ITFakultetet

Introduksjon til Big Data

Terje Berg-Hansen

Stavanger 28. august 2019

ITFakultetet

Hva er Big Data?

«Big Data» is like Teenage Sex:

- Everybody is talking about it
- Nobody knows how to do it
- Everybody assumes that all the others are doing it
- So they claim they, too are doing it

ITFakultetet

Litt mer formelle definisjoner

- «Big Data inkluderer vanligvis datasett som er større enn det vanlig brukte verktøy kan håndtere når det gjelder datafangst, konvertering og prosessering innen rimelige tidsrammer»
- «Big Data er data som karakteriseres av så stor mengde (**volume**), hastighet (**velocity**) og mangfold (**variety**) at det er nødvendig med spesielle teknologier og analysemetoder for å skape verdi av dem»
- «Big Data er data som må behandles ved parallell prosessering over flere maskiner»

ITFakultetet

Big Datas fire V'er

Volume

- Mengden data som produseres og lagres. Mengden avgjør verdien og potensiale for innsikt, og om det kan kalles Big Data eller ikke. Terabytes og Petabytes med data er vanlig.

Variety

- Datas type og natur. Big Data hentes fra tekst, bilder, audio, video osv. - og har evnen til å fylle inn manglende biter gjennom datafusjon.

ITFakultetet

Big Datas fire V'er

Velocity

- Hastigheten data genereres og prosesseres med. Hvor ofte genereres data, og hvor ofte behandles og publiseres data? Big Data er ofte tilgjengelig i sanntid og produseres kontinuerlig.

Veracity

- «Sannhetsgehalt». Utvidet definisjon av Big Data som refererer til datakvalitet og dataverdi. Kvaliteten i datafangsten kan variere og påvirke analysene i stor grad

ITFakultetet

Typer Big Data

- **Strukturerte data**
  - Databaser, Loggfiler, Sensordata osv
- **Semi-strukturerte data**
  - Twitter-feeds, Data som XML / JSON osv
- **Ustrukturerte data**
  - Tekstmeldinger, Eposter, Tekstdokumenter osv.
  - Stadig større andel av datamengden er ustrukturert

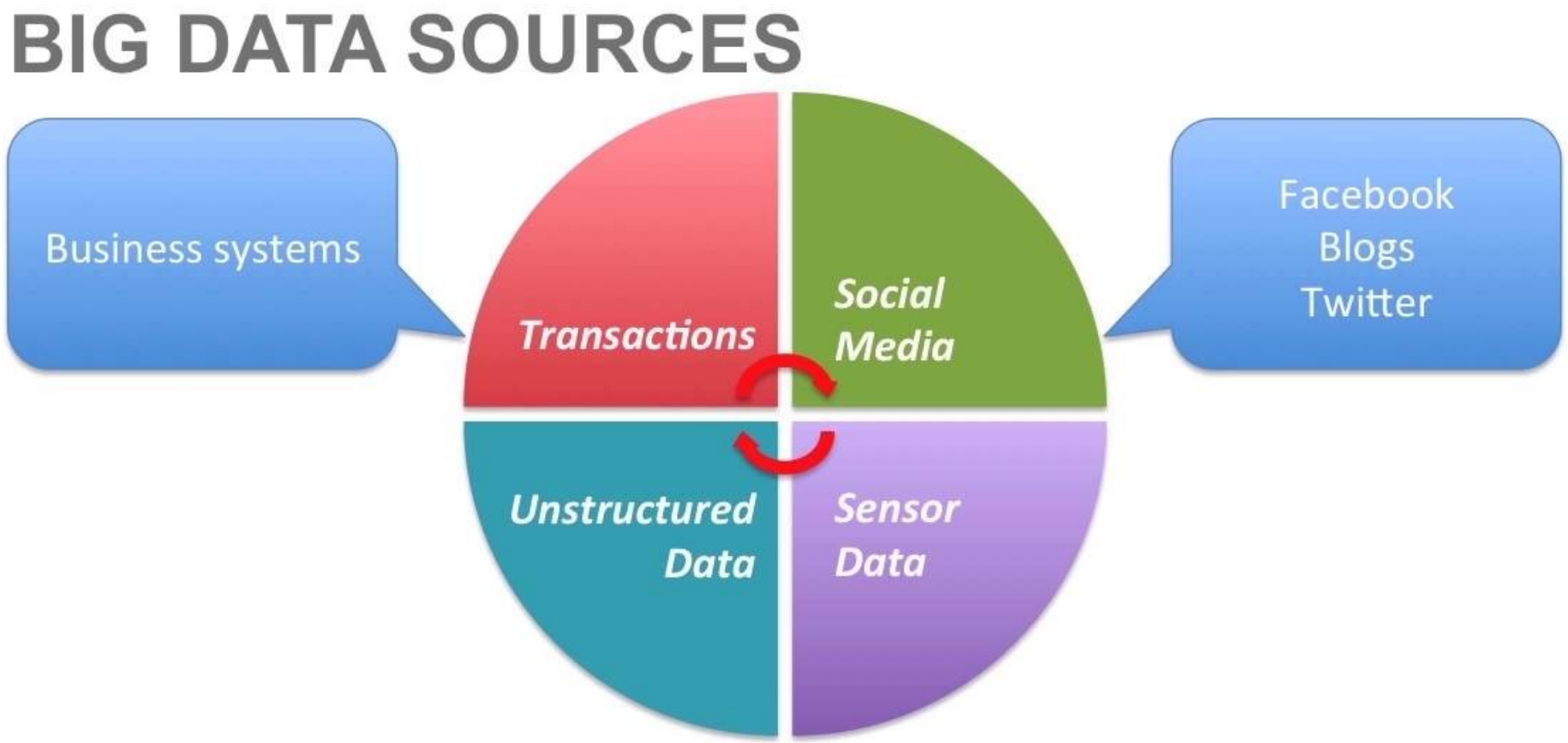
ITFakultetet

To hovedtyper data

- «Data at rest»
  - Data fra Databaser og data-filer
  - Data fra lagrede dokumenter, bilder, audio, video
- «Data in Motion»
  - Strømmer av Logger, web clicks, twitter-meldinger
  - Sanntids annonse-matching
  - Sensordata (IoT)
  - Online kredittkort-sjekker

ITFakultetet

Big Data-kilder



ITFakultetet

IDCs Prediksjoner for 2025

- Data går fra forretnings-kritisk til livs-kritisk.
  - I 2025 vil nær 20% av data globalt være kritiske for menneskers dagligliv og nær 10% av dette vil være livskritiske data.
- Innebygde systemer og Internet of Things (IoT).
  - I 2025 vil en gjennomsnitts-person interagere med oppkoblede gjenstander nær 4.800 ganger daglig – dvs en interaksjon hvert 18. sekund.
- Maskinlæring endrer landskapet
  - IDC beregner at mengden i den globale data-sfæren som blir analysert vil ganges med 50 til 5.2 ZB i 2025.

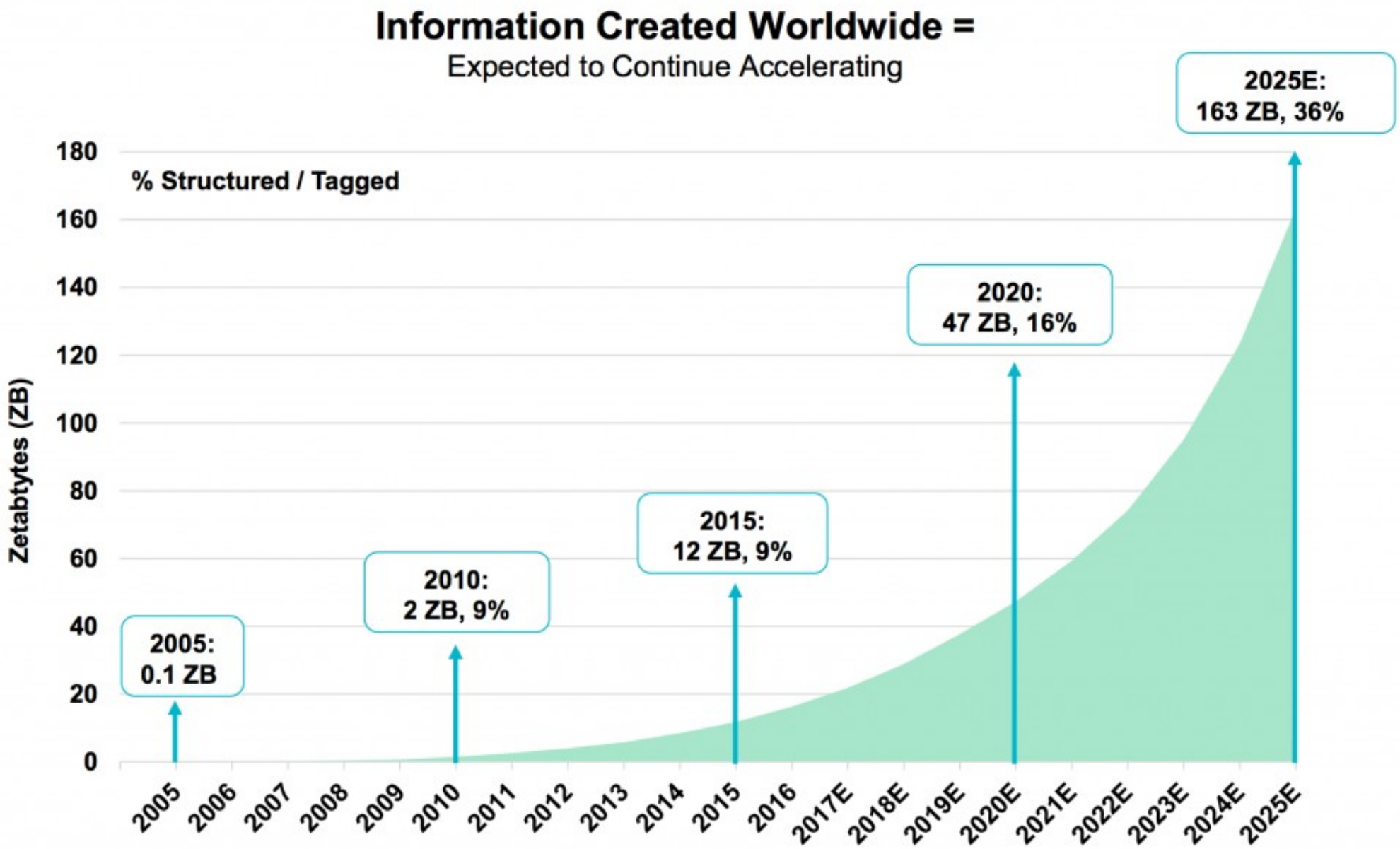
ITFakultetet

IDCs Prediksjoner for 2025

- Data i sanntid.
  - I 2025 vil mer enn en fjerdedel av alle genererte data være sanntidsdata, og 95% av dette vil være sanntidsdata fra IoT.
- Automatisering og maskin-til-maskin teknologier vil generere brorparten av data.
  - Mens veksten i data generert de siste 10 årene primært har kommet fra underholdnings-industrien, vil neste tiår se et skifte mot produktivitets-data og data fra innebygde enheter, som sensordata, video fra overvåknings-kameraer osv.

ITFakultetet

IDCs Prediksjoner for 2025



ITFakultetet

Trender i Big Data

- Strømming av Internet of Things for Maskinlæring
  - Bruk av IoT til å kombinere strøm-analyser med maskinlæring
  - Maskinlæring bruker typisk lagrede data til trening, i et kontrollert læringsmiljø. Med denne nye modellen brukes strømmede data fra «Internet of Things» til å tilby maskinlæring i sanntid i et mindre kontrollert miljø.

ITFakultetet

Trender i Big Data

- Kunstig Intelligens-plattformer
  - Bruk av Kunstig Intelligens-plattformer til å prosessere Big Data for å generere Business Intelligence.
  - KI-plattformer består av 5 logiske lag:
    - Data & Integrasjons-laget gir tilgang til data
    - Eksperimenterings-laget lar Data Scientists utvikle, teste, og bevise hypoteser.
    - Drift og Provisjonerings-laget gir tilgjengelighet
    - Intelligens-laget gir intelligente AI-tjenester .
    - Erfarings-laget interagerer med brukere via f.eks.: «augmented reality», «conversational UI» og «gesture control».

Noen hovedeffekter av BD

- Distribuerte systemer er blitt nødvendig

  Skalere *ut* istedenfor å **skalere *opp***

  - Distribuerte filsystemer
  - Distribuerte Databaser
  - Distribuert parallell-prosessering for Data Science og Data Wrangling
- Bruken av skytjenester har økt dramatisk
  - Offentlige skytjenester
  - In-house skytjenester
  - Hybride skytjenester

ITFakultetet

Trender Big Data har skapt

- Ny IT-rolle: Datakurator
  - Ansvar for å organisere bedriftens Metadata, Datasikkerhet, «Data Governance» og Datakvalitet.
  - Datakuratoren har ansvar for å forstå hvilke analysetyper ulike grupper i organisasjonen har behov for, hvilke datasett som er egnet for disse analysene, og hvilke trinn som er involvert i å få data fra sin rå form til den form og tilstand som trengs for den jobben en datakonsument skal gjøre. Datakuratoren bruker systemer som «self-service data platforms» for å fasilitere brukernes tilgang til data uten å lage uendelige kopier av datasettene.

Trender Big Data har skapt

- Funksjonell programmering
  - Big Data medfører parallell-prosessering i nettverk med mange servere. Funksjonell programmering er spesielt godt egnet til dette, f.eks. ved å vektlegge: «Minimize Mutable State».

    Ved å bruke konstanter istedenfor variabler unngår man bieffekter som kan ødelegge prosesseringen.
  - Java, Scala, Python osv. Støtter funksjonell programmering i større grad i hver ny versjon.

ITFakultetet

Trender Big Data har skapt

- Paradigmeskifte i utvikling og bruk av programvare
  - Programvare utvikles av IT-avdelinger i bedrifter som ikke lever av å selge programvare, som Facebook, Twitter, LinkedIn, Google osv.
  - Utstrakt samarbeid og deling via **Open Source – organisering og -lisensiering**, som regel gratis.
  - Pragmatisk valg av programvare:
    - Fra «Vi bruker Oracle til alt» til «Vi bruker den databasen som passer best til dette formålet»

ITFakultetet

Trender Big Data har skapt

- Paradigmeskifte i bruk av Maskinvare
  - Billige, standard servere og komponenter
  - «Redundancy» og «High availability» ved ekstensiv replikering i klynger av servere med distribuerte filsystemer.
  - «Sharding» ved partisjonering av datasett over mange harddisker, mange servere og mange datasentere.
  - Bruk og kast – kjøp billig maskinvare og bytt ut når de går istykker.