Lectio praecursoria

Satunnaistusalgoritmeja tiedonlouhinnan tulosten merkitsevyyden arviointiin

Markus Ojala

12. marraskuuta 2011

Käsitteet

Satunnaistusalgoritmeja tiedonlouhinnan tulosten merkitsevyyden arviointiin

1. Tiedonlouhinnan tulos
2. Merkitsevyyden arviointi
3. Satunnaistusalgoritmi

Käsitteet

Satunnaistusalgoritmeja tiedonlouhinnan tulosten merkitsevyyden arviointiin

1. Tiedonlouhinnan tulos
2. Merkitsevyyden arviointi
3. Satunnaistusalgoritmi

Tiedonlouhinta

Tiedonlouhinta

Hand et al. 2001: Tiedonlouhinta on (usein isojen) havaintoaineistojen analysointia, jonka tarkoituksena on löytää odottamattomia riippuvuussuhteita aineistossa sekä esittää tietoa uusilla ymmärrettävillä ja hyödyllisillä tavoilla aineiston omistajalle.

Tiedonlouhinta: 1. Esimerkki

Tuotteiden hinnat eri kaupoissa (€/ kg)

	Maito	Leipä	Banaani	Juusto	Kinkku
Kauppa 1	0.69	2.45	0.99	5.49	6.45
Kauppa 2	1.25	4.29	2.49	8.95	9.35
Kauppa 3	0.79	2.45	1.25	6.39	7.59
Kauppa 4	1.19	4.59	1.95	8.45	8.55
Kauppa 5	0.65	2.75	1.15	6.65	7.13
Kauppa 6	1.25	3.95	2.19	7.75	9.99

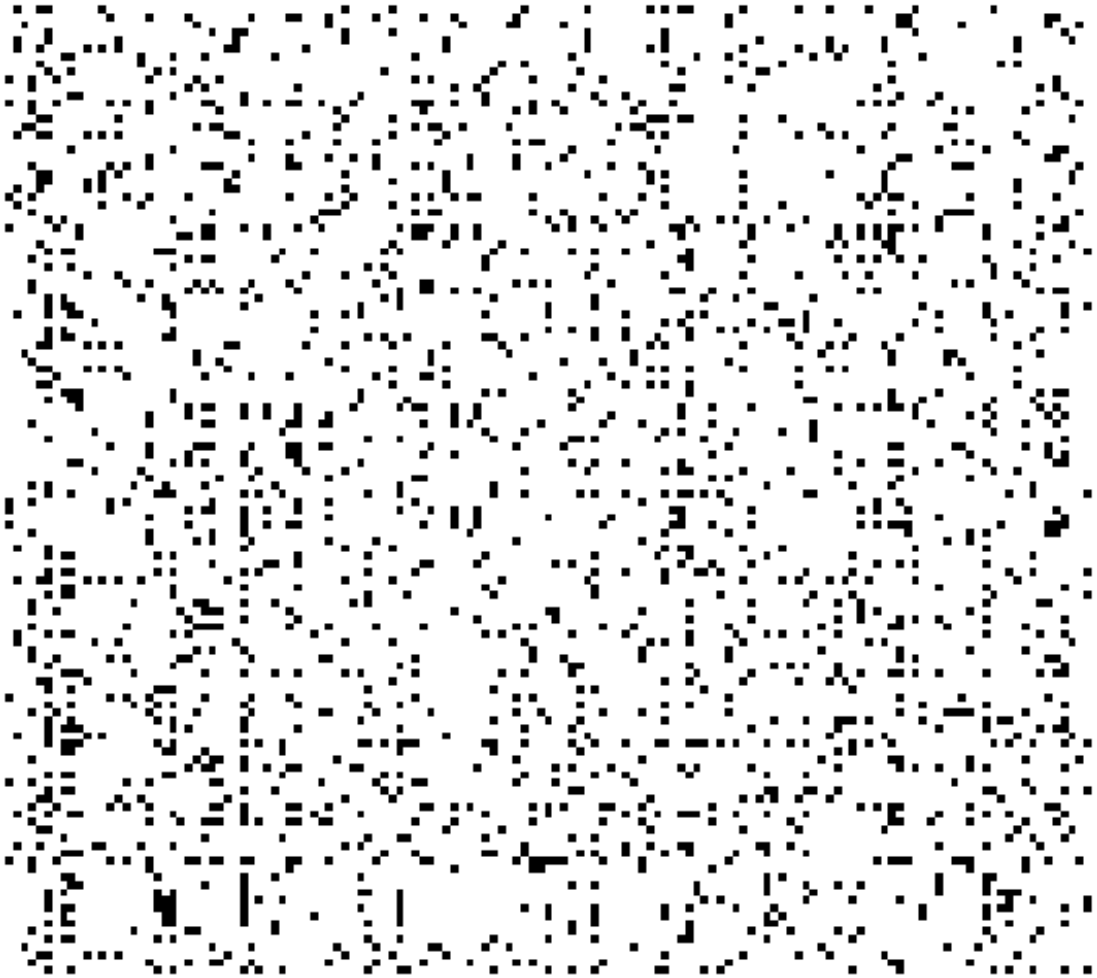
Tiedonlouhinta: 1. Esimerkki

Samanlaisten kauppojen tunnistaminen

	Maito	Leipä	Banaani	Juusto	Kinkku
Kauppa 1	0.69	2.45	0.99	5.49	6.45
Kauppa 3	0.79	2.45	1.25	6.39	7.59
Kauppa 5	0.65	2.75	1.15	6.65	7.13
Kauppa 2	1.25	4.29	2.49	8.95	9.35
Kauppa 4	1.19	4.59	1.95	8.45	8.55
Kauppa 6	1.25	3.95	2.19	7.75	9.99

Tiedonlouhinta: 2. Esimerkki

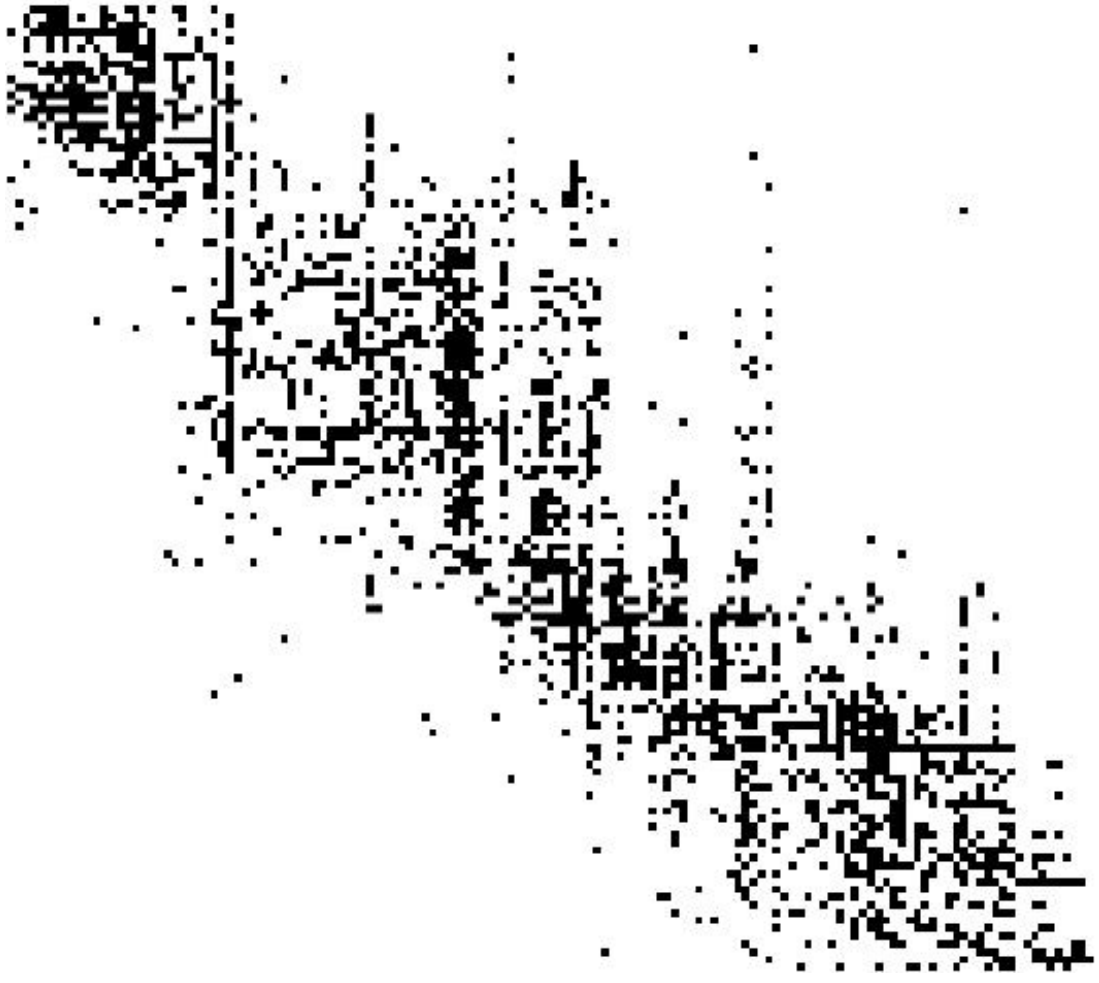
Fossiiliaineisto: (löytöpaikka, eläinlaji)



Fortelius, Jernvall, Gionis, Mannila, Paleobiology 32 (2006)

Tiedonlouhinta: 2. Esimerkki

Fossiiliaineisto: spektraalijärjestäminen



Käsitteet

Satunnaistusalgoritmeja tiedonlouhinnan tulosten merkitsevyyden arviointiin

1. Tiedonlouhinnan tulos
2. Merkitsevyyden arviointi
3. Satunnaistusalgoritmi

Merkitsevyyden arviointi

Tulos on tilastollisesti merkitsevä, jos

- on epätodennäköistä, että tulos on sattumaa
- eli vain pienessä osassa satunnaisia aineistoja havaitaan vastaava tai parempi tulos

P-arvo: Osuus satunnaisista aineistoista, joissa havaitaan vastaava tai parempi tulos

Esim. 1%, 5%

Merkitsevyyden arviointi

Mikä on satunnainen aineisto?

- sattumassakin on rakennetta
- joitain yhteisiä valittuja ominaisuuksia alkuperäisen aineiston kanssa
- muuten täysin satunnainen
- erilaiset satunnaistukset tuottavat erilaisen vertailupohjan

Merkitsevyyden arviointi: 1. Esimerkki

Alentaako lääke X verenpainetta?

	Verenpaineen muutos	Keskiarvo	Ero
Lääke X	-20, -5, 3, -10, -7	-7.8	-9.4
Plasebo	-3, 2, -5, 4, 10	1.6	

Merkitsevyyden arviointi: 1. Esimerkki

Alentaako lääke X verenpainetta?

	Verenpaineen muutos	Keskiarvo	Ero
Lääke X	-20, -5, 3, -10, -7	-7.8	-9.4
Plasebo	-3, 2, -5, 4, 10	1.6	

Yksi satunnaistettu aineisto

	Verenpaineen muutos	Keskiarvo	Ero
“Lääke X”	10, 4, -10, -5, -3	-0.8	4.6
“Plasebo”	2, 3, -5, -7, -20	-5.4	

Merkitsevyyden arviointi: 1. Esimerkki

Alentaako lääke X verenpainetta?

	Verenpaineen muutos	Keskiarvo	Ero
Lääke X	-20, -5, 3, -10, -7	-7.8	-9.4
Plasebo	-3, 2, -5, 4, 10	1.6	

Yksi satunnaistettu aineisto

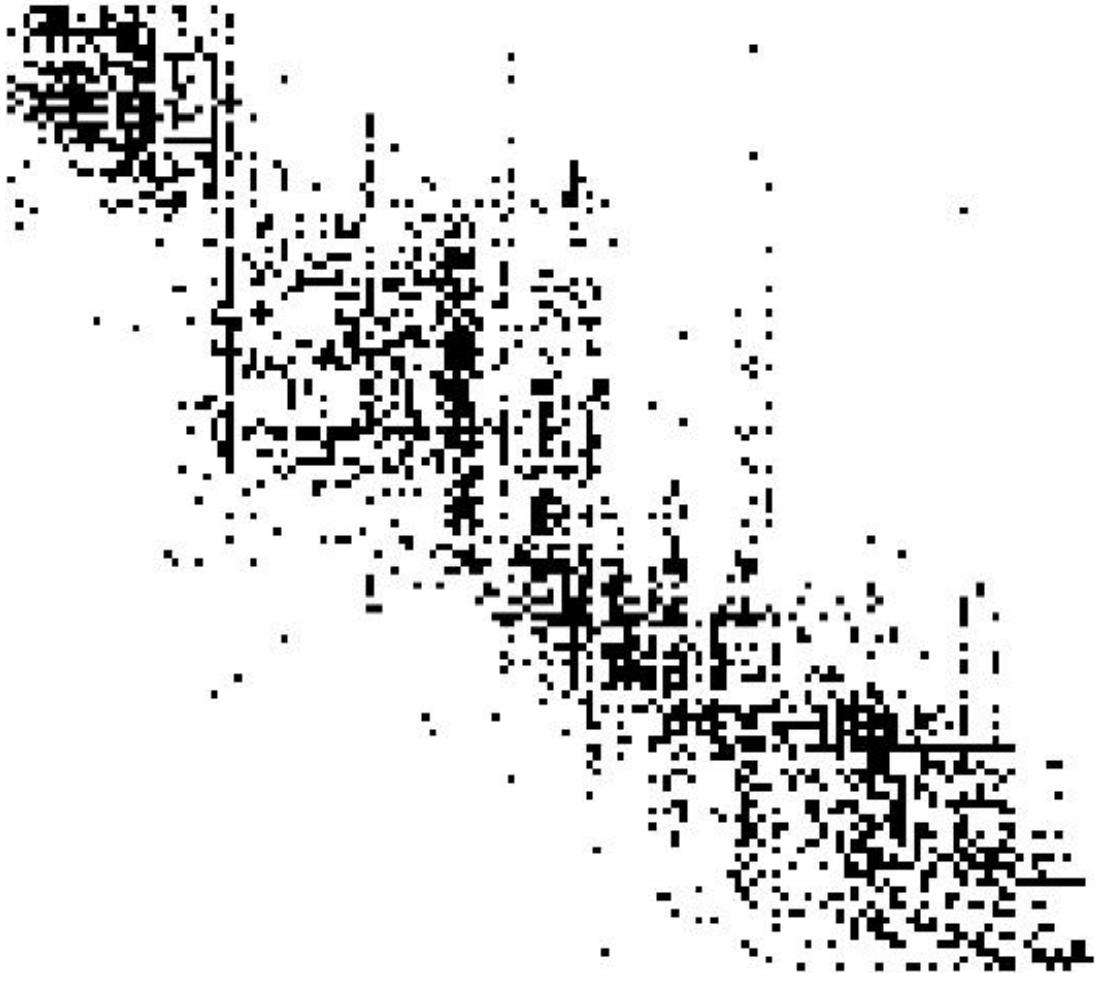
	Verenpaineen muutos	Keskiarvo	Ero
“Lääke X”	10, 4, -10, -5, -3	-0.8	4.6
“Plasebo”	2, 3, -5, -7, -20	-5.4	

Merkitsevyystestaus (1000 satunnaistusta)

$p = 3.6\%$ (tulos on merkitsevä)

Merkitsevyyden arviointi: 2. Esimerkki

Järjestetty alkuperäinen fossiiliaineisto



Merkitsevyyden arviointi: 2. Esimerkki

Satunnaistettu aineisto + järjestäminen

Käsitteet

Satunnaistusalgoritmeja tiedonlouhinnan tulosten merkitsevyyden arviointiin

1. Tiedonlouhinnan tulos
2. Merkitsevyyden arviointi
3. Satunnaistusalgoritmi

Satunnaistusalgoritmi

Satunnaistusalgoritmi

- Menetelmä satunnaistusten tuottamiseen
- Säilyttää halutut ominaisuudet
- Satunnaistaa muut ominaisuudet

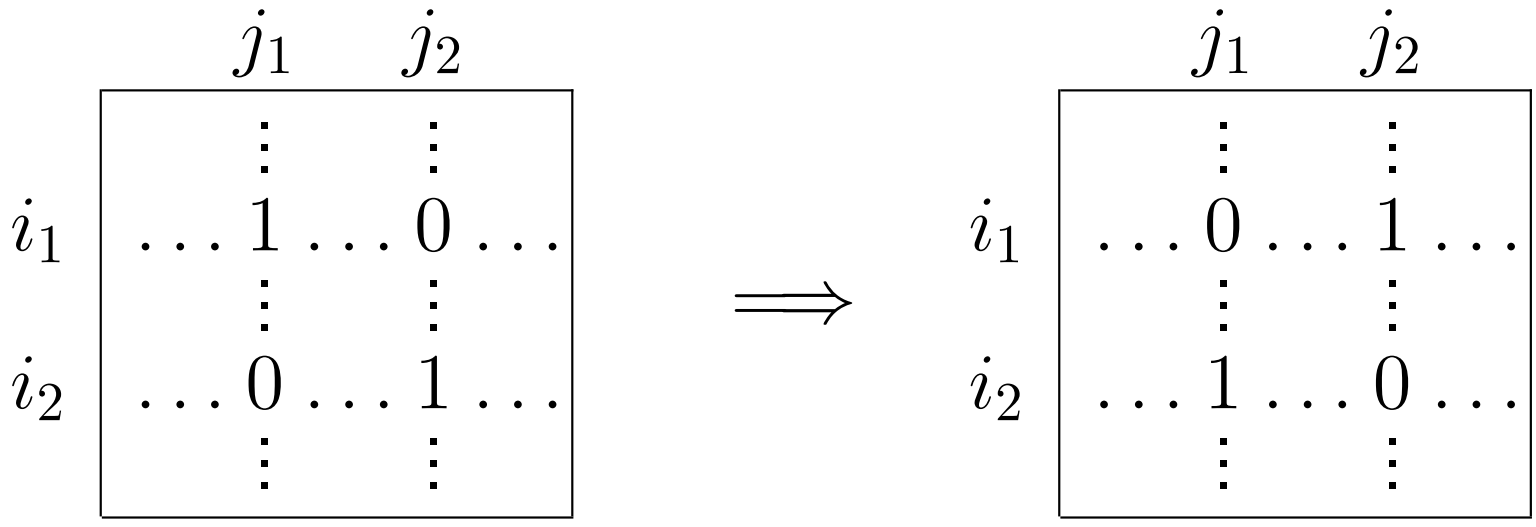
Kolme eri lähestymistapaa

1. Suora tuottaminen
2. Virheettömät paikalliset muutokset
3. Virheelliset paikalliset muutokset + virheen kontrollointi

Satunnaistusalgoritmi: Esimerkki

Binäärimatriisien satunnaistaminen

- säilytetään rivi- ja sarakesummat
- virheettömät paikalliset muutokset



Väitöskirjan sisältö

Väitöskirjan kontribuutiot

Uusia satunnaistusmenetelmiä:

1. Matriiseille
2. Toistuvaan ja tarkentuvaan tiedonlouhintaan
3. Relaatiotietokannoille
4. Luokitelluille aineistoille

Kontribuutiot: Esimerkki

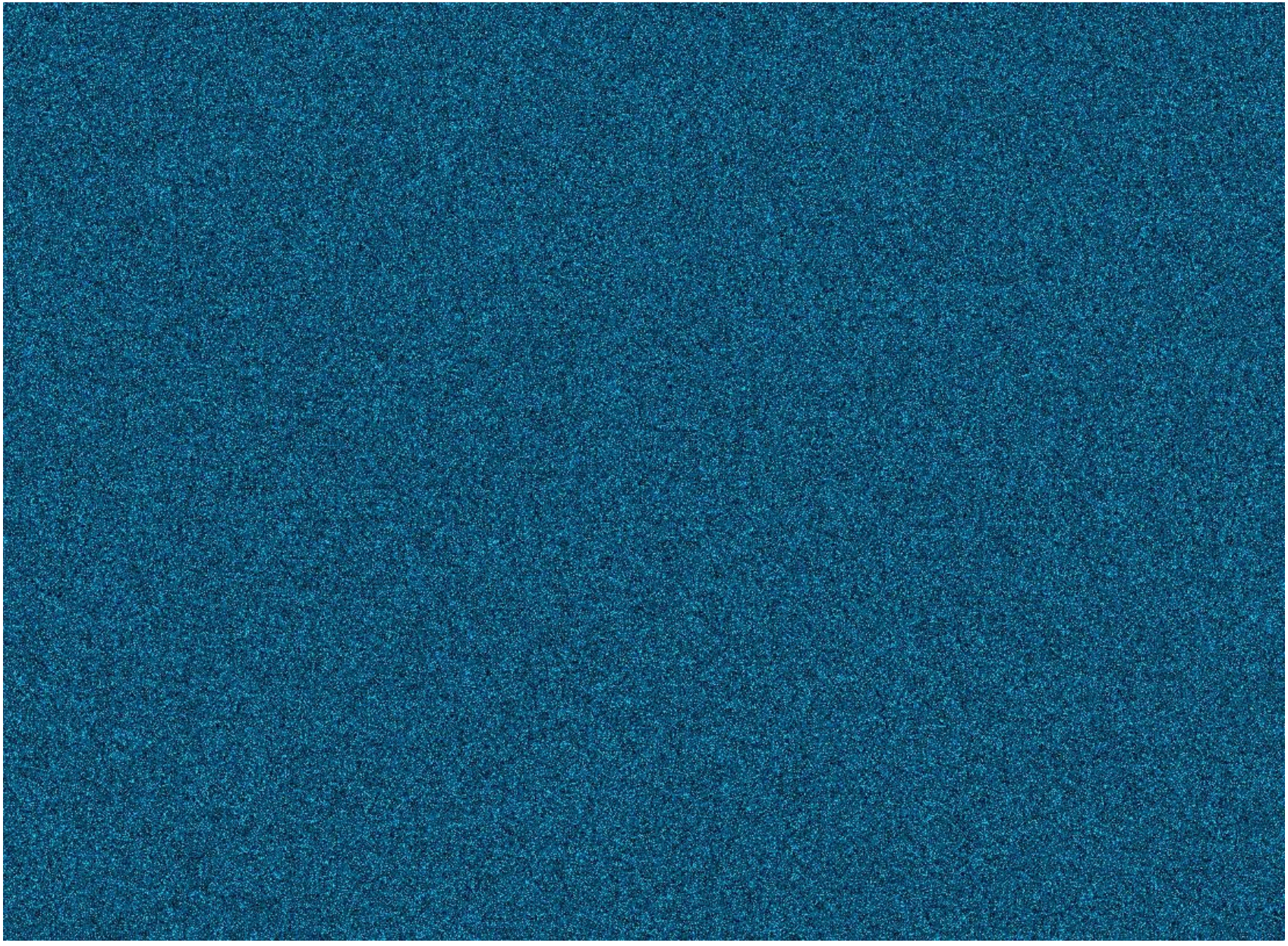
Matriisien satunnaistus

- Säilytetään rivien ja sarakkeiden arvojakaumat satunnaistuksessa *likimääräisesti*
- Tulosta pidetään mielenkiintoisena, jos rivien ja sarakkeiden arvojakaumat eivät selitä havaintoa

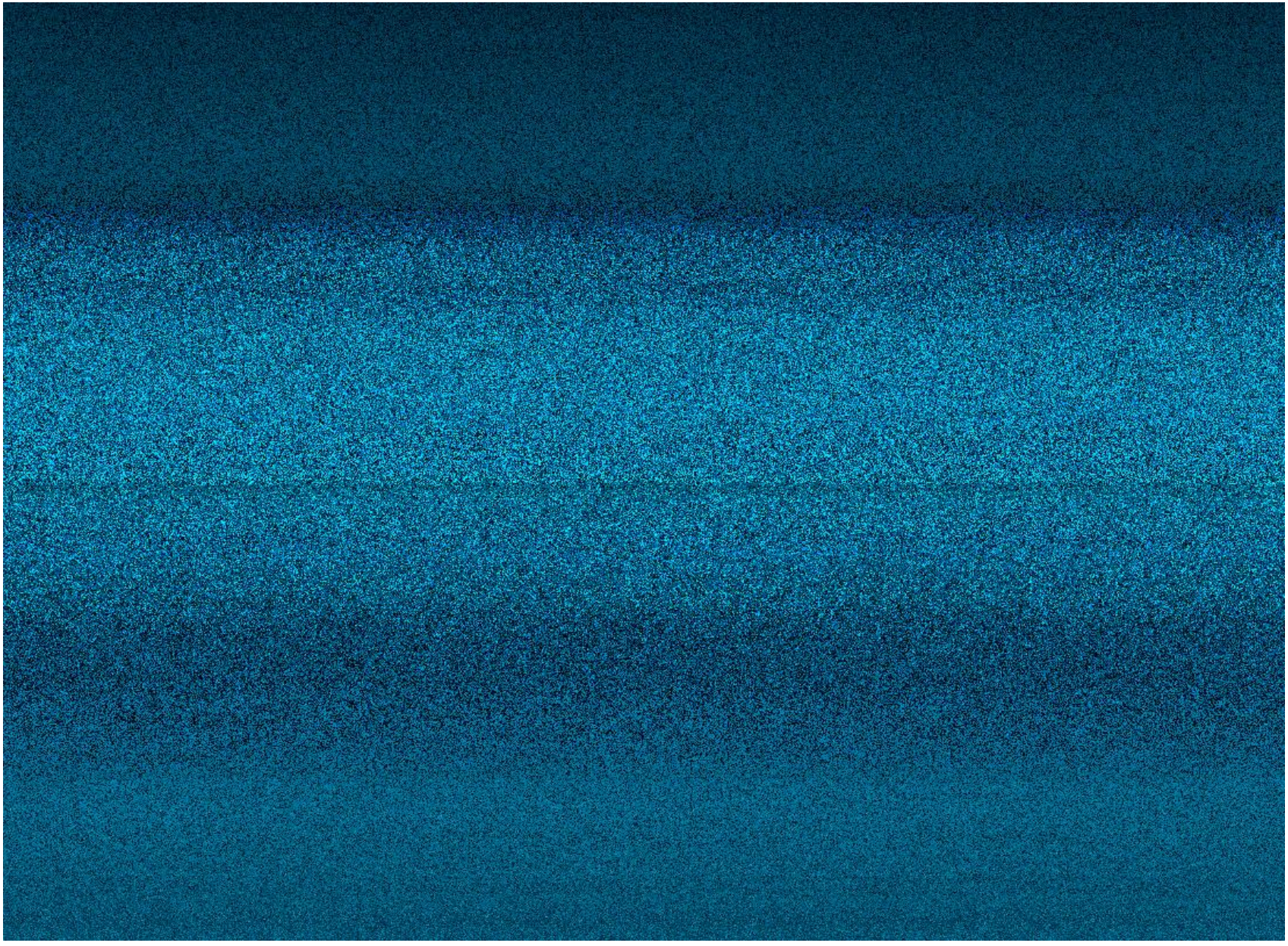
	Maito	Leipä	Banaani	Juusto	Kinkku
Kauppa 1	0.69	2.45	0.99	5.49	6.45
Kauppa 2	1.25	4.29	2.49	8.95	9.35
Kauppa 3	0.79	2.45	1.25	6.39	7.59
Kauppa 4	1.19	4.59	1.95	8.45	8.55
Kauppa 5	0.65	2.75	1.15	6.65	7.13
Kauppa 6	1.25	3.95	2.19	7.75	9.99

Kansikuva: Alkuperäinen

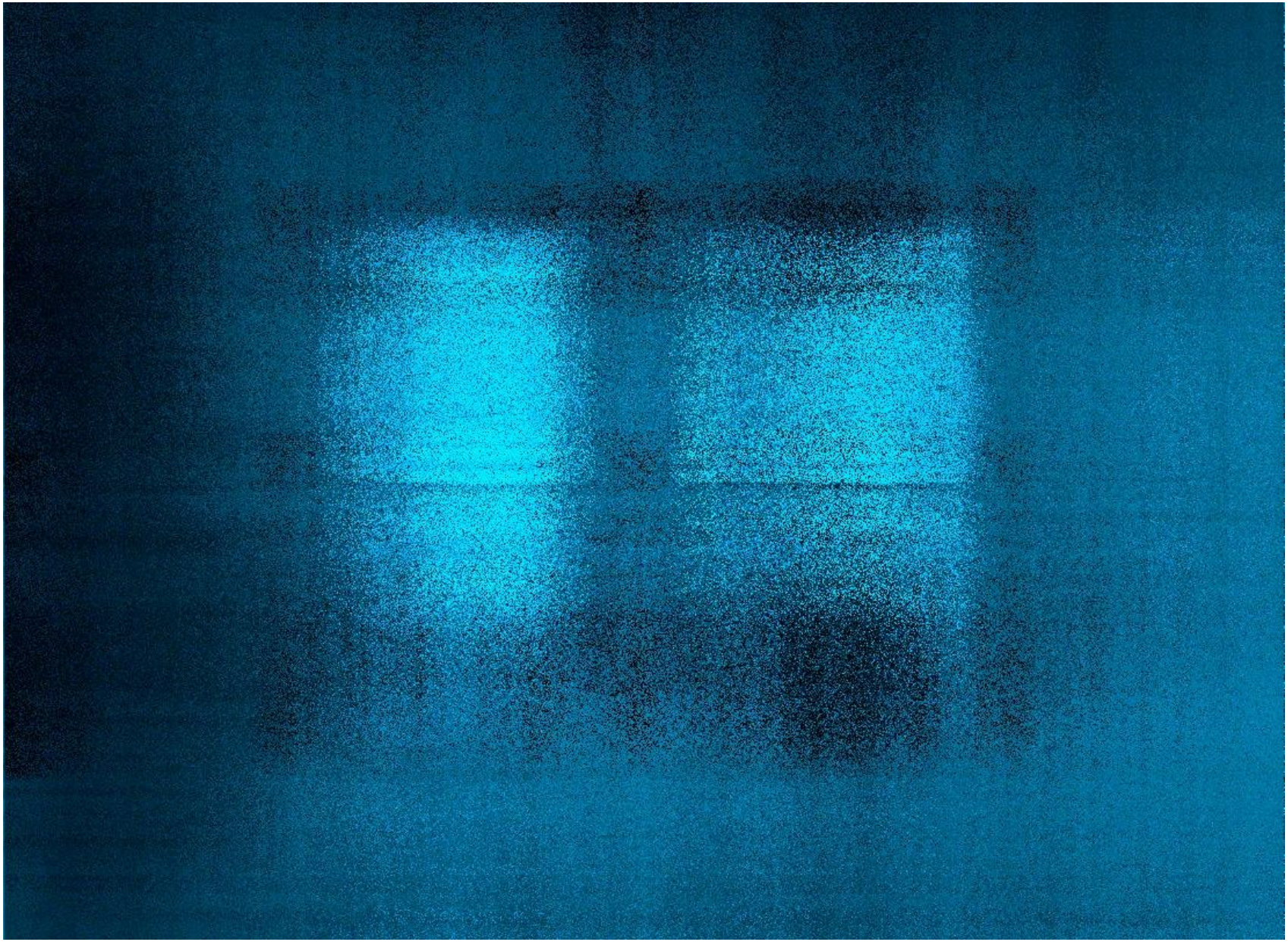
Kansikuva: Täyspermutaatio



Kansikuva: Rivipermutaatio



Kansikuva: Arvojakaumat



Yhteenveto

Merkitsevyystestaus satunnaistamalla

- Helppo ja monikäyttöinen
- Soveltuu erilaisiin tiedonlouhinnan ongelmiin

Väitöskirja

- Uusia satunnaistamistapoja tiedonlouhintaan
- Tulevaisuus: satunnaistuksen käyttö erilaisissa sovelluksissa

Seuraavaksi

Väitöskirjan merkitsevyyden testaaminen