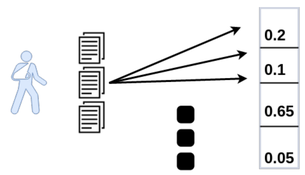
Emerging Technologies for Efficient Clinical, Genetic, and Omics Inference

Oliver Bear Don't Walk IV[1*], Yosuke Tanigawa[1*], Alice Yu[1], Adam Lavertu[1], Sandeep Ayyar[1], Manuel Rivas[2]
[1]Biomedical Informatics, Stanford University, [2]Department of Biomedical Data Science, Stanford University
[*]These authors contributed equally to this work

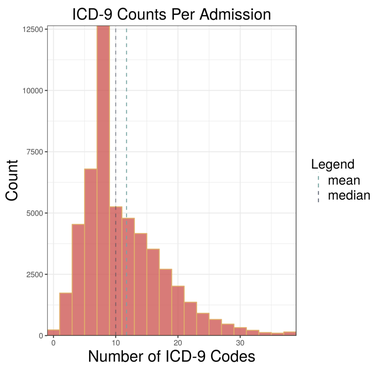
Clinical Note Tagger

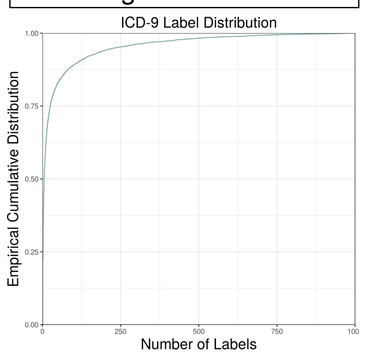
Manual feature extraction has struggled to extract the large amounts of rich data in text while deep learning has seen great success in NLP tasks because of its ability to extract features from raw text. Our goal is to make use of these features by tagging clinical notes with ICD-9 codes via a deep learning approach.

In the past this problem has been dominated by rule based methods, and machine learning has yet to become a viable alternative.



Work so far:
- Initial exploration of clinical notes
- Converted words to distributed vector representations
- Deep learning architecture in place for training



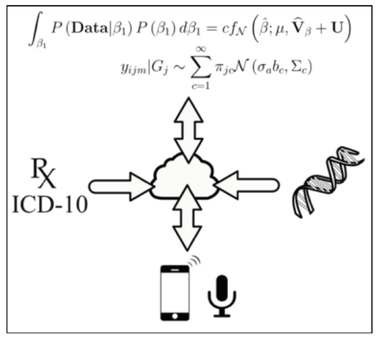


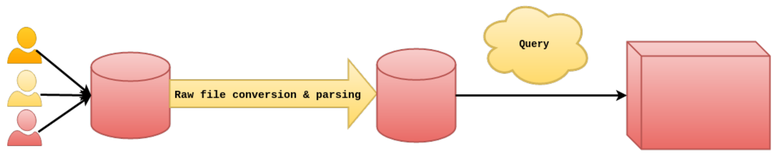
Future Work

The strength of deep learning is that it is able to automatically create and blend features from raw data. We can help with this process by making sure the raw features are relevant, and that the model architecture can effectively summarize clinical notes.

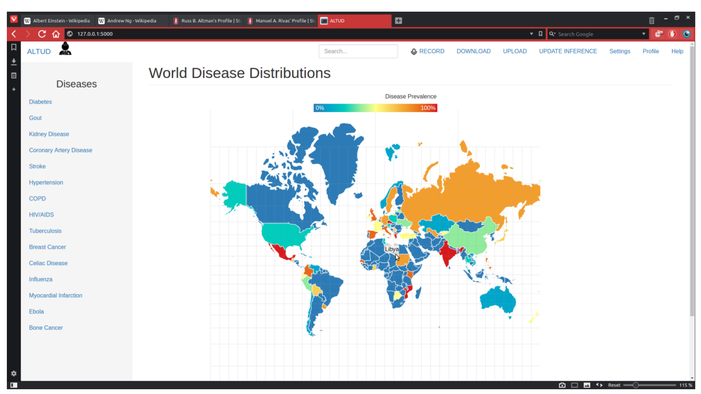
ALTUD: AnaLysis of Text and speech for Understanding Diseases

ALTUD provides a platform for researchers, physicians and patients to upload their clinically relevant data. Raw data can be stored and as technology improves more information can be extracted in order to combine EHR and genomic data.





Work so far:
- During the Health++ hackathon we built a prototype which takes in dictation, stores the text, extracts disease mentions and allows the user to visualize global disease counts.
- Added the ability to uploading physical notes
- Improved OCR capabilities
- Integrated Google Drive into the workflow



Future Work

- Integrate different health record modalities
  - Sequencing data
  - Environmental factors
- Integrate tools to sit on top of ALTUD
  - Clinical Note Tagger
  - Haplotype Inference and Genome Compression

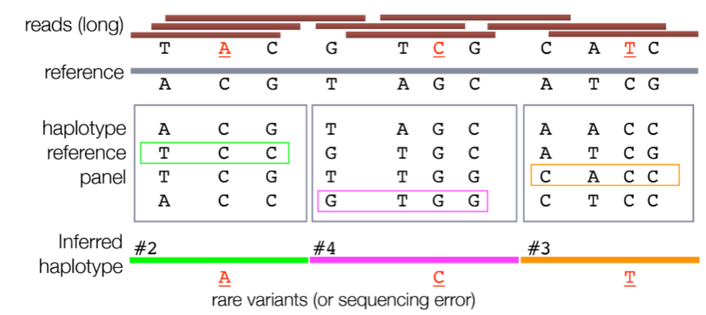
Online Haplotype Inference and Compression of Personal Genomes

Research Questions

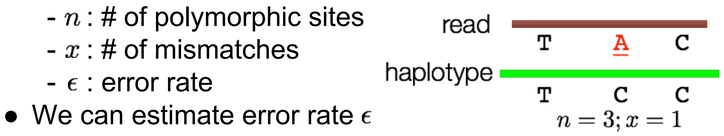
- Can we infer haplotypes of personal genomes **on the fly** with portable sequencers?
- What would be the **compact & efficient** representation of personal genomes?
- How can people **share personal genomes** with researchers while maintaining participants's **privacy**?

Oxford Nanopore Technologies SmidgION

[Methods] Genome = haplotype indices + rare variants

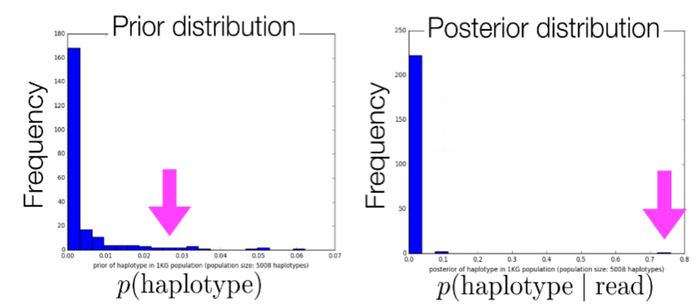


- Population reference captures most of the common variants
- Long-reads contain multiple polymorphic sites
- $p(\text{haplotype} \mid \text{read}) \propto p(\text{read} \mid \text{haplotype}; \epsilon) p(\text{haplotype})$
- Prior $p(\text{haplotype})$: Maximum Likelihood Estimate
- Likelihood $p(\text{read} \mid \text{haplotype}; \epsilon)$:
  Binomial model $p(\text{read} \mid \text{haplotype}; \epsilon) = \epsilon^x (1-\epsilon)^{n-x}$
  - n : # of polymorphic sites
  - x : # of mismatches
  - ϵ : error rate



- We can estimate error rate ϵ
  - $\hat{\epsilon}$: mismatch rate on non-polymorphic sites

[Preliminary Results] Rapid haplotype inference



- Data: Nanopore consortium NA12878 & 1KG population
- Example read: chr20; 30kb long; 15 SNPs
- Prior = 0.0263 —> Posterior = 0.7596
- Only one read is sufficient to find the correct haplotype