A computational method for the extraction of pharmacogenomic relationships from text

Adrien Coulet[1,2], Nigam Shah[2], Yael Garten[2], Mark Musen[2], Russ Altman[2]

1 LORIA, INRIA Nancy – Grand-Est

2 Stanford University

The NCBO and PharmGKB

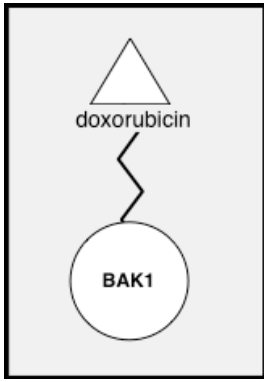
- A joint project

NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY & PharmGKB

- Content of PharmGKB
  - Current:

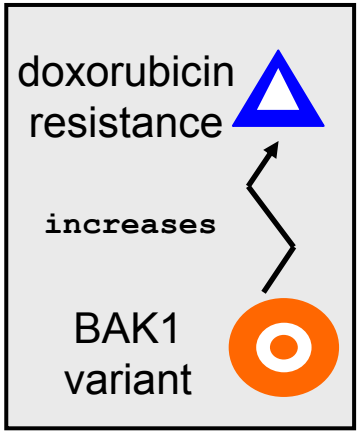
    pharmacogenomics (PGx) relationships

    Gene – Drug ; Gene – Disease ; Drug – Disease



  - Goal:

    to provide more precise relationships



Population of PharmGKB

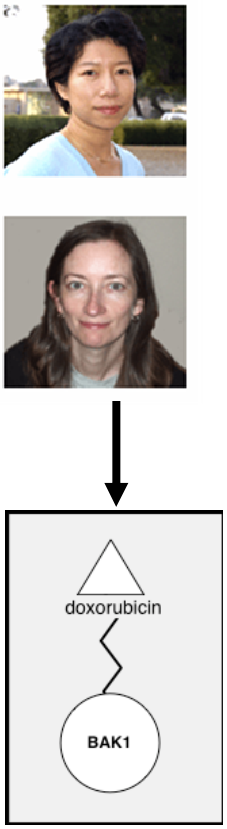
Sentence 1 : Doxorubicin induces BAK1 activity.

Sentence 2 : Doxorubicin transcriptionally activates BAK1.

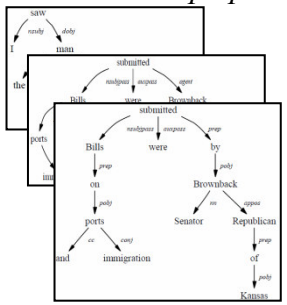
Sentence 3 : BAK1 gene polymorphism affects doxorubicin resistance.

Scientific literature

PharmGKB curators

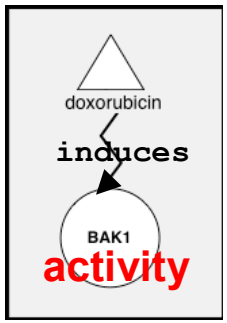


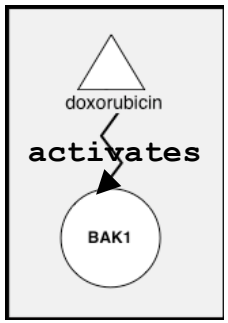
Dependency Graph parsing

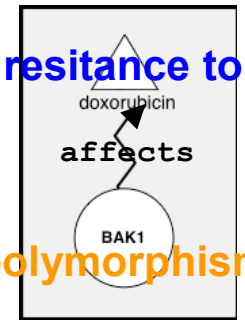


Dependency Graphs of sentences

Relation extraction







Population of PharmGKB

Sentence 1 : Doxorubicin induces BAK1 activity.

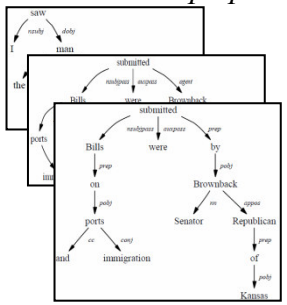
Sentence 2 : Doxorubicin transcriptionally activates BAK1.

Sentence 3 : BAK1 gene polymorphism affects doxorubicin resistance.

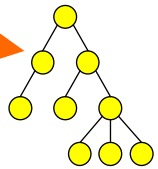
Scientific literature

PharmGKB curators

Dependency Graph parsing

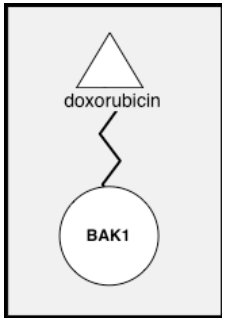


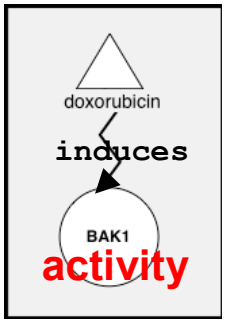
Dependency Graphs of sentences

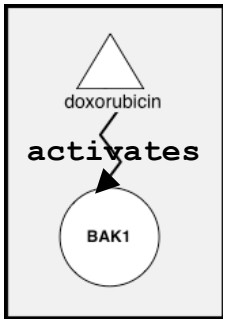


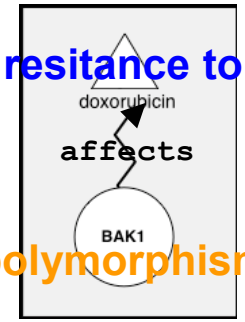
ontology

Relation extraction









Outline

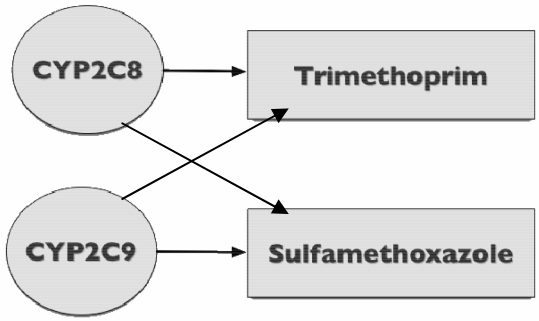
1. Limitations of co-occurrences

2. Construction of semantic network
    1. Algorithm to extract raw relationships
    2. Semi-automated ontology building
    3. Comprehensive knowledge network from 1 & 2

Limitations of co-occurrence (that we wanted to solve)

1. Avoid false positive connections

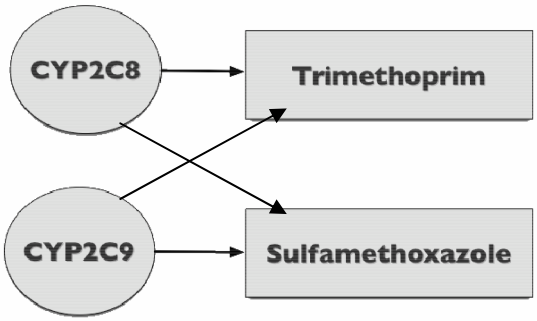
"Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity."



Limitations of co-occurrence (that we wanted to solve)

1. Avoid false positive connections

"Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity."



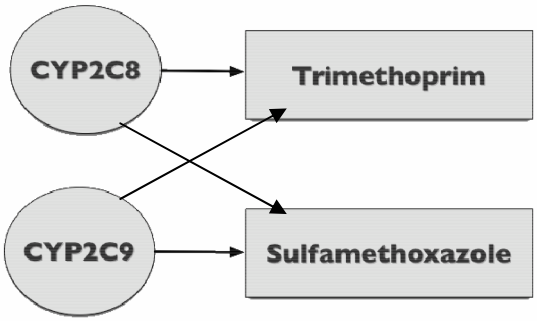
2. Characterize fine-grain semantics of relationships

"CYP3A4 mRNA expression was increased significantly by rifampicin exposure in human hepatocytes."

Limitations of co-occurrence (that we wanted to solve)

1. Avoid false positive connections

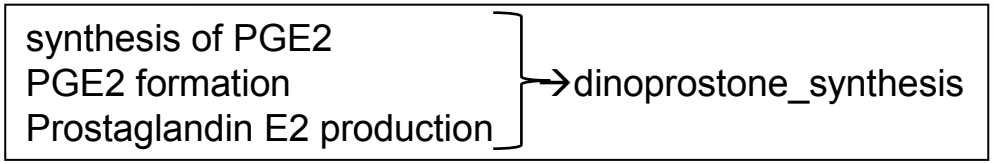
"Trimethoprim inhibits activity of CYP2C8 while sulfamethoxazole inhibits CYP2C9 activity."



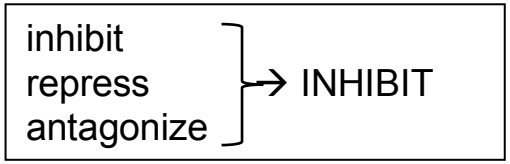
2. Characterize fine-grain semantics of relationships

"CYP3A4 mRNA expression was increased significantly by rifampicin exposure in human hepatocytes."

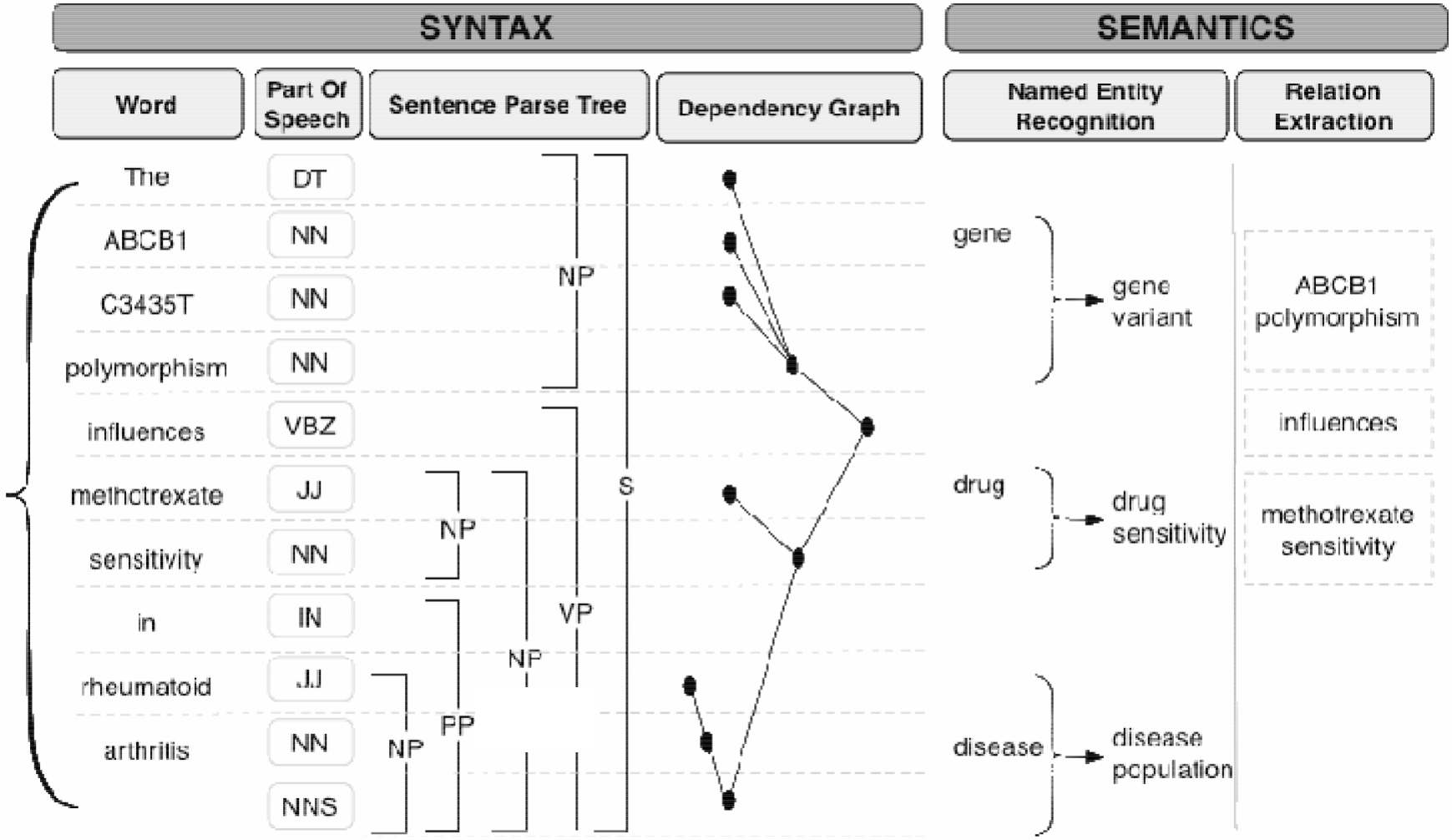
3. To consolidate synonyms (normalize):
   - Between complex entity names:



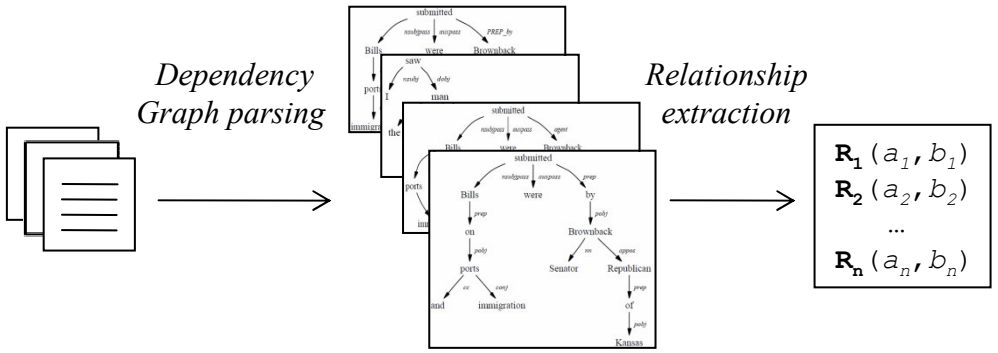
   - Between relationships:



Several steps of text processing enable extracting relationship semantics



The method extracts high quality typed relationships



MEDLINE abstracts

~17,000,000 abstracts

Dependency Graphs of sentences

~87,000,000 dependency graphs

Raw relationships

~41,000 raw relationships

Evaluation:
Randomly selected 220 raw relationships: classified into 3

"polymorphisms in VKORC1 are associated with warfarin dose."

•associated(VKORC1_polymorphisms,warfarin_dose)
= true and complete

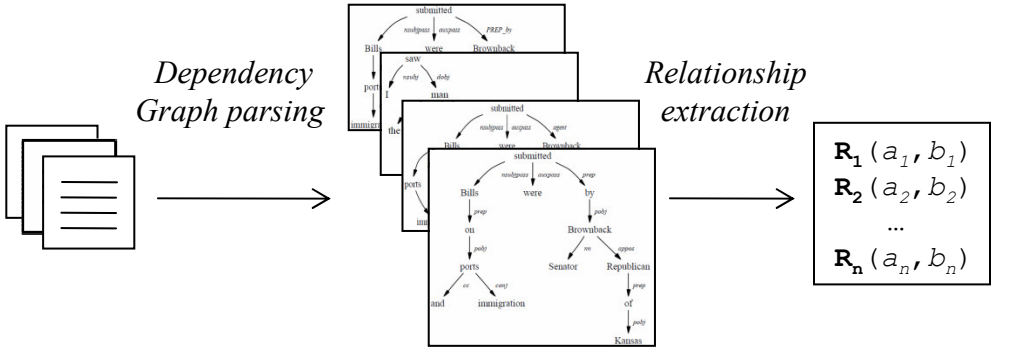
• associated (VKORC1_polymorphisms, warfarin)
= true and incomplete

•polymorphisms (VKORC1, warfarin_dose)
= false

Results:

- 87.7% were complete or incomplete true positives
  - 70% true and complete
  - 17.7% true and incomplete
- 12.3% were false positives

Issue: we extracted heterogeneous relationships



MEDLINE abstracts

Dependency Graphs of sentences

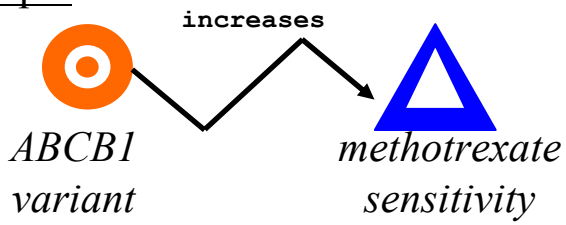
Raw relationships

~17,000,000 abstracts

~87,000,000 dependency graphs

~41,000 raw relationships

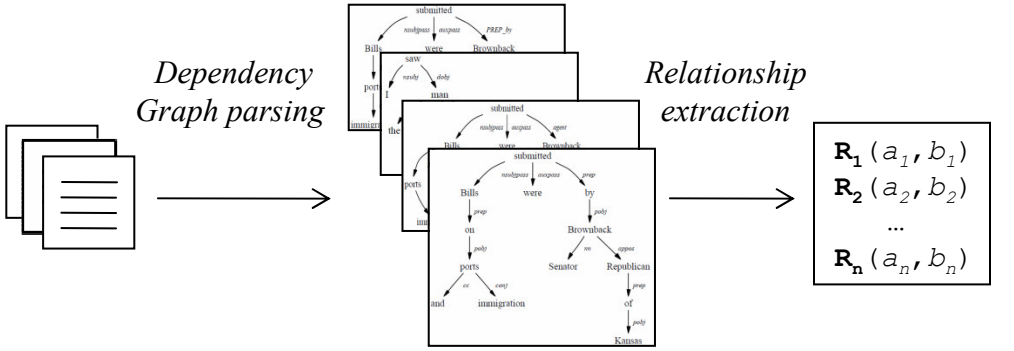
Example:



increase(P-gp_variants, methopterin_intolerance)

augment(ABCB1_SNPs, methotrexate_sensitivity)

Issue: we extracted heterogeneous relationships



MEDLINE abstracts

Dependency Graphs of sentences

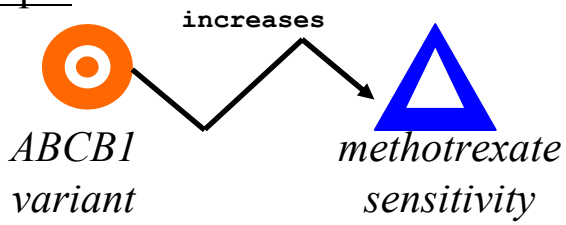
Raw relationships

~17,000,000 abstracts

~87,000,000 dependency graphs

~41,000 raw relationships

Example:

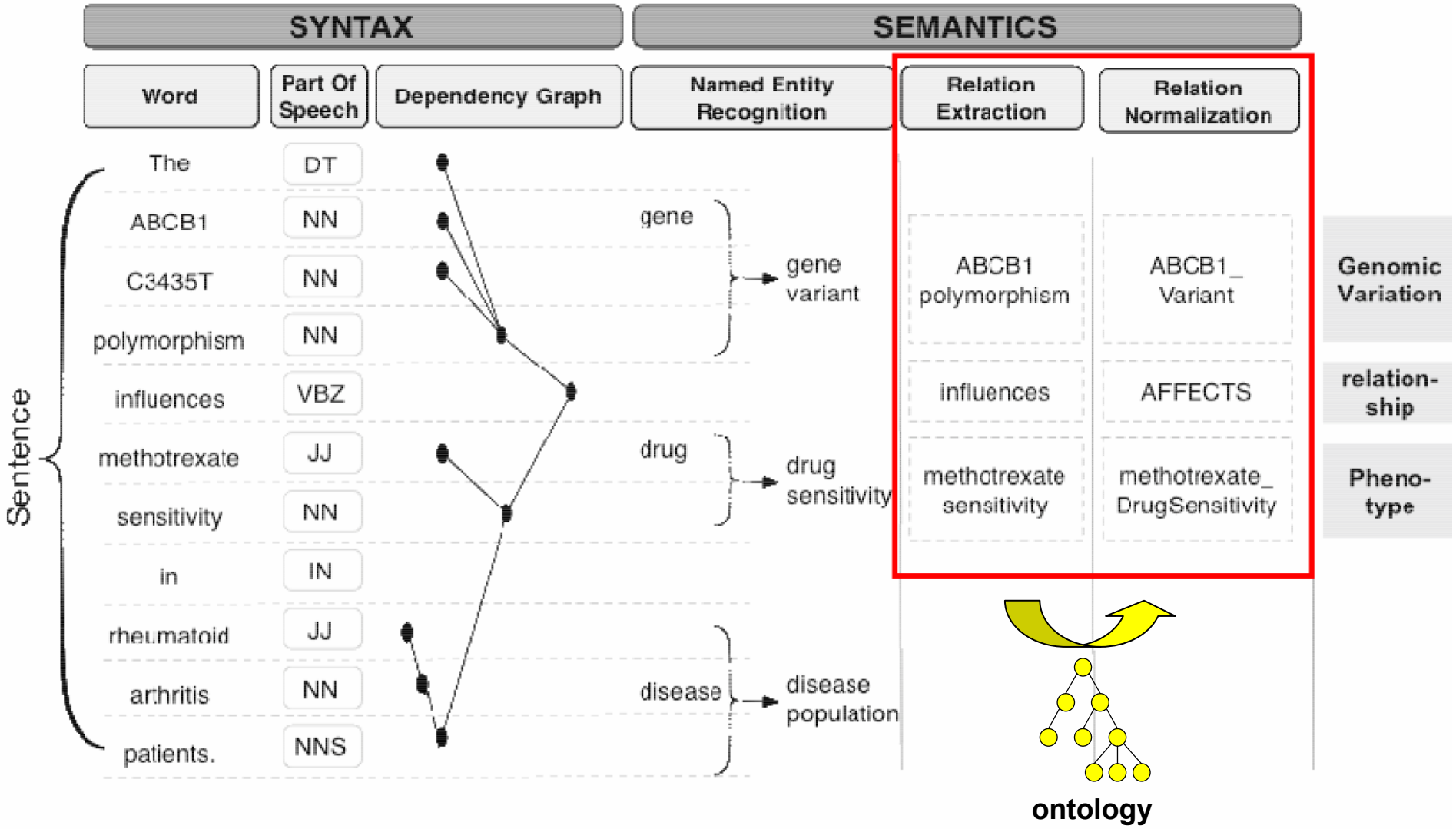


increase(P-gp_variants, methopterin_intolerance)

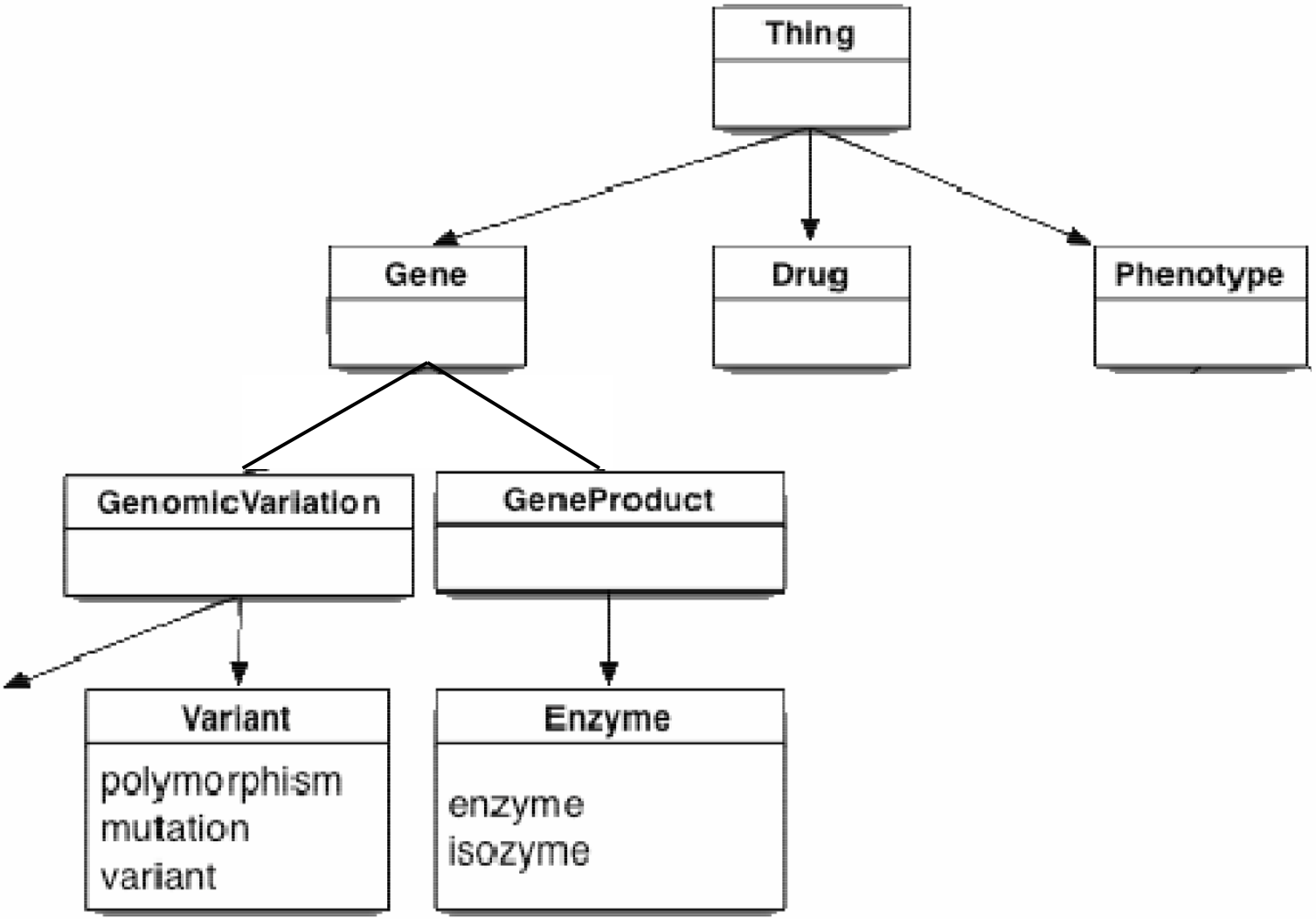
augment(ABCB1_SNPs, methotrexate_sensitivity)

- There is no relation ontology for most of specialized domains
- We created one from extracted relationships

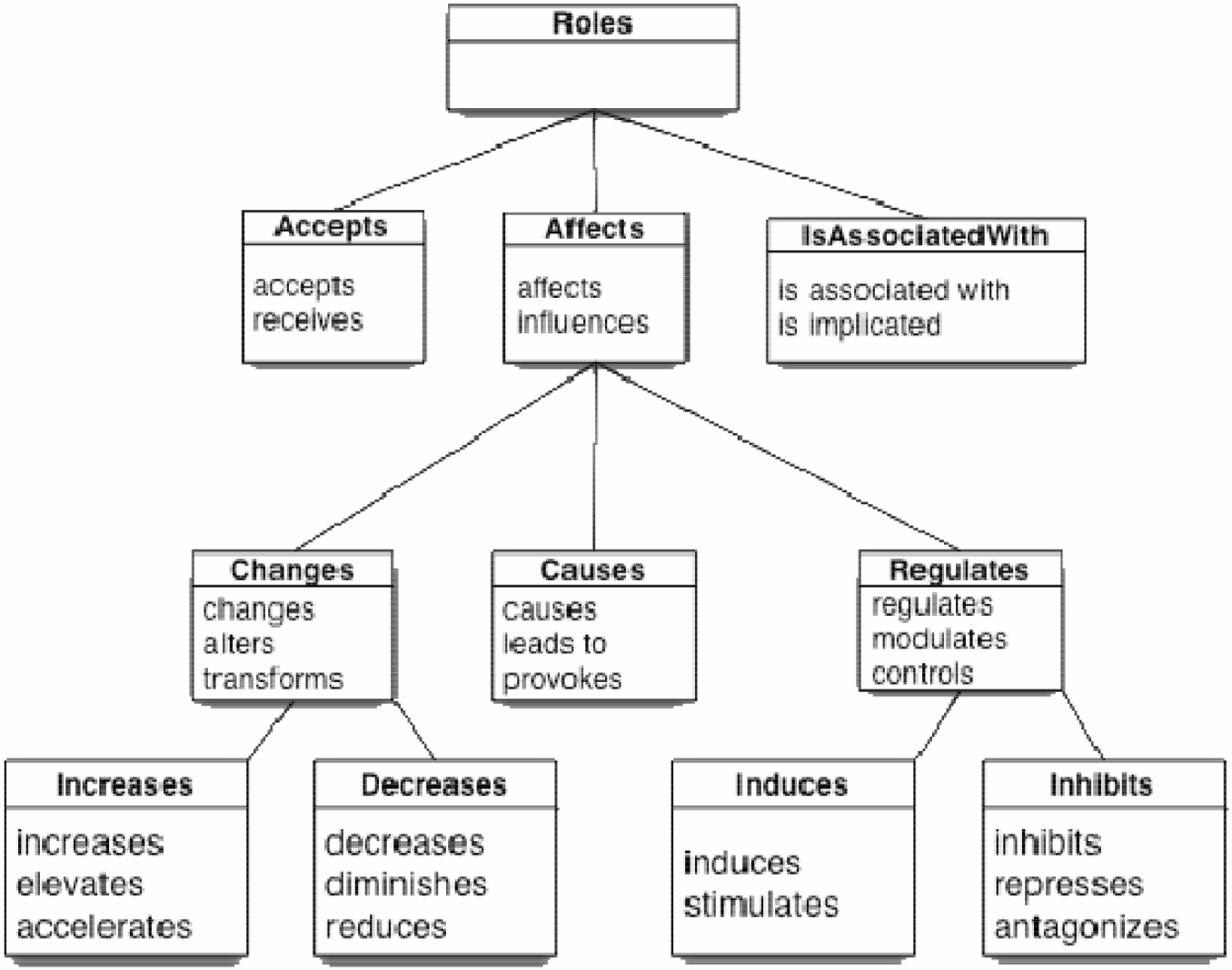
We built and use an ontology to normalize relationships



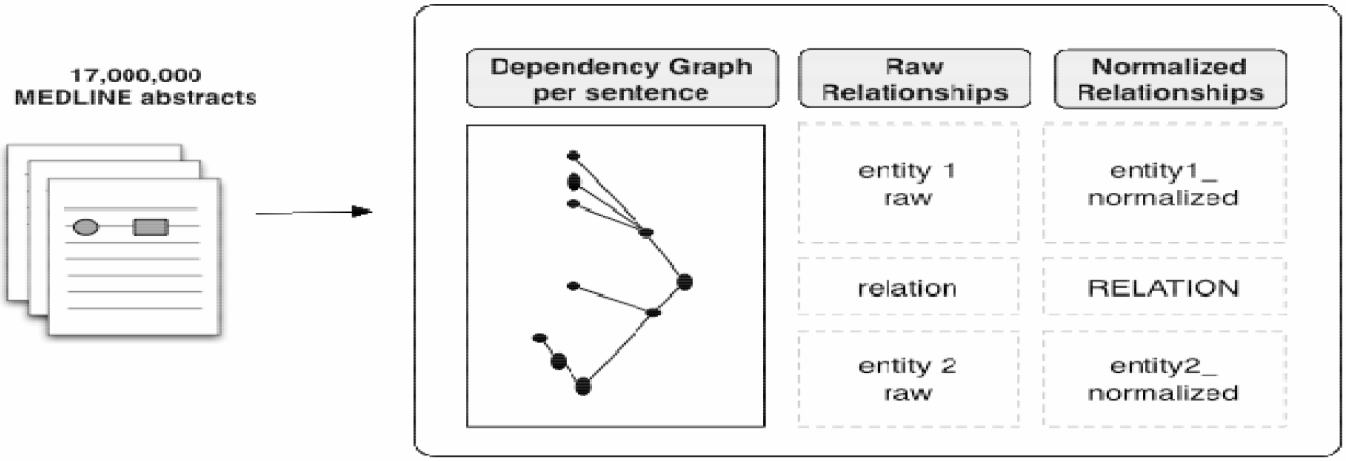
An ontology organizes the "world" into concepts and roles (1/2)



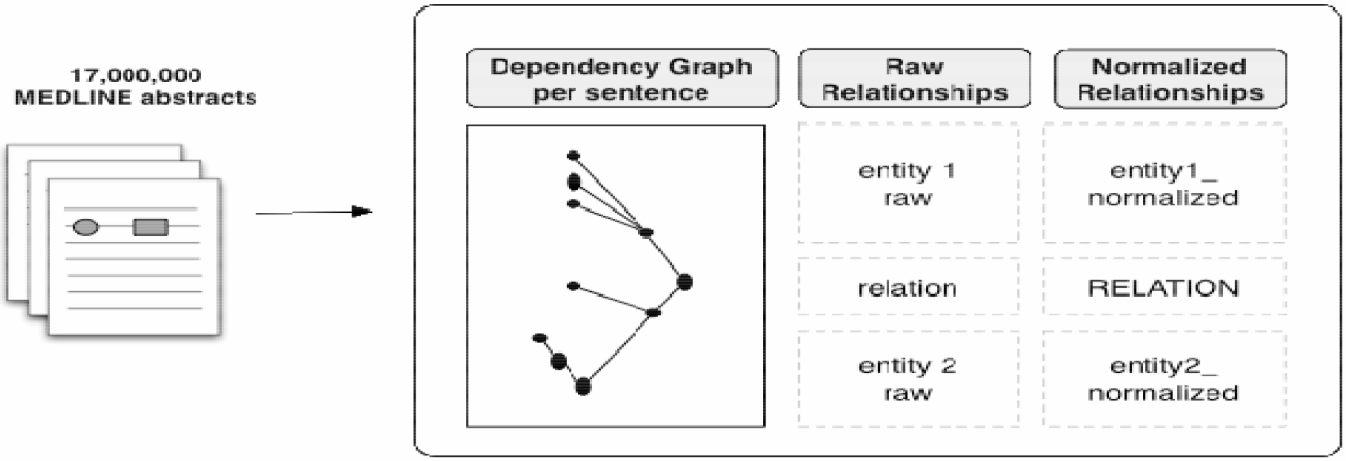
An ontology organizes the "world" into concepts and roles (2/2)



We manually created a PGx ontology "bottom-up"

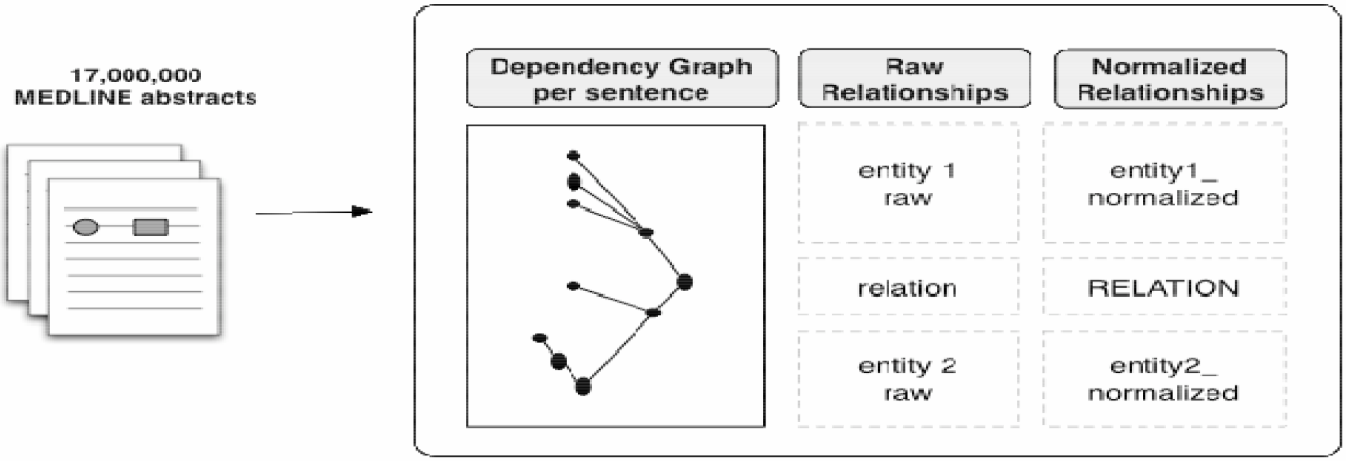


We manually created a PGx ontology "bottom-up"



Relationship types	Entities modified by		
	Genes	Drugs	Phenotypes
2538 associate	1237 *gene*	377 *metabolism*	304 *cell*
1017 increase	1000 *inhibitor*	358 *activity*	114 *line*
985 inhibit	935 *polymorphism*	298 *inhibitor*	101 *patient*
825 induce	775 *expression*	267 *effect*	71 *risk*
763 metabolize	773 *activity*	263 *administration*	35 *tissue*
666 involve	689 *mutation*	246 *channel*	34 *specimen*
643 reduce	685 *genotype*	242 *treatment*	33 *case*
547 catalyze	393 *inhibition*	193 *antagonist*	27 *treatment*
515 cause	329 *level*	178 *concentration*	26 *rate*
509 affect	245 *gene_mutation*	172 *dose*	26 *effect*

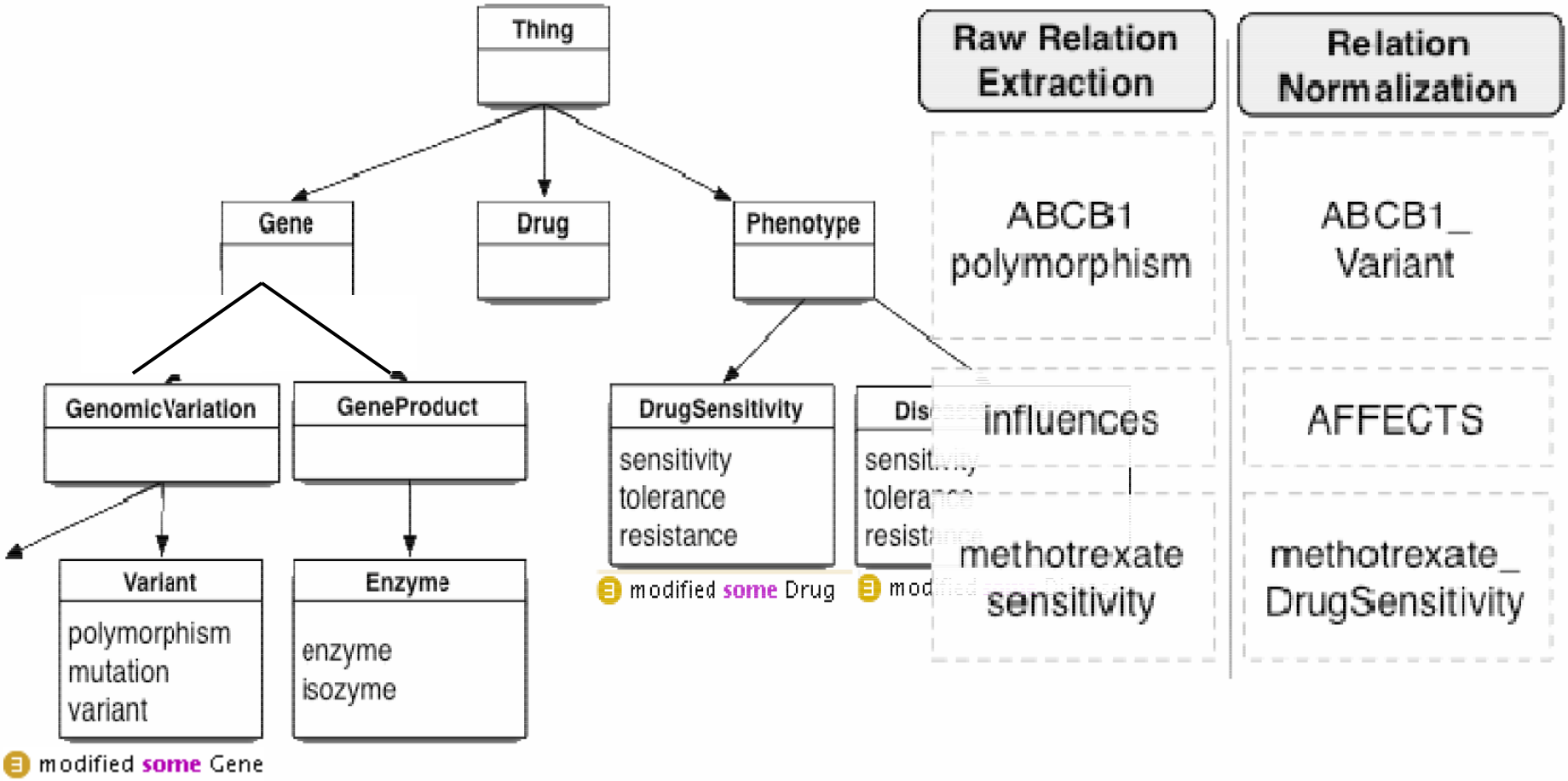
We manually created a PGx ontology "bottom-up"



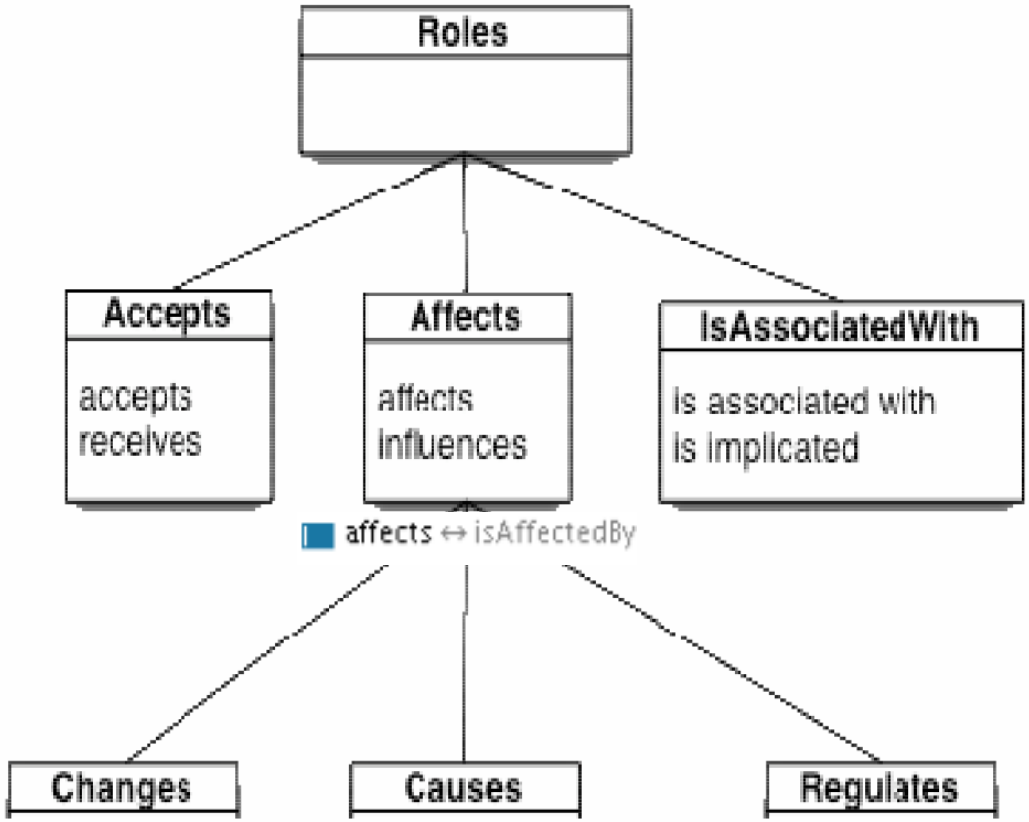
Relationship types	Entities modified by		
	Genes	Drugs	Phenotypes
2538 associate	1237 *gene*	377 *metabolism*	304 *cell*
1017 increase	1000 *inhibitor*	358 *activity*	114 *line*
985 inhibit	935 *polymorphism*	298 *inhibitor*	101 *patient*
825 induce	775 *expression*	267 *effect*	71 *risk*
763 metabolize	773 *activity*	263 *administration*	35 *tissue*
666 involve	689 *mutation*	246 *channel*	34 *specimen*
643 reduce	685 *genotype*	242 *treatment*	33 *case*
547 catalyze	393 *inhibition*	193 *antagonist*	27 *treatment*
515 cause	329 *level*	178 *concentration*	26 *rate*
509 affect	245 *gene_mutation*	172 *dose*	26 *effect*

237 concepts
76 roles

We use the ontology to normalize the raw relationship (subject, relation and object)



We use the ontology to normalize the raw relationship (subject, relation and object)

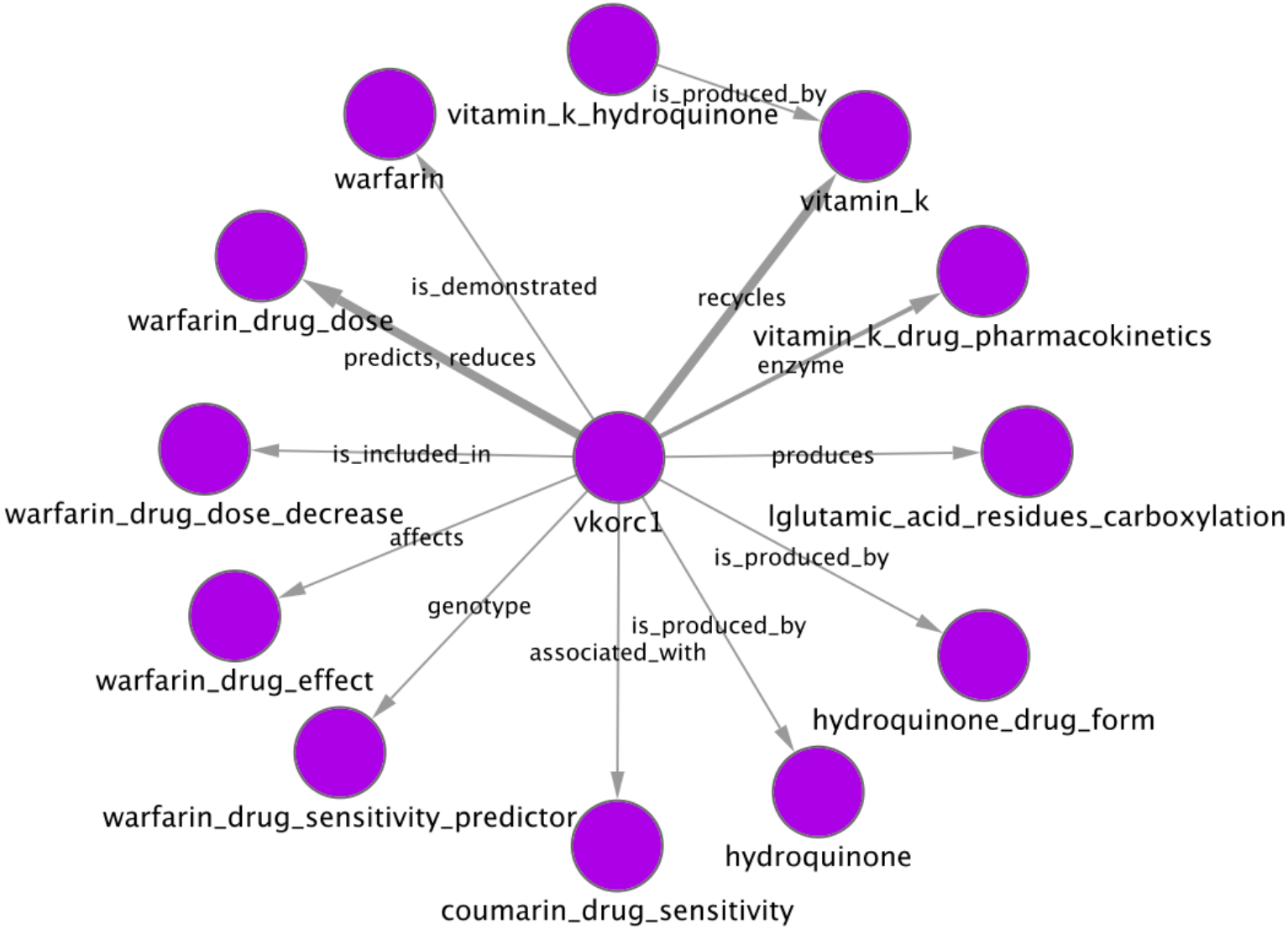


Raw Relation Extraction	Relation Normalization
ABCB1 polymorphism	ABCB1_ Variant
influences	AFFECTS
methotrexate sensitivity	methotrexate_ DrugSensitivity

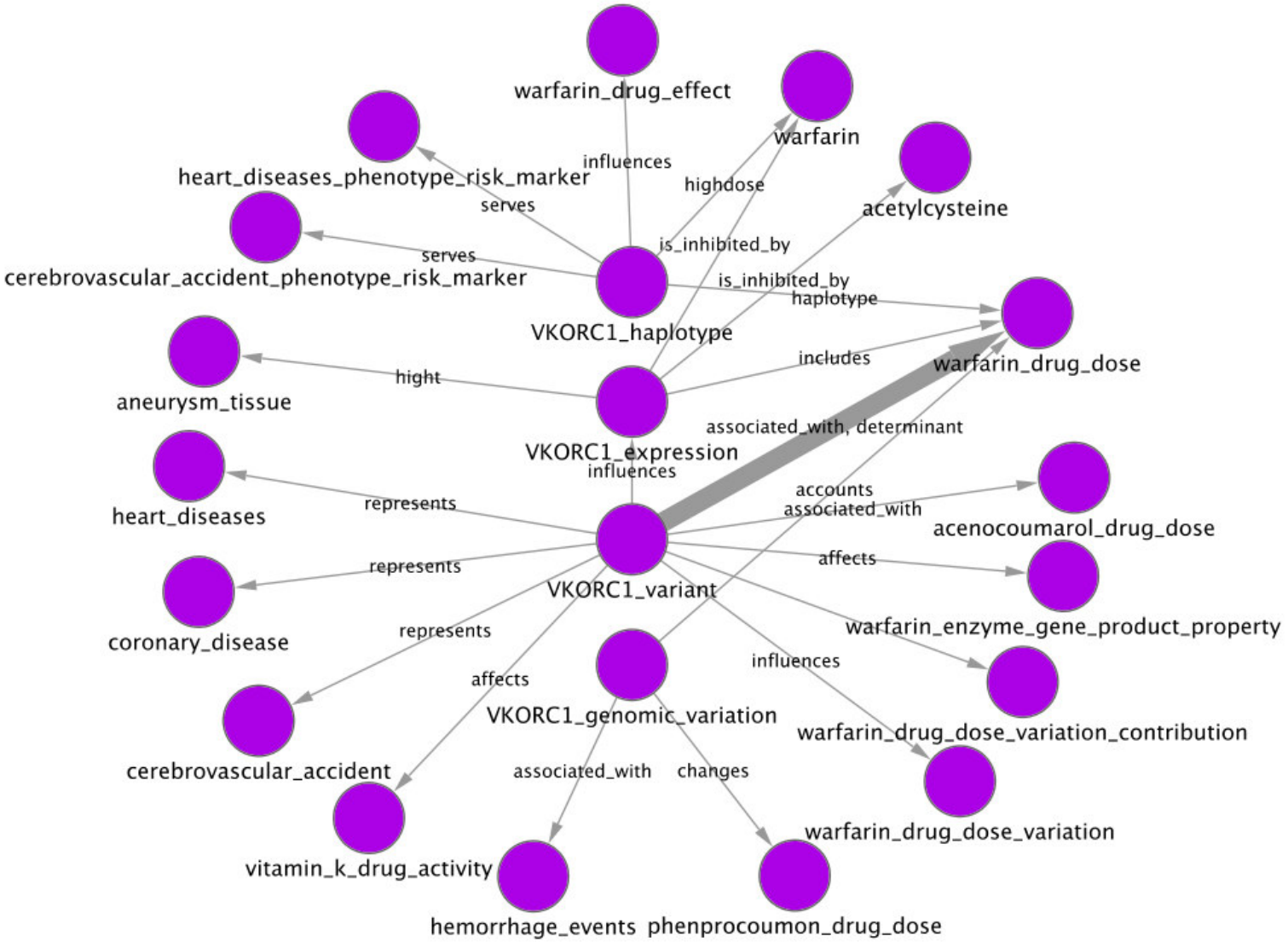
Example: two sentences but one fact

raw text	sentence	The ABCB1 C3435T polymorphism influences methotrexate sensitivity in rheumatoid arthritis patients.	A variant C3435T allele of the MDR1 gene affects methotrexate tolerability.
raw relationship	entity 1	ABCB1 polymorphism	allele of the MDR1 gene
	relationship	influences	affects
	entity2	methotrexate sensitivity	methotrexate tolerability
normalized relationship	entity 1	ABCB1_Variant	ABCB1_Variant
	relationship	AFFECTS	AFFECTS
	entity2	methotrexate_DrugSensitivity	methotrexate_DrugSensitivity

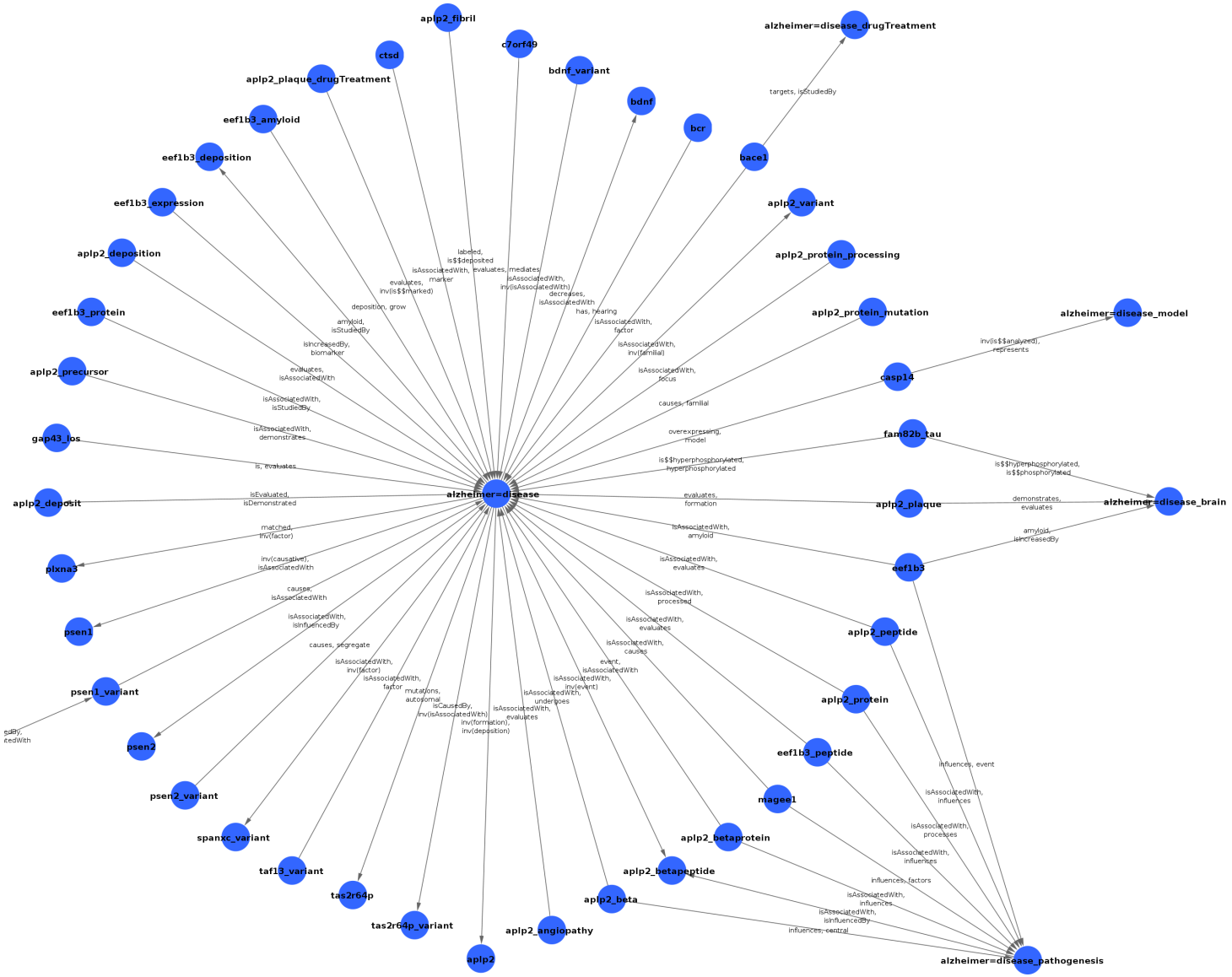
Example of network (1/3): VKORC1



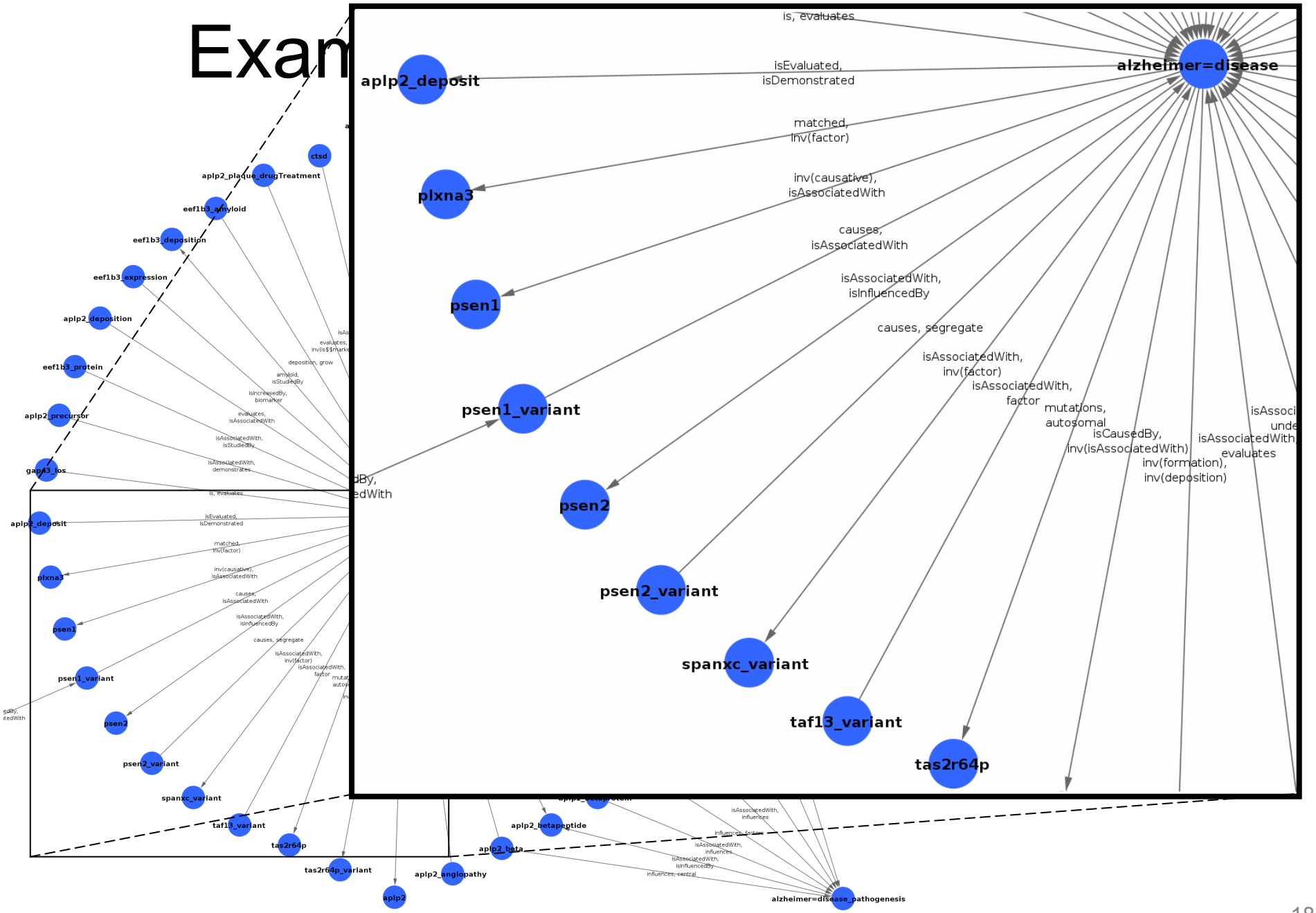
Example of network (2/3): VKORC1 something



Example of network (3/3): AD



Exam

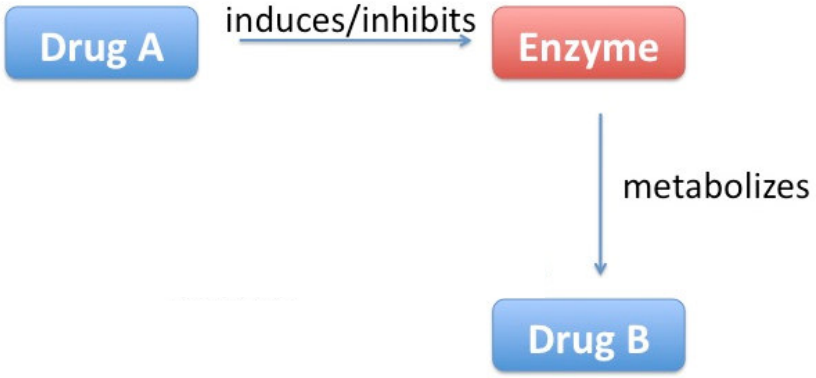

alzheimer=disease
is, evaluates
isEvaluated,
isDemonstrated
matched,
inv(factor)
inv(causative),
isAssociatedWith
causes,
isAssociatedWith
isAssociatedWith,
isInfluencedBy
causes, segregate
isAssociatedWith,
inv(factor)
isAssociatedWith,
factor
mutations,
autosomal
isCausedBy,
inv(isAssociatedWith)
inv(formation),
inv(deposition)
isAssociatedWith,
evaluates
aplp2_deposit
plxna3
psen1
psen1_variant
psen2
psen2_variant
spanxc_variant
taf13_variant
tas2r64p
ctsd
aplp2_plaque_drugTreatment
eef1b3_amyloid
eef1b3_deposition
eef1b3_expression
aplp2_deposition
eef1b3_protein
aplp2_precursor
gapA3_los
tas2r64p_variant
aplp2_angiopathy
aplp2
aplp2_betapeptide
aplp2_beta
alzheimer=disease_pathogenesis

Conclusion

- A method to build semantic network
- Used in PGx:
  - For curation and knowledge summarization
    @PharmGKB
  - For knowledge discovery

    e.g.Predicting Drug-Drug interaction

    =>Yael Garten's PhD thesis

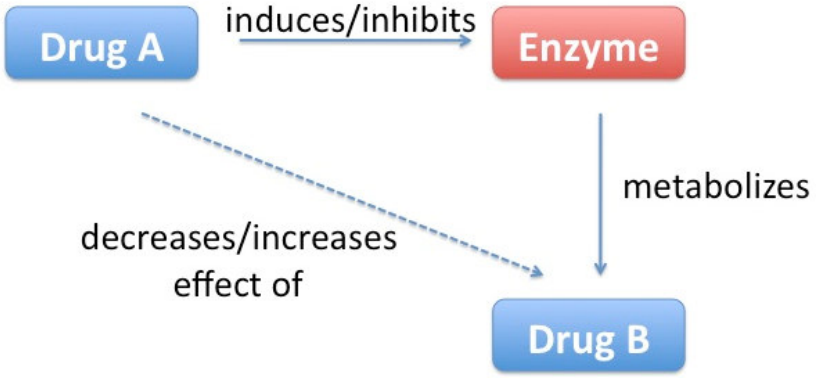


Conclusion

- A method to build semantic network
- Used in PGx:
  - For curation and knowledge summarization
    @PharmGKB
  - For knowledge discovery

    e.g.Predicting Drug-Drug interaction

    =>Yael Garten's PhD thesis



Questions?

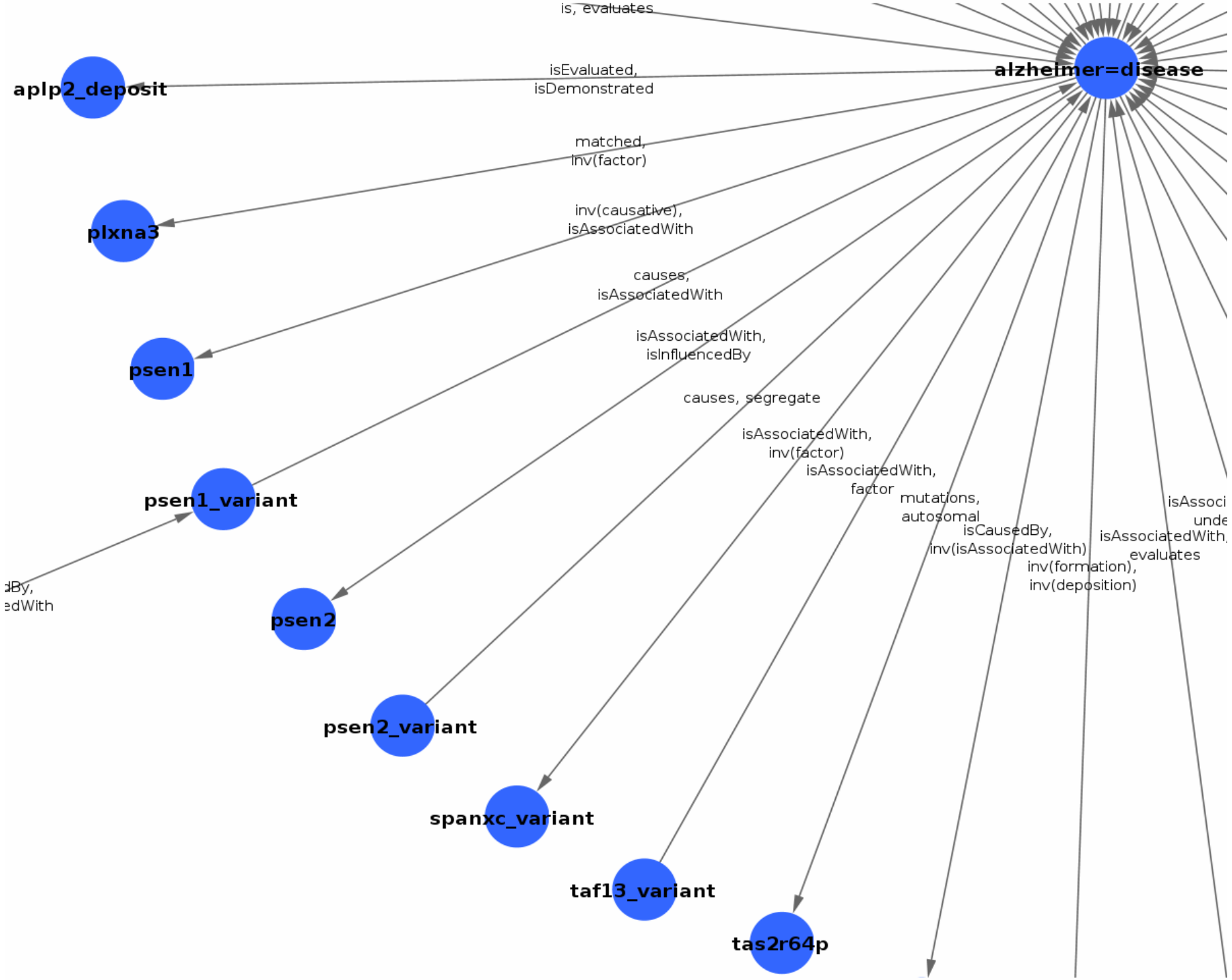
Coulet *et al. Journal of Biomedical Informatics*, In Press, 2010

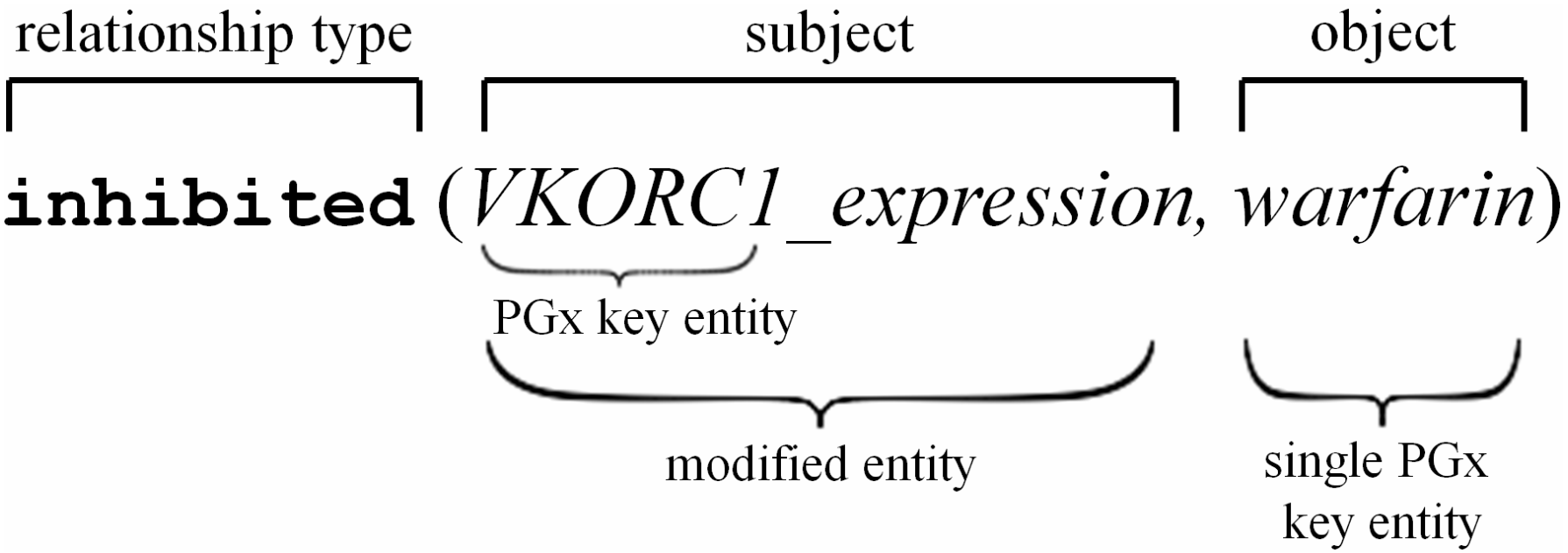
or

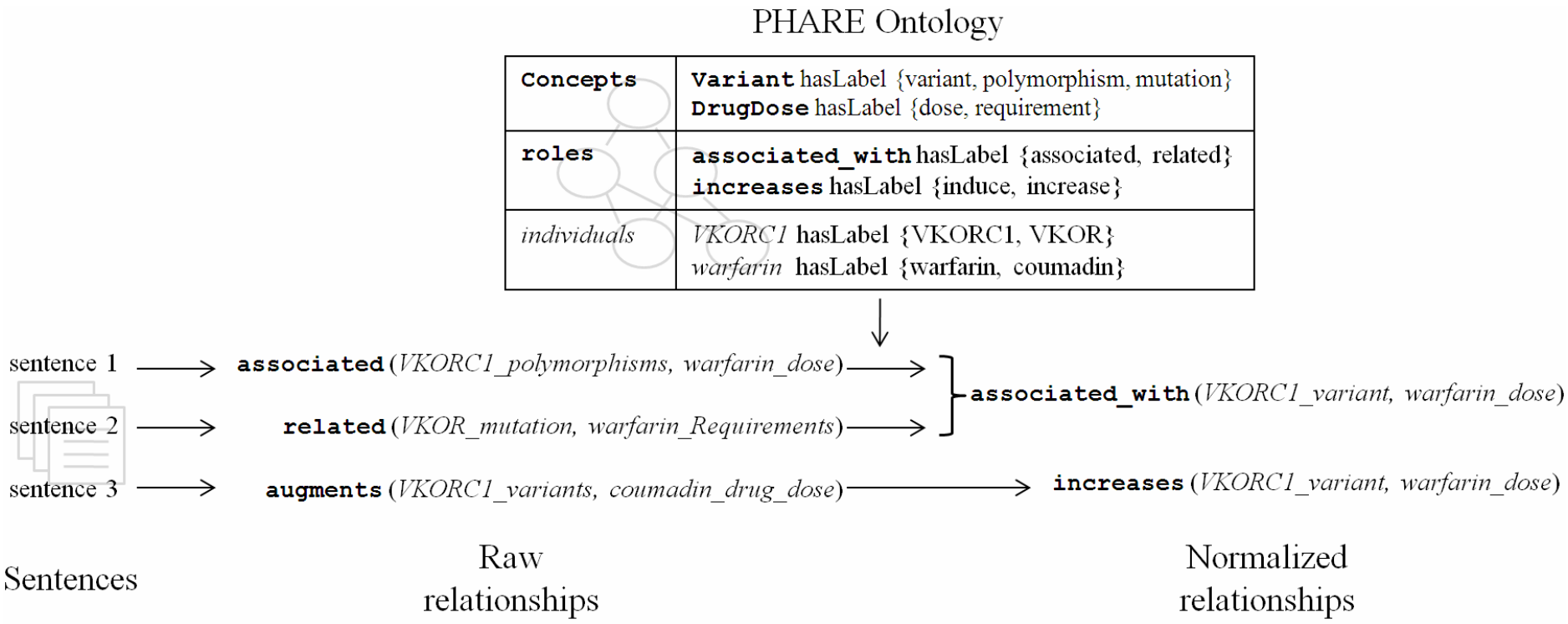
adrien.coulet@loria.fr

Thanks

And thanks to Yael Garten for many slides


is, evaluates
alzheimer=disease
aplp2_deposit
isEvaluated,
isDemonstrated
matched,
inv(factor)
plxna3
inv(causative),
isAssociatedWith
causes,
isAssociatedWith
psen1
isAssociatedWith,
isInfluencedBy
causes, segregate
isAssociatedWith,
inv(factor)
isAssociatedWith,
factor
psen1_variant
mutations,
autosomal
isCausedBy,
inv(isAssociatedWith)
isAssociatedWith,
evaluates
inv(formation),
inv(deposition)
psen2
psen2_variant
spanxc_variant
taf13_variant
tas2r64p


relationship type
subject
object
inhibited (VKORC1_expression, warfarin)
PGx key entity
modified entity
single PGx key entity


PHARE Ontology
Concepts
Variant hasLabel {variant, polymorphism, mutation}
DrugDose hasLabel {dose, requirement}
roles
associated_with hasLabel {associated, related}
increases hasLabel {induce, increase}
individuals
VKORC1 hasLabel {VKORC1, VKOR}
warfarin hasLabel {warfarin, coumadin}
sentence 1
associated(VKORC1_polymorphisms, warfarin_dose)
sentence 2
related(VKOR_mutation, warfarin_Requirements)
associated_with(VKORC1_variant, warfarin_dose)
sentence 3
augments(VKORC1_variants, coumadin_drug_dose)
increases(VKORC1_variant, warfarin_dose)
Sentences
Raw relationships
Normalized relationships