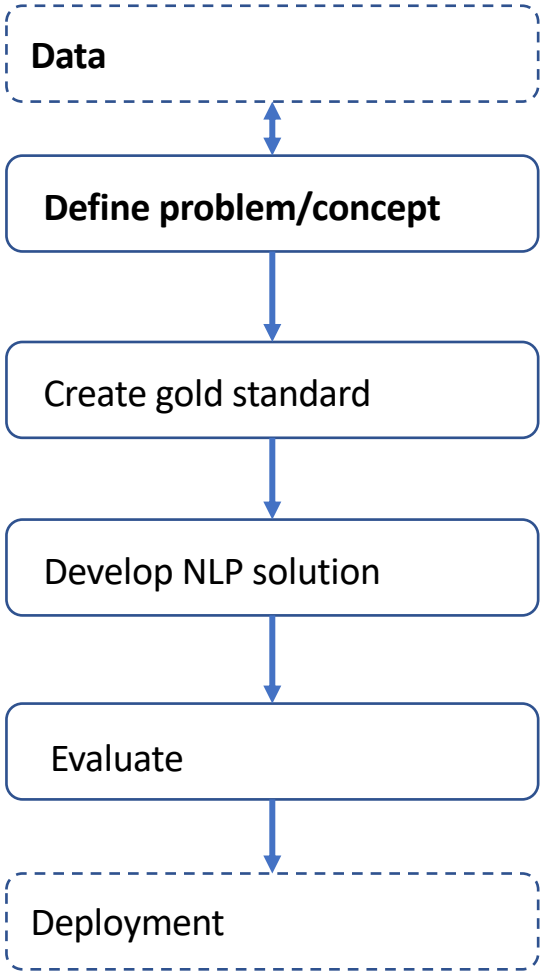
NIHR | Maudsley Biomedical Research Centre

Steps to set up a clinical NLP study

Sumithra Velupillai

CLINICAL RECORD INTERACTIVE SEARCH

Natural Language Processing - workflow



- Understand your data - What data do you have?
- How much data do you have?
- Where does the data come from?
- What do you want to model/extract?
- Research the field:
  - Are there any studies where this problem has been modeled already?
  - Are there any existing resources (lexicons, algorithms, etc) that have already modeled something similar? Can these be useful for your problem?

- Is your data appropriate for this problem?

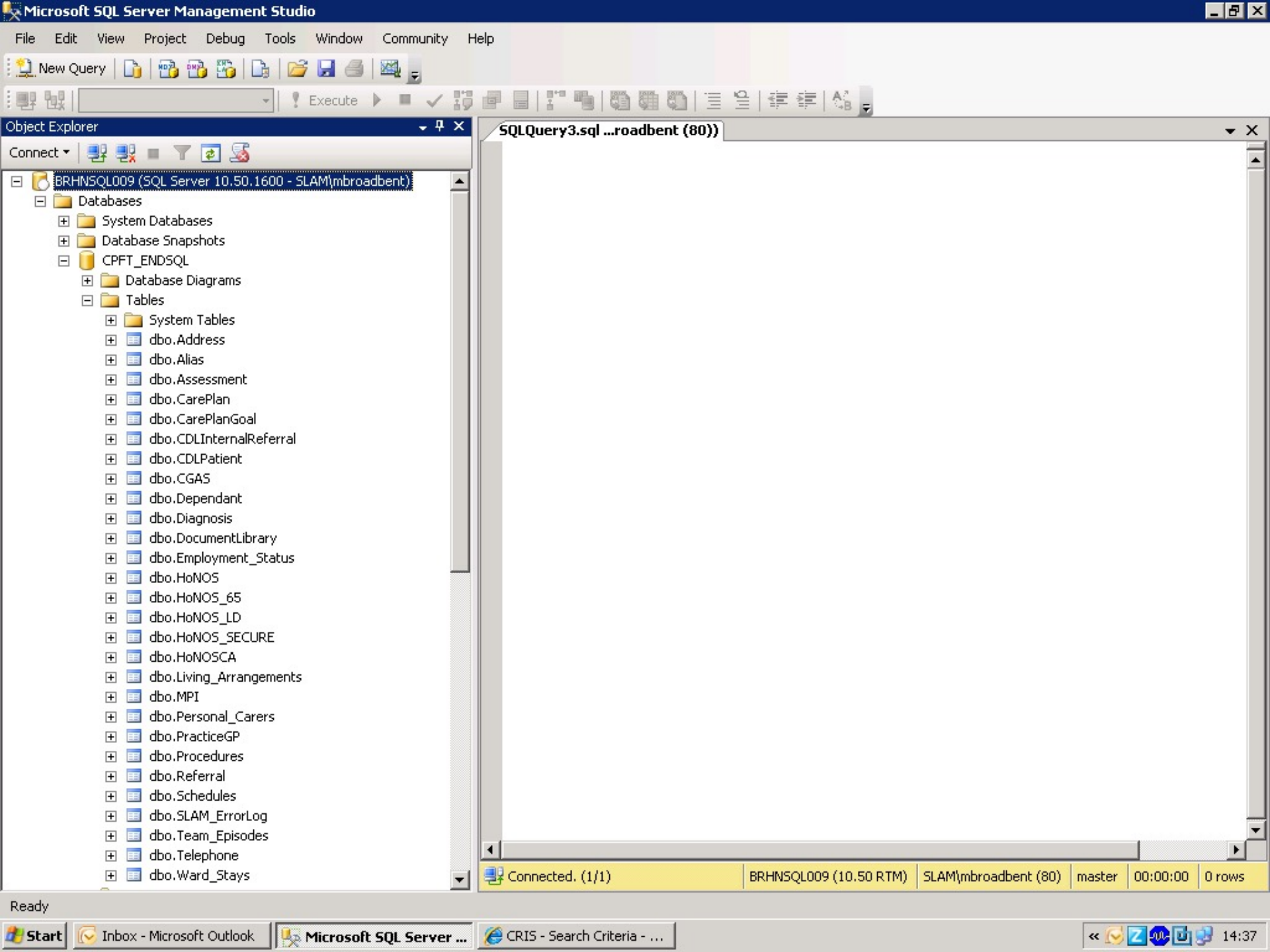
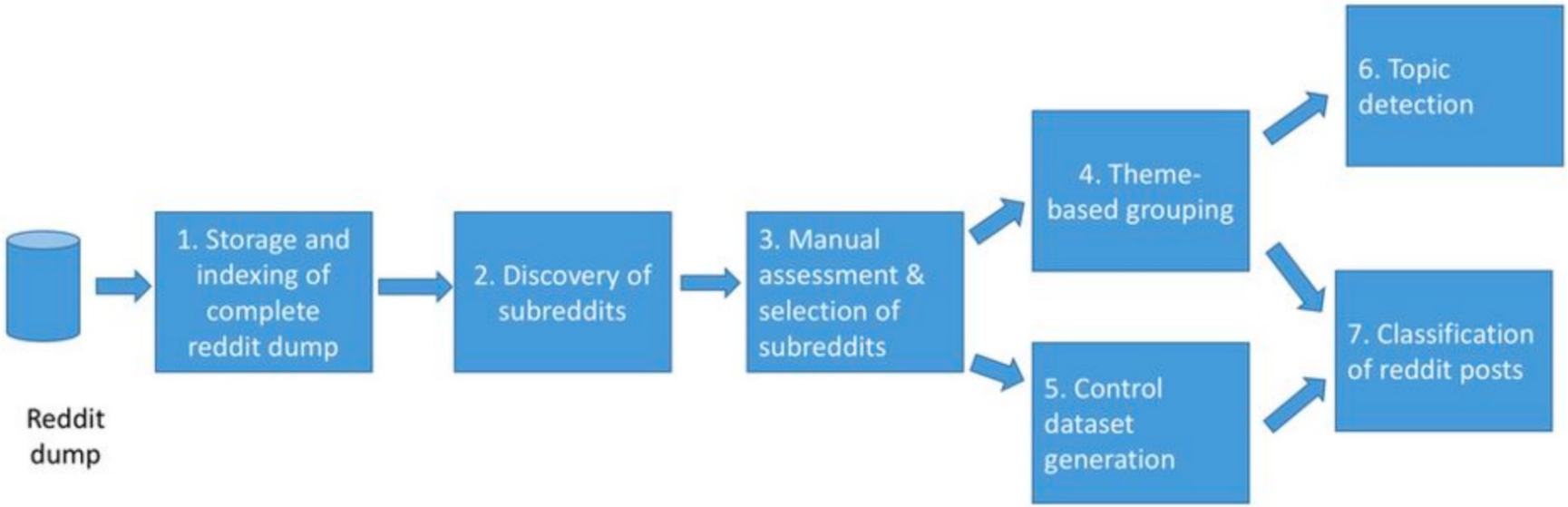

Microsoft SQL Server Management Studio
File Edit View Project Debug Tools Window Community Help
New Query
Execute
Object Explorer
Connect
BRHNSQL009 (SQL Server 10.50.1600 - SLAM\mbroadbent)
Databases
System Databases
Database Snapshots
CPFT_ENDSQL
Database Diagrams
Tables
System Tables
dbo.Address
dbo.Alias
dbo.Assessment
dbo.CarePlan
dbo.CarePlanGoal
dbo.CDLInternalReferral
dbo.CDLPatient
dbo.CGAS
dbo.Dependant
dbo.Diagnosis
dbo.DocumentLibrary
dbo.Employment_Status
dbo.HoNOS
dbo.HoNOS_65
dbo.HoNOS_LD
dbo.HoNOS_SECURE
dbo.HoNOSCA
dbo.Living_Arrangements
dbo.MPI
dbo.Personal_Carers
dbo.PracticeGP
dbo.Procedures
dbo.Referral
dbo.Schedules
dbo.SLAM_ErrorLog
dbo.Team_Episodes
dbo.Telephone
dbo.Ward_Stays
SQLQuery3.sql ...roadbent (80))
Connected. (1/1)
BRHNSQL009 (10.50 RTM)
SLAM\mbroadbent (80)
master
00:00:00
0 rows
Ready
Start
Inbox - Microsoft Outlook
Microsoft SQL Server ...
CRIS - Search Criteria - ...
14:37

Figure 1 : Overall workflow of our approach.

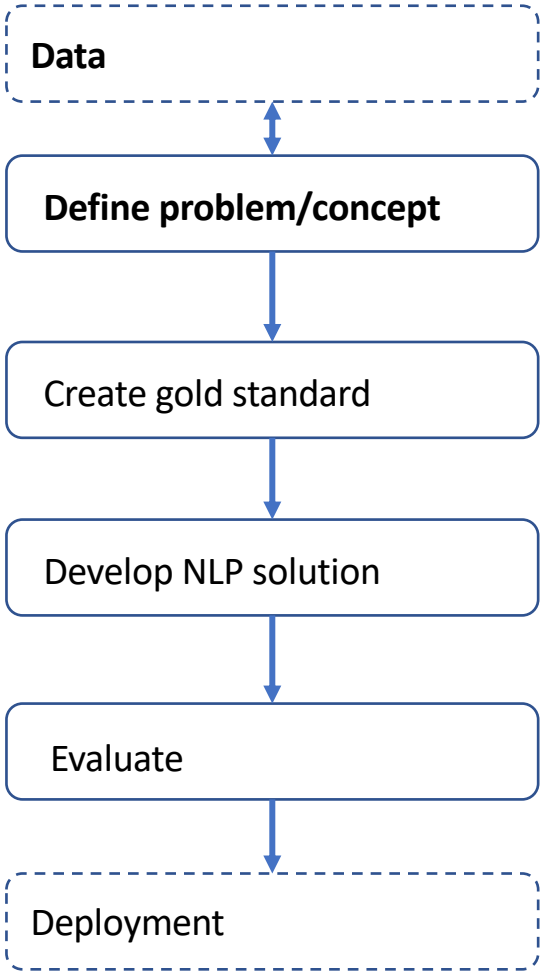
From: Characterisation of mental health conditions in social media using Informed Deep Learning



Back to article page »

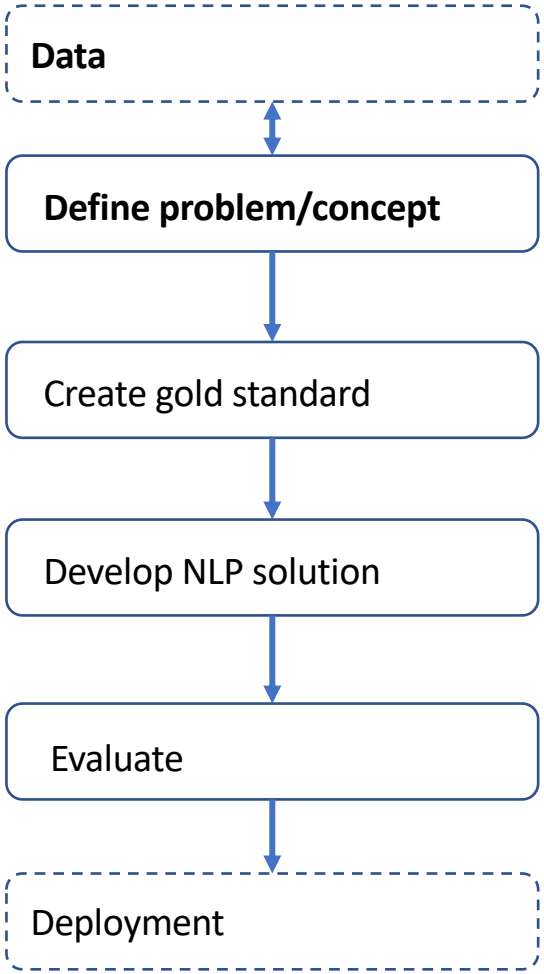
Gkotsis et al. Characterisation of mental health conditions in social media using Informed Deep Learning.
Scientific Reports **volume 7**, Article number: 45141 (2017)

Natural Language Processing - workflow



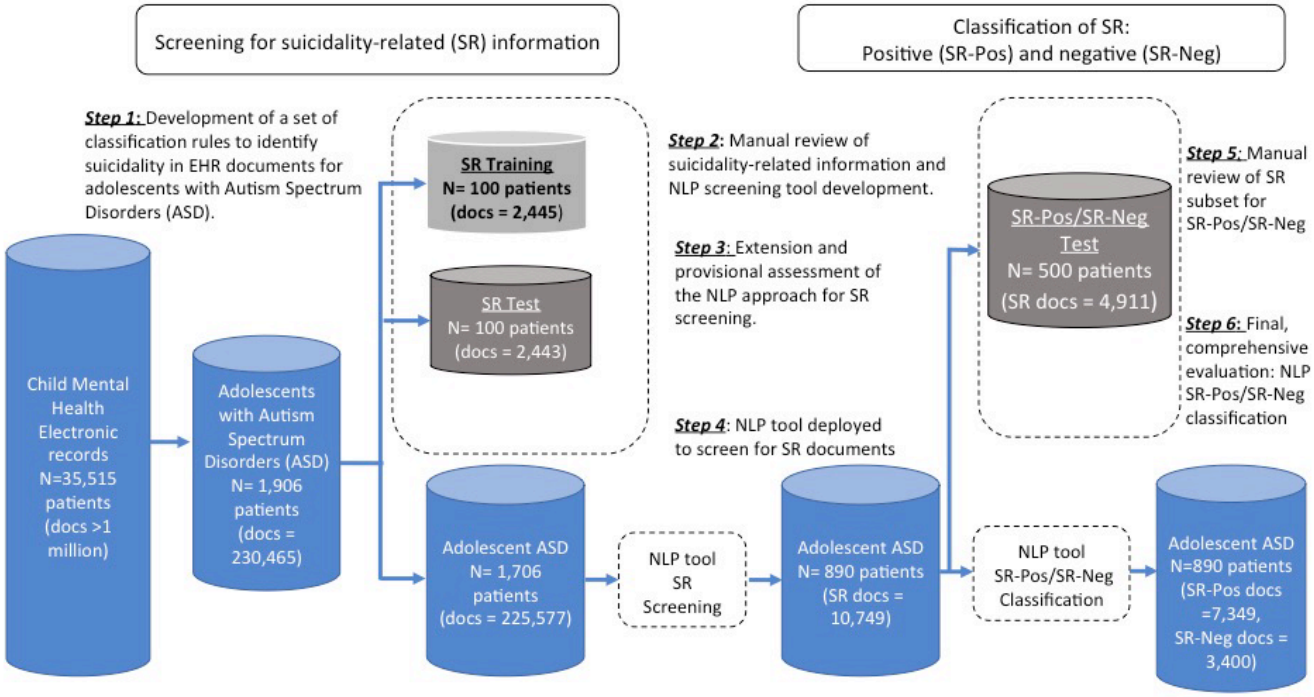
- Define the problem/concept
  - Relate to existing definitions (if any)
- Determine what unit you need to model this problem
  - Patient? Document? Phrase? Something else?
- Is NLP needed for all, or parts of the problem?
- Construct your dataset based on your requirements
  - Look at distributions, characterise the data

Natural Language Processing - workflow

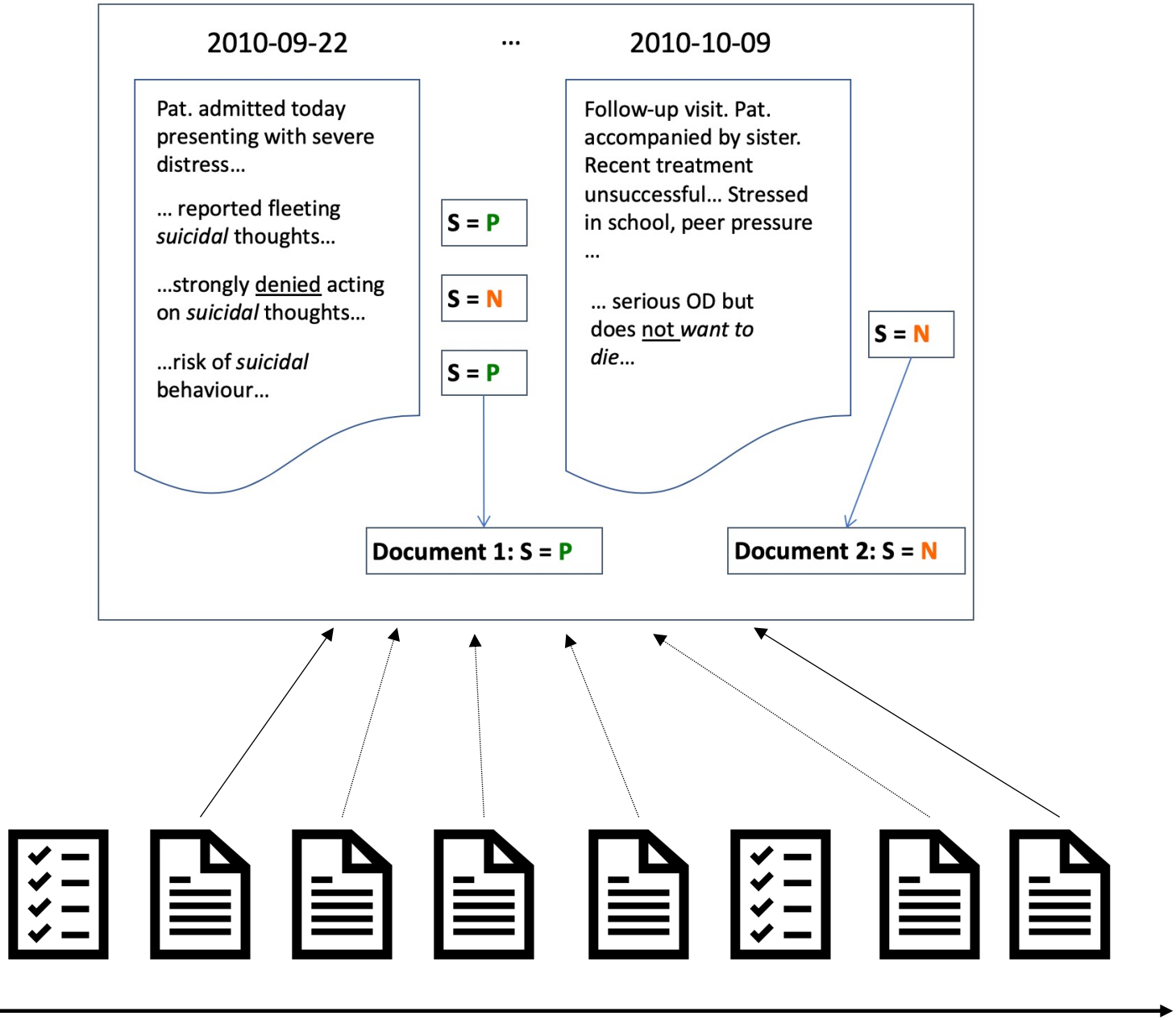


- If you need to use NLP for your use-case:
  - Where do you expect to find the information (all documents, some documents, some parts of some documents)?
  - How many documents?
  - How long are the documents?
- If your use-case also relies on other data (e.g. structured data from other database tables), is this information necessary to take into account when building your dataset?

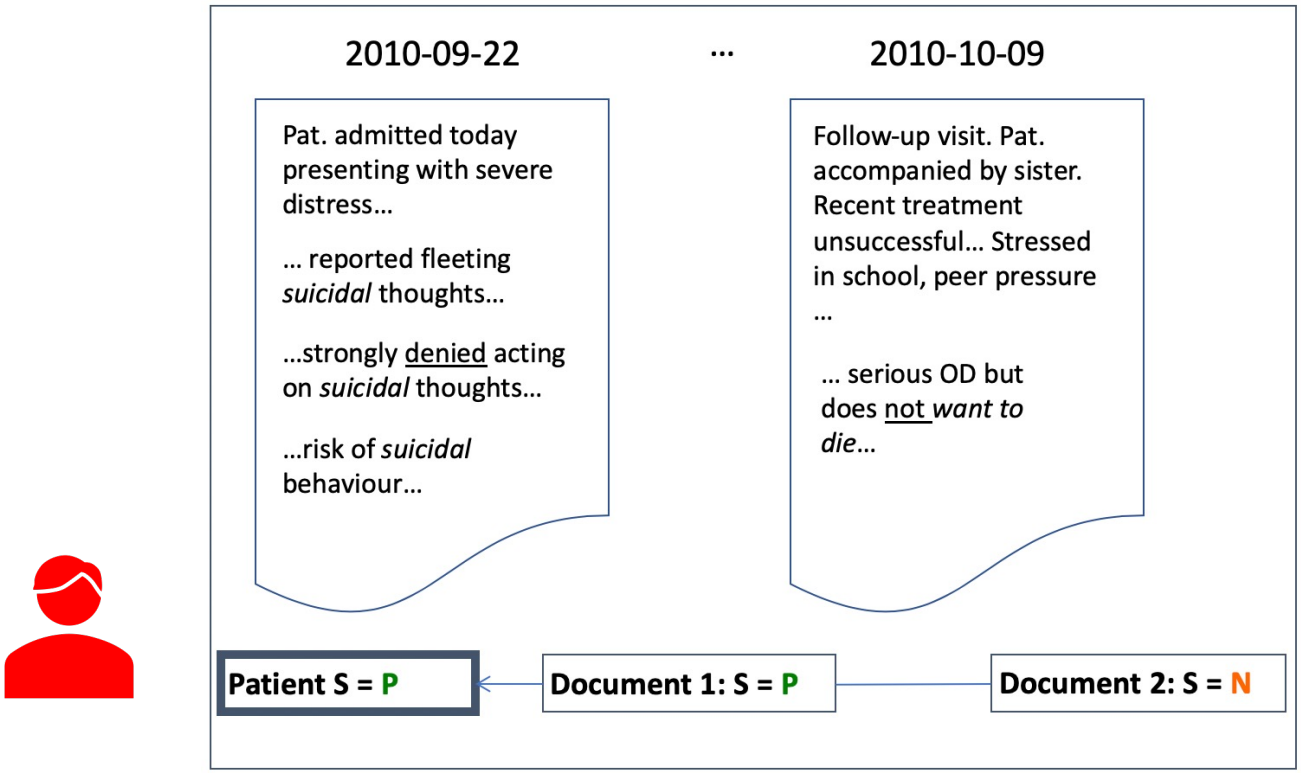
- Consult with domain experts!



Downs et al., 2017. Detection of Suicidality in Adolescents with Autism Spectrum Disorders: Developing a Natural Language Processing Approach for Use in Electronic Health Records. In AMIA 2017 Proceedings.

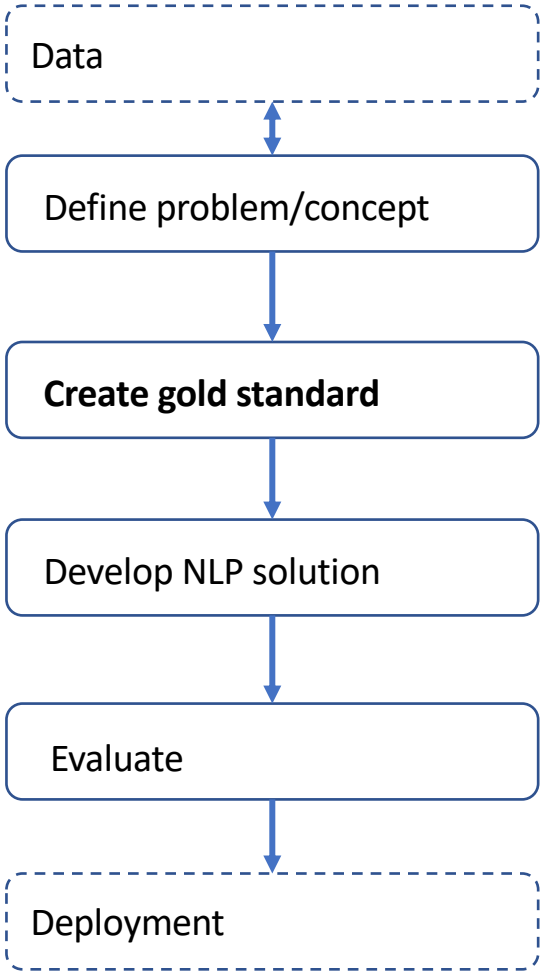


Velupillai S, Epstein S, Bittar A, et al. Identifying Suicidal Adolescents from Mental Health Records Using Natural Language Processing. Stud Health Technol Inform. 2019 Aug 21;264:413-417. doi: 10.3233/SHTI190254.

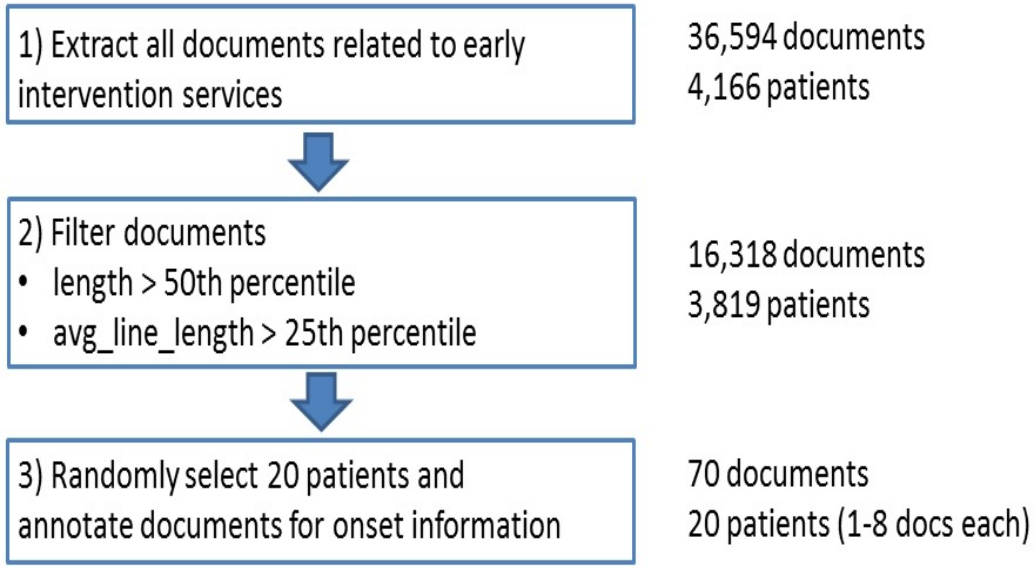


Velupillai S, Epstein S, Bittar A, et al. Identifying Suicidal Adolescents from Mental Health Records Using Natural Language Processing. Stud Health Technol Inform. 2019 Aug 21;264:413-417. doi: 10.3233/SHTI190254.

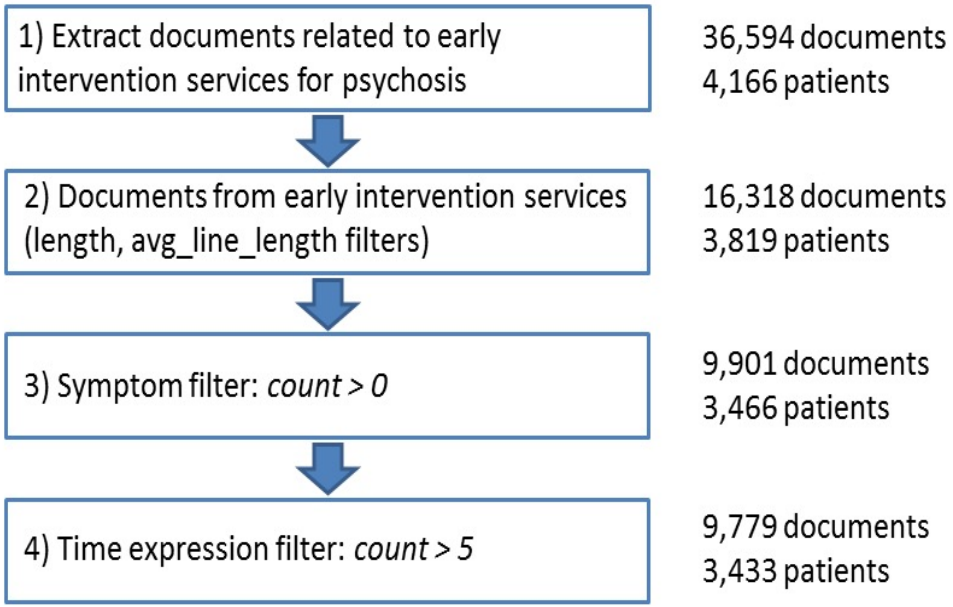
Natural Language Processing - workflow



- Random sample for initial review
  - Worth getting a sense of the potential complexity of the task
- Apply an off-the-shelf tool and look at distributions
  - Maybe pre-annotations can be used?
- Annotation
  - Randomly sample the entities you need (all documents for one patient? Random documents? Random sentences?)
    - Extract a dataset for a first annotation round, initiate guideline development
    - Iterate until you've reached acceptable agreement and guidelines are clear.



Viani et al. 2019. Annotating Temporal Relations to Determine the Onset of Psychosis Symptoms. Medinfo 2019.



Viani et al. 2019. Temporal Information Extraction from Mental Health Records to Identify Duration of Untreated Psychosis. Under review

Developing a gold/reference standard corpus: Annotation

- Gold/Reference standard – why?
  - How can manual annotations inform NLP development? (training & development)
  - How do we know if our NLP system does what we expect it to do? (evaluation)
  - How hard/easy is the task (annotator agreement, upper performance bound)

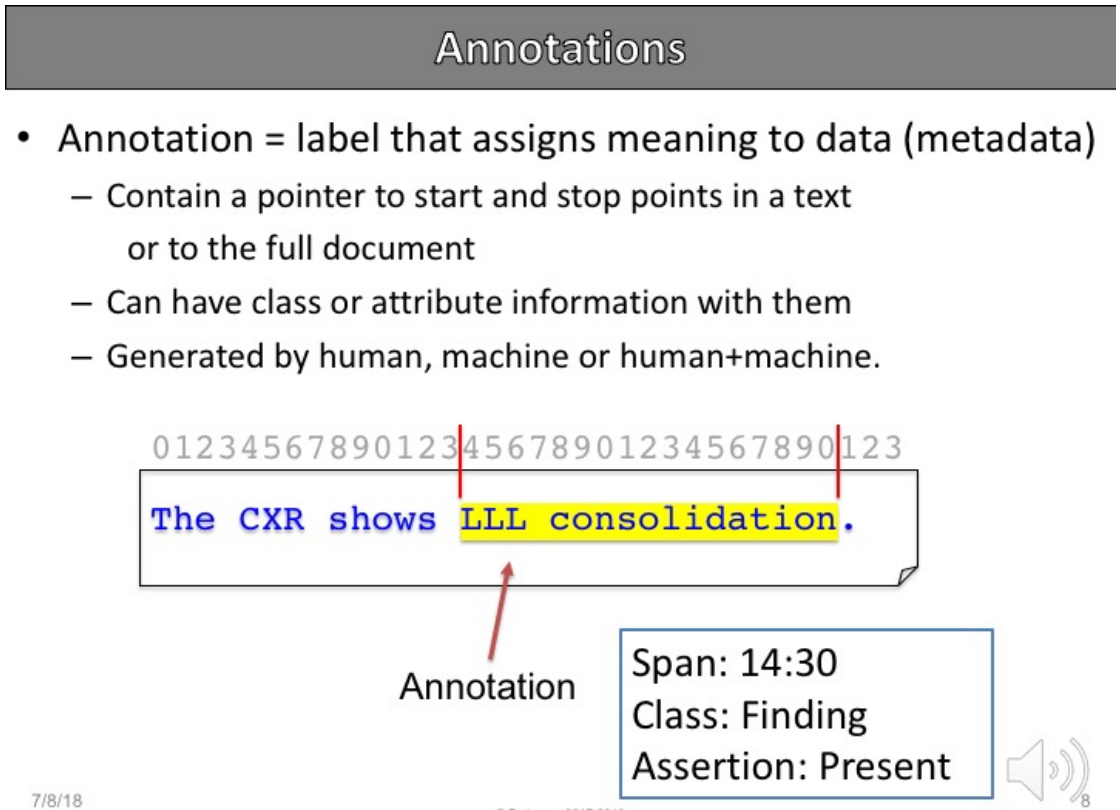
Developing a gold standard corpus: Annotation

- Don’t underestimate this step!
  - “analytics tasks are often talked about as being **80%** data preparation”*
  - Example: Penn TreeBank – widely used for NLP development
    - **8** years of operation (1989-1996)
    - 7 million words of PoS-tagged text, and more.
      - Taylor A., Marcus M., Santorini B. (2003) The Penn Treebank: An Overview. In: Abeillé A. (eds) Treebanks. Text, Speech and Language Technology, vol 20. Springer, Dordrecht

*https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html

Developing a gold standard corpus: Annotation

- What is annotation?



Acknowledgements: Uni. of Utah DeCART summer school: https://github.com/jianlins/AnnotationNLP/blob/master/01_Introduction_to_Annotation.ipynb

Annotation guidelines / Define concept

- Markables – elements to be annotated
  - Annotation type = label to be assigned to a segment of text
  - Relationship = link between instances of annotations
  - Attributes = features of annotation types and relationships

Acknowledgements: Uni. of Utah DeCART summer school: https://github.com/jianlins/AnnotationNLP/blob/master/01_Introduction_to_Annotation.ipynb