UCSF Information Technology

Machine-Redacted Notes & EMERSE

EMERSE WEBINAR – MARCH 3rd

Lakshmi Radhakrishnan – Data Scientist, Bakar Institute
Eric Meeks – CTO, CTSI Research Technology

Outline

- Introduction
- Redaction Process
  - COVID-19 DM & Notes
- EMERSE at UCSF
  - How we chose EMERSE
  - Effort to implement and operate EMERSE
  - Impact of EMERSE
- Questions

Introduction

Clinical Notes

Clinical notes are unstructured data with valuable information relevant to multiple research projects.

Some Ongoing Research Project at UCSF Using Clinical Notes:

- Extracting grade and metastases for urothelial cancers
- Critical marker values and doctor sentiments in gliomas and glioblastomas
- Prostate Pathology, Epidemiology and predictive modelling
- Social Determinants of Lower Back Pain / COCOA, Predictive modelling
- Antacid use as a risk factor of COVID-19 infection
- Differential diagnoses and patient similarity in neurodegenerative conditions
- Analysis of provider sentiment in oncology treatment
- Hip fracture detection

Accessing Clinical Data for Research - IRB

IRB stands for The Institutional Review Board

It is an administrative body established to protect the rights and welfare of human research subjects.

- Any research project that requires the use of clinical data like notes or structures data needs IRB approval
- IRB Approval for a project can take anywhere from 2 to 5 weeks and costs few thousand dollars

Machine-Redacted Notes

Clinical notes extracted from UCSF Electronic Health Record, Epic, and processed by software to remove 18 PHI elements. Each patient's records are assigned synthetic identifiers that maintain the link between notes and the patient's structured data in DEID CDW. There are two datasets of UCSF Machine Redacted Notes, known as 1) General Set and 2) COVID-19 Set.

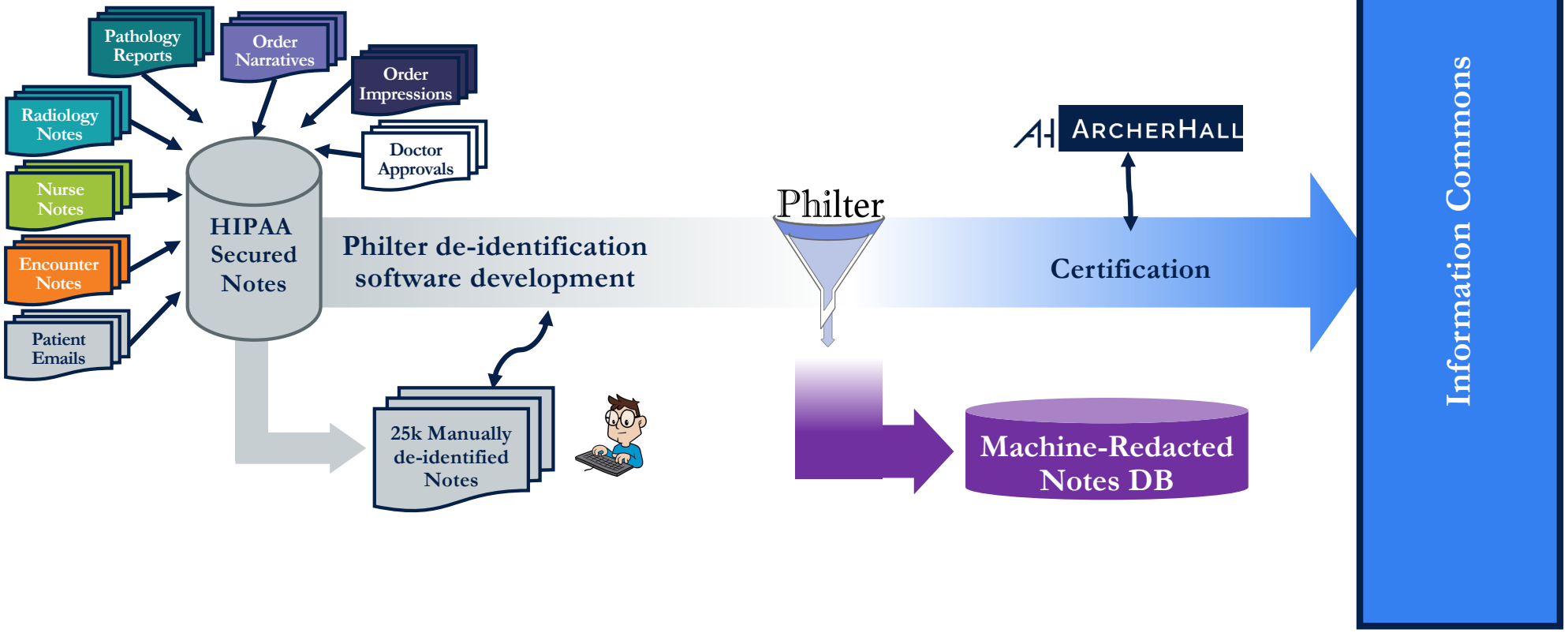
EMERSE: **E**lectronic **Me**dical **R**ecord **S**earch **E**ngine

An open-source software tool to search and browse notes data, without programming experience. UCSF has installed two instances of EMERSE here at UCSF against our 1) machine-redacted notes and 2) COVID-19 notes. EMERSE was created by the University of Michigan.

Notes Data and Deidentification

Clinical Data and Notes

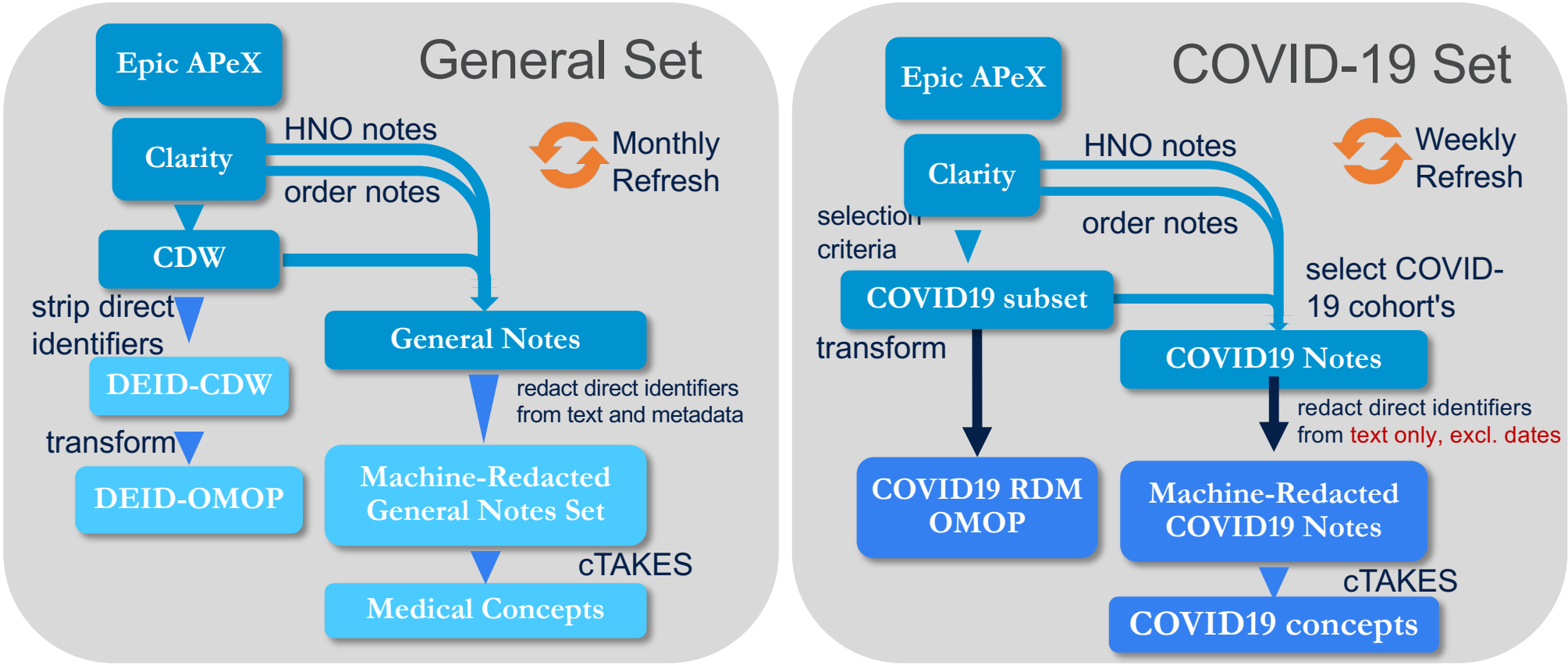
115 million clinical notes



UCSF

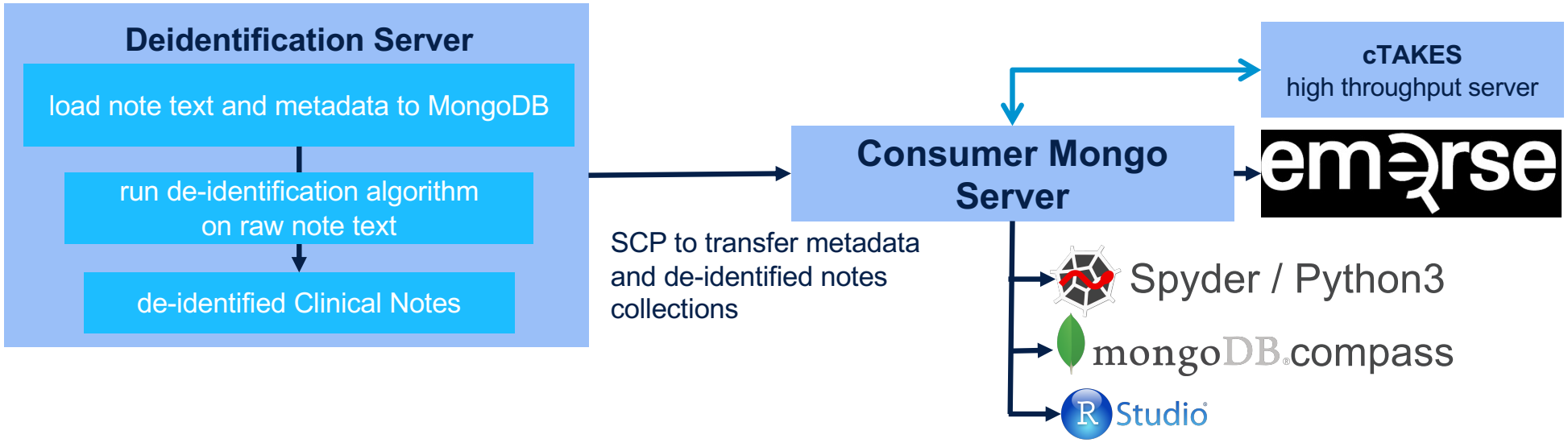
*B. Norgeot, *et al.*, NPJ Digital Medicine, 2020

UCSF Notes



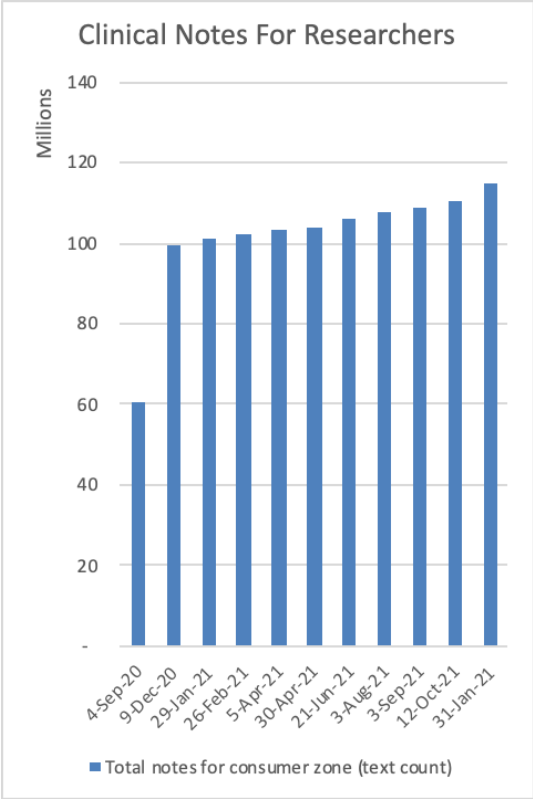
UCSF

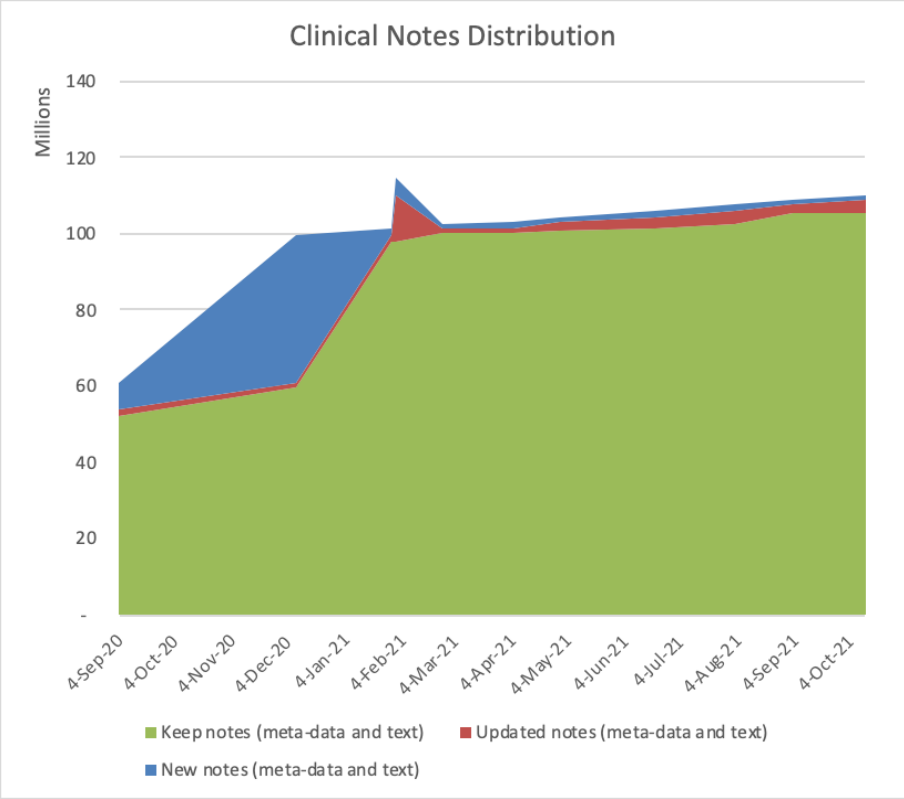
M-R Notes Data Flow



UCSF

Monthly Notes Refresh Cycle





Sample Note

Before De-identification

A900001 2011-1-20 18:40:00.000 7465789 1
EMERGENCY DEPARTMENT CHART [FINAL] ---------
- VISIT DATA ----------
---- Patient SMITH, MARY (317 62 10-1), visit number 189487645, arrived at the ED at 1/20/2011 5:35:18 PM with a complaint at registration of PAIN ARM/LEG-EXTREMITIES. ---------- HISTORY OF PRESENT ILLNESS ---------- (Triage complaint: brick fell on RT foot yesterday at work. now c/o rt big toe pain. no swelling UC), 29 y.o. M c/o of R big toe pain 1/10 mild after a brick fell on his R foot yesterday. He denies any other injury or pain or complaint. No fall. Able to walk without pain or difficulty. No pain or complaint proximally up his RLE. No numbness/tingling., --Medications None--Allergies None --Past Medical History rt eye surgery --Tobacco: None --Alcohol: None --Recreational drugs: denies Other medical charts and records reviewed and interpreted: Triage and nursing notes ---------- REVIEW OF SYSTEMS ---------- --GENERAL ROS comments / Constitutional: negative, fever, chills. --EYES: negative . --EAR NOSE MOUTH THROAT: see HPI

After de-identification

DB35CD2D4DD54A 01/05/2011 18:40:00.***** ***** 1
EMERGENCY DEPARTMENT CHART [FINAL] ----------
VISIT DATA ----------
Patient *****, ***** (***** ***** *****-*****), visit number *****, arrived at the ED at 01/05/2011 5:35:18 PM with a complaint at registration of PAIN ARM/LEG-EXTREMITIES. ---------- HISTORY OF PRESENT ILLNESS ---------- ---------- (Triage complaint: ***** fell on RT foot yesterday at work. now c/o rt big toe pain. no swelling UC), 29 y.o. M c/o of R big toe pain 1/10 mild after a ***** fell on his R foot yesterday. He denies any other injury or pain or complaint. No *****. Able to walk without pain or difficulty. No pain or complaint proximally up his RLE. No numbness/tingling., --Medications None--Allergies None --Past Medical History rt eye surgery --Tobacco: None --Alcohol: None --Recreational drugs: denies Other medical charts and records reviewed and interpreted: Triage and nursing notes ---------- REVIEW OF SYSTEMS ---------- -- GENERAL ROS comments / Constitutional: negative, fever, chills. --EYES: negative . --EAR NOSE MOUTH THROAT: see HPI.

Certification Process

- HIPPA's Privacy Rules:
  - Safe Harbor
  - Expert Determination
- ArcherHall -
  - Forensic data collection, analysis **expert**
  - Iterative process

ArcherHall Checks

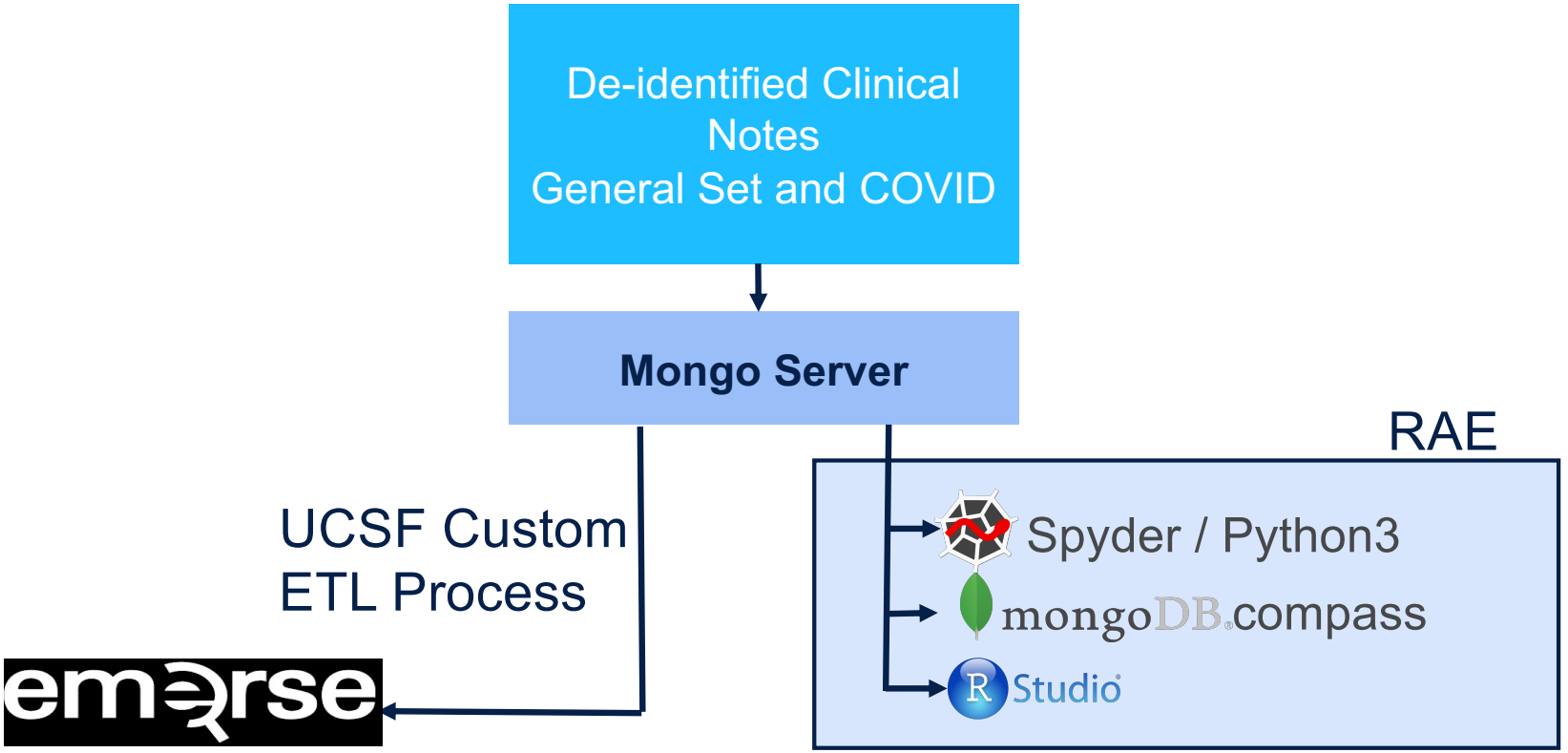
- 10M Clinical notes for validation
- Date Fields
  - Expert Determination approach
  - Less than 0.025% of patients at risk of re-identification
- Non-Date Fields
  - Safe Harbor
  - No instances of PII in the de-identified notes field

Certified De-identified Notes Access Requirements

- Currently only for UCSF and UC Berkley
- Service Now – De-identified data assets request form
- Approver – To sanction access (Faculty, Director)
  - No director approval for UCSF Faculty
- Required courses and Training:
  - UCSF Cyber Security Awareness Fundamentals (required yearly)
  - HIPAA 101 – Privacy and Security for New UCSF Faculty, Staff, Trainees, Students and Volunteers
  - Your Responsibilities While Working with Clinical Data (required every two years)
- Finally Data Use Agreement
  - UCSF Required Agreement on use, management and sharing of de-identified patient data

How to Access

Accessing the Notes Data



UCSF

EMERSE

The Electronic Medical Record Search Engine

UCSF

How we chose EMERSE, a bit of history

Conversations with David H about EMERSE

UCSF's algorithm (Philter) for machine-redaction (M-R) of notes is ready

Feb: UCSF will use EMERSE with M-R notes

Apr: Sub-award received from UMich for UCSF EMERSE Team

Apr: Decision to launch an additional COVID- EMERSE instance

Apr-May: EMERSE is running for internal QA

Jul: Launch!
EMERSE & COVID-EMERSE

Feb: Campus-wide webinar

Nov-Dec: Re-load **certified** de-id notes into EMERSE

No IRB approval required for use!

Effort to Implement and Operate EMERSE

- Implementation
  - **Develop a custom ETL process to import notes, patient data and users**
  - Add custom code/configuration to support UCSF Single Sign On (SSO) which is based on SAML/Shibboleth
  - **Manage programmatic access to the SOLR API**
- Operations
  - Perform QA to see how well EMERSE works with DEID data
  - Properly size server to handle data and traffic
  - Create processes/scripts to keep EMERSE running

Custom ETL for UCSF EMERSE

If you want to install EMERSE at your institution, building an ETL process is the biggest tech lift (and it's not that bad)

- Over 100 million notes to transfer from Mongo into SOLR
  - Takes us about 4 to 5 days to do a full transfer
- Nearly 2.5 million patients from DEID CDW to Oracle XE
  - Takes a few hours
- Handful of user to import from an LDAP group

Much Smaller numbers of notes and patients for the COVID-EMERSE instance

Custom ETL code is < 1000 lines of Java

Access to the Solr API

In addition to the EMERSE UI, UCSF users have access to the Solr API

- UCSF always planned to give direct programmatic access to de-id notes to support research
- We asked our researchers if programmatic access to the EMERSE data was something they would be interested in. The answer was a loud **YES!**
- Researchers were not able to access the notes via the original methods we were considering
- Solr provides efficient and flexible text search and retrieval functionality through programmatic access

Tech Aspects of Managing Access to SOLR

Wanted to control access to SOLR via the EMERSE/Spring security mechanism. Needed a proxy to SOLR that would integrate into EMERSE

Found and then modified a Java based proxy solution at:

https://github.com/mitre/HTTP-Proxy-Servlet

- Uses basic auth (programmer friendly unlike Shibboleth/SAML). Users login with their standard unique "UCSF network" credentials.
- Integrating the proxy into EMERSE required no Java development (in EMERSE) and could be configured with a few simple edits to web.xml and security.xml
- All access is logged to be HIPAA compliant
- We then use one "system level" researcher login into SOLR, and that is shared by all EMERSE users who authenticate into the Proxy API. This is the exact some model that most applications (including EMERSE) use to connect to a database
- Special thanks to David Smiley for developing the proxy!

Impact of EMERSE: Benefiting UCSF Research Teams

1. **Preliminary data review.** Easily review representative notes up front to assess whether information of interest is present and how it is encoded so we could then write code to programmatically pull it in bulk
2. Quickly see the **distribution of relevant terms**. The color coding permitted us to see distributions by how many colors showed up on the summary page
3. **QC during data review.** QC of training data for deep learning models. Data discrepancies were easily reviewed by pulling up notes to pinpoint areas needing attention
4. Enabled easy **self-serve check** on what de-id data are available before going to the trouble of pulling data from additional *identified* data sources.

Impact of EMERSE: Thoughts and anecdotes

1. EMERSE is a part of a de-identified suite of products and services that UCSF has been working on for several years. Some measures of EMERSE at UCSF are likely to be a measure of the complete suite of services, but EMERSE made the notes aspects of our DEID data come to value.
2. Any problems with EMERSE or SOLR are VERY QUICKLY sent to the support team (aka me). Research projects are quickly becoming dependent on EMERSE data.
3. "Hi Leslie, Hope you are well! We are finally writing up the manuscript about breast cancer and pseudocirrhosis, which is very exciting! How do you prefer that we cite or acknowledge EMERSE? Thanks! Laura"
4. Our COVID-EMERSE instance, on the other hand, *currenlty* appears to be getting minimal use.

Thank you to the University of Michigan EMERSE Team!

From the UCSF EMERSE Team


Lakshmi


Eric


Leslie


Oksana


Gundolf

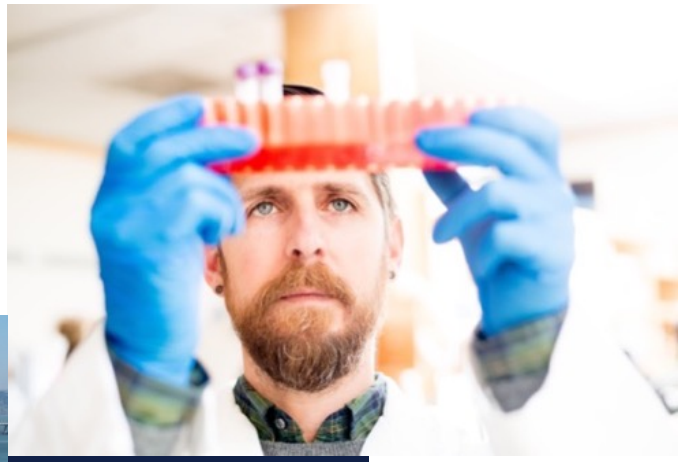

Sharat


Jason


Dima

Questions & Answers

Please follow the link above to view FAQ on the wiki



UCSF