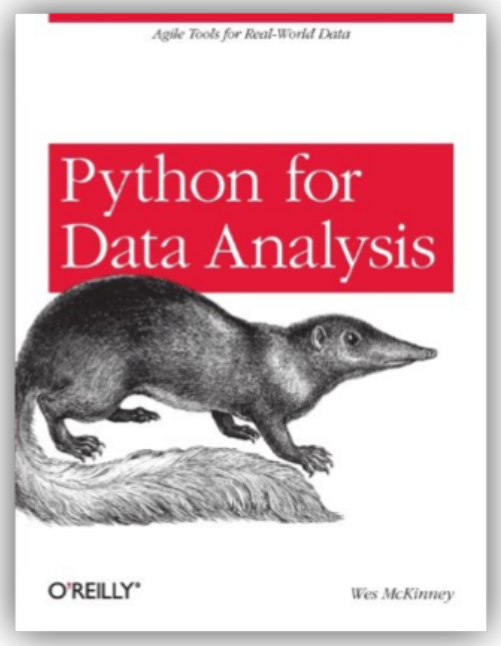
Saturday Morning Keynote

Wes McKinney

@wesmckinn

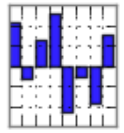
PyCon APAC 2016 (Seoul)

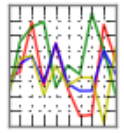
Me

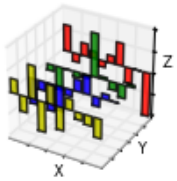


pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$







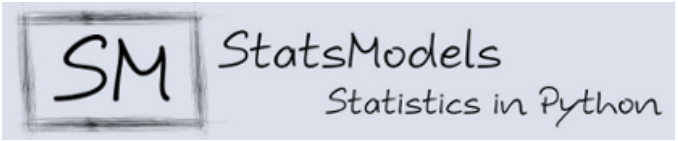
Apache Arrow

Parquet

Feather

ibis

The Apache Software Foundation
Community-led development since 1999.



DataPad

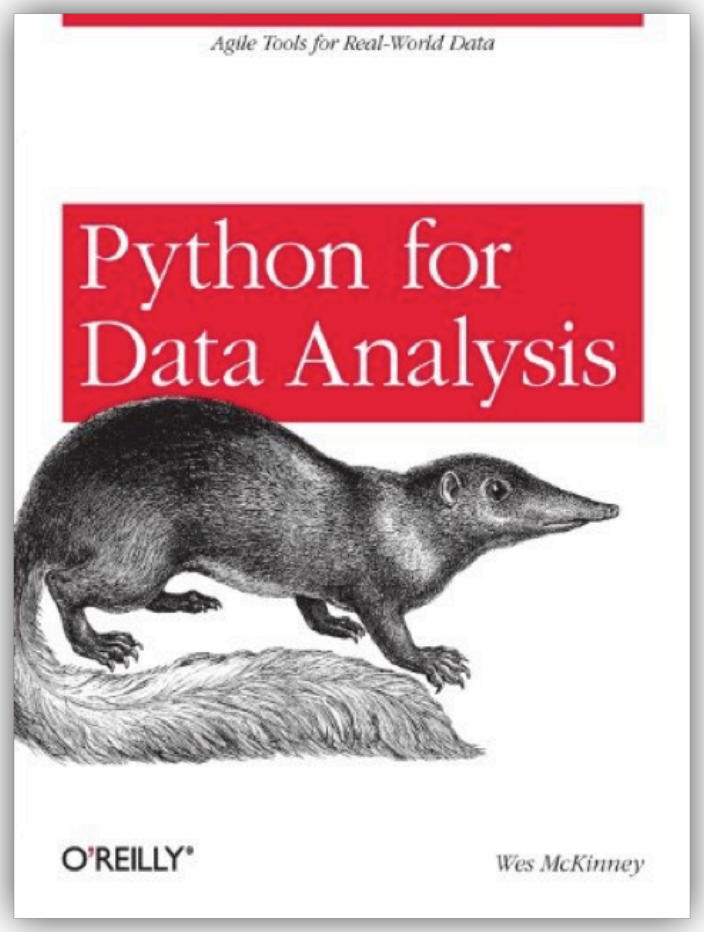
NUMFOCUS
OPEN CODE = BETTER SCIENCE

cloudera

In process:
Python for Data Analysis: 2nd Edition

Coming 2017

(in English ☺)



Q: What brings you here?

Our shared values

Pride in software craftsmanship

My story

- Accidental software developer
- 2007: My first job (financial research analyst)

- I started writing Python libraries to do my own work better
- Soon I was helping my colleagues work better, too

Tools

Tools

Empathy

the feeling that you understand and share another person's experiences and emotions : the ability to share someone else's feelings

Source: Merriam-Webster's Learner's Dictionary

Open source is wonderful…

Open source is wonderful…but it can also be frustrating

Sustainable open source

- How to keep contributors from drowning / burning out?
- How to fund the work?
- How to protect and serve the community?

The Grind

“The grind is an endless stream of bug reports, requests, demands, questions, and occasional inquisitions.”

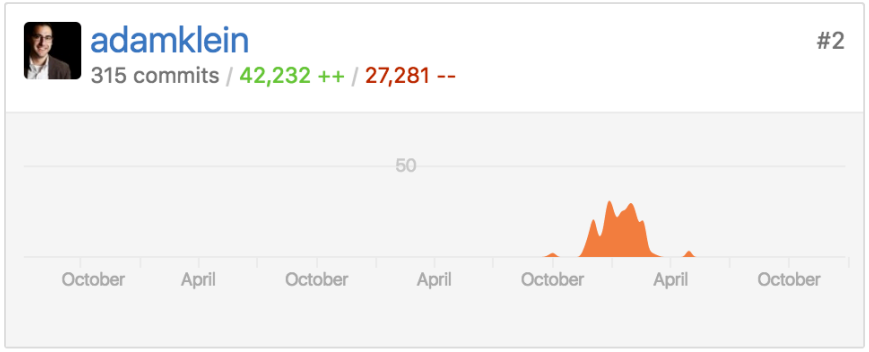
DHH, Creator of Ruby on Rails

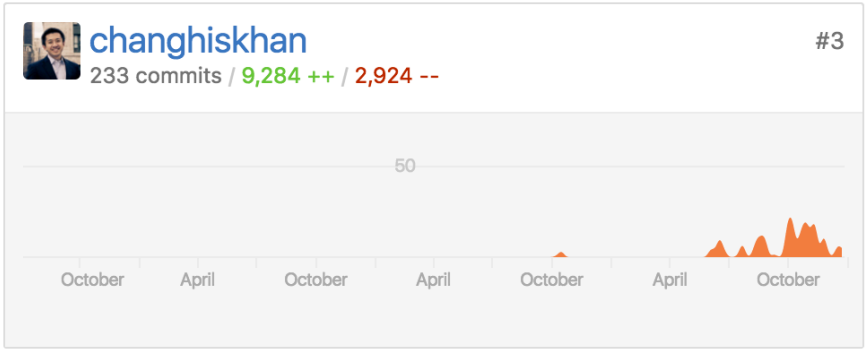
pandas, the open source project

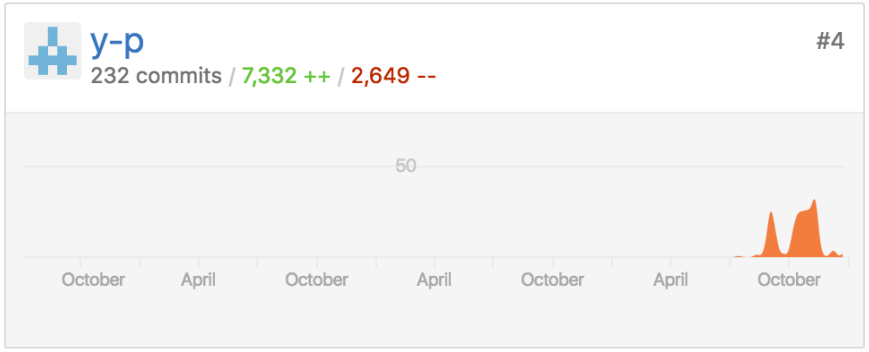
- Parts of code date back to April 2008
- Over 600 unique contributors on GitHub

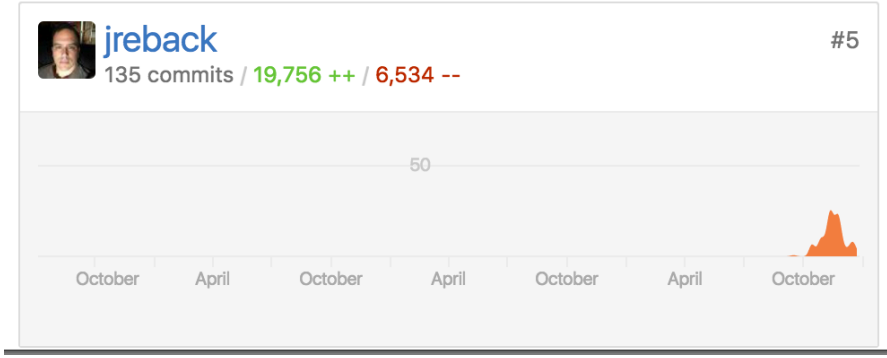
- Active project maintainers range from 4-7 people
- > 6900 Closed Issues
- > 5100 Pull Requests

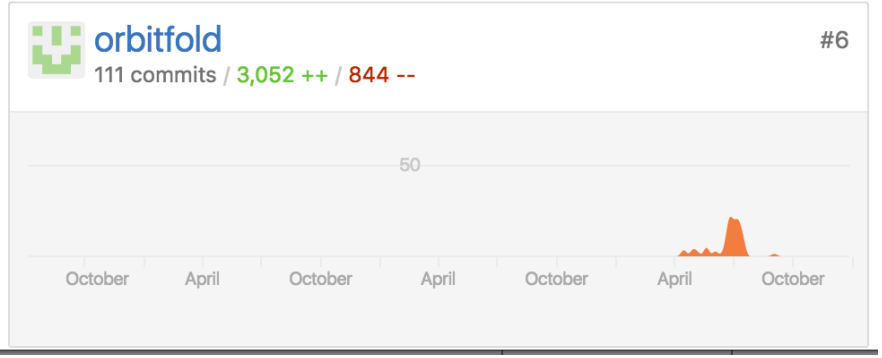
pandas at end of 2012











April 7, 2014

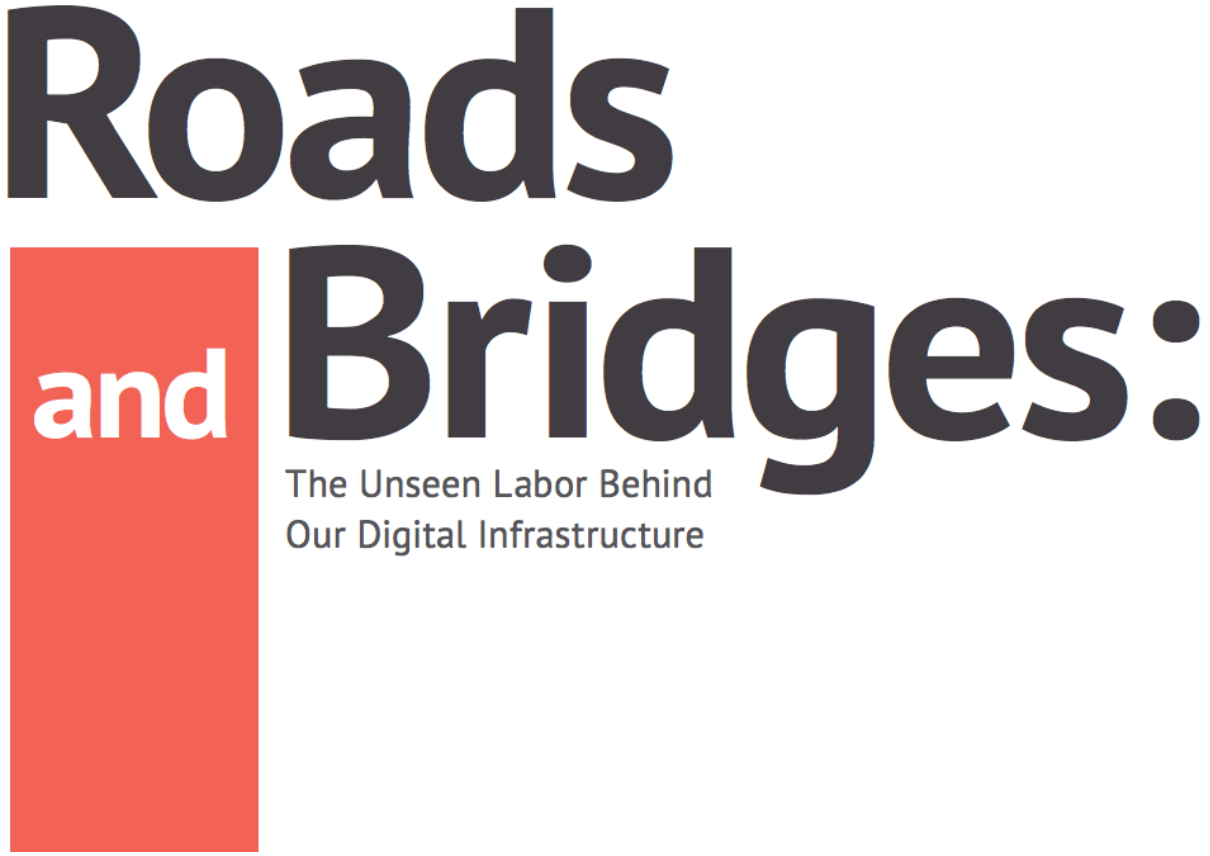
"Some might argue that [Heartbleed] is the worst vulnerability found (at least in terms of its potential impact) since commercial traffic began to flow on the Internet."

Joseph Steinberg, Forbes cybersecurity columnist

“ There should be at least…[6] full time OpenSSL team members, not just one, able to concentrate … without having to hustle commercial work. If you’re a … in a position to do something about it, give it some thought. Please. I’m getting old and weary and I’d like to retire someday.”

Steve Marquess, OpenSSL team

For more on this



By Nadia Eghbal, supported by the Ford Foundation

“The Cathedral and the Bazaar”

Python's normalization in industry

- Python has become a leading language instead of something "experimental" or "risky"
- Many businesses founded on the growth of the Python user base

- See Paul Graham's 2004 essay "The Python Paradox" — how things have changed!

Governance

"the processes of interaction and decision-making among the actors involved in a collective problem…"

M. Hufty (via Wikipedia)

Openness and Transparency

Consensus

Some example governance documents

- **NumPy** (see the docs)
- **IPython / Jupyter** governance
  - github.com/jupyter/governance
- **pandas**
  - github.com/pydata/pandas-governance
  - Modeled after Jupyter governance

http://numfocus.org

http://apache.org


CONDA-FORGE
ABOUT
CONTRIBUTE
CONDA-FORGE
A community led collection of recipes, build infrastructure and distributions for the conda package manager.

conda-forge

- Community-curated conda package channel (hosted on anaconda.org)
- Reproducible build infrastructure (Docker + Circle CI + Travis CI + Appveyor)
- Automated GitHub helper tools

```
conda config --add channels conda-forge
```

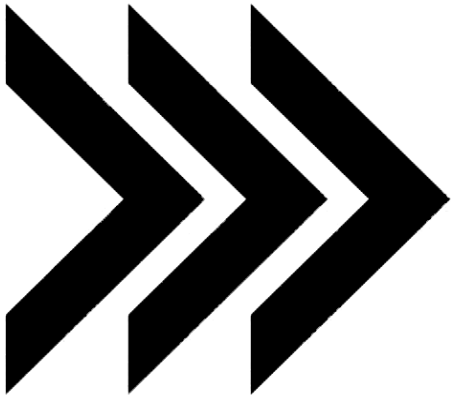
What is next for pandas?

- pandas 1.0
  - A stable, maintenance-only release
- Beginning “pandas 2.0”
  - Planning significant refactoring on the internals of Series, DataFrame

Why pandas 2.0?

- Some changes difficult/impossible to do in an incremental way
- pandas's relationship with the ecosystem has evolved over the last 5 years

- Make pandas
  - Faster and use less memory
  - Fix long-standing limitations / inconsistencies
  - Easier interoperability / extensibility

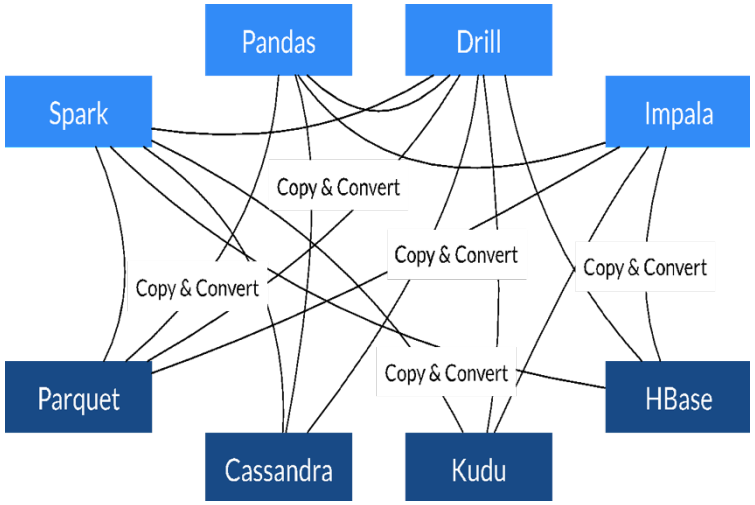


Apache Arrow

http://arrow.apache.org

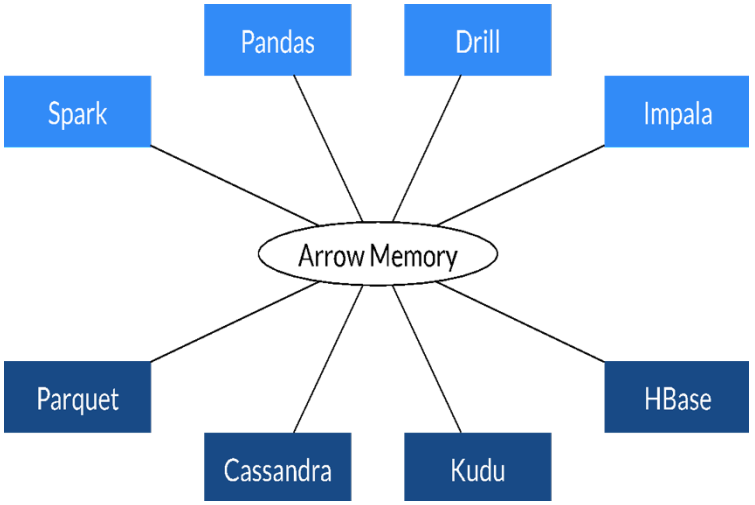
High Performance Sharing & Interchange

Today



- Each system has its own internal memory format
- 70-80% CPU wasted on serialization and deserialization
- Similar functionality implemented in multiple projects

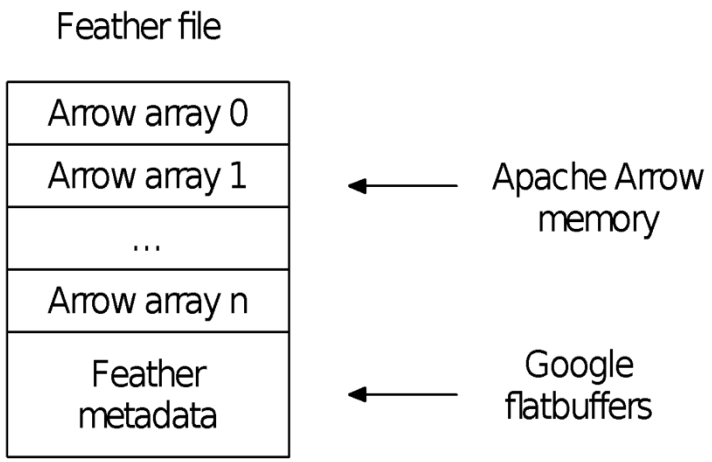
With Arrow



- All systems utilize the same memory format
- No overhead for cross-system communication
- Projects can share functionality (eg, Parquet-to-Arrow reader)

Feather File Format for Python and R

- Problem: fast, language-agnostic binary data frame file format
- By Wes McKinney (Python) and Hadley Wickham (R)
- Read speeds close to disk IO performance
- Leverages Apache Arrow



Thank you

@wesmckinn

http://wesmckinney.com

pandas sprint on Monday!