MVE420:
Maskininlärning och etik

Vilhelm Verendel
2019-05-14

Mina antaganden

- Ni har alla programmerat i ett eller flera programmeringsspråk och skrivit flertalet datorprogram

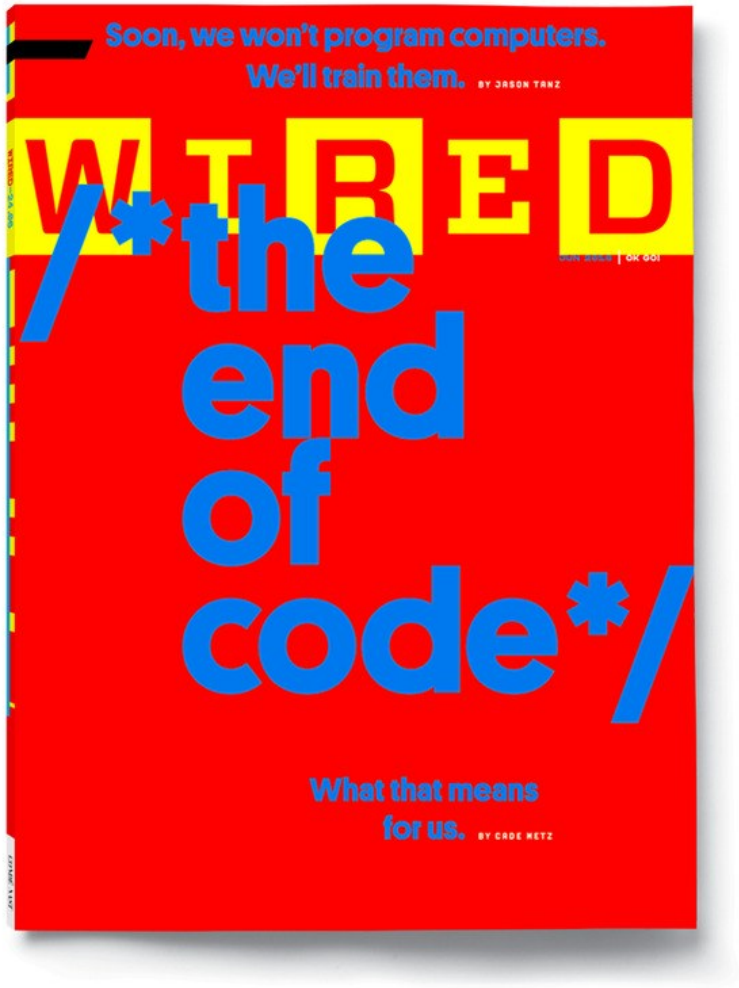
Mina antaganden

- Ni har alla programmerat i ett eller flera programmeringsspråk och skrivit flertalet datorprogram
- Ni vet att programmering är en generell och kraftfull teknologi som kan skapa många framsteg ... och ofta krävs bra programmeringskod

Mina antaganden

- Ni har alla programmerat i ett eller flera programmeringsspråk och skrivit flertalet datorprogram
- Ni vet att programmering är en generell och kraftfull teknologi som kan skapa många framsteg ... och ofta krävs bra programmeringskod
- ... Men kanske finns teknologier med minst lika stor potential att leda till liknande innovation?

Maskininlärning: The end of code? (The Wired, 2016)



Programmera datorer på ett nytt sätt: Träna dem med datamängder

Citat ur artikeln

In this world, the ability to write code has become not just a desirable skill but a language that grants insider status to those who speak it. They have access to what in a more mechanical age would have been called the levers of power.

...

Citat ur artikeln

In this world, the ability to write code has become not just a desirable skill but a language that grants insider status to those who speak it. They have access to what in a more mechanical age would have been called the levers of power.

...

But whether you like this state of affairs or hate it—whether you're a member of the coding elite or someone who barely feels competent to futz with the settings on your phone—don't get used to it. Our machines are starting to speak a different language now, one that even the best coders can't fully understand.

Citat ur artikeln

The implications of an unparsable machine language aren't just philosophical. For the past two decades, learning to code has been one of the surest routes to reliable employment—a fact not lost on all those parents enrolling their kids in after-school code academies.

...

Citat ur artikeln

The implications of an unparsable machine language aren't just philosophical. For the past two decades, learning to code has been one of the surest routes to reliable employment—a fact not lost on all those parents enrolling their kids in after-school code academies.

...

Analysts have already started worrying about the impact of AI on the job market, as machines render old skills irrelevant. Programmers might soon get a taste of what that feels like themselves.

Exempel: Facebook

Låt säga att du just nu postar ett inlägg på Facebook

- Tror du att systemet omedelbart delar detta, längst upp i flödet, med alla dina kontakter?

Undersökning från 2013

Karrie Karahalios på University of Illinois genomförde tidigt en undersökning[1] om Facebooks's nyhetsalgoritm (n=40), och upptäckte att

- 62% av deltagarna var omedvetna om att Facebook påverkar nyhetsflödet
- De trodde att systemet omedelbart skickar ut allt de postar till sina kontakter

[1] Victor Luckerson, *Here's How Facebook's News Feed Actually Works*, 2015, `http://time.com/collection-post/3950525/facebook-news-feed-algorithm/`

Hur skulle ni göra själva?

Hur skulle välja att sprida meddelanden om ni fick programmera en liknande algoritm?

Hur skulle ni göra själva?

Hur skulle välja att sprida meddelanden om ni fick programmera en liknande algoritm?

Diskutera i 2 minuter med 1-2 närmsta bänkgrannar

Några alternativ

- Visa allt för alla i slumpmässig ordning?
- Presentera allt i tidsordning?
- Någon annan strategi?

Några alternativ

- Visa allt för alla i slumpmässig ordning?
- Presentera allt i tidsordning?
- Någon annan strategi?

Ett alternativ till att själv programmera detaljerna är att ge datorn ett väldefinierat mål, och att använda *maskininlärning* för att hitta fram

Några alternativ

- Visa allt för alla i slumpmässig ordning?
- Presentera allt i tidsordning?
- Någon annan strategi?

Ett alternativ till att själv programmera detaljerna är att ge datorn ett väldefinierat mål, och att använda *maskininlärning* för att hitta fram

Ett möjligt mål skulle kunna vara att se till att personer fortsätter vara inloggade, eller att logga in ofta, och anpassa strategin därefter ...

Exempel på maskininlärning

- Problemet kallas sedan länge för *binär klassificering* och förekommer även inom maskininlärning: Att söka och välja ut en funktion som till en vektor av attribut x har ett binärt utfall, så som

$$f(x) \rightarrow \{\text{“intressant”},\text{”ointressant”}\}$$

Exempel på maskininlärning

- Problemet kallas sedan länge för *binär klassificering* och förekommer även inom maskininlärning: Att söka och välja ut en funktion som till en vektor av attribut x har ett binärt utfall, så som

$$f(x) \rightarrow \{\text{“intressant”}, \text{”ointressant”}\}$$

- Tänkbara möjligheter: Klassificera inlägg som “intressant” ifall
  - Avsändaren gillar eller kommenterar dina inlägg
  - Avsändaren själv har skrivit liknande inlägg
  - Liknar dig på andra sätt
  - Har en nära relation till dig
  - ...
  - Kanske en komplicerad blandning av liknande saker?

Exempel på maskininlärning

- Problemet kallas sedan länge för *binär klassificering* och förekommer även inom maskininlärning: Att söka och välja ut en funktion som till en vektor av attribut x har ett binärt utfall, så som

$$f(x) \rightarrow \{\text{“intressant”}, \text{”ointressant”}\}$$

- Tänkbara möjligheter: Klassificera inlägg som “intressant” ifall
  - Avsändaren gillar eller kommenterar dina inlägg
  - Avsändaren själv har skrivit liknande inlägg
  - Liknar dig på andra sätt
  - Har en nära relation till dig
  - ...
  - Kanske en komplicerad blandning av liknande saker?

- Att automatiskt söka och välja ut funktioner (givet en lista av potentiellt viktiga faktorer) är en vanlig fråga inom maskininlärning

Varför använda maskininlärning för sådana fall?

Att välja kriterier för hand är det klassiska sättet att utforma algoritmer som klassificerar: Låt programmerarna välja ut vad som verkar bra, sätt systemet i produktion, och utvärdering sker därefter gradvis...

Varför använda maskininlärning för sådana fall?

Att välja kriterier för hand är det klassiska sättet att utforma algoritmer som klassificerar: Låt programmerarna välja ut vad som verkar bra, sätt systemet i produktion, och utvärdering sker därefter gradvis...

- Big data-ekonomin leder till nya utmaningar: Stora högdimensionella datamängder som beskriver snabba förändringar kommer in i hög takt

Varför använda maskininlärning för sådana fall?

Att välja kriterier för hand är det klassiska sättet att utforma algoritmer som klassificerar: Låt programmerarna välja ut vad som verkar bra, sätt systemet i produktion, och utvärdering sker därefter gradvis...

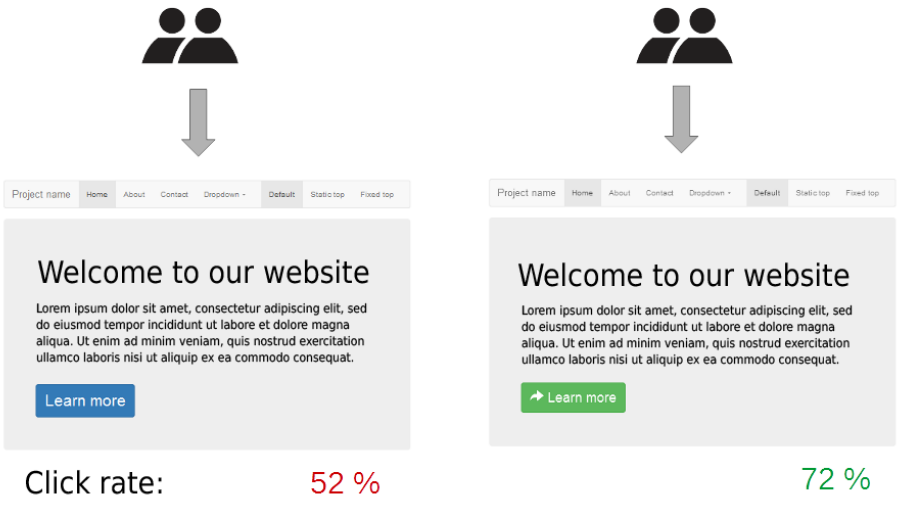
- Big data-ekonomin leder till nya utmaningar: Stora högdimensionella datamängder som beskriver snabba förändringar kommer in i hög takt
- Vi behöver metoder för att snabbt och automatiskt hitta mönster i denna flod av data: Programmerarens roll blir att sätta upp detta?

Varför använda maskininlärning för sådana fall?

Att välja kriterier för hand är det klassiska sättet att utforma algoritmer som klassificerar: Låt programmerarna välja ut vad som verkar bra, sätt systemet i produktion, och utvärdering sker därefter gradvis...

- Big data-ekonomin leder till nya utmaningar: Stora högdimensionella datamängder som beskriver snabba förändringar kommer in i hög takt
- Vi behöver metoder för att snabbt och automatiskt hitta mönster i denna flod av data: Programmerarens roll blir att sätta upp detta?
- En möjlighet är att använda metoder från maskininlärning: Men hur ger vi datorn ett mål? I vissa fall finns naturliga mål

Enkelt exempel: Design av en websida



Enkelt exempel 2: Annonsplacering

Websidor med hög besöksfrekvens kan göra små variationer i annonser automatiskt, t ex vad gäller

- Färg
- Storlek på teckensnitt
- Andra parametrar så som motiv, beskrivning...

Enkelt exempel 2: Annonsplacering

Websidor med hög besöksfrekvens kan göra små variationer i annonser automatiskt, t ex vad gäller

- Färg
- Storlek på teckensnitt
- Andra parametrar så som motiv, beskrivning...

Det stora antalet besökare gör att

- Annonser kan provas snabbt på tusentals personer
- Det är rimligt att förlita sig på statistiska metoder för att välja form av annons
- Gradvis kan den vinnande annonsen väljas ut, eventuellt efter anpassning till användarprofilen

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

- Vad är en relevant funktion att uppskatta? Beror på tillämpningen:

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

- Vad är en relevant funktion att uppskatta? Beror på tillämpningen:
  - Visa meddelanden (x_i) relaterat till längden på användarens uppmärksamhet (y_i) som proxy för intresse ...

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

- Vad är en relevant funktion att uppskatta? Beror på tillämpningen:
  - Visa meddelanden (x_i) relaterat till längden på användarens uppmärksamhet (y_i) som proxy för intresse ...
  - Visa annonser (x_i) som har en hög chans att få klick (y_i) ...

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

- Vad är en relevant funktion att uppskatta? Beror på tillämpningen:
  - Visa meddelanden (x_i) relaterat till längden på användarens uppmärksamhet (y_i) som proxy för intresse ...
  - Visa annonser (x_i) som har en hög chans att få klick (y_i) ...
  - ... Kräver att vi definierar vad vi önskar att systemet kan göra

Maskininlärning: Övervakat lärande

- Den mest välstuderade formen av maskininlärning: Observationer uppmärkta med resultat (x_i, y_i) möjliggör sökning efter samband:

$$y_i = f(x_i)$$

  Vi känner inte till f, men söker uppskatta f med hjälp av uppmärkt data

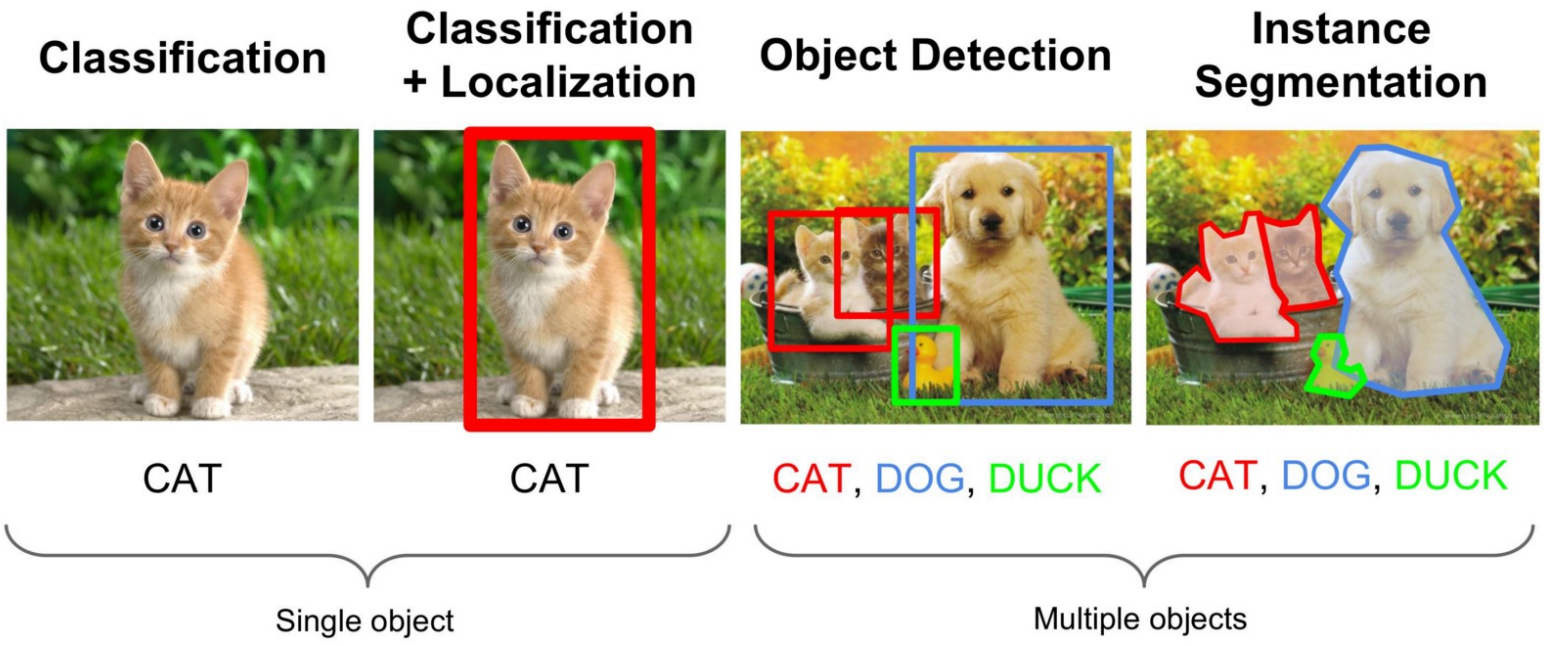
- Vad är en relevant funktion att uppskatta? Beror på tillämpningen:
  - Visa meddelanden (x_i) relaterat till längden på användarens uppmärksamhet (y_i) som proxy för intresse ...
  - Visa annonser (x_i) som har en hög chans att få klick (y_i) ...
  - ... Kräver att vi definierar vad vi önskar att systemet kan göra
- Ofta är datorimplementationen av f mycket komplicerad...

Teknikområden för maskininlärning

Specifika teknikområden där maskininlärning används i allt ökande grad:

1. Bildanalys och objektklassificering
2. Röst- och taligenkänning
3. Automatisk textanalys
4. Självkörande bilar
5. Automatiserade datorspel som schack och Go
6. ...

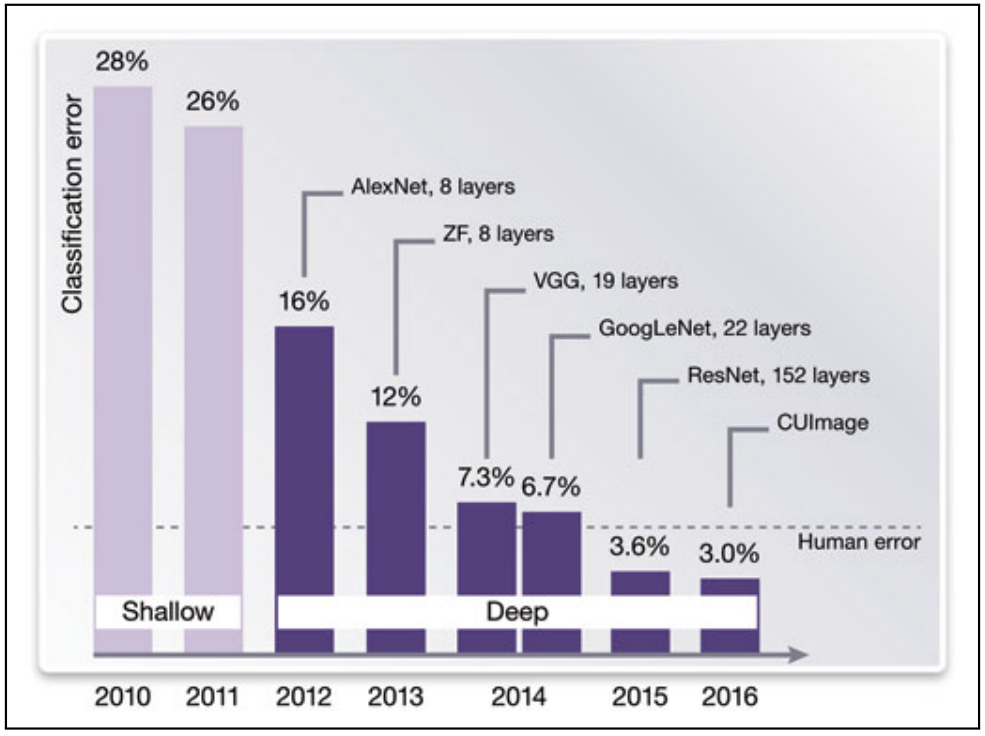
Olika delar av bildanalys



De stora förbättringarna kan förstås som en framgång för djupa neurala nätverk

ML för bildanalys och objektklassificering

Resultat på *ImageNet competition*:



Framgångarna beror på stora databaser som innehåller miljontals bilder med uppmärkt data

Vad är Deep Learning?

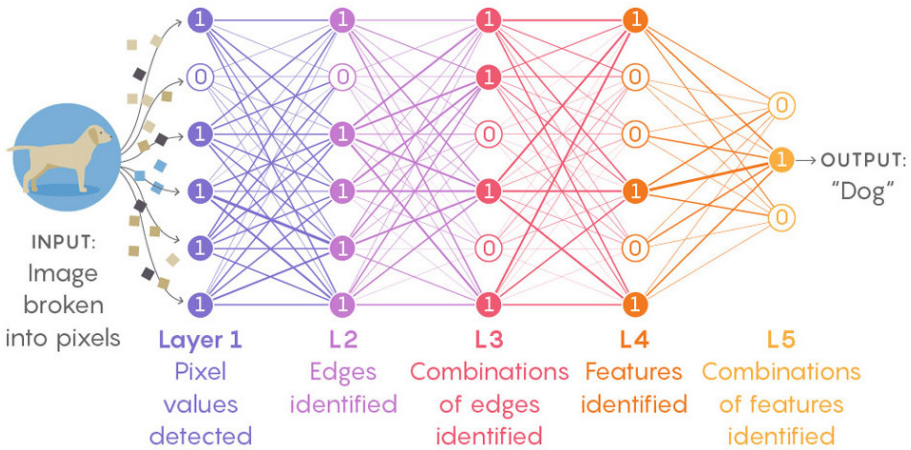
En modell inspirerad av hjärnan

[2] www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec3.pdf

Vad är Deep Learning?

En modell inspirerad av hjärnan

Komplicerad representation: $y_i = f(x_i)$

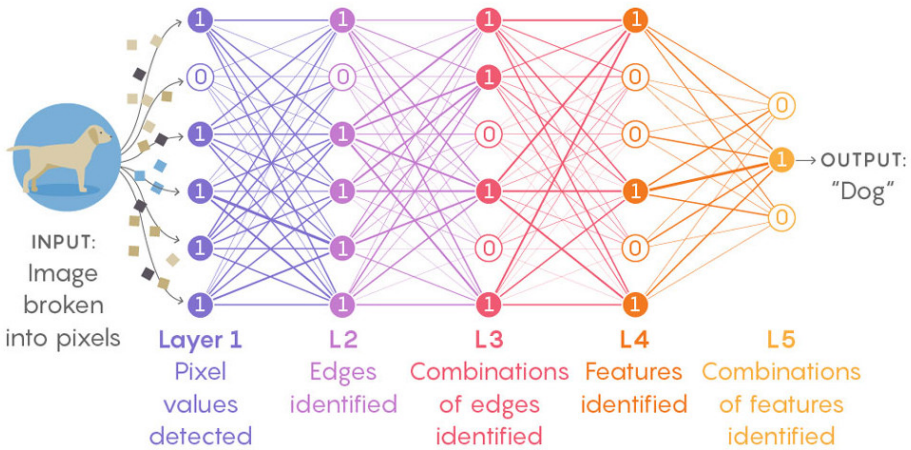


[2] www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec3.pdf

Vad är Deep Learning?

En modell inspirerad av hjärnan

Komplicerad representation: $y_i = f(x_i)$



Modellparametrar:

- Alla olika kopplingar är parametrar i en matematisk modell
- De neurala nätverk som fungerar bra har ofta tiotals/hundratals lager av artificiella neuroner, totalt med miljontals parametrar
- Parametrar väljs ofta med optimeringsmetoden *back propagation*[2]

[2] `www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec3.pdf`

Begränsningar med Deep Learning

- En begränsning i deep learning är svårigheten att sammanfatta och förklara resultaten på ett enkelt sätt

Begränsningar med Deep Learning

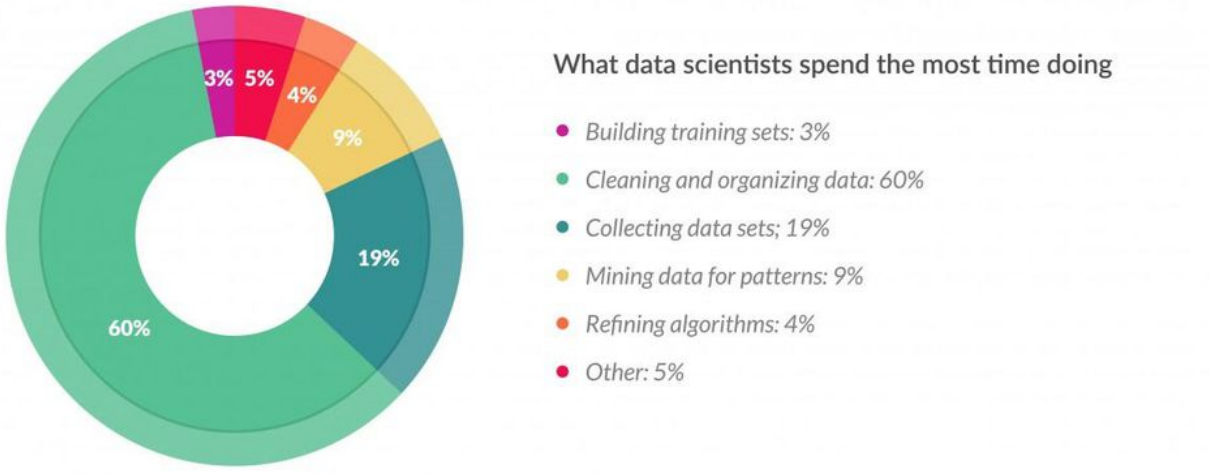
- En begränsning i deep learning är svårigheten att sammanfatta och förklara resultaten på ett enkelt sätt
- Bör vi lita på metoderna utan några garantier, i säkerhetskritiska system?

Begränsningar med Deep Learning

- En begränsning i deep learning är svårigheten att sammanfatta och förklara resultaten på ett enkelt sätt
- Bör vi lita på metoderna utan några garantier, i säkerhetskritiska system?
- Fooling Image Recognition with Adversarial Examples:

`https://www.youtube.com/watch?v=piYnd_wYlT8`

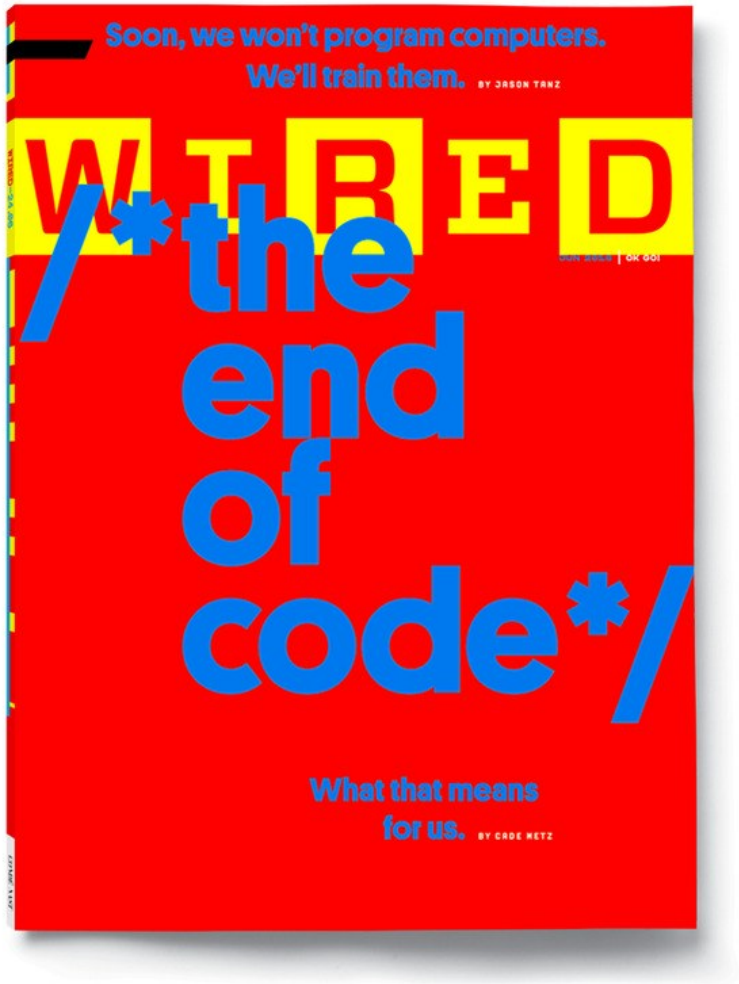
Ersätta programmerare? Finns några bra motargument?



Källa: Forbes/CrowdFlower survey of data scientists, 2016

Liknande resultat i undersökning gjord 2017

Ersätta programmerare? Finns några bra motargument? (2)



Källa: The Wired, 2016

Ersätta programmerare? Finns några bra motargument? (2)

Just as Newtonian physics wasn't obviated by the discovery of quantum mechanics, code will remain a powerful, if incomplete, tool set to explore the world. But when it comes to powering specific functions, machine learning will do the bulk of the work for us.

...

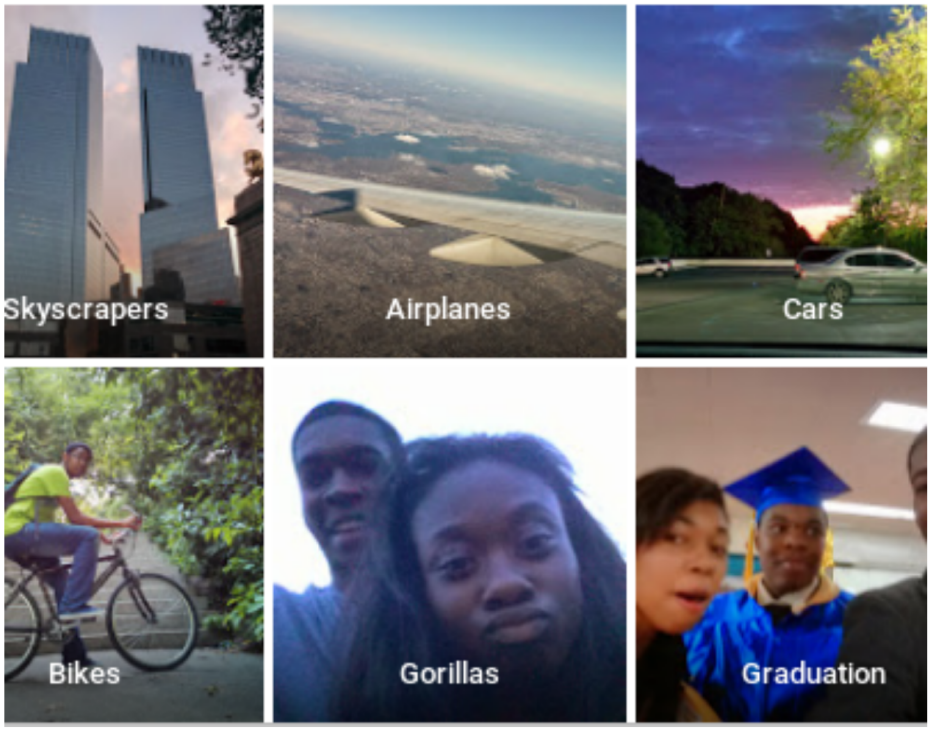
Ersätta programmerare? Finns några bra motargument? (2)

Just as Newtonian physics wasn't obviated by the discovery of quantum mechanics, code will remain a powerful, if incomplete, tool set to explore the world. But when it comes to powering specific functions, machine learning will do the bulk of the work for us.

...

Already the companies that build this stuff find it behaving in ways that are hard to govern. Last summer, Google rushed to apologize when its photo recognition engine started tagging images of black people as gorillas. The company's blunt first fix was to keep the system from labeling anything as a gorilla.

Ersätta programmerare? Finns några bra motargument? (2)



In 2015, a black software developer named Jacky Alciné **revealed** that the image classifier used by Google Photos was labeling black people as "gorillas."

Google apologized profusely and set to work on the bug. Two years later, Google has simply erased gorillas (and, it seems, chimps and monkeys) from the lexicon of labels its image classifier can apply or be searched with, rendering these animals unsearchable and in some sense invisible to the AI that powers Google's image searching capabilities.

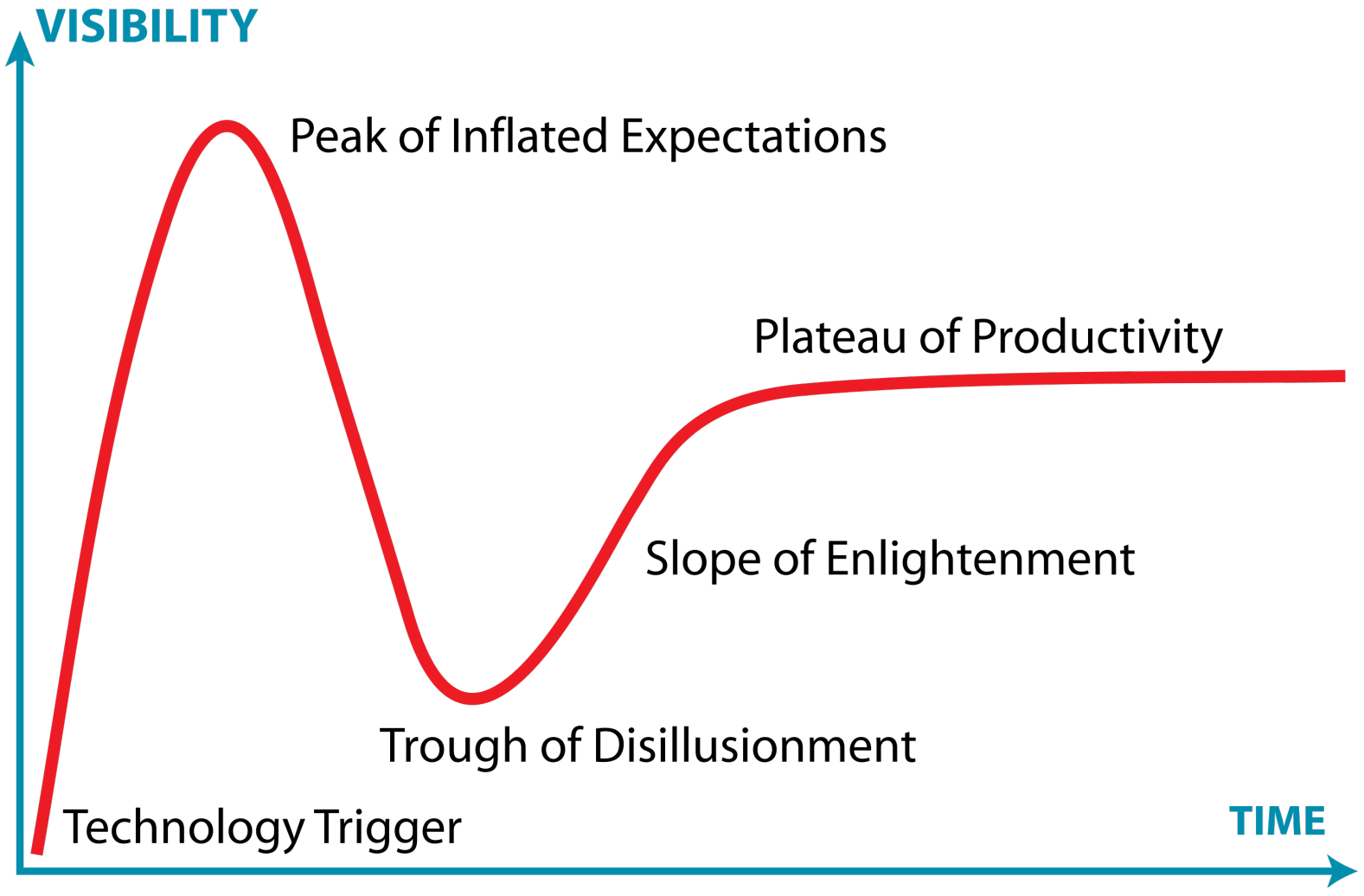
Källa: Boing boing, 2018-01-11

Höga förväntningar: Kanske för höga?

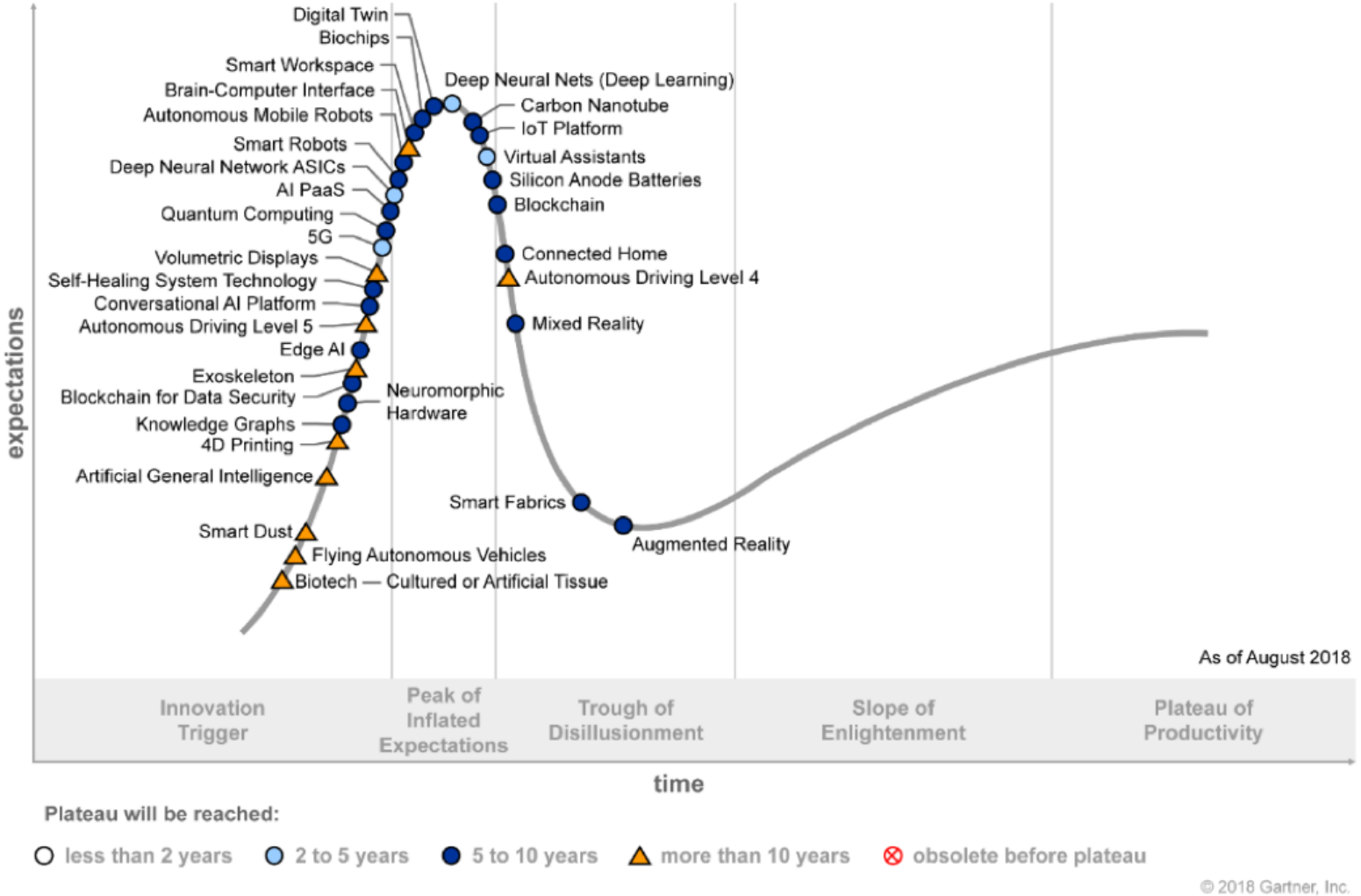
"Machine learning will be the engine of global growth"
– Erik Brynjolfsson, Financial times, July 2018

The Hype Cycle

(Source: Gartner)



The Hype Cycle: 2018 (Source: Gartner)



Självkörade bilar: Tidiga tecken på AI-hype

- Uber, Tesla, och andra företag har tagit tillbaka signifikanta löften de senaste åren
- Förhoppningar inkluderade t ex att självkörande bilar enkelt skall kunna transportera från kust till kust i USA (under 2017 och 2018), vilket inte skedde
- Självkörande bilar verkar vara svårare att utveckla än vad man trott (det finns många ovanliga specialfall)

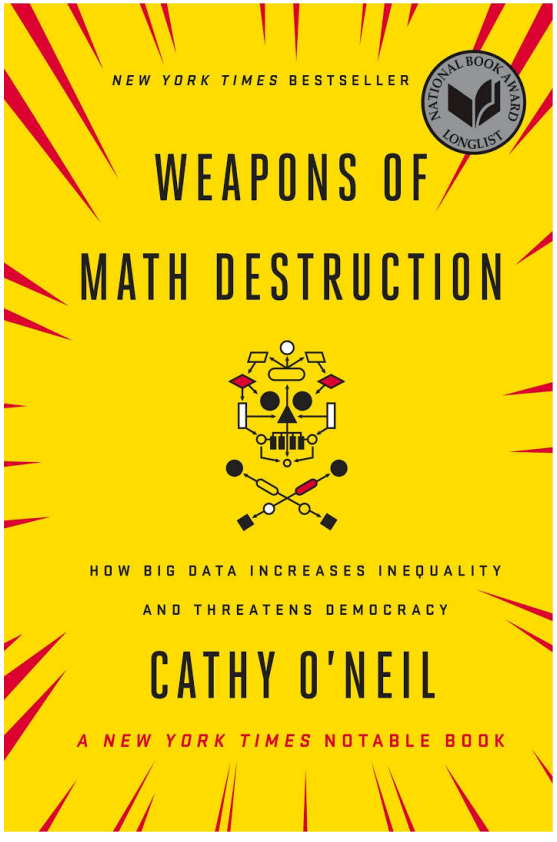
Källor:

https://www.theverge.com/2018/8/1/17641186/tesla-elon-musk-self-driving-coast-to-coast-delay

https://www.npr.org/2018/07/31/634331593/

uber-parks-its-self-driving-truck-project-saying-it-will-push-for-autonomous-car

Transparens: Bok som problematiserar bilden



O'Neil problematiserar maskininlärning med icke-transparenta modeller

Weapons of Math Destruction?

The math-powered applications powering the data economy were based on choices made by fallible human beings. Some of these choices were no doubt made with the best intentions. Nevertheless, many of these models encoded human prejudice, misunderstanding, and bias into the software systems that increasingly managed our lives.

... Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, were beyond dispute or appeal. And they tended to punish the poor and the oppressed in our society, while making the rich richer.

I came up with a name for these harmful kinds of models: Weapons of Math Destruction, or WMDs for short.

Facebook och maskininlärning: Cathy O'Neil skriver

As you know by now, I am outraged by all sorts of WMDs. So let's imagine that I decide to launch a campaign for tougher regulations on them, and I post a petition on my Facebook page. Which of my friends will see it on their news feed?

I have no idea. As soon as I hit send, that petition belongs to Facebook, and the social network's algorithm makes a judment about how to best use it. It calculates the odds that it will appeal to each of my friends. Some of them, it knows, often sign petitions, and perhaps share them with their own networks. Others tend to scroll right past. At the same time, a number of my friends pay more attention to me and tend to click the articles I post. The Facebook algorithm takes all of this into account as it decides who will see my petition. For many of my friends, it will be buried so low on their news feed that they'll never see it.

En aktuell fråga

Support The Guardian Subscribe Find a job Sign in Search
International edition
The Guardian
News Opinion Sport Culture Lifestyle More
UK ▸ UK politics Education Media Society Law Scotland Wales Northern Ireland

Cambridge Analytica

Far more than 87m Facebook users had data compromised, MPs told

Former Cambridge Analytica employee gives evidence before parliamentary committee

Alex Hern
@alexhern
Tue 17 Apr 2018 13.23 BST

715

▲ Brittany Kaiser said Cambridge Analytica had a suite of personality quizzes designed to extract personal data from the social network. Photograph: HANDOUT/Reuters

Far more than 87 million people may have had their Facebook data harvested by Cambridge Analytica, according to evidence from former employee Brittany Kaiser.

Speaking to the Commons digital, culture, media and sport select committee, Kaiser said Cambridge Analytica had a suite of personality quizzes designed to extract personal data from the social network, of which Aleksandr Kogan's This Is Your Digital Life app was just one example.

Advertisement


Exempel på där maskininlärning används

O'Neils bok problematiserar hur algoritmer och maskininlärning i ökande grad används för

- Reklam: Många sidor på Internet
- Politiska kampanjer: Anpassar budskapet efter mottagare
- Arbetsmarknaden: Förstärker ojämlikheter
- Finansiella beslut, Försäkringsbolag, Brottsbekämpande myndigheter...

O'Neil menar: Maskininlärning ger oss fördelar, men även många nya problem (O'Neils bok fokuserar på den sidan)

Relevant för den här kursen

Kunna se och diskutera några exempel på nya etiska dilemman som kan uppstå av att maskininlärning tillämpas ute i samhället

Exempel: Privata preferenser[3]

UK world sport football opinion culture business lifestyle fashion environment tech travel

home › tech

Artificial intelligence (AI)

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions

This article is 2 months old

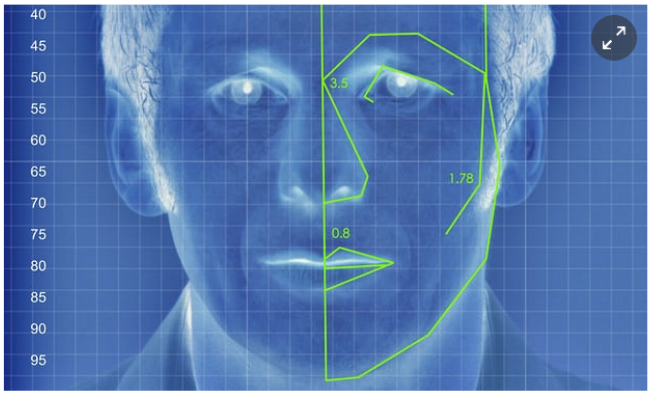
59,261

Sam Levin in San Francisco

@SamTLevin

email

Friday 8 September 2017 00.46 BST



An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

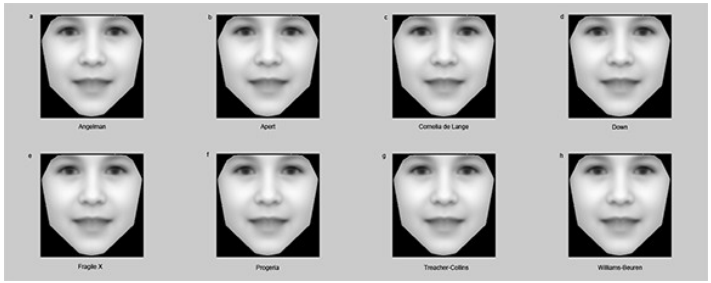
Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.

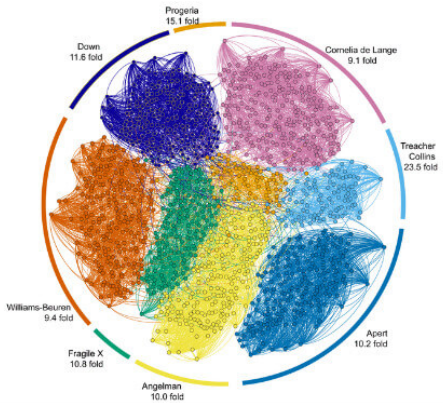
The study from Stanford University - which found that a computer algorithm could correctly distinguish between gay and straight men 81% of the time, and 74% for women - has raised questions about the biological origins of sexual orientation, the ethics of facial-detection technology, and the potential for this kind of software to violate people's privacy or be abused for anti-LGBT purposes.

The machine intelligence tested in the research, which was published in the Journal of Personality and Social Psychology and first reported in the Economist, was based on a sample of more than 35,000 facial images that men and women

[3] Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, 2017. `https://osf.io/zn79k/`

Exempel: Ovanliga sjukdomar[4]





[4]Diagnostically relevant facial gestalt information from ordinary photos, eLife, Journal of Personality and Social Psychology, 2017.
`https://dx.doi.org/10.7554/eLife.02020`

Exempel: Självkörande bilar

JACK STEWART TRANSPORTATION 03.30.18 10:34 PM

TESLA'S AUTOPILOT WAS INVOLVED IN ANOTHER DEADLY CAR CRASH

TESLA

TESLA NOW HAS another fatality to hang on its semi-autonomous driving system. The company just revealed that its Autopilot feature was turned on when a Model X SUV slammed into a concrete highway lane divider and burst into flames on the morning of Friday, March 23. The driver, Wei Huang, died shortly afterwards at the hospital.

Moraliska dilemman

Definition på *moraliskt dilemma*: När två eller flera etiska aspekter kommer i konflikt med varandra

Moraliska dilemman

Definition på *moraliskt dilemma*: När två eller flera etiska aspekter kommer i konflikt med varandra

Två vanliga aspekter som kommer i konflikt

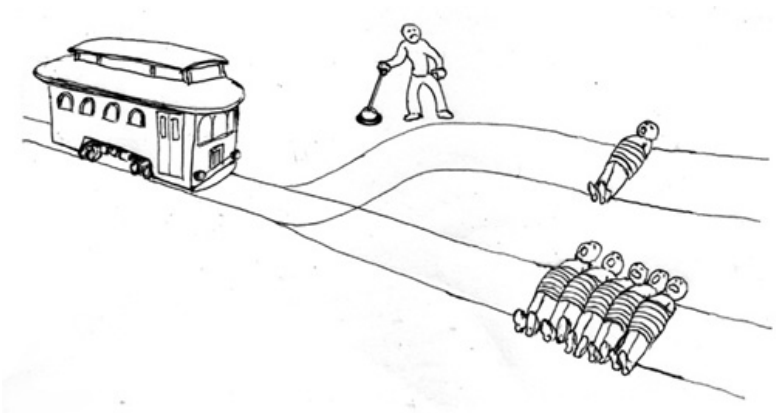
- Den totala nyttan för alla
- Individens frihet, plikter och rättigheter

Moraliska dilemman

Definition på *moraliskt dilemma*: När två eller flera etiska aspekter kommer i konflikt med varandra

Två vanliga aspekter som kommer i konflikt

- Den totala nyttan för alla
- Individens frihet, plikter och rättigheter



Diskussion: Är detta moraliska dilemman?

1. Ansiktsigenkänning: Klassificera sexualitet
2. Klassificera ovanliga sjukdomar

Diskutera i smågrupper i 2 minuter: Finns några moraliska aspekter i dessa två exempel som skulle kunna komma i konflikt med varandra?

Transparens i maskininlärningsmodeller

Bra modeller kan vara svårbegripliga

Transparens i maskininlärningsmodeller

Bra modeller kan vara svårbegripliga

Fyra möjliga skäl till det kan vara svårt att få insyn i modeller

1. Tillgång till data: Integritetsmässigt komplicerat med privatpersoners data, eller begränsad tillgång av andra skäl
2. Vi kanske inte ser begränsningarna i hur data samlades in
3. Tillgång till modellen: Proprietära system med företagshemligheter
4. Komplicerad statistisk modell: Många modeller är tillräckligt komplicerade att programmerarna/matematikerna själva inte förstår varför resultatet fungerar så bra

En annan kategori maskininlärning: Oövervakat lärande

Vi tänker oss att en maskin får någon sekvens av indata $x_1, x_2, x_3, \ldots$: Det skulle kunna vara värden för pixlar i bilder, tal som representerar ljudvågor, text, eller någon annan sorts information, så som produkter i en varukorg eller händelser på sociala medier.

En annan kategori maskininlärning: Oövervakat lärande

Vi tänker oss att en maskin får någon sekvens av indata $x_1, x_2, x_3, \ldots$: Det skulle kunna vara värden för pixlar i bilder, tal som representerar ljudvågor, text, eller någon annan sorts information, så som produkter i en varukorg eller händelser på sociala medier.

Det passar ofta att dela upp maskininlärning i två kategorier:

En annan kategori maskininlärning: Oövervakat lärande

Vi tänker oss att en maskin får någon sekvens av indata $x_1, x_2, x_3, \ldots$: Det skulle kunna vara värden för pixlar i bilder, tal som representerar ljudvågor, text, eller någon annan sorts information, så som produkter i en varukorg eller händelser på sociala medier.

Det passar ofta att dela upp maskininlärning i två kategorier:

1. *Övervakat lärande*: Maskinen får även ett antal önskade utdata $y_1, y_2, y_3, \ldots$ kopplat till indata: Målet för maskinen är att hitta en representation av sambandet $y_i = f(x_i)$ som ger rätt utdata för nya indata

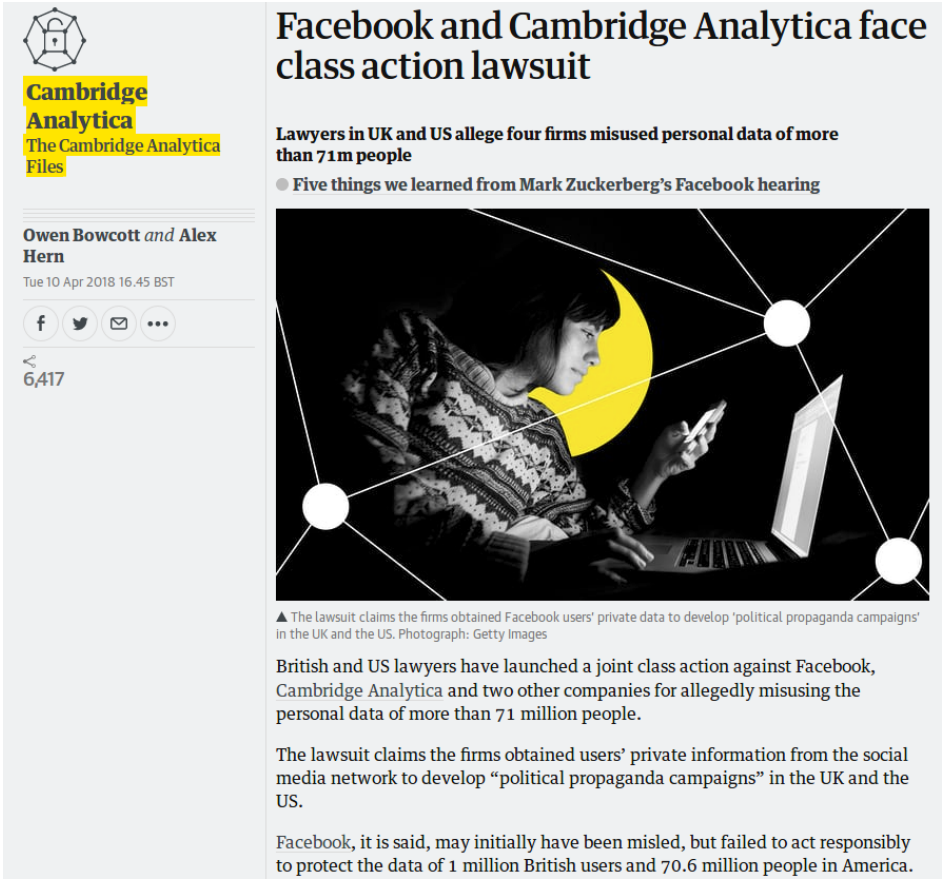
En annan kategori maskininlärning: Oövervakat lärande

Vi tänker oss att en maskin får någon sekvens av indata $x_1, x_2, x_3, \ldots$: Det skulle kunna vara värden för pixlar i bilder, tal som representerar ljudvågor, text, eller någon annan sorts information, så som produkter i en varukorg eller händelser på sociala medier.

Det passar ofta att dela upp maskininlärning i två kategorier:

1. *Övervakat lärande*: Maskinen får även ett antal önskade utdata $y_1, y_2, y_3, \ldots$ kopplat till indata: Målet för maskinen är att hitta en representation av sambandet $y_i = f(x_i)$ som ger rätt utdata för nya indata
2. *Oövervakat lärande*: Systemet får ingen information om vad som är rätt respons (om det ens är meningsfullt): Men maskinen kan ändå tjäna på att hitta mönster i data för att ha en (i) Effektiv representation av indata, (ii) Prediktera framtida indata, (iii) Gruppera och summera indata. (iv) Anomalidetektering.

Storskalig effekt av maskininlärning?



Cambridge Analytica

The Cambridge Analytica Files

Facebook and Cambridge Analytica face class action lawsuit

Lawyers in UK and US allege four firms misused personal data of more than 71m people

Five things we learned from Mark Zuckerberg's Facebook hearing

Owen Bowcott *and* Alex Hern

Tue 10 Apr 2018 16.45 BST

6,417

▲ The lawsuit claims the firms obtained Facebook users' private data to develop 'political propaganda campaigns' in the UK and the US. Photograph: Getty Images

British and US lawyers have launched a joint class action against Facebook, Cambridge Analytica and two other companies for allegedly misusing the personal data of more than 71 million people.

The lawsuit claims the firms obtained users' private information from the social media network to develop "political propaganda campaigns" in the UK and the US.

Facebook, it is said, may initially have been misled, but failed to act responsibly to protect the data of 1 million British users and 70.6 million people in America.

Denna fråga kan förstås utifrån *både* övervakat och oövervakat lärande ...
Men är Facebook verkligen en politiskt relevant plattform?

Facebook: Politisk maktfaktor?

- Facebook har ihop med forskare själva släppt vissa studier
- Tillgång till över en miljard användare: Vad som är släppt som forskningen är troligtvis toppen på isberget
- En bra fråga är: Skulle Facebook och Google kunna påverka politiken?

Studie 1: Påverka valdeltagande via vänner[5]

Forskningsfråga: Kan Facebook öka valdeltagandet?

- Valen i USA 2010 och 2012: Små förändringar i gränssnittet
  Försöket nådde ut till över 61 miljoner amerikaner
- Använde ett verktyg de kallade "Voter megaphone", där användare klickade i att de röstat
- Metod: Undersök grupptryckets effekt på att gå och rösta (genom att tala om för andra att man röstat)

Forskarnas uppskattning är att valdeltagandet ökade med 340,000 personer

[5] Bond et al., *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, Nature, 2012

Studie 2: Påverka valdeltagande med nyheter[6]

Forskningsfråga: Kan politiska nyheter från vänner ändra politiskt beteende?

- Valet 2012: Romney vs Obama
- Två miljoner personer fick fler nyheteslänkar jämfört med annat (så som bilder på hundar och katter)
- Metod: Be personer att självrapportera om de röstat eller inte

Forskarnas uppskattning är att valdeltagandet i denna grupp gick upp från 64 till 67 procent

[6] Bond et al., *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, Nature, 2012

Studie 3: Påverkan via sökmotorers ranking[7]

Forskningsfråga: Kan en sökmotors konfiguration påverka hur personer röstar?

- Två Google-forskare (Robert Epstein och Ronald E. Robertson) bad osäkra väljare i USA och Indien att använda en sökmotor för att bilda sig en uppfattning om valet
- Sökmotorerna var specialprogrammerade till att framhäva ett parti framför ett annat

Forskarnas uppskattning är att röstningspreferenserna kunde ändras i runt 20% av fallen

[7] Bakshy et al., *Exposure to Ideologically Diverse News and Opinion*, Science, 2015

Studie 3: Kritik

- Inget talar för att Google faktiskt försöker påverka valutgången
- 73% av amerikander trodde 2012 att sökresultaten är både korrekta och opartiska[8]

[8] Purcell et al., *Search Engine Use 2012*, Pew Research Center, 2012

Inga tecken på att företagen själva använder detta

Cathy O'Neil kallar inte dessa algoritmer för *WMDs*, men menar att potentialen finns för missbruk...

> I wouldn't yet call Facebook or Google's algorithms political WMDs, because I have no evidence that the companies are using their networks to cause harm. Still, the potential for abuse is vast. The drama occurs in code and behind imposing firewalls. And as we'll see, these technologies can place each of us into our own cozy political nook.

Utesluter inte att andra kan använda plattformen

Cathy O'Neil reflekterar

In *The Selling of the President*, which followed Richard Nixon's 1968 campign, the journalist Joe McGinniss introduced readers to the political operatives working to market the presidential candidate like a consumer good. By using focus groups, Nixon's campaign was able to hone his pitch for different regions and demographics.

...

Cathy O'Neil reflekterar (forts.)

The convergence of Big Data and consumer marketing now provides politicians with far more powerful tools. They can target microgroups of citizens for both votes and money and appeal to each of them with a meticulously honed message, one that no one else is likely to see. It might be a banner on Facebook or a fund-raising email.

...

Cathy O'Neil reflekterar (forts.)

The convergence of Big Data and consumer marketing now provides politicians with far more powerful tools. They can target microgroups of citizens for both votes and money and appeal to each of them with a meticulously honed message, one that no one else is likely to see. It might be a banner on Facebook or a fund-raising email.

...

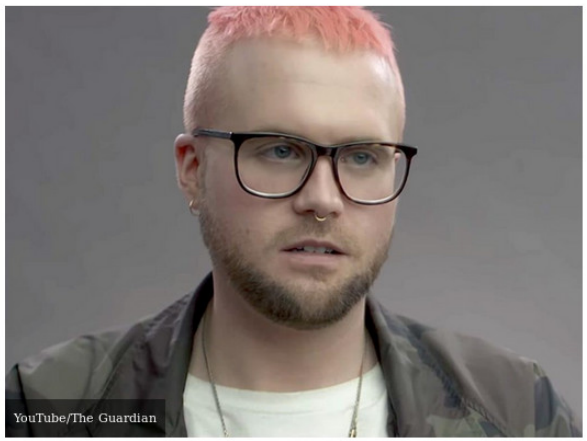
But each one allows candidates to quietly sell multiple versions of themselves and it's anyone's guess which version will show up for work after inaugeration.

Storskalig påverkan via Facebook[9]

Here's the personality test Cambridge Analytica had millions of Facebook users take

Follow @BiNordic Follow @BINordic 2,484 followers

Erin Brodwin | 19 Mar 2018 9:01 PM | 1216



- **Reports published this weekend from The New York Times and The Observer revealed that personality-profiling company Cambridge Analytica harvested data from millions of users and potentially used it in the most recent US presidential election.**
- **The personality test that the company used gives users a score called their "OCEAN" score, referring to how it calculates their performance on a measure of Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.**
- **The test is available for free online at the University of Cambridge Psychometrics Center and forms the basis of many existing psychological studies on happiness and longevity.**

A personality-profiling company called Cambridge Analytica harvested personal data from millions of users and may have used that data to sway voters during the 2016 US presidential election, according to reports published this weekend from The New York Times and The Observer.

[9] http://nordic.businessinsider.com/facebook-personality-test-cambridge-analytica-data-trump-election-2018-3?

Psykografiska modeller

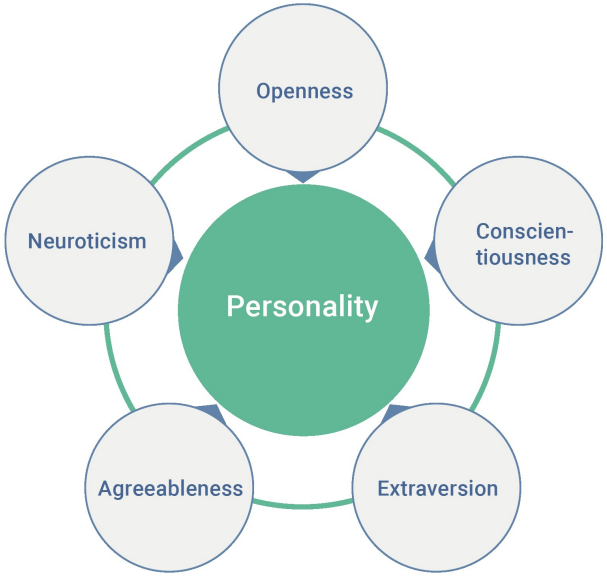
- Av det som framkommit har man använt något de kallat "psykografiska modeller"
- Har sitt ursprung i modeller av människors personlighet
- Upptäcktes ursprungligen med statistiska metoder som inom maskininlärning kan kallas för *oövervakat lärande*

Personlighet

- Personlighet: Stabila dispositioner över situation och tid

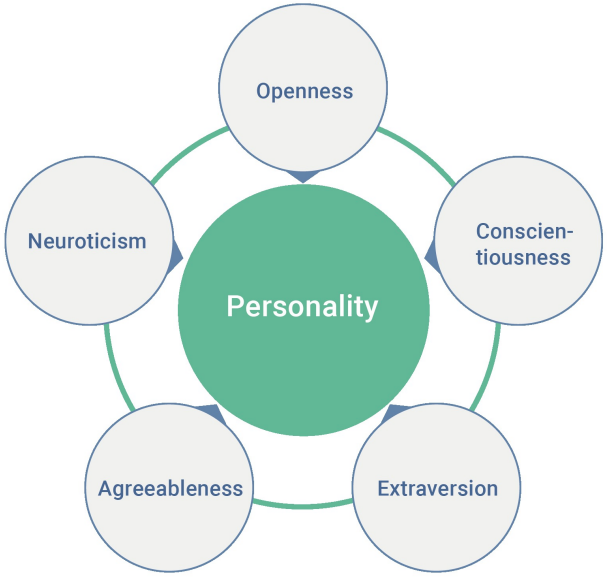
Personlighet

- Personlighet: Stabila dispositioner över situation och tid
- En vanlig modell: OCEAN, delar grovt sett upp personer i $2^5 = 32$ typer



Personlighet

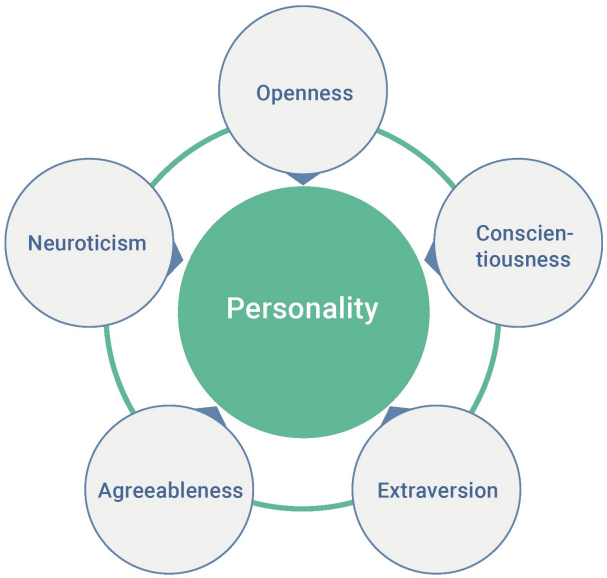
- Personlighet: Stabila dispositioner över situation och tid
- En vanlig modell: OCEAN, delar grovt sett upp personer i $2^5 = 32$ typer



- Varför dessa faktorer? Kommer från *faktoranalys*, en form av oövervakat lärande för att hitta struktur i data

Personlighet

- Personlighet: Stabila dispositioner över situation och tid
- En vanlig modell: OCEAN, delar grovt sett upp personer i $2^5 = 32$ typer



- Varför dessa faktorer? Kommer från *faktoranalys*, en form av oövervakat lärande för att hitta struktur i data
- Tillämpat på stora insamlade datamängder: Frågebatterier om egenskaper hos personer - vilka sorters svar samvarierar?

Kopplingen till övervakat lärande

En möjlig väg har varit att:

- Be en mindre population av personer att på olika sätt göra personlighetstester (gruppera person i som personlighetsklass y_i från deras OCEAN-värden), samla samtidigt in annan information som de har uppgett i sina profiler (x_i)

Kopplingen till övervakat lärande

En möjlig väg har varit att:

- Be en mindre population av personer att på olika sätt göra personlighetstester (gruppera person i som personlighetsklass y_i från deras OCEAN-värden), samla samtidigt in annan information som de har uppgett i sina profiler (x_i)
- Använd övervakad inlärning för att uppskatta sambandet $y_i = f(x_i)$

Kopplingen till övervakat lärande

En möjlig väg har varit att:

- Be en mindre population av personer att på olika sätt göra personlighetstester (gruppera person i som personlighetsklass y_i från deras OCEAN-värden), samla samtidigt in annan information som de har uppgett i sina profiler (x_i)
- Använd övervakad inlärning för att uppskatta sambandet $y_i = f(x_i)$
- Samla in likadana x för en mycket större population (miljoner personer, utan direkt vetskap)

Kopplingen till övervakat lärande

En möjlig väg har varit att:

- Be en mindre population av personer att på olika sätt göra personlighetstester (gruppera person i som personlighetsklass y_i från deras OCEAN-värden), samla samtidigt in annan information som de har uppgett i sina profiler (x_i)
- Använd övervakad inlärning för att uppskatta sambandet $y_i = f(x_i)$
- Samla in likadana x för en mycket större population (miljoner personer, utan direkt vetskap)
- Tillämpa $f(x)$ för varje individ, och skicka sedan politisk information som har framgång hos en viss personlighetstyp...

Sammanfattning

- Maskininlärning är ett nytt sätt att programmera datorer: Söker fram mönster med hjälp av stora mängder data
- Individer, företag, men också myndigheter ställs inför nya moraliska dilemman med nya tillämpningar av maskininlärning
- Tekniken är fortfarande i ett tidigt skede, så det kan vara svårt att veta hur väl vi kommer att förstå metoderna på sikt

Vad bör ni ha lärt er?

- Vad maskininlärning är: Två former av maskininlärning
- Definition på moraliskt dilemma
- Känna till några nya moraliska dilemman relaterade till användandet av maskininlärning
- Fyra skäl till att det kan vara svårt få insyn i maskininlärningsmodeller

Kontakt: vive@chalmers.se