AI TECHNOLOGY FOR EXECUTIVE LEADERSHIP

TEKsystems Global Services

WORKFORCE DEVELOPMENT

Intellectual Property

Content Usage Parameters

Content refers to material including instructor guides, student guides, lab guides, lab or hands-on activities, computer programs, etc. designed for use in a training program

1

Content is subject to copyright protection

2

Content may only be leveraged by students enrolled in the training program

3

Students agree not to reproduce, make derivative works of, distribute, publicly perform and publicly display in any form or medium outside of the training program

4

Content is intended as reference material only to supplement the instructor-led training

TEKsystems | Global Services

COURSE UPDATE WEEK 4

Last week, we reviewed some basic use cases and organizational issues involved in AI

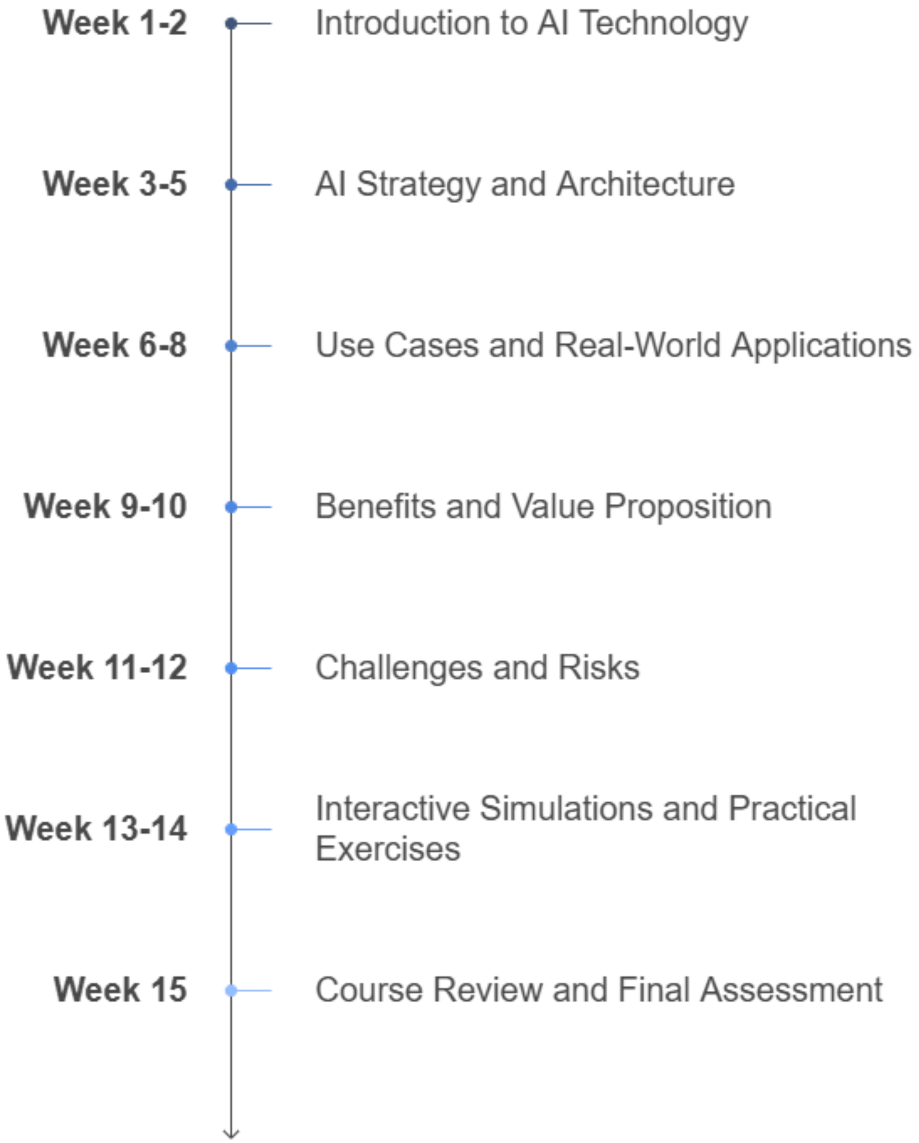
The next two sessions deal with AI infrastructure

This week

- Data Management Overview
- Data Cleaning and Engineering
- Feature Engineering
- Training Models

We will build on these topics in following weeks when we explore

- Monitoring and evaluating models
- AI Training Tools
- Data Governance and Risk



Week 1-2 Introduction to AI Technology

Week 3-5 AI Strategy and Architecture

Week 6-8 Use Cases and Real-World Applications

Week 9-10 Benefits and Value Proposition

Week 11-12 Challenges and Risks

Week 13-14 Interactive Simulations and Practical Exercises

Week 15 Course Review and Final Assessment

DATA

Stoll Elaborates

Data isn't information

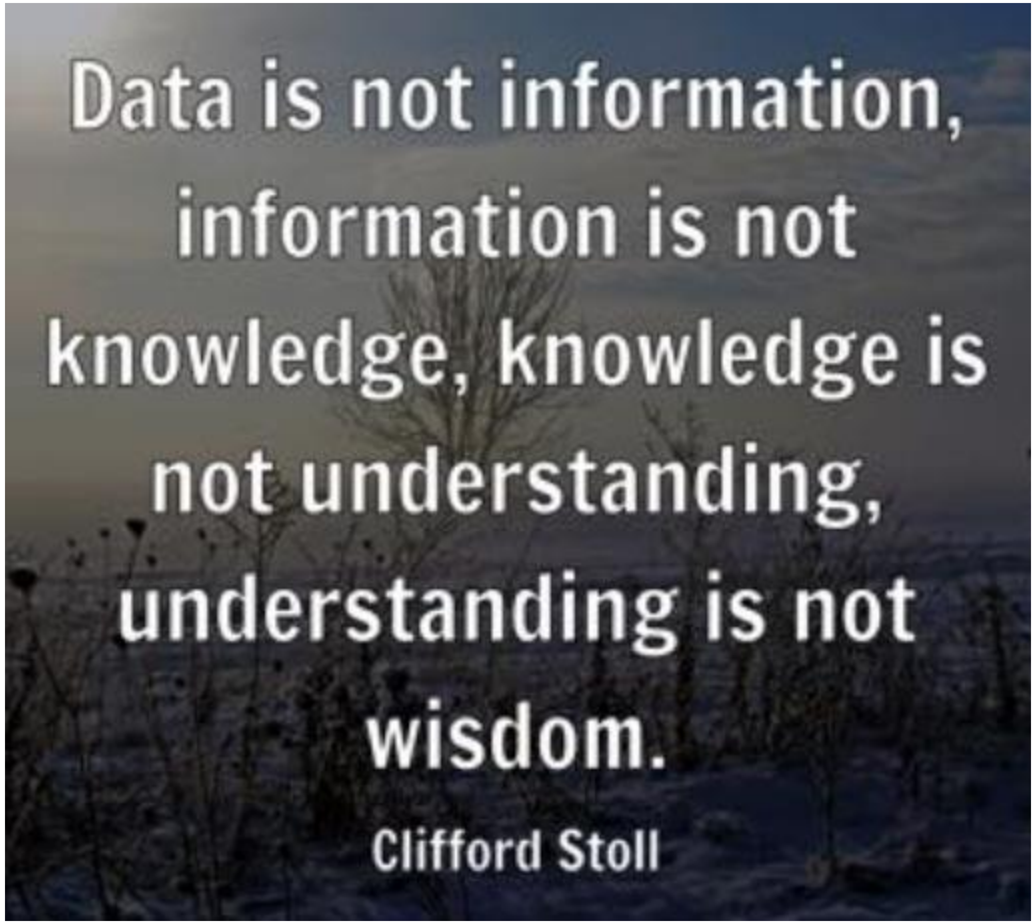
Information, unlike data, is useful.

While there's a gulf between data and information, there's a wide ocean between information and knowledge.

What turns the gears in our brains isn't information, but ideas, inventions, and inspiration.

Knowledge-not information-implies understanding.

And beyond knowledge lies what we should be seeking: wisdom.



Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.

Clifford Stoll

DATA INFORMATION

We can collect data

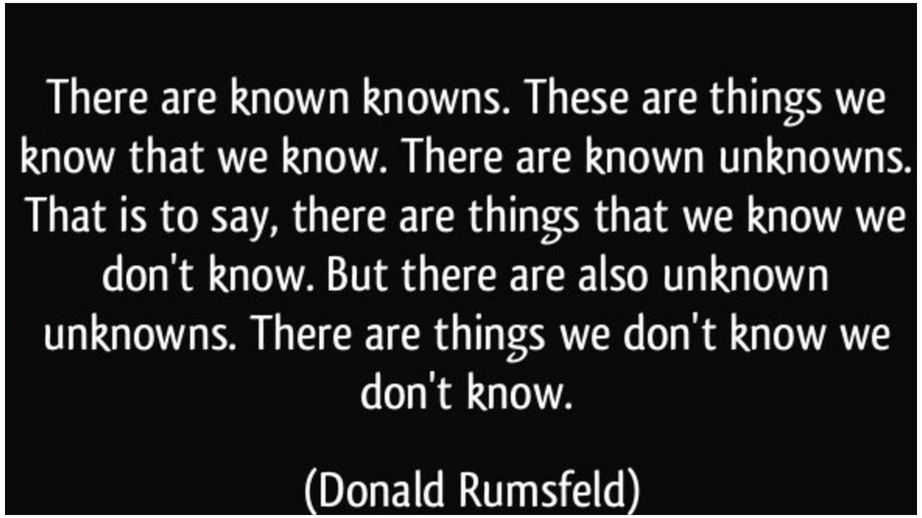
Turning data into information is done by the application of

- Data engineering
- Data cleaning
- Feature engineering

Turning that information into knowledge is the goal of AI

- AI models provide insights
- AI applications provide solutions to problems
- AI systems support decisions, planning and understanding

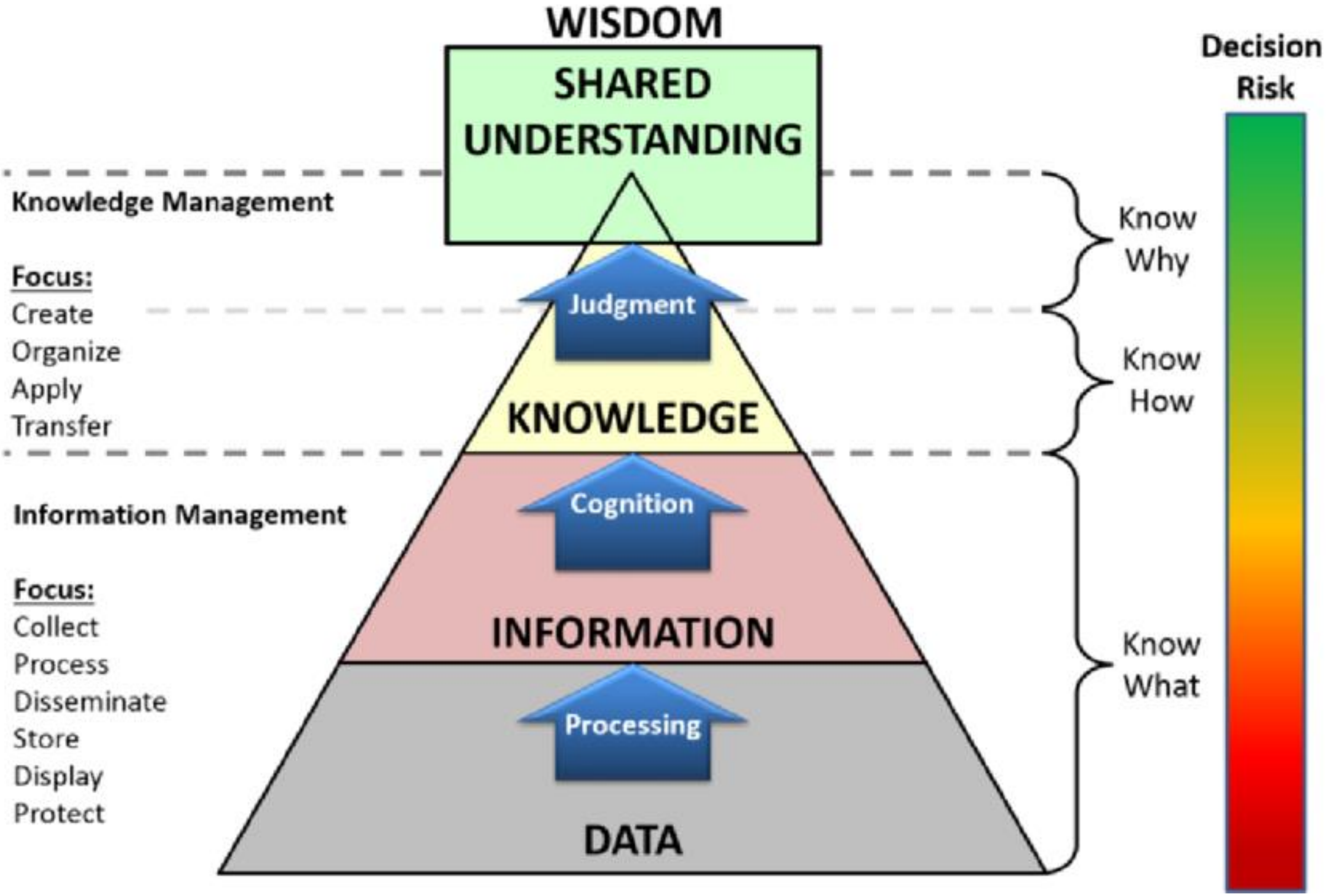
Knowledge tells us what we can do, wisdom is when we decide what we should do



There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

(Donald Rumsfeld)

DATA PYRAMID



DATA

- Raw, unprocessed facts and figures without context. Essentially a mass of measurements
- Characteristics:
  - Discrete entities.
  - Often unorganized and meaningless on its own.
  - Collected through observation, measurement, or generation by devices and systems.
- Example: All the text messages collected by the NSA or all the transactions in a banking system.

INFORMATION

- Data that has been processed, structured, or organized to provide context or meaning.
- Characteristics:
  - Answers questions like who, what, when, and where.
  - Give insights into the behaviors and characteristics of entities in the real world
  - Provides insights but lacks a deeper level of understanding.
- We can start to statistical analyses or create profiles of activities or exemplars of individuals or organization.
- Example:
  - The text messages collected by the NSA are analyzed to see if they are threats and who the sender and receivers are.
  - Or in the banking system, patterns of financial transactions based on customer type
- Information can be thought of as organizing data for use

KNOWLEDGE

- Information that has been analyzed, interpreted, and synthesized to identify patterns, relationships, or context.
- Characteristics:
  - Answers how questions.
  - Requires human or artificial intelligence for interpretation
  - Involves experience, learning, or expertise to add value.
- Examples:
  - Analysis of terrorist related chatter in text messages to predict potential targets and actors.
  - Analysis of bank transactions may help identify pattern indicative of money laundering

WISDOM

- The ability to apply knowledge judiciously and effectively to solve problems or make decisions.
- Characteristics:
  - Answers why questions.
  - Combines ethical judgment, foresight, and insight.
  - Often involves broader understanding and purpose.
- Examples:
  - The understanding of money laundering patters provides guidance in developing new financial systems and products that cannot be used for money laundering.
  - Understanding terrorist communications patters allows for covert infiltration of their networks

DAMA

The sections in this course on data will refer to the DAMA BOK (in the repository)

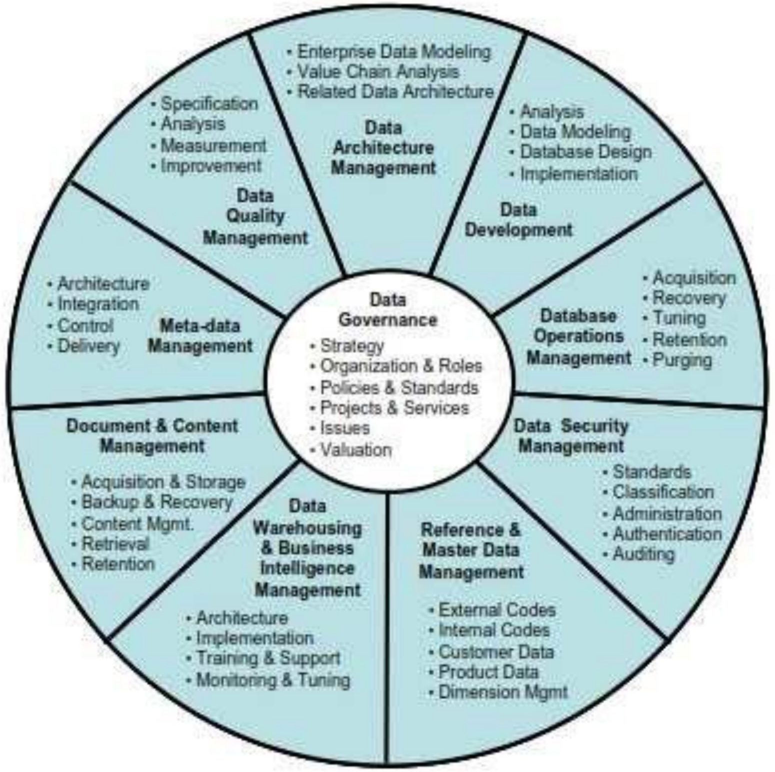
DAMA defines several of dimensions of data management as shown in the diagram.

We will be touching on each of these areas as we go through the course.

Today’s content will intersect a number of the slices in the DAMA pie

DAMA International

The Global **Data Management** Community

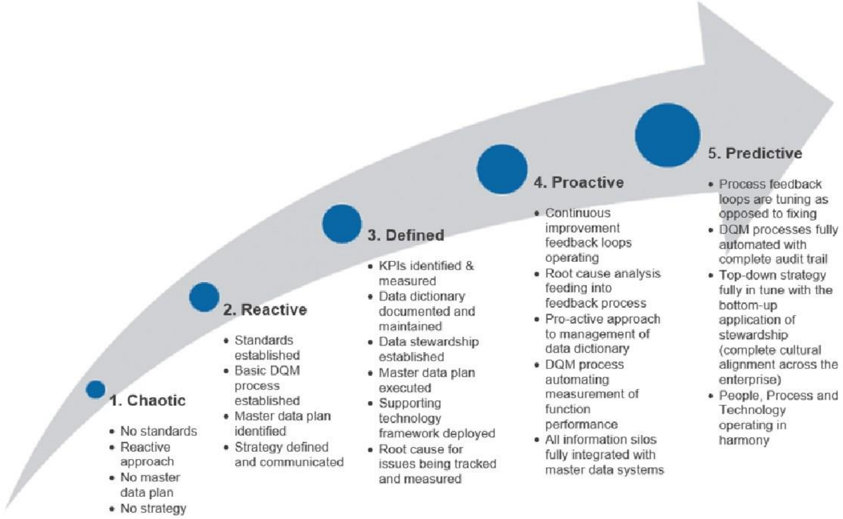


DATA MATURITY MODELS

Just like with CMMI, there are a number of data maturity models in use

We will come back to these in later lectures

There is a lot of different approaches to data maturity depending on what aspects of data are being considered



1. Chaotic
- No standards
- Reactive approach
- No master data plan
- No strategy

2. Reactive
- Standards established
- Basic DQM process established
- Master data plan identified
- Strategy defined and communicated

3. Defined
- KPIs identified & measured
- Data dictionary documented and maintained
- Data stewardship established
- Master data plan executed
- Supporting technology framework deployed
- Root cause for issues being tracked and measured

4. Proactive
- Continuous improvement feedback loops operating
- Root cause analysis feeding into feedback process
- Pro-active approach to management of data dictionary
- DQM process automating measurement of function performance
- All information silos fully integrated with master data systems

5. Predictive
- Process feedback loops are tuning as opposed to fixing
- DQM processes fully automated with complete audit trail
- Top-down strategy fully in tune with the bottom-up application of stewardship (complete cultural alignment across the enterprise)
- People, Process and Technology operating in harmony

DATA MANAGEMENT

- DAMA definition
  - Data Management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles.
  - Data management is the practice of collecting, organizing, protecting and storing an organization’s data so it can be analyzed for business decisions.
- Grouped into different disciplines which overlap and interact
  - **Data Governance** – Establishes policies, procedures, and responsibilities for managing data as an asset.
  - **Data Architecture** – Defines the structure, integration, and management of data across an organization.
  - **Data Modeling & Design** – Creates conceptual, logical, and physical models to structure and organize data.

DATA MANAGEMENT

- **Data Storage & Operations** – Manages the storage, retrieval, and maintenance of data across systems.
- **Data Security** – Ensures data protection, privacy, and compliance with regulatory requirements.
- **Data Integration & Interoperability** – Facilitates seamless data exchange between different systems and platforms.
- **Document & Content Management** – Organizes and controls unstructured data, including documents and media.
- **Reference & Master Data Management** – Maintains consistent, accurate, and authoritative reference data across systems.
- **Data Warehousing & Business Intelligence** – Supports data analytics, reporting, and decision-making through structured data storage.
- **Metadata Management** – Captures, organizes, and maintains metadata to enhance data usability and governance.
- **Data Quality Management** – Ensures data accuracy, completeness, consistency, and reliability.
- **Big Data & Data Science** – Manages large-scale data processing, analytics, and machine learning applications.

TERMINOLOGY

- There are several related terms that are often confused
- Data Engineering
  - Function and responsibilities:
    - Building and managing data pipelines, databases, and data infrastructure.
    - Ensuring data is clean, structured, and accessible for analysis.
    - Optimizing storage and retrieval for performance.
  - Common Tools & Technologies:
    - Data Storage: SQL, NoSQL, Data Lakes (S3, Delta Lake)
    - ETL (Extract, Transform, Load): Apache Airflow, Talend, dbt
    - Big Data Processing: Apache Spark, Hadoop
    - Cloud Platforms: AWS (Redshift, Glue), Azure, GCP (BigQuery)

TERMINOLOGY

  - Example: DE builds a pipeline to
    - Collect raw customer interactions from a website, mobile app, and transactions.
    - Transform the data into a structured format (e.g., removing duplicates, handling missing values).
    - Store the processed data in a data warehouse like Amazon Redshift or Google BigQuery, making it accessible for analysts and scientists.

- Data Analytics
  - Functions and responsibilities
    - Analyzing past data to identify trends and patterns.
    - Creating reports, dashboards, and visualizations.
    - Helping businesses make data-driven decisions.
    - This function in enhanced and supported with AI models and tools
  - Common Tools & Technologies:
    - Supervised and unsupervised learning models
    - Visualization & BI Tools: often used for data exploration (what does our data look like?)
    - Statistical Analysis: Python (Pandas, Matplotlib), R

TERMINOLOGY

- Example: DE builds a pipeline to
  - Collect raw customer interactions from a website, mobile app, and transactions.
  - Transform the data into a structured format (e.g., removing duplicates, handling missing values).
  - Store the processed data in a data warehouse like Amazon Redshift or Google BigQuery, making it accessible for analysts and scientists.

- Data Analytics
  - Functions and responsibilities
    - Analyzing past data to identify trends and patterns.
    - Creating reports, dashboards, and visualizations.
    - Helping businesses make data-driven decisions.
    - This function in enhanced and supported with AI models and tools
  - Common Tools & Technologies:
    - Supervised and unsupervised learning models
    - Visualization & BI Tools: often used for data exploration (what does our data look like?)
    - Statistical Analysis: Python (Pandas, Matplotlib), R

TERMINOLOGY

- Example: A DA at a subscription-based streaming service:
  - Uses historical data to train a recommendation system (like Netflix's movie recommendations).
  - Develops a machine learning model to predict customer churn and suggest retention strategies.
  - Implements NLP tools to analyze customer feedback and improve content offerings.

Data Science

- Functions and responsibilities
  - Conducting deep statistical and exploratory analysis.
  - Identifying appropriate AI tools for a specific problem area
  - Exploring the capabilities and limits of how data can be used
  - Often lumped in together with Data Analytics
- Common Tools & Technologies:
  - Programming & ML Libraries: Python (Scikit-Learn, TensorFlow, PyTorch), R
  - Big Data & ML Pipelines: Spark ML, Databricks
  - Deep Learning: Keras, Hugging Face, OpenAI

TERMINOLOGY

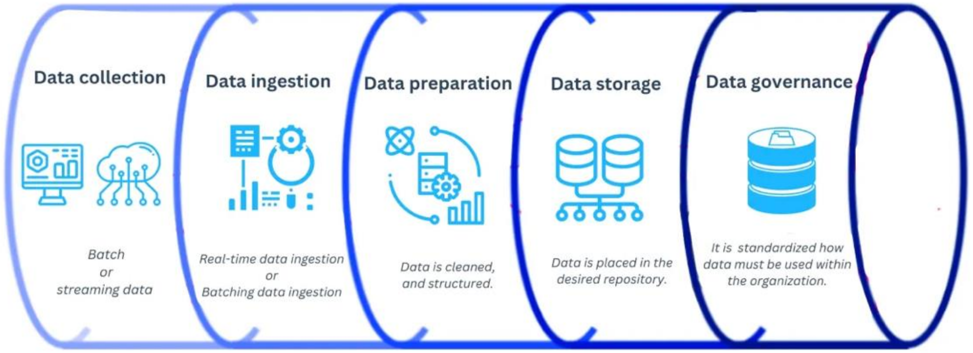
  - Example: A DS at a video streaming service:
    - Uses historical data to train a recommendation system (like Netflix's movie recommendations).
    - Develops a machine learning model to predict customer churn and suggest retention strategies.
    - Implements NLP (Natural Language Processing) to analyze customer feedback and improve content
- These are general categories
  - Generally, the DE prepares data for the use of the DA and DE
  - The DA uses various tools, including AI to convert the data into information
  - The DS uses information to develop models with predictive and/or explanatory power (knowledge)
  - The actual roles and responsibilities overlap with each other
- The rest of today will focus on the data engineering

DATA PIPELINE

A data pipeline is a set of automated processes that move and transform data from one system to another

Typically moving data from raw data sources to a destination such as a data warehouse, database, or analytics platform.

It feeds into the data sinks used as inputs for data analytics, BI and machine learning


Data collection
Batch or streaming data
Data ingestion
Real-time data ingestion or Batching data ingestion
Data preparation
Data is cleaned, and structured.
Data storage
Data is placed in the desired repository.
Data governance
It is standardized how data must be used within the organization.

STAGES IN A DATA PIPELINE

- Data Ingestion
  - Collects data from multiple sources, such as databases, APIs, logs, or streaming services.
- Data Processing (Transformation, Cleansing, and Enrichment)
  - Converts raw data into a usable format by filtering, aggregating, standardizing, and enriching it.
- Data Storage
  - Saves processed data in a data lake, data warehouse, or database for further analysis.
- Data Orchestration
  - Manages the workflow, scheduling, and dependencies between different pipeline components.
- Data Validation & Quality Checks
  - Ensures accuracy, consistency, completeness, and integrity before further use.
- Data Loading & Distribution
  - Moves data to its final destination
- Monitoring & Logging
  - Tracks pipeline performance, detects failures, and ensures data consistency.

PIPELINE BEST PRACTICES

- Scalability
  - Design pipelines to handle increasing data volumes using distributed computing (e.g., Spark, Kafka).
- Modular Design
  - Break the pipeline into reusable and independent components for easy maintenance.
- Data Quality Checks
  - Implement validation rules to catch errors before processing.
- Error Handling & Retry Mechanisms
  - Ensure automatic retries and logging for failed jobs.
- Data Security & Compliance
  - Encrypt sensitive data and enforce access controls (GDPR, HIPAA compliance).
- Version Control & CI/CD
  - Use Git and automated testing for continuous integration and deployment.
- Monitoring & Alerting
  - Track latency, failures, and data anomalies using logging and alert systems.

V'S OF BIG DATA

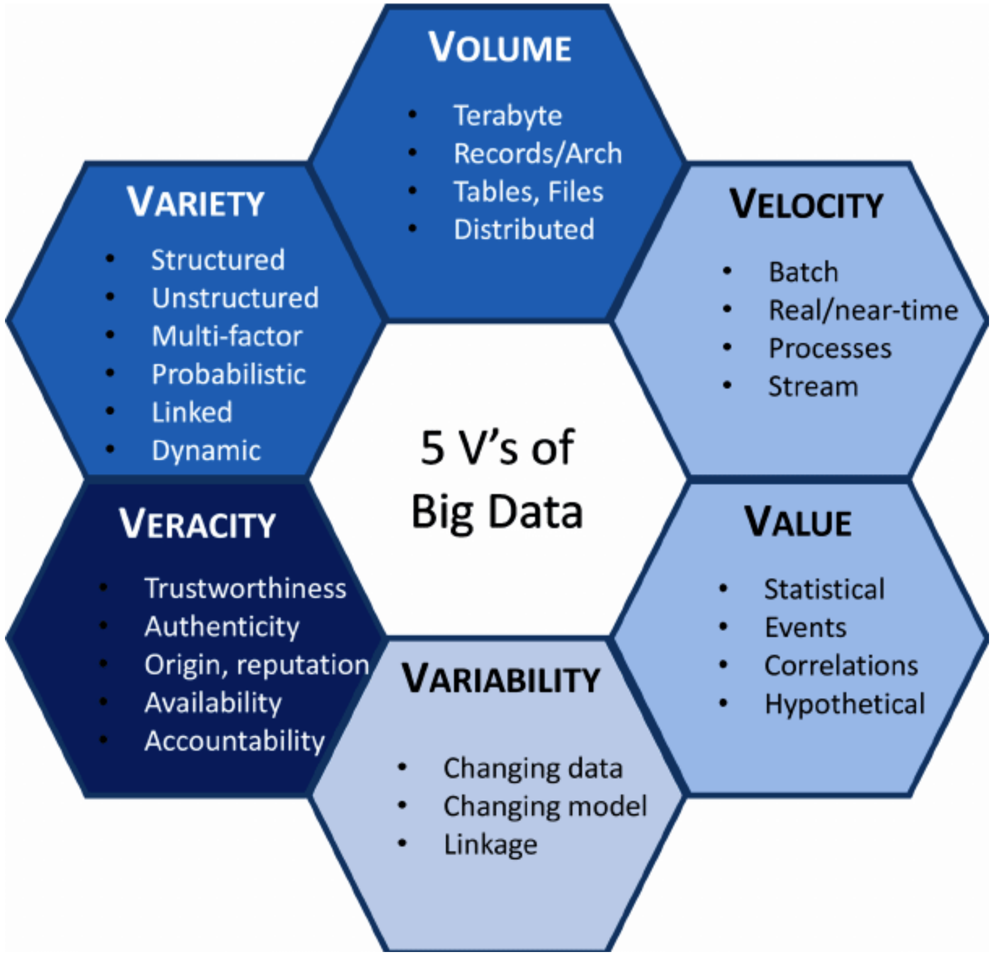
Volume – The massive amount of data generated from various sources, including social media, IoT devices, and transaction records.

Velocity – The speed at which data is generated, processed, and analyzed in real-time or near real-time.

Variety – The diversity of data formats, including structured (databases), semi-structured (JSON, XML), and unstructured (videos, images, text).

Veracity – The trustworthiness and quality of data, ensuring accuracy and consistency for decision-making.

Value – The usefulness of data in driving insights, innovation, and business growth.



COMMON PROBLEMS IN DATA EXTRACTION

- When extracting data from source systems, poor data quality can affect the reliability and accuracy of downstream processes.
- Missing Data:
  - Some records may have missing values due to incomplete sources, system errors, or inconsistent data entry.
- Duplicate Data:
  - Redundant or repeated records can appear due to multiple data sources, poor deduplication logic, or system glitches.
- Inconsistent Data Formats:
  - Variations in date formats, numerical representations, or text capitalization can create inconsistencies across datasets.

COMMON PROBLEMS IN DATA EXTRACTION

- **Incorrect or Inaccurate Data:**
  - Errors from manual data entry, sensor malfunctions, or system bugs can lead to incorrect values.
- **Outliers and Anomalies**:
  - Unexpected extreme values can distort analysis and may indicate data corruption or fraud.
- **Schema Drift:**
  - Changes in source data structures like new columns, renamed fields
- **Data Integrity Violations:**
  - Referential integrity issues, such as missing foreign key relationships, can cause inconsistencies.
- **Encoding and Character Set Issues**:
  - Differences in character encoding (e.g., UTF-8 vs. ASCII) can cause unreadable text or corruption.
- **Truncated or Corrupt Data**:
  - Errors during extraction or transmission may result in incomplete or corrupted records.

DATA CLEANING

The process of identifying and correcting errors, inconsistencies, and inaccuracies

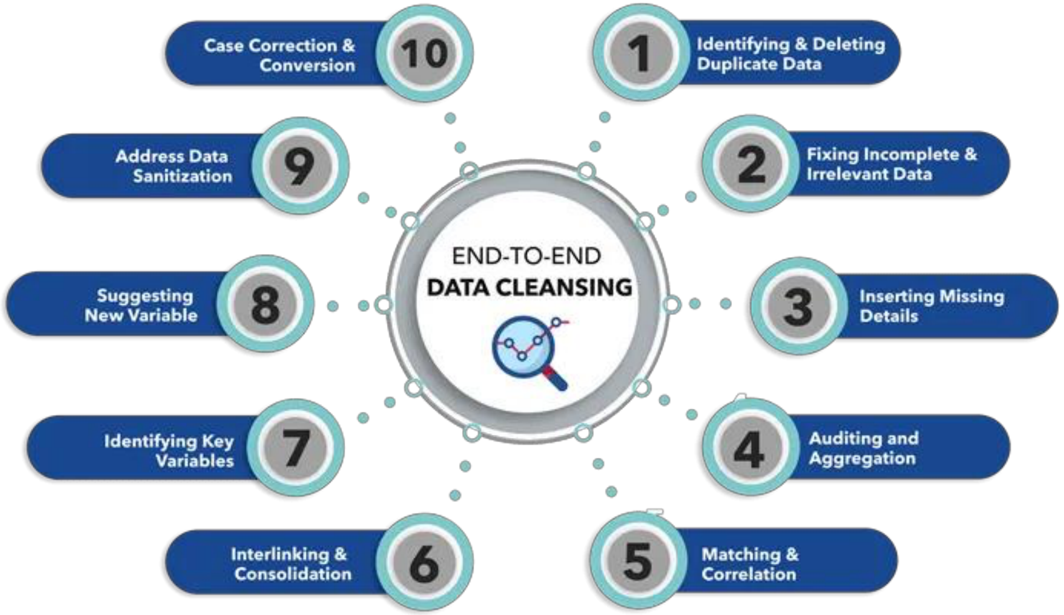
Improves data quality and ensure reliability for analysis, machine learning, and decision-making.

Data cleaning does not transform data but focuses only on data quality

Transforming the data means changing the data values

Think of transforming the data as semantic, like editing the text in a document, which changes how the data is represented

Cleaning the data is like spell checking and fixing the punctuation and formatting the layout of the document


END-TO-END
DATA CLEANSING
1 Identifying & Deleting Duplicate Data
2 Fixing Incomplete & Irrelevant Data
3 Inserting Missing Details
4 Auditing and Aggregation
5 Matching & Correlation
6 Interlinking & Consolidation
7 Identifying Key Variables
8 Suggesting New Variable
9 Address Data Sanitization
10 Case Correction & Conversion

DATA CLEANING

- Common issues for data cleaning
  - **Misspelled Entries**:
    - Typos and spelling mistakes can lead to categorization errors.
    - “United States” versus “Untied States”
  - **Inconsistent Formats**:
    - Dates, numbers, or categories might be represented differently within the same dataset.
    - May 6, 2025 as 05/06/2025 or 06/05/25 or 2025-05-05
    - US number 5,456.99 as opposed to European number 5 456,99
  - **Outliers and Errors:**
    - Unusual or erroneous entries can lead to inaccurate analysis.
    - Extreme values – unless they can be validated are usually treated as errors
    - Person weight is listed as as 845 lbs
  - **Duplicate Records:**
    - Redundant data can lead to inaccurate statistics and conclusions.
    - If the data conflict, there needs to be a reconciliation process

DATA CLEANING

- Common issues for data cleaning
  - **Null or Missing Values:**
    - Incomplete data may lead to gaps in analysis and can lead to inaccurate and/or limited insights.
    - This can be historical where certain data was not collected prior to a specific date
    - This can also be the result of changes in the source data collection
  - **Inaccurate Data:**
    - Incorrect or outdated information can lead to inaccurate decisions.
    - Can result from errors in sensors or recording data – some reliability measure is often used
  - **Unstandardized Units:**
    - Different units of measurement can create data inconsistency issues, particularly when comparing or aggregating data.
    - This may not show up until the system is in production
  - **Incompatible Data:**
    - Conflicting data from different sources can cause discrepancies in data integration and analysis.
    - Note, these are not duplicate records but different records of the same entity

MARS CLIMATE ORBITER

The Mars Climate Orbiter was lost on September 23, 1999, due to a unit conversion error between two engineering teams.

Cause of Failure:

- The NASA navigation team used the metric system (Newtons, meters, kilograms).
- The Lockheed Martin engineering team used imperial units (pound-force, feet, pounds).
- This mismatch led to incorrect thrust calculations for trajectory adjustments.
- As a result, the spacecraft entered too low an altitude in Mars' atmosphere (~57 km instead of the intended ~150 km).
- The orbiter burned up or broke apart due to atmospheric stress.

DATA CLEANING EXAMPLES

- Missing Data:
  - Some records may have empty or null values.
  - Solution:
    - Fill in missing values using imputation (mean, median, mode)
    - Or remove incomplete records.
    - The solution will depend on the how the data will be used
    - Removing incomplete records could skew the data
    - For example, income missing for some women in New York, deleting the records might make the resulting data non-representative of the population
- Duplicate Records
  - The same data entry appears multiple times, leading to redundancy.
  - Solution:
    - Use deduplication techniques to remove repeated entries based on unique identifiers.
    - If there are no unique identifier, other methods have to be used

DATA CLEANING EXAMPLES

- Inconsistent Formatting
  - Data fields may have varying formats, like different date formats.
  - Solution:
    - Standardize formats across all records.
- Incorrect Data Entries
  - Human errors or system glitches introduce incorrect values (e.g., a phone number entered in a name field).
  - Solution:
    - Use validation rules and data type enforcement.
    - If there are type rules (like data patterns (xxx) xxx-xxxx), we can often correct the data

DATA CLEANING EXAMPLES

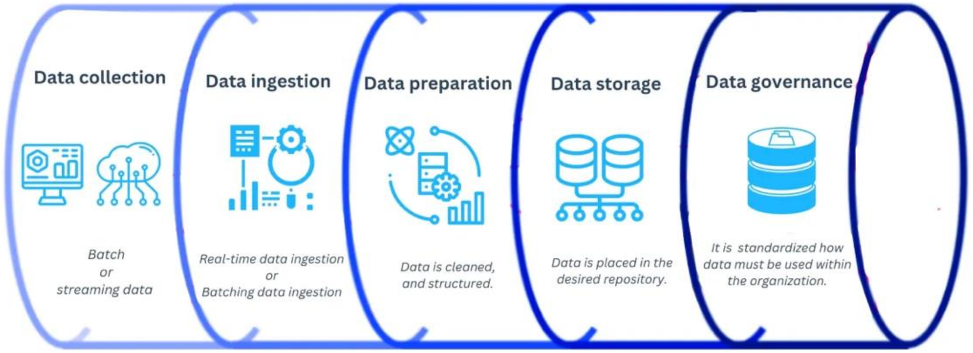
- Outliers and Anomalies
  - Extreme values may distort analysis (e.g., an age recorded as 300 years).
  - Solution:
    - Detect outliers using statistical methods (Z-score, IQR)
    - Decide whether to correct or remove them.
    - May require going back to the source system to validate the data
- Inconsistent Categorical Labels
  - Variations in category names (e.g., "New York" vs. "NY" vs. "N.Y.").
  - Solution:
    - Normalize categorical values using mapping or reference tables.
    - Create master data sets that are used across all data for consistency
- Typos and Spelling Errors
  - Text data may contain spelling mistakes or variations (e.g., "Jonh" instead of "John").
  - Solution:
    - Use text preprocessing techniques like fuzzy matching or spell checkers.

DATA PIPELINE

Cleaning the data is the first step of data preparation

Data transformation is the process of converting raw data from its source format into a structured, clean, and optimized format suitable for analysis, reporting, or machine learning.

It ensures consistency, accuracy, and usability before loading data into the target system like a machine learning pipeline


Data collection
Batch or streaming data
Data ingestion
Real-time data ingestion or Batching data ingestion
Data preparation
Data is cleaned, and structured.
Data storage
Data is placed in the desired repository.
Data governance
It is standardized how data must be used within the organization.

DATA TRANSFORMATION EXAMPLES

- Data Filtering
  - Removing irrelevant or unnecessary data.
  - Example:
    - Filtering out inactive customer records or transactions older than five years.
    - Dropping data that might fail accuracy or reliability measures
- Data Aggregation
  - Summarizing data to reduce granularity.
  - Example:
    - Computing total sales per region instead of storing every transaction.
    - Binning values in to ranges, example, income or age

DATA TRANSFORMATION EXAMPLES

- Data Normalization and Standardization
  - Converting data into a common reference unit or scale.
  - Example:
    - Converting multiple currency into USD
    - Converting prices into a reference year “In 1980 dollars”
- Data Type Conversion
  - Changing data types to ensure consistency.
  - Example:
    - Converting currency amounts into decimal or floats.
    - Converting phone number from numeric into strings

DATA TRANSFORMATION EXAMPLES

- Data Enrichment
  - Enhancing data by integrating additional information.
  - Example:
    - Adding geographic coordinates based on customer addresses
    - Appending demographic details to customer profiles.
- Pivoting and Unpivoting
  - Reshaping data for better analysis.
  - Example:
    - Transforming multiple columns (e.g., "Jan Sales," "Feb Sales") into a single column ("Month" with corresponding sales values).
    - Splitting on column into many – address into street, city, state

DATA TRANSFORMATION EXAMPLES

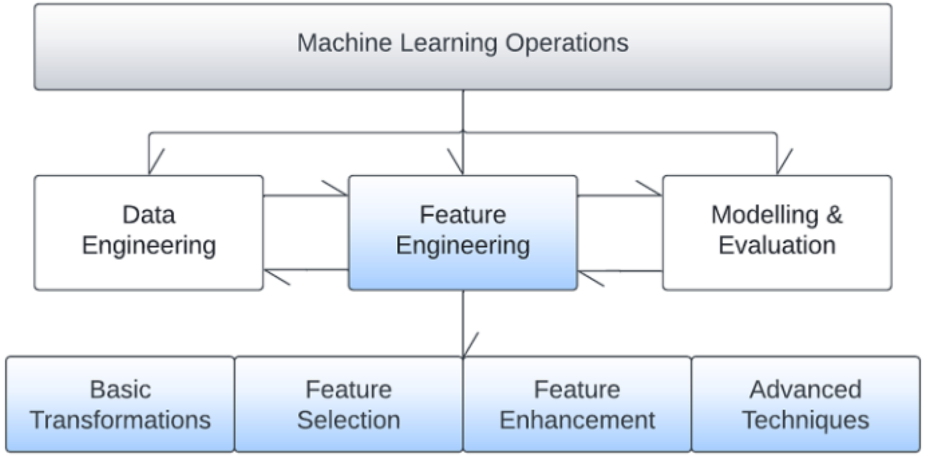
- Key Mapping and Joining
  - Combining data from multiple sources based on common keys.
  - Example:
    - Merging customer data from a CRM system with transaction data from an e-commerce platform using customer IDs.
- Derivation
  - Creating pre-computed attributes for more efficient processing
  - Example
    - Create an age value computed from date of birth

FEATURE ENGINEERING

Beyond ETL transformations, ML also applies transformations as part of the ML process

Feature engineering is a preprocessing step in supervised machine learning and statistical modeling

It transforms raw data into a more effective set of inputs for ML to reduce bias and variance and improve model accuracy

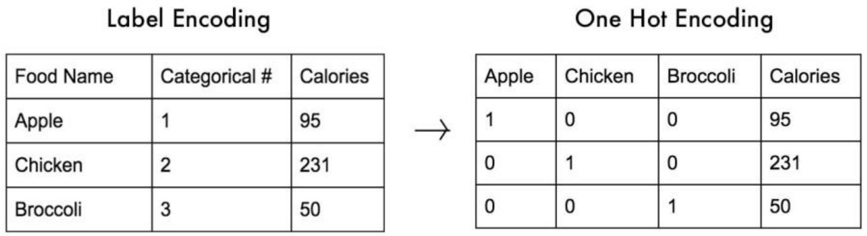


HANDLING CATEGORICAL VARIABLES

Non-numeric values have to be converted into numbers

- One-Hot Encoding: Converts categorical values into binary vectors.
- Label Encoding: Assigns a unique integer to each category (for tree-based models).
- Ordinal Encoding: Applies numerical encoding for ordered categories (e.g., Low=1, Medium=2, High=3).
- Binary Values: True/False, Yes/No mapped to 1/0

Example: Colors mapped to an RGB vector



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

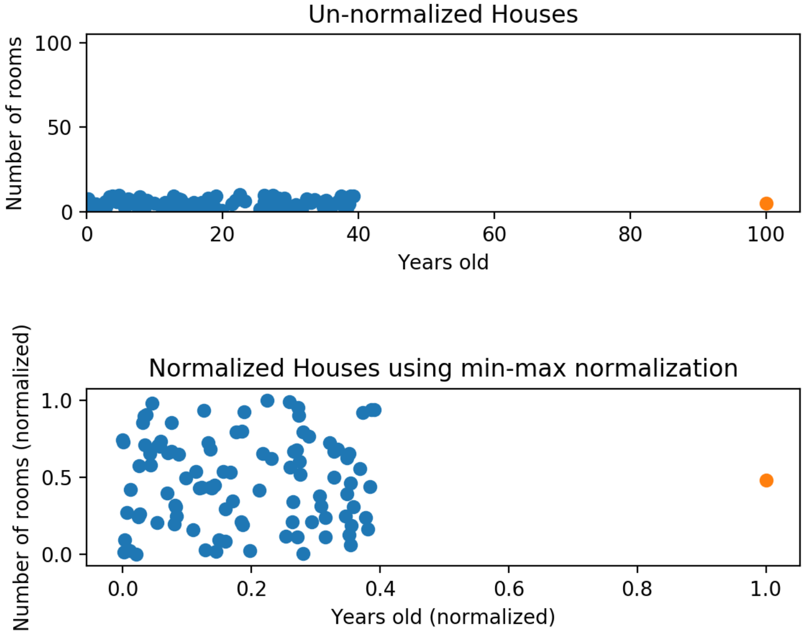
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

FEATURE SCALING & NORMALIZATION

Essential in machine learning to ensure that different numerical features contribute equally to model training.

Many ML algorithms perform better when input features have a similar range.

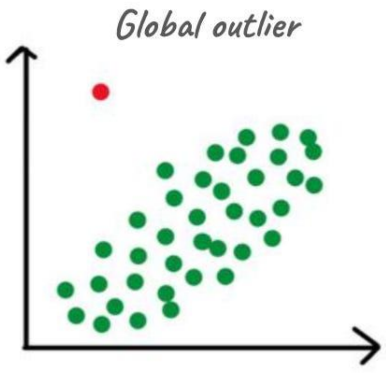
- Avoids Bias Toward Large-Scale Features
  - Some models (e.g., linear regression, k-NN, neural networks) give more weight to features with larger magnitudes.
  - Scaling prevents this imbalance.
- Improves Model Convergence
  - Gradient-based algorithms (gradient descent, deep learning) converge faster when input data is scaled, reducing training time.
- Enhances Distance-Based Model Performance
  - Models like k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and clustering algorithms rely on distance calculations, which are affected by scale differences..



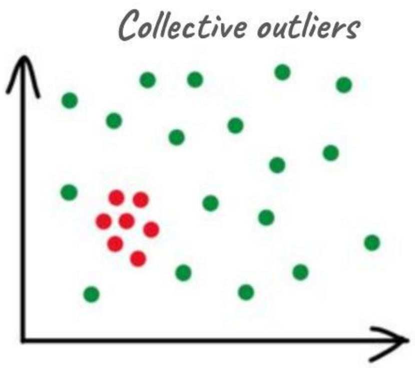
HANDLING OUTLIERS

Outliers are extreme values that deviate significantly from the rest of the data.

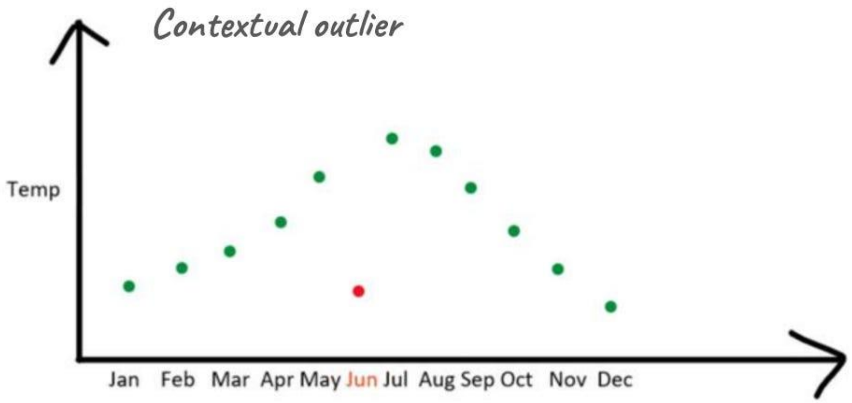
- Prevents Model Distortion
  - Outliers can skew statistical measures like the mean and standard deviation, affecting models that rely on them (e.g., linear regression).
  - Example: A few extremely high sales values may inflate the average, giving a misleading impression of business performance.
- Improves Machine Learning Model Accuracy
  - Some algorithms (linear regression, k-means, PCA) assume normal data distribution and can be heavily influenced by outliers.
  - Removing or transforming outliers improves model generalization and reduces errors.



The red data point is a global outlier.



The red data points as a whole are collective outliers.

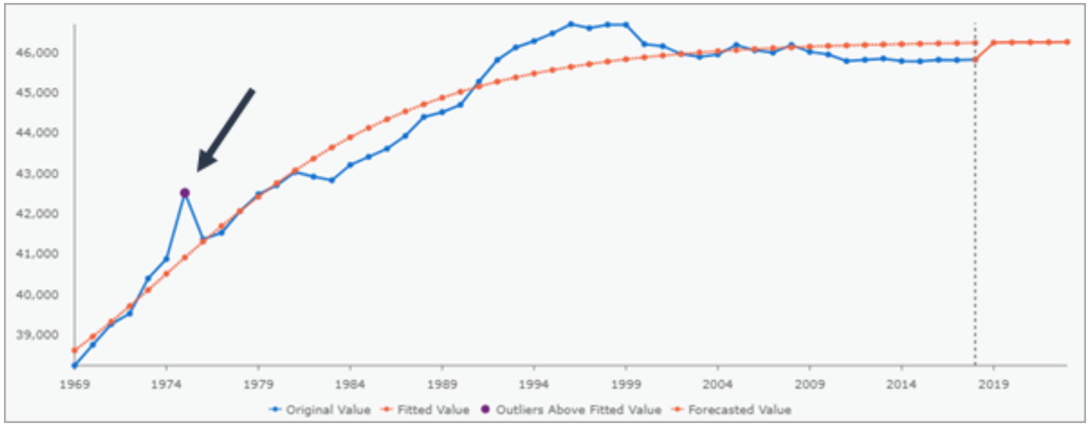


A low temperature value in June is a contextual outlier because the same value in December is not an outlier.

HANDLING OUTLIERS

Outliers are extreme values that deviate significantly from the rest of the data.

- Avoids Bias in Distance-Based Models
  - Algorithms like k-NN, clustering (K-means), and SVM rely on distance metrics, which outliers can disproportionately affect.
  - Example: One extreme value can change cluster assignments drastically in k-means clustering.
- Reduces Impact on Feature Scaling
  - When performing feature scaling (e.g., Min-Max scaling), outliers can stretch the range, making most values close to 0.
- Ensures Robust Statistical Analysis
  - outliers affect measures like mean, variance, and correlation, potentially leading to incorrect conclusions.
  - Example: A single outlier in a small dataset can make a correlation appear stronger or weaker than it actually is.



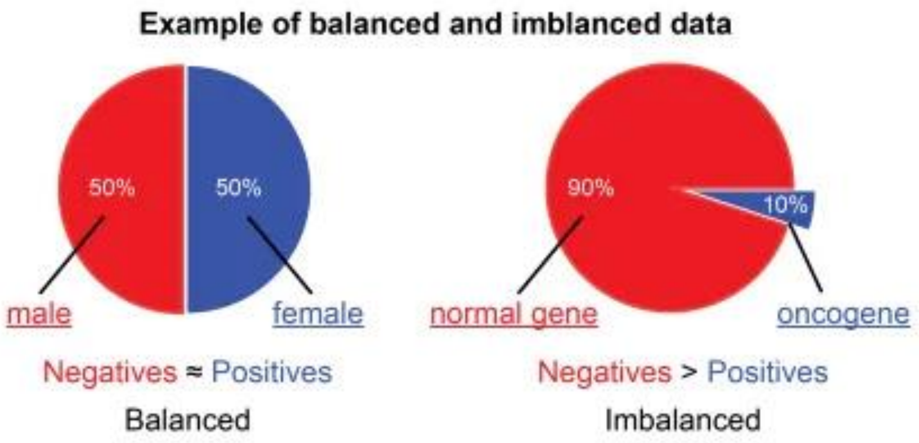
DATA BALANCING FOR CLASSIFICATION

Data imbalance occurs when one class significantly outnumbers others in a classification dataset.

This can lead to biased models that favor the majority class, reducing predictive performance on the minority class.

Data balancing techniques help improve model fairness and accuracy.

In the example shown, many samples would not contain a data from the positives subset



SMOTE

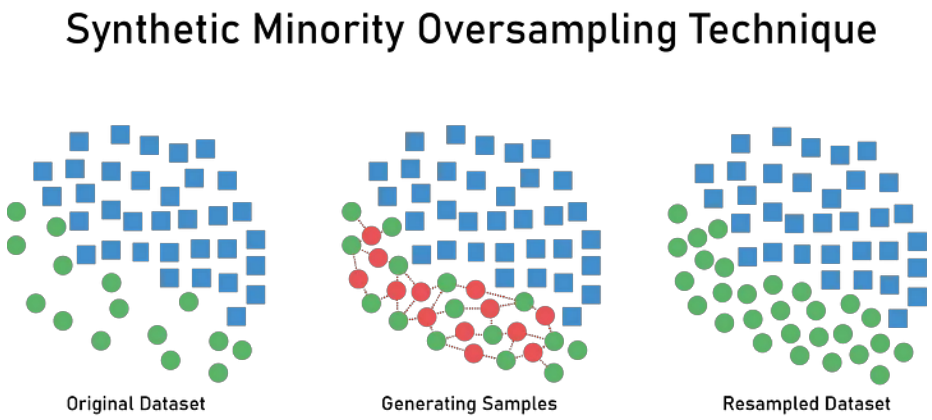
SMOTE - Synthetic Minority Over-Sampling Technique

Instead of duplicating existing minority class samples, SMOTE generates synthetic samples by interpolating between real ones.

It works by selecting a minority class instance, finding its nearest neighbors, and generating new samples along the vector between them.

When to use:

- When the dataset is small, and you want to increase the representation of the minority class.
- When duplicating data would lead to overfitting.



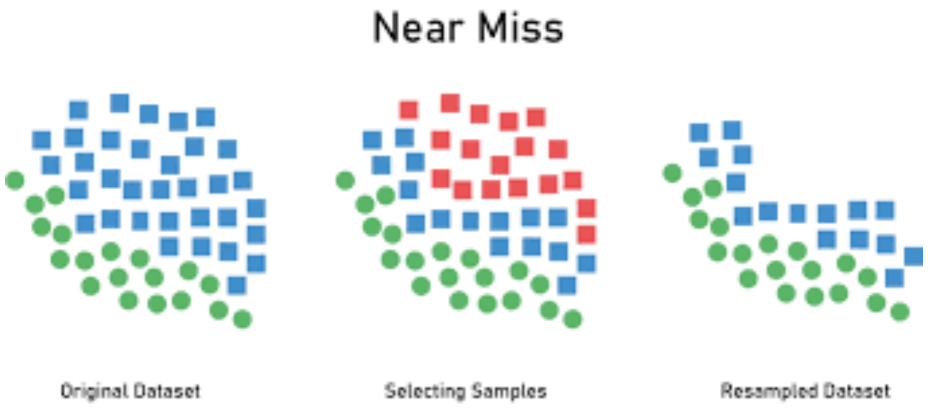
UNDERSAMPLING

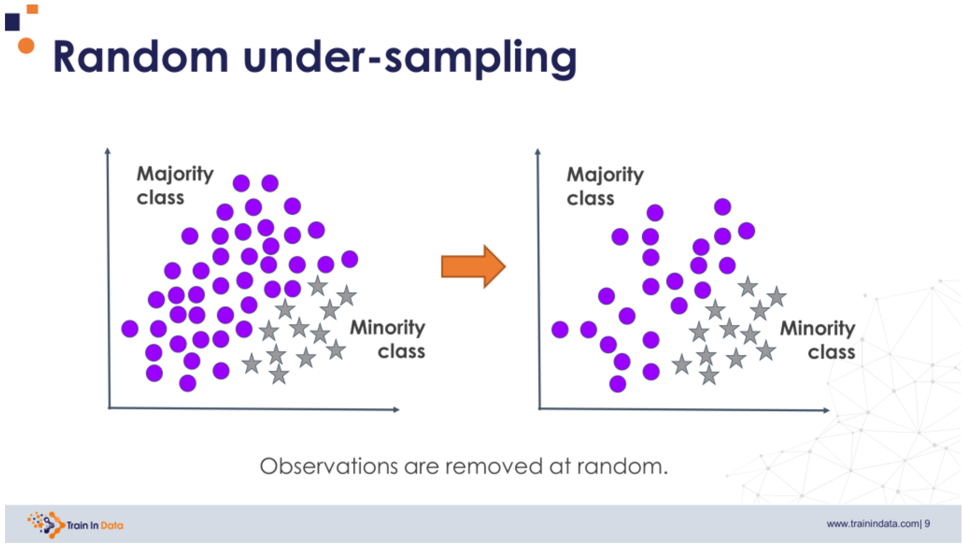
Reduces the number of majority class samples by randomly removing instances to balance class distribution.

Can be done using Random Undersampling (RUS) or more advanced techniques like NearMiss (selecting majority samples closest to the minority).

When to use:

- When the dataset is large, and keeping all majority class samples is unnecessary.
- When SMOTE is not suitable due to feature complexity or dataset structure.




Random under-sampling

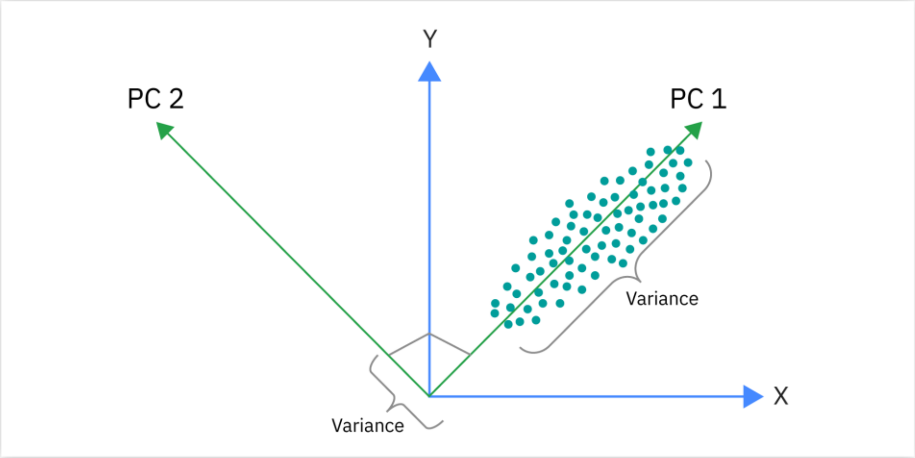
Observations are removed at random.

DIMENSIONALITY REDUCTION

Dimensionality reduction helps improve machine learning models by reducing the number of input features while preserving as much relevant information as possible.

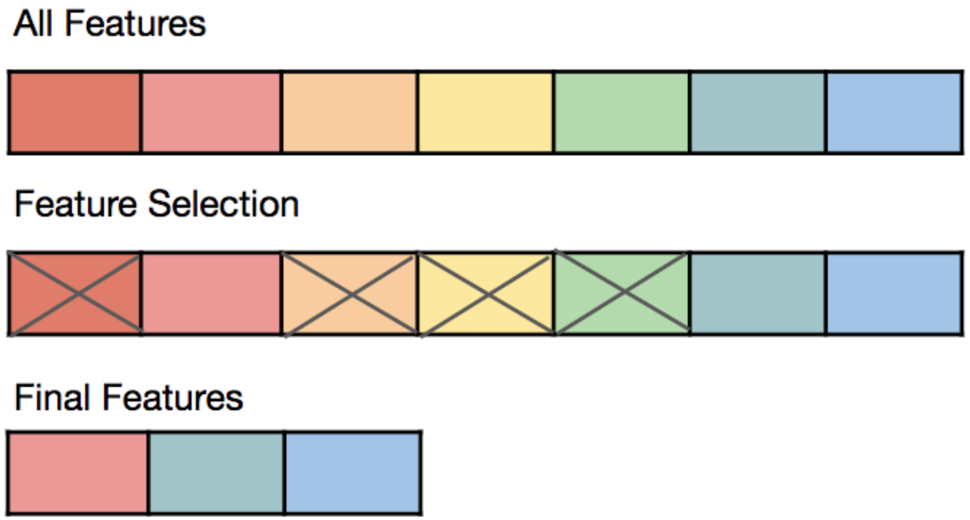
This improves efficiency, reduces overfitting, and enhances model interpretability.

- Principal Component Analysis (PCA)
  - Transforms high-dimensional data into a lower-dimensional space by finding new axes (principal components) that capture the most variance in the data.
  - It helps remove redundant features while keeping key patterns intact.



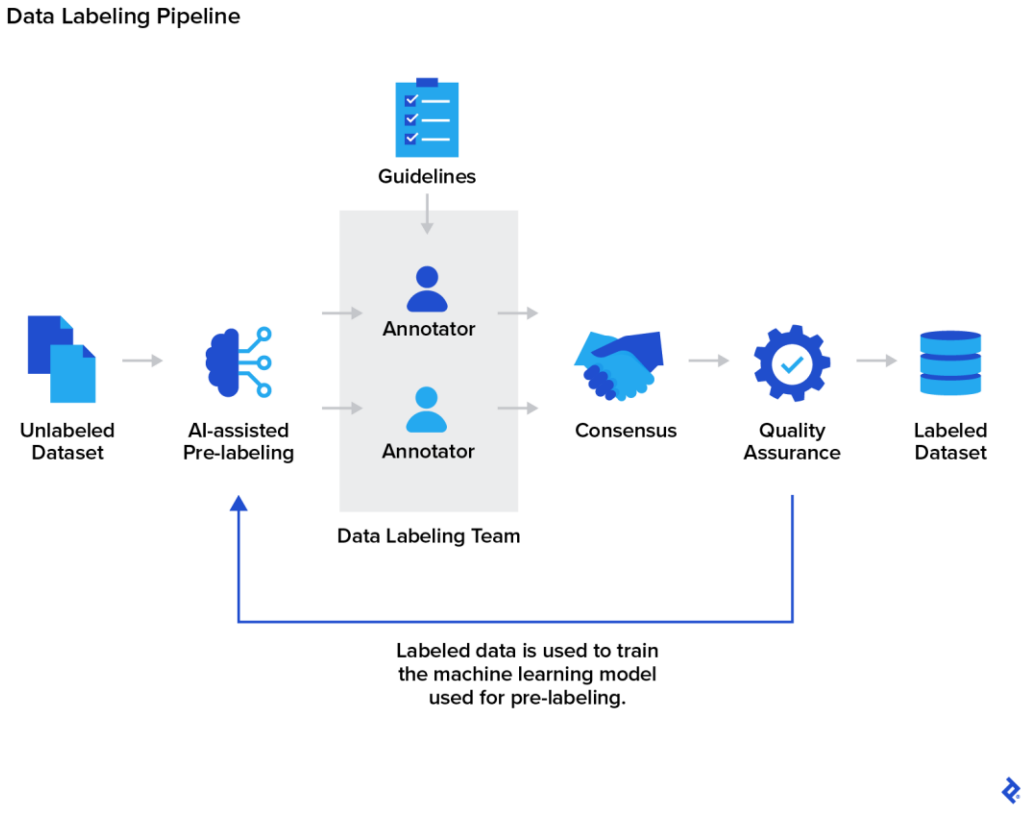
DIMENSIONALITY REDUCTION

- Feature Selection
  - Instead of transforming features (like PCA), feature selection directly removes irrelevant, redundant, or less important features to improve model efficiency.
  - It ensures that only the most relevant features contribute to model training.
- Types of Feature Selection:
  - Filter Methods: Selects features based on statistical properties.
  - Wrapper Methods: Uses model performance to evaluate feature importance.
  - Embedded Methods: Feature selection is built into model training.
  - Example: Decision trees automatically rank important features.



DATA LABELLING

- For supervised and semi-supervised learning labels are needed
- If labels are missing, they have to be added
- Labels may also need to be data cleaned
- Labels may need to be transformed



DATA LINEAGE IN DATA ENGINEERING

- Refers to the tracking and visualization of data’s flow from its origin (source) to its final destination
  - Includes all transformations, processing, and how data is moved between each step
  - It provides a clear map of where data comes from, how it moves, how it changes, and where it is used.
- Why is Data Lineage Important?
  - *Data Governance & Compliance*: Meets regulations like GDPR, HIPAA, and CCPA by tracing sensitive data.
  - *Debugging & Issue Resolution:* Identifies where data errors originate (e.g., incorrect calculations in ETL processes).
  - *Impact Analysis*: Assesses how changes in a data source affect downstream systems.
  - *Data Trust & Transparency*: Ensures data is reliable and can be confidently used for analytics and decision-making.
  - *Optimizing Data Pipelines*: Helps improve performance by identifying bottlenecks or redundant steps.

METADATA CAPTURED IN DATA LINEAGE

- Technical Metadata (System-Level Information)
  - Metadata that describes the technical properties of data.
    - Source & Destination – Where the data originates (databases, files, APIs) and where it is stored (data warehouses, data lakes).
    - Schema Information – Table names, column names, data types, constraints.
    - ETL & Transformation Logic – How data changes (aggregations, joins, filters, mappings).
    - Timestamps & Versioning – When data was created, modified, or updated.
    - Data Flow Information – Relationships between datasets, dependencies, and data movement patterns.
- Business Metadata (Context & Meaning)
  - Metadata that helps end-users understand the business context of the data.
    - Business Glossary – Definitions of business terms (e.g., "Customer ID" vs. "Client Number").
    - Data Ownership & Stewardship – Who is responsible for managing the data.
    - Sensitivity & Classification – Data privacy classifications (e.g., "PII", "Confidential").

METADATA CAPTURED IN DATA LINEAGE

- Operational Metadata (Performance & Monitoring)
  - Metadata related to execution, performance, and data quality.
    - Processing Time & Latency – How long ETL jobs take to execute.
    - Error Logs & Alerts – Records of data failures, errors, or anomalies.
    - Data Quality Metrics – Completeness, accuracy, consistency, freshness.

REFERENCE AND MASTER DATA

- Master Data
  - Provides the context for business activity data in the form of common and abstract concepts that relate to the activity.
  - It includes the details (definitions and identifiers) of internal and external objects involved in business transactions, such as customers, products, employees, vendors, and controlled domains (code values)
  - Master Data Management is the process of creating and maintaining a single master record - or single source of truth - for each person, place, and thing in a business.
  - Through MDM, organizations gain a trusted, current view of key data that can be shared across the business and used for better reporting, decision-making, and process efficiency.
- Reference Data
  - For example, code and description tables, is data that is used solely to characterize other data in an organization, or solely to relate data in a database to information beyond the boundaries of the organization.

Q&A AND OPEN DISCUSSION