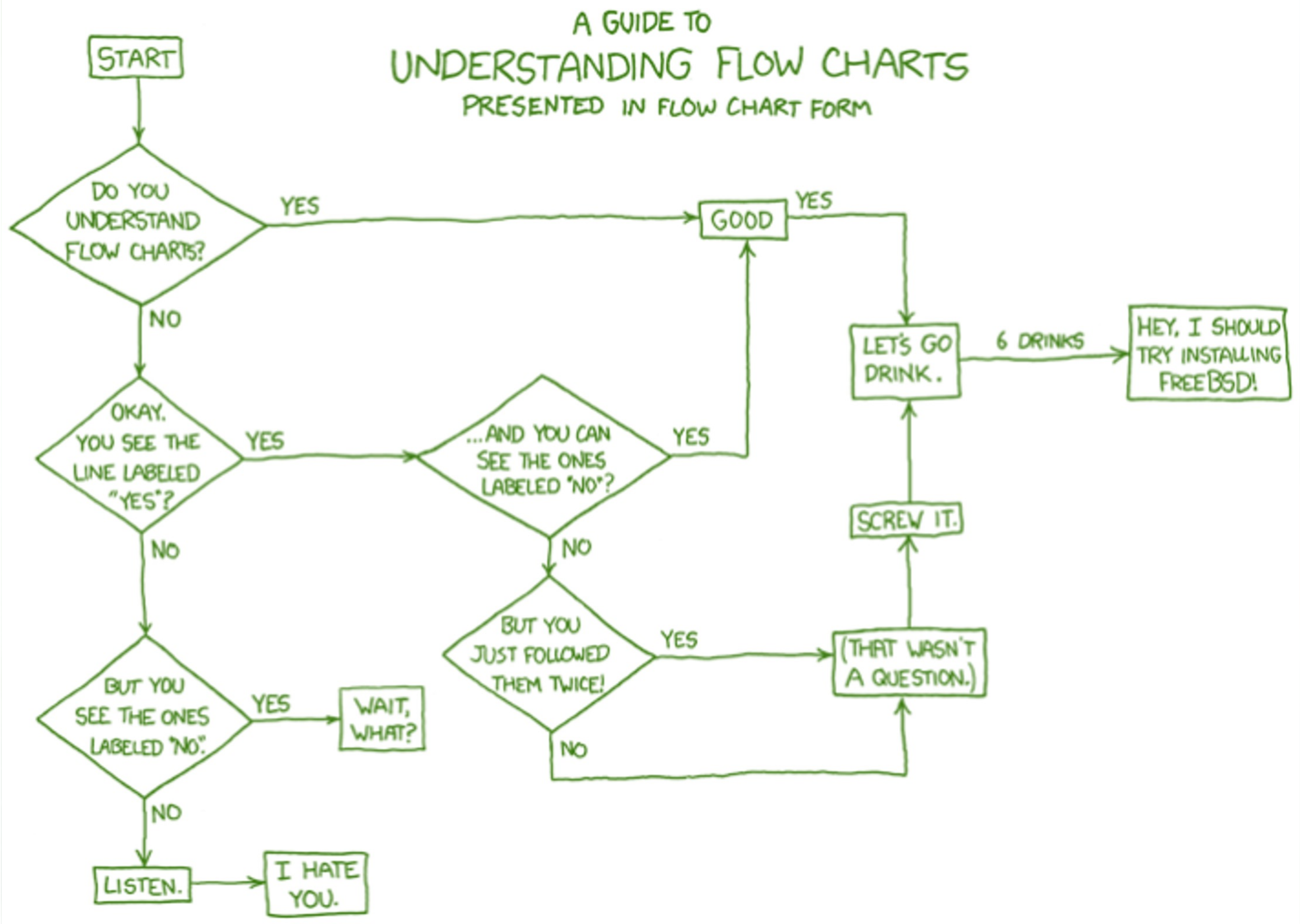


5. Analytics Workflows

Analytics Workflows

You are probably sick of **discussions about context** and would rather move to data analysis proper.

Very soon. One last thing, then: the **project context**.

Data science is more than just the analysis of data; this is apparent when we look at the typical steps involved in a **data science project.**

Data analysis pieces take place within this larger project context, as well as in the context of a larger **technical infrastructure** or **pre-existing system**.

The "Analytical" Method

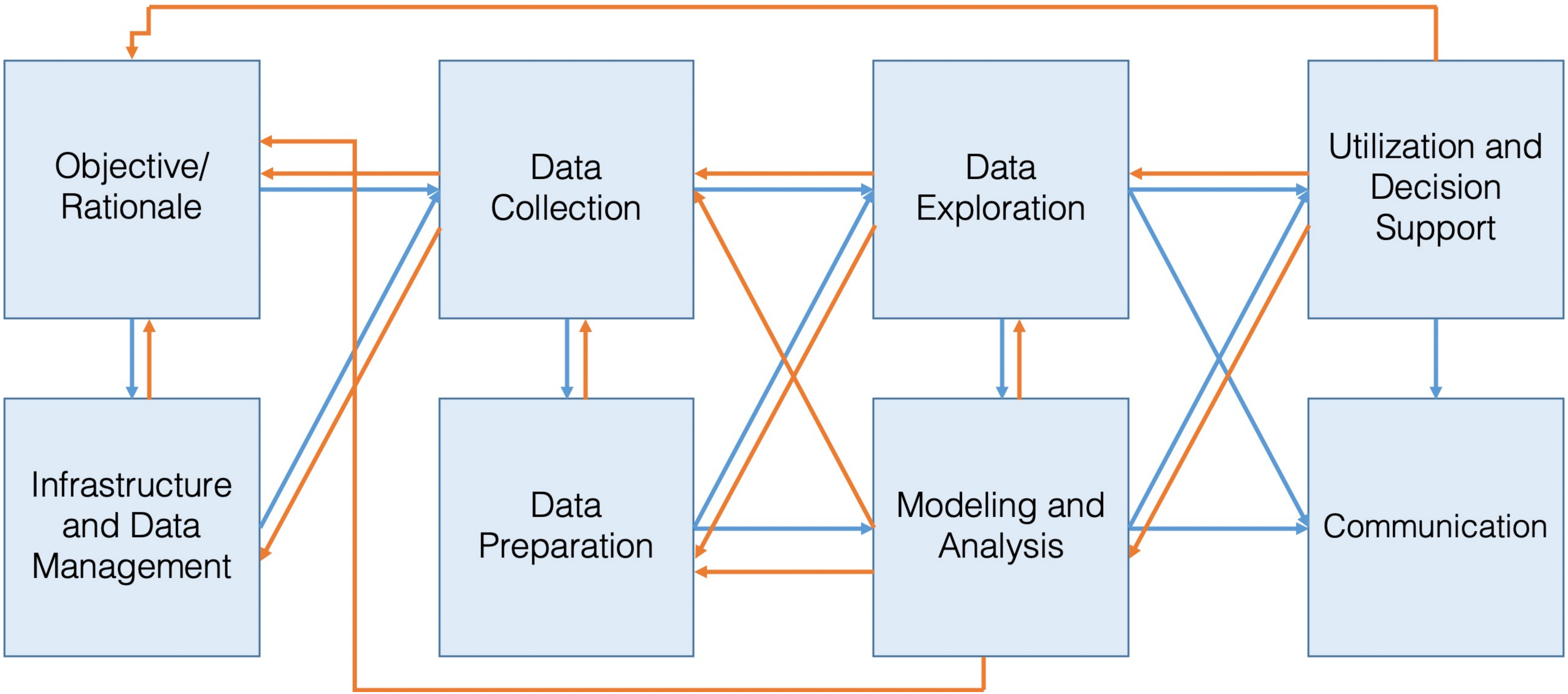
As with the **scientific method**, there is a "step-by-step" guide to data analysis:

- 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.

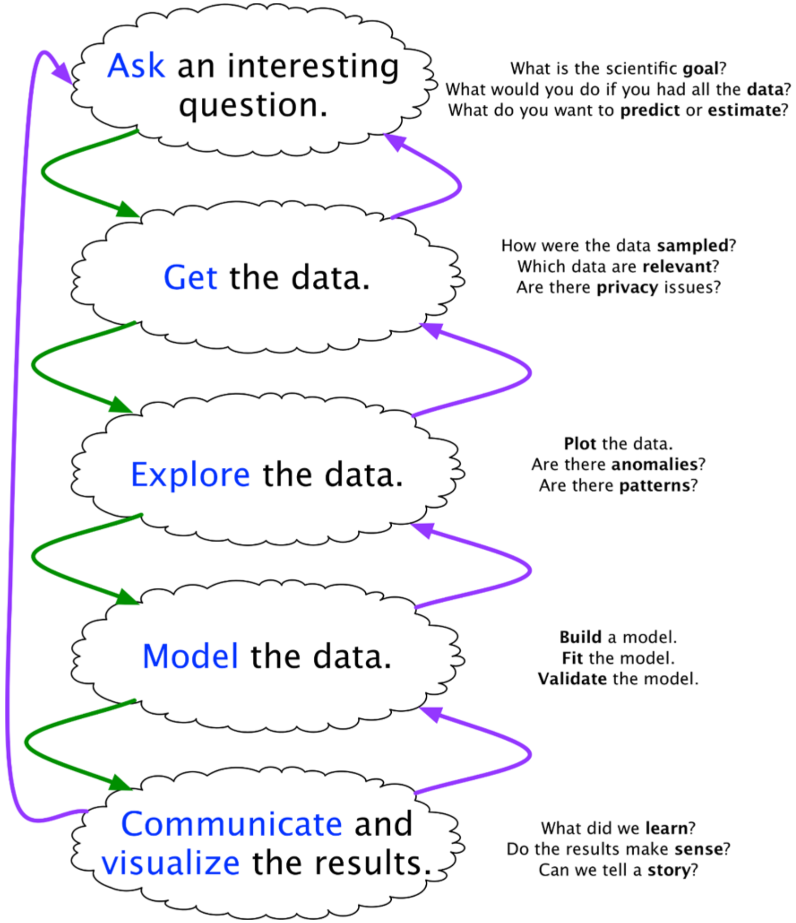
In practice, the process is quite often **messy**, with steps added in and taken out of the sequence, repetitions, re-takes, etc.

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

[CRISP-DM framework (left); Blitzstein and Pfister's data science workflow (right)]

The "Analytical" Methods



The "Analytical" Methods

In practice, data analysis is often corrupted by:

- lack of clarity
- mindless rework
- blind hand-off to IT
- failure to iterate

All approaches have a common core

- data science projects are **iterative**
- (often) **non-sequential**.

Helping stakeholders recognize this **central truth** makes it easier for data scientists to:

- plan the **data science process**
- obtain **actionable insights**

Take-away: there is a lot to consider in advance of modeling and analysis

- **data analysis is not just about data analysis**.

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 it is 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: simply 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**.

Model Assessment and Life After Analysis

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

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**

- 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.

[http://dailycatdrawings.tumblr.com; #309]

Real World

Theory

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

Model

Model Assessment and Life After Analysis

When an analysis or model is 'released into the wild', it often takes on a life of its own. When it inevitably ceases to be **current**, there may be little that data scientists can do to remedy the situation.

How do we determine if the current data model is:

- **out-of-date**?
- no longer **useful**?
- how long does it take a model to react to a **conceptual shift**?

Regular audits can be used to answer these questions.

Model Assessment and Life After Analysis

Data scientists rarely have full control over **model dissemination**.

- results may be misappropriated, misunderstood, shelved, or failed to be updated
- can conscientious analysts do anything to prevent this?

There is no easy answer: analysts should not only focus on the analysis, but also recognize opportunities that arises to **educate stakeholders** on the importance of these auxiliary concepts.

Due to **analytic decay**, the last step in the analytical process is not a **static dead end**, but an invitation to re-iterate to the beginning of the process.

Data Pipelines (First Pass)

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

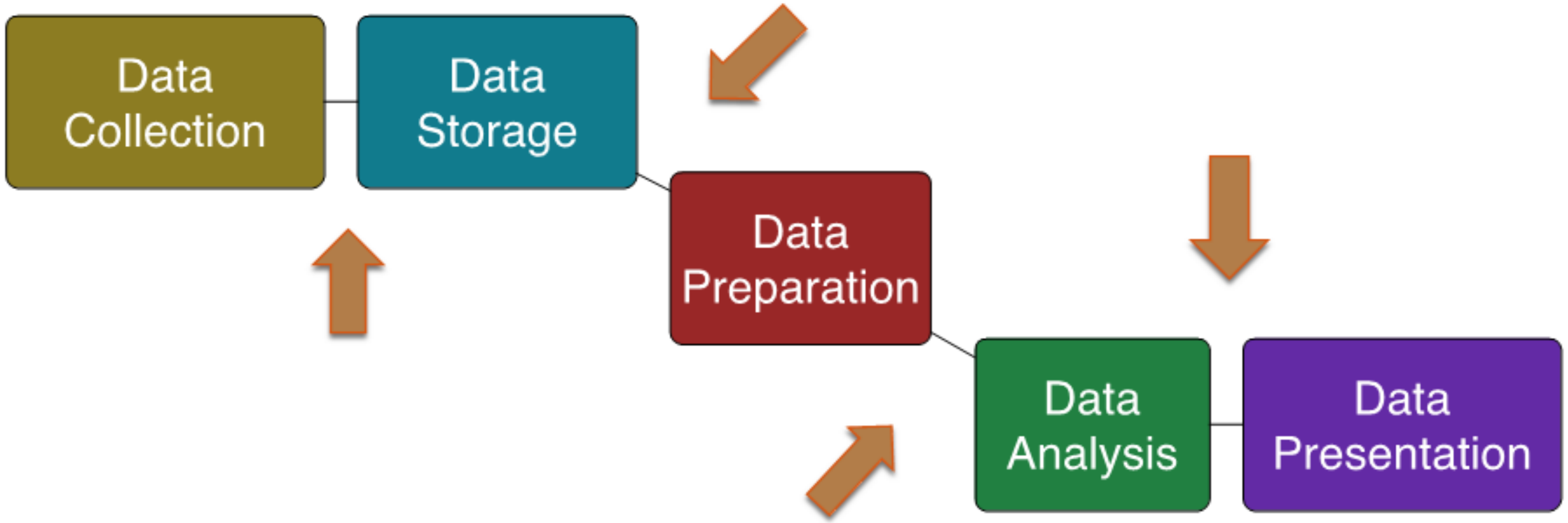
Data pipelines usually consist of 9 components (5 **stages** and 4 **transitions**):

- 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 has to be done **manually** before the implementation is completed.



Suggested Reading

Analytics Workflows

Data Understanding, Data Analysis, Data Science

Data Science Basics

Analytics Workflows

- The "Analytical" Methods
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercises

Analytics Workflows

1. Install R / RStudio (Posit), and packages from the list the instructor will provide.
2. Test the installation with examples from the Programming Primer (sections 2 – 4) to make sure that the software performs as expected.