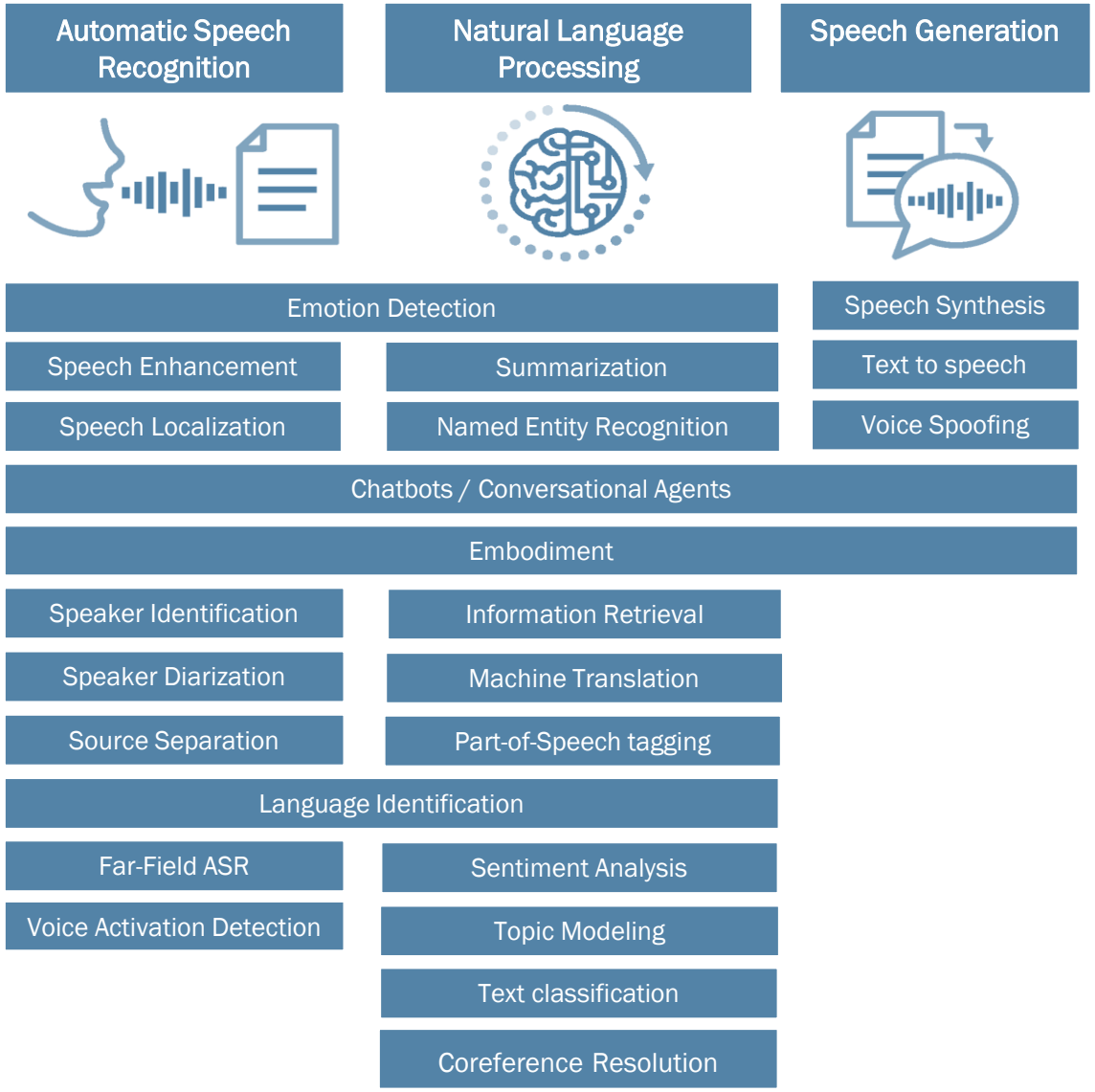

TNO innovation for life
HSD
securitydelta.nl
NL AI Coalitie
Zuid-Holland AI
LANDSCHAP NEDERLANDSTALIGE TAAL- EN SPRAAKTECHNOLOGIE
19 OKTOBER 2021

TAAL- EN SPRAAKTECHNOLOGIE



AANLEIDING 1/2
STARTNOTITIE NAIN

Versie: 0.9, 5 mei 2020

Auteurs
Erwin van der Eijk, NFI
Lisanne van Dijk, NFI
Frans Nauta, Data Science Initiative
Sander Ruiter, Nederlandse AI Coalitie

Een grote belemmering voor de benutting van AI in Nederland is dat bestaande algoritmen niet goed getraind zijn op de Nederlandse taal. […] Dit probleem […] geldt voor de gehele publieke sector en voor alle Nederlandstalige interacties in de markt.

Individuele organisaties ontwikkelen soms deeloplossingen voor specifieke domeinen, maar zonder een overkoepelend idee omdat daarvoor onvoldoende geld is. Om die reden ziet de werkgroep veiligheid het NAIN als een flagship voor de NLAIC.

AANLEIDING 2/2
STARTNOTITIE NAIN

Versie: 0.9, 5 mei 2020

Auteurs
Erwin van der Eijk, NFI
Lisanne van Dijk, NFI
Frans Nauta, Data Science Initiative
Sander Ruiter, Nederlandse AI Coalitie

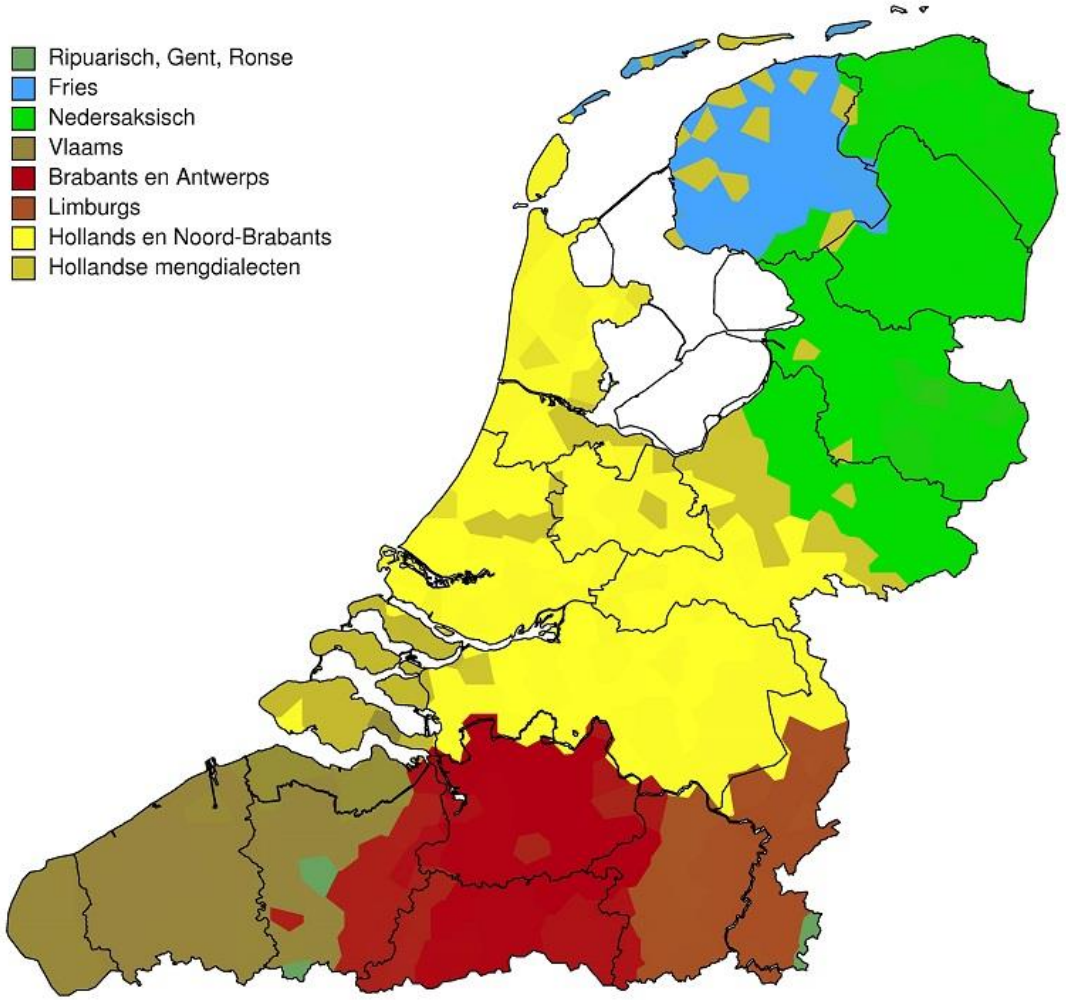
[...] de resultaten van dit project [zijn] overal in de Nederlandse samenleving bruikbaar, iedere dag, honderden miljoenen keren. Bij het registreren van zorghandeling zonder dat een zorgverlener haar handen van het bed hoeft te halen, bij het bellen van 112 om een ongeluk te melden, op de Twitter-feed van KLM om klanten snel en adequaat te woord te staan, voor het analyseren van terroristische dreigingen op een WhatsApp chat, automatische (en betrouwbare) ondertiteling van Nederlands beeld- en audiomateriaal, automatische transcripties van pathologisch onderzoek, van vergaderingen, enz.

[Dit project] maakt een enorme diversiteit aan toepassingen mogelijk [...], met grote publieke en economische waarde.

WAAROM NAIN?

- Matige prestatie TST op tal van Nederlandse dialecten, straattaal, kindertaal, spraakbeperkingen, andere niches, bestaande modellen van tech-giganten werken niet voldoende
- Beperkingen rond het delen van data met behoud van privacy en IP
- Europese/Nederlandse waarden rond bias, inclusie en uitlegbaarheid worden niet voldoende gewaarborgd in huidige oplossingen
- Soevereiniteit van Nederlandse taal en Taal en Spraaktechnologie, geen afhankelijkheid van buitenlandse multinationals
- Probleem en behoefte is te groot om door een/individuele partijen op te pakken, daarom is samenwerking gewenst

NEDERLANDSE DIALECTEN



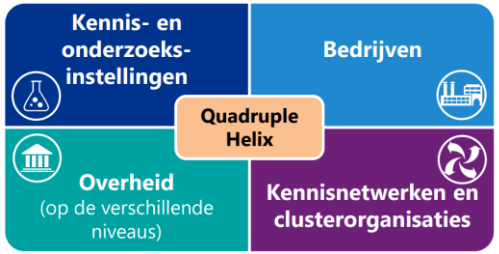
Kaart uit Dialectatlas: Peter Kleiweg en John Nerbonne

"Er zijn meer dan voldoende gemeenschappelijke elementen en belangen om verder samen te werken in de Lage Landen aan AI en taalgerelateerde thema's."

Programma sessies Vlaams-Nederlandse AI Workshop, EWI, EZK, OCW


NAIN
textgain
/instituut voor de Nederlandse taal/
AI Hub Zuid-Holland
nederlandse publieke omroep
Universiteit Leiden
UNIVERSITY OF TWENTE.
KATHOLIEKE UNIVERSITEIT LEUVEN
OPEN SPRAAK TECHNOLOGIE
NL AI Coalitie
taal: unie
TU Delft
LUMC Leids Universitair Medisch Centrum
Utrecht University
university of groningen
TNO innovation for life
Meertens Instituut
BEELD EN GELUID
CLST | Centre for Language and Speech Technology
Radboud University
GHENT UNIVERSITY
De Nederlandstalige Spraak Coalitie
POLITIE
VREDE EN RECHT
Den Haag
HSD
securitydelta.nl
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid
Ministerie van Justitie en Veiligheid
AVANTOPY
kpn
VNG
DEPARTEMENT ECONOMIE WETENSCHAP & INNOVATIE
Openbaar Ministerie
Ministerie van Sociale Zaken en Werkgelegenheid
CBS
Telecats
FIOD
Belastingdienst
Ministerie van Defensie
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties
UbiOps
shopping tomorrow

BETROKKEN PARTIJEN



Instituut voor Beeld en Geluid
NL spraak coalitie
NPO

Meertens Instituut

Amberscript B.V.

LUMC
Universiteit Leiden
INT

NLAIC
VNG
Taaluni

TNO
HSD

Politie
Openbaar Ministerie
KMAR
CBS
NFI
Gemeente Den Haag
Min Sociale Zaken/Opsporing SZW
Min Justitie en Veiligheid
Min Binnenlandse Zaken
Min Defensie

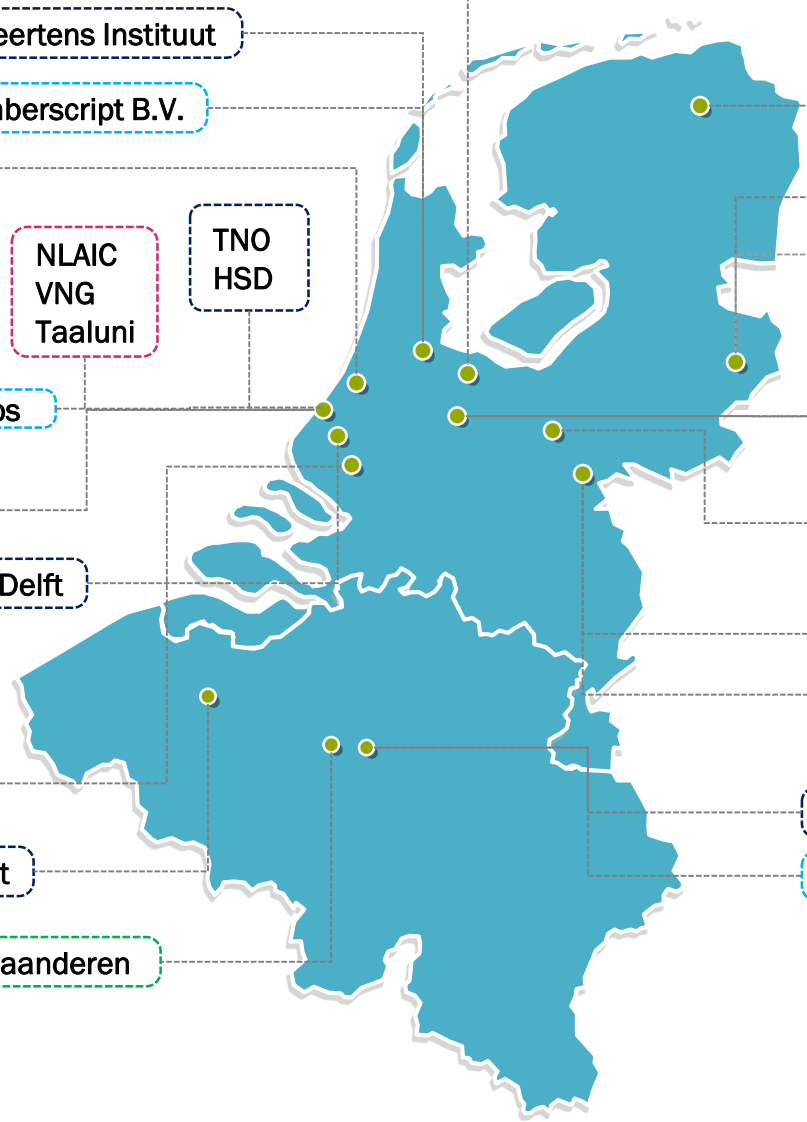
UbiOps

TU Delft

KPN

Universiteit Gent

EWI Vlaanderen



Universiteit Groningen

Universiteit Twente

Telecats

Universiteit Utrecht

FIOD

Thuiswinkel / Shopping Tomorrow

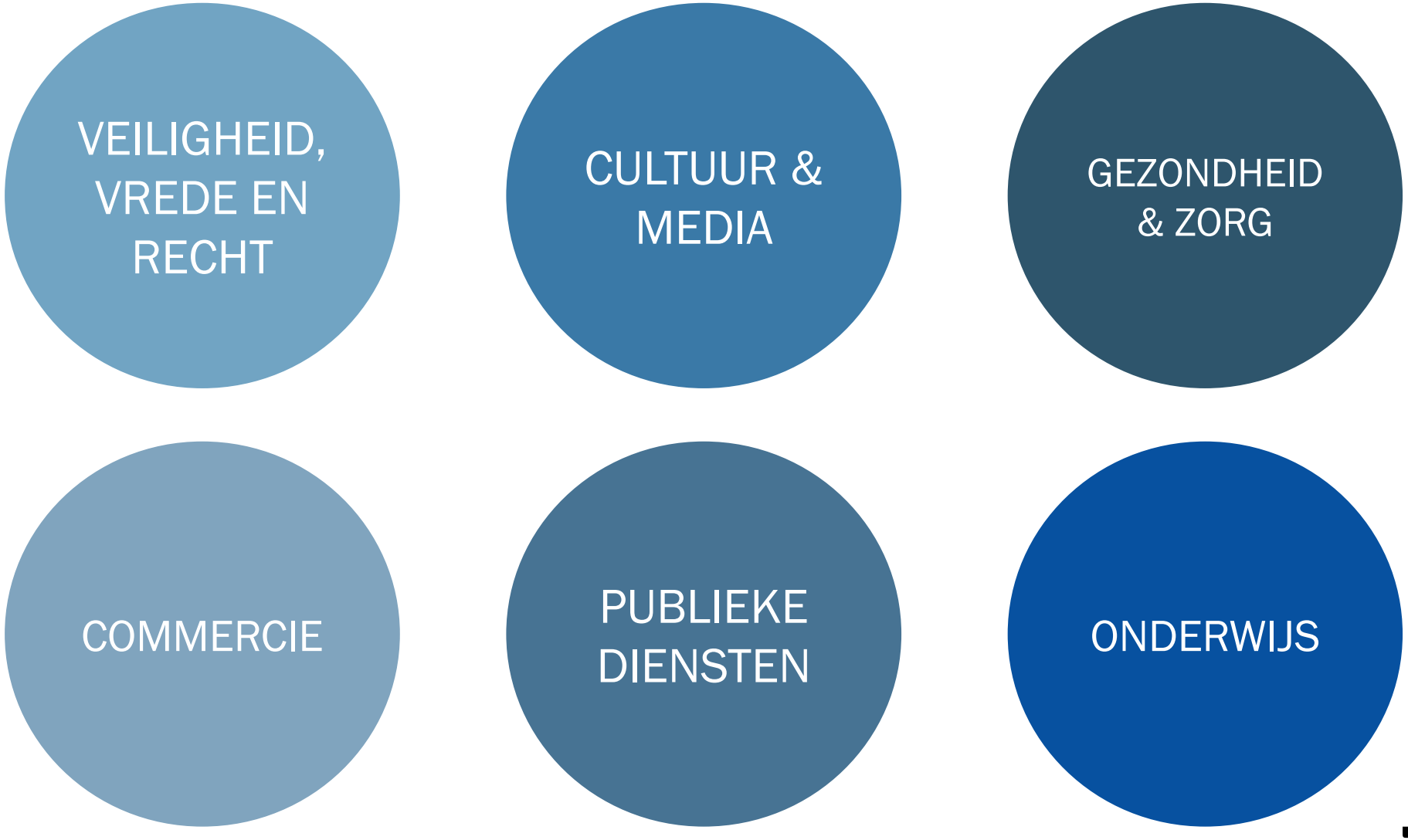
CLST Radboud Universiteit

Stichting Open Spraaktechnologie

Koninklijke Universiteit Leuven

Avantopy

TOEPASSINGS DOMEINEN



SAMENHANG ANDERE INITIATIEVEN

Onderzoeksprogramma's

2004-2010: STEVIN

- Nederlands-Vlaamse stimuleringsprogramma 11,4 miljoen voor de taal- en spraaktechnologie
- 11,4 miljoen euro
- Ontwikkeling corpora zoals het CGN en SoNaR
- Tooling zoals Spelspiek, Audiokrant, Kentekenlijn, Rechtsorde, Klinkende taal, AUTONOMATA, HATCI

VPVM

- Onderzoeksprogramma Vraaggestuurde Programma Veilige Maatschappij (VPVM)
- Doel: de samenleving rechtvaardiger en veiliger maken
- Taal- en spraaktechnologie één van de onderzoeksthema's van komende jaren

Tooling en infrastructuur

2012-nu: CLARIN ERIC

- CLARIN: digitale infrastructuur met data, tools en services om onderzoek gebaseerd op taaldata te ondersteunen
- CLARIAH: nationale tegenhanger CLARIN. Biedt een gedistribueerde onderzoeksinfrastructuur voor de geestes- en sociale wetenschappen

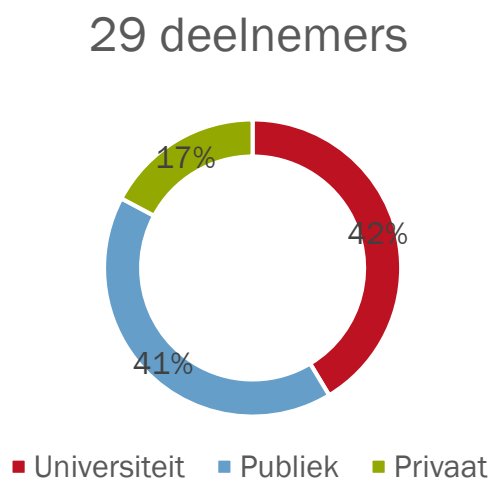
2014-nu: ELRC-SHARE repository

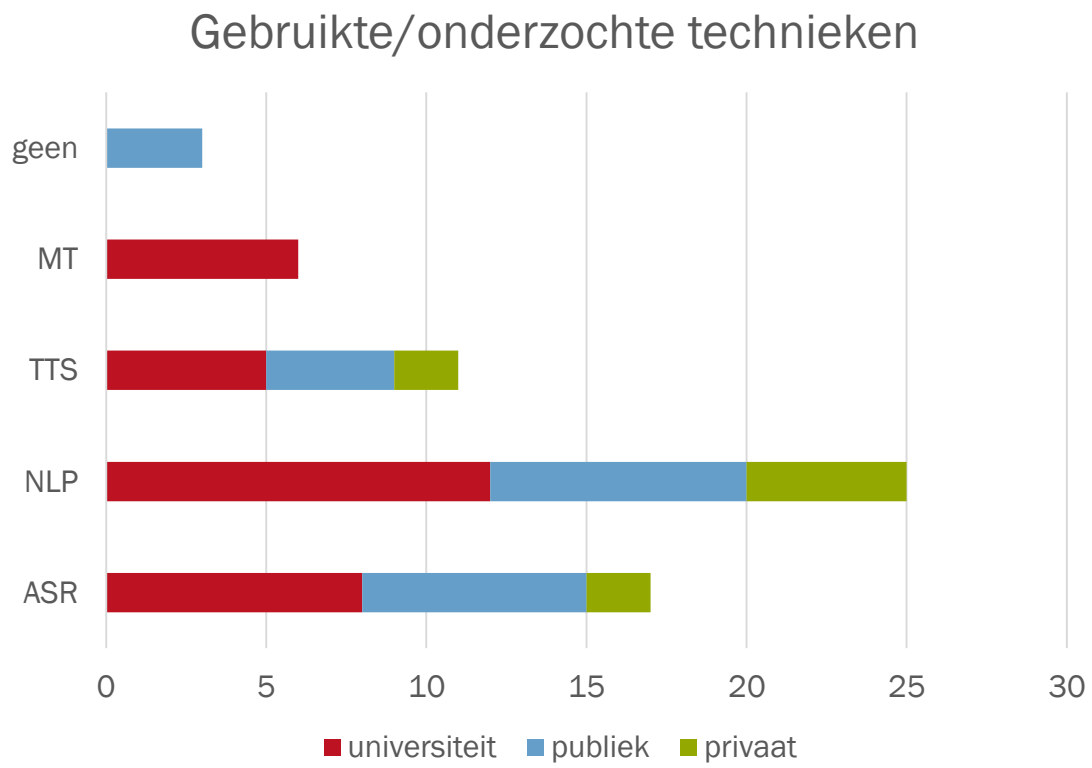
- overall goal ELRC: improve the quality, coverage and performance of the CEF Automated Translation platform in the context of current and future trans-European digital online public services;
- repository is used for documenting, storing, browsing and accessing Language Resources

2019-2021: European Language Grid

- Sharing platform for exchange of technologies, language data and resources, community
- https://www.european-language-grid.eu/

VRAGENLIJST NEDERLANDSE TST NEDERLAND EN VLAANDEREN



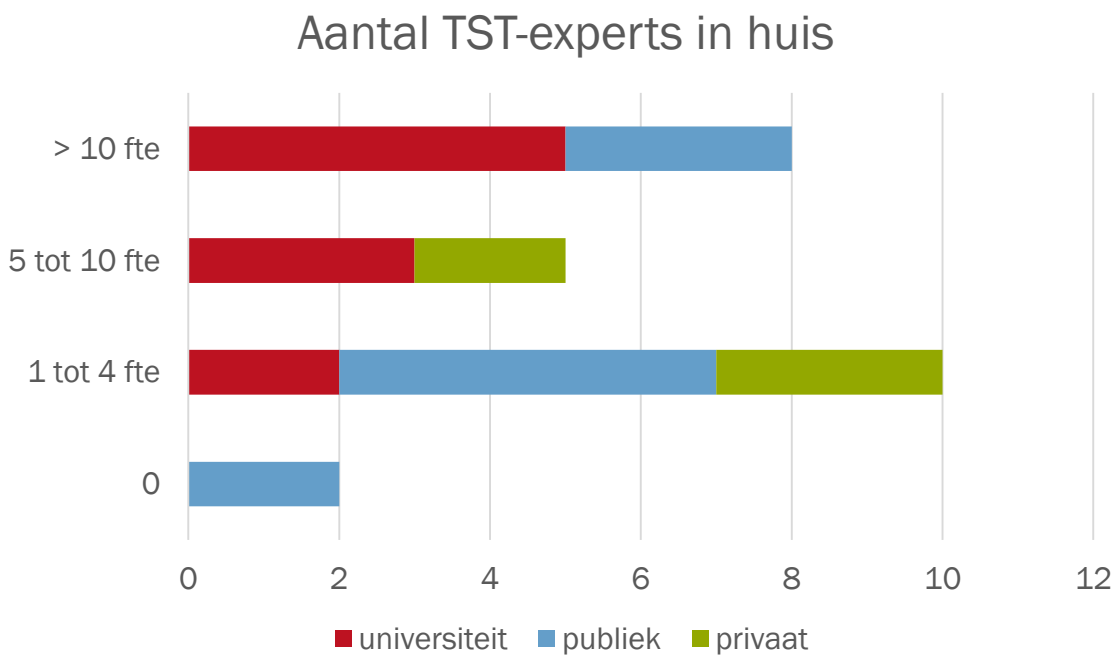


VERWACHTE GEBRUIKERS EN GROEI

- **Verwachte gebruikers**
  - Mogelijke afzetmarkt van ~200.000 personen
  - Intern honderden medewerkers
  - Grote groep potentiële gebruikers in de zorg
  - Miljoenen minuten spraak per jaar
  - over 10 jaar maken ~100.000 medewerkers gebruik van goed werkende TST, geen handmatige verslaglegging meer
  - Uiteindelijk heel Nederland

- **Verwachte groei**
  - Historisch zie ik een groei van rond de 15% per jaar wat we aan projecten extra per jaar doen, de afgelopen 5 jaar
  - Markt heeft potentie van miljarden per jaar in NL alleen al
  - Een heel grote groei, steeds meer bedrijven en de overheid willen NL taal- en spraaktechnologie gebruiken in hun organisaties
  - redelijke groei, maar het hangt ervan af wat er gaat gebeuren, wie er mee gaan doen, hoeverre de markt "rijp" is om door te zetten etc.
  - Zonder serieuze investering in onderzoek of gebruik van beschikbare databanken door lokale partijen, blijft NL TST afhankelijk van de markt die gedomineerd wordt door Big Tech.
  - NLP is nog net niet goed genoeg voor de dagelijkse praktijk. Met één goede verbeterslag kan dit omslagpunt echter wel bereikt worden, waardoor de groeipotentie enorm groot is.

BEMENSING



Zijn dit genoeg mensen voor de komende 5 jaar?

"Nee" (20x)

KNELPUNTEN

- Slechte performance bij
  - jargon
  - dialecten, accenten, straattaal, etc.
  - atypische spraak
  - slechte opname/geluidskwaliteit
- Van spraak > transcript gaat goed, maar van transcript naar informatie is een uitdaging
- Groot gebrek aan vrij beschikbare data
- Geen infrastructuur om gevoelige/enterprise data te delen

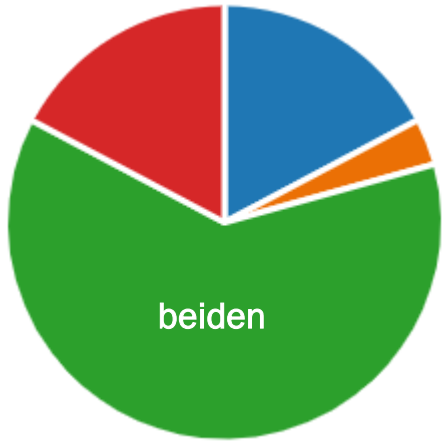
De meerderheid van de respondenten gaf aan deze uitdagingen niet zelf op te kunnen lossen

“[…] ik denk dat de Nederlandse markt te klein is om enkel op commerciële oplossingen te willen leunen. Een publiek/collectief initatief kan in potentie meer bereiken én breder inzetbaar zijn (ook voor kleinere partijen die geen commerciële tarieven kunnen betalen).”

VEEL DATA…. MAAR NIET ZOMAAR DEELBAAR

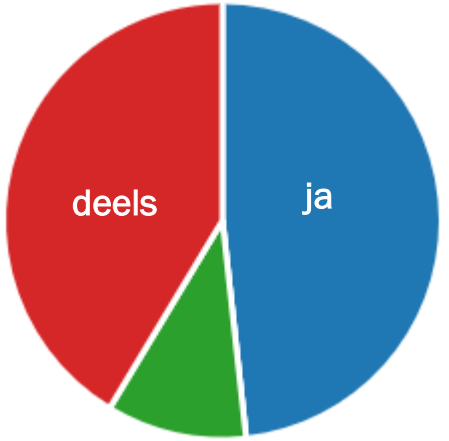
Welke data?

Tekstdata	5
Spraakdata	1
Beide	18
Other	5



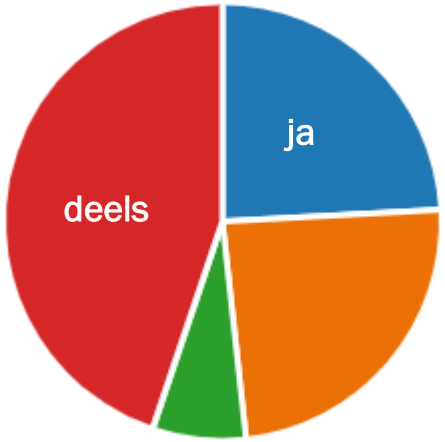
Privacy-gevoelig?

Ja	14
Nee	0
Dat weet ik niet.	3
Other	12



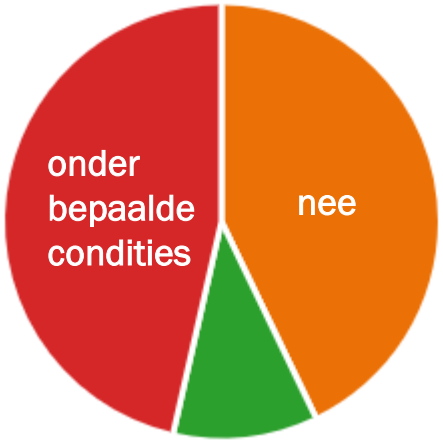
Gelabeld?

Ja	7
Nee	7
Dat weet ik niet.	2
Other	13



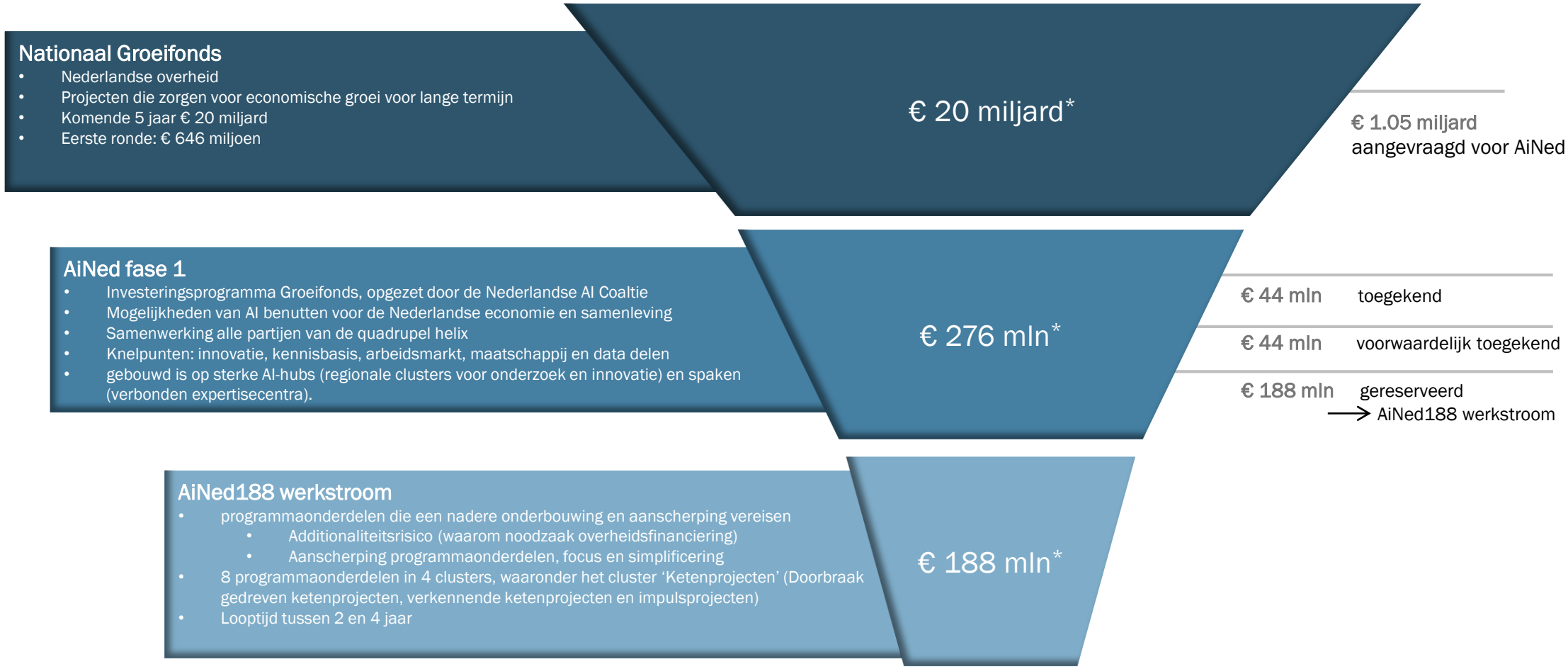
Mag het gedeeld worden?

Ja	0
Nee	12
Dat weet ik niet.	3
Other	13



GROEIFONDS, AINED, AINED188 WERKSTROOM

*obv 50-70% private matching/reguliere publieke instrumenten



TNO innovation for life

INTERREG 2021-2027

Interreg Europe
European Union | European Regional Development Fund

- Interreg programma Vlaanderen-Nederland subsidieert grensoverschrijdende projecten voor slimme, groene en inclusieve groei
- Projecten typisch 3-jarig, budget 3-5 miljoen

Doelen binnen BD Innovatie

- ontwikkelen en versterken van onderzoeks- en innovatiecapaciteit en invoering van geavanceerde technologieën
- ontwikkelen van vaardigheden voor slimme specialisatie, industriële overgang en ondernemerschap

Voorwaardes

- Partners uit grensregio
- Duidelijke maatschappelijk doel
- Grensoverschreidend belang

Tijdslijnen

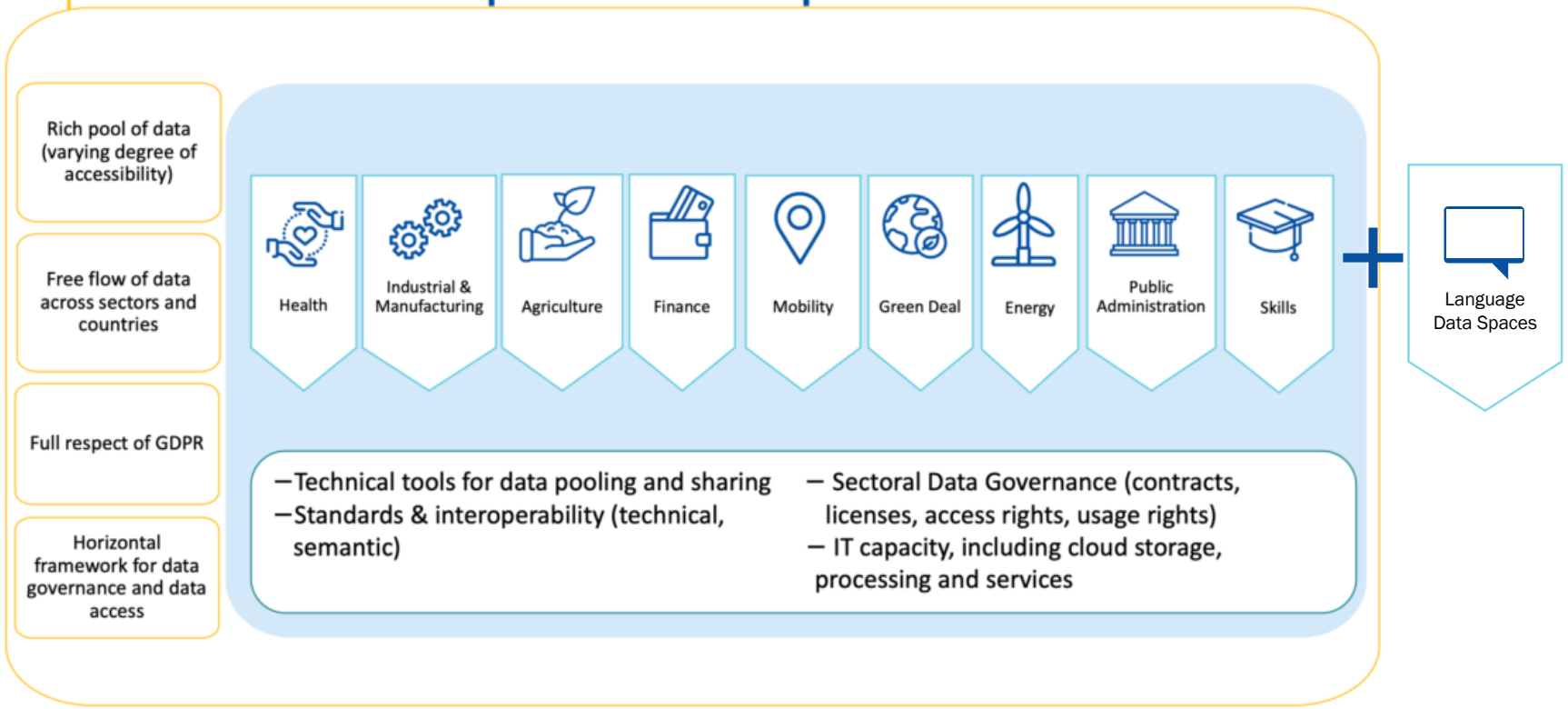
- november 2021: Deadline aanmelding project
- Februari 2022: Uitslag preselectie
- Juni 2022: Deadline projectaanvraag
- Startdatum mogelijk vanaf september 2022
- Najaar 2026: verwacht einde projecten

TNO innovation for life

DIGITAL EUROPE

DIGITALEUROPE

Common European data spaces



- Doel van de Europese data strategie: creeëren van een markt voor data
  - private en publieke partijen controleren gebruik van hun data
  - Zowel private als publieke sector toegang tot grote hoeveelheden data van hoge kwaliteit
- The European data strategy announced the development of 9 initial data spaces, indicating that the list is open, so other data spaces can be added.
- **Language Data Spaces:** The objective is to deploy a Language Data Space for the collection, creation, sharing and reuse of multimodal language data and to deploy large multimodal language models and a wide range of AI language technologies services to be offered through the AI platform.

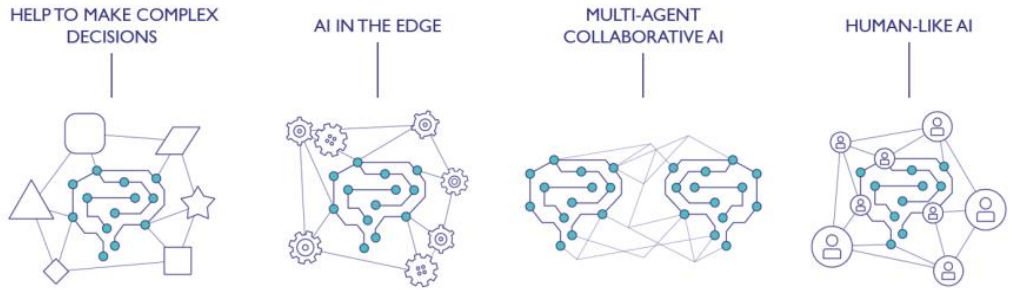
VLAAMS BELEIDSPLAN AI

- (1) **Top strategisch basisonderzoek** voor het gericht ontwikkelen van nieuwe kennis, wetenschappelijke doorbraken en talent op wereldniveau daar waar Vlaanderen reeds excellent presteert én waar synergie kan bekomen worden met de vraaggedreven implementatie-agenda van het Vlaamse bedrijfsleven.
- (2) **Een centrale focus op de implementatie van AI-toepassingen in het bedrijfsleven.** Een vraaggedreven agenda vanuit het bedrijfsleven moet via open, goed georganiseerde kanalen en netwerken gebracht worden tot bestaande overheidsinstrumenten van voornamelijk het Vlaams Agentschap Innoveren en Ondernemen (VLAIO) en relevante instellingen.
- (3) **Een sterk flankerend beleid** waarin naast de significante opleidingsnoden gericht op de arbeidsmarkt ook op het vlak van juridische, ethische, democratische en socio-economische aspecten van AI wordt gewerkt. En waarbij de focus ligt op een correcte doch ambitieuze outreach naar de bevolking zodat vernieuwende technologieën niet louter als exogene maar eerder als endogene, versterkende evoluties worden beschouwd, waaraan Vlaamse actoren actief kunnen meewerken.

Onderzoeksprogramma AI Vlaanderen

3 toepassingsdomeinen

4 uitdagingen



E-JUSTICE

ONDERDEEL VAN THE EUROPEAN COMMISSION JUSTICE PROGRAMME

European Commission

EUROPEAN e-JUSTICE

- Twee calls: eerste in Q1 2021 is al gepasseerd, tweede call verwacht in Q1 2022.
- 3 budgetlines (2022):

1. Judicial cooperation:
promote judicial cooperation in civil matters and contribute to the effective and coherent application and enforcement of EU instruments.

€7,45 miljoen

2. Judicial training:
promote judicial cooperation in criminal matters and contribute to the effective and coherent application of EU mutual recognition instruments in criminal matters.

€16,2 miljoen

3. Access to justice (e-Justice):
contribute to achieving the goals of the Commission Communication on the Digitalisation of Justice in the EU and the Council European e-Justice Strategy and Action Plan 2019-2023 by supporting the implementation of e-Justice projects at the European and national level

€8,5 miljoen totaal, €2,8 miljoen e-Justice

- ‘‘The Justice Programme will finance pilot activities and a study with respect to the use of artificial intelligence technology in the justice field. ’’
- ‘‘Priority will also be given to support for the development of concrete use cases based on artificial intelligence and distributed ledger technology in the justice area.’’

TNO innovation for life

IN HET KORT: AINED, INTERREG, D-E, VB AI, E-JUSTICE

	AINed	Interreg	Digital Europe	Vlaams Beleidsplan AI	e-Justice
Voor wie	• Publieke kennisinstellingen • Private kennisinstellingen • Bedrijven • Intermediaire organisaties (zoals innovatiecentra, ontwikkelingsmaatschappijen, etc.)	• Publieke kennisinstellingen • Private kennisinstellingen • Bedrijven, i.h.b. MKB/KMO • Intermediaire organisaties ***> iig partners uit Zuid-Nederland en Vlaanderen***	• large industrial entities • public stakeholders • small-and-mid-size enterprise stakeholders	• samenwerking tussen overheid, kennisinstellingen en bedrijven • open indiening via bestaande steunkanalen als met open calls/aanbestedingen • Voor het luik 'top strategisch basisonderzoek': programmatorische aanpak	• Publieke of private partijen, opgericht in een van de EU landen. • International Organisations. Met winst-oogmaerk in samenwerking met publieke partijen. • Private non-profit International Organisations.
Budget	Totaal: 188 miljoen Typisch project: ?	Totaal: 60-65 miljoen Typisch project: 3-5 miljoen	Project: 6 miljoen	Jaarlijks: 32 miljoen	e-Justice call: 2,8 miljoen
Cofinanciering	50%	50% obv declaraties	50%	Hefboomwerking andere bronnen: verzekerde cofinanciering KPI	Max. 90%
Andere voorwaarden	• Ketenproject • Sectoroverschreidend	• Partners uit Zuid-Nederland en Vlaanderen • Maatschappelijk(e) doel(en) • Grensoverschrijdend belang	• Procurement		• EU Grant aanvraag van minimaal €75,000 • Voorkeur projecten met Europese samenwerking
Deadlines	• **21 augustus 2021** aanmelding project • ?	• **21 november 2021** aanmelding project • **Februari 2022:** uitslag preselectie • **Juni 2022:** Deadline projectaanvraag	• Call 1: **Q4 2021** • Call 2: **Q3 2022** • Call 3: **Q4 2022**		• ~~Call 1: Q1 2021~~ • Call 2: **Q1 2022**
Startdatum	?	vanaf september 2022	Q2 2022, Q1 2023, Q2 2023		Vanaf Q1 2021, nieuwe call in Q1 2022

KERNTHEMA'S KOMENDE 5 JAAR

Missie: *Ontwikkeling van state-of-the-art soevereine Nederlandstalige taal- en spraaktechnologie, die inclusief, divers, (transparant,) en uitlegbaar is, en waar domeinspecifieke extenties aan gekoppeld kunnen worden*

- Algoritmiek waarbij onderzoek naar inclusieve algoritmes, nieuwe architecturen, en (on)mogelijkheden van kleine datasets voorop staat

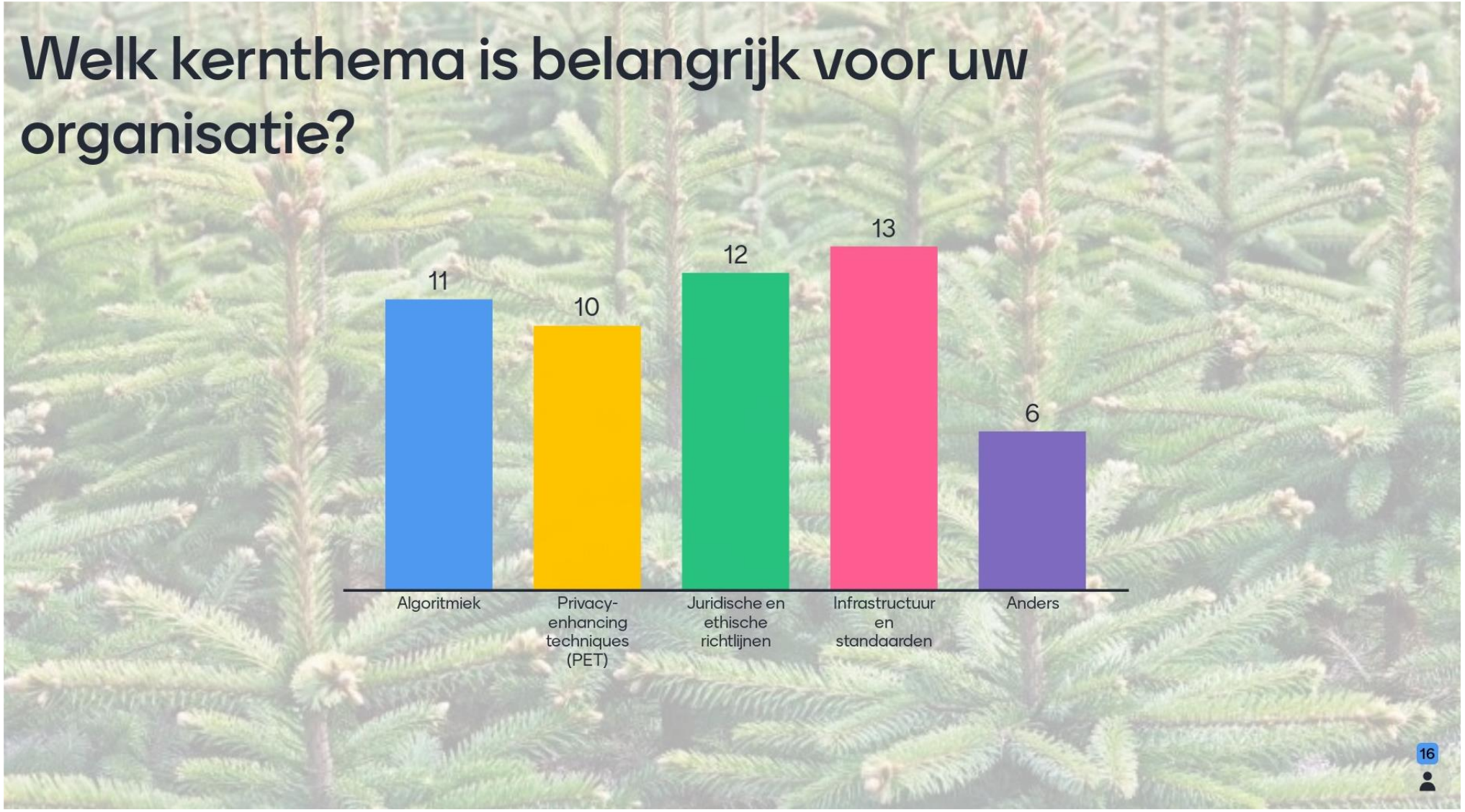
- Juridische en ethische richtlijnen voor het gebruik, ontwikkelen en delen van TST, rekening houdende met Europese AI regelgeving

- Privacy-enhancing techniques (PET) om spraak- en tekstdata te kunnen delen voor trainingsdoeleinden

- Infrastructuur en standaarden om data te verzamelen, data, modellen, en algoritmiek uit te kunnen wisselen en beschikbaar te maken

- Embedding en organisatie om maatschappelijke toepassingen te realiseren, bestuurlijk draagvlak creëren, acceptatie van boven én onderaf, disseminatie over successen

UITKOMSTEN MENTI DEELNEMERS SESSIE 27-9-2021



USE CASE IDEEËN

Transcriptie

Spraak naar tekst

Transcriptie (S2T)

Goede transcripties

Kant en klaar ASR model/service

Verslaglegging, transcriptie

Auto-transcript, auto verslag (samenvatting), auto beslissingen Autoclassificatie

Verhoor, bevindingen, notulen, tapgesprekken

Meenemen ongestructureerde teksten in medische besluitvorming

Anonimiseren van tekst- en spraakdata

Taalherkenning

Speaker labeling nauwkeuriger

Herken van laaggeletterden in digitale domein

Spreker identificatie

Functionaliteiten

Spraak dienstverlening van de ns/gemeente die werkt

Spraak en tekst coach/buddy tbv preventie bij kwetsbare groepen

Informatieontsluiting & service

Betere (data)journalistiek door gebruik meerdere databases op basis van taal

Conversatie in call center in real time analyseren om de agent optimaal ondersteunen in het leiden naar en goede resultaat voor de klant

Meten van nieuwe vormen van criminaliteit obv processen verbaal

Vonnissen via rechtspraak.nl openbaar beschikbaar maken

Assistant interoperability

Feitelijke vragen kunnen stellen (“question answering”) over een hele grote bak (niet netjes geschreven) tekst

Bewaakt inclusie, bias, en privacy, en promoot gebruik van gedeelde modellen en data

Effectiviteit en bereik gesproken beleidstaal

Tools om de data die we hebben te kunnen inzetten voor trainings doeleinden

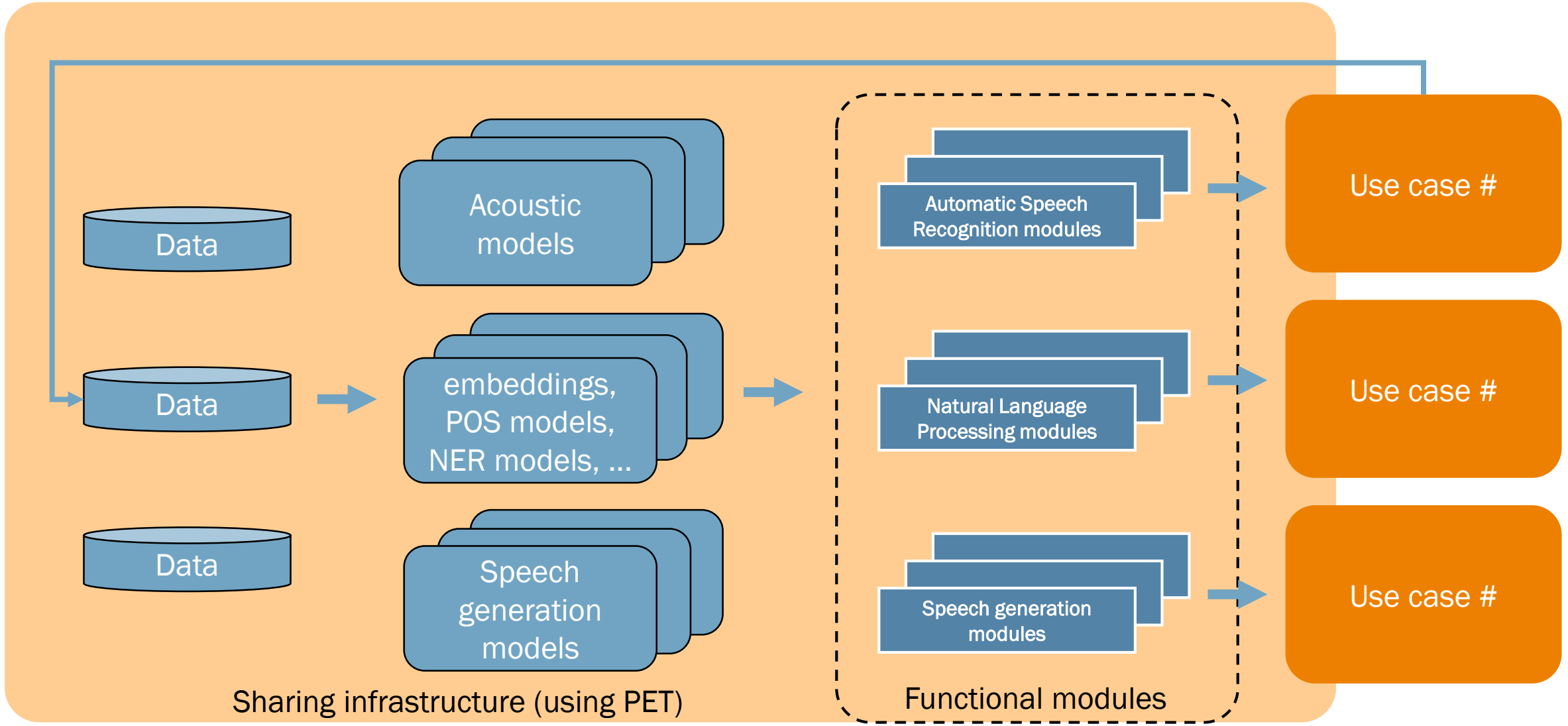
Infrastructuur om algemene language model te bouwen obv van datasets van meerdere bedrijven en instellingen (met PET). Dan infrastructuur om Domain specific fine-tuning te doen

Algemene NL modellen voor transcriptie radio/tv, (gesproken) ondertiteling, metadata extractie, stemgebaseerd zoeken & navigeren

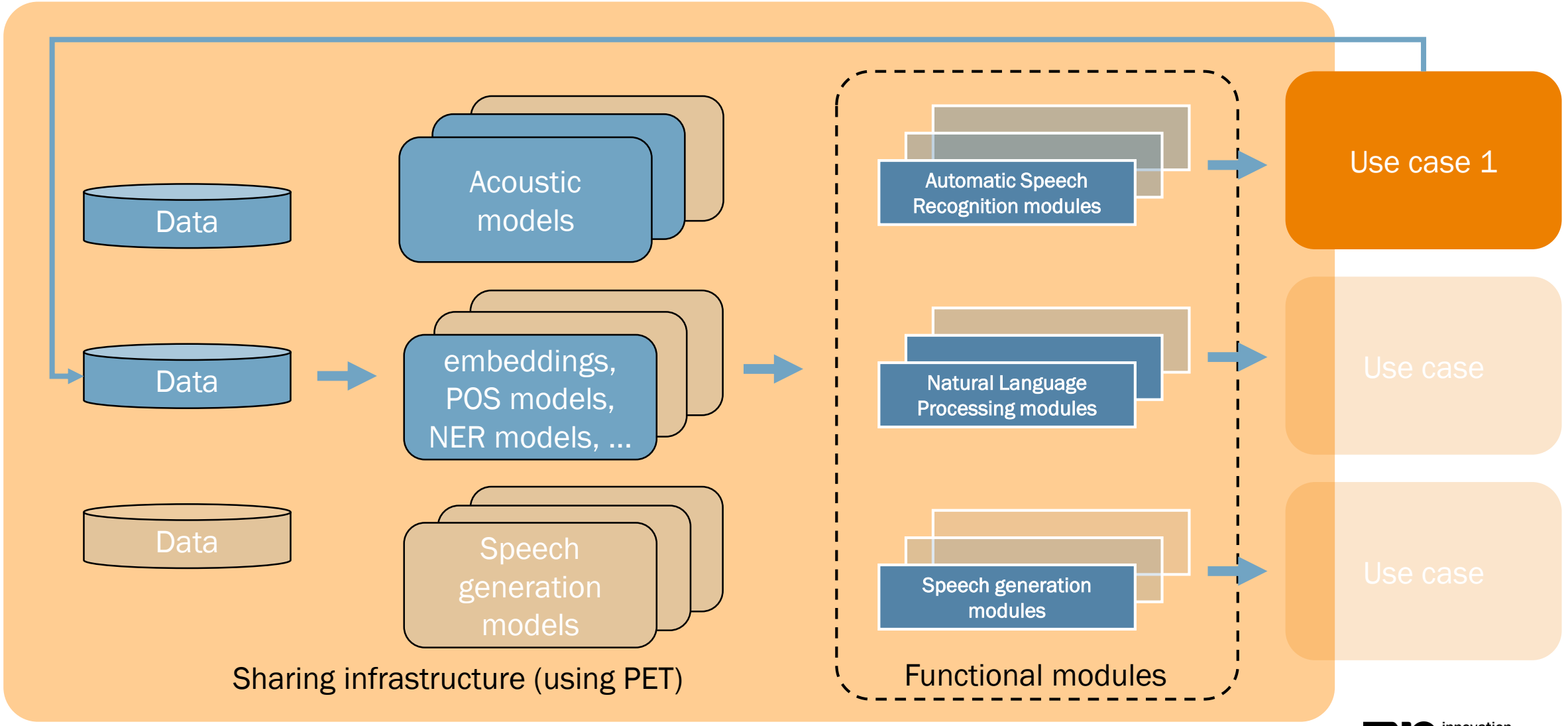
Modellen voor niet-standaard ABN (kinderen, ouderen, behinderden)

Modellen, tools & infrastructuur

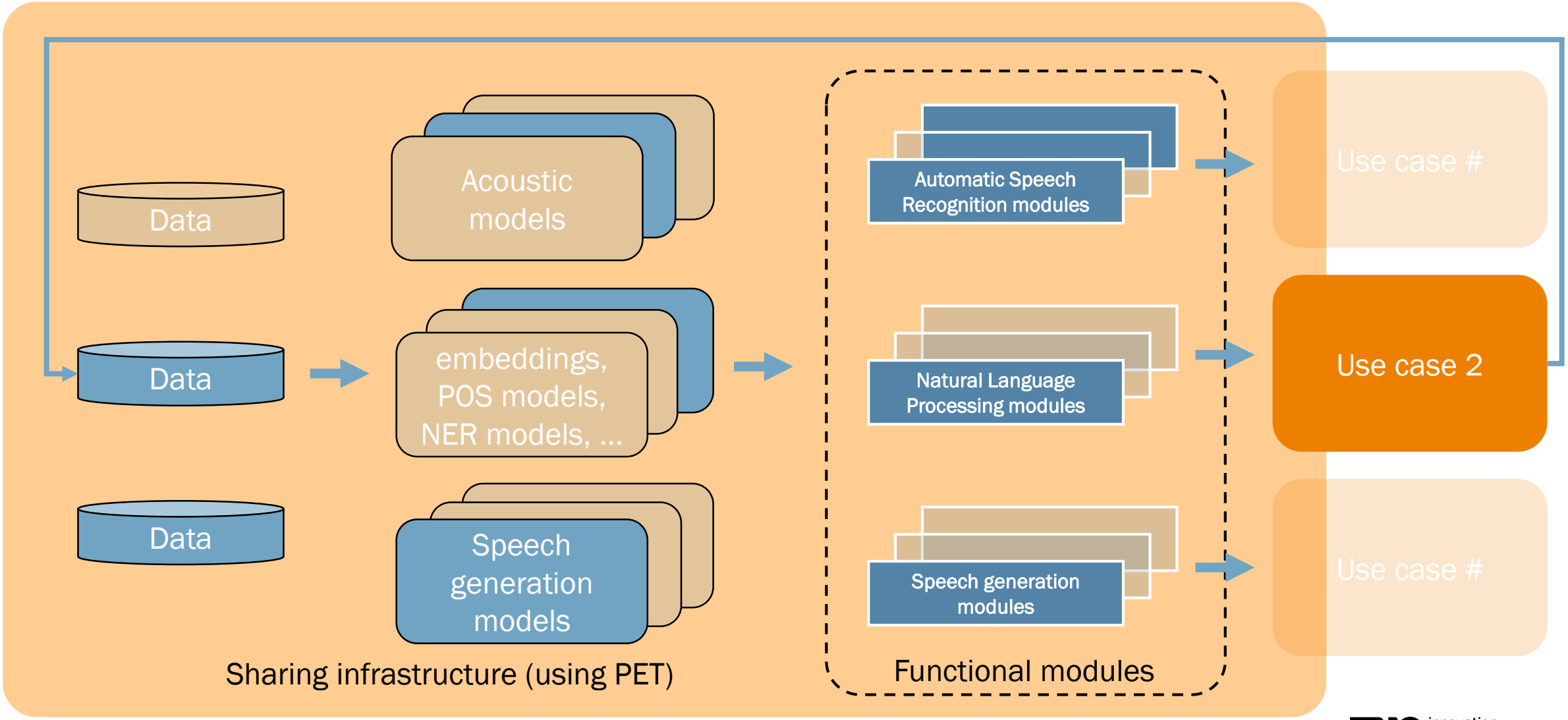
NAIN DRAFT ARCHITECTUUR



NAIN DRAFT ARCHITECTUUR, USE CASE 1



NAIN DRAFT ARCHITECTUUR, USE CASE 2, ETC



NAIN KERNTEAMS EN CONSORTIUM

- Kernteam AiNED
  - Lead: TNO (Saskia Lensink en Joachim de Greeff)
  - Doel: AiNed188 ketenproject voorstel schrijven

- Kernteam Interreg
  - Lead: LUMC (Marco Spruit)
  - Doel: Interreg voorstel schrijven

- Brede NAIN-consortium
  - Lead: HSD (Marlou Snelders en Rosa Edema)
  - Doel: betrokken blijven en informatie delen

Initiatiefnemers

HSD securitydelta.nl

TNO innovation for life

Mede mogelijk gemaakt door Zuid-Holland AI

Zuid-Holland AI

Eindredactie:
Saskia Lensink, Joachim de Greeff, Rosa Edema en Marlou Snelders

Contact NAIN Consortium
Email: marlou.snelders@nlaic.nl

Contact AiNed
E-mail: saskia.lensink@tno.nl / joachim.degreeff@tno.nl

Oktober 2021