





Moving from Narrative to Design with Open Source and Open Data

Jeff Hammerbacher

Chief Scientist and Vice President of Products, Cloudera

April 23, 2010

Presentation Outline

- Who am I and what am I talking about?
 - My Background
 - Defining “Narrative” and “Design”
- Two stories
 - Finance
 - Web
- Making the Sage story a success
 - Open Source
 - Open Data
 - The Unreasonable Effectiveness of Data

My Background

Thanks for Asking

- **hammer@cloudera.com**
- Studied Mathematics at Harvard
- Worked as a Quant at **Bear Stearns**
- Conceived, built, and led Data team at **Facebook**
 - Nearly 30 amazing engineers and data scientists
 - Several open source projects and research papers
- Founder of **Cloudera**
 - Vice President of Products and Chief Scientist
 - Also, check out the book “Beautiful Data”



Tell me. Maybe Iz help.

ICANHASCHEEZBURGER.COM 🐱 🐱

“Genomic Advances of the 2000s Will
Demand an Informatics Revolution in the
2010s”

-- Eric Schadt

From Narrative to Design

Learning from past revolutions

- Leaving technology aside, once you have a narrative:
 - collect and structure data
 - build tools, models, and a domain vocabulary (“understand”)
 - use effective models to achieve your goal (“automate”)
- Pinnacle: Boeing 777, designed entirely on a computer
- Thanks to “Biology is Technology” by Robert Carlson for the “Narrative” and “Design” terminology

Two Stories



Finance

A Cautionary Tale

- Strengths
 - Complex models
 - Low latency decisions
 - Compute grids
- Weaknesses
 - Data needed for decisions is expensive
 - Code unrelated to automating decisions is highly guarded
 - Storing and manipulating large data sets
- Narrative to Design: trading strategies, financial products



The Web

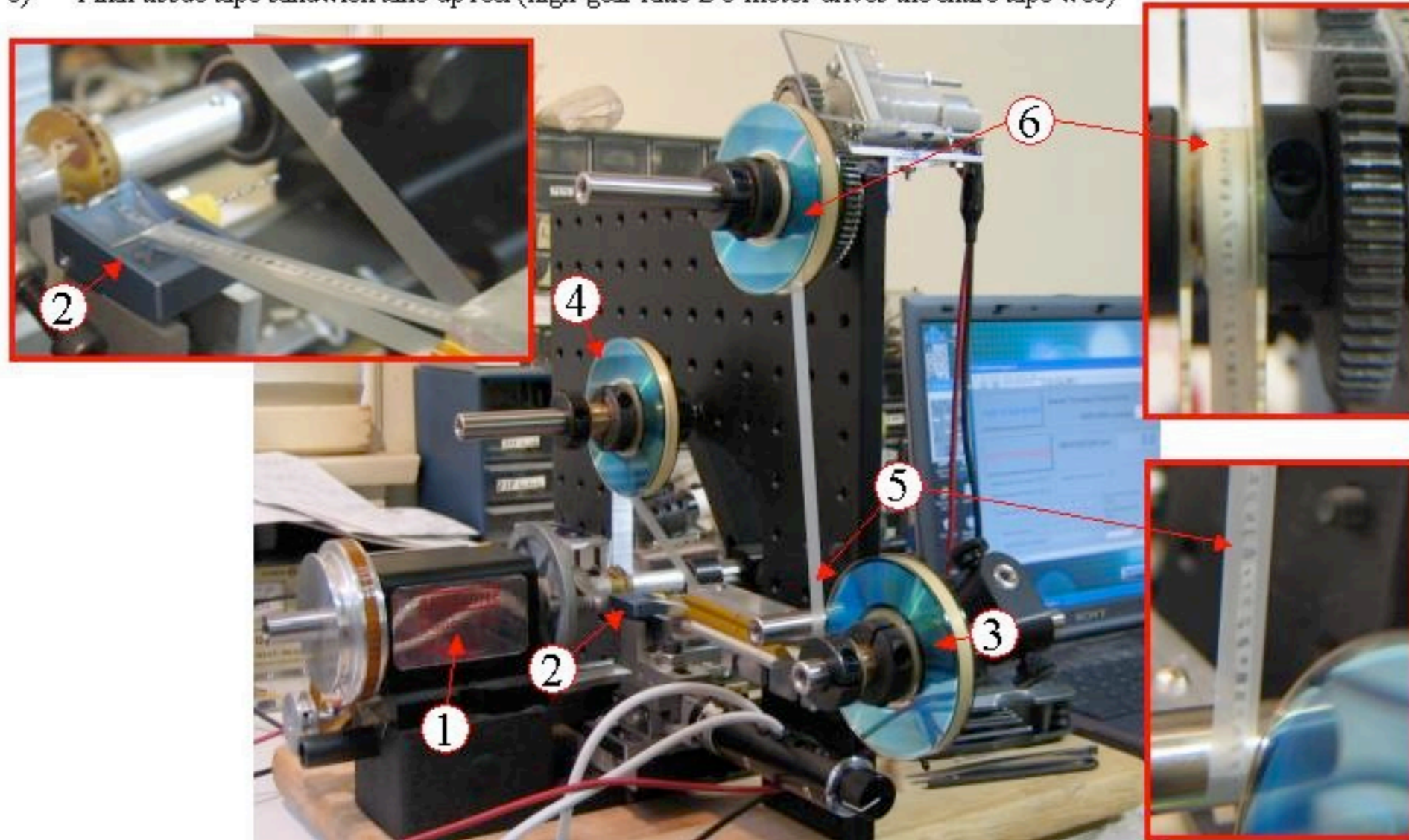
A Success Story

- Strengths
 - Open data
 - Open source
 - Scalable data management and analysis
- Weaknesses
 - Low latency
 - Complex models
 - Style of dress
- Narrative to Design: search quality, advertising targeting

Another Story

Prototype Automatic Taping Lathe-Microtome with Tape Web (conveyor-belt design)

- 1) Lathe spindle and base
- 2) Water-filled knife boat with partially-submerged conveyor belt composed of clear Mylar tape
- 3) Mylar tape feed reel (feeds conveyor belt)
- 4) Mylar tape feed reel (provides cover tape)
- 5) Tissue tape sandwich (inset shows 10 μ m thick ribbon sliced off block and sandwiched between two Mylar tapes)
- 6) Final tissue tape sandwich take-up reel (high-gear ratio DC motor drives the entire tape web)



Extreme Neuroanatomy

DIY Brain Imaging

- Built by Ken Hayworth, ex-JPL engineer
- Now a critical contributor to “Connectome” project
 - Synapse-resolution connectivity diagram of the brain
- Image data for a single rat brain: 900 PB
- This looks familiar!
 - hackers
 - large data sets
 - machine learning

A New Story

Commons

Background
Repository
Forum
Congress

Training

Presentations
Center for Cancer
Systems Biology

Company

Careers
Directors
Partners
News
FAQs

Research

Case Studies
Publications
Resources
Tools



Biology

Finance or the Web?

- Large drug companies not that different from banks
- Recent infusion of open source and open data initiatives!
- Learn from the web
 - Open source
 - Open data
 - “The Unreasonable Effectiveness of Data”
- Use existing open source software
 - Cambrian explosion in open source data management
 - Data storage and processing at petabyte scale is solved

Last Story



High-Energy Physics

Science at the Petabyte scale

- Tremendous investment in **measurement**
 - Large Hadron Collider cost estimated at \$4.4B
- Teams spend years performing **analysis**
 - Shared toolset for analysis and visualization has emerged
- Lessons for Sage
 - Invest in new measurement technologies and store everything
 - Use existing open source tools and share those you build

The Promise of Sage

Delivering on the Excitement

- An information platform for disease modeling and drug discovery
- More importantly: a community of developers and researchers
 - Developers, developers, developers!
 - Wait, that's you guys -- be hackers
- The most important person in an open source project
 - Not the creator
 - The first power user with no ties to the creator

Action Items

1. Share Data

2. Share Tools

3. Share Results



(c) 2009 Cloudera, Inc. or its licensors. "Cloudera" is a registered trademark of Cloudera, Inc.. All rights reserved. 1.0