

Entities, Relations, Events: Representing Biomolecular Semantics

Sampo Pyysalo

Tsujii lab, University of Tokyo

Workshop on Advances in Bio Text Mining (BioTM 2010)

May 10, 2010 Ghent, Belgium



Introduction

- “Biomolecular semantics”: the meaning of texts describing biomolecules (genes, proteins) and their associations

We previously reported that the role of reactive oxygen intermediates (ROIs) in NF-kappaB activation by proinflammatory cytokines was cell specific. However, the sources for ROIs in various cell types are yet to be determined and might include 5-lipoxygenase (5-LOX) and NADPH oxidase. 5-LOX and 5-LOX activating protein (FLAP) are coexpressed in lymphoid cells but not in monocytic or epithelial cells. Stimulation of lymphoid cells with interleukin-1beta (IL-1beta) led to ROI production and NF-kappaB activation, which could both be blocked by antioxidants or FLAP inhibitors, confirming that 5-LOX was the source of ROIs and was required for NF-kappaB activation in these cells. IL-1beta stimulation of epithelial cells did not generate any ROIs and NF-kappaB induction was not influenced by 5-LOX inhibitors. However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

- Semantic representation: a formalization of a way to capture (relevant aspects of) meaning



Introduction

- “Biomolecular semantics”: the meaning of texts describing biomolecules (genes, proteins) and their associations

We previously reported that the role of reactive oxygen intermediates (ROIs) in NF-kappaB activation by proinflammatory cytokines was cell specific. However, the sources for ROIs in various cell types are yet to be determined and might include 5-lipoxygenase (5-LOX) and NADPH oxidase. 5-LOX and NADPH oxidase are coexpressed in lymphoid cells but not in monocytes. IL-1beta stimulates 5-LOX activity in lymphoid cells with interleukin-1beta (IL-1b) **Activate (5-LOX, NF-kappaB)** NF-kappaB activation, which could both be blocked by antioxidants or FLAP inhibitors, confirming that 5-LOX was the source of ROIs and was required for NF-kappaB activation in these cells. IL-1beta stimulation of epithelial cells did not generate any ROIs and NF-kappaB induction was not influenced by 5-LOX inhibitors. However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

- Semantic representation: a formalization of a way to capture (relevant aspects of) meaning



Introduction

- ❑ Information extraction: identify pieces of information from text and represent them in structured form
 - ✓ Unstructured text → Structured data

- ❑ Text mining: automatically derive information relevant to user information needs from a body of text
 - ~ Information Extraction + Data Mining
 - ✓ Often involves a requirement of novelty and/or summarization of data across multiple documents

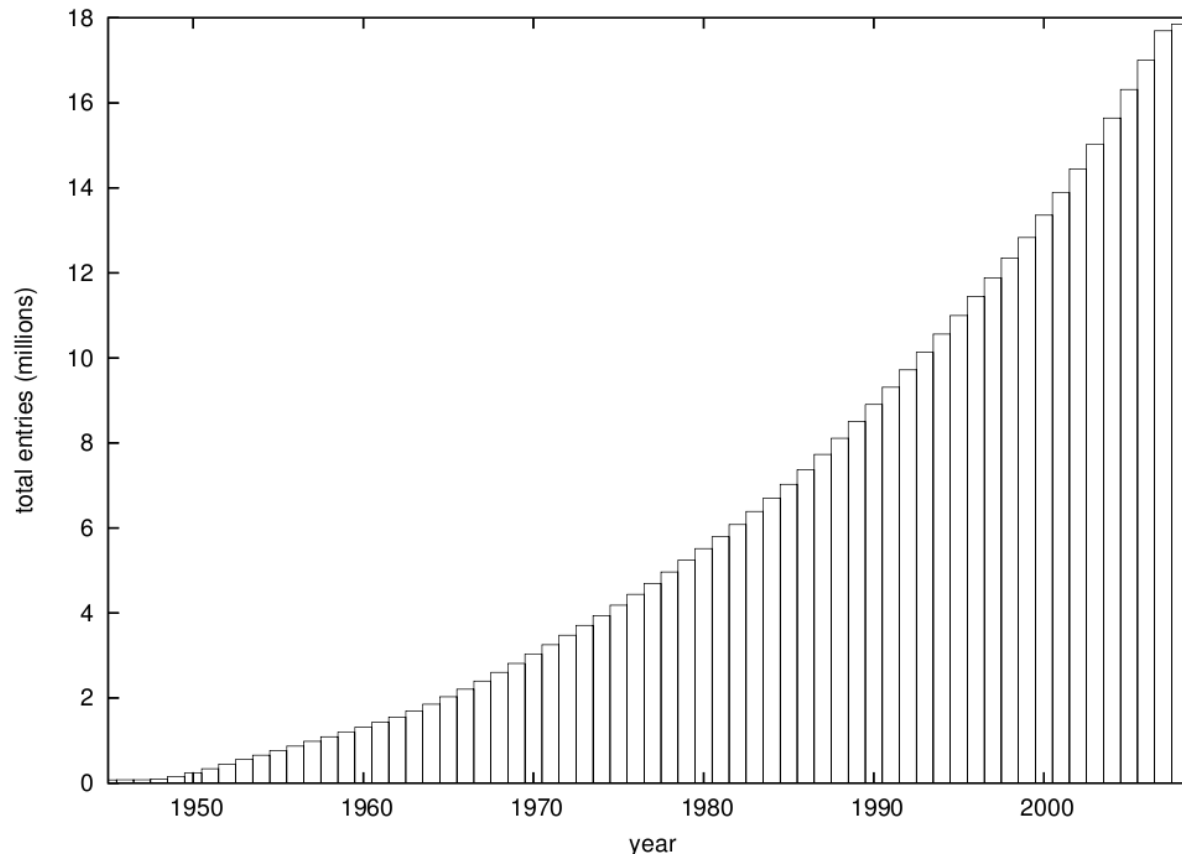


Outline

- Introduction and motivation
- Entities
- Relations
- Events
- Relations (again)
- Where next?

Motivation

- The biomedical literature is growing exponentially
 - ✓ Biology is also changing; the text mining targets are not fixed





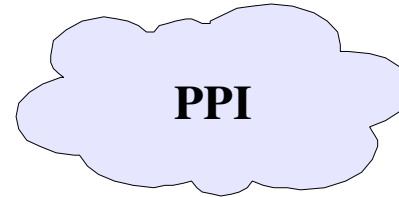
Motivation

- ❑ After more than a decade of concentrated efforts, biomolecular text mining has addressed basic tasks in detail
 - ✓ Dozens of systems proposed for e.g. protein name recognition and protein-protein interaction extraction
- ❑ ... but seen only limited success at advanced tasks
 - ✓ Great number of potential text mining targets
 - ✓ Limited resources and size of research community
- ❑ Focus largely on simple representations and tasks that can be modeled using such representations

- ❑ How can limited resources be used efficiently to address advanced biomedical text mining goals?

Motivation

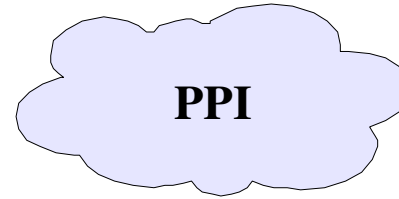
- Aiming for a text mining goal
 - ✓ e.g. “identify evidenced statements of novel protein-protein interactions from PubMed”



TM Goals

□ Aiming for a text mining goal

- ✓ e.g. “identify evidenced statements of novel protein-protein interactions from PubMed”
- ✓ Manual annotation-based approach: formalize goal, mark documents do identify the “correct” extraction result



TM Goals



Motivation

PPI

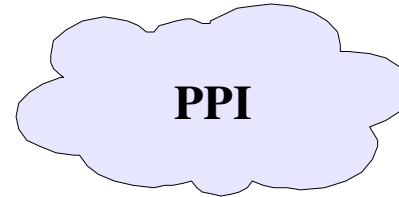
TM Goals

We previously reported that the role of reactive oxygen intermediates (ROIs) in NF-kappaB activation by proinflammatory cytokines was cell specific. However, the sources for ROIs in various cell types are yet to be determined and might include 5-lipoxygenase (5-LOX) and NADPH oxidase. 5-LOX and 5-LOX activating protein (FLAP) are coexpressed. Stimulation of lymphoid cells with interleukin-1beta (IL-1beta) led to ROI production and NF-kappaB activation. IL-1beta stimulation of epithelial cells did not generate any ROIs and NF-kappaB induction was not influenced by 5-LOX inhibitors. However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

Text



Motivation



TM Goals

We previously reported that the role of reactive oxygen intermediates (ROIs) in NF-kappaB activation by proinflammatory cytokines was cell specific.

However, the sources for ROIs in various cell types are yet to be determined and might include 5-lipoxygenase (5-LOX) and NADPH oxidase. 5-LOX and 5-LOX activating protein (FLAP) are coexpressed.

Stimulation of lymphoid cells with interleukin-1beta (IL-1beta) led to ROI production and NF-kappaB activation.

IL-1beta stimulation of epithelial cells did not generate any ROIs and NF-kappaB induction was not influenced by 5-LOX inhibitors.

However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

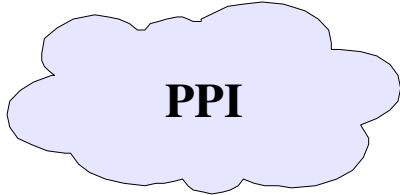
Sentence & word segmentation

Text structure

Text



Motivation



TM Goals

We previously reported that the role of reactive oxygen intermediates (ROI) in NF-kappaB activation by proinflammatory cytokines was cell-specific. Interleukin-10 (IL-10) inhibits nuclear factor-kappaB (NF-kappaB) activation in human monocytes. However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

Parts-of-speech / morphology

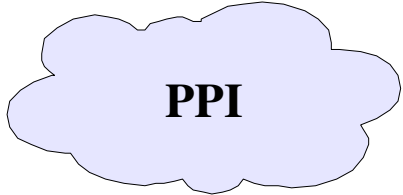
Sentence & word segmentation

Text structure

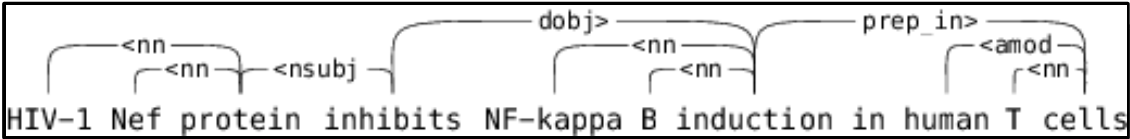
Text



Motivation



TM Goals



Ho
and
and
Sti
pre
IL-1b
kappaB induction was not influenced by 5-LOX inhibitors.

However, reintroduction of a functional 5-LOX system in these cells allowed ROI production and 5-LOX-dependent NF-kappaB activation.

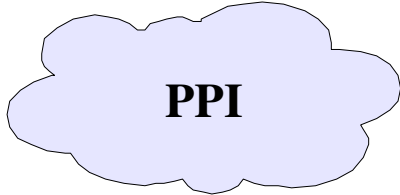
- Syntactic analysis
- Parts-of-speech / morphology
- Sentence & word segmentation

Text structure

Text



Motivation



TM Goals

And many more: chunking, deep syntactic analysis, coreference resolution, term dictionary lookup...

HIV-1 Nef protein inhibits NF-kappa B induction in human T cells

inhibits/VBZ nuclear/JJ factor/NN kappa/NN B/NN (/ (NF/NN kappa/NN B/NN)/) activation/NN in/IN human/JJ monocytes/NNS ./.

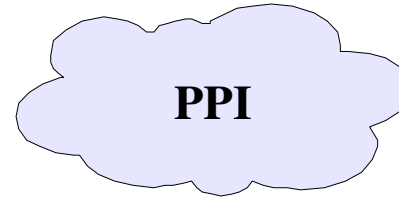
Analysis of text structure

Text structure

Text



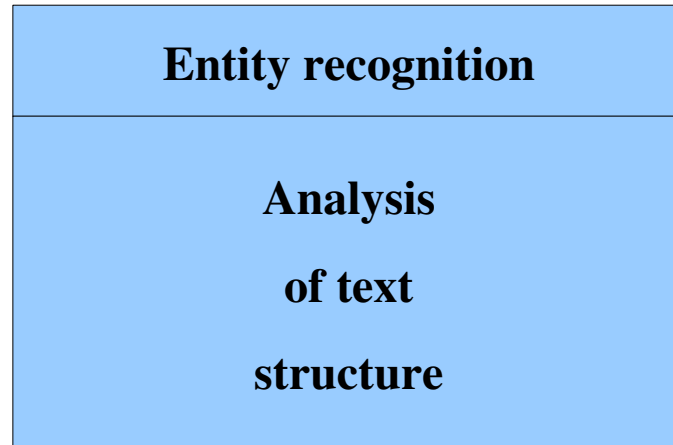
Motivation



TM Goals



Text semantics



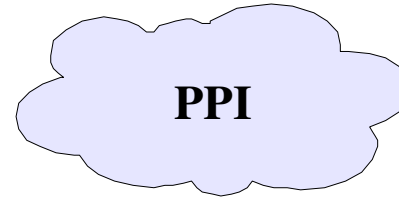
Text structure



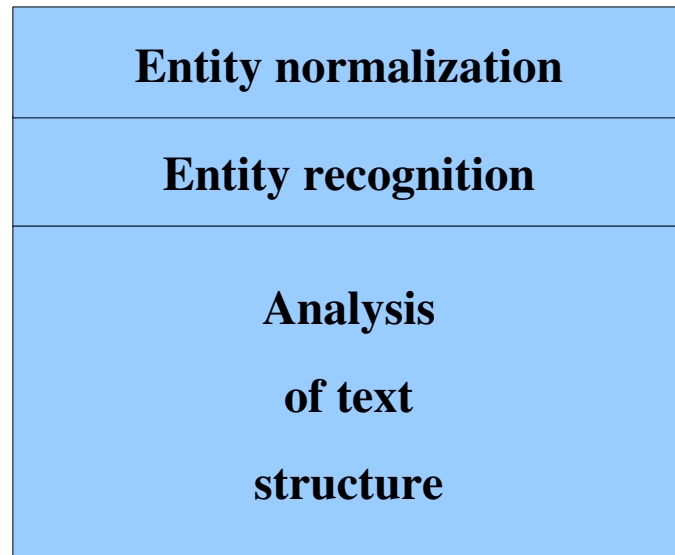
Text



Motivation



TM Goals



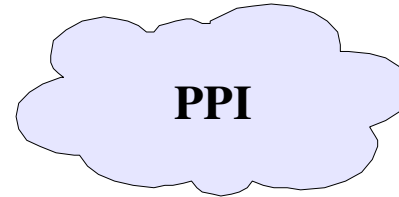
Text semantics

Text structure

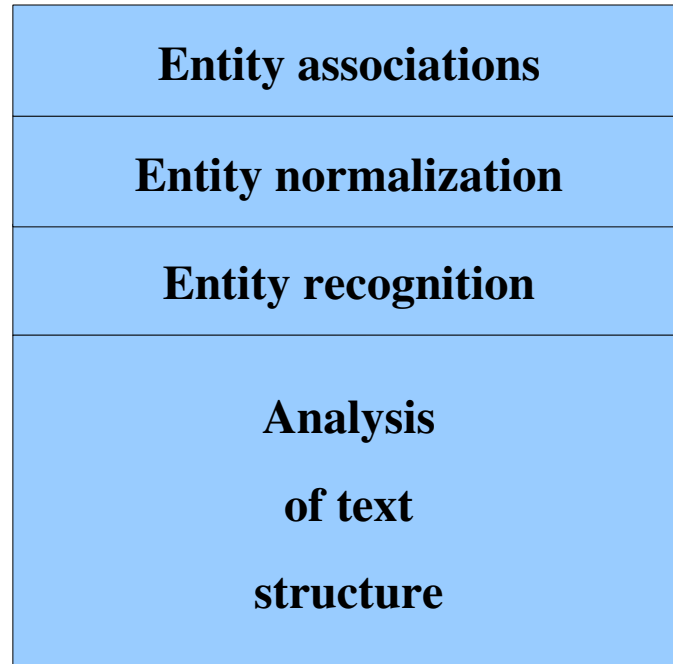
Text



Motivation



TM Goals



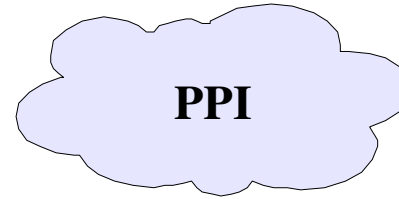
Text semantics

Text structure

Text



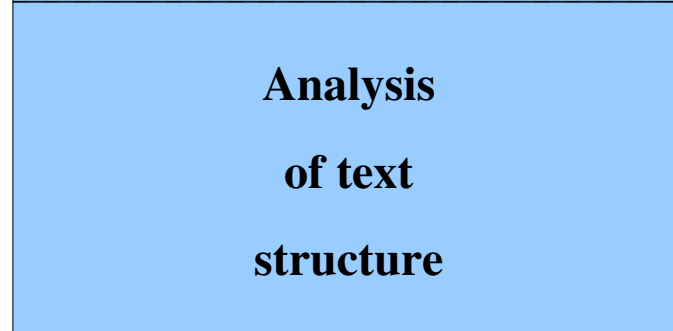
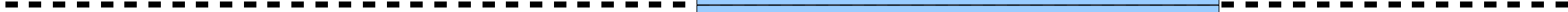
Motivation



TM Goals



Text semantics

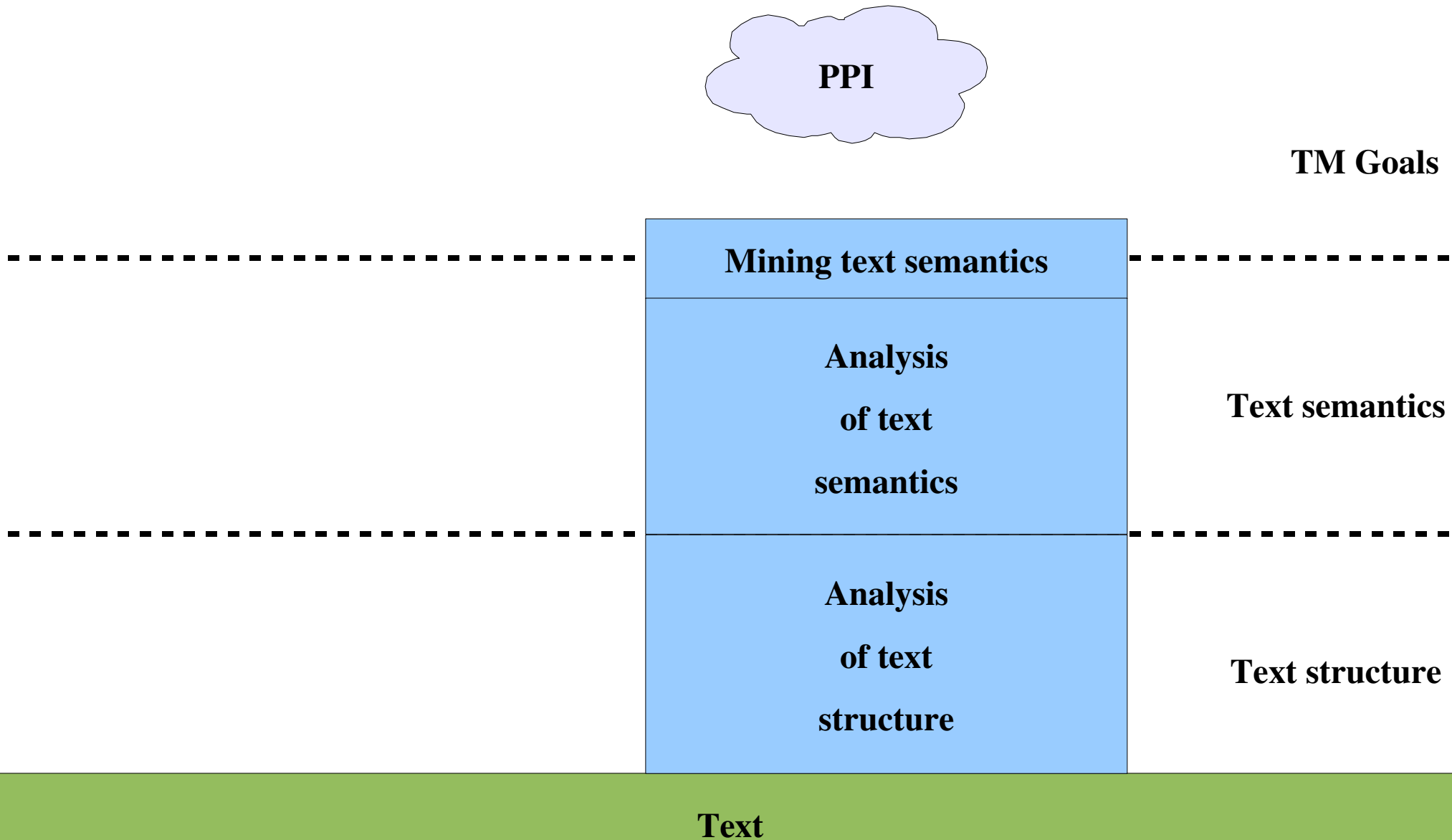


Text structure



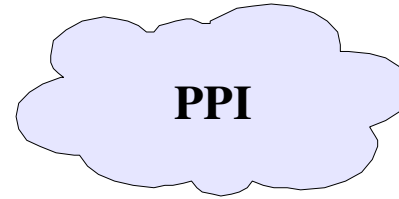
Text

Motivation





Motivation



TM Goals

Mining

**Analysis
of text
semantics**

Text semantics

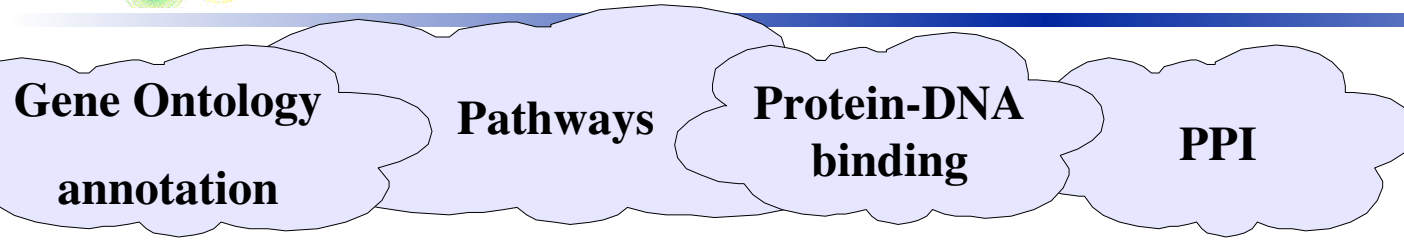
**Analysis
of text
structure**

Text structure

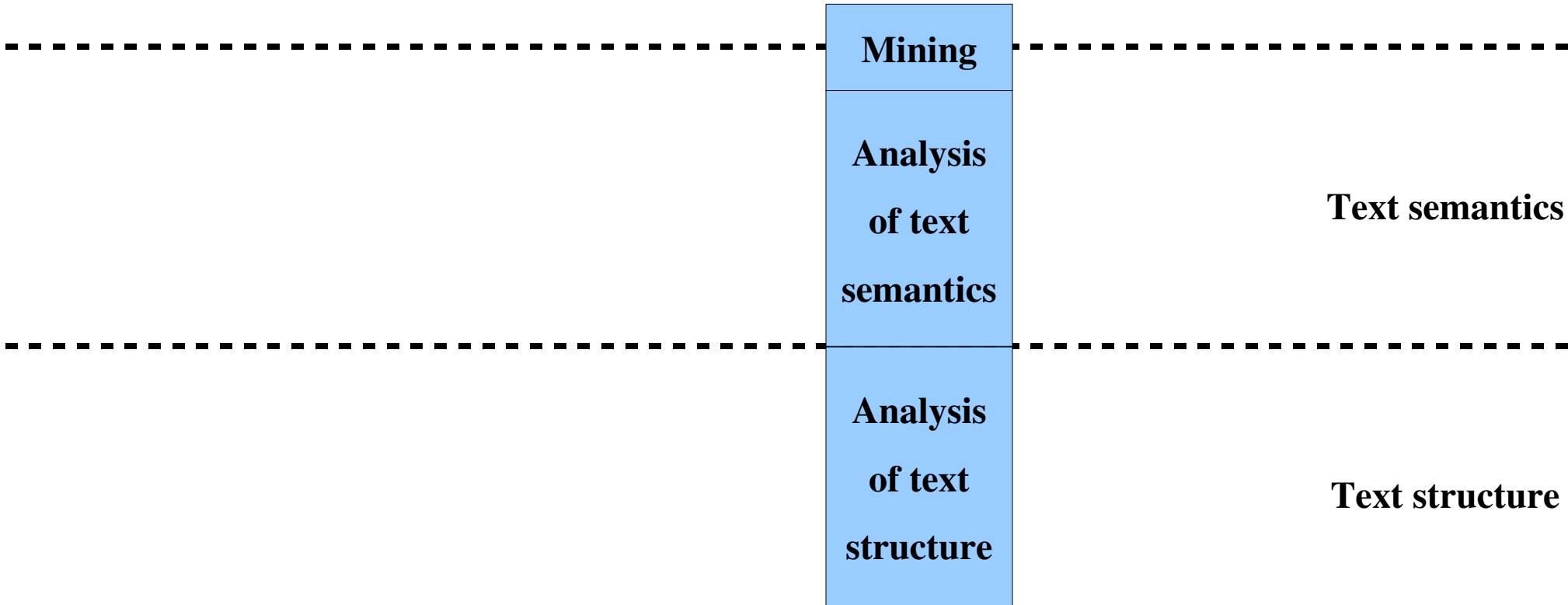
Text



Motivation



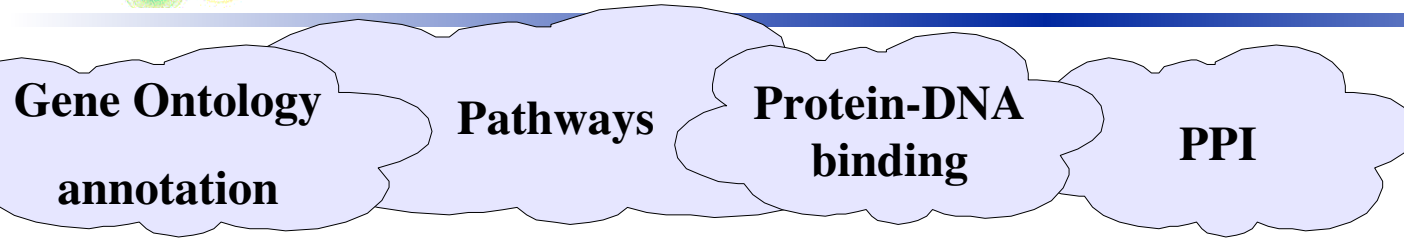
TM Goals



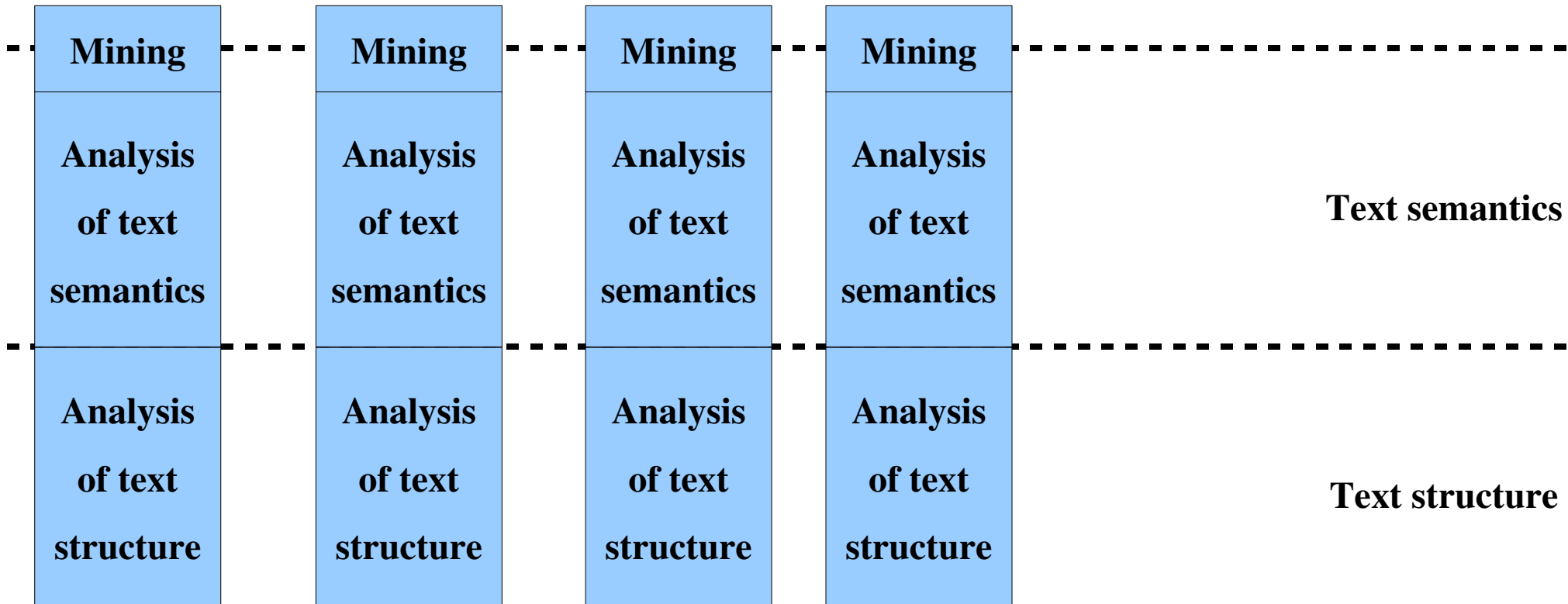
Text



Motivation



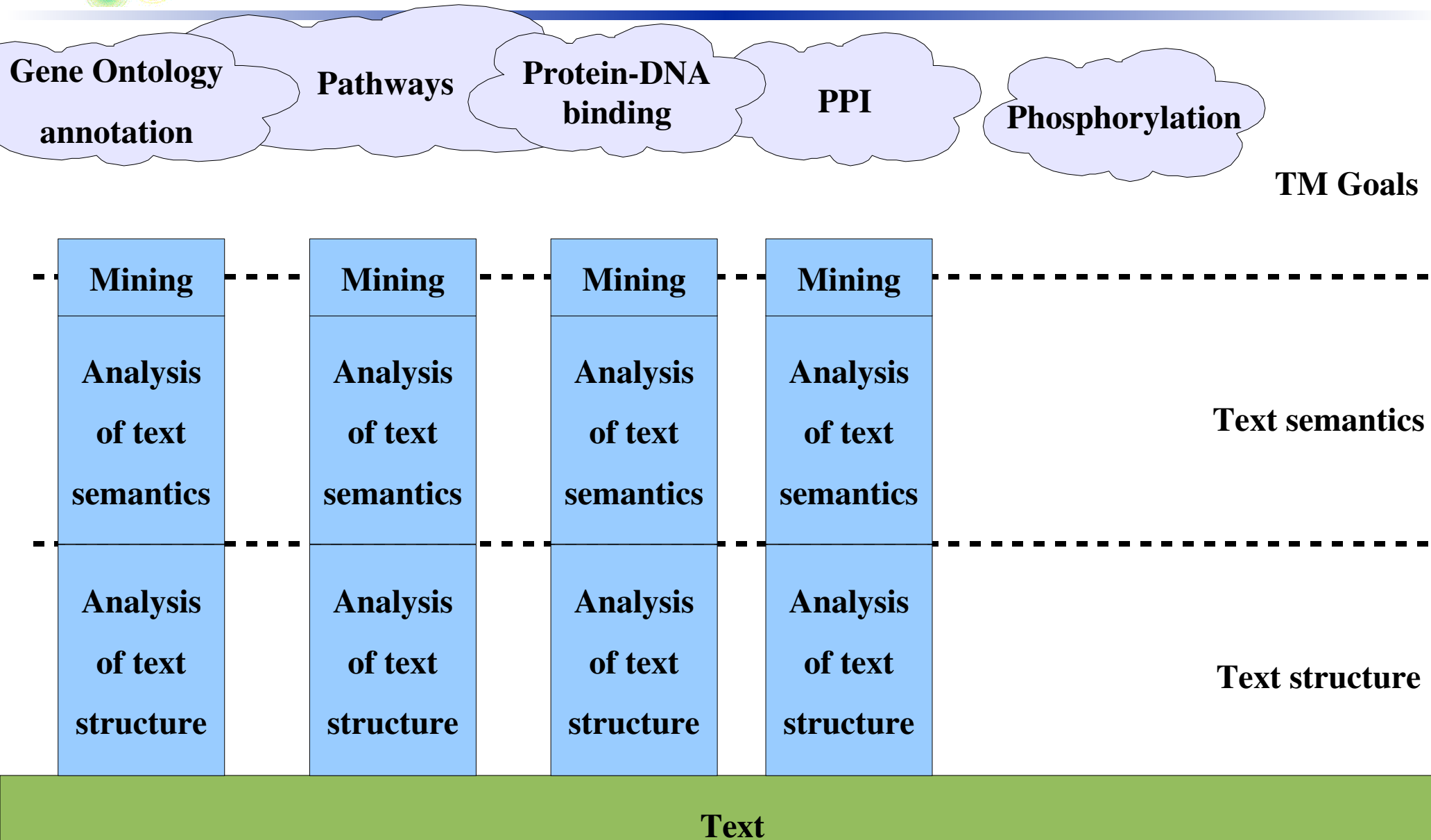
TM Goals



Text

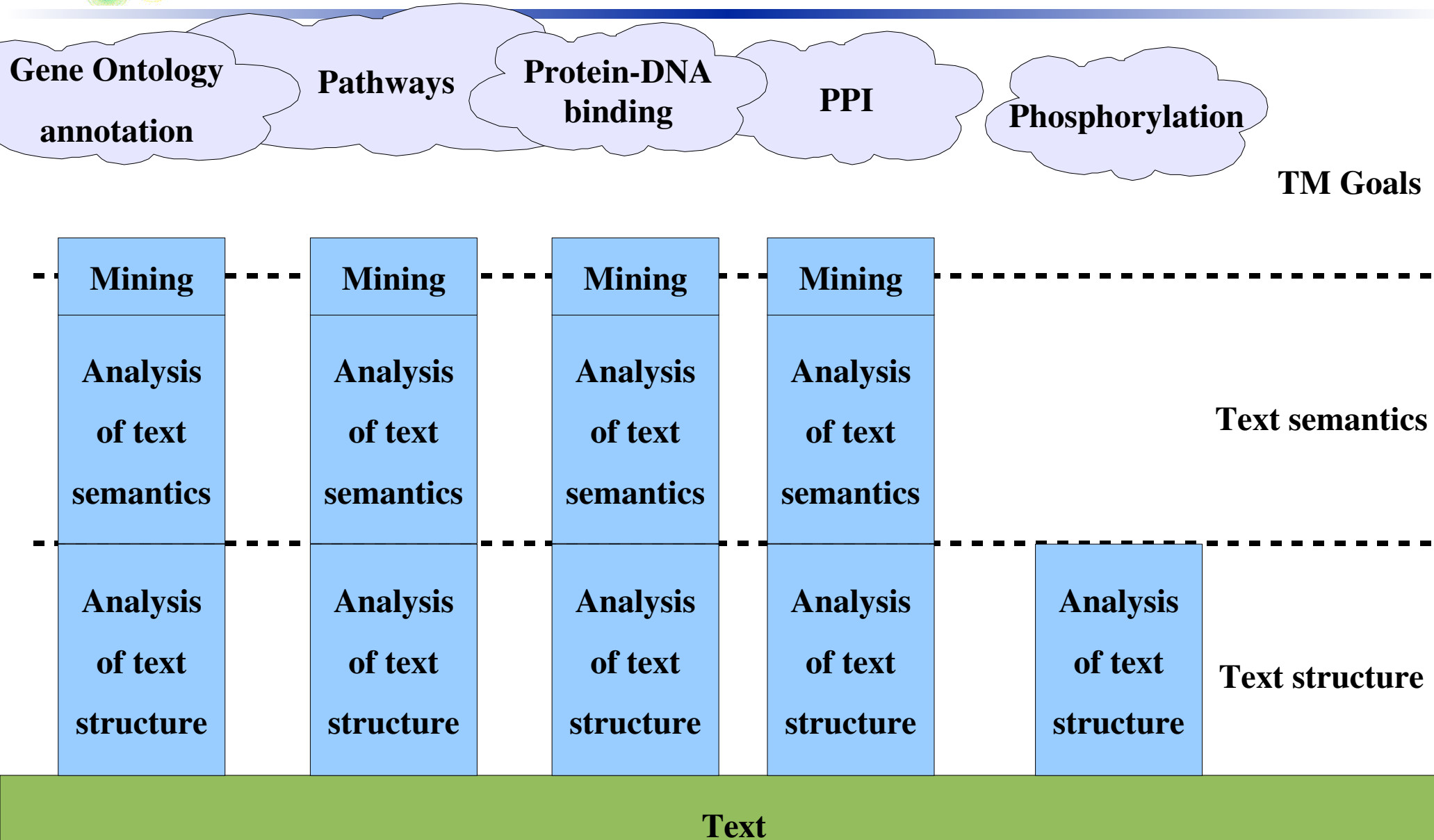


Motivation



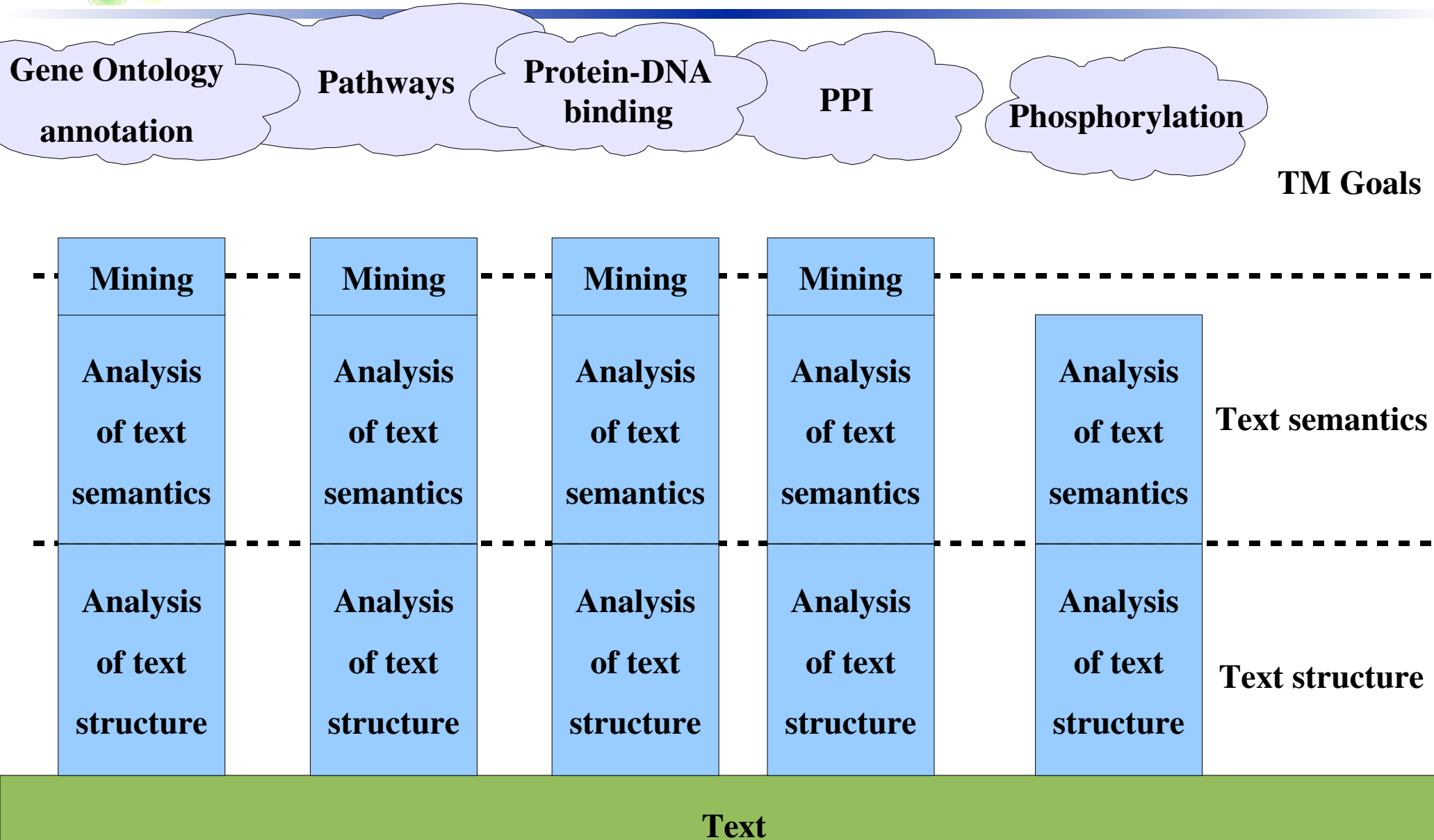


Motivation



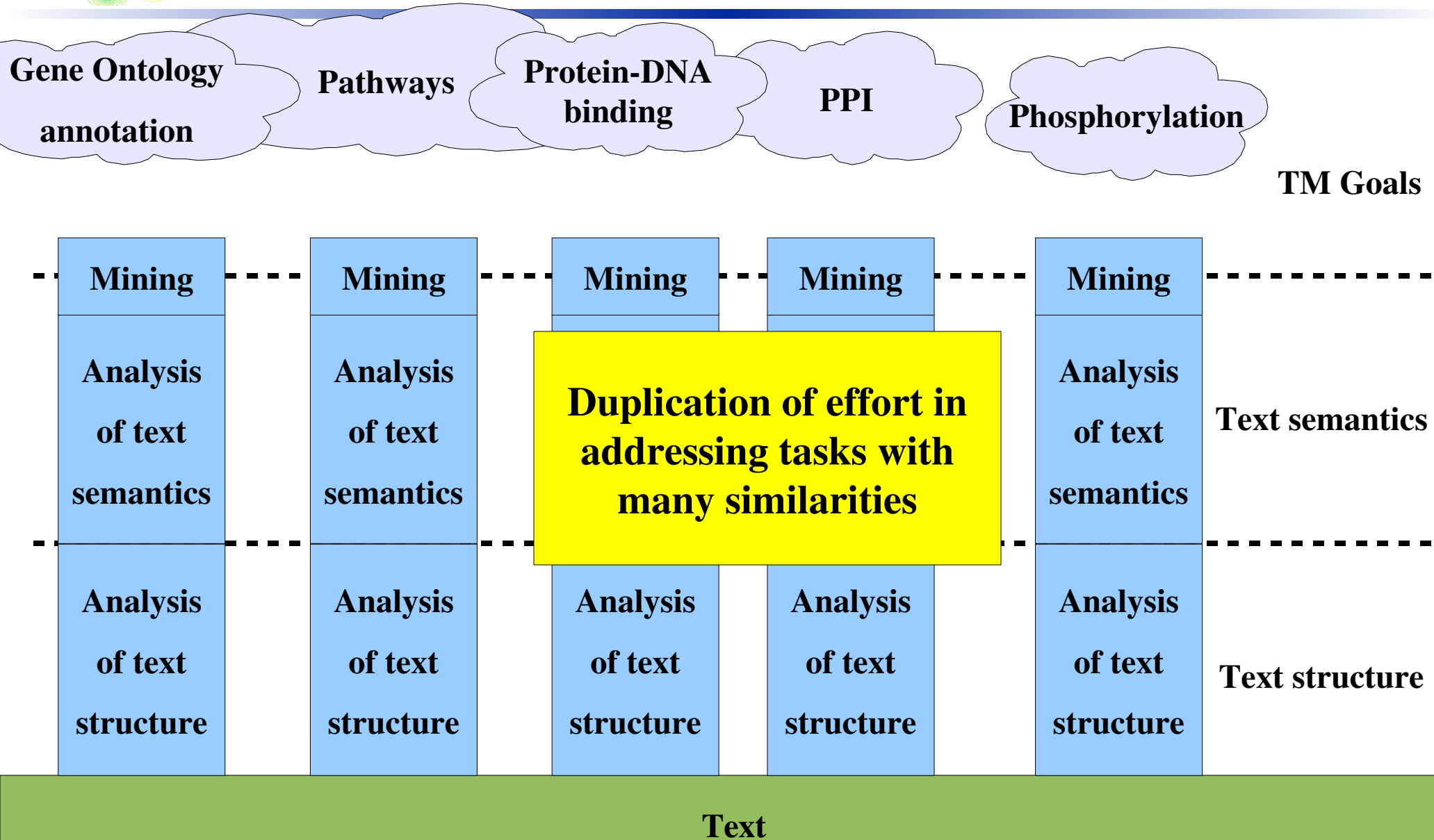


Motivation



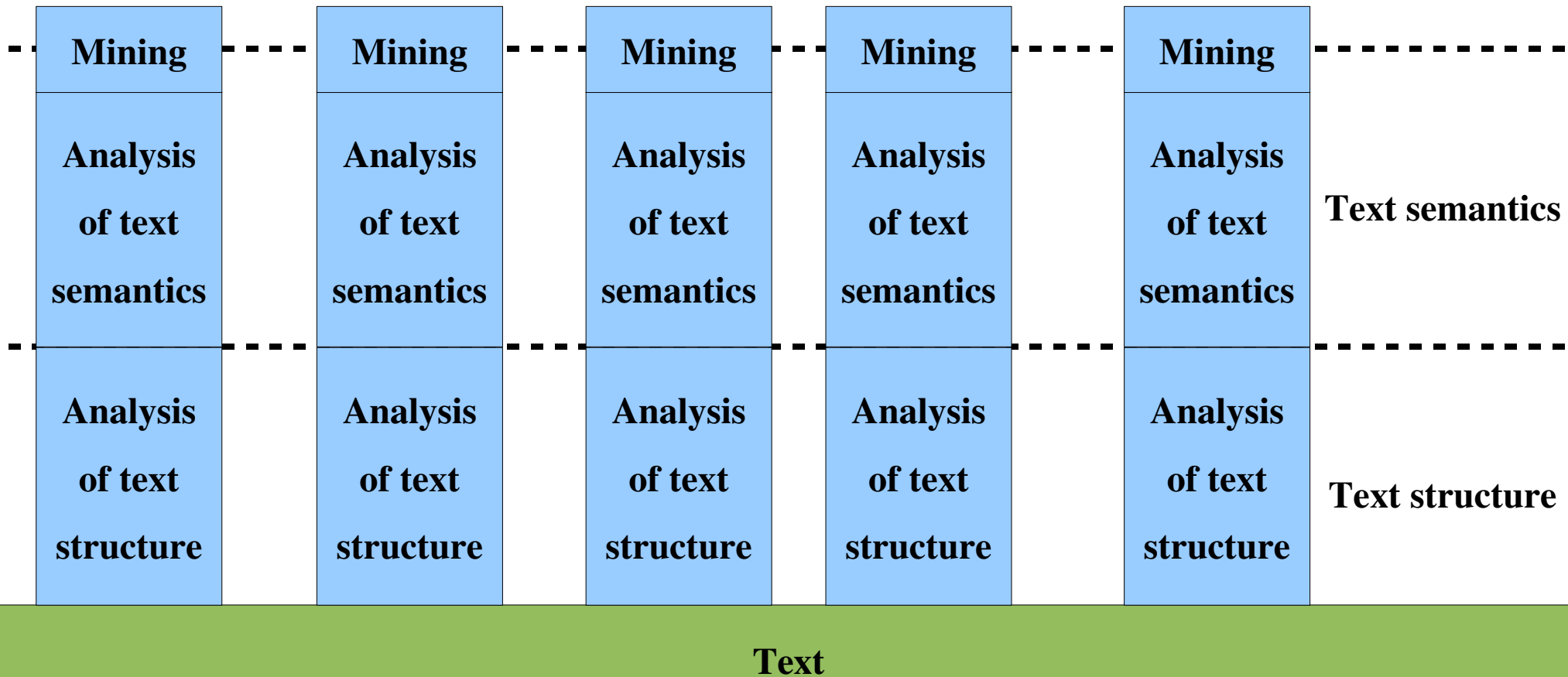
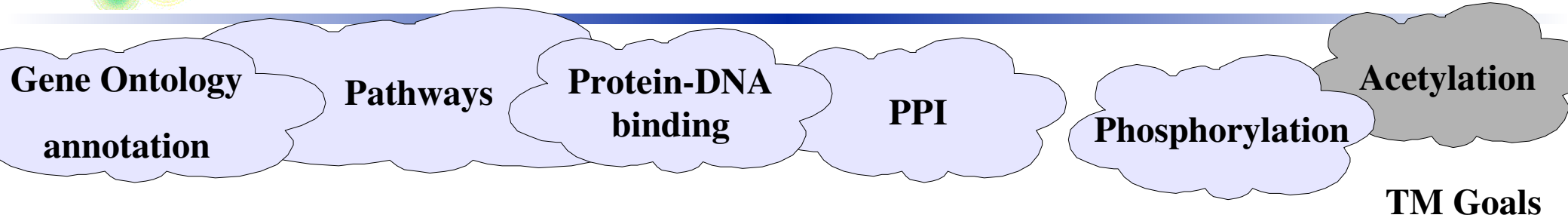


Motivation



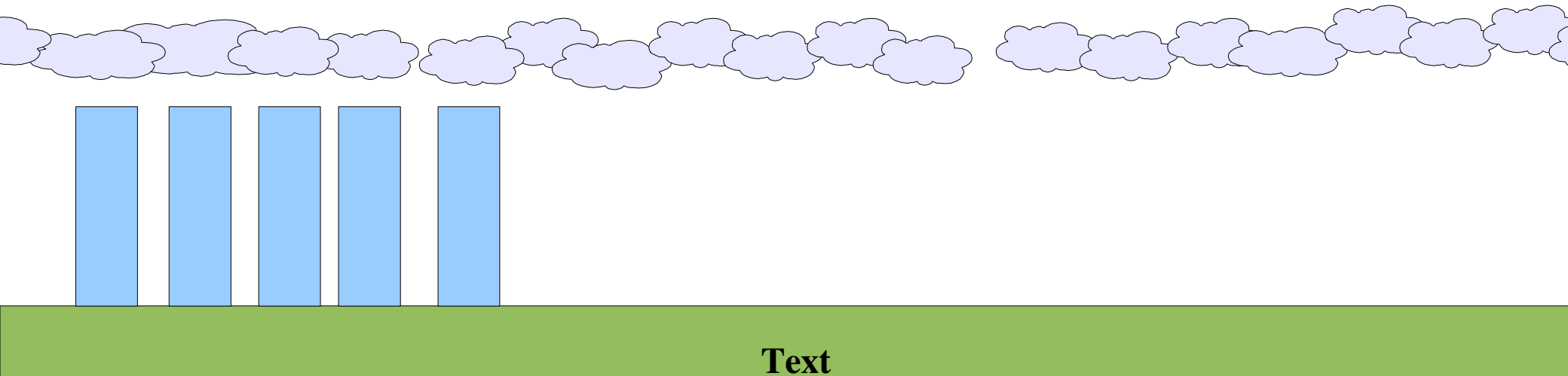


Motivation



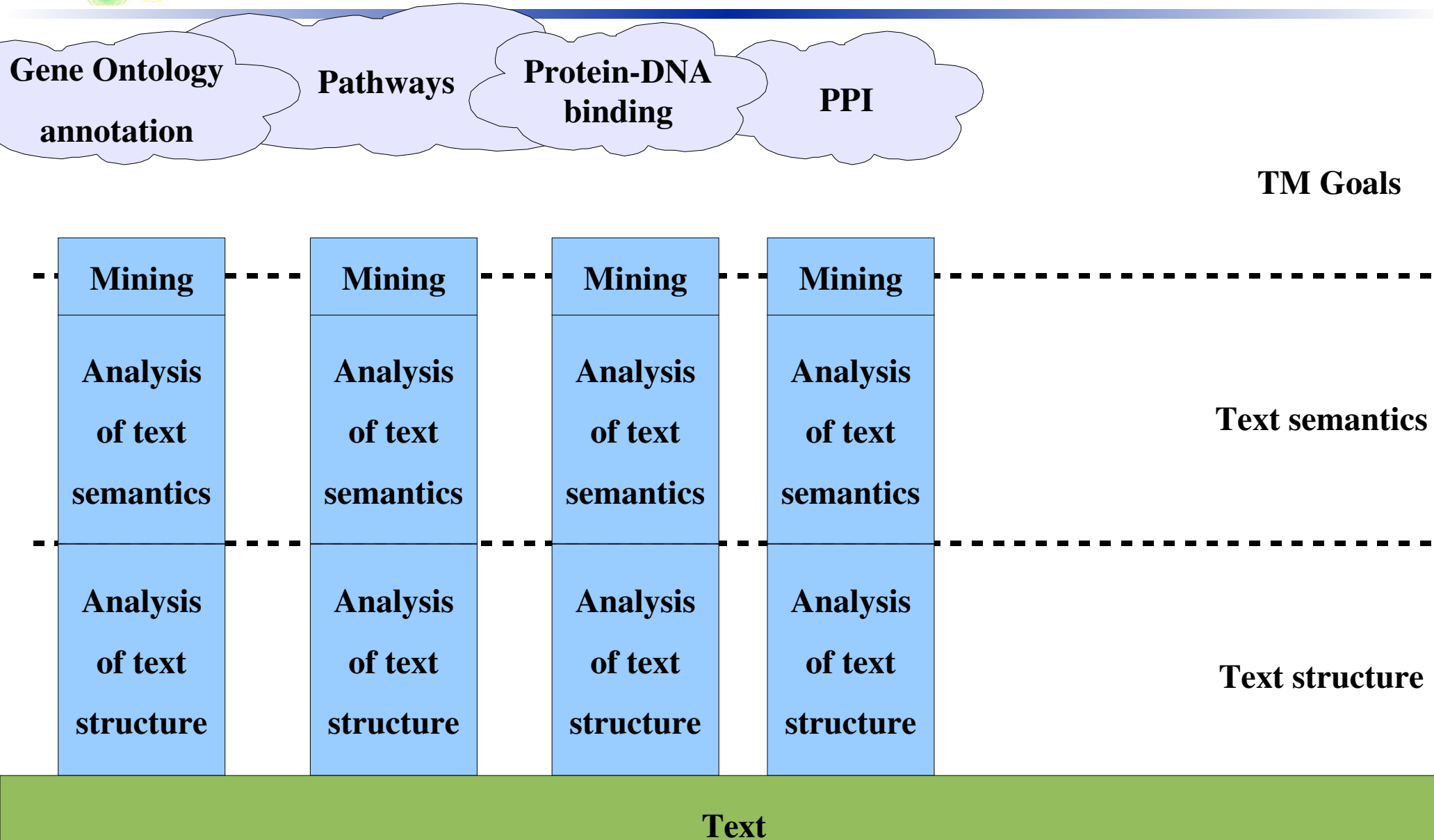
Motivation

- e.g. more than 300 types of post-translational modifications ...



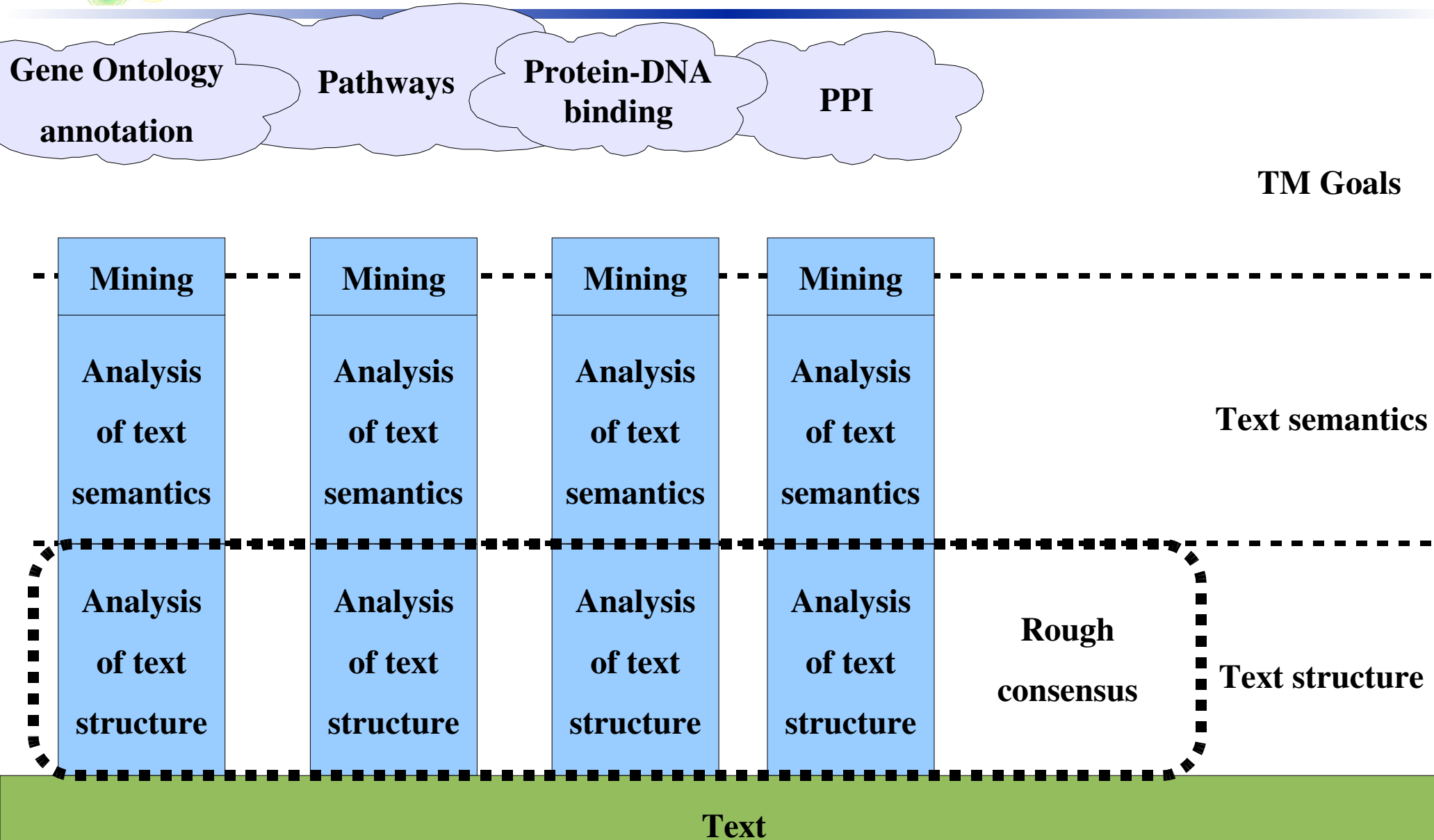


Motivation



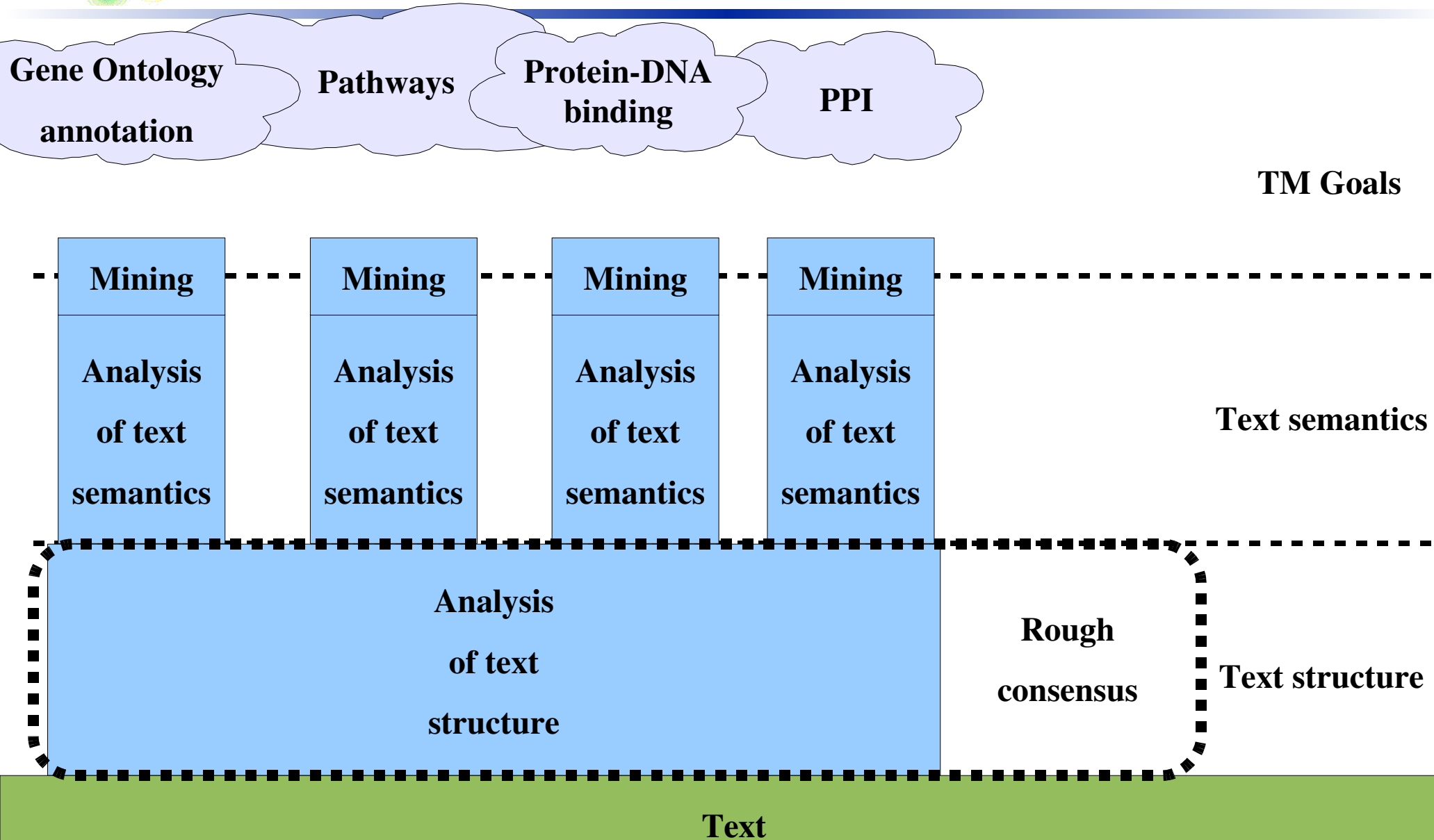


Motivation





Motivation





Motivation

- Example: what does it mean to do “PPI extraction”?
- Representation:
 - ✓ Extract (protein, protein) pairs?
 - ✓ Extract (protein, protein) pairs and assign them a type?
 - ✓ Extract (protein, interaction word, protein) triples?
- Criteria:
 - ✓ Extract all mentions?
 - ✓ Extract affirmatively stated mentions?
 - ✓ Extract novel mentions for which evidence is provided?
- Scope:
 - ✓ Extract all mentions?
 - ✓ Extract unique mentions within each sentence?
 - ✓ Extract unique mentions within each document?



Motivation

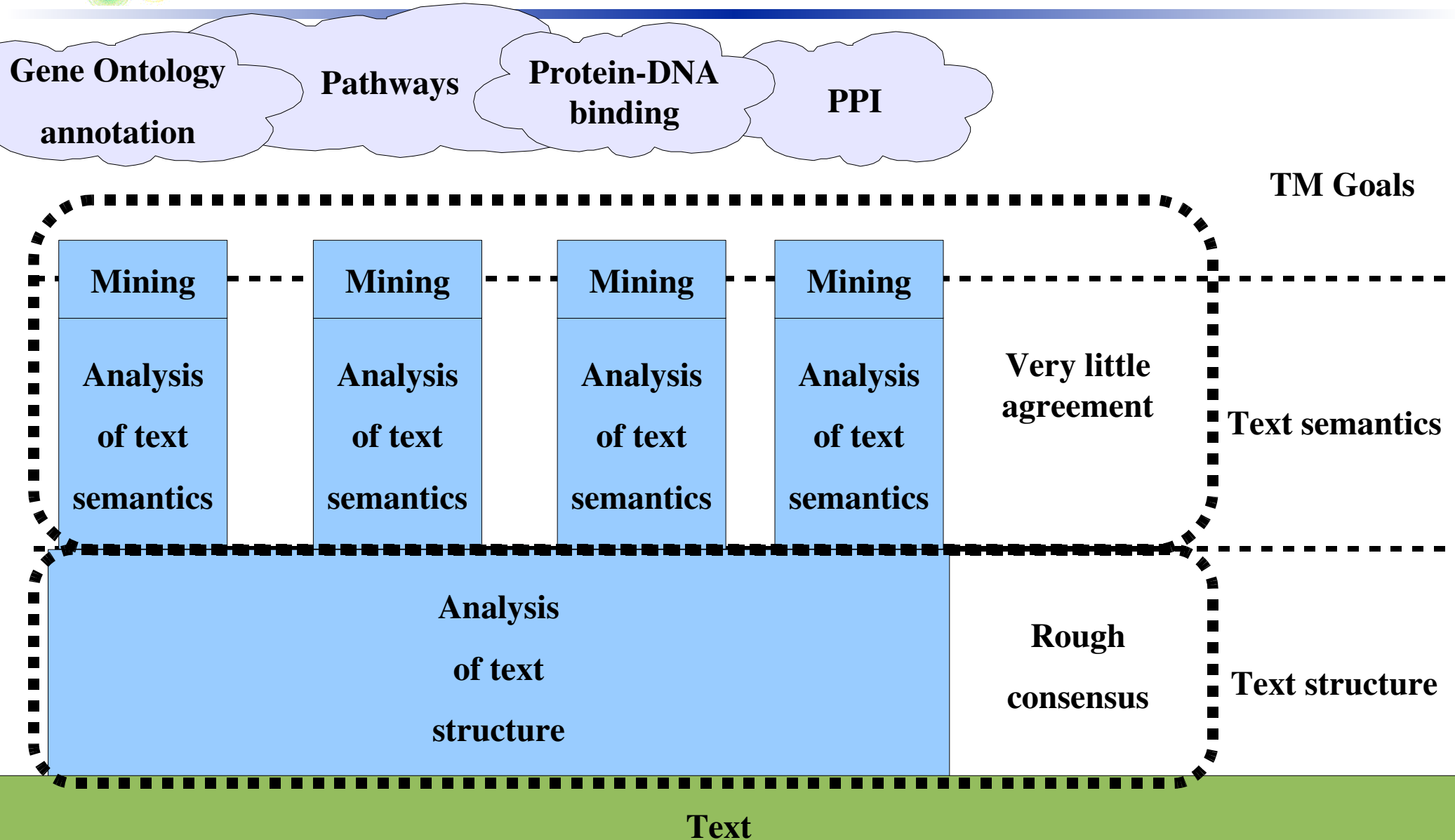
- ❑ Example: what does it mean to do “PPI extraction”?
- ❑ Representation:
 - ✓ Extract (protein, protein) pairs?
 - ✓ Extract (protein, protein) pairs and assign them a type?
 - ✓ Extract (protein, interaction word, protein) triples?
- ❑ Criteria:
 - ✓ Extract all mentions?
 - ✓ Extract affirmatively stated mentions?
 - ✓ Extract novel mentions for which evidence is provided?
- ❑ Scope:
 - ✓ Extract all mentions?
 - ✓ Extract unique mentions within each document?
 - ✓ Extract unique mentions within each document?

- Differences in ultimate information needs are unavoidable

- ... But many issues relating to representation could be answered by a “superset” approach

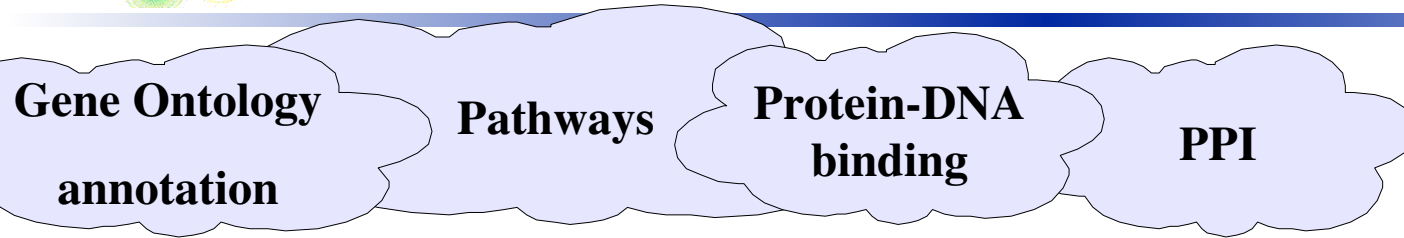


Motivation

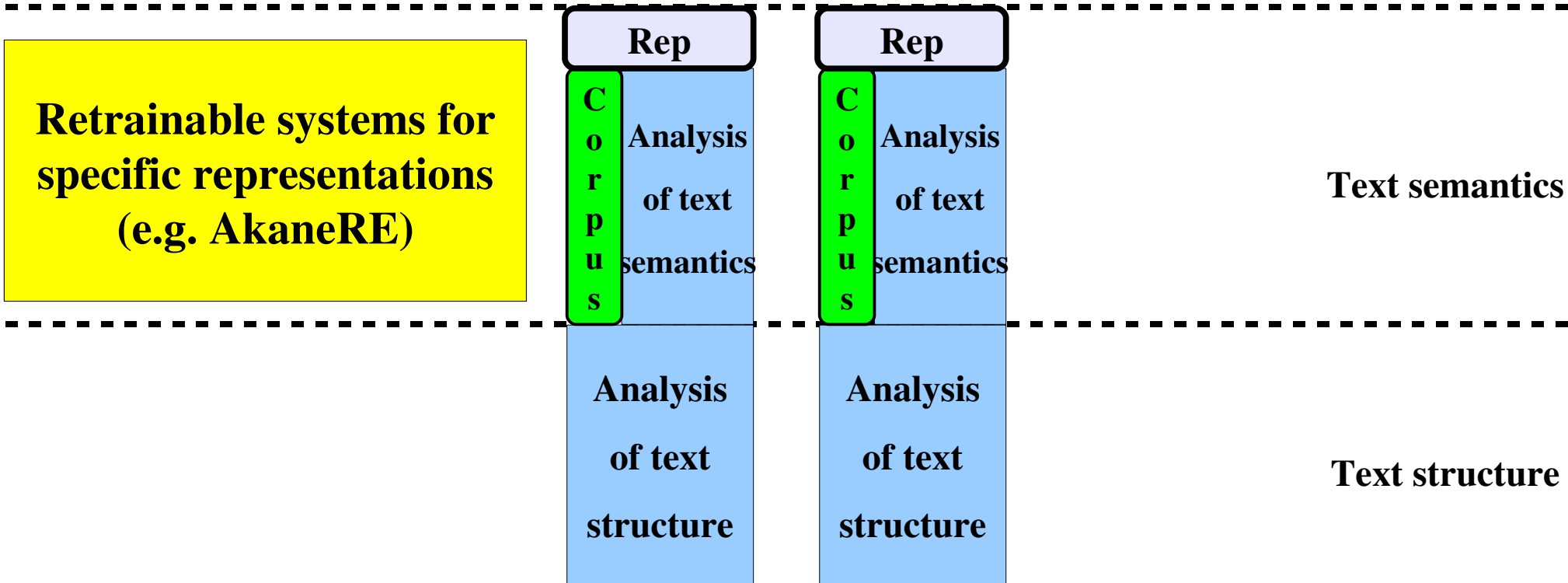




Motivation



TM Goals



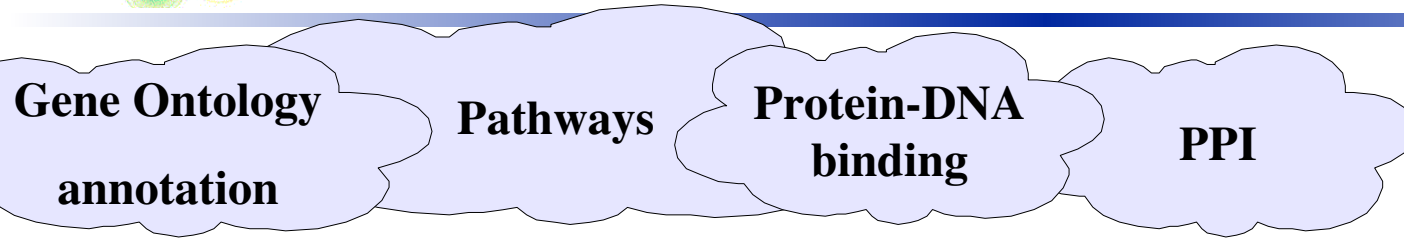
Text semantics

Text structure

Text



Motivation



TM Goals

Expressive representation of text semantics

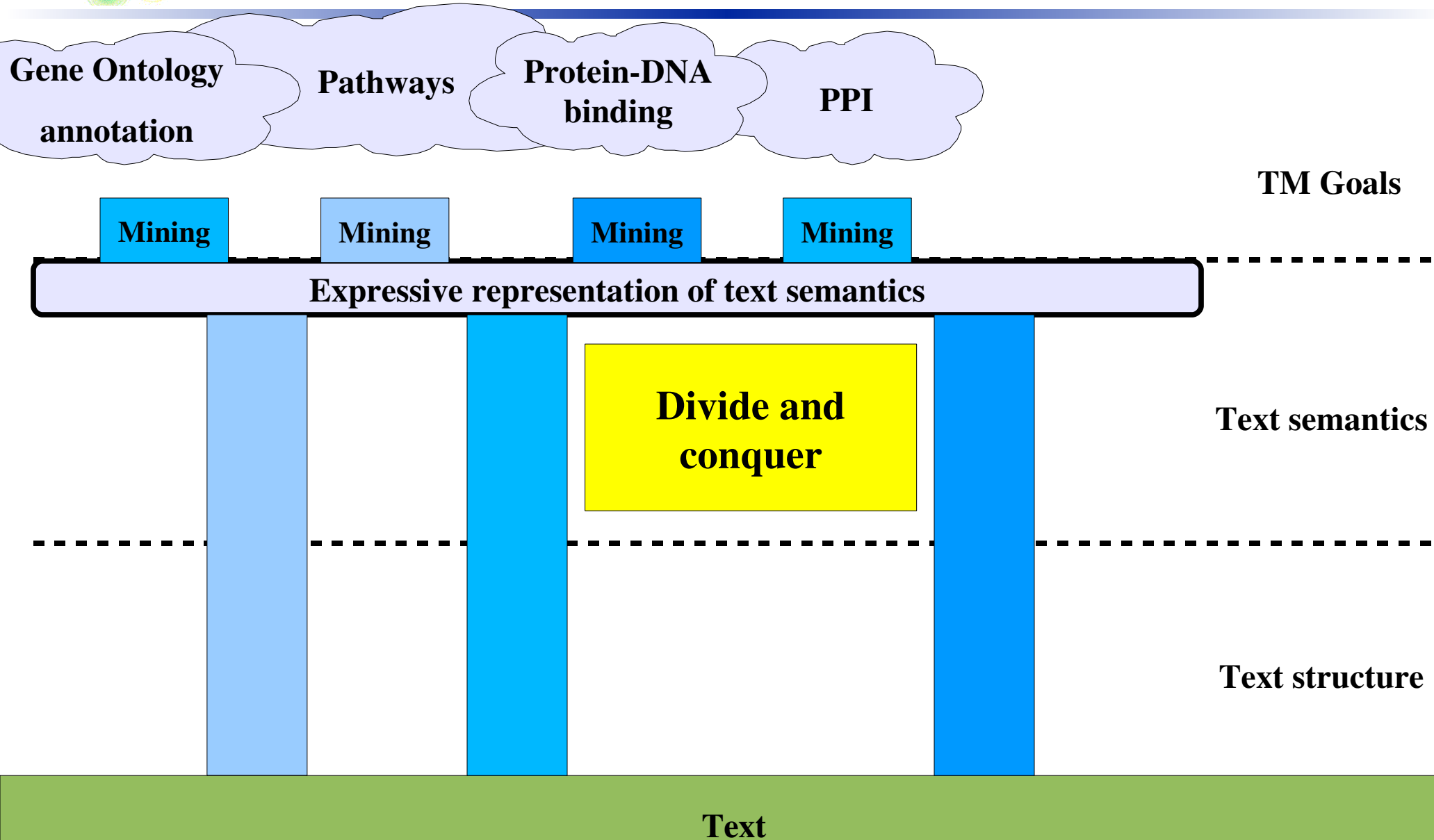
Text semantics

Text structure

Text

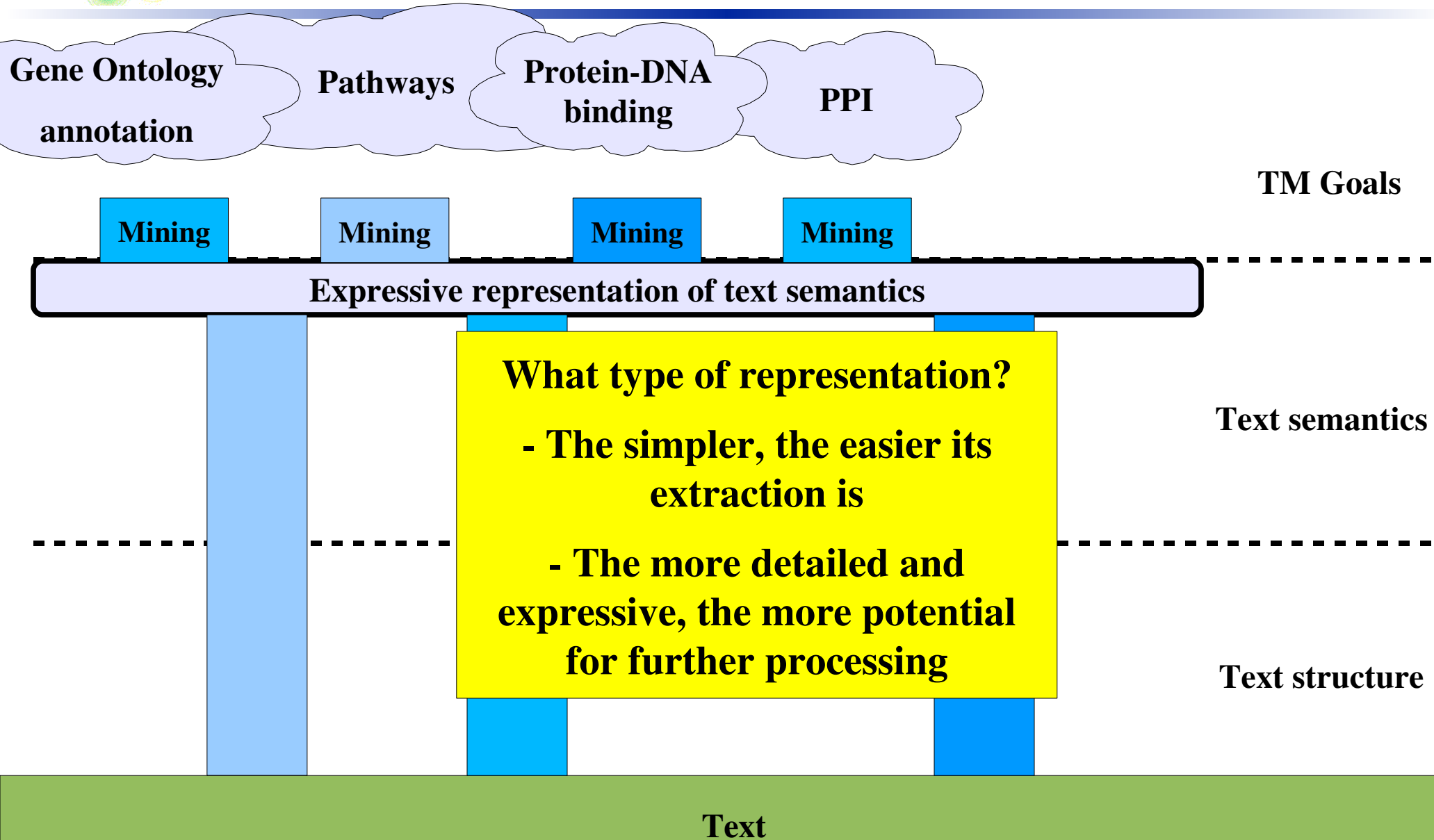


Motivation





Motivation



Representation

- What do we want to identify in texts?
 - ✓ Mentions of things of interest, their associations, and their properties

- Entities

- ✓ “Things that are”
 - ✓ The basis of the representation

binding of MAD-3 to p65

- Relations

- ✓ Representation for n-ary associations
 - ✓ e.g. (MAD-3, p65)



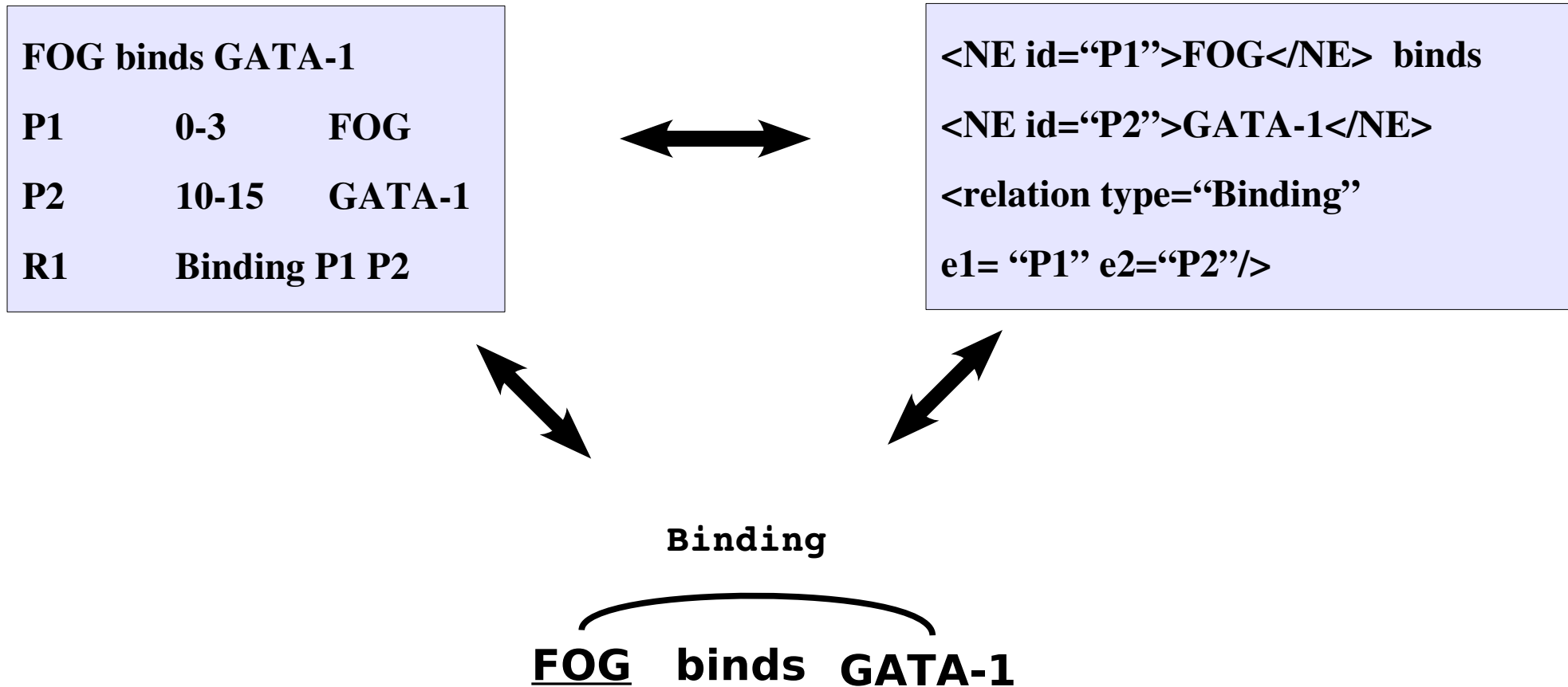
- Events

- ✓ “Things that happen”
 - ✓ Also a class of representation

binding of MAD-3 to p65

Representation

□ Representation vs. format





Outline

- Introduction and motivation
- Entities**
- Relations
- Events
- Relations (again)
- Where next?

Representation: entities

- A pair (begin, end) representing a span is the minimal representation for identifying a thing of interest in text

Here's one!

transcriptional regulation of the human Fas promoter

- When looking for more than one category of thing, triples of (Type, begin, end) suffice

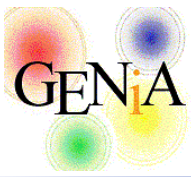
T-cell expression of the human GATA-3 gene

- This representation is used to mark entities



Representation: entities

- ❑ Named Entities
 - ✓ Mentions of names of specifically identifiable real-world entities
- ❑ Serve as the “anchors” connecting the text to reality
 - ✓ “FOG binds GATA-1” is interesting in a way that “A protein binds another protein” is not
- ❑ Names largely continuous, nonoverlapping strings (98% in GENIA)
 - ✓ Exceptions: “alpha- and beta-catenin”
- ❑ Normalization: for each entity mention, determine the specific entry in an appropriate database referred to
 - ✓ e.g. Uniprot for protein name mentions
 - ✓ Not generally possible for non-names



Representation: entities

- ❑ Non-name terms frequently embed other entities (nesting)
 - ✓ e.g. “[alpha-catenin] promoter”
- ❑ ... and are discontinuous in text
 - ✓ e.g. “region of the anti-apoptotic protein bcl-xL”
- ❑ (Some work on the extraction of embedded entities)

→ NEs can be represented as nonoverlapping (start, end) spans,
non-names add nesting



Outline

- Introduction and motivation
- Entities
- Relations**
- Events
- Relations (again)
- Where next?



Representation: relations

- A pair (Entity1, Entity2) identifying two entities is the minimal representation for capturing some association between two things in text

These two

Expression of FOG, which binds GATA-1...

- (Again, some type assumed, and explicit types can be added when multiple classes of relations are of interest)
- (pairwise) relation representation



Representation: relations

- Entities as untyped continuous strings + pair-of-entities relation representation (AIMed, PMID 10206993)

**This study demonstrates that IL - 8 recognizes and
P1
activates CXCR1, CXCR2, and the Duffy antigen by
P2 P3 P4
distinct mechanisms .**

Relations: (P1, P2)
(P1, P3)
(P1, P4)

Representation: relations

- Entities as untyped continuous strings + pair-of-entities relation representation (AIMed, PMID 10206993)

**This study demonstrates that IL - 8 recognizes and
P1
 activates CXCR1, CXCR2, and the Duffy antigen by
P2 P3 P4
 distinct mechanisms .**

Relations: (P1, P2)
 (P1, P3)
 (P1, P4)

**-> no distinction between “recognizes”
 and “activates”, no way to express
 that two relations are stated**



Representation: relations

- ❑ Simplest possible representation for capturing associations in text is a reasonable match for what is required in text mining for protein-protein interactions (DBs such as DIP)
 - ✓ Single type of entity (“protein”) and single type of relation (“interaction”)

- ❑ Easy point of entry to biomedical text mining and overwhelmingly the most popular task
 - ✓ More than a dozen annotated corpus resources (AIMed, BioCreative-PPI, BioInfer, BioRAT, HPRD50, IEPA, ITI TXM, PDG/PICorpus, Wisconsin, ...)
 - ✓ ... and dozens of studies (100+?) targeting PPI



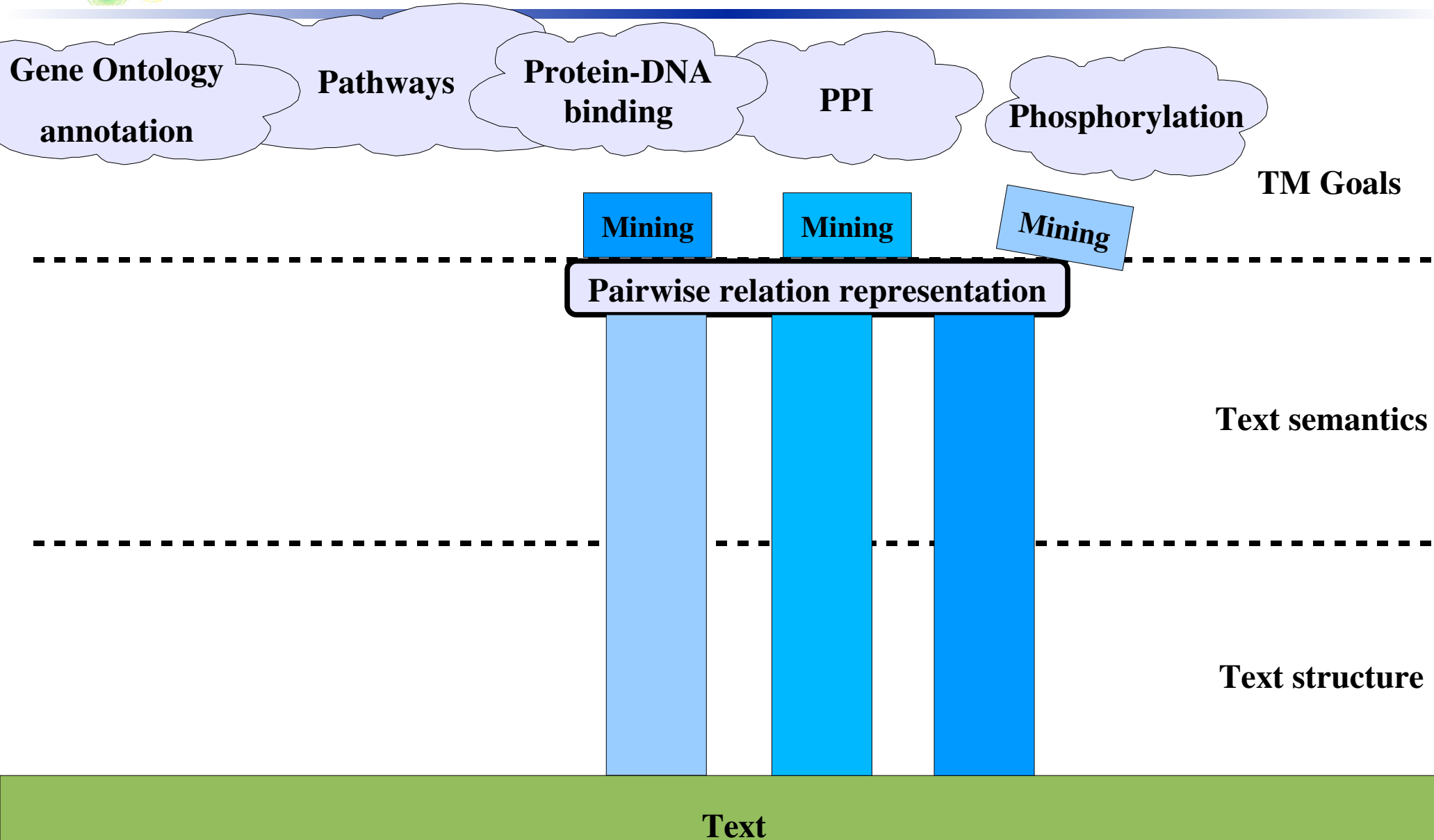
Representation: relations

- The pair-of-entities relation representation has also been applied to a fair number of tasks besides PPI
 - ✓ Protein-gene binding
 - ✓ Gene-disease associations
 - ✓ Gene-drug associations
 - ✓ Regulation relations
 - ✓ Gene-mutation associations
 - ✓ Protein localization

- The requirements for performing these tasks share much in common
 - ✓ ... but proposed methods largely task-specific



Motivation





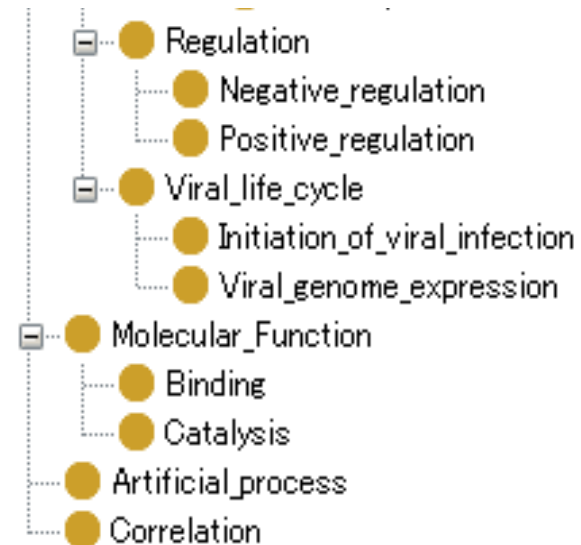
Relation representation: limitations

- Typeless relations can only support the task they were defined for
 - ✓ A representation of both “P1 binds P2” and “P1 inhibits P2” as (P1, P2) is unhelpful if the distinction between binding and inhibition matters

Relation representation: limitations

- Typeless relations can only support the task they were defined for
 - ✓ A representation of both “P1 binds P2” and “P1 inhibits P2” as (P1, P2) is unhelpful if the distinction between binding and inhibition matters

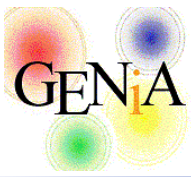
- Assign types to relations
 - Binding(P1, P2)
 - Negative_regulation(P1, P2)





Relation representation: limitations

- Many interesting associations involve more than two participants
 - ✓ “Binding of SNAP23, syntaxin-4 and VAMP-2”
 - ✓ “TR6 inhibits the binding of LIGHT with HVEM”



Relation representation: limitations

- Many interesting associations involve more than two participants
 - ✓ “Binding of SNAP23, syntaxin-4 and VAMP-2”
 - ✓ “TR6 inhibits the binding of LIGHT with HVEM”

- Decomposition and approximation ...
 - Binding (SNAP23, syntaxin-4), Binding (SNAP23, VAMP-2), Binding(syntaxin-4, VAMP-2)
 - Inhibit (TR6, LIGHT) Inhibit (TR6, HVEM) Binding (LIGHT, HVEM)

- ... or n-ary relations and structured relation types ... ?
 - Binding (SNAP23, syntaxin-4, VAMP-2)
 - Inhibit-Binding (TR6, LIGHT, HVEM)
(roles of participants implied by position in relation)



Relation representation: limitations

- Many interesting associations involve more than two participants
 - ✓ “Binding of SNAP23, syntaxin-4 and VAMP-2”
 - ✓ “TR6 inhibits the binding of LIGHT with HVEM”

- Decomposition and approximation ...
 - Binding (SNAP23, syntaxin-4), Binding (SNAP23, VAMP-2), Binding(syntaxin-4, VAMP-2)
 - Inhibit (TR6, LIGHT) Inhibit (TR6, HVEM) Binding (LIGHT, HVEM)

- ... or n-ary relations and structured relation types ... ?
 - Binding (SNAP23, syntaxin-4, VAMP-2)
 - Inhibit-Binding (TR6, LIGHT, HVEM)
(roles of participants implied by position in relation)

←Frequently adopted in e.g. PPI corpora



Relation representation: limitations

- ❑ Not all potentially relevant participants are necessarily mentioned
 - ✓ “translocation of p65 from the cytoplasm to the nucleus”
 - ✓ “translocation of p65 from the cytoplasm”
 - ✓ “translocation of p65 to the nucleus”



Relation representation: limitations

- Not all potentially relevant participants are necessarily mentioned
 - ✓ “translocation of p65 from the cytoplasm to the nucleus”
 - ✓ “translocation of p65 from the cytoplasm”
 - ✓ “translocation of p65 to the nucleus”

- Subdivide relations by participants ...?
 - Localize-from-to (p65, cytoplasm, nucleus)
 - Localize-from (p65, cytoplasm)
 - Localize-to (p65, nucleus)



Relation representation: limitations

- Stretching the representation past its breaking point
 - ✓ “binding of MAD-3 to p65 causes the localization of p65 from the nucleus to the cytoplasm”



Relation representation: limitations

- Stretching the representation past its breaking point
 - ✓ “binding of MAD-3 to p65 causes the localization of p65 from the nucleus to the cytoplasm”
 - Binding-causing-localization-from-to (MAD-3, p65, p65, nucleus, cytoplasm) ... ?

(Not only is this intuitively not the right thing to do, without explicit structure this type of representation effectively ceases to be computationally accessible)



Relation representation: limitations

□ The broader issue

Need to capture *structured* associations of *varying* and (potentially) *arbitrary numbers* of participants in *different roles*.

Binding-causing-localization-from-to



(MAD-3, p65, p65, nucleus, cytoplasm)



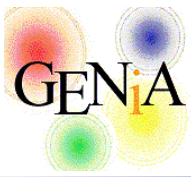
Outline

- Introduction and motivation
- Entities
- Relations
- Events**
- Relations (again)
- Where next?



Event representation

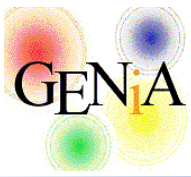
- ❑ A number of recent efforts have proposed expressive representations and corresponding annotated resources
 - ✓ BioInfer (2007), GENIA Event (2008), BioNLP ST (2009), GREC (2009)
- ❑ Here, these will be termed event representations
- ❑ Some “rough consensus” properties of events
 - ✓ Connected to a specific expression in text (“event trigger”)
 - ✓ Given a type defined by an ontology (e.g. GO)
 - ✓ Associates variable numbers of participants in specific roles
 - ✓ Primary objects of annotation, allowing events to participate in other events (structured representation)



Event representation

□ Event representation by example

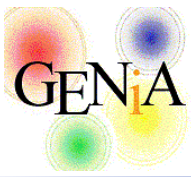
binding of MAD-3 to p65 initiates the movement of p65 from the nucleus to the cytoplasm



Event representation

- Mentions of **entities** are identified

binding of **MAD-3** to **p65** initiates the movement of **p65** from the **nucleus** to the **cytoplasm**



Event representation

- Mentions of **entities** are identified and assigned **types**

binding of **Protein** **Protein** **MAD-3** to **p65** initiates the movement of **Protein** **p65** from the **Cell comp.** **nucleus** to the **Cell comp.** **cytoplasm**

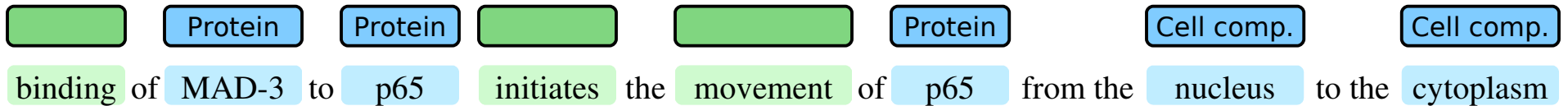
□ Events

binding of **Protein** **Protein** **MAD-3** to **p65** initiates the movement of **Protein** **p65** from the **Cell comp.** **nucleus** to the **Cell comp.** **cytoplasm**

Event representation

□ Events ...

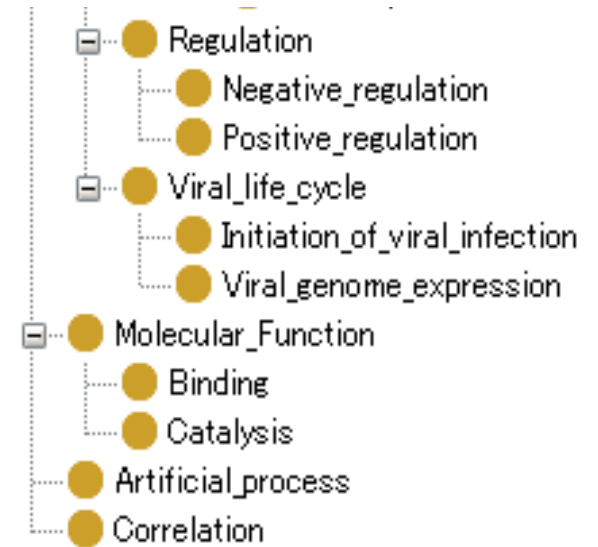
- ✓ are connected to **specific expressions** in text



Event representation

□ Events ...

- ✓ are connected to specific expressions in text
- ✓ are given **types** defined by an ontology

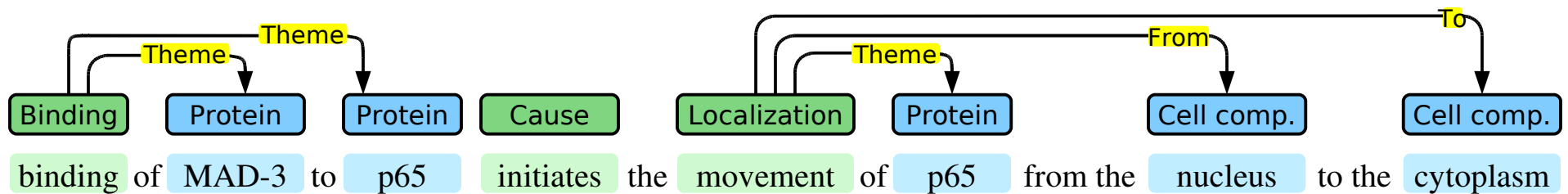


Binding Protein Protein Cause Localization Protein Cell comp. Cell comp.
 binding of MAD-3 to p65 initiates the movement of p65 from the nucleus to the cytoplasm

Event representation

□ Events ...

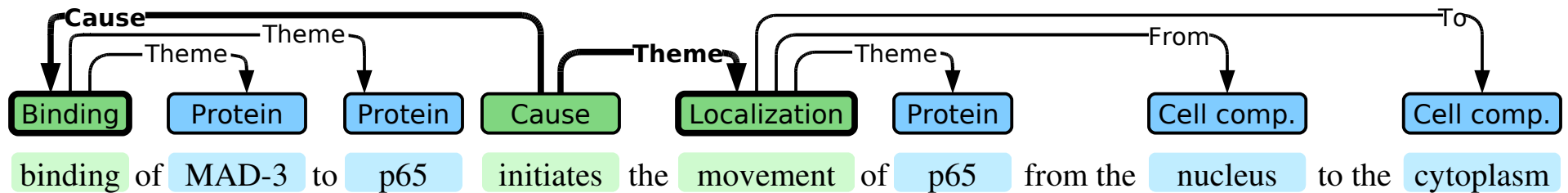
- ✓ are connected to specific expressions in text
- ✓ are given types defined by an ontology
- ✓ associate variable numbers of participants in specific **roles**



Event representation

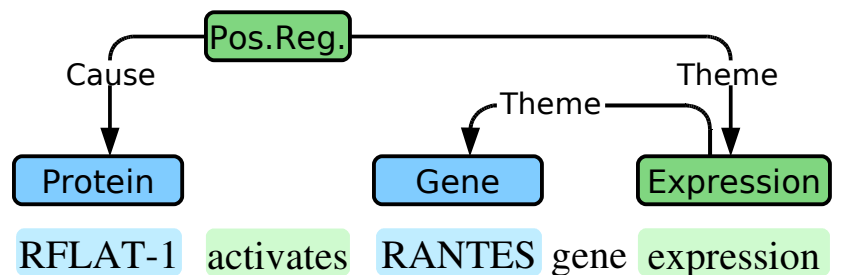
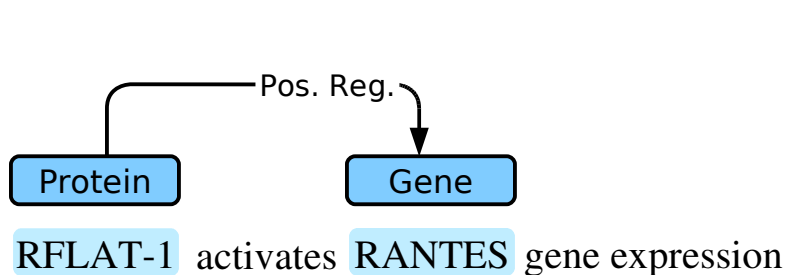
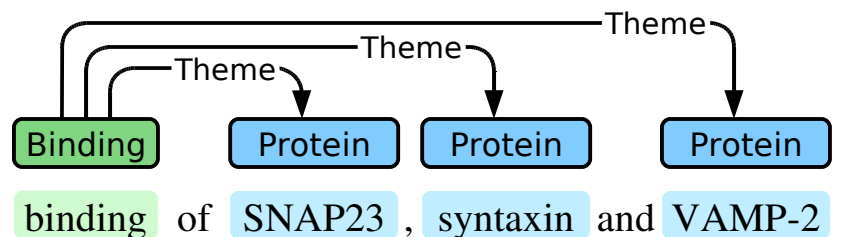
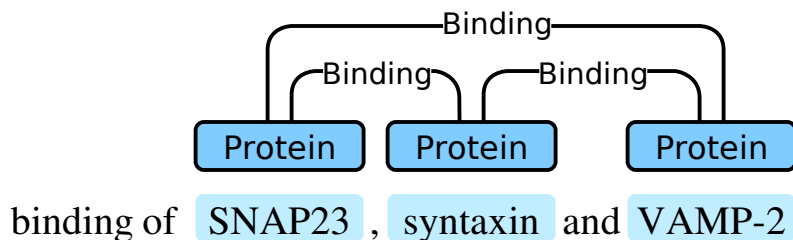
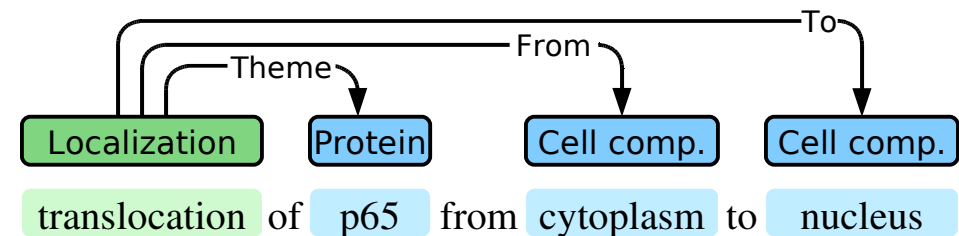
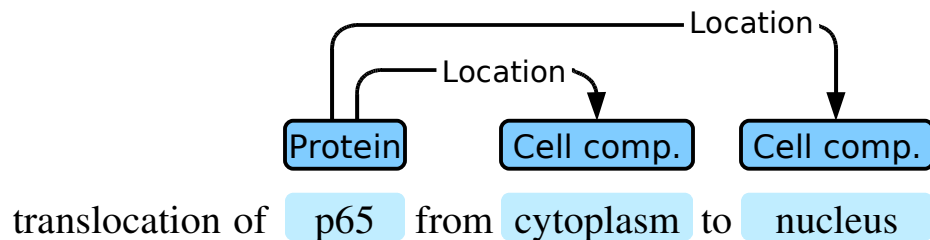
□ Events ...

- ✓ are connected to specific expressions in text
- ✓ are given types defined by an ontology
- ✓ associate variable numbers of participants in specific roles
- ✓ are primary objects of annotation and can participate in other events



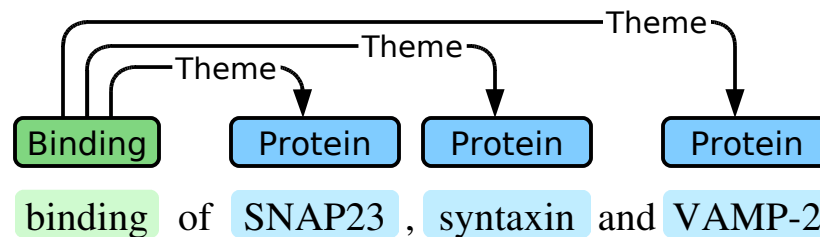
Event representation

□ Pairwise relation representation vs. event representation



Event representation

- A set of events is not equivalent to a set of trigger – participant relations
 - ✓ Alternative groupings of event participants
 - ✓ Multiple events can be triggered by a single word





Event representation

□ Event representations in biomedicine

- ✓ First roughly equivalent representation proposed as a formal extraction target by Hirschman et al. (2002)
- ✓ Corpus resources, extraction efforts and shared tasks focus on entity and relation extraction ...
 - JNLPBA (2004), LLL (2005), BioCreative (2005, 2007, 2009)
- ✓ BioInfer, first larger resource incorporating event-type annotation, published in 2007
 - ~1000 sentences, ~1500 events (“causal relationships”)
- ✓ Event annotation for GENIA, the most widely applied corpus in BioNLP, in 2008
 - ~9400 sentences, ~36000 events
- ✓ BioNLP'09 shared task on event extraction (2009)
 - First practical implementations and available systems



BioNLP'09 Task on Event Extraction

□ Nine event types (drawn from GO)

✓ Protein production and breakdown:

- Gene_expression
- Transcription
- Protein_catabolism

✓ Protein modification:

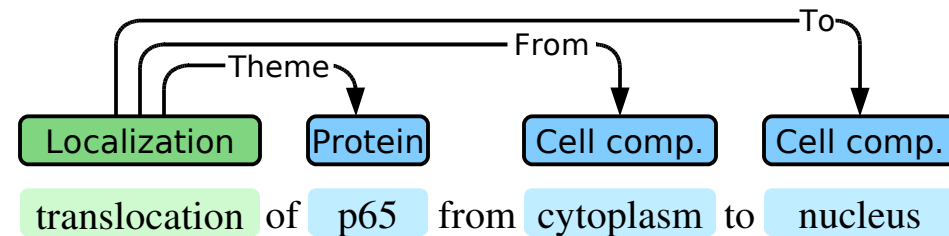
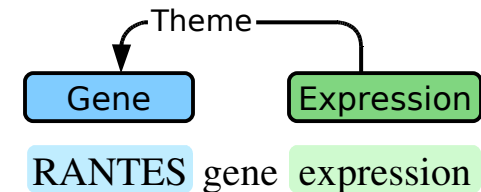
- Phosphorylation

✓ Fundamental molecular events:

- Localization
- Binding

✓ Regulatory events (+general causality):

- Regulation
- Positive_regulation
- Negative_regulation





BioNLP'09 Task on Event Extraction

- ❑ BioNLP'09 Shared Task on Event Extraction
- ❑ Entities identifying gene / protein names provided
- ❑ Three tasks:
 - ✓ Task 1: Core event extraction
 - ✓ Task 2: Event enrichment
 - ✓ Task 3: Negation and speculation recognition

we hypothesized that the phosphorylation of **TRAF2** inhibits its binding to the **CD40** cytoplasmic domain

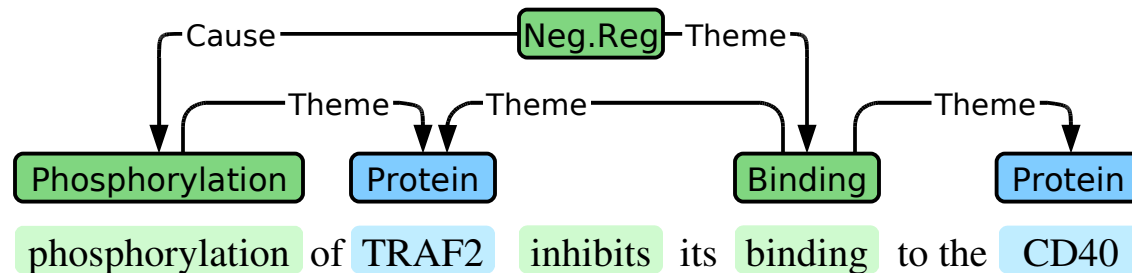
Protein

Protein



BioNLP'09 Task on Event Extraction

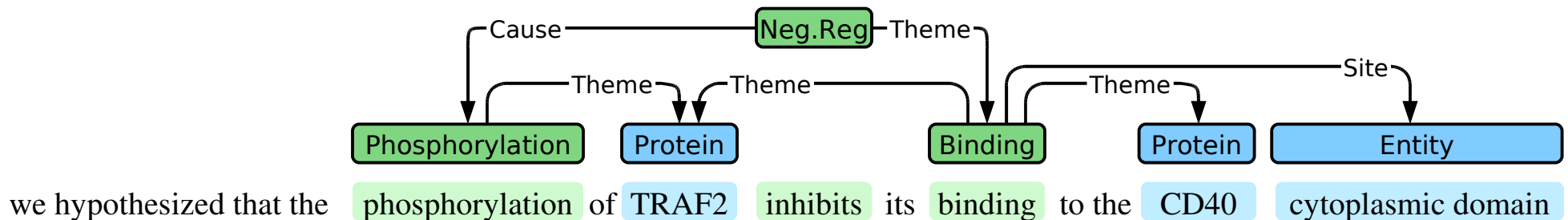
- ❑ BioNLP'09 Shared Task on Event Extraction
- ❑ Entities identifying gene / protein names provided
- ❑ Three tasks:
 - ✓ **Task 1: Core event extraction**
 - ✓ Task 2: Event enrichment (incl. non-NE entity recognition)
 - ✓ Task 3: Negation and speculation recognition





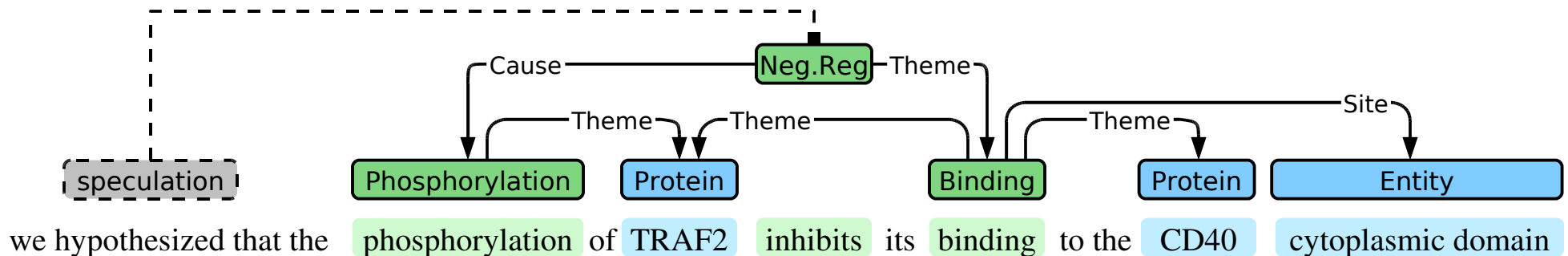
BioNLP'09 Task on Event Extraction

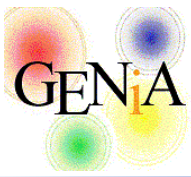
- ❑ BioNLP'09 Shared Task on Event Extraction
- ❑ Entities identifying gene / protein names provided
- ❑ Three tasks:
 - ✓ Task 1: Core event extraction
 - ✓ **Task 2: Event enrichment (incl. non-NE entity recognition)**
 - ✓ Task 3: Negation and speculation recognition



BioNLP'09 Task on Event Extraction

- ❑ BioNLP'09 Shared Task on Event Extraction
- ❑ Entities identifying gene / protein names provided
- ❑ Three tasks:
 - ✓ Task 1: Core event extraction
 - ✓ Task 2: Event enrichment (incl. non-NE entity recognition)
 - ✓ **Task 3: Negation and speculation recognition**



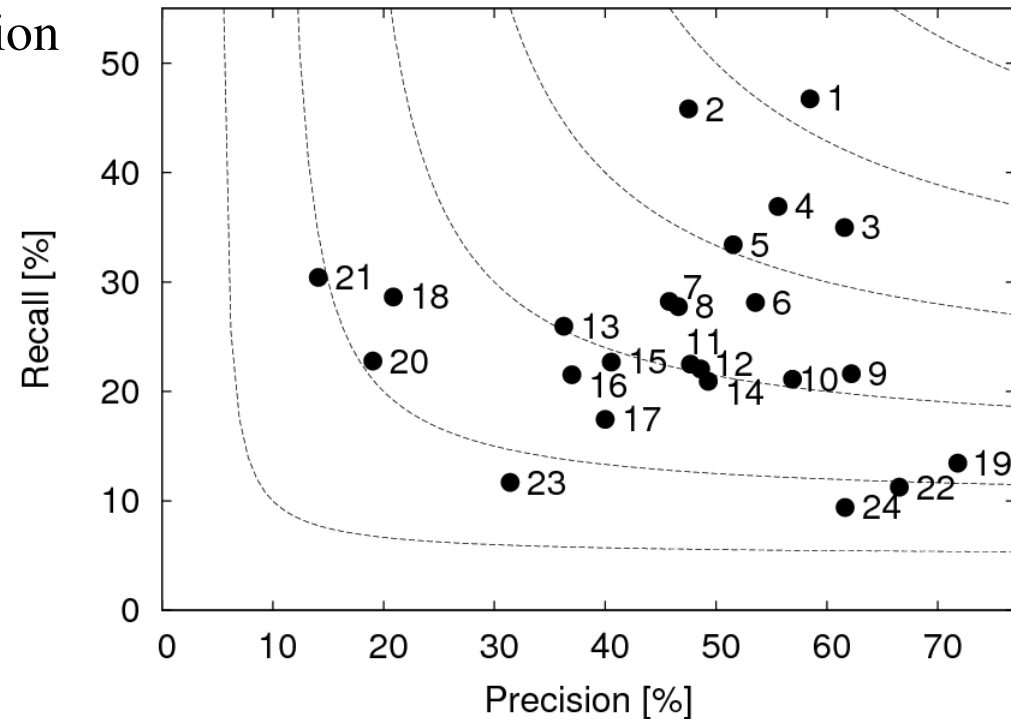


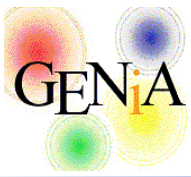
BioNLP'09 Task on Event Extraction

- 24 teams (84 researchers in total) submitted final results
 - ✓ 24 teams for task 1 (mandatory)
 - ✓ 6 teams for task 2
 - ✓ 6 teams for task 3

- Participation comparable to “general-domain” shared tasks
 - The community interested in biomedical IE is not small!

- ❑ Large spread in results
 - ✓ systems pursuing high-precision, high-recall and balanced extraction
- ❑ Task 1: best 52% F-score
 - ✓ University of Turku
- ❑ Task 2: best 43% F-score
 - ✓ U Tokyo & DBCLS
- ❑ Task 3: best 43% F-score
 - ✓ Concordia University





BioNLP'09 Task on Event Extraction

- ❑ Continued efforts on the event extraction task:
 - ✓ Special Issue of Computational Intelligence on Extracting Bio-Molecular Events from Literature (in press)
 - ✓ Numerous other studies continuing to advance the state of the art; Current best task 1 result 56% F-score (Miwa et al. 2010)
 - ➡ ~ 10% relative reduction in error from shared task best result

- ❑ Release of systems and results
 - ✓ Integration of systems with web-accessible interfaces using UIMA
 - ✓ Some systems available for download
 - ✓ Systems outputs becoming available both separately and through online search interfaces



Event representation: limitations

- ❑ The BioNLP'09 task entities
 - ✓ Continuous, non-overlapping strings
 - ✓ Gene/protein named entities (NEs) as basis of core task
- ❑ Statements of events do not necessarily (directly) involve gene/protein NEs ...
 - ✓ “X binds proximal Y promoter”
 - ✓ “X binds negative mutant of Y”
 - ✓ “X affects the IκB family of inhibitors, including IκBa”
- ❑ ... and may involve several
 - ✓ “X binds complex of c-Rel and p50”
- ❑ In cases such as these, the representation necessarily involves approximation



Event representation: limitations

- ❑ X binds dominant negative mutant of Y
 - X binds Y
- ❑ X affects F family members, including Y
 - X affects Y
- ❑ X binds complex of Y and Z
 - X binds Y
 - X binds Z

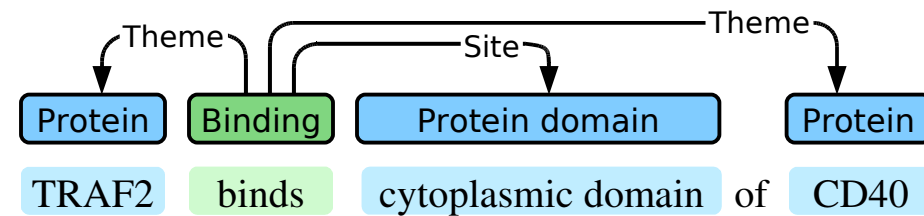
- ❑ Whether approximations of this type are acceptable depends on the application
 - ✓ As relation models preserve considerably less detail, the event model is applicable to at least tasks addressed with relations
 - ✓ ... but a generally applicable representation should offer a way to avoid the approximation

Event representation: limitations

□ Introducing more detail in the representation

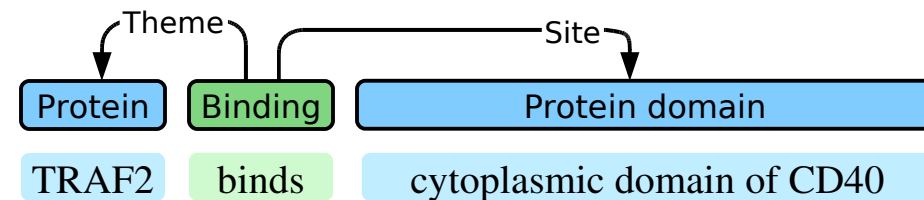
□ Additional information in events

- ✓ e.g. Binding event Site
- ✓ Already implemented ...
- ✓ ... but only a partial solution



□ Increase entity extent

- ✓ Preserves context for humans
- ✓ Loss of NE “anchoring”
- ✓ Blurs entity semantics

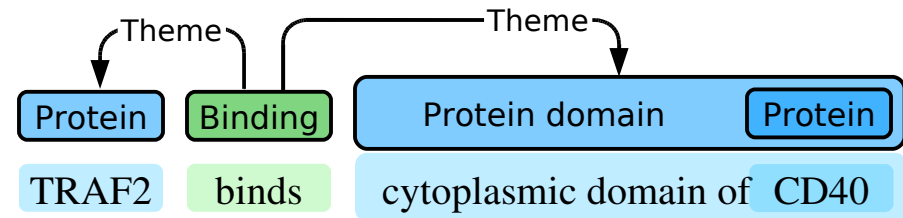


Event representation: limitations

❑ Introducing more detail in the representation

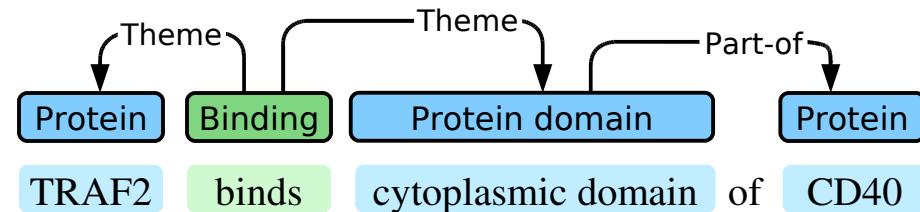
❑ Nested entities

- ✓ Preserves context for humans
- ✓ Applied in some corpora
- ✓ ... but semantics of nesting undefined



❑ Explicit entity associations

- ✓ Can define semantics
- ✓ Can supplement entity nesting
- ✓ Representation?





Outline

- Introduction and motivation
- Entities
- Relations
- Events
- Relations (again)**
- Where next?



Static relations

- ❑ “Static relations”
 - ✓ Entity relations that hold between entities without implication of change or causal connection
 - i.e. mutually exclusive with events
 - Located-in, Part-of, Member-of
- ❑ Rarely considered in biomolecular IE efforts
 - ✓ Interest in claims concerning “things that happen”
 - ✓ Not a primary target: knowledge regarding relations often either in existing databases or obvious
- ❑ ... but studied in “general domain” IE (ACE, SemEval)
 - ✓ Located-in, Part-whole



Static relations

- ❑ Static relation annotation of the GENIA corpus
 - ✓ Relations between gene/protein NEs and other terms
 - ✓ Focused subset of annotation published (BioNLP'09), full-corpus annotation in 2010 (pre-release data available on request)
- ❑ Annotated relation types
 - ✓ Equivalence
 - ✓ Variants
 - ✓ Component-Object
 - ⇒ “Complex of NE1 and NE2”
 - ✓ Object-Component
 - ⇒ “cytoplasmic domain of CD40”
 - ✓ Member-Collection
 - ⇒ “immediate-early response genes such as egr-1”

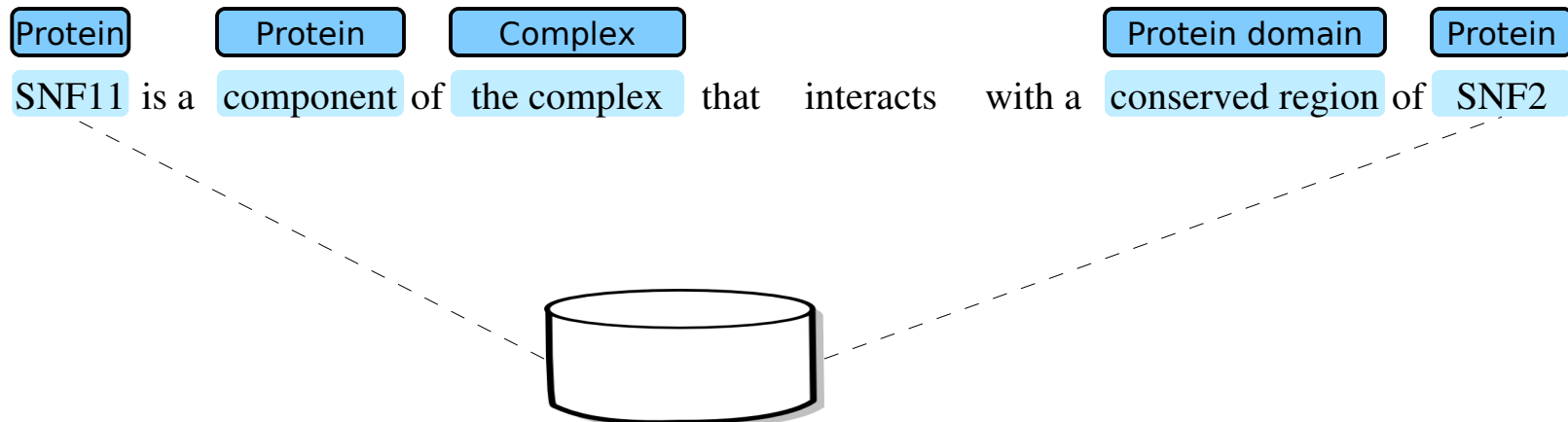
Static relations

- Connecting entities to reality

Protein SNF11 is a Protein Complex component of the complex that interacts with a Protein domain Protein conserved region of SNF2

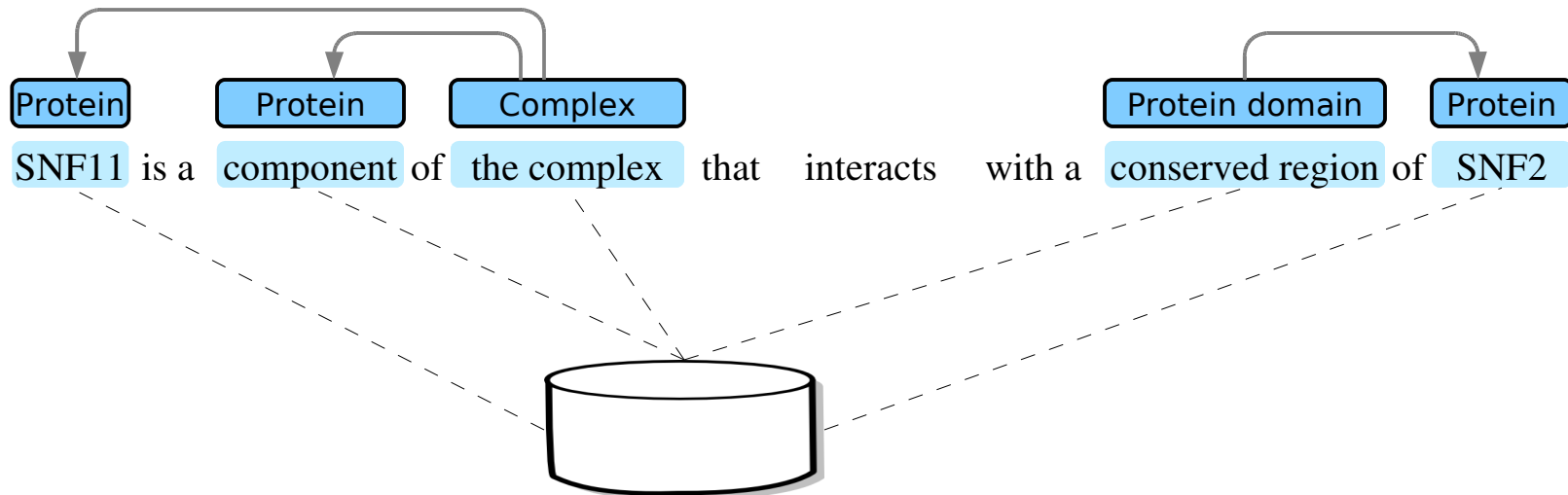
Static relations

- ❑ Connecting entities to reality
 - ✓ Named entities normalized to DB entries



Static relations

- ❑ Connecting entities to reality
 - ✓ Named entities normalized to DB entries
 - ✓ Others related through NEs and relations





A model for biomolecular semantics

□ Entities

- ✓ NEs: typed, continuous, non-overlapping strings (normalizable)
- ✓ Non-NE terms: typed, continuous, potentially nested strings

Protein **Protein** **Protein family** **Protein** **Protein** **Complex**
p27 (Kip1), a cyclin-dependent kinase inhibitor, binds to the cyclin E / CDK2 complex

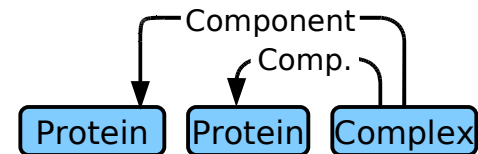
A model for biomolecular semantics

□ Entities

- ✓ NEs: typed, continuous, non-overlapping strings (normalizable)
- ✓ Non-NE terms: typed, continuous, potentially nested strings

□ Relations

- ✓ Typed, directed binary associations of entities in fixed roles
- ✓ Not separately bound to text, do not participate in relations



p27 (Kip1), a cyclin-dependent kinase inhibitor, binds to the cyclin E / CDK2 complex

A model for biomolecular semantics

□ Entities

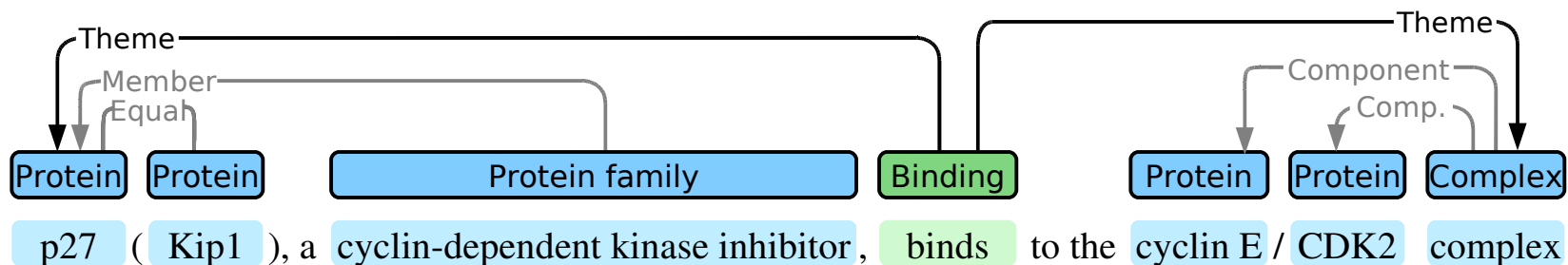
- ✓ NEs: typed, continuous, non-overlapping strings (normalizable)
- ✓ Non-NE terms: typed, continuous, potentially nested strings

□ Relations

- ✓ Typed, directed binary associations of entities in fixed roles
- ✓ Not separately bound to text, do not participate in relations

□ Events

- ✓ Typed associations of variable numbers of participants in different roles
- ✓ Bound to (triggered by) specific expressions in text, can participate in events



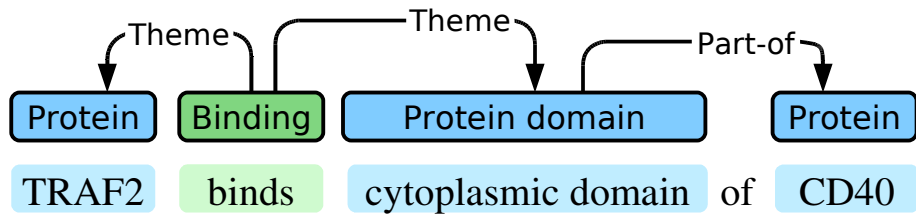


Outline

- Introduction and motivation
- Entities
- Relations
- Events
- Relations (again)
- Where next?**

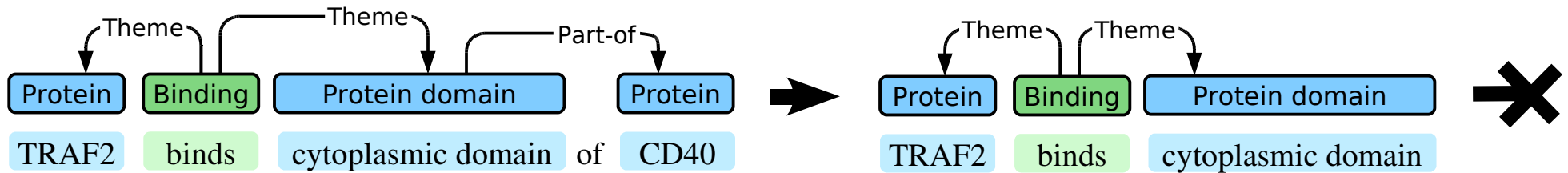
Mapping and inference

□ Mappings of the representation



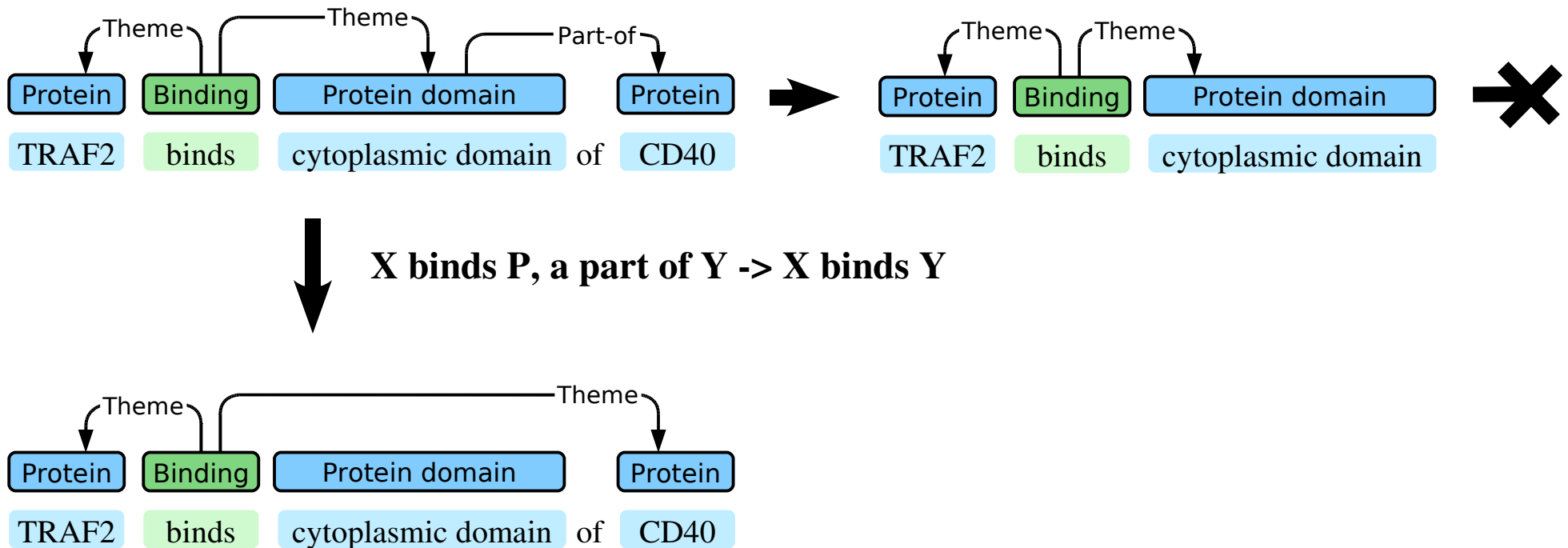
Mapping and inference

□ Mappings of the representation



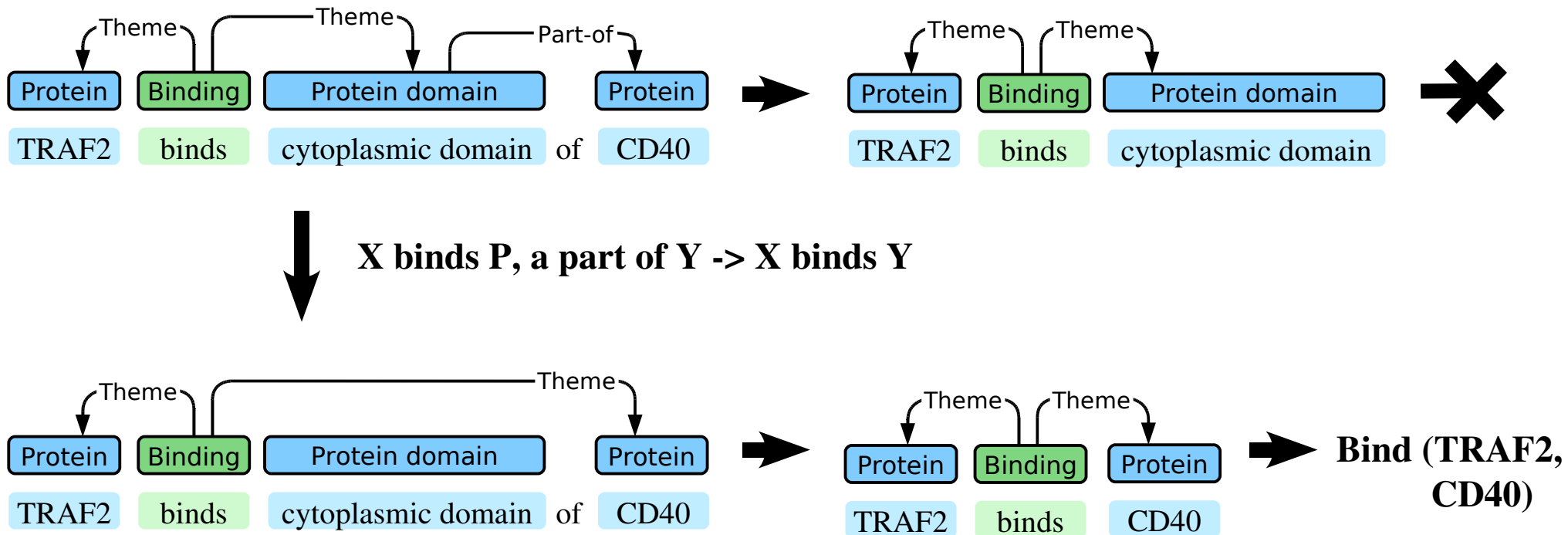
Mapping and inference

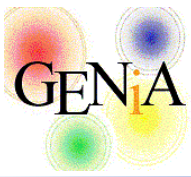
□ Mappings of the representation



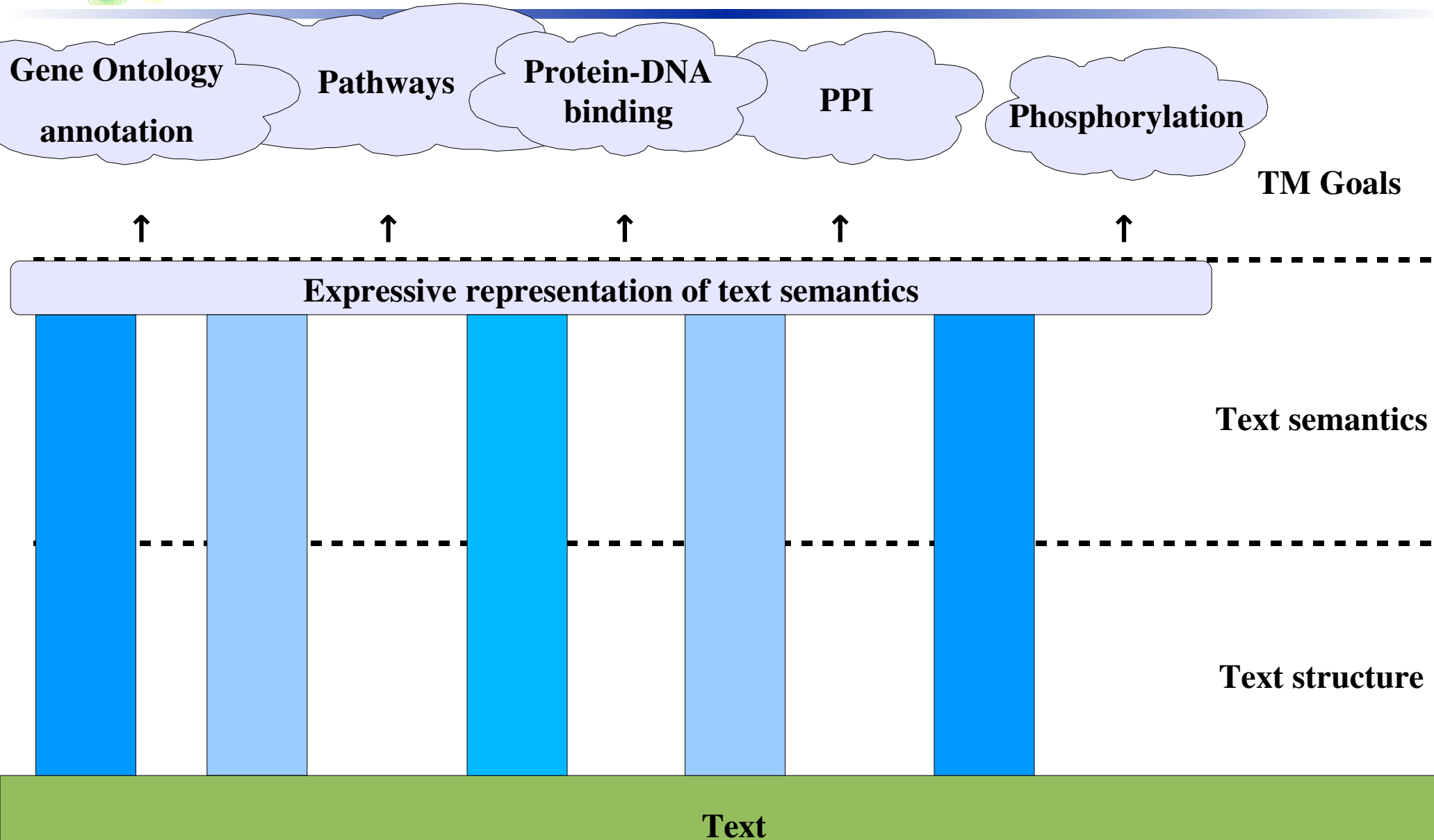
Mapping and inference

□ Mappings of the representation



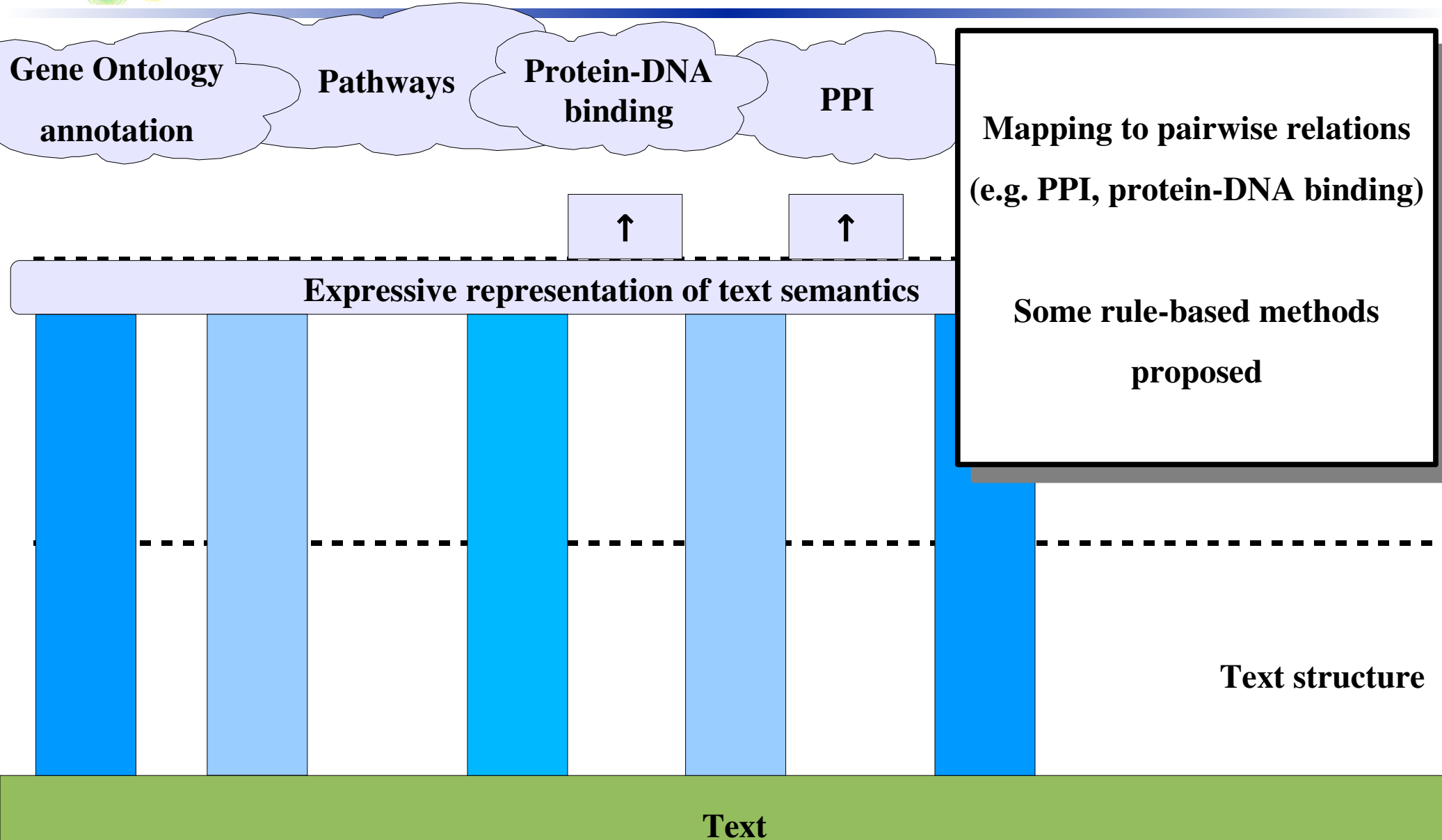


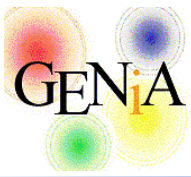
Mapping and inference



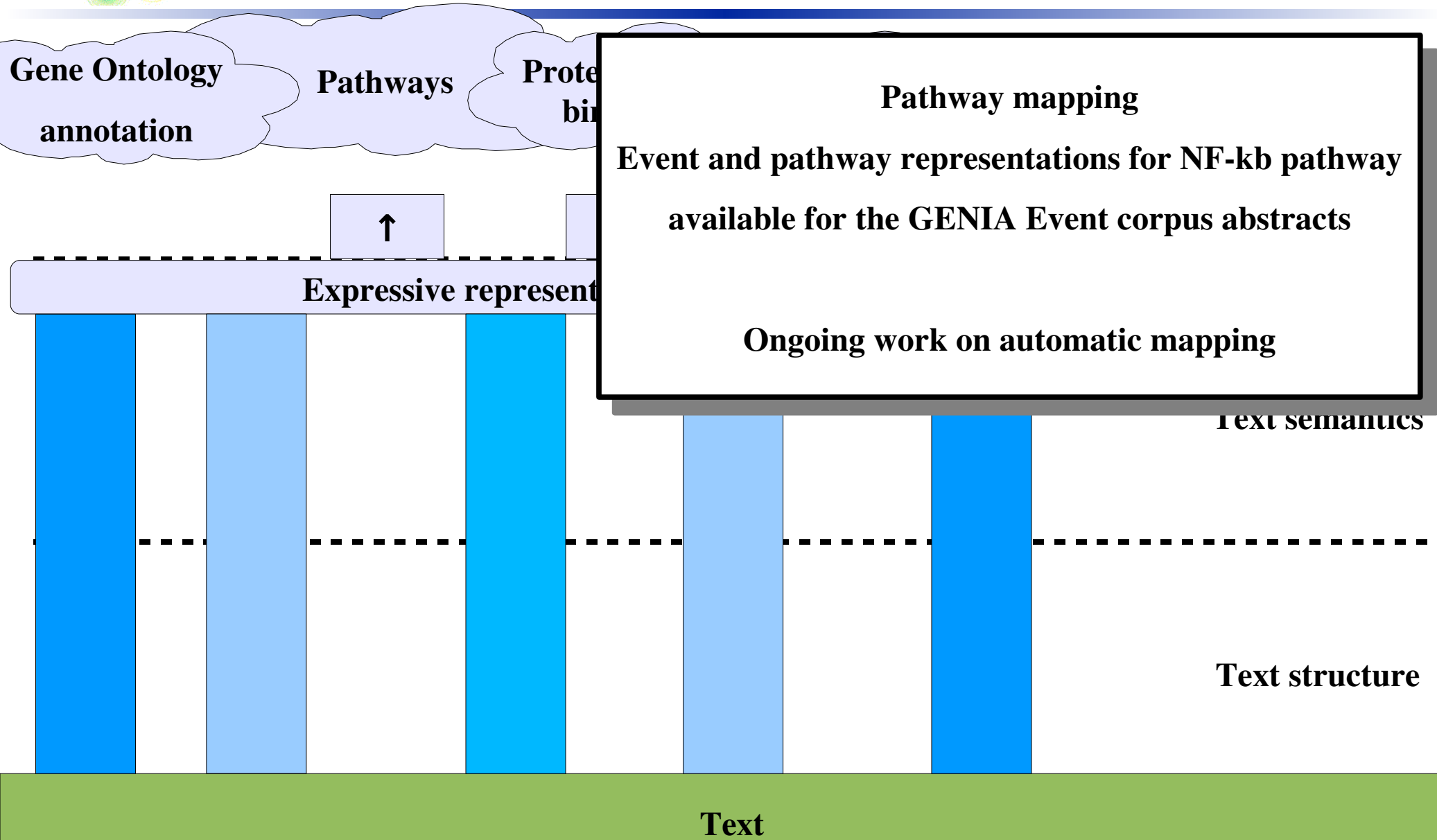


Mapping and inference

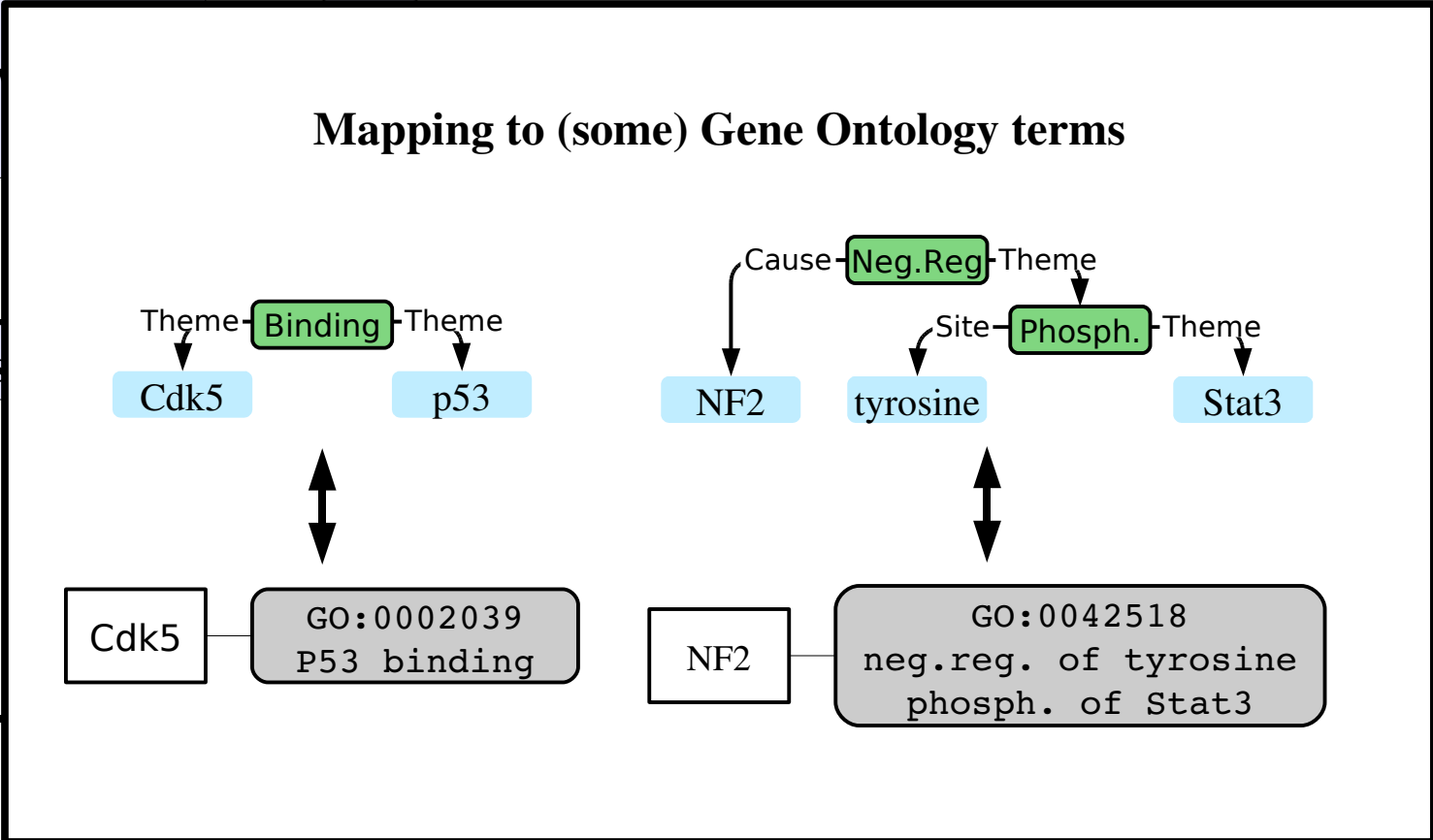
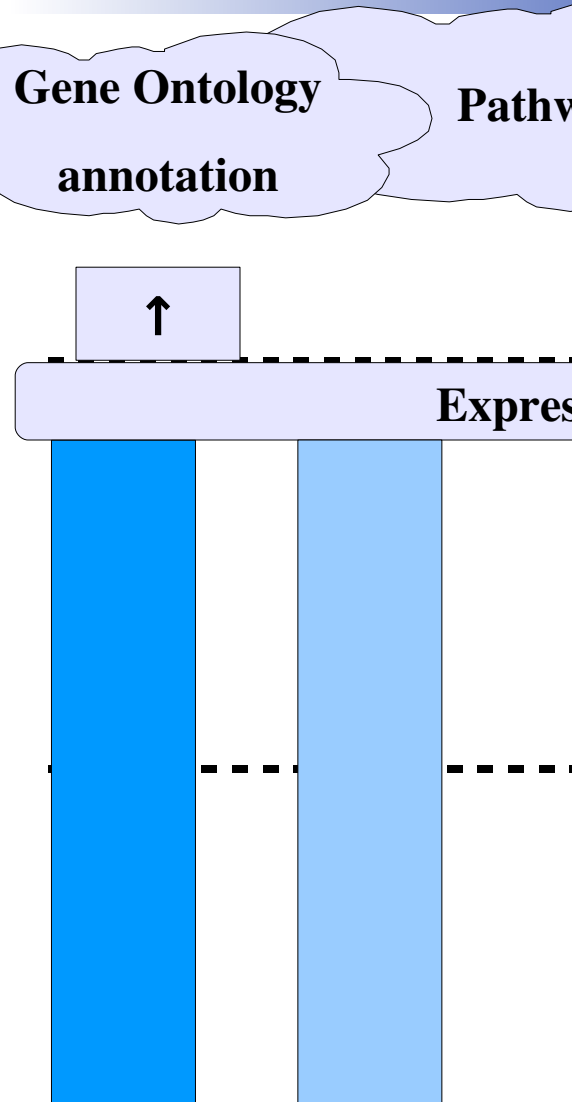




Mapping and inference

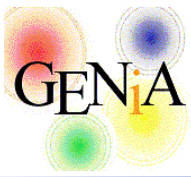


Mapping and inference



Text structure

Text



Where are we now

Entity recognition

- ✓ Protein / gene named entities ~ 90% F-score (GENETAG)
- ✓ Protein / gene normalization ~80% (single species; BioCreative)

Relations

- ✓ Static relation extraction ~70% F (BioNLP'09)

Events

- ✓ Event extraction ~55% (BioNLP'10)



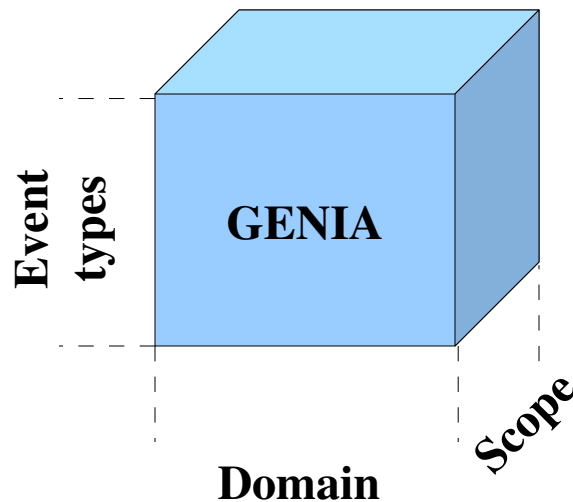
Where are we now

- ❑ Entity, Relation and Event extraction systems available
 - ✓ Only partial integration
- Integrated, retrainable systems needed

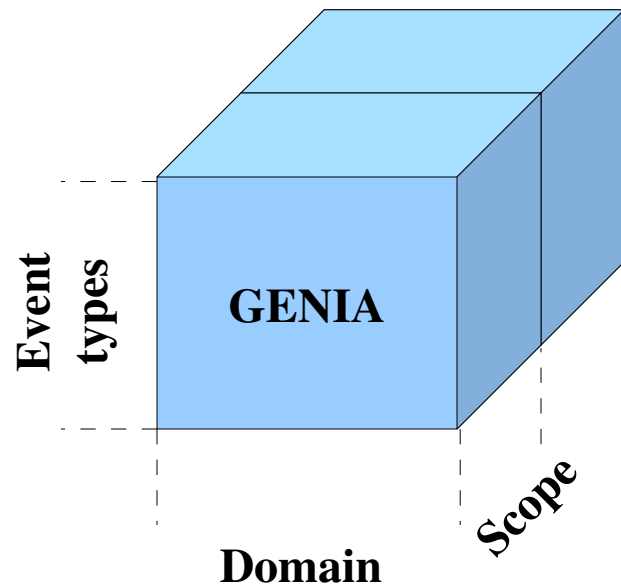
- ❑ BioNLP'09 model defines two modifications that can be assigned to events: negation and speculation
 - ✓ General representation for event properties, but limited scope
- Need to extend to identify e.g. the experimental support of events

Where are we now

- ❑ The scope of the data used in recent event extraction studies (GENIA corpus) is limited
 - ✓ Abstracts, not full texts
 - ✓ Specific subdomain of bio publications
 - ✓ Event types selected for subdomain

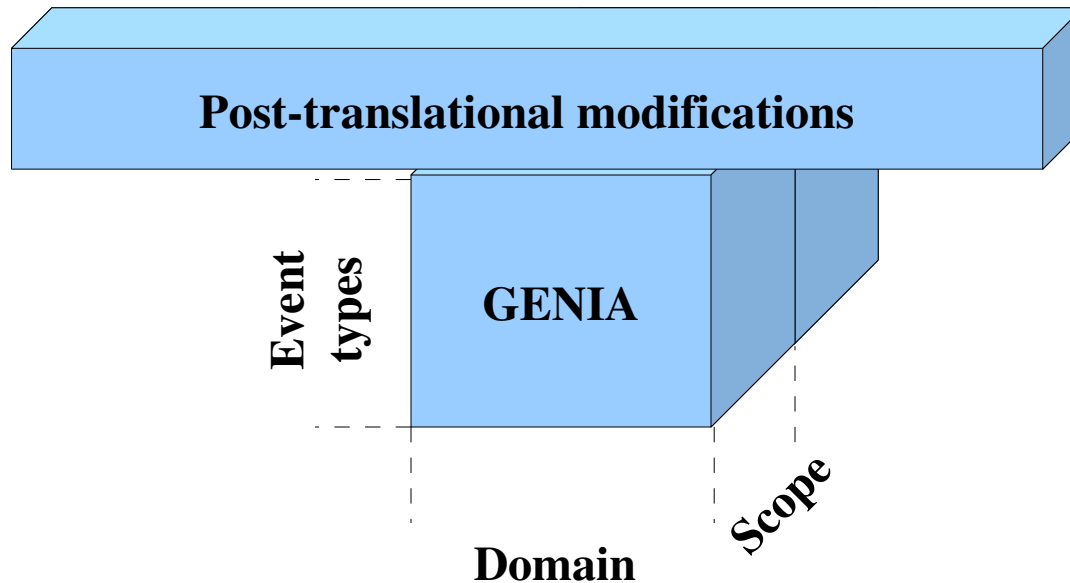


- ❑ GENIA full texts (corpus in preparation)
 - ✓ Full text data
 - ✓ GENIA event types
 - ✓ GENIA domain



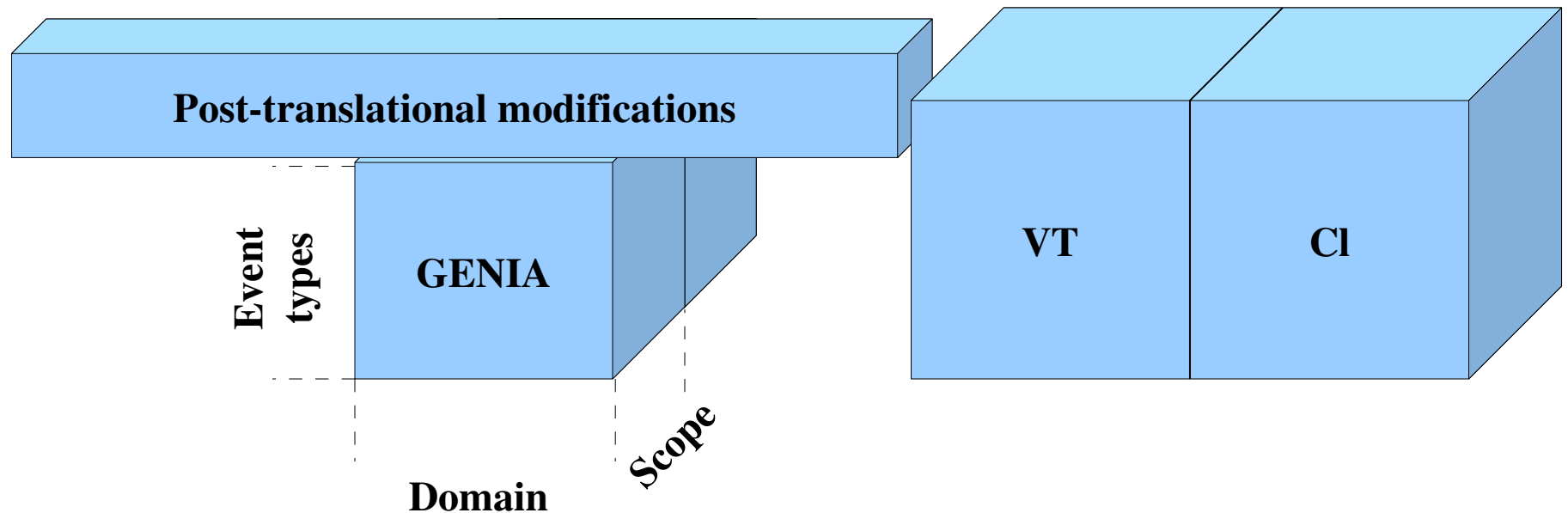
□ GENIA PTM corpus

- ✓ Annotation of selected post-translational modification events
 - Methylation, Acetylation, Glycosylation, ...
 - Together with existing annotation ~90% mention-level coverage of PTM
- ✓ Documents selected for PTM, otherwise open-domain
- ✓ Abstracts



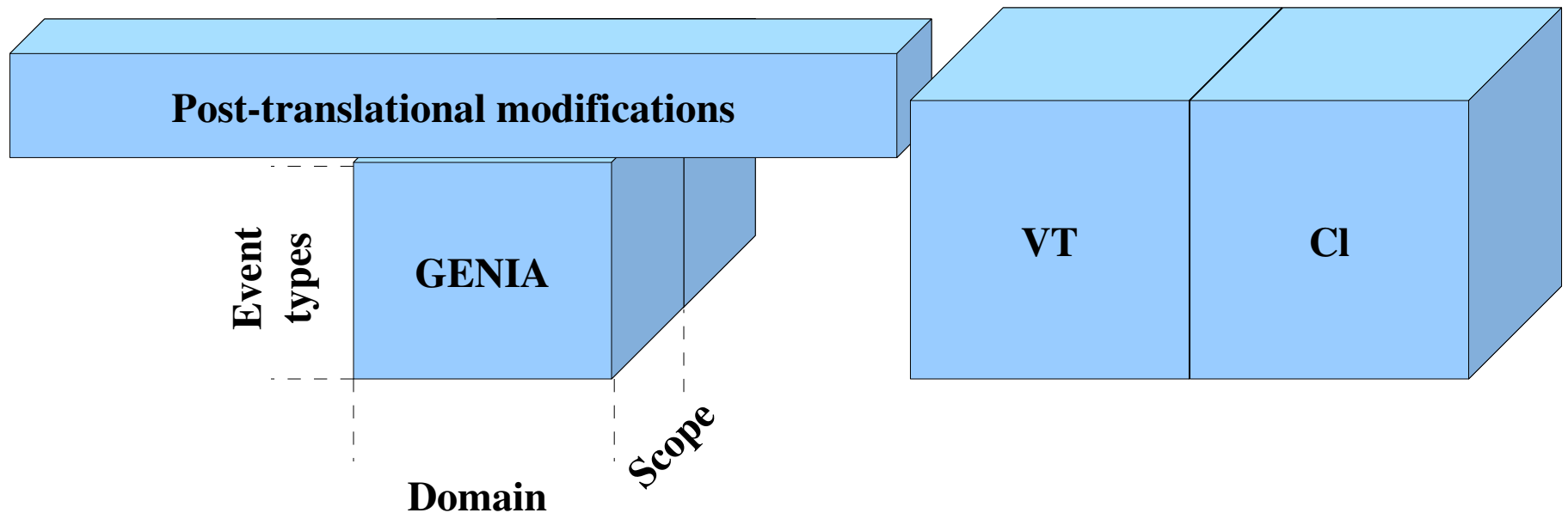
What's next

- Additional corpora
 - ✓ Independent domains
 - ✓ Extension of event types to domain
 - ✓ Full text documents



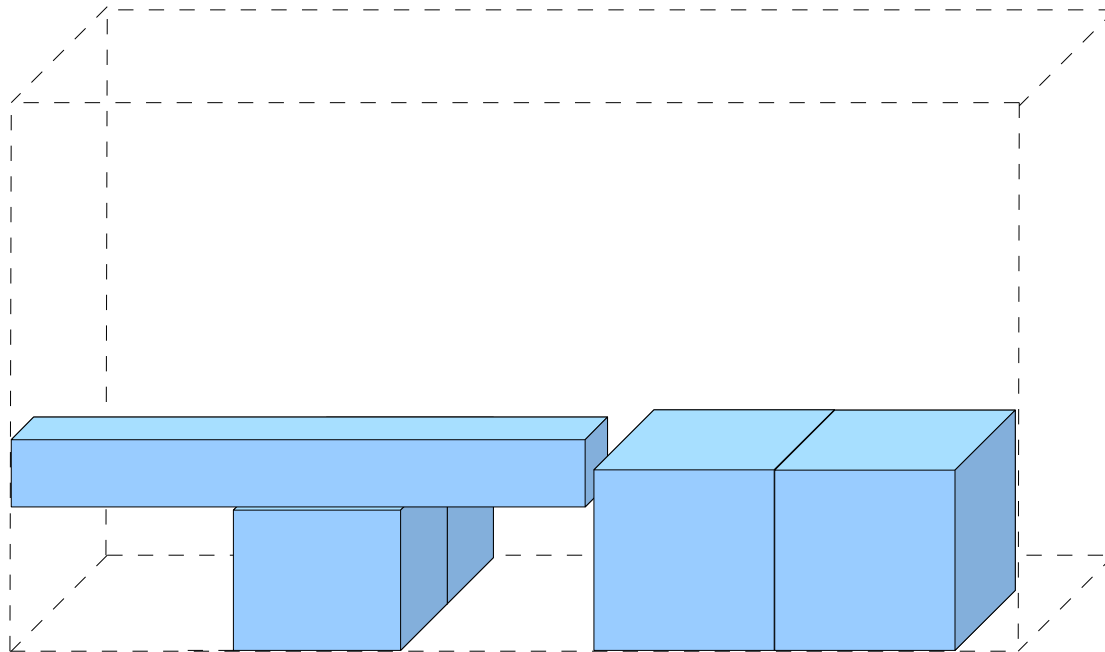
What's next

- ❑ BioNLP'11 Shared Task on Event Extraction (tentative)
 - ✓ Task definitions & sample data fall 2010
 - ✓ Training data available late 2010
 - ✓ Test data and submissions early 2011
 - ✓ **Workshop summer 2011**



What's next

- ❑ Long term: open-domain extraction of all major biomolecular events typed from full-text publications





Acks & thanks

- BioInfer & GENIA teams
- BioNLP'09 Shared Task participants
- BioTM'10 organizers

THANK YOU FOR LISTENING!