Introductie van de case study

Sunil Choenni
Niels Netten

Programma – dag 1

09:15 – 09:30	Inloop en terugblik	
9:30 – 09:50	Welkomstwoord	Jaap van den Herik
9:50 – 11:00	Introductie data science binnen de overheid	Jaap van den Herik
11:15 – 12:30	Introductie case study	Sunil Choenni & Niels Netten
12:30 – 13:30	Pauze	
13:30 – 15:00	Werken aan de case study	Sunil Choenni & Niels Netten
15:00 – 15:15	Pauze	
15:30 – 16:00	Vervolg werken aan de case study	Sunil Choenni & Niels Netten
16:00 – 17:00	Gezamenlijke presentatie	Jaap van den Herik & Clint Pennings

Big Data: potentials en uitdagingen

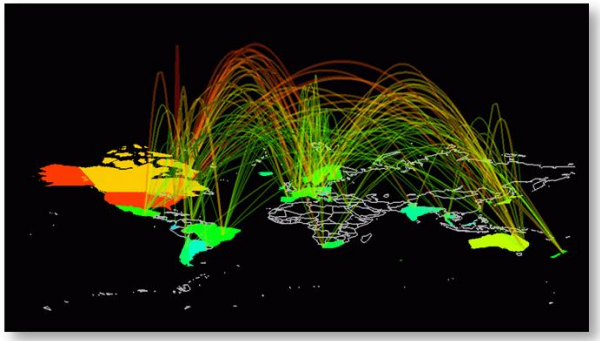
Sunil Choenni

Overzicht

- Introductie
- Potentials
- Uitdagingen
  - Privacy
  - Interpretatie
- Discussie

Big & Open Data

- Combinatie van real time, verschillende types en veel data
  - sensor en streaming data
  - audio, video, natuurlijke taal
  - terra- of pettabytes
- Steeds meer apparaten en (mensen) hangen aan Internet
- Integratie van real-time, multimedia en zeer grote database technologie



Mogelijkheden (ten dienste van de mens)

- Automatisch signaleren wat er in je koelkast moet
- Simuleren alsof je thuis bent terwijl je op vakantie bent
- Auto’s die met elkaar communiceren
- Betere en actueler voorspellingen
- Ouderen helpen navigeren
- Veiligheid
- Serious Games, benadert de werkelijkheid

Big data en opsporing

- Huisjesmelkers
  - GBA koppelen met energie- en waterleveranciers
- Wietplantages
- Getuigen (benut sensor data)
- Fraude en frauderende studenten
- ….

Uitdagingen: softe kant

- Welke data mag je met elkaar combineren?
- Hoe ga je verkregen combinaties duiden?
- Hoe behoud je de relatie met de echte werkelijkheid?
- Privacy wordt een issue; prijsdifferentiaties nu ook al bij AH

Uitdagingen

- Privacy/vrijheid
- Duiding van resultaten
- Data kwaliteit
- https://www.youtube.com/watch?v=jIv1HH2LWlU&feature=youtu.be
- https://www.youtube.com/watch?v=K9ghFuyd2Tk&feature=youtu.be
- Interoperabiliteit

Uitdagingen: privacy

- Door de grote stromen van data en de combinatie van de data is de kans op privacy-schending groter
- Stel een school publiceert gemiddelde examencijfers, uitgesplitst naar jongens en meisjes; via social media weten we dat maar twee meisjes aan een vak deelnemen en een van de meisjes had een 9 → cijfer van andere meisje is dan ook bekend

Uitdagingen: privacy

Grote aantallen kunnen ook onthullend zijn.

- Stel we hebben een database beschikbaar van autorijders die bij een aanrijding betrokken zijn. Met data mining algoritmes kunnen we alternatieve beschrijvingen van de database vragen.

- Vb. Jonge mannen met een lease auto zijn voor 80% van de aanrijdingen verantwoordelijk. Kan stigmatiserend werken, omdat de relatie met het totaal aantal autorijders niet wordt gelegd.

Uitdagingen: privacy

- Data die stigmatiseren niet meenemen voor mining kan leiden tot red lining
  - Ras en woonwijken zijn in den regel onafhankelijke attributen
  - Ras nemen we niet mee in onze mining database, maar woonwijken wel
- Als in bepaalde woonwijken veel allochtonen wonen en mining resultaten hebben betrekking op die woonwijken, dan is attribuut ras indirect toch meegenomen
- Bepaalde attributen dragers van een aantal onafhankelijk attributen?
- Allochtonen dragers van lage opleiding, hoge werkloosheid, …?
- Geen dragers meenemen in mining database

Human Values/Privacy

- Open voor verschillende interpretaties, afhankelijk van de context
- Het geloof dat een specifieke gedragswijze/eindtoestand persoonlijk of sociaal de voorkeur heeft boven een tegenovergestelde of omgekeerde modus
- Waarden verwijzen naar gewenste doelen, gesorteerd op hun belang
- Relatief belang van meerdere waarden stuurt de acties van mensen

Informatie classificatie: identificeerbaarheid, sensitiviteit en abstractie

- Anoniem
  - Kan niet worden herleid naar een individu
  - Bijvoorbeeld: aantal producten verkocht
- Pseudoniem
  - Kan worden herleid naar een individu
  - Bijvoorbeeld: Cookie ID, hashed kentekenplaat
- Persoonlijk
  - Kan worden gerelateerd naar een natuurlijk persoon
  - Bijvoorbeeld: IP adres, kentekenplaat

Informatie classificatie: sensitiviteit

- Public
  - Informatie die publiekelijk beschikbaar is
  - E.g. Een artikel dat iemand heeft geschreven; Jaarrapporten
- Private
  - Informatie over individuen/organisaties dat niet gedeeld is
  - E.g. Corporate strategische plannen, e-mails
- Sensitief (geheim) (WBP artikel 16)
  - Direct/indirect onthullen van sensitieve informatie over persoon
  - E.g. Medische gegevens, ras, sexualiteit, justitiele gegevens

Informatie classificatie: metadata

Metadata versus content

- Metadata
  – Data beschrijving over een of meer aspecten van de data
  – Bijvoorbeeld: tijd, locatie, bellers in een telefoongesprek
- Content
  – De uiteindelijke content van de data
  – Bijvoorbeeld: Belgegevens

  Beide metadata en content kunnen *sensitieve* en *persoonlijke* informatie onthullen!

Privacy by design – 3 voorbeelden

1. Openbare veiligheid

- Doel: beleidsmakers een hulpmiddel bieden waarmee ze mashups kunnen maken, d.w.z. in staat zijn om gegevens uit verschillende bronnen te combineren en hun eigen inhoud te creëren.

- Vereiste: voorkom ongewenste effecten
  - schending van privacy
  - verkeerde interpretatie van statistieken
  - openbaarmaking van de identiteit van een groep individuen

Meten van veiligheid

Fenomenen en variabelen m.b.t. de openbare veiligheid:

- Hebben benut de literatuur over criminologie en openbare veiligheid
- Hebben benut de domeinkennis en databases

Fenomenen gerelateerd aan openbare veiligheid (± 1500 variabelen)

Misdaad - geregistreerde misdaad, slachtoffers, preventieve maatregelen

Handhaving - politiecontacten, verdachten, opgeloste misdaad

Sanctie - boetes, gevangenisstraffen, gerechtelijke zaken

Politie en justitie - gevangeniscapaciteit, politieagenten

Informatiebehoefte

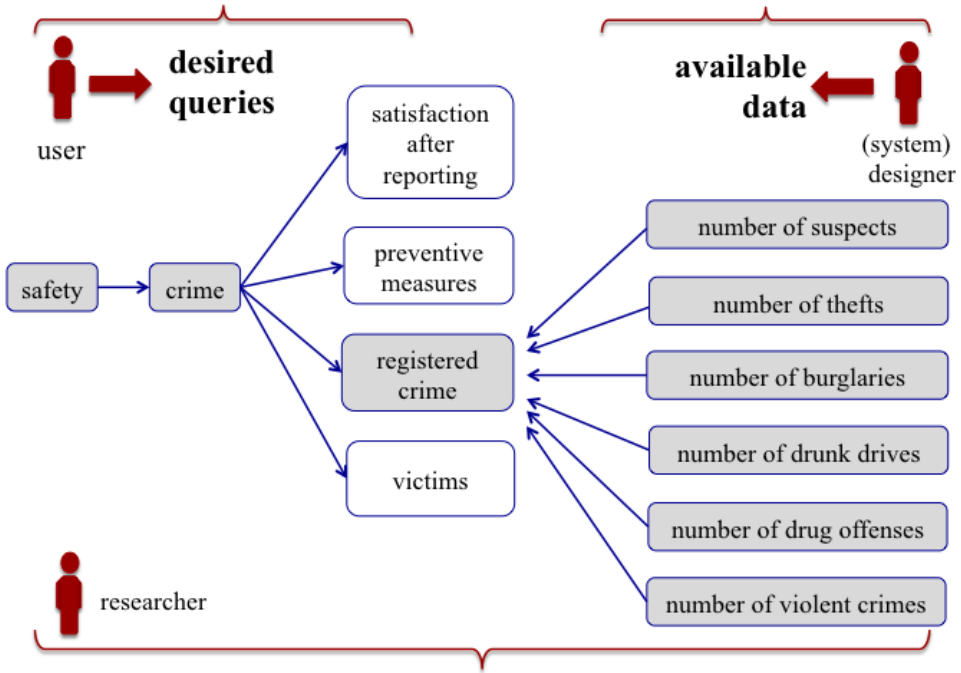
- Op basis van twee workshops: ongeveer 30 mensen, variërend van junior beleidsmakers tot directeuren.

- Enkele individuele vergaderingen na de workshops.

Resultaten

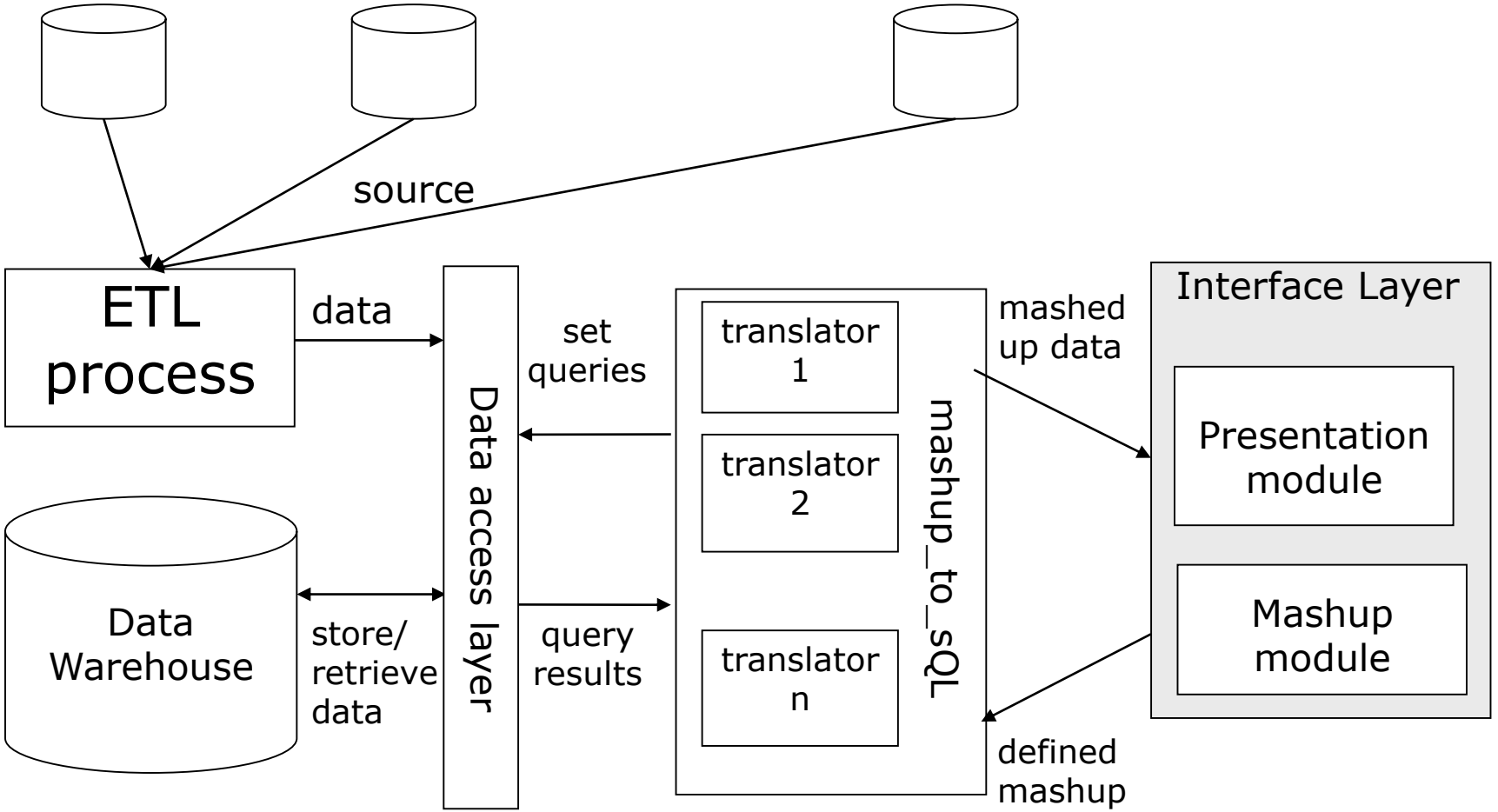
- Drie type vragen;
- Contextuele gegevens zijn ook vereist;
- Benodigdheden voor de tool.

Schets van het fenomeen *misdaad* gepresenteerd tijdens workshops



Drie type vragen

- *Eenvoudige queries*. Hoeveel mensen in een regio binnen een periode reageerden bijvoorbeeld op een specifieke manier op een specifieke enquêtevraag?
- *Context van een quantifier*. Hoe verhoudt bijvoorbeeld de groei of daling van een bepaald cijfer in een geografische regio zich tot een ander cijfer? Een toename van fietsdiefstallen in een wijk kan bijvoorbeeld een relatieve terugval worden als de lokale bevolkingsgroei groter is dan de groei. Deze contextualisering moet worden overwogen om de gegevens over de openbare veiligheid te begrijpen.
- *Similarity queries*, d.w.z. zoeken naar regio's die in zekere zin dezelfde context delen. Na het opvragen van een specifieke dataset waarin sommige aantallen op de een of andere manier opvallen, kan de gebruiker vragen stellen over andere regio's met vergelijkbare aantallen of trends.


source
ETL process
data
Data access layer
set queries
translator 1
translator 2
translator n
mashup_to_sql
mashed up data
Interface Layer
Presentation module
Mashup module
defined mashup
Data Warehouse
store/ retrieve data
query results

Architectuur: om schending van privacy te voorkomen

- Er worden alleen attributen in het systeem opgeslagen die in overeenstemming zijn met de Nederlandse Wet Bescherming Persoonsgegevens, d.w.z. geen gegevens over iemands religie of levensovertuiging, politieke overtuiging, gezondheid, seksuele geaardheid, etnische afkomst
- Alleen geaggregeerde gegevens worden opgeslagen
- Mashups die resultaten bevatten die de privacy schenden, worden niet getoond door de presentatiemodule (bijvoorbeeld als er slechts 2 veroordeelde personen zijn voor een misdaadtype X, wordt dit niet getoond.) Ook als 90% van de mensen in een regio betrokken zijn bij misdaad, dit wordt ook niet getoond
- Een uitgebreide verklaringsmodule om de interpretatie te vergemakkelijken

2. Gezondheidszorg

- Kleinschalige huisvesting voor dementiepatiënten neemt toe
- Doel: de kwaliteit van leven verhogen door het aanbieden van alternatieven van het traditionele verpleeghuis
- Huizen uitgerust met infraroodsensoren om personeel te waarschuwen als patiënten hulp nodig hebben, b.v. uit bed vallen kan leiden tot breuken (duur om te herstellen voor ouderen)
- Veel valse alarmen als gevolg van vallende dekens en / of dekbedden

Gezondheidszorg

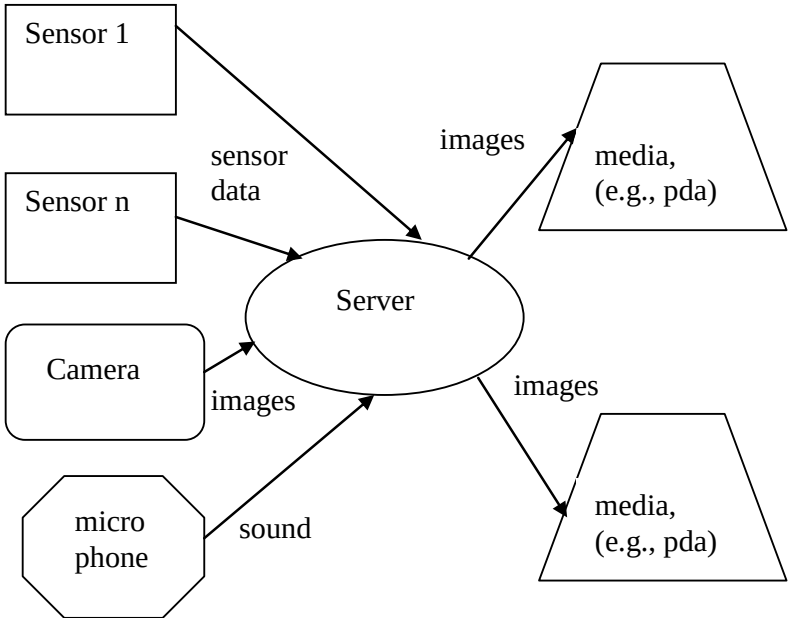


Figure 1: Design of a health care system

Gezondheidszorg: privacy en vertrouwen

- Bezwaren tegen privacy werden weggenomen omdat de server beslist of het bewegingen van patiënten naar de verzorgers zal sturen.

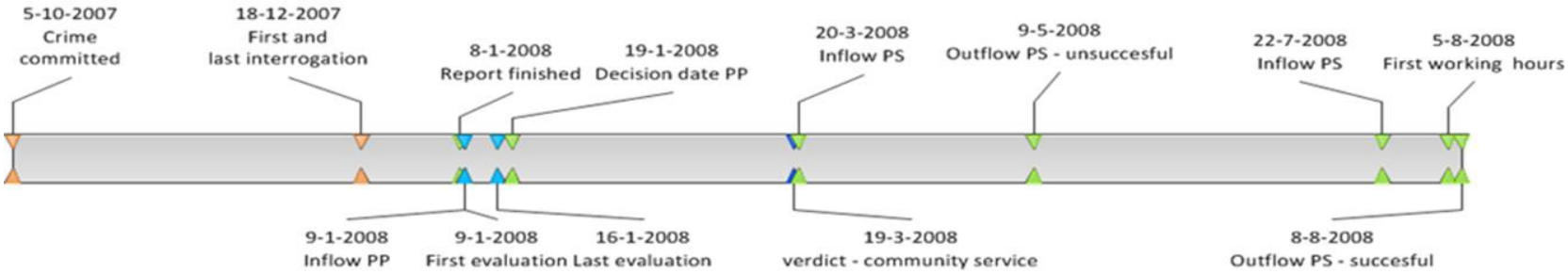
- Wanneer een verzorger de camera in een huis inschakelt, wordt deze door het systeem vastgelegd en doorgegeven aan hoger management.

- Vertrouwen tussen patiënt en zorgverlener blijft van belang, aangezien patiënten het systeem beschouwen als een uitbreiding van het personeel en niet als een vervanging.

3) Doorlooptijden

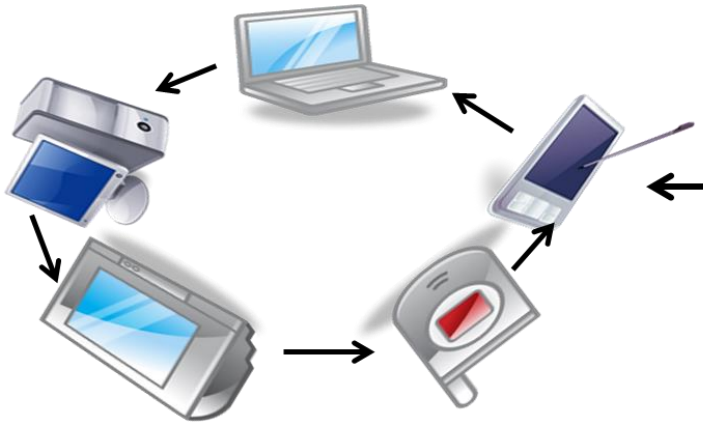
POLICE	Suspect ID	Report ID	Crime committed	First Interrogation	Last interrogation	Report Finished	
	1122	22-1234-33	5-10-2007	18-12-2007	18-12-2007	8-1-2008	
	1124	22-1234-33	5-10-2007	22-12-2007	22-12-2007	8-1-2008	
PP	Case nr.	Report ID	Inflow date	First evaluation	Last evaluation	Case decision	Decision date
	23333	22-1234-33	9-1-2008	9-1-2008	16-1-2008	Subpoena	17-1-2008
Courts	Case nr.	Report ID	Trail	Verdict	Punishment	Verdict	
	23333	22-1234-33	19-3-2008	Guilty	Community service	19-3-2008	
PS	Person ID	Case nr	Date inflow	First working hours	Product	Date outflow	Result
	997788	22-1234-33	20-3-2008		Community service	9-5-2008	No contact with client
	997788	22-1234-33	22-7-2008	5-8-2008	Community service	8-8-2008	Finished community service

Koppelen en transformeren van datasets



Duiding

Environment



Information systems

Dataprocessing

- classification
- rearranging /sorting
- aggregating
- summarizing
- calculations on data
- selection of data

Data

Information

Over de interpretatie van data-analytics resultaten

De systeemwerkelijkheid kan afwijken van de echte realiteit

- Databases kunnen grote hoeveelheden data bevatten die verzameld en opgeslagen zijn in het (verre) verleden.
- Wanneer legacy databases worden gebruikt, dan houden de verkregen resultaten door analyse, (e.g. data mining) niet altijd stand in de echte realiteit van vandaag te dag.
- De resultaten kunnen gelden voor het verleden op het moment dat de data was verzameld. Analyse van de klachten over de afgelopen 35 jaar van de Nationale Ombudsman resulteerde in

Mannen die goed opgeleid zijn en in stedelijke gebieden wonen → hebben een grotere kans om een klacht in te dienen

Over de interpretatie van Data-analytics resultaten

Statistische waarheden:

- Vereenvoudiging: Jones heeft 80% kans om betrokken te zijn bij een auto-ongeluk

Subjectieve benadering: p = gekwantificeerd oordeel

- Voorgaande kans (mogelijk inclusief "frequentistische benadering")
- Interpretatie verschillend voor ontvanger en kans genererende eenheid.

Casus: Leefbaarheid in Haagse wijken

Niels Netten

Casus - inleiding

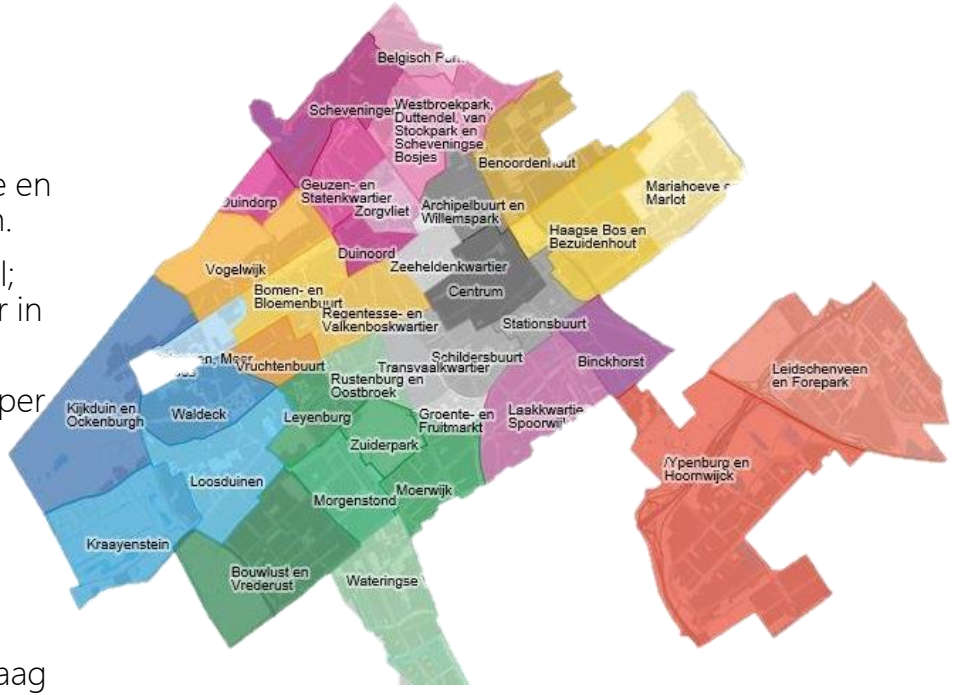
De gemeente Den Haag heeft de ambitie om kennis-gedreven te werken door data, informatie en kennis over en uit de wijken bij elkaar te brengen.

De mening van de bewoners speelt een grote rol; wat is belangrijk, wat gaat goed, wat kan er beter in de wijk?

De fysieke en sociale leefbaarheid verschilt vaak per wijk:

- Zelfredzaam wijken;
- Aandachtswijken;
- Kantelwijken

Naast de bewoners zelf, wil de gemeente Den Haag ook de beschikbare databronnen benutten om zo een completer beeld te vormen van de situatie in de wijken.



De 44 Haagse wijken (CBS-indeling)

Casus - datasets

Dataselectie van de gemeente Den Haag m.b.t. de dimensies handhaving, veiligheid en leefbaarheid;

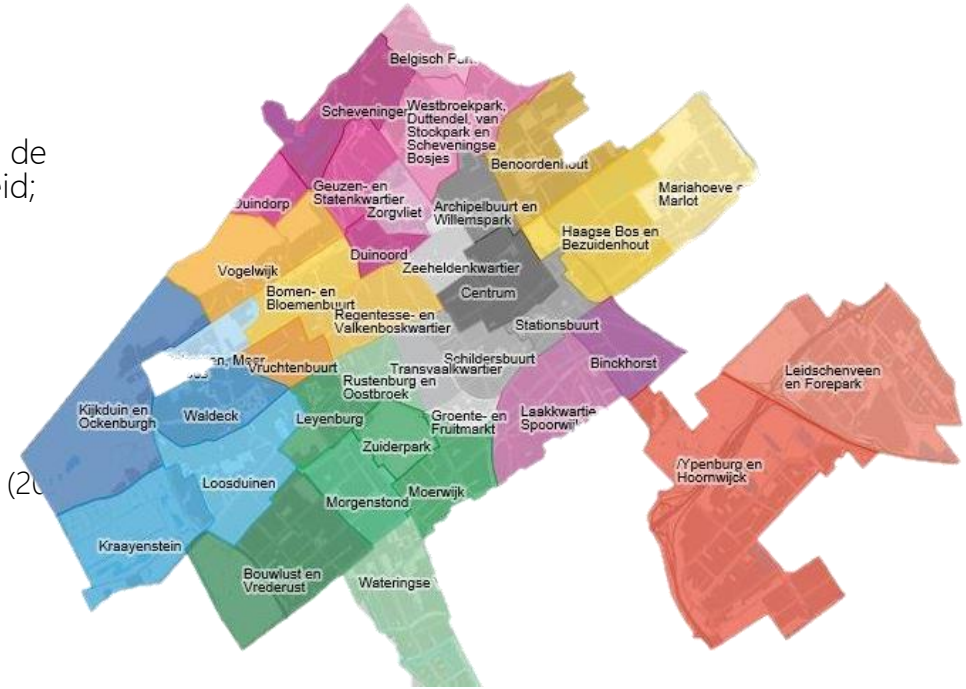
Alleen gestructureerde data;

Handhaving

- Bonnen en waarschuwingen (2014-2016)

Leefbaarheid

- Groen → Bomen
- Leefbaarometer → Wijk stand/ontwikkeling (2
- Voorzieningen (buurtmonitor Den Haag)
- Parkeerdrukte
- Maatschappelijke index (2012-2016)
- Wijkcodes NL
- Veiligheid → Veiligheidsmonitor (2015)



De 44 Haagse wijken (CBS indeling)

Casus – ongestructureerde data

Voorbeelden bronnen:

- Klachtenmeldsysteem

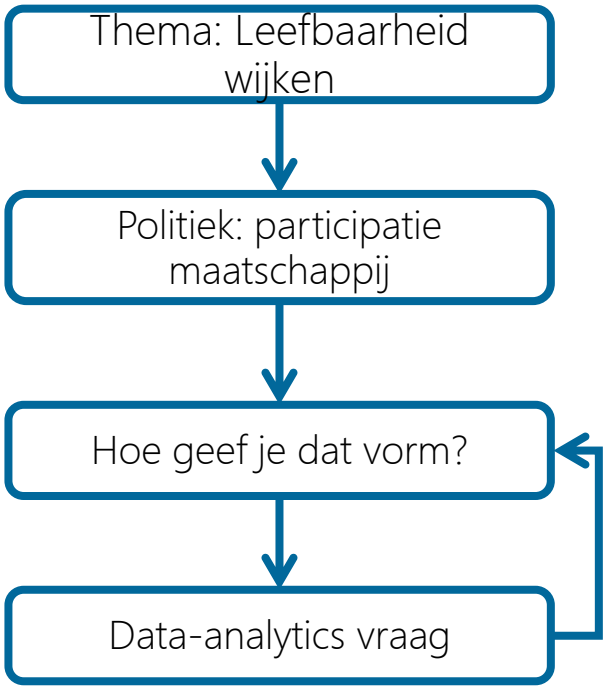
'In mijn straat staan 2 bomen die erg veel bladeren verliezen in de winter en dat waait in mijn tuin. Ik heb vrijdag 4 volle zakken weggegooid en nu ligt het weer vol. Er zijn maar een paar mensen die ook opruimen, maar alles op de hoek van de tuin vegen. Vandaag kwam een veegwagen van jullie die, zo lijkt het, alles in mijn tuin heeft geveegd/geblazen. **Graag een oplossing, de boom is niet van mij, maar ik heb wel de troep.**

- Social media (Twitter, Facebook, enz.)

N.B: ongestructureerde data buiten scope van deze casus.

Casus - vraagstellingen

- Wat verstaat u onder leefbaarheid?
- Definieer aan de hand van de bronnen aandachtswijk, kantelwijk en zelfredzame wijk.
- Wat voor inzicht zou u willen hebben m.b.t. de drie typen wijken?
- Gegeven de datasets wat is uw inschatting dat u dat inzicht kunt verkrijgen uit de datasets?
- Tot welke data-analytics vraagstukken leiden de antwoorden op bovenstaande vragen?



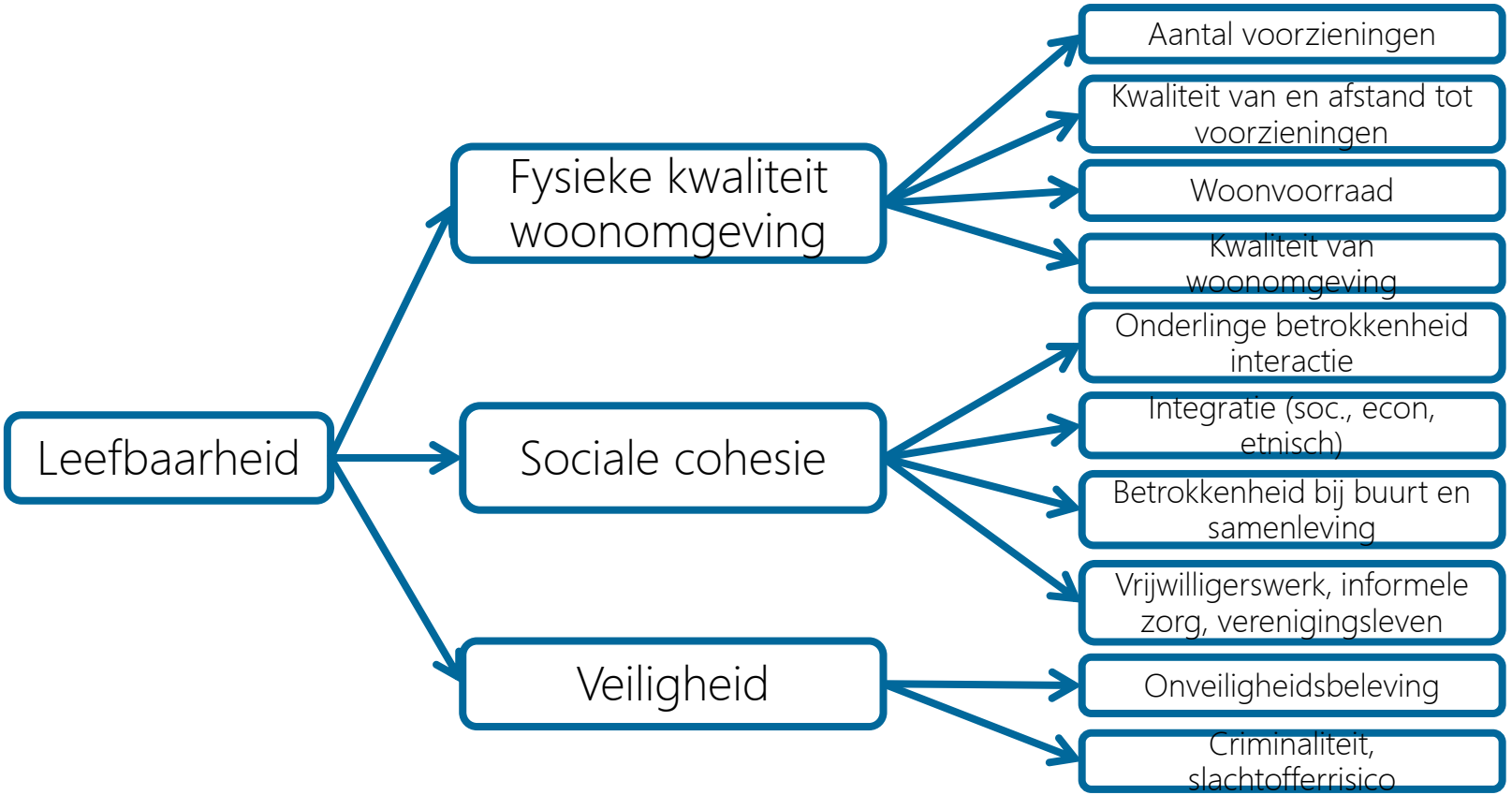
Werken aan de casus

Mogelijke invulling casus

Stappen

1. Operationaliseren Leefbaarheid
2. Definitie van kantel/aandachtswijken
3. Data-analytics vraag formuleren.

1. Operationaliseren: leefbaarheid



2. Definieer kantel/aandachtswijk

Gebruik bijvoorbeeld de veiligheidsmonitor voor inzicht in tevredenheid/ontevredenheid en leid wijktype af.

WIJKCODE	WIJKNAAM	Wijktype
WK051801	Wijk 01 Oostduinen	Z
WK051802	Wijk 02 Belgisch Park	Z
WK051803	Wijk 03 Westbroekpark en Duttendel	Z
WK051804	Wijk 04 Benoordenhout	Z
WK051805	Wijk 05 Archipelbuurt	Z
WK051806	Wijk 06 Van Stolkpark en Scheveningse B	K
WK051807	Wijk 07 Scheveningen	K
WK051808	Wijk 08 Duindorp	K
WK051809	Wijk 09 Geuzen- en Statenkwartier	Z
WK051810	Wijk 10 Zorgvliet	K
WK051811	Wijk 11 Duinoord	K
WK051812	Wijk 12 Bomen- en Bloemenbuurt	K
WK051813	Wijk 13 Vogelwijk	Z
WK051814	Wijk 14 Bohemen en Meer en Bos	Z
WK051815	Wijk 15 Kijkduin en Ockenburgh	A
WK051816	Wijk 16 Kraayenstein en de Uithof	Z
WK051817	Wijk 17 Loosduinen	A
WK051818	Wijk 18 Waldeck	Z
WK051819	Wijk 19 Vruchtenbuurt	Z
WK051820	Wijk 20 Valkenboskwartier	Z
WK051821	Wijk 21 Regentessekwartier	Z
WK051822	Wijk 22 Zeeheldenkwartier	Z
WK051823	Wijk 23 Willemspark	Z
WK051824	Wijk 24 Haagse Bos	Z
WK051825	Wijk 25 Mariahoeve en Marlot	Z
WK051826	Wijk 26 Bezuidenhout	Z
WK051827	Wijk 27 Stationsbuurt	A
WK051828	Wijk 28 Centrum	Z
WK051829	Wijk 29 Schildersbuurt	A
WK051830	Wijk 30 Transvaalkwartier	A

3. Welke data is beschikbaar

- Voorzieningen (per wijk)
  - Afstanden tot scholen
  - Afstanden tot zwembad
  - Aantal winkels
  - …
- Maatschappelijk index (per wijk)
  - Verdeling op leeftijd
  - Verdeling op inkomen
  - Aantal woningen
  - ……
- Bonnen en waarschuwingen
  - Aantal foutparkeerders
  - Aantal hondenpoep overtredingen
  - …
- Bomen
  - Aantal bomen
  - Aantal jonge bomen
  - Aantal oude bomen
  - ….

WIJKCODE	WIJKNAAM	afstand_tot attractie	Afstand tot bioscoop	Afstand tot zwembad	aantal_overlast jeugd	aantal_overlast drugs	aantal_overlast_ personen	aantal_overlast_ geluid	aantal_verkoopunte n	...	...	Wijktype
WK051801	Wijk 01 Oostduinen	-	-	-	1	0	3	0				Z
WK051802	Wijk 02 Belgisch Park	1,1	0,7	1,1	13	3	67	125	123			Z
WK051803	Wijk 03 Westbroekpark en Duttendel	0,9	1,7	1,7	22	0	15	14	1			K
WK051804	Wijk 04 Benoordenhout	2,1	2,8	2,6	44	3	56	61	89			K
WK051805	Wijk 05 Archipelbuurt	1,3	1,8	2,9	46	6	54	50	40			K
WK051806	Wijk 06 Van Stolkpark en Scheveningse Bo	1,3	1,4	1,1	2	0	5	1	2			Z
WK051807	Wijk 07 Scheveningen	1,2	1,1	0,9	163	9	279	376	187			Z
WK051808	Wijk 08 Duindorp	3,5	2,2	3	44	3	48	86	9			Z
WK051809	Wijk 09 Geuzen- en Statenkwartier	2,3	1,2	2	109	1	62	93	154			Z
WK051810	Wijk 10 Zorgvliet	2,2	0,6	2,6	14	0	15	4				Z
WK051811	Wijk 11 Duinoord	2,9	1	2,7	43	8	63	84	54			Z
WK051812	Wijk 12 Bomen- en Bloemenbuurt	4,1	2,1	2,1	209	11	77	178	148			Z
WK051813	Wijk 13 Vogelwijk	4,4	2,4	2,3	71	2	36	56	3			Z
WK051814	Wijk 14 Bohemen en Meer en Bos	4,2	4,4	1,2	35	1	39	42	11			Z
WK051815	Wijk 15 Kijkduin en Ockenburgh	3,9	5,3	2,4	7	1	28	27	32			Z
WK051816	Wijk 16 Kraayenstein en de Uithof	2	7	3,1	24	6	34	38	10			Z
WK051817	Wijk 17 Loosduinen	2,6	5,4	2	129	11	115	122	105			A
WK051818	Wijk 18 Waldeck	3,3	5	1,3	134	12	120	213	68			K
WK051819	Wijk 19 Vruchtenbuurt	4,5	3,4	0,8	92	5	47	129	62			K
WK051820	Wijk 20 Valkenboskwartier	3,9	2,3	2,1	87	23	151	400	121			K
WK051821	Wijk 21 Regentessekwartier	3,2	1,7	2	85	20	165	232	119			K
WK051822	Wijk 22 Zeeheldenkwartier	2,6	1,2	2	48	9	111	277	135			K
WK051823	Wijk 23 Willemspark	1,8	1,5	2,7	23	0	21	37	54			K
WK051824	Wijk 24 Haagse Bos	3,4	1,9	1,5	4	0	13	4	3			K
WK051825	Wijk 25 Mariahoeve en Marlot	5	3,8	1,4	91	21	135	320	38			K
WK051826	Wijk 26 Bezuidenhout	3,2	2	1	72	11	213	186	120			K
WK051827	Wijk 27 Stationsbuurt	1,6	0,8	1,5	77	20	250	274	118			A
WK051828	Wijk 28 Centrum	2,3	0,6	1,2	194	43	824	669	857			K
WK051829	Wijk 29 Schildersbuurt	1,6	1,6	0,8	453	34	418	481	137			A
WK051830	Wijk 30 Transvaalkwartier	2,3	2,2	1,3	209	17	342	350	150			A
WK051831	Wijk 31 Rustenburg en Oostbroek	3,1	2,8	1,8	166	37	196	452	118			K
WK051832	Wijk 32 Leyenburg	3,6	3,8	1,6	175	13	118	213	46			Z
WK051833	Wijk 33 Bouwlust en Vrederust	2	6,3	2,9	229	27	278	441	68			K
WK051834	Wijk 34 Morgenstond	3,2	4,8	1,5	136	24	310	438	171			Z
WK051835	Wijk 35 Zuiderpark	3,1	3,6	1,2	9	0	19	3				K
WK051836	Wijk 36 Moerwijk	2,7	3,9	1,5	161	24	246	407	76			K
WK051837	Wijk 37 Groente- en Fruitmarkt	1,9	2,7	1,5	45	6	50	63	25			Z
WK051838	Wijk 38 Laakkwartier en Spoorwijk	1,3	2,4	2,4	352	51	488	960	247			K
WK051839	Wijk 39 Binckhorst	1,8	2,2	3,2	8	5	30	12	21			Z
WK051840	Wijk 40 Wateringse Veld	2	6,3	2,5	205	12	50	154	41			K
WK051841	Wijk 41 Hoornwijk	0,8	4,3	2,6	16	7	18	8	10			Z
WK051842	Wijk 42 Ypenburg	3,6	4,1	2,2	216	26	102	227	32			K
WK051843	Wijk 43 Forepark	3	5,6	2,9	10	1	13	6	11			K
WK051844	Wijk 44 Leidschenveen	4,9	7	2,9	182	5	62	144	23			Z

4. Data analytics vraag

- Data-analytics vraag:

  wat zijn de onderscheidende kenmerken van kantelwijken of aandachtswijken op het gebied van leefbaarheid?

- Profiling → zoek de onderscheidende kenmerken die te maken hebben met de leefbaarheid van een kantelwijk of aandachtswijk.