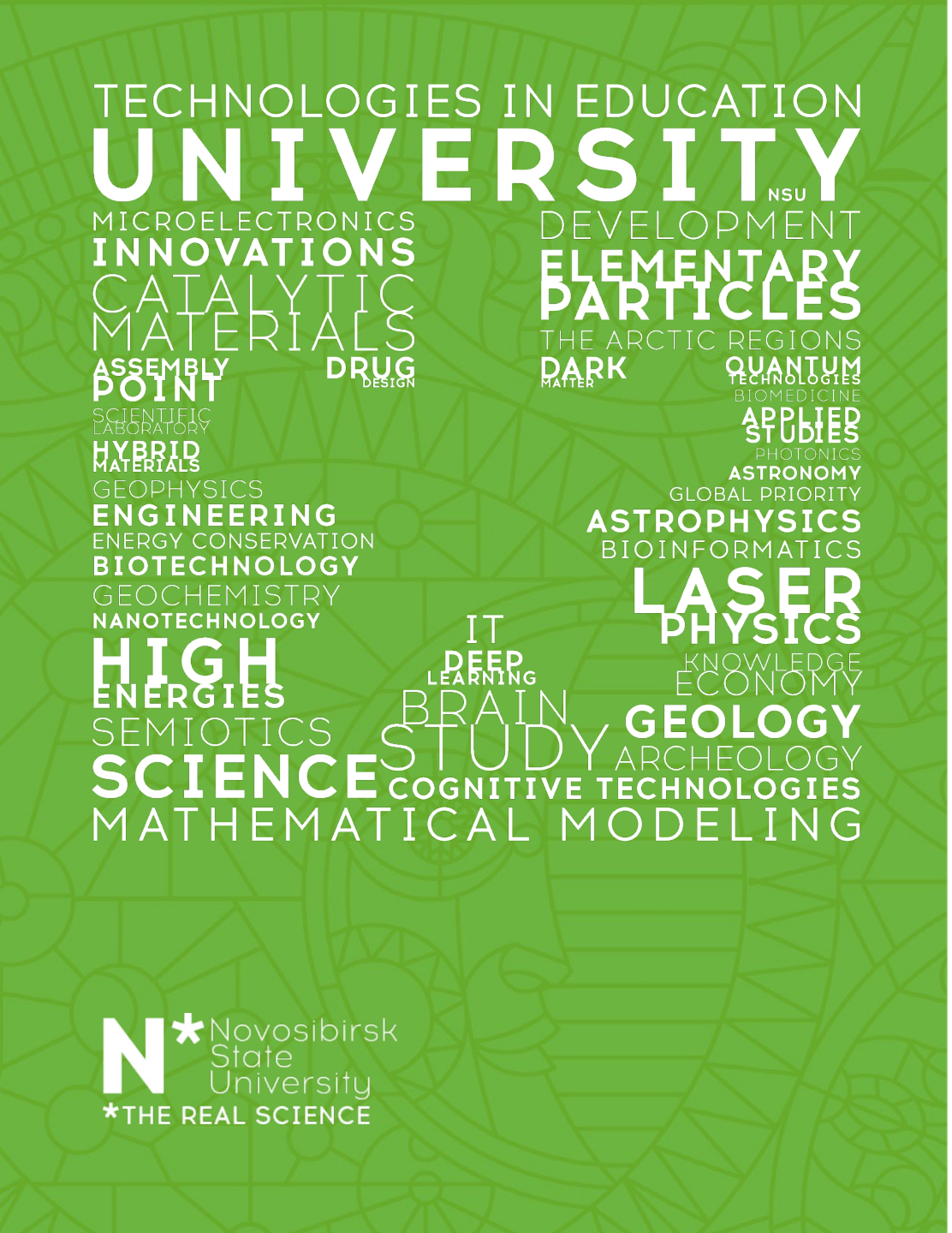


Data Preprocessing

Luu Minh Sao Khue
Department of Mathematics and Mechanics

Overview

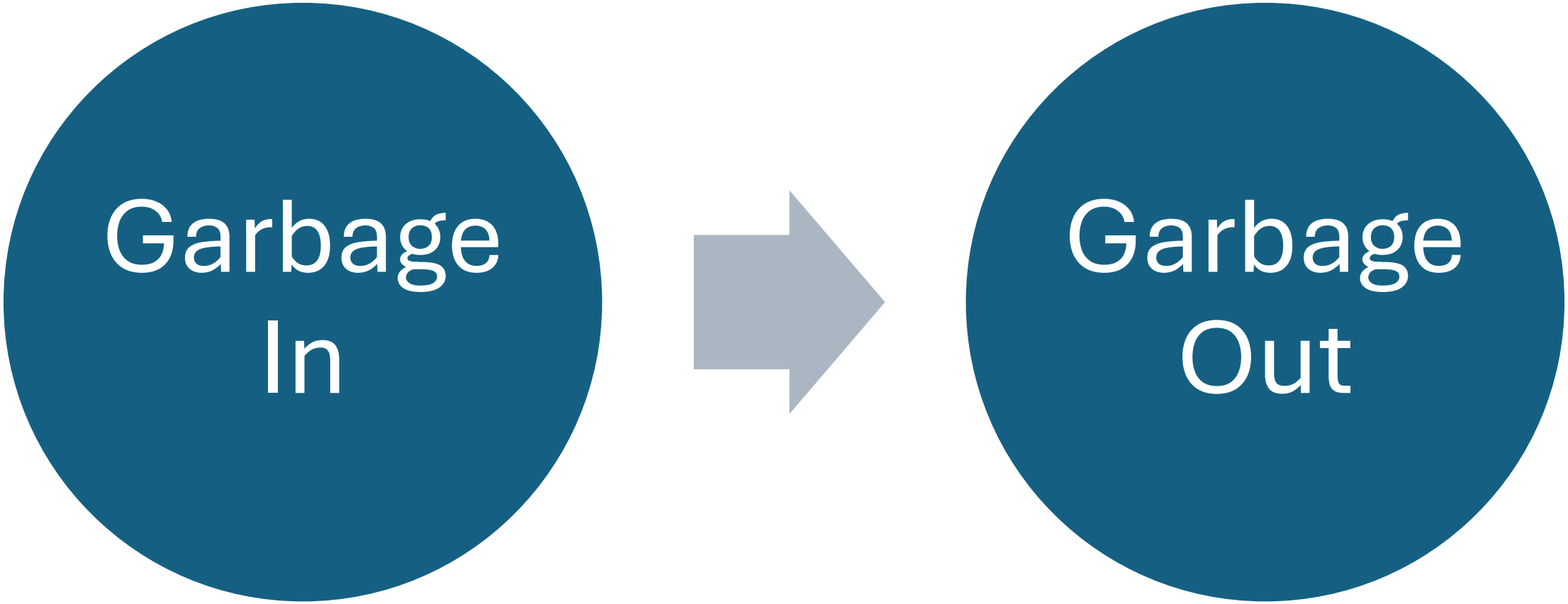
- Data challenges in Machine Learning
- Data preprocessing techniques

Data in Machine Learning

The Importance of Data in Machine Learning

- Data is the Foundation
  Models rely on data to learn and make predictions.
- Quality Matters
  High-quality data = accurate models; poor data = inaccurate, biased results.
- Better Decisions
  Clean data leads to useful insights that support better decision-making.
- Reduces Bias
  Representative data helps ensure fair and ethical AI outcomes.
- Takeaway: "Good data = Good models."

The Importance of Data in Machine Learning



Data Challenges

1. Quality and Integrity Challenges

- Missing Data: Data is often incomplete due to unrecorded values or errors during collection.
- Noisy Data: Data with random errors or noise, often requiring cleaning or denoising techniques.
- Outliers: Extreme values that can distort models, particularly in sensitive algorithms.
- Duplicate Data: Multiple records for the same entity can bias the model by giving undue weight to certain information.
- Incorrect Labels: Mislabeling in supervised datasets can confuse models and reduce accuracy.

2. Data Distribution Challenges

- Imbalanced Data: Disproportionate representation of classes, which can bias models toward the majority class.
- Skewed Data Distribution: Highly skewed features can impact model accuracy, especially for algorithms assuming normal distributions.
- Class Boundary Overlap: Poorly defined boundaries between classes make classification challenging.
- Temporal Drift (Concept Drift): Statistical properties of data change over time, causing models to become outdated.

3. Feature-Related Issues

- Irrelevant Features: Non-informative features add noise to model training.
- High Dimensionality (Curse of Dimensionality): Too many features increase computational complexity and the risk of overfitting.
- Feature Interactions: Complex relationships between features can be hard to capture.
- Class Boundary Overlap: Overlapping features make it hard for models to distinguish between classes.

4. Data Representation and Format Issues

- **Inconsistent Formatting:** Variations in data formats (e.g., dates or units) lead to processing errors.
- **Heterogeneous Data Sources:** Data from different sources with varied formats and structures complicate integration.
- **Unstructured Data:** Text, images, or audio need special processing for machine learning.

6. **Data Privacy and Security**

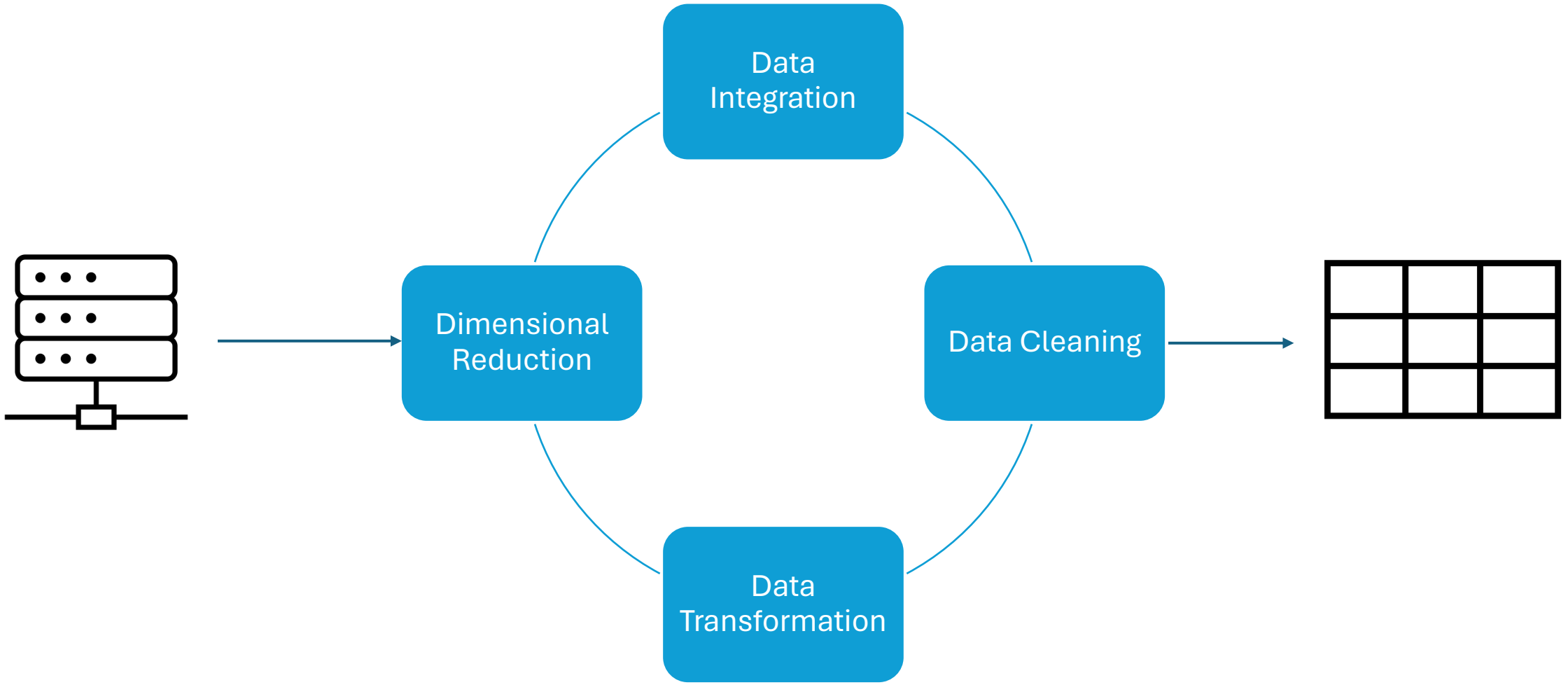
- **Privacy Concerns:** Sensitive information must be anonymized or handled carefully.
- **Data Protection Laws:** Compliance with regulations like GDPR or HIPAA is necessary.
- **Security Issues:** Data breaches or leaks can compromise sensitive datasets.

7. Labeling and Annotation Challenges

- **Inadequate Labeling:** Small or incorrectly labeled datasets reduce model quality.
- **Expensive Labeling Processes:** Human annotations, especially in specialized fields (e.g., medical), are costly.
- **Class Ambiguity:** Ambiguous labels make it hard for models to learn accurately.

Data Preprocessing

What is data preprocessing?



Lab

Link:
https://github.com/luumsk/NSU_ML/blob/main/Labs/lab2a.ipynb

References

- https://en.wikipedia.org/wiki/Missing_data#:~:text=In%20statistics%2C%20missing%20data%2C%20or,be%20drawn%20from%20the%20d
- https://www.scalablepath.com/data-science/data-preprocessing-phase