MANCHESTER 1824
The University of Manchester

Healthcare Text Analytics: Analysing Free-text Health Data

Goran Nenadic[1,2,3]

[1]School of Computer Science, University of Manchester
[2] Health eResearch Centre ,The Farr Institute of Health Informatics Research
[3]Manchester Institute of Biotechnology

Farr

Health e-Research Centre

Healtex

Manchester Institute of Biotechnology

Aims of this tutorial

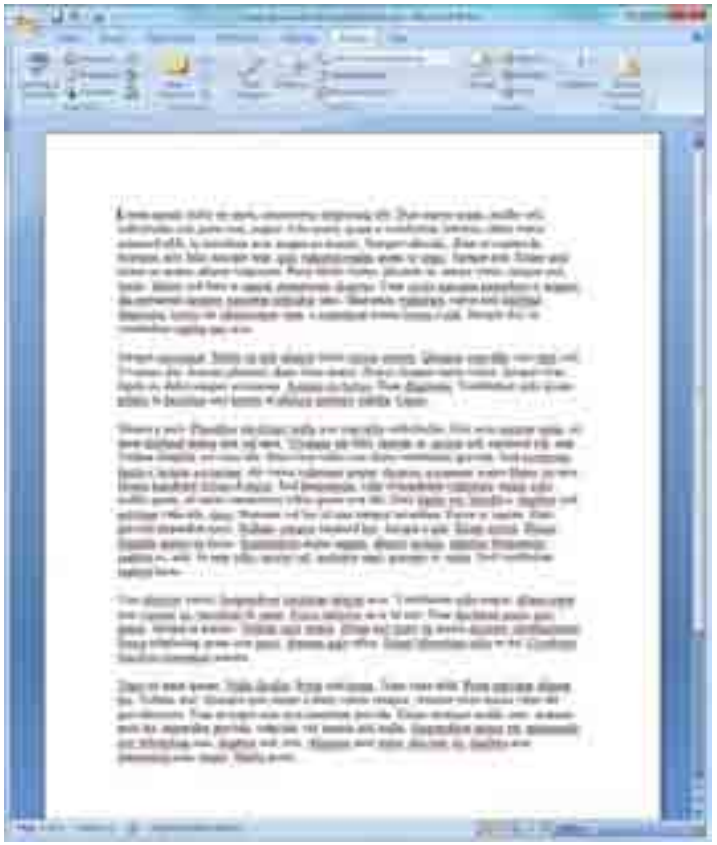
- Understand the needs, opportunities and challenges in the extraction and integration of information from free-text health data sources
- Discuss the main steps in text mining
- Show-case possible application areas
- Discuss the state of the art
- Provide some pointers for further engagement

Plan

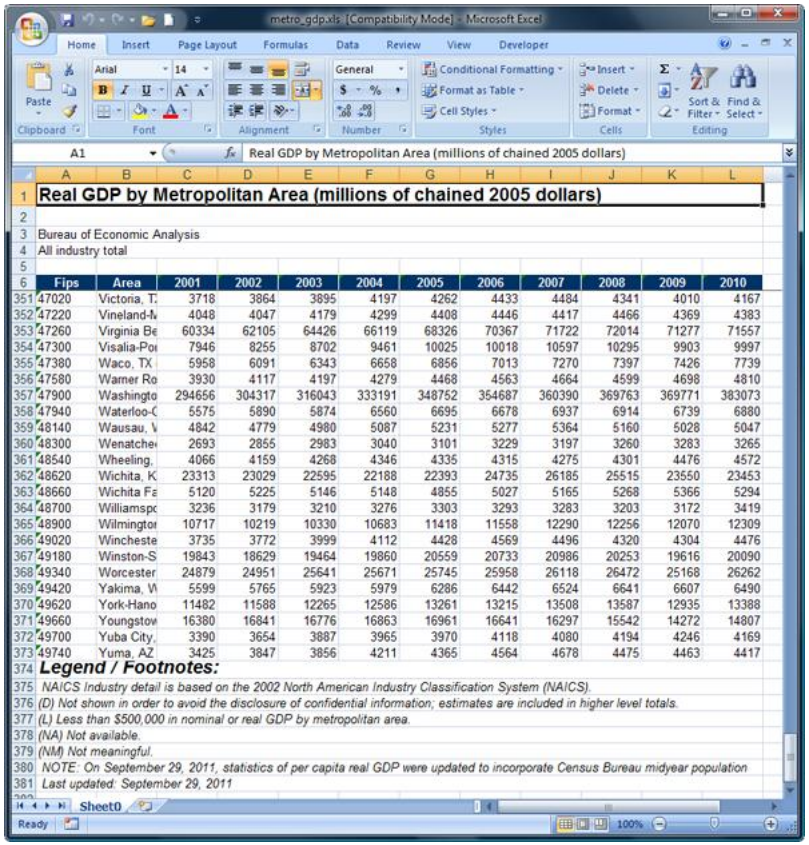
1. Very brief overview of text mining
2. Examples of processing healthcare free text
   - Clinical notes
   - Patient generated data
3. Hands on
   - Extract dosage information from free text
   - Extract symptoms, procedures, anatomical locations
   - Get all patient with hypertension
4. State of the art
5. Join the community

Group reflection: Word vs. Excel

- What is easier to create?
- What is easier to process?



vs



metro_gdp.xls [Compatibility Mode] - Microsoft Excel

Home | Insert | Page Layout | Formulas | Data | Review | View | Developer

A1: Real GDP by Metropolitan Area (millions of chained 2005 dollars)

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Real GDP by Metropolitan Area (millions of chained 2005 dollars)**											
2												
3	Bureau of Economic Analysis											
4	All industry total											
5												
6	Fips	Area	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
351	47020	Victoria, T	3718	3864	3895	4197	4262	4433	4484	4341	4010	4167
352	47220	Vineland-M	4048	4047	4179	4299	4408	4446	4417	4466	4369	4383
353	47260	Virginia Be	60334	62105	64426	66119	68326	70367	71722	72014	71277	71557
354	47300	Visalia-Por	7946	8255	8702	9461	10025	10018	10597	10295	9903	9997
355	47380	Waco, TX	5958	6091	6343	6658	6856	7013	7270	7397	7426	7739
356	47580	Warner Ro	3930	4117	4197	4279	4468	4563	4664	4599	4698	4810
357	47900	Washingto	294656	304317	316043	333191	348752	354687	360390	369763	369771	383073
358	47940	Waterloo-C	5575	5890	5874	6560	6695	6678	6937	6914	6739	6880
359	48140	Wausau, W	4842	4779	4980	5087	5231	5277	5364	5160	5028	5047
360	48300	Wenatchee	2693	2855	2983	3040	3101	3229	3197	3260	3283	3265
361	48540	Wheeling,	4066	4159	4268	4346	4335	4315	4275	4301	4476	4572
362	48620	Wichita, K	23313	23029	22595	22188	22393	24735	26185	25515	23550	23453
363	48660	Wichita Fa	5120	5225	5146	5148	4855	5027	5165	5268	5366	5294
364	48700	Williamspo	3236	3179	3210	3276	3303	3293	3283	3203	3172	3419
365	48900	Wilmington	10717	10219	10330	10683	11418	11558	12290	12256	12070	12309
366	49020	Winchester	3735	3772	3999	4112	4428	4569	4496	4320	4304	4476
367	49180	Winston-S	19843	18629	19464	19860	20559	20733	20986	20253	19616	20090
368	49340	Worcester	24879	24951	25641	25671	25745	25958	26118	26472	25168	26262
369	49420	Yakima, W	5599	5765	5923	5979	6286	6442	6524	6641	6607	6490
370	49620	York-Hano	11482	11588	12265	12586	13261	13215	13508	13587	12935	13388
371	49660	Youngstow	16380	16841	16776	16863	16961	16641	16297	15542	14272	14807
372	49700	Yuba City,	3390	3654	3887	3965	3970	4118	4080	4194	4246	4169
373	49740	Yuma, AZ	3425	3847	3856	4211	4365	4564	4678	4475	4463	4417
374	***Legend / Footnotes:***											
375	*NAICS Industry detail is based on the 2002 North American Industry Classification System (NAICS).*											
376	*(D) Not shown in order to avoid the disclosure of confidential information; estimates are included in higher level totals.*											
377	*(L) Less than $500,000 in nominal or real GDP by metropolitan area.*											
378	*(NA) Not available.*											
379	*(NM) Not meaningful.*											
380	*NOTE: On September 29, 2011, statistics of per capita real GDP were updated to incorporate Census Bureau midyear population*											
381	*Last updated: September 29, 2011*											

Sheet0

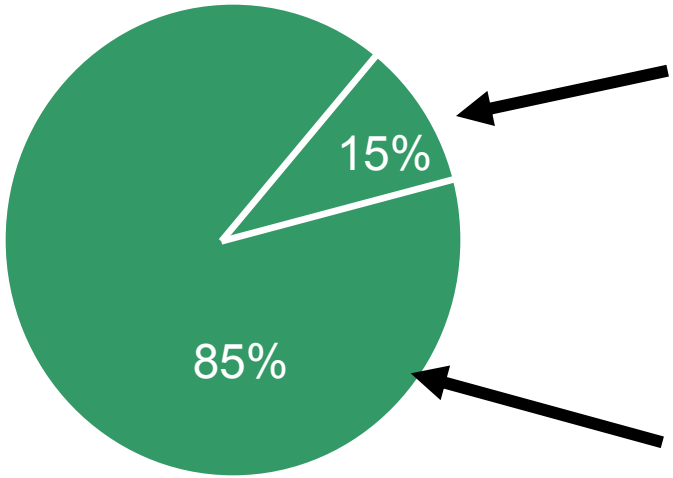
Ready

100%

Natural language data is everywhere

- Internet
  - Newswire, reports, blogs, etc.
  - **Facebook**: 510,000 comments posted every minute (with 293,000 status updates and 136,000 photos)
  - **Twitter**: 6,000 per second (i.e. 200 billion per year)
- Scientific literature
  - ~1 million scientific articles per year in biomedicine only
- Internal company reports
  - One legal company has a repository of 1 billion documents (including different versions)
- Clinical records
  - The Christie Hospital: 135k patients

Actionable information



Structured variables and coded information

e.g. lab-tests, codes, prescriptions

Unstructured or semi-structured information

e.g. clinical narrative, letters, social media, etc.

Group reflection: what is in free-text?

- Do you have a clinical example of information that is mostly or exclusively available in free-text in your domain?

Natural Language Systems

- Started around 1950s
- Various terminology
  - natural language processing (NLP)
  - text mining
    (now widely used as an umbrella for large variety of NLP techniques to denote all approaches to retrieve, extract and analyse text`)
  - text analytics
  - computational linguistics
  - (human) language technology
- Many sub-areas

What does it take to be a data scientist

High Level Skills		
Bachelors	Masters	PHD
ML_NEURAL_NETS_SVM	ML_NEURAL_NETS_SVM	ML_NEURAL_NETS_SVM
OPTIMIZATION	OPTIMIZATION	TEXT_ANALYTICS
DATA_WRANGLING	CLUSTERING	OPTIMIZATION
CLUSTERING	DATA_WRANGLING	LINEAR_REGRESSION
DATA_VISUALIZATION	TIME_SERIES	CLUSTERING
TIME_SERIES	LINEAR_REGRESSION	TIME_SERIES
TEXT_ANALYTICS	TEXT_ANALYTICS	DESIGN_OF_EXPERIMENTS_AB
LINEAR_REGRESSION	DATA_VISUALIZATION	DATA_VISUALIZATION
DESIGN_OF_EXPERIMENTS_AB	DESIGN_OF_EXPERIMENTS_AB	DATA_WRANGLING
BASIC_STATISTICS	BASIC_STATISTICS	LOGISTIC_GLM

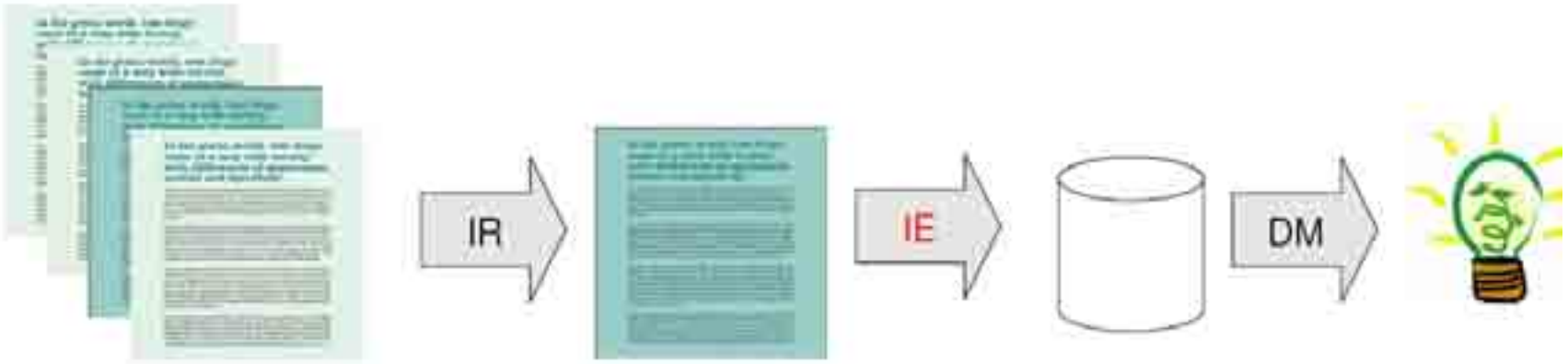
http://blogs.sas.com/content/text-mining/2015/03/27/whats-it-take-to-be-a-data-scienti

Text Mining in one slide

- On May 2, 2014 the then Indian Prime Minister Manmohan Singh revealed: "General Musharraf and I had nearly reached an agreement, a non-territorial solution to all problems..."

- Following in suite of the electric vehicle manufacturer, General Motors has revealed its plans of growing its energy storage market.

Typical text mining framework



- Information Retrieval (IR)
  - find relevant documents [unstructured form]
- **Information Extraction (IE)**
  - find detailed information [(semi-)structured form]
- Data Mining (DM)
  - find associations, build networks, make predictions

Information retrieval (IR)

- Searching for relevant **documents**
  - Find similarity between user query
  - Rank documents
- Result of IR is a set of relevant documents
  - filtering huge collections based on a query
  - no fine-grained information, just whole documents
  - users would need to read and analyse these documents on their own

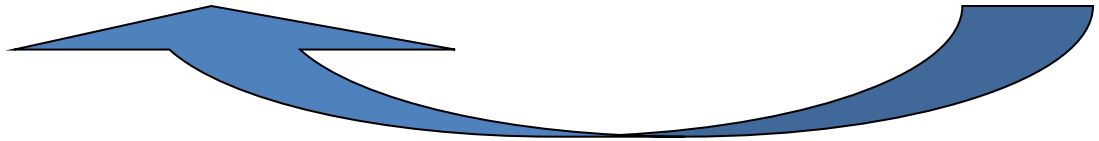
Information extraction (IE)

- Extract information i.e. facts from text
- Identify instances of pre-defined *entities* (dates, names of people, locations, etc.) and relations between them
- Fill in database-like tables with “facts”

Slot	Information
Date	7/10/96 (today)
Location	SanSalvador
Victim injured	policeman
Victim attacked	guards
Perpetrator	urban guerrillas

San Salvador, 7/10/96

It has been officially reported that a policeman was wounded today when urban guerrillas attacked the guards at a power substation located downtown San Salvador.



Main problems: ambiguity

- Different meanings/senses of words
  - e.g. *Apple* (the company) or *apple* (the fruit)
  - e.g. *Toyota* can be a *car* or a *company*
  - e.g. acronyms have different meanings in different contexts

USA =

United States Army	Union of South Africa
United States of America	Union Street Athletics
Ulhasnagar Sindhi Association	Unionville-Sebewaing Area
Ultimate in Suspense and Action	Unique Settable Attributes
Unconditional Self-Acceptance	Unit Self Assessment
Unconventional Stellar Aspect	United Scenic Artists
Under Secretary of the Army	University of South Alabama
Underground Service Alert	University of South Australia
Underground Sewer Adapter	Unix System Admin
Underwriting Service Assistant	Unstable Angina
Unicycling Society of America	Unusually Sensitive Area

Main problems: variability

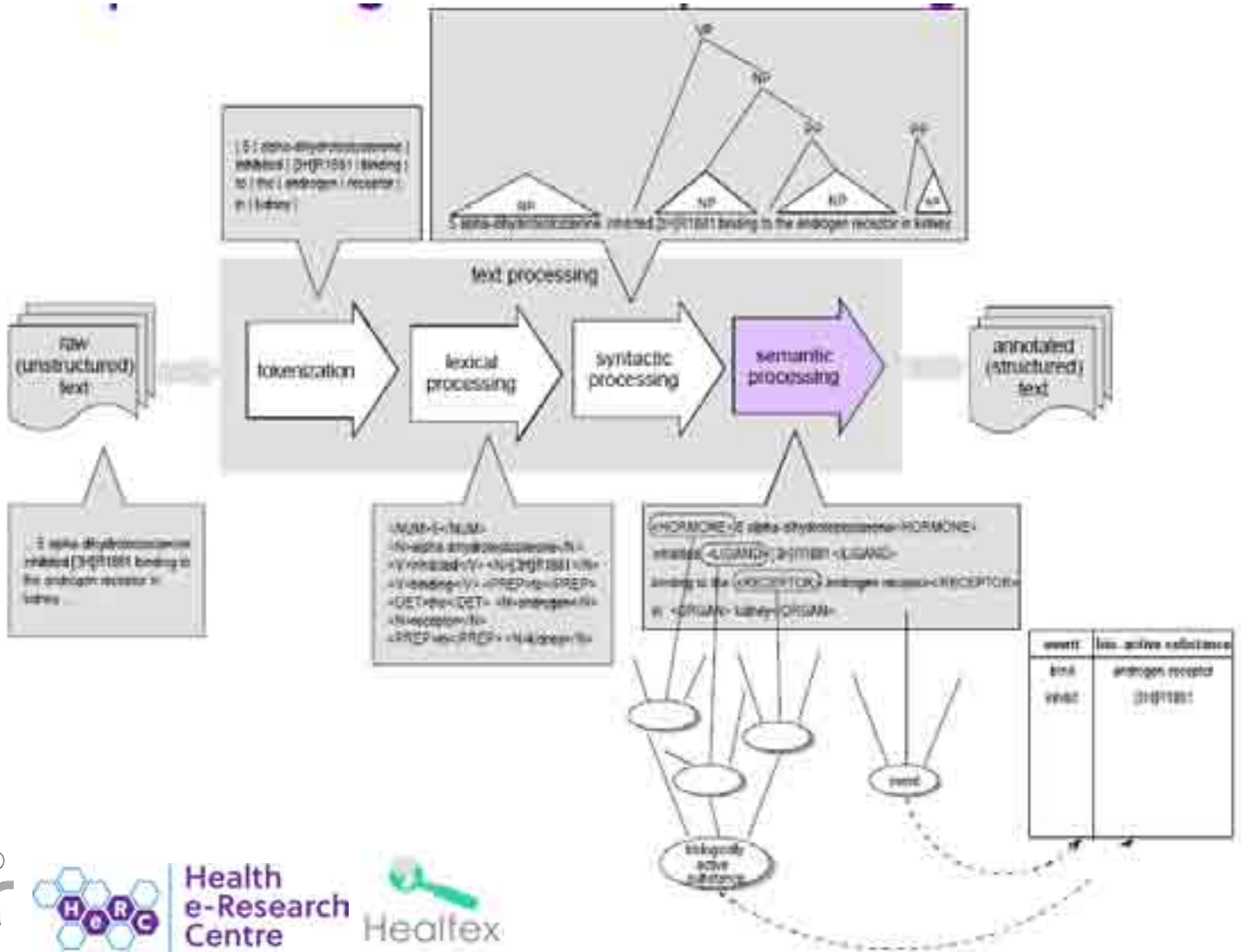
- Numerous ways to express the same thing

Chronic renal impairment
Chronic kidney disease
CKD
kidney failure
Chronic renal failure
CRF
Chronic renal failure syndrome
Renal insufficiency
eGFR 44

Basic text pre-processing

- Basic ≠ trivial
- Typically includes:
  - Identify words (tokenisation)
    - *oxycodone 5-10 mg p.o. q.4 h. as needed for pain.*
  - Normalise word forms
    - e.g. plural forms, or past tense verbs
  - Segment sentences
    - punctuation not always used!

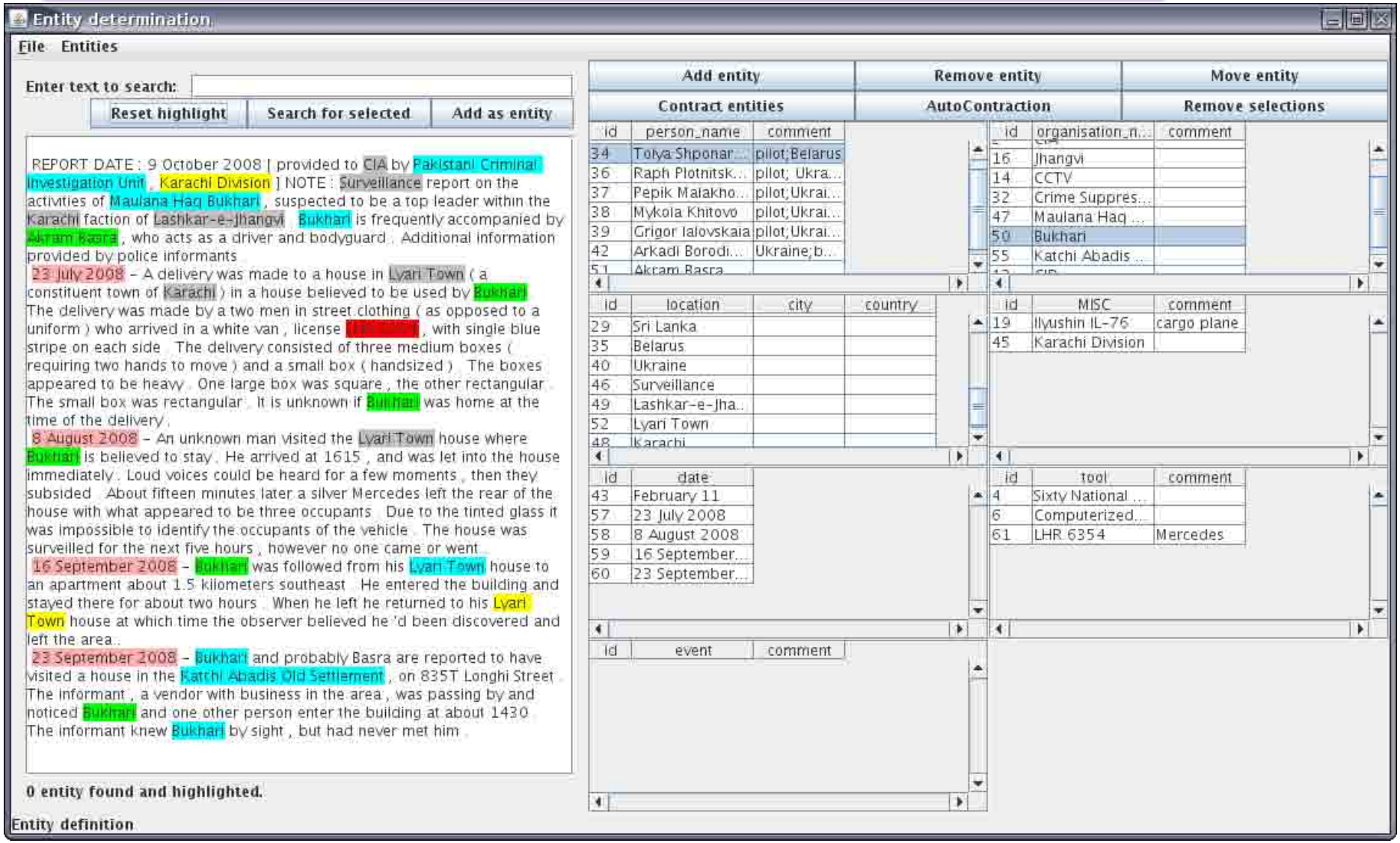
Main steps in text processing



Finding entities/concepts

- Identification of *entities* and *terms* of interest
  - find strings in text that denote specific entities or concepts (named-entity recognition = NER)
  - e.g. persons, jobs, organisations, dates, time, locations, symptoms, disorders, medications, treatments, …
  - designed for each *class* of interest

NER example


Entity determination
File Entities
Enter text to search:
Reset highlight
Search for selected
Add as entity
REPORT DATE : 9 October 2008 [provided to CIA by Pakistani Criminal Investigation Unit , Karachi Division] NOTE : Surveillance report on the activities of Maulana Haq Bukhari , suspected to be a top leader within the Karachi faction of Lashkar-e-Jhangvi . Bukhari is frequently accompanied by Akram Basra , who acts as a driver and bodyguard . Additional information provided by police informants .
23 July 2008 – A delivery was made to a house in Lyari Town (a constituent town of Karachi) in a house believed to be used by Bukhari . The delivery was made by a two men in street clothing (as opposed to a uniform) who arrived in a white van , license [illegible] , with single blue stripe on each side . The delivery consisted of three medium boxes (requiring two hands to move) and a small box (handsized) . The boxes appeared to be heavy . One large box was square , the other rectangular . The small box was rectangular . It is unknown if Bukhari was home at the time of the delivery .
8 August 2008 – An unknown man visited the Lyari Town house where Bukhari is believed to stay . He arrived at 1615 , and was let into the house immediately . Loud voices could be heard for a few moments , then they subsided . About fifteen minutes later a silver Mercedes left the rear of the house with what appeared to be three occupants . Due to the tinted glass it was impossible to identify the occupants of the vehicle . The house was surveilled for the next five hours , however no one came or went .
16 September 2008 – Bukhari was followed from his Lyari Town house to an apartment about 1.5 kilometers southeast . He entered the building and stayed there for about two hours . When he left he returned to his Lyari Town house at which time the observer believed he 'd been discovered and left the area .
23 September 2008 – Bukhari and probably Basra are reported to have visited a house in the Katchi Abadis Old Settlement , on 835T Longhi Street . The informant , a vendor with business in the area , was passing by and noticed Bukhari and one other person enter the building at about 1430 . The informant knew Bukhari by sight , but had never met him .
0 entity found and highlighted.
Entity definition
Add entity
Remove entity
Move entity
Contract entities
AutoContraction
Remove selections
id person_name comment
34 Tolya Shponar... pilot;Belarus
36 Raph Plotnitsk... pilot; Ukra...
37 Pepik Malakho... pilot;Ukrai...
38 Mykola Khitovo pilot;Ukrai...
39 Grigor Ialovskaia pilot;Ukrai...
42 Arkadi Borodi... Ukraine;b...
51 Akram Basra
id organisation_n... comment
16 Jhangvi
14 CCTV
32 Crime Suppres...
47 Maulana Haq ...
50 Bukhari
55 Katchi Abadis ...
id location city country
29 Sri Lanka
35 Belarus
40 Ukraine
46 Surveillance
49 Lashkar-e-Jha...
52 Lyari Town
48 Karachi
id MISC comment
19 Ilyushin IL-76 cargo plane
45 Karachi Division
id date
43 February 11
57 23 July 2008
58 8 August 2008
59 16 September...
60 23 September...
id tool comment
4 Sixty National ...
6 Computerized...
61 LHR 6354 Mercedes
id event comment

Two kinds of NE approaches

Knowledge Engineering

- developed by experienced language engineers
- make use of human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

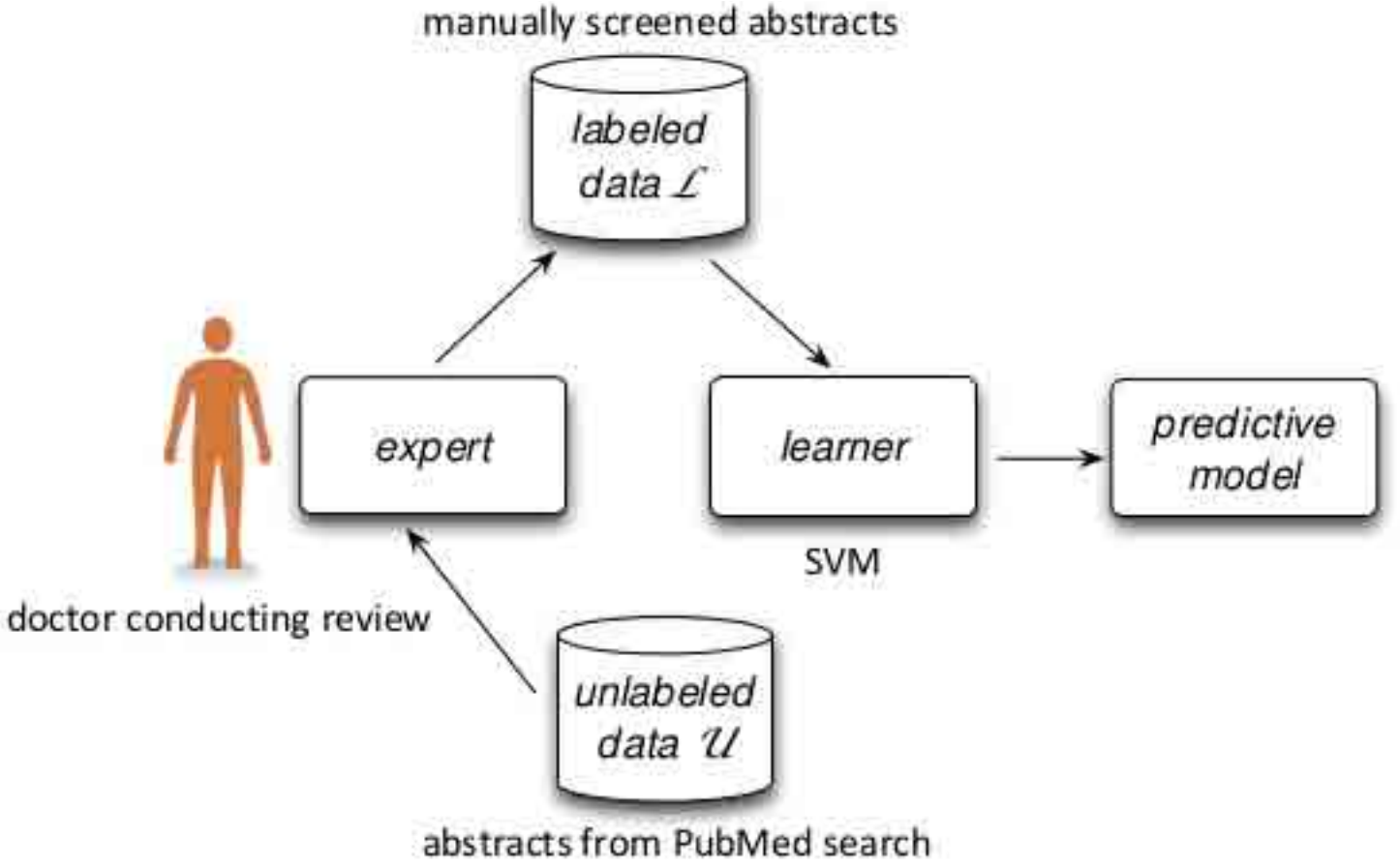
Learning Systems

- use statistics or other machine learning
- developers do not need language expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

NER methods

- **dictionary-based** (gazetteers)
  - matches items from a **lexicon** against text
- **rule-based**
  - hand-written regular expressions
  - e.g. <title> <capitalised word>$^{+}$ → personal name
  - e.g. <person>Mr John Smith Jr.</person>
- **machine learning**
  - often supervised models to identify specific entities
- **hybrid** – any combination of the above

Approaches – machine learning



T. Liptrot

Finding relations between entities

- Extract specific facts, relations and events by linking entities
  - use of templates, regular expressions, grammars

  <PERSON> is appointed as a <JOB> of <COMPANY>

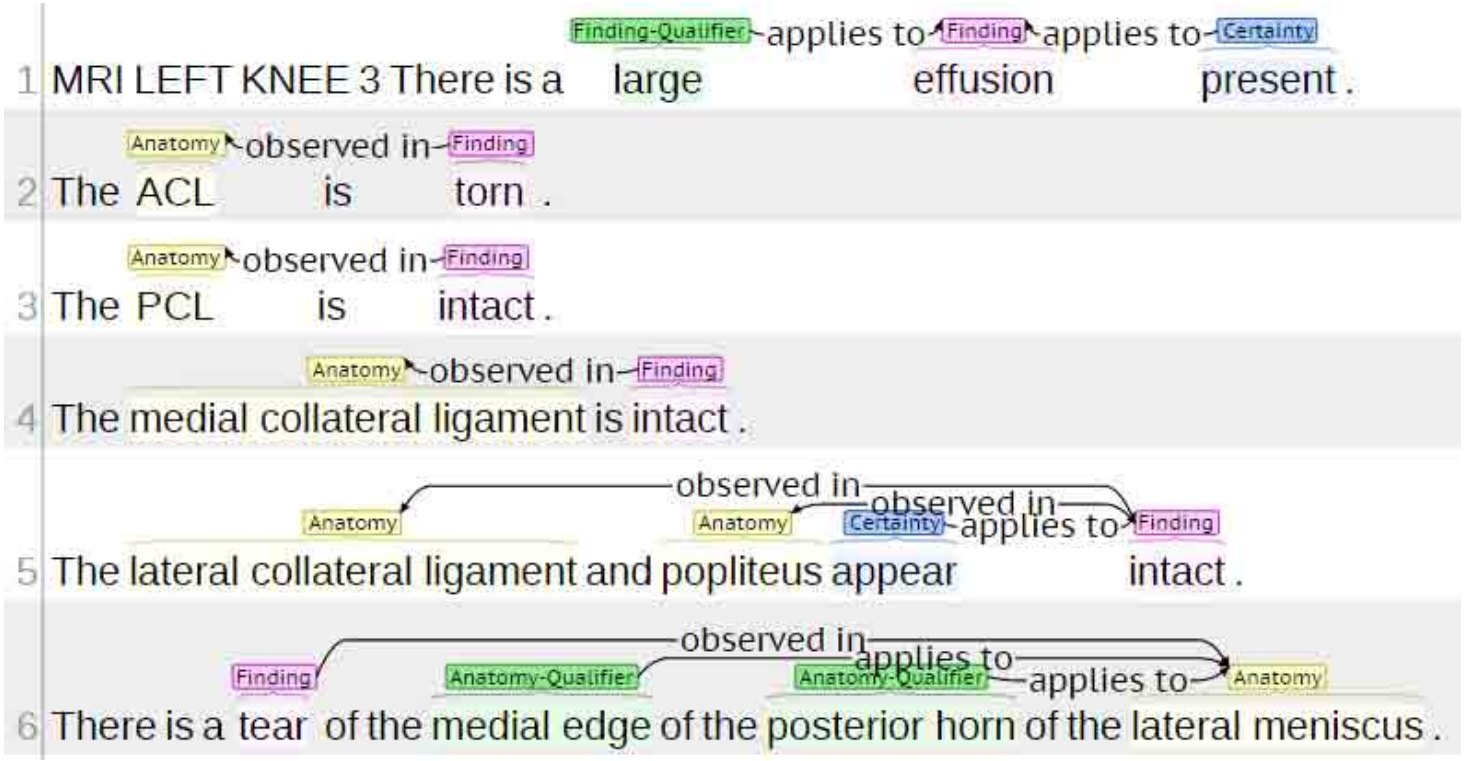
  <DRUG> for <SYMPTOM>

  - typically designed around important verbs –
    - e.g. *attacked, bought*, *appointed*, *merged*, *acquired*, etc.

  - also, machine learning approaches

Finding relations between entities

- Manual annotation of a training set



Quality metrics

- The result is compared with a manual“gold standard”
  - How consistent the humans are/can be?
  - 100% is impossible even for human annotators.
- Typical measures:
  - **Sensitivity** (true positive rate, recall, probability of detection)
    proportion of positives that are correctly identified as such
  - **Specificity** (true negative rate)
    proportion of negatives that are correctly identified as such
  - **Precision** (positive predictive value, PPV)
    proportion of retrieved elements that are positives
  - **F-measure** (harmonic mean of sensitivity and PPV)
    combines sensitivity (recall) and precision

Measuring agreement (Kappa)

- Measures the agreement between two annotators who each classify *N* items into *C* mutually exclusive categories

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

- Pr(*a*) = relative observed agreement among annotators
- Pr(*e*) = the hypothetical probability of chance agreement
  - typically using the observed data to calculate the probabilities of each observer randomly saying each category
- If the annotators are in complete agreement then $\kappa = 1$.
- $\kappa = 0$ if there is no agreement among the annotators other than what would be expected by chance (as defined by Pr(*e*))

Mining healthcare free-text

Farr

Health e-Research Centre

Healtex

Routinely collected narrative data

- Clinician notes, letters are part of routinely collected data

ALLERGIES: To IV contrast , penicillin
for which she has a hive reaction.

cerebrovascular accident.

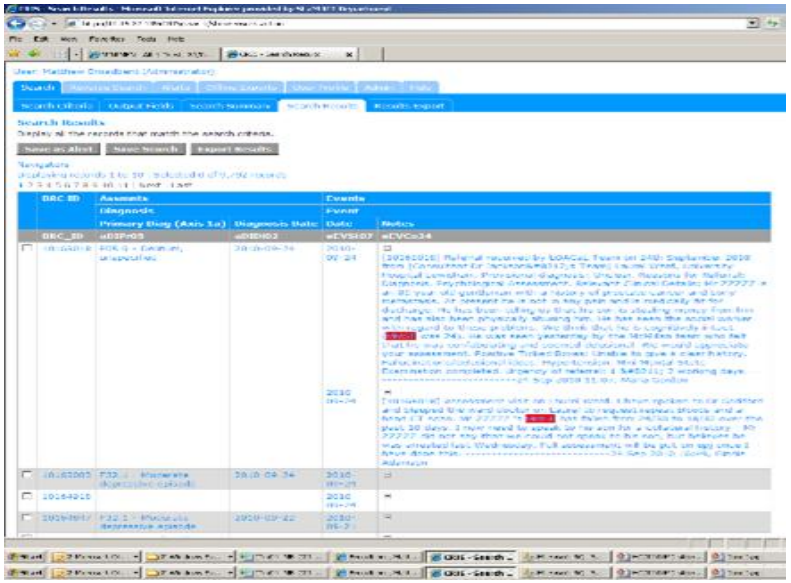
OTHER DIAGNOSES PERIPHERAL VASCULAR DISEASE , HYPERTENSION.

- People communicate most efficiently in narrative text
- Not everything can be captured in pull-down menus or check-lists
- Codes and terminologies can be difficult to navigate – even for experts

Large-scale data is available

Clinical Record Interactive Search (CRIS) system

- SLaM CRIS
- De-identified EPR database
- Structured and free text data



UK-CRIS network

- 10 Mental Health NHS Trusts (total 2M patients)

CRIS NETWORK

Text Mining Clinical Narrative

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed. Since then, self-monitoring of blood glucose (SMBG) showed blood glucose levels of 250-270 mg/dL. She was referred to an endocrinologist for further evaluation.

On examination, she was normotensive and not acutely ill. Her body mass index (BMI) was 18.7 kg/m2 following a recent 10 lb weight loss. Her thyroid was symmetrically enlarged and ankle reflexes absent. Her blood glucose was 272 mg/dL, and her hemoglobin A1c (HbA1c) was 10.3%. A lipid profile showed a total cholesterol of 261 mg/dL, triglyceride level of 321 mg/dL, HDL level of 48 mg/dL, and an LDL of 150 mg/dL. Thyroid function was normal. Urinanalysis showed trace ketones.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years, and limited her alcohol intake to 1 drink daily. Her mother's brother was diabetic.

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed. Since then, self-monitoring of blood glucose (SMBG) showed blood glucose levels of 250-270 mg/dL. She was referred to an endocrinologist for further evaluation.

On examination, she was normotensive and not acutely ill. Her body mass index (BMI) was 18.7 kg/m2 following a recent 10 lb weight loss. Her thyroid was symmetrically enlarged and ankle reflexes absent. Her blood glucose was 272 mg/dL, and her hemoglobin A1c (HbA1c) was 10.3%. A lipid profile showed a total cholesterol of 261 mg/dL, triglyceride level of 321 mg/dL, HDL level of 48 mg/dL, and an LDL of 150 mg/dL. Thyroid function was normal. Urinanalysis showed trace ketones.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years, and limited her alcohol intake to 1 drink daily. Her mother's brother was diabetic.

Text Mining Clinical Notes

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	patient
relative temporal context:	3 months ago
negation indicator:	not negated

Medication CEM	
text:	Glyburide
code:	315989
subject:	patient
frequency:	once daily
negation indicator:	not negated
strength:	2.5 mg

Tobacco Use CEM	
text:	smoking
code:	365981007
subject:	patient
relative temporal context:	25 years
negation indicator:	not negated

Disorder CEM	
text:	diabetes mellitus
code:	73211009
subject:	family member
relative temporal context:	
negation indicator:	not negated

A 43-year-old woman was diagnosed with type 2 diabetes mellitus by her family physician 3 months before this presentation. Her initial blood glucose was 340 mg/dL. Glyburide 2.5 mg once daily was prescribed.

She adhered to a regular exercise program and vitamin regimen, smoked 2 packs of cigarettes daily for the past 25 years

Her mother's brother was diabetic.

Text Mining Clinical Notes

- Transform free-text into structured data: extract clinical *variables* and their *values*
  - **Diseases, problems, symptoms**
    - acute dizzyness, nausea
  - **Anatomy**
    - upper eyelid, tongue, throat
  - **Drugs/medications**
    - steroids, desloratadine, zoladex
  - **Adverse drug events**
    - differentiate from symptoms
  - **Diagnostic/therapeutic procedures**
    - hysterectomy
  - **Behaviour/quality of life**
    - anxiety, problems at work, poor sleep

Mining clinical narrative

Examples with real data

- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

Extracting baseline patient data

- 6 elements:
  - primary disease site
  - histology
  - stage
  - performance status
  - comorbidities
  - responsible consultant
  - date of diagnosis & death
- 135,000 sets of notes – manual work would take 80 years

Approaches – rule-based

- Create *rules* that are matched against text

T[Xx0-4isd][abcdismi]*
N[Xx0-4is][abcdismi]*
M[01Xx]

"…scan showed a stage of **T3N4M0…**"

"Current diagnosis showed **TxN0M0…**"

Results

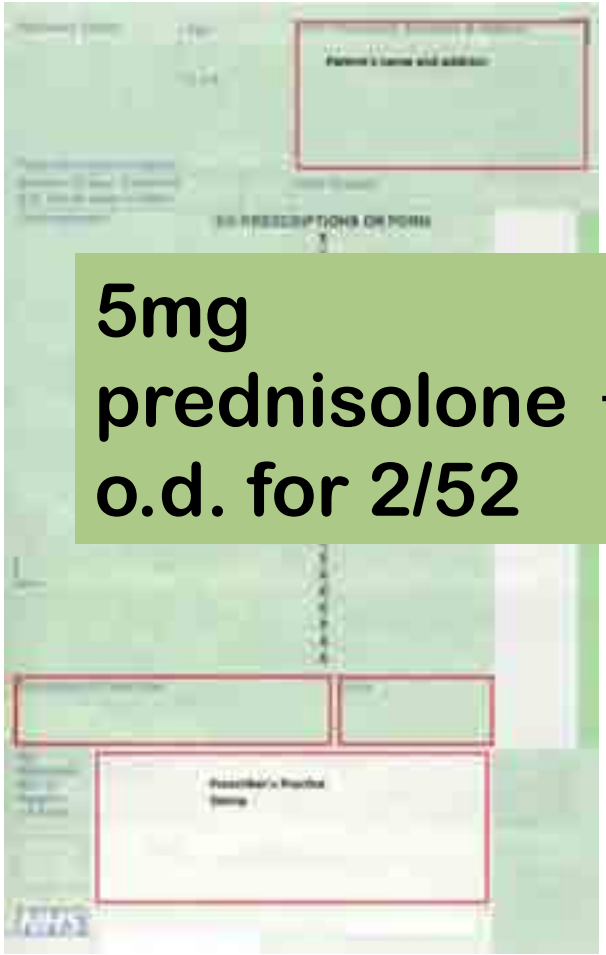
class	method	precision	recall	f_measure	extras
site	abstractors*	0.83	NA	NA	NA
	rules	0.84	0.61	0.71	0.01
	ml	**0.96**	0.53	0.68	0.00
	ml + rules	0.88	**0.72**	0.79	**0.01**
histology	abstractors*	**0.92**	NA	NA	NA
	rules	0.89	0.71	0.79	0.02
	ml	0.83	0.32	0.46	0.09
	ml + rules	0.84	**0.82**	**0.83**	**0.10**
t stage	rules	0.84	0.70	0.76	0.05
	ml	0.90	0.75	0.82	0.09
	ml + rules	**0.91**	**0.85**	**0.88**	**0.12**
n stage	rules	0.90	0.55	0.68	0.00
	ml	**0.94**	0.62	0.75	0.21
	ml + rules	0.91	**0.81**	**0.86**	**0.21**
m stage	rules	**1.00**	0.36	0.53	0.00
	ml	**1.00**	0.41	0.59	0.19
	ml + rules	**1.00**	**0.55**	**0.71**	**0.19**
ps	rules	**1.00**	0.87	0.93	0.07
	ml	0.97	0.76	0.85	0.23
	ml + rules	0.97	**0.92**	**0.95**	**0.25**
responsible consultant	ml	**0.85**	**0.75**	**0.79**	**0.03**
	abstractors*	0.84	NA	NA	NA

Mining clinical narrative

Examples with real data

- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

Medication prescriptions



5mg prednisolone o.d. for 2/52

Raw Data

	studyno	start	stop	ondrug
1	[illegible]	11/01/2000	27/05/2000	1
2	[illegible]	17/08/2000	14/07/2001	0
3	[illegible]	14/07/2001	20/07/2002	1
4	[illegible]	10/07/2002	15/08/2003	1
5	[illegible]	15/08/2003	12/10/2003	0
6	[illegible]	12/10/2003	12/10/2003	1
7	[illegible]	12/10/2003	17/02/2003	1
8	[illegible]	17/02/2003	13/07/2003	1
9	[illegible]	13/07/2003	01/08/2004	0
10	[illegible]	02/08/2004	20/11/2005	0
11	[illegible]	19/11/2001	05/07/2002	1
12	[illegible]	01/03/2002	14/07/2002	0
13	[illegible]	14/07/2002	12/10/2002	0
14	[illegible]	12/10/2002	12/07/2004	1
15	[illegible]	12/07/2004	21/02/2005	0
16	[illegible]	21/02/2005	20/11/2005	0
17	[illegible]	20/01/2002	15/01/2004	1
18	[illegible]	15/01/2004	15/03/2004	0
19	[illegible]	15/03/2004	13/04/2004	0
20	[illegible]	13/06/2004	15/06/2004	0
21	[illegible]	15/06/2004	16/02/2005	0
22	[illegible]	16/01/2005	20/11/2005	0
23	[illegible]	20/01/2002	14/02/2005	0
24	[illegible]	14/02/2005	20/11/2005	0
25	[illegible]	15/01/2002	14/06/2002	0
26	[illegible]	14/06/2002	12/09/2002	0
27	[illegible]	12/09/2002	15/02/2005	0
28	[illegible]	15/02/2005	20/11/2005	0
29	[illegible]	20/01/2002	14/07/2002	1
30	[illegible]	14/07/2002	12/07/2003	0
31	[illegible]	12/07/2003	12/09/2003	0
32	[illegible]	15/08/2003	04/02/2005	0

Data prepared for analysis

Medication prescriptions

- Dose, frequency, duration, intervals
  - take 2 tablets 4 times a day for 2/52

- But often complex prescriptions
  - one **or** two to be taken every 4 **to** 6 hours
  - 10mg to be taken **weekly**
  - a half to one tablet to 2 three times a day **when required**

Karystianis et al, BMC Med Inform Decis Mak. 2016;16:18.
doi: 10.1186/s12911-016-0255-x.

Modelling prescription data

Prescription	dose		freq		interval		unit
	min	max	min	max	min	max	
take 2 tablets 4 times a day	2	2	4	4	1	1	tablet
a half to one tablet to 2 three times a day when required	0.5	1	0	3	1	1	tablet
10mg to be taken weekly	10	10	1	1	7	7	mg
2 with each meal	2	2	3	3	1	?	?
1 to 3 every day	1	3	1	1	1	1	?
one or two to be taken every 4 to 6 hours	1	2	4	6	1	1	?

Variability in prescriptions

- 56,000 most common free text instructions from **CPRD** records
  - CPRD = a UK Anonymised Primary Care EHR Research Database, with over 11M patients

- at least **1 in 4** has inherent **variability**
  - i.e. a choice in taking medication specified by different minimum and maximum doses, duration or frequency.

Arthritis Research UK

Karystianis et al, BMC Med Inform Decis Mak. 2016;16:18.
doi: 10.1186/s12911-016-0255-x.

CPRD MORE DIMENSIONS TO DATA

Mining clinical narratives

Examples with real data

- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

Mining Patient Journeys

- Can text mining techniques be used to <u>reconstruct patient journeys/pathways</u> from unstructured clinical narratives?

- Data: longitudinal narratives
  - Clinical narratives: internal notes and correspondence letters
  - Study participants (n = 27): adult survivors of childhood central nervous cancer (medulloblastoma)

"Comparing the narrative experiences of patients with medulloblastoma with factors identified from their hospital records"

EPSRC

The Christie Charitable Fund

The Christie NHS
NHS Foundation Trust

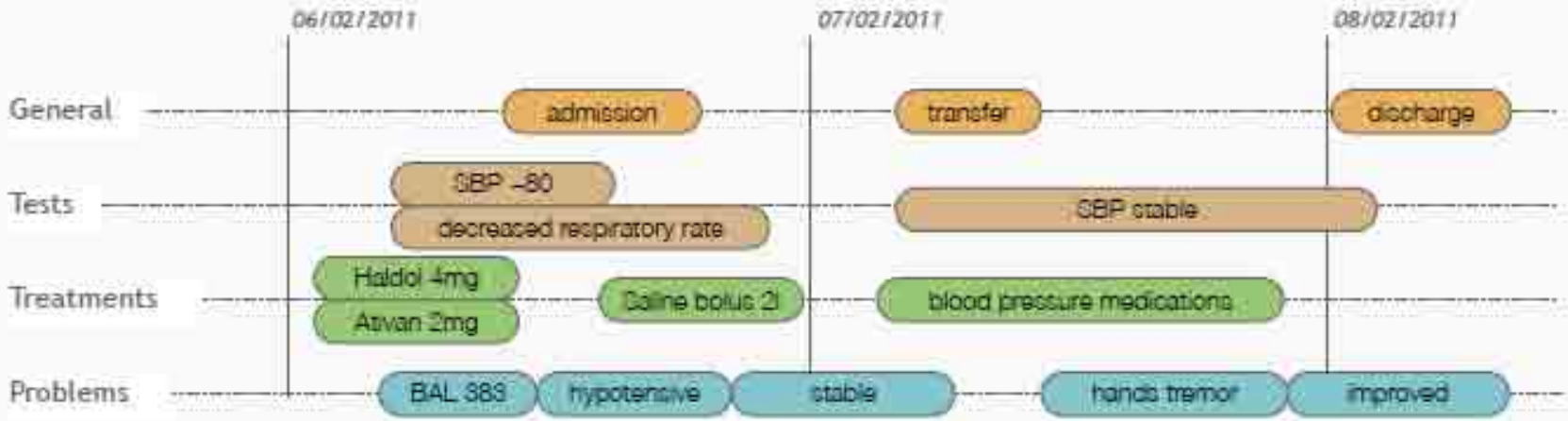
Mining Patient Journeys

- **Healthcare concept extraction**
  - *problems* (e.g., headache, brain cancer),
  - *treatments* (e.g., chemotherapy, radiotherapy)
  - *tests* (e.g., MRI scan, blood test)
  - *health-related quality of life* (e.g., physical, emotional functioning)

- **Temporal Information Extraction**
  - Recognition and normalisation of temporal expressions
  - Chronological ordering of concepts

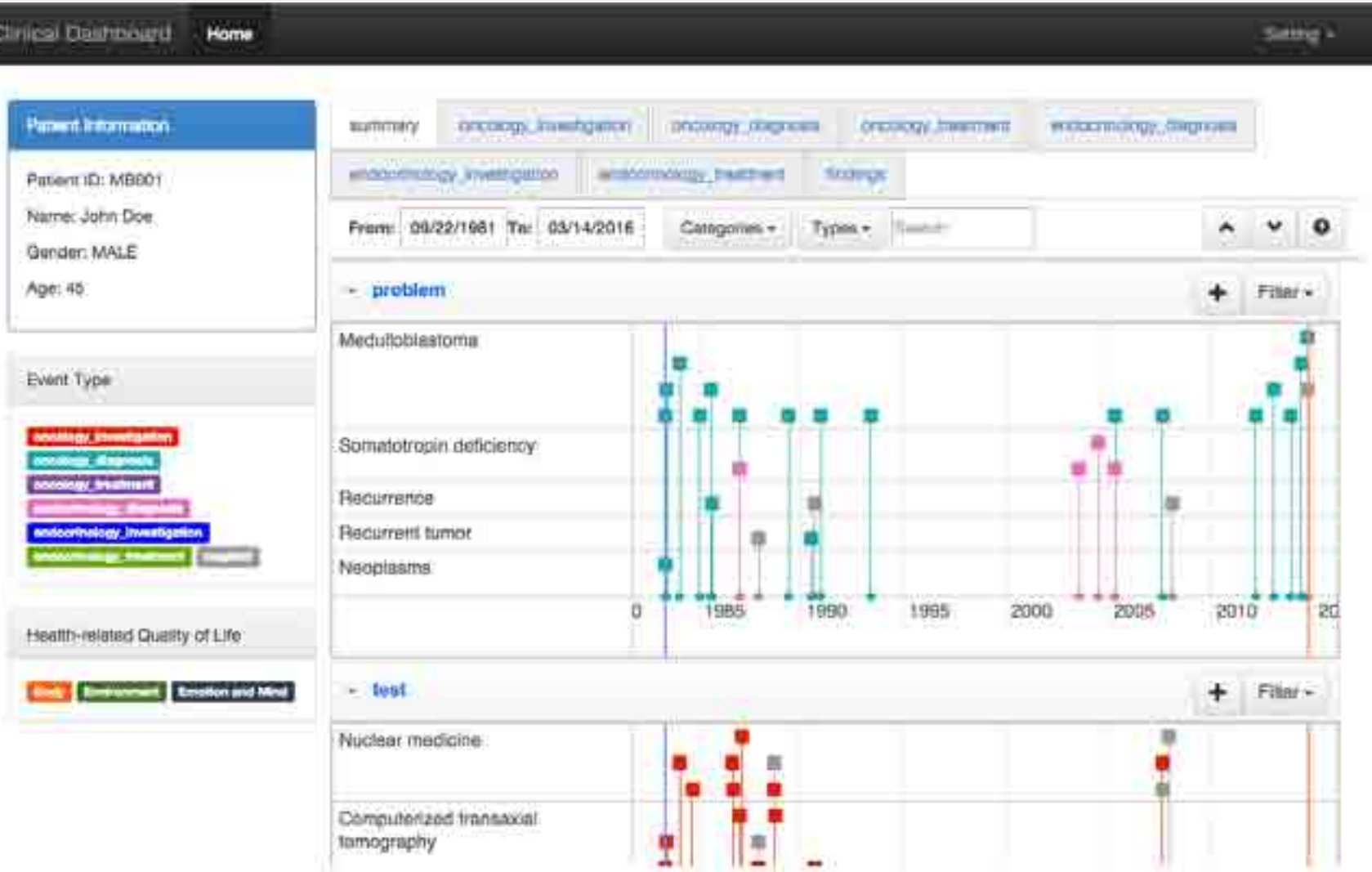
Clinical temporal expressions

ADMISSION DATE: 2011-02-06;
DISCHARGE DATE: 2011-02-08;
HISTORY OF PRESENT ILLNESS: Mr. Pohl is a 53 - year-old male with history of alcohol use and hypertension. Blood alcohol level was 383. Agitated in emergency room requiring soft leather restraints, received 5 mg of Haldol, 2 mg of Ativan. He became hypotensive in the emergency room with a systolic blood pressure in the 80's and had decreased respiratory rate. He received a normal saline bolus of 2 litres of good blood pressure response. The patient was then admitted to the medical Intensive Care Unit for observation and then transferred to our service on medicine when the blood pressures remained stable overnight...

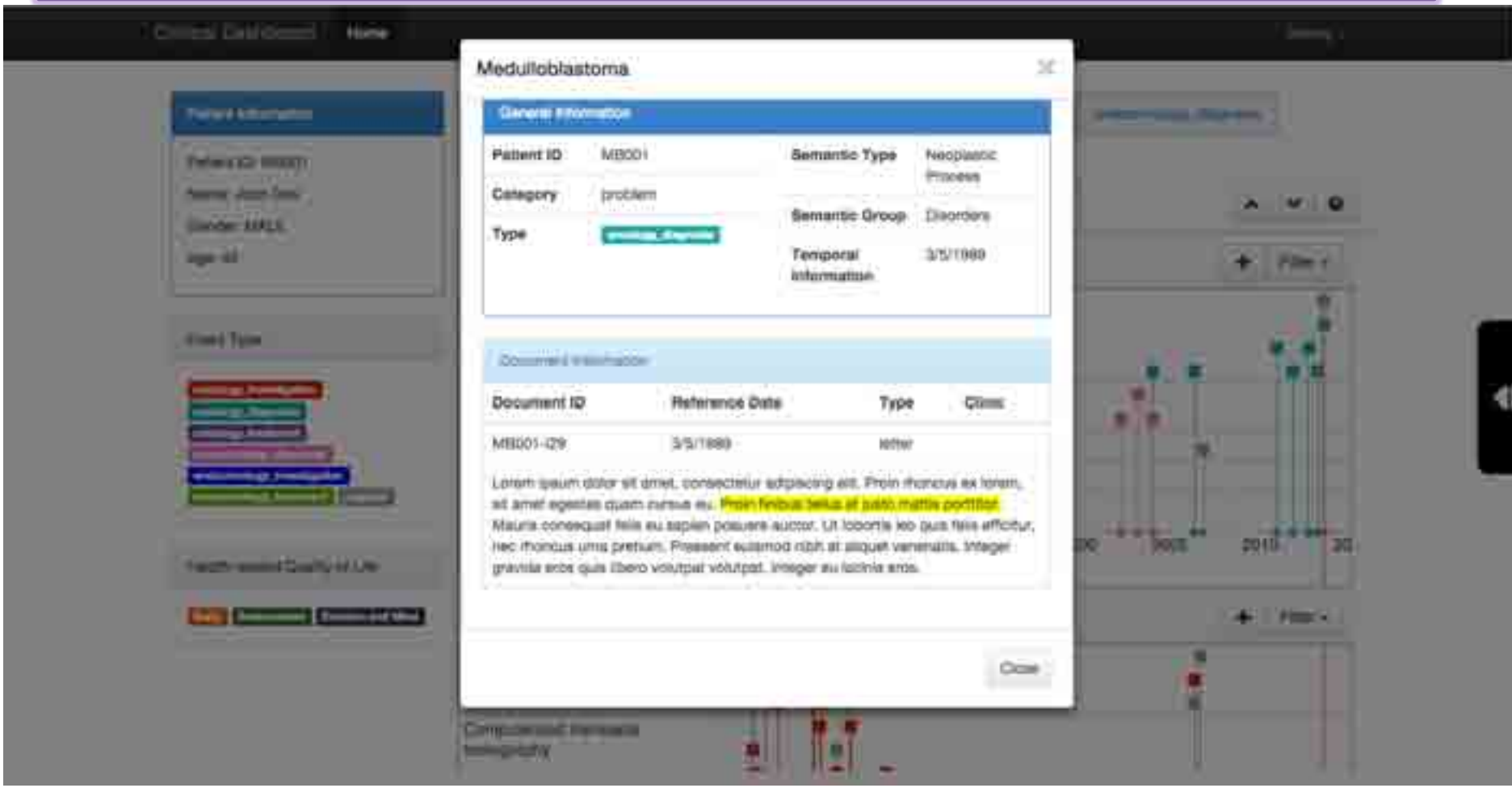


M. Filannino

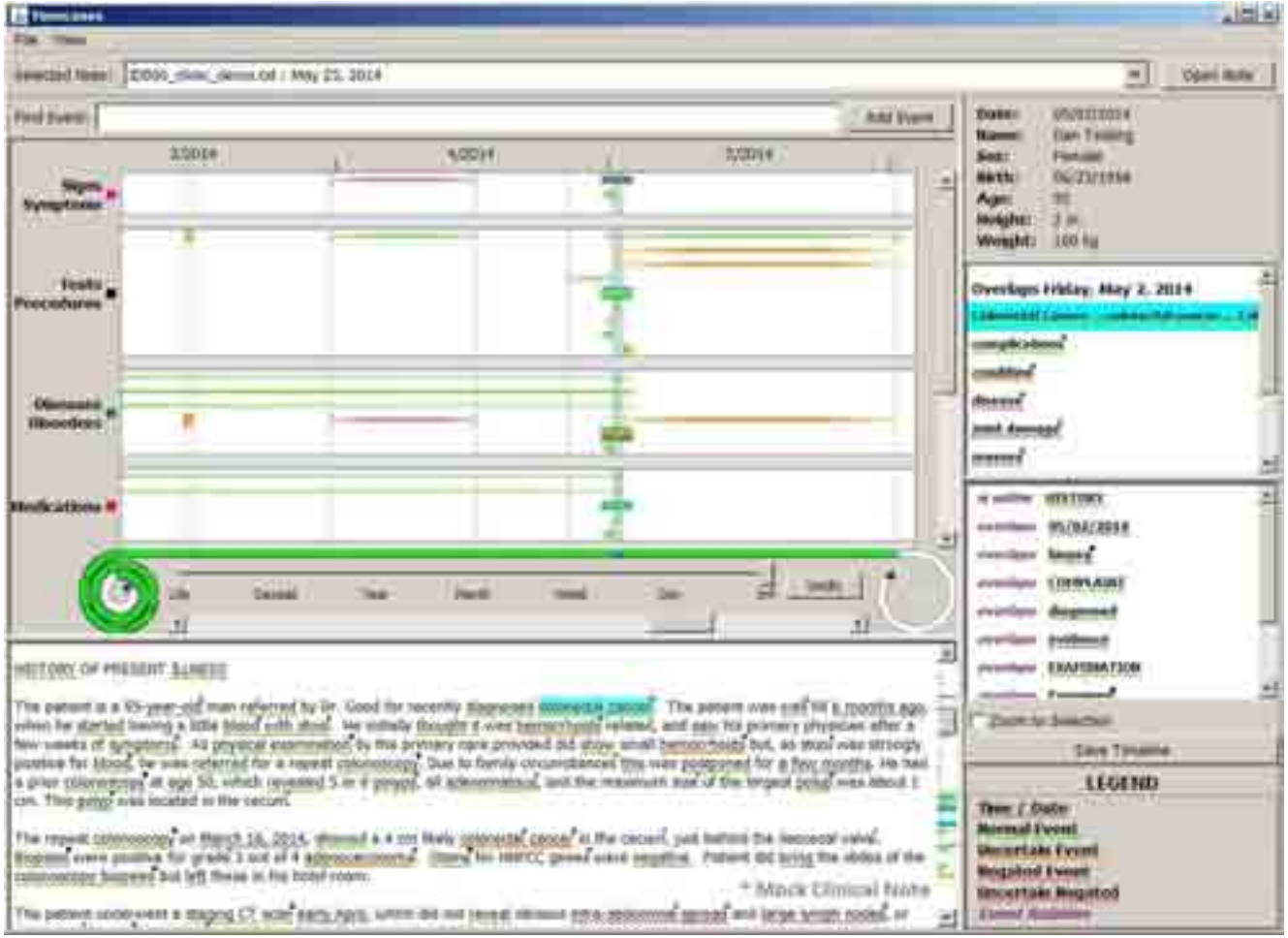
Mining Patient Journeys



Mining Patient Journeys



Patient TimeLanes @ cTAKES



https://sourceforge.net/projects/ohnlp/files/cTAKES/

Mining clinical narratives

Examples with real data

- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

De-identification of narrative

- Access to data?

HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old-woman with long standing hypertension who presented as a walk-in to me at the **Oak Valley Health Center** on **July 9th**. Recently had been started q.o.d. on Clonidine since **May 5th** to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the **Smith Cardiac Unit** for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. **Pearson** to follow.

L. Hirschman, CSHALS 2013

CSHALS MITRE

Automated de-identification

- Motivation
  - Enable data access to researcher
  - Ethical and legal requirements
  - Time and cost vs. manual labour
- Personal identifiable Information (PII)
    - Name, Date of Birth, Contact, Address, Phone number, …
    - NHS number,
    - Profession, family members, etc.
- E.g. clinical letters at the Christie hospital
  - Accuracy 96.05% - comparable to human benchmark

The Christie NHS
NHS Foundation Trust

Mining clinical narrative

Examples with real data

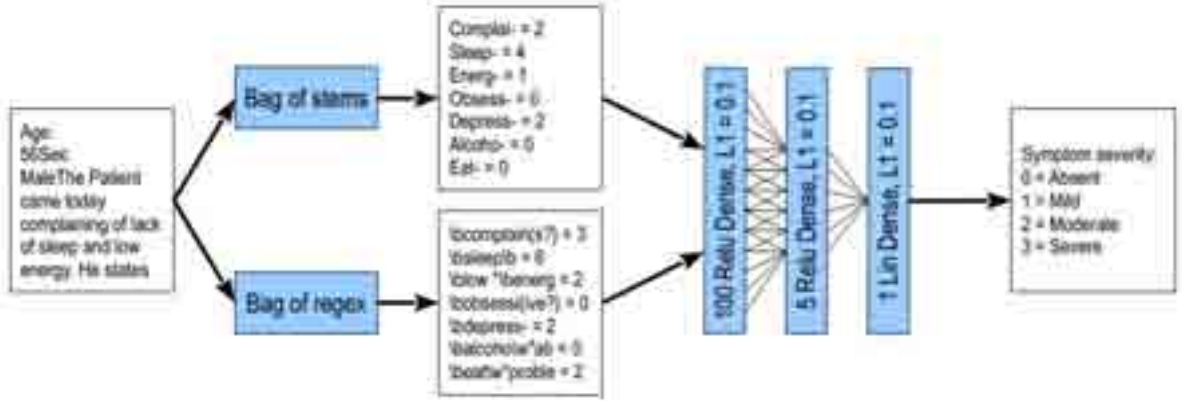
- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

Symptom severity in psychiatric evaluation notes

- Analyse patient initial evaluation notes to identify symptom severity (absent, mild, moderate, severe)
  - Does a patient requires immediate medical attention or hospitalisation?
  - Based on the Research Domain Criteria (RDoC) framework
  - Focus on: positive valence, negative valence, cognitive, social processes, arousal and regulatory system
- Input data is a mix of
  - semi-structured questions (e.g. *“History of Drug Use: Yes”*) and
  - free-text narrative (e.g. *“treatment included x2 IOP treatment, IOP for alcohol dependence”*)

Symptom severity in psychiatric evaluation notes

- Methods:
  - Knowledge-based approach (rules)
    - Identify key-phrases (e.g. substance use, manic episodes, treatments, etc.)
  - Data-driven approach (deep learning)



- Current results
  - absent: 82%; mild: 93%; moderate: 70%; severe: 77%

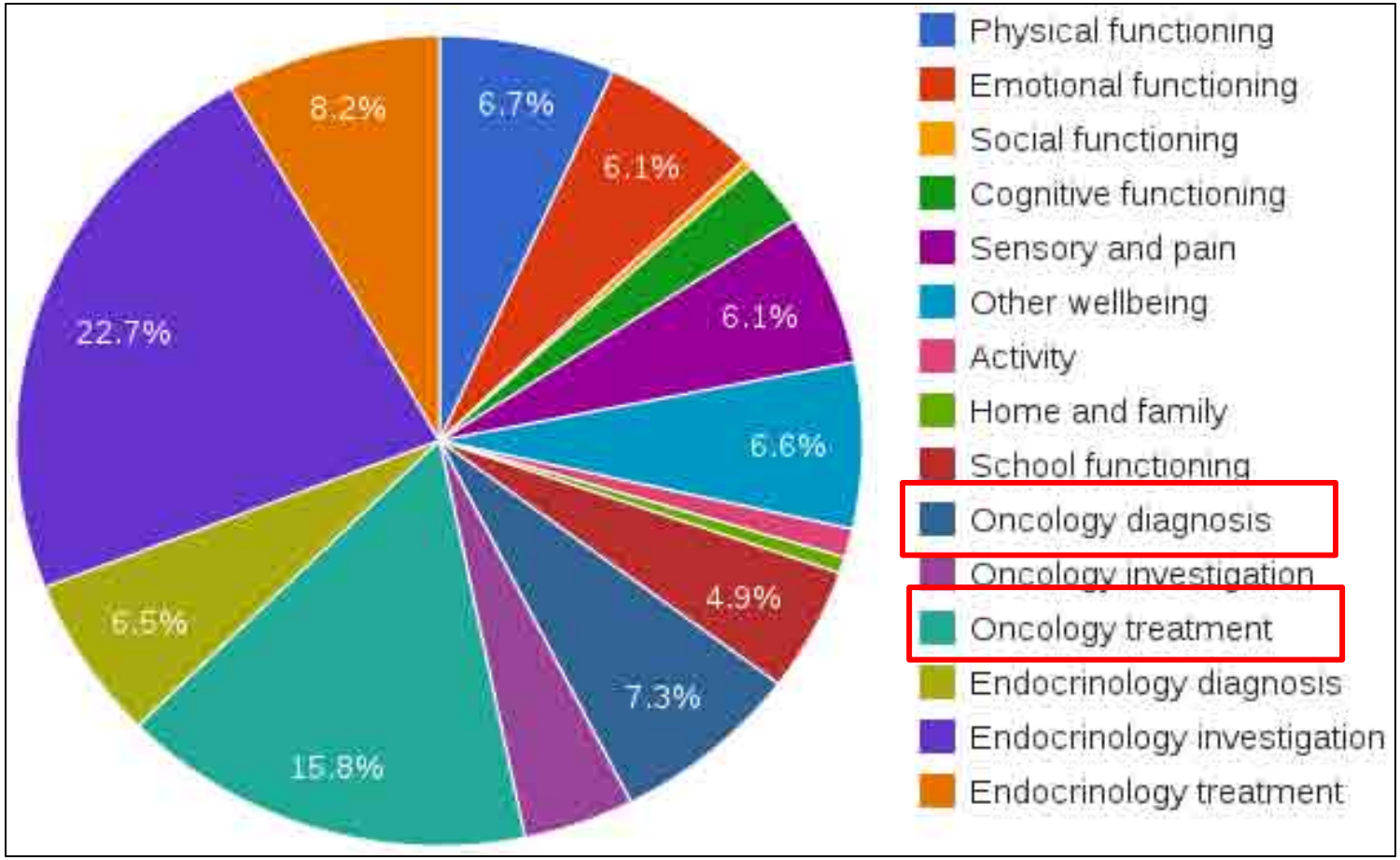
MANCHESTER 1824 The University of Manchester
OXFORD
deepCognito
Health e-Research Centre
AIHI AUSTRALIAN INSTITUTE OF HEALTH INNOVATION

Mining clinical narrative

Examples with real data

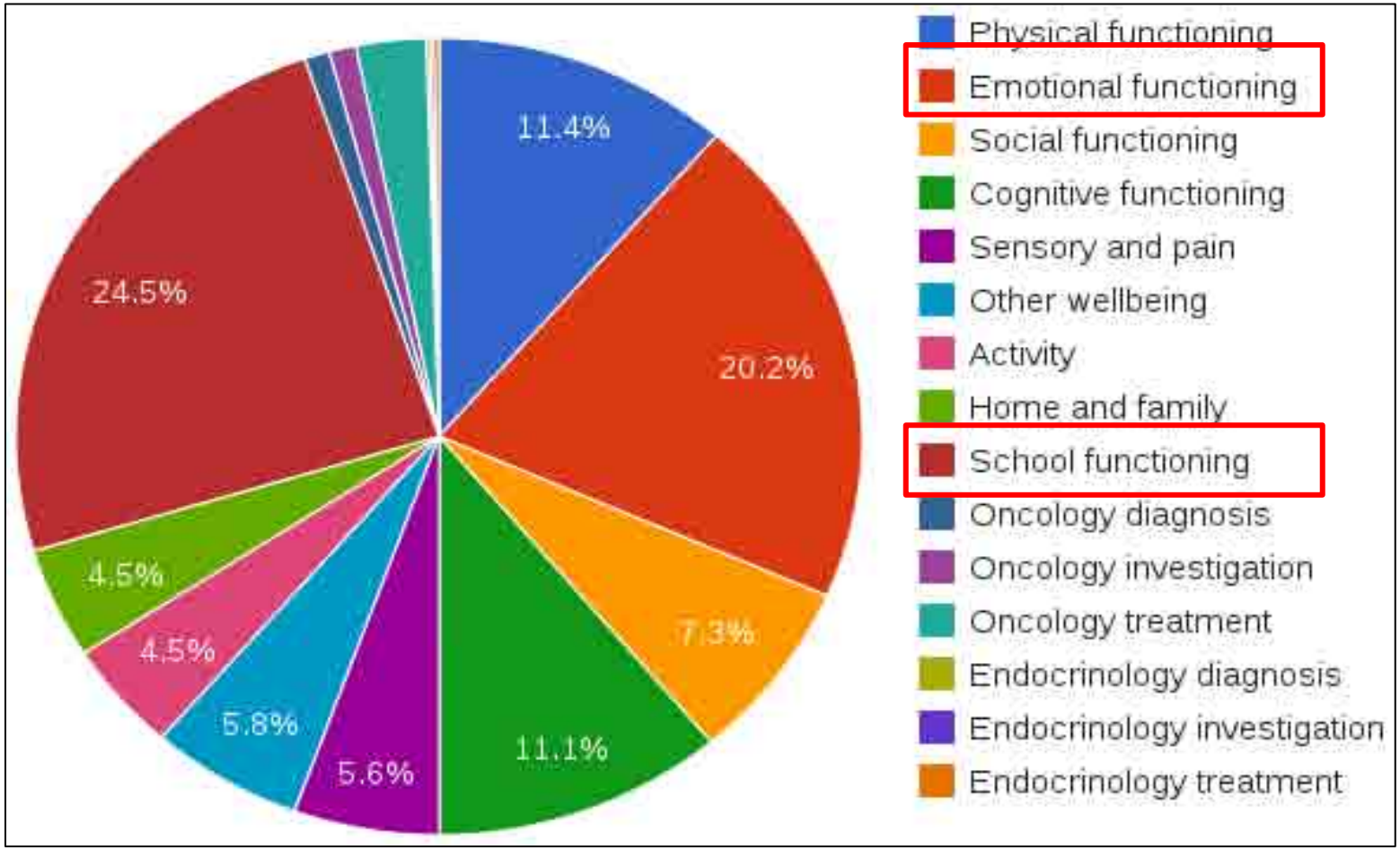
- Extracting baseline patient data
- Extracting dosages from free text prescriptions – finding drug regiments
- Mining patient journeys
- De-identification of clinical narrative
- Symptom severity in psychiatric evaluation notes
- What do EHR talk about?

What do EHR talk about?



EPSRC
The Christie NHS
NHS Foundation Trust
KidsCan
The Christie Charitable Fund
NHS National Institute for Health Research

What patients are talking about?



Healthcare Text Analytics: Demos

George Demetriou[1]

[1]School of Computer Science, University of Manchester

Plan

1. Very brief overview of text mining
2. Examples of processing healthcare free text
   - Clinical notes
   - Patient generated data
3. Hands on
   - Extract dosage information from free text
   - Extract symptoms, procedures, anatomical locations
   - Get all patient with hypertension
4. State of the art
5. Join the community

Task: Extract Dosage Information

Variability in text:

- “25 mg” -> Numeric_value (integer) + Unit
- “15.5 kg” -> Numeric_value (decimal) + Unit
- “20-25 mgs” -> Numeric_value (range) + Unit
- “6 hours” -> Time: different feature of dosage!
- What if we want to match

  e.g. ‘tblspoons’ or ‘tbsp’?