

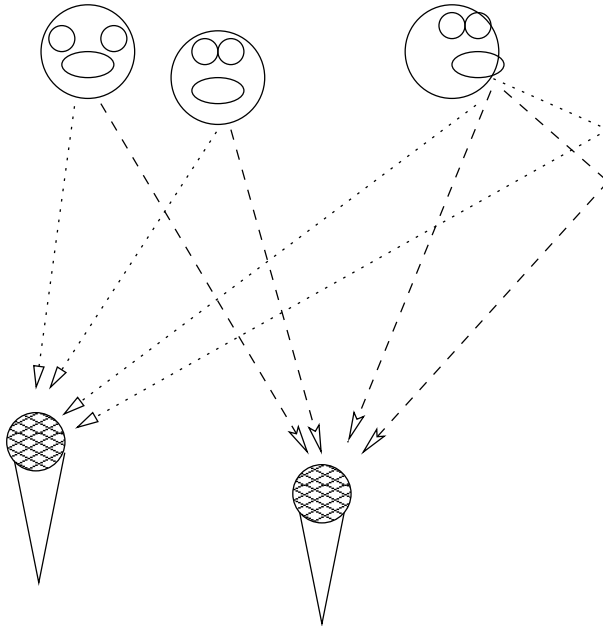
# Riippumattomien komponenttien analyysin laajennuksia ja sovelluksia tiedon louhintaan

# Ongelmanasettelu

- **Dataa on niin paljon, että sen käsittelyyn tarvitaan tehokkaita menetelmiä.**
- **Data: numeroiksi muutettavissa olevaa mittaus- tai muuta havaintoaineistoa.**  
**Esitetään usein taulukkomuodossa.**
- **Ajatellaan että kokonaisuus on muodostunut tuntemattomien osatekijöiden eli komponenttien yhteisvaikutuksena. Selvitetään mitä nämä ovat.**

- **Komponentit antavat usein yksinkertaisemman ja selkeämmän kuvan havaintoaineistosta:**
  - iso tekstiaineisto koostuu erilaisista aiheista, jotka luettelemalla saadaan yleiskuva aineistosta
  - cocktailkutsuilla puheensorinasta voidaan erotella yksittäisten puhujien äänet
- **Komponentit ovat toisistaan riippumattomia (siitä nimi riippumattomien komponenttien analyysi).**
- **Seuraavassa eräs laajennus menetelmään.**

# Kompleksiarvoisten signaalien analyysi



- Matkapuhelinsignaalit tai äänisignaalit saapuvat usein mittauspisteeseen erilaisilla viiveillä tai heijastuvat matkalla.
- Muunnetaan mitatut signaalit kompleksiarvoiksi, jolloin viiveitä on helpompi käsitellä.
- Ratkaistaan riippumattomat komponentit kompleksiarvoista mittauksista.

# Tiedon louhinta

- Dataa kertyy automaattisesti, eikä se ole alunperin tarkoitettu analysoitavaksi.
- Analysoimalla dataa voidaan oppia ilmiöstä, joka on datan tuottanut.
- Tiedon louhinta on joukko matemaattisia, tilastollisia ja tietojenkäsittelytieteellisiä menetelmiä suurten datajoukkojen tehokkaaseen analysointiin.

- **Esimerkki riippumattomien komponenttien analyysin soveltamisesta tiedon louhintaan: Vaihtuva-aiheinen keskustelu.**
- **Internetin chat-keskusteluryhmässä on monta keskustelua meneillään yhtä aikaa, ja keskustelujen aiheet muuttuvat ajan myötä.**

<Haley-CNN> Gallup Poll's Newport will join the #CNN\_Newsroom at 10:00 a.m. EST to discuss: the public views the Bush transition

<TomM--\_> Fatboy....smart!!!

<Gary> ya bush at 60% and Gore at 78% still

<Nailheads> .....I thought not.

<Whitikau-CNN> jeter: Does that mean that Bush will need to work harder than previous Presidents to gain the respect and trust of the people?

<alfie> jeter, according to Louis Lapham, neither candidate was worth dooky.

<DUFUS> That's because when asked why they voted for him.....the answer was: "He is a good old boy."

<DUFUS> Silvertone.....HELLO?

<Nailheads> ....ha ha. if that's all you got...b.s...then okay. sling away.

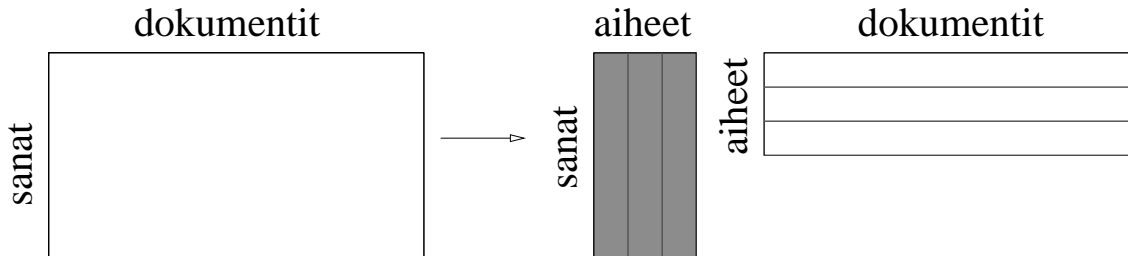
<stella> DUFUS MAKE SURE YOU HAVE THE RIGHT FACT OK BEFORE YOU BLAME HIM EVERY TIME YOU BE IN THIS CHANNEL

<Moro^^^> daddy Bush promised Powell Sec of State...if he sat this election out. This is a grand conspiracy head by the elder Bush. It is his second administration.

<jeter> Nailheads, 60%...LOLOLOLOL!!! Are you nuts? Maybe that was a question on him personally. But on handling the job...thus far, it's less than 50%.



- Chat-keskusteluvirta pilkotaan lyhyiksi pätkiksi. Kukin pätkä ajatellaan yhdeksi tekstidokumentiksi.
- Tekstidokumenttikokoelma muutetaan numeroiksi:
  - Luodaan aineiston pohjalta sanasto aineiston tärkeimmistä sanoista.
  - Luetellaan jokaisen dokumentin kohdalla, miten monta kertaa sanaston eri sanat esiintyvät siinä.
- Ratkaistaan riippumattomat komponentit



---

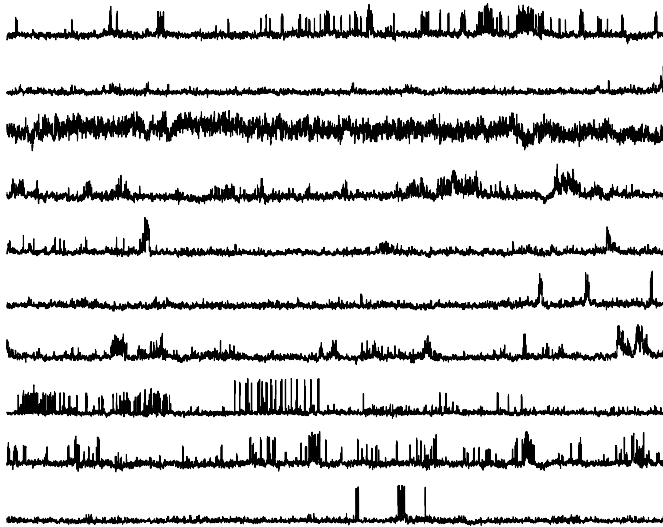
Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
jackson	site	bush	religion	violenc
sharpton	web	ashcroft	god	report
child	net	vote	jesu	youth
stori	word	kennedi	bibl	children
drudg	parent	presid	religi	gun
rainbow	nanni	cnn	life	point
monei	internet	time	follow	home
mistress	block	gore	read	drug
coalition	kid	question	stori	famili
tonight	system	elect	univers	satcher
pregnant	access	god	exist	health
affair	child	senat	faith	risk
black	base	power	man	factor
chenei	chat	thing	book	surgeon
jessi	page	fact	earth	prevent

---

---

Aihe 6	Aihe 7	Aihe 8	Aihe 9	Aihe 10
flag	california	join	tax	free
move	power	discuss	cut	liber
citi	electr	est	exempt	opinion
ncaa	energi	tonight	monei	religion
offici	blackout	room	gop	form
atlanta	state	studio	hous	polit
count	deregul	cnn	congress	conserv
game	compani	conserv	pay	birth
night	crisi	american	interest	philosophi
georgia	price	nea	recess	establish
chang	plant	union	payer	narrow
lose	util	keen	secur	restrict
confeder	order	type	henri	independ
hehe	home	chat	hypocrit	orthodox
chenei	cost	newsroom	hyde	bound

---



**Vaaka-akselilla on aika (24 tuntia).**  
**Eri käyrät vastaavat eri keskustelunaiheita.**  
**Käyrässä näkyy piikki, kun aiheesta keskustellaan.**

**Kiitos**