textanalys på stor skala

jussi karlgren

Gavagai och KTH

april 2017

Jussi Karlgren

- distributionell semantik på realistisk skala

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- och en hel del som adjungerad professor i språkteknologi på KTH

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- och en hel del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar

Gavagai

- analyserar stora mängder strömmande text på massa språk

Gavagai

- analyserar stora mängder strömmande text på massa språk
- bygger på distributionell semantik

Gavagai

- analyserar stora mängder strömmande text på massa språk
- bygger på distributionell semantik
- grundat 2008

Gavagai

- analyserar stora mängder strömmande text på massa språk
- bygger på distributionell semantik
- grundat 2008
- av två lingvister härifrån och tre datavetarkollegor till dem

Distributionell semantik

- distributionella modeller bygger på (sam)förekomststatistik ...

Distributionell semantik

- distributionella modeller bygger på (sam)förekomststatistik ...
- ... av observerbara språkliga företeelser ...

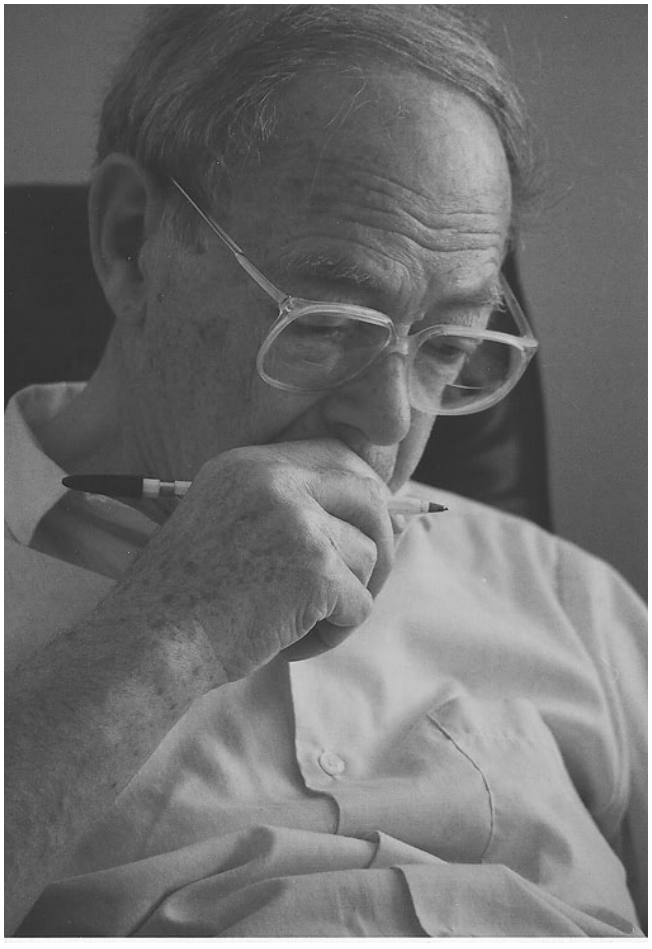
Distributionell semantik

- distributionella modeller bygger på (sam)förekomststatistik ...
- ... av observerbara språkliga företeelser ...
- ... med hänsyn tagen till kontext.

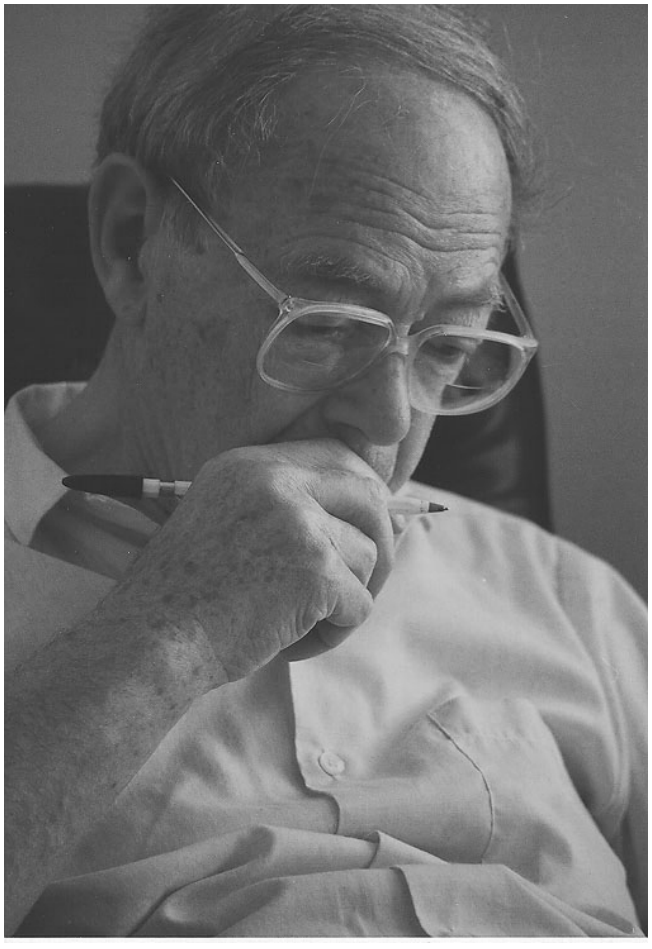
Distributionell semantik

- distributionella modeller bygger på (sam)förekomststatistik ...
- ... av observerbara språkliga företeelser ...
- ... med hänsyn tagen till kontext.

(det är många frågor de där begreppen väcker, eller hur?)



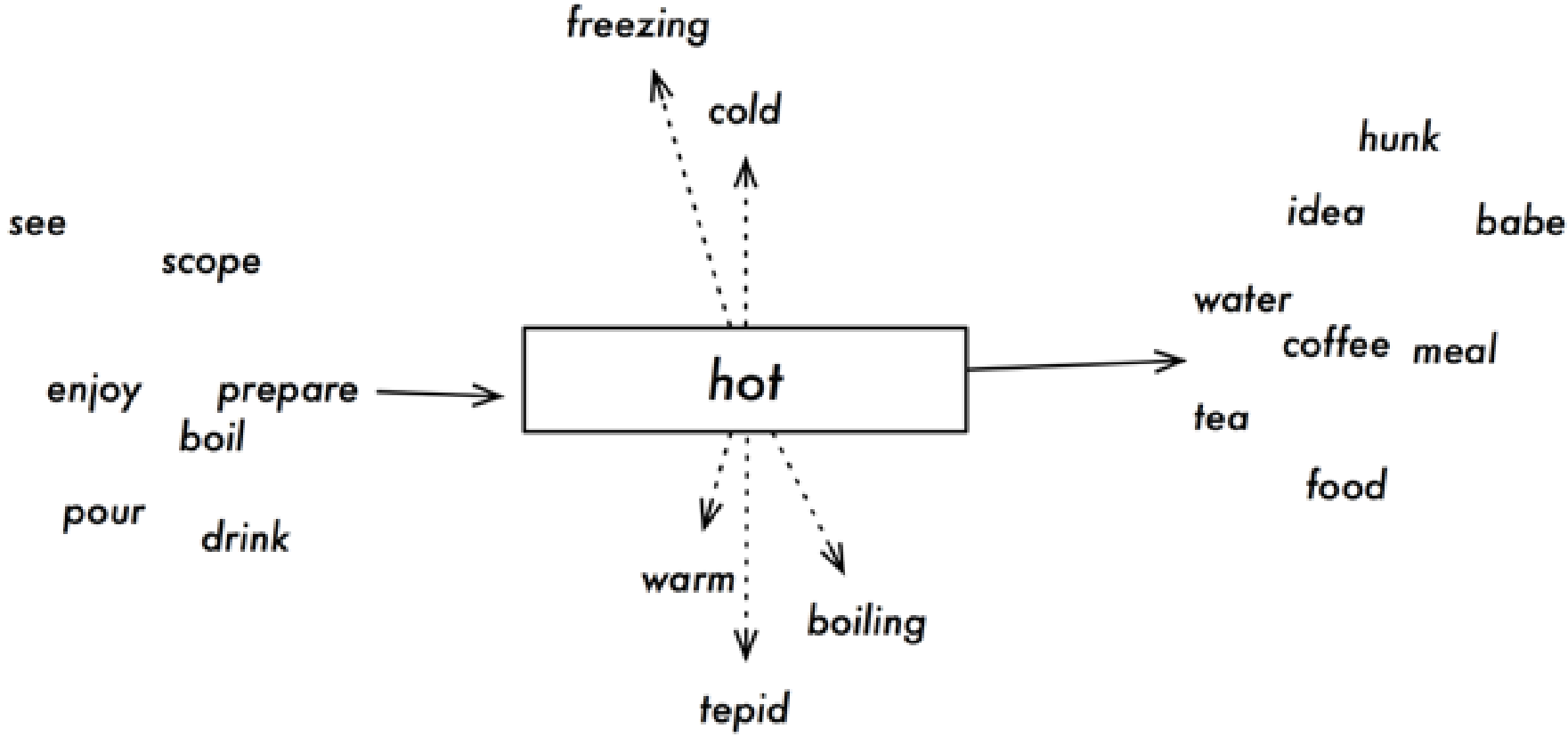
termer med liknande distributionella egenskaper har liknande betydelse
''den distributionella hypotesen''



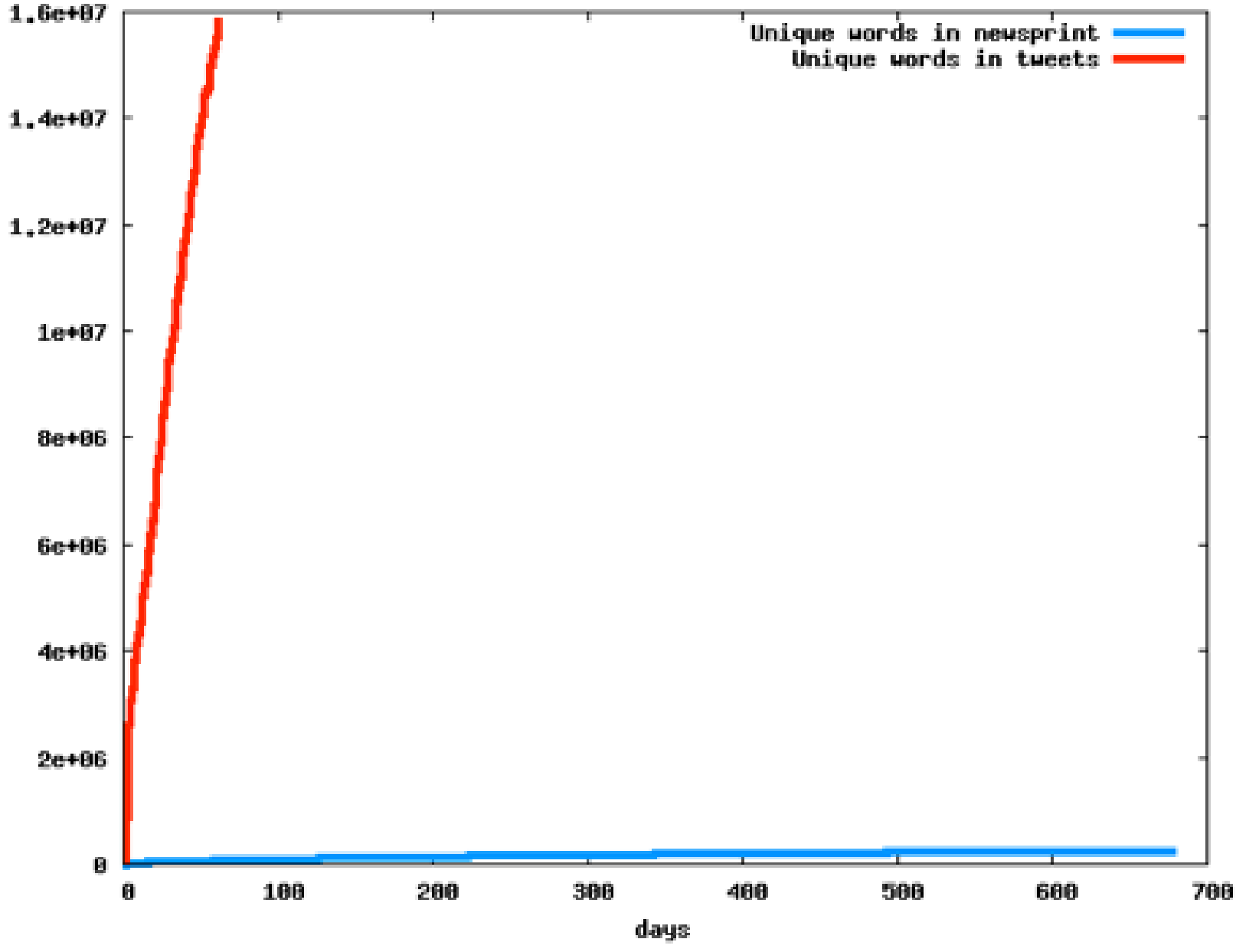
termer med liknande distributionella egenskaper har liknande betydelse
''den distributionella hypotesen''

(fler frågor väcks förstås här, eller hur?)

the weather is great in barcelona
the weather is hot in brownsville
the weather is gray in stockholm
the climate is passable in nice
the weather is chilly in helsinki
the weather is nippy in moscow
the weather is nice in hong kong
the weather in syktyvkar is balmy
the climate is chilly at the office
the tea is hot
i drink tea
a hot meal will make you feel better
enjoy your hot beverages


freezing
cold
hunk
idea
babe
see
scope
water
coffee
meal
enjoy
prepare
hot
boil
tea
pour
drink
food
warm
boiling
tepid

hur ser då språkliga data ut?

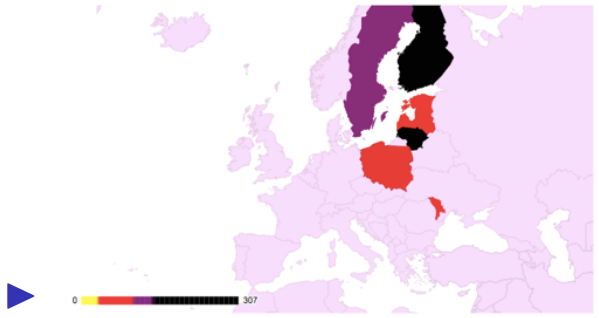

1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
0
100
200
300
400
500
600
700
days
Unique words in newsprint
Unique words in tweets

statistisk observation: stor volym data; massa massa särdrag; gles samförekomstmatris

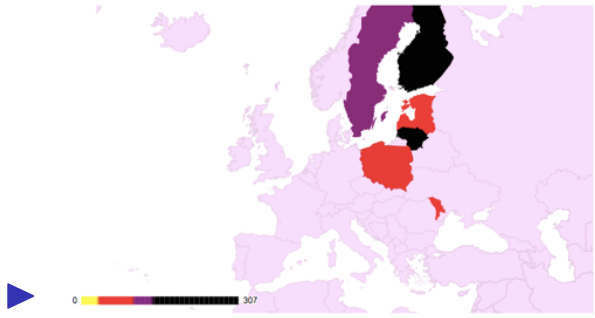
statistisk observation: stor volym data; massa massa särdrag; gles samförekomstmatris
men vi vet bättre

lexicon.gavagai.se

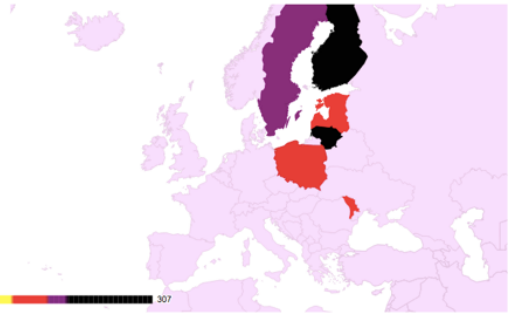
- geopolitiska kartor och säkerhetstil-lämpningar



- geopolitiska kartor och säkerhetstil-lämpningar
- finansiella tillämpninar

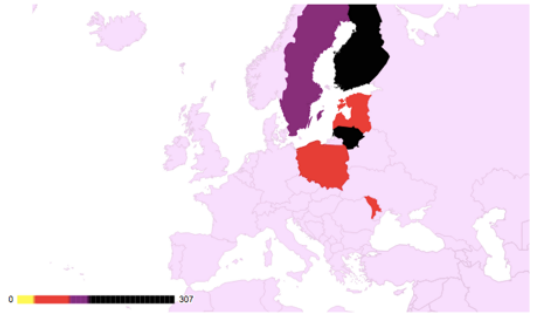


- geopolitiska kartor och säkerhetstil-lämpningar
- finansiella tillämpninar
- hatspråksvarnare

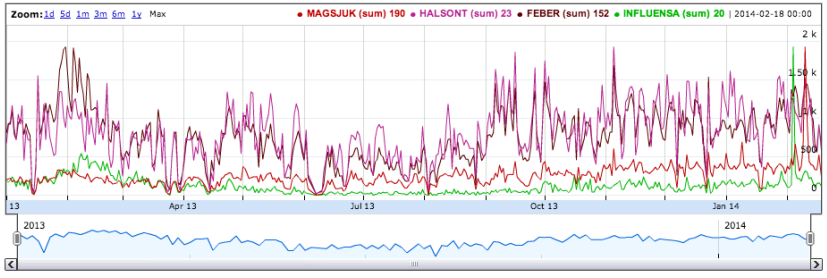


- kulturskillnader!

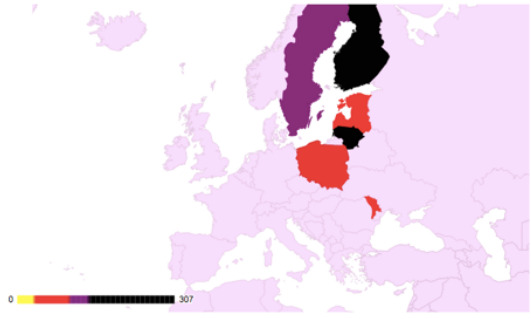
- geopolitiska kartor och säkerhetstil-lämpningar
- finansiella tillämpninar
- hatspråksvarnare
- jag-barometern



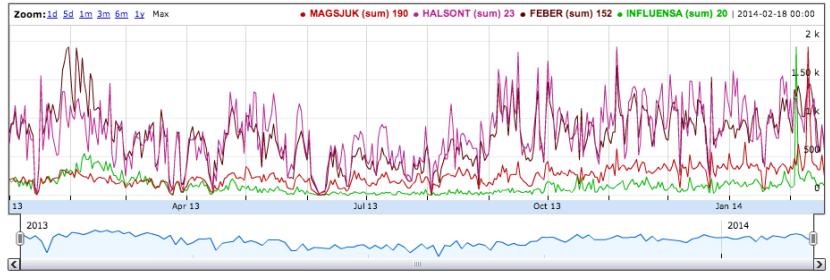
- kulturskillnader!
-



- geopolitiska kartor och säkerhetstil-lämpningar
- finansiella tillämpninar
- hatspråksvarnare
- jag-barometern



- kulturskillnader!
-



inga insikter utan analytiker!

''data mining''?

- datorminne är billigt och hyrbart

''data mining''?

- datorminne är billigt och hyrbart
- systemen är uppkopplade

''data mining''?

- datorminne är billigt och hyrbart
- systemen är uppkopplade
- sakernas internet

''data mining''?

- datorminne är billigt och hyrbart
- systemen är uppkopplade
- sakernas internet
- nya (gamla!) datamängder blir inlästa och uppkopplade

''data mining''?

- datorminne är billigt och hyrbart
- systemen är uppkopplade
- sakernas internet
- nya (gamla!) datamängder blir inlästa och uppkopplade
- encyklopediska data blir inlänkade

''data mining''?

- datorminne är billigt och hyrbart
- systemen är uppkopplade
- sakernas internet
- nya (gamla!) datamängder blir inlästa och uppkopplade
- encyklopediska data blir inlänkade
- allt möjligt loggas

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer
- befolkningsstatistik

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer
- befolkningsstatistik
- tidslinjer av massa olika slag

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer
- befolkningsstatistik
- tidslinjer av massa olika slag
- texter

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer
- befolkningsstatistik
- tidslinjer av massa olika slag
- texter

språk är den mest rimliga abstraktionsnivåhanteraren om människor ska fundera på data --- vem ska jobba med den språkliga representationen?

kunskap dyker nu upp på alla möjliga abstraktionsnivåer!

- sensormätvärden såsom temperaturer
- befolkningsstatistik
- tidslinjer av massa olika slag
- texter

språk är den mest rimliga abstraktionsnivåhanteraren om människor ska fundera på data --- vem ska jobba med den språkliga representationen?
(helt oväldigt påstående)

vad är semantik?

vad är semantik?
semantik kopplar kunskapsrepresentationer till varandra

vad är semantik?
semantik kopplar kunskapsrepresentationer till varandra

- en representation, t ex observationer

vad är semantik?
semantik kopplar kunskapsrepresentationer till varandra

- en representation, t ex observationer
- en annan (användbar) representation

vad är semantik?
semantik kopplar kunskapsrepresentationer till varandra

- en representation, t ex observationer
- en annan (användbar) representation
- relationer mellan dem

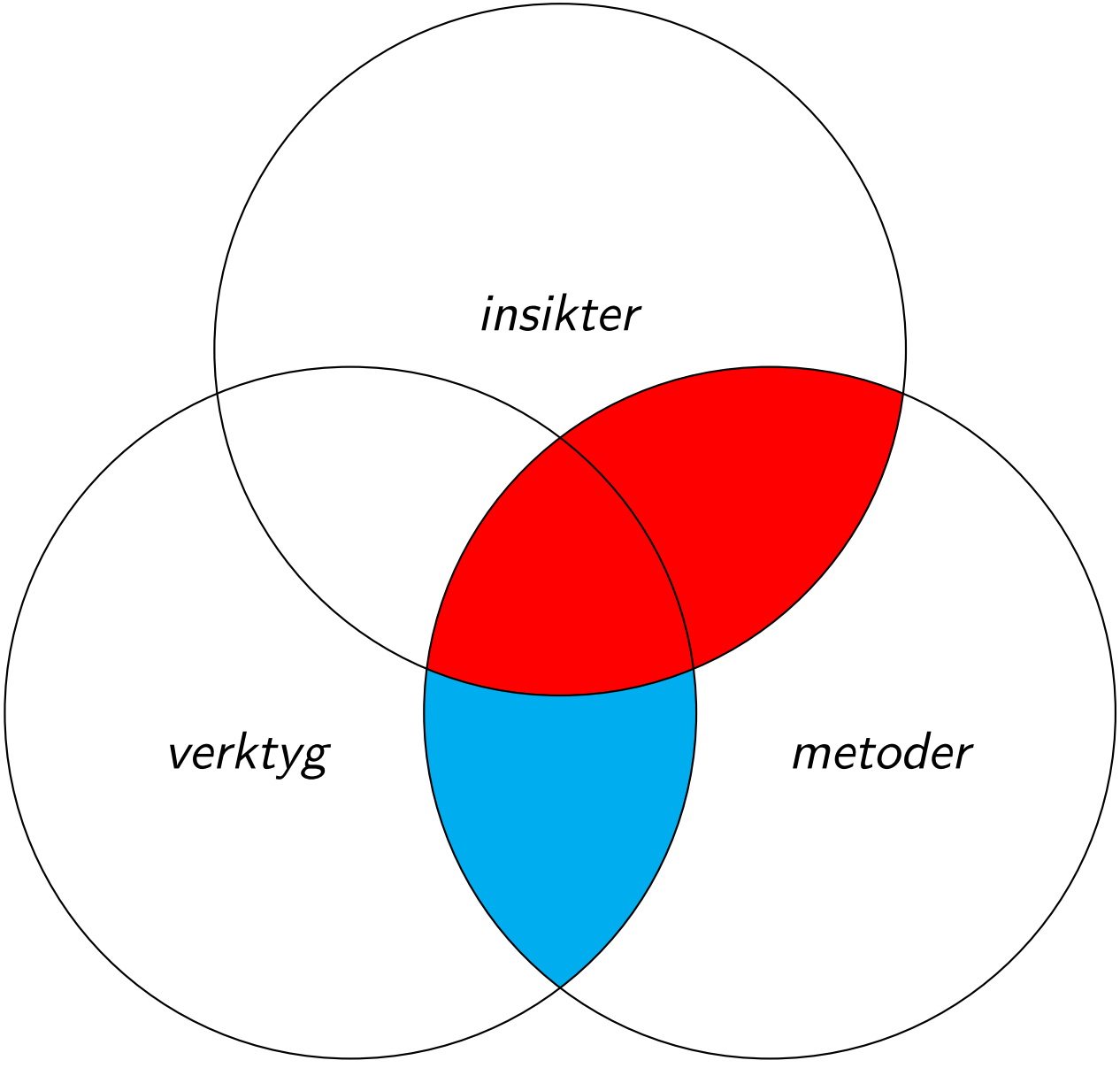
vad är semantik?
semantik kopplar kunskapsrepresentationer till varandra

- en representation, t ex observationer
- en annan (användbar) representation
- relationer mellan dem

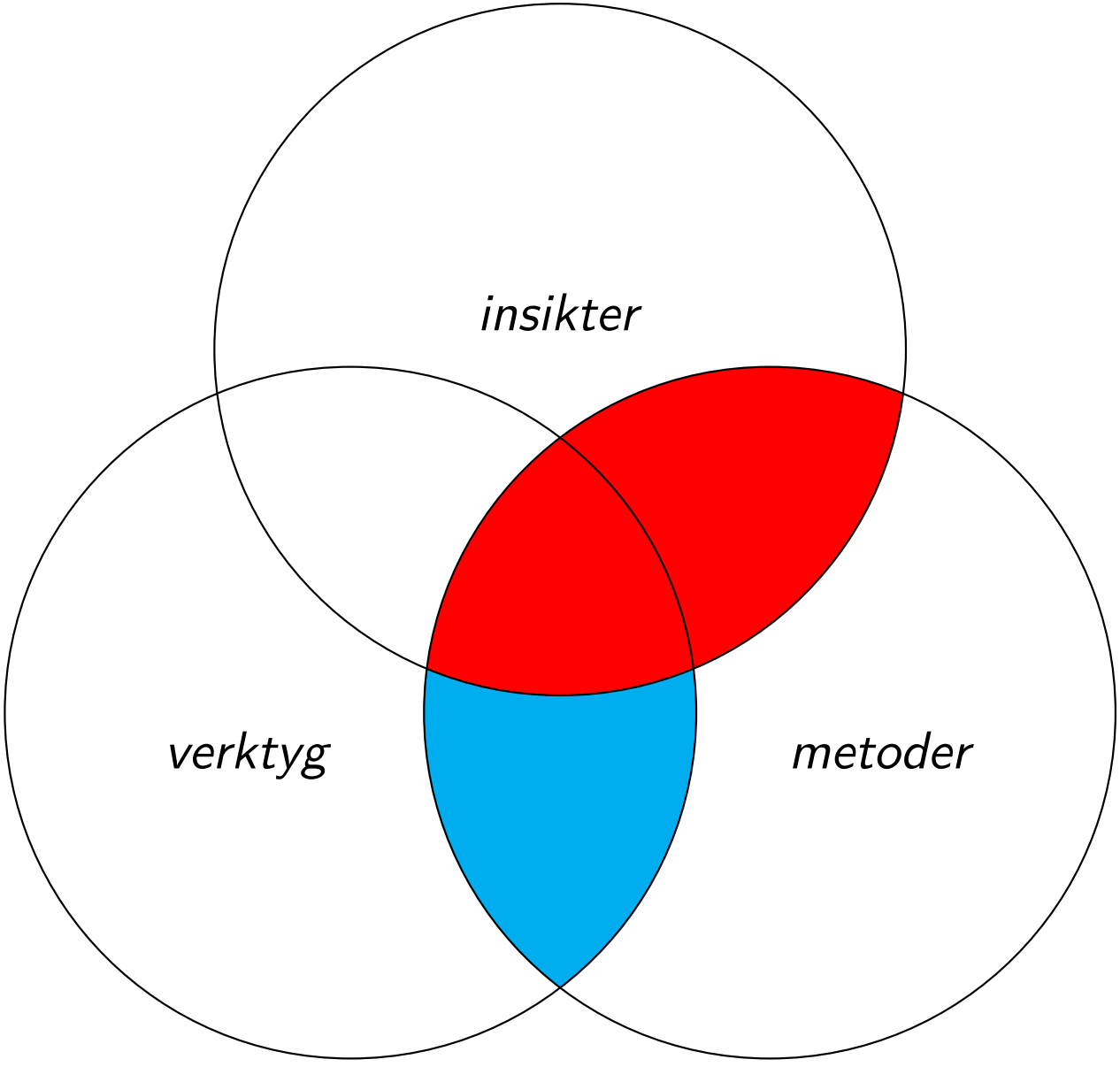
det här är ungefär det som data mining handlar om

dags för venndiagram!

dags för venndiagram!



dags för venndiagram!



äsch

humaniora studerar numera (ganska) stora och (snabbt) växande datamängder

humaniora studerar numera (ganska) stora och (snabbt) växande datamängder

de verktyg som finns kan skräddarsys för behov, men bara på beställning

humaniora studerar numera (ganska) stora och (snabbt) växande datamängder

de verktyg som finns kan skräddarsys för behov, men bara på beställning

någon måste göra beställningen

ilsken fråga:
var är humanisterna?

att ta med hem:

att ta med hem:
insikter kräver analys

att ta med hem:
insikter kräver analys

analys kräver kunskapsrepresentation

att ta med hem:
insikter kräver analys

analys kräver kunskapsrepresentation

kunskapsrepresentationsbygge kräver hypoteser och kunskap om det som behandlas

att ta med hem:
insikter kräver analys

analys kräver kunskapsrepresentation

kunskapsrepresentationsbygge kräver hypoteser och kunskap om det som behandlas

vem ska formulera hypoteserna?

att ta med hem:
insikter kräver analys

analys kräver kunskapsrepresentation

kunskapsrepresentationsbygge kräver hypoteser och kunskap om det som behandlas

vem ska formulera hypoteserna?

identifiera och förädla hantverkskunnandet!

(hos forskare, hos analytiker såväl som hos verktygskonstruktörer; inget av detta är trivialt, inget av detta är magi.)