

Scalable Computing for Individuals

Joshua Cook

November 2, 2016

```
% pandoc -t beamer habits.txt -o habits.pdf
```

Scalable Computing for Individuals

Problem: Medium-Sized Data

- ▶ Kaggle Problems; Datasets from UCI
- ▶ Small enough to work with using standard database tools (Postgres, Mongo)
- ▶ Large enough to be unwieldy; feature engineering and training is extremely slow
- ▶ Advantage of working as an individual can be lost (creativity, rapid innovation)
- ▶ Especially, difficulties in using Jupyter with medium to large data sets

Solution: Infrastructure as Code

Use `docker` and `docker-compose` to define a multi-container system for processing data.

Considering Docker best-practice, one process per container, our system uses the following container types:

- `Jupyter` primary interface to system

- `Postgres` database

- `Redis` memory cache

- `Webserver` basic webserver designed for monitoring worker health

- `Worker` dedicated python processor

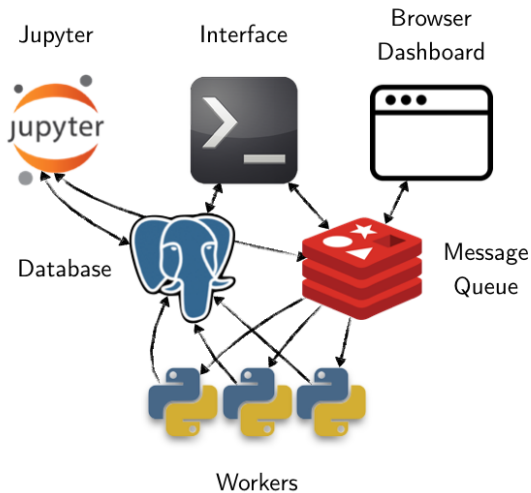


Figure 1: Infrastructure