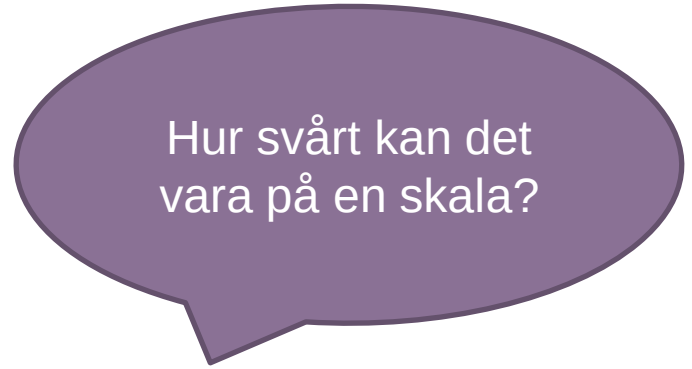
Avidentifiering – testdata på Skatteverket

2022-11-07

Agenda



- Bakgrund och behov
- Några begrepp och begreppsdefinitioner
- Riskhantering
- Informationsklassificering & Konsekvensbedömningar
- Fiktivt testdata från Navet
- Tjänsten (Shared service): Avidentifiering – testdata
- Fallgropar och utmaningar
- Hur vi jobbar med avidentifiering idag
- Testdata kommer i tre storlekar (S/M/L) på Skatteverket

Skatteverket

Varför behövs produktionsdata för testsyften?

- För enkelt eller tillrättalagt testdata
- Svårt och dyrt att skapa tillräckligt med testdata (volymer och produktionslikt)
- Tillgång till konsistent testdata i stora volymer
- Större variation och mer produktionslikt testdata behövs för att säkerställa kvaliteten på leveranser till produktion
- Träningsdata för AI-lösningar
- …

Skatteverket

Vilka är de främsta behoven idag på Skatteverket?

Avidentifierat produktionsdata:

- Pilottester inom ÅB (Årlig beskattning)
- Prestandatestester och migreringstester inom stora AO (Applikations områden)

Delmängd av avidentifierat produktionsdata – ”Subsetting”:

- Mastertester/Systemintegrationstester inom ÅB

Skatteverket

Varför kan vi inte bara använda oss av produktionsdata (personuppgifter) i testmiljöer?

- Samma skydd och åtkomst som i produktion
- Värna om den personliga integriteten
- Samtycke från alla alltid (Artikel 4, 6, 7, 8, 9 och 18)

Skatteverket

Först några begrepp i Dataskyddsförordningen (GDPR)

- Vad är en personuppgift?
  - Med personuppgifter avses varje upplysning som avser en identifierad eller identifierbar fysisk person. Avgörande är att uppgiften, enskilt eller i kombination med andra uppgifter, kan knytas till en levande person.
- Vad betyder personuppgiftsbehandling?
  - Behandling är ett brett begrepp och innefattar allt som kan göras med personuppgifter. Till exempel kan man samla in, registrera, lagra, lämna ut eller radera dem. När någon behandlar dina personuppgifter ska de följa reglerna i dataskyddsförordningen (GDPR)
- Informationssäkerhet
  - De företag och andra organisationer som hanterar dina personuppgifter ska skydda uppgifterna mot obehörig eller otillåten behandling, förlust, förstöring eller skada genom olyckshändelse.

Skatteverket

Sekretesshantering (Känslig information)

- Uppgifterna i en handling som omfattas av sekretess till skydd för enskilda kan alltså grovt delas upp i två typer
  1. De känsliga uppgifterna som t.ex:
     - etniskt ursprung
     - politiska åsikter
     - religiös eller filosofisk övertygelse
     - medlemskap i fackförening
     - uppgifter som rör hälsa eller sexualliv
     - genetiska uppgifter
     - biometri
  2. Uppgifterna som identifierar den enskilde som personnummer, namn och adress mfl.

Skatteverket

Centrala begrepp

- Anonymisering
- Pseudonymisering
- Avidentifiering
- Pseudonymisering kontra Avidentifiering

Skatteverket

Definitioner - Anonymisering

- Principerna för dataskyddet bör därför inte gälla för anonym information, nämligen information som inte hänför sig till en identifierad eller identifierbar fysisk person, eller för personuppgifter som anonymiserats på ett sådant sätt att den registrerade inte eller inte längre är identifierbar.
- För anonymisering krävs att två förutsättningar är uppfyllda:
  - Anonymiseringen är oåterkallelig
  - Anonymiseringen har gjorts på ett sådant sätt att det inte går att identifiera den fysiska personen ifråga.

Skatteverket

Definitioner - Pseudonymisering

Artikel 4:5 Behandling av personuppgifter på ett sätt som innebär att personuppgifterna inte längre kan tillskrivas en specifik registrerad utan att kompletterande uppgifter används, under förutsättning att dessa kompletterande uppgifter förvaras separat och är föremål för tekniska och organisatoriska åtgärder som säkerställer att personuppgifterna inte tillskrivs en identifierad eller identifierbar fysisk person.

Skatteverket

Definitioner - Avidentifiering

Artikel 29-arbetsgruppen för skydd av personuppgifter.

Anonymisation techniques
Avidentifieringsmetoder/Avidentifieringstekniker

Ovan beskrivs och utvärderas

Skatteverket

Pseudonymisering kontra Avidentifiering

Pseudonymisering - ersättning av en identitetsbeteckning med en pseudonym. Detta används bland annat i medicinsk och samhällsvetenskaplig forskning. Man byter ut namn, personnummer eller annan beteckning som är knuten till en bestämd person mot ett namn eller nummer som inte kan knytas till den personen. Det är däremot möjligt att ta reda på vilken person en pseudonym står för om man har tillgång till extra information, som ska hållas hemlig och åtskild från det undersökningsmaterial som bearbetas.

Pseudonymisering är alltså inte samma sak som avidentifiering av material; avidentifiering innebär att all information som kan användas för att identifiera personer tas bort permanent.

Skatteverket

Pseudonymisering kontra Avidentifiering

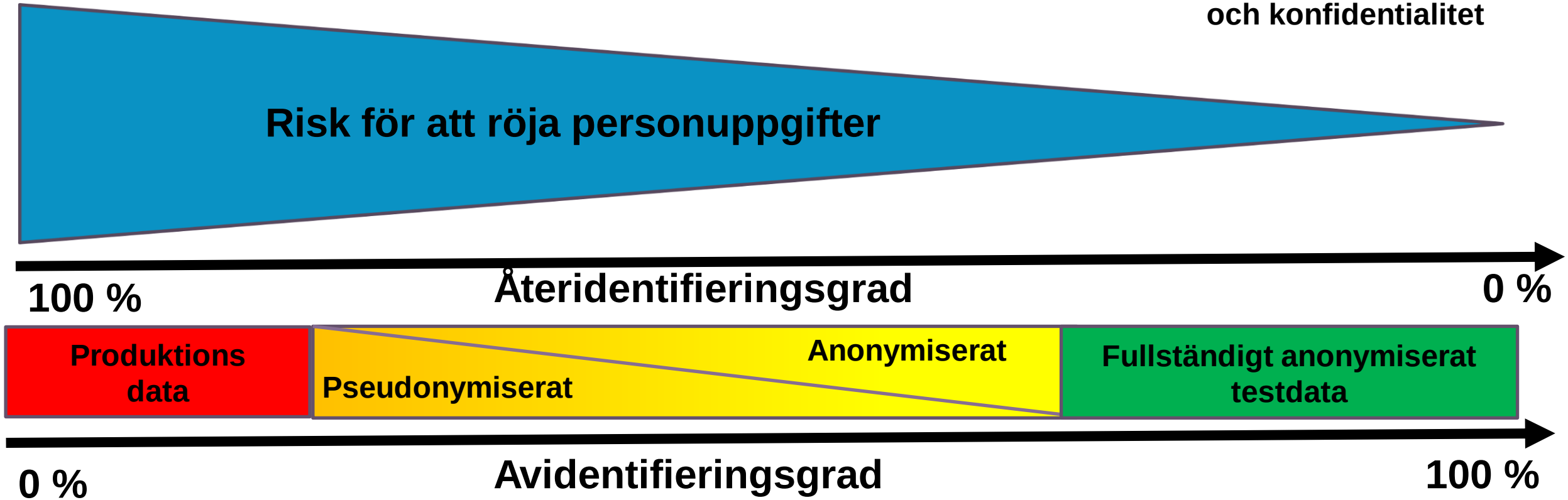
- Förstöra/ta bort
- Byta ut (t ex genom perturbation eller inte)
- Förändra (Smutsa ner/lägga på brus)

Skatteverket

Skala – Avidentifierings- / Återidentifieringsgrad

Låg personlig integritet och konfidentialitet

Hög personlig integritet och konfidentialitet



Får inte användas som testdata i testmiljöer (GDPR)

Bibehållet informationsvärde

Tillfredställande nivå av avidentifiering (pseudonymisering/ anonymisering)

Är inte en personuppgift och behöver inte underkasta sig GDPR

Är antagligen inte så användningsbart som testdata

Skatteverket

Etablera en process för adekvat nivå av avidentifiering

Faktorer att ta hänsyn till:

- Sannolikheten för återidentifiering
- Sannolikheten för att ett återidentifieringsförsök skulle lyckas
- De avidentifierings- metoder/tekniker som kan användas
- Kvaliteten på testdata efter det att det blivit avidentifierat och att det möter organisationens behov och krav på testdata

Skatteverket

Testdata från Folkbokföringens aviseringssystem - NAVET

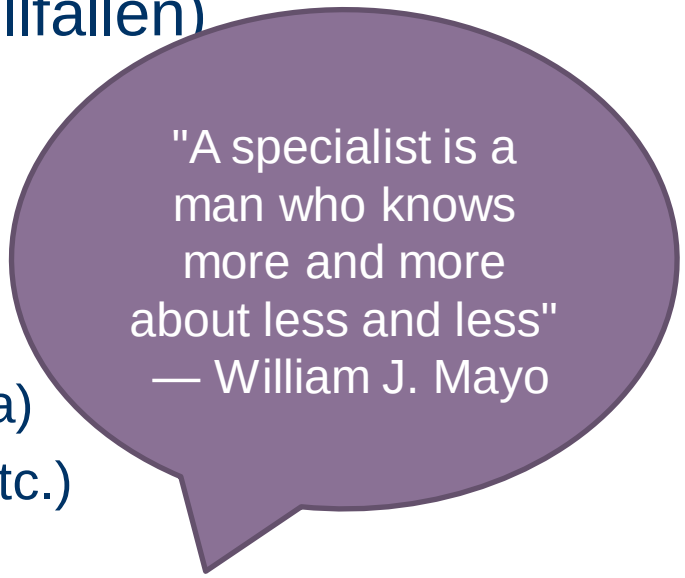
Testdata drygt 8 000 st fiktiva personnummer inkl. övriga personuppgifter i excelformat som kan hämtas och användas kostnadsfritt för olika testsyften från Skatteverkets hemsida: Teknisk information | Skatteverket

- https://www.skatteverket.se/offentligaaktorer/informationsutbyte/navethamtauppgifteromfolkbokforing/tekniskinformation.4.2106219b17988b0d23160f.html?q=testdata

Skatteverket

Tjänsten (Shared service): Avidentifiering – testdata

- Nyutveckling/vidareutveckling
  - Tar fram avidentifieringslösningar för befintliga system
  - Deltahantering (förändring mellan avidentifieringstillfällen)
- ”Subsetting”
  - ”Skär ut” en delmängd av avidentifierat data
- Förvaltning
  - Uppdaterar och förbättrar avidentifieringslösningar (inkl. delta)
  - Effektivisering (Automatisering, Fast Recovery, Optimering etc.)



Skatteverket

Hur jobbar vi med avidentifiering på Skatteverket?

- Analys – (Vad behöver/måste avidentifieras)
  - Informationsklassificering och Konsekvensbedömning
  - Databaser och scheman
  - Tabeller och kolumner (ADM – Application Data Model)
- Implementation och verifiering - (Hur)
  - Avidentifieringslösning med testdata i testmiljö
  - Val av eller kombination av avidentifieringstekniker
  - Lösningsmönster med avidentifieringsfunktioner (återanvändning)
  - Nycklar/korsreferenser (Referentiell integritet)
  - Automatisering
- Validering
  - Avidentifieringslösning med produktionsdata

Skatteverket

Juridiska-, organisatoriska-, säkerhets-, och tekniska aspekter

- Regulatoriska krav och principer - GDPR
  - Principer för behandling av personuppgifter och säkerhet (Artikel 5, 6 och 32)
  - Inbyggt dataskydd och dataskydd som standard (Artikel 25)
  - Uppgiftsminimering, lagringsminimering, mm. (Artikel 47 d och Artikel 87)
- Bestäm den säkerhetsnivå som är lämplig i förhållande till:
  - **tillgänglig teknik**
  - **kostnaden** för åtgärderna
  - om det finns några särskilda **risker** med behandlingen
  - hur pass **känsliga** uppgifterna är

"... och det vi inte kan prata om måste vi vara tysta."
- Ludwig Wittgenstein

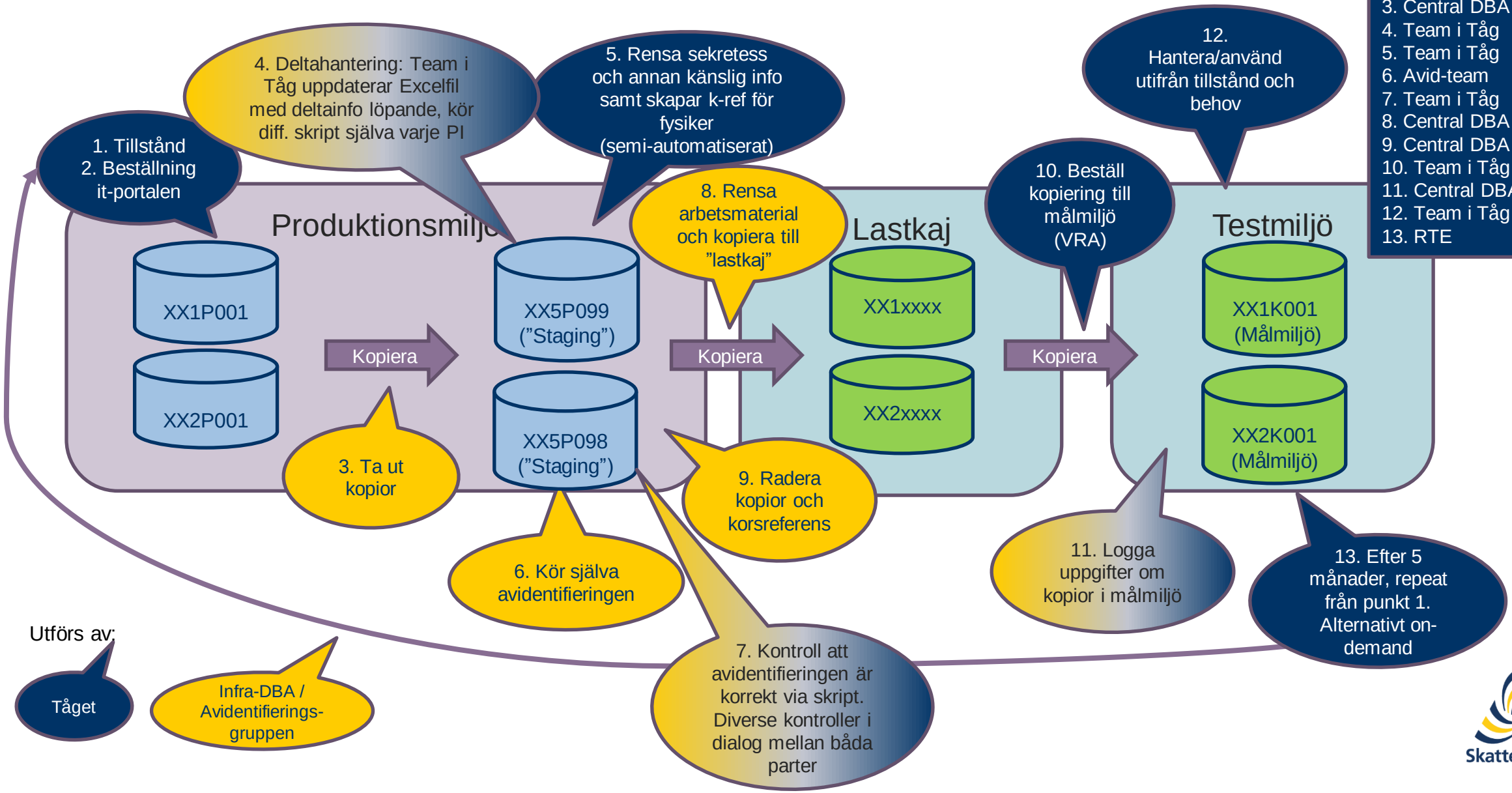
"... och det som vi inte kan preudonymisera/anonymisera/ avidentifiera på ett tillräckligt säkert sätt måste vi låta bli."
- x x

Några exempel hur vi avidentifierar olika uppgifter

Uppgift (direkta-, indirekta personuppgifter)	**Avidentifieras genom (Metod/Teknik)**
Sekretess, Skyddad Folkbokföring	Tas bort (alla poster som hör till person)
Personnummer (Avlidna)	Byts ut mot lediga personnummer (Födelseår behålls)
Namn/Namndelar/Aviseringsnamn/Företagsnamn	Byts ut mot andra befintliga namn/namndelar *
Adresser (Gatuadress, Box, Postnr etc.)	Hela Sveriges befolkning flyttar runt * (lägenhet till annan lägenhet, villa till..)
Organisationsnummer	Byts ut inom jurformen *
Ärendenummer/Diarienummer	Byts ut *
Datum (t ex händelser/händelsesamband)	Smutsar ner/ändrar datum
Fritextfält (ostrukturerad information) och t ex scannade handlingar (bilder med personuppg.)	Byts ut mot ”dummy” information i samma storlek
*) Olika vid varje avidentifieringstillfälle	Randomiseras/slumpas

Skatteverket

Avidentifiering framåt (2.1) - Exekveringsprocess



Vem gör vad?

1. Linjen/BO för Tåg
2. RTE för Tåg
3. Central DBA
4. Team i Tåg
5. Team i Tåg
6. Avid-team
7. Team i Tåg
8. Central DBA
9. Central DBA
10. Team i Tåg
11. Central DBA/Team i Tåg
12. Team i Tåg
13. RTE

Skatteverket

Erfarenheter

- Grädde på moset/Lök på laxen – positiva bieffekter
  (Registerproblem/Registervård/Registerkvalitet)

- ”Norgehistorien” – Lessons learned
  (Balans mellan användbarhet och återidentifieringsrisk)

Skatteverket

Tio små fallgropar…

1. Avsaknad av bra avidentifieringsmiljö och verktyg
2. Återstartspunkter
3. Missar att identifiera personuppgifter/identifikatorer
4. Missar att avidentifiera eller ta bort personuppgifter
5. Kontroller avskilda från specifikation
6. Defaultvärden om ej lyckad avidentifiering
7. Kan ej säkerställa den referentiella integriteten eller att avidentifierat data fungerar i systemet/applikationen
8. Olika databaser avidentifierade på olika sätt eller vid olika tidpunkter så att avidentifierat data inte hänger ihop över system/applikationer
9. Överskattar förmågan och underskattar riskerna

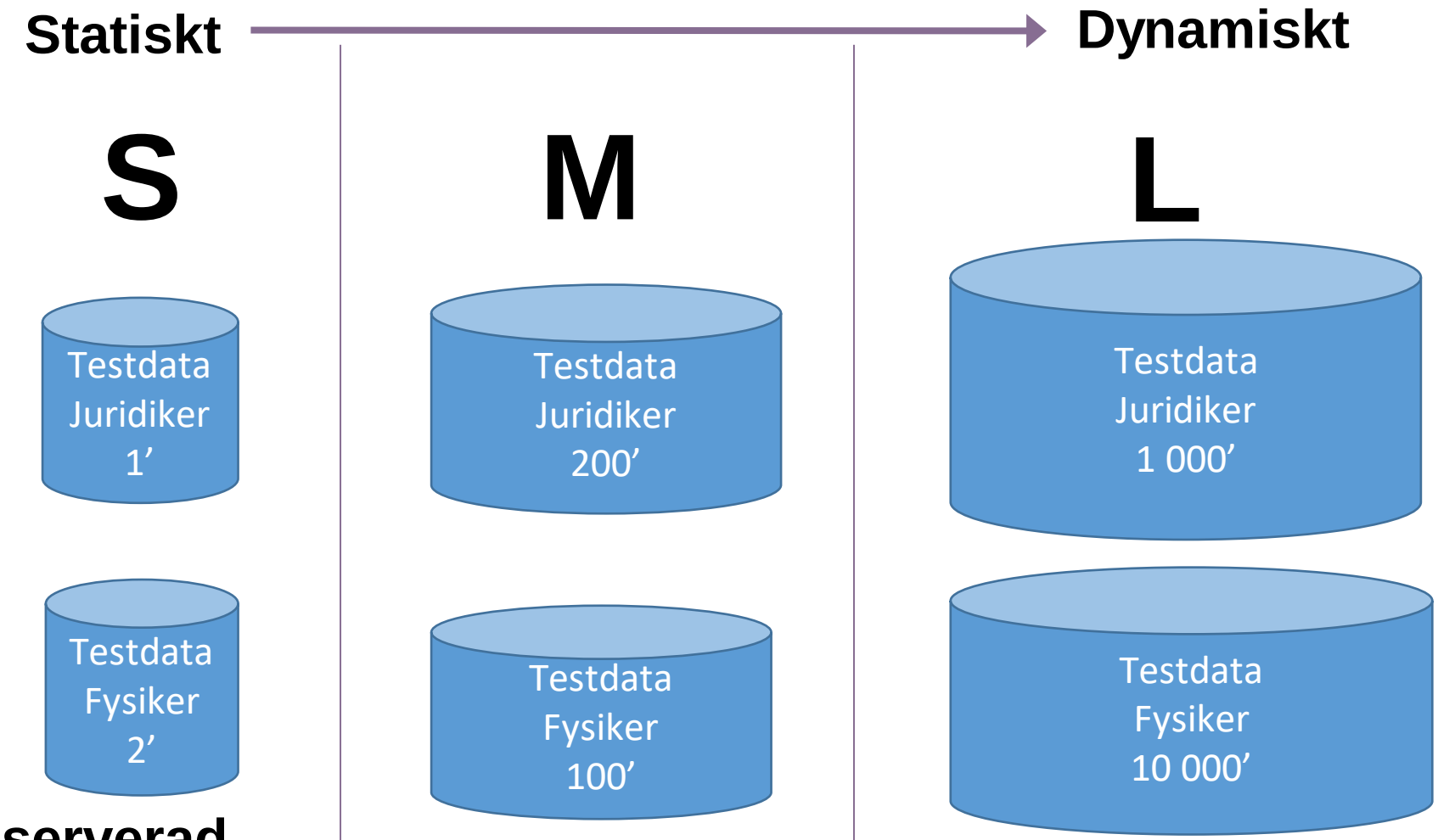
Skatteverket

Summering eller tips & råd

- Informationsklassificering och konsekvensbedömning - lär känna ditt data
- Säkerhet/behörighet (t ex hantering av nycklar/k-ref) - skydda
- Sekretess eller annan känslig information – ta bort/utelämna/ersätt
- Ostrukturerade uppgifter (Fritext, dokument/handlingar etc.) – ta bort ersätt
- Indirekt utpekning (Avlidna personer, Ärende-/diarienr,…) – minimera
- ”Svansar” (Månggifte, Hundraåringar som försvinner, Adoptivbarn mfl.) – eliminera det som sticker ut
- Mönster/Kvasiidentifierare (Händelsetidpunkter, Händelsesamband, …) – Smutsa ner

Skatteverket

Testdata kommer i tre storlekar på Skatteverket!



Skatteverket

Referenser/länkar till nyttig information om avidentifiering

- Anonymisation: Managing data protection risk code of practice (ICO = Information Commissioner´s Office)
- ARTIKEL 29-ARBETSGRUPPEN FÖR SKYDD AV PERSONUPPGIFTER (0829/14/EN WP216)
- Dataskyddsförordningen (IMY = Integritetsskyddsmyndigheten)
- eSam "Vägledning om pseudonymisering av personuppgifter" ES2022-01

Skatteverket