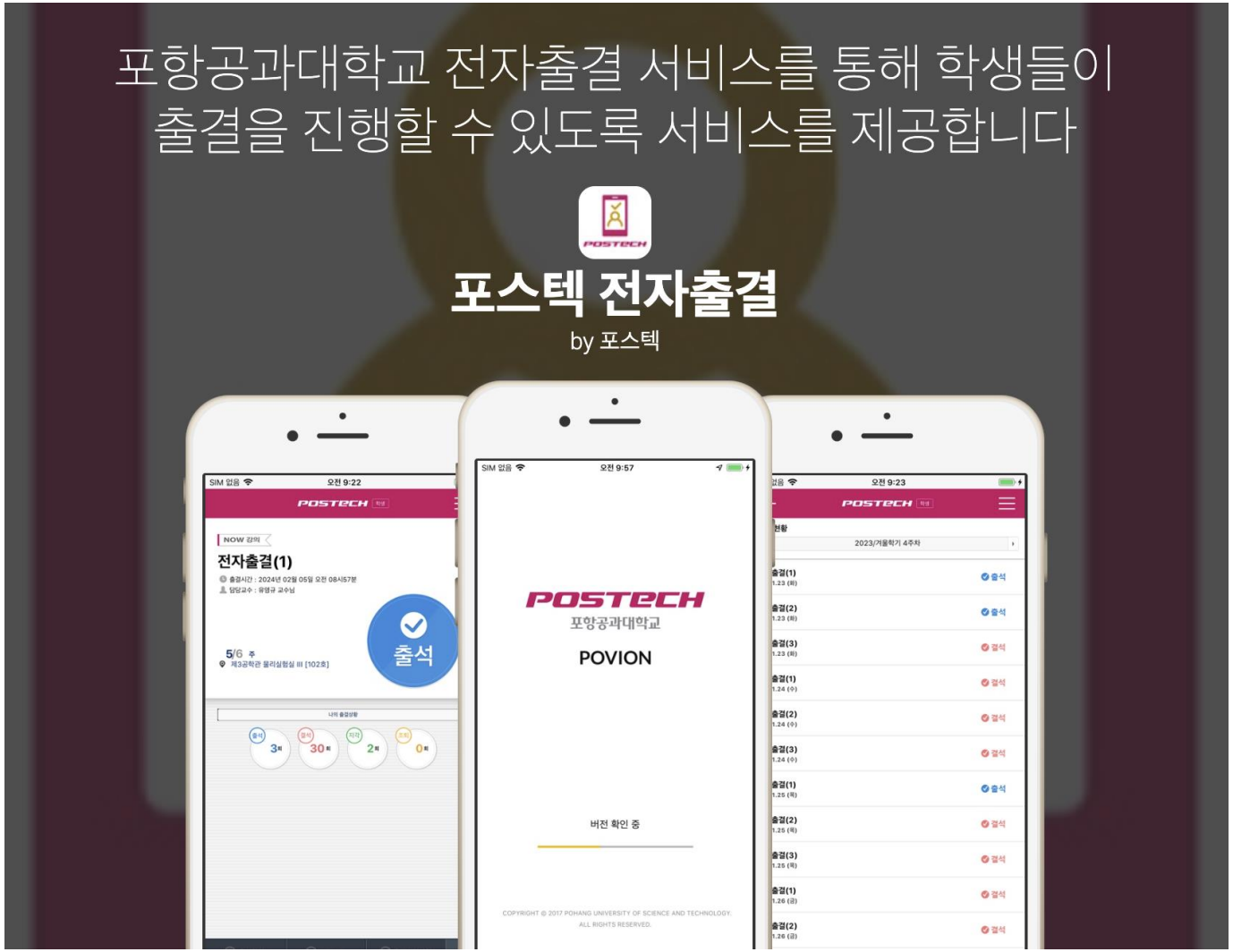
IMEN891K/IDSC501 Data Engineering

4. **Pre-processing I**: Cleaning & Integration

Fall 2025

Instructor: **Sangwoo Mo**

Recap: Attendance

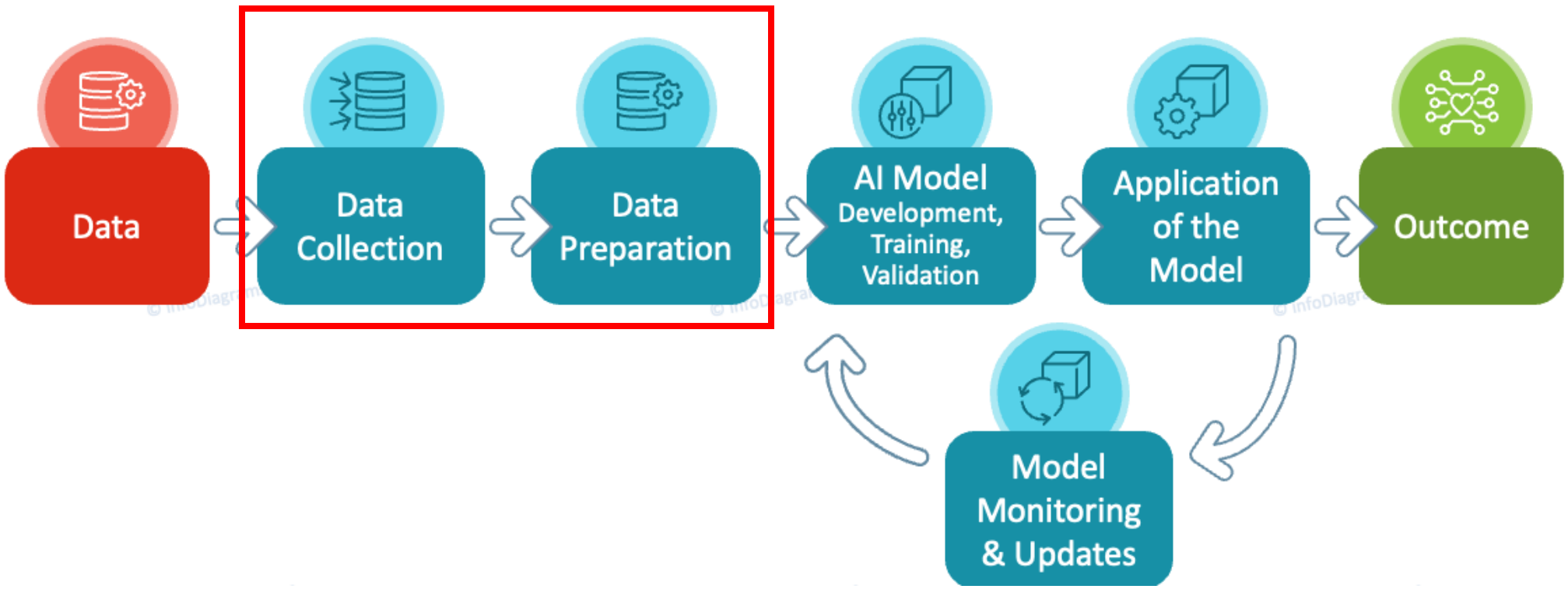


Recap: Team Project

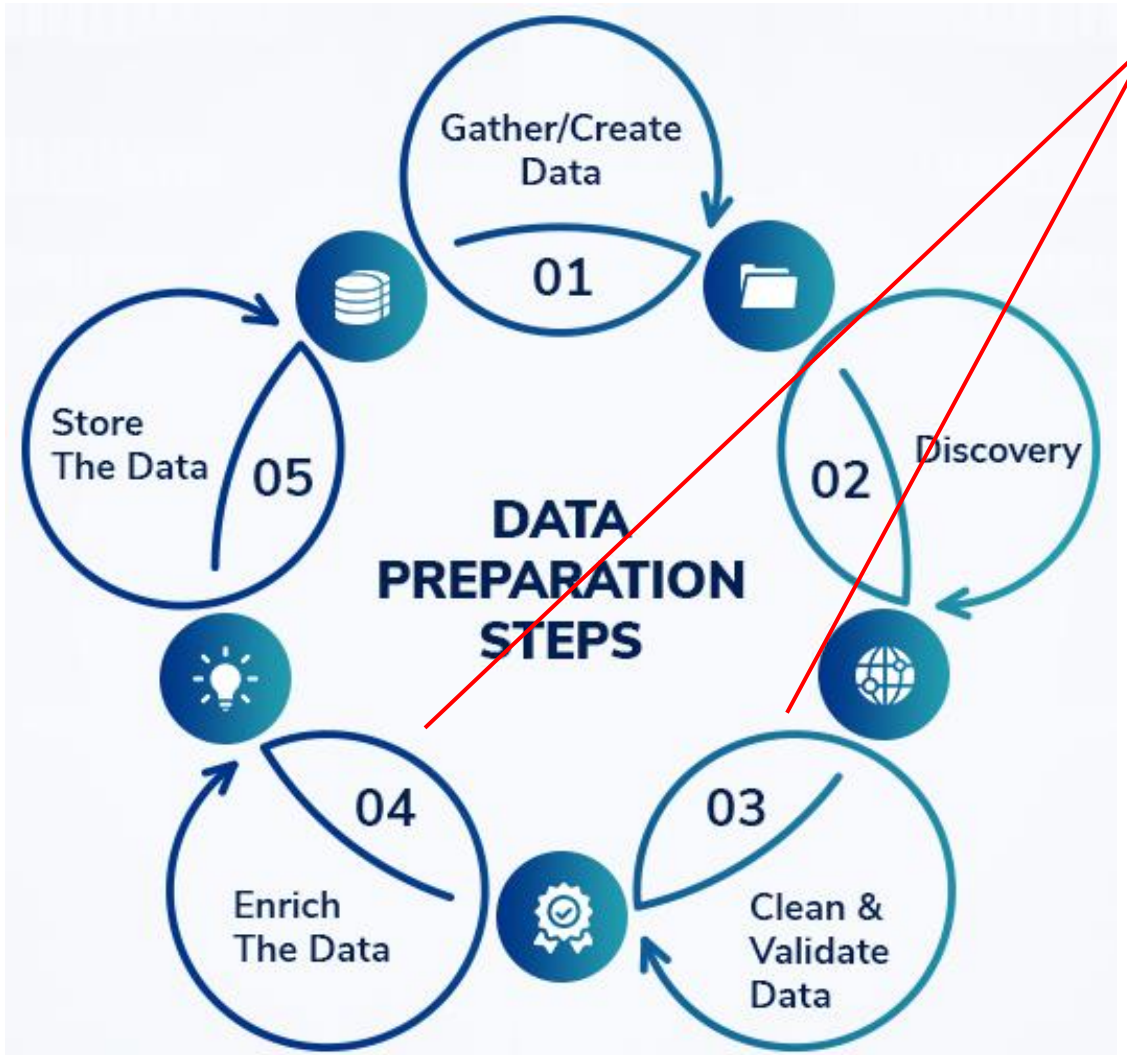
Build your team! **(Deadline: Sep 18)**



Recap: Course Overview

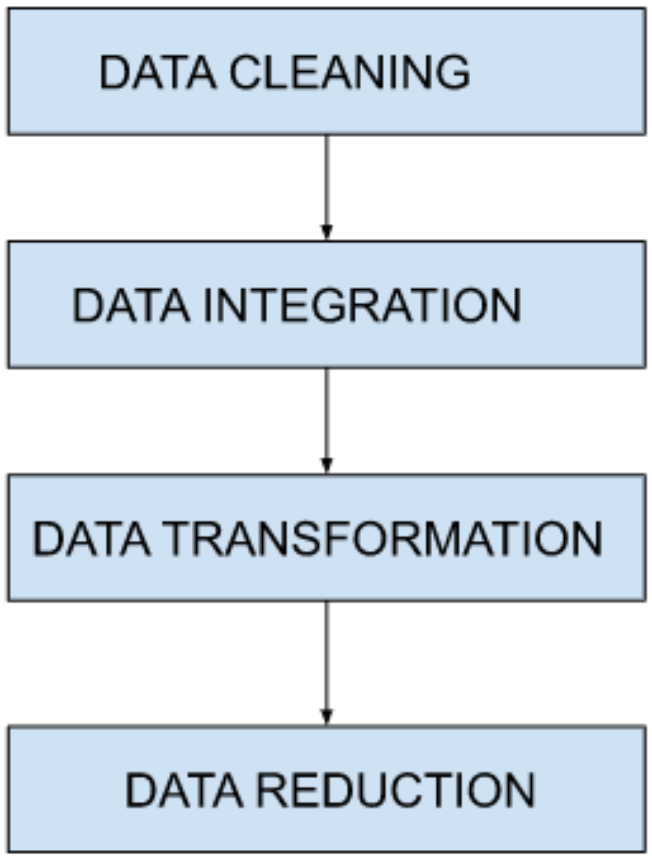


Recap: Course Overview

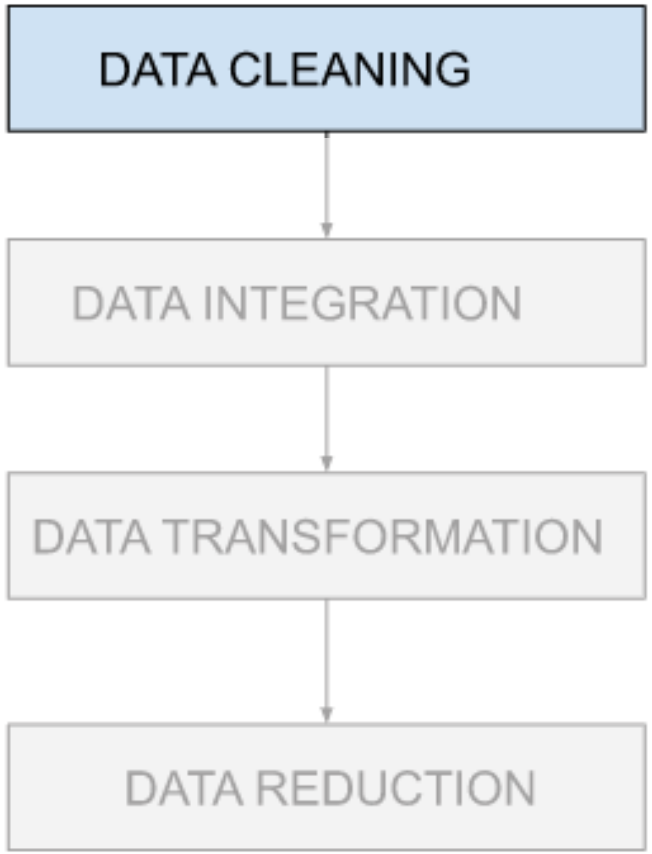


Lecture 4-6

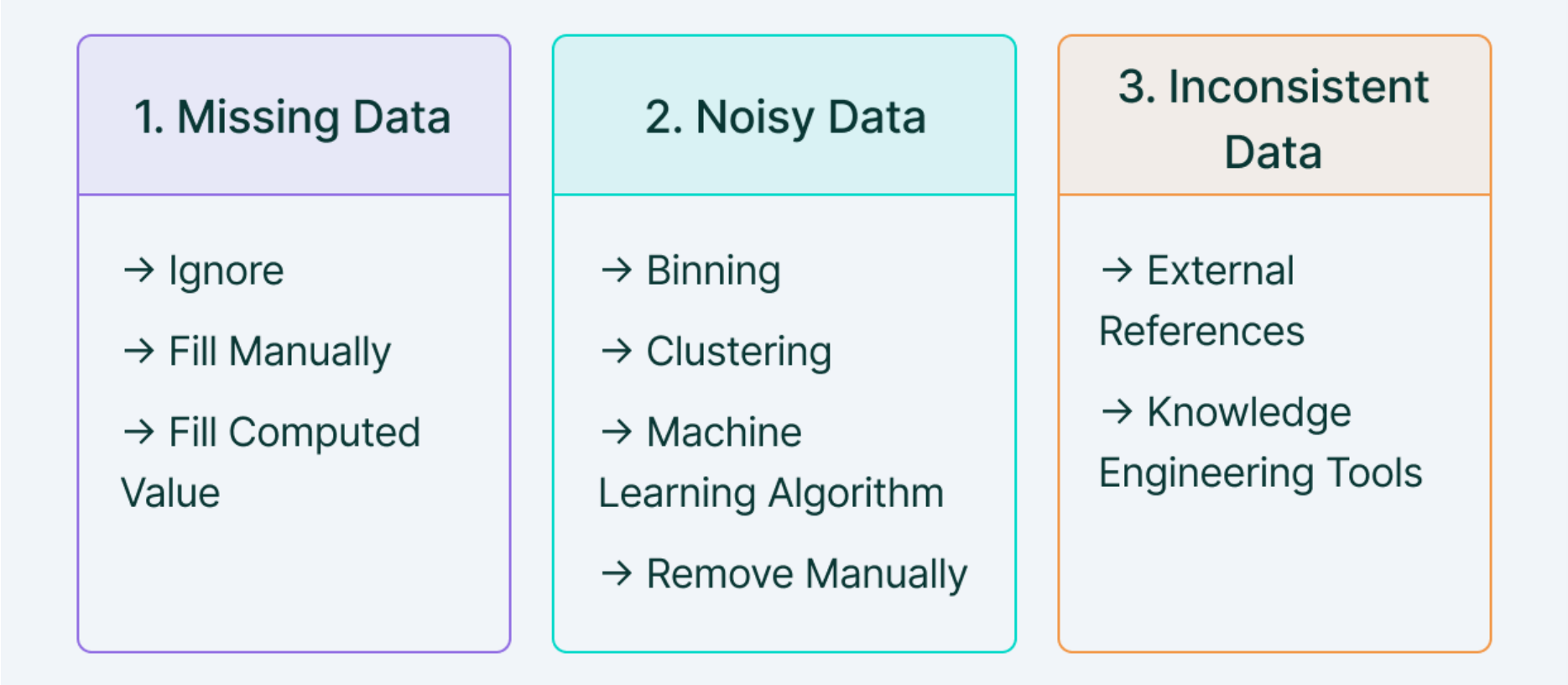
Outline: Data Pre-processing



Outline: Data Pre-processing



Data Cleaning

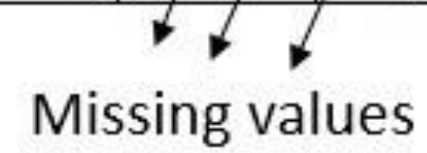


Missing Data

1. Missing Data	2. Noisy Data	3. Inconsistent Data
→ Ignore → Fill Manually → Fill Computed Value	→ Binning → Clustering → Machine Learning Algorithm → Remove Manually	→ External References → Knowledge Engineering Tools

Missing Data

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX		Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX		Mid
6	CA	76600	High
7	NY	85000	High
8	CA		Entry
9	CT	45000	Entry



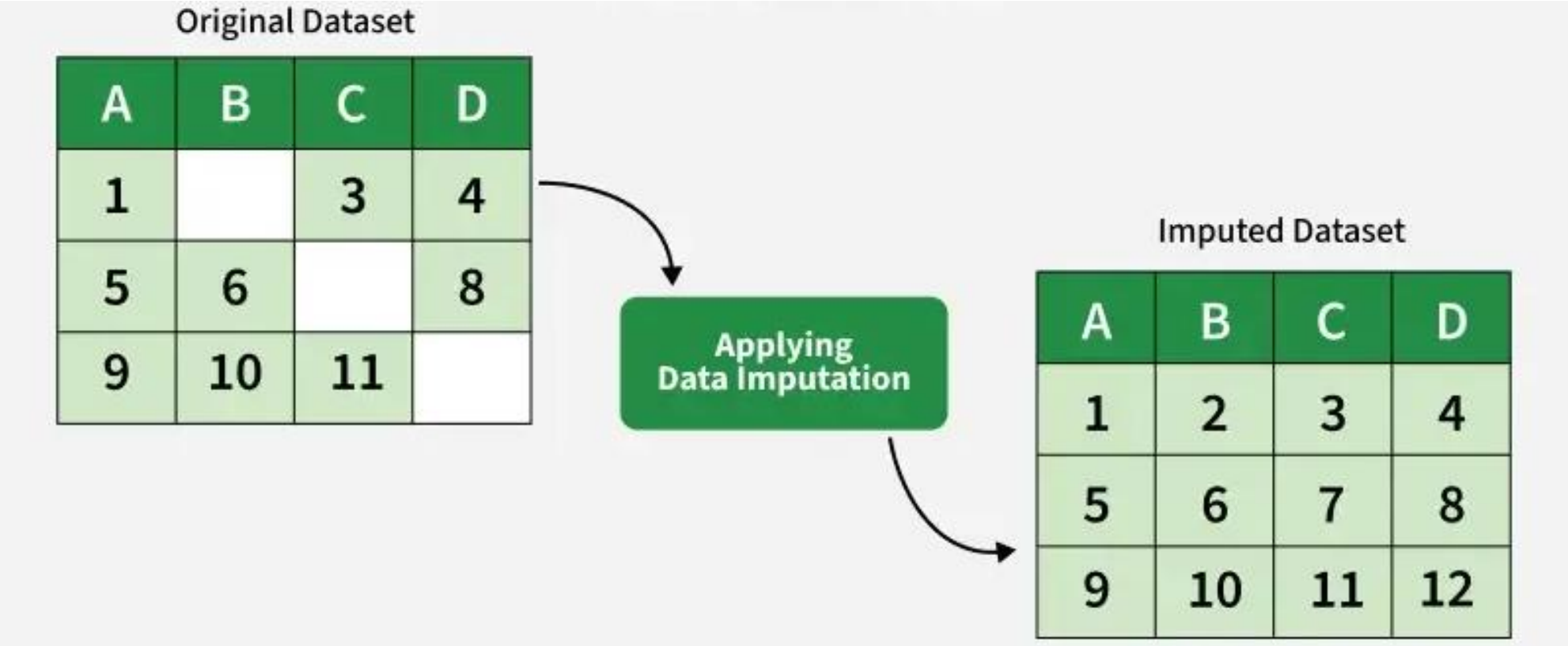
Missing Data

1. **Ignore** the tuple
2. Use a **global constant**
3. Use a **measure of central tendency**
4. Use the **most probable** value

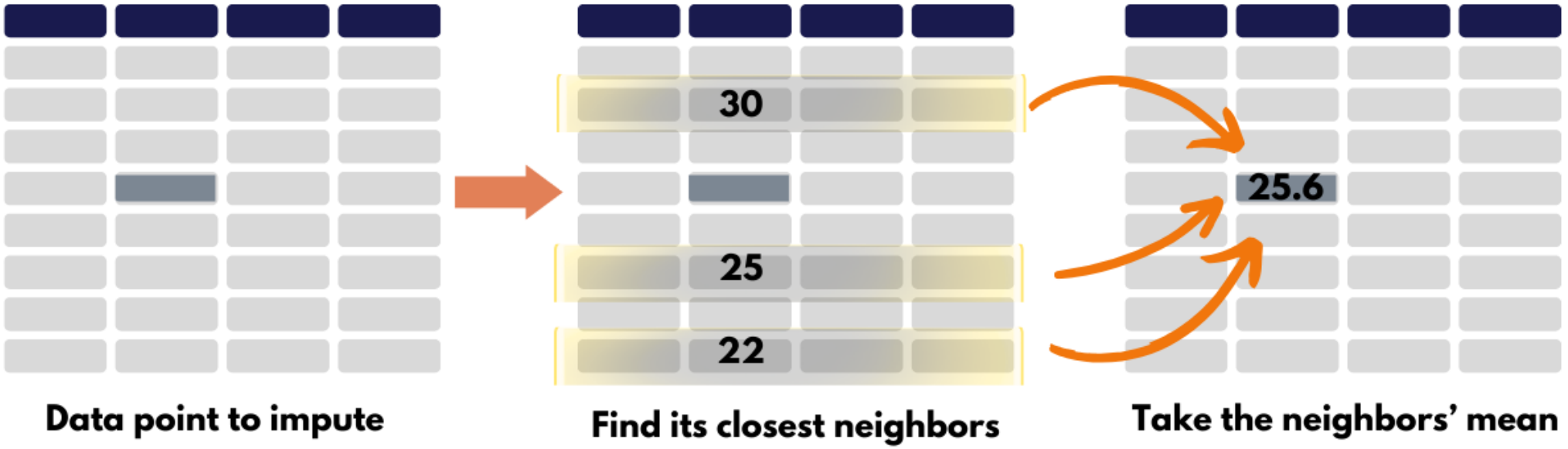
Missing Data

1. **Ignore** the tuple
2. Use a **global constant**
3. Use a **measure of central tendency**
4. Use the **most probable** value

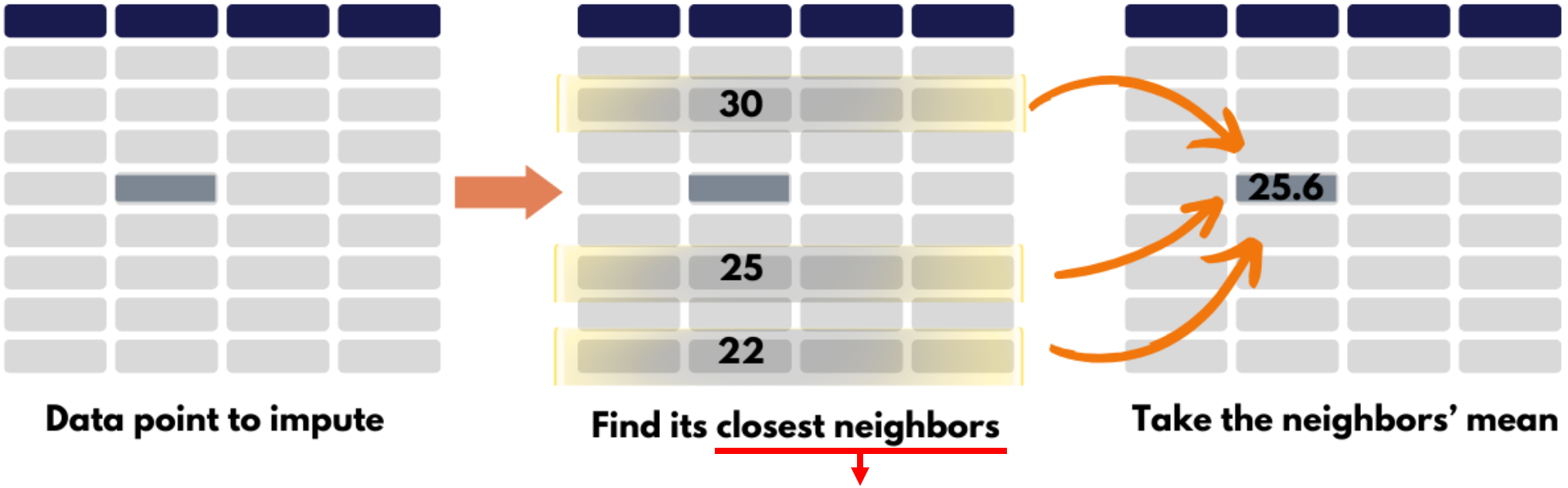
Data Imputation



Data Imputation – Nearest Neighbor

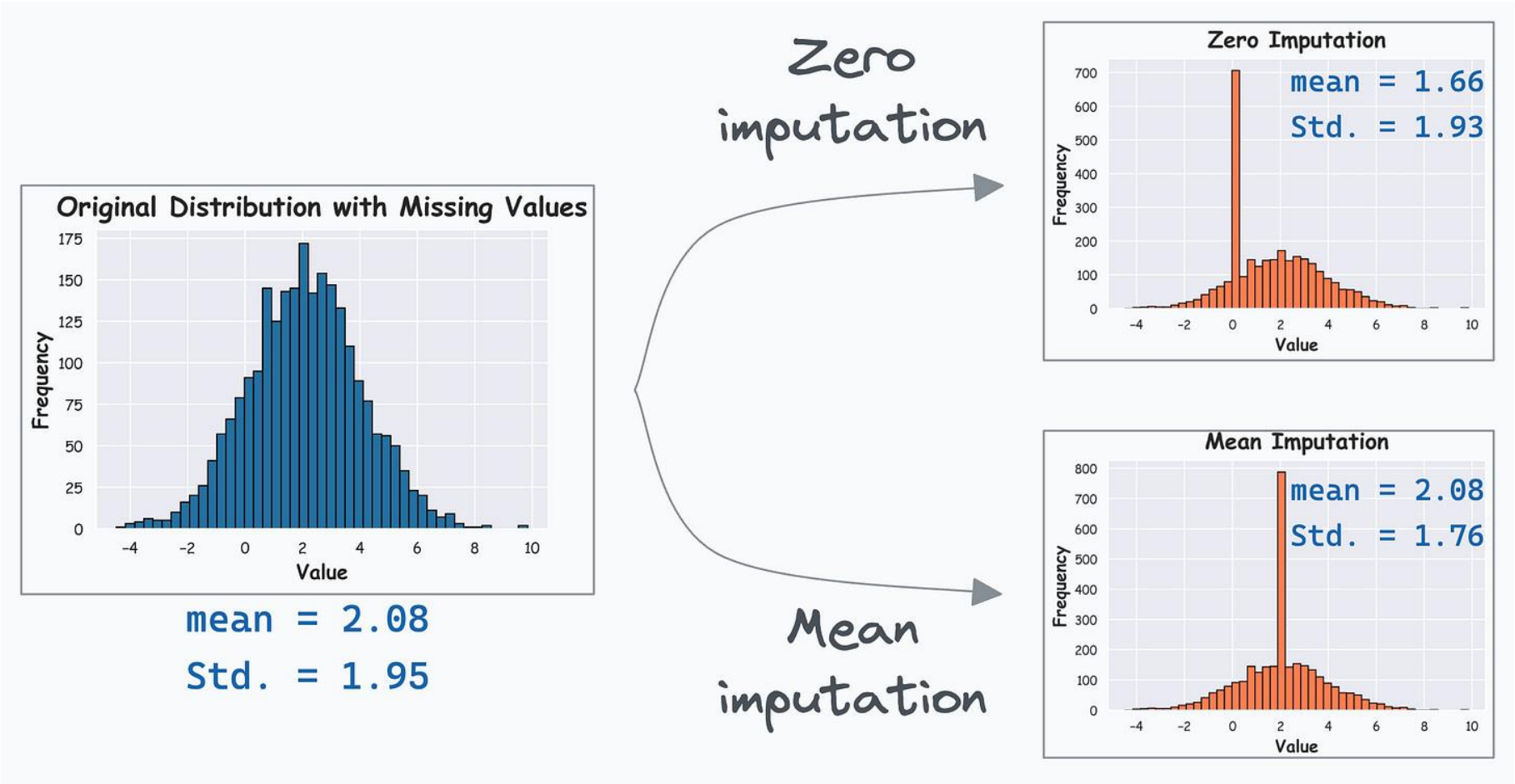


Data Imputation – Nearest Neighbor

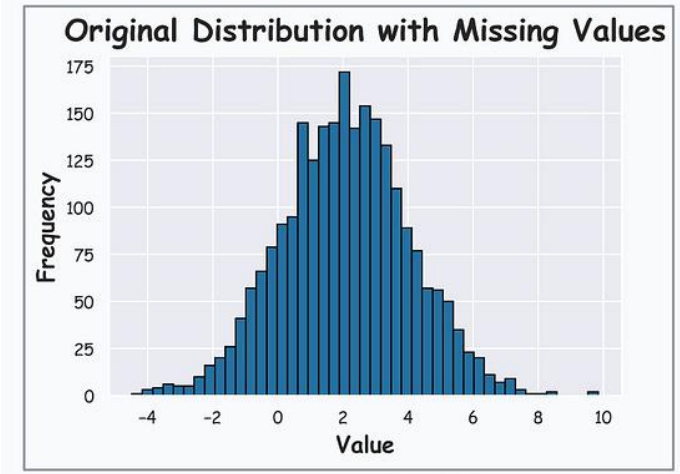


Features: Other values, or their embeddings

Data Imputation – Nearest Neighbor



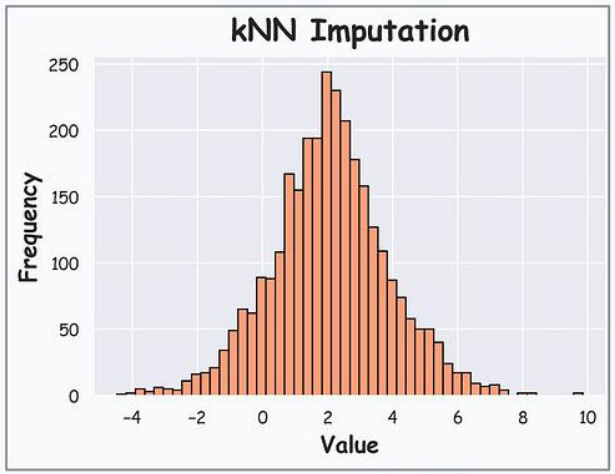
Data Imputation – Nearest Neighbor



mean = 2.08

Std. = 1.95

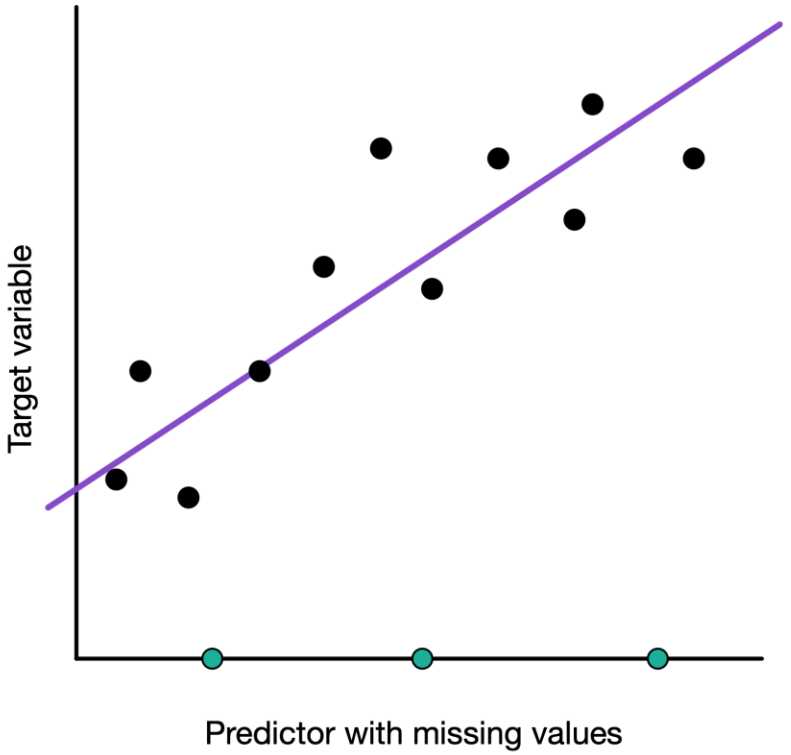
kNN imputation

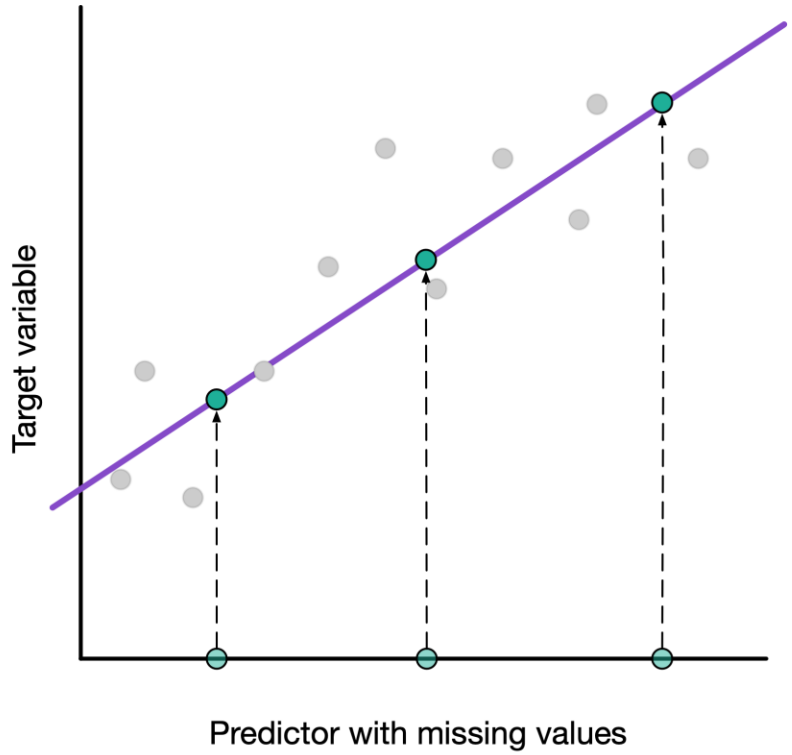


mean = 2.07

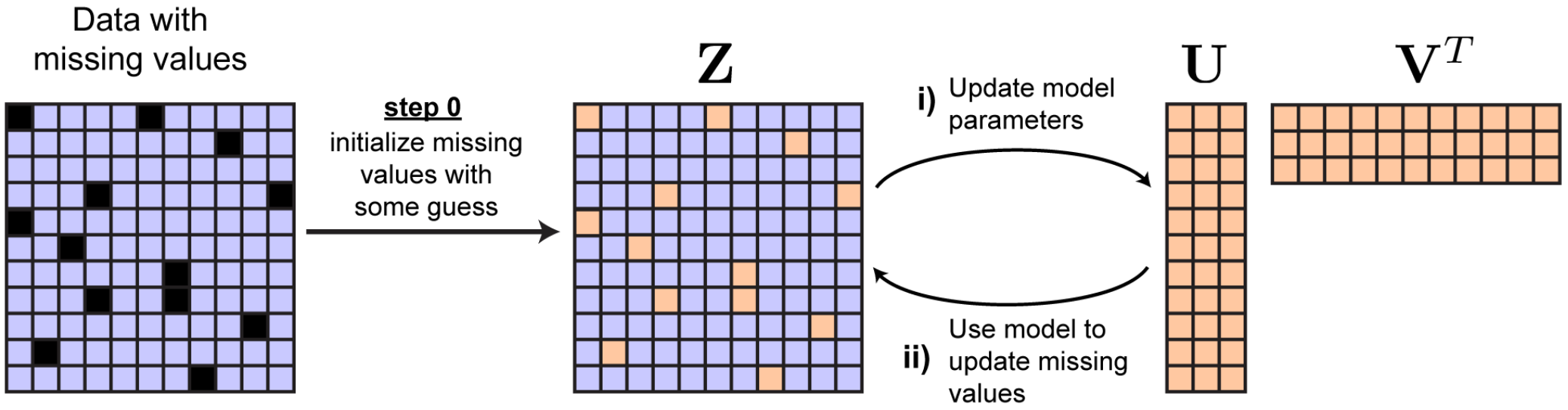
Std. = 1.90

Data Imputation – Regression





Data Imputation – Matrix Factorization

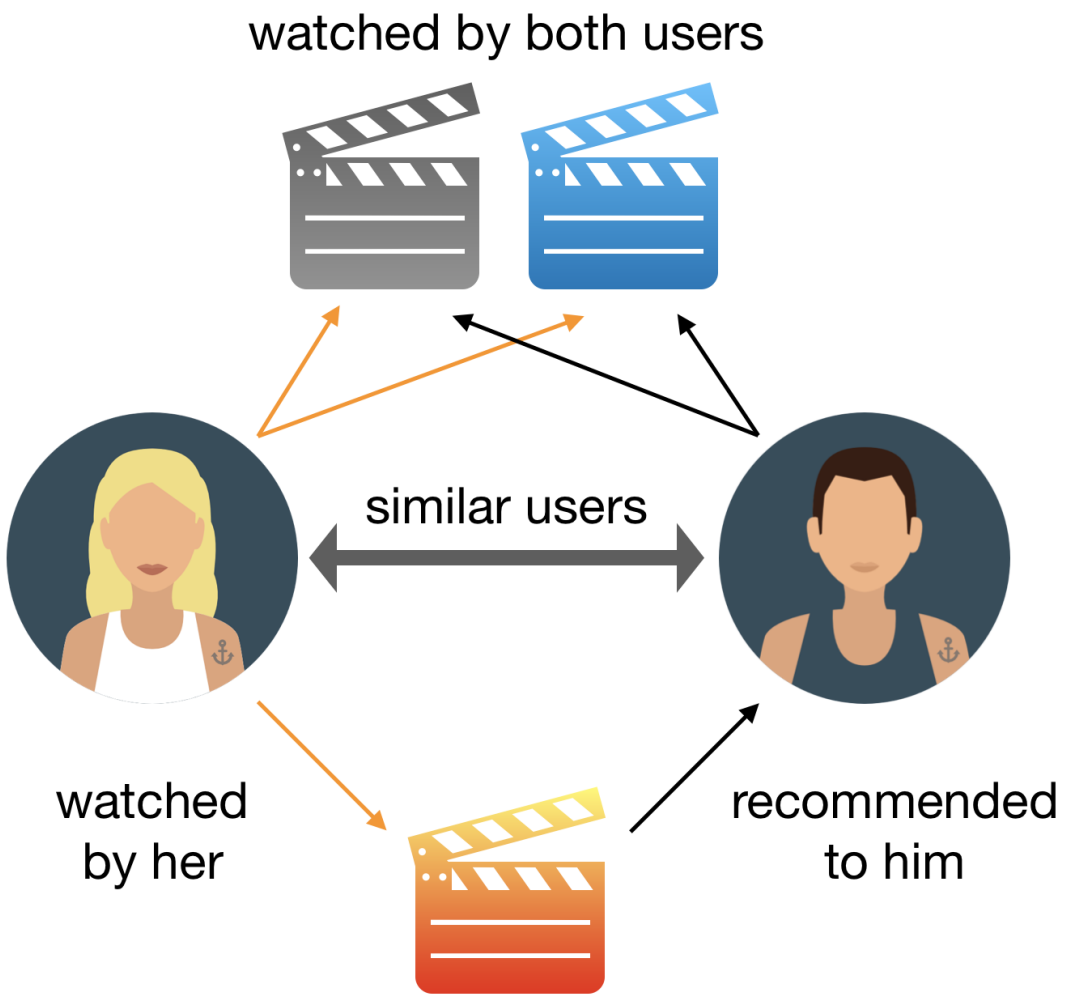


Data Imputation – Matrix Factorization

Case Study: The Netflix Prize

Data Imputation – Matrix Factorization

Case Study: The Netflix Prize



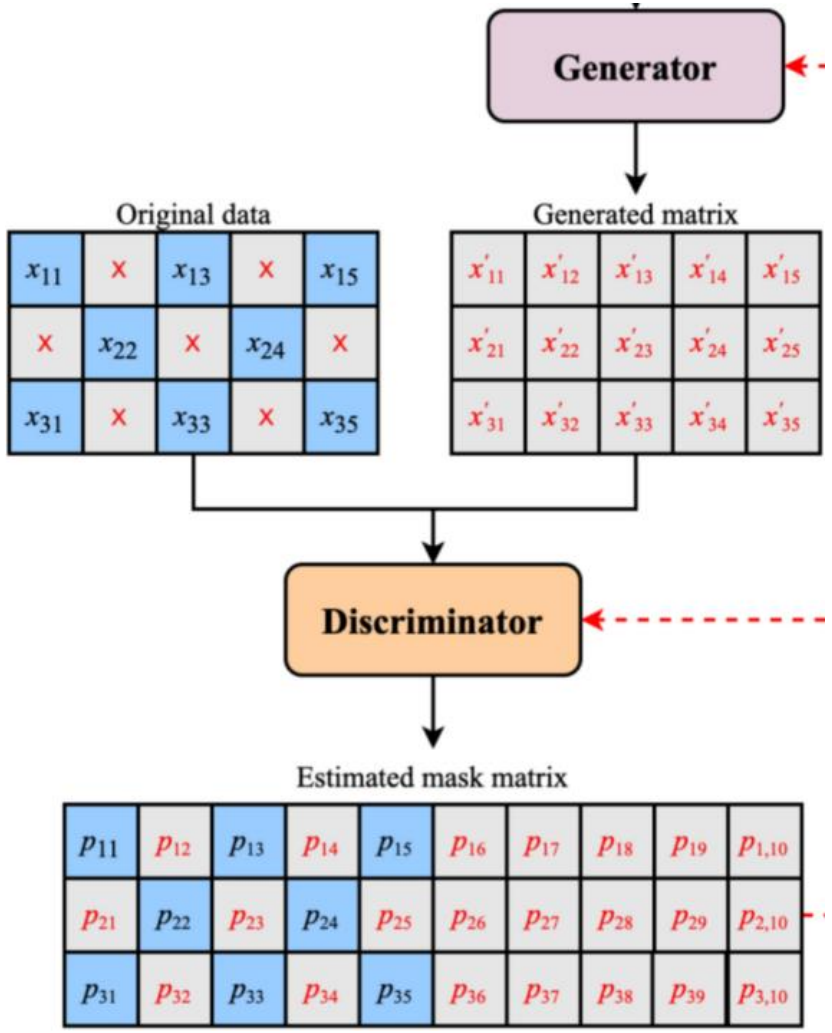
Recommendation system
– Collaborative filtering

Data Imputation – Matrix Factorization

Case Study: The Netflix Prize



Data Imputation – Generative Models



GAIN: Missing Data Imputation using Generative Adversarial Nets. ICML, 2018.

Caveat: Missing at Random?

Caveat: Missing at Random?

Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ		121	91		110		118	93			116	

vs.

Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ							118	93	116	141	97	104

Caveat: Missing at Random?

Missingness is independent

Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ		121	91		110		118	93			116	

vs.

Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ							118	93	116	141	97	104

Missingness is dependent to the data

Caveat: Missing at Random?

Missingness is independent

Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ		121	91		110		118	93			116	

vs.

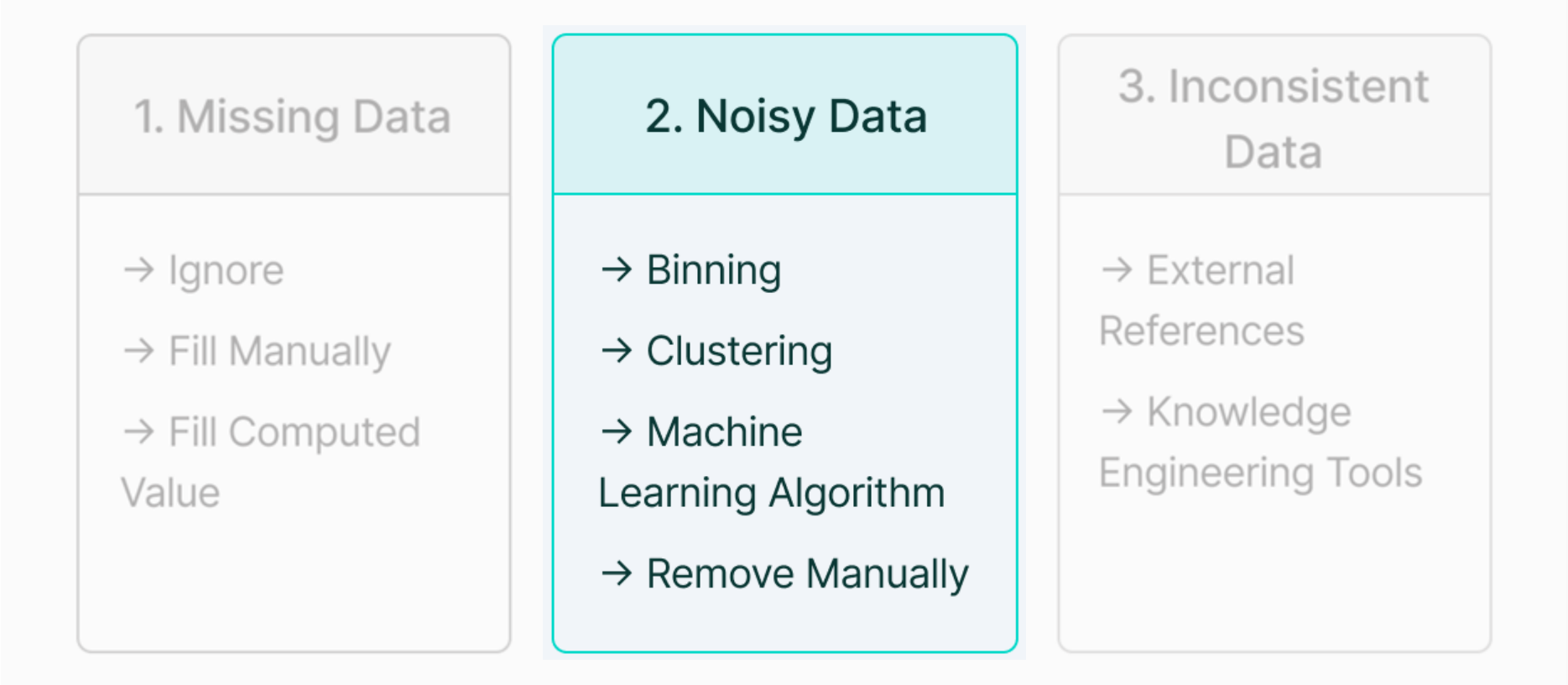
Age	25	26	29	30	30	31	44	46	48	51	52	54
IQ							118	93	116	141	97	104

Missingness is dependent to the data

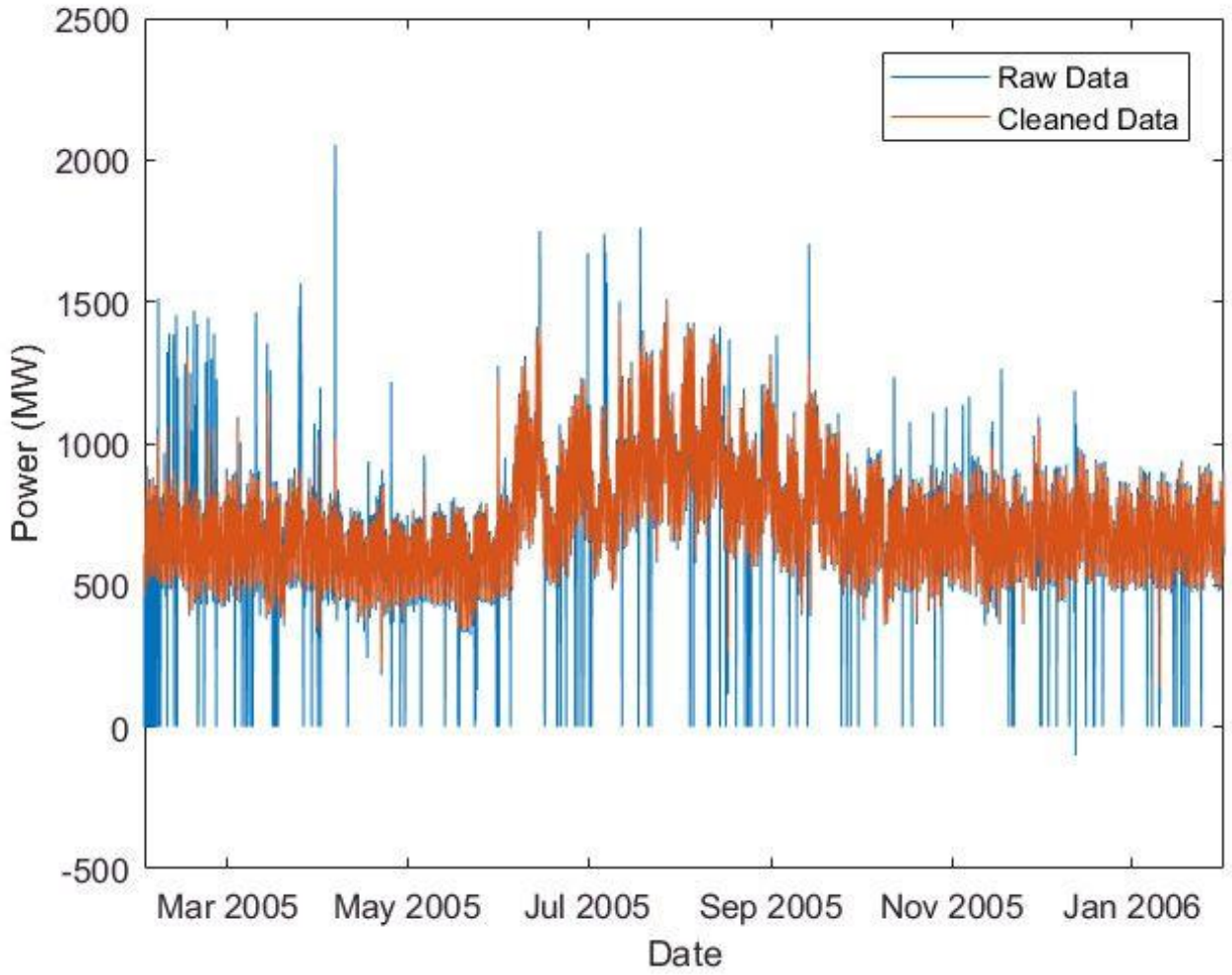


Imputation can be **biased**

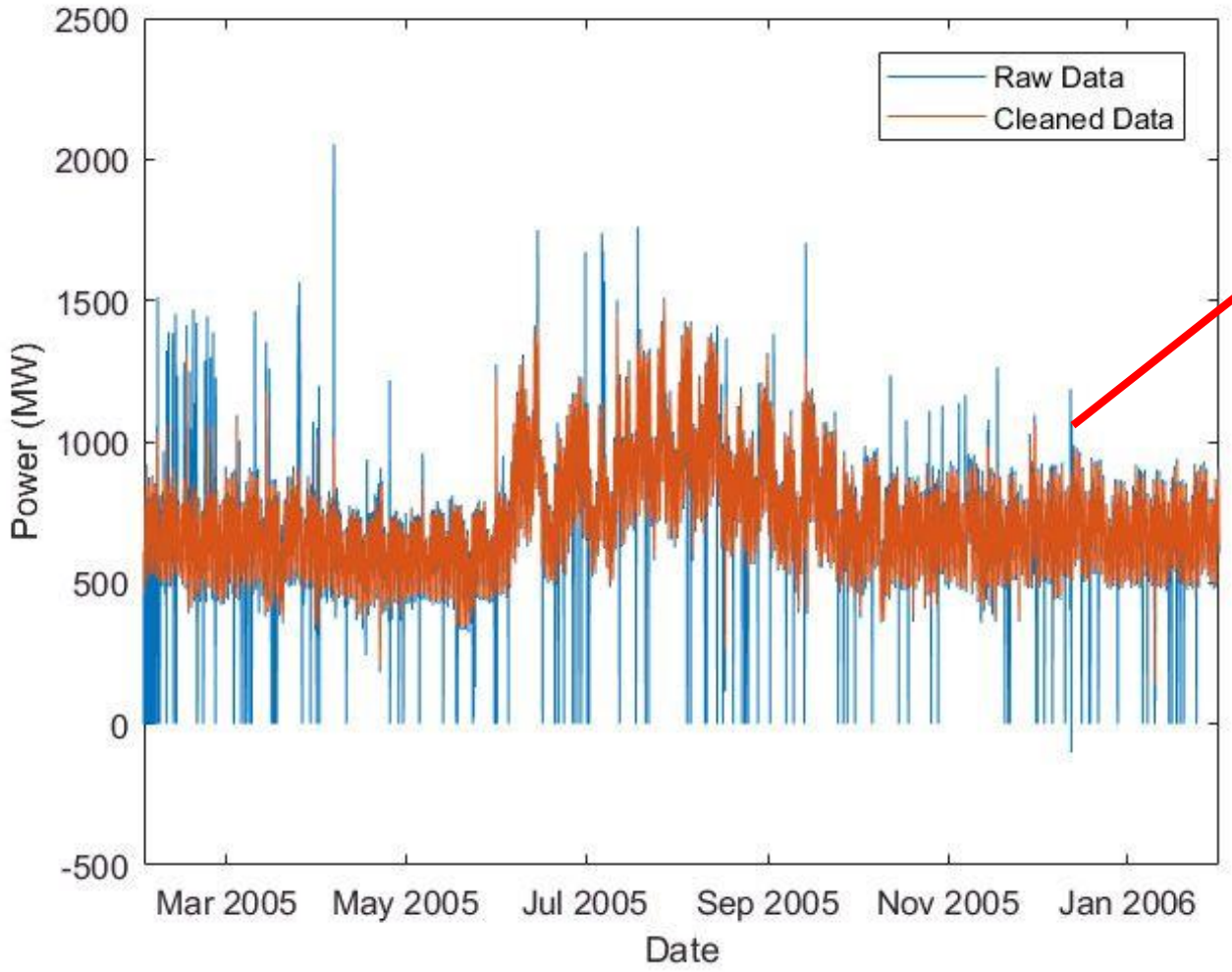
Noisy Data


1. Missing Data
→ Ignore
→ Fill Manually
→ Fill Computed Value
2. Noisy Data
→ Binning
→ Clustering
→ Machine Learning Algorithm
→ Remove Manually
3. Inconsistent Data
→ External References
→ Knowledge Engineering Tools

Noisy Data


2500
2000
1500
1000
500
0
-500
Power (MW)
Raw Data
Cleaned Data
Mar 2005
May 2005
Jul 2005
Sep 2005
Nov 2005
Jan 2006
Date

Noisy Data



Smoothing

Data Smoothing – Binning

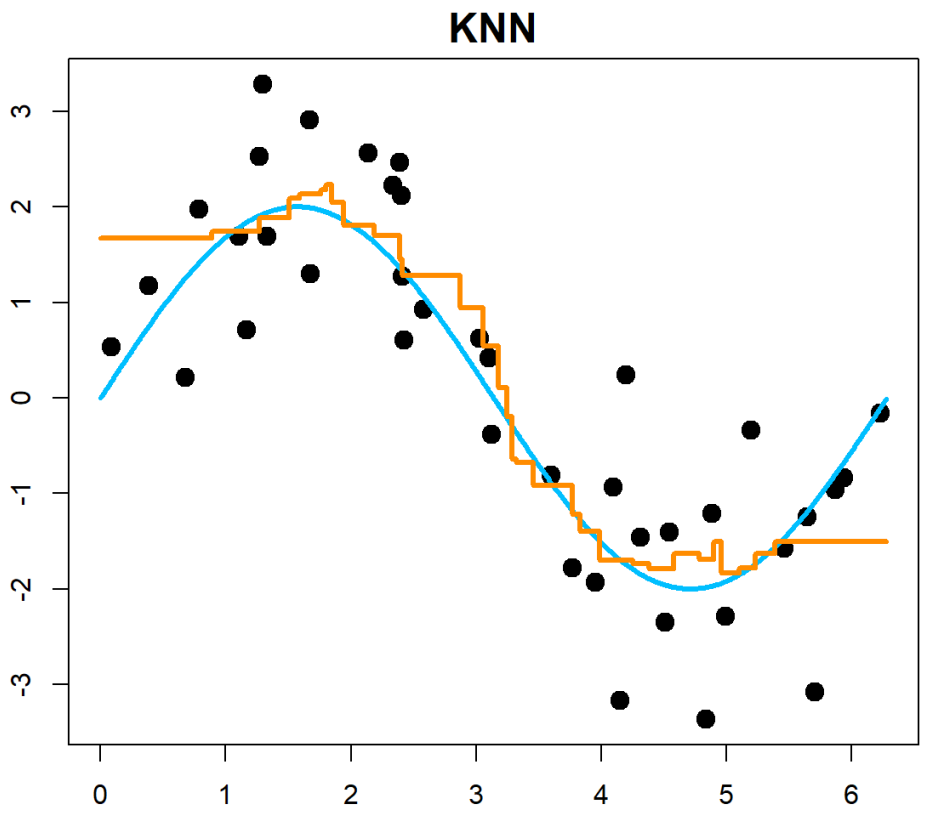
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

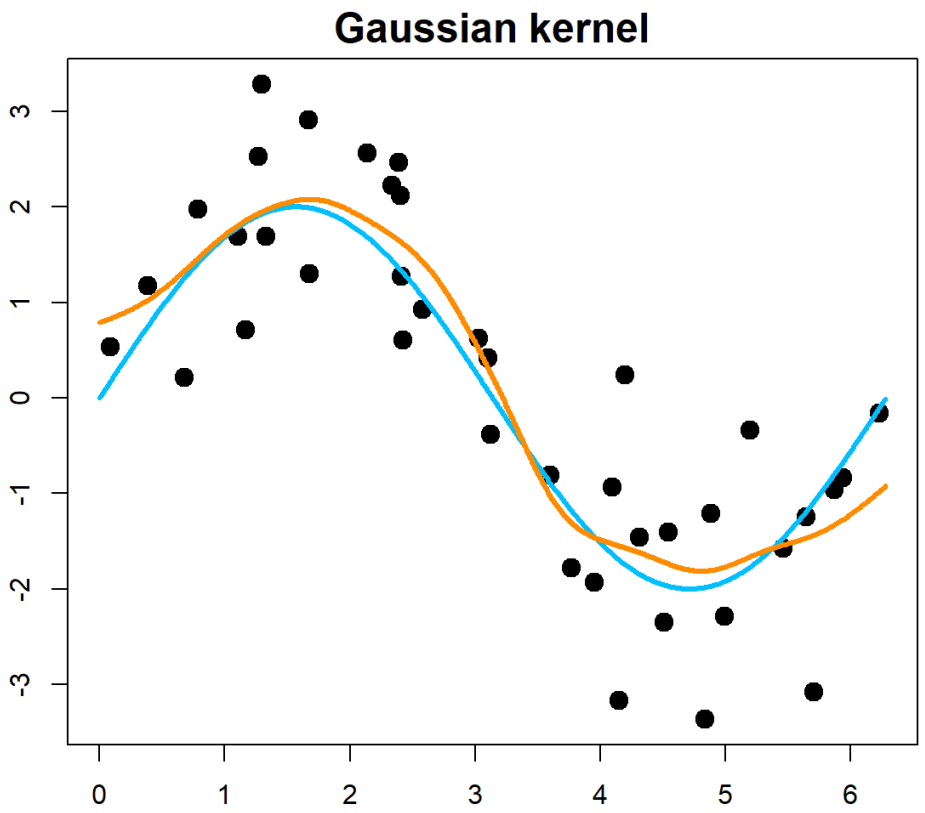
- Bin 1: 4, 8, 9, 15 mean: 9
- Bin 2: 21, 21, 24, 25 mean: 23
- Bin 3: 26, 28, 29, 34 mean: 29

Smoothed Data by bin means:

9, 9, 9, 9, 23, 23, 23, 23, 29, 29, 29, 29

Data Smoothing – Weighted Average





Data Smoothing – Moving Average

Simple $\frac{A_1 + A_2 + ... + A_n}{n}$

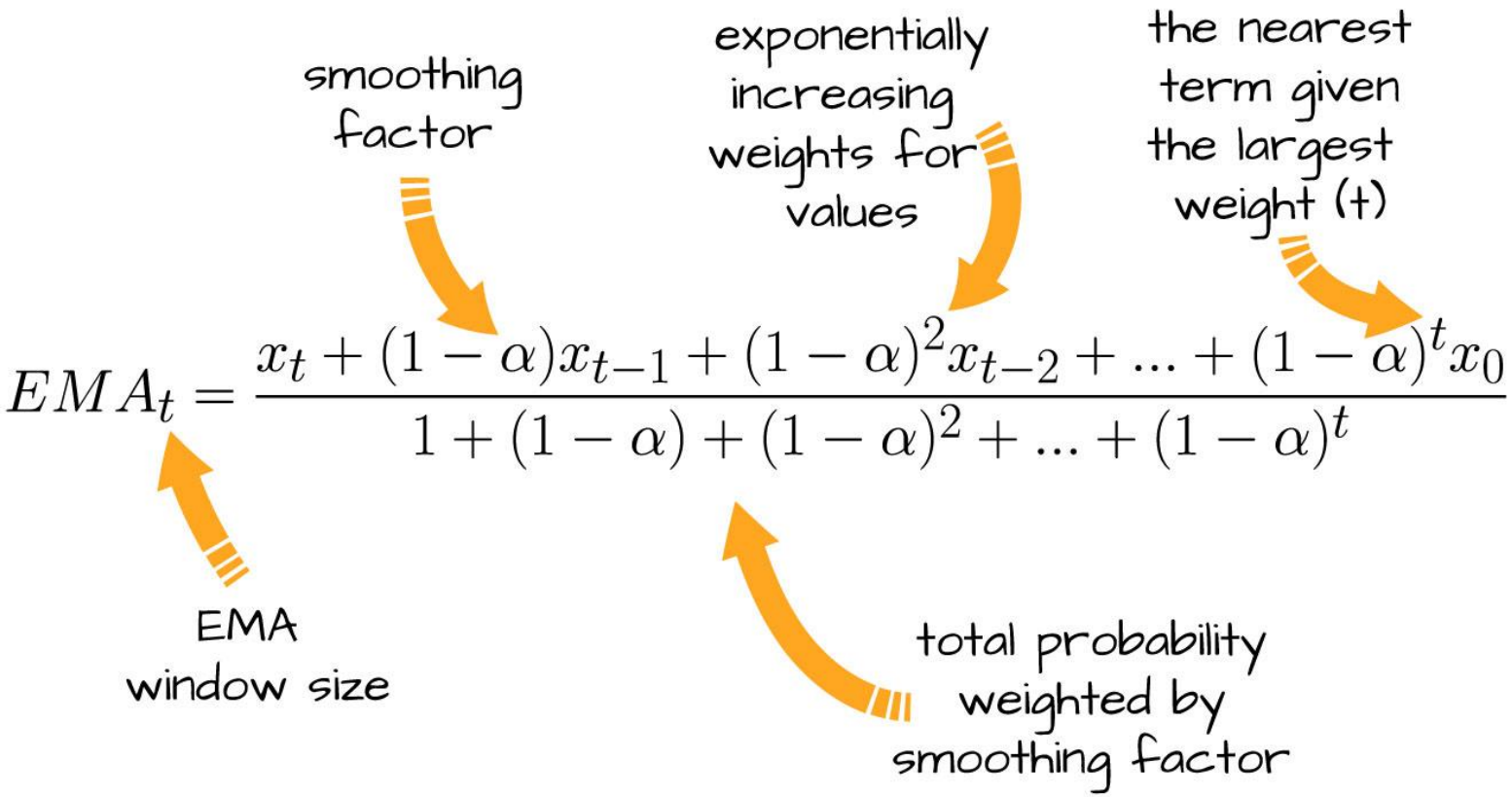


Data Smoothing – Moving Average

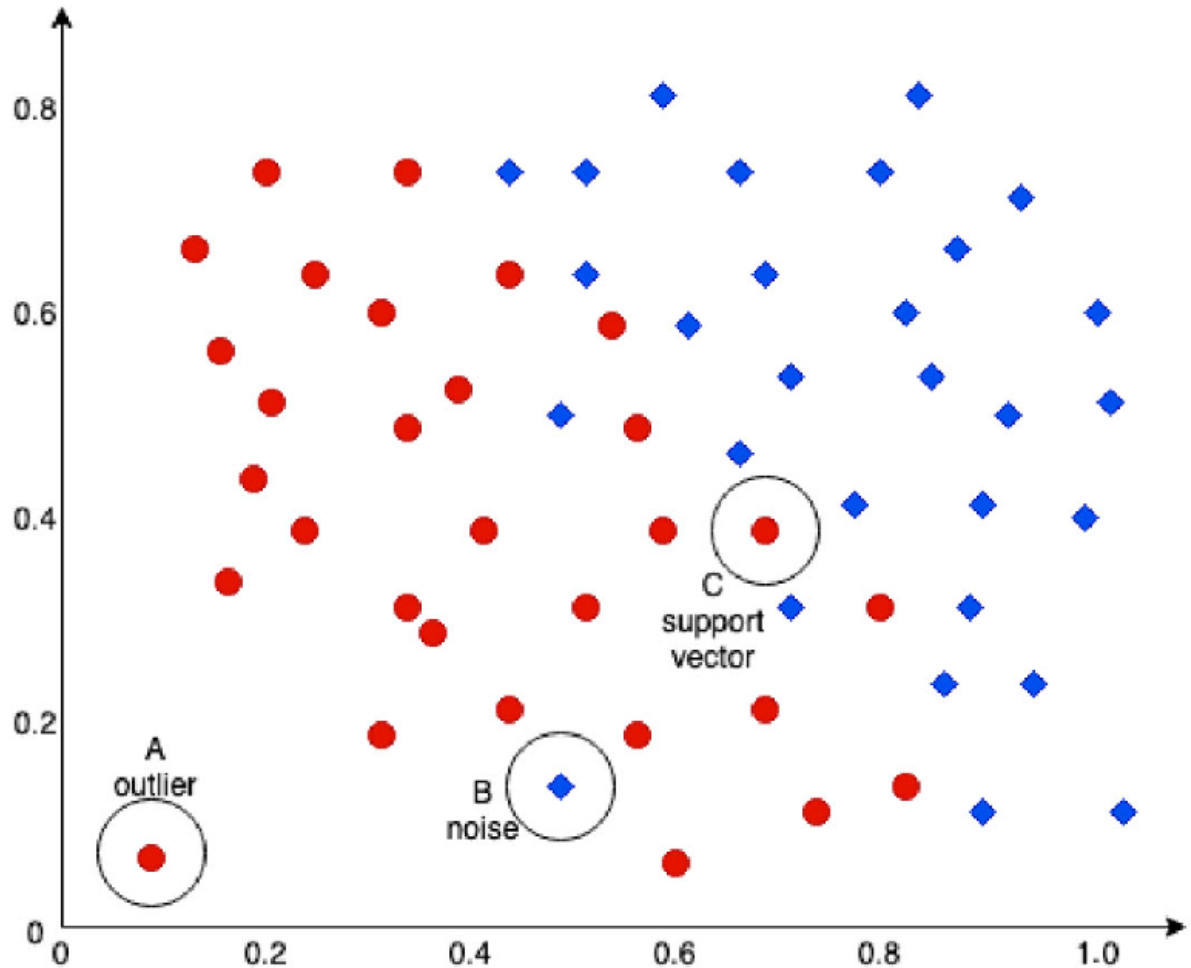
Exponential Simple $\frac{A_1 + A_2 + ... + A_n}{n}$



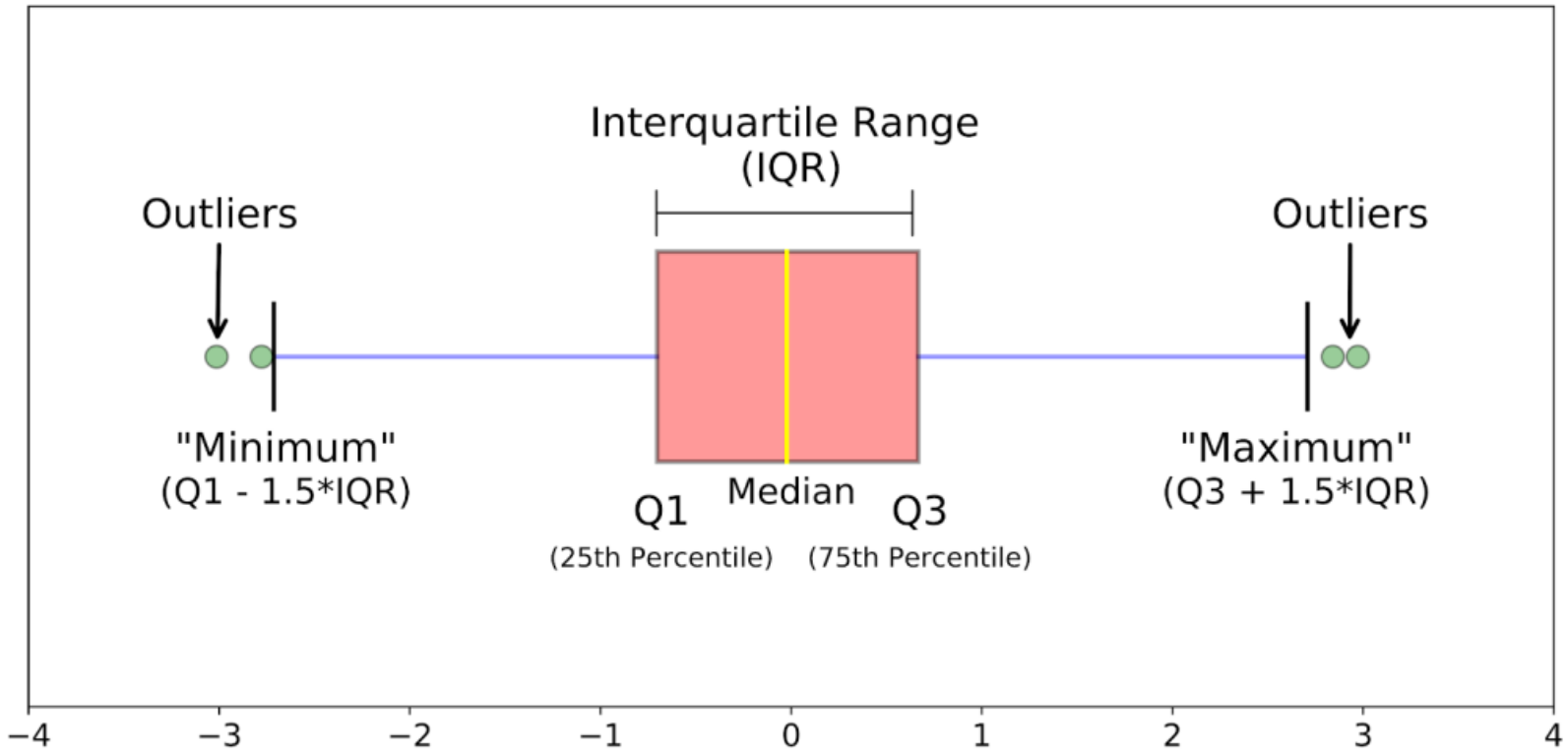
Data Smoothing – Moving Average



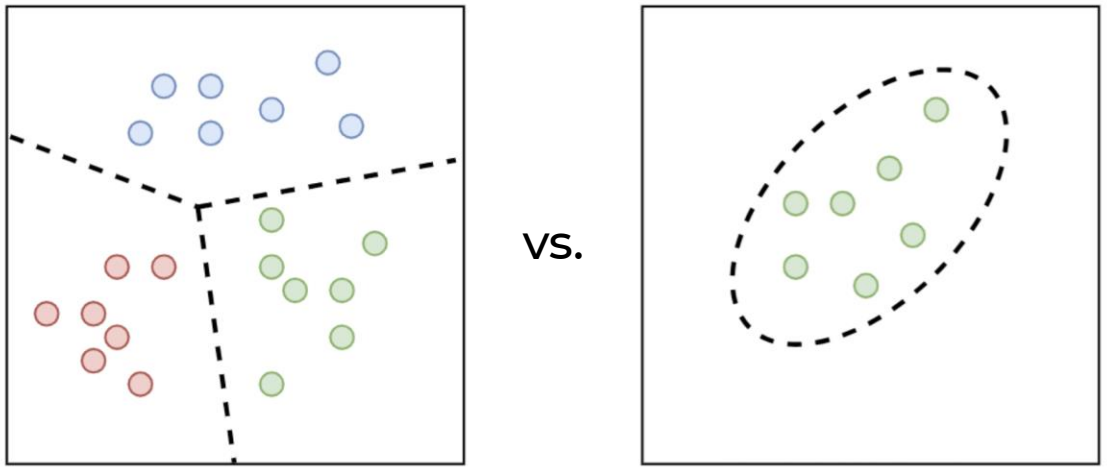
Noisy Data vs. Outlier



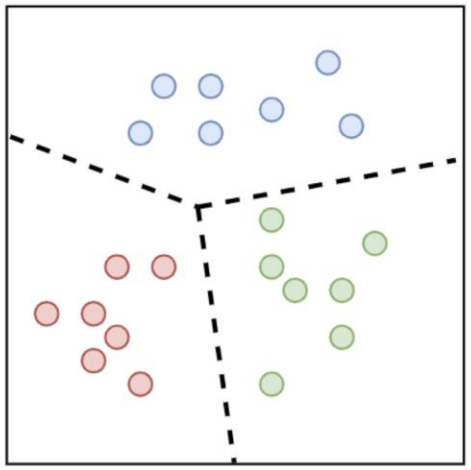
Outlier Detection – Statistics



Outlier Detection – One-class Classification

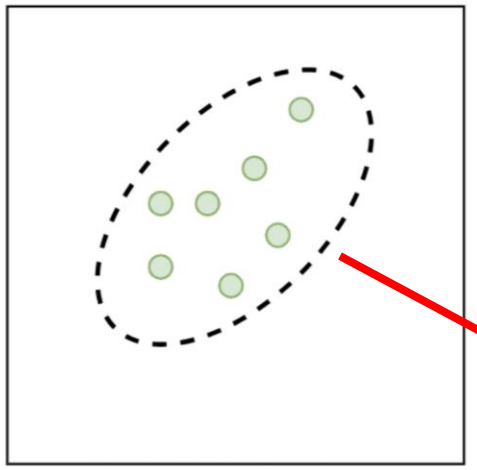


Outlier Detection – One-class Classification



Multi-class Classification

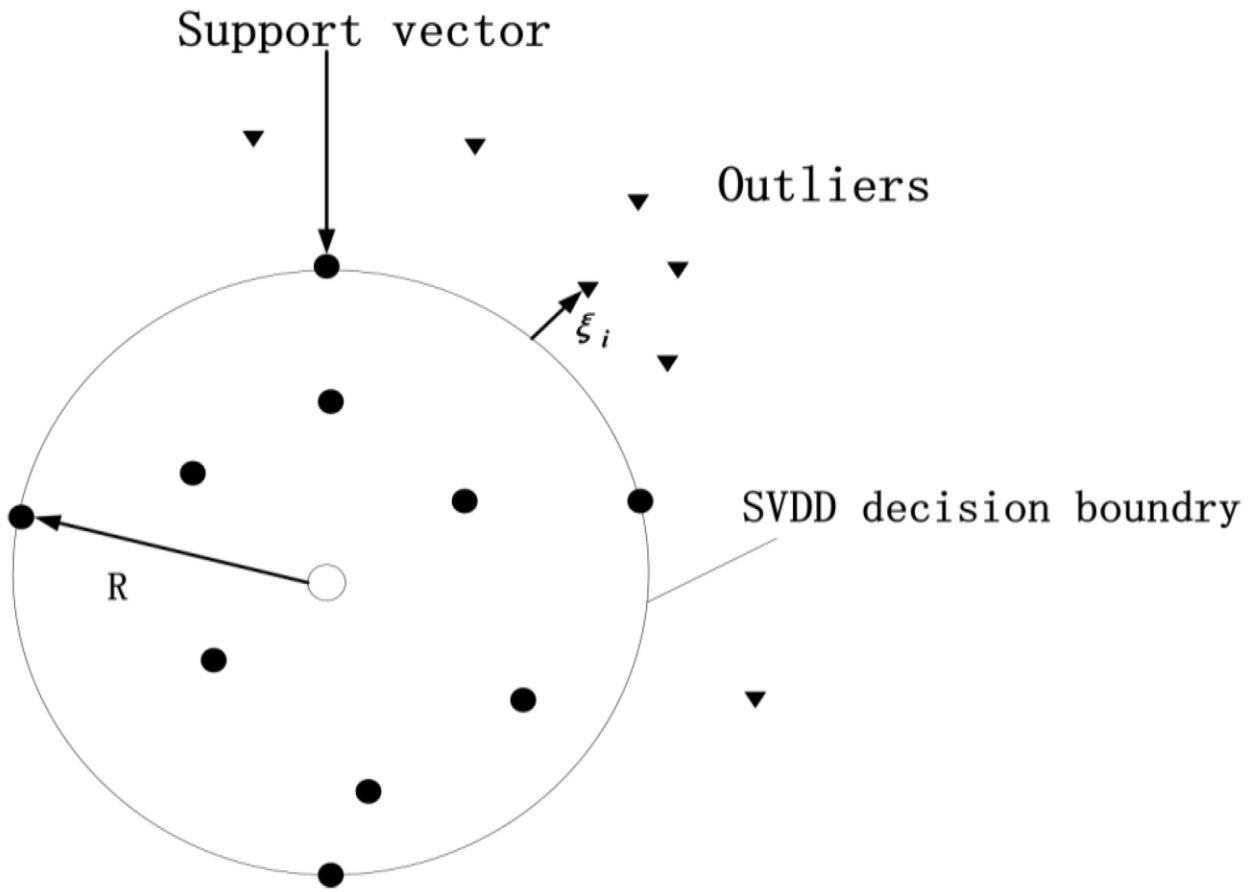
vs.



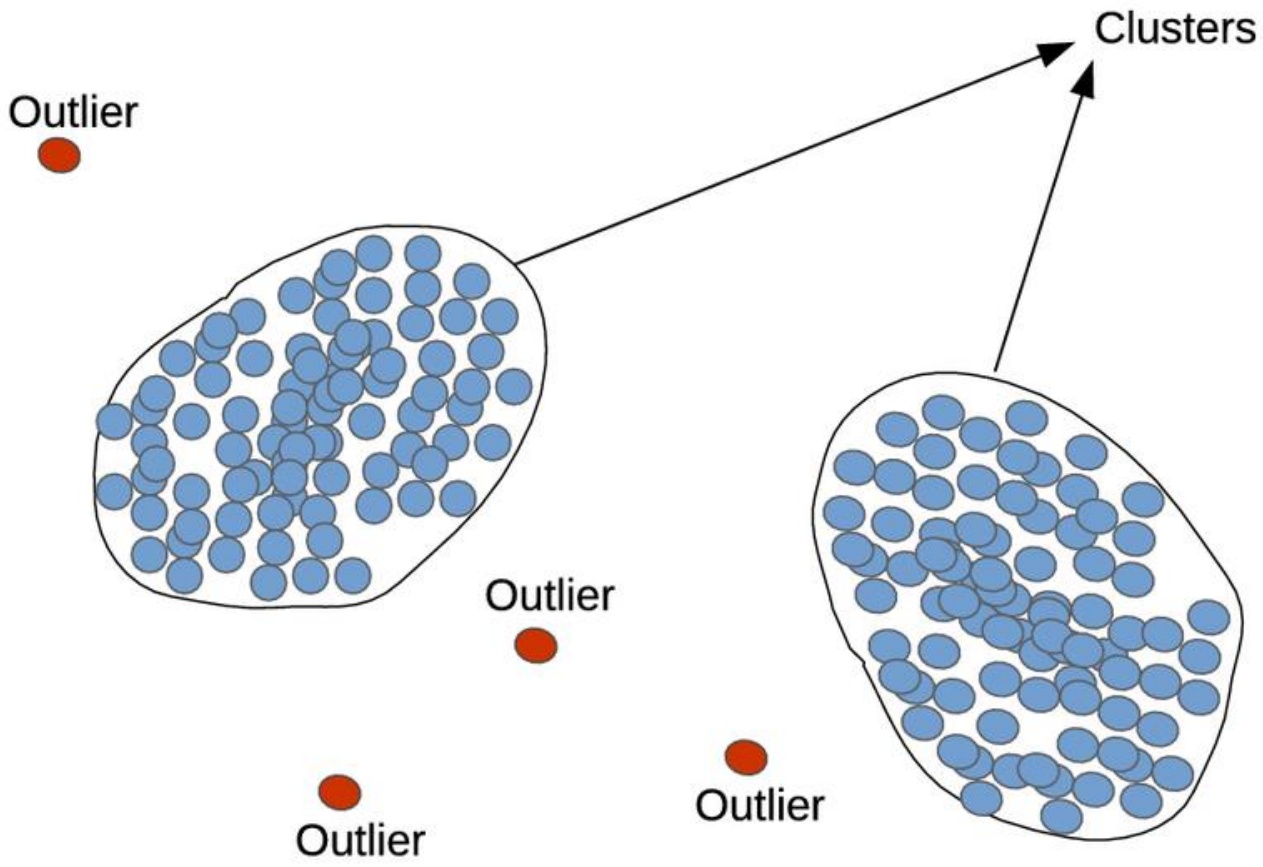
One Class Classification

Given only “normal” data, detect outliers

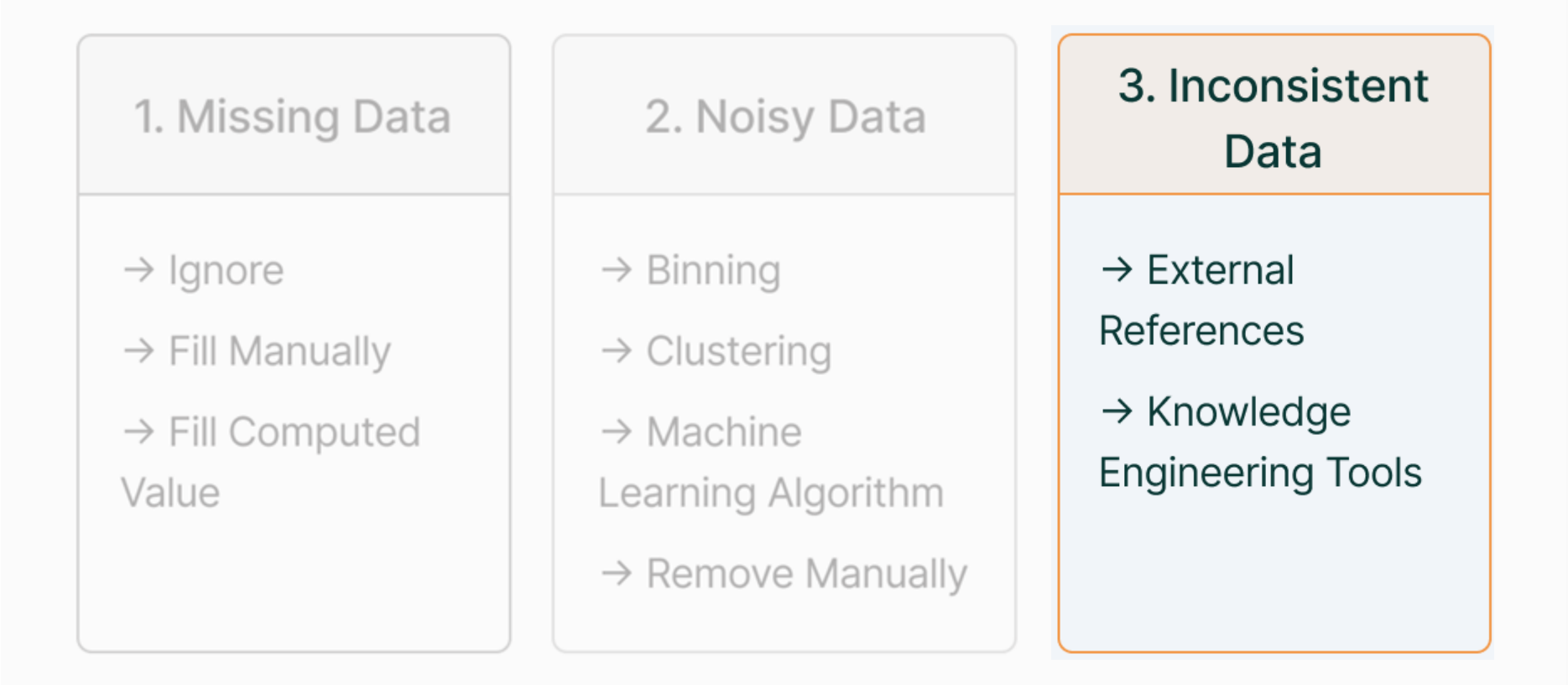
Outlier Detection – One-class Classification



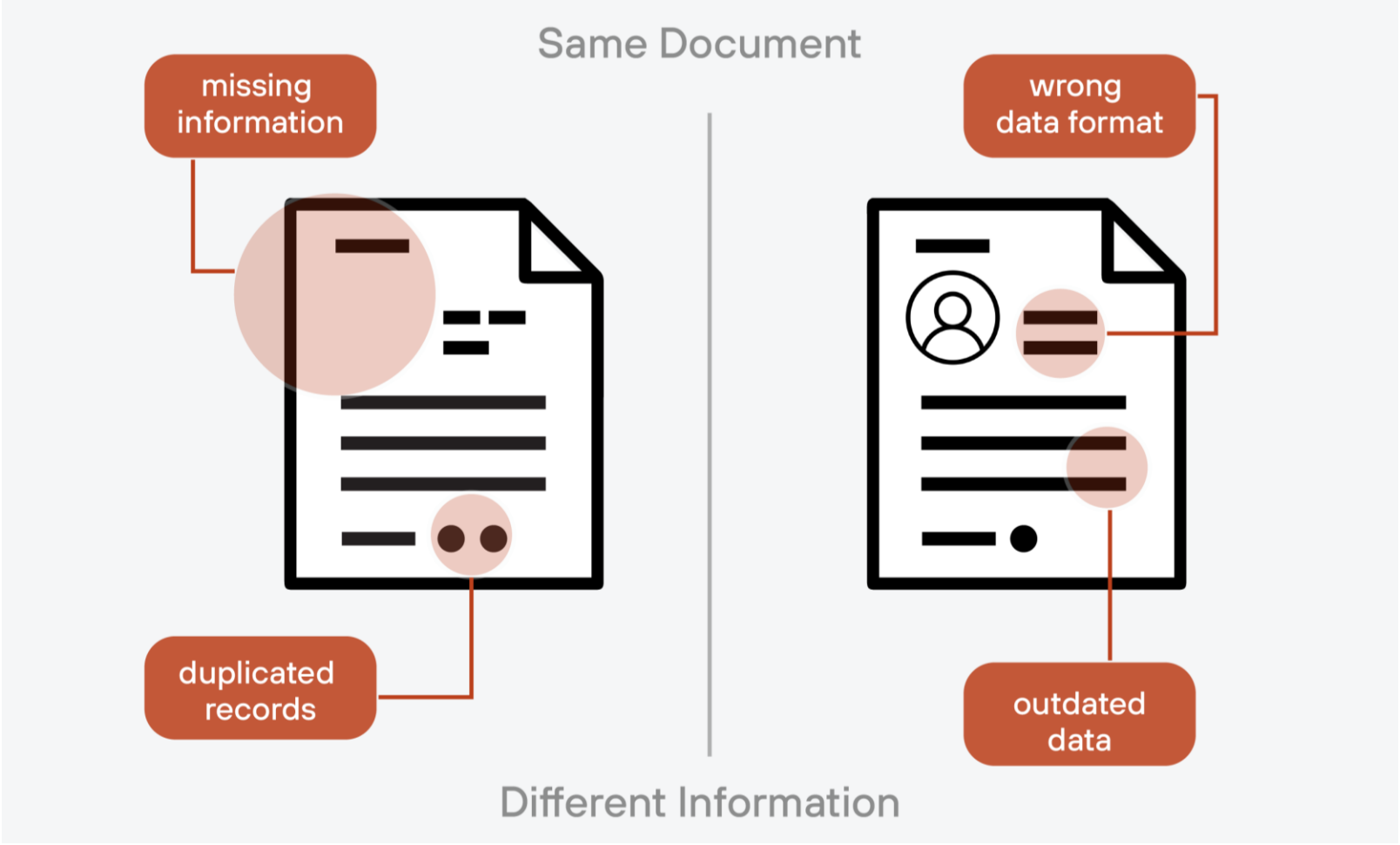
Outlier Detection – Density Estimation



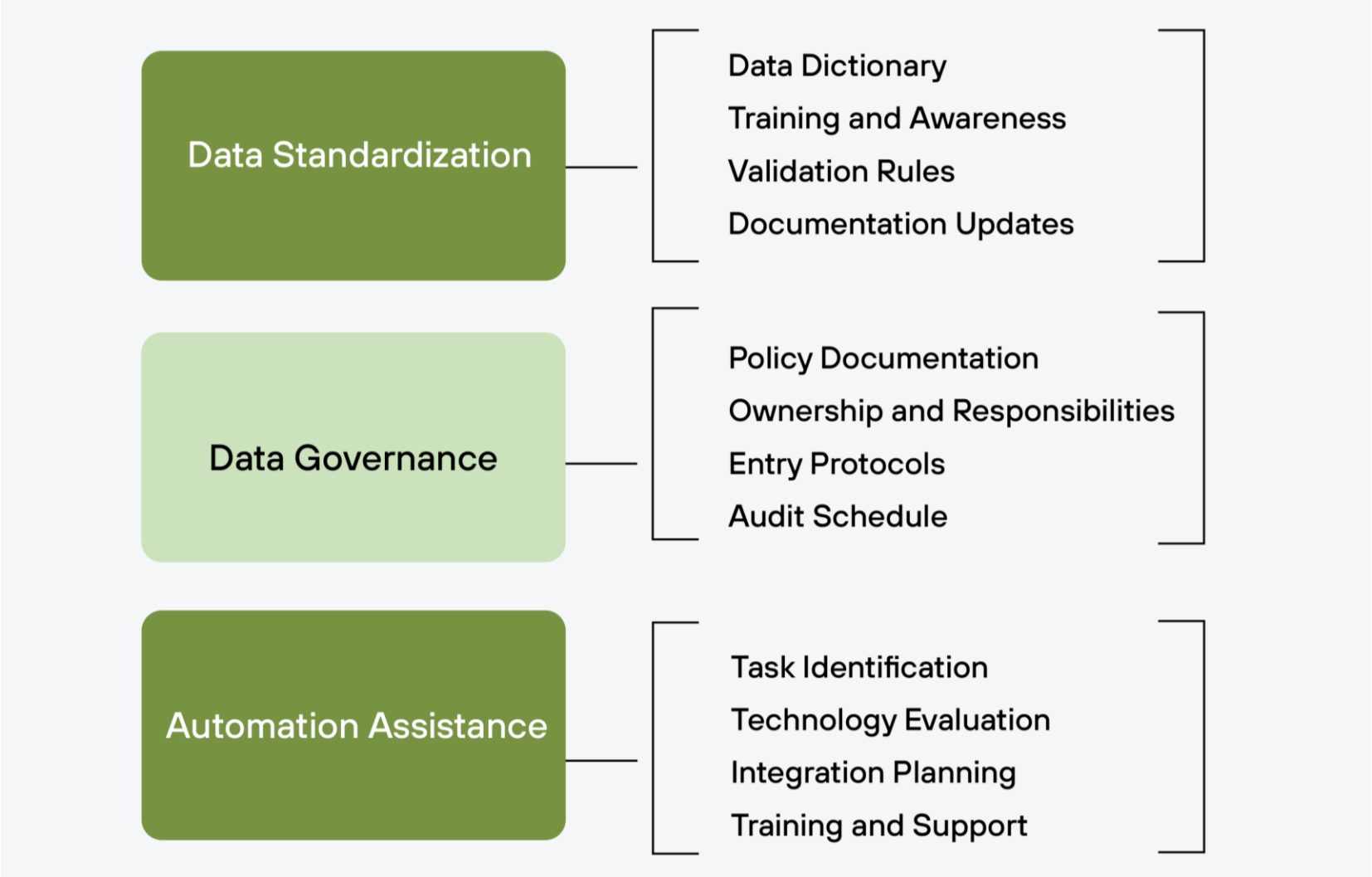
Inconsistent Data



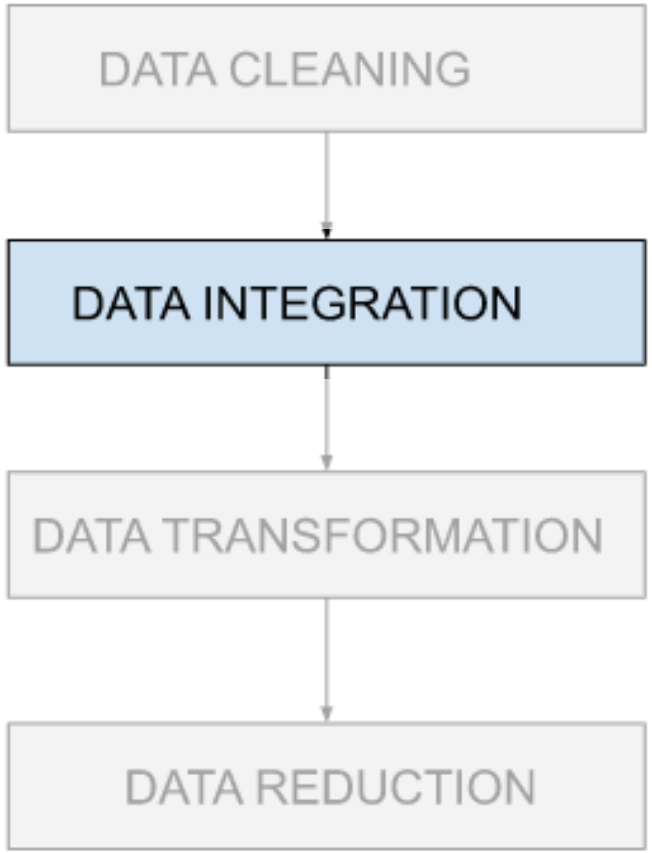
Inconsistent Data



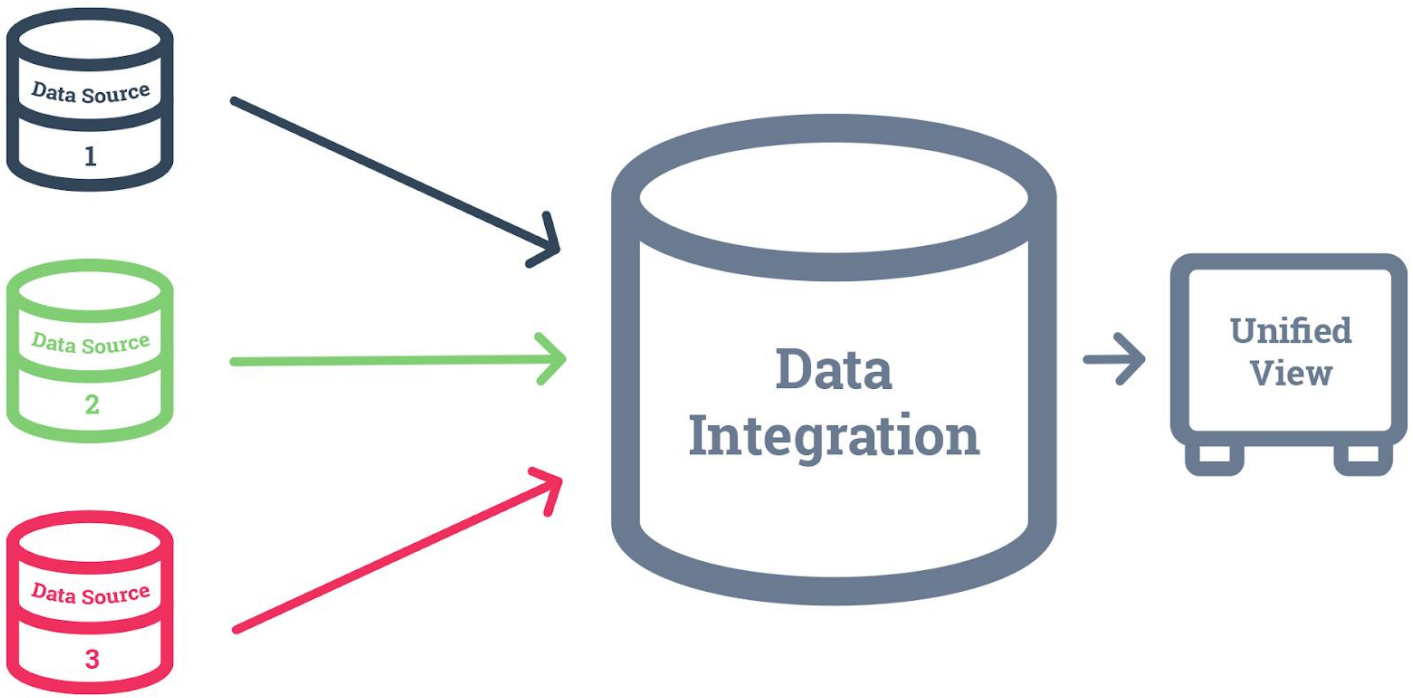
Mitigating Inconsistent Data



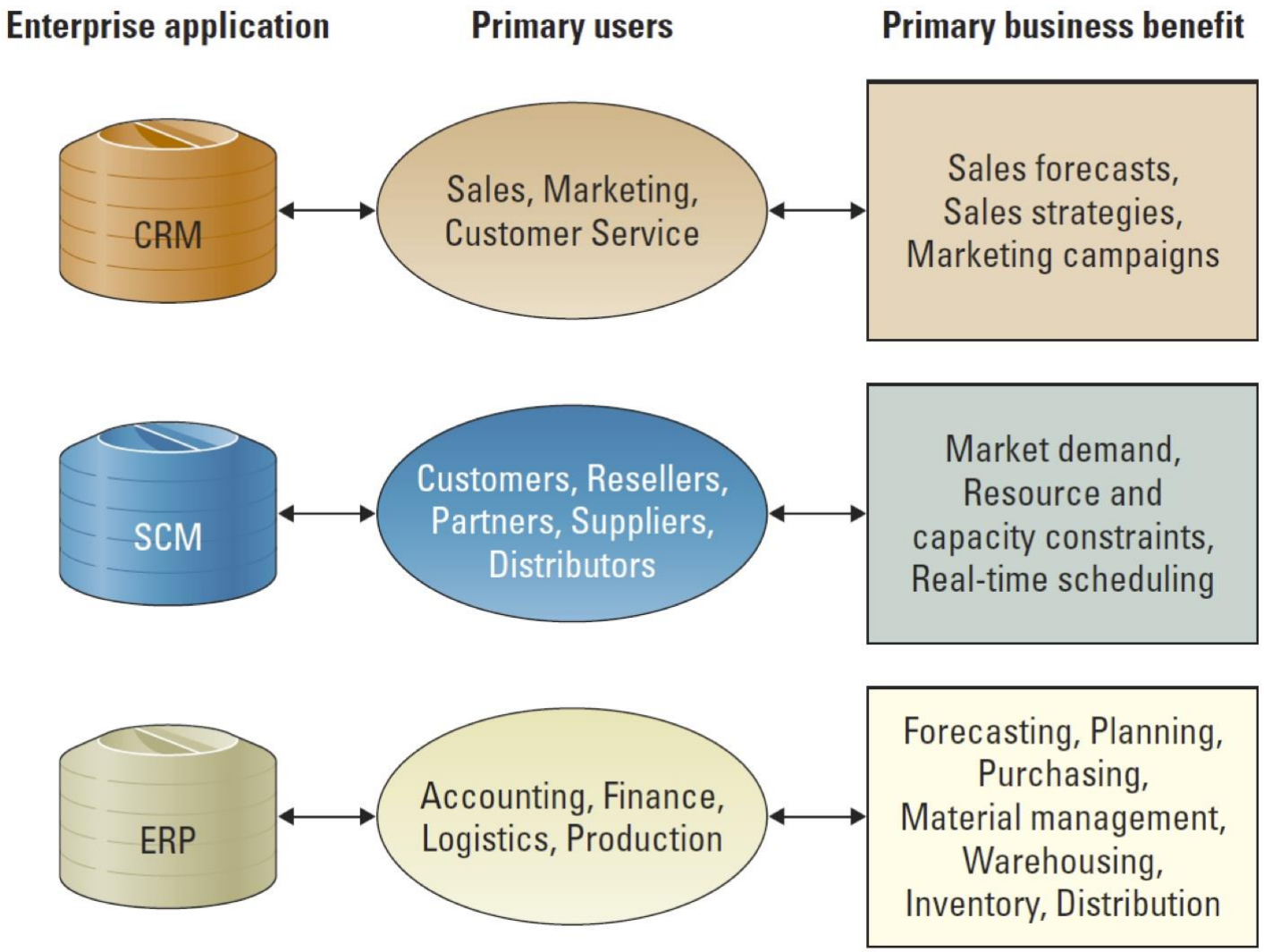
Outline: Data Pre-processing



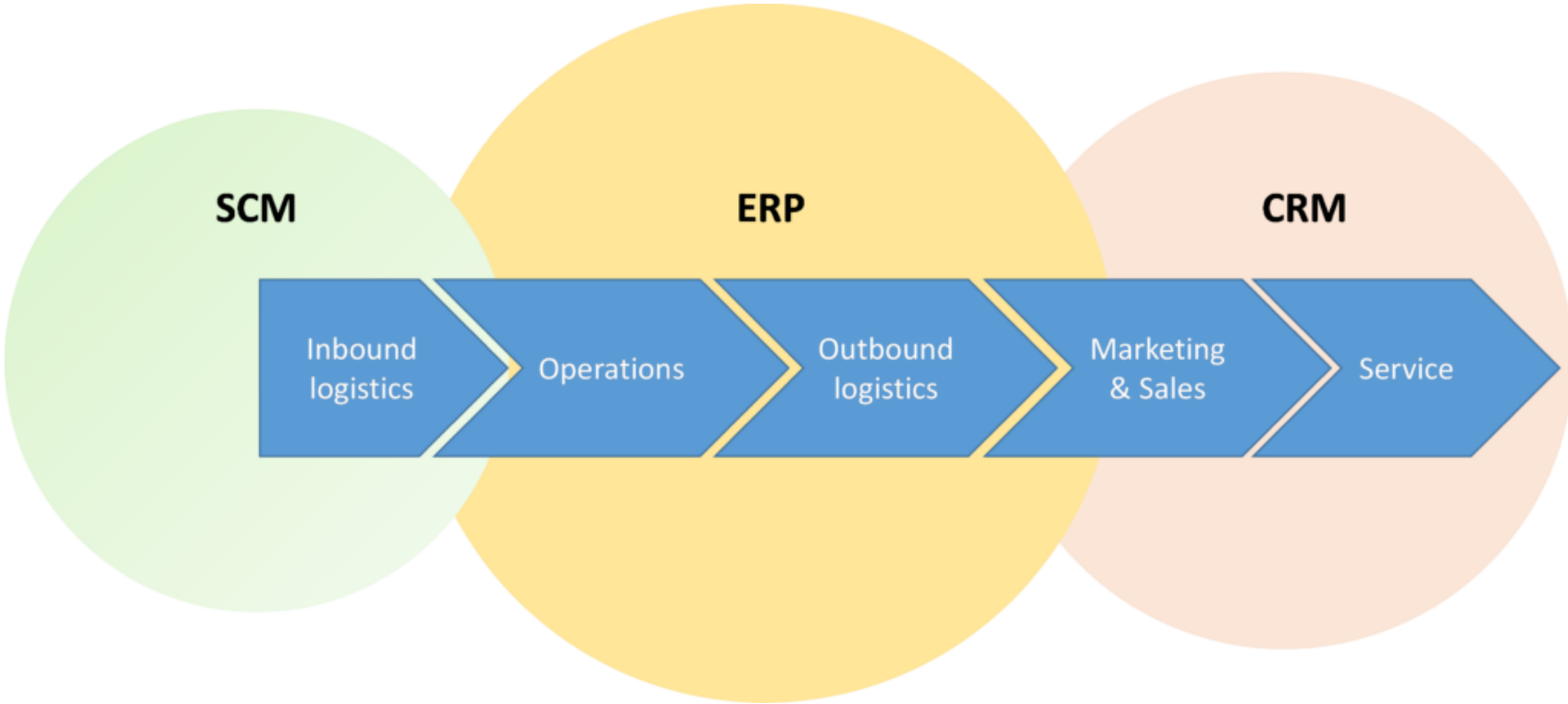
Data Integration



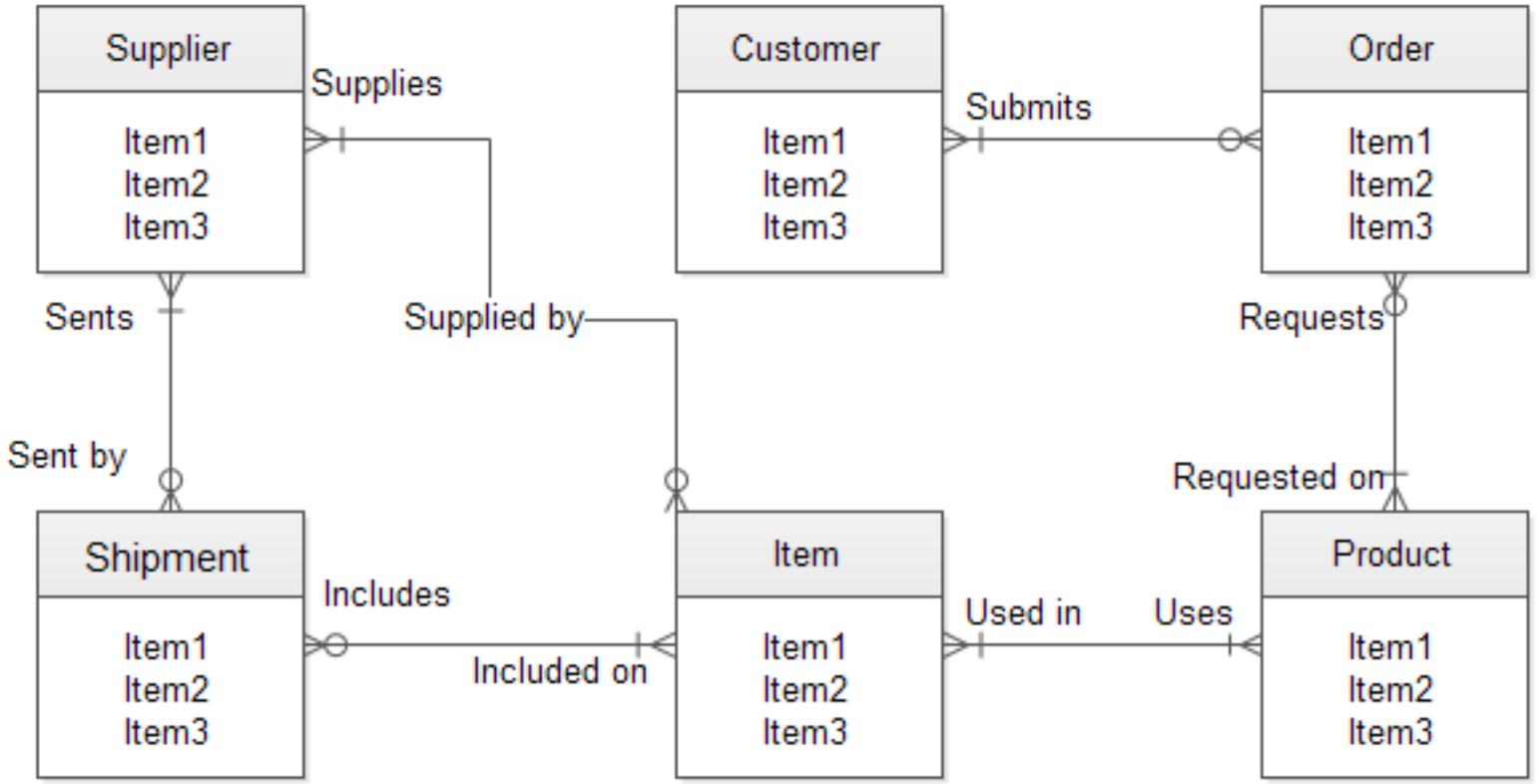
Data Integration



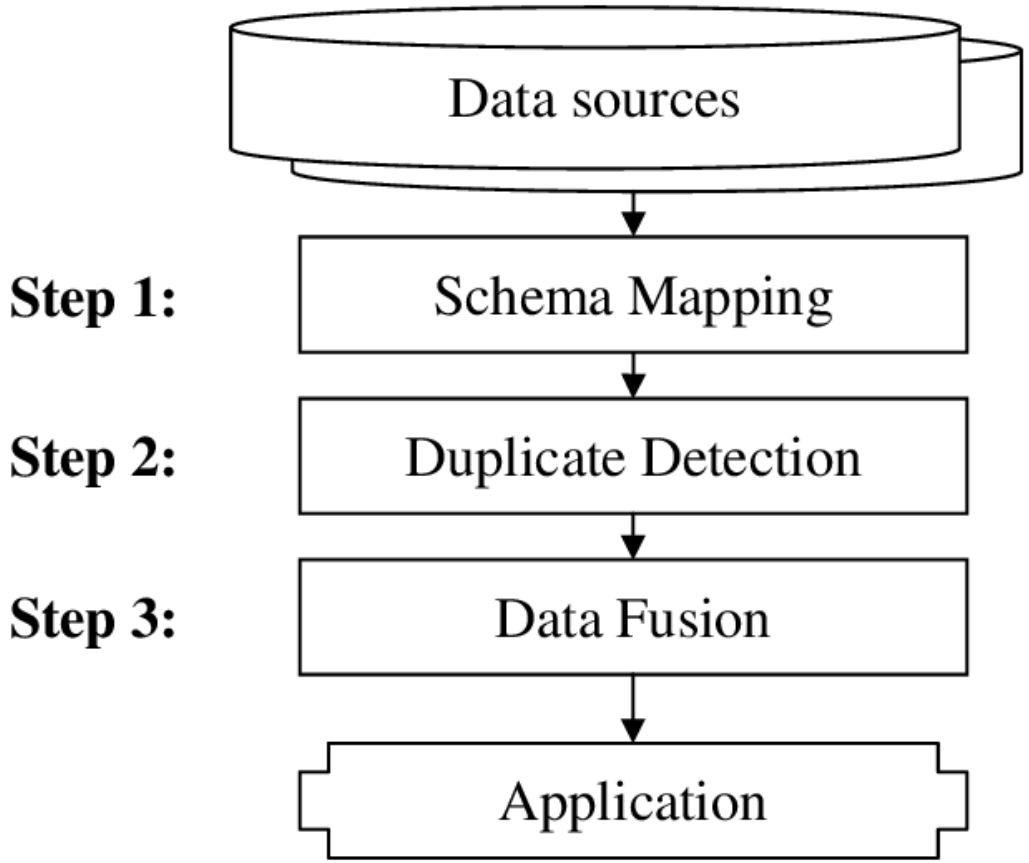
Data Integration



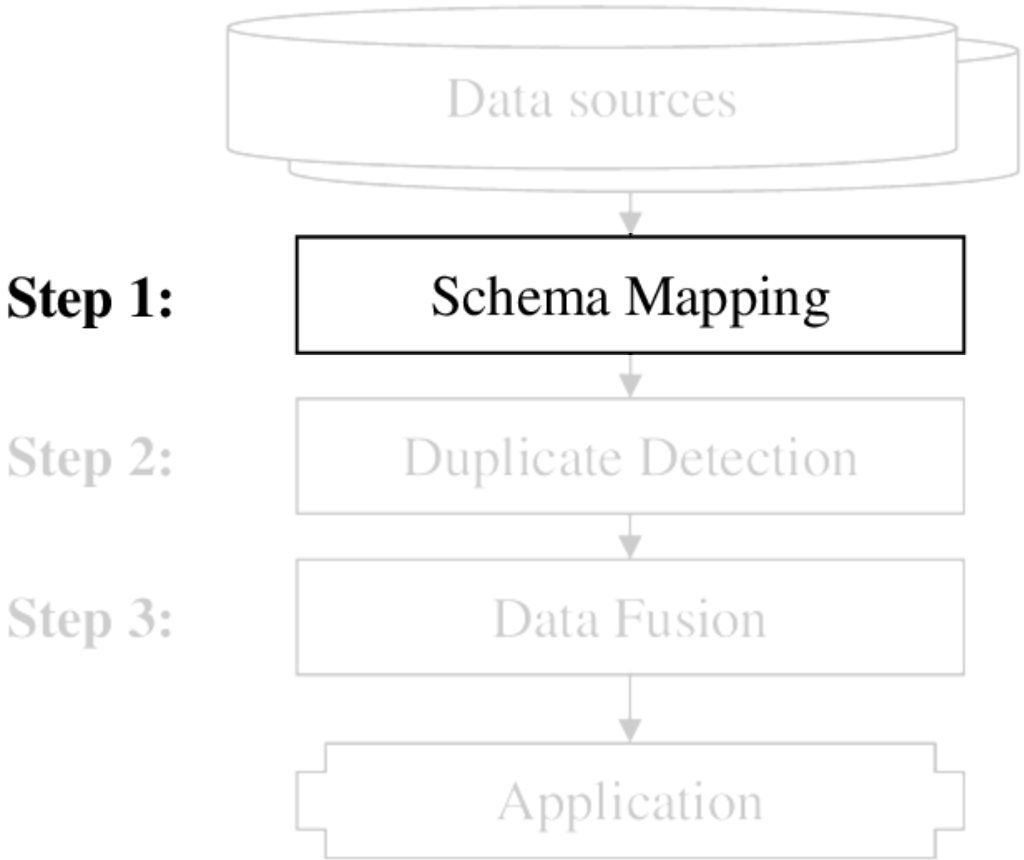
Data Integration



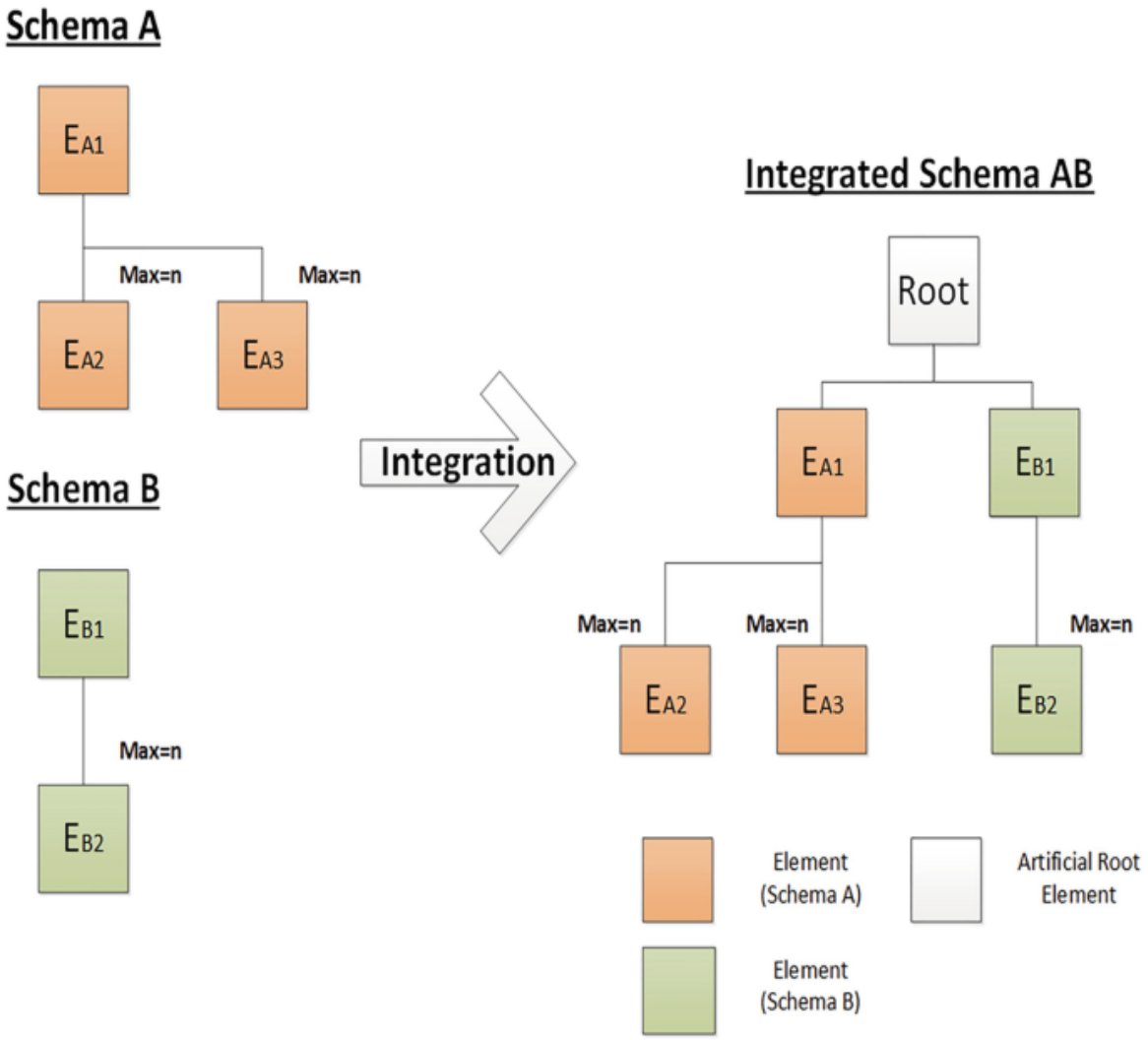
Data Integration



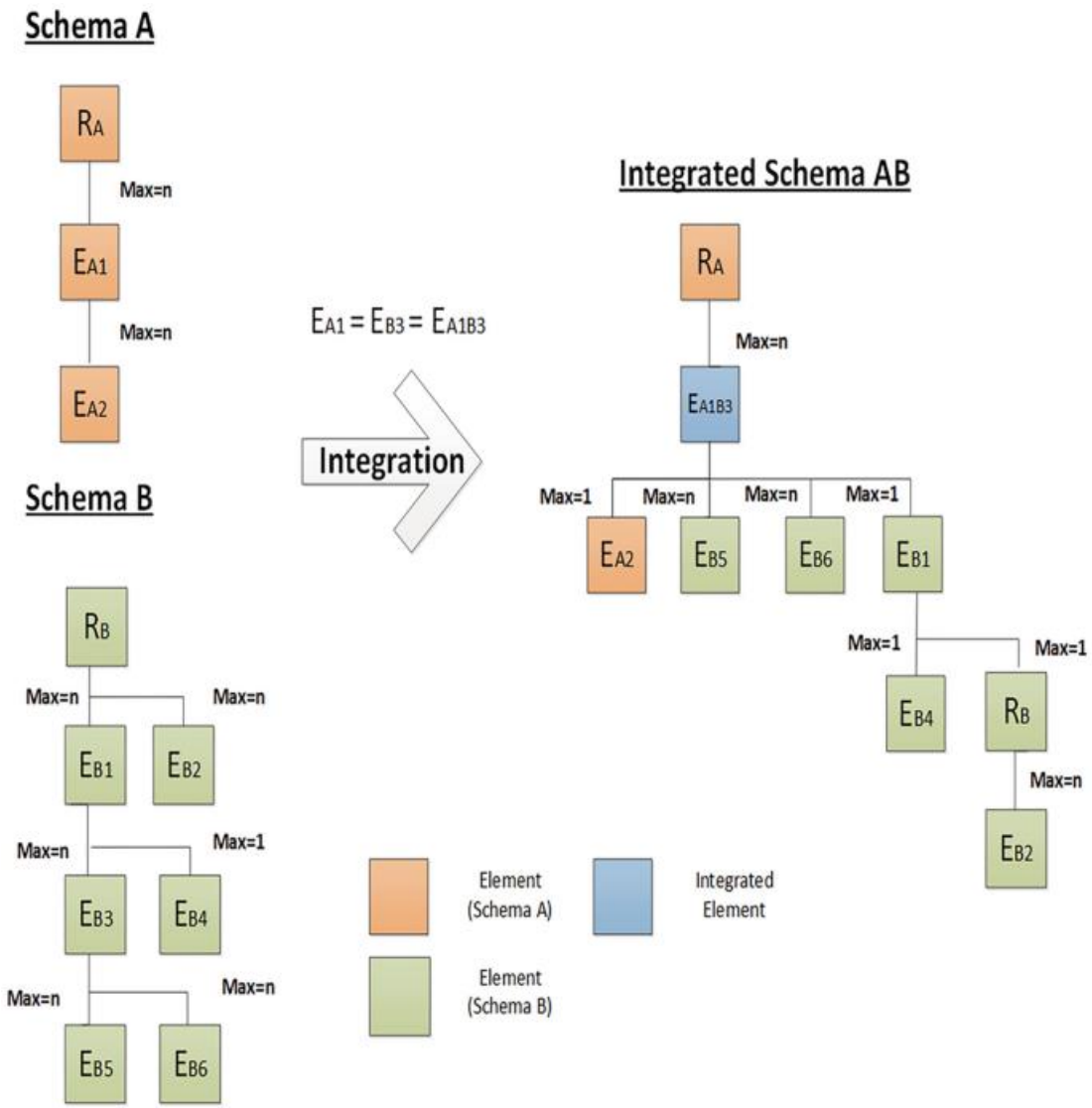
Schema Mapping



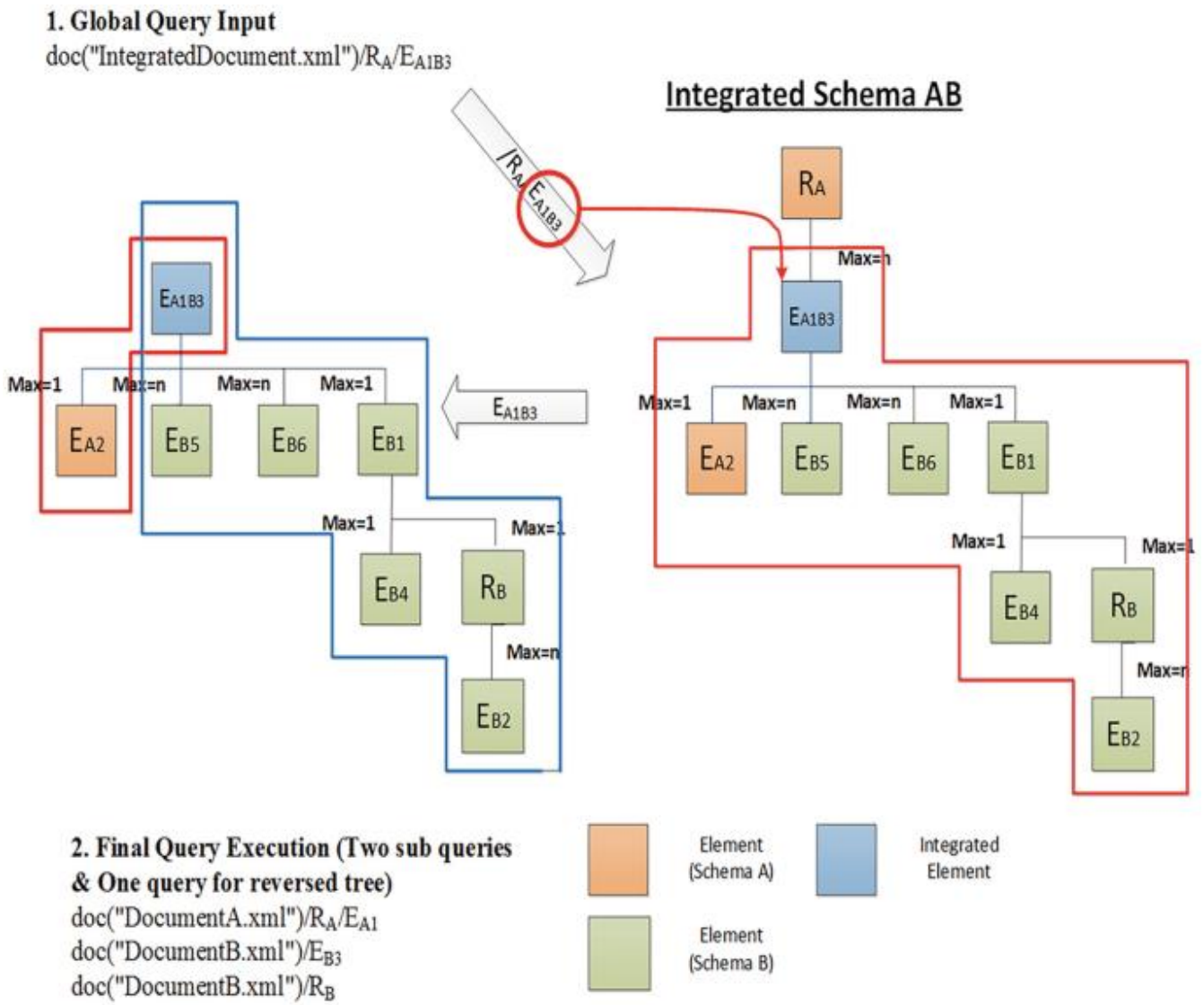
Schema Mapping



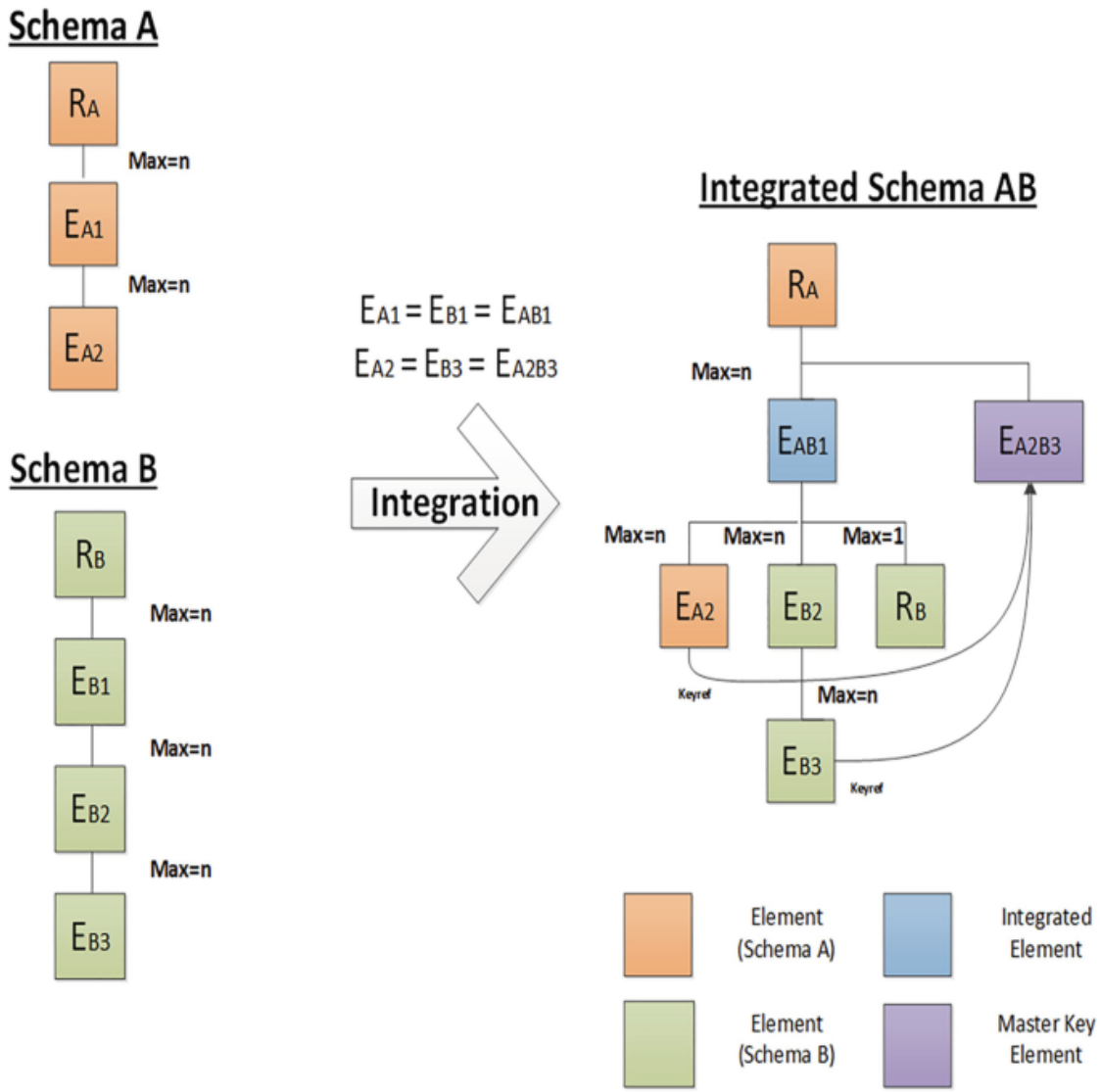
Schema Mapping



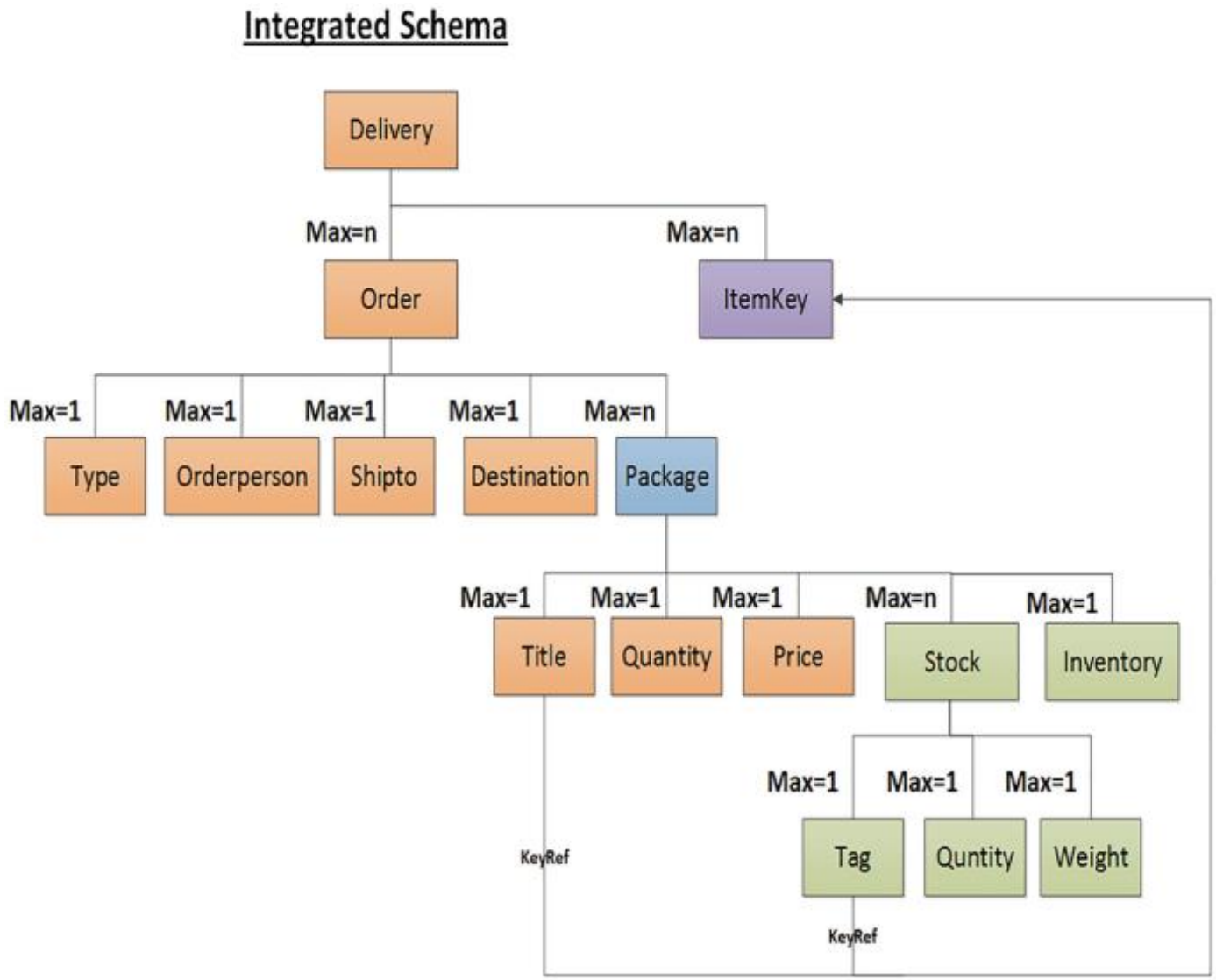
Schema Mapping



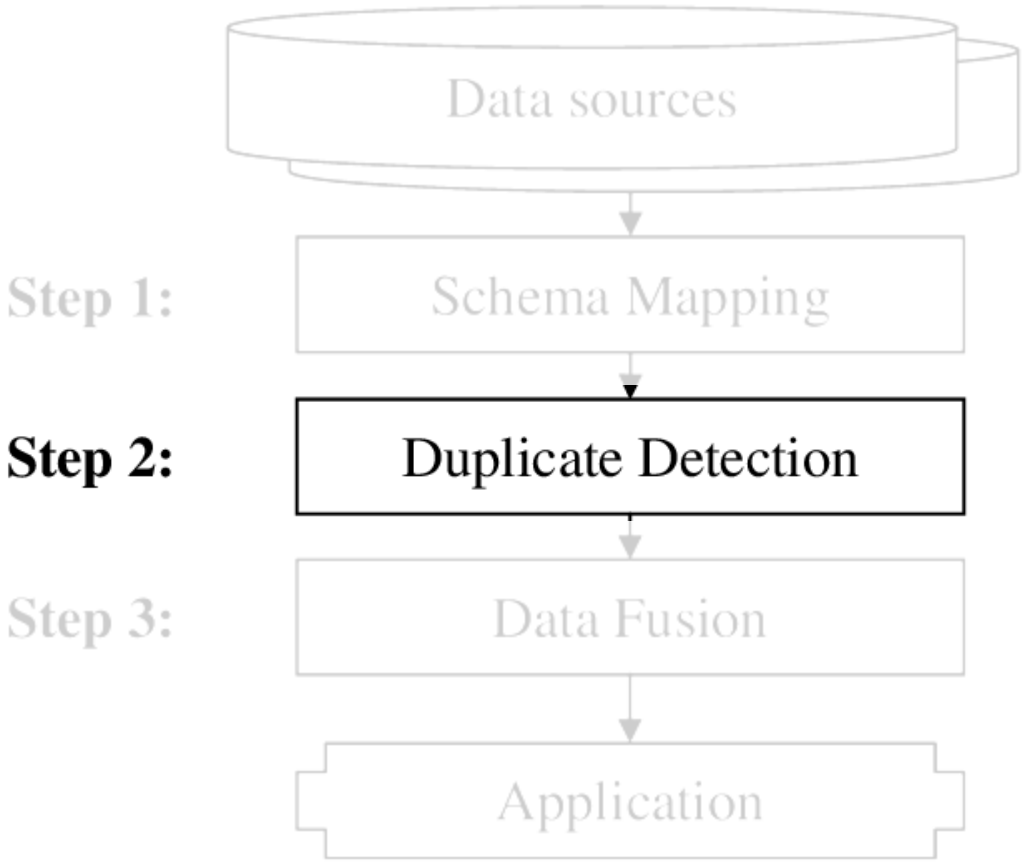
Schema Mapping



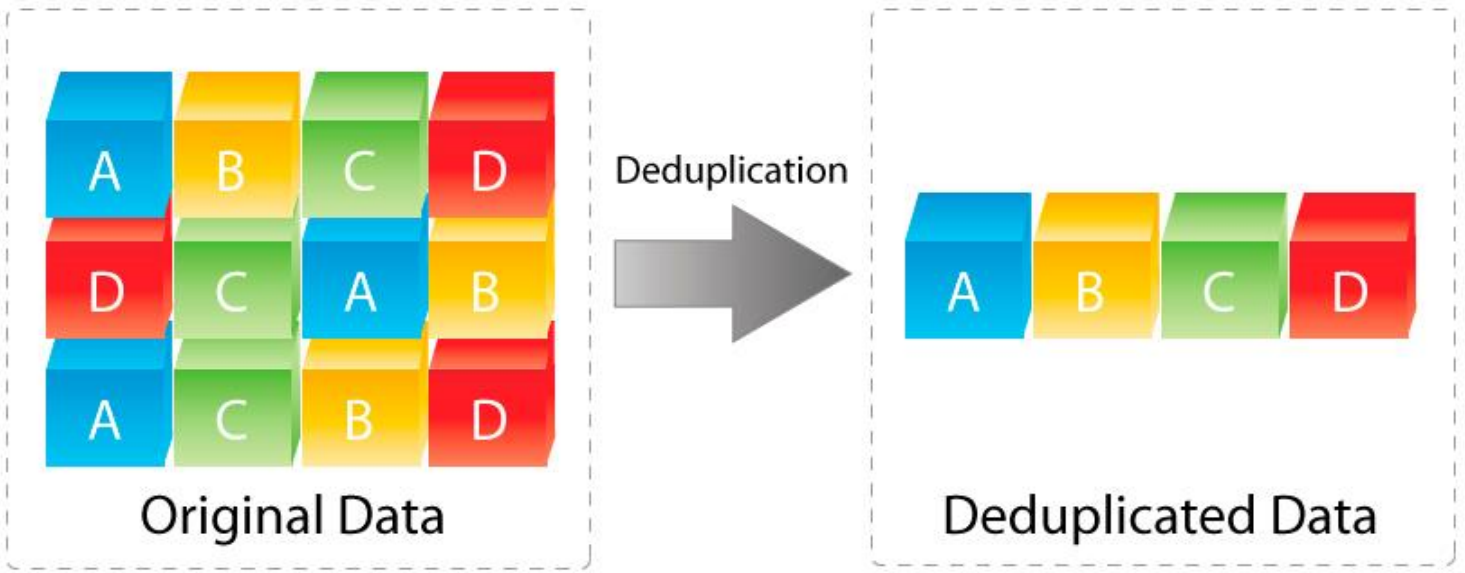
Schema Mapping



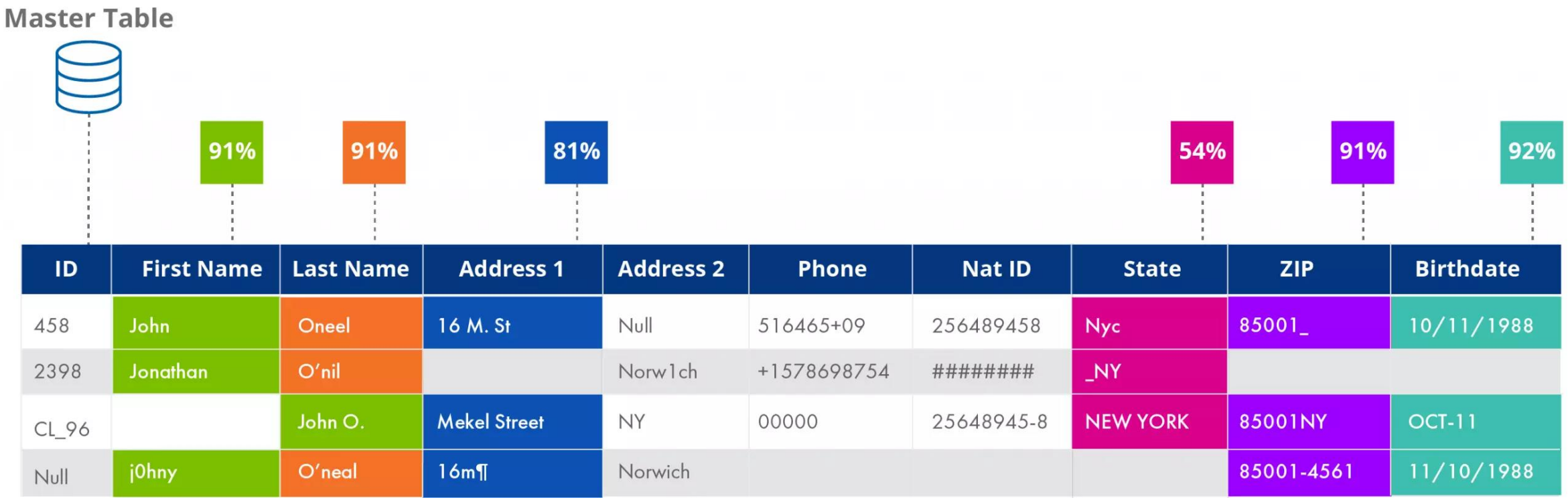
Duplicate Detection



Duplicate Detection

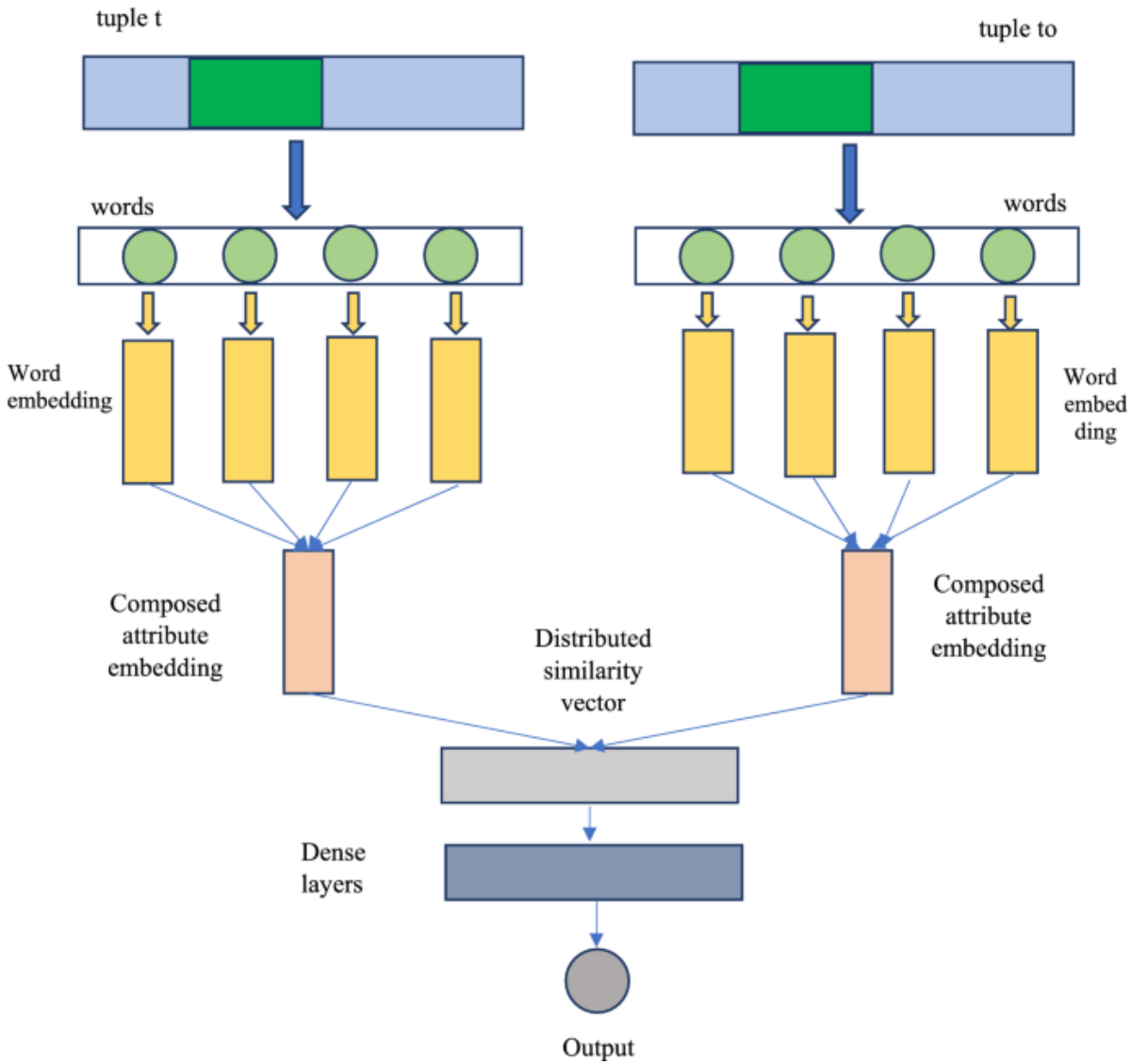


Duplicate Detection – Fuzzy Matching

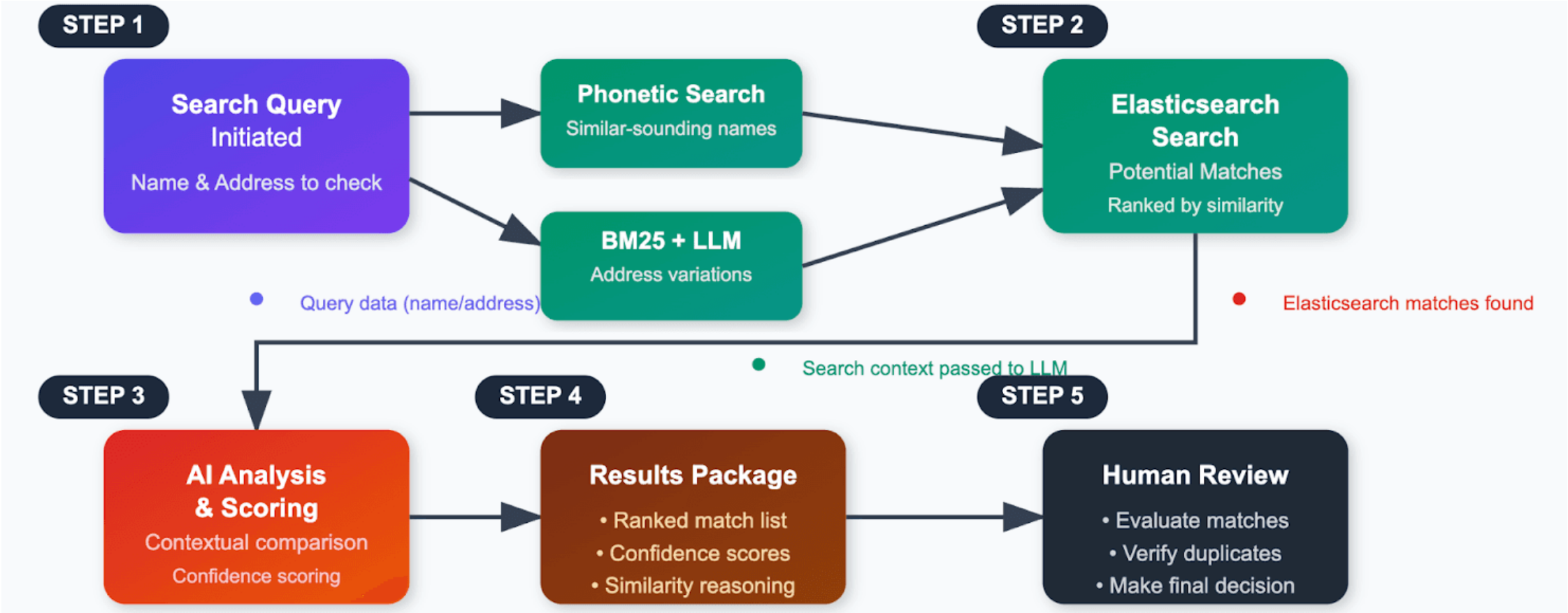


ID	First Name	Last Name	Address 1	Address 2	Phone	Nat ID	State	ZIP	Birthdate
458	John	Oneel	16 M. St	Null	516465+09	256489458	Nyc	85001_	10/11/1988
2398	Jonathan	O'nil		Norw1ch	+1578698754	########	_NY		
CL_96		John O.	Mekel Street	NY	00000	25648945-8	NEW YORK	85001NY	OCT-11
Null	j0hny	O'neal	16m¶	Norwich				85001-4561	11/10/1988

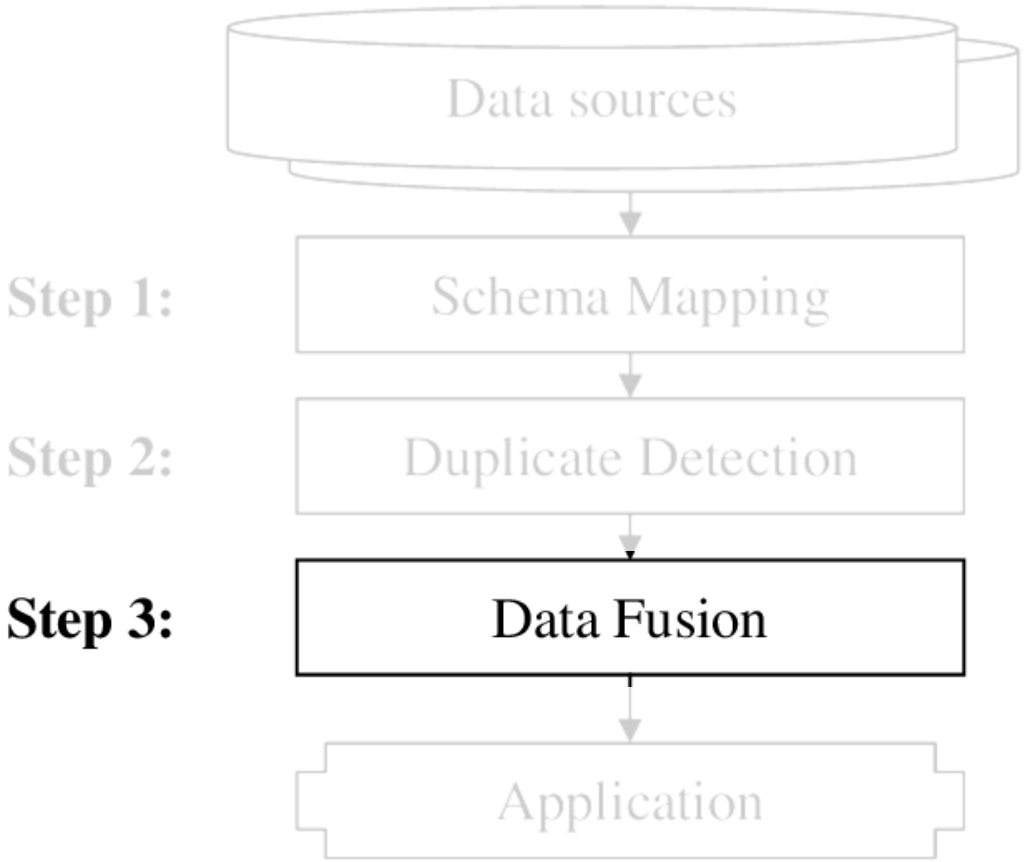
Duplicate Detection – Machine Learning



Duplicate Detection – LLM



Data Fusion


Data sources
Step 1:
Schema Mapping
Step 2:
Duplicate Detection
Step 3:
Data Fusion
Application

Data Fusion

ID	First Name	Last Name	Address 1	Address 2	Phone	Nat ID	State	ZIP	Birthdate
458	John	Oneel	16 M. St	Null	516465+09	256489458	Nyc	85001_	10/11/1988
2398	Jonathan	O'nil		Norw1ch	+1578698754	########	_NY		
CL_96		John O.	Mekel Street	NY	00000	25648945-8	NEW YORK	85001NY	OCT-11
Null	j0hny	O'neal	16m¶	Norwich				85001-4561	11/10/1988

Data Fusion

ID	First Name	Last Name	Address 1	Address 2	Phone	Nat ID	State	ZIP	Birthdate
458	John	Oneel	16 M. St	Null	516465+09	256489458	Nyc	85001_	10/11/1988
2398	Jonathan	O'nil		Norw1ch	+1578698754	########	_NY		
CL_96		John O.	Mekel Street	NY	00000	25648945-8	NEW YORK	85001NY	OCT-11
Null	j0hny	O'neal	16m¶	Norwich				85001-4561	11/10/1988

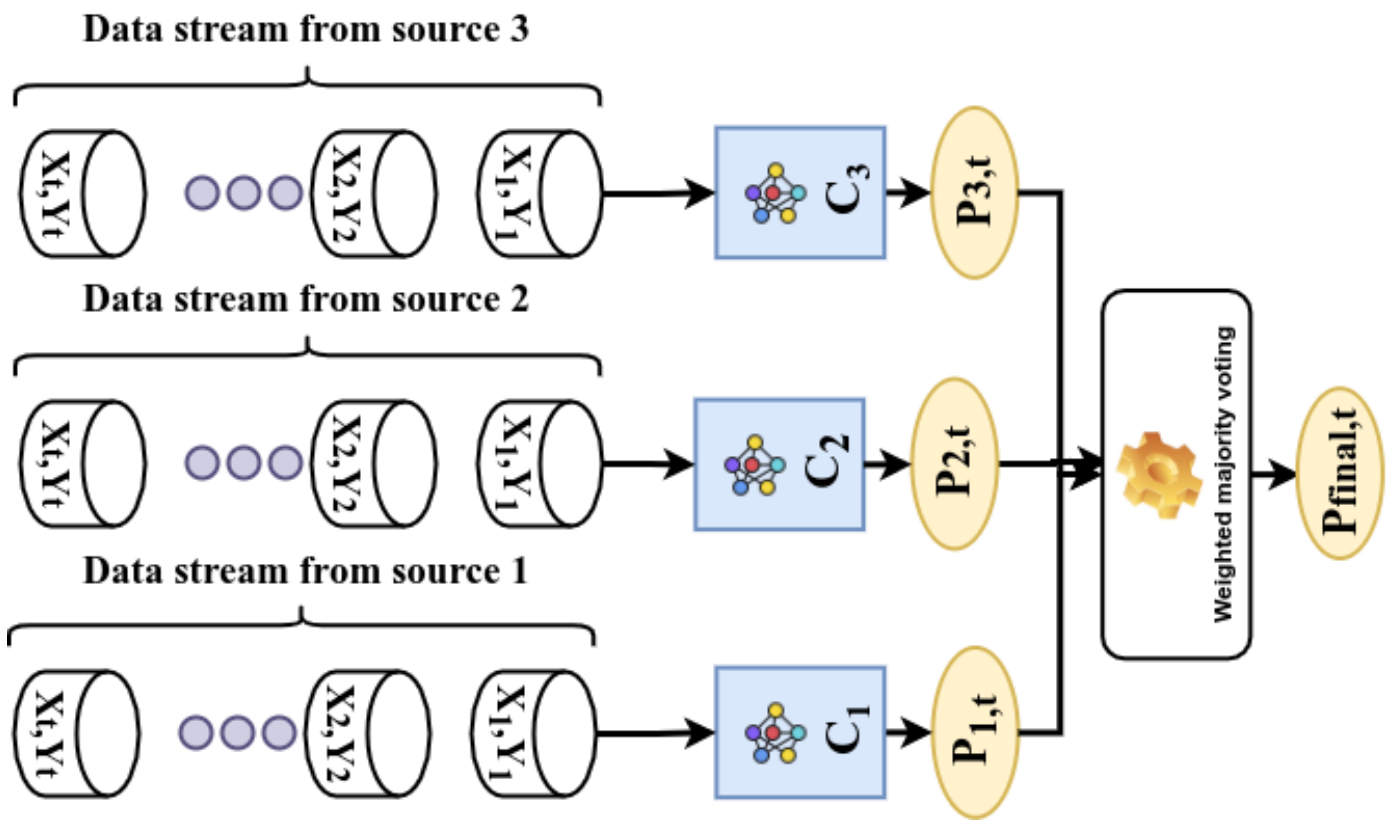


Which value to use?

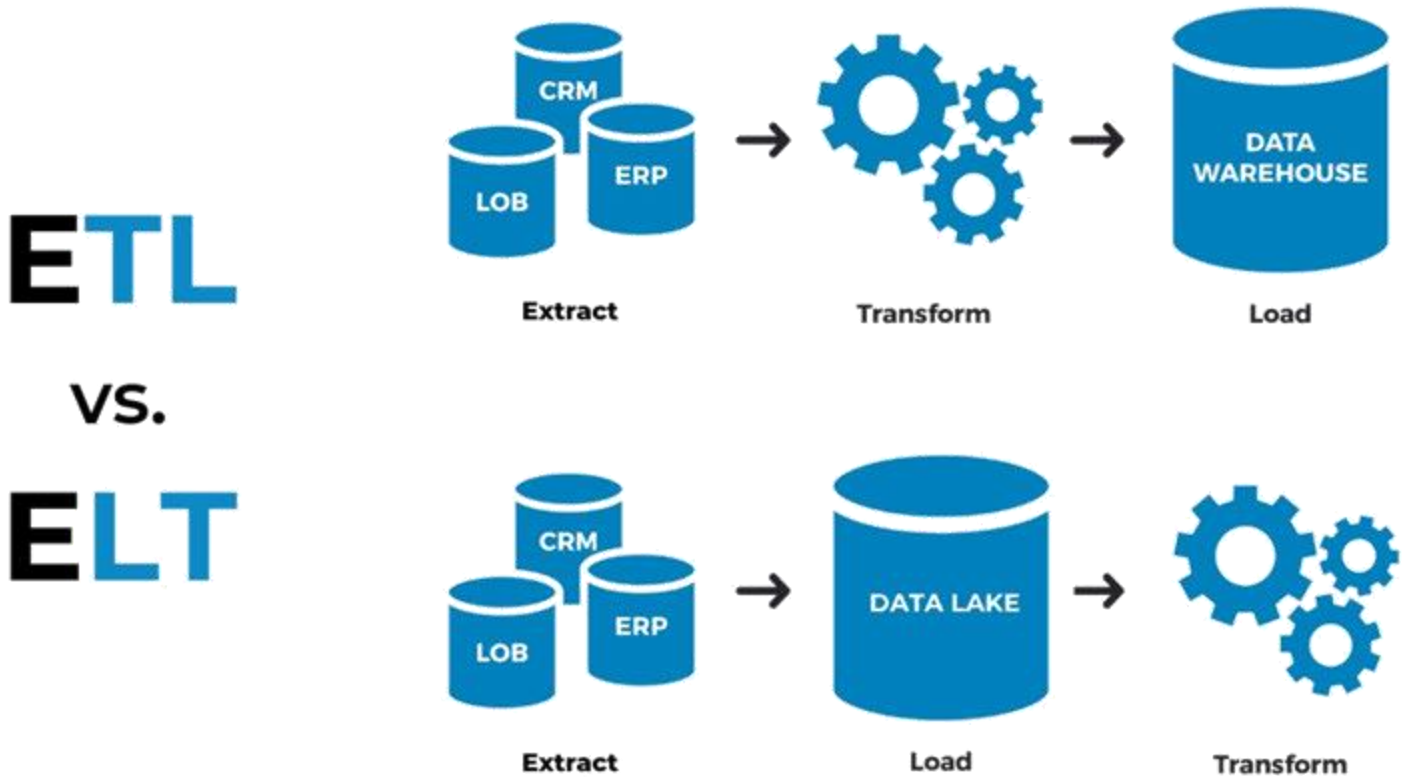
Data Fusion

1. Source trustworthy
2. Recency
3. Majority voting
4. Machine learning

Data Fusion – Majority Voting



Q. When to Integrate?



Q. When to Integrate?

Aspect	ETL (Extract → Transform → Load)	ELT (Extract → Load → Transform)
Transform timing	Transform **before** loading	Transform **after** loading
Best suited for	Traditional on-premise Data Warehouses	Cloud Data Warehouses / Big Data platforms
Advantages	- Only clean data stored in DW - Easier governance & compliance	- Scales well for large data - Raw data retained for reuse - Faster initial loading
Disadvantages	- Transformation server can be a bottleneck - Limited scalability	- Raw data in DW poses security/governance risks - Transformation can increase query/storage cost - Risk of “data swamp” if unmanaged