



# Exploratory Data Analysis: from *insights* to *storytelling*

**Patrick Marcel - Verónica Peralta**

**LIFAT - University of Tours – France**

**eBISS 2022 - Cesena**

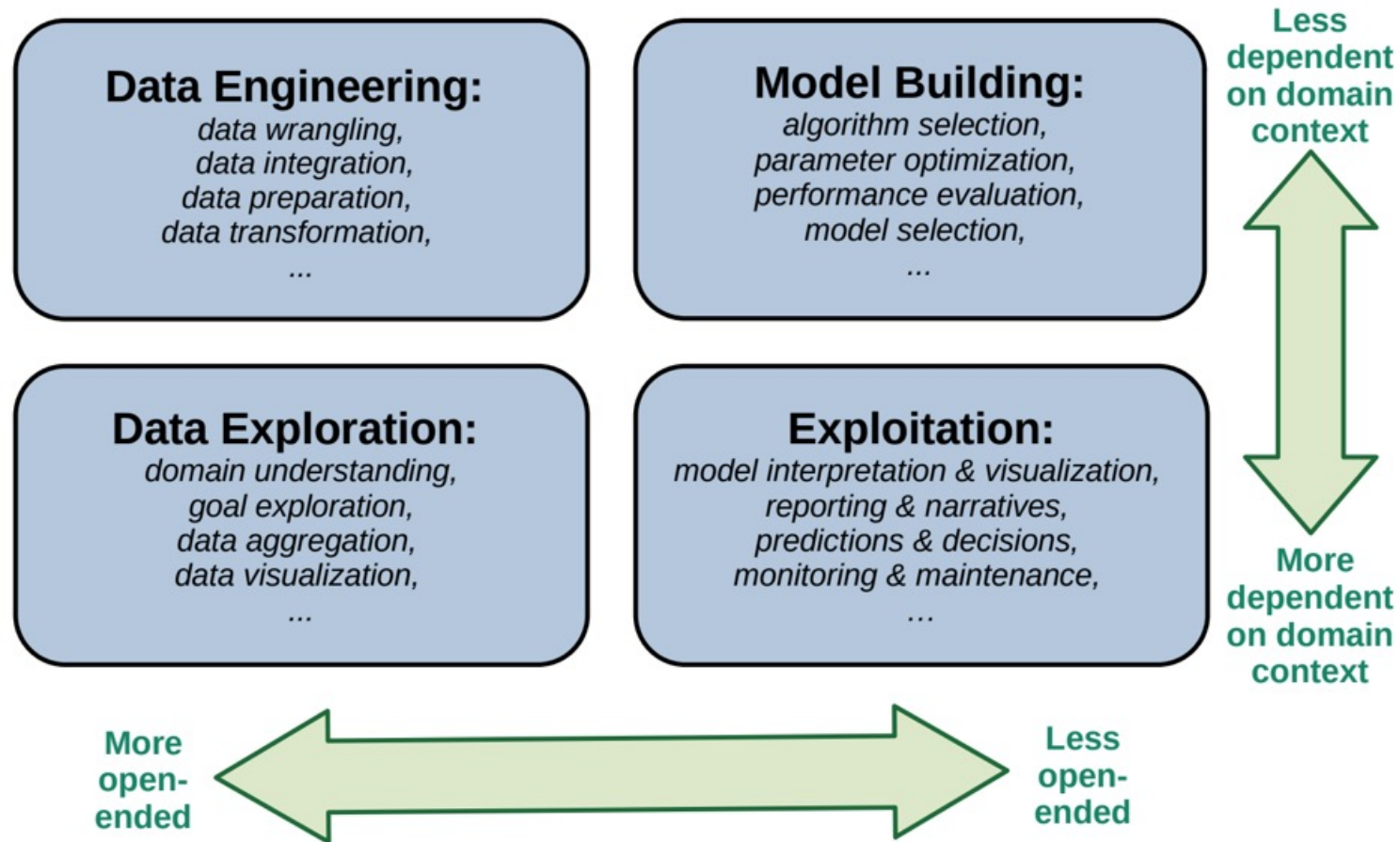
# University of Tours

- ❑ **31 510 students**
  - 3 255 international students, 131 nationalities
- ❑ **1 200 teaching & research staff**
- ❑ **1 300 technical & administrative staff**
- ❑ **35 research laboratories**
- ❑ **3 major fields of Research**
  - Sciences & Technology
  - Life & Health Sciences
  - Human & Social Sciences

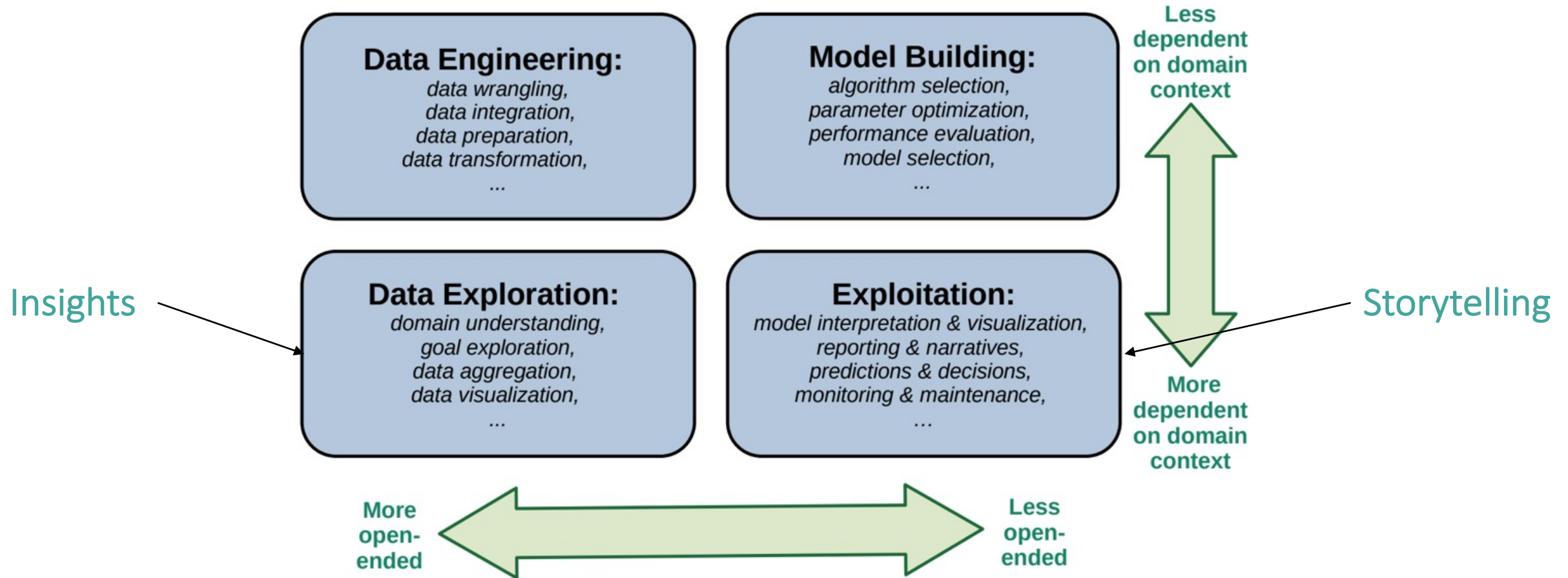
In the heart of the **Loire Valley**  
*The Cradle of the French and the Garden of France*  
Notable for its *historic towns, architecture, and wines*  
In the UNESCO list of *World Heritage Sites*



# The 4 data science quadrants [CACM 2022]



# The 4 data science quadrants [CACM 2022]

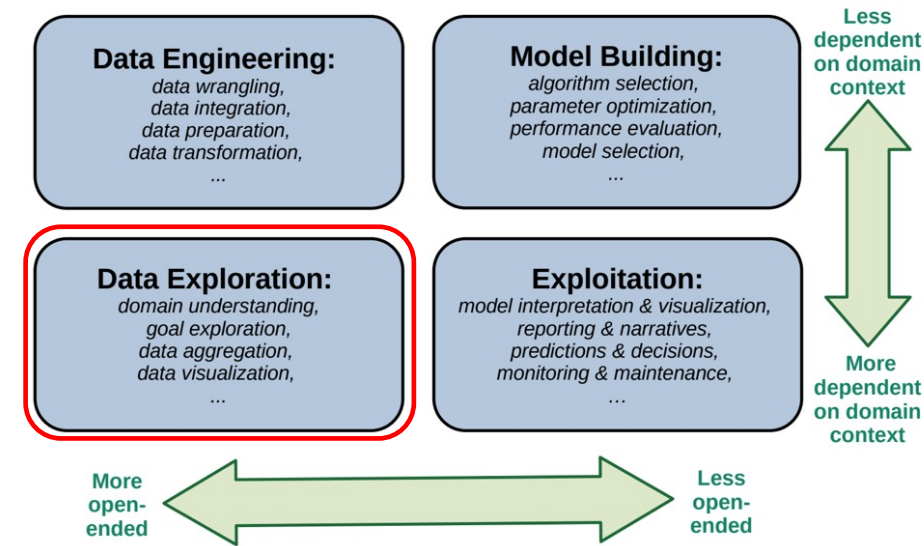




# The data exploration quadrant

## Data Exploration

- ❑ **Interactively** analyzing dataset to gain insights [Tukey 1977]
- ❑ Notoriously **tedious**
- ❑ **Background knowledge** and **human judgement** are key to success
- ❑ Poses the **greatest challenges for automation**
  - Understanding the data analyst's intentions, preferences, perception, cognition capacities, ...



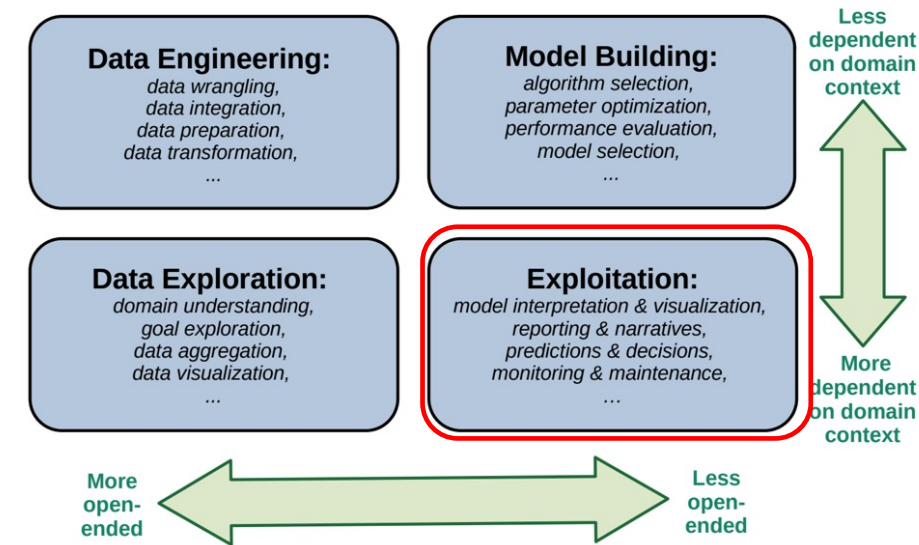
## Five subtasks

1. Form of the patterns
2. Interestingness
3. Algorithmic strategy
4. Presentation
5. Interaction

# The data exploitation quadrant

## Data Exploitation

- ❑ **Understanding** of insights and models produced in the earlier stages
- ❑ **Publishing** them as building blocks for **decisions** and **new discoveries**
- ❑ Some **specific activities** can be automated to a high degree
  - E.g., reporting
- ❑ But **external validation** poses additional challenges
  - E.g., trade-off between accuracy and fairness



# Outline

## □ Part 1: Insights

- The highlights of the 2015 Sigmod tutorial
- What is the problem?
  - and how is it solved?
- Insights
  - and their interestingness
- Human in the loop
  - Declarative languages

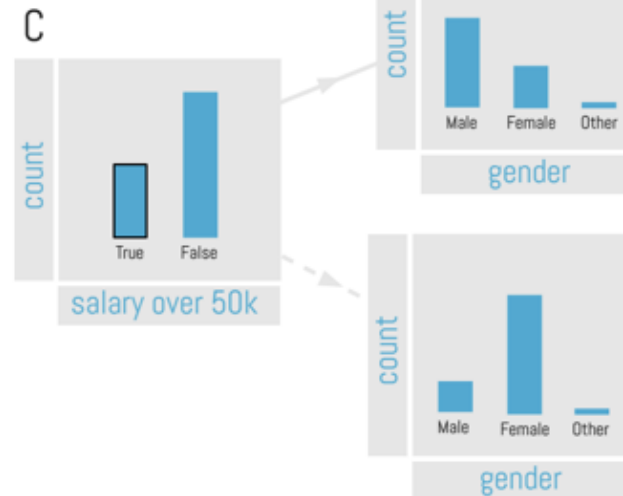
## □ Part 2: Storytelling

- What is a data narrative?
  - Definition and examples
  - Conceptual model
- Crafting process
  - focus on storytelling activities
- Automation
- Perspectives

# Part 1: Insights



# Exploratory Data Analysis



# A long time ago...

- ❑ An overview of data exploration techniques by Idreos et al. [SIGMOD 2015]
- ❑ Sheds a light on the different approaches to support the exploration of large datasets
- ❑ What is new since then?

## Overview of Data Exploration Techniques

Stratos Idreos  
Harvard University  
stratos@seas.harvard.edu

Olga Papaemmanouil  
Brandeis University  
olga@cs.brandeis.edu

Surajit Chaudhuri  
Microsoft Research  
surajitc@microsoft.com

### ABSTRACT

Data exploration is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for. In this tutorial, we survey recent developments in the emerging area of database systems tailored for data exploration. We discuss new ideas on how to store and access data as well as new ideas on how to interact with a data system to enable users and applications to quickly figure out which data parts are of interest. In addition, we discuss how to exploit lessons-learned from past research, the new challenges data exploration crafts, emerging applications and future research directions.

### 1. INTRODUCTION

**Assumptions in Traditional Systems.** Traditional data management systems assume that when users pose a query a) they have good knowledge of the schema, meaning and contents of the database and b) they are certain that this particular query is the one they wanted to pose. In short, we assume that users know what they are looking for. In response, the system always tries to produce correct and complete results.

Traditional DBMSs are designed for static scenarios with numerous assumptions about the workload. For example, state-of-the-art systems assume that there will be a tuning phase where a database administrator tunes the system for the expected workload. This assumes that we know the workload, we know that it will be stable and we have enough idle time and resources to devote to tuning.

**Modern Exploration-driven Applications.** The above assumptions were valid for the static applications of the past and they are still valid for numerous applications today. However, as we create and collect increasing amount of data, we are building more dynamic data-driven applications that do not always have the same requirements that database systems have tried to address during the past five decades. Indeed, managing an employee or an inventory database is a

drastically different setting than looking for interesting patterns over a scientific database.

Consider an astronomer looking for interesting parts in a continuous stream of data (possibly several TBs per day): they do not know what they are looking for, they only wish to find interesting patterns; they will know that something is interesting only after they find it. In this setting, there are no clear indications about how to tune a database system or how the astronomer should formulate their queries. Typically, an exploration session will include several queries where the results of each query trigger the formulation of the next one. This data exploration paradigm is the key ingredient for a number of discovery-oriented applications, e.g., in the medical domain, genomics and financial analysis.

**Database Systems for Data Exploration.** Such novel requirements of modern exploration driven interfaces have led to rethinking of database systems across the whole stack, from storage to user interaction. Visualization tools for data exploration (e.g., [38, 49, 66]) are receiving growing interest while new exploration interfaces emerged (e.g., [18, 32, 45, 57]) aiming to facilitate the user's interactions with the underlying database. In parallel, numerous novel optimizations have been proposed for offering interactive exploration times (e.g., [6, 36, 37]) while the database architecture has been re-examined to match the characteristics of the new exploration workloads (e.g., [8, 27, 28, 39]). Together, these pieces of work contribute towards providing data exploration capabilities that enable users to extract knowledge out of data with ease and efficiently.

**Tutorial Outline.** This tutorial gives a comprehensive introduction to the topic of data exploration, discussing state-of-the-art in the industry and in the academic world. Specifically, it includes the following sections.

**1. Introduction:** We start with an introduction of the concept of data exploration and an overview of the new challenges presented in the era of "Big Data" which make data exploration a first class citizen for query processing techniques. In this part, we also discuss the support available in today's products and services for data exploration techniques and what is still missing.

**2. User Interaction:** We take an in-depth look the advanced visualization tools and alternative exploration interfaces for big data exploration tasks. We further divide this last topic into three sub-categories: a) systems that assist SQL query formulation, b) systems that automate the data exploration process by identifying and presenting relevant data items and c) novel query interfaces such as keyword search queries over databases and gestural queries.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.  
SIGMOD '15, May 31–June 4, 2015, Melbourne, Victoria, Australia.  
Copyright © 2015 ACM 978-1-4503-2758-9/15/05 ...\$15.00.  
http://dx.doi.org/10.1145/2723372.2731084  
This work is partially supported by NSF grants IIS-1253196 and IIS-1452595.

# The 2015 tutorial

**Idreos, Papaemmanouil, Chaudhuri. Overview of Data Exploration Techniques.  
Sigmod 2015**

# Classification of approaches

<i>User Interaction</i>	<b>Data Visualization</b> [38]	Visual Optimizations [11, 12, 49, 66]	Visualization Tools [40, 48, 61, 62]	
	<b>Exploration Interfaces</b> [14]	Automatic Exploration [18, 20]	Assisted Query Formulation [3, 4, 13, 21, 52, 57, 58, 64, 51]	Novel Query Interfaces [32, 44, 45, 47]
<i>Middleware</i>	<b>Interactive Performance Optimizations</b>	Data Prefetching [36, 37, 41, 63]	Query Approximation [16, 5, 6, 7, 24, 25]	
<i>Database Layer</i> [27, 39]	<b>Indexes</b>	Adaptive Indexing [26, 29, 30, 31, 33, 22, 23, 50]	Time Series [68]	Flexible Engines [17, 42, 43, 34]
	<b>Data Storage</b>	Adaptive Loading [28, 8, 2, 15]	Adaptive Storage [9, 19]	Sampling [59, 60, 35]

- ❑ Distinguished contributions at 3 levels:
  - GUI
  - Middleware
  - DB engine

# Salient perspectives of the 2015 tutorial

- ❑ A system **should be able** to provide answers **instantly** even if they are not complete
  - Data system **architectures** **should** inherently support exploration
- ❑ The overall vision is to achieve data navigation systems that **automatically steer** users towards **interesting** data
- ❑ Still lack **declarative “exploration” languages** to present and reason about popular navigational idioms
  - Future directions include processing **past user interaction histories**

What is the problem?

Insights and interestingness

Human in the loop

# What is the problem?

**And how is it solved?**



# Approaches

- ❑ **Generate and select**

- [SIGMOD 2017, 2019, 2020, 2021, DOLAP 2020, EDBT 2022]

- ❑ **Guided EDA**

- [VLDB 2020, CIKM 2021]

# Chain Composite Items (CCI)

## Tutorial at [EDBT 2018]

- ❑ Retrieval of items that should be recommended together
  - E.g., travel itinerary recommendation
- ❑ Usually expressed as a **constrained optimization problem**
- ❑ Chain shaped CIs are traditionally defined in terms of:
  - **compatibility** (e.g., geographic distance),
  - **validity** (e.g., the total cost of an itinerary is within budget)
  - **maximality** (e.g., the itinerary should be of the highest value in terms of its POIs popularities),
    - ❑ often used as the objective function.
- ❑ Usually **NP-hard**
  - reduced to TSP or orienteering problems

Problem Name	CI Shape	Hardness	Algorithm Strategy
k-Package (RecSys 2010)	star	NP-hard	Fagin Style Algorithm – instance optimal
Chain retrieval (ICDE 2011)	chain	NP-hard	Rooted Orienteering type of greedy heuristic algorithm
KOR Query (VLDB 2012)	chain	NP-hard	Approximation algorithm with guarantees, greedy heuristic
k-CIs retrieval (TKDE 2014)	snowflake	NP-hard	Clustering based heuristic algo
Customized k-CIs retrieval (DSAA 17)	snowflake	NP-hard	Fuzzy clustering based heuristic
TourRec: Additive Tour (WSDM 2014)	chain	NP-hard	Traveling Salesman Problem (TSP)
TourRec: CoveringTour	snowflake	NP-hard	Dynamic programming based solutions
Star retrieval (SIGMOD 2010)	star	#P-Complete	random walk over lattice Reduction from the Set Cover. approximation algorithm
a. Maximal star		NP-hard	greedy heuristics
b. Summarization		NP-hard	
c. Diversification			

# Traveling Analyst Problem (TAP) [DOLAP 2020]

- ❑ Computation of a **sequence of interesting queries** over a dataset
  - Given a **time budget** on the execution **cost**
  - Minimizing the **distance between queries**
- ❑ Differs from the classical **orienteering** problem
  - Adds a **knapsack** constraint to it
  - No starting or finish point for the sequence
  - Distance cannot be made analogous to a time or to a physical distance
  - Impossible to merge action cost and travel time budget

Given

- ▶ a set of queries  $Q$
- ▶ a cost  $cost(q_i) \forall q_i \in Q$
- ▶ an interestingness score  $interest(q_i) \forall q_i \in Q$
- ▶ a metric  $dist(q_i, q_j) \forall q_i, q_j \in Q$
- ▶ a time budget  $\epsilon_t$

find a sequence  $\langle q_1, \dots, q_M \rangle$  of queries in  $Q$ , such that:

1.  $\max \sum_{i=1}^M interest(q_i)$
2.  $\sum_{i=1}^M cost(q_i) \leq \epsilon_t$
3.  $\min \sum_{i=1}^{M-1} dist(q_i, q_{i+1})$ .

TAP is strongly NP-hard [3]

# Finding Top-k insights [SIGMOD 2017]

- ❑ **Sibling group**: subspaces that differ on a single dimension only
- ❑ **Extractor**: basic analysis operation on a sibling group
- ❑ **Insight**: result of a composite extractor on a sibling group
- ❑ Problem: Given a dataset and **composite extractor depth**
  - find top-k insights with the **highest scores**
  - among all possible combinations of sibling groups, composite extractors, and insight types

Sib. group $SG(S, D_x)$	Measure $S_c.\mathcal{M}$	Derived measure $S_c.\mathcal{M}'$ for			
		Rank	%	$\Delta_{avg}$	$\Delta_{prev}$
$\langle 2010, F \rangle$	13	4	15%	-4.4	
$\langle 2011, F \rangle$	10	5	11%	-7.4	-3
$\langle 2012, F \rangle$	14	3	16%	-3.4	4
$\langle 2013, F \rangle$	23	2	27%	5.6	9
$\langle 2014, F \rangle$	27	1	31%	9.6	4

# Markov decision problem [VLDB 2020]

- ❑ The **guided EDA** problem: a Markov decision process
  - State: displays several sets of objects
  - Transition: the application of an **exploration action** to a chosen set
  - **Utility**: reward obtained by **transitioning**
- ❑ Exploration **session**: a sequence of exploration actions
- ❑ Exploration **policy**: a function that maps a state to an action
  - And generates an exploration
- ❑ Problem: finding a **policy that maximizes utility**

$$\pi^* = \operatorname{argmax}_{\pi} p\_utility(\pi, s_1, \mathcal{U}_t), \forall s_1 = \langle g_1, \mathcal{G}_{k1} \rangle$$

# Algorithmic strategies

- ❑ **Exact solutions**
  - Unfeasible for real life problems/datasets
- ❑ **Heuristics**
  - Greedy algorithms, dynamic programming or dedicated TSP strategies
- ❑ **Machine learning**
  - Active learning, reinforcement learning, etc.
  - Tutorial on Automating Exploratory Data Analysis via Machine Learning [SIGMOD 2020]
- ❑ **Pattern mining**
  - Survey on exploring data using patterns [DOLAP 2021]

Module	System Type	Exploration Type	Personalization
EDA Recommender Systems	Data-Driven	Tuples Recommendation [11], Data Cube/OLAP [20, 38], Visualizations [46, 48]	No
	Log-Based	SQL [13] OLAP [1, 17, 49],	Yes
	Hybrid	Generic EDA [29, 31]	Yes
Predicting/Modeling Users' Interest	Dynamic Measure Prediction (kNN-based Classification)	Generic EDA [28]	Yes
	Modeling (Active Learning)	SQL/ Tuples Recommendations [10, 18]	Yes
	Modeling (Learning-to-Rank)	Visualizations [26]	No
Fully-Automated EDA	Seq2seq RNN	Visualizations [9]	No
	Deep Reinforcement Learning	Generic EDA [2, 30]	No

Method	Approach	Applications
CAPE [14]	Contrast	Explaining queries
Data Auditor [9]	Coverage	Data quality analysis
Data X-ray [19]	Contrast	Data quality analysis
DIFF [2]	Contrast	Outlier analysis
Explanation tables [7]	Information	Feature selection
Macrobase [1]	Contrast	Outlier analysis
MRI [5]	Coverage	Explaining queries
RSExplain [15]	Contrast	Explaining queries
Scorpion [20]	Contrast	Outlier analysis
Shrink [10]	Information	Explaining queries
Smart Drilldown [11]	Coverage	Explaining queries
SURPRISE [16]	Information	Explaining queries



# In all cases...

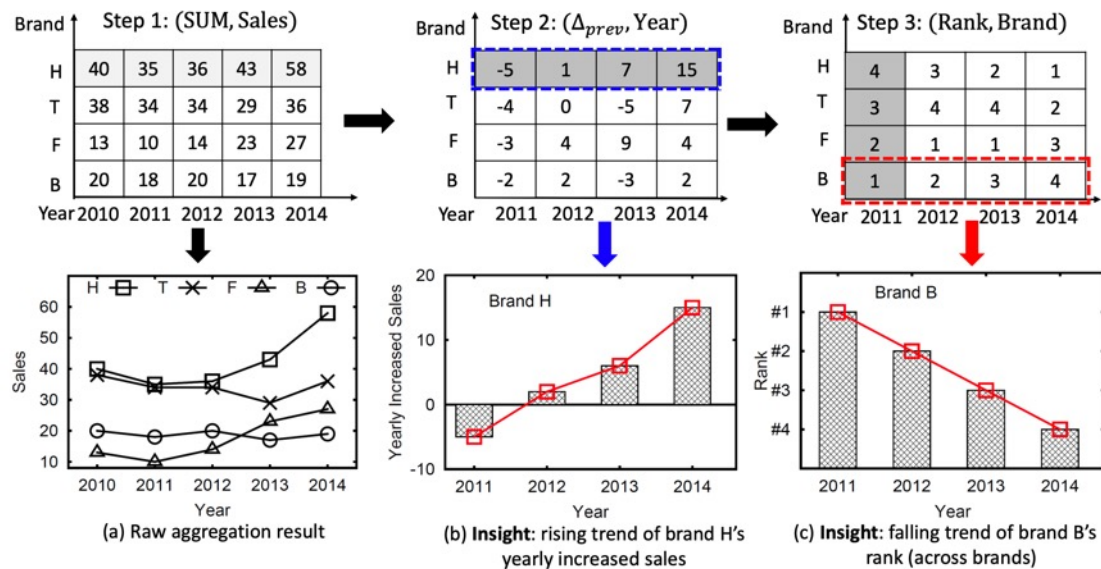
- ❑ There is a need to define
  - what an **insight** is
  - what its **interestingness** is...

# Insights

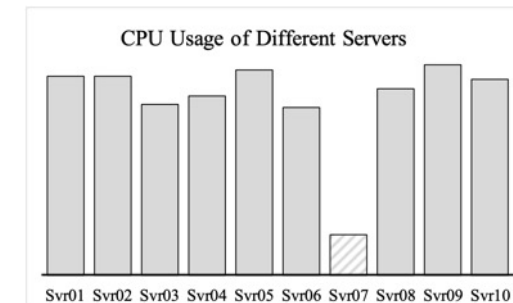
**And their interestingness**

# Different forms of insights

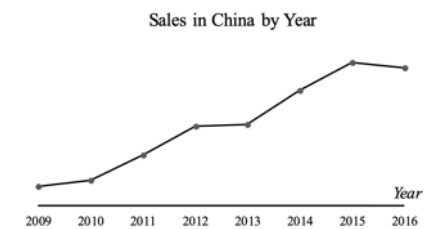
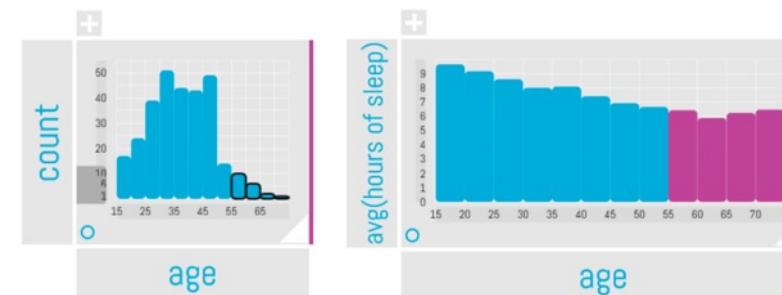
- ❑ [SIGMOD 2017, 2019, 2021, CHI 2018, EDBT 2021, 2022] (at least)
  - Many synonyms in the literature: insights, highlights, findings, discoveries, etc.



Continent	April	May
Africa	31598	92626
America	1104862	1404912
Asia	333821	537584
Europe	863874	608110
Oceania	2812	467



(a) "People over the age of 55 seem to sleep, on average, less than younger people."



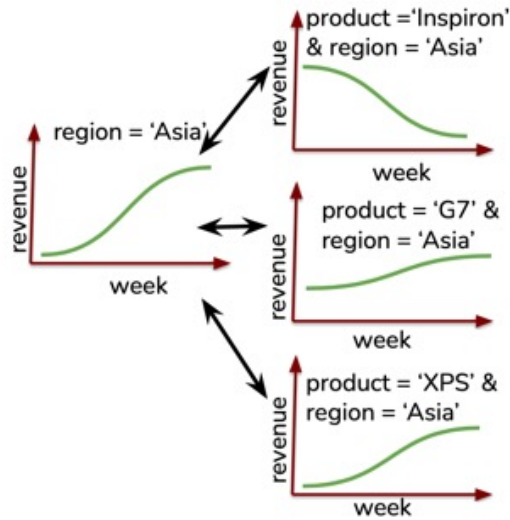
# A focus on comparison insights

(a) "People over the age of 55 seem to sleep, on average, less than younger people."

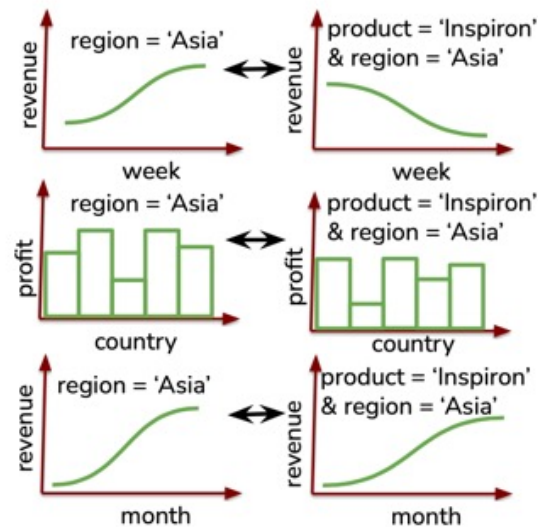


- ❑ One of the most popular [CHIRA 2020, VLDB 2021]
- ❑ 60% of spurious user-reported insights [CHI 2018]
  - Hence the need for systems able to automatically characterize insights

# Comparison insights

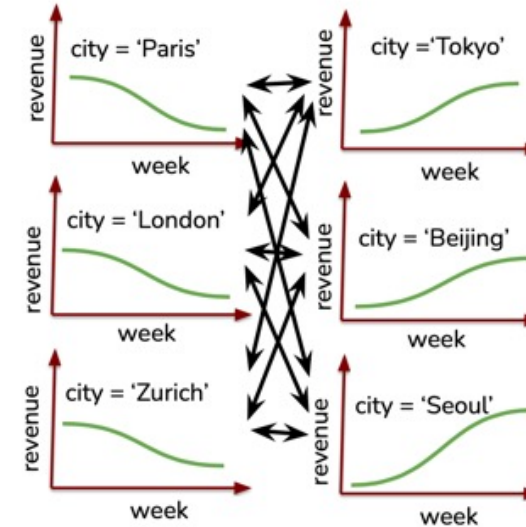


1a: One to many comparisons over fixed X and Y attributes

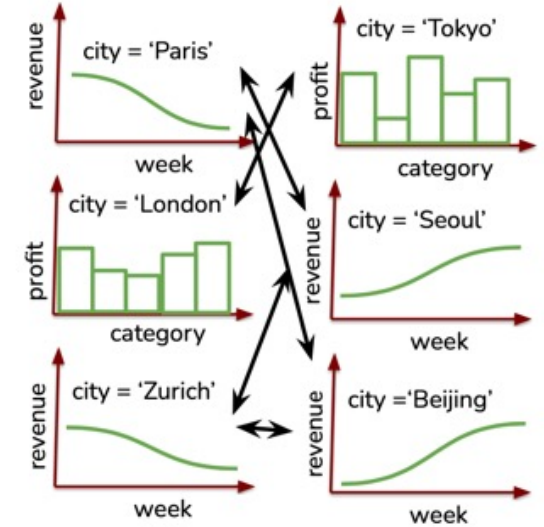


1b: One to one comparisons over varying X and Y attributes

↔ denotes comparison



2a: Many to many comparisons over fixed X and Y attributes



2b: Many to many comparisons over varying X and Y attributes

□ Plenty of them... [VLDB 2021]

# Generation of comparison notebooks [EDBT 2022]

Continent	April	May
Africa	31598	92626
America	1104862	1404912
Asia	333821	537584
Europe	863874	608110
Oceania	2812	467



On average, there were more cases in May compared to April

- ❑ Is this actually true?
- ❑ How to generate comparison queries that convey only statistically significant insights?



# Definition of comparison insight [EDBT 2022]

Extended relational algebra queries of the form:

$$\tau_A((\gamma_{A,agg}(M) \rightarrow val(\sigma_{B=val}(R))) \bowtie (\gamma_{A,agg}(M) \rightarrow val'(\sigma_{B=val'}(R))))$$

- ▶ over schema  $R[A_1, \dots, A_n, M_1, \dots, M_m]$
- ▶  $A, B$  are categorical attributes in  $\{A_1, \dots, A_n\}$
- ▶  $M$  is a measure attribute
- ▶  $agg$  is an aggregate function
- ▶  $val, val' \in dom(B)$

A tuple  $i = (M, B, val, val', p)$  where

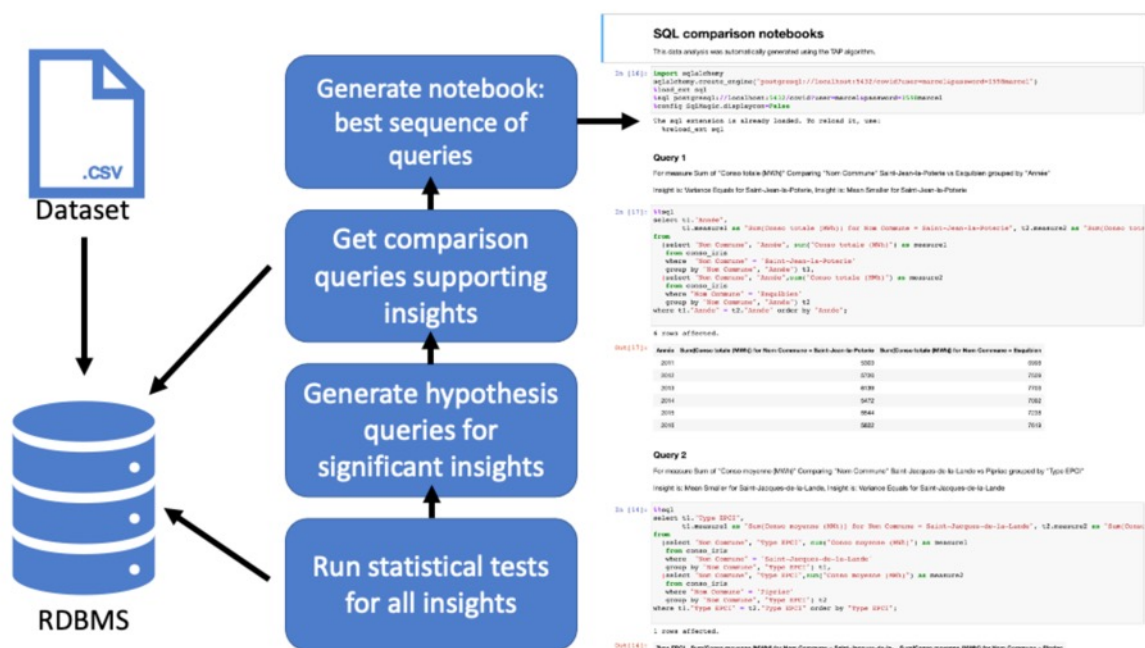
- ▶  $M$  is a measure attribute
- ▶  $B$  is a categorical attribute  $B$
- ▶  $val, val' \in Dom(B)$
- ▶  $p$  is a selection predicate
  - ▶  $avg(val) > avg(val')$  (*mean greater*),
  - ▶  $variance(val) > variance(val')$  (*variance greater*)

Each insight is associated with a statistical test

# Notebook generation approach [EDBT 2022]

## □ Bottleneck

- generate all hypothesis queries and run non parametric statistical tests



with comparison as

(select t1.continent, April , May  
from

(select month, continent, sum(cases) as April  
from covid where month = '4'  
group by month, continent) t1,  
(select month, continent, sum(cases) as May  
from covid where month = '5'  
group by month, continent) t2  
where t1.continent = t2.continent  
order by t1.continent)

select 'mean greater' as hypothesis from comparison  
having avg(April)<avg(May);

# Algorithmic strategy [EDBT 2022]

## Naive approach (sketch)

- ❑ **Generate all possible insights**
- ❑ **Loop over all insights**
  - Compute significance
  - If significant and supported by a hypothesis query
    - ❑ Add the comparison query to the set  $Q$
- ❑ **Solve the TAP for  $Q$**

Given

- ▶ a set of queries  $Q$
- ▶ a cost  $cost(q_i) \forall q_i \in Q$
- ▶ an interestingness score  $interest(q_i) \forall q_i \in Q$
- ▶ a metric  $dist(q_i, q_j) \forall q_i, q_j \in Q$
- ▶ a time budget  $\epsilon_t$

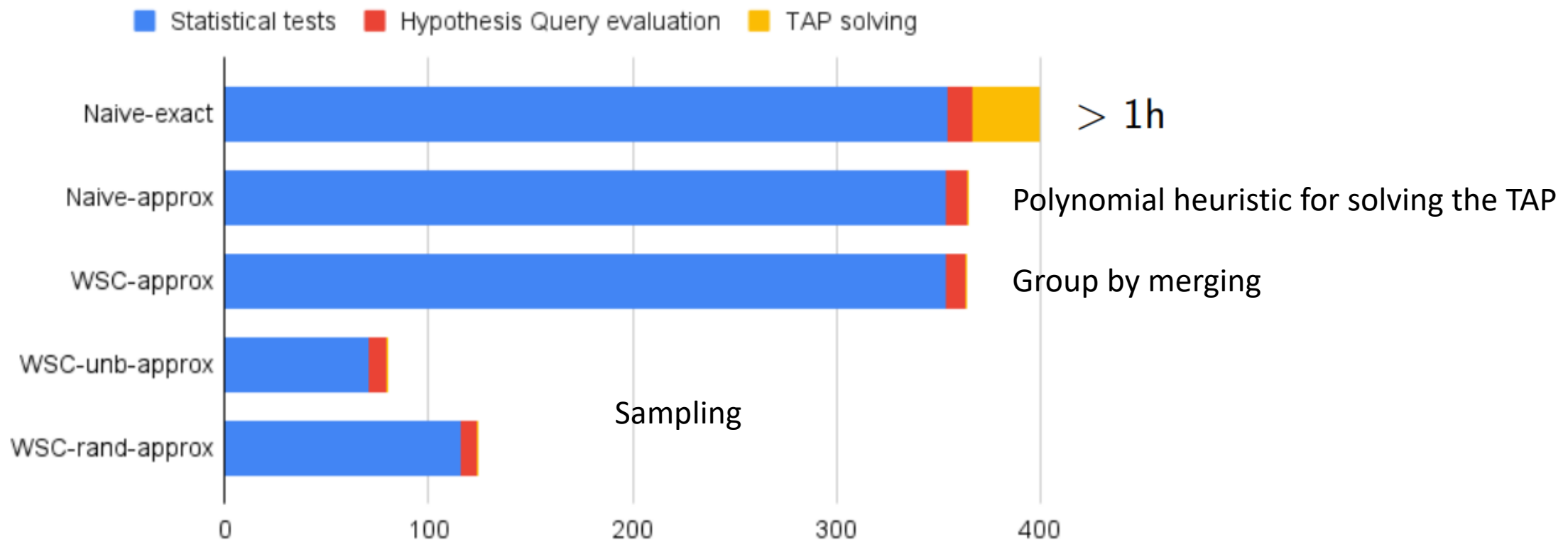
find a sequence  $\langle q_1, \dots, q_M \rangle$  of queries in  $Q$ , such that:

1.  $\max \sum_{i=1}^M interest(q_i)$
2.  $\sum_{i=1}^M cost(q_i) \leq \epsilon_t$
3.  $\min \sum_{i=1}^{M-1} dist(q_i, q_{i+1})$ .

TAP is strongly NP-hard [3]

# Various optimizations [EDBT 2022]

## Runtime breakdown (s)



# Insight interestingness

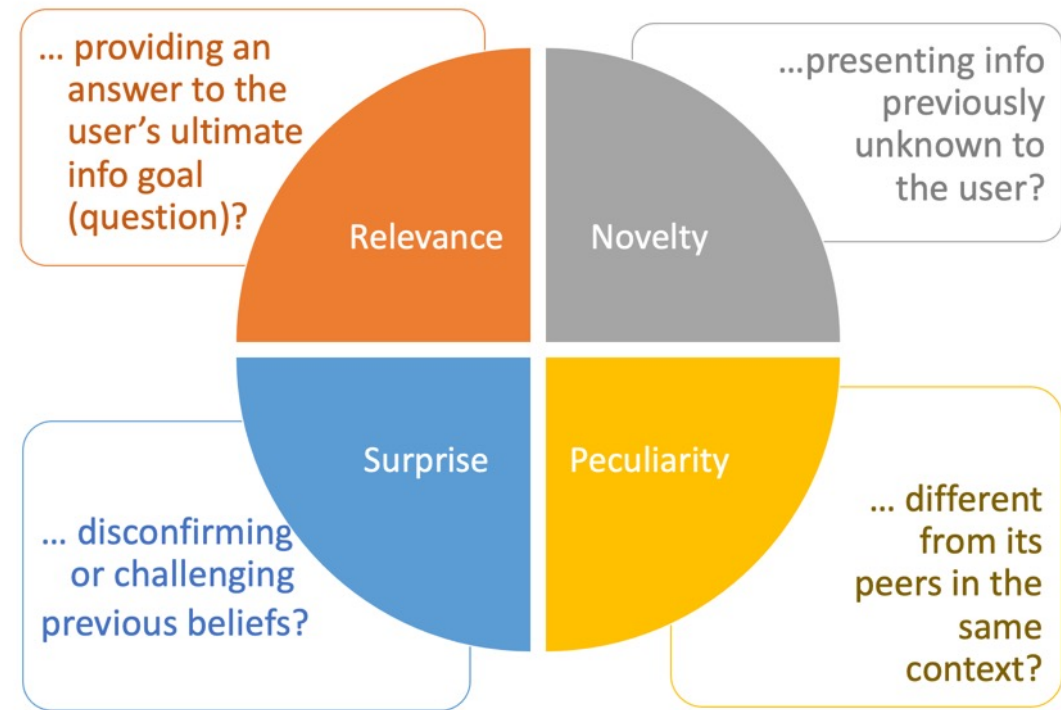
## ❑ 2 approaches

- Heuristic measures
  - ❑ Many heuristic measures proposed
    - Capturing a **different facet** of the **broad concept**
    - Helping to understand the **nature** of interestingness
  - ❑ But studies have shown that
    - there **is no single measure** that consistently outperforms the rest
    - interestingness is often **subjective** and changes **dynamically**
- Machine learning [SIGMOD 2020]
  - ❑ Dynamic selection of interestingness measures
  - ❑ ML-based models for users' interest
    - Active learning, learning-to-rank

# Interestingness (heuristic)

- ❑ 4 main dimensions of interestingness [ADBIS 2019]
  - **Relevance**: data vs goal
  - **Novelty**: data vs history
  - **Peculiarity**: data vs data
  - **Surprise**: data vs belief
- ❑ Misses some aspects
  - **Presentation**: is the insight presentation intelligible enough?

Interestingness aspects = To what extent is a piece of info ...





# Interestingness (heuristic) - exercise

I'd like to know more  
about Covid19 cases  
in Spring...

...I know nothing  
about it...

... but I expect that  
there is on average  
less cases as summer  
approaches...

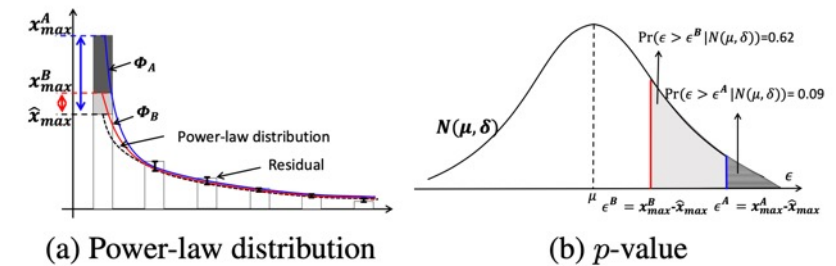


Continent	April	May
Africa	31598	92626
America	1104862	1404912
Asia	333821	537584
Europe	863874	608110
Oceania	2812	467

Relevance?  
Novelty?  
Peculiarity?  
Surprise?  
Presentational?

# Peculiarity: significance

- ❑ Insights turned into **hypothesis testing** [SIGMOD 2017, CHI 2019, EDBT 2022]
- ❑ Allows to:
  - Use **p-value** for significance
  - Define false discoveries (type 1 errors)
    - ❑ Visualization supporting a non significant insight
  - Define false omissions (type 2 errors)
    - ❑ Visualization non supporting a significant insight
  - Credibility
    - ❑ %age of visualizations supporting an insight
- ❑ The risk of type 1 error increases as more than one hypothesis is considered at once
  - **Correction** is needed

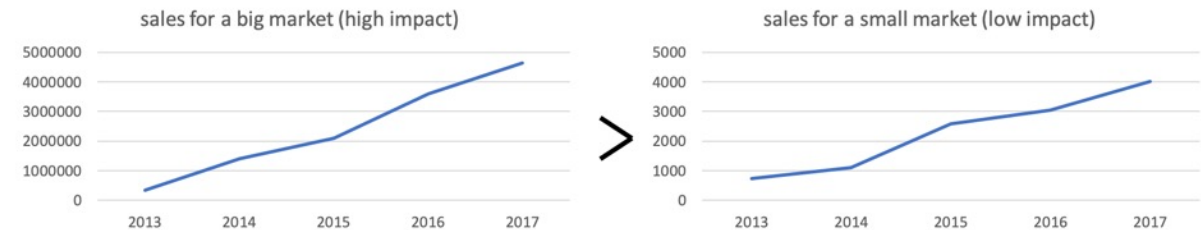


```
{
  "dimension": "hours_of_sleep",
  "dist_alt": "75 < age >= 55",
  "dist_null": "55 < age >= 15",
  "comparison": "mean_smaller"
}
```

Insight Class	Null Hypothesis	Permutation $\pi$	Test Statistic
Mean	$E[X] = E[Y]$	$X \cup Y$	$ \mu_X - \mu_Y $
Variance	$var(X) = var(Y)$	$X \cup Y$	$ \sigma_X^2 - \sigma_Y^2 $
Shape	$P(X Y = y_1) = P(Z Y = y_2)$	$Y$	$\ P(X Y = y_1) - P(Z Y = y_2)\ $
Correlation	$X \perp Y$	$X$	$ \rho(X, Y) $
Ranking	$X \sim Unif(a, b)$	$\pi \sim Unif(a, b)$	$\begin{cases} 1 & rank(X_\pi) = rank(X_{obs}) \\ 0 & \text{else.} \end{cases}$

# Peculiarity: coverage, impact

- Importance of the subject of an insight **against the entire dataset** [SIGMOD 2017, 2019, 2021]
- Anti-monotonic condition
  - if the subject of insight A is a superset of the subject of insight B, then impact of A should be no less than impact of B



$$\text{Impact}_{ds} = \frac{m_{\text{Impact}}(ds.\text{subspace})}{m_{\text{Impact}}(\{*\})} \in [0, 1]$$

# Peculiarity: coherency, distance

- ❑ **Coherency:** is a given EDA operation **coherent at a certain point?**
  - Learned with heuristic classification rules [SIGMOD 2020]
    - ❑ **General properties** of the operations sequence
      - E.g., a group-by on a continuous, numerical attribute is incoherent
    - ❑ Dependent on the input **dataset's semantics**
      - E.g., If the user focuses on flight delays, aggregating on the columns “departure-delay time” is preferred
- ❑ **Distance:** comparisons of exploration actions
  - Weighted Hamming distance of relational query parts [EDBT 2022]

$$\sigma_{que}(q, q') = \alpha \cdot \sigma_{gbs}(q, q') + \beta \cdot \sigma_{sel}(q, q') + \gamma \cdot \sigma_{meas}(q, q')$$

# Novelty: diversity, curiosity

- ❑ **Diversity:** induce new observations [SIGMOD 2020]
  - Minimal Euclidean distance between the current observation and **all previous displays obtained**
- ❑ **Curiosity:** going further in the exploration [CIKM 2021]
  - inversely proportional to the **number of times a result is encountered**
  - keep a counter for each seen result  $s$ 
    - ❑  $curiosity(s)=1/counter(s)$

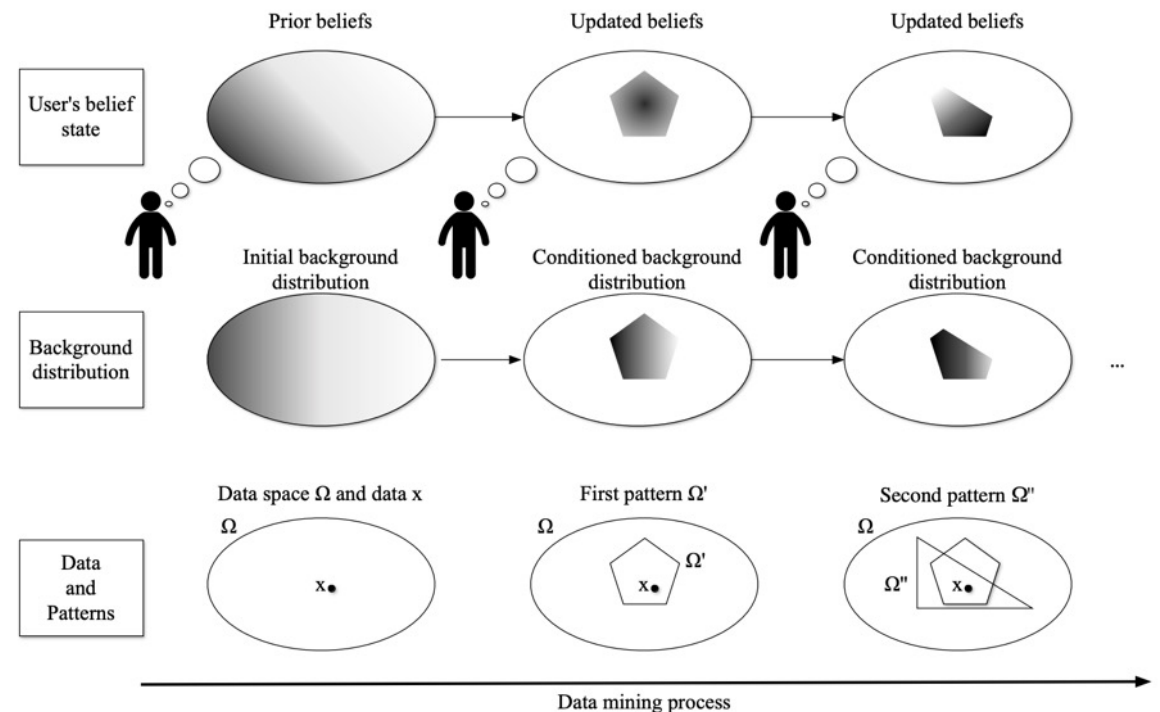
# Relevance: familiarity

- Familiarity : **concentration ratio of target objects in a set** [CIKM 2021]
  - Given a target set of **familiar objects**  $T$
  - Expected to be higher as the EDA session goes
    - to avoid "over-exploiting" a set of familiar objects
  - E.g., variant of the Jaccard index

$$Familiarity(s_i, T) = \sum_{O \in sets(s_i)} \frac{|O \cap T|^2}{|O| \times |T|}$$

# Surprise: information content

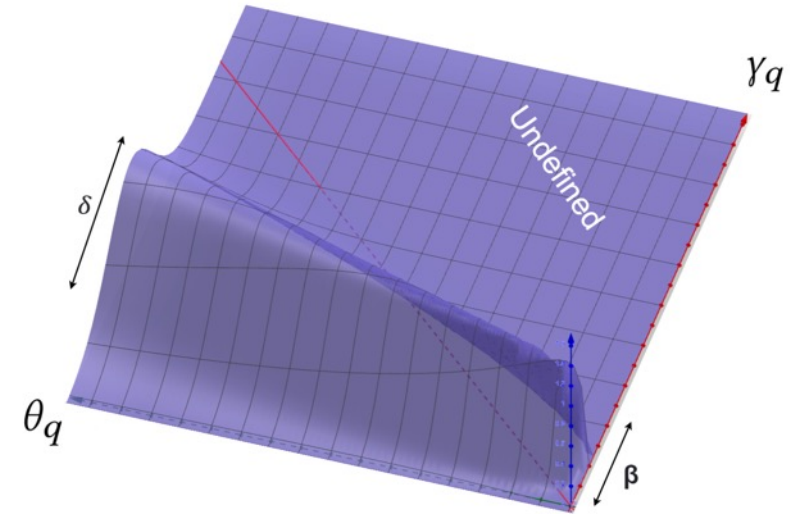
- ❑ **Information theoretic approach [IDA 2013]**
- ❑ Interactive exchange of information between data and user, accounting for the **user's prior belief state**
- ❑ **Background distribution:** probability measure over the exploration results
  - Approximates the belief that the user would attach to the result being expected





# Presentation: Conciseness, descriptonal complexity

- Conciseness: favoring insights being both **informative** and **easy to understand** [SIGMOD 2020, SIGMOD 2021, EDBT 2022]
  - E.g., compact group-by results covering many tuples
  - Sigmoid or non monotonic function of the number of groups and the number of the underlying tuples
- **Descriptonal complexity** [IDA 2013]
  - E. g., the more items a set contains, the more complex it is to assimilate



$$\text{conciseness}(\theta_q, \gamma_q) = e^{-\frac{1}{\theta_q \delta_q} (\gamma_q - \theta_q \alpha)^2}$$

# Interestingness: combination

## □ Product, weighted sum, ratio...

Interestingness		Relevance	Novelty	Peculiarity	Surprise	Presentation
Information content / descriptive complexity	IDA 2013				X	X
Significance * coverage	SIGMOD 2017, 2019			X		
(Conciseness or distance) + diversity + coherency	SIGMOD 2020		X	X		X
Familiarity + curiosity	CIKM 2021	X	X			
Novelty + peculiarity + surprise	IS 2021		X	X	X	
Conciseness * coverage	SIGMOD 2021			X		X
Significance * conciseness * credibility	EDBT 2022			X		X
Coverage + diversity	SIGMOD 2022		X	X		

# Human in the loop

**Declarative languages**

# Languages for EDA

## ❑ **EDA operations**

- ATENA
  - ❑ Filter, group, back
- DORA primitives
  - ❑ explore-around, explore-within, by facet, by-distribution, by-topic

## ❑ **Insight specific primitives**

- COMPARE, ASSESS, VCA: Comparison insight operators

## ❑ **High level algebra for analytical intentions**

- Intentional OLAP




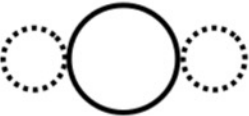
# ATENA [SIGMOD 2020]

- ❑ **3 atomic and easy ways to compose actions allowing to gradually form complex displays**
  - **FILTER: select** data tuples that match a criteria
  - **GROUP: group by** and **aggregate** the data
  - **BACK: backtrack** to the previous display to take an alternative exploration path

# DORA [CIKM 2021]

## □ Set-oriented higher-level exploration operations

- Find subsets that have the same value for some attributes
- Find sets that are similar to/different from an input set
- Find sets similar to/different from an input set in terms of their distributions
- Find  $k$  diverse sets that overlap with an input set
- Find  $k$  subsets that maximize the coverage of the input set

Operator	RCC8 Formalism [15]	Output description
$\text{by-facet}(D, A)$	NTPPi 	returns as many subsets of $D$ as there are combinations of values of attributes in $A$
$\text{by-superset}(D)$	NTPP 	returns the $k$ smallest supersets of input set $D$
$\text{by-distribution}(D)$	DC 	returns $k$ sets that are distinct from the input set $D$ and have the same distribution
$\text{by-neighbors}(D, a)$	EC 	returns 2 sets that are distinct from the input set $D$ and that have the previous and next values of attribute $a$

# COMPARE [VLDB 2021] and ASSESS [EDBT 2021]

- ❑ **COMPARE**: semantically equivalent to a relational expression consisting of multiple sub-queries with unions, group-bys, and joins
  - In DB engine
- ❑ **ASSESS**: semantics defined in terms of a logical cube algebra
  - middleware
- ❑ In both cases, logical/physical optimizations are proposed

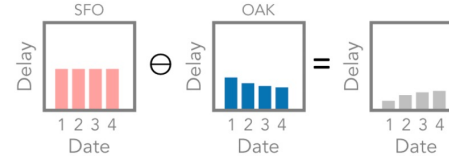
```
SELECT R1, P, W, V, score
FROM sales R
COMPARE [((R.region = Asia) AS R1) <--> (R1, R.product AS P)]
          [R.week AS W, AVG (R.revenue) AS V]
USING SUM OVER DIFF(2) AS score
```

```
with SALES
for year = '2019', product = 'milk'
by year, product
assess quantity against 1000
using ratio(quantity, 1000)
labels {[0, 0.9): bad, [0.9, 1.1]: acceptable, (1.1,inf): good}
```

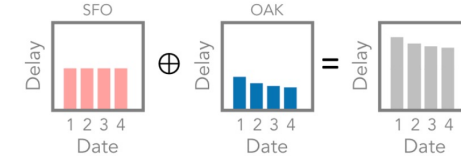


# VCA: View Comparison Algebra [TVCG 2022]

Set of composition operators that **summarize, compute differences, merge, and model** their operands



(a) Statistical Composition: minus



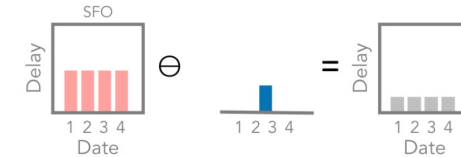
(b) Statistical Composition: add



(c) Union: juxtaposition



(d) Union: superposition



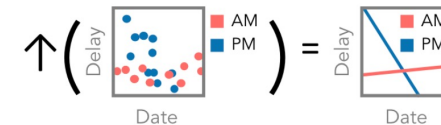
(e) Nonexact Schema Composition



(f) Nonexact Schema Composition



(g) Viewset Statistical Composition

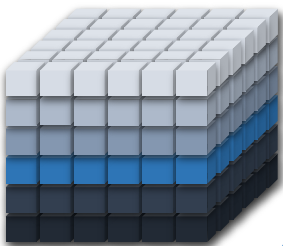


(h) Lift

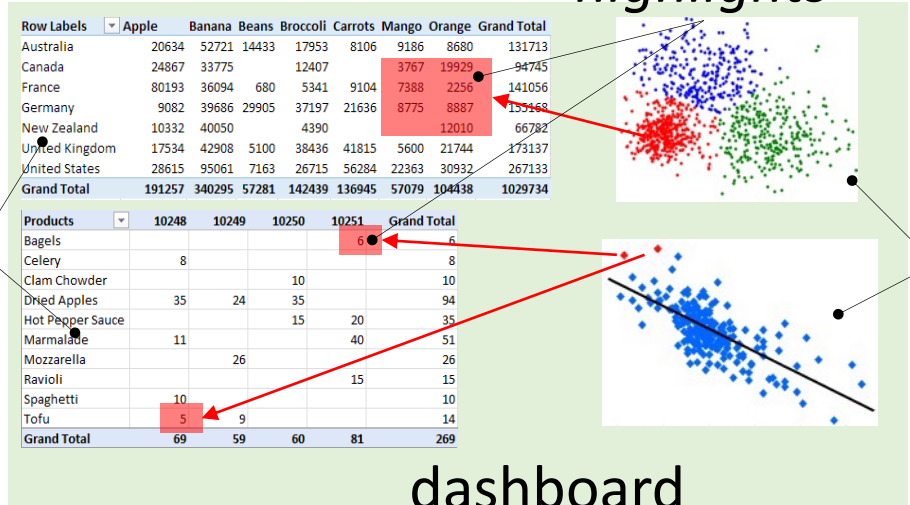
Name	Arity	Notation	Description
Stat Comp	Binary	$\odot_h(V_1, V_2)$	Compute difference of matching rows.
Union	Binary	$\cup(V_1, V_2)$	Superpose or Juxtapose marks.
Stat Comp	Nary	$\odot_f(\{V_1, \dots\})$	Aggregate matching rows from set of views.
Union	Nary	$\cup(\{V_1, \dots\})$	Superpose or Juxtapose marks.

Name	Arity	Notation	Description
Extract	Unary	$\downarrow(V, p)$	Derive subview w/ rows matching predicate $p$ .
Explode	Unary	$\Xi_A(V)$	Facet into small multiples w/ attributes $A$ .
Lift	Unary	$\uparrow(V)$	Fit model to view data.

# Intentional OLAP [IS 2019]



data



dashboard

Till now...	We advocate...
Query Operators: which <b>data</b> to bring...	Query Operators: user <b>intentions</b> ...
... as explicitly dictated by the <b>user</b> ...	... <b>automatically translated</b> by the <b>system</b> to queries
...for an answer being a set of <b>tuples</b>	...for an answer being a set of <b>tuples</b> + <b>models</b> + <b>highlights</b>

NEW answers:

Data

Models

Highlights

# In summary

- ❑ **Some answers were brought to the 2015 tutorial's perspectives**
- ❑ **In terms of:**
  - Automation
  - Interestingness
  - Languages
  - In-DB engine support

# End of part 1, thank you for your attention!

Wake-up, it's Q&A time 😊



Also, we have PhD positions, so...



# With a little help from...



Marie  
Chagnoux



Thomas  
Devogele



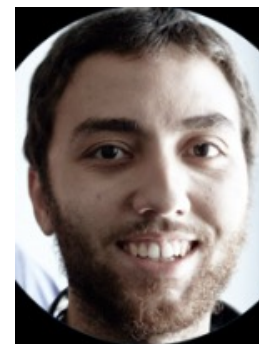
Matteo  
Golfarelli



Nicolas  
Labroche



Stefano  
Rizzi



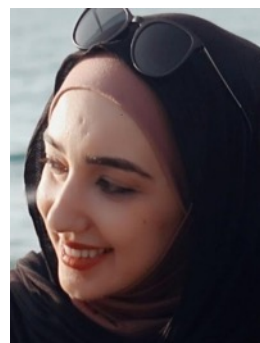
Raphaël  
da Silva



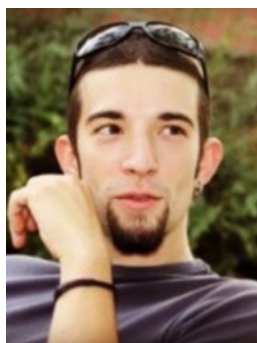
Panos  
Vassiliadis



Alexandre  
Chanson



Faten  
El Outa



Matteo  
Francia



Lucile  
Jacquemart



Raymond  
Ondzigue Mbenga



Patrick.Marcel@univ-tours.fr  
Veronika.Peralta@univ-tours.fr

# References



# Introduction

- ❑ [Tukey 1977] John W. Tukey. Exploratory Data Analysis. Pearson, 1977.
- ❑ [SIGMOD 2015] Stratos Idreos, Olga Papaemmanouil, Surajit Chaudhuri: Overview of Data Exploration Techniques. SIGMOD 2015: 277-281
- ❑ [CACM 2022] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, Christopher K. I. Williams: Automating Data Science: Prospects and Challenges. CACM 65(3): 76-87

# What is the problem?

- ❑ [SIGMOD 2017] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, Dongmei Zhang: Extracting Top-K Insights from Multi-dimensional Data. SIGMOD 2017: 1509-1524
- ❑ [EDBT 2018] Sihem Amer-Yahia, Senjuti Basu Roy: Interactive Exploration of Composite Items. EDBT 2018: 513-516
- ❑ [DOLAP 2020] Alexandre Chanson, Ben Crulis, Nicolas Labroche, Patrick Marcel, Verónica Peralta, Stefano Rizzi, Panos Vassiliadis: The Traveling Analyst Problem: Definition and Preliminary Study. DOLAP 2020: 94-98
- ❑ [VLDB 2020] Mariia Seleznova, Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Eric Simon: Guided Exploration of User Groups. VLDB 13(9): 1469-1482 (2020)
- ❑ [SIGMOD 2020] Tova Milo, Amit Somech: Automating Exploratory Data Analysis via Machine Learning: An Overview. SIGMOD 2020: 2617-2622
- ❑ [DOLAP 2021] Lukasz Golab, Divesh Srivastava: Exploring Data Using Patterns: A Survey and Open Problems. DOLAP 2021: 116-120



# Insights (1)

- ❑ [IDA 2013] Tijl De Bie. Subjective interestingness in exploratory data mining. IDA 2013: 19–31.
- ❑ [SIGMOD 2017] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, Dongmei Zhang: Extracting Top-K Insights from Multi-dimensional Data. SIGMOD 2017: 1509-1524
- ❑ [CHI 2018] Emanuel Zgraggen, Zheguang Zhao, Robert C. Zeleznik, Tim Kraska: Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. CHI 2018: 479
- ❑ [SIGMOD 2019] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, Dongmei Zhang: QuickInsights: Quick and Automatic Discovery of Insights from Multi-Dimensional Data. SIGMOD 2019: 317-332
- ❑ [ADBIS 2019] Patrick Marcel, Verónica Peralta, Panos Vassiliadis: A Framework for Learning Cell Interestingness from Cube Explorations. ADBIS 2019: 425-440
- ❑ [SIGMOD 2020] Ori Bar El, Tova Milo, Amit Somech: Automatically Generating Data Exploration Sessions Using Deep Reinforcement Learning. SIGMOD 2020: 1527-1537

# Insights (2)

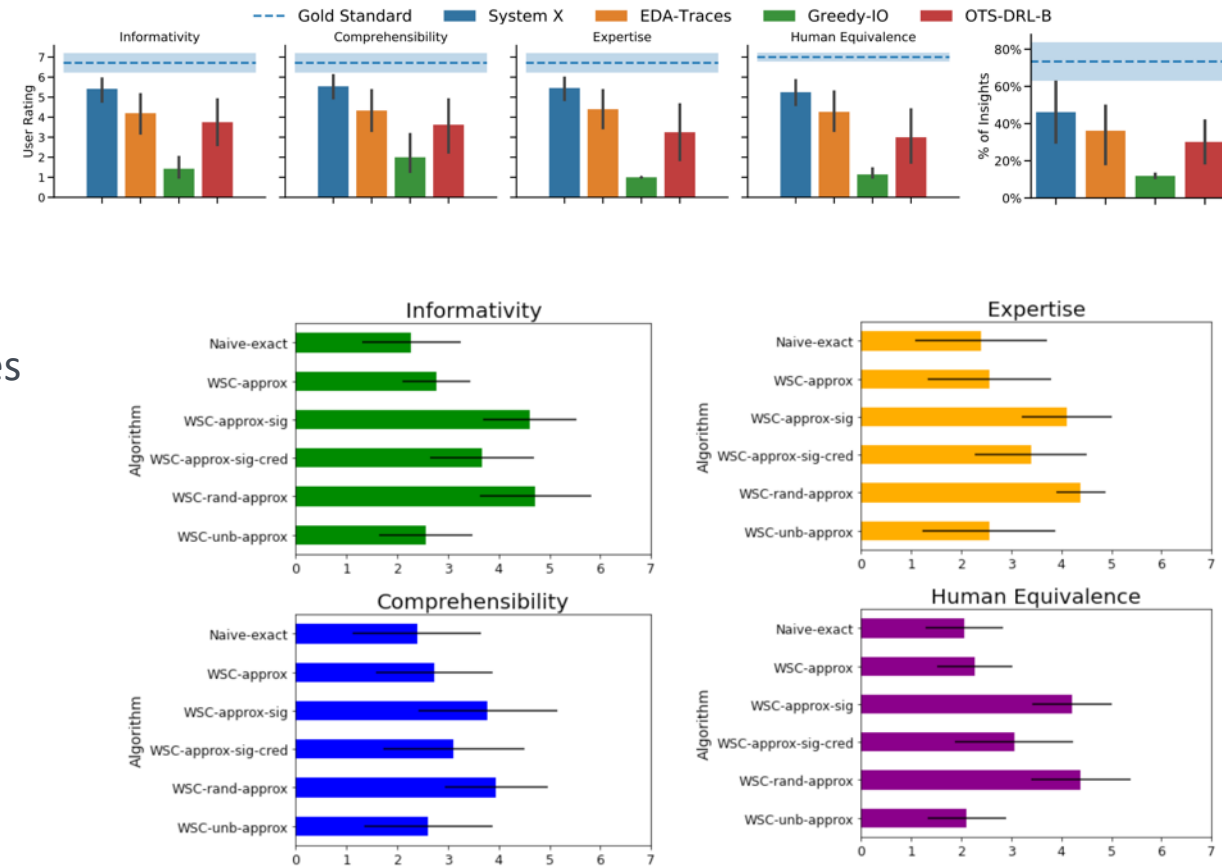
- ❑ [CHIRA 2020] Tom Blount, Laura Koesten, Yuchen Zhao, Elena Simperl: Understanding the Use of Narrative Patterns by Novice Data Storytellers. CHIRA 2020: 128-138
- ❑ [SIGMOD 2021] Pingchuan Ma, Rui Ding, Shi Han, Dongmei Zhang: MetaInsight: Automatic Discovery of Structured Knowledge for Exploratory Data Analysis. SIGMOD 2021: 1262-1274
- ❑ [CIKM 2021] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, Srividya Subramanian: DORA THE EXPLORER: Exploring Very Large Data With Interactive Deep Reinforcement Learning. CIKM 2021: 4769-4773
- ❑ [VLDB 2021] Tarique Siddiqui, Surajit Chaudhuri, Vivek R. Narasayya: COMPARE: Accelerating Groupwise Comparison in Relational Databases for Data Analytics. VLDB 14(11): 2419-2431 (2021)
- ❑ [EDBT 2021] Matteo Francia, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, Panos Vassiliadis: Assess Queries for Interactive Analysis of Data Cubes. EDBT 2021: 121-132
- ❑ [EDBT 2022] Alexandre Chanson, Nicola Labroche, Patrick Marcel, Stefano Rizzi, Vincent T'Kindt: Automatic generation of comparison notebooks for interactive data exploration. EDBT 2022: 2:274-2:284
- ❑ [SIGMOD 2022] Kathy Razmadze, Yael Amsterdamer, Amit Somech, Susan B. Davidson, Tova Milo. SubTab: Data Exploration with Informative Sub-Tables. SIGMOD 2022: 2369-2372

# Human in the loop

- ❑ [IS 2019] Panos Vassiliadis, Patrick Marcel, Stefano Rizzi: Beyond roll-up's and drill-down's: An intentional analytics model to reinvent OLAP. Inf. Syst. 85: 68-91 (2019)
- ❑ [SIGMOD 2020] Ori Bar El, Tova Milo, Amit Somech: Automatically Generating Data Exploration Sessions Using Deep Reinforcement Learning. SIGMOD 2020: 1527-1537
- ❑ [SIGMOD 2020] Philipp Eichmann, Emanuel Zgraggen, Carsten Binnig, Tim Kraska: IDEBench: A Benchmark for Interactive Data Exploration. SIGMOD 2020: 1555-1569
- ❑ [VLDB 2021] Tarique Siddiqui, Surajit Chaudhuri, Vivek R. Narasayya: COMPARE: Accelerating Groupwise Comparison in Relational Databases for Data Analytics. VLDB 14(11): 2419-2431 (2021)
- ❑ [CIKM 2021] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, Srividya Subramanian: DORA THE EXPLORER: Exploring Very Large Data With Interactive Deep Reinforcement Learning. CIKM 2021: 4769-4773
- ❑ [EDBT 2022] Alexandre Chanson, Nicola Labroche, Patrick Marcel, Stefano Rizzi, Vincent T'Kindt: Automatic generation of comparison notebooks for interactive data exploration. EDBT 2022
- ❑ [TVCG 2022] Eugene Wu: View Composition Algebra for Ad Hoc Comparison. IEEE Trans. Vis. Comput. Graph. 28(6): 2470-2485 (2022)

# Evaluation

- ❑ Qualitative human evaluation
  - [SIGMOD 2020, EDBT 2022]
- ❑ Users inspect automatically generated notebooks and rate them according to
  - **Informativity**
    - ❑ How informative the notebook is and how well does it capture dataset highlights?
  - **Comprehensibility**
    - ❑ To what degree is the notebook comprehensible and easy to follow?
  - **Expertise**
    - ❑ What is the level of expertise of the notebook composer?
  - **Human Equivalence**
    - ❑ How closely the notebook resembles a human-generated session?



# Evaluation

- ❑ IDEBench: A Benchmark for Interactive Data Exploration [SIGMOD 2020]
- ❑ Well adapted for approximate query answering
- ❑ Main metrics:
  - **Time Requirement Violations:** boolean value indicating whether a query exceeded the time requirement
  - **Mean relative error:** error between the latest result of an approximate aggregate query and its ground-truth
  - **Missing Bins/Groups:** completeness for an aggregate query result
  - **Cosine:** how much the “shape” of an aggregate result deviates from the shape of the groundtruth

	<i>TPC-H</i>	<i>TPC-DS</i>	<i>SSB</i>	<i>IDEBench</i>
<i>Schema</i>	snowflake	snowflake	star	star (default)
<i>Data Origin</i>	synthetic	synthetic	synthetic	real-world
<i>Data Distributions</i>	uniform	skewed	uniform	real-world
<i>Data Scaling</i>	yes	yes	yes	yes
<i>Iterative Query Formulation</i>	no	4 out of 99	no	yes
<i>Multi-Query Execution</i>	no	no	no	yes
<i>Think Time</i>	no	no	no	yes
<i>Metrics</i>	Time-based	Time-based	Time-based	Quality, Time

# Next hot topics in EDA and data narrations

## ❑ Integrated approaches to

- Explore and analyze datasets
- And then craft, share, query, reuse data narratives

## ❑ Explainability

- Explanations of insights
- Data narratives as explanations

## ❑ Personalization

- Leveraging user's preferences, background knowledge, intentions
- Monitoring their learning curves
- Personalization of data narratives

Example of research challenges from [CACM 2022]

- Generating **collaborative** reports and presentations, facilitating the **interrogation, validation** and **explanation** of models and results.
- Doing data science as **querying** or **programming** may help bridge the composition and mechanization forms of automation.