Through the clouds

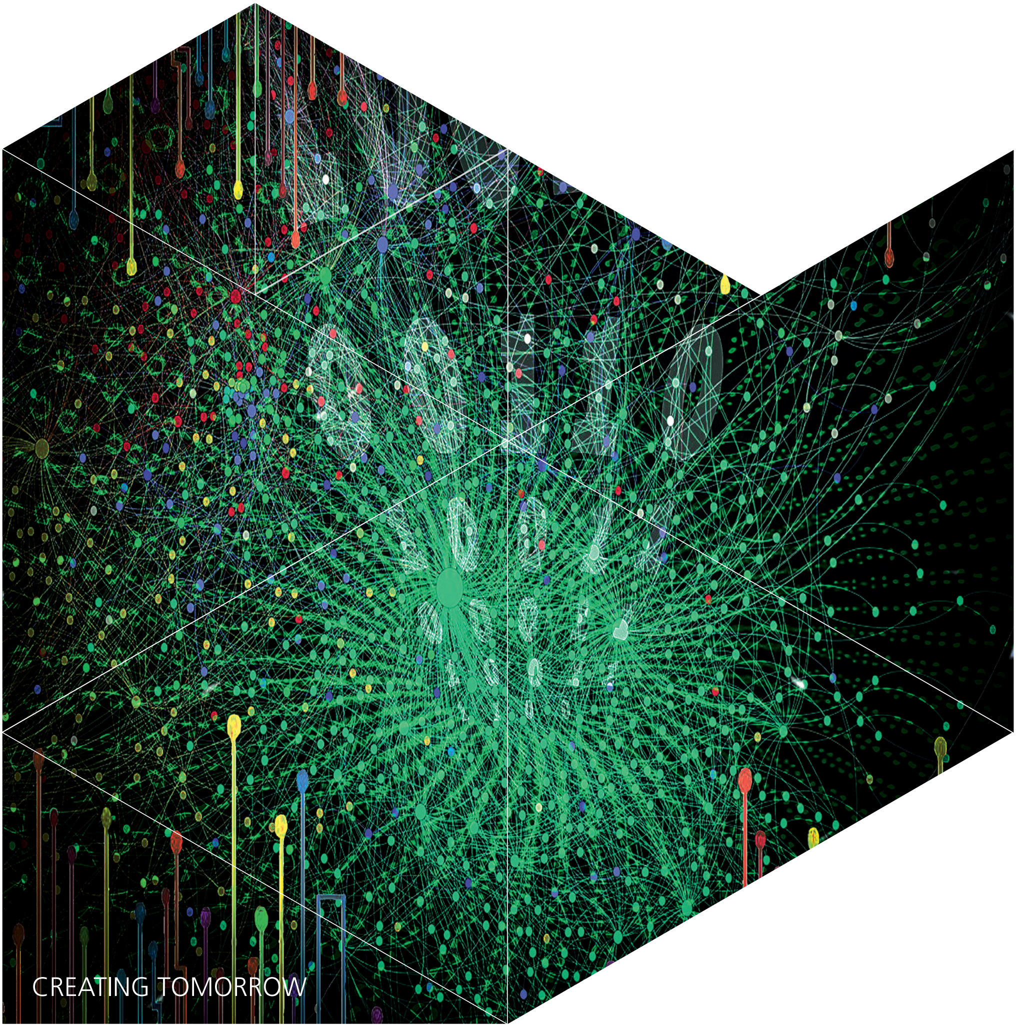
Piersma, N.

Hogeschool van Amsterdam

THROUGH THE CLOUDS

URBAN ANALYTICS FOR SMART CITIES

Dr. Nanda Piersma



CREATING TOMORROW

Through the Clouds

Through the Clouds

Urban Analytics for Smart Cities

Inaugural Lecture

Tuesday 17 April 2018

Nanda Piersma

Professor of Urban Analytics
Amsterdam University of Applied Sciences
and
Centrum Wiskunde & Informatica

Cover image: BertZuiderveen.nl

Graphic design: Textcetera, The Hague

ISBN: 978-94-6301-181-5

A PDF edition of this publication is freely downloadable from our website.
www.hva.nl/urban-analytics

Contents

This is my father, Jan Piersma (1930-2016).

He owned a car repair store for almost 50 years. He had no computer system for inventory, no cash register and no customer relations system. When asked for a spare part from a motor block Vauxhall model 1972 he would immediately respond, “Second aisle left, top shelf, next to the model 1968.”
He based his business decisions on knowledge and insight: “Do not take in the Toyota models for spare parts, but do so take them for the Volkswagen; they always break.”

He memorized it all. Now, the digital society is here.

Introduction

Large cities provide an exciting and creative environment but are also perceived to be expensive and congested. Cities host tech innovations, startups and (new) business development. They are also known for their busy streets, pollution, loneliness, and crime. Citizens may demand good city services, inexpensive and accessible health and school facilities, and political and social engagement. The challenge that cities face is to realize livability, social cohesion, efficient public services, economic growth and prosperity in a compact urban space.

Data has been collected for centuries, but in recent years technical innovations have enabled us to collect exponentially growing amounts of data through the use of sensors, smart devices and other sources. The walls literally have ears, noses and eyes, and citizens' movements, moods and interactions are recorded. This data is used to create an efficient, effective and inclusive environment to guide the digital transformation of citizens without compromising their privacy.

Mobile devices enable citizens to access data (services) anytime and anywhere. This development has a huge impact on our physical environment: phone cells, traffic direction signs and public transport information boards, for instance, are disappearing from the physical environment and replaced by apps and online assistance tools. Businesses are strongly affected by the availability of online and real-time information on their sales, inventory levels, finances and performance.

The digital transformation is an enormous challenge for citizens, companies and governments, and is having a major impact on daily life. We already rely on computer screens, historic performance data, trend reports and numerous other data sources. The rise of smart complex and self-organizing systems may guide a new level of human decision making, through automated decision making and systems that adapt to new situations through learning models. We need explainable, transparent and secure (data) models to include all citizens in this transformation.

Data analytics and algorithms are at the heart of the complex systems. Artificial intelligence and machine learning are used for understanding the data and for decision making. There is a paradigm shift in which insight and knowledge are constantly fed with data. Simple mechanical jobs are increasingly guided by online manuals and real-time diagnostic technology. We use calculators for basic calculations, and without a TomTom we would be lost before reaching the end of our street. We rely on algorithms for route planning or selecting a new wardrobe, and use automatically generated recommendations for choosing or for recommending a Netflix series that we may like.

Artificial intelligence is being introduced to support human intelligence. On multiple applications the artificial intelligence has shown to be superior to the human decision making. Can we control the artificial brain, or are we (as humans) becoming the future work drones for artificial intelligent systems?

This book is both an introduction to the world of Big Data and Smart Cities, and an assessment of the role that data analytics is playing in the digital transformation in our cities. The book is complemented with MOOCs, references to other online courses and technical reports for people new to this field (see www.hva.nl/urban-analytics). The chapter “Urban Analytics” discusses the activities of the Urban Analytics Research Group, with examples that explain the possibility to combine academic and applied research.

1 Data in Urban Environments

There has been data since the beginning of mankind. It has been used to count the number of people and animals in a community or to keep track of economic transactions and possessions. "Data are the raw material produced by abstracting the world into representational forms: numbers, characters, symbols, images, sounds, electromagnetic waves, bits and so on in order to categorize and measure it" (Kitchin, 2014).

Only through an ordering of the raw data into ordered facts does data becomes meaningful and capable of being used to create understanding. Data as a means to communicate information can add understanding, to create knowledge:

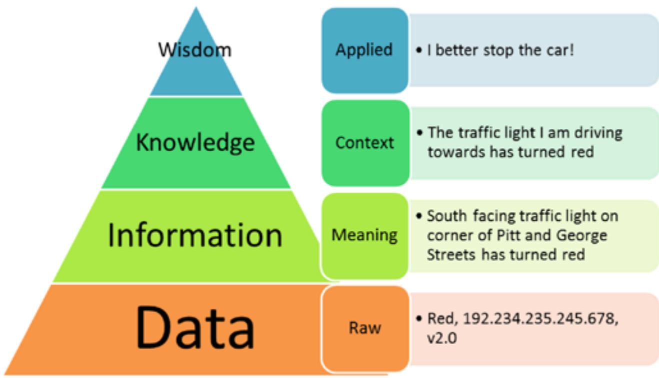


Figure 1 Wisdom-Knowledge-Information-Data Pyramid[1]

Raw data is the collection of low-level information units such as sensor readings, but it can also be a photo or the (geo-)location of a tree. From the raw data the information is derived by aggregating the facts through:

- (Subcategory) counts and measures
- Scaling and normalization (relative measures, counts per unit)
- Statistical (descriptive) analysis (averages, standard deviation, anomalies, outliers)

This aggregated data loses the information of the individual records but gives an overview of the information in the raw data. To make use of the data, it needs to be accurate (Kitchin, 2014):"Good quality data are discrete and intelligible (each datum is individual, separate and separable, and clearly defined), aggregative (can be built into sets), have associative metadata (data about data) and can be linked to other datasets to provide insights not available from a single dataset."

Data collection

Communication devices and the internet cause humans to produce enormous amounts of structured and unstructured raw data streams consisting of messages, pictures or geo locations. Smart devices, such as Fitbit trackers, TomTom, and mobile phones also track the activities and the whereabouts of their users.

The use of technical systems (sensors, cameras, and so on) for data collection is an emerging field. With the ability to gather data with smart devices, we can overcome cumbersome human counting processes (cars that pass a street, people in a room). Counting tends to be done for short intervals, is time-consuming and prone to error. Professional counting bureaus increasingly use automated data collection. Combined with the use of smart algorithms, for instance artificial intelligence algorithms for image recognition, they can program the raw data (images) into the required data (for instance, the number of cars that pass a street).

Other techniques include web scraping, where data is collected from websites. Open data sets are shared through an application programming interface (API) that allows programmers to import data directly into a data analytics environment. These techniques are publicly available, they improve the computational power of projects, and enable analysts to combine data from different sources in order to find relationships, correlations and causality in the data projects.

Big Data

The amount of (raw) data that is produced is growing exponentially and is denoted as BIG DATA. In 2013, it was estimated that the amount of data produced each day was 2 exabytes (10^{18} bytes, IBM, 2017). Big data are large volumes of heterogeneous data that have hidden structures that need to be extracted. The value of the information may hold only for a short time and is found only in fragmented small subsets of the data. For instance, the measurement of passing vehicles in a street to detect traffic jams is of interest only in the local environment and only during rush hour.

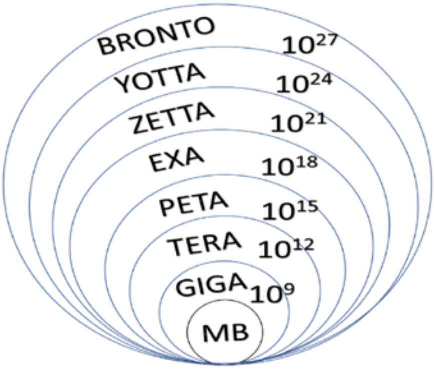
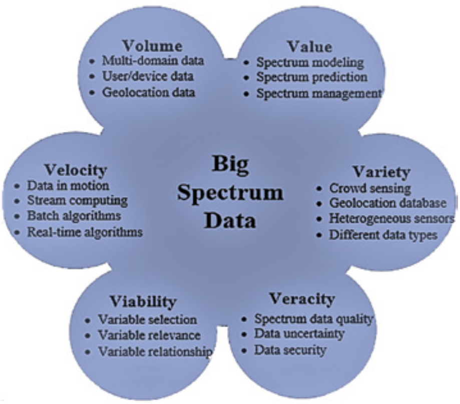


Figure 2 Data volumes

The "Age of Big Data" is characterized not only by large volumes, but also by other characteristics (Bahrami, 2015).



Volume indicates a very large volume of data.
Velocity indicates the speed for data processing in terms of response time. This response time could be a batch, real-time or stream response-time.
Variety indicates heterogeneity in data that we have collected for processing and analysis. This data variety includes structured, unstructured and semi-structured data.
Veracity is the accuracy of the data. For example, a sensor that generates data can have an incorrect value that does not provide accurate data.
Value of data refers to the information that can be extracted from data.
Viability means that the data should be filtered to find the most relevant factors and attributes and to detect hidden relationships in data.

Figure 3 Characteristics of Big Data[2]

High volumes of structured data can be stored using traditional relational databases (SQL). Big data sets such as text documents with no inherent structure, images or videos can also be NoSQL (Sakr, 2016).

Societal impact of data

The societal impact of using data has three important elements: privacy, security and responsibility.

Privacy is a person's right to be unobserved and uncontrolled by third parties. Data sets that reveal information about an individual, or data sets that can be combined in a way that allows that individual to be identified are allowed only with the explicit permission of each individual in the dataset. The new European General Data Protection Regulation (GDPR) will take effect on May 28, 2018. The GDPR brings a new set of digital rights for EU citizens in an age when the economic value of personal data is increasing in the digital economy.

> The bulk collection of telephony metadata in the United States under the USA Patriot Act in 2013 (USA, 2013) metadata caused a debate about individual privacy. According to the Act: "This information is limited to telephony metadata, which includes information about what telephone numbers were used to make and receive the calls, when the calls took place, and how long the calls lasted. Importantly, this information does not include any information about the content of those calls—the Government cannot, through this program, listen to or record any telephone conversations."

In the Netherlands, there is also a misconception about privacy-preserving metadata or data that can be traced backed to individuals. The 2017 Dutch Privacy Authority (Autoriteit Persoonsgegevens) identified data breaches at Airbnb, Uber, WHOIS data, Microsoft and Facebook.[3]

Security of (personal) data is of utmost importance to preserve privacy. Data ownership should be clear, and the owner should take sufficient measures to protect the data. Numerous stakeholders can breach data security: datacenters and platform owners (through the use of cryptography on the data, secured ports, firewalls, data management protocols with permissions and passwords), service providers (data management) and users of the services (passwords). Cryptography of the data can secure forbidden or unwanted access to data (in storage or in use). Since the days of Alan Turing, people have been trying to break the codes of cryptographed data (Hodges, 1983).

Hacking is done in two ways: by breaking into the datacenter (through an insufficiently secured port, cracking the coded data) or by using permissions and passwords of one of the users to access the data. Systems are most vulnerable to social hacking, with people sharing passwords, service desks that share passwords over the phone, people showing passwords on desktop screens, leaving laptops accessible, and so on. In addition, malware and phishing emails sent to end users can infect computers with software viruses that take control of them. Some of this software is programmed to send huge amounts of data to online platforms so that they block access to the platform (DDos attacks). This software does not hack the platform itself but disrupts it by preventing access.

New discussions in the data field relate to responsible or FAIR (Findable, Accessible, Interoperable, and Re-usable) data and FACT (Fairness, Accuracy, Confidentiality, and Transparency). The scientific approach to responsible data focusses on four research questions (Aalst, 2017):

Table 1 FACT

fairness	data science without prejudice – how can we avoid unfair conclusions even if they are true?
accuracy	data science without guesswork – how can we answer questions with a guaranteed level of accuracy?
confidentiality	data science that ensures confidentiality – how can we answer questions without revealing secrets?
transparency	data science that provides transparency – how can we clarify answers so that they become indisputable?

Responsible data science is one of the foundations of the digital transformation. Customers, patients, citizens and other stakeholders should be able to understand the purpose and the use of the data models in order to trust the outcomes “Rather than to avoid the use of data altogether, we strongly believe that data science techniques, infrastructures and approaches need be made responsible by design” (Aalst, 2017).

Smart Cities frameworks

The purpose of using data for city services is to reduce costs and resource consumption and to allow for effective and efficient city services while engaging citizens in their social living environment. Using data for this purpose can enhance the understanding of city service process, enable efficient decision making and create transparency about the drivers of city services. It has been estimated that cities generate more than 4.1 TB/day/km^2 of urbanized land area (Ciobanu, 2014).

Categories of data in urban context are

Table 2 Urban Data

General information	Registration records about citizens, companies and government organizations.
Infrastructure of a city	Houses and environments, transportation systems, geography, energy and water systems, green spaces and public spaces, air pollution monitoring, garbage collection, nature.
Living	Health, safety and security, education, income and expenditure, leisure and culture, tourism.
Economics and politics	Labor market, social services, price levels, elections, city (economic) investments.
Citizens' data	Social media (Twitter, Instagram, Facebook, Snapchat), customer relations (online shopping, subscriptions), GEO tracking (such as Facebook check-ins).

Urban data sources are available to citizens and other shareholders. Many open datasets are available (https://data.amsterdam.nl/, www.cbs.nl) from public and private sources. The City of Amsterdam is frontrunner in the collection and publication of CITY DATA.

Equally interesting are real-time data, such as cameras and sensors that track citizens' movements by car, bicycle or on foot, or that monitor air and water quality. Real-time data is used for monitoring and detection. This data typically holds all characteristics of big data such as veracity, volume and variety. Real-time data can be used only with the appropriate technical infrastructure support.

In an IBM corporate document (Harrison, Eckman, Hamilton, Hartswick, Kalagnanam, Paraszczak, 2010), a *Smart City* is an "instrumented, interconnected and intelligent city." *Instrumented* refers to the capability of capturing and integrating live real-world data through the use of sensors, meters, appliances, personal devices, and the like. *Interconnected* means the integration of these data into a computing platform that allows the communication of such information among the various city services. *Intelligent* is the inclusion of complex analytics, modelling, optimization, and visualization services to make better operational decisions.

The Smart City builds upon the combination of improved data storage capability, enlarged computational power and the availability of advanced data collection devices. It comes to life, however, only through data analytics and data-based service providing.

Urban computing is an interdisciplinary research field in which ICT systems, IoT systems, mathematical modelling and data are used to resolve the issues in an urban setting.

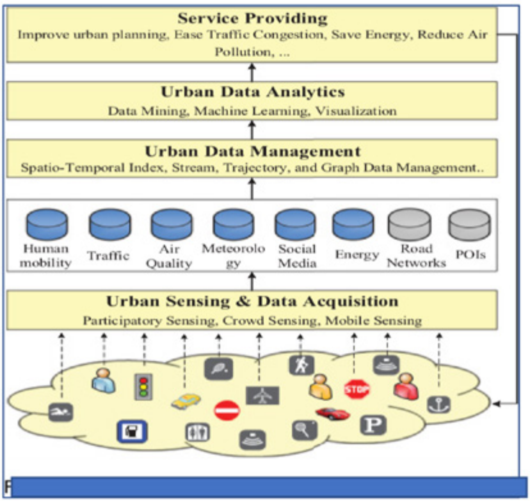


Figure 4 General framework of urban computing (Zheng, 2014)

Urban computing connects urban sensing, data management, data analytics, and service provision to improve the quality of life, city operation systems, and the environment. Governments and industry take the lead in implementing urban computing services and applications.

From a holistic view, a city functions better when urban computing becomes a public utility, just like water, electricity and other utilities. A *City Brain* uses data to create a centralized dashboard view of the sensors deployed across a distributed network in the city. In theory, municipal administrators could use a city brain to check on a wide variety of conditions detected by millions of low-cost wireless sensors.

At a presentation at the 2017 World AI Summit in Amsterdam (Hua, 2017), Alibaba AI manager Xian-Sheng Hua demonstrated the City Brain as a smart traffic system for the city of Macau. The data platform constantly monitors citizens and uses applied AI, deep learning and real-time data analytics to optimize the traffic streams. The City Brain can be extended to fields such as garbage collection, parking, safety, or crowd control. It can even monitor social behavior (detecting loneliness, mental problems, health crises). Singapore is implementing a City Brain and Cisco is working with ten cities (including Paris and Copenhagen) on the Cisco Kinetics for Cities, a Smart+Connected™ Digital Platform for monitoring and for improving operational effectiveness.[4]

Ranking and Scaling

On a global scale, many cities have the same urban data available, making it possible to compare their performance. Because cities face growing competition for investors, qualified employees and tourists, the rankings can reflect the competitiveness of a city, assist with strategic positioning to achieve city goals and profiling. Ranking cities may also give an indication of the progress in the digital transformation of cities. Giffinger (2007; 2010) identifies the six domains of a smart city:

Table 3 Six Domains of a Smart City (Giffinger, 2007)

Smart Economy	Competitiveness
Smart People	Social and Human Capital
Smart Governance	Participation
Smart Mobility	Transport and ICT
Smart Environment	Natural Resources
Smart Living	Quality of Life

Using weights for the importance of the indicators in each domain, the score of cities is calculated for each ranking with a specific focus (such as sustainability, economic growth).

Many global rankings are published, for instance by IBM, the McKinsey Global Institute, Roland Berger, EY, and PricewaterhouseCoopers. On a global scale, data is collected by the UN http://urbandata.unhabitat.org/ providing data about 741 cities with 103 indicators. In the CITYKeys project (Horizon project 2015-2017) more than 40 EU frameworks for smart city rankings are compared, resulting in a set of indicators for assessing the success of smart city projects.

Giffinger (2010) distinguishes four types of rankings based on the Acting Institute: commissioned economy consultancy-based, commissioned expert panels/private research institutes, magazines and NGOs, and sponsored university or economic research institutes. The differences in the ranking methodology among these institutes is characterized in, among others, the quality of documentation, transparency, and number of cities.

From an academic viewpoint, the principle of scaling in cities is of interest. Are there universal rules to describe the services in a city in relation to its size? How does the infrastructure of a city scale with the number of citizens or with the (area) size of the city? Consider parking: will larger cities need extra parking space at a rate that grows linear with the number of citizens, or will there be a tipping point in the size of the city population that makes people choose alternative means of transportation?

Bettencourt (2013) claims that cities evolve according to a basic set of local principles. Measures of urban efficiency are independent of the size of the city (usually measured by a city's population). This result can tilt the balance between socioeconomic outputs and infrastructure costs. Van Raan (2015) and West (2017) show that city infrastructures grow either sublinearly or super linearly with the population, depending on the purpose.

The introduction of socio-technical systems to support citizens will have an impact on the city infrastructure and may result in new scaling principles. Smart technology and digital platforms will enable citizens to use their environment more efficiently, but each city will adopt smart city technology in its own way. What drives the adoption of smart technology? Is it possible to find scaling laws for smart cities? Scaling laws may help to understand both smart cities and the influence of new technologies on livability for citizens.

Digital transformation

People are relying on smart consumer products more than ever. Among these are technical systems, such as smart TVs or smart energy meters. There is the fast-moving embedding of smart technologies into everyday life. The expansion of the accessibility and bandwidth of ICT networks and the replacement of traditional systems by smart systems has given rise to pervasive and ubiquitous computing.

Pervasive computing is the addition of computational power and access to ICT networks to everyday fixed objects and environments to make them interactive and "smart" (Dourish, 2001). Ubiquitous computing is the computational power (such as smart phones) that accompanies a person independent of location and environment. Ubiquitous computing has been called the third wave in computing; increasingly the technology will recede into the background of our lives and citizens are becoming both the owners and the users of the city data infrastructure.

Socio-technical systems involve the interaction of humans with complex technical systems and infrastructures. The technical systems have a "front end" for humans and a "back-end" where collected data are shared over the internet and stored in data warehouses owned by the product companies to be studied and used. For instance, a smart energy meter provided by the energy company is used by a homeowner to set the thermostat. The energy company records the settings to balance the regional use of energy against the supply. This dual purpose of technical systems is under some scrutiny. The Samsung Smart TV, for instance, provides a YouTube app on the TV, but does not keep the app upgraded. People cannot install or upgrade apps on the smart TV. Within a couple of years, the TV will be less "smart." Meanwhile Samsung can monitor the use of each TV for its own business development goals.

One of the goals of research into socio-technical systems is to jointly optimize technical excellence and quality of life: how does the technology actually affect the human experience?

There is a paradigm shift where (personal) insight and knowledge are constantly fed with new data, presented to humans on complex socio-technical systems (smart devices). The use of digital technology transforms the context for humans.

*The **digital transformation** is defined by the use of digital technology for the transformation of business and organizational activities, processes, competencies and models (Industry 4.0), and its accelerating impact on human society (Society 5.0, sometimes denoted as "The paperless society" or "The knowledge society").*

The transformation results in a connected city, or rather in connected citizens, with services that are more cost-effective, have improved planning, improved designs and faster delivery of transport, and an infrastructure and housing that ensures a healthy, sustainable, resilient and prosperous living environment.

Data supports Bottom up and Top down processes

Citizens have access to the smart devices and technical systems for their own benefit and to contribute to society. People can control their own data stream, with data from apps like Fitbit, Eetmeter, Facebook, and LinkedIn. This is *small data*, the mathematical denotation of n=1 (one person). Only if information is shared by enough people can it be used to study citizens' behavior from a top-down perspective. Cooperation and (decentralized) exchange through virtual communities and social platforms, are typically situated in the not-for-profit sector, with smart living models and sustainability (Florida, 2017). The data shared on these platforms can help to start movements, to connect people and to form inclusive social practices (e.g., *stadsdorpen*).

Urban data systems under top-down control rarely connect to these social platforms. The role of collaborative commons in smart cities is underestimated and not visible enough in the current review of smart cities. If used appropriately, urban data can help bridge the gap between top-down and bottom-up processes, whilst helping stakeholders to recognize that cities are complex systems that operate through various spatial scales of urban form (Grey, 2017).

Critical publications and groups fault the top-down implementation of data-driven models (O'Neil, 2016). The triple helix models (government, research institutes and business) should be replaced by a quadruple helix model with the addition of the collaborative commons as a partner in smart cities projects (Florida, 2017). Bottom-up initiatives are studied, for instance, by the research group Create IT of the AUAS, which focusses on citizens' empowerment, with new techniques to create inclusive digital platforms.

2 Computer systems and data processing

The exponential growth of data was possible with the construction of more powerful computer infrastructures to store, access and process the data. When we store big data, we need to extract, transport and load (ETL) the data in order to analyze it and to extract information. Accessibility has a tradeoff with storage; some storage models use more computer space but provide faster access to specific data units, where other smaller storage models need more time for data processing.

The ability to record, store, access and process data in real time is the driver of the big data age.

When data is too big for our laptop

Traditional computers use a series of logic gates that transform different inputs into a predictable output. There are three computer architecture dimensions: storage (the amount of data stored), memory (the amount of data that is active) and data processing units (the data processing speed).

Computing storage is measured in bytes and computer memory is measured in RAMs (Read Access Memory). Data processing is measured in numbers of processing units CPU (central processing unit) or GPU (graphical processing unit) with the related processing speed MIPS (million instructions per seconds) or FLOPS (floating-point operations per second). Hardware-accelerated computing is the use of hardware to perform some functions more efficiently than is possible in software running on a more general-purpose processing unit. The cost of the architecture are determined by the amount of storage.

The first computers could store very few characters (Selectron tube 1946: 256 to 4096 bits), and until 1980 the maximum data storage of a mainframe computer was limited to gigabytes (IBM Model 3380 1980: 2.52 GB). The largest computer storage machines that can now store about 20 petabytes. Meanwhile the cost of computers has fallen dramatically. The costs to store 3 TB on a community computer is now approximately €100; in 2000 it was €33.000 and in 1990 €3.300.000.

Supercomputers (Scale up)

A supercomputer has a much higher level of computing power than a general-purpose computer. The first supercomputers introduced in the 1960s were highly tuned conventional designs that ran faster than their general-purpose contemporaries. With new capabilities such as vector computing, JIT, quantum computing and massive parallelization, supercomputers are more powerful than ever. Other new capabilities include optical and DNA computing.

In June 2016, the fastest supercomputer on the TOP500 supercomputer list was the Sunway TaihuLight, in China, with a LINPACK benchmark score of 93 PFLOPS and 20 PB storage capacity.

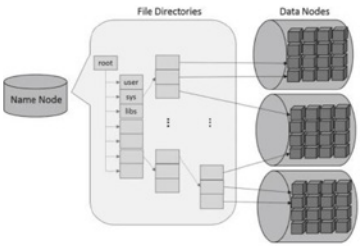
Active	June 2016
Operators	National Supercomputing Center in Wuxi
Location	National Supercomputer Center, Wuxi, Jiangsu, China
Architecture	Sunway
Power	15 MW (Linpack)
Operating system	Sunway RaiseOS 2.0.5 (based on Linux)
Memory	1.31 PB (5591 TB/s total bandwidth)
Storage	20 PB
Speed	1.45 GHz (3.06 TFlops single CPU, 105 PFLOPS Linpack, 125 PFLOPS peak)
Cost	1.8 billion Yuan (US$273 million)
Purpose	Oil prospecting, life sciences, weather forecast, industrial design, pharmaceutical research
Web site	http://www.nsccwx.cn/wxcyw/

Figure 5 Supercomputer

Distributed computer models (scale out)

Analyzing huge volume datasets use both the “scale up” (building larger computer systems) and “scale out” techniques. The latter involved the clustering of commodity machines that act as an integrated work unit; the dataset is distributed over multiple machines.

Distribution models are used to split the dataset into data blocks. Traditional relational databases use metadata as keys to relate SQL databases. Non-structured data (NoSQL) models are based on block storage, file storage or object storage principles. Block storage is the lowest level without structure and is related to hard disks. File storage uses a file directory structure and object storage uses nodes that are accessed on meta-data keys. Metadata is the frame of reference that gives data its context and meaning.



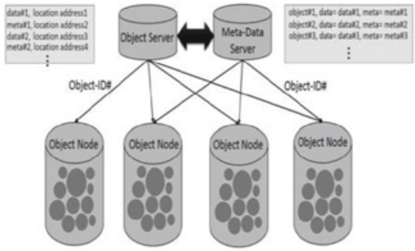


Figure 6a File storage

Figure 6b Object storage[5]

All major data storage systems use one of these models: the Amazon EBS is block storage model; Google file system (GFS) and Hadoop distribution file system (HDFS) are file storage systems; Amazon S3 and Open Stack Swift are object storage systems (Wu, Sakr, Zhu, 2017).

Principles that can be used to enhance computing power are faster memory processing (data to algorithm) or algorithm to data (MapReduce) techniques.

MapReduce is a famous big data solution: split the data set, perform parallel algorithms on the subsets of data (distributed computer systems) and combine the resulting output.

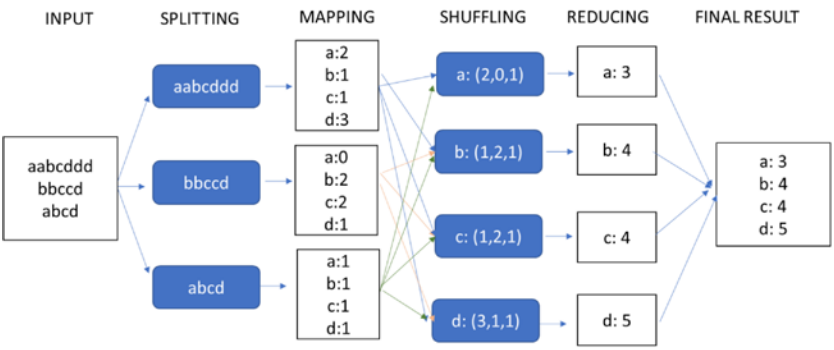


Figure 7 MapReduce technique for the word count example (Hadoop Apache website)

New developments for computer infrastructures evolve along the lines of scaling up and scaling out in combination with alternative use of computer memory and hardware (Al Ars, 2017):
Until 2005: High-performance machines
Until 2010: Hadoop-enabled distributed storage and processing
Until 2015: Faster in-memory processing Spark, HBase, Hana

Today there are ongoing developments in storage techniques (Quantum, DNA) and advances in data processing techniques such as JIT (just-in-time compiling), optical computing, TensorFlow (numerical computation using data flow graphs).

Computer power is said to develop towards the singularity point, a hypothetical point in the future when artificial intelligence will surpass human intelligence. The futurist Kurzweil (2006) predicted that computer power will be as powerful as the human brain in 2020, and reach that of the total human population in 2080.

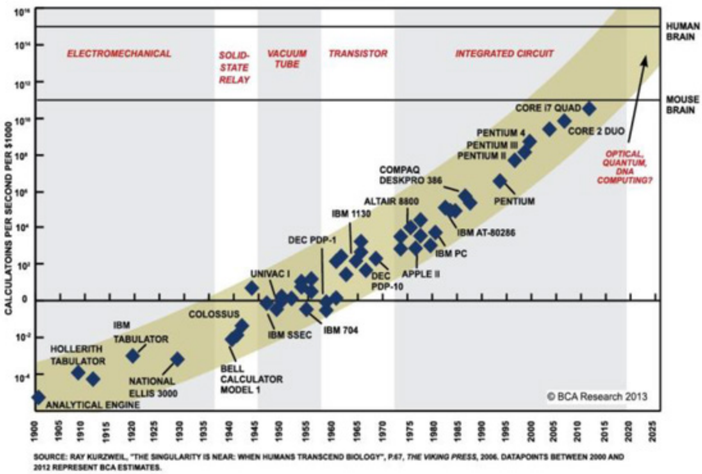


Figure 8 Computer power moving to the singularity point (Kurzweil, 2006)

Cloud platforms

Another way to scale out the computer infrastructure is cloud computing, which shifts the location of the data infrastructure (hardware and software) to more centralized and larger-scale data centers. Users can access data via the internet using thin of flat client platforms (laptops, mobile phones). A second important feature provided by cloud computing technology is the ability to grow (or shrink) computing and storage needs on demand. These features allow customers to pay the infrastructure costs of storing and computing based on their current capacity of big data and transactions (Bahrami, 2015; Sakr, 2016).

The pay-per-use principle has multiple forms, or “As a Service” components:

Table 4 Cloud Services

IaaS	Infrastructure as a Service	IT infrastructure (servers and virtual machines (VMs), Storage, Networks, Operating Systems)
PaaS	Platform as a Service	Platforms for developing, testing, delivering and controlling software applications
SaaS	Software as a Service	SaaS enables cloud providers to host software applications and its IT infrastructure, to perform administrative tasks for control and maintenance such as installing software updates and security patches.
DaaS	Data as a Service (not Desktop or Diagnostics)	This is a combination of IT infrastructure, data cleaning tools and SaaS

Cloud platforms have been used for a decade in Business2Business, enabling companies to outsource their IT infrastructures, software and increasingly their services. For example, printers are now on-site devices as part of a “Printing as a Service” contract that a company has with the printer supplier. The company does not buy the printers; it buys the print service with a cost structure based on print volume.

People also intensively use cloud platforms and their services.

Many organizations create their customer relations and support services on cloud platforms. Customers create an account on the website of the company and use it to communicate with the company. Chatbots and recommendation systems enable companies to automate customer relations interactions, enabling a company to expand without having to invest in its back-end.

People are using cloud platforms for online shopping, to interact with product and services companies (kitchen appliances support or to make online appointments with dentists or hairdressers) and for government and public functions (online tax applications, registration of energy and water use readings).

Social media platforms such as Facebook, Instagram and Twitter are cloud-based services where individuals can create an account and use the platform. These services are free, in the sense that there is no cost to subscribe.

There is some concern about the dominance of the companies that own the computer infrastructures (Amazon, Google, IBM and Microsoft) and companies offering social media platforms (such as Facebook, Twitter, YouTube). All of these companies support the connectivity of multiple dimensions of data that people are sharing on social media platforms, the customer support platforms and their private cloud environments, including passwords and personal data. The data submitted by account users is owned by the platform company, which uses the data for profiling and advertisements. As the saying goes, "If a product is free, you are the product."

Disappearing computer

In the past 40 years, one by one the boundaries of communication were lifted (Ballon, 2016):
- The "tele" age introduced communication hardware (telegraph, telephone, television) to overcome distance.
- The "e" age (e.g., e-business, e-commerce) is characterized by the internet as a communication platform: the information can be accessed anytime.
- In the "i" age (iPad, iPod, iPhone) the communication platform is personalized. Without time and space boundaries, information can now be directed to the individual need to know.

Without much thought, people are using integrated systems on their personal devices. Hardly anyone looks at train schedules on information boards in the station. The railway apps automatically personalize the information based on GPS tracking of the location of the mobile phone and the time of the inquiry. Only trains leaving from the nearest train station in the coming hour are shown. Travelers expect this information to show first.

As computing power and computing speed increase, computer chips are becoming smaller. The "disappearing computer" has been integrated into our televisions, wearable computers, mobile phones, actuators, wireless communication equipment, Radio Frequency Identifier (RFID), tags and cards and Personal Digital Assistants (PDAs).

Ubiquitous computing technologies integrate computers seamlessly into the physical world. In addition, sensor technologies such as QR (Quick Response) Code, GPS (Global Positioning System), Light, Pressure and Humidity Sensors, Magnetic Sensors, Orientation Sensors, Gyroscope, Microphones, Accelerometers, Clocks, Cameras and NFC (Near Field Communication) are used to acquire information in a ubiquitous environment, making them both a service and a data collection device.

IoT (internet of things) devices are also programmable to deliver services through real-time data analysis. Simple examples are a Fitbit tracker that signals when a heart rate threshold has been surpassed, or a security gate delivering an alarm when it detects safety breaches.

The further embedding of integrated systems will support services on appliances and products without human interaction. Examples are car systems that automatically summon an ambulance when an air bag in the car deploys, or mobile phones that shut down when they detect that a car is moving.

The abundance of intelligent devices creates a complex system of agents interacting with each other. Intelligent and Autonomous (complex) Systems are studied by the IAS group at CWI, focusing on mechanisms that enable the emergence of various degrees of organization, intelligence and autonomy in such systems, and apply them to problems of societal relevance.

3 Data analysis and mathematical modelling

Data science is the process of adding meaning to data in order to create knowledge and understanding. Companies have traditionally applied business intelligence to monitor and to manage their business processes. Data science that extends beyond the goals of business intelligence use data for understanding, forecasting and decision making. The GoDataDriven Big Data survey 2017 (GoDataDriven, 2017) states that business companies see a difference between business intelligence and data science. Eighty-seven percent of the companies use data for dashboards and reports, but 45% of the respondents says that they do not yet use data to develop predictive models. Some perceptions are that data science "takes over" the managerial business development models, the way we do business. However, data science is an additional way to manage businesses, combining data with business instincts and human knowledge.

Designing a data science project: Business Analytics

Business understanding is important for the setting of a data science project. The urban analytics group offers business analytics workshops through the Smart City Academy[6] to identify how data can support business development. In this workshop we use business development techniques (Nieuwenhuyse & Vanhoudt, 2008) to design a data science project that has an actionable impact on the company. *Actionable* means that the project is executed in a controlled setting, with a limited number of known stakeholders and several well-defined goals. The workshop associates the business project goals with possible interventions. Crucial in the workshop is to define critical process indicators from the data that monitor the outcomes of the project. Only when the goals of the project are clear and the use of the data is defined, one should begin with collecting (raw) business data and start the data project.

Example: The use of data to improve satisfaction with the restrooms in the Amsterdam Arena.

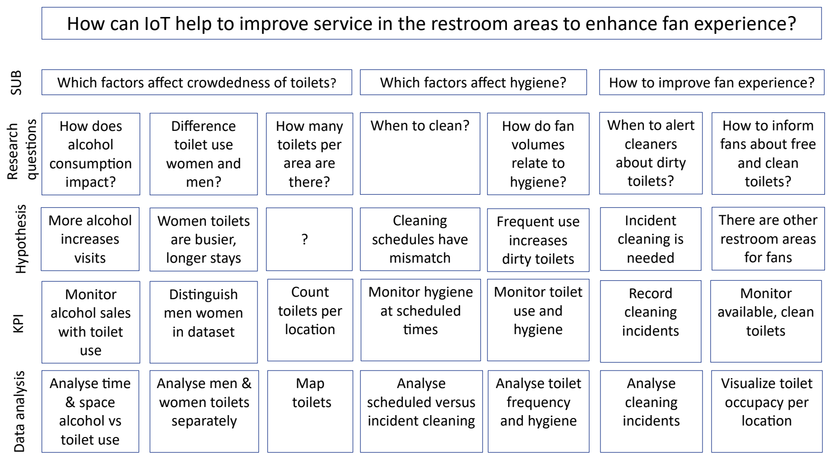


Figure 9 Business analytics for the use of restrooms at Amsterdam Arena[7]

The business terminology used here to describe a data science project also holds for many other projects, including bottom-up citizens development. For instance, when defining a data project to improve social engagement in a neighborhood, it is still important to identify stakeholders (who), actionable process interventions (how) and key performance indicators (goals).

Data science project

A data science process starts with a precise research question, its relation to the business context and the scope of the project: these are the required outcomes of the project. The business analytics workshop helps to define the research question. The research question also determines the approach to the project. Sometimes the project is more a business intelligence project (visualizing business data) than a data science project (adding meaning to data to create knowledge).

The data science project follows these steps: define the research question, extract data, analyze the data, visualize and implement the results. The details of the data science project are given in figure 10:

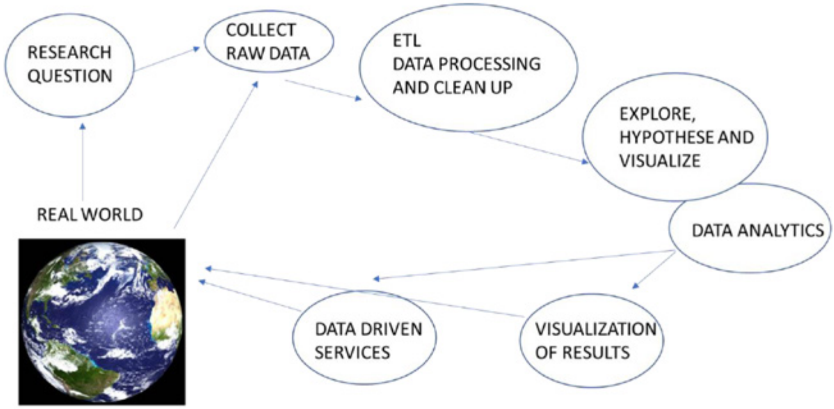


Figure 10 Data science project

ETL (extract, transform and load)

The collection of raw data will require some loops to determine the data quality, to store the data and to make the data accessible. Data acquisition (raw data), establishing data connections (API, LoRa) and sensoring and managing IoT data streams require some technical skills. The urban analytics group is writing tech reports to help acquire the basic skills for data acquisition (see e.g. reports on Wi-fi scraping, Web scraping, API setup, UA-Ubuntu images for Raspberry Pi on www.hva.nl/urban-analytics).

From an ICT perspective there are many challenges in handling and using the acquired data, for instance developing high dimensional data structures, the fast exchange of large data sets, or the optimal use of computational power. Big data requires accurate data engineering, both for research purposes and for their applications. Not only do big data sets need to be (stored and) processed to generate the information and the knowledge, but due to the possible flaws in the data collection the data needs pre-processing to identify errors, missing data fields and other irregularities.

IDOLAAD project

The Energy research group of the Urban Technology research program of the AUAS hosts the data of all charging sessions of electric vehicles in the public streets in Amsterdam, Rotterdam, Utrecht, The Hague, and the Metropole Region Amsterdam (MRA). The dataset requires many cleaning steps to handle erroneous charging sessions, such as double records, empty records, no kWh charged, zero connection time, and unidentifiable charge station. For the 2012-2014 datasets of the City of Amsterdam cleaning the datasets reduced the number of records (sessions) from approximately 400.000 records to less than 250.000.

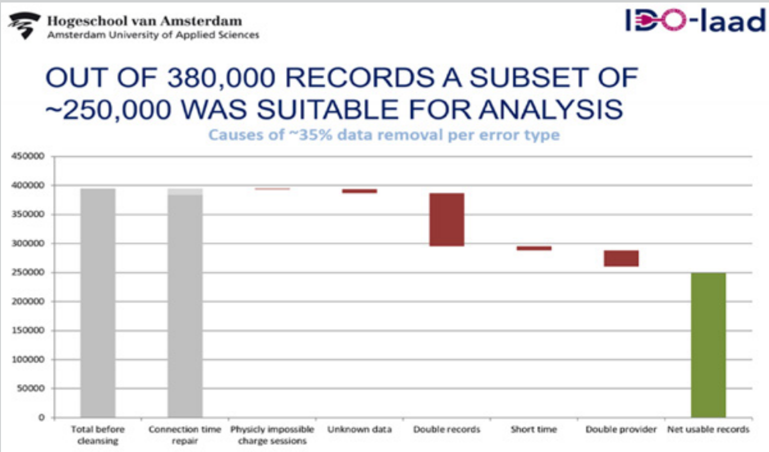


Figure 11 Data cleansing of the CHIEF data sets

Visualization

Visualization is used in the exploration phase of a data science project. In Tufte (1985) basic principles and examples of visuals that help the data exploration are explained, and the famous Anscombe data representation example (Anscombe, 1973) shows the importance of visualization. At the end of a data science project, visualization is used to communicate the results to stakeholders, for instance, through dashboards (both static or interactive and real-time), by graphical representations or by using storytelling tools.

The IDOLAAD project uses visualization techniques for monitoring the charge infrastructure and reporting to the municipalities.

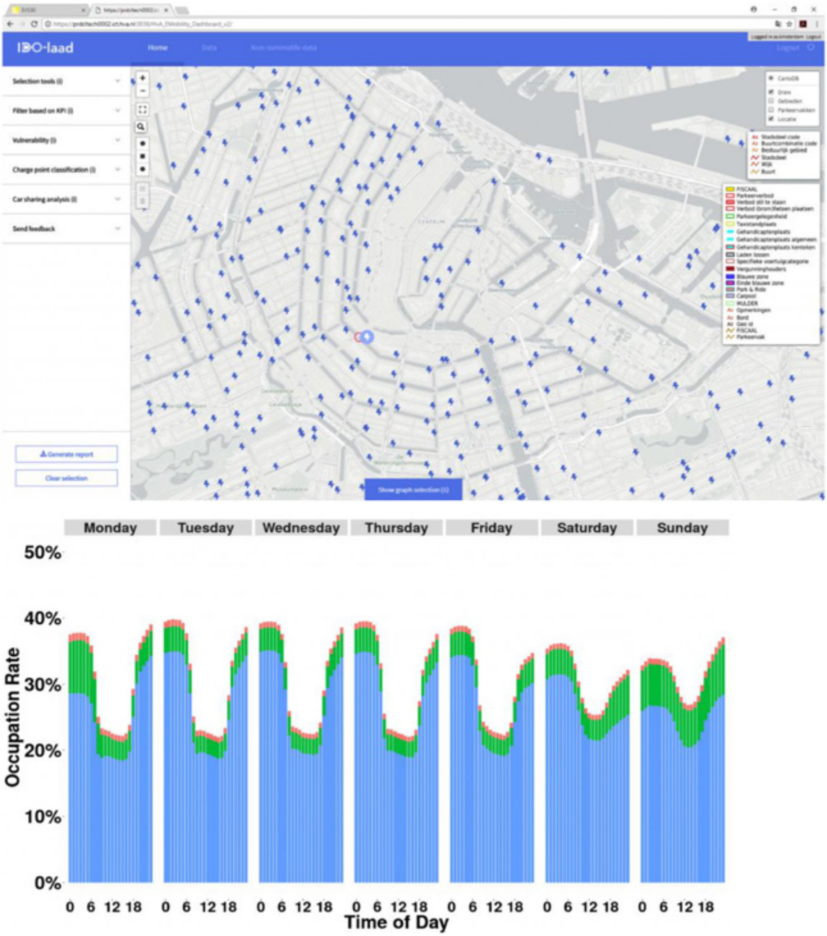


Figure 12 Charge Infrastructure Amsterdam: Example dashboards (IDOLAAD.nl)

Visual analytics (Keim et al., 2008) is a research field that uses visualization as the prime analytic technique to retrieve knowledge from the data. Examples are geospatial analytics, cognitive and perceptual science and knowledge representation.

Data analysis

The analysis of the data to understand the underlying process can be characterized into four groups: what happened, why did it happen, what will happen in the future and how can we make it happen.

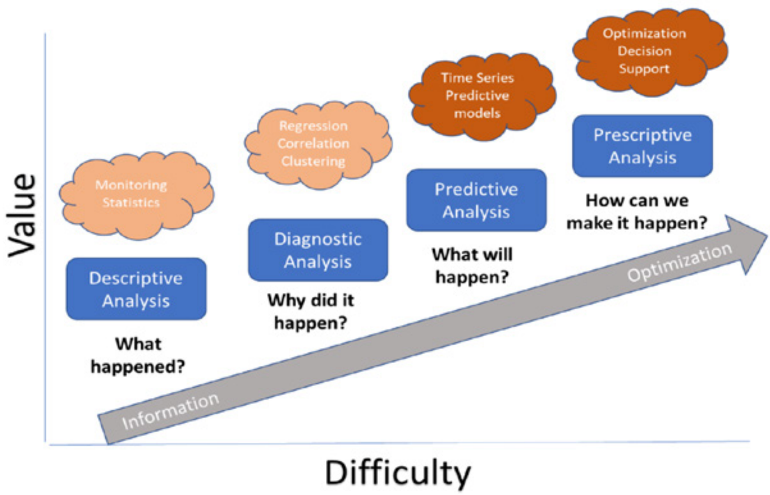


Figure 13 Data analytics[8]

Data analysis can be one or any combination of the techniques in Figure 13. Descriptive analysis is closely related to business intelligence and visualization techniques: it uses statistics, dashboards, reports and other visualizations to give the user insight into the data. As noted at the beginning of this chapter, 87% of the companies use their data for descriptive analysis.

Diagnostic analysis consists of a wide range of sometimes already long-established techniques and models such as correlation and mathematical models to find causal effects. Big data adds a new dimension to this analysis: large high dimensional data streams are now used for the data analysis and automation of data exploring algorithms. It's difficult to establish a ranking or hierarchy between data mining, artificial intelligence, data analytics and machine learning. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data (Han, 2012). It consists of automatic (machine-driven) algorithms for finding patterns or clusters in a dataset, and for finding causal relations. Combining multiple sources of data is an important part in data mining. Data mining is a foundation for machine learning and artificial intelligence; it can be used to formulate hypotheses about undiscovered structures in the data. Machine learning is a more generic process of training using (historical) data to predict the outcome for new instances (predictive analysis), typically used for, among others, recommendation systems. Artificial intelligence (AI) is a field that uses data for automatic logic, perception, reasoning, learning, and actions. AI is used, for instance, for language understanding, image processing or complex automated systems and robots.

Machine learning and AI can often perceive, understand, predict, and manipulate real-world research questions faster and more precisely than humans. However, they cannot exist without a proper interpretation of the application and its context. Barthelemy (2016, section 2.5.3) gives an example where the CO2 emissions levels from vehicles in cities are measured. The understanding of the city size (defined by land use or by population density) gives a radically different context of a city, leading to different conclusions. Depending on the definition of a city, larger cities are considered to be either less green (urban areas) or greener (population density).

The predictive and prescriptive analysis thrives on the understanding and modelling of the patterns in the data (diagnostics analysis). Data analysis of this kind supposes that the data can be divided into patterns that can be found and modelled ($f(X, \beta)$) and noise (ε) that is data that the mathematical model at hand cannot explain:

$$Y = f(X, \beta) + \varepsilon$$

Here Y is the characteristic under study, X is the set of characteristics related to Y through a functional form ($f(X, \beta)$) with a weight factor β. For instance the amount of kWh energy usage in a household Y can be explained by other characteristics (X) in the dataset, such as household size, and a score of the quality of the building. This will hardly ever be a perfect functional form. The residue [formule] is noise, the part that cannot be explained. When the mathematical form is acceptable (e.g. the noise is small enough) we can predict future behavior of Y by applying the functional form to predicted values of X.

Data-driven services

A data science project uses data to help understand and act on a business process. The results of the data science project will enable the use of new data-driven services for monitoring, optimization and digital platform (communication) tools.

Data-driven products of the IDOLAAD project

Some successful data science projects in the IDOLAAD project for the charge infrastructure of electric vehicles are (see also http://www.idolaad.nl/):

* CHIEF dashboards: an interactive dashboard with the monthly performance of the charge stations (see figure 11, Maase et al., 2017)
* Public parking space reservation for electric vehicles (Piersma et al., 2017)
* A fuzzy model for measuring the pressure on charge station networks (Piersma et al., 2018).
* Profiling EV chargers with machine learning (Van den Hoed & Helmus, 2015).
* Predicting future charging demand in neighborhoods (Steenbrink et al., 2016).
* Web-based applications:
* Free charge stations (https://www.oplaadpunten.nl/, https://www.amsterdam.nl/parkeren-verkeer/amsterdam-elektrisch/kaart-oplaadpunten/) Social charging app (https://www.social-charging.com/)

Mathematical Modelling versus Data Mining

There is a major difference between studying real-life processes using data discovery techniques or by fitting a mathematical model to the process. Data mining finds patterns in data and models the observed (cor)relations; the mathematical form is therefore a result of the data discovery process. Mathematical modelling presupposes a structural form of the real-life process and tries to fit this model to the process by finding parameter values. Examples of mathematical models for prescriptive analysis are decision models, complex systems and simulation models.

Simulation models are mathematical models that are used to predict the outcome of structural changes in the system that cannot be observed real time. By statistically modelling the actors in the current systems, and by simulating alternative setups of the system with the same actors, one can derive performance measures for the alternative setups.

Modelling the interaction among multiple identities to understand the complex system in an urban environment is not easy. But practical applications may need very complex mathematical models to understand their dynamics, relying on many data sources to capture the dynamics. The IAS group at CWI studies complex systems and their algorithms.[9] Translating these mathematical models into practical applications is one of the challenges of the Urban Analytics group.

The different approaches to modelling data processes can strengthen each other. For instance, for the IDOLAAD project, in (Steenbrink, 2016, Berkelmans, 2018) the predicted electrical vehicle charge demand per neighborhood is modelled as a discrete choice model (EV users choose a neighborhood for charging), with a multi-nested logit model as the statistical model for the probability of choosing a specific neighborhood. This model is then applied to a dataset of charge sessions to find the parameters of the mathematical model. By distinguishing different users (taxis make choices for charging that home-based chargers or commuters do not) we find different models for each user type. These profiles of EV users were found in the data by using machine learning (Van den Hoed & Helmus, 2015). The total charge demand in the neighborhoods is then calculated as a weighted combination of the choice models for the distinct user types, resulting in a more precise model for the prediction of future demand.

Explainable Artificial Intelligence (AI)

AI algorithms learn from historical data. They possess the ability to reason, to draw inferences based on the situation and context and to analyze and solve complex problems while handling constantly changing structured and unstructured data. The field of artificial intelligence attempts to build intelligent entities operating independently in a complex environment. Examples are self-organizing systems (such as robots), self-learning systems (such as social media recommendation systems) or deep learning algorithms (such as camera recognition systems).

Citizens' interactions with (artificial) intelligence systems will occur through the increasing availability of smart devices and designs. These socio-technical systems only support urban living if citizens understand and accept the algorithms. The citizens have to trust the way the algorithms work and what implications the results have for the application.

For instance, false positives (erroneous findings in the data) and false negatives (failure to find a result in the data) may have different implications.

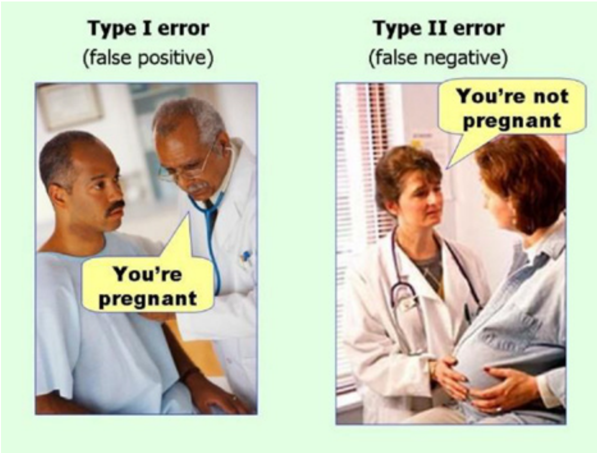


Figure 14 False positive and false negative for pregnancy[10]

If an algorithm is 98% precise in predicting an outcome, in an academic sense the algorithm is considered to be of good quality. Other indications are sensitivity, specificity and accuracy: measures that try to explain the cause of the errors (false positive or a false negative).

This leads to discussions in public space: how much do we care about false positives or false negatives? A false positive usually costs money in the forms unnecessary additional research or counter measures. In an ethical sense, we find that false negatives have more serious implications.

Artificial intelligence strengthens the public debate, since the algorithms uses logic, perception, and intelligence (learning) to guide the real-life processes. Translating these indicators into real consequences for performance is an emerging field called *explainable AI*.

Since the amount of urban data increases exponentially, we will have a greater need for AI algorithms to make sense of them. This is especially relevant because Amsterdam wants to become the prime AI ecosystem in EU.

Applied versus Academic Research

The urban environment can be studied from a well-defined actionable perspective or by using more complex mathematical modelling. These approaches can simultaneously be applied to applications in the urban context.

Consider the delivery of goods in a shopping district. From the perspective of the transport company the actionable problem is to prevent delivery delays. Its goal is to develop efficient planning tools, where driving distances are matched with estimated driving times based on historic data. The historical driving times are used to estimate future driving times which are then used as input for a planning optimization model.

A second, more academic project can be to find the causes of the delays, using data from other sources (e.g. weather data) to give better estimations of future driving times, and to build dynamic planning systems with real-time driving time estimates.

Even beyond that, an artificial intelligent system may add automated route planning in the trucks of the transport company that adapts to the traffic pressure in the neighborhood, interacting with the location of other street participants.

An agent-based simulation model is a fourth approach. It models the interaction of the transport company with other participants in the streets (tourists, commuters, pedestrians, other vehicles) and the use the existing infrastructure. The model can try to find interventions in the infrastructure that reduces the delays for all participants of the traffic system. All four projects impact the delivery problem for the transport company, but for different stakeholders and with different actionable partners.

From a scientific viewpoint we may also want to find mathematical laws of the traffic situation that provide knowledge about the patterns of the street traffic, such as chaos theory, tipping points, laws of catastrophe and so on. This approach is theory driven, the knowledge provided by the studies do not translate into the actionable problems, but can help to drive new approaches to street traffic (such as negotiation models for time windows for delivery and possibilities for automatic driving).

The goal of the Urban Analytics team is to identify both the actionable research in a well-defined setting (applied research) and the underlying scientific modelling for extending the setup of the research environment.

4 Urban Analytics

For the digital transformation of citizens and the emerging smart cities, urban computing and socio-technical interaction systems in urban environments are key elements of the surrounding of urban analytics. Any data-driven approach relies on technical infrastructures, but also on business analytics, data analytics, and mathematical modelling to apply services that support efficient and livable cities. At the cutting edge of complex systems are autonomous and intelligent algorithms, enabling citizens to use self-organizing entities, multi-agent and self-negotiating systems that can be applied to, for instance, energy systems, predictive maintenance services and collaborative commons.

The Urban Analytics research group will pursue four research directions:

- Business analytics and actionable models
  These are well-defined research topics that can be modelled with business analytics and solved with data analytics. The research consists of developing scalable techniques for business analytics activities that can be transferred to each business domain.
- Mathematical modelling for multi-agent's systems
  These research topics require understanding and explaining complex interactions among stakeholders, requiring mathematical modelling on a higher level of urban challenges.
- Design smart systems for socio-technological interaction
  This research direction involves real-time visualization and decision-making supported by a technical device. Smart explainable algorithms programmed on the device should be transparent and understandable for the users.
- Universal laws of scaling in cities
  From a meta perspective, the complexity of urban computing/analytics research questions will scale with the size of a city (and the number of citizens). Understanding the impact of the city scale on smart cities is an unexplored research field. In this era of exponential growth of data and digital transformation the theory of scaling is especially interesting.

These research directions have applications in three domains in urban environments: energy, urban living and mobility.

Energy

In the energy domain, the digital transformation is complemented by the transition to sustainable energy sources. Energy grids are becoming complex systems that allow for bi-directional energy flows, for instance to load solar energy into the energy grid. Smart grid systems are communication networks, using data-driven energy-balancing algorithms to transform the conventional energy grids into an intelligent and adaptive energy delivery network (for an overview of communication networks for smart grids, see Kahn, 2013).

Smart meters and energy-balancing algorithms can be applied to optimize energy usage. Predictions of energy demand for the next day or even hour is done by the energy company for all its customers in a designated region. But balancing the energy use can also be done locally (smart buildings) or by household (smart homes), using smart meters and price elasticity algorithms. Energy optimization, prediction models for future demand on different time scales, and the participation of consumers in the demand response to price incentives and to sustainable alternative (local) energy sources are active research fields. Mocanu (2017) has applied machine learning algorithms to smart grids.

Energy companies are incorporating smart grids into their business models. Smart grids are being pilot-tested worldwide to see how consumer participation influences the energy nets:

- Social platforms and tools for citizens can help to balance household energy usage, by arranging for home people to use energy at certain times. For instance, the social charging app for electrical vehicles will use charging stations more efficiently by sharing a charging station among several vehicles and by charging cars when it is less expensive.
- There are already pilots with local energy nets: privately owned energy systems such as solar or wind systems that can produce energy locally. The supply energy production of privately owned solar panels can be stored in batteries or returned to the energy net. The communication and administration of this system is supported by a socio-techno platform.

We see three uses of data analytics for the energy sector: digital platforms to support energy communication, optimization of energy use and the prediction of future energy usage. Many datasets are available about energy consumption in the traditional (fossil fuel based) grids. Data about both traditional and sustainable smart energy systems are less available. The Urban Technology group wants to initiate a data platform for open energy data.

The Urban Analytics group will participate in three projects in this field of smart energy systems.

Project Smart Charging (TKI, approved)
Charging electric vehicles (EV) is closely related to the project (smart) energy systems. Each EV in itself is a battery for energy that can be used for driving or for (temporarily) storing energy. The number of EVs is growing, along with the understanding of the charging and driving behavior of the EVs. Intelligent systems for smart charging and for the support of energy exchange between citizens using EVs is an interesting addition to the energy system study in Cities. The study includes:

– smart charging algorithms
– social charging platforms
– prediction models for charging demand
– matching algorithms for charging and household's energy demand

Project Energy Intranets (NWO, approved)
As the share of renewables grows, so does the decentralization of production and control, making it more difficult to match supply and demand. New control mechanisms and ancillary services will need to be developed to dampen the inevitable fluctuations. This project proposes to investigate the use of data-driven demand and supply matching in challenging use cases.

Project citizens empowerment (Interreg, in application)
Our aim is to allow citizens in our cities to benefit from the energy transition by empowering them. This means a better choice of supply, access to reliable energy price comparison tools and the possibility to produce and sell their own electricity. Increased transparency and better regulation give more opportunities for civil society to become more involved in the energy system and respond to price signals.
Our approach follows four focal points:

– Analysis of specific governance/regulatory/technical challenges
– Development of optimal regional methodologies
– Testing of the methodologies in eight pilot projects
– Communication

Urban Living

Governments and municipalities try to organize the cities such that it has high economic, livable and sustainable standards. Urban living is the way in which humans interact with a city environment. Cities are complex systems of (technical) infrastructures, human activities and human technical interactions. Systems for housing, mobility, food, energy and water are basic subsystems in cities, the city infrastructure. In Chapter 2 a City Brain was introduced as part of the city infrastructure, consisting of data to make the city's work efficient and effective.

City governments aim to improve prosperity, in economic status, in health, in sustainability and in social cohesion. Well-being and social cohesion are important, but they are complex goals that are difficult to measure (see ranking of cities Chapter 1). The livability for citizens in urban environments has many indicators, such as clean air, no pollution, green spaces and good recreational facilities, good schools, inclusive society, good health services, and good mobility.

The interaction between bottom-up and top-down approaches to a smart city are visualized in digital platforms. Citizens' interaction systems and smart designs are used for visualization of information and insights from the data, to interact with citizens and display the results of analysis. Often the data itself is shared on the digital platforms. The citizens themselves monitor their environment and share their data. Sensing the city is done to detect anomalies, to create inclusive communities and to match city services with the needs of its citizens.

The European Horizon program mentions "Digital innovation hubs and platforms" as an innovation theme for the next five years, with these subfields:

- Open data markets/exchange hubs
- Blockchains
- Big data platforms
- Citizens' data lab (participating citizens)

In GUST (2017) the characteristics, practices and examples of urban living labs are discussed, a collaborative, transdisciplinary co-creation form of experimenting, designing and evaluating new ways in urban living styles. They show three forms of living labs: strategic (led by governments or large private actors), civic (led by urban actors such as universities, cities or urban developers) and grassroots (led by urban actors in civil society and not-for-profit organizations).

The Create IT research group hosts the civic form with the Citizens Data Lab, where inclusive data-collecting practices are studied. De Gezonde Stad hosts grassroots labs such as the zero-waste lab. Both are citizen-empowering labs from a bottom-up perspective. In A Lab of Labs (Ferri & de Waal, 2017) the Waag Society explains how it uses game and narratives to design inclusive interactions with the citizens of Amsterdam. The Urban Analytics studies the city data in the urban living labs of these partners.

Project "Design thinking for the Circular Economy"
(Raak MKB, in preparation)
The design of technical systems for sustainable and social neighborhoods. Small Business Enterprises develop the technical solutions for citizens' communities. Citizens' participation in local production chains (reuse of waste products, energy, food) need digital (communication) platforms, technology and sustainable business models to create an inclusive successful circular economy. This project studies how design methodologies can help to introduce socio-technical systems in local communities to build a circular economy.

Project Monitor "Gezonde Stad" (ongoing collaboration)
Every April a monitor of the sustainability goals of the city of Amsterdam is presented. It consists of five fields: Green spaces, Waste, Clean Air, Energy and Food. The monitor is updated annually (by the Gezonde Stad and the Urban Analytics team) and presented at a festive meeting ("Wij maken de Stad") in April.

Mobility

Busy streets are a serious concern for citizens and the municipal governments. People and goods compete for space in city streets. The new subway line "Noord Zuid lijn" will start running in 2018, with major consequences for public transport in Amsterdam. City services such as garbage collection cause traffic streams in all parts of the city at all hours. Online shopping reduces shop delivery logistics, but results in many diesel fuel buses servicing the last mile of the supply chains. Ploos van Amstel (2015) shows that 10-15% of the traffic kilometers in the city are freight traffic, with examples of 40% during morning and evening rush hours.

Temporal and spatial studies with dynamic (real time) analysis of the use of the public domain are possible through the use of data from sensors, cameras and from smart devices brought by the participants such as mobile phones, GEO and Wi-Fi trackers. The Urban Analytics team will study the logistics in city shopping districts, a study of the impact of the Noord Zuid subway line and dynamic planning of garbage collection.

Project Noord-Zuid Lijn
Joint project plan with Rob van der Mei (CWI), Elenna Dugundji (CWI/VU), GvB, Municipality of Amsterdam, UvA. This project monitors the interaction of citizens (travelers) with the new subway line in the city of Amsterdam. The subway line will open in 2018 and will replace some of the current public transport routes. The research explores what the new transport behavior of citizens will be.

Project Logistic profiling of shopping streets
Dapperbuurt, Museumkwartier, Haarlemmerstraat, current project with Municipality of Amsterdam (Logistics group).
These streets face many actors that interact, such as delivery of goods, attractive environments, customers' access, garbage collection, and parking.

Project Dynamic data-driven planning of garbage collection
Current project with Municipality of Amsterdam (Garbage project team) and Walther Ploos van Amstel.
Efficient dynamic route planning, in combination with smart solutions for reduce of waste, separate collection of reducible waste products and the optimal location of collection stations all rely on data and data analysis. For instance, accurate prediction of the fill rate of underground garbage containers is a crucial element for the dynamic route planning algorithms.

These projects are complemented with an academic study on scaling of cities. With the students of the AUAS, the Urban Analytics team will use urban data to study the city of Amsterdam and to apply scaling methodology by West (2017) as described in Chapter 1. The rapid introduction of the digital society will probably disrupt these universal laws of scaling. Based on the scaling laws found so far, we will study the effect of the digital transformation in the field of energy, livability and mobility.

The Big Data Station

In the Faculty of Technology there are three connections with educational programs that form the basis of the future plans. We did start small with a data science track in the bachelor program Applied Mathematics. Next, the thesis-lab Data Science was developed for fourth-year students of all bachelor programs in the Faculty of Technology. In this lab students work together on data projects from industry, and each student writes a thesis. Last year the Big Data in Urban Technology minor was launched, a 20-week specialization track in Data Science, open to all bachelor students of the AUAS. The goal is to upscale relations between research education and applications and to start a co-creation lab on data science, the Big Data Station:

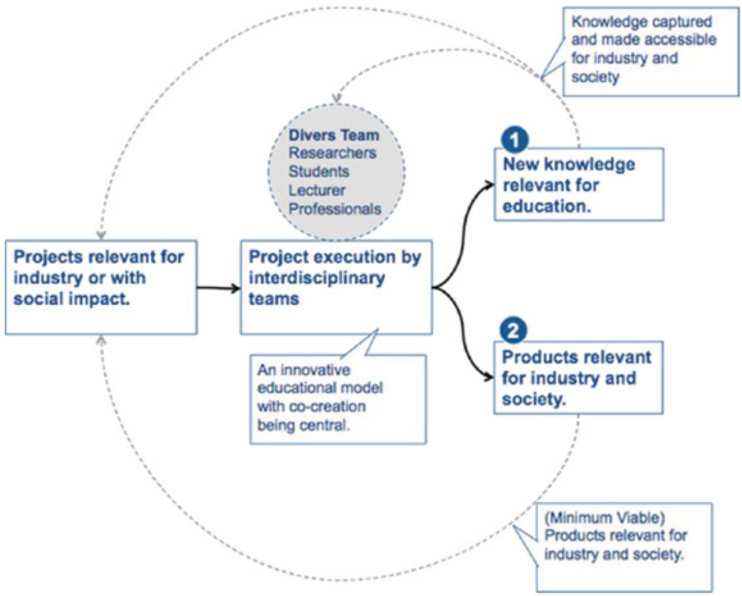


Figure 15 Big Data Station principle

The ambition of the Urban Analytics group is to become a “one stop shop” for data science projects.

The aim is to create a permanent environment (STATION) in the Faculty of Technology in collaboration with the bachelor programs. Students, researchers and industry professionals work together on Big Data challenges. Students in interdisciplinary teams can explore the viability of their ideas and find ways to turn them into fully functioning prototypes.
The collaborative environment is case-driven, which means that students and researchers will be working on real-life cases, with real data from industry and society. Activities range from complete education programs (e.g., educational track Big Data) to co-curricular activities (e.g., inspiration sessions and hackathons).

Lecturers are invited to translate the activities of students into curriculum points, researchers (and students) bring new knowledge and techniques into the station and industry share their projects and challenges. In this environment, learning is stimulated and facilitated by the researchers and lecturers of the faculty using online material on data science. The online training can be used by students and professionals and tailored to suit the individual needs. The Big Data Station provides innovative education for the Faculty of Technology in three ways.
First, the lecturers will be able to keep their knowledge up to date by working on Big Data challenges. Second, the curriculum becomes flexible. Finally, students develop life-long learning skills with an active role in the co-creation process.

The Urban Analytics research group will first connect and disclose expertise, and then develop new expertise as needed. The station builds knowledge cells, but can also scale down by deserting knowledge cells that are no longer useful in the research projects. The research environment connects to the partners’ living labs. If needed, the projects can be done in the urban analytics research environment, an experimental research environment that is available for participants in the projects.

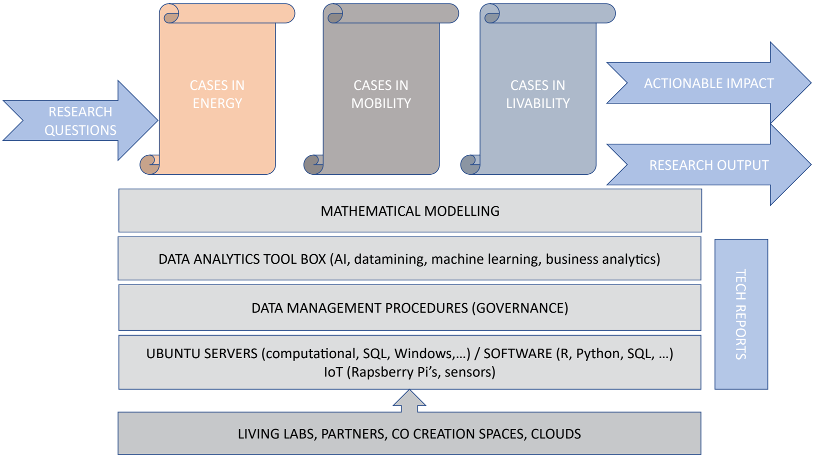


Figure 16 Architecture of the Urban Analytics Research Environment

The faculty of DMCI has invested in data labs, such as Citizen Data Lab, MediaLAB Amsterdam, Makers Lab, Publishing Lab, Interaction and Games Lab, where students and researchers are working on visualization techniques and design studies. The Faculty of Technology has invested in a Sensor lab, a Simulation lab and the Data Science thesis lab.

These labs can be partners of the Big Data Station, as independent labs, but with shared projects and knowledge transfers.

The station (or lab) is a place where students, lecturers, researchers and industry partners can collaborate on data projects. It supports a data infrastructure and has data science knowledge in different sections.

Smart prototypes

Communicating the results of the data analysis can be the end of the data project, and very likely the start of a new cycle. With smart devices, data can be collected and analyzed in real time on the device itself, even resulting in actionable devices (such as a smart device that turns the light off if the device detects no humans in the room). In this way we can combine data analytics and mathematical modelling with the technical infrastructure of a city.

In autonomous systems, one can program negotiation, auction, stochastic, and simulation models. The socio-technical systems with the technical prototypes can only work properly when (artificial) intelligence is programmed on it. This can be simple add-ons (recognizing malfunction, postponing the start of a load session for an EV) but more interesting applications are the recognition of busy streets, with suggestions for alternative routes for approaching cars, reservations and negotiation models for parking spots and loading zones, and exchange of (solar) energy among partners in an energy grid.

The AUAS can build prototypes of technical equipment with intelligent algorithms developed at CWI. With its business partners, the sensor lab of the Urban Technology group, the engineering department and the Smart City Academy this creates an interesting application environment for the intelligent algorithms of the CWI.

Another advantage of the use of smart devices is that it can solve privacy issues in data collection, for instance camera detection devices that record the number of people in a room, but do not save images; this is an example of privacy by design. It requires computational power and algorithms that operate on the technical device, making it smart.

Digital Society School

In February 2018 the the "Data-driven Transition" track of the Digital Society School will start. Within this track we will study, develop and conduct applied research on how society can use data and gather new insights to help or accelerate societal transformations. Participants in the Digital Society School will learn how (digital) technology can have a positive impact on society. It will complement the development of technical systems by implementing technologies in society. The Urban Analytics team will be part of the developing team of the track Data-driven Transition. This track is closely connected to the Big Data Station and the projects in the energy domain, creating a co-creation community with different expertise and experience.

Urban Analytics will have to choose the project wisely. There are three domains (energy, mobility and livability), three levels of engagement (academic and applied research, education) and many partners in business, government and citizens.

Literature

van der Aalst, W. M. P., Bichler, M. & Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering 59* (5), p. 311-313.

Anscombe, F.J. (1973). Graphs in statistical analysis, *The American Statistician, 27*(1), p.17-21.

Al Ars, Z. (2017). Hardware-accelerated big data analytics for genomics analysis. Presentation at Big Data Alliance meeting, https://www.bigdata-alliance.org/wp-content/uploads/2017/04/Presentatie-Zaid-Al-Ars-Big-Data-Alliance-Member-Meeting.pdf

Bahrami, M., & Singhal, M. (2015). The Role of Cloud Computing Architecture in Big Data, In W. Pedrycz & S.M. Chen (Eds.). *Information Granularity, Big Data, and Computational Intelligence*. Switzerland: Springer International Publishing.

Ballon, P. (2016). *Smart Cities: Hoe technologie onze steden leefbaar houdt en slimmer maakt* (in Dutch), Uitgeverij Lannoo nv, Tielt.

Barthelemy, M. (2016). *The structure and dynamics of cities*. Cambridge: Cambridge University Press.

Berkelmans, G., Berkelmans, W., Piersma, N., van der Mei, R. & Dugundji, E. (2018). *Predicting electric vehicle charging demand using mixed generalized extreme value models with panel effects*. Report, Amsterdam University of Applied Science, submitted.

Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science, 340*, pp. 1438-1441.

Ciobanu, R.I., Cristea, V., Dobre, C. & Pop, F. (2014). Big data platforms for the Internet of Things, In N. Bessis & C. Dobre (Eds.), *Big data and Internet of Things:A roadmap for smart environments, studies in computational intelligence 546*. Switzerland: Springer International Publishing.

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. Cambridge: MIT Press.

Ferri, G., & de Waal, M., Eds. (2017). *A Lab of Labs: Methods and Approaches for a Human-Centered Design*. Amsterdam: Amsterdam Creative Industries Publishing.

Florida, R. (2017). *The new urban crisis: How our cities are increasing inequality, deepening segregation, and failing the middle class and what we can do about it*. New York: Basic Books.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanović, N., & Meijers, N. (2007). *Smart cities, ranking of European medium-sized cities*. Report of the Research Project, www.smart-cities.eu.

Giffinger, R., Haindlmaier, G., & Kramar, H. (2010). The role of rankings in growing city competition, *Urban Research and Practice 3*(3), pp. 299-312.

GoDataDriven (2017). How European organizations increase their innovation speed with smart data applications, *Big Data Survey 2017*, godatadriven.com

Grey, T., Dyer, M. & Gleeson, D. (2017). Using big and small urban data for collaborative urbanism. In *Springer Tracts in Civil Engineering Citizen Empowerment and Innovation in the Data-Rich City*, C. Certomà, M. Dyer, L. Pocatilu, & F. Rizzi, Switzerland: Springer International Publishing.

GUST (2017). The emerging landscape of urban living labs: Characteristics, practices and examples. EU project Governance of Urban Sustainability Transitions, www.urbanlivinglabs.com

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. 3rd ed. San Francisco, CA: Morgan Kaufmann Publishers.

Harrison, C., Eckman, B., Hamilton, R., Hartswick, P., Kalagnanam, J., Paraszczak, J. & Williams, P. (2010). Foundations for smarter cities. *IBM Journal of Research and Development 54*(4), pp. 1–16.

Hodges, A. (1983). *Alan Turing: The enigma*. New York: Simon and Schuster.

Hua, X.S. (2017). *Hangzhou City Brain – using video and image recognition technologies to create the world's most advanced smart traffic management system*. Presentation at the World AI Summit, October 2017, Amsterdam.

IBM (2017). *10 key marketing trends for 2017 and ideas for exceeding customer expectations*. White paper, https://www-07.ibm.com/in/solution/10_Key_Marketing_Trends_for_2017.html

Kahn, R.H., & Khan, J. (2013). A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network. *Computer Networks 57*, pp. 825-845.

Keim, D.A., Mansmann, F., Schneidewind, J. Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. In S.J. Simoff, M.H. Böhlen,A. Mazeika (Eds.): *Visual data mining, LNCS 4404* (pp. 76-90). Berlin and Heidelberg: Springer-Verlag.

Kitchin, R. (2014). *The data revolution, big data, open data, data infrastructures and their consequences*. London, Los Angeles, New Delhi, Singapore, Washington D.C.: Sage.

Kurzweil, R (2006). *The singularity is near*. New York: Penguin Putnam Inc.

Leopold, G. (2014 May 6). 'Big spectrum data' could help predict wireless net usage, https://www.datanami.com/2014/05/06/big-spectrum-data-help-predict-wireless-net-usage/.

Maase, S.J.F.M., Dilrosun, X.F., Kooi, M.J.W., & Hoed, van den, R. (2017). Performance of EV charging infrastructure: A decision support tool based on charging data. Proceedings of the 30th International Electric Vehicle Symposium, Stuttgart, Germany.

Mocanu, E. (2017). *Machine learning applied to smart grids*. PhD dissertation, Eindhoven University of Technology.

Van Nieuwenhuyse, D., & Vanhoudt, D. (2008). *Performance management, Van prestatie-meting naar prestatiemanagement door de toepassing van analytische intelligentie* (in Dutch), LannooCampus.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishers.

Piersma, N., van Elzakker, F., van der Poel, G., Kluitenberg, J., van de Reijt, A., Talens, H., van den Hoed, R., Voogt, N., & Ferwerda, R. (2017, May). KENCIJFERS Openbare Laadinfrastructuur Elektrisch Vervoer, onderzoeksverslag Nationaal Kennisplatform Laadinfrastructuur NKL (in Dutch), http://www.nklnederland.nl/nieuws/kencijfers-laadinfrastructuur-nu-beschikbaar/.

Piersma, N., Van der Poel, G., Gozems, N., & Van den Hoed, R. (2018). *Monitoring the usage of an EV charge infrastructure, a fuzzy model for neighborhoods in The Hague, the Netherlands* (in Dutch), Report of the Amsterdam University of Applied Science.

Ploos van Amstel, W. (2015). CityLogistiek, op weg naar een duurzame stadslogistiek voor aantrekkelijke steden, Inauguration lecture 2015 (in Dutch), Amsterdam University Press.

Sakr, S. (2016). Big data 2.0 processing systems: A survey, *SpringerBriefs in Computer Science*, Switzerland: Springer International Publishing.

Steenbrink, M., Jochemsen, T., Piersma, N., van der Mei, R., & Dugundji, E. (2016, September). A prediction model for the demand of charging stations for electric vehicles in the municipality of Amsterdam, *EV Heart 2016 Proceedings*, Delft.

Thierauf, R. J. (1999). *Knowledge management systems*. New York: Quorum Books.

Tufte, E. (1985). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

USA Patriot Act (2013). Bulk collection of telephony metadata under section 215 of the USA Patriot Act, Administration White Paper.

Van den Hoed, R., & Helmus, J. (2015). Unraveling user type characteristics: Towards a taxonomy for charging infrastructure. *Proceedings of the 28th International Electric Vehicle Symposium*, Kintex, Korea.

Van Raan, A.F.J., van der Meulen, G., & Goedhart, W. (2015). Urban scaling of cities in the Netherlands. Report of the Centre for Science and Technology Studies. The Netherlands: Leiden University.

West, G. (2017). *Scale, the universal laws of life and death in organisms, cities and companies*. London: Weidenfeld & Nicolson (Orion Publishing Group Ltd.).

Wu, D., Sakr, S. & Zhu, L. (2017). Big data storage and data models, In Zomaya, A., Sakr, S. (Eds.) *Handbook of big data technologies*. Switzerland: Springer International Publishing.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing. *ACM Transactions on Intelligent Systems and Technology*, *5*(3) pages.

Notes

1. Angus McDonald, the Wisdom-Knowledge-Information-Data Pyramid originates from Thierauf (1999).
2. Graph by Leopold (2014).
3. https://autoriteitpersoonsgegevens.nl/en/news.
4. See https://www.cisco.com/c/en/us/solutions/industries/smart-connected-communities/kinetic-for-cities.html.
5. Graphics from Wu et al., 2017.
6. Contact smartcityacademy@hva.nl.
7. Developed by Bas Breur, Hans den Boer, Sterre van der Poll, Thairis da Silva Faria as part of the minor Big Data in Urban Technology.
8. Graph by Gartner, see Gartner.com.
9. See an introduction of the IAS group at https://www.youtube.com/watch?v=kjmPrN5iWw0&index=9&list=PLtBl3dSNhbhgXwyNPSE8hJgySy6d18RWZ&t=5s
10. Images from the Marginal Revolution, Type I and Type II Errors Simplified, by Alex Tabarrok on May 10, 2014. http://marginalrevolution.com/marginalrevolution/2014/05/type-i-and-type-ii-errors-simplified.html

Curriculum Vitae

Nanda Piersma (b. 1963) is Professor of Urban Analytics at the Amsterdam University of Applied Science and Centrum Wiskunde & Informatica.

She finished her PhD in 1994 on Stochastic Optimization, and she worked on optimization models at the Econometric Institute at the Erasmus University for ten years. She supervised many master students and PhD projects in domains such as health, marketing, and logistics.

Professional education is her passion, preparing students for their first job and working with industry partners in projects and lifelong learning activities. Working in a number of positions, Nanda gained experience and expertise in developing and managing education programmes in Mathematics and Logistics and in interacting with industries. At the Amsterdam University of Applied Science, she managed the bachelor's programmes in Applied Mathematics, Computer Science, International Business Management Studies, Logistics Engineering and Logistics Management. With industry partners and academics, she was one of the founders of the KennisDC Logistiek and part of the Human Capital group in Logistics which set the research agenda for the Topsector in Logistics.

With her return to data-related research in 2014, she started the research group Urban Analytics as part of the Urban Technology research program of the Faculty of Technology at AUAS. This is a group of data lovers who research how organizations in cities can gain better insights from their data through data analysis, machine learning and business analytics. The City of Amsterdam is her second passion (de "Lieve Stad"), from the canals to Ajax, and all its residents.

Managing her time between CWI, the Faculty of Digital Media and Creative Industries and the Faculty of Technology, she works on data-driven solutions for city challenges, with Amsterdam as her main living lab.

She is a board member of Amsterdam Data Science (ADS), and part of the coordinator team of Amsterdam Data Exchange initiative (AMDEX).