# Data Science Generalist

so much to know

July 23, 2020

Rory Hartong-Redden
roryhr@gmail.com

https://github.com/roryhr/data_science_generalist

# Agenda

- What is a data scientist?
- Pandas + Stats
- Python

# What is a data scientist?

- A data scientist solves with data
  - Recommendation algorithms, A/B testing, clustering, anomaly detection
  - Data exploration and research
  - Results used by software
- Python, Stats, SQL, Spark

- An analyst solves problems with data
- Tableau, Excel, SQL, Python
- Results used by people

# My path into data science

A series of unfortunate events

- BS Physics and Mechanical Engineering, UC Santa Barbara (2006-2010)
  - 4 quarters of summer work
  - Funded by my parents
- PhD Physics, Northwestern (2010-2012, dropped out)
  - Funded by research and teaching
- MS Mechanical Engineering, UC Santa Barbara (2012-2014)
  - Funded by research and teaching
- startup.ml, San Francisco (2015)
  - Funded by my grandparents

# Statistics

```
il_df = pd.read_csv('../data/All Data by State/Illinois/PPP Data up to 150k - IL.csv')
```

```
il_df.describe()
```

| | LoanAmount | Zip | NAICSCode | JobsRetained |
|---|---|---|---|---|
| **count** | 174745.000000 | 174738.00000 | 167686.000000 | 165186.000000 |
| **mean** | 31679.864620 | 60850.03036 | 537892.053332 | 4.854001 |
| **std** | 32616.795136 | 841.79601 | 197290.458456 | 10.158962 |
| **min** | 1.000000 | 29488.00000 | 111110.000000 | 0.000000 |
| **25%** | 9000.000000 | 60176.00000 | 447110.000000 | 1.000000 |
| **50%** | 20080.000000 | 60605.00000 | 541211.000000 | 2.000000 |
| **75%** | 42220.000000 | 61265.00000 | 624410.000000 | 6.000000 |
| **max** | 149999.000000 | 83108.00000 | 999990.000000 | 500.000000 |

---

- What does this tell us?

# Plots I - Gravitational Waves

https://physics.aps.org/featured-article-pdf/10.1103/PhysRevLett.116.061102

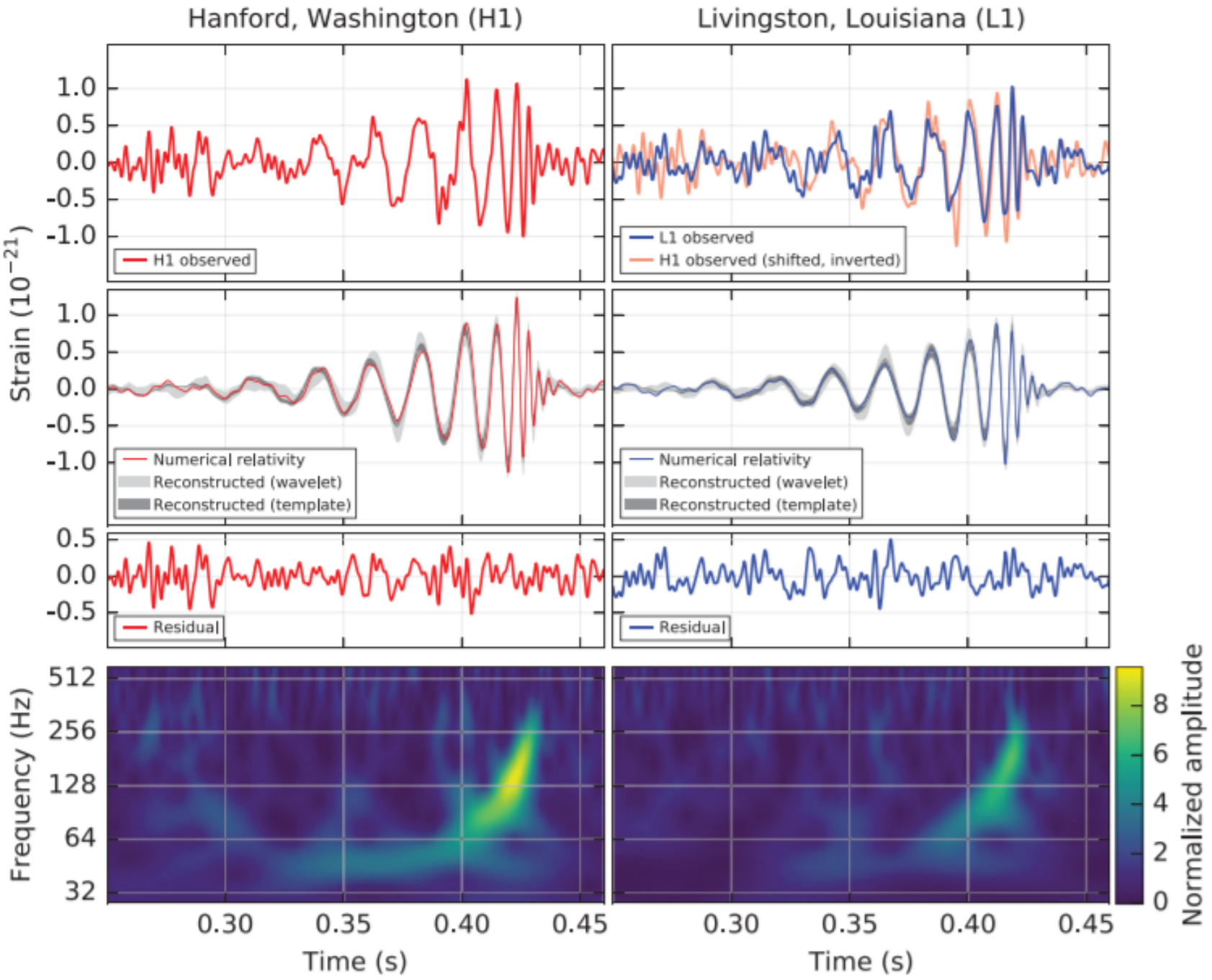
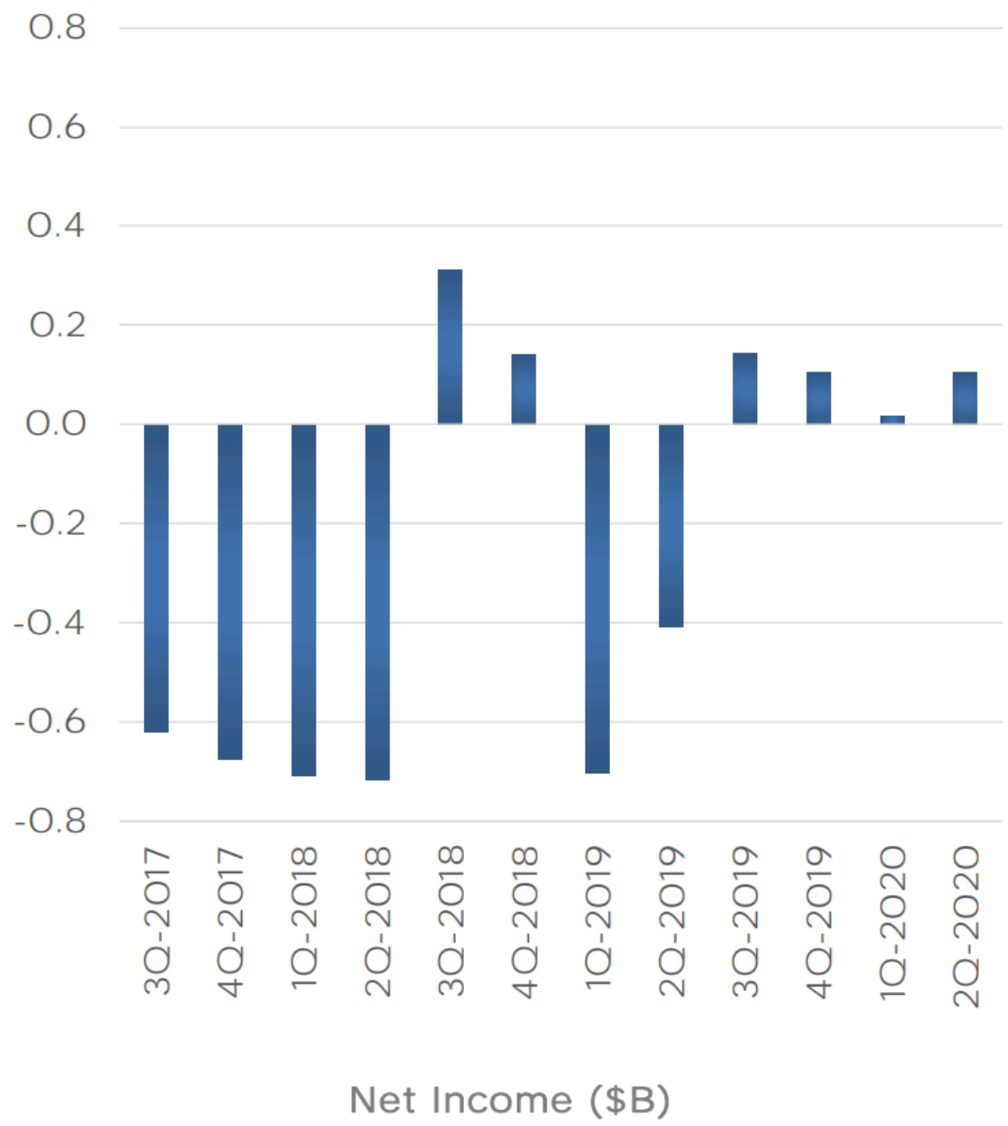



FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors' most sensitive frequency band, and band-reject filters to remove the strong instrumental spectral lines seen in the Fig. 3 spectra. *Top row, left:* H1 strain. *Top row, right:* L1 strain. GW150914 arrived first at L1 and $6.9^{+0.5}_{-0.4}$ ms later at H1; for a visual comparison, the H1 data are also shown, shifted in time by this amount and inverted (to account for the detectors' relative orientations). *Second row:* Gravitational-wave strain projected onto each detector in the 35–350 Hz band. Solid lines show a numerical relativity waveform for a system with parameters consistent with those

# Plots II - Tell a story

- Simple chart from Tesla Q2 earnings update
- Note label with units
- Imagine you're Elon, explain what's going on



# Navigating open source

- Keeping up to date is hard
- Tech Talks and Twitter have a bias for novelty
- So much of what I learned was from osmosis
- Boring and old is good
- Spend time learning tools and paradigms that'll stick around
  - Unix philosophy: pipes, one thing well
  - Web apps with Django, MVC

| Tech | Year | Comments |
|---|---|---|
| SQL | 1974 | The language for data |
| Make | 1976 | Build recipes |
| vi | 1976 | Still the best text editor |
| ssh | 1976 | Get in there |
| Bash | 1989 | Still the best shell |
| Python | 1990 | Programming for humans |
| HTTP | 1991 | Deliver that HTML |
| HTML | 1993 | Websites |
| PostgreSQL | 1996 | Excellent database |
| curl | 1997 | download stuff |
| matplotlib | 2003 | plots |
| git | 2005 | version control |

# Software development

- Git
  - Pull request workflow on Github
- Packaging
- Dependencies
- Testing
- Code formatting
- Continuous Integration (CI)
- Web development
  - HTTP actions
  - Model View Controller
- Security
  - SQL injection!

# Questions?

- @rory_h_r on Twitter