# From Origins to Destinations: The Past, Present and Future of Visualizing Flow Maps

MATTHEW CLAUDEL, TILL NAGEL and CARLO RATTI

*Flow maps are an established cartographic method to depict movements over time and space. In recent years, the exponential increase of geospatial information – what we call* urban 'big data' *– has introduced new uses and highlighted the need to expand cartography. In this paper, we define existing visualization strategies and tools, and examine their characteristics. From this, we identify challenges and opportunities for data-driven flow maps and suggest future developments. Specifically, we apply a new taxonomy to compare several geospatial data visualizations from the MIT Senseable City Lab and extract principles that can define the capabilities of a new interactive flow mapping tool. We have begun to work on such a tool – called the* Datacollider *– that is public, powerful, intuitive, and scalable. In the latter portion of this paper, we describe the* Datacollider, *detail its limitations, and outline directions for future development. We conclude by extrapolating broader trends for the field of geospatial data visualization. We articulate a shift from visualization as a set of graphic tools for representing found insights, to visualization as a way of engaging with data and deriving knowledge.*

Flow maps emerged during the early 1800s as a new form of cartographic representation (Robinson, 1967). These introduced techniques for documenting geospatial data with time-varying properties – the movements and transfers of such things as people, objects, goods and disease – as information overlaid on a standard bird's-eye view projection. Pioneering graphic techniques at the time undertook legibility while negotiating many simultaneous data types. These documents represented a new tool, a *geospatial infographic*. Maps augmented with information layers introduced a rich spectrum of possibilities beyond traditional cartography.

Charles Joseph Minard was an early pioneer of data-centric mapping, best known for his iconic documentation of Napoleon Bonaparte's march to Moscow (a map that plots Napoleon's number of troops, distance marched, temperature, latitude and longitude, direction of travel, and location with respect to time). Minard's body of infographic work also includes documents showing the geographic extents of cattle raised to feed Paris in 1858, worldwide immigration patterns, and global wine exports from France (Robinson, 1967). An even earlier visualization of traffic flows in the Pale of Dublin in 1837 produced by Henry Harness for the British Army prior to the decision to build a railway in Ireland and later work on migration in the UK by Ernst Georg Ravenstein in the late 1880s reveal that flow maps have always been popular ways of picturing and communicating such movements, despite the difficulties of doing so (Batty, 2015).

Minard produced documents – a novel

approach to information-augmented cartography, certainly, but with the structure of a traditional map. In anticipation of the Paris World Exhibition of 1900, Elisée Reclus, a geographer, and Patrick Geddes, an urban planner and sociologist, proposed an entirely new visualization device for geospatial data. Their giant Paris Globe project was intended as an educational tool capable of displaying global information (Meller, 1990). Decades later, during the 1960s, Buckminster Fuller proposed a similar geospatial infographic device, called *Geoscope*. His globe, 200 ft (61 m) in diameter, would be studded with coloured lights, dynamically showing information in real time. If realized, Geoscope would have been able to accept a variety of datasets with geospatial dimensions (such as transportation, economic activity or demographic shifts) and display their planetary flows. 'With the Geoscope humanity would be able to recognize formerly invisible patterns and thereby to forecast and plan in vastly greater magnitude than heretofore' (Fuller, 1981). Both concepts proved too ambitious to be realized during their time.

These conceptual projects anticipated the Internet, and the unprecedented amount of data that is now generated at every moment. The miniaturization of computing and the ubiquity of telecommunications networks have added another layer to physical space, enriching the physical world with a wealth of virtual bits. More and more material movements – whether people, vehicles or goods – can now be tracked and tagged in real time: their aggregate sum is an unprecedented 'big data'. What Charles Joseph Minard spent years painstakingly recording with analogue tools can now be generated on a massive scale to describe planetary flows. In this article, we encounter and explore mapping in the hybrid digital-physical condition.

The contemporary data deluge – in addition to new tools for analysis and representation – demands a reconsideration of practices and applications for geospatial data visualization. To date, a wide variety of new techniques has been developed to contend with datasets, analytical tools and representation platforms (Andrienko and Andrienko, 2012). Given the concentration of data associated with urban agglomerations (owing to population density and a higher rate of technology adoption), as well as greater opportunities for application of results (geospatial infographics can readily inform policy at the metropolitan scale), many of these techniques focus on the urban condition. It is in this domain that a line of our research at the MIT Senseable City Lab has focused over the past 10 years. Simply as a reference, we evaluate a sequence of visualization projects from this time frame:

1. *Real Time Rome* was among the first explorations of aggregated data from cell phones (through Telecom Italia's LocHNESs platform), buses, and taxis in Rome (Calabrese *et al*., 2011). Researchers sought a better understanding of urban dynamics, specifically surrounding major events. Their analysis revealed the pulse of the city, demonstrating that collective emotion, action and interaction are written in geospatial data.

2. *Live Singapore* brought together diverse datasets, including weather, energy use, social media, public and private transit and telecommunications, to enrich a portrait of the city (Kloeckl *et al*., 2012). The resulting visualizations have reached a broad audience not only through scientific papers but also exhibitions in the city-state.

3. *A Tale of Many Cities* investigated one category of data – telecommunications – with an in-depth analysis and comparison of urban areas around the globe, finding important patterns and characteristics through juxtaposition (Grauwin *et al*., 2014). Crucially, it is visualized as a public web app for browsing data.

These projects represent a development from the geospatial infographic to interactive geovisualizations. This is a move from

animated maps to dynamic representations of time-based data, in which a user can interact by filtering content according to place, time, dimension, characteristic, and more. The examples progress from narrative representation of relationships or changes, to enabling non-linear user exploration and engagement (a detailed review follows in 'State of the Art' below).

While interactive geovisualizations are a powerful tool, they are nonetheless trammelled by significant constraints. State-of-the-art geovisualization remains limited in its technical capacity and in its accessibility. More specifically, visualizations are made by highly trained experts, using prepared datasets, over a long period of time. Particularly in the urban context, a broad range of stakeholders stand to gain from geospatial infographics, not only technically capable geospatial scientists. Commuters, entrepreneurs, academics, government officials and business leaders create data every day, and could potentially benefit from using geospatial data tools for decision-making and communication.

In this paper, we begin with an overview of the state-of-the-art in the field of flow mapping, including data structure, software, capacities and limitations, to establish clearly the gaps in contemporary approaches. Next we present several examples of geospatial data visualization from the Senseable City Lab – focusing on the development of cartographic practice over time; we summarize the main contributions of selected projects, and discuss these in the context of the aforementioned usage scenarios. We synthesize our experience with the limitations we have identified in contemporary tools, in order to inform a path forward.

Together, these factors suggest a clear vector: to put the tools of data analysis and visualization into public hands, with a system that is powerful, intuitive, and scalable, at little to no cost – spanning the edge of MacEachren's conceptual diagram of the 'cartography cube' which we show in figure 1 (MacEachren, 1994). Importantly, this will not sacrifice the extreme conditions, or 'corners' of the cube (specialists who require advanced tools and laypersons who require simple, intuitive functionality). We articulate our first attempt at such a tool, called the *Datacollider*. We detail the *Datacollider*'s structure and function, its current limitations and our plans for its future development. In conclusion, we discuss the impact and subsequent research directions for the broader field of data-driven geospatial flow mapping. We also include an Appendix which summarizes the key terminology related to visualization in this domain.

## Context and Users

By definition, big data – in its raw form – cannot be understood without analysis, synthesis and context (Snijders *et al*., 2012). The term 'big data' was first used in a 1997 paper (Cox and Ellsworth, 1997) to indicate the limit at which NASA scientists faced computational challenges in the process of visualization. The authors assert that 'in the area of scientific visualization, input datasets are often very large', and propose a segmented memory system whose 'techniques effectively reduce the total memory required by visualization at run-time'. In the years since this initial characterization, 'big data' has continued to be vexed by a number of limitations in the process of visualization. As datasets have become larger, computational demands of processing for visualization have only become greater.

Urban agglomerations generate a tremendous amount of data and, furthermore, a diverse set of stakeholders in the metropolitan context demand synthesis and analysis to communicate insights. Particularly in the realm of so-called 'smart cities', analysis and visualization are well understood as a necessary corollary to sensors, urban operations and urban big data itself (Townsend, 2014). At stake is the challenge of turning *raw data* into *information*. A cadre of specialized

geographers, geographic information scientists, and data visualization experts has emerged to produce complex geospatial data visualizations, yet the demand for simple and direct ways of creating such documents outstrips the capacity of the specialist coterie.

With the establishment of digital cartography and interactive geovisualization, a number of taxonometric schemes (e.g. Crampton, 2002; Roth, 2013) have been proposed to systematically classify tools and visualization techniques. The cube schema by MacEachren (MacEachren, 1994) describes a number of interrelated spectra: tasks (presenting knowns; revealing unknowns), interaction levels (passive; active), and audiences (public/laypeople; specialist/experts) (see figure 1). An additional axis through the middle of the cube spans from Analysis to Synthesis to Presentation. MacEachran and Kraak (1997) further elaborate the scope of Presentation, specifying that the viewing audience may gain new insights that had only been 'largely known to the information designer'. Furthermore, presentation may be interactive while remaining 'mostly' on the Presenting Knowns extreme of the Task axis.

This is a useful classification tool, but in the past decade the meanings of 'tasks', 'interactions' and 'users' have changed with respect to a new and expanded digitally-literate audience. With the broad adoption of location-aware mobile devices, and the everyday use of personal navigation applications, a large segment of the population is geospatial data-savvy, and more accustomed to interactive maps. As a result, there is a need for new, simplified data management and visualization tools (Andrienko *et al*., 2010). While existing tools typically fit in one of the cube's corners, opposed across the diagonal axis, we extend the concept of cartography. Rather than creating targeted tools that are specialized for extremes of the cube's axes, we propose an alternative goal: to deliver software that spans the entire edge, from layperson to specialist, in a way that is highly interactive and reveals unknowns. Existing solutions are based either on advanced software development skills (for example, programming custom software), or on phrasing complex queries in a coding language. Furthermore, as detailed below, a vast amount of urban data has inherent temporal properties that are difficult or impossible to visualize with current tools.

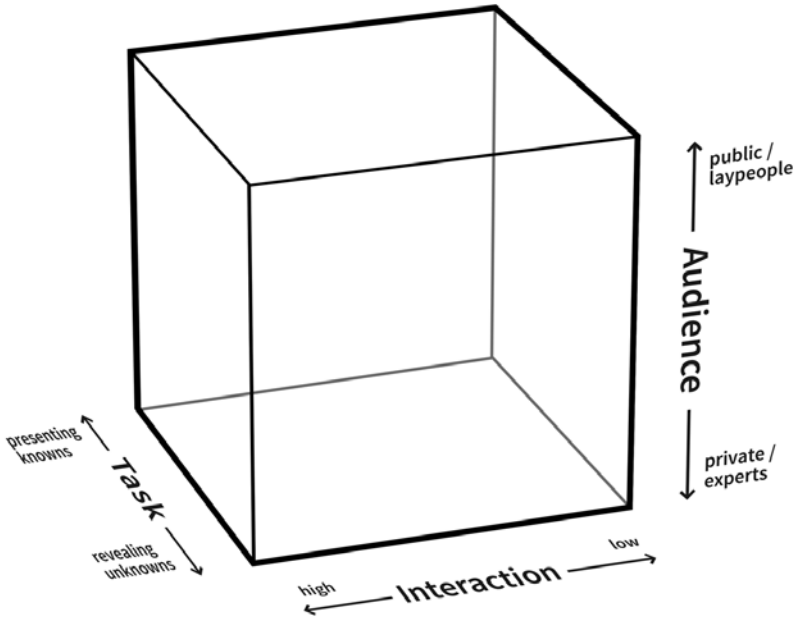


Figure 1. Cartography cube (based on MacEachren 1994).

## The State-of-the-Art

### *Existing Tools*

Geographic Information Systems (GIS) software is the industry standard for the majority of professional geospatial data processing. Of existing tools, the two most powerful are the proprietary *ArcGIS* software package by ESRI (Environmental Systems Research Institute), and the open source software *QGIS* (Quantum GIS). These allow users to import, analyze, process and display geospatial data. The software accepts input data (both tabular and cartographic) and can commit a variety of computational processes on that data, both raster-based (*map algebra*) and vector-based (*geoprocessing*). This geospatial computation is defined by calculations and user input within the software, driven predominantly by internal processing and to a lesser extent by connection to existing programming languages (e.g. Python for *ArcGIS*).

Independent of GIS software, there are a variety of programming languages (including R, Python and Matlab) and libraries that allow a wide range of operations, from specialized geospatial analysis to data processing. Using these tools, entirely new visualizations can be created, yielding results more varied than GIS software – but requiring significant technical expertise. To address this, some scripting environments – for example, D3 and Processing – simplify analysis and visualization, but both require coding skills and do not offer data management functionalities.

There exist software tools for geovisualization that are capable of importing and displaying data in a variety of ways. This third group differs from GIS applications in that they are generally designed to be intuitive for a non-specialized audience, and from programming languages in that they are user-facing software. These range from general data tools (Tableau or Excel), to geovisualization applications (CartoDB or Mapbox). CartoDB simplifies map-making through an intuitive web application, but does not provide special support for large or time-stamped datasets. Software such as Tableau, Qlikview or Omniscope have made a step forward in usability by introducing a graphical interface to manage and process big databases, but they still rely on traditional visualization types.

### *A Taxonomy Matrix*

We propose a basic taxonomy of functionality for geospatial data visualization tools, based on the MacEachren's cube diagram. This provides a rubric for classifying existing tools and identifying the advantages and drawbacks of different approaches. We first describe the input (data), the development (processing procedure), and the output (visualization) of each. Subsequent discussion is grounded in this taxonomic model.

*Input (data)*: Data can come from many different sources. Some online tools allow users to edit or copy+paste data into a simple text field. Essentially all tools allow files to be uploaded, with some providing input from cloud services such as Google Drive or Dropbox. Each of these tools supports many and varied file formats, ranging from structured text files such as CSV, XML, and JSON to standardized formats for geospatial data such as KML, GeoJSON and Shapefiles. Furthermore, tools can be connected to external sources and accept data directly. Using a database as an external source typically requires standard spatial queries, while using APIs (such as Twitter) involves more complex programming mechanisms.

*Development*: Visualization tools can be classified according to the way a user brings raw data to visualization: textual programming, visual programming, or a tailored user interface. This strongly defines the audience that can interact with a given tool, according to technical expertise. Tools amenable to non-experts often guide users in choosing the appropriate representation strategy.

*Output (visualization)*: While datasets can be depicted in a variety of ways, some require specific representations. A variety of visualization types span this breadth of output. Users can engage in a range of interactions, from observation of a single document (Static), to watching a visualization change in time (Dynamic), to manipulating different aspects of the representation (Interactive):

(*a*) Static graphical representation of geospatial data: Export: (JPEG, PNG, TIFF).

(*b*) Dynamic representations take the form of video, showing animated geospatial data. A subset of dynamic visualizations is customizable representations: these are not truly interactive but grant users a measure of control, usually over graphic characteristics: styles, labels, colours, or base-map. Dynamic, non-interactive visualizations require the user to structure the output as a sequential narrative, to communicate with the end viewer. Export: (MP4, AVI, GIF).

(*c*) Interactive representations of geospatial data are structured to dynamically query the dataset based on viewer input. Common cartographic interaction types include pan, zoom, select geographical objects, get details on demand. Some allow more advanced interactions such as automatically zooming and panning so that selected objects fit in the view frame. Geospatial data commonly has temporal properties, placing importance on time-based interactions: such as play, stop, zoom (in time) and identifying temporal patterns (rhythms). There are few standardized platforms for presenting interactive visualizations, and generally must be custom built. Export: (web-app in a browser or a smartphone app).

### *An Overview*

Overall, visualizations of large datasets with temporal and spatial properties are useful in a variety of domains, yet the tools that can be employed to generate such visualizations are either too complex for a broad user base other than technical specialists, or lack appropriate functionalities. Through this detailed examination of software and tools for the creation of interactive geovisualizations, it is clear that the existing offering only partly fulfils the variety of requirements and use cases arising today. We will continue this with a discussion of projects by the Senseable City Lab as case studies, identifying their advantages and drawbacks to arrive at a better understanding of how to structure a much more comprehensive interactive geospatial analyses and visualization tools.

## Geospatial Data Visualization at the MIT Senseable City Lab

Since the Lab's founding in 2005, its multidisciplinary group of researchers has designed a variety of projects using geospatial data. Working with vastly heterogeneous datasets, led by a broad spectrum of goals, the lab has produced many different kinds of outputs. From this portfolio, we detail five examples that showcase different strategies and encompass a chronological development. Drawing on our taxonomic framework, we describe each project in terms of interaction, visualization, task and audience, which categorize the main features and objectives of Senseable City Lab projects.

### *Geospatial Data Visualization*

*Real Time Rome* (2006) aimed to imagine and present the implications of ubiquitous connectivity in the urban environment. With temporal data from many different sources, what was, at the time, a new form of cartography synchronized databases to create a portrait of urban behaviour patterns. Data were aggregated on the Localizing and Handling Network Event Systems (LocHNESs) platform, a system developed by Telecom Italia.

The project was presented at the 2006 Venice Biennale of Architecture in the form

of seven animated visualizations that showed urban activity in Rome, including traffic congestion, human mobility, and the exact movements and locations of buses and taxis (using GPS). Analysis revealed consistent patterns of activity, as well as anomalies. Mega-events such as the football World Cup were expressed in the data, as the city's population flooded through different districts to celebrate after the match. These maps show aggregated movements of mobile phone users (in figure 2) and public transportation (in figure 3). The raw data was analyzed and visualized by experts, resulting in carefully chosen animations to present a story and answer specific pre-selected questions.

The first animation in figure 2 shows gatherings of people for special events. It visualizes mobile phone users as a 3D heat map on top of a satellite map with the height and density coding the amount of cellphone activity. In this way, areas with dense activity are revealed. The visualization is non-interactive, and animates through one day, in this case, through the day of the Madonna concert in 2006 in Rome.

The second animation in figure 3 combines two datasets, showing mobile phone users and public transit vehicles. The density of mobile phone users is visualized as a heat map overlaid onto a satellite map, while buses are shown with their current position as white dots and trails. This visualization is non-interactive, and animates through a representative sample day. *Real Time Rome* suggested a future in which the pioneering

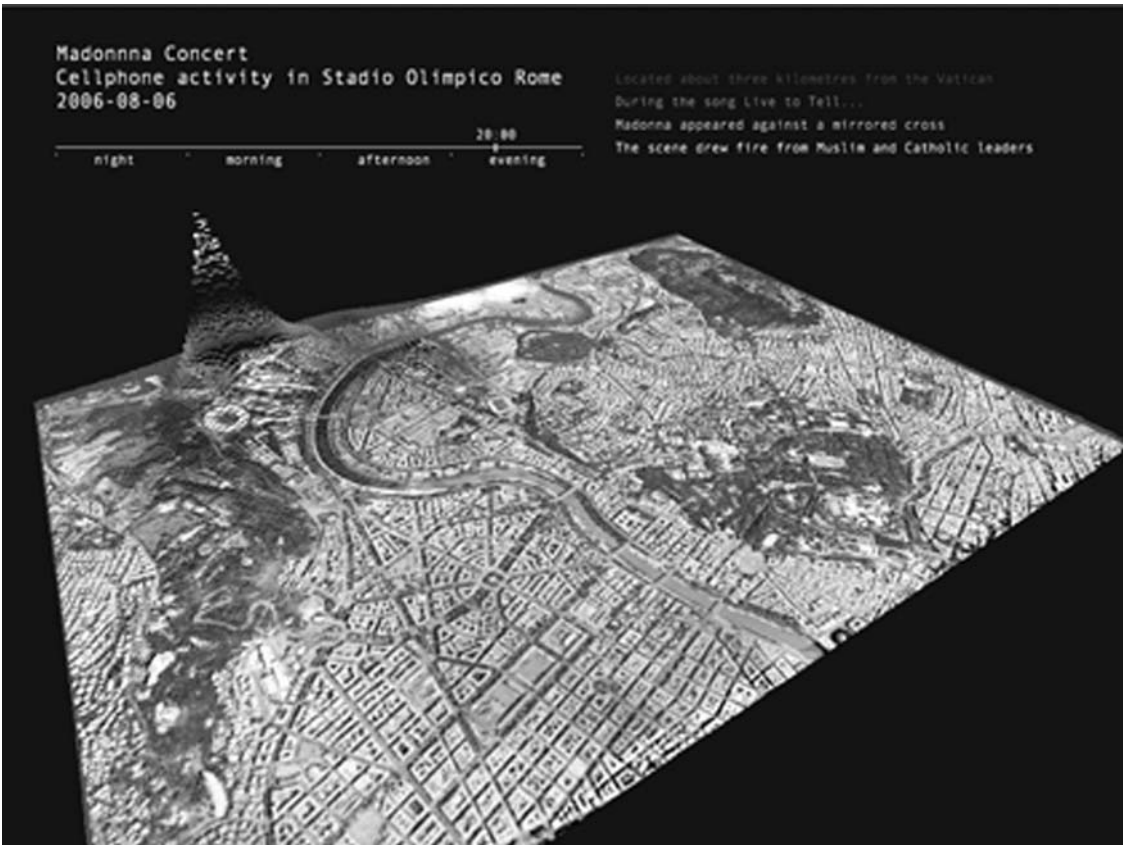


Figure 2. *Real Time Rome*: Gatherings. *Interaction*: None (Animation). *Visualization*: 3D heat map showing mobile phone users.


Figure 3. *Real Time Rome*: Connectivity. *Interaction*: None (Animation). *Visualization*: Flow map showing bus routes; Heat map showing mobile phone users. *Task:* Presenting knowns: is public transportation where the people are? *Audience*: Public.

methodologies of urban science and the tools of big data analytics allow officials to target the inefficiencies of urban systems and open the way to a more sustainable future.

### *Linking and Comparing Datasets*

*Live Singapore* is a multi-year research initiative that aims to give different stakeholders in Singapore real-time access to information about their city. The goals of this initiative are manifold, ranging from acquiring and analyzing data, to creating urban data visualizations as prototypes, to exhibiting urban demos to communicate to stakeholders the value of visualization and to realize that value. In a later stage, a major objective is to marry these approaches in order to create a unified data platform where policy-makers and citizens alike can easily create interactive apps visualizing diverse datasets (see 'Synthesis: Powerful and Public Visualization' below). This aims at empowering citizens and officials to explore and understand their own data, ultimately to gain insights into Singapore's complex urban system.

*Raining Taxis* shows how weather affects taxi usage on the island. This visualization acts as an example of linking datasets in new ways in order to investigate a local phenomenon. Users can rotate the map to orient the 2.5D view and move over the histogram to set the time of day for the animation. The histogram

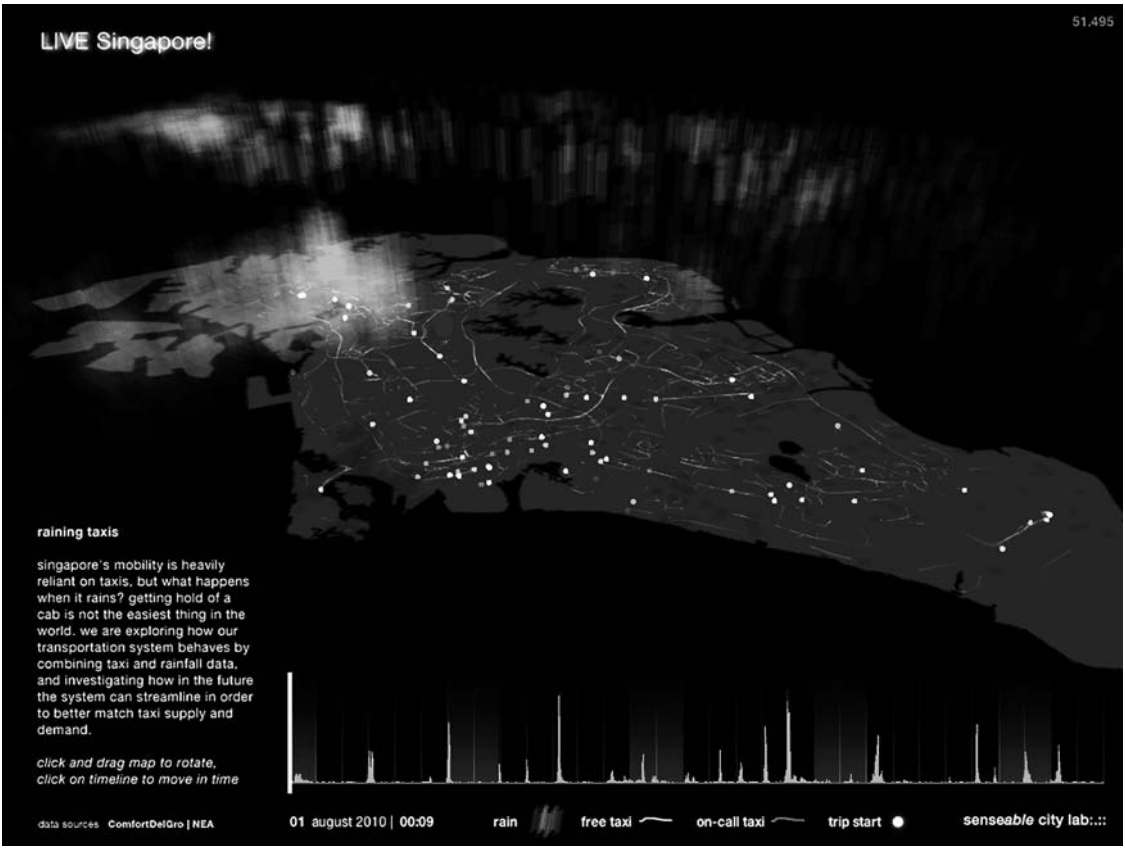


Figure 4. *Live Singapore*: Raining Taxis. *Interaction:* Rotate the map; move in time. *Visualization:* Histogram showing amount of rain over time; dot density showing trip starts; flow map showing free and busy taxi routes; grid map showing rain density. *Task:* Presenting knowns: More rain results in less free taxis.

shows the amount of rain over time as a bar chart. On the visually reduced base map of Singapore, white dots show the start locations of taxi trips, thus giving an impression of taxi use distribution. A flow map depicts the origin, path and destination of each taxi, with bolder lines showing free taxis, and finer lines showing on-call taxi trips. Over the map, a second layer shows rain density in light blue as a gridded map of vertical bars.

*Touching Transport* is a multitouch visualization of the public transit network in Singapore (Nagel *et al*., 2014). It supports the exploration and understanding of complex tempo-spatial data for experts and non-experts. The system provides multiple perspectives of the data and consists of three interactive visualization modes conveying chrono-spatial patterns as map, arc view, and time-series, as shown in figure 5.

In the larger process of working towards an informed discourse on smart cities, we believe that reaching out to both experts and citizens with a tool that facilitates exploring data, gaining knowledge of urban activity and soliciting feedback is an important next step. One way of achieving that goal is to exhibit publicly visualizations of urban data in a demonstration, with an aim of providing visual and interactive access to large urban datasets, engaging public and industry part-


Figure 5. *Touching Transport. Interaction*: Zoom and pan the map; move in time; select time range (aggregate data); switch views (different perspectives); select bus line + direction (filter data); details on demand (filter data). *Visualization*: Glyphs showing boarding and alighting pax per station; glyphs on a map showing spatio-temporal relations (cluster, areas); time-series with glyphs showing spatio-temporal relations (at a glance); arc diagram showing flow of pax (OD, clusters); histogram showing pax over time. *Task*: Presenting knowns: revealing unknowns: personally relevant insights. *Audience*: Public: laypeople and experts.

ners, and collecting feedback and ideas from users. Within *Live Singapore*, a variety of urban demos have been created.

*Evaluation and Comparison Across Cities*

An analysis carried out during 2014 in partnership with the Ericsson Company (a multinational digital services provider) looked specifically at telecommunications patterns in cities around the world. The findings were presented as an interactive web application, titled *Many Cities*, and first presented at the New Cities Foundation Summit on 18 June 2014. This tool maps mobile phone usage data in cities around the world (London, New York, Hong Kong, Los Angeles), allowing the user to browse data (calls, SMS, data traffic) with high spatial granularity (districts, neighbourhoods) over time (days, weeks) as shown in figure 6.

The project was conceptualized with two main aims, specific to two intended user groups.

1. Providing the researcher with an easy-to-use tool for quick, intuitive visualizations of spatio-temporal data (in this case, mobile phone usage data, a source that is increasingly used by researchers for many purposes, from analysis of human mobility to demographic and economic insights). Cities on different continents have characteristic and distinguishable activity patterns due to local


Figure 6. The *Many Cities* web application. *Interaction*: Select city (filter data); switch views (different perspectives); select city districts (filter and compare data); details on demand (filter data); select time range (aggregate data); share views and save comments (communicate). *Visualization*: Choropleth map; cluster map showing spatial clusters of similar activities; time-series of mobile phone activities.

economic-cultural conditions (e.g. United States teenagers sending more SMS in the evening than elsewhere in the world), and the financial centres of cities share a common pattern that is strikingly similar in all cities. As such, mobile phone data suggests that globalization (in the sense of uniformization) exists mainly in the business and finance sector.

2. Providing the citizen with a user-friendly tool that provides a better understanding of individual and collective urban dynamics. For example, specific events (concerts, storms, exhibitions, holidays) can be directly identified by deviations from the average mobile phone usage pattern.

*Many Cities* is a visualization tool for analyzing data. Yet the data has been managed or pre-analyzed, and represented in a customized web-based interface. This allows users to browse (not manipulate) data, such that they can find and explore their own insights. For example, citizens may compare the typical patterns in the area they live with those of other areas in the same city or in cities around the world. Most uniquely for a browser of this type is the capacity to export and share any given session. Users can save, share, annotate and compare individual 'explorations'. A rich and expanding set of interpretations reveals how users – from researchers to laypeople – make sense of complex datasets.

## Synthesis: Powerful and Public Visualization

As a response to the barriers of technical specialization, comprehensive utility and limited accessibility identified in our review and taxonomy of geospatial software tools, we propose a powerful and publicly access-

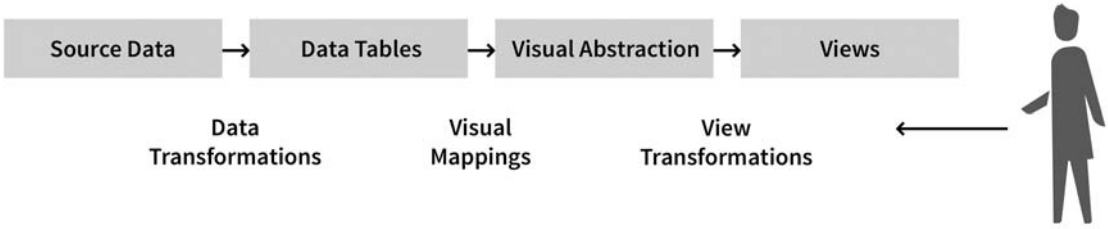


Figure 7. The visualization pipeline.

ible system. Such a software tool would effectively span the edge of the 'cartography cube' without sacrificing functionality at the corners.

With this agenda, the Senseable City Lab is currently developing a new tool called the *Datacollider* – a web application that provides a library of visualization types (and the capacity for specialists to create custom visualization types) along with an interface that simplifies the process of uploading, structuring and processing heterogeneous datasets for use within those visualization frameworks (Senn *et al*., 2015). In this section, we begin with a summary of key elements of the *Datacollider*'s front-end interface and its back-end architecture. We then articulate the path from the raw data to visualization as four sequential steps. This demonstrates the tool's utility, specifically for processing time-and-location-dependent data, and ultimately producing interactive geospatial infographics.

In the classic visualization pipeline (Card *et al*., 1999) shown in figure 7, the raw data is parsed and transformed, mapped, and displayed in some visual form. The user is able to interact with the visualization by adapting each stage of the pipeline. This, however, was intended for expert analysts. As we have explained above, there is an increasing demand from laypeople for urban visualization tools of varying complexity. In figure 8, we identify interaction mechanisms when matched to the capabilities of Senseable City Lab projects. The *Datacollider* aims to provide functionality for the full spectrum of visualization adaptation.

### *The Datacollider*

The user interface is designed as a seamless tool for querying and visualizing geospatial datasets, while the backend framework is robust enough to manage large amounts

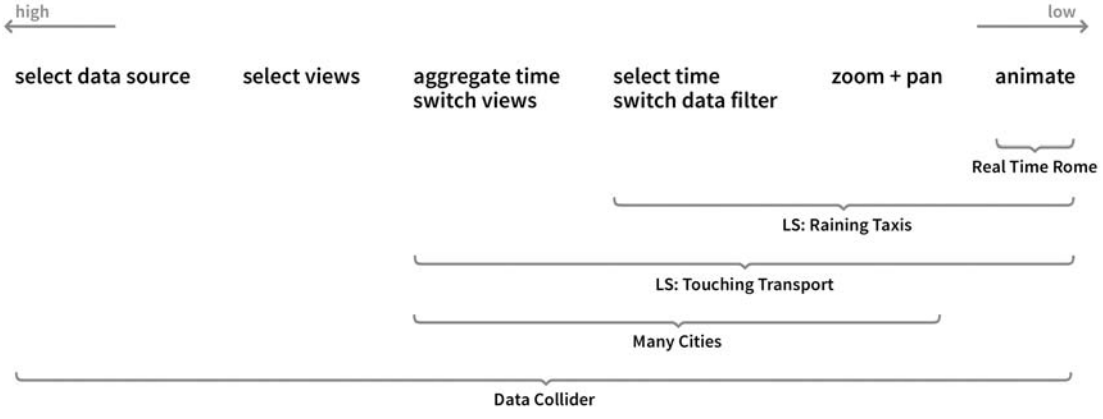


Figure 8. Selected interaction mechanisms for the Senseable City Lab projects.

of data. The goal of the *Datacollider* is to provide high-level functionality (in the form of processing power for large and heterogeneous data with a temporal dimension) without sacrificing intuitive workflow and ease of use – as previously stated, to span the edge of the 'cartography cube'. The *Datacollider* dramatically reduces time required to glean and represent insights from large datasets. Furthermore, a large (and growing) library of visualization types contains many different representation strategies, amenable to a broad variety of datasets, and seeks to convey information in a way that is interactive and intuitive for non-experts. For all of these reasons, both the process and the product of the tool span a broad user and audience demographic. However, the *Datacollider* lacks important functionalities that will be the subject of its continued development. These may also inform the broader field of geospatial data visualization.

*The Work Flow*

We will outline the four stage workflow which constitutes the foundation of the *Datacollider*:

(*a*) Upload *data*: Users begin by logging in and uploading data to their repository. The *Datacollider* accepts two data formats: GeoJSON and delimiter-separated files with spatial identifiers. Having uploaded a desired dataset (or using a dataset previously uploaded), the user then creates a new project.

(*b*) *Select Time Frame and Resolution*: The user must select a time window (when the visualization will begin and end) and the level of resolution (5m, 10m, 15m, 30m, 1h, 4h, 6h, 12h, 24h, 7d, 30d, 90d, 180d and 365d). This is done on a simple timeline with extent-pointers that can be dragged to demarcate the desired time frame.

(*c*) *Analyze Data*: The user can then manipulate the data through a range of SQL-like operators, for example: group, filter, join, etc. There are specific operators to perform transformations, for example aggregation into grids/polygons. Multiple operations can be chained together to form a workflow. This is achieved through a graphical user interface and dragging/dropping items. Furthermore, the user is simultaneously presented with a preview of the data to show implications of adding a given operator to the dataset.

(*d*) *Visualize*: After processing the data, the user can render a visualization using any of the templates provided (see map and chart types, below). Each template presents a list of requirements that can be mapped to data fields from a selection panel. Specific attributes of objects are fully customizable, for example, colours, sizes, heights, etc. after they have been linked to a specific data dimension. Workflow integration allows the user to review the data and operators as necessary, and, if changes are desired, to return to the data operations window for additional or alternate processing. The user can change the imagery of the maps, add additional GeoJSON layers, add contextual elements, or animate the visualization as a narrative.

The current offering of map and chart types:

- Map – Scaled bars
- Map – Scaled circles
- Map – Heat map with bars
- Map – Heat map with spheres
- Map – Origin destination flow
- Chart – Scatter plot
- Chart – Bar chart
- Chart – Pie chart

*Datacollider Limitations*

The *Datacollider* aims to overcome many of the identified limitations of contemporary GIS software, and it constitutes a first step towards an openly available, intuitive and powerful real-time geospatial visualization

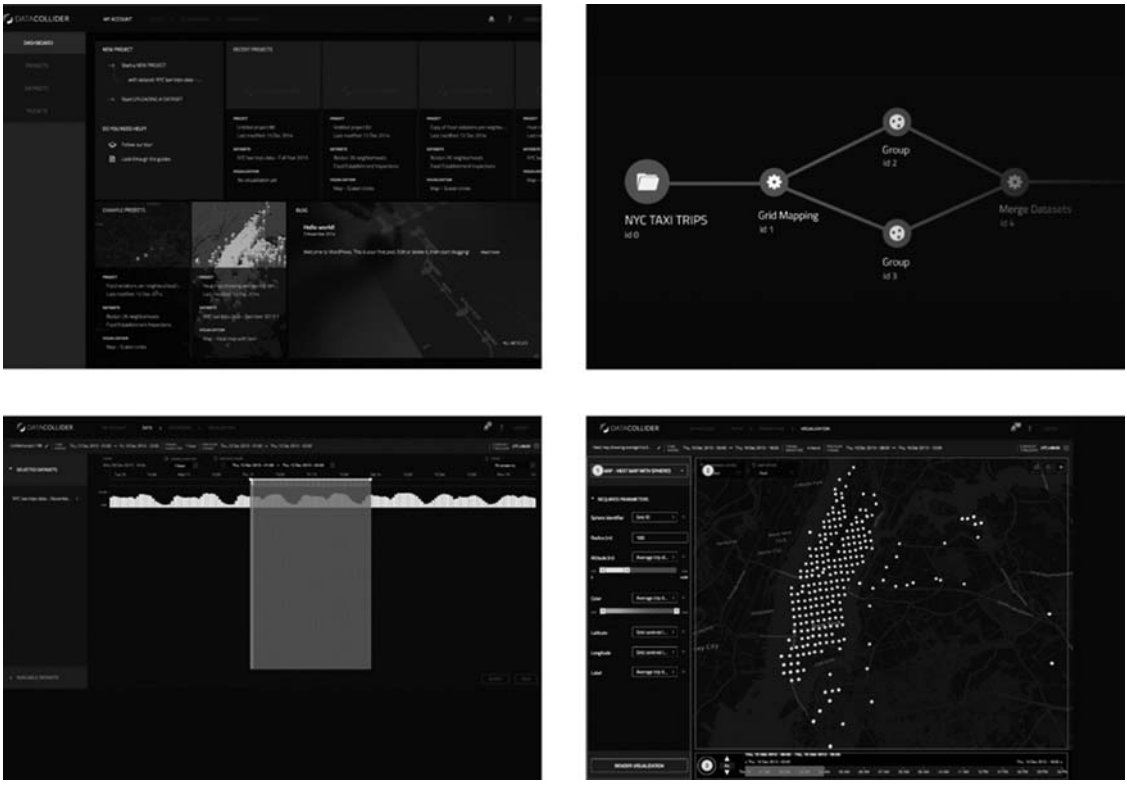


Figure 9. Screenshots illustrating the *Datacollider* work flow. *Top left*: Upload Data; *top right*: Select time frame and resolution; *bottom left*: Analyze data; *bottom right*: Visualize data.

tool. However, the software has limitations that should be acknowledged, and which may inform future research:

*Real-time Functionality*: The *Datacollider* does not support real-time input. Data must be uploaded as a cohesive and singular dataset for processing with the *Datacollider*'s structuring tool.

*Data Structure*: Input data can be saved natively in a variety of formats, but must be: (*a*) delimited file; (*b*) every record associated with a time stamp; and (*c*) every record associated with a latitude/longitude (for use with cartographic visualizations).

*Data Size*: Maximum size is limited (although it is variable, dependent upon the number of operations in a workflow – there are no universal bounds).

*Data Hosting*: When uploaded, data is hosted on servers in Singapore, which may raise political concerns for some users.

*Visualization Export*: The *Datacollider* currently exports high-resolution (2560 × 1920px approximately 2.5MB) images in JPG and PNG format. Animated visualizations can be captured using screencast software such as Quicktime. However, future developments will simplify and improve the export process.

## Future Work

Building on our evaluation of contemporary GIS software and our experience with the *Datacollider*, there are several clear avenues for future development – both of the tool

itself and to progress the general field of geospatial data visualization software.

In terms of technical capabilities, our priority for the *Datacollider* is to enable real-time function, such that data streams can be plugged into the tool, rather than uploaded as a discrete dataset. This will allow stakeholders to understand the real-time operation of systems, rapidly gain insight and take action. Simultaneously, we propose to continue expanding and improving the library of visualization types. This will be significantly advanced as the tool is released to the broader public, and developers are invited to create and upload custom visualization types using open documentation of the application programming interface or API. Finally, we intend to increase the scale and scope of the tool. Larger servers, more file upload types, and locations outside Singapore will allow the *Datacollider* to process more data types, more data, more quickly, and in a politically agnostic way.

Beyond the *Datacollider* itself, we wish to acknowledge policy standards as an important aspect of continued development in the field of geospatial data visualization. This is particularly relevant in the urban context, where activity happens primarily at the intersection of government, industry and academia. Each of these players currently works with proprietary standards and structures, limiting or precluding any transfer of data, workflows or general best practice. Developing schema and protocols for real-time data will allow every potential actor to produce a wider variety of datasets that can be more easily aggregated or streamed, manipulated, and visualized. Examples include: comparison of ocean patterns between coastal cities to better understand sea level rise and inform urban resilience strategies; analysis of energy use patterns across many scales (individual homes, cities, regions) and across private energy providers, to inform and manage transition to a renewable energy grid; and aggregating real-time transportation data from many modes to drive a city-scale 'ambient mobility' platform, linking and optimizing urban mobility options.

More broadly, we imagine that an important development will simply be the wider application of geospatial data visualization tools. This will include existing datasets that have not been visualized and entirely new datasets (created with or without the intention of visualization). We also suggest the representation and use of geospatial data visualizations in new contexts, for example, decision-making in business strategy – these will only increase in the years to come.

## Conclusion

In this paper, we have outlined the state-of-the-art in geospatial data visualization, described a new synthetic approach to visualization called the *Datacollider*, and suggested avenues for future development of the field. These are valuable for the improvement of software and for the general advancement of knowledge, but a widely accessible tool for real-time flow mapping also has the potential to radically transform urban operations by engaging every segment of the population. Urban complexities – formerly the domain of specialists or limited stakeholders – can generate concrete insights that anyone can grasp intuitively. This is crucial for involving communities of citizens who constitute the city itself, and whose behaviours define its overall function. A real-time omni-modal transportation analytics platform, for example, could benefit individuals by elucidating the fastest, most comfortable, or least expensive options, while simultaneously rebalancing the overall system to reduce congestion. This generally represents a shift from visualization as a set of graphic tools for representing *previously found* insights, to visualization as a way of engaging with data and *actively deriving* insights: from post-rationalization to interrogation.

Visual representations may help urban inhabitants to become more aware of issues and patterns in the city around them by

providing a rich array of communication tools (Vande Moere and Hill, 2012). Visualization platforms can transfer information from the top down (government to citizens; businesses to subscribers) or horizontally among citizens (peer to peer). As they become increasingly intuitive they will reach a larger number of people across a broader demographic and be more desirable for everyday use. The goal for geospatial data visualization tools is to add value and enrich life across a significant segment of the population. At this point, people will be empowered to use such tools creatively and in a personally meaningful way. We suggest that there is an urgent need for a simple, intuitive and powerful geospatial and temporal data exploration and visualization tool.

Through our evaluation of the state-of-the-art in geospatial visualization tools, as well as a review of our own work at the Senseable City Lab, we have identified important best practices, limitations, and trends in geospatial data visualization. Considering the contemporary influx of data, increased availability of datasets, and wider use of data-driven cartography in a broad range of decision-making applications, we contend that there is a need for simple, unrestricted and powerful geospatial data visualization tools that span the edge of MacEachren's (1994) 'cartography cube'. We demonstrate that our *Datacollider* is a first step in this direction.

Flexible data analysis and visualization tools, amenable to many and varied real-time data streams (which, themselves, are governed by overarching standards) may become an integral aspect of urban life. Such tools span the edge of the conceptual 'cartography cube', and foreground the *process* of interacting with data. A visualization may be the end product, but new insights will be generated through providing broad public access to an interrogative tool for interacting with data. Geospatial data science becomes an active endeavour rather than a specialist's post-rationalization. We submit that this should be the goal for data-driven flow mapping, as the contemporary urban condition generates enormous real-time datasets. Ultimately, a wide community of interested participants can benefit from such exploration, synthesis and visualization.

## REFERENCES


Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I. and Tominski, C. (2010) Space, time and visual analytics. *International Journal of Geographical Information Science*, **24**(10), pp. 1577–1600.

Andrienko, N. and Andrienko, G. (2012) Visual analytics of movement: an overview of methods, tools and procedures. *Information Visualization*, **12**(1), pp. 3–24.

Batty, M. (2015) Data about Cities: Redefining Big, Recasting Small. Paper prepared for the *Data and the City Workshop*. National University of Ireland, Maynooth. Abstract and video available at: http://progcity.maynoothuniversity.ie/2015/09/data-and-the-city-workshop-session-4-videos/ accessed/.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D. and Ratti, C. (2011) Real-time urban monitoring using cell phones: a case study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, **12**(1), pp. 141–151.

Card, S.K., Mackinlay, J.D., and Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufmann.

Chen, J., Guo, D. and MacEachren, A. (2005) Space-Time-Attribute Analysis and Visualization of US Company Data. Minneapolis, MN: IEEE Symposium on Information Visualization, Compendium Volume, pp. 23–25.

Cox, M. and Ellsworth, D. (1997) Application-Controlled Demand Paging for Out-of-Core Visualization. Report NAS-97-010, MS T27A-2. Moffett Field, CA: NASA Ames Research Center.

Crampton, J.W. (2002) Interactivity types in geographic visualization. *Cartography and Geographic Information Science*, **29**(2), pp. 85–98.

Fuller, R.B. (1981) *Critical Path*. New York: St. Martins Press.

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I. and Ratti, C. (2014) Towards a comparative science of cities: using mobile traffic records in New York, London, and Hong Kong. *Com-*

*putational Approaches for Urban Environments*, **13**, pp. 363–387.

Kloeckl, K., Senn, O. and Ratti, C. (2012). Enabling the real-time city: LIVE Singapore! *Journal of Urban Technology*, **19**(2), pp. 89–112.

MacEachren, A.M. (1994) *Visualization in Modern Cartography: Setting the Agenda*. Oxford: Pergamon.

MacEachren, A.M., and Kraak, M. (1997) Exploratory cartographic visualization: advancing the agenda. *Computers and Geosciences*, **23**(4), pp. 335–343.

Meller, H. (1990) *Patrick Geddes, Social Evolutionist and City Planner*. London: Routledge.

Nagel, T., Maitan, M., Duval, E., Moere, A.V., Klerkx, J., Kloeckl, K. and Ratti, C. (2014) Touching Transport – A Case Study on Visualizing Metropolitan Public Transit on Interactive Tabletops. *ACM Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14*, pp. 281–288

Robinson, A.H. (1967) The thematic maps of Charles Joseph Minard. *Imago Mundi*, **21**(1), pp. 95–108.

Roth, R.E. (2013) An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, **19**(12), pp. 2356–2365.

Senn, O., Khairul, M., Maitan, M., Pribadi, R., Shah, M. and Sivaprakasam, R. (2015) Datacollider, in Proceedings of SIGGRAPH Asia 2015 Visualization in High Performance Computing, Kobe, Japan.

Snijders, C., Matzat, U. and Reips, U.-D. (2012) 'Big Data': big gaps of knowledge in the field of internet. *International Journal of Internet Science*, **7**, pp. 1–5.

Townsend, A.M. (2014) *Smart Cities: Big Data, Civic Hackers, and The Quest for a New Utopia*. New York: W.W. Norton.

Vande Moere, A. and Hill, D. (2012) Designing for the situated and public visualization of urban data. *Journal of Urban Technology*, **19**(2), pp. 25–46.

van Wijk, J.J. (2005) The value of visualization. *VIS 05 IEEE Visualization*, pp. 79–86

## Appendix: Terminology for Visualizations

*Map*: A static tool for representing collected information about what exists in physical territory. Intended to show one or several dimensions of physical space.

Maps depict geospatial information, and are a classic, centuries old technique to represent our physical environment. For example: a maritime navigational document showing land masses, water bodies and ports.

*Infographic*: A visual representation of data. Intended to communicate information to a defined audience.

Infographics (or 'information graphics') are graphical representations of data or information, in order to communicate some facts, stories or insights visually. They are – in contrast to data visualization – typically static or at least non-interactive. While they can consist of multiple visualizations or graphs side-by-side, they tend to be visually less complex, and often use a broad design style. For example: two side-by-side coca cola bottles filled up to different heights, indicating average consumption in two different countries.

*Geospatial Infographic*: A visual representation of data that incorporates a map. Intended to communicate information from basic geospatial data to a wide audience.

Geospatial infographics are graphical representations of geospatial data or information. Maps typically are a part of a geospatial infographic. For example: a map of United States voting trends, showing red and blue states for a given year.

*Geospatial Data Visualization (Geovisualization, Flow Map)*: A visual representation of dynamic (time-based) data across physical territory. Intended to show relationships or changes in geospatial data.

Information visualization focuses on graphical representations of data to help

people understand and analyze data. This foregrounds the question of how to use visualization techniques to effectively and efficiently reveal the internal structure of the data (Van Wijk, 2005). While information visualization typically focuses on abstract data (Chen *et al*., 2005), geovisualization, or geographic information visualization deals with data that has physical and spatial correspondence. However, it covers not only geospatial data, but also multivariate georeferenced data, typically including time-varying properties. Geovisualization is a visual analysis of patterns, relationships, and trends, and allows exploration and understanding of spatio-temporal information.

Flow maps are a subset of this category: visualizations depicting movement between geographical locations on a map. They are a combination of flow charts and maps, typically showing a background map overlaid with lines connecting the flow origins with the destinations, often with the flow magnitudes mapped to line thickness. For example: a map of telecommunication patterns between one city and other cities, animated to show how those connections change throughout the day.

*Interactive Geovisualization*: A visual representation of dynamic (time-based) data across physical territory, which can be filtered by a user (according to place, time, dimension, characteristic). Intended to show relationships or changes in geospatial data, and can be browsed by a wide variety of users.

Data is pre-processed (by visualization and/or data science experts) to derive specific insights that can be communicated to users. Advanced data analysis techniques can be applied to the raw data, and the interactive tool can focus on communicating those insights in a personally relevant way. Interactivity can further two main goals:

(*a*) Presenting a larger amount of information without overwhelming users (particularly non-specialists). For example, an Interactive Geovisualization may show all transportation data, including car, bus, taxi, bike, train, metro, but only display one of those at a time.

(*b*) Allowing users to find personally relevant insights. In the above example, a user may only be interested in exploring the spatial relationship between bike routes and public transit.

For example: a browser for all transportation data in a city.

*Geospatial Data Tool*: A tool for collecting, structuring, manipulating, analyzing, and visualizing time-based geospatial data. Outputs can be interactive, to be browsed by a wide variety of users. Intended to empower anyone, regardless of background and expertise, to work with data, find, and visualize insights.

A geospatial data tool will be applicable for every step in the production and consumption chain, from raw data to visualization. Stakeholders may apply this tool for the entire process, or interact with it at different stages. For example, an urban researcher may upload, analyze and visualize data, or a graphic designer may enter at the final stage to turn insights into a compelling dynamic graphic. The greatest value of this tool will be in *finding* insights and *presenting* them. That is, its strength is in enabling anyone to actively and productively explore raw data. For example: the *Datacollider*.