Classification of UMLS Source Vocabularies through Semantic Group Profiles

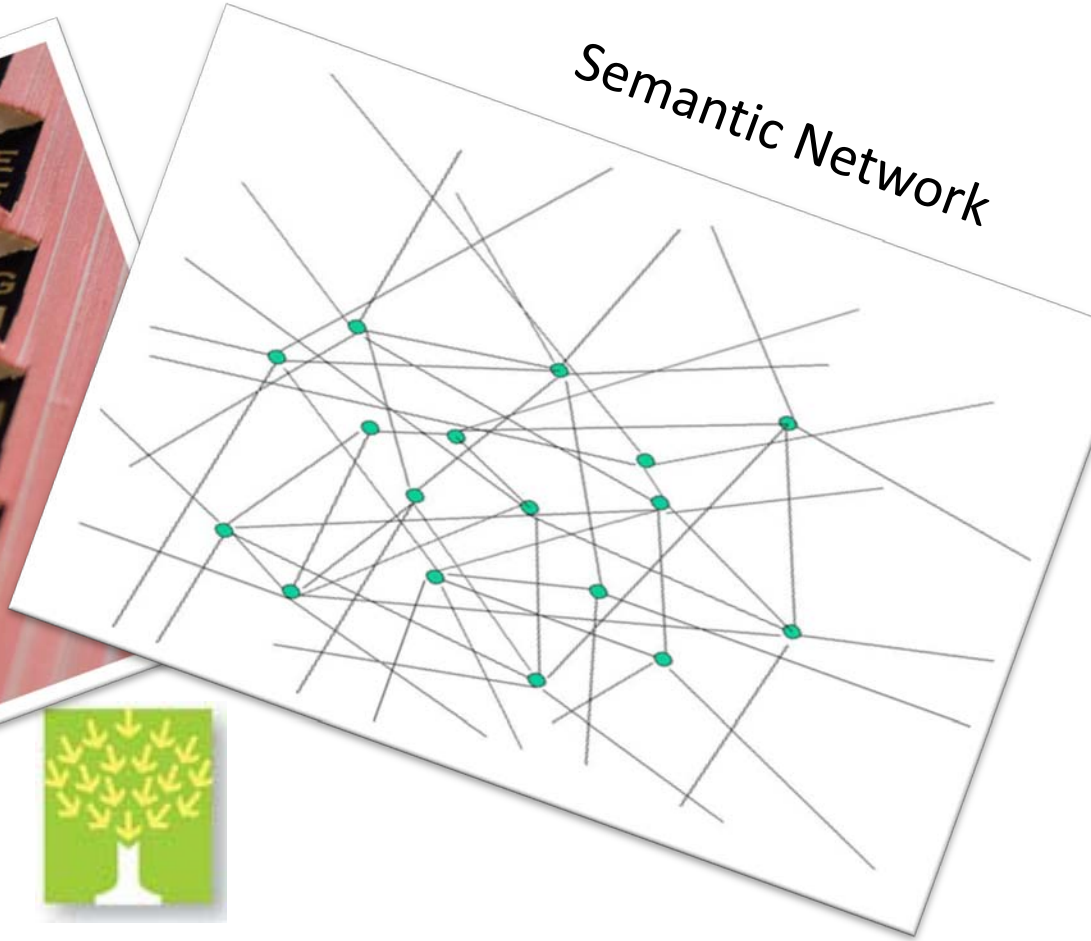
Thai Le
University of Washington
NLM Trainee Summer Rotation
8/1/11

Biomedical literature is vast.

What is the UMLS?

MLS?


Metathesaurus


Semantic Network

SPECIALIST Lexicon

155 Source Vocabularies

Clinical Procedural Terminology

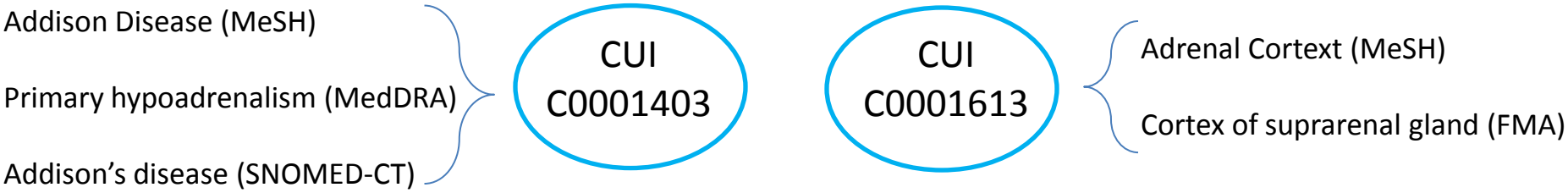
ICD 10

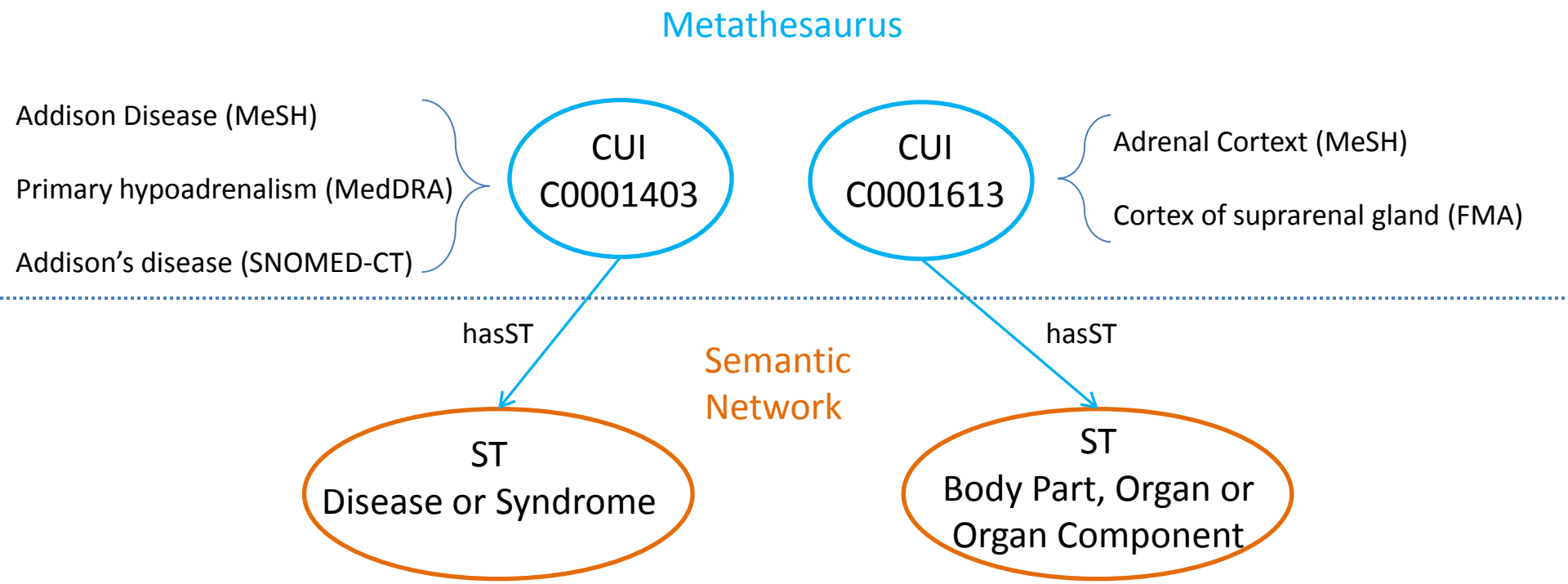
155 Source Vocabularies

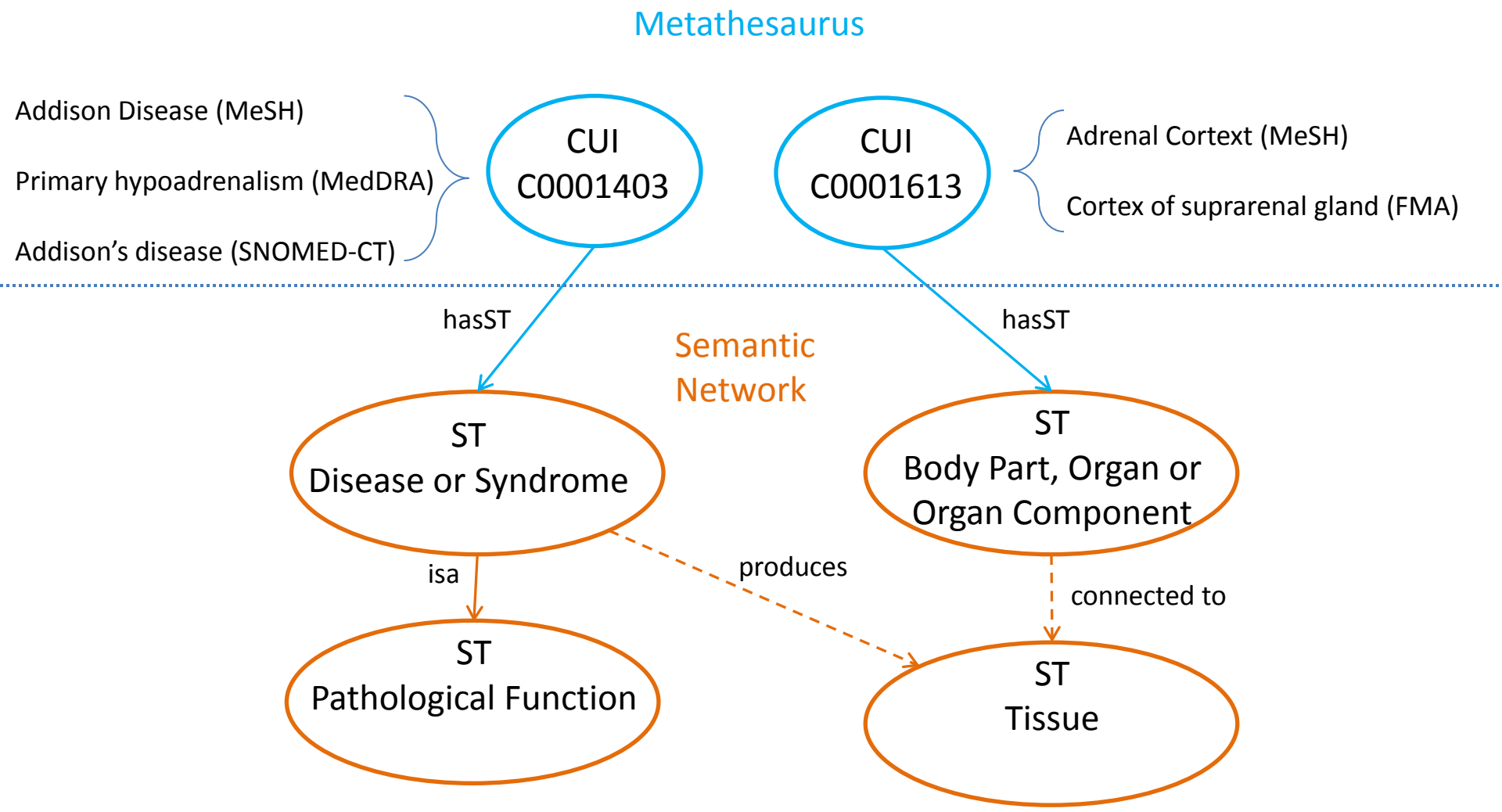
Gene Ontology

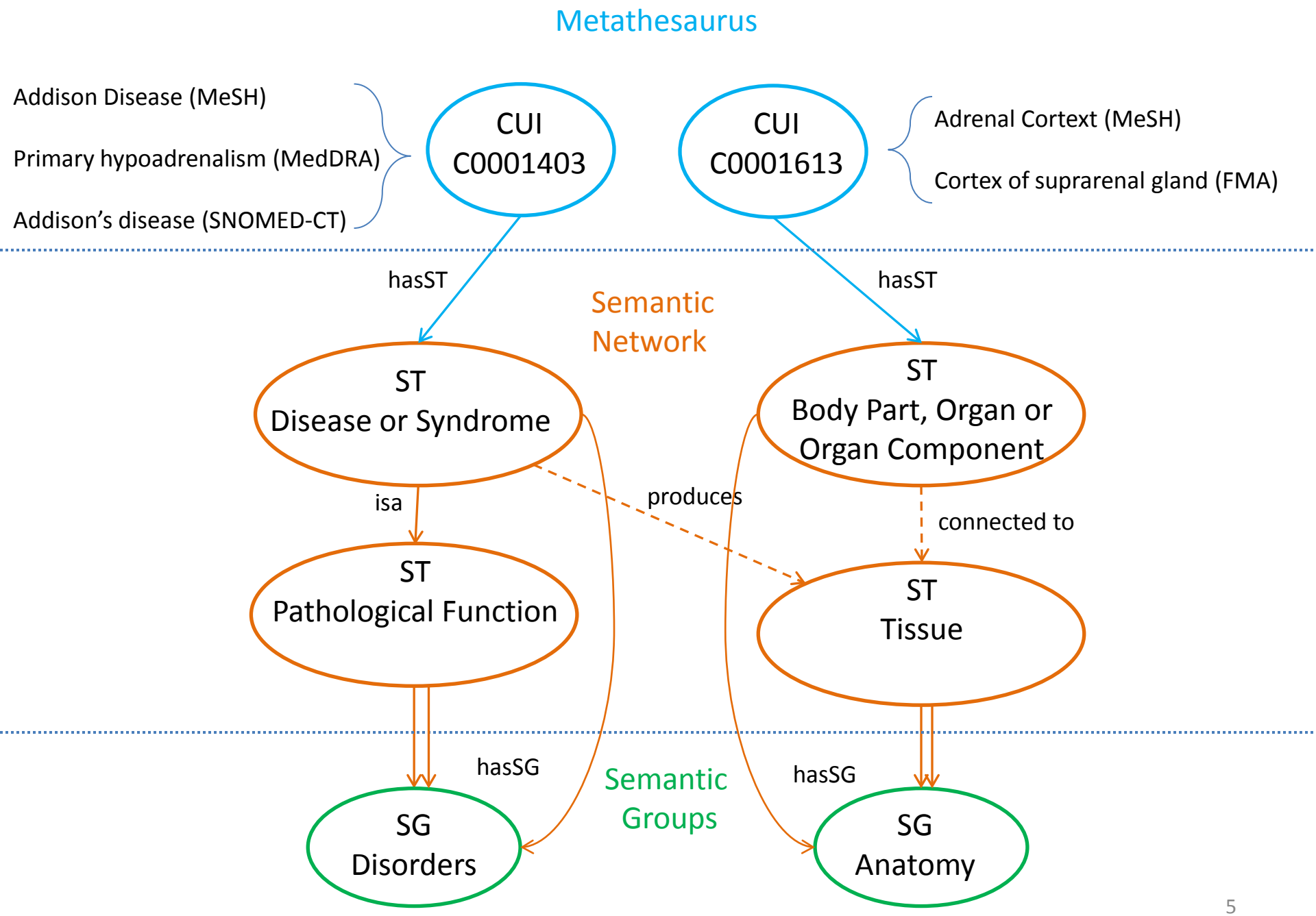
RxNorm

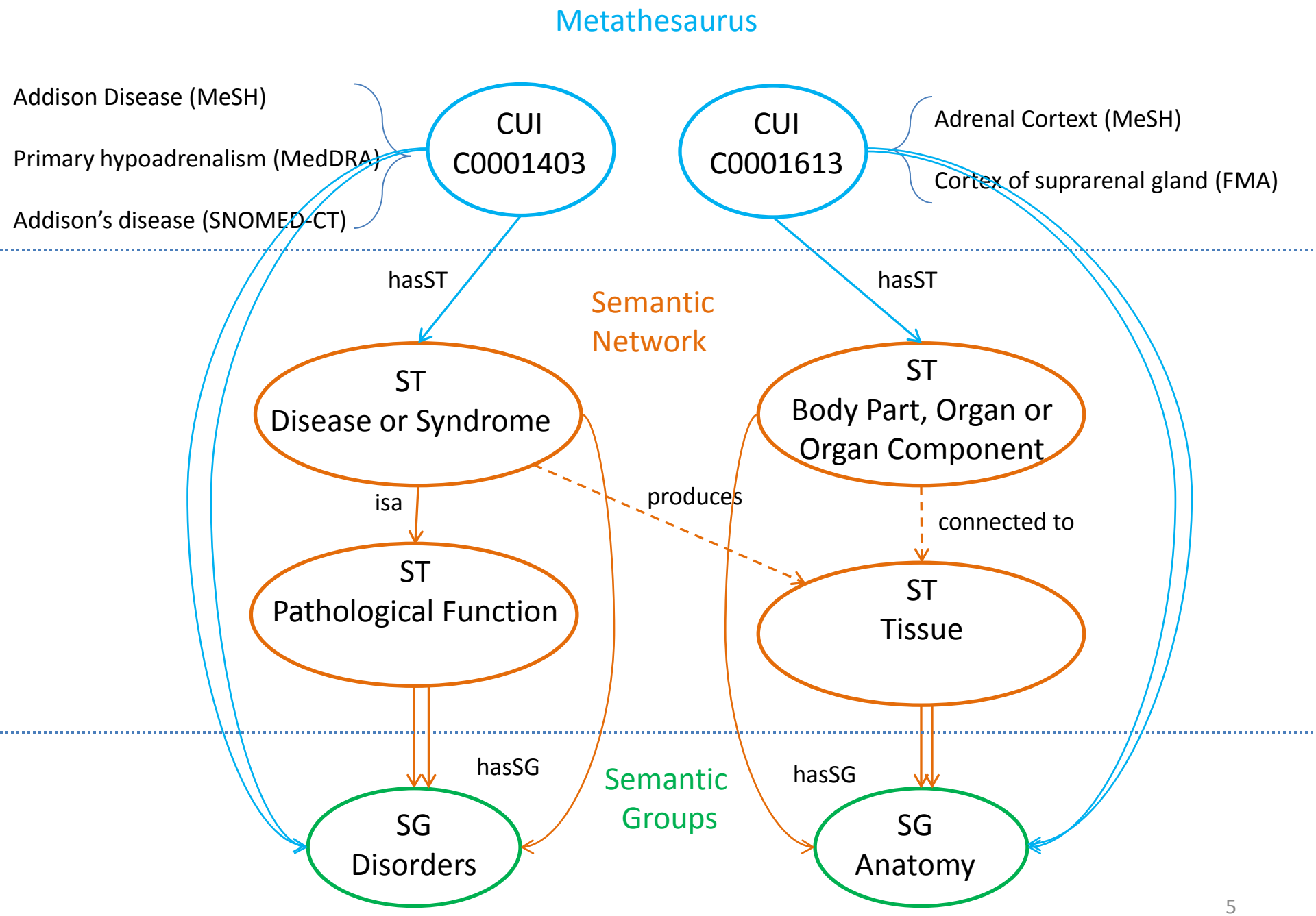
Metathesaurus




Metathesaurus
Addison Disease (MeSH)
Primary hypoadrenalism (MedDRA)
Addison's disease (SNOMED-CT)
CUI
C0001403
CUI
C0001613
Adrenal Cortext (MeSH)
Cortex of suprarenal gland (FMA)
hasST
hasST
Semantic
Network
ST
Disease or Syndrome
ST
Body Part, Organ or
Organ Component


Metathesaurus
Addison Disease (MeSH)
Primary hypoadrenalism (MedDRA)
Addison's disease (SNOMED-CT)
CUI
C0001403
CUI
C0001613
Adrenal Cortext (MeSH)
Cortex of suprarenal gland (FMA)
hasST
hasST
Semantic
Network
ST
Disease or Syndrome
ST
Body Part, Organ or
Organ Component
isa
produces
connected to
ST
Pathological Function
ST
Tissue


Metathesaurus
Addison Disease (MeSH)
Primary hypoadrenalism (MedDRA)
Addison's disease (SNOMED-CT)
CUI
C0001403
CUI
C0001613
Adrenal Cortext (MeSH)
Cortex of suprarenal gland (FMA)
hasST
hasST
Semantic
Network
ST
Disease or Syndrome
ST
Body Part, Organ or
Organ Component
isa
produces
connected to
ST
Pathological Function
ST
Tissue
hasSG
hasSG
Semantic
Groups
SG
Disorders
SG
Anatomy


Metathesaurus
Addison Disease (MeSH)
Primary hypoadrenalism (MedDRA)
Addison's disease (SNOMED-CT)
CUI
C0001403
CUI
C0001613
Adrenal Cortext (MeSH)
Cortex of suprarenal gland (FMA)
hasST
hasST
Semantic
Network
ST
Disease or Syndrome
ST
Body Part, Organ or
Organ Component
isa
produces
connected to
ST
Pathological Function
ST
Tissue
hasSG
hasSG
Semantic
Groups
SG
Disorders
SG
Anatomy

What are semantic groups?

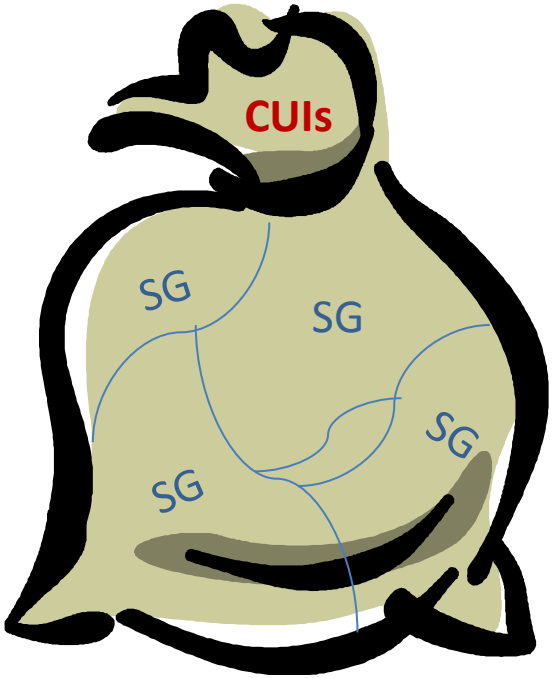
Semantic Group	Abbreviation
Activities & Behavior	ACTI
Anatomy	ANAT
Chemicals & Drugs	CHEM
Concepts & Ideas	CONC
Devices	DEVI
Disorders	DISO
Genes and Molecular Sequences	GENE
Geographic Areas	GEOG
Living Beings	LIVB
Objects	OBJC
Occupations	OCCU
Organizations	ORGA
Phenomena	PHEN
Physiology	PHYS
Procedures	PROC

What are semantic groups?

Semantic Group	Abbreviation
Activities & Behavior	ACTI
Anatomy	ANAT
Chemicals & Drugs	CHEM
Concepts & Ideas	CONC
Devices	DEVI
Disorders	DISO
Genes and Molecular Sequences	GENE
Geographic Areas	GEOG
Living Beings	LIVB
Objects	OBJC
Occupations	OCCU
Organizations	ORGA
Phenomena	PHEN
Physiology	PHYS
Procedures	PROC

Semantic Types
Acquired Abnormality
Anatomical Abnormality
Cell or Molecular Dysfunction
Congenital Abnormality
Disease or Syndrome
Experimental Model of Disease
Finding
Injury or Poisoning
Mental or Behavioral Dysfunction
Neoplastic Process
Pathological Function
Sign or Symptom

2.4 million concepts → 133 semantic types → 15 semantic groups



Concepts can be categorized into more than one semantic group (1029 such concepts)

Semantic groups form a 99.96 % partition of all concepts.

Project Goals

1. Demonstrate feasibility of using semantic group profiles to classify source vocabularies

Project Goals

1. Demonstrate feasibility of using semantic group profiles to classify source vocabularies

2. Compare with a functional classification of source vocabularies

Source Vocabulary Filtering

Difficult to display all source vocabularies

Filtered based on criteria:

-Greater than 1000 CUIs

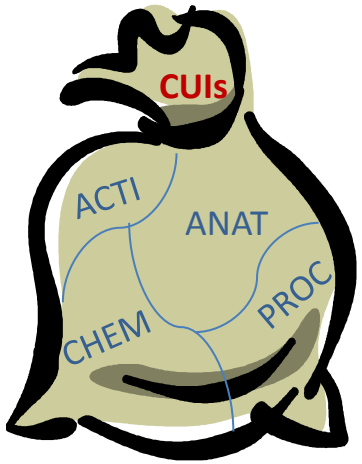
-Only English vocabularies (removes redundancy)

155 source vocabularies → 57 filtered source vocabularies

Classification based on:

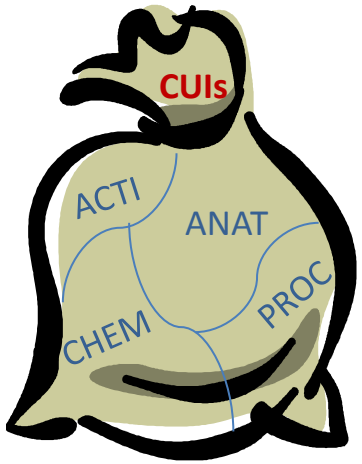
Semantic Group Profiles

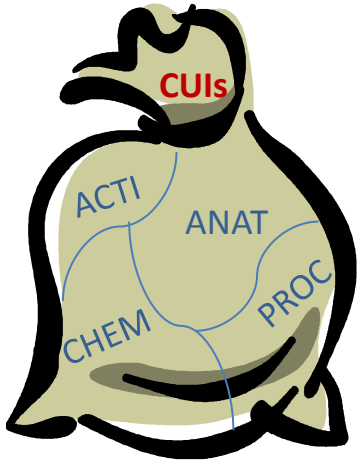
Generating Semantic Group Profiles



Source	CONC	PHEN	CHEM	…	PROC
CSP	4.02	2.11	30.1	…	11.11

Generating Semantic Group Profiles





Source	CONC	PHEN	CHEM	…	PROC
CSP	4.02	2.11	30.1	…	11.11
PSY	18.23	1.71	8.26	…	12.33

Define Euclidian Distance Matrix for Hierarchical Clustering

Source	CONC	PHEN	CHEM	LIVB	ACTI	DISO	GEOG	ANAT	GENE	OCCU	OBJ	DEVI	ORGA	PHYS	PROC
CSP	4.02	2.11	30.1	11.5	2.08	18.1	0.58	7.15	0.73	2.11	1.72	1.12	0.51	7.09	11.11
PSY	18.23	1.71	8.26	8.9	10.77	14.44	0.17	4.72	0.06	2.71	2.65	0.49	1.51	13.05	12.33

Define Euclidian Distance Matrix for Hierarchical Clustering

Source	CONC	PHEN	CHEM	LIVB	ACTI	DISO	GEOG	ANAT	GENE	OCCU	OBJ	DEVI	ORGA	PHYS	PROC
CSP	4.02	2.11	30.1	11.5	2.08	18.1	0.58	7.15	0.73	2.11	1.72	1.12	0.51	7.09	11.11
PSY	18.23	1.71	8.26	8.9	10.77	14.44	0.17	4.72	0.06	2.71	2.65	0.49	1.51	13.05	12.33

$$\sqrt{(18.23-.38)^2 + (8.26-0.55)^2 + (8.9-0)^2 + \ldots\ldots + (12.33-0)^2}$$

Euclidian distance = 28.65

Define Euclidian Distance Matrix for Hierarchical Clustering

Source	CONC	PHEN	CHEM	LIVB	ACTI	DISO	GEOG	ANAT	GENE	OCCU	OBJ	DEVI	ORGA	PHYS	PROC
CSP	4.02	2.11	30.1	11.5	2.08	18.1	0.58	7.15	0.73	2.11	1.72	1.12	0.51	7.09	11.11
PSY	18.23	1.71	8.26	8.9	10.77	14.44	0.17	4.72	0.06	2.71	2.65	0.49	1.51	13.05	12.33

$$\sqrt{(18.23-.38)^2 + (8.26-0.55)^2 + (8.9-0)^2 + \ldots\ldots + (12.33-0)^2}$$

Euclidian distance = 28.65

-	-	-	0	Source 4
-	28.65	0	-	CSP
-	0	28.65	-	PSY
0	-	-	-	Source 1
Source 1	PSY	CSP	Source 4	

Hierarchical Clustering

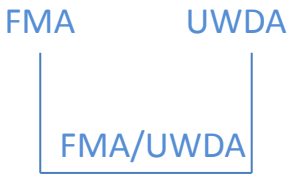
100.1	100.2	0	**PSY**
0.236	0	100.2	**UWDA**
0	0.236	100.1	**FMA**
FMA	**UWDA**	**PSY**	

Hierarchical Clustering

100.1	100.2	0	**PSY**
0.236	0	100.2	**UWDA**
0	0.236	100.1	**FMA**
FMA	**UWDA**	**PSY**	

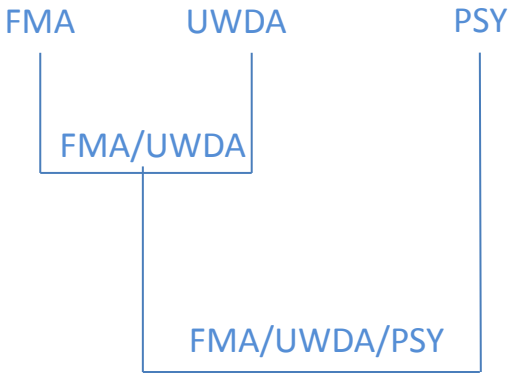
Hierarchical Clustering

100.1	100.2	0	**PSY**
0.236	0	100.2	**UWDA**
0	0.236	100.1	**FMA**
FMA	**UWDA**	**PSY**	



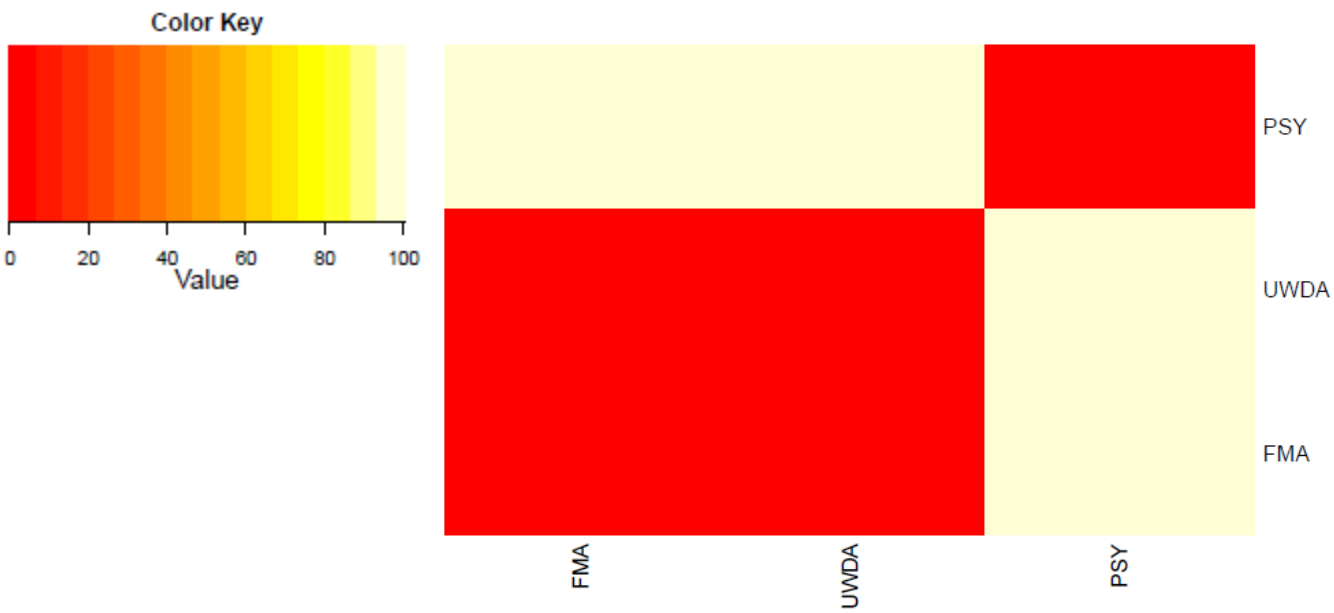
Hierarchical Clustering

100.2	0	PSY
0	100.2	UWDA/ FMA
FMA/UWDA	PSY	



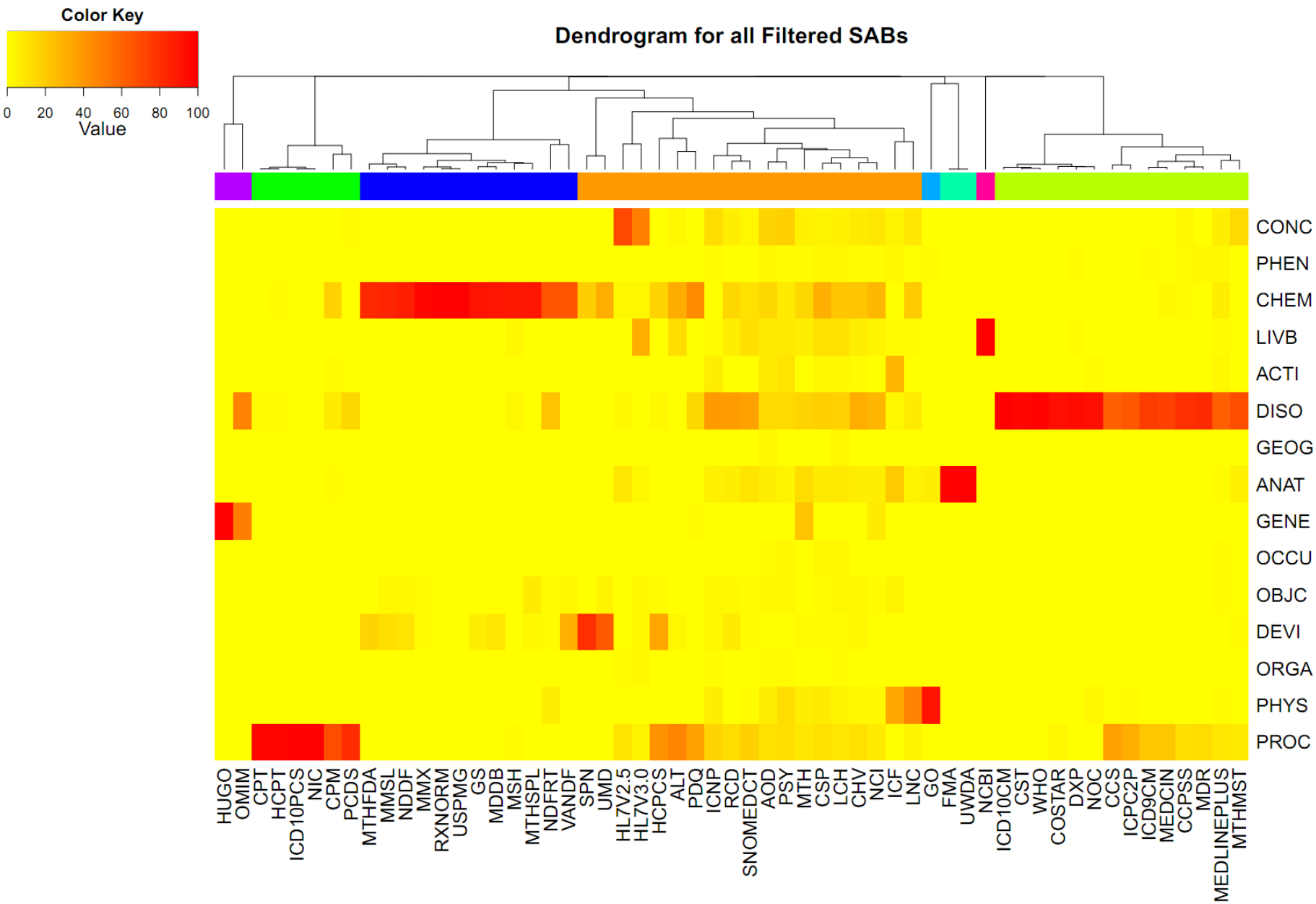
Hierarchical Clustering

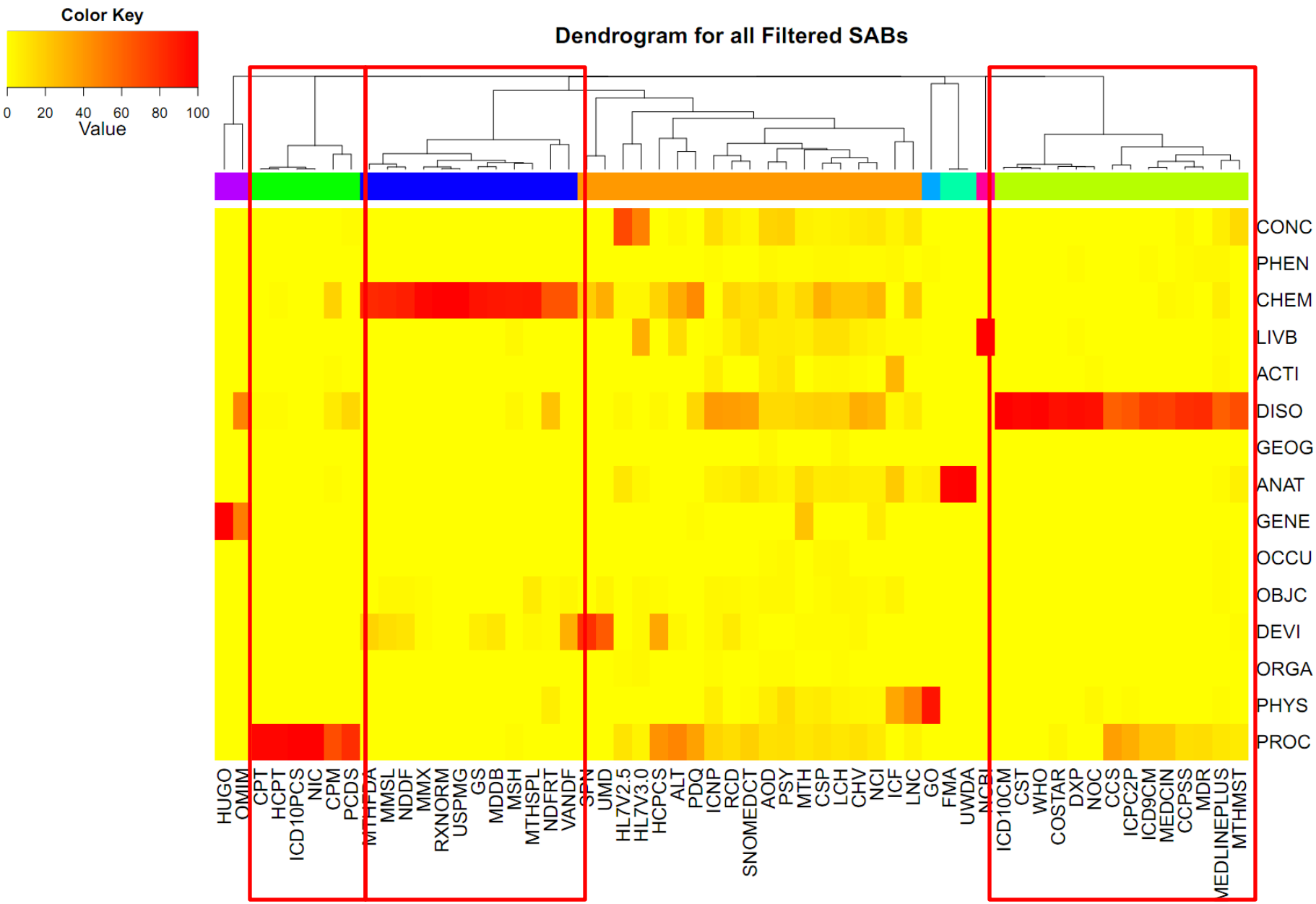
100.1	100.2	0	**PSY**
0.236	0	100.2	**UWDA**
0	0.236	100.1	**FMA**
FMA	**UWDA**	**PSY**	

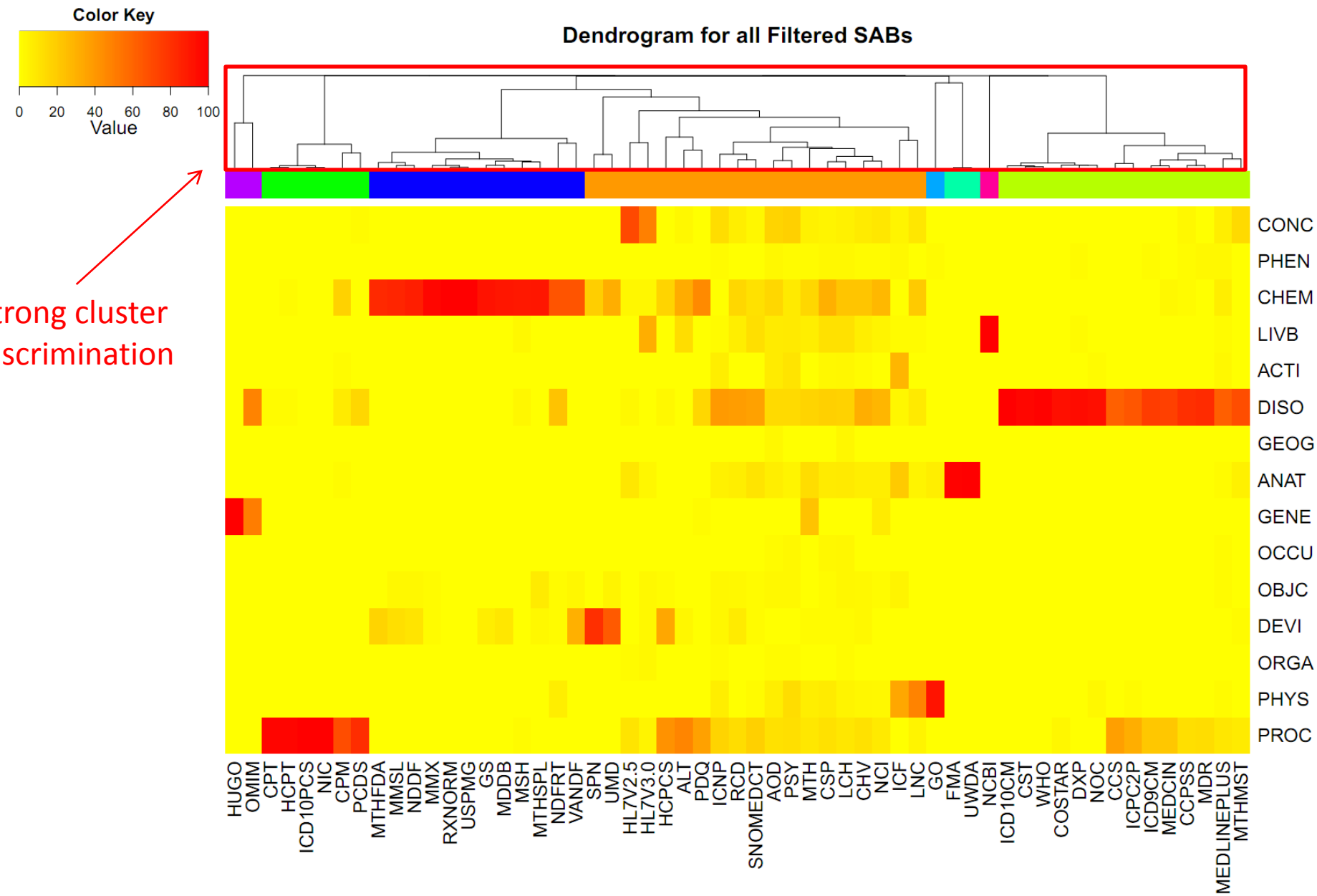


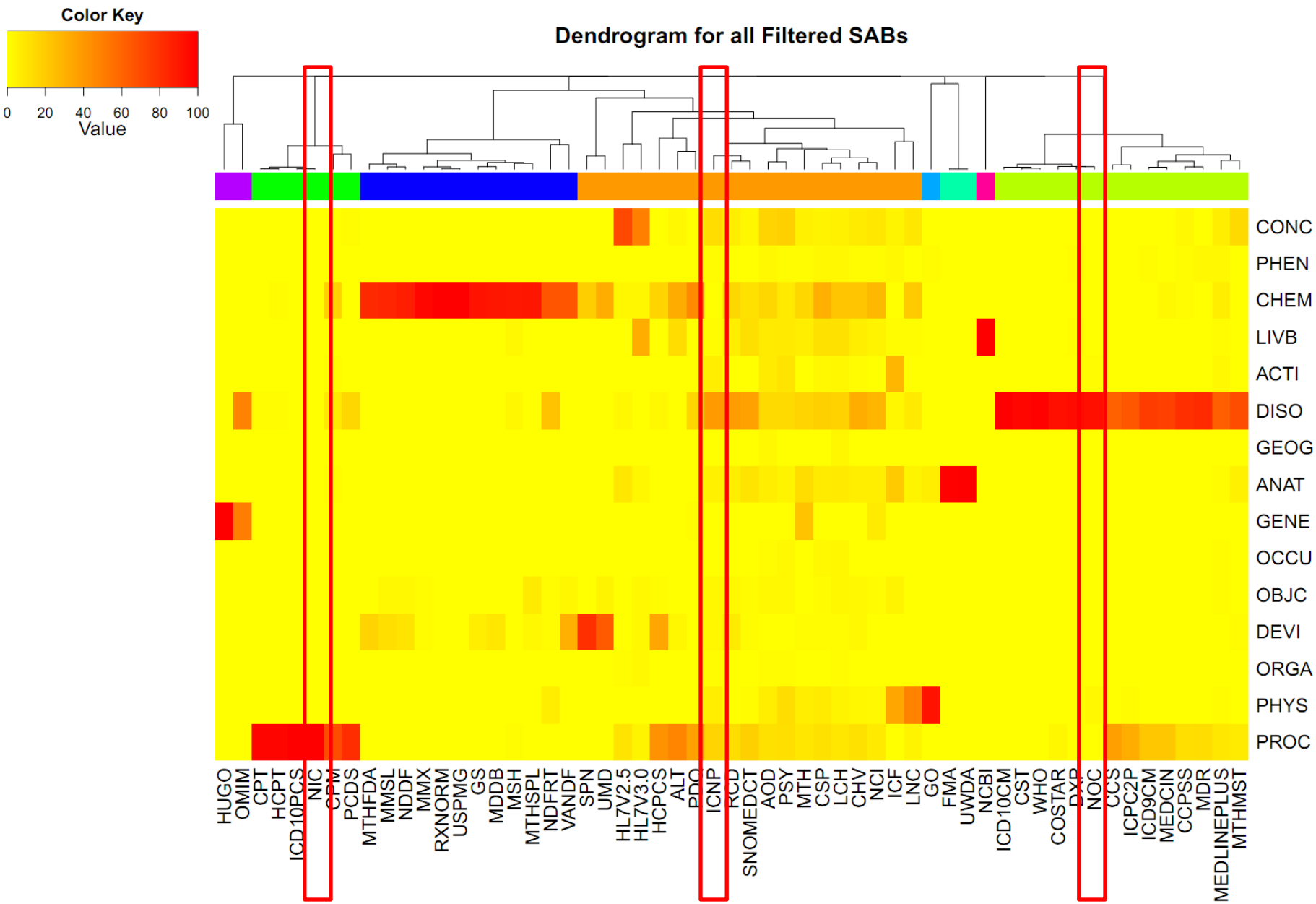
Results and analysis:

Semantic Group Profile


Color Key
0 20 40 60 80 100
Value
Dendrogram for all Filtered SABs
CONC
PHEN
CHEM
LIVB
ACTI
DISO
GEOG
ANAT
GENE
OCCU
OBJC
DEVI
ORGA
PHYS
PROC
HUGO
OMIM
CPT
HCPT
ICD10PCS
NIC
CPM
PCDS
MTHFDA
MMSL
NDDF
MMX
RXNORM
USPMG
GS
MDDB
MSH
MTHSPL
NDFRT
VANDF
SPN
UMD
HL7V2.5
HL7V3.0
HCPCS
ALT
PDQ
ICNP
RCD
SNOMEDCT
AOD
PSY
MTH
CSP
LCH
CHV
NCI
ICF
LNC
GO
FMA
UWDA
NCBI
ICD10CM
CST
WHO
COSTAR
DXP
NOC
CCS
ICPC2P
ICD9CM
MEDCIN
CCPSS
MDR
MEDLINEPLUS
MTHMST


Color Key
0 20 40 60 80 100
Value
Dendrogram for all Filtered SABs
CONC
PHEN
CHEM
LIVB
ACTI
DISO
GEOG
ANAT
GENE
OCCU
OBJC
DEVI
ORGA
PHYS
PROC
HUGO
OMIM
CPT
HCPT
ICD10PCS
NIC
CPM
PCDS
MTHFDA
MMSL
NDDF
MMX
RXNORM
USPMG
GS
MDDB
MSH
MTHSPL
NDFRT
VANDF
SPN
UMD
HL7V2.5
HL7V3.0
HCPCS
ALT
PDQ
ICNP
RCD
SNOMEDCT
AOD
PSY
MTH
CSP
LCH
CHV
NCI
ICF
LNC
GO
FMA
UWDA
NCBI
ICD10CM
CST
WHO
COSTAR
DXP
NOC
CCS
ICPC2P
ICD9CM
MEDCIN
CCPSS
MDR
MEDLINEPLUS
MTHMST
Procedures
Drug vocabularies
Disorders


Color Key
0 20 40 60 80 100
Value
Dendrogram for all Filtered SABs
Strong cluster discrimination
CONC
PHEN
CHEM
LIVB
ACTI
DISO
GEOG
ANAT
GENE
OCCU
OBJC
DEVI
ORGA
PHYS
PROC
HUGO
OMIM
CPT
HCPT
ICD10PCS
NIC
CPM
PCDS
MTHFDA
MMSL
NDDF
MMX
RXNORM
USPMG
GS
MDDB
MSH
MTHSPL
NDFRT
VANDF
SPN
UMD
HL7V2.5
HL7V3.0
HCPCS
ALT
PDQ
ICNP
RCD
SNOMEDCT
AOD
PSY
MTH
CSP
LCH
CHV
NCI
ICF
LNC
GO
FMA
UWDA
NCBI
ICD10CM
CST
WHO
COSTAR
DXP
NOC
CCS
ICPC2P
ICD9CM
MEDCIN
CCPSS
MDR
MEDLINEPLUS
MTHMST



Nursing vocabularies are in distinct clusters

Classification based on:

Source Usage

UMLS Functional Classification

Biomedical vocabularies in the UMLS include:

"thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research."

Reference: US National Library of Medicine. UMLS Metathesaurus Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Gene Ontology (GO)

Purpose

GO is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The current ontologies of the GO project are molecular function, biological process, and cellular component.

Audience

GO is intended for use by system developers who are addressing the need for consistent descriptions of gene products in different databases.

Functional Classification

Basic Research

Gene Ontology (GO)

Purpose

GO is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The current ontologies of the GO project are molecular function, biological process, and cellular component.

Audience

GO is intended for use by system developers who are addressing the need for consistent descriptions of gene products in different databases.

Functional Classification

Basic Research

Healthcare Common Procedure Coding System (HCPCS)

Purpose

HCPCS is a collection of standardized codes that represent medical procedures, supplies, products and services. The codes are used to facilitate the processing of health insurance claims by Medicare and other insurers.

Audience

HCPCS is used by physicians and other health care professionals and insurance programs.

Functional Classification

Health Services Billing

Comparisons across:

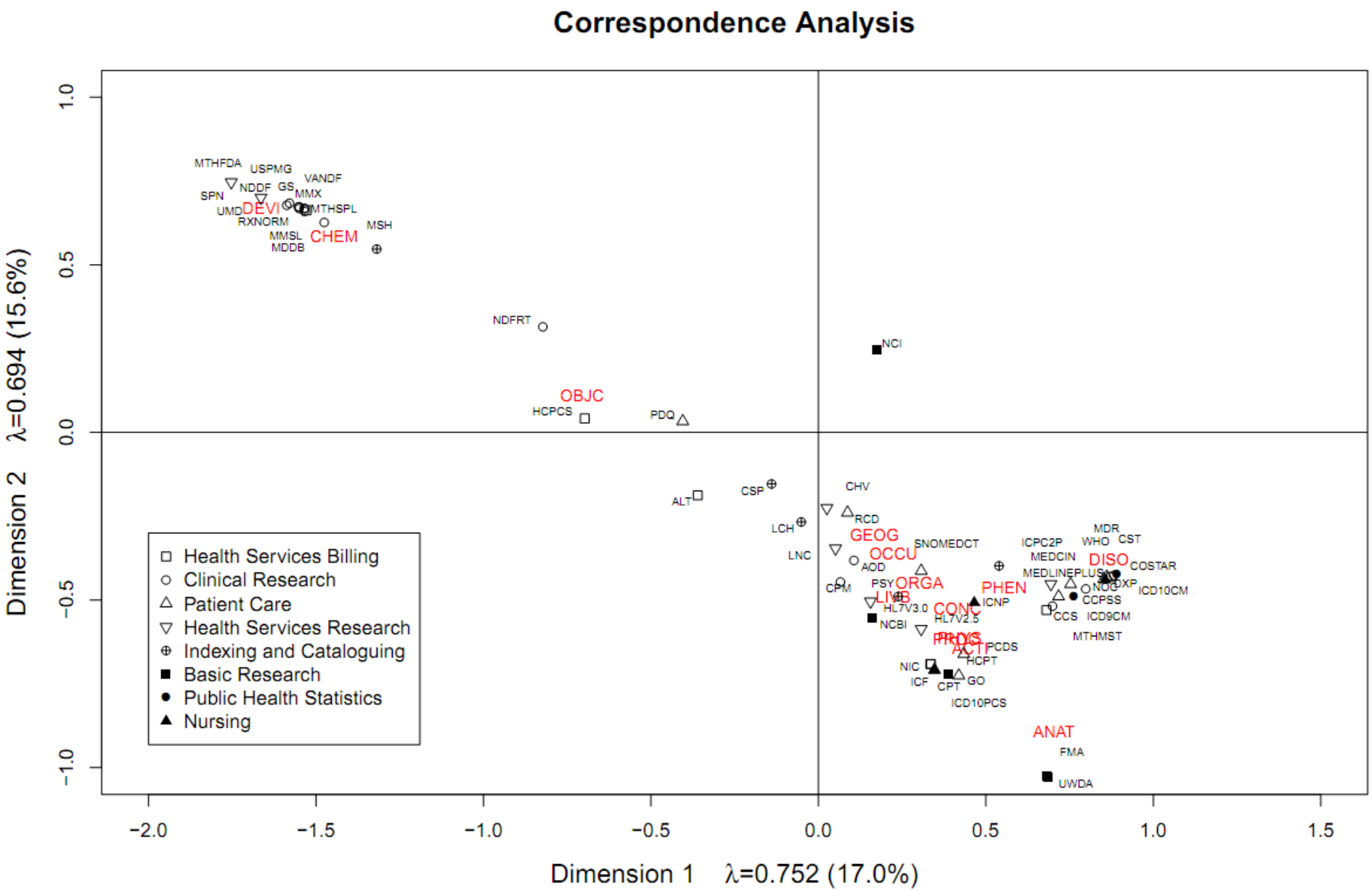
Source Classifications

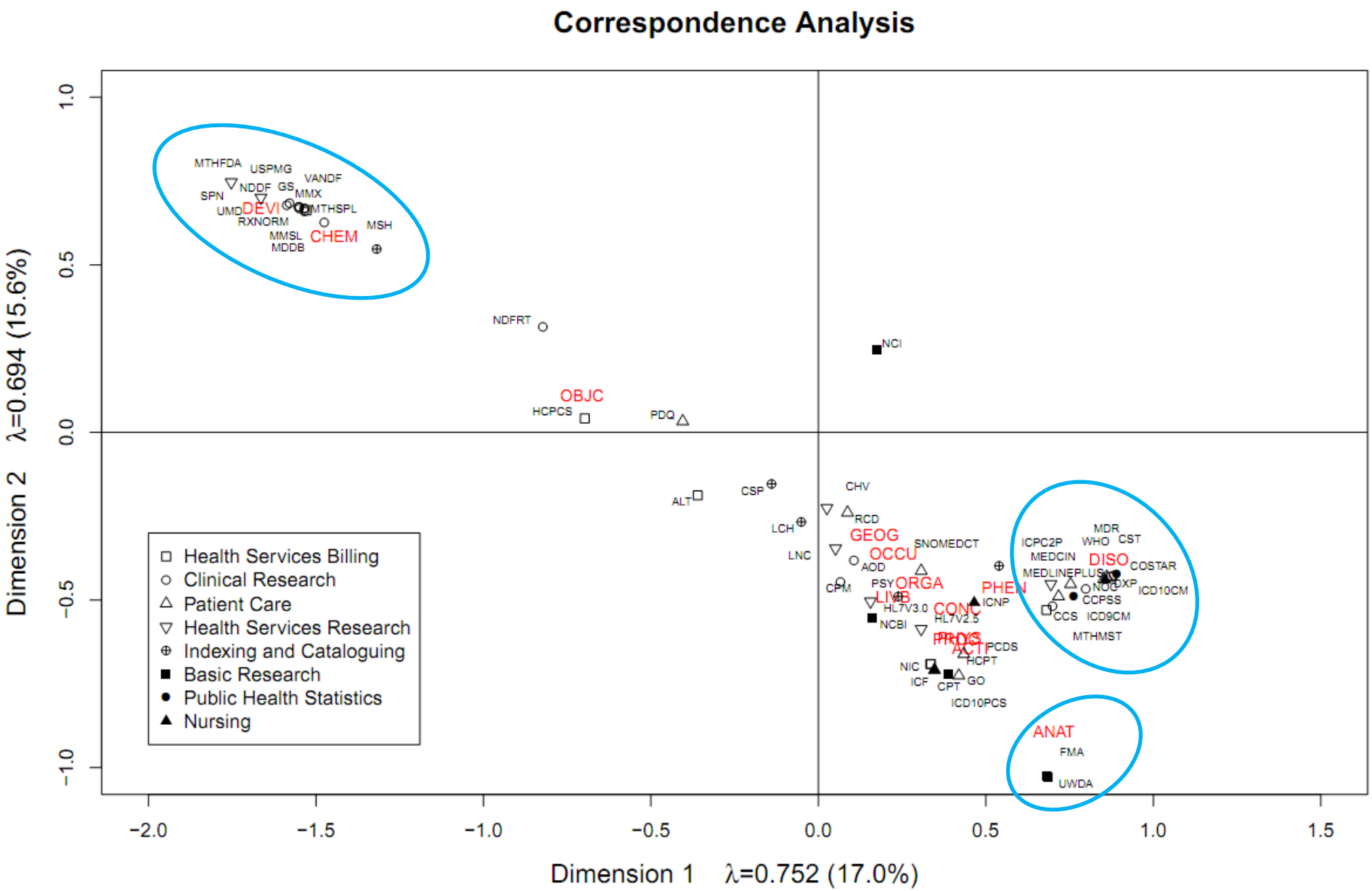
Correspondence Analysis

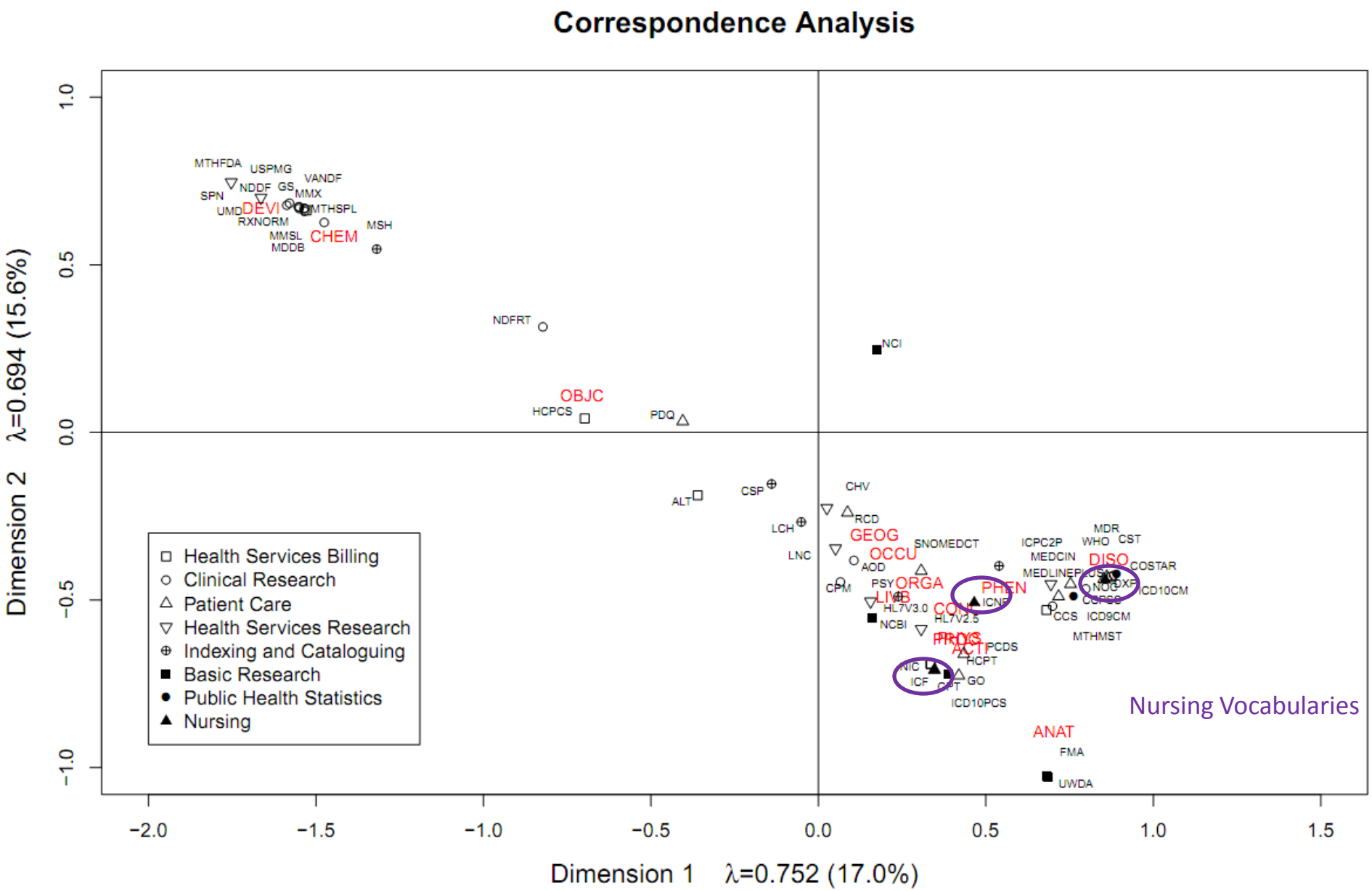
Similar to principal component analysis

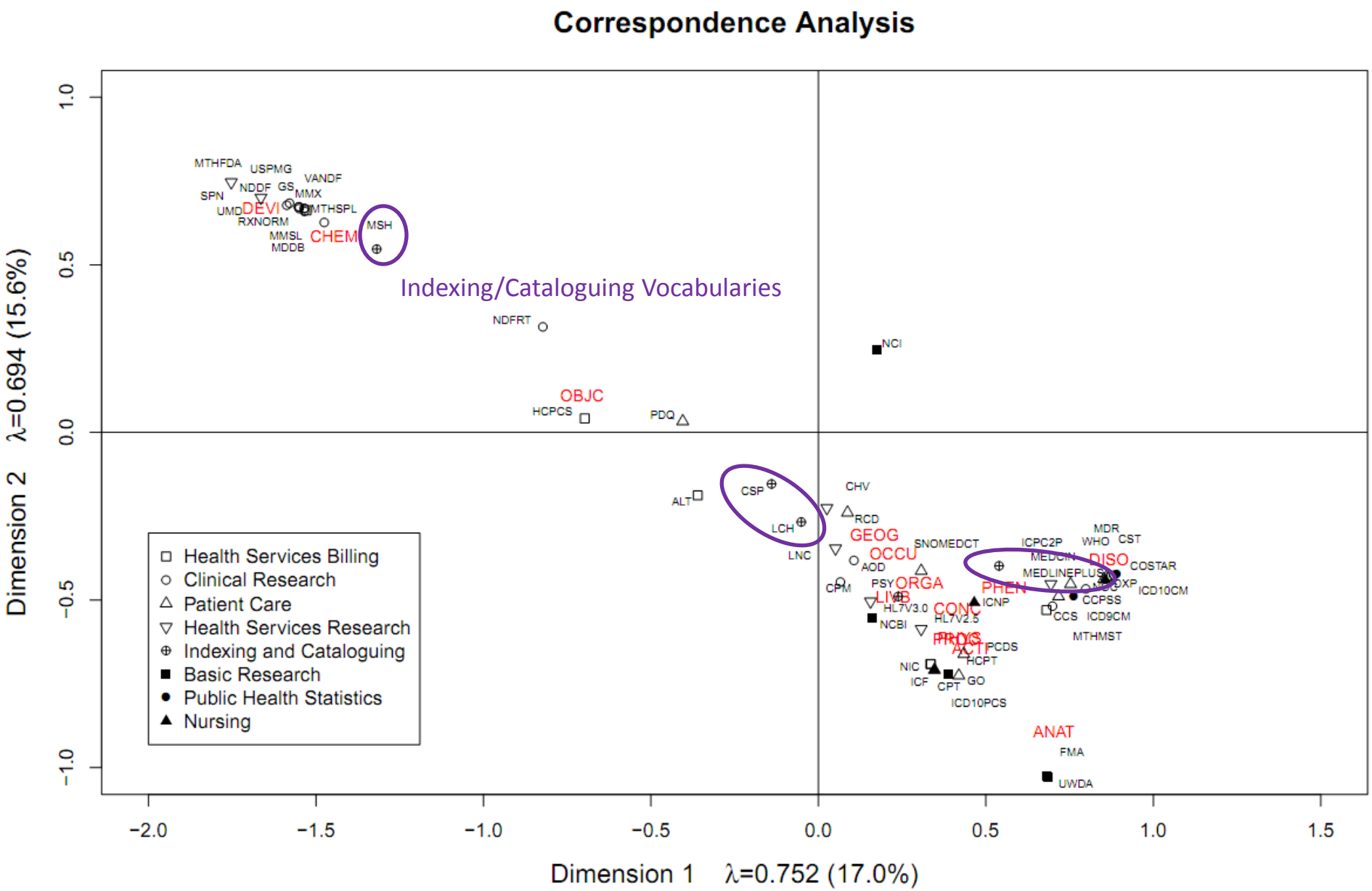
Compares row and column profiles for significance

Projects onto 2D plot


Correspondence Analysis
Dimension 2 λ=0.694 (15.6%)
Dimension 1 λ=0.752 (17.0%)
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
MTHFDA
USPMG
VANDF
NDDF
GS
SPN
MMX
UMD
DEVI
MTHSPL
RXNORM
MSH
MMSL
CHEM
MDDB
NDFRT
NCI
OBJC
HCPCS
PDQ
ALT
CSP
CHV
LCH
RCD
GEOG
LNC
SNOMEDCT
OCCU
AOD
CPM
PSY
ORGA
LIVB
HL7V3.0
CONC
NCBI
HL7V2.5
PHEN
ICNP
PCDS
HCPT
NIC
ICF
CPT
GO
ICD10PCS
ICPC2P
MDR
WHO
CST
MEDCIN
DISO
COSTAR
MEDLINEPLUS
NCI
DXP
ICD10CM
CCPSS
CCS
ICD9CM
MTHMST
ANAT
FMA
UWDA
Health Services Billing
Clinical Research
Patient Care
Health Services Research
Indexing and Cataloguing
Basic Research
Public Health Statistics
Nursing


Correspondence Analysis
Dimension 2 λ=0.694 (15.6%)
Dimension 1 λ=0.752 (17.0%)
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
MTHFDA
USPMG
VANDF
NDDF
GS
MMX
SPN
UMD
DEVI
MTHSPL
RXNORM
MSH
MMSL
CHEM
MDDB
NDFRT
NCI
OBJC
HCPCS
PDQ
ALT
CSP
CHV
RCD
LCH
GEOG
SNOMEDCT
LNC
OCCU
AOD
CPM
PSY
ORGA
LIVB
HL7V3.0
CONC
NCBI
HL7V2.5
PHEN
ICNP
PROC
ACTI
PCDS
HCPT
NIC
ICF
CPT
GO
ICD10PCS
MDR
WHO
CST
ICPC2P
MEDCIN
DISO
COSTAR
MEDLINEPLUS
NCI
DXP
ICD10CM
CCPSS
CCS
ICD9CM
MTHMST
ANAT
FMA
UWDA
Health Services Billing
Clinical Research
Patient Care
Health Services Research
Indexing and Cataloguing
Basic Research
Public Health Statistics
Nursing


Correspondence Analysis
Dimension 2 λ=0.694 (15.6%)
Dimension 1 λ=0.752 (17.0%)
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
MTHFDA
USPMG
VANDF
NDDF
GS
MMX
SPN
UMD
DEVI
MTHSPL
RXNORM
MSH
MMSL
CHEM
MDDB
NDFRT
NCI
OBJC
HCPCS
PDQ
ALT
CSP
CHV
RCD
LCH
GEOG
LNC
SNOMEDCT
OCCU
AOD
CPM
PSY
ORGA
LIVB
HL7V3.0
NCBI
PHEN
ICNP
CONC
HL7V2.5
PROC
ACTI
PCDS
HCPT
NIC
ICF
GO
ICD10PCS
ICPC2P
MEDCIN
MDR
WHO
CST
DISO
COSTAR
MEDLINEPLUS
NDC
DXP
ICD10CM
CCS
ICD9CM
MTHMST
ANAT
FMA
UWDA
Health Services Billing
Clinical Research
Patient Care
Health Services Research
Indexing and Cataloguing
Basic Research
Public Health Statistics
Nursing
Nursing Vocabularies

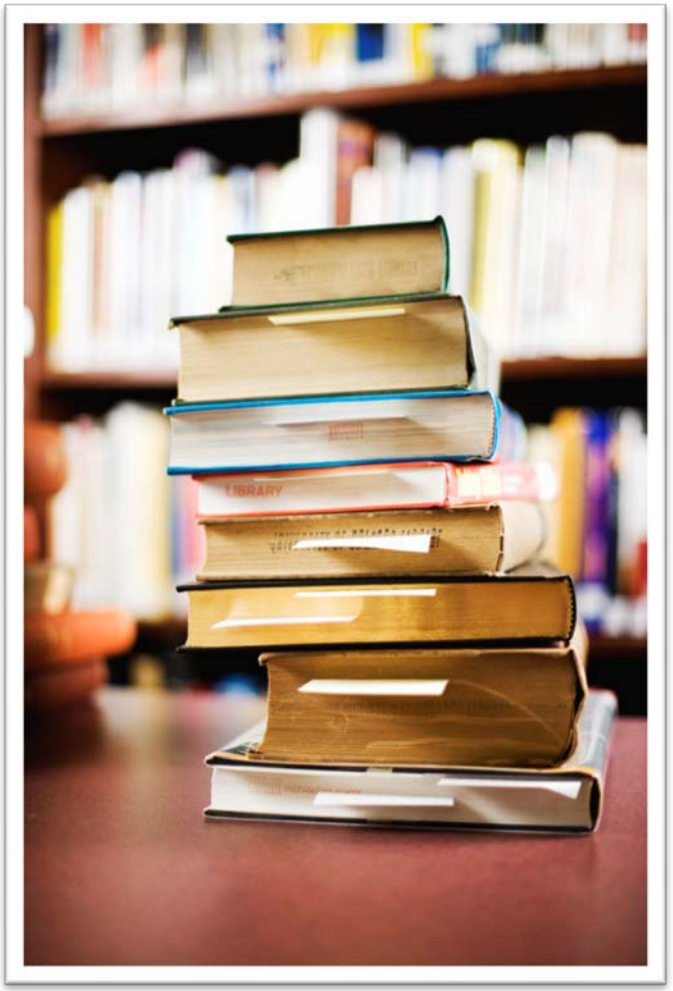

Correspondence Analysis
Indexing/Cataloguing Vocabularies
Health Services Billing
Clinical Research
Patient Care
Health Services Research
Indexing and Cataloguing
Basic Research
Public Health Statistics
Nursing
MTHFDA
USPMG
VANDF
NDDF
GS
MMX
SPN
UMD
DEVI
MTHSPL
RXNORM
MMSL
MDDB
CHEM
MSH
NDFRT
NCI
OBJC
HCPCS
PDQ
ALT
CSP
LCH
LNC
CHV
RCD
GEOG
SNOMEDCT
OCCU
AOD
CPM
PSY
ORGA
LIVB
HL7V3.0
NCBI
CONC
HL7V2.5
PHEN
ICNP
ICPC2P
MEDCIN
MEDLINEPLUS
MDR
WHO
CST
DISO
COSTAR
DXP
ICD10CM
CCPSS
CCS
ICD9CM
MTHMST
PCDS
HCPT
NIC
ICF
CPT
GO
ICD10PCS
ANAT
FMA
UWDA
Dimension 1 λ=0.752 (17.0%)
Dimension 2 λ=0.694 (15.6%)
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0

Conclusion

Combine complementary source classification techniques

Support system developers selecting source vocabularies

Other use case scenarios?



Thank You!

Dr. Olivier Bodenreider
Dr. Bastien Rance
Ms. May Cheh