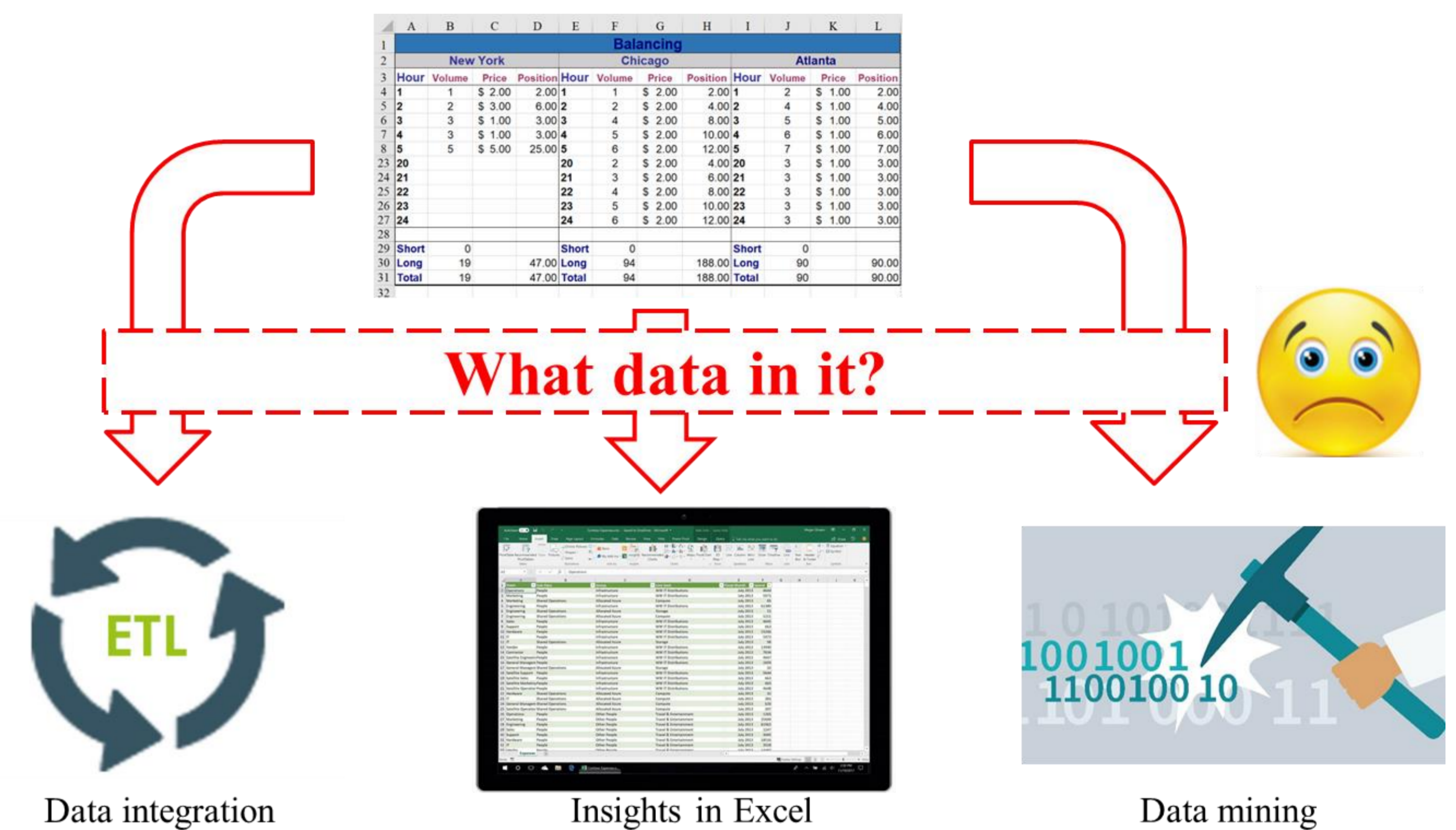# 电子表格中可扩展单元组识别

窦文生, 韩石, 徐良, 张冬梅, 魏峻

Spreadsheet structure information is missing

Spreadsheets are designed to be **interpreted by human**, rather than data analysis applications.

	Relational data	Spreadsheet
Data schema	Explicit	Implicit
Structure	Two-dimensional	Multi-dimensional
Data relationship	Foreign key	Formula / No relationship
Table definition	Explicit	Implicit

Analysis-friendly **Human-friendly**

What data in it?

Data integration   Insights in Excel   Data mining

Expandable group

Expandable group: A basic unit and all its repeating units form an expandable group.

➤ The corresponding cells in an expandable group have the similar formats and semantics, e.g., cells [A2:D31] and [E2:H31].

➤ Based on identified expandable groups, we can transform spreadsheet data into relational data, which can be used by other data analysis tools.
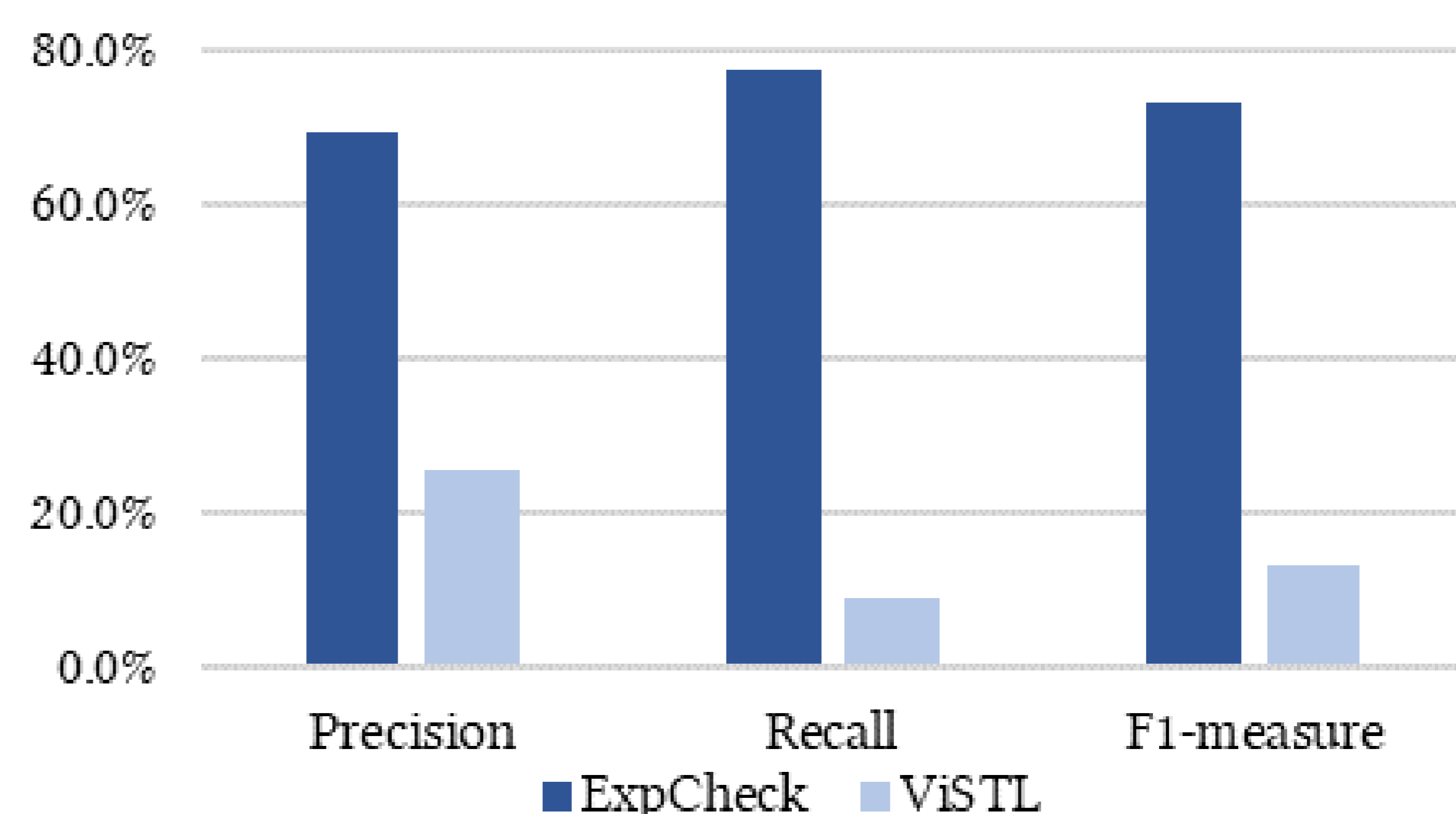

City	Hour	Volume	Price	Position
New York	1	1	2.00	2.00
New York	2	2	3.00	6.00
New York	3	3	1.00	3.00
Chicago	1	1	2.00	2.00
Chicago	2	2	2.00	4.00
Chicago	3	4	2.00	8.00
Atlanta	1			2.00
Atlanta	2	4	1.00	4.00
Atlanta	3	5	1.00	5.00

Analysis-friendly

ExpCheck

Key observation: Units in an expandable group have the similar formats and semantics.

Similar semantics

Same formats

Evaluation

ExpCheck performs much better than existing techniques, e.g., ViSTL[1].

[1] R. Abraham et al., Inferring Templates from Spreadsheets, ICSE 2006.