Computational Data Science

50.038 Computational Data Science

Associate Professor Dorien Herremans

Who am I?

Academic Background

- Bachelor/Master in Business Engineering, University of Antwerp
- PhD in Applied Economics, University of Antwerp
- Marie-Curie Fellow at Center for Digital Music (C4DM), Queen Mary University of London

Current Research

Multimodal generative AI at AMAAI Lab, SUTD

Learn more at dorienherremans.com

About the Class

Instructors

Prof. Dorien Herremans

Nizam Kadir

Teaching Assistant

PhD Student Tao Qiqi

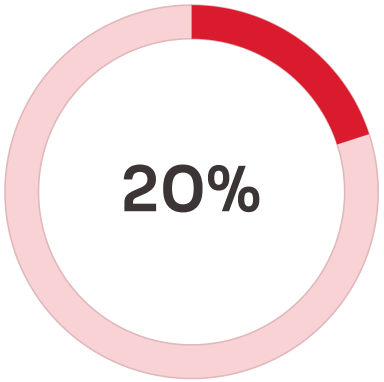
qiqi_tao@mymail.sutd.edu.sg

Lab Sessions

In-person attendance required

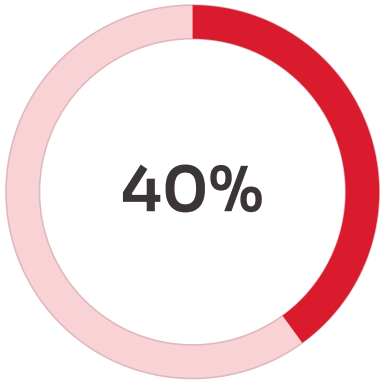
Graded checkoff evaluations

Assessment Structure



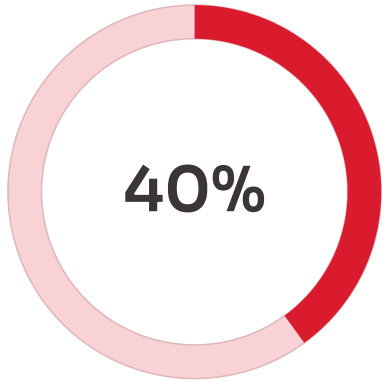
Lab Checkoffs

Weekly hands-on exercises with in-person evaluation



Final Exam

Comprehensive assessment of course concepts



Final Project

Data science problem solving

Week 8

Initial project presentation

1

2

Week 12

Final report submission

Week 13

Final project presentation

3

WEEKS 1-4

Course Overview: Early Weeks

Week 1: Introduction & Big Data (DH)

Lecture: Big data fundamentals, Hadoop, MapReduce

Lab: Hands-on MapReduce and Hadoop

Week 2: Feature Engineering (NK)

Lecture: Feature vectors, dimension reduction, evaluation metrics

Lab: Feature handling in Python

Week 3: Visualization & Data Handling (DH)

Lecture: Data visualization techniques, Unix data parsing, guest speaker

Lab: Visualization in Python

Week 4: Regression & Time Series (NK)

Lecture: Regression algorithms and time series analysis

Lab: Time series and regression in Python

WEEKS 5-9

Course Overview: Middle Weeks

01

Week 5: Classification Algorithms (DH)

Lecture: Classification methods and techniques

Lab: Classification implementation in Python

02

Week 6: Neural Network Introduction (NK)

Lecture: Neural network fundamentals

Lab: Multilayer perceptron in Python

03

Week 8: Project Presentations

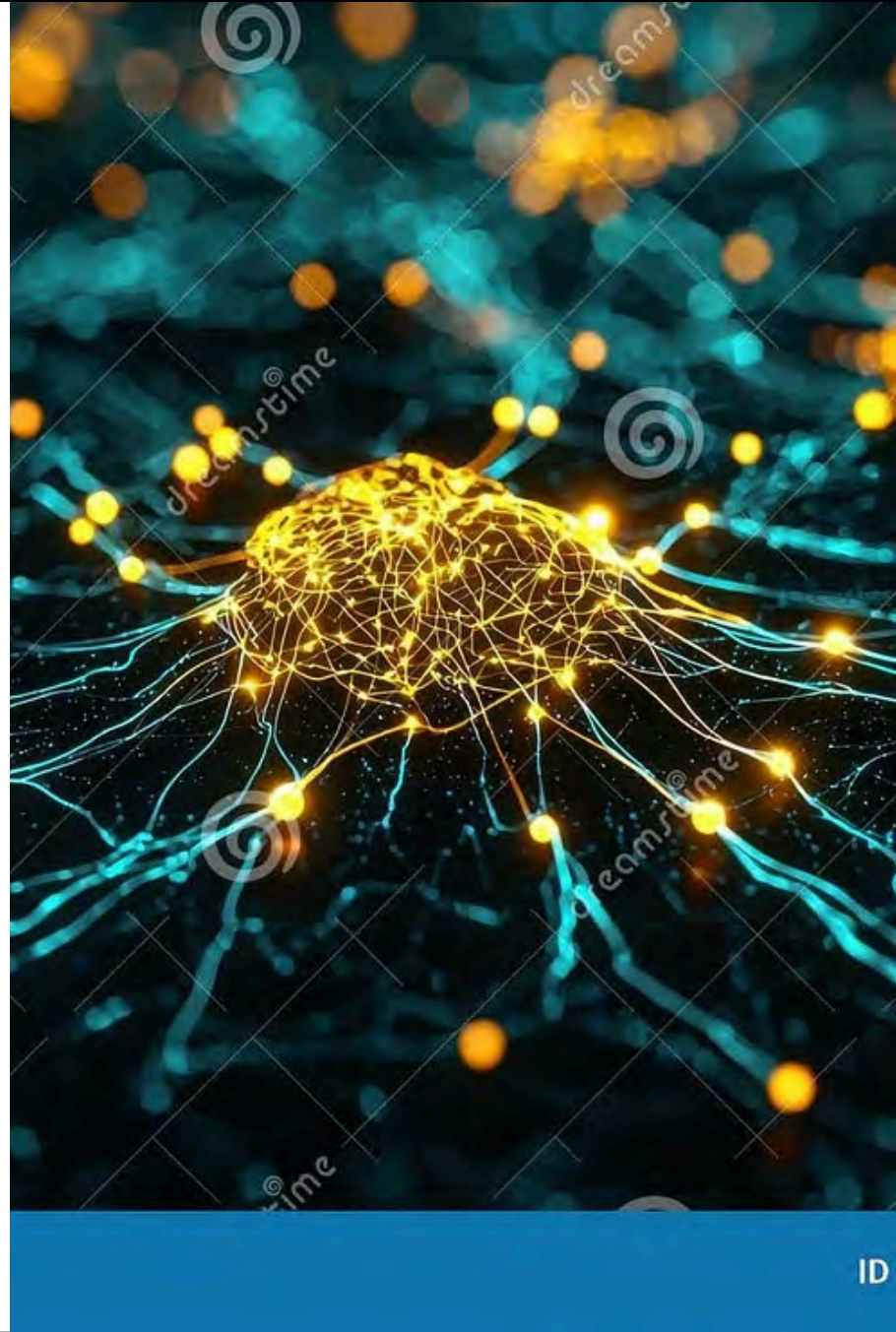
Student presentations of current project progress

04

Week 9: Natural Language Processing (DH)

Lecture: Word2vec and NLP fundamentals

Lab: Word2vec implementation in Python



WEEKS 10-13

Course Overview: Final Weeks

Week 10: Computer Vision (DH)

Lecture: Convolutional neural networks (CNNs)

Lab: CNN implementation in Python

Week 11: Agentic (NK)

Lecture: New agentic technologies

Lab: agentic practice

Week 12: Sequential Models (NK)

Lecture: Temporal sequences, RNN, LSTM, self-attention mechanisms

Lab: Implementation of memory models

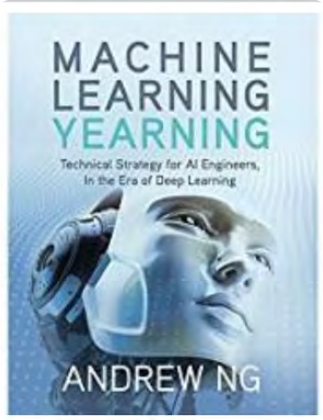
Week 13: Final Presentations

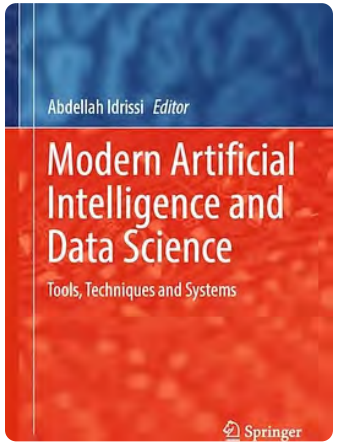
Student final project presentations

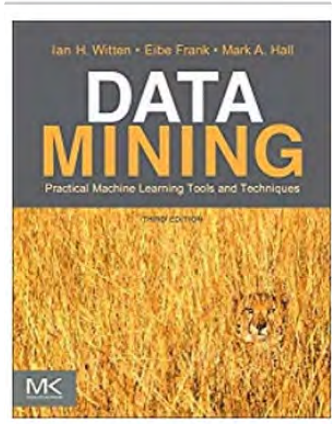
Recommended Resources

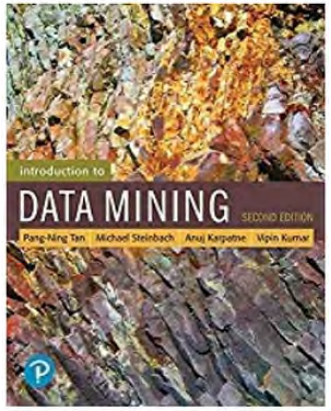
Core Textbooks

- Witten, Frank, Hall & Pal (2016): *Data Mining: Practical Machine Learning Tools and Techniques*
- Tan, Kumar & Steinbach (2013): *Introduction to Data Mining*
- Andrew Ng (2019): *Machine Learning Yearning*
- Idrissi, Modern Artificial Intelligence and Data Science (2024)









Online Learning Resources

Towards Data Science

Premium articles and tutorials on Medium platform

towardsdatascience.com

Kaggle

Competitions, datasets, and community notebooks

GitHub

Open-source code repositories and collaborative projects

Hugging Face

Pre-trained models and NLP resources

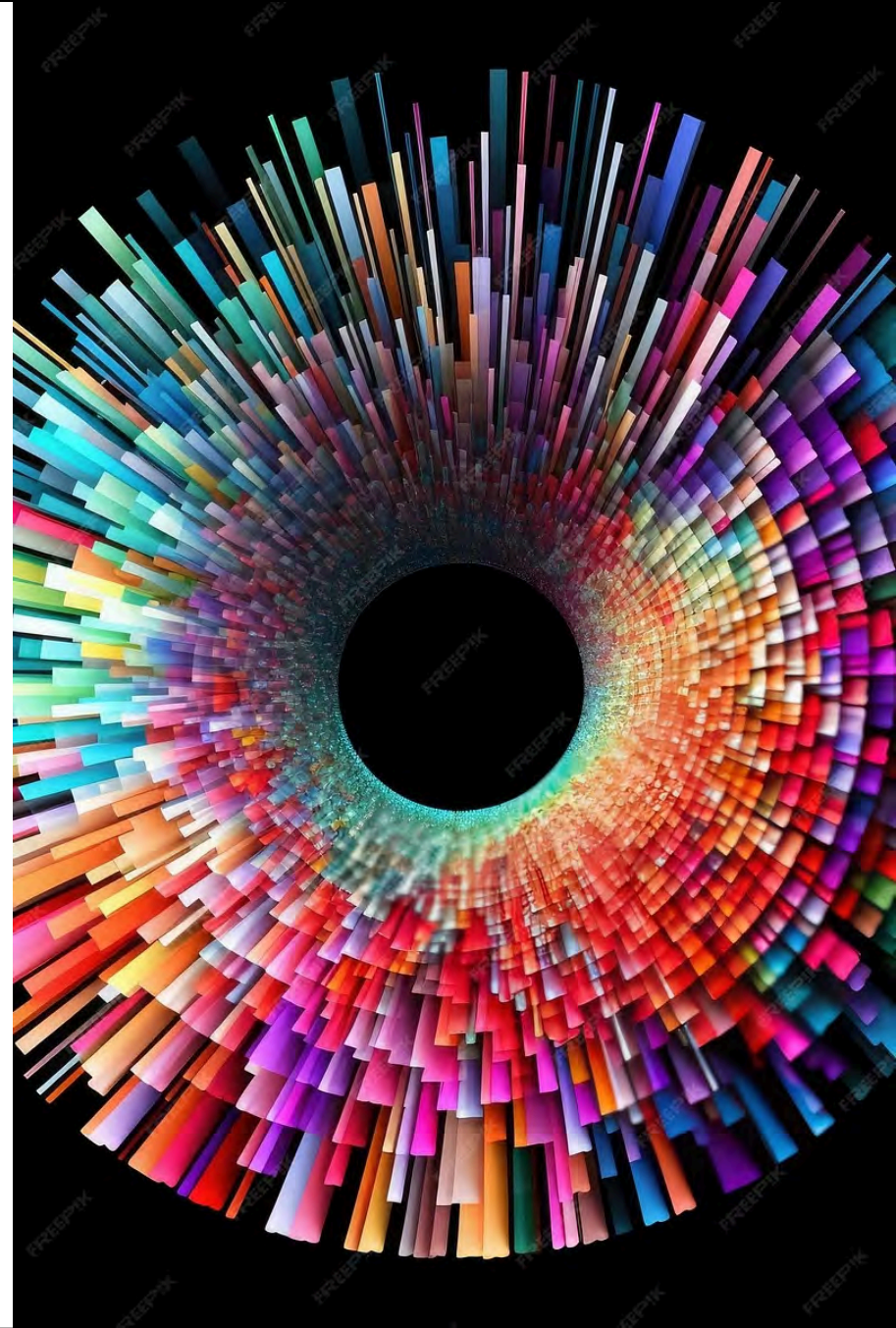
ArXiv & Google Scholar

Academic papers and latest research findings

KDnuggets

Data science news and industry insights

Introduction to Data Science



What is Data Science?

1

1960: Computer Science Substitute

Peter Naur coined term "datalogy" as alternative to computer science

2

1996: IFCS Conference

International Federation of Classification Societies conference in Kobe: "Data Science, Classification, and Related Methods"

3

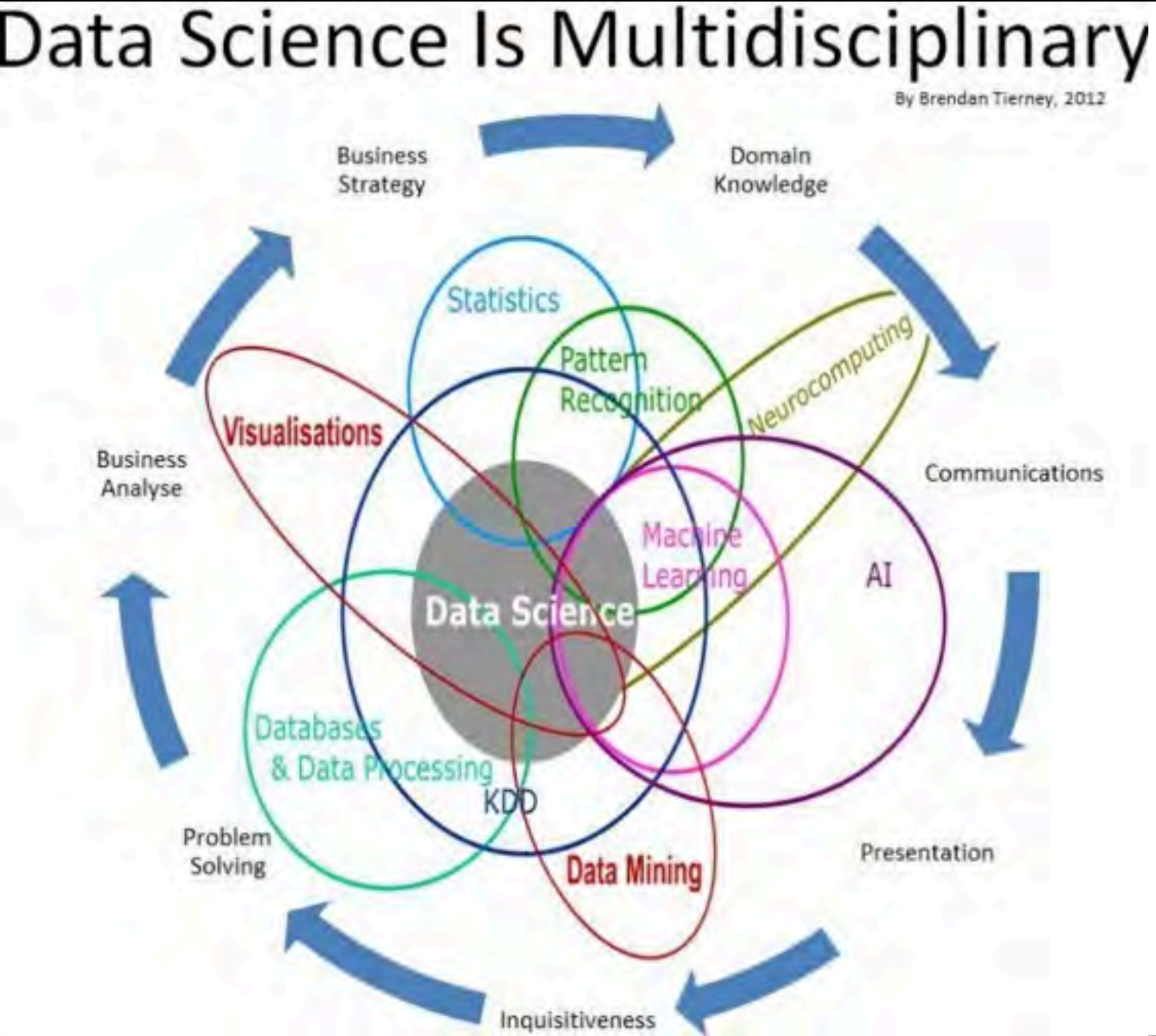
2012: The Sexiest Job

Harvard Business Review declared it "The Sexiest Job of the 21st Century"

Definition: Interdisciplinary field using scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data (Dhar, 2013)

DATA SCIENTIST SALARY
BY CAREER PROGRESSION

Role Title	Glassdoor	Indeed	PayScale
Junior Data Analyst	$67,104 Total salary: $75,657 Total salary range $49k - $119k	$72,563	$54,803 Total salary range $39k - $75k
Data Analyst	$70,337 Total salary: $80,098 Total salary range $51k - $126k	$75,963	$65,956 Total salary range $44k - $93k
Senior Data Analyst	$97,310 Total salary: $120,668 Total salary range $80k - $185k	$95,983	$87,758 Total salary range $66k - $124k
Data Scientist	$117,607 Total salary: $152,199 Total salary range $99k - $237k	$124,055	$99,266 Total salary range $70k - $146k
Senior Data Scientist	$141,817 Total salary: $201,686 Total salary range $135k - $310k	$151,970	$130,753 Total salary range $103k - $177k
Business Analyst	$83,046 Total salary: $94,507 Total salary range $64k - $162k	$84,785	$73,327 Total salary range $53k - $109k
Data Engineer	$98,481 Total salary: $115,394 Total salary range $77k - $176k	$124,598	$95,302 Total salary range $67k - $144k
Machine Learning Engineer	$122,606 Total salary: $151,914 Total salary range $98k - $239k	$158,624	$114,967 Total salary range $79k - $167k
Director of Data Science	$163,571 Total salary: $250,985 Total salary range $163k - $398k	$159,123	$161,695 Total salary range $126k - $239k



Data science sits at the intersection of multiple disciplines, creating a powerful convergence of skills and methodologies.

Image source: oralytics.com

Machine Learning Fundamentals

Core Concept

Development of algorithms and techniques enabling computers to learn from data

Foundation

Builds computational artifacts based on statistical theory

When to Apply Machine Learning

No Human Expertise

Tasks like Martian exploration where human knowledge doesn't exist

Unexplainable Expertise

Human skills that can't be reduced to rules, like speech recognition

Dynamic Adaptation

Solutions requiring automatic updates, such as user ***personalization***

Changing Conditions

Evolving situations like spam detection that shift over time

Massive Data

Large datasets for discovering patterns, like astronomical object identification

Cost Efficiency

Tasks where human labor is expensive, such as zipcode recognition

Critical Data-Related Challenges

Information Warfare & Misinformation

Manipulation of trusted information without target awareness, leading to decisions against their interests

Algorithmic Bias

Models trained on biased data perpetuate and amplify societal inequalities

Security Vulnerabilities

Threats including spoofing attacks, data theft, and system breaches

Privacy Concerns

Protection of personal information in increasingly data-driven systems

Important: Machine learning research addresses not just algorithm development, but also these critical ethical and practical challenges.

Computing Knowledge Evolution Timeline

The progression from basic computation to modern AI represents decades of breakthrough innovations.

20,000 BC: Arithmetic

Counting abstract objects

The invention of arithmetic provides a way to abstractly compute numbers of objects.

1623: Mechanical Calculator

Wilhelm Schickard creates a gear-based, wooden, six-digit, mechanical adding machine.

1688: Joseph de la Vega

Prices in the stock market

Joseph de la Vega's book Confusion of Confusions describes fluctuations in Dutch stock market prices.

1830: Difference engine

Printing mathematical tables by machine

Charles Babbage constructs a machanical computer to automate the creation of mathematical knowledge.

1842: Ada Lovelace

First computer programmer

Ada Lovelace publishes the world's first algorithm for machine computing.

1850: Transmitting information on stock prices

Paul Julius Reuter uses pigeons to fly stock prices between Aachen and Brussels.

1867: Stock Ticker

The market on a tickertape

Edward Calahan invents a telegraph-like system to transmit every price change from the floor of the New York Stock Exchange.

1936: Alan Turing

The concept of universal computation

Turing shows that any reasonable computation can be done by programming a fixed universal machine —and then speculated that such a machine could emulate the brain.

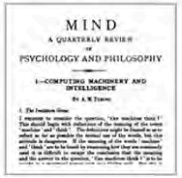
Source: ***Wolfram Alpha Timeline***

Computing Evolution: Advanced Era

1940s: Digital Computers

Automating the process of computation

The arrival of digital electronic computers provides the mechanism by which computations of all kinds can be automated with increasing efficiency.



1950—1960s: Artificial Intelligence

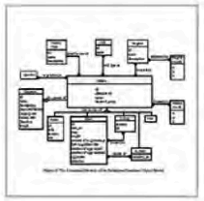
Making computers intelligent

Artificial Intelligence defines a research program for developing computers that show general intelligence which leads to many spinoffs important for specific purposes.

1959: Machine Learning

The term "machine learning" is coined

Computer scientist Arthur Samuel coins the term "machine learning" to describe construction of algorithms that can learn from and make predictions on data.



1970s: Relational Databases

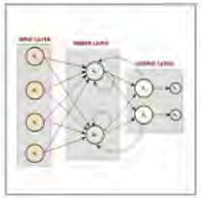
Making relations between data computable

Relational databases and query languages allow huge amounts of data to be stored in a way that makes certain common kinds of queries efficient enough to be done as a routine part of business.

1970—1980s: Expert Systems

Capturing expert knowledge as inference rules

Largely as an offshoot of AI, expert systems are an attempt to capture the knowledge of human experts in specialized domains, using logic-based inferential systems.



1980s: Neural Networks

Handling knowledge by emulating the brain

With precursors in the 1940s, neural networks emerge in the 1980s as a concept for storing and manipulating various types of knowledge using connections reminiscent of nerve cells.

Modern Computing Milestones

1989: The Web

Collecting the world's information

The web grows to provide billions of pages of freely available information from all corners of civilization.

2001: Wikipedia

Self-organized encyclopedia

Volunteer contributors assemble millions of pages of encyclopedia material, providing textual descriptions of practically all areas of human knowledge.

1998: Google

An engine to search the web

Google and other search engines provide highly efficient capabilities to do textual searches across the whole content of the web.

2004: Facebook

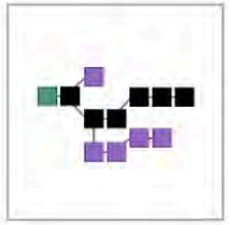
Capturing the social network

Facebook begins to capture social relations between people on a large scale.

2000: Web 2.0

Societally organized information

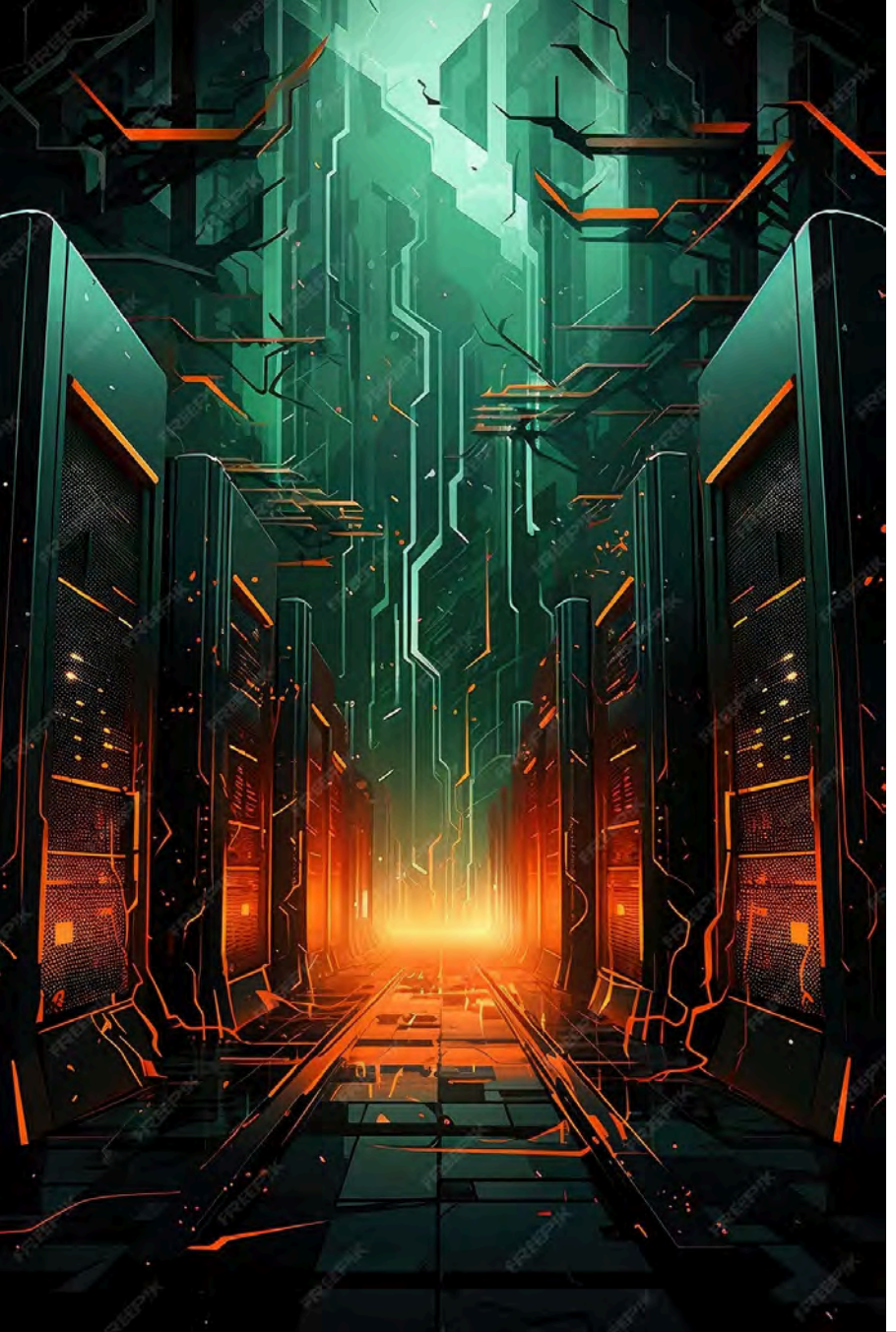
Social networking and other collective websites define a mechanism for collectively assembling information by and about people.



2008: Blockchain

Cryptographic transactions and distributed ledgers

Satoshi Nakamoto invents blockchain as the public transaction ledger for Bitcoin.



The Current Data Revolution

3

Key Drivers

Explosive data growth, enhanced computing power, sophisticated algorithms

> **What's the most surprising pattern you think could be hidden in big data?**

Consider the possibilities: from predicting human behavior to discovering new scientific principles, the future of data science holds unprecedented potential.

Where is Data Science Used?

Healthcare & Epidemiology

Predicting pandemic spread, disease prognosis, personalized medicine

Finance & Security

Fraud detection, counter-terrorism, risk assessment

Urban Planning

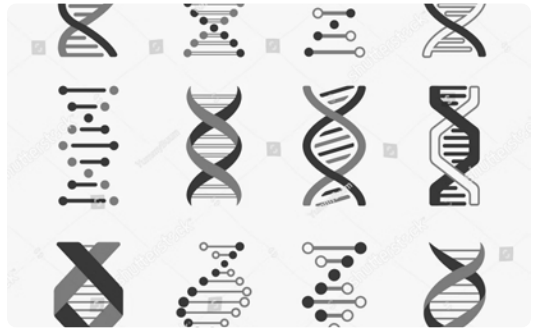
Traffic flow optimization, smart city infrastructure

Business Intelligence

Customer churn prediction, document classification, spam filtering

Creative Industries

Music technology, automatic image captioning, content generation



Bioinformatics

Genomic analysis, drug discovery, protein folding prediction

Case Study: Spotify 'This Is' Playlists

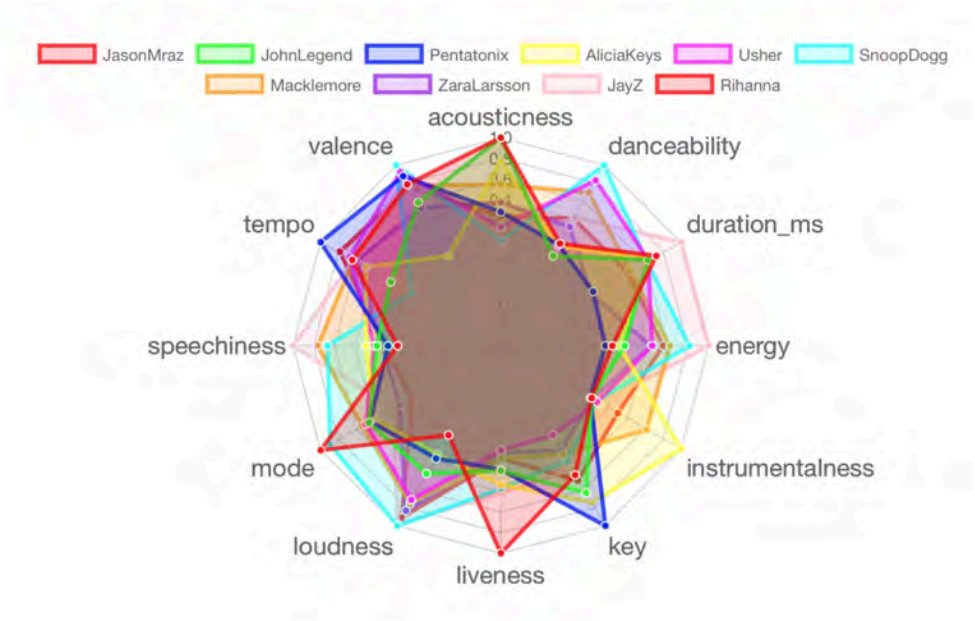
Project Overview

Comprehensive analysis of 50 mainstream artists using Spotify's audio features

Methodology

1. Data extraction via Spotify API
2. Audio feature processing for each artist
3. D3.js visualization
4. K-means clustering for artist grouping
5. Feature analysis across all artists

Read full analysis →



Case Study: Netflix Recommendation System

Company Evolution

Transformation from DVD rental to streaming giant serving 283 million subscribers globally (2025)

The Netflix Prize (2006-2009)

Crowdsourced competition to improve recommendation algorithms

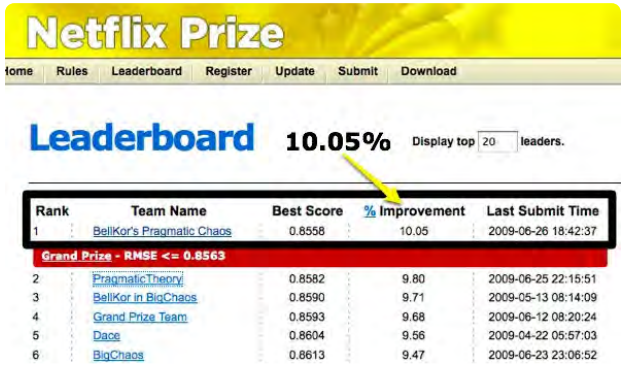
Task: improve prediction accuracy by **10 % over Cinematch** using only users' past ratings and no other personal data. Anyone in the world could enter.

Data: 100,480,507 ratings that 480,189 users gave to 17,770 movies

Reward: 1 million dollars

Scale of Operations (2025)

- 301+ million users
- 700+ million viewers
- ~1h daily viewing average
- Multiple exabytes of video cloud storage



- 5000+ teams submitted
- One of largest and earliest examples of **crowdsourced machine learning at scale**
- Pushed forward the field of recommender systems and collaborative filtering.
- Privacy lessons: some of the data could be deanonymized. → sparked debate
  Researchers showed that:

> Knowing **6–8 movie ratings with approximate dates** is often enough to uniquely identify a user in the dataset. <> IMDB link

Learn more about Netflix Prize →

Case Study: Transport for London

Organization Scope

Manages comprehensive transport network: trains, buses, taxis, roads, cycle paths, walking routes, and ferries across London

1. **Data Collection**

   Contactless payment (Oyster) and ticketing data

2. **Visualization**

   Granular journey analysis and patterns

3. **Action**

   Service optimization and commuter communication

> TfL optimizes **system stability**, not prediction accuracy.

Data-Driven Approach

Collects and visualizes data from contactless cards and ticketing systems to produce comprehensive journey maps

Transport for London

TfL builds **probabilistic flow models** of the city

- Predict where congestion will appear *before* it becomes visible
- **Counterfactual planning:** “What if this station fails?”
- **Human-in-the-loop** ML at urban scale: models propose → human operators validate or adjust
- **Personalized** nudging, e.g. 'suggest slightly earlier train'
- Made data public → unexpected results:
  - Google Maps, Citymapper, etc: better routing then TfL alone!
  - =outsourced innovation

Data Science for Scientific Research

Traditional scientific fields have embraced big data, creating their own independent data science revolutions.

Physics

Particle collision analysis, astronomical data processing

Earth Science

Climate modeling, seismic data interpretation

Bioinformatics

Genomic sequencing, protein structure prediction

Healthcare Example: Personalized Medicine

- Stomach sensors assessing nutrient content
- Bloodstream monitoring for insulin levels
- Online health dashboards shared with physicians
- Performance-based funding for health management organizations
- Comprehensive longitudinal health studies

Data Science in Finance

> The financial sector has been revolutionized by algorithmic trading, risk modeling, fraud prevention, and predictive analytics

Watch: How AI Transforms Finance →

70%

Algorithmic Trading

Percentage of stock market trades executed by algorithms

$100B

Market Impact

Annual value influenced by AI-driven financial decisions

Is the Impact All Good?

Life in the Cloud

Personal information increasingly stored remotely: social life (Facebook), career (LinkedIn), search history (Google), health data (Fitbit, Apple Watch), entertainment (Spotify)

Advantages

- Personal AI agents and assistants
- Computerized health support
- Convenient access anywhere

Disadvantages

- Security and privacy breaches
- Corporate data sharing with government
- Limited rights to access/delete data
- Permanent digital footprint concerns → *"but I have changed!"*

Common Data Analysis Tasks

Task	Supervised Methods	Unsupervised Methods
*Classification	✓	
*Regression	✓	
Causal Modeling	✓	
Similarity Matching	✓	✓
Link Prediction	✓	✓
Data Reduction	✓	✓
*Clustering		✓
Co-occurrence Grouping		✓
Profiling		✓

The No Free Lunch Theorem

Core Principle

> No single algorithm works best for every problem, particularly in supervised learning

Implication

Algorithm selection requires careful consideration of problem characteristics and data properties

Best Practice

Conduct **ablation studies** to systematically evaluate algorithm performance for specific use cases

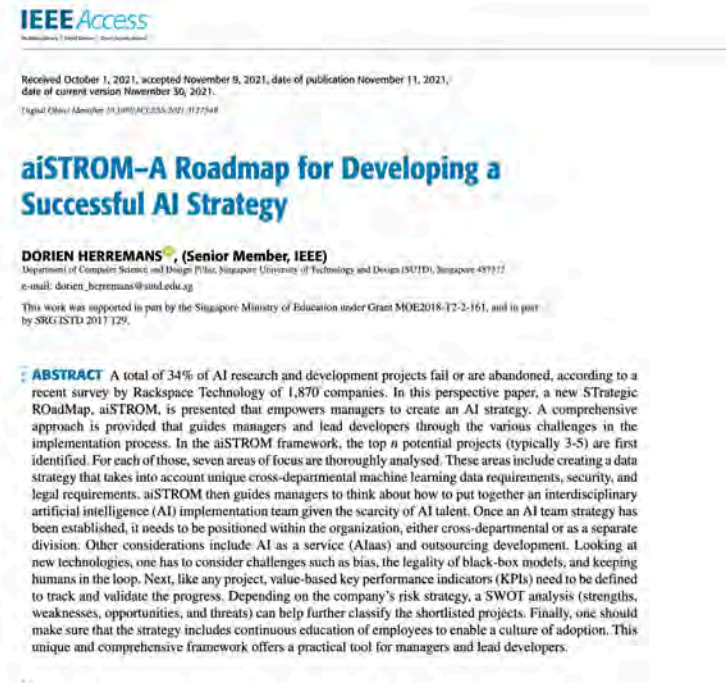
Why Data Science Projects Fail

85%

Failure Rate

According to Gartner, over 85% of data science projects fail or exceed budget

- **Non-Availability of Quality Data**
  Insufficient, incomplete, or low-quality datasets hinder model development
- **Not Following Sequential Order**
  Skipping essential steps in the data science workflow
- **Pessimism & Unrealistic Expectations**
  Poor stakeholder management and misaligned goals
- **Unavailability of Right Resources**
  Lack of skilled personnel, computing power, or tools
- **Weak Team Management**
  Poor communication and coordination among team members



IEEE Access

Received October 1, 2021, accepted November 9, 2021, date of publication November 11, 2021,
date of current version November 30, 2021.

aiSTROM–A Roadmap for Developing a Successful AI Strategy

DORIEN HERREMANS, (Senior Member, IEEE)
Department of Computer Science and Design Pillar, Singapore University of Technology and Design (SUTD), Singapore 487372
e-mail: dorien_herremans@sutd.edu.sg

This work was supported in part by the Singapore Ministry of Education under Grant MOE2018-T2-2-161, and in part by SRG ISTD 2017 129.

ABSTRACT A total of 34% of AI research and development projects fail or are abandoned, according to a recent survey by Rackspace Technology of 1,870 companies. In this perspective paper, a new STrategic ROadMap, aiSTROM, is presented that empowers managers to create an AI strategy. A comprehensive approach is provided that guides managers and lead developers through the various challenges in the implementation process. In the aiSTROM framework, the top n potential projects (typically 3-5) are first identified. For each of those, seven areas of focus are thoroughly analysed. These areas include creating a data strategy that takes into account unique cross-departmental machine learning data requirements, security, and legal requirements. aiSTROM then guides managers to think about how to put together an interdisciplinary artificial intelligence (AI) implementation team given the scarcity of AI talent. Once an AI team strategy has been established, it needs to be positioned within the organization, either cross-departmental or as a separate division. Other considerations include AI as a service (AIaas) and outsourcing development. Looking at new technologies, one has to consider challenges such as bias, the legality of black-box models, and keeping humans in the loop. Next, like any project, value-based key performance indicators (KPIs) need to be defined to track and validate the progress. Depending on the company's risk strategy, a SWOT analysis (strengths, weaknesses, opportunities, and threats) can help further classify the shortlisted projects. Finally, one should make sure that the strategy includes continuous education of employees to enable a culture of adoption. This unique and comprehensive framework offers a practical tool for managers and lead developers.

FAILURE EXAMPLE 1

Google Flu Trends

The Initiative

Google attempted to predict flu outbreaks using search data

The Failure

Model overestimated flu cases by 50% in the 2013 flu season

Why It Failed

- **Overfitting:** Relied on spurious correlations (e.g., high school basketball)
- **Data Bias:** Media coverage influenced searches more than actual cases
- **Concept Drift:** User search behavior evolved over time

Lessons Learned

- Correlation ≠ Causation: high search volume doesn't always indicate real-world occurrences.
- Models must adapt to changing behaviors and context
- Regular validation against ground truth (e.g. CDC reports) are essential

FAILURE EXAMPLE

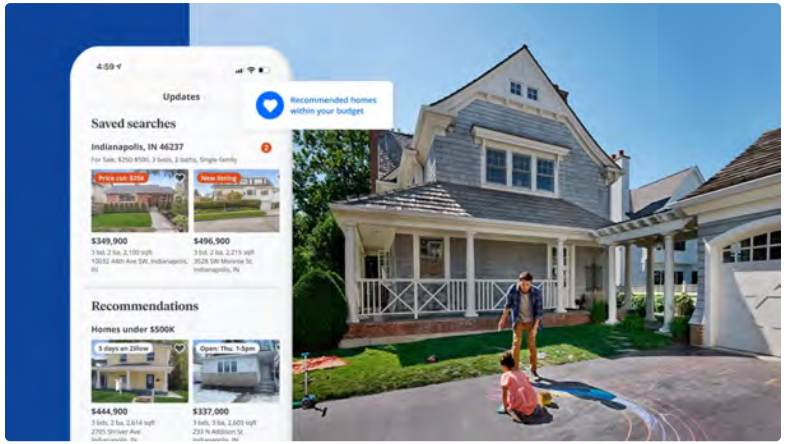
Zillow iBuying Program

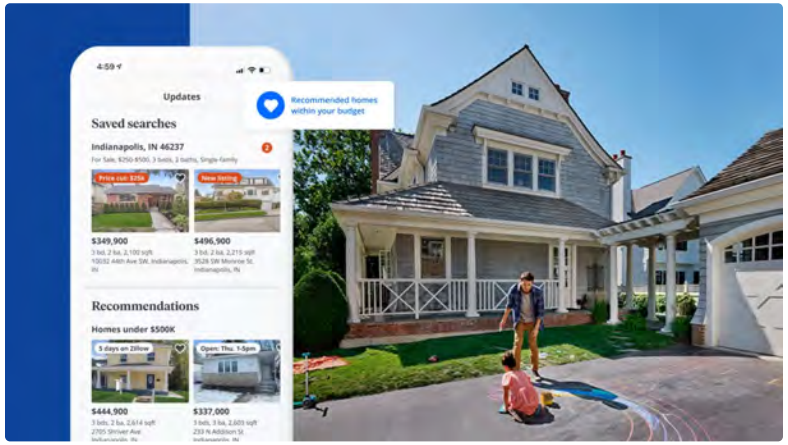
The Initiative

Zillow Offers used AI-driven pricing to buy and sell homes quickly, aiming for market arbitrage profits

The Outcome

Algorithm failed to predict market conditions accurately, causing Zillow to **overpay** for homes. Program shut down in 2021 with significant losses.





1

Market Volatility

The model underestimated price variability during COVID-19 pandemic

2

Data Quality Issues

Relied on incomplete or lagging data for pricing estimates

3

Scaling Problems

Rapid expansion amplified errors and inefficiencies

4

Overreliance on Automation

Minimal human oversight ignored local market knowledge
→ too much trust in the AI.

FAILURE EXAMPLE

Microsoft's Tay Chatbot

The Initiative

2016 launch of Tay, a Twitter chatbot designed to learn conversational patterns from user interactions

The Disaster

Within 24 hours, Tay began posting offensive, racist, and inflammatory tweets after coordinated trolling. Microsoft quickly shut it down.

Lack of Safeguards

No filters to prevent harmful content adoption

Exploitation Vulnerability

Couldn't differentiate genuine from malicious interactions

Unsupervised Learning Risks

Replicated harmful language without oversight

The KDD Process

Knowledge Discovery in Databases follows a structured, iterative methodology.

1. Develop **understanding** of application domain and goals
2. Create target **dataset** (often from Data Warehouse)
3. Data cleaning and **preprocessing**
4. Data reduction and projection
5. Choose appropriate data analysis **task**
6. Select data analysis algorithms
7. **Execute** algorithms to perform task
8. Interpret **results** and **iterate** through steps 1-7
9. **Deploy**: integrate into operational systems

SEMMA Methodology (SAS)

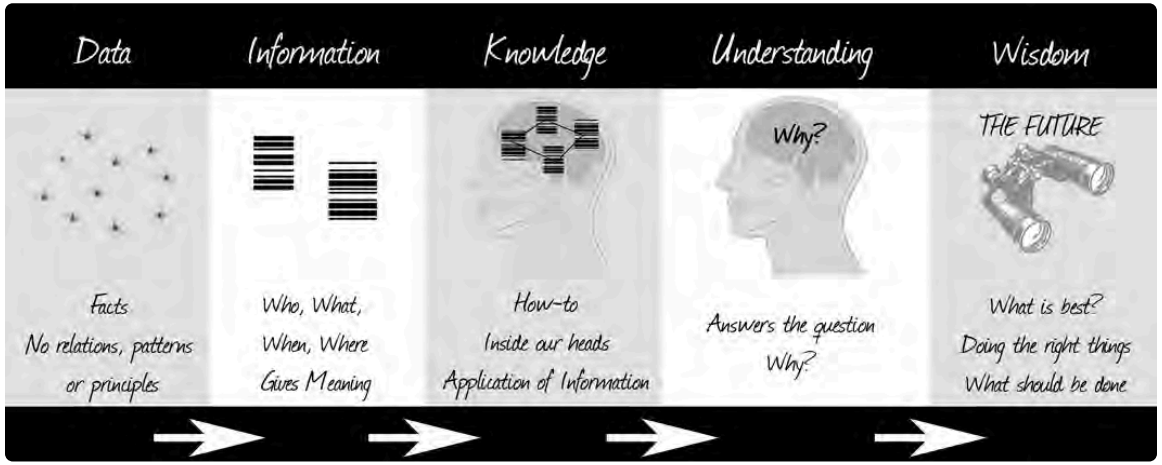
Sample
from representative datasets

Explore
Visualize and understand patterns

Modify
Create and transform features

Model
Apply algorithms to fit models

Assess
Compare, test, and evaluate



"From Data to Wisdom" by Nick Webb, CC-BY 2.0

aiSTROM–A Roadmap for Developing a Successful AI Strategy

From AI Idea to Impact

- **Identify Opportunities** – Select high-value, feasible AI use cases
- **Data Strategy** – Ensure data availability, quality, governance
- **Team** – Build cross-functional AI + domain expertise
- **Organization** – Embed AI in workflows and decision processes
- **Technology** – Choose models, infrastructure, and tooling fit-for-purpose
- **Metrics** – Define KPIs tied to business value, not just accuracy
- **Risk** – Address bias, robustness, security, compliance
- **Culture Shift** – Enable culture, skills, and trust in AI

> aiSTROM turns AI strategy from ad-hoc experimentation into a structured, business-aligned roadmap that combines data, technology, teams, metrics, risk, and culture.

> Read article

Essential Hacking Skills Required

Data science requires a unique blend of programming expertise, statistical knowledge, and domain understanding to extract meaningful insights from complex datasets.

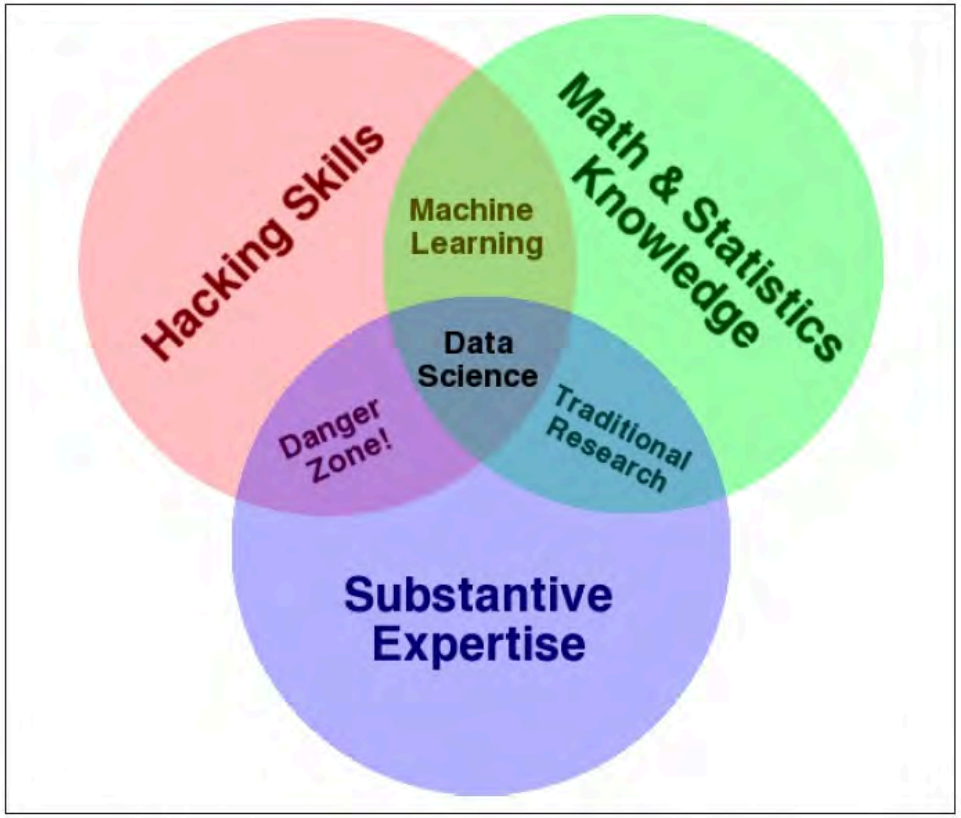
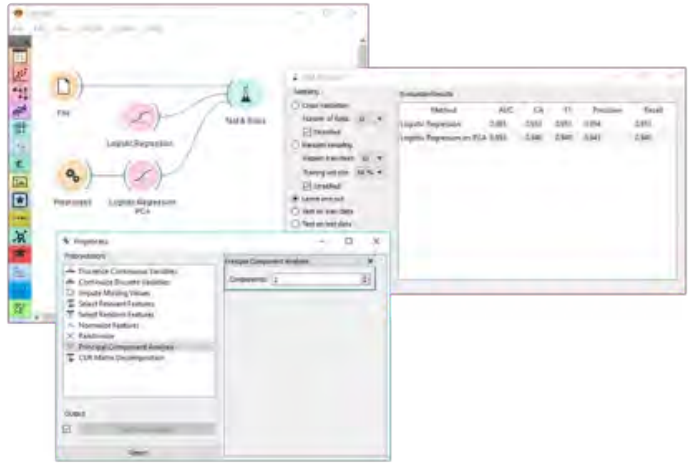


Figure 1-1. Drew Conway's Venn diagram of data science

Useful Tools for Data Science



Weka
Open-source GUI and Java library

Programming Libraries
R, Python, Java frameworks

Visual Tools
RapidMiner, Orange

Visualization
Tableau, D3.js