

Health Search

From Consumers to Clinicians

Slides available at

<https://ielab.io/russir2018-health-search-tutorial/>

Guido Zuccon
Queensland University of Technology

 @guidozuc



Outline

- Dealing with the **semantic gap**: exploiting the semantics of medical language
 - concept based search & inference, query expansion, learning to rank
- Dealing with the nuances of **medical language**
 - negation, family history, understandability
- Understanding and aiding **query formulation**
 - query variations, query reformulation, query clarification, query suggestion, query intent, query difficulty, task-based solutions

Dealing with the semantic gap

Exploiting semantics of medical language

- What are medical concepts, where are they defined
- Why use concepts
- Why concepts and terms

Medical concepts

- Medical concepts are defined in domain knowledge resource
- Capture the key aspects of the domain or some specific sub-domain
- Relationships between concepts capture associations

Implicit VS Explicit Semantics

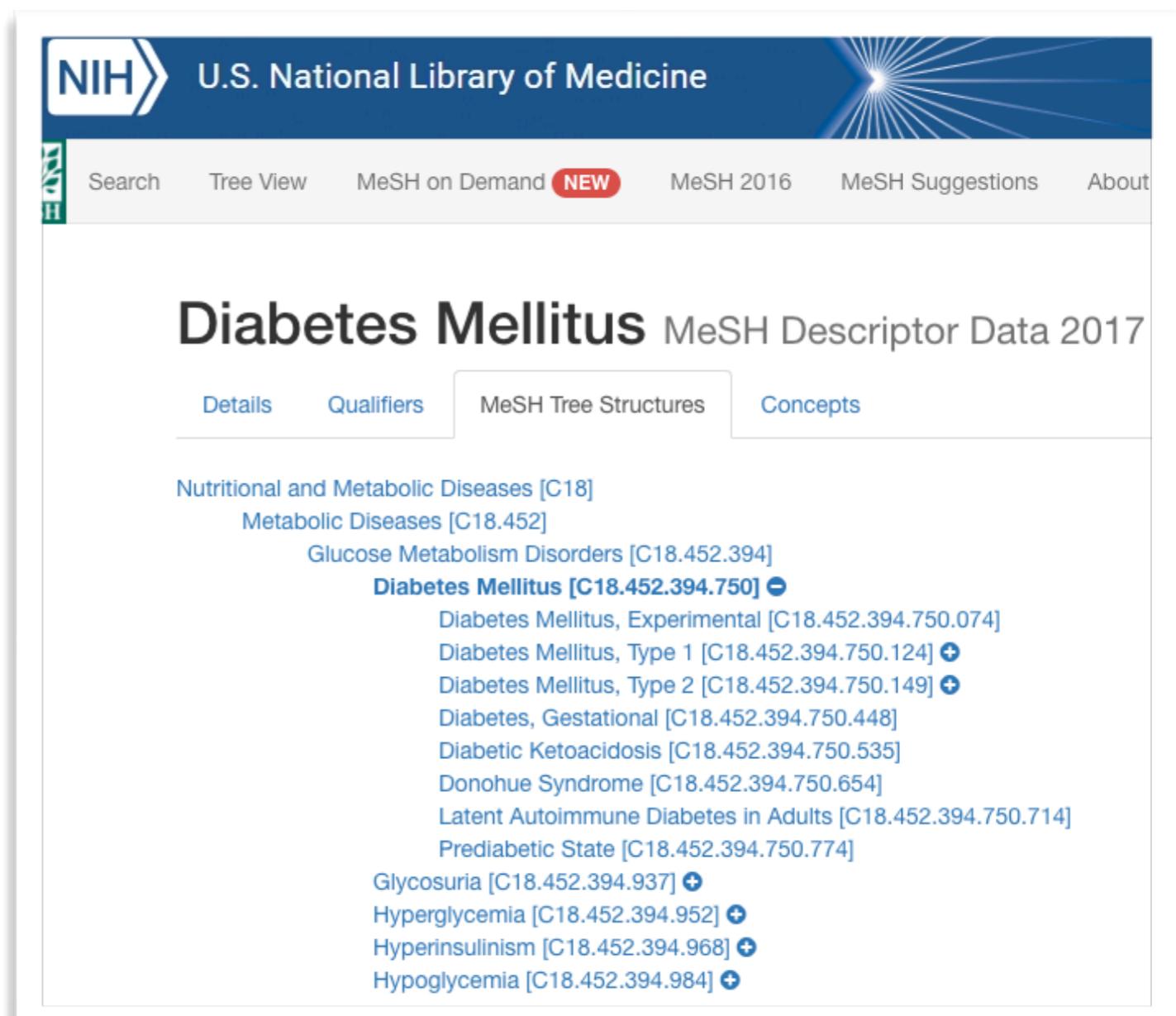
- Explicit semantics: structured human representation of knowledge and its concepts
 - e.g., medical terminologies
- Implicit Semantics: draw representation of words/concepts from data
 - e.g., distributional/latent semantic models

Key Medical Terminologies

Medical Subject Headings (MeSH)

Controlled vocabulary for
indexing journal articles

Mainly used by researchers
and clinicians searching the
literature.



The screenshot shows the MeSH Descriptor Data 2017 interface for the term "Diabetes Mellitus". The top navigation bar includes the NIH logo, the U.S. National Library of Medicine, and links for Search, Tree View, MeSH on Demand (NEW), MeSH 2016, MeSH Suggestions, and About. Below the navigation, the main title "Diabetes Mellitus" is displayed, followed by "MeSH Descriptor Data 2017". A horizontal menu bar offers options for Details, Qualifiers, MeSH Tree Structures, and Concepts. The "Concepts" tab is currently selected. A list of related concepts is shown, each with a link to its descriptor data, such as "Nutritional and Metabolic Diseases [C18]", "Metabolic Diseases [C18.452]", "Glucose Metabolism Disorders [C18.452.394]", and "Diabetes Mellitus [C18.452.394.750]".

Diabetes Mellitus MeSH Descriptor Data 2017

Details Qualifiers MeSH Tree Structures **Concepts**

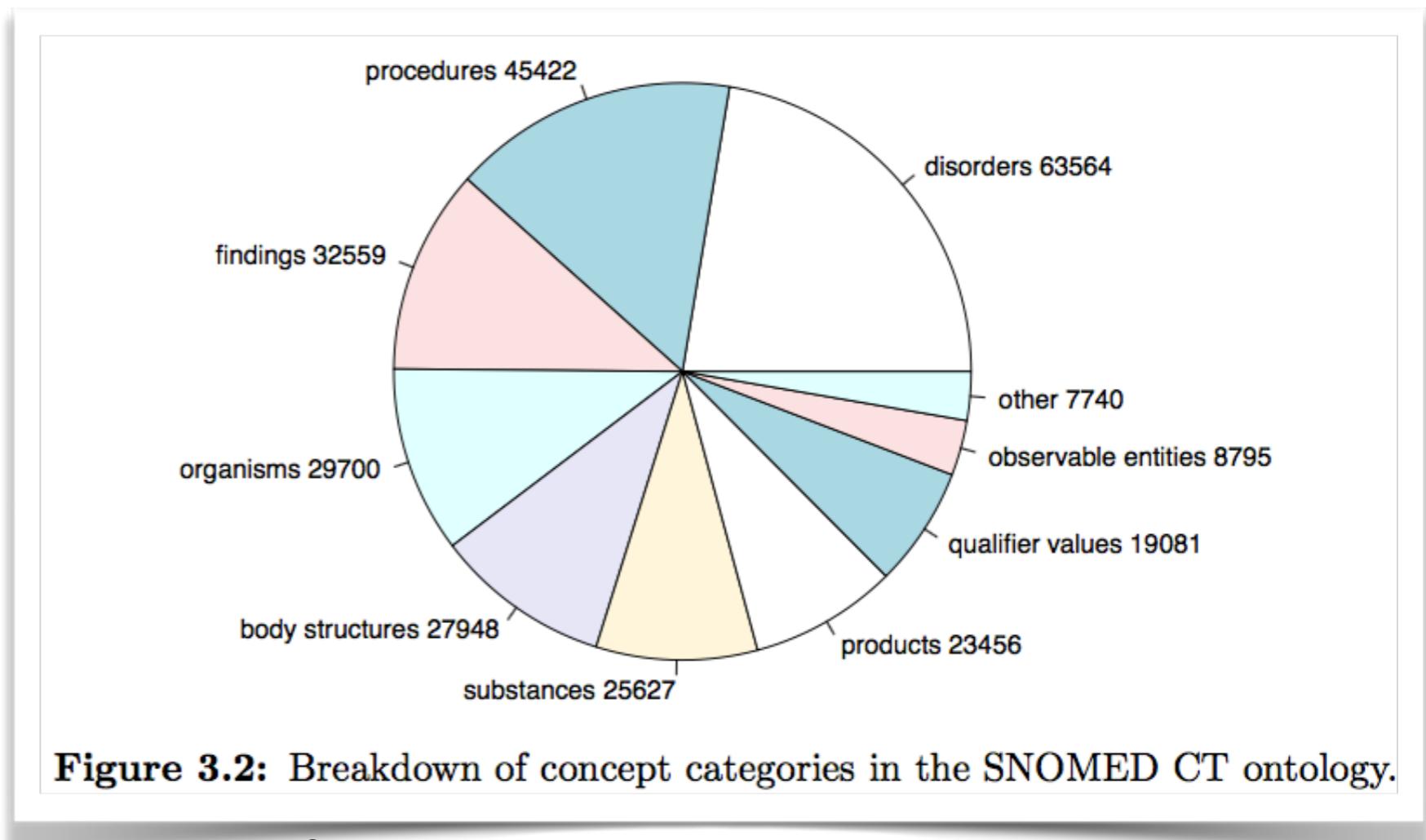
Nutritional and Metabolic Diseases [C18]
Metabolic Diseases [C18.452]
Glucose Metabolism Disorders [C18.452.394]
Diabetes Mellitus [C18.452.394.750] 
 Diabetes Mellitus, Experimental [C18.452.394.750.074]
 Diabetes Mellitus, Type 1 [C18.452.394.750.124] 
 Diabetes Mellitus, Type 2 [C18.452.394.750.149] 
 Diabetes, Gestational [C18.452.394.750.448]
 Diabetic Ketoacidosis [C18.452.394.750.535]
 Donohue Syndrome [C18.452.394.750.654]
 Latent Autoimmune Diabetes in Adults [C18.452.394.750.714]
 Prediabetic State [C18.452.394.750.774]
 Glycosuria [C18.452.394.937] 
 Hyperglycemia [C18.452.394.952] 
 Hyperinsulinism [C18.452.394.968] 
 Hypoglycemia [C18.452.394.984] 

SNOMED CT

Formal medical ontology: ~500,000 concepts ~3,000,000 relationships

Becoming de-facto mean of formally representing clinical data.

Adopted by software vendors



SNOMED CT

Formal medical ontology: ~500,000 concepts ~3,000,000 relationships

Bec
Add
ven

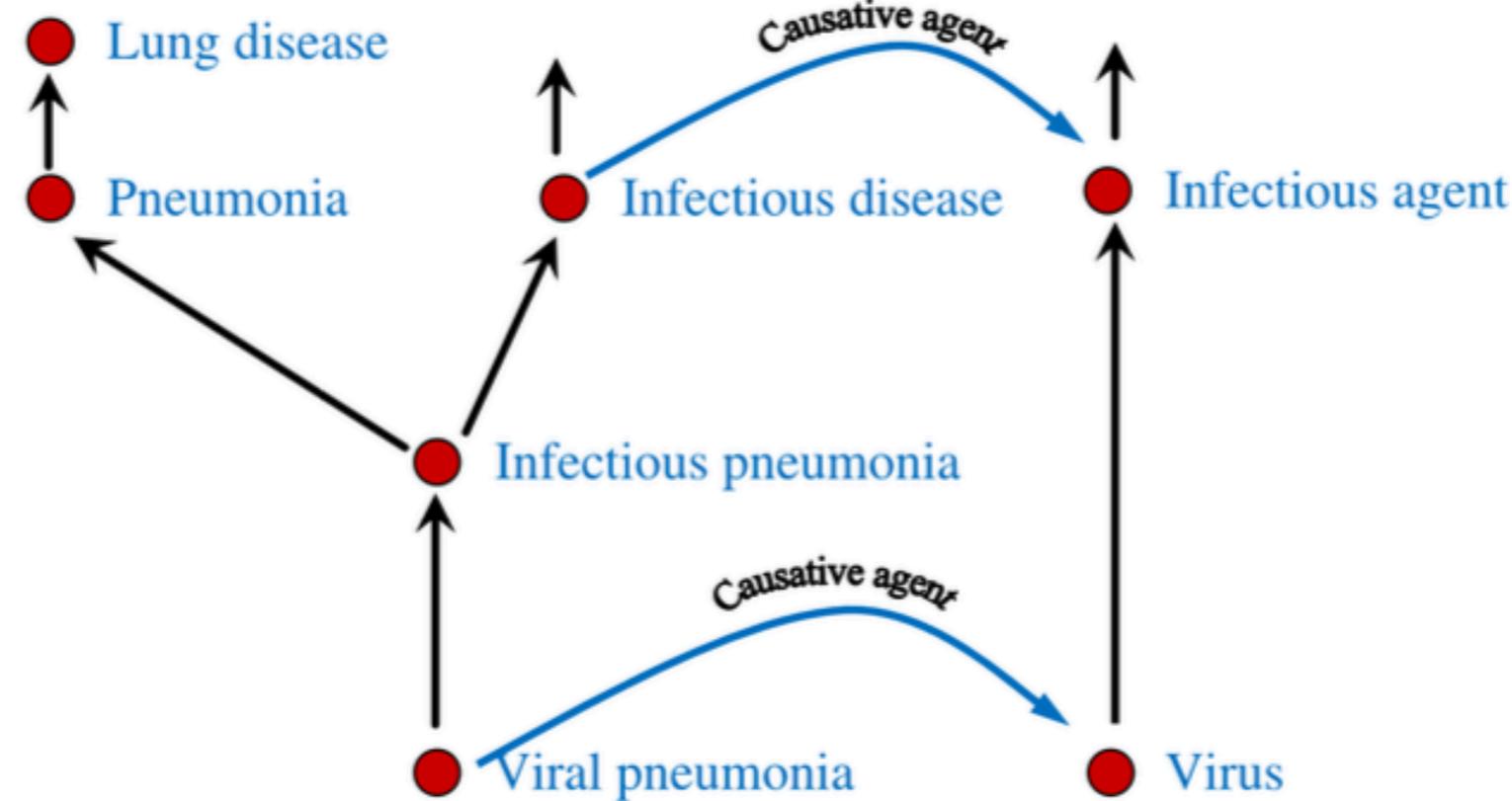


Figure 3.3: Concept hierarchy for *Viral pneumonia*.

data.

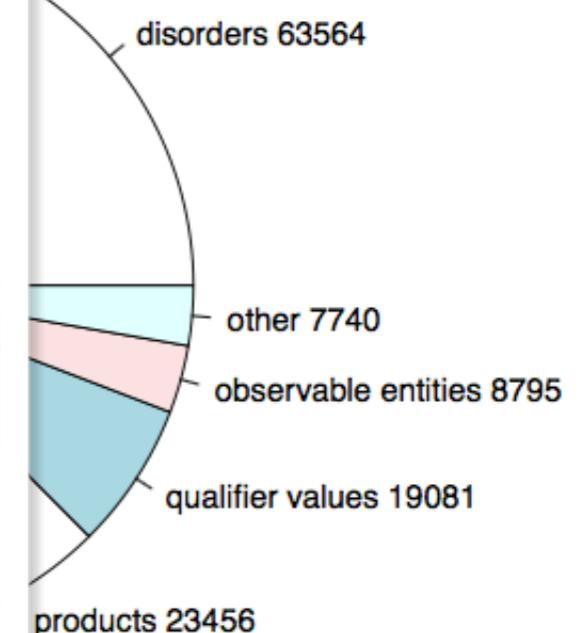


Figure 3.2: Breakdown of concept categories in the SNOMED CT ontology.

ICD

International Statistical
Classification of Diseases and
Related Health Problems
(ICD)

Diagnosis classification from
World Health Organisation

Used extensively in **billing**

International Statistical Classification of Diseases and
Related Health Problems 10th Revision

Chapter	Blocks	Title
I	A00– B99	Certain infectious and parasitic diseases
II	C00– D48	Neoplasms
III	D50– D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00– E90	Endocrine, nutritional and metabolic diseases
V	F00– F99	Mental and behavioural disorders
VI	G00– G99	Diseases of the nervous system
VII	H00– H59	Diseases of the eye and adnexa
VIII	H60– H95	Diseases of the ear and mastoid process
IX	I00–I99	Diseases of the circulatory system
X	J00– J99	Diseases of the respiratory system
XI	K00– K93	Diseases of the digestive system
XII	L00– L99	Diseases of the skin and subcutaneous tissue
	M00–	Diseases of the musculoskeletal system

Unified Medical Language System (UMLS)

- UMLS is a compendium of many controlled vocabularies in the biomedical sciences
- **Combined many terminologies under one umbrella**
- UMLS concept grouped into higher level semantic types
 - Concept: *Myocardial Infarction* [C0027051] of type *Disease or Syndrome* [T047]
 - <https://uts.nlm.nih.gov//metathesaurus.html>



An important note

- These resources contain information that can help characterise medical language
 - Synonyms of a term
 - Relationship between terms/concepts
- Rarely do these resources contain information that directly answers questions like
 - What is the drug of choice for condition x?
 - What is the cause of symptom x?
 - What test is indicated in situation x?
 - How should I treat condition x (not limited to drug treatment)?
 - How should I manage condition x (not specifying diagnostic or therapeutic)?
 - What is the cause of physical finding x?
 - What is the cause of test finding x?
 - Can drug x cause (adverse) finding y?
 - Could this patient have condition x?
- That is, they **do not directly resolve the clinical questions** presented in [Ely et al., 2000] taxonomy
- They capture truisms/**universal facts**, not subjective knowledge/things that could change over time

Convert Terms to Concepts (aka Concept Mapping)

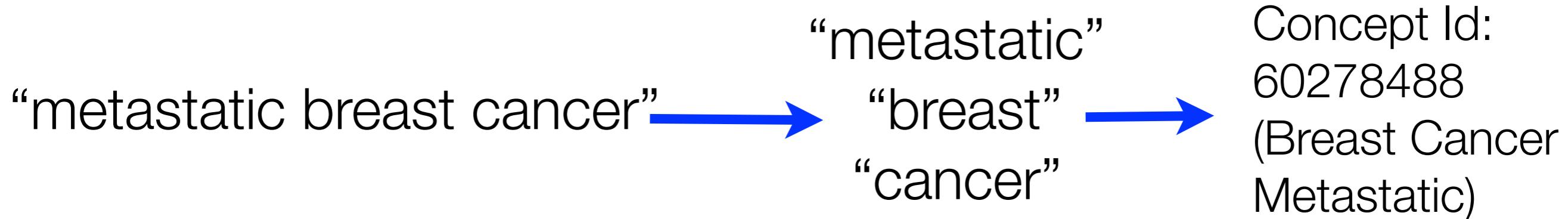
Convert Terms to Concepts (aka Concept Mapping)

“metastatic breast cancer”

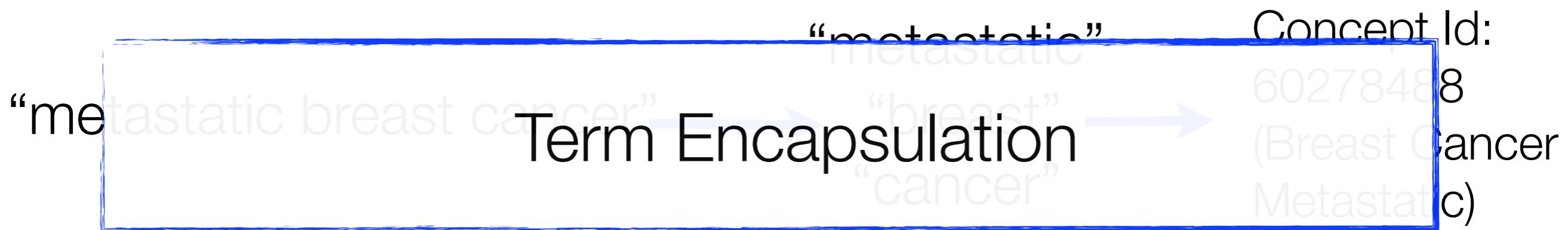
Convert Terms to Concepts (aka Concept Mapping)

“metastatic breast cancer” → “metastatic”
“breast”
“cancer”

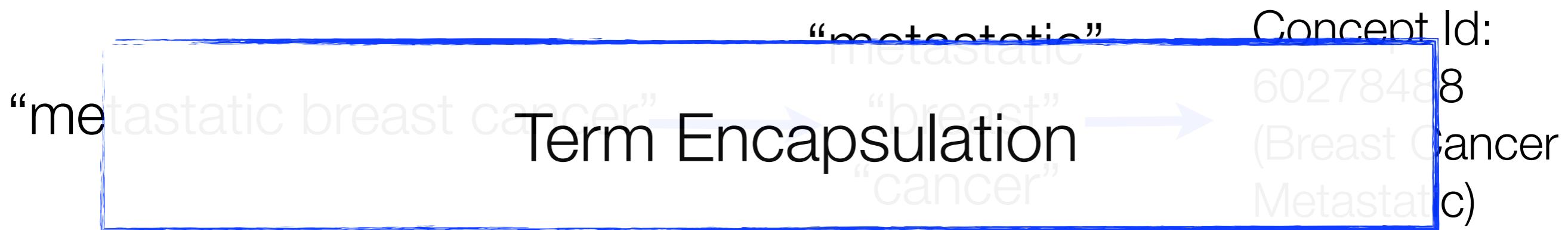
Convert Terms to Concepts (aka Concept Mapping)



Convert Terms to Concepts (aka Concept Mapping)



Convert Terms to Concepts (aka Concept Mapping)



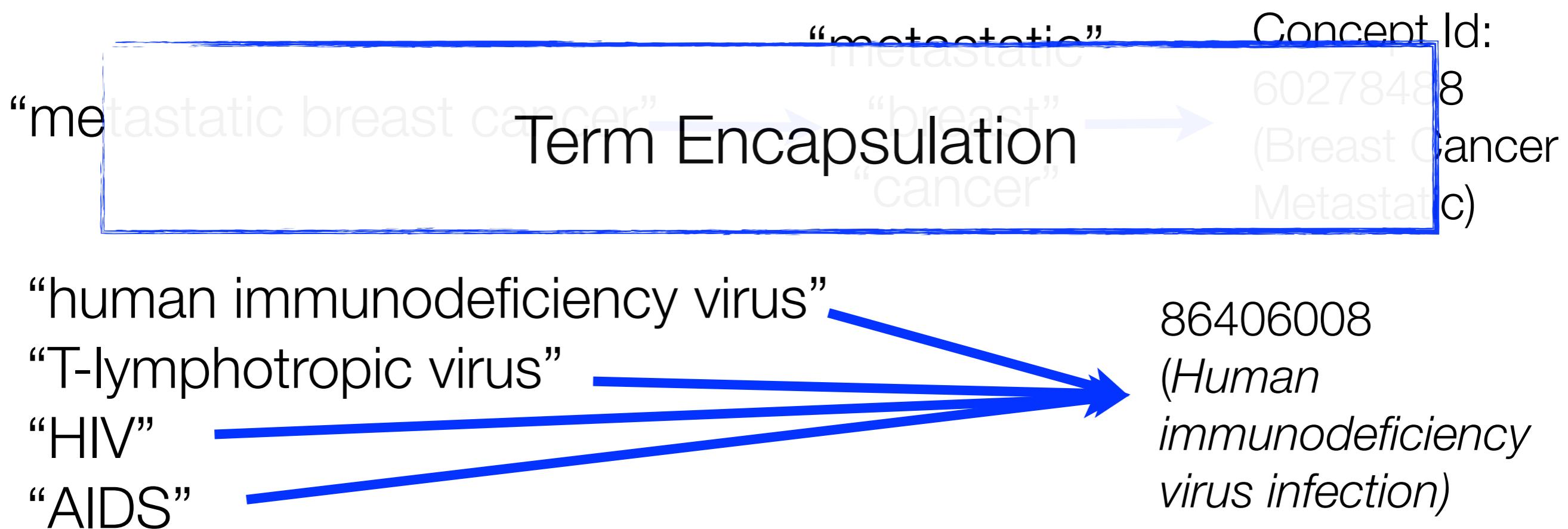
“human immunodeficiency virus”

“T-lymphotropic virus”

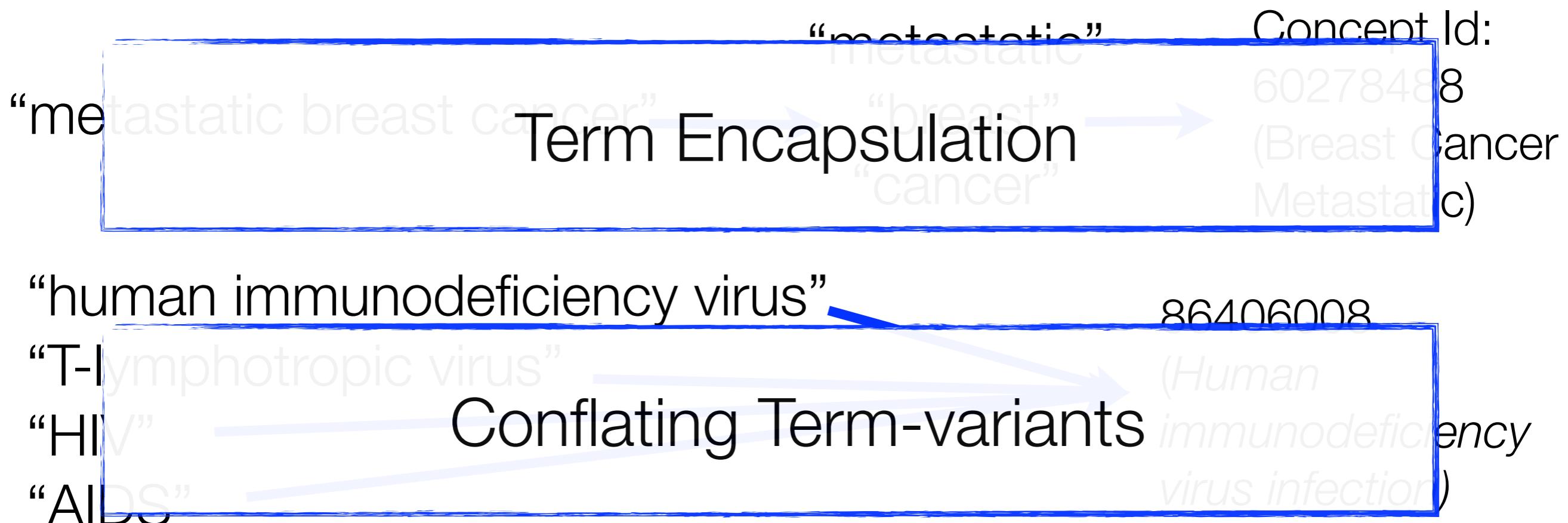
“HIV”

“AIDS”

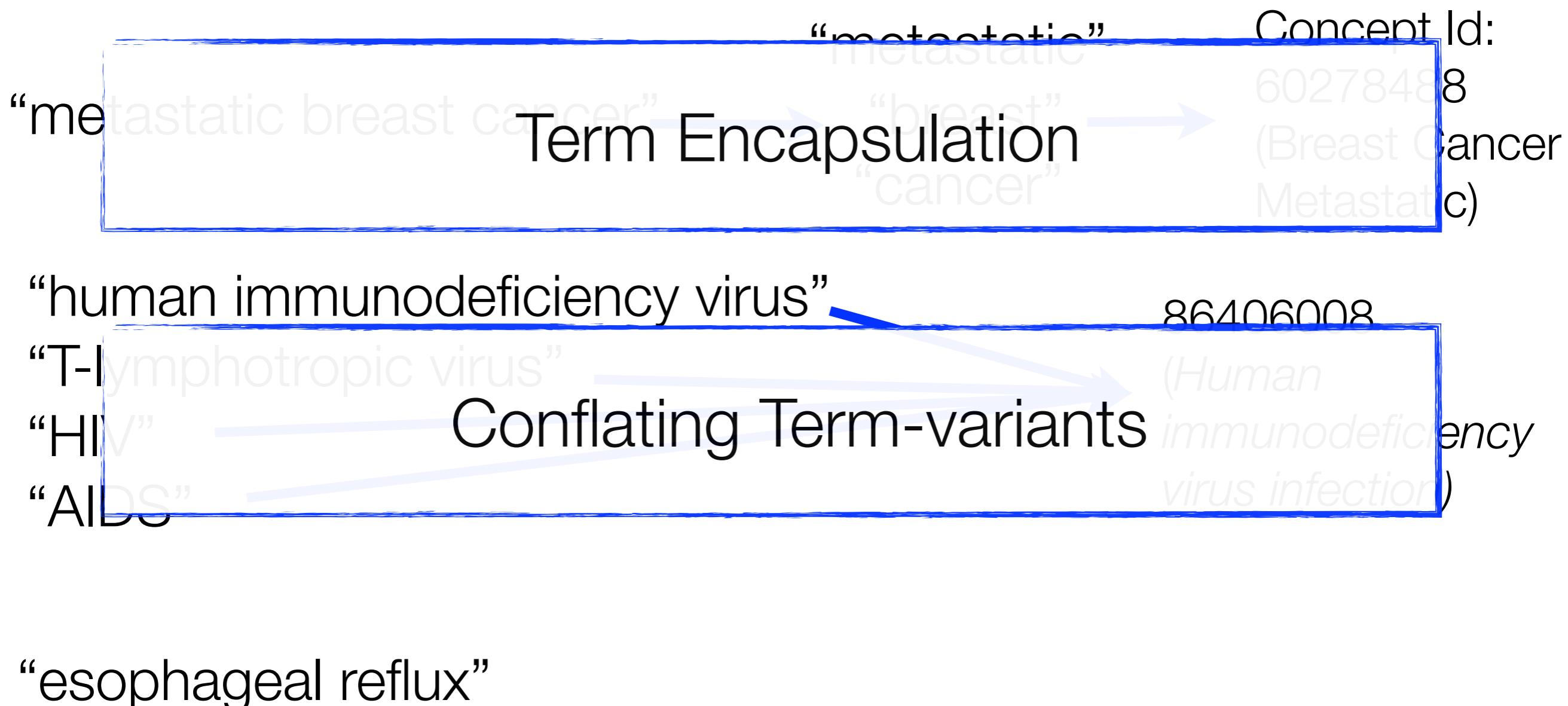
Convert Terms to Concepts (aka Concept Mapping)



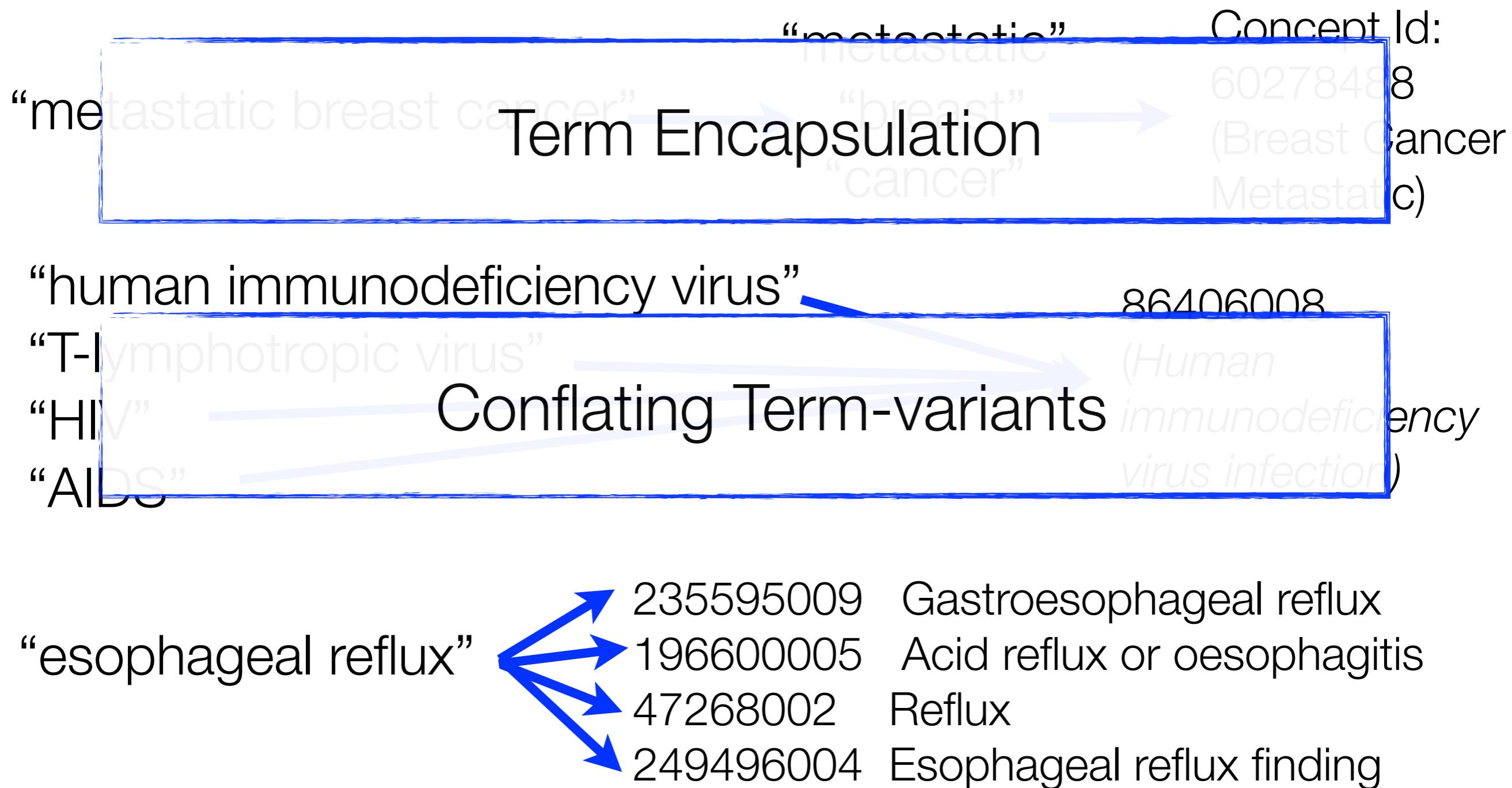
Convert Terms to Concepts (aka Concept Mapping)



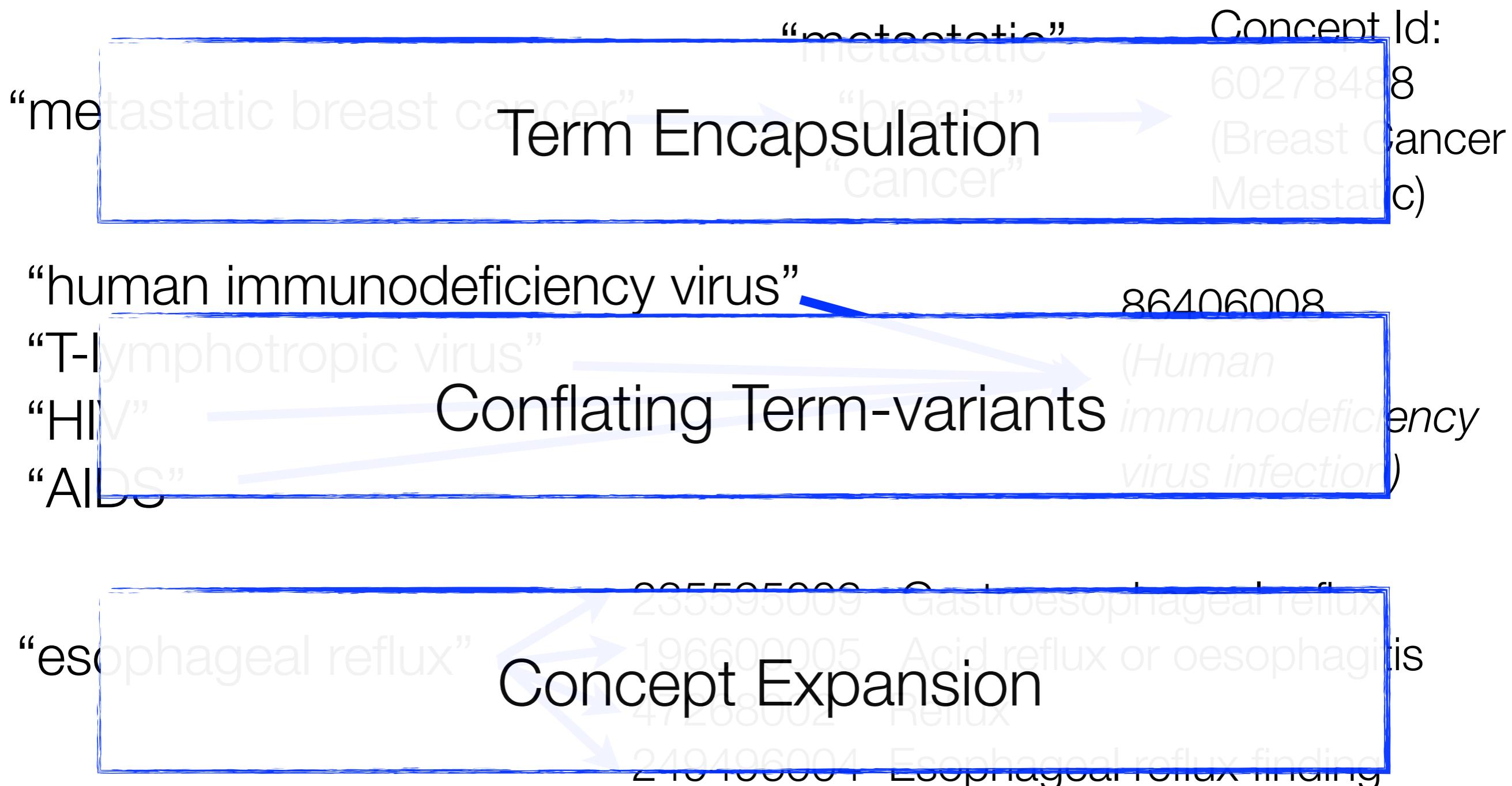
Convert Terms to Concepts (aka Concept Mapping)



Convert Terms to Concepts (aka Concept Mapping)



Convert Terms to Concepts (aka Concept Mapping)

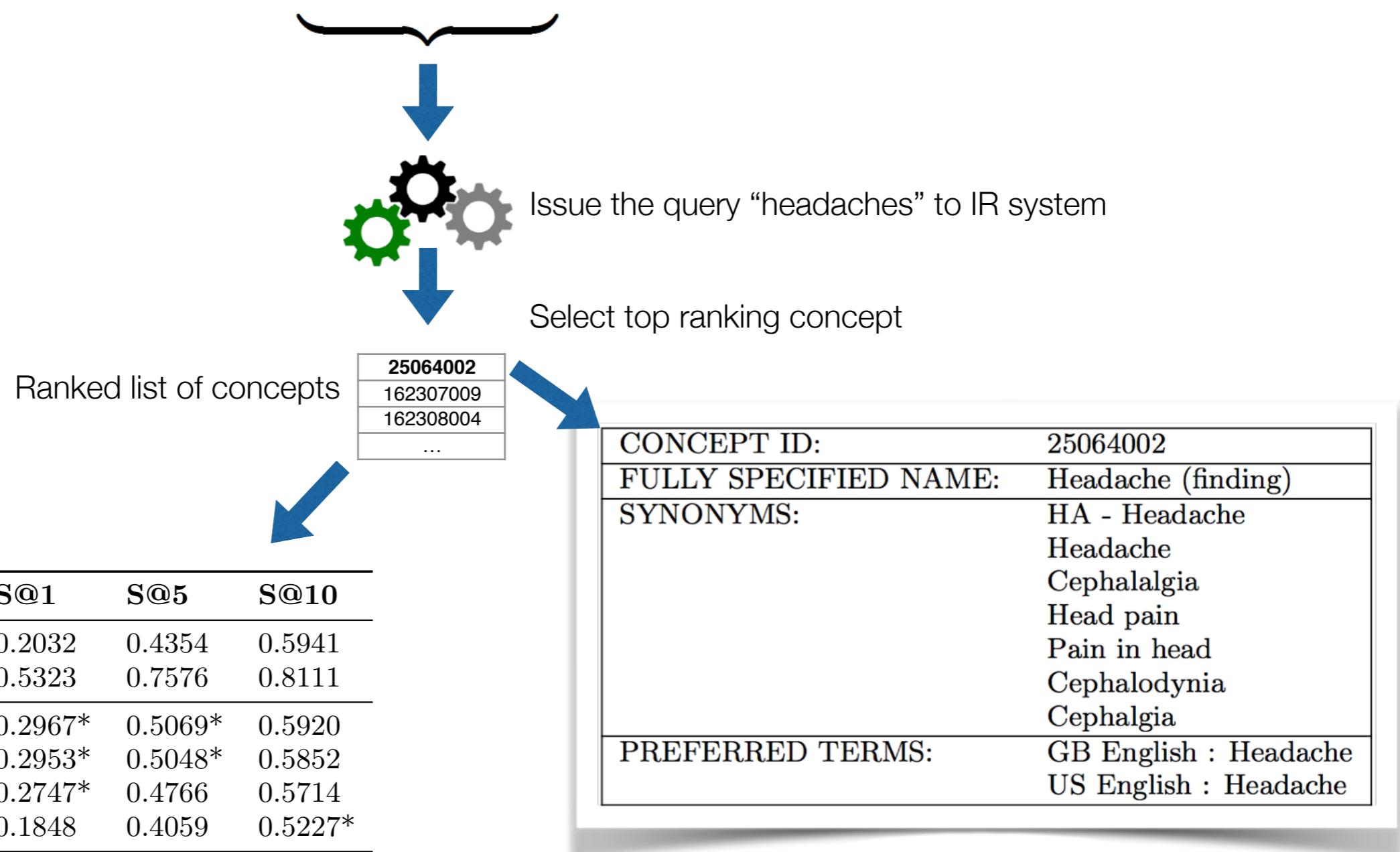


Concept extraction/mapping tools

- **Metamap** — National Library of Medicine [[Aronson&Lang, 2010](#)]
 - Extensive configuration option; but: default options tuned for biomedical literature, not necessarily websites or clinical text
 - Can be slow and unstable
- **QuickUMLS** [[Soldaini&Goharian, 2016](#)]
 - Modern computationally efficient mapper
 - Shown in the hands-on session
- **SemRep** — to extract relations between concepts [[Rindflesch&Fiszman, 2003](#)]
 - <subject, object, relation> from 27.9M PubMed articles stored into SemMedDB: <https://skr3.nlm.nih.gov/SemMedDB/>
 - Others exist: cTakes [[Savova et al., 2010](#)], Ontoserver [[McBride et al., 2012](#)], etc.

Concept Mapping as an IR problem

“...the patient had headaches and was home...”



(when retrieval methods are able to generate at least one mapping)

[Mirhosseini et al., 2014]

Practical - part 1

- In this hands-on session, we will:
 1. Take a collection of clinical trials, annotate them with medical concepts, producing documents with both term and concept representation.
- In part 2, we will use these results to:
 2. Index these documents in Elasticsearch with multi term/concepts fields.
 3. Search Elasticsearch with either term or concept, demonstrating semantic search capabilities.
 4. Play a bit more (maybe)
- Instructions: <https://ielab.io/russir2018-health-search-tutorial/hands-on/>

Implicit Medical Concept

Representations: Word Embeddings

- [Pyysalo et al., 2013]: word2vec and random indexing on very large corpus of biomedical scientific literature. <http://bio.nlplab.org>
- [De Vine et al., 2014]: word2vec on medical journal abstracts (embedding for UMLS)
 - Learns embedding of a concept, from co-occurrence with concepts
- [Zuccon et al., 2015, b]: word2vec on TREC Medical Records Track.
<http://zuccon.net/ntlm.html>
- [Choi et al., 2016]: word2vec on medical claims (embedding for ICD), clinical narratives (embedding for UMLS) <https://github.com/clinicalml/embeddings>
- [Beam et al., 2018]: cui2vec (variation of word2vec) on 60M insurance claims + 20M health records + 1.7M full text biomedical articles.
<https://figshare.com/s/00d69861786cd0156d81>
- Nuances of medical word embeddings:
 - [Chiu et al., 2016]: bigger corpora do not necessarily produce better biomedical word embeddings

Concept-based IR

Two types for Concept-based Retrieval

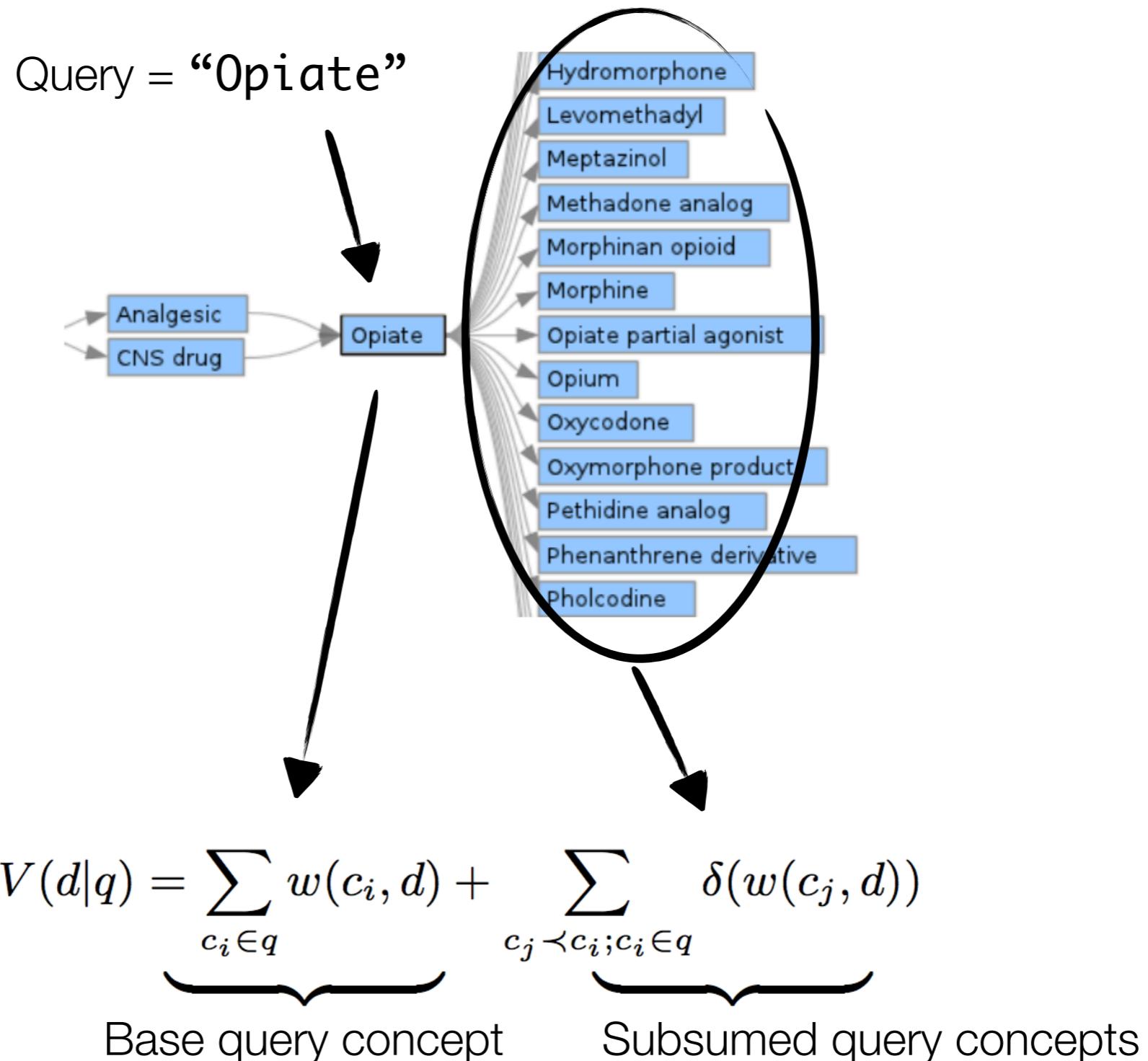
- Concept Augmented Term-based Retrieval
 - e.g. [[Ravindran&Gauch, 2004](#)]
 - Maintain the original term representation of documents.
 - Use a concept-based approach to improve the query representation.
- Pure Concept-based Retrieval
 - Map the terms in documents to higher-level concepts
 - Retrieval is then done in ‘concept space’ rather than ‘term space’
 - SAPHIRE system [[Hersh&Hickam, 1995](#)]
 - Language modelling concepts [[Meij et al., 2010](#)]

Combining Text and Concept Representations

[Limsopatham et al., 2013c]: learning framework that combines bag-of-words and bag-of-concepts representations on per-query basis

1. Linear combination model for merging scores from the two representations
2. Features: QPPs for both representations
3. Regression to infer model parameters (Gradient Boosted Regression Trees)

Exploiting concept hierarchies



Semantic Inference for IR

Concept-based retrieval that exploits ontology relationships

- Inferring conceptual relationships [[Limsopatham et al., 2013](#)]
- Information Retrieval as Semantic Inference [[Koopman et al., 2016](#)]
- both: expand queries by inferring additional conceptual relationships from KB, but in different ways
- [[Limsopatham et al., 2013](#)] also infers relationships
 - from collection of medical free-text, and
 - via PRF

“This is a 62-year-old gentleman who has Type I DM and is on hemodialysis. He is currently taking Avapro”

“This is a 62-year-old gentleman who has Type I DM and is on hemodialysis. He is currently taking Avapro”

- Hemodialysis ✓

“This is a 62-year-old gentleman who has Type I DM and is on hemodialysis. He is currently taking Avapro”

- Hemodialysis ✓
- DM? Diabetes mellitus?

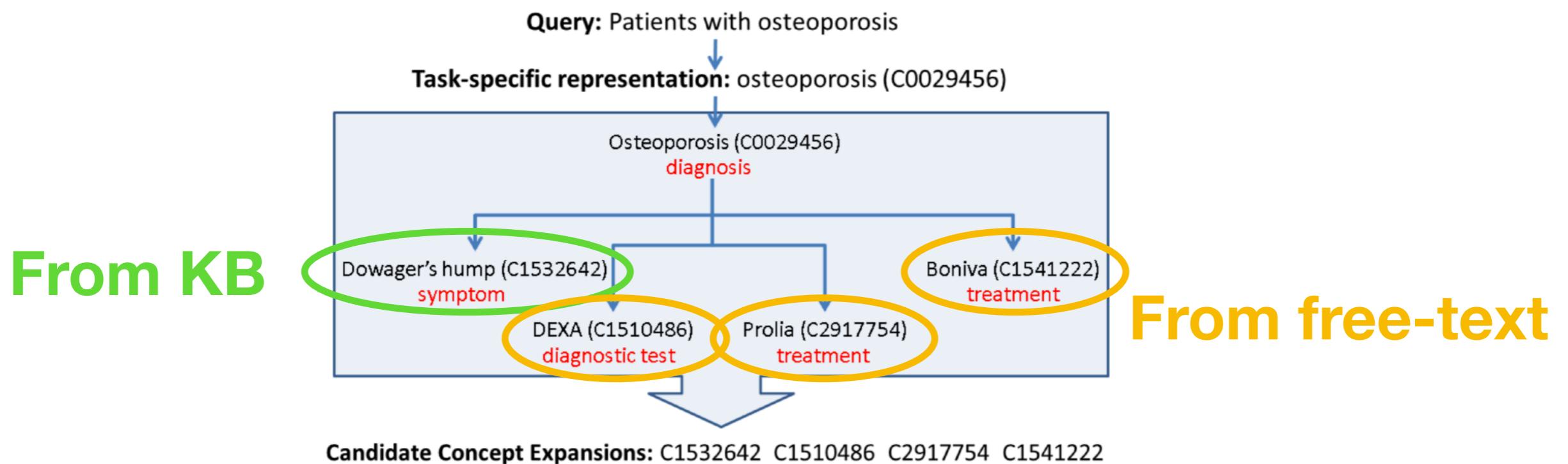
“This is a 62-year-old gentleman who has Type I DM and is on hemodialysis. He is currently taking Avapro”

- Hemodialysis ✓
- DM? Diabetes mellitus?
- Avapro? Hypertension!

Inferring conceptual relationships

[Limsopatham et al., 2013]

- For KB: use semantic relationships of concepts to represent the relationships between concepts.
- For free-text: MetaMap to identify concepts from the free-text, then infer relationships by co-occurrence/association rules



[Koopman et al., 2016]

q
“Patients with diabetes
and renal failure”

d

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

[Koopman et al., 2016]

q
“Patients with diabetes
and renal failure”

d

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

$$P(d|q) = 0$$

[Koopman et al., 2016]

q
“Patients with diabetes
and renal failure”

d

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

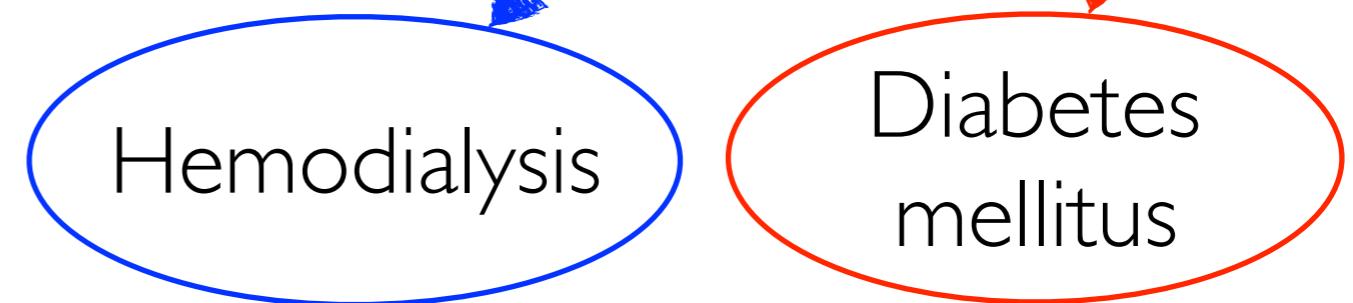
Graph Inference Model

[Koopman et al., 2016]

d

“Patients with diabetes
and renal failure”

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

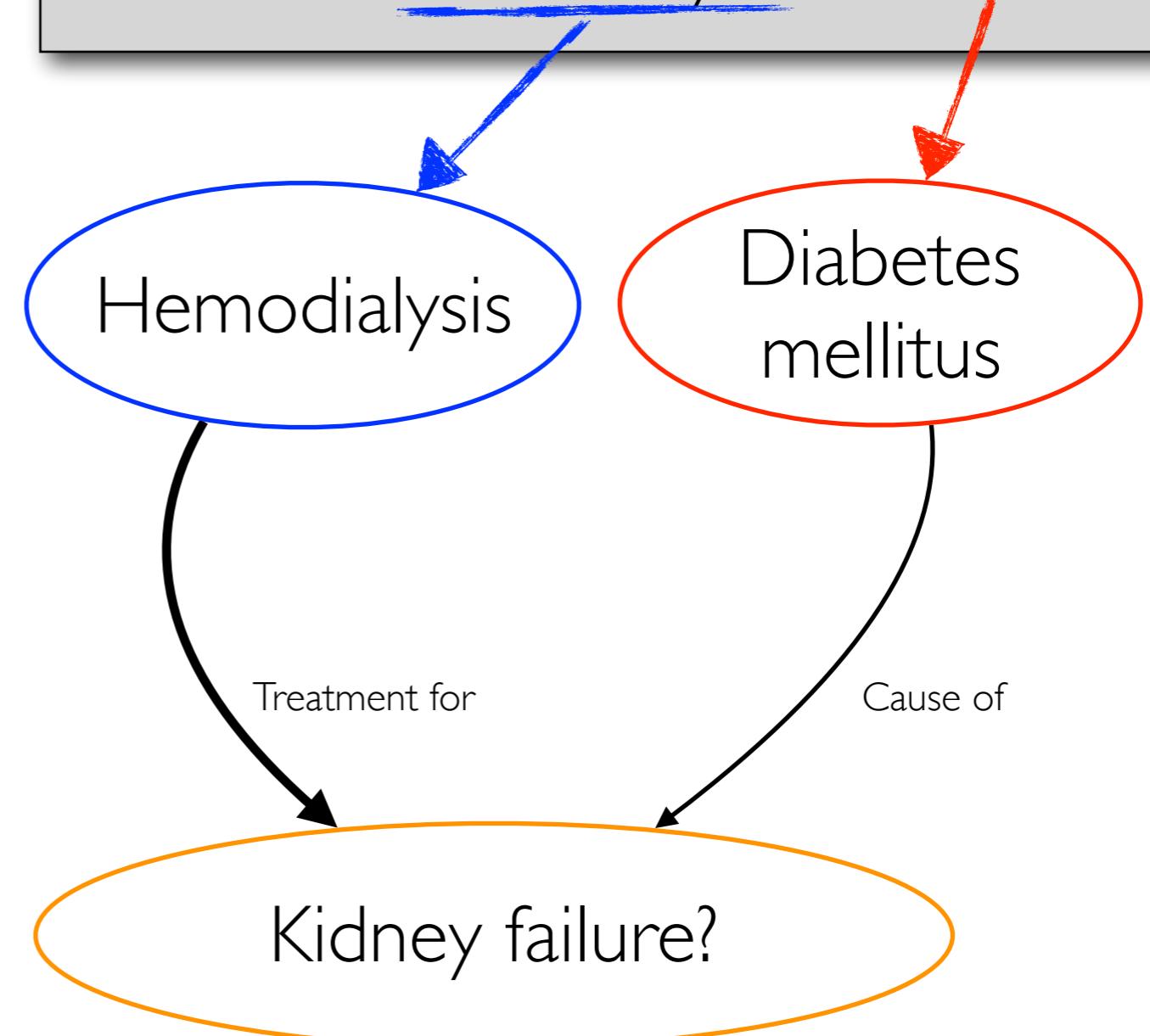


[Koopman et al., 2016]

q “Patients with diabetes and renal failure”

d

“This is a 62-year-old gentleman who has history of Type I DM and is on hemodialysis.”

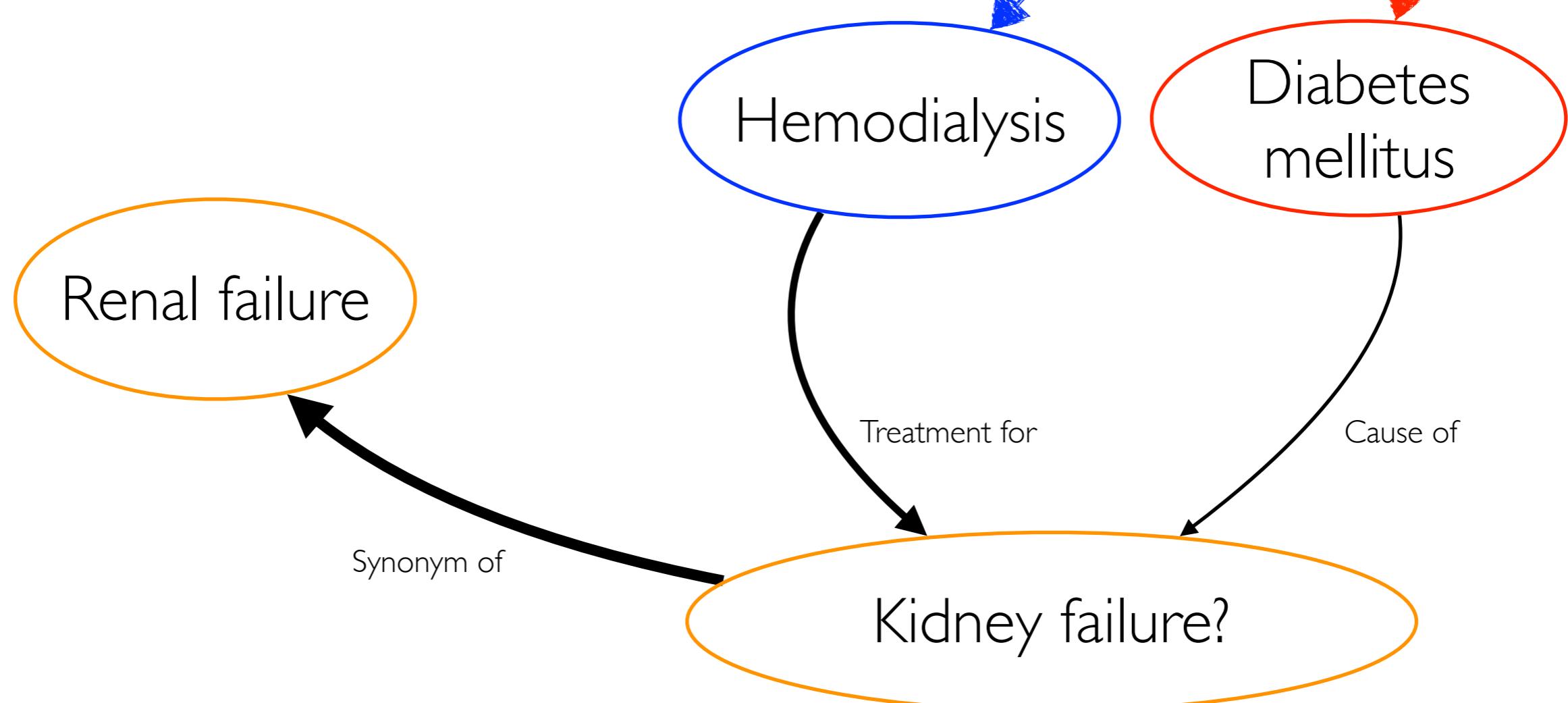


[Koopman et al., 2016]

q
“Patients with diabetes
and renal failure”

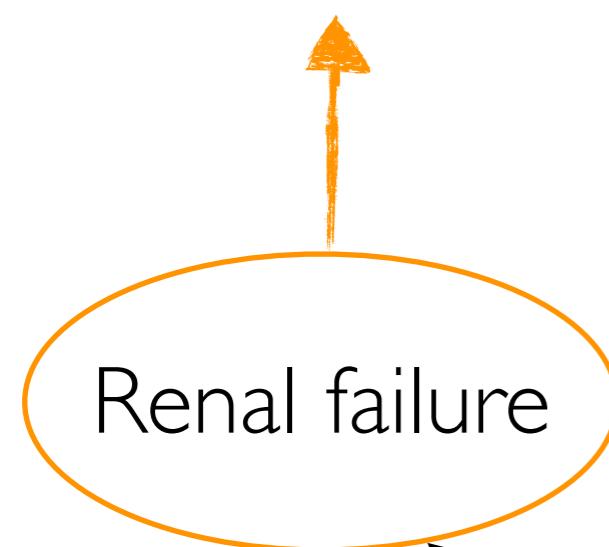
d

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

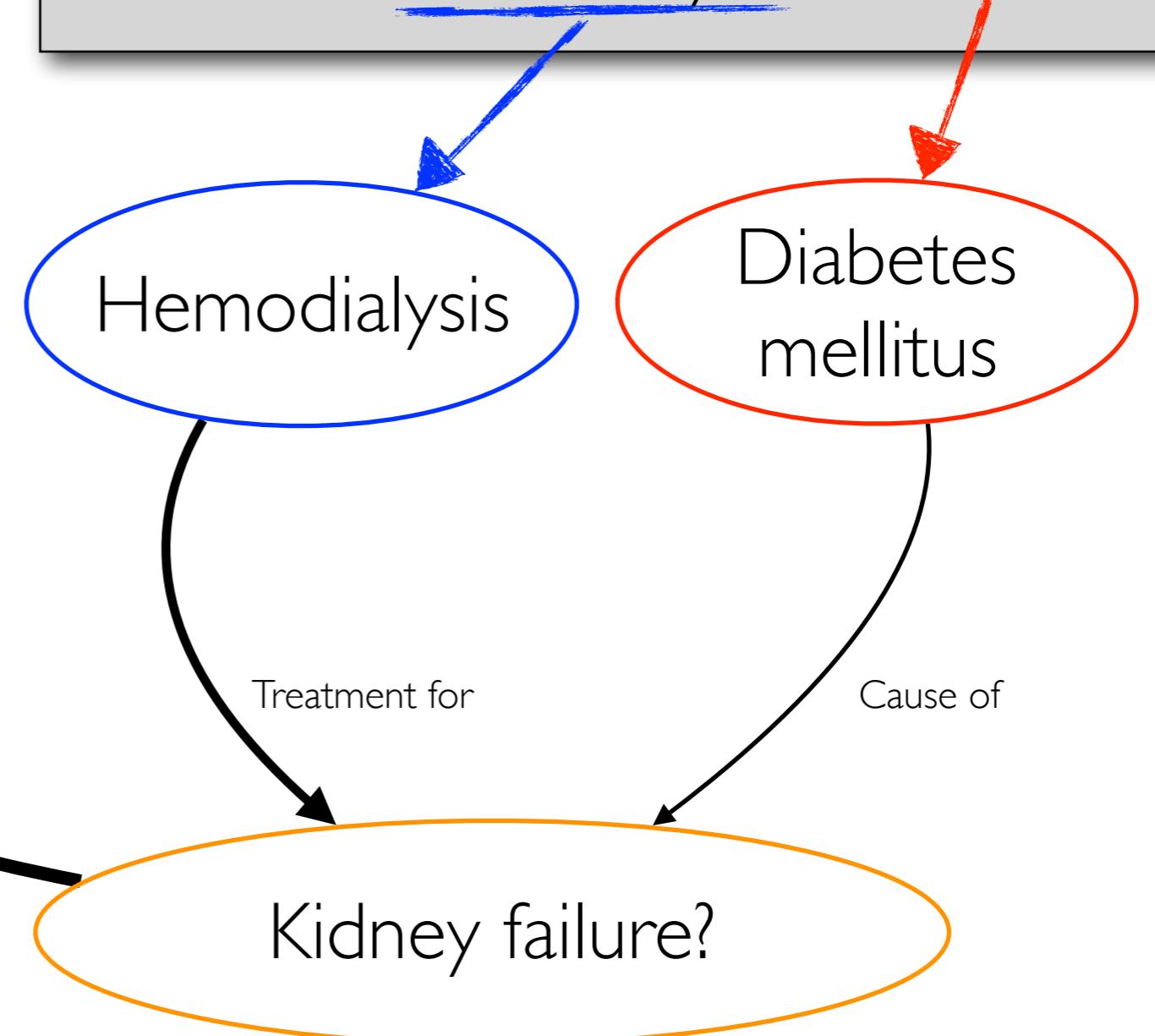


[Koopman et al., 2016]

q “Patients with diabetes and renal failure”



d “This is a 62-year-old gentleman who has history of Type I DM and is on hemodialysis.”

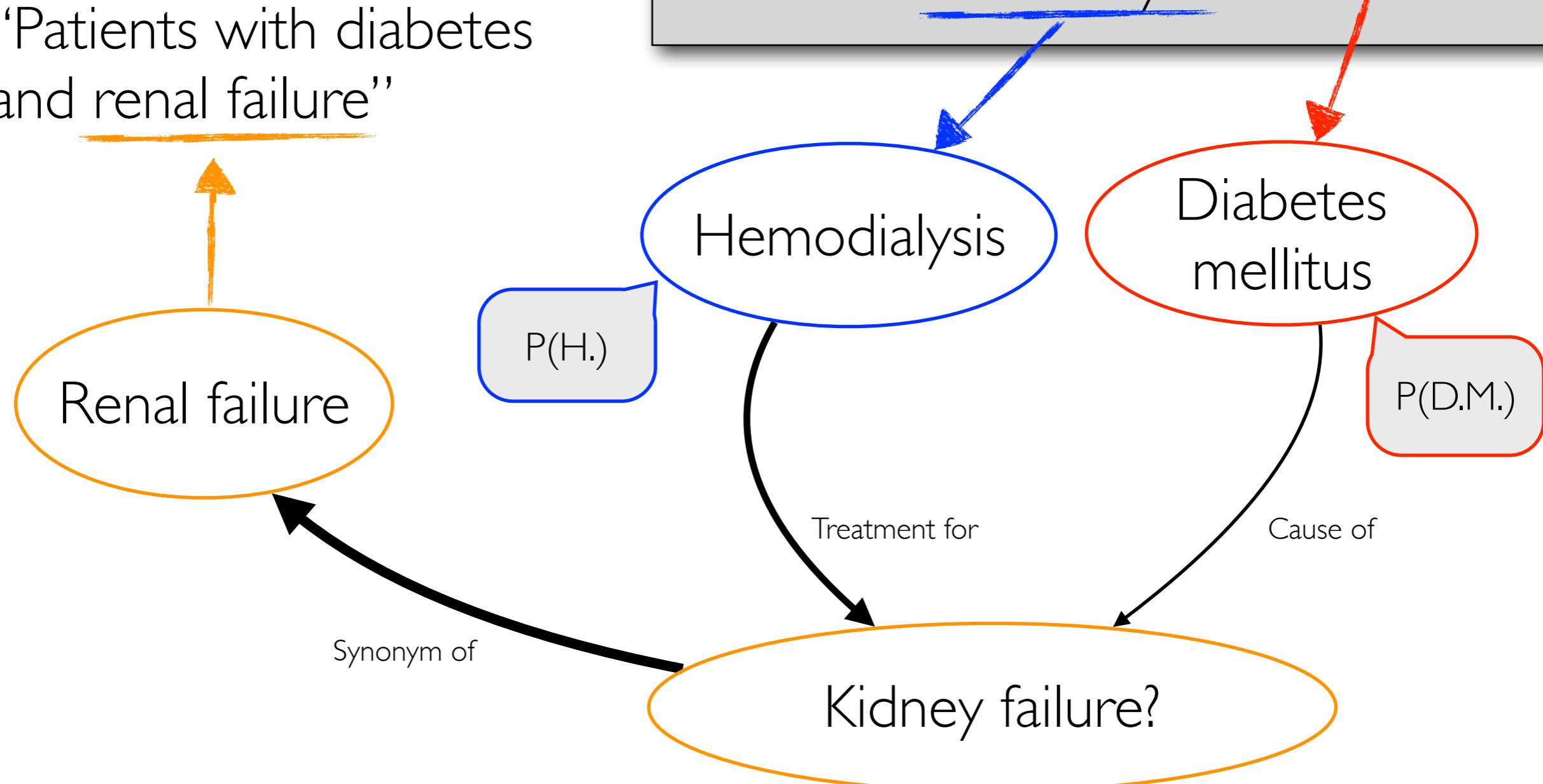


[Koopman et al., 2016]

q “Patients with diabetes and renal failure”

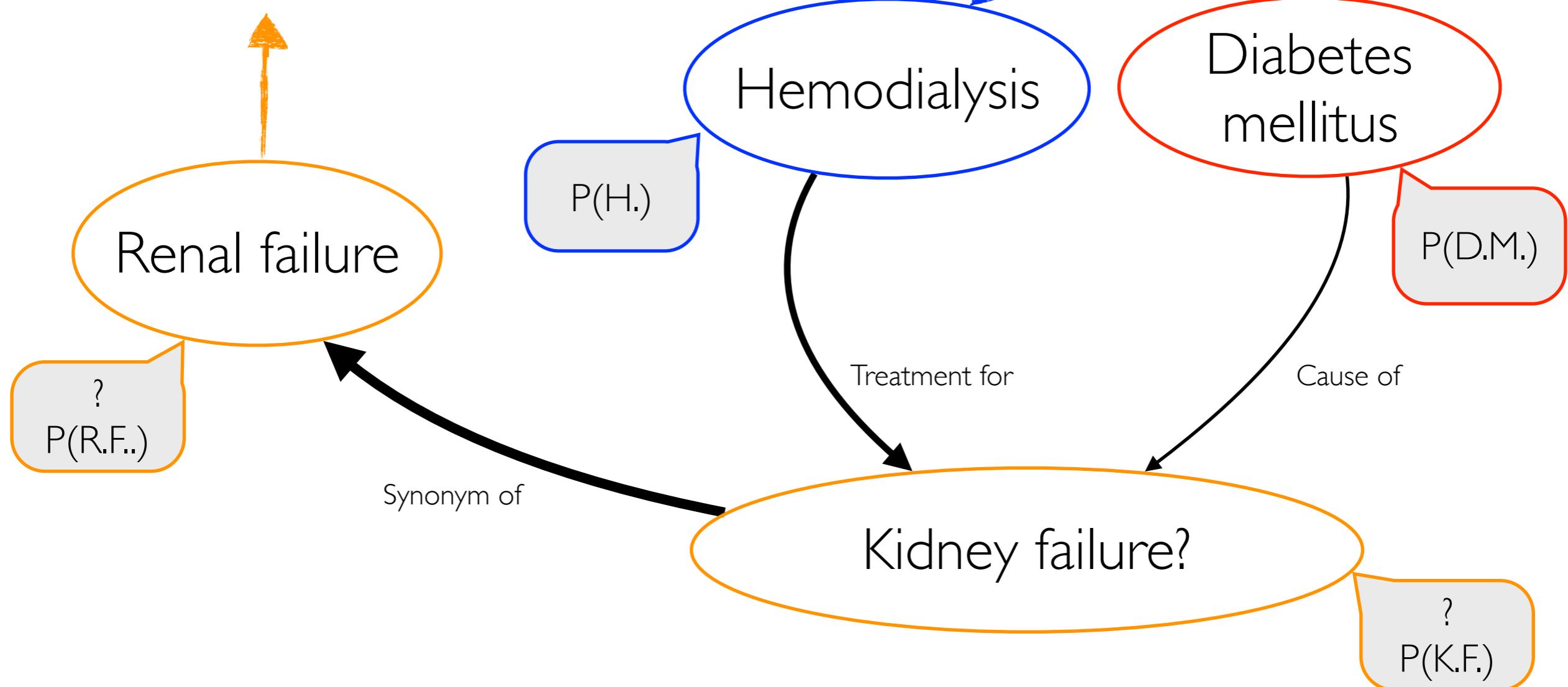
“Patients with diabetes and renal failure”

d “This is a 62-year-old gentleman who has history of Type I DM and is on hemodialysis.”



[Koopman et al., 2016]

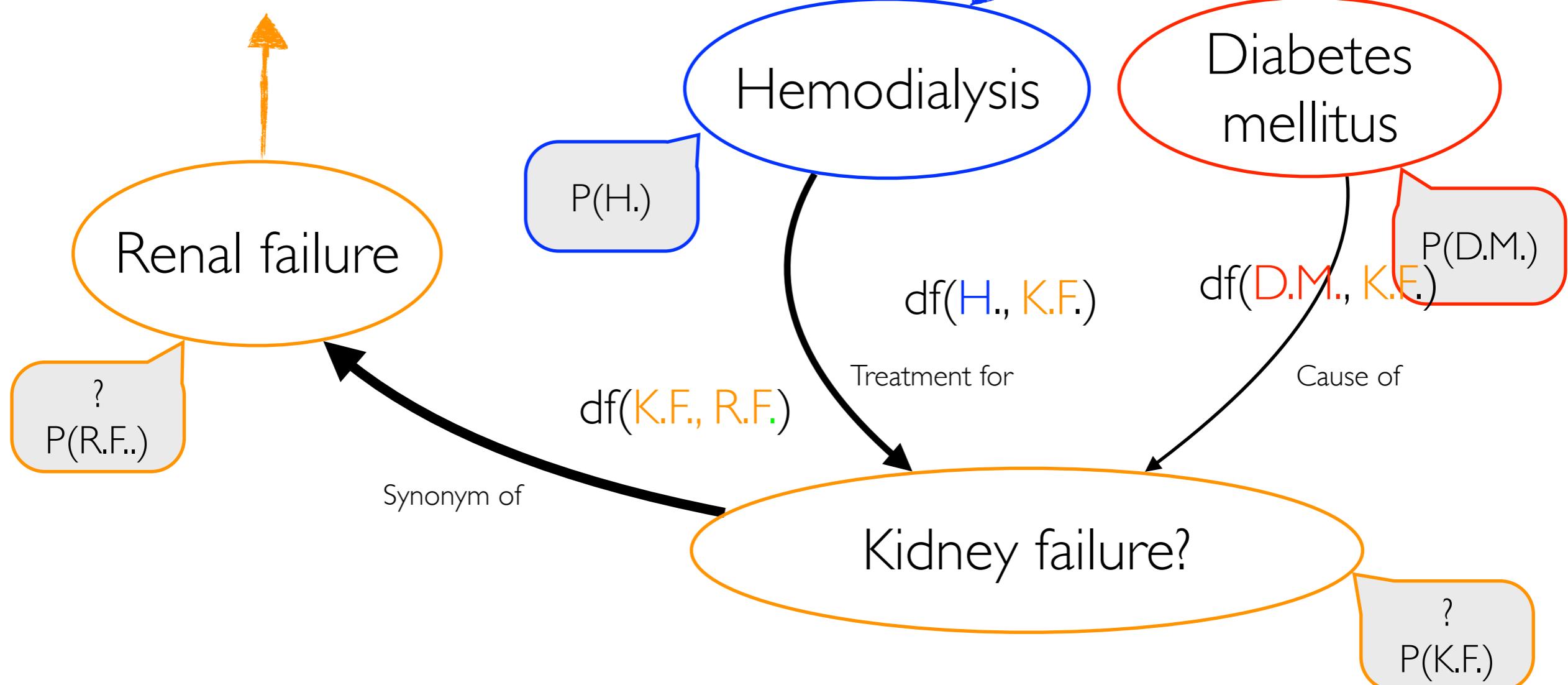
q
“Patients with diabetes
and renal failure”



d
“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

[Koopman et al., 2016]

q
“Patients with diabetes
and renal failure”



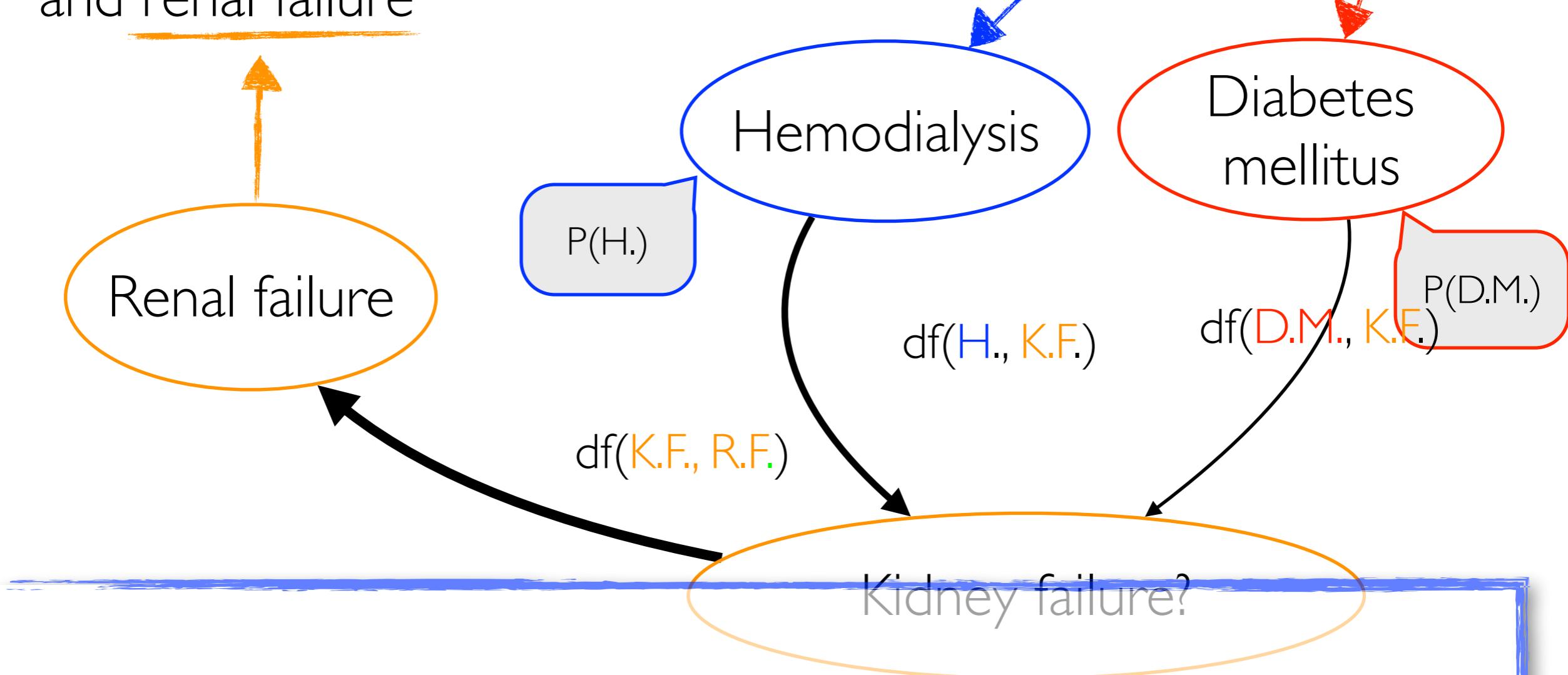
d
“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

[Koopman et al., 2016]

q

“Patients with diabetes
and renal failure”

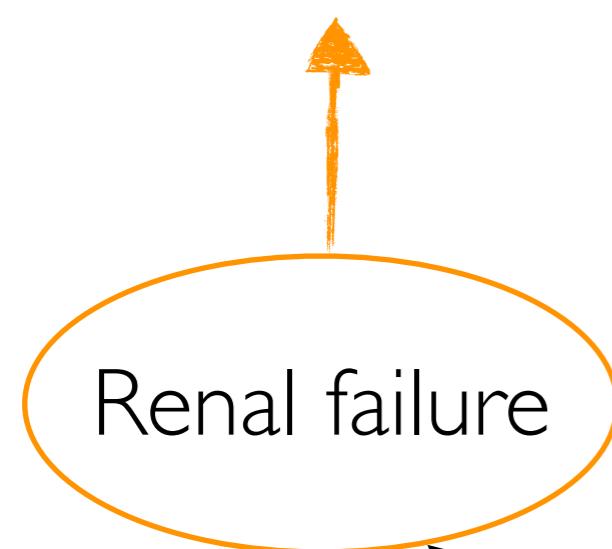
d
“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”



[Koopman et al., 2016]

q

“Patients with diabetes and renal failure”



$$P(d|q) = 0$$

d

“This is a 62-year-old gentleman who has history of Type I DM and is on hemodialysis.”



P(H.)

df(H., K.F.)



P(D.M.)

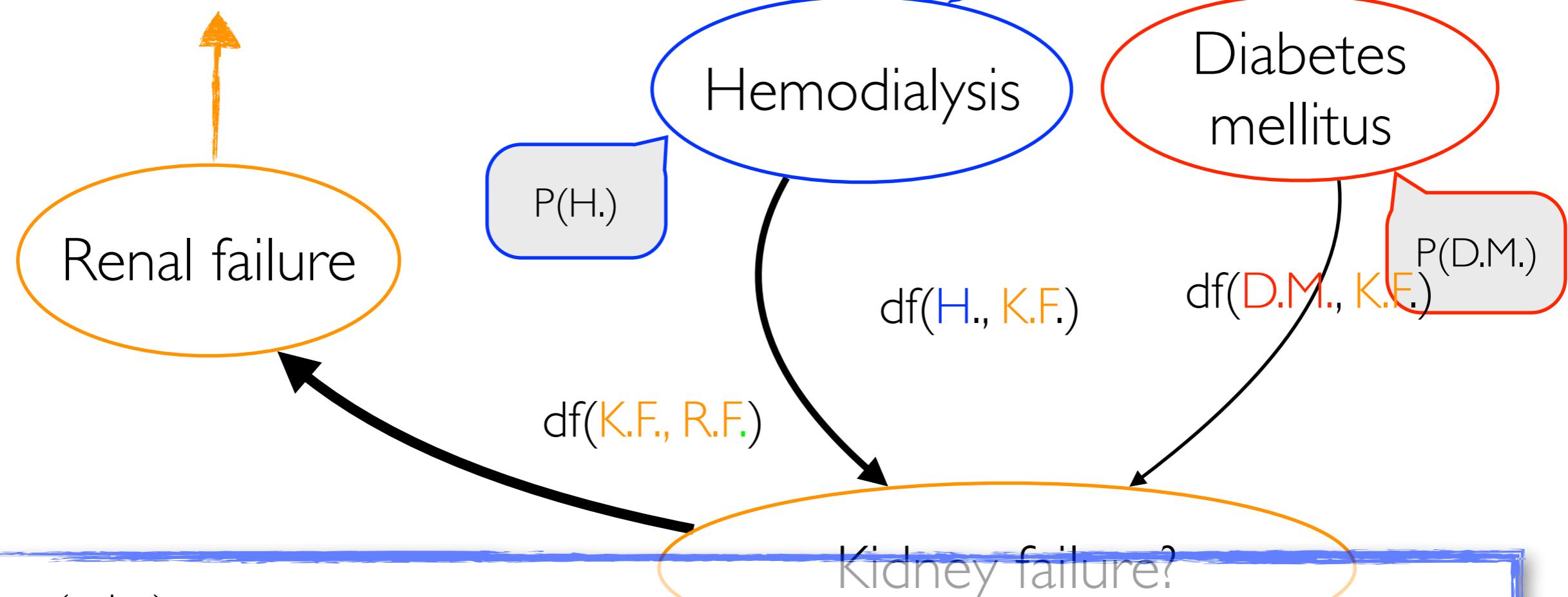
df(D.M., K.F.)

Kidney failure?

[Koopman et al., 2016]

q

“Patients with diabetes
and renal failure”



$$P(d|q) = 0$$

$$P(d \rightarrow q)$$

d

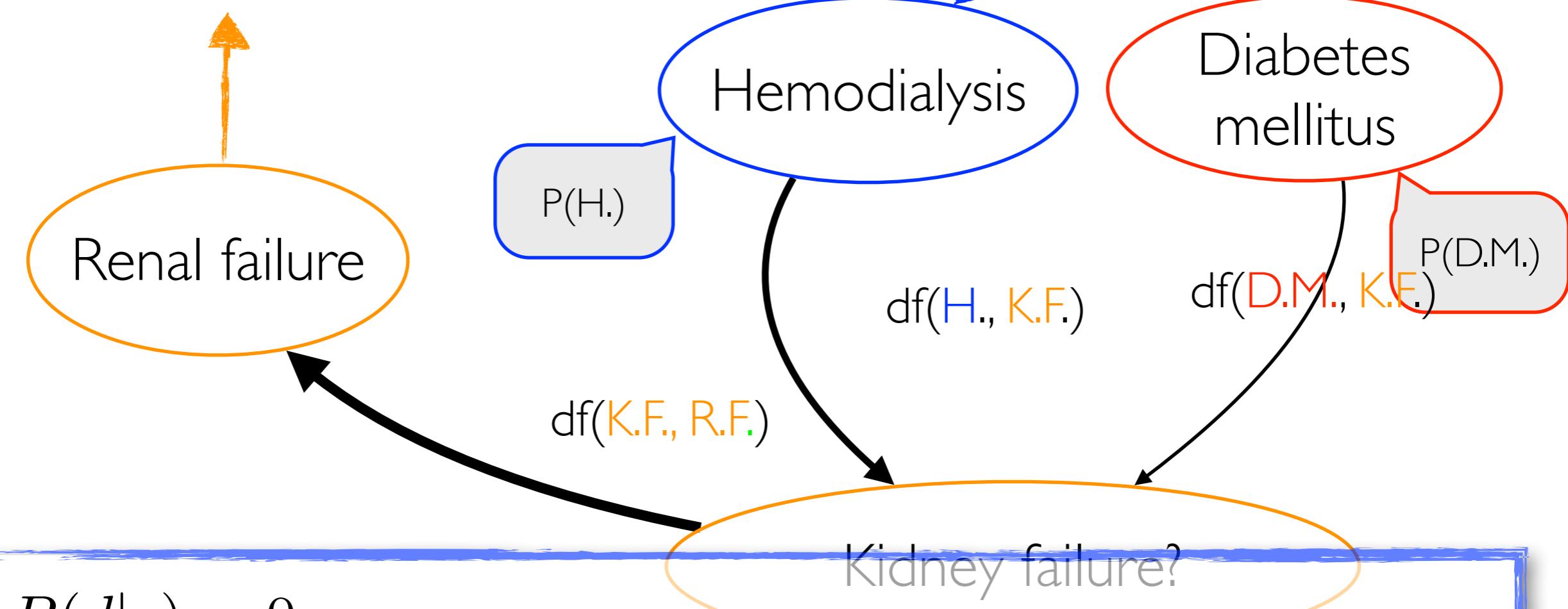
“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”

[Koopman et al., 2016]

d

q
“Patients with diabetes
and renal failure”

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”



$$P(d|q) = 0$$

$$P(d \rightarrow q) \approx P(D.M.) * df(D.M., K.F.)$$

[Koopman et al., 2016]

q

“Patients with diabetes
and renal failure”



d

“This is a 62-year-old gentleman
who has history of Type I DM
and is on hemodialysis.”



P(H.)

df(H., K.F.)

P(D.M.)

df(D.M., K.F.)

df(K.F., R.F.)

Kidney failure?

$$P(d|q) = 0$$

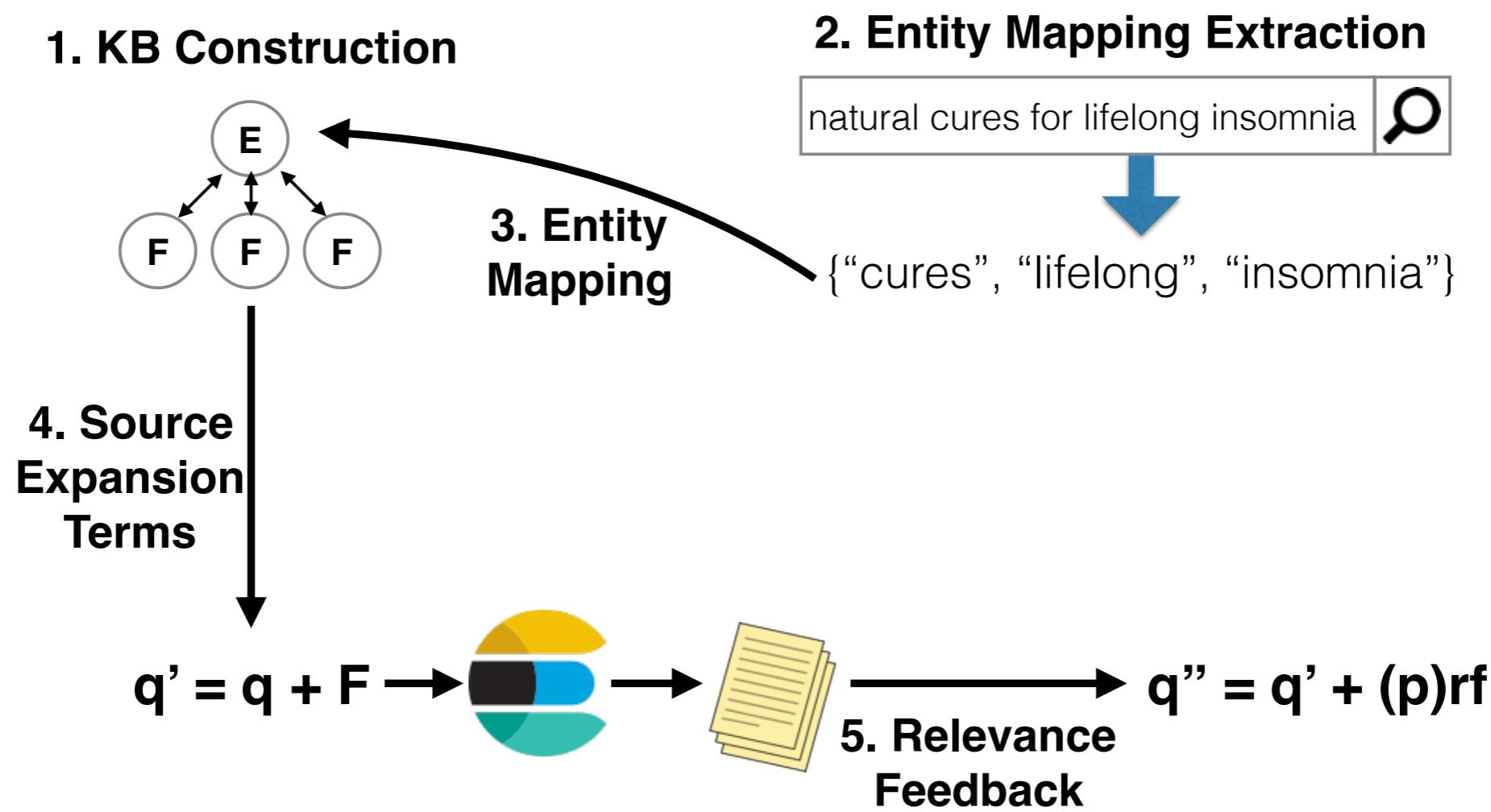
$$P(d \rightarrow q) \approx P(D.M.) * df(D.M., K.F.) + P(H.) * df(H., K.F.)$$

Practical - part 2

- Let's resume from where we left in part 1, and let's do:
 1. Index these documents in Elasticsearch with multi term/concepts fields.
 2. Search Elasticsearch with either term or concept, demonstrating semantic search capabilities.
 3. Play a bit more (maybe)
- Instructions: <https://ielab.io/russir2018-health-search-tutorial/hands-on/>

Choices in KB Query Expansion

- Many other approaches to do inference over KB data
- [Jimmy et al., 2018] consider the Entity Query Feature Expansion model [Dalton et al., 2014] and the influence settings choices have



Choices in KB Query Expansion

Findings for CHS

- For CHS, EQFE based on **UMLS** is more effective than based on Wikipedia.
 - Choice 1: Index **all UMLS concepts**
 - Choice 2: Use **all uni-, bi-, and tri-grams** of the original **queries**
 - Choice 3: **Map** mentions to **UMLS aliases**
 - Choice 4: Source **expansion** from the **UMLS title**
 - Choice 5: Add **relevance feedback** terms