Attendance code:

Some data fundamentals

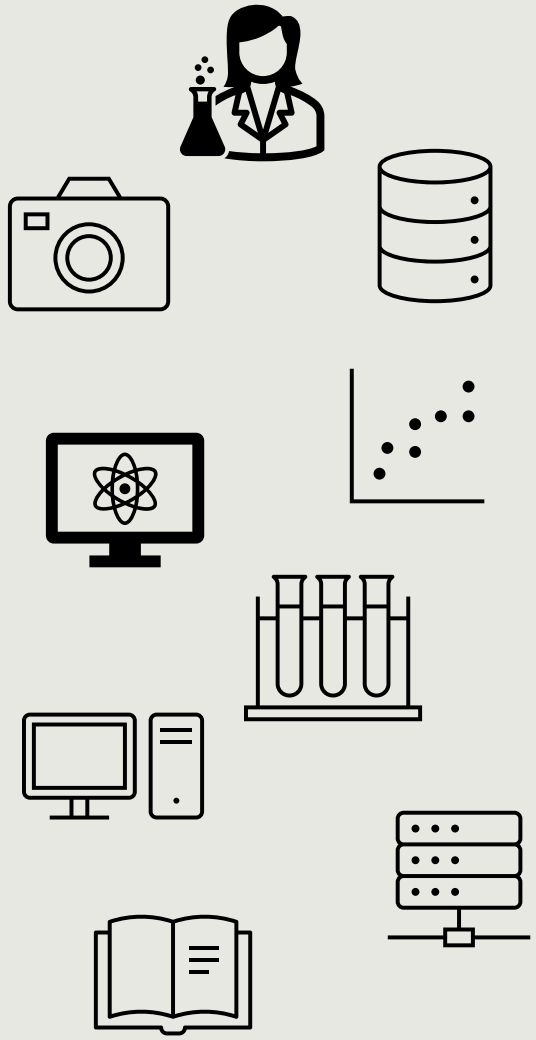
CHEM502 – Chemical Data, Discovery and Design
Sam Chong s.chong@liverpool.ac.uk

What is data?

Data is a collection of facts, numbers, words, observations or other useful information. Through data processing and data analysis, organizations transform raw data points into valuable insights that improve decision-making and drive better business outcomes.

https://www.ibm.com/think/topics/data, accessed 10/01/2025

Some data fundamentals

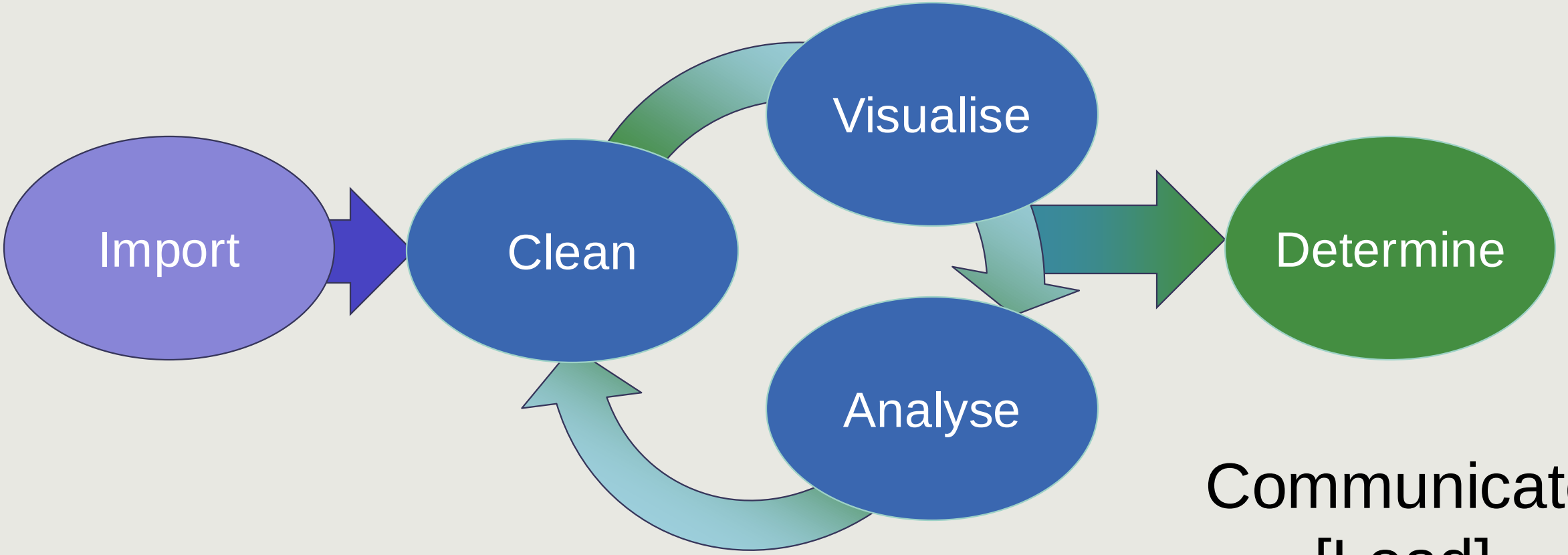


1102	0.232	True	AATCG
1103	0.324	True	CTACG
1104	0.278	False	ACTGA
1105	1.234	True	CCATG
1106	1.990	False	ATAAC
1107	0.288	False	AGTGA
1108	2.234	True	CAATG
1109	1.910	False	ATAAG

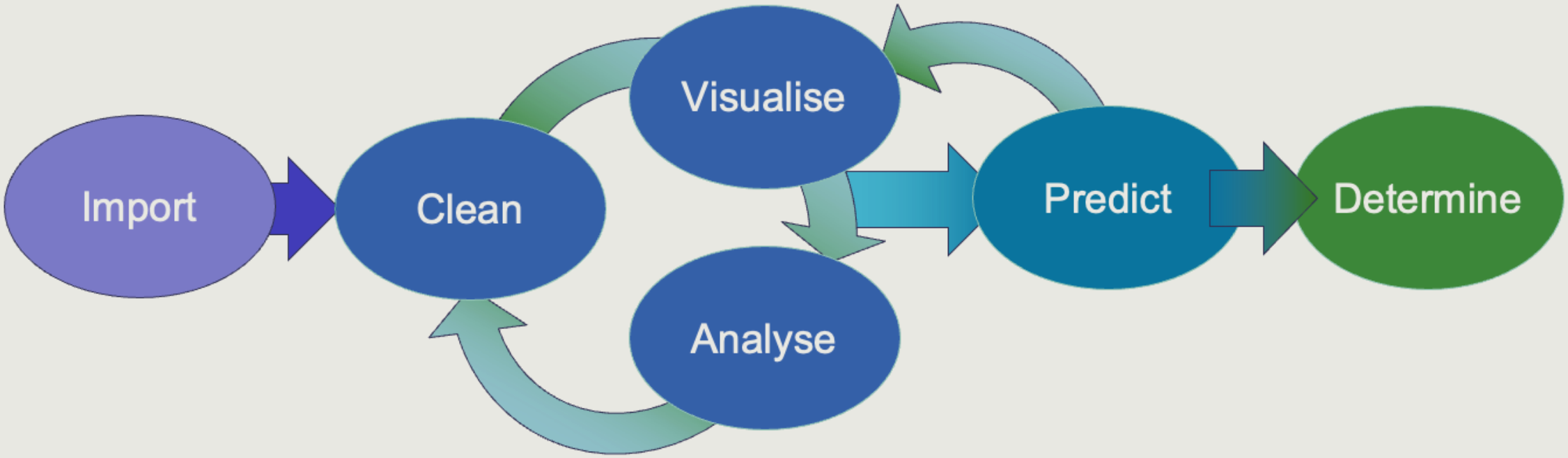
Extract

Transform

Load



Data → Knowledge → Understanding → Progress



Data → Knowledge → Understanding → Progress

How good is my data?

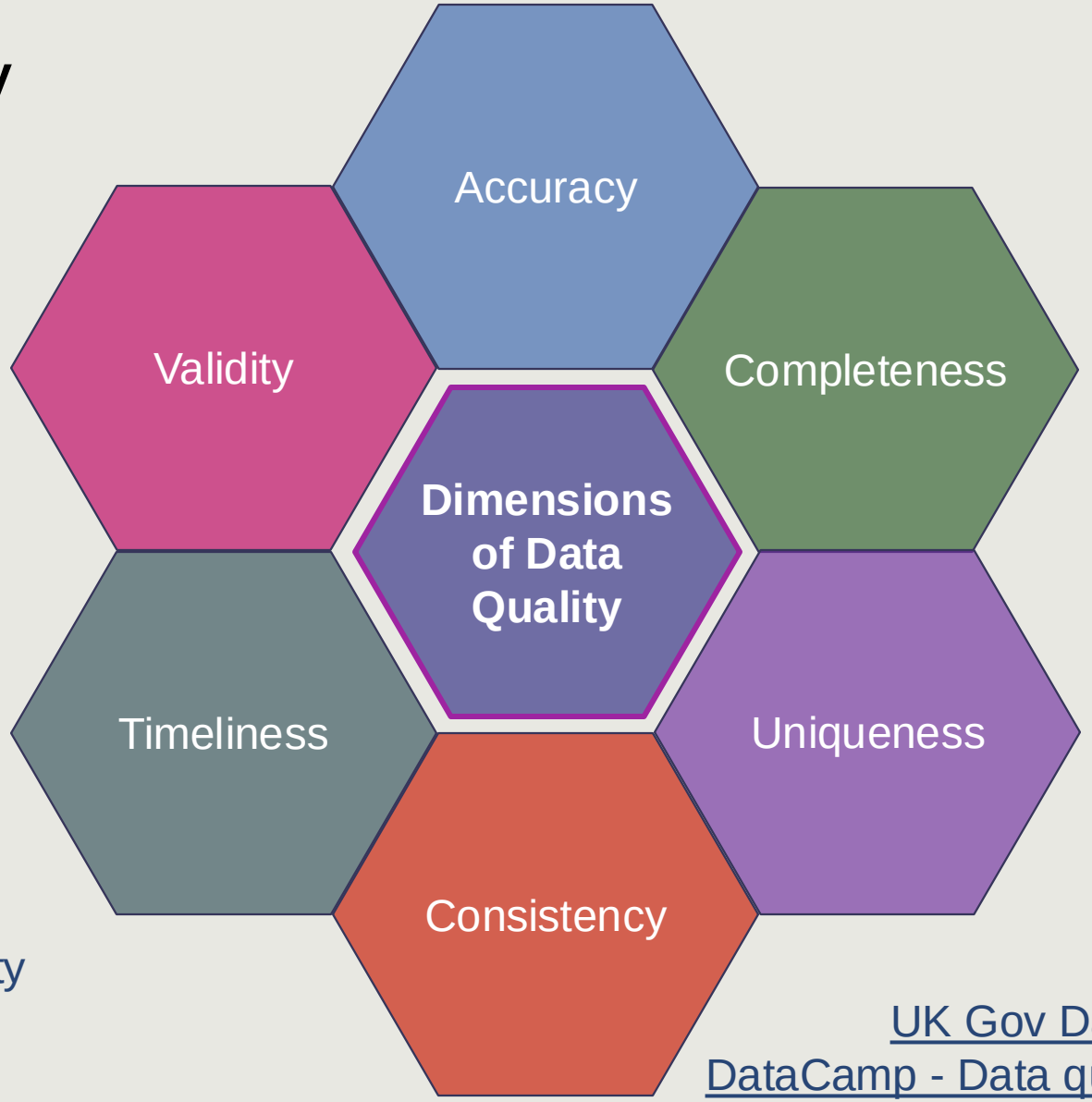
- Quality

  Cleanliness/tidiness

  Diversity

- Quantity

Data quality



Dimensions of data quality
Wikipedia
IBM

UK Gov Data Framework – Data quality
DataCamp - Data quality dimensions cheat sheet

Data bias

- Systematic errors or distortions in a dataset that can lead to misleading conclusions.
- Can lead to incorrect models, poor predictions and, ultimately, flawed scientific interpretations.

Data quality and AI – mitigating bias and error to protect fundamental rights (EU Agency for Fundamental Rights)

Types of data bias

- **Sampling bias**
- **Measurement bias**
- **Selection bias**
- **Confirmation bias**
- **Processing bias**
- **Algorithmic bias**

Types of data bias

	Cause
Sampling bias	Certain data points are overrepresented while others are missing
Measurement bias	Inaccuracies in data due to instrument calibration, human error, or environmental conditions.
Selection bias	Data points are chosen based on certain criteria, excluding others.
Confirmation bias	Interpreting or selecting data that supports pre-existing beliefs.
Processing bias	Data is transformed, cleaned, or analysed in a way that introduces distortions.
Algorithmic bias	The way algorithms and models use data produces results that can reinforce pre-existing biases towards a particular group or outcome.

Types of data bias

- **Sampling bias**
- **Measurement bias**
- **Selection bias**
- **Confirmation bias**
- **Processing bias**
- **Algorithmic bias**

Examples in chemistry

Collecting data only from successful experiments while ignoring failed ones

A study on reaction efficiency tests only inexpensive or commonly available reagents

A machine learning model trained only on well-behaved reaction datasets

Excessive smoothing on spectroscopic data, removing small but meaningful peaks

Discarding data points that do not fit an expected trend.

A balance that consistently measures 2 mg above the true mass.

Types of data bias

	Examples in chemistry
Sampling bias	Collecting data only from successful experiments while ignoring failed ones.
Measurement bias	A balance that consistently measures 2 mg above the true mass.
Selection bias	A study on reaction efficiency tests only inexpensive or commonly available reagents
Confirmation bias	Discarding data points that do not fit an expected trend.
Processing bias	Excessive smoothing on spectroscopic data, removing small but meaningful peaks
Algorithmic bias	A machine learning model trained only on well-behaved reaction datasets

Types of data bias

	Effect
Sampling bias	Leads to incorrect estimates of overall trends and variability.
Measurement bias	Systematic deviation from true values, leading to incorrect conclusions.
Selection bias	Conclusions may not generalise to the expected range of possible scenarios.
Confirmation bias	Overfitting models to expectations rather than reality.
Processing bias	Loss of important relevant information.
Algorithmic bias	Models produce misleading predictions, overconfident classifications, or fail to generalise.

Identifying and reducing bias

- Perform **exploratory data analysis** (EDA) to check for missing data, outliers, and unexpected trends.
- Be aware of the need for **diverse and representative sampling** of chemical data when designing experimental studies and selecting other data sources.
- Regularly **calibrate instruments** and verify against standards.
- Consider **blind analysis** to avoid confirmation bias.
- **Document and justify data cleaning and processing steps** to maintain transparency.

Exploratory Data Analysis (EDA)

What is EDA?

- A first “triaging” step in a workflow to assess a set of data.
- Process of **examining**, **summarising** and **visualising** data
- Aims to uncover patterns, detect anomalies, and gain insights before applying formal models or statistical tests.

What can we get from EDA?

- **Understand data structure** – Identify key variables, distributions, and relationships.
- **Detect errors and biases** – Spot missing values, outliers, or inconsistencies.
- **Generate hypotheses** – Reveal trends that can guide further investigation.
- **Choose appropriate models** – Decide whether data fits assumptions for statistical methods or machine learning models.

Examples of EDA in chemical data

- Reaction kinetics: plotting **rate vs. concentration** may reveal **non-linear trends** that indicate appropriate rate-order models.
- Spectroscopy: visualising raw **IR** or **NMR spectra** can help identify **baseline shifts** requiring additional processing or **unexpected peaks** before analysis.
- Materials science: initial exploration of **crystal structure datasets** may reveal possible **correlations between lattice parameters and properties**.

EDA techniques

Initial assessment of

- Whether values are “reasonable” or as expected.
- Shape of data and distributions.
- Relationships between variables (features).

EDA techniques

- **Descriptive statistics:** Mean, median, standard deviation
- **Visualisations:** Histograms, scatter plots, box plots
- **Missing data checks:** Heatmaps, bar charts of missing values
- **Correlation analysis:** Pairwise plots, correlation matrices

EDA techniques

- **Descriptive statistics:** Mean, median, standard deviation

  Quantitative ways to assess and compare central tendency (e.g. mean/median/mode)

  Spread of data (e.g. range, standard deviation, percentile ranges)

- **Visualisations:** Histograms, scatter plots, box plots

- **Missing data checks** Heatmaps, bar charts of missing values

- **Correlation analysis:** Pairwise plots, correlation coefficients/matrices

  Quantify strength and direction of relationships between variables

Why bother with visualisation?

Question

- If two datasets have the same mean, variance, and correlation, does that mean they are similar?

  Notebook