Hensikten med Disruptive Perspektiver:

Når vi analyserer ulike temaer bruker vi mye tid og mange verktøy (kvartalsrapporter, analyser, dialog med selskapene, bedriftsbesøk, excel, kalkulator og ordmodeller). Ofte lager vi små notater, og noen ganger store notater som vi tenker på som perspektiver. Vi er gamle nok til å vite at det sjeldent finnes sannheter, ofte bare ulike perspektiver.

Disruptive Perspektiver har kun én hensikt: Å dele våre perspektiver på temaer som former vår fremtid. Dette er ikke akademiske notater, innlegg til et leksikon eller anbefalinger om å gjøre noe, kjøpe eller selge noe. Kun god gammeldags informasjonsdeling for å synliggjøre hvordan vi ser på ulike temaer på publiseringstidspunktet. Perspektiver blir ikke mindre, kanskje heller mer, når man deler det. Med det utgangspunktet; ha en fin reise i våre perspektiver.

Innhold

Disclaimer

Innholdet i denne artikkelen er ikke ment som investeringsråd eller anbefalinger. Har du noen spørsmål om fondene det refereres til, bør du kontakte en finansrådgiver som kjenner deg og din situasjon. Husk også at historisk avkastning i fond aldri er noen garanti for fremtidig avkastning. Fremtidig avkastning vil blant annet avhenge av markedsutvikling, forvalterens dyktighet, fondets risiko, samt kostnader ved kjøp, forvaltning og innløsning. Avkastningen kan også bli negativ som følge av kurstap.

1.0 Data som den nye oljen i AI-æraen

I en tid der AI transformerer utviklingen innen selvkjøring og humanoids, er dataen ikke lenger bare en ressurs. Den er selve drivkraften bak innovasjon og konkurransefortrinn. Som Larry Ellison, grunnlegger av Oracle, har sagt: «*It's all public data from the internet … So they're all basically the same. And that's why they're becoming commoditized so quickly*». Hans uttalelser belyser skillet mellom offentlige og private data, og hvordan det påvirker AI-modellers verdi på kort og lang sikt. For å gjøre dette mer konkret presenterer vi en 2x2-matrise basert på Jay Barneys *Resource-Based View* (RBV). Ifølge denne teorien kommer varige konkurransefortrinn ut av ressurser og kapabiliteter som er verdifulle, sjeldne, vanskelige å imitere og ikke lett kan erstattes. På skolen er dette kjent som *VRIN*-rammeverket (eller VRIO). Den viser at private data kan være en kilde til varige konkurransefortrinn, mens offentlige data oftere bidrar til at selskaper blir mer like.

Postcards From The Future

Data: Type data og verdi

	Offentlig data Delt & Generisk	**Private data** Proprietære domenspesifikke
Kort Sikt Operasjonell verdi	**Høy tilgjengelighet for rask modelltrening (f.eks. Wikipedia-data for generalist-AI); lav kostnad, men begrenset differensiering**	**Begrenset bruk pga. privacy, men umiddelbar nytte i nisje-applikasjoner (f.eks. bankdata for fraud-deteksjon).**
Lang sikt Strategisk verdi	**Leder til commoditization og konvergering av modeller, som Ellison advarer.**	**Skaper VRIN-ressurser for disrupsjon, f.eks. via RAG for sikre analyser.**

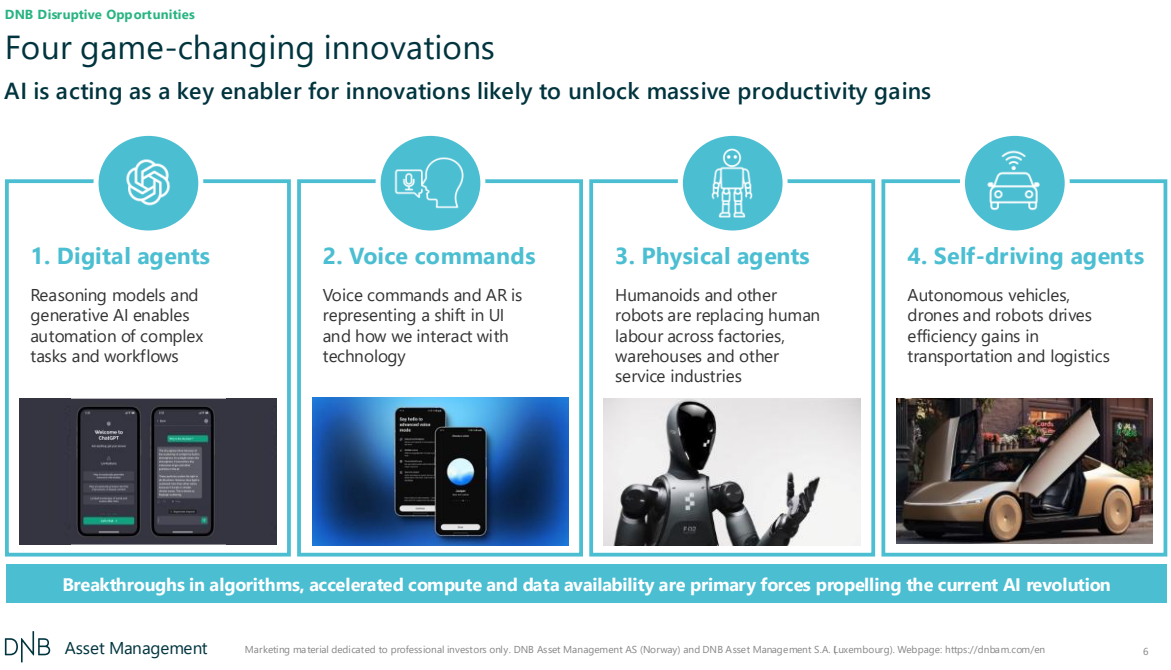
ent

Denne matrisen setter grunnlaget for vår analyse, som bygger på diskusjoner om data-kategorisering, Big Data-prinsipper og AI-applikasjoner. Vi vil videre utforske hvordan data «mates» til nevrale nettverk, risikoen ved «*garbage in, garbage out*» (GIGO), og et integrert rammeverk for fremtidig verdi.

2.0 Teoretisk Perspektiv

I innledningen etablerte vi dataens rolle som den nye «oljen» i AI-æraen. Dette kapittelet handler om det overordnede teoretiske perspektivet for å forstå temaet. I en eksponentiell verden, der teknologisk vekst følger Moores lov og dobler kapasiteten hvert annet år, blir lineære forretningsmodeller, som antar at fortid = fremtid, disruptert.

Digitale og fysiske agenter «spiser» og vokser på tre essensielle ressurser. Det er silisium (halvledere), data og energi. Vi fokuserer på data som sentral ressurs, og integrerer teori fra Schumpeter, Christensen, *Blue Ocean Strategy*, Porter, nettverkseffekter, *flywheel* og Kuhn. Disse teoriene belyser hvordan data driver disrupsjon, skaper nye markeder og akselererer eksponentiell vekst, mens de utfordrer tradisjonelle lineære tilnærminger. Vi tar oss friheten til en kort oppsummering.



Schumpeters *Kreative ødeleggelse*

Joseph Schumpeter introduserte begrepet *creative destruction* i 1942 for å beskrive hvordan innovasjon river ned gamle strukturer og baner vei for nye. For Schumpeter var dette selve kjernen i kapitalismen. I en eksponentiell verden gjelder denne dynamikken også for data. Offentlige datasett (som innhold på det åpne internettet) svekkes gradvis av dataforurensning fra AI-generert materiale, mens private og syntetiske data danner grunnlaget for nye verdikjeder.

Lineære modeller, som bygger på antakelsen om stabil vekst og historiske mønstre, blir utfordret av eksponentielle innovasjoner. Et godt eksempel er RAG-teknologi, som står for *Retrieval-Augmented Generation*. Enkelt forklart betyr det at modellen ikke bare svarer ut fra det den har lært på forhånd, men også kan hente inn relevant informasjon fra et privat datalager i sanntid (f.eks interne dokumenter, kundedata eller selskapskunnskap). På den måten kan virksomheter bruke egne data til å gjøre AI'en langt smartere og mer presis, uten at informasjonen må deles eller trenes åpent inn i modellen. Det gjør private data til en strategisk *moat*. Data blir dermed en sentral drivkraft i det Schumpeter kalte *industrial mutation*. AI-agenter vokser på data som drivstoff, bryter opp tradisjonelle industrier og legger grunnlaget for nye økosystemer.

Christensens *Disruptiv innovasjon*

Clayton Christensen definerer disrupsjon som en prosess der enklere, billigere produkter starter i lav-ende nisjer og beveger seg oppover, og utfordrer etablerte ledere. I konteksten av data gjelder dette skiftet fra offentlige, generiske datasett (*inkrementell innovasjon*) til private og syntetiske data (*disruptiv*). Lineære modeller, som baserer seg på historisk dataanalyse, blir disruptert av eksponentielle teknologier som edge AI, der agenter prosesserer data i sanntid uten sentralisert trening. For eksempel starter syntetiske data i nisjer som f.eks simulering av mindre vanlige

drone-scenarier, før de skalerer og gjør tradisjonelle datainnsamlingsmetoder overflødig. Dette akselererer etterspørselsveksten for silisium (GPU'er for generering) og energi (compute-intensiv prosessering), og understreker at data er en disruptiv kraft som omformer markeder.

Blue Ocean Strategy

Blue Ocean Strategy handler om å skape nye markeder gjennom differensiering og lav kostnad, i stedet for å konkurrere i *red oceans* av blodig rivalisering. Anvendt på data innebærer dette å flytte fra *commoditized* offentlige datasett til nye markeder der private data kombineres med syntetisk for å skape unik verdi, som i RAG-baserte systemer for enterprise-AI. Eksponentiell vekst disrupter lineære modeller ved å gjøre data til en såkalt «*value innovation*», for eksempel ved å bruke multimodal data (video + sensor) for humanoids, som åpner nye markeder uten direkte konkurranse. Agenter vokser her ved å bruke data som drivstoff, mens silisium og energi støtter skaleringen, og skaper bærekraftig vekst uten å slåss om eksisterende ressurser.

Porters *Fem krefter* og konkurransestrategi

De fem kreftene fra Porter analyserer industristruktur gjennom trusler fra nye aktører, leverandører, kjøpere, substitutter og rivalisering mellom dem, for å forstå konkurranseintensitet og profittpotensial. I en eksponentiell data-drevet verden øker trusselen fra nye aktører (startups med syntetisk data), mens leverandørkraft (databaser som Oracle) styrkes av privat data-moat. Lineære modeller undervurderer substitutter som AI-generert data, som disrupter tradisjonell innsamling. Porters ramme viser hvordan data endrer konkurranselandskapet. Agenter som bruker data reduserer rivalisering ved å skape differensierte nisjer, mens halvledere og energi blir kritiske input som påvirker leverandørkraft og profitt.

Metcalfes lov og *Nettverkseffekter*

Metcalfes lov sier at en nettverks verdi er proporsjonal med antall brukere av plattformen, opphøyd i andre (n^2). Poenget er eksponentielle nettverkseffekter. For data gjelder dette økosystemer der flere brukere (eller agenter) genererer mer data, som akselererer vekst. En positiv *feedback*-loop. I AI-verdenen vokser agenter på data som nettverksverdi. Flere sensorer i droner skaper eksponentielt mye bedre forståelse, mens halvledere og energi støtter skaleringen. Dette forklarer hvorfor data-plattformer som Oracle eller Tesla vokser raskere enn lineære konkurrenter, og understreker dataens rolle i å skape «*winner-takes-all*»-markeder.

Jim Collins' *Flywheel-effekt*

Flywheel-effekten beskriver hvordan små, konsistente fremskritt bygger momentum over tid, som et tungt svinghjul som akselererer eksponentielt. I data-kontekst starter det med innsamling av primærdata, som bygger på seg selv til å skape *moats*. Som Amazons *flywheel* der mer data fører til bedre AI, som tiltrekker mer data. Eksponentiell vekst disrupter lineære modeller ved å gjøre data til en selvforsterkende syklus.

Thomas Kuhns *Paradigmeskift*

Kuhn beskrev paradigmeskifter som revolusjoner der gamle vitenskapelige rammer erstattes av nye når anomalier akkumuleres, ofte drevet av sosiale faktorer. Anvendt på data representerer skiftet fra lineære, historiske analyser til eksponentielle, data-drevne AI et paradigmeskifte. Gamle modeller kollapser under anomalier som model collapse (altså at kvaliteten svekkes når modellene i økende grad trenes på AI-generert data), mens nye paradigmer (f.eks RAG) dominerer.

Samlet sett integrerer disse teoriene et perspektiv der data i en eksponentiell verden disrupter lineære modeller, og posisjonerer våre fire agenter som sentrale aktører som vokser på silisium, data og energi. Dette kapittelet setter scenen for de påfølgende analysene av data-kategorisering og AI-applikasjoner, og understreker at suksess krever å omfavne disrupsjon fremfor å motstå den.

3.0 Definisjoner og kategorisering av data

For å forstå hvorfor data er blitt en av de viktigste ressursene i AI-økonomien, må vi først skille mellom ulike typer data. Ikke all data har samme verdi. Noen data er lett tilgjengelige for alle, andre er beskyttet, proprietære og vanskelige å kopiere. Noen data er ryddige og strukturerte, andre ligger gjemt i tekst, bilder, video, lyd og sensorer. Og nettopp disse forskjellene avgjør hvor nyttige dataene er for kunstig intelligens.

En grunnleggende inndeling går mellom offentlige og private data. Offentlige data er informasjon som er fritt tilgjengelig, ofte hentet fra åpne kilder som Wikipedia, Reddit, offentlige databaser eller det åpne internettet. Slike data har vært avgjørende for treningen av store språkmodeller, fordi de finnes i enorme mengder og dekker et bredt spekter av temaer. Samtidig har de klare begrensninger. De kan være fulle av feil, støy, skjevheter og gjentakelser. Når mange modeller trenes på de samme åpne datakildene, øker også risikoen for at modellene blir mer like hverandre. Data som alle har tilgang til, gir sjelden varig konkurransefortrinn (moat).

Private data er annerledes. Dette er data som eies av individer, selskaper eller institusjoner, og som ofte er beskyttet av lover, avtaler eller interne sikkerhetskrav. Eksempler kan være medisinske journaler, finansielle transaksjoner, kundedata, produksjonsdata, forsyningskjededata eller interne dokumenter. Slike data er ofte mer verdifulle fordi de er domenespesifikke, ferskere og tettere på virkelige beslutningsprosesser. De kan gi AI-systemer en dypere forståelse av et konkret

problem, en bestemt bransje eller en spesifikk organisasjon. Samtidig er de vanskeligere å bruke, nettopp fordi de kan være sensitive og underlagt regelverk som GDPR eller HIPAA.

En annen viktig dimensjon er forskjellen mellom strukturert og ustrukturert data. Strukturert data er data som er organisert i et fast format, for eksempel tabeller med rader og kolonner i en database. Dette kan være salgsdata, lagerstatus, regnskapstall eller kundetransaksjoner. Fordelen med strukturert data er at den er relativt enkel å søke i, analysere og bruke i tradisjonelle modeller. Den passer godt til presise beregninger, prognoser og automatiserte beslutninger.

Ustrukturert data er mer kaotisk, men ofte langt rikere. Dette er data som ikke følger et fast skjema, som tekst, bilder, video, lyd, e-poster, PDF-er, sosiale medier-innlegg og sensordata. For mennesker er denne typen informasjon ofte lett å tolke, men for maskiner har den historisk vært vanskeligere å utnytte. Det er her moderne AI virkelig har endret spillereglene. Store språkmodeller, datavisjon og multimodale modeller gjør det mulig å hente verdi ut av data som tidligere var for rotete, fragmentert eller kompleks til å analysere effektivt. For selvkjørende biler, droner og humanoide roboter er dette helt avgjørende. De må kunne tolke verden gjennom kameraer, radar, lidar, mikrofoner og andre sensorer. Da holder det ikke med pene regneark. Modellene må forstå ustrukturert og multimodal data fra den fysiske verden.

Mellom strukturert og ustrukturert data finner vi semi-strukturert data. Dette er data som ikke passer helt inn i tradisjonelle tabeller, men som likevel har en viss organisering. Eksempler er JSON-filer, XML-filer, loggfiler og mange typer IoT-data. Semi-strukturert data har blitt stadig viktigere i moderne programvare, fordi den kombinerer fleksibiliteten fra ustrukturert data med noe av søkbarheten fra strukturert data.

Data kan også kategoriseres etter kilde. Primærdata er data en aktør samler inn selv. Det kan være kameradata fra en bil, sensordata fra en drone, bruksdata fra en digital plattform eller produksjonsdata fra en fabrikk. Fordelen med primærdata er at den

ofte er unik, relevant og direkte knyttet til aktørens egne produkter eller tjenester. Den kan dermed bli en strategisk ressurs. Ulempen er at den krever investeringer i infrastruktur, sensorer, programvare og systemer for innsamling og lagring.

Sekundærdata er data som kommer fra eksterne kilder. Det kan være offentlige datasett, kjøpte datasett, bransjerapporter eller aggregerte datakilder. Slike data er ofte billigere og raskere å få tilgang til, men de er sjelden like unike. De kan være nyttige som supplement, men gir normalt svakere konkurransefortrinn enn data man har samlet inn selv.

Til slutt må data vurderes etter sensitivitet og personvern. Noen data er *ikke-sensitive*, som værdata, trafikkmønstre eller anonymiserte statistikker. Andre data er *sensitive*, som helseopplysninger, finansiell informasjon, persondata eller forretningskritiske dokumenter. Jo mer sensitiv dataen er, desto større krav stilles det til sikkerhet, tilgangskontroll og juridisk etterlevelse. Mange virksomheter klassifiserer derfor data i nivåer som public, internal, restricted og confidential.

Dette avgjør hvem som kan bruke dataene, hvordan de kan brukes, hvor verdifulle de er, og om de kan skape varige konkurransefortrinn. I AI gjelder fortsatt det gamle prinsippet «garbage in, garbage out» (GIGO). Dårlige, skjeve eller irrelevante data gir svake modeller. Gode, relevante og proprietære data kan derimot bli selve drivstoffet som gjør en AI-modell bedre, mer presis og vanskeligere å kopiere.

Som Geoffrey Moore observerer i konteksten av Big Data:

"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway."

Dette understreker hvor viktig det er å forstå data langs flere dimensjoner, blant annet hvordan de er strukturert, hvor de kommer fra, og hvor sensitive de er. I praksis

handler det om forskjellen mellom strukturert og ustrukturert data, primærdata og sekundærdata, samt ikke-sensitive og sensitive data. For AI-systemer som selvkjørende biler, droner og roboter er dette avgjørende. Videodata er for eksempel både ustrukturert og multimodal, fordi den kombinerer bilde, bevegelse, tid og kontekst. Slike data må ha høy kvalitet og troverdighet for at modellen skal lære riktig. Hvis inputen er dårlig, skjev eller mangelfull, blir også outputen upålitelig (GIGO).

I AI-sammenheng er disse kategoriene helt sentrale fordi nevrale nettverk lærer av data. Man kan si at de «spiser» data for å forstå mønstre, sammenhenger og avvik. Offentlige og ustrukturerte data kan gi modellene bred kunnskap om verden, men private og mer spesifikke data er ofte det som gir presisjon i konkrete bruksområder. Det kan være diagnostikk i helse, svindeldeteksjon i finans, prediktivt vedlikehold i industri eller bevegelsesplanlegging i roboter. Jo bedre dataene speiler problemet modellen skal løse, desto mer nyttig blir AI-systemet.

Her spiller også dataklassifisering en viktig rolle. Ved hjelp av maskinlæring kan rådata tagges, sorteres og organiseres automatisk basert på innhold, kontekst og sensitivitet. Det gjør det lettere å skille nyttige signaler fra støy, beskytte sensitiv informasjon og bygge mer pålitelige datasett for trening og analyse. Rådata i seg selv er sjelden helt nok. Verdien oppstår først når dataene blir strukturert, forstått og gjort anvendbare.

4.0 Big Data sine fire V'er

Som en naturlig videreføring av data-kategoriseringen kan Big Data forstås gjennom fire sentrale dimensjoner. På engelsk omtales de ofte som Volume, Velocity, Variety og Veracity. På norsk kan vi oversette dem til volum, hastighet, mangfold og pålitelighet.

Dette rammeverket gir en nyttig måte å vurdere data på, ikke bare etter hvor mye data man har, men også hvor raskt dataene oppstår, hvor ulike de er, og hvor mye man kan stole på dem. I en AI-kontekst er dette helt avgjørende. Nevrale nettverk og språkmodeller blir ikke bedre bare fordi de mates med mer data. De blir bedre når dataene er relevante, varierte, oppdaterte og pålitelige.

"Big data is at the foundation of all the megatrends that are happening"

- Chris Lynch

1. **Volum**

   Handler om de store datamengdene som skapes, lagres og behandles hver dag. Viktig fordi nevrale nettverk ofte trenger store datasett for å lære mønstre og bli mer presise.

   Et eksempel er helsesektoren, der genomiske datasett kan brukes til å identifisere genetiske markører for sykdom og utvikle mer tilpasset behandling. Et annet eksempel er selvkjøring, som samler inn store mengder sensor- og kameradata for å forbedre objektgjenkjenning og kjøreadferd.

   Utfordringen er at høyt datavolum krever lagring, regnekraft og god struktur. Uten dette kan dataene skape mer støy enn verdi.

2. **Velocity (Hastighet)**

   Handler om hvor raskt data skapes, samles inn og analyseres. I mange AI-systemer må dette skje i sanntid eller nesten sanntid.

   Dette er avgjørende i bruksområder som krever umiddelbar respons. I helsevesenet kan AI analysere pasientdata fortløpende for å oppdage tegn på

akutte situasjoner. For droner og selvkjørende biler må sensordata behandles på millisekunder for å unngå kollisjoner. Digitale agenter bruker også kontinuerlige datastrømmer fra brukere for å tilpasse svar og handlinger med én gang.

Høy hastighet krever ofte teknologi som edge computing, der data behandles nærmere kilden (i sluttproduktene). Det reduserer latens (forsinkelsen), men gjør også datahåndteringen mer kompleks.

3. **Variety (Mangfold)**
   Handler om at data kommer i mange ulike former, som strukturert data i tabeller, semi-strukturert data som JSON-filer, og ustrukturert data som tekst, bilder, lyd og video.

   I helse kan for eksempel elektroniske journaler, medisinske bilder og data fra smartklokker kombineres for å gi mer helhetlige diagnoser. For droner og roboter kan videodata, sensordata og tekstbaserte instruksjoner kobles sammen for å forstå situasjoner bedre.

   Et godt eksempel på dette er forskjellen mellom kamera og lidar i selvkjørende biler, droner og roboter. Kameraer samler inn visuelle data, omtrent slik mennesker ser verden, med farger, skilt, veimerking, objekter og bevegelser. Dette gir rike data for AI-modeller som skal forstå kontekst, men kameraer kan være sårbare for dårlig lys, regn, snø og blending. Lidar fungerer annerledes. Den sender ut laserpulser og måler avstanden til objekter, og lager dermed et presist tredimensjonalt kart over omgivelsene. Lidar gir svært god dybdeforståelse, men mindre semantisk informasjon enn kamera.

Dette er kjernen i multimodal AI, hvor modeller lærer fra flere typer data samtidig. Utfordringen er at ulike dataformater må tolkes, renses og kobles sammen på en måte som faktisk gir mening.

4. **Veracity (Pålitelighet)**

   Hvor nøyaktige, konsistente og troverdige dataene er. I AI er dette avgjørende, fordi modeller lærer av dataene de får. Hvis dataene er feil, skjeve eller fulle av støy, vil modellen også kunne gi svake eller misvisende resultater.

   Et eksempel er AI i helse, der medisinske data må valideres nøye før de brukes til diagnostikk. Et annet er roboter og cobots, der sensordata må renses og kontrolleres for å unngå feil handlinger i den fysiske verden.

   Dette blir enda viktigere med syntetiske data. Slike data kan være nyttige, men hvis de ikke speiler virkeligheten godt nok, kan de svekke modellen over tid. Derfor krever høy pålitelighet god datastyring, kvalitetssikring og kontinuerlig verifisering.

Til slutt utvides rammeverket ofte med en femte V, nemlig verdi. Dette handler om å gjøre data om til innsikt, beslutninger og konkrete forbedringer.

I AI ser vi dette tydelig i personaliserte anbefalinger, som hos Netflix og Amazon. Ved å analysere brukerdata kan de foreslå filmer, serier eller produkter som oppleves mer relevante for hver enkelt bruker. Det samme gjelder i finans, helse og industri, der gode data kan gi bedre risikomodeller, mer presise diagnoser eller mer effektiv drift.

Uten verdi er de andre V'ene lite nyttige. Store, raske, varierte og pålitelige data betyr lite hvis de ikke kan brukes til noe. Det er først når data forbedrer produkter, beslutninger eller forretningsmodeller at AI skaper reell verdi.

Med de 4 V-er (+ Verdi) som grunnmur, beveger vi oss mot ulikhetene i treningsgrunnlag og modellkonvergering, der teoretiske rammeverk som *Disruptive Innovation* hjelper oss å forstå dataens transformative potensial.

Eksempler på syntetiske data

Syntetiske data er kunstig genererte data som etterligner ekte data. De brukes ofte i AI når ekte data er vanskelig å få tak i, for sensitive til å deles, eller for skjeve til å gi gode modeller. Slike data kan lages gjennom statistiske modeller, simuleringer eller generative modeller.

I helsevesenet kan syntetiske pasientjournaler brukes til forskning og AI-trening uten å eksponere ekte pasientdata. De kan simulere symptomer, behandlinger og utfall, og dermed gjøre det mulig å utvikle bedre diagnostiske modeller innenfor rammene av personvernregler som GDPR og HIPAA. Syntetiske helsedata kan også brukes til å fylle hull i datasett, for eksempel for svært sjeldne sykdommer eller grupper som er dårlig representert i eksisterende data.

I finans brukes syntetiske data til å simulere transaksjoner, kundeadferd og markedsforhold. Dette gjør det mulig å teste nye løsninger, trene modeller for svindeldeteksjon og analysere risiko uten å bruke sensitive kundedata. Banker kan dermed bygge og teste AI-systemer i trygge miljøer før de tas i bruk i virkeligheten.

For selvkjørende biler, droner og roboter er syntetiske data særlig viktige. I stedet for å vente på at sjeldne situasjoner oppstår i virkeligheten, kan man simulere dem. Det kan være farlige trafikkhendelser, dårlig vær, mørke, uventede hindringer eller komplekse bymiljøer. Slik kan AI-systemer trenes på langt flere scenarioer enn det som ville vært praktisk eller trygt å samle inn fysisk.

Også innen media og bildebehandling brukes syntetiske data. Kunstig genererte bilder og videoer kan hjelpe datavisjonsmodeller med å kjenne igjen objekter, mennesker, bakgrunner eller bevegelser. På samme måte kan syntetisk tekst og lyd

brukes til å trene språkmodeller, taleassistenter og systemer for oversettelse eller kundeservice.

Poenget er at syntetiske data kan gjøre AI-utvikling raskere, billigere og mer personvernvennlig. Samtidig krever det høy kvalitet. Hvis de syntetiske dataene ikke ligner nok på virkeligheten, kan modellen lære feil mønstre. Da går syntetiske data fra å være en styrke til å bli en kilde til støy. Derfor blir balansen mellom skala, realisme og pålitelighet helt avgjørende.

5.0 Ulikheter i treningsgrunnlag og modellkonvergering

Etter å ha sett på ulike datakategorier og Big Data sine fire V'er, blir neste spørsmål hva slags data AI-modeller faktisk trenes på. Det er her skillet mellom offentlige og private data blir avgjørende. Offentlige data kan gi brede og nyttige modeller, men fordi mange aktører bruker de samme datakildene, kan modellene også begynne å ligne mer på hverandre. Private data fungerer annerledes. De er vanskeligere å få tak i, mer domenespesifikke og ofte langt mer verdifulle. Derfor kan de bli en strategisk moat.

Treningsgrunnlaget er datasettet som brukes til å utvikle og forbedre AI-modeller. Store språkmodeller som ChatGPT, Gemini og Llama er i stor grad bygget på enorme mengder offentlig tilgjengelig data fra internett, som nettsider, Wikipedia, bøker, kode, forum og åpne datasett. Dette gir modellene en bred forståelse av språk og verden, men det gjør dem også til generalister. De kan skrive, oppsummere, resonnere og svare på mange typer spørsmål, men de mangler ofte dyp innsikt i

interne prosesser, spesifikke kunder, medisinske pasienthistorikker, industrielle sensordata eller en bedrifts unike beslutningsgrunnlag.

Når flere modeller trenes på mye av det samme åpne internettmaterialet, oppstår det en form for modellkonvergering. Modellene lærer mange av de samme mønstrene, får lignende styrker og svakheter, og produserer etter hvert resultater som kan oppleves ganske like. Over tid kan dette gjøre generelle AI-modeller mer som råvarer eller infrastruktur. De blir nyttige, kraftige og billige, men vanskeligere å differensiere. Litt som skylagring eller betalingsinfrastruktur. Viktig for alle, men ikke nødvendigvis unikt i seg selv.

Det betyr ikke at offentlige data er uviktige. Tvert imot har de vært selve grunnlaget for den første bølgen av moderne AI. Men offentlige data gir først og fremst bredde. Private data gir dybde. I helse kan det være pasientjournaler, bildediagnostikk og behandlingshistorikk. I finans kan det være transaksjoner, kundeadferd og risikodata. I industri kan det være produksjonsdata, vedlikeholdslogger og sensordata fra maskiner. Slike data er ofte mer presise, mer relevante og tettere knyttet til konkrete beslutninger.

Dette er grunnen til at private data kan skape divergens mellom AI-systemer. To selskaper kan bruke den samme grunnmodellen, men få helt ulike resultater dersom den ene har tilgang til bedre, mer relevant og mer strukturert data. I praksis flyttes konkurransefortrinnet da fra selve modellen til dataene rundt modellen. Modellen blir motoren, men dataene bestemmer hvor godt den kjører i et konkret terreng.

Her blir RAG-teknologi viktig. *Retrieval-Augmented Generation* gjør det mulig å koble en AI-modell til private dokumenter, databaser eller kunnskapskilder uten at modellen nødvendigvis må trenes direkte på dem. En bank kan bruke AI til å analysere interne lånedata, compliance-dokumenter og kundehistorikk. Et sykehus kan bruke AI til å hente relevant informasjon fra pasientjournaler og medisinske retningslinjer. En industribedrift kan la en modell hente innsikt fra vedlikeholdslogger,

sensordata og manualer. På den måten kan generelle modeller bli domenespesifikke uten at all privat informasjon må eksponeres eller bygges inn i selve modellen.

Syntetiske data forsterker dette bildet ytterligere. De kan brukes til å fylle hull der ekte data er vanskelig, dyrt eller risikabelt å samle inn. For selvkjørende biler og droner kan man simulere sjeldne hendelser som ulykker, ekstremvær eller uventede hindringer. For roboter kan man trene i virtuelle miljøer før systemene testes fysisk. Syntetiske data kan dermed fungere som en bro mellom offentlige og private data, men bare dersom kvaliteten er høy. Hvis de syntetiske dataene ikke speiler virkeligheten godt nok, kan modellen lære feil mønstre.

Disrupsjon handler ikke nødvendigvis om den mest avanserte teknologien fra start. Ofte begynner den i nisjer, med løsninger som er enklere, billigere eller mer tilpasset et konkret behov (Christensens Disruptive Innovasjon). Offentlige data og generelle AI-modeller driver mye av den inkrementelle forbedringen hos de store teknologiselskapene. Private data, RAG og domenespesifikke AI-systemer kan derimot åpne for mer disruptive løsninger i bransjer der presisjon, sikkerhet og kontekst er viktigere enn generell språkforståelse.

For AI kan dette bety at verdien gradvis flyttes fra å ha den største grunnmodellen til å ha det beste datagrunnlaget, den beste datastyringen og den sterkeste koblingen mellom modell og virkelig bruk. For digitale agenter, autonome systemer, roboter og helseteknologi blir dette avgjørende. Den som eier de mest relevante dataene, kan bygge systemer som lærer raskere, handler mer presist og blir vanskeligere å kopiere.

Dermed blir modellkonvergering og datadivergens to sider av samme utvikling. På overflaten blir grunnmodellene mer like. Under overflaten blir forskjellen mellom aktørene større, fordi private, proprietære og høykvalitetsdata avgjør hvem som faktisk klarer å skape verdi. Det er her data går fra å være råmateriale til å bli en strategisk ressurs.

6.0 Data Verdi Pyramiden (DVP)

Som et resultat av våre diskusjoner, introduserer vi *data verdi pyramiden*. Et hierarkisk rammeverk som kategoriserer data fra base (kilde, struktur) til topp (bruk). I basen finner vi selskaper som Oracle (ORCL) for strukturert privat data, mens midt-nivået inkluderer Nvidia (NVDA) for syntetisk multimodal data innen robotikk. På toppen vurderes verdien. Tesla (TSLA) gir moat via primær video-data for humanoids. RBV støtter dette, som Barney noterer:

"To attract the kinds of resources that can generate profits, managers must recognize that stakeholders… have claims on the profits."

Fremtiden peker mot AI factories der syntetisk data dominerer, som Ellison forutsier:

"*The future lies in leveraging private enterprise data."*

Om dette er riktig vet vi ikke, men det er kraftfullt perspektiv. Og der er jo det vi leter etter i våre perspektivnotater.

Nederst i pyramiden finner vi de grunnleggende kategoriene. Her vurderes data etter kilde, struktur og sensitivitet. Er dataene primære eller sekundære? Er de strukturert, semi-strukturert eller ustrukturert? Er de offentlige, interne eller sensitive? Et selskap som Oracle er relevant i denne delen av pyramiden, fordi mye av verdien deres ligger i lagring, organisering og bruk av strukturerte, private bedriftsdata. Dette er ikke den mest spektakulære delen av AI, men det er ofte fundamentet alt annet bygges på.

Midt i pyramiden finner vi den AI-spesifikke utvidelsen. Her blir spørsmålene mer avanserte. Er dataene ekte eller syntetiske? Er de tekstbaserte, visuelle, sensorbaserte eller multimodale? Hvor høy er kvaliteten, og hvor godt speiler dataene

virkeligheten? Dette nivået blir særlig viktig for robotikk, droner, selvkjørende biler og humanoids. Nvidia er et godt eksempel på en aktør som passer inn her, fordi selskapet ikke bare leverer regnekraft, men også bygger økosystemer for simulering, syntetiske data og fysisk AI. Når roboter kan trenes i virtuelle miljøer før de testes i virkeligheten, endres hele læringskurven.

Øverst i pyramiden handler det ikke lenger bare om hva slags data man har, men hva dataene faktisk kan brukes til. Skaper de bedre modeller? Lavere kostnader? Raskere produktutvikling? Sterkere kundeinnsikt? Eller en moat konkurrentene ikke lett kan kopiere? Tesla er et godt eksempel på denne delen av pyramiden. Dersom selskapets enorme mengder video- og sensordata fra biler kan brukes til å trene selvkjøring, humanoids og fysisk AI, kan dataene bli en langsiktig strategisk ressurs. Verdien ligger ikke bare i bilene, men i læringssløyfen mellom produkter i bruk, nye data, bedre modeller og bedre produkter.

Dette kan også forstås gjennom Resource-Based View. I RBV er de mest verdifulle ressursene de som er sjeldne, vanskelige å imitere og godt organisert i selskapet. Proprietære data med høy kvalitet kan oppfylle nettopp disse kriteriene. Offentlige data kan alle bruke. Private, ferske og domenespesifikke data er langt vanskeligere å kopiere.

I dag brukes data primært til å trene, finjustere og forbedre AI-modeller. Selvkjørende biler bruker kamera-, radar- og lidar-data til objektgjenkjenning og bevegelsesplanlegging. Roboter bruker bevegelsesdata og sensordata til å lære interaksjon med den fysiske verden. Digitale agenter bruker tekst, dokumenter og brukerhistorikk til å forstå kontekst og handle mer presist. Verdien ligger i bedre beslutninger, mer automatisering og høyere operasjonell effektivitet.

Fremover kan verdien bli enda større. Mot 2030 peker utviklingen mot AI factories, der data ikke bare brukes til å trene modeller én gang, men inngår i kontinuerlige læringssløyfer. Ferske data fra virkelige interaksjoner kan bli viktigere enn gamle datasett. Syntetiske data kan fylle hull i sjeldne scenarioer, som ulykker, ekstremvær

eller farlige robotsituasjoner. Edge AI kan gjøre det mulig å behandle data nærmere kilden, med lavere forsinkelse og bedre personvern.

Samtidig er dette ikke uten risiko. Hvis syntetiske data brukes ukritisk, kan modellene lære feil mønstre. Hvis private data håndteres dårlig, kan det skape personvernproblemer og regulatorisk risiko. Hvis dataene er ufullstendige eller gamle, kan de gi modeller som virker intelligente, men tar dårlige beslutninger.

DVP viser derfor at det ikke er nok å ha mye data. Man må ha riktig data, pålitelig data, fersk data og evnen til å omdanne data til bedre AI-systemer. Fremtidens vinnere blir ikke nødvendigvis de som har de største modellene, men de som har de beste læringssløyfene mellom data, modeller og virkelige produkter. Det er her data går fra å være digitalt råstoff til å bli en strategisk ressurs.

Postcards From The Future

Data: Real vs. Syntetisk Data Matrix (GIGO og applikasjoner)

	Real Data (Autentisk, høy veracity)	**Syntetisk Data** (Generert, fleksibel)
Høy Verdi (Lang Sikt, Strategisk)	**Proprietær video fra humanoids: Bygger moat, men privacy-risiko.**	**GAN-generert data F.eks for droner, fyller gap, disrupter markedet (80% av AI-data i 2028?)**
Lav Verdi (Kort Sikt, Operasjonell)	**Offentlig tekstdata Commodity, konvergering.**	**Lav-kvalitets syntetisk Risiko for model collapse i agenter.**

ent

Postcards From The Future

Data: Modalitet vs. Tidshorisont Matrix (utvidelse med video)

	Unimodal (Tekst & Tall)	**Multimodal** (Video + Andre)
Nåtid	**Grunnleggende trening for digitale agenter: Effektiv, men begrenset**	**Video for selvkjørende biler: Rik, men prosesseringskrevende. Edge AI prosessering er enabler**
Fremtid Lang sikt	**Commodity i AI factories.**	**Dominerende for humanoids & cobots: Real-time video-integrasjon via edge AI, høy markedsvekst**

ent

Dette rammeverket gir en analytisk linse for å maksimere dataens verdi, mens det adresserer risikoer. For implementering, prioriter syntetisk data for testing og video for visjon-baserte systemer. Fremtiden tilhører de som dyrker frem «god» data.

6.1 Relevante selskaper og aksjer

Med DVP som rammeverk kan vi plassere ulike selskaper og aksjer etter hvilken rolle de spiller i AI-økosystemet. Listen er ikke ment å være komplett, men gir eksempler på aktører som er relevante for datainnsamling, datalagring, datakvalitet, AI-trening og praktiske anvendelser innen selvkjørende biler, droner, humanoids, cobots og digitale agenter.

Poenget er ikke bare hvem som «driver med AI», men hvor i verdikjeden de skaper verdi. Noen selskaper kontrollerer grunnleggende datainfrastruktur. Andre samler inn proprietære data fra produkter i bruk. Noen gjør dataene anvendbare gjennom lagring, strukturering og governance. Andre bruker dataene til å trene modeller,

simulere scenarioer eller bygge nye AI-produkter. DVP-rammeverket hjelper oss å skille mellom disse rollene.

Basen

Nederst i pyramiden finner vi selskaper som håndterer data på grunnivå. Dette handler om kilde, struktur, sensitivitet og behandling. Her ligger mye av fundamentet for Big Data sine V-er, særlig volum, hastighet, mangfold og pålitelighet.

Tesla (TSLA) er et eksempel på en aktør med store mengder primærdata. Selskapet samler inn video- og sensordata fra biler i bruk, noe som kan brukes til trening av selvkjøring og på sikt fysisk AI som Optimus. Verdien ligger i at dataene er proprietære, nye og hentet fra virkelige situasjoner.

Figure AI (ikke notert) er et humanoid-selskap og er avhengige av store mengder primærdata fra fysisk interaksjon med verden. Når robotene beveger seg, griper objekter, samarbeider med mennesker og utfører oppgaver i fabrikker eller lagerbygg, samles det inn verdifulle sensor, video og bevegelsesdata. Slike data kan brukes til å forbedre robotenes motorikk, beslutninger og evne til å håndtere nye situasjoner. Verdien ligger i at dataene er proprietære, praktiske og hentet fra ekte arbeidsmiljøer, noe som kan gi Figure et viktig treningsgrunnlag for fysisk AI over tid.

Alphabet og Waymo (GOOGL) har også verdifulle primærdata fra selvkjørende biler. Waymos flåter samler inn data fra kameraer, lidar og sensorer i komplekse trafikkmiljøer. Dette gir et sterkt treningsgrunnlag for autonome systemer.

Snowflake (SNOW) og Amazon (AMZN) er mer relevante på sekundær- og aggregeringssiden. Snowflake hjelper virksomheter med å samle, organisere og bruke data på tvers av skymiljøer. Amazon bruker enorme mengder e-handels-, logistikk- og kundeinteraksjonsdata, samtidig som AWS er en sentral plattform for selskaper som bygger AI-løsninger.

Oracle (ORCL) sitter tett på strukturerte enterprise-data gjennom databaser og forretningskritiske systemer. Dette er kanskje mindre synlig enn de store AI-modellene, men svært viktig for RAG-baserte løsninger, der private bedriftsdata kobles til AI-modeller.

Salesforce (CRM) passer også inn her, fordi selskapet sitter på strukturerte CRM-data knyttet til kunder, salg og arbeidsflyt. Slike data kan bli verdifulle når de brukes til agentic AI i salg, service og kundedialog.

For ustrukturert data er Meta (META) et godt eksempel. Selskapet har enorme mengder tekst, bilder, video og sosial interaksjon, noe som er relevant for multimodale modeller. Seagate (STX) kan plasseres i samme lag fra en annen vinkel, som leverandør av lagringskapasitet for de enorme datamengdene AI krever.

Innen sensitiv data og datasikkerhet er SentinelOne (S) og Rubrik (RBRK) relevante eksempler. SentinelOne bruker AI på cybersikkerhetsdata for å oppdage trusler, mens Rubrik er eksponert mot backup, databeskyttelse og governance. I en verden der AI kobles tettere på private data, blir denne typen sikkerhet og kontroll stadig viktigere.

Palantir (PLTR) passer særlig godt inn i kategorien databehandling og justering. Selskapet arbeider med å samle, rense, strukturere og gjøre komplekse datasett anvendbare, spesielt i forsvar, myndigheter, industri og helse. Dette er et godt eksempel på at verdien ofte ligger i å gjøre rådata beslutningsklare.

Midten

Midt i pyramiden finner vi selskaper som er mer direkte knyttet til AI-trening, syntetiske data, multimodalitet og fysisk AI. Her handler det ikke bare om å lagre eller strukturere data, men om å gjøre dem nyttige for nevrale nettverk.

Nvidia (NVDA) er den tydeligste aktøren på dette nivået. Selskapet leverer ikke bare GPU'er til datasentre, men bygger også økosystemer for simulering, robotikk og

syntetiske data. Gjennom plattformer som *Omniverse*, *Isaac* og *Cosmos* kan Nvidia bidra til å skape virtuelle treningsmiljøer for roboter, droner og autonome systemer. Dette gjør selskapet sentralt i overgangen fra ren databehandling til fysisk AI.

Ambarella (AMBA) er relevant innen multimodal prosessering, særlig for kamera- og videobaserte AI-systemer. Selskapets halvledere brukes i løsninger der visuell informasjon må behandles effektivt, som i droner, sikkerhetskameraer, roboter og autonome kjøretøy.

Ouster (OUST) er eksempel på selskap som leverer sensorteknologi og fysisk data til autonome systemer. Deres lidar og sensordata kan brukes til kartlegging, objektgjenkjenning og navigasjon. Dette er særlig viktig i fysiske AI-applikasjoner der modellen må forstå verden i tre dimensjoner.

Apple (AAPL) passer inn i multimodalitet gjennom kombinasjonen av tekst, tale, bilde, video, sensorer og brukerdata i sitt økosystem. Selskapet har en unik posisjon fordi det sitter tett på sluttbrukeren og kan integrere AI direkte i enheter.

ServiceNow (NOW) representerer en mer enterprise-orientert form for AI-data. Selskapet bruker arbeidsflytdata og prosessdata til å bygge digitale agenter som kan automatisere oppgaver internt i virksomheter. Her er verdien ikke nødvendigvis multimodal video, men kontekstuell forståelse av hvordan organisasjoner faktisk jobber.

Symbotic (SYM) kombinerer sensordata, bevegelsesdata og operasjonelle data for å optimalisere logistikk og robotstyring.

Toppen

Øverst i pyramiden vurderes hvilke selskaper som kan omdanne data til varige konkurransefortrinn. Her blir spørsmålet om dataene er sjeldne, vanskelige å kopiere og tett integrert i selskapets produkter og forretningsmodell.

Tesla (TSLA) er et av de tydeligste eksemplene på en mulig langsiktig data-moat. Dersom selskapet klarer å bruke video- og sensordata fra bilflåten til å forbedre selvkjøring og fysisk AI, kan det skape en sterk *flywheel*. Flere produkter i bruk gir mer data, mer data gir bedre modeller, bedre modeller gir bedre produkter, og bedre produkter kan igjen gi flere brukere.

Nvidia (NVDA) har en annen type *moat*. Her ligger verdien i infrastrukturen rundt AI-trening, simulering og akselerert databehandling. Dersom fremtidens AI i økende grad trenes i virtuelle miljøer og AI-fabrikker, kan Nvidia bli en sentral plattform for både ekte og syntetiske data.

TSMC (TSM) og Arm (ARM) ligger lenger ned i den fysiske verdikjeden, men er likevel strategisk viktige. TSMC produserer mange av de mest avanserte brikkene som AI-økosystemet er avhengig av. Arm leverer arkitektur som er viktig for energieffektiv databehandling, særlig på edge-enheter, mobile enheter og etter hvert robotikk.

Microsoft (MSFT) og Amazon (AMZN) har sterke posisjoner gjennom sky, enterprise-kunder og AI-plattformer. De kontrollerer ikke bare infrastruktur, men også distribusjonskanaler inn i bedrifter. Det gir dem en viktig rolle i hvordan private data kobles til modeller og agenter.

Palantir (PLTR) kan også plasseres høyt i pyramiden fordi selskapet fokuserer på å gjøre komplekse, sensitive og ofte fragmenterte data operasjonelle. I bransjer som forsvar, helse og industri kan dette gi betydelig verdi dersom AI-systemene blir en del av kritiske beslutningsprosesser.

Implikasjoner

Rammeverket viser at AI-verdi ikke bare skapes i selve modellen. Den skapes i hele kjeden rundt modellen. Fra datainnsamling og lagring til strukturering, kvalitetssikring, simulering, trening og praktisk anvendelse.

Tesla illustrerer verdien av primære og proprietære real-world data. Nvidia illustrerer verdien av syntetiske data, simulering og AI-infrastruktur. Oracle, Snowflake og Salesforce illustrerer verdien av strukturerte enterprise-data. Palantir og Rubrik viser betydningen av datagovernance, sikkerhet og anvendbarhet. Ambarella og Ouster viser hvor viktig sensor- og videodata blir for fysisk AI.

Det viktigste poenget er at fremtidens AI-vinnere ikke nødvendigvis bare blir de som har de største modellene. Det kan like gjerne bli de som eier de beste dataene, har de sterkeste læringssløyfene og klarer å gjøre data om til konkrete produkter, bedre beslutninger og varige konkurransefortrinn.

7.0 Fremtiden

Dette notatet har navigert gjennom dataens evolusjon i AI, fra grunnleggende kategorier til strategiske rammeverk, og understreket at god data er essensiell for robuste modeller i applikasjoner som droner og cobots. Som Christensen oppsummerer:

> *«First, disruptive products are simpler and cheaper. They generally promise lower margins, not greater profits»*

Vi ser mot en fremtid der syntetisk og privat data disrupterer tradisjonelle paradigmer. For å visualisere dette, presenterer vi en avsluttende 2x2-matrise som integrerer *Disruptive Innovation Theory* med data-typer, og viser potensialet for innovasjon mens vi advarer mot risikoer som bias og commoditization.

Postcards From The Future

Data: Data Type vs. Disruptivt Potensial

	Real Data (Autentisk, Høy Veracity)	**Syntetisk Data** (Generert, Fleksibel)
Lavt Disruptivt Potensial (Sustaining Innovation)	**Offentlig real data: Forbedrer eksisterende modeller, men leder til konvergering (f.eks. generelle AI-agenter).**	**Lav-kvalitets syntetisk:** **Fyller midlertidige gap, men risikerer model collapse.**
Høyt Disruptivt Potensial (New-Market Disruption) .	**Privat real data** **Skaper moats i nisjer (f.eks. video for humanoids, biler, ai briller og droner).**	**Disrupter ved å simulere sjeldne scenarier (f.eks. drone-trening), potensielt 80% av fremtidig AI-data.**

ent

For å konkludere, så ligger dataens verdi i dens evne til å utvikle etisk og effektiv AI. Ved å investere i kvalitetsdata og teoretiske rammeverk, kan vi navigere mot en fremtid der AI ikke bare er smart, men også bærekraftig.

Takk for tiden.

Vi sees i fremtiden 😊