

Analytics

Ilya Kavalero

University of Maryland
Electrical and Computer Engineering

ilyak@umiacs.umd.edu

November 18, 2015

About me

- Previously at Artsy (backend engineer, prediction)
- PhD Student at EECE department at UMD, Signal Processing

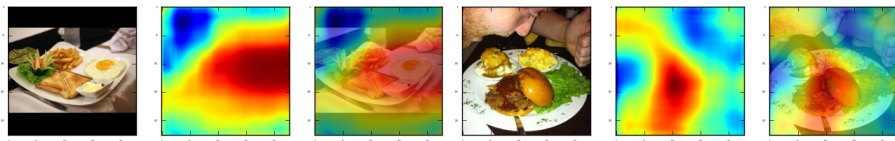


Figure: Egg and Burger detection

- Why be data driven?
- New kinds of problems introduced by learning
- Some simple examples of using typical business data

Data at work

- Data driven/Data-curious culture (SQL and graph friendly)
- Kaizen
- Why the "big data" hype?
 - Magic bullet image
 - abundance of fuel (**heuristics** become harder, **learning** becomes more effective) eg. Face recognition, 30TB = 1yr Duke heart center, 1 day Fb, 1 s CERN



- Extract new forms of value from information

How others use data: The job

- "Data scientist"
 - 2008 from LinkedIn Job posting by DJ Patil and Jeff Hammerbacher
 - 2012 Wikipedia Entry, HBR's [Sexiest job of 21st century](#)
 - Statistics, Data munging, Visualization

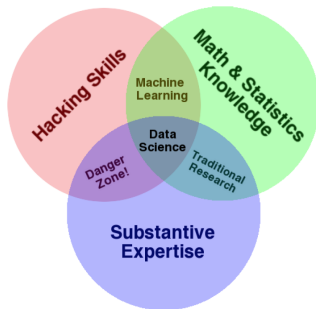


Figure: Drew Conway's diagram

How others use data: Learn a model

- Maximum Likelihood: Maximize probability of your data given a model
- iid assumption:

$$p(x_1, x_2, \dots, x_n | \Theta) = p(x_1 | \Theta) \times p(x_2 | \Theta) \times \dots \times p(x_n | \Theta)$$

$$\Theta^* = \operatorname{argmax}_{\Theta} \prod_{i=1}^N p(x_i | \Theta) = \operatorname{argmax}_{\Theta} \sum_{i=1}^N \log p(x_i | \Theta)$$



Dangers: Machine learning disrupts software engineering

- Abstraction = leverage works of others
- Engineer complex artifacts (Airplane: 3M parts, Debian: 0.4B LOC)
- We expect to use programs much like we use math theorems (design contracts)
- Abstraction leaks limit complexity (Snowplow)
 - Sorting has few assumptions on input (only obvious failure is performance)
 - Using a learned model has huge assumptions on input (understanding these is the hard part)

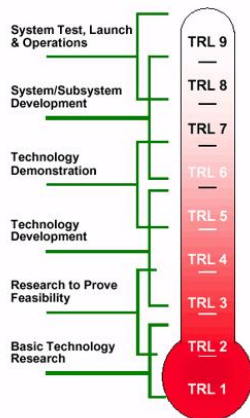


Figure: NASA
Technology Readiness
Levels

In short: Don't go overboard

- "Great products are a convergence of the right set of technologies" - Jon Rubinstein (EE behind the first iPod)
- Do simple things with few assumptions for less surprises
- Track everything, ask questions later

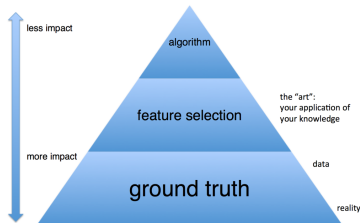


Figure: Chart inspired by a slide from Alex Pinto's talk, "[Secure Because Math: A Deep-Dive on ML-Based Monitoring](#)"


- Iterative/Continuous improvements in the startup spirit


Conditional Probability & Baye's rule


- $p(X|Y) = \frac{p(X,Y)}{p(Y)}$
- $p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}$
- $\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$
- We'll use this in the following example: Subscription tiers on BriskIT, Free, \$Monthly, and \$\$Yearly. Say we want to run a promotion on the yearly subscription, we decide to A/B Test 2 methods to convert users


BriskIT Subtle Banner


Click [here](#) to upgrade to a Yearly Subscription




 SOLVE

 SOLVING 3


 SOLVED

 Log Out

 Current sessions


request

Hardware

 m


test


Hardware

 hannah

someone taking

Hardware

 jadami10

 H

BriskIT Pop-up

The screenshot displays the BriskIT web application interface. On the left is a dark sidebar with the 'BriskIT' logo at the top. Below the logo are three menu items: 'SOLVE' with a puzzle piece icon, 'SOLVING' with a speech bubble icon and a red badge containing the number '3', and 'SOLVED' with a checkmark icon. At the bottom of the sidebar is a 'Log Out' button with a right-pointing arrow icon. The main content area has a grey header with the text 'Current sessions' and a user profile icon labeled 'H'. Below the header, there are three session cards. Each card has a title, a 'Hardware' button, and a user icon with a name. The first card is titled 'request' and is associated with user 'm'. The second card is titled 'test' and is associated with user 'hannah'. The third card is titled 'someone taking' and is associated with user 'jadami10'. A large, light blue pop-up window is centered over the 'request' session card. The pop-up has a close button (an 'X' icon) in the top right corner. Inside the pop-up, the text reads: 'Before continuing, please consider:' followed by a white rectangular button that says 'Upgrade to Yearly'.

BriskIT

Current sessions

SOLVE

SOLVING 3

SOLVED

Log Out

request

Hardware

test

Hardware

someone taking

Hardware

Before continuing, please consider:

Upgrade to Yearly

m

hannah

jadami10

Simpson's Paradox: 1

- Subscription tiers on BriskIT, Free, \$Monthly, and \$\$Yearly. Say we want to run a promotion on the yearly subscription, we get the results of an A/B Test on 2 methods to convert users:

	Pop Up	Subtle Banner
Users (Converted/Total)	78% (273/350)	83% (289/350)

- Looks like subtle banner wins. But with more features ...

Simpson's Paradox: 2

- With more features:

	Pop Up	Subtle Banner
Monthly Subscribers	93% (81/87)	87% (234/270)
Free Users	73% (192/263)	69% (55/80)
Users (Converted/Total)	78% (273/350)	83% (289/350)

- The less effective method (Subtle Banner) is tested on the easier cases more often \nrightarrow Subtle Banner is more effective

Simpson's Paradox: 3

- Okay it's intuitive, our bad pitch (subtle banner) got the easy targets (already subscribers), but let's express this with a probability.

	Pop Up	Subtle Banner
Monthly Subscribers	93% (81/87)	87% (234/270)
Free Users	73% (192/263)	69% (55/80)
Users (Converted/Total)	78% (273/350)	83% (289/350)

- $p(\text{Monthly}) = \frac{87+270}{87+270+263+80} = 0.51$

Bizsouk: Material sourcing platform for industries
Visualize last 1000 trades

Demo

Zipf's Law

- Total income a function of heaviest hitter's income A :
$$\frac{A}{1} + \frac{A}{2} + \dots + \frac{A}{n}$$
- Upward convexity is a condition of surfeit (too few are getting too much, diversification will follow), downward convexity is a condition of deficiency
- Long tail, can be similar to log normal

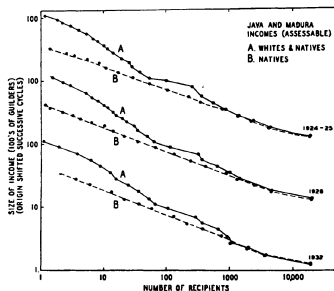



Fig. 11-9. Incomes: Java and Madura (Netherlands East Indies).

Figure: From Zipf's Principle of least effort. Power changed hands by 1949.

Artrank

Is Frank Stella undervalued according to Artsy Auction Prices?

Demo



[ARCHIVE](#)
[POSTS](#)
[FAQ](#)
[SUBSCRIBE](#)

Collect smarter.

Artrank™ gives you an unprecedented data-driven advantage in art collecting.

BUY UNDER \$10,000	BUY UNDER \$30,000	BUY UNDER \$100,000	SELL / PEAKING	EARLY BLUE CHIP	UNDERVALUED BLUE CHIP
1. Dora Budor	1. Calvin Marcus	1. Harold Ancart	1. Justin Adian	1. Jonas Wood	1. Helen Frankenthaler
2. Haley Mellin	2. Ian Cheng	2. Avery Singer	2. Torrey Thornton	2. Cory Arcangel	2. Frank Stella
3. Przemek Pynaszek	3. Sanyia Kantarovskiy	3. Aaron Garber-Malkovska	3. Mike Bouchet	3. Joe Bradley	3. Simon Hantai
4. Emily Mae Smith	4. Borna Sammak	4. Math Bass	4. Alex Isreal	4. Jordan Wolfson	4. Sam Francis
5. Paul Kremer	5. Tabor Robak	5. Katherine Bernhardt	5. Jeff Elrod	5. Ella Kruglyanskaya	5. Larry Rivers
6. Gerasimos Floratos	6. Artie Vlerkant	6. Petra Cortright	6. Wyatt Kahn	6. Oscar Murillo	6. Bernar Venet
7. Michael Rey	7. Max Hooper Schneider	7. Jamian Juliano-Villani	7. Secundino Hernández	7. Danh Vø	7. Jules Olitski
8. Katja Novitskova	8. Elaine Cameron-Weir	8. Ryan Gander		8. Nicole Eisenman	8. Robert Motherwell
9. David Rappeneau	9. Jonathan Gardner			9. Eddie Peake	9. Sol LeWitt

[Browse Index](#)

Artrank™ index last update: Q3/2015. Next update: Q4/2015 (1 December 2015!)

GitRank: Featured OS Libraries

Find undervalued game engines

Demo

[GitRank](#)

Login

Welcome to GitRank

A platform where you can easily feedback and discover open source projects

twbs/bootstrap ★ 89137
Last Updated on Nov 18, 2015
The most popular HTML, CSS, and JavaScript framework for developing responsive, mobile first projects on the web.
+ Give Feedback

vhf/free-programming-books ★ 45709
Last Updated on Nov 18, 2015
Freely available programming books
+ Give Feedback

angular/angular.js ★ 44340
Last Updated on Nov 18, 2015
HTML enhanced for web apps
+ Give Feedback

mbostock/d3 ★ 43521
Last Updated on Nov 18, 2015
A JavaScript visualization library for HTML and SVG.

nodejs/node-v0.x-archive ★ 37652
Last Updated on Nov 18, 2015
Moved to <https://github.com/nodejs/node>

jquery/jquery ★ 36780
Last Updated on Nov 18, 2015
jQuery JavaScript Library

Ilya Kavalero (UMD)

Cornell CS5356 Fall 2015

November 18, 2015

18 / 21

Conclusion

- Data driven decision making can lead to quicker/more continuous improvements
- It's easy to fall into doing research, keep it simple
- Some examples similar to what I saw while working

Thank You!

References & Further reading

- Spotify Labs puzzles
- Dataquest visualization
- Dive into Machine Learning
- Doing data science: Straight talk from the frontline
- Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology