Lecture 12: Databases & Information Extraction

Olexandr Isayev

Department of Chemistry, CMU

olexandr@cmu.edu

HW5: Regression is Available Now

Kaggle competition

Get data from Kaggle

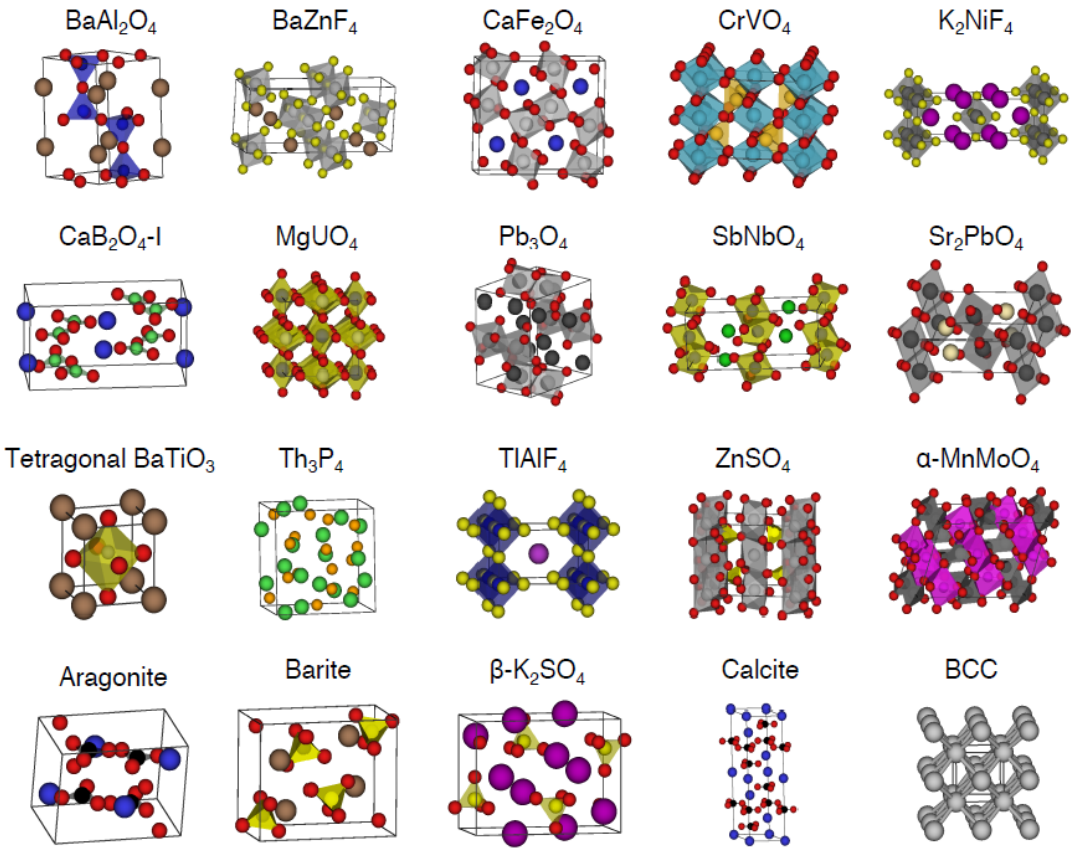
Solve the problem

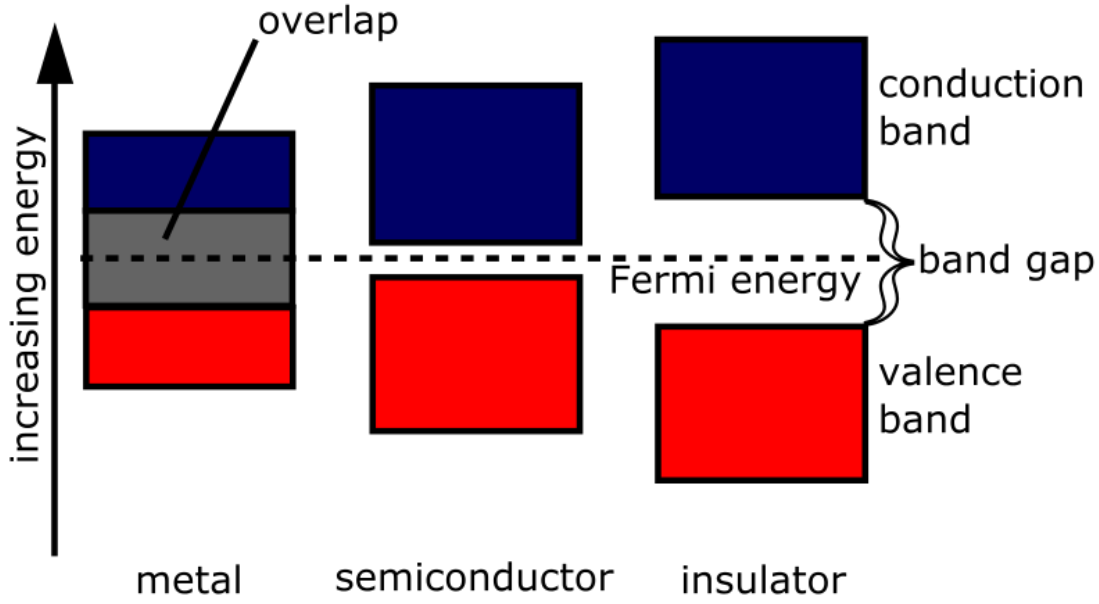
Submit solution for autograding to Kaggle

Submit IPYNB file to canvas

Due: Sun December 3

Problem: predict band gap in solid state





HW5: Regression

Build any regression model, focus on model engineering

Find BEST model

Use your model to score Ytest on Leaderboard

Recital on Thu after Thanksgiving

Class project presentation

Thursday, December 8

- Plan for ~10 min presentation
- Peer grading!

Class project presentation

- Background
- Scientific Problem
- Dataset
- Data curation and processing (if any)
- ML experiments
- Obtained results
- Lessons Learned

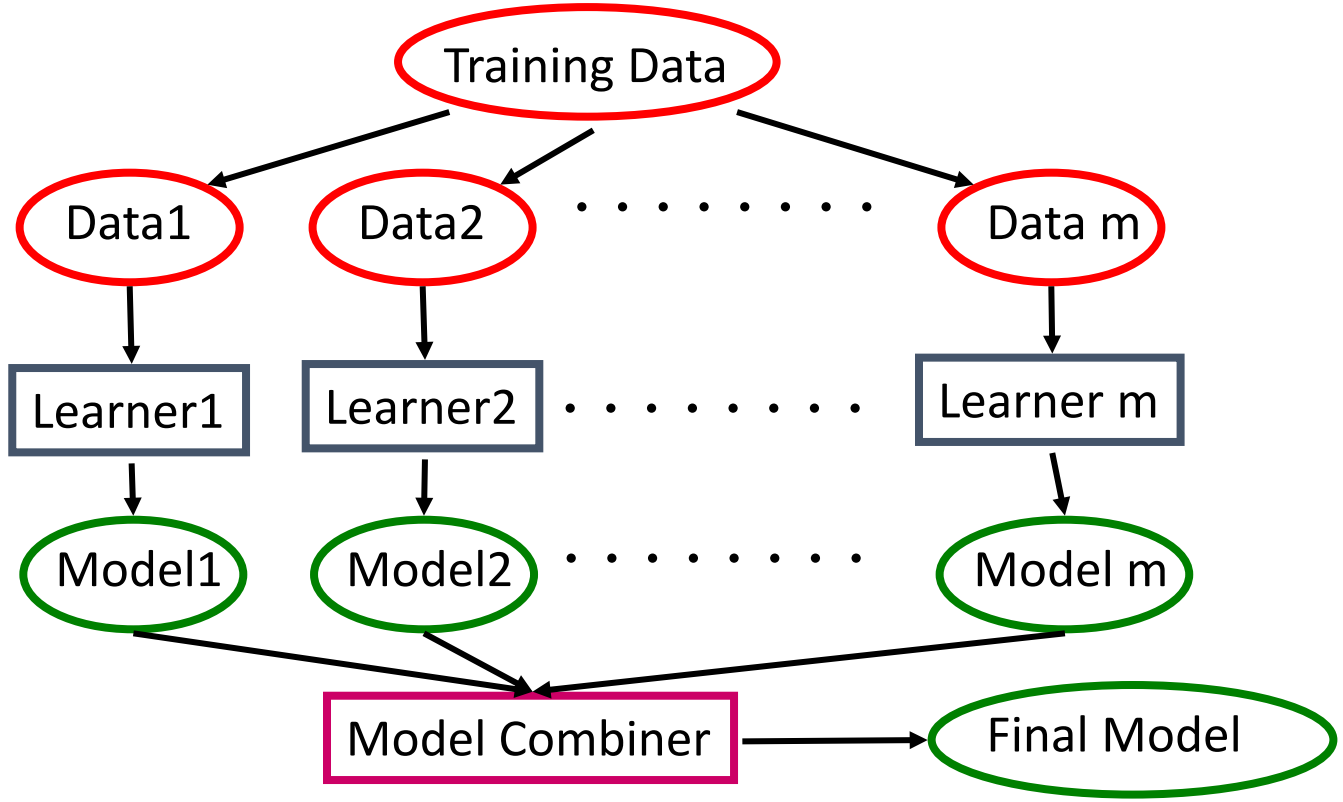
- Go into Jupyter and show important moments if you like

Project Deliverables

- By December 10, please submit to Canvas:
- Code implementing your project experiments
- Presentation slides
- Project report ~2 pages
  - Summarize your project experience, results, and lesson learned

Learning Ensembles

- Learn multiple alternative definitions of a concept using different training data or different learning algorithms.
- Combine decisions of multiple definitions, e.g. using weighted voting.



Source: Ray Mooney

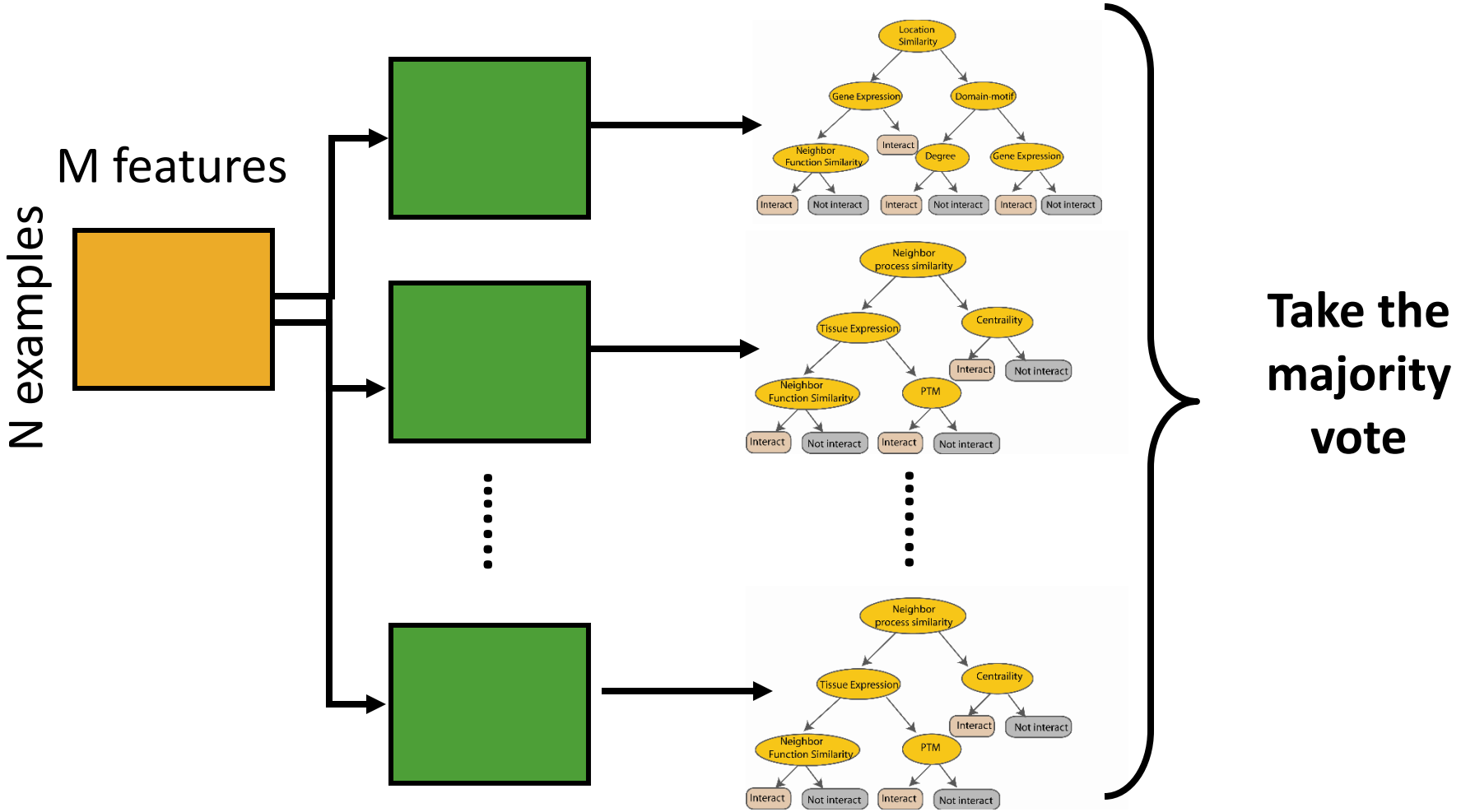
Bagging

- Bagging or *bootstrap aggregation* a technique for reducing the variance of an estimated prediction function.

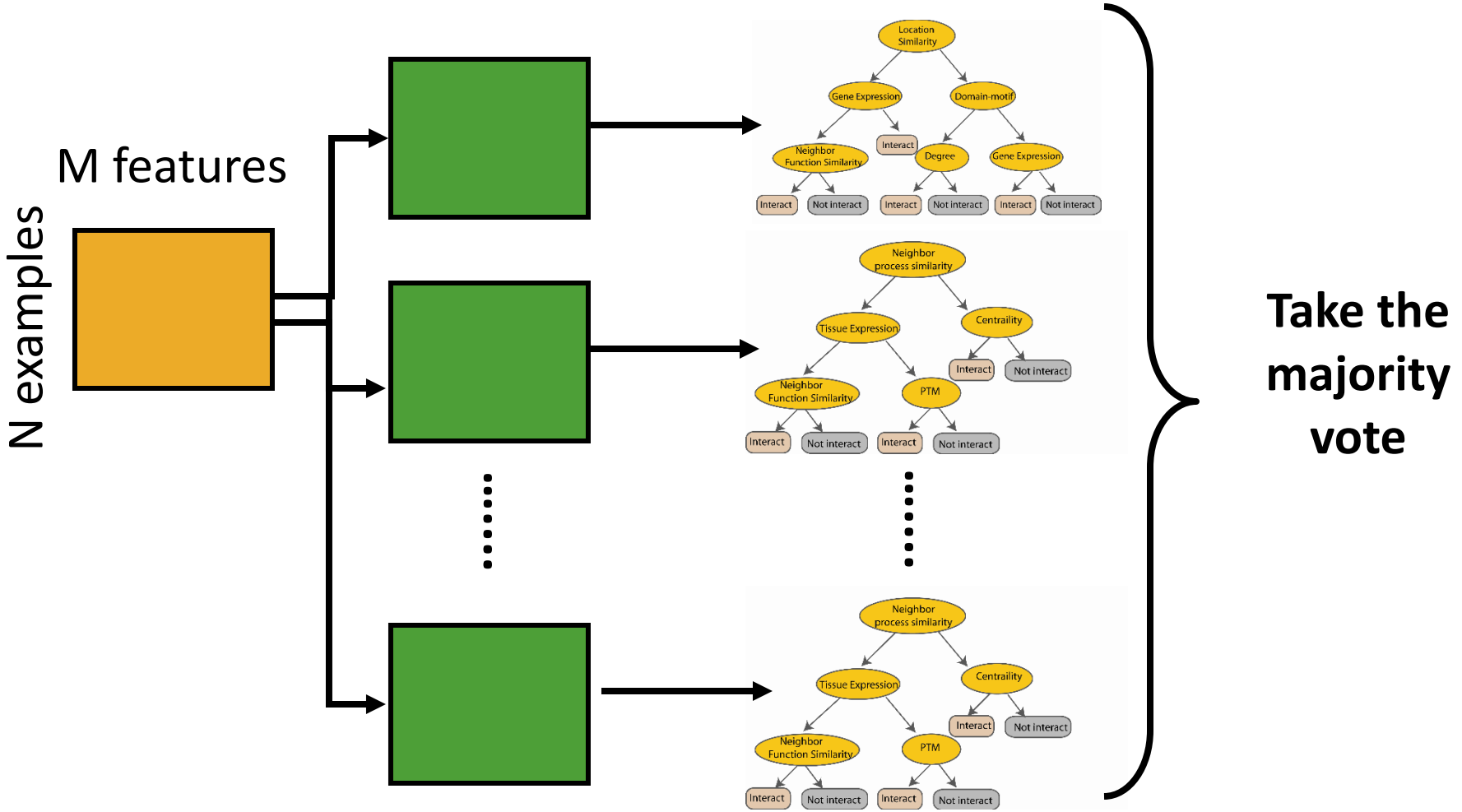
- For classification, a *committee* of trees each cast a vote for the predicted class.

- For regression: average/mean of predicted numeric values

Bagging Classifier



Random Forest Classifier



RF Limitations

- Can’t extrapolate
- Fundamentally discrete algorithm. Not very good for smooth & continuous outputs
- Not good for Oblique/curved frontiers
  - Staircase effect
  - Many pieces of hyperplanes

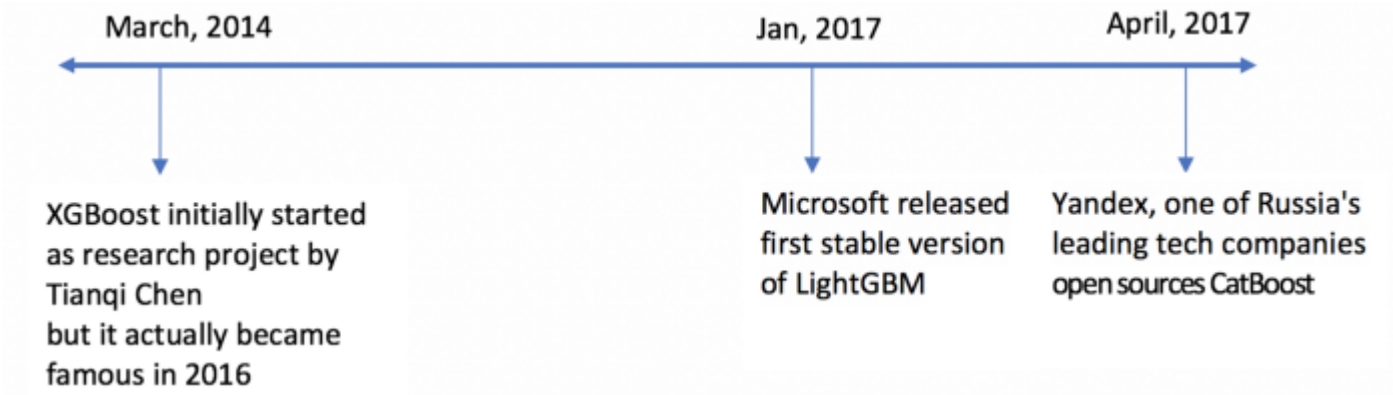
Boosting

- A sequential and adaptive method
- The previous model informs the next.
- Initially, all observations are assigned the same weight. If the model fails to classify certain observations correctly, then these cases are assigned a heavier weight so that they are more likely to be selected in the next model.
- In the subsequent steps, each model is constantly revised in an attempt to classify those observations successfully.
- While bagging requires many independent models for convergence, boosting reaches a final solution after a few iterations.

Comparing bagging and boosting

	Bagging	Boosting
Sequent	Two-step	Sequential
Partitioning data into subsets	Random	Give misclassified cases a heavier weight
Sampling method	Sampling with replacement	Sampling without replacement
Relations between models	Parallel ensemble: Each model is independent	Previous models inform subsequent models
Goal to achieve	Minimize variance	Minimize bias, improve predictive power
Method to combine models	Weighted average or majority vote	Majority vote
Requirement of computing resources	Highly computing intensive	Less computing intensive

Development of Boosting Machines

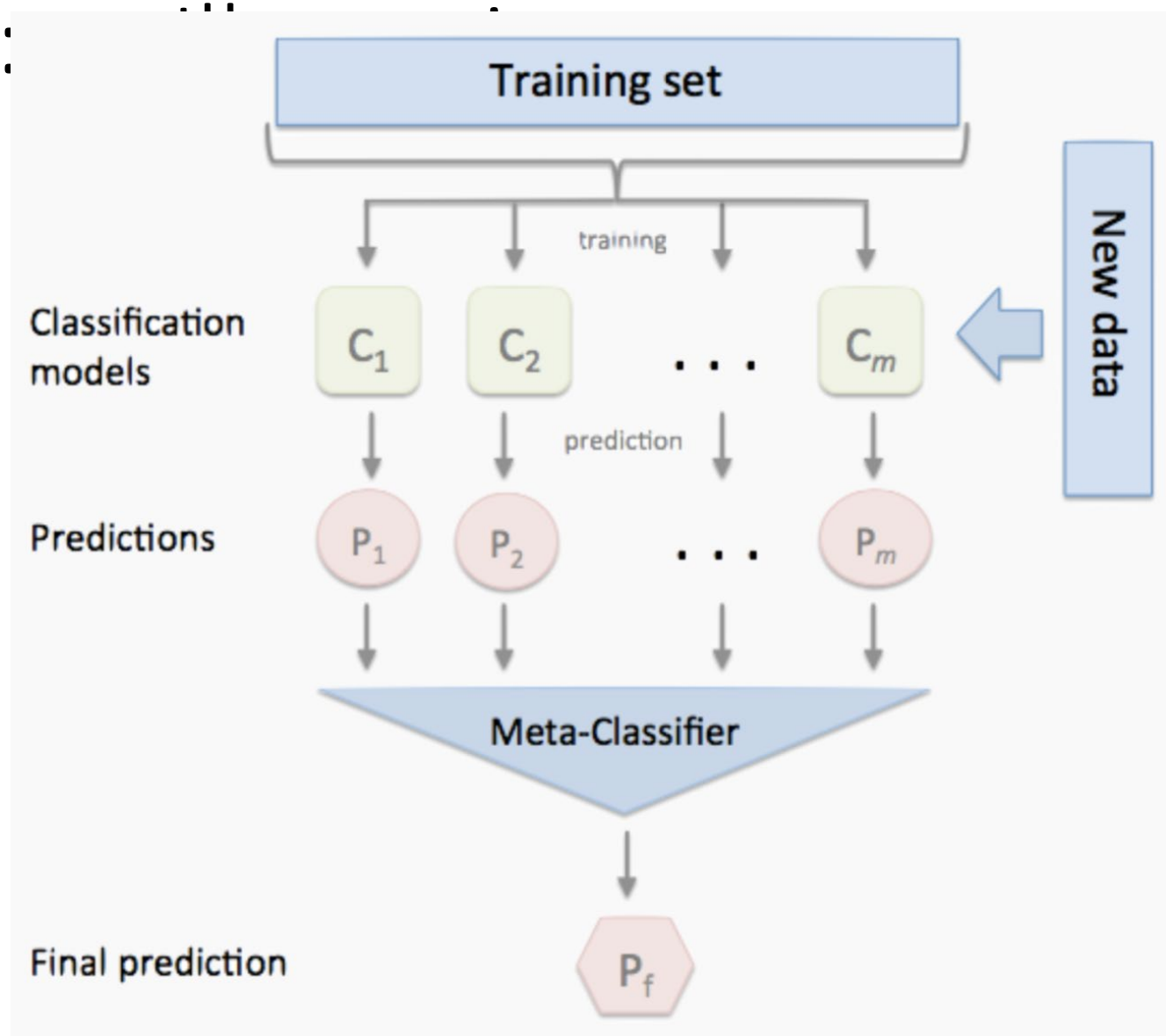


Gradient Boosting Machine (GBM)
GBM combines predictions from multiple decision trees, and all the weak learners are decision trees. The key idea with this algorithm is that every node of those trees takes a different subset of features to select the best split. As it's a Boosting algorithm, each new tree learns from the errors made in the previous ones.

Light Gradient Boosting Machine (LightGBM)
LightGBM can handle huge amounts of data. It's one of the fastest algorithms for both training and prediction. It generalizes well, meaning that it can be used to solve similar problems. It scales well to large numbers of cores and has an open-source code so you can use it in your projects for free.

Categorical Boosting (CatBoost)
This particular set of Gradient Boosting variants has specific abilities to handle categorical variables and data in general. The CatBoost object can handle categorical variables or numeric variables, as well as datasets with mixed types. That's not all. It can also use unlabeled examples and explore the effect of kernel size on speed during training.

Stacking:


Training set
training
Classification models
C_1
C_2
. . .
C_m
New data
prediction
Predictions
P_1
P_2
. . .
P_m
Meta-Classifier
Final prediction
P_f

Information Extraction

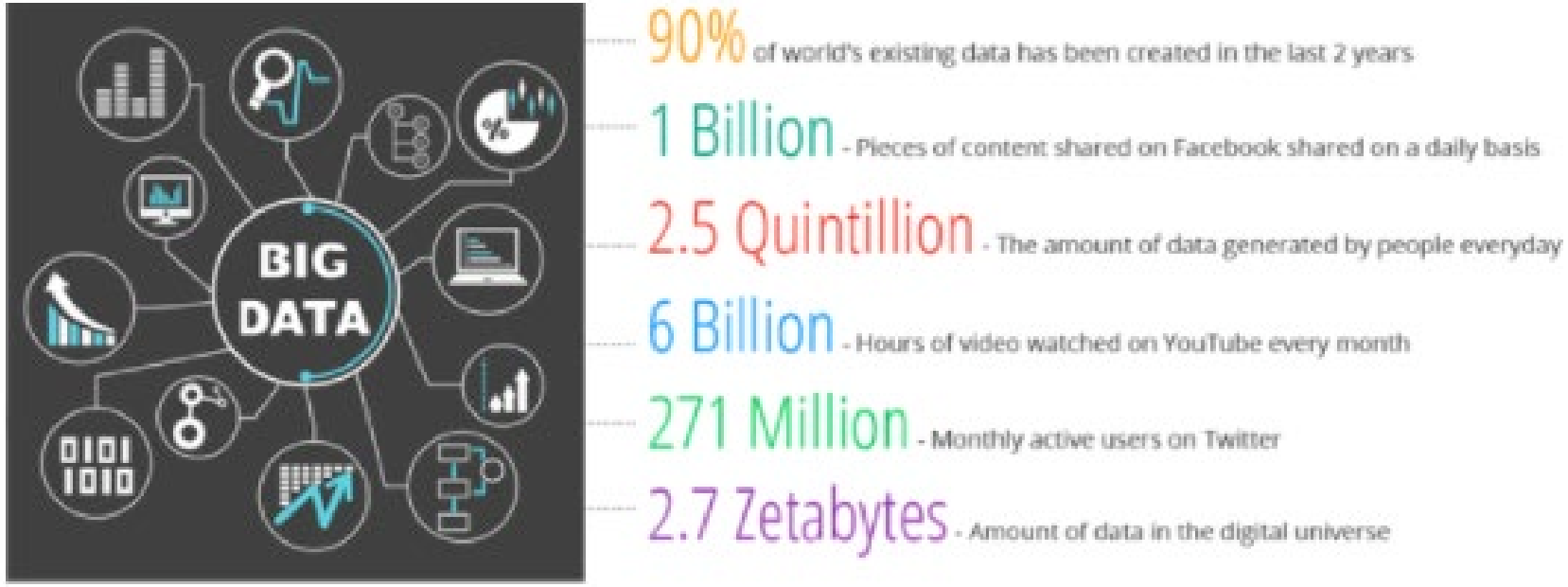
1. Extracting knowledge from unstructured data (e.g., text)

2. Recognize Named Entities in unstructured data

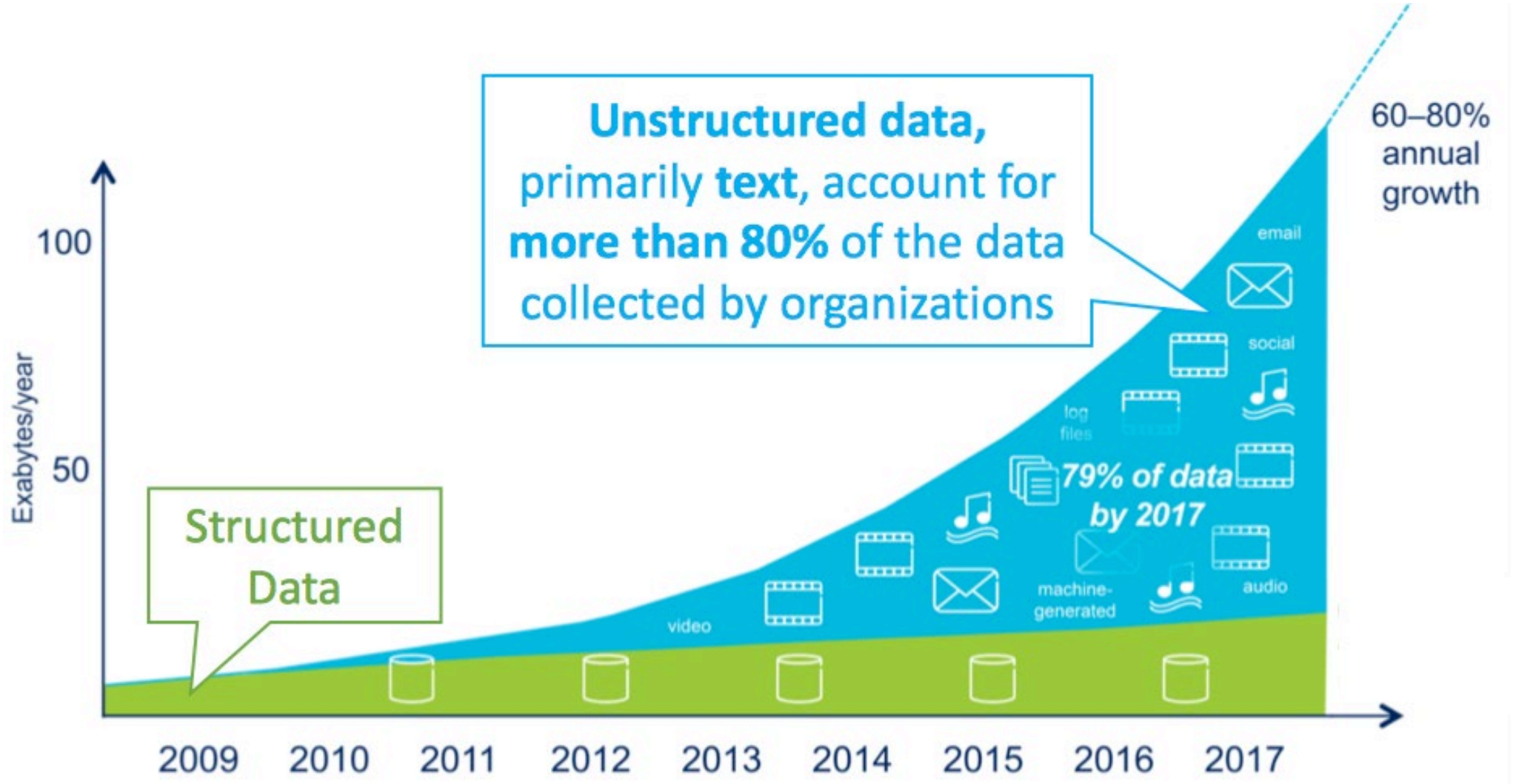
3. Clean and normalize extractions

What is Information Extraction?

Goal: Mine knowledge from unstructured data



Growth of Unstructured Text Data



Knowledge in unstructured data

...

News
Social media post
Web pages
...

Get overview of recent news events

...

Financial reports
Medical records
Legal acts
...

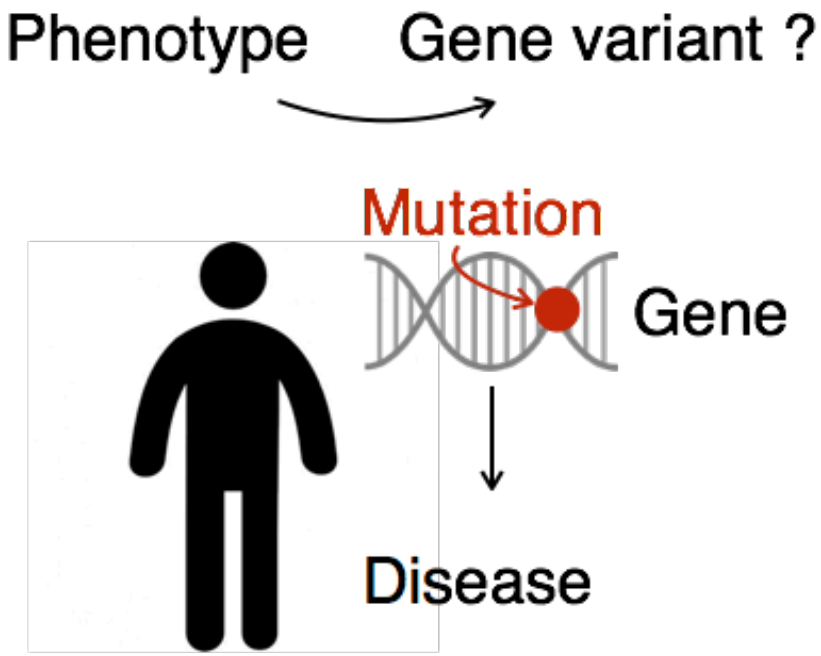
Obtain insights from data for decision support

...

Customer reviews
Tech support memos
Field service notes

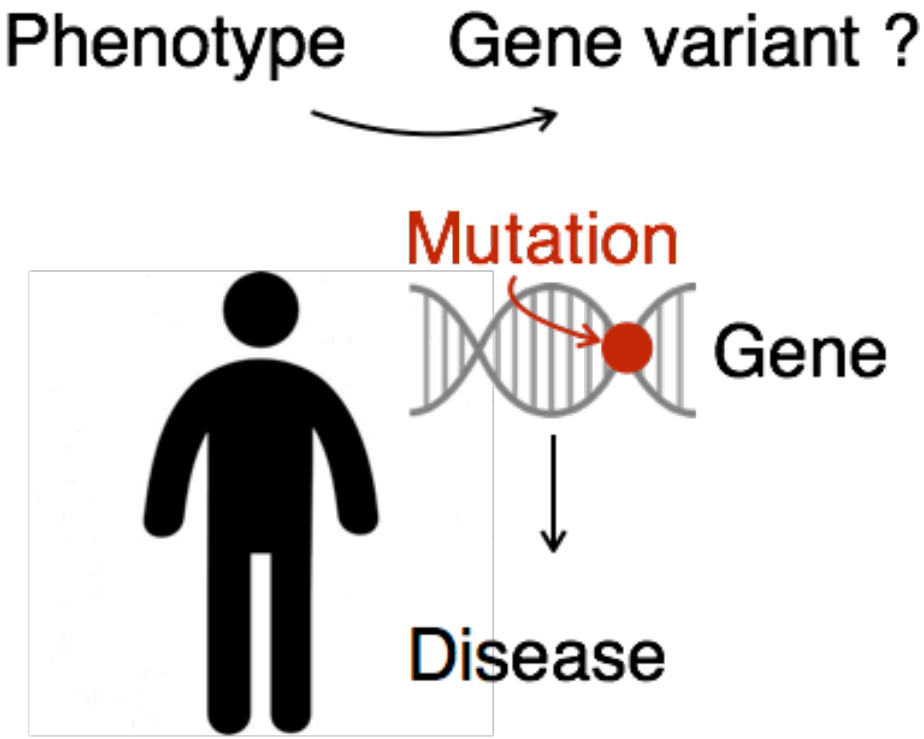
Summarize user feedbacks for quality control

Knowledge from Unstructured Data (Example)



STANFORD
HOSPITAL & CLINICS
Stanford University Medical Center

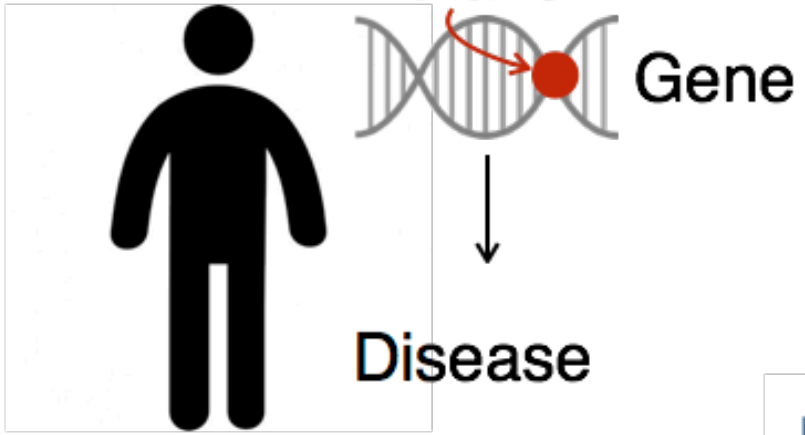
Personalized medicine



Which gene is at fault?

Personalized medicine

Phenotype → Gene variant ?



Find right article (1hr/variant)

Which gene is at fault?

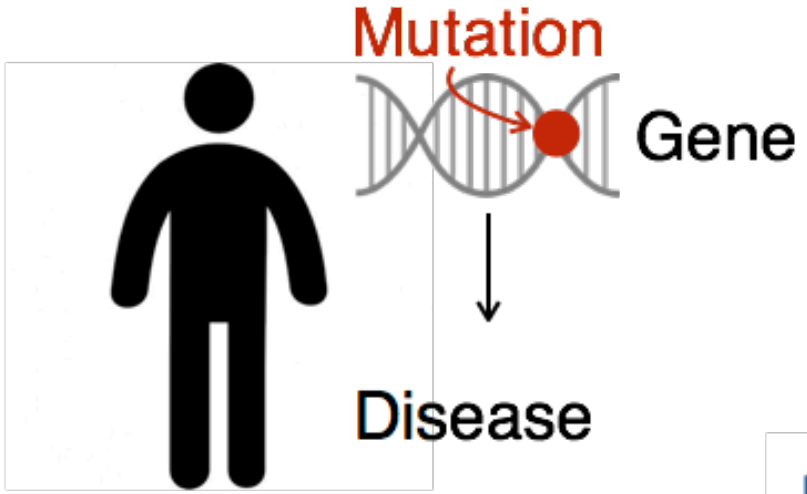
PubMed
National Library of Medicine NLM

25 million articles

Doctor Google

Personalized medicine

Phenotype → Gene variant ?



Find right article (1hr/variant)

Which gene is at fault?

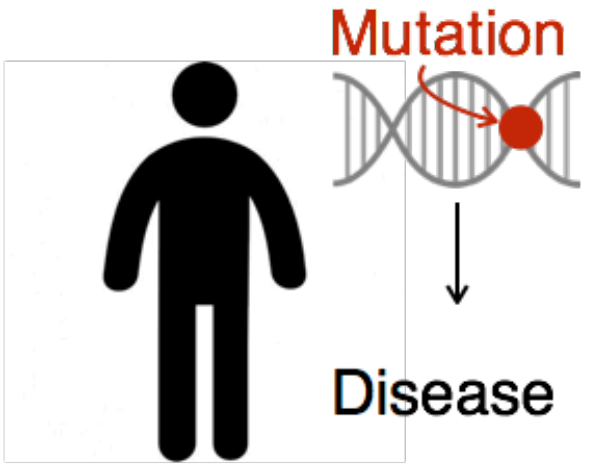
PubMed
National Library of Medicine NLM

25 million articles

Can we build a machine to read these articles?

Personalized medicine

Phenotype → Gene variant ?



Query KB
(Instantaneous)

Which gene is at fault?

Knowledge Base

Gene	Phenotype
DEAF1	Intellectual Disability

Knowledge Base Construction (KBC)

DeepDive

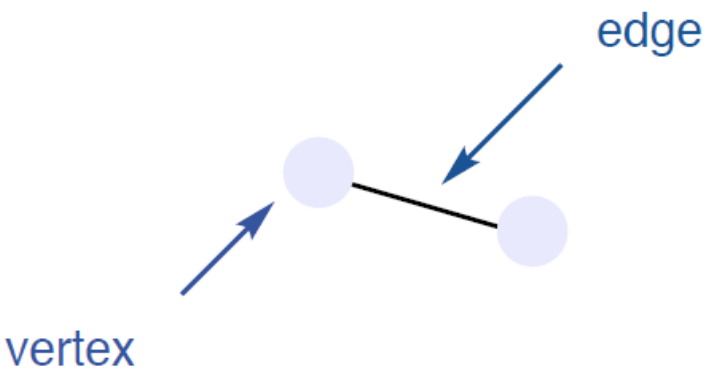
Cheaper

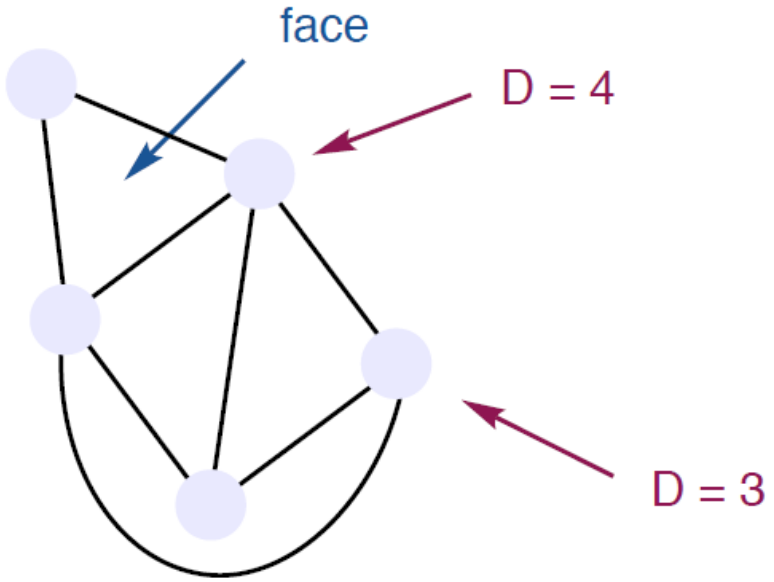
Faster

Scalable

What is a graph?

A graph is defined as a non empty set of vertices and a set of edges



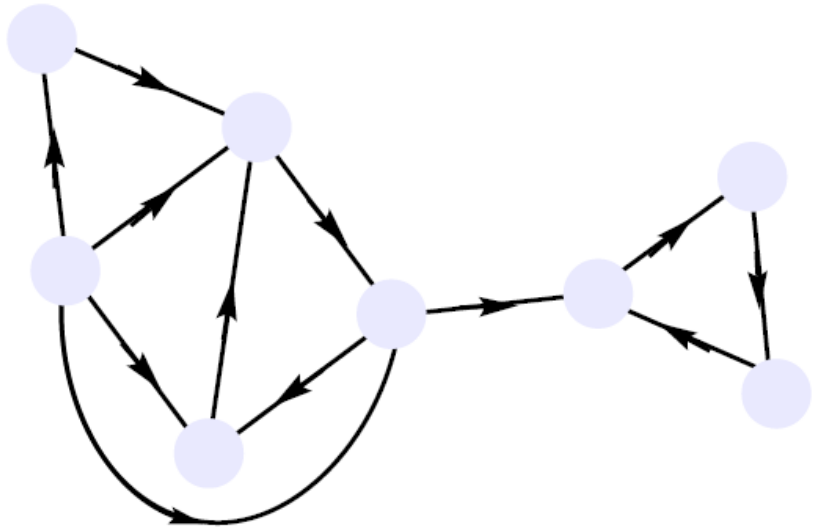


Order of the graph (N) = # of vertices

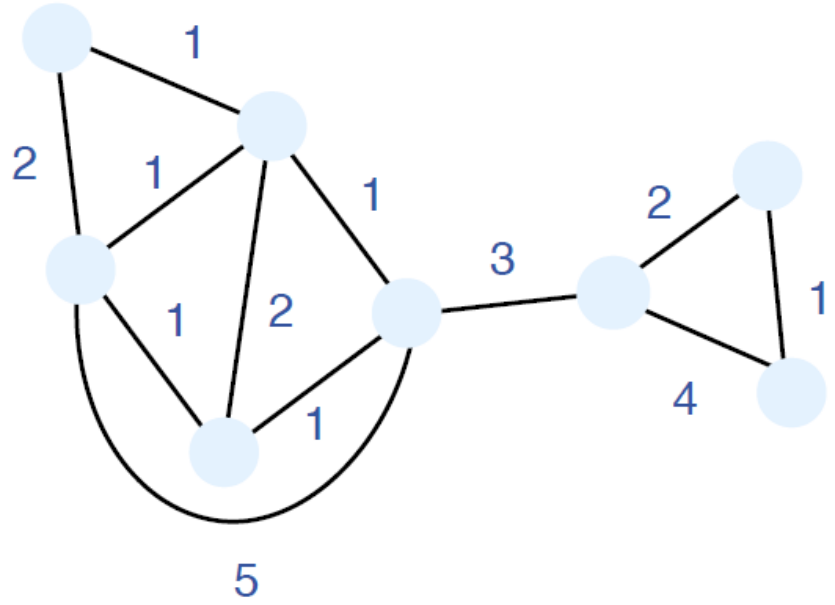
Size of the graph (M) = # of edges

Degree of a vertex (D) = # of edges incident with a vertex

Digraphs, line graphs, and weighted graphs



Digraph

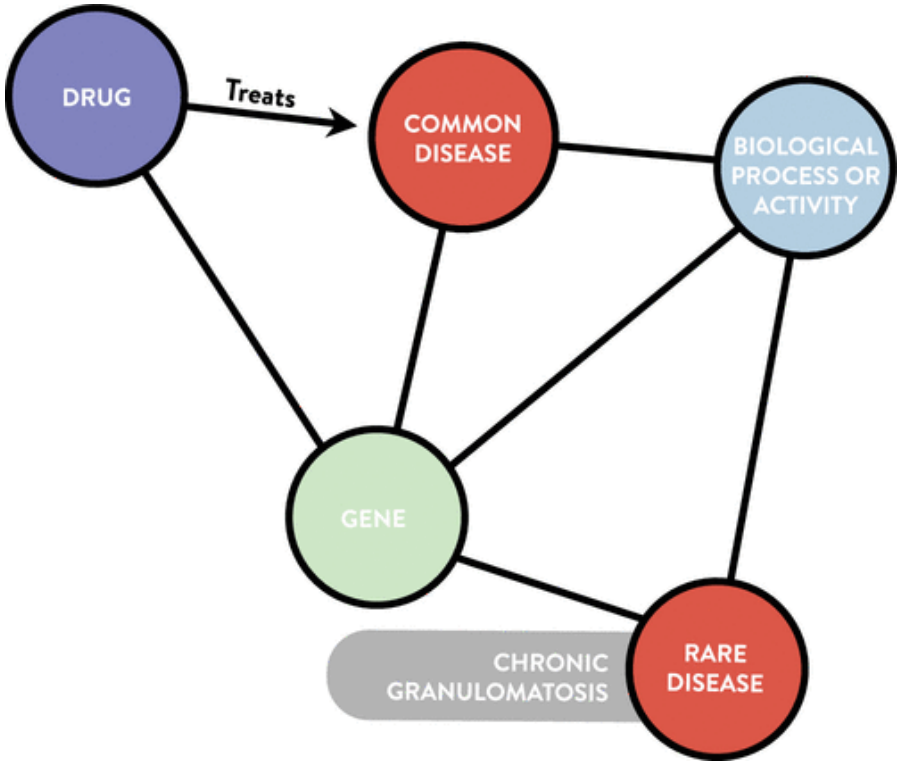


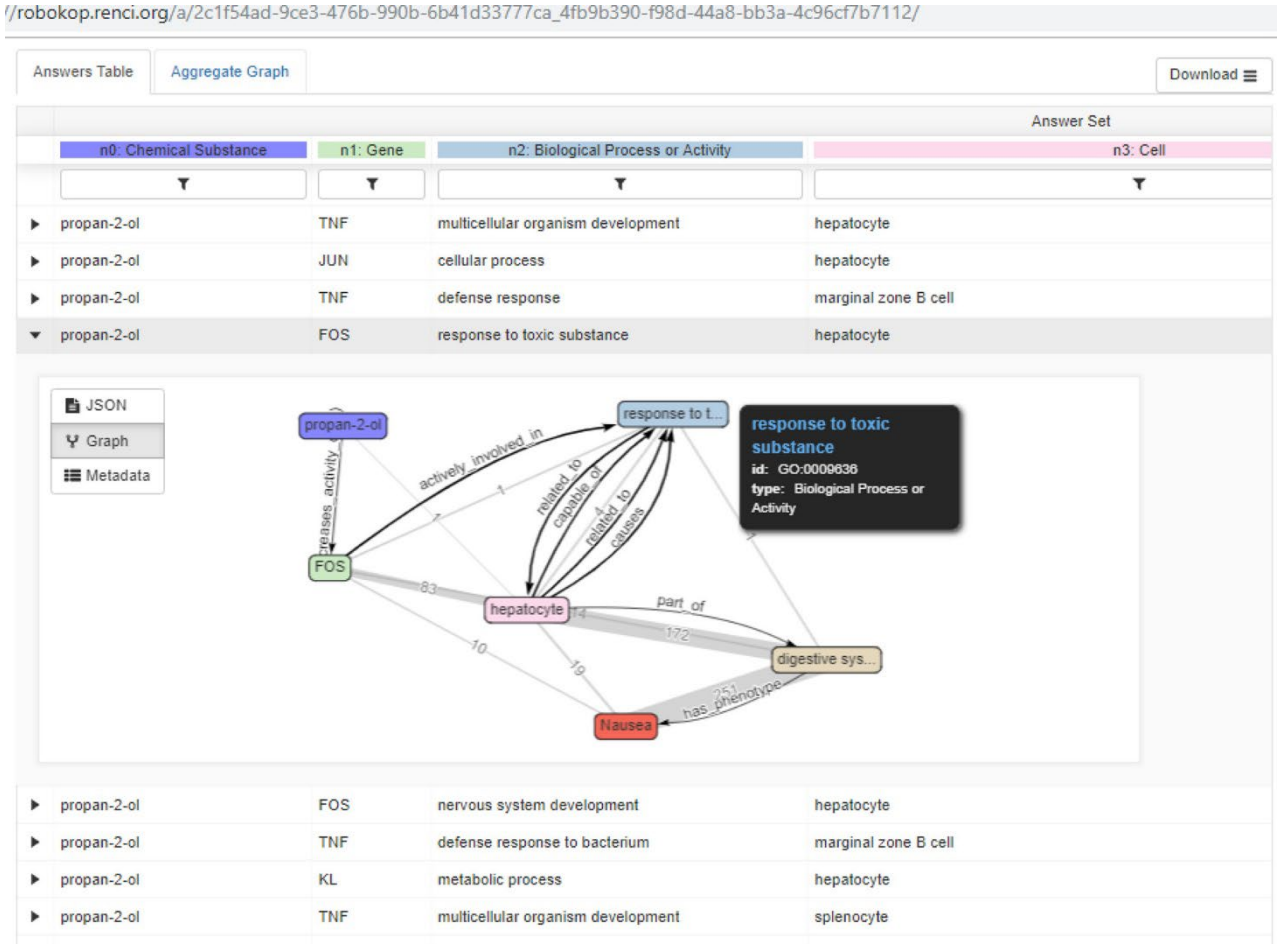
Weighted Graph

Knowledge Extraction from Unstructured Data

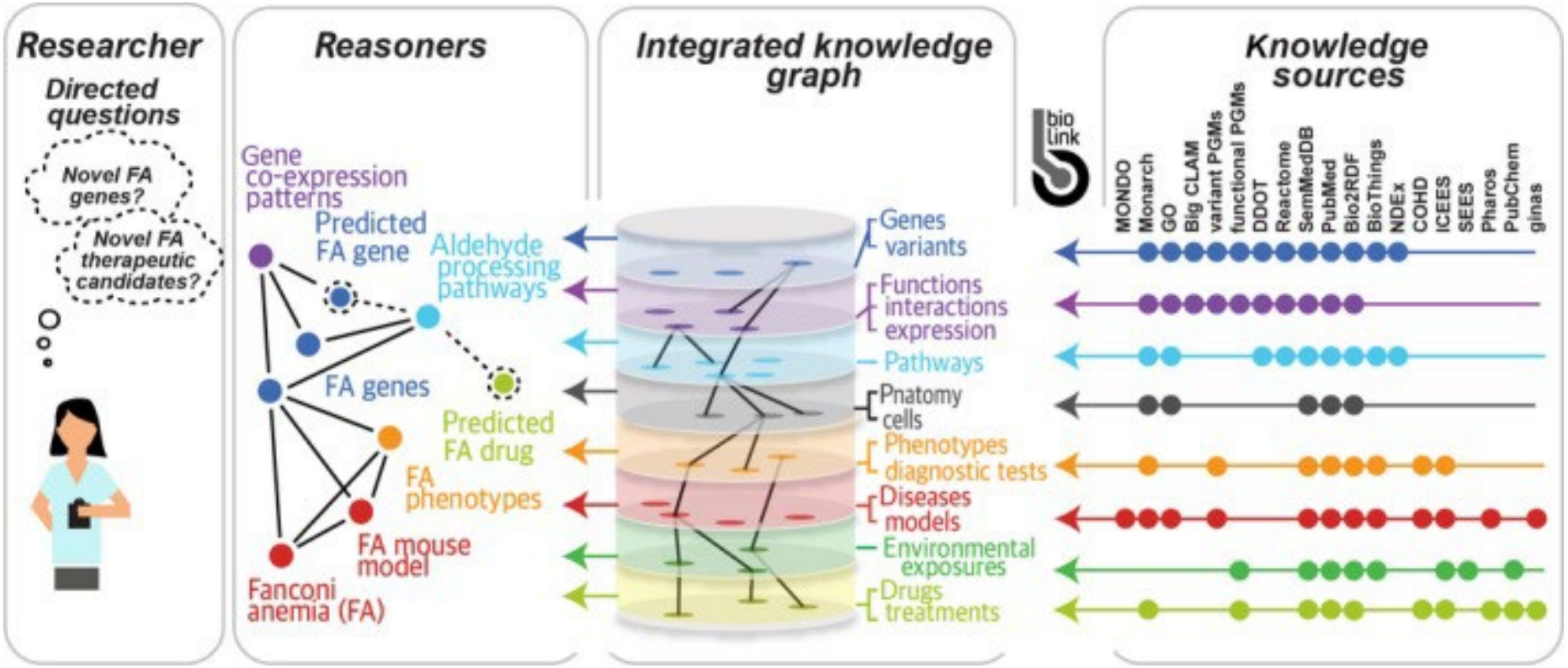
1. Step 1: Identify Entities of interest
2. Step 2: Identify relations that these entities participate in
3. Step 3(*): Identify events

Knowledge Graph Mining





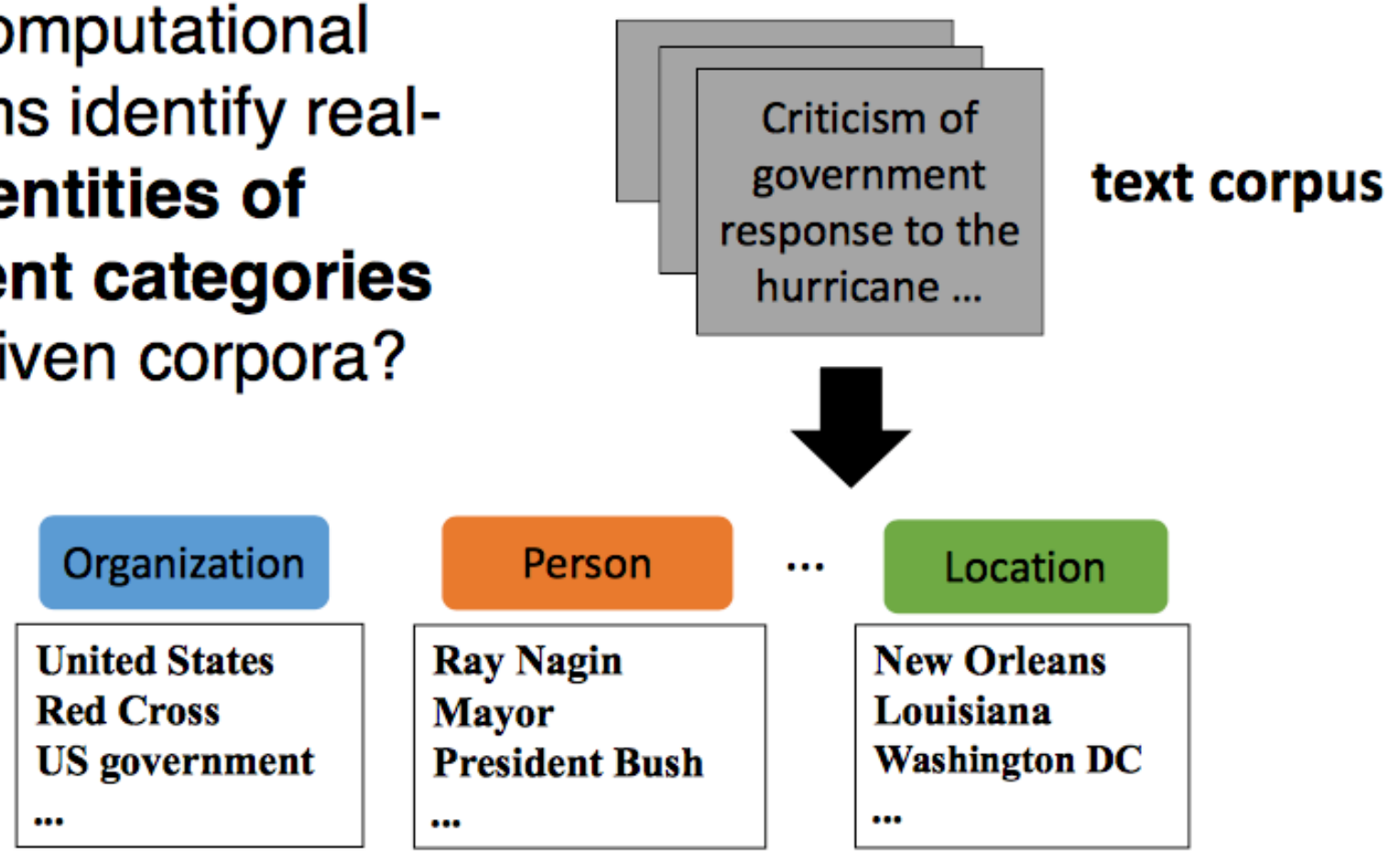
NIH Biomedical Data Translator



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6440568/

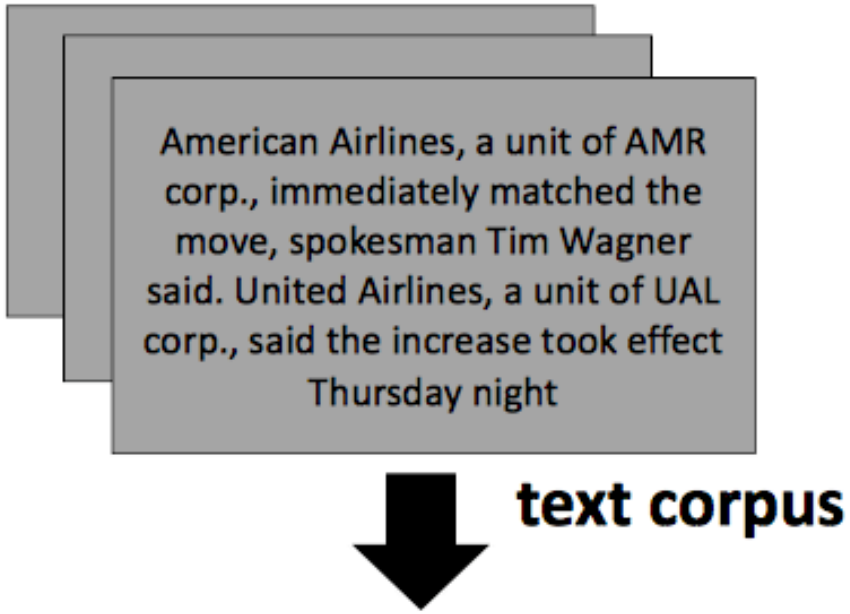
Entities

Can computational systems identify real-world **entities of different categories** from given corpora?



Relations

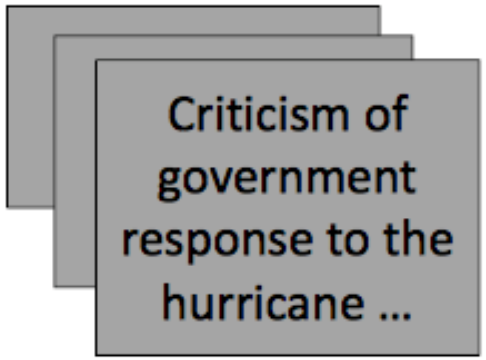
Can computational systems capture **different relations between the entities** from given corpora?



Entity 1	Relation	Entity 2
American Airlines	is_subsidiary_of	AMR
Tim Wagner	is_employee_of	American Airlines
United Airlines	is_subsidiary_of	UAL
...	...	...

Events

Can computational systems identify real-world **event of different types** from given corpora?



text corpus

Terrorism Template

LOCATION	TYPE	…	Date
CHILE: MOLINA (CITY)	*ROBBERY*		*07 JAN 90*

What is Information Extraction

As a task: Filling slots in a database from sub-segments of text.

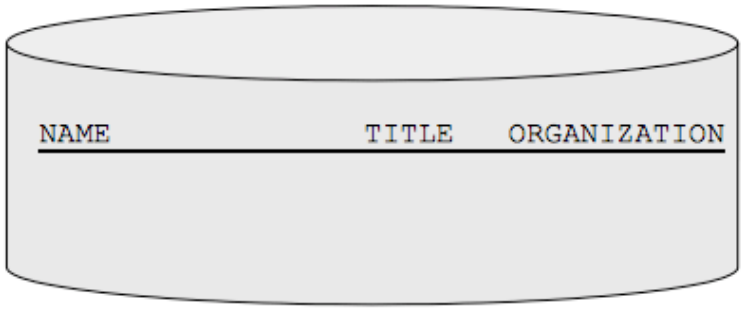
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



What is Information Extraction

As a task: Filling slots in a database from sub-segments of text.

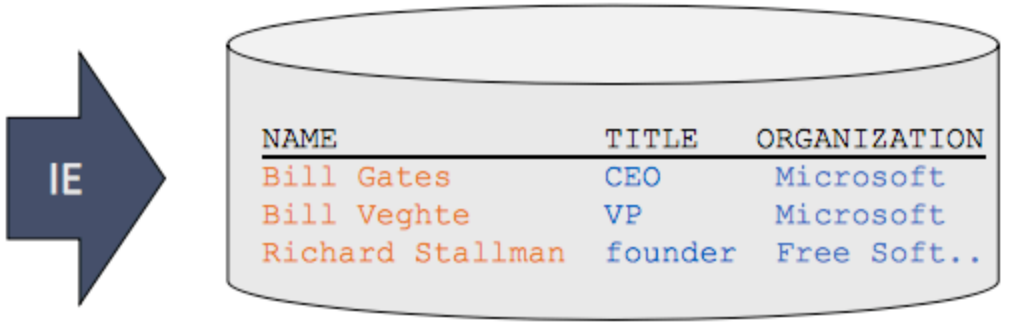
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



What is Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka "named entity extraction"

What is Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

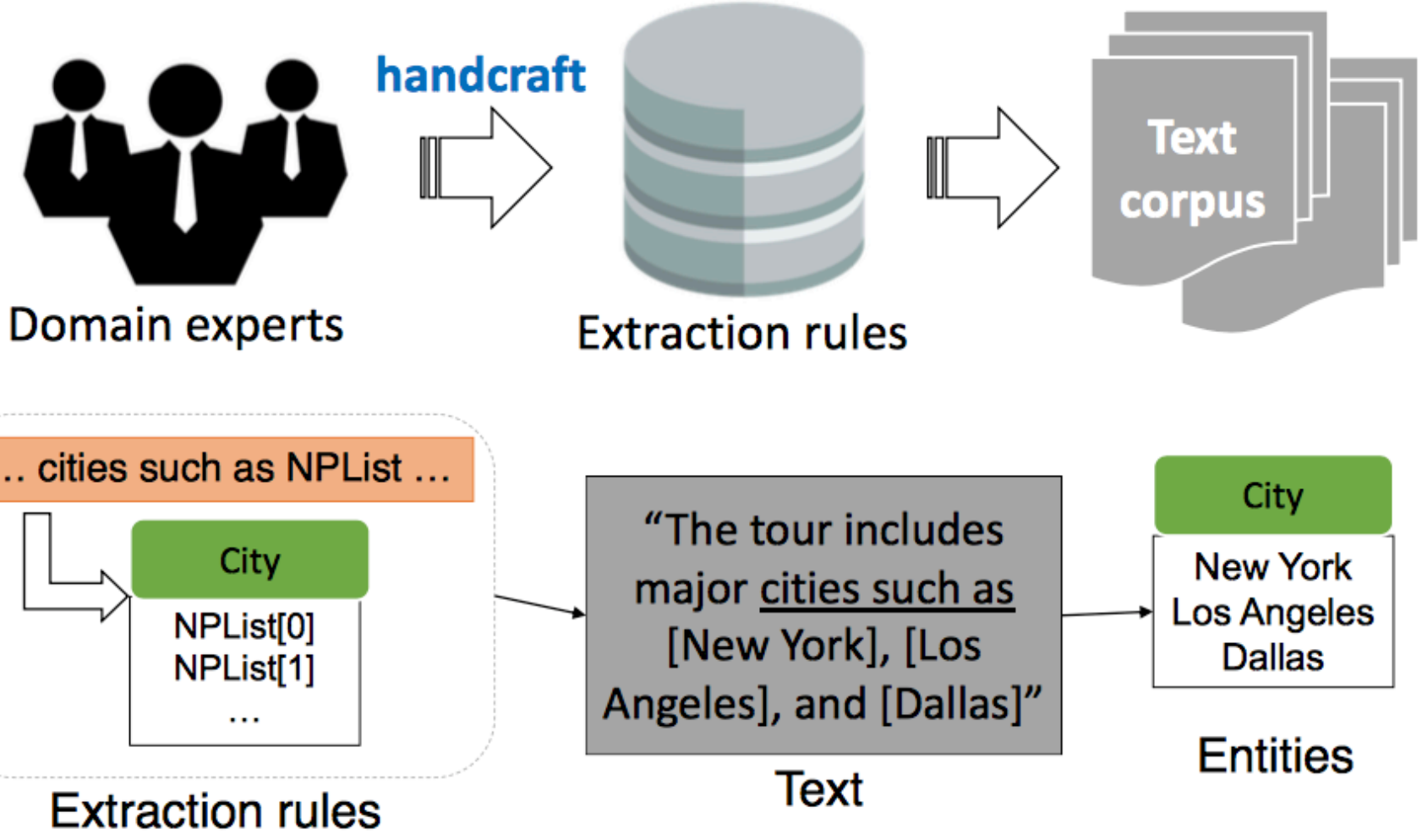
* Microsoft

* Bill Veghte
Microsoft
VP

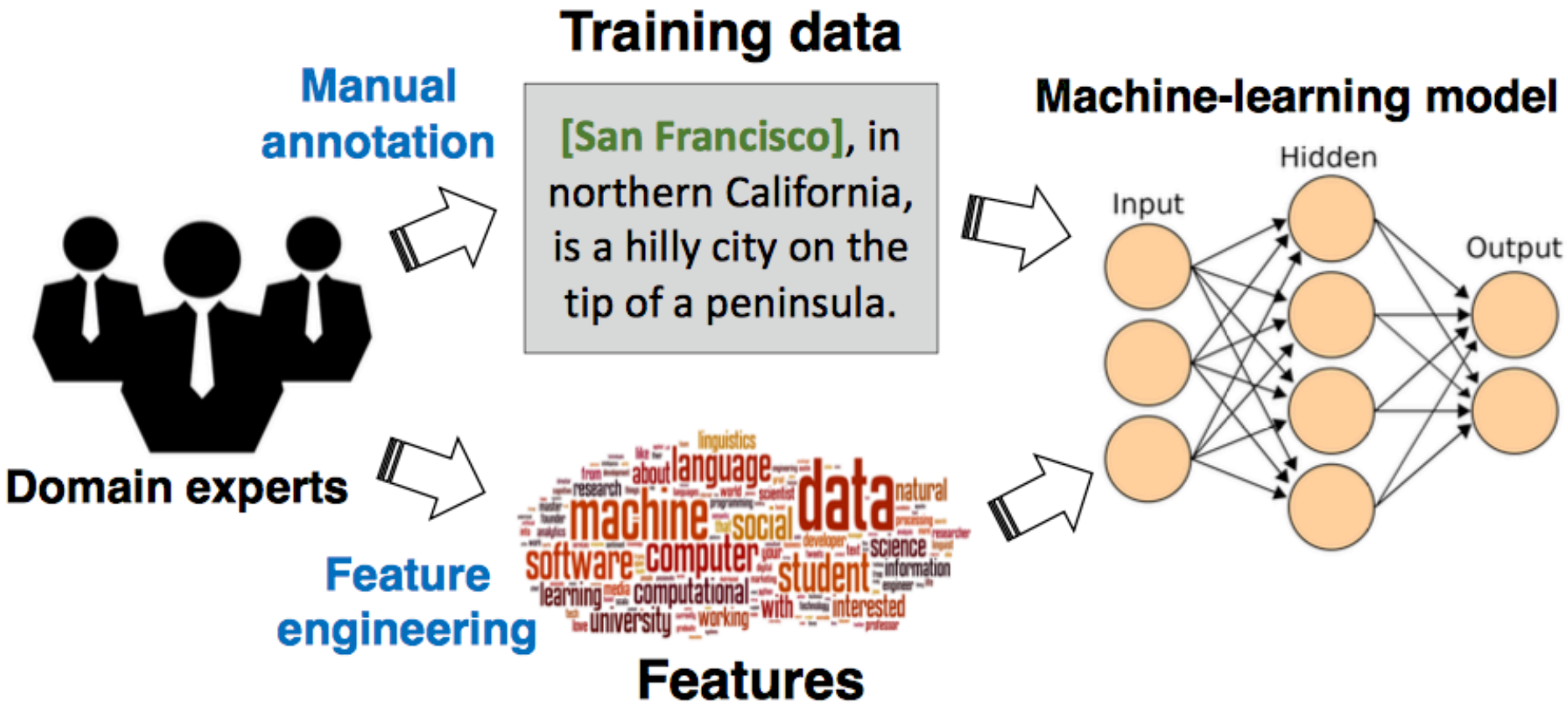
Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

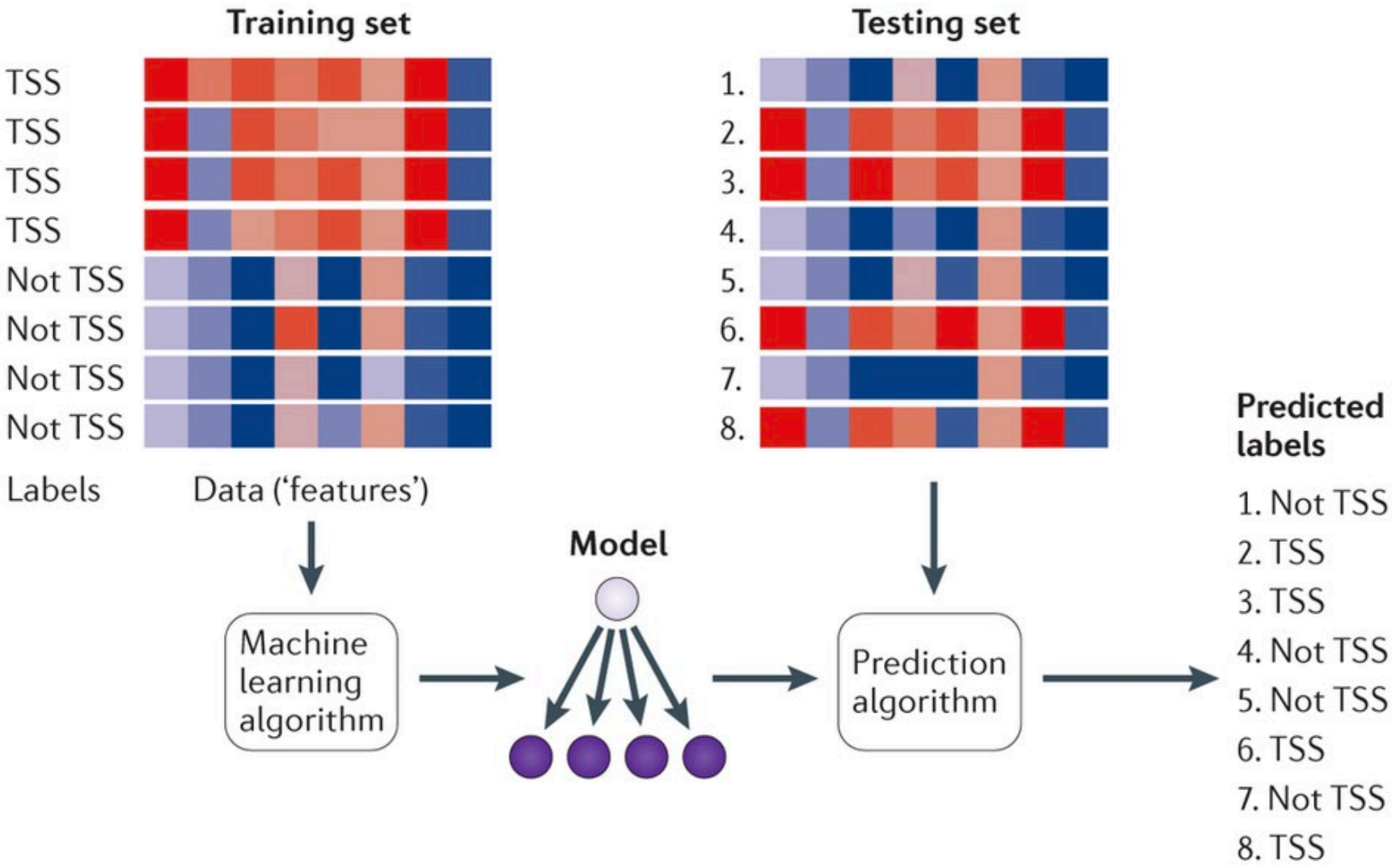
Traditional Rule-Based Systems



Supervised Machine Learning-Based Systems (state-of-the-art)



IE as Supervised Learning



IE as Supervised Learning

News articles

Analyze Input

"**President Barack Obama** and and his wife **Michelle**…"

Learning and Inference

Structured Output

Mention1	Mention2	IsSpouse
Michelle Obama	Barack Obama	T

Candidate Extraction

Sentences

id	content
	Michelle Obama married to President Barack Obama.

Michelle Obama is married to President Barack Obama.

StanfordCoreNLP

Mention	Type
Michelle Obama	PERSON
Barack Obama	PERSON
President	TITLE

User Defined Function

Mention1	Mention2	HasSpouse
Michelle Obama	Barack Obama	

Feature Extraction

Previously users would write features by hand

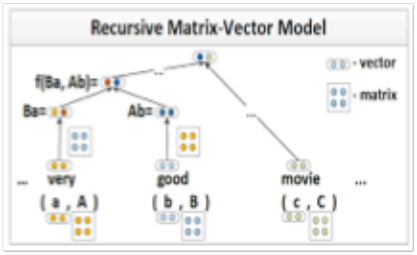
Michelle Obama **is married to** President Barack Obama.

- `Word_in_between["marry"]`
- `Distance<=5`

…

Now, most users rely on **automated** methods

Recursive Neural Networks (RNNs)



Treedlib (our library)

…However, these automated methods all rely on having a **large** (but noisy?) labeled training set!

Distant Supervision

Leverage existing knowledge bases, dictionaries to obtain training data via matching to the input corpus

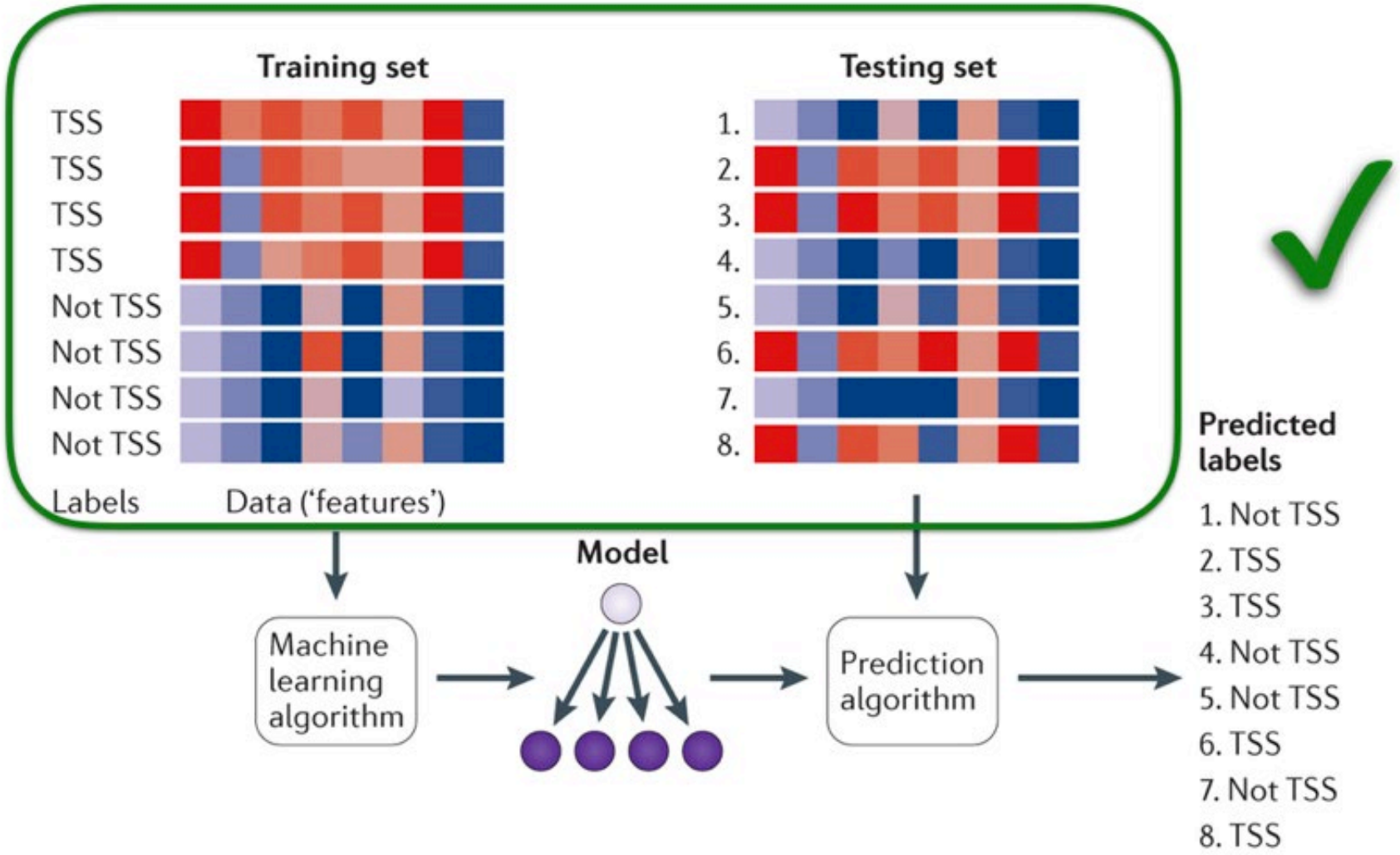
Michelle Obama is married to President Barack Obama.

Positive Example

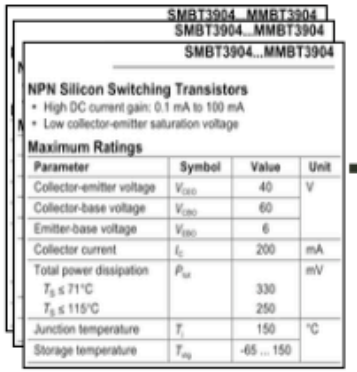
Spousal Relationship

Person 1	Person 2
Barack Obama	Michelle Obama
Nicolas Sarkozy	Carla Bruni
Hillary Clinton	Bill Clinton

IE as supervised learning



Fonduer: An example state-of-the-art system


SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mW
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Data Input

Preprocess data

Candidate Generation

Generate Training data

Featurization & Classification

HasCollectorCurrent

Transistor Part	Current
SMBT3904	200mA
MMBT3904	200mA

Output

Fonduer: An example state-of-the-art system

Richly formatted data

SMBT3904...MMBT3904

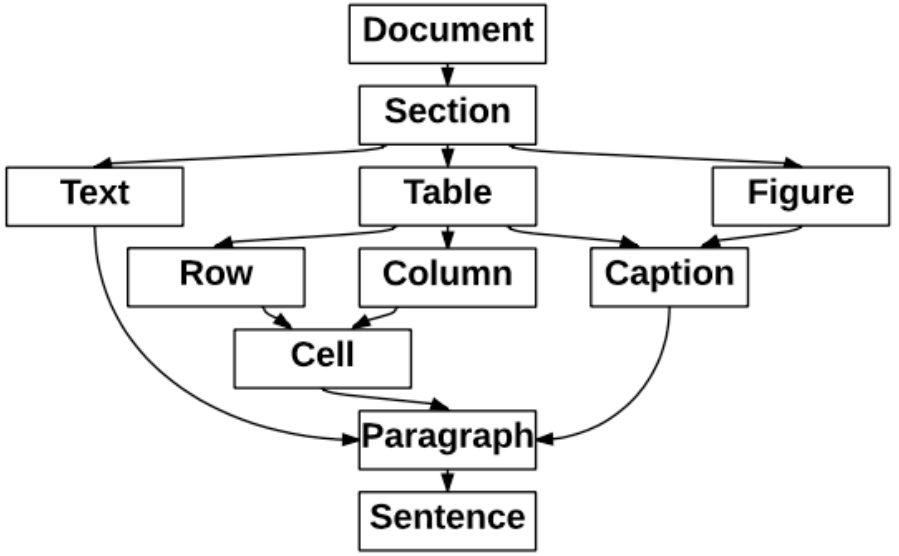
NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Data model



Fonduer automatically parses the richly formatted data into the data model that:

- ❑ Preserves structure/semantics across modalities
- ❑ Unifies a diverse variety of formats and styles
- ❑ Serves as the formal representation in KBC

Databases of Scientific Information

Just FYI

No need to memorize

Could be useful for your research project or classes

PubMed

- PubMed is the number one resource for anyone looking for literature in medicine or biological sciences.
- PubMed stores abstracts and bibliographic details of more than 30 million papers and provides full text links to the publisher sites or links to the free PDF on PubMed Central (PMC).

https://pubmed.ncbi.nlm.nih.gov/

PubChem: chemical information repository at the U.S. NIH

PubChem (https://pubchem.ncbi.nlm.nih.gov) is a public repository of information on small molecules and their biological activities, developed and maintained by the National Library of Medicine (NLM), an institute within the U.S. National Institutes of Health (NIH).

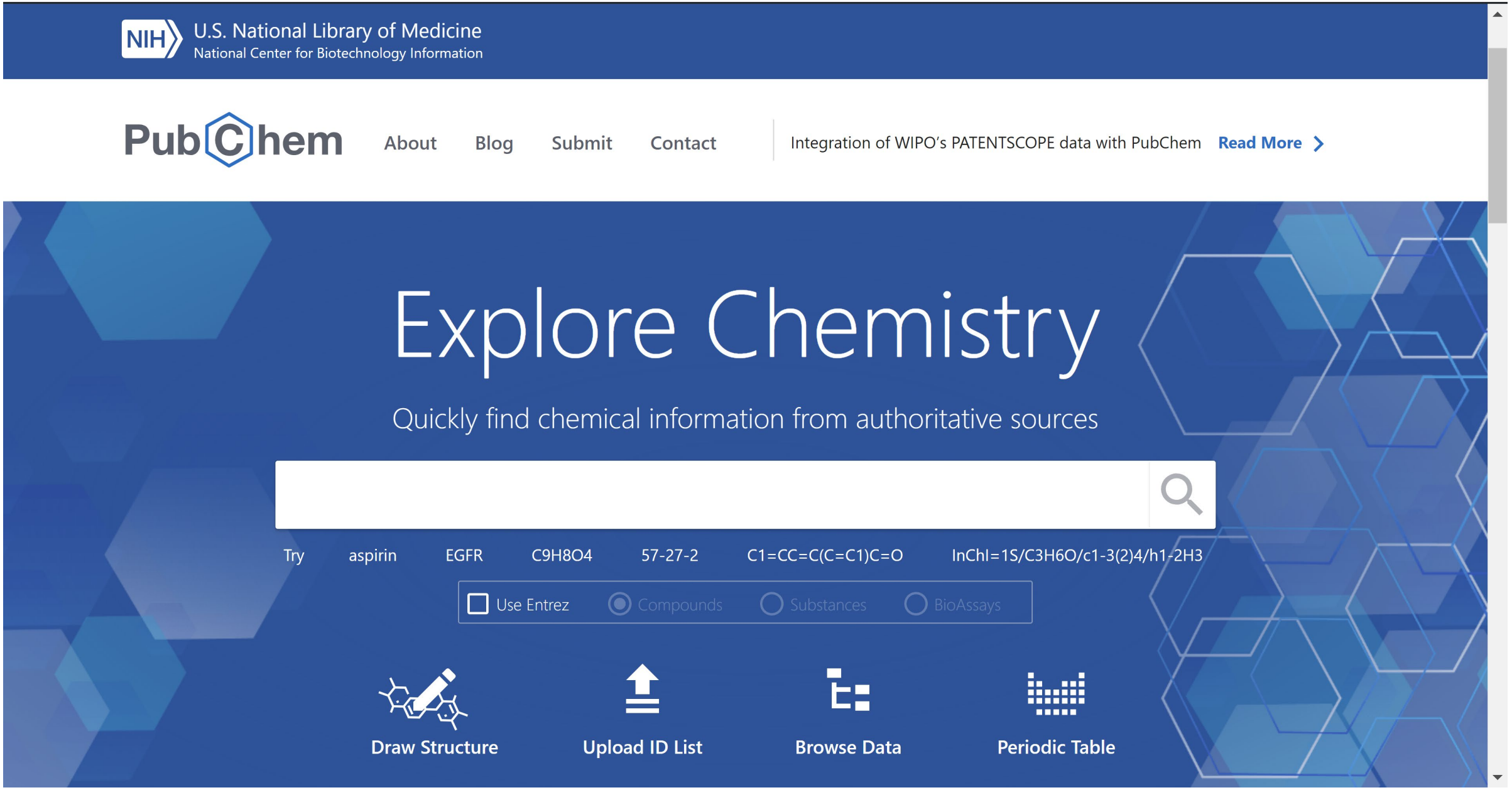
Since its launch in 2004 as a component of the NIH's Molecular Libraries Roadmap Initiatives, it has been rapidly growing, and now serves as a key chemical information resource for researchers in many biomedical science areas, including cheminformatics, chemical biology, and medicinal chemistry.

PubChem publications

PubChem Substance and Compound databases
S. Kim *et al.*, Nucleic Acids Research **2016**, *44*, D1202-D1213
(https://doi.org/10.1093/nar/gkv951)

PubChem BioAssay: 2017 update
Wang Y. *et al.* Nucleic Acids Research **2017**, *45*, D955-D963
(https://doi.org/10.1093/nar/gkw1118)

Getting the most out of PubChem for virtual screening
S. Kim, Expert Opin. Drug Discov. **2016**, *11*, 843-855
(http://dx.doi.org/10.1080/17460441.2016.1216967)

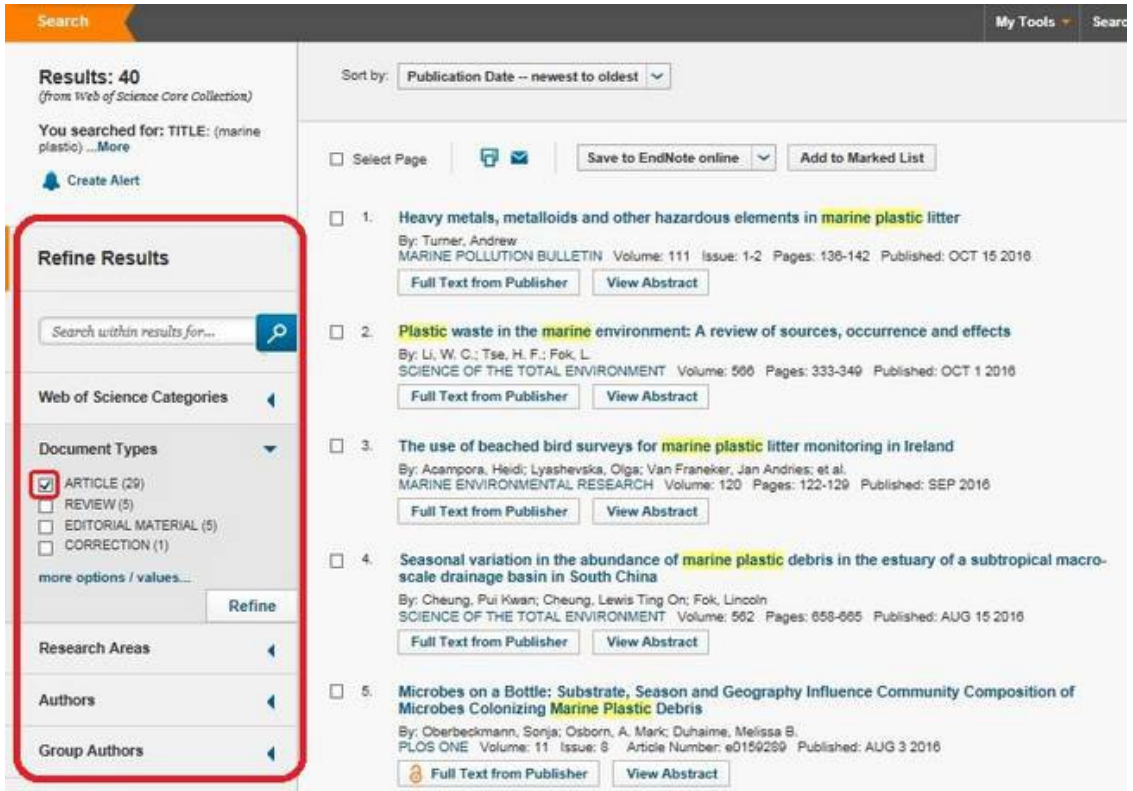

U.S. National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Blog
Submit
Contact
Integration of WIPO's PATENTSCOPE data with PubChem
Read More
Explore Chemistry
Quickly find chemical information from authoritative sources
Try
aspirin
EGFR
C9H8O4
57-27-2
C1=CC=C(C=C1)C=O
InChI=1S/C3H6O/c1-3(2)4/h1-2H3
Use Entrez
Compounds
Substances
BioAssays
Draw Structure
Upload ID List
Browse Data
Periodic Table

Web of Science

Web of Science also known as Web of Knowledge is the second big bibliographic database. Usually, academic institutions provide either access to Web of Science or Scopus on their campus network for free.

- Coverage: approx. 100 million items
- References: 1.4 billion
- Discipline: Multidisciplinary
- Access options: institutional subscription only

https://apps.webofknowledge.com

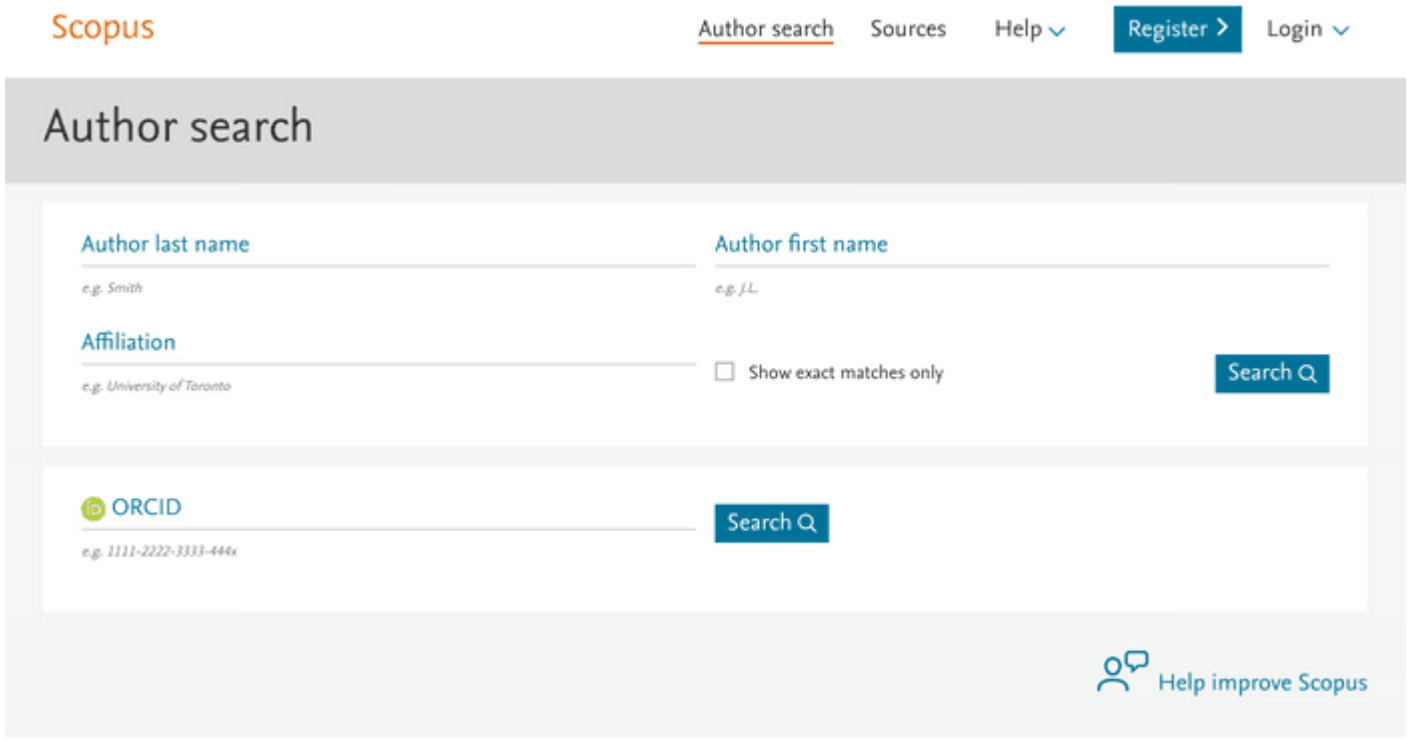


Scopus

Scopus is one of the two big commercial, bibliographic databases that cover scholarly literature from almost any discipline. Beside searching for research articles, Scopus also provides academic journal rankings, author profiles, and an h-index calculator.

- Coverage: approx. 71 million items
- References: 1.4 billion
- Discipline: Multidisciplinary
- Access options: Limited free preview, full access by institutional subscription only
- Provider: Elsevier

https://www.scopus.com/



ChEMBL: literature-extracted biological activity information

ChEMBL (https://www.ebi.ac.uk/chembl/) is a large bioactivity database, developed and maintained by the European Bioinformatics Institute (EBI), which is part of the European Molecular Biology Laboratory (EMBL).

The core activity data in ChEMBL are “manually” extracted from the full text of peer-reviewed scientific publications in select chemistry journals, such as *Journal of Medicinal Chemistry*, *Bioorganic Medicinal Chemistry Letters*, and *Journal of Natural products*.

From each publication, details of the compounds tested, the assays performed and any target information for these assays are abstracted.

ChEMBL publication

Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. *Nucleic Acids Res.* **2017**, *45*, D945.

ChEMBL

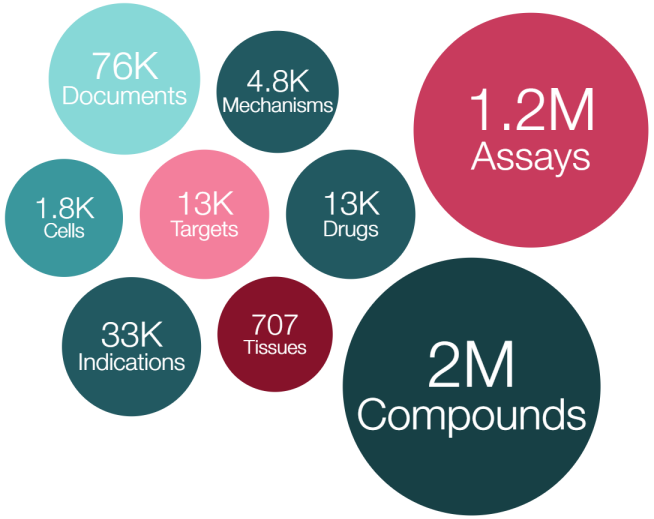
Search in ChEMBL

Examples: Imatinib erbB2 brain MDCK c1ccccc1N

Draw a Structure | Enter a Sequence

As a result of planned maintenance that is due to end on 4th April 2020, there may be unexpected disruption caused to some of the services we host. If you experience any issues, please contact the relevant help desk directly for support.

ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.



Explore ChEMBL

Description: Shows a summary of the ChEMBL entities and quantities of data for each of them.

Instructions: Click on a bubble to explore a specific ChEMBL entity in more detail.

NIST Webbook: thermodynamic and spectroscopic data of chemicals

The U.S. National Institutes of Standards and Technology (NIST) compiles chemical and physical property data for chemical species and distributes them through the web site called the NIST Chemistry WebBook (http://webbook.nist.gov)

These data include thermochemical data (e.g., enthalpy of formation, heat capacity, and vapor pressure), reaction thermochemistry data (e.g., enthalpy of reaction and free energy of reaction), spectroscopic data (e.g., IR and UV/Vis spectra), gas chromatographic data, ion energetics data, and so on.

Linstrom, P. J.; Mallard, W. G. *J. Chem. Eng. Data* **2001**, *46*, 1059.

NIST Chemistry WebBook

NIST Standard Reference Database Number 69

Last update to data: 2018 DOI: https://doi.org/10.18434/T4D303

View: Search Options, Models and Tools, Special Data Collections, Documentation, Changes, Notes

▷Credits

NIST reserves the right to charge for access to this database in the future.

NIST recently released a new version of the NIST Inorganic Crystal Structure Database (ICSD). For more information visit the ICSD web site.

Search Options

General Searches

- Formula
- Name
- IUPAC identifier
- CAS registry number
- Reaction
- Author
- Structure

Physical Property Based Searches

- Ion energetics properties
- Vibrational and electronic energies
- Molecular weight

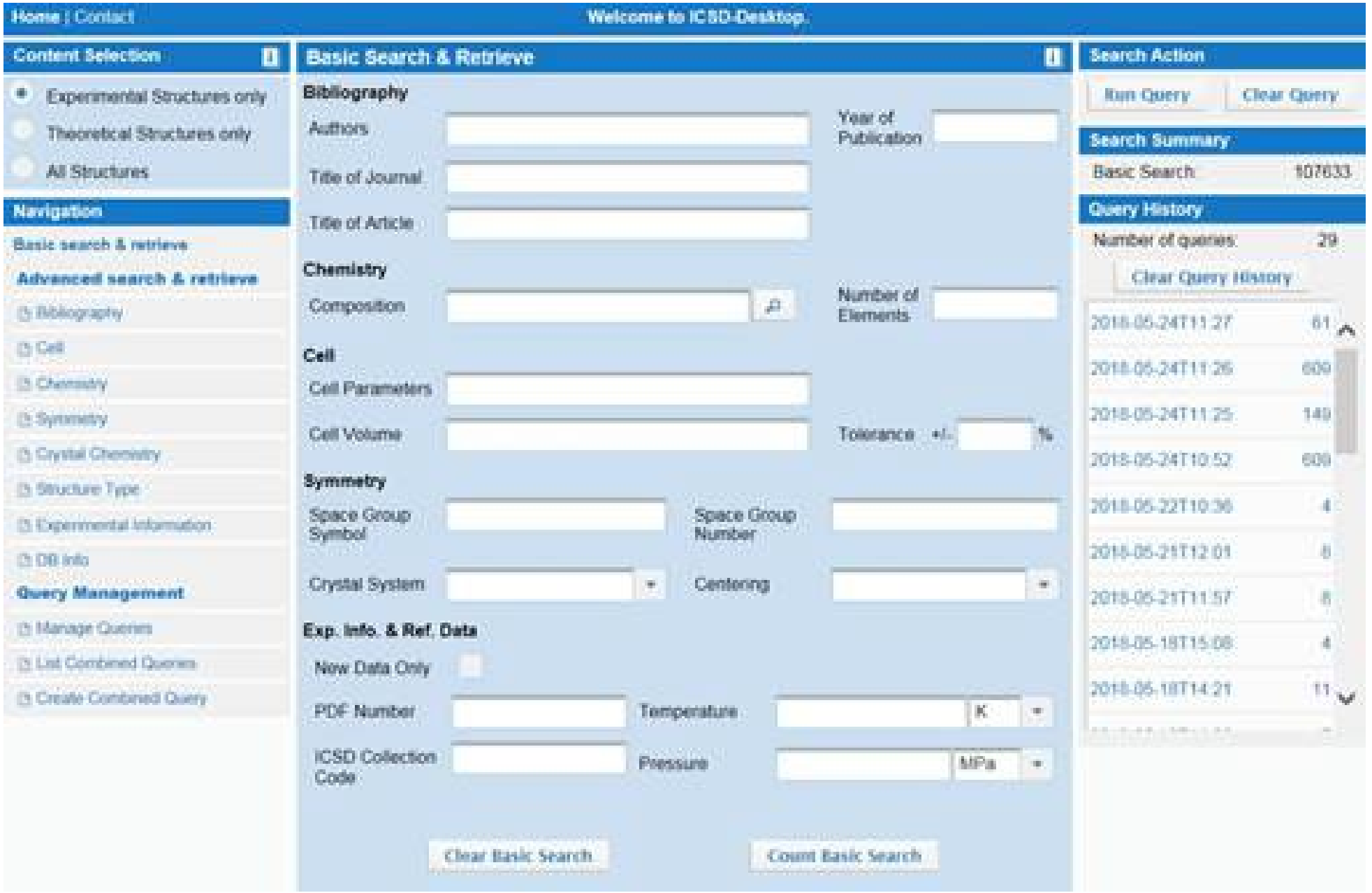
INORGANIC CRYSTAL STRUCTURE DATABASE (ICSD)

The Inorganic Crystal Structure Database (ICSD) is the world's largest database of fully determined inorganic crystal structures, from elements to quintenary compounds. It contains about 185,000 structures with 6,000 added annually.

Each record includes crystallographic data as well as chemical/physical property data and bibliographic information for the journal article referencing the structure.

ICSD is commercial by subscription ($$$$)

https://icsd.products.fiz-karlsruhe.de/ or CD (!)


Home | Contact
Welcome to ICSD-Desktop.
Content Selection
Experimental Structures only
Theoretical Structures only
All Structures
Navigation
Basic search & retrieve
Advanced search & retrieve
Bibliography
Cell
Chemistry
Symmetry
Crystal Chemistry
Structure Type
Experimental Information
DB Info
Query Management
Manage Queries
List Combined Queries
Create Combined Query
Basic Search & Retrieve
Bibliography
Authors
Year of Publication
Title of Journal
Title of Article
Chemistry
Composition
Number of Elements
Cell
Cell Parameters
Cell Volume
Tolerance +/-
%
Symmetry
Space Group Symbol
Space Group Number
Crystal System
Centering
Exp. Info. & Ref. Data
Now Data Only
PDF Number
Temperature
K
ICSD Collection Code
Pressure
MPa
Clear Basic Search
Count Basic Search
Search Action
Run Query
Clear Query
Search Summary
Basic Search 107633
Query History
Number of queries: 29
Clear Query History
2018-05-24T11:27 61
2018-05-24T11:26 609
2018-05-24T11:25 149
2018-05-24T10:52 609
2018-05-22T10:36 4
2018-05-21T12:01 8
2018-05-21T11:57 8
2018-05-18T15:08 4
2018-05-18T14:21 11

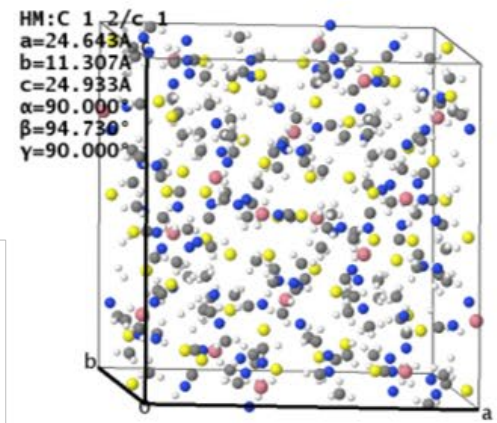
Summary

Collection Code 250888

Struct.formula	(N (C H3)4)2 (Co (N C S)4)
Space Group	C 1 2/c 1 (15)
Unit Cell	24.6433(2) 11.30658(11) 24.9333(3) 90 94.7301(9) 90
Cell Volume	6923.55 A³
Formula Units per Cell	12
Temperature	295 K
Pressure	atmospheric
PDF-Numbers	
R-Value	0.0521
Remark	
	High Quality Data

Author	Shurdha, Endrit; Moore, Curtis E.; Rheingold, Arnold L.; Lapidus, Saul H.; Stephens, Peter W.; Arif, Atta M.; Miller, Joel S.
Title of Article	First row transition metal(II) thiocyanate complexes, and formation of 1-, 2-, and 3-dimensional extended network structures of M(NCS)2(solvent)2 (M = Cr, Mn, Co) composition
Reference	Inorganic Chemistry (2013) 52, (18) p10583-p10594
Warnings & Comments	1 Warnings / 1 Comments

Published Crystal Structure



Interactive Visualization

The Cambridge Structural Database (CSD)

https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/

Established in 1965, the CSD is the world's repository for small-molecule organic and metal-organic crystal structures. Containing over one million structures from x-ray and neutron diffraction analyses, this unique database of accurate 3D structures has become an essential resource to scientists around the world.

1M+ crystal structures!

Web access by the CSD code and Python API

CCDC FIZ Karlsruhe Leibniz Institute for Information Infrastructure

Access Structures

Sign In

Simple Search | Structure Search | Unit Cell Search | Formula Search

Entry search

Welcome to Access Structures, the CCDC's and FIZ Karlsruhe's free service to view and retrieve structures. Please use one or more of the boxes to find entries. If you enter details in more than one field the search will try to find records containing all the terms entered. More information and search help

More advanced search functionality and additional curated data for the Cambridge Structural Database (CSD) and the Inorganic Crystal Structure Database (ICSD) is available through the CSD-System and ICSD, respectively. Click here for more information.

Identifier(s) CCDC Number(s), CSD Number(s), CSD Refcode(s) or ICSD Number(s)

Compound name e.g. sulfadiazine

DOI A single publication DOI, CSD DOI or ICSD DOI

Authors e.g. F.H.Allen

Journal e.g. Journal of the American Chemical Society

Publication details Year | Volume | Page

Database to search (•) Entire published collection ( ) CSD ( ) ICSD ( ) Teaching subset

Search Clear

Advanced Search

Zillions of other databases!

Let me know if I missed something useful for your research!

Protein Data Bank - https://www.rcsb.org/

Molecular Spectroscopic Data - https://www.nist.gov/pml/molecular-spectroscopic-data