

# Machine Learning Workflows

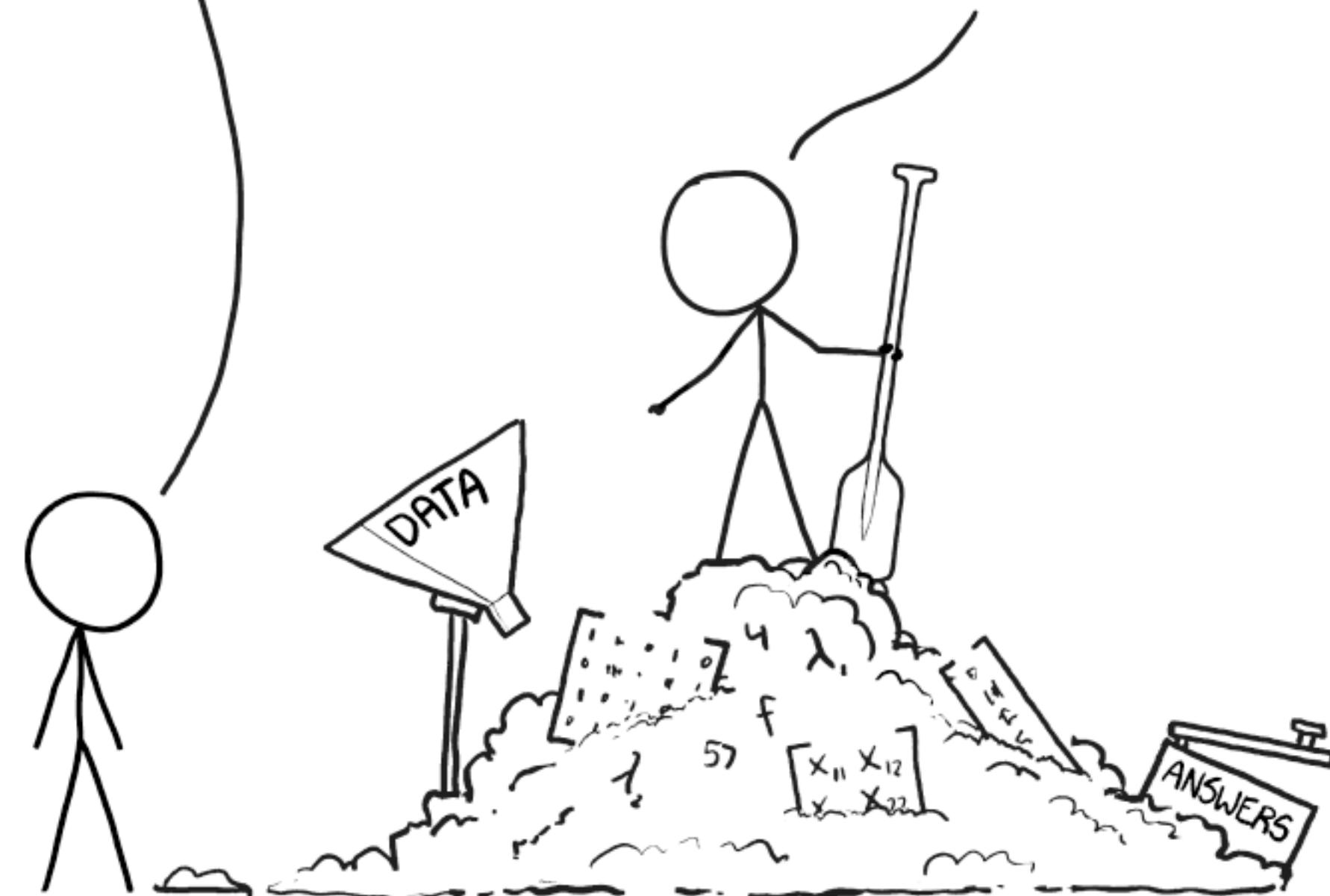
Brendan Herger, <https://www.hergertarian.com/>  
Slides: TODO

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



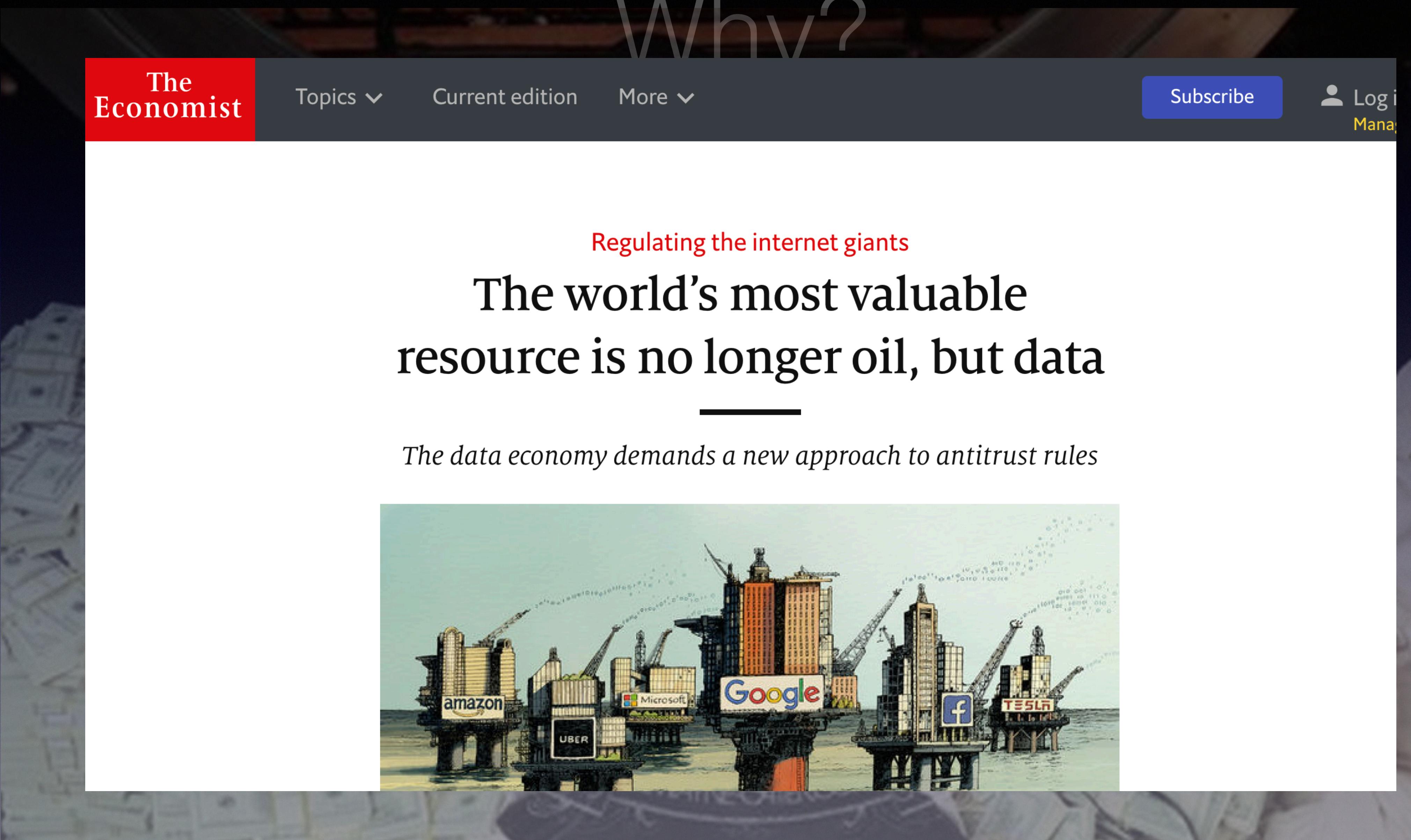


Intro  
Workflow  
Demo & pro-tips

# Intro

# Why?





The  
Economist

Topics ▾

Current edition

More ▾

Subscribe

Log in  
Manage

# Tools

- **Pandas:** Tabular data store
- **Scikit-Learn:** Data science & machine learning library

# Terms

- **DataFrame:** Tabular data store (i.e. like a SQL or excel table)
- **Transformer:** Accepts data, transforms it (e.g. fills null values)
- **Machine learning model:** Accepts data, makes a numerical prediction (regression) or categorical prediction (classification)
- **Pipeline:** A list of transformers / ML models

# Workflow

# ET(M)L

- Extract
- Transform
- (Model)
- Load

# ET(M)L

- **Extract:** Pull anything you need from the world
- **Transform:** Feature engineering
- **(Model):** Model the data
- **Load:** Store anything the world needs from you

Demo & pro-tips

The screenshot shows a PyCharm IDE interface with the following details:

- Project Tree:** The project is named "Example\_Data\_Science\_Project". It contains a "bin" directory which includes files: \_\_init\_\_.py, code\_template.py, lib.py, and main.py. Other directories like "conf", "data", and "docs" are also present.
- Code Editor:** The file "main.py" is open. The code is a Python script for a data science project, specifically for extracting the titanic dataset, performing feature extraction, and using a pipeline. It includes imports for logging, pandas, and scikit-learn, as well as definitions for extract(), transform(), and model() functions.
- Toolbars and Status Bar:** The top bar shows standard icons for file operations. The status bar at the bottom indicates the file is saved ("Git: master"), the encoding is UTF-8, and the Python version is 3.7.
- Bottom Navigation:** Icons for Version Control, Python Console, Terminal, Docker, and TODO list are visible.

A large, semi-transparent white watermark with the word "Demo" is centered over the code editor area.

[https://github.com/bjherger/Example\\_Data\\_Science\\_Project](https://github.com/bjherger/Example_Data_Science_Project)

# Pro tips

- **Consistency:** If your process is an assembly line, you can write robust code quickly. Create a repo with your project ‘template’
- **Iteration:** Set time aside to iterating on your workflow, not just the task at hand
- **Accept inconsistency:** Know when and why to break the rules

Intro  
Workflow  
Demo & pro-tips

# Thanks!

Brendan Herger, <https://www.hergertarian.com/>  
Slides: TODO