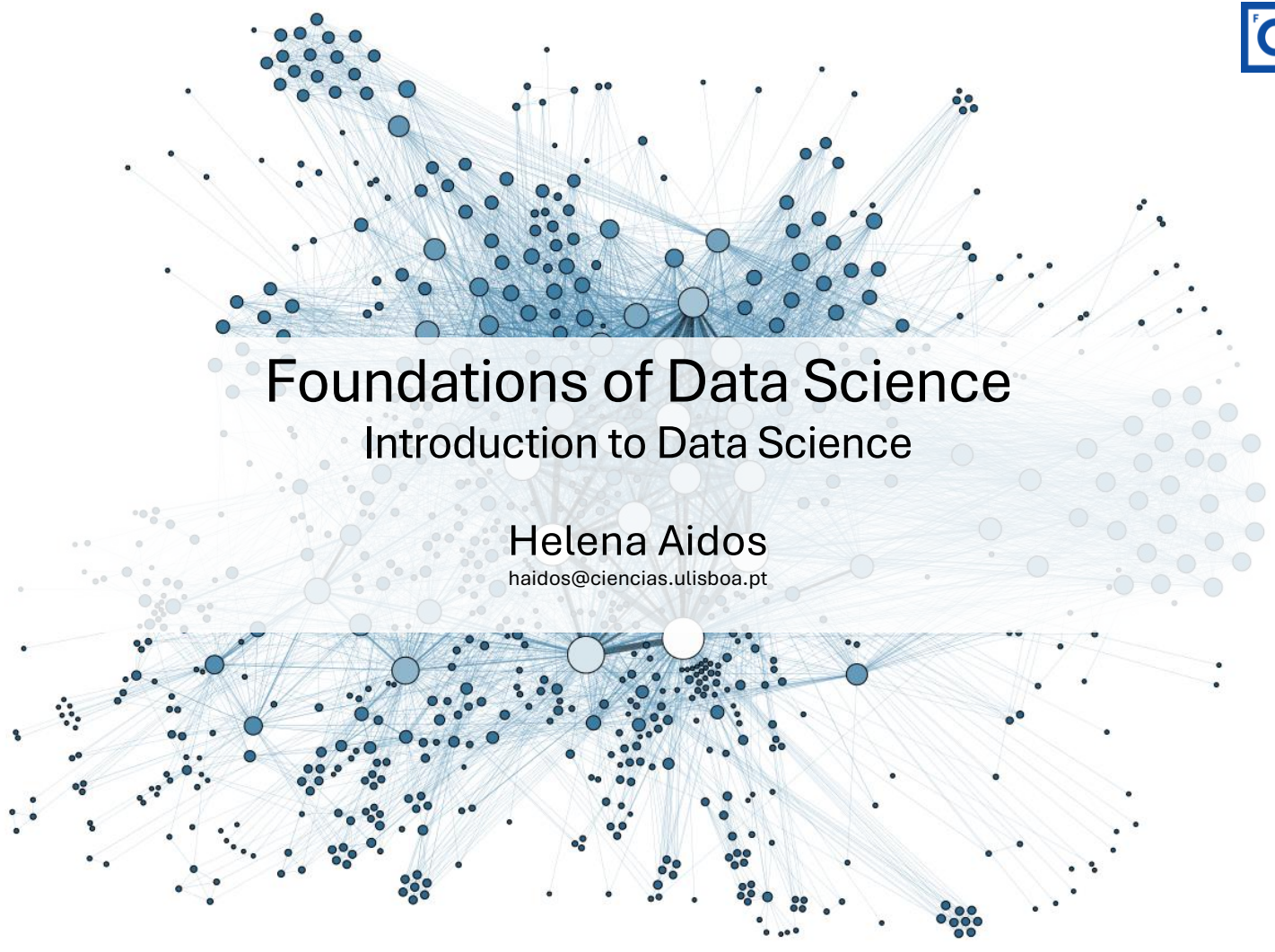


A complex network graph with numerous nodes of varying sizes and colors (blue, grey, white) connected by thin lines, forming a dense web. The graph is centered on the slide, with the title and author information overlaid on it.

# Foundations of Data Science

## Introduction to Data Science

Helena Aidos

[haidos@ciencias.ulisboa.pt](mailto:haidos@ciencias.ulisboa.pt)

# Why Data Science?

# Why Data Science?



Online  
activity

Sensors



Scientific  
experiments



**WE ARE GENERATING MORE  
DATA THAN EVER BEFORE**



Health  
records

Images &  
videos



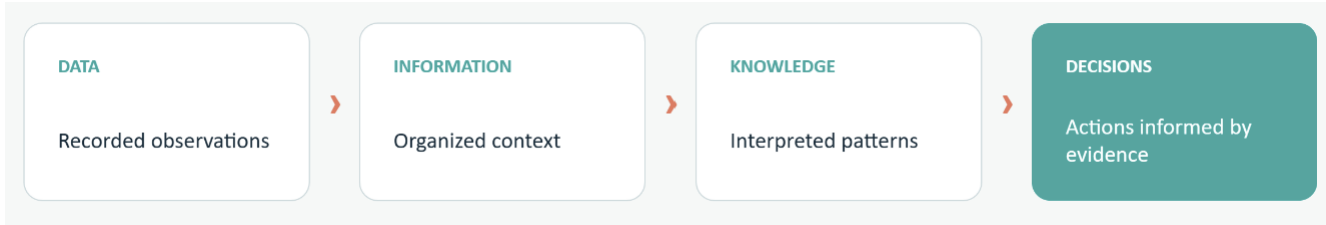
**But does data necessarily mean more knowledge?**

# From data to knowledge

*We are drowning in information  
and starving for knowledge.  
—Rutherford D. Roger*



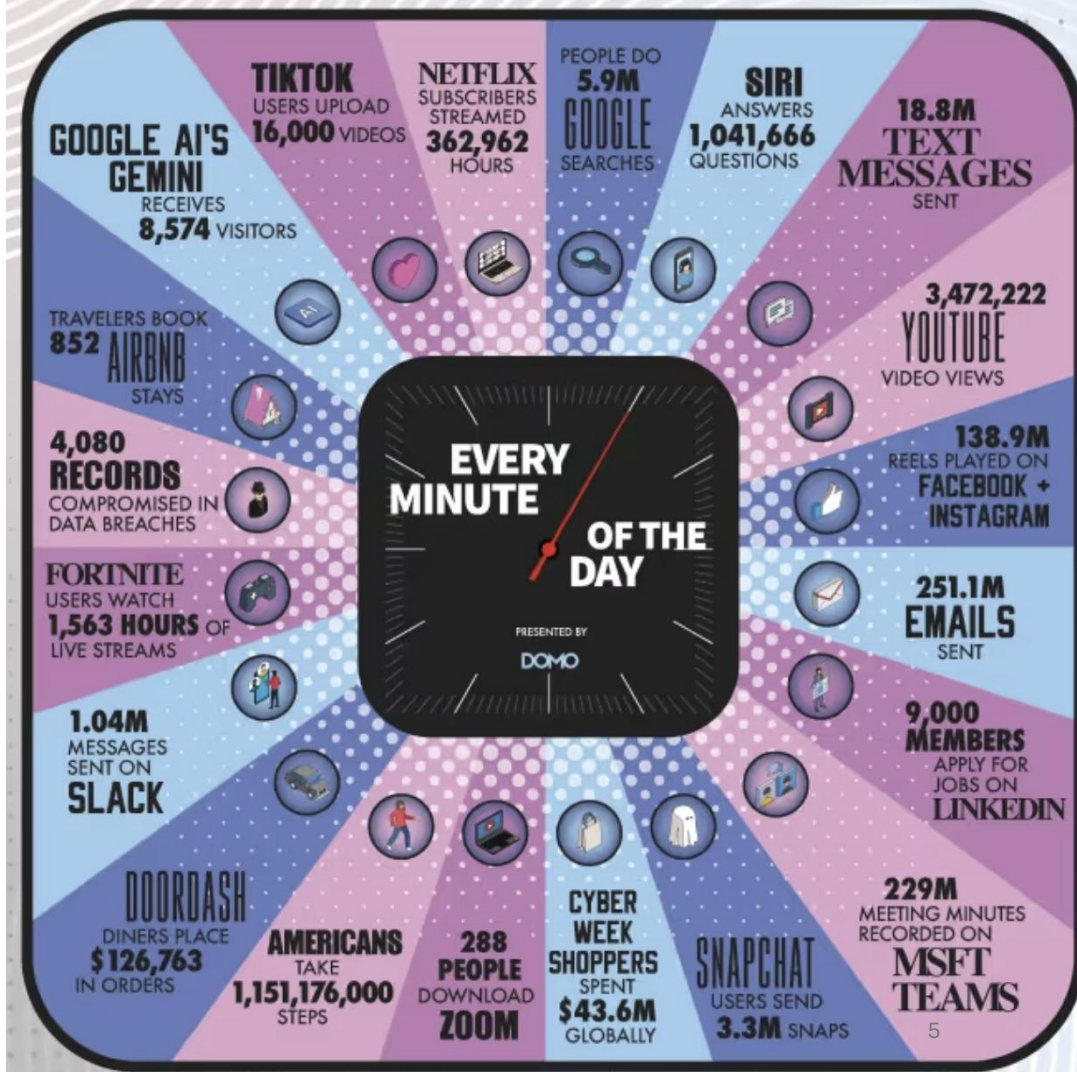
**Having data is not enough.**



The challenge is to transform increasingly large and complex data into  
**useful knowledge and informed decisions.**

# Data is everywhere

Data is generated continuously, across almost every domain.



# What can we do with data?

## **What happened?**

Understand and describe

How did energy consumption change?

## **Why did it happen?**

Discover patterns and relationships

What factors are associated with disease progression?

## **What will happen?**

Predict and forecast

Which machines are likely to fail?

## **What should we do?**

Support decisions and actions

Which intervention should we recommend?

# More data, more opportunities and more challenges

**More data expands what we can learn, but it also raises harder questions.**

## OPPORTUNITIES

-  **Understand complex phenomena**
-  Discover useful patterns
-  Anticipate future outcomes
-  Improve decisions and services

## CHALLENGES

-  **Is the data reliable and representative?**
-  Can we trust the conclusions?
-  Are decisions fair and ethical?
-  Can people use the results effectively?

**Turning data into value requires more than algorithms.**

# What is Data Science?

# What is Data Science?

Data Science is the study and practice of **extracting value from data**.

## DATA

observations and evidence

## ANALYSIS

methods for learning from  
data

## CONTEXT

meaning, constraints and  
purpose

The value does not come from the data alone. It comes from how the data is interpreted and used.

# Insights, decisions and actions

**Data Science can deepen understanding or guide what happens next.**

## INSIGHTS

|                  |                                |
|------------------|--------------------------------|
| <b>Describe</b>  | summaries and trends           |
| <b>Discover</b>  | patterns and relationships     |
| <b>Structure</b> | groups and latent organization |
| <b>Explain</b>   | hypotheses and mechanisms      |

## DECISIONS AND ACTIONS

|                  |                             |
|------------------|-----------------------------|
| <b>Predict</b>   | future outcomes             |
| <b>Classify</b>  | cases or events             |
| <b>Recommend</b> | possible choices            |
| <b>Optimize</b>  | decisions under constraints |

**Some projects aim to understand. Others aim to predict or intervene. Many do both.**

# A Data Science example

## EXAMPLE CASE STUDY:

**Which students may need additional academic support?**



**A useful model output does not guarantee a successful project.**

**What happens if the prediction is accurate but no effective support follows?**

# Data Science and related fields

The fields overlap, but their central emphasis differs.

## STATISTICS

Learning from data while accounting for variability and uncertainty

## MACHINE LEARNING

Developing models that learn patterns and generalize to new data

## ARTIFICIAL INTELLIGENCE

Building systems capable of intelligent behaviour

## DATA SCIENCE

Using data and analytical methods to generate knowledge and support decisions

A machine-learning model may predict readmission. A Data Science project also asks whether the prediction supports better care.

# Where does Data Science begin?

Which of these situations count as Data Science?

- 01 Calculate the average grade of a class
- 02 Create a dashboard of hospital admissions
- 03 Investigate factors associated with readmission
- 04 Develop a system to identify patients who may benefit from follow-up care

**There is no sharp boundary.**

Scope, purpose and context matter more than the use of a particular algorithm.

# A working definition for this course

**Data Science is an interdisciplinary process that uses data to understand phenomena, answer questions and support decisions or actions.**

01

## **A meaningful problem**

What are we trying to understand or improve?

02

## **Relevant data**

What observations can inform the problem?

03

## **Appropriate analysis**

How can we obtain reliable conclusions?

04

## **Useful outcomes**

How will the results create value, and for whom?

**What knowledge and skills does this process require?**

# Data Science as an interdisciplinary field

# Why is Data Science interdisciplinary?

**Data Science problems connect technical questions with domain needs, human decisions and societal consequences.**

## EXAMPLE:

**A university wants to identify students who may need additional support.**

### DOMAIN

What does “at risk” mean, and what support is appropriate?

### DATA

Which observations are relevant, reliable and ethically available?

### ANALYSIS

Which methods can reveal patterns or estimate risk?

### USE

**How should the university interpret and act on the results?**

**No single discipline can answer all four questions.**

# Contributions from different disciplines

**Each area contributes a different form of knowledge or capability.**

STATISTICS AND MATHEMATICS

Uncertainty, inference, modelling and validation

COMPUTER SCIENCE

Algorithms, software, databases and computation

MACHINE LEARNING AND AI

Prediction, pattern discovery and automated decisions

DATA ENGINEERING AND INFORMATION SYSTEMS

Data acquisition, integration, storage and pipelines

DOMAIN KNOWLEDGE

Meaningful questions, constraints and interpretation

SOCIAL SCIENCES, ETHICS AND LAW

Human behaviour, fairness, privacy and accountability

**Communication and visualization connect the analysis with the people who must understand and use it.**

# One problem through different lenses

## EXAMPLE:

**Which students may need additional academic support?**

### EDUCATION

Which behaviours genuinely indicate academic difficulty?

### STATISTICS

How uncertain are the estimates?

### MACHINE LEARNING

Will the model generalize to future students?

### COMPUTER SCIENCE

Can the system integrate and process the available data?

### ETHICS AND LAW

Should these data be used for this purpose?

### COMMUNICATION

How should teachers and students receive the result?

**Each perspective reveals assumptions or risks that the others may overlook.**

**Which question disappears if we treat this only as a machine-learning problem?**

# The role of domain knowledge

Domain expertise influences every stage, not only the initial question.

01

PROBLEM FORMULATION

Which question  
is meaningful?

02

DATA

Which variables  
and populations  
are relevant?

03

INTERPRETATION

Do the patterns  
make sense in  
the domain?

04

EVALUATION

Which errors and  
outcomes matter?

05

ACTION

Which intervention  
is feasible and  
appropriate?

**The same analytical method can produce useful or misleading results depending on how well the domain is understood.**

# Interdisciplinary work requires collaboration

**Data Science rarely depends on one person possessing every form of expertise.**

## DATA SCIENTISTS CONTRIBUTE

-  Analytical reasoning
-  Computational methods
-  **Integration across the project**

## OTHER SPECIALISTS CONTRIBUTE

-  **Domain expertise**
-  Engineering and infrastructure
-  Legal and ethical guidance
-  Decision-making authority
-  Communication with affected communities

**The challenge is to connect these perspectives throughout the project, rather than consult them at the end.**

Within this collaboration, what does a data scientist actually do?

# The role of a Data Scientist

# The role of a data scientist

**A data scientist uses data and analytical methods to investigate questions and support decisions.**

## **FRAME THE PROBLEM**

Translate a real-world need into questions that data can help answer.

## **WORK WITH DATA**

Understand, prepare and evaluate the available data.

## **DEVELOP AND EVALUATE**

Select appropriate methods and assess the resulting evidence.

## **COMMUNICATE AND SUPPORT USE**

**Explain the results and connect them with decisions or actions.**

**The role connects the problem, the data, the analysis and its use.**

# The Data Science lifecycle

The work progresses through connected and revisited stages.

01

## PROBLEM

Define the need, question and intended use

02

## DATA

Identify sources, provenance, limitations and relevant populations

03

## PREPARATION

Clean, integrate, transform and examine the data

04

## ANALYSIS

Apply suitable statistical, computacional or machine learning methods

05

## EVALUATION

Test validity, generalization, usefulness and potential harms

06

## COMMUNICATION

Present the findings and connect them with a decision or action

07

## MONITORING

Observe performance, consequences and changes over time

The lifecycle is iterative: Later findings often require earlier decisions to be reconsidered.

# From a real-world need to an analytical problem

## EXAMPLE:

### REAL-WORLD NEED

**Reduce student dropout**

### INITIAL REQUEST

**“Build a model to predict which students will drop out.”**

## Questions a data scientist must ask

What counts as dropout?

How early must the prediction be available?

Which errors have the greatest consequences?

Which students and time period matter?

**What action will follow the prediction?**

How will we know whether the project helped?

**A technically solvable problem may still be the wrong problem.**

# Understanding the data

**Data preparation requires decisions about meaning, not only technical cleaning.**

## PROVENANCE

Where did the data come from, and why was it collected?

## QUALITY

What is missing, noisy, inconsistent or unreliable?

## REPRESENTATION

Which people, situations or periods does the dataset represent?

## MEANING

**What do the variables measure, and what might they fail to capture?**

# Analysis, modeling and evaluation

**Modelling is one part of the analytical work.**

01

## **ESTABLISH A BASELINE**

What can a simple approach achieve?

02

## **SELECT APPROPRIATE METHODS**

Does the methods fit the question, data and constraints?

03

## **COMPARE ALTERNATIVES**

Which approach performs reliably under relevant conditions?

04

## **ANALYZE ERRORS AND UNCERTAINTY**

When does the method fail, and how confident are its outputs?

05

## **EVALUATE USEFULNESS**

Does the result support the intended decision?

**The most complex model is not necessarily the most useful model.**

Would you accept slightly lower predictive performance for a model that people can understand and use appropriately?

# From analytical to real-world use

## EXAMPLE:

### MODEL OUTPUT

Estimated risk of dropout

### INTERPRETATION

Identify factors and uncertainty associated with the estimate

### DECISION

Determine whether additional support should be offered

### ACTION

Contact the student and provide appropriate support

### OUTCOME

Improve the student's academic experience and progress

## After implementation



Monitor data and performance



Examine unintended consequences



Collect feedback from users and affected people



**Update or withdraw the system when necessary**

**A model can perform well while the overall intervention fails.**

# The role depends on the project and organization

**There is no single job description for a data scientist.**

## RESEARCH

Formulating questions, analyzing evidence and scientific communication

## SMALL ORGANIZATION

Working across most stages of the lifecycle

## LARGE TECHNOLOGY TEAM

Experimentation, modeling and collaboration with specialist engineers

## PUBLIC SECTOR OR HEALTHCARE

**Validity, accountability and domain constraints and consequences**

## CONSULTING

Problem formulation, rapid analysis and client communication

**How does Data Science work change across application domains?**

# Applications across domains

# Data Science across domains

Different domains use data to understand phenomena, anticipate outcomes and support decisions.

## HEALTHCARE

Which patients may need additional care?

## EDUCATION

Which students may benefit from support?

## TRANSPORTATION

Where and when will congestion occur?

## RETAIL

Which products may be relevant to a customer?

## INDUSTRY

Which machines are at risk of failure?

## FINANCE

Which transactions require investigation?

What is common to all six examples?

# Domain, data and intended use

| DOMAIN         | EXAMPLE DATA                      | POSSIBLE OUTPUT      | INTENDED USE                |
|----------------|-----------------------------------|----------------------|-----------------------------|
| Healthcare     | Clinical records and tests        | Risk estimate        | Plan follow-up care         |
| Education      | Grades, attendance and activity   | Student risk profile | Offer academic support      |
| Transportation | GPS, sensors and weather          | Congestion forecast  | Adapt traffic management    |
| Retail         | Browsing and purchases            | Product ranking      | Personalize recommendations |
| Industry       | Sensors and maintenance records   | Failure probability  | Schedule maintenance        |
| Finance        | Transactions and account activity | Fraud score          | Prioritize investigation    |

**A similar analytical output may support very different actions.**

# Common tasks across different domains

## DESCRIPTION

**What happened?**

---

Hospital activity  
Student participation  
Energy consumption

## PREDICTION AND FORECASTING

**What may happen next?**

---

Disease progression  
Machine failure  
Traffic congestion

## CLASSIFICATION AND DETECTION

**Which cases require attention?**

---

Fraudulent transactions  
Abnormal sensor readings  
Harmful content

## RECOMMENDATION AND OPTIMIZATION

**What should be considered?**

---

Products or treatments  
Transport routes  
Allocation of resources

**Methods transfer across domains, but their interpretation and use remain domain-specific.**

# Prediction has different consequences

The same technical error can have very different practical effects.

## APPLICATION

## FALSE POSITIVE

## FALSE NEGATIVE

**Hospital readmission**

Unnecessary follow-up

A vulnerable patient receives no additional care

**Student dropout**

A student receives unnecessary attention

A struggling student receives no support

**Machine failure**

Unnecessary maintenance

An unexpected breakdown occurs

**Fraud detection**

A legitimate transaction is blocked

**Fraud remains undetected**

**Which error is more costly?**

The answer depends on the domain, stakeholders and available actions.

# Context changes the Data Science problem

The same method may require different assumptions, evidence and safeguards.

01

DATA

Availability, quality,  
scale and sensitivity

02

TIME

Real-time decisions,  
short forecasts or  
long-term outcomes

03

ERRORS

Different costs and  
consequences

04

STAKEHOLDERS

Different people  
benefit, decide or  
experience harm

05

CONSTRAINTS

Regulation, resources,  
infrastructure and  
professional practice

**A method that works in one domain cannot simply be transferred without reconsidering its assumptions and consequences.**

**What can prevent a Data Science project from producing reliable and beneficial results?**

# Challenges in Data Science

# Challenges in Data Science

A project can fail even when its model performs well.

## DATA

- Quality and missingness
- Representativeness
- Access and integration
- Provenance and change

## ANALYSIS AND MODELS

- Validity and generalization
- Uncertainty
- Interpretability
- Changing relationships

## PEOPLE AND SOCIETY

- Privacy
- Fairness
- Accountability
- Misuse and unintended effects

## PRACTICE

- Reproducibility
- Communication
- Deployment and monitoring
- Organizational adoption

Challenges arise throughout the lifecycle and often interact.

# Is the data suitable for the question?

## QUALITY

Are values missing, noisy, inconsistent or incorrectly recorded?

## REPRESENTATION

Does the dataset reflect the population and situations where results will be used?

## MEASUREMENT

Do variables measure the concepts that we care about?

## PROVENANCE

Why, when and how was the data collected?

## Example

**Platform activity records behaviour within one system. It does not directly measure motivation, effort or learning.**

**More data does not automatically mean better evidence.**

# Will the conclusions remain valid?

01

## GENERALIZATION

Will the result apply to new people, locations or time periods?

02

## UNCERTAINTY

How confident should we be in the estimate?

03

## ROBUSTNESS

Does the result change under small variations in data or assumptions?

04

## INTERPRETABILITY

Can people understand its basis and limitations?

05

## CHANGE OVER TIME

What happens when behaviour or data-generating processes change?

### Example

**A dropout model trained during online teaching may behave differently after students return to primarily in-person teaching.**

**Good performance on historical data does not guarantee reliable future use.**

# Who benefits, and who may be harmed?

## PRIVACY

Should these data be collected, combined or used for this purpose?

## FAIRNESS

Do errors or benefits affect groups differently?

## ACCOUNTABILITY

Who is responsible for the decision and its consequences?

## AUTONOMY AND TRANSPARENCY

Do affected people know how data influences decisions about them?

## Example

**If students know every platform interaction contributes to a risk profile, could this affect their behaviour or trust?**

**Technical validity does not determine whether a use of data is acceptable.**

# Can the result work in practice?

01

## REPRODUCIBILITY

Can another person obtain and verify the same result?

02

## COMMUNICATION

Can decision-makers understand the evidence and limitations?

03

## INTEGRATION

Does the analysis fit existing systems and workflows?

04

## MONITORING

Will declining performance or unintended effects be detected?

05

## ADOPTION

Do intended users trust the result and find it useful?

### Example

**A student-risk model has little value if the output arrives too late, nobody knows how to interpret it, or the university cannot provide support.**

**A useful analysis must fit the technical and organizational environment in which it will operate.**

# One problem can reveal several challenges

## EXAMPLE:

### Student support system

#### DATA

Platform activity provides an incomplete view of engagement.

#### MODEL

Performance may change across courses or academic years.

#### DECISION

The university has limited capacity to contact students.

#### SOCIETY

Monitoring may affect privacy, autonomy and trust.

#### OUTCOME

Contact does not guarantee that the offered support is effective.

**Fixing one component does not guarantee that the project will succeed.**

# Questions for responsible Data Science

01 PROBLEM

Is this the right question?

02 DATA

Is the evidence suitable and representative?

03 ANALYSIS

Are the methods and conclusions reliable?

04 EVALUATION

Does assessment reflect real use and consequences?

05 COMMUNICATION

Are uncertainty and limitations clear?

06 USE

Who decides, benefits or may experience harm?

07 MONITORING

What could change after implementation?

**Reliable Data Science requires critical questions, not only better tools.**

**How are new technologies changing the available tools and the work of Data Science?**

# Emerging trends

# Emerging trends in Data Science

## CURRENT DIRECTIONS IN 2026

**The tools and scale of Data Science are changing rapidly.**

### FOUNDATION AND MULTIMODAL MODELS

Models increasingly work across text, images, audio, structured data and scientific measurements.

### AI-ASSISTED WORKFLOWS

Generative models and agents increasingly support coding, exploration, visualization and modelling.

### SCIENTIFIC DISCOVERY

AI systems increasingly help researchers find patterns, formulate hypotheses and design experiments.

### RESPONSIBLE AND EFFICIENT SYSTEMS

Evaluation, governance, resource use and monitoring receive greater attention as systems become more capable.

**The field is moving from isolated models toward integrated analytical systems.**

# Models that work across data types

CURRENT DIRECTIONS IN 2026

Foundation and multimodal models broaden what one system can process.

## EARLIER PATTERN

**A specialized model for one dataset and one task**

Separate pipelines for text, images, tables and signals

## EMERGING PATTERN

**A broadly trained model adapted to several tasks and data types**



Text and documents



Images and video



Speech and audio



Tables and time series



Scientific measurements

More flexible models create new possibilities, but also make evaluation and control more difficult.

# AI systems increasingly participate in the workflow

## CURRENT DIRECTIONS IN 2026

### ASSISTANT

Suggests code or explanations while a person directs each step.

### COLLABORATOR

Plans and performs connected tasks with human review.

### AGENT

**Uses tools and executes longer workflows with greater autonomy.**

## Possible contributions



Plan an analysis



Inspect and transform data



Generate and execute code



Compare models



Create visualizations



**Document decisions**

**Automation changes the data scientist's work, but it does not remove the need for judgment.**

# Data Science as a partner in scientific discovery

CURRENT DIRECTIONS IN 2026

01

## OBSERVE

Integrate large and heterogeneous datasets

02

## DISCOVER

Find structures and candidate relationships

03

## HYPOTHESIZE

Propose explanations or experimental directions

04

## TEST

Design experiments and analyse evidence

05

## REFINE

Update models as new evidence appears

### Examples

Medicine and biology

Climate and Earth observation

Materials and engineering

Social and behavioural science

**Data Science increasingly supports the full cycle of discovery, not only the final analysis.**

# Capability increases the need for responsibility

## CURRENT DIRECTIONS IN 2026

### NEW POSSIBILITIES

- Analyse larger and more diverse datasets
- Automate repetitive analytical work
- Accelerate experimentation
- Make analytical tools accessible to more people

### CONTINUING RESPONSIBILITIES

- Verify data and generated outputs
- Evaluate across populations and contexts
- Protect privacy and sensitive information
- Document assumptions and human oversight
- Monitor performance and unintended effects

**Powerful tools make critical evaluation more important.**

# What remains fundamental?

**Data Science will continue to change, but its central questions remain stable.**

- |    |             |                                     |
|----|-------------|-------------------------------------|
| 01 | PROBLEM     | What are we trying to solve?        |
| 02 | DATA        | Are the data suitable?              |
| 03 | ANALYSIS    | What can the evidence support?      |
| 04 | RELIABILITY | How reliable are the conclusions?   |
| 05 | USE         | Who will use the result?            |
| 06 | IMPACT      | Who may benefit or experience harm? |
| 07 | AFTERWARD   | What happens after implementation?  |

**Data Science turns data into knowledge and responsible action.**

# Reading

Data Science in Context: Foundations, Challenges, Opportunities  
Alfred Z. Spector, Peter Norvig, Chris Wiggins, Jeannette M. Wing

Read Section 1.1

<https://datascienceincontext.com/manuscript/>