Data Science and Beyond:
Complex data types & Artificial Intelligence

Traditional data science

Traditional data science refers to working with structured datasets that are organized in **rows** and **columns**, similar to how data is presented in spreadsheets or databases.

ID	Age	Job	Education	Loan
1	42	IT	College	Yes
2	38	HR	College	Yes
3	26	BD	Secondary	No
4	39	HR	High school	No
5	55	SE	College	Yes

A traditional tabular dataset.

Introduction to complex data types

Complex data types go beyond traditional tabular data and include unstructured data like images, text, and audio.

Examples:
- Images (e.g., photos, medical scans)
- Text (e.g., documents, reviews)
- Audio (e.g., speech, music)

Complex data types matter because:
- With the rise of **multimedia** platforms, vast amounts of unstructured data are generated.
- These data types unlock **new insights** in fields like healthcare (e.g., medical imaging), customer sentiment (e.g., text reviews), and voice recognition.

Image data

Image data refers to visual information that is stored in a digital format, typically represented as an array of pixels.

A dataset for animal classification
Reference: https://dm.kaist.ac.kr/datasets/animal-10n/

Text data

Text data refers to any information stored in a textual format, typically consisting of characters, words, sentences, or paragraphs.

Reviews	Label
This was an absolutely **terrible movie**. Don't be lured in by Christopher Walken or Michael Ironside. Both are great actors, but this must simply be their **worst role** in history. Even their great acting could not redeem this movie's **ridiculous storyline**. This movie is an early nineties US propaganda piece. The most pathetic scenes were those when the Columbian rebels were making their cases for revolutions. Maria Conchita Alonso appeared phony, and her pseudo-love affair with Walken was nothing but a pathetic emotional plug in a movie that was devoid of any real meaning. I am disappointed that there are movies like this, ruining actor's like Christopher Walken's good name. I could barely sit through it.	0
I have been known to **fall asleep** during films, but this is usually due to a combination of things including, really tired, being warm and comfortable on the sette and having just eaten a lot. However on this occasion I **fell asleep** because the film was **rubbish**. The plot development was constant. Constantly **slow and boring**. Things seemed to happen, but with no explanation of what was causing them or why. I admit, I may have missed part of the film, but i watched the majority of it and everything just seemed to happen of its own accord without any real concern for anything else. I **cant recommend** this film at all.	0
This is the kind of film for a snowy Sunday afternoon when the rest of the world can go ahead with its own business as you descend into a big arm-chair and mellow for a couple of hours. **Wonderful performances** from Cher and Nicolas Cage (as always) gently row the plot along. There are no rapids to cross, no dangerous waters, just a warm and witty paddle through New York life at its best. A family film in every sense and one that **deserves the praise** it received.	1
As others have mentioned, all the women that go nude in this film are mostly absolutely **gorgeous**. The plot very ably shows the hypocrisy of the female libido. When men are around they want to be pursued, but when no "men" are around, they become the pursuers of a 14 year old boy. And the boy becomes a man really fast (we should all be so lucky at this age!). He then gets up the courage to pursue his true love.	1

A dataset for movie review sentiment analysis.

Question: Can you think of an application where both *image* and *text* data are involved?

Audio data

Audio data refers to sound recordings stored in digital format. It captures variations in sound waves over time. Audio data can include speech, music, environmental sounds, and more.

Here's a text version of my voice - because even my words need subtitles to keep up with how fast I talk!

Speech-to-text recognition.
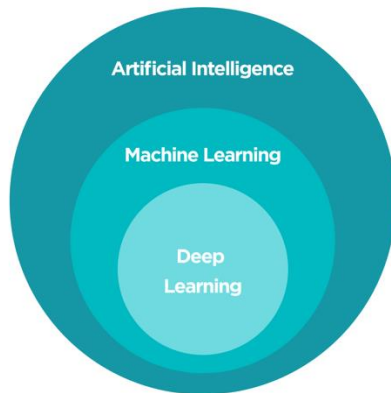**Question**: How does the training set of speech-to-text application look like?

Artificial Intelligence (AI)

Artificial Intelligence (AI): a field of computer science enabling machines to imitate human intelligence.

**Machine Learning** (ML): a branch of AI providing machines ability to learn with data.

Deep Learning (DL): subfield of ML exploiting Deep Neural Network (DNN).

AI and complex data types

AI can handle complex data types in:

- **Computer Vision** (CV): involves enabling machines to interpret and understand *visual data* from images or videos.
- **Natural Language Processing** (NLP): focuses on enabling machines to understand, interpret, and generate *human language*.
- **Speech and Sound Processing**: processes *audio data* to recognize patterns, convert speech to text, and generate sound.
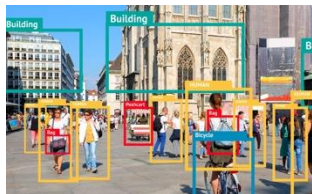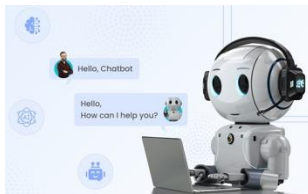
Computer Vision (CV)   Natural Language Processing (NLP)   Speech and Sound Processing

Question: Could you name other applications using image, text and audio data?

AI applications

Healthcare

Automobile

Surveillance

Banking, Finance

Robotics

Social Media

Entertainment

E-commerce

Automobile
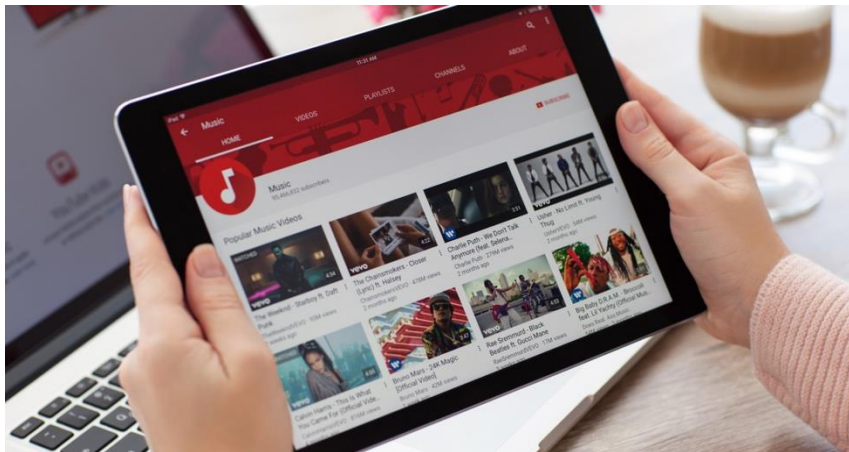
- Autonomous driving (a.k.a. self-driving) cars: avoid obstacles, communicate with near-by cars, etc., by applying Machine Vision and IoT technologies.

Entertainment

- User data collection and analysis.
- Recommendation engine.

Stock price prediction
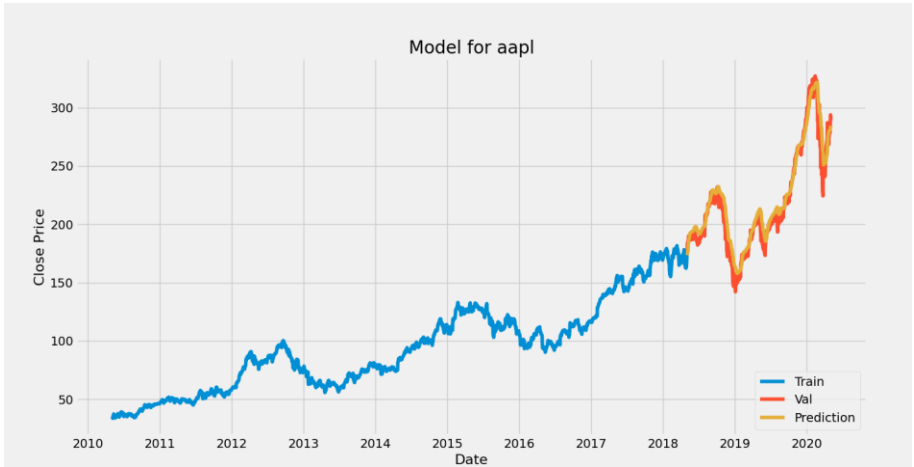
Model for aapl

Data generation

- AI generates data samples similar to what it has seen (e.g., GAN).

Data generation as Deep fake

Computer Vision - Image data

Image data is typically represented as a grid of pixels, where each pixel holds color or intensity information.

- Each pixel contains three values representing the intensity of red, green, and blue colors (**RGB**). Each value is between 0 and 255.
- Images are usually defined by width, height, and color channels (e.g., 32x32x3 for an RGB image.

Question:
What is the size, in bits, of a 32x32 RGB image with 3 color channels?
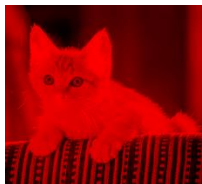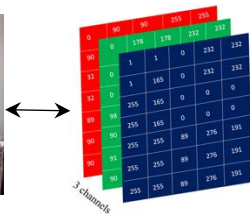
Computer Vision - Image classification

label = 5	label = 0	label = 4	label = 1	label = 9
label = 2	label = 1	label = 3	label = 1	label = 4
label = 3	label = 5	label = 3	label = 6	label = 1
label = 7	label = 2	label = 8	label = 6	label = 9
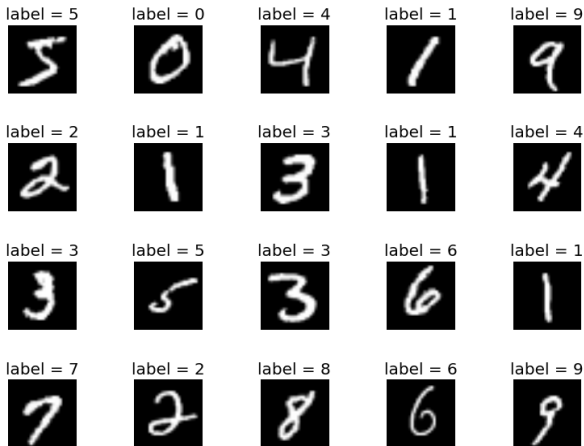
MNIST dataset for handwritten digits recognition.
Question: What are input features and output labels of this dataset?

Computer Vision - Image data

Data features are gray-scale images and labels are corresponding digits.
Question: What type of learning is it?

$$h_\theta(x)$$

$$h_{\theta^*}(x)$$

7   2   8   6   9   9

Training   Prediction
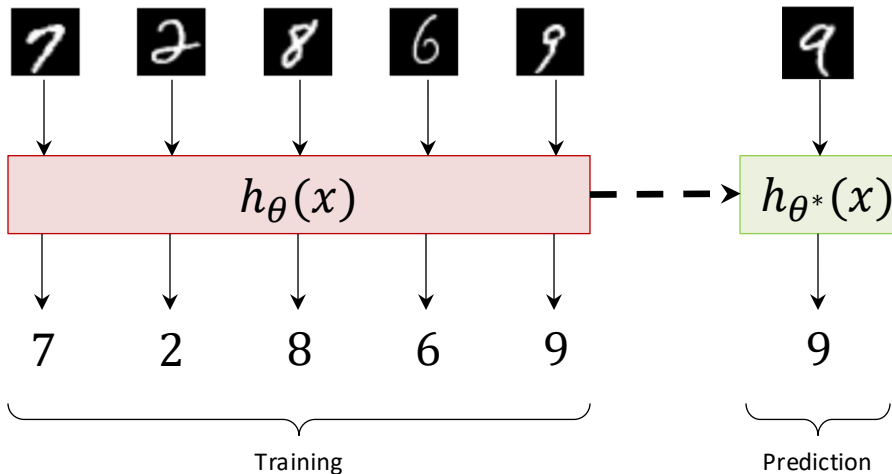
Applications of AI

AI applications

Healthcare	Education
Automobile	Marketing
Banking, Finance	Gaming
Surveillance	Robotics
Social Media	Government
Entertainment	E-commerce

Healthcare

- Wearable devices that can analyze sleep patterns, calories burned, heart rate.
- Medical image diagnosis.

Automobile

- Autonomous driving (a.k.a. self-driving) cars: avoid obstacles, communicate with near-by cars, etc., by applying Machine Vision and IoT technologies.
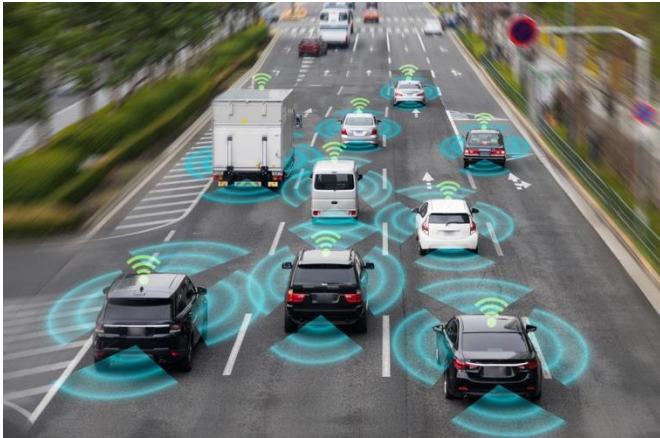
Banking and Finance

- Credit scoring.
- Automatic customer support.
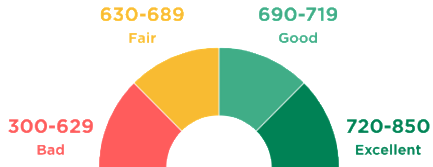
Surveillance

- Facial recognition.
- Public safely.
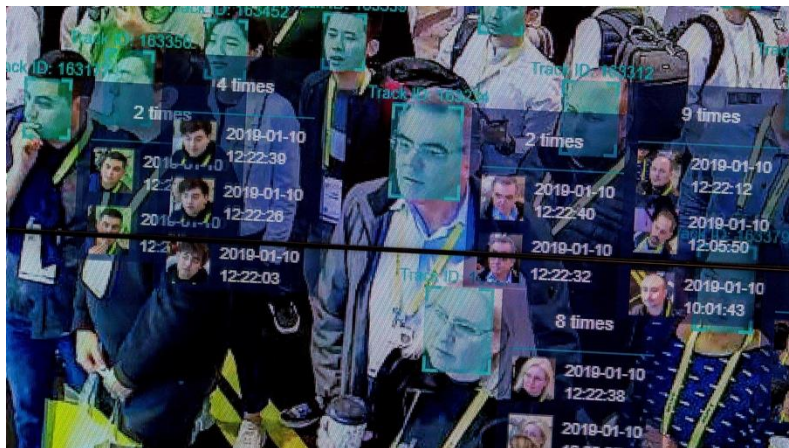
Social media

- Social trends detection (social listening).
- User behavior understanding (for marketing/advertising purpose).

Entertainment

- User data collection and analysis.
- Recommendation engine.
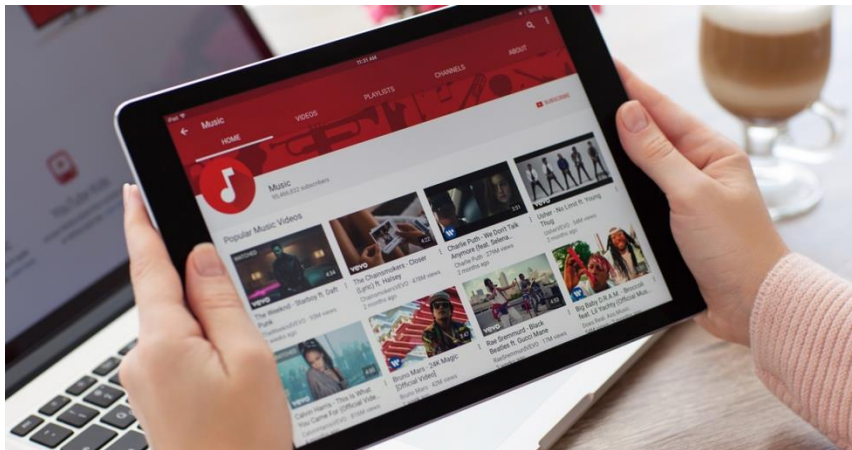
Education

- AI-powered tutoring.
- Personalized education.

Marketing

AI can analyze customer behavior, segment audiences, create campaigns, optimize pricing, and increase conversions.

Geographic
Location-based, such as from IP address or user-provided address

Demographic
Age, marital status, sex, education level, occupation, income, religion, and nationality

Psychographic
Lifestyles, personalities, interests, beliefs, attitudes, and values

Behavioral
Online shopping habits, loyalty, usage frequency, and website actions

Gaming

- Reinforcement Learning

Robotics

- Automate the works in factory.

Agriculture

- Use of drone for spraying insecticide.
- Detect weed formation in large farms.

E-commerce

- Logistics.
- Demand forecasting.
- Intelligent marketing

VerkoopHoeveelheid per YearWeek en Type
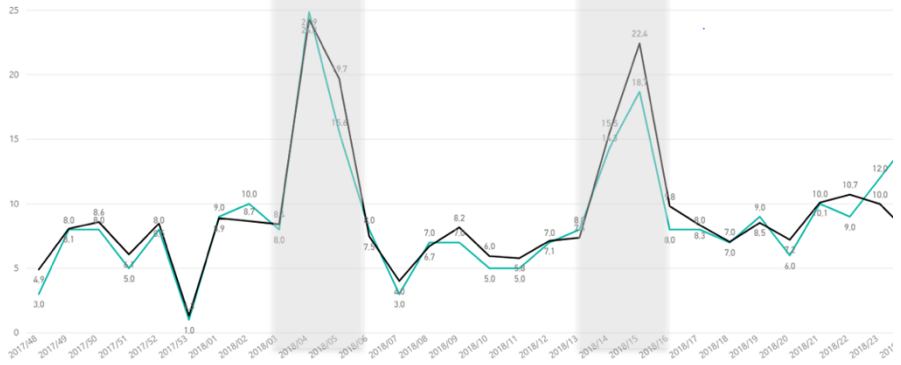
Real examples of AI

Object detection & classification
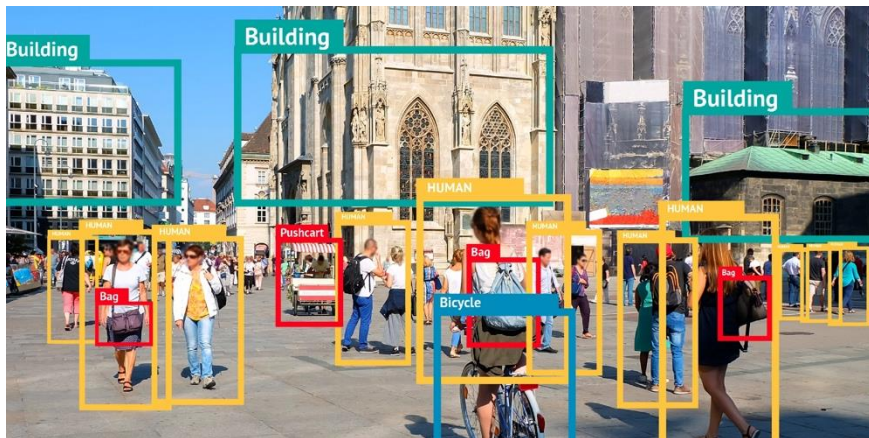
Object segmentation

- Better understanding of relation and interaction between objects.

Data generation

- AI generates data samples similar to what it has seen (e.g., GAN).

Data generation

- AI can combine features of data samples to generate new ones.

StyleGAN - NVDIA
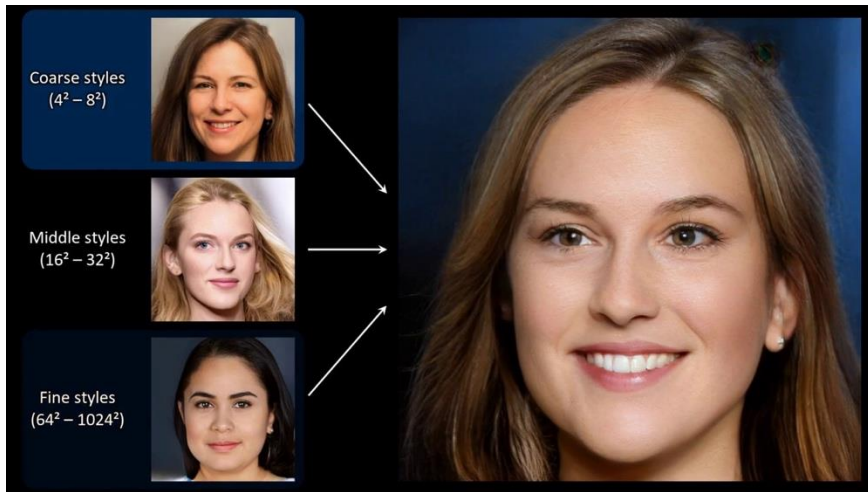
Image caption generation

- AI can generate caption for photos (combination of CV and NLP).

Only G: a car is driving down a street.

G+SA: a car is driving down a street with a traffic light.

G+SA+OA: a car is driving down a street with a group of people in the background.

(a)
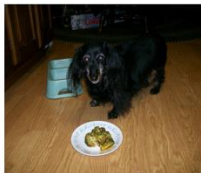
Only G: a black dog sitting on the floor.

G+SA: a black dog sitting on the floor with a plate of food.

G+SA+OA: a black dog sitting on a wooden floor next to a plate of food.

(b)

**Only G:** a passenger jet sitting on the ground.

G+SA: a large passenger jet sitting on top of an airport tarmac.

G+SA+OA: a large passenger jet sitting on top of an airport tarmac next to a man.

(c)

Only G: a fire hydrant sitting on a sidewalk.

G+SA: a fire hydrant sitting in the middle of a sidewalk.

G+SA+OA: a yellow fire hydrant sitting on the side of a road.

(d)

G: global image feature; **SA**: spatial attention; **OA**: object attention.

AI & Data Science in production

AI project management

We need to manage AI projects because they are different from traditional projects and pose unique challenges and opportunities:

- AI projects are unpredictable and constantly **evolving**. They involve handling data that frequently changes and applying algorithms that **continuously learn** and adapt.
- AI projects are complex and **interdisciplinary**. AI projects require a combination of technical, business, and domain expertise, as well as **collaboration** among various stakeholders, such as data scientists, engineers, analysts, managers, customers, etc.

AI project management

```
┌──────────────┐   ┌──────────────┐   ┌──────────────┐   ┌──────────────┐   ┌──────────────┐
│ Define project│──▶│  Acquire &   │──▶│  Model data  │──▶│ Interpret &  │──▶│ Implement,   │
│  objectives   │   │ explore data │   │              │   │ communicate  │   │ document &   │
│              │   │              │   │              │   │              │   │  maintain    │
└──────────────┘   └──────────────┘   └──────────────┘   └──────────────┘   └──────────────┘
```
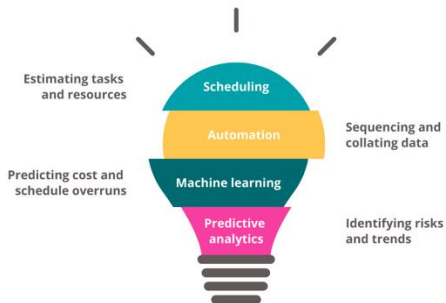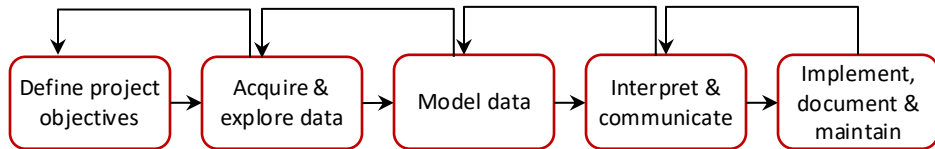
AI lifecycle is the **iterative** process of moving from a business problem to an AI solution that solves the problem. Each of the steps in the life cycle is **revisited many times** throughout the design, development, and deployment phases.

AI project management
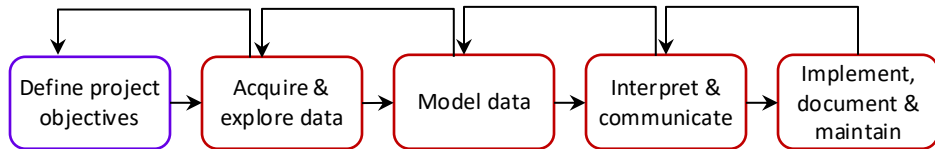
Define project objectives
- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

AI project management

Define project objectives
- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

Acquire and explore data
- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
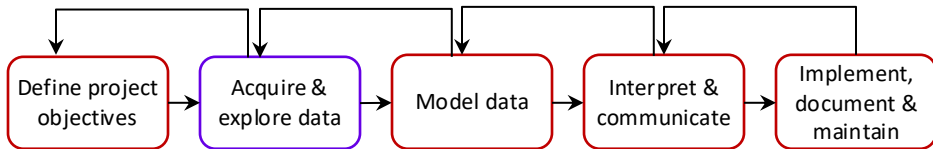- Feature engineering

AI project management

Define project objectives
- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

Acquire and explore data
- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Model data
- Variable selection
- Build candidate model
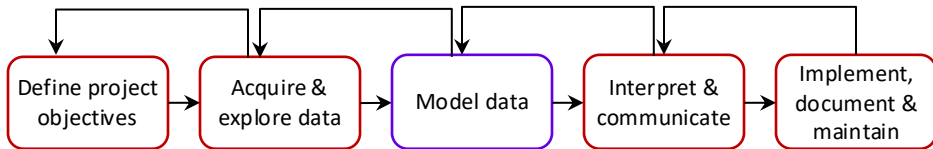- Model validation and selection

AI project management

Define project objectives
- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

Acquire and explore data
- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Model data
- Variable selection
- Build candidate model
- Model validation and selection

Interpret and communicate
- Interpret model
- Communicate model insights

AI project management

Define project objectives
- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
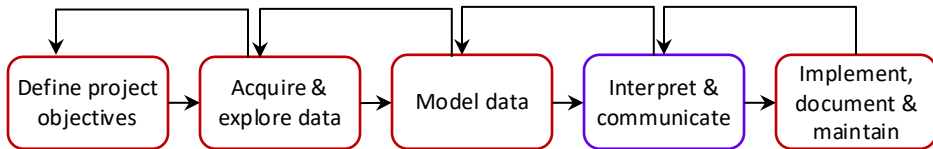- Consider risks and success criteria
- Decide whether to continue

Acquire and explore data
- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
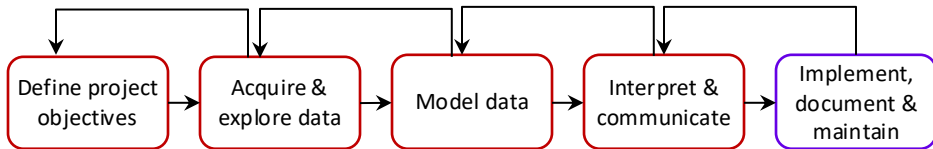- Feature engineering

Model data
- Variable selection
- Build candidate model
- Model validation and selection

Interpret and communicate
- Interpret model
- Communicate model insights

Implement, document and maintain
- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance phase

Data collection & annotation

- Data collection and annotation require much time and effort.
- Domain-expertise is necessary for labeling the data.
- **How much data is enough?**
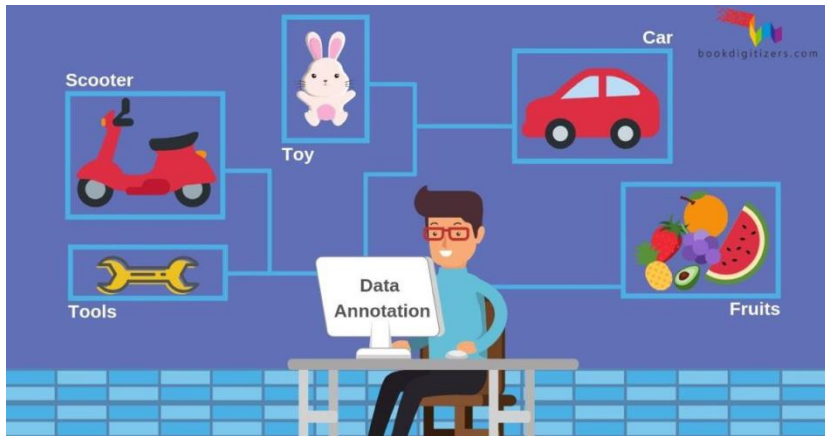
Model training & validation

- Train the model
- Validate the model to find the best set of hyper-parameters.
- Test the model for accuracy evaluation.
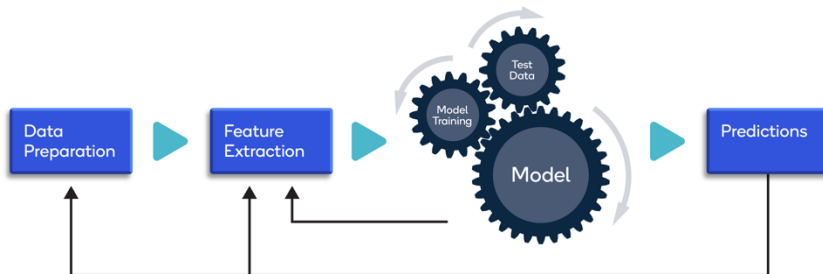
Deployment

- Deployment on-prem(ises): in-house deployment
- Deployment on **cloud infrastructure**: GCP, AWS, Azure, etc.
- Serving the models using Tensorflow Serving, Model Serving in PyTorch, TensorRT, Multi Model Serving (MMS), and using **REST API** or **GRPC** for calling.
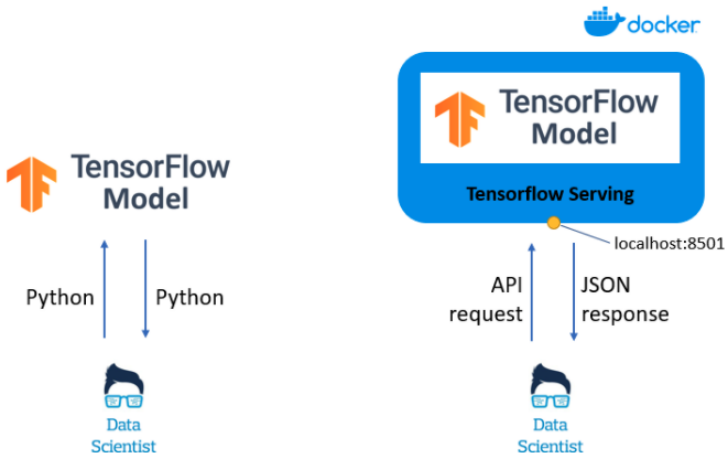
Deployment on Mobile Device

- Deep neural network compression:
  - Knowledge distillation.
  - Network quantization.
  - Neural Architecture Search (NAS).
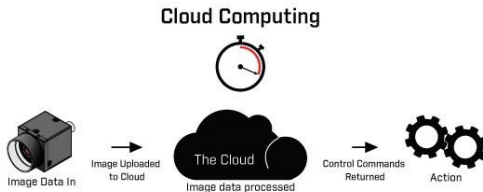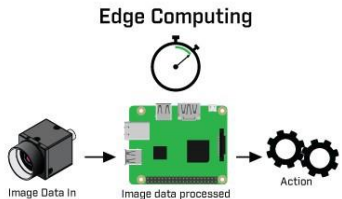
Data privacy and security

Data is confidential and data privacy is a big concern.
- Data encryption and secured data transfer (client <-> prediction server).
- Data anonymization (i.e., hide all sensitive information).

Question: Is it possible to train a deep learning model without touching raw data?

Data Privacy

Focus on how to **collect, process, share, store** and **delete** data.

Data Security

Focus on how to **prevent unauthorized access** to data.

Careers in AI and Data Science

Roles in AI and Data Science

- Business Analyst (BA)
- Data Analyst (DA)
- AI project manager
- AI and Data scientist
- AI and Data engineer
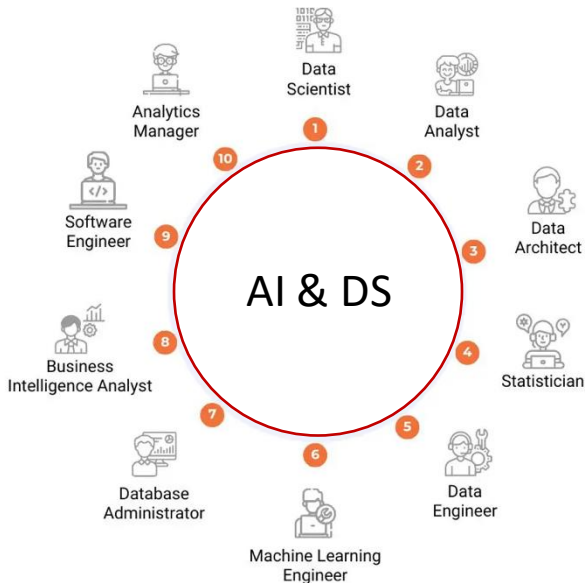- AI and Data Quality Control
- ...

Business Analyst (BA)

A business analyst is a professional who **analyzes businesses** and helps them improve their efficiency, productivity, and profitability.

Some of the responsibilities of a business analyst are:
- Using SQL and Excel to analyze large data sets
- Compiling charts, tables, and other elements of **data visualization**
- Understanding **business strategies**, goals, and requirements
- Creating financial models to support **business decisions**

Some of the skills and qualifications of a business analyst are:
- Technical writing to document processes, requirements, and specifications
- Programming languages such as SQL, Python, R, or Java to manipulate and **analyze data**
- **Database** proficiency to store, query, and manage data
- **Statistics** and **probability** to apply mathematical methods and models to data
- **Communication** to present and explain findings and recommendations to stakeholders
- Subject matter expertise to understand the **business domain** and context

Business Analyst (BA)
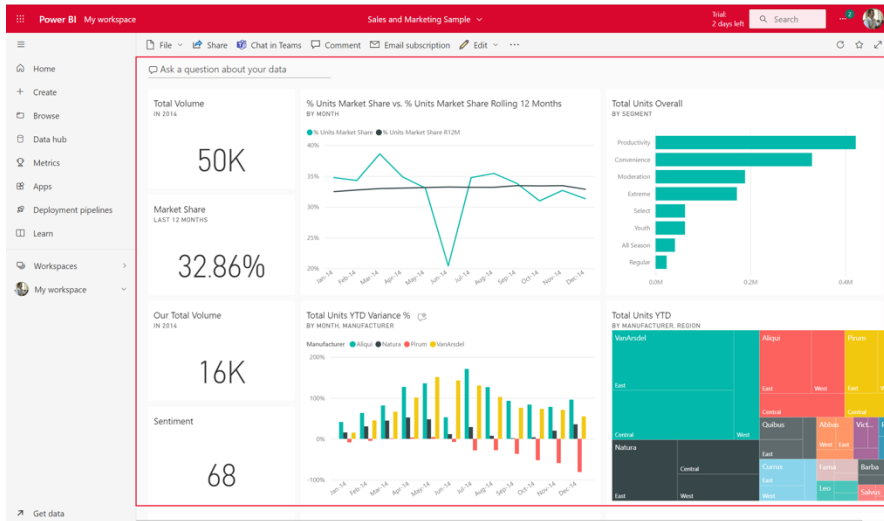
A Business Intelligence (BI) dashboard.

Data Analyst (DA)

A data analyst is a person who **collects**, **cleans**, and **interprets datasets** in order to answer a question or solve a problem.

Some of the responsibilities of a data analyst are:
- Collecting data from a variety of sources, such as **databases**, surveys and web analytics
- Cleaning the data to **remove errors** or inconsistencies, ensuring data quality
- Organizing the data in the **proper formats** for analysis
- Analyzing the data using **technical methods** to draw insights and conclusions
- Interpreting and presenting the data to **audiences**, such as executives or stakeholders.
- Making **recommendations** on how a business can improve its operations based on data insights.

Some of the skills and qualifications of a data analyst are:
- **Data collection** and manipulation skills, such as SQL, Excel, Python, R, or Java
- Data analysis and **visualization** skills, such as statistics, machine learning, text mining, predictive analytics, charts, tables, and graphs
- **Data reporting** and communication skills, such as writing, speaking, presenting, and storytelling
- Attention to detail and **quality control** skills, such as checking data accuracy and validity

SQL database query.

Business Analyst (BA) vs. Data Analyst (DA)

Business Analyst (BA)	Data Analyst (DA)
Analyze data	Gather data
Analyze business	Clean data
Communicate findings	Analyze data
Recommend strategies	Compile report
Business-focused	**Data-focused**
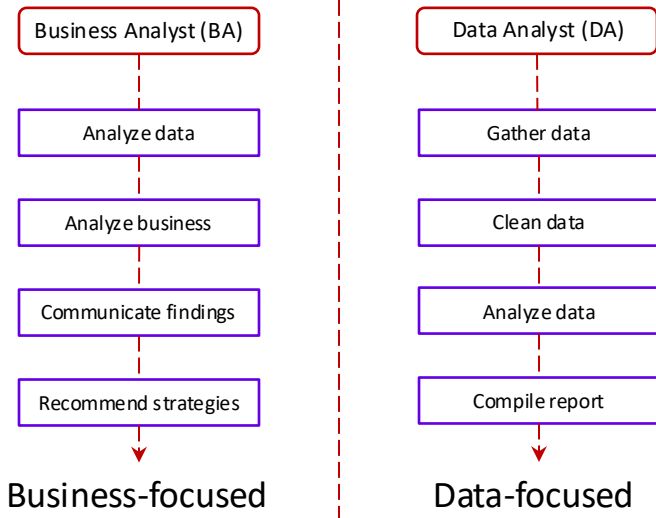
AI and Data Science project manager

An AI and Data Science project manager is a professional who **oversees projects** related to artificial intelligence.

Some of the responsibilities of an AI project manager are:
- Defining the **scope**, **objectives**, and **deliverables** of AI projects.
- Planning and managing the **budget**, **timeline**, and **resources** of AI projects.
- **Coordinating** and communicating with **project stakeholders**, such as clients, sponsors, team members, vendors, etc.
- Monitoring and controlling the **quality**, **performance**, and **risks** of AI projects.
- Implementing and **evaluating** AI solutions and ensuring they meet the project requirements and expectations.
- Reporting and presenting the **progress**, **results**, and lessons learned of AI projects.

Some of the skills and qualifications of an AI project manager are:
- Knowledge and experience in **artificial intelligence** concepts, techniques, and applications
- Ability to understand and translate **business** needs into AI solutions.
- Ability to lead and motivate cross-functional teams and collaborate with diverse stakeholders.
- Ability to analyze data, solve problems, and make decisions under uncertainty.

AI and Data Science project manager
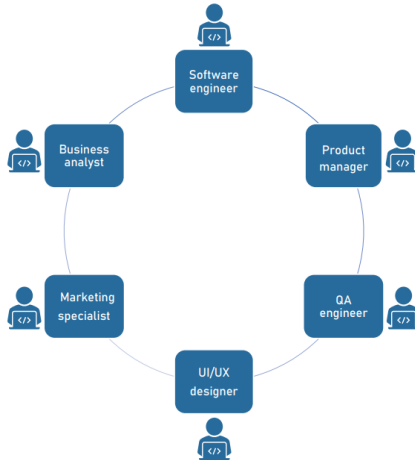
AI and Data Science project manager works with many teams in a project.

AI and Data scientist

An AI and data scientist is a person who uses **data** to create solutions for various business problems or opportunities.

Some of the responsibilities of an AI project manager are:
- Identify and mine reliable internal and external **data sources**
- Analyze and interpret data using statistical methods and **machine learning** algorithms
- Create and present data **visualizations**, reports, and dashboards
- Make **recommendations** on how a business can leverage data insights for innovation or optimization

Some of the skills and qualifications of an AI project manager are:
- Programming skills, such as Python, R, or SQL.
- **Statistics** and **probability** skills, such as linear regression, hypothesis testing, probability distributions, or Bayesian inference
- **Machine learning** and **deep learning** skills, such as supervised learning, unsupervised learning, neural networks, or natural language processing
- Data analysis and **visualization** skills, such as Excel, Tableau, Power BI, Matplotlib, or Seaborn
- Communication and **presentation** skills, such as writing, speaking, storytelling, or public speaking

AI and Data scientist
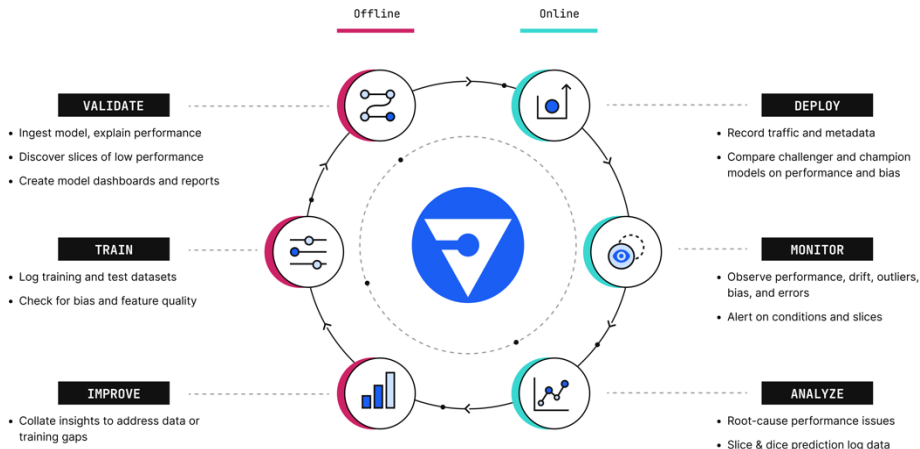
Example of day-to-day tasks of AI and data scientist.

AI and Data engineer

An AI and data engineer is a professional who builds and maintains **data systems** that enable the use of artificial intelligence and machine learning for various business purposes.

Some of the responsibilities of an AI project manager are:
- Designing and managing the development and production of AI **infrastructure**
- Transforming machine learning models into specific applications using **APIs** or embedded code
- **Automating** infrastructure for data collection, transformation, ingestion, and analysis
- Developing **data pipelines** for data extraction, loading, and processing
- Testing and **deploying** data systems and ensuring their reliability, security, and scalability

Some of the skills and qualifications of an AI project manager are:
- Programming skills, such as Python, R, Java, C++, or Scala
- **Data engineering** skills, such as SQL, NoSQL, Hadoop, Spark, or Airflow
- **Data science** skills, such as statistics, probability, machine learning, deep learning, or natural language processing
- **AI skills**, such as computer vision, speech recognition, or neural networks,
- **Cloud computing** skills, such as AWS, Azure, or Google Cloud Platform
- Data analysis and **visualization** skills, such as Excel, Tableau, Power BI, or Matplotlib.

AI and Data engineer
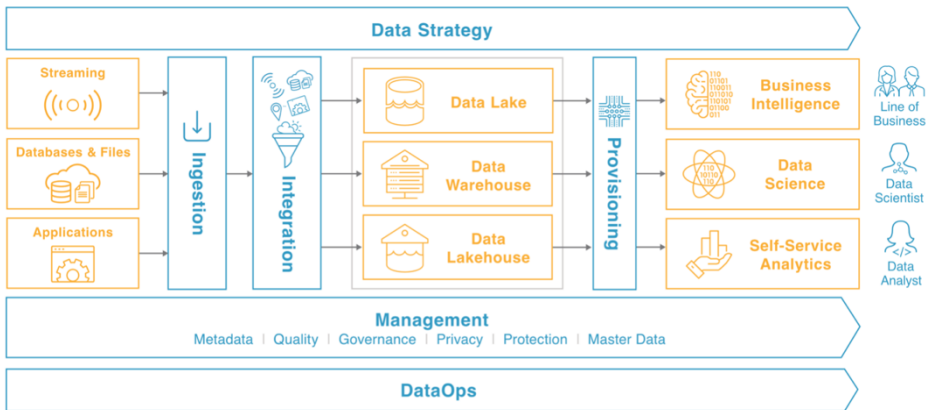
Example of a data pipeline built and managed by a data engineer.

AI and Data scientist vs. engineer

THE DATA SCIENCE
HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

@mrogati
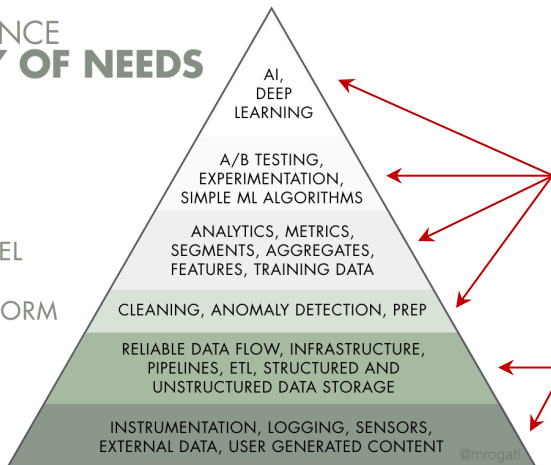
AI and Data scientist

AI and Data engineer
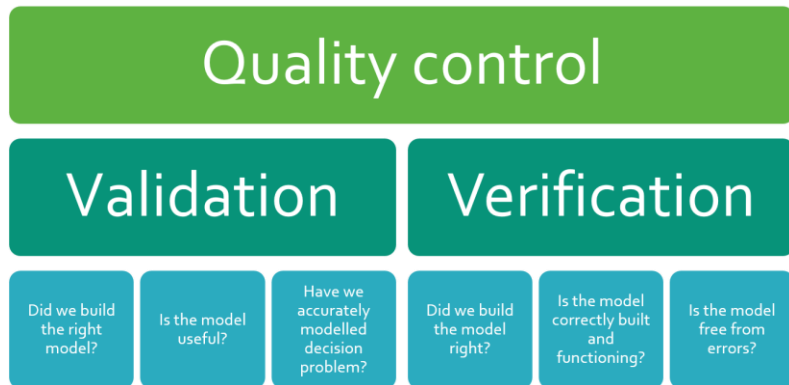
AI and Data quality control

AI and data quality control is a process of ensuring that the data used for AI and machine learning purposes meets a set of **predefined standards** and requirements.

Some of the responsibilities of an AI project manager are:
- Defining and communicating the requirements and standards for **data quality** management projects
- Creating and implementing **policies** or **processes** that will result in improved data quality
- Monitoring and reporting on the progress and outcomes of data quality management projects
- **Collaborating** with other team members, such as data engineers, data scientists, AI architects, and machine learning engineers
- Ensuring that the data systems and AI models are secure, **reliable**, scalable, and compliant with regulations.

Some of the skills and qualifications of an AI project manager are:
- Data quality skills, such as data profiling, cleansing, **validation**, verification, auditing, and governance
- **Programming** skills, such as Python, R, Java, C++, or Scala

Quality control is to ensure successful project outcomes.

Qualification

- Degree: Computer Science, Mathematics, Economics, …
- Having CS and/or Math background is a big plus.
- Hands-on experience is really appreciated.

Summary

- Tabular data.
- Complex data types:
    - Image
    - Text
    - Audio
- Artificial Intelligence (AI)
- AI applications.
- AI & Data Science in production.
- Career in AI and Data Science.

Q&A

Thank you