Challenges in representing *Rich* Data and Annotations

PCREF Workshop, 2020

Sameer S. Pradhan
Linguistic Data Consortium
University of Pennsylvania

Deep Learning Revolution

NLP—No longer *only* Language

- More accurate ***computer vision*** and ***speech recognition*** models
- Not just written language, but ***Multimodal*** understanding
- Representing data is already challenging
- Representing multi-***modal***, multi-***layered*** metadata (annotations, in our case) which remains *in sync* with the data and ***maintain consistency within and across layers*** can be quite ***challenging***

Underlying Assumption…

Metadata in the form of ***annotations***—
Linguistic or *otherwise*—***play an important role***
in the underlying research

An Example Scenario… Only Written Text

Robustness to Tokenization

A Grant is (*finally*) Funded!

- **Phase I**
  - Use existing Treebank-ed text (**=** *use existing **trees** and **tokens***)
  - Add a few layers of *rich* annotation
    - Word Sense (depends *only* on tokens)
    - Named Entities (depends *only* on tokens)
    - Propositions (depends on Tree structure)
    - Coreference (depends on Tree structure)
- **Phase II**
  - It is found to be very important to make ***minor changes*** to…
    - Treebank and PropBank **layer guidelines** so they more are *in sync*
    - A minor change in **tokenization** to *split* on *some hyphens*

As a result, some tokens are split into multiple tokens

New-York-based (single token)

↓

New-York	(first token)
-	(second token)
based	(third token)

Now, ***Update*** existing Annotations

Easy, Right?

well, Not ***Necessarily***

Factors determining the Difficulty

- How the annotation layers are ***represented***?
- How tight is the ***data coupling*** *between* ***the layers***?
- How detailed are the ***specifications***?

  a. ***within*** each layer

  b. ***between*** the layers

- Depending on the answers to the above questions (and maybe a few more)
  - It might be a ***nightmarish scenario***, or
  - It might be a ***reasonable task***
- ***Both*** options will very likely require ***human intervention*** (*annotator*)
- The ***degree of that intervention*** and the ***complexity of the task*** will be determined to a large degree by the ***above design decisions***

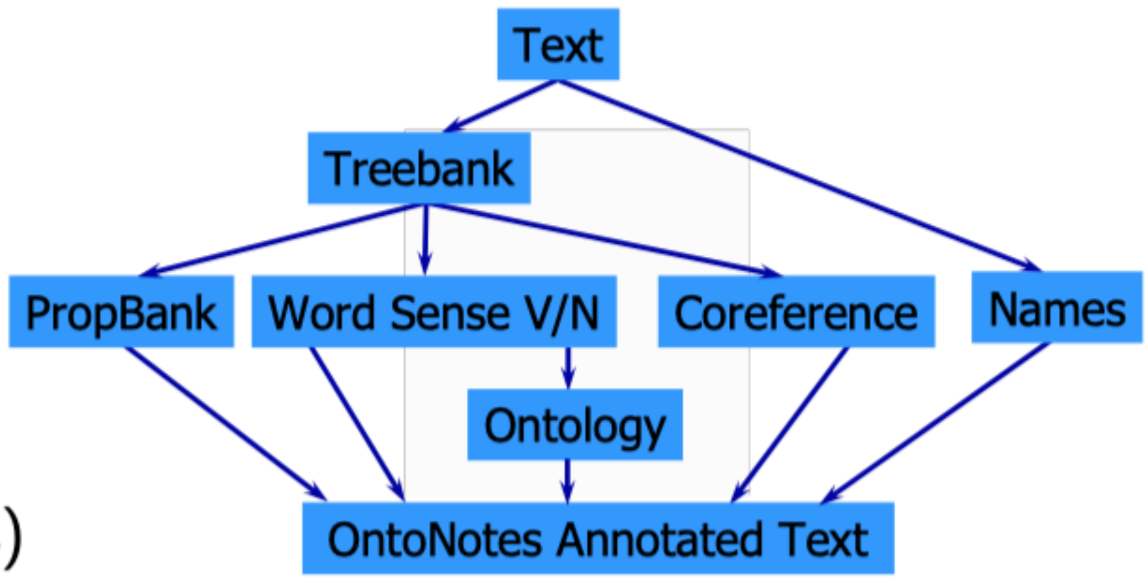
This was *not* a Hypothetical scenario

It *happened* in the OntoNotes project

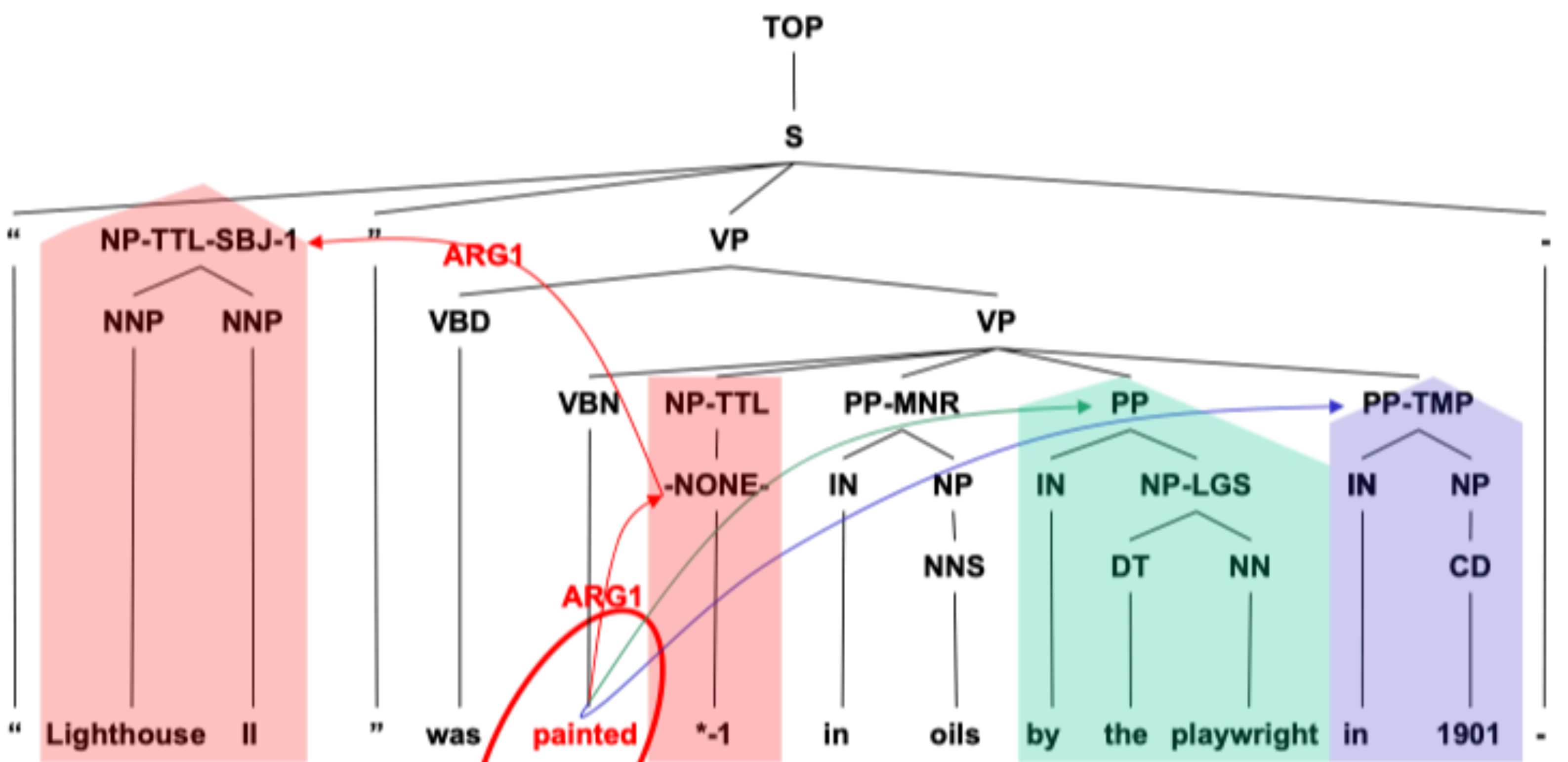
- Owing to the design of the underlying representation, it was...
  - a ***reasonable task***
    - Each layer had a ***detailed specification***
    - The layers—both *inter*- and *intra*- used a ***relational data model***
    - The layers were ***not too tightly coupled***

Multiple-Layers in OntoNotes

- Multiple layers of annotation
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
  - Ontology
- Multilingual resource
  - English (∼1M words)
  - Chinese (∼1M words)
  - Arabic (∼1M words)
- Parallel Data



Interpreting Tree Pointers



wsj/00/wsj_0037.mrg 67 5 gold set.02 ----- 0:2-ARG0 5:0-rel 6:1-ARG1 10:2-ARGM-TMP
wsj/00/wsj_0037.mrg 68 5 gold paint.01 ----- 5:0-rel 1:1*6:0-ARG1 8:1-ARG2-in 10:1-ARG0-by 12:1-ARGM-TMP
wsj/00/wsj_0037.mrg 69 21 gold exchange.01 ----- 17:2-ARG0 21:0-rel 22:1-ARG1 23:1-ARGM-TMP
wsj/00/wsj_0037.mrg 69 35 gold say.01 ----- 31:1-ARG0 35:0-rel 0:2*37:0-ARG1

Propbank Frames

```
wsj_0037.mrg 67 5 gold set.02 ----- 0:2-ARG0 5:0-rel 6:1-ARG1 10:2-ARGM-TMP
wsj_0037.mrg 68 5 paint.01  5:0-rel 1:1*6:0-ARG1 8:1-ARG2-in 10:1-ARG0 12:1-ARGM-TMP
wsj_0037.mrg 69 21 gold exchange.01 ----- 17:2-ARG0 21:0-rel 22:1-ARG1 23:1-ARGM-TMP
wsj_0037.mrg 69 35 gold say.01 ----- 31:1-ARG0 35:0-rel 0:2*37:0-ARG1
```

```
<!DOCTYPE frameset SYSTEM "frameset.dtd">
<frameset>
<predicate lemma="paint">
<note>
Frames file for 'paint' based on sentences in wsj and automatic expansion via verbnet.
</note>

<roleset id="paint.01" name="put paint on a surface" vncls="25.1">
<roles>
<role descr="agent, painter" n="0"> <vnrole vncls="25.1" vntheta="Agent"/></role>
<role descr="surface" n="1"><vnrole vncls="25.1" vntheta="Destination"/></role>
<role descr="explicit mention of paint" n="2> <vnrole vncls="25.1"
vntheta="Theme"/> </role>
</roles>
```

Word Sense Inventories

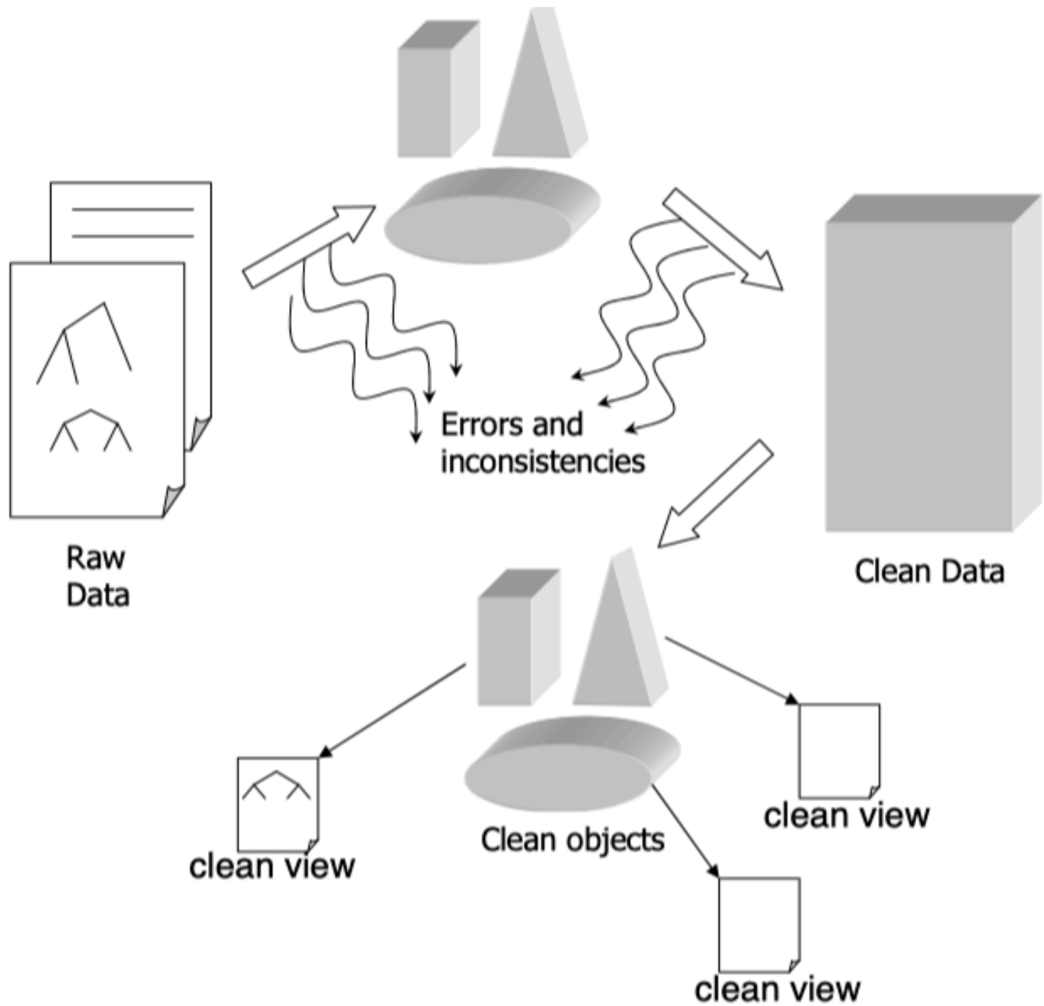
PRO Judging from the Americana in Haruki Murakami 's " A Wild Sheep Chase " (Kodansha, 320 pages, $18.95 *U*) , baby boomers on both sides of the Pacific have a lot in common .

wsj/00/wsj_0037.mrg	0	1	judge-v	2
wsj/00/wsj_0037.mrg	0	36	lot-n	1

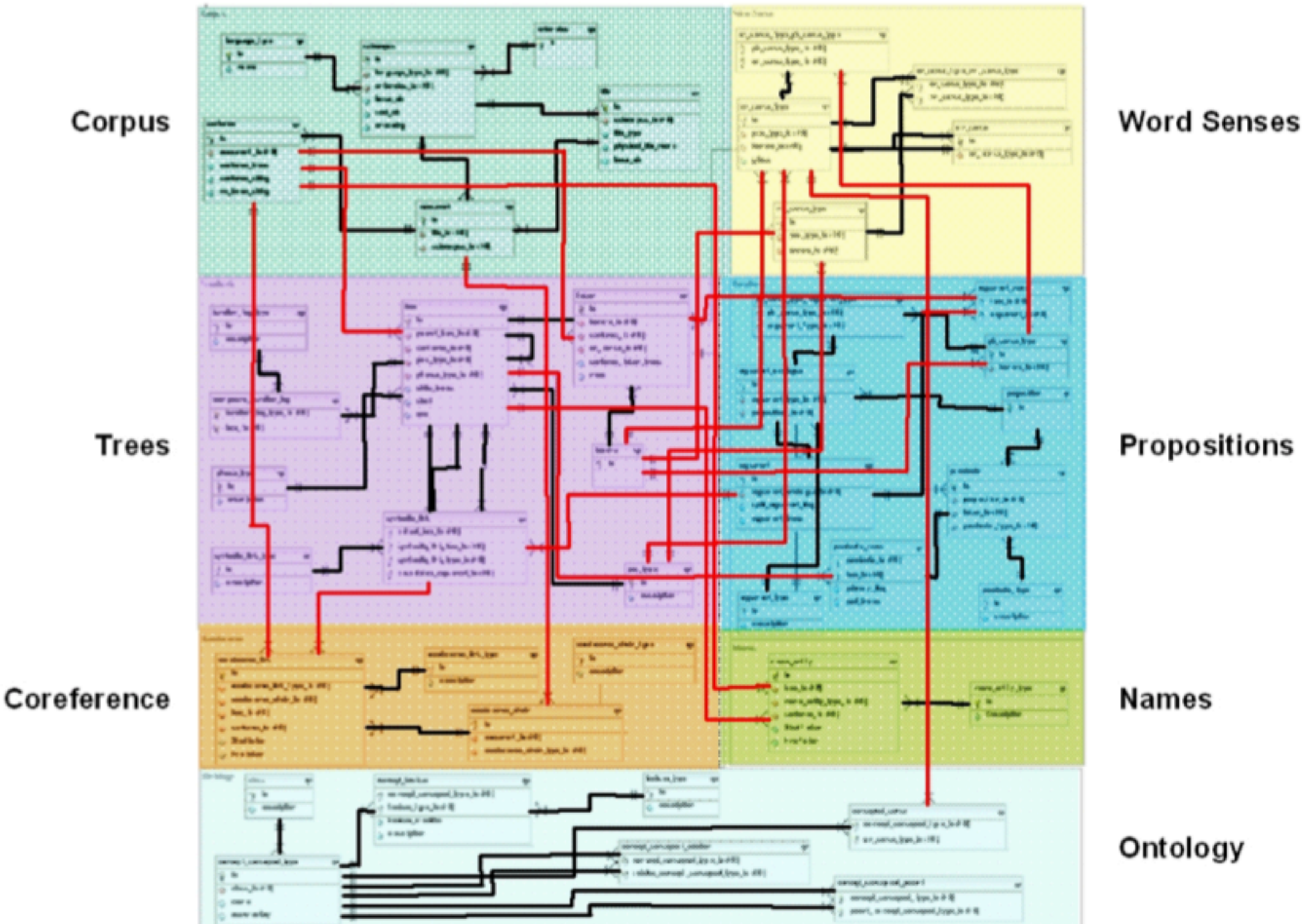
Sense Number

```
<?xml version="1.0" ?>
<!DOCTYPE inventory SYSTEM "inventory.dtd">
<inventory lemma="judge-v">
<sense group="1" n="1" name="act as an official judge>
<examples> She was asked to judge the fancy-dress competition. </examples>
<mappings>  <wn version="2.1">1,5</wn>  <pb>judge.01</pb>  </mappings
</sense>
<sense group="1" n="2" name="form an opinion, or conclusion>
<examples> They quickly judged him unfit to join the team. </examples>
<mappings> <wn version="2.1">2,3,4</wn>  <pb>judge.01</pb> </mappings>
</sense>
</inventory>
```

Annotation Lifecycle

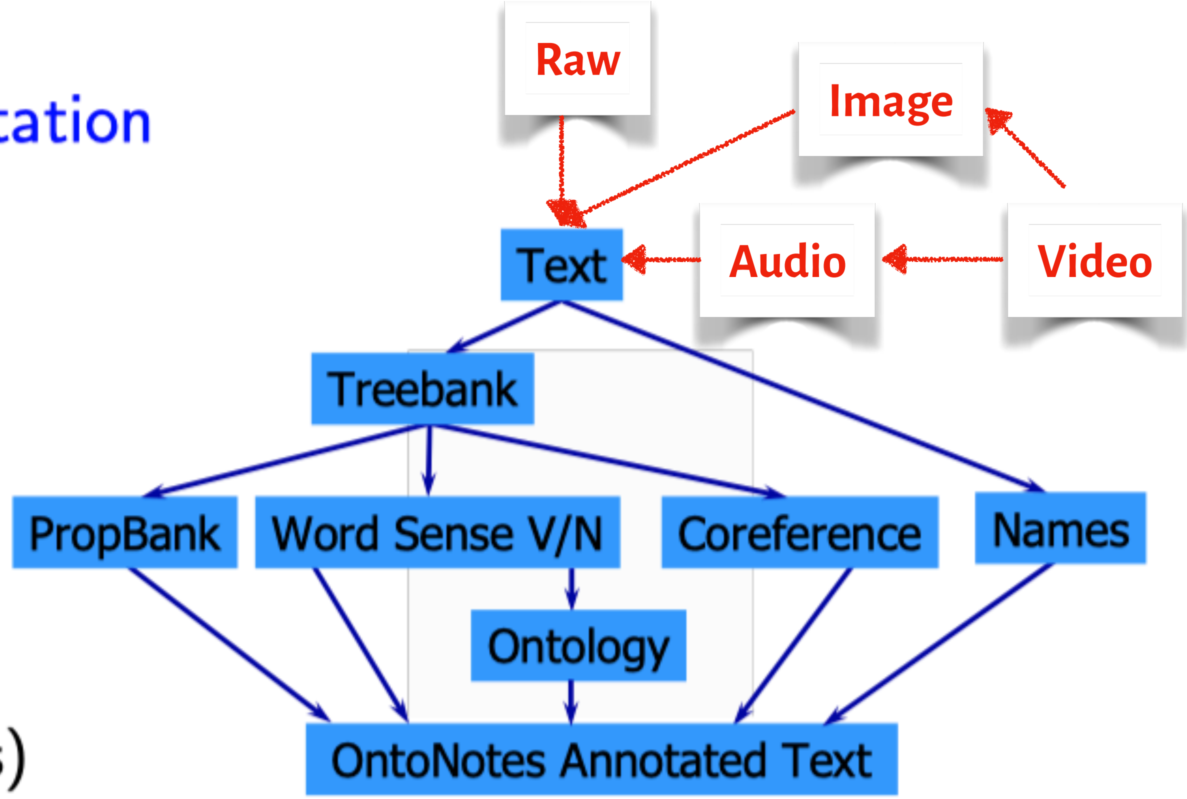


Entity-Relationship Diagram



Multiple-Layers in OntoNotes

- Multiple layers of annotation
  - Syntax
  - Propositions
  - Word sense
  - Coreference
  - Names
  - Ontology
- Multilingual resource
  - English (~1M words)
  - Chinese (~1M words)
  - Arabic (~1M words)
- Parallel Data

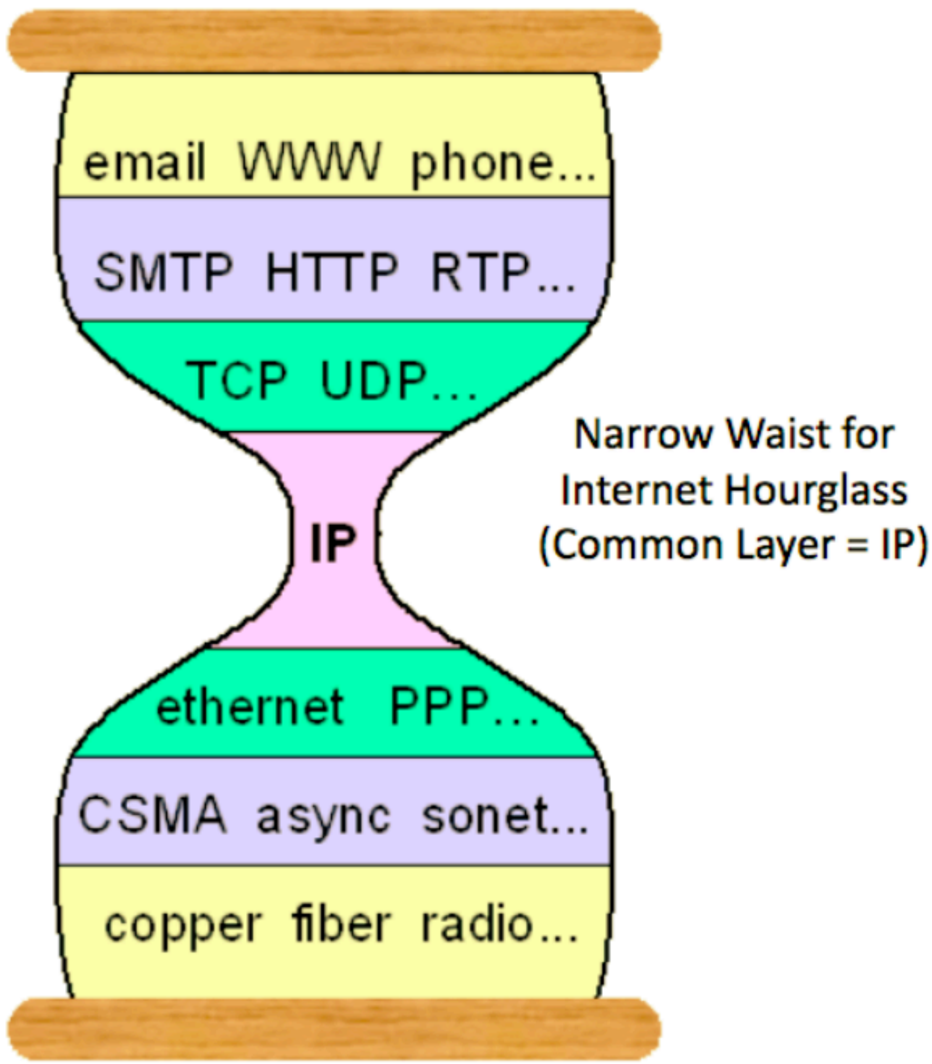


But isn't SQL last-century?

But don't take our word for it. Take Google's.

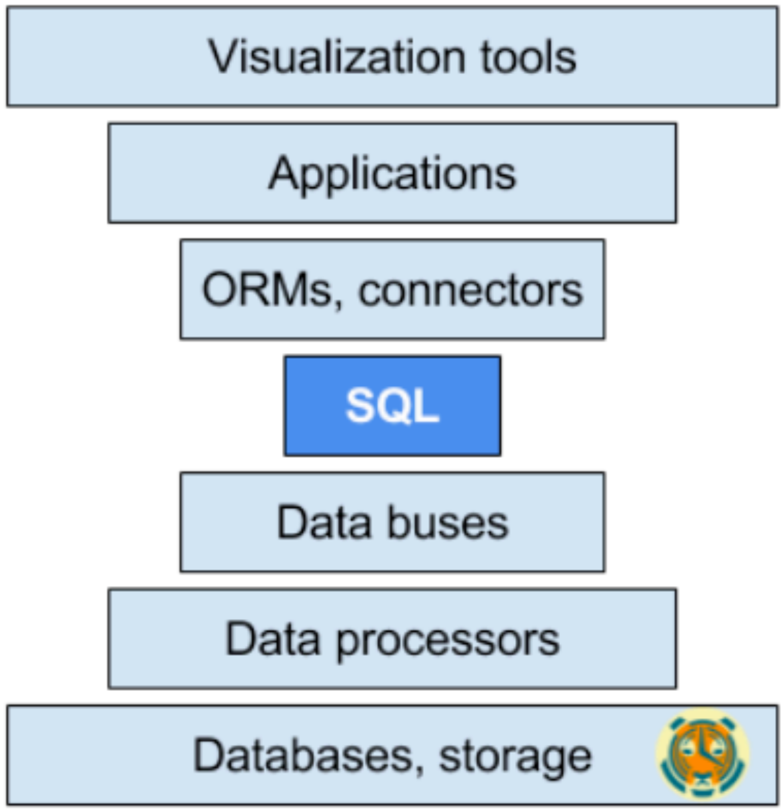
Take a look at Google's second major **Spanner** paper (Spanner: Becoming a SQL System, May 2017), and you'll find that it bolsters our independent findings.

Lessons from the Internet



IP as the Networking Universal Interface (source).

SQL as the slim-waist



The Data Universal Interface

Most of the issues are carried over to Clinical Narrative
We are adding other modalities involved…

Audio,
Images,
Videos
…

The same task can become exponentially harder or impossible

Annotation as Code

What if...

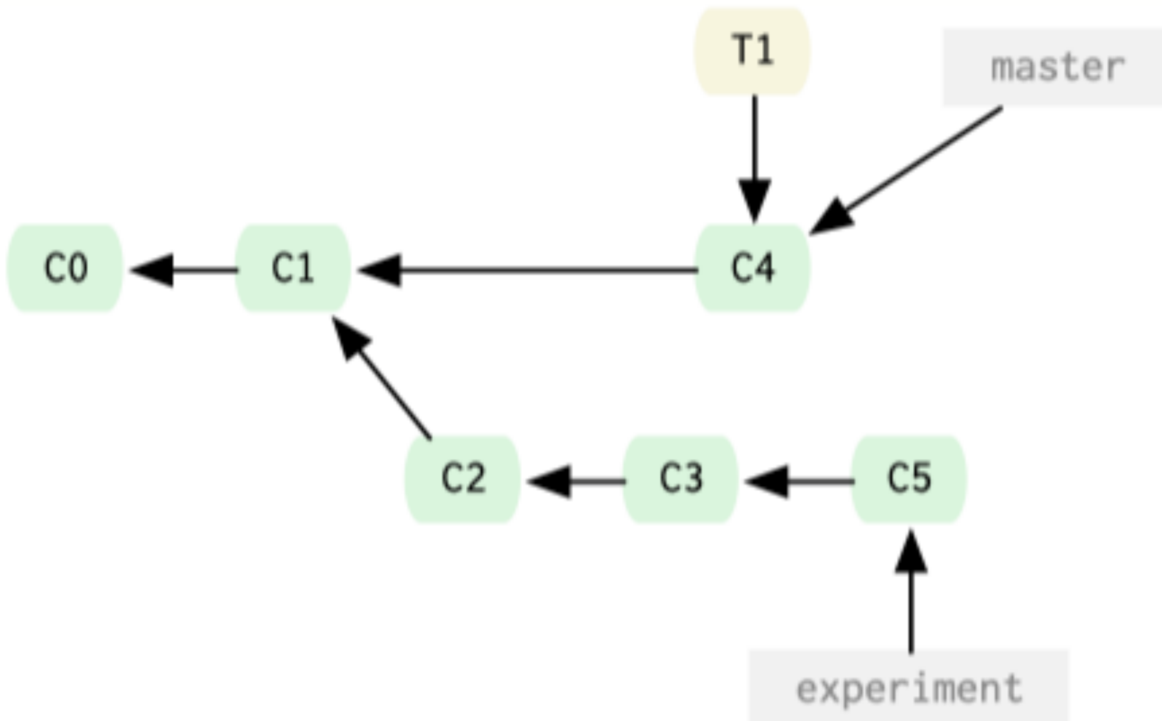
- Annotations are represented as we represent and manage ***source-code***?
- One peculiarity—***increased complexity of the semantics*** for such “language”
- We might benefit from a ***version control*** mechanism

Functional Data Structures

Ideas from *git*

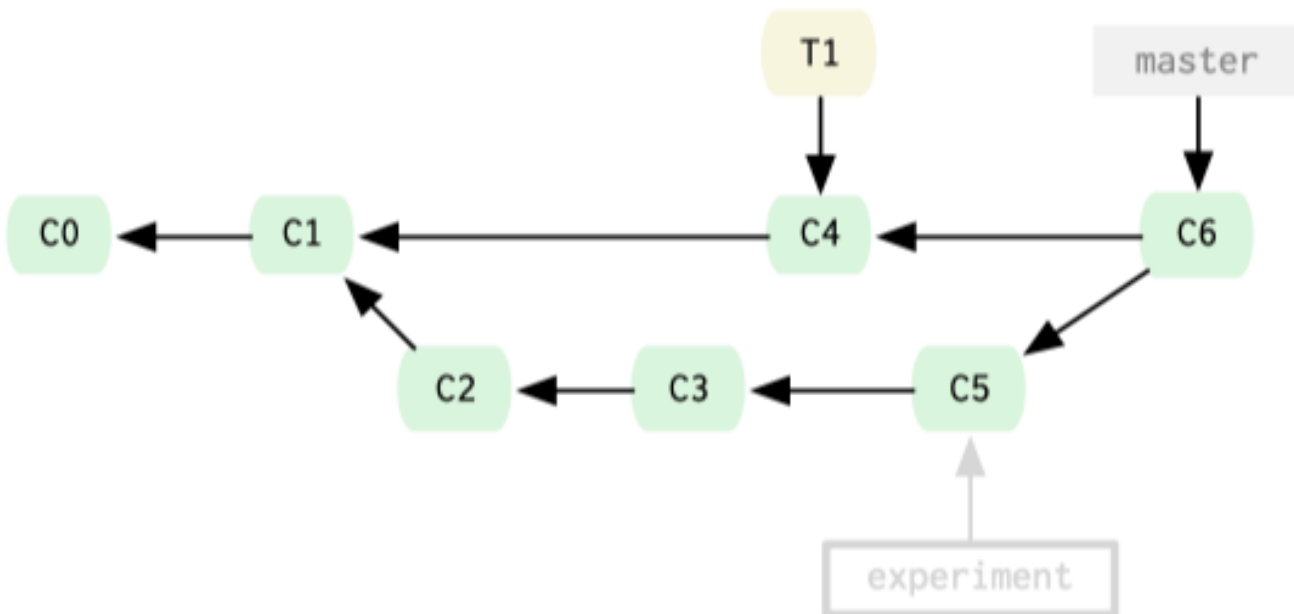
- ***Each version*** of ***each layer*** of annotation is an ***incremental operation*** on top of the ***earlier version***.
- Try to maximize ***deterministic bi-directional*** transformations
- Minimize ***lossy uni-directional*** transformations
- Track ***annotation version*** and the ***guideline specification*** dependencies
- Create new ***annotation snapshots***
- Consistency checks using ***content hashes***

Past Annotations Reachable



Annotation Guidelines can be *kept in sync*

Ease of Experimentation

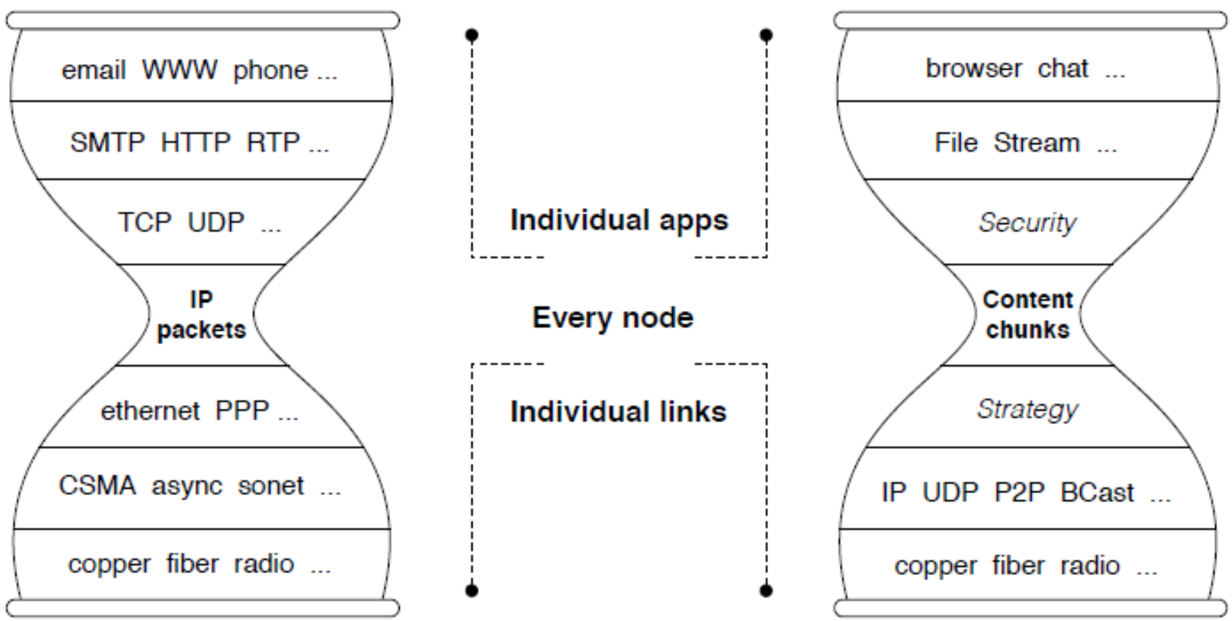


Very important to know exactly what guidelines were used for a particular set of annotations

Internet Philosophy

Next Generation of the Internet

- **Plain text** files where possible—UTF-8 for serialization or even base64
- **Media Containers**—akin to the next generation of internet that focuses on content—Named Data Networking (NDN) or Content Centric Networking (CCN)



Layer Tags

When a `.parse` is not <u>the</u> `.parse`

- File extensions are typically used to determine content, but when it comes to layers of annotation, things can easily get complicated
- Example, a `.parse` might contain one of many *kinds* or *qualities* of parses
  - **Dependency parse**
    - Universal Dependency v2.0
    - Custom dependency
  - **Constituency parse**
    - Gold parse (with traces)
    - Automatically generated
      - Using Charniak parser, model (A)
      - Using Charniak parser, model (B)
      - Using Berkley parser
      - ...
- Similarly the columns in a `.conll` file might be interpreted differently depending on the year and task involved

When a `.parse` is not <u>the</u> `.parse`

- File extensions are typically used to determine content, but when it comes to layers of annotation, things can easily get complicated
- `oohggg_parse` might contain one of many *kinds* or *qualities* of parses
- `cnnnnn_parse` might contain one of many *kinds* or *qualities* of parses
  - **Gold or Automatic**
    - `o` = Gold parse (using OntoNotes guidelines);
    - `t` = Gold parse (using original Treebank guidelines)
    - `c` = Charniak parser (Automatic);
    - `b` = Berkeley parser (Automatic)
  - **Traces**
    - `o` = original traces;
    - `n` = no traces
  - **Hyphens**
    - `h` = split-at-hyphens;
    - `n` = not-split at hyphens
  - ...

Cannonical, Compositional Representation

The case of Chinese Characters

Graphical Features—

- Recent work by Wang *et al.*, (2019) has shown that using the ***radicals*** in Chinese characters contain semantic information similar to the notion of ***subwords*** and ***suffixes*** in English and can be used to improve unsupervised learned representations that can improve ***named entity tagging***

Character	Primary Radical
病(illness)	疒(sickness)
痨(tuberculosis)	疒(sickness)
痛(pain)	疒(sickness)
肝(liver)	月(moon)/肉(meat)
胸(chest)	月(moon)/肉(meat)
脑(brain)	月(moon)/肉(meat)

Multiple representations—

- ***Pinyin*** representations of Chinese characters also help…

Character	Pinyin
病(illness)	bìng
痨(tuberculosis)	láo
痛(pain)	tòng
肝(liver)	gān
胸(chest)	xiōng
脑(brain)	nǎo

Interoperability Matters

Case of COVID-19

- Shah and Curtis (2020) identify the *limitations of current EHR systems*
- Difficulties in pooling multiple data sources owing to missing mapping between different medicine nomenclatures
- For a simple query—***Find me patients using Hydroxychloroquine***
  - — EHR (A) used **National Drug Code**
  - — EHR (A') used **Medi-Span**
- A **Common Data Model** (CDM) would have helped bridge the two variations of the same EHR system and allowed for better and quicker data analysis
- CDM is ***not automatic*** and ***not static***, but a better tracking system can be used to manage mapping across multiple versions and nomenclatures.

Learning CDM (Mappings)

- Dong et al. (2020), shows the significance of mapping types of **COVID-19** ***tests*** using **LOINC** codes
  - ~600 Manually mapped codes
  - 99.3% ITA (Cohen's kappa)
  - 98.9% automatic mapping accuracy
- Allowed finer grained analysis of COVID-19 testing data across 8 sites

LOINC Code	Total	Percentage	LOINC Long Common Name
Molecular			
94759-8	240	42.25	SARS-CoV-2 (COVID19) RNA [Presence] in Nasopharynx by NAA with probe detection
94500-6	202	35.56	SARS-CoV-2 (COVID19) RNA [Presence] in Respiratory specimen by NAA with probe detection
94309-2	75	13.20	SARS-CoV-2 (COVID19) RNA [Presence] in Unspecified specimen by NAA with probe detection
94502-2	13	2.29	SARS-related coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection
94660-8	11	1.94	SARS-CoV-2 (COVID19) RNA [Presence] in Serum or Plasma by NAA with probe detection
Antibody			
94563-4	10	1.76	SARS-CoV-2 (COVID19) IgG Ab [Presence] in Serum or Plasma by Immunoassay
94564-2	4	0.70	SARS-CoV-2 (COVID19) IgM Ab [Presence] in Serum or Plasma by Immunoassay