methods like **expert judgment** or **consensus validation** are used
- **Example Method:** Collect multiple expert labels and use majority vote or confidence-weighted consensus to approximate ground truth

Interpretability

- ML models trained on qualitative data may lack interpretability
- Understanding how and why a model makes decisions is crucial, especially in sensitive or high-stakes applications
- Techniques for explaining and interpreting ML models need to be adapted for qualitative data analysis

Data Complexity

- Qualitative data can be highly complex and multi-dimensional
- ML algorithms may struggle to capture the nuanced relationships and patterns present in qualitative data
- Advanced techniques, such as deep learning or ensemble methods, may be necessary to handle data complexity effectively

Conclusion on qualitative data

- Machine learning analysis of qualitative information presents several challenges
- Addressing these issues requires interdisciplinary collaboration and innovative methodological approaches
- Overcoming these challenges can unlock valuable insights and applications in fields such as natural language processing, social sciences, and healthcare