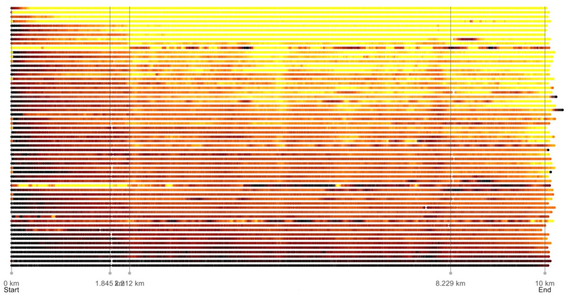
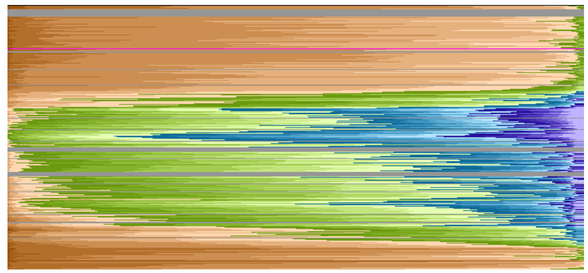
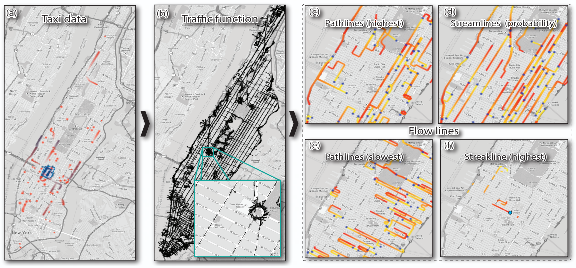
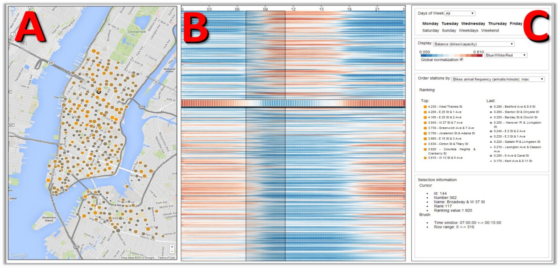
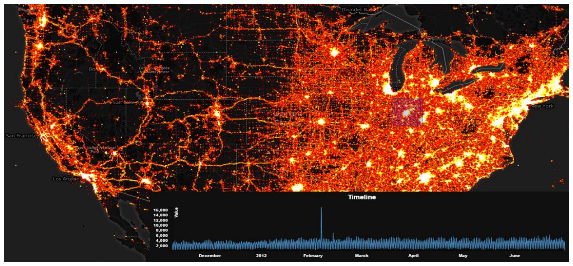
Visual Data Analysis of Unstructured and Big Data
João Comba

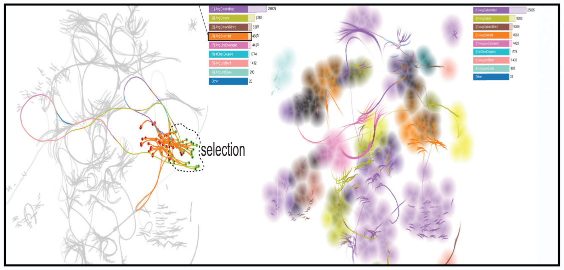












Outline

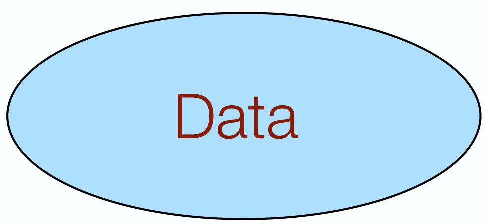
- Basic Visual Data Analysis Concepts
- 12 Examples of Interesting Problems
  - High-level problem description
    - Data
    - Important questions to answer
    - Video illustrating solution

Data is everywhere

- Social: every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. source: http://www.internetlivestats.com
- Urban: NYC Taxi, Metro and Bike data, San Francisco open data, etc.
- Sensor: various types
  - fitness (Garmin, Nike, Polar, Fitbit, GPS trackers, etc)
  - health (Withings, phone apps, etc)
- Scientific: simulation, medicine, etc

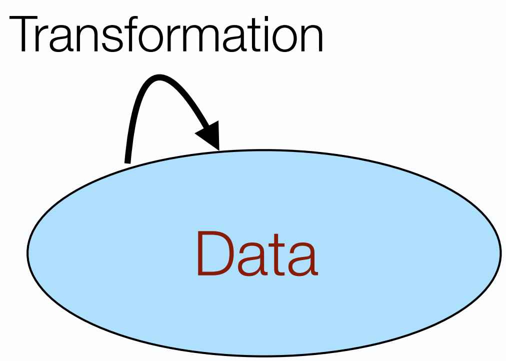
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



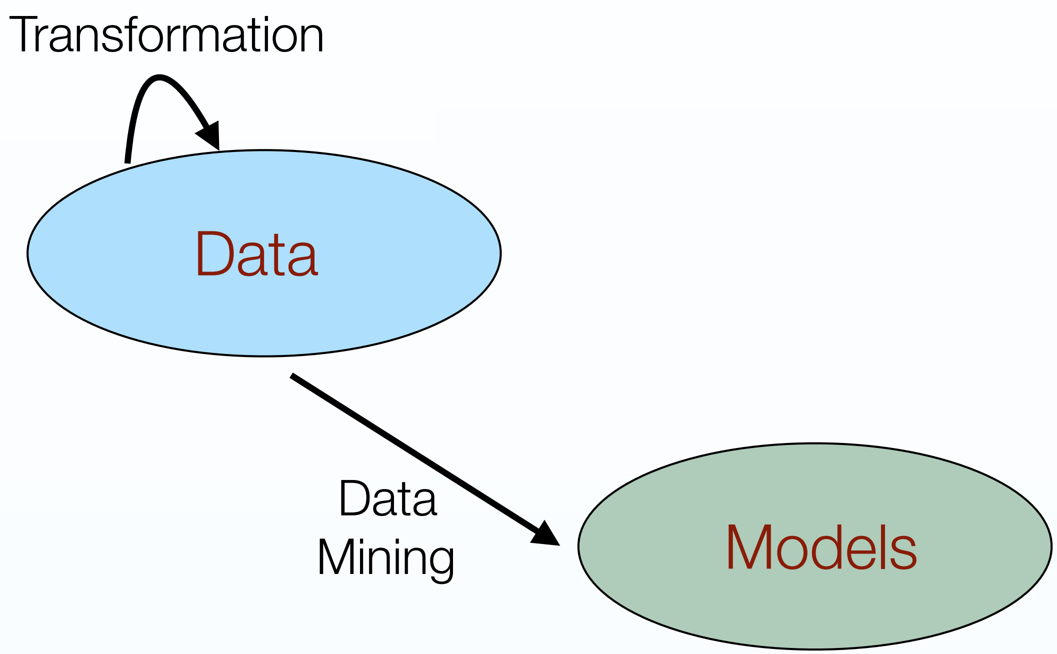
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



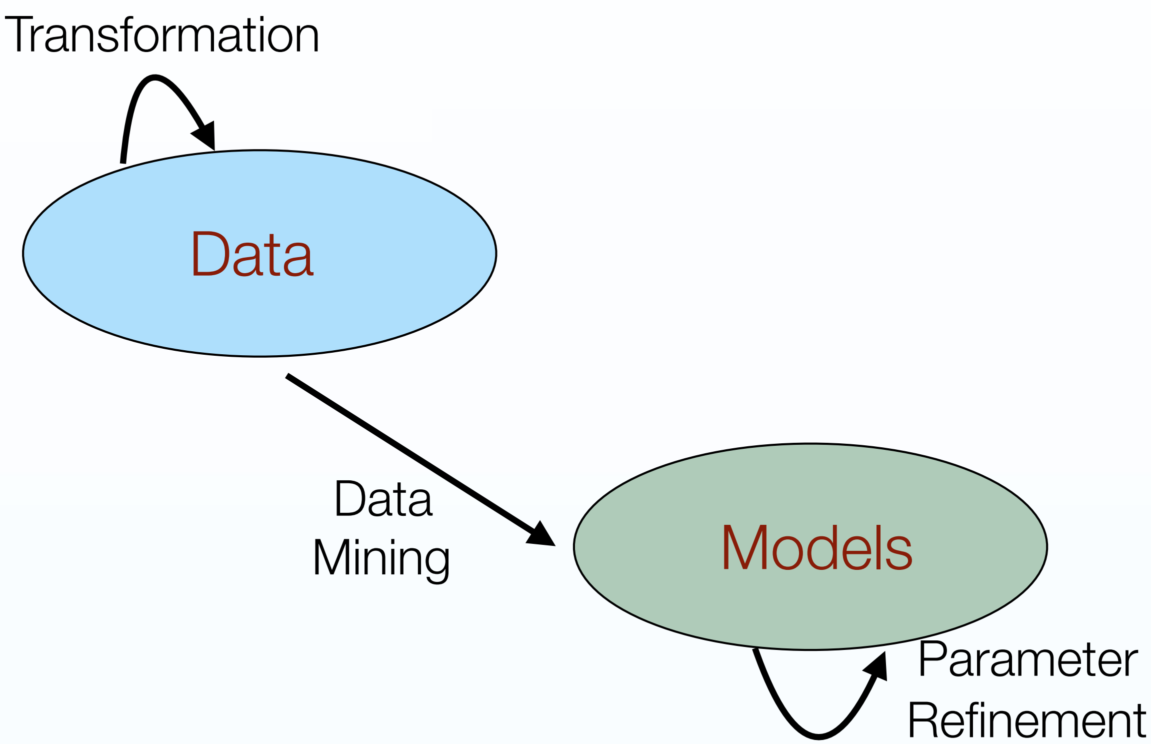
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



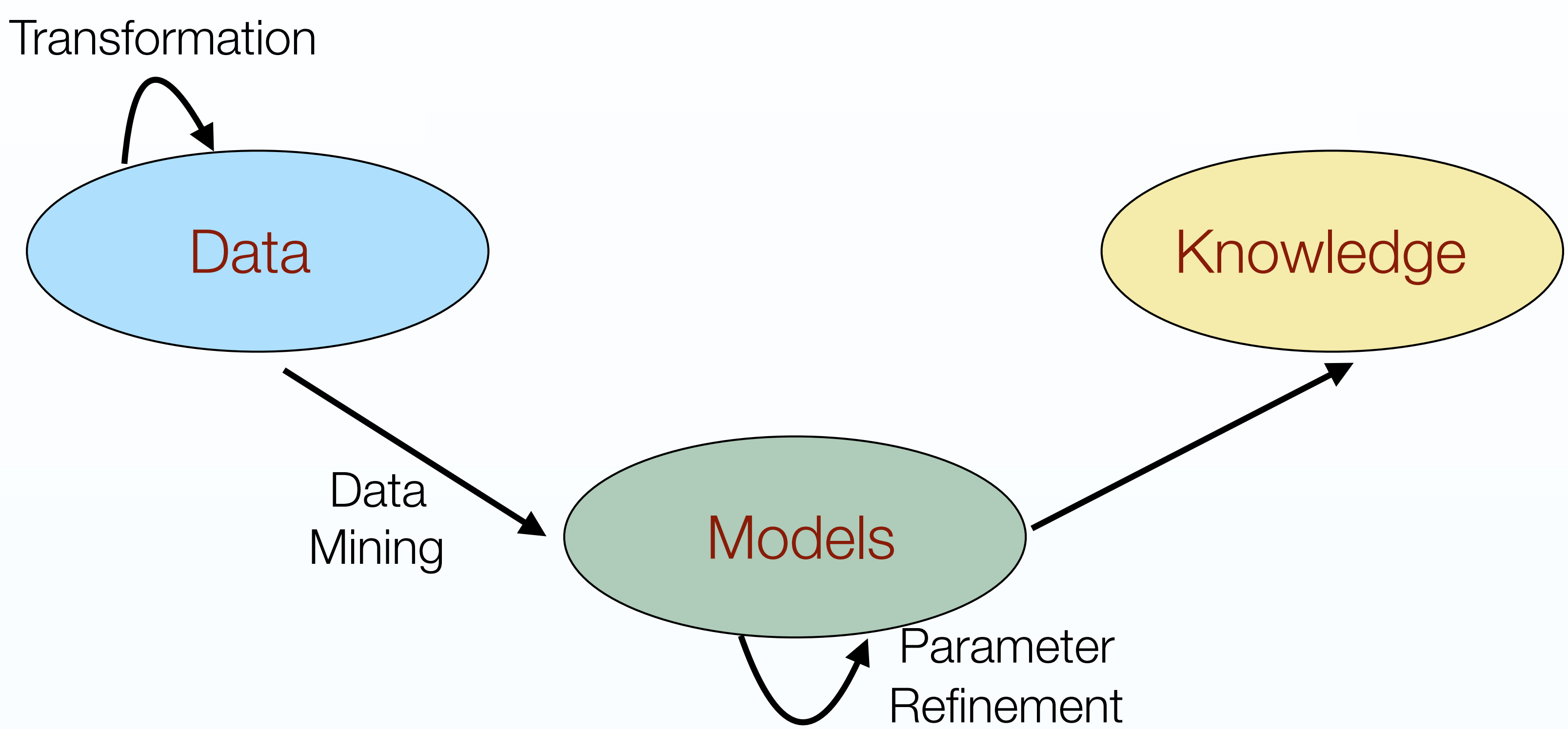
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



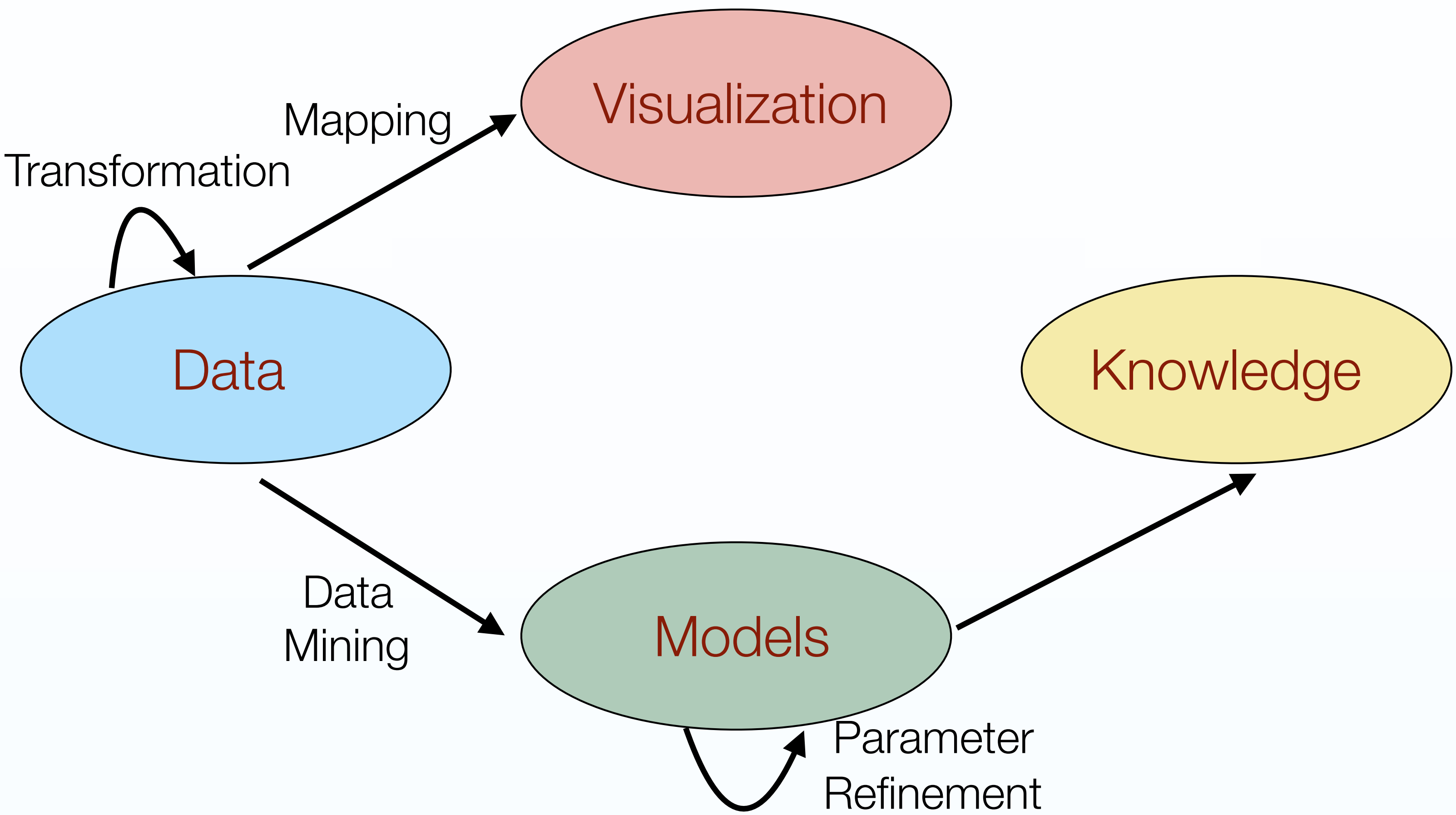
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



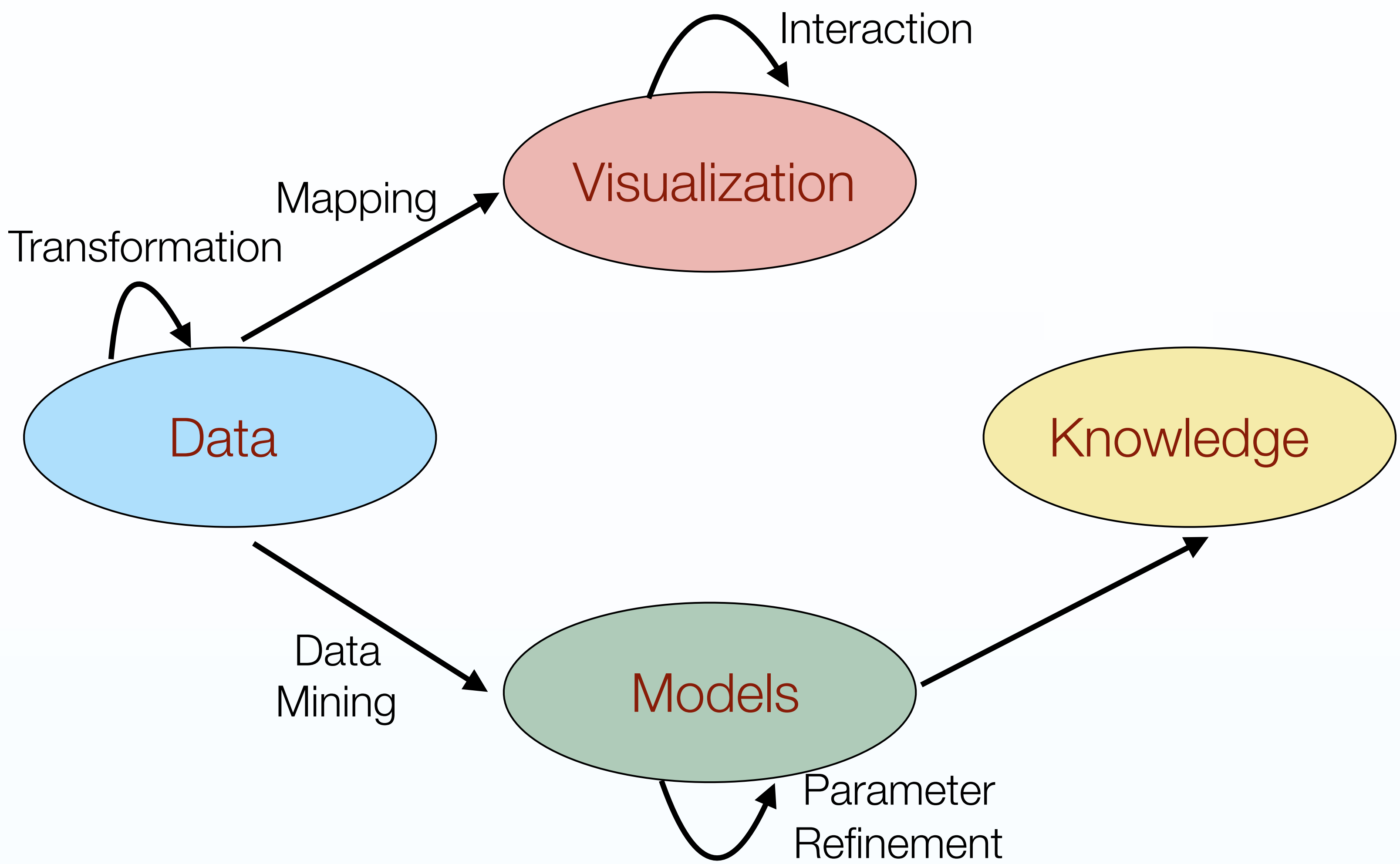
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



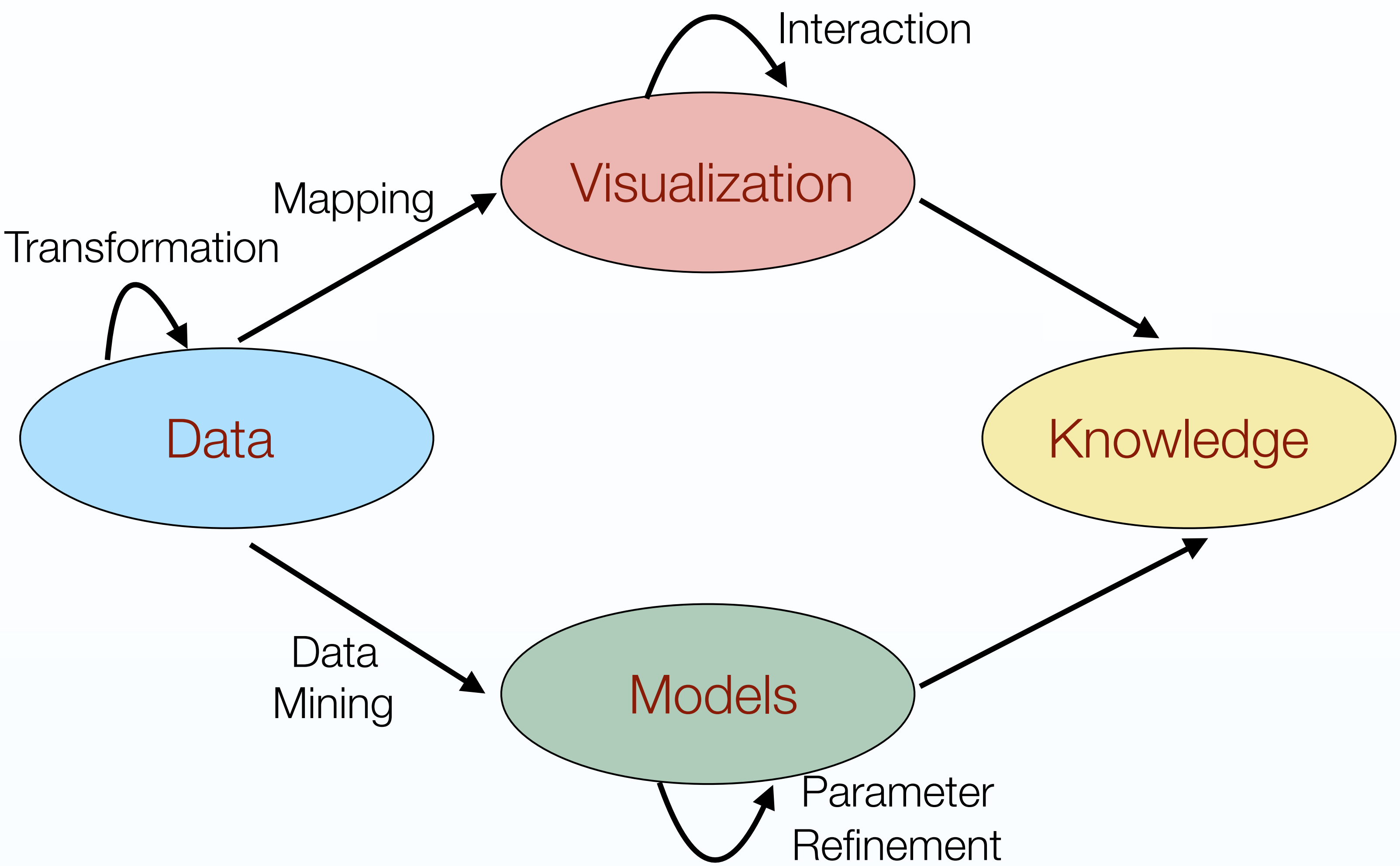
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



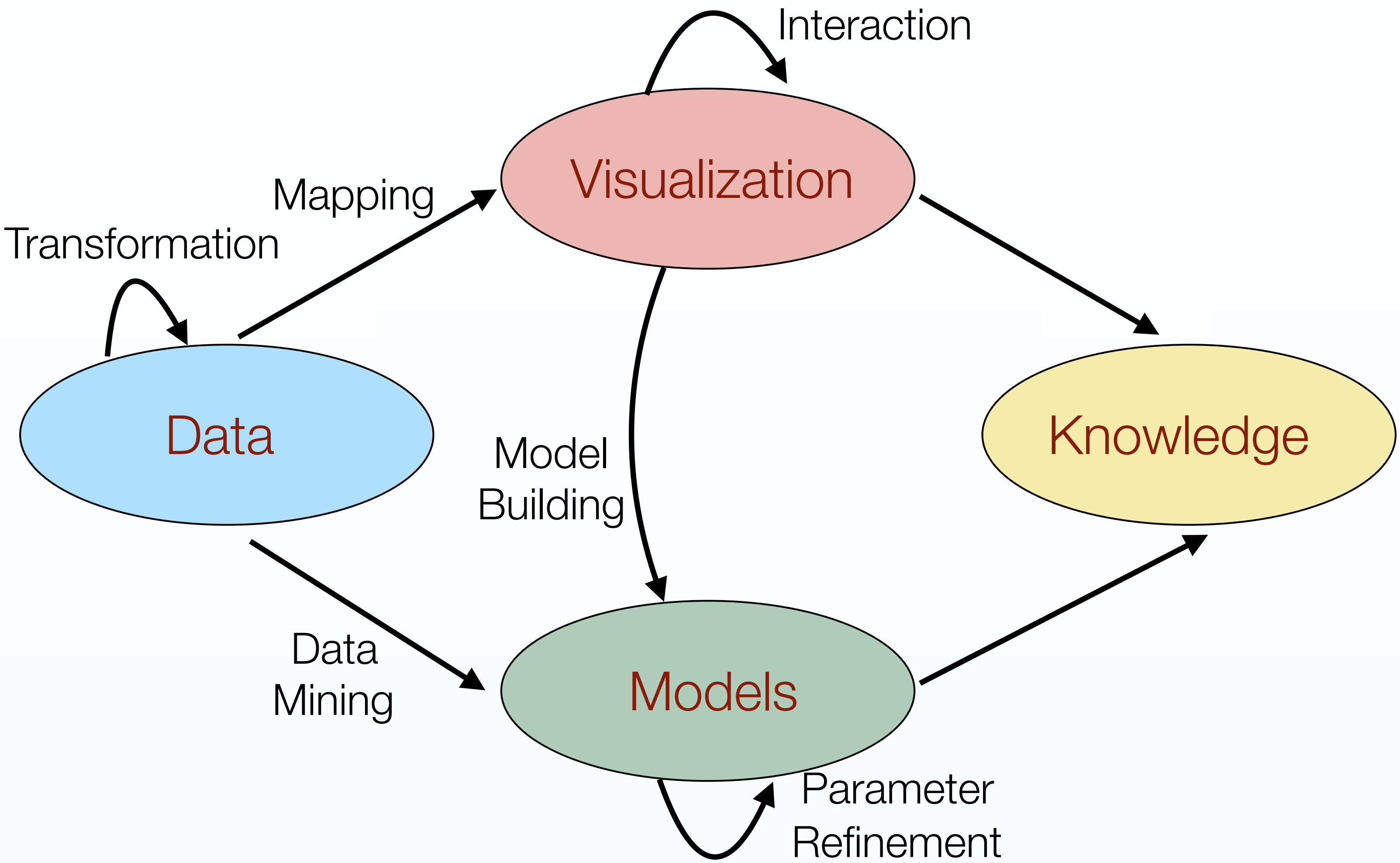
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



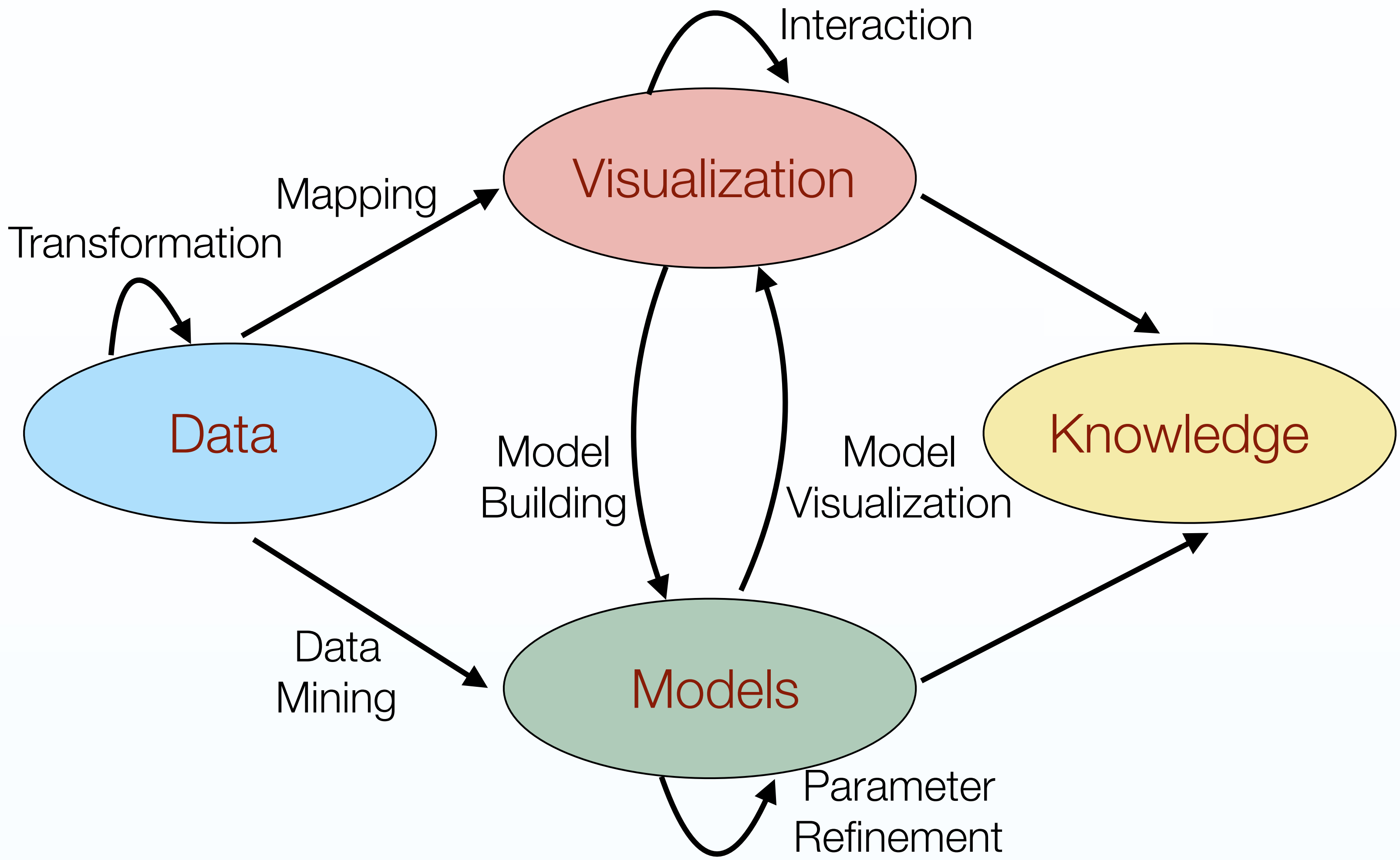
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



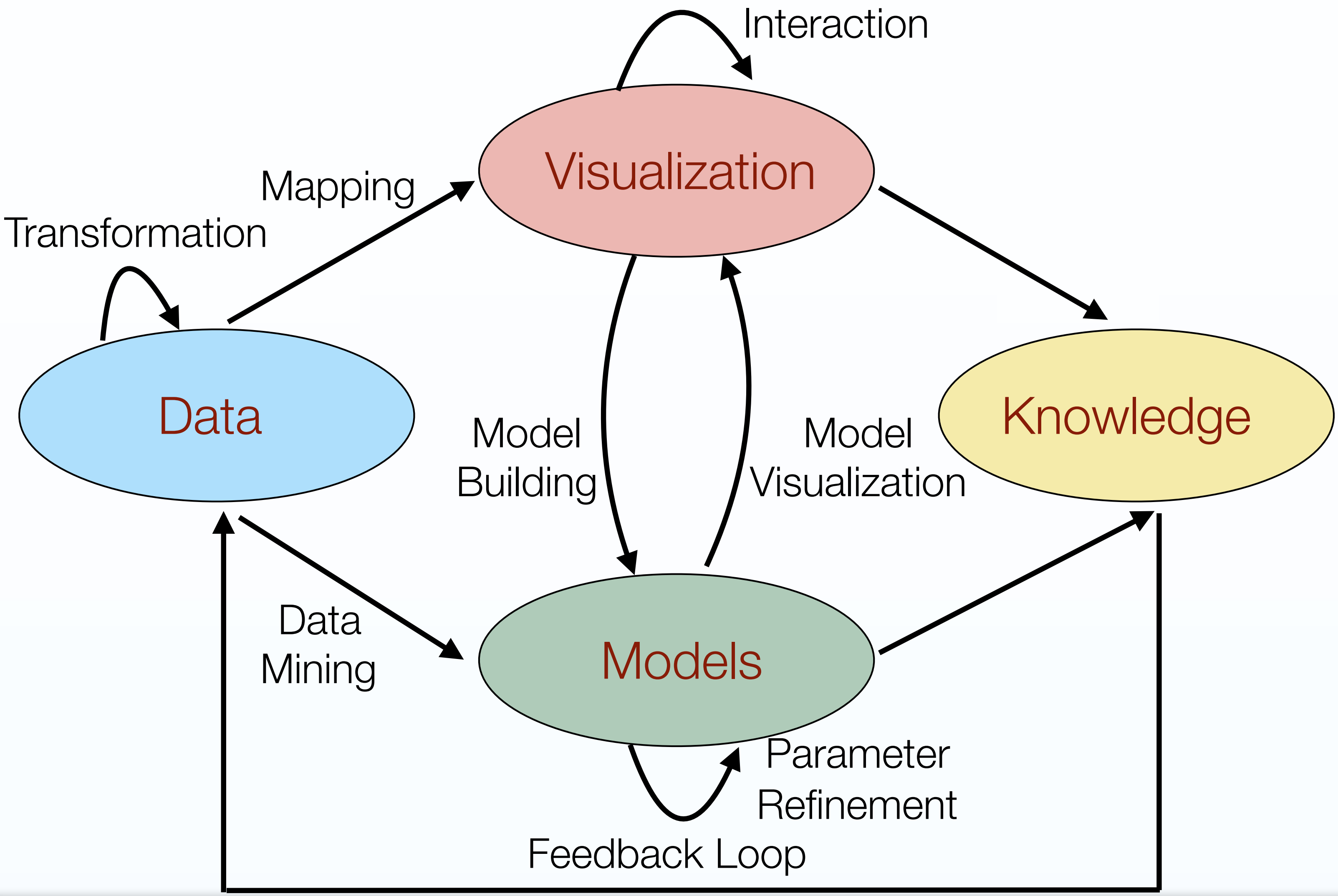
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



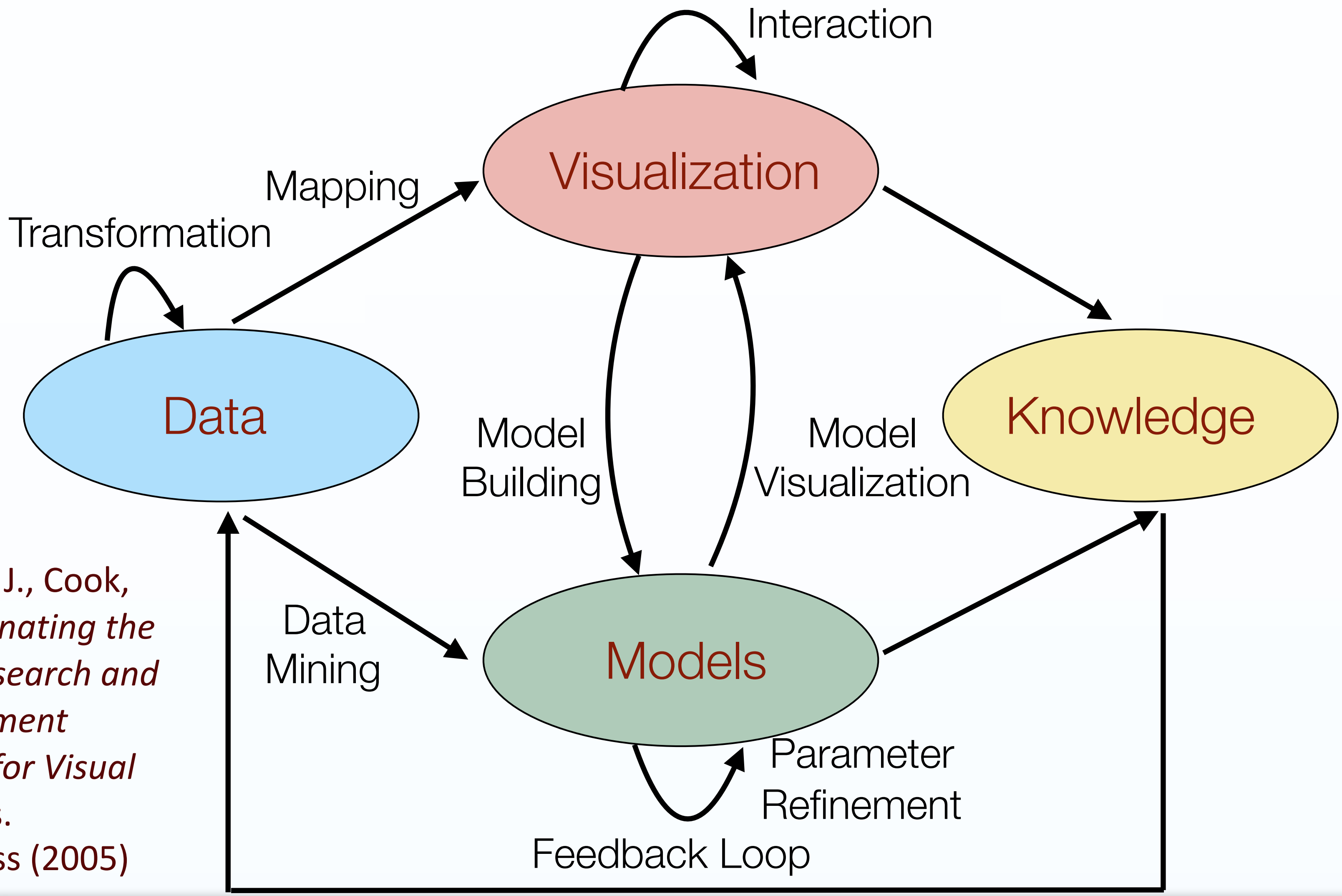
Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



Visual Data Analysis: The role of Data Analysis

- Data is big, unstructured, and often complex.
- Finding patterns, associations, or relationships in data using visualization, mining and analytical tools



Thomas, J., Cook, K.: *Illuminating the Path: Research and Development Agenda for Visual Analytics.* IEEE-Press (2005)

Visualization Reveals Data

- show the data
- induce the viewer to think about the substance rather about methodology, graphic design …
- avoid distorting what the data have to say
- present many numbers in small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from broad view to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set

Edward Tufte. *The Visual Display of Quantitative Information*. 1983 (p. 13)

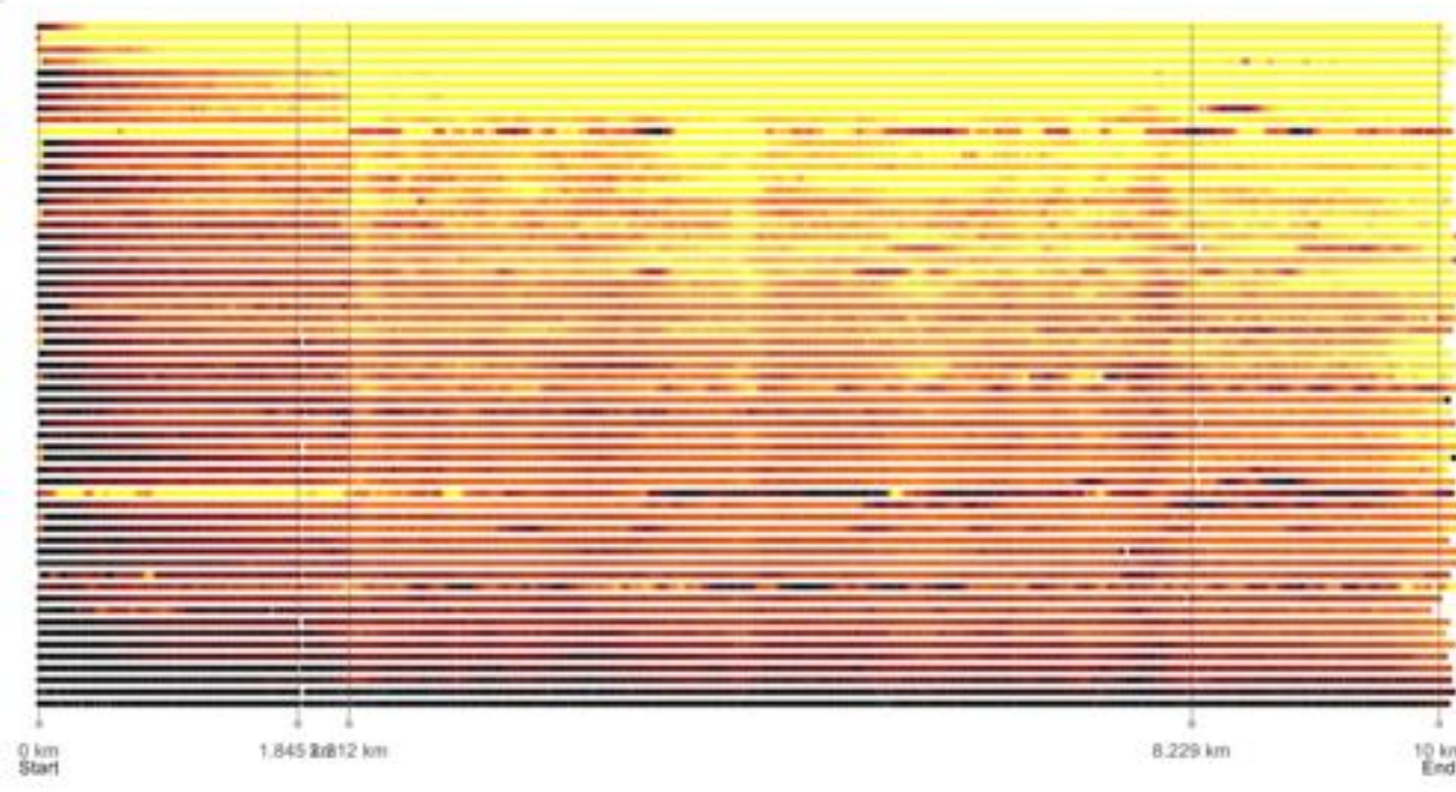
Examples of Interesting Problems

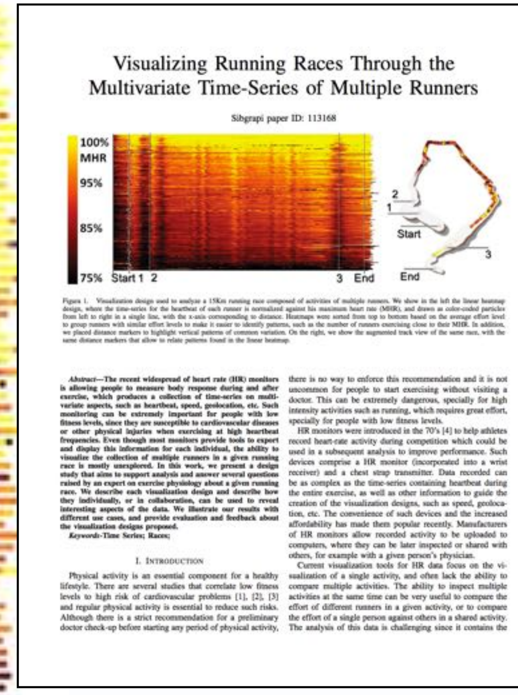
- Published Work

1. Visualization of Running Races
2. Environmental Phenological Analysis (Biology)
3. Traffic Analysis in Urban Environments
4. Visual Analysis of Bike Sharing Systems
5. Real-Time Visual Exploration of Big Data
6. Visual Exploration of Software Repositories

Published Work

Visualizing Running Races Through the Multivariate Time-Series of Multiple Runners





Visualizing Running Races Through the Multivariate Time-Series of Multiple Runners

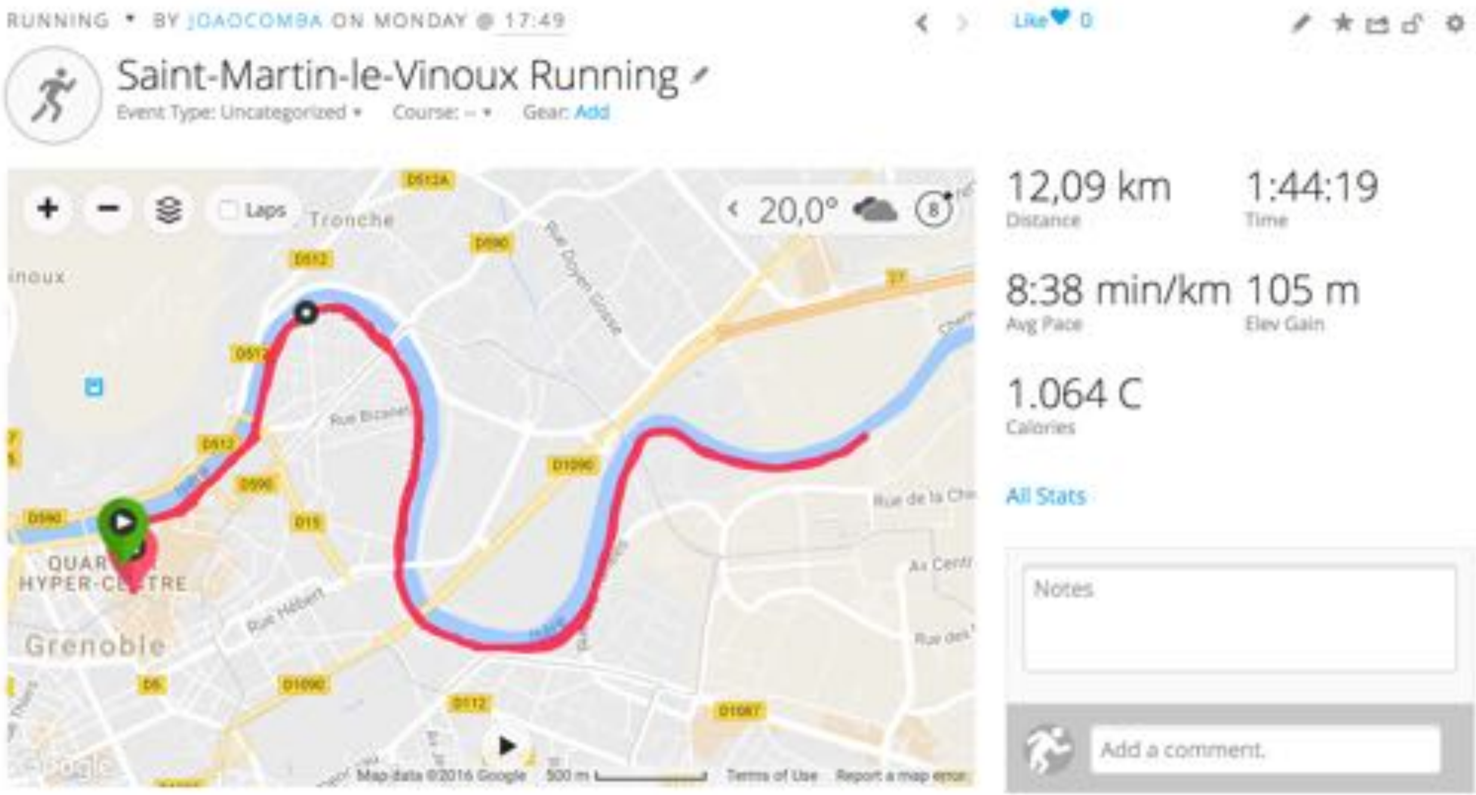
Sibgrapi paper ID: 113168

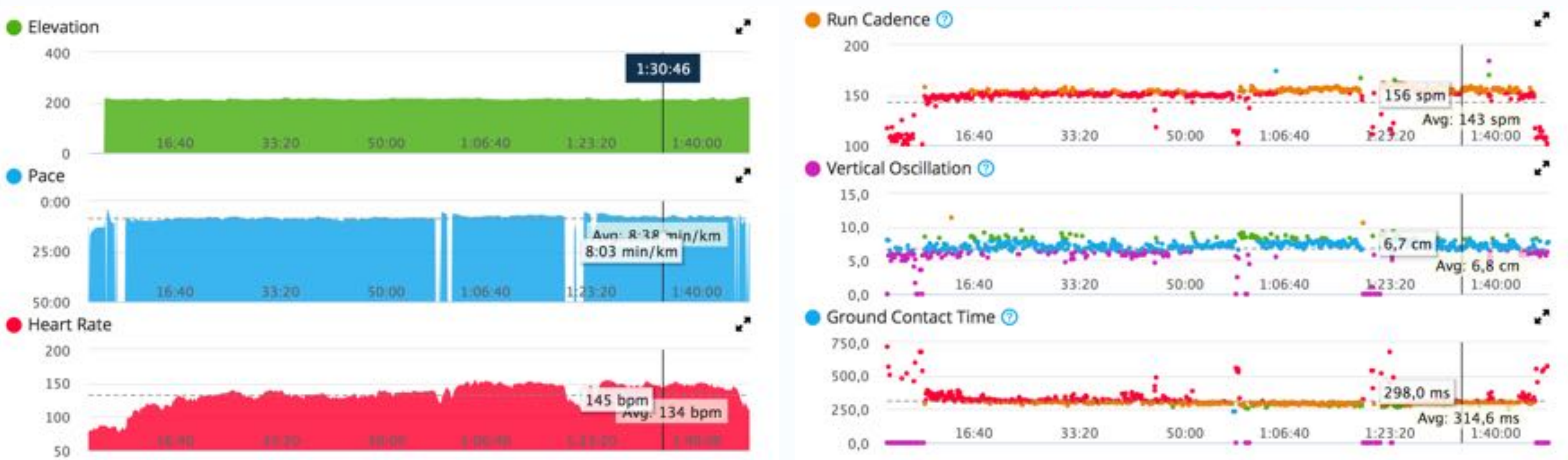
Visualizing Running Races Through the Multivariate Time-Series of Multiple Runners

Guilherme Oliveira, João Comba, Rafael Torchelsen, Claudio Silva, Maristela Padilha

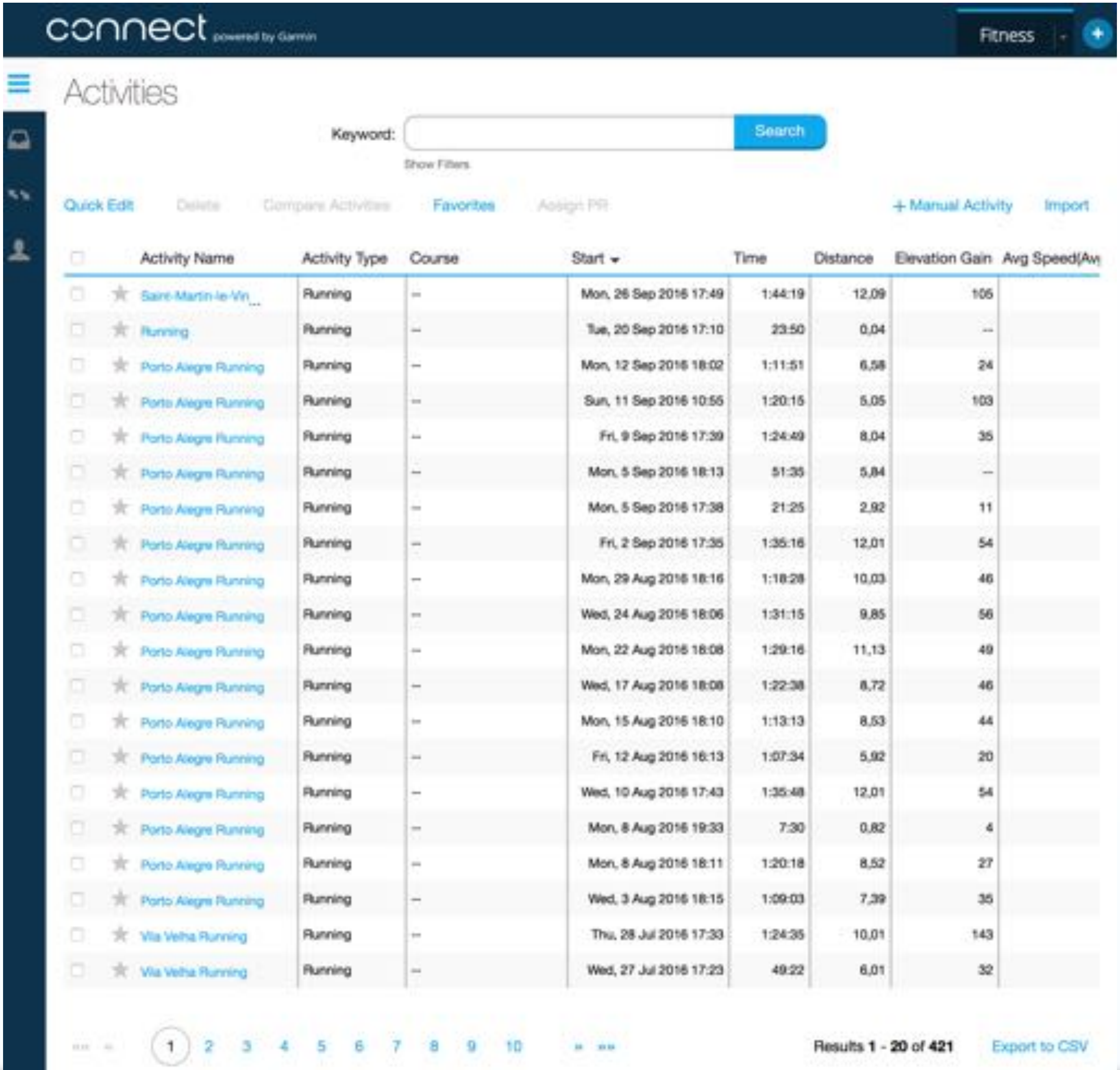
Proceedings of Conference on Graphics, Patterns and Images (SIBGRAPI 2013).

GPS and Heart-Rate Monitors





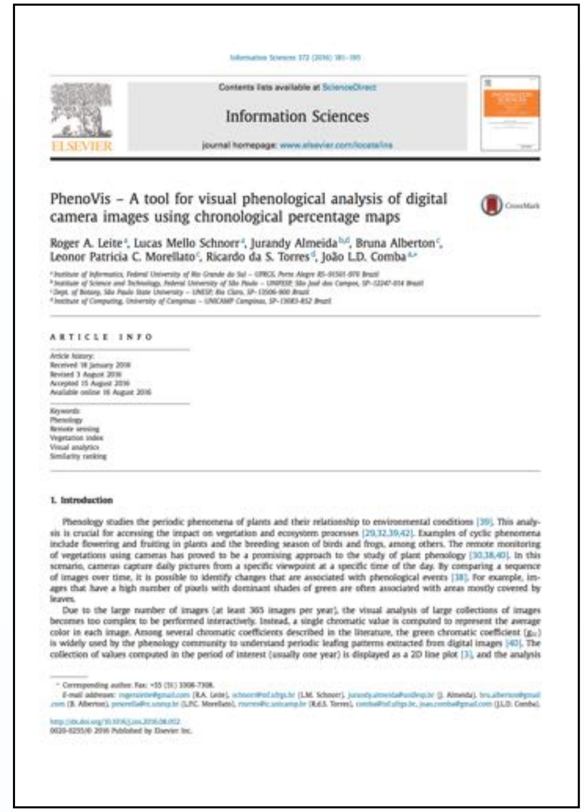
GPS and Heart-Rate Monitors



Activity Name	Activity Type	Course	Start	Time	Distance	Elevation Gain	Avg Speed(Av
Saint-Martin-le-Vin...	Running	--	Mon, 26 Sep 2016 17:49	1:44:19	12,09	105	
Running	Running	--	Tue, 20 Sep 2016 17:10	23:50	0,04	--	
Porto Alegre Running	Running	--	Mon, 12 Sep 2016 18:02	1:11:51	6,58	24	
Porto Alegre Running	Running	--	Sun, 11 Sep 2016 10:55	1:20:15	5,05	103	
Porto Alegre Running	Running	--	Fri, 9 Sep 2016 17:39	1:24:49	8,04	35	
Porto Alegre Running	Running	--	Mon, 5 Sep 2016 18:13	51:35	5,84	--	
Porto Alegre Running	Running	--	Mon, 5 Sep 2016 17:38	21:25	2,92	11	
Porto Alegre Running	Running	--	Fri, 2 Sep 2016 17:35	1:35:16	12,01	54	
Porto Alegre Running	Running	--	Mon, 29 Aug 2016 18:16	1:18:28	10,03	46	
Porto Alegre Running	Running	--	Wed, 24 Aug 2016 18:06	1:31:15	9,85	56	
Porto Alegre Running	Running	--	Mon, 22 Aug 2016 18:08	1:29:16	11,13	49	
Porto Alegre Running	Running	--	Wed, 17 Aug 2016 18:08	1:22:38	8,72	46	
Porto Alegre Running	Running	--	Mon, 15 Aug 2016 18:10	1:13:13	8,53	44	
Porto Alegre Running	Running	--	Fri, 12 Aug 2016 16:13	1:07:34	5,92	20	
Porto Alegre Running	Running	--	Wed, 10 Aug 2016 17:43	1:35:48	12,01	54	
Porto Alegre Running	Running	--	Mon, 8 Aug 2016 19:33	7:30	0,82	4	
Porto Alegre Running	Running	--	Mon, 8 Aug 2016 18:11	1:20:18	8,52	27	
Porto Alegre Running	Running	--	Wed, 3 Aug 2016 18:15	1:09:03	7,39	35	
Via Velha Running	Running	--	Thu, 28 Jul 2016 17:33	1:24:35	10,01	143	
Via Velha Running	Running	--	Wed, 27 Jul 2016 17:23	49:22	6,01	32	

GPS and Heart-Rate Monitors

Phenological Analysis Using Chronological Percentage Maps



Information Sciences 372 (2016) 181–195

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

PhenoVis – A tool for visual phenological analysis of digital camera images using chronological percentage maps

Roger A. Leite[a], Lucas Mello Schnorr[a], Jurandy Almeida[b,d], Bruna Alberton[c], Leonor Patricia C. Morellato[c], Ricardo da S. Torres[d], João L.D. Comba[a,*]

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 3 August 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
Phenology
Remote sensing
Vegetation index
Visual analytics
Similarity ranking

1. Introduction

Phenology studies the periodic phenomena of plants and their relationship to environmental conditions [39]. This analysis is crucial for accessing the impact on vegetation and ecosystem processes [29,32,39,42]. Examples of cyclic phenomena include flowering and fruiting in plants and the breeding season of birds and frogs, among others. The remote monitoring of vegetations using cameras has proved to be a promising approach to the study of plant phenology [30,38,40]. In this scenario, cameras capture daily pictures from a specific viewpoint at a specific time of the day. By comparing a sequence of images over time, it is possible to identify changes that are associated with phenological events [38]. For example, images that have a high number of pixels with dominant shades of green are often associated with areas mostly covered by leaves.

Due to the large number of images (at least 365 images per year), the visual analysis of large collections of images becomes too complex to be performed interactively. Instead, a single chromatic value is computed to represent the average color in each image. Among several chromatic coefficients described in the literature, the green chromatic coefficient (g_{cc}) is widely used by the phenology community to understand periodic leafing patterns extracted from digital images [40]. The collection of values computed in the period of interest (usually one year) is displayed as a 2D line plot [3], and the analysis

* Corresponding author. Fax: +55 (51) 3308-7308.

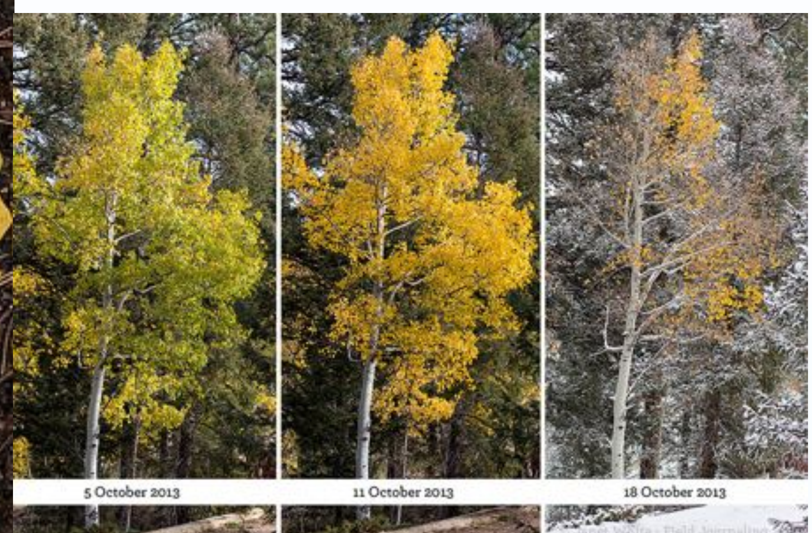
PhenoVis – A Tool for Visual Phenological Analysis of Digital Camera Images Using Chronological Percentage Maps

Guilherme Oliveira, Lucas Schnorr, Jurandy Almeida, Bruna Alberton, Leonor Patricia Morellato, Ricardo Torres, João Comba

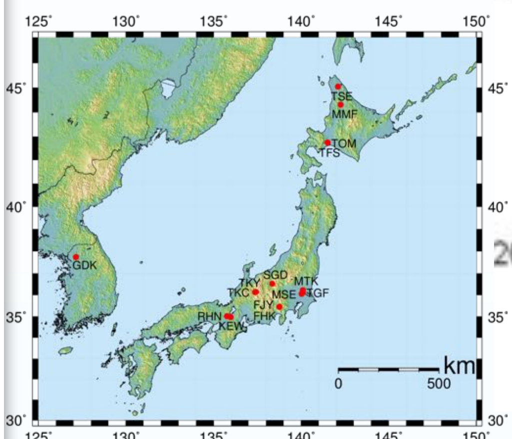
Information Sciences 372 (2016) 181–195

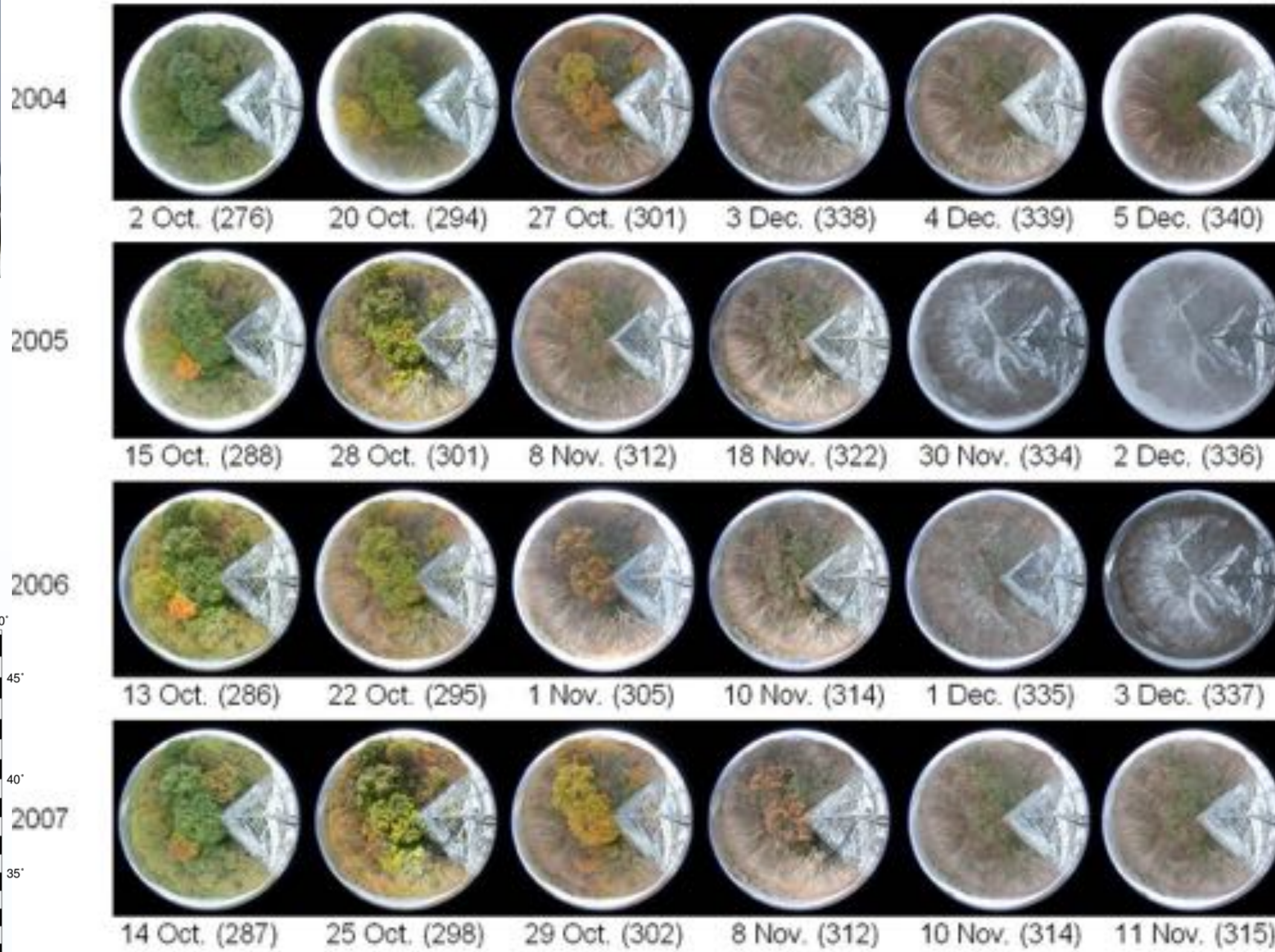
Phenology

- Study of periodic plant and animal life cycle events and how these are influenced by seasonal and inter-annual variations in climate



DATA: Continuous, Long Term and Multi-ecosystem

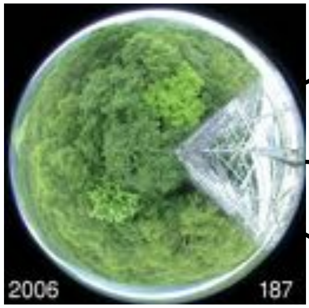




Phenological Eyes Network (PEN)
http://pen.agbi.tsukuba.ac.jp/index.html

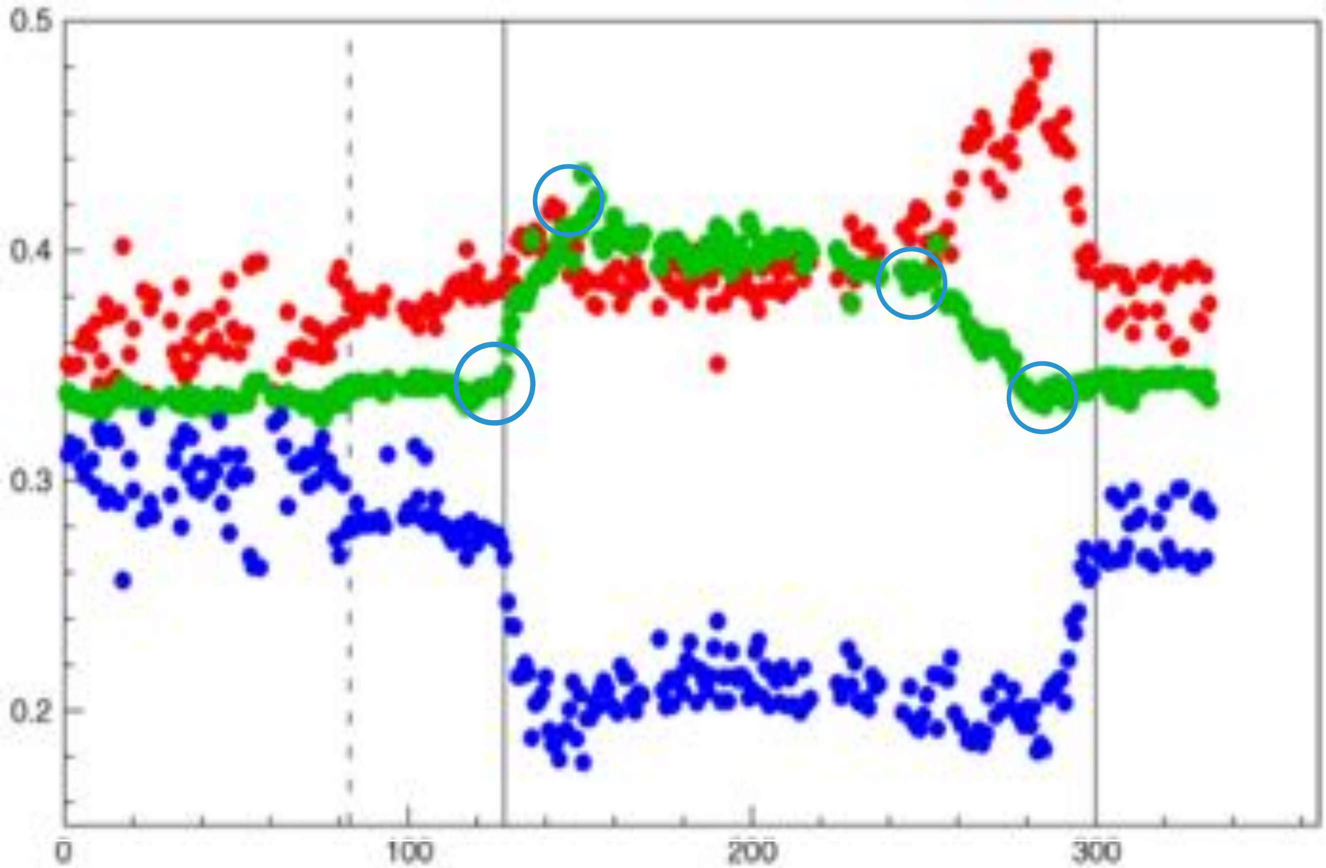
Phenological phases

- Average used to evaluate phenological variations in a year



- %r = r / (r+g+b)
- %g = g / (r+g+b)
- %b = b / (r+g+b)

Time: 1 year



1
2
3
4
5
6
7
8
9
10
11
12

The problem with the average

2004
day 281

2007
day 28

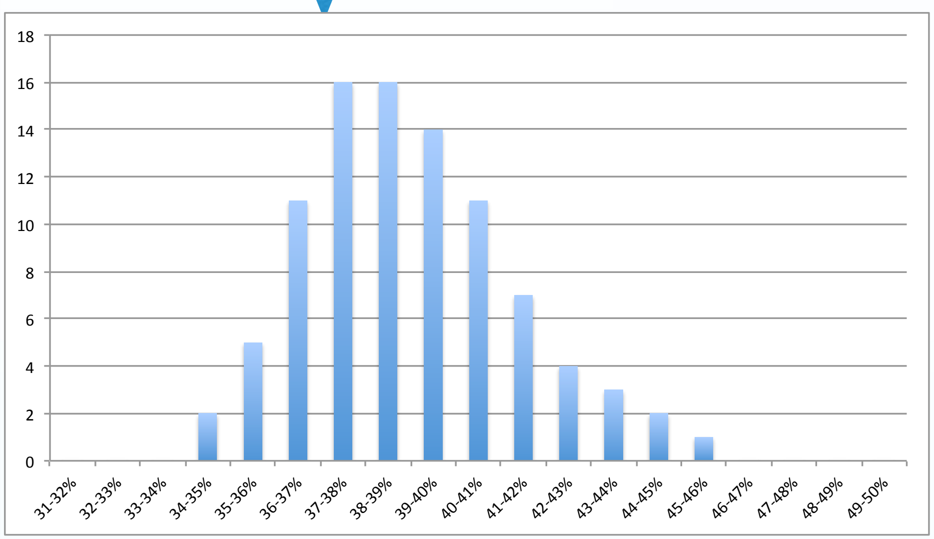
2012
day 293

2013
day 29

$\%g_{cc} = G / (R+G+B) = \mathbf{0.3905}$

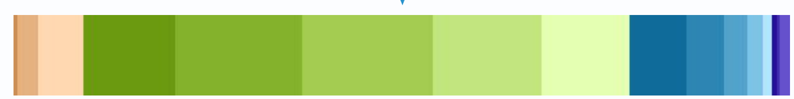
Percentage Maps

Compute for each pixel a value (e.g. gcc) and generate a percentage distribution



Given a pre-defined color map

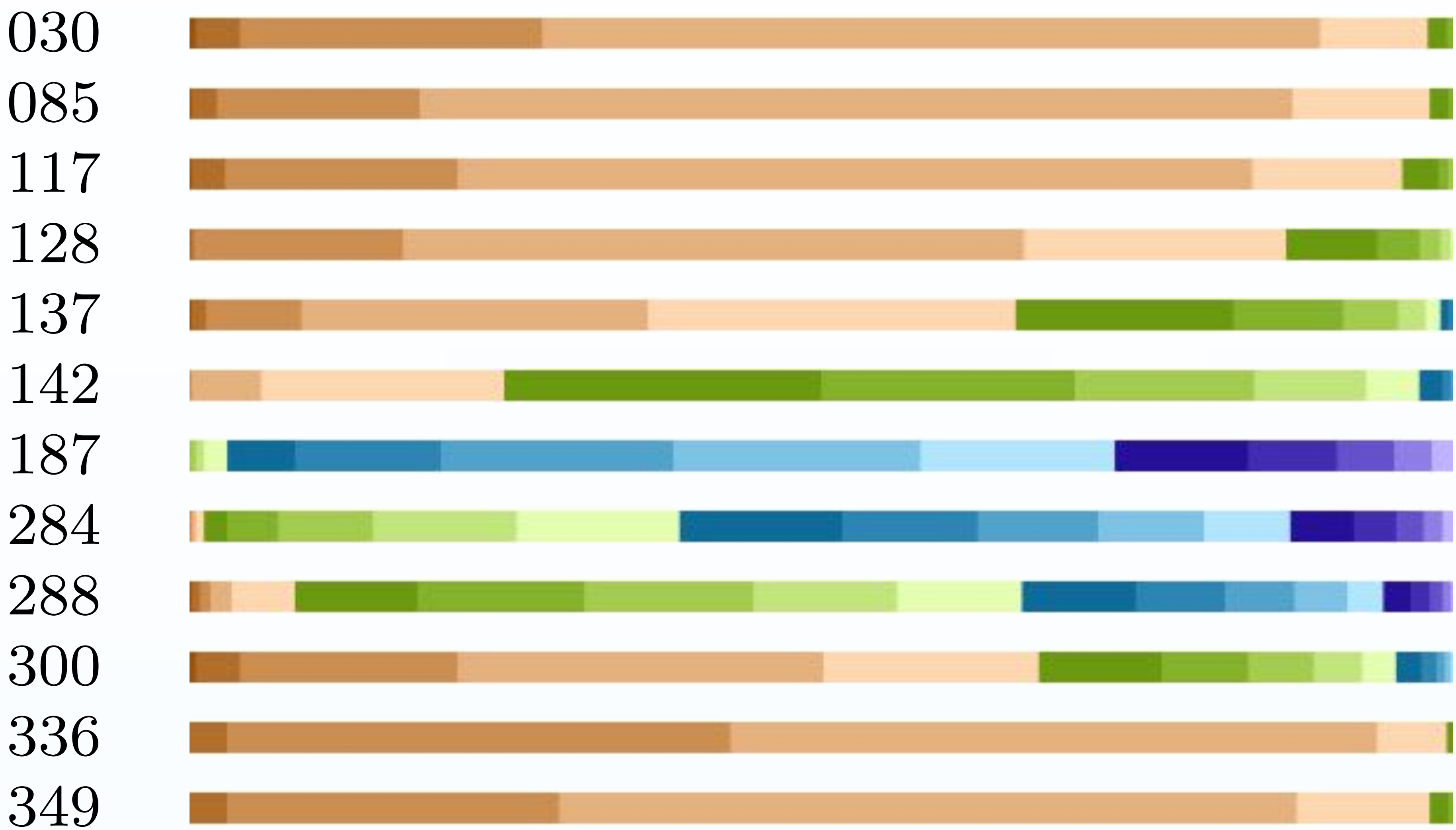
Scale it proportionally to the percentage distribution



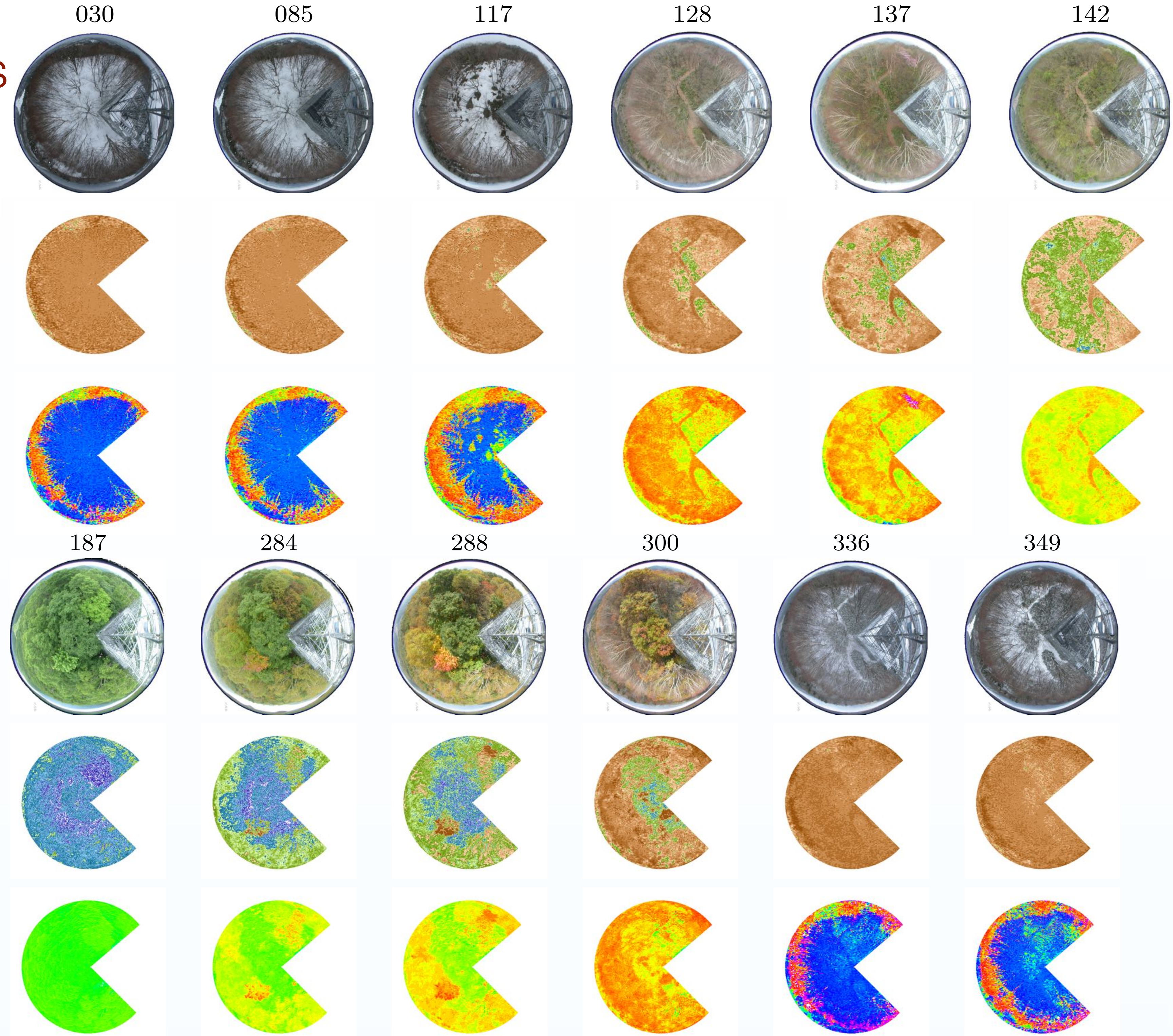
Percentage Map

Chronological Percentage Maps (CPMs)

Stacks of percentage maps in chronological order

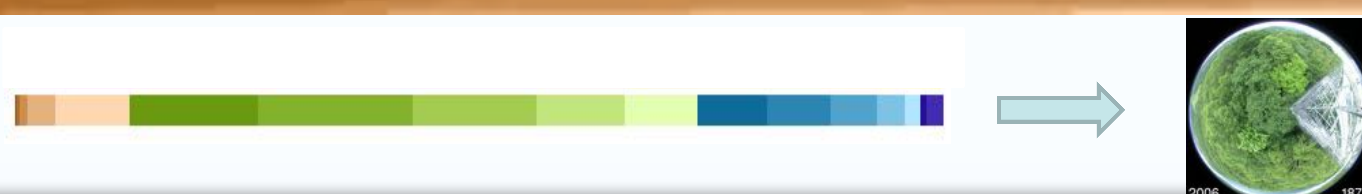


CPMs

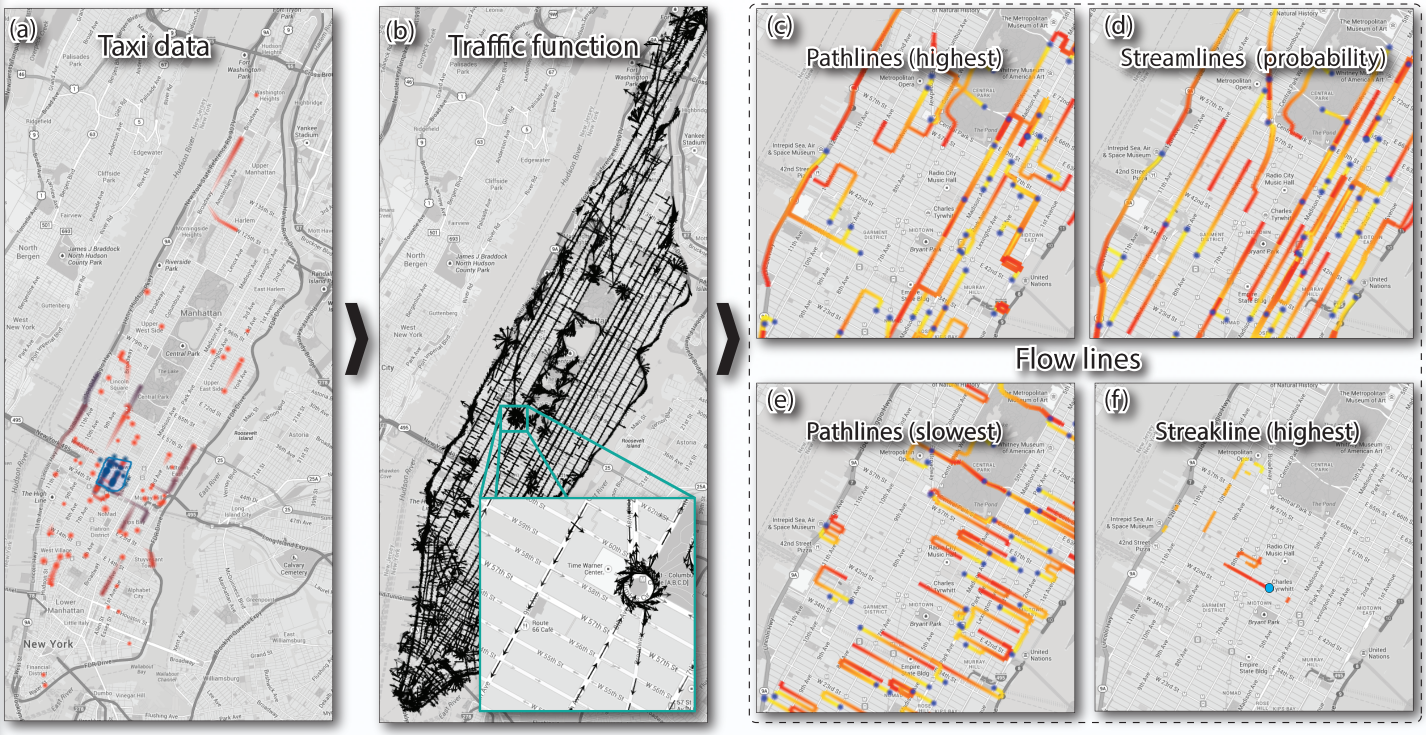


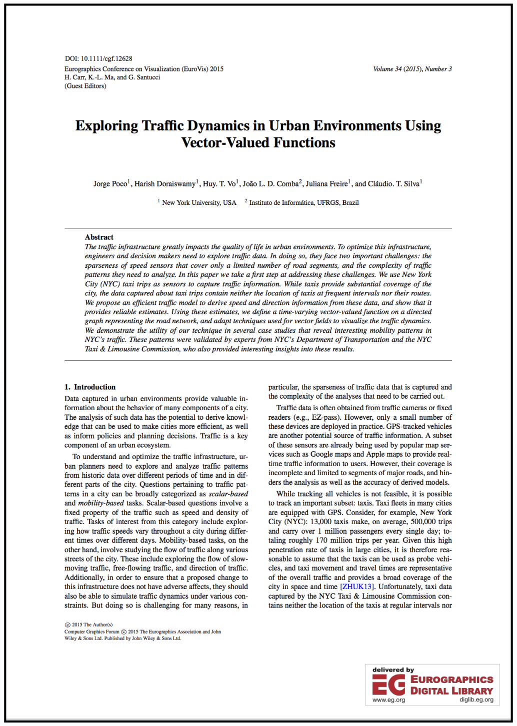
Chronological Percentage Maps (CPMs)

Stacks of percentage maps in chronological order



Using Probabilistic Vector-Valued Functions to Explore Traffic Dynamics in Urban Environments




DOI: 10.1111/cgf.12628
Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)
Volume 34 (2015), Number 3

Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions

Jorge Poco[1], Harish Doraiswamy[1], Huy. T. Vo[1], João L. D. Comba[2], Juliana Freire[1], and Cláudio. T. Silva[1]

[1] New York University, USA [2] Instituto de Informática, UFRGS, Brazil

Abstract
The traffic infrastructure greatly impacts the quality of life in urban environments. To optimize this infrastructure, engineers and decision makers need to explore traffic data. In doing so, they face two important challenges: the sparseness of speed sensors that cover only a limited number of road segments, and the complexity of traffic patterns they need to analyze. In this paper we take a first step at addressing these challenges. We use New York City (NYC) taxi trips as sensors to capture traffic information. While taxis provide substantial coverage of the city, the data captured about taxi trips contain neither the location of taxis at frequent intervals nor their routes. We propose an efficient traffic model to derive speed and direction information from these data, and show that it provides reliable estimates. Using these estimates, we define a time-varying vector-valued function on a directed graph representing the road network, and adapt techniques used for vector fields to visualize the traffic dynamics. We demonstrate the utility of our technique in several case studies that reveal interesting mobility patterns in NYC's traffic. These patterns were validated by experts from NYC's Department of Transportation and the NYC Taxi & Limousine Commission, who also provided interesting insights into these results.

1. Introduction

Data captured in urban environments provide valuable information about the behavior of many components of a city. The analysis of such data has the potential to derive knowledge that can be used to make cities more efficient, as well as inform policies and planning decisions. Traffic is a key component of an urban ecosystem.

To understand and optimize the traffic infrastructure, urban planners need to explore and analyze traffic patterns from historic data over different periods of time and in different parts of the city. Questions pertaining to traffic patterns in a city can be broadly categorized as *scalar-based* and *mobility-based* tasks. Scalar-based questions involve a fixed property of the traffic such as speed and density of traffic. Tasks of interest from this category include exploring how traffic speeds vary throughout a city during different times over different days. Mobility-based tasks, on the other hand, involve studying the flow of traffic along various streets of the city. These include exploring the flow of slow-moving traffic, free-flowing traffic, and direction of traffic. Additionally, in order to ensure that a proposed change to this infrastructure does not have adverse affects, they should also be able to simulate traffic dynamics under various constraints. But doing so is challenging for many reasons, in particular, the sparseness of traffic data that is captured and the complexity of the analyses that need to be carried out.

Traffic data is often obtained from traffic cameras or fixed readers (e.g., EZ-pass). However, only a small number of these devices are deployed in practice. GPS-tracked vehicles are another potential source of traffic information. A subset of these sensors are already being used by popular map services such as Google maps and Apple maps to provide real-time traffic information to users. However, their coverage is incomplete and limited to segments of major roads, and hinders the analysis as well as the accuracy of derived models.

While tracking all vehicles is not feasible, it is possible to track an important subset: taxis. Taxi fleets in many cities are equipped with GPS. Consider, for example, New York City (NYC): 13,000 taxis make, on average, 500,000 trips and carry over 1 million passengers every single day; totaling roughly 170 million trips per year. Given this high penetration rate of taxis in large cities, it is therefore reasonable to assume that the taxis can be used as probe vehicles, and taxi movement and travel times are representative of the overall traffic and provides a broad coverage of the city in space and time [ZHUK13]. Unfortunately, taxi data captured by the NYC Taxi & Limousine Commission contains neither the location of the taxis at regular intervals nor

Using Probabilistic Vector-Valued Functions to Explore Traffic Dynamics in Urban Environments
Jorge Poco, Harish Doraiswamy, Huy Vo, João Comba, Juliana Freire, Cláudio Silva
Computer Graphics Forum, Volume 34 (2015), Number 3

Visual Analysis of Bike-Sharing Systems


ARTICLE IN PRESS

Computers & Graphics

Special Section on SIBGRAPI 2016

Visual analysis of bike-sharing systems

Guilherme N. Oliveira, Jose L. Sotomayor, Rafael P. Torchelsen, Cláudio T. Silva, João L.D. Comba

ARTICLE INFO

ABSTRACT

1. Introduction

Visualizing the Dynamics of Bike-Sharing Systems

Guilherme Oliveira, Jose L. Sotomayor, Rafael Torchelsen, João Comba

Computers&Graphics, 2016 (available online)

Video

- Bike-Sharing Systems

Video

- Bike-Sharing Systems

Full

Empty

Video

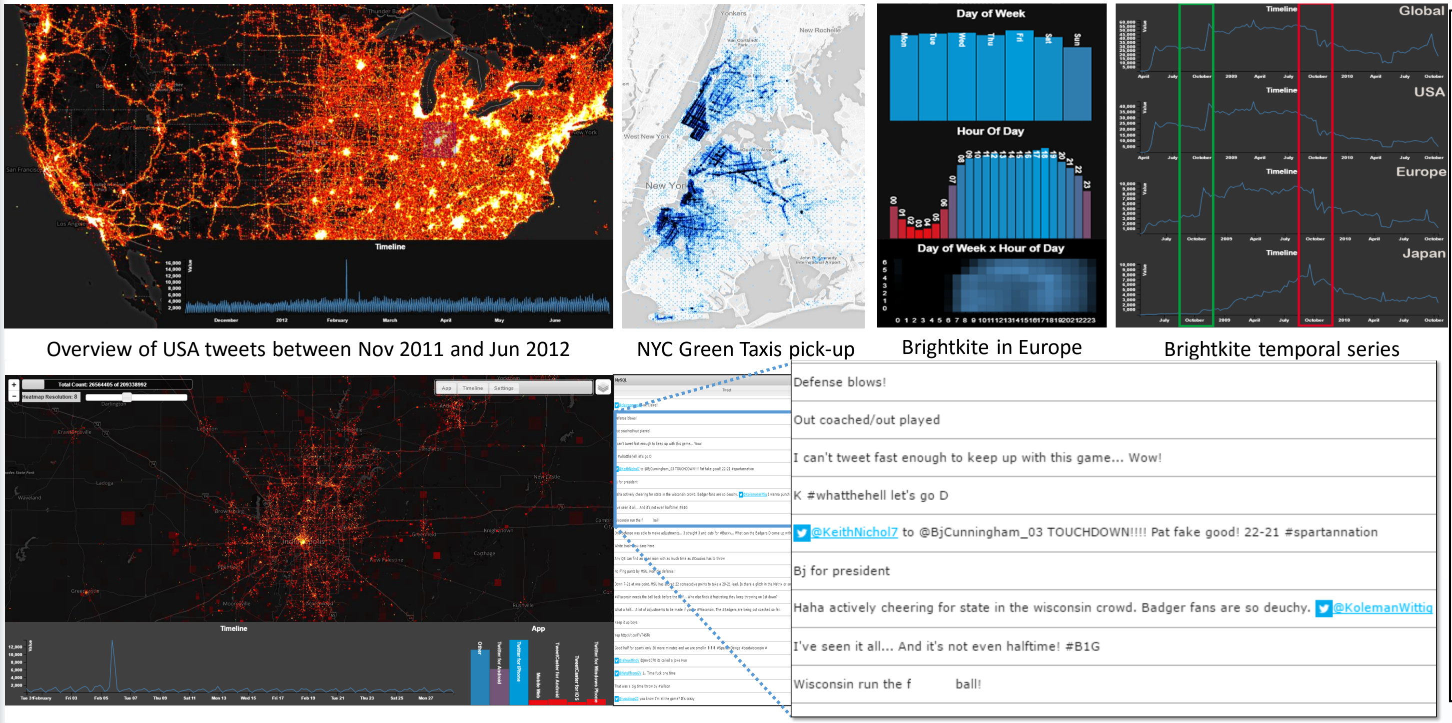
- Bike-Sharing Systems

Full

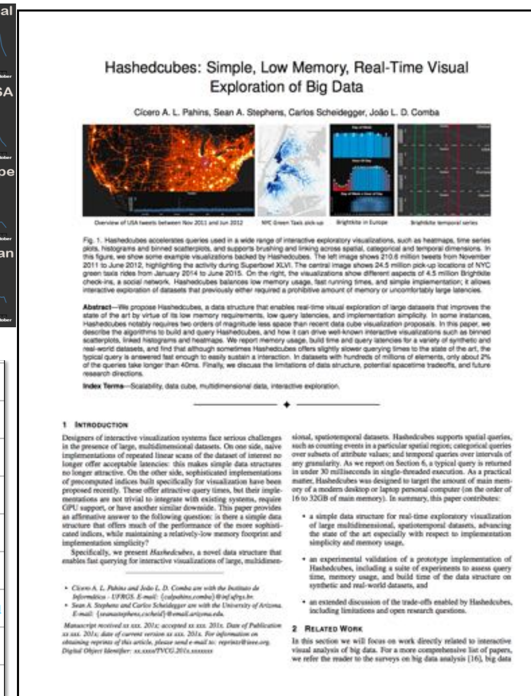
Empty

Rebalancing

Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data



Overview of USA tweets between Nov 2011 and Jun 2012 | NYC Green Taxis pick-up | Brightkite in Europe | Brightkite temporal series


Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data

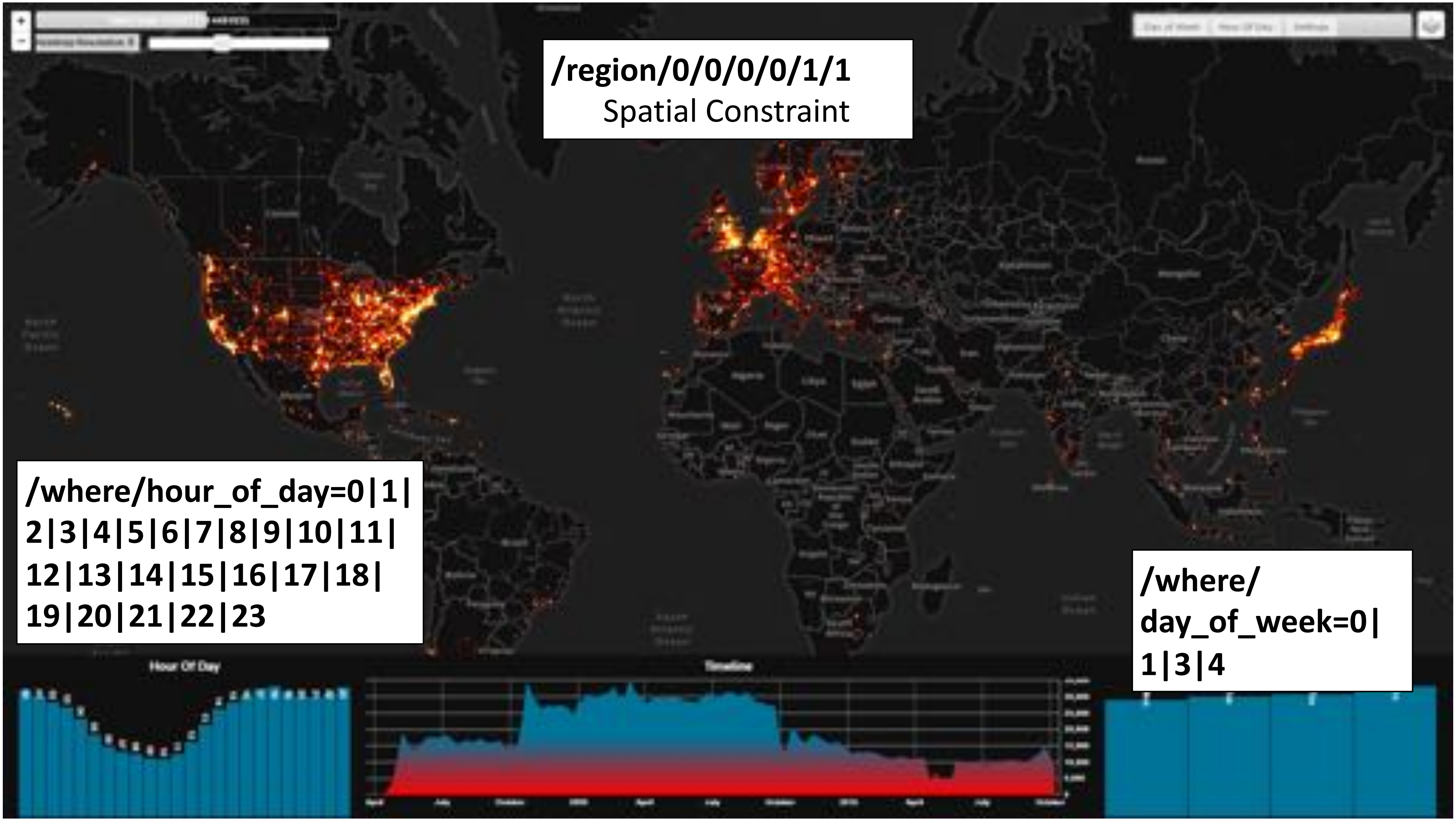
Cicero A. L. Pahins, Sean A. Stephens, Carlos Scheidegger, João L. D. Comba

Hashedcubes: Simple, Low Memory, Real-Time Visual Exploration of Big Data

Cicero Pahins, Sean , Carlos Scheidegger, João Comba

IEEE Transaction on Visualization 2016 (to appear), Proceedings of IEEE InfoVis 2016

Hashedcubes Query



Hashedcubes Query

- **/brightkite**
  - Dataset
- **/region**
  - Query type
- **/region/0/0/0/0/1/1**
  - Spatial Constraint
- **/where/day_of_week=0|1|3|4**
  - Categorical Constraint
- **/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23**
  - Categorical Constraint

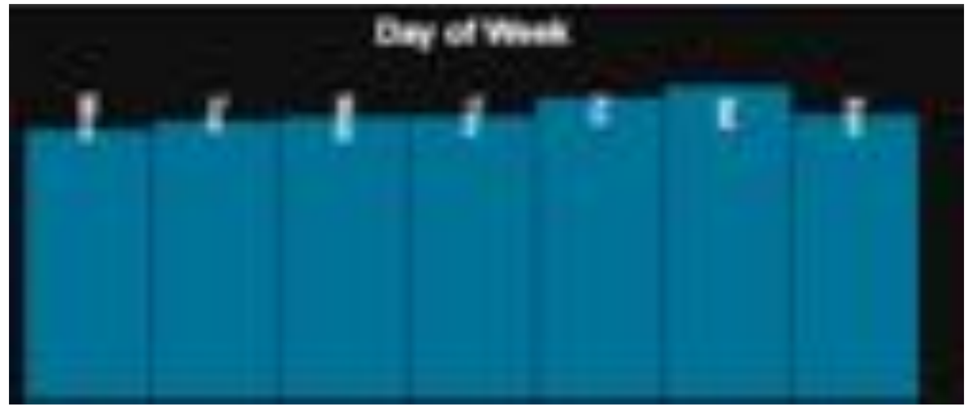
Heatmap

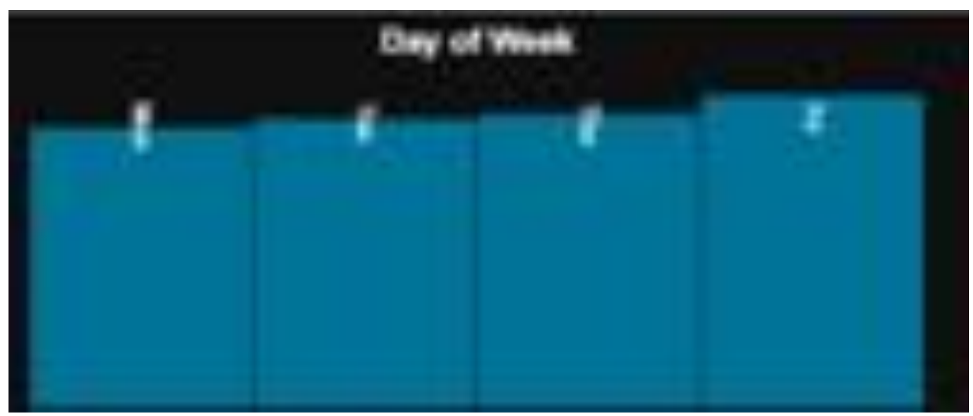
- **No constraint**

- **/region/0/0/0/0/1/1**
  - Spatial Constraint
- **/where/day_of_week=0|1|3|4**
  - Categorical Constraint
- **/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23**
  - Categorical Constraint

Histogram

- **No constraint**



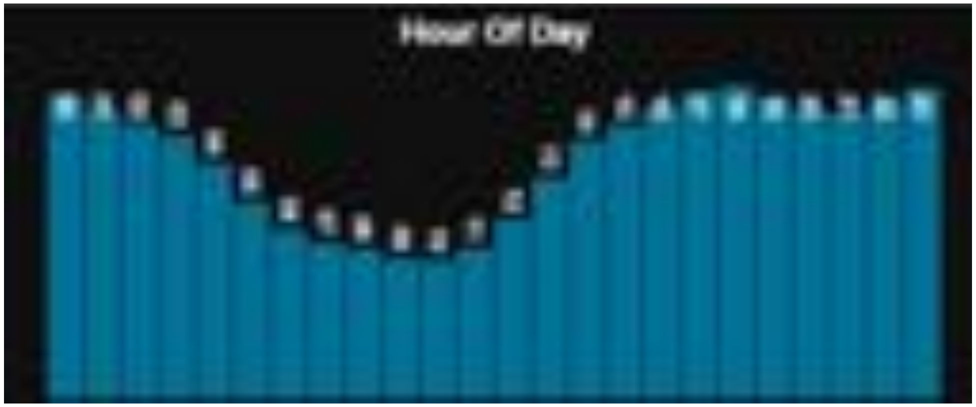


- **/region/0/0/0/0/1/1**
  - Spatial Constraint
- **/where/day_of_week=0|1|3|4**
  - Categorical Constraint
- **/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23**
  - Categorical Constraint

Histogram

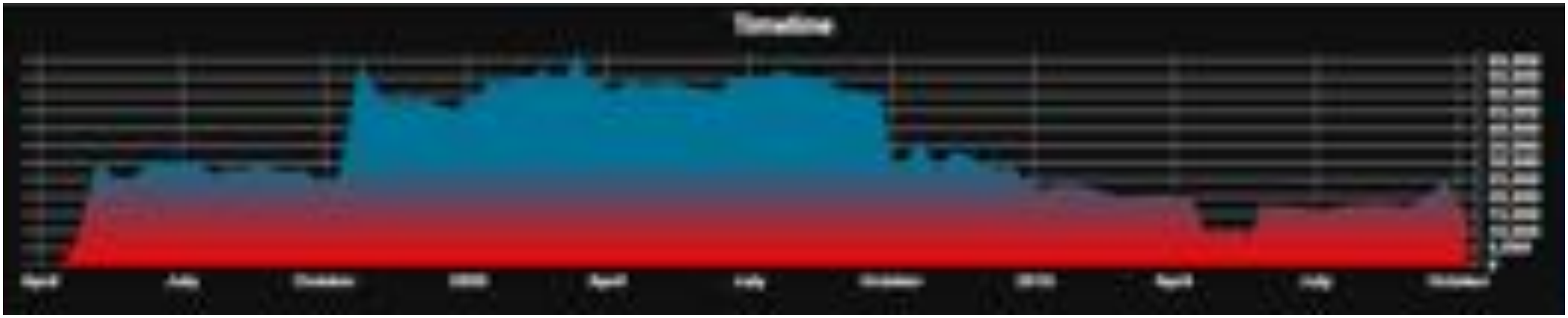
- **No constraint**

- **/region/0/0/0/0/1/1**
    - Spatial Constraint
- **/where/day_of_week=0|1|3|4**
    - Categorical Constraint
- **/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23**
    - Categorical Constraint

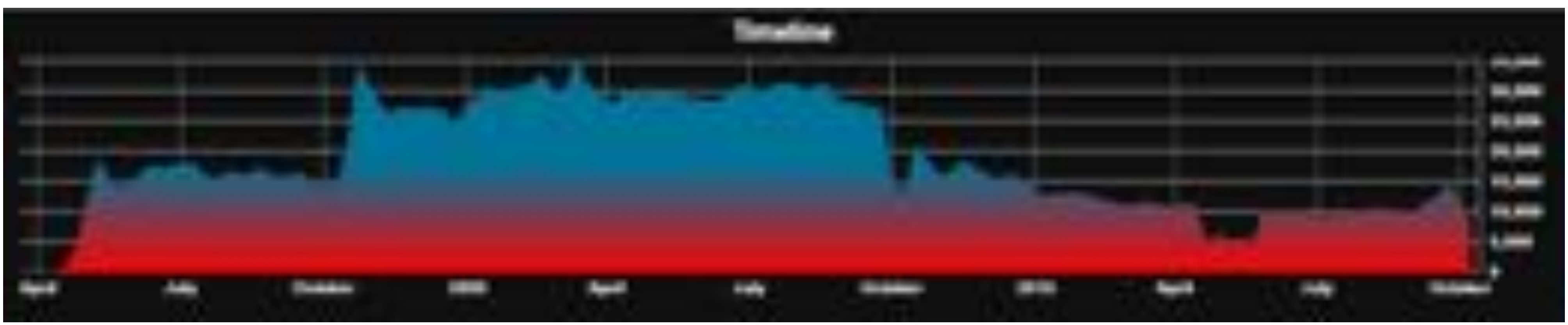


Line Chart

- **No constraint**



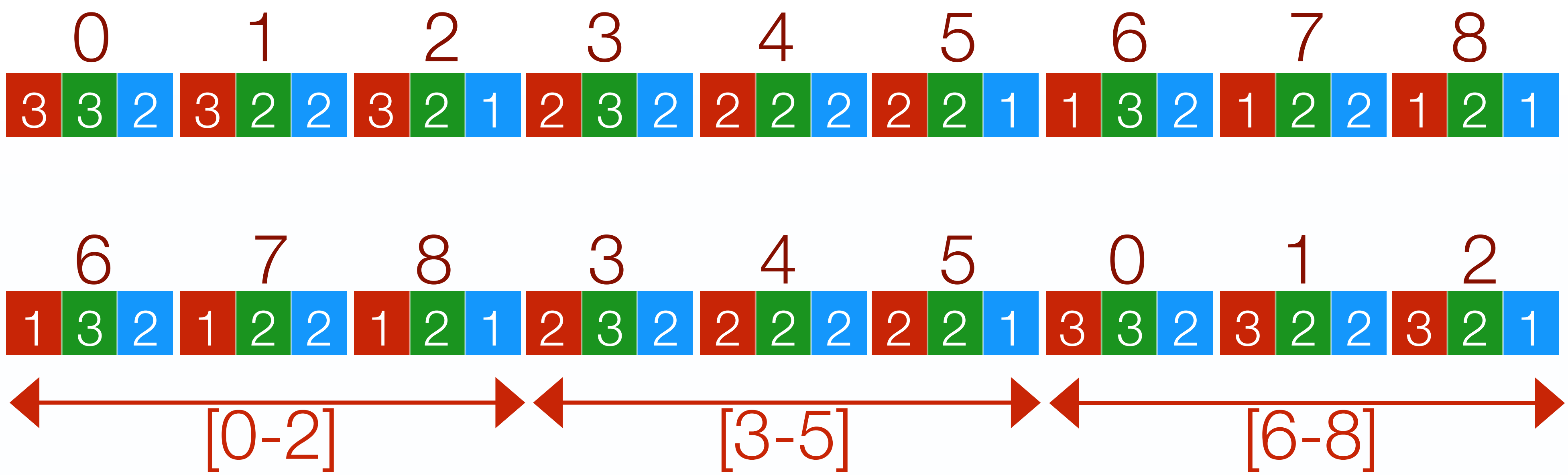
- **/region/0/0/0/0/1/1**
  - Spatial Constraint
- **/where/day_of_week=0|1|3|4**
  - Categorical Constraint
- **/where/hour_of_day=0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23**
  - Categorical Constraint



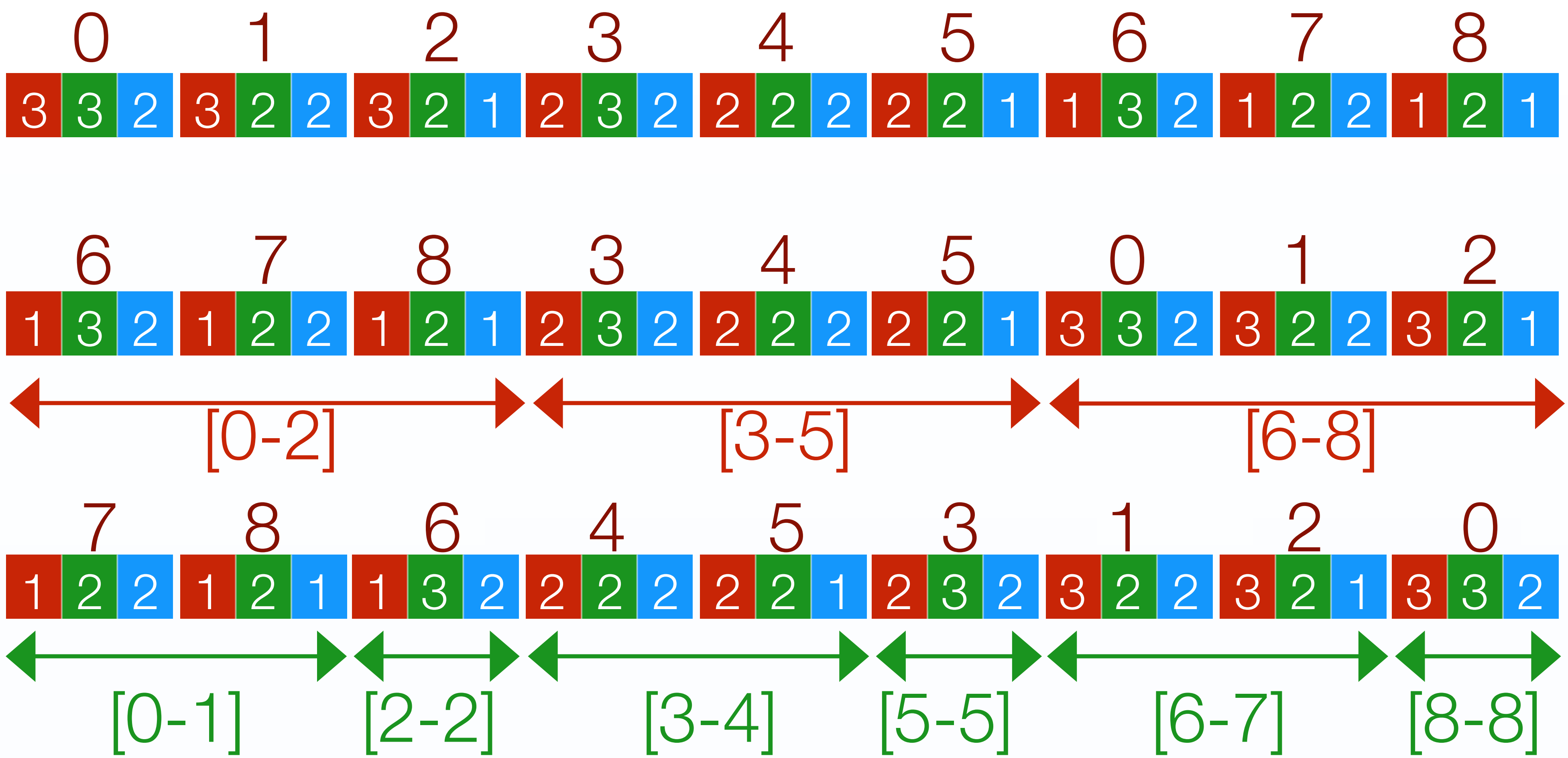
Pivot Concept and Hierarchy

0	1	2	3	4	5	6	7	8
3 3 2	3 2 2	3 2 1	2 3 2	2 2 2	2 2 1	1 3 2	1 2 2	1 2 1

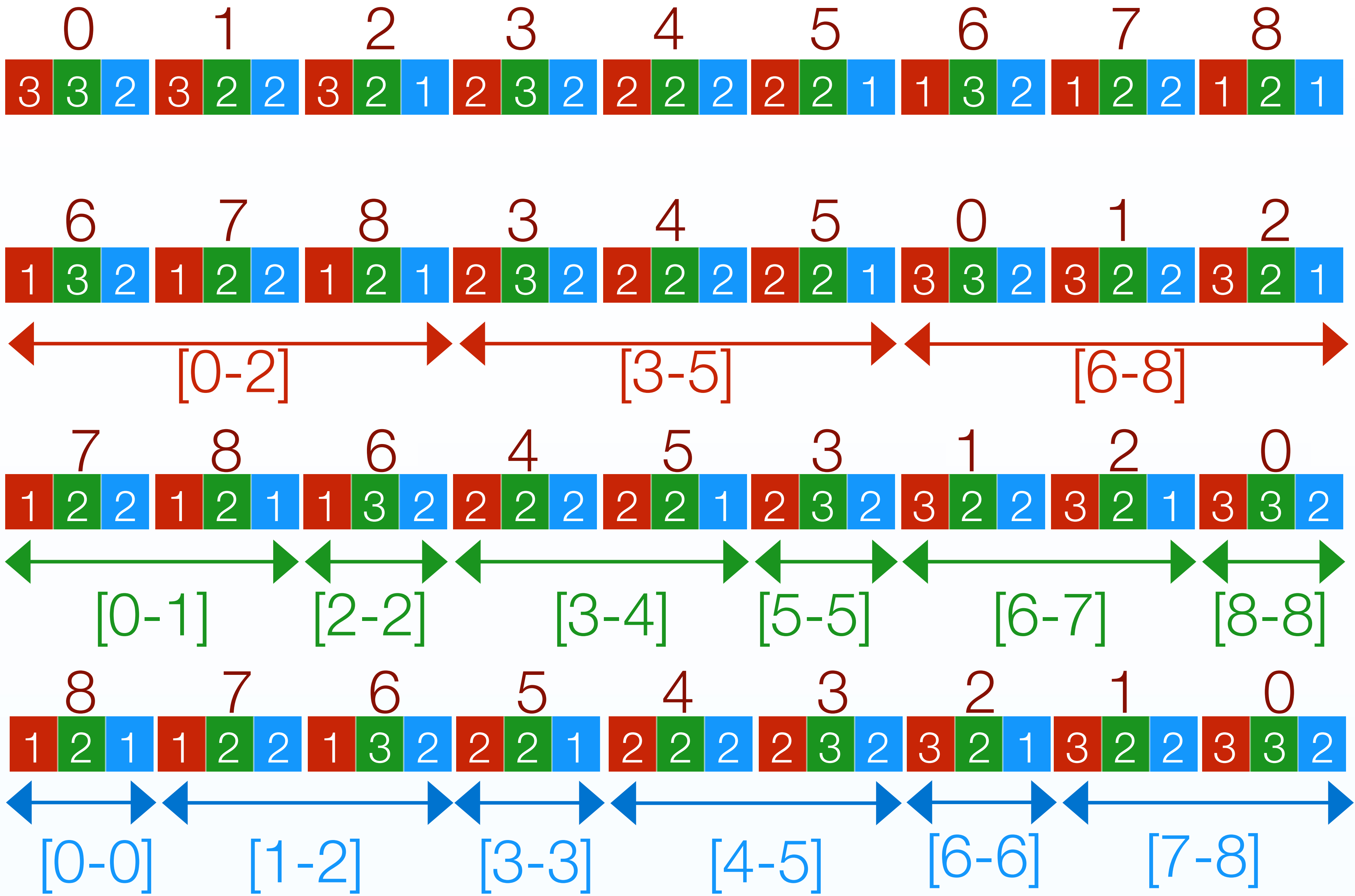
Pivot Concept and Hierarchy



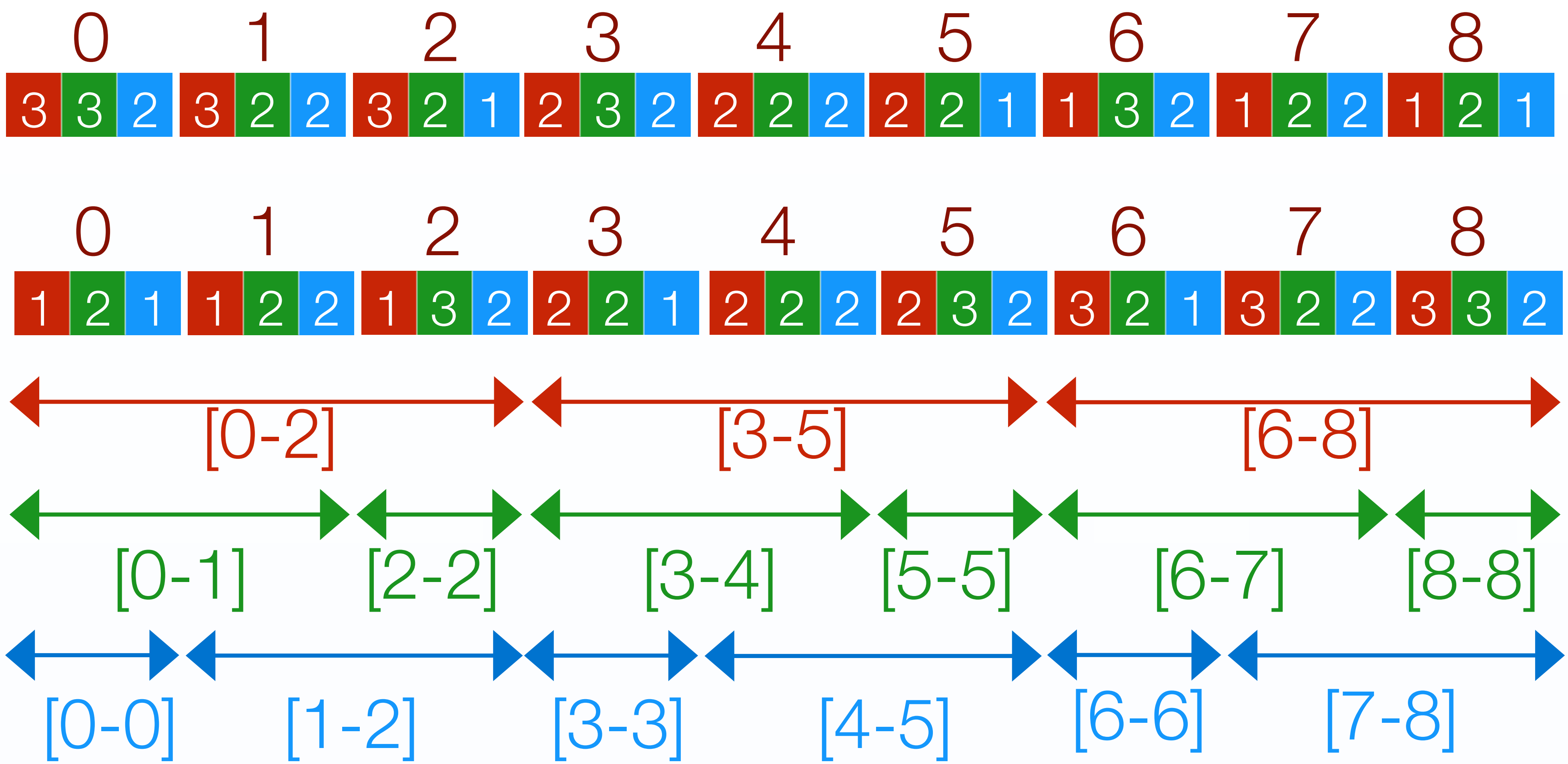
Pivot Concept and Hierarchy



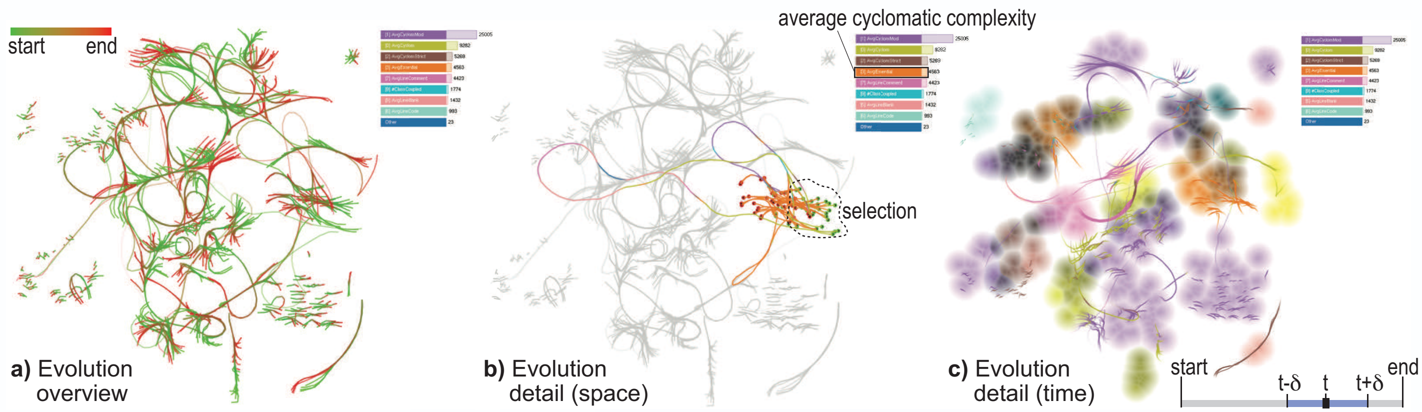
Pivot Concept and Hierarchy

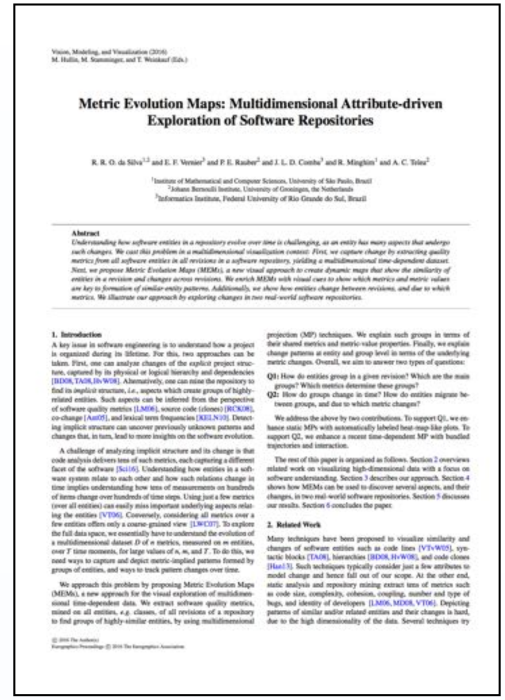


Pivot Concept and Hierarchy



Metric Evolution Maps: Multidimensional Attribute-driven Exploration of Software Repositories



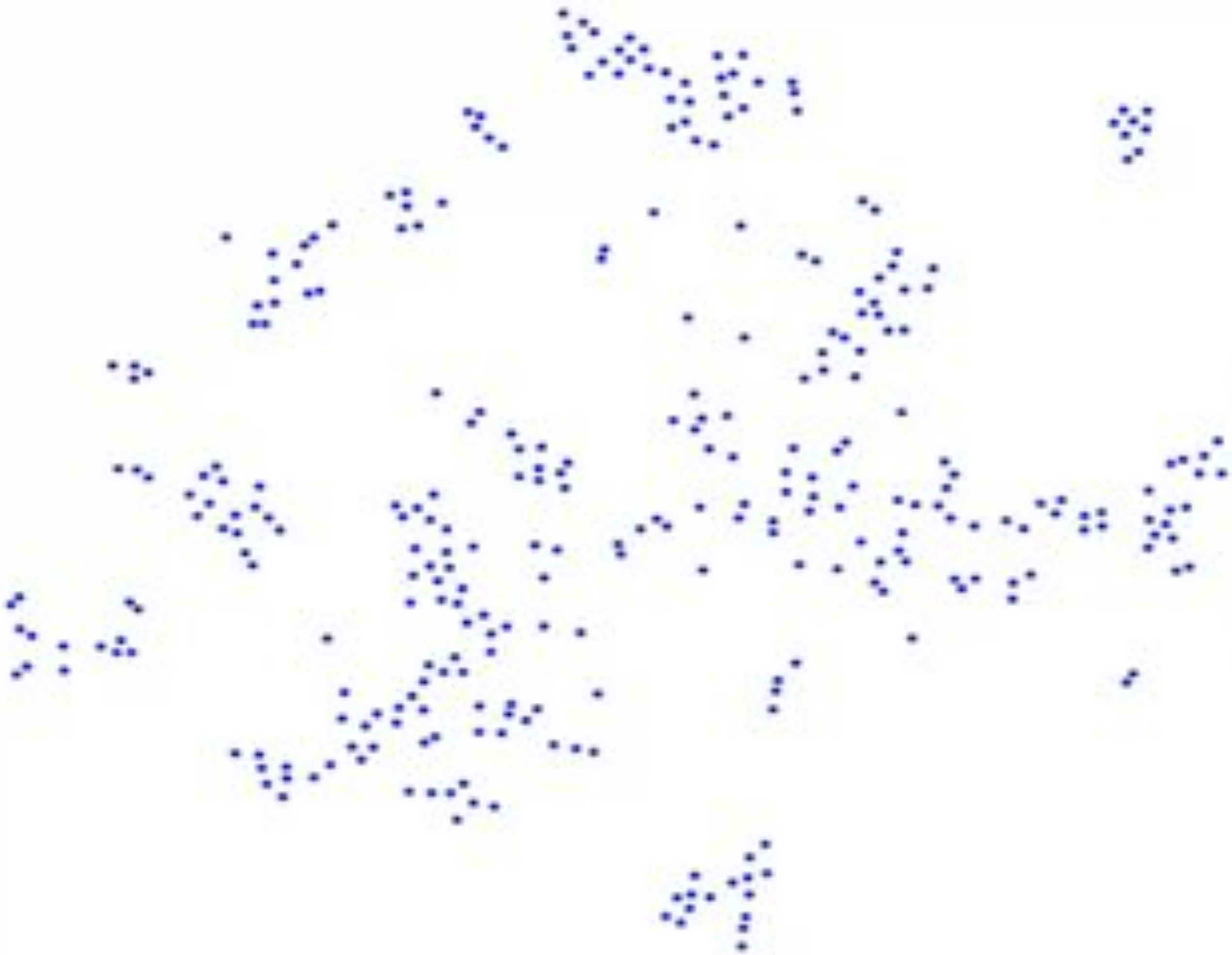


Metric Evolution Maps: Multidimensional Attribute-driven Exploration of Software Repositories

Renato Silva, Eduardo Vernier, P. Rauber, João Comba, Rosane Minghim, Alexandru Telea

VMV 2016

Video



Multidimensional Projection
Group classes with similar quality metrics in the same visual neighborhood

Conclusions

- Visual Data analysis has several interesting problems
  - Data Mining Algorithms
  - Visualization Techniques
  - High-Performance Computing
  - Spatial Data Structures and Geometric Algorithms
  - Machine Learning Algorithms
  - Mathematical and Statistical Analysis
  - Software implementation (web, prototypes)

Questions ?

e-mail: joao.comba@gmail.com

web: http://www.inf.ufrgs.br/~comba

Short Bio

I am an associate professor in the Computer Graphics Group at the "Instituto de Informática" of the Federal University at Rio Grande do Sul (UFRGS), Brazil. I received a Ph.D in Computer Science from Stanford University under the supervision of Leonidas J. Guibas . Before that, I received a masters degree in Computer Science from the Federal University of Rio de Janeiro (UFRJ), Brazil, working with Ronaldo Marinho Persiano. My bachelor's degree in Computer Science was given by the Federal University of Rio Grande do Sul , Brazil.

Publications by type

- Journals
- Conferences
- Books
- Book Chapters
- Patents
- Technical Reports
- Conferences/TRs (Portuguese)
- Ph.d. Thesis
- M.Sc. and B.Sc. Dissertations (Portuguese)

Publications by topic

- Visualization
- Geometric Algorithms/Data Structures
- High Performance and Parallel Computing
- Graphics Hardware and Games

Supervised Ph.D. and M.Sc.

- Ph.D. Advisor Co-advisor
- M.Sc. Advisor Co-Advisor

Publications by year

2016	2015		
2014	2013	2012	2011
2010	2009	2008	2007
2006	2005	2004	2003
2002	2001	2000	1999
1996	1994	1993	1991
1990	1989	1988	1987
All Years			

CV and other statistics

- Curriculum Lattes (CNPq-Brazil)
- Google Scholar
- DBLP

Funding

- CNPq

International Collaborations

- USA: CNPq-NSF
- Germany: CAPES-DAAD
- Netherlands: CAPES-BRANETEC

Professional Activities

- Conference Chair
- Program Committee
- Organizing Committee
- Co-Chair of CiSE VizCorner