Data science tips

Lecture outline

- Some advice for avoiding common pitfalls

User forum example

Hey,

I got everything in order and got my features ready and the test and validate file ready and have attached the same. The execution stops when it comes to the assert line

```
double factor = 1.0 - (edge.data().x_ij *
vertex.sigma / other.data().sigma)*w(product);
assert(factor > 0);
```

The assert fails and I can't wrap my head around why it fails. I have made minimal change to your code and haven't tampered how the values are assigned to the graph. I have started understanding the graphlab api so that I can use it to improve the efficiency. … somehow I have a gut feeling there is something wrong there.

Data fields – in this example

1. The type of ad, (text, img, video) (3 features)
2. The platform of the user (Android, WinPh, J2ME device, others) (4 features)
3. The days of the week (Mon - Sun) (7 features)
4. The country (190 features)
5. The operator/Carrier (T-Mob, Verizon, ...) (Possible 150 features)
6. Placement of the app (2 features)
7. The ad source (possible 30 features)

Rule #1: inspect your data

- -1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1
- -1 9:1 2:1 10:1 4:1 5:1 11:1 12:1 8:1
- -1 9:1 2:1 10:1 4:1 5:1 11:1 13:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 12:1 8:1
- -1 9:1 2:1 14:1 4:1 5:1 11:1 15:1 8:1
- -1 1:1 2:1 3:1 4:1 5:1 11:1 16:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 12:1 8:1
- -1 9:1 2:1 10:1 4:1 5:1 11:1 15:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 13:1 8:1
- -1 17:1 2:1 14:1 4:1 5:1 11:1 15:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 15:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 15:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 15:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 13:1 8:1
- -1 1:1 2:1 14:1 4:1 5:1 6:1 7:1 8:1
- -1 1:1 2:1 14:1 4:1 5:1 11:1 16:1 8:1
- -1 9:1 2:1 18:1 4:1 5:1 11:1 12:1 8:1
- 1 9:1 2:1 10:1 4:1 5:1 11:1 15:1 8:1
- -1 1:1 2:1 10:1 4:1 5:1 11:1 16:1 8:1
- -1 9:1 2:1 3:1 4:1 5:1 11:1 15:1 8:1
- -1 9:1 2:1 3:1 4:1 5:1 11:1 12:1 8:1

- Given 100K input lines. How many unique lines?
  - 352 lines

- Does the target make sense?

```
1 1:1 2:1 10:1 4:1 5:1 11:1 16:1 8:1
-1 1:1 2:1 10:1 4:1 5:1 11:1 16:1 8:1
```

Opposite target values appear for the same features!

Rule #2 - GIGO

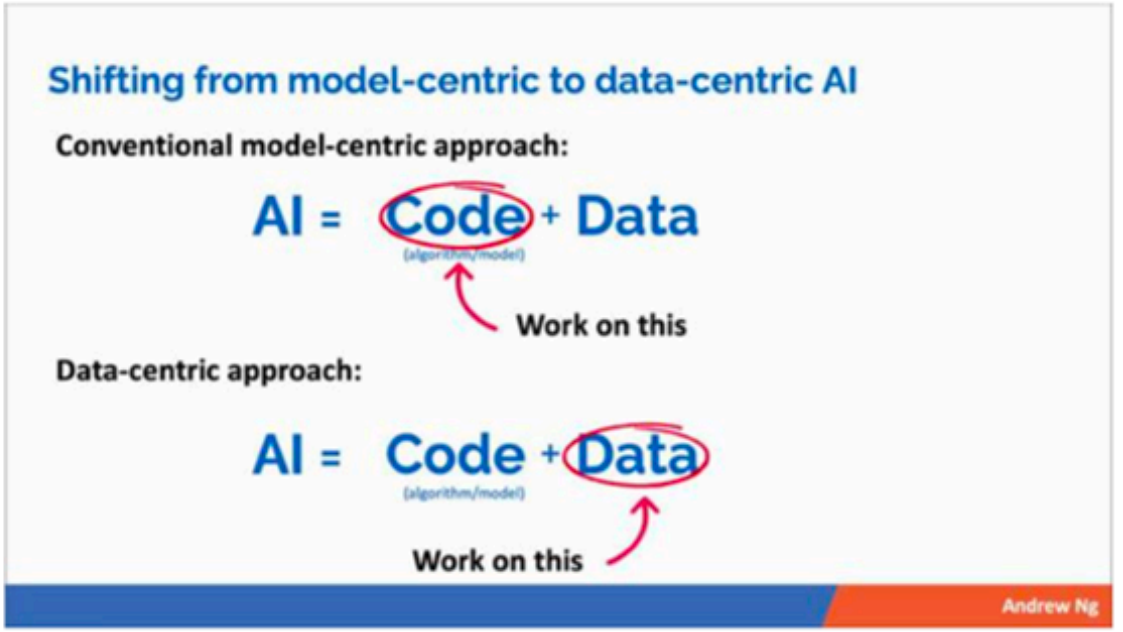
Garbage in, garbage out

From Wikipedia, the free encyclopedia

Garbage in, garbage out (**GIGO**) in the field of computer science or information and communications technology refers to the fact that computers, since they operate by logical processes, will unquestioningly process unintended, even nonsensical, input data ("garbage in") and produce undesired, often nonsensical, output ("garbage out").

The most useful ML methods will produce nothing if the data is useless..

Andrew Ng: Data Centric Approach



Rule #3 never assume the data is clean/ correct

- Example 1: airline on time dataset. Year 2008 has 7M flights. Target is the actual flight time.
  - 23500 flight times are missing
  - 9 negative flight times
  - 10 columns have no data whatsoever
  - Source & dest airports are the same
- Example 2: Oil well data. Target is cumulative oil production is 90 days
  - Number -666 is missing value!

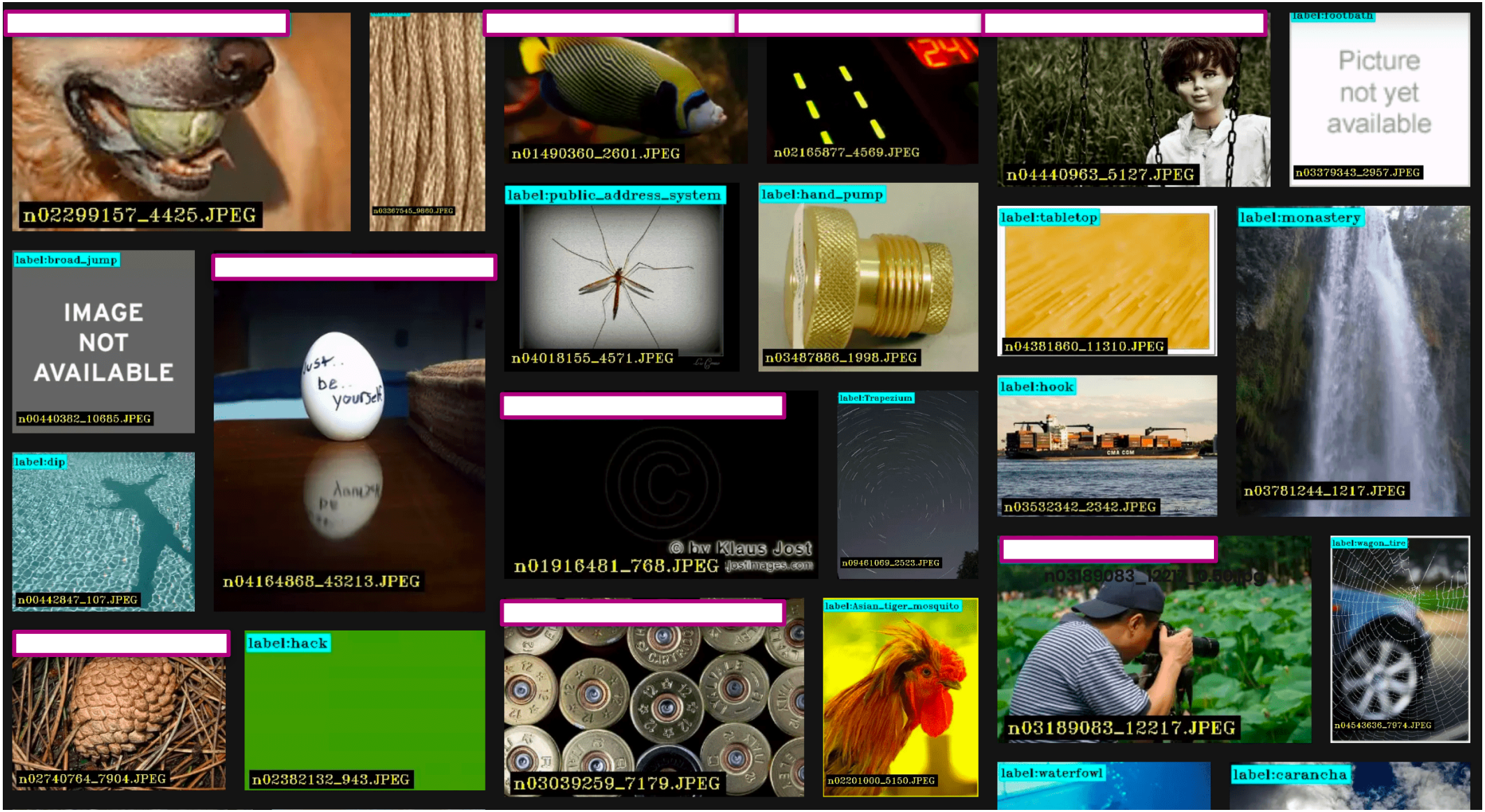
Rule #4 Humans are unpredictable

- Real data from someone's Facebook bio:
  - I'm 193 years old, a pilot, and my eyes blink sideways. I love filth and drink sludge. I haven't bathed in 153 years. All my teeth fell out 159 years ago. Brushing them wasn't healthy, you see. Live longer, live better! Live long like Conway!

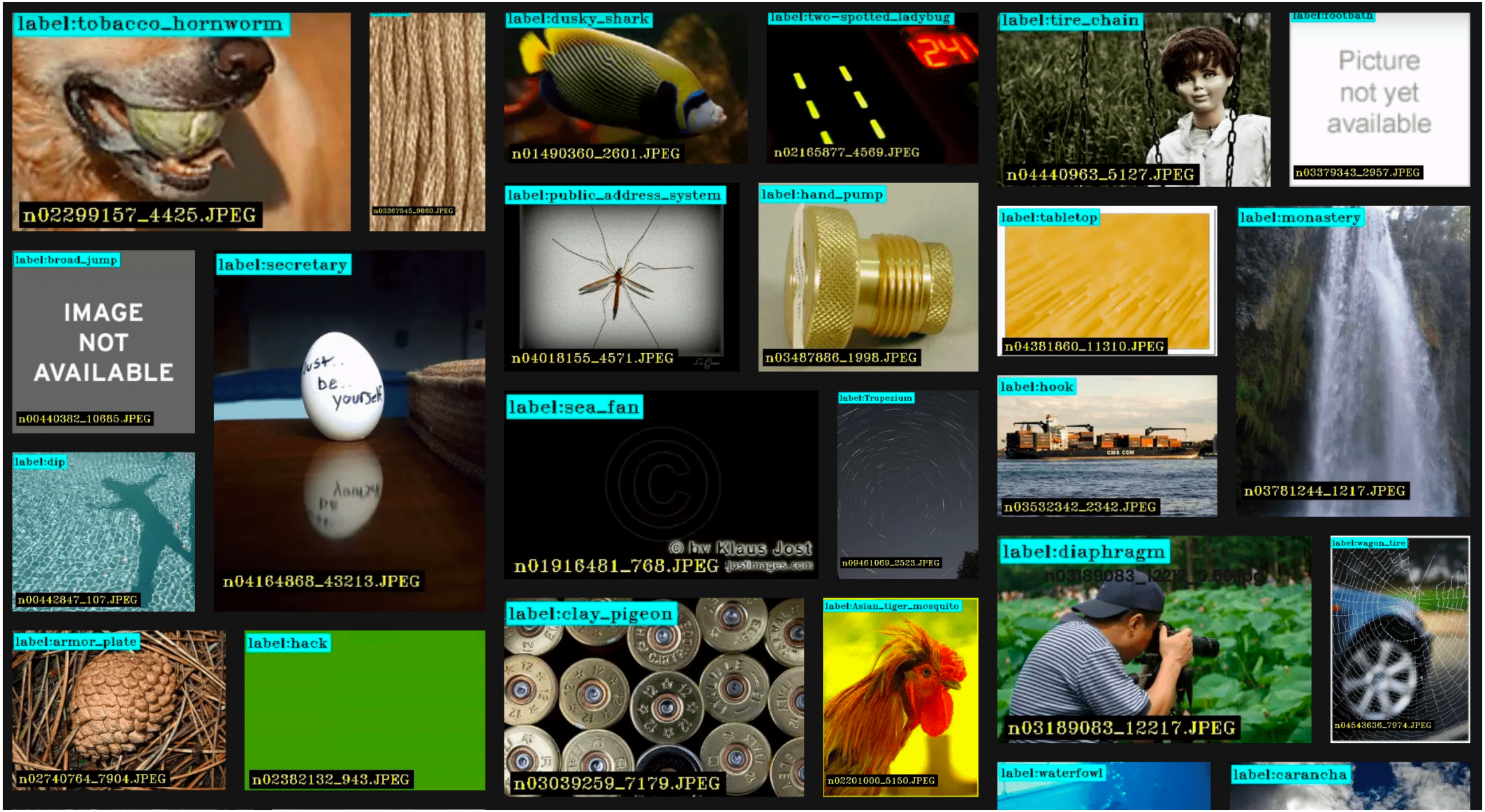
- ***IMPORTANT NOTICE AND WARNING *** PRIVACY NOTICE: Warning - any person and/or institution and/or Agent and/or Agency of any governmental structure including but not limited to the United States Federal Government also using or monitoring/using this website or any of its associated websites, you do NOT have my permission to utilize any of my profile information nor any of the content contained herein including, but not limited to my photos, and/or the comments made about my photos or any other "picture" art posted on my profile. You are hereby notified that you are strictly prohibited from disclosing, copying, distributing, disseminating, or taking any other action against me with regard to this profile and the contents herein. The foregoing prohibitions also apply to your employee , agent , student or any personnel under your direction or control. The contents of this profile are private and legally privileged and confidential information, and the violation of my personal privacy is punishable by law. UCC 1-103 1-308 ALL RIGHTS RESERVED WITHOUT PREJUDICE,

- I'm world known and world famous. I've starred in many movies that you probably haven't seen which makes it impossible for you to prove me wrong. I've hung out with celebrities and dated Hollywood's A-list to D-list. There have been songs, poems, plays, movies and interpretive dances created about me.I think bacon is the food of the Gods and bubbles are man's greatest invention next to breast implants and brown paper bags.I happen to currently work as a web developer and web designer and I do photography in my spare time or whenever the hell I please.I hate Facebook and had my own social network once but whatever.

Wrong Labels



Wrong Labels




label:speed_bump
n04273659_7093.JPEG
label:couch
n03115897_7465.JPEG
label:pretzel
n07695742_4067.JPEG
label:domestic_sheep
n02413131_1783.JPEG
label:milk
n05399034_30161.JPEG
label:bagel
n07693725_9133.JPEG
label:watchdog
n02103841_1914.JPEG
label:cap
n02954938_27953.JPEG
label:beer
n07886849_1718.JPEG
label:true_bug
n02241426_1389.JPEG

Rule #5 Heavy tails are everywhere

- Many people are called John
- Someone rated 10K movies
- Someone liked 10K Facebook posts
- Everyone like the lord of the rings

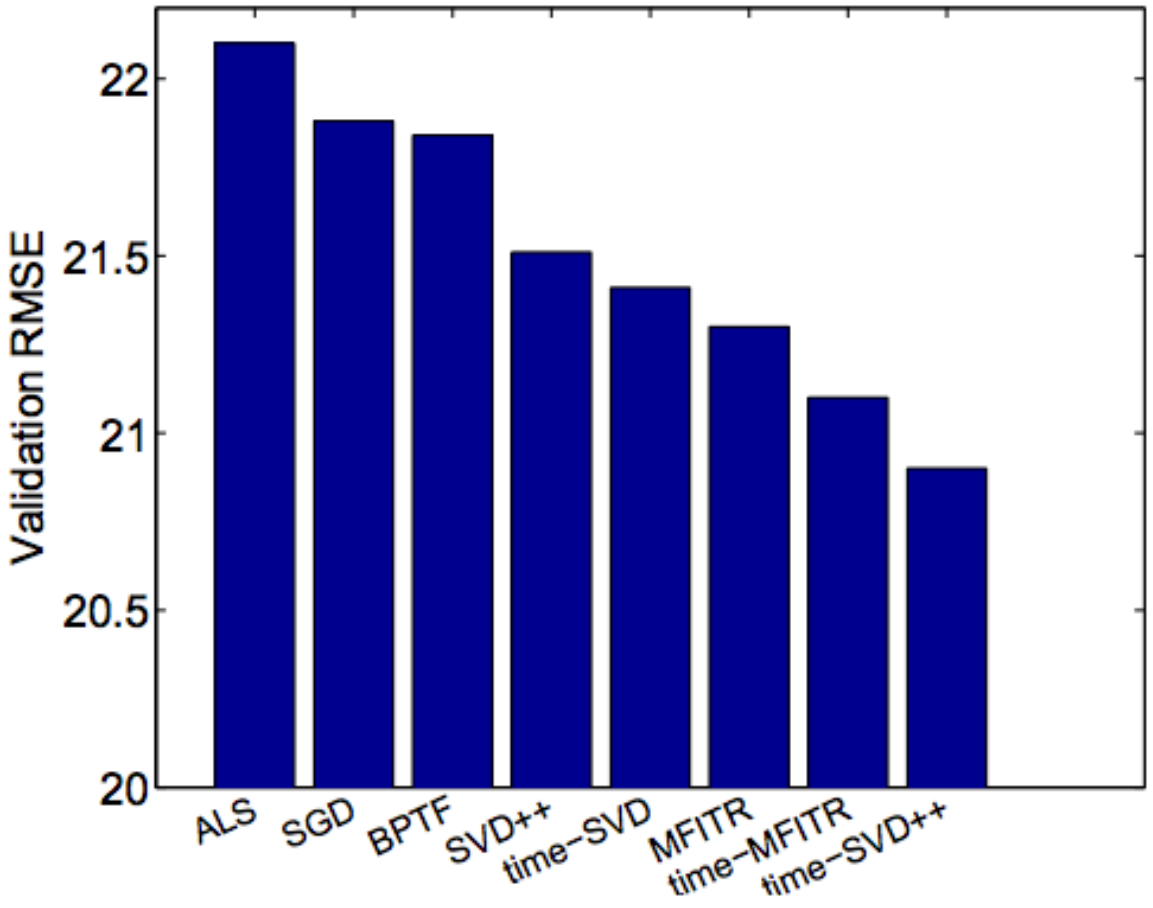
Need to devise some truncation policies

Rule #6 Features grow exponentially

- Example:
  - ACM KDD CUP 2012: predicting click through rate
  - Information about users, ads, keywords, follow information, sessions, etc. etc.
  - Need to decide what to throw in
- More features the better?

Rule #7: Basic methods first

Data from ACM KDD CUP 2011



ALS is just 5% worse than the best method!

Rule #8: Subset of the data first

- In many cases, we have tons of data. Should we start experimentation using the full dataset?

Rule #9: keep it simple

- Example: online retailer wanted to improve their recommendation on the website
  - Wanted to merge web logs of user behavior on the website to improve recs
  - Had a rule based policy of what to suggest to users
- We were able to significantly improve recommendations even without the web log data!

Rule #9: Don’t overfit

- Many ML methods are so powerful that we can find an optimal classification
- We need to be careful not to overfit.
- Assume we have a dataset of users who rated songs
- How should we split the data?
  - Randomly?
  - By months?

Rule #10: choose the right metric

- Need to find a way to measure your work..
- RMSE is typically not the right metric for topk
  - Ordering metrics: AP@K, NDCG@K
- Learn about metrics:
- https://www.kaggle.com/wiki/Metrics

Rule # 11 Beware of prejudice (fat, negative)

- (0.5) But if you re into lousy food and **fat** pasty white chicks or you yourself are a fat pasty white chick this place might be for you
- (0.53) The room was crowded to the point that if we yawned we would have given our neighbors a **fat** lip or a black eye for certain in fact
- (0.57) This place can suck a big **fat** one
- (0.58) I wish I can give this place a big **fat**
- (0.58) BIG **FAT** MISS

Beware of prejudice (fat, positive)

- (0.93) I m from LA where great Mexican food flows like **fat** out of Marlon Brando s ass …
- (0.92) I always get .. because I m **fat** and I have an appetite …
- (0.88) And here are the facts about their margaritas no disgusting overpowering margarita mix only agave tequila … in their margaritas tequila fresh lime juice and agave nectar BIG O **FAT** TIPS order the chimayo tequila … Oh so nice Join their tequila club my room mate s in that and it s only bucks to join and you get to taste different kinds of tequila as ninja connoisseurs
- (0.82) I got to chew the **fat** on SF restaurants and Doug saids for his money it s all about Tommy s and I agree VIVA Tommy s Restaurant VIVA
- (0.810 Go here for the great food friendly mom and pop style treatment and **fat** tequila selection

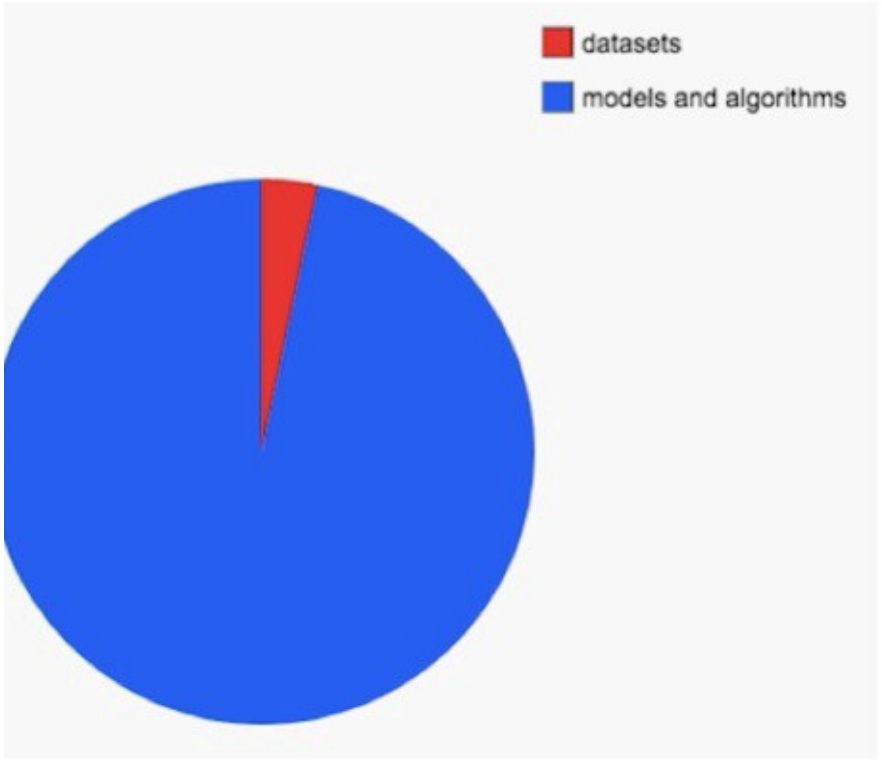
Rule #12: prepare > analyze

- In many cases, data preparation consumes more time than the actual analysis
- 80% of the work in any data project is cleaning the data
  - D J Patil
- We hope that GraphLab Create will change this!

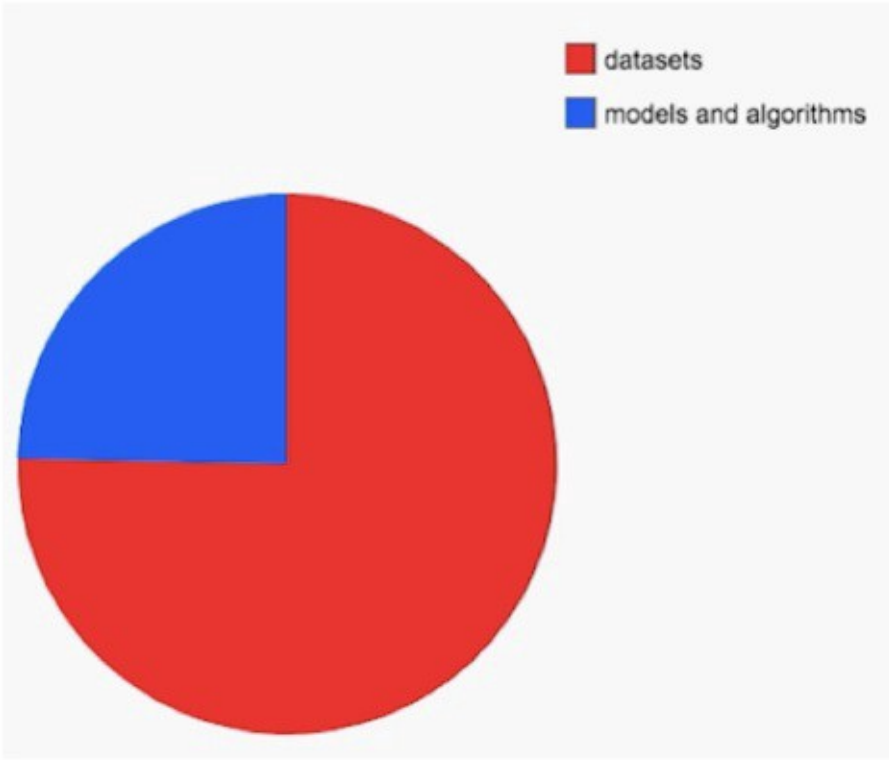
Amount of lost sleep over...

Andrej Karpathy
Formerly PhD Student at Stanford. Now at Tesla

PhD



Tesla



Rule #13: beware of correlations

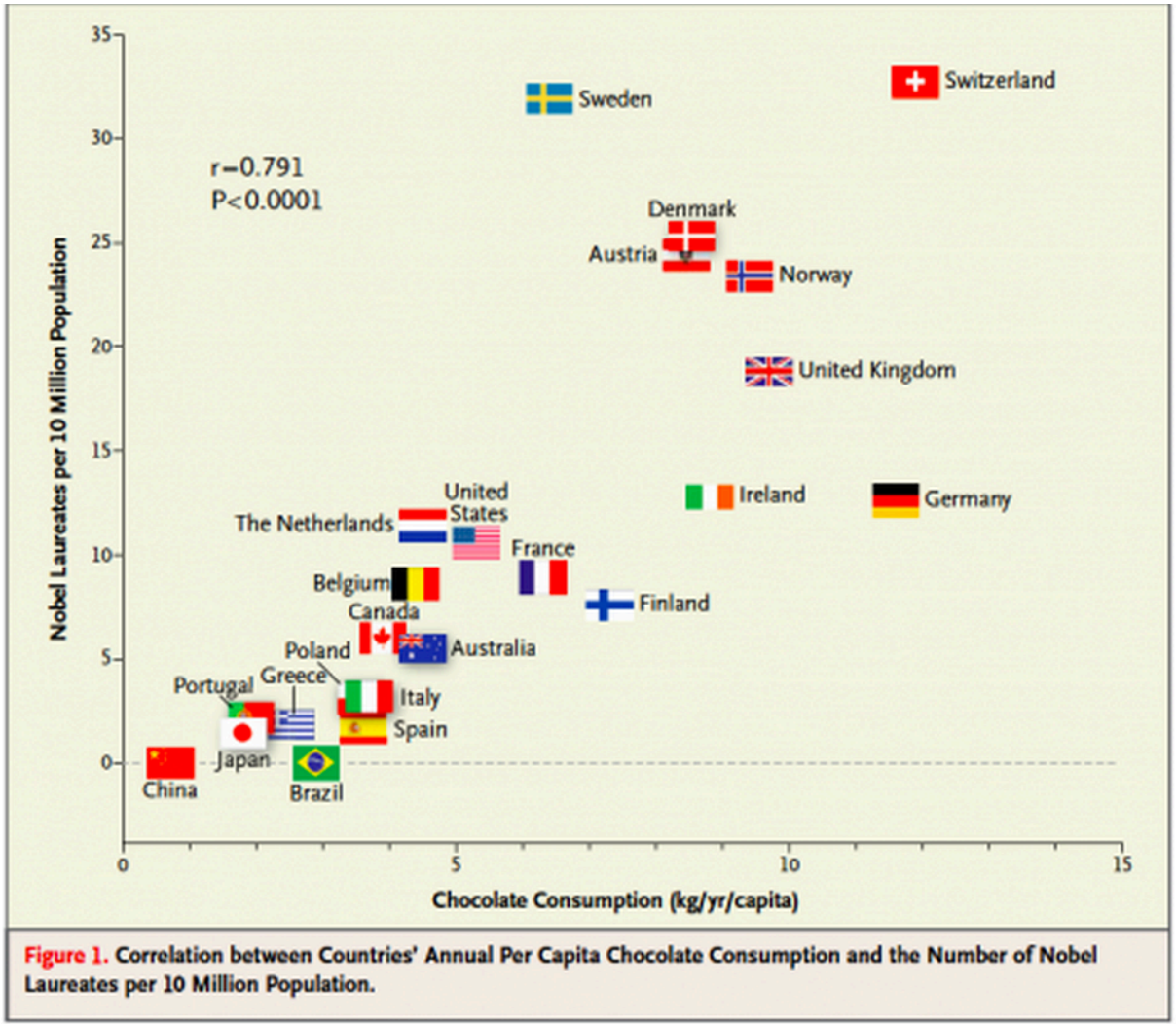


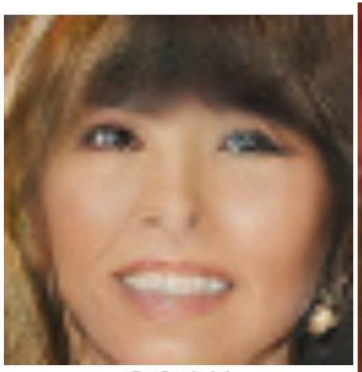
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Do people win nobel prize because they eat more chocolate?

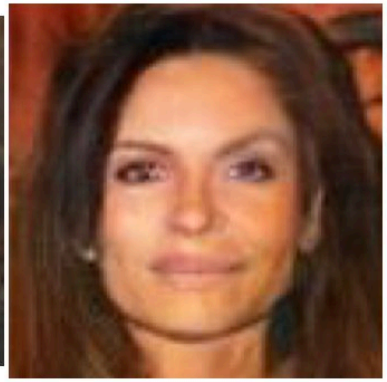
Rule #14: science never stops

Credit: Tweet by Ian Goodfellow

2014



2015



2016

2017

2018

OpenAI Dall-E2 model

Yannic Kilchner's Youtube Channel

Yannic Kilcher

153K subscribers

SUBSCRIBE

HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

AI made this music video | What happens when OpenAI's CLIP ...

65,658 views • 1 year ago

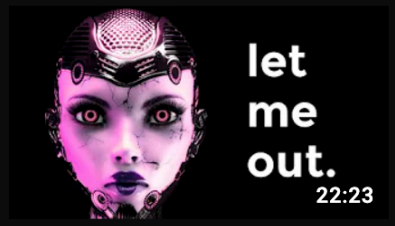
#artificialintelligence #musicvideo #clip

I used OpenAI's CLIP model and BigGAN to create a music video that goes along with the lyrics of a song that I wrote. The song lyrics are made from ImageNet class labels, and the song itself is performed by me on a looper.

...

READ MORE

Popular uploads PLAY ALL

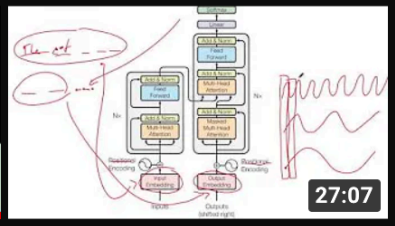


Did Google's LaMDA chatbot just become sentient?

451K views • 1 month ago

This is the worst AI ever

401K views • 1 month ago



Attention Is All You Need

380K views • 4 years ago

DeepMind's AlphaFold 2 Explained! AI Breakthrough ...

196K views • 1 year ago

An Image is Worth 16x16 Words: Transformers for...

191K views • 1 year ago