Surprising Course of Knowledge & Innovation

Ellen Friedman
Berlin Buzzwords
6 Jan 2016 #bbuzz

Contact Information

Ellen Friedman

Solutions Consultant, MapR Technologies

Committer: Apache Drill and Apache Mahout

O'Reilly author

Email ellenf@apache.org

efriedman@maprtech.com

Twitter @Ellen_Friedman

Hashtag today #bbuzz

Take time to think

Focus in on what matters

Ellen Friedman Berlin Buzzwords 2015
Image credit newthinking communications http://bit.ly/berlin-buzzwords-2015-ef

Get past the details

Discover the key concepts (not just generalizations)

- Helps you be clever

- Helps you communicate

- Helps you innovate

Innovation in big data (and open source)

A data success story, with a twist…

- Big data
- Open source
- Crowd source
- Time series data
- Data-driven decisions that saved lives & made people rich

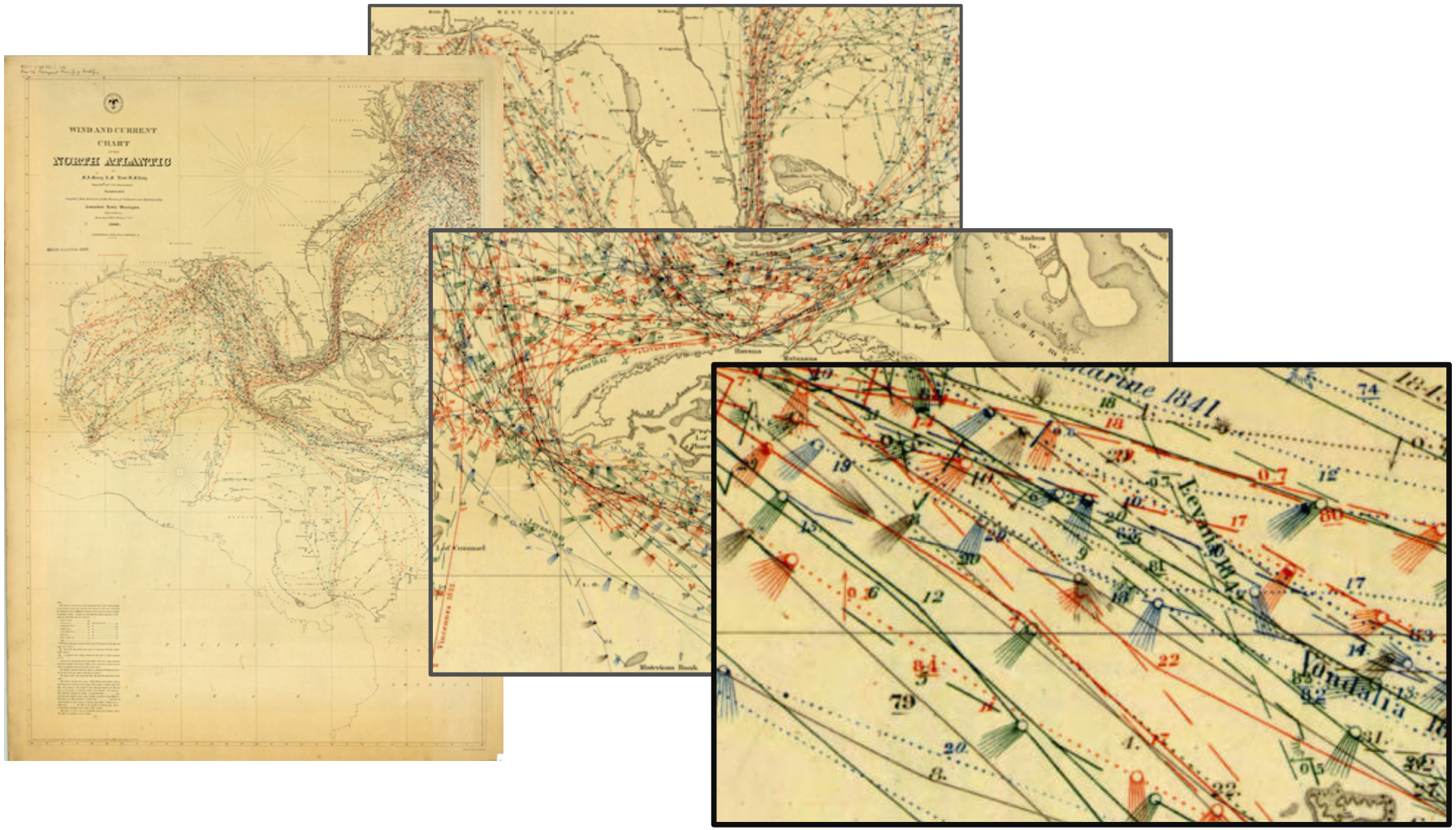
It all happened in the 19th century…

Matthew Fountain Maury was an officer in the US Navy in the 1830s

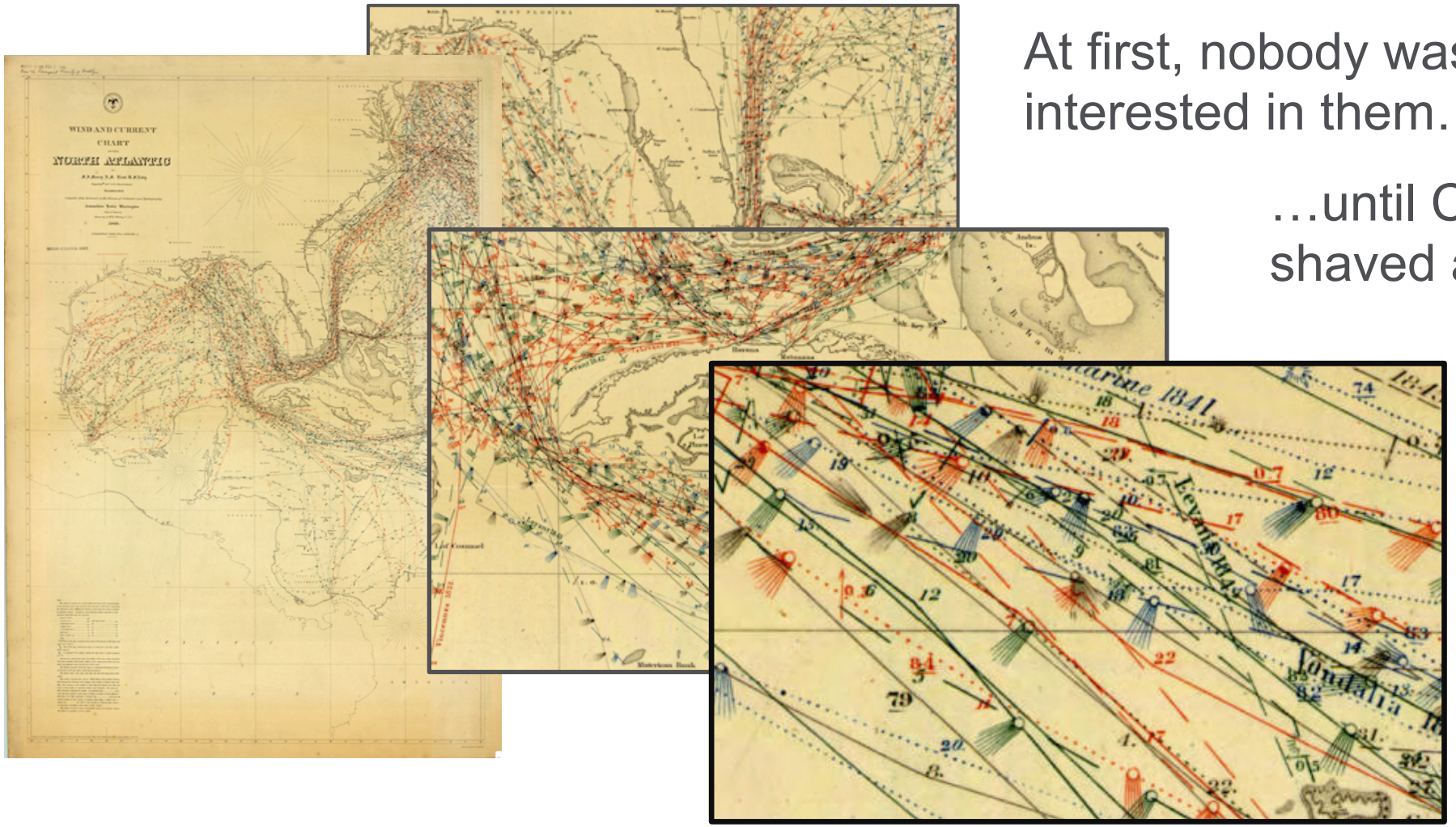
He injured his leg, so the Navy gave him a “desk job”

Oddly, that’s where his real adventure started

Big data project: Maury's *Wind and Currents* charts



Big data project: Maury's *Wind and Currents* charts



At first, nobody was interested in them…

…until Captain Jackson shaved a month off the run from Baltimore to Rio de Janeiro

Then everybody wanted one!

Lessons:

- Aggregate big data to add value greater than individual parts
- Take human motivation into account for successful projects
- Get to the essential concepts that matter – best way to innovation

Un-discovering the cure for scurvy

The Cure for Scurvy: Vitamin C

We all know:

- Horrible disease “scurvy” is cured by ascorbic acid (Vitamin C) in small amounts
- Vitamin C is found in fresh foods, especially citrus, watercress and to some extent meat.

The Cure for Scurvy: Timeline of Discovery

- 1747 Lind discovers fresh citrus cures scurvy
- 1753 Lind publishes *Treatise of the Cure for Scurvy*
- 1799 Royal British Navy ships required to use lemon juice

- 1933 Szent-Gyorgyi identifies Vitamin C – ascorbic acid

The Cure for Scurvy: Timeline of Discovery

- 1747 Lind discovers fresh citrus cures scurvy
- 1753 Lind publishes *Treatise of the Cure for Scurvy*
- 1799 Royal British Navy ships required to use lemon juice

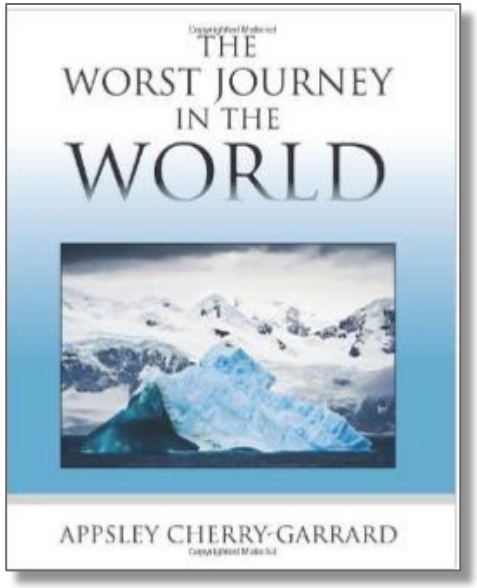
- 1911 Cure for scurvy now unknown
  Thought to be caused by darkness, cold and hard work

- 1933 Szent-Gyorgyi identifies Vitamin C – ascorbic acid

heh?

Un-discovering the Cure for Scurvy

Story of Scott's 1911 expedition to the South Pole



The Worst Journey in the World

Book by Apsley Cherry-Garrard originally written in 1922. Current edition 2011

http://bit.ly/worst-journey-book

"Idle Words" blog on lost knowledge of cure for scurvy March 2010
http://bit.ly/scurvy-blog

The Cure for Scurvy: Timeline of Discovery

- 1747 Lind discovers fresh citrus cures scurvy
- 1753 Lind publishes *Treatise of the Cure for Scurvy*
- 1799 Royal British Navy ships required to use lemon juice

- 1860 British Navy substitutes West Indies bottled lime juice
  ineffective against scurvy, but people didn't notice

- 1911 Cure for scurvy now unknown
  Thought to be caused by darkness, cold and hard work

- 1933 Szent-Gyorgyi identifies Vitamin C – ascorbic acid

heh? again

Un-discovering the Cure for Scurvy

How did people “lose” the knowledge?

- 1860 Shifted from Sicilian lemons to West Indies limes (more acid, less Vitamin C)
- Using bottled juice instead of fresh citrus; exposed to air in big vats; piped through copper
- Faster transport by steam-driven ships: less time to develop disease
- Long polar expeditions began to suffer from scurvy even with lime juice

.

A Modern Laboratory Story

- A standard procedure in molecular biology fails one day
- Technician is frustrated but carefully repeats procedure: fails again
- Technician repeats procedure many times: considers alternative career.
- Better:
  - Ask, “What changed?”
  - If answer appears to be “nothing changed”, assume laws of physics still OK and look again. “What changed?”
  - Answer: One reagent was from a new lot.
  - New lot was contaminated.

When did things last work?

What changed?

What about you? “It worked on my machine.”

- Different environment going from development to production

- Possibilities:
  - Less memory
  - More memory
  - Too many threads
  - Different coordination
  - <insert your experience here>

Consider alternative explanations

How do you know what you know?

Dog food testing:
Which bowl?

Image © Ellen Friedman 2015

A/B test for new dog food flavor:

Old flavor in one bowl; new flavor in the other.

Which does dog prefer?

How do you know you're testing his preference for flavor and not just for the position of the bowls?

How do you know what you know?

Dog food testing:
Which bowl?

Data scientists:
Which book?

Be careful how you design tests:

Do you know what factor you are testing?

Are you testing what you think you're testing?

Consider alternative explanations

What about you? Are you using the right data?

Video Recommender

- Input data as clicks: recommender gives poor performance
- Input data as 1st 30 seconds viewing: recommender good!

First case, testing how well people liked the titles

Second case, testing how well people liked the videos

What about you? Target Leak

News groups classifier

- Even articles were used to train; odd articles used to test
- Result: REALLY really good

  In fact, too good to be true…

Problem:

Many articles were followed by one with only slight modifications; so test data overlapped with training data

Clue:

When data was split into before and after blocks, the classifier broke

If it’s too good to be true…

Consider alternative explanations

Early Chemistry – 18th century experiments of Lavoisier

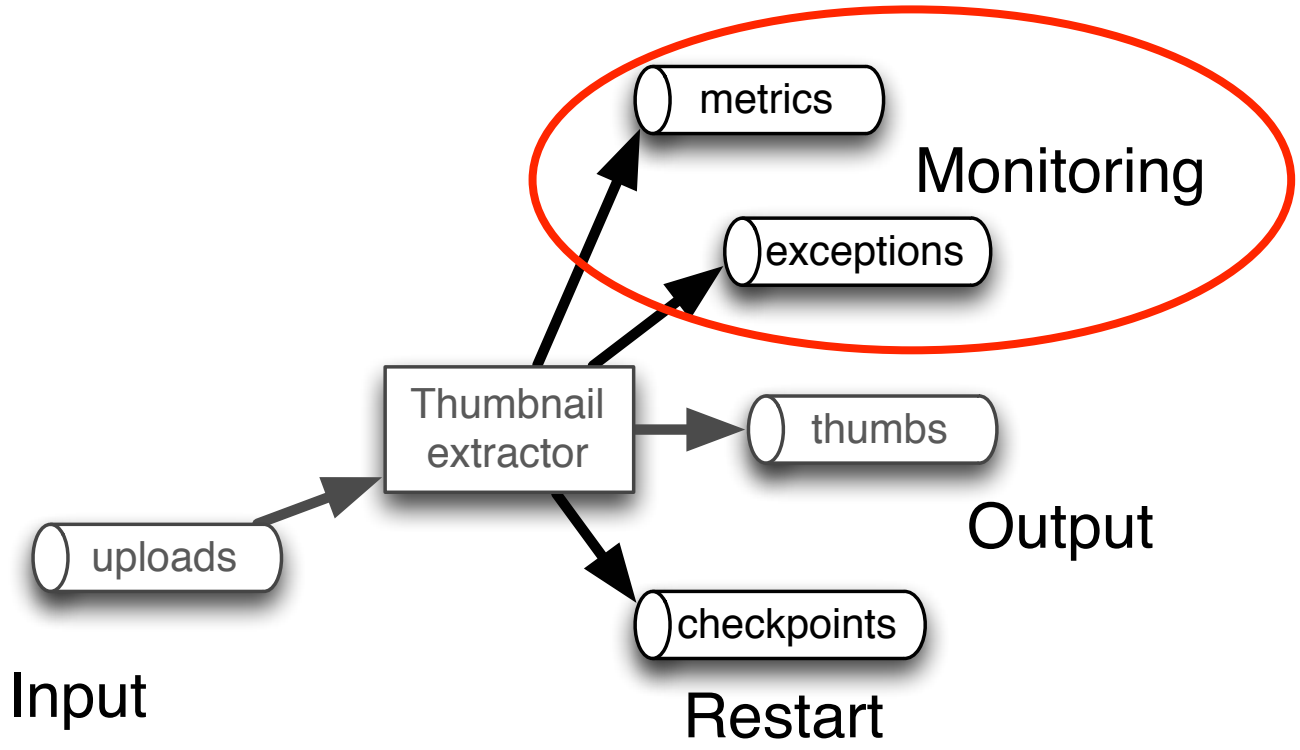
- Carefully weighed materials in combustions inside sealed jars to include solids & gases
- Supported idea of conservation of matter
- Killed idea that matter contained “phlogiston”
- Discovered & named oxygen

Image public domain
http://bit.ly/wiki-lavoisier-lab

Careful measurements are important

What about you?

Modern stream-based architectures support widespread metrics



From Chap 3 *Streaming Architecture*
by Ted Dunning & Ellen Friedman (O'Reilly) 2016.

Metrics, metrics, metrics

What about you? Detecting Security Attacks

Example - doing it right at a bank:

- Saved headers for web site requests
- Detected anomaly in headers for the attackers

Lesson:

- Keeping careful history paid off
- You may not know what you'll need to know later

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

What about you? Long term re-playable log

- You can save years of data streams
- Surprising: A topic partition in MapR Streams is distributed across the entire cluster, not limited to one machine

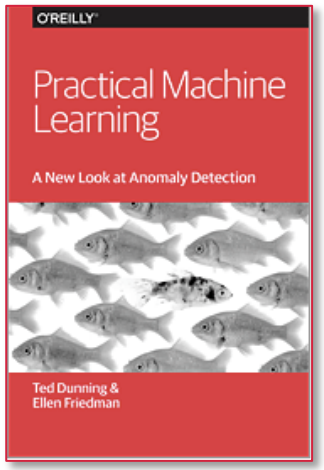
Lessons - Summary

- Take time to think: Find the essential concepts in what you do
- Things break. Ask: When did they last work? What changed?
- Look for alternative explanations
- Are you testing what you think you're testing?
- When it's "too good to be true", look for alternative explanations
- Careful measurements, metrics, monitoring help
- Save data: you don't know what you'll want to know later

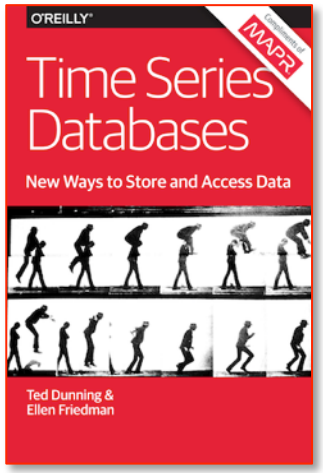
Short Books by Ted Dunning & Ellen Friedman

Free pdf download courtesy of MapR www.mapr.com/ebook

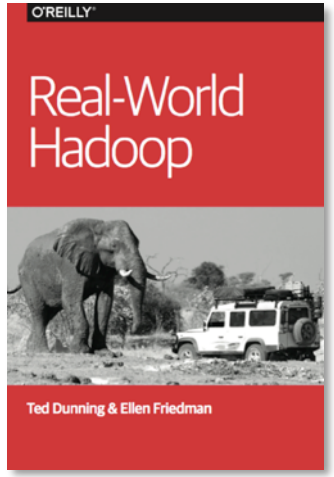
http://bit.ly/recommendation-ebook



http://bit.ly/ebook-anomaly



http://bit.ly/mapr-tsdb-ebook



http://bit.ly/ebook-real-world-hadoop

http://bit.ly/mapr-ebook-sharing-data

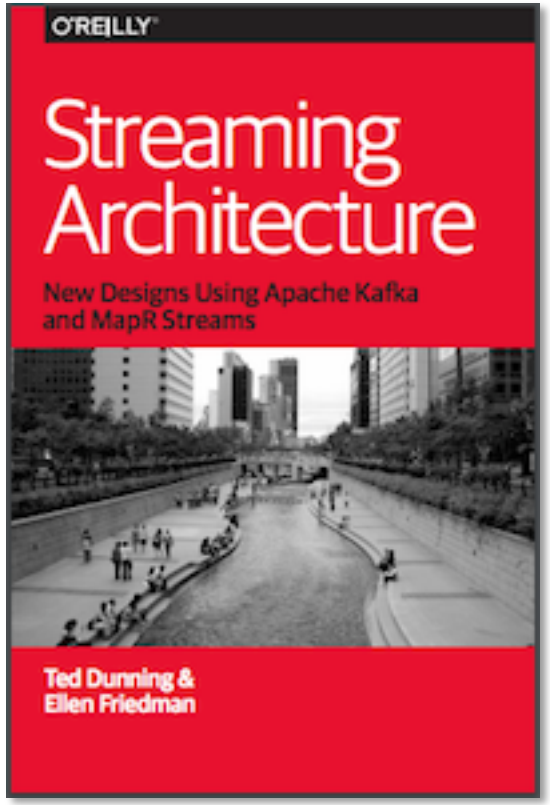
You're invited to a free copy of our new book

Streaming Architecture:

New Designs Using Apache Kafka & MapR Streams

Book signing: Ted Dunning & Ellen Friedman

Free online at www.mapr.com/ebooks



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

Thank you !

Contact Information

Ellen Friedman

Solutions Consultant, MapR Technologies
Committer: Apache Drill and Apache Mahout
O'Reilly author

Email ellenf@apache.org
efriedman@maprtech.com

Twitter @Ellen_Friedman

Hashtag today #bbuzz