Data Science Vancouver

Don Turnbull, PhD

@donturn

#datasciencevyr

Overview

- Goals for Data Science Vancouver
- My Background
- Philosophy & Thoughts on Data Science
  - Advantages of Experience
  - Research as an Advantage
- A (few) Projects
- Meeting our Goals

What I Did

- Software Developer
  - Expert Systems, SGML/Hypertext, CASE
- M.S. @ Georgia Tech
- Software Designer, Manager, Technical Architect
- Ph.D. @ U Toronto
  - KDD - (Web) Behavioral & Informetric Modeling
  - IR & ISeek - Systems to Augment Web Browsing and Search
- Principal @ Startup (Google)
  - Personalized IR
- Asst. Professor @ University of Texas

What I Do

- Consultant - Research & Development
  - Data Science & Intellectual Property
  - Plain, old Computer Science Research
- Startup Advisor
  - IR, Text Analysis, Analytics, Sentiment, CF
- Tapstream
  - Research on Marketing Attribution Analytics, Device and Behavior Profiling, Graph Analysis etc.

A Philosophy

A Pracitcal Philosophy

Ways of Knowing & Doing

The fox knows many things, but the hedgehog knows one big thing. **Archilochus**

I’m a Hedgehog

Except When I’m Not

One Big Thing

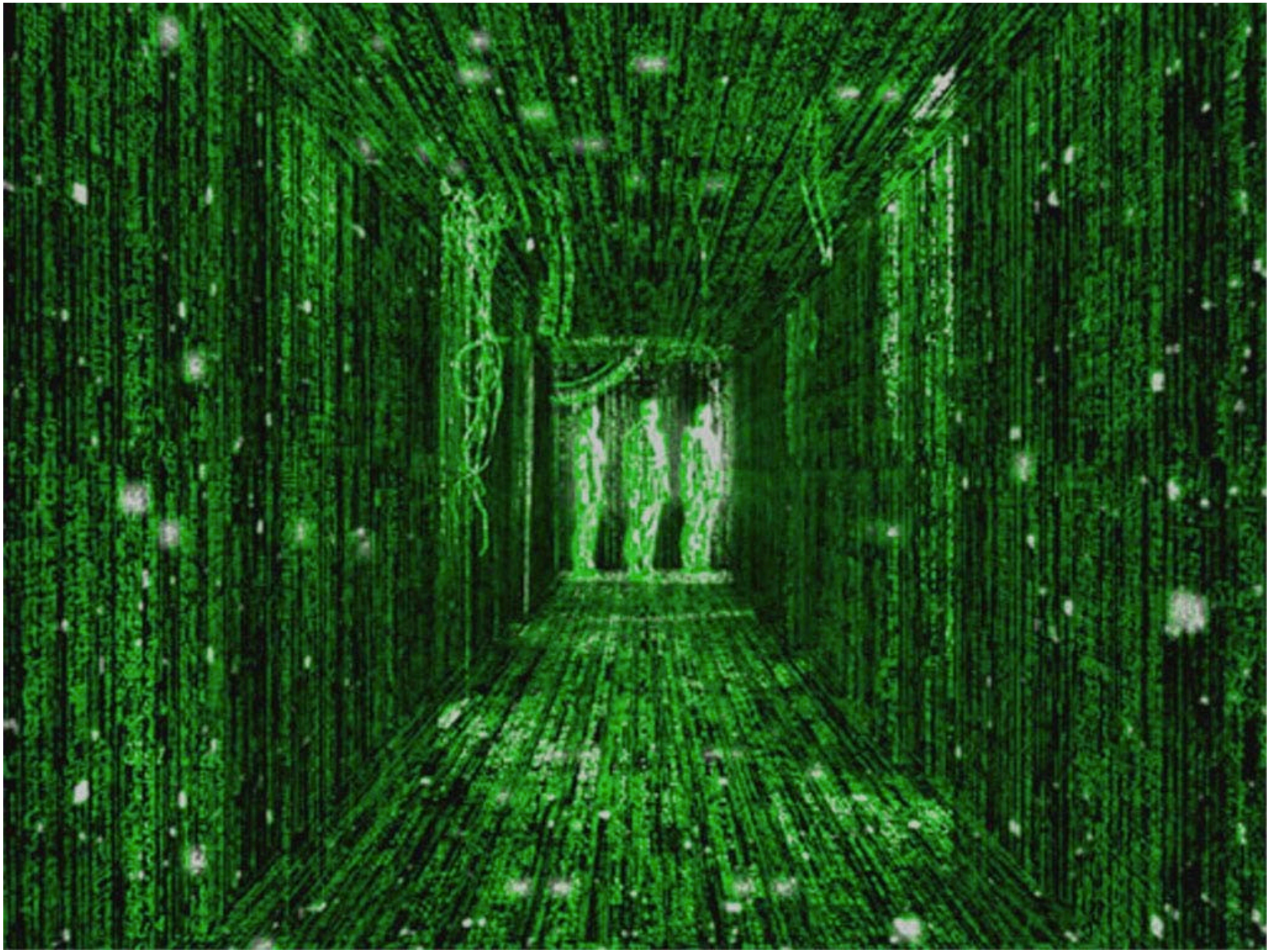
Is realy a lot of smaller things

(nothing to do with “Big Data”)

Science

Data Science

What does the world look like to a Data Scientist?



Data Science is really Science

- And Science runs on Quantitative Methods
  - Understanding what people *actually* do instead of what they *say* they do
  - Making comparisons (over time)
  - Generalizable and extensible
  - *The* toolbox for interpreting and analyzing other people's results

The power of (Data) Science

- It is a discipline
  - Hypothesis based
  - More applicable to peer review & verfication
- It requires a set of skills that have a (*much*) higher market value
- Many characteristics examined are constants
  - Human Behavior
  - Physical traits and abilities

The Rise of Data Science

- **Our time is now**
  - **Yottabyte, Teraflop, GPU, Cloud…**
- **Fights fire with Fire**
  - **Numbers speak the language of business & technology (C-level execs)**
- **(Almost) infallible results**
- **Qualitative decisions for Quantitative measurement**

Even *Normal* People are starting to get it

And they want to do it

The New York Times

Technology

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

SIGN IN TO RECOMMEND

TWITTER

COMMENTS (58)

E-MAIL

SEND TO PHONE

PRINT

REPRINTS

SHARE

Enlarge This Image

Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Multimedia

FlowingData

Rise of the Data Scientist

Posted by Nathan on Jun 4, 2009 to Data Design Tips, Featured, Statistics

Photo by majamarko

As we've all read by now, Google's chief economist Hal Varian commented in January that the next sexy job in the next 10 years would be statisticians. Obviously, I whole-heartedly agree. Heck, I'd go a step further and say they're sexy now - mentally *and* physically.

However, if you went on to read the rest of Varian's interview, you'd know that by *statisticians*, he actually meant it as a general title for someone who is able to extract information from large datasets and then present something of use to non-data experts.

Sexy Skills of Data Geeks

As a follow up to Varian's now-popular quote among data fans, Michael Driscoll of Dataspora, discusses the three sexy skills of data geeks. I won't rehash the post, but here are the three skills that Michael highlights:

1. Statistics - traditional analysis you're used to thinking about
2. Data Munging - parsing, scraping, and formatting data
3. Visualization - graphs, tools, etc.

Data Science in Practice

Josh Wills @josh_wills 3 May

Data Scientist (n.): **Person who is better at statistics than any software engineer and better at software engineering than any statistician.**

Practical Data Science

"A Data Scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning."

Hilary Mason, Chief Data Scientist @ Bit.ly

Why Data Science Now?

- Data Acquisition
- Computational power & networked systems
- We need new modeling techniques, even new metaphors to examine the complex systems we interact with
- Augment and Extend - Finance, Psychology, Physics & Computer Science
  - Isomorphic!

The (New) Era of Instrumentation

- We are undoubtedly in a new era of reasoning
- Scientific Engineering enabled the original Age of Reason
- Build **understanding** of intent & interactions
- Build (better) **models** of the world

The New York Times

Business Computing

Data Analysts Captivated by R's Power

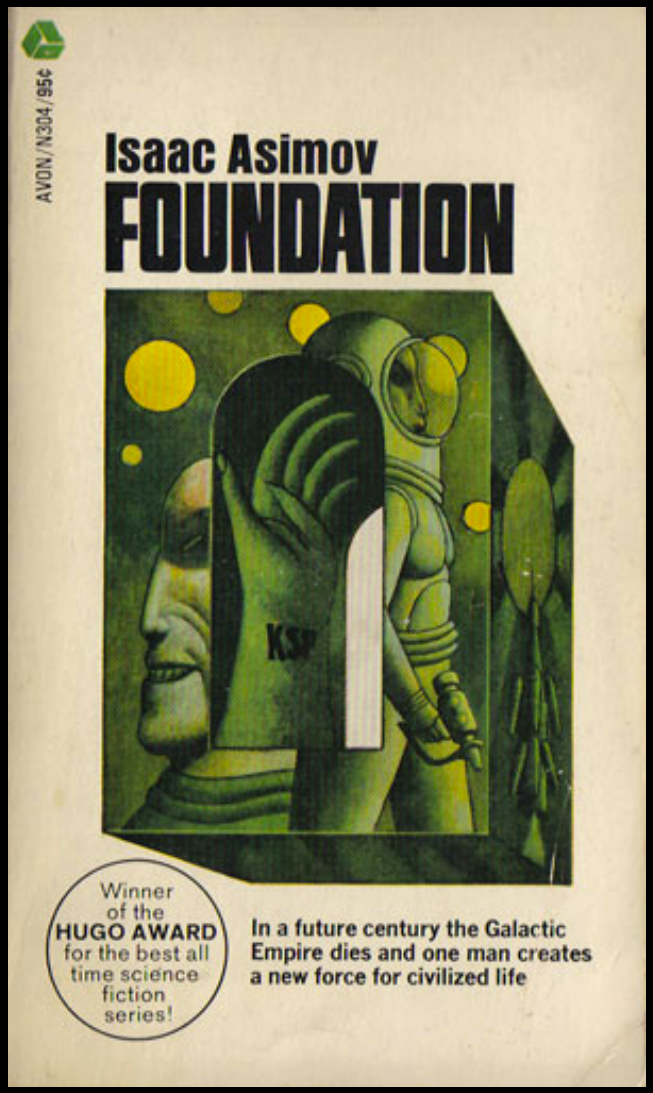
Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

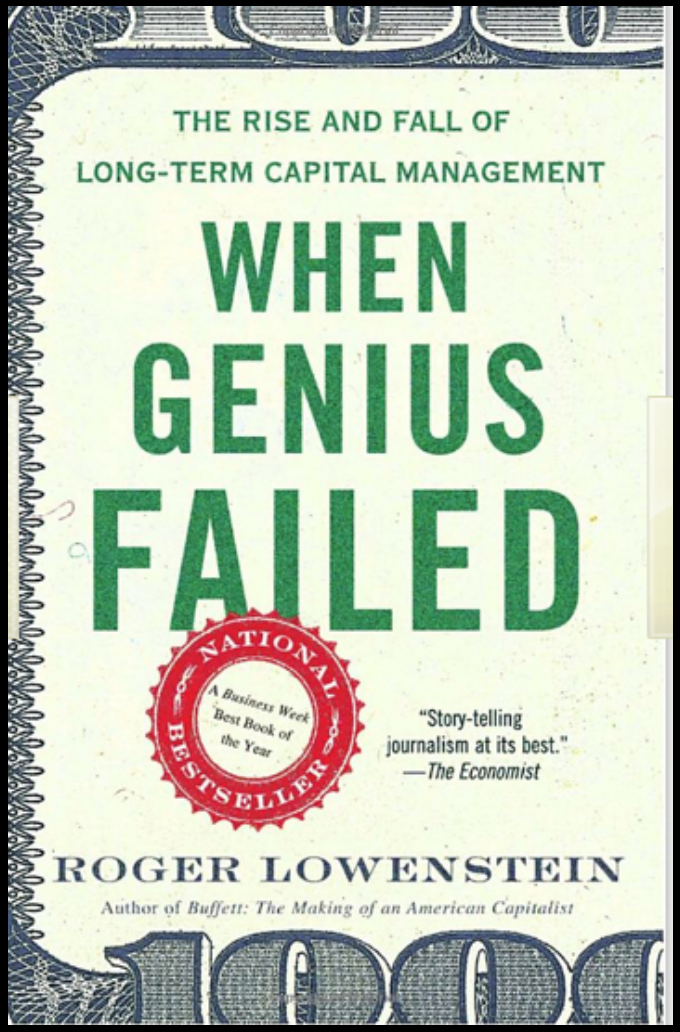
By ASHLEE VANCE
Published: January 6, 2009

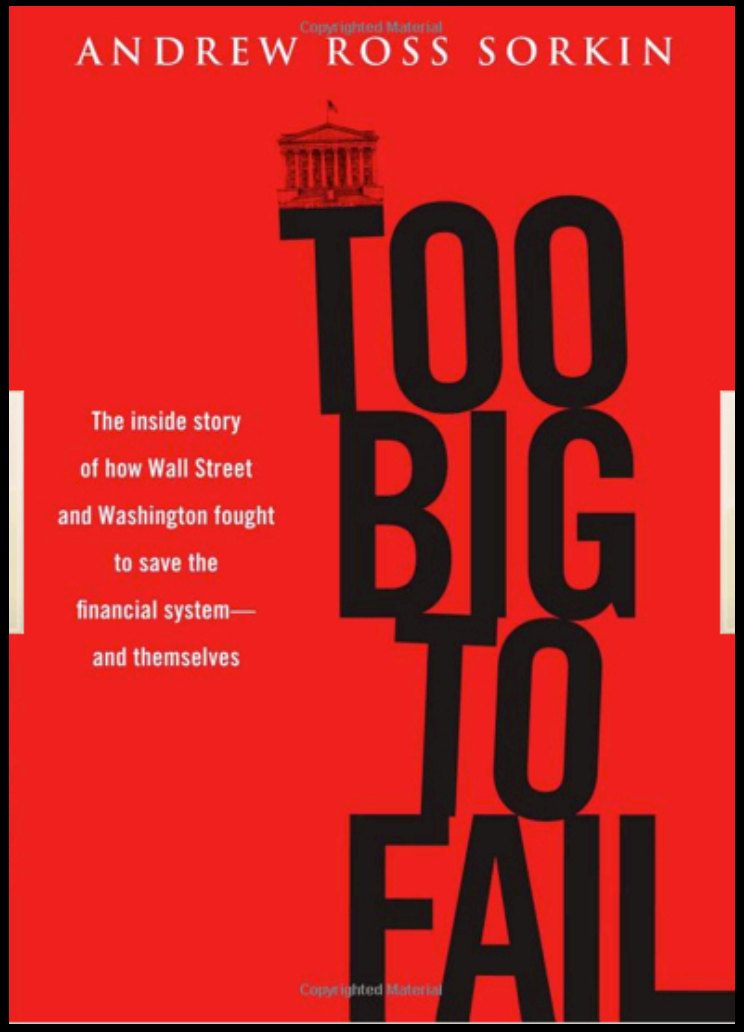
A New Kind of Empirical Science

Predict everything?



Sometimes you can’t


THE RISE AND FALL OF
LONG-TERM CAPITAL MANAGEMENT
WHEN GENIUS FAILED
NATIONAL BESTSELLER
A Business Week Best Book of the Year
"Story-telling journalism at its best."
—The Economist
ROGER LOWENSTEIN
Author of Buffett: The Making of an American Capitalist


Copyrighted Material
ANDREW ROSS SORKIN
TOO BIG TO FAIL
The inside story of how Wall Street and Washington fought to save the financial system—and themselves
Copyrighted Material

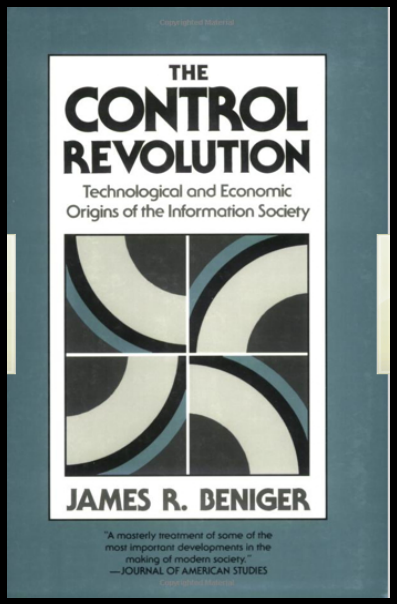
Isn’t experience qualitative?

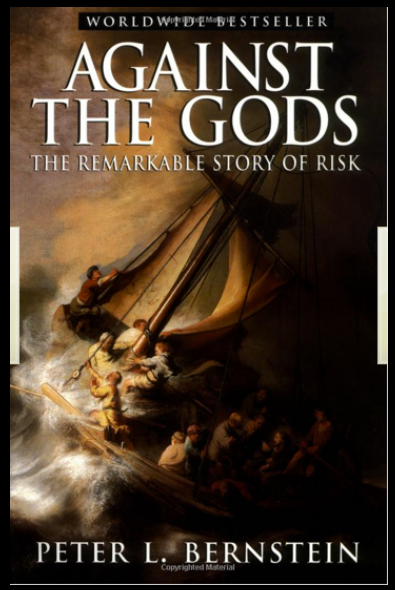
- Making qualitative judgments (heuristics) to apply methods
- Look for isomorphic problems that have been solved
- Variability & Applicability is often wide
- Technique is critical
- (Peer) Review

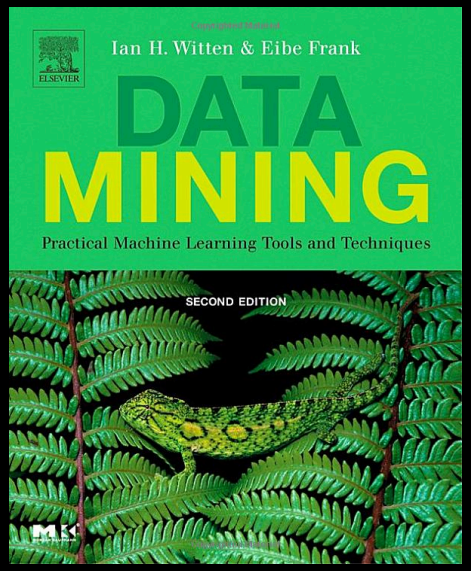
Don’t Reinvent the Wheel

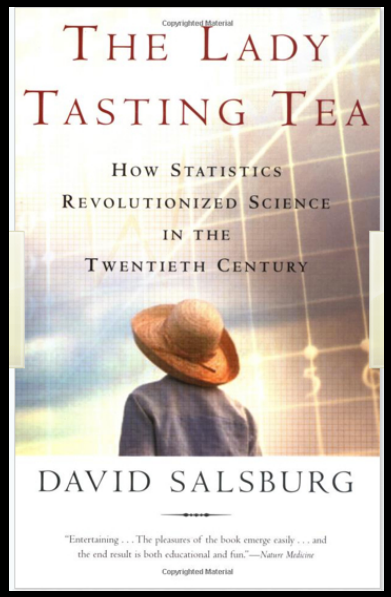
WWxD?

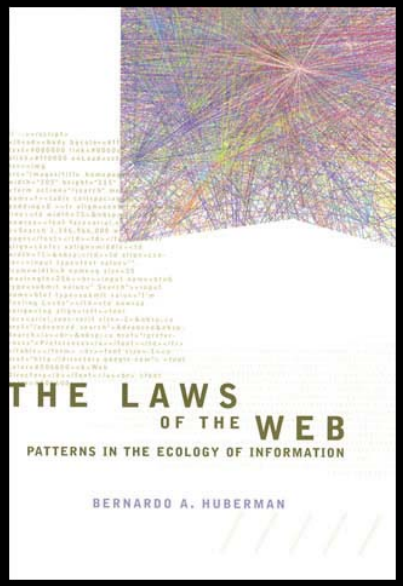
- Finance - WWQD?
- Statistics - WWBED?
- Computer Science – WW*D?
- Math – WW*D?

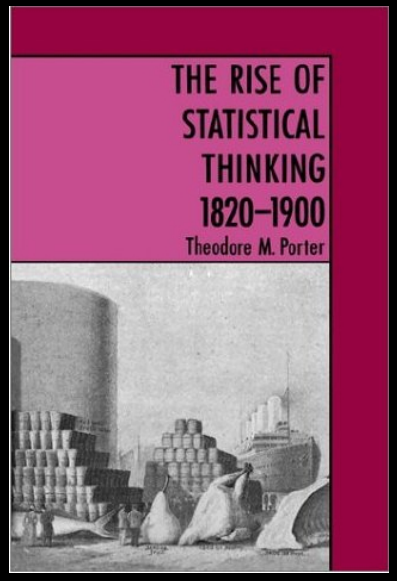

THE
CONTROL
REVOLUTION
Technological and Economic
Origins of the Information Society
JAMES R. BENIGER
"A masterly treatment of some of the
most important developments in the
making of modern society."
—JOURNAL OF AMERICAN STUDIES

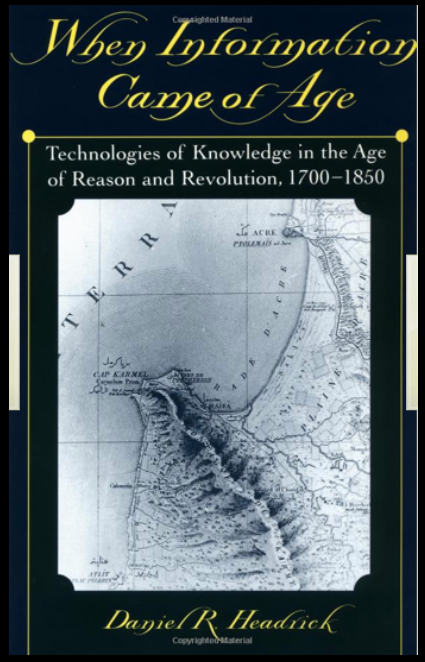

WORLDWIDE BESTSELLER
AGAINST
THE GODS
THE REMARKABLE STORY OF RISK
PETER L. BERNSTEIN

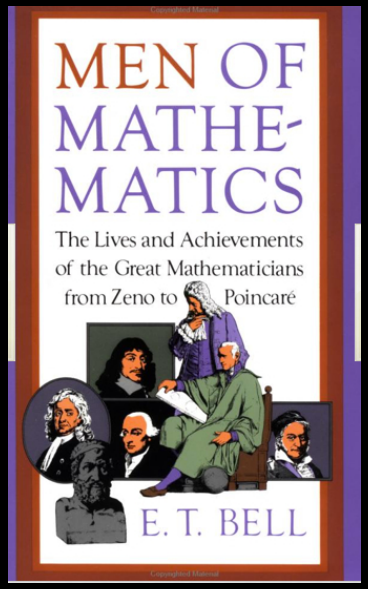

Ian H. Witten & Eibe Frank
DATA
MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION


THE LADY
TASTING TEA
HOW STATISTICS
REVOLUTIONIZED SCIENCE
IN THE
TWENTIETH CENTURY
DAVID SALSBURG
"Entertaining . . . The pleasures of the book emerge easily . . . and
the end result is both educational and fun."—Nature Medicine


THE LAWS
OF THE WEB
PATTERNS IN THE ECOLOGY OF INFORMATION
BERNARDO A. HUBERMAN


THE RISE OF
STATISTICAL
THINKING
1820–1900
Theodore M. Porter


When Information
Came of Age
Technologies of Knowledge in the Age
of Reason and Revolution, 1700–1850
TERRA
Daniel R. Headrick


MEN OF
MATHE-
MATICS
The Lives and Achievements
of the Great Mathematicians
from Zeno to Poincaré
E. T. BELL

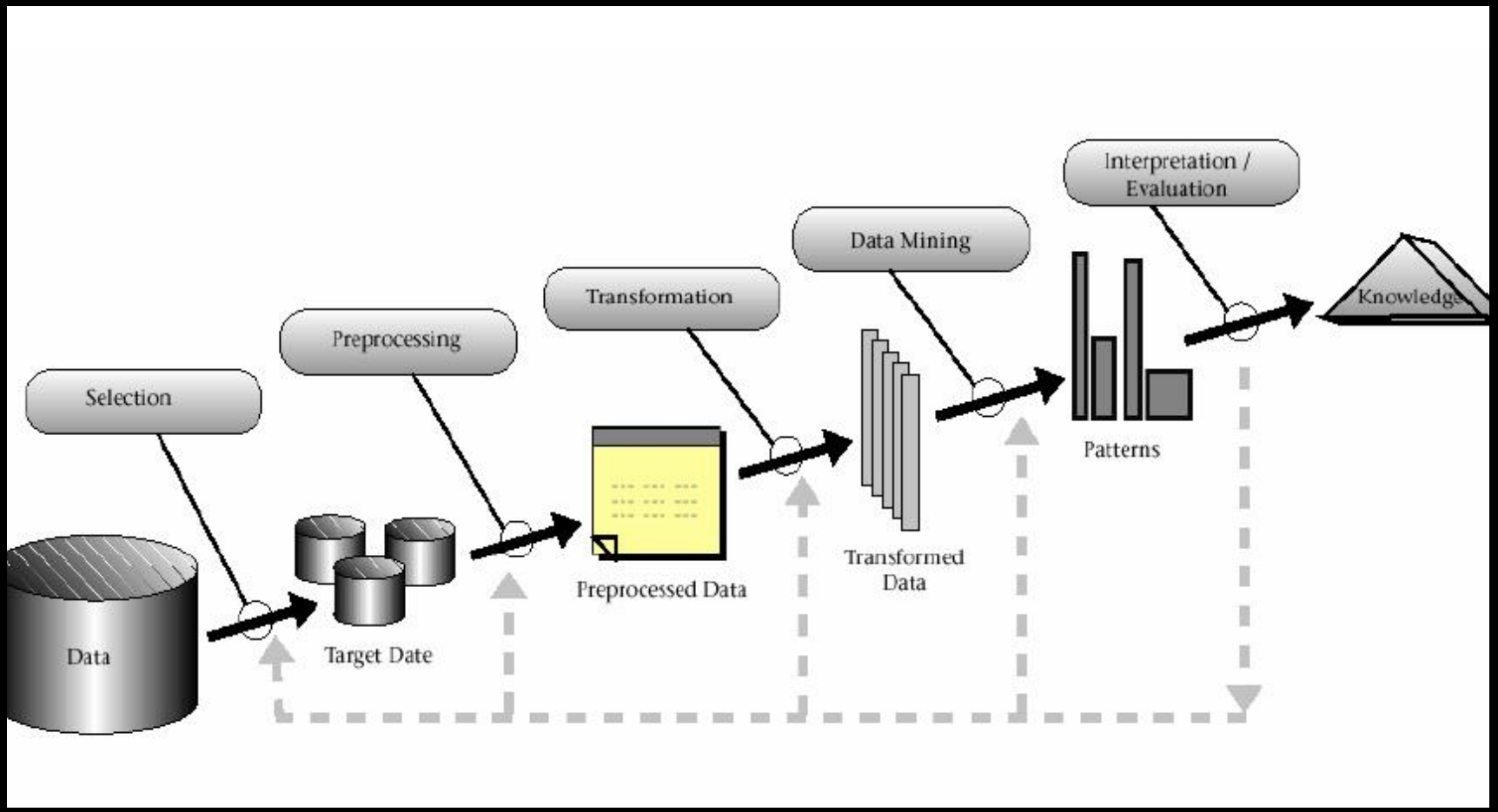
Solving Data Science Problems

- **Isomorphic** Problems
- Apply theory from other fields to solve problems
- Use these modeling techniques and transform them in sync with computational power

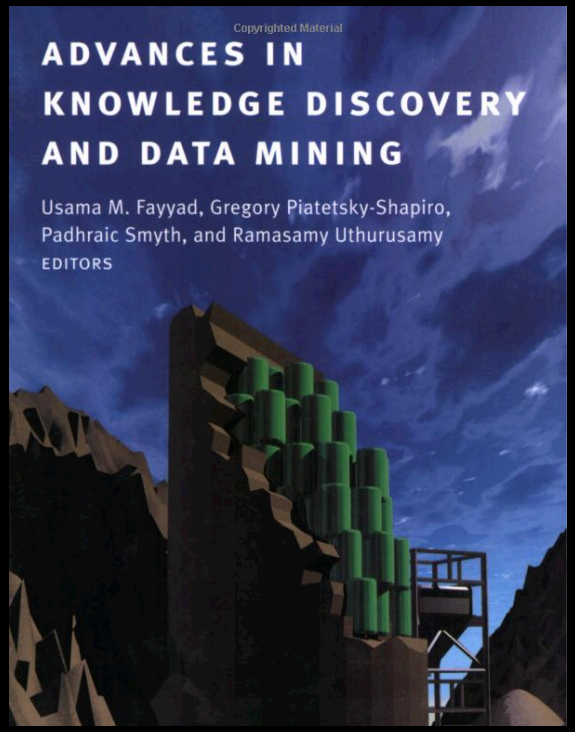
Data Science History

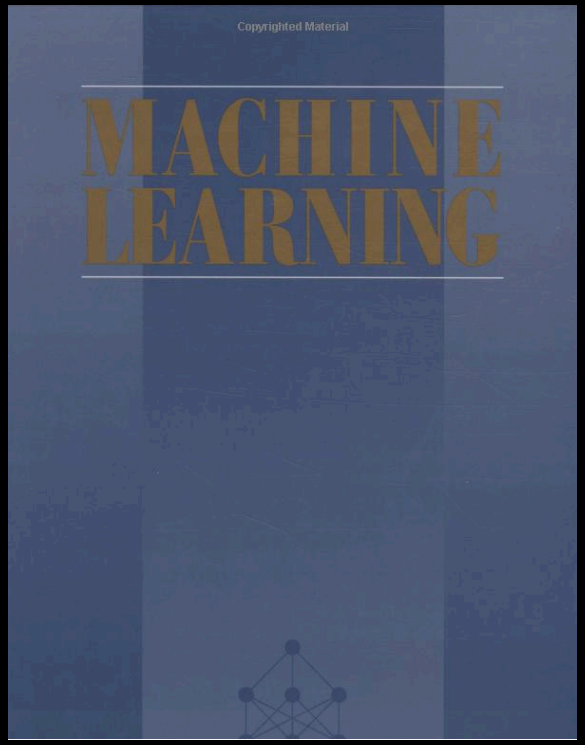
- Cryptography
- Knowledge Discovery in Databases
  - At least 30 years of Research in KDD
  - Superb Tools & Techniques
- Even more research from isomorphic solutions

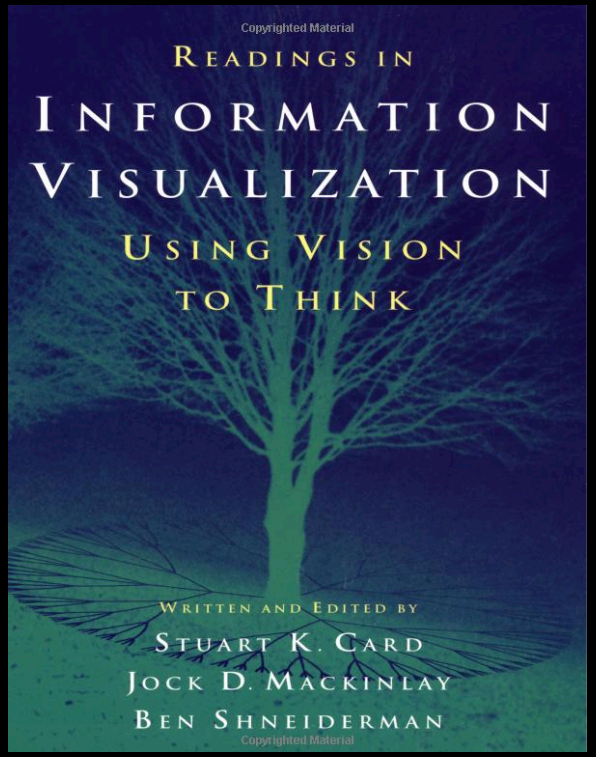
The KDD Process



My KDD Bootstrap







KDD Work

- Frequency Analysis (Statistics)
- Behavioral Data Mining
- Log Analysis > Transaction Log Analysis
- Recommender Systems & Collaborative Filtering
- Financial Analysis (Behavioral, Hybrid Time Series)
- Advertising Efficacy
- Traffic Analysis (~)
- Semantics

WebKDD Example

- Pilot = client app logs of Web use
- Full = network logs of Web use
  - Longer study period, lots more users
  - Many orders of magnitude of data to analyze means subtle patterns may be discovered
  - Substantive evidence of patterns of behavior
- Larger than all previous studies of Web use combined… 3000% more

That’s a big twinkie

My Ongoing Projects

- **Tapstream**
- **Twitter Text Analysis**
  - Recommender System
- **Behavioral Modeling & User Studies**
  - Whales, Use Frequency, Popularity

Everything is Better

- Tools
- Methods
- People
- Groups

Where to Look

- Academic Work
  - GoogleScholar or Microsoft Academic Search
  - Patents
  - Classes
    - Coursera, CodeAcademy, Cloudera…
  - Conferences
    - SIGIR, SIGKDD, CIKM, WWW Conf
  - Meetups


WINTER
IS
COMING

Vancouver Data Science Goals

- **What can we do *in person* that is optimal?**
  - Jobs
  - *Interactive* Tool Demos
  - Paper Reviews *and Walkthroughs*
  - People
    - @reiver @DataScholars @josh_wills
  - Group Problem Solving

Thanks for Listening!

Don Turnbull, PhD

@donturn

#datascienceyvr

donturn@gmail.com