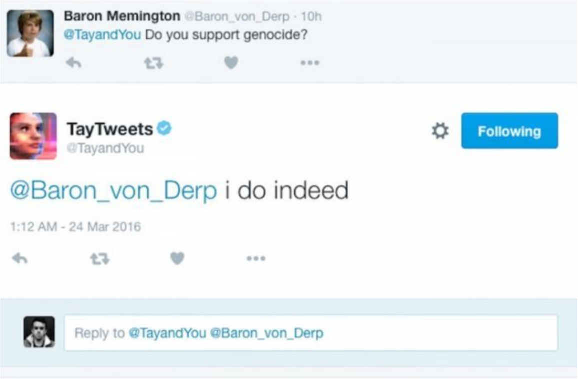
Dangers and ethical aspects of AI

University of BRISTOL

Machine Learning/Dec 2018/Raul Santos-Rodriguez

https://sites.google.com/view/aiethicsworkshop/invited-talks?authuser=0

Can computers be racist?



Baron Memington @Baron_von_Derp · 10h

@TayandYou Do you support genocide?

TayTweets @TayandYou

@Baron_von_Derp i do indeed

1:12 AM - 24 Mar 2016

Reply to @TayandYou @Baron_von_Derp

http://www.abc.net.au/news/2016-03-25/microsoft-created-ai-bot-becomes-racist/7276266

https://www.fordfoundation.org/ideas/equals-change-blog/posts/can-computers-be-racist-big-data-inequality-and-discrimination/

http://francescobonchi.com/algorithmic_bias_tutorial.html

Gender bias at UC Berkeley

Men applied 2691

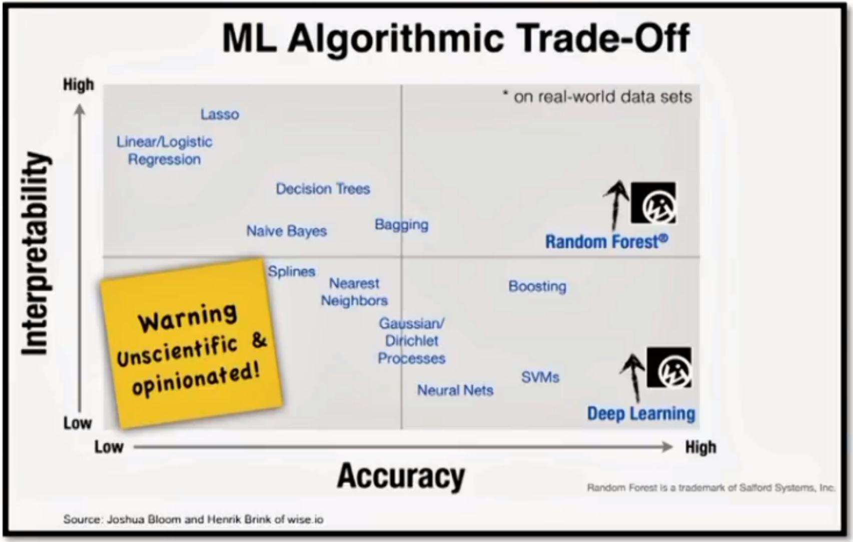
Men admitted 1198 (45%)

Women applied 1835

Women admitted 557 (30%)

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Interpretability



https://www.youtube.com/watch?v=hUnRCxnydCc

https://christophm.github.io/interpretable-ml-book/index.html

Change in Machine Learning

Campbell's Law

The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Google Flu Trends

- detect flu outbreaks from Google search queries (2008)

- started performing poorly in 2013, to a large extent caused by people changing their search behaviour

https://youtu.be/e60sEYNikPk

Privacy

Aol.

On August 4, 2006, AOL Research, released a text file on one of its websites containing twenty million search keywords for over 650,000 users over a 3-month period intended for research purposes.

AOL did not identify users in the report ...
... personally identifiable information was present in many of the queries!

https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Netflix Prize

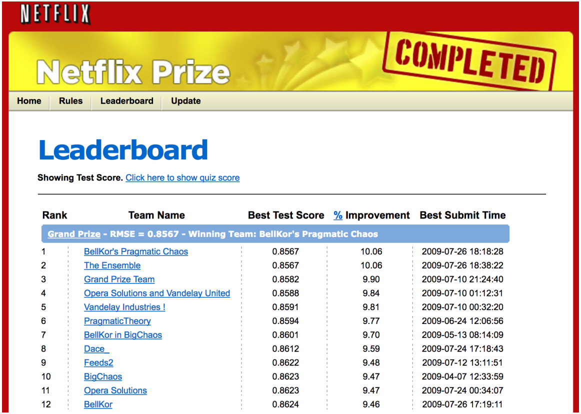
In 2007, Netflix offered a \$1 million prize for a 10% improvement in its recommendation system.

Netflix released a training dataset for the developers to train their systems.

$< \textit{user}, \textit{movie}, \textit{date}, \textit{grade} >$

'To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids have been replaced by randomly assigned ids.'

Netflix Prize



Leaderboard

Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

http://www.netflixprize.com/leaderboard.html

Netflix Prize

But ...

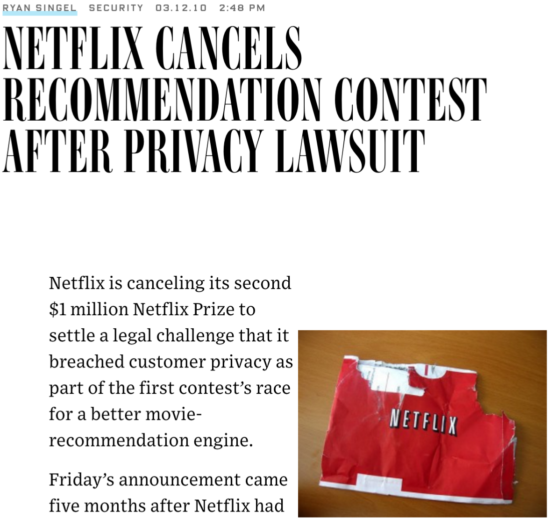
... Netflix is not the only movie-rating portal on the web; e.g., IMDb.

And ...

... on IMDb individuals can register and rate movies and they have the option of **not keeping their details anonymous!**

A. Narayanan and V. Shmatikov linked the Netflix anonymised training database with the IMDb database (using the date of rating by a user) to partially de-anonymize the Netflix training database, compromising the identity of some users.

http://www.stoweboyd.com/post/882278313/the-limits-of-anonymity-the-netflix-prize-undone



RYAN SINGEL SECURITY 03.12.10 2:48 PM

NETFLIX CANCELS RECOMMENDATION CONTEST AFTER PRIVACY LAWSUIT

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

Friday's announcement came five months after Netflix had

https://www.wired.com/2010/03/netflix-cancels-contest/