WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

CS639:
Data Management for Data Science

Lecture 21: Information Extraction

Theodoros Rekatsinas

So far…

1. Manage data of various forms (structured, key-values, documents)
    1. RDBMS
    2. MadReduce
    3. Key-value Stores

2. How to learn models that capture the distribution of observed data
    1. Statistics and Statistical Inference
    2. Linear Classifiers
    3. Decision Trees
    4. Unsupervised/Supervised learning
    5. Optimization

Until the end of the semester…

1. Information extraction and Data Integration

2. Communicating insights
    1. Visualizations and Privacy

Information Extraction

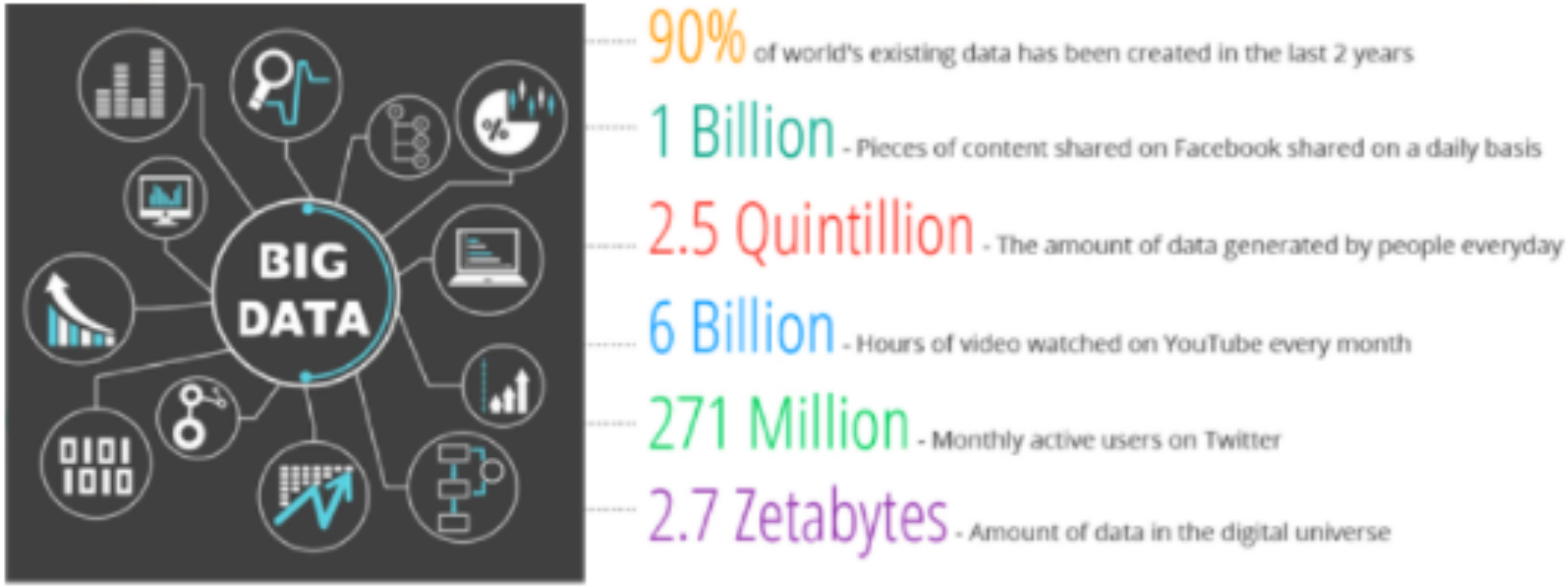
1. Extracting knowledge from unstructured data (e.g., text)

2. Recognize Named Entities in unstructured data

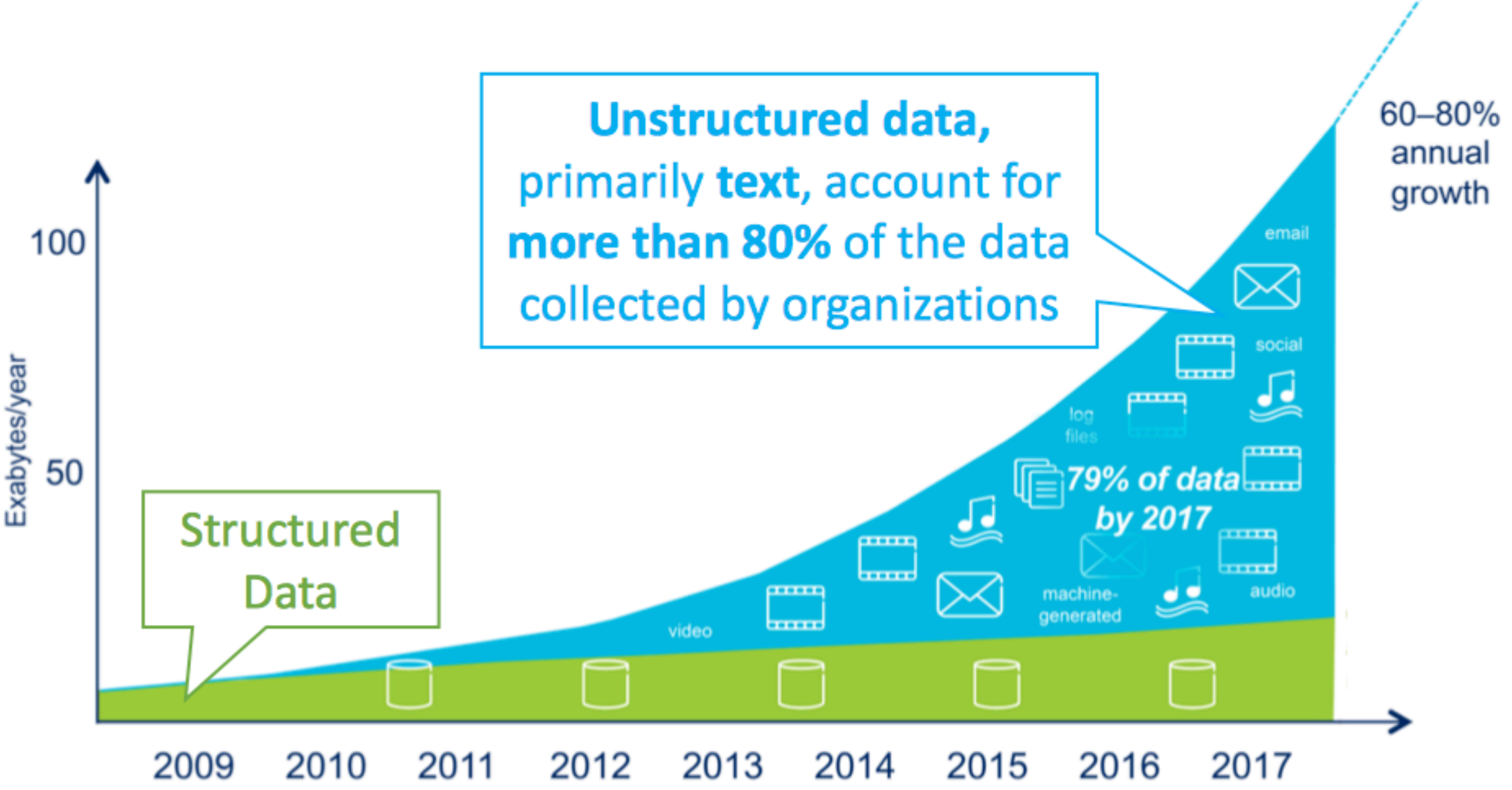
3. Clean and normalize extractions

What is Information Extraction?

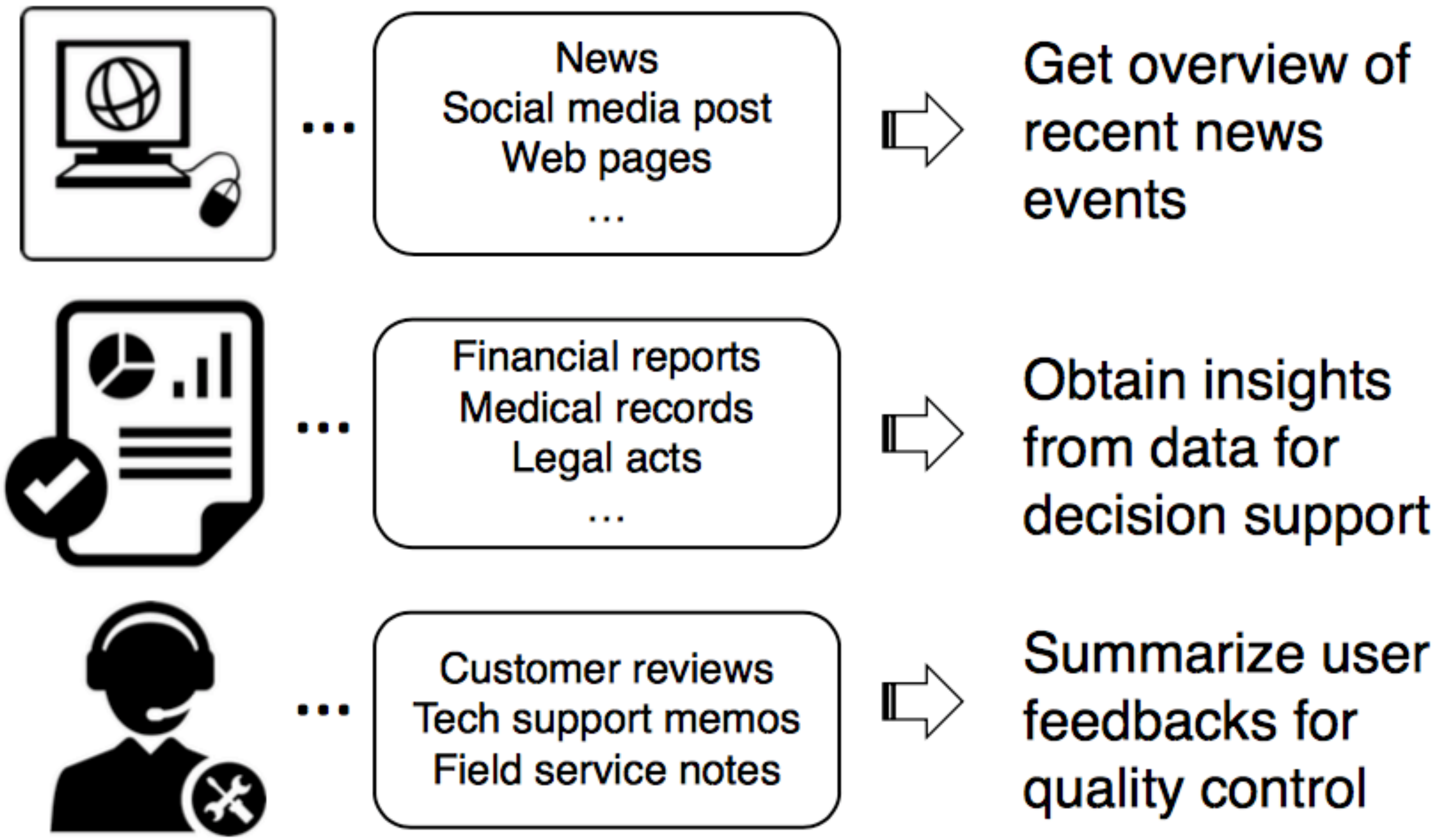
Goal: Mine knowledge from unstructured data



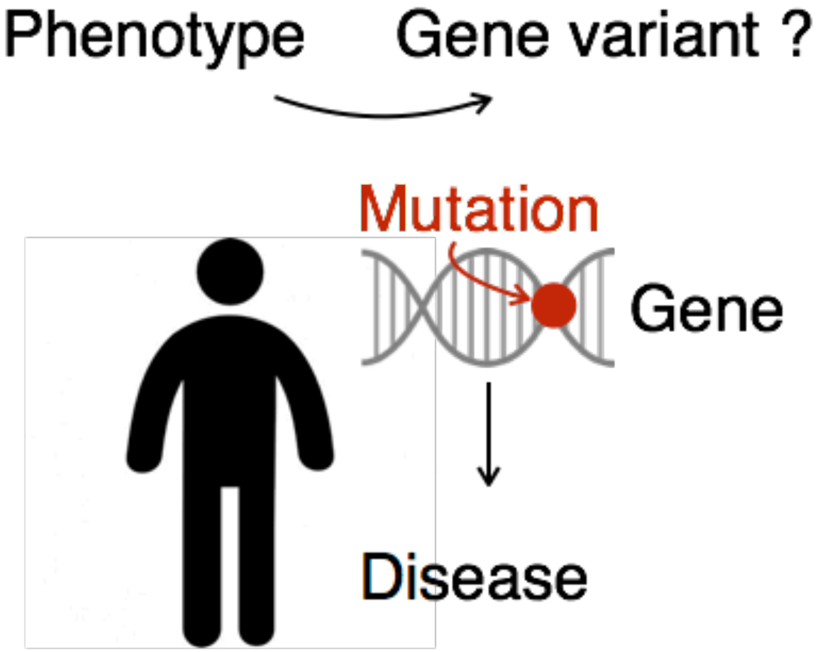
Growth of Unstructured Text Data



Knowledge in unstructured data

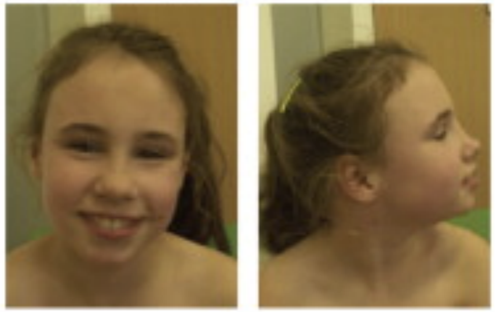


Knowledge from Unstructured Data (Example)



STANFORD
HOSPITAL & CLINICS
Stanford University Medical Center

Personalized medicine

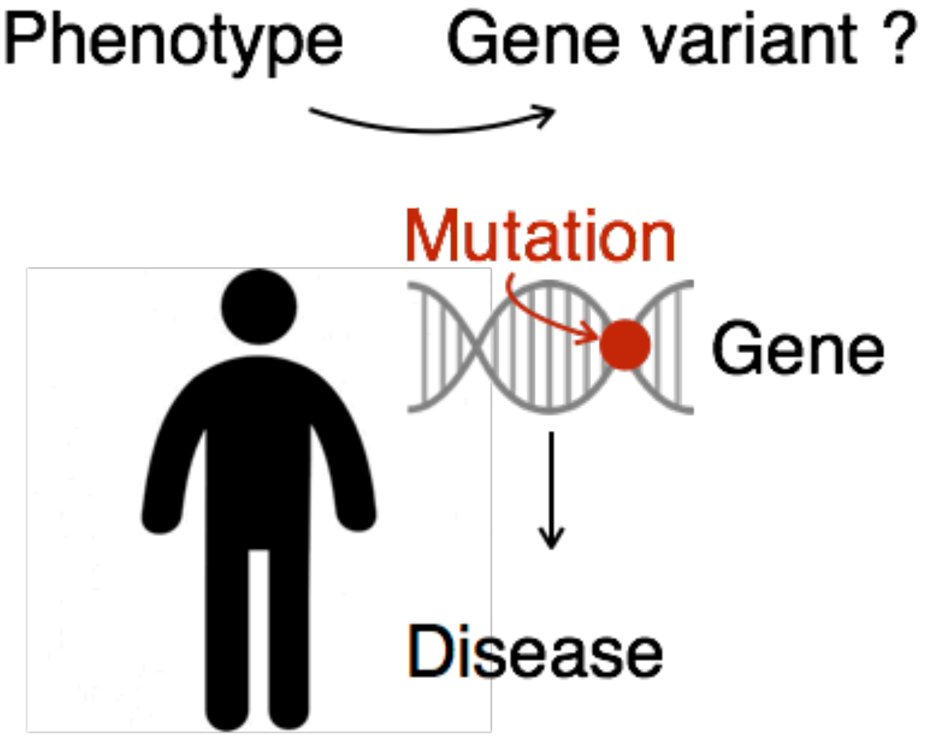


intellectual disability with impaired speech development and aggressive behavior

83 candidate genes in her exome with rare variants

AC018470.1, ACAP3, ADAP1, AMPD1, ASPM, ASXL2, BAZ1B, BHLHE22, BTBD9, C17orf104, C17orf74, C19orf26, C1orf87, C2orf81, CCNL2, CDH10, CHD6, CNOT3, COL6A5, DCHS2, DEAF1, DNM1, FAM216B, FAM73B, FAM83H, FAM84B, FAT3, FBXO25, FCRLB, FLJ00104, FRS2, GRK7, HEPHL1, HOXD11, IL19, INSRR, IQCC, KIAA0825, LAMA5, LAMC3, LGR6, MAST4, MBD6, MBLAC2, MCM10, MDH2, METRN, MSL2, N4BP3, NCKAP5, NUP50, NYNRIN, ORC3, PDCD2L, PDXP, PLEKHG1, PLIN2, POU3F2, PXMP2, RAB11FIP1, RASSF1, RIMS1, RTKN2, SASS6, SERPINA3, SH3BP1, SHB, SLC2A9, SLC38A8, SON, SP8, SPTBN5, SRRM2, TAAR1, TARSL2, TET2, TRIM72, TSPAN15, TSPYL4, WDR20, XPNPEP1, ZFYVE16, ZNF469, ZSCAN29

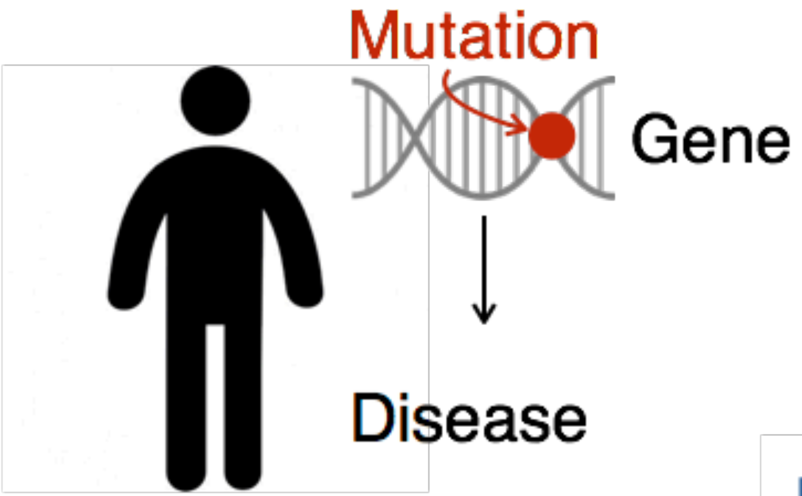
Personalized medicine



Which gene is at fault?

Personalized medicine

Phenotype Gene variant ?



Find right article (1hr/variant)

Which gene is at fault?

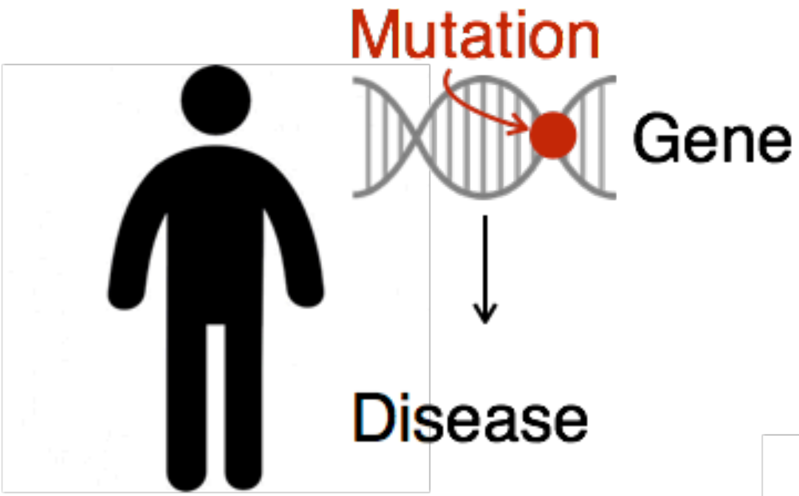
PubMed

National Library of Medicine NLM

25 million articles

Personalized medicine

Phenotype Gene variant ?



Find right article (1hr/variant)

Which gene is at fault?

PubMed

National Library of Medicine NLM

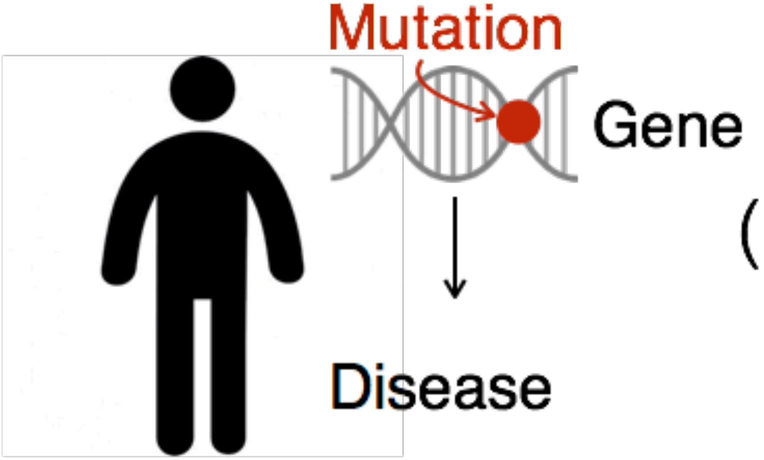
25 million articles

Doctor Google

Can we build a machine to read these articles?

Personalized medicine

Phenotype → Gene variant ?



Query KB
(Instantaneous)

Cheaper

Faster

Scalable

Knowledge Base Construction (KBC)

DeepDive

Which gene is at fault?

Knowledge Base

Gene	Phenotype
DEAF1	Intellectual Disability

Knowledge Extraction from Unstructured Data

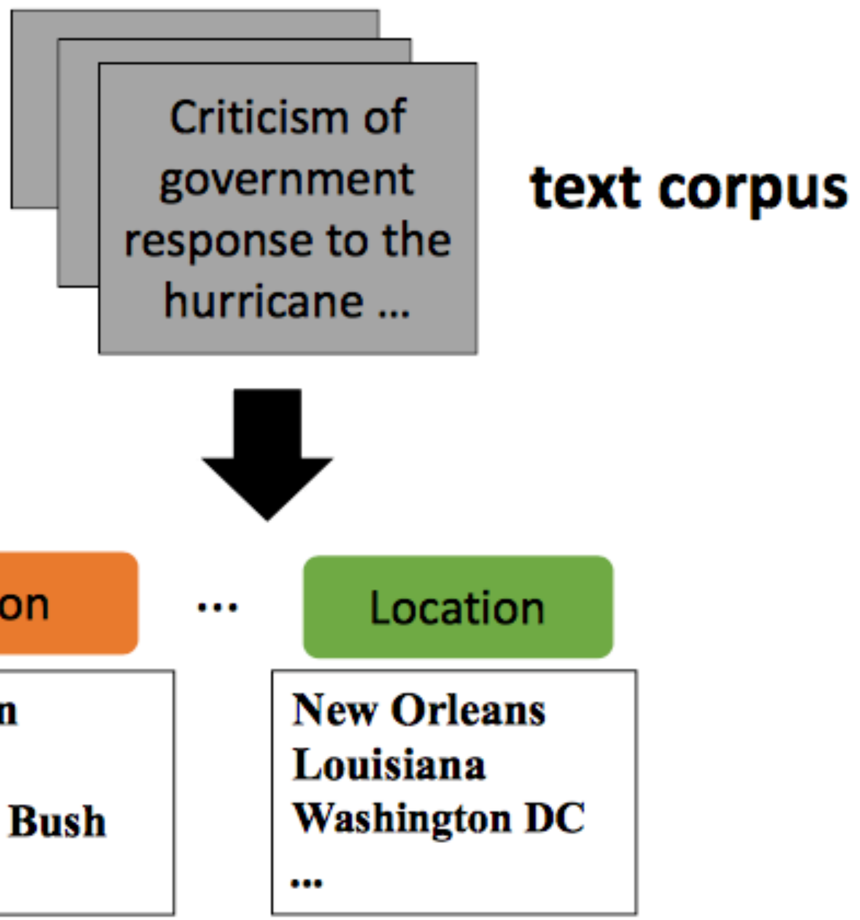
1. Step 1: Identify Entities of interest

2. Step 2: Identify relations that these entities participate in

3. Step 3(*): Identify events

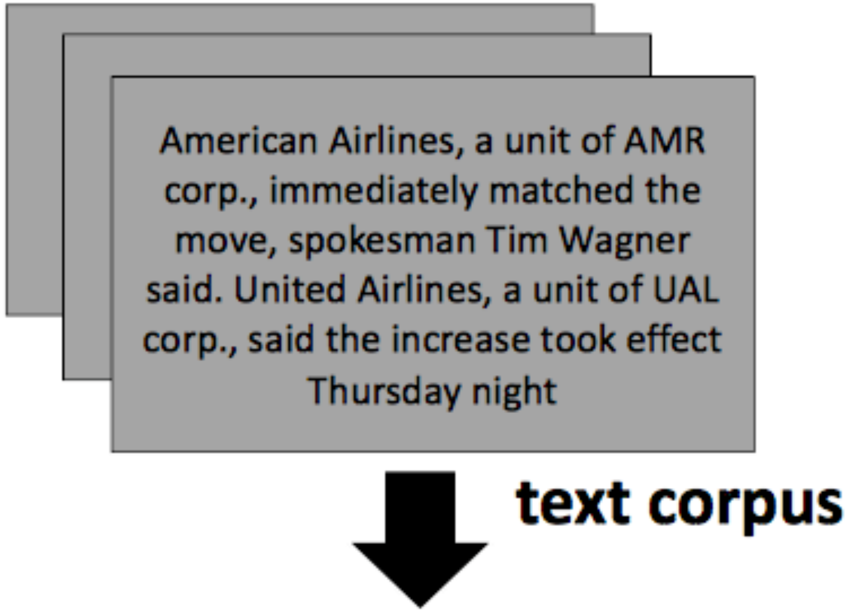
Entities

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



Relations

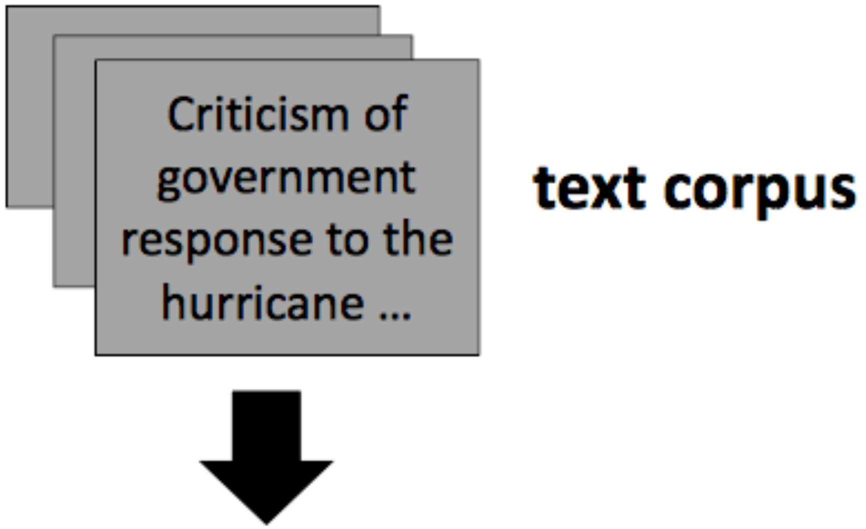
Can computational systems capture **different relations between the entities** from given corpora?

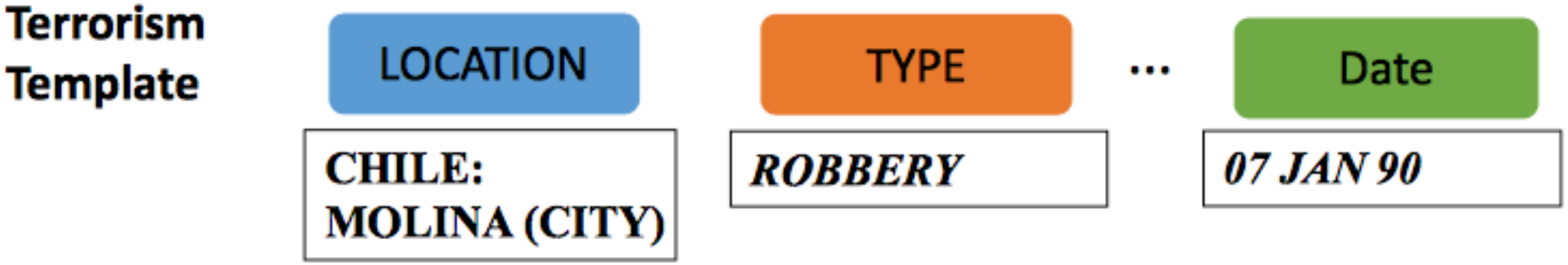


Entity 1	Relation	Entity 2
American Airlines	is_subsidiary_of	AMR
Tim Wagner	is_employee_of	American Airlines
United Airlines	is_subsidiary_of	UAL
...	...	...

Events

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





What is Information Extraction

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

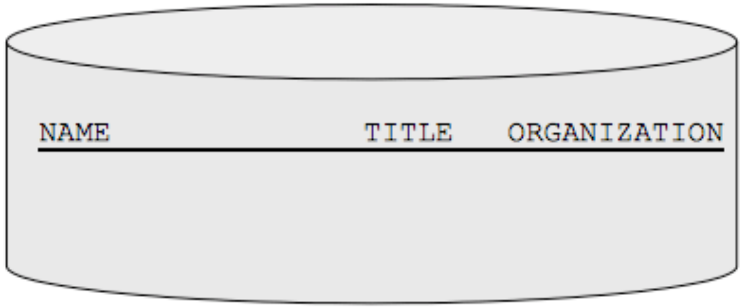
October 14, 2002, 4:00 a.m. PT

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…

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

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

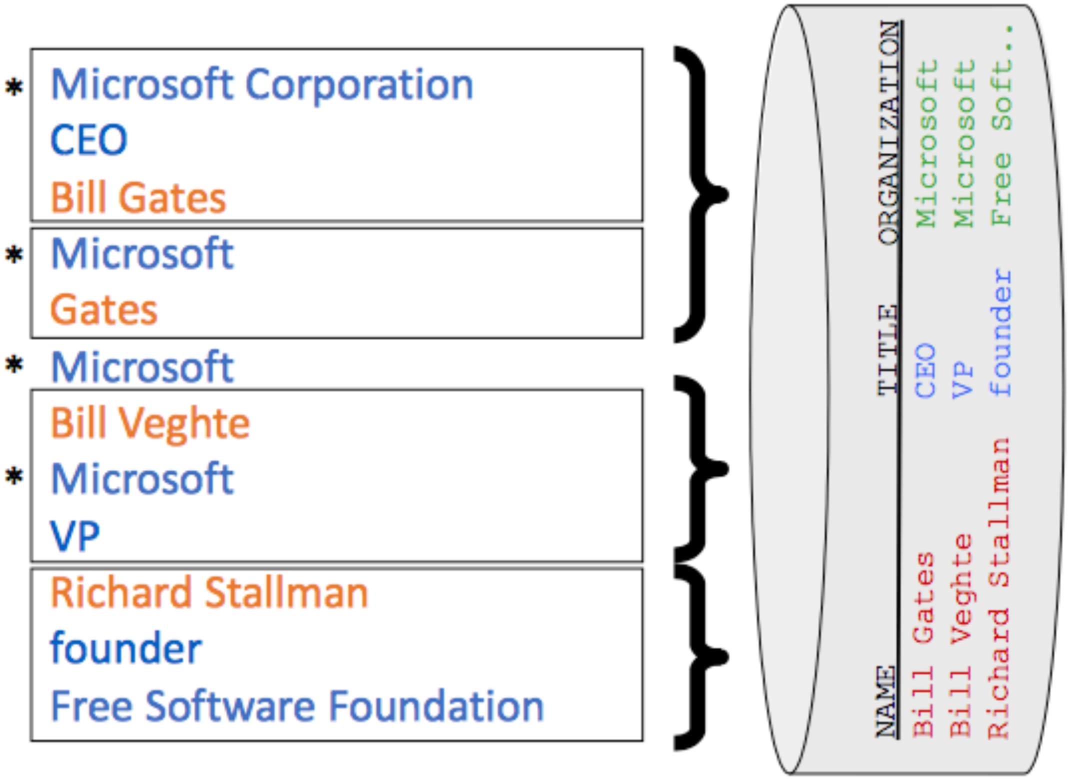
October 14, 2002, 4:00 a.m. PT

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.

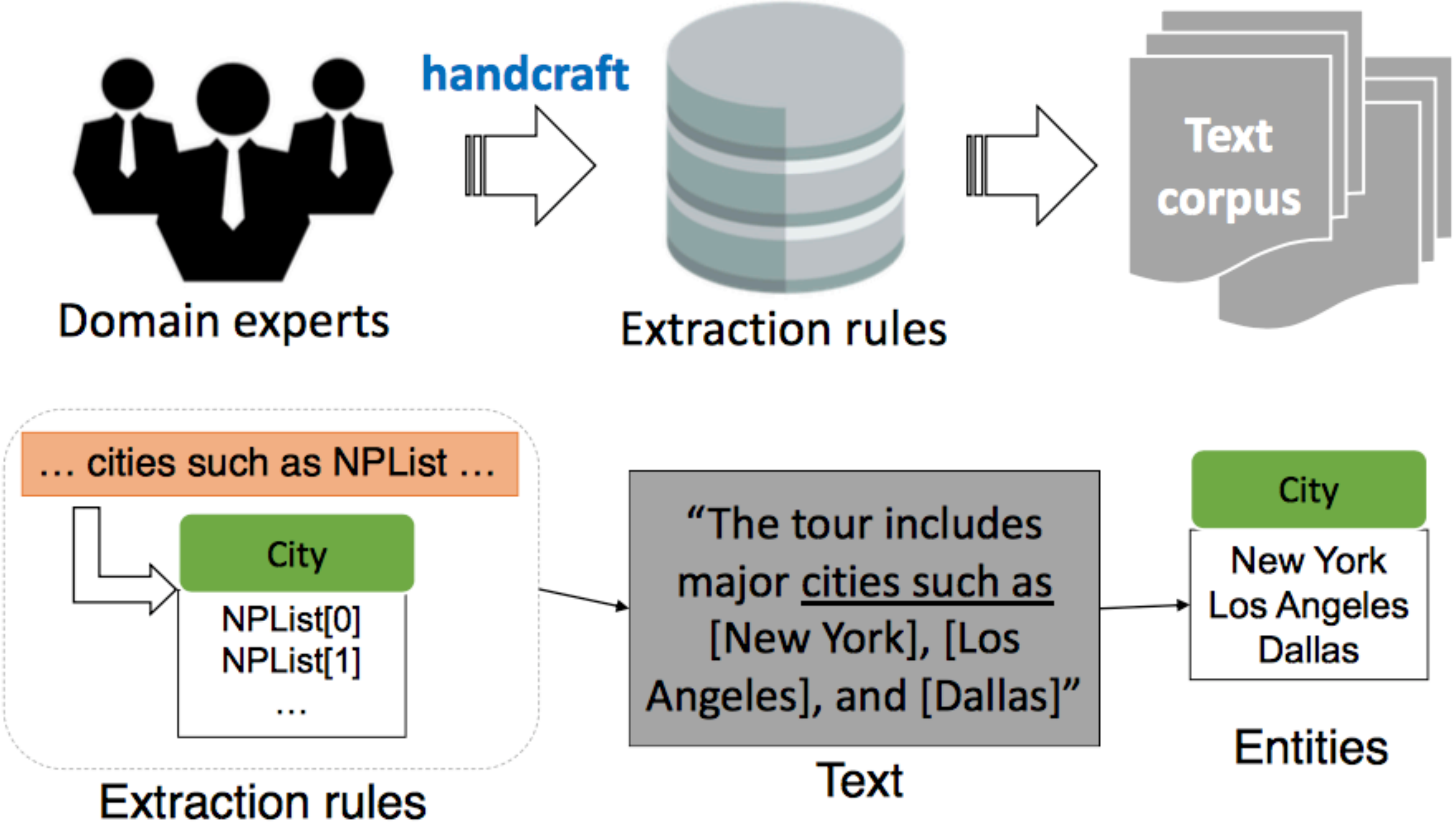
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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."

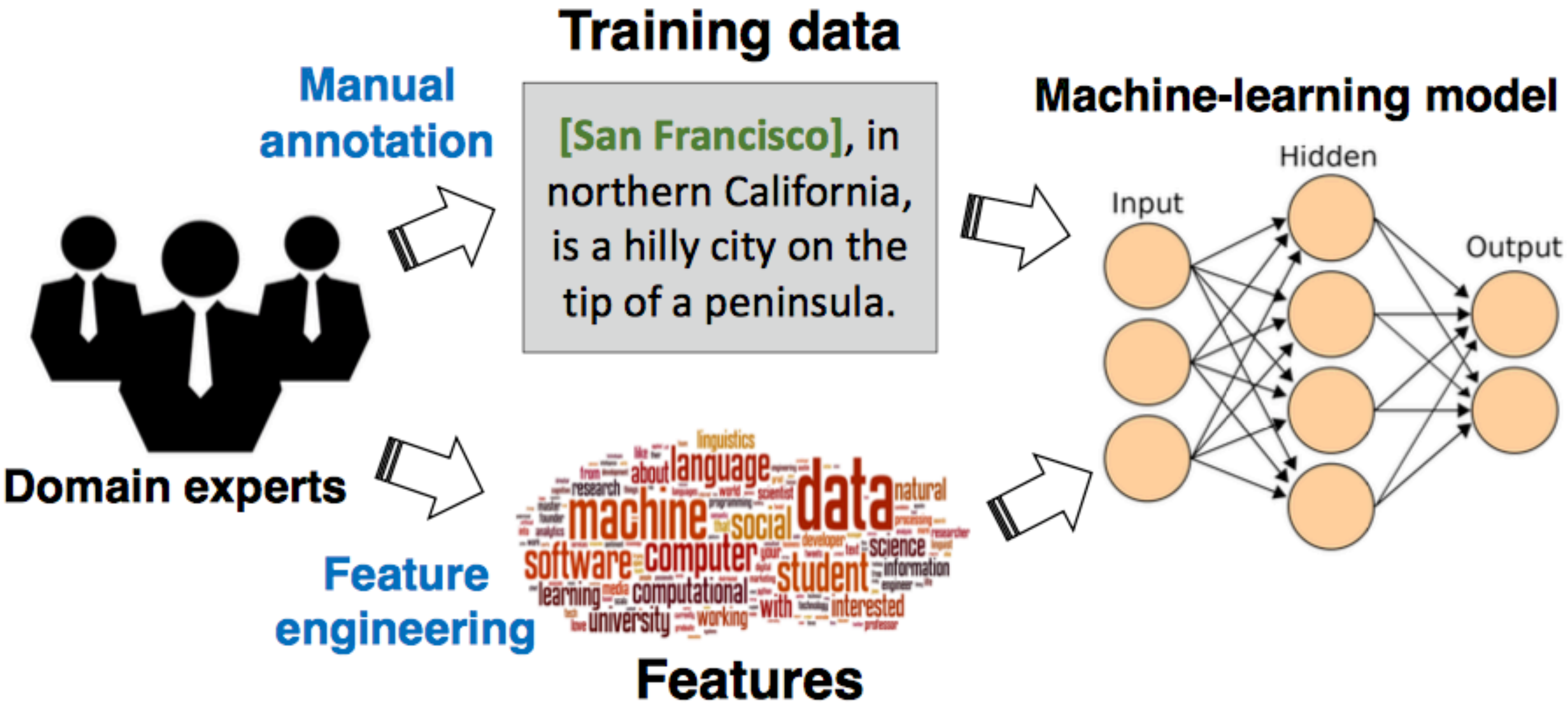
Richard Stallman, founder of the Free Software Foundation, countered saying…



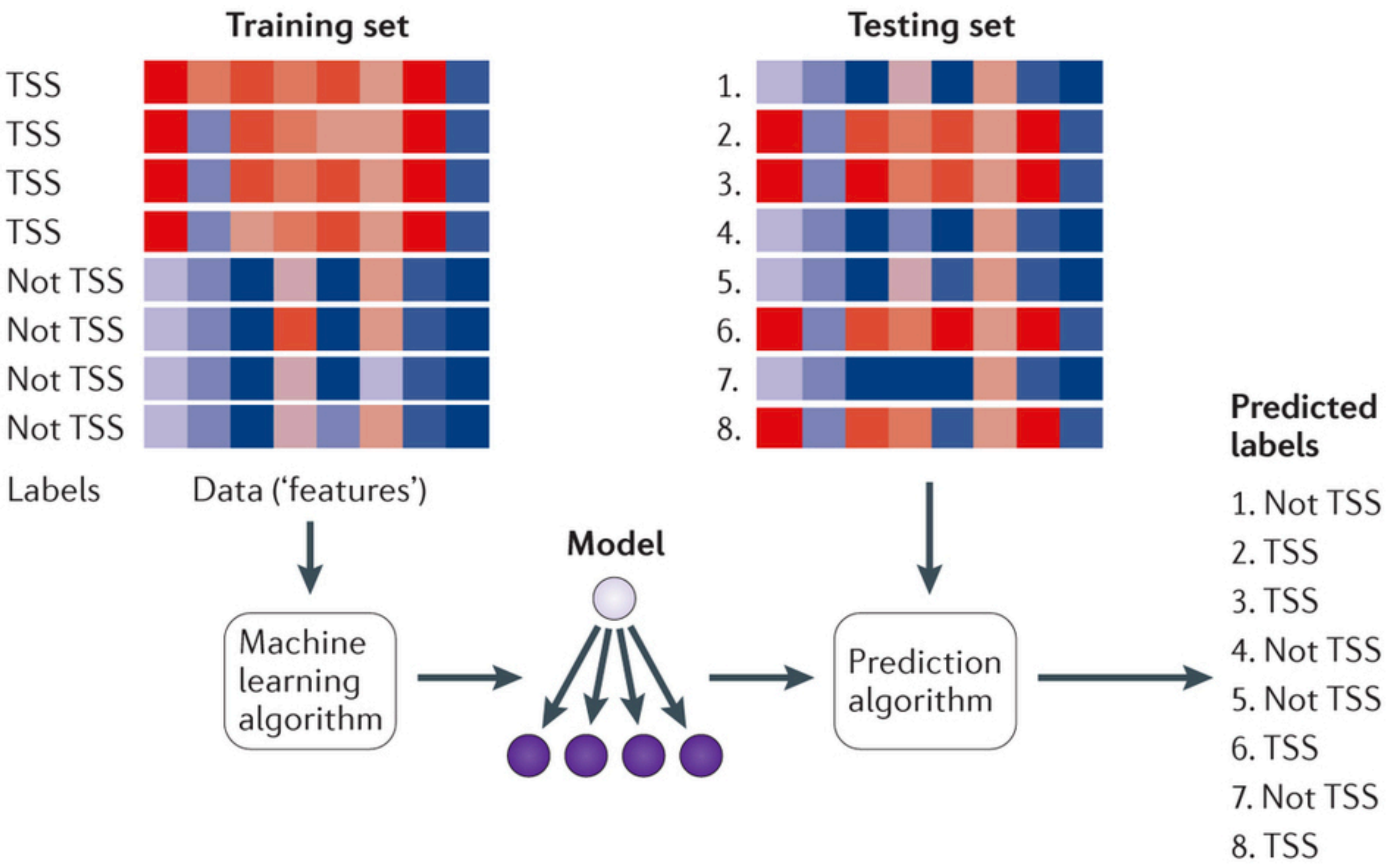
Traditional Rule-Based Systems



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	

Candidate Extraction++

Rheumatoid Arthritis [MalaCards] [LLD]

Network | Comorbidity | GWAS | OMIM | DEG | GeneRIF | GeneWays | miRNA | Drug

Genes that are relevant to **Rheumatoid Arthritis** based on the OMIM Gene Map.

GENE	OMIM ID	OMIM RECORD
CIITA	600005	Bare lymphocyte syndrome, type II, complementation group A Rheumatoid arthritis, susceptibility to
PTPN22	600716	Diabetes, type 1, susceptibility to Rheumatoid arthritis, susceptibility to Systemic lupus erythematosus susceptibility to
IL10	124092	HIV-1, susceptibility to Graft-versus-host disease, protection against Rheumatoid arthritis, progression of
HLA-DRB1	142857	Pemphigoid Sarcoidosis, susceptibility to, 1 Multiple sclerosis, susceptibility to, 1 Rheumatoid arthritis, susceptibility to
CD244	605554	Rheumatoid arthritis, susceptibility to
NFKBIL1	601022	Rheumatoid arthritis, susceptibility to
SLC22A4	604190	Rheumatoid arthritis, susceptibility to
DHX40	605347	Rheumatoid arthritis, susceptibility to
PADI4	605347	Rheumatoid arthritis, susceptibility to
MIF	153620	Rheumatoid arthritis, systemic juvenile, susceptibility to

Regular expressions

`/^#?([a-f0-9]{6}|[a-f0-9]{3})$/`

Remember:
The goal is to maximize recall !

Feature Extraction

Michelle Obama is married to President Barack Obama.

Mention1	Mention2	HasSpouse
Michelle Obama	Barack Obama	

Feature Extraction

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

Mention1	Mention2	HasSpouse
Michelle Obama	Barack Obama	

Mention1	Mention2	feature
M. Obama	B. Obama	PERSON - mary - PERSON
M. Obama	B. Obama	Distance=3

Feature Extraction

Previously users would write features by hand

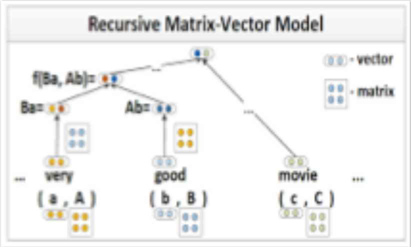
Michelle Obama **is married to** President Barack Obama.

- `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.

WIKIPEDIA
The Free Encyclopedia

Freebase

Positive Example

Spousal Relationship

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

Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

[Adapted example from Luke Zettlemoyer]

Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

[Adapted example from Luke Zettlemoyer]

Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page ...

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

[Adapted example from Luke Zettlemoyer]

Distant Supervision

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page ...

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

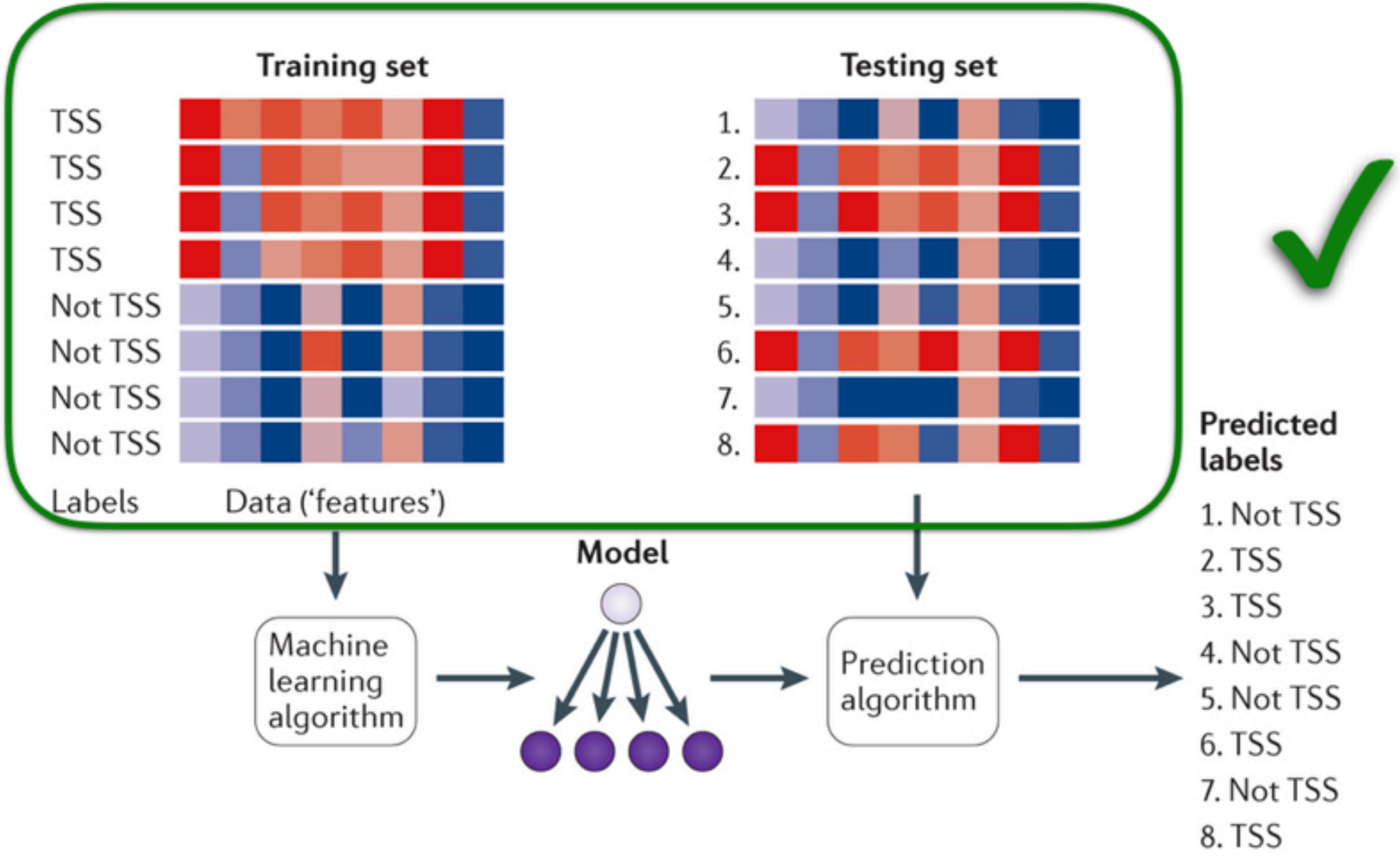
(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

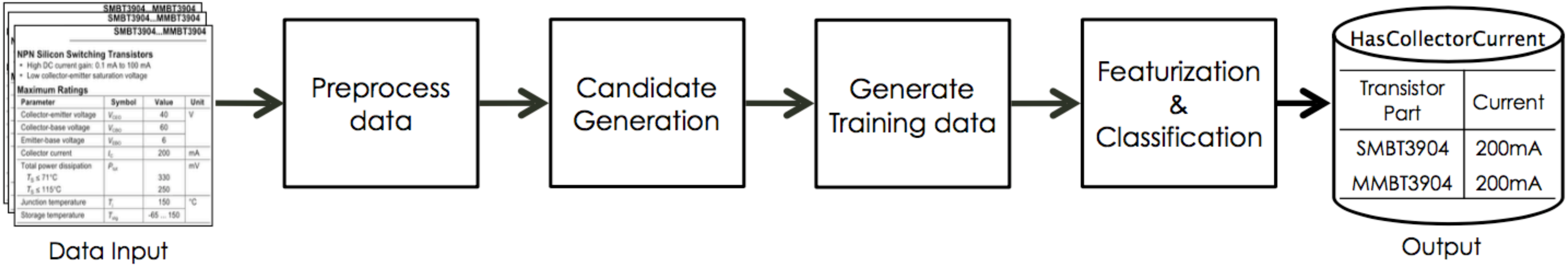
For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]

IE as supervised learning



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



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

Richly formatted data

SMBT3904...MMBT3904

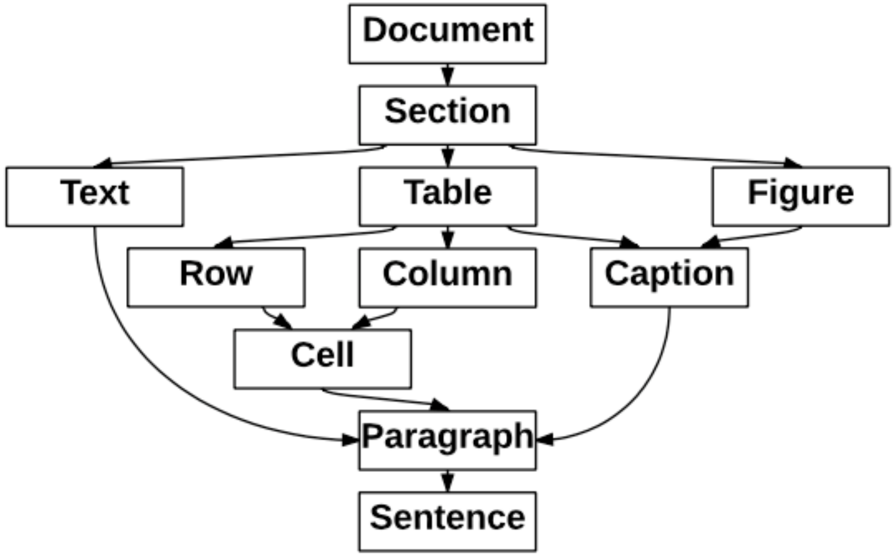
NPN Silicon Switching Transistors

- 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

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

Signals from different modalities can be useful to find the information.