Intro to Coding with Python– Visual Analytics

Dr. Ab Mosca (they/them)

Plan for Today

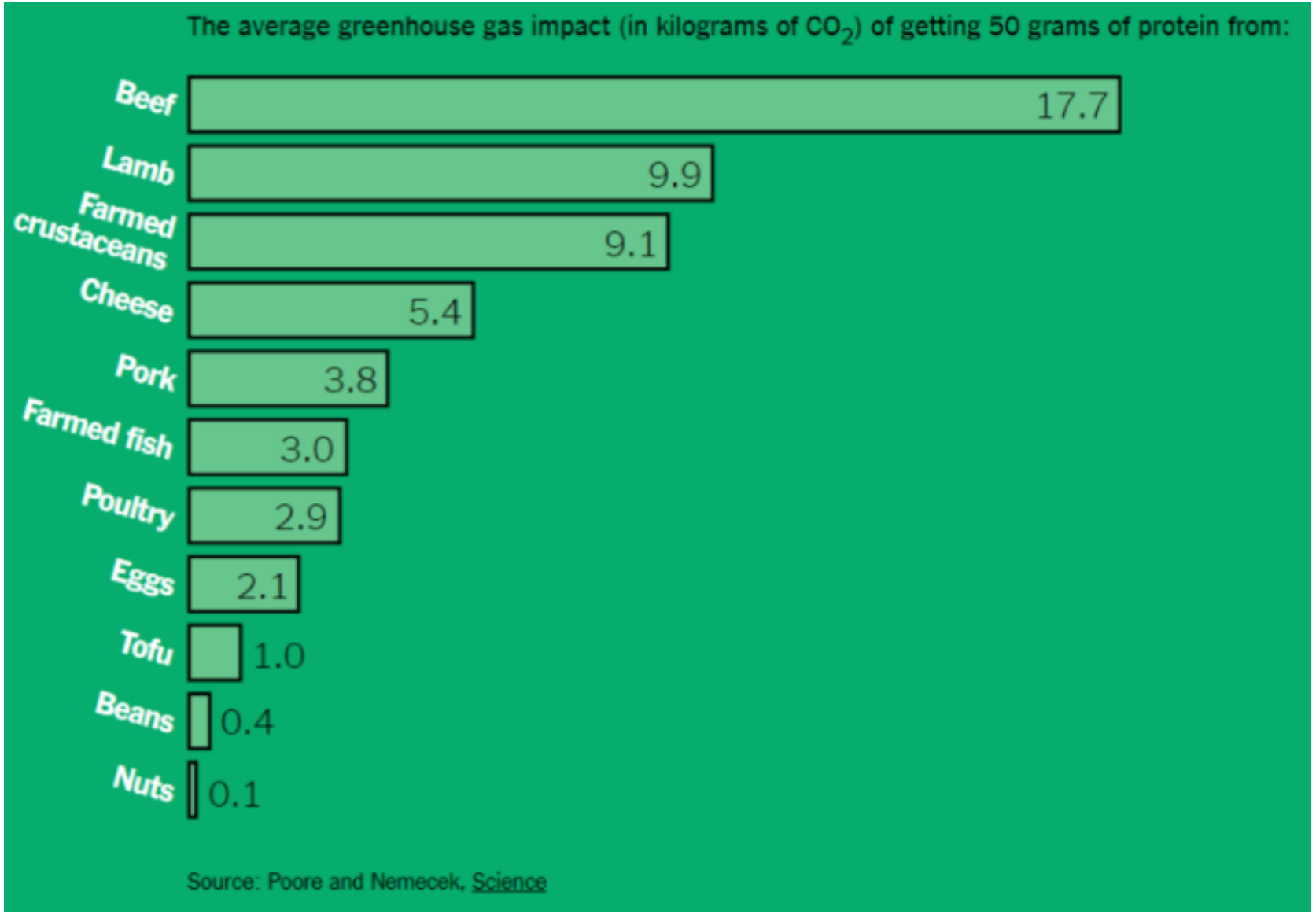
- Advanced Topic: Visual Analytics

What is Visualization?

Definition: Visualization

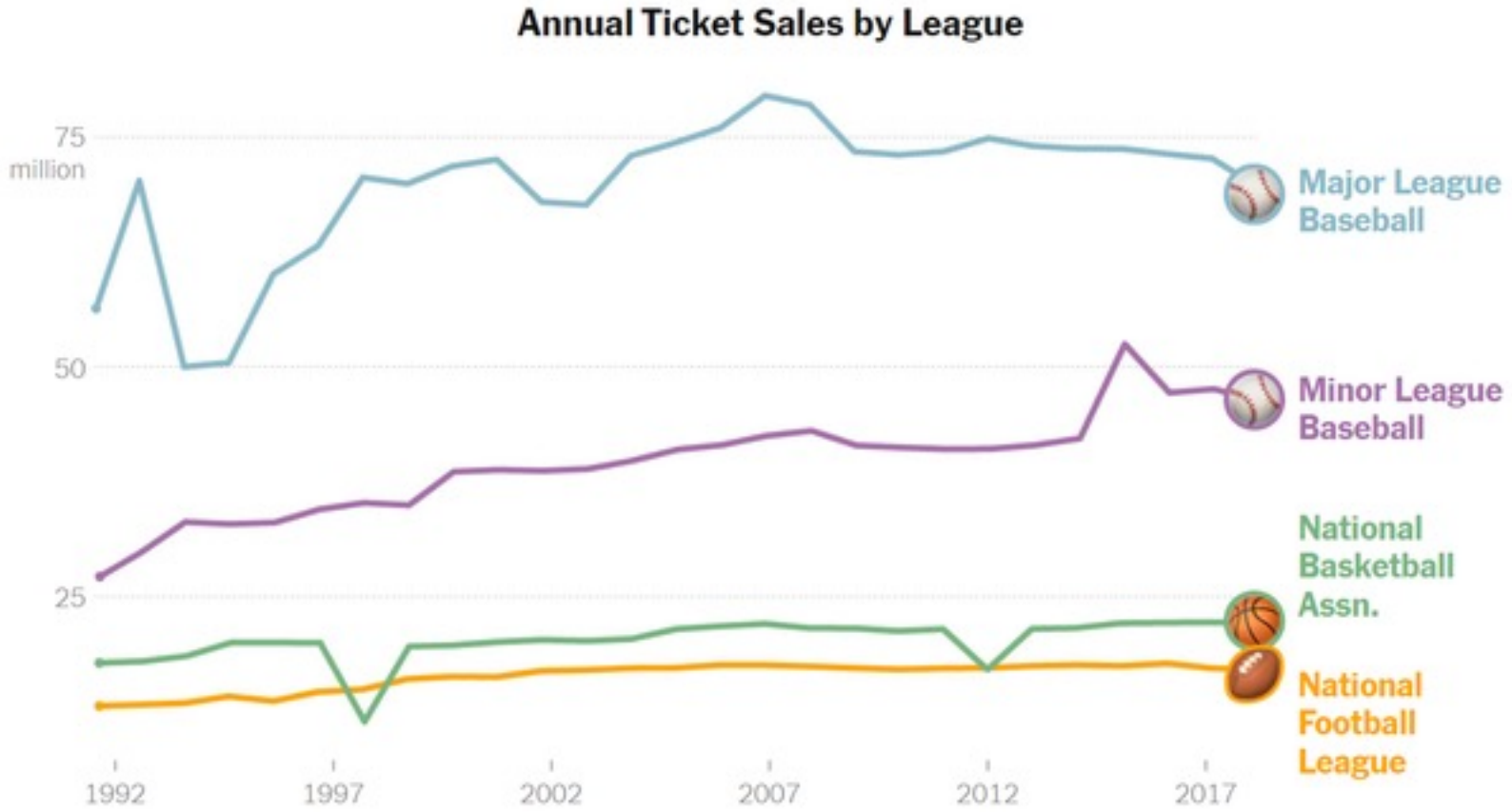
- The graphical representation of information and data.

Example



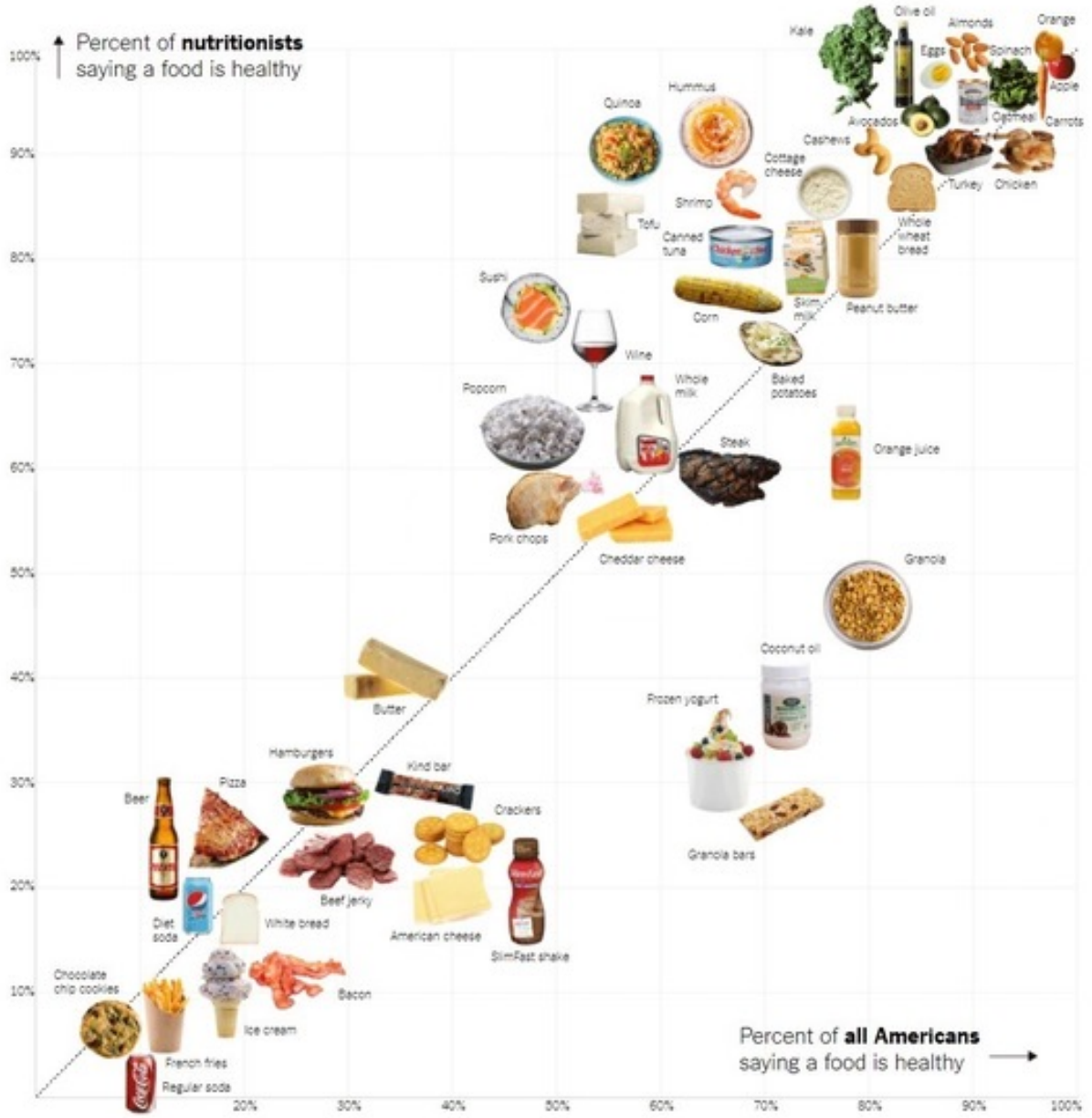
https://www.nytimes.com/2019/12/05/learning/whats-going-on-in-this-graph-dec-11-2019.html

Example



https://www.nytimes.com/2019/10/24/learning/whats-going-on-in-this-graph-oct-30-2019.html

Example



https://www.nytimes.com/2017/10/09/learning/whats-going-on-in-this-graph-oct-10-2017.html

What is Visual Analytics?

Definition: Visual Analytics

- The science of analytical reasoning facilitated by visual interactive interfaces

James J. Thomas and Kristin A. Cook (Ed.) (2005). *Illuminating the Path: The R&D Agenda for Visual Analytics* National Visualization and Analytics Center

Definition: Visual Analytics

What is analytical reasoning?

- The science of analytical reasoning facilitated by visual interactive interfaces

James J. Thomas and Kristin A. Cook (Ed.) (2005). *Illuminating the Path: The R&D Agenda for Visual Analytics* National Visualization and Analytics Center

Example 1: "How the Swedes Book their Summer Vacations"

- Tripadvisor (circa 2015)
  - "Tell me if there's anything interesting about the way the Swedes book their summer vacations"

Example 1: "How the Swedes Book their Summer Vacations"

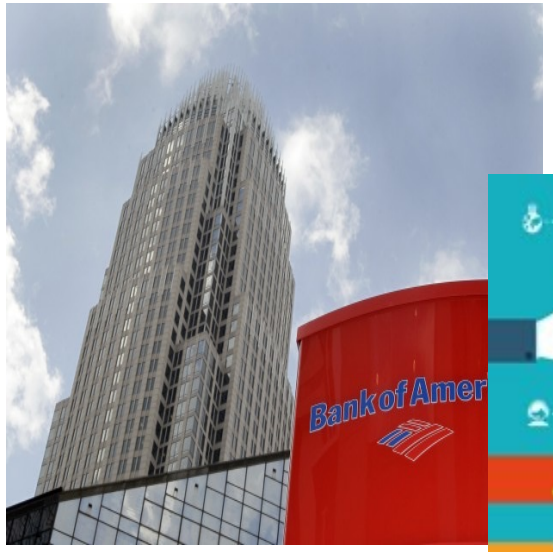
- Context
  - The boss was very interested in the answer to this problem because it represented a business advantage
  - (The student working on this problem quit after 2 weeks of banging his head against this)
  - Tripadvisor just merged with another company. So data resided in multiple (unconnected) databases
  - The sizes of the databases are massive. They don't fit in a typical hard drive of a desktop computer
- How would you go about solving this problem?

Example 1: "How the Swedes Book their Summer Vacations"

- Why is this hard?
  - Problem is ill-defined. What does "interesting" mean?
    - Is "interesting" how Swedes book summer versus winter vacations?
    - Or "interesting" when compared to other Europeans in their summer vacations?
    - Or against Americans or others around the world?
  - Data is too disparate
    - How do I extract the right data from multiple databases when I don't know what the question is?
  - Data is too large
    - Can't fit the whole database into (Excel | R | memory | hard drive), so how do I find a representative sample?
  - No idea what's in the data or what it looks like
    - Unclear what kinds of statistics (or machine learning) is needed
    - Don't know how to visualize this data

Example 2: "Financial (Wire) Fraud"

- Bank of America (circa 2007)
  - Legal responsibility for banks to report suspicious financial transactions (money laundering, supporting terrorist organizations, etc)



Example 2: "Financial (Wire) Fraud"

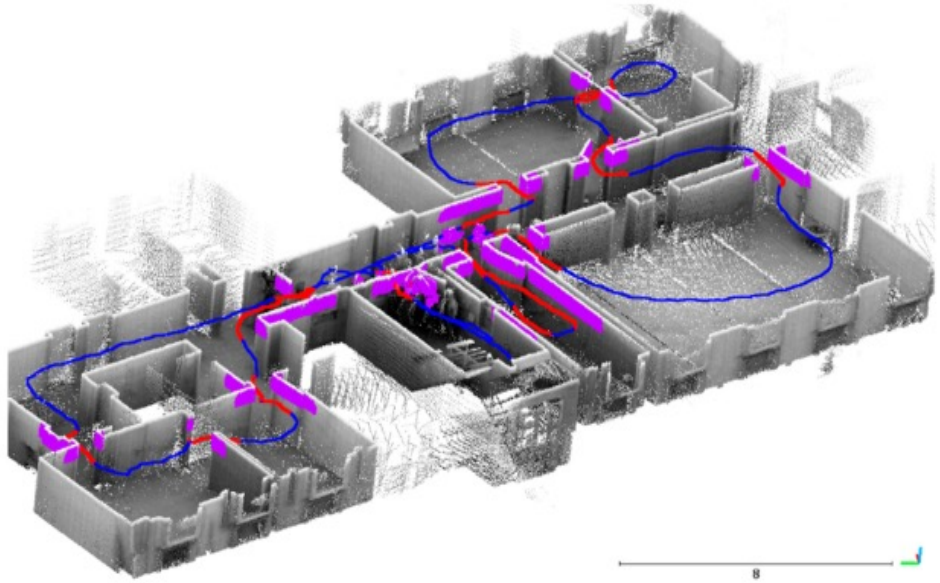
- Context
  - All banks in the US are required to report suspicious financial transactions, in this case, wire transactions
  - The size of the dataset is massive. There are hundreds of thousands to millions of wire transactions per day that go through a large US bank
  - At Bank of America, around 2007, 10-15 analysts (the WireWatch group) were responsible for monitoring and reporting of ***all*** of Bank of America's suspicious activity reports (SARs) on wire transactions
- How would you coordinate a team of analysts?

Example 2: "Financial (Wire) Fraud"

- Why is this hard?
  - Problem is ill-defined. What does "suspicious" mean?
    - No single transaction by itself is inherently fraudulent (there are exceptions). What do "fraudulent patterns" look like?
  - Cat-and-Mouse game
    - Bad guys are smart. They change tactics. So automated detection systems are (by and large) only good at catching senseless bad guys.
  - Data is too large and complex
    - How does someone examine hundreds of millions to billions of transactions (over a year) to find patterns?
  - Lack of ground truth
    - No one knows how well the detection algorithm or method is working because no one has "ground truth" data
  - Coordination and collaboration is hard
    - Analysts were divided into regions in the world, but nowadays fraud occurs between regions. So who's responsible for what, and how do they coordinate?

Example 3: "Disease Spread within a Hospital"

- Mass General Hospital (MGH), (circa 2019)
  - "Do we have a problem of disease spreading within the hospital? If so, what happened?"



Example 3: "Disease Spread within a Hospital"

- Context
  - Doctors and administrators at MGH are wondering if a recent increase in common colds, flu, etc. in the patient population of a particular floor at MGH is caused by practices within the hospital
  - Personnel movements within a hospital is continuous, fluid, and can be chaotic. There are lots of people (patients, nurses, doctors, visitors, etc.) and lots of places of interactions (rooms, hallways, nurse stations, bathrooms, etc.)
  - Data is available but disjointed. While per-second location information isn't immediately available, they can be extracted if needed (e.g. from camera)

- What data would you use to start your analysis? How would you prepare your data?

Example 3: "Disease Spread within a Hospital"

- Why is this hard?
  - Unlike the previous problems, in this case the problem is better defined. However, the challenge is finding a "smoking gun"
    - For example, when two trajectories cross each other, does that mean the disease has spread?
  - Unlike the previous problems, the data size here is much smaller in comparison. However, data quality is a concern
    - Aside from trajectory information, there are issues relating to dwell, amount of (physical) interactions, accuracy of location, uncertainty regarding whether each person carries some known or unknown disease
  - The number of hypotheses and "analysis trails" is very large
    - Partially due to the issues relating to data quality, there are many possible explanations for each phenomenon
  - Ultimately, the "cost" of **making a wrong decision** is very high

Reflecting on the Three Examples

- What did they NOT have in common?

Reflecting on the Three Examples

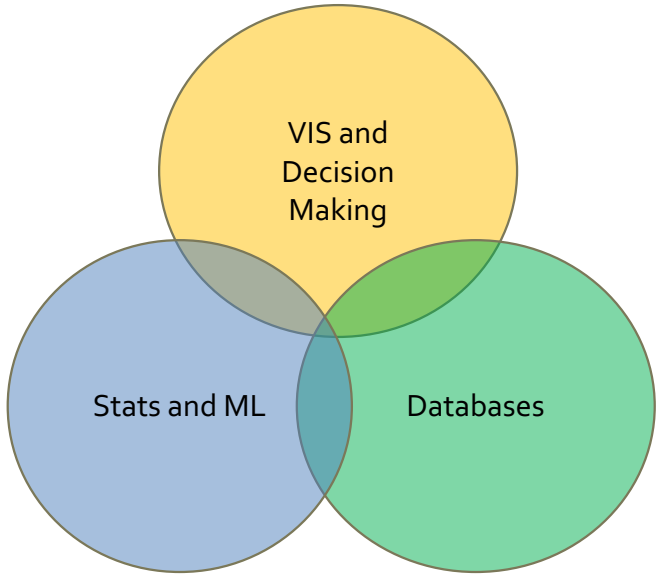
- What did they NOT have in common?
  - Data sizes differ
    - MGH example has much less data than Tripadvisor or Bank of America
  - Analysis methods differ
    - Stats or ML methods for analyzing vacation-booking patters are different from trajectory analyses for a hospital
  - Users are different
    - Financial analysts, data science interns, and hospital administrators have different backgrounds and needs

Reflecting on the Three Examples

- What did they have in common?
  - "Wicked problem": problem that is impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize. (source: wiki)
- Computer scientists hate these types of problems…

Reflecting on the Three Examples

- Visual Analytics is the science of analytical reasoning facilitated by visual interactive interfaces (Thomas and Cook, 2004)
- *"Detect the Expected, Discover the Unexpected"*



- "The key is to **let computers do what they are good at**, which is trawling these massive data sets for something that is mathematically odd," said Daniel Gruhl, an IBM researcher whose recent work includes mining medical data to improve treatment. "And that **makes it easier for humans to do what they are good at** — explain those anomalies."[1]
  - Note the emphasis on "phenomenon" and not data

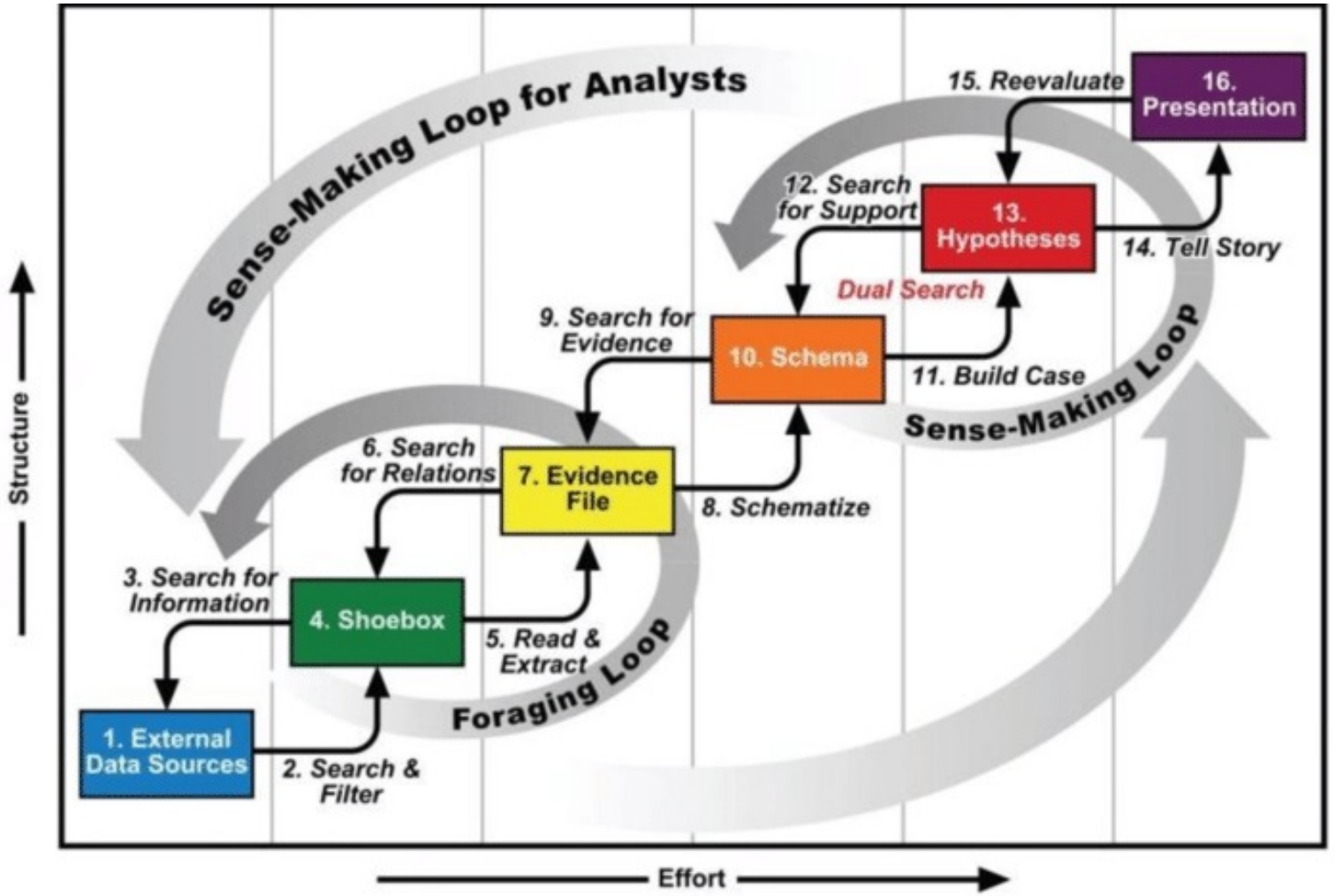
1. New York Times. "For Today's Graduate, Just One Word: Statistics ", August 5, 2009.

Why Visual Analytics (and not ML or Stats or Data Science)?

- In Stats, you learn how to model a problem mathematically
- In ML, you learn how to use computation to fit a model to data
- In both cases, the premise is that the problem is relatively well-defined
- The community of visual analytics wants to tackle the problem space that sits above Stats and ML
  - "Tell me something interesting about…" is a very common (and relatively logical) request, but it isn't immediately obvious how Stats/ML can be applied

BIG Visual Analytics Question

How do we help people "find something interesting"?



Pirolli and Card's sensemaking model

Definition: Visual Analytics

- The science of analytical reasoning facilitated by visual interactive interfaces

James J. Thomas and Kristin A. Cook (Ed.) (2005). *Illuminating the Path: The R&D Agenda for Visual Analytics* National Visualization and Analytics Center

Definition: Visual Analytics

- The science of analytical reasoning facilitated by visual interactive interfaces

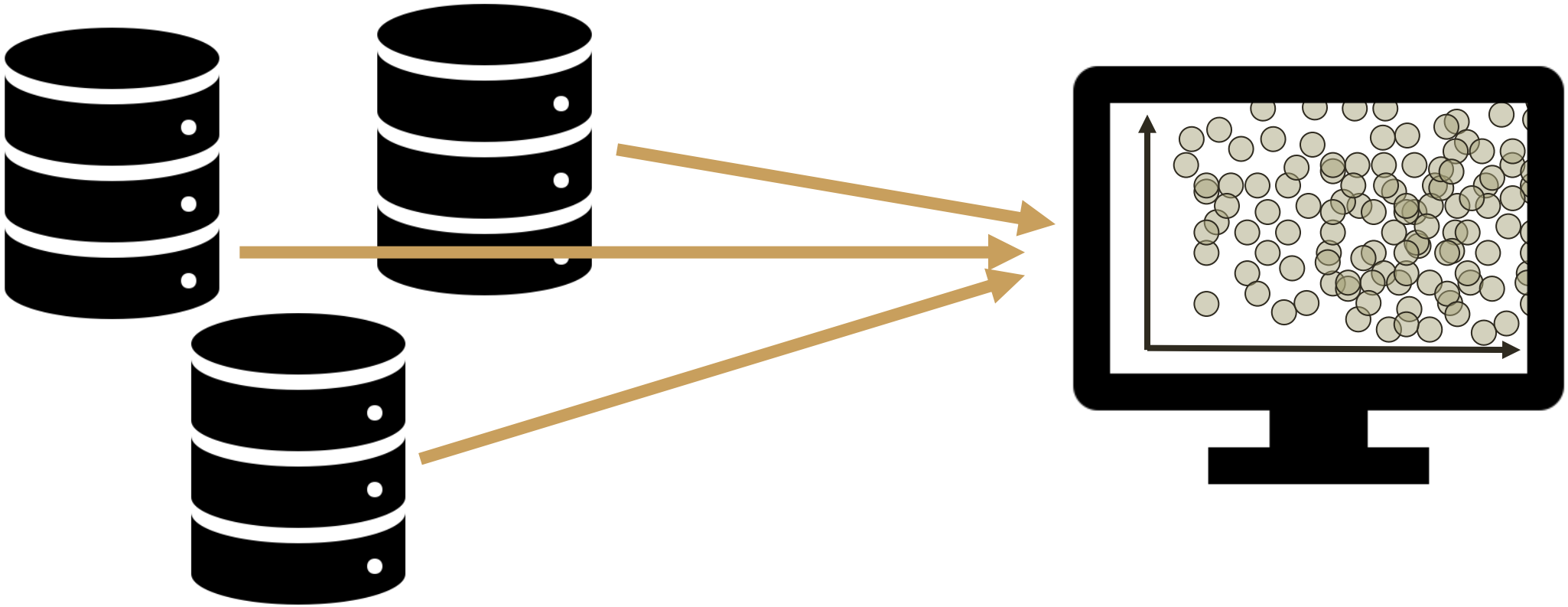
What are **high level** interactions for analytical reasoning?

James J. Thomas and Kristin A. Cook (Ed.) (2005). *Illuminating the Path: The R&D Agenda for Visual Analytics* National Visualization and Analytics Center

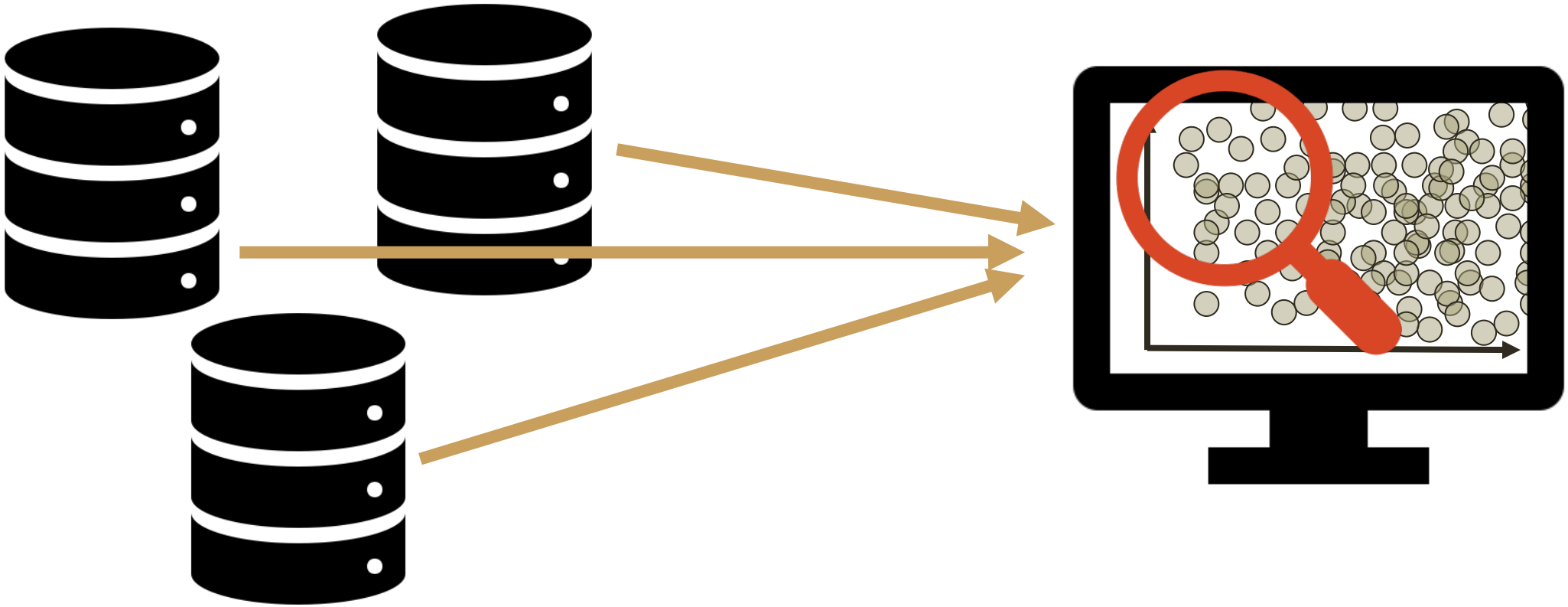
Visual Analytics

- Okay, but why is interaction helpful?

When we have A LOT of data to show, interaction is essential



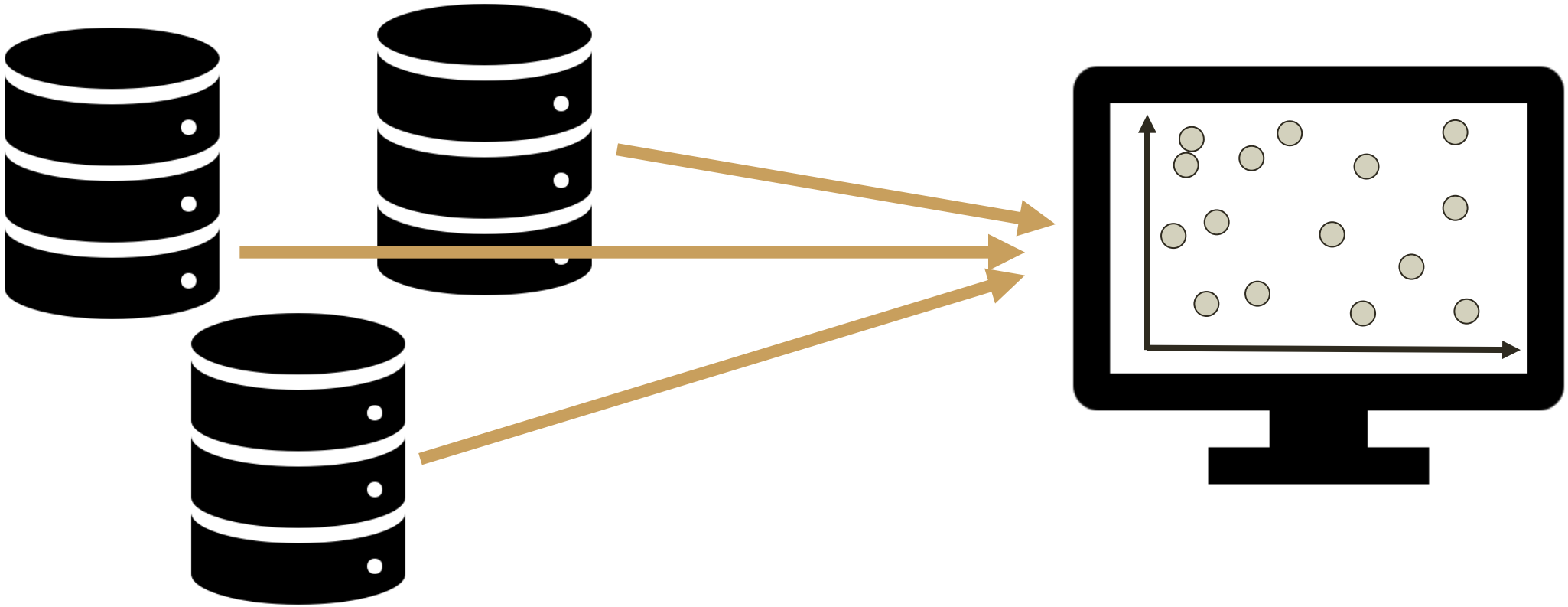
When we have A LOT of data to show, interaction is essential



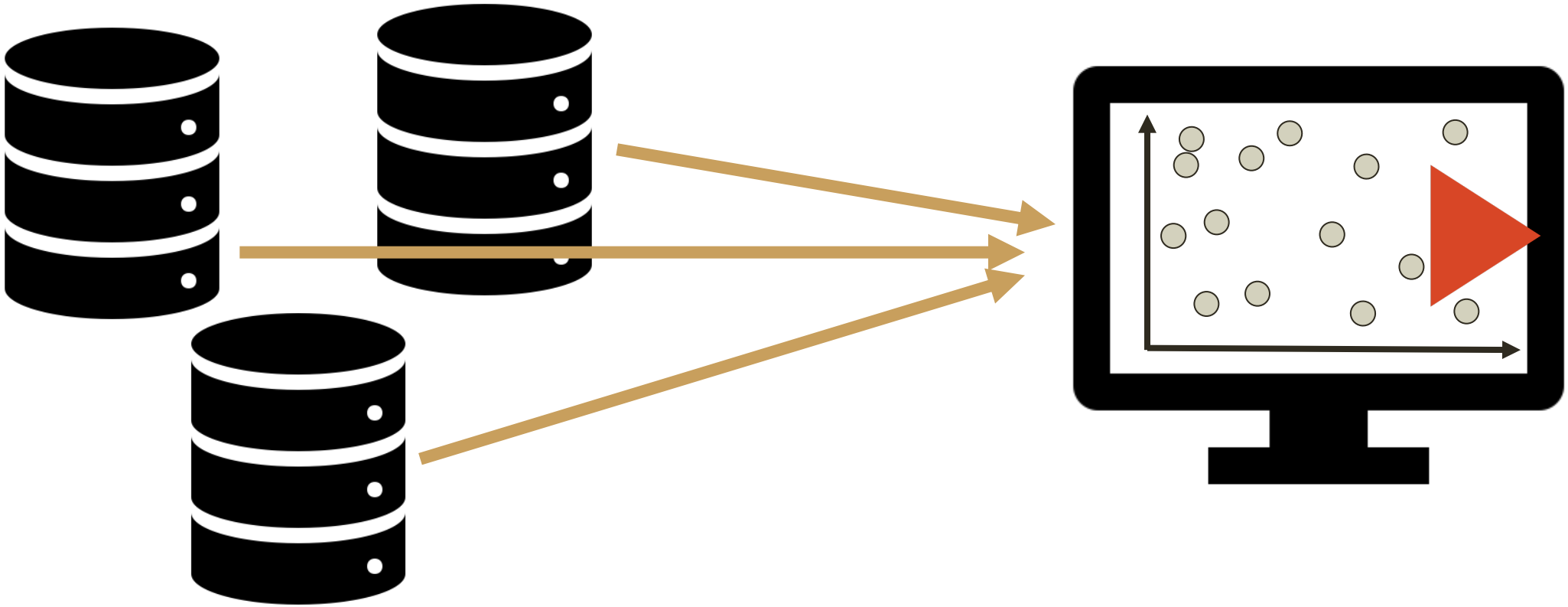
Visual Data Analysis Group

[vlalt

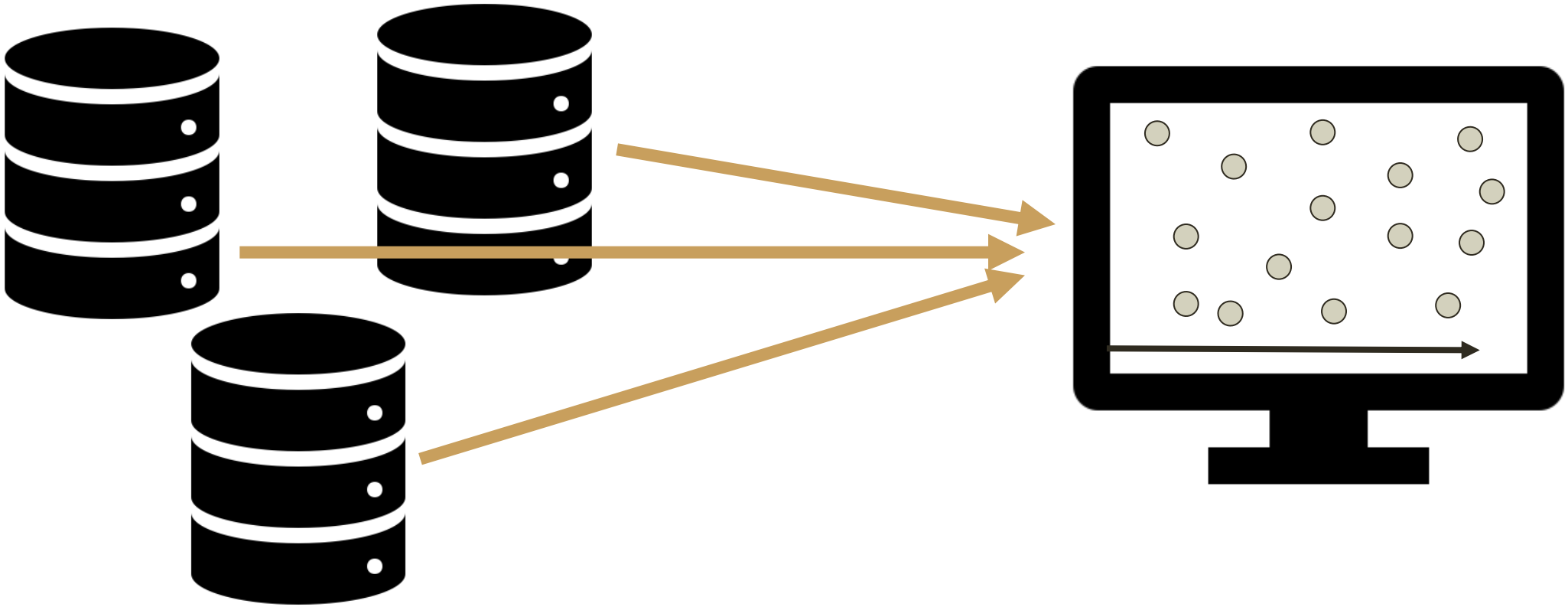
When we have A LOT of data to show, interaction is essential



When we have A LOT of data to show, interaction is essential



When we have A LOT of data to show, interaction is essential



Examples of Visual Analytics Systems

Voyager

- https://vega.github.io/voyager/
- Load the Barley dataset
- Find something interesting

Examples of Visual Analytics Systems

Tableau

- Main Air Pollutants, by the European Environment Agency
- https://public.tableau.com/app/profile/european.environment.agency/viz/Emisions/MainAirPollutants
- Find something interesting

Examples of Visual Analytics Systems

RNNbow