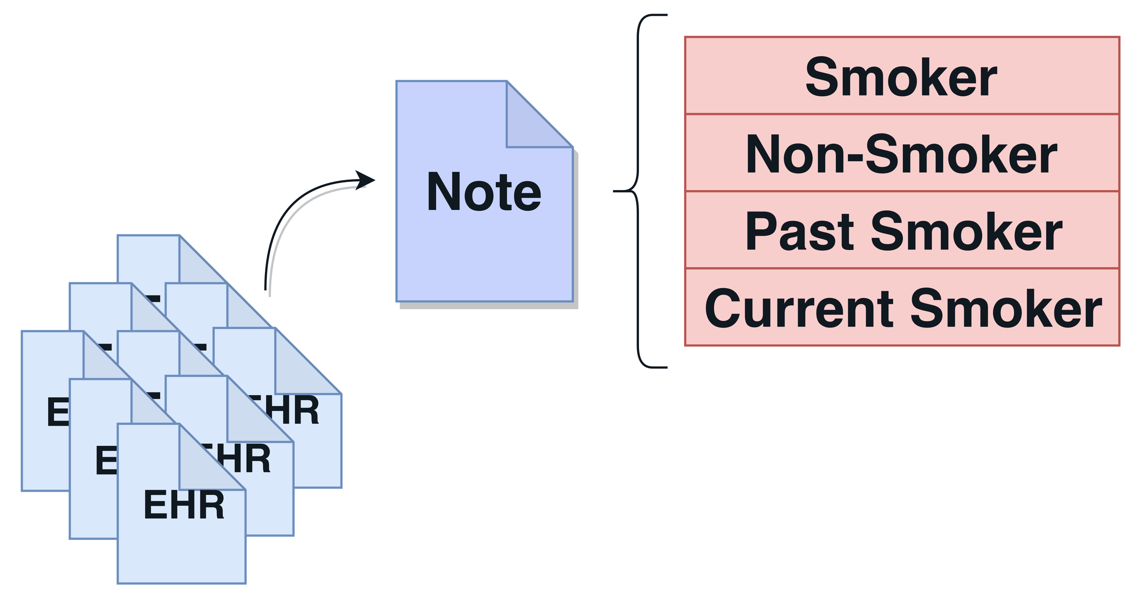
Phenotyping of Clinical Notes with Contextualized Neural Language Models

Andriy Mulyar[1], Elliot Schumacher[2], Masoud Rouhizadeh[2], and Mark Dredze[2]

Virginia Commonwealth University[1], Johns Hopkins University[2]

What is Clinical Note Phenotyping?



- **Clinical phenotyping**: extract patient conditions or traits from unstructured clinical text.
- Automated phenotyping of clinical notes:
  - Adds structured information to electronic health records.
  - Enhances the productivity and accuracy of medical coders.
  - Provides information for downstream clinical decision support tasks.

Why is Phenotyping Difficult?

- A given phenotype may not be present in a specific clinical note.
- A phenotype may be present, but not explicitly stated.

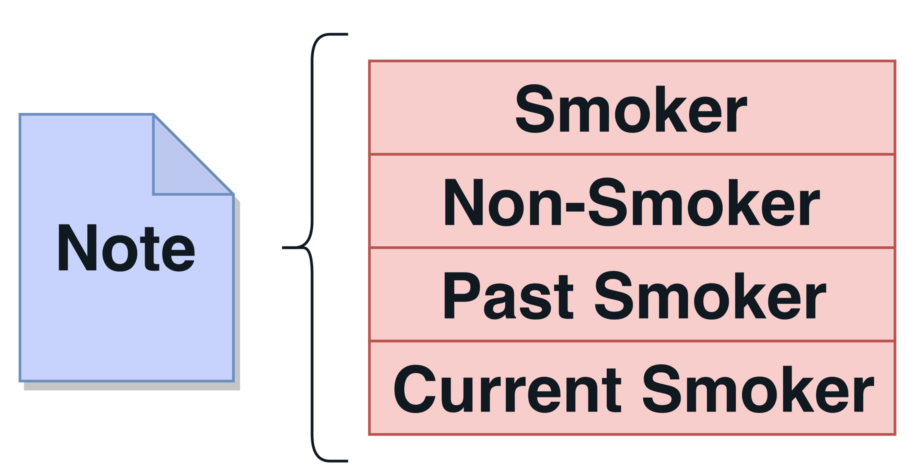
"...The patient currently resides on a friend's couch ..."	**Homeless**

- Signal of a phenomic trait may be present anywhere within a clinical note.

Phenotyping Benchmarks

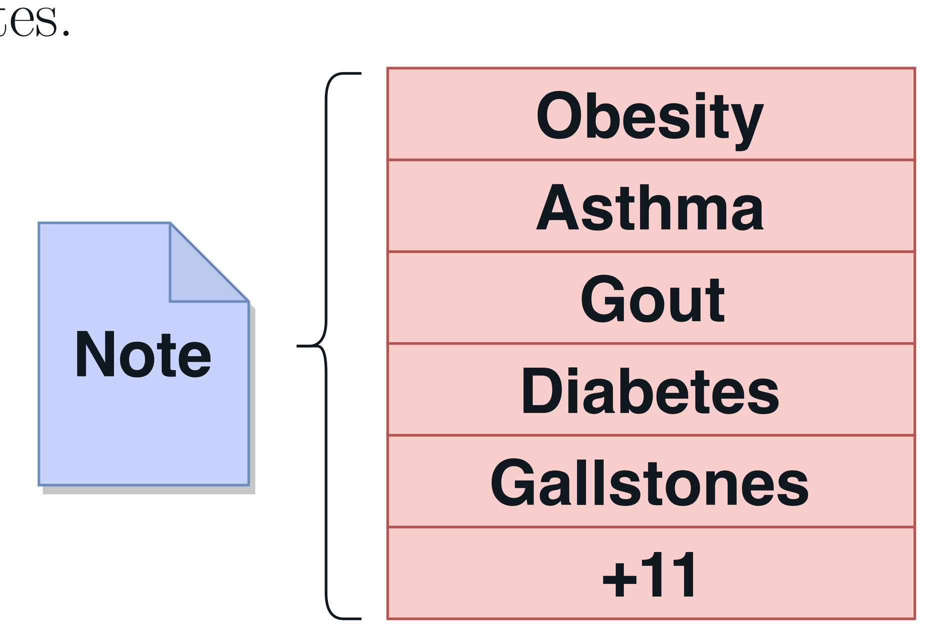
I2B2 2006: Smoking

Smoking status identification.



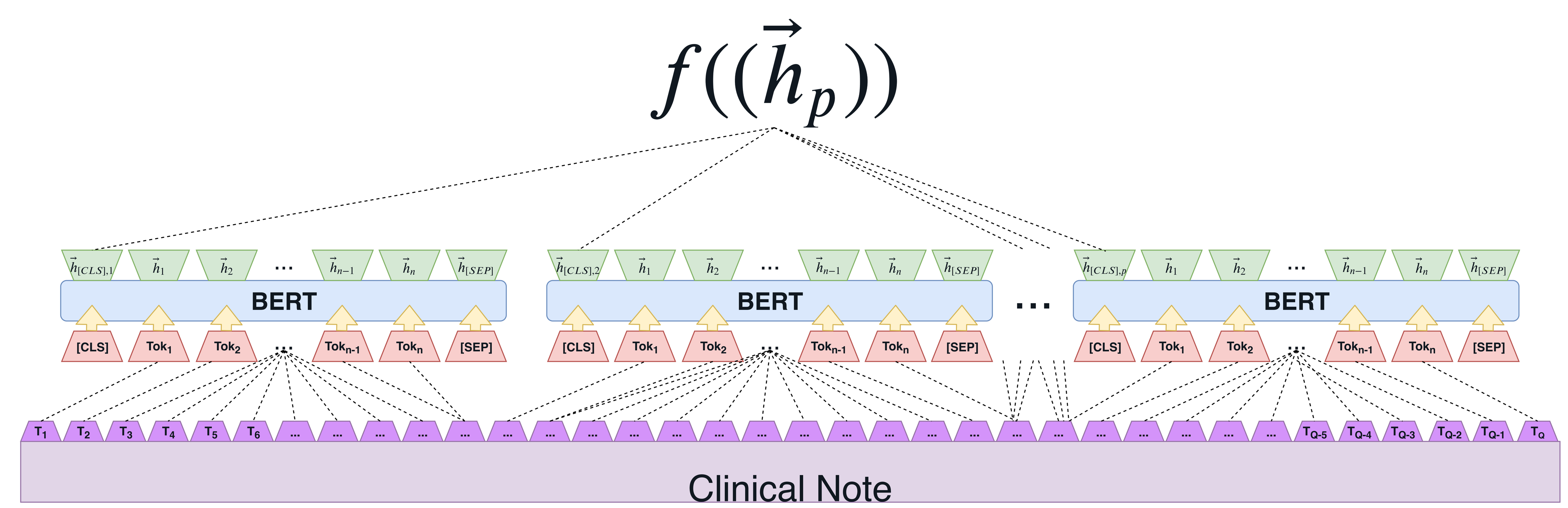
I2B2 2008: Obesity

Disease presence identification in discharge notes.



End-to-End Phenotyping with Language Models

- Frame phenotyping as a document classification task.
- Unroll a pre-trained, fixed context window language model (BERT) to generate a sequence of locally contextualized classifier token representations.
- Encode a global document context vector from local classifier token sequences and train with cross-entropy classification head.



Experiments

- Global Context Heads $f((\vec{h}_p))$
  - Dimension-wise max
  - Concatenation (identity)
  - Transformer Encoder
  - LSTM Encoder

	I2B2 2006: Smoking	I2B2 2008: Obesity
$f_{\max}$	50.3	74.7
f_I	82.9	76.8
$f_{\text{Transformer}}$	75.9	**97.7**
f_{LSTM}	**98.1** (97.1 ± .48)	**99.7** (93.9 ± .59)
Shared Task 1st Place	90.0	95.0
Majority Label Baseline	81.0	74.4
DocBert	80.2	67.6
CNN	77.0	–
CNN + Rules	–	96.2

- **State-of-the-art** performance with LSTM encoder.
- Outperforms rule based techniques with no task specific dictionary searches or rules.
- Sequence of classifier tokens has average length ~ 7 across both tasks.
  - Transformer Encoder and LSTM Encoder treat global context encoding as sequence compression problem.