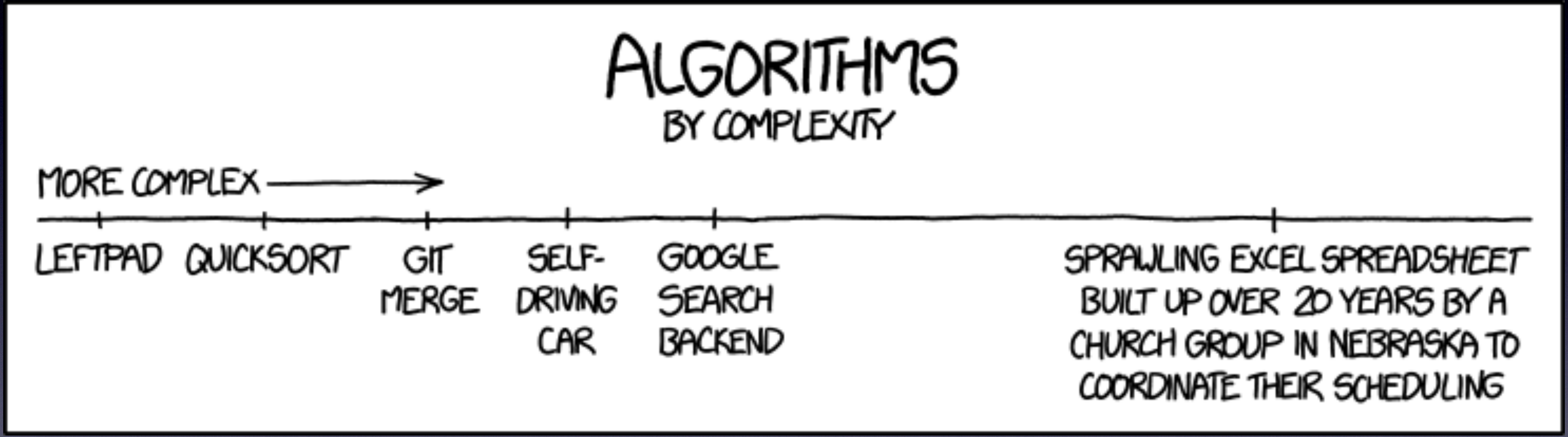
From Python data stack to PySpark

Brendan Herger, hergertarian.com

Slides: https://goo.gl/Waz4kM


ALGORITHMS
BY COMPLEXITY
MORE COMPLEX
LEFTPAD
QUICKSORT
GIT MERGE
SELF-DRIVING CAR
GOOGLE SEARCH BACKEND
SPRAWLING EXCEL SPREADSHEET BUILT UP OVER 20 YEARS BY A CHURCH GROUP IN NEBRASKA TO COORDINATE THEIR SCHEDULING

https://imgs.xkcd.com/comics/algorithms.png

Intro

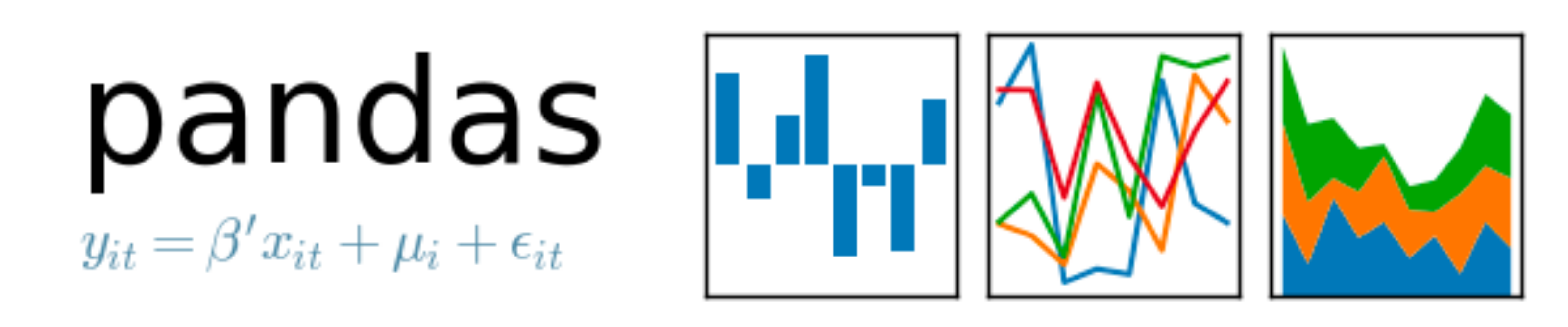
Spark & PySpark

PySpark Components

Pro tips

Recap

Intro


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

scikit
learn

pandas

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

- Data connections
- Data engineering

scikit learn

- Data engineering
- Machine learning

Opportunities

- Scale
- Speed
- Integration
- Development speed

Spark & PySpark

Easy things should be easy,

and hard things should be possible.

– Larry Wall,
Father of Perl programming language

Spark

Spark is an open-source distributed computing platform. It is:

- **Fast:** Spark is memory based, which is 10-100 times faster than Hadoop (which is disk based)
- **Scalable:** Can handle arbitrarily large data sets (GB to TB)
- **Usable:** Allows users to rapidly write R+D or production jobs

Spark use cases

- **Fast:** If other pipelines are too slow or not parallelizable
- **Scalable:** If our dataset is too large for RAM on a single machine
- **Usable:** If other pipelines (such as map reduce) require reinventing the wheel

PySpark

PySpark is an Python wrapper for Spark. It is usually about one release behind core Spark.

PySpark Components

PySpark Components

- **DataFrames:** Tabular data store, based partially on Pandas DataFrames
- **Spark ML: Machine learning library, modeled after SKLearn**
- **Spark SQL:** SQL interface, for ease of use and contributors who don't know how to code
- **GraphX:** Graph database & algorithms

PyData stack similarities

- **DataFrames:** Similar column operations
- **Spark ML:** Same algorithms, same `.fit` and `.predict`
- **Spark SQL:** Vanilla SQL

Pro tips

Pro tips

- **Deployment:** Setting up a Spark cluster takes work. Let someone else do it by using DataBricks, or AWS's EMR

- **Learning curve:** It's like learning a related language. Set aside time to learn the similarities and differences

- **Re-invest, don't re-implement:** Rather than 1-for-1 porting your existing code, write Spark code the 'Spark way'

Recap

Agenda

- Intro
- Spark & PySpark
- PySpark Components
- Pro tips
- Recap

Takeaways

- **Opportunity:** Python data stack is limited to a single thread, on a single machine

- **Scale:** PySpark can handle arbitrarily large data sets, with as many machines as you've got

- **Similarity:** PySpark is designed after the SKLearn and Pandas workflow

- **Re-invest, don't re-implement:** Rather than 1-for-1 porting your existing code, write Spark code the 'Spark way'

Thanks!

Brendan Herger, hergertarian.com
Slides: https://goo.gl/Waz4kM