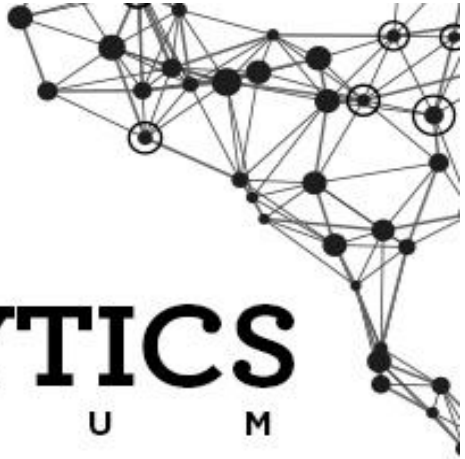

BIG DATA ANALYTICS

S Y M P O S I U M

UCF

UCF

Data Science Ops

Robert Norberg

Key Points

- What is Data Science Ops?
- How do I implement Data Science Ops best practices?
- Why is Data Science Ops important?

Introduction

Introduction

So you're ready to graduate and you know how to:

- manipulate and prepare data
- choose and apply the correct hypothesis test/confidence interval
- train a good regression/classification model
- design a good experiment
- build a good visualization
- etc

And a company offers you a job...

Introduction

Your new #1 priority is to PROVIDE VALUE TO YOUR COMPANY.

You will wield your skills to report, analyze, and predict for your company.

You will communicate your work to the rest of your company via reports, visualizations, and models.

What is Data Science Ops?

What is Data Science Ops?

All of the things that go into supporting data science products (reports, visualizations, and models)

- hardware
- software
- deployment
- maintenance

In many organizations, the data science team is responsible for its own software, deployment, and maintenance.

Hardware

- Computers/servers used by data scientists
- Servers used to deploy data science products
  - E.g dashboard server
- Typically supported by the organization's IT infrastructure

Software

- Software used by data scientists
  - SAS, R, Python, Tableau, etc
  - Packages/libraries
  - Database(s)
- Software used to deploy data science products
  - SAS Web Report Server, Tableau Server, etc
- Code written by data scientists

UCF

Software

- IT probably manages the software used by data scientists
- The data science team may be responsible for the software used to deploy their products
- The data science team is responsible for the code they write

Deployment

- Sharing, scheduling, distributing reports
- Hosting dashboards
- Serving model predictions

Maintenance

- Updating software
- Monitoring model performance
- Retraining models
- Updating reports, visualizations

How do I implement Data Science Ops best practices?

How do I implement Data Science Ops best practices?

- Use a well defined stack
- Use version control for code
- Require code reviews
- Keep all of your (reports/models/visualizations) in one place
  - Version them

A well defined stack

- At most 2 (data science) languages
  - A style guide for each
- A central code repository
- Deployment tools for reports, visualizations, and models

Version control for code

Git is the most popular version control system. It integrates well with R, Python, and SAS.

Require code reviews

Data science products are created with code. No product should leave the data science group without its code being thoroughly reviewed by a second data scientist.

Data Science code reviews do not just focus on code, they also focus on methodology.

Keep all of your reports in one place

- SharePoint, Dropbox, Google Drive, etc
- Do not use email attachments - send links instead
- Minor changes to the link's endpoint are OK
- Major changes should result in an incremented version number
  - My Report v2.0
- When a report is refreshed with new data, the old report should be archived

Keep all of your visualizations in one place

- Serve them from a web server
  - This will support static and interactive visualizations
  - Other parts of the business can embed the visualization in their products
- Do not use email attachments - send links instead

Keep all of your models in one place

- Easiest if model output is served via an API
  - Offer individual and batch scoring
- For every prediction that goes out, an actual outcome must come in
  - These should all be stored in the same place too
- When a model is retrained, do not delete the old model
  - Increment the model's version number instead

Why is Data Science Ops important?

Why is Data Science Ops important?

- Reduce your team's "bus factor"
- Increase collaboration
- Reduce errors
- Spend more time creating and less time maintaining
- Present a professional, predictable interface to the rest of the business

Reduce your team's "bus factor"

Inevitably, people will leave (and join!) your team.
It is easier to transfer responsibilities when:

- Everyone uses the same suite of tools
- Code is properly reviewed and versioned
- Software packages are properly versioned

Increase collaboration

Collaboration among data scientists is valuable to the data scientists and to the final product.
It is easier for data scientists to collaborate when:

- Everyone uses the same suite of tools
- Code is properly versioned
- Code review is required

Reduce errors

- A well defined stack will have fewer technical issues
- Code reviews catch human errors
- Standardized model deployment makes monitoring models easy
  - Easy model monitoring will help you catch a bad model in production before too much damage is done

Spend more time creating and less time maintaining

Standardized deployment and maintenance processes make supporting things easy.

The less time you spend supporting things, the more time you can spend creating things.

Present a neat interface to the rest of the business

Standardized deployment isn't just neater and easier for the data science team.
It makes interacting with the data science team easier for the rest of the organization.

Recap

- What is Data Science Ops?
  - All of the “other stuff” that isn’t data science
- How do I implement Data Science Ops best practices?
  - Use a well defined stack
  - Version code, models, reports, and visualizations
  - Require code reviews
  - Keep things in one place
- Why is Data Science Ops important?
  - Increases quality and quantity of work possible
  - Makes work more rewarding for the data scientists

UCF

Questions or comments?

robert c norberg @gmail.com
r-norberg.blogspot.com

UCF