

# Data Mining em Ensaios Clínicos

Sérgio Matos, [aleixomatos@ua.pt](mailto:aleixomatos@ua.pt)  
<http://bioinformatics.ua.pt>

Simpósio CEIC

Ensaios Clínicos: novos desafios, papel social e centros de ensaio

Lisboa, 22 Novembro 2016

# Data mining em ensaios clínicos

[J Med Internet Res](#). 2016 Jul; 18(7): e185.  
Published online 2016 Jul 6. doi: [10.2196/jmir.5549](#)

PMCID: PMC4954919

## Are Randomized Controlled Trials the (G)old Standard? From Clinical Intelligence to Prescriptive Analytics

The incidence of renal replacement therapy (RRT) on ICU was 8% last year	Clinical Intelligence
We intend to treat at least 79 patients with RRT next year	BI-style Intelligence
Propability for 4 new RRT patients this weekend is 80%	Predictive Analytics
Mr. Smith is at high risk for renal failure. Recommend volume expansion	Prescriptive Analytics

From clinical intelligence to prescriptive analytics. BI business intelligence; ICU: intensive care unit.

“Given the constraints on clinical trials, for a majority of clinical questions, **the only relevant data available to aid in decision making are based on observation and experience.**

(...)

We conclude that randomized controlled trials are not at risk for extinction, but **innovations in statistics, machine learning, and big data analytics may generate a completely new ecosystem for exploration and validation.”**



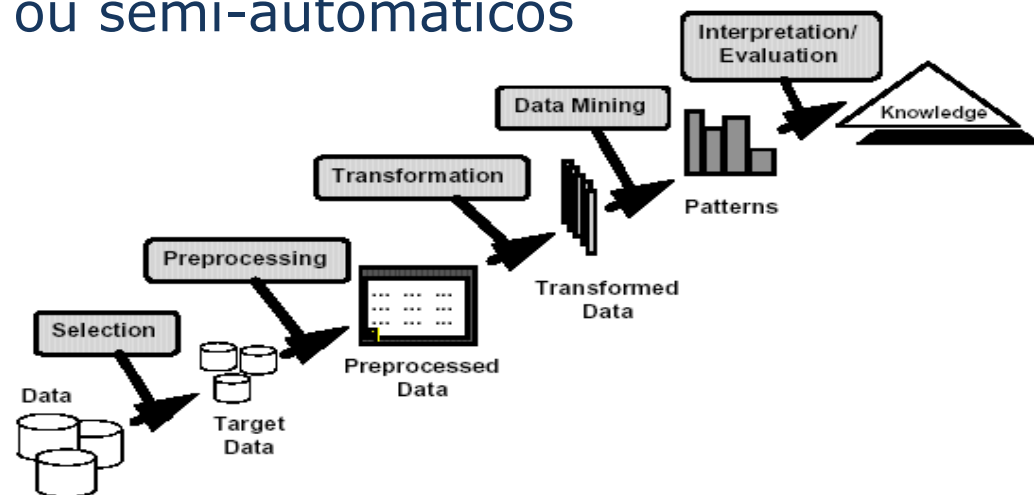


- O que é “Data Mining”?
  - Descoberta de informação não evidente e potencialmente útil, em grandes quantidades de dados
  - Descoberta de padrões através da exploração e análise de grandes quantidades de dados, usando métodos automáticos ou semi-automáticos



# Data Mining

- O que é “Data Mining”?
  - Descoberta de informação não evidente e potencialmente útil, em grandes quantidades de dados
  - Descoberta de padrões através da exploração e análise de grandes quantidades de dados, usando métodos automáticos ou semi-automáticos







- O que é “Data Mining”?
  - Distinto de abordagens tradicionais
    - Quantidade de dados
    - Dimensionalidade
    - Heterogeneidade
  - Tarefas / objetivos
    - Exploração
    - Geração de hipóteses
    - Previsão
    - Classificação / agrupamento de dados
    - Descoberta de associações
    - Detecção de anomalias





- *Data mining* vs. análise estatística
  - Estatística:
    - Experiência definida de forma a validar uma hipótese
    - Dados são recolhidos de acordo com o problema / hipótese
    - Modelo é definido à partida
    - Hipótese é verificada de acordo com o modelo
  - Data mining
    - Processo de análise e modelo são definidos pelos dados
    - “Os meus dados ajudam a explicar esta observação?”
    - “Existem padrões nos dados que permitam melhorar processos?”
    - “É possível construir um modelo para prever resultados futuros?”
    - Uso de aprendizagem automática (*machine learning*)

# Ensaio clínicos



- ↑ Eficácia
- ↓ Risco
- ↓ Custo / “time-to-market”



# Ensaio clínicos



↑ Eficácia  
↓ Risco  
↓ Custo / “time-to-market”



Fatores  
Genéticos  
Ambientais  
Estilo de vida

Diagrama adaptado de: FDA







- Desafios
  - Identificação de alvos / compostos
  - Seleção de participantes
    - Exposição / resultado clínico
    - Diversidade
  - Estratificação
    - Sub-grupos com resposta distinta
    - Quais os fatores?
  - Combinação de ensaios
  - Monitorização pós-aprovação
  - Interação entre fármacos



- Oportunidades
  - Disponibilização de dados de investigação
  - Adoção de EHRs
  - Avanços na sequenciação
  - Publicação de dados de ensaios clínicos

# Data mining em ensaios clínicos

- Oportunidades
  - Disponibilização de dados de investigação
  - Adoção de EHRs
  - Avanços na sequenciação
  - Publicação de dados de ensaios clínicos



The screenshot displays the EMA website with the following elements:

- Logo:** European Medicines Agency, Science Medicines Health.
- Text size:** A A A
- Navigation bar:** Home, Find medicine, Human regulatory, Veterinary regulatory, Committees, News & events, Partners & stakeholders.
- Left sidebar:** Pre-authorisation, Post-opinion, Post-authorisation, What we publish.
- Breadcrumbs:** Home > Human regulatory > Clinical data publication.
- Section Header:** Clinical data publication.
- Main Content:** As of October 2016, the European Medicines Agency (EMA) publishes clinical data submitted by pharmaceutical companies to support their regulatory applications for human medicines under the centralised procedure. This is based on EMA's flagship policy on the publication of clinical data.

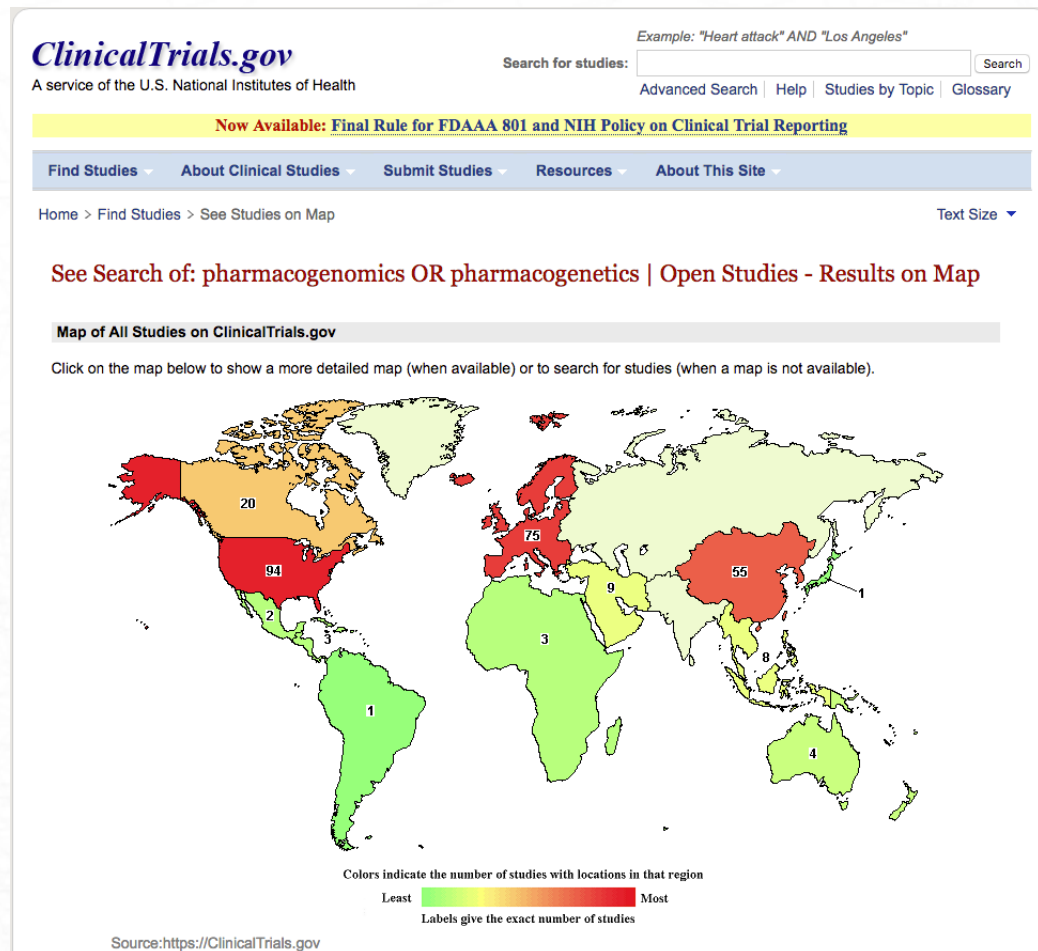




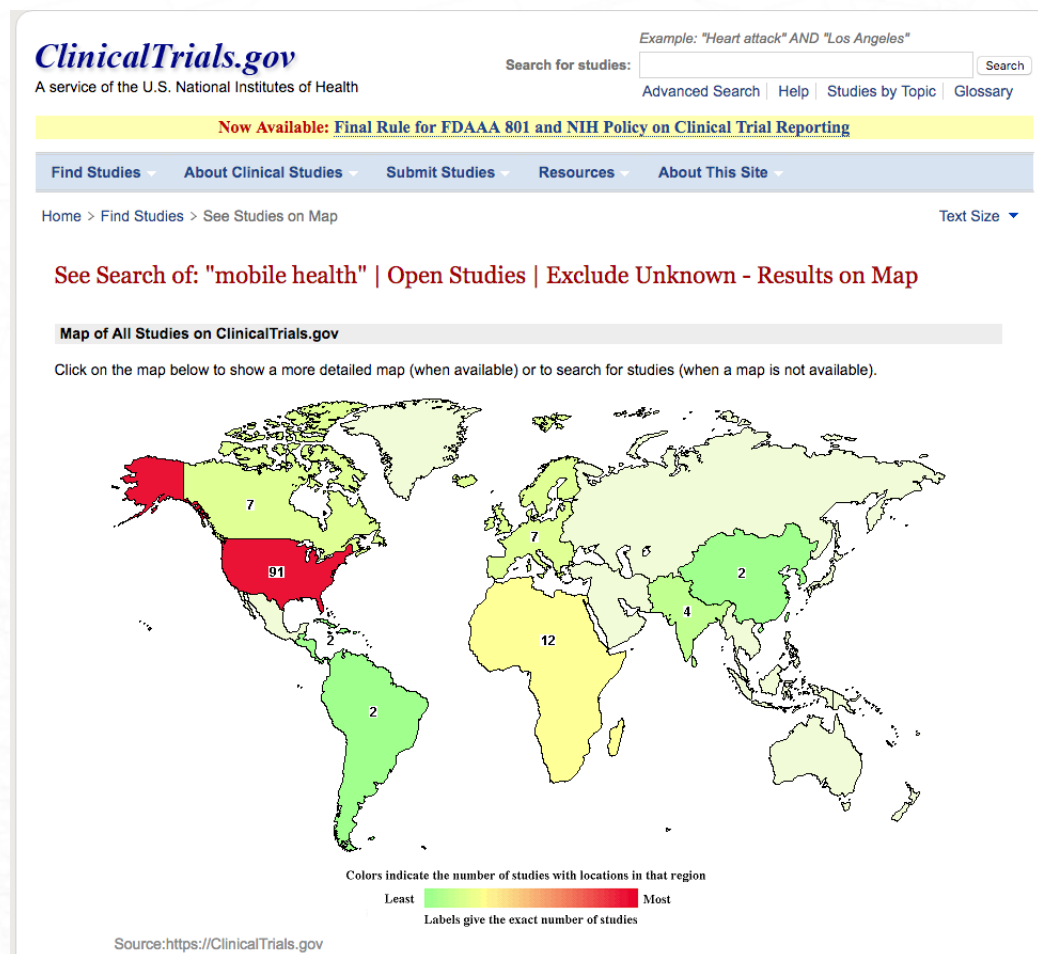
- Oportunidades
  - Disponibilização de dados de investigação
  - Adoção de EHRs
  - Avanços na sequenciação
  - Publicação de dados de ensaios clínicos
  - Partilha de informação pessoal
    - Fóruns de pacientes
    - Redes sociais
    - Dispositivos de monitorização “ligados” (mHealth)
      - Parâmetros cardíacos, respiratórios, atividade física, sono, etc.



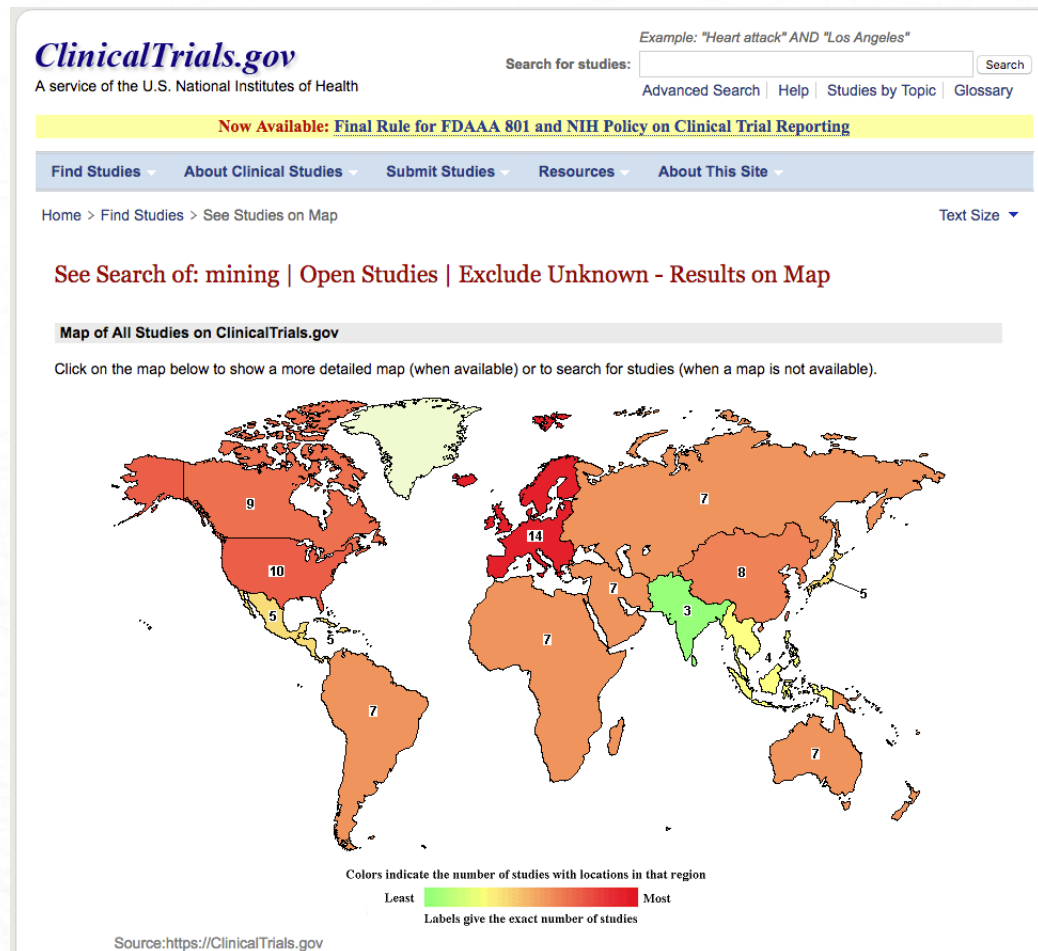
# Data mining em ensaios clínicos



# Data mining em ensaios clínicos



# Data mining em ensaios clínicos





- Oportunidades
  - Disponibilização de dados de investigação
  - Adoção de EHRs
  - Avanços na sequenciação
  - Publicação de dados de ensaios clínicos
  - Partilha de informação pessoal
- **Torna-se necessário integrar toda esta informação**



# Exemplos de aplicação





- Mineração EHR
  - Exposição, presença de doença / sintoma, histórico de família, comorbilidades... informação genética?
  - Identificação de participantes
  - Efeitos adversos
  - Interação entre fármacos
- Acesso a dados distribuídos, privacidade
- Mapeamento de diferentes vocabulários (ICD, SNOMED, ...)
- Mineração texto livre
  - multilingue

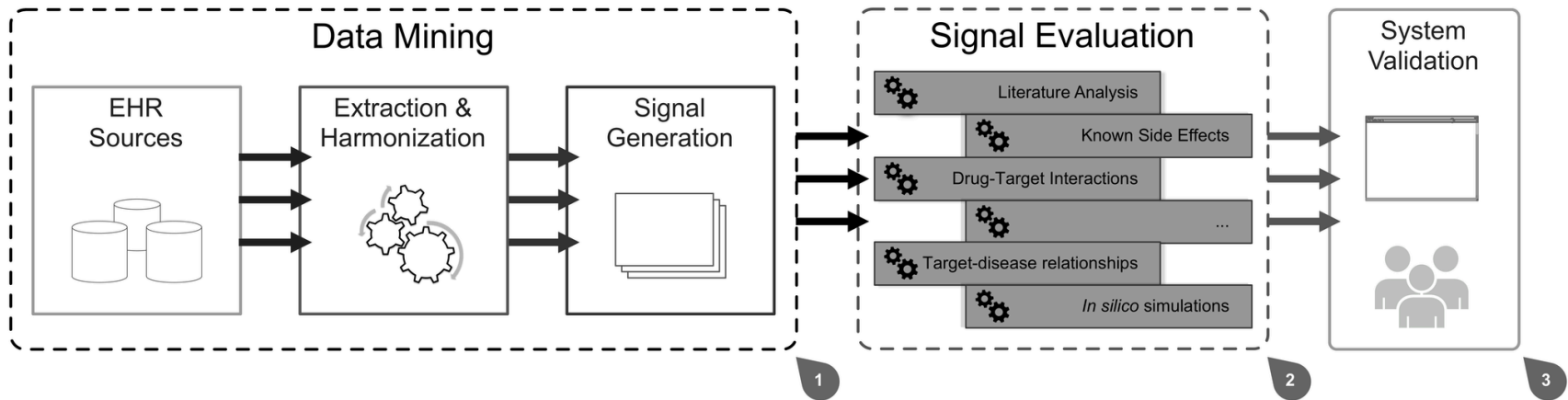
# Mineração EHR

Identificação e substanciação de efeitos adversos

eu-adr



Exploring and Understanding Adverse Drug Reactions By Integrative Mining of Clinical Records and Biomedical Knowledge





## Kibana Drilldown

- Rapid population stats
- Physicians/researchers can quickly analyze data
- Integration with health records
  - Demographics
  - Laboratory testing
  - Comorbidities
  - Treatment information





# Mineração EHR

## Integração de informação farmacogenética

A.

\*\*\*PHARMACOGENETICS CONSULT FOR\*\*\*  
\*CYP2D6 GENOTYPE\*

Sample for CYP2D6 Genotype Obtained: 03/10/2012  
PG4KDS CYP2D6 Genotype Result: \*1/\*1(3N)

Based on the genotype result this patient is predicted to be an ultrarapid metabolizer of CYP2D6 substrates. This patient may require either a dose adjustment to any drug metabolized by CYP2D6 or a therapeutic alternative.

This result signifies that the patient has 3 copies of a wild-type (normal function) allele. This patient may be at risk for an adverse or poor response to medications that are metabolized by CYP2D6. To avoid an untoward drug response, dose adjustments or alternative therapeutic agents may be necessary for medications metabolized by the CYP2D6 enzyme pathway. If codeine is prescribed to an ultrarapid metabolizer, toxic side effects are likely; therefore a therapeutic alternative is recommended. The diplotype result yields a CYP2D6 activity score of 3. For more information about specific medications metabolized by CYP2D6, please go to [www.stjude.org/pg4kds](http://www.stjude.org/pg4kds).

Jane Smith, Pharm.D., pager 1234

Phenotype  
Assignment

Diplotype  
Interpretation

Dosing  
Recommendations

Activity Score

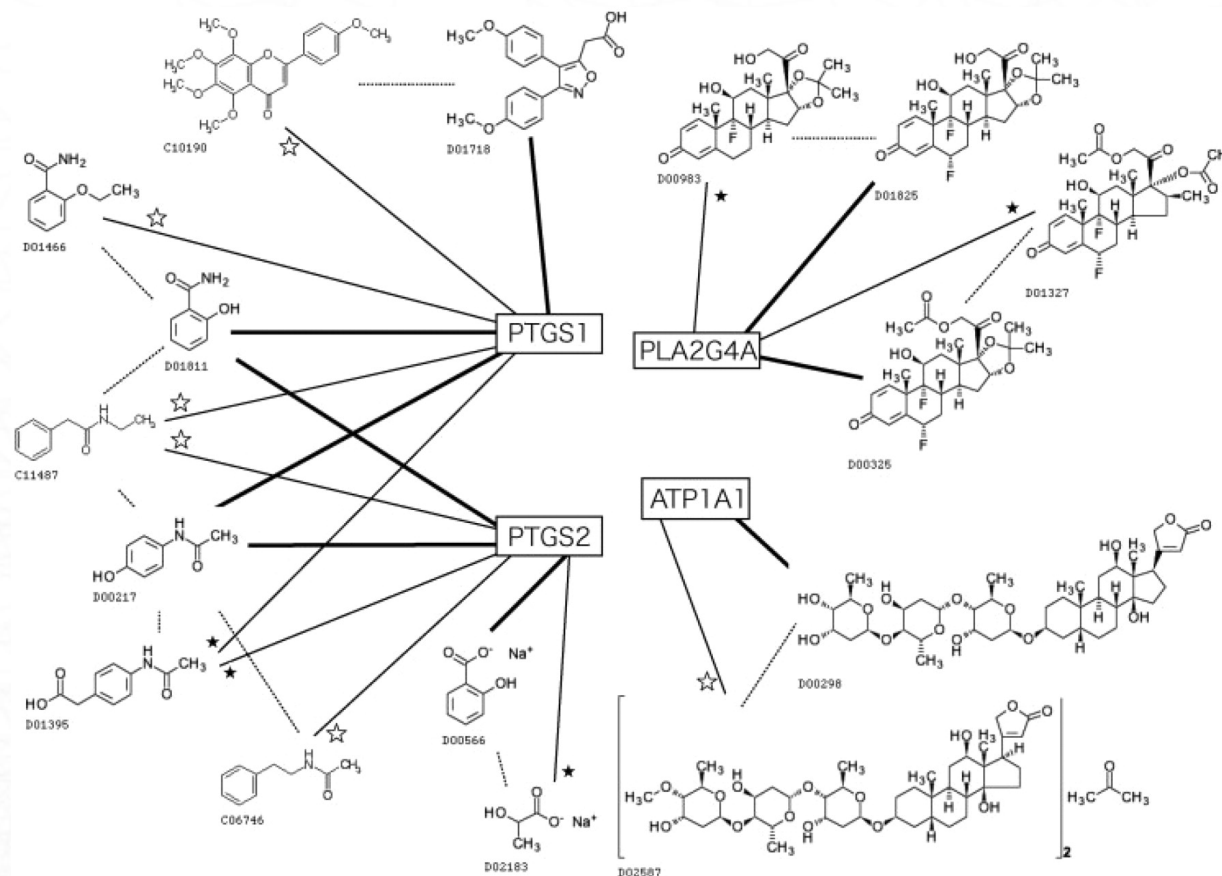
Educational  
Link



# Mineração de dados

Integração e mineração de dados químicos, genômicos, e farmacológicos

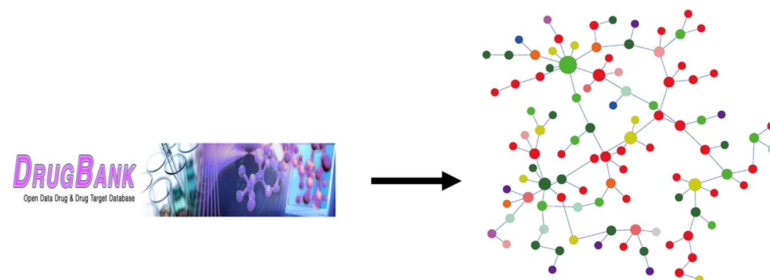
## Examples of the proposed drug–target interactions.



Yoshihiro Yamanishi et al. *Bioinformatics* 2010;26:i246-i254

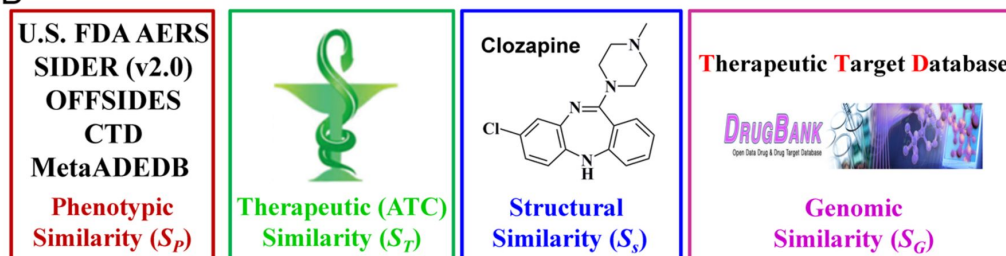
# The heterogeneous network-assisted inference (HNAI) framework for predicting drug–drug interactions (DDI).

A



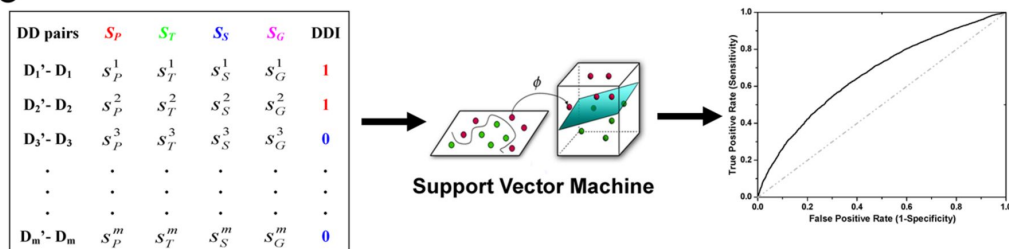
Construction of drug-drug interaction (DDI) network

B



Calculation of drug phenotypic, therapeutic, structural, and genomic similarities

C

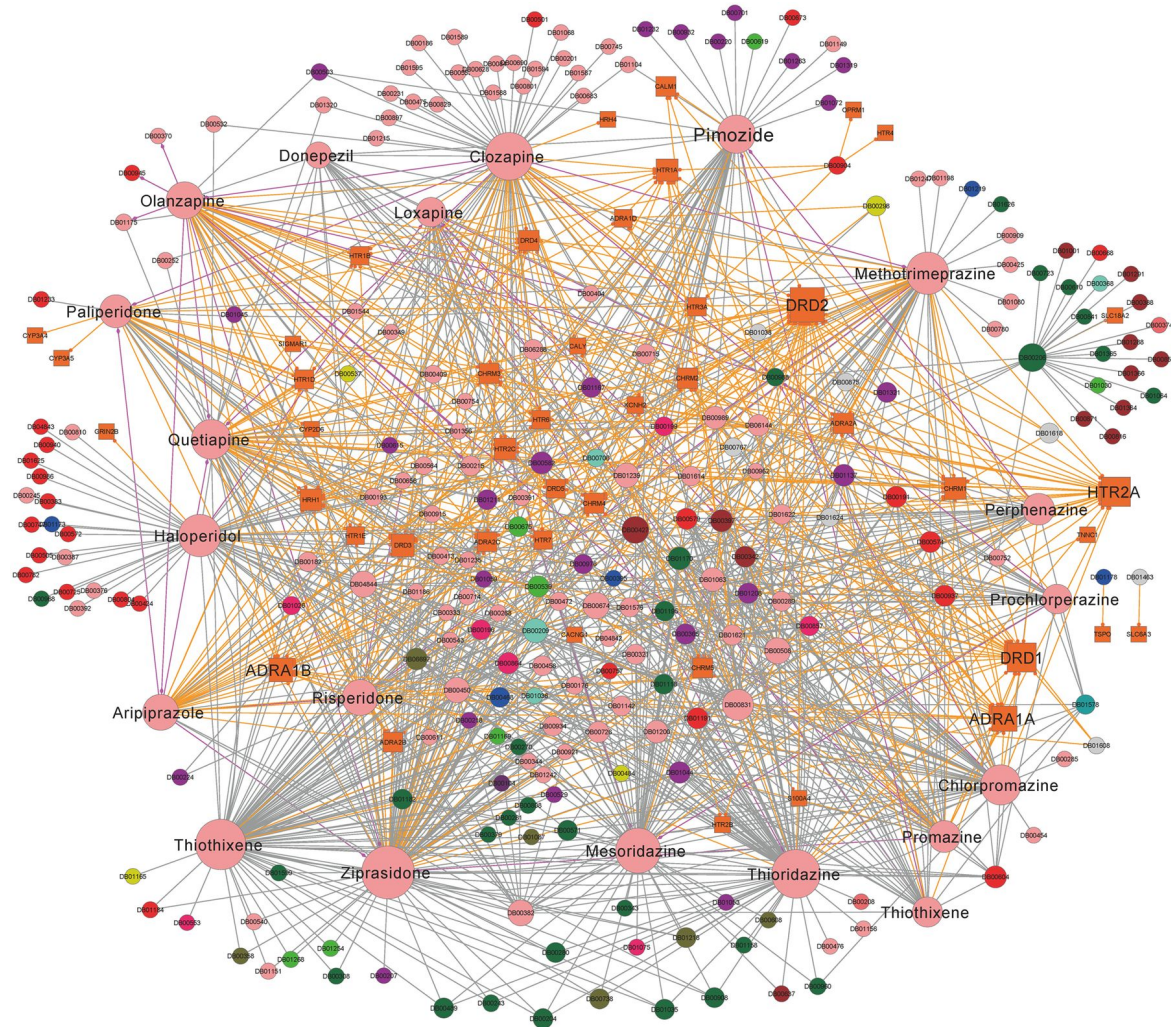


Building heterogeneous network-assisted inference (HNAI) models for DDI prediction

Feixiong Cheng, and Zhongming Zhao J Am Med Inform Assoc 2014;21:e278-e286



## Drug–drug and drug–target interaction network for antipsychotic drugs.

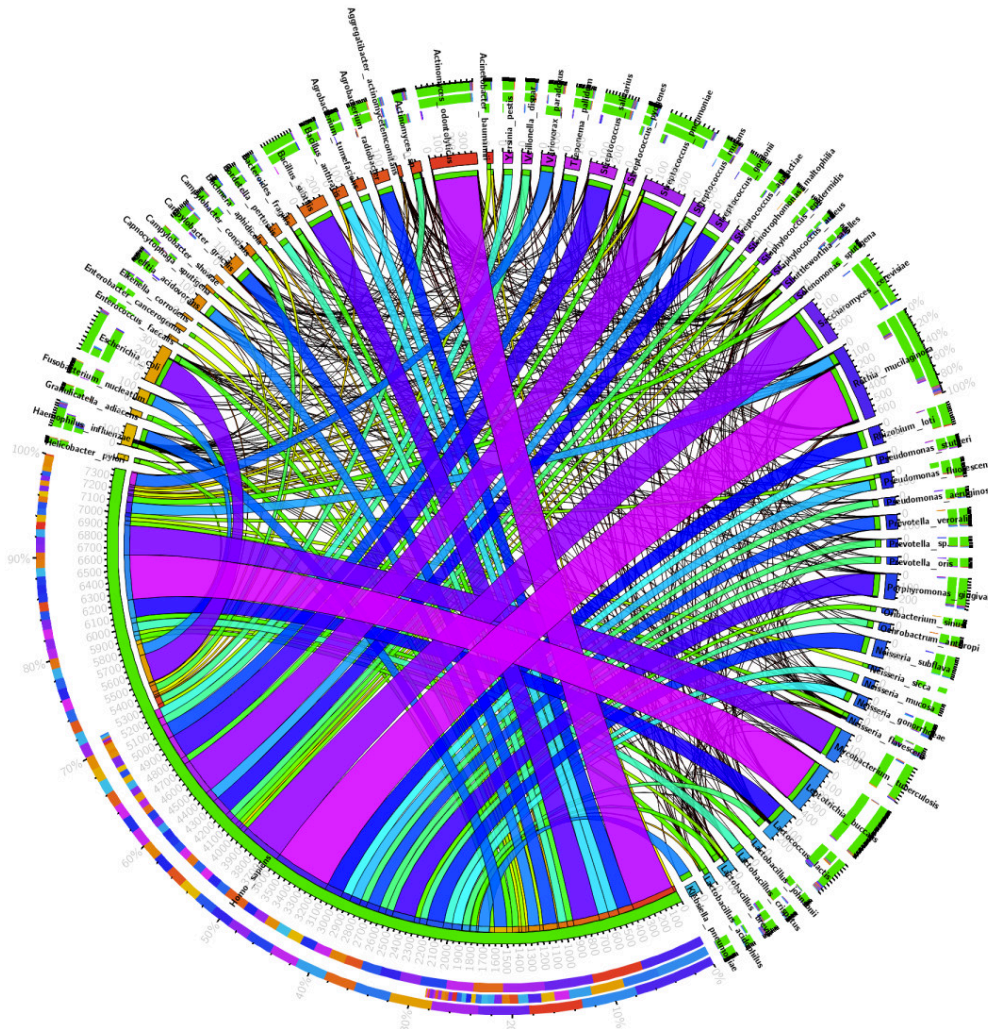


**Feixiong Cheng, and Zhongming Zhao** *J Am Med Inform Assoc* 2014;21:e278-e286



# Mineração de dados

Descoberta de interações entre proteínas



human-microbial oral interactome

Feature	AUC
+ Literature	0.781
+ Sequence	0.877
+ GO	0.817
+ COGs	0.663
+ DDIs	0.620
Final Model	0.926



# Mineração de texto

- Texto clínico
  - Informação em texto livre, não codificada

The CT showed mainly anechoic (cystic) collection in the deep subcutaneous tissues of the lower midline abdominal wall, which measures up to 3.4 cm sagittal x 3 cm AP x 3.8 cm transverse. Some heterogenous hyperechoic material is noted within this mainly cystic collection an abdominal abcess was noted.

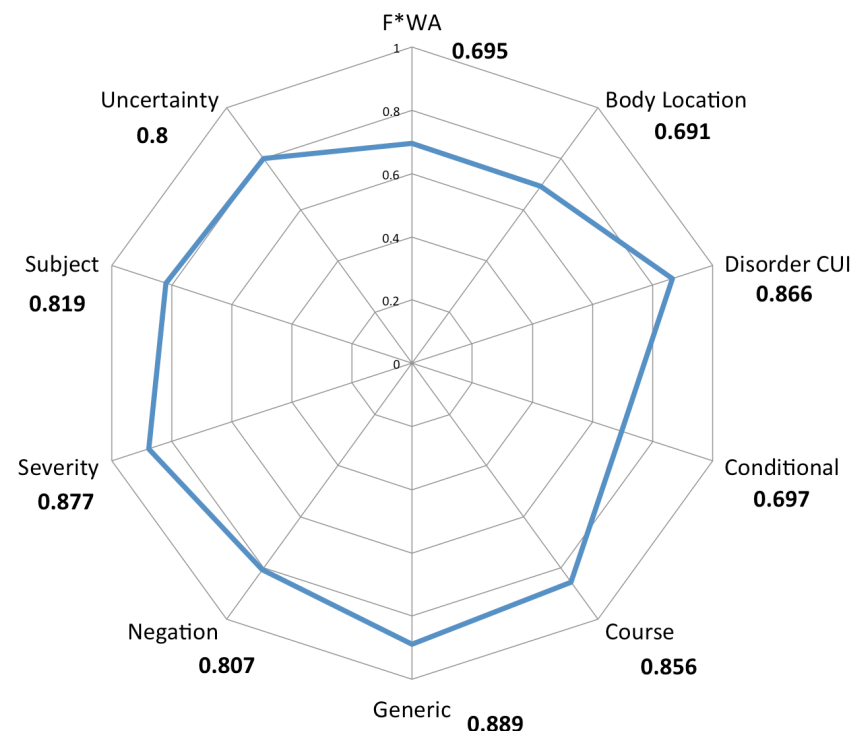
#### Past Medical History:

Crohns s/p 17 surgeries  
Ileostomy 1993; Left abdominal stoma  
ventral Hernia repain x 2  
peripheral neuropathy  
anxiety  
depression

#### Social History:

Disabled  
No ETOh, no smoking (quit 3 months ago; 30 pack year history)

Disease template filling performance





# Mineração de texto

- Texto clínico
  - Informação em texto livre, não codificada

The CT showed **mainly anechoic (cystic) collection** in the **deep subcutaneous tissues** of the **lower midline abdominal wall**, which measures up to **3.4 cm sagittal x 3 cm AP x 3.8 cm transverse**. Some heterogenous hyperechoic material is noted within this mainly cystic collection an **abdominal abcess** was noted.

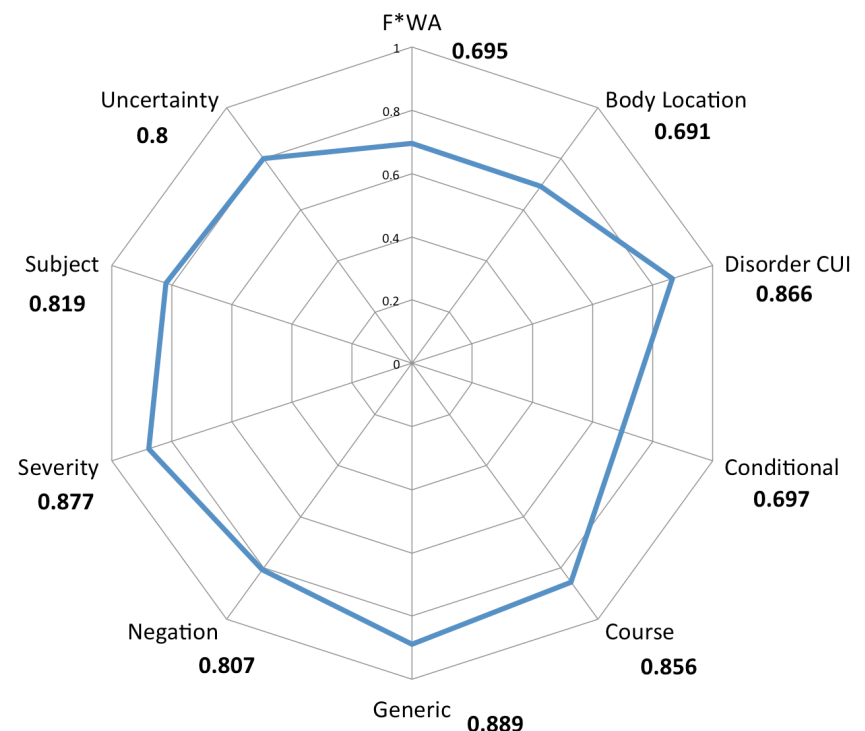
Past **Medical History**:

Crohns s/p 17 surgeries  
Ileostomy 1993; Left abdominal stoma  
ventral Hernia repain x 2  
peripheral neuropathy  
anxiety  
depression

Social History:

Disabled  
No ETOh, no smoking (quit 3 months ago; **30 pack year history**)

Disease template filling performance



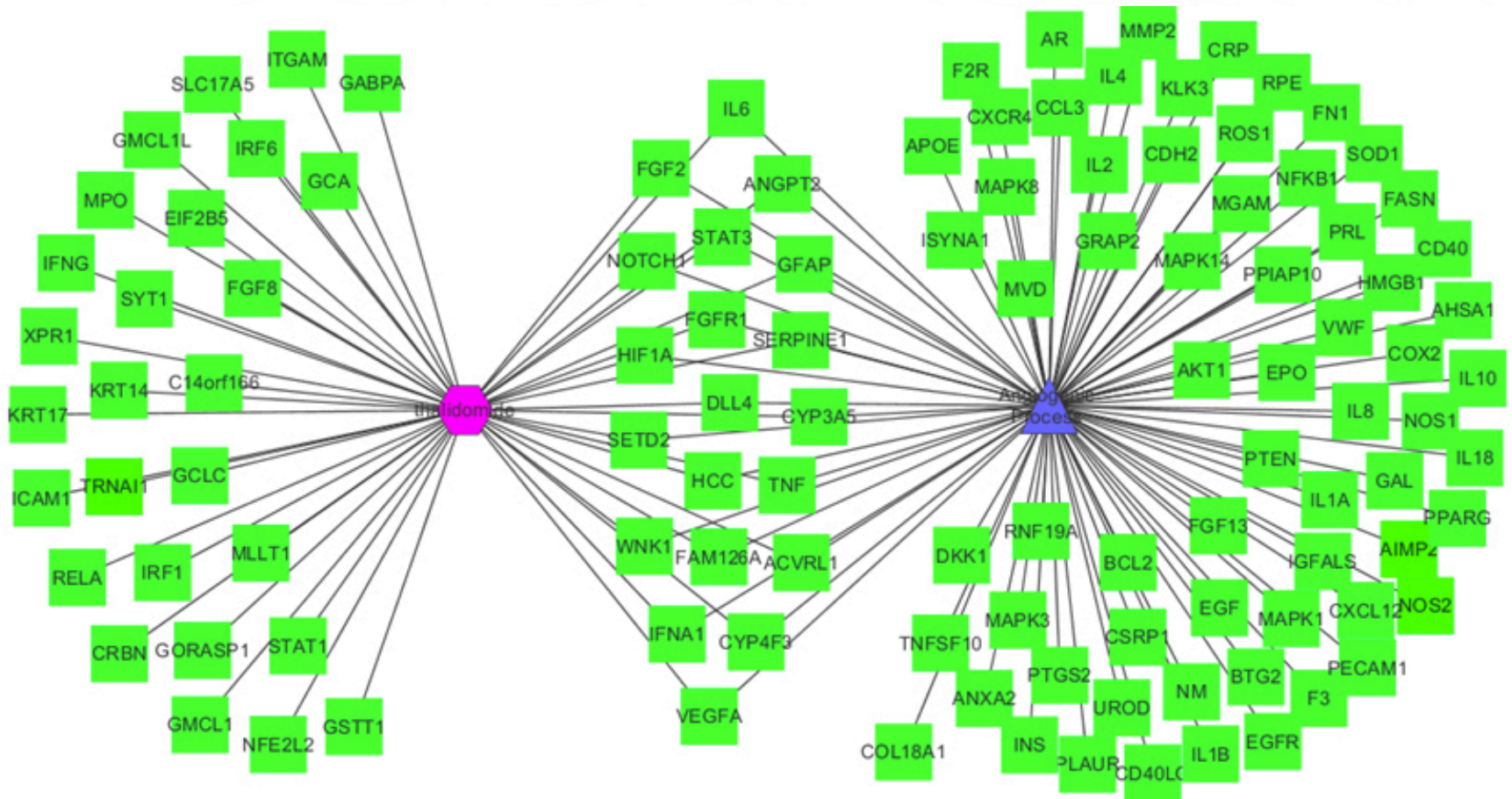


- Texto clínico
  - Informação em texto livre, não codificada
- Literatura científica, patentes
  - Complementar informação das bases de dados
  - Identificação de possíveis alvos terapêuticos
  - Interações entre fármacos
  - Similaridade fármaco-doença
    - Reposicionamento



# Mineração de texto

Identificação de interações e reposicionamento de fármacos



# Mineração de texto

## Enriquecimento semântico

**becas** Annotate Help API Widget About Contact

**HIGHLIGHT**

All None

- ☒ Anatomy
- ☒ Disorders
- ☒ Pathways
- ☒ Chemicals
- ☒ Enzymes
- ☒ Genes and Proteins
- ☒ Molecular Functions
- ☒ Ambiguous

New to becas? [Take the tour](#) »

⊕ Load text

⊕ Expand All ⊖ Collapse All ⇅ Toggle All

- ⊕ Anatomy ( 1 )
- ⊕ Disorders ( 4 )
- ⊕ Pathways ( 1 )
- ⊕ Chemicals ( 9 )
- ⊕ Enzymes ( 2 )
- ⊕ Genes and Proteins ( 4 )
- ⊕ Molecular Functions ( 1 )

Carbamate analogs of thiaphysovene, pharmaceutical compositions, and method for inhibiting cholinesterases Substituted carbamates of tricyclic compounds which have a cyclic sulfur atom, having the formula:(See formula I) wherein R1 is H or a linear or branched chain C1- C10 alkyl group; and R2 is selected from the group consisting of a linear or branched chain -C1-C10 alkyl group, and (See formula I) wherein R3 and R4 are independently selected from the group consisting of H and a linear or branched chain C1-C10 - alkyl group; and with the proviso that when one of R1 or R2 is a H or a methyl group the other of R1 or R2 is not H and optical isomers of the 3aS series, provide highly potent and selective cholinergic agonist and blocking activity and are useful as pharmaceutical agents. Cholinergic disease are treated with these compounds such as glaucoma, acetylcholinesterase and butyryl - cholinesterases.

**Myasthenia Gravis** (1 occurrence) ×

Matches 1 concept with 4 external references.

**Disorders ( 1 )**

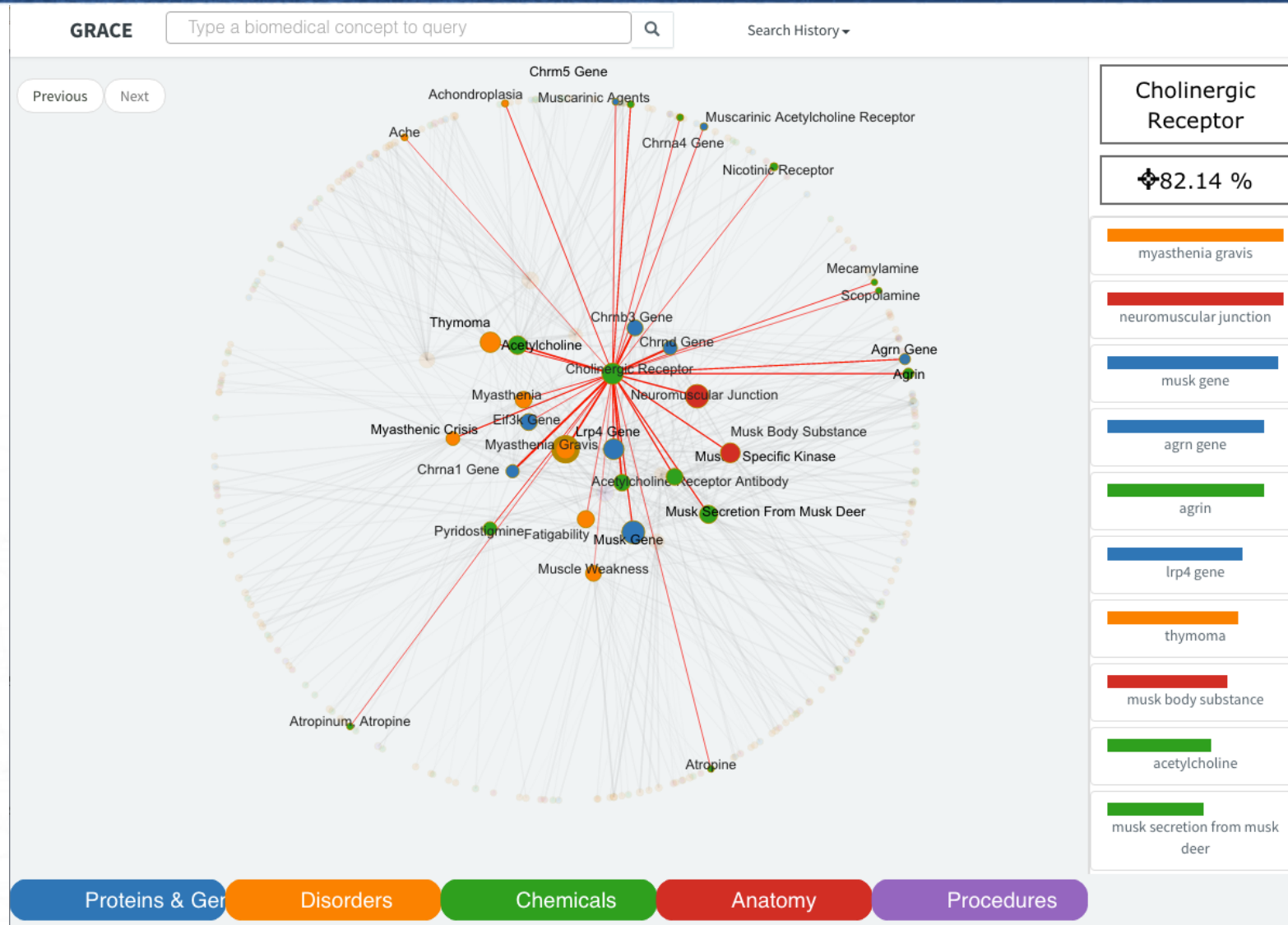
- **Myasthenia Gravis ( 4 )**
- 🔗 NCI:C60989
- 🔗 NCIm:C0026896
- 🔗 SNOMEDCT:91637004
- 🔗 omim.org:254200

📄 Export ▾

**Concept Tree**

# Mineração de texto

## Descoberta de associações





# Mineração de texto

## Identificação de associações

Egag v2.0 © 2015 BMD Software

HGMD IAT Gold Set PMID10072423

Documentation Tools Administrator View Sérgio Matos

1 A wide variety of mutations in the **parkin** gene are responsible for autosomal recessive parkinsonism in Europe.

2 French Parkinson's Disease Genetics Study Group and the European Consortium on Genetic Susceptibility in Parkinson's Disease.

3

4 Autosomal recessive juvenile parkinsonism (AR-JP, PARK2; OMIM 602544), one of the monogenic forms of Parkinson's disease (PD), was initially described in Japan.

5 It is characterized by **early onset** (before age 40), marked response to levodopa treatment and levodopa-induced dyskinesias.

6 The gene responsible for **AR-JP** was recently identified and designated **parkin**.

7 We have analysed the 12 coding exons of the **parkin** gene in 35 mostly European families with **early onset** autosomal recessive parkinsonism.

8 In one family, a **homozygous deletion of exon 4** could be demonstrated.

9 By direct sequencing of the exons in the index patients of the remaining 34 families, eight previously undescribed point mutations (homo

10 The mutations segregated with the disease in the families and were not detected on 110-166 control chromosomes.

11 Four mutations caused truncation of the **parkin** protein.

12 **Gene or Protein**  
Sérgio M. PARK2 HGNC:8607  
Definition: No information available.  
Synonyms: AR-JP; E3 ubiquitin ligase; parkin; parkin RBR E3 ubiquitin protein ligase; PDJ

13 , 255delA and 321-322insGT) and one a nonsense mutation (Trp453Stop).

14 s (Lys161Asn, Arg256Cys, Arg275Trp and Thr415Asn) that probably affect amino acids that a

15 Mutations in the **parkin** gene are therefore not invariably associated with **early onset parkinsonism**.

16 In many patients, the phenotype is indistinguishable from that of **idiopathic PD**.

17 This study has shown that a wide variety of different mutations in the **parkin** gene are a common cause of **autosomal recessive parkinsonism** in Europe and that different types of point mutations seem to be more frequently responsible for the disease phenotype than are deletions.

Concept type: Age, Age of Onset, Disease, Gene or Protein, InDel, Inheritance, Penetrance

Normalization: PARKINSON DISEASE 2, AUTOSOMAL RECESSIVE JUVENILE 600118, PARKINSON DISEASE 14, AUTOSOMAL RECESSIVE 612953, PARKINSON DISEASE 7, early onset autosom

Definition: No information available  
Synonyms: EPDF; PARK2; PARKINSON DISEASE, JUVENILE, AUTOSOMAL RECESSIVE; PARKINSONISM, EARLY-ONSET, WITH DIURNAL FLUCTUATION; PDJ

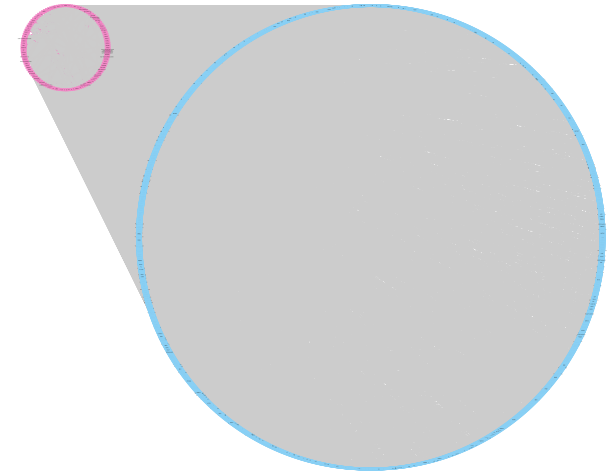
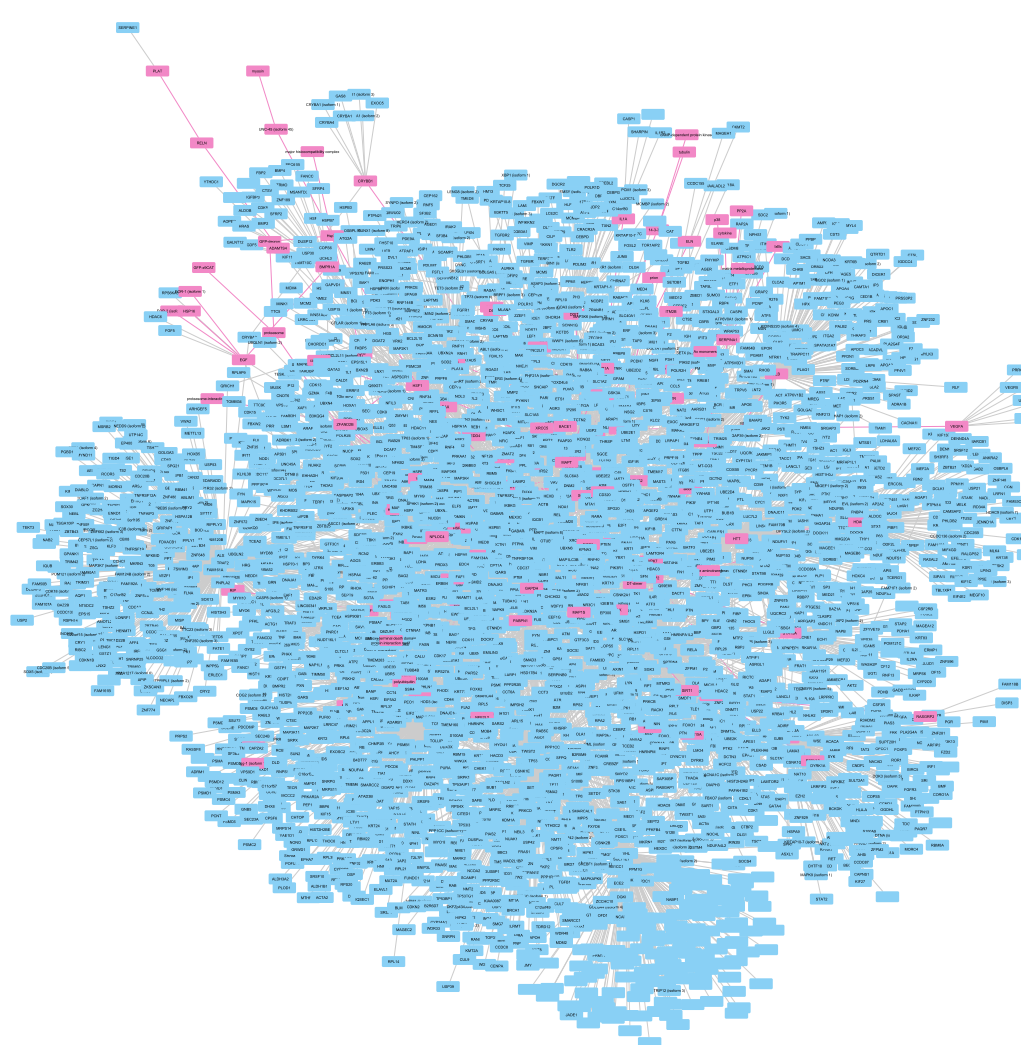
same phenotype as truncating mutations or





# Mineração de texto

## Rede interações entre proteínas



Anotação assistida de 108 resumos MEDLINE  
88 novas PPI envolvendo 116 proteínas



# Mineração de redes sociais

## Reações adversas a medicamentos

	Drug	Disorder
1	caffeine	caffeine stimulant related disorder
2	glipizide	Prostate carcinoma
3	glipizide	Malignant neoplasm of prostate
4	miconazole	Yeast infection
5	glipizide	Memory impairment
6	sertraline	Self hatred
7	cyanocob(III)alamin	Chronic Fatigue Syndrome
8	infliximab	Crohn Disease
9	sertraline	Loss of interest
10	sildenafil	Premature Ejaculation
11	infliximab	Cancer Remission
12	adalimumab	Crohn Disease
13*	glipizide	Psoriasis
14*	glipizide	insanity
15	clomiphene	Infertility

	Drug	Symptom
1	cyanocob(III)alamin	Trembling
2	sildenafil	Memory Loss
3	glipizide	Memory Loss
4	infliximab	Flare
5	cyanocob(III)alamin	Weakness
6	glipizide	Back Pain
7*	melatonin	Trembling
8	adalimumab	Flare
9	cortancyl	Flare
10	cyanocob(III)alamin	Lassitude
11	glipizide	Sleeplessness
12	glipizide	Stomach ache
13	cyanocob(III)alamin	Asthenia
14	testosterone	Blurred vision
15	glipizide	Ache





- Monitorização clínica e pessoal (atividade física)
- Integração com EHR
- Monitorização em ensaios clínicos
- Exemplo: alteração na frequência da tosse (min. 24h) como indicador primário em terapias para doenças respiratórias

- Monitorização clínica e pessoal (atividade física)
- Integração com EHR
- Monitorização em ensaios clínicos
- Exemplo: alteração na frequência da tosse (min. 24h) como indicador primário em terapias para doenças respiratórias



Clinical, Fitness, Wellness

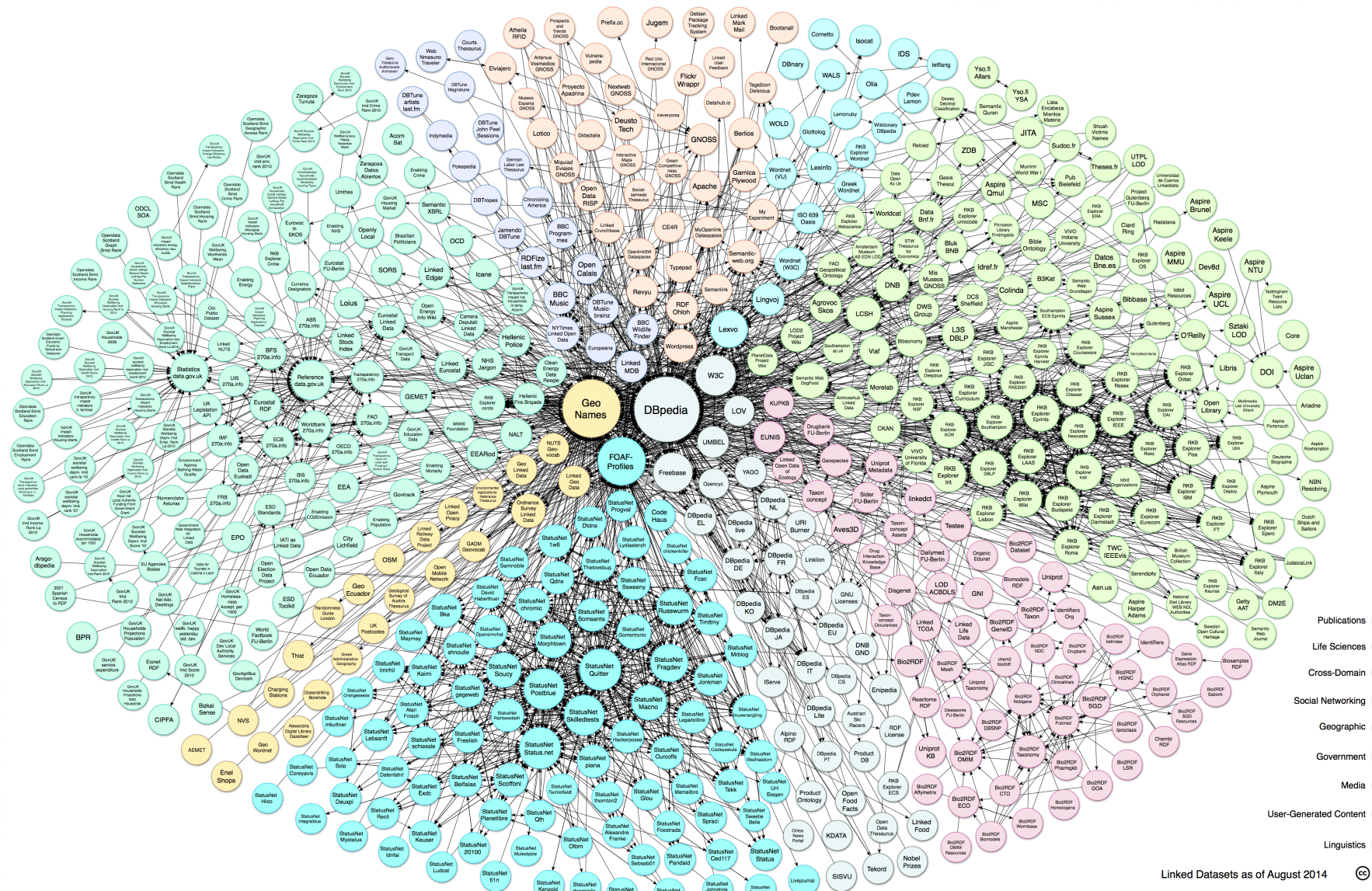






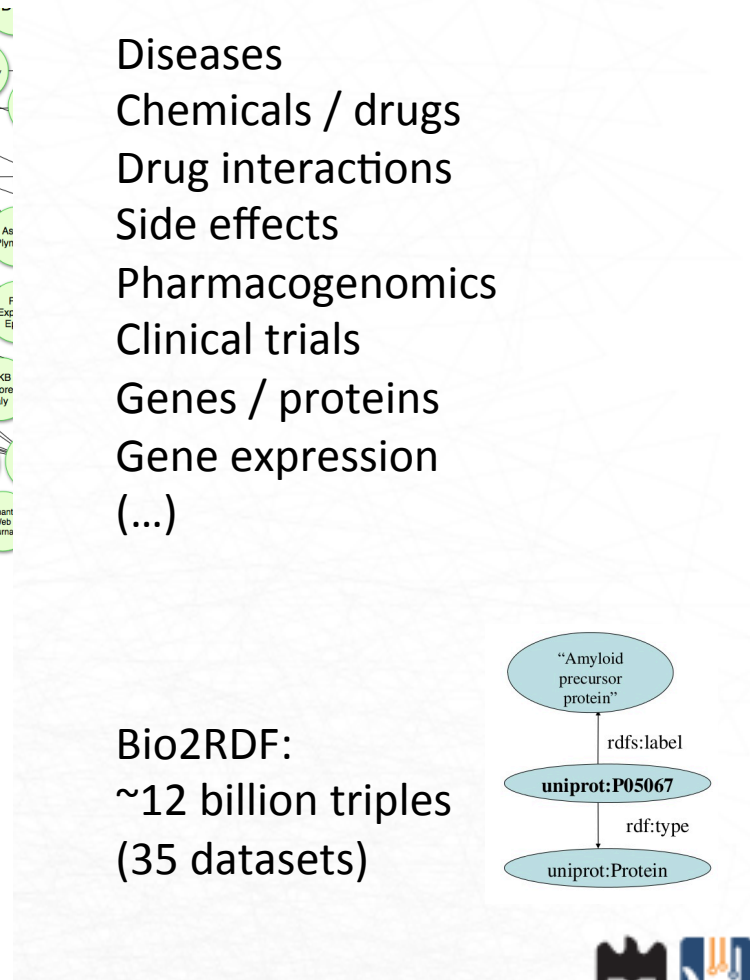
- Oportunidades
  - Disponibilização de dados de investigação
  - Adoção de EHRs
  - Avanços na sequenciação
  - Publicação de dados de ensaios clínicos
  - Partilha de informação pessoal
- **Torna-se necessário integrar toda esta informação**

# Linked Data

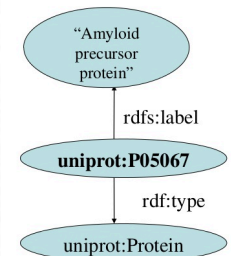




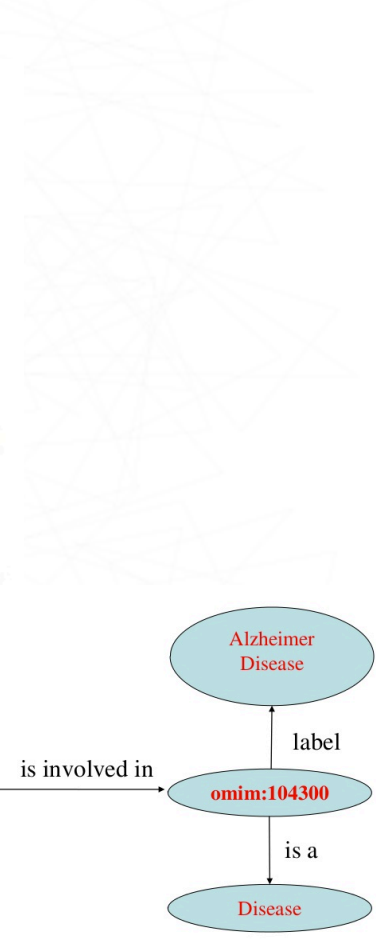
\_\_\_\_\_



Bio2RDF:  
~12 billion triples  
(35 datasets)



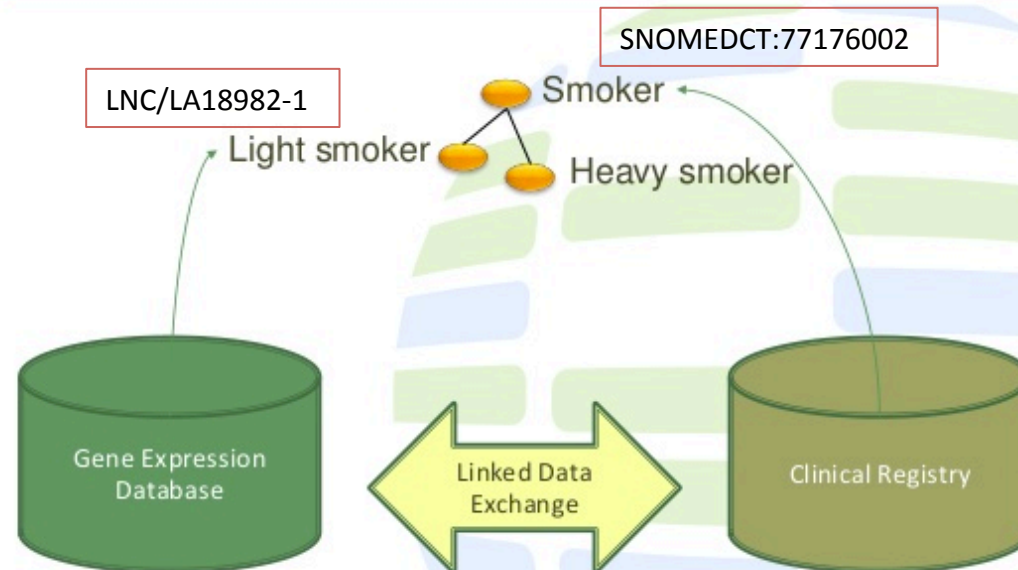






Most common use: common reference

61





- Ensaios clínicos
  - Uso de informação genética, estilo de vida, ambiental
  - Análise exploratória dos dados
    - Melhores resultados, redução de risco
- Mineração de EHR
  - Identificação de participantes em ensaios clínicos
  - Monitorização após aprovação
  - Verificação de interações entre medicamentos
- Mineração da literatura e dados científicos
  - Identificação de potenciais alvos (hipóteses)
  - Complementar e guiar análise de resultados



# Conclusões

Computers might not find the solutions to our problems, but they would be able to do the bulk of the legwork required, assist our human minds in intuitively finding ways through the maze.

Tim Berners-Lee

[sendablequotes.com](http://sendablequotes.com)



# Obrigado!

Sérgio Matos, [aleixomatos@ua.pt](mailto:aleixomatos@ua.pt)  
<http://bioinformatics.ua.pt>

Simpósio CEIC  
Ensaio Clínicos: novos desafios, papel social e centros de ensaio

Lisboa, 22 Novembro 2016

