
ONS VERHAAL: OP WEG VAN DATAVERZAMELING NAAR DATA SCIENCE
INNOVATIEVE DATAVERZAMELING, MACHINE LEARNING EN DATA SCIENCE BINNEN DE SOCIAL SCIENCES
LENNARD KUIJTEN – CENTERDATA & JADS

EVEN VOORSTELLEN

- LENNARD KUIJTEN
- STICHTING CENTERDATA – TILBURG, NEDERLAND
  - INSTITUUT VOOR DATAVERZAMELING EN ONDERZOEK
  - WETENSCHAPPELIJK – STICHTING – 35 MENSEN
  - GELIEERD AAN TILBURG UNIVERSITY – ZITTEN OP DE CAMPUS
  - BESTAAT 20 JAAR
- WETENSCHAPPELIJK SOFTWARE ONTWIKKELAAR
  - WETENSCHAPPELIJK OPGELEID
  - SOFTWARE, CODE, ALGORITMIEK, APPLICATIES, APPS, DATABASES, API'S, E.D.
  - FACILITEREND EN ONDERSTEUNING IN ONDERZOEK – AUTOMATISERING
- JADS

CENTERDATA



- DIVERSE KERNEN
  - SURVEY ONDERZOEK – BELEIDSONDERZOEK EN PROGNOSES – ICT
  - CONSUMENTENONDERZOEK – DATA SCIENCE
- PANELS
  - LISS PANEL – CENTERPANEL
- WETENSCHAPPERS EN IT'ERS – DIR. PROF. MARCEL DAS
- ONDERZOEK – PUBLICATIES – DATADISSEMINATIE
  - O.A. WWW.LISSDATA.NL (MESS) / WWW.DHSDATA.NL / SHARE-PROJECT.CENTERDATA.NL
- NATIONALE EN INTERNATIONALE SAMENWERKING
  - EU PROJECTEN – NATIONALE EN INTERNATIONALE INFRASTRUCTUUR SOCIAL SCIENCES

WAAR NIET OVER HEBBEN



- BIG DATA
  - 3 V'S, 4 V'S, 7 V'S, .. 8 V'S, 6 D'S, ..
- ELKE 2 JAAR VERDUBBELT HOEVEELHEID DATA
  - MOORE'S LAW HARDWARE OOK VAN TOEPASSING OP HOEVEELHEID DATA
  - ..EXPONENTIEEL
- PETABYTES, EXABYTES, ZETTABYTES, YOTTABYTES, ETC.
- DATA VERZAMELEN NIET MEER HET PROBLEEM
  - VAKGEBIED.. ASTRONOMIE, GENETICA, MANUFACTURING, ONLINE, ETC.
  - SOCIAL SCIENCES – GEDRAGSWETENSCHAPPEN – SURVEYS
- PRIVACY EN WETGEVING EN DATA
- VAN SUBJECTIEF NAAR OBJECTIEF

WAAR WEL OVER HEBBEN

- ONZE ‘QUEST’: WERKEN MET SUBJECTIEVE DATA NAAR ÓOK OBJECTIEVE DATA
- VOORBEELDEN INNOVATIEVE DATAVERZAMELING LISS PANEL
  - 2011
- NIEUWE TECHNIEKEN EN METHODEN
  - TECHNISCH
  - (NIET ONDERZOEKSRESULTATEN)
  - DATA ENGINEER – DATA SCIENTIST
- OP WEG NAAR TOEGEPASTE DATA SCIENCE
  - SOCIAL SCIENCES
  - ANALYSES
  - EXPLORATIEF – JADS

CENTERDATA – LISS PANEL

- LISS PANEL – MESS PROJECT – SINDS 2007 – NWO
  - 8.000 PANELLEDEN (>= 16 JAAR)
  - 5.000 HUISHOUDENS
  - PROBABILITY SAMPLE NEDERLANDSE BEVOLKING (CBS)
  - INCENTIVES (~15 EURO / UUR)
  - (~30MIN) MAANDELIJKS VRAGENLIJST(EN)
  - LONGITUDINALE KERNVRAGENLIJSTEN
  - > 200 SPECIFIEKE ONDERZOEKSPROJECTEN (WWW.LISSDATA.NL)
  - > 450 PUBLICATIES
- CENTERPANEL – (O.A. DNB HOUSEHOLD SURVEY)
- AL ONZE KERNEN VERZAMELEN STEEDS MEER (EN OOK COMPLEXE) DATA



INNOVATIEVE DATAVERZAMELING

- SELF-REPORTED + OBJECTIEVE DATA
- SINDS 2011
- EYE-TRACKING
- INTERNET WEEGSCHALEN – DRAADLOOS NAAR ROUTER
- ACCELEROMETERS
- SMARTPHONES – APPS – GPS – UI-I – BROWSER DATA
- WEBCAM – GEZICHTSUITDRUKKING – EMOTIES – MICRO-EXPRESSIES
- (2011) SHARE BIOMETRIE (MARKERS & GRIP STRENGTH)

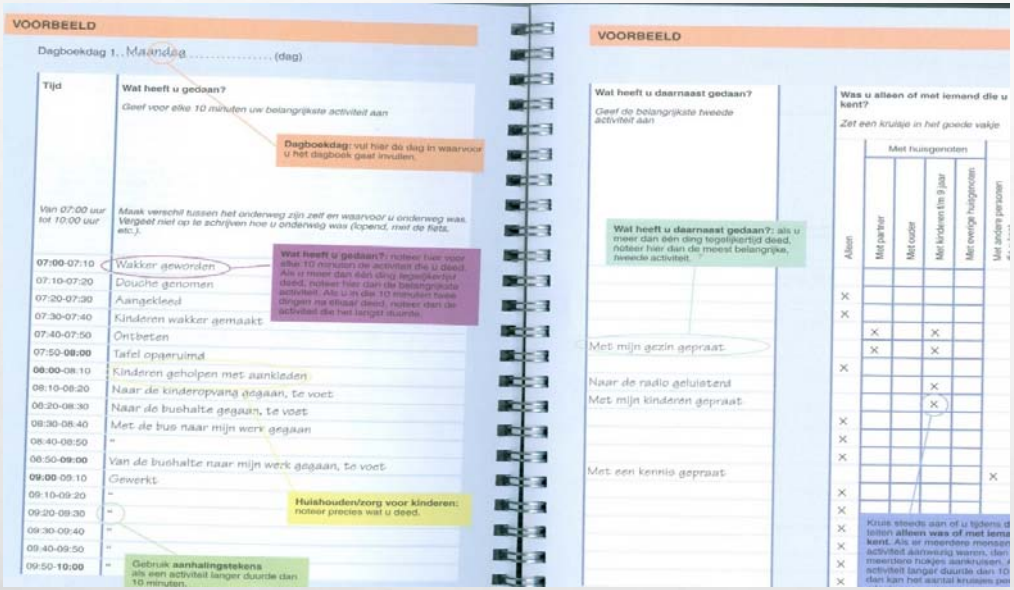


MEER DATA – ANDERE DATA

- MEER DATA – ANDERE DATA
  - MB → GB → TB
  - NIET MEER OPENEN / REKENEN
- COMPLEXE DATA
  - SENSOR- EN SIGNAALDATA (I.O.T.)
  - GEORGRAFISCH
  - TEXTUEEL (VRIJE INVOER)
- RUW – GRANULARITEIT
  - NIET EENVOUDIG TE STRATIFICEREN
- PRIVACY WAARBORGING (GPS)



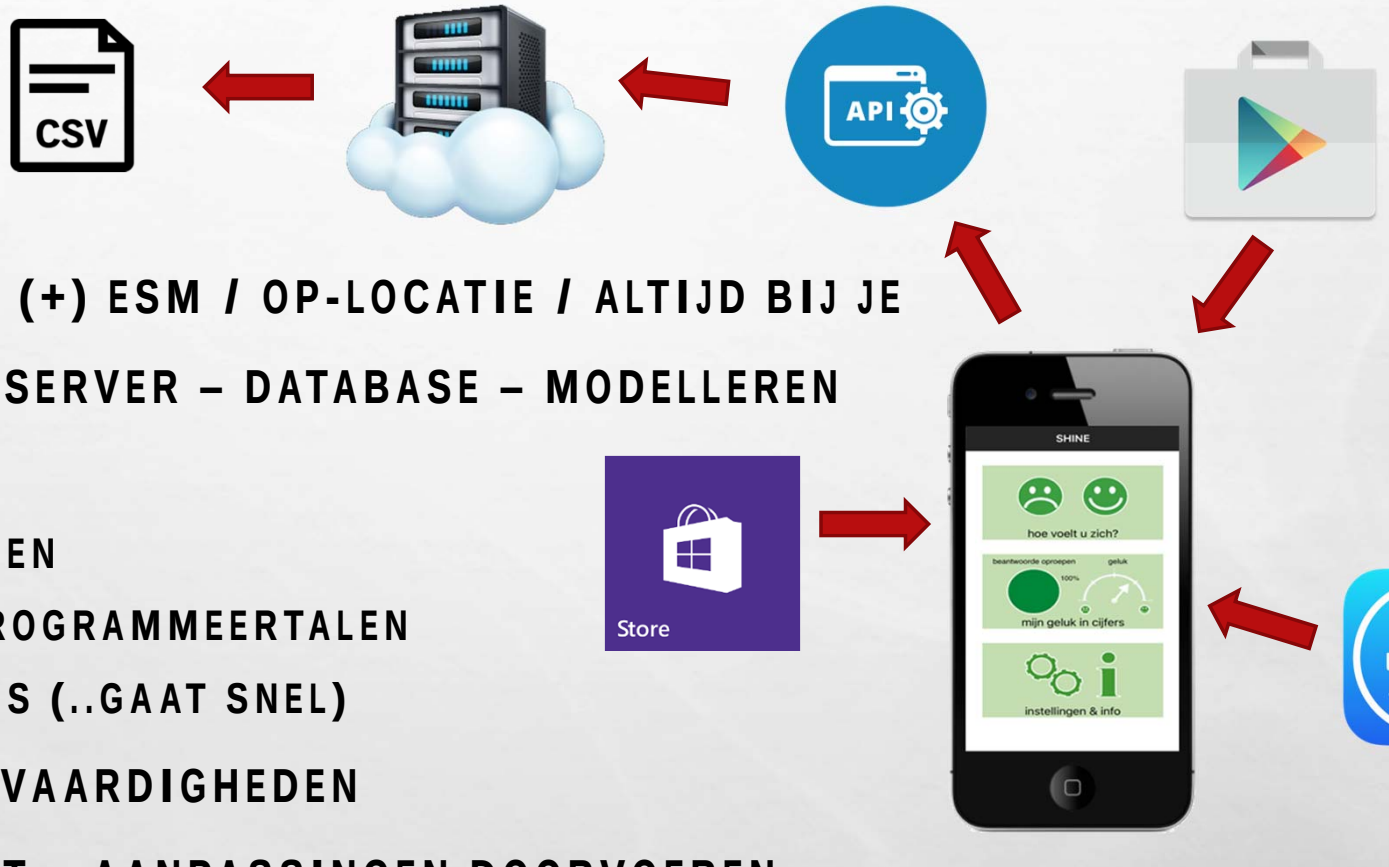
TIJDSBESTEDINGSONDERZOEK





- REAL-TIME
- API

APPS



- (6-TAL APPS)
- SOCIAL SCIENCES: (+) ESM / OP-LOCATIE / ALTIJD BIJ JE
- BACKEND: API'S – SERVER – DATABASE – MODELLEREN
- FRONEND: APPS
  - UIX – PLATFORMEN
  - PAKKETTEN & PROGRAMMEERTALEN
  - MOBIELE VERSIES (..GAAT SNEL)
- SPECIALISTISCHE VAARDIGHEDEN
- DOORLOOP TRAJECT – AANPASSINGEN DOORVOEREN
  - ONDERZOEKSOPZET – SENSOREN – TESTEN – TESTERS – APPSTORES – REGELS

WEGEN = WETEN

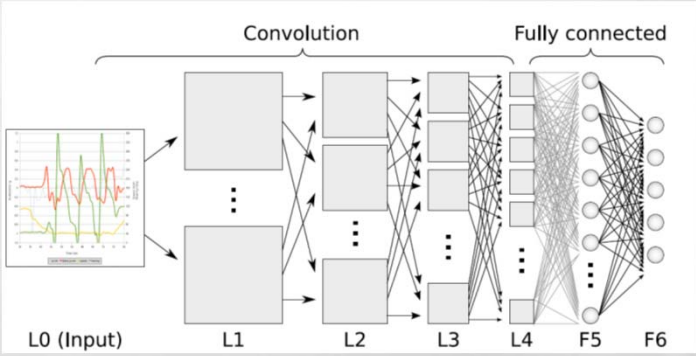
- 1.250 WEEGSCHALEN
- STUDIE 3 JAAR
- TENMINSTE 1 JAAR IN HUIS
- ‘DAGELIJKS’ METINGEN
- GEHELE HUISHOUDEN
- AUTOMATISCHE DETECTIE – WIE
- STROMENDE DATA – DIRECTE FEEDBACK
  - WEEGSCHAAL > DATABASE > ANALYSE > PANELLID LOGIN PAGINA
- (-/- 0.8KG, +/+ 0.2KG ..)

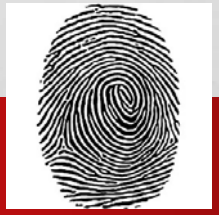


BE-WEGEN = WETEN



- ACCELEROMETRIE – 60 HZ & 10-12 DAGEN
- 886 PANELLEDEN
- 886 X ~62 MILJOEN RECORDS = ~ 55 MILJARD RECORDS MET N-DATAPUNTEN
- CSV ~3.6 TB
- 886 LOSSE 4GB BESTANDEN, OOK NIET TE DOEN
- BINAIRE OPSLAG (~900 GB) – MATLAB (~1.4 TB)
- EPOCH DATA
- NIET GELABELDE DATA (!)
- UNSUPERVISED LEARNING – SUPERVISED LEARNING – DEEP LEARNING (~48 BILJOEN INVOER)





EXPONENTIELE ONTWIKKELING



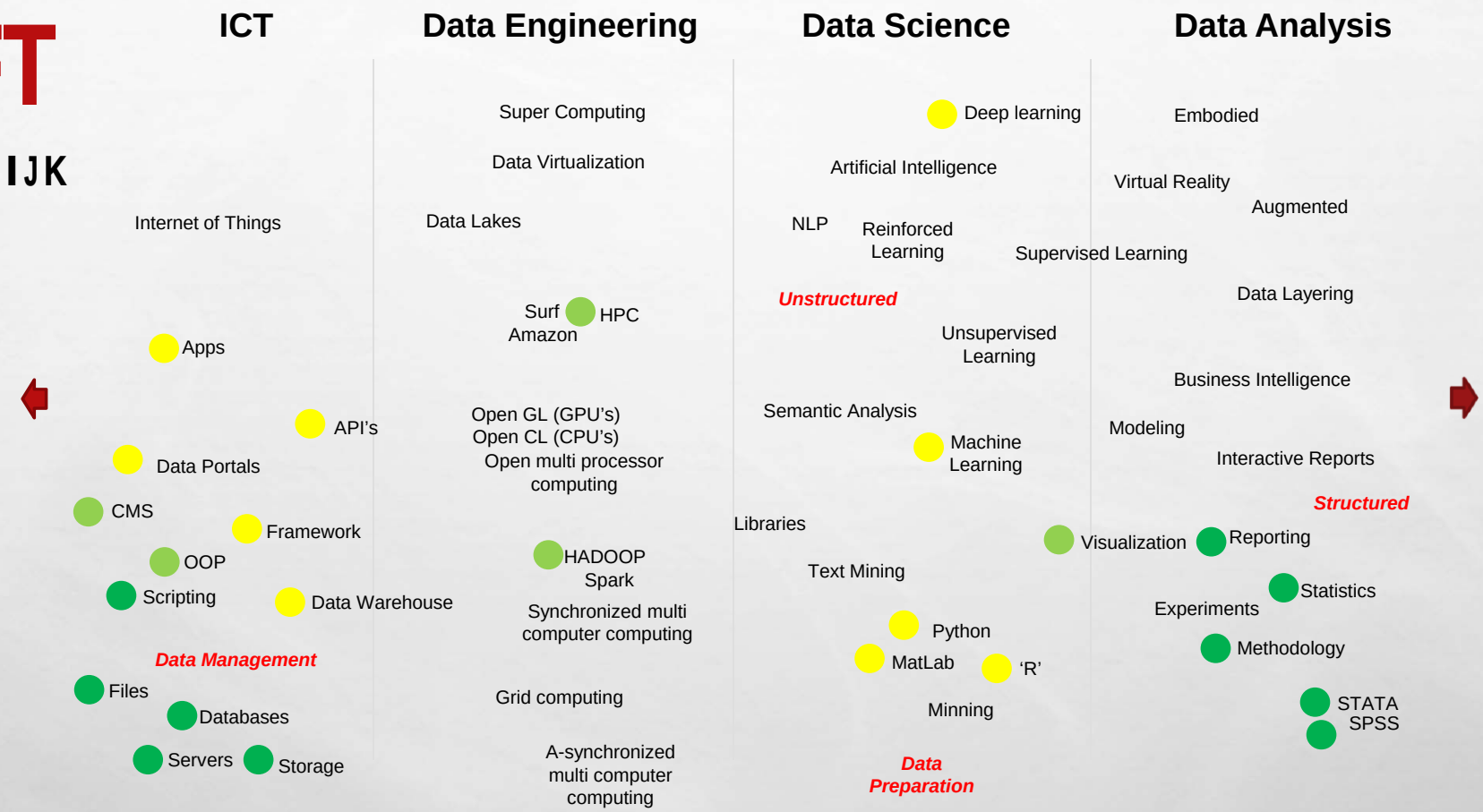
ANDERS – SURVIVAL OTF



- MEER VERZAMELEN?
- PLATSLAAN? (EPOCH – 60 HZ > 1.667^{E}-2 HZ)
- REPRESENTEREN (GPS / ACTIVITEIT / VERPLAATSING)
- VERLIES WAARDEVOLLE KENNIS?
- UITDAGING, NIET BEHAPBAAR MAKEN: A.I. – DATA SCIENCE – MULTI-DIMENSIONEEL
- ANDERE INFRASTRUCTUUR
  - OPSLAG (GEDISTRIBUEERD)
  - REKENEN GEDISTRIBUEERD (GRID, HADOOP, SPARK)
  - REKENCAPACITEIT (OPEN-CL, OPEN-MV, GPU CLUSTERS, HPC)
- ANDERE TOETSING > PLUG & PLAY MACHINE LEARNING
  - SCRIPTING, ALGORITMIEK, SIGNAALDATA, NLP, MACHINE LEARNING, DEEP LEARNING

SHIFT

- GELEIDELIJK
- NIEUWE MENSEN

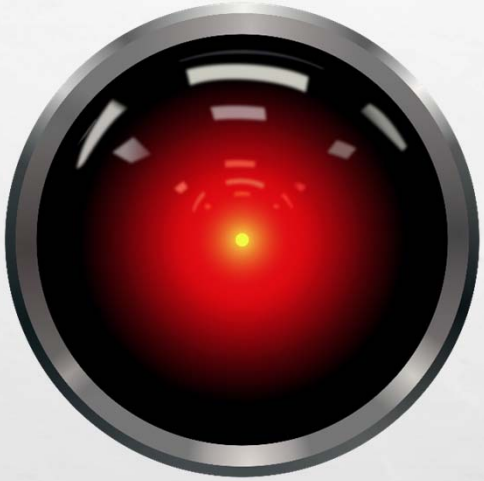


EXPLORATIE EN NIEUWE TOEPASSINGEN

- JADS: M-C-Q GENERATOR
  - ZINNEN – GATEN – DISTRACTOREN
- JADS: MICROFINANCIERING (1.3M FOTO'S GEZICHTEN)
- TWITTE-R (VERKIEZINGSDATA)
  - HARVESTEN – API
  - TEXT MINING – SEMANTISCHE ANALYSE (NLP)
  - SENTIMENT ANALYSE
  - NETWERK DATA – CLUSTEREN (NIET HANDMATIG > A.I.)
  - VISUALISATIE: NETWERK – CHORD CHARDS – GRAFEN
- JADS: SURFSARA CLOUD (VMS) EN LISA & CARTESIUS
- JADS: BIG (OPEN) DATA INFRASTRUCTUUR EN DATA VIRTUALISATIE



PRNTLN : $.VAR = { “EINDE”};



L.KUIJTEN@UVT.NL / WWW.CENTERDATA.NL / WWW.JADS.NL

BATHROOM SCALE STUDY

- WIRELESS BATHROOM SCALES
- USES RADIO SIGNALS THROUGH "BRIDGE" AND INTERNET TO SEND THE DATA
- MEASURES WEIGHT AND BODY FAT
- VARIABILITY OVER TIME: MORE ACCURATE AND MORE FREQUENT



DESIGN STUDY

- LONGITUDINAL STUDY (APPROX. 3 YEARS)
- 950 SCALES, IN LAST YEAR OF STUDY 300 NEW SCALES ADDED
- ***LEND TO HOUSEHOLD FOR AT LEAST 1 YEAR***

BATHROOM SCALE STUDY

- FIRST EMPIRICAL ANALYSIS IS BASED ON ALMOST 80,000 MEASUREMENTS COLLECTED IN 2011
- SUBJECTIVE SELF-REPORTS VERSUS OBJECTIVE MEASUREMENTS?

Average self-reported weight is 0.9 kilograms lower than average actual weight for men and 0.7 kilograms lower for women



MORE RESULTS

- KOOREMAN, P. AND A. SCHERPENZEEL (2014). "HIGH FREQUENCY BODY MASS MEASUREMENT, FEEDBACK, AND HEALTH BEHAVIORS," *ECONOMICS & HUMAN BIOLOGY*, 14, PP 141-153.
- HIGHLIGHTS
- WEIGHT IS 0.2 KILOGRAMS LOWER ON FRIDAYS THAN ON MONDAYS; BMI IS 0.06 LESS AND FAT PERCENTAGE IS 0.03 LOWER ON FRIDAYS.
- THE FAT-BASED MEASURE OF OBESITY INDICATES A THREE TIMES LARGER PREVALENCE OF OBESITY (53%) THAN THE BMI-BASED MEASURE (17%).
- A FEEDBACK THAT INCLUDES A RECOMMENDED WEIGHT RANGE INCREASES THE TEMPORAL VARIATION IN INDIVIDUAL BODY WEIGHT BY ABOUT 10%.

RECRUITMENT EXPERIMENT – LISS PANEL

Experimental design

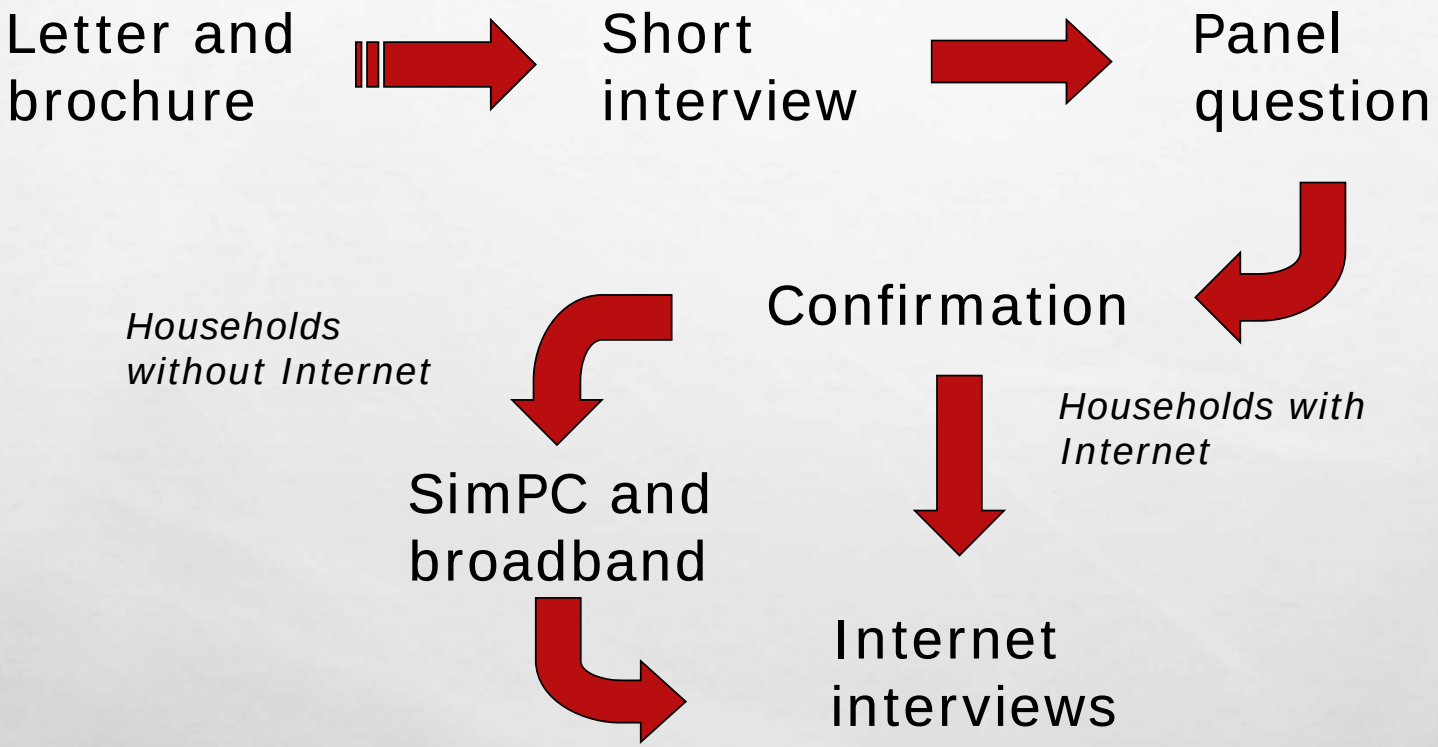
		CATI		CAPI		N
		Special letter	Standard letter	Special letter	Standard letter	
Prepaid incentive	Introduce panel in letter	**0 euro**	**10 euro**			200
	Introduce panel after interview		**20 euro**	**50 euro**		300
Promised incentive	Introduce panel in letter			**0 euro**	**10 euro**	400
	Introduce panel after interview	**50 euro**			**20 euro**	300
N		200	200	400	400	1200

Scherpenzeel, A. and V. Toepoel (2012). "Recruiting a Probability Sample for an Online Panel: Effects of Contact Mode, Incentives, and Information," *Public Opinion Quarterly* 76(3), pp 470-490.

EFFECT OF INCENTIVES

- INCENTIVES INCREASE RESPONSE RATES
- - EFFECT IS LARGE FROM 0 TO 10 EURO (INCREASE OF 15%)
- - EFFECT IS NOT SIGNIFICANT FROM 10 TO 20,
  20 TO 50 EURO (ABOUT 3%)
- PREPAID WORKS BETTER THAN PROMISED

RECRUITMENT



ATTRITION LISS HOUSEHOLDS 2008-2016

