DATA ANALYTICS

WORKFLOWS AND PIPELINES

THE "ANALYTICAL" METHOD

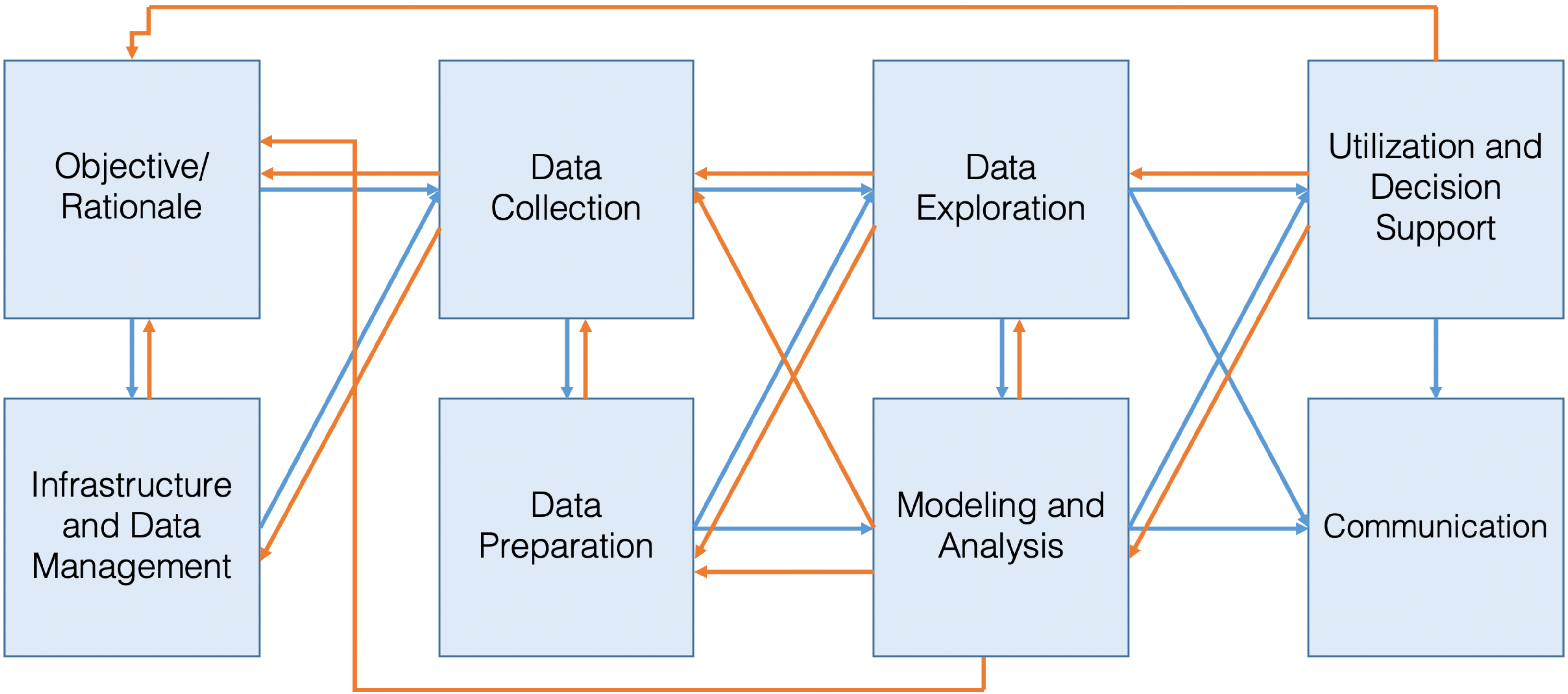
As with the **scientific method**, there is a "step-by-step" guide to data analysis:

- statement of objective
- data collection
- data clean-up
- data analysis/analytics
- dissemination
- documentation

Notice that **data analysis** only makes up a small segment of the entire flow.

In practice, the process is quite often **messy**, with steps added in and taken out of the sequence, repetitions, re-takes, etc.

Surprisingly, it tends to work… when **conducted correctly**.


Objective/ Rationale
Data Collection
Data Exploration
Utilization and Decision Support
Infrastructure and Data Management
Data Preparation
Modeling and Analysis
Communication

DATA PIPELINES (FIRST PASS)

In the **service delivery context**, the data analysis process is implemented as an **automated data pipeline** to enable automatic runs.

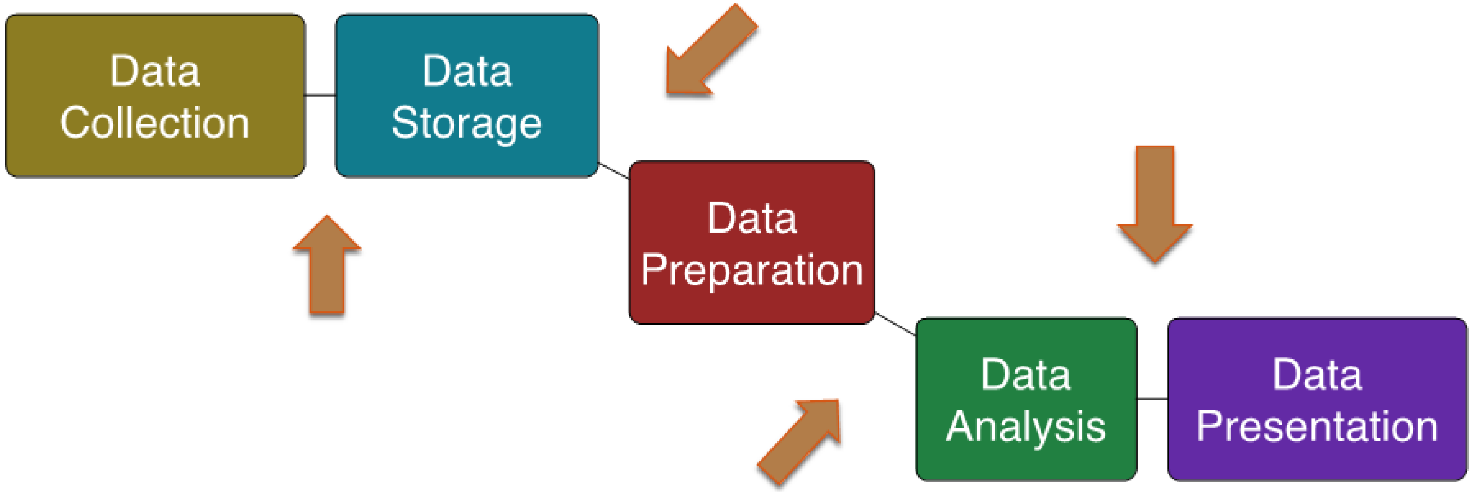
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- data collection
- data storage
- data preparation
- data analysis
- data presentation

DATA PIPELINES (FIRST PASS)

Each components must be **designed** and then **implemented**.

Typically, at least one data analysis pass process must be done **manually** before the implementation is complete.



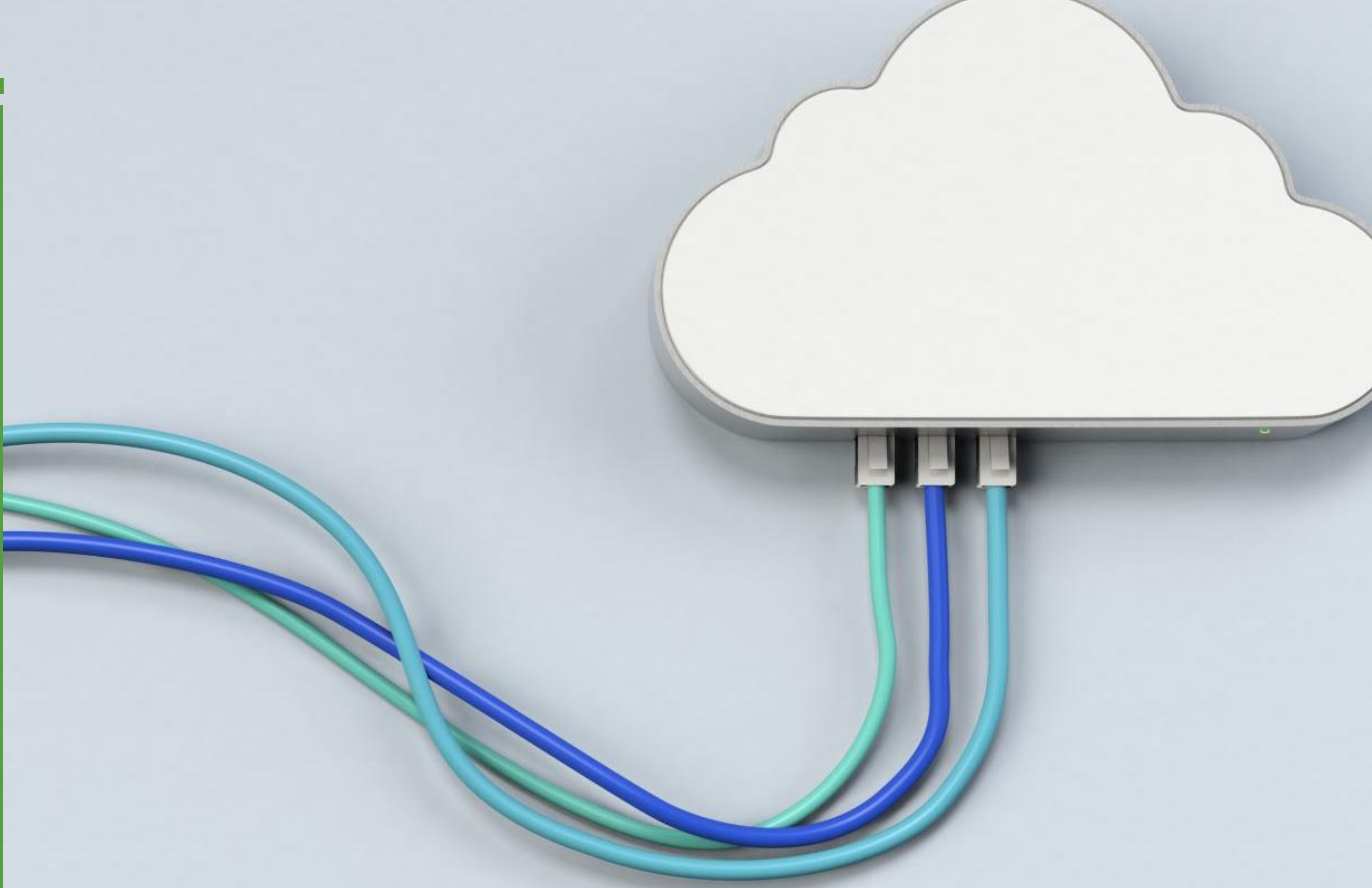
DATA COLLECTION

Data enters the **data science pipeline** by being **collected**.

There are various ways to do this:

- data may be collected in a **single pass**;
- it may be collected in **batches**;
- it may be collected **continuously**.

The **mode of entry** may have an impact on the subsequent steps, including how frequently models, metrics, and other outputs are **updated**.



DATA STORAGE

Once collected, data must be **stored**.

Choices related to storage (and **processing**) must reflect:

- how the data is collected (**mode of entry**);
- how much data there is to store and process (**small vs. big**);
- the type of access and processing that will be required (**how fast**, **how much**, **by whom**).

Stored data may go **stale** (*figuratively* and *literally*); regular data audits are recommended.

DATA PROCESSING

The data must be **processed** before it can be analyzed.

The key point is that **raw data** has to be converted into a format that is **amenable to analysis**, by:

- identifying **invalid**, **unsound**, and **anomalous** entries
- dealing with **missing values**
- **transforming** the variables so that they meet the requirements of the selected algorithms

The **analysis** itself is almost anti-climactic: run the selected methods or algorithms on the processed data.

MODELING

Data science teams should know:

- data cleaning
- descriptive statistics and correlation
- probability and inferential statistics
- regression analysis
- classification and supervised learning
- clustering and unsupervised learning
- anomaly detection and outlier analysis
- big data/high-dimensional data analysis
- stochastic modeling, etc.

These only represent a **small slice** of the analysis pie (see earlier slide).

No one analyst/data scientist could master all (or even a majority of them) at any moment, but that is one of the reasons why data science is a **team activity**.

ASSESSMENT AND LIFE POST ANALYSIS

Before applying findings, we must first confirm that the model is reaching **valid conclusions** about the system.

Analytical processes are **reductive:** raw data is transformed into a small(er) **numerical summaries**, which we hope is **related** to the system of interest.

Data science methodologies include an **assessment phase**, an analytical sanity check: is anything **out of alignment?**

Beware the **tyranny of past success:** even if the analytical approach has been vetted and has given useful answers in the past, it may not always do so.

Real World **Model**

Theory

Identification of details relevant to **description** and **translation** of real-world objects into model variables

DATA ANALYTICS

WORKFLOWS AND PIPELINES