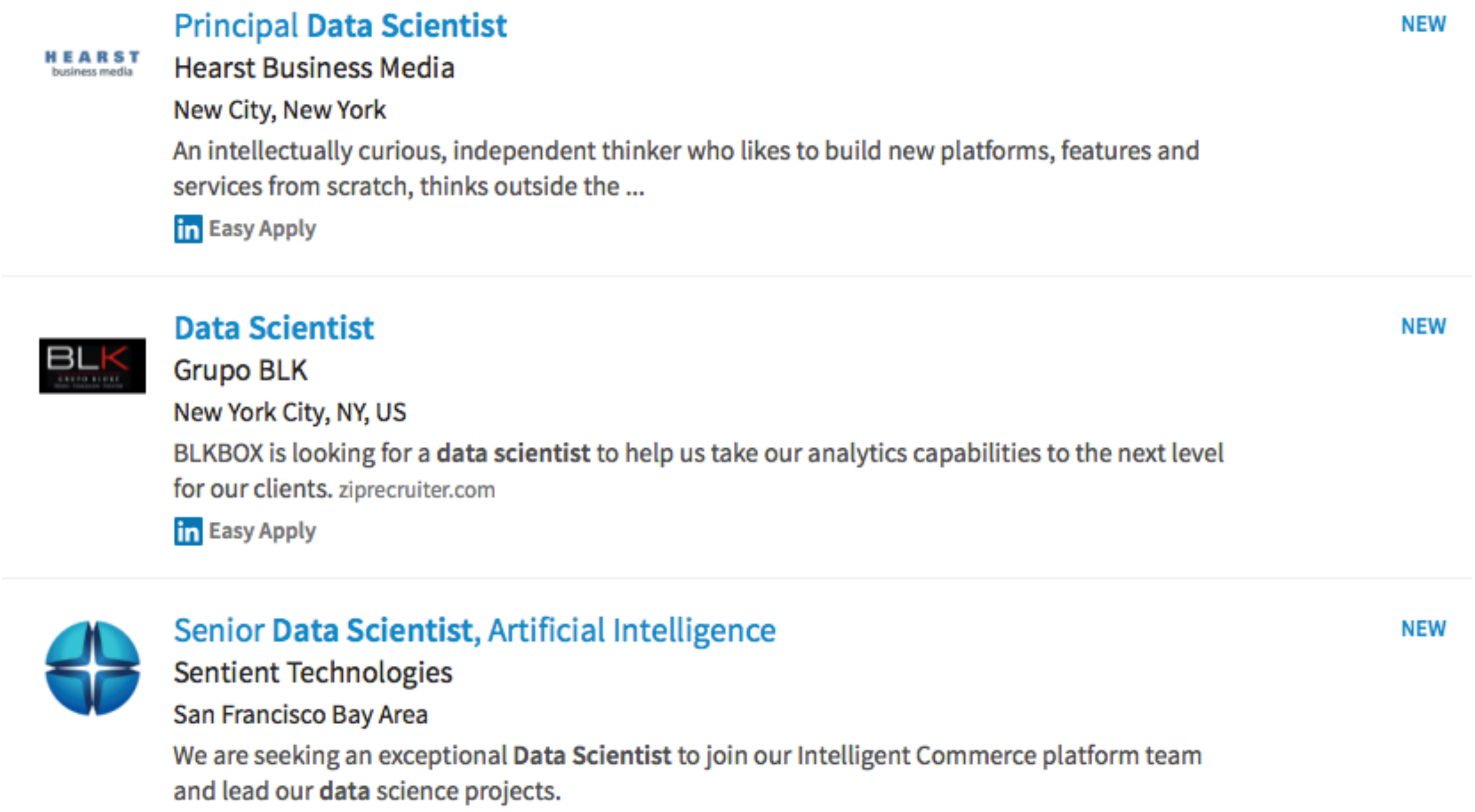
WELCOME TO STK-INF3000/4000

Selected topics in Data Science

Dirk Hesse

DATA SCIENCE?



Principal **Data Scientist**

Hearst Business Media

New City, New York

An intellectually curious, independent thinker who likes to build new platforms, features and services from scratch, thinks outside the ...

Easy Apply

NEW

Data Scientist

Grupo BLK

New York City, NY, US

BLKBOX is looking for a **data scientist** to help us take our analytics capabilities to the next level for our clients. ziprecruiter.com

Easy Apply

NEW

Senior **Data Scientist**, Artificial Intelligence

Sentient Technologies

San Francisco Bay Area

We are seeking an exceptional **Data Scientist** to join our Intelligent Commerce platform team and lead our **data** science projects.

NEW

Jobs

THE SCIENTIFIC METHOD

Ibn al-Haytham (965-1040), Wikimedia

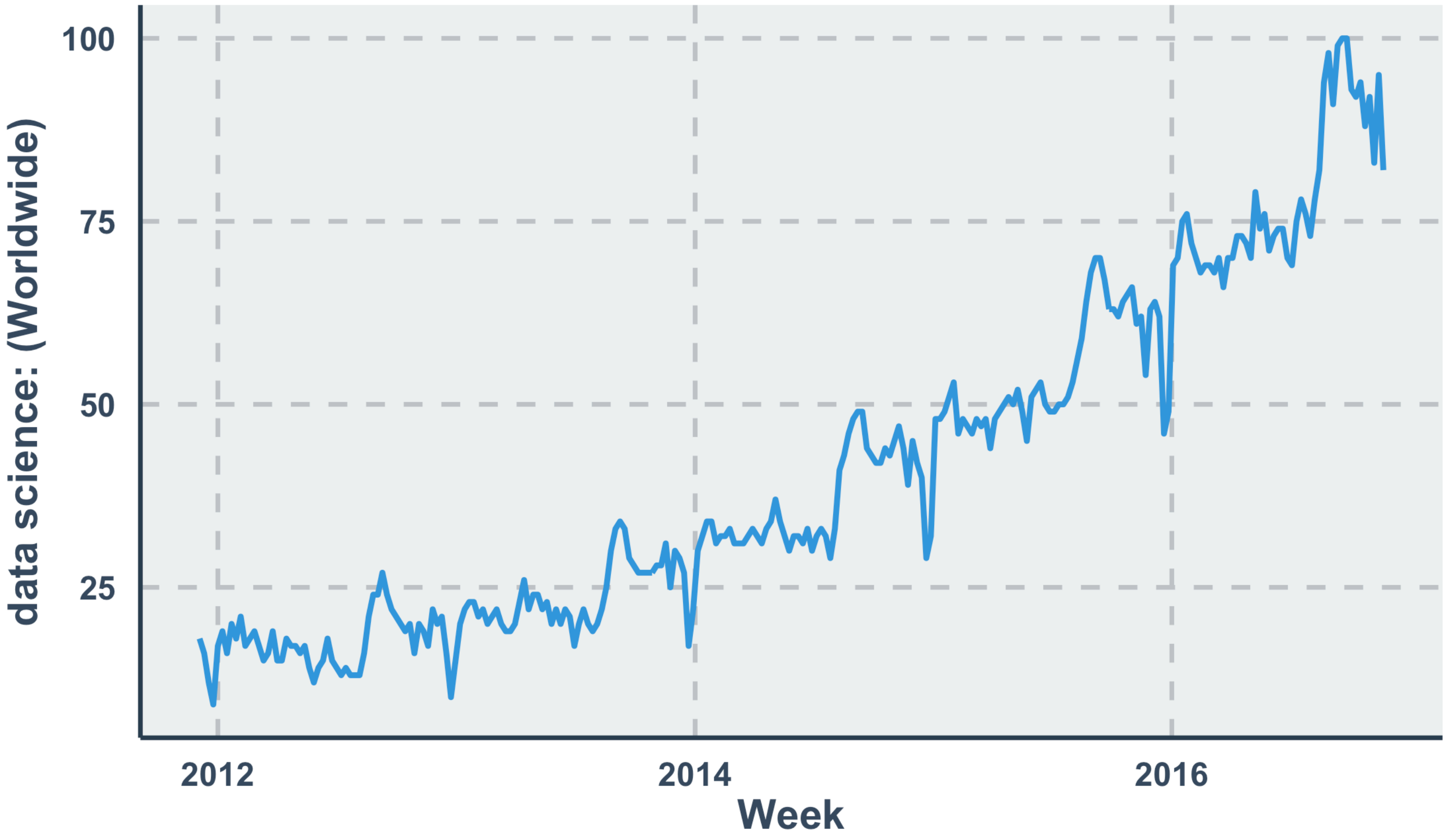
1. **Observation**
2. Question
3. Hypothesis
4. Prediction
5. Testing
6. Analysis

HYPOTESIS TESTING

Peter Higgs, by Bengt Nyman

- Sometimes hypotheses can't be tested due to **technical restrictions**.
- Consider the Higgs boson (postulated 1964, discovered 2012).

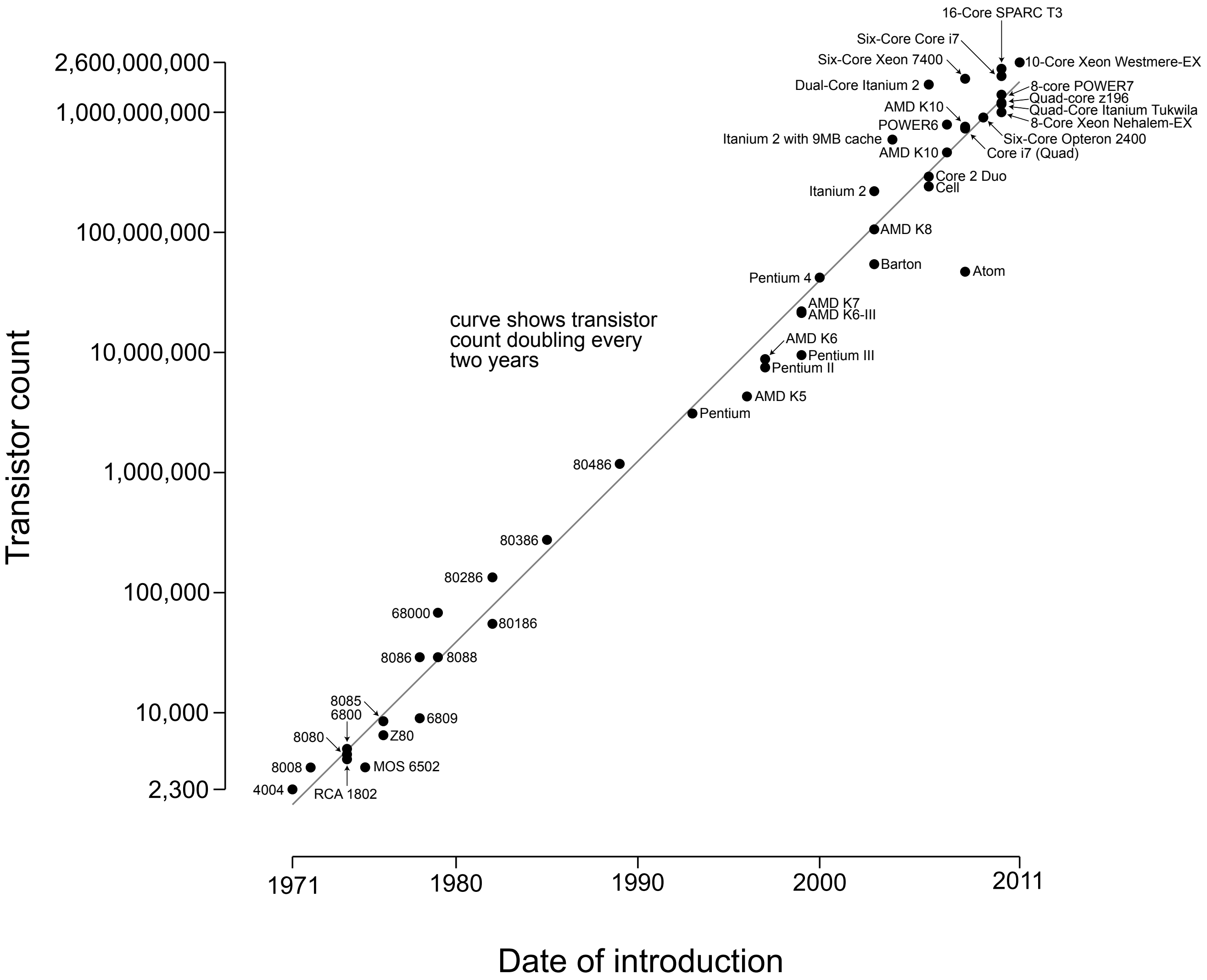
WHY NOW?



Google trends

MOORE'S LAW

Microprocessor Transistor Counts 1971-2011 & Moore's Law

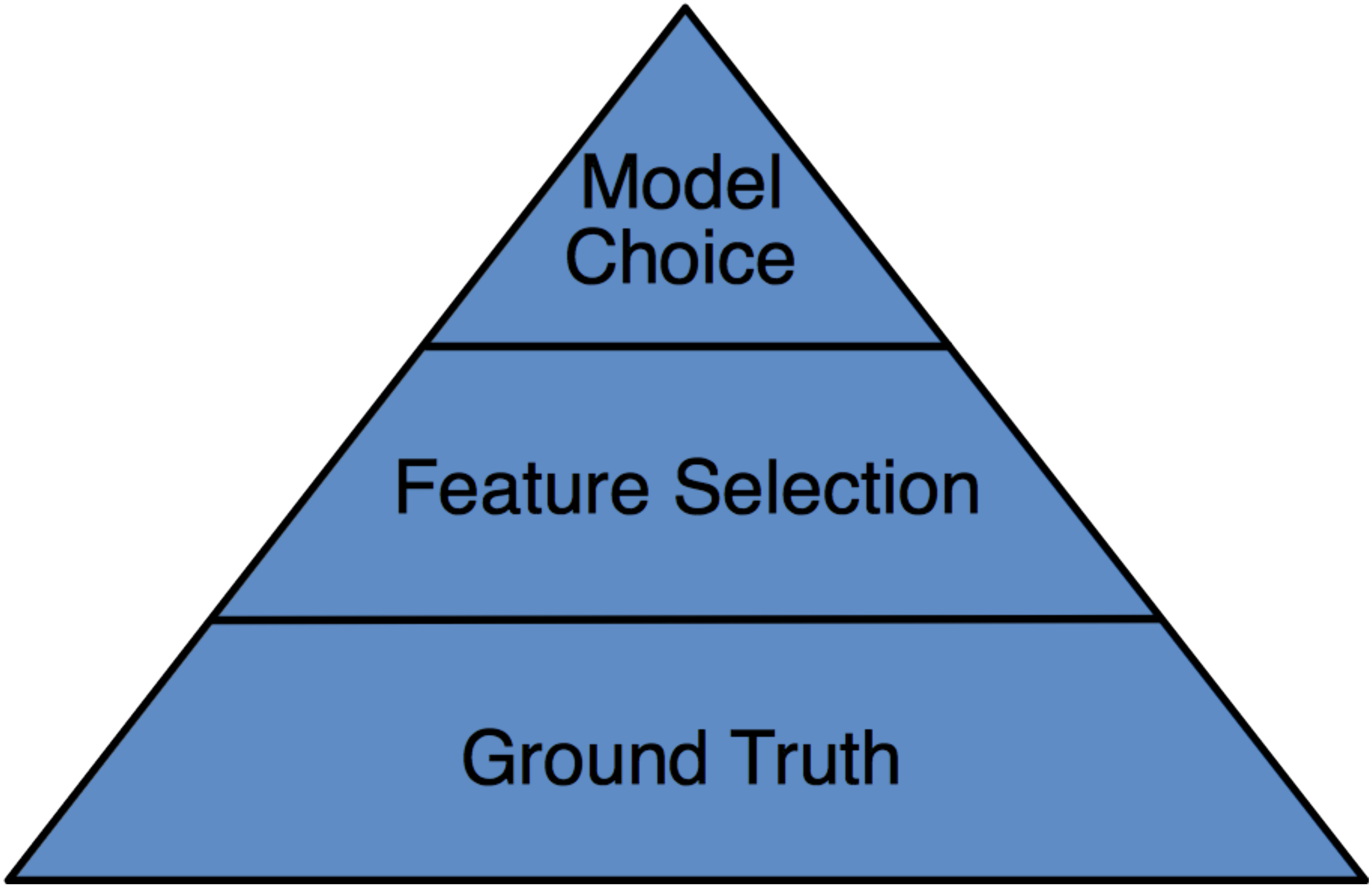


By Wgsimon - Own work, Wikimedia

TOOLS

Apache Foundation

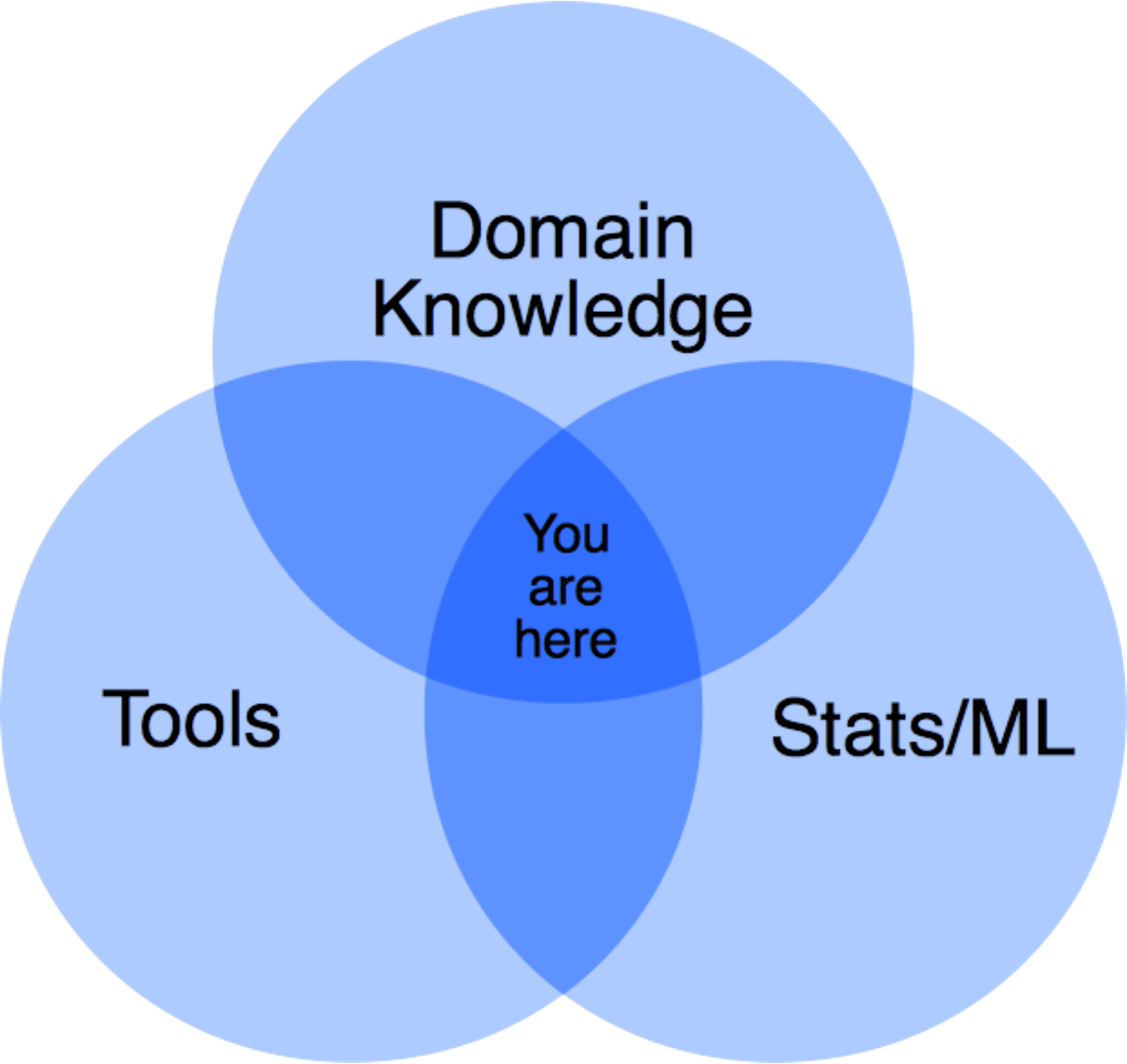
TOOLS ... AREN'T EVERYTHING


Model Choice
Feature Selection
Ground Truth

SO, WHAT DOES A DATA SCIENTIST DO?

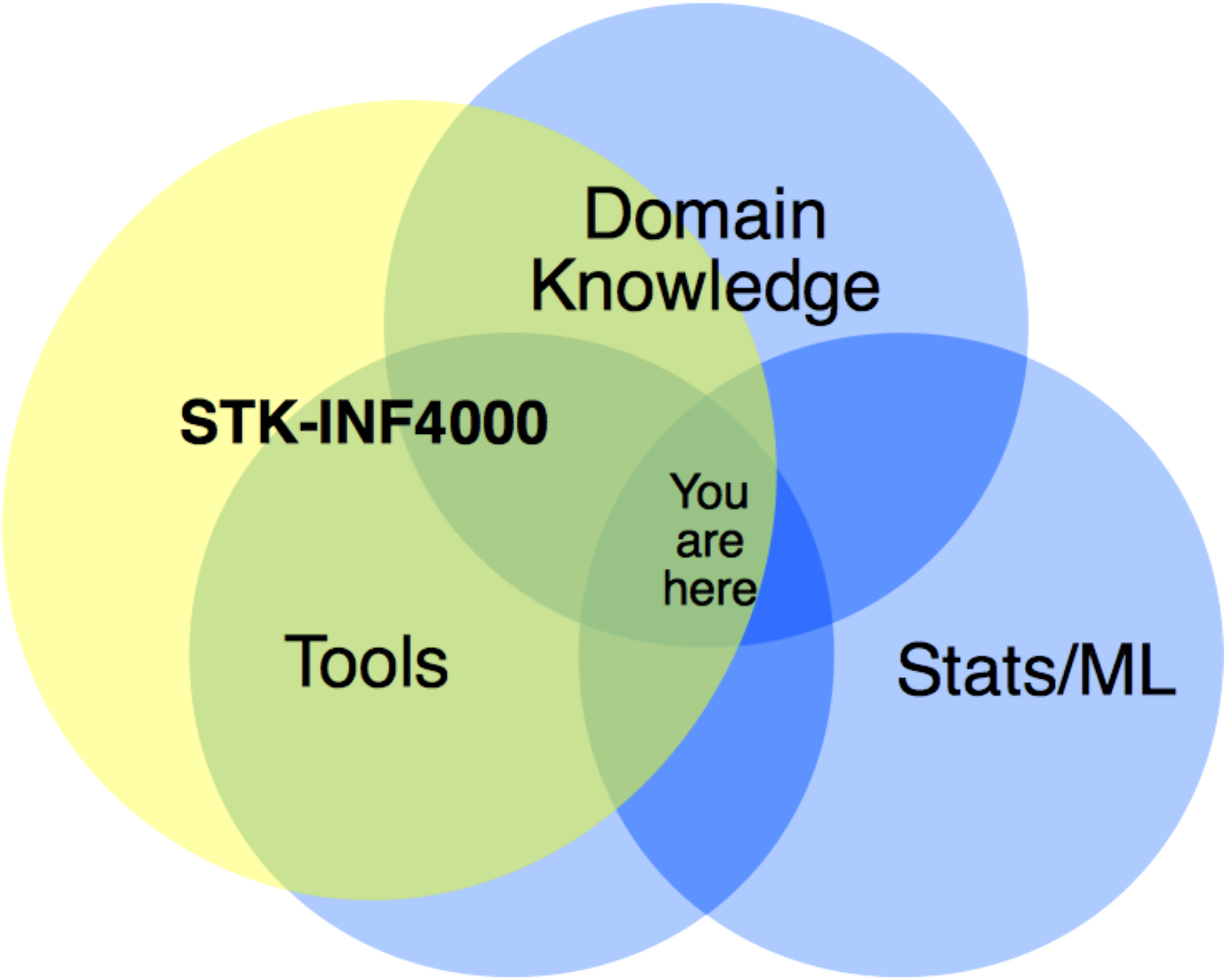
- Talk to managers, try to understand the business.
- Find room for improvements, new projects.
- Use data to implement those.
    - Plain statistics.
    - Machine learning.
    - Big data projects.
    - Lots of coding.
- Present findings, convince people to act on them.

DATA SCIENCE



DS Venn Diagram

THE IDEA



Course Content

(SELECTED) PROBLEM DOMAINS

- Customer relationship management (CRM).
    - Churn prediction.
    - Case prioritization.
    - Campaign optimization.
- Fraud detection.
    - Credit card fraud.
    - Intrusion detection.
- Recommender systems.
- Non-profit/NGO.
    - Disaster prediction/reaction optimization.
    - Conflict analysis.

- Online-ads.
- Transportation.
    - Route optimization.
    - Traffic flow optimization.
- **Many** more.

OUTLINE

- Python for data analysis.
  - A tour of python.
  - Visualization.
- Data from the web.
  - REST APIs.
  - Crawling.
- More on python.
  - Numpy/scipy.
  - Machine learning in `scikit-learn` (maybe).
  - Programming style.
  - Testing your code.

OUTLINE (CONT.)

- Git and github.
- Storing data (MongoDB and friends).
- Strategies for dealing with big data quantities.
- Apache Spark.
- Machine Learning in Spark with sample data sets.
  - Classification and regression.
  - Data quality and features.
  - Time series.
  - Clustering.
  - Frequent pattern mining.
  - Anomaly detection.
  - Streaming data.
- Publishing web data: Flask.

COURSE MECHANICS

- 3 lectures / week (Mondays, me).
- 2 computer labs (Tuesdays, Håvard Kvamme).
- Homework (voluntary).
- Project work (mandatory).
- Examination.
  - Mid-term oral exam (**30%**).
  - Final oral exam (**30%**).
  - Final written exam (**40%**).

THE PROJECT

Propose a project that involves

- Ingesting data.
- Processing data.
- Making predictions.
- Presenting the results.

And a motivation.

- Who will use/buy it?

PYTHON FOR DATA SCIENCE

- Why Python?
  - Easy to learn.
  - Powerful.
  - Widely spread.
  - Lots of useful packages.
- Why coding?
  - Data science means a lot of coding.
  - This code should be 'production grade'.
    - Readable.
    - Reliable.

WHAT YOU'LL NEED

- Python
- pip
- virtualenv

LINUX

```
sudo apt install python
sudo apt install python-pip
pip install --upgrade pip
```

MAC

- install Homebrew (https://brew.sh)

And then:

```
brew install python
```

```
pip install virtualenv
cd my_project
virtualenv venv
source venv/bin/activate
pip install matplotlib jupyter
jupyter notebookbu
deactivate
```

HOMEWORK

https://dhesse.github.io/STK-INF4000-hw/

SLACK - PAGE

STK-INF4000.slack.com