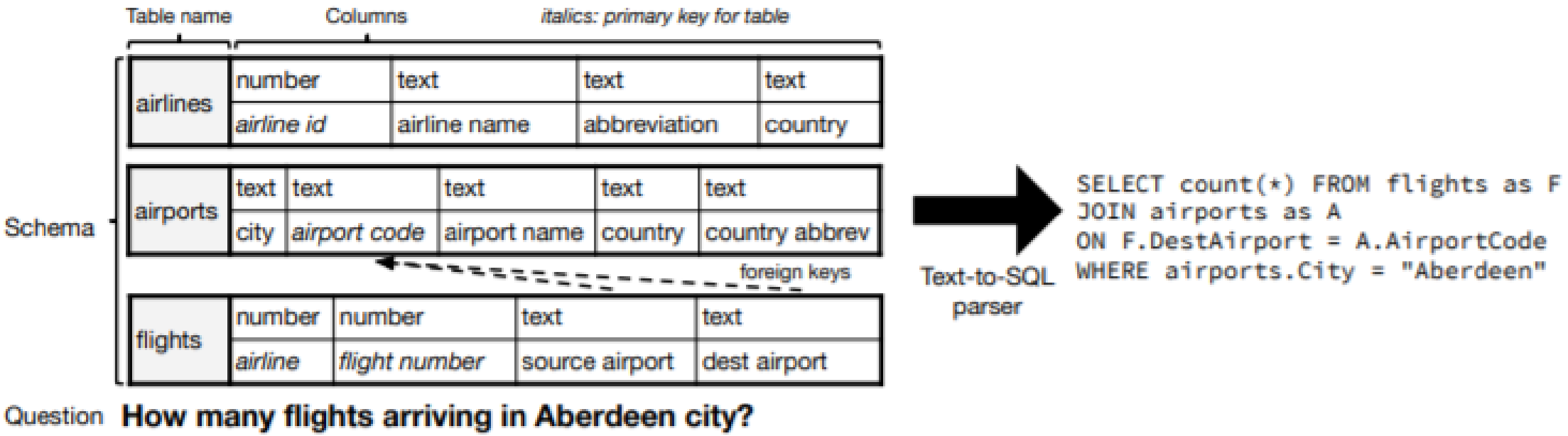
This Photo by Unknown author is licensed under CC BY-NC.

NL2SQL: An Overview

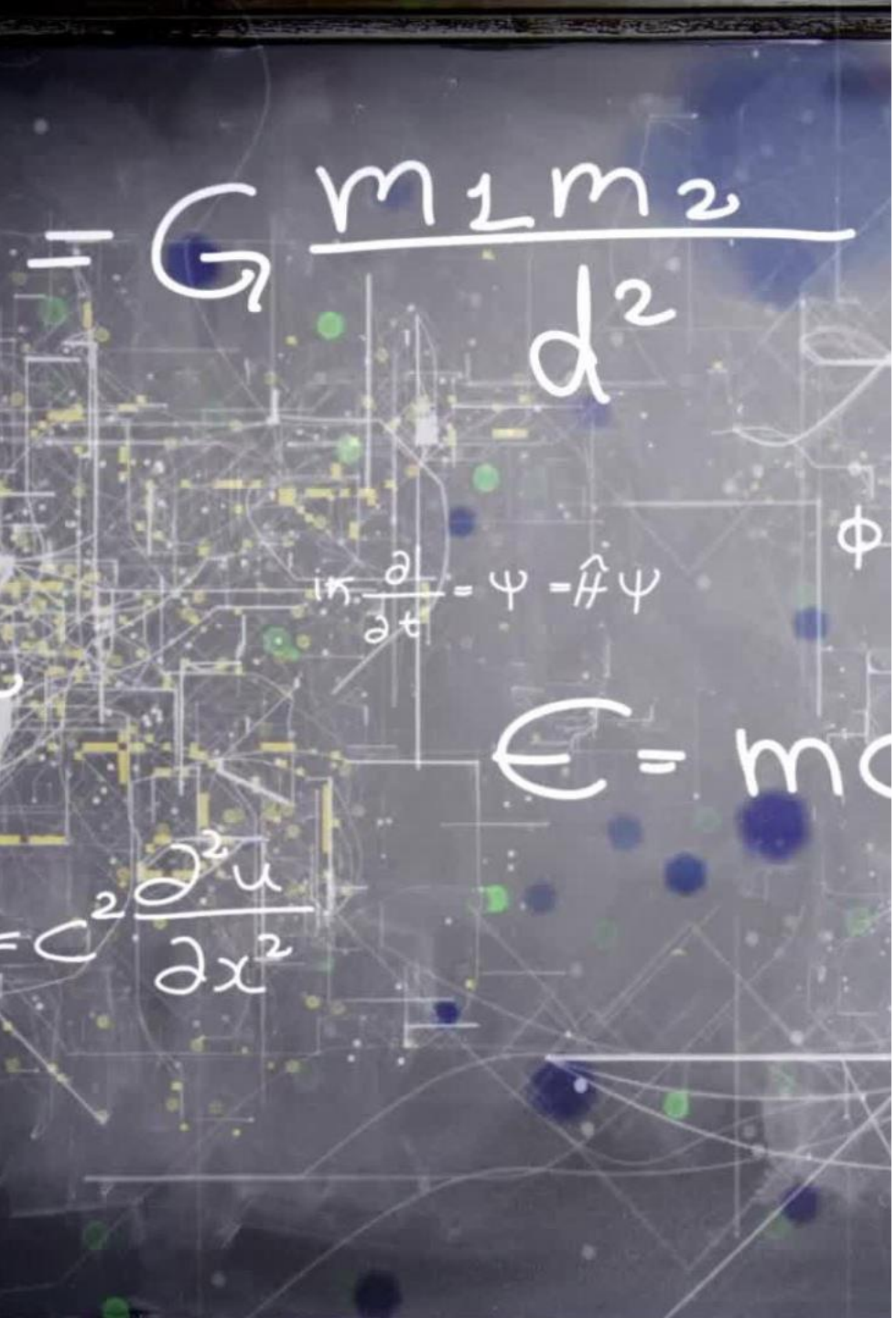
By Raymond Tomo

03/29/2022



Example of NL2SQL System

- In concept, it will be able to take a plain-text question written in a natural language from a user.
- It will then look at the schema, compare key words in the syntax of the natural language sentence, then convert it into a query as shown in the example above.



Early History

- 1980s-2000s
  - First concepts of NL2SQL were being proposed., first systems (LUNAR, CHAT-80) developed later on
  - Without the means to create a direct system yet, most of these proposed methods used an intermediary rather than a direct system through defined rules
  - For example, it would take a natural language question, convert it into a logical schema, then use that logical schema as an input, rather than a direct conversion
- 2000s-2020s
  - More advanced systems rolled out, using natural language parsers, but still relied on manually defined rules rather than being able to adapt
  - NaLIR (2014) - Introduced User Interaction to the system
  - ATHENA(2016) - Implemented domain specific concepts to help specify results

Recent History

With the rise of deep learning algorithms in many fields of computer science, NL2SQL has shifted in this direction

However, this has faced some challenges due to the lack of *training data,* which improves through repeated use of the system and feedback to enforce correct results.

WikiSQL now serves as a benchmark for new deep learning-based languages, encouraging meta-learning (systems learning from the output of other systems) and is available through github.

TypeSQL (2018) tags each word with a data type to describe it, then organizes it to form a query.

Current Systems + Capabilities: What they can do

- Currently, the more recent models of NL2SQL systems are capable of the following:
- Processing Simple Queries effectively: Most systems are capable of processing simple queries with between a 69-80% accuracy according to experiments performed on several systems.
- Occasionally processing more complex queries: While still lacking in many areas, some of the current systems can process more complex natural language queries on occasion.
- Follow learning protocols to improve: More recent models such as TypeSQL and WikiSQL can improve through the meta-learning discussed earlier.

Current Systems + Capabilities: How they do it

- At it's core, the systems follow the following steps to produce a query in the following order
  - Input- The natural language question such as 'Which flights took off at 8 am today?"
  - Input enrichment- Linking certain phrases in the text to parts of the schema in the database, or constant values (in this example, items such as 8 am or flights)
  - Translation- This is the step where the program either applies a fixed set of rules (Like Athena or NaLIR) or a Deep Learning method, determines how to generate this into SQL queries
  - Post-translation- This serves as an additional stage to complete translation
  - Training- Learning algorithms help improve the system by recognizing patterns in natural language to improve future results.
  - Output- This outputs the expected query based on the natural language input.

Research Opportunities

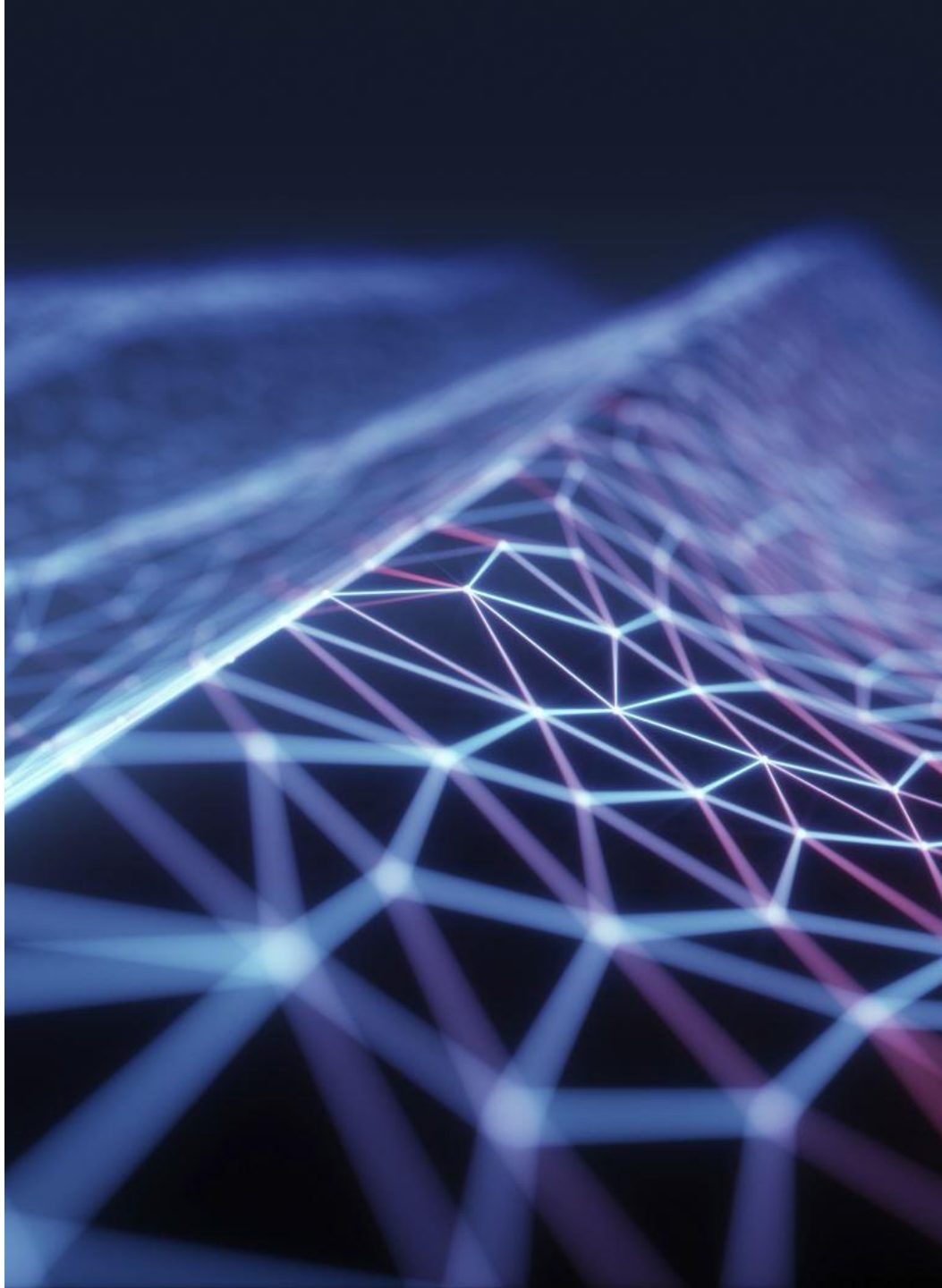
- The following fields in NL2SQL need improvement and are struggling to progress
  - Table/column reference alignment- methods that allow for improved matching of direct references to tables or columns
  - Extracting constant values- most methods available today are not able to extract constants from more complex queries.
  - Adaptability- Some patterns that have been
  - Deep Learning- Due to the small amount of training data available, a project to develop more training data for the sake of meta-learning may be promising.

What's next for NL2SQL?

One developing field takes NL2SQL a step further, using it as a basis for data visualization as proposed in the following paper.

Many language specific versions of nl2sql have been rolling out, such as mRAT-SQL+GAP, a Portuguese text to SQL transformer.

Along with this, new parsing techniques such as RaSaP (Relation Aware Semi-autoregressive Semantic Parsing) have made progress in teaching DL based systems better word representation, as show in this paper here.



What would I do?

I think one field that could prove extremely promising is deep learning, due to the prominence it has had on the area of research within a relatively short amount of time

Using machine learning to allow a system to understand human speech has already been done in a number of programs that are being widely used by many software companies

It will continue to get better at recognizing patterns through reinforced learning, causing imperfections to decrease overtime with proper reinforcement

References

- http://www.vldb.org/pvldb/vol13/p1737-kim.pdf
- https://arxiv.org/pdf/2108.00804.pdf
- https://github.com/salesforce/WikiSQL
- https://luoyuyu.vip/files/nvBench-SIGMOD21.pdf
- https://www.arxiv-vanity.com/papers/2110.03546/