Percent百分点

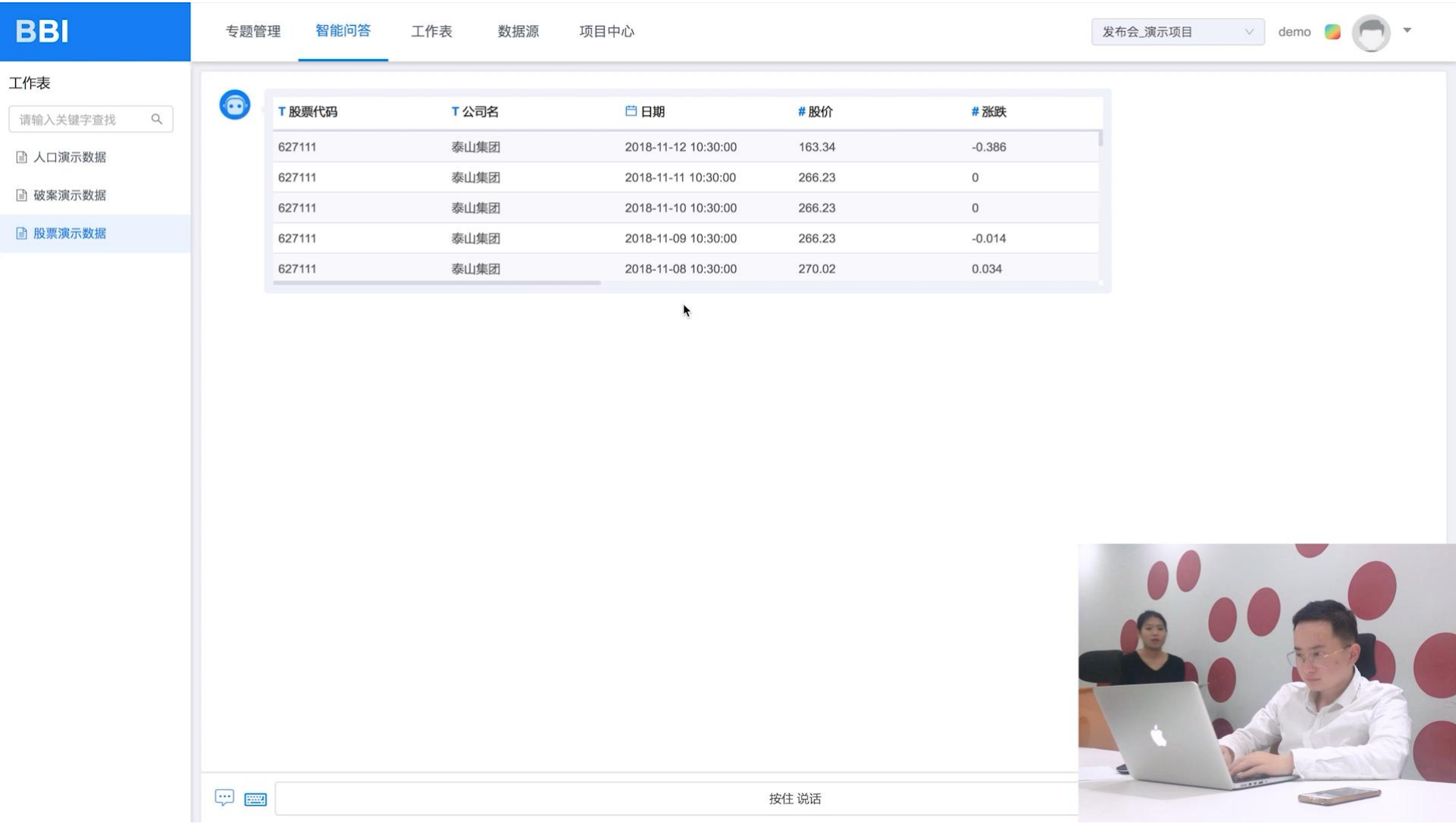
基于自然语言处理和深度学习的NL2SQL技术及其在BI增强分析中的应用

百分点　刘译璟

2019.12.01

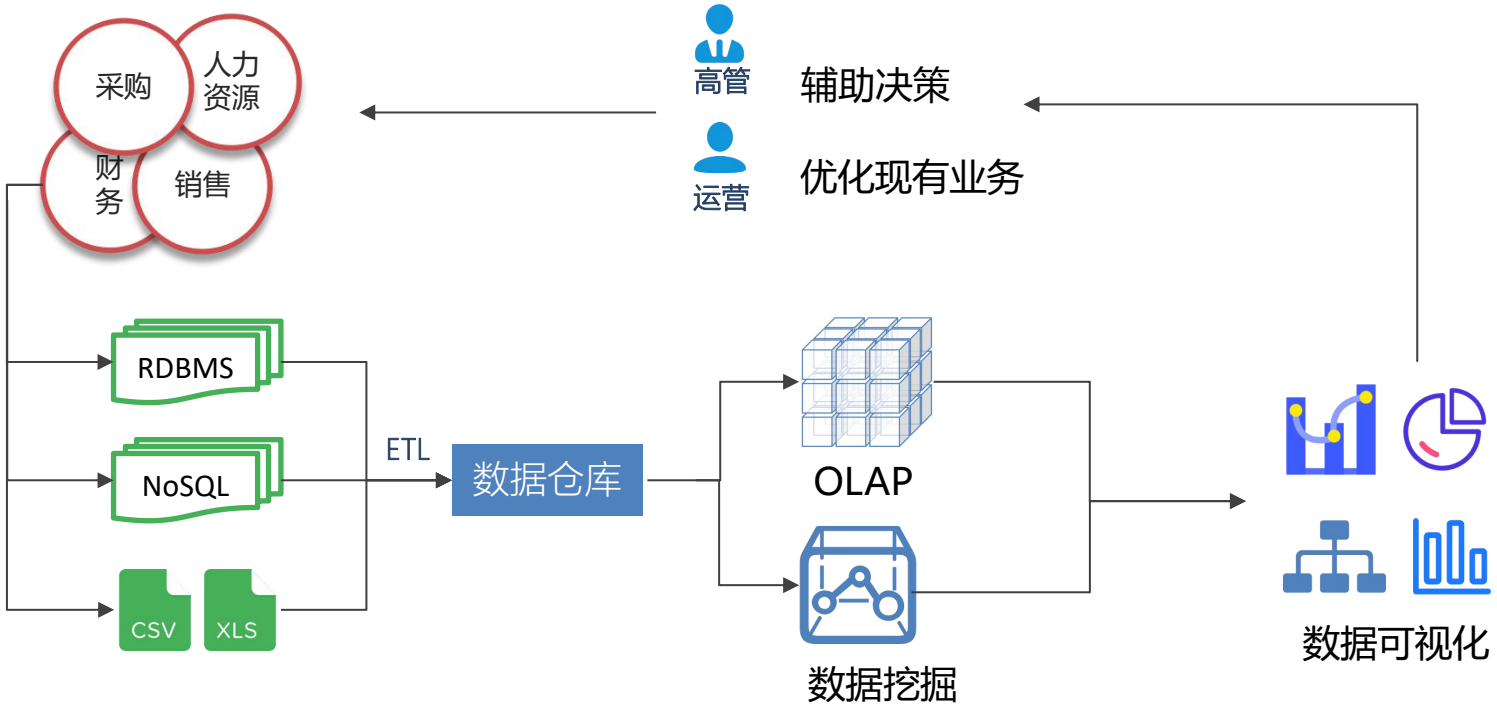
Percent百分点

CleverBI智能问答演示



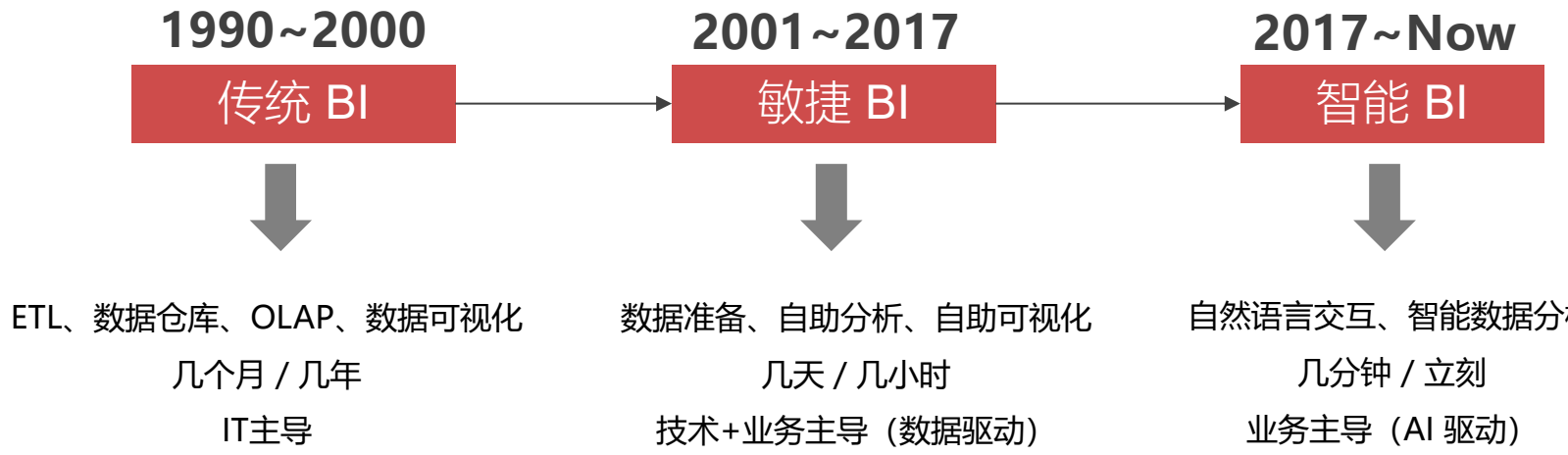
什么是BI

BI（Business Intelligence）即商业智能，用现代**数据仓库**技术、**联机分析处理**技术、**数据挖掘**和**数据展现**技术进行**数据分析**以实现**商业价值。**



Percent百分点

BI的发展阶段



Percent百分点

BI和增强分析

Gartner 将**增强分析**定义为使用机器学习和 AI 来进行智能的数据分析，以及利用自然语言更自然的与系统进行交互等，即在这些常见**分析**过程中使用自动化来**增强**它们。

主要驱动力

到 2020 年，增强分析将成为新用户购买 BI 产品、数据科学和机器学习平台、以及嵌入式分析的主要驱动力。

50%

到 2020 年，有 50% 的分析查询会通过搜索、自然语言处理或语音生成，或者自动生成。

3倍

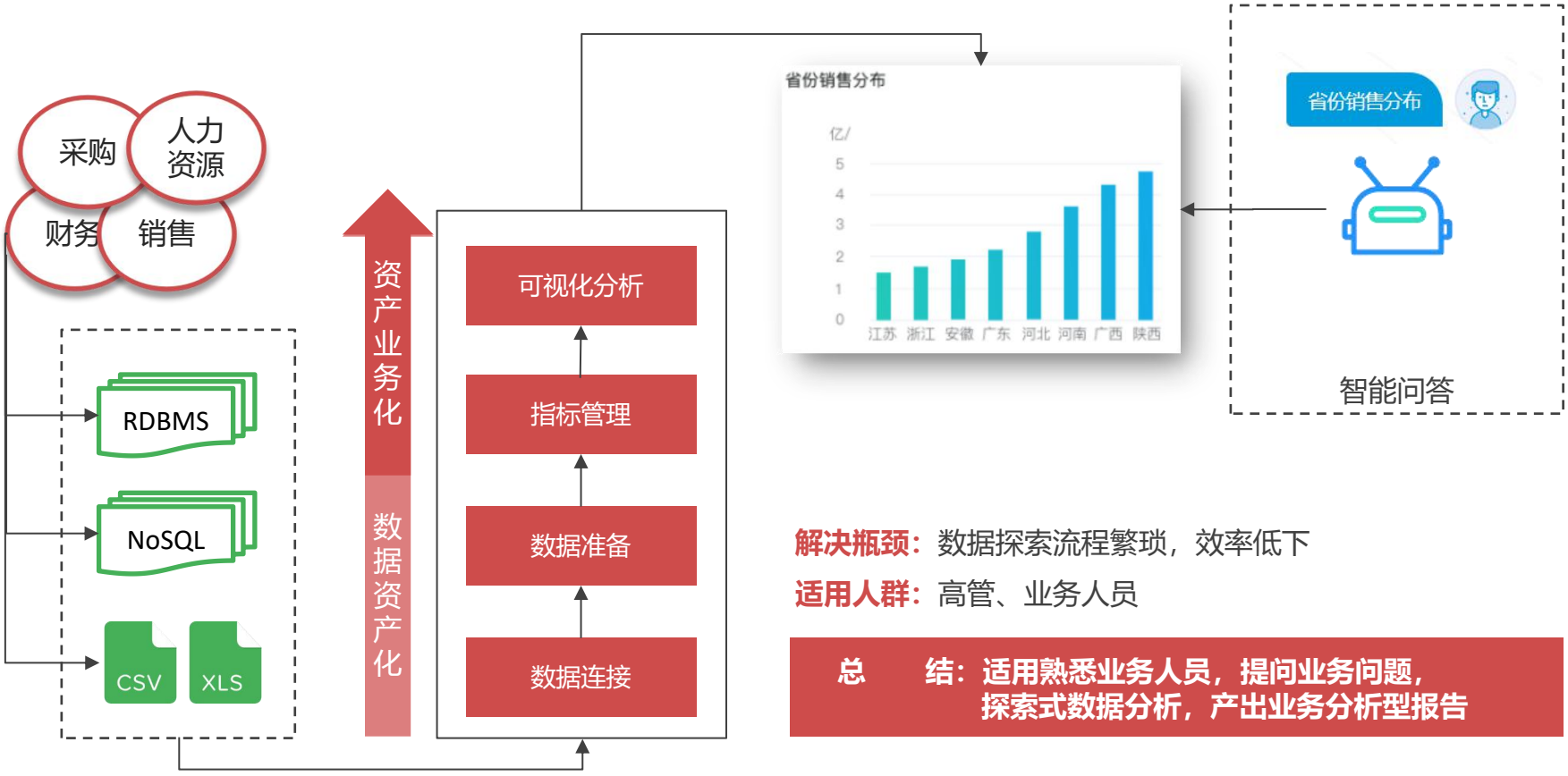
到 2020 年，业务部门的数据和分析专家数量的增速将是 IT 部门专家的 3 倍，这会迫使企业重新考虑其组织模式和技能。

50%↑

到 2021 年，自然语言处理和会话分析这两个功能，会在新用户、特别是一线工作人员中，将分析和商业智能产品的使用率从 35% 提升到 50% 以上。

Percent百分点

CleverBI中的智能问答



解决瓶颈：数据探索流程繁琐，效率低下

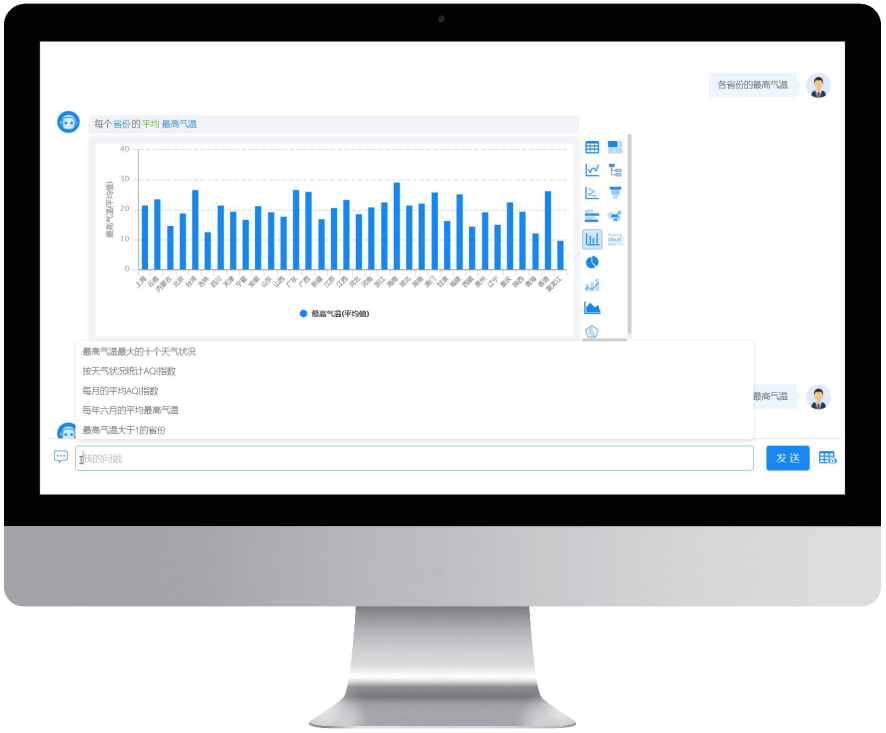
适用人群：高管、业务人员

总　　结：适用熟悉业务人员，提问业务问题，探索式数据分析，产出业务分析型报告

Percent百分点

CleverBI中的智能问答

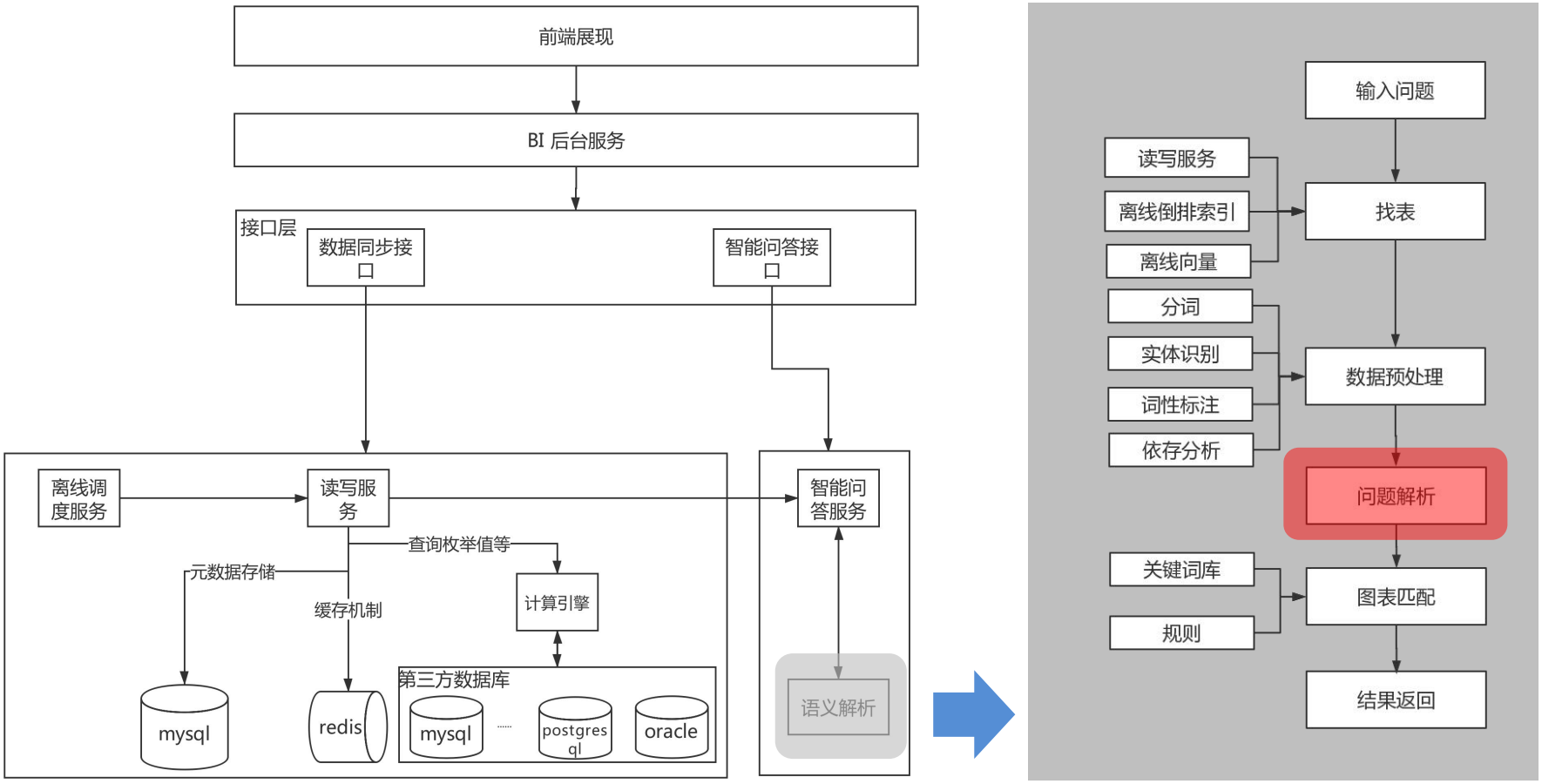
用户在分析问题时，可以使用智能问答更高效的探索数据。



- 自然语言转SQL
- 推荐聚合函数
- 推荐图表类别
- 交互式复述
- 多轮对话
- 问题生成

Percent百分点

智能问答关键架构



问题解析原理：NL2SQL

WikiSQL数据集

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...	...	...	...

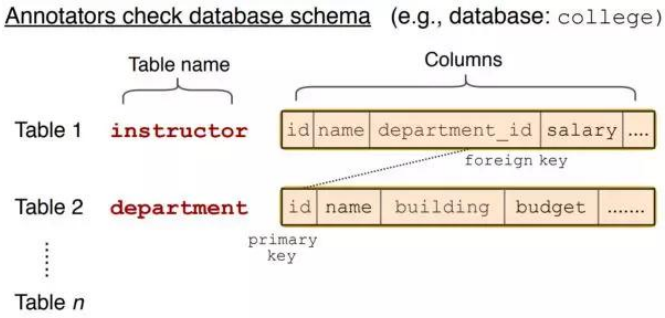
Question:
How many CFL teams are from York College?

SQL:
```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = "York"
```

Result:
2

Figure 2: An example in WikiSQL. The inputs consist of a table and a question. The outputs consist of a ground truth SQL query and the corresponding result from execution.

Spider数据集



Annotators create:

Complex question: What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex SQL:
```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

Coarse-to-Fine

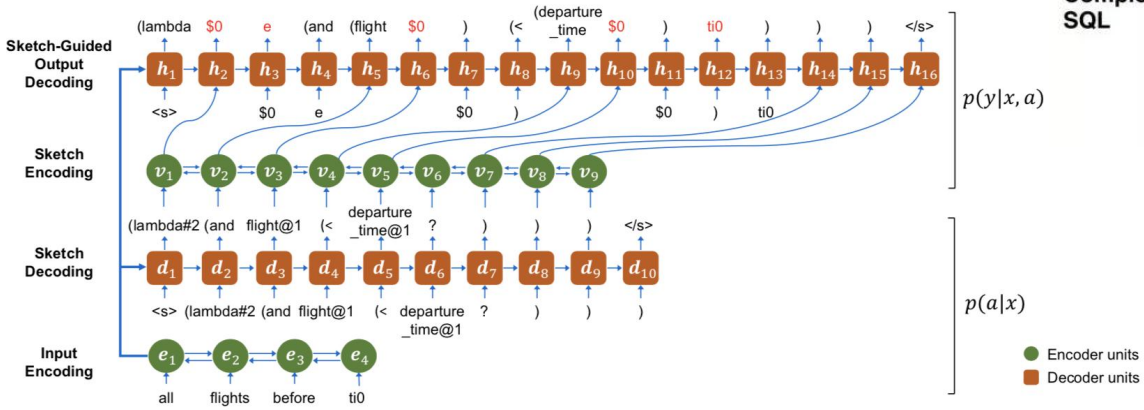


Figure 1: We first generate the meaning sketch a for natural language input x. Then, a fine meaning decoder fills in the missing details (shown in red) of meaning representation y. The coarse structure a is used to guide and constrain the output decoding.

Percent百分点

问题解析原理：NL2SQL

WikiSQL数据集

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...			

Question:
How many CFL teams are from York College?

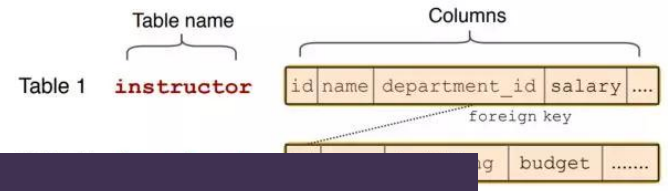
SQL:
```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = "York"
```

Result:

Figure 2: An exar
of a ground truth S

Spider数据集

Annotators check database schema (e.g., database: college)



of the departments
eater than the

department as
2.id

(SELECT avg(salary) FROM instructor)

现有NL2SQL的局限

缺乏大规模的中文标注集
复杂问题的解析准确度很低

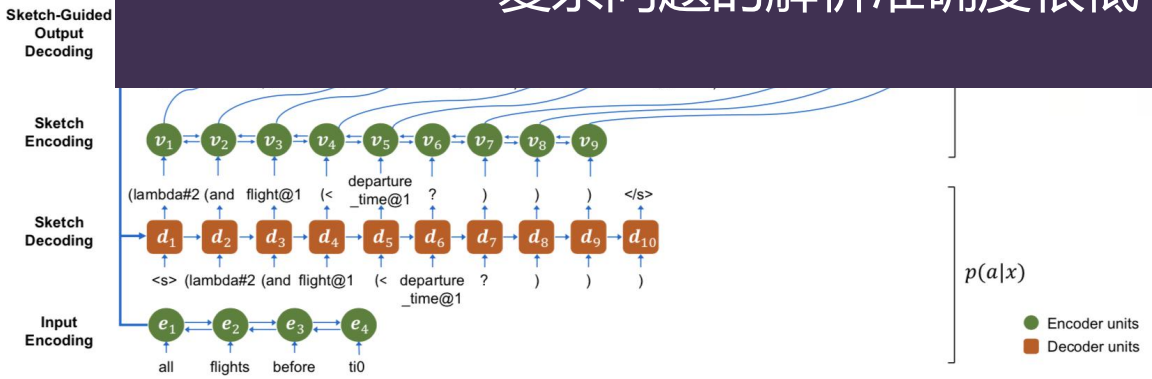
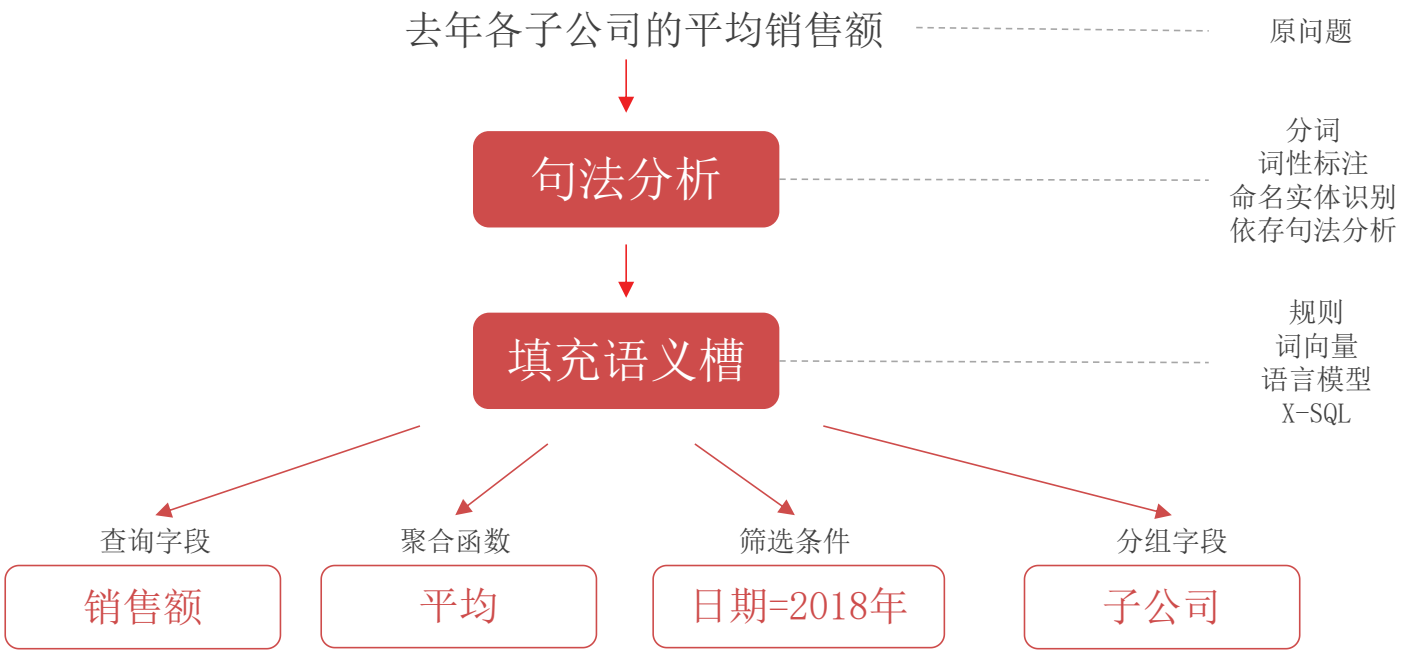


Figure 1: We first generate the meaning sketch a for natural language input x. Then, a fine meaning decoder fills in the missing details (shown in red) of meaning representation y. The coarse structure a is used to guide and constrain the output decoding.

Percent百分点

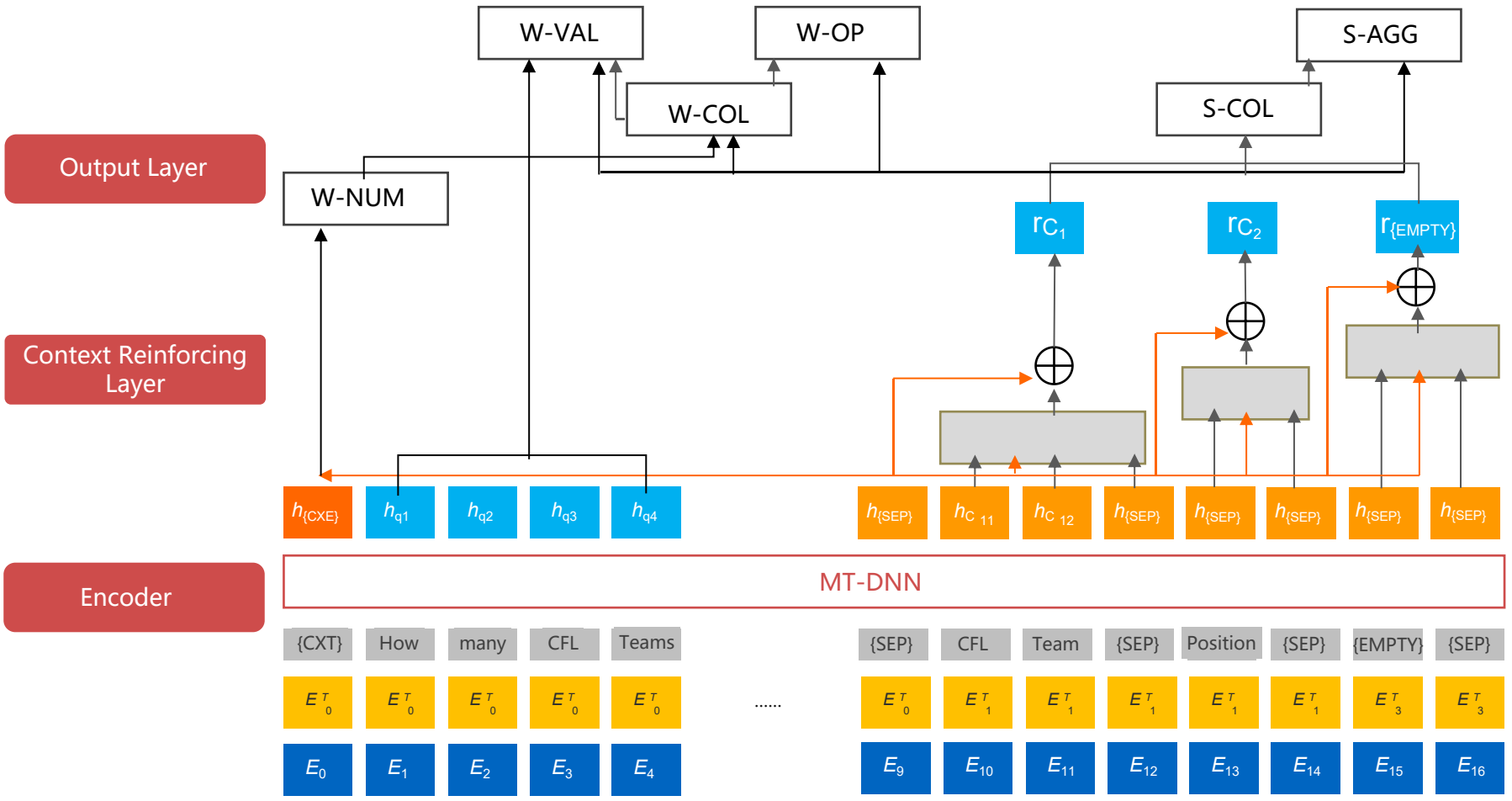
CleverBI NL2SQL原理




```
select 子公司 , AVG( 销售额 ) from 业绩表 where 日期=‘2018年’ group by 子公司
```

Percent百分点

基于槽位预测的X-SQL模型



Percent百分点

基于X-SQL和依存句法的NL2SQL新算法

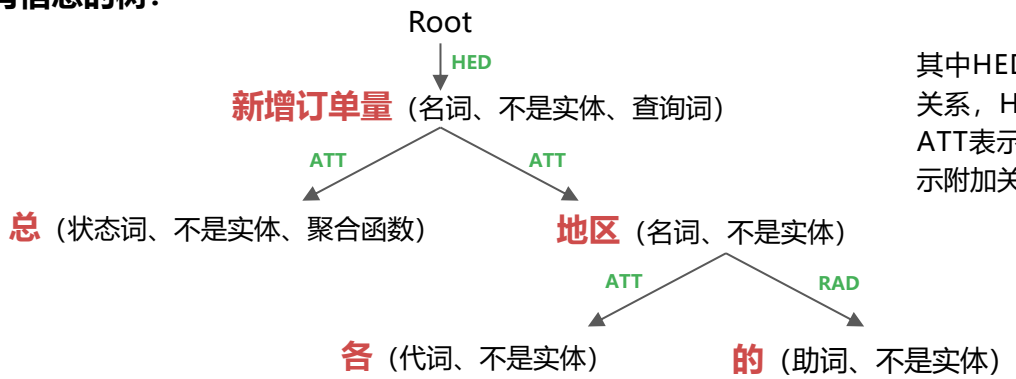
问题：**各地区的总新增订单量**

步骤1

分词后的结果（需要考虑字段名，X-SQL的结果）：**各　地区　的总新增订单量**

步骤2

得到的聚合了所有信息的树：



其中HED、ATT等表示依存关系，HED表示核心关系，ATT表示定中关系，RAD表示附加关系。

步骤3

通过词库以及后序遍历解析依存树

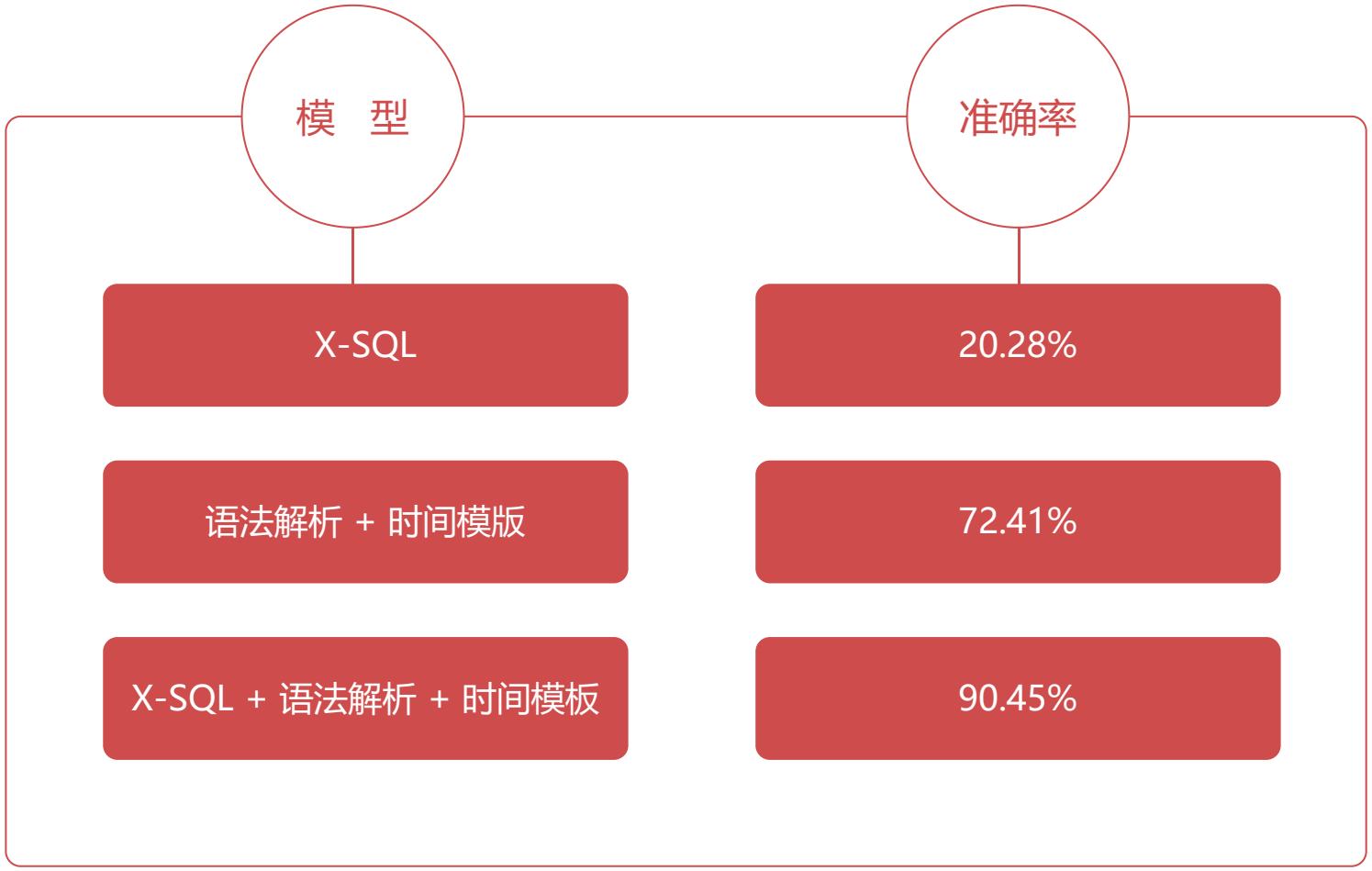
1.首先遍历到“**总**”。由X-SQL得知这是聚合函数。
2.遍历到“**各**”。得到这个是一个分组描述符。
3.遍历到“**的**”。得到这是一个无意义的词。
4.遍历到“**地区**”。从表中匹配得知这是一个字段名称，从孩子节点处得到的信息及ATT的关系，得知这是一个分组字段。
5.遍历到“**新增订单量**”。由X-SQL得知这是查询词，并且结合孩子节点得知聚合函数是“总”，分组词是“地区”。
6.遍历到root，得到最终结果：select内容为总新增订单量，分组字段为地区。

步骤4

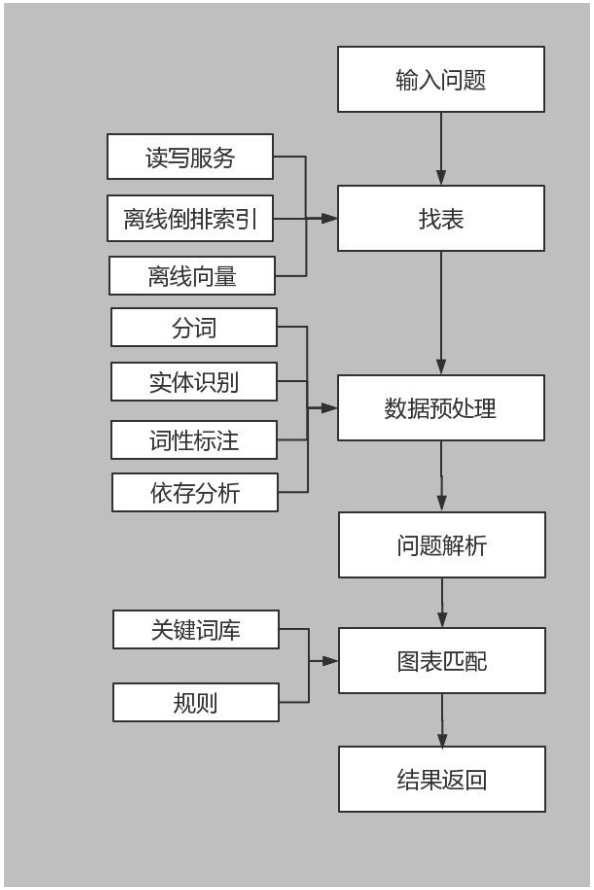
最终得到解析结果

```
时间模版：
{"template":"近${num1}天",
    "result":{"value":[{"year":"NOW_YEAR","month":"NOW_MONTH","day":"NOW_DAY-${num1}"}]}
}
```

实验效果



未来计划



- 在实际应用中大量收集数据，形成类似甚至超越Spider的中文NL2SQL标准数据集；
- 增加知识库优化问题解析效果；
- 找表、预处理和图表匹配步骤还需要大量的人工规则，利用半监督和无监督算法替换现有的算法，让整个过程更加灵活。

用数据智能推动社会进步

Percent百分点