

# AI 驱动 软件研发 全面进入数字化时代

中国·北京 08.18-19

AI+  
software  
Development  
Digital  
summit



## InsightPilot: Towards LLM-Empowered Automated Data Exploration

丁锐 微软

# 科技生态圈峰会 + 深度研习 ——1000+ 技术团队的选择



2023K+  
全球软件研发行业创新峰会  
上海站

会议时间 | 06.09-10



2023K+  
全球软件研发行业创新峰会  
北京站

会议时间 | 07.21-22



2024K+  
全球软件研发行业创新峰会  
深圳站

会议时间 | 05.17-18



K+峰会详情



会议时间 | 08.18-19

AiDD AI+软件研发数字峰会  
北京站



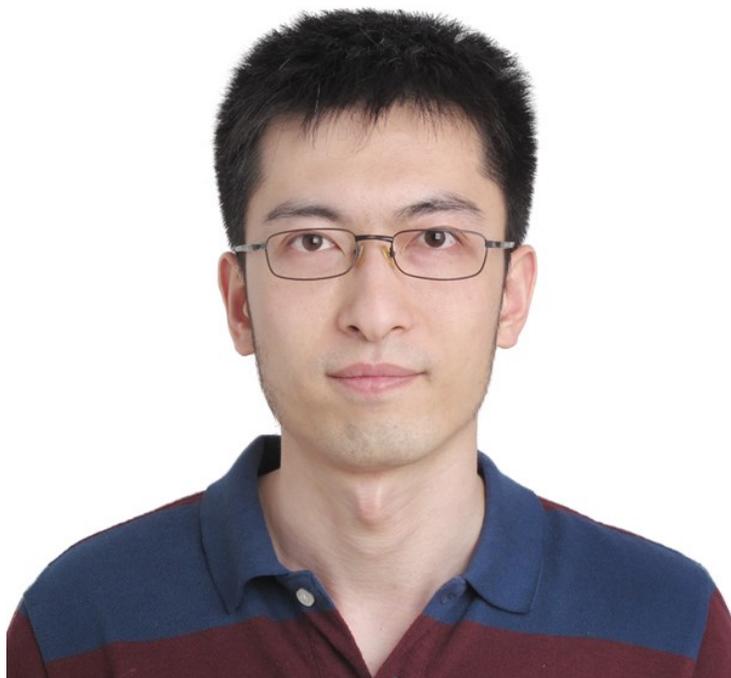
会议时间 | 11.17-18

AiDD AI+软件研发数字峰会  
深圳站



AiDD峰会详情

## ▶ 演讲嘉宾



### 丁锐

微软 首席研究员

丁锐是微软的数据、知识和智能（DKI）团队的首席研究员。丁锐一直致力于数据分析当中的洞见（insights）研究，这对于在商业和日常生活中理解数据及有效决策至关重要。丁锐的研究主要集中在两个主题上。第一个主题是如何将洞见概念转化为可计算的数据实体，这是洞见发现（即检测和挖掘）的基础问题。另一个主题是数据分析的可解释性以及因果性在其中的重要角色，这也是使洞见具有解释性，可靠性及泛化性的关键。丁锐的研究成果主要发表在SIGMOD和SIGKDD等会议上。此外，与洞见相关的研究在微软也有一系列产品转化，作为微软产品中用于数据分析的功能，包括Power BI的QuickInsights、Excel的Analyze Data和Forms Insights。

# 目录

## CONTENTS

1. **Insight-Based Exploratory Data Analysis**
2. **InsightPilot: LLM-Empowered Automated Data Analysis Paradigm**

## **PART 01**

# **Insight-Based Exploratory Data Analysis**

# ▶ Outline

- The concept and formulation of insight
- The analysis space established from insight
  - MetaInsight: Enriching the intension of insight
  - XInsight: Expanding extension of insight

# ▶ What is Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a process of analyzing data to summarize its main characteristics, for the purpose of

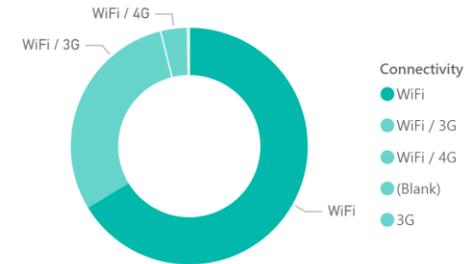
- Gaining knowledge from data
- Facilitating further in-depth data analysis



# ▶ Importance of Interesting Pattern Discovery

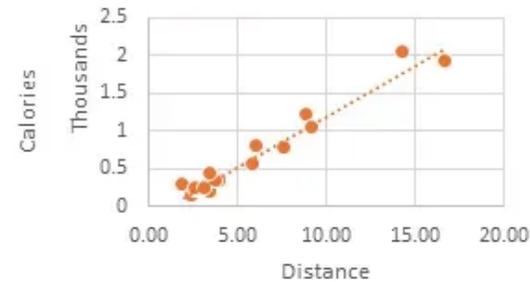
- Increasingly popular in the era of big data
- Values of discovering interesting data pattern
  - Data understanding
  - Knowledge discovery
  - Further drill-down data analysis
- Common practice
  - Exploratory data analysis
  - Visual/interactive data analysis

Value (US\$M) (USD)  
BY CONNECTIVITY

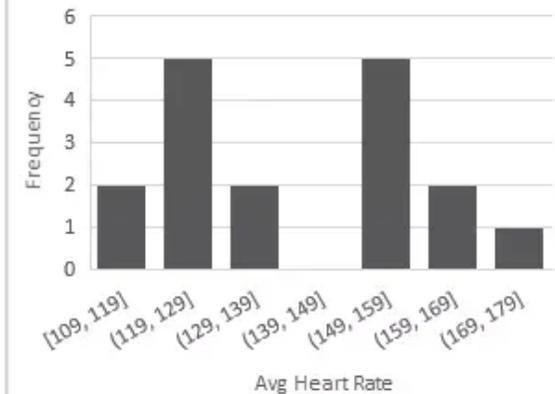


'WiFi' accounts for the majority of Value (US\$M) (USD) for OS 'Android'.

Field: **Distance** and Field: **Calories** appear highly correlated.



Frequency of 'Avg Heart Rate'



# ► Challenges of Interesting Pattern Discovery

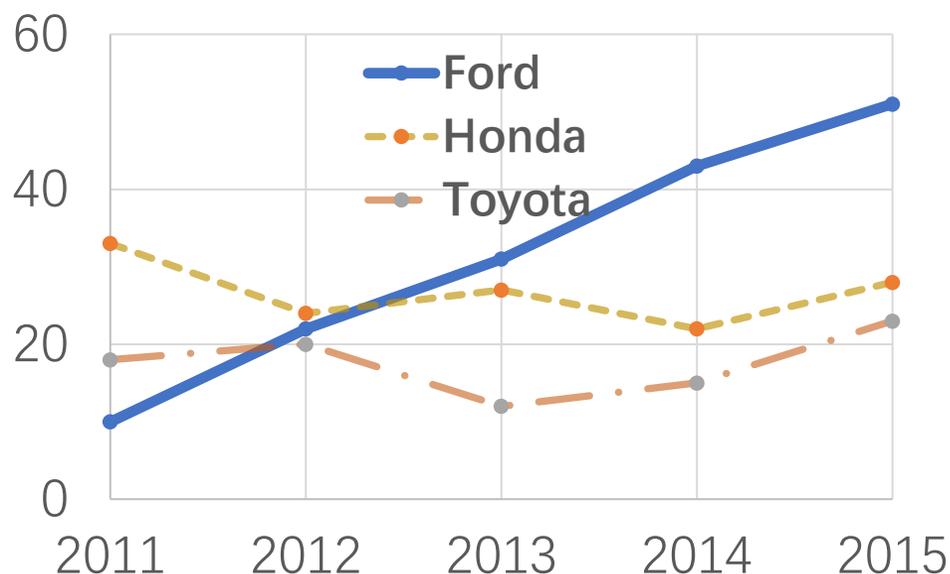
- Existing work
  - Mainly focus on dealing with individual types of interesting patterns
  - Lack of unified formulation of “interesting patterns”
  - Lack of general mining frameworks
- User’ s target is often broad or vague
  - Insights hidden amongst subspaces across different semantic levels
    - E.g., Ford vs. Ford SUV vs. Ford SUV China
  - Insights hidden amongst different measure columns
    - E.g., price, volume, revenue

## ▶ Key problems to be solved

- **Insight Definition:** What are the general abstraction & tangible form of interesting patterns?
- **Insight Scoring:** What are the factors & criteria for quantifying insight?
- **Insight Mining:** How to discover desirable insights efficiently?

## ▶ Intuition of insight concept (I)

An insight typically reflects the **interestingness** of a **subject** or a **relationship** among a set of subjects from certain **perspective**



Example: Ford's sales is increasing steadily over years

- Subject(s): Sales of Ford, grouped-by Year
- Perspective: trend
- Interestingness: increasing steadily

## ▶ Intuition of insight concept (II)

- Analysis entity: abstraction of subject or relationship among subjects
  - Specifies the content of interests for data analysis purpose
    - E.g., Brand = Ford, Measure = Sales, Group-by = Year
  - Corresponds to a specific raw data distribution
    - E.g., (2011, 15,000), (2012, 21,000), (2013, 27,000), (2014, 32,000), (2015, 38,000)
- Analysis semantic: facilitating analysis needs
  - Captures essential characteristics of raw data distributions
    - E.g., a trend is appeared in the time series
  - In the form of symbolic representation
    - E.g., Increasing = True, Steadiness = True

# ▶ Mapping to Multi-Dimensional Data Model

- Analysis entity: a 3-tuple  $\langle \text{Subspace}, \text{Breakdown}, \text{Measure} \rangle$ 
  - Indicates a sibling group with corresponding aggregate values on the measure.
  - Corresponds to a data cube, with the raw data distribution
- Analysis semantic
  - Perspective: Materialize as different insight types
  - Essential characteristics: captured by insight properties
- Interestingness:  $\text{Evaluate}(AE, type) = \text{true}$

Note: AE is short for 'Analysis Entity'

## ▶ Insight definition

- In multi-dimensional data model, an insight is a tuple of

$$Insight := \langle AE, Type, Property \rangle$$

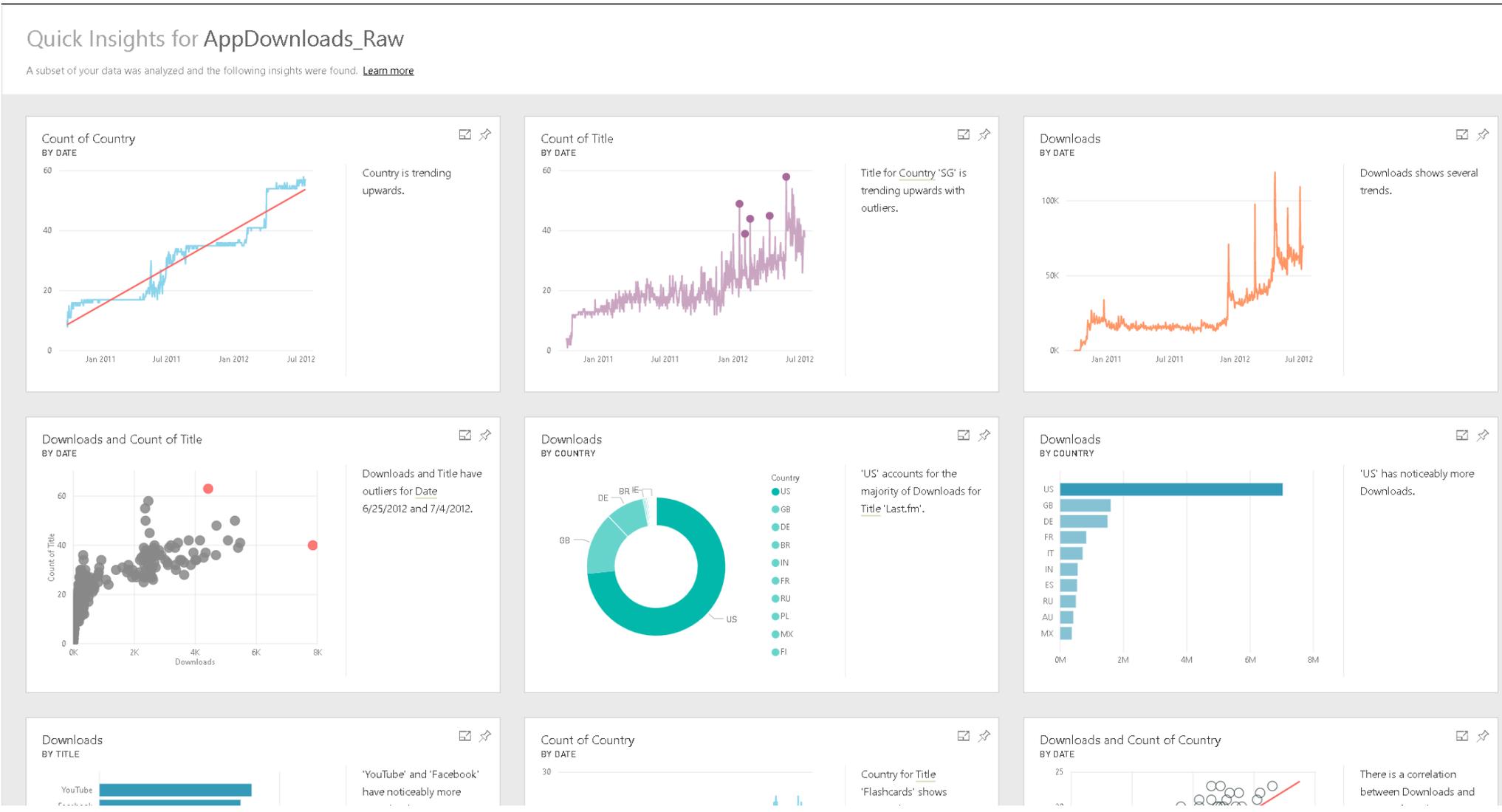
$$\text{where } AE := \langle Subspace, Breakdown, Measure \rangle$$

with non-trivial (significant) evaluation result

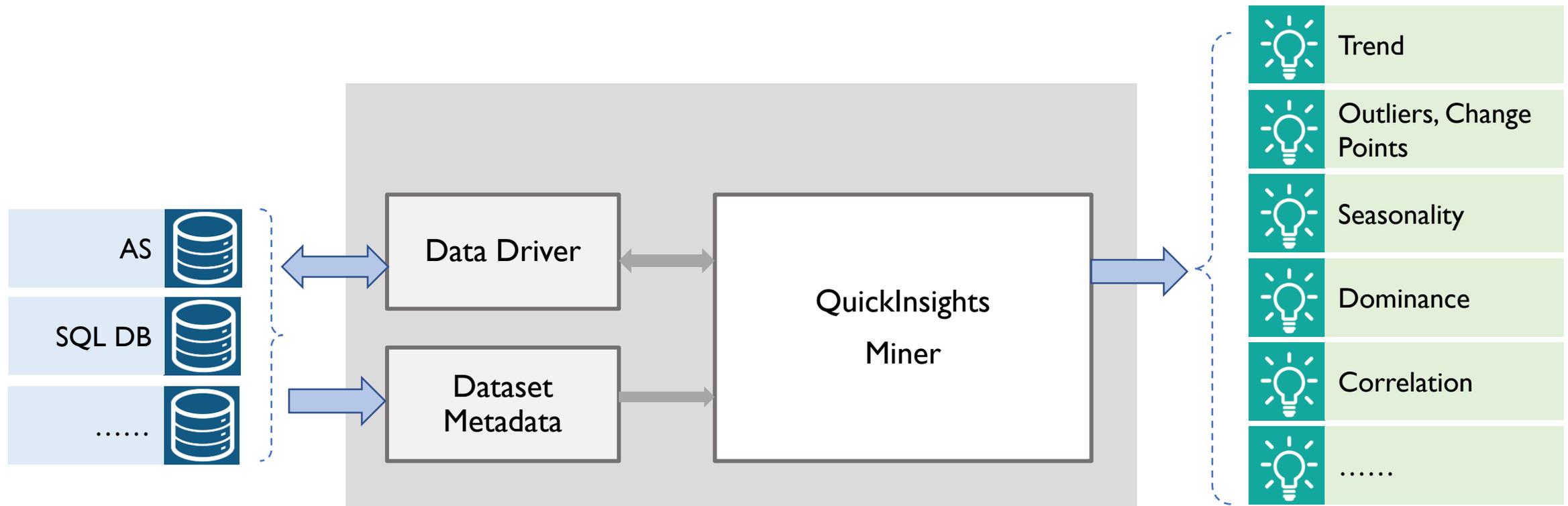
# ▶ The pivoting role of AE (Analysis Entity)

- It reflects the typical query operation in OLAP
  - Filtering / group-by / aggregation
- It acts as a bridge between data manipulation operations and insights
  - Insight is evaluated from an AE against a specific perspective
- It has a natural mapping to visual charts
  - X-axis values: values of breakdown dimension
  - Y-axis values: aggregation values from the measure
  - Filter: the subspace

# ▶ Example of insights

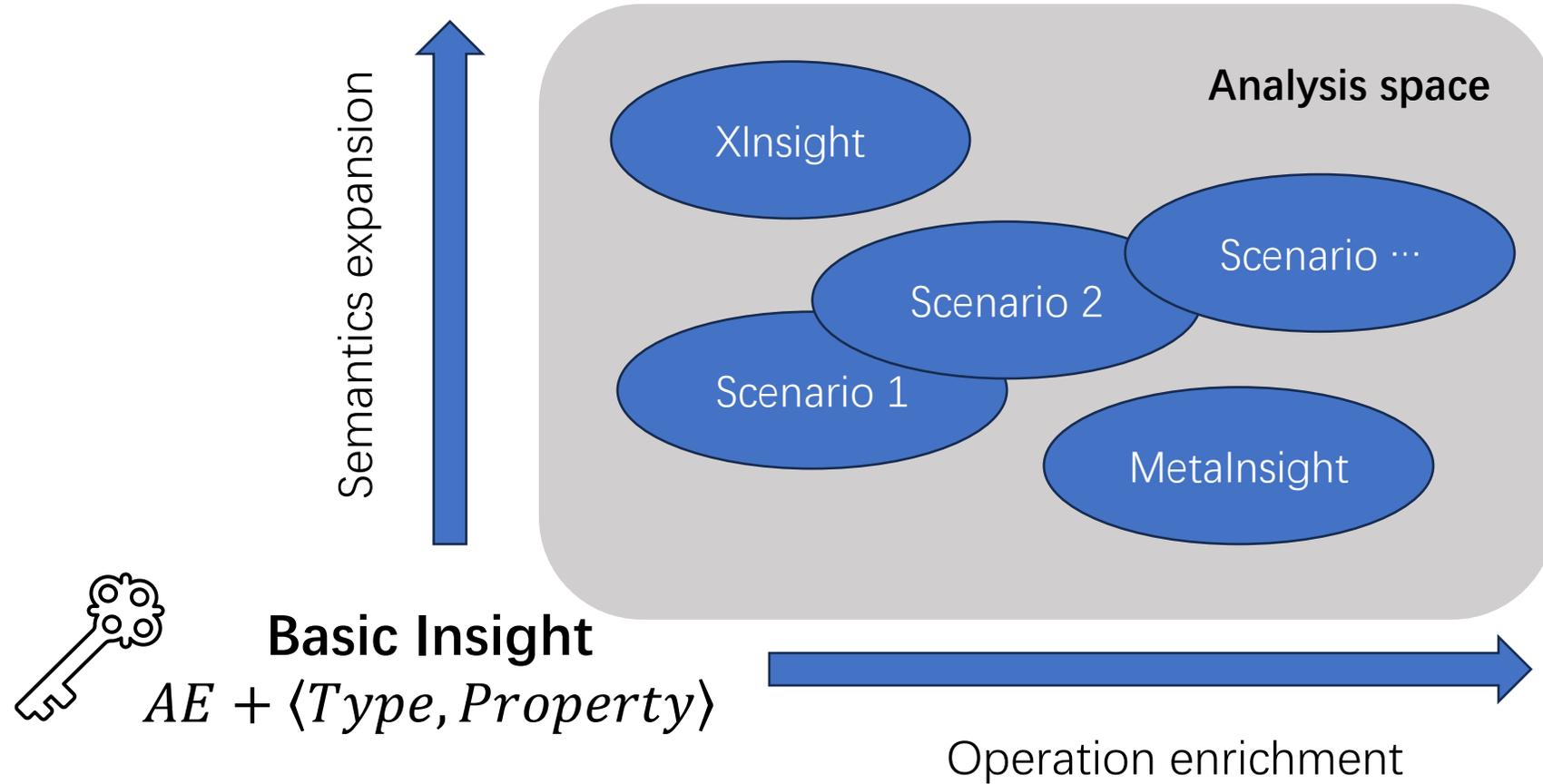


# ▶ System: QuickInsight\*



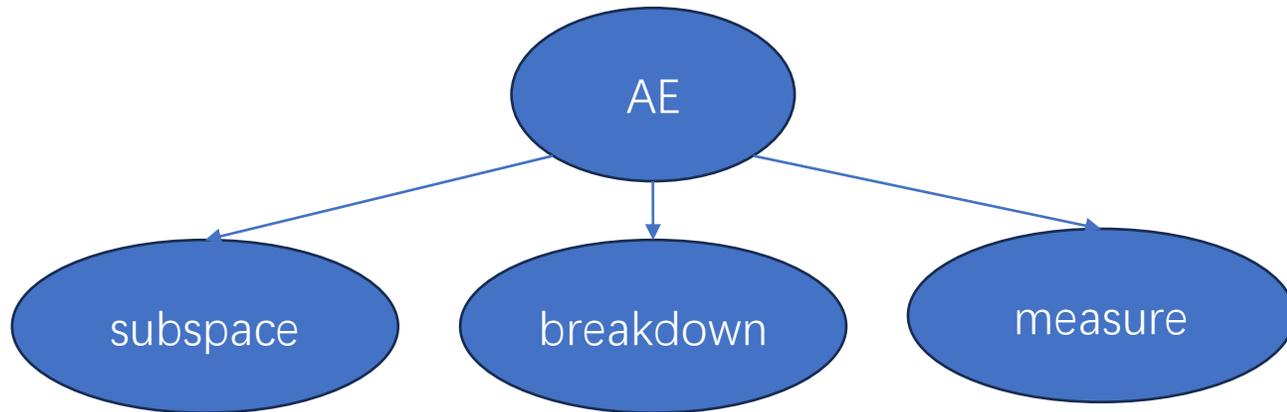
\* Released into Microsoft Power BI (2015) and Excel Analyze Data (2019)

# ▶ Establishing insight-based analysis space



# ▶ MetaInsight: composition from basic insights

- Observation: there exists semantically meaningful operations over AE:



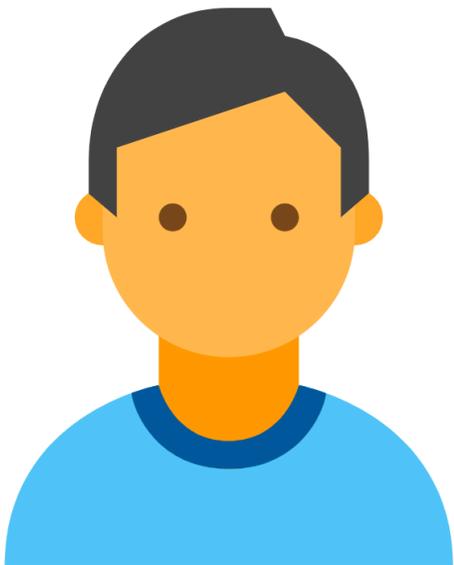
Operations →

- Siblings
- Children
- Parents

- Inclusion / exclusion
- Different granularity

- Inclusion / exclusion
- Homogenous
- ...

# ▶ A typical EDA iteration

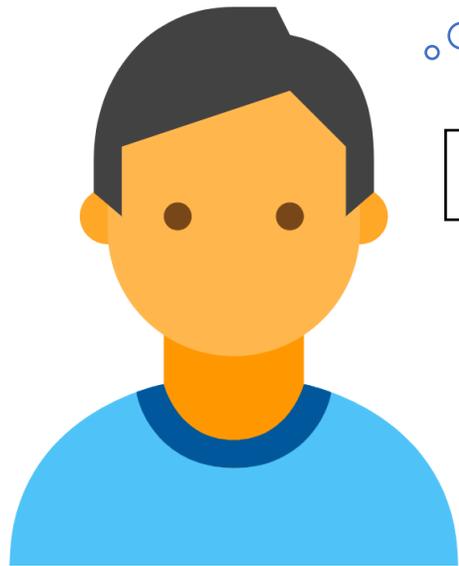


A real estate agent -- Bob

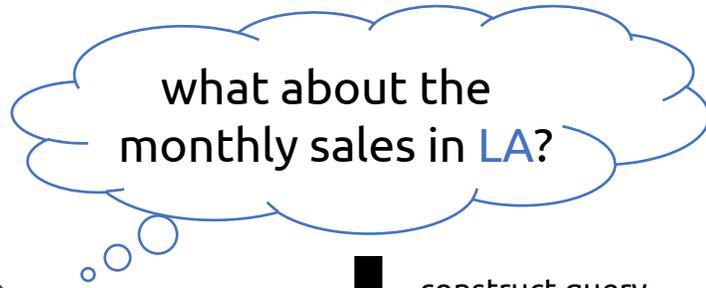
City	House Style	Month	Sales	...
Los Angeles	2Story	Jan	208,500	...
...	...	...	...	...
Los Angeles	1.5Fin	Dec	163,200	...
...	...	...	...	...
Yuba	1.5Unf	Dec	118,000	...

House Sales in California

# ▶ A typical EDA iteration



A real estate agent -- Bob



construct query

```
SELECT Month, SUM(Sales) FROM DATASET WHERE  
City=="Los Angeles" GROUP BY Month
```

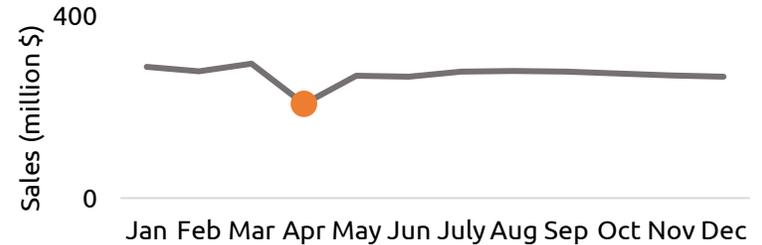
query result

City	Month	Sales
Los Angeles	Jan	288,231,200
	...	...
	Dec	269,562,200

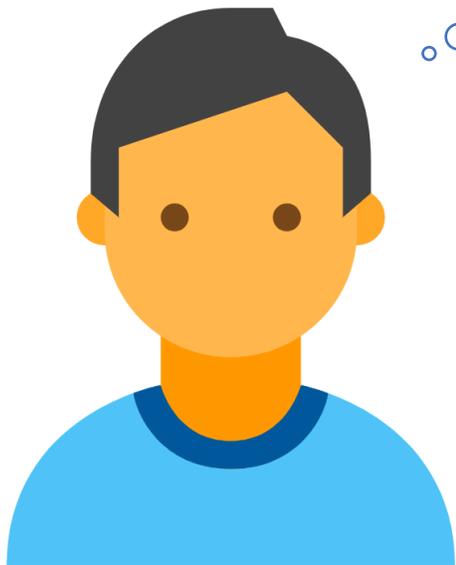
visualization



For Los Angeles, Apr has minimum Sales

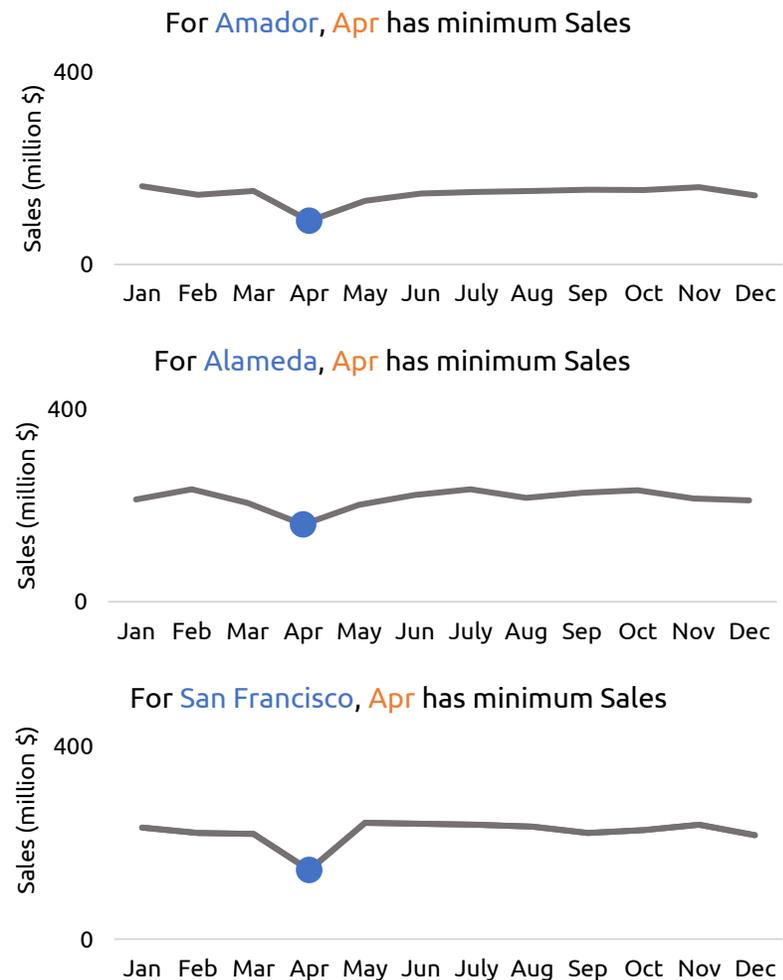


# ▶ A typical EDA iteration

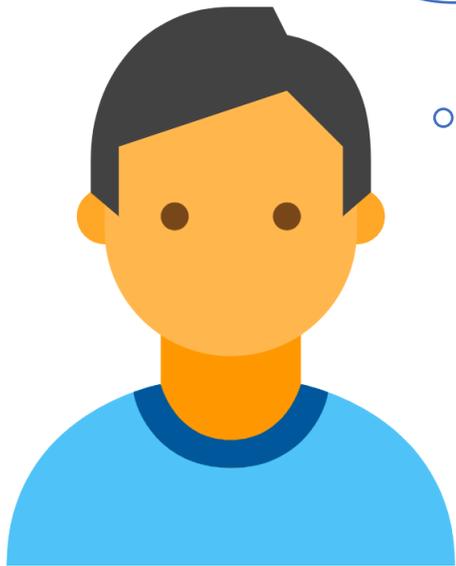


A real estate agent -- Bob

what about the monthly sales in other cities?



# ▶ A typical EDA iteration



A real estate agent -- Bob

did **ALL** cities have bad sales in **April** or are there any exceptions?

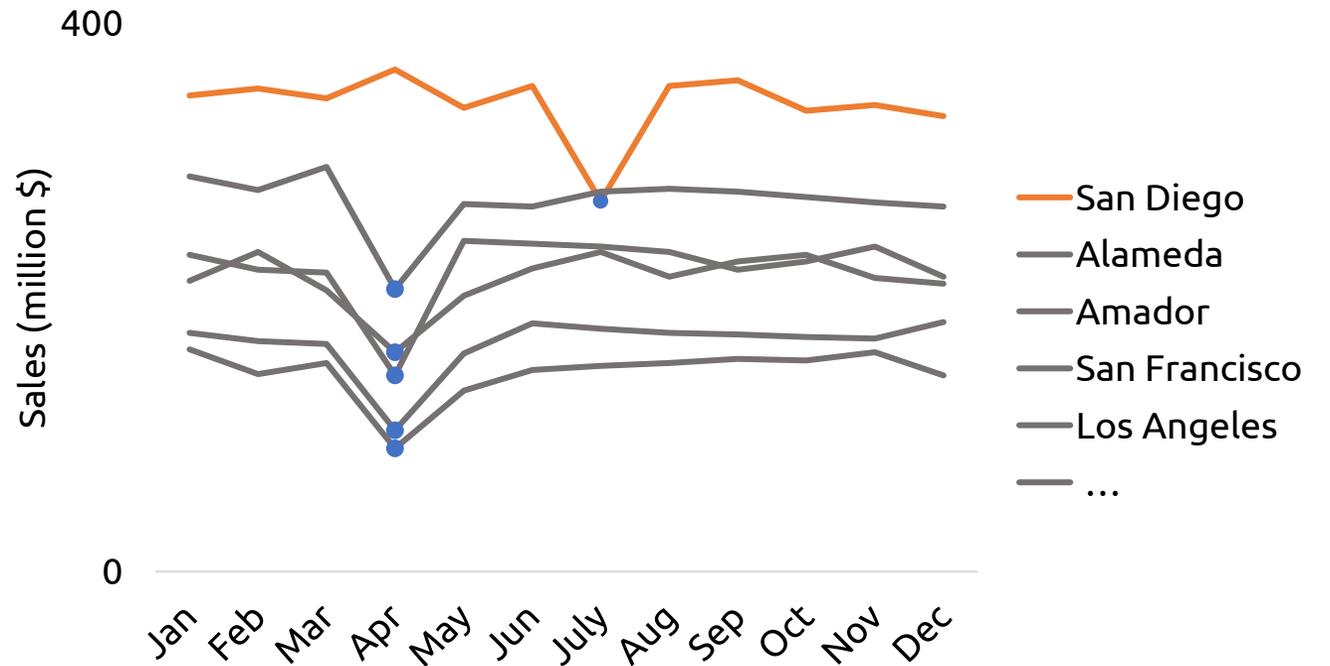


# ▶ A typical EDA iteration



A real estate agent -- Bob

*MetInsight:* For most Cities in California, Apr has minimum Sales; except San Diego



# ▶ Sensemaking mechanisms of human EDA iteration

- Mechanism #1 [1]

*knowledge extraction*: essential characteristics of raw data distribution

e.g., April has **BAD** Sales.

- Mechanisms #2 [2]

*inductive hypothesis*: the generality of characteristics of a basic data pattern

e.g., did **ALL** cities have bad sales in April?

- Mechanisms #3 [2]

*validity inquiry*: existence of unusual cases and how they differ from the general knowledge

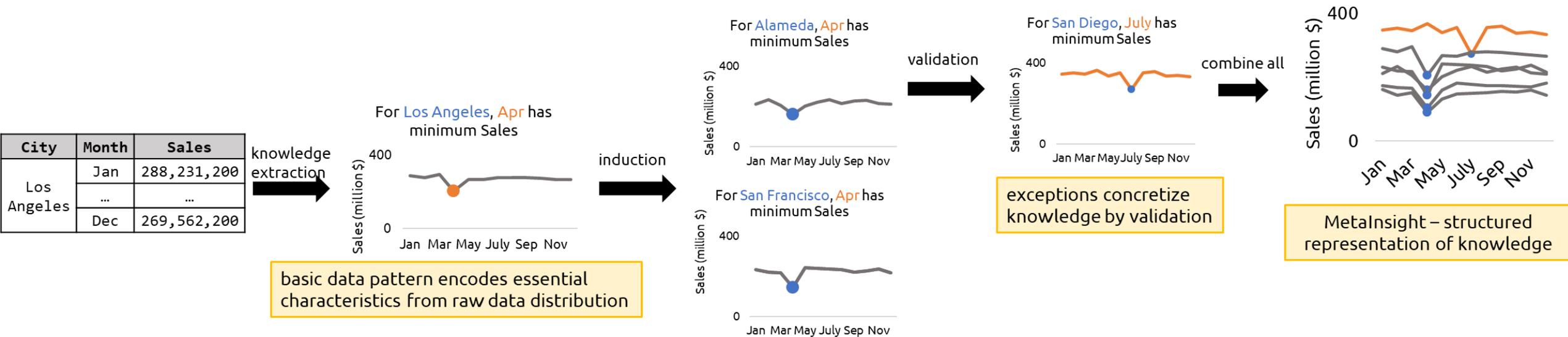
e.g., is there any city that does **NOT** have bad sales in April?

[1] Ding, Rui, et al. "QuickInsights: Quick and automatic discovery of insights from multi-dimensional data." *SIGMOD*. 2019.

[2] Zhang, Pengyi, and Dagobert Soergel. "Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking." *Journal of the Association for Information Science and Technology* 65.9 (2014): 1733-1756.

# ▶ MetaInsight Overview

- *MetaInsight* is a **structured representation of knowledge** extracted from multi-dimensional data.
- How to constitute a MetaInsight?



# ▶ How to generate a data pattern?

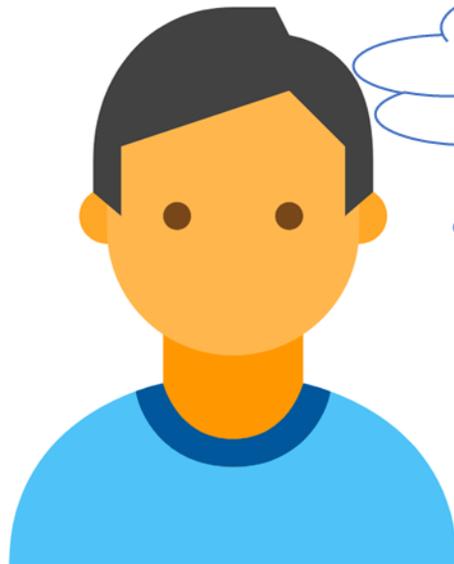
$ae =$

City	Month	Sales
Los Angeles	Jan	288,231,200
	...	...
	Dec	269,562,200

evaluate patterns

Insight Type	Evaluation Result
Trend	False
...	...
Unimodality	True

generate valid pattern



is there an increasing trend?  
is there a valley?  
...

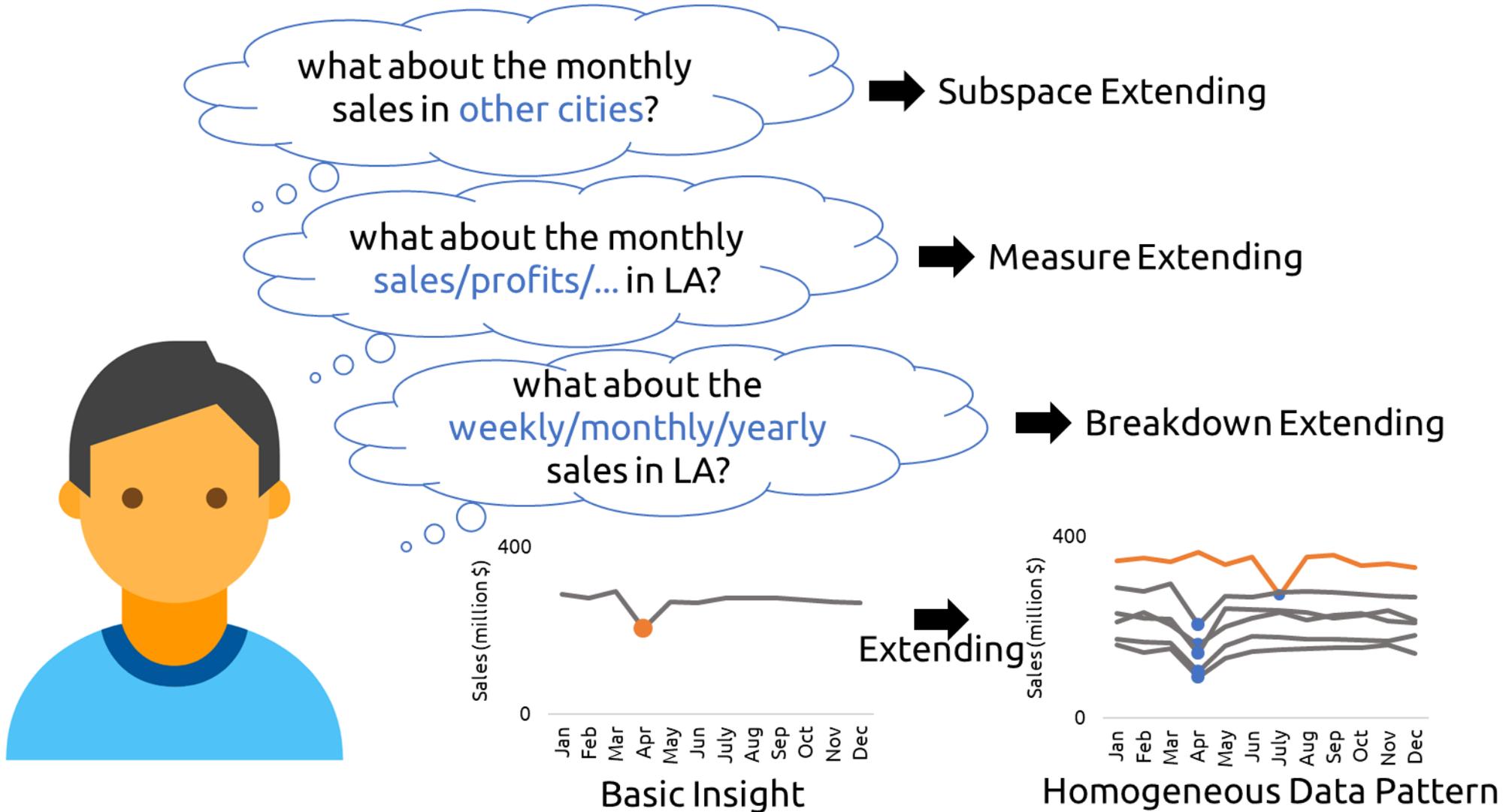
type-induced data pattern  
(a.k.a. insight)

For Los Angeles, Apr has minimum Sales



$insight = \langle ae, \underset{\text{type}}{\text{Unimodality}}, \langle \text{April}, \underset{\text{Property}}{\text{Valley}} \rangle \rangle$

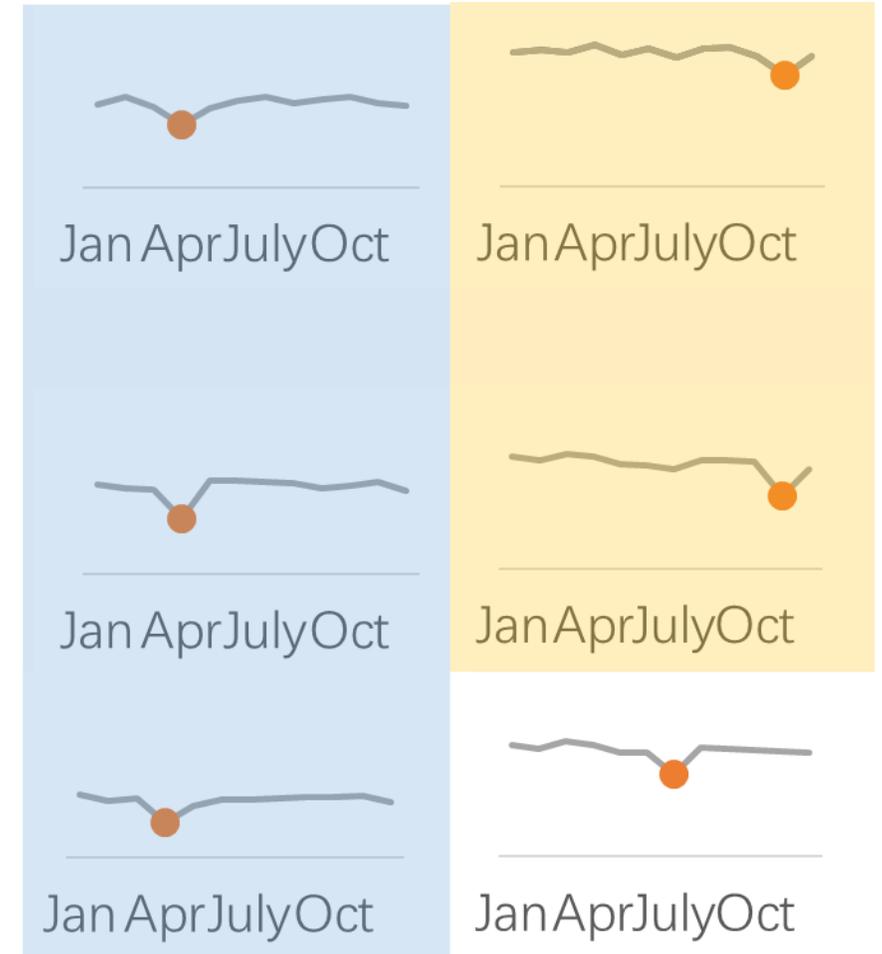
# ▶ How to extend an (basic) insight?



# ▶ How to organize basic insights?

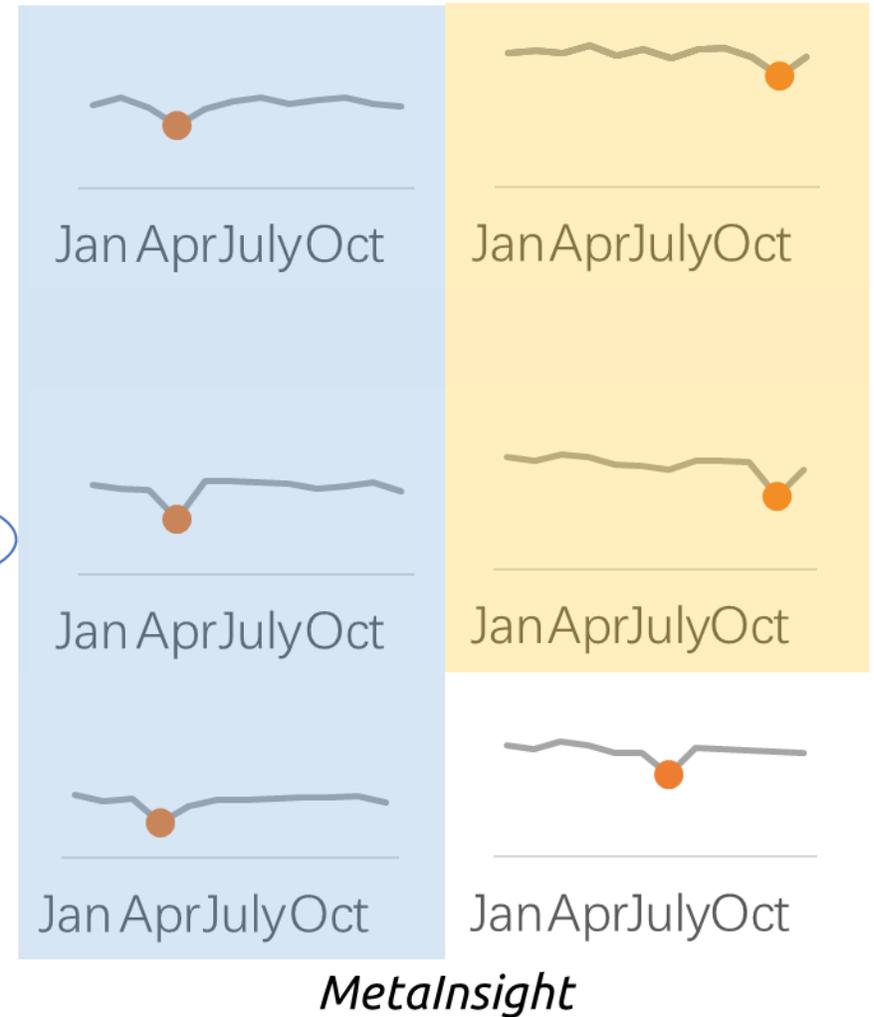
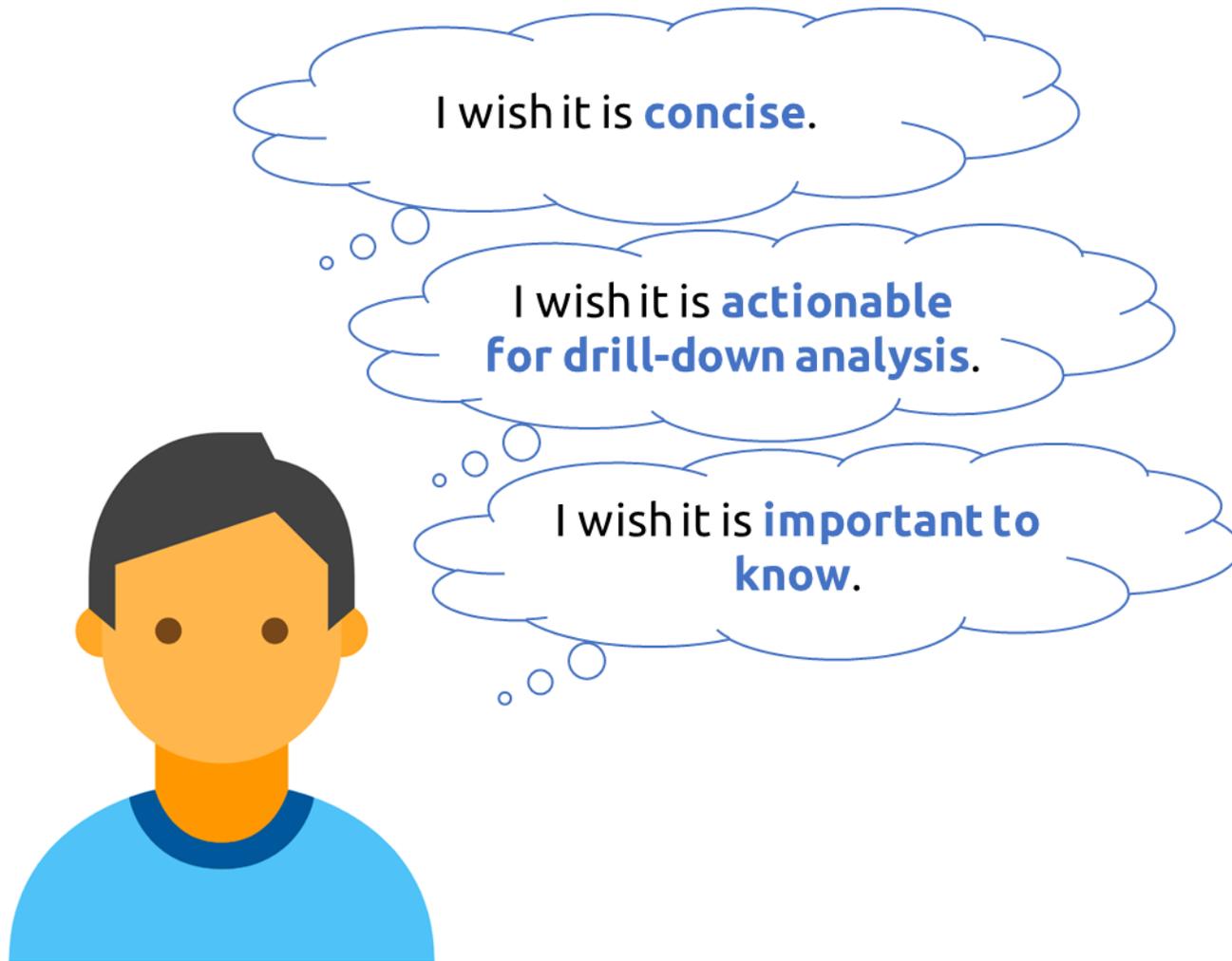
## Organize Rules

- Rule #1
  - Insights with **same type and highlight** are grouped to form **commonness** if their ratio exceeds  $\tau$
- Rule #2
  - Remaining insights are marked as **exceptions**



Homogeneous Data Pattern → *MetaInsight*

# ▶ How to score MetaInsight?

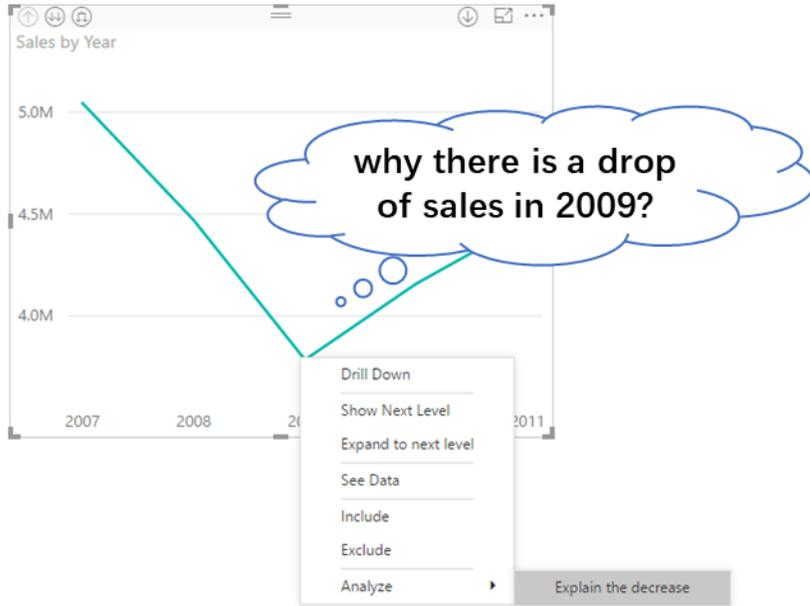


# ▶ XInsight: eXplainable Data Analysis

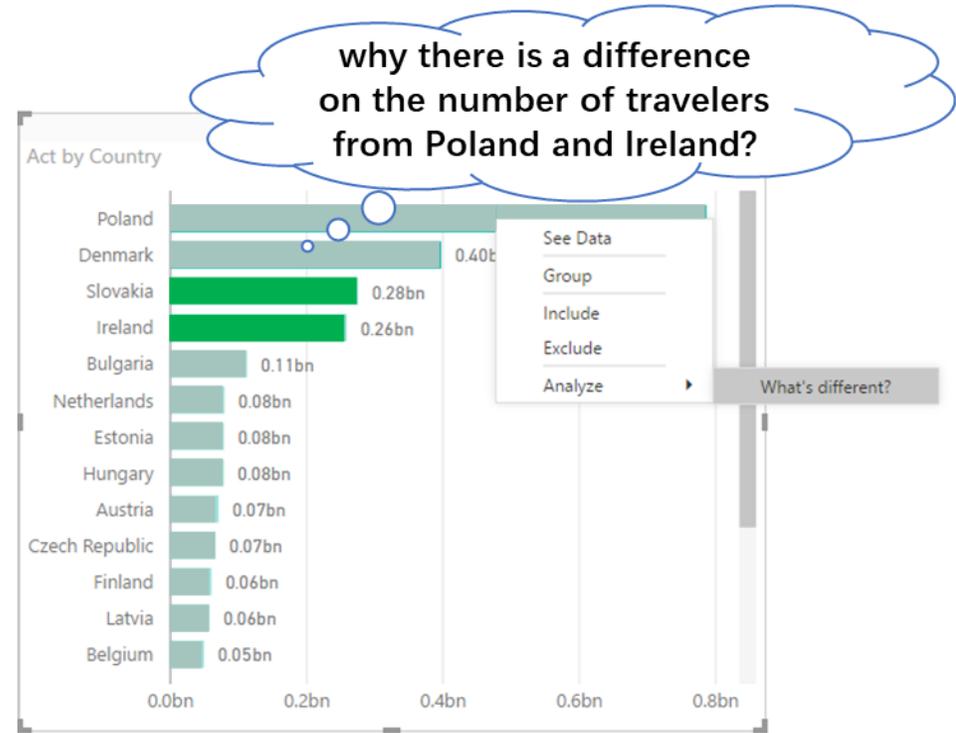
The practical needs of eXplainable Data Analysis (XDA)

- What do users want from EDA: justify and rely on knowledge and conclusions →
- XDA: is proposed to deliberate data facts and enhance user comprehension. XDA advances data analysis by providing users with effective explanations. By suggesting and justifying choices to alter outcomes, XDA helps users comprehend and trust phenomena emerging from data; as a result, it facilitates real-world decision making

# ▶ What Does User Ask in Daily Data Analysis Tasks?

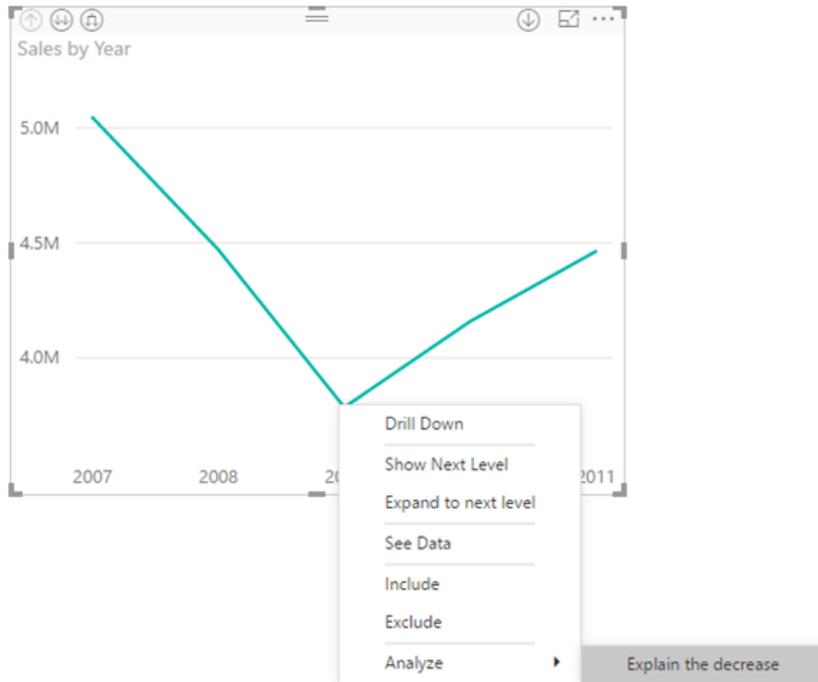


Car Sales Data



Hawaii Traveler Data

# ▶ What Does User Expect For An Answer?

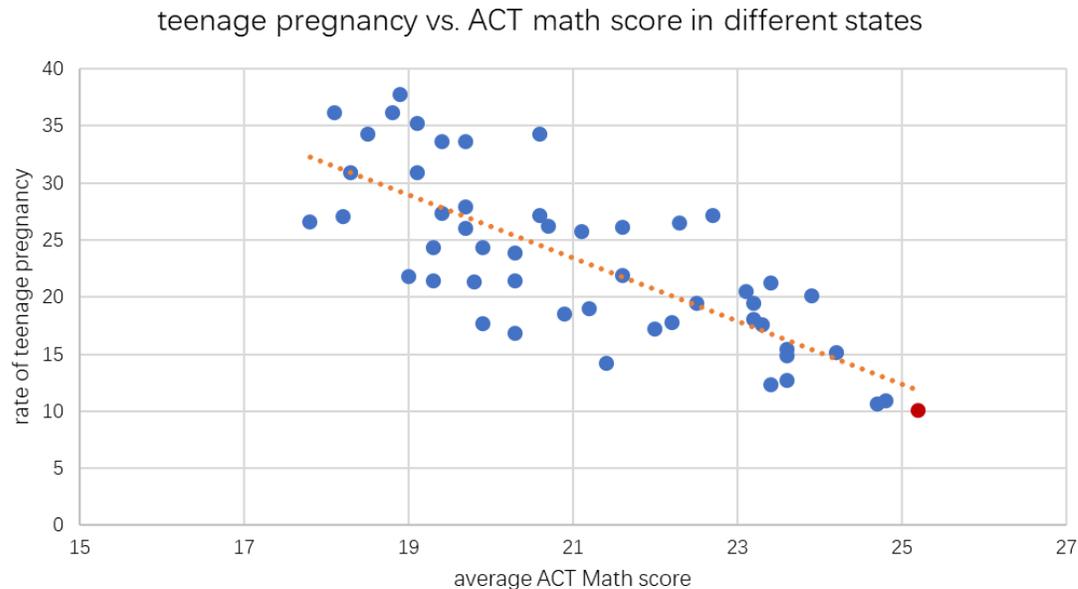


Sales  
BY YEAR AND BRAND

'Toyota' and 'Honda' accounted for the majority of the decrease among Brand. The relative contributions made by 'Hyundai', 'GMC', and 'Ford' changed the most.

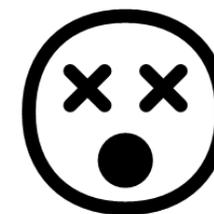


# ▶ Problematic Explanation Can Backfire Human Cognition



Why teenage pregnancy rate is low in Massachusetts?

Because it has high average ACT math score.



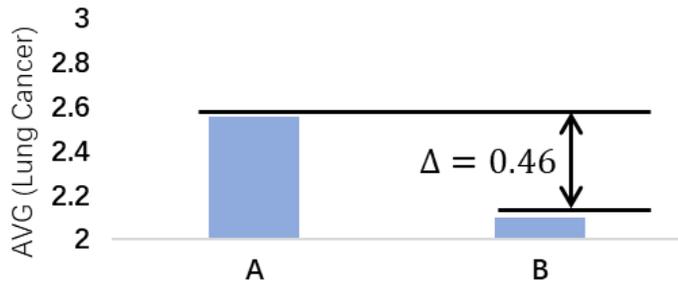
# ▶ Formulation of XInsight

- Why-Query: composition of two AEs
  - $ae_1 = \langle subspace_1, null, measure \rangle$
  - $ae_2 = \langle subspace_2, null, measure \rangle$
  - Where  $subspace_1$  and  $subspace_2$  are within a sibling group
- $XInsight := \langle ae_1, ae_2, Type, Property_X \rangle$  where
  - $Property_X := \langle Predicate, Responsibility \rangle$

# ▶ An example of XInsight for XDA

Location	Stress	Smoking	Lung Cancer	Surgery	5Y Survival
A	High (3)	Yes (1)	Severe (3)	Yes (1)	No (0)
...	...	...	...	...	...
B	Low (1)	No (0)	Mild (1)	No (0)	Yes (1)

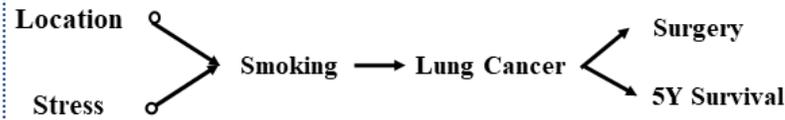
(a) raw data



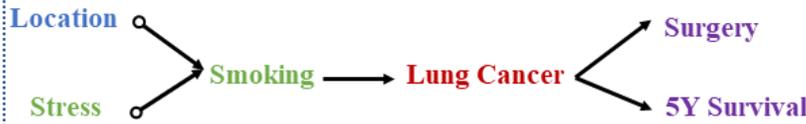
Why Lung Cancer (Severity) in Location=A is notably higher than Location=B?

WHY QUERY

(b) typical EDA output and the derived WHYQUERY



(c) qualitative causal relations learned by XLEARNER



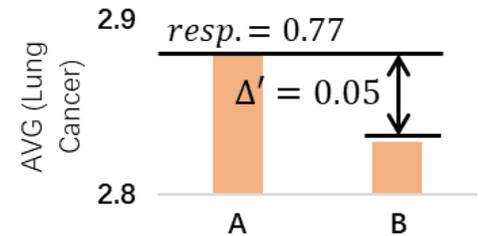
(d) causal and non-causal semantics judged by XTRANSLATOR

Type	Predicate	Responsibility
Causal	Smoking = Yes	0.77
Causal	Mid (2) ≤ Stress ≤ High (3)	0.61
...	...	...
Non-causal	Surgery = Yes	0.73

(e) explanations identified by XPLAINER

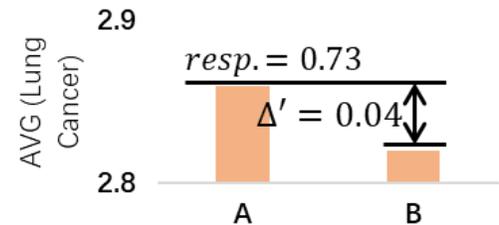
Three modules of XINSIGHT

**Causal Explanation:** “Factor=Smoking. Smoking=Yes” explains the difference on Lung Cancer between Location=A and Location=B.



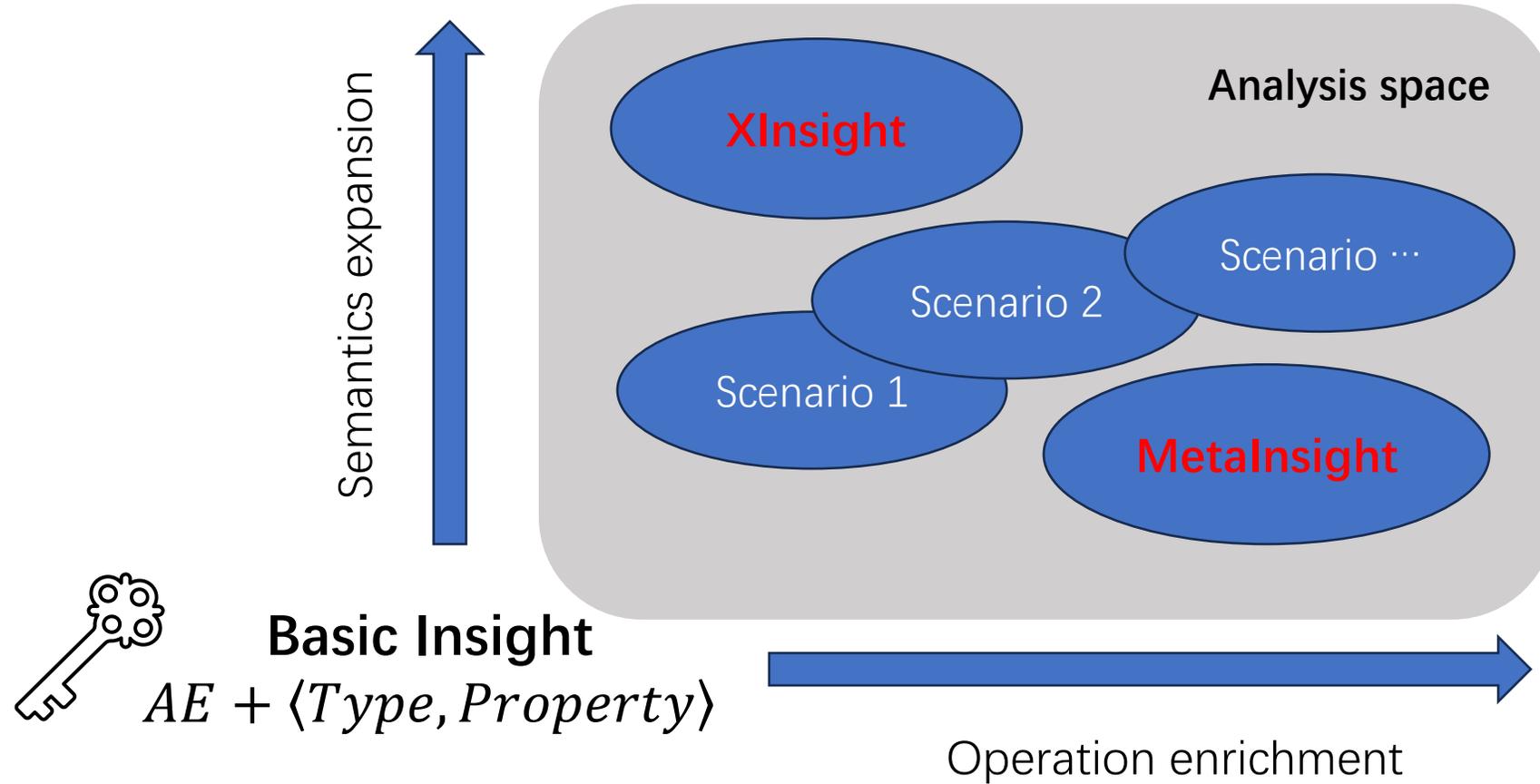
(f) example of causal explanation

**Non-causal Explanation:** “Factor=Surgery. Surgery=Yes” is relevant to the difference on Lung Cancer between Location=A and Location=B.



(g) example of non-causal explanation

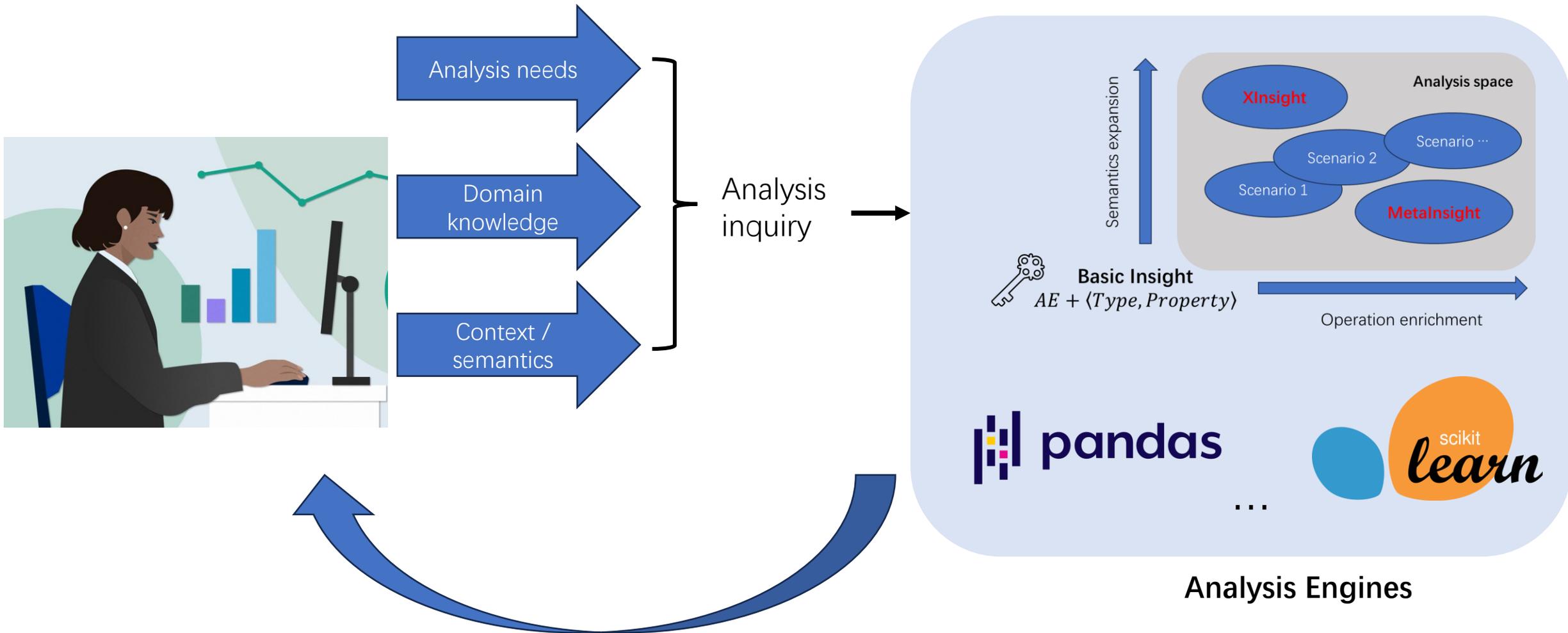
# ▶ Recap: Establishing insight-based analysis space



## **PART 02**

# **InsightPilot: LLM-Empowered Automated Data Analysis Paradigm**

# ▶ Today's data analysis is human centric



# ▶ Automated data analysis: the holy grail

- Human centric data analysis → automated data analysis can hardly happen
- LLM has potential to play the data analyst' s role → automated data analysis could happen!

## ▶ The potential of LLM

LLM can play at least as a modest data analyst to “drive” the analysis, because of

- Broad domain knowledge
- Can understand data context / semantics
- Fast response

A case study: ask LLM to conduct data analysis on two tasks

- House price data, and weasel vs. stoat data.

# ▶ Our goal: unleash the power of LLM but also harness it

- LLM' s risks: Unsoundness / Explainability / Hallucination

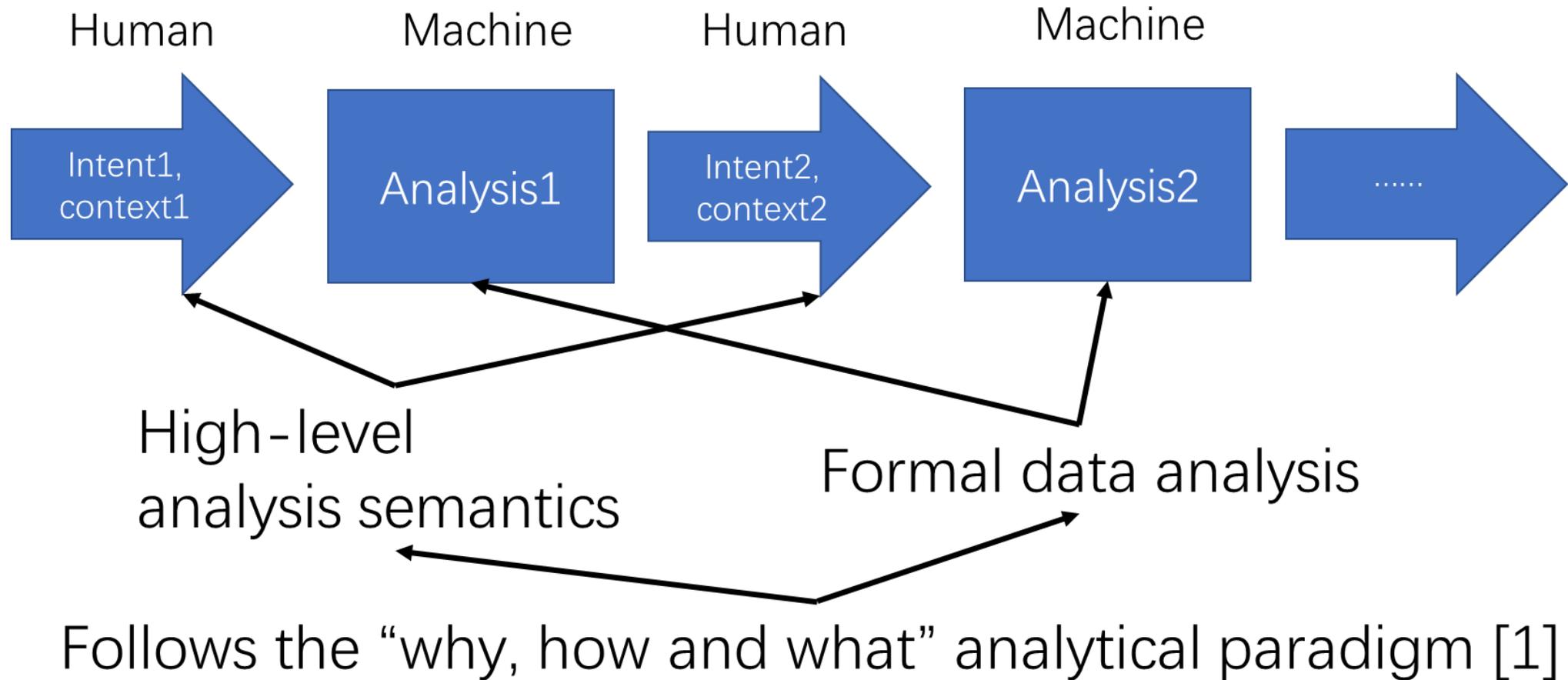


- LLM could play the role of a data analyst, but cannot replace the role of analysis engines



- InsightPilot: let LLM and insight engine works synergically!

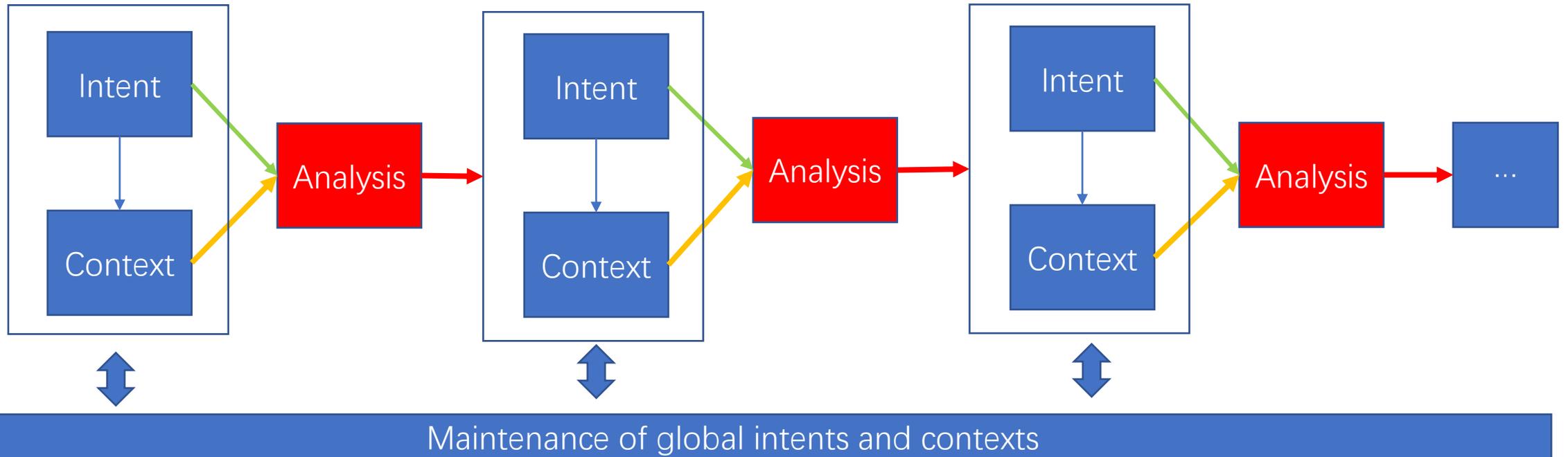
# ▶ Revisit the cognitive process of data analysis flow



[1] Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12), 2376-2385.

# ▶ Framework of InsightPilot

 :Provide by Insight  
 :Provide by LLM/Human



-  Analysis to intent: use current analysis result to detect the next analysis intent
-  Intent to analysis: use current intent to pick the suitable analysis engine to conduct next-step analysis
-  Context to analysis: use current context to feed suitable parameters to the analysis engine

# ▶ Analysis intent selection

- Problem: given current analysis result (insight) and analysis history (context), what is the next analysis intent?
- Domain of analysis intent
  - {Understand, Summarize, Explain, Compare, ...}
  - Extensible to support broader scenarios
- Solution:  $Intent = LLM(insight, context, metadata)$ 
  - E.g., given a dataset about the math' s scores of each student in a school (metadata), now an insight shows that ClassA has highest average math' s score (insight), given that the math' s scores have been explored by different classes (context). Now LLM suggests that the next step is Intent= "Explain" , corresponding to the intent that to explain why ClassA has such highest score.

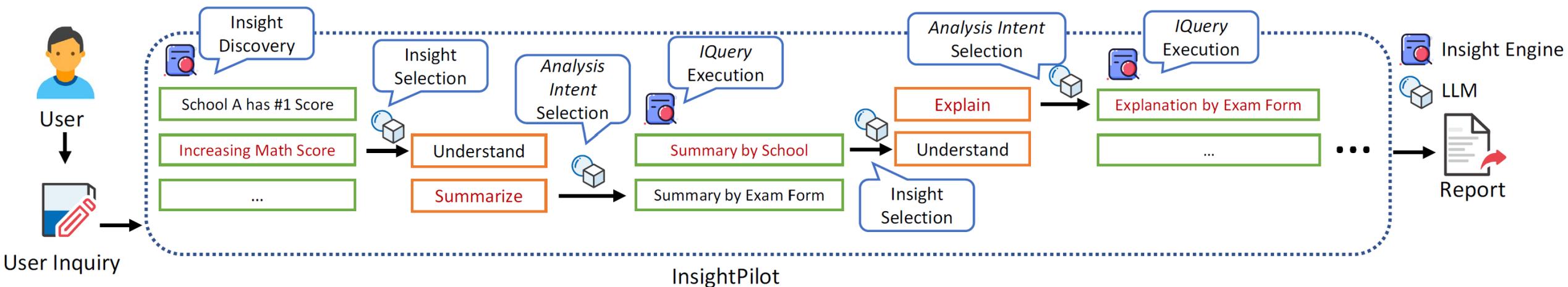
# ▶ Extracting parameters to trigger insight engine

- Problem: given an analysis intent and previous insight, how to parse suitable parameters to trigger the insight engine?
- Key observations
  - The parameters are structured due to the formulation of insight
  - The parameters are highly associated with the “property” tuple of the previous insight
  - Easy to provide few shot examples to LLM
- Solution:  $params = LLM(intent, insight, metadata)$ 
  - E.g., the intent is “Explain” , and the  $insight = \{*, Class, AVG(Score), Rank, "Top1 = ClassA"\}$  (ClassA has highest average math’ s score among all classes), the suitable parameters can be
    - Module = XInsight
    - $AE_1 = \{ClassA, null, AVG(Score)\}$
    - $AE_2 = \{Other, null, AVG(Score)\}$  or  $AE_2 = \{ClassB, null, AVG(Score)\}$
    - Other: the other classes, ClassB: the class with second highest math’ s score

# ▶ Summary of prompt engineering

- Initialization
  - Data, metadata, initial insights
- Analysis intent selection
  - Insight, context, metadata
- Parameter extraction
  - Insight, intent, metadata
- Insight selection
  - A set of insights, context, metadata
- Report generation
  - A sequence of explored insights, metadata

# ▶ Current implementation



## Typical tasks between LLM and Insight Engine

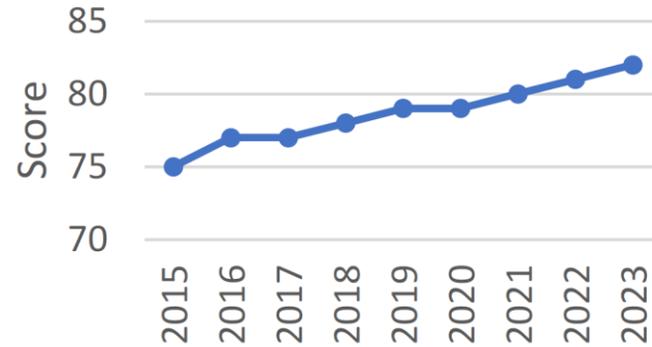
- Intent selection: select what is the high-level intent of next analysis step
  - Based on context and semantics of currently seen insights
- Parameters extraction: extract proper parameters which are feasible to feed into Insight Engine
- Insight selection: select a most suitable insight that the analysis proceed

# ▶ A case study

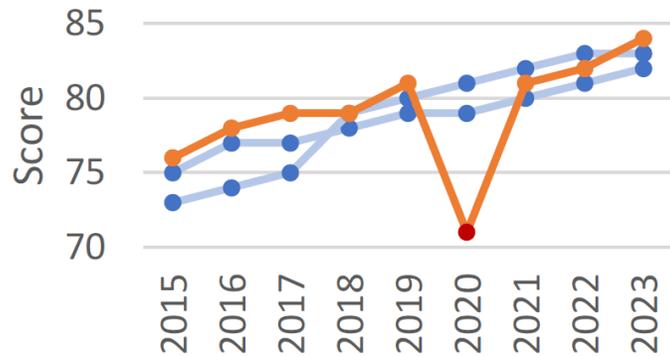
raw data

Student	School	Score	Year	Subject	Exam Form
Bob	A	87	2018	Math	Multi-Choice
Tom	B	67	2018	History	Open-book
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

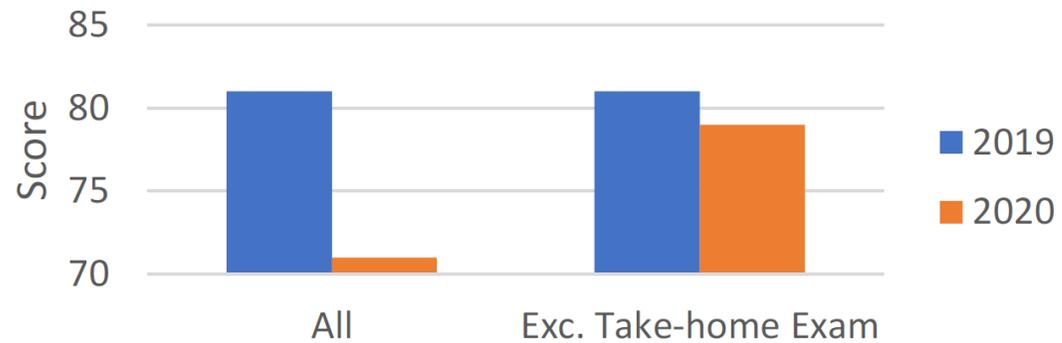
Subject=Math has an **increasing** trend over years.



When Subject=Math, most School have an **increasing** trend while School=C has an **outlier** in Year=2020.



Exam Form=Take-home explains the **outlier** in Year=2020.



# ▶ Summary

- Insight-Based Exploratory Data Analysis
  - The concept and formulation of insight
  - The analysis space established from insight
  - Case studies: MetaInsight, XInsight
- InsightPilot: LLM-Empowered Automated Data Analysis Paradigm
  - The opportunities of unleashing the power of LLM
  - Synergistic Integration of LLM with InsightEngine
  - Towards automated data analysis

# 感谢聆听

