Computational Data Analysis (02582)

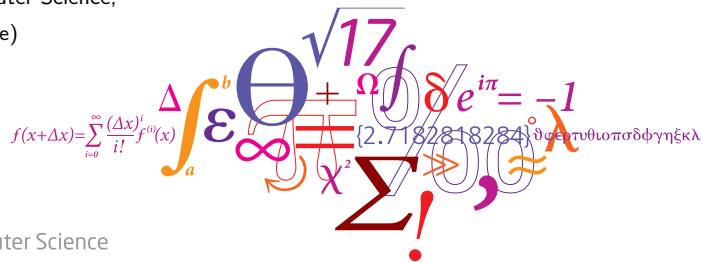
INTRODUCTION

Sneha Das

Assistant Professor

Department of Applied Mathematics and Computer Science,

Technical University of Denmark (DTU Compute)



DTU Compute

Department of Applied Mathematics and Computer Science

Outline I

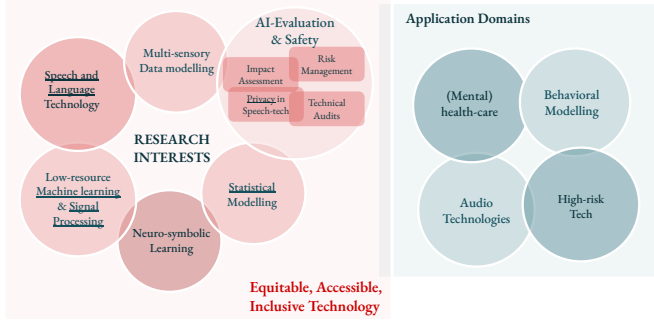
- Course details
- Part 1 - CDA and a very brief history

Meet the Instructors

Sneha Das (Course-responsible), Assistant Professor, DTU Compute (sned@dtu.dk)

- **Speech Tech:** Low-resource speech recognition and generative modeling.
- **AI Safety:** Alignment, robustness and safety in small-sample environments.
- **Applications:** Modeling behaviors via multisensory sensor data (psychiatry, mental-health).

PhD in speech communication technology from Aalto University, Finland | Interested in learning from finite, high-stakes data while ensuring the 'safety' of applying AI models.



Meet the Instructors

Line H Clemmensen (Co-responsible)
Professor, DTU Compute (lkhc@dtu.dk)

- Evaluation and fairness AI
- Machine learning in Life-science
- Learning from small and large sample sizes
- Applications in psychiatry, drug discovery, and more

Principal Data Scientist, Mærsk Digital
PhD in Statistical Image Analysis from DTU

The Support Ecosystem - TA team

Emil A. Jensen
PhD Student in Multi-modal models.

Tova Alenfalk
PhD Student in Sustainability Performance of Bio-based Solutions.

Martin Bertelsen
PhD Student in Intelligent hydroacoustics).

Eva Paraschou
PhD Student in AI alignment and safety.

Mustafa Jawad El-Madani
Masters student and TA.

Format & Communication

- Lecture from 08:00-10:00.
- Exercises from 10:00-12:00.
- Outstanding questions: Send email.

Course overview

NOTE: The course overview is subject to changes (mostly in topic grouping and ordering)
Refer to Learn under 'Course overview'

Examination & Projects

Component	Description	Weight
Case 1	Statistical Engineering baseline + report hand-in	15%
Case 2	(Poster & Peer Review) + report hand-in	25%
Final Exam	Written Examination	60%

Mandatory Requirement

Passing both cases is mandatory to enter the final exam.
Pass is 7.5/15 points for case 1 and 12.5/25 points for case 2

Written examination

Multiple choice and open questions.

Course material

1. Slides and exercises

- Defines the course
- Our selection of topics and connections
  - Broad overview (macroscopic view)
  - Theoretical level (microscopic view on few essential topics)

2. The Elements of Statistical Learning

- Book by T. Hastie, R. Tibshirani and J. Friedman (2nd edition)
- Available online `https://hastie.su.domains/ElemStatLearn/`
- (Fairly) Statistical perspective.
- Supplementary/suggested papers and book chapters will be given

3. Computer exercises and cases

- Theory and application - hands on
- Software: **Python**.

Our aim

After the course you should (hopefully)...

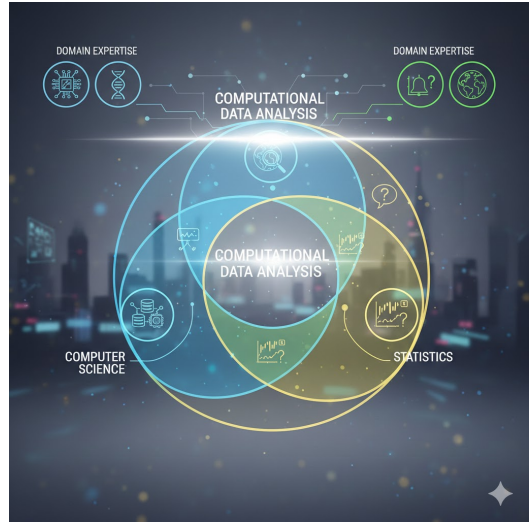
- Say that it was **tough** but you learned a lot.
- Have a **toolbox** of methods to use.
- Know the methods such that you can,
  - **demonstrate** understanding of their usage,
  - **evaluate** methods and their theories,
  - **motivate** choices of methods and parameters.
- Be **ready** to learn more.

INTRODUCTION

What is CDA

Computational Data Analysis is the intersection of three vital pillars:

- **Computer Science**
  - Algorithms & Data Structures
  - Scalability & Performance
- **Statistics**
  - Mathematical Framework
  - Quantifying Uncertainty
- **Domain Expertise**
  - Contextual Meaning
  - 'The story' behind the data



The Modern Workflow: A Pipeline Perspective

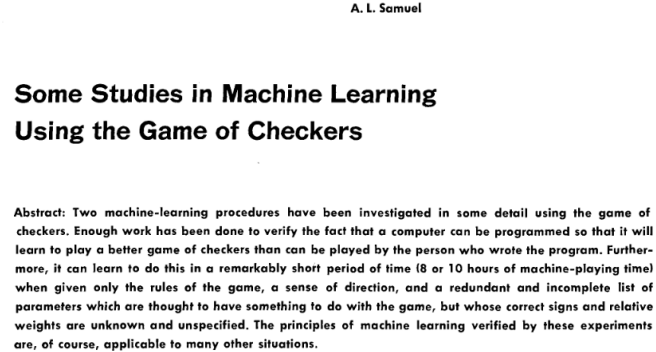
Computational Data Analysis is a multi-stage **pipeline**. The 'analysis' isn't just the final model; it's the architecture supporting every step. It is a complex system

Stage	The Computational Challenge
Ingestion	Handling high-velocity streaming data or scraping massive web repositories.
Cleaning	Automated 'wrangling' to handle missing values and outliers at scale.
Exploration	Using high-dimensional visualization to find hidden patterns.
Inference	Applying ML or Statistical Modeling to predict and explain.
Deployment	Turning a model into a functional tool, dashboard, or API.

Table: The CDA Pipeline Stages

Efficiency at scale requires computational automation at every node.

The Evolution of Learning



A. L. Samuel

Some Studies in Machine Learning Using the Game of Checkers

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.

A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," in IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229, July 1959, doi: 10.1147/rd.33.0210.

1959: The 'Explicit' Barrier

Arthur Samuel defined the field as:

'The ability to learn without being explicitly programmed.'

- Transition from logic gates to pattern recognition.
- Foundations of modern multisensory analysis.

Experience, Task, Performance

1997

Tom Mitchell's Axiom:
A computer program learns from **Experience (E)** with respect to **Task (T)** and **Performance (P)** if its performance at tasks in T, as measured by P, improves with experience E.

Modern CDA Context

Computational Data Analysis helps us maximize information extraction from limited or "expensive" samples.

2001: The Two Cultures

Leo Breiman's Wake-up Call

Statistical modeling vs. Algorithmic modeling.

- **Traditional:** Start with data, assume a stochastic model (e.g:, OLS).
- **Algorithmic:** The 'Black Box' approach. Accuracy over interpretability.

'If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.'

This course lives at the intersection: Understanding the box while respecting the math.

August 2001

Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)

Leo Breiman

Statist. Sci. 16(3): 199-231 (August 2001). DOI: 10.1214/ss/1009213726

ABOUT | FIRST PAGE | CITED BY | REFERENCES

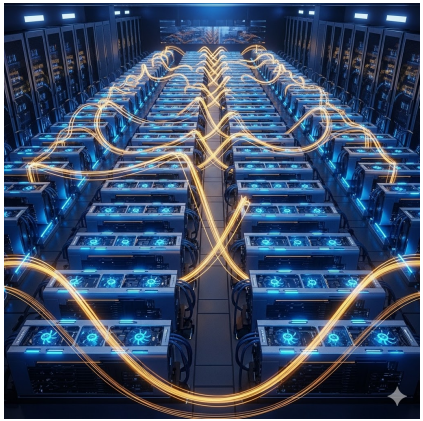
Abstract

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Citation Download Citation

Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." Statist. Sci. 16 (3) 199 - 231, August 2001. https://doi.org/10.1214/ss/1009213726

2012: The Era of 'Big Data' and Deep learning



The Hardware-Algorithm Handshake
The explosion of compute (GPUs) met high-capacity models (Neural Nets).

- **Result:** A decade obsessed with 'More Data'.
- **The Catch:** Most high-stakes problems (Medicine, Safety) don't have infinite data.

CDA remains critical here

CDA in the Era of Deep Learning (DL)

If we have 'Black Box' DL, why learn traditional CDA?

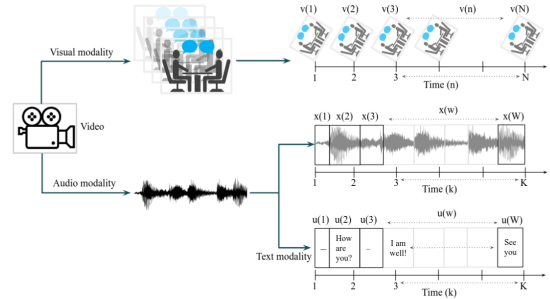
- **Understanding vs. Prediction:** DL is excellent at *predicting* what happens; CDA is essential for *explaining why* it happens (Causal reasoning to model outcomes).
- **Data Efficiency:** DL is 'data-hungry'. Many sensitive domains are data scarce (amount and quality). CDA provides the tools to extract maximum signal from small, noisy, or highly structured datasets.
- **The First Principles Argument:** Deep Learning *is* CDA. Neural networks are simply massive compositions of linear algebra, calculus, and statistical optimization. $\rightarrow$ Impactful innovation in AI goes back to the first principles.

CDA in the Multi-Sensory Frontier

Beyond Flat Tabular Data

CDA in **Multisensory High-Stakes ML**.

- **Unimodal:** Images, text, speech and audio
- **Multimodal:** Fusing audio, physiological sensors, and clinical text.
- **Small-Sample:** Learning efficiently from rare events or niche languages.
- **Safety-Critical:** Models that provide uncertainty and don't fail silently.



Lønfeldt, N. N., Das, S., Mora-Jensen, A. R. C., Pagsberg, A. K., & Clemmensen, L. (2023). Scaling-up Behavioral Observation with Computational Behavior Recognition.

Application: CDA in diagnosis of neurodegenerative disorders from voice

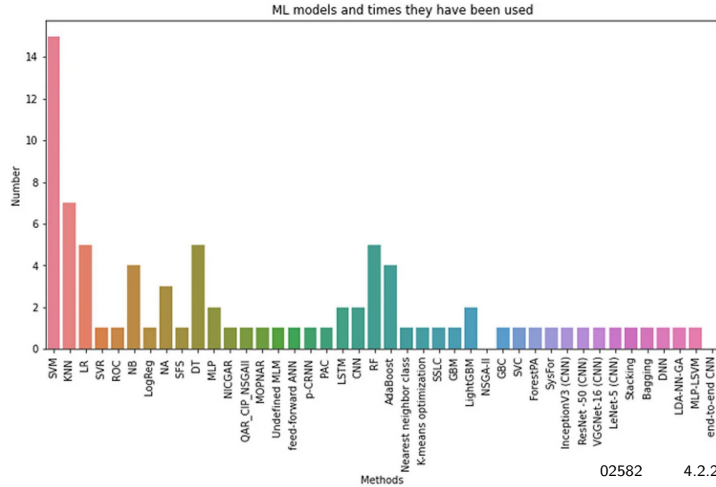
Input is a voice signal
Difference in the voice signals of patients with Parkinson's and control participants

- **Challenge:** Small samples, high noise, missing data.
- **CDA Utility:** Feature selection and robust covariance estimation.
- **Goal:** Objective diagnostic tools.

Das, S. (2026). AI Models for Voice Disorders: Considerations from Development to Deployment. In Voice-related Biomarkers (pp. 73-85). Cham: Springer Nature Switzerland.

Fig. 6.2

From: AI Models for Voice Disorders: Considerations from Development to Deployment



Usage of ML Models. Distribution showing the different models. (Based on an overview of papers in collaboration with M. Pedersen)

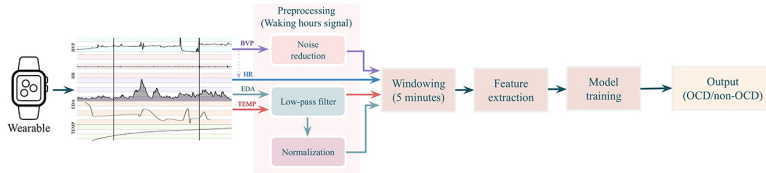
Application: Multi-modal Modeling

Case: High-Stakes Psychiatry

Using wearable sensor data to model behavior and clinical outcomes.

Predicting obsessive-compulsive disorder episodes in adolescents using a wearable biosensor-A wrist angel feasibility study

- **Challenge:** Small samples, high noise, missing data.
- **CDA Utility:** Feature selection and robust covariance estimation.
- **Goal:** Objective diagnostic tools.



Lønfeldt, N. N., Olesen, K. V., Das, S., Mora-Jensen, A. R. C., Pagsberg, A. K., & Clemmensen, L. K. H. (2023). Predicting obsessive-compulsive disorder episodes in adolescents using a wearable biosensor—A wrist angel feasibility study. Frontiers in Psychiatry, 14, 1231024.

Logistic regression, Random forest (RF), Feedforward neural networks, and Mixed-effect random forest (MERF)

CDA: The Triple Convergence + domain expertise

We are at a unique historical moment where three forces have met:

Ubiquity
Sensors are everywhere (IoT, Phones, Satellites).

Compute
GPU and Cloud accessibility is at an all-time high.

Openness
Open-source libraries (Python/R) have democratized the 'tools'.

The Missing Ingredient: Domain (Human) Insight.