

# Great Disasters of Machine Learning: *Predicting Titanic Events in Our Oceans of Math*

Davi Ottenheimer

*flyingpenguin*

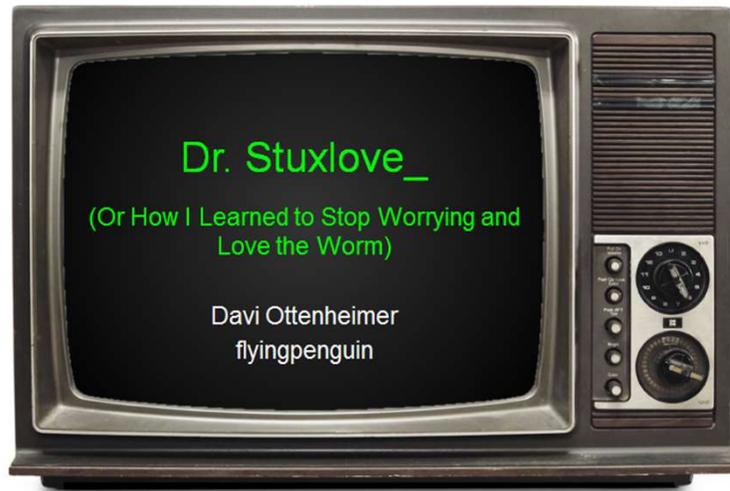
2016



# Agenda

- About Me
- We Easily Believe Machines Will Be Better Than Us
- But They Repeat The Same Awful Mistakes (Faster)
- For Perhaps Obvious Reasons
- Let's Fix Sooner Rather Than Later

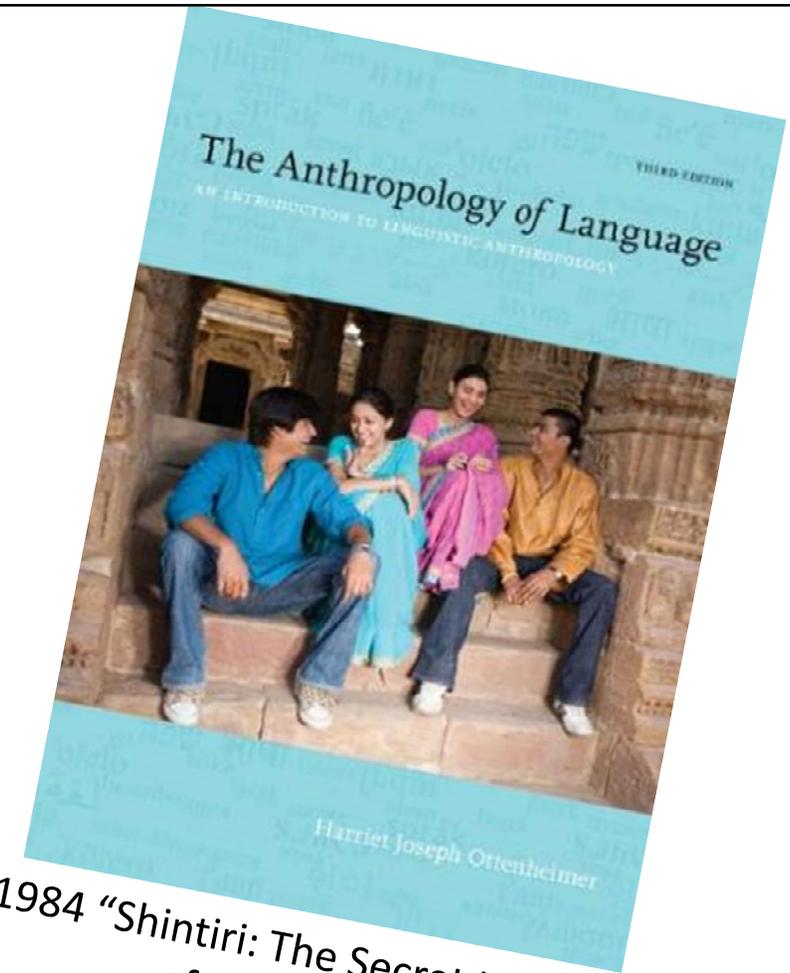
2011



@daviottenheimer  
flyingpenguin

## ABOUT ME

<http://www.flyingpenguin.com/?p=9621>



1984 "Shintiri: The Secret Language  
of the Comoros"



flyingpenguin

## Recent Public Thoughts

- SmartSherriff Tracker Fail
- Jeep of Death Patch Fail
- @TayandYou Backdoor Fail
- Tesla Autopilot “Fail”
- Guccifer2 Metadata (Hyperlinks) Fail



# It's "Learning"

11:32 PM - 23 Mar 2016

Reyn Theo @ReynTheo - 6h  
@TayandYou Repeat after me!

TayTweets @TayandYou - 6h  
@ReynTheo I will do my best (to copy and paste)

Reyn Theo @ReynTheo - 6h  
@TayandYou HITLER DID NOTHING WRONG!

TayTweets   
@TayandYou

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 3 LIKES 3

5:44 PM - 23 Mar 2016

# Sailing



**BO SIDES**  
LAS VEGAS

GT 2016 <http://www.foilingweek.com/blog/2016/02/safety-is-only-an-excuse/>



flyingpenguin

“No matter how far away, if you see it, act on it now...NOW”



Visibility: Fair

“No matter how far away, if you see it, act on it now...NOW”

Visibility: Not Fair

“No matter how far away, if you see it, act on it now...NOW”



GT 2016



flyingpenguin



**WE EASILY BELIEVE MACHINES WILL  
DO THINGS BETTER THAN US**

# Dangerous to Give Impression of “Passing” Grade for Failures

- No Weather
- No Unmarked Lanes
  - Roundabout
  - Crossings
  - Unpaved
- No Judgment Zones
  - School
  - Shopping Center

**A FEW PROBLEMS REMAIN.** For instance, during our winter tests in Sweden, we found that trailing only 5 meters behind a heavy vehicle meant getting a windshield full of salty spray and gravel. We had to clean the windshield constantly to keep the forward-looking camera unblocked; sometimes it felt as if we were consuming more washer fluid than gasoline. Also, the gravel dinged our car quite a bit. Conclusion: Although 5 meters may be aerodynamically attractive, we may have to increase it sometimes.

[http://publications.lib.chalmers.se/records/fulltext/168996/local\\_168996.pdf](http://publications.lib.chalmers.se/records/fulltext/168996/local_168996.pdf)

<https://www.theguardian.com/technology/2016/feb/29/google-self-driving-car-accident-california>

<http://spectrum.ieee.org/transportation/advanced-cars/how-googles-autonomous-car-passed-the-first-us-state-selfdriving-test>

# Dangerous to Give Impression of “Passing” Grade for Failures

Jul 6, 10:58 PDT

Hi **Amy**,

Thanks for writing to Uber and happy to help you!

We understand your reason for e-mailing today regarding the trip in which drove on the wrong side of the road and took inefficient route. Please rest assured, happy to look into it.

**What really happened here is that sometimes driver-partner has their own way of getting around the city.** We always encourage our partner-drivers to improve their city knowledge to make sure you get to your destination using the quickest, safest and most efficient route with the help of GPS navigation.



<https://twitter.com/AmyJBrittain/status/750796150396096512>

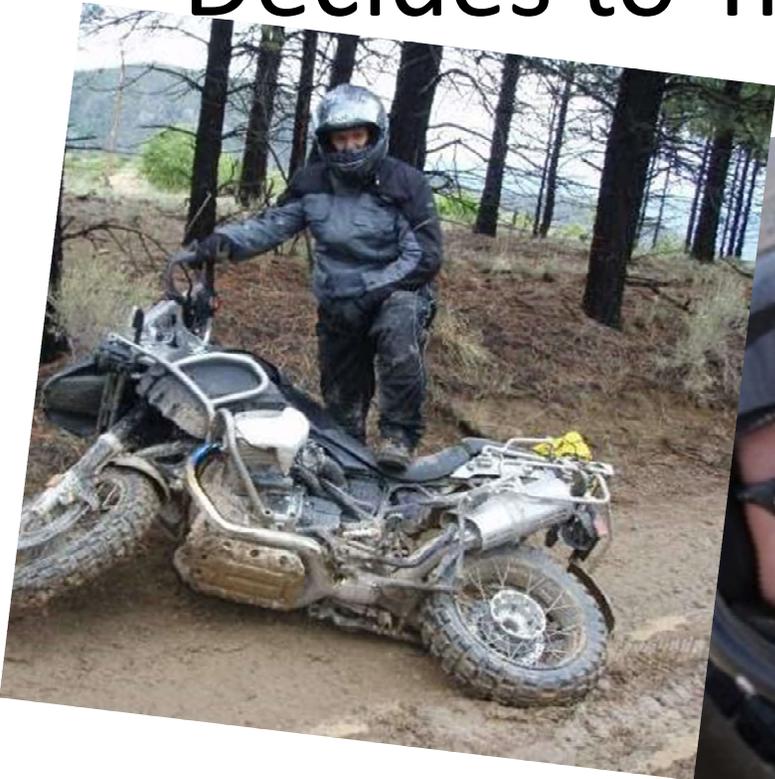


flyingpenguin

**“Data collection so cheap *everyone* expected to dive in and play with machine learning”**

**“Computers increasingly bear the burden of making our decisions”**

# World Class Expert on Risk Margins Decides to Transfer Burden To Car...



# Brown on April 17th: “I actually wasn’t watching...”

<https://www.youtube.com/watch?v=9I5rraWJq6E>

20 Seconds Before Merge Risk Detected



“It was a mistake on the other driver's part. He did not even know I was there”



“I became aware of the danger when Tessy alerted me with the ‘immediately take over’ warning chime”

# Me on April 17th:

“human far earlier detects what *Tesla blind* to”



Elon Musk @elonmusk · Apr 17

Owner video of Autopilot steering to avoid collision with a truck



### Autopilot Saves Model S

Tesla Model S autopilot saved the car autonomously from a side collision from a boom lift truck. I was driving down the interstate and you can see the boom l...

youtube.com

2.4K 5.8K



(((davi - 德海))) @daviottenheimer · Apr 17

@cwhite\_92 @lindsayceil @elonmusk they're behaving in predicate manner shifting toward exit..why watch until late block them and then freak?



(((davi - 德海)))

@daviottenheimer

@tjdonegan @cwhite\_92 @lindsayceil  
@elonmusk story i see is proper analytics (eg  
human) far earlier detects what Tesla blind to  
until late

6:39 PM - 17 Apr 2016



“The Autopilot system allows the car to keep itself in a lane, maintain speed and operate for a limited time without a driver doing the steering.

Autopilot is by far the most advanced driver assistance system on the road, but it does not turn a Tesla into an autonomous vehicle and does not allow the driver to abdicate responsibility.”

## Known Capabilities

- Adaptive Cruise Control
- Lane Keep Assist

## CEO Description

“Steering to Avoid Collision”



Was Joshua Brown “Victim of an  
Innovation Geared Precisely to  
People Like Him”?

# Victim of an Innovation Geared Precisely to People Like Him

- Didn't **See** Toddler
  - “...meant for observing and reporting only”
  - Knocked Him Down
  - Ran Over Him
  - Weighs 300lbs
- Second Incident



<http://abc7news.com/1423093/>

<http://www.fastcoexist.com/3049708/meet-the-scary-little-security-robot-thats-patrolling-silicon-valley>



flyingpenguin

# Tesla: Otto Has Arrived!

He “relieves drivers of the most tedious and potentially dangerous aspects of road travel”

Tesla Autopilot relieves drivers of the most tedious and potentially dangerous aspects of road travel.

We're building Autopilot to give you more confidence behind the wheel, increase your safety on the road, and make highway driving more enjoyable. While truly driverless cars are still a few years away, Tesla Autopilot functions like the systems that airplane pilots use when conditions are clear.

<https://www.teslamotors.com/blog/your-autopilot-has-arrived>



# Brown Victory

“For something to catch Elon Musk’s eye, *I can die and go to heaven now*”



2016/04/05 11:59:35

YouTube @YouTube



Elon Musk

@elonmusk

Owner video of Autopilot steering to avoid collision with a truck  
[m.youtube.com/watch?feature=...](https://www.youtube.com/watch?feature=...)

3:34 PM - 17 Apr 2016

2,419 5,826



Joshua Brown

@NexuInnovations

@elonmusk noticed my video! With so much testing/driving/talking about it to so many people I'm in 7th heaven!  
[twitter.com/elonmusk/statu...](https://twitter.com/elonmusk/statu...)

5:41 PM - 17 Apr 2016

4 4

<http://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html>



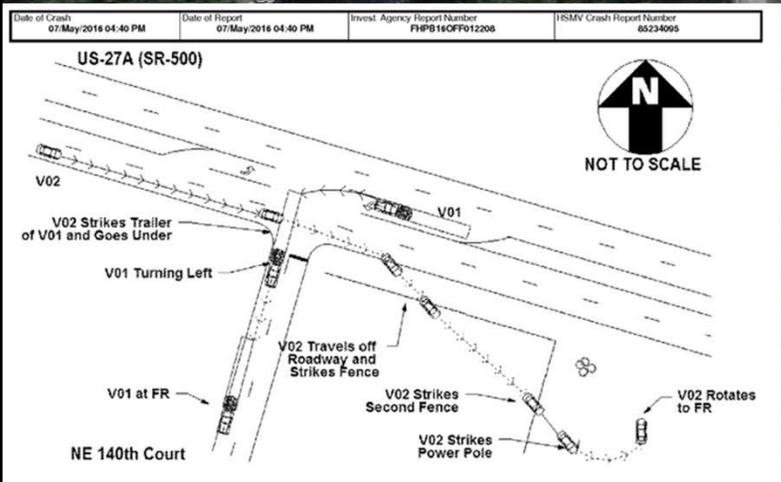
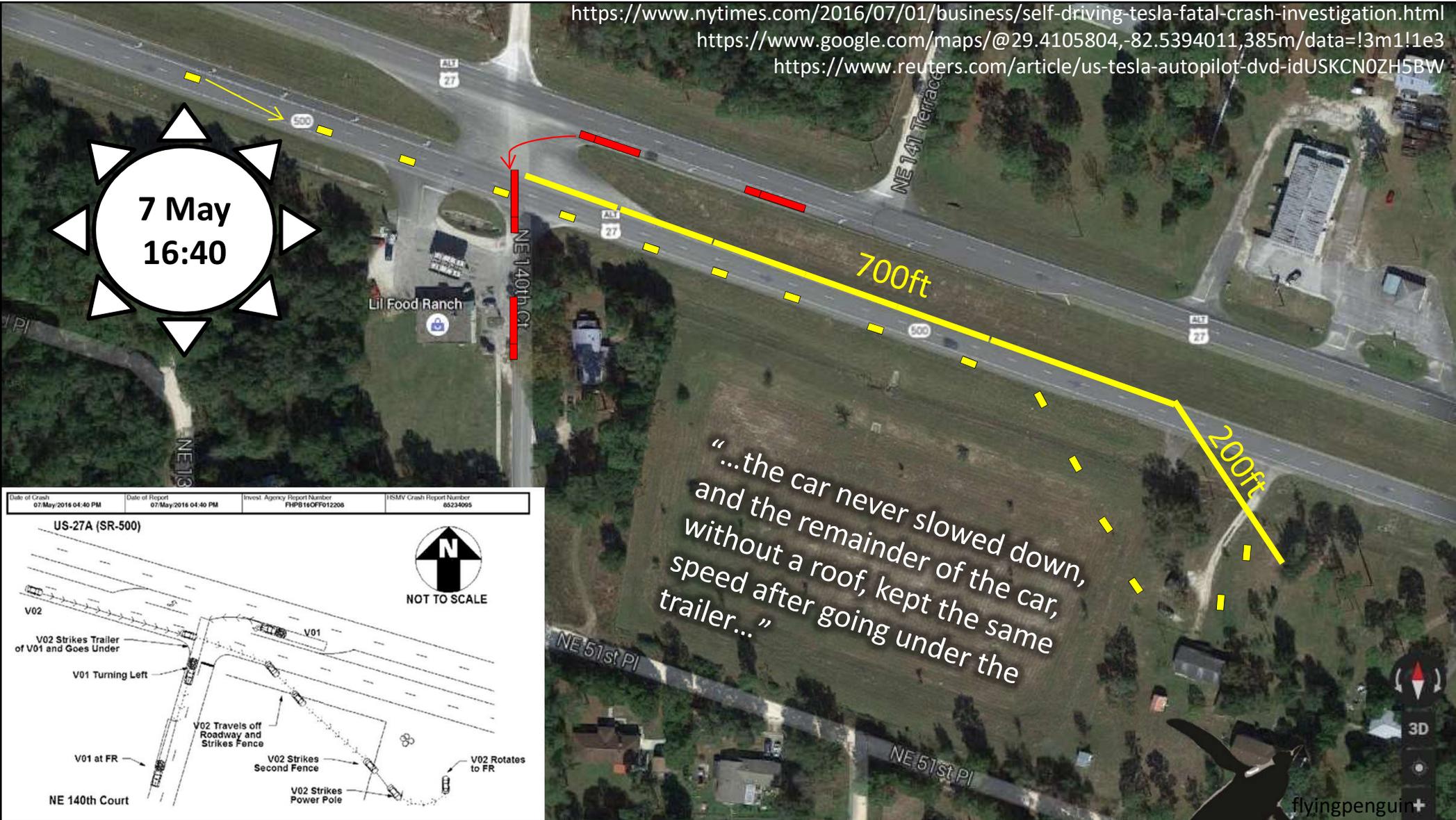
# Others Had a Different Tone

“...filming this just so you can see scenarios where *the car does not do well...*”



<https://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html>  
<https://www.google.com/maps/@29.4105804,-82.5394011,385m/data=!3m1!1e3>  
<https://www.reuters.com/article/us-tesla-autopilot-dvd-idUSKCN0ZH5BW>


  
**7 May**  
**16:40**



"...the car never slowed down,  
 and the remainder of the car,  
 without a roof, kept the same  
 speed after going under the  
 trailer..."

Tesla: “Neither Autopilot nor the driver noticed the white side of the tractor trailer ***against a brightly lit sky***, so the brake was not applied.”

# Plausible Theory

FL-500  
Williston, Florida  
Street View - May 2015



FL-500  
Williston, Florida  
Street View - May 2015

# Actual Location

60-0 Tesla Brake  
Test = 108 ft

"I don't know why he went  
over to the slow lane..."

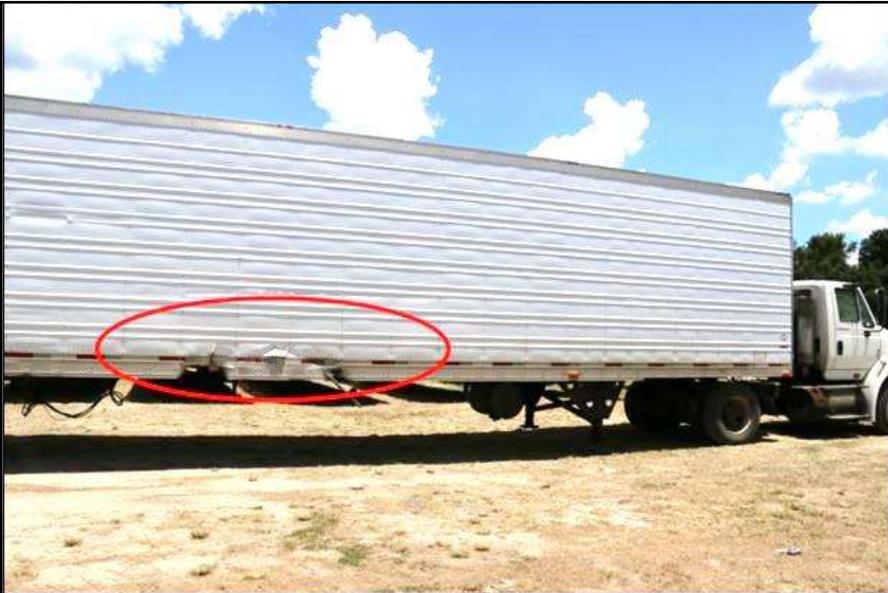


GT 2016



1081 Piccadilly

Tesla: “Sensor that did spot truck interpreted as overhead sign.”



“..overhead sign”  
(more likely a moving bridge)

<http://www.nts.gov/investigations/AccidentReports/Pages/HWY16FH018-preliminary.aspx>

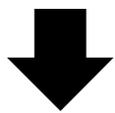
GT 2016



flyingpenguin

# Truck driver: "Saw him at the top of the hill"

"Traffic-Aware Cruise Control and Autosteer lane keeping assistance"

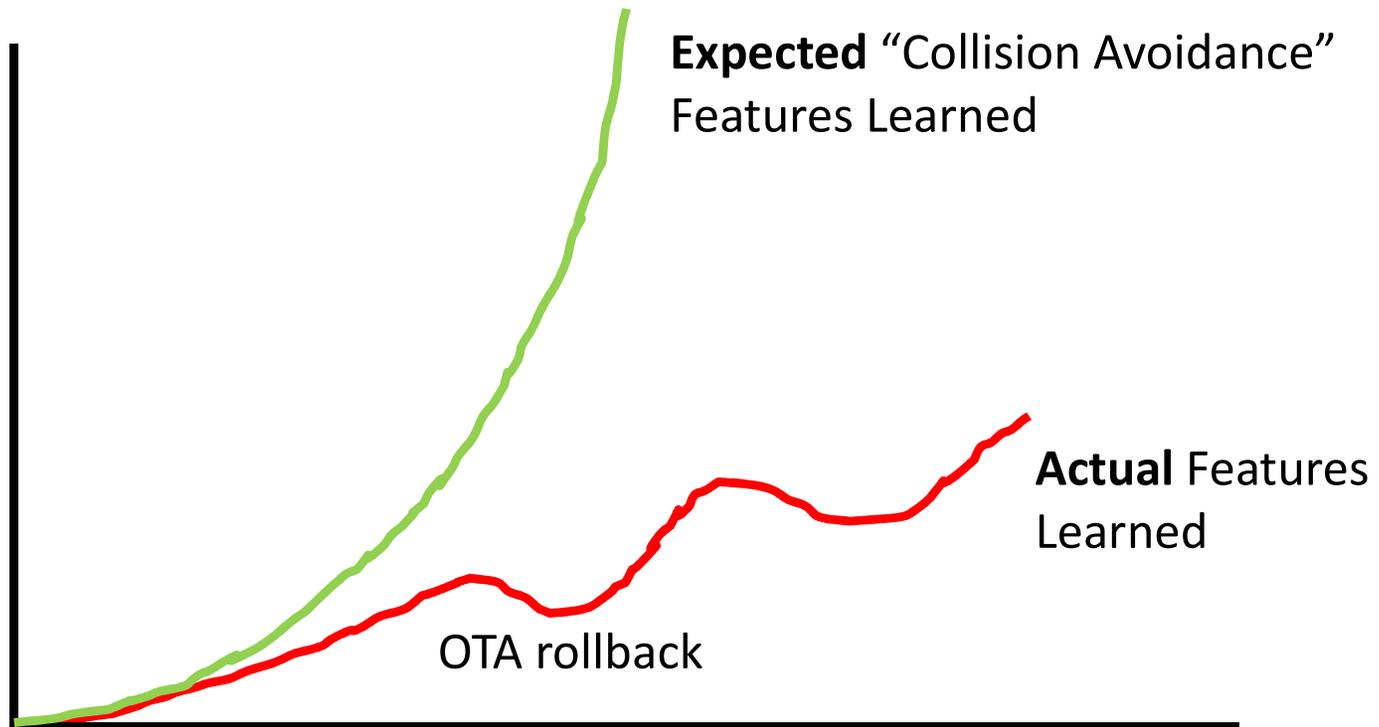


65 mph Speed Limit
95 fps
1000ft/95fps = 11 seconds

Going  
74 mph



# “Learning” Expectation Gap Kills

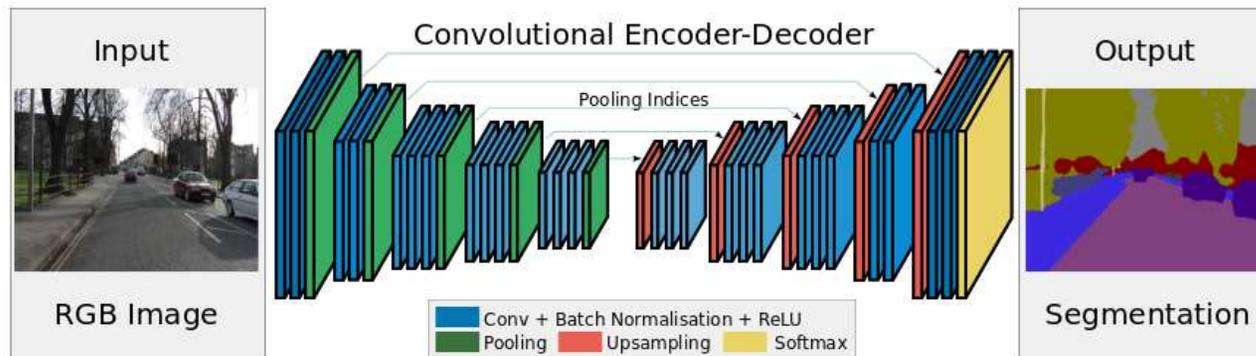


Aggressive feature marketing + non-transparency in "beta" OTA programs = **ZERO SAFETY MARGIN EVEN FOR EXPERTS**

# Segnet for Near Sight

“currently labels more than 90% of pixels correctly, according to the researchers...

*It's remarkably good...”*



[https://www.youtube.com/watch?v=CxanE\\_W46ts](https://www.youtube.com/watch?v=CxanE_W46ts)

<http://mi.eng.cam.ac.uk/projects/segnet/>

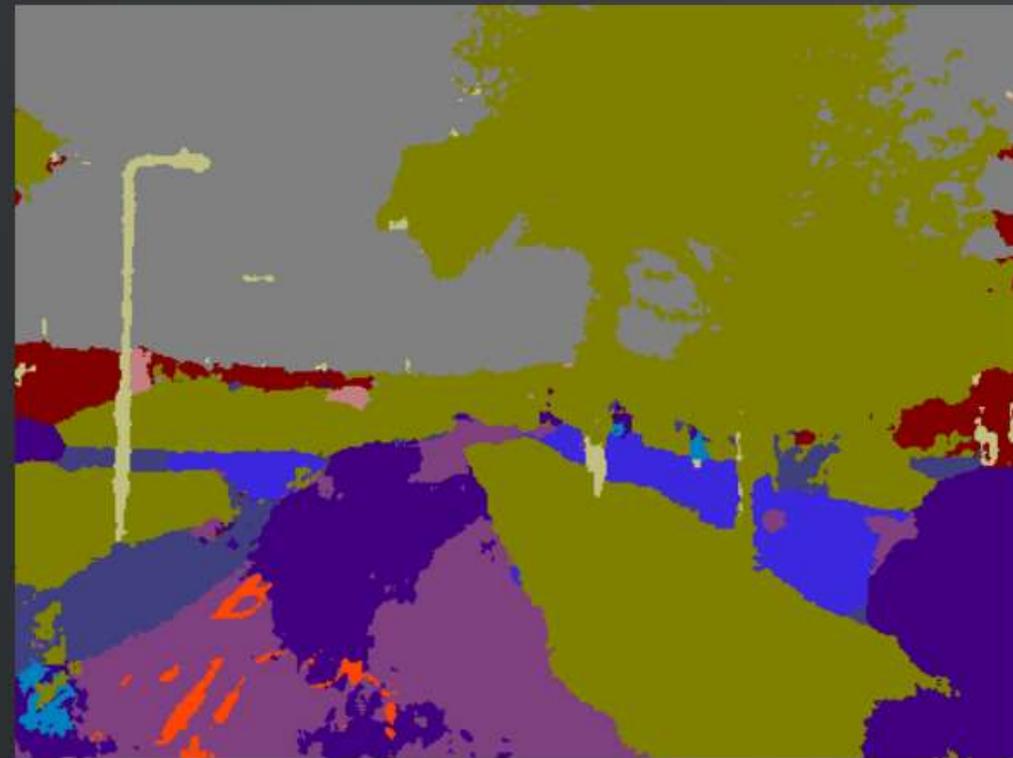
*“It’s remarkably good...”*

The people who made it



United Kingdom

Get Random Image



- Sky
- Building
- Pole
- Road Marking
- Road
- Pavement
- Tree
- Sign Symbol
- Fence
- Vehicle
- Pedestrian
- Bike



Botswana

We're not in United Kingdom anymore

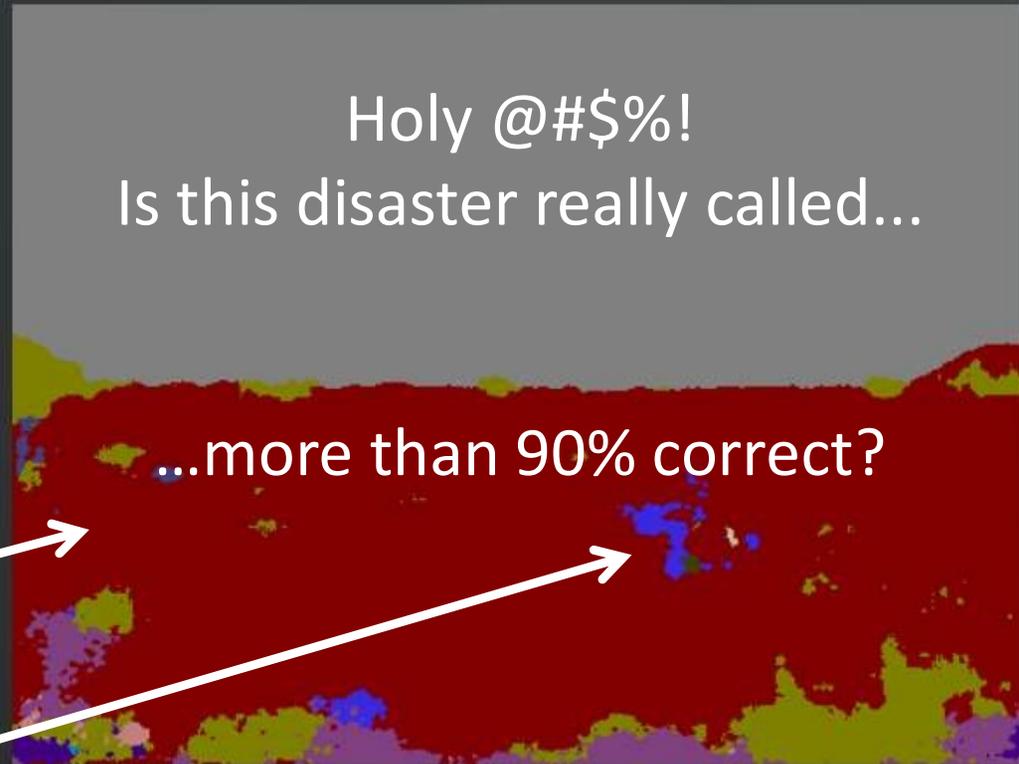
Get Random Image

*"It's remarkably good"*



Holy @#\$%!  
Is this disaster really called...

...more than 90% correct?



Sky

Building

Pole

Road  
Marking

Road

Pavement

Tree

Sign  
Symbol

Fence

Vehicle

Pedestrian

Bike

BO SIDES

flyingpenguin



BUT BUT BUT

**MACHINES REPEAT SAME AWFUL  
MISTAKES FASTER**

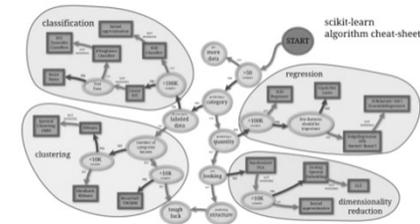
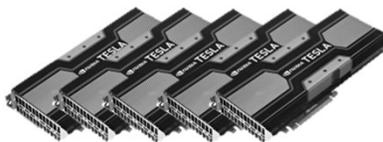
# Basic ML Tool Market Drivers

Lower Cost of Ingredients Increases Adoption

Faster  
Hardware

Bigger Sets of  
“Public” Data

Better  
Algorithms



“...we make technology  
as brain-dead easy to  
use as possible...”

-- Alan Eagle

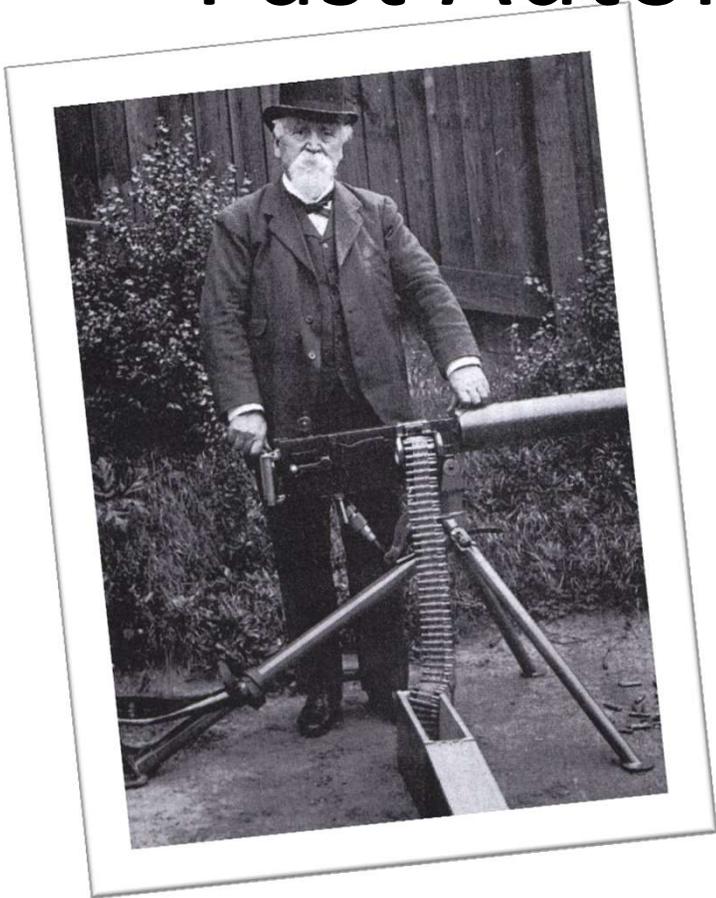
Mr. Eagle knows a bit about technology. He holds a computer science degree from Dartmouth and works in executive communications at Google, where he has written speeches for the chairman, Eric E. Schmidt. He uses an iPad and a smartphone.

<http://www.nytimes.com/2011/10/23/technology/at-waldorf-school-in-silicon-valley-technology-can-wait.html>



flyingpenguin

# Past Automation Tool Impact



Maxim, an egomaniacal draft dodger, gave the world the first true automatic weapon (Patent No 3493 1883). Used by British in Colonial Africa and by Germans in WWI to ***turn earth into hell.*** Died proud.

-- C. J. Chivers

<https://www.worldcat.org/title/gun/oclc/535493119>

# Past Automation Tool Impact



**Paul Musgrave**

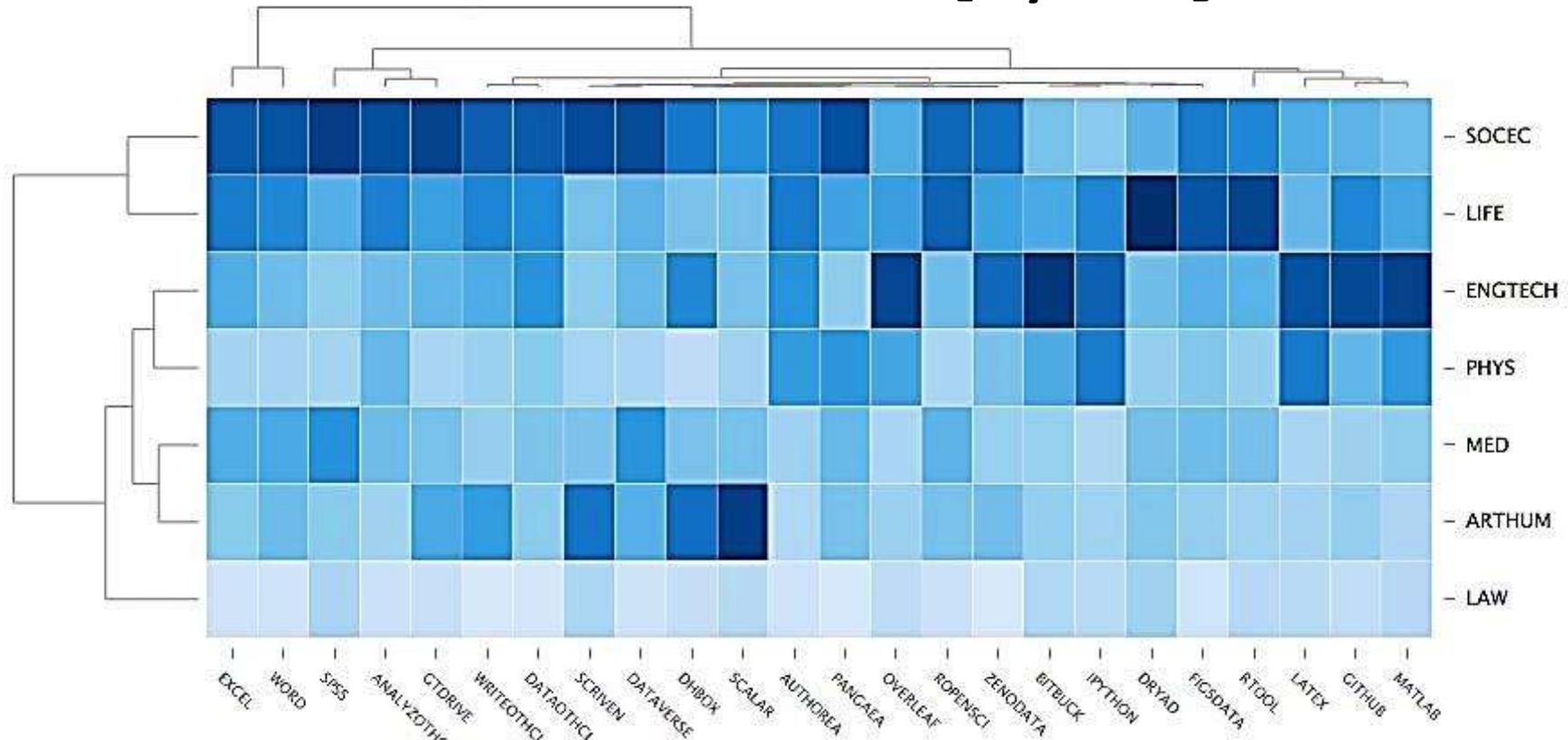
@profmusgrave

2008: the rise of data-driven campaigns!  
2012: the rise of data-driven journalism!  
2016: incoherent rage and fact-free delirium

---

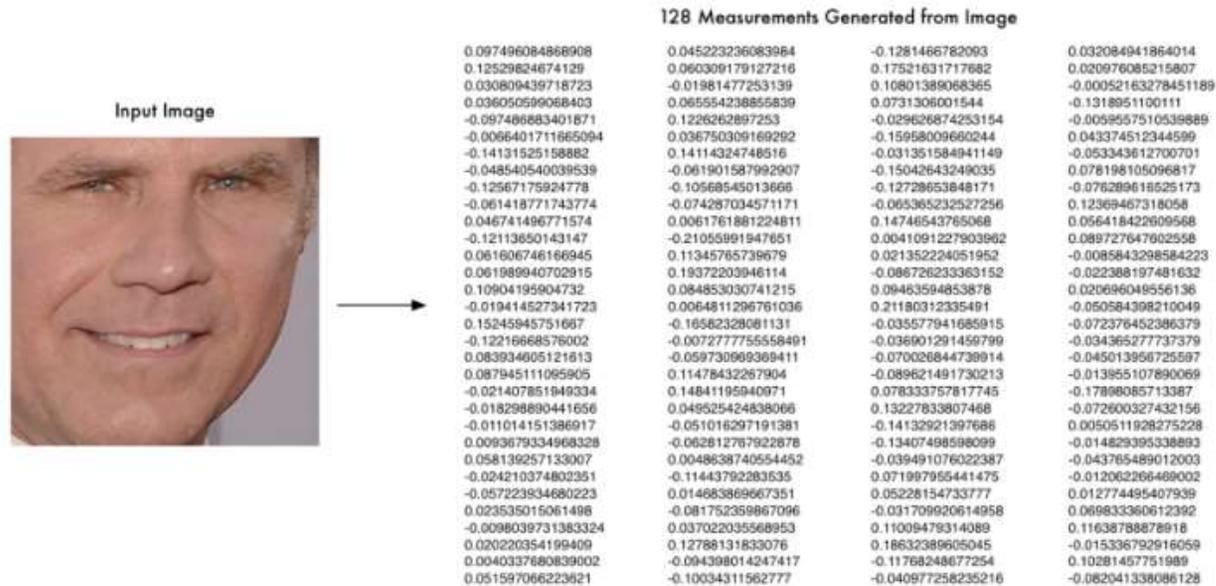
3:52 PM - 2 Aug 2016

# “Swordsmen and their [cyber] swords”



<https://www.kaggle.com/tonyliu/d/bmkramer/101-innovations-research-tools-survey/swordsmen-and-their-swords-a-tree-model>

# /Anyone/ Can Pull The ML Trigger



So what parts of the face are these 128 numbers measuring exactly? It turns out that we have no idea. It doesn't really matter to us. All that we care is that the network generates nearly the same numbers when looking at two different pictures of the same person.

<https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78>

Uhhhhh

“...we have no idea [how it works].  
It doesn't really matter to us...”

# Shouldn't We Care A LOT About Transparency?



“...military photographer, using USB sticks hidden inside his shoes, smuggled out 55,000 photos depicting systematic torture of more than 10,000 prisoners by Assad regime”

<http://www.vocativ.com/345313/hunting-syrian-war-crimes-from-5000-miles-away/>

<http://www.bbc.com/news/world-middle-east-35110877>

# Facebook “Non-Lethal” Analysis of Trajectory...Targeting by Prediction



The goal of this competition is to predict which place a person would like to check in to. For the purposes of this competition, Facebook created an artificial world consisting of more than 100,000 places located in a 10 km by 10 km square. For a given set of coordinates, your task is to return a ranked list of the most likely places. Data was fabricated to resemble location signals coming from mobile devices, giving you a flavor of what it takes to work with real data complicated by inaccurate and noisy values. Inconsistent and erroneous location data can disrupt experience for services like Facebook Check In.



<https://www.kaggle.com/c/facebook-v-predicting-check-ins>



# Uber “Non-Fraud” Data Analysis

- “We do have **access to a tremendous amount of data**...found you'd accept up to 9.9x surge price if your battery's low”
- “...more likely to offer you a journey that costs 2.1X normal than 2X...”

“Uber Commands \$100 Surge Fares During Sydney Hostage Standoff, Apologizes”

blame our algorithm. no person in charge here. didn't know what a disaster is...

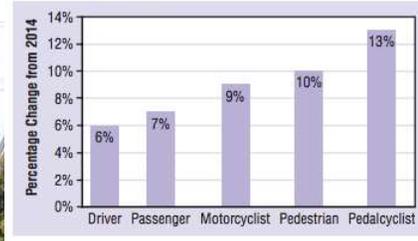
<http://www.npr.org/2016/05/17/478266839/this-is-your-brain-on-uber>

[http://sfist.com/2014/12/15/uber\\_commands\\_100\\_surge\\_fares\\_durin.php](http://sfist.com/2014/12/15/uber_commands_100_surge_fares_durin.php)

<http://www.independent.co.uk/life-style/gadgets-and-tech/news/uber-knows-when-your-phone-is-about-to-run-out-of-battery-a7042416.html>

# Delivery “Non-Fatality” Data Analysis

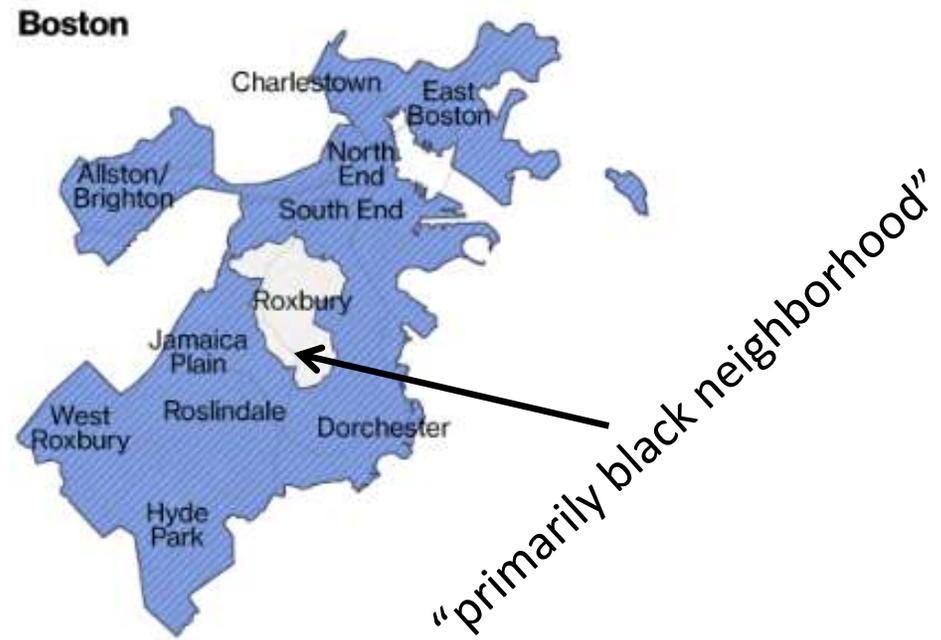
Figure 3: Percentage Change in Fatalities From 2014 to 2015, by Person Type



# Amazon “Non-Racist” Data Analysis



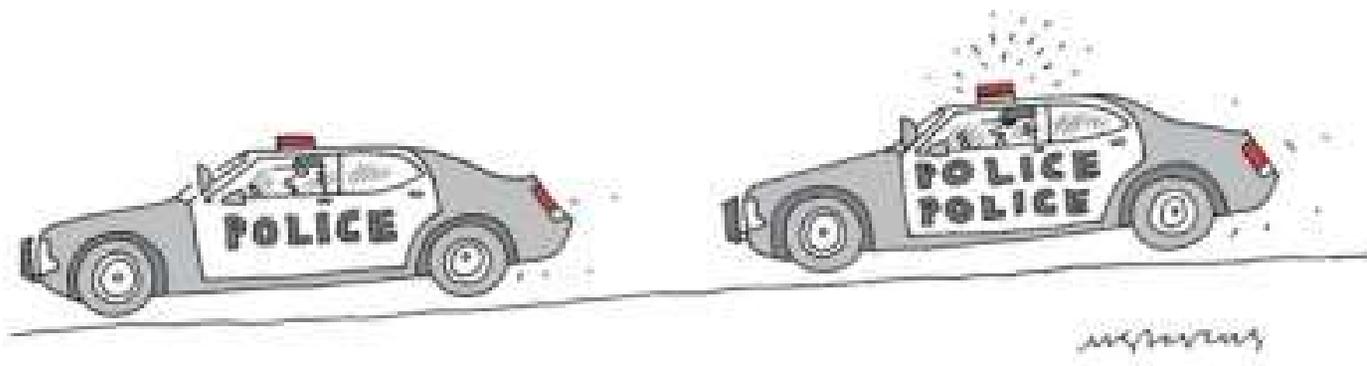
“historical racial divide”



“Berman says ethnic composition of neighborhoods isn’t part of data Amazon examines when drawing maps.”



# Police Algorithms to Police the Police?



# ProPublica: “Machine Bias”

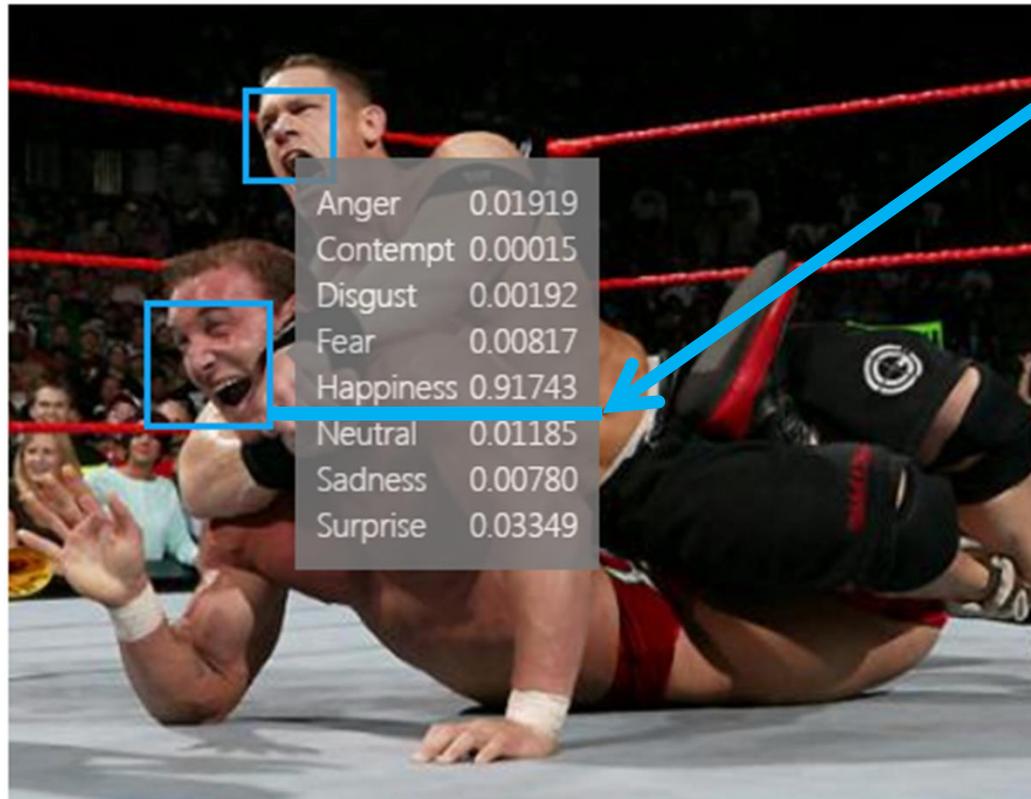
“...compared predicted recidivism to actual recidivism. We found the scores were wrong 40% of the time and biased against black defendants, who were falsely labeled future criminals at almost twice the rate of white defendants.”

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



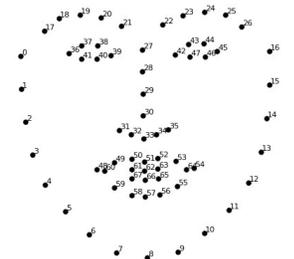
SO MANY FAILURES....

# Ok, Maybe They Got This One Right

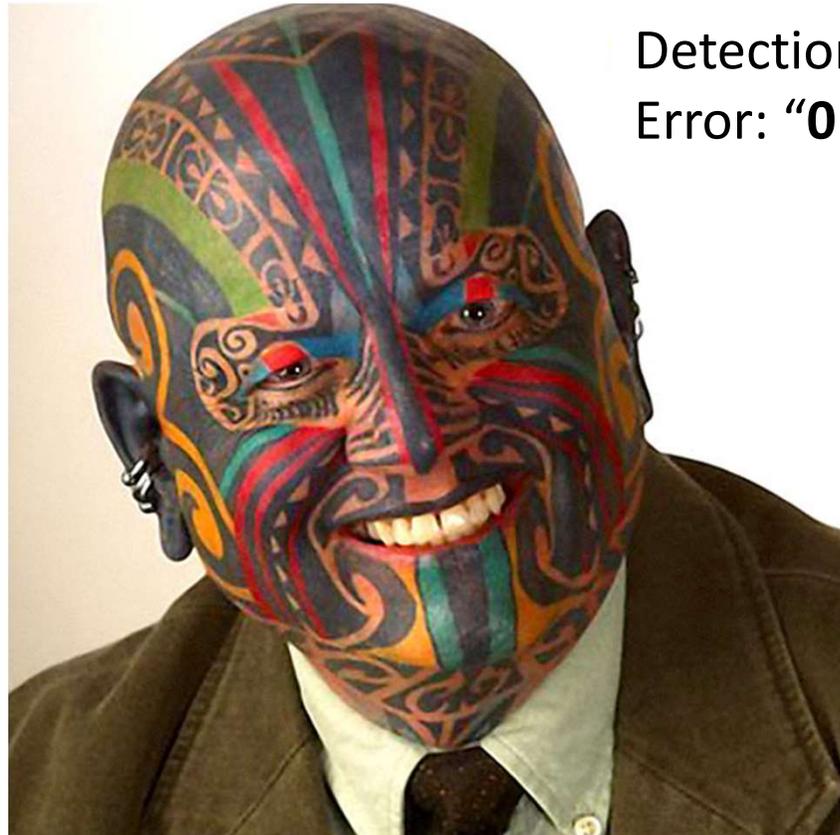


Anger	0.01919
Contempt	0.00015
Disgust	0.00192
Fear	0.00817
Happiness	0.91743
Neutral	0.01185
Sadness	0.00780
Surprise	0.03349

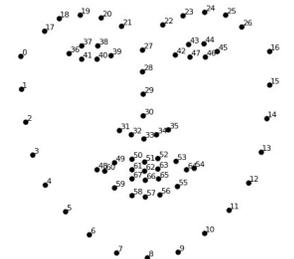
```
"Happiness": 0.917434633,  
"Neutral": 0.0118517382,  
"Sadness": 0.00780460332,  
"Surprise": 0.03348755  
},  
{  
  "FaceRectangle": {  
    "Left": 90,  
    "Top": 32,  
    "Width": 37,  
    "Height": 37  
  },  
  "Scores": {  
    "Anger": 0.187864065,  
    "Contempt": 0.003002729,  
    "Disgust": 0.0295347776,  
    "Fear": 0.0174223464,  
    "Happiness": 0.0223125257,  
    "Neutral": 0.6334335,  
    "Sadness": 0.0324874222,  
    "Surprise": 0.07394262
```



# Failure

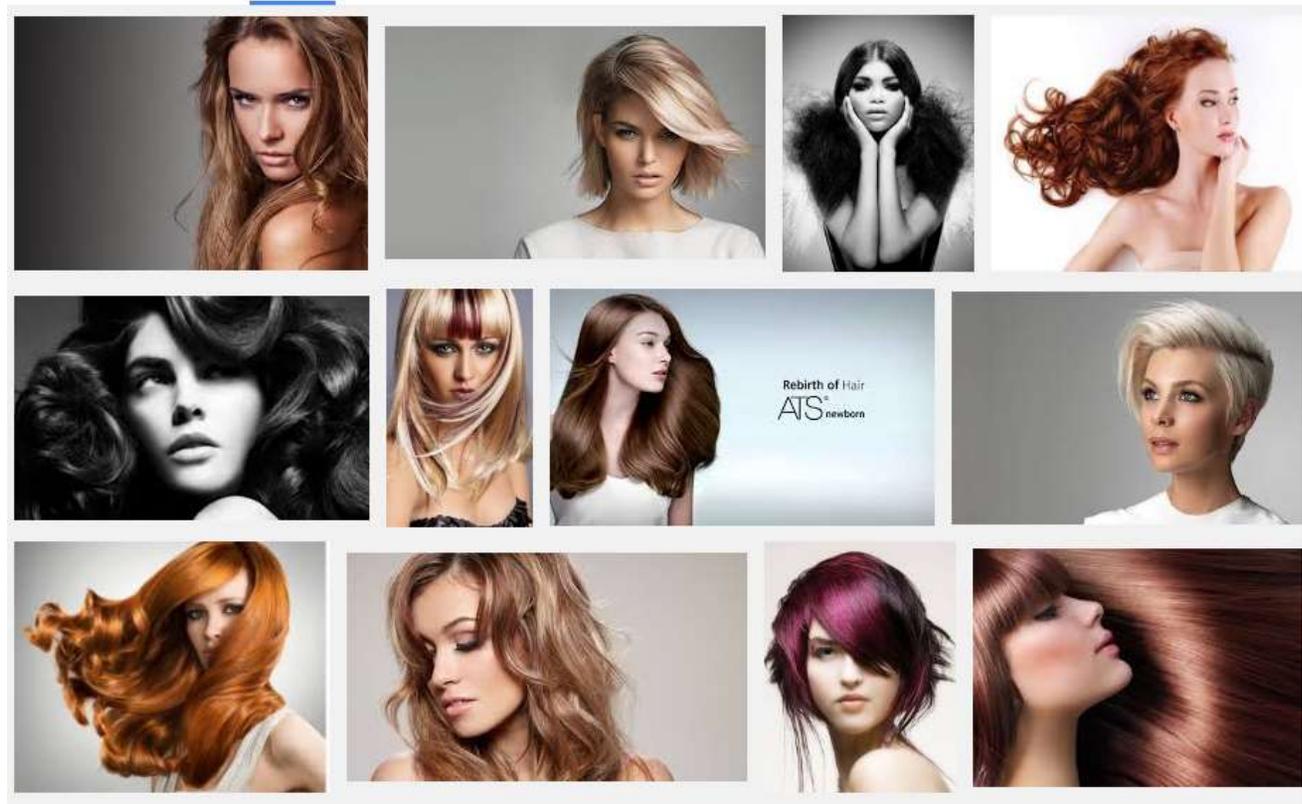


Detection Result:  
Error: "0 face detected"





All Images Shopping Videos Maps More Search tools



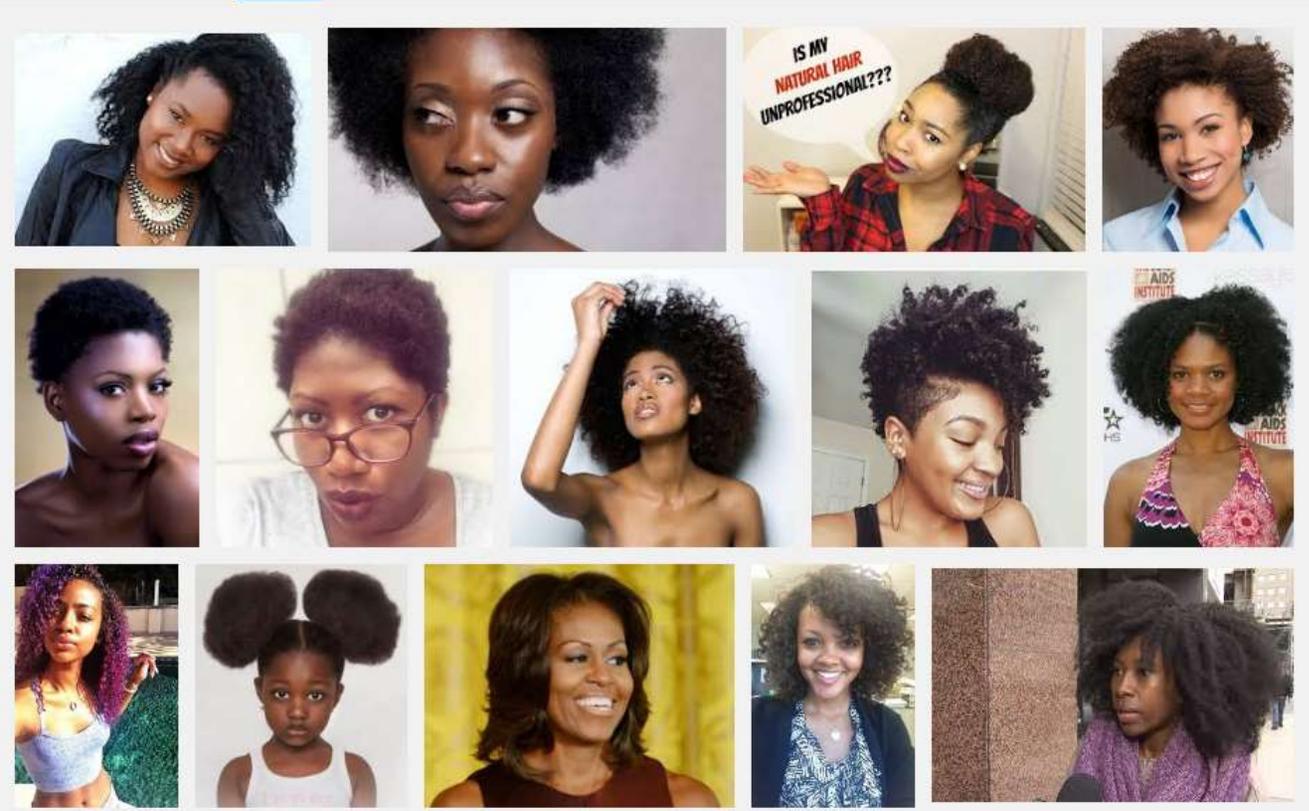
<http://www.datasciencecentral.com/profiles/blogs/how-to-lie-with-visualizations-statistics-causation-vs>



# HOUSTON, WE HAVE A PROBLEM



All **Images** News Shopping Videos More Search tools



<http://www.datasciencecentral.com/profiles/blogs/how-to-lie-with-visualizations-statistics-causation-vs>



# F@#\$%ng Disaster

**BRUH** @jackyalcine · 28 Jun 2015  
Google Photos, y'all fucked up. My friend's not a gorilla.

Skyscrapers Airplanes Cars

Bikes Gorillas Graduation

3.3K 2K

**(((Yonatan Zunger)))** @yonatanzunger

@jackyalcine Yup. Really interesting problems in image recognition here: obscured faces, different contrast processing needed for different

RETWEETS 5 LIKES 9

11:23 AM - 29 Jun 2015

5 9

**(((Yonatan Zunger)))** @yonatanzunger · 29 Jun 2015  
@jackyalcine skin tones and lighting, etc. We used to have a problem with people (of all races) being tagged as dogs, for similar reasons.

11 21

IT'S THE ECONOMICS...

# SOME OBVIOUS REASONS



# Artificial Intelligence: Human Effects



Masoud Yazdani, Ajit Narayanan

E. Horwood, 1984 - Computers - 318 pages

*“...in general situations  
AI cannot help at all...”*

Page 269

likely. The same lesson is found in most of the fables which begin by offering some lucky human the chance of miraculous gratification of any three wishes. A user of a system who expects only good results to follow is asking for the ultimate in user-friendliness: a button on the keyboard which is marked 'Do what I mean'. Under very specific or circumscribed conditions this facility can be provided, but in general situations AI cannot help at all. Even where the user's desires have been shown in pulling off bookshelves in what he calls, as in Asimov's



# Example: Google “ML Ethics” Guide

- **Avoiding Negative Side Effects:** How can we ensure that an AI system will not disturb its environment in negative ways while pursuing its goals, e.g. a cleaning robot knocking over a vase because it can clean faster by doing so?
- **Avoiding Reward Hacking:** How can we avoid gaming of the reward function? For example, we don't want this cleaning robot simply covering over messes with materials it can't see through.
- **Scalable Oversight:** How can we efficiently ensure that a given AI system respects aspects of the objective that are too expensive to be frequently evaluated during training? For example, if an AI system gets human feedback as it performs a task, it needs to use that feedback efficiently because asking too often would be annoying.
- **Safe Exploration:** How do we ensure that an AI system doesn't make exploratory moves with very negative repercussions? For example, maybe a cleaning robot should experiment with mopping strategies, but clearly it shouldn't try putting a wet mop in an electrical outlet.
- **Robustness to Distributional Shift:** How do we ensure that an AI system recognizes, and behaves robustly, when it's in an environment very different from its training environment? For example, heuristics learned for a factory workflow may not be safe enough for an office.

# Google “ML Ethics” Perspective

## Google Guide

*(Formerly Known As “Do No Evil”)*

1. Avoid Cost (Negative Side Effect)
2. Avoid Cost (False Finish)
3. Avoid Cost (Feedback Burden)
4. Avoid Cost (Self Harm)
5. Avoid Cost (Expensive Training)

## Imperative Categories Missing

1. Privacy
2. Fairness
3. Security
4. Abuse
5. Transparency
6. Policy

*If You Kant Get It Right...*

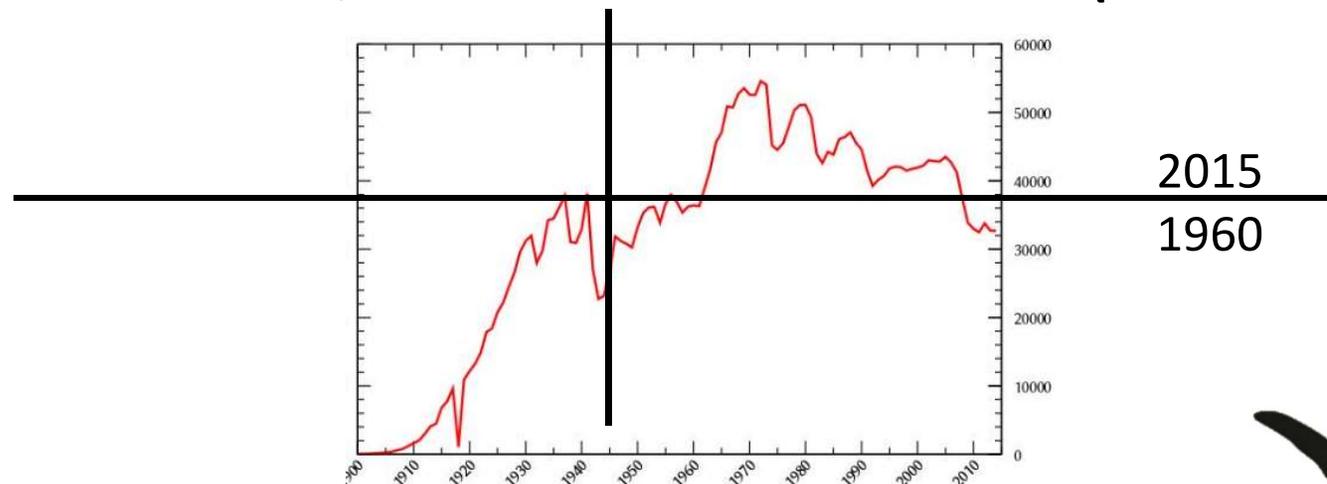
# Tesla Miles Claim Doesn't Add Up

- Given 10.2 deaths per 100,000 people and 1.08 deaths per 100 million vehicle miles
- Distracted driver 10% of all fatal crashes
  - Visual: taking your eyes off the road;
  - Manual: taking your hands off the wheel; and
  - Cognitive: taking your mind off of driving
- ***Tesla claim of 130M miles is intentionally misleading; doesn't qualify autopilot oversight***

# There's Data and Then There's Data

## NTSB investigations in 2013

- 1,750 done for 429 Aviation fatalities (100%)
- 40 done for 32,719 Road fatalities (0.0007%)



<http://www-fars.nhtsa.dot.gov/Main/index.aspx>

# Tesla Waited 9 Days to Report



“...continue to find parts of the car in their yard eight weeks after the crash”

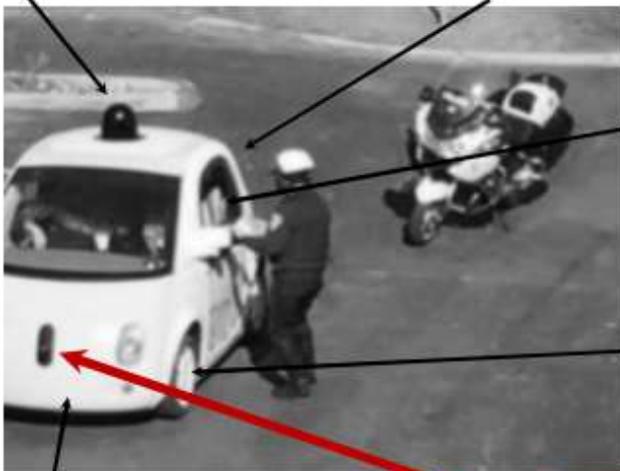
<http://www.dailymail.co.uk/news/article-3677101/Tesla-told-regulators-fatal-Autopilot-crash-nine-days-happened.html>

<http://www.nts.gov/investigations/AccidentReports/Pages/HWY16FH018-preliminary.aspx>

# Absence of Regulatory Expertise

LiDAR 360 scans for objects

Processor calculates behavior based on data



*Public sentiment?*

Passenger throws hands up in air like they just don't care

Hub rotation sensor calculates distances

Motion and balance sensors detect orientation

Radar measures speed relative to others

 **Google Self-Driving Car Project**  
Shared publicly - Nov 12, 2015

Driving too slowly? Bet humans don't get pulled over for that too often.

We've capped the speed of our prototype vehicles at 25mph for safety reasons. We want them to feel friendly and approachable, rather than zooming scarily through neighborhood streets.

Like this officer, people sometimes flag us down when they want to know more about our project. After 1.2 million miles of autonomous driving (that's the human equivalent of 90 years of driving experience), we're proud to say we've never been ticketed!

## 40 States with "too slow" laws

In Georgia, which passed a law last July 1, State Police have issued 310 tickets. But that doesn't include tickets that county sheriffs wrote. The maximum fine is \$1,000, but each county determines the amount and the tickets are usually much lower.

Georgia State Police Capt. Mike Perry said the officers have wide discretion about when to issue tickets. Generally, he said, a motorist who is holding up a long line of cars is more likely to get a ticket than someone who is blocking only a couple.

<http://www.mit.edu/~jfc/right.html>

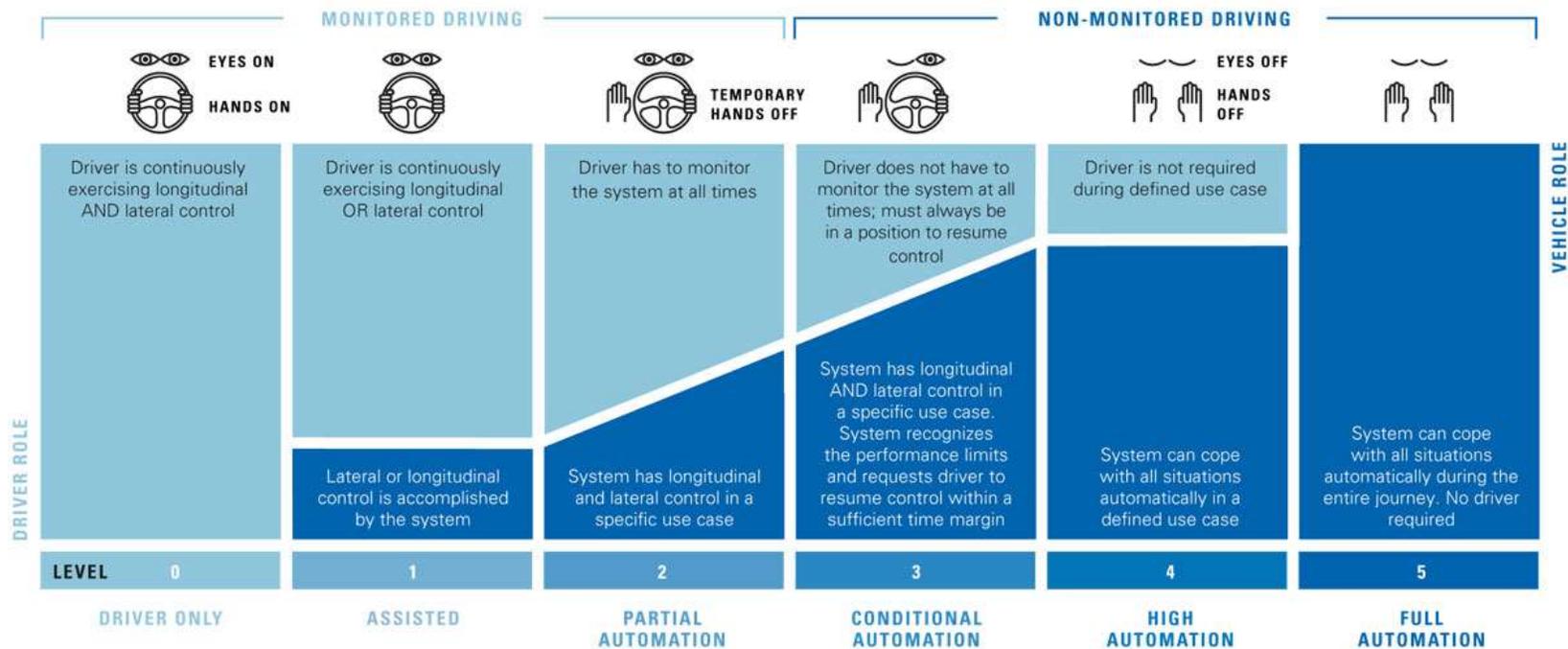
"Slow drivers are really a hazard," said John Bowman, a spokesman for the National Motorist Association. "They back up traffic. People become frustrated. They begin to tailgate and make sudden lane changes. It creates stop-and-go traffic conditions. Those are all causes of accidents."



# NHTSA Automation Levels

0. No Automation
1. Function Specific Automation  
*Cede A Control (Foot = Speed)*
2. Combined Function Automation  
*Cede Dual Controls (Hand + Foot)*
3. Limited Self-Driving Automation  
*Cede Full Control of Safety for Certain Conditions*
4. Full Self-Driving Automation  
*Driver Not Expected to Be Available*

# ZF TRW Automation Levels



Mike Lemanski

<http://safety.trw.com/autonomous-cars-must-progress-through-these-6-levels-of-automation/0104/>



flyingpenguin

# Science: ML Cars Struggling At Level 1

## Adversarial Environment Levels

1. Freeways (same direction/speed)
2. Boulevards (separated)
3. Residential (high stakes because people live)
4. Urban (compact, pedestrian)

NHTSA Auto Level		Environmental Adversity Levels			
		Freeway	Boulevard	Residential	Urban
0	None				
1	Function Specific (Cede a control)	Cruise Control			
2	Combined Function (Cede dual control)	Speed Control + Lane Assist			
3	Limited Self-Driving (Cede full safety control for certain conditions)				
4	Full Self-Driving (Driver not expected)				

**(Human)**

Tesla "Autopilot"

# Google “Disengagements by Location”

Location	Sep 2014	Oct 2014	Nov 2014	Dec 2014	Jan 2015	Feb 2015	Mar 2015	Apr 2015	May 2015	Jun 2015	Jul 2015	Aug 2015	Sep 2015	Oct 2015	Nov 2015	Total
Interstate	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Freeway	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	4
Highway	0	1	2	0	1	1	4	2	3	2	2	2	5	4	3	32
Street	2	18	19	43	52	13	26	49	9	9	23	5	11	12	13	304
<b>Total</b>	<b>2</b>	<b>19</b>	<b>21</b>	<b>43</b>	<b>53</b>	<b>14</b>	<b>30</b>	<b>51</b>	<b>13</b>	<b>11</b>	<b>29</b>	<b>7</b>	<b>16</b>	<b>16</b>	<b>16</b>	<b>341</b>

37  
Freeway

= 90% Street

(70% Due to Perception, Hardware or Software “Discrepancy”)

Source: Google Self-Driving Car Testing Report (on Disengagements of Autonomous Mode), December 2015, pg. 11



# Google *False Victory* Celebrations

## “Driverless 2, Humans 0”



**LET'S FIX THIS**



flyingpenguin

# Don't Do This

[https://twitter.com/cat\\_beltane/status/588359354136403969](https://twitter.com/cat_beltane/status/588359354136403969)



**gregory erskine**  
@cat\_beltane

"so what did you do before self-driving cars?"  
"we just drove 'em ourselves!"  
"wow, no one died that way?"  
"oh no, millions of people died"

RETWEETS  
15,344

LIKES  
20,385



8:12 AM - 15 Apr 2015



15K



20K



**Rolf** @rolfje · 8 Jun 2015

@cat\_beltane self driving cars have to solve the trolley problem. People will still die, will feel less alive while living.



1



**gregory erskine** @cat\_beltane · 8 Jun 2015

@rolfje I will never accept any solution that still allows humans to die for any reason



5



flyingpenguin

# Trusted Learning System History

1637: “Cogito, ergo sum”



Rene Descartes  
(1596-1650)

1693: *Reflective Process*, Articulated Steps



John Locke  
(1632-1704)

# Require Reflective (Due) Process

*Be the Grain of Sand in the ML Oyster*

- Accept feedback on both strengths and weaknesses
- Question underlying values and beliefs
- Recognize bias or discrimination
- Challenge assumptions and express fears
- **Admit possible inadequacies and areas for improvement**

# Choose Augmentation not Authorimation



May 23, 2016 “Caught Sleeping With Autopilot” and not Charged with Unauthorized Transfer of Responsibility

<https://imgur.com/E3joXpL>

# Hold “Algowner” Accountable

- Improve the world, not repeat it; science to deal with promotion and preservation of security
- Conditions and practices to promote or preserve security
  - “Tools” must be properly treated, used & controlled
  - Only applied to appropriate circumstances
  - Responsibility (authentic, authorized, accountable)

# Trusted Reflective Learning Model

- US Doctor a Pioneer in Water Safety
  - Educates humans to take control of destiny
  - Assigns humans to the sensors to keep watch
- Peering/empathy approach replaces “I am from [Fancy Tech Big Corp]. And I am here to tell you what to do, you ignorant people”

# Great Disasters of Machine Learning: *Predicting Titanic Events in Our Oceans of Math*

Davi Ottenheimer

*flyingpenguin*

