Data-analyysi tieteenalana

Professori, laitosjohtaja Sasu Tarkoma
Tietojenkäsittelytieteen laitos
Helsingin yliopisto

Sisällys

Johdanto

Data-analyysi

Datatiede

Big Data –ratkaisut

Tutkimusesimerkkejä

Digitalisaatio

Teknologia mahdollistaa esineiden ja asioiden reaali-aikaisen seurannan ja säätämisen

Tilat, liikenne, teollisuus, ketjut ja verkostot

Kehittyneet data-analytiikkaratkaisut mahdollistavat uudenlaisen lisäarvon löytämisen datasta

Kone-oppiminen ja tekoäly tulevat muuttamaan toimialoja ja luomaan uusia. ETLA ennustaa, että 36% työnimikkeistä katoaa Suomessa tämän muutoksen seurauksena. Samalla syntyy myös uusia nimikkeitä, kuten datatieteen asiantuntija (data scientist)

Data-analyysi

Data sisältää informaatiota liittyen asiaan tai ilmiöön.

Tieto on hyvin perusteltu väite maailman tilasta.

Data-analyysi käsittää menetelmiä ja metodeja kerätyn **datan** jalostamiseen korkeamman tason **tiedoksi** tavoitteena hyödyllisten **johtopäätösten** tekeminen.

Data-analyysiä käytetään monilla alueilla hyödyntäen erilaisia menetelmiä ja lähestymistapoja.

Data-analyysin monet kasvot

Data-analyysi luo pohjan **datatieteelle**, joka käsittää sekä menetelmät että niiden soveltamisen eri alueilla.

Data-analyysin ja datatieteen tutkimusaloja ovat mm. hahmontunnistus (pattern recognition), tiedon louhinta (data mining) koneoppiminen (machine learning), bioinformatiikka (bioinformatics).

Moderni datatiede ja data-analyysi keskittyy “big data” aineistoihin, joissa keskiössä ovat erittäin suuret järjestelemättömät ja jatkuvasti lisääntyvät data-aineistot.

Data-analyysi

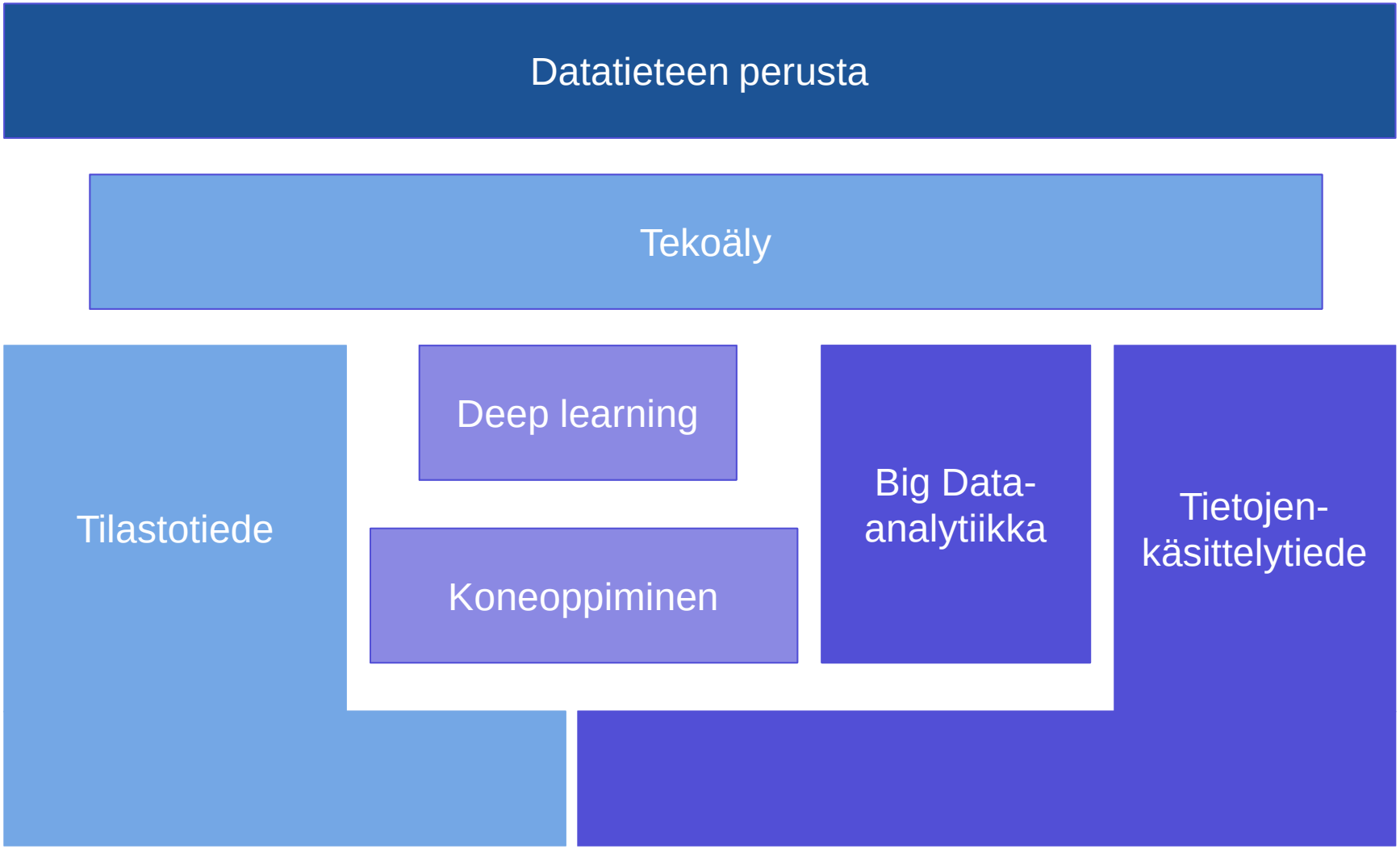
Tilastotiede

- Havaintoaineiston tilastollinen analyysi
- Tilastolliset mallit ja estimointimenetelmät
- Tutkitaan ilmiöitä, joissa esiintyy vaihtelua (variation) ja pyritään erottelemaan vaihtelun systemaattinen ja satunnainen osa toisistaan

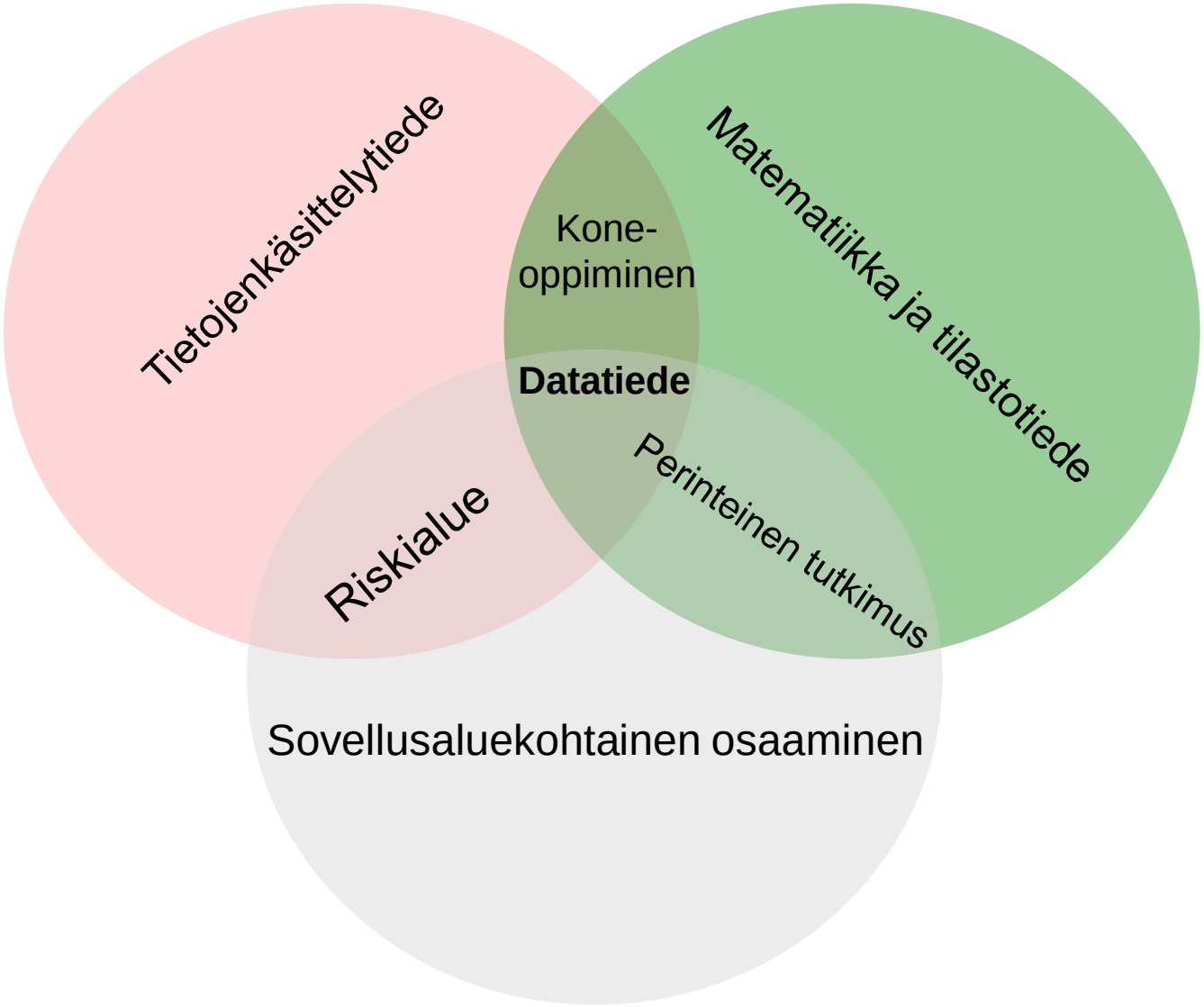
Tietojenkäsittelytiede

- Tiedon louhinta
- Koneoppiminen
- Neuroverkot, Deep learning
- Big Data

Käsitteitä


Datatieteen perusta
Tekoäly
Tilastotiede
Deep learning
Koneoppiminen
Big Data-analytiikka
Tietojen-käsittelytiede

Kohti datatieteen määritelmää

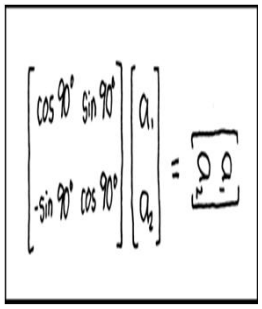


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Datatieteen käytäntö

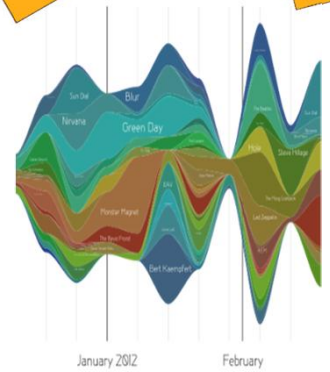
Datan tutkiminen

Datan esikäsittely



Mallin rakentaminen

Suuren mittakaavan hyödyntäminen



Evaluaatio
Tulkinta

Source: Databricks introduction slides

Data-analyysin haasteita

Skaalautuvuus ja Big Data

Menetelmien automatisointi

Reaaliaikaisuus

Tietosuoja

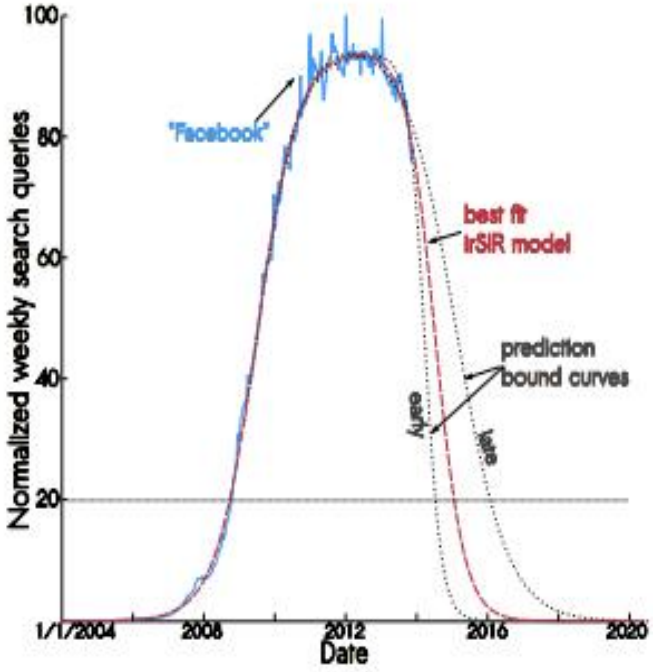
Keskitetty vs. hajautettu toiminta

Datan moniulotteisuus, rakenteettomuus, hälyisyys

Suuressa data-aineistossa (Big Data) on arvoa
Pieni data on myös arvokasta!

Korrelaatio ja kausaatio

- Princetonin tutkijat ennustivat Google hakudatan (MySpace hakusanana) rakennetun mallin avulla, että **Facebookin käyttö vähenee nopeasti** muutamassa vuodessa
- Facebook käytti samaa mallia: **Princetonilla ei ole opiskelijoita 2021**
  - http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/
- Google Flu Trends toimi alkuun, mutta sittemmin ennusteet eivät ole toimineet hyvin.
  - Hakusanapohjainen malli on liian yksinkertainen.
- Pelkkä datavetoisuus ei riitä, tarvitaan tilastollinen validius ja merkitsevyys



Haut sanalle “Facebook”

Laskentakeskukset ja pilvilaskenta

TRADITIONAL WORKLOADS

CLOUD WORKLOADS

Palvelimet

Palvelintehokkuuden lisääminen (scale up)

Kalliit työvälineet luotettavuuden lisäämiseen

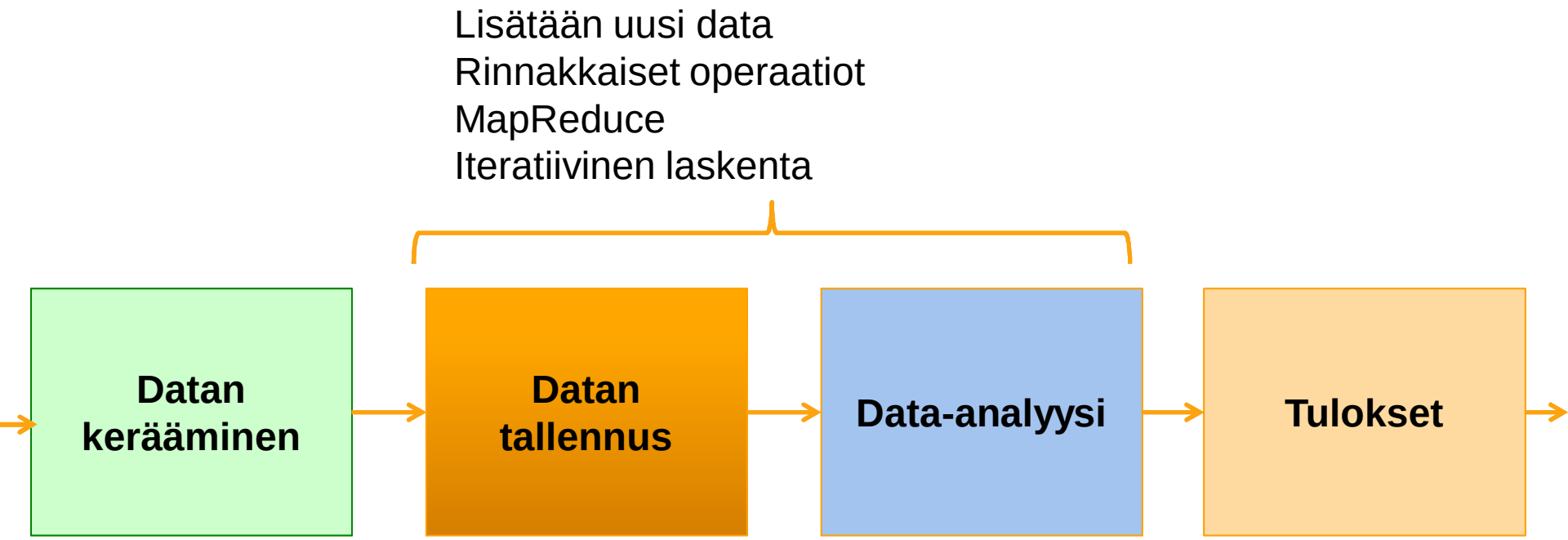
Virtualisoidut elastiset resurssit

Hajautetut konesalit

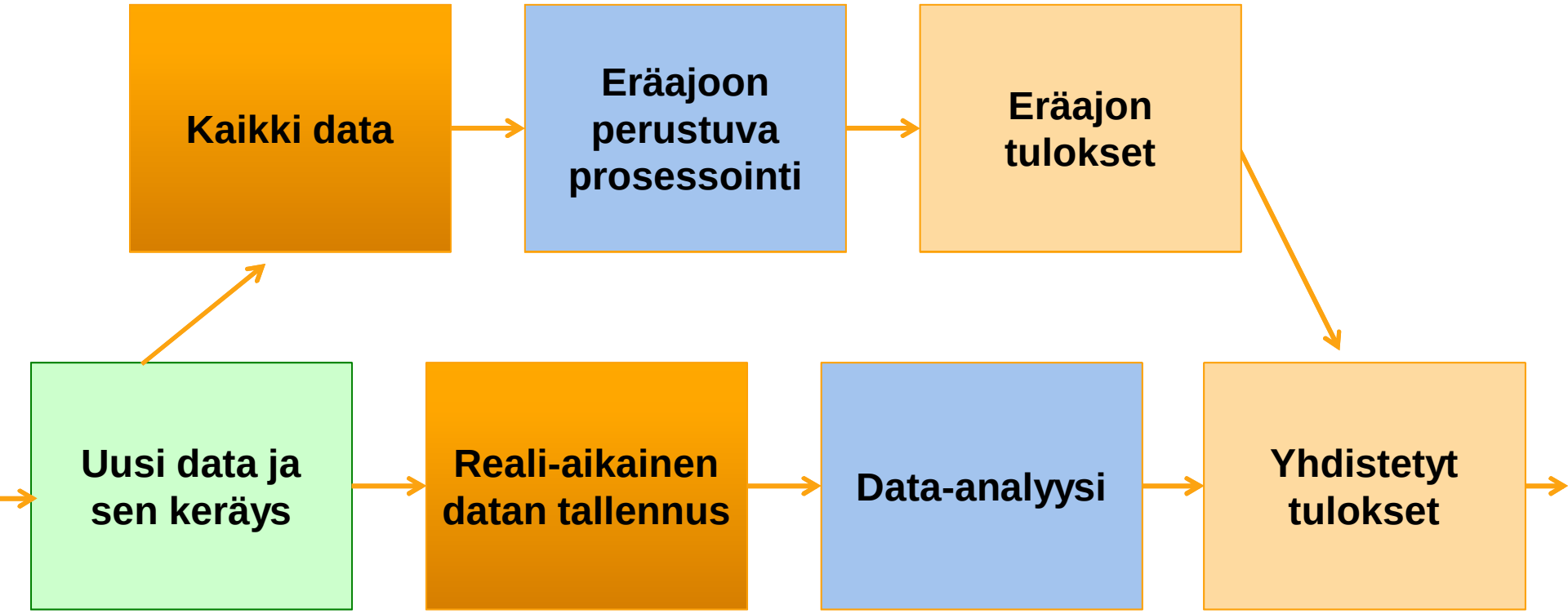
Vikasietoisuus sisäänrakennettuna

Sovellukset skaalautuvat virtualisoidussa ympäristössä (scale out)

Esimerkkejä Big Data -ratkaisuista

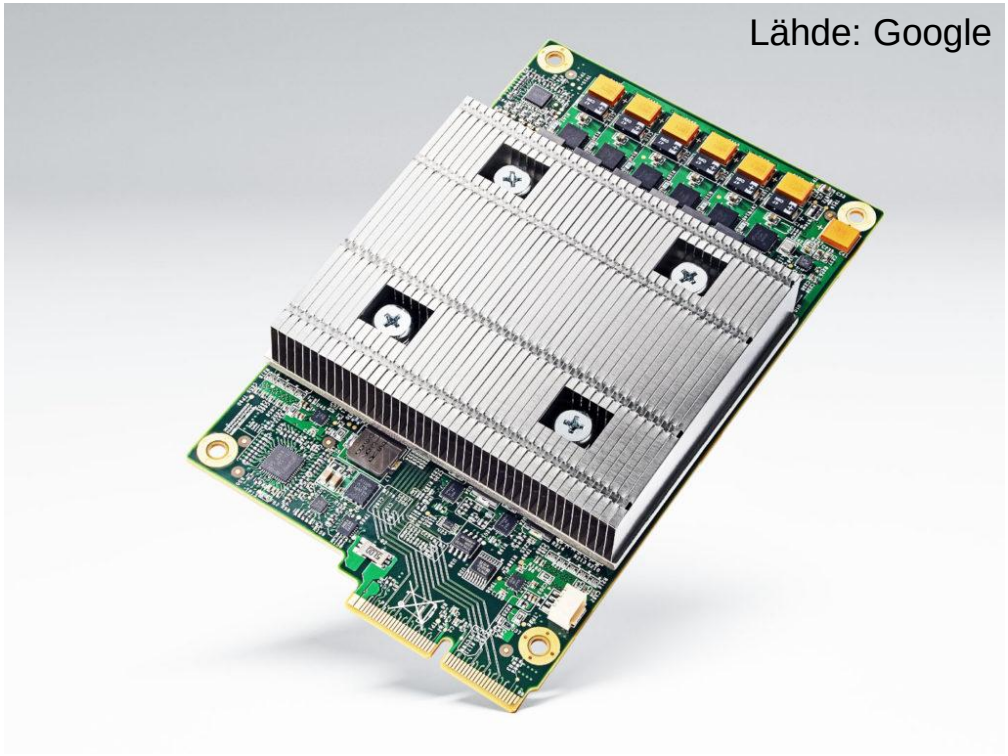


Lambda-arkkitehtuuri



Integroituna esimerkiksi Apache Sparkissa

Google Tensor Processing Unit



15-30x nopeampi kuin perinteinen CPU deep learning tehtäviin

Data-analyysin automatisointi

Uudet kehittyneet menetelmät pyrkivät data-analyysin automatisointiin

Tiedon keruun automatisointi

Algoritmien valinta ja parametrisointi

Tulosten hyödyntäminen reaali-aika sovelluksissa

Tutkimusesimerkkejä

Kohti esineiden Internetin reaaliaikaista analytiikkaa

Reunalla tapahtuva analytiikka

Big Data -kehikot

Reaali-aikainen käsittely

Reaali-aikainen käsittely

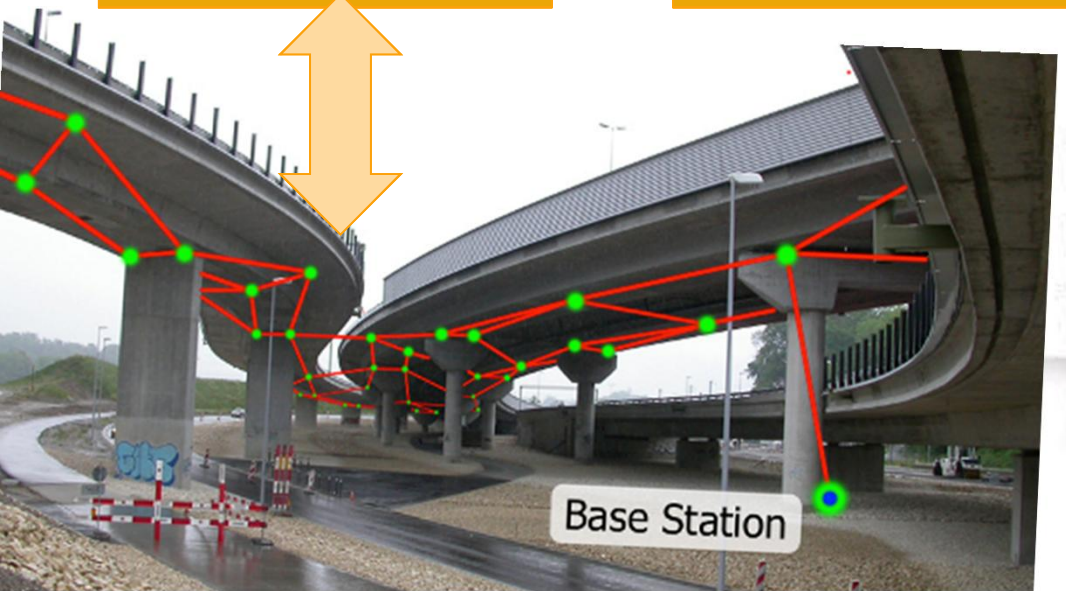
Reaali-aikainen käsittely

Eräajo-käsittely

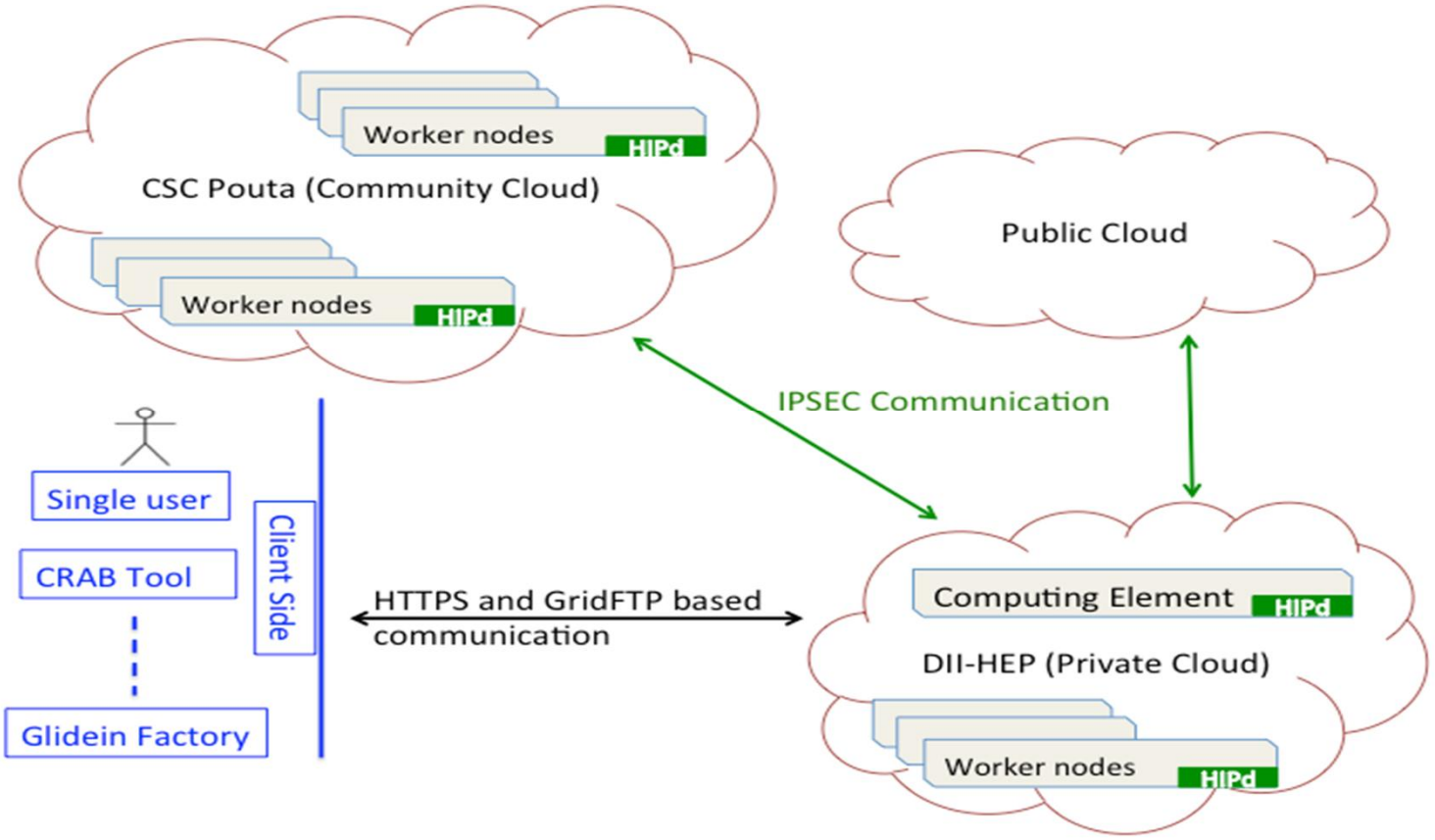
Datan kerääminen ja prosessointi

Datan välitys ja prosessointi verkossa (4G ja 5G)

Datan prosessointi pilvialustan ja Big Data –kehikon avulla hajautetussa ympäristössä



Platform View: Secure Scientific Cloud for High Energy Physics



L. Osmani, S. Toor, M. Komu, M. J. Kortelainen, T. Lindén, J. White, R. Khan, P. Eerola, S. Tarkoma. **Secure Cloud Connectivity for Scientific Applications**. IEEE Transactions on Services Computing, 2015.

Carat

- Yhteistyöprojekti UC Berkeleyn ja Helsingin yliopiston välillä
- Ilmainen mobiilisovellus Androidille sekä iOSille
- Yli 850 000 käyttäjää
- Yli 5 vuotta dataa
  - >2,5 TB dataa, > 250 miljoonaa näytettä
- Yli 450 000 eri sovellusta
- Tutkimusprojektilla on monta suuntaa
- http://carat.cs.helsinki.fi

UNIVERSITY OF HELSINKI

amplab UC BERKELEY

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

data mining
data analysis
statistics
machine learning
distributed computing

Methods

natural sciences and engineering
life sciences and medicine
humanities, social sciences
pedagogics, economics

Applications

HiDATA

Helsinki Centre for Data Science

Research results

Education of experts

Innovations

Information for the general public

Yhteenveto

Rakennamme datavetoista reaaliaikaista digitaalista infrastruktuuria.

Datatiede kehittää menetelmiä ja mahdollistaa niiden laajan monitieteisen soveltamisen.

Data-analytiikan ja tekoälyn alustat ovat kehittyneet paljon viime vuosina.

Data-analytiikan ja datatieteen automatisointi on vasta alussa.