HiOA Big Data Course

Session 1 - Python for Analytics

Dirk Hesse

Mechanics

- Some (few) slides.
- Mainly jupyter-based teaching.
  - You'll get a copy of those.
  - Take notes.
- Materials: https://github.com/dhesse/HIOA-2017
- Homework: https://dhesse.github.io/HIOA-2017/

Why Python?

The Ecosystem

- Want a web server? `Flask`.
- Want analytics? `scikit-learn`.
- Want hardcore ML? `statsmodels`.
- Want a MongoDB connection? `pymongo` .
- Web scraping? `scrapy`.

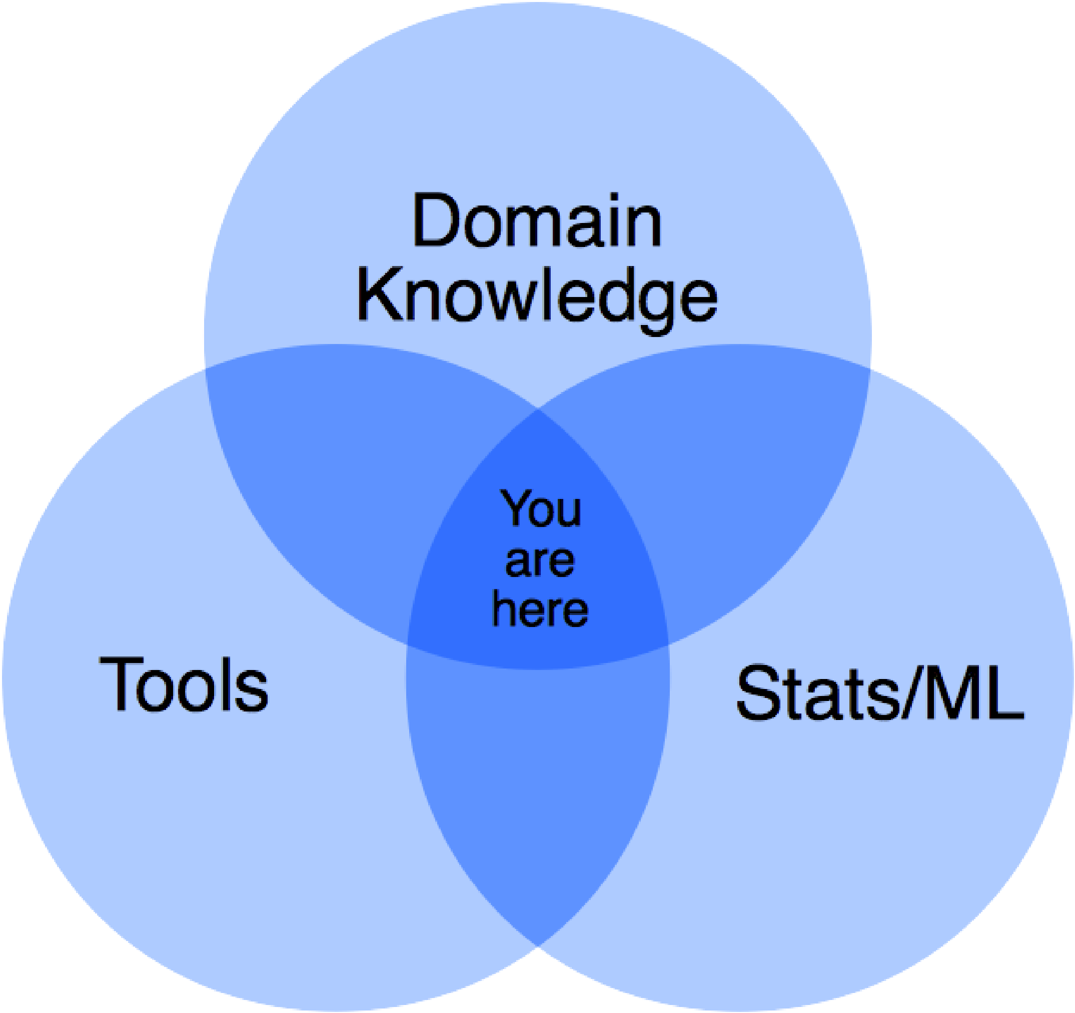
So It Has Many Packages

... yes, but it's *also* a **real programming language.**

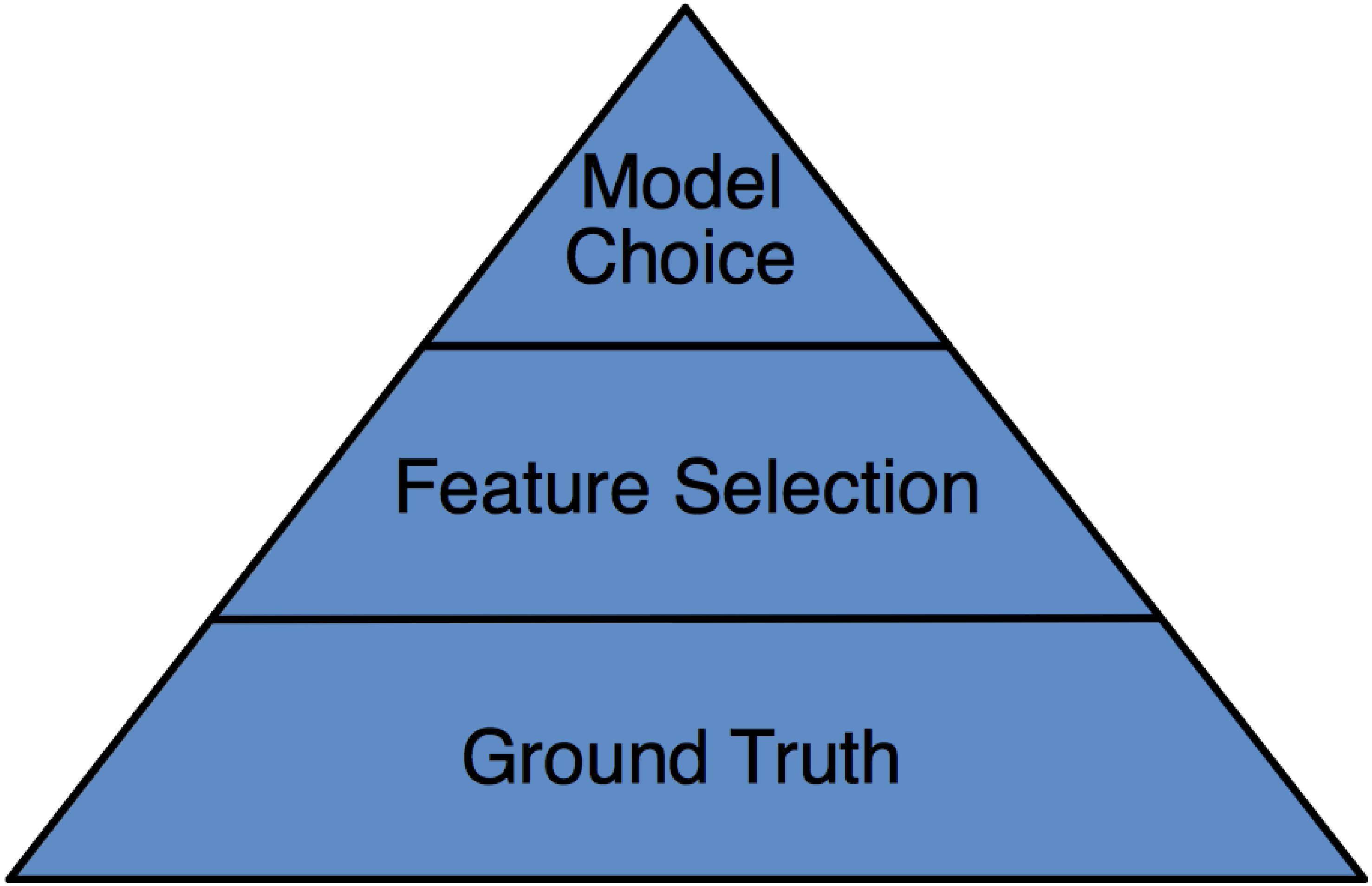
Made with Python:

- Youtube
- Dropbox
- Reddit
- Spotify
- And many more!

Why Do We Need To Program Anyway?



There's more!



Model Choice

Feature Selection

Ground Truth

Ground Truth?

The **ground truth**, the "*thing that*'*s* ***actually*** *going on*" is the single **most important** reason for success and failure of your analytics project. **But** it is often

- messy,
- hard to access,
- only hidden in external data ...
- ... that is not in your data warehouse.

Data Scientists are Analysts are Software Engineers

My team eats new programming languages for breakfast.

W. Whipple Neely, Director of Data Science at Electronic Arts

Let's Code.