Machine Learning for Harvesting Health Knowledge

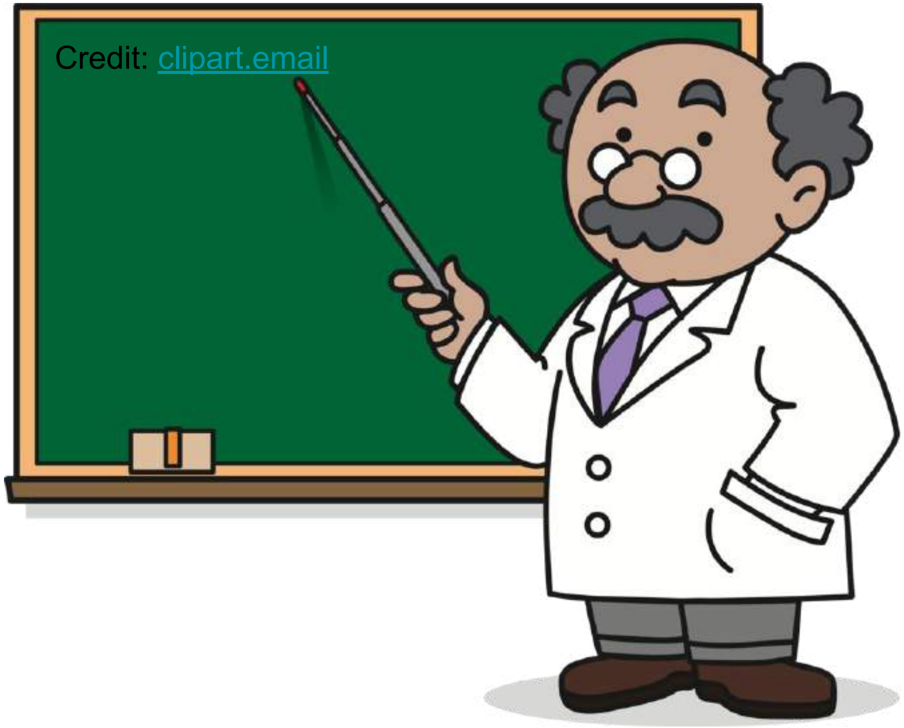
Block Seminar - Saarland University

7 May 2020

max planck institut informatik

Instructors

1. Patrick Ernst
2. Erisa Terolli
3. Andrew Yates



Erisa Terolli

- Short CV
  - Computer Engineering Diploma from Epoka University, Albania.
  - PhD in Computer Science, Sapienza University of Rome, Italy.
  - Post-Doc Researcher at MPII.
- Research Interest
  - IR for Biomedicine.
  - Graph Mining.
  - Social web data modeling and analysis.
- Email
  - eterolli@mpi-inf.mpg.de

Patrick Ernst

- Short CV
  - Master of Science from University of Kaiserslautern
  - PhD in Computer Science, University of Saarland/MPII
  - Post-Doc Researcher at MPII
  - Machine Learning Scientist with Amazon
- Research Interest
  - Knowledge Bases and IR for Biomedicine
  - Personalization
- Email
  - peernst@mpi-inf.mpg.de

Andrew Yates

- Short CV
  - Computer Science BSc from Illinois Institute of Technology, Chicago, IL, USA
  - Computer Science PhD from Georgetown University, Washington, DC, USA
  - Senior Researcher at MPII
- Research Interest
  - Information Retrieval: biomedical applications, neural methods, and personalization
  - NLP: biomedical applications, personal knowledge base construction, and credibility analysis
- Email
  - ayates@mpi-inf.mpg.de

Basic Seminar Info

- Type: Block Seminar
- Number of credits: 7 ECTS
- Lecture/Meeting:
  - 7 May 2020 - Introductory Lecture
  - August 2020 - 2 day block seminar (TBD)
- Room: Zoom until a further notice
- Materials: will be put on the seminar web-page

Main Blocks

- Five Topics
  - Information Retrieval, Automatic Health Assessment, Social Media Analysis, Information Extraction, Conversational AI
- Two scientific publications
- Written report
  - Hand in your write-up in pdf format before the specified deadline.
  - 8 pages including references.
  - Obeye the scientific standards and avoid plagiarism!
  - Compulsory midterm meeting with instructor.
- Peer Review report
  - Hand in your review in pdf format before the specified deadline.
- Oral Presentation
  - 25 minutes plus 10 minutes discussion.
  - Compulsory. You fail if you do not show-up for the oral presentation.

Topics Distribution

- Express your topic preferences.
  - Pick your top three topics by Saturday (May 9) at https://forms.gle/ERTNXz5N53rzbBcm9
  - Map each students with their top preferences
  - Conflict: Break the ties arbitrarily
- Each student will be matched with a primary topic.
- Each student will be given a secondary topic for peer reviewing.
- Each student will be matched with one instructor.
- All assignments will be made by May 11.

Machine Learning for Harvesting Health Knowledge (Block Seminar)

Choose your top 3 preferences for the topics of the Machine Learning For Harvesting Health Knowledge Block Seminar. Deadline for filling is this form is May 10, 2020 at 23:59.

Full Name *

Short answer text

Matriculation Number *

Short answer text

Email Address *

Short answer text

What is your 1st preference? *

- Information Retrieval
- Automatic Health Assessment
- Social Media Analysis for Health Care
- Information Extraction
- Conversational AI

What is your 2nd preference? *

- Information Retrieval
- Automatic Health Assessment
- Social Media Analysis for Health Care
- Information Extraction
- Conversational AI

What is your 3rd preference? *

- Information Retrieval
- Automatic Health Assessment
- Social Media Analysis for Health Care
- Information Extraction
- Conversational AI

Seminar Timeline

- May 9: Students pick their top 3 preferential topics.
- May 11: Topic Distribution.
- June 16: Midterm Meeting with Instructors.
- July 16: Technical Report Submission Deadline.
- August 6: Review Submission Deadline.
- August 20: Final Report Submission Deadline.
- August: Two day block seminar for oral presentations (TBD).

Evaluation

1. Technical Report (max 50 points)
2. Oral Presentation (max 30 points)
3. Peer Review (max 20 points)

Grades

- >= 90: 1
- >= 80: 2
- >= 70: 3
- >= 60: 4
- < 60: 5

https://www.pngfuel.com/

What makes a good technical report?

- Should NOT be just a summary of your assigned papers.
- Review the literature for your assigned topic.
- Contextualize general approaches of your topic to the medical domain.
- Accurate
- A fluent narrative
- Concise and Clear
- Comprehensive

A good review should be:

- Focused
    - Focus on the most important elements of the report.
- Reasonable
    - Make realistic requests that are relevant to the report. Avoid “Nice to have” changes.
- Critical but Constructive
    - Address problems clearly.
    - Write suggestions on why and how could the suggested problems should be tackled.
- Structured
    - Write a brief summary: Shows you got the key points.
    - Address problems on Major vs Minor Points.
    - Ideally write a paragraph for each Major Point.
- Polite and Professional
    - Express your views fairly but POLITELY.

www.chemistryworld.com

Preparing your oral presentation

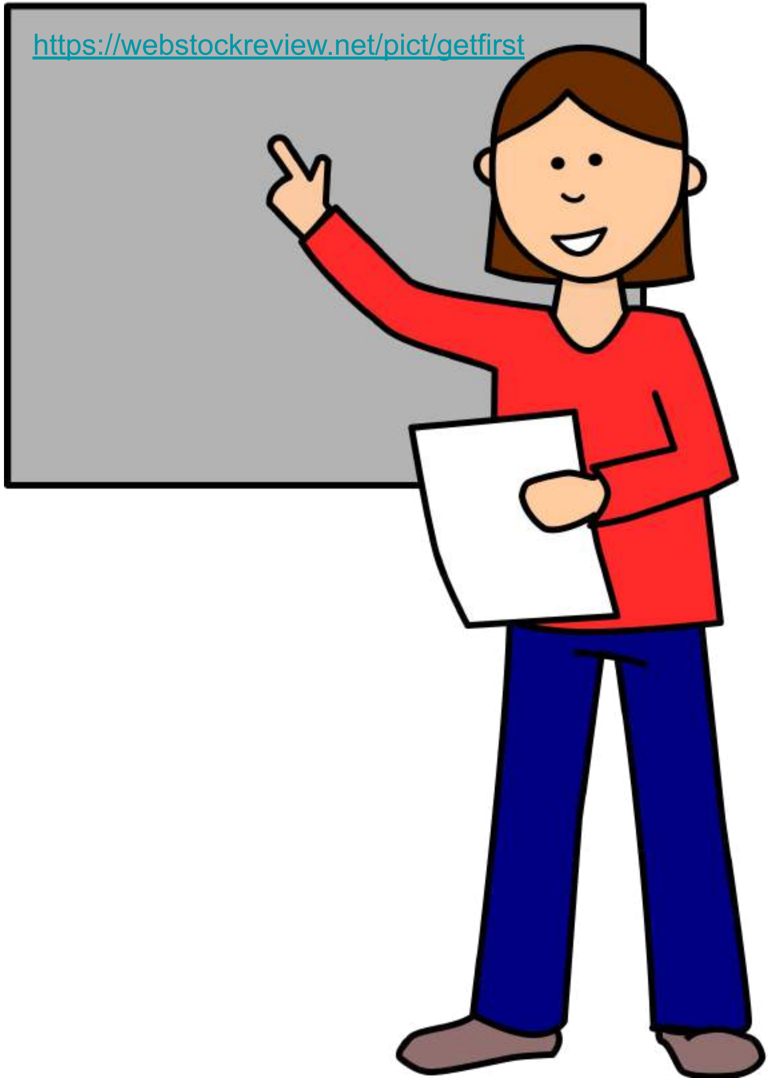
- Communicate some information to an audience.
- A presentation should be: Informative and Interesting.

Tips:

- Organize your thoughts
  - Start with an outline and develop good transition between sections.
- Have a strong opening
  - Why should people listen to you?
- Finish with a bang
  - Finish with a couple of sentences that sum up the importance of your work.
- Time yourself
- Practice a lot

Presenting...

- Excitement
- Speak with confidence
- Make eye contact with the audience
- Avoid reading your presentation
- Leave some time for QA



Resources

- Seminar web-page:
  - https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/teaching/summer-semester-2020/machine-learning-for-harvesting-health-and-life-science-knowledge/
- Topics Preferences Form:
  - https://forms.gle/ERTNXz5N53rzbBcm9
- Technical Report Template:
  - https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sig-proceedings-template/bmvfhcdnxfty
- Peer Review Report Template:
  - https://docs.google.com/document/d/13I1Kao4elsDBKv205Gy6snetoe8ML4JRjz69LgN65W8/edit?usp=sharing

Questions?

Topic Explanations

Five Topics

- Information Retrieval
- Social Media Analysis
- Automatic Health Assessment
- Information Extraction
- Conversational AI

Information Retrieval

IR: finding resources to satisfy a user's information needs

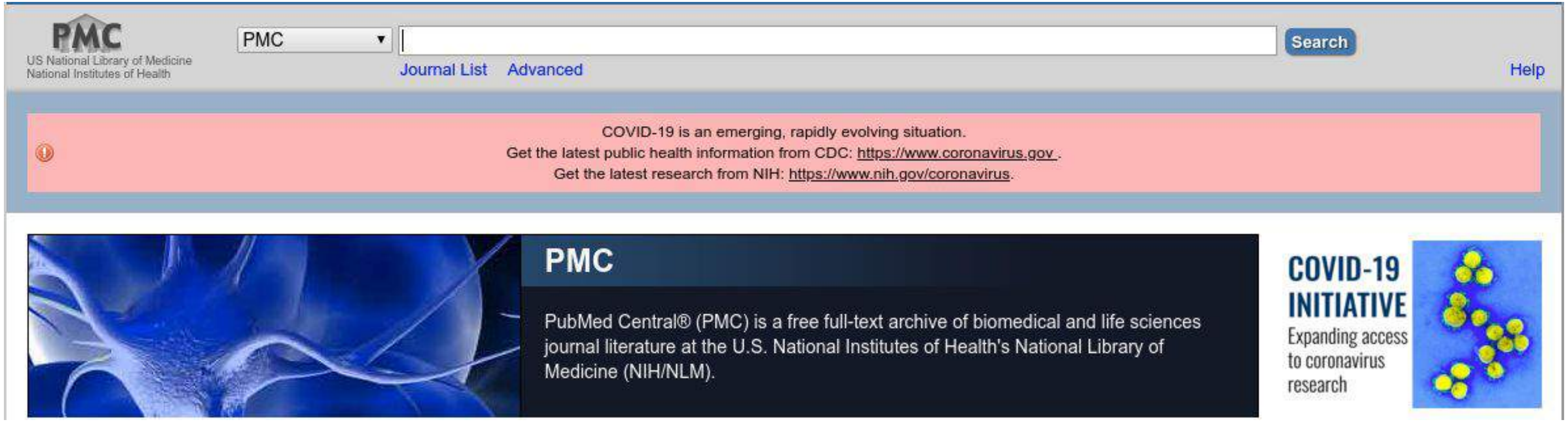
In the context of health/medicine, this is often finding relevant biomedical literature

- *remdesivir severe acute respiratory syndrome*

...or finding credible articles written for laypeople (non-experts)

- *"What are the symptoms of COVID-19?"*
- *"coronavirus symptoms"*

Information Retrieval



PubMed: a repository of biomedical literature used by experts

Information Retrieval

Journal List > J Glob Health > v.10(1); 2020 Jun > PMC7125419

The University of Edinburgh
Edinburgh University Global Health Society
journal of global health

J Glob Health. 2020 Jun; 10(1): 011001.
Published online 2020 Mar 31. doi: 10.7189/jogh.10-011001
PMCID: PMC7125419
PMID: 32257173

J Glob Health

An evidence-based framework for priority clinical research questions for COVID-19

Carlyn Harris,[1,2] Gail Carson,[3,4] J Kenneth Baillie,[4,5] Peter Horby,[3,4] and Harish Nair[2]

▸ Author information ▸ Copyright and License information Disclaimer

Abstract

Go to:

Background

On 31 December, 2019, the World Health Organization China Country Office was informed of cases of pneumonia of unknown aetiology. Since then, there have been over 75 000 cases globally of the 2019 novel coronavirus (COVID-19), 2000 deaths, and over 14 000 cases recovered. Outbreaks of novel agents represent opportunities for clinical research to inform real-time public health action. In 2018, we conducted

Formats:

Article | PubReader | ePub (beta) | PDF (343K) | Citation

Share

Facebook Twitter Google+

Save items

Add to Favorites

Similar articles in PubMed

World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). [Int J Surg. 2020]

COVID-19: The outbreak caused by a new coronavirus. [Bol Med Hosp Infant Mex. 2020]

The novel zoonotic COVID-19 pandemic: An expected global health concern. [J Infect Dev Ctries. 2020]

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (CO [Int J Antimicrob Agents. 2020]

Novel coronavirus 2019-nCoV: prevalence, biological and clinical characteristics comparison wi [Eur Rev Med Pharmacol Sci. 2020]

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7125419/

Information Retrieval

```xml
<topic number="1">
  <disease>melanoma</disease>
  <gene>BRAF (E586K)</gene>
  <demographic>64-year-old female</demographic>
</topic>

<topic number="4">
  <disease>Breast cancer</disease>
  <gene>FGFR1 Amplification, PTEN (Q171)</gene>
  <demographic>67-year-old female</demographic>
  <other>Depression, Hypertension, Heart
  Disease</other>
</topic>
```

Queries from *TREC Precision Medicine*

Information Retrieval

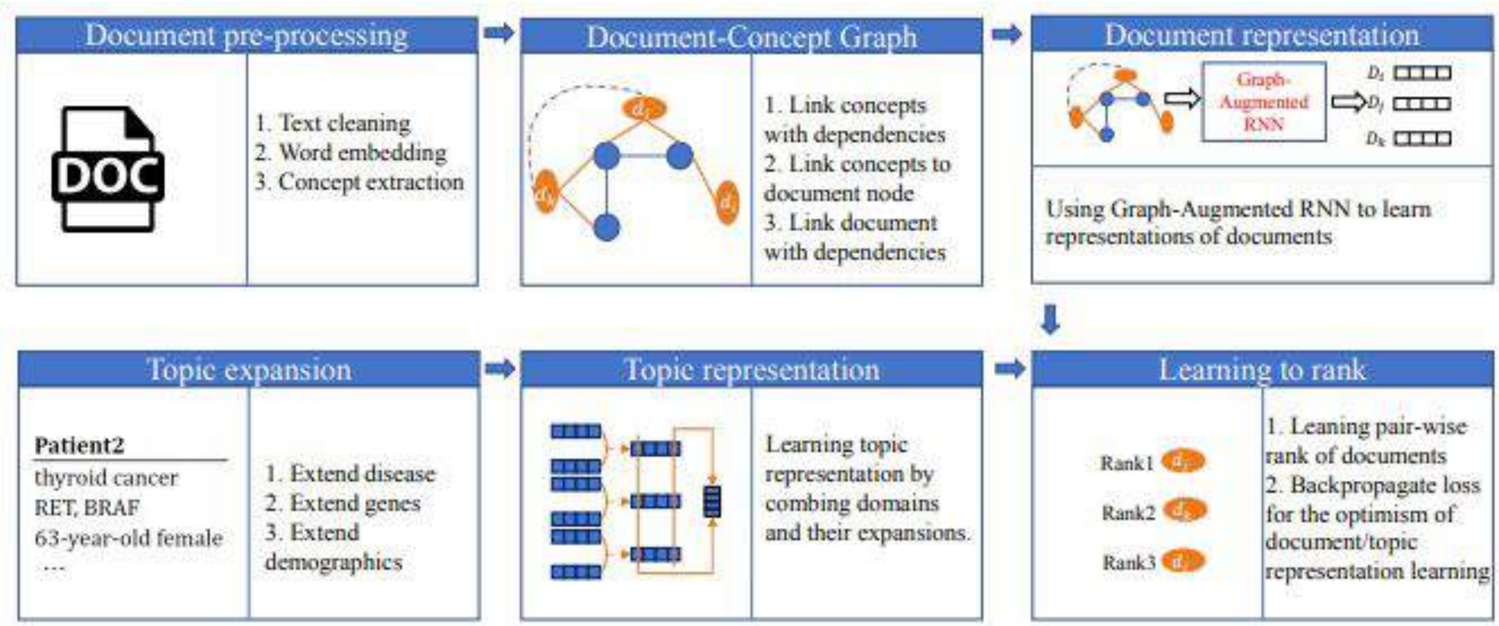
```
<topic number="30">
  <query>coronavirus
  remdesivir</query>
  <question>is remdesivir an effective
  treatment for COVID-19</question>
  <narrative>
    seeking specific information on
    clinical outcomes in COVID-19
    patients treated with remdesivir
  </narrative>
```

```
<topic number="18">
  <query>masks prevent coronavirus</query>
  <question>
    what are the best masks for preventing
    infection by Covid-19?
  </question>
  <narrative>
    What types of masks should or should not
    be used to prevent infection by Covid-19?
  </narrative>
```

Queries from *TREC COVID Challenge*

Information Retrieval: Biomedical Literature

(Zhao et al 2019) propose a neural framework for retrieving biomedical literature



Information Retrieval: Clinical Decision Support

(Alsulmi and Carterette 2016) investigate query reformulation strategies for improving Clinical Decision Support (CDS) search to identify relevant articles

- In CDS, a clinical case report is the query
- Often a vocab mismatch between the query and relevant scientific literature

SAMPLE TOPICS FOR TREC CDS TRACK SHOWN AS CLINICAL CASES.

Topic Type	Clinical Summary
Diagnosis	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.
Test	40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.
Treatment	63-year-old heavy smoker with productive cough, shortness of breath, tachypnea, and oxygen requirement. Chest x-ray shows hyperinflation with no consolidation.

Information Retrieval: Conclusion

Finding documents to satisfy a user's biomedical information needs

- What literature is available about this disease given patient's characteristics?
- Given a clinical case report, what articles support a treatment/test/diagnosis?
- What articles address a layperson's query?

Key point: biomedical queries to retrieve biomedical information,
which may be written for experts or for lay people

Social Media Analysis

Large & growing amount of health-related information on social media

- 8% of US adult internet users “have posted a health-related question or comment online within the past year” (Survey by Pew Research)

Social Media (Twitter, Reddit, specialized forums, etc) provide unique opportunities to observe users’ behavior:

- “*I’ve had trouble sleeping since starting Prozac (fluoxetine)*”
- “*Zoloft is making my depression worse, so I’m changing meds next week*”

Idea: use this observational data to enable applications, such as

- Assessing drug effectiveness
- Discovering unknown drug side effects
- Estimating disease prevalence

Social Media Analysis

Social media also brings unique difficulties, such as

- Colloquial terminology / Layperson vocabulary (that is often verbose)
  - "*heart palpitations*" (expert term)
    MayoClinic: *feelings of having a fast-beating, fluttering or pounding heart*
  - "*my heart is beating fast*" (colloquial)
  - "*my chest is pounding*" (colloquial)
  - "*pain in my chest*" (different)
- Causality: are the palpitations a side effect or a symptom of health condition?
- Credibility/Accuracy: is it truthful and relevant for the intended use case?
  "*My heart is beating fast -- yours could be too with a cup of Folgers coffee!*"

Social Media Analysis: Adverse Drug Events

Adverse Drug Event (ADE): “an injury caused by taking medication” (Wikipedia)

- i.e., a negative drug side effect. Also called Adverse Drug Reactions (ADRs)
- *Pharmacovigilance* is the monitoring of ADEs

(Lee et al. 2017) considers post-market pharmacovigilance using Twitter

Table 1: Example of ADE and non-ADE Tweets

Class	Tweet
ADE	Oh yay, **Niaspan** reaction. *Face burning up.*
Non-ADE	*My face is on fire* and **Tylenol** isn’t helping. *I’m burning up.* Fingers crossed I’m not getting sick.

Approach: classification task using a
semi-supervised neural network (CNN)

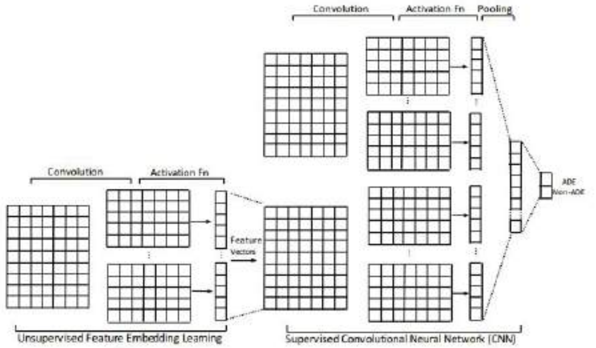


Figure 1: Semi-Supervised CNN.

Social Media Analysis: Drug Effectiveness

Drug Effectiveness: a drug’s ability to cure a disease

- i.e., whether taking a drug helped the patient

(Chai et al 2019) study drug efficacy by performing *relation extraction* on tweets

Tweets	Relations
(i) I started off on Zoloft, I’m going next week to get changed. it helps my **ocd** and **anxiety**, but made my **depression** worse.	**better:** <ocd, Zoloft>, <anxiety, Zoloft> **worse:** <depression, Zoloft>
(ii) No **cold** medicine has ever cured a **cold**. # Mucinex # Robitussin.	**maintain:** <cold, Muxinex>, <cold, Robitussin>
(iii) OK I must sleep now. Despite all the normal meds I take at bedtime I had to add in an Imitrex for the **migraine** I feel coming.	No relation exists (since the effect of “migraine” is not mentioned).

...using a graph of chemical (drug) and disease mentions

Social Media Analysis: Conclusion

Using social media to *learn about health topics through observational studies*

- What claims do people make?
- How do the claims relate to information from other sources?

Key point: aggregating information across users to study a topic

Automatic Health Assessment

Social Media and other user-generated data can also be used to assess a user

- The goal is to assess a given user, whereas in the previous topic the goal was to conduct observational studies across users

Mental health in particular has a unique connection to language

- Can we tell when someone is depressed? Or at risk of self-harm?
- …without an explicit mention of either?

When someone makes a health-related claim, is it accurate?
"I just had a heart attack"

Example Application

ReachOut.com Forums

ReachOut Forums is a supportive, safe and anonymous space where people care about what's happening for you, because they've been there too.

Read what others are saying about similar situations:

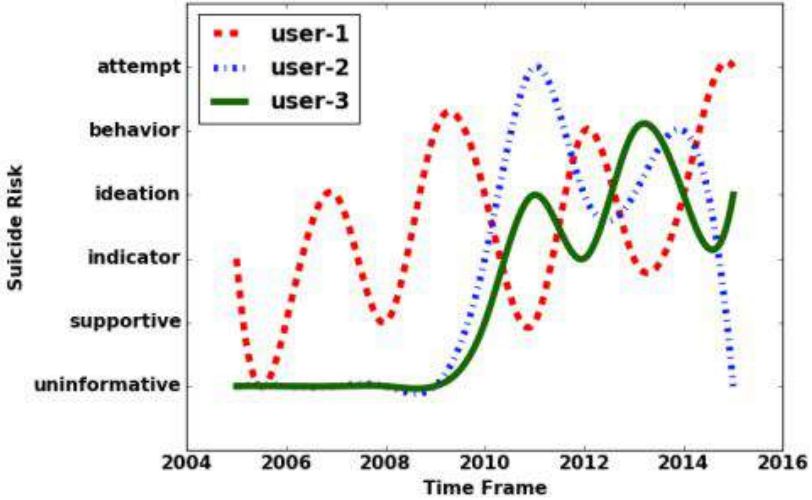
- Get insight into what's happening for you
- Ask questions if you want to

Is a given user at risk?

GREEN	AMBER	RED
I'm proud that I was able to call and keep up a phone conversation with my mum.	There are so many stuff I'm thinking about, but my medications are slowing my thoughts down and making it more manageable	I feel helpless and things seem pointless. I hate feeling so down

Auto Health Assessment: Severity of Suicide Risk

(Gaur et al 2019) automatically determine whether a user is at risk of suicide



…by identifying mentions of suicidal thoughts and actions in the user's posts and using a neural network for text classification .

Auto Health Assessment: Personal Health Mentions

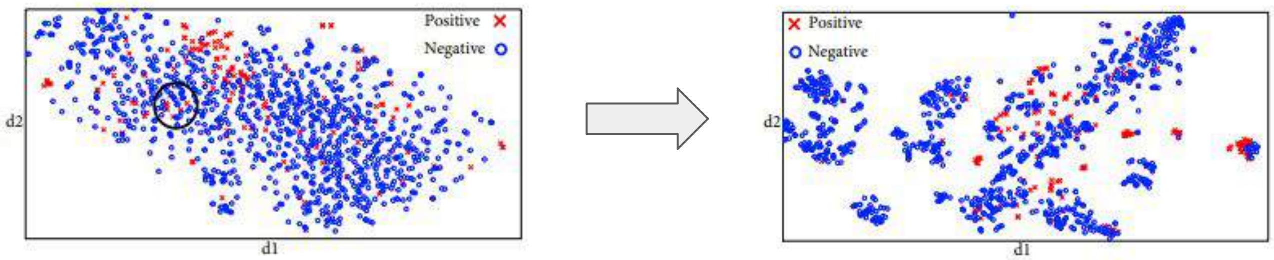
Personal Health Mentions may indicate a user has experienced a condition or event

- Previously, the assessment was an inference based on the user's data
- This assessment is of whether the user is describing a real event

(Karisani and Agichtein 2018) detect whether text contains a personal health event

- *"I almost had a heart attack when I found out they're doing a lettering workshop at @heathceramics"*
- *"My mom died to lung cancer thanks to smoking for like 40 years."*

Approach: represent as word embeddings; modify embedding space to improve classification



Auto Health Assessment: Conclusion

Making predictions about a user's health status

- Can we infer that a user has some health condition, is at risk, etc?
- Is a user stating that they have some health condition?
  (or is making some other health-related claim?)

Key point: assessing a **user's** activity to learn about the user's health

References

From seminar website:

- Information Retrieval
  - Mohammad Alsulmi and Ben Carterette. 2016. Improving clinical case search using semantic based query reformulations. In Bioinformatics and Biomedicine (BIBM'16).
  - Sendong Zhao, Chang Su, Andrea Sboner, and Fei Wang. 2019. GRAPHENE: A Precise Biomedical Literature Retrieval Engine with Graph Augmented Deep Learning and External Knowledge Empowerment. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM ’19).
- Automatic Health Assessment
  - Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In The World Wide Web Conference (WWW ’19).
  - Payam Karisani and Eugene Agichtein. 2018. Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media. In Proceedings of the 2018 World Wide Web Conference (WWW ’18).
- Social Media Analysis for Health Care
  - Kathy Lee, Ashequl Qadir, Sadid A. Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017. Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In Proceedings of the 26th International Conference on World Wide Web (WWW ’17).
  - Zi Chai, Xiaojun Wan, Zhao Zhang, and Minjie Li. 2019. Harvesting Drug Effectiveness from Social Media. In Proceedings of the 42nd International ACM SIGIR Conference (SIGIR’19).

Information Extraction

Goal

Extract structured information from noisy, highly-unstructured input data

Facilitates:

- Information Retrieval
- Reasoning
- Information Discovery

Encyclopedias

Scientific Literature

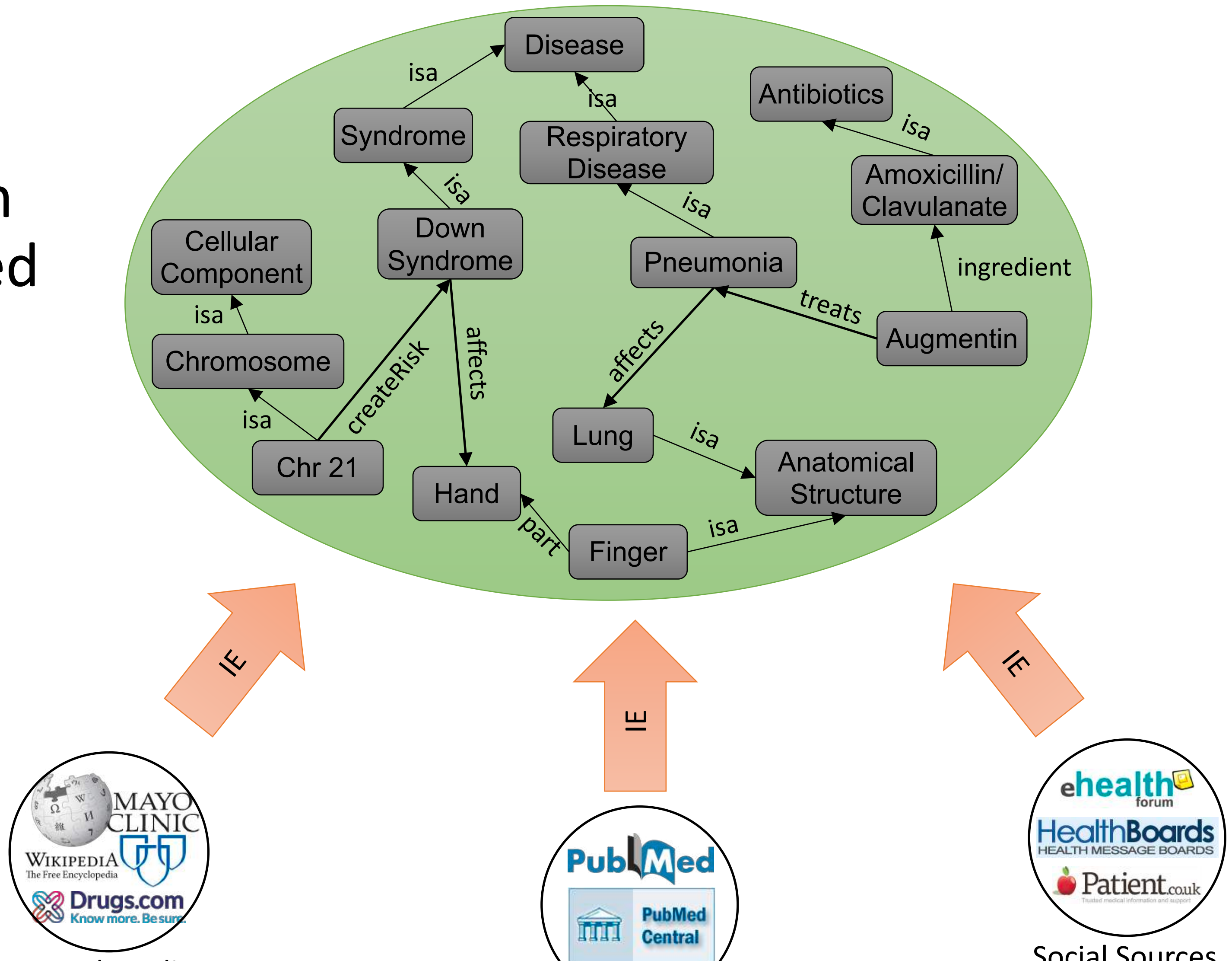
Social Sources

Goal

Extract structured information from noisy, highly-unstructured input data

Facilitates:

- Information Retrieval
- Reasoning
- Information Discovery



Encyclopedias

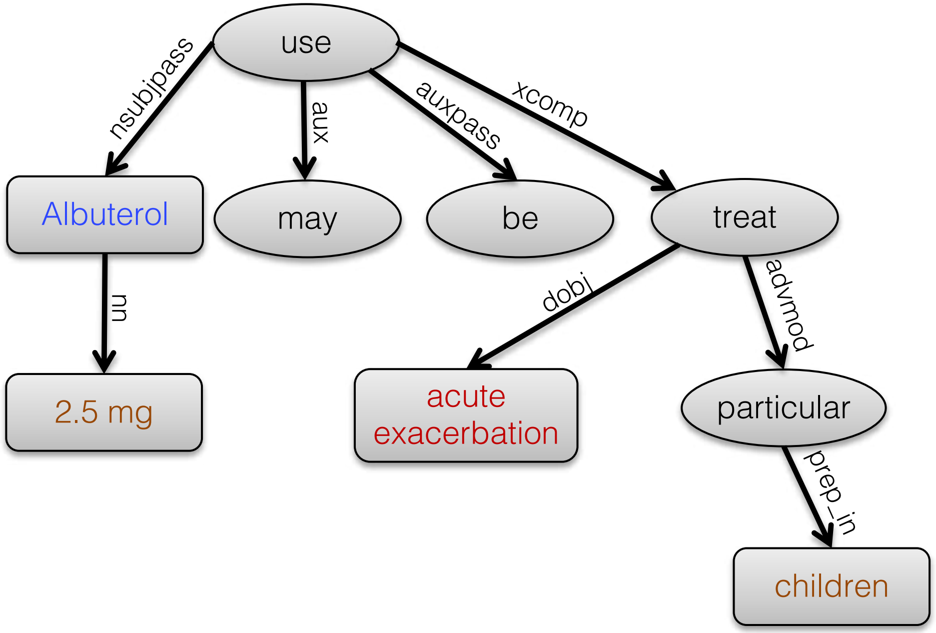
Scientific Literature

Social Sources

Ambiguity

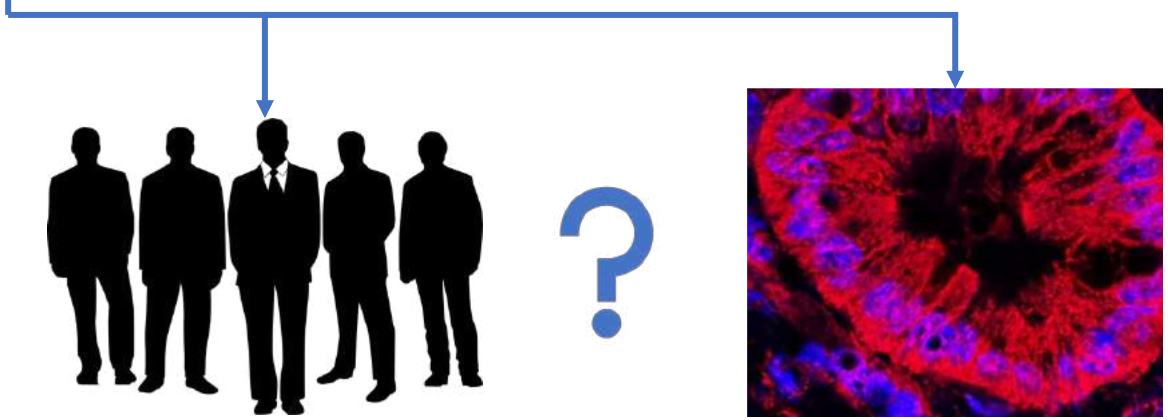
Syntactical ambiguity: finding the correct grammatical or structural interpretation of human text

2.5 mg Albuterol may be used to treat acute exacerbations, particularly in children.



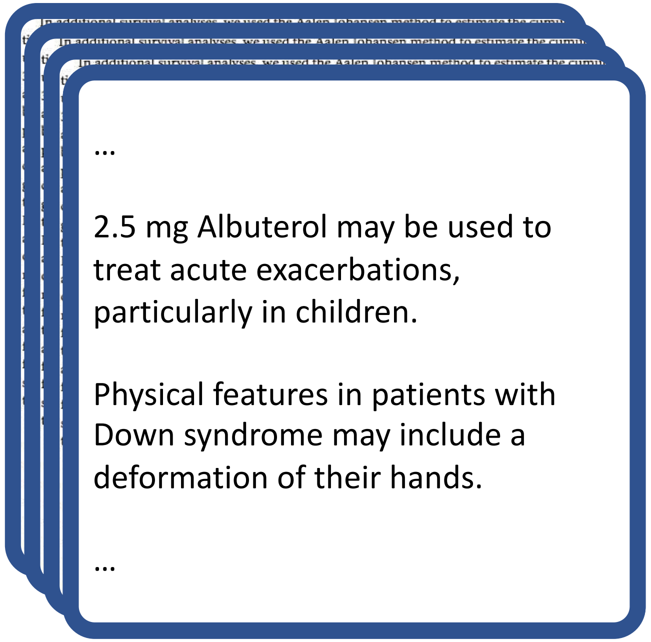
Semantic ambiguity: finding the right interpretation of human text given context

men is a disease in which one or more of the endocrine glands are overactive or forms a tumor.



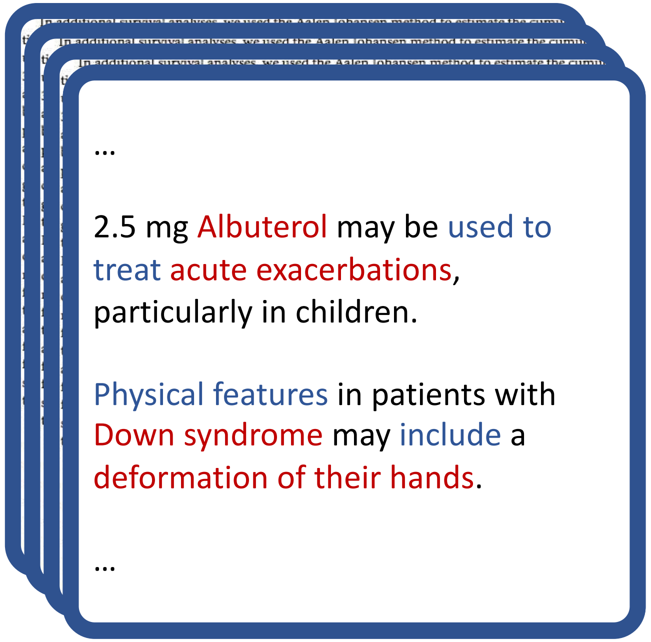
How much do I need to know?

- Unsupervised (Open Information Extraction): just relies on a large input corpus without any annotations
- Supervised: relies on a large input corpus with full annotation
- Distantly supervised: the middle ground – large input corpus with few annotations derived from external source



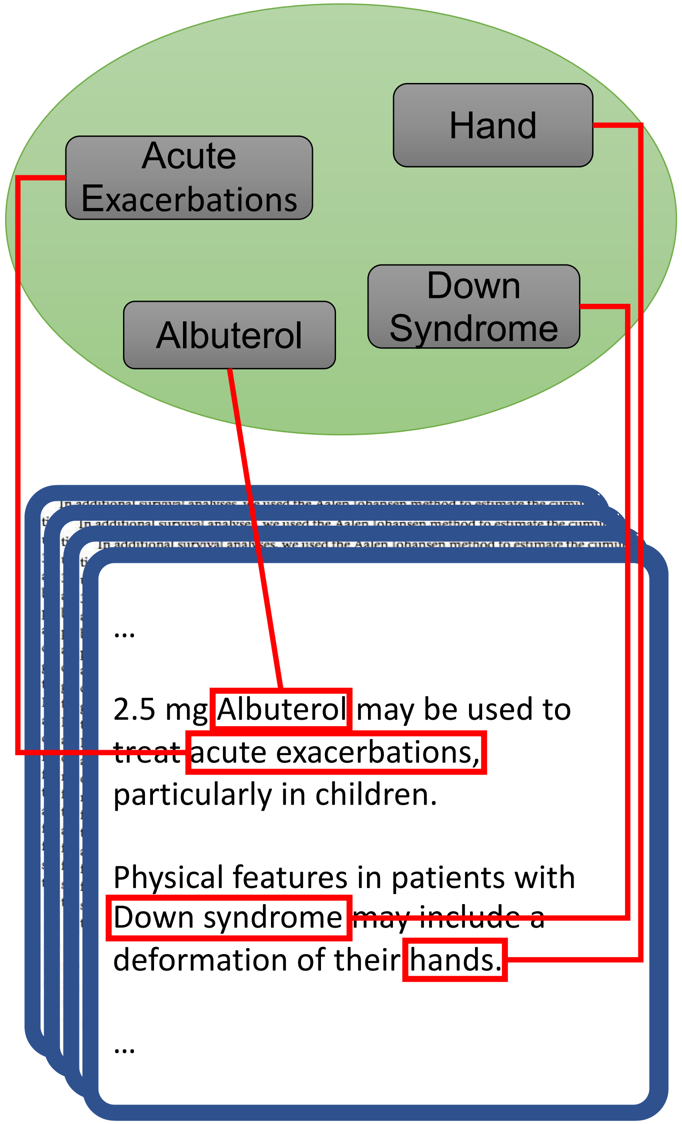
How much do I need to know?

- Unsupervised (Open Information Extraction): just relies on a large input corpus without any annotations
- Supervised: relies on a large input corpus with full annotation
- Distantly supervised: the middle ground – large input corpus with few annotations derived from external source



How much do I need to know?

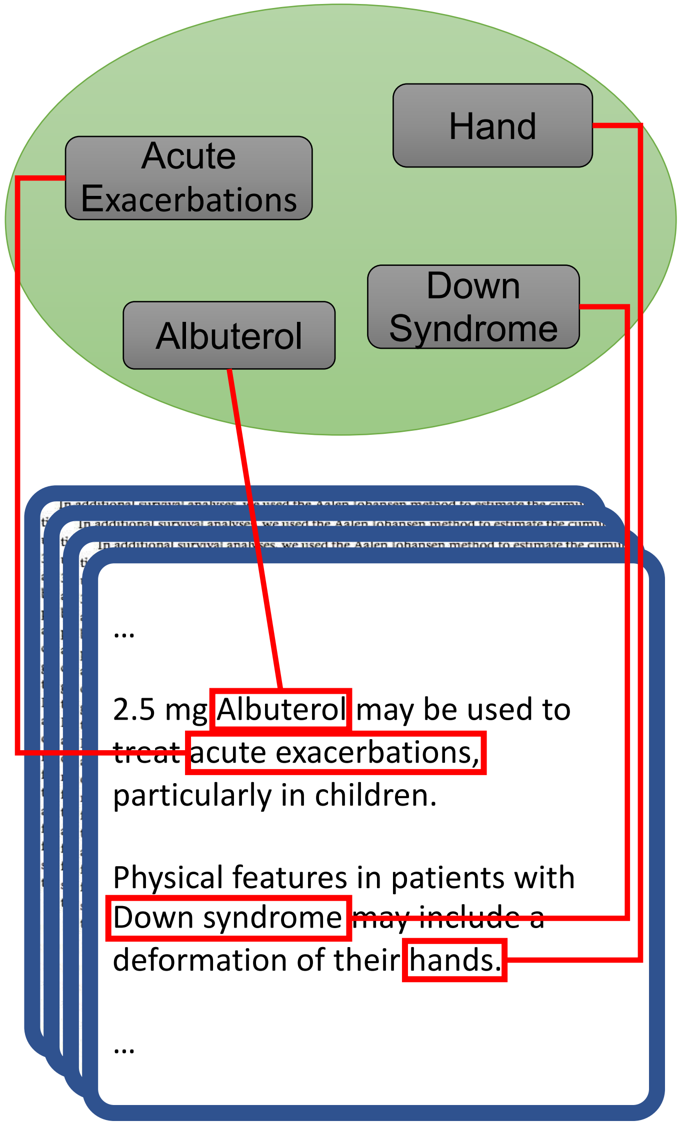
- Unsupervised (Open Information Extraction): just relies on a large input corpus <u>without any annotations</u>
- Supervised: relies on a large input corpus with <u>full annotation</u>
- Distantly supervised: <u>the middle ground </u>– large input corpus with few annotations derived from external source



Entity Extraction

- An *entity* is a collection of all possible mentions that refer to the identical real-world object or abstract concept.

- Named Entity Recognition and Disambiguation (NERD):
  - Detection of mentions of entities (Recognition)
  - Resolving the ambiguity of these mentions to canonical entities (Disambiguation)



Relations and Facts Harvesting

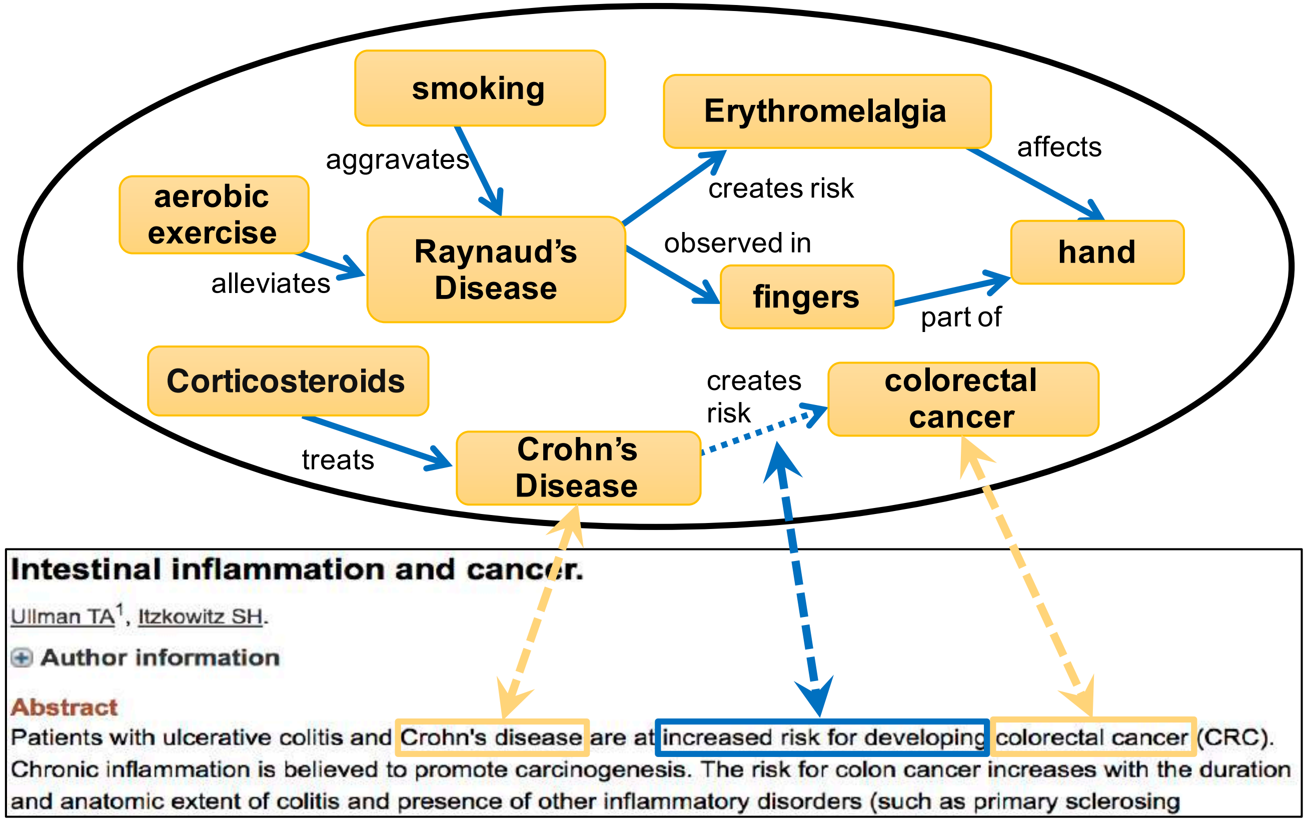
A *fact* is an instance of an n-ary relation:

$$R(a_1, \ldots, a_n)$$

where *R* is an n-ary relation and $a_1, \ldots, a_n$ are constants (e.g. entities)

Fact harvesting:

- aims to identify new relation mentions to harvest new facts.
- A relation mention is a piece of text expressing a relation between a tuple of entities



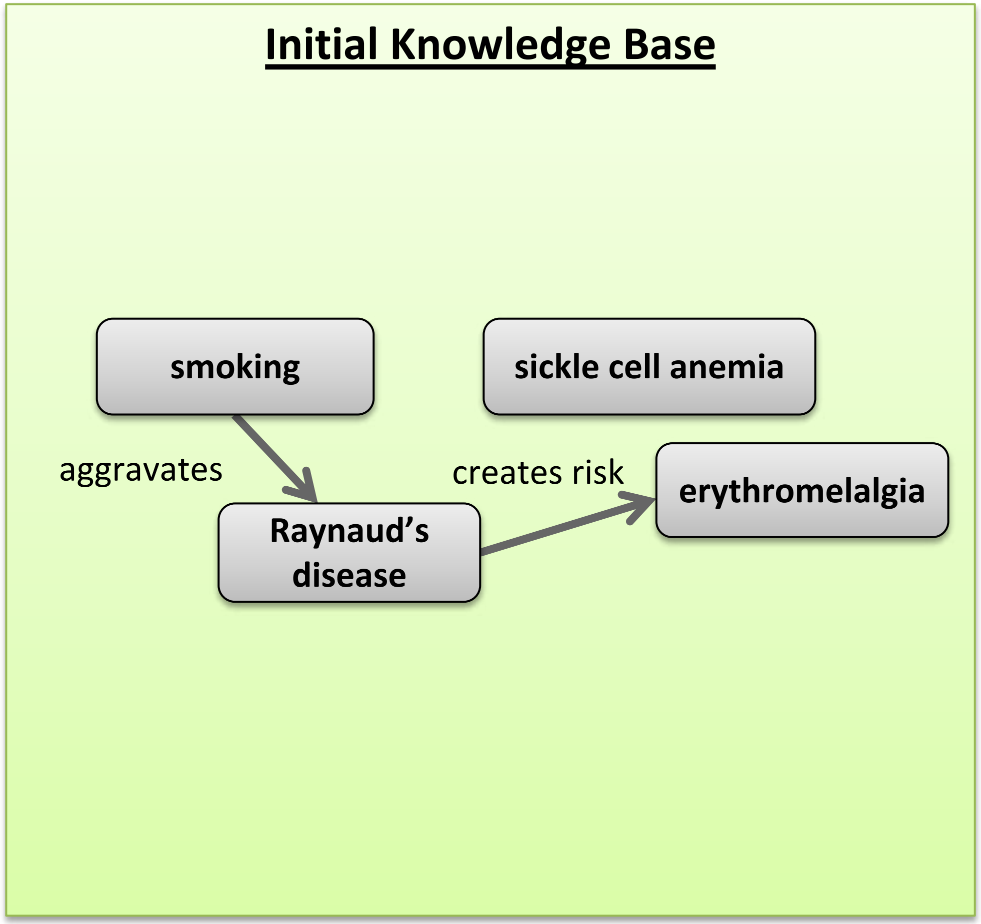
Knowledge Base Construction

Knowledge Base Construction (KBC) is the process of populating a Knowledge Base with entities, facts or rules harvested from large amounts of input data.

Social Sources

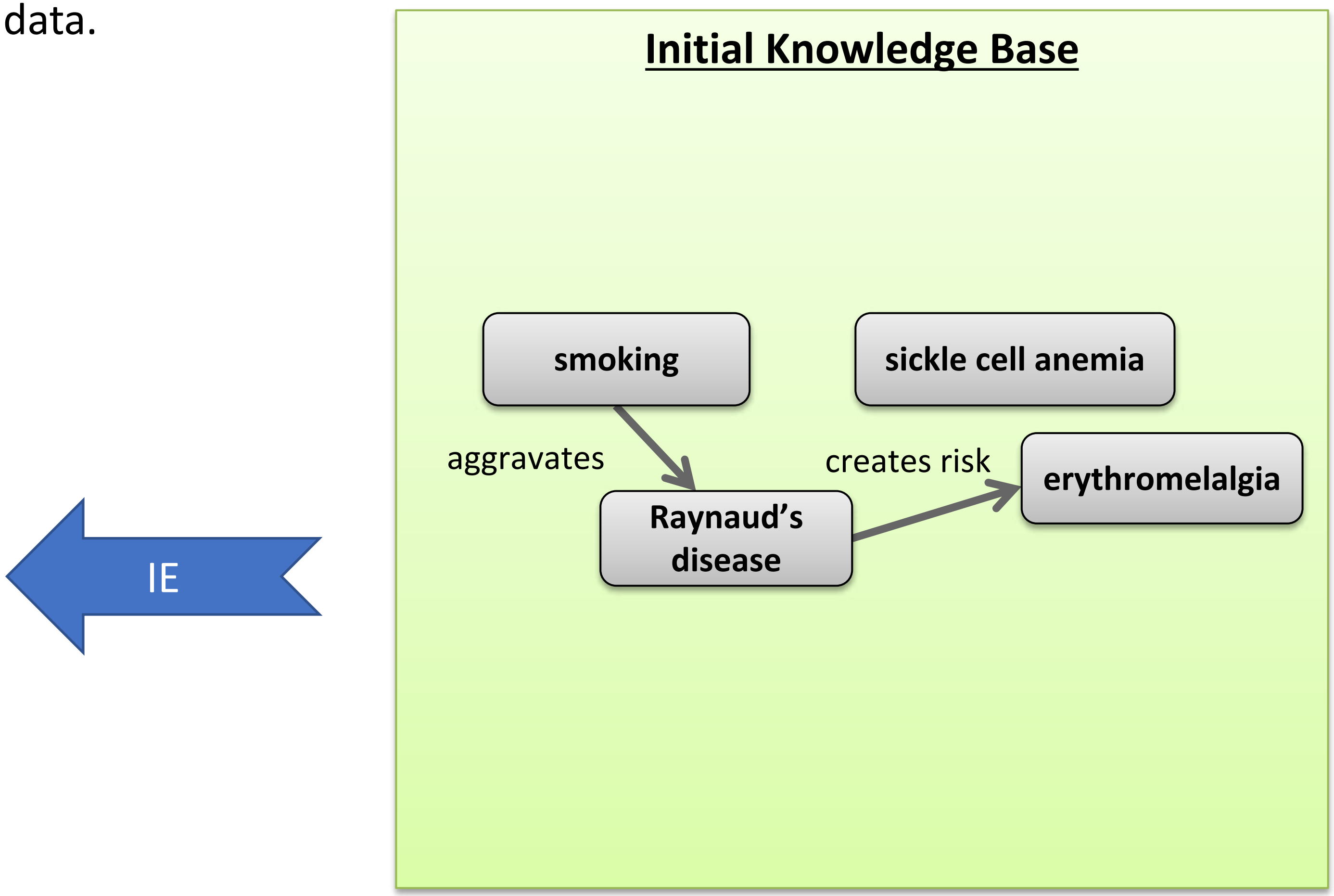
Scientific Literature

Encyclopedias



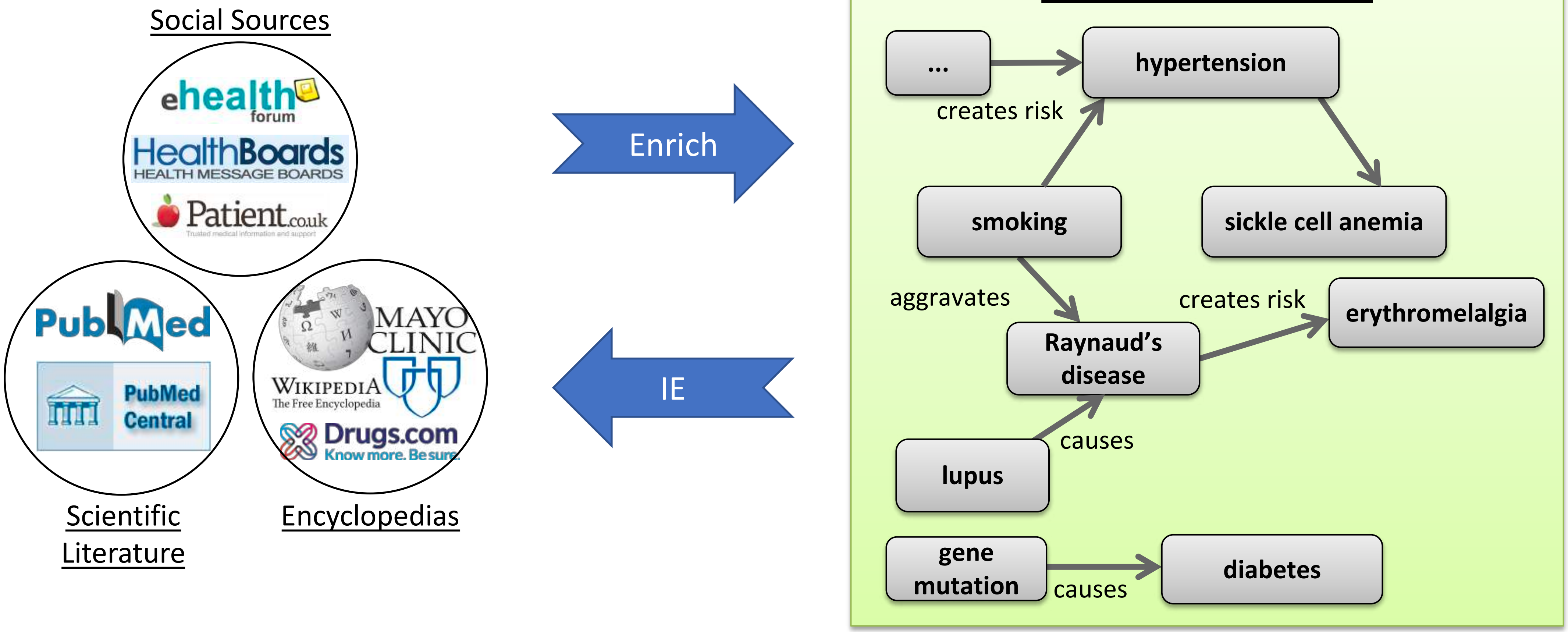
Knowledge Base Construction

Knowledge Base Construction (KBC) is the process of populating a Knowledge Base with entities, facts or rules harvested from large amounts of input data.



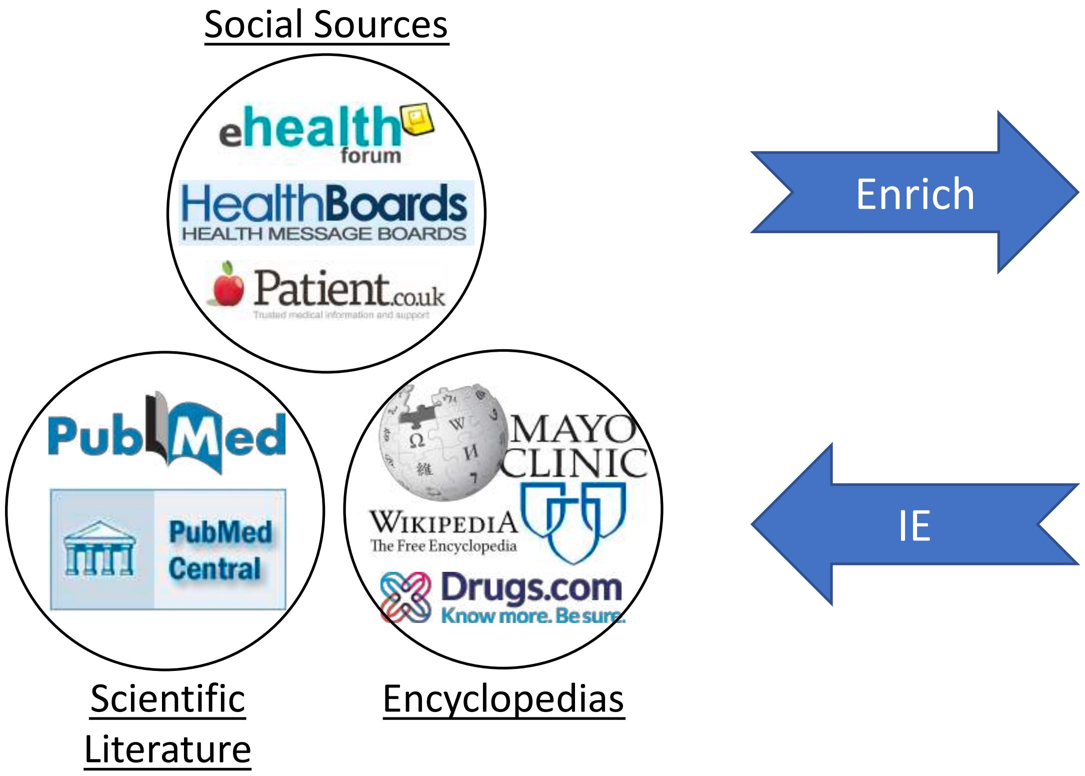
Knowledge Base Construction

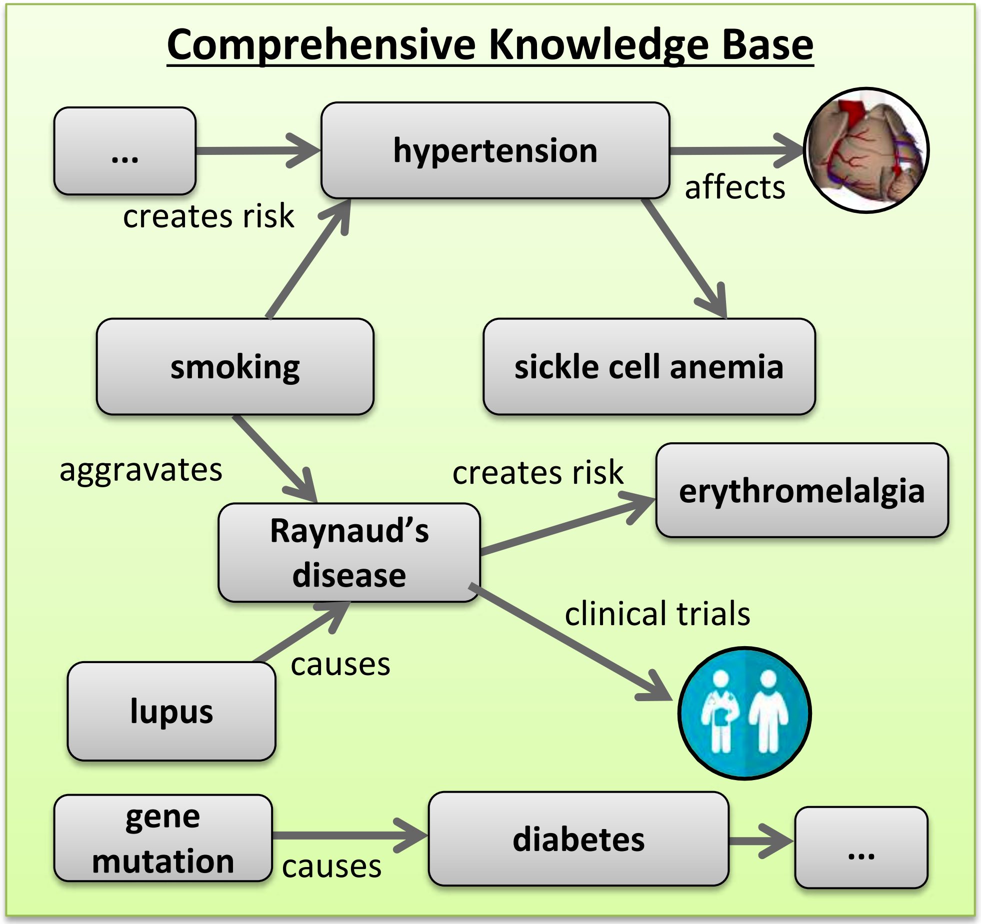
Knowledge Base Construction (KBC) is the process of populating a Knowledge Base with entities, facts or rules harvested from large amounts of input data.



Knowledge Base Construction

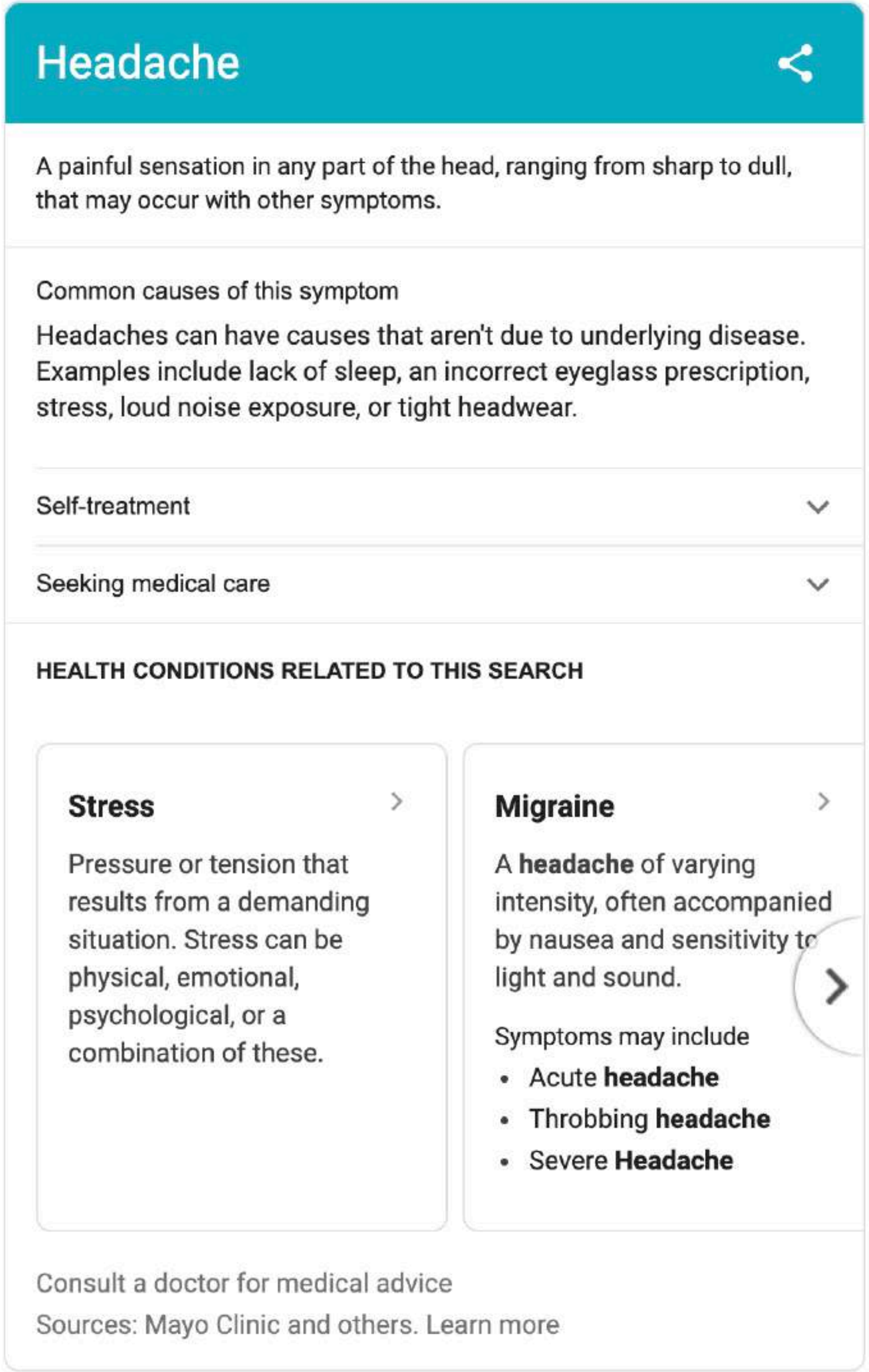
Knowledge Base Construction (KBC) is the process of populating a Knowledge Base with entities, facts or rules harvested from large amounts of input data.





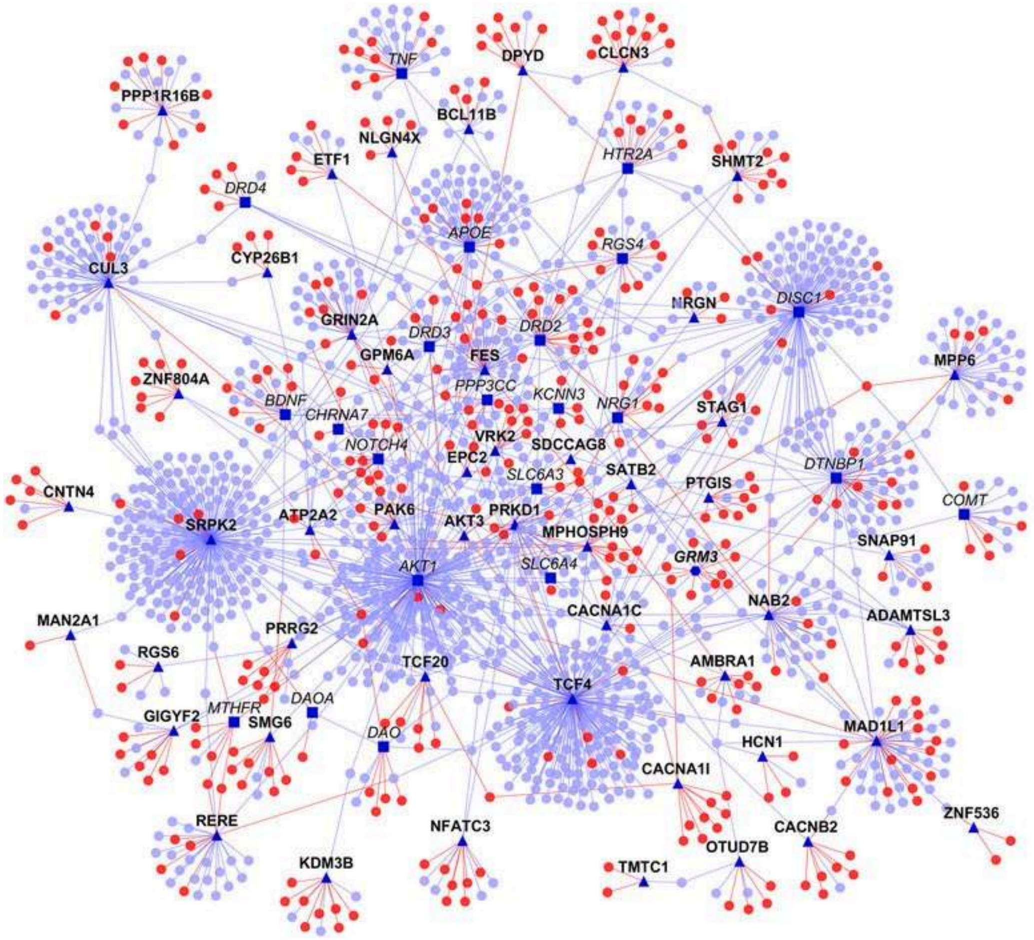
A biomedical perspective

- Google Health Knowledge Graph
- Protein Interaction (PPI) Databases
- Unified Medical Language System
- …



A biomedical perspective

- Google Health Knowledge Graph
- Protein Interaction (PPI) Databases
- Unified Medical Language System
- …



A biomedical perspective

- Google Health Knowledge Graph
- Protein Interaction (PPI) Databases
- Unified Medical Language System
- …

UMLS

Conversational AI

Alexa – What is Conversational AI?

„is the study of techniques for software agents that can engage in natural conversational interactions with humans“

```
  Welcome to
                    EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Khatri, Chandra, et al. "Alexa Prize—State of the Art in Conversational AI." *AI Magazine* 39.3 (2018): 40-55.

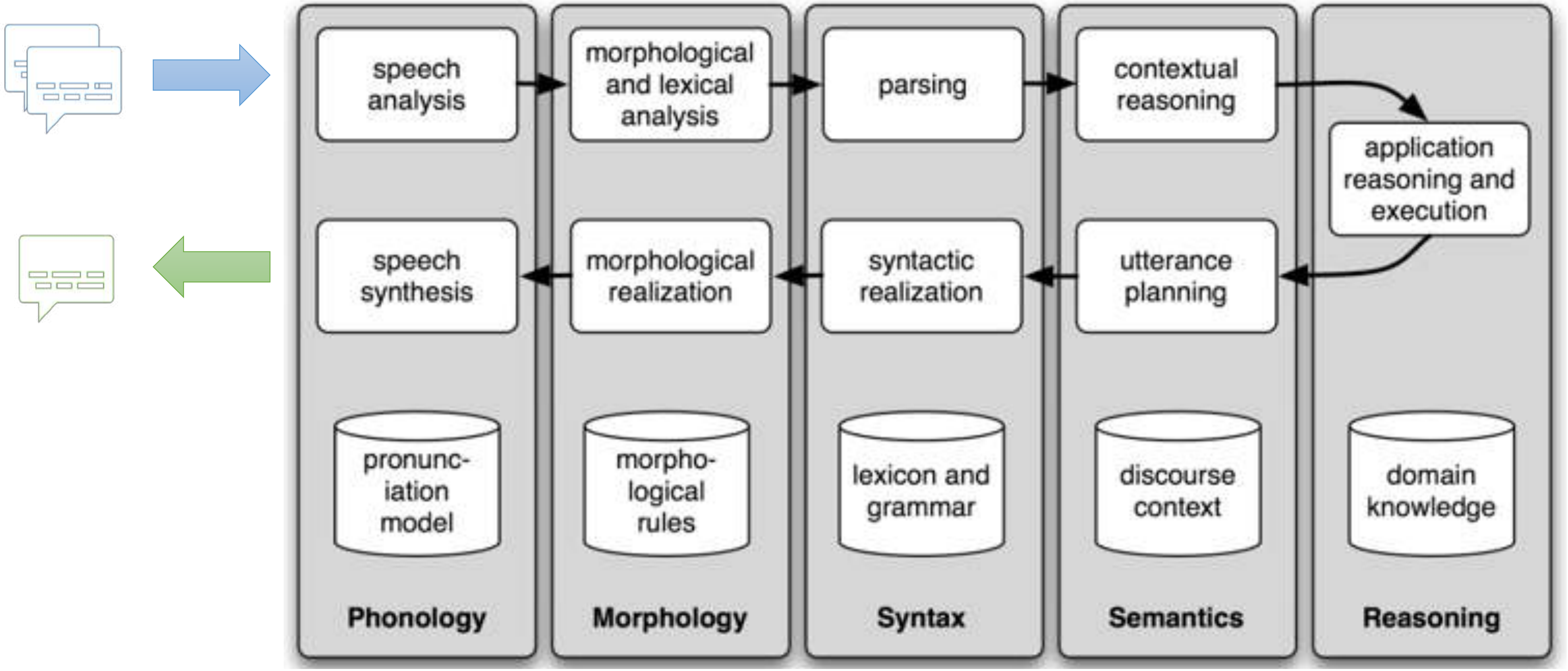
Alexa – What is Conversational AI?

- **Question Answering:** providing concise, direct answers to user queries: general (weather, sport results) and domain-specific symptoms of disease, business acquisitions

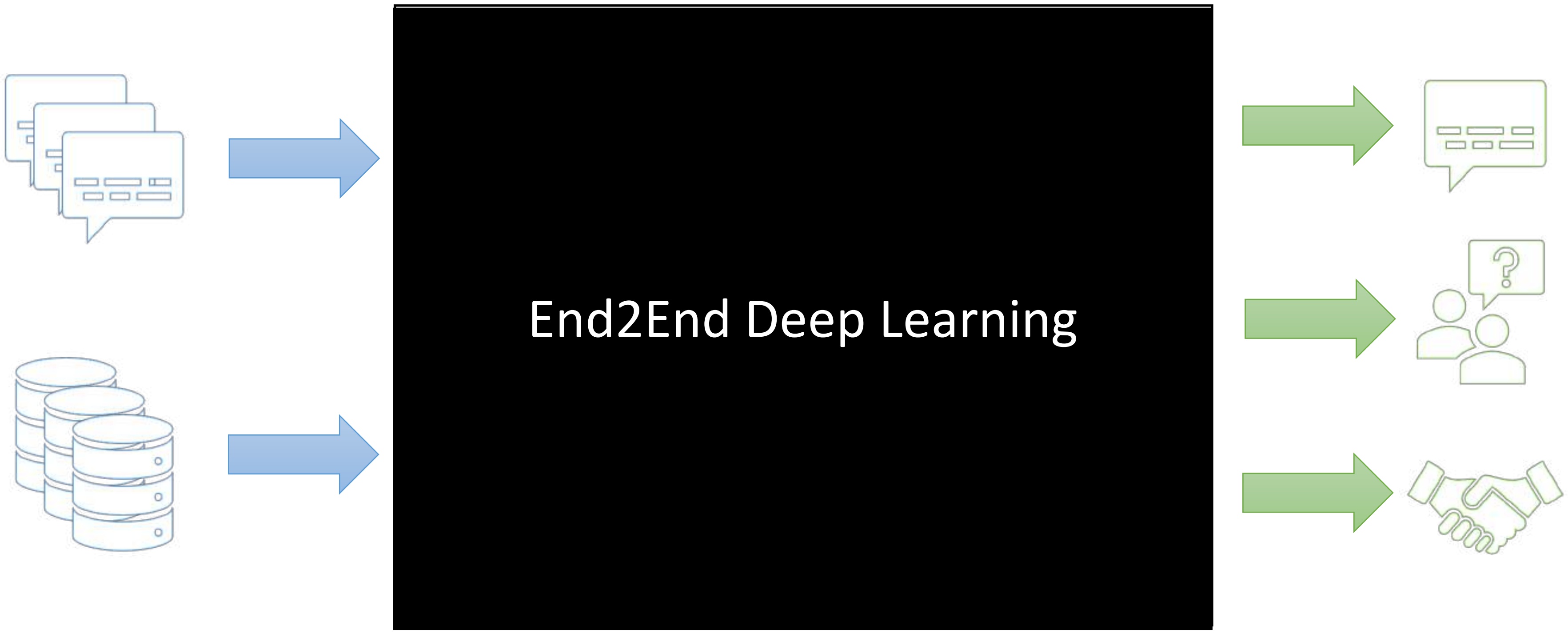
- **Task completion**: accomplishing of user actions: reservations, meeting scheduling, handling of order returns

- **Social chat**: conversing seamlessly and appropriately with users

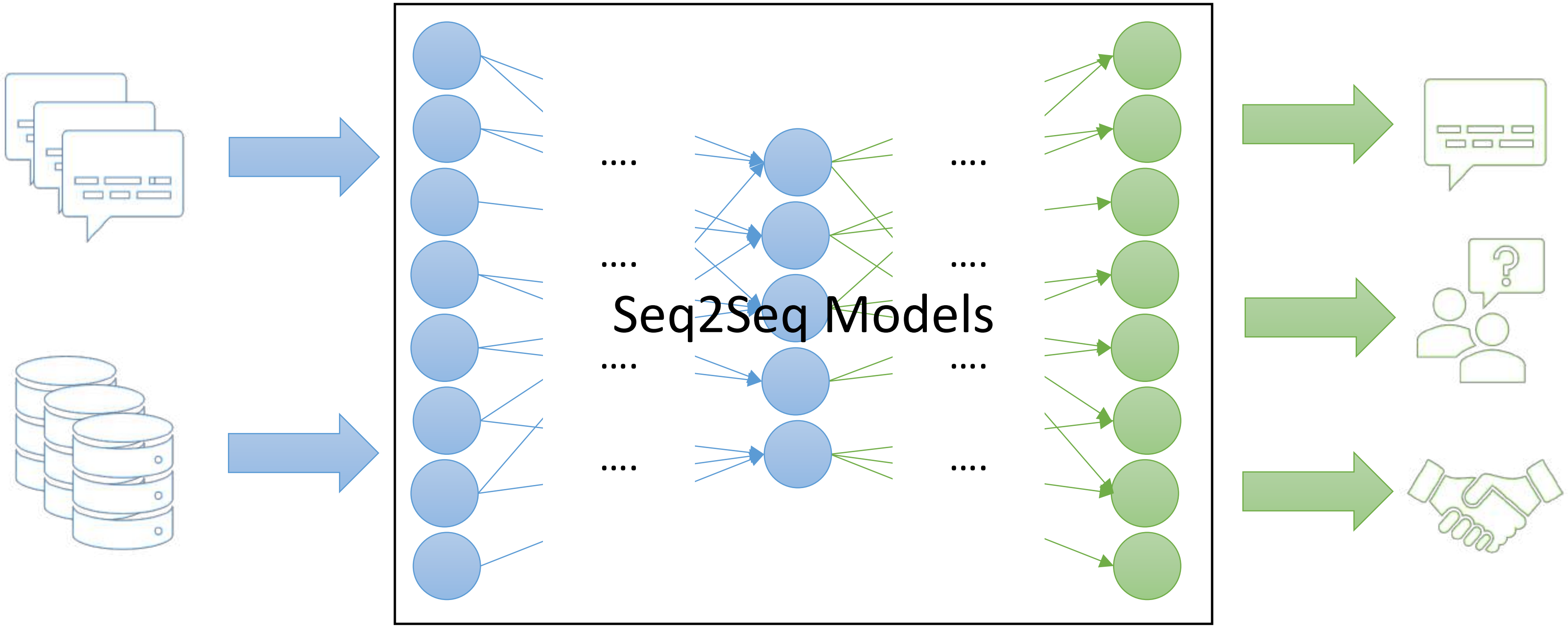
Ok Google – How do we build conversational AIs?



Neural Approaches to Conversational AI



Neural Approaches to Conversational AI



Siri – Tell me the Open Challenges

- Specificity: generate uninformative responses such as "I don't know" or "Alright"

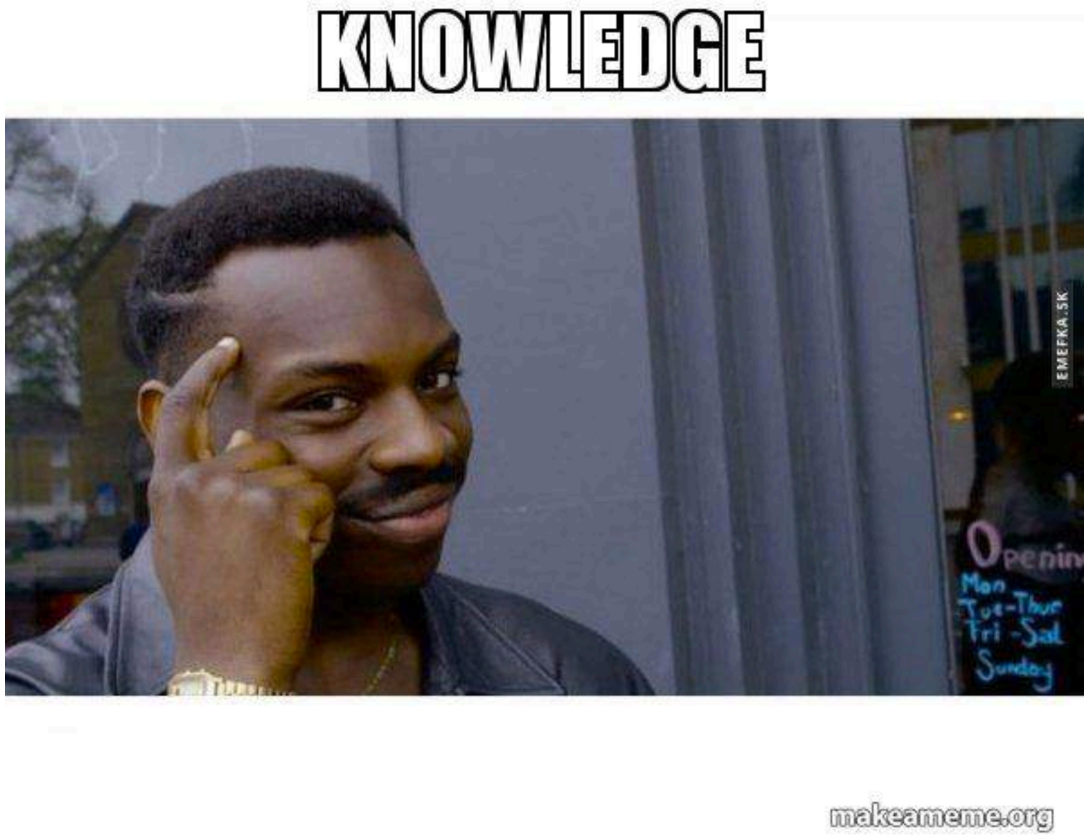
Siri – Tell me the Open Challenges

- Specificity: generate uninformative responses such as “I don’t know” or “Alright”

- Consistency: trained from chats with multiple personas

Siri – Tell me the Open Challenges

- Specificity: generate uninformative responses such as “I don’t know” or “Alright”
- Consistency: trained from chats with multiple personas
- Knowledge Access



Siri – A Medical Outlook

- **Question Answering:** sideeffect of drugs, allergies, symptom check
- **Task completion**: telemedicine to cover general checkups
- **Social chat**: social skill training, behaviour analysis

Questions?