Natalia Andrienko
Gennady Andrienko · Georg Fuchs
Aidan Slingsby · Cagatay Turkay
Stefan Wrobel

Visual Analytics for Data Scientists

Springer

Visual Analytics for Data Scientists

Natalia Andrienko • Gennady Andrienko
Georg Fuchs • Aidan Slingsby • Cagatay Turkay
Stefan Wrobel

Visual Analytics for Data Scientists

Springer

Natalia Andrienko
Fraunhofer Institute Intelligent
Analysis and Information Systems IAIS
Schloss Birlinghoven
Sankt Augustin, Germany

Department of Computer Science
City, University of London
Northampton Square, London, UK

Gennady Andrienko
Fraunhofer Institute Intelligent
Analysis and Information Systems IAIS
Schloss Birlinghoven
Sankt Augustin, Germany

Department of Computer Science
City, University of London
Northampton Square, London, UK

Georg Fuchs
Fraunhofer Institute Intelligent
Analysis and Information Systems IAIS
Schloss Birlinghoven
Sankt Augustin, Germany

Aidan Slingsby
Department of Computer Science
City, University of London
Northampton Square, London, UK

Cagatay Turkay
Centre for Interdisciplinary Methodologies
University of Warwick
Coventry, UK

Stefan Wrobel
Fraunhofer Institute Intelligent
Analysis and Information Systems IAIS
Schloss Birlinghoven
Sankt Augustin, Germany

University of Bonn
Bonn, Germany

ISBN 978-3-030-56145-1 ISBN 978-3-030-56146-8 (eBook)
https://doi.org/10.1007/978-3-030-56146-8

To our families, friends, colleagues and partners

Preface

There are several disciplines concerned with developing computer-oriented methods for data analysis: statistics, machine learning, data mining, as well as disciplines specific to various application domains, such as geographic information science, microbiology, or astronomy. Until recent years, it was customary to believe in numbers and rely in data analysis solely on mathematical and computational techniques. Visualisation appeared after the end of the analysis processes for making illustrations for reports. Currently, the situation is different, and the importance of visualisation as a means to convey information to humans for involving human understanding and reasoning in the processes of analysing complex data and solving complex problems is acknowledged. As a sign of this, data visualisation is included in the list of the skills required for being a data scientist.

Visual analytics is a research discipline that is based on acknowledging the power and the necessity of the human vision, understanding, and reasoning in data analysis and problem solving. It aims at developing methods, analytical workflows, and software systems that can support unique capabilities of humans by providing appropriate visual displays of data and involving as much as possible the capabilities of computers to store, process, analyse, and visualise data. During the time of the existence of visual analytics, the researchers have created a large amount of knowledge on human-computer data analysis that could be used by practitioners, particularly, by data scientists. There is a problem, however: the practicable knowledge is scattered over a great number of scientific papers, which are not only abundant but also not practitioner-oriented. It is very hard for non-researchers to find what can be useful to them in this ocean of publications.

At the same time, the use of visual representations in data analysis becomes a widely accessible and seemingly easy activity due to the appearance of the Python, R and Javascript languages and proliferation of open-access packages with codes for data processing, analysis, modelling, and visualisation. Creation and execution of analytical workflows involving both computations and visualisations is supported by the Jupyter Notebook application, and there are myriads of analytical notebooks created by various people and published on the Web. These notebooks are often taken by other people for being adapted to their needs or as examples of what is possible.

While this is a very positive development, it has a back side. The notebooks are often created or adapted by people having quite little idea of how to choose appropriate visualisation techniques and design correct and effective visualisations of the data they deal with, and also have no good understanding of why, when, and how visualisations need to be used in analysis. Some visualisations occurring in the publicly accessible example notebooks may look impressive and convincing to non-specialists, but, in fact, they may communicate spurious patterns in inadequate ways. Those who view these visualisations and think of doing the same for their data and tasks often lack knowledge that would enable critical assessment and understanding of the suitability of the techniques. Other notebooks include only basic graphics having little analytical value, whereas better ways exist for representing the relevant information.

Besides poor visualisation literacy, detrimental for analysis is a propensity to uncritically trust computers and take the outcome of a single run of an analysis algorithm with default parameter settings, or with settings previously used by someone else, as the final result. Naive analysts may not realise that a slight change in the data or parameters can sometimes significantly change the result; therefore, they may not bother to examine the reaction of the algorithm to such changes and to check results of several runs for consistency. More experienced and critically minded analysts, who usually take the trouble to evaluate and compare what they get from computers, may tend to rely solely on statistical measures rather than trying to gain better understanding with the help of visualisations.

Visual analytics has not only generated the body of knowledge on how to create meaningful visualisations and how to use them effectively in data analysis together with computer operations but also developed a philosophy that should underlie analytical activity. The main principles are the primacy of human understanding and reasoning and awareness of the weaknesses of computers, which cannot see, understand, and think, and thus need to be led and controlled by humans. Both the knowledge and the philosophy should be transferred to practitioners to help them to do better analyses and come to valid conclusions. With this textbook, we make an attempt to do this.

In this book, we do not aim to present the latest results of the visual analytics research and the most innovative and advanced techniques and methods. Instead, we shall present the main principles and describe the techniques and approaches that are ready to be put in practice, which means that they, first, proved their utility and, second, can be reproduced with moderate effort. We put emphasis on describing examples of analyses, in which we explain the need for and the use of visualisations.

Sankt Augustin,
London
June 2020

Natalia Andrienko
Gennady Andrienko
Georg Fuchs
Aidan Slingsby
Cagatay Turkay
Stefan Wrobel

Acknowledgements

This book is a result of our collaboration with many different groups of partners. We are thankful to

- visual analytics researchers who developed methods and analytical workflows presented and discussed in out book;
- students of the master program on data science[1] at the City, University of London, whose feedback on our visual analytics module helped us to understand what and how needs to be taught to them;
- numerous project partners in a series of research projects funded by EU (European Union) and DFG (German Research Foundation), who introduced us to a variety of application domains of visual analytics, helped us to shape and develop our vision, and critically evaluated our approaches;
- our colleagues at Fraunhofer Institute IAIS – Intelligent Analysis and Information Systems[2], especially the excellent team of the KD – Knowledge Discovery department[3], and at City, University of London[4], specifically, the dynamic and vibrant team of GICentre[5];
- our colleagues Linara Adilova and Siming Chen - for careful reading of different versions of our manuscript and providing helpful comments and constructive critiques.

[1] https://www.city.ac.uk/study/courses/postgraduate/data-science-msc

[2] www.iais.fraunhofer.de/en.html

[3] https://www.iais.fraunhofer.de/en/institute/departments/knowledge-discovery-en.html

[4] www.city.ac.uk

[5] www.gicentre.net

Writing of the book was financially supported by the German Priority Research Program SPP 1894 on Volunteered Geographic Information, EU projects Track&Know and SoBigData++, EU SESAR project TAPAS, and Fraunhofer Cluster of Excellence on “Cognitive Internet Technologies”.

Contents

Acronyms

ACF	**AutoCorrelation function**: correlation of a signal with a delayed copy of itself as a function of delay
AI	**Artificial Intelligence**
ASW	**Average Silhouette Width**: a cluster quality measure
AUC	**Area Under the Curve**: a measure of the quality of a classifier; see ROC.
CMV	**Coordinated Multiple Views**: two or more visual displays where special techniques support finding corresponding pieces of information.
DTW	**Dynamic Time Warping**: a distance function for measuring similarity for time series data.
GPS	**Global positioning system**: a satellite-based navigation system that enables a GPS receiver to obtain location and time data without a requirement to transmit any data.
IQR	**interquartile range**: the difference between the upper (third) and the lower (first) quartiles of a set of numeric values
KDE	**Kernel Density Estimation**: a statistical technique used for data smoothing
LDA	**Latent Dirichlet Allocation**: a topic modelling method
LSA	**Latent Semantic Analysis**: a topic modelling method
MDS	**MultiDimensional Scaling**: a method for data embedding
ML	**Machine Learning**
NER	**Named Entity Recognition**: a text processing method
NMF	**Nonnegative Matrix Factorization**: a topic modelling method

OSM **OpenStreetMap**: a database of crowdsourced worldwide geographic information and a set of services, including generation and provision of map tiles

PCA **Principal Component Analysis**: a method for dimensionality reduction

PLSA **Probabilistic Semantic Analysis**: a topic modelling method

POI **Point Of Interest**, or **Place Of Interest**: a place in the geographical space.

RMSE **Root Mean Squared Error**: a measure of model error

ROC **Receiver Operating Characteristic** Curve: a plot for assessing classifier's performance

SOM **Self-Organising Map**: a machine learning method that can be used for data embedding

STC **Space-time cube**: a visual display providing a perspective view of a 3D scene where the dimensions represent 2D space and time.

SVM **Support-vector machine**: a machine learning algorithm for classification and regression analysis

t-SNE **T-distributed Stochastic Neighbor Embedding**: a method for data embedding

U-matrix **unified distance matrix**: the matrix of distances between data items represented by nodes (neurons) of a network resulting from SOM

WGS84 **World Geodetic System** established in 19**84**: a standard reference system for specifying geographic coordinates

XAI **eXplainable Artificial Intelligence**

Part I

Introduction to Visual Analytics in Data Science

Chapter 1

Introduction to Visual Analytics by an Example

Abstract An illustrated example of problem solving is meant to demonstrate how visual representations of data support human reasoning and deriving knowledge from data. We argue that human reasoning plays a crucial role in solving non-trivial problems. Even when the primary goal of data analysis is to create a predictive model to be executed by computers, this cannot be done without human reasoning and derivation of new knowledge, which includes understanding of the analysis subject and knowledge of the computer model built. Reasoning requires conveying information to the human's mind, and visual representations are best suited for this. Visual analytics focuses on supporting human analytical reasoning and develops approaches combining visualisations, interactive operations, and computational processing. The underlying idea is to enable synergistic joint work of humans and computers, in which each side can effectively utilise its unique capabilities. The ideas and approaches of visual analytics are therefore very relevant to data science.

1.1 What is visual analytics? (A brief summary)

Visual analytics is the science and practice of analytical reasoning by combining **computational processing** with **visualisation**. These are tightly-coupled using **interactive** techniques so that each informs the other. In this book, we will discuss and illustrate the benefits of this approach for doing analysis.

The involvement of computational processing allows computers to do **what computers are good at**: transforming and summarising data and searching for specific pieces of information. Interactive visual interfaces involve the human analyst, allow him or her do **what human analysts are good at**: interpreting and reasoning. There is a long tradition of using visualisation to help interpret results of computing, but visual analysis develops this further, emphasising the benefits of an iterative cycle of doing computations, understanding and evaluating the results to refine the compu-

N. Andrienko et al., *Visual Analytics for Data Scientists*,
https://doi.org/10.1007/978-3-030-56146-8_1

tation analysis or to investigate complementary findings. In visual analytics, there is close-coupling between the use of computational techniques and the interactive visualisation in the data analysis process undertaken by the human analyst. One of the important principles and benefits of visual analytics is to **not take computational analysis results for granted**.

Computational methods include those that summarise (e.g. summary statistics), find relationships (e.g. correlations), and identify formally specified types of patterns (e.g. groups of objects that have similarities according to a certain computable criterion). Often, the analyst has to make decisions about how these methods are run: the input data (e.g. which subsets of data to use and how the data are processed) and the values of method parameters. Depending on the analytical question and the prior experience of the analyst, it may be more or less obvious as to which of these should be. Visual analytics can help the analyst try different parameter values and see what effects these have on the analysis.

As datasets become larger, computational methods get increasingly important for helping analysts. We cannot just plot everything and look at it, like with a small dataset. Recent advances in machine learning techniques (often referred to as Artificial Intelligence) raise high expectations concerning their power to identify important relationships and useful patterns in extremely large datasets. Many of these newer techniques are black boxes that are often taken for granted. There are plenty of reasons why this is problematic. Visual analytics approaches can be applied to these black boxes as to phenomena whose behaviour needs to be studied and understood. This is similar to studying real-world phenomena by analysing data related to them. In this way, visual analytics can help humans to come to some degree of understanding of these models and judge their appropriateness and reliability.

Interactive visual interfaces enable human analysts to see and interpret the outputs of computational models in the context of the input data and parameters. Crucially, they also offer the ability to tweak parameters or data subsets used by the computational analysis techniques and the ability to show what effects these input data and parameters have on the analysis.

The **visual analytics process** involves human interpretation of visually represented data or model outputs and taking decisions concerning the next analytical steps. The next steps will depend on the interpretation, the current knowledge, and the analytical goals. It might be to vary model inputs or parameters to see how sensitive it is to this variation. It might be more directed, refining the model, for example, by removing one of the less important factors. Visualisation would than help indicate the difference this made. It might prompt an investigation of subsets of the data – e.g., by category, for a particular place, by hour, or according to a natural break in the distribution of one of the variables. If a computational method has identified groups of similar (in some formally defined sense) data items, visual representations characterising each group may help the analyst judge whether these groups are helpful or whether different similarity criteria need to be used.

However, visual analytics is not only the science and conduct of careful and effective use of computational techniques, but it is, first and foremost, the science of human **analytical reasoning**, which does not necessarily require the involvement of sophisticated computing but does require appropriate representation of information to the human, so that it can be used in the reasoning. Visual representation is acknowledged to be the best for this purpose, and it is the task of computers to generate these representations from available data and results of computations. In the following, we present and discuss an example of an analysis process in which visual representations inform human reasoning.

1.2 A motivating example: Investigating an epidemic outbreak

Some readers of this book may have quite vague idea of what visual analytics is and may not understand why and how it can be useful for data scientists. Others may believe that visualisation can be good for communicating ideas and presenting analysis results but may not think of visual displays as tools for doing analysis. To help both categories of readers to grasp the basic idea of visual analytics, we shall start with an example showing visual analytics approaches at work. We shall then discuss this example and outline in a more general way where and how visual analytics fits in the data science workflows. This example comes from the IEEE VAST Challenge 2011 [1]. Although the data are synthetic, they were carefully constructed to resemble real data as much as possible. A good feature of these data is that they contain an interesting and even dramatic story, while similar real data may be either uninteresting or unavailable. Let's dive into the story.

1.2.1 Data and task description

Vastopolis is a major metropolitan area with a population of approximately two million residents. During the last few days, health professionals at local hospitals have noticed a dramatic increase in reported illnesses. Observed symptoms are largely flu-like and include fever, chills, sweats, aches and pains, fatigue, coughing, breathing difficulty, nausea and vomiting, diarrhoea, and enlarged lymph nodes. More recently, there have been several deaths believed to be associated with the current outbreak. City officials fear a possible epidemic and are mobilising emergency management resources to mitigate the impact.

We have two datasets. The first one contains microblog messages collected from various devices with GPS capabilities, including laptop computers, hand-held com-

puters, and cellular phones. The second one contains map information for the entire metropolitan area. The map dataset contains a satellite image with labelled highways, hospitals, important landmarks, and water bodies (Fig. 1.1). There are also supplemental tables for population statistics and observed weather data.

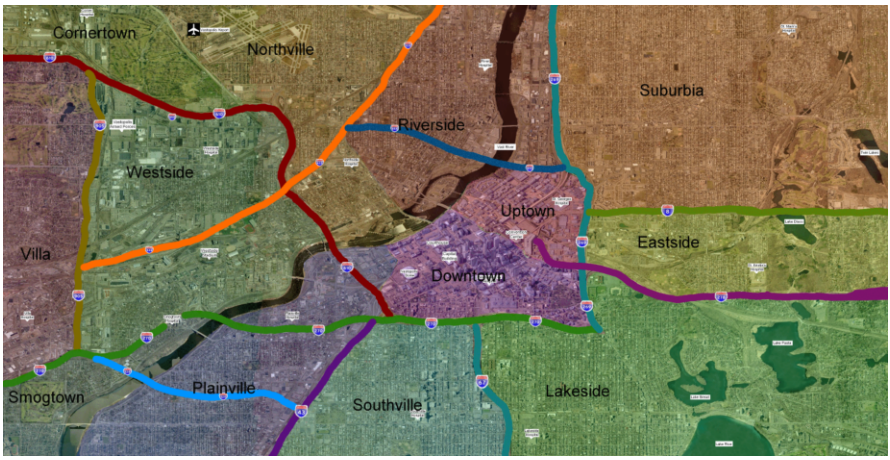

Fig. 1.1: A satellite image-based map of Vastopolis.

We need to find answers to the following questions:

- Identify approximately where the outbreak started (ground zero location). Outline the affected area.
- Present a hypothesis on how the infection is being transmitted. Is the method of transmission person-to-person, airborne, waterborne, or something else?
- Is the outbreak contained? Is it necessary for emergency management personnel to deploy treatment resources outside the affected area?

1.2.2 Data properties

We need to acknowledge that the available data do not directly represent disease occurrences; they just contain texts that may mention disease symptoms. We should not assume that the locations and times specified in the microblog records mentioning disease symptoms are the actual locations and times of disease occurrences. People may write about their health condition not necessarily immediately after getting sick and not necessarily from the location where they first felt some health problems. We should also keep in mind that not everyone who gets sick would send

a message about it, whereas some people may send more than one message. People may also write about someone else being sick. Besides, messages mentioning disease symptoms may appear not only during the time of epidemic outbreak but also at any other time. These specifics have the following implications for the investigation we need to do:

- We should keep in mind that the distribution of the microblog posts can give us only a rough approximation of the distribution of the disease cases.
- The epidemics may be manifested as patterns of increased temporal frequency and spatial density of disease-mentioning messages. This is what we shall try to find.

Hence, among all data records contained in the dataset, we need to identify the subset of the data that are related to the epidemic. This subset has two major characteristics: first, the texts of the messages include disease-related terms; second, the temporal frequency of posting such messages is notably higher than usual.

1.2.3 Data preparation

First we need to select the data that are potentially relevant to the analysis goals, that is, the messages mentioning health disorders. The task description lists some symptoms that were observed: fever, chills, sweats, aches and pains, fatigue, coughing, breathing difficulty, nausea and vomiting, diarrhoea, and enlarged lymph nodes. These keywords can be used in a query for extracting potentially relevant data records.

We perform querying in an interactive way. We start with putting the keywords from the task description in the query condition. After the query selects a subset of messages that include any of these keywords, we apply a tool that extracts the most frequent terms from these messages (excluding so called "stop words" like articles, prepositions, pronouns, etc.) and creates a visual display called text cloud, or word cloud (Fig. 1.2) using font size to represent word frequencies. In this display, we find other disease-related terms (e.g., flu, stomach, sick, doctor) that occur in the selected messages together with the terms that have been used in the query condition. We extend the query condition by adding these terms; the query extracts additional messages; in response, the word cloud display is updated to show the frequent words and word combinations from the extended subset of messages. We also find that some frequently used words shown in the word cloud are irrelevant (e.g., come, case, today, day, night, etc.), add them to the list of stop words, and make the word cloud update after exclusion of these words (Fig. 1.3).

Now we notice word combinations that appear irrelevant to the epidemic: "chicken flu" and "fried chicken flu" (Fig. 1.3). We apply another query to the selected sub-

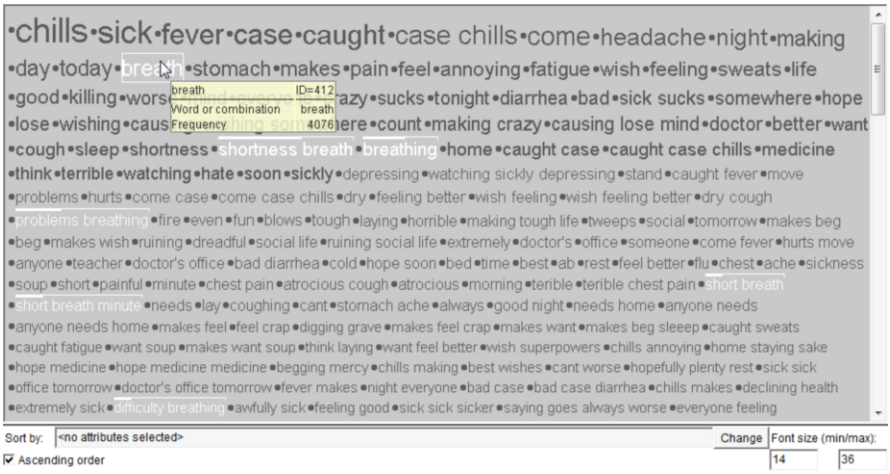

Fig. 1.2: Frequent terms extracted from the messages satisfying filter conditions.

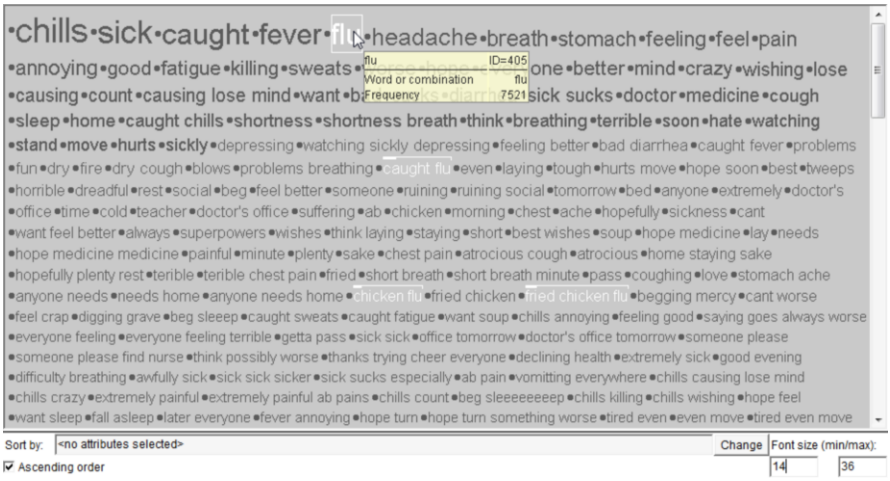

Fig. 1.3: The word cloud display has been updated in response to changing the query condition and extending the list of stop words.

set of messages, which selects only the messages containing the terms ‘chicken’ and ‘flu’. The word cloud changes as can be seen in Fig. 1.4. We also compare the temporal frequency distribution of all messages containing some disease-related terms and the messages containing the terms ‘chicken’ and ‘flu’. For this purpose, we use an interactive filter-aware time histogram, as in Fig. 1.5. The upper image shows the state of the time histogram after selecting the subset of messages containing any of the disease-related terms. Each bar corresponds to one day. The whole bar height is proportional to the total number of messages posted on that day whereas the dark

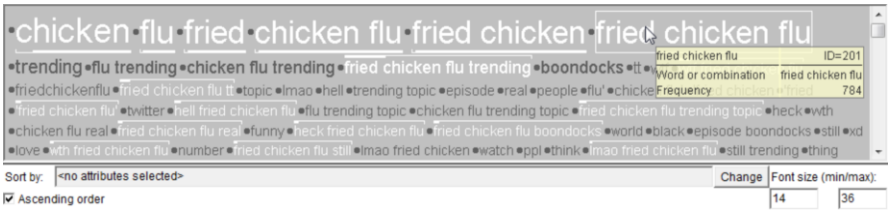

Fig. 1.4: Frequent words appearing in the messages containing the terms ‘chicken’ and ‘flu’.

segment represents the number of messages satisfying the current filter condition, i.e., containing any of the disease-related terms. We see that the frequency of such messages notably increases in the last three days. This corresponds to the statement in the task description: “During the last few days, health professionals at local hospitals have noticed a dramatic increase in reported illnesses”.

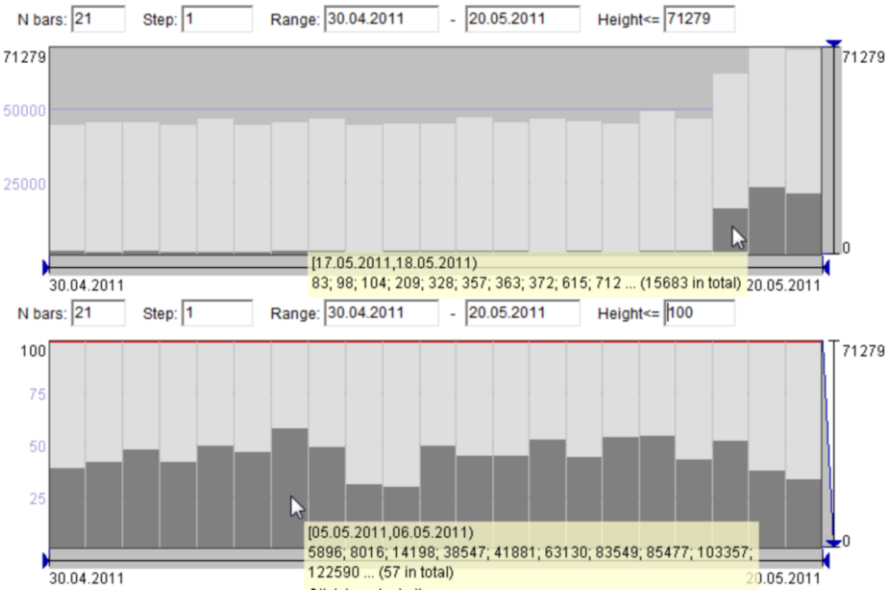

Fig. 1.5: Top: The time histogram shows the temporal frequency distribution of the massages containing disease-related terms. Bottom: The time histogram shows the temporal frequency distribution of the messages containing the terms ‘chicken’ and ’flu’.

The lower image in Fig. 1.5 shows the state of the time histogram after selecting the messages containing the words ‘chicken’ and ‘flu’. Since the dark segments were small and highly visible (because of low proportions of the selected messages

among all messages), we have changed the vertical scale of the histogram using an interactive focusing operation. We see that the messages related to the chicken flu are distributed more evenly throughout the time period covered by the data. The highest frequency of such messages was attained on the seventh day, i.e., long before the increase of the number of disease-related messages. This indicates that the messages mentioning the chicken flu are indeed irrelevant to the analysis task and should be filtered out. So, we exclude these messages from the further consideration. The remaining set consists of 79,579 messages, which is 7.8% out of the original set of 1,023,077 messages.

1.2.4 Analysing the temporal distribution

The temporal histogram (Fig. 1.5, top) shows us that the epidemic happened in the last three days, which are represented by the three rightmost bars of the histogram. More specifically, 59,761 out of the 79,579 disease-related messages (75%) were posted in the last three days. We can conclude that the epidemic happened in the last 3 days; however, we want to identify the time of the epidemic start more precisely. We use a time histogram with hourly temporal resolution, i.e., each bar corresponds to a time interval of 1 hour length (Fig. 1.6), where we see that the temporal frequency of the disease-related messages increased starting from 1 o'clock on May 18, then a very high increase happened at 9 o'clock of the same day, and a high peak occurred at 18 o'clock of that day. In the remaining days, the frequency was stably high except for drops in the night times (between 0 and 2 o'clock), which give us some evidence that the observed frequency increase at 1 o'clock on May 18 is indeed due to the epidemic outbreak start.

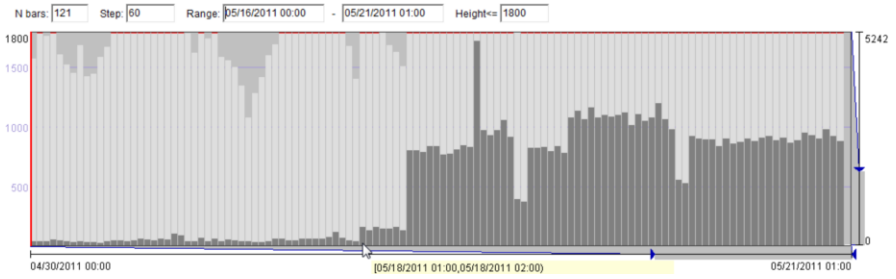

Fig. 1.6: A time histogram shows the temporal frequency distribution of the disease-related messages in the last 5 days by hourly intervals.

1.2.5 Analysing the spatial distribution

To analyse the spatial distribution of the outbreak-related messages (i.e., the messages mentioning disease symptoms that were posted in the last 3 days), we use a *dot map* (Fig. 1.7, top) in which the messages are represented by dots (small circles) in yellow. We observe quite prominent spatial patterns, namely, spatial clusters, which appear as areas with high density of the circle symbols. *Please note that in this and following maps of the message distributions we adjust the transparency of the symbols so that the patterns are best visible.*

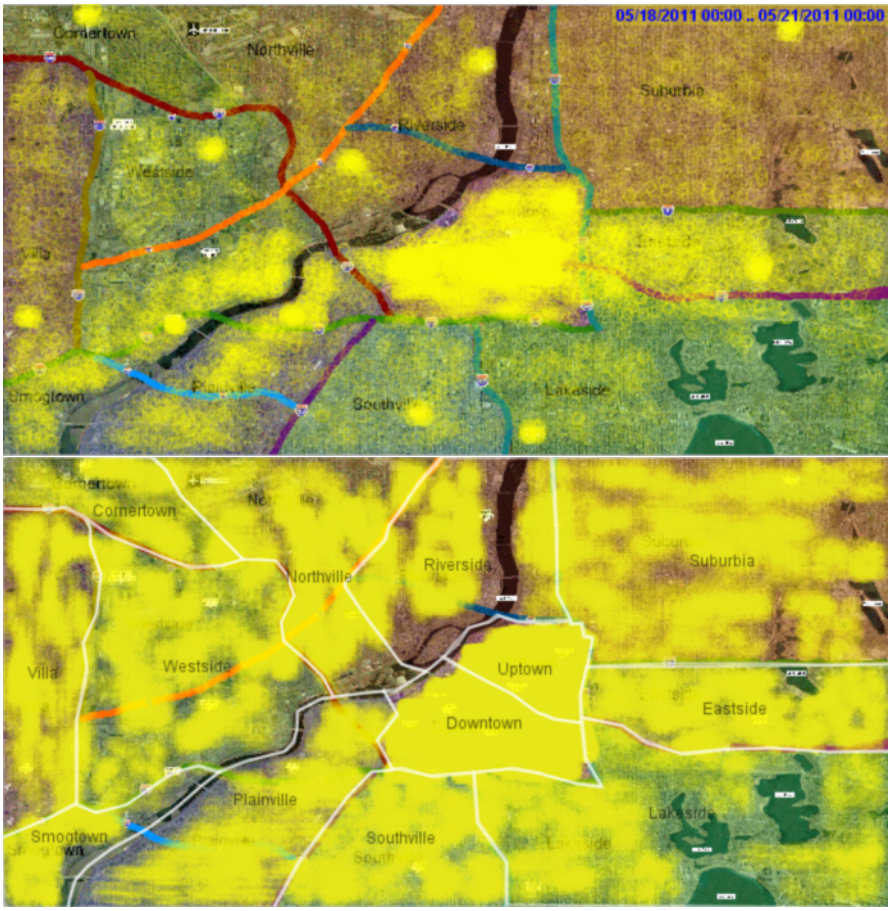

Fig. 1.7: Top: The dot map shows the spatial distribution of the epidemic-related messages. Bottom: The dot map shows the spatial distribution of the disease-unrelated messages.

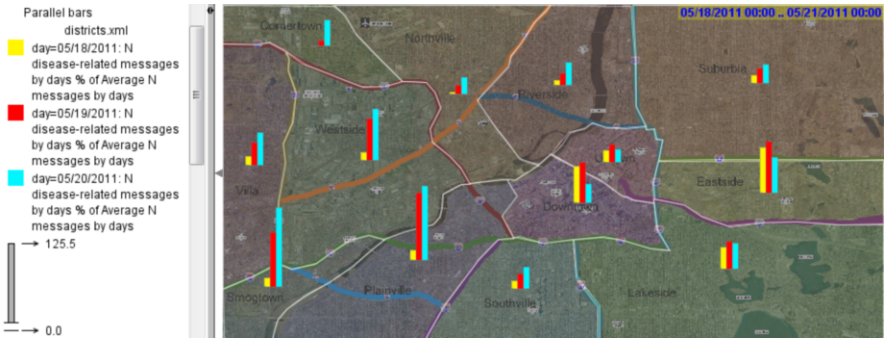

Fig. 1.8: The bar diagrams drawn within district boundaries show the ratios of the numbers of the disease-related messages in the three days of the epidemics to the average daily numbers of the messages posted in the districts before the epidemic outbreak.

Can the distribution of the outbreak-related messages be indicative of the distribution of the disease occurrences? To check this, we need to compare the spatial distribution of the outbreak-related messages to the spatial distribution of the unrelated messages. If these distributions look very similar, there would be no ground for taking the distribution of the messages as a proxy for the distribution of the disease occurrences. The lower dot map in Fig. 1.7 shows the spatial distribution of the disease-unrelated messages, which differs much from the distribution in the upper map. However, we notice a similarity: the density of the messages in the city centre is very high in both maps. Therefore, we cannot be sure that the dense cluster of disease-related messages in the city centre observed in the upper map is a hot spot of the disease outbreak or it corresponds to the generally high density of messages posted in this area.

1.2.6 Transforming data to verify observed patterns

To check whether the high density of the outbreak-related messages in the centre is due to the high spread of the disease in this area or due to the usual high message posting activity, we perform some calculations based on the available data. Using the boundaries of the city districts (visible in Fig. 1.7, bottom), we compute the average daily number of the messages posted in each district before the beginning of the outbreak. We also compute the number of disease-related messages posted in each of the three days of the epidemics. From these numbers, we compute the ratios of the numbers of the epidemic-related messages to the average daily message counts. The computed numbers are represented by bar diagrams in the map in Fig. 1.8.

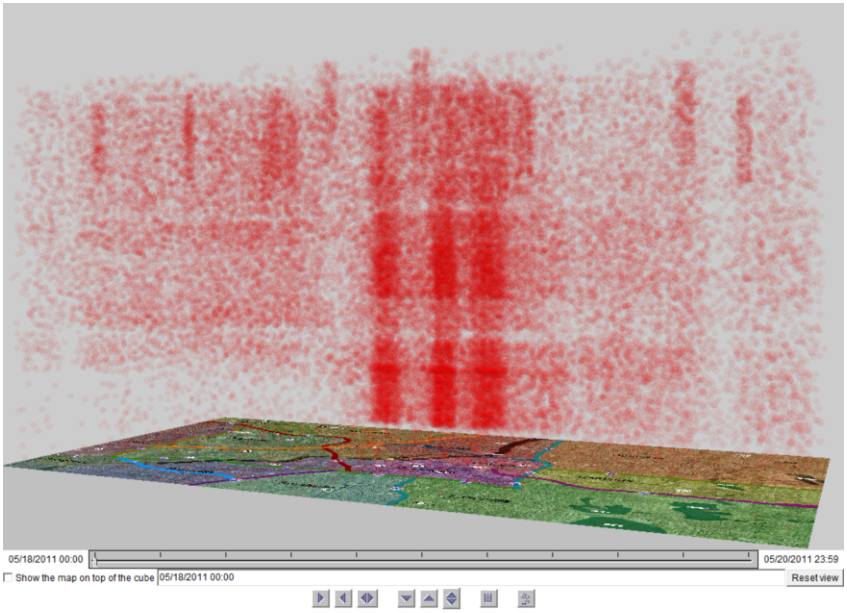

Fig. 1.9: The red dots are put in a space-time cube, where the horizontal plane represents the geographic space and the vertical dimension time, according to the spatial locations and posting times of the outbreak-related messages. The cube thus shows the spatio-temporal distribution of the messages.

The yellow, red, and cyan bars correspond to the first, second, and third day of the outbreak, respectively.

We see that one of the two central districts, called Downtown, has notably higher relative numbers of disease-related messages in the first day of the outbreak than the other districts, except the ones on the east and southeast. Hence, it can be concluded that this district was indeed hit by the outbreak on the first day. The other central district, called Uptown (northeast of Downtown), has only a slightly higher relative number of outbreak-related messages than other districts. However, this district covers a relatively small part of the dense cluster of disease-related messages. Hence, we can conclude that we see the cluster in the city centre because this area was hit by the outbreak and not just because it usually has high message posting activity.

1.2.7 Exploring the spatio-temporal distribution

Already the bar diagram map in Fig. 1.8, indicates that the spatial distribution of the outbreak-related messages was not the same during the three days of the epidemic. To see the evolution of the spatial distribution in more detail, we use a space-time cube (STC) display (Fig. 1.9). It is a perspective view of a 3D scene where the horizontal plane represents the geographic space and the vertical dimension represents time going in the direction from the bottom to the top. The epidemics-related messages are represented by dots (in red, drawn with low degree of opacity) positioned in the cube according to their spatial locations and posting times.

The observed gaps along the vertical dimension of the STC (i.e., time intervals of low density of the dots) correspond to the night drops in the message numbers observed earlier in a time histogram (Fig. 1.6). These gaps separate the three days of the outbreak.

We see that three very dense spatio-temporal clusters of messages, i.e., very high concentrations of messages in space and time, emerged on the first day of the epidemics. We use a dot map to see better the spatial footprints of the clusters (Fig. 1.10, top). It appears that the disease might originate from these three places, or, even more probably, these were areas visited by many people on the first day of the outbreak. Relatively high message density was also on the east of the three central clusters. By the end of day 1 and during day 2, the spatial spread of the messages increased; in particular, the density of the messages increased on the southwest of the city. In the third day, multiple spatially compact clusters emerged. The map in Fig. 1.10, bottom, shows that these clusters are located around hospitals, which indicates that ill people came to hospitals.

1.2.8 Revealing the patterns of the disease spread

Ill people might have posted messages concerning their health condition multiple times. To see how the disease spread, it is reasonable to look only at the distribution of the messages where disease symptoms were mentioned for the first time. To separate such messages from the rest, we apply another transformation to the data. Each record contains an identifier of the person who posted the message. We link the disease-mentioning messages of each person into a chronological sequence. There are 27,446 such sequences, and this is the number of individuals who supposedly got sick (it is 37% of the 73,928 distinct individual's identifiers occurring in the dataset). The lengths of the sequences vary from 1 to 6. Now we take only the first message from each sequence and look at the spatio-temporal distribution of these messages using a space-time cube display, as in Fig.1.11. The distribution differs from that of all outbreak-related messages (Fig.1.9). Most notably, we don't

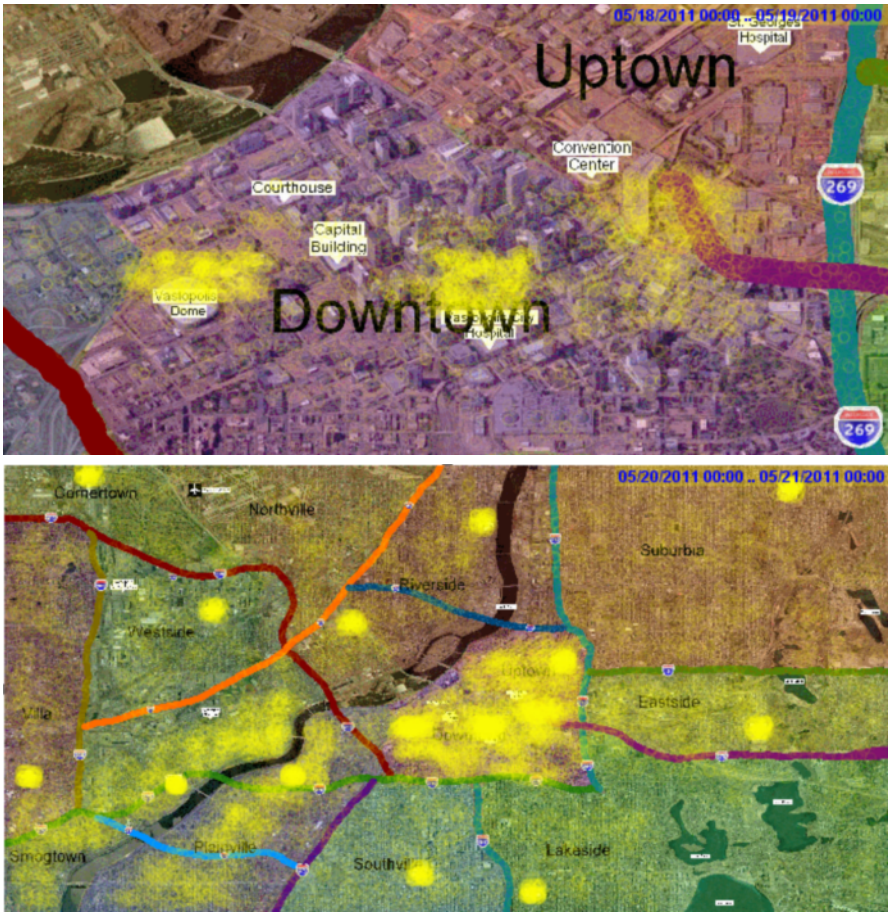

Fig. 1.10: Top: Three dense clusters that emerged in the city centre on the first day of the outbreak. Bottom: The distribution of the disease-related messages on the third day of the outbreak.

see the hospital-centred clusters on the third day. The three very dense clusters that emerged in the city centre on the first day dissolved on the second day. We see a zone of increased message density stretching from the centre to the east on the first day and another zone that formed on the second day on the southwest of the city.

Based on these observations, we conclude that the disease started in the centre and spread to the east on the first day. On the second day, the outbreak hit the southwestern part of the city. In the city centre, the frequency of the disease cases remained quite high during the second and third days. Beyond the observed clusters, the remaining messages were scattered over the whole territory.

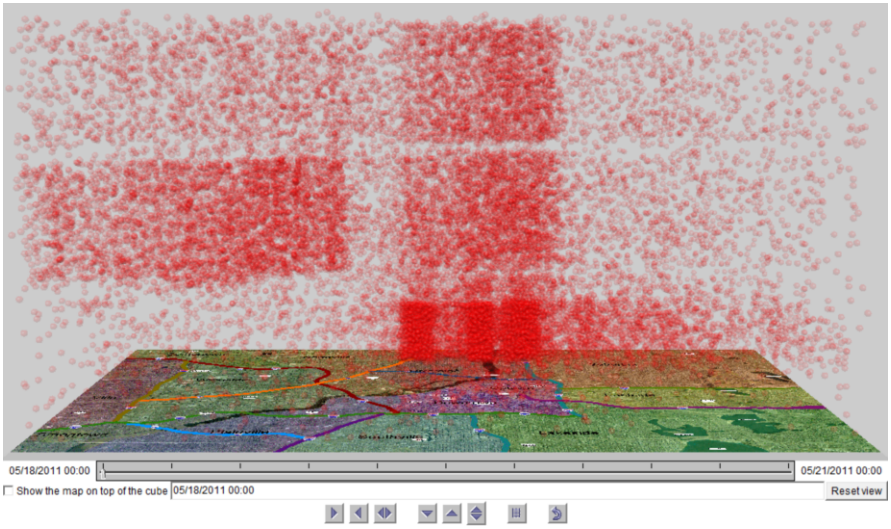

Fig. 1.11: The space-time cube shows the spatio-temporal distribution of the messages where people mention disease symptoms for the first time.

1.2.9 Identifying the mechanisms of the disease transmission

If we consider the first two days of the epidemic, when the disease was spreading (we can disregard the third day when the messages mostly concentrated around the hospitals), we basically see two major areas highly affected by the outbreak: the centre and east of the city and the southwest. The latter area was affected later than the former. We need to understand why it was so. From the weather data provided together with the messages, we learn that on May 18 (i.e., on the first day of the outbreak), there was wind from the west and on the next day from the west and northwest. This could explain the propagation of the disease to the east but not to the southwest. We wonder whether the disease symptoms were the same in the two areas. The illustrations in Fig. 1.12 show that this is not so. We have used an interactive spatial filtering tool for selecting the messages from the central-eastern area and from the southwestern area and looked at the corresponding frequent keywords in the word cloud display. The most frequent keywords in the centre and on the east were chills, fever, caught, headache and other words indicating flu-like symptoms. In the southwestern area, the most frequent symptoms were stomach disorders. Hence, these two areas were hit by different diseases, which, probably, have different transmission mechanisms. It is the most likely that the flu-like disease was transmitted from the centre to the east by the western wind, but this does not apply to the stomach disorders.

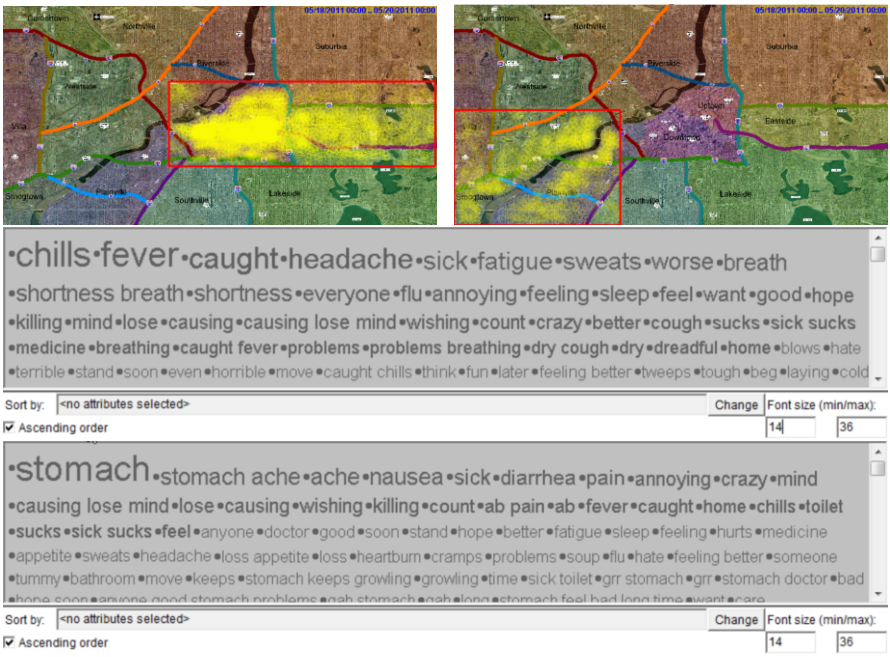

Fig. 1.12: Top: Selecting messages from two outbreak-affected areas by a spatial filter. Middle and bottom: The frequent keywords that occurred in the messages in these two areas.

On the map, we can observe that the dense message clusters on the southeast are aligned along a river; hence, the stomach disease could be transmitted by the water flow in the river.

Could both diseases have a common origin? On the upper map in Fig. 1.7, the central and southwestern clusters seem to emanate from a point where a motorway represented by a thick dark red line crosses the river. It is probable that there was a common reason for both diseases. We come to a hypothesis that some event might have happened on or near the motorway bridge before the 18th of May causing toxic or infectious substances to be discharged in the air and in the river. This probable event might leave traces in the microblog messages. To check this, we apply spatial filtering to select the area around the bridge and temporal filtering to select the day before the outbreak started. The word cloud display (Fig. 1.13) indicates that a truck accident occurred in this place causing a fire and spilling of cargo. Evidently, the fire produced some toxic gas that contaminated the air, and the spilled cargo contained some toxic substance that contaminated the water.

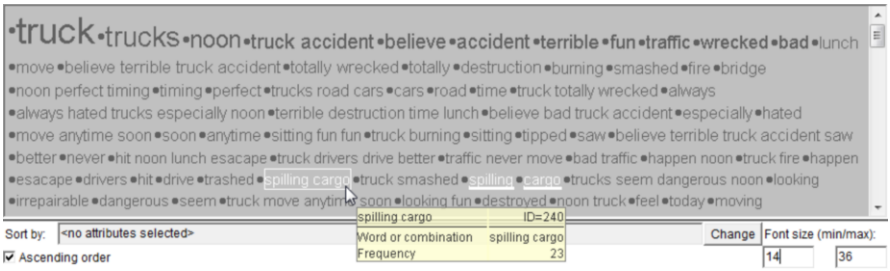

Fig. 1.13: The frequent words and combinations from the messages posted near the motorway bridge on May 17.

1.2.10 Identifying the epidemic development trend

What is the tendency in the outbreak development? Does the disease continue spreading? Are any actions required to stop the spread, or the health professionals mainly need to help people who already got sick? To answer these questions, we first look at the time histogram of the frequencies of the messages that mention disease symptoms for the first time (Fig. 1.14). The histogram bars are divided into segments of two colours, red for the messages mentioning digestive disorders and blue for the remaining messages. We see that the overall frequency of the messages gradually decreases, meaning that the outbreak goes down. This also indicates that the disease, most likely, is not transmitted from person to person; otherwise, we would observe an increasing rather than decreasing trend.

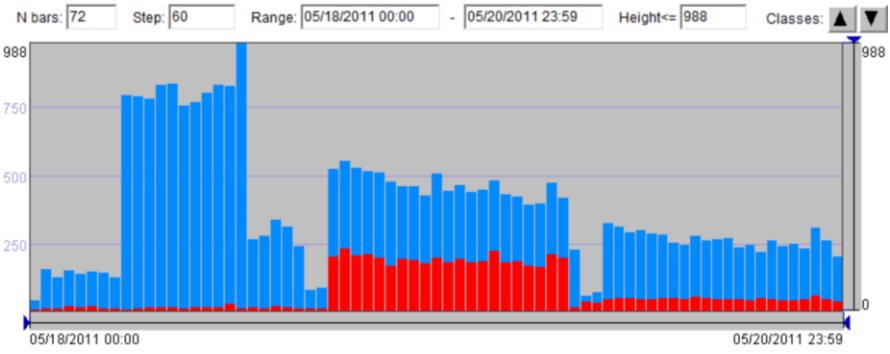

Fig. 1.14: The time histogram of the frequencies of the first mentioning of the disease symptoms. The red bar segments correspond to the messages mentioning digestive disorders and the blue segments to the remaining messages.

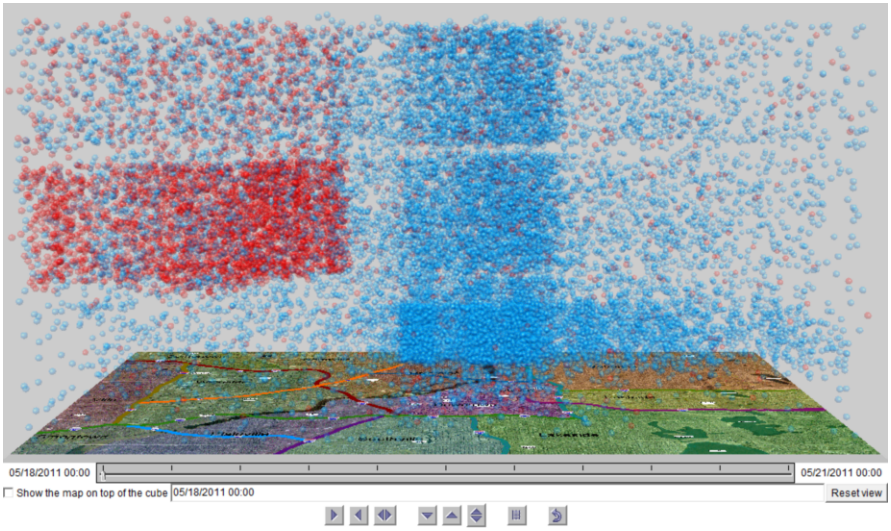

Fig. 1.15: The spatio-temporal distribution of the messages mentioning health problems for the first time. The messages mentioning digestive disorders are represented in red and the remaining messages in blue.

Looking separately at the red and blue segments, we can notice that the frequencies of the new messages mentioning digestive disorders (red) are much lower on the third day than on the second day, whereas the frequencies of the messages mentioning flu-like symptoms are almost as high as on the second day, which means that the morbidity rate does not decrease as fast as it would be desired. We propagate the red-blue colouring also to the space-time cube (Fig. 1.15). We see that new mentions of digestive disorders appear mainly on the southwest, as in the second day, which means that the water remains contaminated. The new mentions of the flu-like symptoms are scattered everywhere but the highest concentration is in the centre. It may mean that some traces of contamination still remain in this area, and it would be good to clean it somehow. It is also reasonable to warn the population about the risks of contact with the water in the river.

1.2.11 Summary: The story reconstructed

We seem to have plausibly reconstructed the story of what happened in Vastopolis. On May 17, around the noon time, a traffic accident on the motorway bridge in the city centre caused a fire on a truck carrying a toxic substance. The western wind moved the smoke containing toxic particles to the central and eastern areas of the city. The cargo from the damaged car spilled into the river and was moved by the

river flow towards the southwest. Many people in the city centre and on the east inhaled toxic particles on the next day after the accident and got ill. The main symptoms were chills, fever, caught, headache, sweats, and shortness of breath, similarly to flu. Unfortunately, since the city centre is the busiest and most crowded area, quite many people were affected. The toxic spill in the river also had sad consequences, which were mainly observed on May 19. It affected people who were on the river banks and, possibly, had direct contacts with the water. Toxic particles somehow got to their stomachs and led to disorders of the digestive system. The morbidity rate of the digestive system disease has notably decreased on May 20. New cases of people feeling flu-like symptoms continued to appear on the second and third day after the accident. The morbidity rate decreases quite slowly, calling for some measures to clean the territory. Fortunately, there is no evidence that the disease can be transmitted through personal contacts; therefore, there is no need to isolate affected people from others and examine everyone who was in contact with them.

1.3 Discussion: How visual analytics has helped us

We have reconstructed the story and answered the questions of the challenge by means of analytic reasoning, which is the principal component of any analysis. To be able to reason, we need to ingest information into our brain. What is the best way to do this? Can we just read all data records? Even if we could read the available 1,023,077 records in a reasonable time, would this help us to understand what was going on? It seems doubtful. Throughout the whole process of analysis, we used visual representations, simply speaking, pictures.

You probably heard the idiom "A picture is worth a thousand words"[1]. In our example, a picture can be worth more than a million records. Instead of reading the records, we could catch useful information in just one look. Pictures were the main sources providing material for our reasoning. This material was put in such a form that could be very efficiently transferred to our brain using the great power of our vision. What our vision does is not just transmitting pixels. Psychological studies show that human vision involves abstraction [26]. Seeing actually means subconsciously constructing patterns and extracting high-level features, and it is these patterns and features that we use as material for our reasoning. Moreover, perceiving patterns and features inevitably triggers our reasoning. Hence, whenever a task cannot be fulfilled by routine computer processing but requires human reasoning, visual representations can effectively convey relevant information to the human mind. Of course, the visual representations need to be carefully and skilfully constructed to be really effective and by no means mislead the viewers by conveying false pat-

[1] https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words

terns. It is one of the goals of this book to explain how to construct such representations.

Although one picture can be worth million records, a single picture may not be sufficient for solving a non-trivial problem. Thus, our analysis consisted of multiple steps:

- data preparation, in which we selected the subset of potentially relevant records;
- analysis of the temporal distribution of the records, in which we identified the start time of the epidemic;
- analysis of the spatial distribution, in which we identified the most affected areas;
- verification of observed patterns, in which we checked whether the high density of the disease-related messages is not just proportional to the usual density of the messages;
- analysis of the spatio-temporal distribution, in which we identified how the outbreak evolved, discovered differences between the temporal patterns in two most affected areas, and came to a hypothesis that they could be affected by different diseases;
- comparison between the texts of the messages posted in the two most affected areas, in which we confirmed our hypothesis of the existence of two different diseases;
- reasoning about the disease transmission mechanisms, in which we related the observed patterns to the context information concerning the weather (the wind) and the geographic features (the river);
- hypothesising about a common source of the two diseases based on the observation of the spatial patterns;
- finding relevant information for explaining the reasons for the epidemic outbreak;
- putting our findings together into a story that gives answers to the questions of the challenge.

In each step of this process, we used visual aids for our thinking. Besides, we used various operations: data selection, spatial and temporal filtering, extraction of frequent terms, derivation of secondary data, such as district-based aggregates, constructing record sequences, and extracting sequence heads. Throughout the process, we continuously interacted with our computer, which performed these operations upon our request and also produced the visual representations that we used for our reasoning. The whole process corresponds to the definition of visual analytics: "*the science of analytical reasoning facilitated by interactive visual interfaces*" [133, p. 4]. An essential idea of visual analytics is to combine the power of human reasoning with the power of computational processing for solving complex problems.

At the same time, our analysis process also corresponds well to what data scientists usually do: select and process data, explore the data to identify patterns, verify the patterns, develop models, and communicate results. These activities cannot be done without human reasoning, and, as we showed and discussed earlier, visual representations can be of great utility. Perhaps, the main difference between visual analytics and data science is their focusing on different aspects of the analytical process, which is performed by a joint effort of the human and the computer. Data science focuses on the computer side, on computational processing and derivation of computer models (here and further throughout the book, we use the term *computer model* to refer to any kind of model that is meant to be executed by computers, typically for the purpose of prediction). Visual analytics focuses on the human side, on reasoning and derivation of knowledge.

Visual analytics considers the **knowledge** generated by the human to be an essential result of the analytical process, irrespective of whether a computer model is built or not. This knowledge can also be seen as a kind of model of the subject that has been analysed. It is a **mental model**[2], that is, a representation of the subject in the mind of the human analyst. In our example, we have constructed such a model in the process of the analysis. It can not only explain what happened but also predict how the situation will develop and tell what actions should be taken. Hence, it may not always be necessary to develop a computer model, yet whenever a computer model is required, visual analytics can greatly help in building it. Moreover, a computer model cannot be appropriately used without human knowledge of what it is supposed to do and when and how to apply it. This knowledge of the computer model is, like the knowledge of the phenomenon that is analysed and modelled, an important outcome of the analytical process. Therefore, the scope of visual analytics includes not only derivation of mental models (i.e., new knowledge represented in the analyst's mind) but also conscious development of computer models that are well considered and well understood.

As you see, visual analytics aptly complements data science, and, moreover, it is instrumental for doing **good data science**, because non-trivial and non-routine analysis tasks require joint efforts of computers and humans. Visual analytics provides a way to perform data science so that the power of human vision and reasoning are effectively utilised.

1.4 General definition of visual analytics

As we already mentioned, visual analytics has been defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [133, p. 4]. This definition emphasises a certain kind of activity (analytical reasoning) and a certain tech-

[2] https://en.wikipedia.org/wiki/Mental_model

nology (interactive visual interfaces) supporting this activity. We have discussed in the previous section why visual representations are essential for human analytical reasoning. However, the definition also contains the keyword "interactive". What does it mean, and why do visual interfaces need to be interactive?

When we consider the use of a picture representing data, it is quite likely that this picture does not include all information that we may need for our reasoning, or some information may not be easily perceivable because our attention is attracted to stronger visual stimuli. In such cases, we need to modify the picture or to obtain something in addition to it. Here are some examples of what we may need:

- get a more detailed view of some part of the picture (*zooming*);
- hide information that we currently do not need for our reasoning (*filtering*);
- see exactly what data records stand behind some elements of the picture (*querying*);
- create additional representations showing different facets of the same data (*multiple views*);
- find corresponding pieces of information in two or more displays (*linking multiple views*).

All these are examples of interactive operations. Usually, to gain knowledge from non-trivial data, it is not enough just to look at a single static (non-interactive) picture, even when it is perfectly designed. Therefore, the definition of visual analytics emphasises the importance of interaction for analytical reasoning.

While the human brain is a powerful instrument for analysis and knowledge building, it has its limitations, mostly regarding the memory capacity and the speed of operation. In these respects, computers are immensely more powerful. Then, why not to combine the strengths of the humans and computers? This is what visual analytics aims at! It develops such approaches to data analysis and knowledge building in which the labour is distributed between humans and computers so that they can effectively and synergistically collaborate utilising their unique capabilities, some of which are listed below.

Humans	**Computers**
flexible and inventive, can deal with new situations and problems	can handle huge amounts of data
can associate diverse information pieces and "see the forest for the trees"	can do fast search
can solve problems that are hard to formalise	can perform fast data processing
can cope with incomplete/inconsistent information	can interlink to extend their capacities
can see and recognise things that are hard to compute or formalise	can render high quality graphics

So, it is sensible to put these great capabilities together and let them work jointly. This requires communication between the human and the computer, and the most convenient way for the human is to do this through an interactive visual interface. The book "Mastering the Information Age : Solving Problems with Visual Analytics" [79] says: "The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data". As a schematic representation of this statement, Figure 1.16 shows how a human and a computer work together to analyse data and generate knowledge. The computer performs various kinds of automated data processing and derives some artefacts, such as transformed data, results of queries and calculations, statistical summaries, patterns, or models. The computer also produces visualisations enabling the human to perceive original data as well as any further data and information derived by means of computational processing. The human uses the information perceived for reasoning and knowledge construction. The human determines and controls what the computer does by selecting data subsets to work on, choosing suitable methods, and setting parameters for processing. Based on the current knowledge, the human may refine what the computer has produced, for example, discard some artefacts as uninteresting and apply further processing to interesting stuff, or partition the input data into subsets to be processed separately.

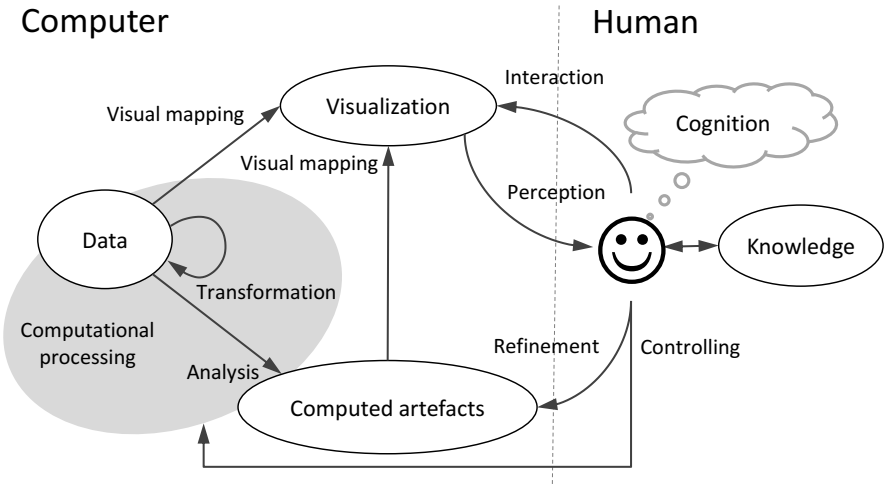

Fig. 1.16: Schematic representation of the visual analytic activity in which human cognition is combined with computational processing. Adapted from [78, 79]

During this activity, the human constantly uses and enriches the knowledge existing in the mind. The human begins the analysis having some prior knowledge and constructs new knowledge as the analysis goes on. The activities of the humans are supported by interactive visual representations created by the computers, and the computers also help the humans by handling data and deriving various kinds

of information by means of algorithmic methods. The visual analytics technology thus includes principles and methods of visual representation of information, techniques for human-computer interaction, and algorithmic methods for computational processing. The following chapter provides an overview of these technical means.