Affecting Change with Data

A Primer on Hype and Reality in Data Mining

Patrick Hall

H_2O.ai

July 26th, 2018

Agenda

Disambiguation

Hype and Reality

Short History

Affecting Change
- Past
- Present
- Future

Precautions

Preface

"We are drowning in information and starving for knowledge."

— Rutherford Roger, *Librarian (1915 - 2015)*

"Information is not knowledge. Knowledge is not wisdom. Wisdom is not truth. Truth is not beauty."

— Frank Zappa, *Musician (1940 - 1993)*

"All models are wrong, but some are useful."

— George Box, *Statistician (1919 – 2013)*

Buzzword Disambiguation

Well-defined	Well-understood	Fake News
Computer Science	Analytics, Data Mining, Predictive Modeling	Big Data
• Artificial Intelligence (AI)	Deep Learning	Data "Science"
• Strong AI		
• Weak AI		
• Machine Learning		
Reporting		
Statistics		

Hype and Reality

Image credit: Shutterstock — KimSongsak

- Data mining has been used for decades in traditional industries to improve financial margins.
- Big tech. made a windfall by overselling new but minor conveniences often at the cost of our personal privacy.
- As usual, we *may* be on the brink of a nonlinear leap in technology.

A Very Short History of Machine Learning

1962
J.W. Tukey: "The Future of Data Analysis"

1958
Rosenblatt's Perceptron

1965
DENDRAL expert system program begins

1974
"Data Science" used freely in Naur's *Concise Survey of Computer Methods*

1974 - 1980
First AI Winter
Numerous failures in the face of unrealistic expectations cause US and UK agencies to cut funding.

1980
DEC uses XCON expert system to recommend mainframe configurations

1987 - 1993
Second AI Winter
Commercial attention turns away from specialized systems to the PC.

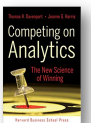
1989
Advent of contemporary FICO scores

Lecun, Bengio and others propose basic deep learning architectures

1993
Match.com founded: Weak AI gets *really* personal.

2001
Statistical learning flourishes at Stanford & Berkeley

2003
Amazon announces mature product recommender
Available at amazon



2005
Davenport et al: *Competing on Analytics*

2011
IBM Watson becomes world's best Jeopardy player

2008
Facebook & Linkedin collaborate on data science functions

2017
Google's AlphaGo becomes world's best Go player; Google & others announce legitimate efforts in automatic machine learning

2018
EU GDPR takes effect

Mass-layoffs in IBM Watson Health division; Chinese government and industry plans to invest double US funding in AI

Affecting Change with Data

Data mining is a blunt tool for revisiting the past, understanding the present, and seeing into the future.

Image credits: Flickr — chengyee, Andy Castro, Roo Reynolds

Learn from the Past

- Often easier to conduct as an individual as tangible results include papers, data visualizations, stories, lessons, and morals.
- Use common sense to learn about human or financial gain or loss.
- Look for trends, groups, and outliers in data.

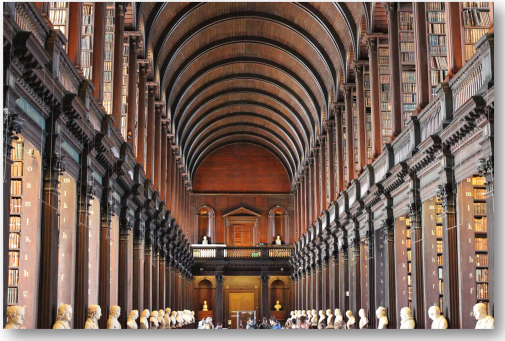


Image credit: Flickr — David P. Whelan

Understand the Present

Image credit: Flickr — Sérgio Bernardino

Typically requires organizational information technology (IT) support as tangible results include real-time reports and alerts created from data and IT systems.

Prepare for the Future

- Usually requires advanced IT support.
- Tangible results are typically decisions about customers, patients, students, equipment, or other valuable entities made by sophisticated computer systems.
- Common commercially viable applications include:
  - Quantifying future risk
  - Personalized promotions
  - Predicting churn
  - Recommending products or content

Take Necessary Precautions

Image credit: Flickr — Thomas Hawk

- Data can be biased, incomplete, or just wrong.
- Results can be ignored or used politically.
- Avoid:
  - Confirmation bias
  - Testing multiple hypotheses
  - Unintended discrimination
  - Privacy violations
  - Unmanageable complexity

Stay in Touch

linkedin.com/in/jpatrickhall

@jpatrickhall