INFO251 – Applied Machine Learning

Lab 3

Satej Soman, Suraj R. Nair

Announcements

- **Problem Set 2 due Feb 11!**

Today

- Any questions from last time?
- PS1 review: predictive power/correlation vs causality
- Vectorized computation + Matrix handling
- Today's programming tool: `numpy`

Predictive power vs causality

- You observe that when sidewalks are wet, the grass near them is also wet. A regression of the number of days with wet sidewalks on the number of days with wet grass has a high R^2. **Is there a causal relationship between the two?**

- You use a treatment vs control design to estimate the effect of a new medication on heart disease incidence. You perfect comparability of treatment and control groups at baseline and perfect compliance with treatment. The coefficient on a treatment indicator in a regression with appropriate controls is -0.001 (taking the new medication reduces the chances of heart disease by 0.1%), and is precisely specified. **Does this coefficient have a causal interpretation?**

Vectorized Computation

- Efficient vectorized computation – operate on arrays of data in one shot
- Creating and manipulating matrices in Python
- Matrix operations: Addition, multiplication, dot product

Today's programming tool: `numpy`

How to make a program run fast

- Choice of programming language
  - **Fast:** C, C++, Java, Ocaml, Rust
  - **Slow:** Julia, Python
  - **Very slow:** R

Video: counting to 1 billion in C++ vs Python

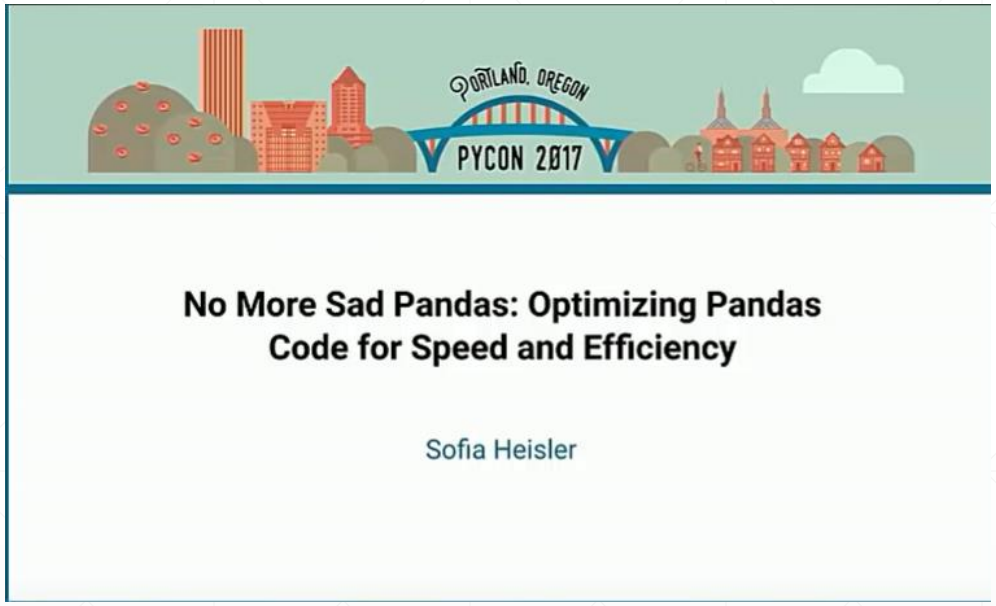
- Writing efficient code
  - For-loops vs. vectorized computation – where `numpy` comes in

- Hardware and parallelization
  - Run parts of a program in parallel on separate cores -- on a single machine or in a distributed system
  - Libraries for parallelizing/speeding up in python:
    `pyspark, dask, multiprocessing`
  - For more: **CS267**

Pandas Optimization

- Avoid for loops / `df.iterrows()`
- If looping is a must, use `df.apply(fn)`.
- Pandas series vectorization
- Vector operations on NumPy arrays are more efficient than on native Pandas series
- Consider parallelized/sped-up alternatives:

  `pandarallel, polars, dask`



Video

live coding