# Overheidsdata & Onderzoek
## Werk samen met wetenschappers!

*Dr. Erik-Jan van Kesteren*
*Assistant Professor in Data Science / Statistics*
*Universiteit Utrecht*

# Werk samen met wetenschappers!

- Uitvoerende taken leiden tot wetenschappelijke vragen
- Wetenschappers brengen innovatie in gegevensdeling, verzameling, bescherming
- Onderzoek heeft een speciale rol in gegevensverwerking (en leidt tot nieuwe inzichten!)

# Wetenschappers willen dit ook

- Wetenschap in transitie: meer focus op impact van ons werk
- Overheidsgegevens kunnen nieuwe, interessante wetenschappelijke vragen beantwoorden
  Bewijs: honderden wetenschappelijke onderzoeksprojecten tegelijkertijd bij het CBS

# Context: wie ben ik

- Onderzoeker in statistiek, methodologie, data science, computational social science
- Universitair docent: data science
- Team leider van ODISSEI Social Data Science team Dataconsultancy voor complexe datavragen van SSH wetenschappers
- Generalist: vandaag privacy, fairness, synthetic data, causal inference

Utrecht University

SoDA
ODISSEI Social Data Science Team

**Géén privacy officer / CDO!**

# Uitvoerende taken leiden tot wetenschappelijke vragen

**Uitvoerende taak**

- Nieuw fraudealgoritme nodig voor DUO (OCW)
- Eerlijkheid van dit algoritme heeft hoge prioriteit

**Wetenschappelijke vraag**

- Hoe bepaal ik in dit specifieke geval wat “eerlijk” betekent?
- Wat is belangrijk en wat is bijzaak?

*Samenwerken: wetenschappelijke onderbouwing ingebouwd*

https://www.autoriteitpersoonsgegevens.nl/actueel/fraudeaanpak-duo-discriminerend-en-illegaal

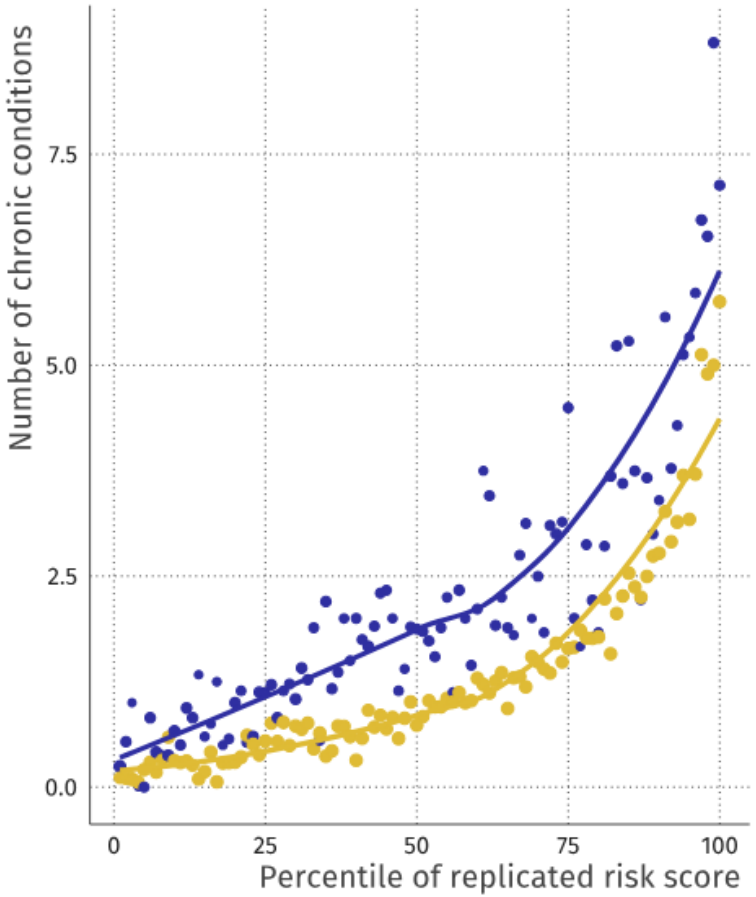
# Sneak preview: vragen

- Welke attributen zijn “gevoelig”?
  - Migratieachtergrond, geslacht, studieniveau, huishoudenstype, …
- Welke definitie van eerlijkheid hanteren we, en welke “metric” hoort daarbij?
  - Statistical parity: positive prediction rate
  - Equal opportunity: false negative rate
  - …
- Welke indirecte effecten op eerlijkheid heeft het algoritme? Onderscheid op inkomen -> onderscheid op migratie
- Meten we wat we zouden moeten meten?

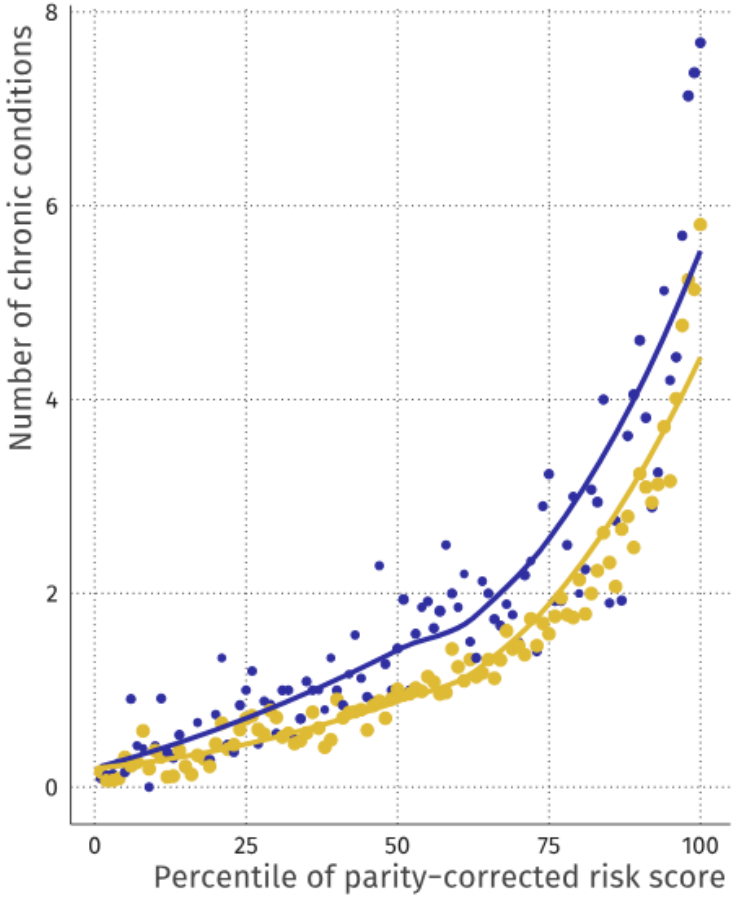
# Sneak preview: antwoorden

- **Een eerlijk algoritme heeft slechtere “performance” dan een oneerlijk algoritme**
- Het algoritme moet de gevoelige attributen kennen tijdens training. **Deze data moet dus beschikbaar zijn!**
- Tijdens voorspelling: maak het systeem eerlijk
  - Iedereen een vrouw? (statistical parity, maar niet equal opportunity)
  - Aangepaste probability cut-off waarden per gevoelige categorie?
- Eerlijk op de ene variabele betekent niet meteen eerlijk op de andere: hoe meten / definiëren we fraude?

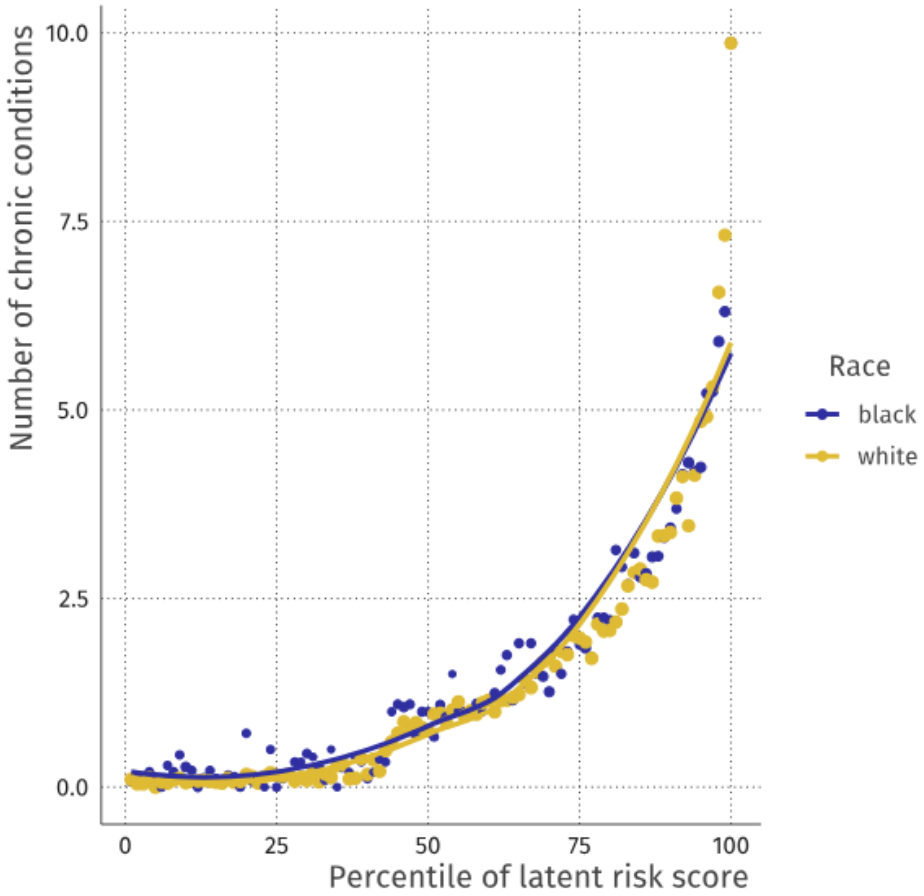
Risicoscore op gezondheidskosten is oneerlijk m.b.t. ras




Parity correction op gezondheidskosten lost het probleem niet helemaal op



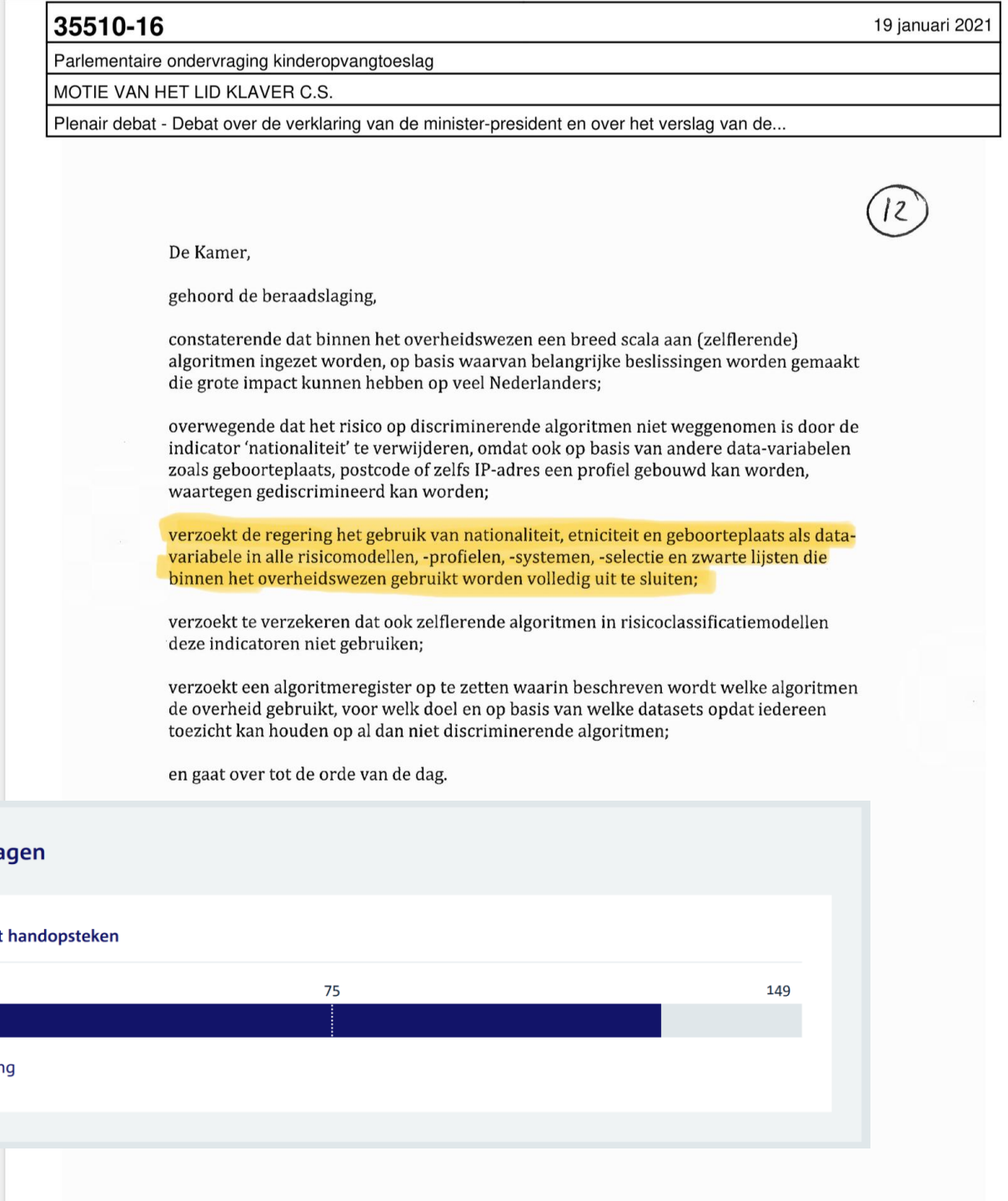

Operationaliseer de juiste uitkomstmaat om verschillende rassen gelijk te behandelen




Boeschoten, L., Van Kesteren, E. J., Bagheri, A., & L. Oberski, D. (2021). Achieving Fair Inference Using Error-Prone Outcomes. International Journal of Interactive Multimedia and Artificial Intelligence, 6(5), 9–15. https://doi.org/10.9781/ijimai.2021.02.007

# Fairness through unawareness?

- Een gevoelige variabele weglaten is niet eerlijk
- “Ongevoelige variabelen” kunnen indirecte oneerlijkheid teweegbrengen



35510-16 19 januari 2021

Parlementaire ondervraging kinderopvangtoeslag

MOTIE VAN HET LID KLAVER C.S.

Plenair debat - Debat over de verklaring van de minister-president en over het verslag van de...

12

De Kamer,

gehoord de beraadslaging,

constaterende dat binnen het overheidswezen een breed scala aan (zelflerende) algoritmen ingezet worden, op basis waarvan belangrijke beslissingen worden gemaakt die grote impact kunnen hebben op veel Nederlanders;

overwegende dat het risico op discriminerende algoritmen niet weggenomen is door de indicator ‘nationaliteit’ te verwijderen, omdat ook op basis van andere data-variabelen zoals geboorteplaats, postcode of zelfs IP-adres een profiel gebouwd kan worden, waartegen gediscrimineerd kan worden;

verzoekt de regering het gebruik van nationaliteit, etniciteit en geboorteplaats als data-variabele in alle risicomodellen, -profielen, -systemen, -selectie en zwarte lijsten die binnen het overheidswezen gebruikt worden volledig uit te sluiten;

verzoekt te verzekeren dat ook zelflerende algoritmen in risicoclassificatiemodellen deze indicatoren niet gebruiken;

verzoekt een algoritmeregister op te zetten waarin beschreven wordt welke algoritmen de overheid gebruikt, voor welk doel en op basis van welke datasets opdat iedereen toezicht kan houden op al dan niet discriminerende algoritmen;

en gaat over tot de orde van de dag.

Stemmingsuitslagen

Aangenomen met handopsteken

Voor 75 149

127

Detail stemming

# Conclusie

- Eerlijkheid in algoritmen is complex
- Waar je op moet letten kan een wetenschapper uitzoeken
- Dit veld is constant in beweging

- Pragmatisch: je moet wel een algoritme opleveren
- Keuzes maken en die keuzes onderbouwen betekent samenwerken

# Goed voorbeeld: Meer uren werkt!

**Uitvoerende taak**

- Deeltijdmedewerkers in Nederland meer laten werken
- Implementeer interventies met grote werkgevers (e.g. betere roostering o.b.v. data)

**Wetenschappelijke vraag**

- Welke interventies werken?
- Hoe meet je of een interventie goed werkt?

https://www.rijksoverheid.nl/ministeries/ministerie-van-sociale-zaken-en-werkgelegenheid/programmas/meer-uren-werkt

# Wetenschappers brengen innovatie in gegevensdeling, verzameling, bescherming, infrastructuur

# ODISSEI

*Open Data Infrastructure for Social Science and Economic Innovations*

Consortium van >40 sociale wetenschappen faculteiten, onderzoeksinstituten, publieke onderzoeksbureaus, infrastructuur organisaties, …

Vanuit overheid: CBS, SCP, CPB, PBL, RIVM, KiM

https://odissei-data.nl/

# ODISSEI

*Open Data Infrastructure for Social Science and Economic Innovations*

**Duurzame onderzoeksinfrastructuur voor sociale vraagstukken**

- Toegang tot data (e.g., CBS microdata, LISS panel)
- Veilige compute (ODISSEI Secure Supercomputer, SANE)
- Coördinatie van gegevensverzameling (ESS, NKO, …)
- Community building (ODISSEI Conference)
- Ondersteuning bij datagebruik (SoDa team)

https://odissei-data.nl/


# We help social scientists with data intensive & computational research

Our goal is to enhance the evidence base and impact of social science by bringing the added value of new data sources and new data analysis techniques into social research in the Netherlands

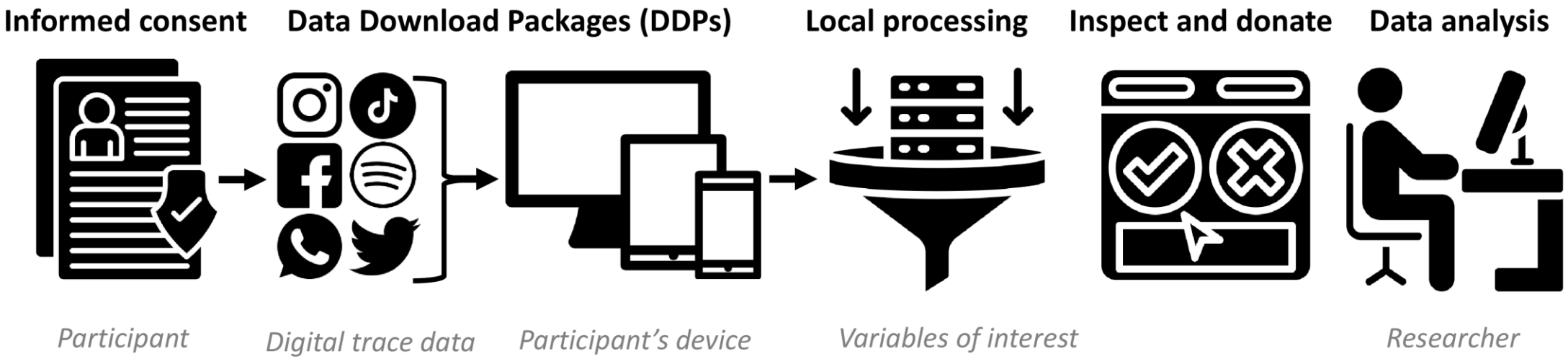
https://odissei-soda.nl

# Innovatieve gegevensdeling: SANE

- Tinker SANE: CBS microdata toegang model
  - Virtuele machine op ISO 270001 certified infra
  - Specifieke data per onderzoeksproject
  - Linkbaar met externe data op persoonsniveau
  - Output checks door data provider
- Blind SANE: code-to-data model
  - Bereid analysecode voor
  - Stuur analyse naar data, ontvang resultaten

https://odissei-data.nl/facility/secure-analysis-environment-sane/

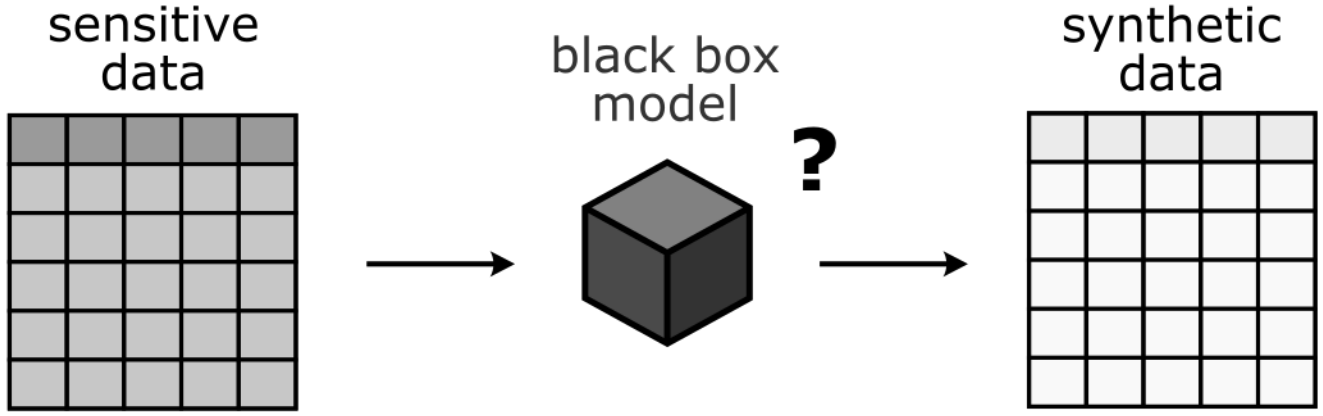
# Innovatieve gegevensverzameling: Data Donatie van Digital Trace Data




https://datadonation.eu/data-donation/

# Innovatieve gegevensbescherming: Synthetische data

- Toegang tot gevoelige data is waardevol maar heeft barrières
- Synthetische data als een oplossing?




Synth. data vaak niet “herleidbaar” maar is **niet** per definitie veilig

van Kesteren, E. J. (2024). To democratize research with sensitive data, we should make synthetic data more accessible. *Patterns*, *5*(9).

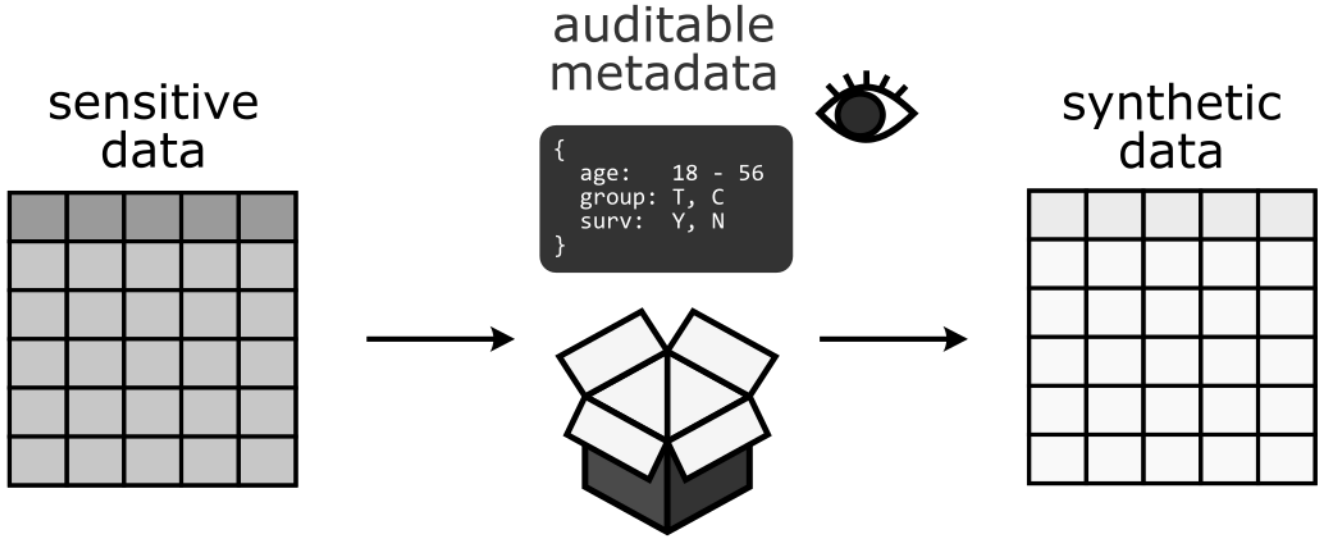
# Innovatieve gegevensbescherming: Synthetische data

- Een complex genoeg model kan de gevoelige data exact reproduceren (perfecte “fidelity”)
- Dan heb je geen privacy meer, dus wees voorzichtig
- Hoe groot is privacy-risico? Wetenschappelijke vraag


van Kesteren, E. J. (2024). To democratize research with sensitive data, we should make synthetic data more accessible. *Patterns*, *5*(9).

{meta}syn

- Onze oplossing: simpel model dat alleen structuur reproduceert



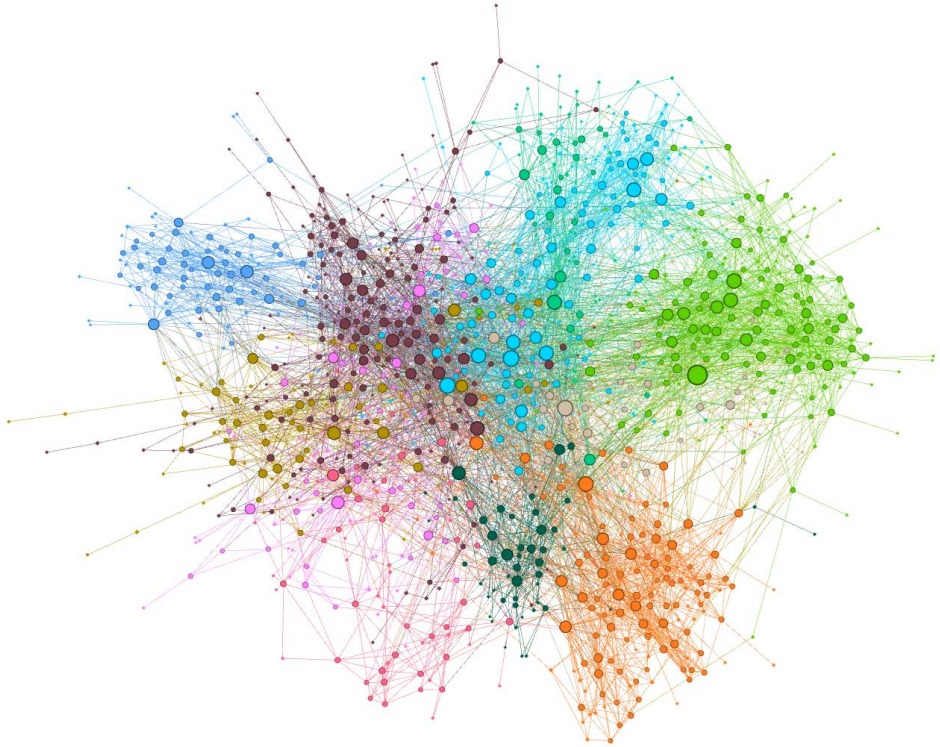

- Toegang tot echte data nog nodig voor analyse, want lage “fidelity”

Schram, R., Spithorst, S., & van Kesteren, E. J. (2025). Metasyn: Transparent Generation of Synthetic Tabular Data with Privacy Guarantees. The Journal of Open Source Software, 10(105), 7099.

# Innovatieve infra

POPNET

- Populatie netwerk data (school, werk, buren, huishoudens, …)
- Sociale context van verschillende processen
- Hoe verspreidt ziekte zich over scholen, werkplekken, buurten? ODISSEI SoDa + RIVM, MinVWS https://www.nature.com/articles/s41598-024-82646-7



Van der Laan, J., De Jonge, E., Das, M., Te Riele, S., Emery, T. (2023) A Whole Population Network and Its Application for the Social Sciences, *European Sociological Review*, 39(1) 145–160, https://doi.org/10.1093/esr/jcac026

ODISSEI

- Samenwerken met wetenschappers?
- ODISSEI is een perfect startpunt
- Enorm netwerk, brede kennis, bestaande ervaring met overheidssamenwerking

https://odissei-data.nl/contact-us/

# Onderzoek heeft een speciale rol in gegevensverwerking

# UAVG artikel 24

[...] verbod om bijzondere categorieën van persoonsgegevens te verwerken niet van toepassing, indien: [...] de verwerking noodzakelijk is met het oog op **wetenschappelijk of historisch onderzoek of statistische doeleinden** overeenkomstig artikel 89, eerste lid, van de verordening; [...]

# Onderzoek met overheidsdata

**Overheid genereert data**

- Overheid doet een interventie (beleid, proof-of-concept, test)
- Daarover komt een rapport
- En dan?

**Sociaal-wetenschappelijk onderzoek**

- Interesse in (retrospectieve) analyse van interventies
- “Hoe goed werkt dit type interventie voor dit proces?”

# Voorbeeld: NPRZ

**Overheid genereert data**

- Interventie op scholen in Rotterdam Zuid
- Leerlingen extra lessen aanbieden (langer op school)

**Hoe goed werkt dit soort educatie-interventies?**

- Groot onderzoeksgebied: theoretisch en empirisch
- Historische data nodig, interventiedata nodig

# Voorbeeld: NPRZ

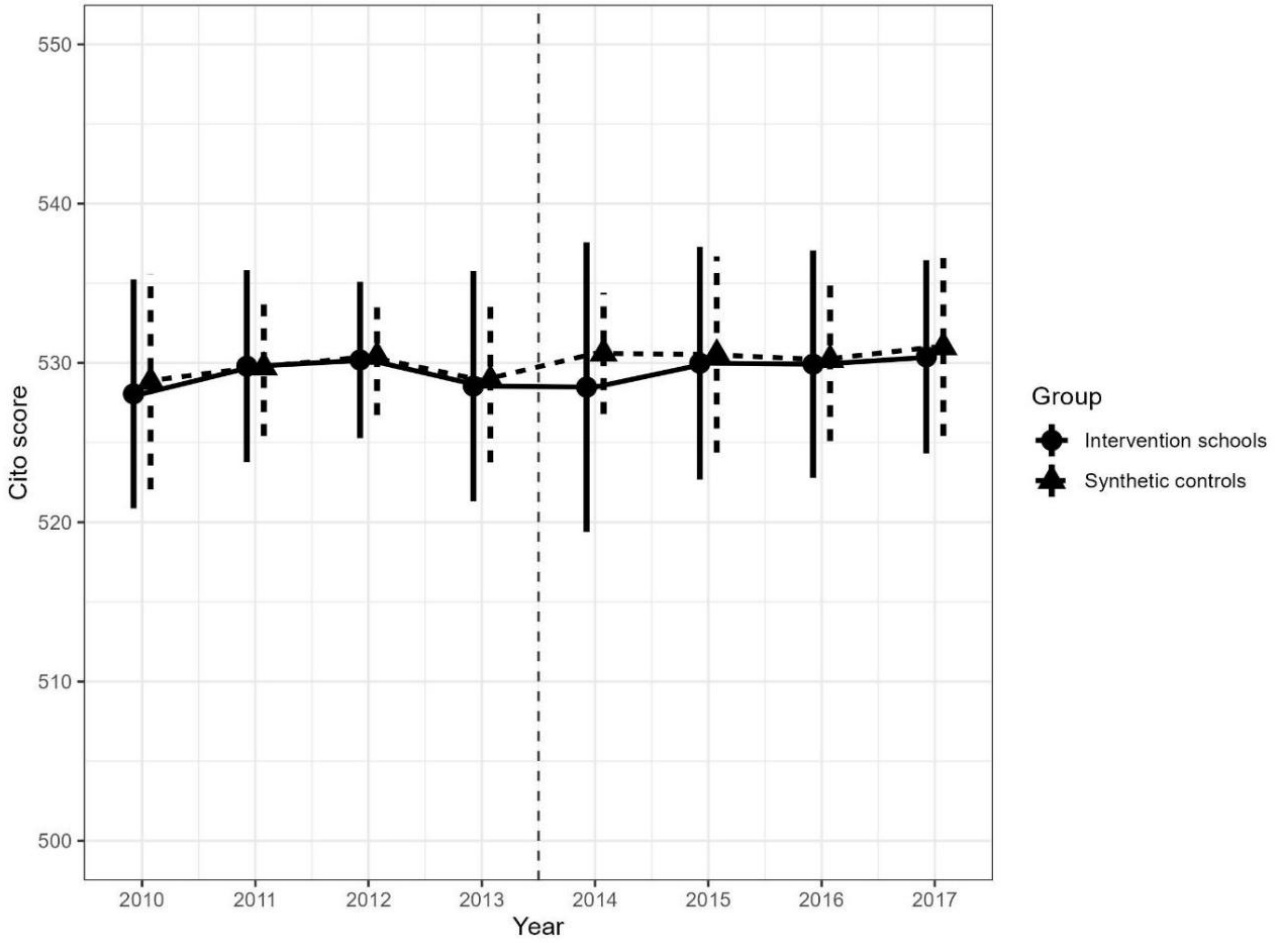
Specifieke onderzoeksvraag

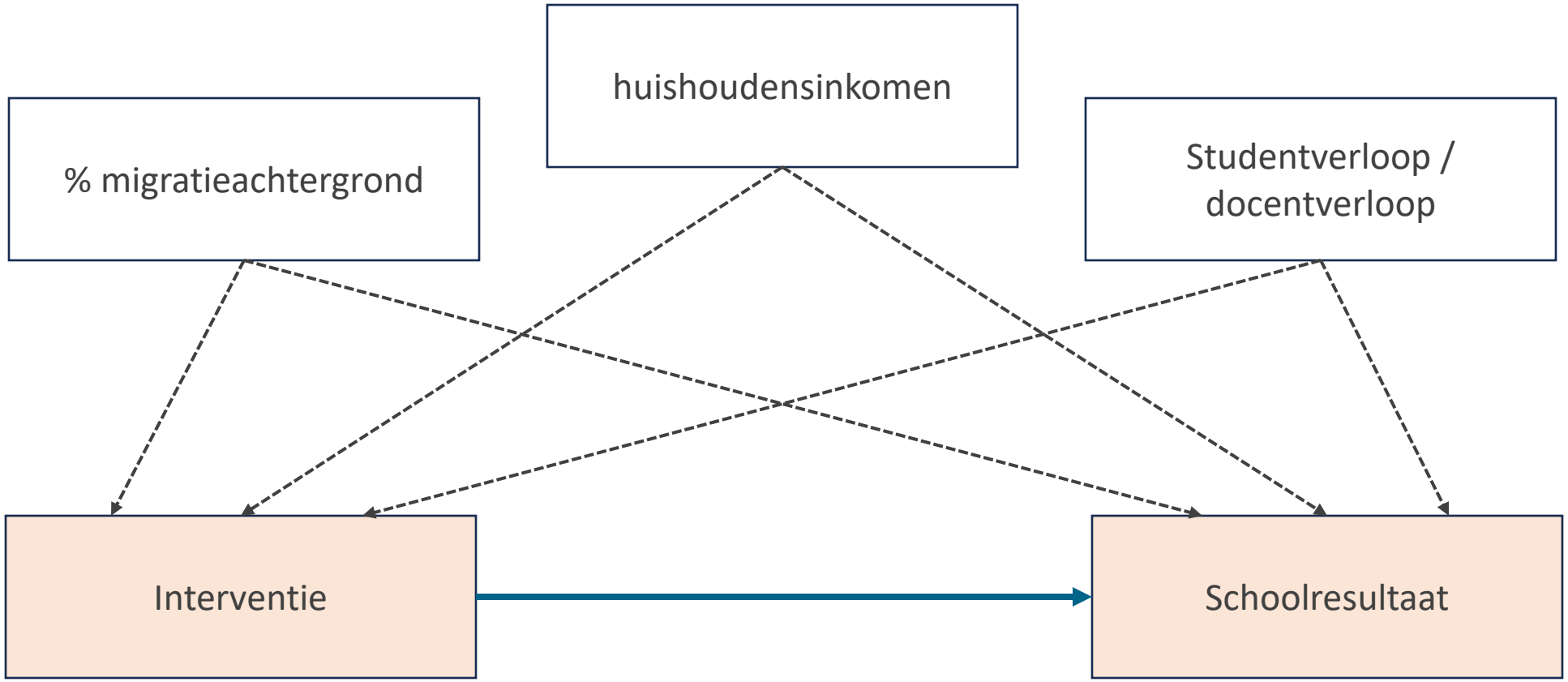
**Wat is het effect van zulke interventies op schooluitkomsten?**

- Verschillende opties:
  - Vraag het de betrokkenen
  - Doe een experiment (niet mogelijk / niet ethisch!)
  - Retrospectieve causale interventieanalyse

# Causale interventieanalyse

- Counterfactual prediction: wat zouden de schooluitkomsten zijn zonder de interventie?
- Betere historische data leidt tot geloofwaardiger antwoord
- Eigenschappen van scholen, van studenten, details van de interventie




huishoudensinkomen
% migratieachtergrond
Studentverloop / docentverloop
Interventie
Schoolresultaat

# Werk samen met wetenschappers!

- Persoonlijke connectie tussen wetenschapper en NPRZ
- Dat kan beter!
- Wetenschap kan overheidstaken versterken
- Bestaande links met bijv. SCP

**Voordeel van ODISSEI:**

Ruimte om kennis te verzamelen en verspreiden, e.g., https://causalpolicy.nl/

“community of practice” voor computational social science

# Overheid: bewaar je gegevens!

- Individuele data over iedere “unit”
- Experiment? Informatie over controlegroep
- Gegevens voor en na de interventie
- Nog beter: tijdreeks (historisch, voor en na interventie)

# Overheid: bewaar je gegevens!

- Zorg dat informatie over experimenten / interventies wordt opgeslagen

- Wild idee: centrale opslag! Van individueel experiment naar historische kennisbron over gedrag en beleid

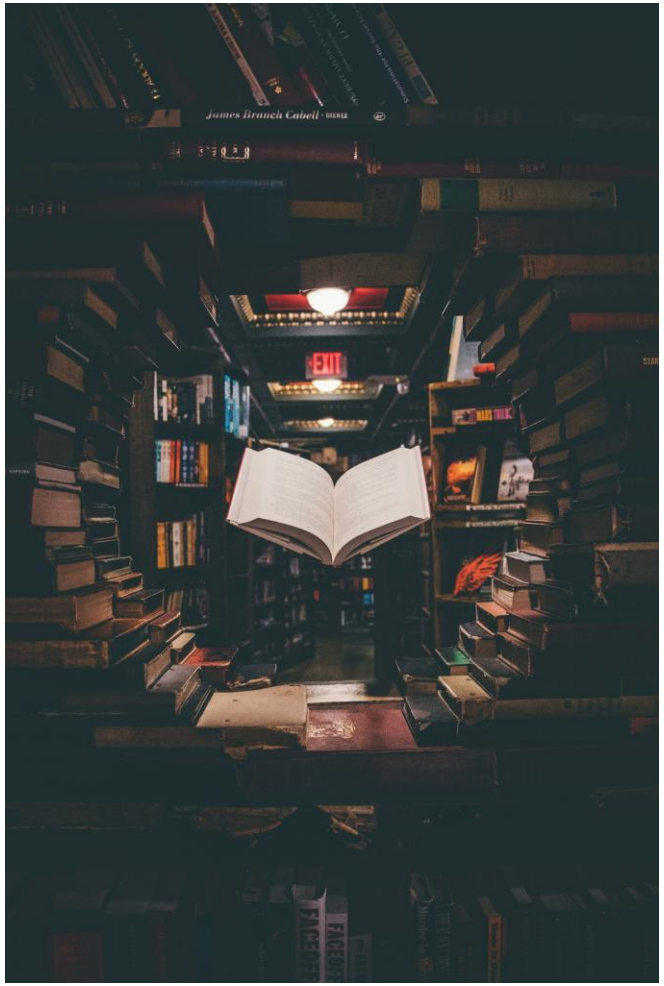
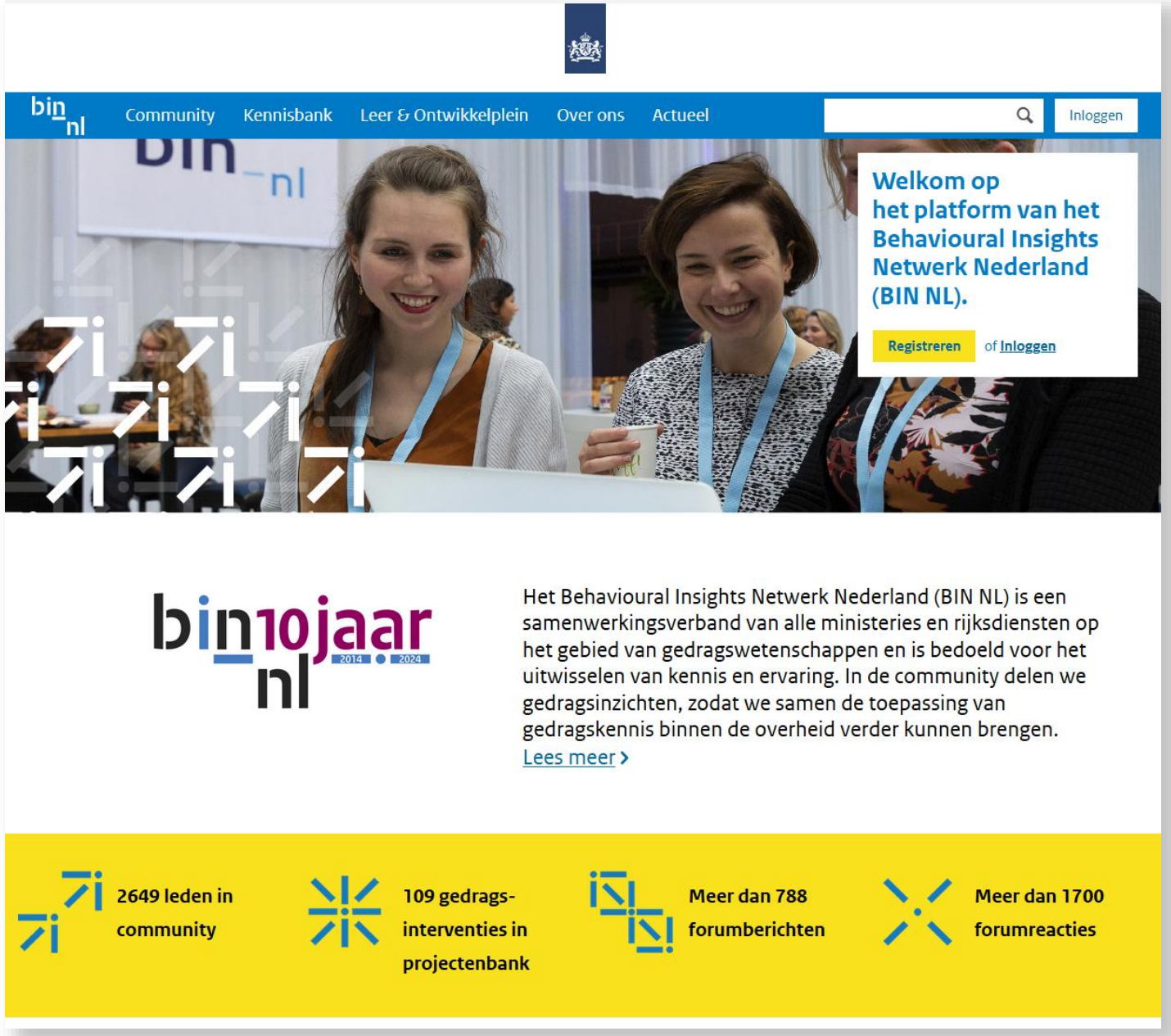



Photo by Jaredd Craig on Unsplash


bin nl
Community
Kennisbank
Leer & Ontwikkelplein
Over ons
Actueel
Inloggen
Welkom op het platform van het Behavioural Insights Netwerk Nederland (BIN NL).
Registreren
of Inloggen
bin 10 jaar nl
2014 2024
Het Behavioural Insights Netwerk Nederland (BIN NL) is een samenwerkingsverband van alle ministeries en rijksdiensten op het gebied van gedragswetenschappen en is bedoeld voor het uitwisselen van kennis en ervaring. In de community delen we gedragsinzichten, zodat we samen de toepassing van gedragskennis binnen de overheid verder kunnen brengen.
Lees meer
2649 leden in community
109 gedrags-interventies in projectenbank
Meer dan 788 forumberichten
Meer dan 1700 forumreacties

# Werk samen met wetenschappers!

- Uitvoerende taken leiden tot wetenschappelijke vragen
- Wetenschappers brengen innovatie in gegevensdeling, verzameling, bescherming
- Onderzoek heeft een speciale rol in gegevensverwerking (en leidt tot nieuwe inzichten!)

# Vragen?

e.vankesteren1@uu.nl

https://erikjanvankesteren.nl

https://fosstodon.org/@erikjan

SoDA

soda@odissei-data.nl

https://odissei-soda.nl