CAN AI ANSWER DOCTORS QUESTIONS

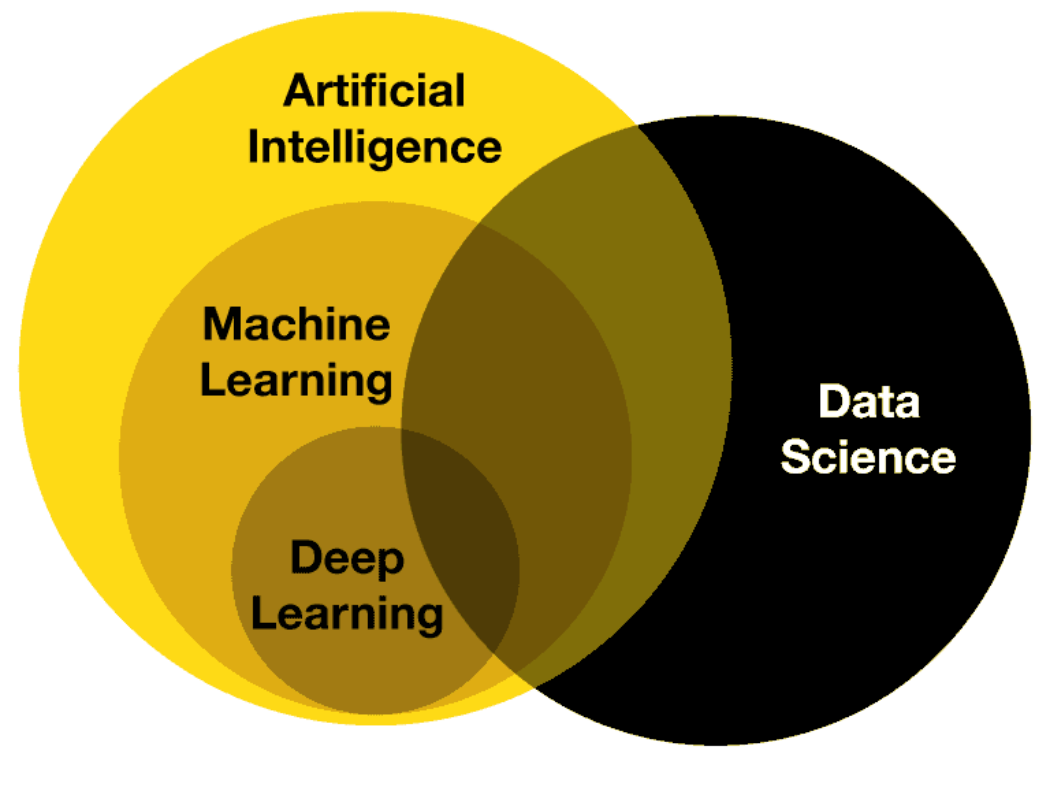
MEDICAL AI NETWORK JOURNAL CLUB #18

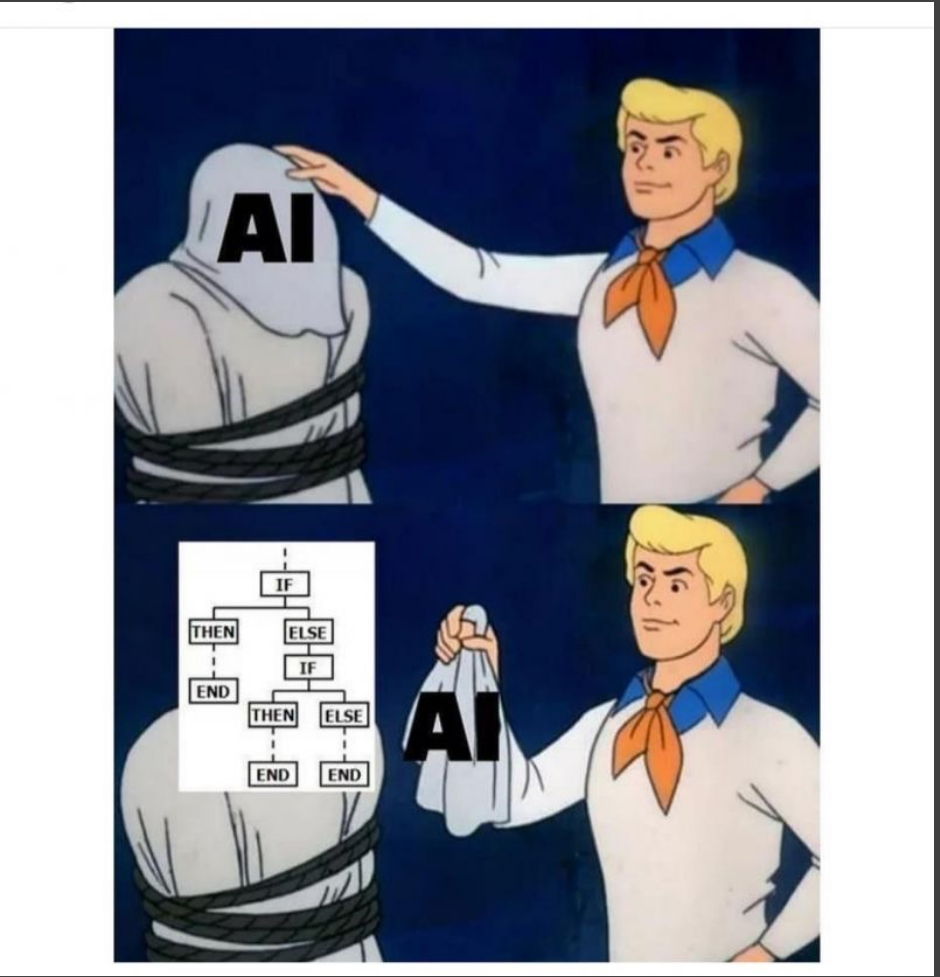
A generic discussion on research paper - Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset – and a high level survey of other emerging research in this area

About Me


ELEMENTARY

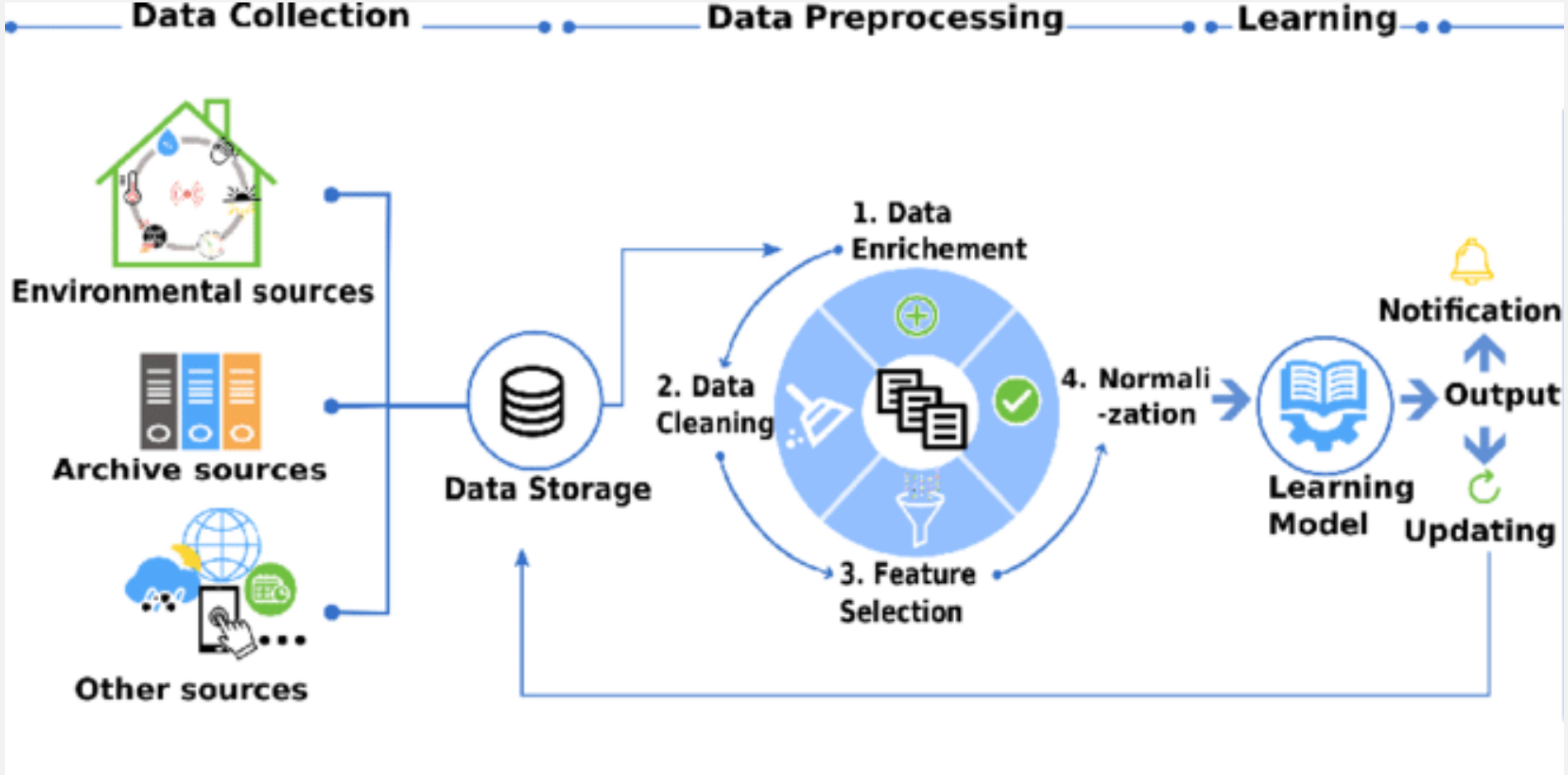
BREAKING DOWN AI SOLUTIONS



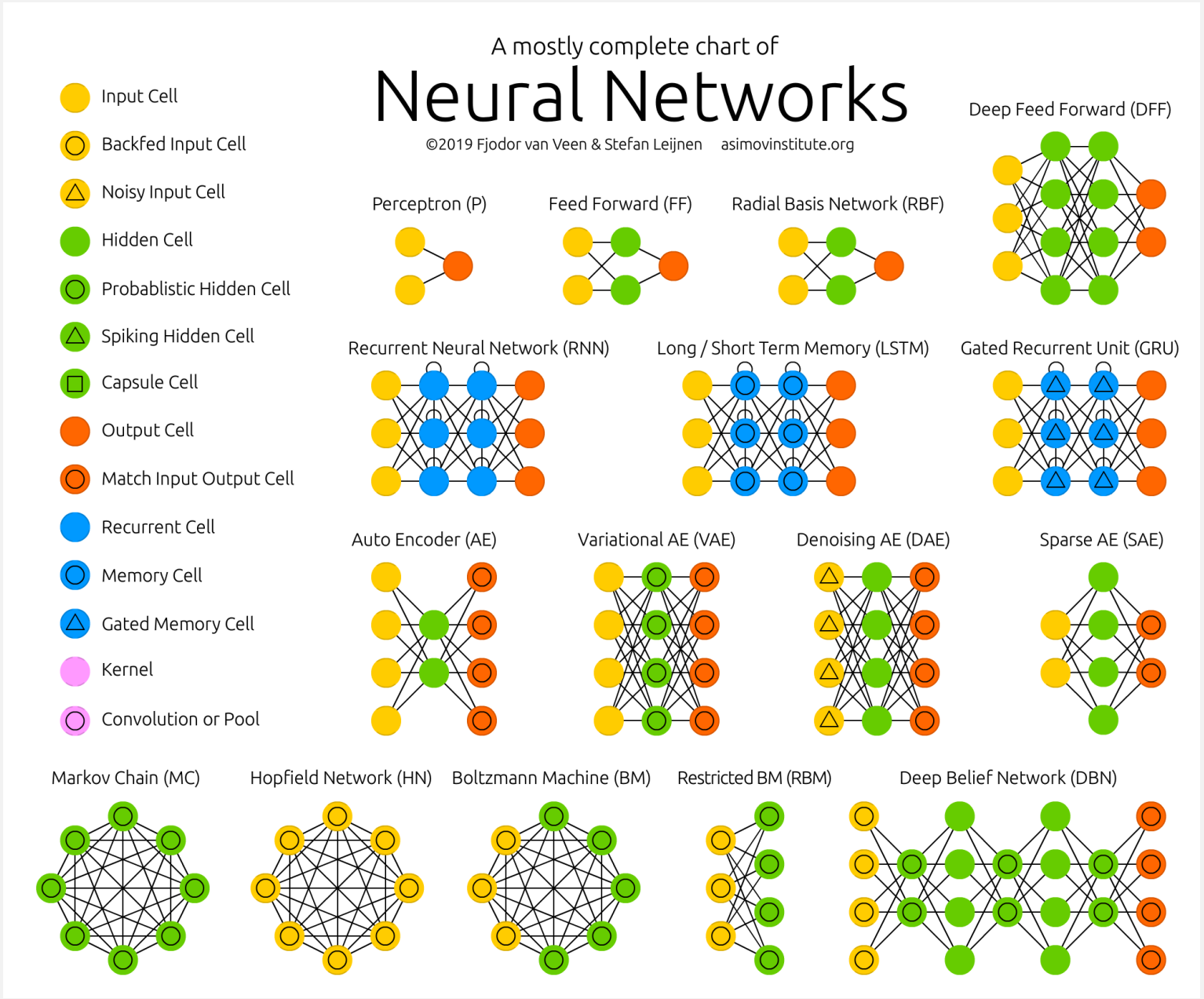


Kulin, Merima & Kazaz, Tarik & Moerman, Ingrid & De Poorter, Eli. (2020). A survey on Machine Learning-based Performance Improvement of Wireless Networks: PHY, MAC and network layer

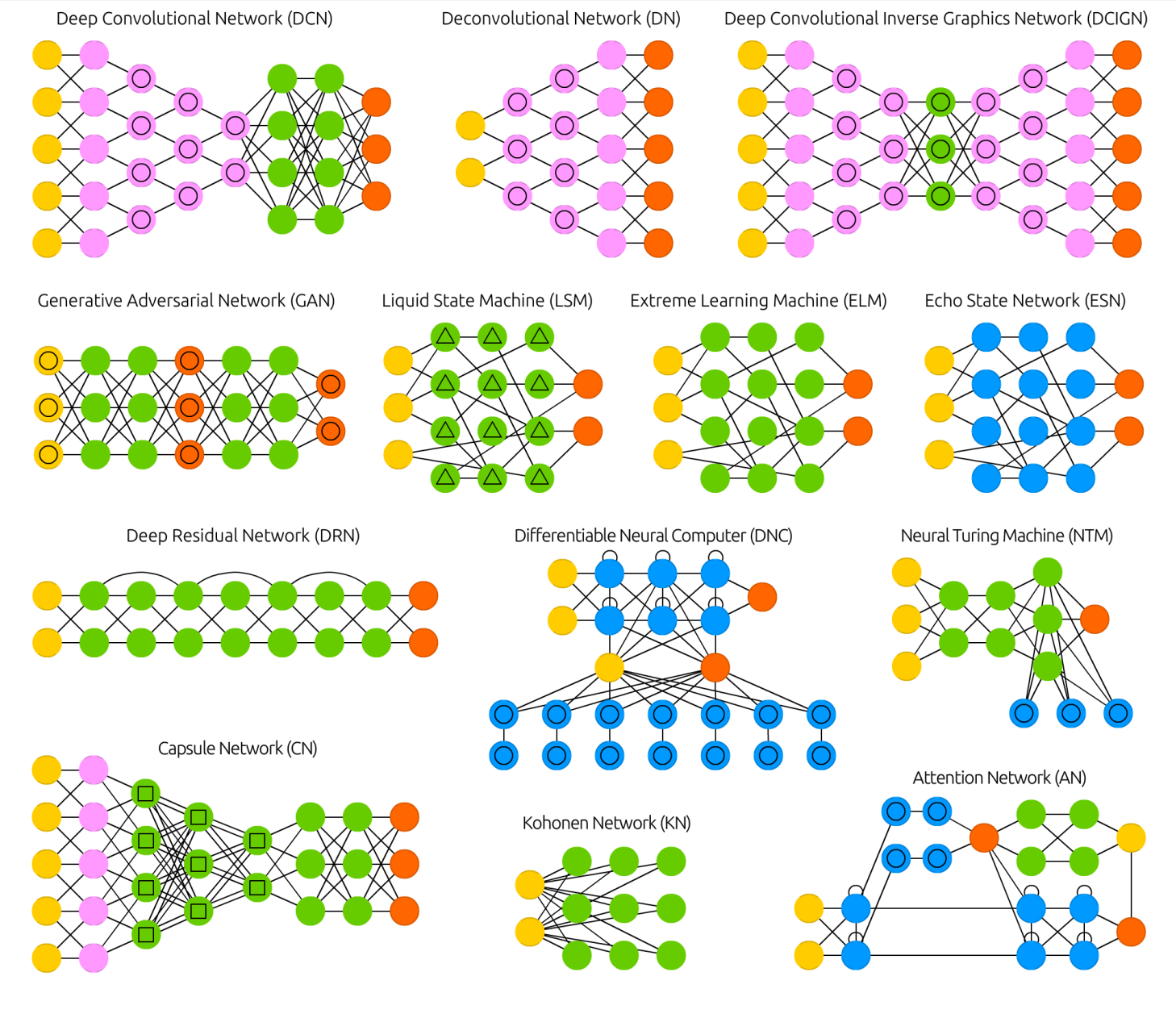
Breaking down Generic AI Solutions



Djenouri, Djamel & Laidi, Roufaida & Djenouri, Youcef & Balasingham, Ilangko. (2019). Machine Learning for Smart Building Applications: Review and Taxonomy. ACM Computing Surveys. 52. 10.1145/3311950.

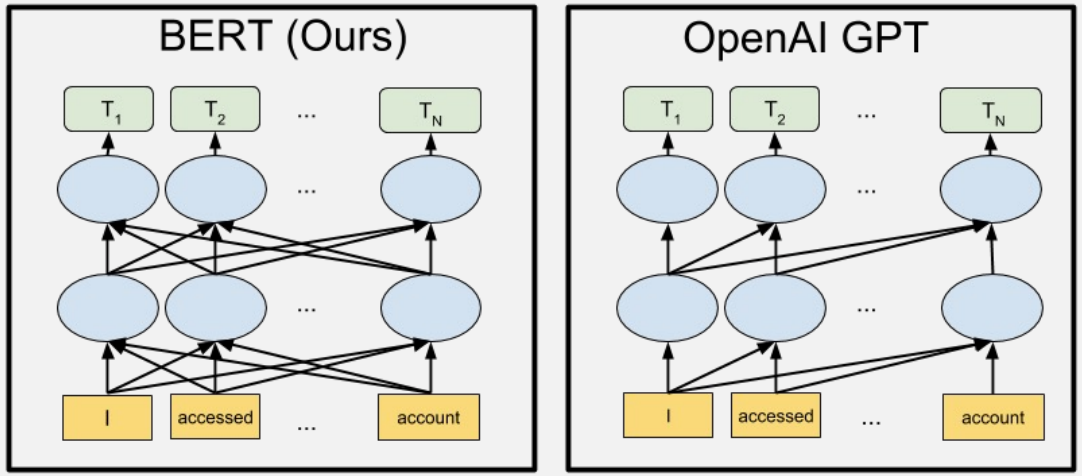


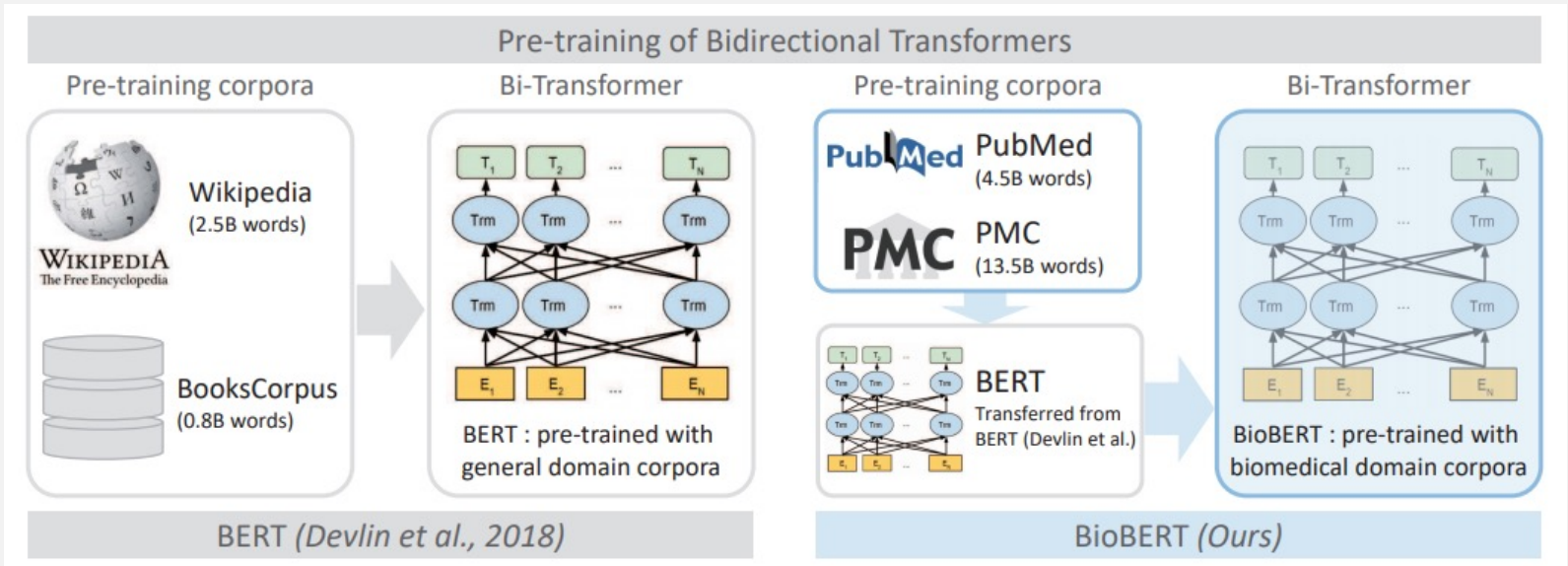
[The Neural Network Zoo - The Asimov Institute]



[The Neural Network Zoo - The Asimov Institute]

BERT





[Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing – Google AI Blog (googleblog.com)]
[BioBERT: the new NLP model giving drug development teams a boost (idalab.de)]

RESEARCH ON EMR /EHR

emrQA: A Large Corpus for Question Answering on Electronic Medical Records

Anusri Pampari* **Preethi Raghavan†♣** **Jennifer Liang†** **and Jian Peng*‡**
♣MIT-IBM Watson AI Lab, Cambridge, MA
†IBM TJ Watson Research Center, Yorktown Heights, NY
*Dept. of Computer Science, University of Illinois Urbana Champaign, IL
‡Carle Illinois College of Medicine, University of Illinois Urbana Champaign, IL
*{pampari2,jianpeng}@illinois.edu †{praghav,jjliang}@us.ibm.com

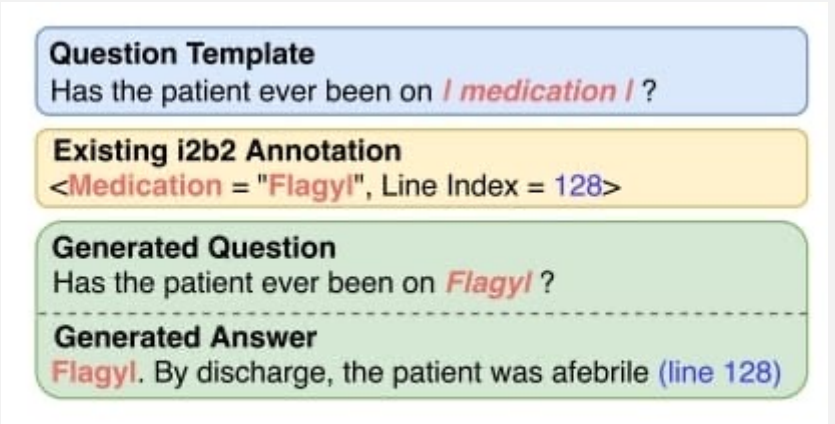
- Novel framework for to generating domain-specific large-scale question answer- ing (QA) datasets was proposed

- Resulting corpus (emrQA) has 1 million question-logical form and 400,000+ question- answer evidence pairs
- It re-purposed existing an- notations from the community shared n2c2 datasets for solving reasoning challenges without heavy reliance on expert input

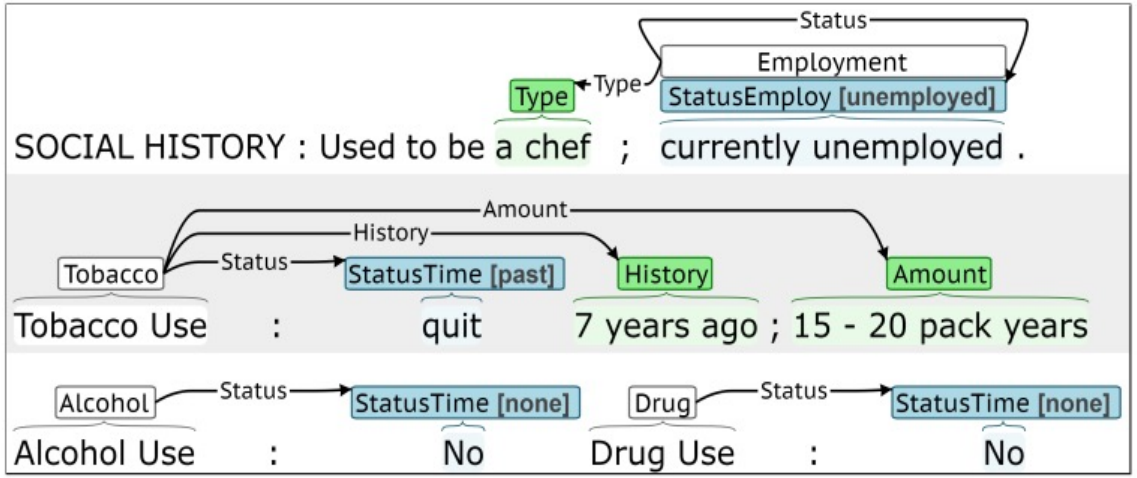
I2b2 || n2C2

- **2006 - Deidentification & Smoking**
  - Evaluating the state-of-the-art in automatic de-identification
  - Identifying patient smoking status from medical discharge records
- **2008 - Obesity**
  - Recognizing Obesity and Co-morbidities in Sparse Data
- **2009 - Medication**
  - Extracting Medication Information from Clinical Text
  - Community Annotation Experiment for Ground Truth Generation for the i2b2 Medication Challenge
- **2010 - Relations**
  - 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text
- **2011 - Coreference**
  - Evaluating the state of the art in coreference resolution for electronic medical records
- **2012 - Temporal Relations**
  - Evaluating temporal relations in clinical text: 2012 i2b2 Challenge
  - Annotating temporal information in clinical narratives
- **2014 - Deidentification & Heart Disease**
  - Creation of a new longitudinal corpus of clinical narratives
  - Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1
  - Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus
- **2018 (Track 1) - Clinical Trial Cohort Selection**
  - Cohort selection for clinical trials: n2c2 2018 shared task track 1
- **2018 (Track 2) - Adverse Drug Events and Medication Extraction**
  - 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records

- I2b2 — Integrating Biology and the Bedside research program National Center for Biomedical Computing (NCBC) based at Partners HealthCare System in Boston
- n2c2—National NLP Clinical Challenges is an outgrowth of this program under stewardship of the Harvard Medical School Department of Biomedical Informatics (DBMI)



Emr QA for Clinical Decision Support



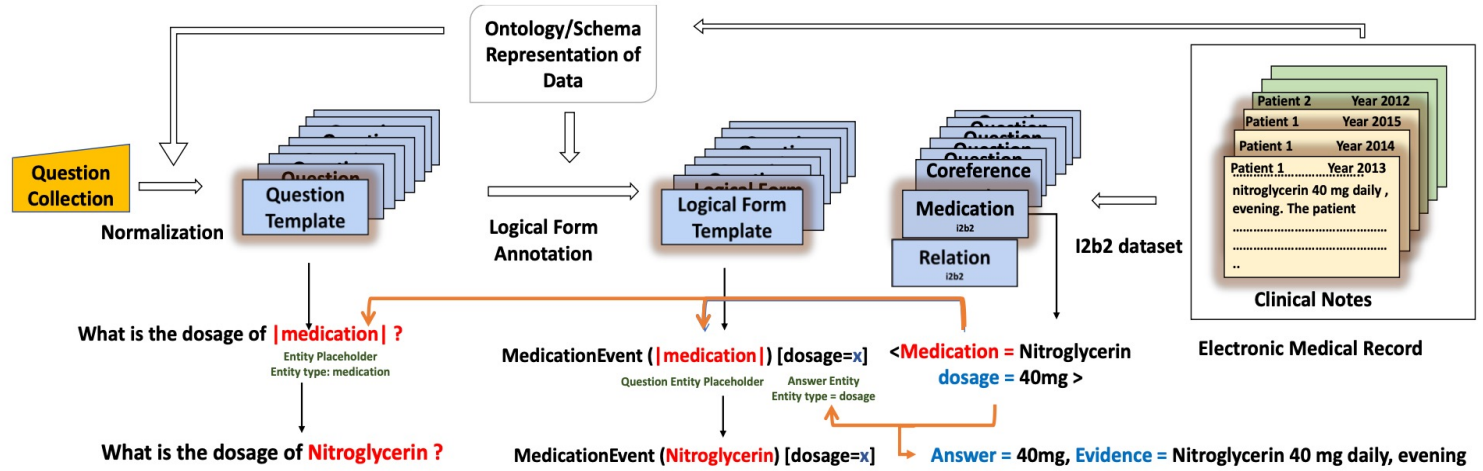
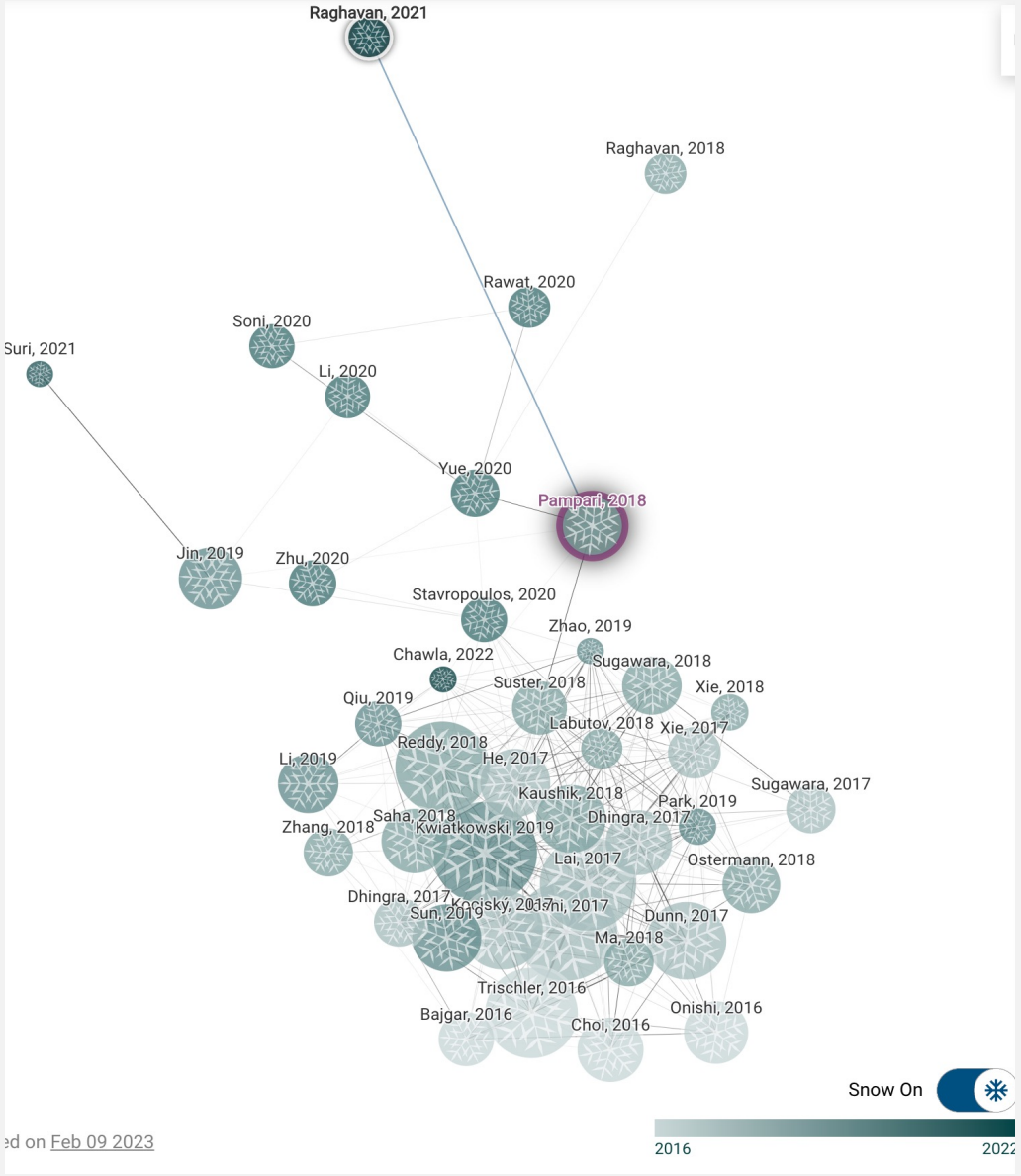


Figure 2: Our QA dataset generation framework using existing i2b2 annotations on a given patient's record to generate a question, its logical form and answer evidence. The highlights in the figure show the annotations being used for this example.



Origin paper

emrQA: A Large Corpus for Question Answering on Electronic Medical Records

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, Jian... 2018

CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension

Simon Suster, Walter Daelemans 2018

Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset

Xiang Yue, Bernal Jimenez Gutierrez, Huan Sun 2020

Quasar: Datasets for Question Answering by Search and Reading

Bhuwan Dhingra, Kathryn Mazaitis, William W. Cohen 2017

Can Machines Learn to Comprehend Scientific Literature?

Donghyeon Park, Yonghwa Choi, Daehan Kim, Minhwan Yu,... 2019

PubMedQA: A Dataset for Biomedical Research Question Answering

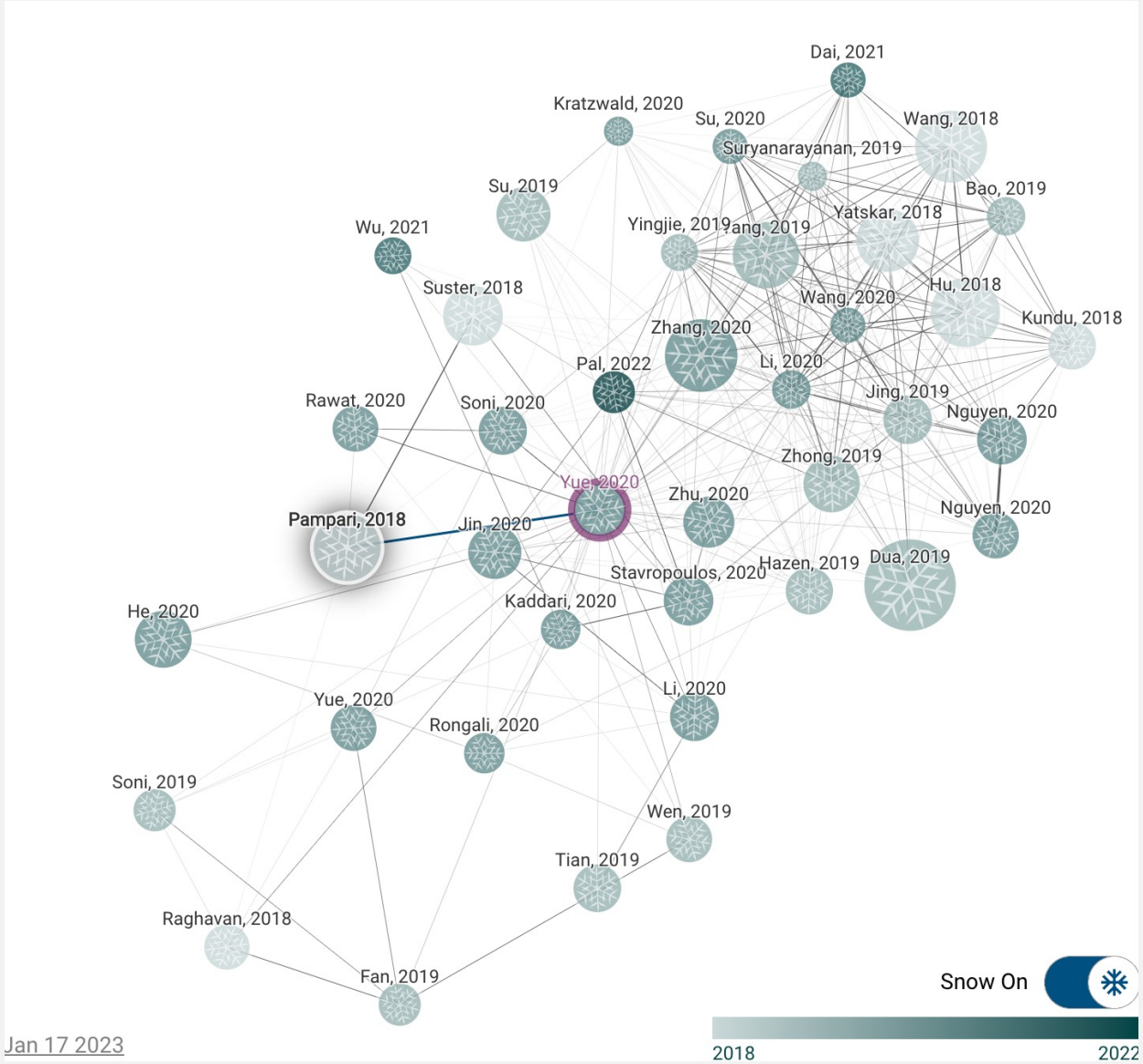
Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen... 2019

Medical Exam Question Answering with Large-scale Reading Comprehension

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, Ying Su 2018

2018
emrQA

Source – [emrQA | Connected Papers Search]



Source – [emrQA | Connected Papers Search]

Origin paper

Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset

Xiang Yue, Bernal Jimenez Gutierrez, Huan Sun 2020

emrQA: A Large Corpus for Question Answering on Electronic Medical Records

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, Jian... 2018

Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical...

Sarvesh Soni, Kirk Roberts 2020

Entity-Enriched Neural Models for Clinical Question Answering

Bhanu Pratap Singh Rawat, W. Weng, Preethi Raghavan,... 2020

Towards Medical Machine Reading Comprehension with Structural Knowledge and Plain Text

Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, Anqi... 2020

CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension

Simon Suster, Walter Daelemans 2018

Annotating Electronic Medical Records for Question Answering

Preethi Raghavan, Siddharth Patwardhan, Jennifer J. Liang,... 2018

What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medica...

Di Jin, Eileen Pan, Nassim Oufattole, W. Weng, Hanyi Fang,... 2020

A Study of the Tasks and Models in Machine Reading Comprehension

Chao Wang 2020

2020 Clinical Reading Comprehension

Indepth Analysis

Error Type	Question	emrQA Answers	Prediction	Error Ratio	
				Medication	Relation
Span mismatch - include key info	Does she have a history of known drug allergies?	ALLERGIES: He had no known drug allergies	He had no known drug allergies	**78%**	**66%**
Span mismatch - miss key info	What is the current dose of lasix?	MEDS: K-Dur 20 BID, Nexium 20, lasix 160 BID	BID	**4%**	**0%**
Ambigious questions	What is the patient's low history?	At the time of discharge, her potassium had been low despite repletion	11) Low grade,anemia	**8%**	**4%**
Incorrect golds	What is the patient's incisions status?	Wash incisions with warm water and gentle soap	Do not apply lotions, creams, ointments or powders to incision	**2%**	**2%**
False negatives	Is there a mention of fluid in the record?	There is some fluid, or mucosal thickening in the ethmoid and sphenoid sinuses	The amount of fluid layering at the apices and the pleural spaces appear slightly decreased	**2%**	**18%**
May need external knowledge	What treatment has the patient had for his CAD?	CAD s/p CABG 2003 s/p	Pt's vancomycin was stopped after 14 days of treatment	**2%**	**2%**
Others	Is the patient's right hand ganglion cyst well-controlled?	right hand ganglion cyst removed	x 3 right hand ganglion cyst	**4%**	**8%**

Table 3: Error analysis on 50 sampled questions from the *Medication* and *Relation* dev sets respectively. Example question, ground truth and prediction from either *Medication* or *Relation* are given for each type of error.

	Medication	Relation
# Question	222,957	904,592
# Context	261	423
# Question Template	80	139
Question: avg. tokens	8.00	7.91
Answers: avg. tokens	9.47	10.41
Context: avg. tokens	1062.66	889.23

Table 1: Statistics of two major subsets, *Medication* and *Relation*, of the emrQA dataset.

emrQA answers are often incomplete

emrQA questions are often answerable without using domain knowledge

SUITABILITY FOR LARGE LANGUAGE MODELS

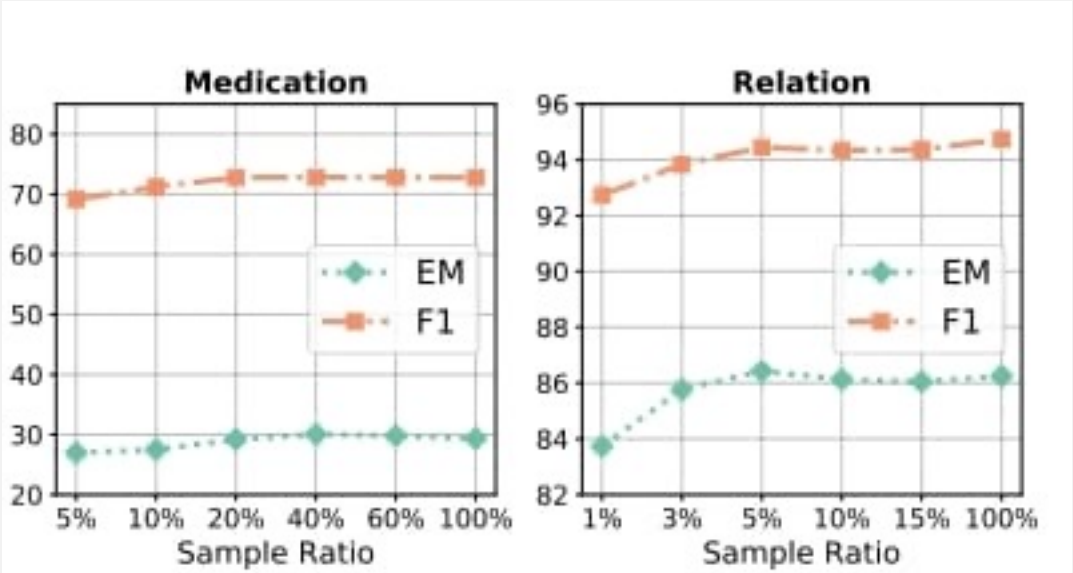
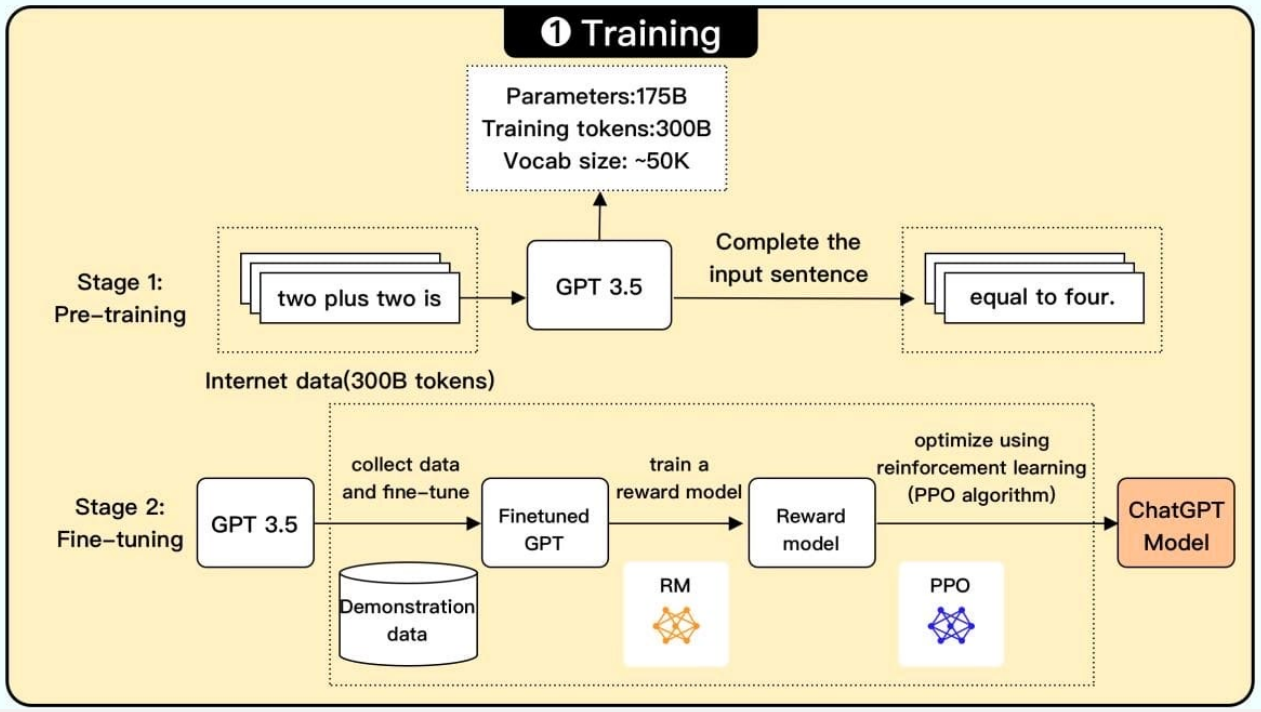


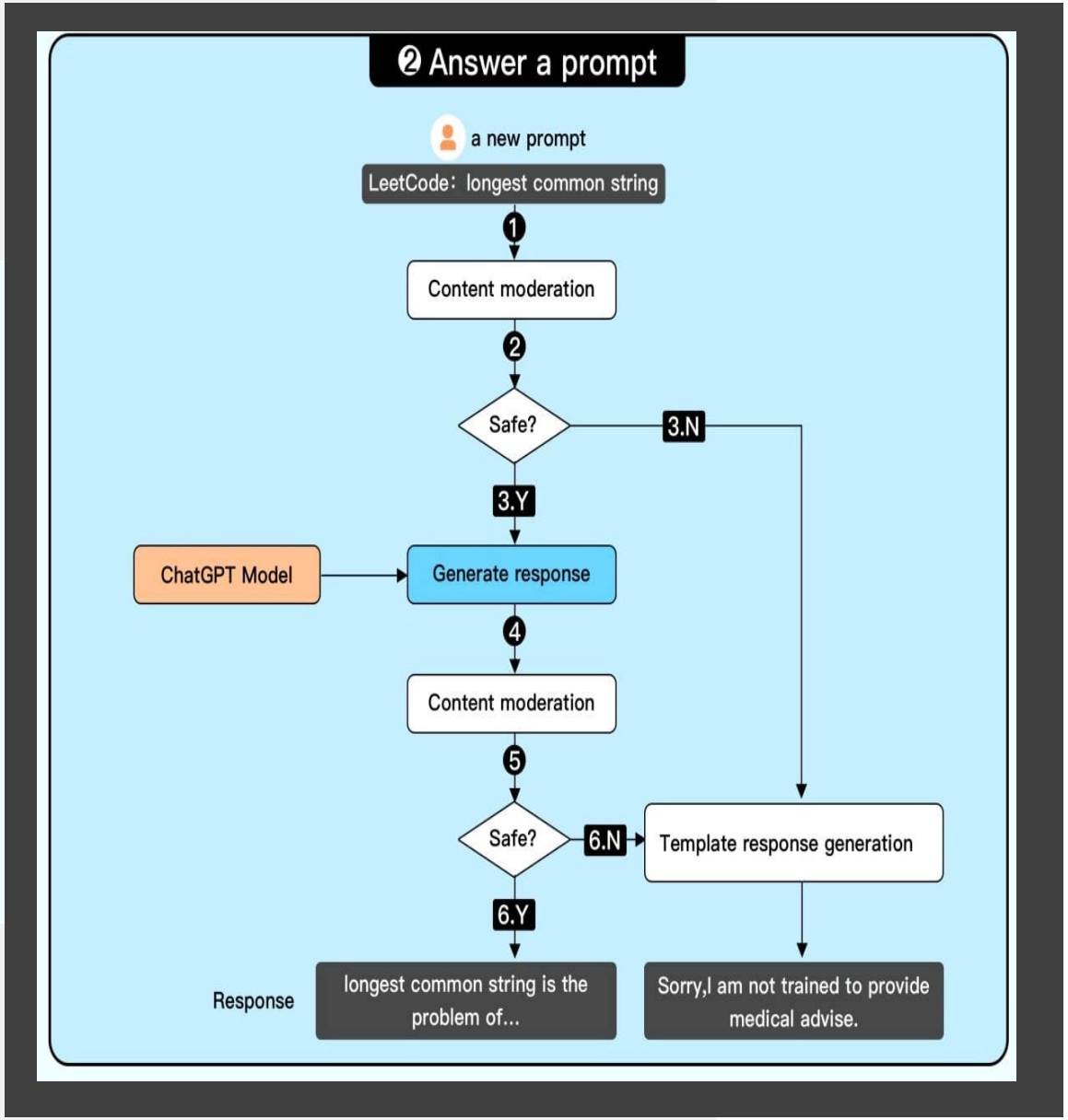
Figure 3: Impact of *training size* on the performance of DocReader (Chen et al., 2017) based on the *Medication* and *Relation* dataset.

- A small sampled subset (5%-20%), can provide roughly equal performance compared to the model trained on the entire dataset and is close to human expert's performance
- BERT models do not beat the best performing clinicalRC base model.

WHERE IS AI HEADING

How ChatGPT works?





Source:
EP 44: How does ChatGPT work? - by Alex Xu (bytebytego.com)

Third AI Winter

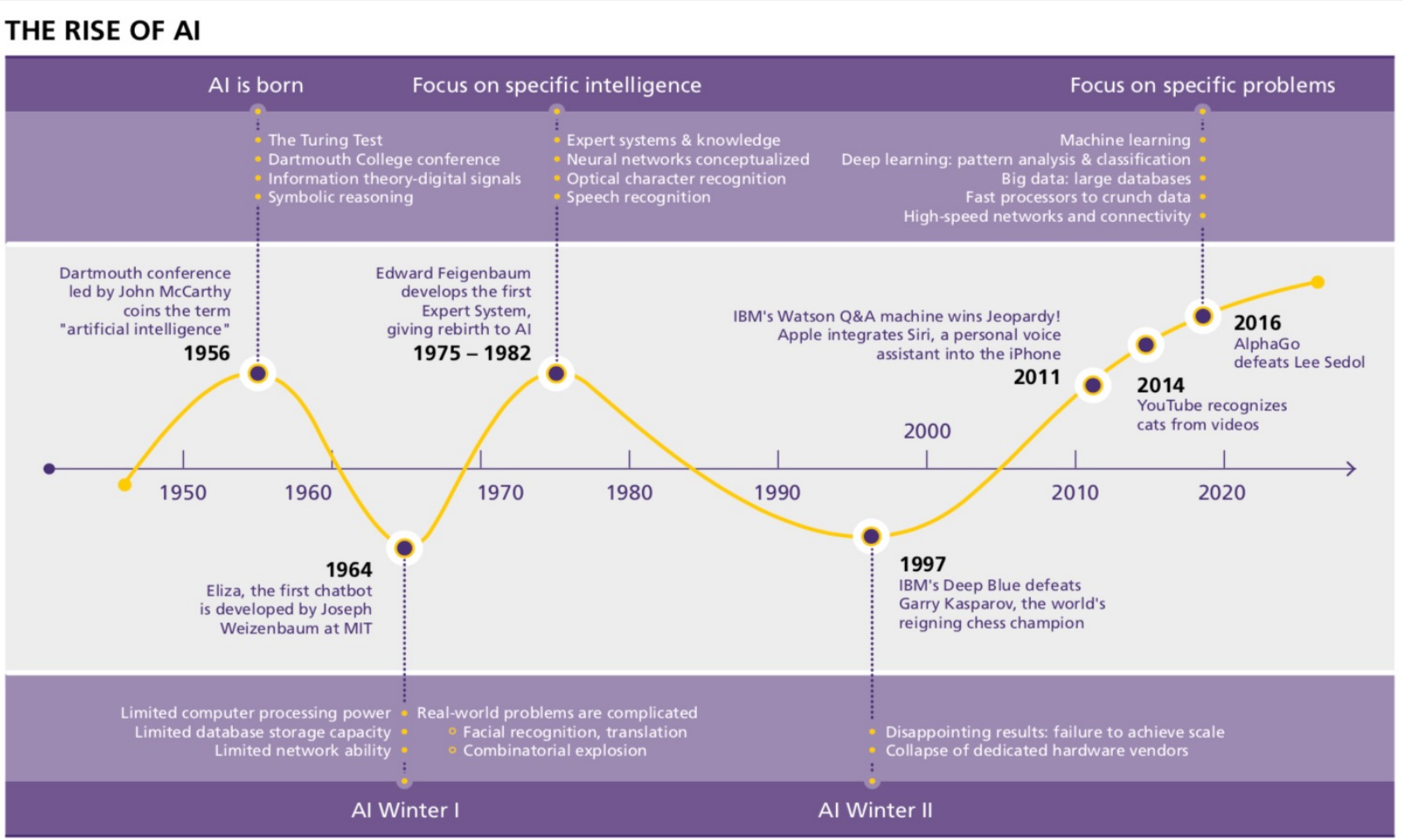
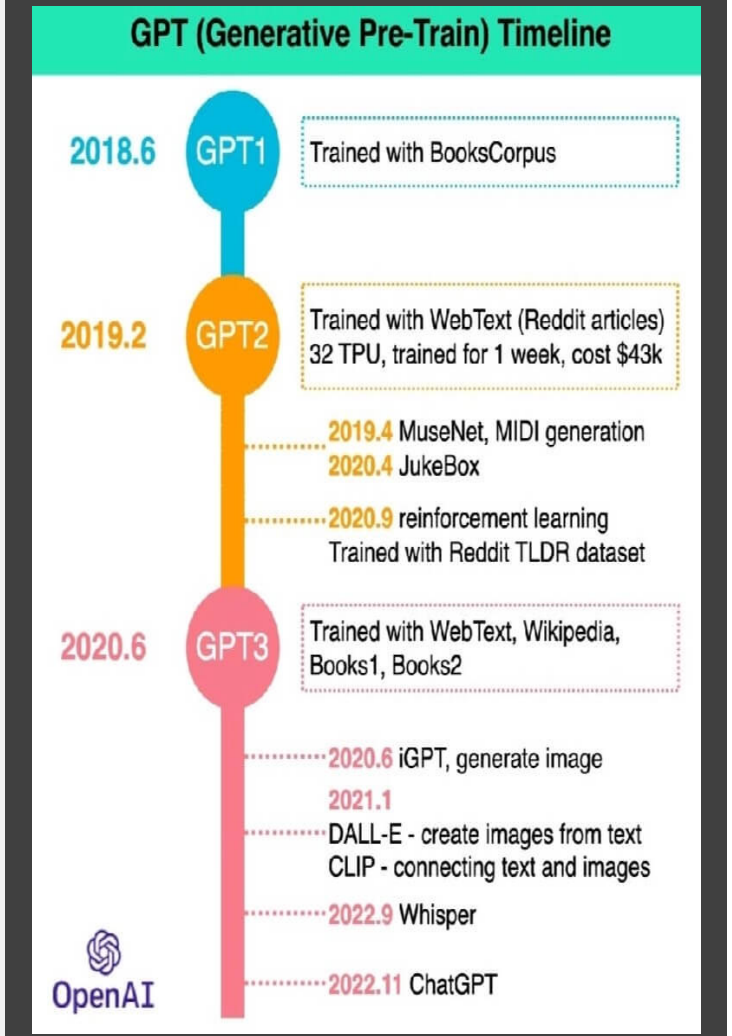


Figure 1: An AI timeline; **Source:** Lavenda, D./Marsden, P.

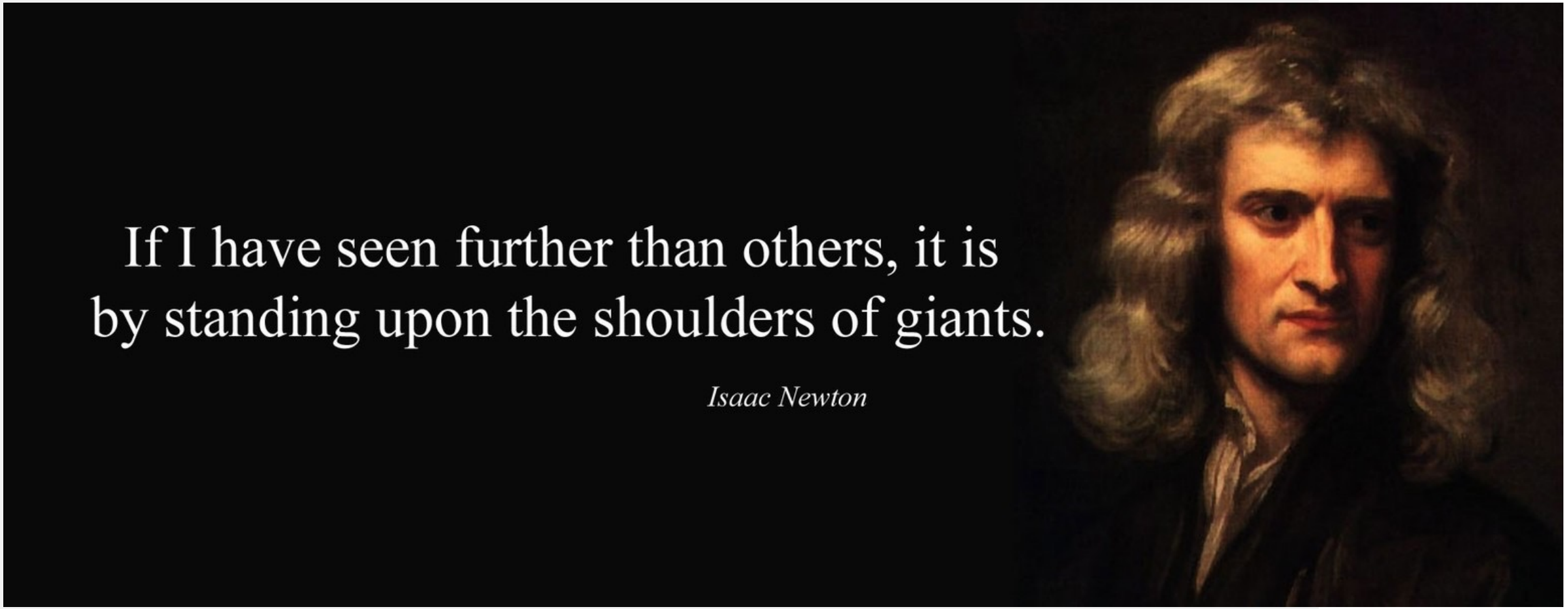
source dhl via @mikequindazzi



Source:

1. Why Did AI Rise for a Third Time After the 1st and 2nd AI Winters? - Brightwork Research & Analysis
2. 10 Questions to ChatGPT: How It Can Change Cybersecurity - SOCRadar

THANK YOU


If I have seen further than others, it is
by standing upon the shoulders of giants.
Isaac Newton