

“Maskinlæring” for Dataanalyse

Eksempler på Anvendelser, Close-up på Spamfilter og Google PageRank

ROBERT JENSSEN

INSTITUTT FOR FYSIKK OG TEKNOLOGI
UNIVERSITETET I TROMSØ

Hovedpunkter

- Data og informasjon!
- Maskinlæring
- Eksempler ...
- Close-up
 - Bayesian søppelfilter
 - Google PageRank
- The 0 \$ eigenvector



Vanvittig mye data!

■ Enorme mengder data:

- Dokumenter
- Sanger, tale
- Nettsider, epost
- Medisinsk
- Biologisk
- Klima etc.

■ Finne **mønstre**, informasjon!

■ Datagruvedrift. Maskinlæring!



Maskinlæring

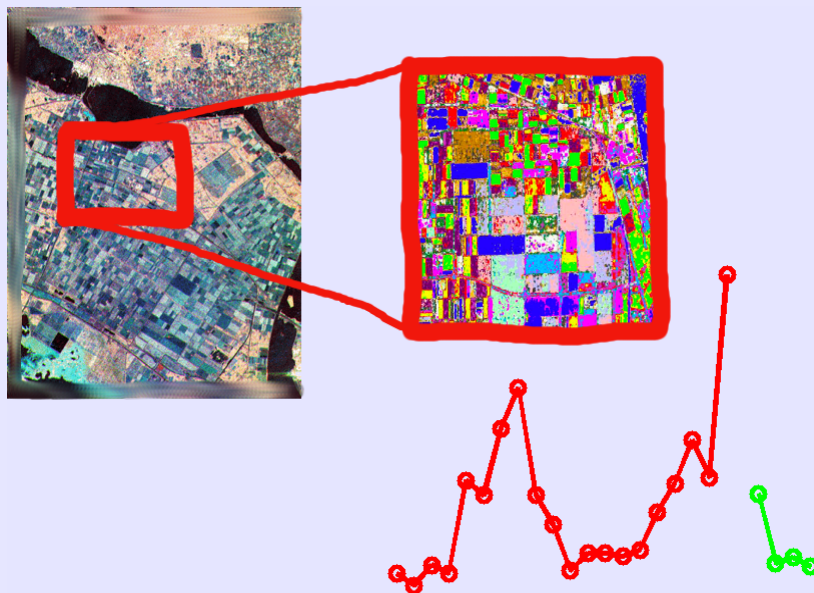
■ Klassifisere
➤ Oppgave...

■ Rangere

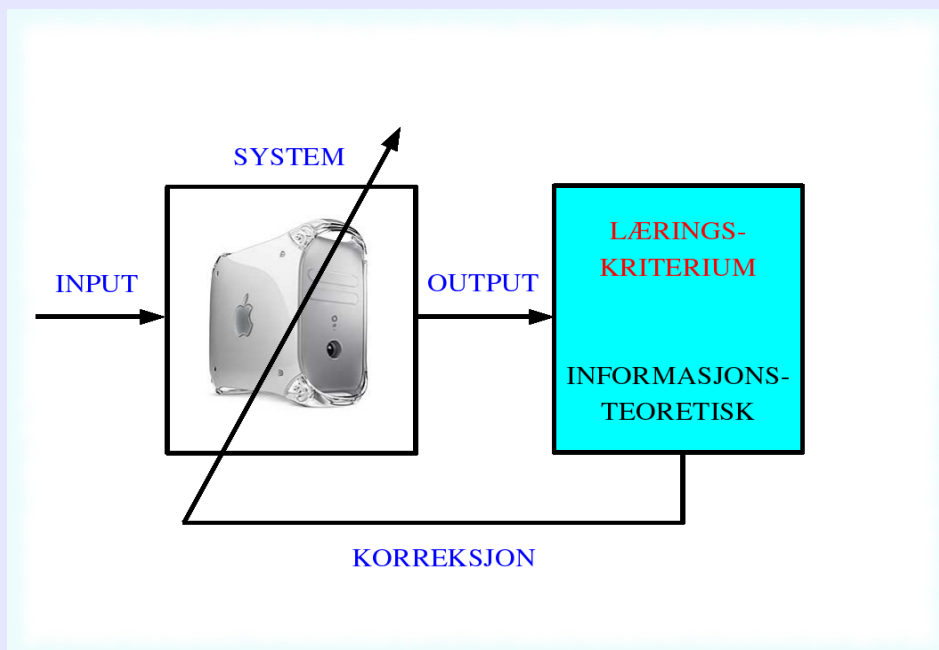
■ Minske

■ Forutsi

■ Selvorganisere



Maskinlæring



Maskinlæring

- Kunstig nevralt nettverk

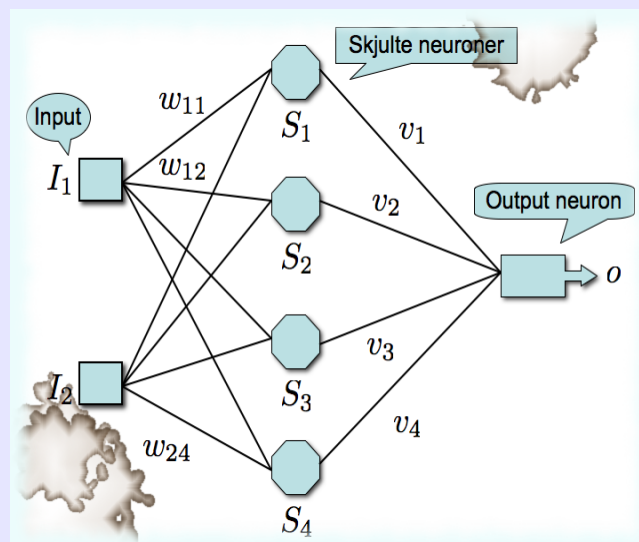
- "Kopi" av hjernen

- Neuroner og synapser

- "Synapse-vektar"

$$w_{11}, w_{12}, \dots, v_1, v_2, \dots$$

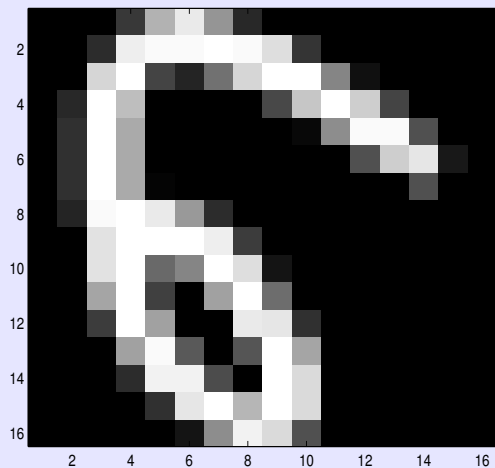
- Må utvikles/læres



Eks.: Nevralt nettverk

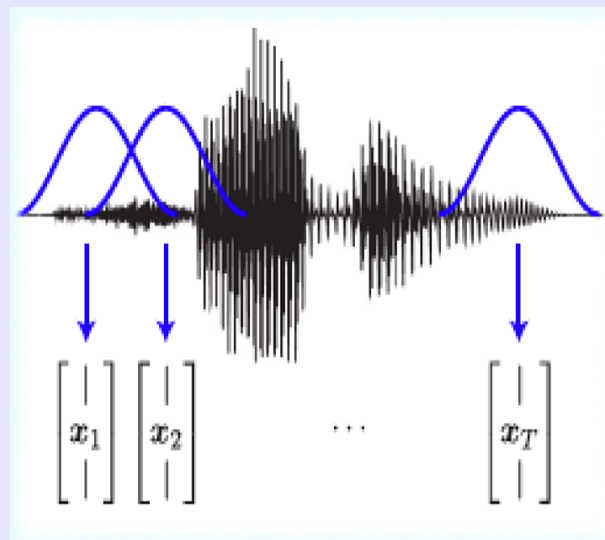
Maskinlæring

- Ofte: Input til maskinen som en rekke av tall per objekt
- Hvordan representere andre typer objekter?



Maskinlæring

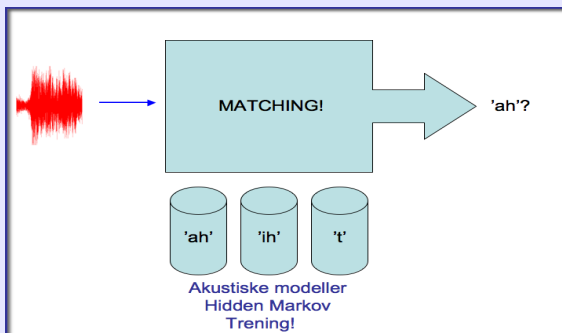
- Lingvistiske byggestener!
- Phoneme (gresk "lyd, tale")
'ih' 'n' 'k' 'r' 'iy' 's' 'ah' 'z'
- Røffly
 - Hvert 10 ms
 - Se på 32 ms
 - Frekvensanalyse!
 - Gir en rekke av tall for hvert vindu.



Tale (og tekst)

■ Talegjenkjenning:

- Kommandoer: Open firefox!
- Diktere dokumenter (f.eks. sykehus)
- Identifikasjon
- Mange bedrifter



Furby

■ Vanvittig populær leke!

■ Kan

➤ Være blid

➤ Bli sur!

➤ Danse og leke

➤ Fortelle vitser

➤ Rape m.m

■ Gjenkjenner tale!

■ "Lærer!"



<http://www.hasbro.com>

Furby

- *Hasbro* om Furby: “Offering the best speech recognition technologies ... includes hidden Markov model-neural nets ...”
- *Newsweek*: “So packed with technology ... if you continue to say one of hundreds of programmed words like "love" and "friend" the FURBY will learn it and repeat it back to you.”
- Kan f.eks. bli fæl til å rape ved positiv oppmerksomhet (“reinforcement learning.”)

Furby

- *Hasbro* om Furby: "Offering the best speech recognition technologies ... includes hidden Markov model-neural nets ..."
- *Newsweek*: "So packed with technology ... if you continue to say one of hundreds of programmed words like "love" and "friend" the FURBY will learn it and repeat it back to you."
- Kan f.eks. bli fæl til å rape ved positiv oppmerksomhet ("reinforcement learning.")
- **Forbudt fra NSA!!** "NSA officials were worried that people would take them home and they'd start talking classified..." (CNN).

Massive software

- Filme et *massivt* slag?!
- Hver “agent” et system som
 - Lærer/selvorganiserer
 - Har “personlighet”
- Filmer & reklame
 - Ringenes Herre, 300, Eragon, King Kong ++
 - Lynx-billions, Nike, Budweiser-wave ++



<http://www.lordoftherings.net>

Spill!

Oversetting

■ Google ++: Maskinlæring for oversetting!

■ Trench: Forente Nasjoners tekstkorpus

And by the way, I think that it is really, really bad that the department of physics and technology is not situated on campus!

And, I think that's really a shame that the Department of Physics and Technology, which is not on campus!

■ Mange bruksområder

➤ Forsvaret

➤ Bruksanvisninger...

**"Displayets
dritinnstillinger"**

I bruksanvisningene hersker et fullstendig språklig sammenbrudd.

<http://www.aftenposten.no>

Sjekker og post

■ Automatisering: Bank og post

■ Handskrevne tall/bokstaver

■ *Helvetica Krankenkasse*

➤ Størst i Sveits

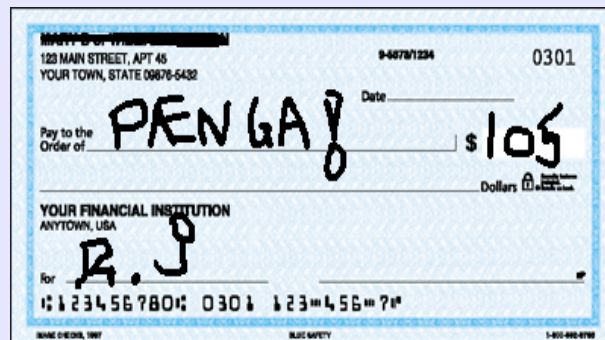
➤ Før: 150 saker daglig

➤ Etter: 400 saker daglig

➤ \$ 2300000 spart årlig

➤ Kostnadseffektivt: 8 mnd

■ 9000 Nordpolen ("Nordlys")



<http://www.imagechecks.com>

Vanskelig felt!
Basert på nevrane nett!
1990-tallet

Bioinformatikk

■ Enormt felt: Maskinlæring + biologi

■ Klassifisere (mikroarray)

➤ Gener

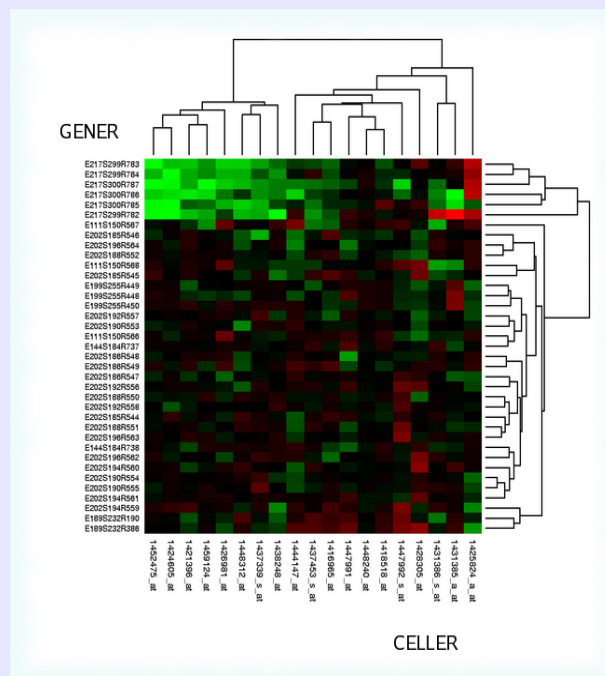
➤ Celler

➤ Proteiner

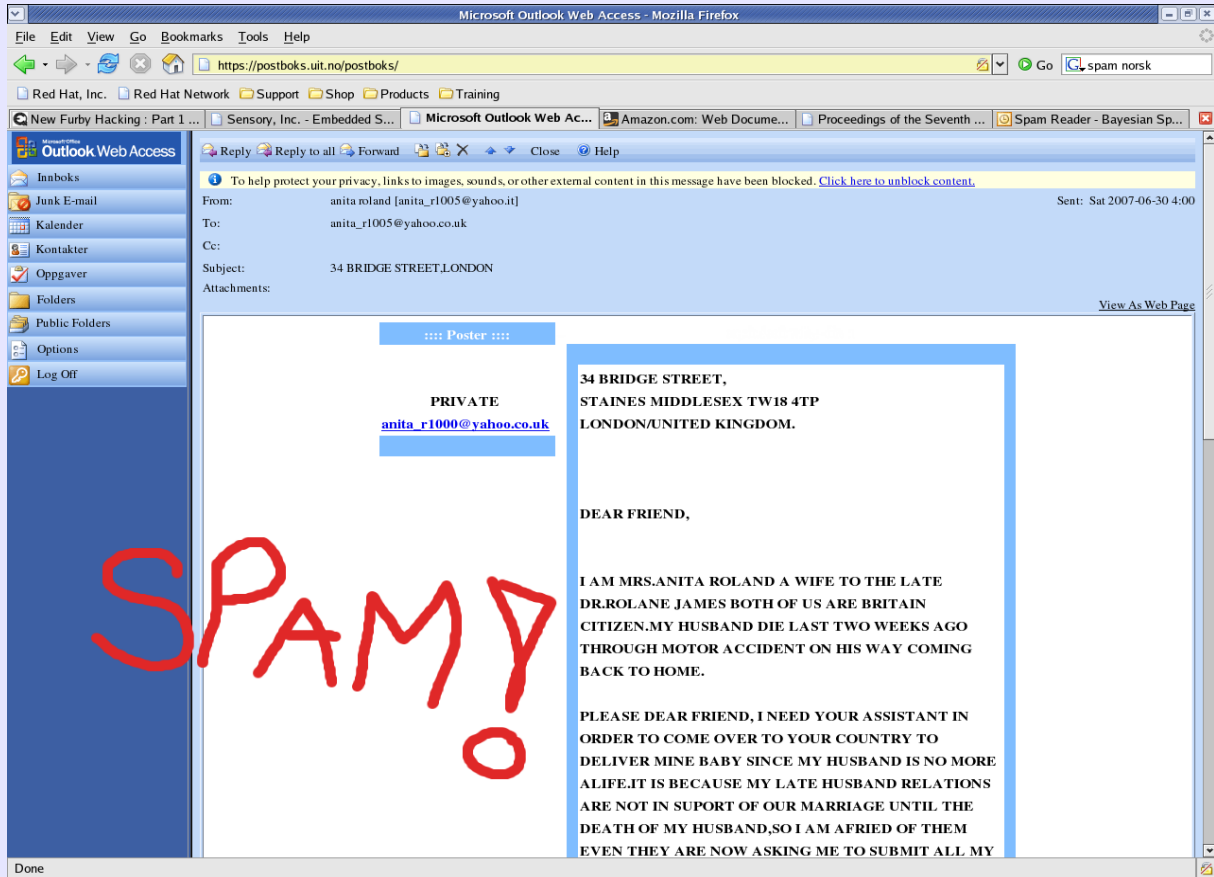
➤ DNA

■ Finne nye medisiner

■ Human Genome Project



LOS
UP



Bayesian søppelfilter

■ Før: Søppel/spam hvis visse ord opptrer: *Dear Friend, Click...*

■ Godt, men ikke godt nok

■ To the rescue: Thomas Bayes!

- o Britisk matematiker (1702-1761)

- o Prest i Skottland (Calvinist)

- o Bayes lov publisert i 1764 (30 år før Tromsø blir by)



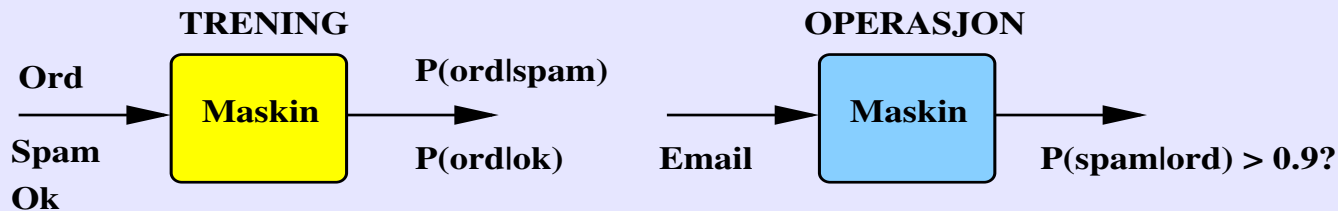
Bayesian søppelfilter

$$P(\text{spam}|\text{viagra}) = \frac{P(\text{viagra}|\text{spam})P(\text{spam})}{P(\text{viagra})}$$

- $P(\text{spam}|\text{viagra})$: Sannsynlighet for at en email er spam gitt at ordet *viagra* finnes i denne emailen.
- $P(\text{viagra}) = P(\text{viagra}|\text{spam})P(\text{spam}) + P(\text{viagra}|\text{ok})P(\text{ok})$
- Maskinlæring for å utnytte dette!

Paul Graham, 2002: A Plan for Spam: “I think it’s possible to stop spam...”

Bayesian søppelfilter



- I bruk i stort sett alle emailprogrammer!
- Kan oppdateres av brukeren
- Ikke bare "vanlige" ord, f.eks $P(\text{spam}|\text{ff0000})$ er høy
- Spammerne: Alife, suport, afried, xxxviagra ++



- Ikke bare matching mot søkeord...
- Rangering av **viktighet** av nettsider! Linkestruktur
- PageRank: Page & Brin, studenter ved Stanford

WWW'98: "In this paper, we present Google..."

Google PageRank

Viktighet **ukjent**: x_1, x_2, x_3 og x_4

❶ Tell “tilbakelenker”

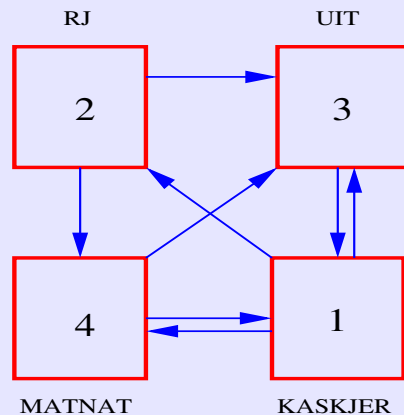
$$x_1 = 2$$

$$x_2 = 1$$

$$x_3 = 3$$

$$x_4 = 2$$

■ Veldig enkelt, men lett å utnytte!



Google PageRank

② "Stemming." Viktige viktigst

$$x_1 = x_3 + x_4$$

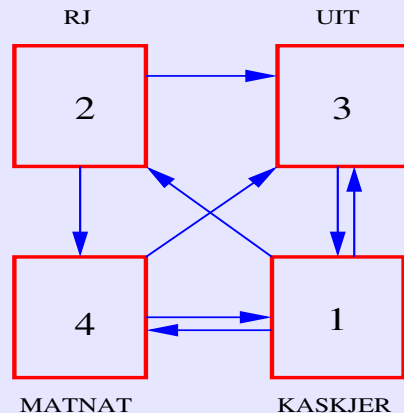
$$x_2 = x_1$$

$$x_3 = x_1 + x_2 + x_4$$

$$x_4 = x_1 + x_2$$

■ Dette er jo et likningssett..! Men det lar seg ikke løse...

■ Ulogisk: Noen "dominerer" ved å "stemme" mye (x_1)



Google PageRank

③ Kun en stemme i sum:

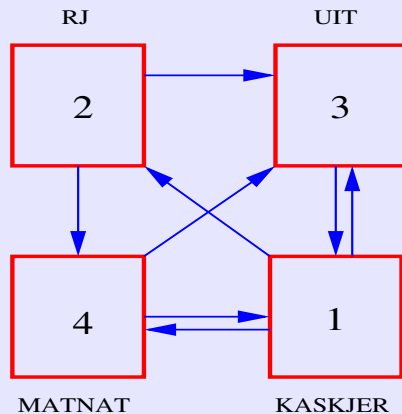
$$x_1 = x_3 + \frac{x_4}{2}$$

$$x_2 = \frac{x_1}{3}$$

$$x_3 = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2}$$

$$x_4 = \frac{x_1}{3} + \frac{x_2}{2}$$

■ Kan vi finne $x_1, \dots, x_4$ som tilfredsstiller dette? JA!



$$x_1 = 0.4$$

$$x_2 = 0.1$$

$$x_3 = 0.3$$

$$x_4 = 0.2$$

Google PageRank

■ Annen form: Tabell **A**

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

■ La

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\begin{aligned} x_3 + \frac{x_4}{2} &= x_1 \\ \frac{x_1}{3} &= x_2 \\ \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2} &= x_3 \\ \frac{x_1}{3} + \frac{x_2}{2} &= x_4 \end{aligned}$$

$$\mathbf{Ax} = 1 \cdot \mathbf{x}$$

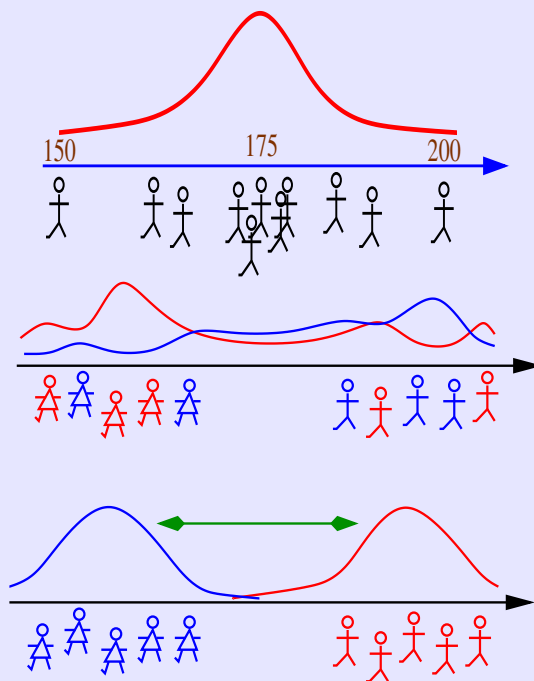
The \$ 25000000000 eigenvector!!

The \$ 0 eigenvector

- Klassifiser slik at en statistisk (informasjonsteoretisk) avstand mellom klassene er størst mulig
- Klassetilhørighet gitt ved x
- Via omveier: Finn x s.a.

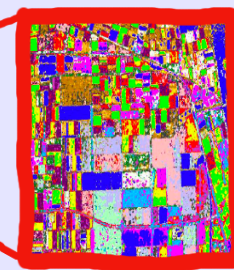
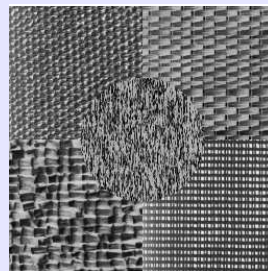
$$Ax = e_{max} \cdot x \quad !$$

A er gitt av avstandsmålet



The \$ 0 eigenvector

- Klassifiseringsmetode i samme “familie” som Google Page Rank og andre metoder brukt f.eks. innen dokumentklassifisering
- Publisert i flere internasjonale tidsskrift og proceedings
- F.eks. ca. 98 % rett på “opp-gaven” (683 objekter)



Stian Anfinsen

Takk for meg!



[HTTP://WWW.PHYS.UIT.NO/~ROBERTJ](http://www.phys.uit.no/~robertj)