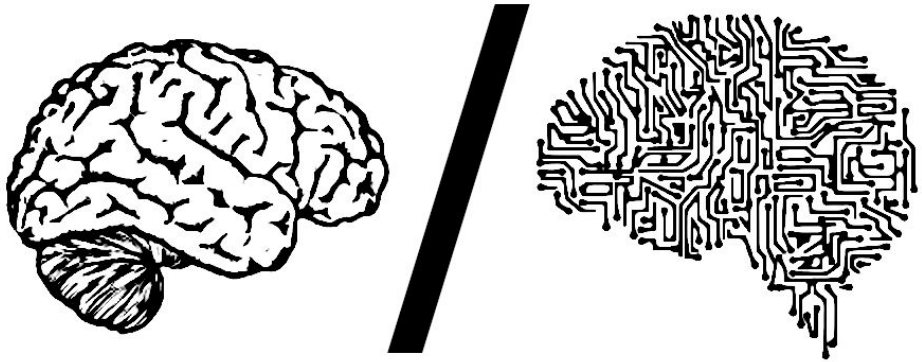
AI i offentlig sektor

Socialchefsdagarna, 2020-09-30

Olof Mogren, PhD
RISE Research Institutes of Sweden

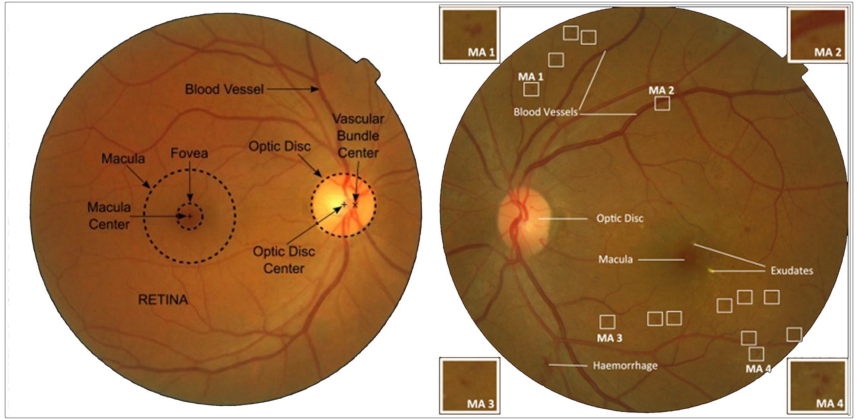
Några genombrott inom AI

- 1995: Deep Blue vs Gary Kasparov (IBM)
- 2012: Bildigenkänning (Krizhevsky et.al.)
- 2013: Ordrepresentationer, (ex. Mikolov, et.al)
- 2015: AlphaGo vs Lee Se-dol (Silver et.al)
- 2017: Språkmodellering, (Vaswani, et.al)
  - Generera språk
  - Översättning
  - Klassificera text
  - Med mera



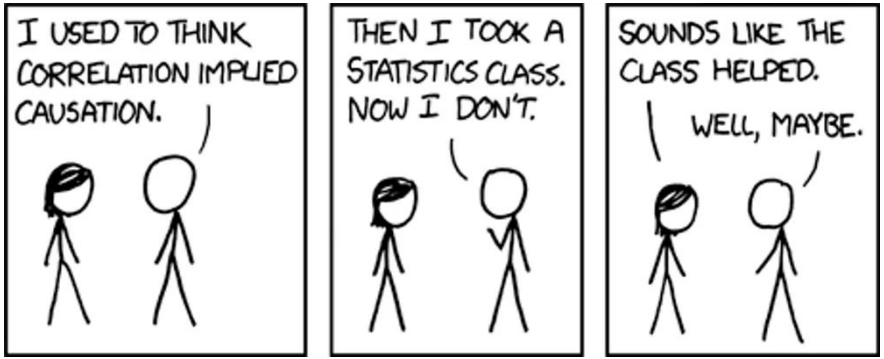
AI är inte som mänsklig intelligens.

Maskiner kan numera mycket



- Köra bil (Volvo, Zenuity, NVIDIA, mfl) *Schmidt-Erfurth, et.al., 2018*
- Förutsäga vissa avancerade mönster (bättre än människor)
- Skriva (nära nog) felfri text (Brown, et.al., 2020)
- Lära sig saker från stora mängder data
- Tagga dina vänner i foton (Facebook, 2011)
- Anomalier i medicinska bilder (ögonfoto, röntgen, etc)
- Betala din kaffe med ett leende (Baidu, Alibaba; Kina)

Maskiner kan (ännu) inte



Munroe, xkcd.com

- Förstå orsakssamband
- Förstå innebörden i text den genererar
- Bry sig om
- Känna
- Avgöra vad i den stora datan som är vettigt, moraliskt, lagligt, mm

“Man is to computer programmer as woman is to homemaker”

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

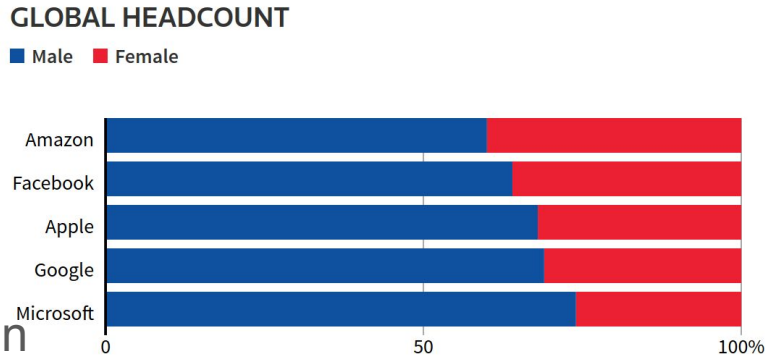
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

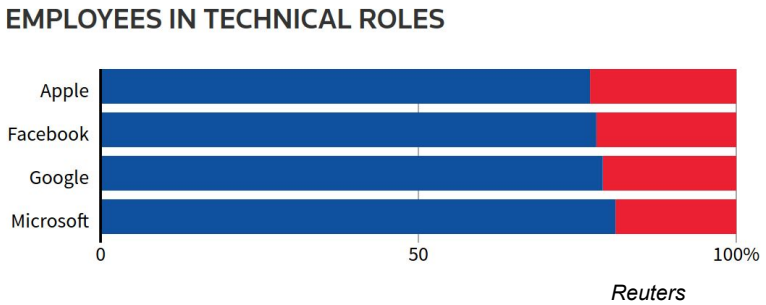
Genus-fördomar i ordrepresentationer

Bolukbasi, et.al., (NeurIPS 2016)

Datadriven AI som beslutsstöd

- Rekrytering (Amazon 2018)
  - Genus-problem
- Microsoft Tay: chatt-bot 2016
  - Togs offline efter 16 timmar
  - Grovt rasistiska uttalanden
- Modellerna speglar fördomar i datan
- Fairness in AI: ett öppet problem





Perspektiv att ta hänsyn till

Social bias

Ex. genus, ras, etc.

Vilka attribut kan vi använda för ett beslut?

Hur kan vi isolera dessa?

Underliggande

Faktorer korrelerar

Hitta underliggande faktorer

Fairness/Rättvisa

Behandlas alla individer rättvist vid ett beslut? (Demografi, genus, etc)

Privacy

Vilken information om mig själv delar jag med andra?

Vilken information använder vi vid ett beslut?

Hur får vi vårt beslutsstöd att reagera på rätt attribut, och inte alla?

RI.
SE

Vad kan vi göra då?

Intelligent filtrering

- “Privacy-bevarande maskininlärning”
- Tränad filtreringsmodul “tar bort” känslig information.

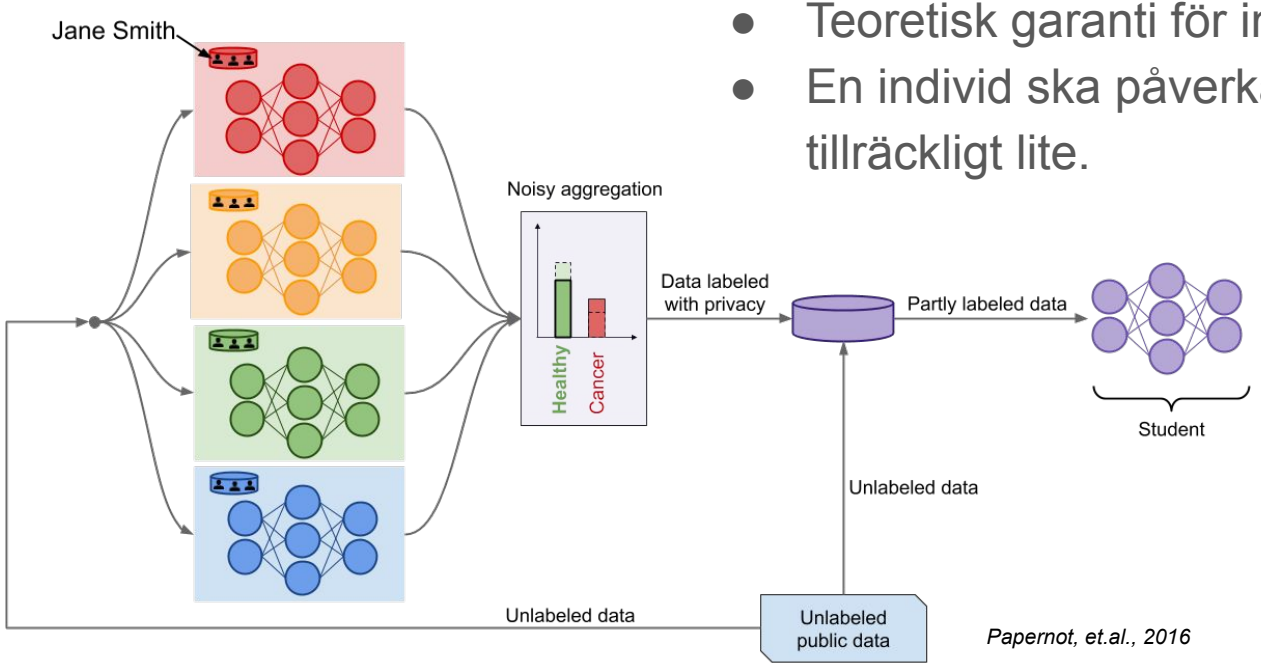
Martinsson, Listo Zec, Gillblad, ***Mogren****, 2020*

Intelligent filtrering

- “Privacy-bevarande maskininlärning”
- Tränad filtreringsmodul “tar bort” känslig information.

Martinsson, Listo Zec, Gillblad, ***Mogren****, 2020*

Differential privacy



- Teoretisk garanti för individer.
- En individ ska påverka resultatet tillräckligt lite.

Papernot, et.al., 2016

AI kommer inte kunna göra vårt jobb

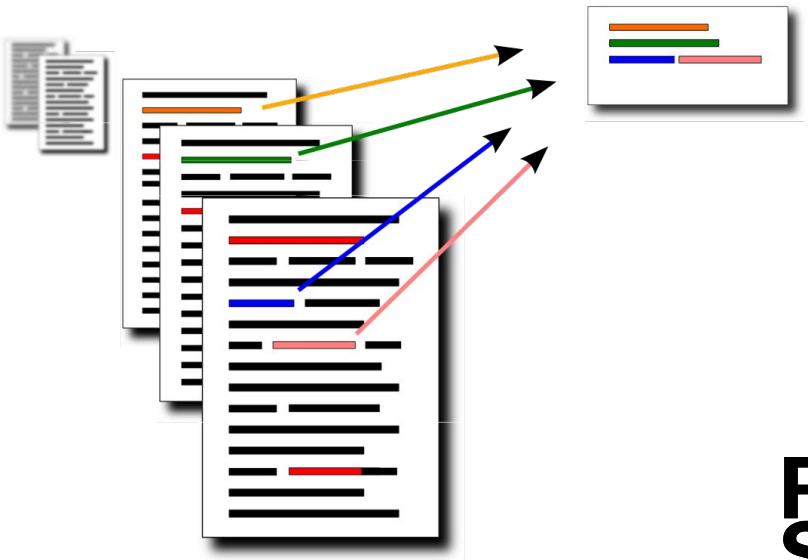
(Än på ett tag).

AI kan dock underlätta vårt jobb

(Redan nu).

Datadriven AI som beslutsstöd

- Att ta rättvisa beslut
  - Bevis behövs innan användning!
- Sammanfatta stora mängder data
- Sortera data
  - Kategorisera dokument
  - Svarsförslag på e-post
  - Hitta rätt mottagare i stor organisation
- Förutsäga händelser
  - Fall-skador för äldre
  - Ekonomisk risk
  - Problem i verksamheten
- Predicera vilka anställda som behöver särskilt stöd och feedback



***Mogren**, 2015*

Tack

Olof Mogren, olof.mogren@ri.se

Referenser

- Brown, et.al., 2020, Language models are few-shot learners. https://arxiv.org/abs/2005.14165
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357). http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119). https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
- Martinsson, J., Listo Zec, E., Gillblad, D.,Mogren, O. (2020) Adversarial representation learning for synthetic replacement of private attributes. https://arxiv.org/abs/2006.08039
- Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K. (2016). Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. https://arxiv.org/abs/1610.05755
- Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., & Bogunović, H. (2018). Artificial intelligence in retina. Progress in retinal and eye research, 67, 1-29., https://www.sciencedirect.com/science/article/pii/S1350946218300119
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
- Silver, D., Huang, A., et.al. (2016) Mastering the game of Go with Deep Neural Networks & Tree Search. Nature. https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

RI.
SE