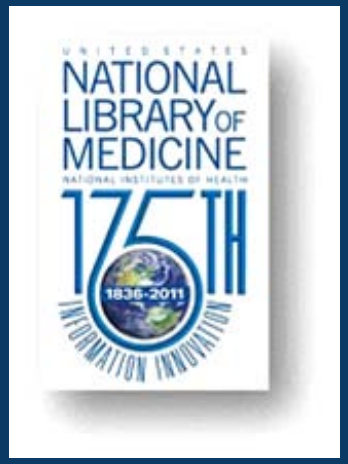


Knowledge-Based NLP in Biomedicine

Thomas C. Rindflesch, Ph.D., FACMI

Lister Hill National Center
for Biomedical Communications

Introduction

- Language complexity is daunting
  - Effective automatic processing must be based on characteristics of language
- Exploit (and adapt) existing resources
  - Linguistic
  - Ontology of the domain
- Abstract away from full interpretation
- Need principled solutions
  - For incremental development
  - For extensibility

Background

- Two general approaches
  - Emphasis on language
  - Emphasis on formalism and statistical methods
- Historical focus on one or the other
- Balanced methods
  - Best prospects for significant advances
- Case study
  - SemRep

SemRep

- Extract semantic predications from biomedical research literature (MEDLINE citations)
- Based on
  - Generalizations about the structure of English
  - Structured domain knowledge: UMLS
- Balances linguistic insight with implementation expediency
  - Underspecified syntax
  - Core predications only
  - Limited by domain

Theoretical framework

- Lexical semantics [Cruse; Pustejovsky]
- Ontological semantics [Nirenburg & Raskin]
  - "Ontology is the main metalanguage of meaning"
- Meaning-text theory [Mel'cuk]
- Dependency grammar [Tesnière; Sleator]

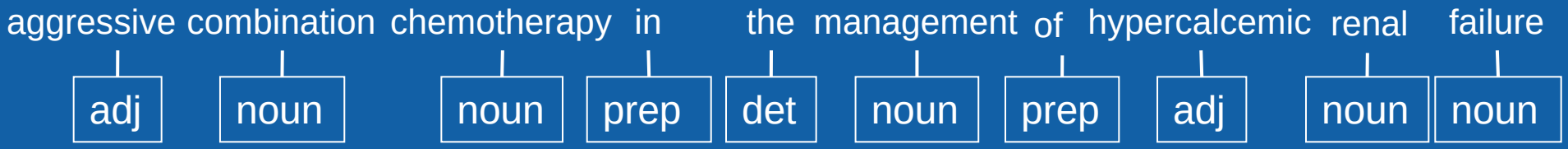
Syntactic phenomena

- Simple noun phrase
- NP – V – NP
- Nominalization
- Comparative (some)
- Relativization
- Argument coordination

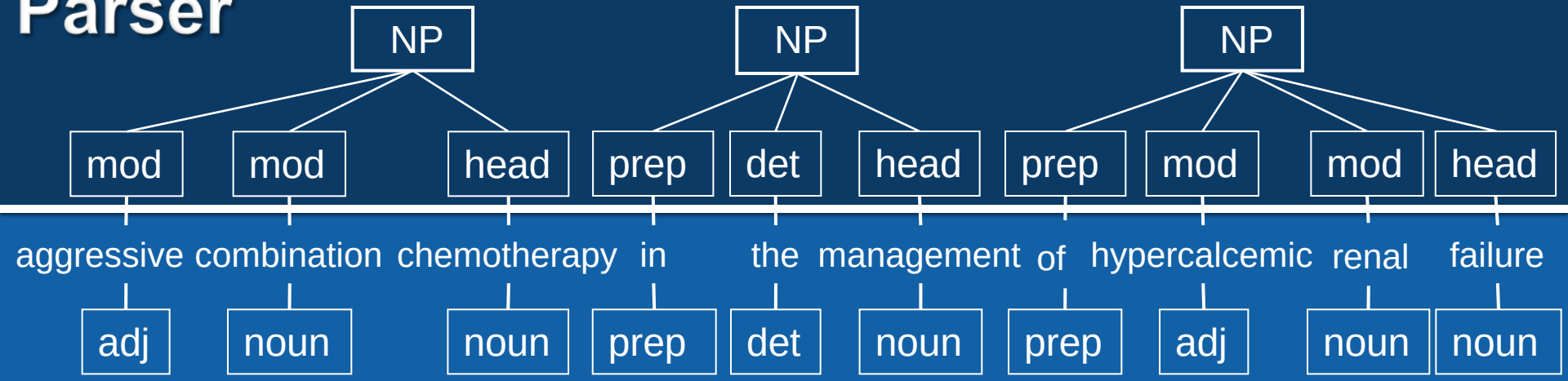
SemRep processing: Input text

aggressive combination chemotherapy in the management of hypercalcemic renal failure

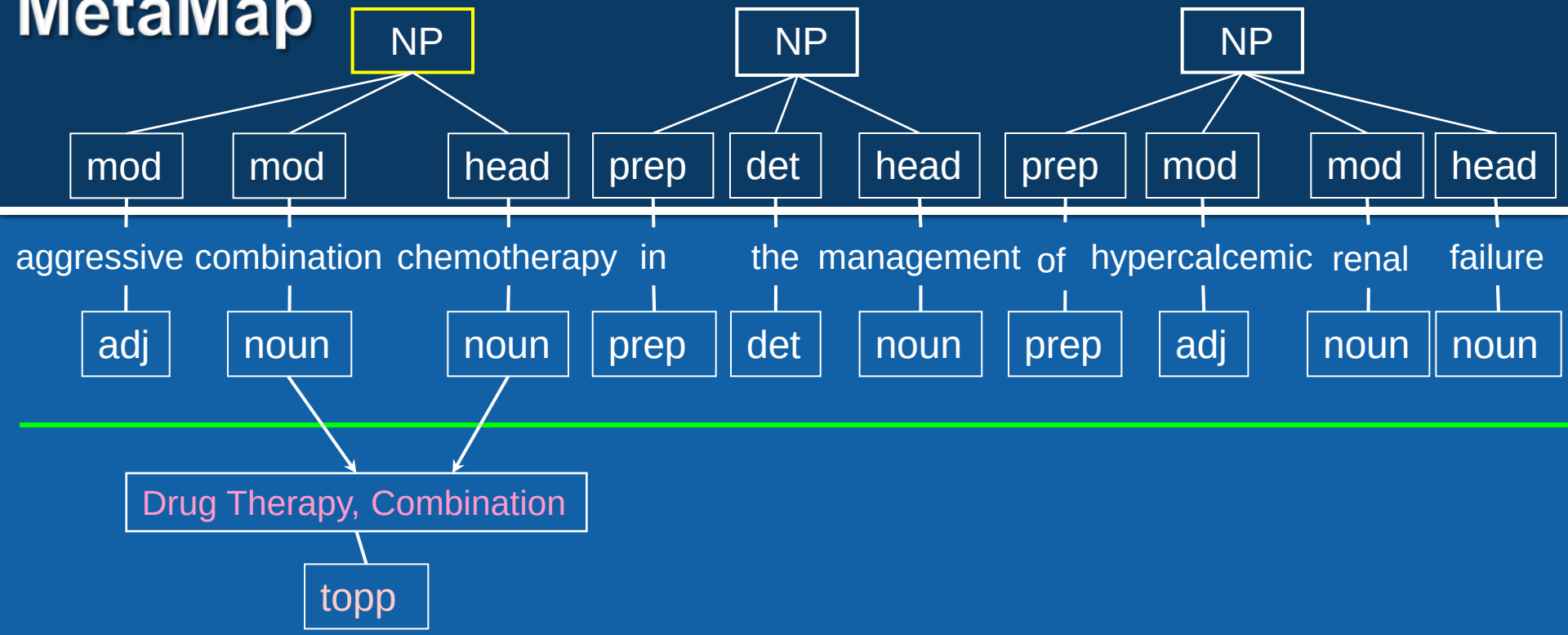
Lexical look-up and tagger



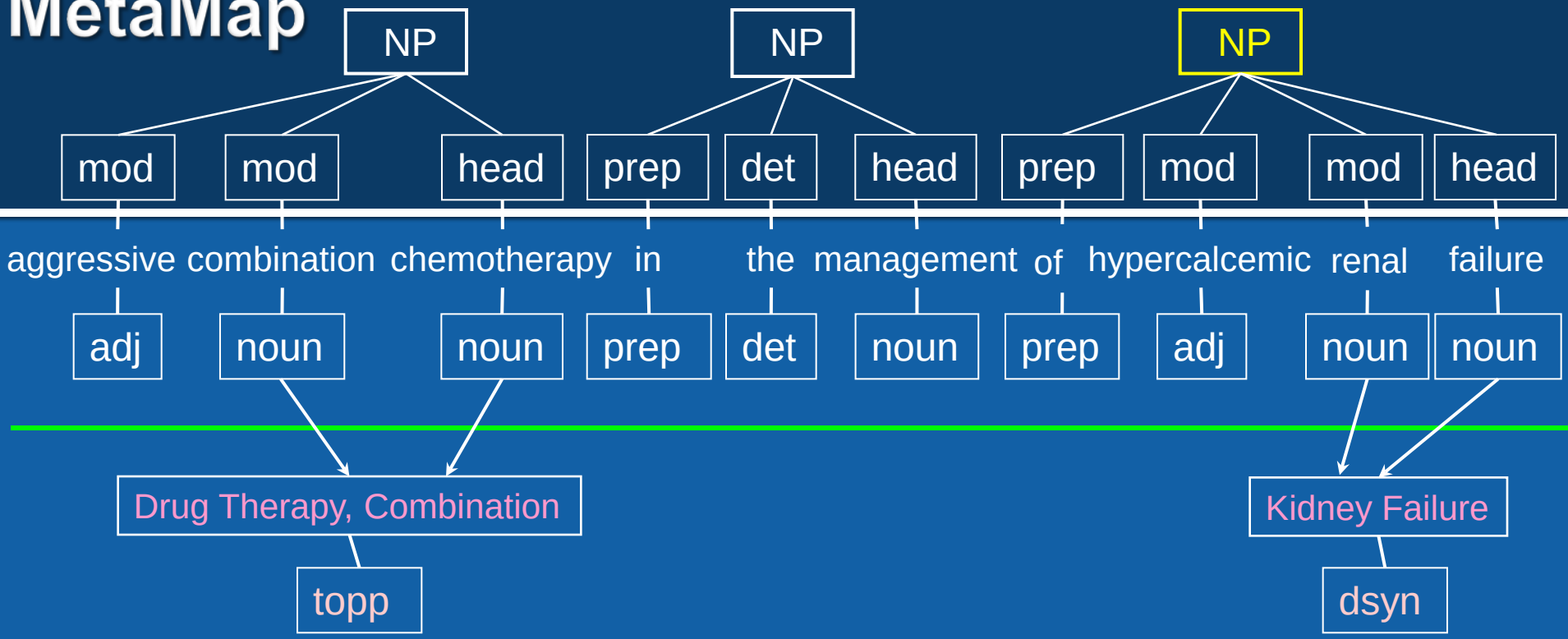
Parser


NP
NP
NP
mod
mod
head
prep
det
head
prep
mod
mod
head
aggressive
combination
chemotherapy
in
the
management
of
hypercalcemic
renal
failure
adj
noun
noun
prep
det
noun
prep
adj
noun
noun

MetaMap

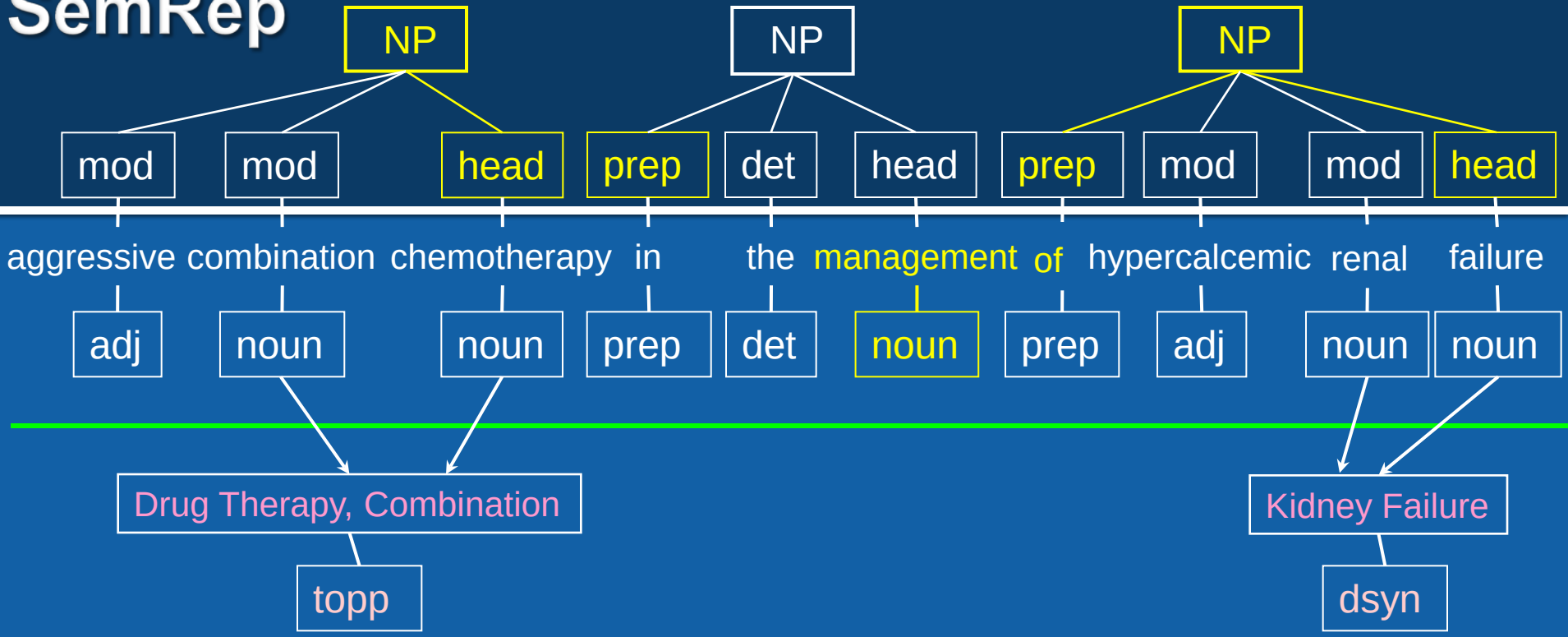


Therapeutic or Preventive Procedure



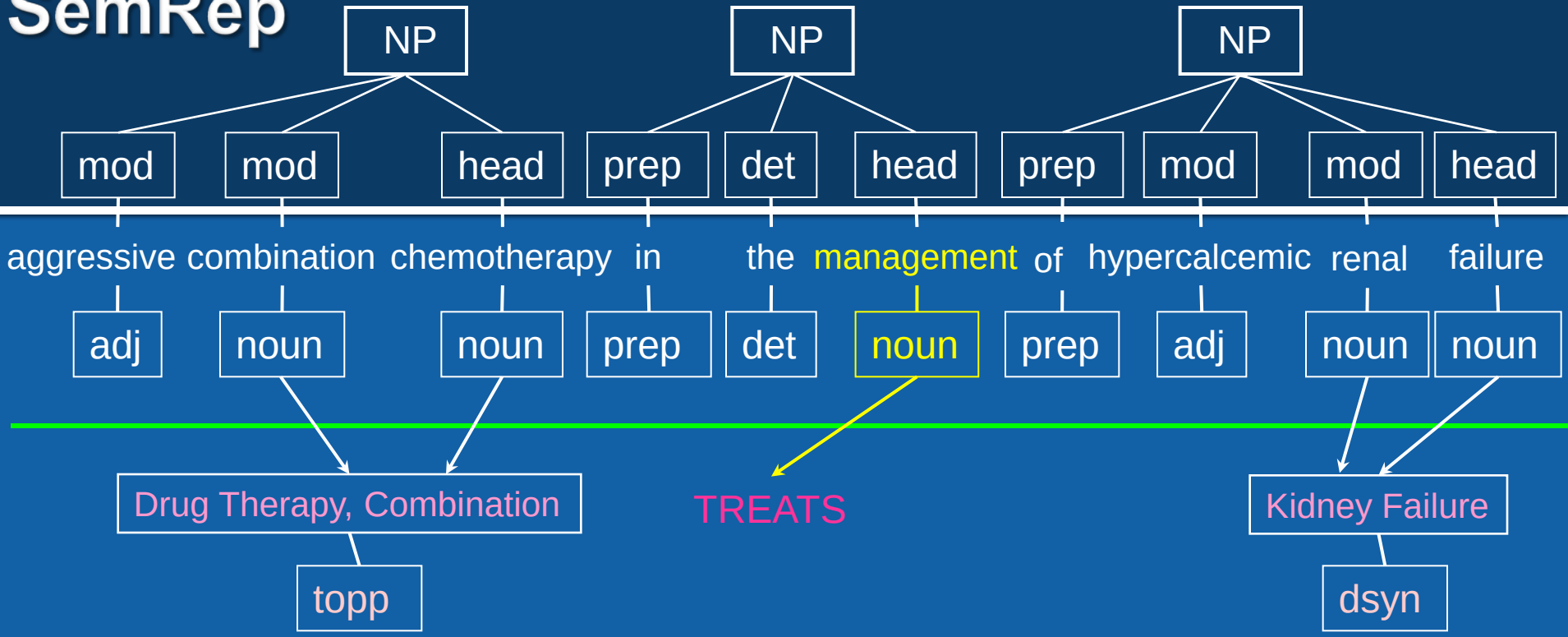
Disease or Syndrome

SemRep



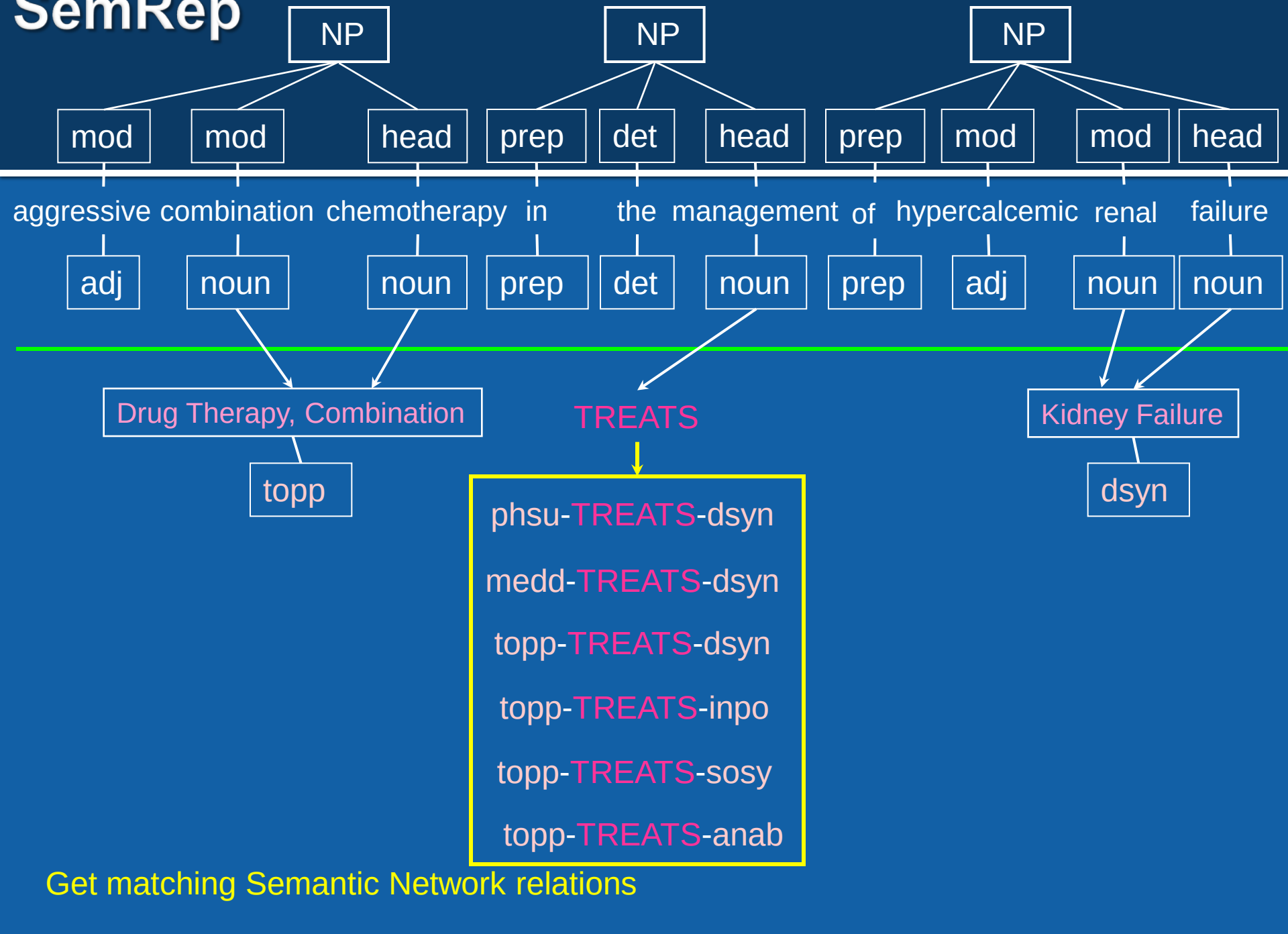
Apply syntactic argument constraints for nominalization

SemRep



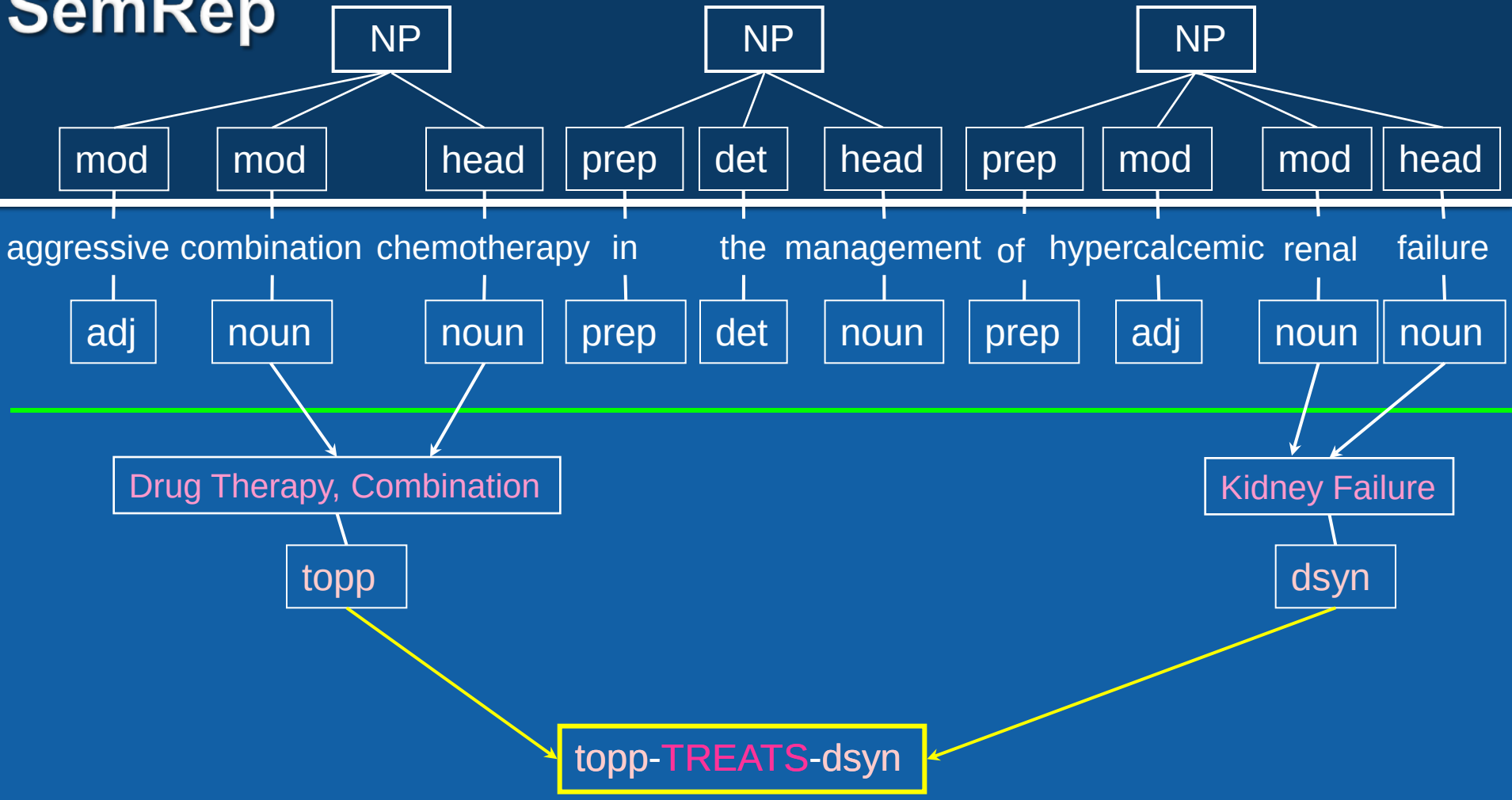
Indicator rule maps nominalization to Semantic Network predicate

SemRep



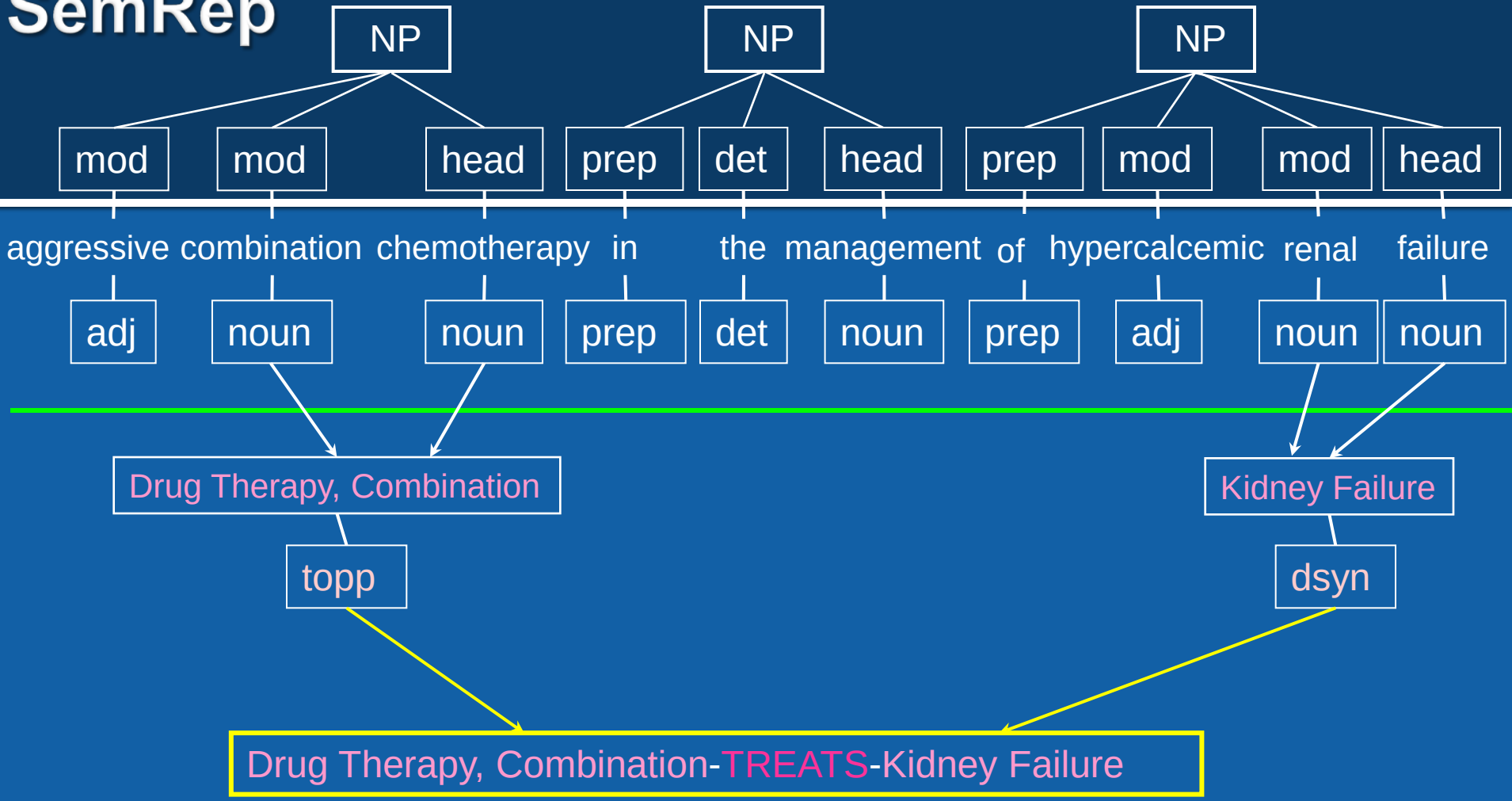
Get matching Semantic Network relations

SemRep



Match semantic types between arguments and Semantic Network

SemRep



Substitute concepts for semantic types

Predication database

- Processed all of MEDLINE
  - More than 21 million citations
  - Titles and abstracts
- SemRep predications extracted
  - 57 million predications (through 03/31/2012)
- Made available to the research community
  - SQL database
  - RDF triples

Domain coverage

- Initially developed for clinical medicine
- Extended to
  - Genetic etiology of disease
  - Substance interactions
  - Pharmacogenomics
- Working on
  - Influenza epidemic preparedness
  - Climate change and health
  - Health promotion
  - Biomedical knowledge processing

Several evaluations

- Focused on biomedical subdomains, e.g.
  - Clinical treatment, genetic etiology of disease, pharmacogenomics
- Focused on structure, e.g.
  - Hypernymic predications, comparatives, nominalizations
- Overall
  - Precision is around 75% (lower for molecular biology)
  - Recall is around 60%

Semantic MEDLINE

PubMed

MEDLINE citations

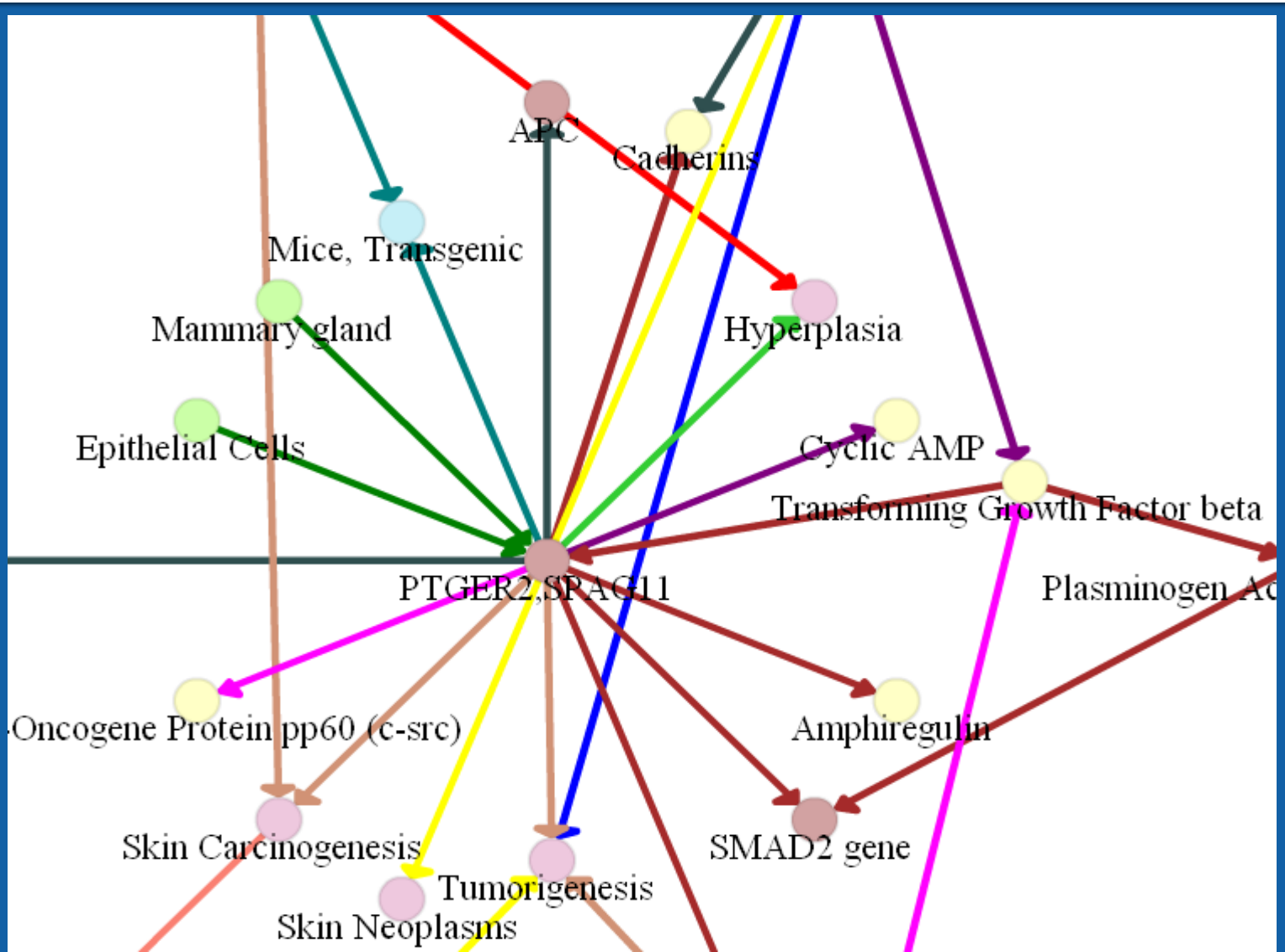
SemRep

Semantic predications

Automatic summarization

Graphical summary

Semantic MEDLINE: ep2 receptor



Literature-based discovery

- Hypogonadism degrades sleep in aging men [SLEEP 35(2):278-85]
- Melanin is crucial in the etiology of restless legs syndrome
- Effective drugs for obstructive sleep apnea
- Discovery browsing
  - Interaction of sleep, inflammation, and norepinephrine in depression
  - Caspase 3 in the development of induced pluripotent stem cells in adults

Information management

- Portfolio analysis of NIH grant applications
  - 2,175,737 predications extracted from 309,515 applications (2008 through 2011) [support from NIH/OD]
- Literature to support clinical practice guidelines
  - Adapted Semantic MEDLINE for production [support from NHLBI]
- In discussion for integration into Data.gov

Collaborations: using predications

- Dimitar Hristovski (U. Ljubljana, Slovenia)
  - Implemented system (BITOLA) for literature-based discovery
- Trevor Cohen (U. Texas-Houston)
  - Mathematical models for inferencing and discovery
- Guilherme Del Fiol (U. Utah)
  - Automatic generation of patient specific knowledge summaries to support clinical decision making
- Serguei Pakhomov (U. Minnesota)
  - Automatic semantic labeling of related terms in clinical text

Beyond propositional meaning

- SemRep focuses on propositional meaning
- Proposition-modifying information
  - Speculations, opinions, evidence, attitudes
- Discourse coherence
  - How is a coherent discourse formed from individual units of meaning?
- Core notions [Kilicoglu 2012]
  - Semantic embedding
  - Compositionality
  - Rule-based implementation

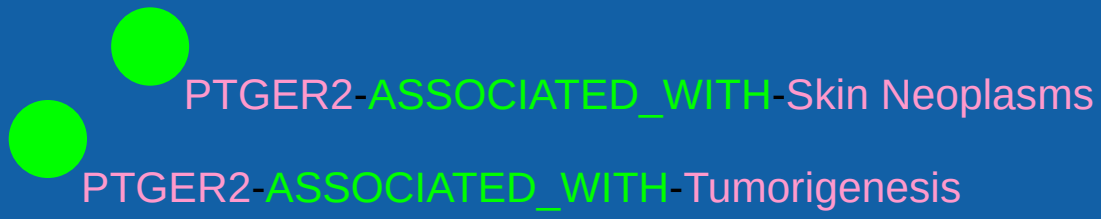
Example

It has been shown that EP2 **plays a critical role** in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.

PTGER2-ASSOCIATED_WITH-Tumorigenesis

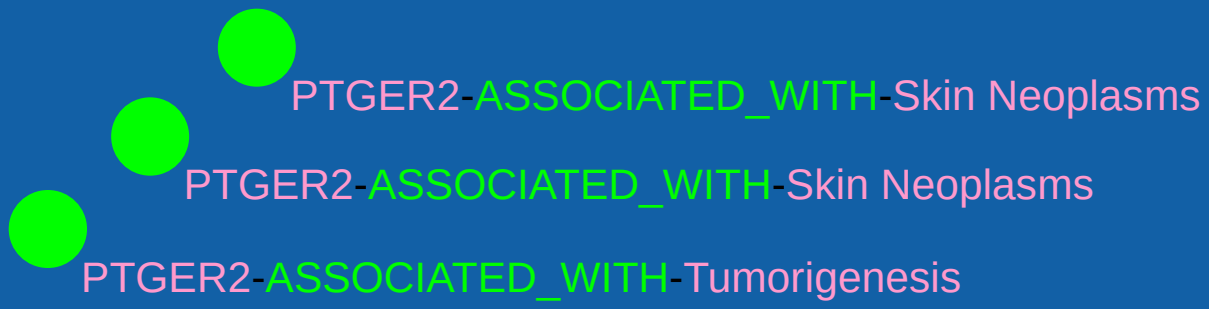
Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



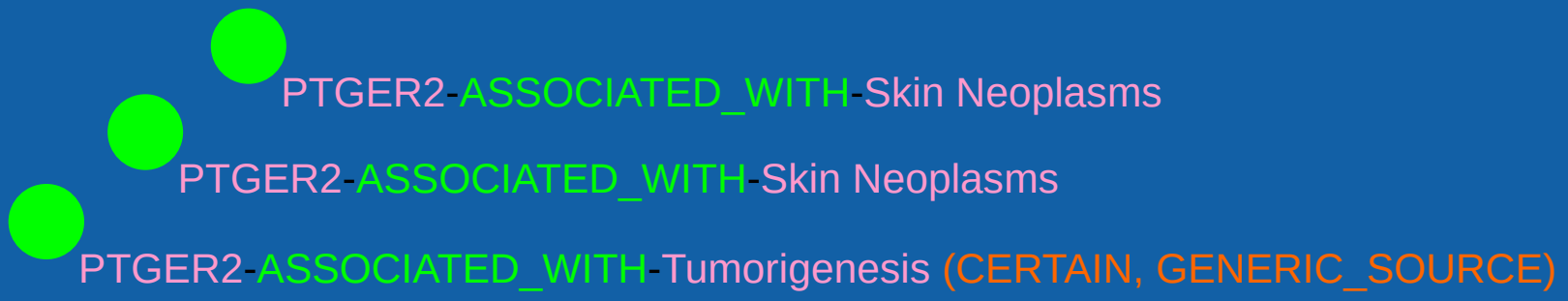
Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



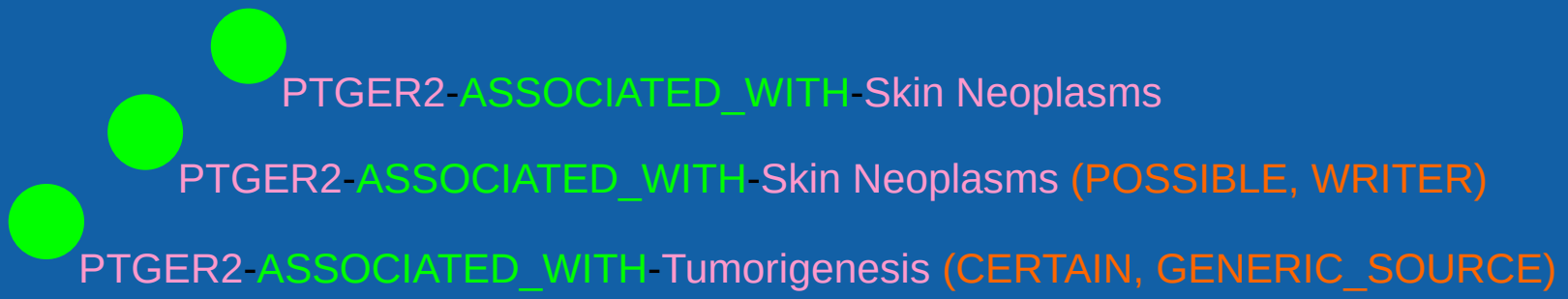
Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



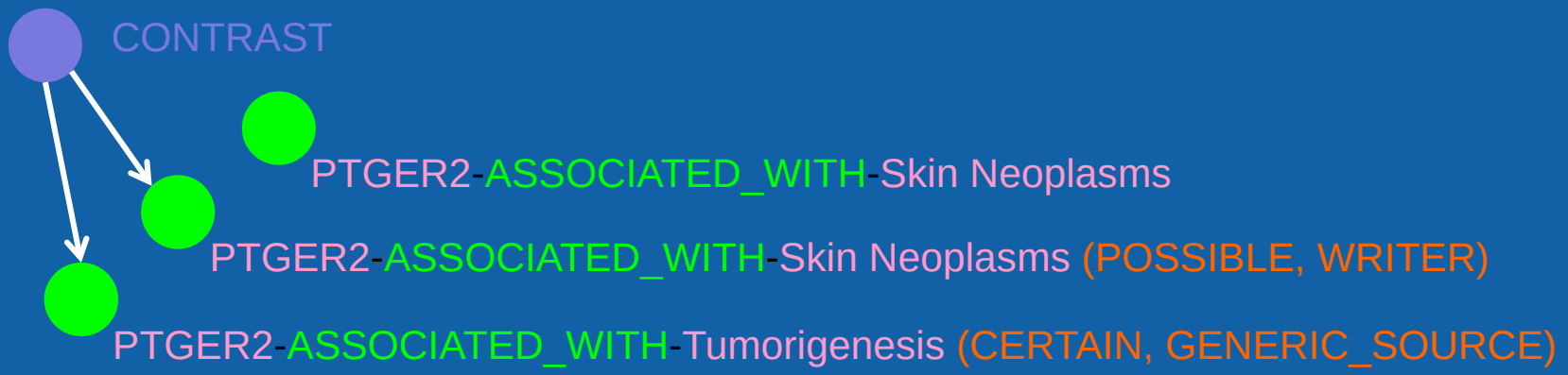
Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



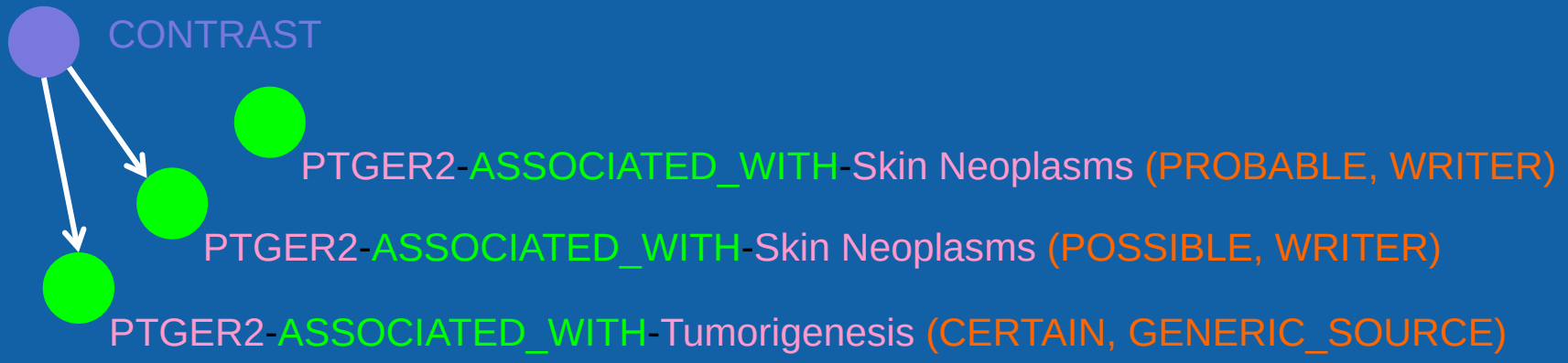
Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



Example

It has been shown that EP2 plays a critical role in tumorigenesis in mouse mammary gland and colon. However, the possibility that the EP2 receptor is involved in the development of skin tumors was unknown…Our data suggest that the EP2 receptor plays a significant role in the protumorigenic action of PGE2 in skin tumor development.



Significance

- Assess the reliability of a scientific claim
- Speculations are important in science
  - Establish current trends
  - Future directions
- Ability to track change in scientific knowledge
  - Ties in with the SemRep predication database
- Discourse coherence relations enhance semantic interpretation
  - Inter-sentential relations
  - Interact with coreference

Conclusion: SemRep

- Balanced NLP for biomedical research literature
- Based on linguistic and domain knowledge
- Principled for extensibility
- Supports incremental development
  - Recent extension to extra-propositional meaning
- Results exploited for biomedical research
- Cooperation of
  - Computer scientists
  - Linguists
  - Domain experts

Acknowledgments

- Michael J. Cairelli, D.O.
- Guocai Chen, Ph.D.
- Marcelo Fiszman, M.D., Ph.D.
- Halil Kilicoglu, Ph.D.
- Christopher M. Miller, M.D.
- Graciela Rosemblat, Ph.D.
- Susan Roy, M.S., M.L.S.
- Dongwook Shin, Ph.D.
- T. Elizabeth Workman, Ph.D.