Democratize Data Science
Natural Language Interface to Data

Xifeng Yan
with Izzedin Gur, Semih Yavuz, Yi Ding, Yu Su
Computer Science
University of California at Santa Barbara

Growing Gap between Human and Data

What disease does the patient have?

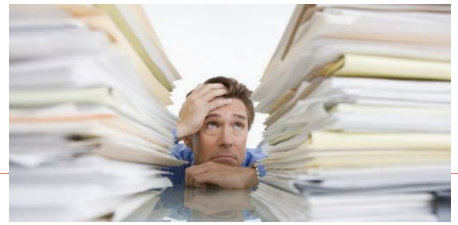
- (EMR) Similar patients?
- (Literature) New findings?
- (Gene sequence) Suspicious mutations?
- … …

Ad-hoc information needs for on-demand decision making

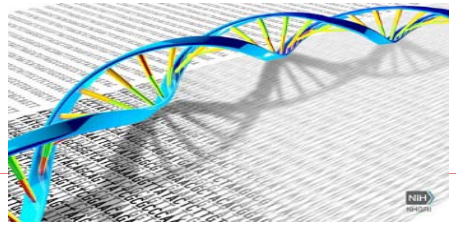
Massive, heterogeneous data

86.9% adoption (NEHRS 2015)

27M+ papers, >1M new/year (PubMed)



$1000 gene sequencing



24x7 monitoring

Data Science

Foundation

- Numerical Linear Algebra
- Optimization
- Stochastic Methods
- Discrete Mathematics and Algorithms

Data Science

- Introduction to Statistical Inference
- Databases and SQL
- Machine Learning
- Data Mining

Scientific Programming

- Data Analysis with Python
- Parallel Computing
- Distributed Algorithms and Optimization

Specialization

- Data Driven Medicine
- Modern Statistics for Modern Biology
- Representations and Algorithms for Computational Molecular Biology

Come On

"find all patients diagnosed with eye tumor"

```
WITH Traversed (cls,syn) AS (
    (SELECT R.cls, R.syn
    FROM XMLTABLE ('Document("Thesaurus.xml")
      /terminology/conceptDef/properties
      [property/name/text()="Synonym" and
      property/value/text()="Eye Tumor"]
      /property[name/text()="Synonym"]/value'
      COLUMNS
      cls CHAR(64) PATH './parent::*/parent::*
                         /parent::*/name',
      tgt CHAR(64) PATH'.') AS  R)
UNION ALL
    (SELECT CH.cls,CH.syn
    FROM Traversed PR,
      XMLTABLE ('Document("Thesaurus.xml")
      /terminology/conceptDef/definingConcepts/
      concept[./text()=$parent]/parent::*/parent::*/
      properties/property[name/text()="Synonym"]/value'
      PASSING  PR.cls AS "parent"
      COLUMNS
      cls CHAR(64) PATH './parent::*/
                         parent::*/parent::*/name',
      syn CHAR(64) PATH'.') AS CH))
SELECT DISTINCT V.*
FROM Visit V
WHERE V.diagnosis IN
  (SELECT DISTINCT syn FROM Traversed)
```

Asking a doctor to write this?

Execution time in mins
Coding in hours
Finding coders in weeks?

NCI thesaurus

"Semantic queries by example", Lipyeow Lim et al., EDBT 2014

What To Do?

Democratize

Data Science

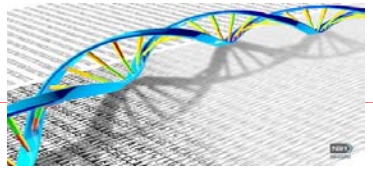
How can NLP/AI Bridge the Gap?

Insights
Discoveries
Solutions

Bottleneck #2: Access

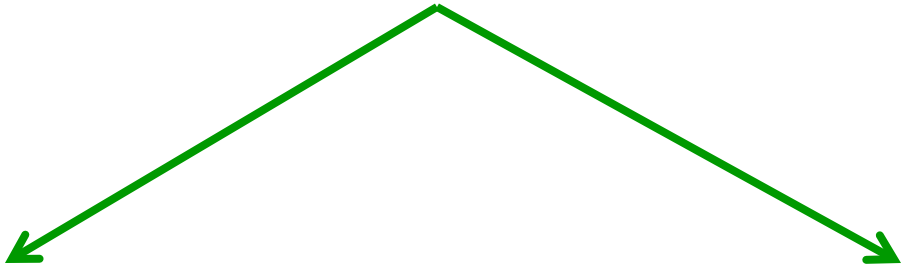
Bottleneck #3: Reasoning

Bottleneck #1: Knowledge



Natural Language Interface Is The Key

- Natural Language Query on Knowledge Graph



- NLI on Relational Data
- NLI on Everything
- Knowledge Graph Construction

Sequence to SQL (SPARQL)

Column Headers	First Name	Last Name	Gender	Age	Salary	Country
Sample Record	Dennis	Wright	Male	34	112,751	Philippines

Q: What is the average age for male employees
A: 44.69

Q: How old are male employees in Philippines?
A: 6196.18

Q: What is the average age for male employees whose salary is great than 100K?
A: *show average of male employees age grouped by salary*

query examples from: http://kueri.me/demo

What Is the Issue?

NLP: We are working on NLI on Databases

DB : It is not new. Can you guarantee the correctness?

NLP: No.

DB : What is your accuracy on WikiSQL [Zhong 2017]?

NLP: 68%.

DB : It is low!

NLP: We can improve it 2% per published paper.

DB : Given a question, how can a user know if the answer is correct or not?

NLP: There is 68% chance for the answer to be correct.

DB : Forget about it!

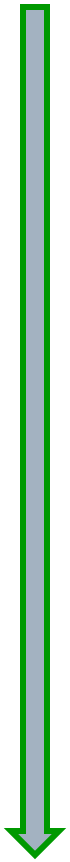
NLP: But at least the user can look at the generated SQL statement?

Some Thoughts

- Two Conclusions
  - In a short period, likely we are not able to break the accuracy barrier, anything close to 100%
  - If we ask a user to check the correctness by reading SQL, we already fail
- Can we still make a dent?
  - Start with simple queries
  - Leverage human intelligence
  - Standardize query benchmark
- Set up am agenda

Level of Driving Automation vs Level of NLIDB

Level	Driving Automation*	NLIDB
Level 0		SQL
Level 1	Hands on	Schema-aware, simple question
Level 2	Hands off	Schema-aware, sequential simple questions
Level 3	Eyes off	Schema-free, sequential simple questions
Level 4	Mind off	Schema-free, composite questions + some assistance
Level 5	Steering wheel optional	Any Question + no assistance



*: from https://en.wikipedia.org/wiki/Autonomous_car

Some Progress From Peers

- Simple Dataset: WikiSQL Data [Zhong 2017]
  - There are a few negative comments on the simplicity of the task
  - But come on, if you think this dataset or query is too simple, show me a query engine with 100% accuracy
- Sequence2SQL: e.g., Seq2SQL [Zhong 2017], SQLNet[Xu 2017], [Wang 2017], etc.
- Break query to small sequential simple queries. e.g. [Iyyer 2017][Saha 2018]
- SQL Statement translated back to natural language question: e.g., [Deutch 2017]

SQLNet (Dawn Song's group)

Year	Award	Category	Nominee
1986	Tony Award	Best Musical	Best Musical
1986	Tony Award	Best Direction of a Musical	Bob Fosse

Question: **Which award has the category of the best direction of a musical?**

Correct SQL:

```
SELECT award
WHERE category = best direction of a musical
```

SQLNet:

```
SELECT award
WHERE category = the best direction award a musical
```

Why Not Use Table Cells?

- You are “cheating.” :(
- Do you mean you can achieve 100% accuracy if you use table cell information?

Models	Dev(where)	Test(where)
Seq2SQL	62.1%	60.2%
SQLNET (Seq2set + CA + WE)	**74.1%**	**71.9%**
Using Table Cells (Baseline)	77.2%	77.1%
Using Table Cells (Upper Bound)	92.1%	91.4%

Error Analysis

Semantic Matching (31%): Our model simply cannot find the correct column semantically better match to the context of the value in question

Question	What elimination move is listed against wrestler Henry, eliminated by Batista?		
Condition Value	Henry	Candidate Columns	{wrestler, eliminated by}
Correct Column	wrestler	Model Prediction	eliminated by

Ambiguous Question (27%): Question does not carry sufficient information

Question	Who was the director of the king's speech?		
Condition Value	the king's speech	Candidate Columns	{winner and nominees, original title}
Correct Column	original title	Model Prediction	winner and nominees

Error Analysis (cont.)

Missing/Extra Value (25%): Missing (or extra) condition value that occurs (or does not occur) as value of a condition in where statement

Sparsity (13%): Candidate column name appears too rarely (less than 5 times) in the training data.

Wrong labeling (4%): Generated (paraphrased) question for SQL query by AMT worker is wrong.

Human and Machine Intelligence Integration

Year	Mens singles	Womens singles	Mens doubles	Womens doubles	Mixed doubles
...	...	...	...	...	...
2001	aivaras kvedarauskas	neringa karosait	aivaras kvedarauskas juozas spelveris	kristina dovidaityt neringa karosait	aivaras kvedarauskas ligita zakauskait
...	...	...	...	...	...

- Asking a computer to completely understand the above table is hard. It is like to ask a 4 year old kid without knowledge about tennis to understand this table
- However the query issuer likely understands the table

Make the Question Simpler

Imperative Query

What is the number of movies featuring Brad Pitt?

- Identify aid that matches Brad Pitt's name in table "actor"
- Search aid in table name "cast"
- Count movies in table name "movie" with matching "msid" in table name "cast"

Imperative Query Model

- SELECT aid FROM actor WHERE name = "Brad Pitt"
- SELECT * FROM cast where aid = ?
- SELECT COUNT(*) FROM movie WHERE msid = ?

SELECT COUNT(*) FROM movie WHERE msid = (SELECT msid FROM cast WHERE aid = (SELECT aid FROM actor WHERE name = "Brad Pitt"))

Standardize Query Benchmarks

- Level 1: Schema-aware, simple question
- Level 2: Schema-aware, sequential simple questions
- Level 3: Schema-free, sequential simple questions
- Level 4, 5: Composite questions

- WikiSQL [Zhong 2017] can be adapted as query benchmark for Level 1
- SequentialQA [Iyyer 2017] good for Level 3
- Level 4, 5: WebQuestions[Berant 2013], WikiTableQuestions [Pasupat 2015], GraphQuestions [Yu 2017]
- We need to characterize the complexity of questions so that we can better understand the QA behavior

Recap

- The first step to democratize data science is to let people freely access their data without knowing programming language
- In order to make it happen, start with simple queries first. Make it work
- Move up the ladder solidly
- Leverage human intelligence