DATA FOR SOCIAL DEVELOPMENT

OPPORTUNITIES, RISKS AND CASE STUDIES OF BIG DATA IN PUBLIC POLICY

Sofie Roehrig

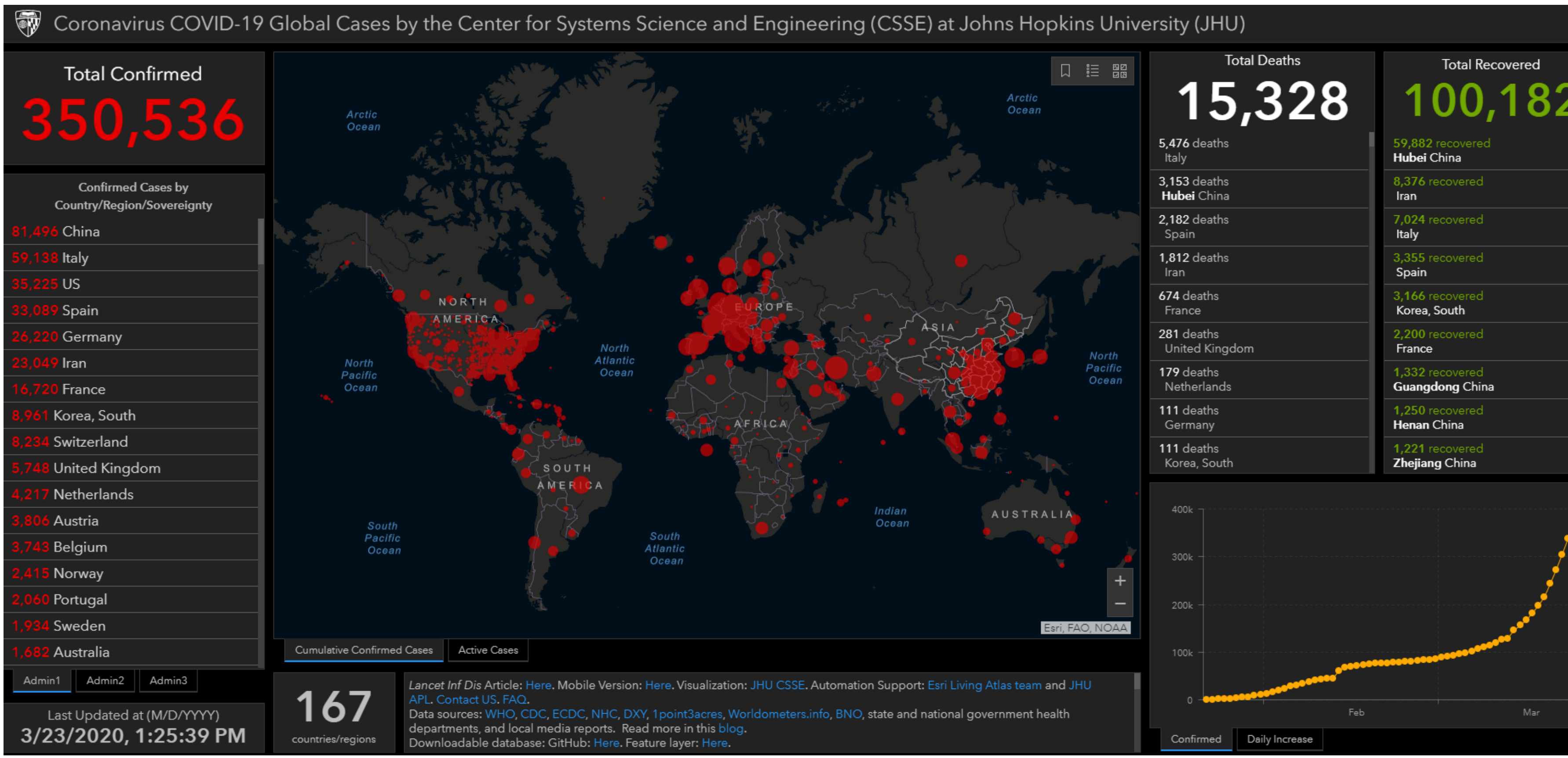
UNU-MERIT/MGSOG

2020

Maastricht, The Netherlands

1. Understanding Big Data
2. Potential of Big Data for development
3. Big Data in public policy process phases
4. Challenges and risks
5. Case studies
6. Methodology
7. Practical exercise

Big Data that we see every day (1)



Defining concepts

Big Data

- Signifies three trends: 1) growing volume of data, 2) new technologies to capture and process, 3) intention to extract insights from those (2)

Artificial Intelligence

- The simulation of human intelligence through machines, mostly through computer systems (3)

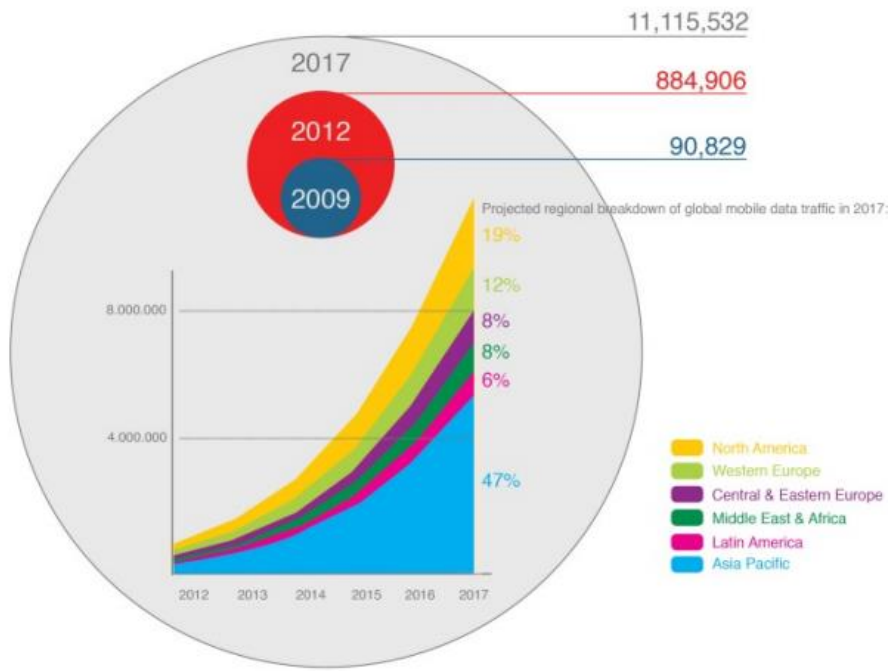
Machine Learning

- A field of AI; construction and study of computer algorithms (procedures for calculations and/or classification) that can teach themselves to grow and change when exposed to new data (2)

The rise of Big Data

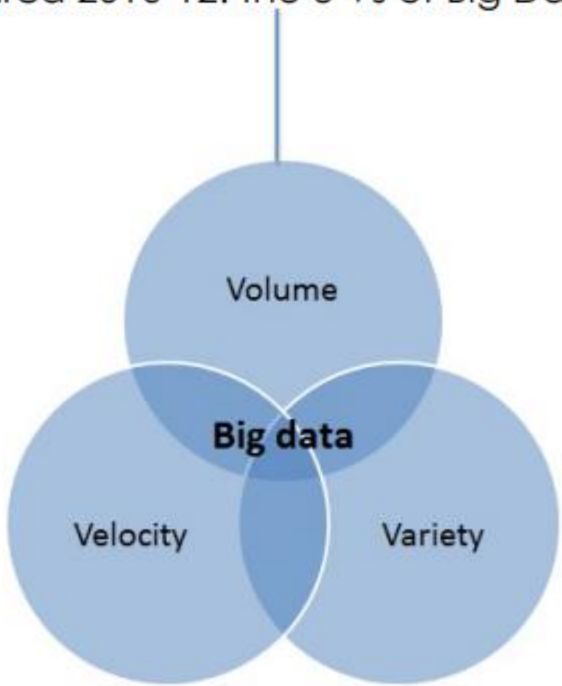
- Devices with permanent internet connection e.g. smartphones, cameras with image recognition and analytics software, sensors, IoT systems etc.) produce vast amounts of different types of electronic data (4)
- New possibilities of information collection, storage, and processing, which increase the data's availability and usability
- New methods of data analysis

Global Mobile Data - Traffic growth & forecast (terabytes per month) (5)



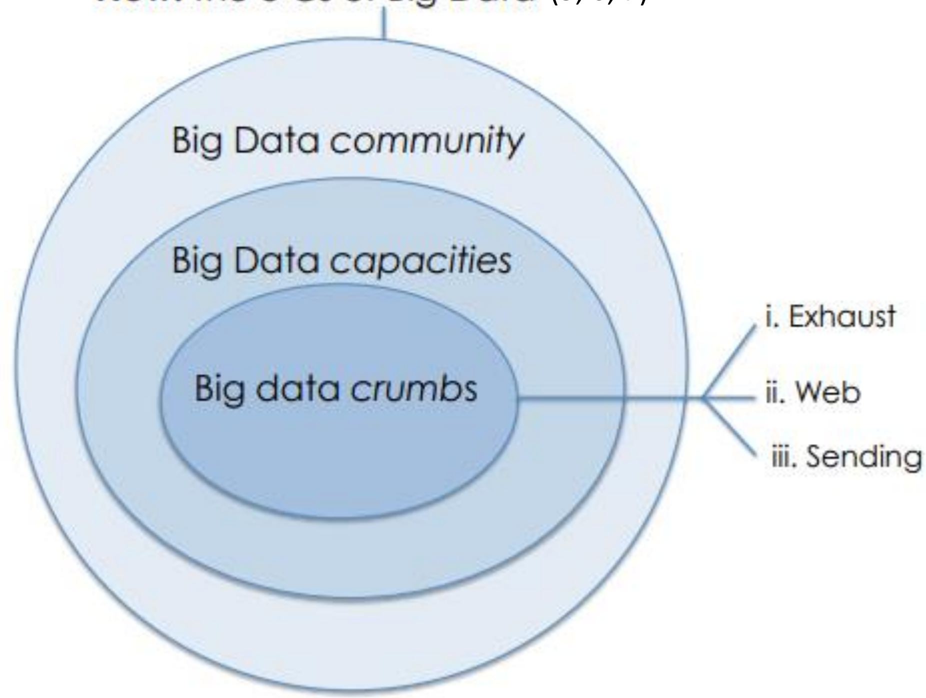
Describing Big Data

Circa 2010-12: the 3 Vs of Big Data (5)



- **Volume**: unprecedented data growth
- **Velocity**: high-speed data creation
- **Variety**: wide variety of data structures

Now: the 3 Cs of Big Data (5, 6, 7)



- **Crumbs**: passively-generated individual and networked "traces of human actions picked up by digital devices"
- **Capacities**: intent and capacity to yield insights through storage, computing and analysis
- **Community**: people and groups having access to and using the crumbs and capacities

Taxonomies of Big Data

Structure (8)

- **Structured data**: organised, clearly identifiable, e.g. database with columns and rows
- **Semi-structured data**: no formal structure but contains "tags" which separate of data records or fields
- **Unstructured data**: no identifiable structure, e.g. texts, photos, videos and audio files

Source (9)

- **Administrative data**: collected for transactional purposes, records of behaviours
- **Digital residues**: digital data series that resemble administrative data but contain a great deal of potentially useful text, image or sound information if decoded

Solidity (10)

- **Hard data**: traditionally collected administrative statistics
- **Soft data**: freely available on the internet, often subject to property rights of public or private actors

Access (11)

- **Personal and proprietary**: controlled by individual or commercial entities with rights to restrict access, e.g. personal health records or credit card information
- **Government-controlled**: government can restrict access, e.g. census data or personal tax or health records
- **Open data commons**: available to all; private, commercial or government controlled, e.g. geographic data

Size, a spectrum (8)

- **Big data**: high volume, high velocity, high variety and high complexity, e.g. linked datasets with different structures
- **Small data**: high volume, low velocity, low variety and low complexity, e.g. land use data for a small city

Big Data value chain

Data acquisition

- Structured data
- Unstructured data
- Event processing
- Protocols
- Real-time
- Data-streams
- Multi-modality

Data analysis

- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked data
- Data discovery
- Whole world semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data curation

- Data-quality
- Trust-provenance
- Annotation
- Data-validation
- Human-data interaction
- Top-down/bottom-up
- Community/Crowd
- Human computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data storage

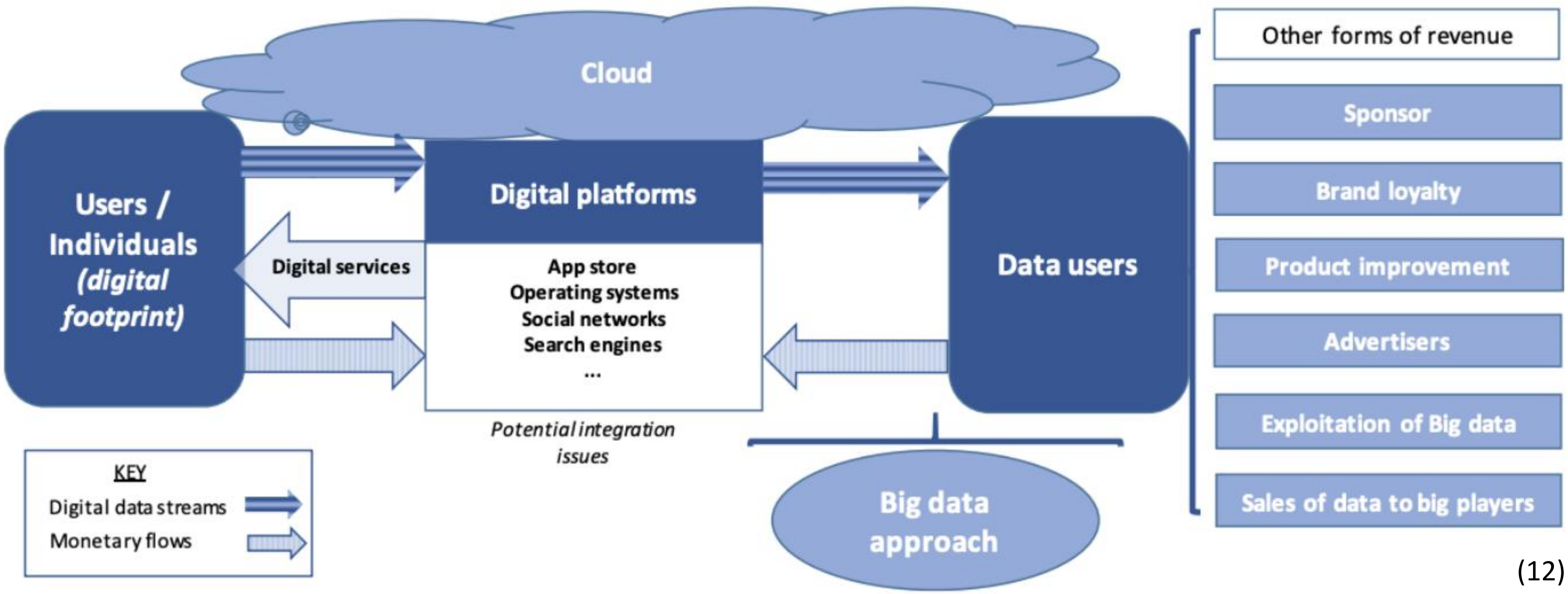
- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and performance
- Data models
- Consistency, Availability, Partition-tolerance
- Security and privacy
- Standardization

Data usage

- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualization
- Modelling
- Control
- Domain-specific usage

(12)

Digital data market and ecosystem

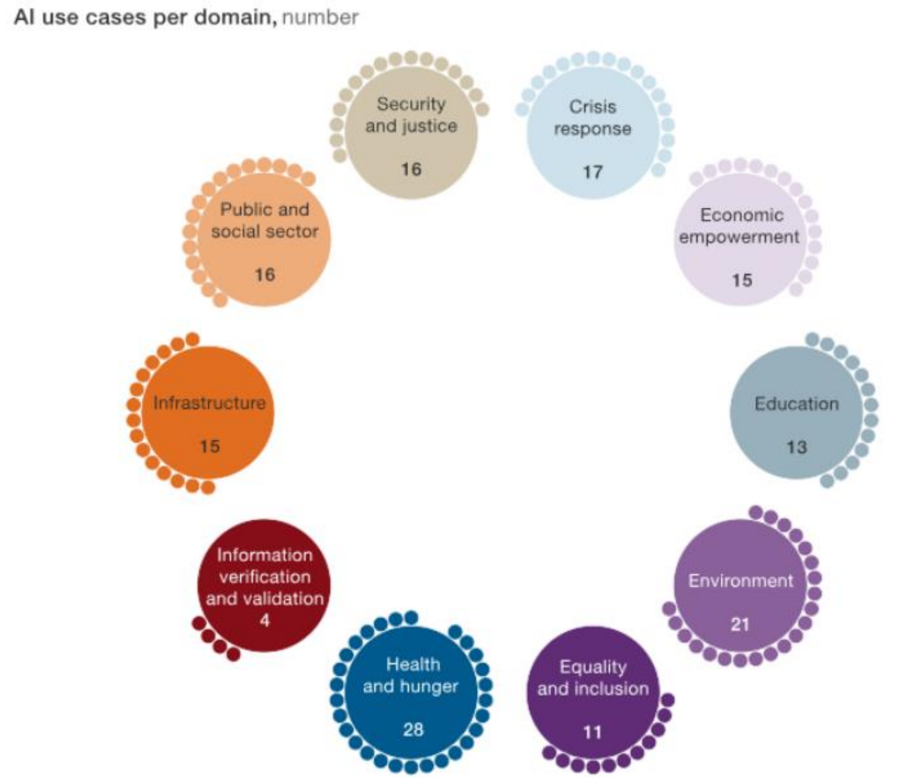


(12)

- Individuals/ groups who generate data, also data providers (12)
- Technology providers, e.g. data management platforms
- Users who use data to generate value
- Data brokers which collect data from different sources and sell it
- Research organisations and companies creating new ways to extract, process and explore data
- Public bodies which regulate data, provide data-based products or use data in their processes

The potential of Big Data and AI for social development

- "New sources of data (…), new technologies, and new analytical approaches, if applied responsibly, can enable **more agile, efficient and evidence-based decision-making** and can **better measure progress** on the Sustainable Development Goals (SDGs) in a way that is both **inclusive** and **fair**.", UN website (13)
- "Big data can **shed light** on disparities in society that were previously hidden.", UN website (13)



Note: Our library of about 160 use cases is not comprehensive and will continue to evolve. This listing of the number of cases per domain should thus not be read as exhaustive.

(14)

Data in policymaking

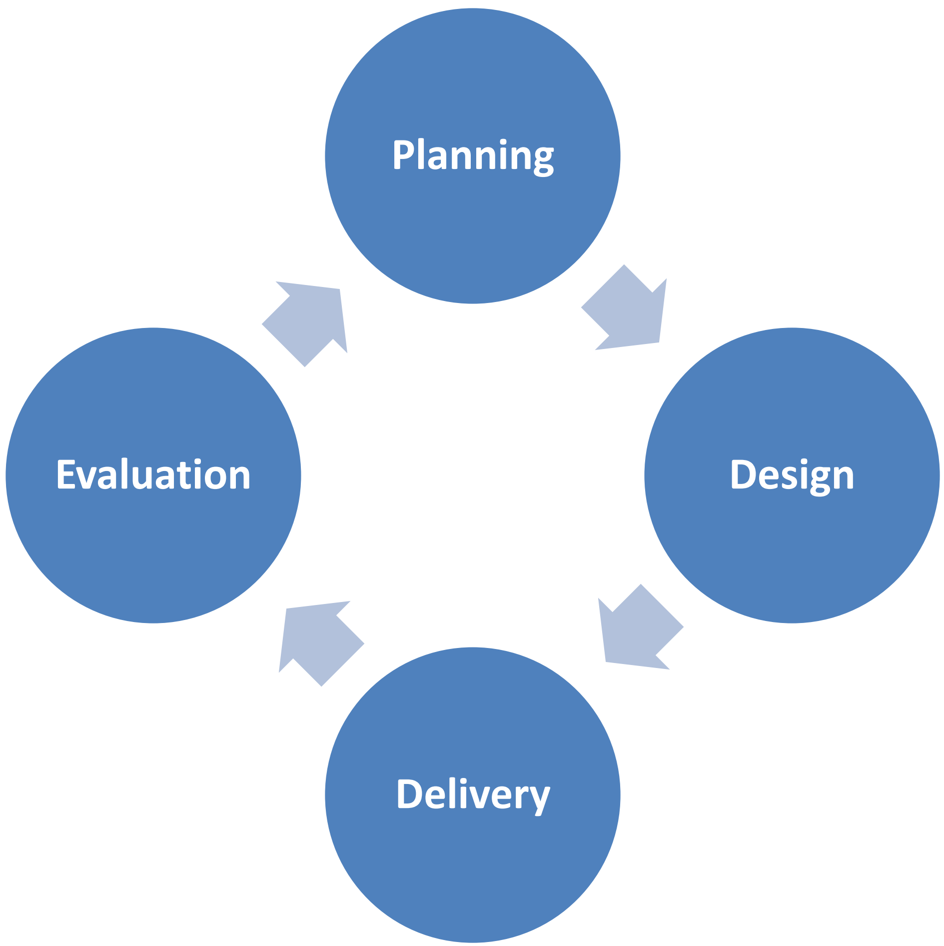
- Evidence-based policymaking (15)
  - Replacing ideologically-driven politics with rational decision making
  - A discourse or set of methods which informs the policy process
  - Rational, rigorous and systematic approach
  - Minimize policy failures caused by a mismatch between government expectations and actual, on-the-ground conditions

- Policy-analytics (16)
  - Development and application of skills, methodologies, methods and technologies, which aim to support relevant stakeholders engaged at any stage of a policy cycle, with the aim of facilitating meaningful and informative hindsight, insight and foresight
    - Meaningful – relevant and adding value to the process
    - Operational – practically feasible
    - Legitimating – ensuring transparency and accountability

Needs and trends in European Public Administrations

- Strategical needs (17)
  - Development of domain specific target and indicator systems
  - Involvement of the public and citizens, as well as the development of citizen-centred policy making
  - Strengthen citizens' trust in public administration
  - Continuous evaluation of policies
  - Take into account local and regional specificities
  - Cross-linked information exchange
- Informational needs
  - Link between impact, quality, performance measurements and financial information
  - Ensure availability of (real-time) information and knowledge

Big Data in the public policy process



In the policymaking context, big data is "the creative application of large transactional data sets generated by the internet (such as comments on social media) to the processes of policymaking" (18, p.221)

UNITED NATIONS UNIVERSITY
UNU-MERIT

Maastricht University

Public policy process – Big Data for policy planning

- For policy planning, big data can add value during (4)
  - Agenda-setting, problem definition, policy discussion, and citizen participation
- Sources / types of data
  - Especially social media: can be used for insights on citizens' policy preferences
- Techniques
  - Sentiment analysis, opinion mining, clustering, machine learning

Planning

Public policy process – Big Data for policy design

- For policy design, big data can add value during (4)
  - Policy formulation and as information-based policy instruments for evidence-based policymaking
- Sources / types of data
  - Mobility data (e.g. geodata), educational data, employer-employee microdata
- Techniques
  - Predictive analytics, network science, data visualization, scenario techniques

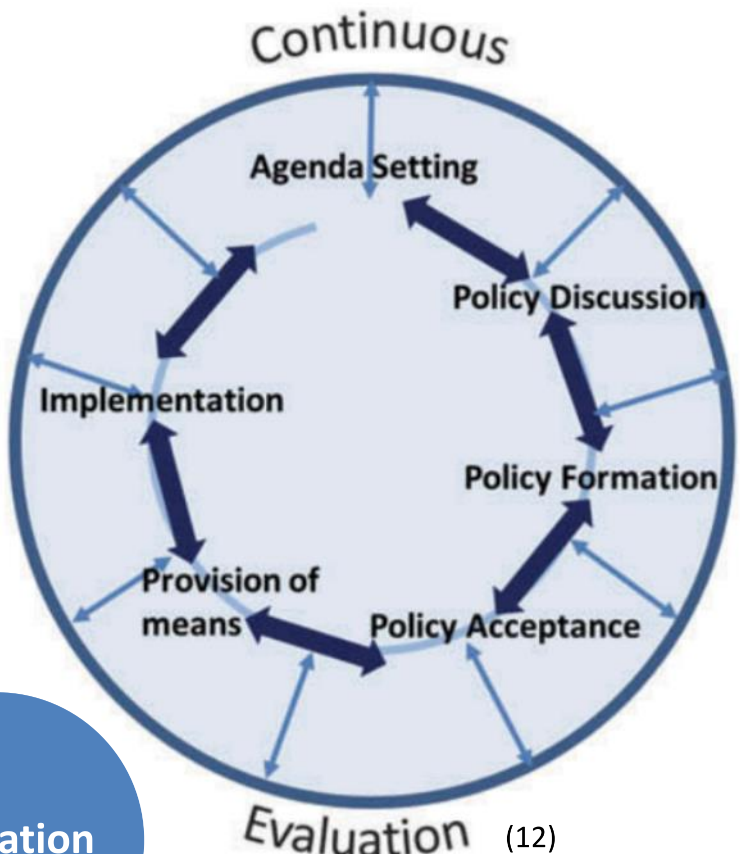
Design

Public policy process – Big Data for policy delivery

- For policy delivery, big data can add value during (4)
  - Public supervision and public regulation; continuous evaluation of policy effectiveness to improve future implementation processes
- Sources / types of data
  - Real-time data with immediate feedback loops
  - Social media data, mobility data, administrative data (e.g. census or budget)
- Techniques
  - Sentiment analysis, network analysis, randomized control trials, machine learning, data mining

Delivery

Public policy process – Big Data for policy evaluation



(12)

- For policy evaluation, big data can add value during (4)
  - The entire policy process, through continuous evaluation of policies
- Sources / types of data
  - Real-time data with immediate feedback loops
- Techniques
  - Sentiment analysis, micro-experimentation
- Transformative
  - Formative (<-> summative)
  - Continuous, iterative feedback (<-> after policy implementation)
  - Bottom-up, participatory

Evaluation

Zooming in on development policy evaluation

- Trends in evaluating development effectiveness (19)
  - Complex problems and complex assessment: e.g. the SDGs' focus on social, economic and environmental impacts rather than a single one
  - Adaptive Management and Doing Development Differently: doing "what works" through agile methodology, iterative learning, modern technology
  - Demand to trace impacts and results through value for money
- Challenges with traditional methods
  - Data gaps in official statistics (disregarded groups, weak infrastructure)
  - Household data often filled by hand causes inevitable time lags
  - Limited population groups and sizes, high cost
- Mixed methods (RCT+) may address those trends and challenges
  - Yield dynamic and rich sets of indicators
  - Enable feedback that may reveal unintended consequences while being more accurate and timely

Evaluation

Risks of using Big Data for development policy

- Selection bias and lack of external validity (2)
  - E.g. data from non-representative population
- Historical bias and lack of internal validity
  - E.g. event data from the past is not representative of events in the present
- Focus on correlation and prediction over cause, causal inference and diagnostics leading to prescriptive insights rather than understanding
  - E.g. predicting high crime rates rather than explaining why crime occurs
- Creating a new digital divide in terms of power and income
  - Big Data requires analytical capability and infrastructures
- Better data does not necessarily lead to better outcomes
  - Problem-solving requires interpretation, understanding, interoperability and acting
- Risk to privacy protection
  - Lack of common rules for data privacy and security requirements
  - Risk of de-anonymization and re-individualization of data
- Machine Learning / AI as black box
  - Lack of transparency and accountability for prescriptions delivered by machines

Challenges and bottlenecks of using Big Data widely in the policy process

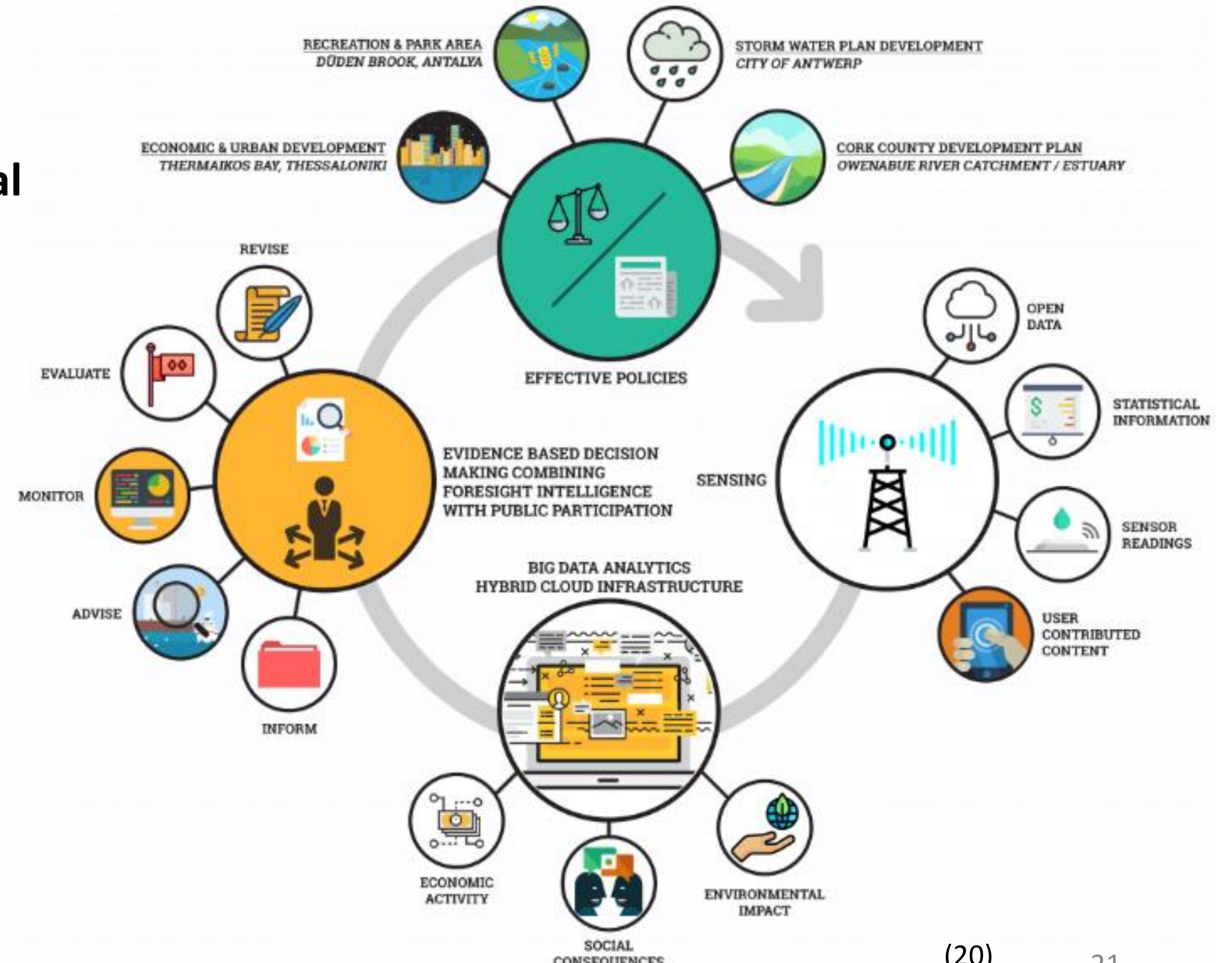
- Low budgets and legacy systems that do not leverage the power of Big Data (12)
- Limited interoperability of data due to lack of common legal, technical, operational and semantic alignment
- Lack of strategic leadership

UNITED NATIONS UNIVERSITY
UNU-MERIT

Maastricht University

Case study 1 - COASTAL URBAN DEVELOPMENT THROUGH THE LENSES OF RESILIENCY (CUTLER)

- EU-funded project (20)
- Sustainable development of **coastal areas** in different European cities
- Positive economic, social, and environmental impact
- Data-driven policymaking methodology
  - Inform
  - Advise
  - Monitor
  - Evaluate
  - Revise



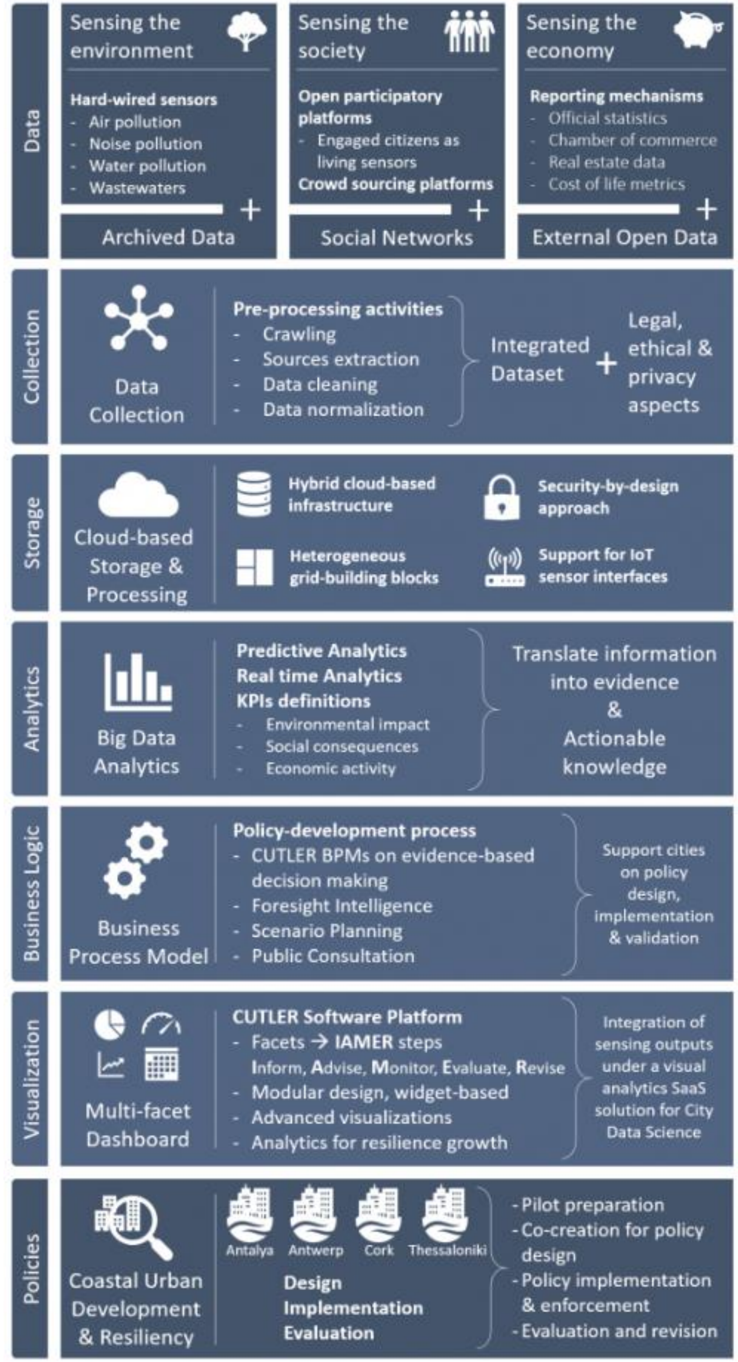
(20)

UNITED NATIONS UNIVERSITY

UNU-MERIT

Case study 1 – CUTLER architecture

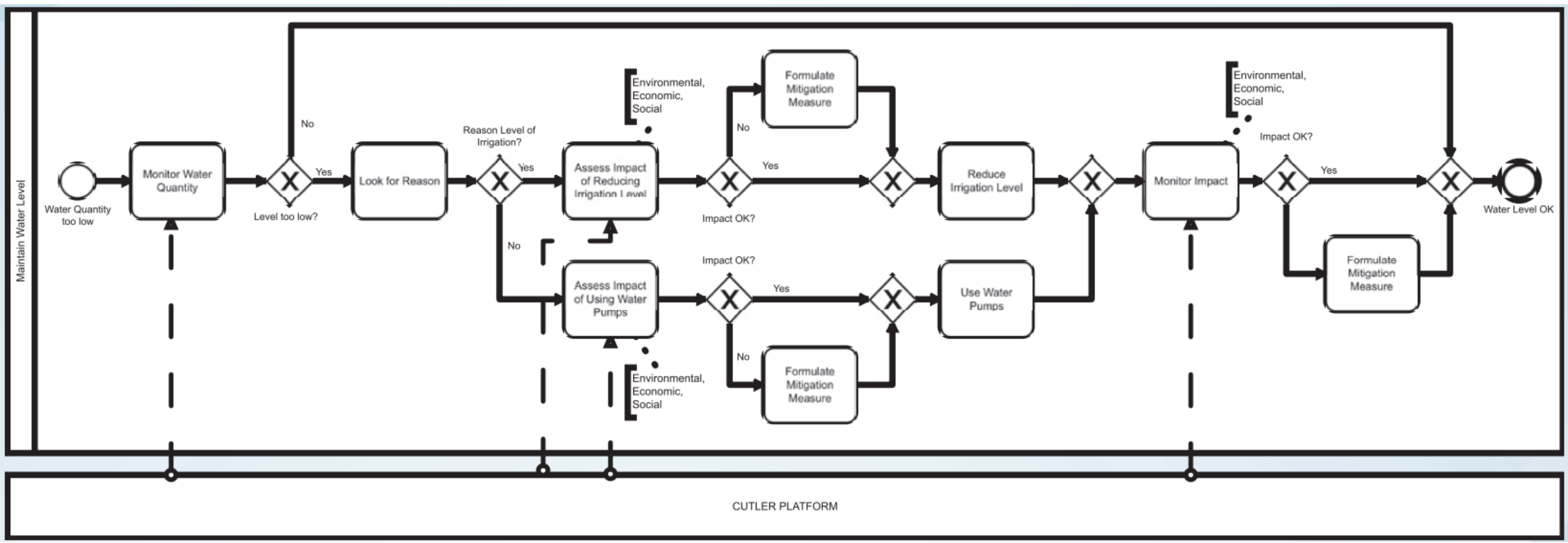
"The cardinal objective of CUTLER is to **shift the existing paradigm of policy making**, which is largely based on intuition, towards an **evidence driven approach enabled by big data**." (20)



sity

(21)

Case study 1 – CUTLER process model



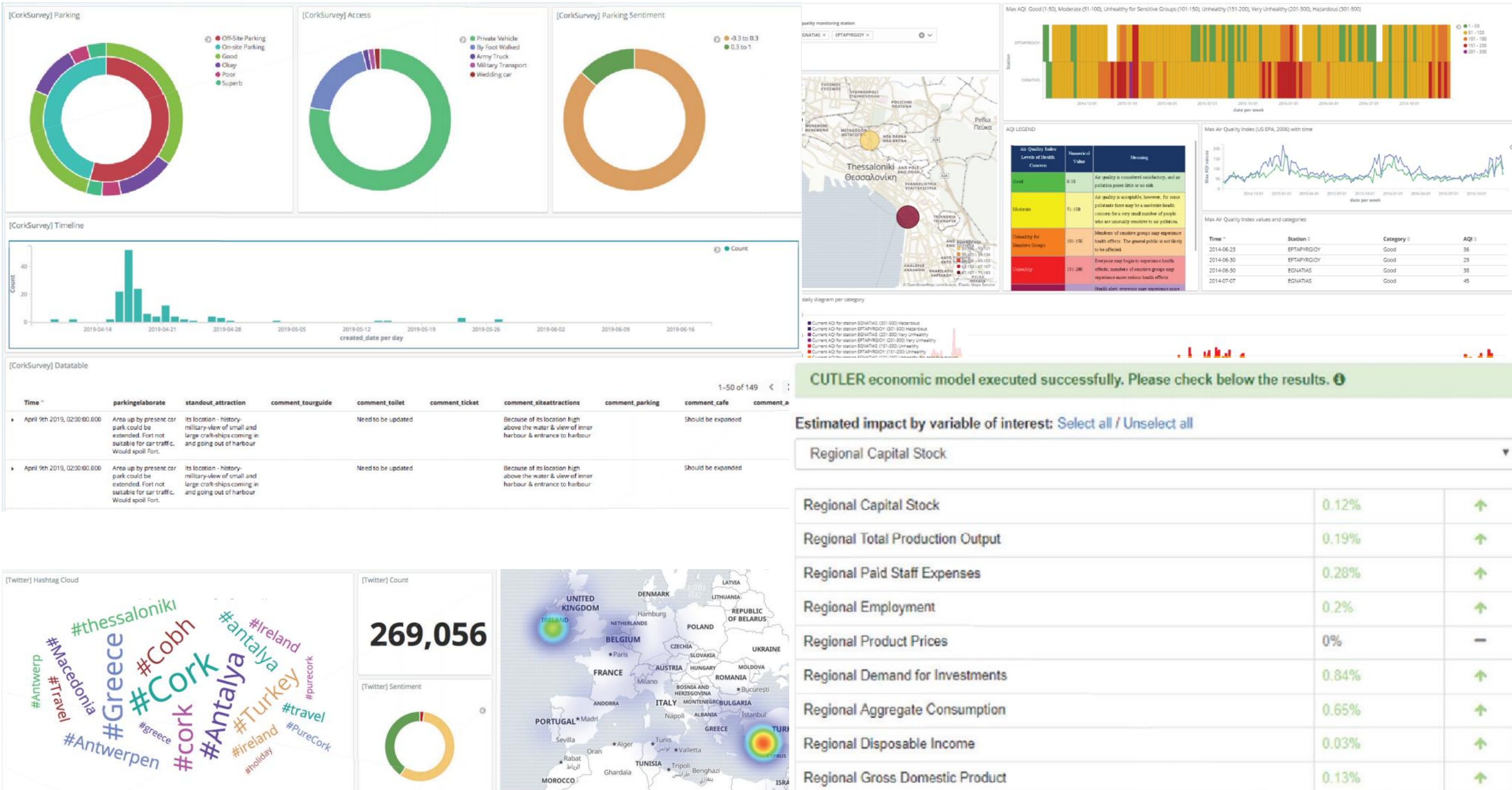
(22)

UNITED NATIONS UNIVERSITY

UNU-MERIT

Maastricht University

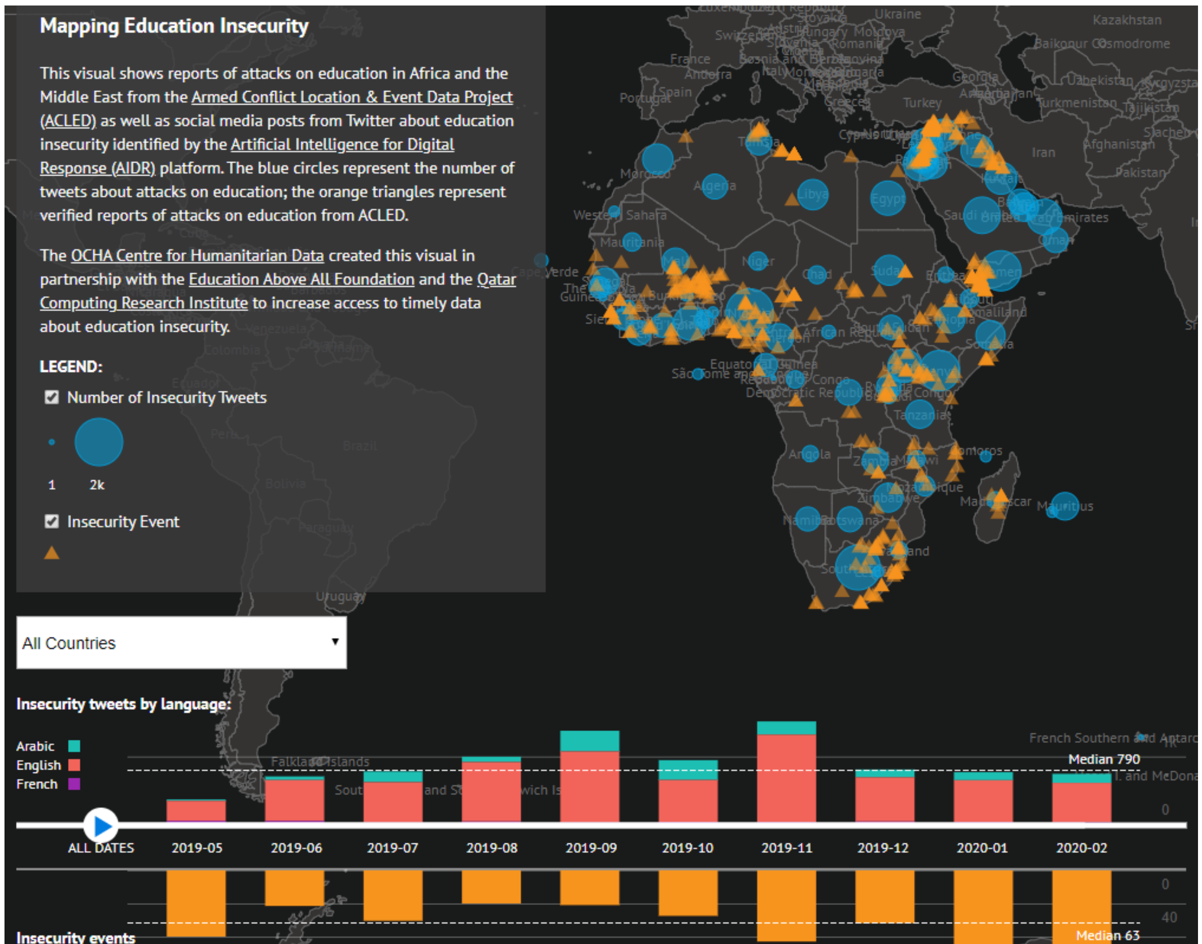
Case study 1 – CUTLER dashboards for social, economic and environmental impact



(23, 24, 25)

Case study 2 – Mapping Education Insecurity

- https://data.humdata.org/visualization/mapping-education-insecurity/



(26)

Methodology of a data project

Step 1: Understand the problem and its context (27)

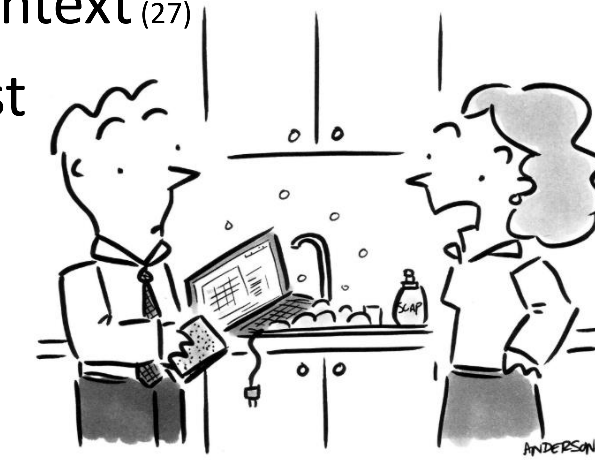
Step 2: Formulate a specific field of interest

Step 3: Get Your Data

Step 4: Explore and clean your data

Step 5: Formulate your question

Step 6: Enrich your dataset



"This is not what I meant when I said 'we need better data cleansing!'"

Step 7: Build visualizations (+ models, analyses, predictions etc.)

Step 8: Place findings in context

Step 9: Formulate actionable policy

Step 10: Iterate

Computational basics

- print()
  - A function in python which writes content

- Data types:
  - Int (integer), e.g. 5
  - Float, e.g. 5.1
  - String, e.g. hello
  - Boolean, either TRUE or FALSE

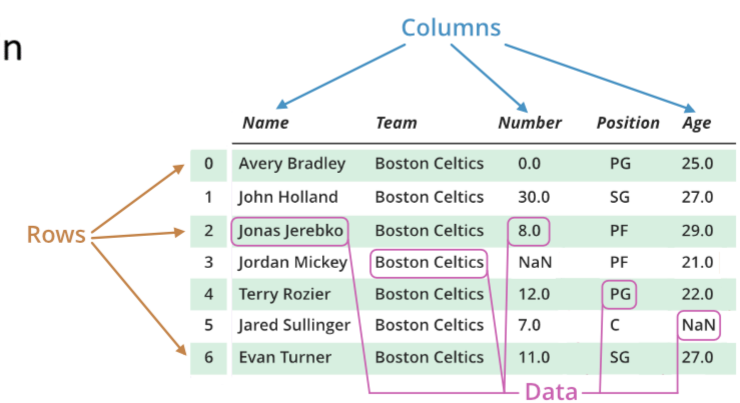
- Operators:

  + - * / %

  != < <= > >=

  is is not in not in

- DataFrame:
  - Two-dimensional tabular data structure with labelled axes (columns and rows)



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

Practical exercises in Kaggle

Introduction to Data Visualisation

- https://www.kaggle.com/learn/data-visualization

Introduction to Geospatial Analysis

- https://www.kaggle.com/learn/geospatial-analysis

UNITED NATIONS UNIVERSITY
UNU-MERIT

Maastricht University

References

1) Johns Hopkins University & Medicine (2020). Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html.
2) Letouzé, E. (2015). Big Data & Development: An Overview. Data-Pop Alliance Primers Series. http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf.
3) Educba (2020). 5 Best Difference Between Big Daga Vs Machien Learning. https://www.educba.com/big-data-vs-machine-learning/.
4) Studinka, J. & Guenduez, A. A. (2018). The Use of Big Data in the Public Policy Process - Paving the Way for Evidence-Based Governance. 2018. - EGPA Conference 2018. - Lausanne.
5) Data-Pop Alliance (2015). Official Statistics, Big Data and Human Development. White Papers Series. http://datapopalliance.org/wp-content/uploads/2015/11/Big-Data-Official-Statistics-FINAL.pdf.
6) Letouzé, E. (2012). Big Data for Development: Challenges and Opportunities. UN Global Pulse. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.
7) Letouzé, E. (2014). Big Data for Development: Facts and Figures. SciDev. www.scidev.net/global/data/feature/big-data-for-development-facts-andfigures.html.
8) Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. Administration & Society, 49(7), 1043-1064. doi: 10.1177/0095399714555751.
9) Dunleavy, P. (2016). 'Big data' and policy learning. In G. Stoker, & M. Evans (Eds.), Evidence-based Policy Making in the Social Sciences: Methods That Matter (pp. 143-151). Bristol: Policy Press. Retrieved 12 March, 2018, from https://www.researchgate.net/profile/Patrick_Dunleavy/publication/299467976_%27Big_data%27_and_policy_learning/links/56fa2cef08ae81582bf4435e/Big-data-and-policylearning.pdf?origin=publication_detail.
10) Severo, M., Feredj, A., & Romele, A. (2016). Data and Public Policy: Can Social Media Offer Alternatives.
11) Heitmueller, A., Henderson, S., Warburton, W., Elmagarmid, A., Pentland, A. S., & Darzi, A. (2014). Developing Public Policy to Advance the Use of Big Data in Health Care. Health Affairs, 33(9), 1523-1530. doi: 10.1377/hlthaff.2014.0771.
12) Big Policy Canvas (BPC). D5.2 Roadmap for Future Research Directions (Pending European Commission Approval). https://www.bigpolicycanvas.eu/sites/default/files/roadmap/BPC_D5.2_Roadmap_for_Future_Research_Directions-Final_Version.pdf.
13) Chui, M., Harrysson, M., Manyika, K., Roberts, R. (2018). Applying Artificial Intelligence for Social Good. McKinsey & Company. https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good.

UNITED NATIONS UNIVERSITY
UNU-MERIT

Maastricht University

References cont'd

14) United Nations, UN (2020). Big Data for Sustainable Development. https://www.un.org/en/sections/issues-depth/big-data-sustainable-development.html.

15) Howlett, M., Ramesh, M., & Perl, A. (2009). Studying Public Policy: Policy Cycles and Policy Subsystems (Vol. 3). Ontario: Oxford University Press.

16) De Marchi, G., Lucertini, G., & Tso, A. (2016). From Evidence-based Policy Making to Policy Analytics. Annals of Operations Research, 236(1), 15-38. doi: 10.1007/s10479-014-1578-6.

17) Big Policy Canvas (BPC). D3.1 Needs and Trends in Public Administration. https://www.bigpolicycanvas.eu/sites/default/files/misc/deliverables/D3.1_Needs_and_Trends_in_Public_Administrations_v1.1.pdf,

18) Bright, J., & Margetts, H. (2016). Big Data and Public Policy: Can It Succeed Where E-Participation Has Failed? Policy & Internet, 8(3), 218-224. doi: 10.1002/poi3.130.

19) Letouzé, E., Stock, M., De Chiara, F., Lizzi, A., Mazariegos, C. (2019). Harnessing Innovative Data and Technology to Measure Development Effectiveness. OCCASIONAL PAPER SERIES N º 54. Southern Voice.

20) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Concept: Overall Concept Underpinning CUTLER. https://www.cutler-h2020.eu/concept/.

21) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Architecture. https://www.cutler-h2020.eu/architecture/.

22) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Dissemination: Modelling decision-making processes in public administration. https://www.cutler-h2020.eu/dissemination/.

23) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Dissemination: Measuring the economic impact of policy making. https://www.cutler-h2020.eu/dissemination/.

24) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Dissemination: Measuring the environmental impact of policy making. https://www.cutler-h2020.eu/dissemination/.

25) Coastal Urban Development Through the Lenses of Resiliency (CUTLER). Dissemination: Measuring the social impact of policy making. https://www.cutler-h2020.eu/dissemination/.

26) The Humanitarian Data Exchange, Humdata. (2020). Mapping Education Insecurity. https://data.humdata.org/visualization/mapping-education-insecurity/.

27) Lau, C. H. (2019). 5 Steps of a Data Science Project Lifecycle. Towards Data Science. https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492.

Annex 1 – Use cases for Big Data in the public policy planning phase

Phase	**Literature**									
Planning Phase	Longo et al. (2017)	Höchtl et al. (2016)	Severo et al. (2016)	Alfaro et al. (2016)	Whitman Cobb (2015)	Lee et al. (2016)	Burnap & Williams (2015)	Schintler & Kulkarni (2014)	Bright & Margetts (2016)	Panagiotopoulos et al. (2017)
Technique			Soft data analysis	Text & sentiment analysis; opinion mining		Machine classification; statistical modeling	Supervised machine learning			Social media analysis; cluster mapping
Data Types		Social media data	Social media data	Social media data	Social media data	Social media data	Social media data		Social media data	Social media data
Goal	Agenda-setting	Agenda-setting; policy discussion	Agenda-setting	Identifying opinion streams and incorporating this feedback into the policy process	Measuring salience	Informing policymaking to formulate policy that is suited to the needs of local people	Monitoring the reaction to large-scale emotive events that may lead to hate crimes	Enhancing stakeholder participation & accountability	Enabling more accessible forms of public (mass) participation in policymaking	Discover conversations and spontaneous reactions in/near real time t
Example		Boston's Street Bump Application			Measuring public opinion of space policy in the U.S.	Measuring the effect of environmental attitudes of citizens on the adoption of green electricity policies in the U.S.	Analyzing the spread of online hate speech immediately after the murder of Drummer Lee Rigby in London			Analyzing collective input by farmers on Twitter for policy activities of the DEFRA

(4)

Annex 2 – Use cases for Big Data in the public policy design phase

Phase	Literature					
Design Phase	Giest (2017)	Williamson (2016)	Höchtl et al. (2016)	Guerrero & Lopez (2017)	De Gennaro et al. (2016)	Semanjski et al. (2016)
Technique	Data visualization	Learning analytics platforms	Advanced predictive analytics methodologies; scenario techniques	Network science; computational methods	Data collection with navigation systems; data processing platform	Smartphone app based data collection process; machine learning
Data Types	Administrative data, complemented with (real-time) data based on social media input, cameras and sensors	Educational data		Big data from administrative records (administrative data)	Datasets of driving and mobility patterns	Mobility data contributed by citizens
Goal	Shaping information-based policy instruments	Providing fine-grained knowledge and intelligence to formulate policy options	Contributing to evidence-based policymaking in the phase of policy formation	Using highly granular big data sets to create better policymaking tools	Increasing the effectiveness of future policies in the fields of transport and energy	Providing insights on mobility indicators and imminent feedback on implemented measures; shorter data collection and processing phases
Example	Providing real-time information on the education system for by creating digital and interactive data visualizations	Learning analytics platforms capture data from children's educational activities to track and algorithmically optimize their educational experience; predicting the future performance of the system and the student		Using employer-employee microdata in the labor flow network (LFN) model to capture labor mobility patterns and construct new labor market measures and policymaking tools for unemployment policies	Using the Transport Technology and Mobility Assessment (TEMA) processing platform s to support the development of effective transport regulation in the EU	Using the application *Routecoach*, 8300 citizens in Leuven voluntarily contribute their mobility data. Their crowd-sourced behavior was mapped to the wider population using machine learning approach

(4)

Annex 3 – Use cases for Big Data in the public policy delivery phase

Phase	**Literature**									
Delivery Phase	Höchtl et al. (2016)	Höchtl et al. (2016)	Höchtl et al. (2016)	Longo et al. (2017)	Dunleavy (2016)	Dunleavy (2016)	Maciejewski (2017)	Maciejewski (2017)	Maciejewski (2017)	Brayne (2017)
Technique				Real-time micro-experimentation	Online randomized control trials (online RCTs)	Sentiment analysis; machine learning; data mining	Sentiment analysis	Big data analytics	Performance monitoring; network analysis	Predictive analytics; network analysis
Data Types	Real-time data		Budgetary data	Real-time data		Social media data	Social media data		Mobility data	
Goal	Testing new policies by using real-time data produced in the execution	Improving the accuracy of information sources for policy implementation, e.g. census data	Improving decisions on required personnel and financial means for policy implementation	Testing policies by manipulating input variables	Evaluating small-scale effects using the availability of huge datasets	Preventive policing to improve arrest records, crime prevention and deterrence effects	Deriving feedback about policies to make an immediate response	1) Public supervision 2) Public regulation	Optimizing the transport infrastructure and commuting patterns	Using big data for a law enforcement-related activities
Example	Reducing crime rates at their origin by focusing an increase in policing more specifically on problem areas		Analysing the data generated from budgetary processes to detect patterns and design more efficient provision of means for a policy;		Using online RCTs to test the design of reminder letters for court fines (and ist influence on peoples' willingness to pay)	1) Police forces in Manchester monitored would-be rioters' chatter on social media and broadcasted own messages 2) Predictive policing in LA	Using the software Vizie (monitoring and analysis tool for social media) to quickly alert decision-makers of any changes that might require their attention	1) Using big data analytics to detect tax fraud, e.g. the British Connect system 2) Using big data to monitor adverse effects of FDA-approved drugs	The Land Transport Authority in Singapore (LTA) applies big data methods to improve public transport by gathering information of daily commuter rides	The LAPD compiles and analyzes big data for predictive policing. Geo-fences are used to generate real-time notifications and ALPR data is used for investigations

(4)

UNITED NATIONS UNIVERSITY

UNU-MERIT

Maastricht University

Annex 4 – Use cases for Big Data in the public policy evaluation phase

Phase	Literature				
Evaluation Phase	Ceron & Negri (2016)	Höchtl et al. (2016)	Schintler & Kulkarni (2014)	Ruggeri et al. (2017)	Lavertu (2014)
Technique	Supervised Aggregated Sentiment Analysis (SASA)				
Data Types	Social media data				
Goal	Ex-post evaluation of policies (reaction of online public opinion on policy alternatives and monitoring mobilization of opposition groups)	Continuous evaluation using big data as an integral part of every policy process phase	Ongoing evaluation of existing policies using big data to empower and engage citizens and stakeholders in the process	Smart regulation: using big data when rolling out policies to revise interventions in real time	External political actors are increasingly able to observe and evaluate the administration of public programs, using performance information
Example	Analysing citizens' opinions on two major public policies in Italy (job market reform & school reform) using Twitter data	The *Automated Continuous Evaluation System* of the U.S. Army uses big data analytics and context aware security to analyze government, commercial, and social media data to uncover patterns of applicants			

(4)

UNITED NATIONS UNIVERSITY

UNU-MERIT

Maastricht University

Annex 5 – Taxonomy for Big Data in development

	Applications	Explanation	Examples	Comments
UN Global Pulse report Taxonomy[1] (Letouzé, 2012)	1. *Early warning*	Early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis	Predictive policing is based upon the notion that analysis of historical data can reveal certain combinations of factors associated with greater likelihood of crime in an area; it can be used to allocate police resources. Google Flu trends is another example, where searches for particular terms ("runny nose", "itchy eyes") are analyzed to detect the onset of the flu season — although its accuracy is debated.	This application assumes that certain regularities in human behaviours can be observed and modelled. Key challenges for policy include the tendency of most malfunction-detection systems and forecasting models to over-predict — i.e. to have a higher prevalence of 'false positives'.
	2. *Real-time awareness*	Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies	Using data released by Orange, researchers found a high degree of association between mobile phone networks and language distribution in Ivory Coast — suggesting that such data may provide information about language communities in countries where it is unavailable.	The appeal for this application is the notion that Big Data may be a substitute for bad or scarce data; but models that show high correlations between 'Big Data-based' and 'traditional' indicators often require the availability of the latter to be trained and built. 'Real-time' here means using high frequency digital data to get a picture of reality at any given time.
	3. *Real-time feedback*	The ability to monitor a population in real time makes it possible to understand where policies and programmes are failing, and make the necessary adjustments	Private corporations already use Big Data analytics for development, which includes analysing the impact of a policy action — e.g. the introduction of new traffic regulations — in real-time.	Although appealing, few (if any) actual examples of this application exist; a challenge is making sure that any observed change can be attributed to the intervention or 'treatment'. However high-frequency data can also contain 'natural experiments' — such as a sudden drop in online prices of a given good — that can be leveraged to infer causality.

(2)

Annex 6 – Taxonomy for Big Data in development cont'd

	Applications	Explanation	Examples	Comments
Alternative taxonomy (Letouzé et al., 2013)	1. *Descriptive*	*Big Data can document and convey what is happening*	This application is quite similar to the 'real-time awareness' application — although it is less ambitious in its objectives. Any infographic, including maps, that renders vast amounts of data legible to the reader is an example of a descriptive application.	Describing data always implies making choices and assumptions — about what and how data are displayed — that need to be made explicit and understood; it is well known that even bar graphs and maps can be misleading.
	2. *Predictive*	*Big Data could give a sense of what is likely to happen, regardless of why*	One kind of 'prediction' refers to what may happen *next* —as in the case of predictive policing. Another kind refers to proxying prevailing conditions through Big Data—as in the cases of socioeconomic levels using CDRs in Latin America and Ivory Coast.	Similar comments as those made for the 'early-warning' and 'real-time awareness' applications apply.
	3. *Prescriptive*	*Big Data might shed light on why things may happen and what could be done about it*	So far there have been few examples of this application in development contexts.	Most comments about 'real-time feedback' apply. An example would require being able to assign causality. The prescriptive application works best in theory when supported by feedback systems and loops on the effect of policy actions.

(2)

Annex 7 – Links to relevant organisations, initiatives and research

- https://www.unglobalpulse.org/
- https://olc.worldbank.org/content/big-data-action-development
- https://data.humdata.org/
- https://datapopalliance.org/
- https://www.opalproject.org/home-en
- https://www.centreforpublicimpact.org/design-thinking-in-policymaking/
- https://www.weforum.org/agenda/2019/01/ai-for-human-development
- https://www.hbs.edu/socialenterprise/blog/post/tech-accelerator-for-nonprofits-supports-impact-that-scales
- https://www.tno.nl/en/focus-areas/strategic-analysis-policy/expertise-groups/strategy-policy/policy-lab-developing-data-driven-policies/
- https://www.bigpolicycanvas.eu/
- https://ec.europa.eu/info/sites/info/files/research_and_innovation/contact/documents/egov_brochure_interactive_1.pdf

Feel free to get in touch

Sofie Roehrig, sofie.roehrig@alumni.lse.ac.uk

Thank you for your attention!

DATA FOR SOCIAL DEVELOPMENT

OPPORTUNITIES, RISKS AND CASE STUDIES OF BIG DATA IN PUBLIC POLICY

Sofie Roehrig

UNU-MERIT/MGSOG

2020

Maastricht, The Netherlands