



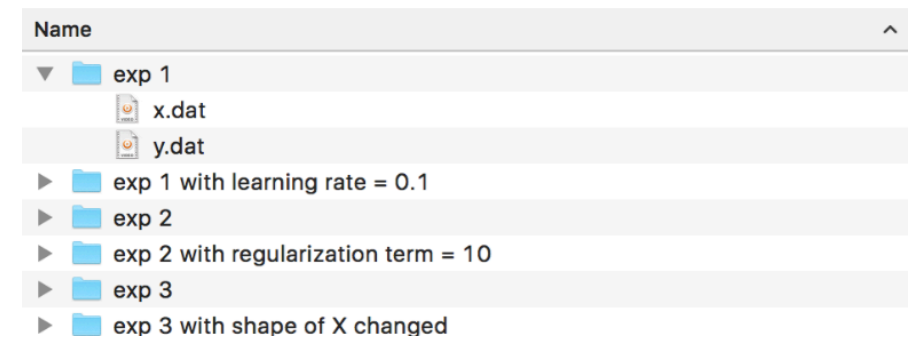
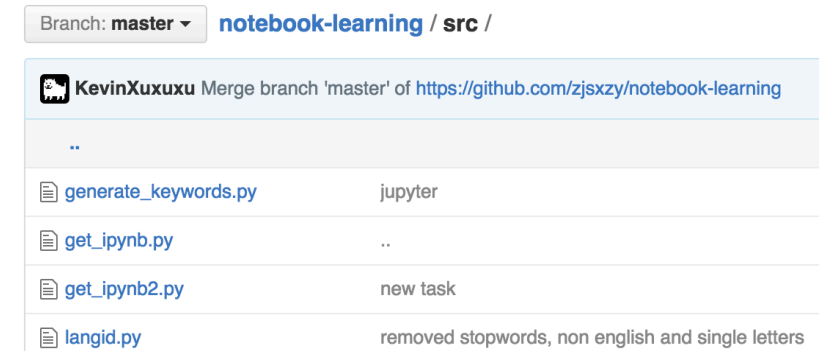
Datalab: A Version Data Management And Analytics System

Yang Zhang, Fangzhou Xu, Erwin Frise, Siqu Wu, Bin Yu, Wei Xu



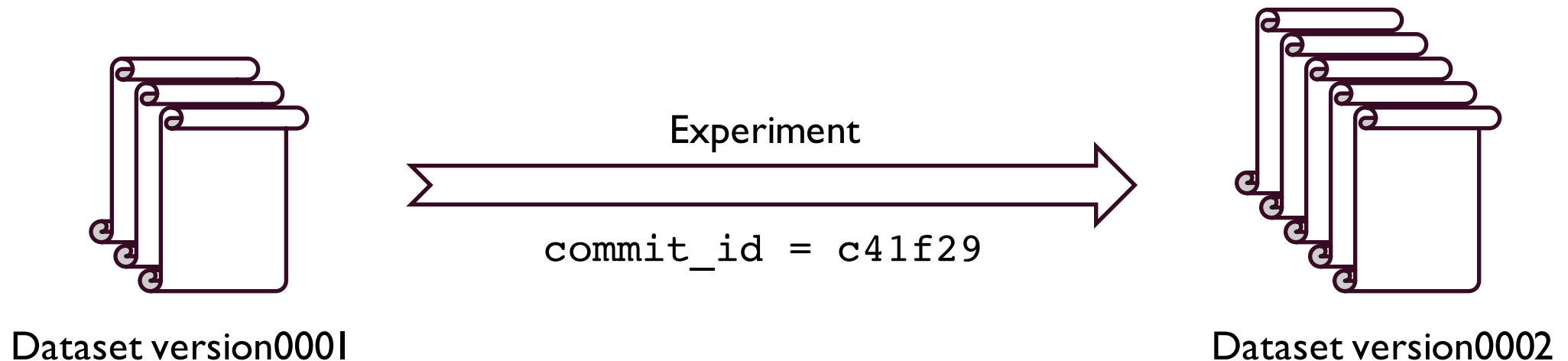
Overview

- Problem: how do we manage code and data with versions?
 - Code version control, e.g. GitHub
 - Data version control, e.g. DataHub^[1]
- But how to combine them in a coherent system?



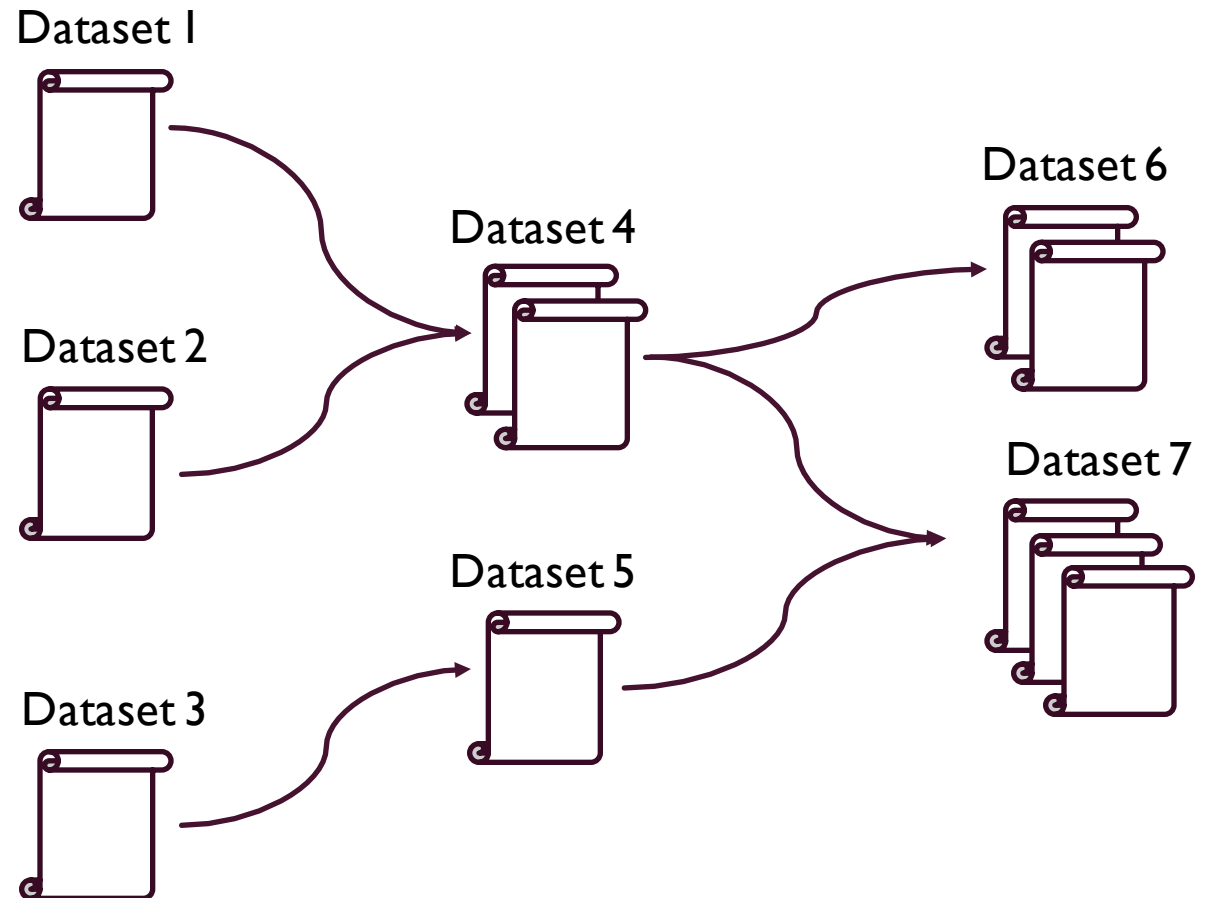
Our Solution

- Version control combining codes and datasets.
- Datasets are generated by execution of codes.
- Two data versions are connected by a code version.

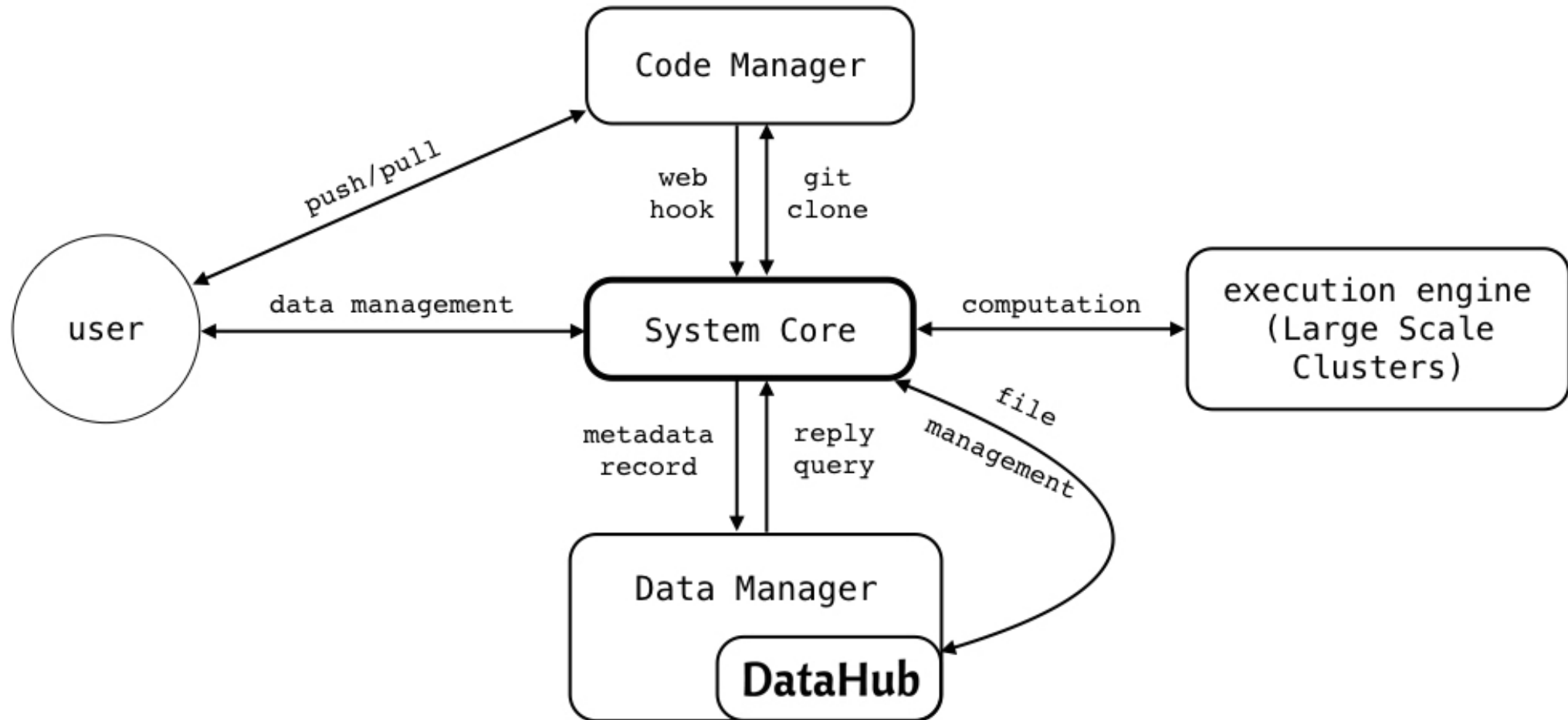


Data Work Flow (DWF)

- Pairs of data versions make up a data work flow (DWF)
- Reconstruct a dataset by re-executing the version of code that generates it

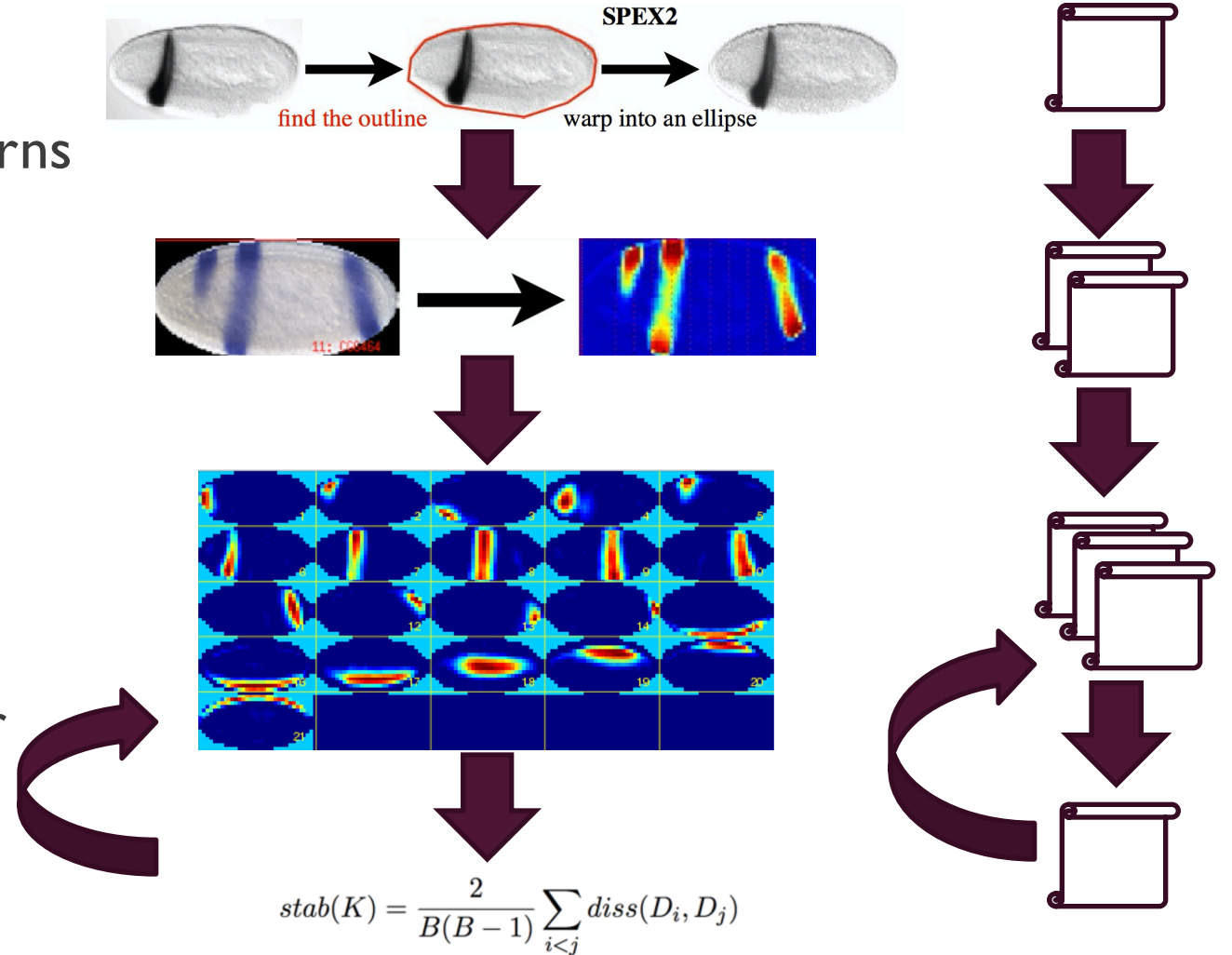


System Architecture



Case Study -- A Biological Data Application

- Goal: find the best K principle patterns
- Procedure:
 - Data preprocessing
 - Feature extraction
 - Non-negative matrix factorization
 - Evaluate K by a stability function
 - Repeat until find the best parameter



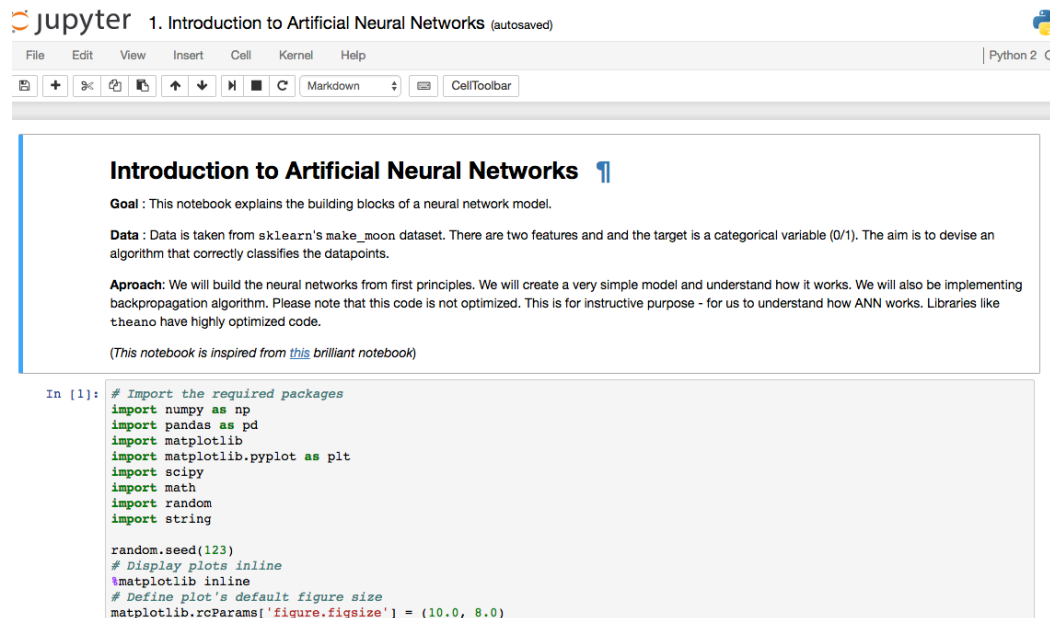
Core APIs

Core APIs

name	functionality	input	output
create	create a project	dataset name	null
inspect	list metadata of a dataset	dataset name	spreadsheet
upload	upload a dataset to a project	file or directory name	null
import	import a dataset to the system	file or directory name	null
merge	merge two dataset into one	two dataset names	null
diff	check out the difference between two datasets	two dataset names	spreadsheet of difference
submit	push codes to system and automatically execute	commit ID	null

Future Work

- Dataset caching
- Online development environment
- Multi-level of interfaces



jupyter 1. Introduction to Artificial Neural Networks (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Introduction to Artificial Neural Networks

Goal: This notebook explains the building blocks of a neural network model.

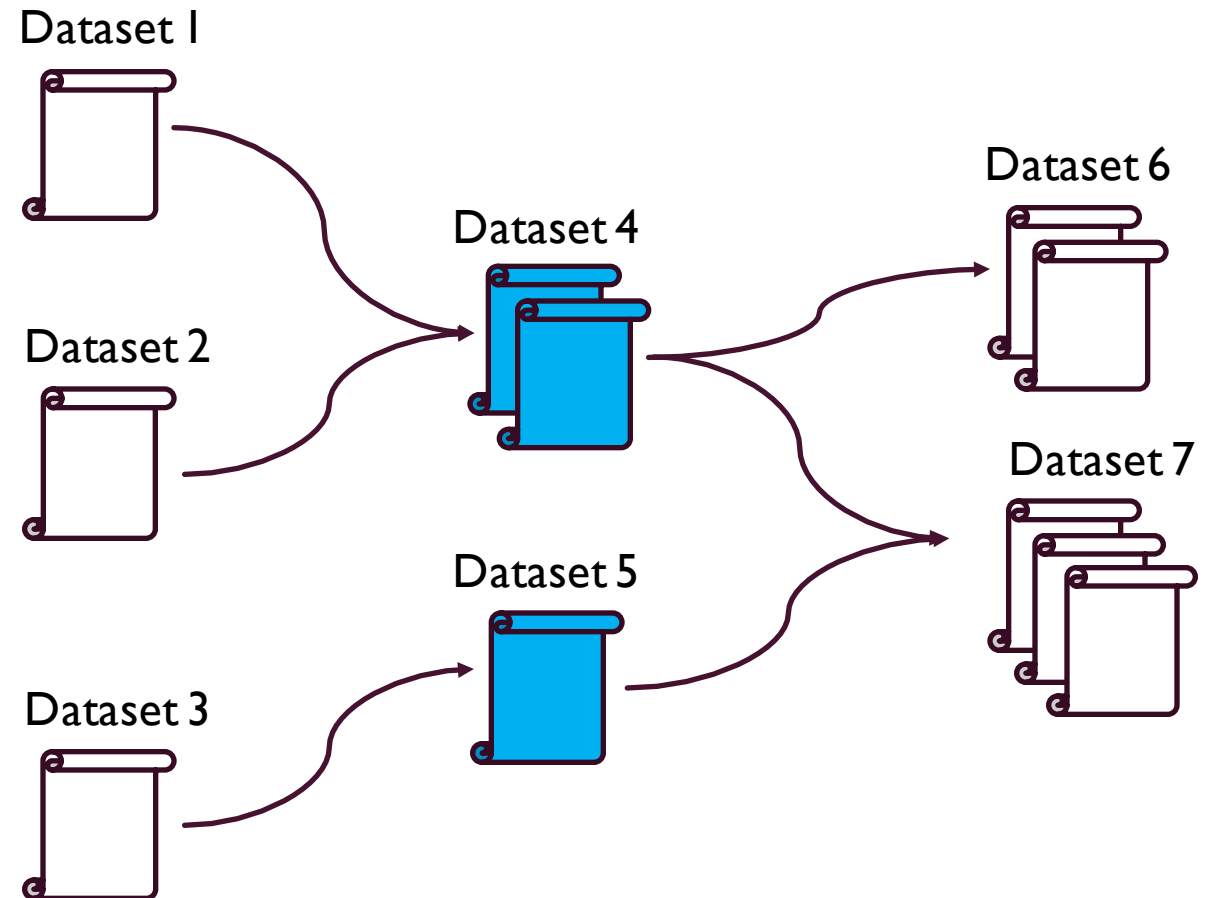
Data: Data is taken from sklearn's make_moon dataset. There are two features and the target is a categorical variable (0/1). The aim is to devise an algorithm that correctly classifies the datapoints.

Approach: We will build the neural networks from first principles. We will create a very simple model and understand how it works. We will also be implementing backpropagation algorithm. Please note that this code is not optimized. This is for instructive purpose - for us to understand how ANN works. Libraries like theano have highly optimized code.

(This notebook is inspired from [this brilliant notebook](#))

```
In [1]: # Import the required packages
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import scipy
import math
import random
import string

random.seed(123)
# Display plots inline
%matplotlib inline
# Define plot's default figure size
matplotlib.rcParams['figure.figsize'] = (10.0, 8.0)
```



Conclusions

- We combine data and code version control
- We propose data work flow
- We improve the efficiency of a data science procedure