Data science at ReverbNation

Ralph Haygood and Kevin Hofmaenner

June 29, 2016

What is "data science"?

A superpower that will supercharge your business!

What is "data science"?

~~A superpower that will supercharge your business!~~

A buzzword that means different things to different people.

What is "data science" *at ReverbNation*?

A superpower that will supercharge your business!

A buzzword that means different things to different people.

Computer programming and applied statistics that turn data into metrics and models to guide decision-making.

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician." *(apocryphal)*

Which data?

- Production database (MySQL slave).

- Mixpanel.

- Google BigQuery.

- Google Analytics.

- Etc.

Which metrics?

- Counts — e.g., verified artists active within the past year and eligible for a trial of Basic or Premium Plan.

- Rates — e.g., sign-ups, trial starts, trial conversions.

- Aggregates — e.g., lifetime values, correlation coefficients.

- Differences — e.g., now vs. then, this split-test group vs. that one, US vs. UK, rock vs. hip hop.

"Data scientists mostly just do arithmetic."
(Noah Lorang, data scientist at Basecamp)

Which models?

- Nonparametric "models", tests, and methods — e.g., histograms, Mann-Kendall, bootstrapping.

- Lines and curves — e.g., linear regression, exponential smoothing, kernel density estimation.

- Machine learning — e.g., random forests, gradient boosted regression trees, AdaBoost.

Which decisions?

- Does something need fixing? (Is it broken?)

- Should we proceed with membership?

- How much should we offer for ▬▬▬▬▬?

- Which opportunities should we promote to a given artist?

- Etc.

Step by step

- Define questions.

- Identify relevant data.

- Extract and filter data.

- Do appropriate analyses.

- Explain results.

"Here's a breakdown of how I spend my time:

- 1/3 talking with others and figuring out how we can use our data to solve problems

- 1/3 up to my elbows in uggo data cleaning and prepping to solve a problem

- 1/3 hunting down data that's logged in some strange way

- .00000001% actual training of models

This is why Kaggle is bullshit."
(John Foreman, data scientist at MailChimp)

What we talk about when we talk about filtering

- Over 4600 page objects are testers.

- Some subscriptions supposedly ended or were paid before they started, were paid after they ended, etc.

- Some Promote It campaign groups contain no campaigns, point to campaigns they don't contain, etc.

- Ad nauseam.

Tools

Mostly Python:

- Python is a pleasant, modern language.

- NumPy accelerates vector and matrix operations.

- pandas and scikit-learn implement statistical methods.

- Jupyter notebooks facilitate exploration.

[Jupyter notebook demo]

Step by step

- Define questions.

- Identify relevant data.

- Extract and filter data.

- Do appropriate analyses.

- Explain results.

Some data are relevant again and again.

So extract and filter them once and for all.

ETL processing

- Tasks run nightly "in the cloud".

- Reads production database, Mixpanel, etc.

- Filters and consolidates recurrently relevant data.

- Writes data warehouse (PostgreSQL) and HDF5 files.

- Luigi coordinates runs to satisfy dependencies.

Step by step

- Define questions.

- Identify relevant data.

- Extract and filter data.

- Do appropriate analyses.

- Explain results.

Some analyses are appropriate again and again.

So do them daily and explain results once and for all.

Chartio

- Is an online, WYSIWYG tool for making online, interactive dashboards.

- Queries the data warehouse and turns results into charts and tables.

- Provides dropdowns, text fields, etc. to modulate queries.

- Reruns queries and updates dashboards automatically.

[Chartio dashboard demo]

Thanks

- David.

- DevOps, especially Justin.

- Chris, Mike D., Steve S., Wes, and product managers.

- All of you.