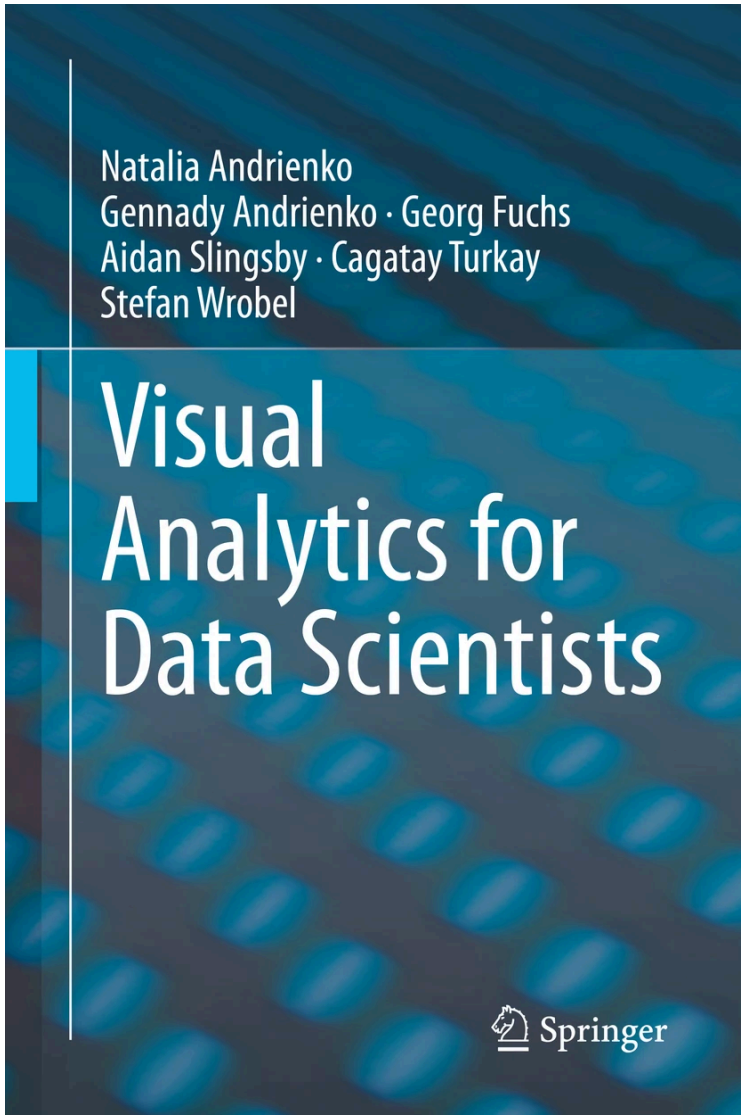
1 · VISUAL ANALYTICS - INTRODUCTION

Combining Human Reasoning with Computational Power

Based on: Chapter 1: Introduction to Visual Analytics by an Example



TEACHER
Salvo Rinzivillo

Università di Pisa · Dipartimento di Computer Science
Master in Big Data Analytics and AI for Society

Who am I?

- **Salvo Rinzivillo** – Senior Researcher at ISTI-CNR, Pisa, Italy.
- **Research Interests:** Visual Analytics, Data Mining, Human-Computer Interaction.
- **Contact:**
  - Email: rinzivillo@isti.cnr.it (use the tag [VA] in the subject)
- Page course: https://didawiki.cli.di.unipi.it/doku.php/magistraleinformaticaeconomia/va/start
- GitHub:
  - https://github.com/va602aa-master
- Telegram: https://t.me/+FKe7v-AZDaU4MzE8

Outline

1. **Part 1: Defining Visual Analytics**
    - What is it?
    - Visual Analytics vs. Data Science
    - The Human-Computer relationship
2. **Part 2: Case Study: The Vastopolis Epidemic**
    - Data Preparation & Cleaning
    - Spatio-Temporal Analysis
    - Hypothesis Generation & Verification
3. **Part 3: The Visual Analytics Framework**
    - Keim's Process Model
    - The Role of Mental Models

Part 1: What is Visual Analytics?

> "Visual analytics is the science and practice of analytical reasoning by combining computational processing with visualisation."

Core Principles:

- **Goal:** Enable synergistic work between humans and computers.
- **Mechanism:** Tight coupling of interactive visual interfaces and automated processing.
- **Role of Visualization:** It is not just for *presenting* results; it is a tool for *doing* the analysis.

Why Humans in the Loop?

Even with advanced AI and Machine Learning, human reasoning is essential for:

1. **Solving Non-Trivial Problems:** Requires understanding the subject matter, not just executing code.
2. **Understanding "Black Boxes":** As datasets grow and models become opaque, we cannot take computational results for granted.
3. **Parameter Tuning:** Analysts need to see how changing inputs or parameters affects the outcome.
4. **Context:** Computers cannot easily handle incomplete or inconsistent information without human guidance.

Visual Analytics vs. Data Science

The fields overlap, but their focus differs:

Feature	Data Science	Visual Analytics
Primary Focus	Computational processing & computer models	Human reasoning & mental models
Outcome	Predictive models (executable by computers)	Knowledge / Mental models (understanding)
Role	Automating analysis	Facilitating reasoning through interaction

Visual Analytics is instrumental for doing good Data Science.

Part 2: A Motivating Example

Investigating an Epidemic Outbreak (IEEE VAST Challenge 2011)

The Scenario:

- **Location:** Vastopolis (Metropolitan area, pop. ~2 million).
- **Event:** Dramatic increase in illnesses (flu-like symptoms) and reported deaths.
- **Goal:** Identify "Ground Zero," the transmission method, and the containment status.

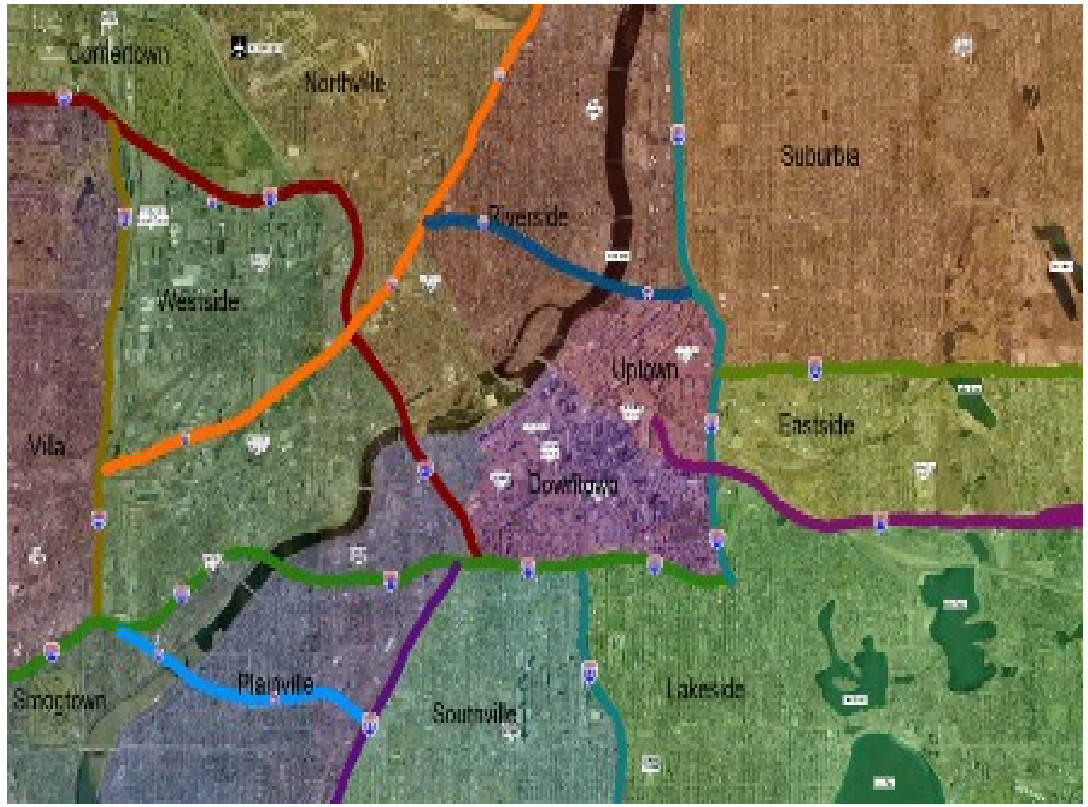


Fig 1.1: The investigation area

The Data

We have access to two primary datasets:

1. **Microblogs (Tweets):**
   - ~1 million records.
   - Contains: Text, User ID, Timestamp, GPS location.
   - *Challenge:* Unstructured text, noise, varying relevance.
2. **Map Data:**
   - Satellite imagery.
   - Labels: Highways, hospitals, rivers, districts.
   - Supplemental: Weather and population density tables.

Data Properties & Challenges

Important Consideration:

- The data does **not** directly represent disease cases.
- It represents **texts** that *might* mention symptoms.

Implications:

- **Time Lag:** People might post *after* getting sick, not immediately.
- **Location Lag:** They might post from home, not where they were infected.
- **Noise:** People might discuss "sick beats" or "love-sick," not the flu.
- **Goal:** Identify the subset of data characterized by disease terms and high temporal frequency.

Step 1: Data Preparation (Keywords)

Objective: Select potentially relevant records.
Method: Interactive Querying.

1. **Initial Filter:** Keywords like *fever, chills, sweats, fatigue*.
2. **Refinement:** Use a **Word Cloud** to see frequent terms in the filtered set.
3. **Discovery:** New relevant terms appear (e.g., *flu, stomach, sick, doctor*).
4. **Iteration:** Add new terms to the query and update the view.

Fig 1.2: Interactive term extraction

Step 1: Cleaning the Noise ("Chicken Flu")

Visual Inspection:

- Word Cloud showed "chicken" and "flu" frequently.
- Context: "Fried chicken flu" and "Chicken flu trending."

Investigation:

- **Time Histogram:** Shows "chicken flu" messages (bottom) are evenly distributed, unlike the disease spike (top).
- **Conclusion:** This is a social trend, not a disease.
- **Action:** Filter out "chicken".
- **Result:** Data reduced to ~80k relevant records.

Fig 1.5: Comparing disease patterns (top) vs. noise (bottom)

Step 2: Temporal Analysis (When?)

Goal: Identify the exact start of the outbreak.

1. **Daily Resolution:**
   - Histogram shows a massive spike in the last 3 days.
2. **Hourly Resolution:**
   - Drill down into the last 5 days.
   - **Observation:** A sharp increase begins at **13:00 on May 18th**.
   - **Validation:** Night-time drops (02:00–05:00) confirm data reflects human cycles.

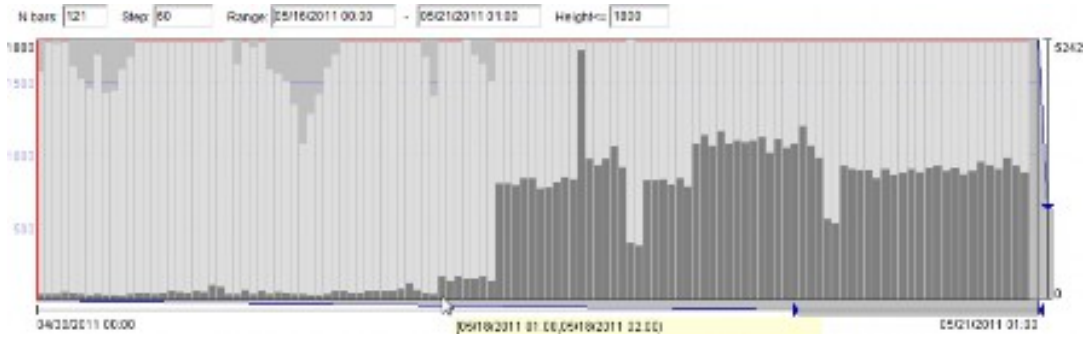


Fig 1.6: Identifying the outbreak start

Step 3: Spatial Analysis (Where?)

Goal: Locate "Ground Zero."

Visual: Dot Map of disease-related messages.
Observation:

- High density in **Downtown** and **Uptown**.
- **Problem:** These are also the most populous areas.
- **Question:** Is this a disease cluster, or just a population map?

(Visual comparison of disease messages vs. all messages is required)

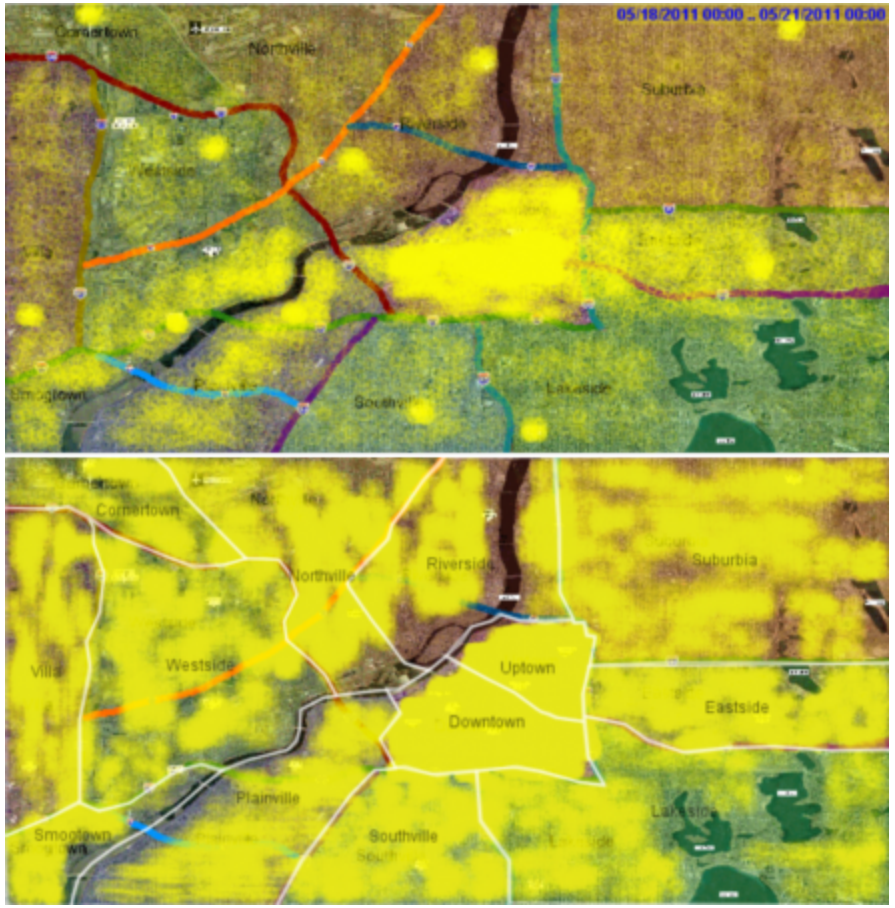


Fig 1.7: Disease messages (Top) vs. Unrelated (Bottom)

Step 3: Verification (Data Transformation)

Method: Normalize data by population density.

1. **Calculate Baseline:** Average daily messages per district (pre-outbreak).
2. **Calculate Outbreak:** Disease messages per district (during outbreak).
3. **Visualization:** Bar charts on the map representing the **Ratio**.

Fig 1.8: Verifying the hotspot with ratios

Result:

- **Downtown** has a significantly higher ratio.
- **Conclusion:** The Downtown cluster is a real hotspot.

Step 4: Spatio-Temporal Analysis

Goal: Understand the spread evolution.
Tool: Space-Time Cube (STC).

- **X/Y:** Geography.
- **Z:** Time (Bottom to Top).

Observations:

1. **Day 1 (May 18):** Three dense clusters in center, spreading East.
2. **Day 2 (May 19):** Spread moves to **Southwest**.
3. **Day 3 (May 20):** Clusters appear around **Hospitals**.

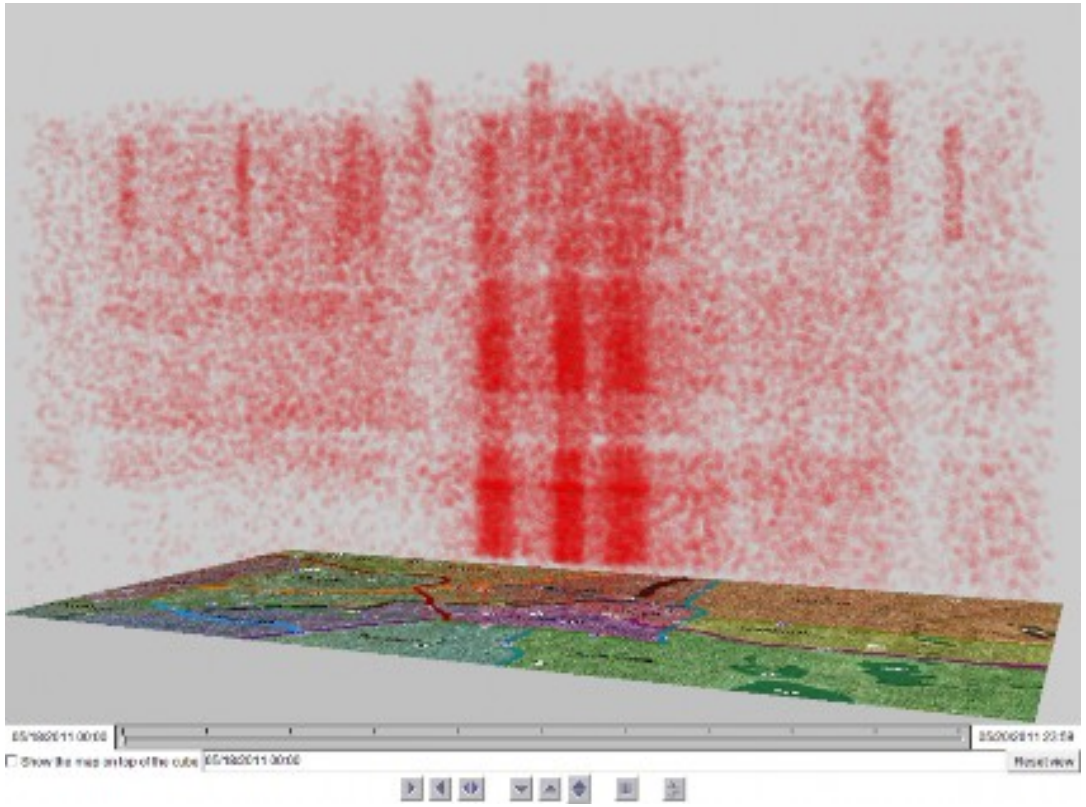


Fig 1.9: Spatio-temporal distribution

Step 5: Refining the Signal (First Mentions)

Problem: Ill people post multiple messages, obscuring the *spread*.
Solution:

- Link messages by **User ID**.
- Extract only the **first** disease-mentioning message per user.

New Insight:

- The "Hospital clusters" (Day 3) disappear.
- The spread pattern is clearer: **Center/East** (Day 1) $\rightarrow$ **Southwest** (Day 2).

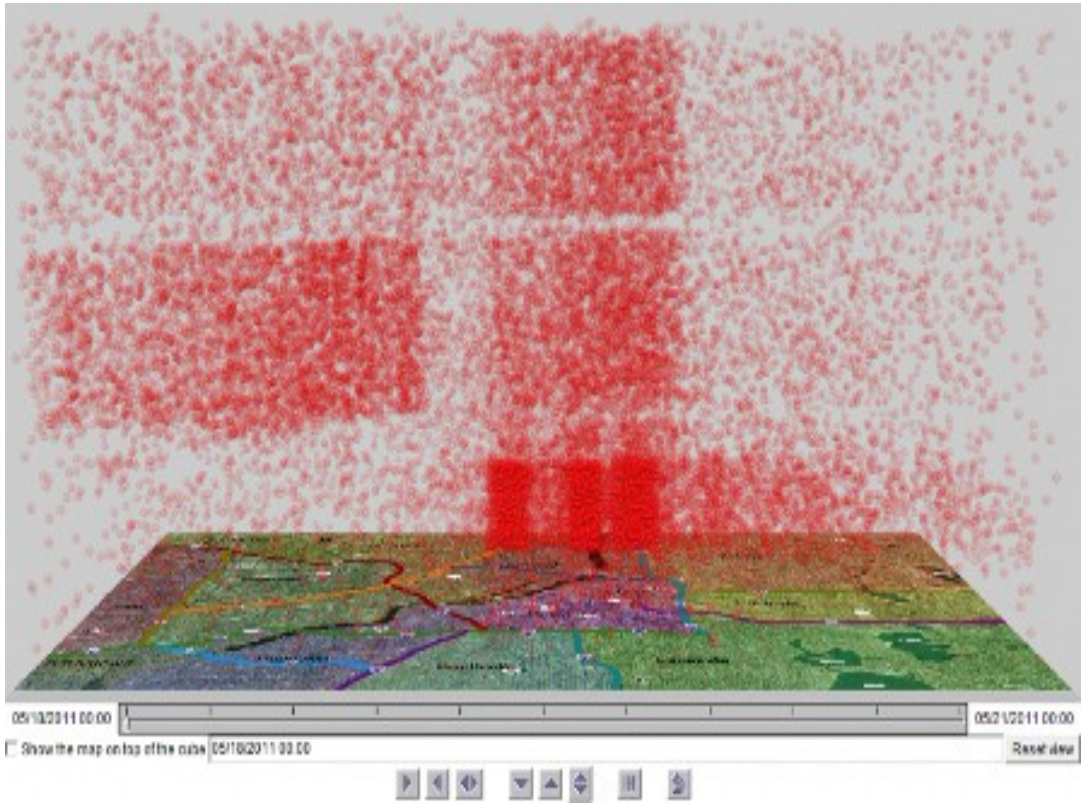


Fig 1.11: First mentions of symptoms

Step 6: Hypothesis Generation

Investigation 1: Weather

- May 18: Wind from **West**.
- *Hypothesis:* Explains spread to the East (Airborne).

Investigation 2: Symptoms by Region

- **East:** *Chills, fever, cough* (Flu-like).
- **Southwest:** *Stomach, nausea, diarrhea* (Waterborne).
- *Conclusion:* **Two different syndromes**.

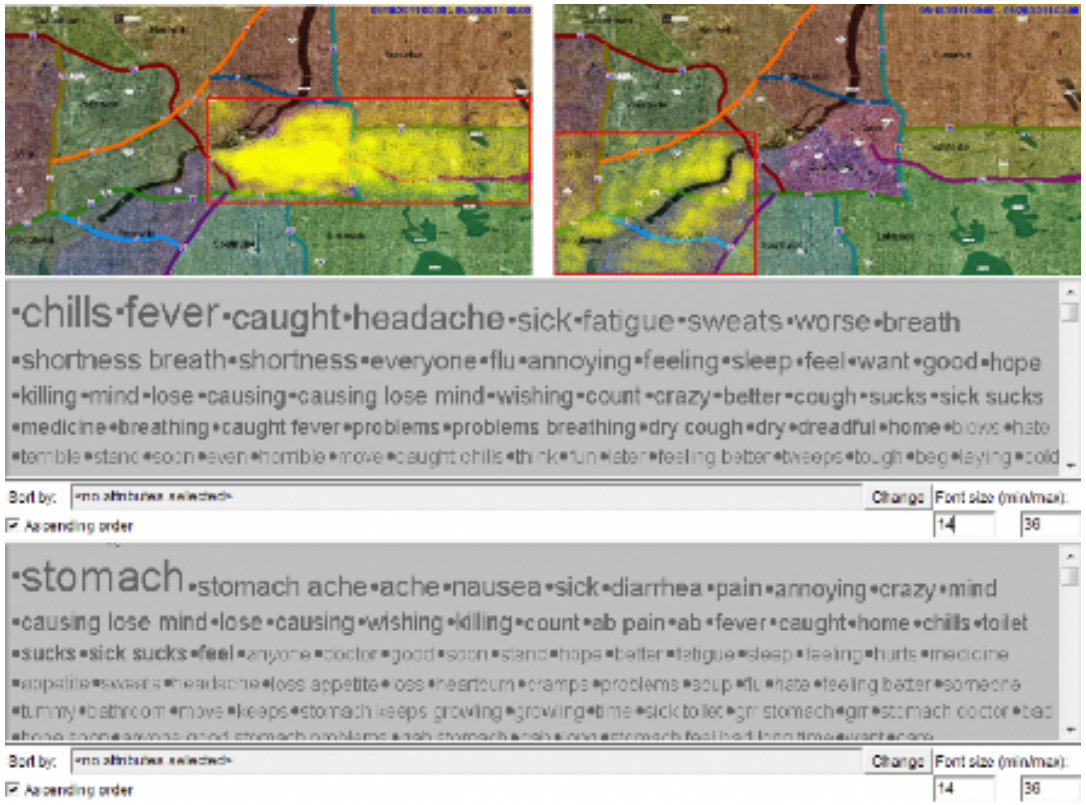


Fig 1.12: Comparing symptoms by area

Step 7: Identifying the Source

The Intersection:

- The Airborne path (East) and Waterborne path (Southwest) emanate from a common point.
- **Location:** Motorway bridge crossing the river.
- **Hypothesis:** An event at the bridge caused both.

Verification:

- Query: Location = Bridge, Time = May 17.
- **Result:** *Truck, Accident, Fire, Spilling Cargo*.

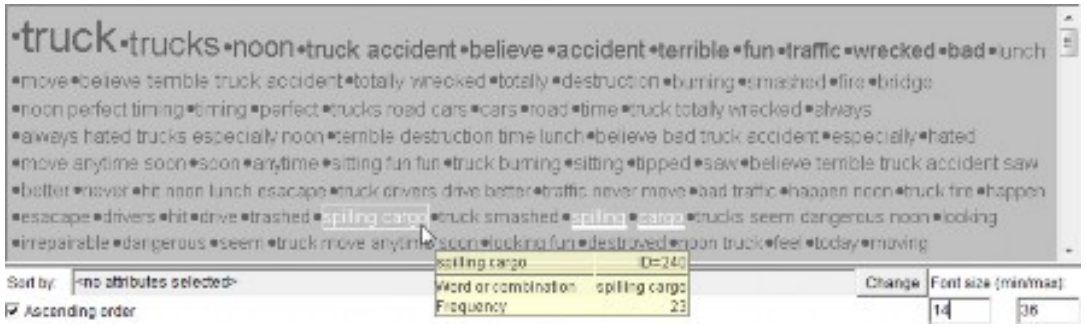


Fig 1.13: Evidence of the accident

Summary: The Reconstructed Story

1. **The Event:** May 17, noon. A truck carrying toxic substance crashes and burns on the bridge.
2. **Transmission A (Air):** West wind carries toxic smoke to the Center and East. People inhale it and get flu-like symptoms (Day 1).
3. **Transmission B (Water):** Spilled cargo enters the river. Current carries it Southwest. People contact water and get digestive symptoms (Day 2).
4. **Conclusion:** The "Epidemic" is actually a toxicity event. It is not contagious (person-to-person).

Part 3: Why Visual Analytics Worked

Efficiency:

- "A picture is worth a million records."
- We replaced reading 1,023,077 texts with recognizing visual patterns.

Cognitive Support:

- **Abstraction:** Vision allows us to subconsciously construct patterns (clusters, gaps, outliers).
- **Reasoning:** Perceiving these patterns triggers hypothesis generation.
- **Interaction:** We didn't just look; we filtered, queried, and transformed data iteratively.

The Visual Analytics Process (Keim's Model)

The process is a feedback loop involving the Analyst and the Computer:

1. **Data:** Input for processing.
2. **Computational Processing:** Automated analysis, mining.
3. **Visualization:** Mapping results to visual forms.
4. **Interaction:** The human refines parameters, selects data, or changes views.
5. **Knowledge:** The ultimate output is a Mental Model.

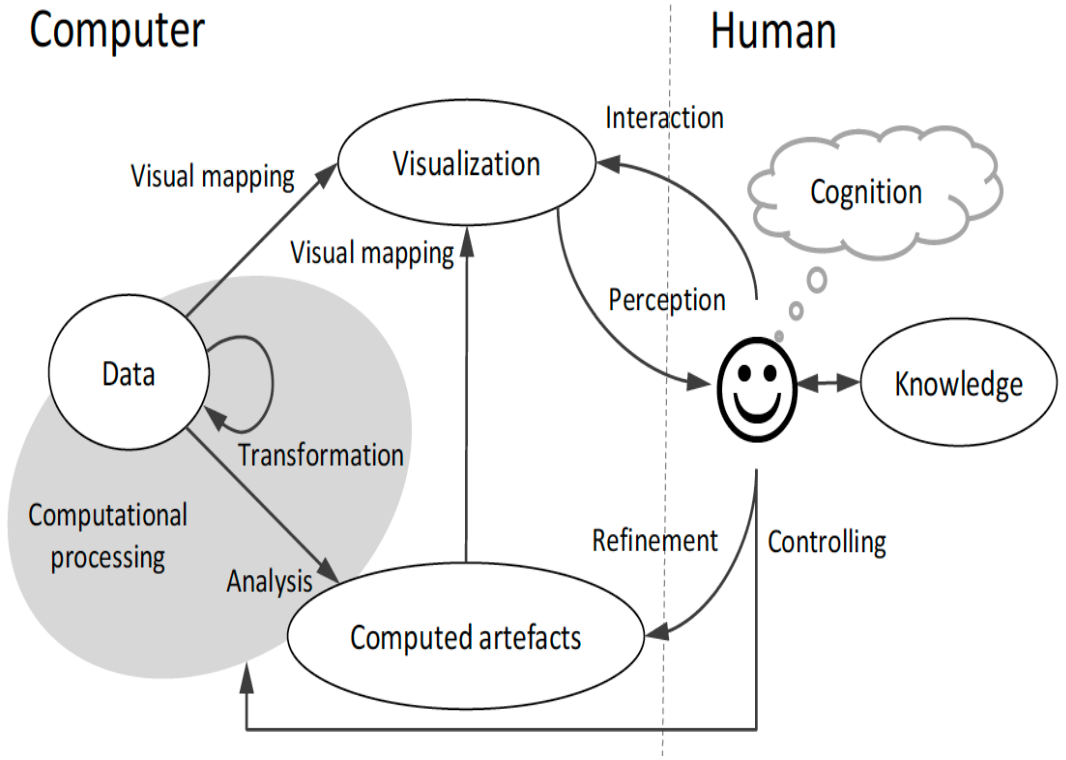


Fig 1.16: The VA Process Model

Human-Computer Collaboration

Computer Strengths:

- Handling massive data volumes.
- Fast searching and processing.
- Consistent execution of algorithms.

Human Strengths:

- Flexible and inventive reasoning.
- Contextual awareness (e.g., recognizing "Chicken Flu" is food, not disease).
- Handling incomplete or inconsistent information.
- "Seeing the forest for the trees."

Conclusion

- **Visual Analytics** facilitates analytical reasoning through interactive visual interfaces.
- It is essential for **Data Science** when exploration, understanding, and verification are required.
- **The Workflow:**
  - **Explore** (Data Prep)
  - **Analyze** (Space/Time patterns)
  - **Verify** (Hypothesis testing)
  - **Conclude** (Story reconstruction)

Next Lecture: We will dive into the specific visualization techniques and tools used to build these systems.