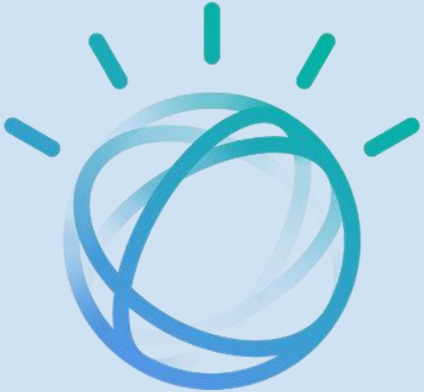
Application of IBM Watson in the Medical Field

Utkarsh Kumar
2020 CSCI Senior Seminar
Division of Science and Mathematics
University of Minnesota Morris

What is Watson?



- Question Answering (QA) computing system
- Open domain datasets
  - Wikipedia
  - Twitter
  - Online Research Datasets

Figure taken from [5]

Outline

- **The Problem**
- Data in published medical research
- How Watson works
- Case Study
- Conclusions

The Problem

- Drug Discovery [1]
  - Massive Investment
  - 80% fail to gain approval of FDA.
- Pressure on Researchers
- A lot more data
  - Limitation: Scalability

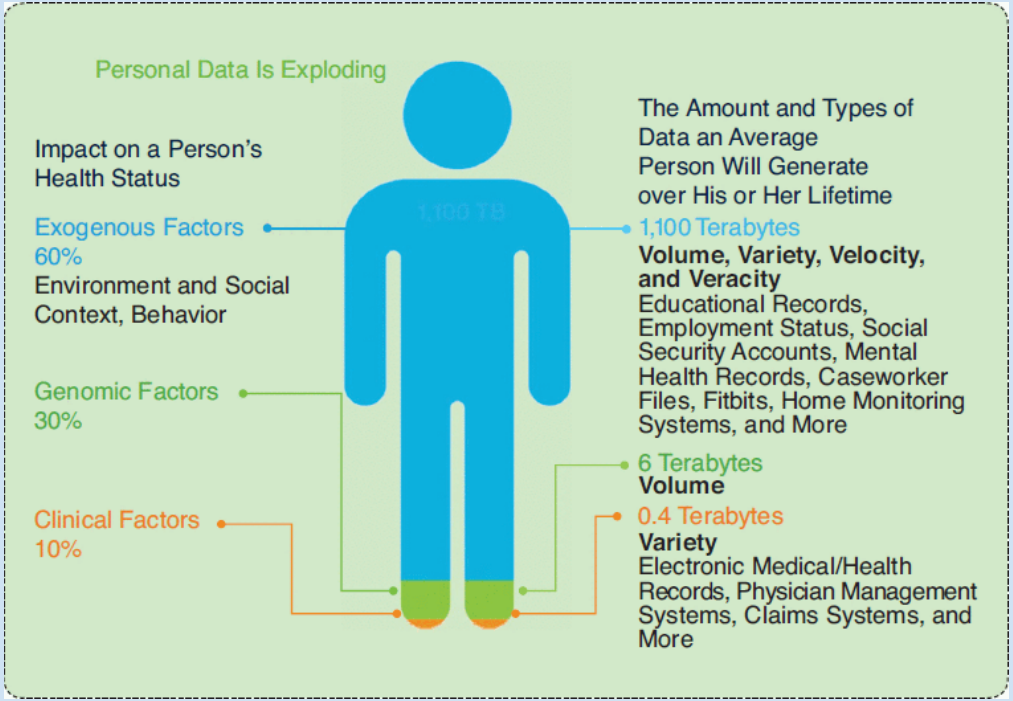

Personal Data Is Exploding
Impact on a Person's Health Status
Exogenous Factors 60%
Environment and Social Context, Behavior
Genomic Factors 30%
Clinical Factors 10%
1,100 TB
The Amount and Types of Data an Average Person Will Generate over His or Her Lifetime
1,100 Terabytes
Volume, Variety, Velocity, and Veracity
Educational Records, Employment Status, Social Security Accounts, Mental Health Records, Caseworker Files, Fitbits, Home Monitoring Systems, and More
6 Terabytes
Volume
0.4 Terabytes
Variety
Electronic Medical/Health Records, Physician Management Systems, Claims Systems, and More

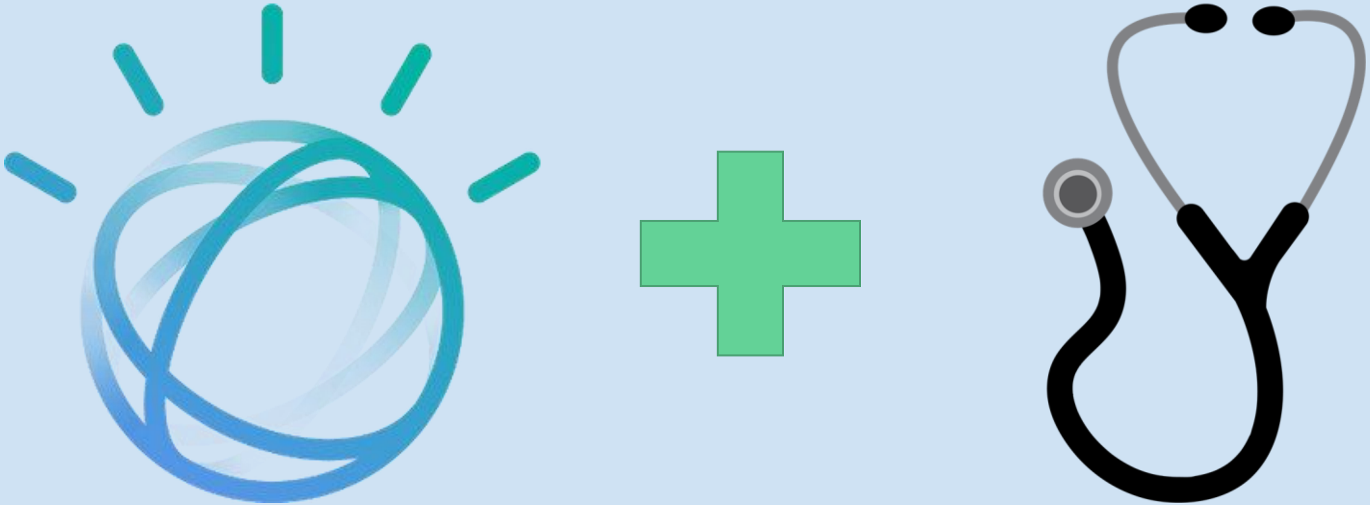
Figure taken from [4]

Limitation: Scalability

- MEDLINE Corpus
    - U.S National Library of Medicine
    - 28 million+ abstracts
    - 5000+ journals
    - 1.8 million abstracts published annually
- Average Researcher
    - 250-300 articles in a given year
        - Time factors limit this

The Need: Solution!!

Let Watson do (most of) the Work

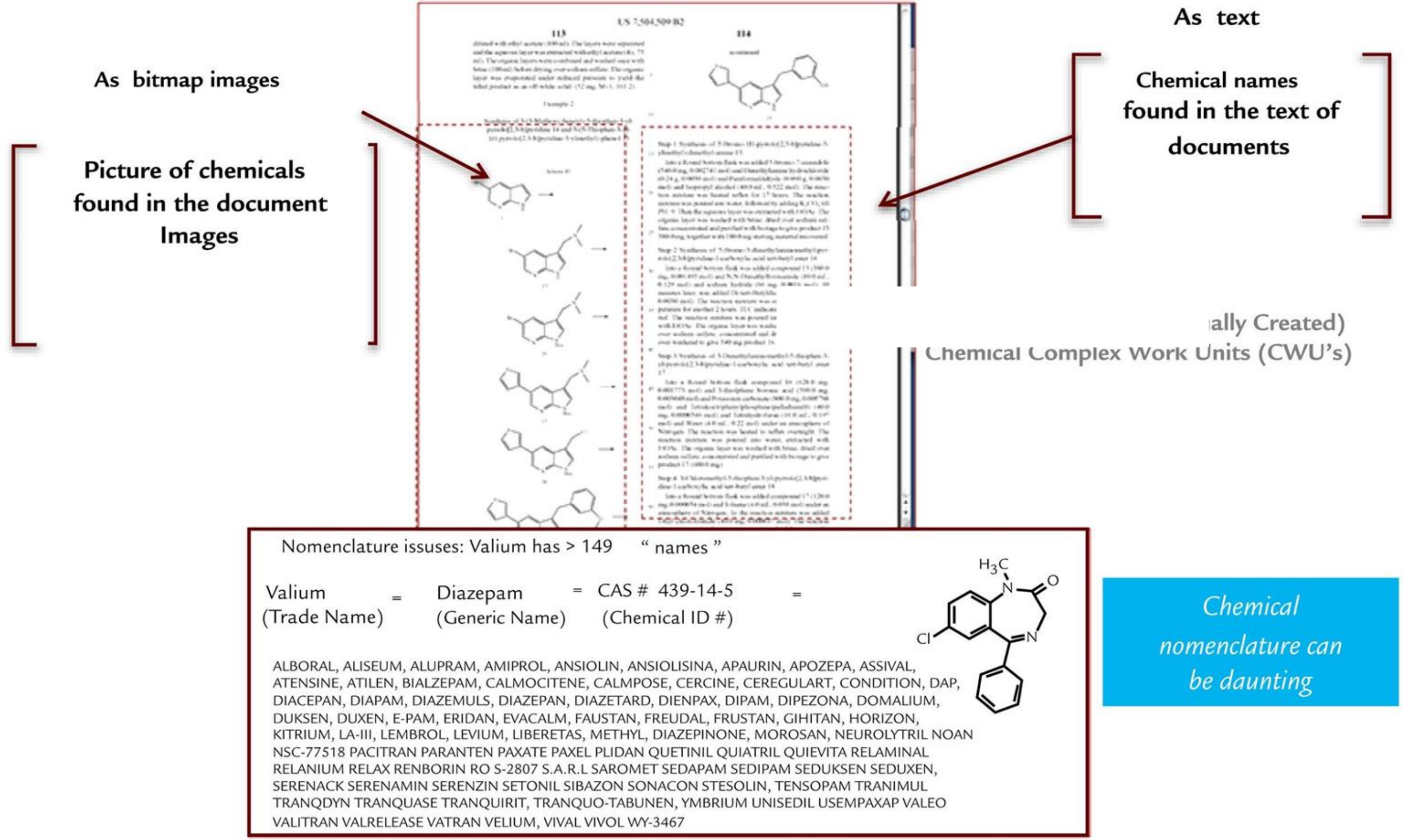


Figures from [5] and [6]

Outline

- The Problem ✓
- **Data in published medical research**
- How Watson works
- Case Study
- Conclusion

Understanding Data: Chemical Nomenclature


As bitmap images
Picture of chemicals found in the document Images
US 7,504,509 B2
113
114
As text
Chemical names found in the text of documents
ally Created)
Chemical Complex Work Units (CWU's)
H_3C
N
O
N
Cl
Chemical nomenclature can be daunting

Nomenclature issues: Valium has > 149 “ names ”

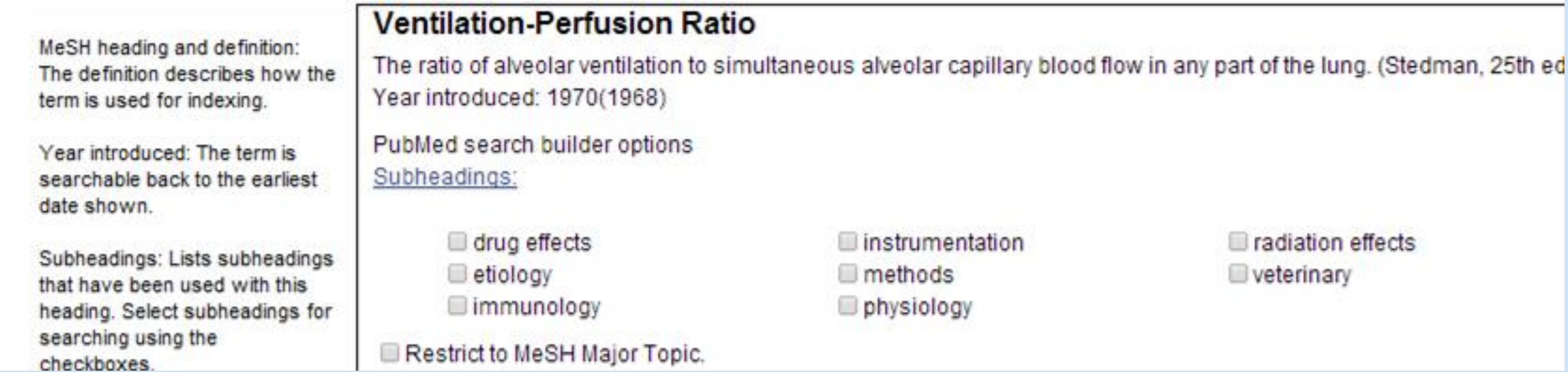
Valium (Trade Name) = Diazepam (Generic Name) = CAS # 439-14-5 (Chemical ID #) =

ALBORAL, ALISEUM, ALUPRAM, AMIPROL, ANSIOLIN, ANSIOLISINA, APAURIN, APOZEPA, ASSIVAL, ATENSINE, ATILEN, BIALZEPAM, CALMOCITENE, CALMPOSE, CERCINE, CEREGULART, CONDITION, DAP, DIACEPAN, DIAPAM, DIAZEMULS, DIAZEPAN, DIAZETARD, DIENPAX, DIPAM, DIPEZONA, DOMALIUM, DUKSEN, DUXEN, E-PAM, ERIDAN, EVACALM, FAUSTAN, FREUDAL, FRUSTAN, GIHITAN, HORIZON, KITRIUM, LA-III, LEMBROL, LEVIUM, LIBERETAS, METHYL, DIAZEPINONE, MOROSAN, NEUROLYTRIL NOAN NSC-77518 PACITRAN PARANTEN PAXATE PAXEL PLIDAN QUETINIL QUIATRIL QUIEVITA RELAMINAL RELANIUM RELAX RENBORIN RO S-2807 S.A.R.L SAROMET SEDAPAM SEDIPAM SEDUKSEN SEDUXEN, SERENACK SERENAMIN SERENZIN SETONIL SIBAZON SONACON STESOLIN, TENSOPAM TRANIMUL TRANQDYN TRANQUASE TRANQUIRIT, TRANQUO-TABUNEN, YMBRIUM UNISEDIL USEMPAXAP VALEO VALITRAN VALRELEASE VATRAN VELIUM, VIVAL VIVOL WY-3467

Figure from [2]

Understanding Data: MeSH

- Medical Subject Headings
- Manually curated series of vocabulary terms
    - National Library of Medicine
- Assigned to articles and books
    - Index citations
    - Facilitate Search health information



MeSH heading and definition: The definition describes how the term is used for indexing.

Year introduced: The term is searchable back to the earliest date shown.

Subheadings: Lists subheadings that have been used with this heading. Select subheadings for searching using the checkboxes.

Ventilation-Perfusion Ratio

The ratio of alveolar ventilation to simultaneous alveolar capillary blood flow in any part of the lung. (Stedman, 25th ed

Year introduced: 1970(1968)

PubMed search builder options

Subheadings:

□ drug effects
□ etiology
□ immunology
□ instrumentation
□ methods
□ physiology
□ radiation effects
□ veterinary

□ Restrict to MeSH Major Topic.

Figure from [3]

Outline

- The Need ✓
- Data in published medical research ✓
- **How Watson works**
- Case Study
- Conclusion

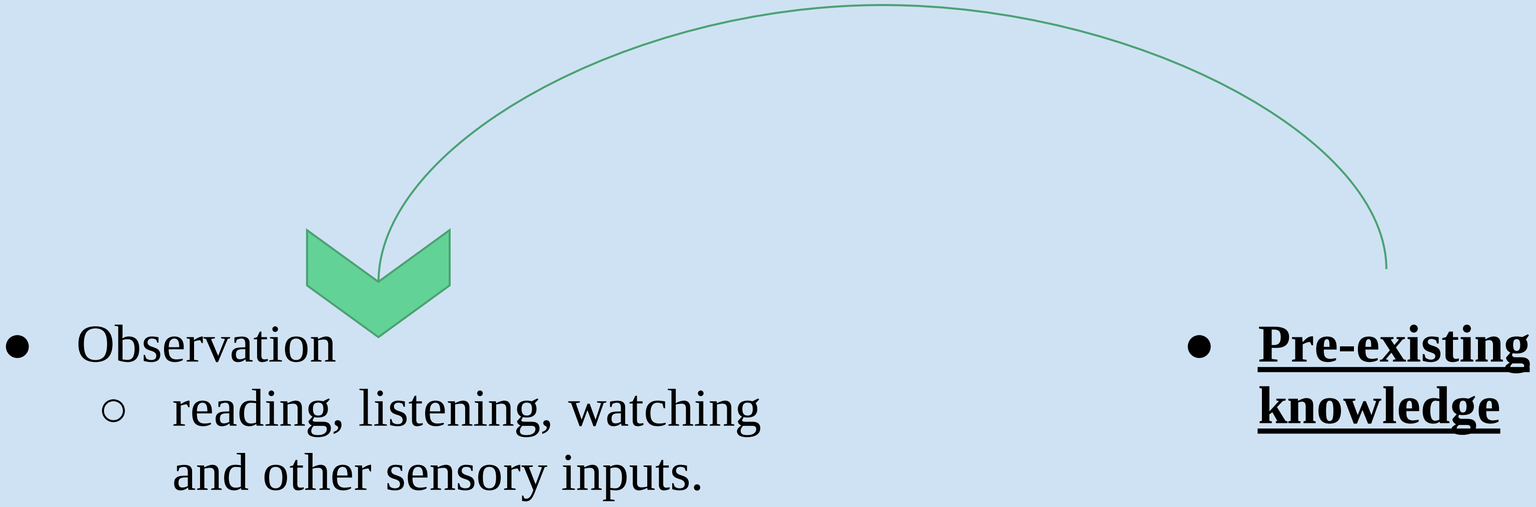
How Watson works

How ~~Watson~~ Humans reason

- Observation
  - reading, listening, watching and other sensory inputs.

- **<u>Pre-existing knowledge</u>**

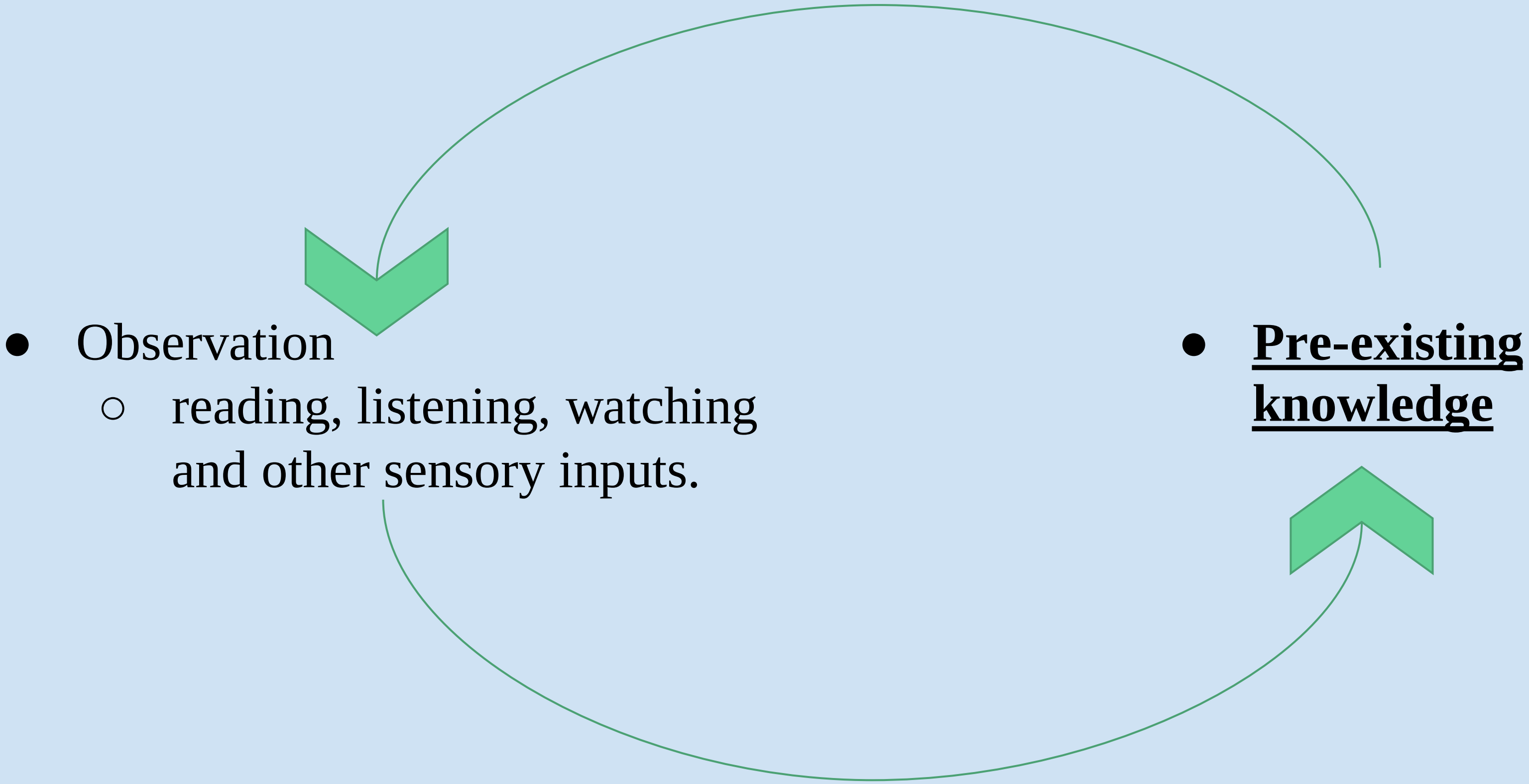
How ~~Watson~~ Humans reason



- Observation
  - reading, listening, watching and other sensory inputs.

- **Pre-existing knowledge**

How ~~Watson~~ Humans reason



- Observation
  - reading, listening, watching and other sensory inputs.

- **Pre-existing knowledge**

Foundational Knowledge

- *Establish a unique **corpus**
  - Dictionaries of domain-specific knowledge
- Key Concepts in the medical field
  - Genes
  - Drugs
  - Diseases
  - Symptoms
  - Chemicals

- **Entity Types** and **Entity**
- Examples
  - List of proteins associated with each gene.
  - Approval status of drugs.
  - Synonyms

Outline

- The Need ✓
- Data in published medical research ✓
- How Watson works
  - Foundational Knowledge ✓
  - **Named Entity Recognition**
  - Named Entity Resolution
  - Semantic Relationship Extraction
- Case Study
- Conclusion

Named Entity Recognition

- 1,3,7-trimethyl-purine-2,6-dione
- CHEMBL113
- “Caffeine is the world’s most widely consumed psychoactive drug…...the oral administration of CHEMBL113 was observed to. . . ”
- Dictionaries
  - compound names
  - synonyms
- Rule-based approach

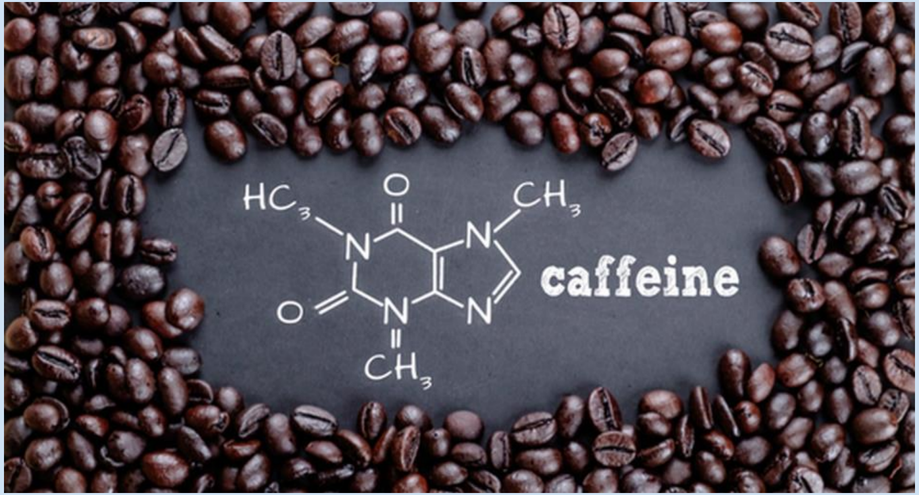


Figure from [7]

Rule based-approach

- Context Rules
  - Prevent subterms to be extracted
    - “Carbon” in context of “Carbon Dioxide”
  - Acronyms
    - Numerous
    - Lack of consistency
    - Temporary definition

Outline

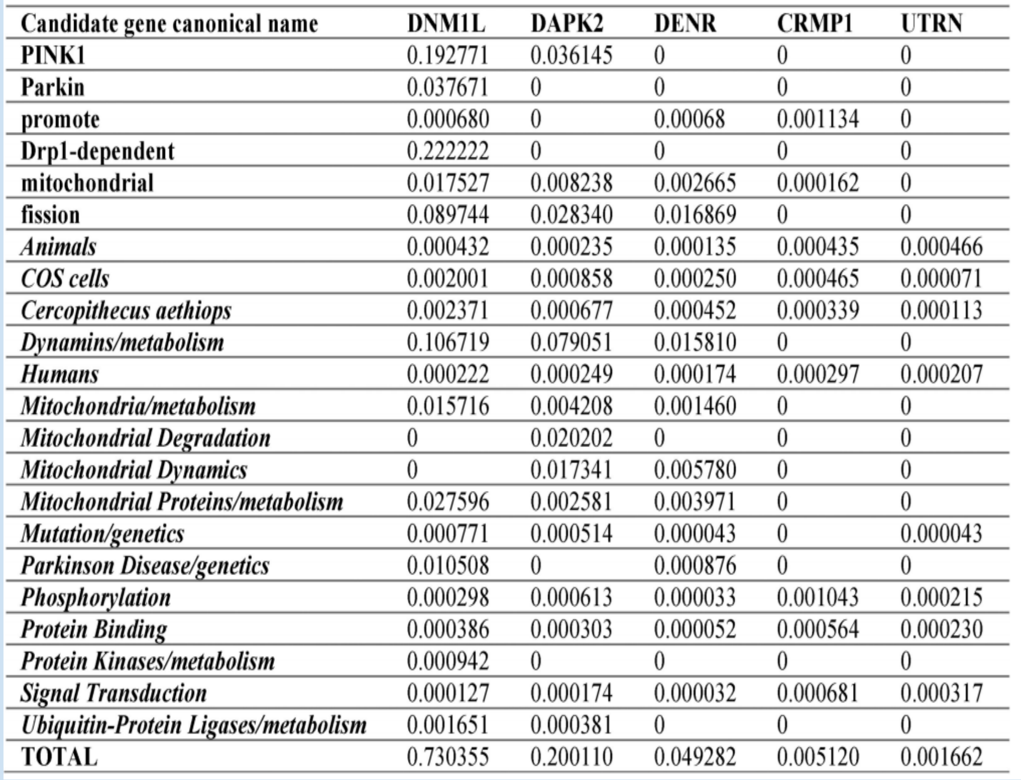
- The Need ✓
- Data in published medical research ✓
- How Watson works
  - Foundational Knowledge ✓
  - Named Entity Recognition ✓
  - **Named Entity Resolution**
  - Semantic Relationship Extraction
- Case Study
- Conclusion

Named entity Resolution

- General Normalization
  - Case normalization
    - Carbon, carbon, CARBON → carbon
  - Accent normalization
    - é → e
- **Canonical form**
- Normalization based on entity types
  - Chemicals, Compounds, **Genes**

Gene Normalization

- “We show that PINK1 and Parkin promote Drp1-dependent mitochondrial fission by mechanism that are least in part independent"
- Context terms
- **MeSH terms**
- Frequency of normalization



Candidate gene canonical name	DNM1L	DAPK2	DENR	CRMP1	UTRN
PINK1	0.192771	0.036145	0	0	0
Parkin	0.037671	0	0	0	0
promote	0.000680	0	0.00068	0.001134	0
Drp1-dependent	0.222222	0	0	0	0
mitochondrial	0.017527	0.008238	0.002665	0.000162	0
fission	0.089744	0.028340	0.016869	0	0
Animals	0.000432	0.000235	0.000135	0.000435	0.000466
COS cells	0.002001	0.000858	0.000250	0.000465	0.000071
Cercopithecus aethiops	0.002371	0.000677	0.000452	0.000339	0.000113
Dynamins/metabolism	0.106719	0.079051	0.015810	0	0
Humans	0.000222	0.000249	0.000174	0.000297	0.000207
Mitochondria/metabolism	0.015716	0.004208	0.001460	0	0
Mitochondrial Degradation	0	0.020202	0	0	0
Mitochondrial Dynamics	0	0.017341	0.005780	0	0
Mitochondrial Proteins/metabolism	0.027596	0.002581	0.003971	0	0
Mutation/genetics	0.000771	0.000514	0.000043	0	0.000043
Parkinson Disease/genetics	0.010508	0	0.000876	0	0
Phosphorylation	0.000298	0.000613	0.000033	0.001043	0.000215
Protein Binding	0.000386	0.000303	0.000052	0.000564	0.000230
Protein Kinases/metabolism	0.000942	0	0	0	0
Signal Transduction	0.000127	0.000174	0.000032	0.000681	0.000317
Ubiquitin-Protein Ligases/metabolism	0.001651	0.000381	0	0	0
TOTAL	0.730355	0.200110	0.049282	0.005120	0.001662

Figure from [1]

Outline

- The Need ✓
- Data in published medical research ✓
- How Watson works
  - Foundational Knowledge ✓
  - Named Entity Recognition ✓
  - Named Entity Resolution ✓
  - **Semantic Relationship Extraction**
- Case Study
- Conclusion

Semantic Relationship Extraction

- Relationship
  - Two distinct entities
    - Agent
    - Target
  - Domain-relevant verb or Trigger word
- Example
  - “The results show that ERK2 phosphorylated p53”.
- Normalization
  - “phosphorylated” → “phosphorylate”
  - “bring” or “overlap” → “association”

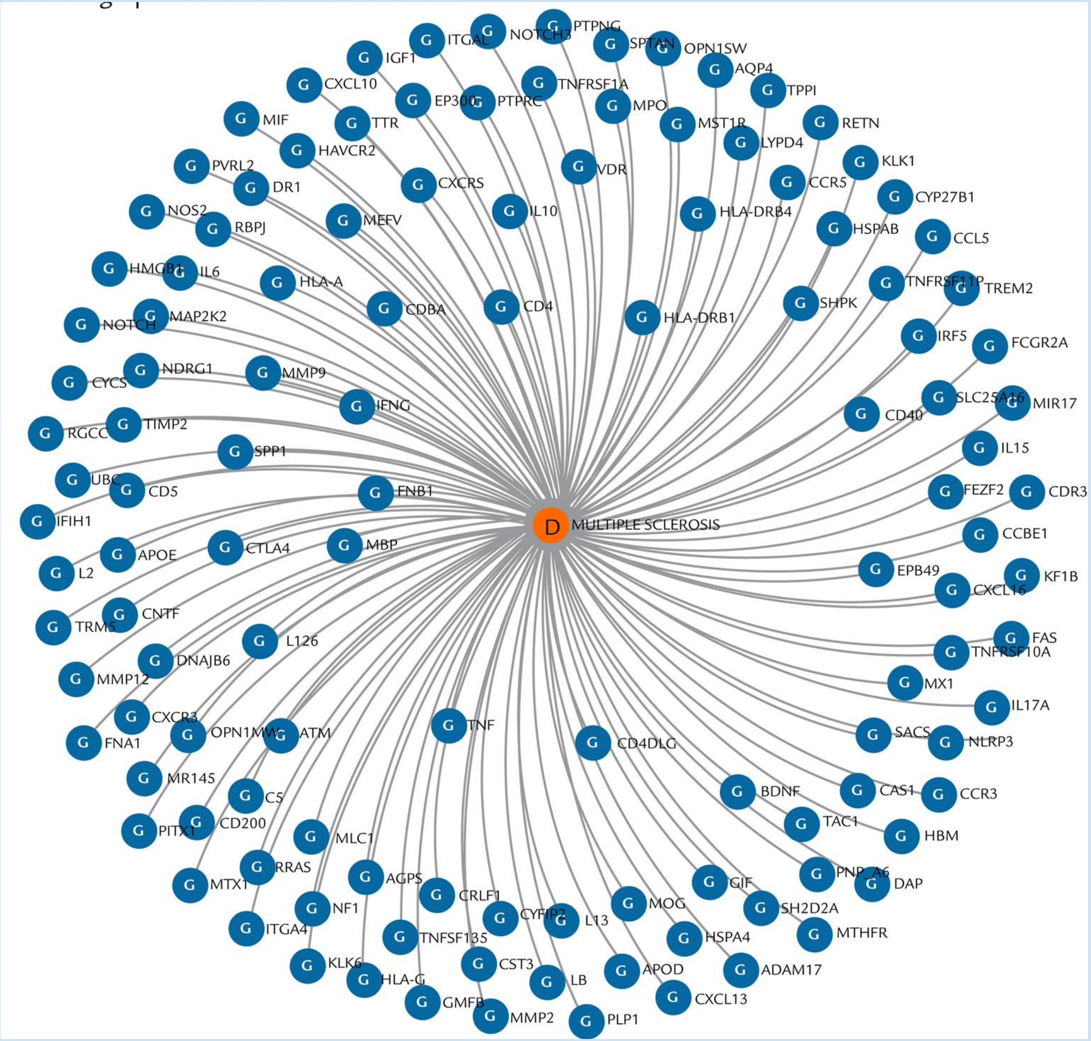


Figure from [2]

Outline

- The Need ✓
- Data in published medical research ✓
- How Watson works ✓
  - Foundational Knowledge ✓
  - Named Entity Recognition ✓
  - Named Entity Resolution ✓
  - Semantic Relationship Extraction ✓
- **Case Study**
- Conclusion

Case Study

- **Artificial intelligence in neurodegenerative disease research:use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis**
  - 2017 study
  - Identifying proteins altered in ALS

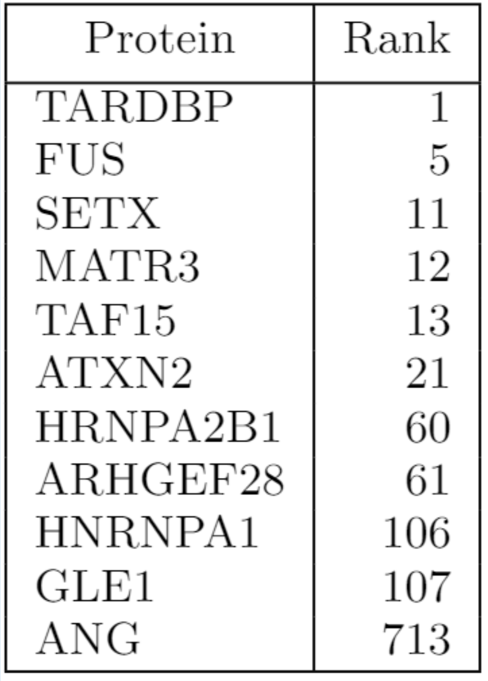
- What is ALS?
  - Disease
    - loss of muscle control
  - No effective treatment
  - Linked to RNA binding proteins(**RBPs)** in patients

Background

- RBPs
    - 1542 RBPs-encoding genes in human genome
    - **11 genes** have shown mutations related to ALS
    - **6 other RBPs** with alterations related to ALS
        - Gene hasn't been linked to a mutation
    - Less than 1% of RBPs have yet to be linked to ALS
- Hypothesis:
    - Additional RBPs contribute to ALS
- Predict potential candidates
- Limitation
    - Only **1,478 RBPs** were mentioned at least once in published abstracts

Validating Watson

- Leave-one-out cross validation (LOOCV)
  - Applied an algorithm 11 times
  - A different RBP from known gene mutation is moved into the candidate set alongside the other 1,478 RBPs
- 90% of the known proteins ranked are in top 7 %

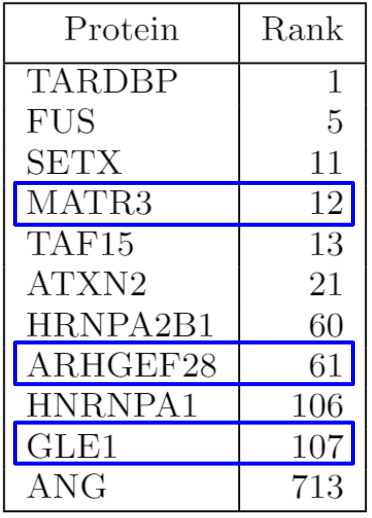


Protein	Rank
TARDBP	1
FUS	5
SETX	11
MATR3	12
TAF15	13
ATXN2	21
HRNPA2B1	60
ARHGEF28	61
HNRNPA1	106
GLE1	107
ANG	713

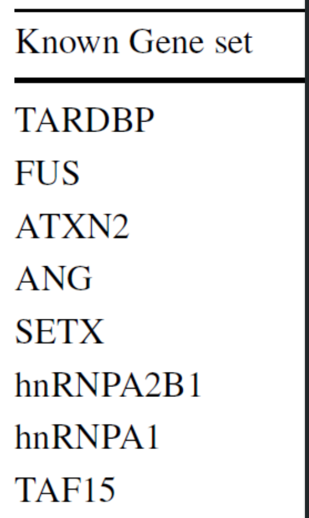
Figure from [1]

Retrospective Study

- Literature published up to 2012
  - 8 known RBPs linked to mutations
  - 1,439 out 1,487 RBPs
- Goal:
  - How would Watson rank the other three ?
    - MATR3, ARGHEF28 and GLE1
    - Found 2013 - 2017



Protein	Rank
TARDBP	1
FUS	5
SETX	11
MATR3	12
TAF15	13
ATXN2	21
HRNPA2B1	60
ARHGEF28	61
HNRNPA1	106
GLE1	107
ANG	713

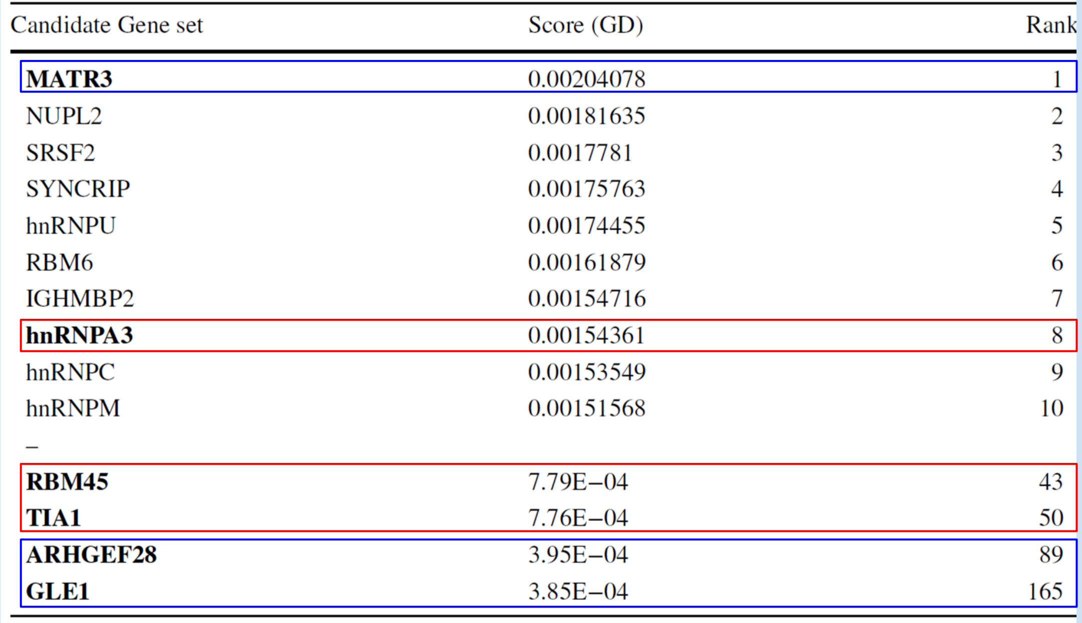


Known Gene set
TARDBP
FUS
ATXN2
ANG
SETX
hnRNPA2B1
hnRNPA1
TAF15

Figures from [1]

Retrospective Study Results

- Blue Box
  - Proteins with known gene mutations
- Red Box:
  - Altered proteins without known gene mutation
- Ranked in top 165 (11%) of candidate gene set
- What if Watson was used in 2012 ?
  - MATR3 → May 2014

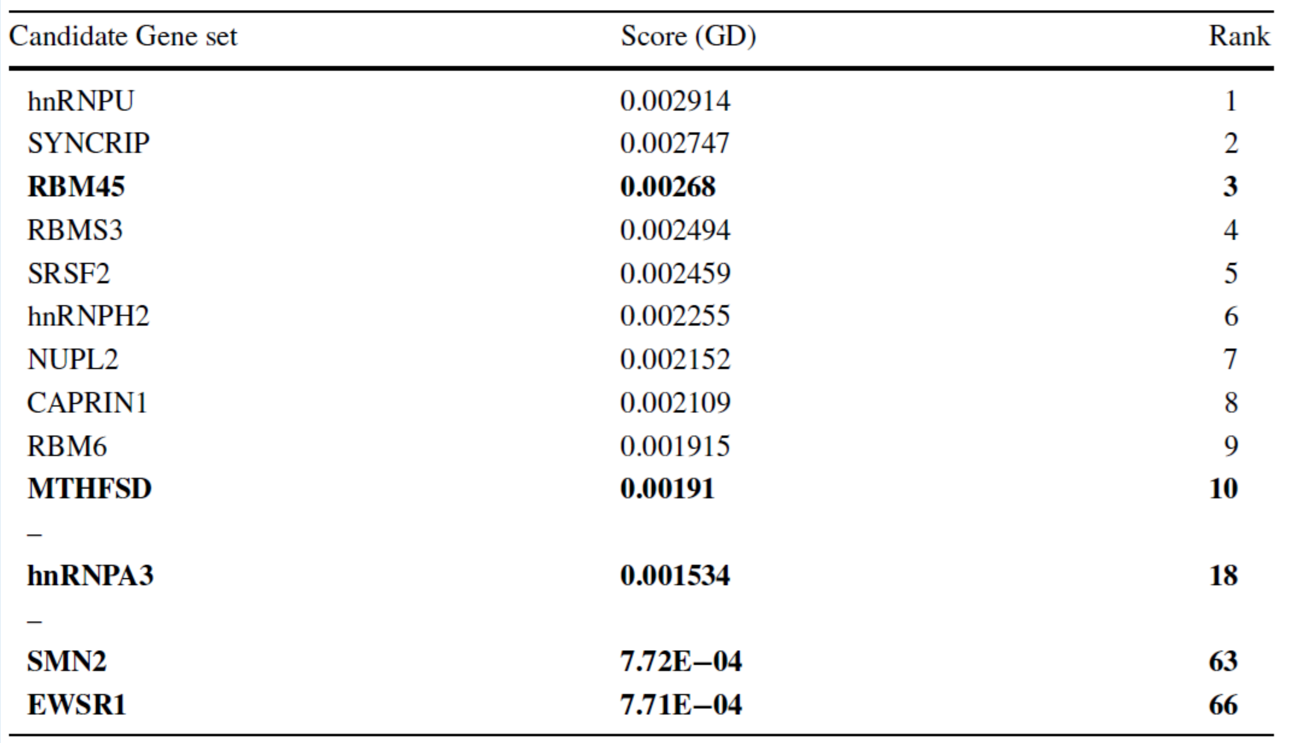


Candidate Gene set	Score (GD)	Rank
MATR3	0.00204078	1
NUPL2	0.00181635	2
SRSF2	0.0017781	3
SYNCRIP	0.00175763	4
hnRNPU	0.00174455	5
RBM6	0.00161879	6
IGHMBP2	0.00154716	7
hnRNPA3	0.00154361	8
hnRNPC	0.00153549	9
hnRNPM	0.00151568	10
–		
RBM45	7.79E–04	43
TIA1	7.76E–04	50
ARHGEF28	3.95E–04	89
GLE1	3.85E–04	165

Figure from [1]

Prospective Study

- 1478 RBPs and 11 known genes



Candidate Gene set	Score (GD)	Rank
hnRNPU	0.002914	1
SYNCRIP	0.002747	2
RBM45	**0.00268**	**3**
RBMS3	0.002494	4
SRSF2	0.002459	5
hnRNPH2	0.002255	6
NUPL2	0.002152	7
CAPRIN1	0.002109	8
RBM6	0.001915	9
MTHFSD	**0.00191**	**10**
–		
hnRNPA3	**0.001534**	**18**
–		
SMN2	**7.72E–04**	**63**
EWSR1	**7.71E–04**	**66**

Altered proteins without known gene mutation

Figure from [1]

Validation and Results

- Validation
  - Positive control: 8 of the top 10 candidates
  - Negative control: Bottom 3 candidates (rank 1476-1478)
  - 4 different biological methods
    - Show significant difference in at least two methods
- Results
  - 5/8 RBPs showed significant alterations.
  - No alternations in bottom RBPs

Outline

- The Need ✓
- Data in published medical research ✓
- How Watson works ✓
  - Foundational Knowledge ✓
  - Named Entity Recognition ✓
  - Named Entity Resolution ✓
  - Semantic Relationship Extraction ✓
- Case Study ✓
- **Conclusion**

Conclusion

- Powerful tool
  - Analyzing published literature at a scale
  - Better selection of candidates for further examination
- Widespread Adoption ?

Acknowledgements

Thanks to Nic Mcphee and Elena Machkasova for their advice.

Questions

?

References

[1] N. Bakkar, T. Kovalik, I. Lorenzini, S. Spangler,A. Lacoste, K. Sponaugle, P. Ferrante, E. Argentinis,R. Sattler, R. Bowser, and et al. Artificial intelligence in neurodegenerative disease research: Use of IBM Watson to identify additional RNA-binding proteins altered in Amyotrophic Lateral Sclerosis.Acta Neuropathologica, 135(2):227–247, 2017

[2] Y. Chen, J. Elenee Argentinis, and G. Weber. IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research.Clinical Therapeutics, 38(4):688 – 701, 2016.

[3] The MeSH database. https://www.nlm.nih.gov/bsd/disted/meshtutorial/themeshdatabase/index.html.

References

[4] M. N. Ahmed, A. S. Toor, K. O’Neil, and D. Friedland.Cognitive computing and the future of healthcare cognitive computing and the future of healthcare: The cognitive power of ibm watson has the potential to transform global personalized medicine.IEEE Pulse,8(3):4–9, 2017.

[5] Wikipedia. Watson (computer) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Watson\%20(computer)&oldid=982177596

[6] Malnutrition - clip art stethoscope png transparent png- full size clipart (265149) - pinclipart. https://www.pinclipart.com/maxpin/iRhRTR/

[7] C. Graham. How much caffeine in a cup!!!! https://halatreecoffee.com/kona-coffee-blog/how-much-caffeine-in-a-cup/, Jul 2019