Analytics

Ilya Kavalerov

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

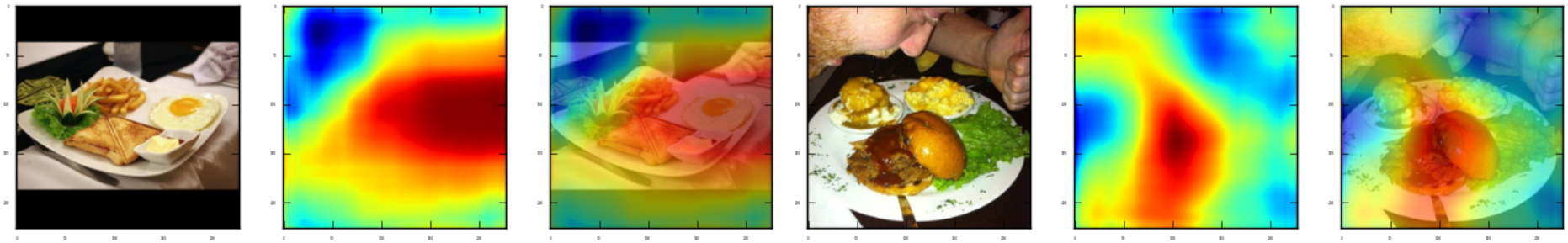


Figure: Egg and Burger detection

Overview

- 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

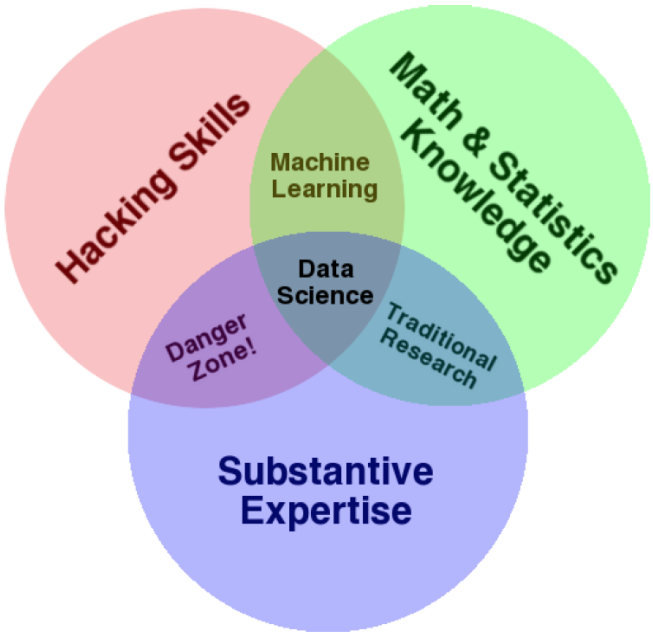


Figure: Drew Conway's diagram

How others use data: Learn a model

- Maximum Likelihood: Maximize probability of your data given a model
- iid assumption:
  $p(x_1, x_2, \cdots, x_n|\Theta) = p(x_1|\Theta) \times p(x_2|\Theta) \times \cdots \times p(x_n|\Theta)$

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

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)

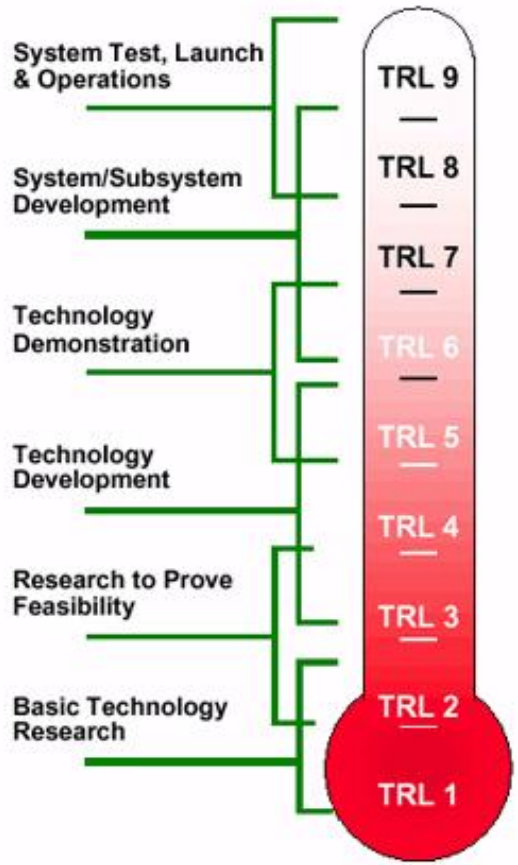


Figure: NASA Technology Readiness Levels

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

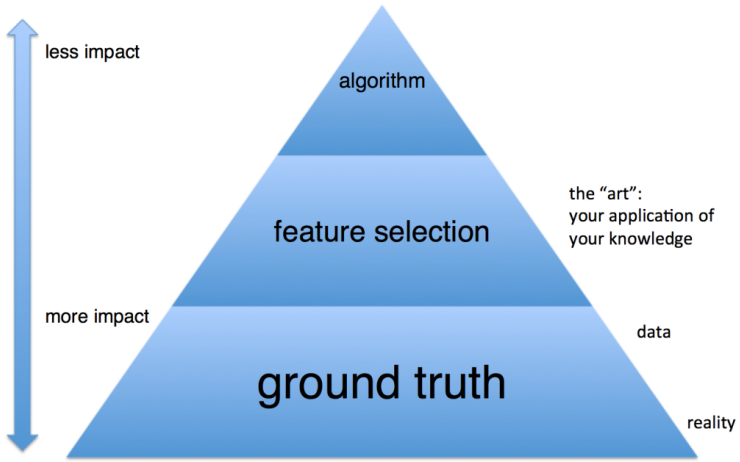


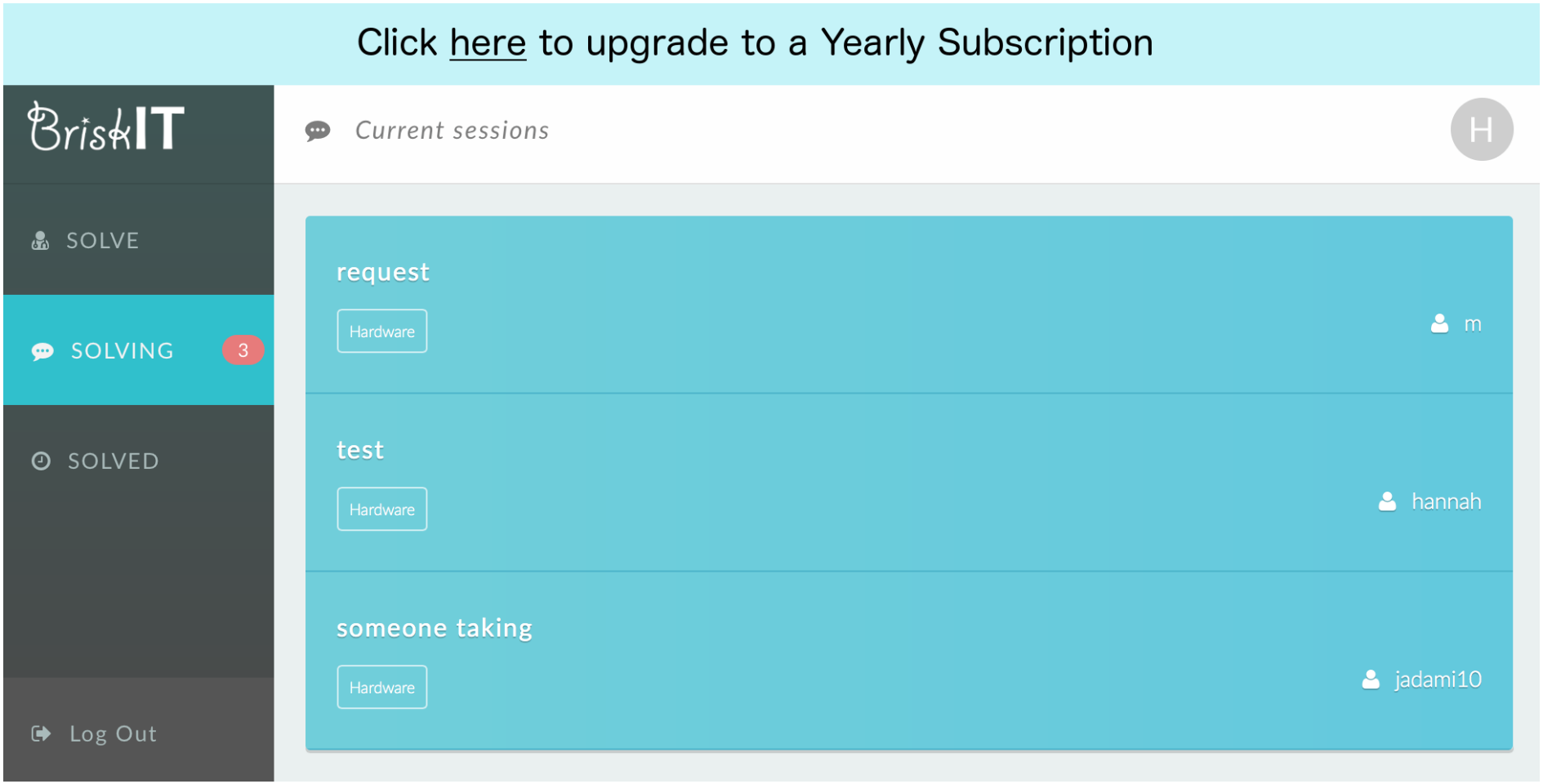
Figure: Chart inspired by a slide from Alex Pinto's talk, "Secure Because Math: A Deep-Dive on ML-Based Monitoring"

- Iterative/Continuous improvements in the startup spirit

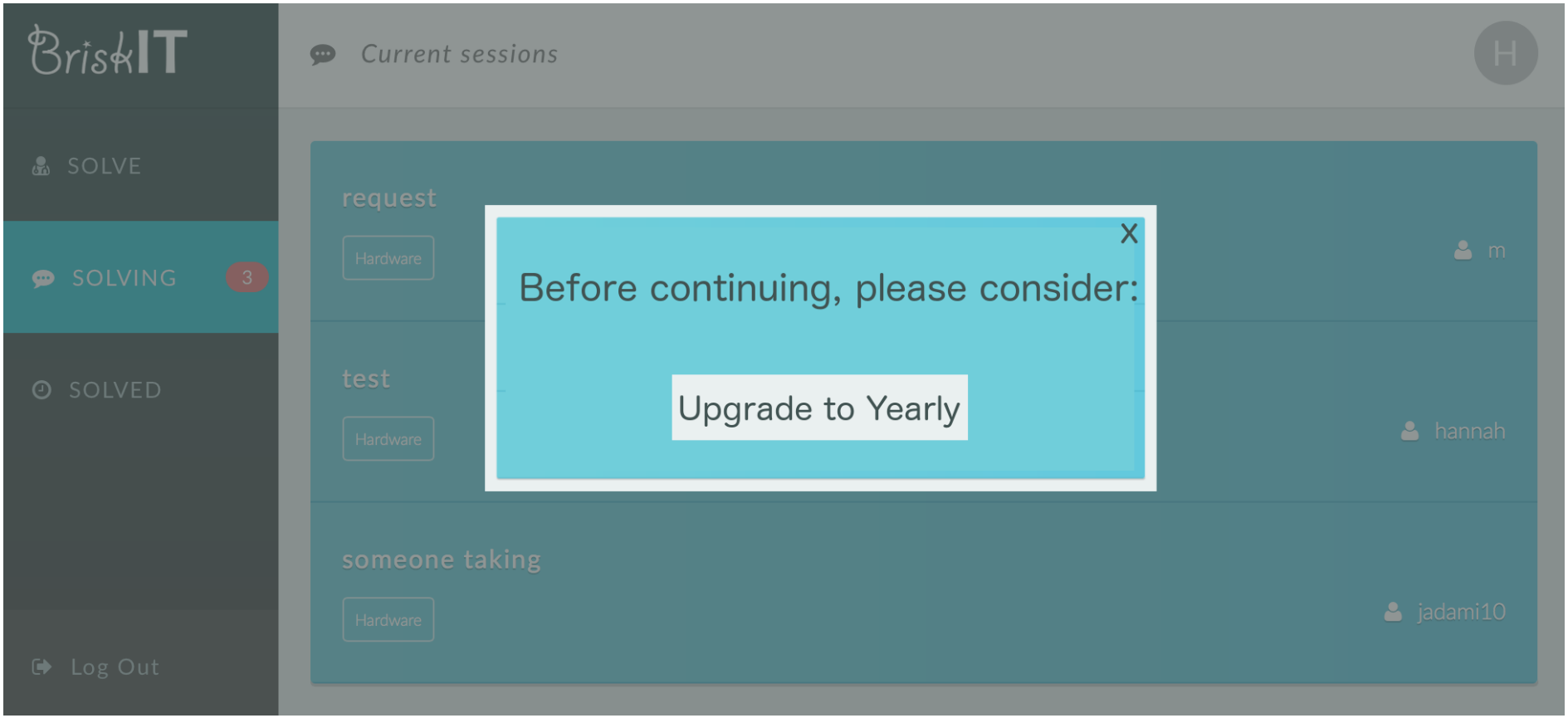
Conditional Probability & Baye's rule

- $p(X|Y) = \frac{p(X,Y)}{p(Y)}$
- $p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}$
- 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



BriskIT Pop-up



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 $\not\Rightarrow$ 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: $\frac{A}{1} + \frac{A}{2} + \cdots + \frac{A}{n}$
- 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

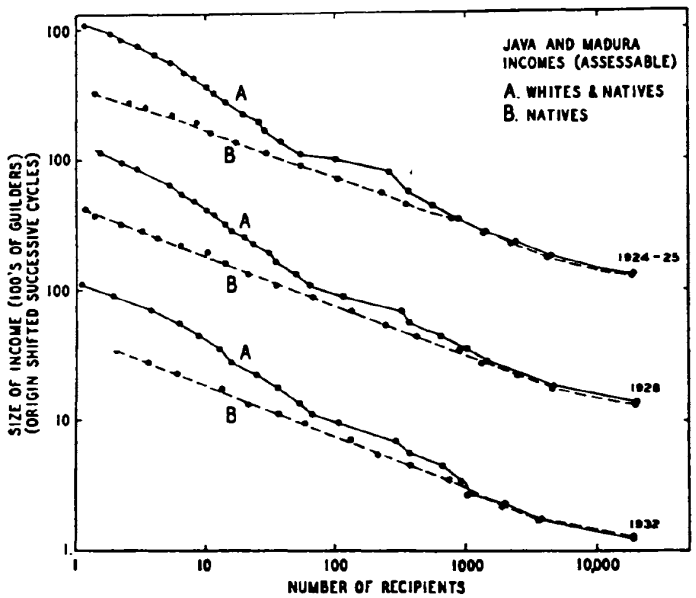


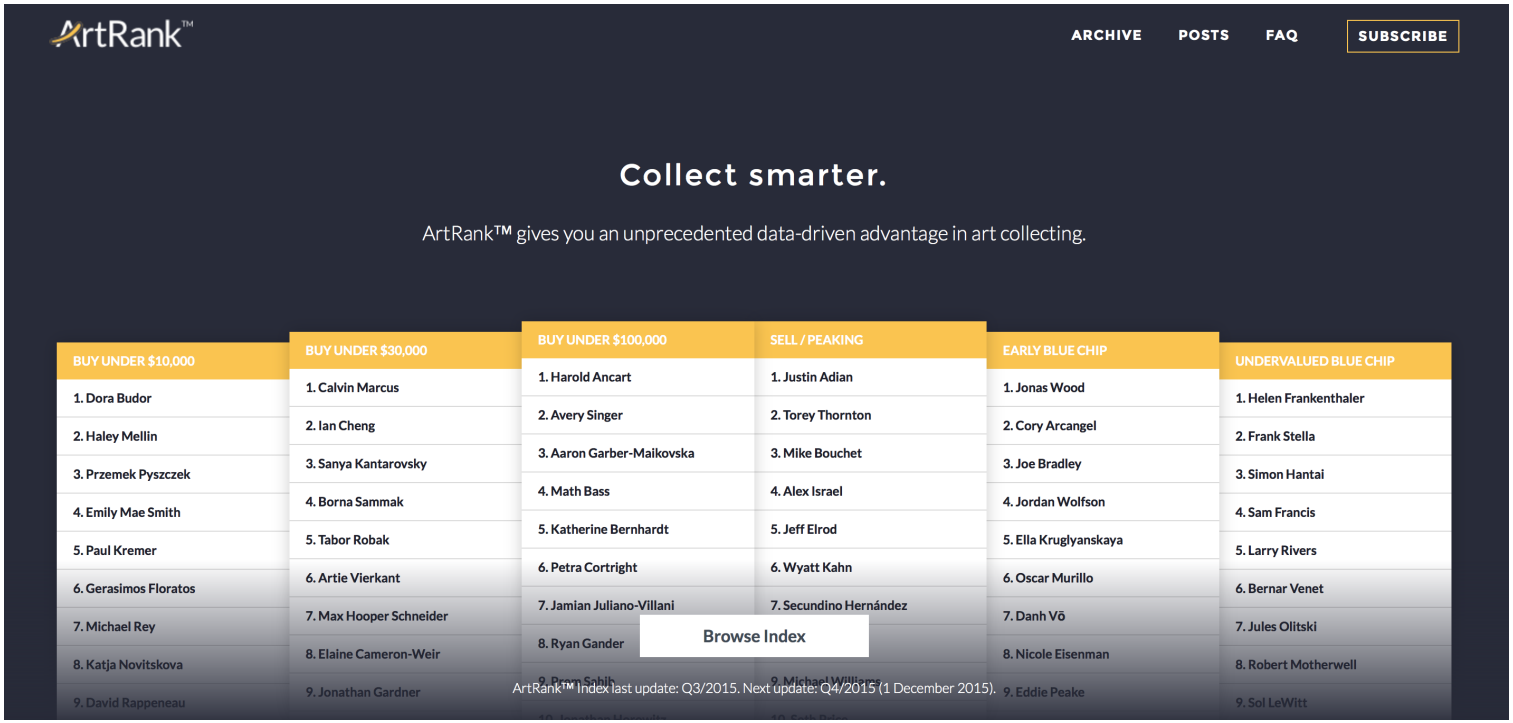
Fig. 11–9. Incomes: Java and Madura (Netherlands East Indies).

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

Artrank

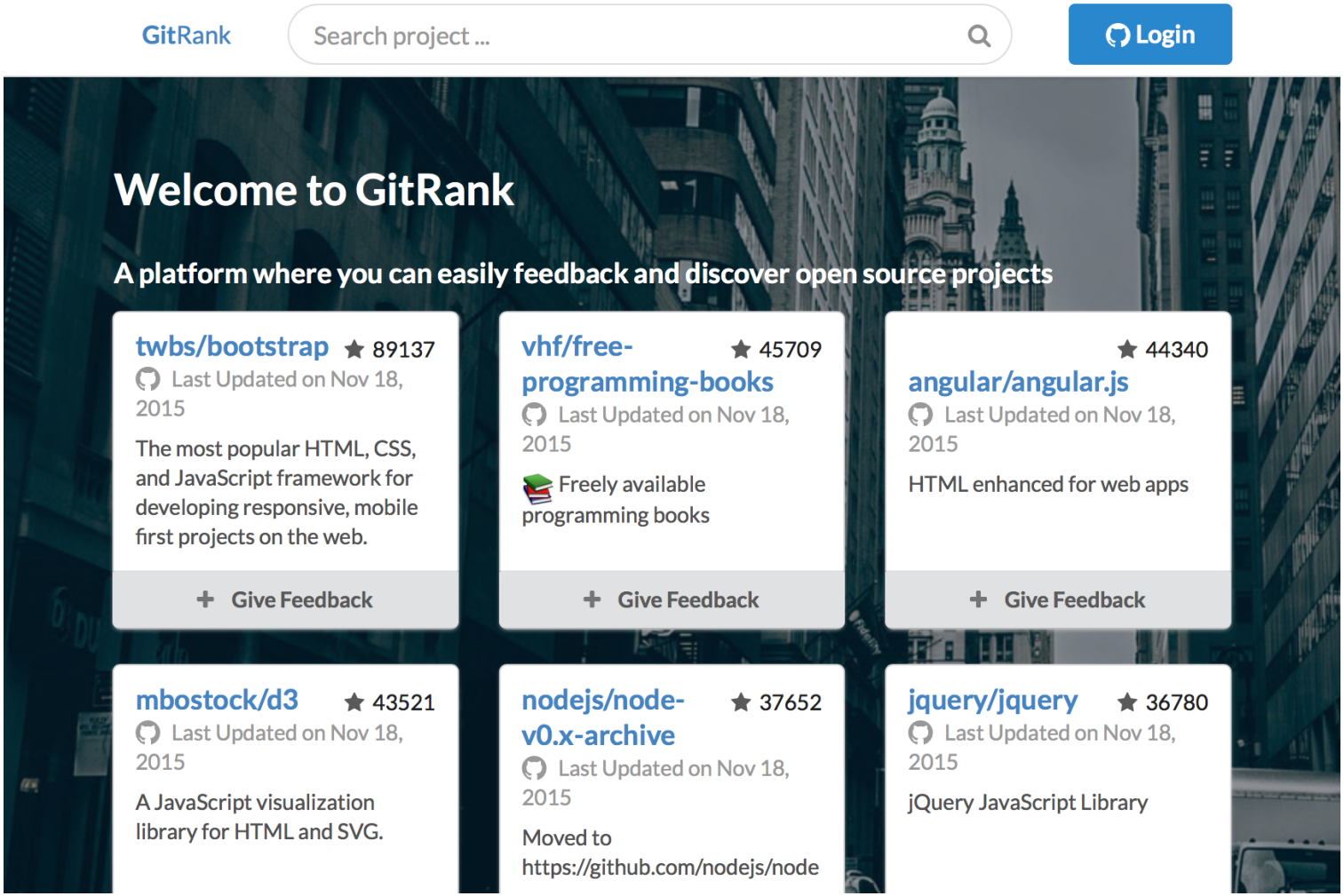
Is Frank Stella undervalued according to Artsy Auction Prices?

Demo



Gitrank: Featured OS Libraries
Find undervalued game engines

Demo



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