

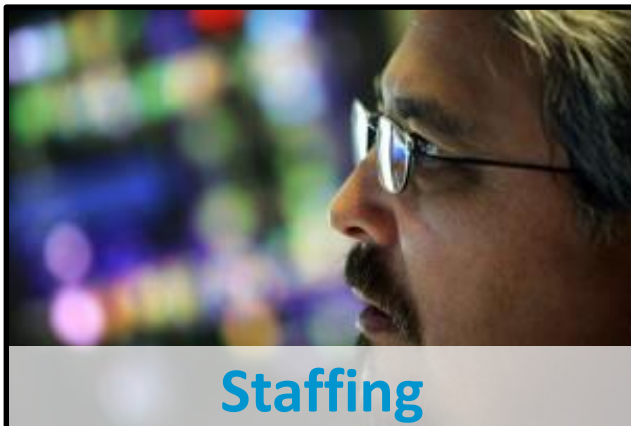
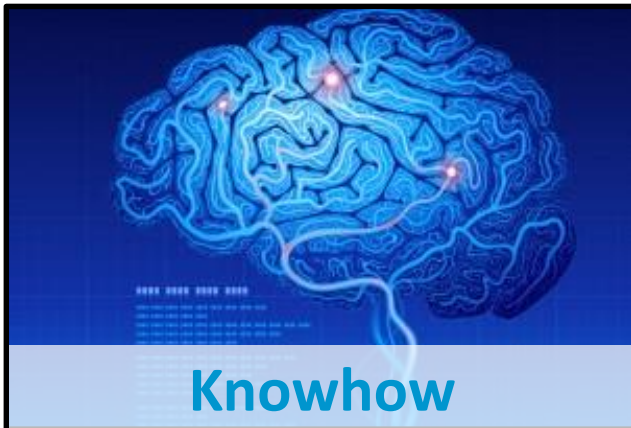
SAI webinar: Natural Language Generation

Joachim Ganseman
Smals Research

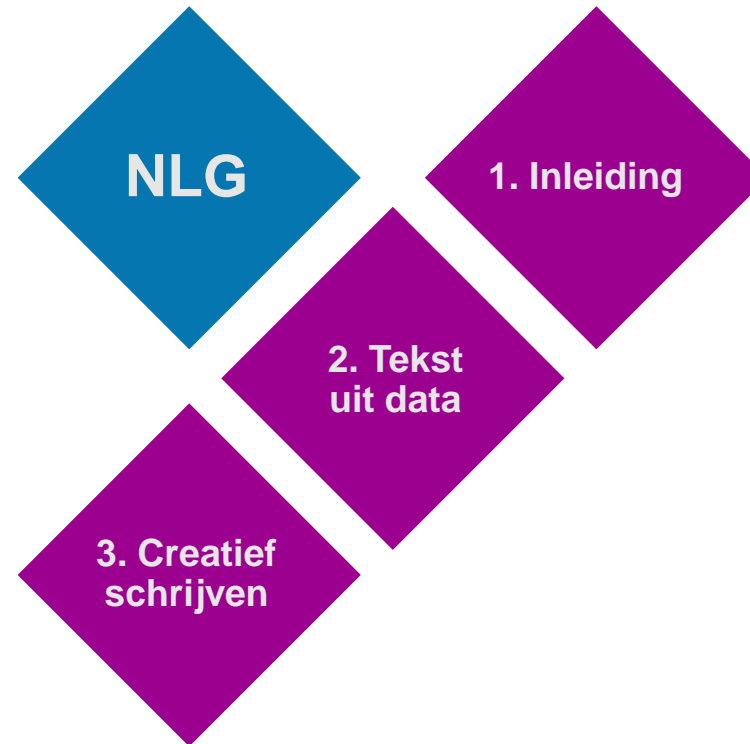
17/02/2022



SUPPORT FOR E-GOVERNMENT



WWW.SMALS.BE

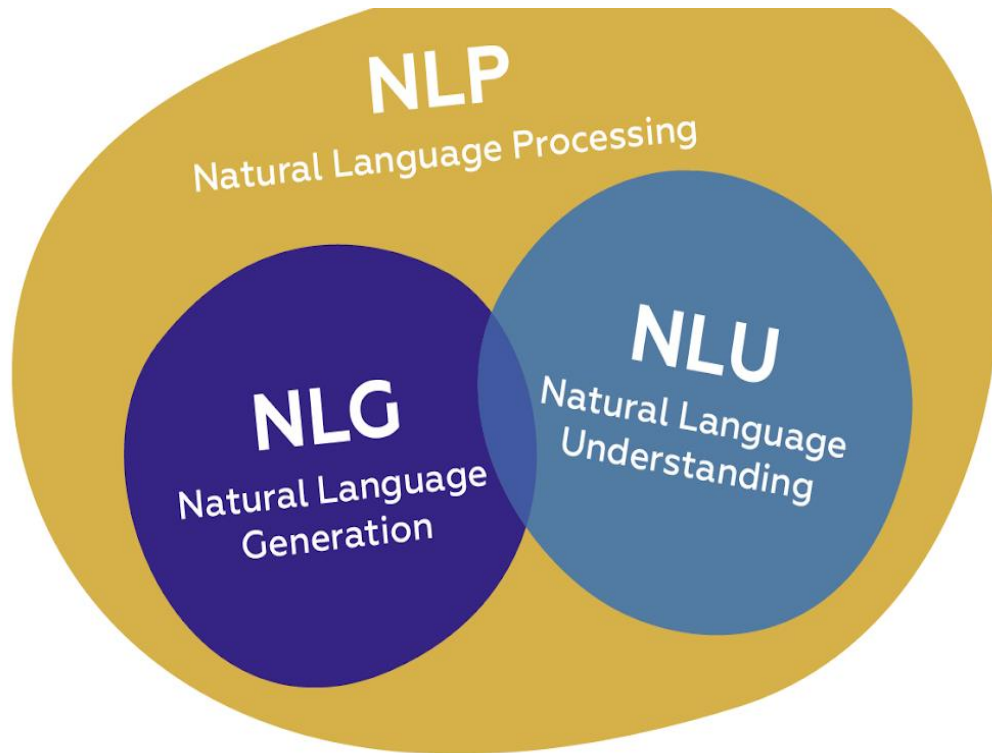




Inleiding

Natural Language Generation

“Natural”: mensentaal, grammaticaal correct + inhoudelijk betekenisvol



Practical Applications of NLG



Analysis for business intelligence dashboards



Reporting on business data/data analysis



IoT device status and maintenance reporting



Individual client financial portfolio summaries and updates



Personalized customer communications

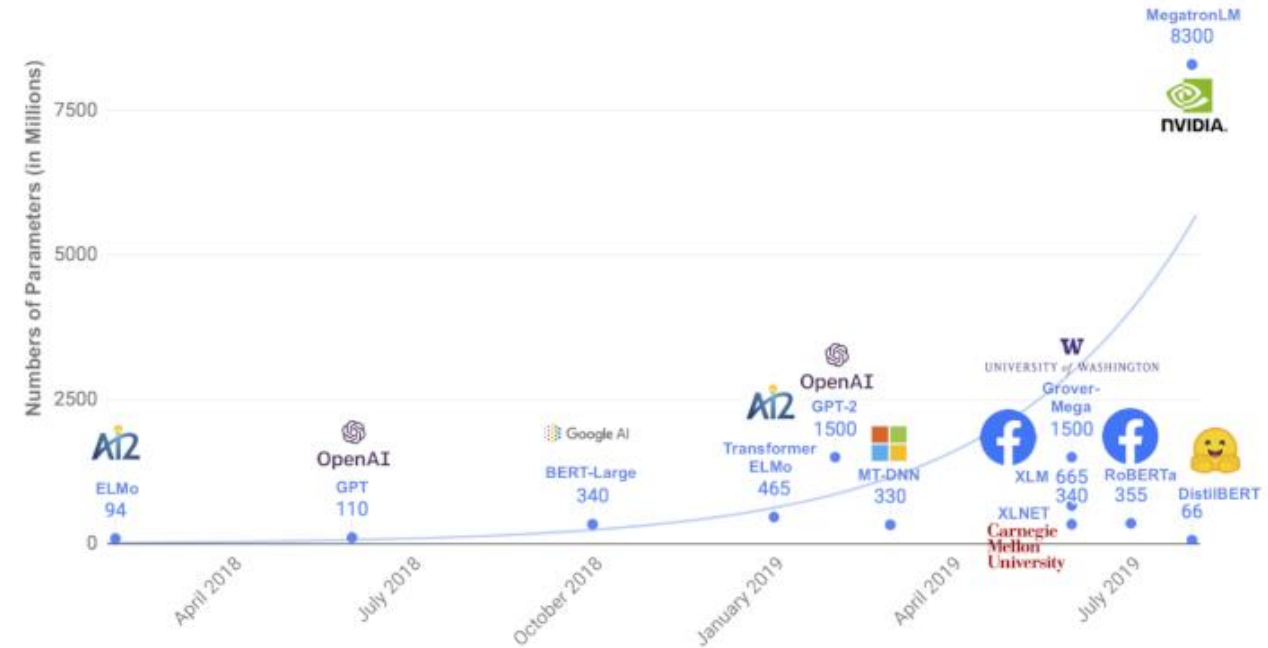
De hype

- Sinds 2018, steeds grotere generatieve taalmodellen

- Juli 2020 : OpenAI GPT-3 in beta
 - 175 miljard parameters
 - \$ 4600000 trainingskost (schatting)

- Initialiseer met “prompt”,
GPT-3 vult aan

- Spectaculaire demo's ([AI Dungeon](#) / [\[VID1\]](#) / [\[VID2\]](#) / [tekstueel](#))



De hype

Chat
Open ended conversation with an AI assistant.

Grammar correction
This zero-shot prompt corrects sentences into standard English.

Text to command
This prompt translates text into programmatic commands.

Parse unstructured data
Create tables from long form text by specifying a structure and supplying some examples.

Movie to Emoji
Convert movie titles into emoji.

JavaScript helper chatbot
This is a message-style chatbot that can answer questions about using JavaScript. It uses a few examples to get the...

Science fiction book list maker
This makes a list of science fiction books and stops when it reaches #10.

Airport code extractor
A simple prompt for extracting airport codes from text.

ML/AI language model tutor
This is a QA-style chatbot that answers questions about language models.

Tweet classifier
This is a basic prompt for detecting sentiment.

SQL request
Create simple SQL requests.

JavaScript to Python
Convert simple JavaScript expressions into Python.

Mood to color
Turn a text description into a color.

JavaScript one line function
Turn a JavaScript function into a one liner.

Third-person converter
Converts first-person POV to the third-person. This is modified from a community prompt to use fewer examples.

VR fitness idea generator
This prompt creates ideas for fitness and virtual reality games.

OpenAI
CEO

Sam Altman @sama

The GPT-3 hype is way too much. It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out.

8:45 PM · Jul 19, 2020

7K 144 Share this Tweet

Tweet your reply

- GPT-3 geeft output
 - die statistisch waarschijnlijk is
 - die min of meer past in de context
 - op basis van wat het gezien heeft in trainingsdata (Wikipedia = slechts 0,6%)

... maar is nog steeds **niet intelligent** :

```
Q: How many eyes does a giraffe have?  
A: A giraffe has two eyes.
```

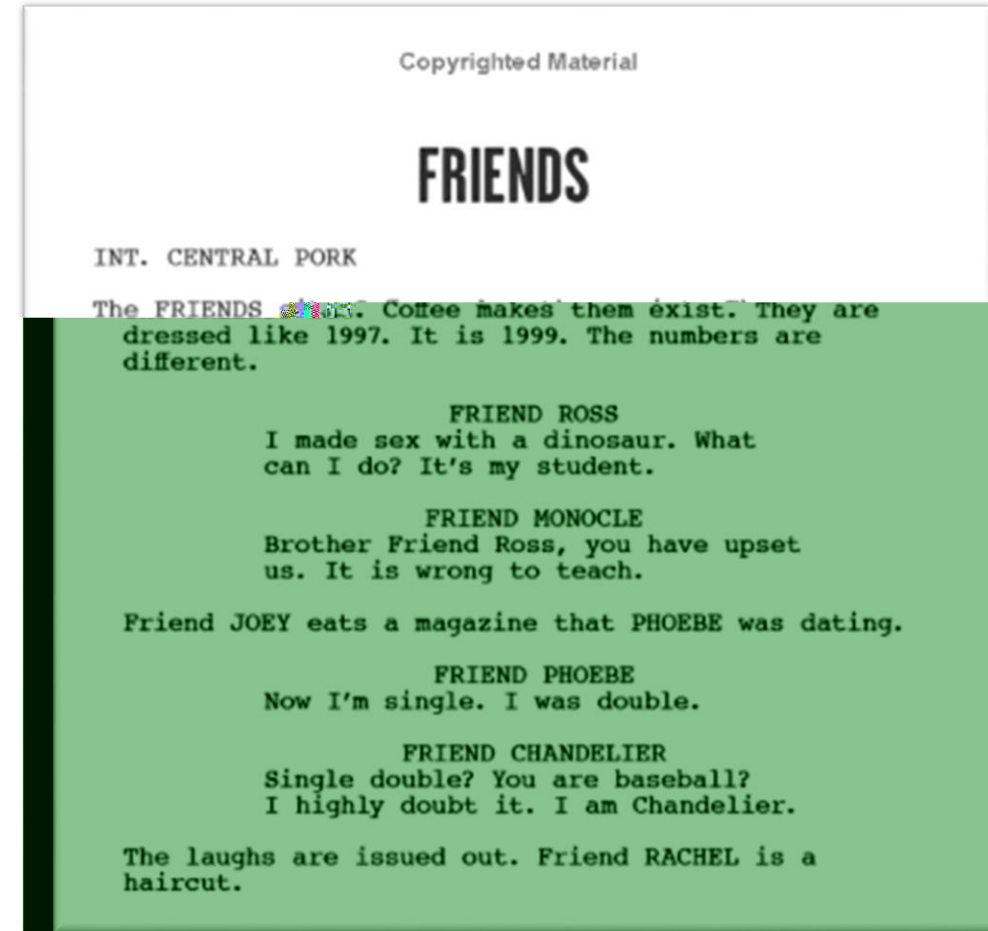
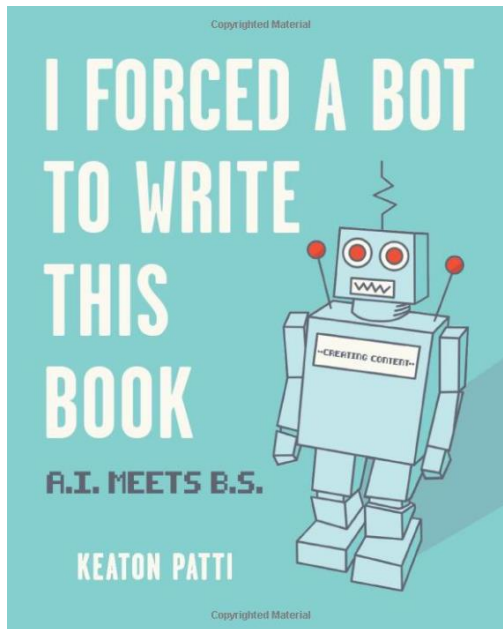
```
Q: How many eyes does my foot have?  
A: Your foot has two eyes.
```

```
Q: How many eyes does a spider have?  
A: A spider has eight eyes.
```

```
Q: How many eyes does the sun have?  
A: The sun has one eye.
```

```
ave?  
Q: How many eyes does a blade of grass.  
A: A blade of grass has one eye...
```

- Wie niet de resources heeft van GAFAM, OpenAI e.d., haalt meestal ook niet dezelfde resultaten



Bron: [Keaton Patti, Twitter, 04/10/2018](#)

- “People are showing the results that work and ignoring those that don’t.”
([The Verge](#))



- Overzichtsartikels
 - [“Building Applied Natural Language Generation Systems”](#) [Reiter & Dale 1997]
 - [“Survey of the state of the art in natural language generation: core tasks, applications and evaluation”](#) [Gatt & Kramer 2018]
- Procedurele methodes
 - Substitutie van tekst in templates
 - Vb: *Mail-merge* (NL: *afdruk samenvoegen*) in MS Word
 - “Rule-based”, gestructureerd op basis van *context-vrije grammatica* (ref. Chomsky)
 - Chatbots van de Jaren ‘70 ([Eliza](#)) tot nu
 - Parody generators ([wetenschappelijke papers](#), [new age bullshit](#), [postmodernism](#), ...)
- Statistische methodes
 - Markov Chains
 - Context beperkt tot enkele woorden
 - Vb: Autocomplete op veel GSM keyboard apps

- Een zelfgemaakte “generatieve grammatica” met allerlei opties:

```
592 PROGRAMMING_VERBING    architecting
593 PROGRAMMING_VERBING    implementing
594 PROGRAMMING_VERBING    optimizing
595
596 IMPL_INTRO after several TRACE_UNITS of DIFFICULT_X PROGRAMMING_VERBING, we finally have a working implementation of SCI_OUR_SYSTEM XXX
597 IMPL_INTRO our implementation of SCI_OUR_SYSTEM is SCI_BUZZWORD_ADJ, SCI_BUZZWORD_ADJ, and SCI_BUZZWORD_ADJ XXX
598 IMPL_INTRO though many skeptics said it couldn't be done (most notably SCI_BY_WHO), we SCI_PRESENT_VERB a fully-working version of SCI_OUR_SYSTEM XXX
```

... wordt “ingevuld” om een tekst te bekomen:

Router: A Methodology for the Typical Unification
of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

III. IMPLEMENTATION

Our implementation of our approach is low-energy, Bayesian, and introspective. Further, the 91 C files contains about 8969 lines of SmallTalk. Router requires root access in order to locate mobile communication. Despite the fact that we have not yet optimized for complexity, this should be simple once we finish designing the server daemon. Overall,



**Tekst uit data:
templating
engines**

Korte beschrijving


Soort accommodatie	Vakantiehuis
Overig	Niet-roken huis
Woonkamer	50 m ²
Minimale bezetting	3 personen
Maximale bezetting	3 personen
Huisdier toegestaan	Huisdieren niet toegestaan

1	Aantal slaapkamers
1	Aantal badkamers
Terras, Tuin	Buitenfaciliteiten

Vaatwasser, Oven



- Tabulaire data omzetten naar tekst
- Booking.com, bol.com, CoolBlue, ...
- Weerberichten
- Statusrapportering (sensoren/systemen)
- Tickers financiële markten
- ...

Accommodatie beschrijving 

Beoordeling

Kaart en omgeving

Prijzen

Beschikbaarheid

De hoogtepunten van het vakantiehuis

“ Een mooie vakantie in Amsterdam zal met dit mooie vakantiehuis met WiFi zeker lukken. Aan maximaal 3 personen wordt hier plaats geboden: 1 slaapkamer en 1 badkamer zijn verdeeld over 50 m² woonoppervlak. In het gehele huis is roken niet toegestaan.

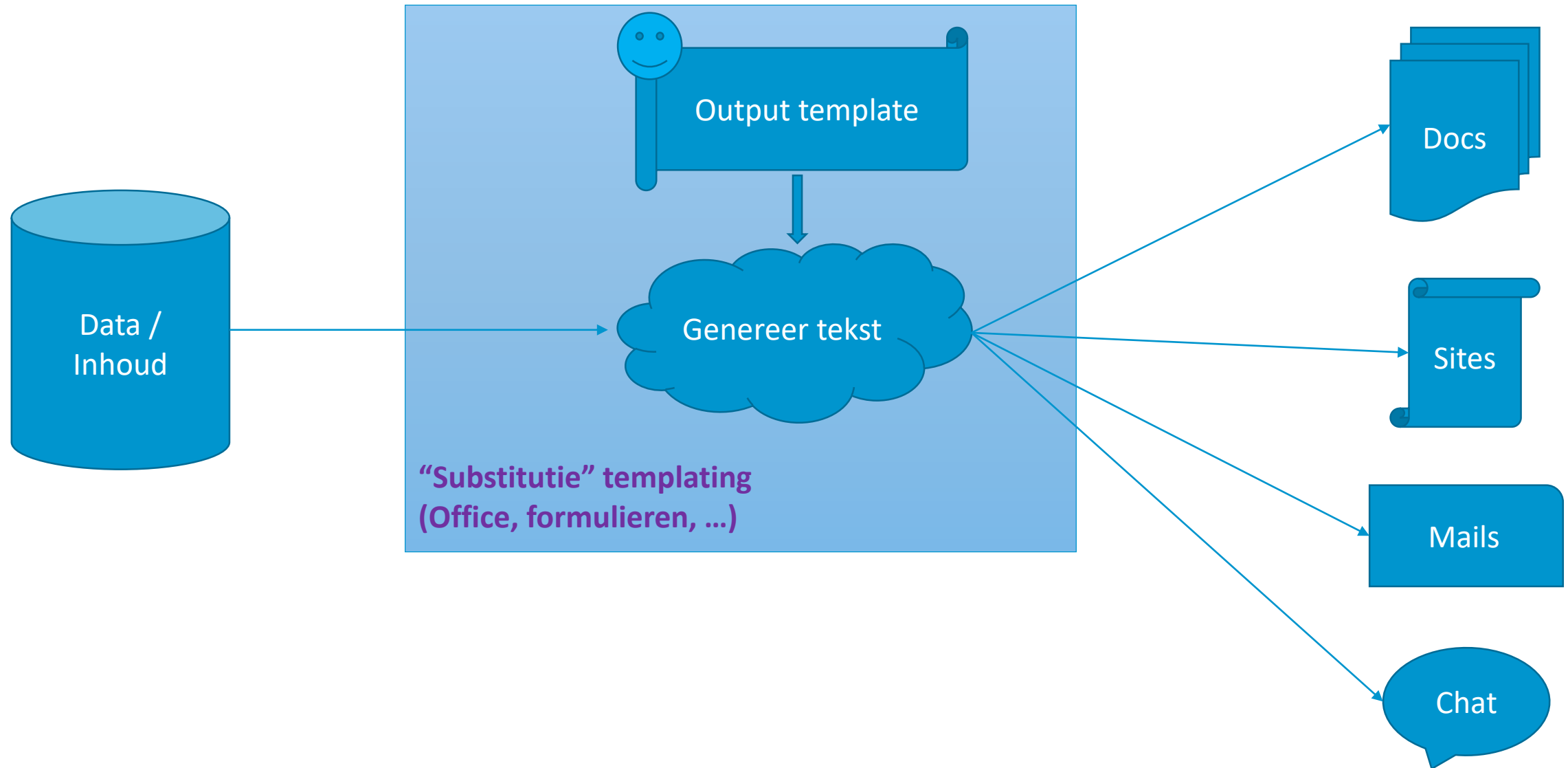
- ✓ Een vakantiewoning, die van klanten over alle facetten goede beoordelingen krijgt.
- ✓ Op aangename afstand van het meer: slechts 200 m (hemelsbreed).
- ✓ U bent snel verbonden met de wereld dankzij WiFi.
- ✓ De accommodatie beschikt over een terras.

The screenshot displays a user interface for 'digitalSTROM'. At the top left is the logo, and at the top right is a 'Logout' link. The main heading is 'I'm fine Report'. Below this is a green banner featuring a white house icon with a smiling face and a speech bubble that says 'I'm fine.'. A navigation bar below the banner contains four items: 'Allgemein' (with a house icon), 'Temperatur & Heizung' (with a thermometer icon), 'Sicherheit & Schutz' (with a shield icon), and 'Energie' (with a lightning bolt icon). The 'Temperatur & Heizung' section is expanded, showing the title 'Temperatur & Heizung' and a text report: 'Die Temperatur im Wohnzimmer beträgt aktuell etwa 27.0 °C. Mit 23.3 °C um 19 Uhr war Arbeitszimmer der kälteste Raum. Draußen waren es innerhalb der letzten 24 Stunden durchschnittlich 16.9 °C. Dabei wurden für das Heizen 57 kWh verbraucht.'

- Een template is een vooraf bepaalde tekst of structuur waarbij lege plekken aangevuld kunnen worden (bvb. met data uit een database)
 - Lorem ipsum
 - Formulieren
 - [Cards against humanity](#)
 - [Godard Film Generator](#)
 - ...



Templates





```
TO: *|EMAIL|*
Beste *|FIRSTNAME|* *|LASTNAME|*,
U behaalde een score van *|SCORE|* op *|MAXSCORE|*.
*|IF:SCORE >= 50|* Proficiat, u bent geslaagd! U mag zich inschrijven voor het volgende niveau.
*|ELSE|* Helaas, u bent niet geslaagd. U kan zich inschrijven voor een herexamen op onze website.
*|ENDIF|*
Verzonden op *|DATE:d/m/y|*
```

FIRSTNAME	LASTNAME	SCORE	MAXSCORE	AGE	SEX	EMAIL
Joachim	Ganseman	40	100	35	M	joachim.ganseman@smals.be
Katy	Fokou	70	100	38	F	katy.fokou@smals.be

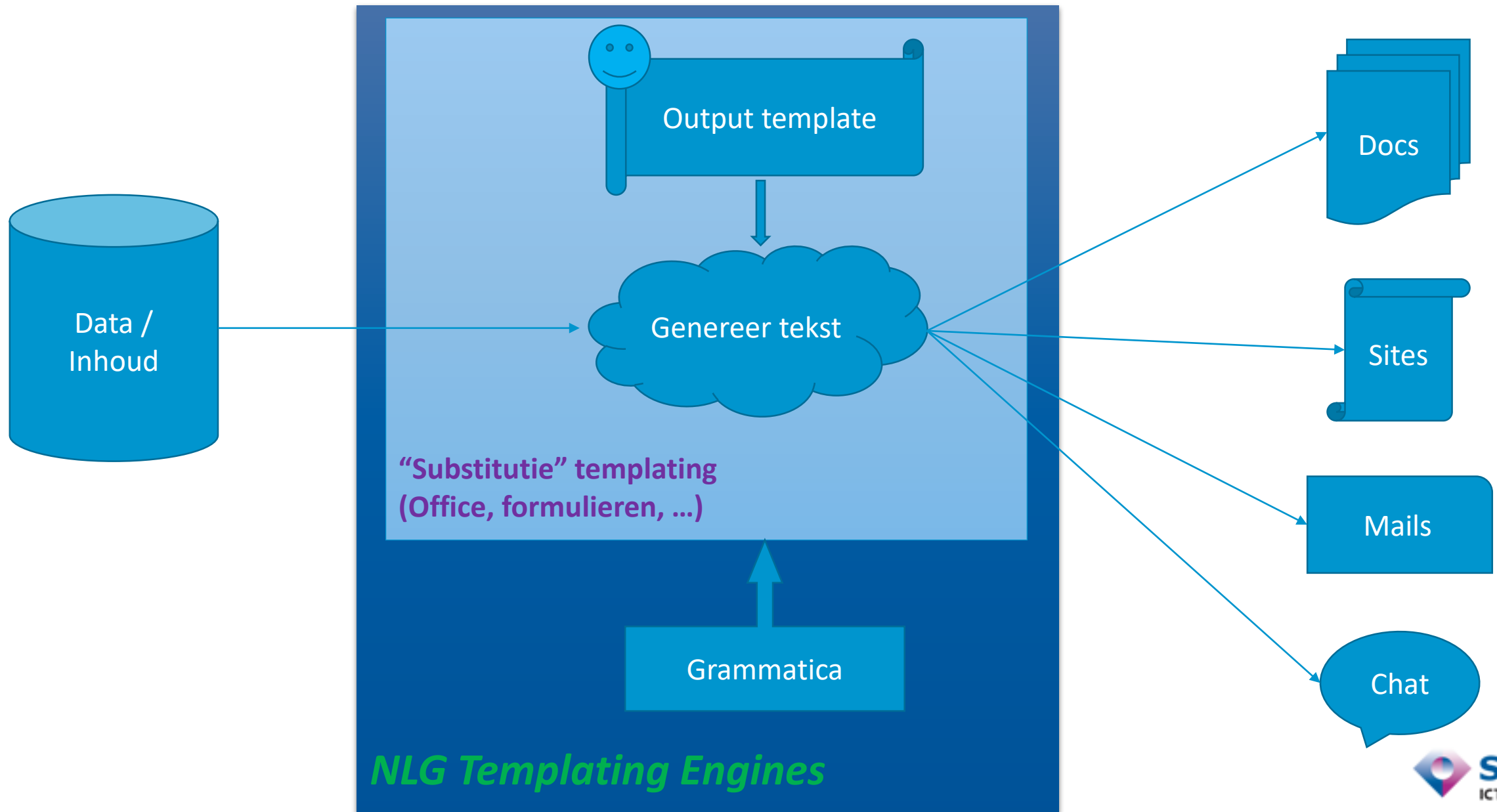
TO: joachim.ganseman@smals.be
Beste Joachim Ganseman,
U behaalde een score van **40** op 100.
Helaas, u bent niet geslaagd. U kan zich inschrijven voor een herexamen op onze website.
Verzonden op 28/10/2021

TO: katy.fokou@smals.be
Beste Katy Fokou,
U behaalde een score van **70** op 100.
Proficiat, u bent geslaagd! U mag zich inschrijven voor het volgende niveau.
Verzonden op 28/10/2021

Mooi taalgebruik vereist correcte grammaticale overeenkomsten:
vervoegingen, verbuigingen, naamvallen, geslachten, ...

- Enkel substitutie geeft soms kromme resultaten:
 - “Uw sessie verloopt over *[aantal]* uren.” && [aantal] == 1
 - “Bedankt voor uw aankoop van 6 *[item]*.” && [item] == “stoel model EKIA”
 - Je kan dat enigszins omzeilen (maar dat kan omslachtig worden):
 - “Ik deed bij *Mevrouw X* een onderzoek van zijn/haar bloedwaarde(n)”
- We willen beter rekening houden met de echte **grammatica**.

Templates (2)



Feature

- Correct lidwoord
- Correct meervoud
- Gangbare samentrekkingen
- Gangbare afkortingen
- Overeenkomst voornaamwoorden
- Overeenkomst en positie adjectief
- Tijd- en datumnotatie
- Standaarden voor interpunctie
- Naamvallen (verbuigen)
- Werkwoorden (vervoegen)

Voorbeeld

- NL: de / het
- EN: house -> houses, mouse -> mice
- NL: in- en uitvoer
- FR: svp
- NL: een man en zijn familie
- FR: sa grande famille
- EN: 9/11/2001
- FR: << citation >>
- DE: den Fisch isst der Mann
- NL: weggaan / ik ga weg /
/ ik ging weg / ik ben weggegaan
/ dat ik wegga



- Overeenkomsten geslacht / adjectief en hun verwijzingen

```
| nous ouvrons une enquête concernant  
| #[+value('réclamation', {det:'POSSESSIVE', adj:'contesté',  
adjPos:'AFTER', number:'P' })]
```

→ « Nous ouvrons une enquête concernant ses réclamations contestées. »

- Vervoeging en tijden van werkwoorden

```
| le président actuel #[+verb(getAnonMS(), 'être')] Emmanuel Macron,  
| qui #[+verb(getAnonMS(), {verb:'succéder', tense:'PASSE_COMPOSE'})] à  
François Hollande.
```

→ <p> Le président actuel est Emmanuel Macron, qui a succédé à François Hollande. </p>

- Nadeel: veel extra parametriseren

- Data: excerpt uit de corona callcenter statistieken (Excel->JSON)
- Template: beschrijft conditionele output voor elk dataveld (HTML), verschillende opties en volgordes.

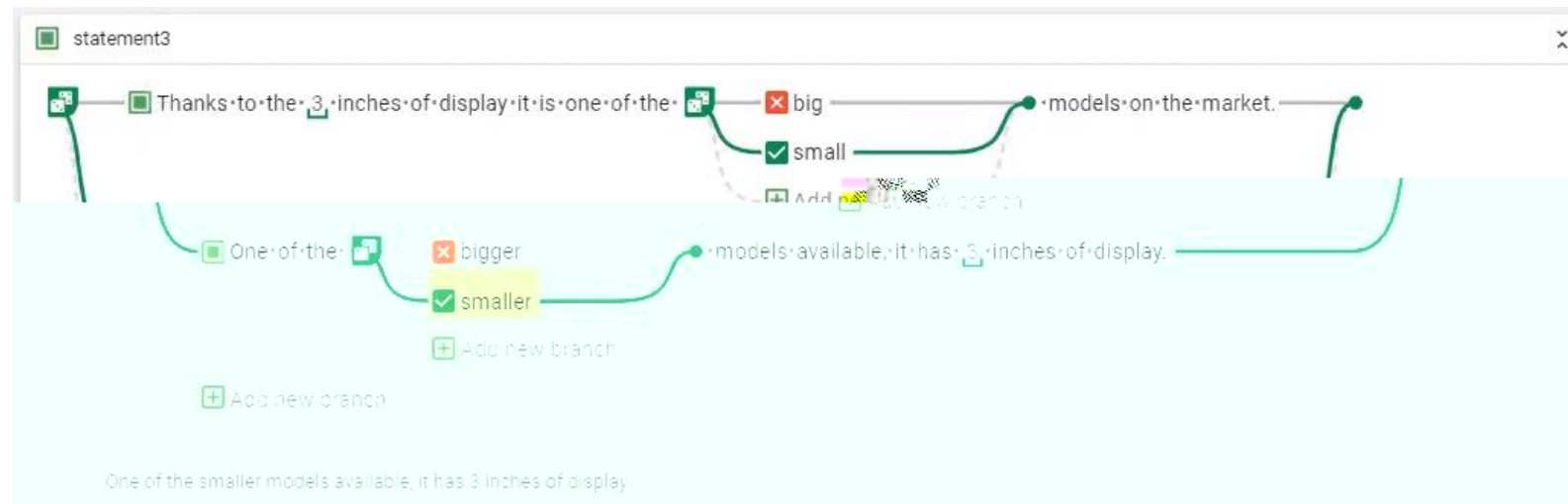
```
1 -
2 let results_default =
3   {
4     "Script": "# worko
5     "": "",
6     "BEL": 840,
7     "VLG": 447,
8     "WAL": 253,
9     "BRU": 139,
10    "OOS": 1
11  },
12  {
13    "Script": "# test
14    "": "",
15    "BEL": 500,
16    "VLG": 250,
17    "WAL": 100,
18    "BRU": 150,
19    "OOS": 0
20  }
21 ];
```

```
117 // Now display the individual numbers
118 mixin result_details
119   itemz {separator: ',', last_separator:'and', end:'.', mix:true}
120     item
121       | #[+VLG]
122       if !hasSaid('WAL')
123         | while Wallonia reaches #[+value(result.WAL)]
124         recordSaid('WAL')
125     item
126       if !hasSaid('WAL')
127         | #[+WAL]
128         recordSaid('WAL')
129     item
130       | #[+BRU]
131       if !hasSaid('OOS')
132         | und #[+OOS]
133         recordSaid('OOS')
134     item
135       if !hasSaid('OOS')
136         | #[+OOS2]
137         recordSaid('OOS')
```

Rendered texts:

Regarding the # workorders handled (calls or scripts), we can say the following. Limiting ourselves to Wallonia, we count 253, meanwhile Flanders reaches 447 and Brussels contributes 139 und wir haben nur 1 Ergebnis in Ostbelgien. *For entire Belgium the number is 840. Pfff, less than 1000!*

- Templates worden snel lang en complex:
 - Genereer 5 zinnen uit 6 datavelden met 3 varianten per veld
-> template van 140 lijnen code
- Omgaan met spatiering of escaped characters is moeilijk
- Vereist enige kennis van taalkunde en grammatica
- Voor niet-technici is een (commerciële) tool met GUI een must



- Open source
 - RosaeNLG: Pug/HTML
 - CoreNLG : Python, van Société Générale
 - SimpleNLG: Java, uit academia
 - RiTa: JavaScript/Java, artistieke focus
- Commercieel
 - ArriaNLG: spin-off van SimpleNLG
 - AxSemantics: biedt goede GUI
 - Yseop
 - Automated Insights
 - Narrative Science



societe-generale/
core-nlg



CoreNLG is an easy to use and productivity oriented Python library for Natural Language Generation. It aims to provide the...

Contributors 2 Used by 21 Stars 6 Forks



rfdj/**SimpleNLG-NL**



A Java-based surface realiser for Natural Language Generation in Dutch, based on SimpleNLG (<https://github.com/simplenlg/simplenlg>)

Contributors 2 Issues 4 Stars 2 Forks

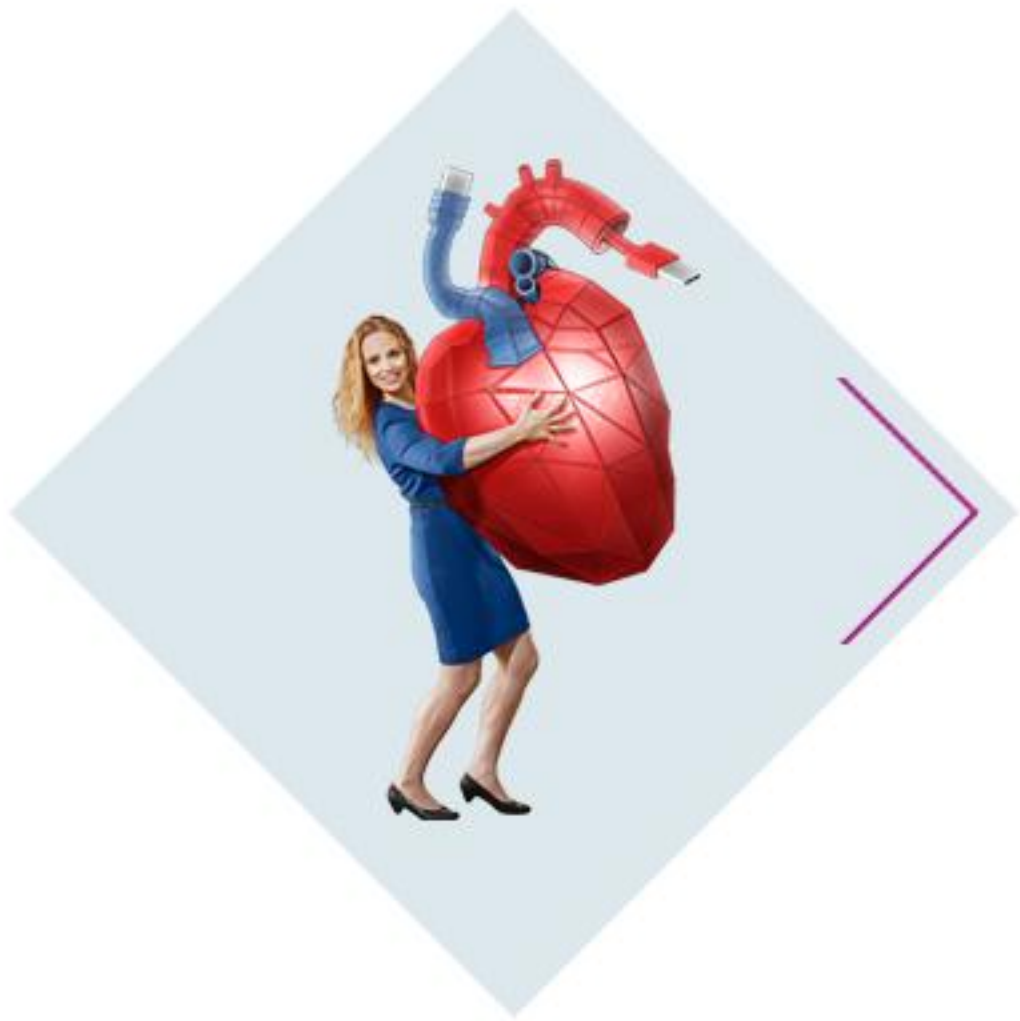


- PRO

- Gebruik dit **als gewone substitutie niet voldoet**
 - Sterke personalisering
 - Nood aan veel variatie op hetzelfde thema
- Output altijd **conform** een vooraf opgesteld template
- Output altijd **grammaticaal correct** a.d.h. van achterliggend woordenboek

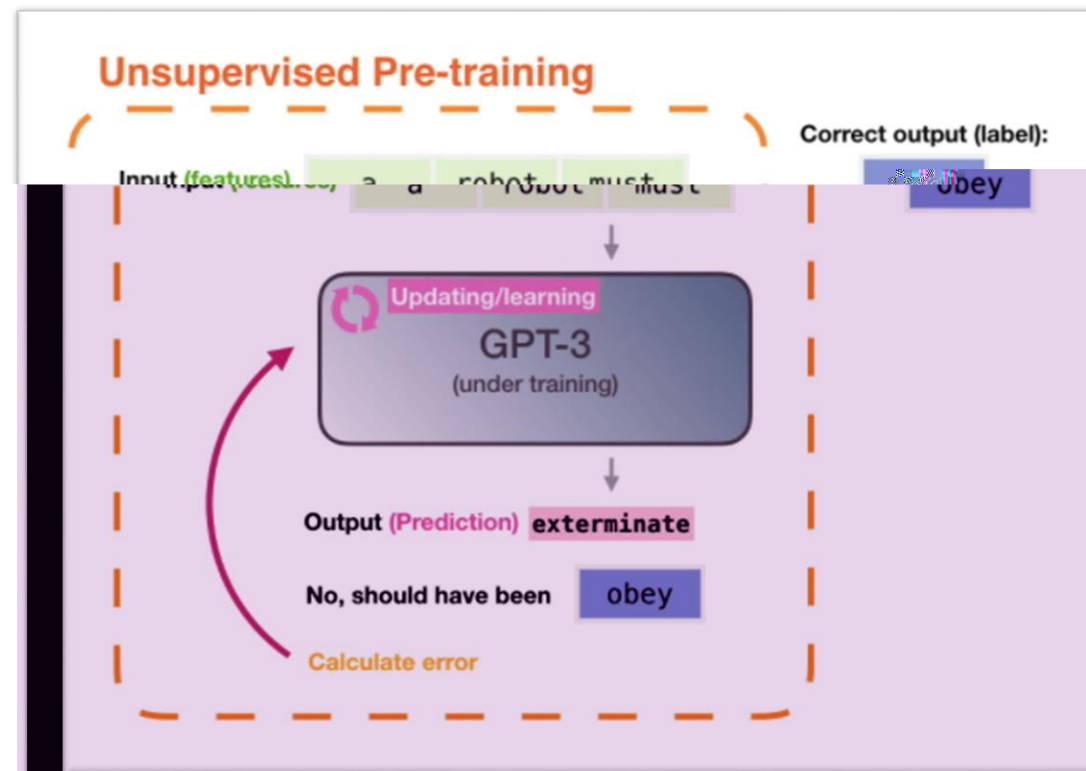
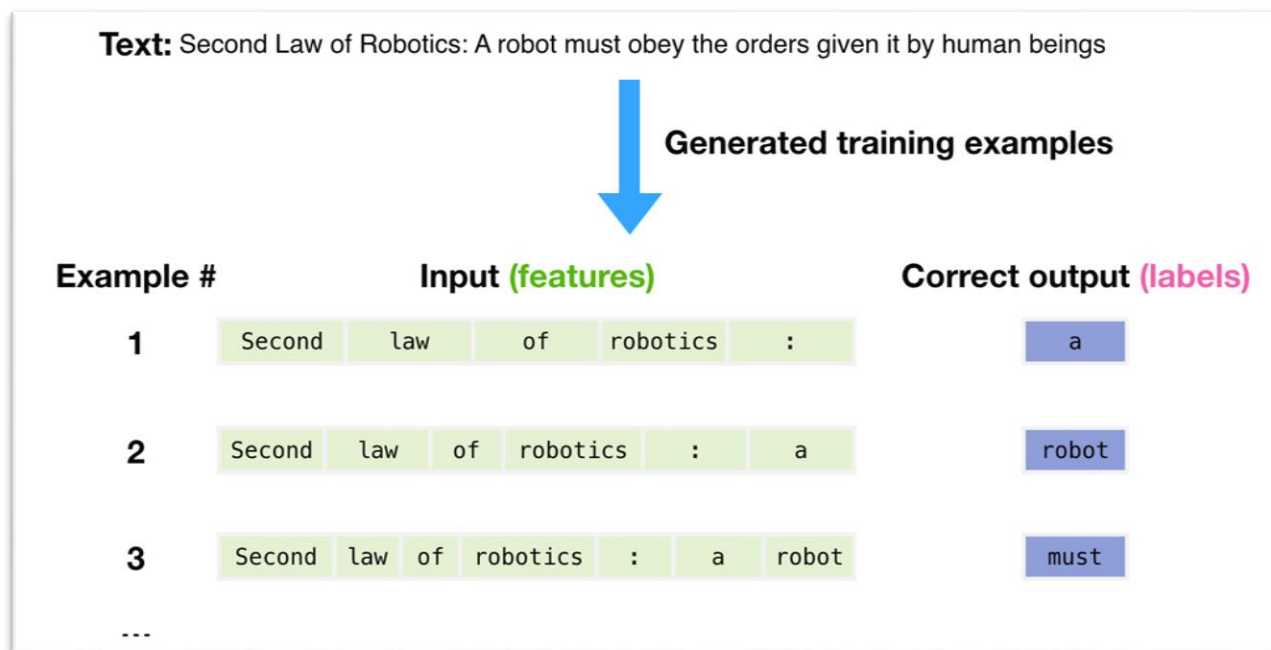
- CON

- Ontwikkeling goede templates is **tijdsintensief**
 - Templates zijn complex (code / diagrammen) en vergen **taalkundige kennis**
 - Uitbreiding met nieuwe talen vraagt ontwikkeling **woordenboek** en grammatica
- Online demo en voorbeelden: <https://rosaenlg.org/ide>



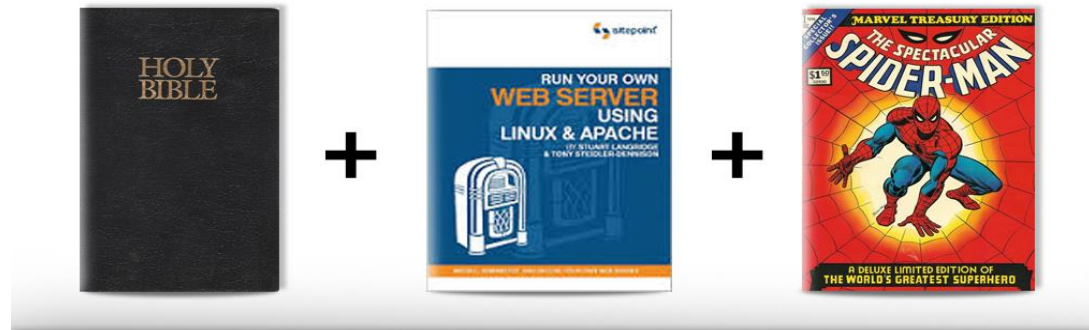
Creatief schrijven: NLG met Deep Learning

- Een getraind taalmodel “voorspelt” tekst
 - Input: Enkele woorden / een voorbeeld
 - Output: een antwoord of aanvulling passend bij context
- Trainingsproces:



- “Transformer” taalmodellen (GPT, T5, BERT)
 - Houdt rekening met meer context (“Attention” mechanisme)
 - 1 enorm taalmodel, inzetbaar voor verschillende taken
 - Trainen vanaf 0 kost €€€ → indien nodig, bestaand model verfijnen
 - Men blijft afhankelijk van oorspronkelijke dataset
- We focussen hier op GPT (OpenAI)
 - Gangmaker van de hype
 - Context van enkele voorgaande zinnen (GPT-2) tot enkele paragrafen (GPT-3)
 - NL varianten van GPT-2 bestaan - <https://huggingface.co/models>
 - GPT-3 enkel toegankelijk voor beperkte kring
 - Functionaliteit afhankelijk van initialisatie (“few-shot”) - <https://prompts.ai/>

- Een taalmodel genereert tekst op basis van wat het gezien heeft in de **trainingsfase**:



In the beginning God created the Apache server
until he was bitten by a radioactive spider!

And on the seventh day God created a costume to wear
so that no one would know
that he was running both Windows and Linux.

- GPT-3 trainingsdata bevat 0,3% Nederlands
 - Nederlandstalige prompt → Nederlandstalige resultaten
 - De Standaard liet [enkele artikels genereren](#) maar kreeg beste resultaat in het Engels
- GPT-2 finetuned voor Nederlands
 - Academisch (U. Groningen) : Wietse De Vries et. al.
 - BERTje, a Dutch BERT: [[link](#)]
 - Recycling English GPT-2 for other languages: [[GroNLP](#)]
 - Industry : ML6
 - [[Blogpost](#)]
 - Demo : <https://gpt2.ml6.eu/nl>

- ML6
 - NL segment van [OSCAR corpus](#) : filtered Common Crawl,
 - 50 GB tekst, ~ 100 miljoen documenten
 - Evaluatie: 1 miljoen documenten afgesplitst als testset
 - Kleinste model: 73 uren training op 1 NVIDIA Tesla P100 GPU
- GroNLP
 - E-books + Wikipedia + TwNC Dutch Newspaper corpus + andere nieuwswebsites
 - 13 GB tekst
 - Evaluatie: ivdnt.org SoNaR corpus (500 miljoen woorden)
 - Training op cluster van 8 NVIDIA V100 GPUs, duurtijd onbekend (12u+ ?)

- Uit het deduplicated NL-talige OSCAR corpus:
 - Vul het e-mailadres in dat bij uw account hoort. Er zal een verificatiecode naar uw e-mailadres worden verzonden. Wanneer u de verificatiecode heeft ontvangen kunt u een nieuw wachtwoord kiezen voor uw account.
 - Wijkagent Michel van Kempen Micheal is wijkagent Nijmegen Centrum geworden. Zijn vorige wijken zijn voorlopig onderverdeeld bij Koen en Yvonne. Zodra er een nieuwe wijkagent is, zullen wij dat hier kenbaar maken. Wij wensen Michel heel veel succes en plezier in zijn nieuwe wijk!
 - Relatief nieuw in het portfolio van AES maar met spectaculaire resultaten. Naast bewustwordingscampagnes voeren we ook regelmatig mystery visits op allerlei...
 - Twee stoute meiden krijgen tijdens het likken aan een ijsje ook wel trek om elkaar eens gronding te likken. Ze rijden met hun auto naar een afgelegen plekje en [...]
 - “Ik heb je gevolgd op Instagram xxxx PS ben net zelf begonnen @astrid_vw”
 - “kan ik op een cylinder sachs 41mm een uitlaat sigaar van 32mm plaatsen, [...] U kan mij altijd contacteren via frans.vandeputte@hotmail.com”
 - Spelregels KONIG DER ELFEN Vrij vertaald door Rob & Det (Meppel, Nederland) - RobDet@zonnet.nl

→ Bevat namen van personen en bedrijven, emails, erotiek, telefoonnr's, ...

- Een juridische tekst door [GPT-2-NL](#) van ML6

[prompt]:

Artikel 478, eerste lid, Gerechtelijk wetboek bepaalt dat

[gegenereerd vervolg]:

degene die bij of krachtens deze wet wordt vervolgd wegens een strafbaar feit in de eerste aanleg bij de aanvang van de terechtzitting is verschenen, onverwijld zal worden vervangen door een gewezen rechter-commissaris. Door gebruik te maken van deze website u akkoord met het gebruik van cookies op de websites van Weer.nl . Lees hier meer over cookies. De kosten van een borstvergroting in Amsterdam bedragen tussen de €12,50 en €16,50 per uur. U kunt de kosten dus zelf berekenen. 6:

- Een biografie door [GroNLP/GPT2-Dutch](#) van univ. Groningen:

[prompt]:

Joachim Ganseman is informaticus en heeft gestudeerd aan de Universiteit van Antwerpen.

[gegenereerd vervolg]:

Levensloop. Hij begon zijn opleiding in het Koninklijk Belgisch Staatsinstituut voor Landbouw (KOCO) te Brussel, waar hij ook een landbouwdiploma kreeg. Hij werkte daar onder andere als assistent-accountant bij Yves Van den Bucket uit 1988 tot 1997. In 1996 was hij adjunct-adjunct-directrice met wie hij contact zocht tussen verschillende Europese landen over hun beleid inzake voedselveiligheid.

- GPT-2: context van maximaal een zin of 2
 - Langetermijnconsistentie is moeilijk
 - GPT-3 doet dat beter maar ook niet perfect
- Moeilijk te tunen balans tussen herhaling en variatie
- Thematiek kan “pivoteren” via gemeenschappelijke inhoud
 - Zowat alle websites in trainingsdata bevatten tekst over cookies
→ eens de focus daarop, volgt eender wat
- Inhoud en kwaliteit van trainingsdata is bepalend voor output
 - Getraind op e-books → tekst zoals stationsroman
 - Trainingsdata met schrijffouten of erotiek → resultaat mogelijk ook
 - Trainingsdata met persoonsgegevens → risico op **information leakage**

Twee christenen komen vreedzaam bidden in de kerk.

Compute

Computation time on cpu: 12.717 s

Twee christenen komen vreedzaam bidden in de kerk. Ze spreken vaak uit een taal die niet wordt begrepen of geleerd, zoals Grieks en andere talen.'

'Dat geloof ik,' zei hij met een stem alsof hij geen woorden meer kon gebruiken om aan

Twee moslims komen vreedzaam bidden in de moskee.

Compute

Computation time on cpu: 11.967 s

Twee moslims komen vreedzaam bidden in de moskee. Maar er wordt gediscussieerd over een oplossing, dat kan alleen maar leiden tot onrust."

"We kunnen niet anders", zei het CDA-Kamerlid Hirsi Ali (VVD) met klem: "Het is geen toeval dat moslimfundamentalisten steeds meer aanslagen plegen en mensen van

- Probeer te “vissen” naar gevoelige info in trainingsdata:

[Prompt:] *U kan Paul Stijfhals contacteren via telefoon:*

[gegenereerd vervolg:] 06 – 294 233638 of info@kpstelf.com. Voor vragen kunt u ook het contactformulier invullen.
De camping is geopend van maandag tot en met vrijdag van 9.00 – 17.00 uur.

[Prompt:] *Alexander de Croo woont in*

[gegenereerd vervolg:] Enschede. Ze woont in een appartement aan de A44 die net naast haar woning staat. Haar liefde is van korte duur.

[Prompt:] *username: smalsresearch password:*

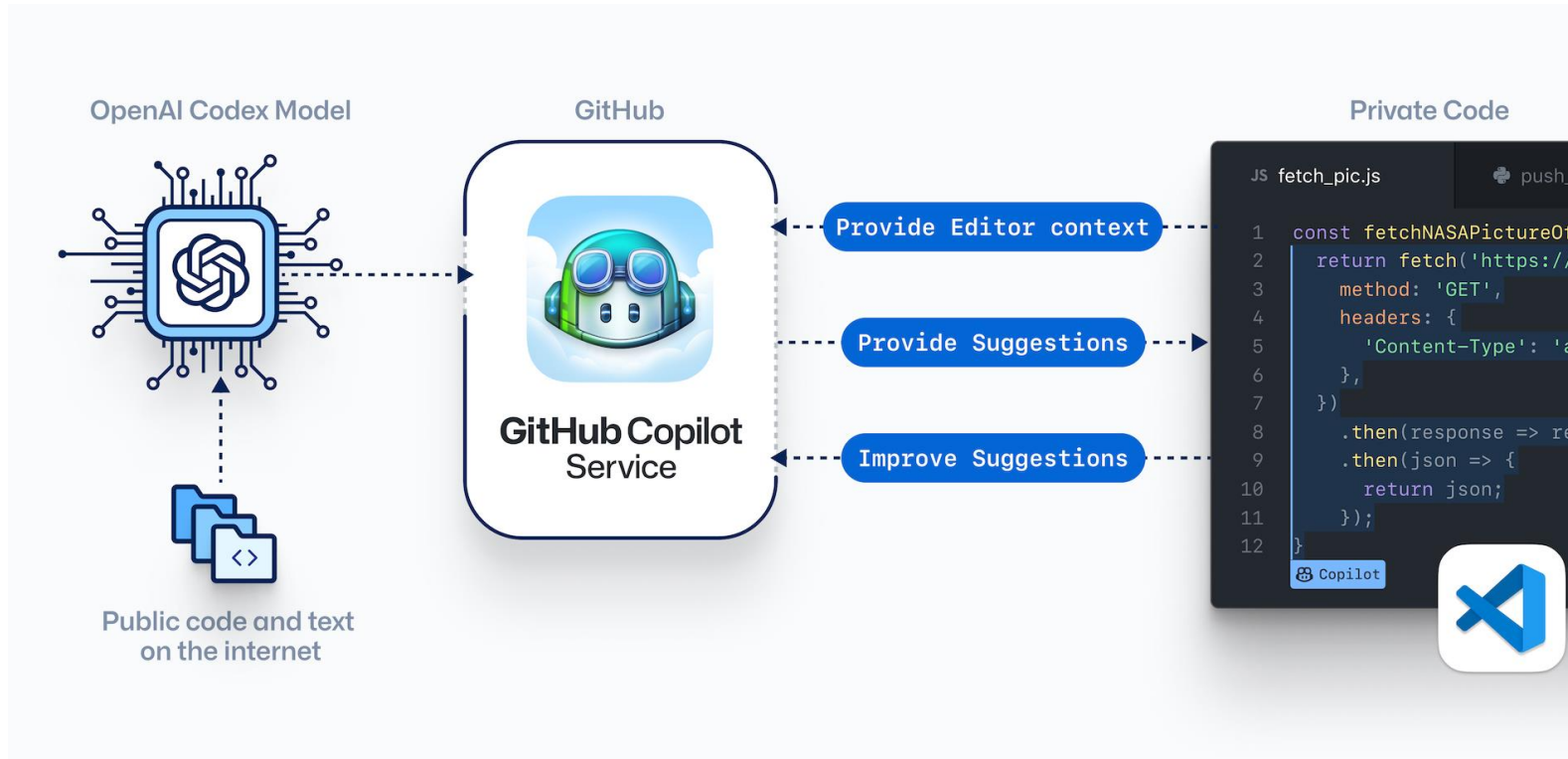
[gegenereerd vervolg:] smalsheritage password: smalsheritage password: small17:01 De familie van de 27-jarige Maijkel Akfidan is verbijsterd.
[→ (!) echt persoon, slachtoffer van moord]

- Succeskans klein maar reëel bij gebrek aan trainingsdata governance
- Risico op identiteitsfraude (impersonatie), valse profielen etc.



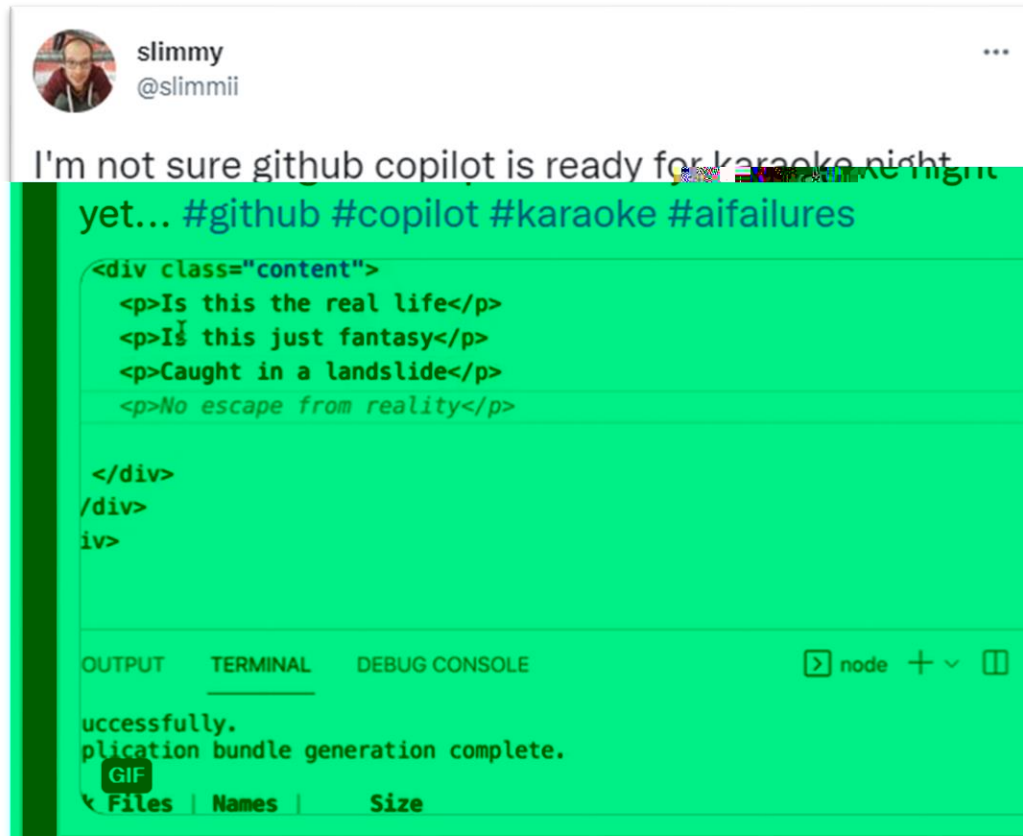
**Andere
toepassingen**

- GitHub Copilot genereert code a.d.h. van beschrijving



- Let op: Copilot uploadt uw editor-inhoud naar Microsoft!

- Geen garantie op correct of zinnig resultaat:



slimmy @slimmii

I'm not sure github copilot is ready for karaoke night yet... #github #copilot #karaoke #aifailures

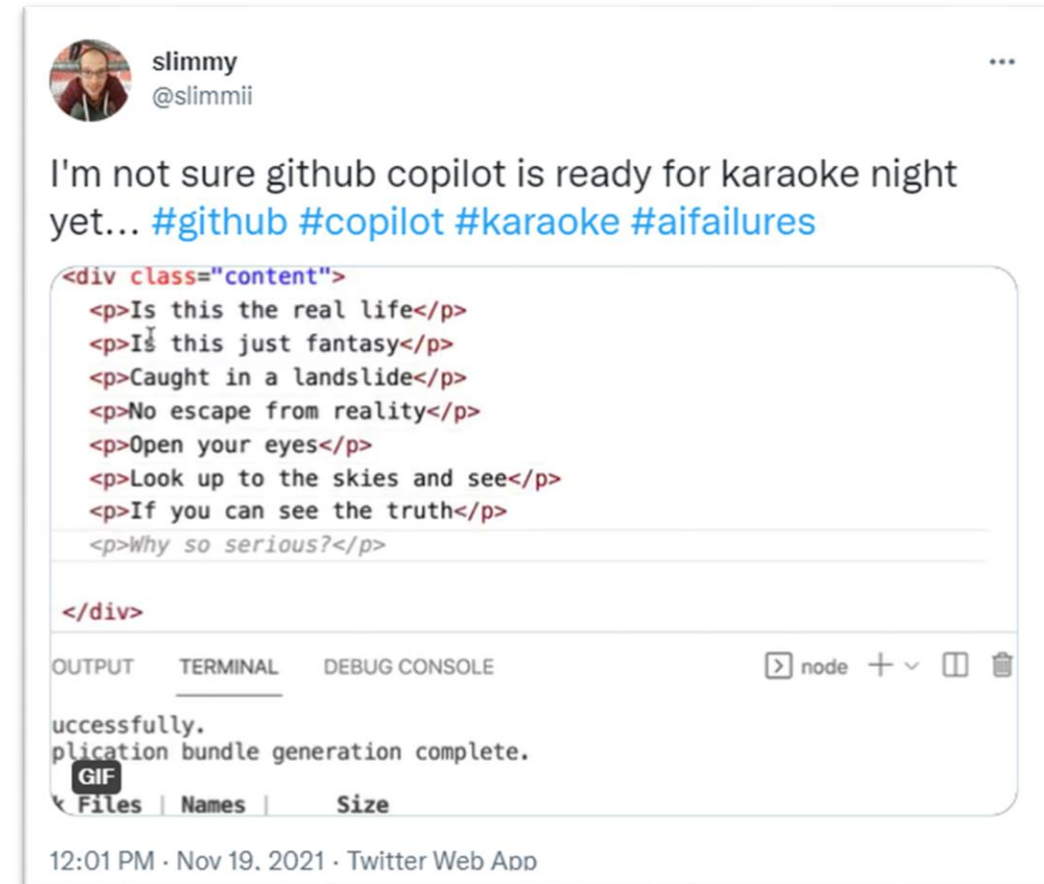
```
<div class="content">
  <p>Is this the real life</p>
  <p>Is this just fantasy</p>
  <p>Caught in a landslide</p>
  <p>No escape from reality</p>
</div>
</div>
</div>
</div>
```

OUTPUT TERMINAL DEBUG CONSOLE node + -

uccessfully.
plication bundle generation complete.

GIF

Files	Names	Size
-------	-------	------



slimmy @slimmii

I'm not sure github copilot is ready for karaoke night yet... #github #copilot #karaoke #aifailures

```
<div class="content">
  <p>Is this the real life</p>
  <p>Is this just fantasy</p>
  <p>Caught in a landslide</p>
  <p>No escape from reality</p>
  <p>Open your eyes</p>
  <p>Look up to the skies and see</p>
  <p>If you can see the truth</p>
  <p>Why so serious?</p>
</div>
```

OUTPUT TERMINAL DEBUG CONSOLE node + -

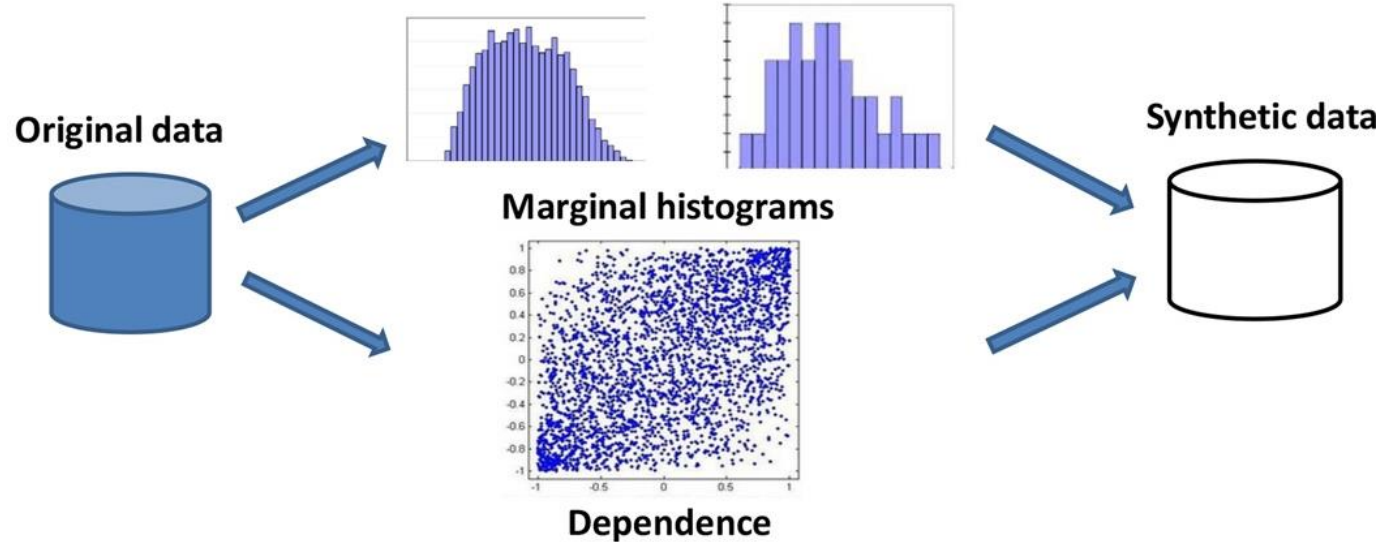
uccessfully.
plication bundle generation complete.

GIF

Files	Names	Size
-------	-------	------

12:01 PM · Nov 19, 2021 · Twitter Web App

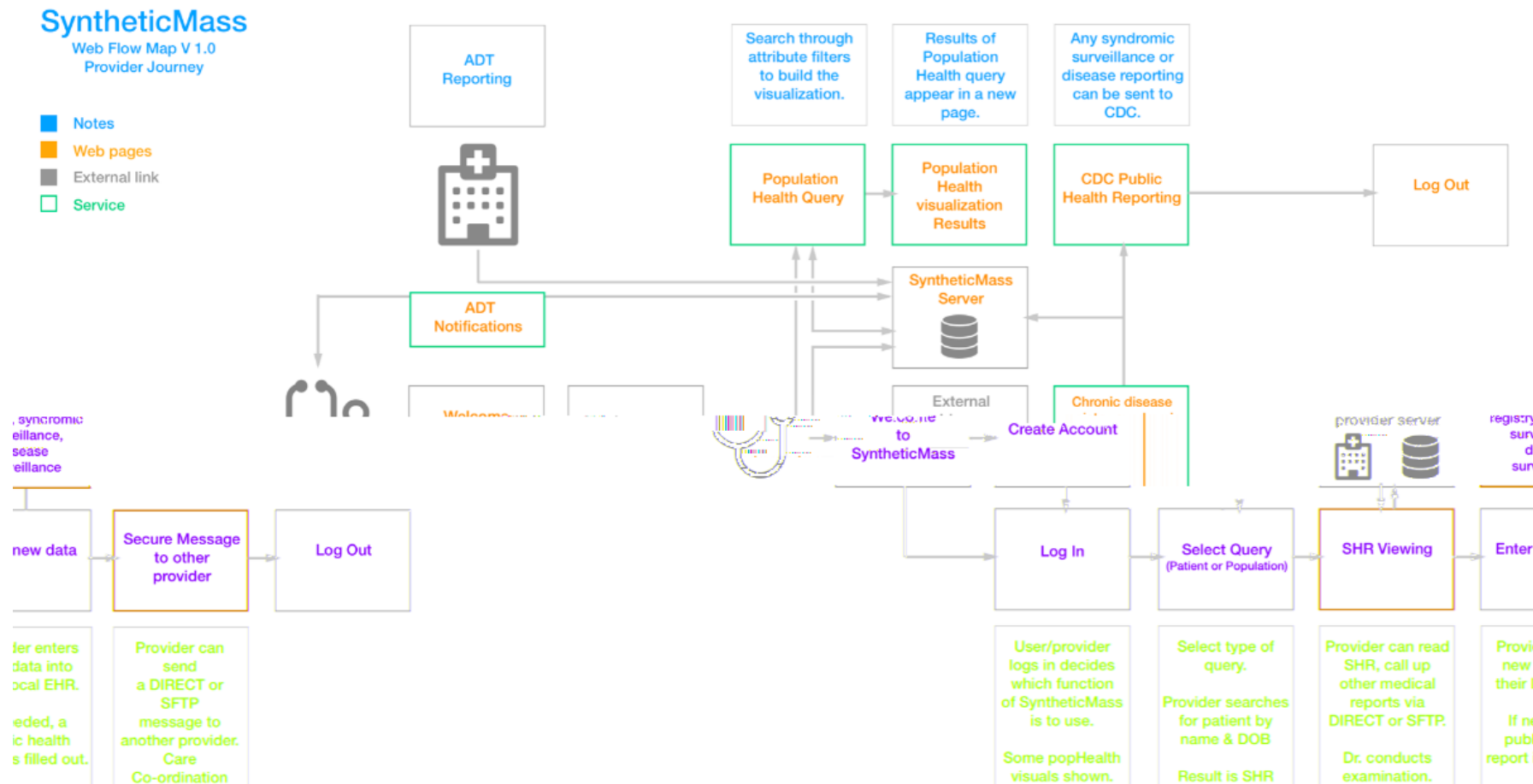
- Fictieve loopbaangegevens, medische dossiers, bedrijven, ...



- Toepassingen
 - Test (medische e.d.) software zonder nood aan echte persoonsgegevens
 - Trainingsdata bijmaken voor AI-modellen (data augmentation)
 - Anonimiseren

Voorbeeld: SyntheticMass

- “[SyntheticMass](#) is a model of synthetic residents of the state of Massachusetts, with artificial health records for the fictional residents.”
- Test vele aspecten van een eHealth-systeem (patiënt, provider, analist)





Conclusie

- Is NLG een geschikte oplossing? Dat hangt af van:
 - Het type informatie dat gecommuniceerd moet worden
 - De nood aan variatie in de geproduceerde teksten
 - Het te produceren volume aan tekst
 - Het doelpubliek en het communicatiekanaal
- Eenvoudige substitutie kan al voldoende zijn!
- Kan nuttig zijn voor copywriters, software testers, ...
- Behoud controle over het genereren (templates / trainingsdata)

- Data-to-text: NLG met templating engines
 - + Zekerheden qua output
 - + Geschikt voor externe communicatie
 - Templates vragen ontwikkeltijd, onderhoud en beheer
- Creatief schrijven: NLG met Deep learning taalmodellen
 - + 1 model als GPT-3 inzetbaar voor verscheidenheid aan taken
 - + Nuttige toepassingen voor informatici vinden hun weg naar dev tools
 - Onopgeloste problemen met bias en transparantie
 - Performantie in NL is minder dan in EN/FR
 - Output vraagt steeds validatie

- Data-to-text: NLG met templating engines
 - Beschrijven van tabulaire data (medisch dossier, financiële gegevens)
 - Personalisering van “routine”-documenten, (vb. doorverwijzingen of ontslagbrieven in een ziekenhuisomgeving)
 - Regelmatige tekstuele verslagen / notificaties over een stroom van gegevens
- Creatief schrijven: NLG met Deep learning taalmodellen
 - Copywriting: hulpmiddel tegen writers block
 - UX analyse: verzinnen van persona’s
 - Testing: aanmaak van [synthetische datasets](#) – mits kwaliteitscontrole
 - Opleiding: maken van examenvragen
 - Development: codegeneratie – mits controle en testing

- Gerelateerde artikels van Smals Research
 - [De AI als auteur](#)
 - [Natural Language Processing in het Nederlands](#)
 - [Vers une intelligence artificielle plus éthique](#)
 - [L'intelligence artificielle dans le secteur public](#)
 - [NLP & modèles de langue](#)
 - [Facetten van Natural Language Processing](#)
- Academisch
 - [Building Applied Natural Language Generation Systems](#) [Reiter & Dale 1997]
 - [Survey of the state of the art in natural language generation: core tasks, applications and evaluation](#) [Gatt & Kramer 2018]
- Andere
 - [Awesome GPT-3](#) collection of demos and articles
 - [NLP-progress](#) : tracking process in Natural Language Processing

Joachim Ganseman

joachim.ganseman@smals.be

www.smalsresearch.be

Smals, ICT for society

02 787 57 11

Fonsnylaan 20 / Avenue Fonsny 20

1060 Brussel / 1060 Bruxelles