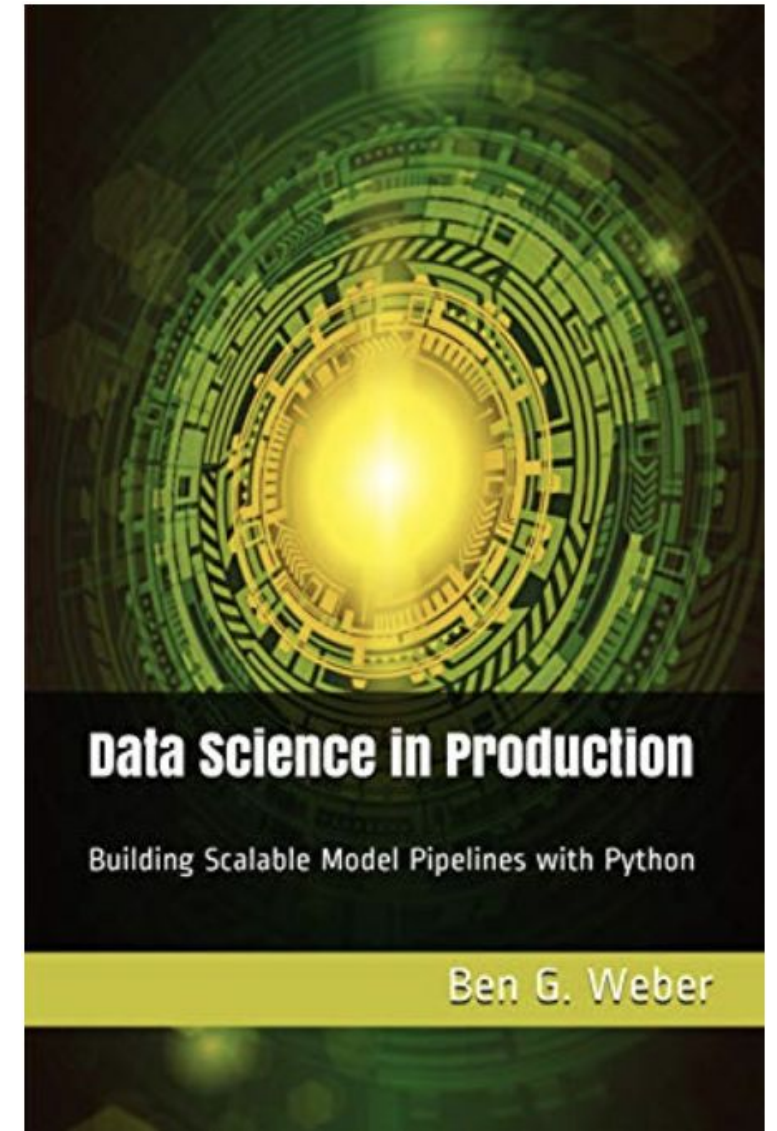DOMINO

**Nikolay Manchev**
Principal Data Scientist for EMEA
Domino Data Lab
@nikolaymanchev

# Building ML pipelines with Jupyter Notebooks
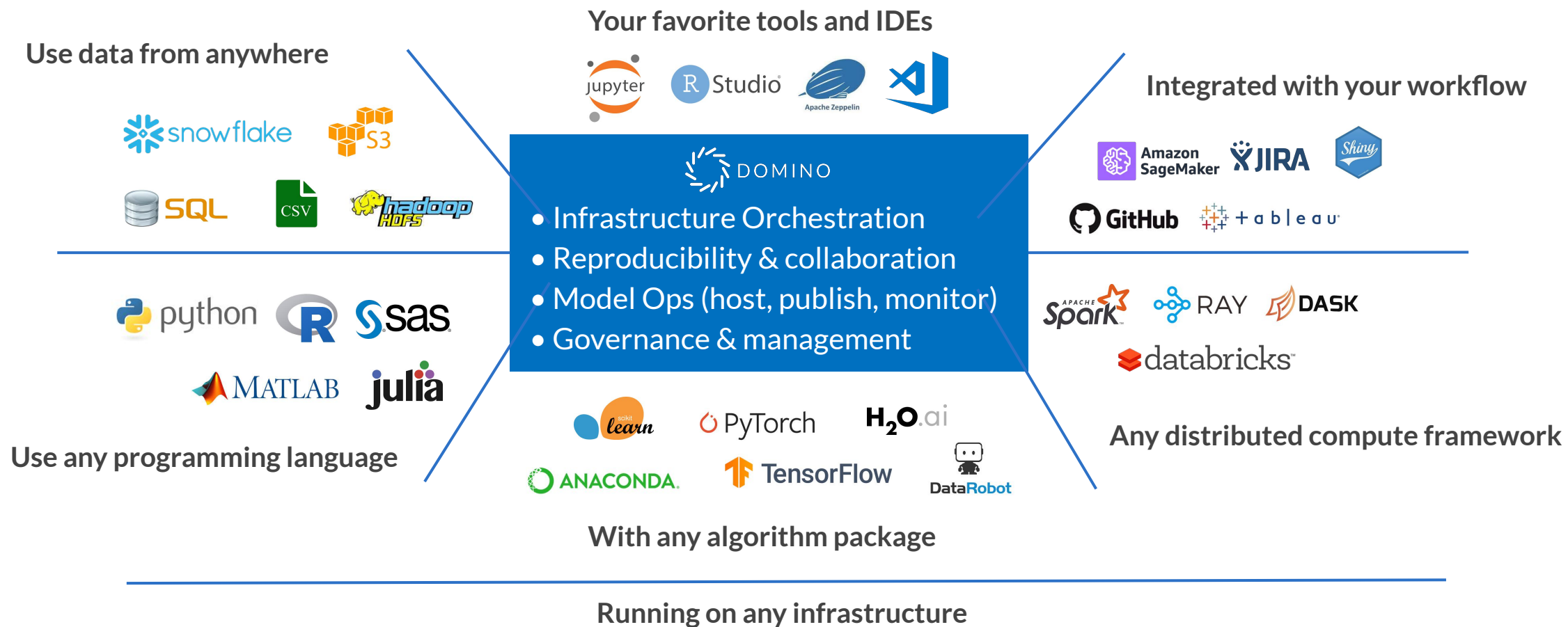
30 APRIL 2020

# Housekeeping

- Moving to virtual events until the Covid-19 situation is resolved
- Next event - 28th May
- I need your help
  - Speakers
  - Topics
- The slide deck will be available on github https://github.com/nmanchev/LondonDSML
- To get in touch with me use **@nikolaymanchev**
- Prize

DOMINO

# Domino is "The Center of the ML Ecosystem"

The **gateway** to data science infrastructure and the **system of record** for work

**Your favorite tools and IDEs**

**Use data from anywhere**

**Integrated with your workflow**

**DOMINO**

- Infrastructure Orchestration
- Reproducibility & collaboration
- Model Ops (host, publish, monitor)
- Governance & management

**Use any programming language**

**Any distributed compute framework**

**With any algorithm package**

**Running on any infrastructure**

DOMINO

**Nikolay Manchev**
Principal Data Scientist for EMEA
Domino Data Lab
@nikolaymanchev

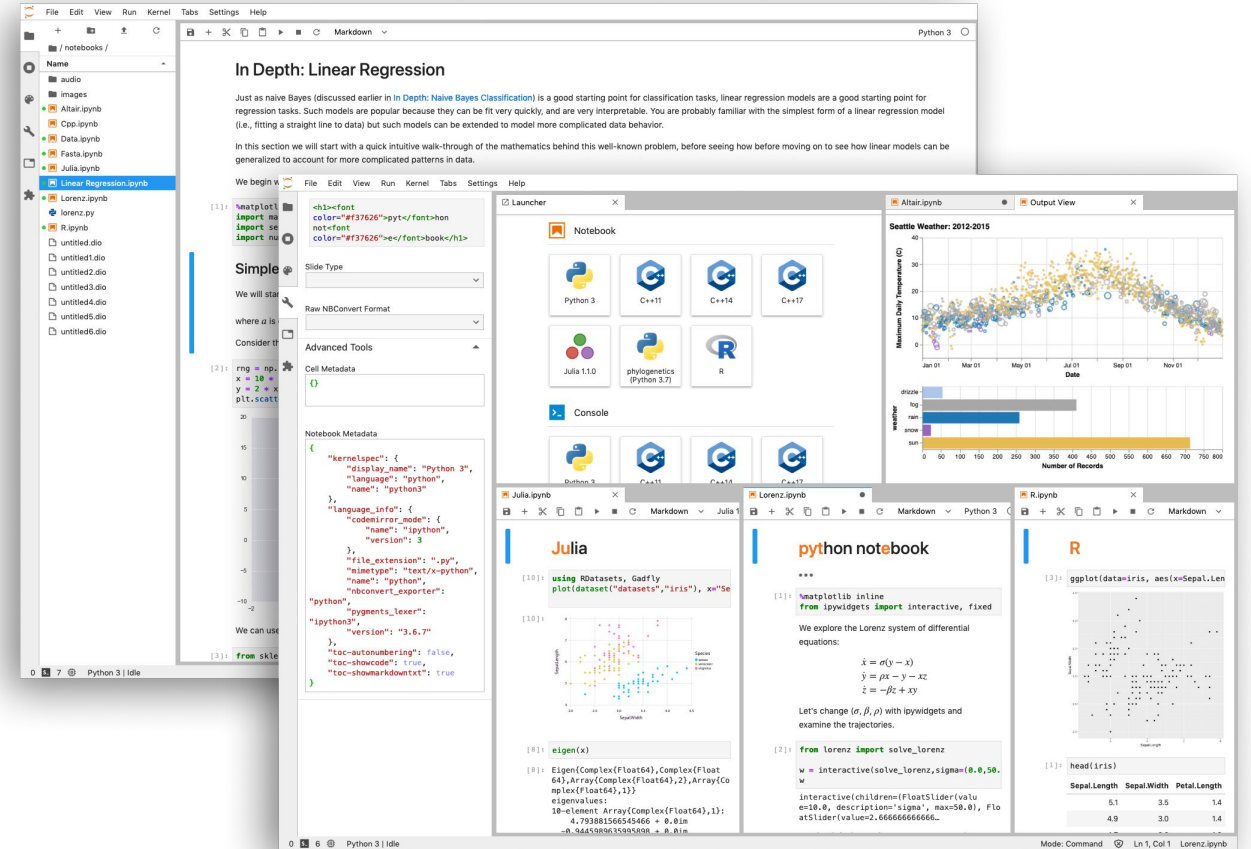# Building ML pipelines with Jupyter Notebooks

30 APRIL 2020

# Why Notebooks

## Because they are cool!

- Great for communicating / showcasing

- Secure (kind of)

- Easy to set up

- Always in the toolbox

- Tons of legacy work

  not cool

DOMINO

# What options are out there

- Manual (the caveman approach)

    o `nbconvert`

- `%run` magic

- Papermill

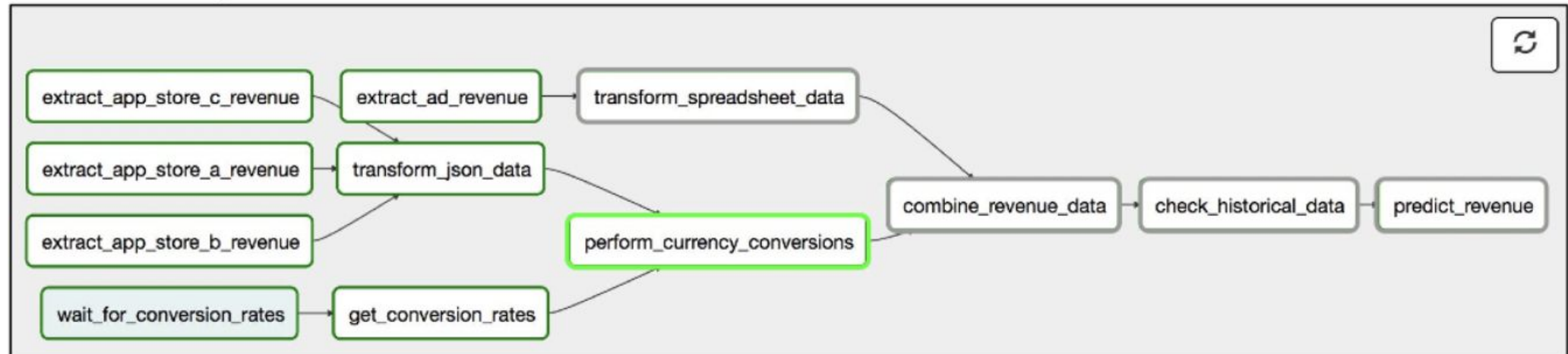- Airflow

# Papermill

- Installation takes seconds

  - `python3 -m pip install papermill`

- Straightforward mechanism

  - Executing notebooks

  - Parameterising notebooks

- Python API and CLI

# Apache Airflow

- Workflows as Directed Acyclic Graphs (DAGs) of tasks



- Hooks (Hive, S3, MySQL, Postgres, HDFS etc.)

- Pools, queues, XCom, variables, SLA's, triggers

- CLI, Python API, REST API

https://www.meetup.com/London-Data-Science-and-Machine-Learning

# Thank You

DOMINO