Doing Data:
Data Prep for Machine Learning

Ellen Friedman, PhD
18 June 2019
Berlin Buzzwords #bbuzz

Contact Information

Ellen Friedman, PhD

Committer Apache Drill & Apache Mahout projects

O'Reilly author

Email ellenf@apache.org

Twitter @Ellen_Friedman

Today: #bbuzz

Bloom County cartoon by Berkeley Breathed
https://www.berkeleybreathed.com/

Bloom County cartoon by Berkeley Breathed
https://www.berkeleybreathed.com/

Sometimes, the most powerful solutions are very basic.

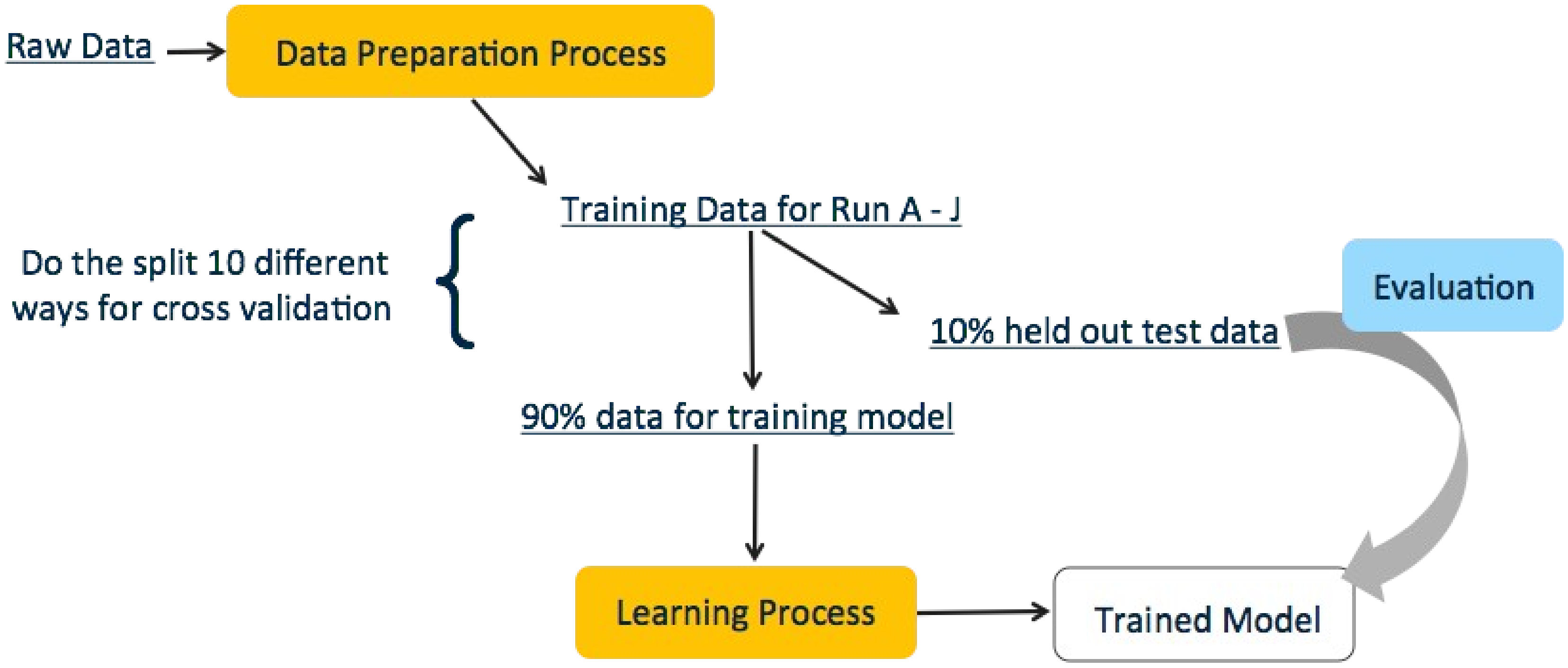
Bloom County cartoon by Berkeley Breathed
https://www.berkeleybreathed.com/

Sometimes, the most powerful solutions are very basic.

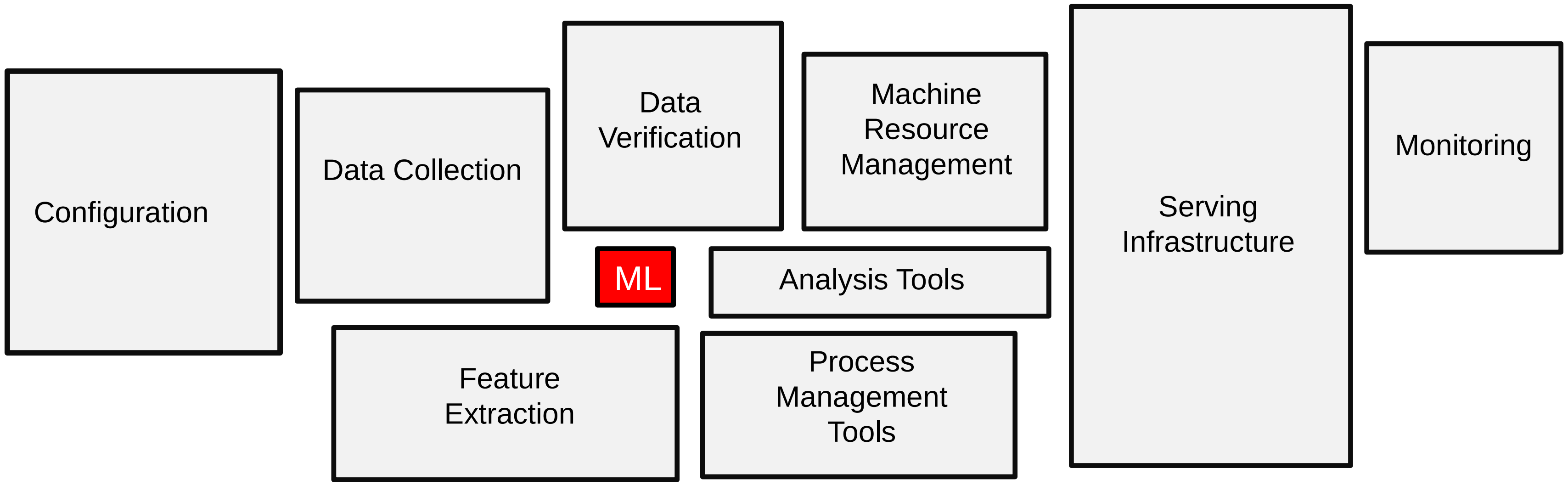
That doesn't necessarily make them easy.

What makes machine learning work?

The data


Raw Data
Data Preparation Process
Training Data for Run A - J
Do the split 10 different ways for cross validation
10% held out test data
Evaluation
90% data for training model
Learning Process
Trained Model

Find the ML code



Only a small part of ML systems is the learning code.
The rest is vast infrastructure of data collection and processing.

*Figure based on **“Hidden Technical Debt in Machine Learning Systems”** by Sculley et al. (Google, Inc)*
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Things That Matter in Data Preparation

I. What’s in your data? (really?)

II. How do you know what features to build?

III. How do you know what you did?

Retroactive Value in Data

Labs in Canada froze blood samples for years in case the samples might contain valuable information

Correlated with outcomes for the donor patients

Retroactive Value in Data

Labs in Canada froze blood samples for years in case the samples might contain valuable information

.

- ✓ They did

Correlated with outcomes for the donor patients

Retroactive Value in Data

Labs in Canada froze blood samples for years in case the samples might contain valuable information

.

- ✓ They did

Modern genetic techniques revealed key disease data

- ✓ Correlated with outcomes for the donor patients

Retroactive Value in Data

Labs in Canada froze blood samples for years in case the samples might contain valuable information

.

- ✓ They did

Modern genetic techniques revealed key disease data

- ✓ Correlated with outcomes for the donor patients

The data was preserved *before* the analysis was even begun.

Retroactive Value in Data

Labs in Canada froze blood samples for years in case the samples might contain valuable information

- ✓ They did

Modern genetic techniques revealed key disease data

- ✓ Correlated with outcomes for the donor patients

frozen

The data was preserved *before* the analysis was even begun

Things That Matter in Data Preparation

I. What's in your data? (really?)

II. How do you know what features to build?

III. How do you know what you did?

Thanks to these data scientists for their stories

Joe Blue
Director Global Data Science,
MapR

Ted Dunning
Chief Technical Officer,
MapR

A loyal fan of Berlin Buzzwords

Ted Dunning, Berlin 2018

Machine Learning in the Real World

Kaggle

vs

https://visibleearth.nasa.gov/view.php?id=56229

I. What’s in Your Data (Really)?

Verify!

- Examine data
- Ask questions (domain knowledge matters)
- Make sure what you say is what they hear

Explore!

- Find out what you’ve got
- Sometimes data exploration gives you the solution
- Visual inspection & draw pictures
- Example tool: Apache Drill

Verify!

Example:

Fraud detection model trained with data from column named “fraud”

Verify!

Example:

Fraud detection model trained with data from column named “fraud”

Oops.

It was the fraud analyst ID, not a flag for known fraud

Verify!

Example:

Fraud detection model trained with data from column named “fraud”

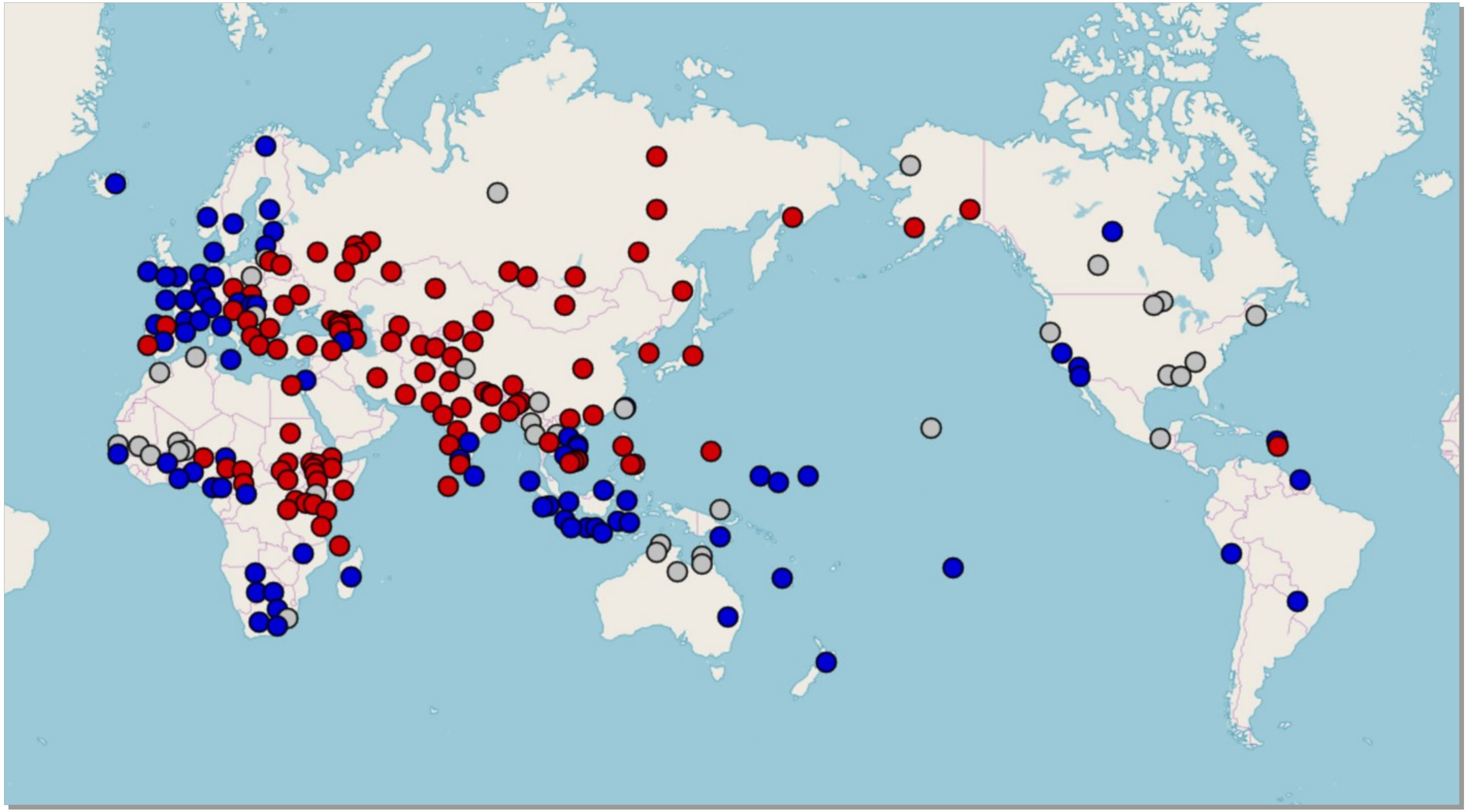
Oops.

It was the fraud analyst ID, not a flag for known fraud

Clear communication is essential.

Fun example: Can you spot the pattern tea vs chai?

- tea
- chai



https://wals.info/chapter/138

Explore!

Example:

Big European service provider had complaints of poor response time.
But average response time in the reports was always fine…?!

Hard problem! Expect to use sophisticated ML to find the problem.

Explore!

Example:

Big European service provider had complaints of poor response time.
But average response time in the reports was always fine…?!

Hard problem! Expect to use sophisticated ML to find the problem.

1st step: explore data using Apache Drill.

Explore!

Example:

Big European service provider had complaints of poor response time.

But average response time in the reports was always fine…?!

Hard problem! Expect to use sophisticated ML to find the problem.

1st step: explore data using Apache Drill.

Immediately discover dropped data. Easy solution: ML not needed.

II. How do you know what features to build?

Features are built, not just chosen

There’s no “right” answer: trial and error (success) to find winners

What makes good features?

Performance

Feasibility

Interpretability

Think through behaviors

Build Features for Fraud Detection

What would you do if you were a fraudster?

Behaviors That Point to Fraud

Fraudster has stolen debit card, but doesn’t know pin number

Tries to use it as credit card with signature: easier to fake

Behaviors That Point to Fraud

Fraudster has stolen debit card, but doesn't know pin number

Tries to use it as credit card with signature: easier to fake

9990 →

Leaves clues you can discover: Make a feature from this change

More Behaviors That Point to Fraud

- Domain expert says “Fraudsters often do a probe transaction at a gas station just before making their big fraud transaction(s).”

- How do you build a feature to detect probe behaviors?

- Risk tables can be constructed:
  - to find a probe event or
  - to find the main fraud event

We Build a "Risk Table"

When they say "gas station", we think "merchant type"

When they say "just before", we think of several possible time periods

Take many transactions grouped by consumer, ordered by time

- For each fraud, count the merchant types in the preceding window of time
- For lots of non-frauds, count the merchant types in the preceding window

A risk table has the (log of the) ratio of the fraud counts to the non-fraud counts for each merchant type, for each window size

Building a Risk Table

Type	*Before frauds (100k samples)*	*Before non-fraud (1M samples)*	*Log Risk Ratio*
gas station	60227	66639	2.20
tea room	157	10087	-1.86
hotel	1633	24720	-0.41
airline	1035	12389	-0.18
pizza delivery	28765	52838	1.69

A bigger positive value for ratio = more risk

Look at recent events for a particular card

```
2019-06-11T14:19:09Z, grocery, -0.1
2019-06-11T20:36:11Z, books, 0.1
2019-06-12T04:42:14Z, restaurant, 0.05
2019-06-12T09:30:08Z, books, 0.1
2019-06-12T12:07:03Z, gas station, 2.2
```

Type	***Before frauds (100k samples)***	***Before non-fraud (1M samples)***	***Log Risk Ratio***
gas station	60227	66639	2.20
tea room	157	10087	-1.86
hotel	1633	24720	-0.41
airline	1035	12389	-0.18
pizza delivery	28765	52838	1.69

2019-06-11T14:19:09Z, grocery, -0.1
2019-06-11T20:36:11Z, books, 0.1
2019-06-12T04:42:14Z, restaurant, 0.05
2019-06-12T09:30:08Z, books, 0.1
2019-06-12T12:07:03Z, gas station, 2.2

Type	***Before frauds (100k samples)***	***Before non-fraud (1M samples)***	***Log Risk Ratio***
gas station	60227	66639	2.20
tea room	157	10087	-1.86
hotel	1633	24720	-0.41
airline	1035	12389	-0.18
pizza delivery	28765	52838	1.69

2019-06-11T14:19:09Z, grocery, -0.1
2019-06-11T20:36:11Z, books, 0.1
2019-06-12T04:42:14Z, restaurant, 0.05
2019-06-12T09:30:08Z, books, 0.1
2019-06-12T12:07:03Z, gas station, 2.2

Type	***Before frauds (100k samples)***	***Before non-fraud (1M samples)***	***Log Risk Ratio***
gas station	60227	66639	2.20
tea room	157	10087	-1.86
hotel	1633	24720	-0.41
airline	1035	12389	-0.18
pizza delivery	28765	52838	1.69

2019-06-12T23:47:22Z, ***2.35***

Data Augmentation

Augment data: add external information

Example:

- You have merchant ID
- Look up store location
- Give model location as a feature

Data Transformation

Simple data transformation can be powerful

Domain knowledge helps you know what to do

Example:

- Data is value for amount of €
- Take log of value because % gives a more meaningful feature

  10 € → 12 €   is very different change than 100 € → 102 €

Velocity as a Feature

Commonly used

How many ways can you describe velocity?

Domain knowledge helps you know what to do

Examples:

- Geo-distance / time
- # events / time
- € / time

III. How do you know what you did?

Code:

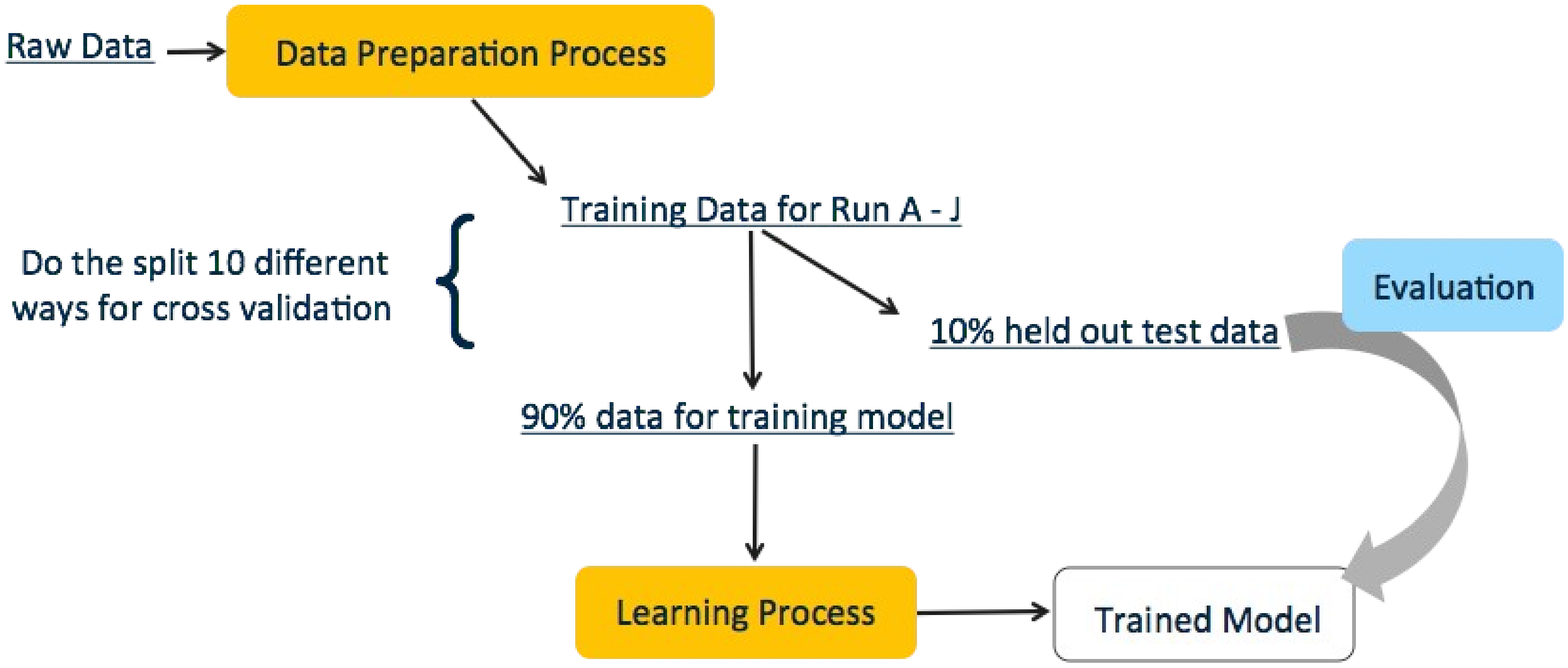
Document the reasoning behind code
Version control for code

Data:

Document how training data was prepared

- Which features? Why?
- How were they built?

Version control for the training data



Raw Data
Data Preparation Process
Training Data for Run A - J
Do the split 10 different ways for cross validation
10% held out test data
Evaluation
90% data for training model
Learning Process
Trained Model

What is the role of data in building an ML model?

Blog post:

“Computer Science vs Data Science” by Ted Dunning

https://mapr.com/blog/data-science-vs-computer-science/

Data makes the model

Different training data, different model

Data makes the model

Different training data, different model

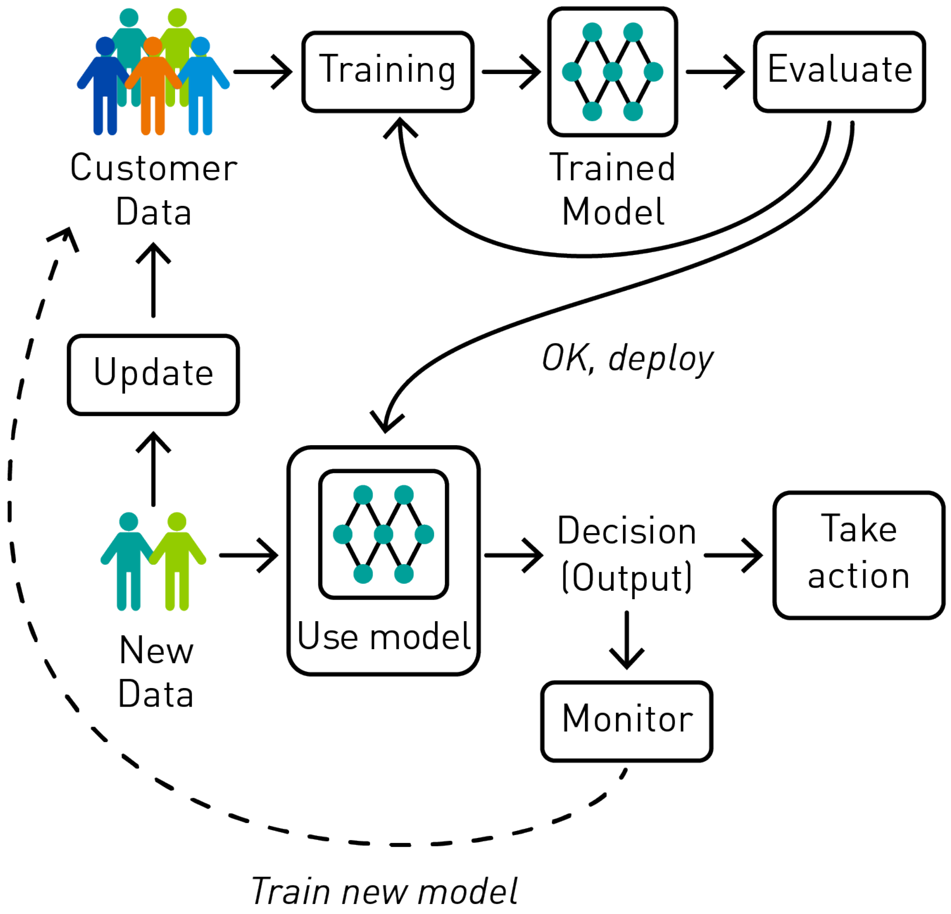
Is it OK if only one person can compile my production source code?

Based on Bloom County cartoon by Berkeley Breathed
https://www.berkeleybreathed.com/

The same applies for data used to build a machine learning model

Notes to Your Future Self

Machine Learning is Iterative Process



- Held-out training data is used for evaluation
- Usually have much more data in training than in production
- Don’t fall for the myth of unitary model: Lots of models, lots of trials

Notebooks are Excellent

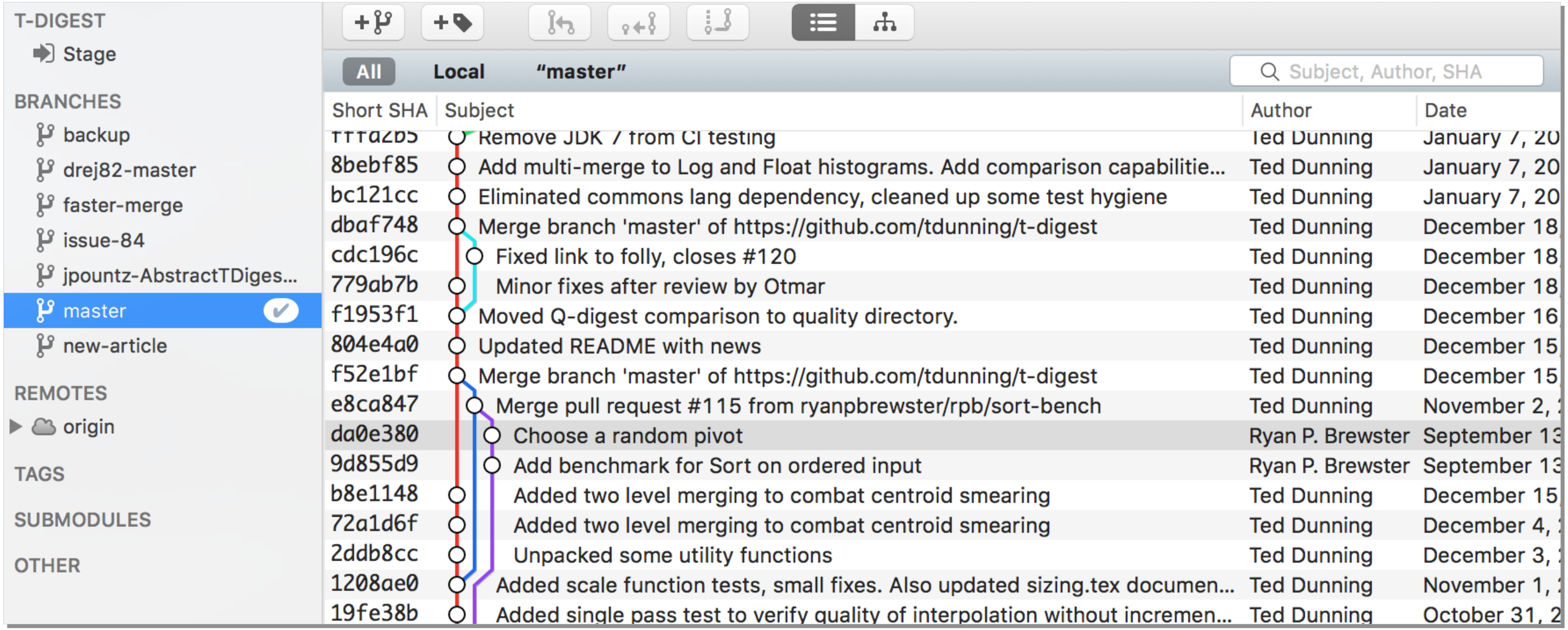
Notes to self: A good way to remember what you've done

A good way to communicate as well

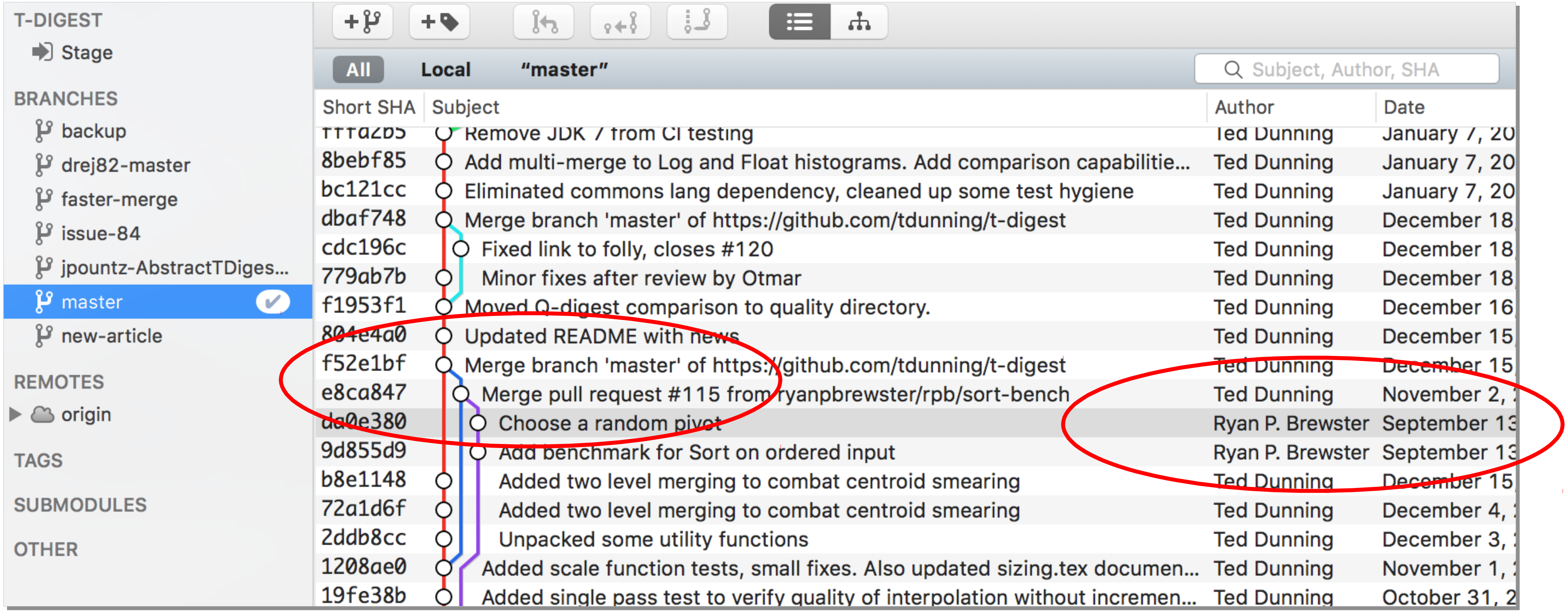
https://jupyter.org/

https://zeppelin.apache.org/

Code Versioning Via Git



Code Versioning Via Git



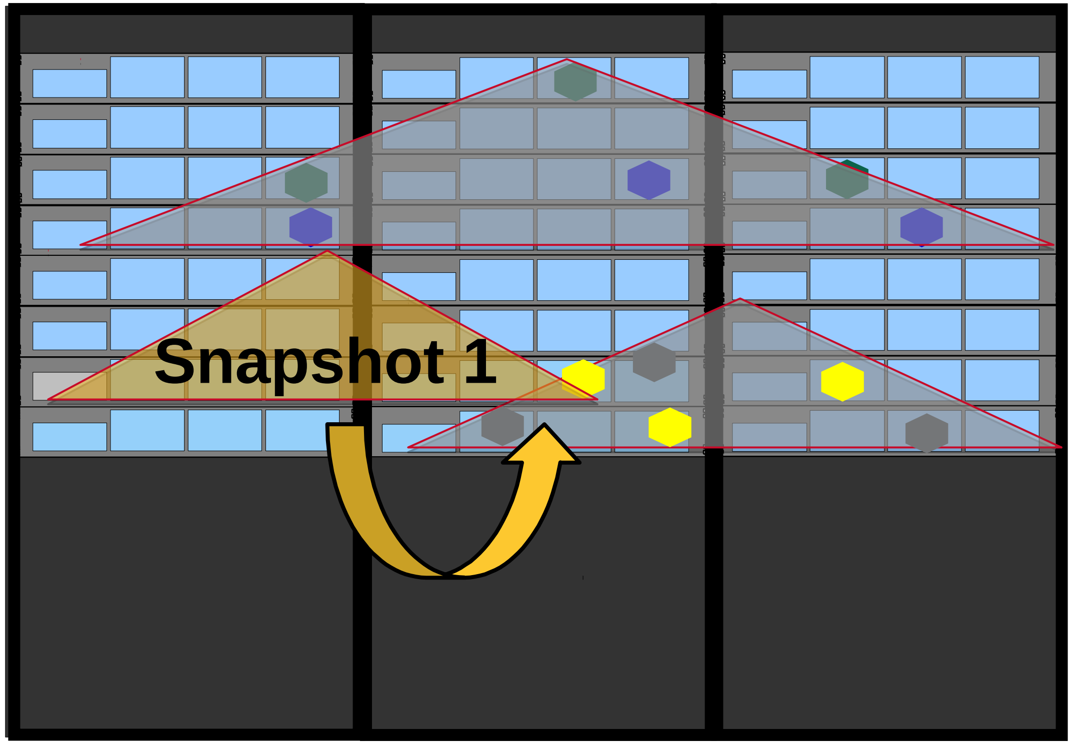
Easy Data Version Control with Snapshots

Snapshots based on MapR volumes

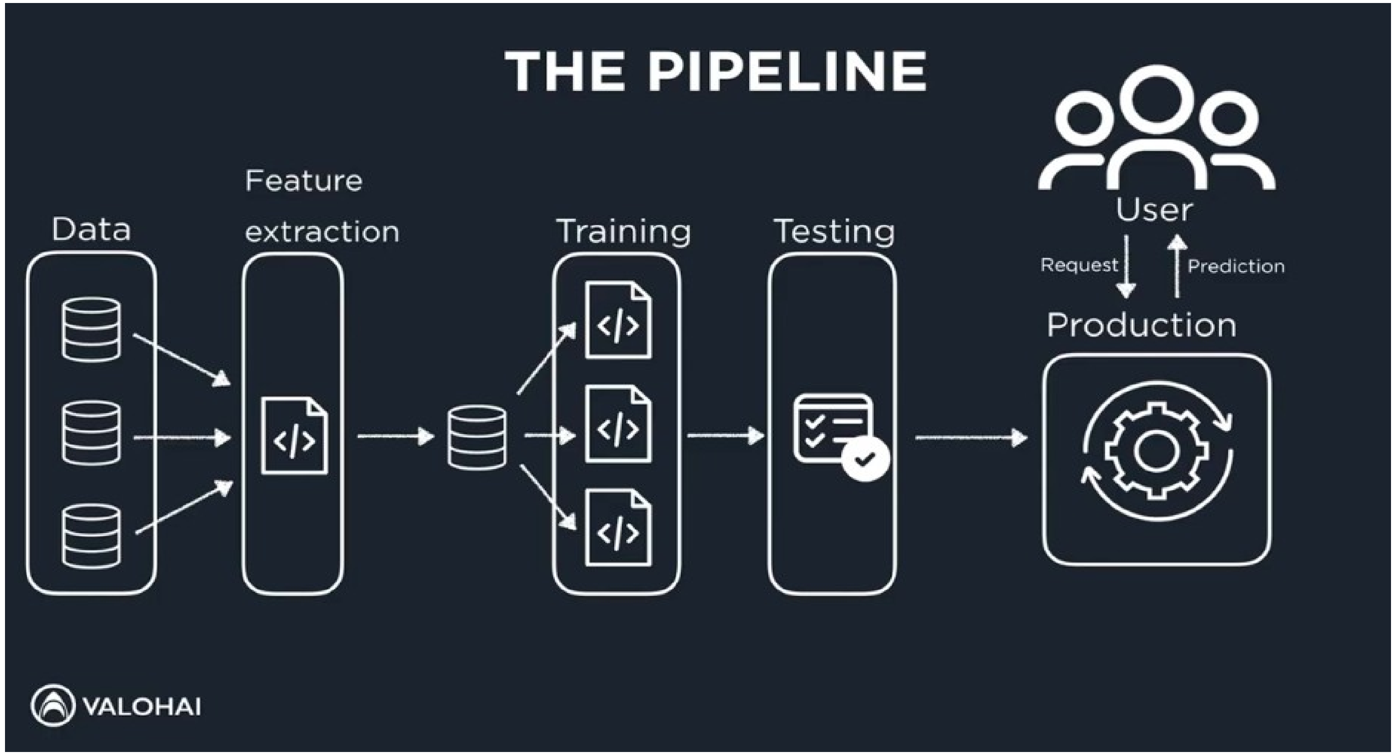
True point-in-time version of data

Less expensive than copying

Fully distributed across cluster



Valohai Uses Snapshots for Their ML Pipeline Service



Eero Laaksonen & Juha Kilii from Finnish company Valohai demonstrate how they do version control using snapshots in this webinar with Ian Downard (MapR):

https://mapr.com/webinars/a-guide-to-version-control-for-machine-learning/

This is What You Track

- <u>For Training Data:</u>

- Pathname of raw data snapshot
- Git reference for data preparation process (feature extraction)

- <u>For Code (Delivered Model):</u>

- The model
- Pathname of training data snapshot
- Git reference for learning script
  - Includes random number seed
  - Includes knob settings for learning process
  - The learning code

Data Unit Testing

Does what you're doing now match what you did before?

Test that: build a way to see if there are changes that matter.

- Test outputs: Maybe what changed doesn't matter
- Test inputs: Another good approach (see Google paper)

Data Validation Article from Google Research

DATA VALIDATION FOR MACHINE LEARNING

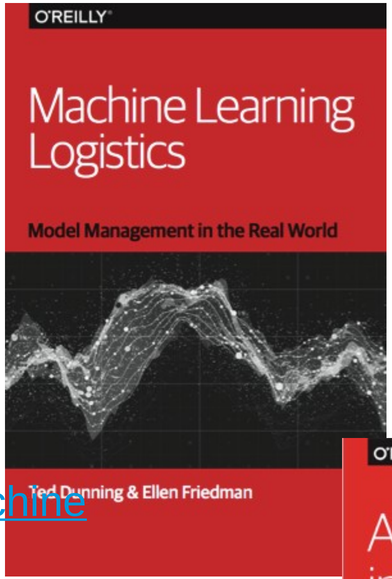
Eric Breck [1] Neoklis Polyzotis [1] Sudip Roy [1] Steven Euijong Whang [2] Martin Zinkevich [1]

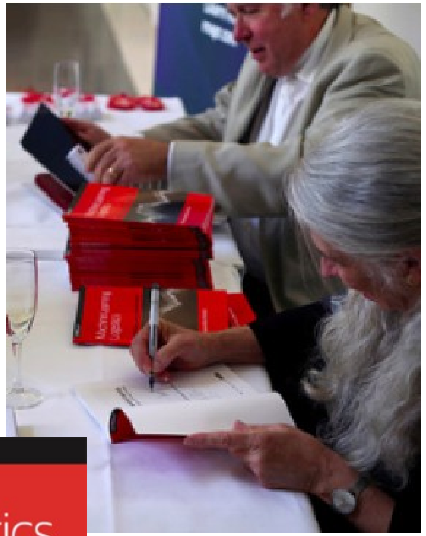
ABSTRACT

Machine learning is a powerful tool for gleaning knowledge from massive amounts of data. While a great deal of machine learning research has focused on improving the accuracy and efficiency of training and inference algorithms, there is less attention in the equally important problem of monitoring the quality of data fed to machine learning. The importance of this problem is hard to dispute: errors in the input data can nullify any benefits on speed and accuracy for training and inference. This argument points to a data-centric approach to machine learning that treats training and serving data as an important production asset, on par with the algorithm and infrastructure used for learning.

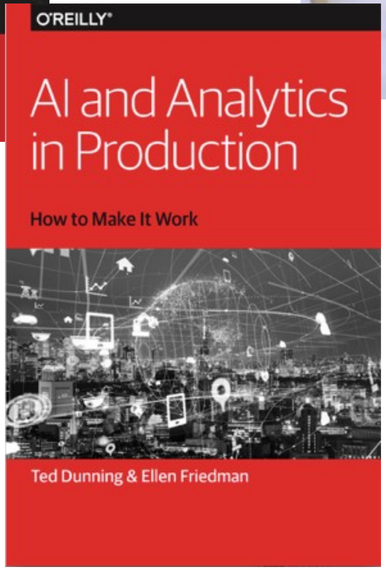
https://www.sysml.cc/doc/2019/167.pdf

Free eBooks courtesy of MapR





https://mapr.com/ebook/machine-learning-logistics/



https://mapr.com/ebook/ai-and-analytics-in-production/

Please support women in tech – help build girls’ dreams of what they can accomplish

#womenintech #datawomen

Thank you !

Contact Information

Ellen Friedman, PhD

Committer Apache Drill & Apache Mahout projects

O'Reilly author

Email ellenf@apache.org

Twitter @Ellen_Friedman

Today: #bbuzz

