vad är (dator)lingvistens roll när maskininlärningssystem verkar göra jobbet?

jussi karlgren

december 2018

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Jussi Karlgren

- distributionell semantik på realistisk skala
- stilistik och genre i text och konversation
- mest på textanalysföretaget Gavagai
- en del som adjungerad professor i språkteknologi på KTH
- examen från dessa salar
- gillar inte termen datorlingvist
- har aldrig riktigt accepterat distinktionerna mellan syntax-semantik-pragmatik

Gavagai

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

Gavagai

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

Gavagai

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

Gavagai

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

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?

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

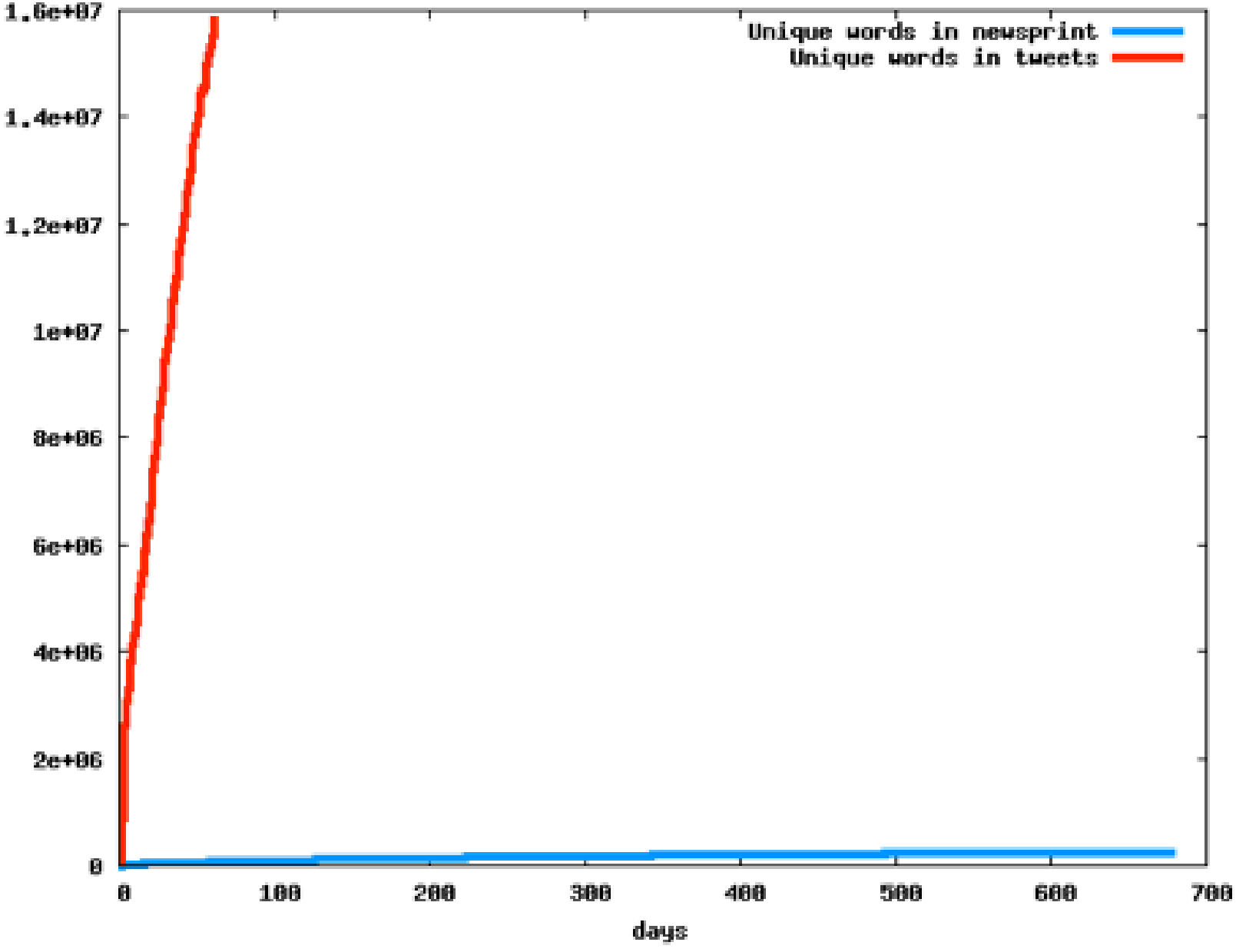
språk är den mest rimliga abstraktionsnivåhanteraren om människor ska fundera på data — vem ska jobba med den språkliga representationen?

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?

hur ser då språkliga data ut?



vad är semantik?

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

det här är ungefär det som data mining handlar om

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

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

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

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

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

geometriska modeller och deras bestickande karaktär

först lite matematisk övning

- semantiska rum är mellanhögdimensionella i ett högdimensionellt observationsrum
- helpopulära just nu
- geometriska modeller är intuitivt lättbegripliga och spännande

---

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

geometriska modeller och deras bestickande karaktär

först lite matematisk övning

- semantiska rum är mellanhögdimensionella i ett högdimensionellt observationsrum
- helpopulära just nu
- geometriska modeller är intuitivt lättbegripliga och spännande

---

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

geometriska modeller och deras bestickande karaktär

först lite matematisk övning

- semantiska rum är mellanhögdimensionella i ett högdimensionellt observationsrum
- helpopulära just nu
- geometriska modeller är intuitivt lättbegripliga och spännande

---

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

geometriska modeller och deras bestickande karaktär

först lite matematisk övning

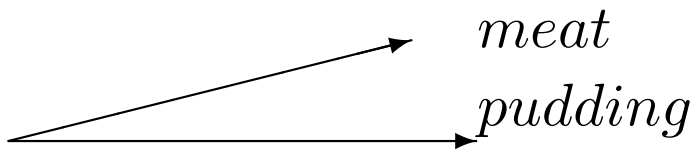
- semantiska rum är mellanhögdimensionella i ett högdimensionellt observationsrum
- helpopulära just nu
- geometriska modeller är intuitivt lättbegripliga och spännande

$$d_{cos}(x, y) = \frac{\vec{x} \cdot \vec{y}}{\mid \vec{x} \mid\mid \vec{y} \mid} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Gerard Salton et al. 1975. A vector space model for automatic indexing. CACM
- Hinrich Schütze. 1993. Word Space. NIPS.
- David Dubin. 2004. „The most influential paper Gerard Salton never wrote". Library Trends.
- Magnus Sahlgren. 2006. The Word-Space Model. Stockholm University.

But

men de hallucinerar precision bortom den semantiska horisonten

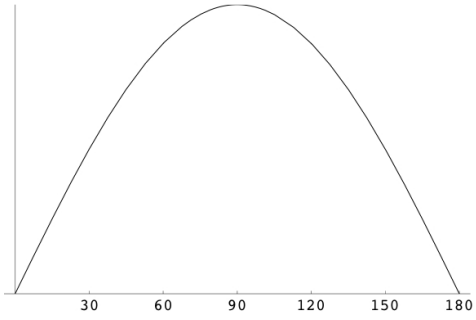


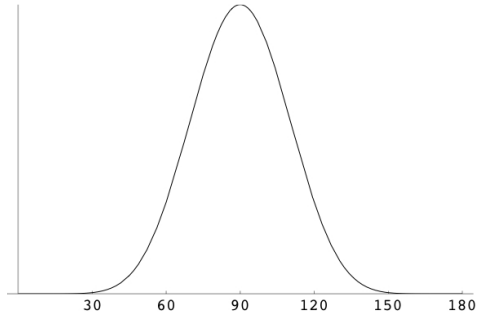
$d_{cos}(\text{"}pudding\text{"}, \text{"}meat\text{"}) \approx cos(\pi/5)$
$d_{cos}(\text{"}cardamom\text{"}, \text{"}tensor\ \ algebra\text{"}) =?$

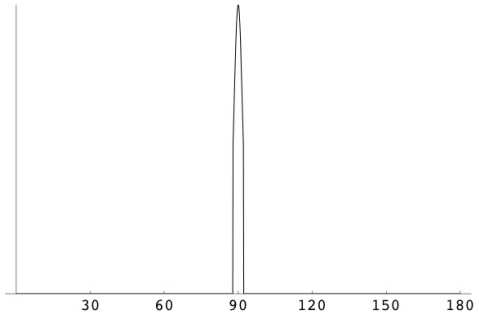
- Jussi Karlgren. 2005. "Meaningful models for information access systems." CSLI

hur ser rummet ut?

centroider funkar inte som man tror







3 dimensions 10 dimensions 1000 dimensions

- Jussi Karlgren et al. 2008. Filaments of Meaning in Word Space. ECIR

och ickelinjära modeller leder oss bortom geometriska modeller

... vilket tar oss tillbaka till frågan jag började med

hur ska vi då göra?

- vi (i dessa salar) har kunskap om hur språkliga uttryck uppvisar vissa regelbundenheter
- andra har kunskap om hur regelbundenheter kan krafsas fram ur observation
- kodning är tröskeln

hur ska vi då göra?

- vi (i dessa salar) har kunskap om hur språkliga uttryck uppvisar vissa regelbundenheter
- andra har kunskap om hur regelbundenheter kan krafsas fram ur observation
- kodning är tröskeln

hur ska vi då göra?

- vi (i dessa salar) har kunskap om hur språkliga uttryck uppvisar vissa regelbundenheter
- andra har kunskap om hur regelbundenheter kan krafsas fram ur observation
- kodning är tröskeln

min processmodell:

- konsumtion
- konstruktivism

min processmodell:

- konsumtion
- konstruktivism

konstruktivism

hyggligt debatterbart påstående:

ett yttrandes konfiguration är ett särdrag med samma ontologiska status som de i yttrandet förekommande termerna

konstruktioner och lexem har begreppslig betydelse (på ett kontinuum)

går inte att dra gräns dem emellan

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

konstruktivism

hyggligt debatterbart påstående:

ett yttrandes konfiguration är ett särdrag med samma ontologiska status som de i yttrandet förekommande termerna

konstruktioner och lexem har begreppslig betydelse (på ett kontinuum)

går inte att dra gräns dem emellan

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

konstruktivism

hyggligt debatterbart påstående:

ett yttrandes konfiguration är ett särdrag med samma ontologiska status som de i yttrandet förekommande termerna

konstruktioner och lexem har begreppslig betydelse (på ett kontinuum)

går inte att dra gräns dem emellan

- Naomi Sager et al. 1965-1998. Linguistic String Project.
- William Croft. 2005. Radical and typological arguments for radical construction grammar. In Construction Grammars John Benjamins.

hur göra detta i praktiken?

- konkatenation
- vektoraddition

hur göra detta i praktiken?

- konkatenation
- vektoraddition

hur göra detta i praktiken?

- konkatenation
- vektoraddition

hur göra detta i praktiken?

- konkatenation
- vektoraddition

(a) ord $\bar{v}_{dog} + \bar{v}_{chew} + \bar{v}_{bone} \sim \bar{v}_{bone}$

hur göra detta i praktiken?

- konkatenation
- vektoraddition

(a) ord $\bar{v}_{dog} + \bar{v}_{chew} + \bar{v}_{bone} \sim \bar{v}_{bone}$

(b) „semantiska roller”:
$\Pi_{agent}(\bar{v}_{dog}) + \Pi_{predicate}(\bar{v}_{chew}) + ...$

hur göra detta i praktiken?

- konkatenation
- vektoraddition

(a) ord $\bar{v}_{dog} + \bar{v}_{chew} + \bar{v}_{bone} \sim \bar{v}_{bone}$

(b) „semantiska roller”:
$\Pi_{agent}(\bar{v}_{dog}) + \Pi_{predicate}(\bar{v}_{chew}) + ...$

(c) konstruktionella element:
$\bar{v}_{uncertain} + \bar{v}_{profanity} + \bar{v}_{utteranceverb} + \bar{v}_{present_tense} \cdots$

massa meningar i ett vektorrum ($d = 2000; k = 10$):

massa meningar i ett vektorrum ($d = 2000; k = 10$):

„I really did not like the clarinet, I am afraid: it sounded weak!”

massa meningar i ett vektorrum ($d = 2000; k = 10$):

„I really did not like the clarinet, I am afraid: it sounded weak!”

(a) ord: My sister plays the clarinet.

massa meningar i ett vektorrum ($d = 2000; k = 10$):

„I really did not like the clarinet, I am afraid: it sounded weak!”

(a) ord: My sister plays the clarinet.

(b) konstruktioner: I’m surrounded by really soft decadent pillows which do not work for me at all.

aktuella och kommande experiment

bygger på konsumtion och konstruktioner

analys av attityd vs ämne vs individuell variation

analys av attityd vs aspektualitet

könsskillnader i text (mycket skeptisk)

yttranderum med referentialitet, TMA, attityd, hållning och sannfärdighet, polaritet, intensitet ...

authorship gender profiling

linguistic theories are not built for this task and are overly specific;
word occurrence models overtrain on topic and do not generalise (but give upwards of 80% accuracy)

- Jussi Karlgren, Lewis Esposito, Chantal Gratton, Pentti Kanerva. Authorship Profiling Without Using Topical Information. Notebook for PAN at CLEF 2018.

authorship gender profiling

linguistic theories are not built for this task and are overly specific;
word occurrence models overtrain on topic and do not generalise (but give upwards of 80% accuracy)

- add all observed features of potential interest into same representation

---

- Jussi Karlgren, Lewis Esposito, Chantal Gratton, Pentti Kanerva. Authorship Profiling Without Using Topical Information. Notebook for PAN at CLEF 2018.

authorship gender profiling

linguistic theories are not built for this task and are overly specific;
word occurrence models overtrain on topic and do not generalise (but give upwards of 80% accuracy)

- add all observed features of potential interest into same representation
- use cosine to test which representations fit best

- Jussi Karlgren, Lewis Esposito, Chantal Gratton, Pentti Kanerva. Authorship Profiling Without Using Topical Information. Notebook for PAN at CLEF 2018.

authorship gender profiling

linguistic theories are not built for this task and are overly specific;
word occurrence models overtrain on topic and do not generalise (but give upwards of 80% accuracy)

- add all observed features of potential interest into same representation
- use cosine to test which representations fit best
- findings (so far):
  - ♀: first person subjects, „truly"-amplifiers, interjections, 1psg
  - ♂: think verbs and hedges, profanity, passives, modal auxes,

---

- Jussi Karlgren, Lewis Esposito, Chantal Gratton, Pentti Kanerva. Authorship Profiling Without Using Topical Information. Notebook for PAN at CLEF 2018.

authorship gender profiling

linguistic theories are not built for this task and are overly specific;
word occurrence models overtrain on topic and do not generalise (but give upwards of 80% accuracy)

- add all observed features of potential interest into same representation
- use cosine to test which representations fit best
- findings (so far):
  - ♀: first person subjects, „truly"-amplifiers, interjections, 1psg
  - ♂: think verbs and hedges, profanity, passives, modal auxes,

non-symmetric relationship between authors

- Jussi Karlgren, Lewis Esposito, Chantal Gratton, Pentti Kanerva. Authorship Profiling Without Using Topical Information. Notebook for PAN at CLEF 2018.

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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

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!
ställ krav på kravspecifikation!

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