Data Science

Collection, Cleaning, and Analysis

Example of Ideal Data (for analysis)

Manufacturer	Name	CPU Frequency	L1 Cache Size	Volatile Memory Size
HP	Pavillion	3.2 GHz	512 KB	4 GB
Acer	Aspire	3.5 GHz	512 KB	4 GB
Dell	Inspiron	4.0 GHz	1024 KB	8 GB
IBM	Watson	5.0 GHz	2048 KB	32 GB
HP	Spectre	2.8 GHz	256 KB	6 GB
Toshiba	Tecra	3.0 GHz	256 KB	4 GB

Example of Ideal Data (for analysis)

Manufacturer	Name	CPU Frequency	L1 Cache Size	Volatile Memory Size
HP	Pavillion	3.2 GHz	512 KB	4 GB
Acer	Aspire	3.5 GHz	512 KB	4 GB
Dell	Inspiron	4.0 GHz	1024 KB	8 GB
IBM	Watson	5.0 GHz	2048 KB	32 GB
HP	Spectre	2.8 GHz	256 KB	6 GB
Toshiba	Tecra	3.0 GHz	256 KB	4 GB

- No missing values

Example of Ideal Data (for analysis)

Manufacturer	Name	CPU Frequency	L1 Cache Size	Volatile Memory Size
HP	Pavillion	3.2 GHz	512 KB	4 GB
Acer	Aspire	3.5 GHz	512 KB	4 GB
Dell	Inspiron	4.0 GHz	1024 KB	8 GB
IBM	Watson	5.0 GHz	2048 KB	32 GB
HP	Spectre	2.8 GHz	256 KB	6 GB
Toshiba	Tecra	3.0 GHz	256 KB	4 GB

- No missing values
- Uniformity (categories are all unique, consistent units, etc.)

Example of Ideal Data (for analysis)

Manufacturer	Name	CPU Frequency	L1 Cache Size	Volatile Memory Size
HP	Pavillion	3.2 GHz	512 KB	4 GB
Acer	Aspire	3.5 GHz	512 KB	4 GB
Dell	Inspiron	4.0 GHz	1024 KB	8 GB
IBM	Watson	5.0 GHz	2048 KB	32 GB
HP	Spectre	2.8 GHz	256 KB	6 GB
Toshiba	Tecra	3.0 GHz	256 KB	4 GB

- No missing values
- Uniformity (categories are all unique, consistent units, etc.)
- Lots (at least 100’s) of examples (rows) and lots of information (cols)

Standard Data Pipeline

Collection

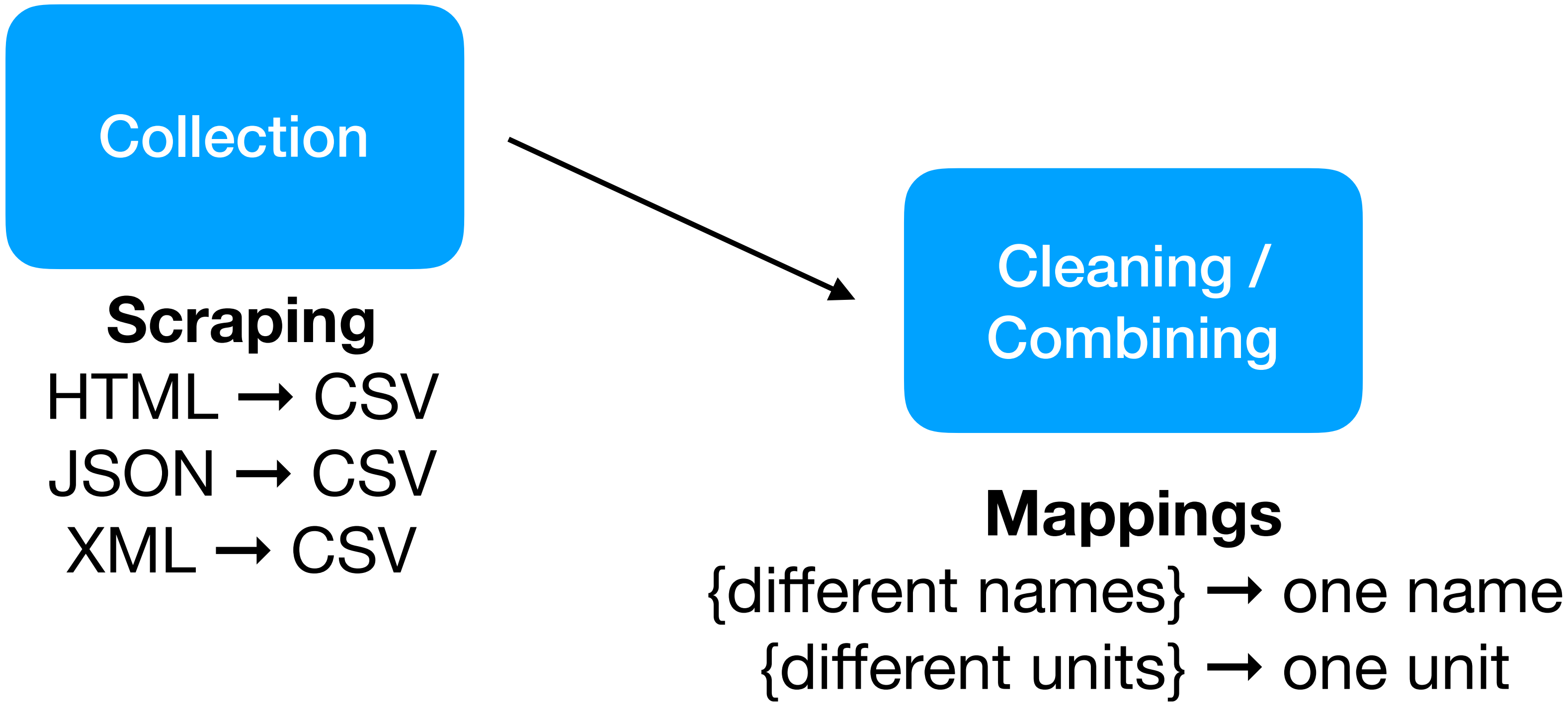
Scraping

HTML → CSV

JSON → CSV

XML → CSV

Standard Data Pipeline



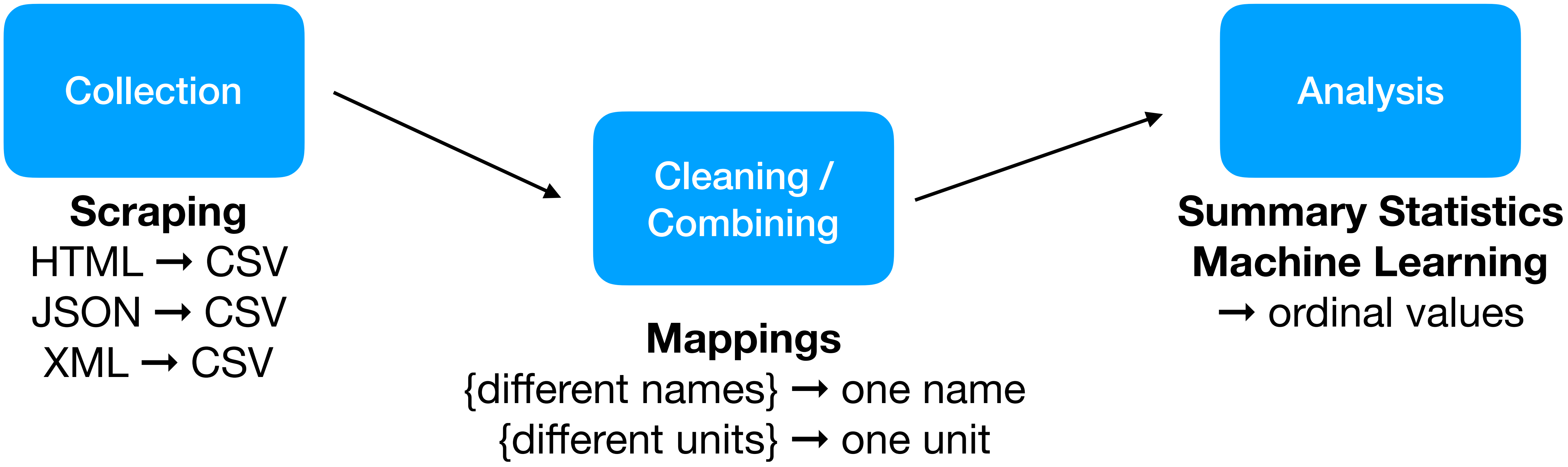
Collection

Scraping
HTML ➞ CSV
JSON ➞ CSV
XML ➞ CSV

Cleaning / Combining

Mappings
{different names} ➞ one name
{different units} ➞ one unit

Standard Data Pipeline



Collection

Scraping
HTML → CSV
JSON → CSV
XML → CSV

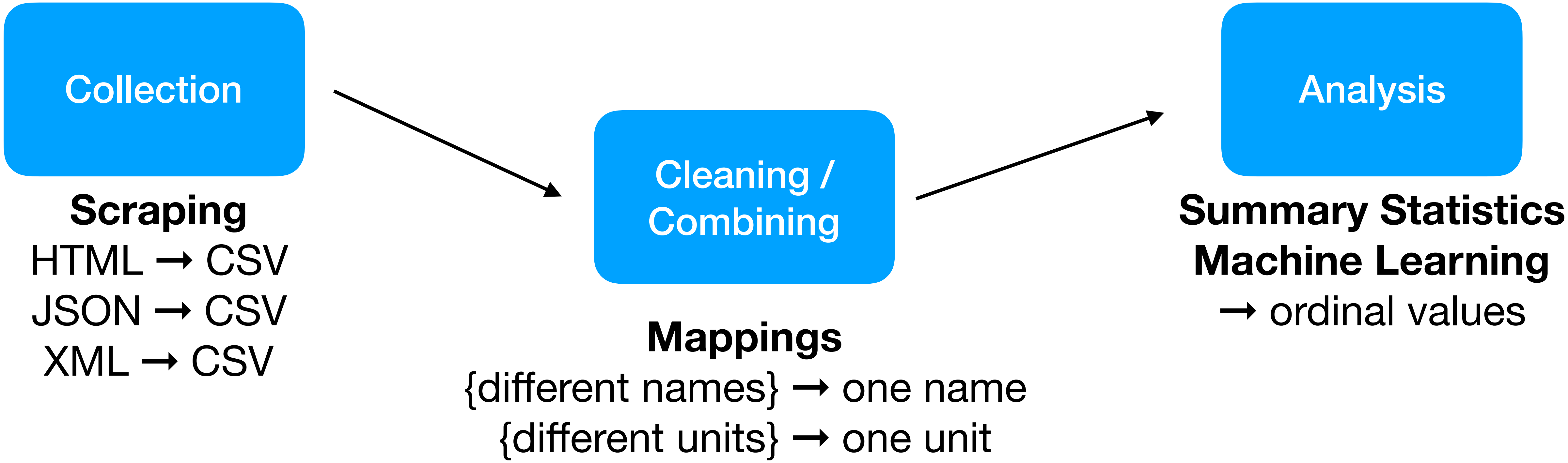
Cleaning / Combining

Mappings
{different names} → one name
{different units} → one unit

Analysis

Summary Statistics
Machine Learning
→ ordinal values

Standard Data Pipeline



Collection

Scraping
HTML → CSV
JSON → CSV
XML → CSV

Cleaning / Combining

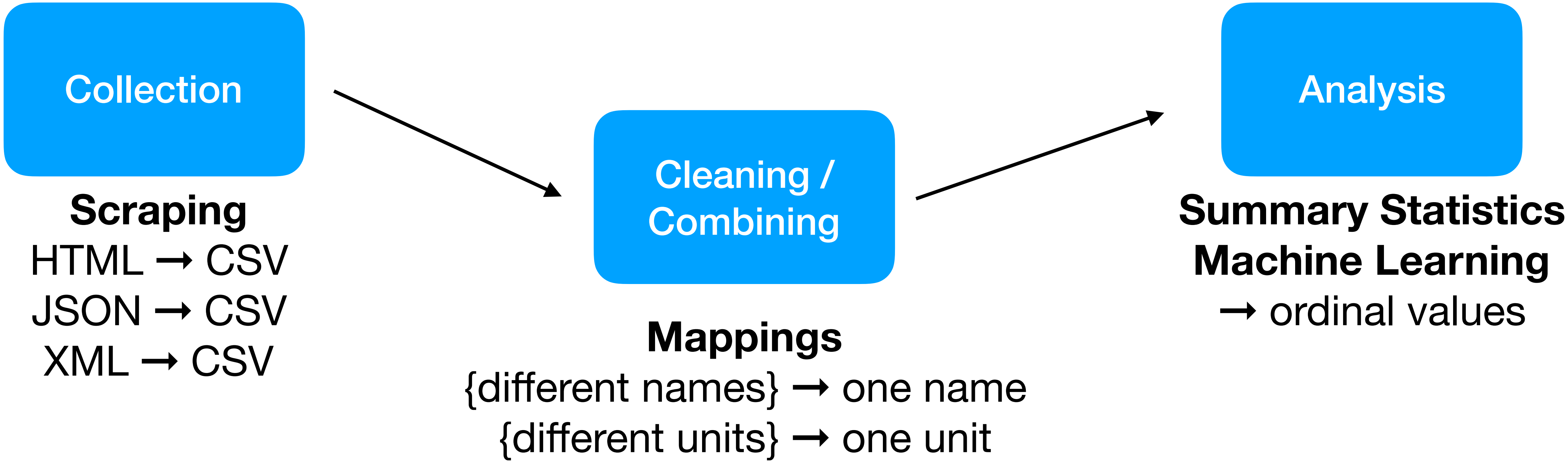
Mappings
{different names} → one name
{different units} → one unit

Analysis

Summary Statistics
Machine Learning
→ ordinal values

Three important data collection and cleaning questions:

Standard Data Pipeline



Collection

Scraping
HTML → CSV
JSON → CSV
XML → CSV

Cleaning / Combining

Mappings
{different names} → one name
{different units} → one unit

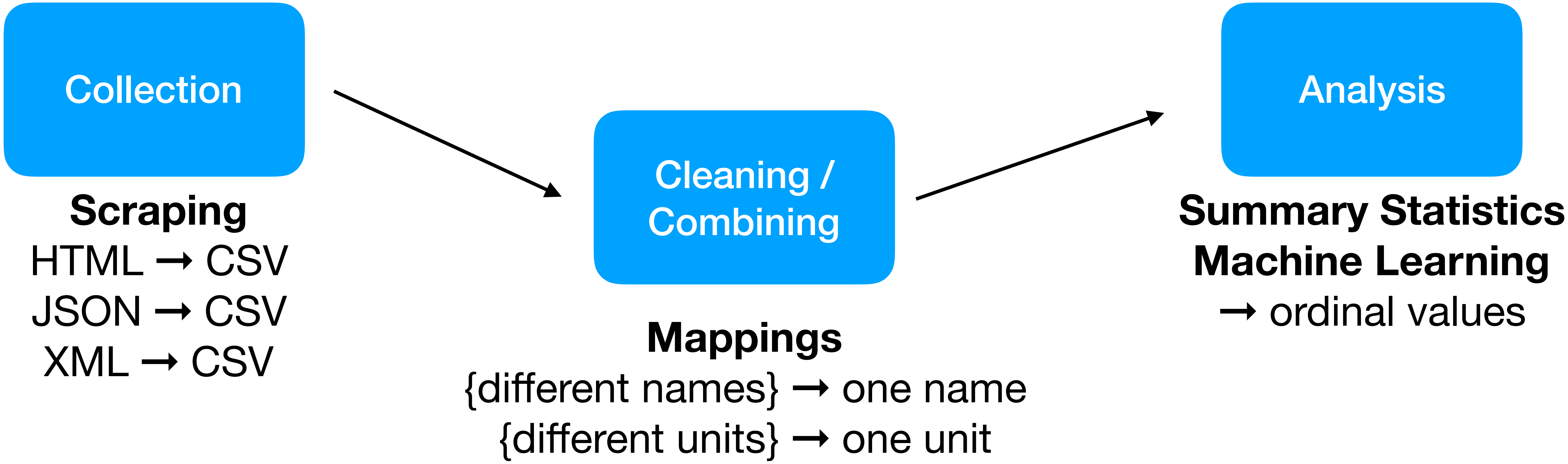
Analysis

Summary Statistics
Machine Learning
→ ordinal values

Three important data collection and cleaning questions:

– What are the sources that you got data from?

Standard Data Pipeline



Collection

Scraping
HTML → CSV
JSON → CSV
XML → CSV

Cleaning / Combining

Mappings
{different names} → one name
{different units} → one unit

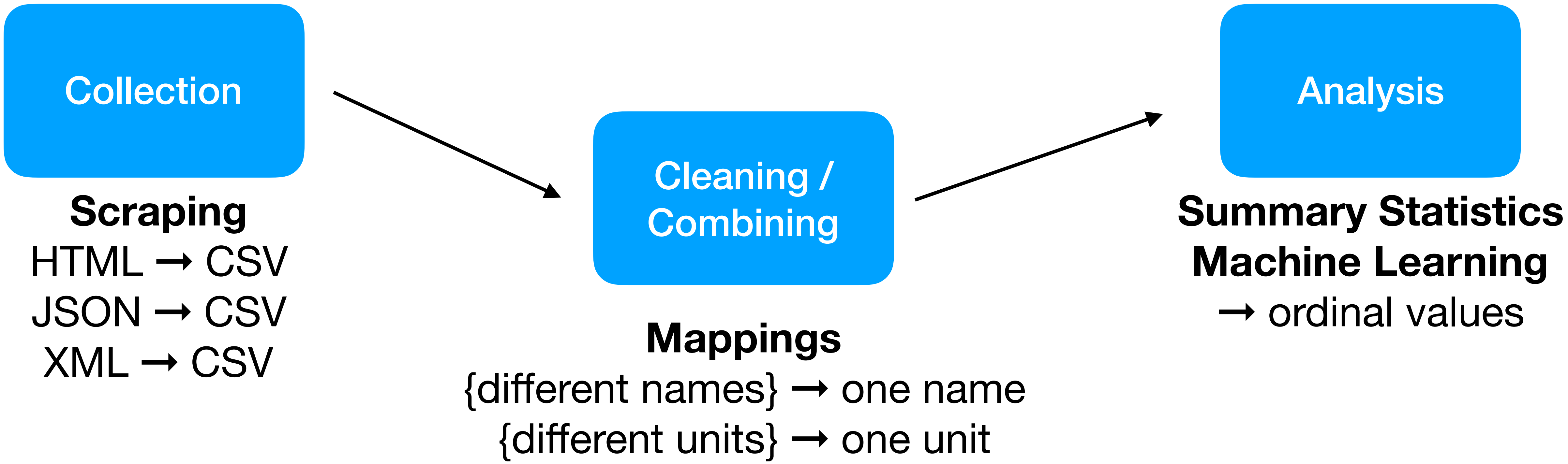
Analysis

Summary Statistics
Machine Learning
→ ordinal values

Three important data collection and cleaning questions:

- What are the sources that you got data from?
- What data did you get from each of the sources?

Standard Data Pipeline



Collection

Scraping
HTML → CSV
JSON → CSV
XML → CSV

Cleaning / Combining

Mappings
{different names} → one name
{different units} → one unit

Analysis

Summary Statistics
Machine Learning
→ ordinal values

Three important data collection and cleaning questions:

- What are the sources that you got data from?
- What data did you get from each of the sources?
- How did you merge that data (name cleaning, etc.)

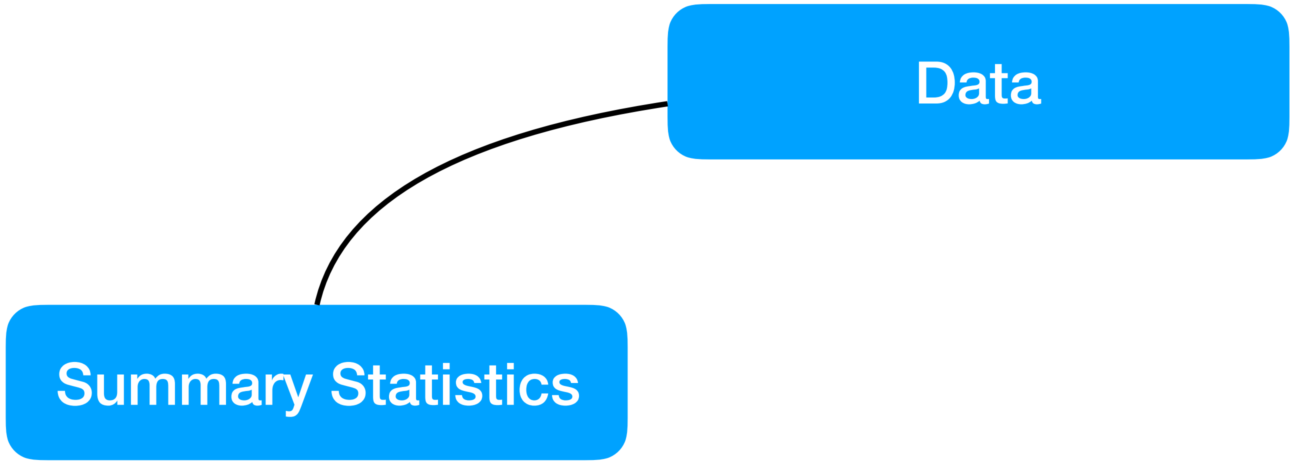
Structure of Analysis Concepts

Why clean?

Data

Structure of Analysis Concepts

Why clean?

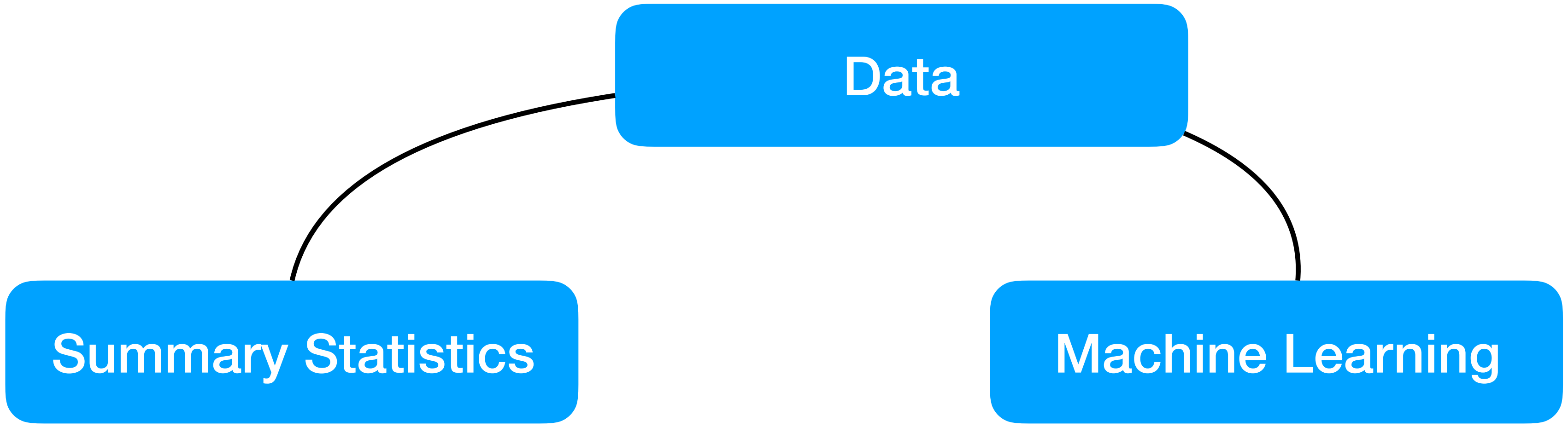


Data

Summary Statistics

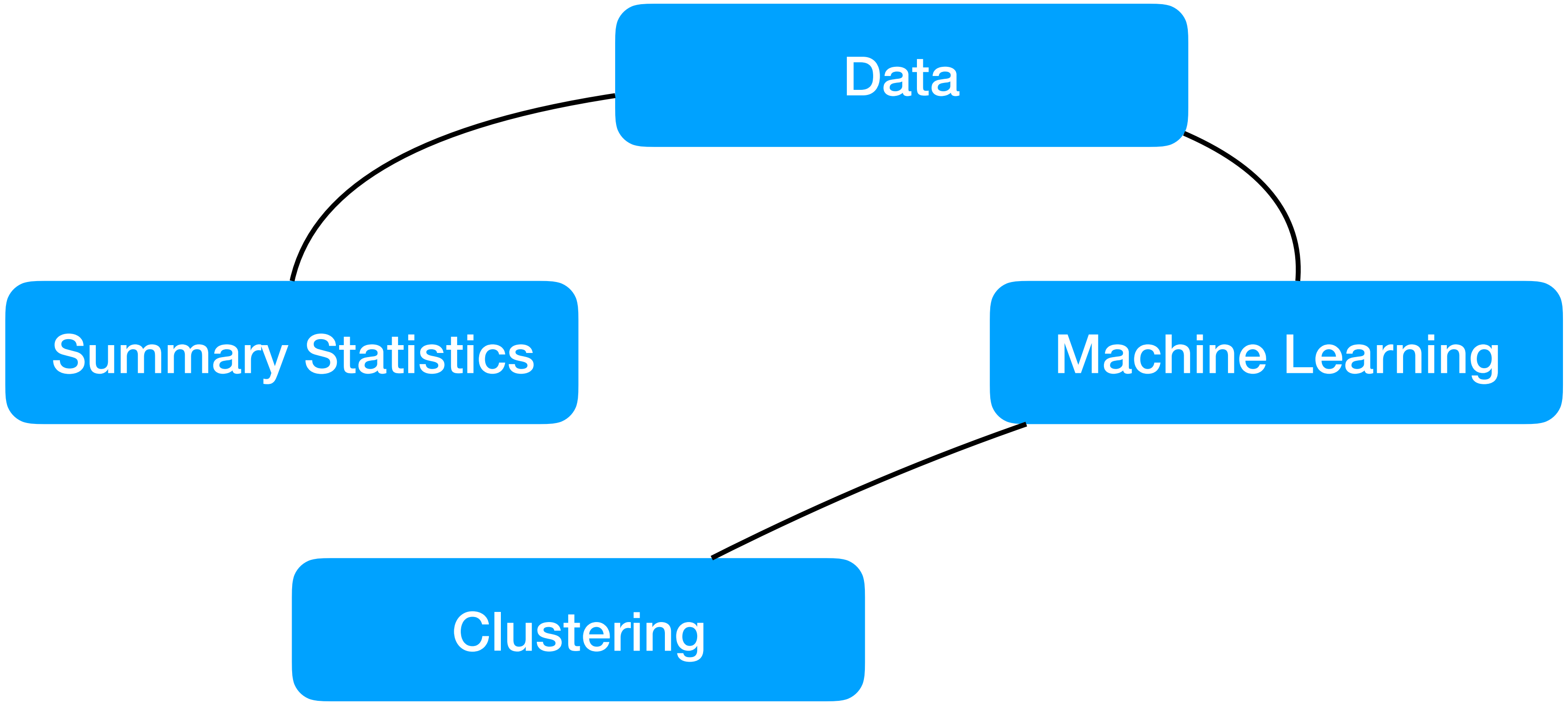
Structure of Analysis Concepts

Why clean?



Structure of Analysis Concepts

Why clean?



Data

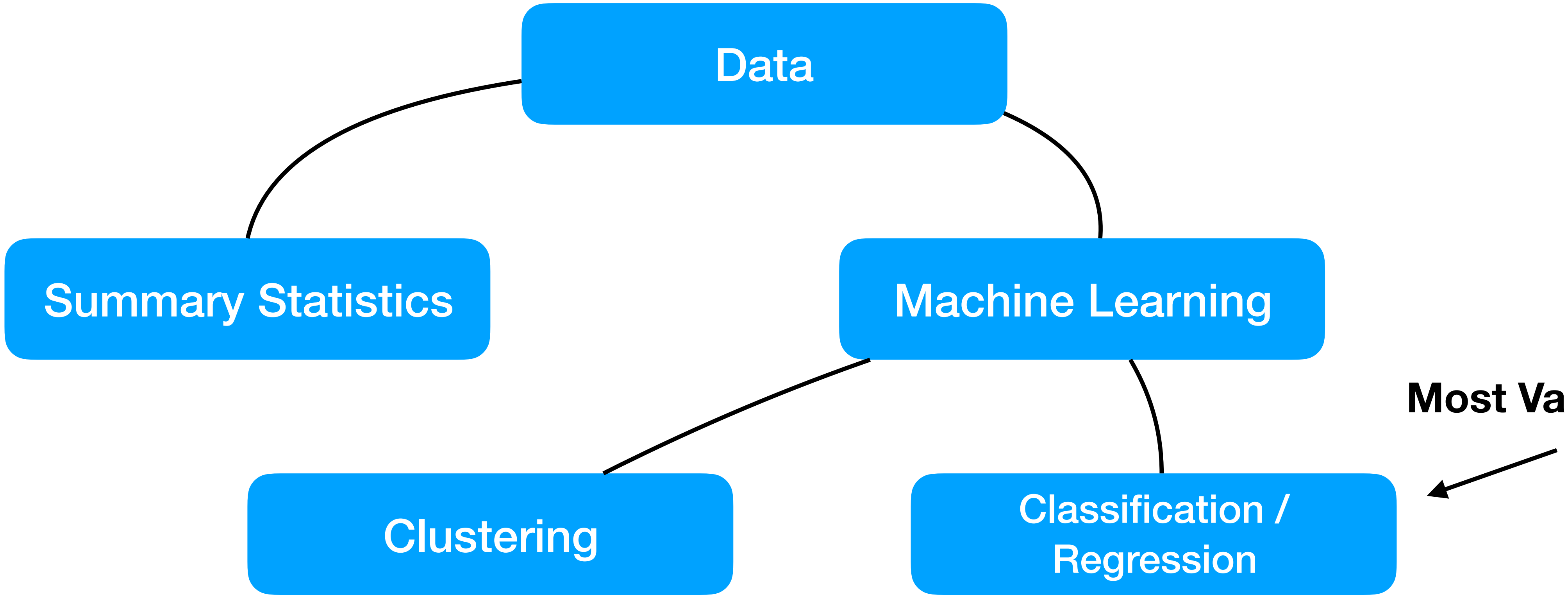
Summary Statistics

Machine Learning

Clustering

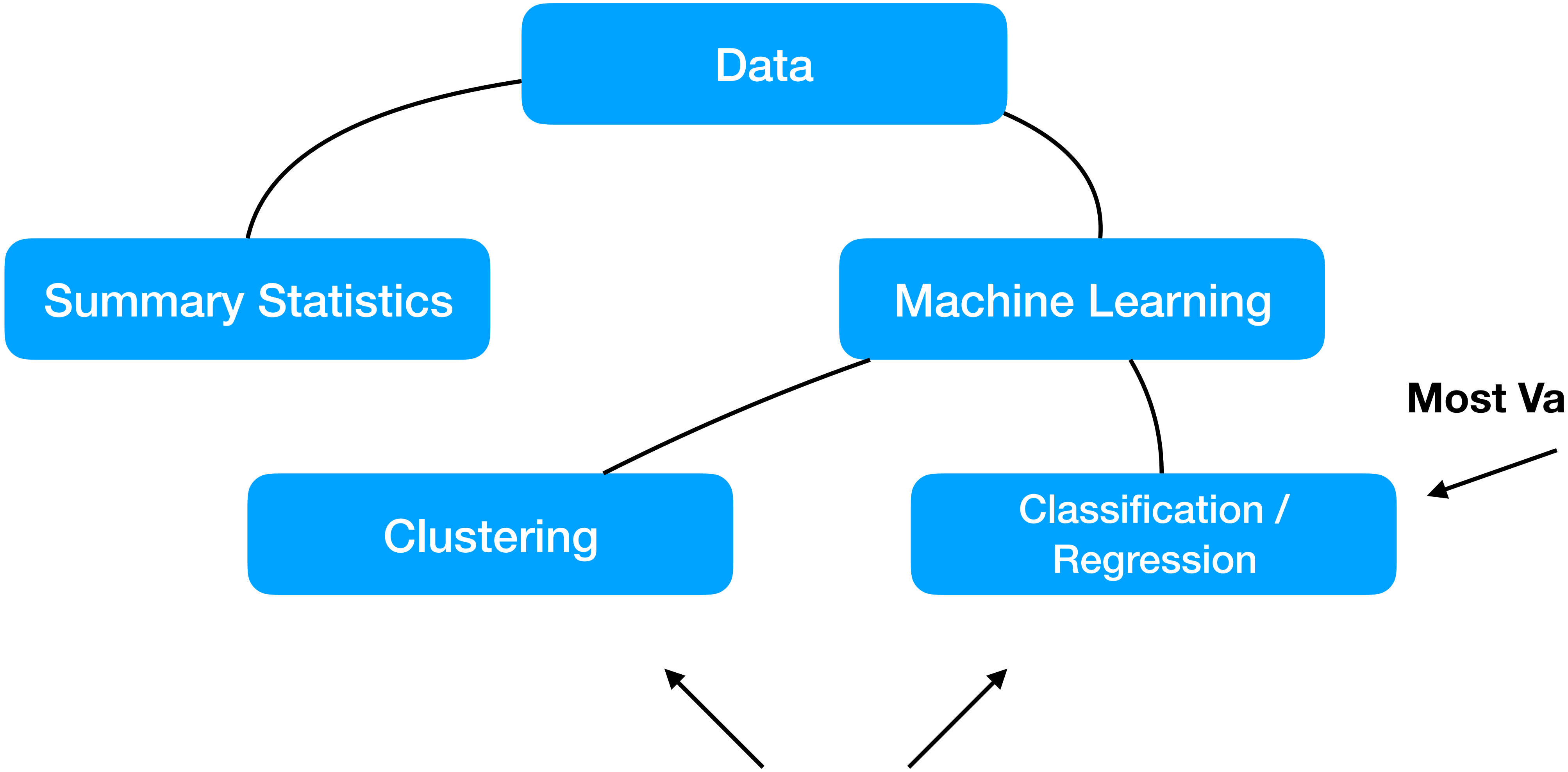
Structure of Analysis Concepts

Why clean?



Structure of Analysis Concepts

Why clean?



Data

Summary Statistics

Machine Learning

Clustering

Classification / Regression

Most VarSys work

(Usually) require real valued vectors!

VarSys Goals

VarSys Goals

– Quantify *variability* across the system stack. (memory, processing, etc.)

VarSys Goals

- Quantify *variability* across the system stack. (memory, processing, etc.)
- Model and understand the key influencers of *variability*.

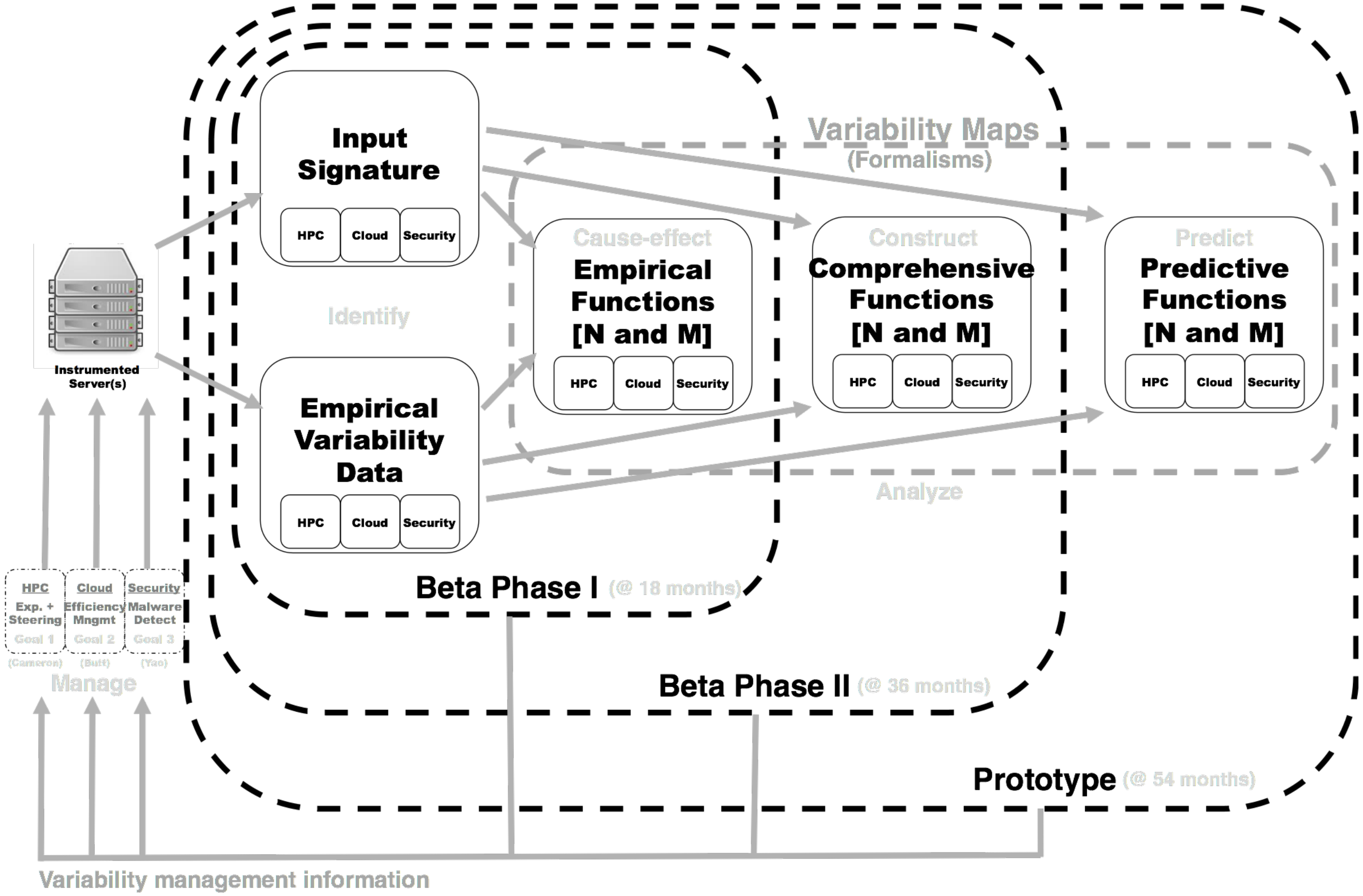
VarSys Goals

- Quantify *variability* across the system stack. (memory, processing, etc.)
- Model and understand the key influencers of *variability*.
- Predict and manage variability at the peta- and exascale level.

VarSys Goals

- Quantify *variability* across the system stack. (memory, processing, etc.)
- Model and understand the key influencers of *variability*.
- Predict and manage variability at the peta- and exascale level.

VarSys Software Framework Evolution



VarSys Data Collection & Cleaning

VarSys Data Collection & Cleaning

I/O Zone

Measuring I/O throughput when reading and writing to disk (or HDD).

6 parameters: 3 ordinal – 3 categorical.

Predicting the distribution of I/O throughput at new parameterizations.

VarSys Data Collection & Cleaning

I/O Zone

Measuring I/O throughput when reading and writing to disk (or HDD).

6 parameters: 3 ordinal – 3 categorical.

Predicting the distribution of I/O throughput at new parameterizations.

CAT

Measuring the number of clock cycles required to check AES key bytes.

4 parameters, all ordinal.

Predicting the baseline time-model for new system parameterizations.

VarSys Analysis & Demo

Details of Analysis

Categorical values are mapped to a regular simplex.

Columns with only one unique value are ignored.

No missing values, if so, those rows with missing entries are ignored.

For >20K data points, Nearest Neighbor is the default algorithm.
(others are Voronoi, Delaunay, MLP, MARS, LSHEP, Decision Tree)