



Beyond the Algorithm: What Makes Machine Learning Work?

Ellen Friedman, PhD

11 June 2018

Berlin Buzzwords #bbuzz

Contact Information

Ellen Friedman, PhD

Principal Technologist, MapR Technologies

Committer Apache Drill & Apache Mahout projects

O'Reilly author

Email efriedman@mapr.com ellenf@apache.org

Twitter @Ellen_Friedman #bbuzz

What makes machine learning
work?



+



=



Data Engineer Ian Downard



Had never tried machine learning, but he caught the bug...

Image Recognition: Which Bird Is This?

Rhode Island Red



Image from Wikipedia & used under Creative Commons https://en.wikipedia.org/wiki/File:Rhode_Island_Red_cock,_cropped.jpg

Buff Orpington



Image from Wikipedia & used under Creative Commons https://upload.wikimedia.org/wikipedia/commons/7/74/Barred_Plymouth_Rock_Rooster_001.jpg

Jay



Image from Wikipedia & used under Creative Commons <https://en.wikipedia.org/wiki/Aphelocoma#/media/File:WesternScrubJay2.jpg>

More to the point...

Chicken



Image from Wikipedia & used under Creative Commons https://en.wikipedia.org/wiki/File:Rhode_Island_Red_cock,_cropped.jpg

Chicken



Image from Wikipedia & used under Creative Commons https://upload.wikimedia.org/wikipedia/commons/7/74/Barred_Plymouth_Rock_Rooster_001.jpg

Not a Chicken



Image from Wikipedia & used under Creative Commons <https://en.wikipedia.org/wiki/Aphelocoma#/media/File:WesternScrubJay2.jpg>

Domain Knowledge Matters

Chicken



Image from Wikipedia & used under Creative Commons https://en.wikipedia.org/wiki/File:Rhode_Island_Red_cock,_cropped.jpg

Chicken



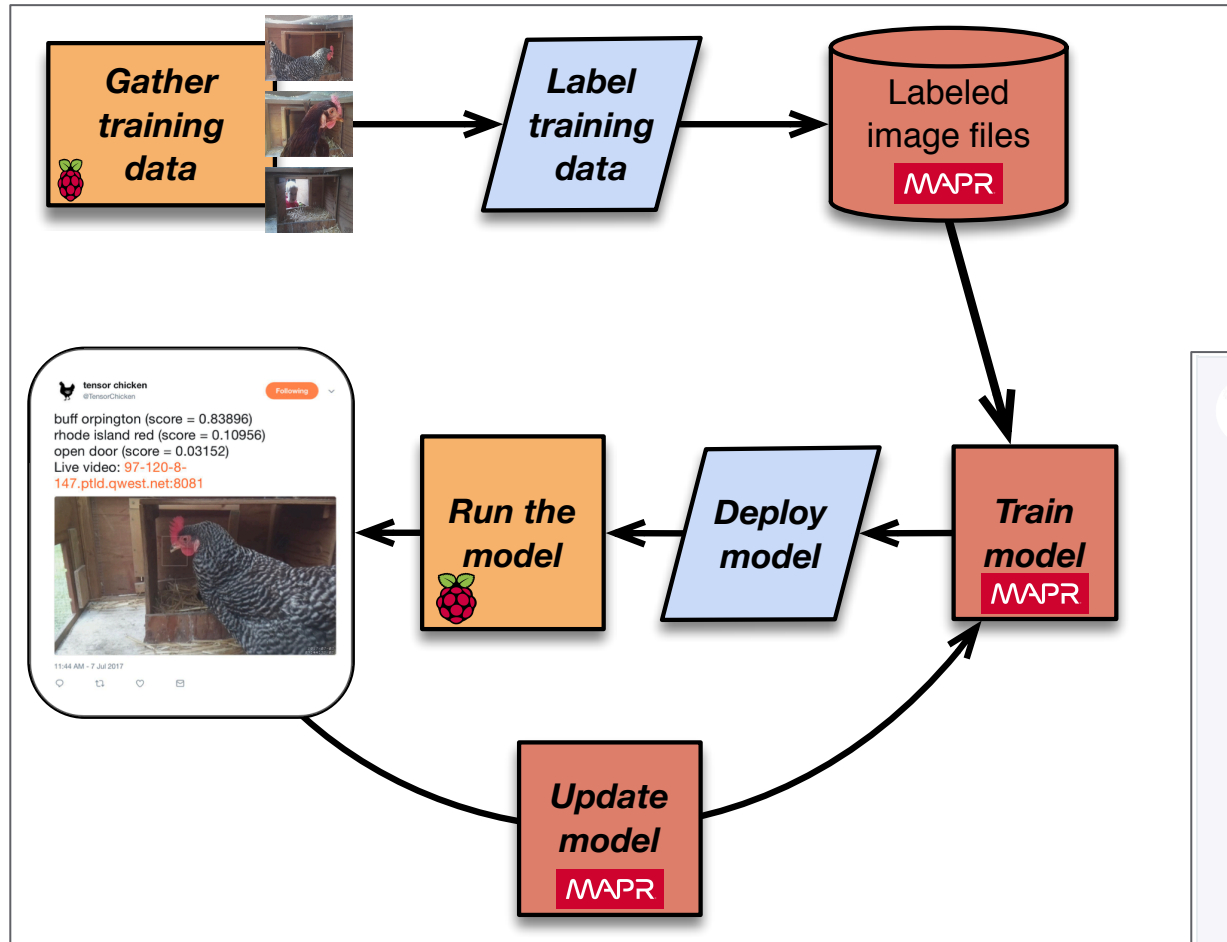
Image from Wikipedia & used under Creative Commons https://upload.wikimedia.org/wikipedia/commons/7/74/Barred_Plymouth_Rock_Rooster_001.jpg

Predator



Image from Wikipedia & used under Creative Commons <https://en.wikipedia.org/wiki/Aphelocoma#/media/File:WesternScrubJay2.jpg>

Tensor Chicken



Deep learning project using Inception v3 model from TensorFlow (see blog + @tensorchicken)



Value from ML: what about SLA's?

How long does it take for image recognition model to classify image?

- ~30 seconds because just running on a Raspberry Pi

How long does it take for Scrub Jay to peck an egg?

- < 30 seconds

- Oops...

Value from ML: what about action?

When image classification indicates a jay in the henhouse, what would you do to chase away the predator?

- Not sure yet. That's a problem...

What makes a difference for impact?

Left: labelled as Buff Orpington.



Image from Wikipedia & used under Creative Commons https://upload.wikimedia.org/wikipedia/commons/7/74/Barred_Plymouth_Rock_Rooster_001.jpg

Right: This what a Buff Orpington really looks like.



Image from Wikipedia used under Creative Commons https://en.wikipedia.org/wiki/Orpington_chicken#/media/File:Coq_orpington_fauve.JPG

But...this error in domain knowledge (wrong name) did not matter for SLAs.

What lessons can we learn from this toy project?

- Image recognition & deep learning are cool
- In some cases, building or training a model is simple
- Domain knowledge matters (really)
- Pay attention to SLAs
- For real business value, have a plan of action in response to machine learning insights (producing a report \neq taking an action)
- Software engineers have a role in machine learning (!)

What about real world
examples?

Domain Knowledge Matters: Video Recommender

- Use clicks as input data: recommender gives poor performance
 - Model is testing the wrong preferences: how well people liked titles
- Use first 30 seconds of viewing as input data: recommender performance is good
 - Model now tests how well people liked the videos, not just the titles

Domain Knowledge Matters: Detecting Security Attacks

Security expert at a bank preserved headers for web site requests

Spot the Important Difference?

```
GET /personal/comparison-table
Host: www.sometarget.com
User-Agent: Mozilla/4.0 (compa
Accept-Encoding: deflate
Accept-Charset: UTF-8
Accept-Language: fr
Cache-Control: no-cache
Pragma: no-cache
Connection: Keep-Alive
```

Attacker request

```
GET /photo.jpg HTTP/1.1
Host: lh4.googleusercontent.
User-Agent: Mozilla/5.0 (Mac
Accept: image/png,image/*;q=
Accept-Language: en-US,en;q=
Accept-Encoding: gzip, defla
Referer: https://www.google.
Connection: keep-alive
If-None-Match: "v9"
Cache-Control: max-age=0
```

Real request

Spot the Important Difference?

```
GET /personal/comparison-table
Host: www.sometarget.com
User-Agent: Mozilla/4.0 (compa
Accept-Encoding: deflate
Accept-Charset: UTF-8
Accept-Language: fr
Cache-Control: no-cache
Pragma: no-cache
Connection: Keep-Alive
```

Attacker request

```
GET /photo.jpg HTTP/1.1
Host: lh4.googleusercontent.
User-Agent: Mozilla/5.0 (Mac
Accept: image/png,image/*;q=
Accept-Language: en-US,en;q=
Accept-Encoding: gzip, defla
Referer: https://www.google.
Connection: keep-alive
If-None-Match: "v9"
Cache-Control: max-age=0
```

Real request

Another Example

```
GET photo.jpg HTTP/1.1
Host: lh4.googleusercontent
User-agent: Mozilla/5.0 (Ma
Accept: image/png,image/*
Accept-language: en-US,en
Accept-encoding: gzip, defl
Referer: https://www.google
Connection: keep-alive
If-none-match: "v9"
Cache-control: max-age=0
```

Real request

```
GET cc/borken.json HTTP/1.1
host: c.qrs.my
user-agent: Mozilla/4.0 (co
accept: application/json, t
accept-language: en-US,en
accept-encoding: gzip, defl
referer: none
connection: keep-alive
if-none-match: "v9"
cache-control: max-age=0
```

Attacker request

Another Example

```
GET photo.jpg HTTP/1.1
Host: lh4.googleusercontent
User-agent: Mozilla/5.0 (Ma
Accept: image/png,image/*
Accept-language: en-US,en
Accept-encoding: gzip, defl
Referer: https://www.google
Connection: keep-alive
If-none-match: "v9"
Cache-control: max-age=0
```

Real request

```
GET cc/borken.json HTTP/1.1
host: c.qrs.my
user-agent: Mozilla/4.0 (co
accept: application/json, t
accept-language: en-US,en
accept-encoding: gzip, defl
referer: none
connection: keep-alive
if-none-match: "v9"
cache-control: max-age=0
```

Attacker request

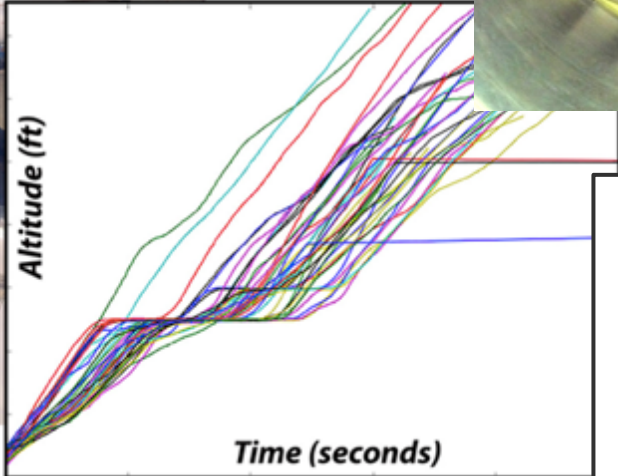
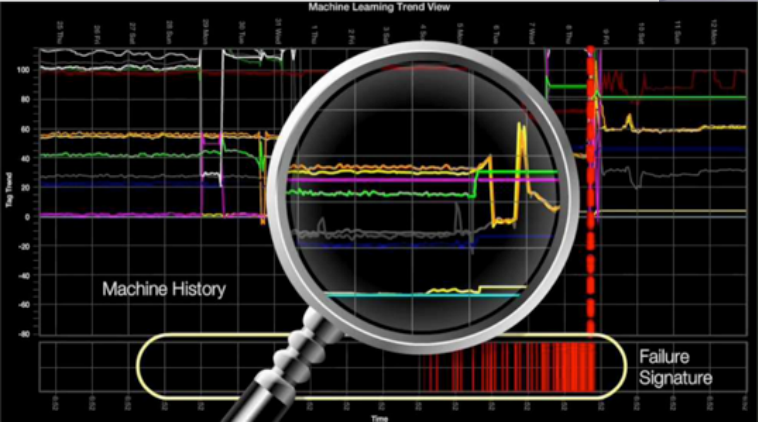
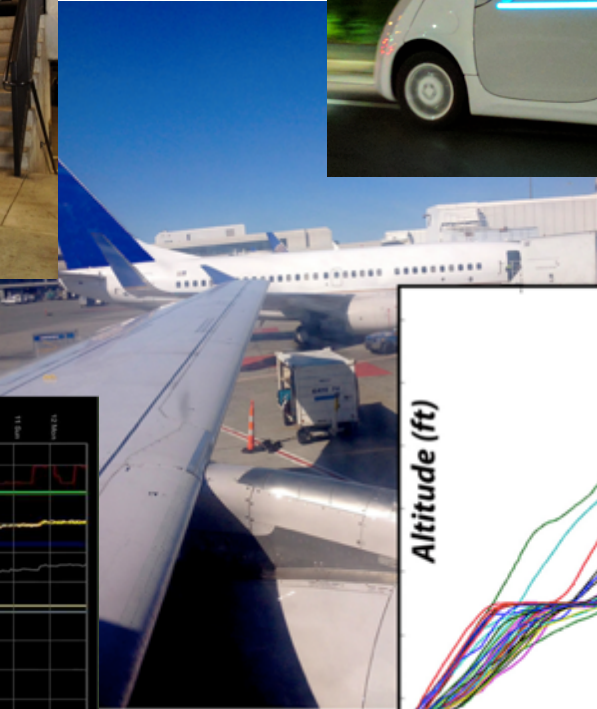
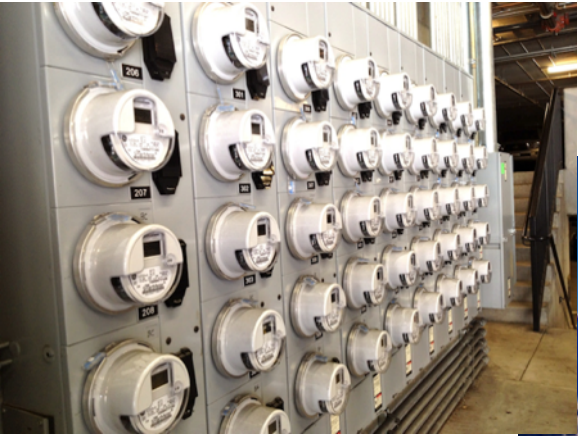
Domain Knowledge Matters: Detecting Security Attacks

Security expert at a bank preserved headers for web site requests

- Detected anomaly in headers for the attackers vs normal (real) requests
- Pattern of behavior for attackers was allowable for headers and it was not predictable: but it was different

Keep data:
You don't know what you'll
need to know later

Big Industry, Big Data, Big Value



All other images © E. Friedman



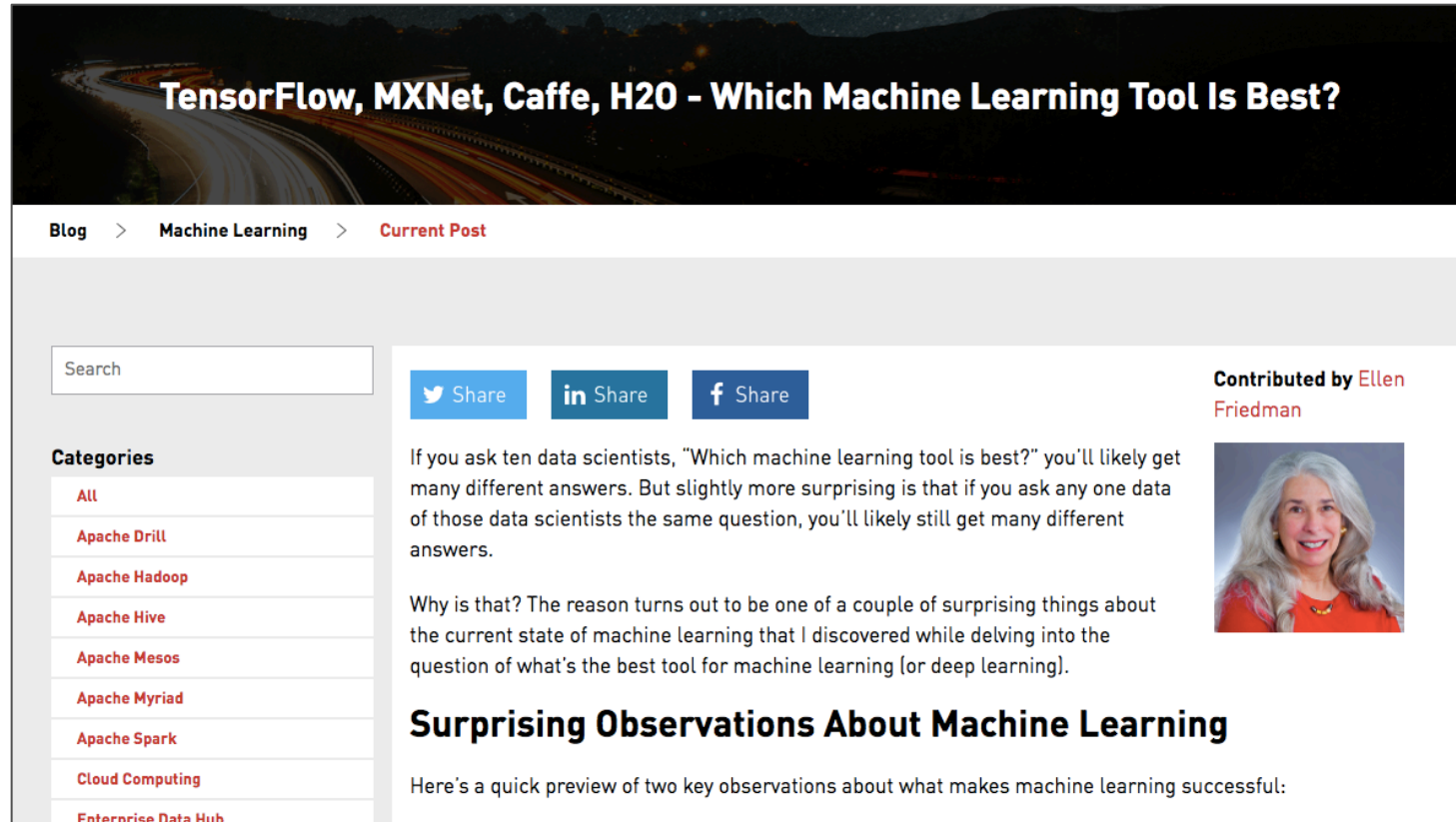
Image courtesy Mtell used with permission

Simple But Valuable: Accounting Audit Targeting

Big industrial company with a lot of machinery

- Tracking actions & parts
 - label as repairs, contracted services, delivery of supplies, etc.
- Which are to be taxed, expensed or counted as revenues
 - Mislabelling can cost millions of dollars
- Use machine learning to target potential mislabelling for audit review
- Relatively simple models (pattern matching; exception detection) deliver a huge business value.

Is it the algorithm? the model? the ML tool?



The screenshot shows a blog post on the MapR website. The title is "TensorFlow, MXNet, Caffe, H2O - Which Machine Learning Tool Is Best?". The author is Ellen Friedman. The post content begins with the text: "If you ask ten data scientists, 'Which machine learning tool is best?' you'll likely get many different answers. But slightly more surprising is that if you ask any one data of those data scientists the same question, you'll likely still get many different answers." This is followed by a sub-section titled "Surprising Observations About Machine Learning" with the introductory text: "Here's a quick preview of two key observations about what makes machine learning successful:". The left sidebar contains a search bar and a list of categories including All, Apache Drill, Apache Hadoop, Apache Hive, Apache Mesos, Apache Myriad, Apache Spark, Cloud Computing, and Enterprise Data Hub. Social sharing buttons for Twitter, LinkedIn, and Facebook are also visible.

TensorFlow, MXNet, Caffe, H2O - Which Machine Learning Tool Is Best?

Blog > Machine Learning > Current Post

Search

Categories

- All
- Apache Drill
- Apache Hadoop
- Apache Hive
- Apache Mesos
- Apache Myriad
- Apache Spark
- Cloud Computing
- Enterprise Data Hub

Share in Share f Share

Contributed by Ellen Friedman

If you ask ten data scientists, "Which machine learning tool is best?" you'll likely get many different answers. But slightly more surprising is that if you ask any one data of those data scientists the same question, you'll likely still get many different answers.

Why is that? The reason turns out to be one of a couple of surprising things about the current state of machine learning that I discovered while delving into the question of what's the best tool for machine learning (or deep learning).

Surprising Observations About Machine Learning

Here's a quick preview of two key observations about what makes machine learning successful:

<https://mapr.com/blog/tensorflow-mxnet-caffe-h2o-which-ml-best/>

90% of the effort in successful
machine learning isn't the
algorithm or the model...

It's the logistics

What Does Streaming Do for You?

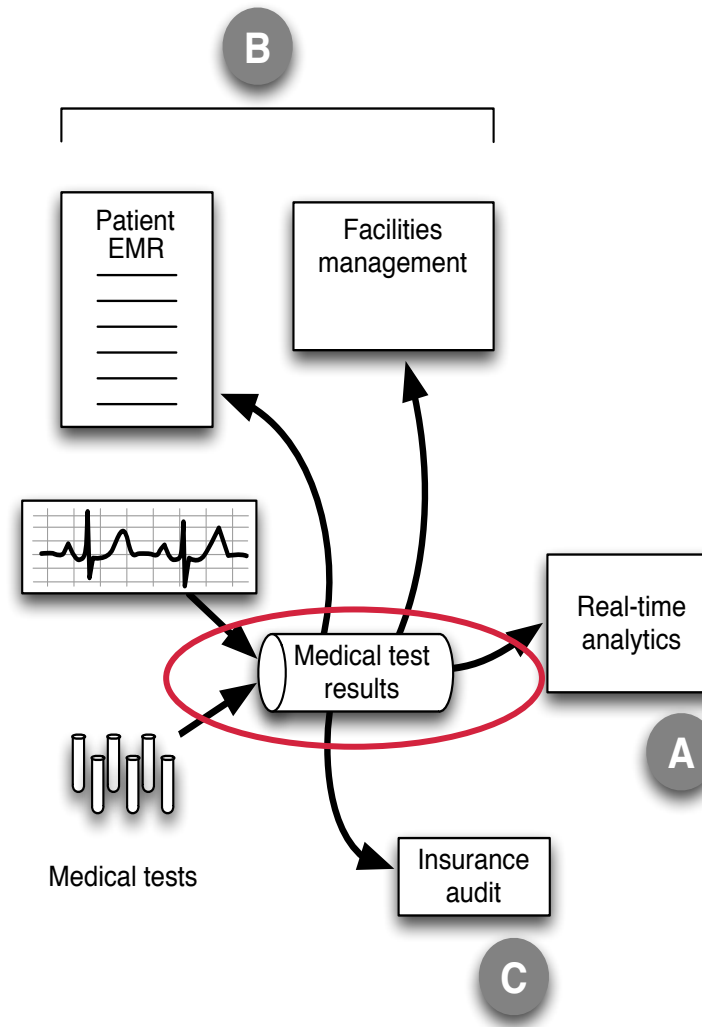


Surfer on standing wave, Munich Image © 2017 Ellen Friedman

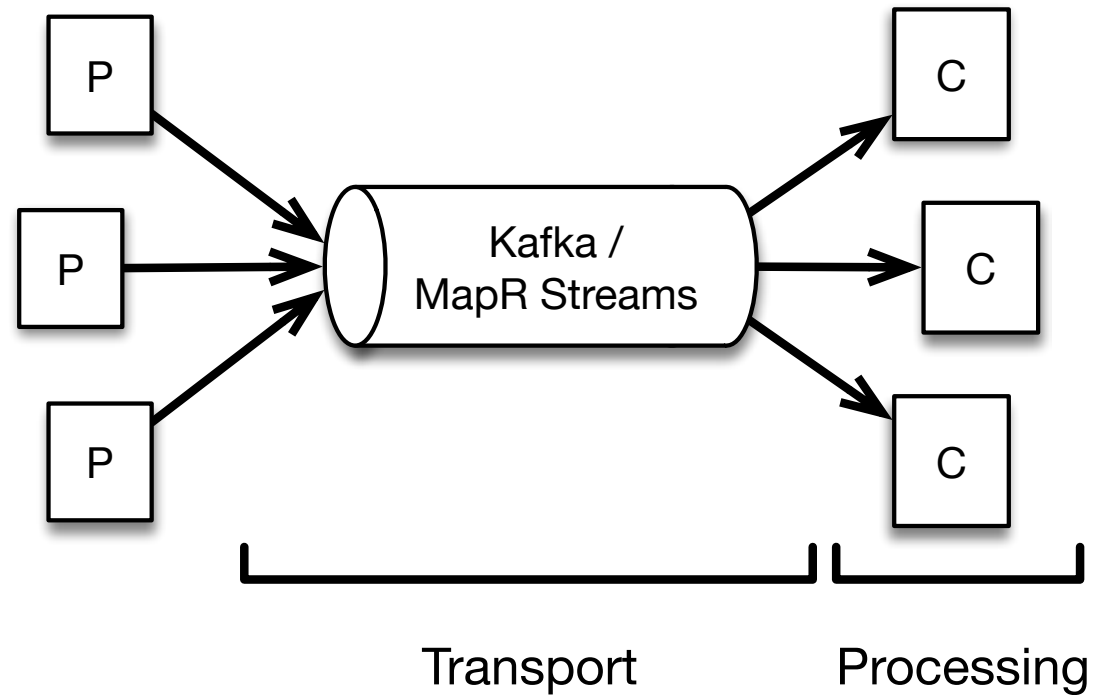
Stream transport supports
microservices

At the Heart: Message Transport

With the right messaging tool at the heart of stream-1st architecture you support other classes of use cases (B & C)



Stream Transport that Decouples Producers & Consumers



Good stream transport is persistent, performant & pervasive!

Streaming Microservices

- “Streaming Microservices” by Ted Dunning & Ellen Friedman, chapter in *Encyclopedia of Big Data Technologies*, Sherif Sakr and Albert Zomaya, editors, © 2018 (Springer International Publishing)
- Chapter 3 of *Streaming Architecture* by Ted Dunning & Ellen Friedman © 2016 (O’Reilly Media)
<https://mapr.com/ebooks/streaming-architecture/chapter-03-streaming-platform-for-microservices.html>

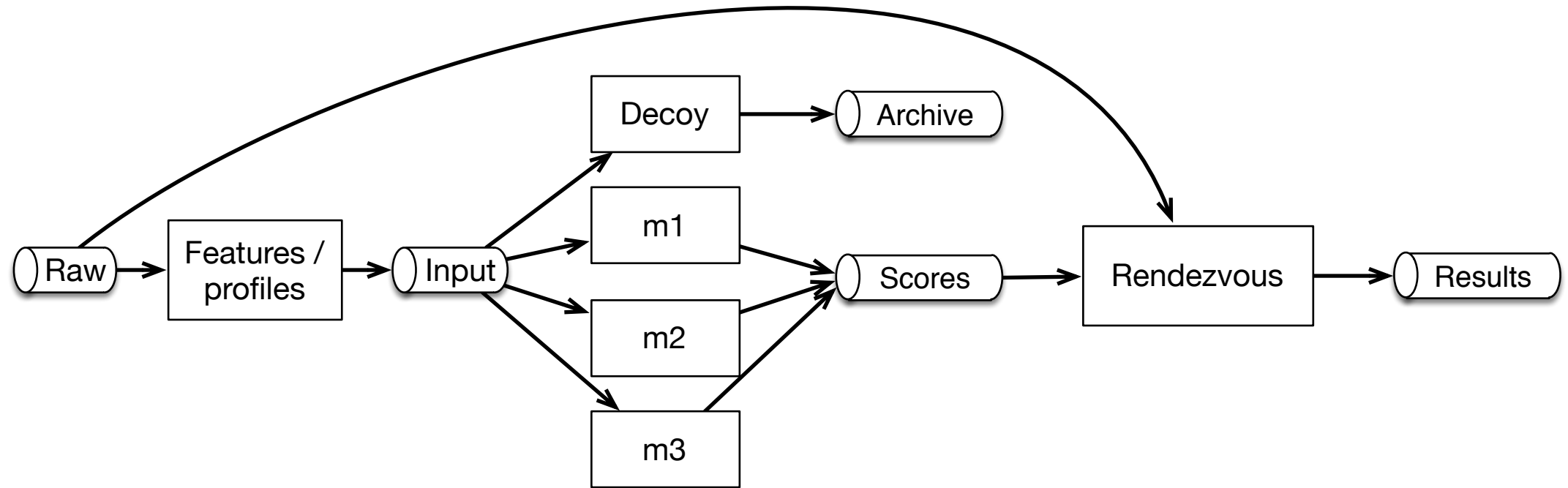
Get rid of the myth of the
unitary model

Streaming microservices provides
flexibility & independence to
manage many models

Logistics for machine learning can be difficult

- Just getting the training & input data is hard
- Many models to manage
- Model-to-model evaluation needs to be convenient & accurate
- Respond as the world changes: Deploy to production with agile roll out & roll
- There's a need for good data engineering (not just data science)

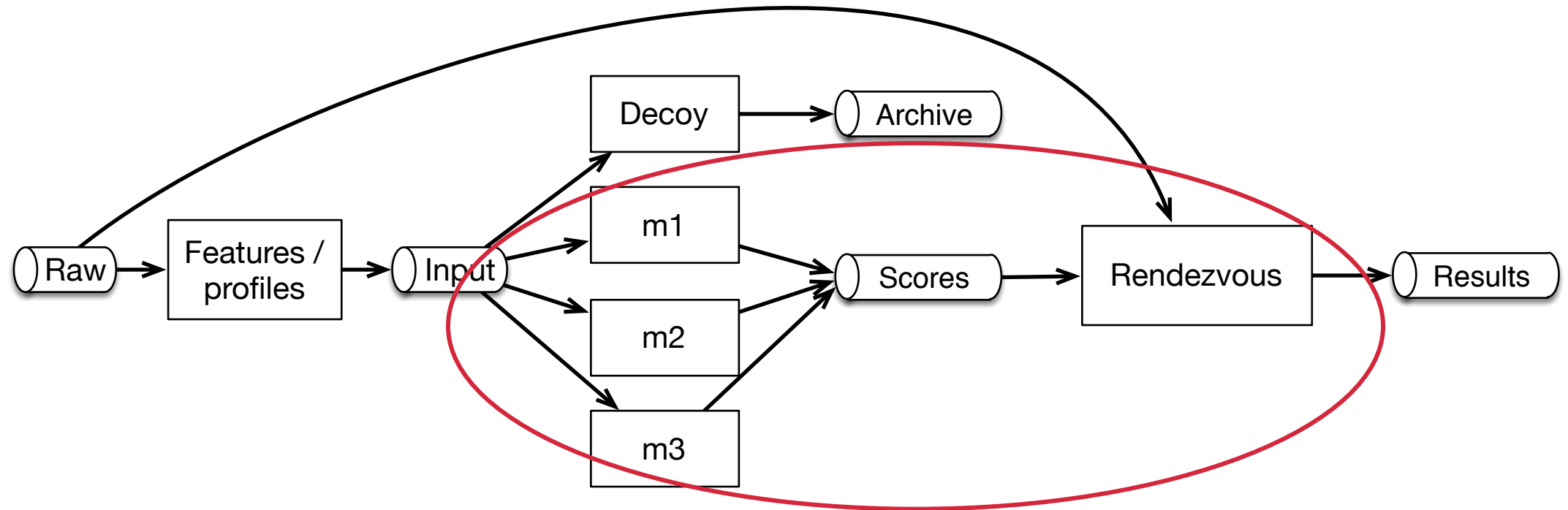
Enter Rendezvous...



Rendezvous Architecture described in:

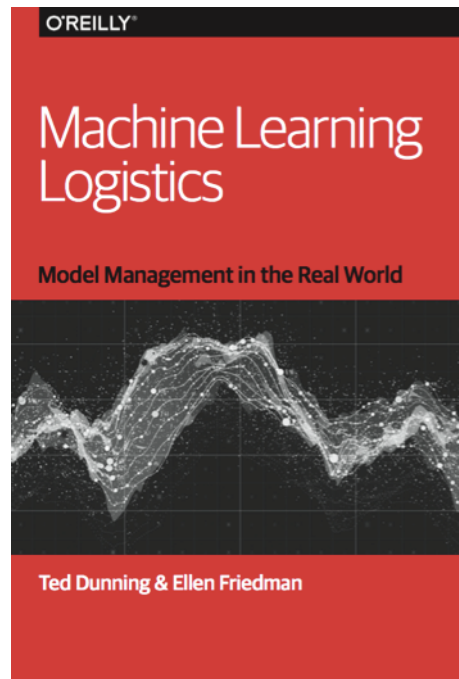
- *Machine Learning Logistics* book by Ted Dunning & Ellen Friedman © 2018 (O'Reilly)
- "Rendezvous Architecture" by Ted Dunning & Ellen Friedman, chapter in *Encyclopedia of Big Data Technologies*. Sherif Sakr and Albert Zomaya, editors. Springer International Publishing, © 2018 in press.

Rendezvous server making continuous decisions



Machine Learning Logistics: Model Management in the Real World

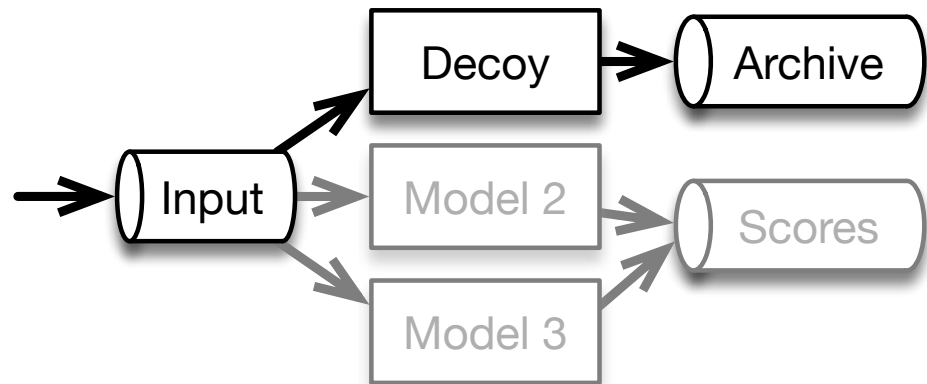
By Ted Dunning & Ellen Friedman © September 2017



Free copy online courtesy of MapR:
<https://mapr.com/ebook/machine-learning-logistics/>

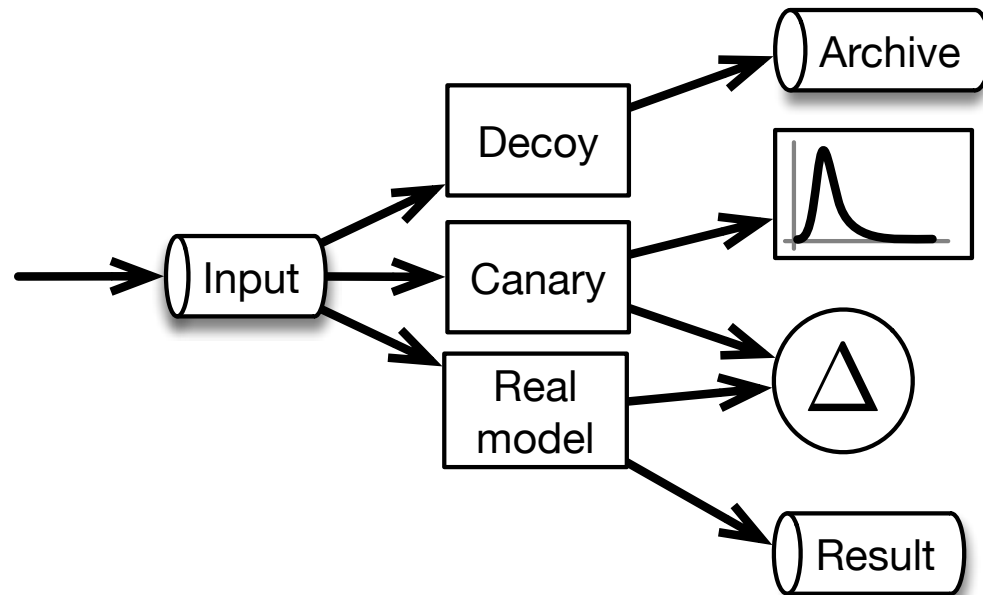
Decoy Model in the Rendezvous Architecture

- Looks like a model, but it just archives inputs
- Safe in a good streaming environment, less safe without good isolation



Canary Model Helps Gauge Performance of New Models

- Acts as a baseline reference for reasonable performance
- Can be compared to performance of newly deployed models

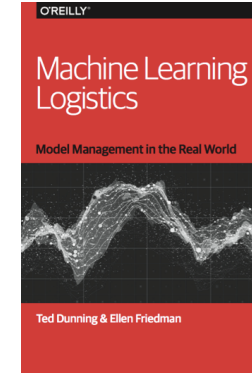


To deploy a new model, just
stop ignoring it

Rendezvous: Mainly for Decisioning Type Systems

- Decisioning style machine learning
 - Looking for “right answer”
 - Simpler than interactive machine learning (think self-driving car)
- Examples: fraud detection, predictive analytics; churn prediction (telecom); deep learning (speech or image recognition)

Update on Machine Learning Logistics



Rendezvous Architecture

- Implemented by Terry McCann (Adatis) and Boris Lublinsky (Lightbend)
- Soon to be put into a product by Finnish company Valonai
- Large US financial company: data science team starting to plan how to build rendezvous servers for each major project
- “Rendezvous Architecture” is an entry in Springer’s *Encyclopedia of Big Data Technologies* (© 2018)

Update on Machine Learning Logistics

- MLFlow

- From DataBricks to focus on early part of logistics

<https://databricks.com/blog/2018/06/05/introducing-mlflow-an-open-source-machine-learning-platform.html>



- Clipper

- From UC Berkeley Rise Lab; has deployment of models; some similar goals to Rendezvous

<https://rise.cs.berkeley.edu/projects/clipper/>

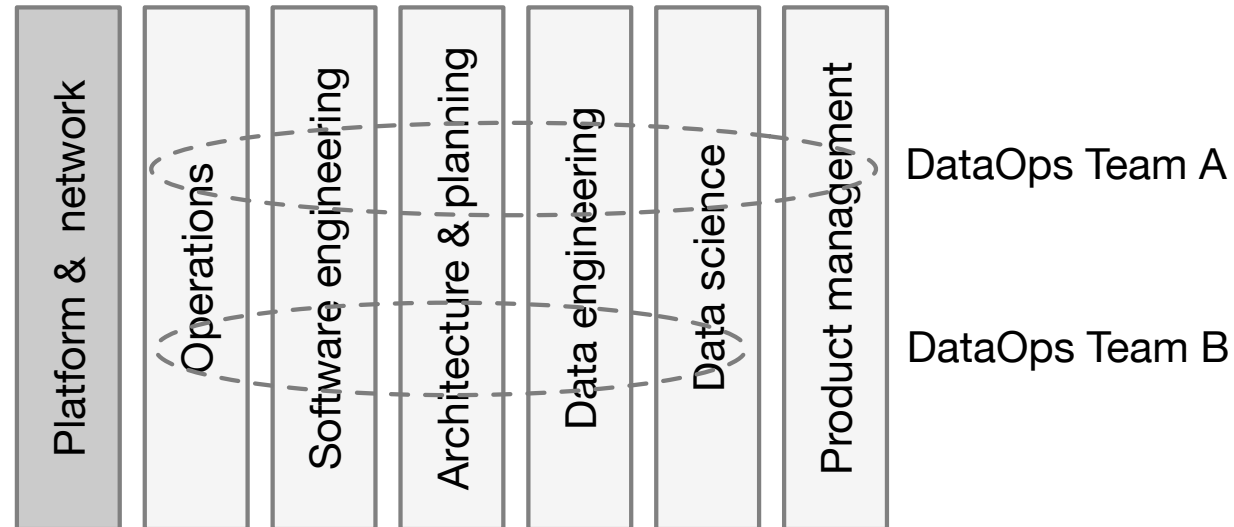


Be ready to adapt:
Conditions will change

DataOps: Brings Flexibility & Focus

- You don't have to be a data scientist to contribute to machine learning
- Software engineer/ developer plays a role: but you need good data skills

Cross functional DataOps teams



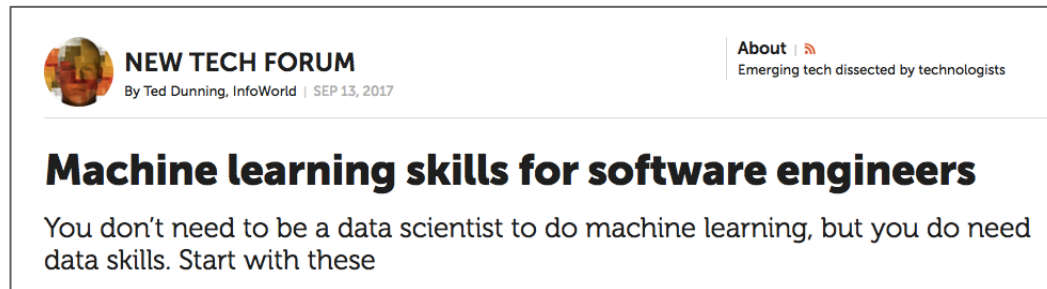
DataOps Principles



“DataOps teams seek to orchestrate data, tools, code and environments from beginning to end.”

Thor Olavsrud interview with Ted Dunning & Ellen Friedman for CIO

<https://www.cio.com/article/3237694/analytics/what-is-dataops-data-operations-analytics.html>



by Ted Dunning 13 Sept 2017 in "InfoWorld"

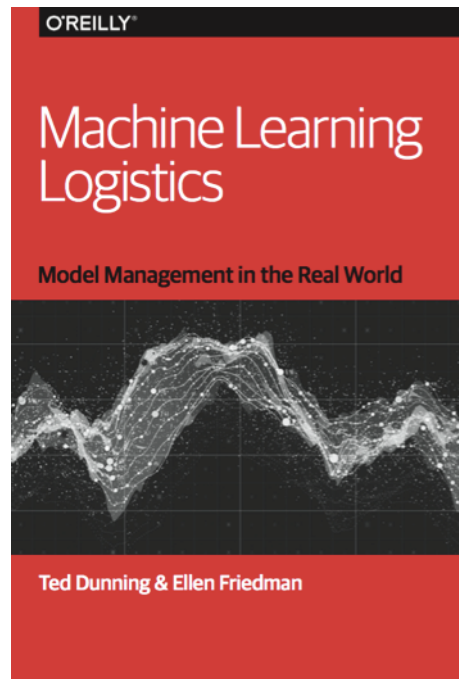
<https://www.infoworld.com/article/3223688/machine-learning/machine-learning-skills-for-software-engineers.html>

Advantages of a DataOps Approach

- Able to pivot & respond to real-world events as they happen
- Improved efficiency and better use of people's time
- Faster time-to-value
- A good fit to working with a global data fabric

Machine Learning Logistics: Model Management in the Real World

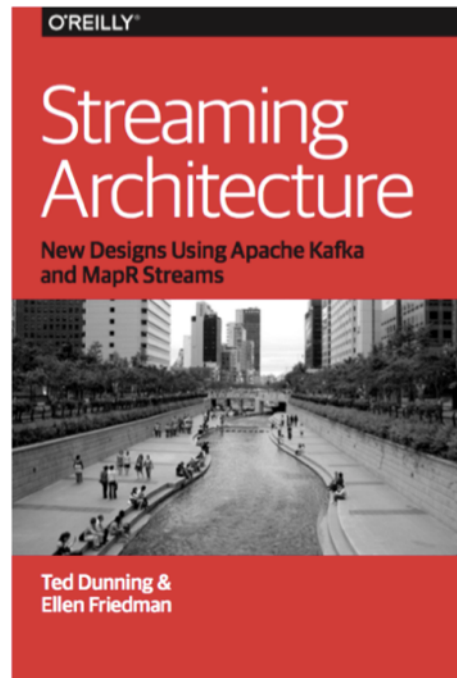
By Ted Dunning & Ellen Friedman © September 2017



Free copy online courtesy of MapR:
<https://mapr.com/ebook/machine-learning-logistics/>

Streaming Architecture: New Designs Using Apache Kafka & MapR Streams

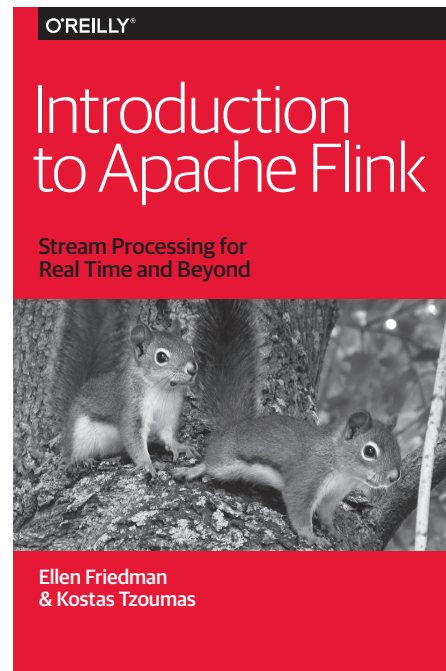
By Ted Dunning & Ellen Friedman © March 2017



Free copy online courtesy of MapR:
<http://bit.ly/mapr-streaming-architecture-book>

Introduction to Apache Flink

By Ellen Friedman & Kostas Tzoumas © September 2016



Free copy online courtesy of MapR:
<https://mapr.com/ebooks/intro-to-apache-flink/>



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

#womenintech #datawomen

© Ellen Friedman 2015



Thank you !

Contact Information

Ellen Friedman, PhD

Principal Technologist, MapR Technologies

Committer Apache Drill & Apache Mahout projects

O'Reilly author

Email efriedman@mapr.com ellenf@apache.org

Twitter @Ellen_Friedman #bbuzz