Effektiv testing med rike anonymiserte testdata

20. september 2016

Helene Aune

Skatteetaten

Erik Rogstad

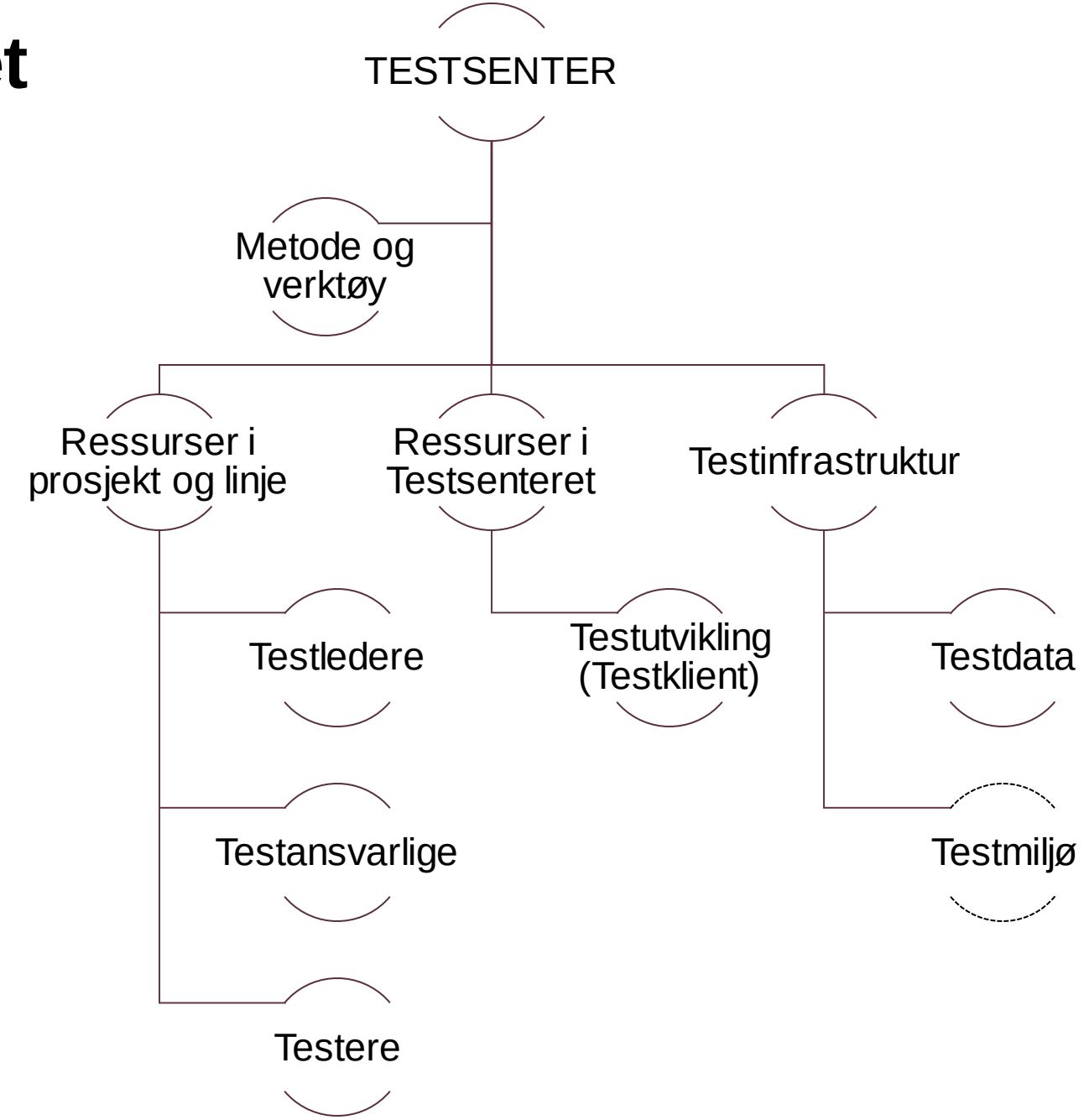
Skatteetaten

23. februar 2020

Skatteetatens IT- og Servicepartner

- Skatteetatens leverandør av IT- og administrative tjenester
- Utvikler, drifter og forvalter Skatteetatens IT-systemer
- Systemutvikling – Prosjektledelse – Infrastruktur – Sikkerhet
- Ca. 900 ansatte fordelt på kontorer i Oslo, Grimstad og Lillehammer

Testsenteret



Innhold

- Anonymiserte data
  - Hvorfor?
  - Til hva?
  - Eksempler på bruk
- Hvordan vi har anonymisert våre data
  - Anonymiseringsnivå
  - Sentrale elementer
  - Erfaringer

Begrepsavklaring

- Anonymiserte testdata

  Produksjonsdata som er anonymisert for å brukes til test

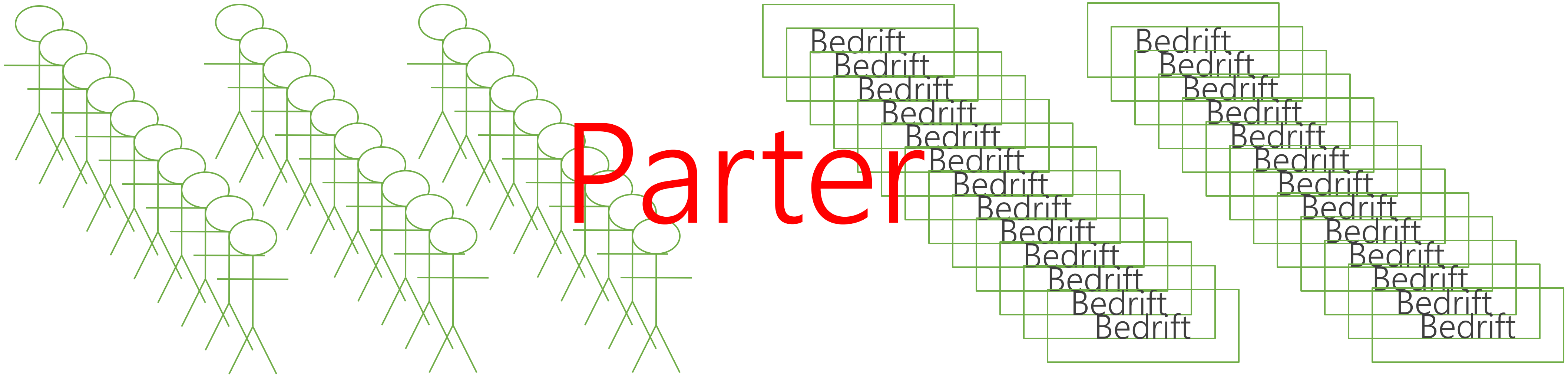
- Syntetiske testdata

  Data som er konstruert uten rot i virkelige data

Domene

Skattedata

- Innskudd, lån og renter
- BSU
- Skattemanntall
- Skatteoppgjør
- Aksjer og fond
- Barnepass
- Selvangivelse
- Skattekort

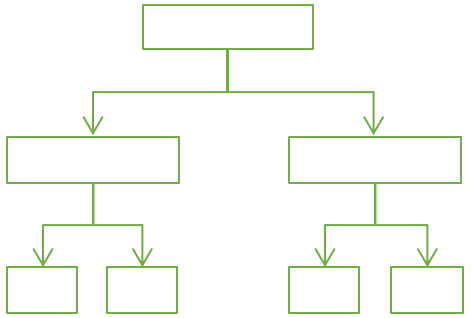


Hva er målet?

- Målet er å teste med mest mulig produksjonslike data for å avdekke realistiske feil
  - Funksjonelle feil - eksempelvis sære men realistiske funksjonelle feil som man kun finner i testdatasett med god spredning
  - Ikke-funksjonelle feil relatert til ytelse og robusthet som best avdekkes ved produksjonslike data og volum
- Komplementært til andre former for testing med syntetiske data

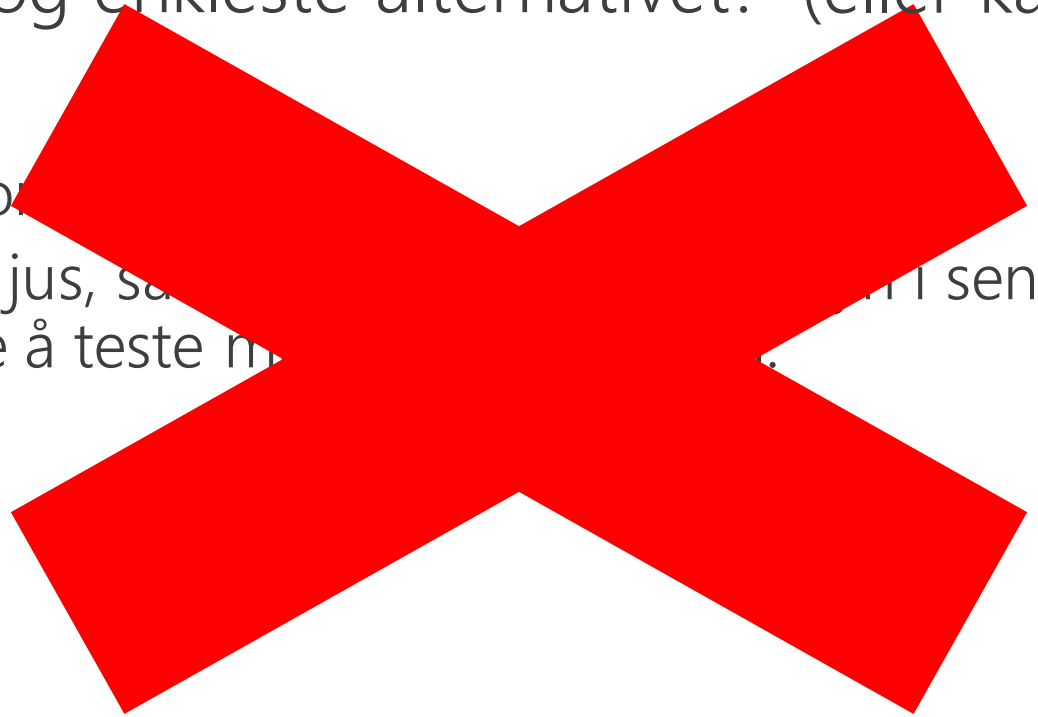
Alternativ 1: Syntetiske testdata

Modell av inputdomenet



Alternativ 2: Skarpe produksjonsdata

- Det billigste og enkleste alternativet? (eller kanskje ikke…?)
- Men NEI
  - Juridisk utfo
  - Og uansett jus, sa sensitive data. De ønsker bare å teste m



Alternativ 3: Anonymiserte produksjonsdata

- Potensielt svært produksjonslike
- Representerer variasjon og særtilfeller fra produksjon
- Er anonyme, men bør behandles med noe mer forsiktighet enn helt syntetiske testdata

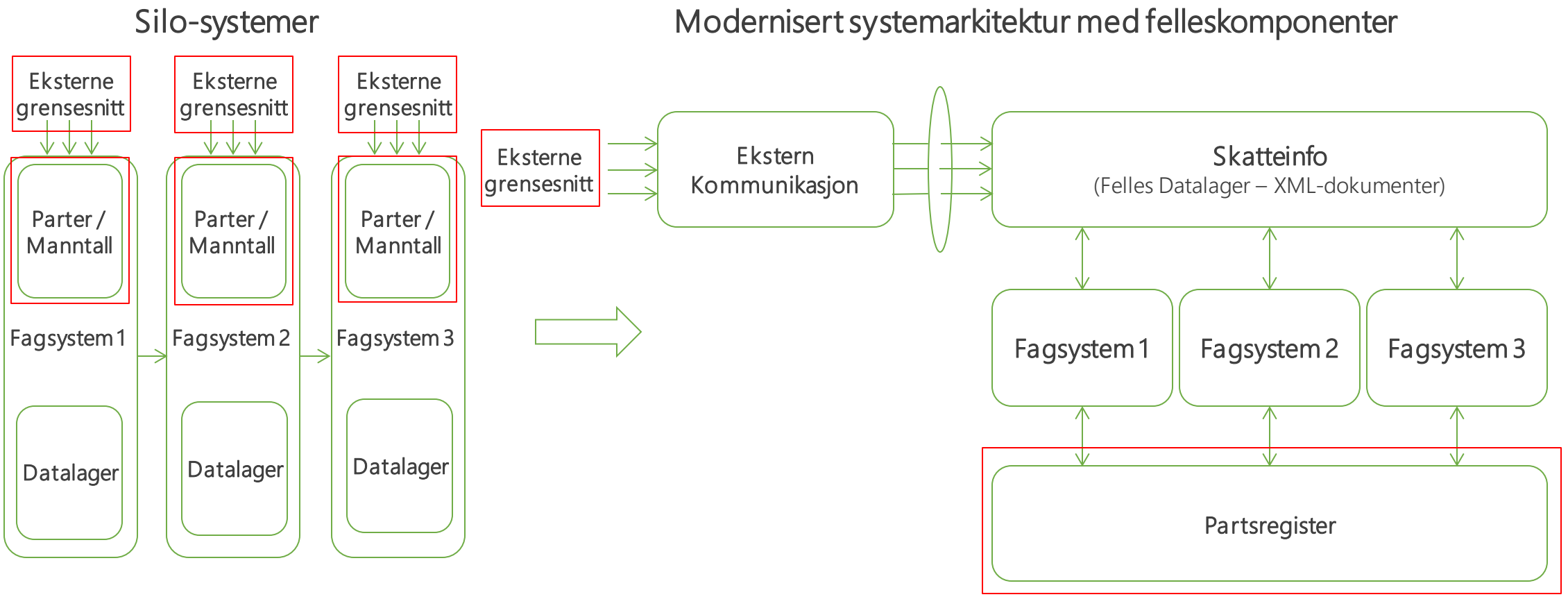
Hva får man med anonymiserte data?

- Rikt soneuavhengig testdatasett
  - Kan teste de samme aspektene uavhengig av sikkerhetssone
  - Bedre forutsetninger for å lykkes med å flytte primæransvaret for test ned i utviklingsteamene
- Mer effektiv manuell test – testdata med komplette sammenhenger som testere kan kjenne seg igjen i
- Muliggjør kontinuerlig regresjonstest av komponenter i akseptansetest-tilstand på ferske produksjonsdata.

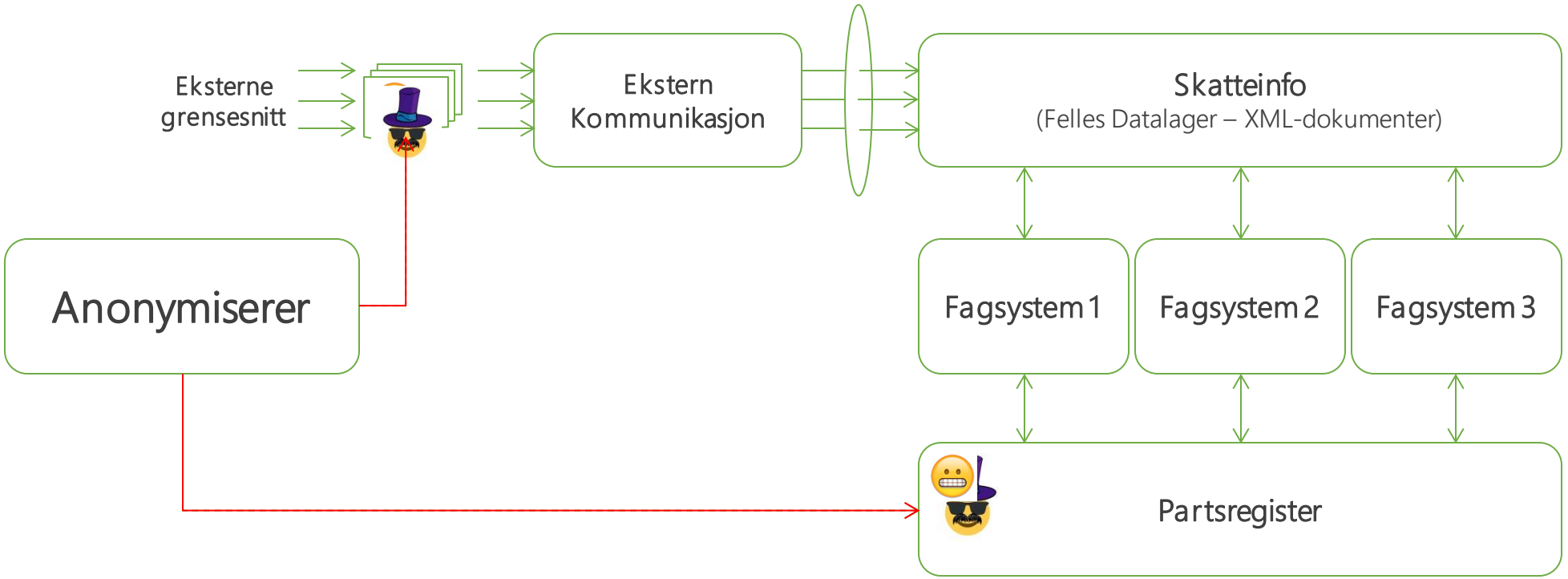
Hvilke data til hvilke type testing?

- Syntetiske data
  - Velegnet til automatiserte tester
  - Nødvendig der hvor det ikke allerede finnes reelle data
  - Test mot eksterne
- Anonymiserte data
  - Velegnet til utforskende funksjonell test på alle testnivåer
  - Velegnet til tester som krever volum og variasjon
  - Regresjonstester på siste testnivå før produksjonssetting
  - Bør primært brukes til test internt

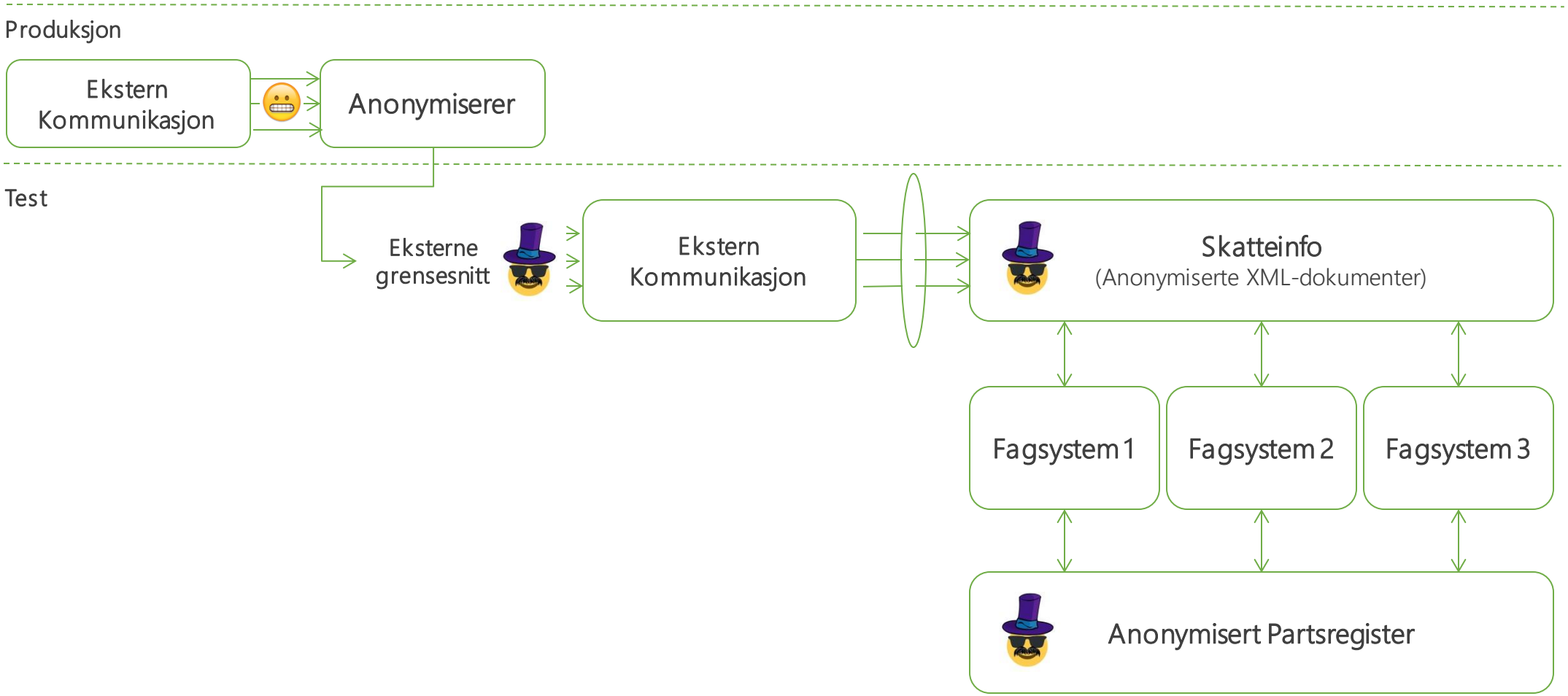
Modernisert systemportefølje



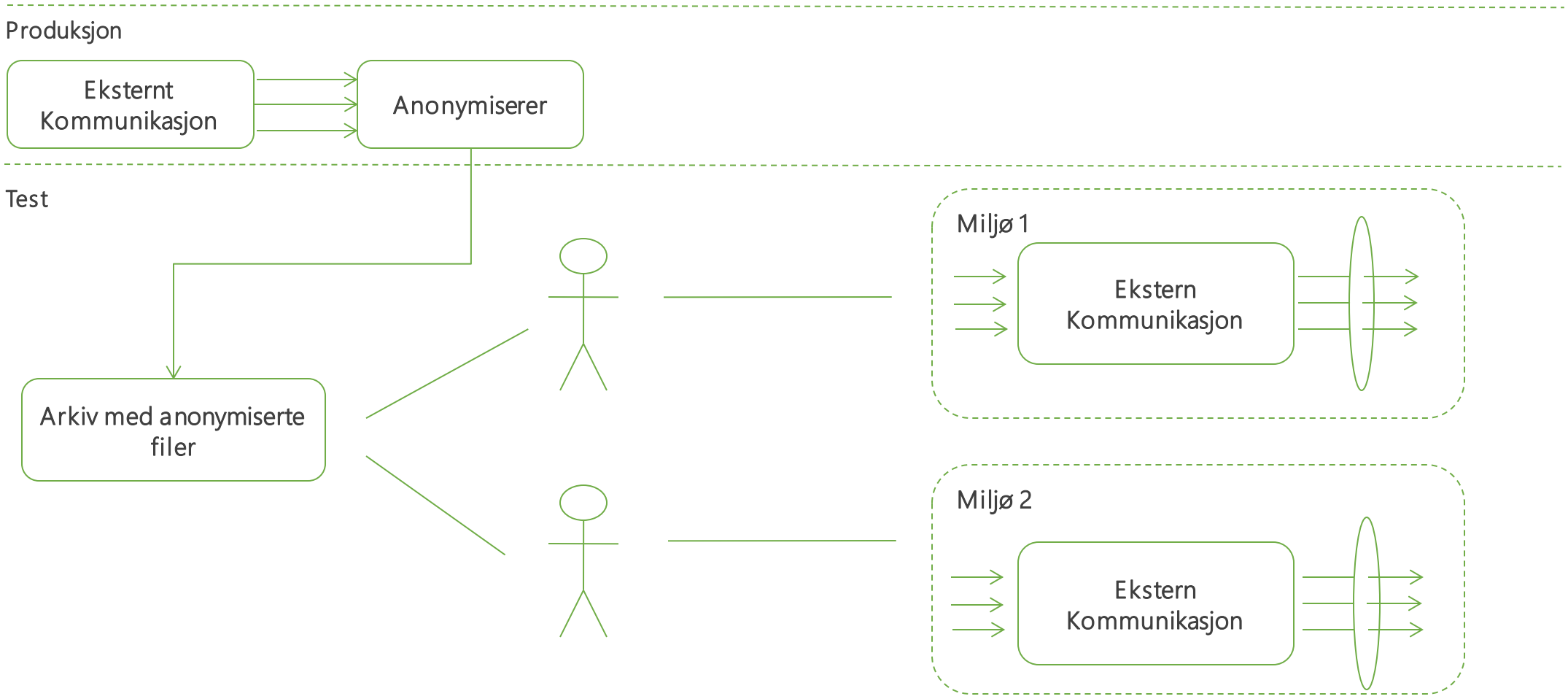
Anonymiserer-komponent



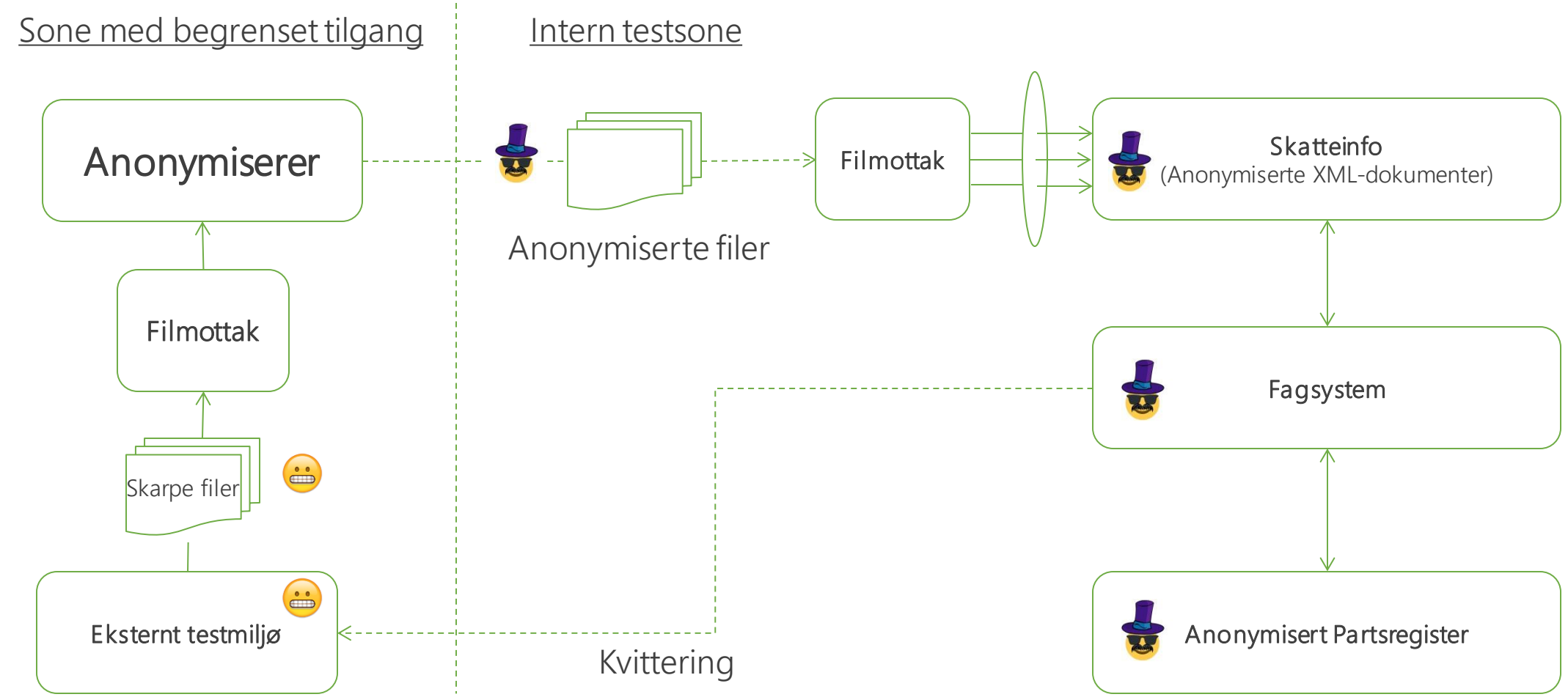
Brukscenario 1: Kontinuering strøm av anonymiserte produksjonsdata



Brukscenario 2: Ad hoc-testing med anonymiserte testdata



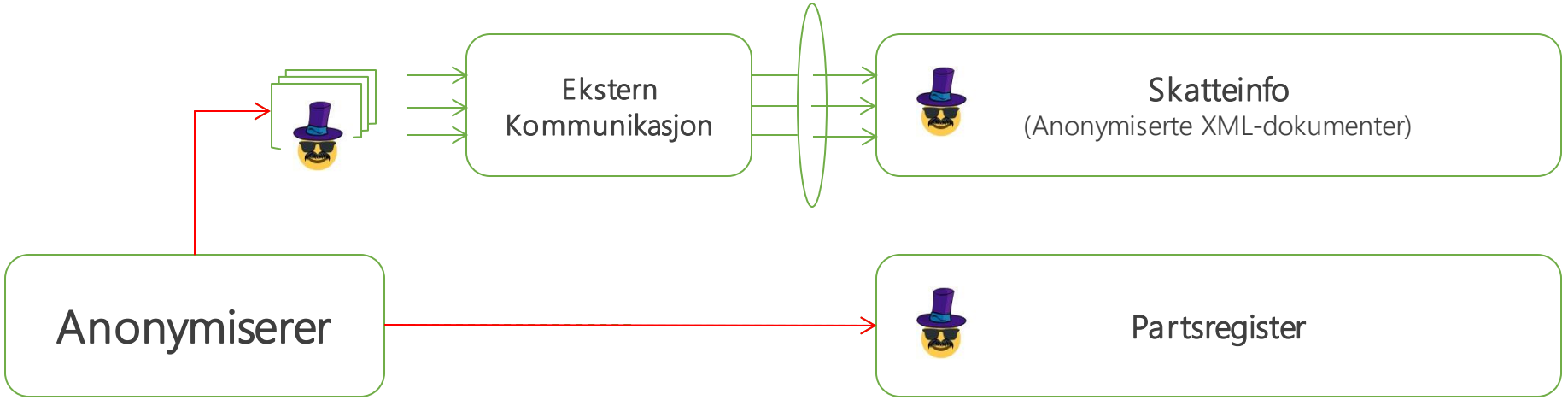
Brukscenario 3: Integrasjonstest mot skarpt miljø



Hvordan har vi anonymisert data?

Konsistente anonymiserte data

- Formål: Anonymisere data på en slik måte at de er konsistente på tvers av komponenter og systemer og følgelig kan benyttes til test av integrasjoner og verdikjeder.



Anonymiseringsnivå

- Overordnet:
  - Anonymiserer all identifiserende informasjon på hver enkelt part, som fnr, dnr, orgnr, navn, adresse, fødselsdato, kontaktinformasjon, etc.
  - Relasjonene i dataene beholdes
- Personer med hemmelig adresse osv. fjernes fra datagrunnlaget
- For utenom partsinformasjon må øvrige identifiserende informasjonselementer anonymiseres, som f. eks. kontonummer i Saldo/Rente-oppgaver fra bankene

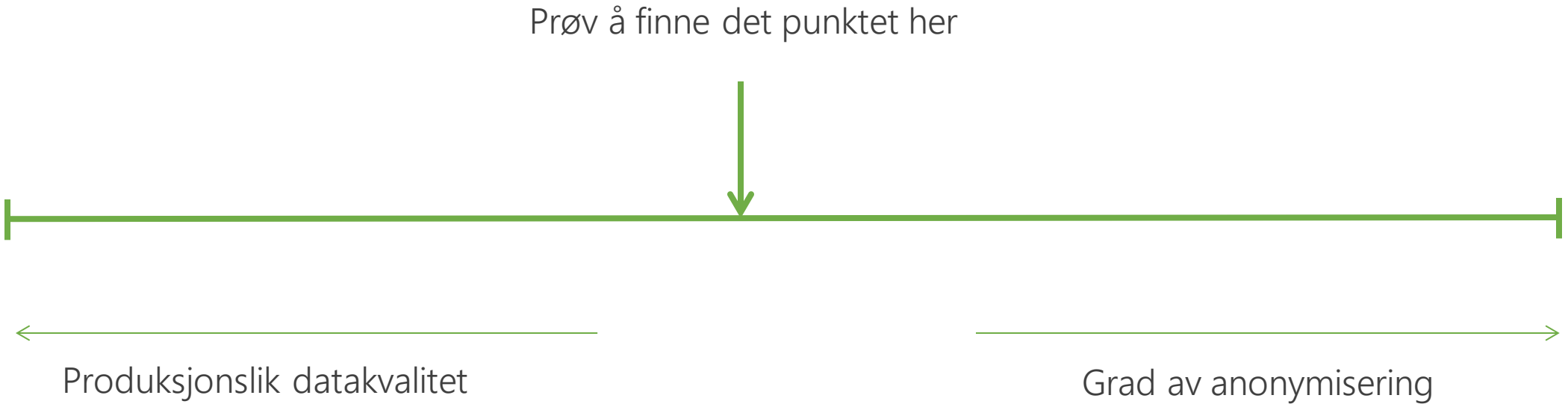
Anonymisering av fnr/orgnr

- Anonymisering av fødselsnummer (og dnr):
  - Anonymiserte fødselsnumre skal være gyldige fødselsnumre (validere)
  - Anonymiserte fødselsnumre kan være i bruk av reelle parter
  - Ivaretar fødselsår og kjønnsopplysning
- Anonymisering av organisasjonsnummer:
  - Validerer med tanke på kontrollsiffer
  - Ellers ingen logikk i organisasjonsnumre

Anonymisering av andre data

- Anonymisering av adresser:
  - Anonymiserer "alle" felter, inkludert kommunenummer og postnummer
- Forsøker å opprettholde kvalitet og distribusjon av verdier, slik at det i størst mulig grad gjenspeiler produksjon
- Regler:
  - Forhåndsdefinerte relasjoner
  - Deterministisk
  - Tilfeldig

Tips


Prøv å finne det punktet her
Produksjonslik datakvalitet
Grad av anonymisering

Kostnad

- Utvikling av løsning:
  - Fire ressurser i 15 måneder.
  - To interne og to eksterne
- Drift og vedlikehold:
  - Halv ressurs årlig
- Videreutvikling
  - Skjer i takt med moderniseringen og bekostes av prosjektene

Skatteetaten