From source data to data narratives:

accompanying users in the way to interactive data analysis

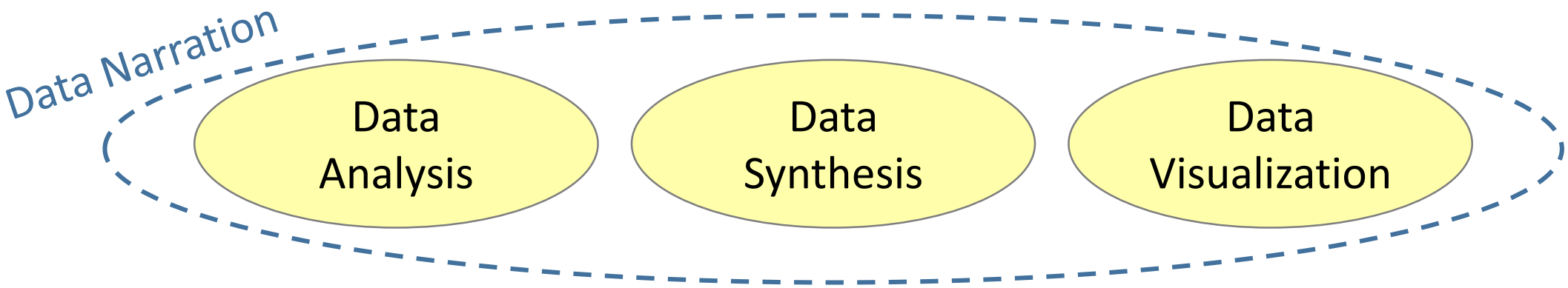
Verónika Peralta

University of Tours – France

ADBIS-TPDL-EDA joint conferences – August 2020

Data narration = Narrating with data visualization

- **The activity of producing narratives supported by facts extracted from data analysis, using interactive visualizations**



📖 J. Hullman, S. Drucker, N. Riche, B. Lee, D. Fisher, E. Adar: "A Deeper Understanding of Sequence in Narrative Visualization", TVCG 19:12, 2013.

📖 S. Carpendale, N. Diakopoulos, N. Riche, C. Hurter: "Data-Driven Storytelling", Dagstuhl Reports 6:2, 2016.

Outline

- **What is a data narrative?**
- **A panorama of tasks and tools for supporting data narration**

  Focus on:
  - Supporting intentional querying
  - Searching interesting findings
- **Open challenges**

What is a data narrative?

A **data narrative** is a **structured** composition of **messages** that

(a) convey **findings** over the **data**, and,

(b) are typically delivered via **visual means** in order to facilitate their reception by an intended audience.

Based on definitions of narrative and storytelling:

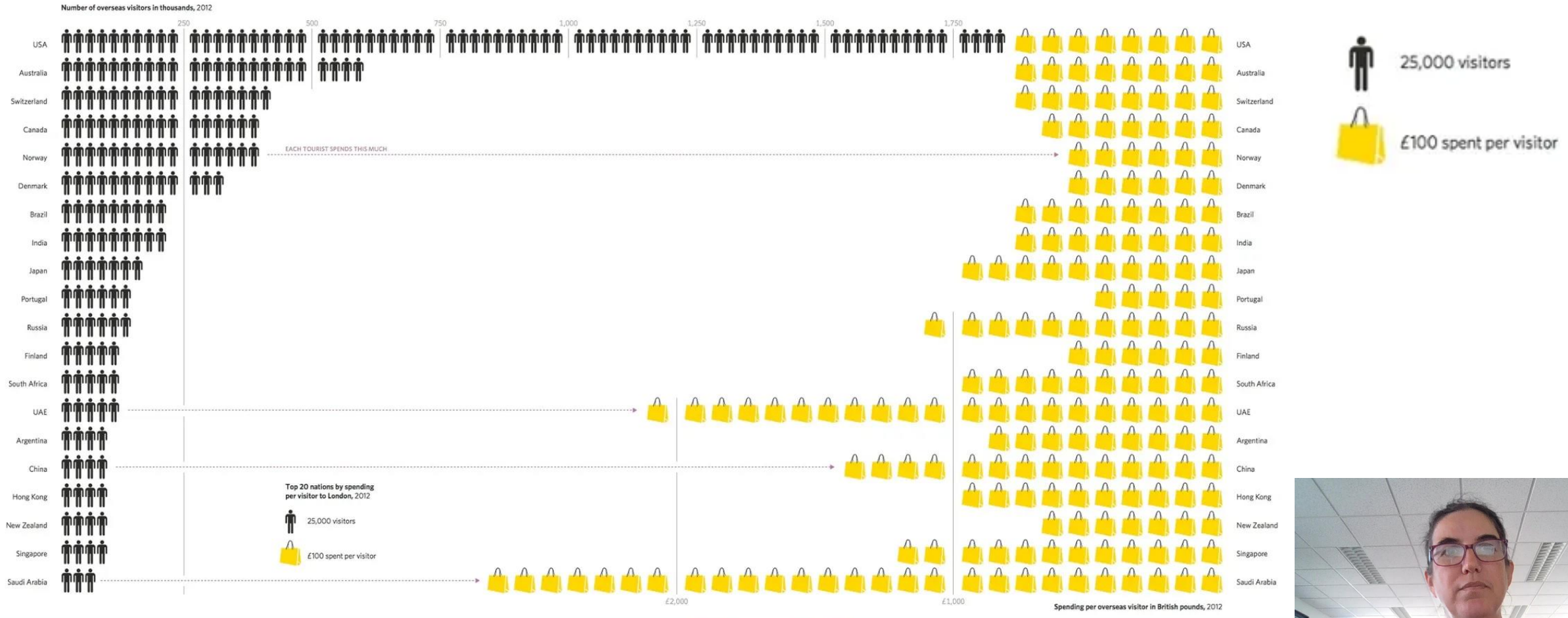
- D. Elson: “Modeling narrative discourse”, Ph.D. thesis, Columbia University, 2012.
- S. Chatman: “Story and Discourse: Narrative Structure in Fiction and Film”, Cornell paperbacks, 1980.
- S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. Nguyen, C. Turkay: “Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling”, TVCG 2018.
- F. El Outa, M. Francia, P. Marcel, V. Peralta, P. Vassiliadis: “A conceptual model of data narrative for exploratory data analysis”, ER 2020.

Examples

Top 20 nations by spending by visitor to London in 2012

Source: The Guardian

URL: https://www.theguardian.com/cities/gallery/2014/oct/28/london-life-mapped-data-visualisation-graphics#img-6



Examples

Covid-19 situation update worldwide

Source: European Centre for Disease Prevention and Control

URL: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

European Centre for Disease Prevention and Control

An agency of the European Union

All topics: A to Z | News & events | Publications & data | Tools | About us

Home > All topics: A to Z > Coronavirus > Threats and outbreaks > COVID-19 > Situation updates on COVID-19 > Situation update worldwide

- Situation updates on COVID-19
- Situation update for the EU/EEA and the UK
- **Situation update worldwide**
- 14-day incidence of reported COVID-19
- COVID-19 country overviews
- Weekly surveillance report on COVID-19
- Data collection

COVID-19 situation update worldwide, as of 28 June 2020

Epidemiological update

The data presented on this page has been collected between 6:00 and 10:00 CET

Disclaimer: *National updates are published at different times and in different time zones. This, and the time ECDC needs to process these data, may lead to discrepancies between the national numbers and the numbers published by ECDC. Users are advised to use all data with caution and awareness of their limitations. Data are subject to retrospective corrections; corrected datasets are released as soon as processing of updated national data has been completed.*

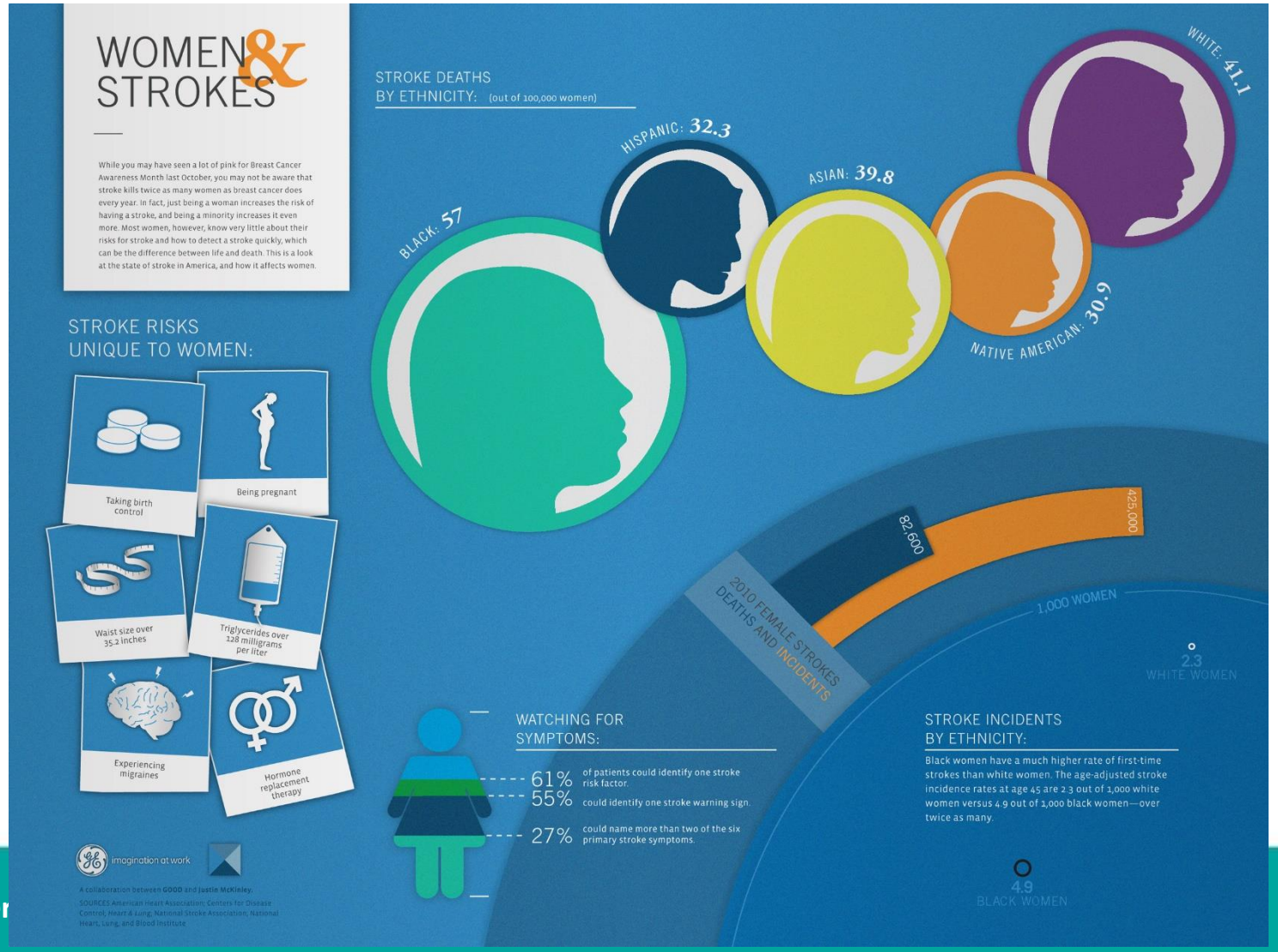
Download today's data | How is the data collected?

Examples

Stroke, a Silent Killer of Women, Facts About Women and Strokes

Source: GOOD

URL: https://www.good.is/infographics/facts-about-women-and-strokes

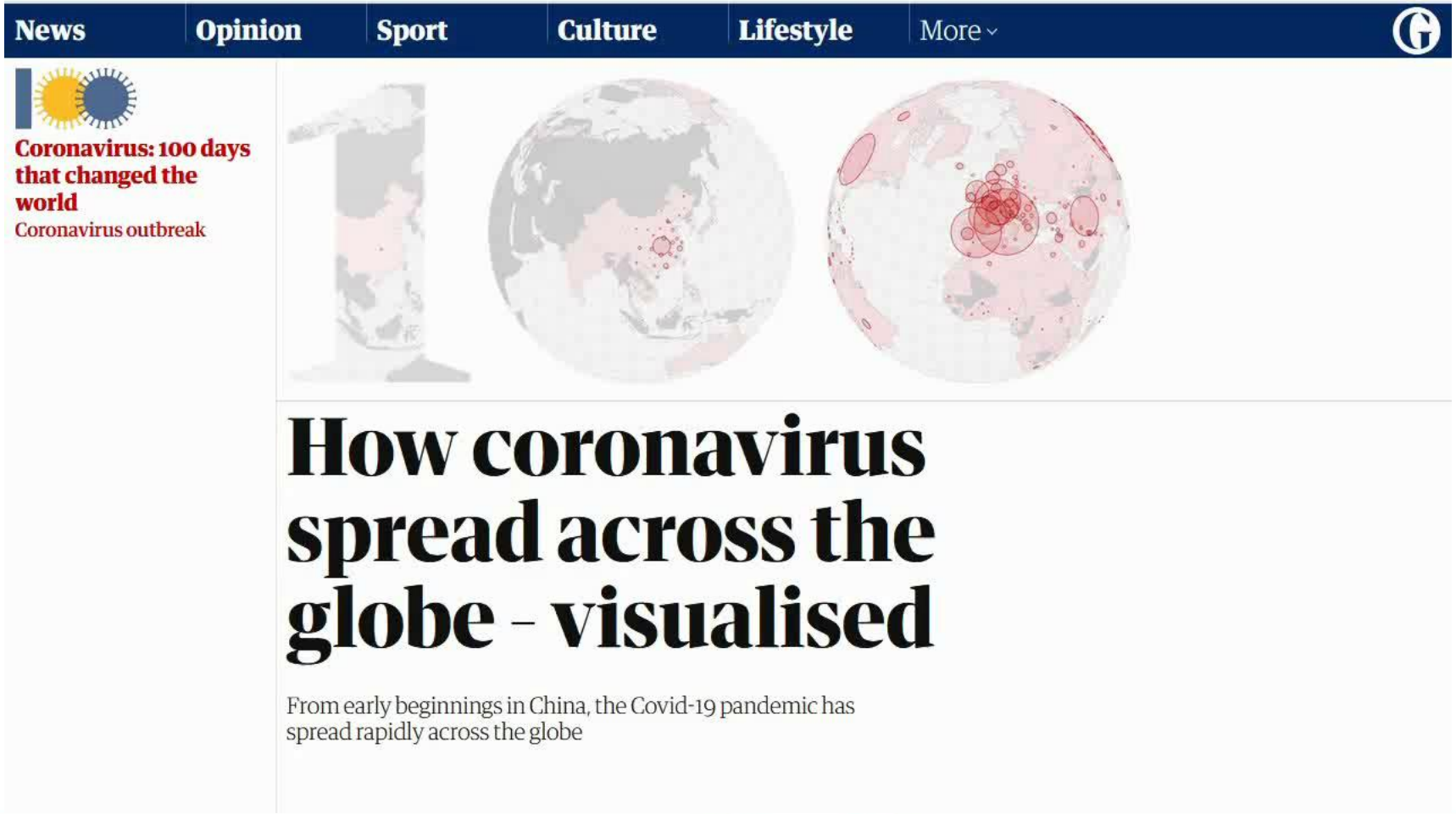


Examples

Coronavirus: 100 days that changed the world

Source: The Guardian

URL: https://www.theguardian.com/world/ng-interactive/2020/apr/09/how-coronavirus-spread-across-the-globe-visualised



Examples

How BuzzFeed News Used Betting Data To Investigate Match-Fixing In Tennis

Source: BuzzFeed News

URL: https://www.buzzfeednews.com/article/johntemplon/how-we-used-data-to-investigate-match-fixing-in-tennis#.vyKWjpWkn

The Tennis Racket

Betting worth billions. Elite players. Vi
And suspicious matches at Wimbledo
tennis authorities have kept secret for

Heidi Blake
UK Investigations Editor, UK

Posted on January 17, 2016, at 4:58 p.m. E

Tweet Share Copy

Secret files exposing evidence of w
of world tennis can today be reveale

The sport's governing bodies have b
players – all of whom have ranked i
and more than half of them will beg

It has been seven years since world t
evidence about a network of players
including Wimbledon following a la
allowed to continue playing.

How BuzzFeed News Used Betting Data To Investigate Match-Fixing In Tennis

With GIFs.

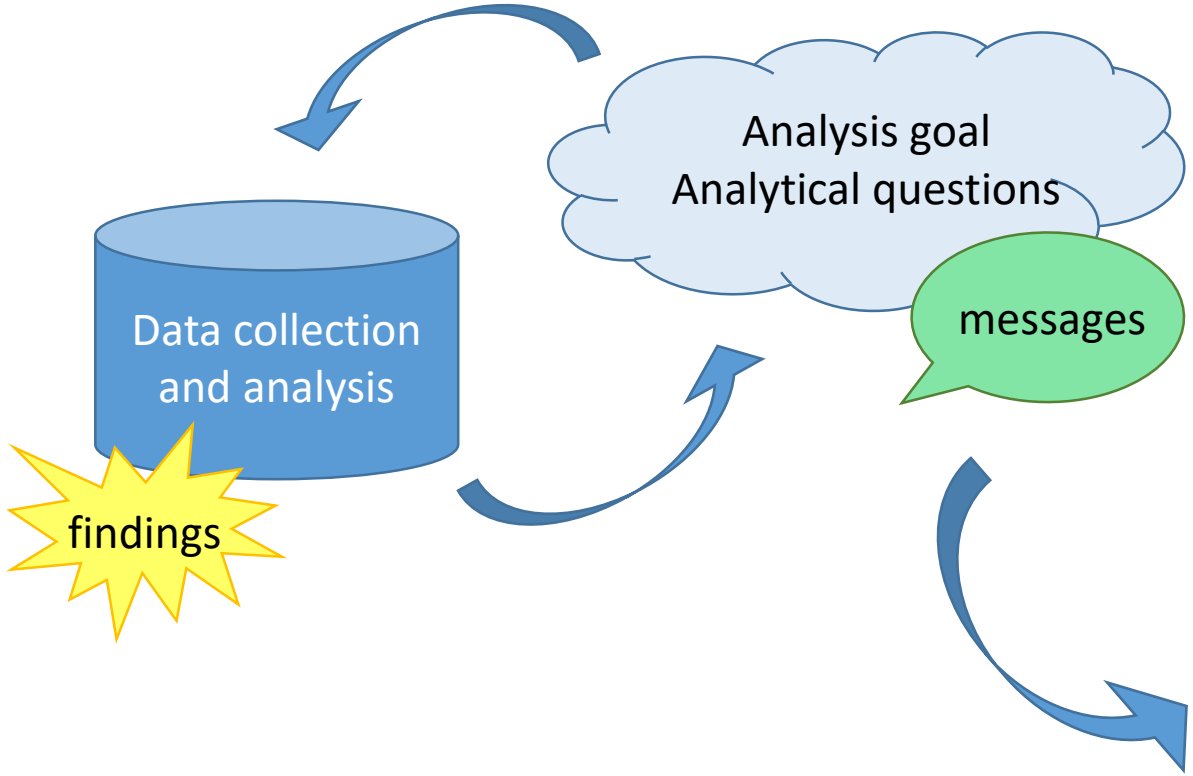
John Templon
BuzzFeed News Reporter

Posted on January 17, 2016, at 5:02 p.m. ET

Tweet Share Copy

I'm John Templon, an investigative data reporter for BuzzFeed News. I spent the past 15 months analyzing tennis betting data to see if I could figure out whether players were fixing matches.

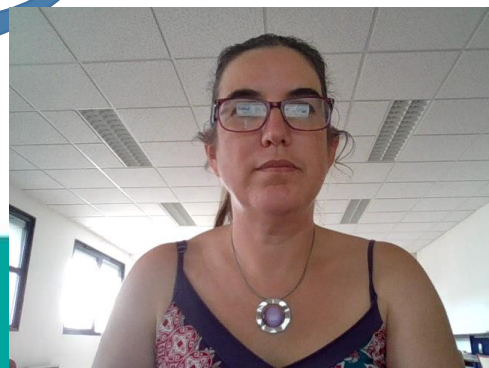
Building data narratives



©elegantthemes.com

©adn-co.news

Structuring



📖 S. Chatman: “Story and Discourse: Narrative Structure in Fiction and Film”, Cornell paperbacks, 1980.
📖 F. El Outa, M. Francia, P. Marcel, V. Peralta, P. Vassiliadis: “A conceptual model of data narrative for exploratory data analysis”, ER 2020.

A model in 4 layers

Story

What is told

Factual

Collecting relevant findings via interactive data exploration

Intentional

Composing messages, based on findings, according to an analysis goal, and identifying characters and measures

Discourse

How is it told

Structural

Organizing messages and structuring the discourse in units (episodes, acts)

Presentational

Providing presentational means to the structured discourse

F. El Outa, M. Francia, P. Marcel, V. Peralta, P. Vassiliadis: “A conceptual model of data narrative for exploratory data analysis”, ER 2020.

A model in 4 layers

Story

Factual

Collecting relevant findings via interactive data exploration

Intentional

Devising messages, based on findings, according to an analysis goal, and identifying characters and measures

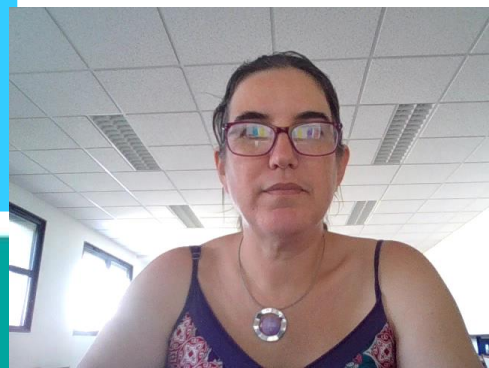
Discourse

Structural

Organizing messages and structuring the discourse in units (episodes, acts)

Presentational

Providing presentational means to the structured discourse

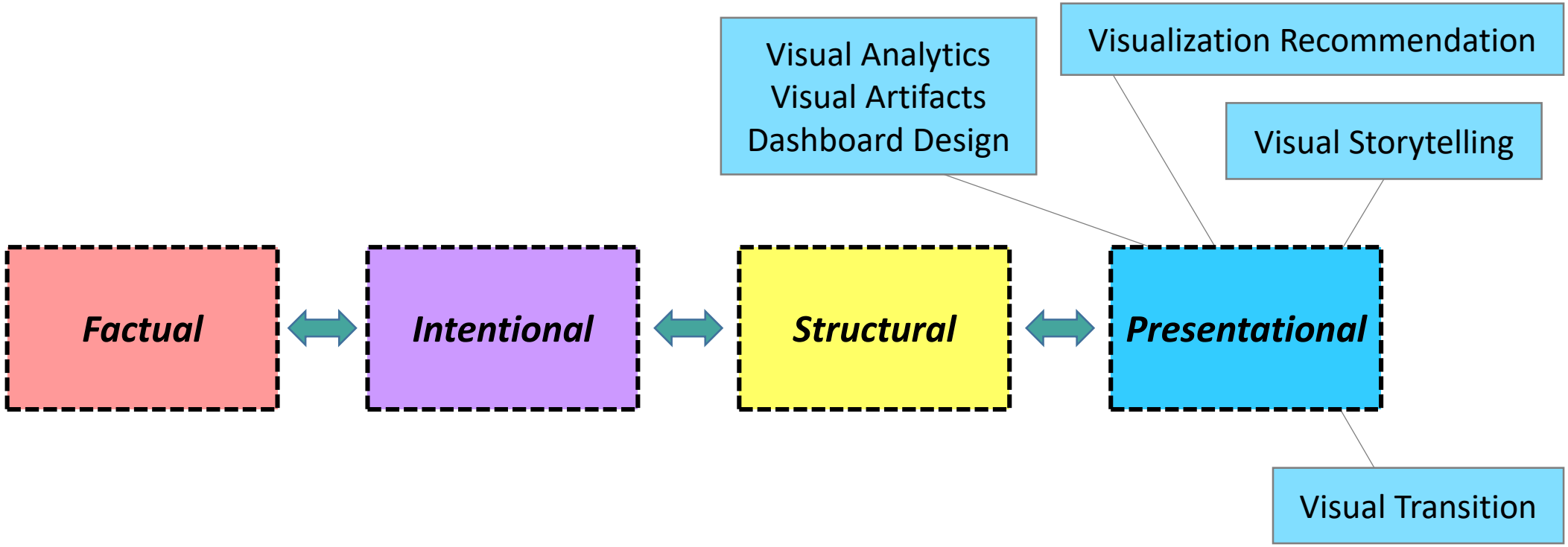


Outline

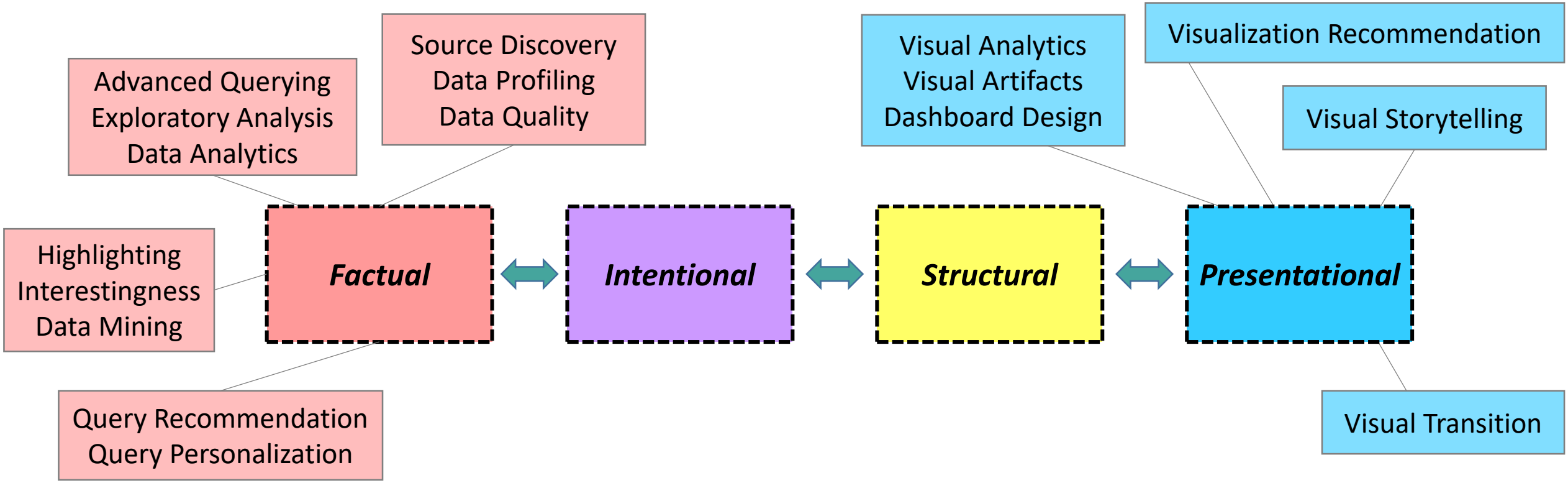
- What is a data narrative?
- **A panorama of tasks and tools for supporting data narration**

  Focus on:
  - Supporting intentional querying
  - Searching interesting findings
- **Open challenges**

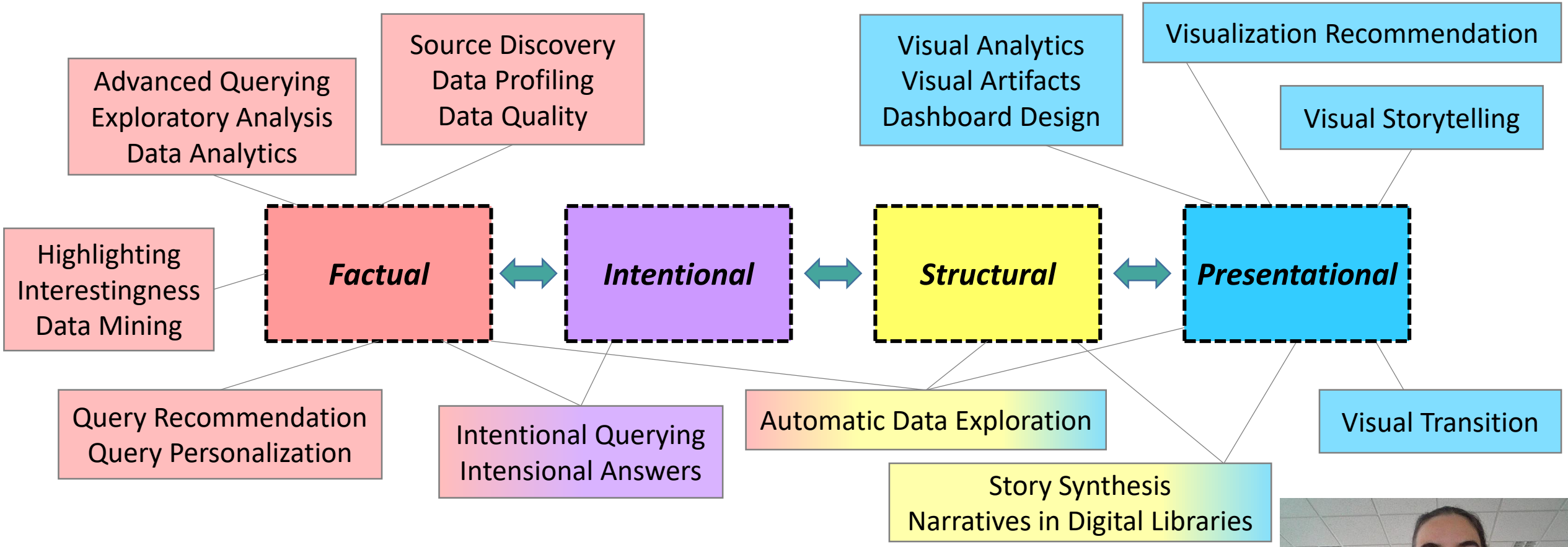
Tasks for supporting data narratives

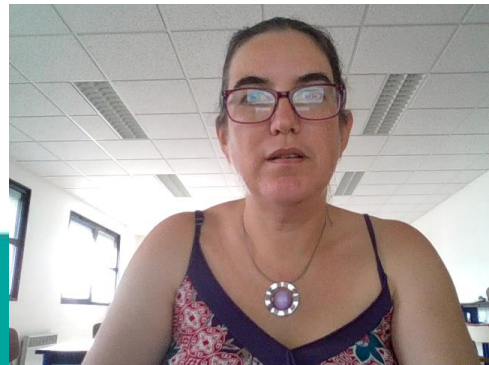


Tasks for supporting data narratives



Tasks for supporting data narratives





Intentional analytics model

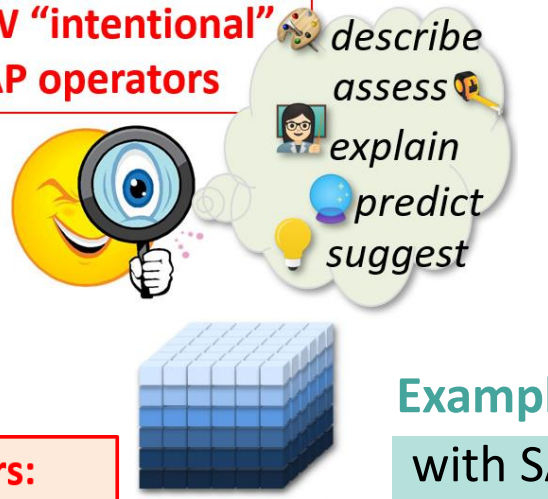
OLAP III: Intentional querying, Intelligent results, Interesting highlights

📖 P. Vassiliadis, P. Marcel, S. Rizzi. “Beyond Roll-Up's and Drill-Down's: An Intentional Analytics Model to Reinvent OLAP”. Inf. Syst. 85, 2019.

📖 A. Chédin, M. Francia, P. Marcel, V. Peralta, S. Rizzi: “The Tell-Tale Cube”, ADBIS 2020.

- **Query operators are user intensions that are automatically translated to queries**
- **Answers are complemented with KDD models and highlights**

NEW “intentional” OLAP operators

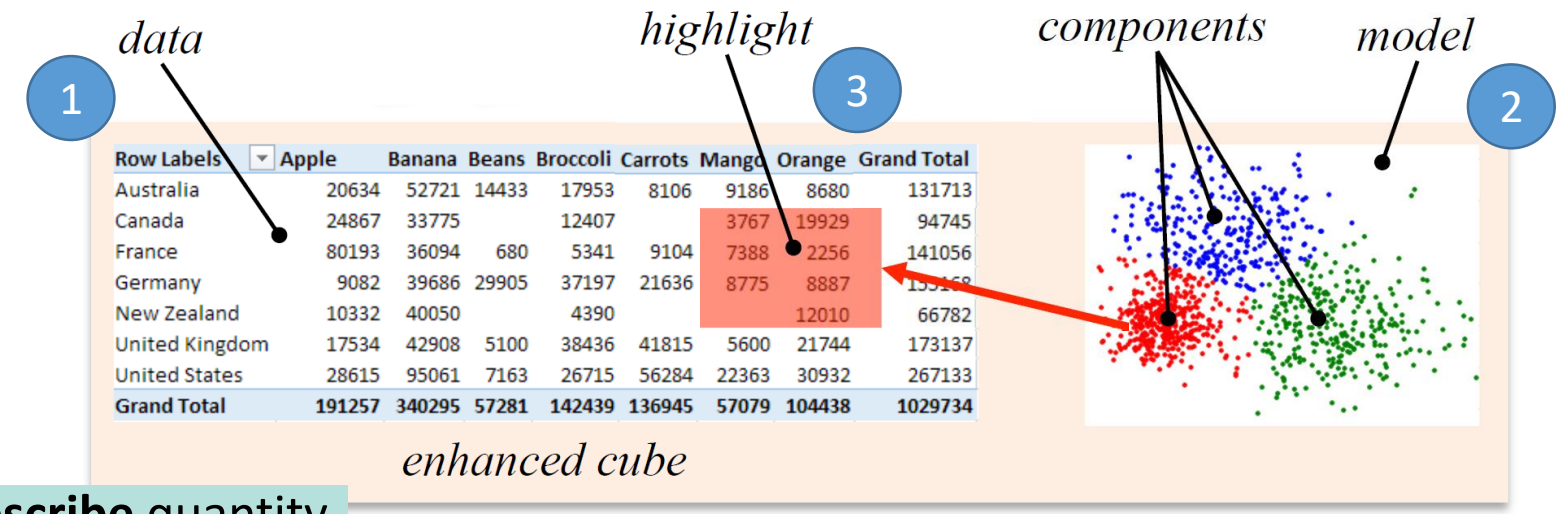


NEW answers:

Data

Models

Highlights



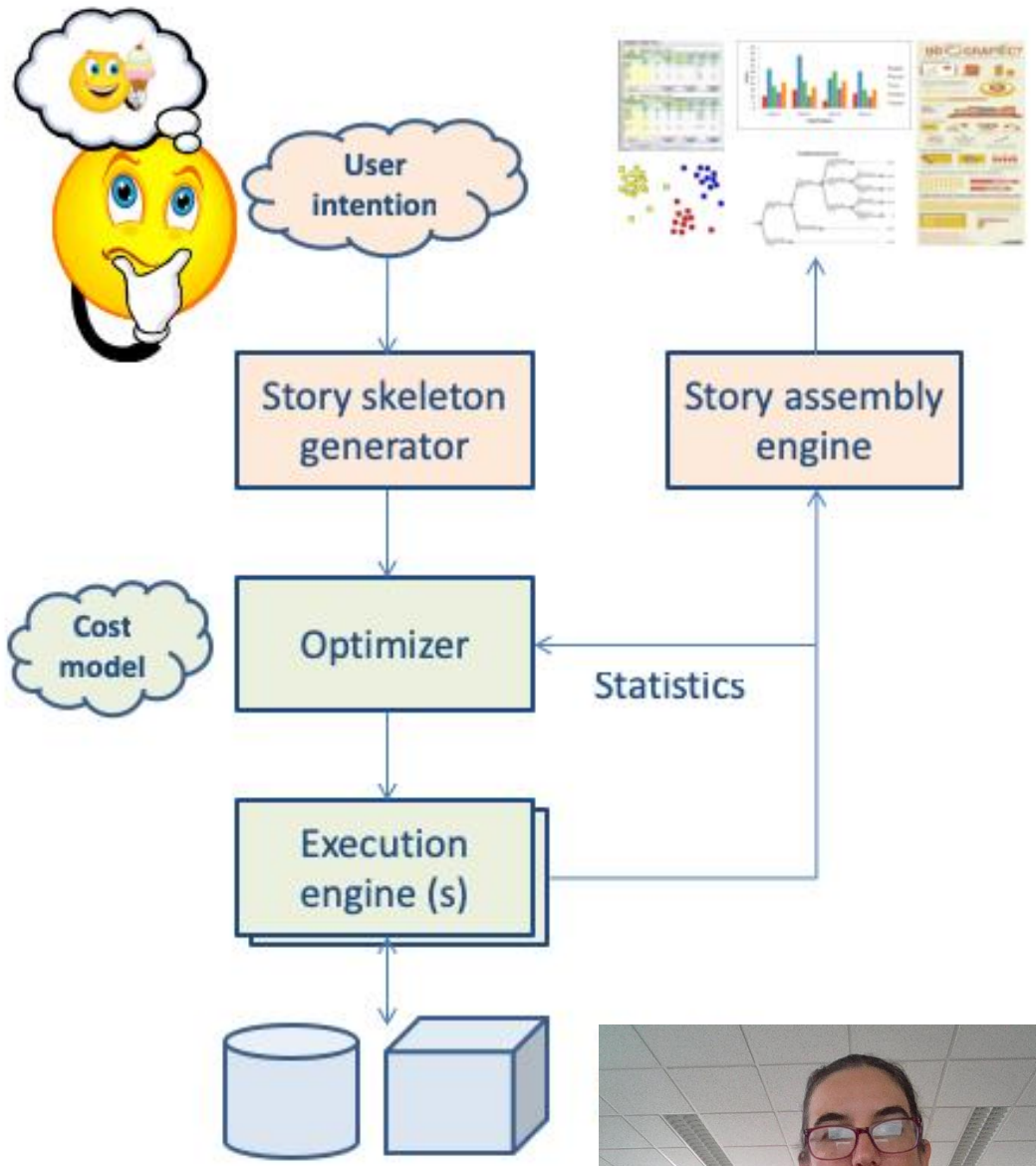
Row Labels	Apple	Banana	Beans	Broccoli	Carrots	Mango	Orange	Grand Total
Australia	20634	52721	14433	17953	8106	9186	8680	131713
Canada	24867	33775		12407		3767	19929	94745
France	80193	36094	680	5341	9104	7388	2256	141056
Germany	9082	39686	29905	37197	21636	8775	8887	[illegible]
New Zealand	10332	40050		4390			12010	66782
United Kingdom	17534	42908	5100	38436	41815	5600	21744	173137
United States	28615	95061	7163	26715	56284	22363	30932	267133
Grand Total	**191257**	**340295**	**57281**	**142439**	**136945**	**57079**	**104438**	**1029734**

Example:

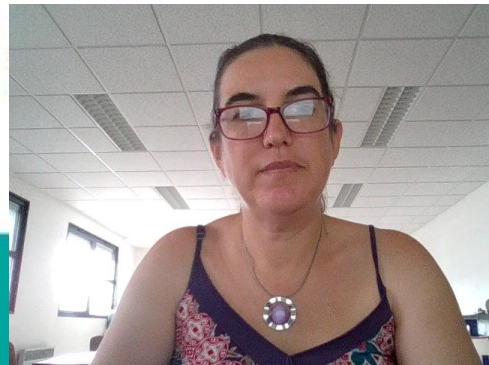
with SALES **describe** quantity
for month = ‘2020-04’
by type and country
using clustering size 3

Challenges

- **Which queries to execute?**
  - Many possibilities (ex. explain operator)
  - It is an optimization problem
- **Which models to execute?**
  - Automatic tuning
- **Which highlights to select?**
  - Interestingness w.r.t. intention
  - Select the most effective visualization
  - Put data and highlights to work together



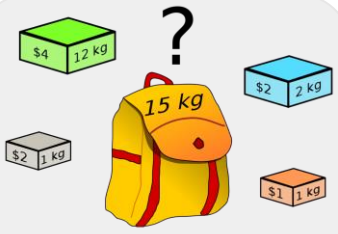
📖 P. Marcel, N. Labroche, P. Vassiliadis: “Towards a benefit-based optimizer for Interactive Data Analysis”, DOLAP 2019.



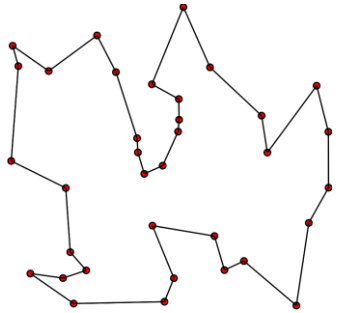
Which queries to execute?

The optimization problem

- **Given**
  - a set of n queries
  - a function estimating an *interestingness* score for a query
  - a function estimating the *execution cost* of a query
  - a function estimating a *cognitive distance* between queries
- **Find a sequence of m ≤ n queries (without repetition) s.t.:**
  - it maximizes the overall interestingness score
  - the sum of the costs does not exceed a time budget
  - it minimizes the overall cognitive distance between the queries



Knapsack Problem

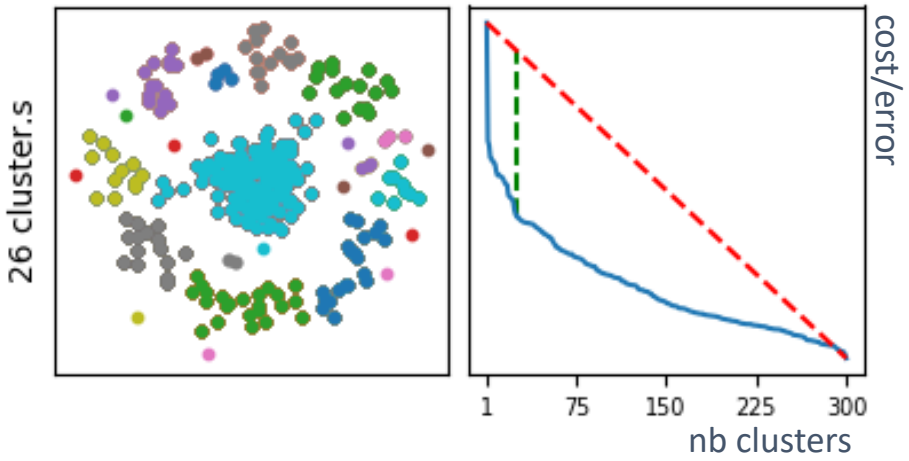


Traveling Salesman Problem

📖 A. Chanson, B. Crulis, N. Labroche, P. Marcel, V. Peralta, S. Rizzi, P. Vassiliadis: "The Traveling Analyst Problem", DOLAP 2020.

Which models to execute?

- **Choosing models:**
  - Examples: top-k, bottom-k, skyline, outliers, clustering
- **Tuning models:**
  - Auto-learning, meta-learning
  - Example: setting clustering size
    - The best separation of clusters can be set to the knee of the evaluation graph of the clustering algorithm
    - Kneedle algorithm



📖 M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter: "Efficient and Robust Automated Machine Learning", Advances in Neural Information Processing Systems 28, 2015.

📖 A. Chédin, M. Francia, P. Marcel, V. Peralta, S. Rizzi: "The Tell-Tale Cube", ADBIS 2020.

📖 V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan: "Finding a kneedle in a haystack: Detecting knee points in system behavior", ICDCS 2011.

Interestingness

To what extent is a piece of info …

… providing an answer to the user's ultimate information goal?

Relevance

Novelty

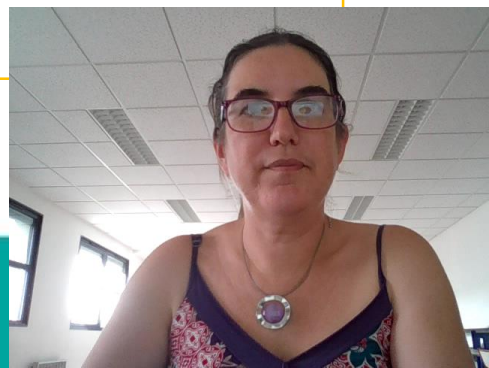
… presenting information previously unknown to the user?

… disconfirming or challenging previous beliefs?

Surprise

Peculiarity

… different from its peers in the same context?



📖 P. Marcel, V. Peralta, P. Vassiliadis: "A framework for learning cell interestingness from cube explorations", ADBIS 2019.

Interestingness

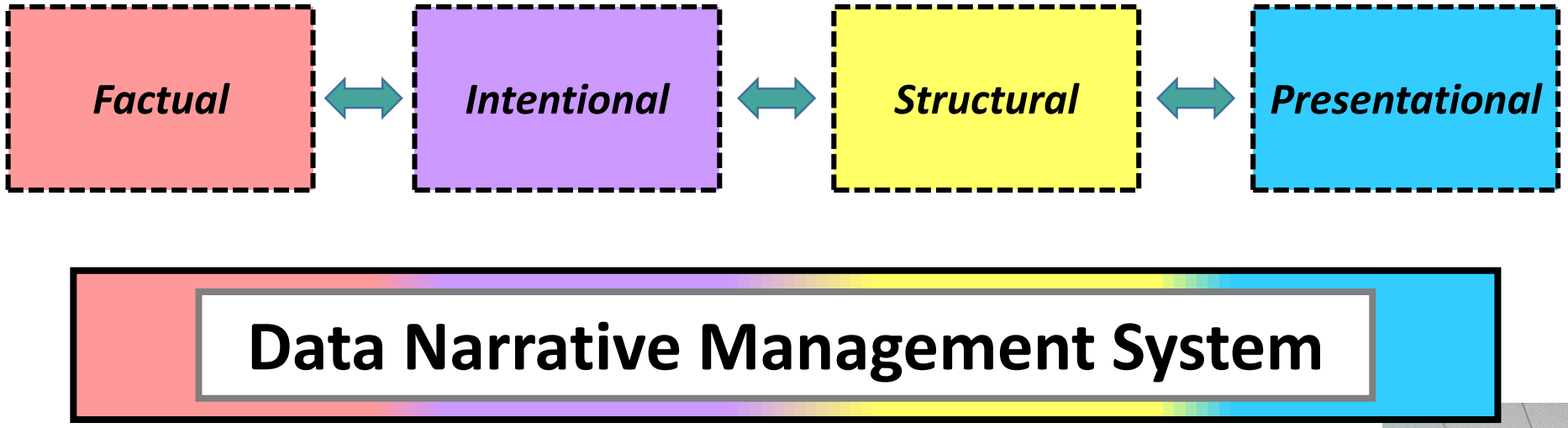
- **A multitude of measures to be reused.** Some surveys:
  - 🕮 L. Geng, H. Hamilton: “Interestingness measures for data mining: A survey”, ACM Comput. Surv. 38(3), 2006.
  - 🕮 K. McGarry: “A survey of interestingness measures for knowledge discovery”, The knowledge engineering review 20(1), 2005.
  - 🕮 M. Kaminskas, D. Bridge: “Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems”, TiiS 7(1), 2017.
  - 🕮 P. Marcel, V. Peralta, P. Vassiliadis: “A framework for learning cell interestingness from cube explorations”, ADBIS 2019.
- **Dynamically selecting the appropriate measure:**
  - Predicting what is interesting
    - 🕮 T. Milo, C. Ozeri, and A. Somech. “Predicting ‘what is interesting’ by mining interactive-data-analysis session logs”, EDBT 2019.
  - ML-based models for learning users’ interest
    - 🕮 E. Huang, L. Peng, L. D. Palma, A. Abdelkafi, A. Liu, Y. Diao: “Optimization for active learning-based interactive database exploration”, VLDB 2018.
    - 🕮 Y. Luo, X. Qin, N. Tang, G. Li: “Deepeye: Towards automatic data visualization”, ICDE 2018.

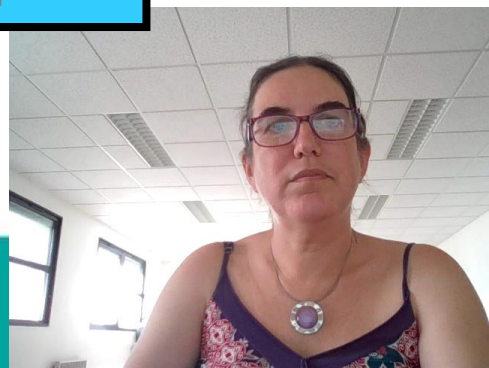
Outline

- **What is a data narrative?**
- **A panorama of tasks and tools for supporting data narration**

  Focus on:
  - Supporting intentional querying
  - Searching interesting findings
- **Open challenges**

More support for data narratives





Conclusion

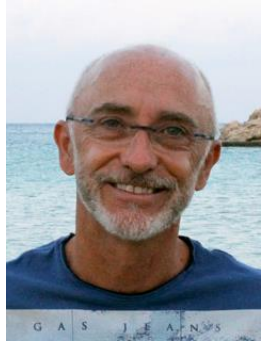
- **A first definition and a conceptual model of data narrative**
- **Complex process combining varied tasks**
  - Good support for data exploration and visualization
  - Precursor works arounds intentional querying, interestingness mining, story synthesis
- **Many open challenges around supporting intentional and structural tasks**

People working around this


Nicolas Labroche


Panos Vassiliadis

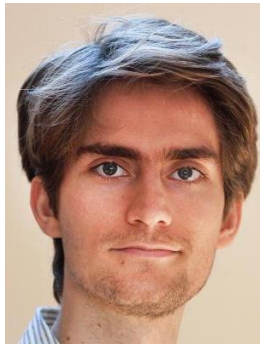

Patrick Marcel


Stefano Rizzi

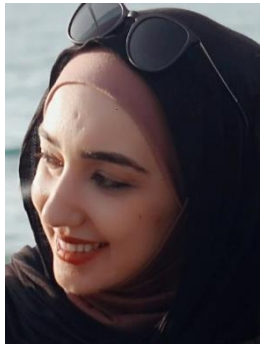

Thomas Devogele


Verónika Peralta


Alexandre Chanson


Antoine Chedin


Clément Moreau


Faten El Outa


Flavia Serra


Matteo Francia


Raymond Ondzigue Mbenga

Thank you for your attention

université de TOURS

Veronika.Peralta@univ-tours.fr

OLAP III

http://www.cs.uoi.gr/~pvassil/projects/olap_III/