Tech Talk

Automate Finding and Harmonizing Data With an Artificial Intelligence Tool: MetaMatchMaker

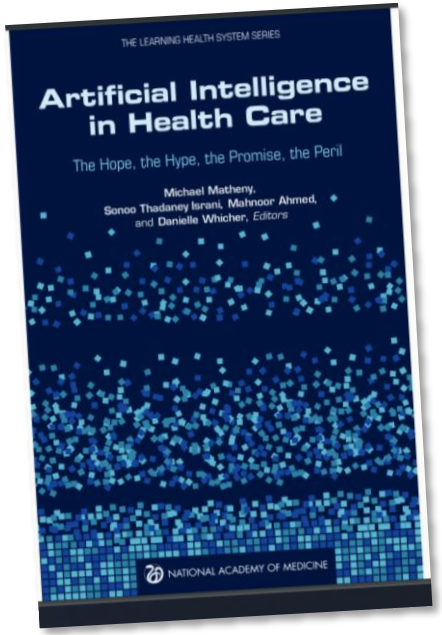
JULY 28, 2022

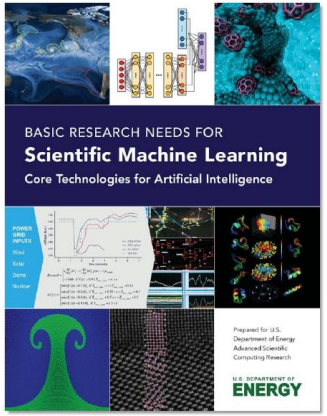
Grier Page, PhD
Eric Earley, PhD

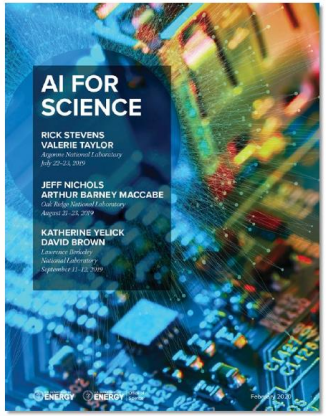
RTI INTERNATIONAL

MetaMatchMaker
Science. Faster.

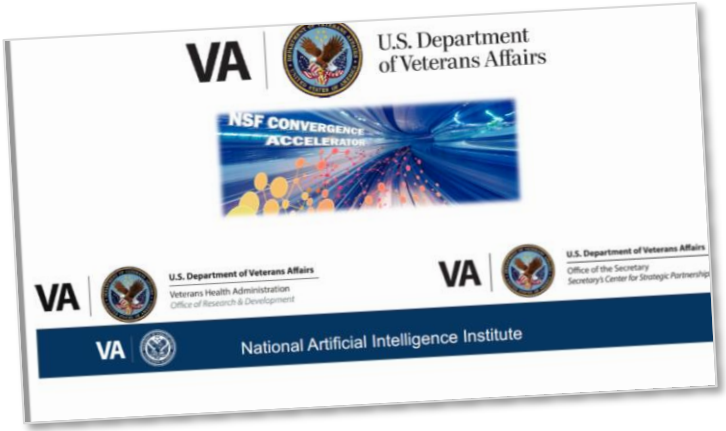
Everyday Artificial Intelligence Applications

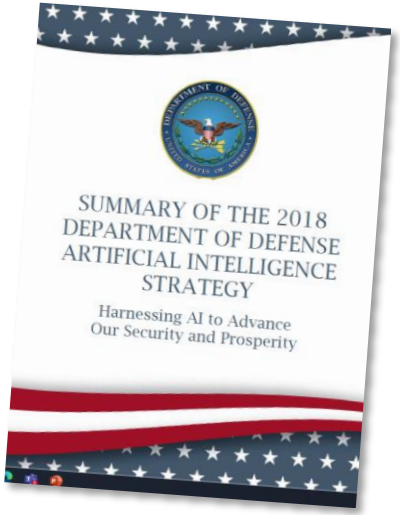

THE LEARNING HEALTH SYSTEM SERIES
Artificial Intelligence in Health Care
The Hope, the Hype, the Promise, the Peril
Michael Matheny, Sonoo Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher, Editors
NATIONAL ACADEMY OF MEDICINE


BASIC RESEARCH NEEDS FOR
Scientific Machine Learning
Core Technologies for Artificial Intelligence
Prepared for U.S. Department of Energy Advanced Scientific Computing Research
U.S. DEPARTMENT OF
ENERGY


AI FOR SCIENCE
RICK STEVENS
VALERIE TAYLOR
JEFF NICHOLS
ARTHUR BARNEY MACCABE
KATHERINE YELICK
DAVID BROWN


CENTER OF EXCELLENCE
ARTIFICIAL INTELLIGENCE
Accelerates adoption of Artificial Intelligence to discover insights at machine speed.
SERVICE OFFERINGS
AI Assessment Of Governance & Enablers
Use Case Discovery & Selection
Workflow Mapping & Process Automation
Lean Innovation Process Design
AI Solutions Identification & Implementation

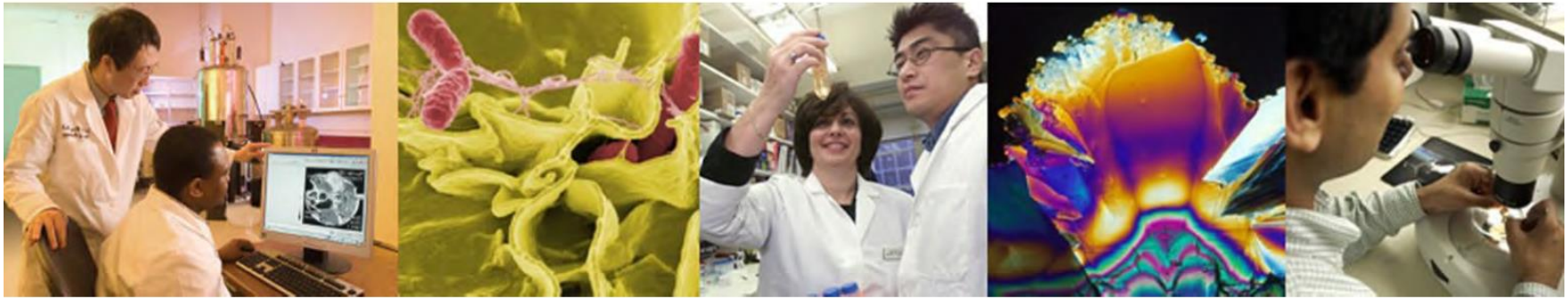

VA
U.S. Department of Veterans Affairs
NSF CONVERGENCE ACCELERATOR
U.S. Department of Veterans Affairs
Veterans Health Administration
Office of Research & Development
U.S. Department of Veterans Affairs
Office of the Secretary
Secretary's Center for Strategic Partnerships
National Artificial Intelligence Institute


SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY
Harnessing AI to Advance Our Security and Prosperity

Artificial Intelligence Working Group Update

119th Meeting of the Advisory Committee to the Director (ACD)

December 13, 2019



David Glazer
Engineering Director, Verily

Lawrence A. Tabak, DDS, PhD
Principal Deputy Director, NIH
Department of Health and Human Services

Artificial Intelligence, Public Trust, and Public Health

September 17, 2020 by Carlos Siordia PhD, Office of Science Fellow and Muin J. Khoury MD, PhD, Director, Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia

As a data-driven agency, CDC has always had highly skilled statisticians and data scientists. As part of the Data Modernization Initiative, CDC is supporting strategic innovations in data science using artificial intelligence and machine learning (Ai/ML). Ai/ML is the practice of using mathematics with computers to learn from a wide range of data and make predictions about the health of populations. By using Ai/ML, CDC can maximize insights from data to improve disease detection, mitigation, and elimination. Ai/ML applications could support public health surveillance, research and, ultimately, decision making, ushering a new era of precision public health.

Here, we provide a quick overview of how CDC scientists are using Ai/ML in public health surveillance and research. We sought to answer three questions:

- What topics have been covered by publications?
- How frequently is open-source software used?
- How often do authors make their algorithms public?

Traditional Approach to Data Science

35-50% of the time and budget of a study is spent on data identification and cleaning. Some studies report as much as 75%.

Kaggle 2018 Survey of Data Scientists

We are fundamentally challenged in the way we do science.

We spend a huge amount of time and budget assembling datasets before any insights can occur.

ECHO
Environmental influences on Child Health Outcomes
A program supported by the NIH

- >300 sites
- 20+ Years of data
- 20,000 clinical variables
- $100+ million annual budget

4.5 years to analyzable datasets

$500,000,000

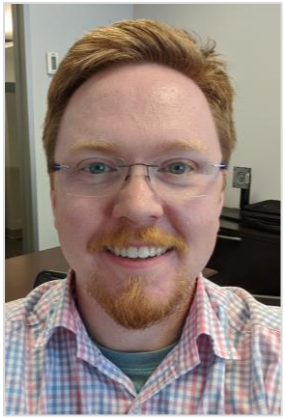
NSF Team – Convergence Accelerator

Grier Page
(PI)
Multi-center Studies
Harmonization

Xiangqin Cui
(co-PI)
EMR Statistics

Ricardo Henao
(co-PI)
Machine Learning

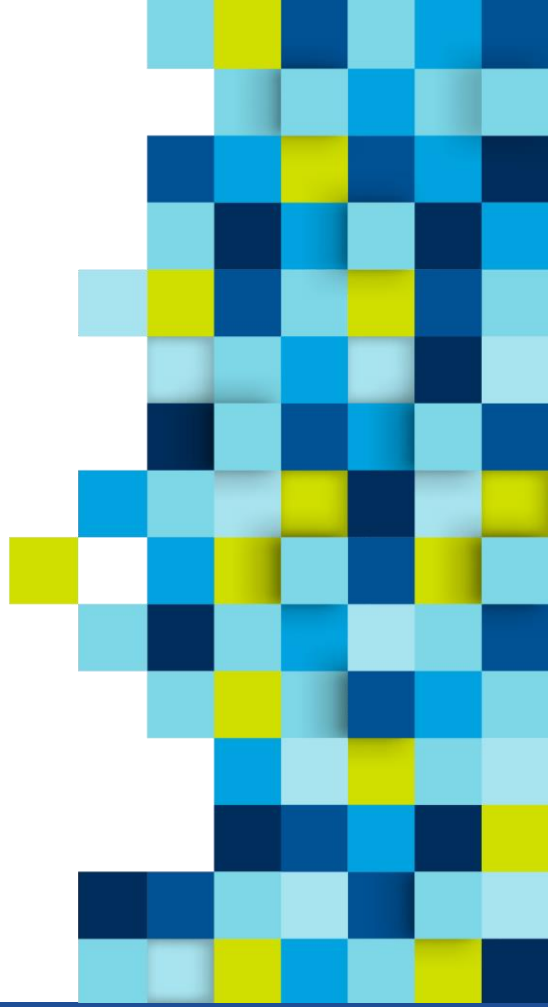
Yun Wang
(co-PI)
Machine Learning



Eric Earley
(PD)
Commercial Software
Development

RTI INTERNATIONAL | EMORY UNIVERSITY | FAVER Foundation for Atlanta Veterans Education and Research | Infinia ML REDEFINE POSSIBLE | COLUMBIA UNIVERSITY | RTI INTERNATIONAL

Use AI to Make Finding and Linking Data Easier, Quicker, and Cheaper

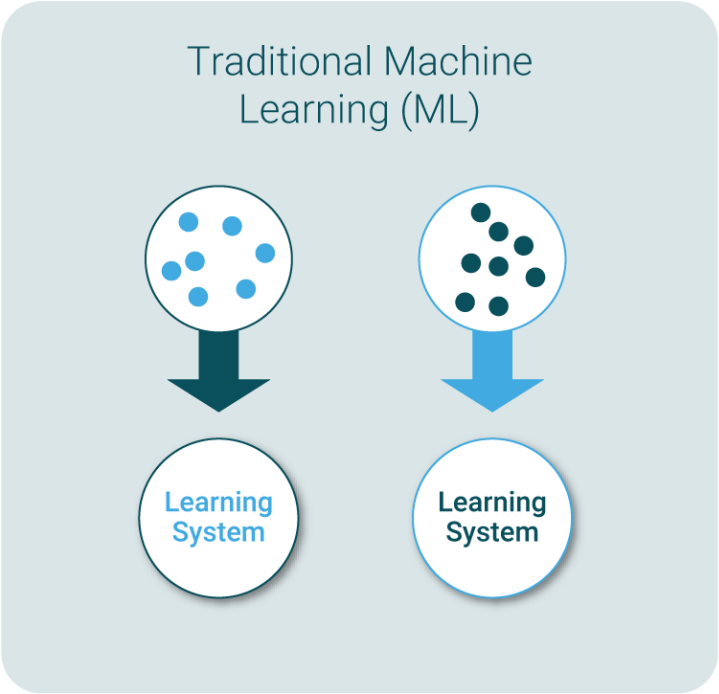


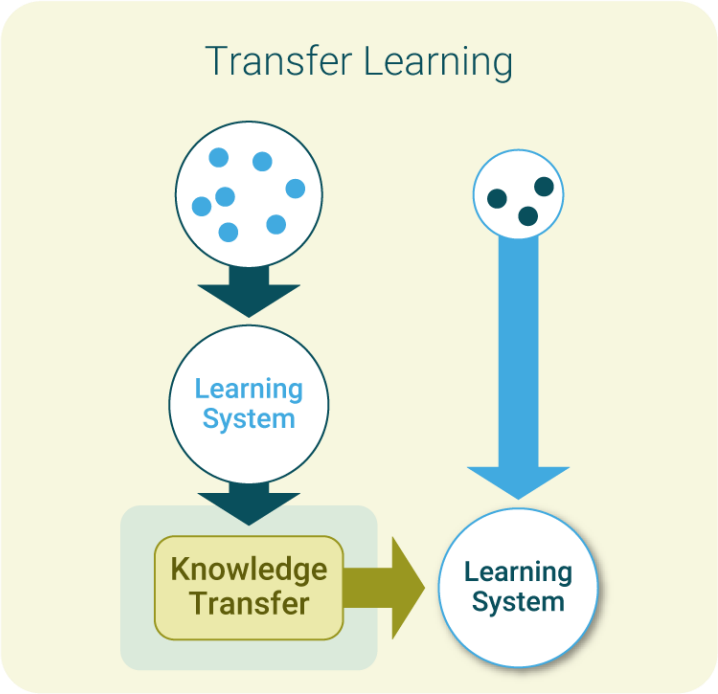
Effort in time and money for data study

Traditional Approach

Meta Match Maker

Transfer Learning





Major themes: From words to words-in-context

Word vectors

cats = [0.2, -0.3, ...]

dogs = [0.4, -0.5, ...]

Sentence / doc vectors

We have two cats. } [-1.2, 0.0, ...]

It's raining cats and dogs. } [0.8, 0.9, ...]

Word-in-context vectors

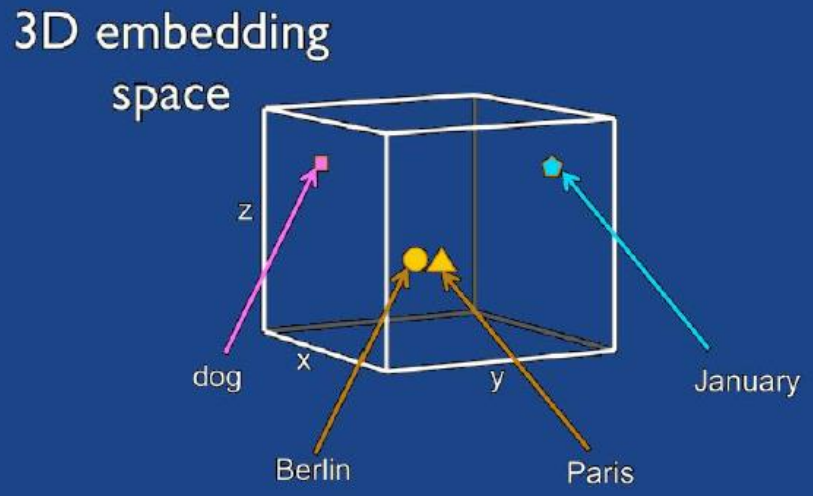
[1.2, -0.3, ...]

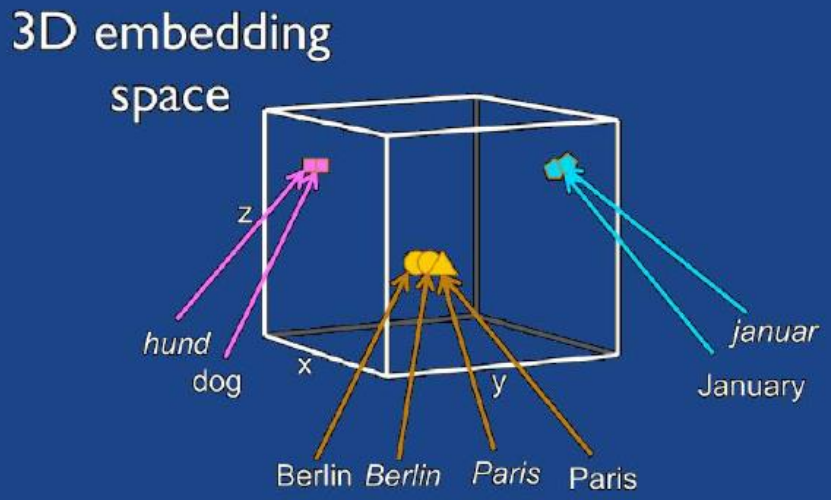
We have two cats.

[-0.4, 0.9, ...]

It's raining cats and dogs.

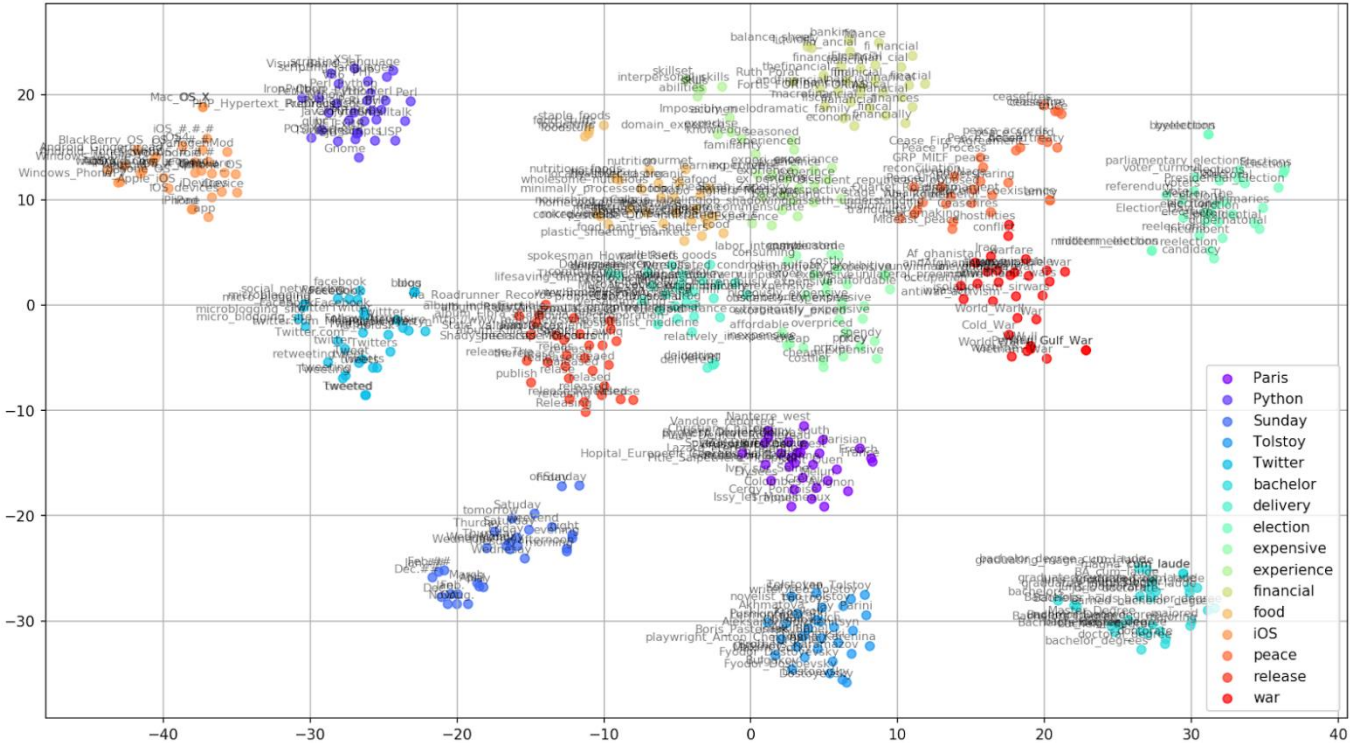
From Sebasitien Ruden


3D embedding space
z
x
y
dog
Berlin
Paris
January

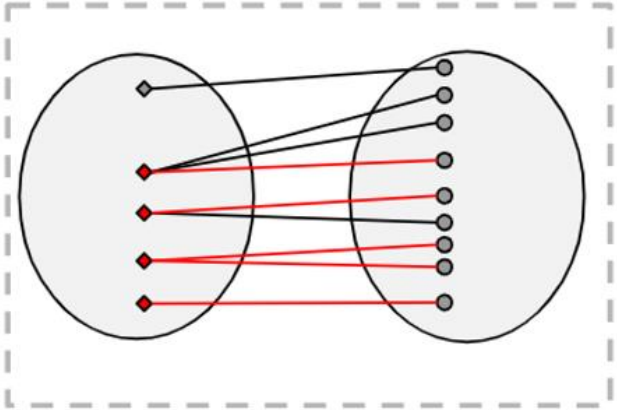

3D embedding space
z
x
y
hund
dog
Berlin
Berlin
Paris
Paris
januar
January

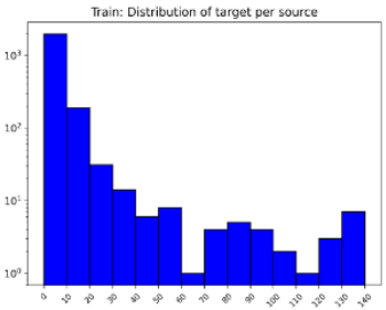
Ruder 2019

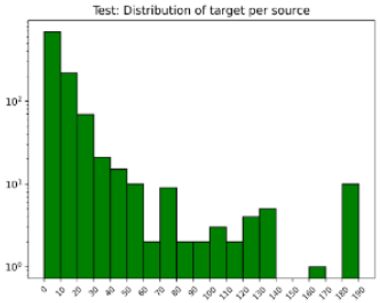
Illustration of Similarity



- Data
  - All source (PhenX) and target (dbGaP) entities are observed during training
    - Number of source entities: 2,549
    - Number of target entities: 9,311
  - Train/ Val/ Test split by linkage
    - Total number of links: 17,639
    - Split fractions:
      Train (70%) / Val (15%) / Test (15%)
    - Number of source entities in Test set: 1,058
      - Evaluate against all target entities
      - Evaluation will include links observed in train set







Manual curation of Phenx and dbGAP variables before July 1, 2018

Processing step

- Text preprocessing (lib BOW factory)
  - Replace camelcase : "camelCase” - "camel case"
  - Lower case: "LOWERCASE" - "lowercase"
  - Replace non alphanumeric: "alphanumeric@123" - "alphanumeric 123"
  - Collapse single character: "c e o" - "ceo"
  - Replace single character: "a” - "'
  - Remove extra space: "extra space" - "extra space"
  - Remove stop words: using nltk (post tokenization)
- Vectorization
  - Term frequency (sklearn CountVectorizer)
    - Remove stop words
    - Tokenize using sklearn tokenizer (must have at least 2 characters, ignore punctuation)
    - Consider only top 10K most frequent tokens
  - Apply global weighting using entropy (instead of idf)

Comparing performance of various classifiers

	swem	attention	rnn	hybrid	Sentence DistilBERT	BioBERT	DistilBERT
best_threshold	0.6906	0.5501	0.4752	0.4876	0.2917	0.4684	0.7410
f1_score	0.9503	0.9161	0.9859	0.9841	0.7531	0.5736	0.5479
precision_score	0.9177	0.8690	0.9775	0.9729	0.6488	0.4516	0.4278
recall_score	0.9854	0.9686	0.9944	0.9956	0.8975	0.7858	0.7620
auc	0.9964	0.9921	0.9988	0.9990	0.9639	0.8752	0.8577
average_precision	0.9751	0.9607	0.9962	0.9933	0.8824	0.7213	0.6585

MetaMatchMaker (M3) Neural Network (NN)

The processes of finding data (variables) and linking data (meta-data) are the same.

2 Tools

M3:Find – Find data

M3:Link – Connect Data

Use M3:Find to Make Finding Data Easy

M3:Find

Meta Match Maker
Science. Faster.

My List | Admin Test

All Sources

pregnancy

search

AND smoking

AND

AND

TAB or click to add term

Studies 5 | (x) Common Data Elements 0

Showing 5 / 5

Minimum Sample Size

0

Database

ALL

Gender

ALL

Study Type

Data Type

STUDY	SAMPLE SIZE	MATCHED VARIABLE COUNT	STUDY TYPE	SOURCE DATABASE	PUBLICATIONS	
Genetic Associations in Idiopathic Talipes Equinovarus (Clubfoot) - GAIT	1903	1	Multiplex Families, Parent-Offspring Trios	dbGap	PubMed	Add to list
Genome-Wide Association Studies of Prematurity and Its Complications (African American)	3478	3	Case-Control	dbGap	PubMed	Add to list

VARIABLE NAME	VARIABLE DESCRIPTION	VARIABLE SAMPLE SIZE	DATASET ACCESSION
site1/3_tobacco	Maternal smoking during pregnancy.	118	pht002254.v1
site1/3_tobacco	Maternal smoking during pregnancy.	16	pht002252.v1
site2_tobacco	Maternal smoking during pregnancy.	73	pht002253.v1

Coronary Artery Risk Development in Young	5115	1	Epidemiology Study	BioLINCC	PubMed	Add to list

Meta Match Maker
Science. Faster.

MetaMatchMaker Demo Video

- Demo: https://www.youtube.com/watch?v=WOW65-JH--s
- Promo: https://www.youtube.com/watch?v=nCM0qQ0aJBU

There Are Many More Scientific Databases

FITBIR
Federal Interagency Traumatic Brain Injury Research
INFORMATICS SYSTEM

Meta Data Providers

dbGaP
GENOTYPES and PHENOTYPES

BioLINCC
Biologic Specimen and Data Repository Information Coordinating Center
Web Access to NHLBI Biospecimens and Data

NATIONAL CANCER DATABASE

Environmental Dataset Gateway
EPA's official open data catalog
Advanced Search

GEO
Gene Expression Omnibus

United States Census 2020

Envirofacts
Your one-stop source for Environmental Information

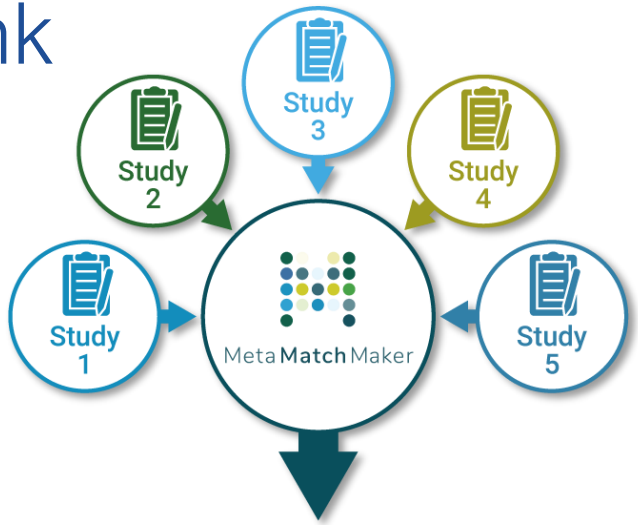
ImmPort
Private Data
Your site for managing ImmPort data uploads

VINCI
VA Informatics and Computing Infrastructure

Data Linkage

> “I wish there was a magic button that would automatically integrate my data for me.”
>
> *- NIH data steward*

M3:Link



Study1	Study2	Study3	Study4	Study5
Age	What age	Subj Age	Participant age	DOB
BMI	Height/weight	BMI	Height/Weight	NA
Heart Attack	Myocard. Inf.	Heart attack	Heart event	M. Infarc
Heart Attack	NA	NA	Heart attack	NA

Meta **Match** Maker

Science. *Faster.*

Predict Matches Between Two Datasets

Upload the source file and target file. Each file must contain a column with the name “text”.

If an input file has a column with the name "id", then the corresponding "id"'s for each comparison will be included in the results.

Source Bulk Upload

Drag and drop file here
Limit 200MB per file • XLSX, CSV
Browse files

Target Bulk Upload

Drag and drop file here
Limit 200MB per file • XLSX, CSV
Browse files

Filters +

Scoring

Matches are assessed based on the overall similarilty of the words used and their ordering. Scores range between 0 and 1, with scores closer to 1 indicating a better match. When two datasets are uploaded, all matches above .75 will be included in the output. The 60% true positive rate was observed at similarity scores of 0.98 or greater.

Note: Similarity scores are not meant to convey probability or confidence in the statistical sense. Instead, the sim. score is a mathematical representation of cosine similarity.

Introduction

MetaMatchMaker uses a machine learning (ML) approach to facilitate rapid mapping of common data elements between two datasets. Analysts and coders currently spend hudreds of hours mapping common data elements, and while it is important to keep humans in the loop we think that machine learning can significantly reduce the time and effort to do this.

http://34.226.96.161:8501

M3 identifies the best match

phenx_text	phenx_meta	dbgap_text	dbgap_meta	score	curation (T/F)
Blood pressure (systolic)	['PX09060224C	Seated Systolic Blood Pressure	['phs000007.v13_pht00	0.997006953	T
Blood pressure (systolic)	['PX09060224C	Seated systolic BP- first reading	['phs000007.v13_pht00	0.995389104	T
Blood pressure (systolic)	['PX09060224C	Seated systolic BP- second reading	['phs000007.v13_pht00	0.992073596	T
Blood pressure (systolic)	['PX09060224C	Seated systolic BP- third reading	['phs000007.v13_pht00	0.995140433	T
Blood pressure (systolic)	['PX09060224C	Systolic blood pressure entered by technician	['phs000007.v13_pht00	0.99785912	T
Blood pressure (systolic)	['PX09060224C	TECHNICIAN SYSTOLIC BLOOD PRESSURE (TO NEAREST 2MM H	['phs000007.v13_pht00	0.991444349	T
Blood pressure (systolic)	['PX09060224C	TECHNICIAN'S SYSTOLIC BLOOD PRESSURE (TO NEAREST 2MM	['phs000007.v13_pht00	0.994368553	T
CIRCUMSTANCES OF HIP FRACTURE:	['PX17090110C	Circumstances of hip fracture	['phs000007.v13_pht00	1.000000477	T
CTR ID Number	['PX19040104C	DUMMY ID NUMBER	['phs000001.v2_pht000	1	F
CTR ID Number	['PX19040104C	FHS Particicpand SHARe ID Number	['phs000007.v13_pht00	0.990215898	F
CTR ID Number	['PX19040104C	Phantom ID Number	['phs000007.v13_pht00	1	F
CTR ID Number	['PX19040104C	SHARE ID NUMBER	['phs000007.v13_pht00	0.99114418	F
CTR ID Number	['PX19040104C	SHARE ID Number	['phs000007.v13_pht00	0.99114418	F
CTR ID Number	['PX19040104C	SHARe ID Number	['phs000007.v13_pht00	0.99114418	F
CTR ID Number	['PX19040104C	SHARe ID number	['phs000007.v10_pht00	0.99114418	F
CTR ID Number	['PX19040104C	Share ID Number	['phs000007.v13_pht00	0.99114418	F
CTR ID Number	['PX19040104C	Share ID number	['phs000007.v13_pht00	0.99114418	F
Check all that apply: Lists of body weight	['PX65050165C	BODY WEIGHT	['phs000007.v13_pht00	0.991760135	T
Chest Discomfort Characteristics Radiation	['PX04060109C	CHEST DISCOMFORT CHARACTERISTICS	['phs000007.v13_pht00	0.999451756	T
Chest discomfort when quiet or resting	['PX04060104C	CHEST DISCOMFORT WHEN QUIET OR RESTING	['phs000007.v13_pht00	0.999999881	T
Chest discomfort when quiet or resting	['PX04060104C	CHEST DISCOMFORT WHEN QUIET OR RESTING?	['phs000007.v13_pht00	0.999999881	T
Chest discomfort when quiet or resting	['PX04060104C	CHEST DISCOMFORT: WHEN QUIET OR RESTING	['phs000007.v13_pht00	0.999999881	T
Chest discomfort with exertion or exciteme	['PX04060103C	CHEST DISCOMFORT WITH EXERTION	['phs000007.v13_pht00	0.990475237	T
Chest discomfort with exertion or exciteme	['PX04060103C	CHEST DISCOMFORT WITH EXERTION OR EXCITEMENT	['phs000007.v13_pht00	0.99999994	T
Chest discomfort with exertion or exciteme	['PX04060103C	CHEST DISCOMFORT WITH EXERTION OR EXCITEMENT?	['phs000007.v13_pht00	0.99999994	T
Chest discomfort with exertion or exciteme	['PX04060103C	CHEST DISCOMFORT WITH EXERTION/EXCITEMENT	['phs000007.v13_pht00	0.996473312	T

Too much promise, not enough delivery.

> “I’m tired of being sold on yet another AI tool. Show me don’t tell me.”
>
> *- VA data steward*

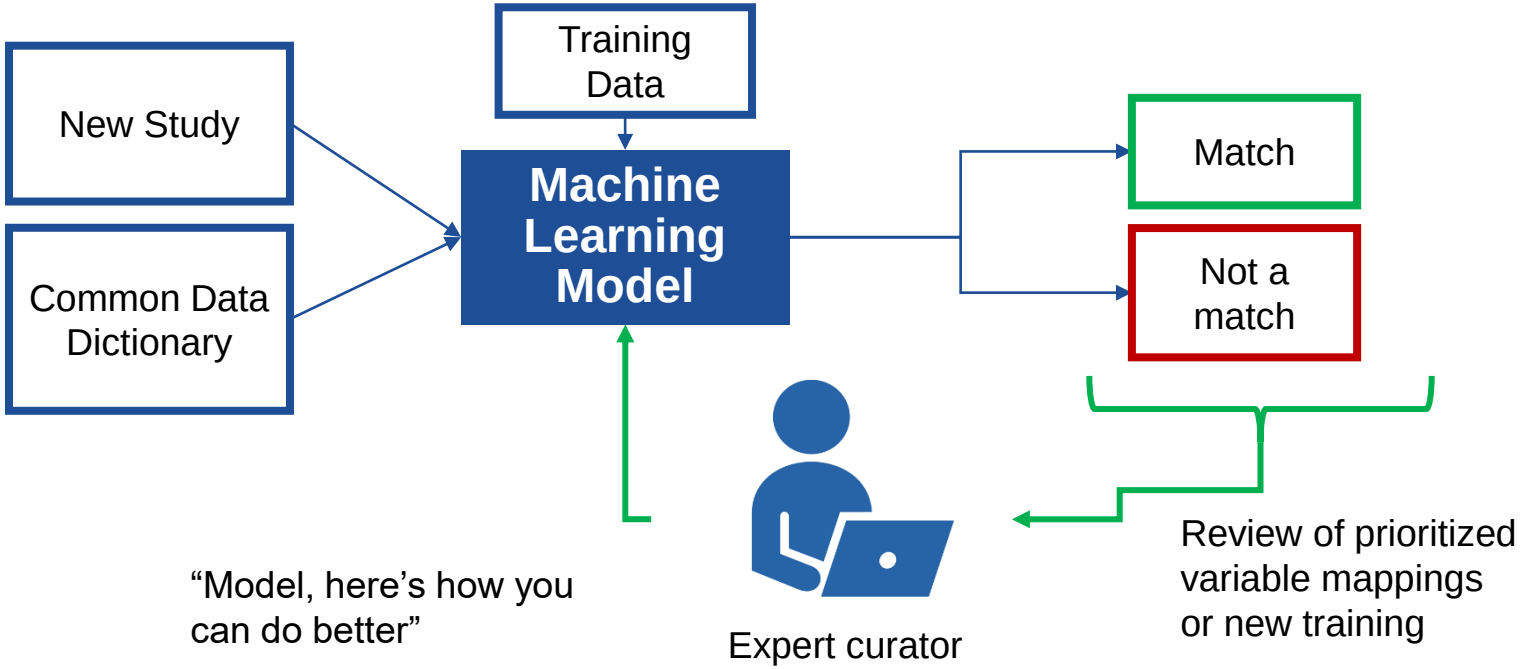
3 different use cases for M3:Link

- Merging 2 or 3 databases together to develop a common database
  - LADDER – Angelman's syndrome
- Linking multiple database to each other
  - DBGAP-PHENX mapping
- Merging multiple databases onto a single database. Submission of data to a repository.
  - PASC - RECOVER - COVID

Not everyone agrees on what match is

Map variables with different wording but the same/similar concept (includes similar time period concepts).
Map variables with a parent-child (or child-grandparent) relationship.
Map cases where one variable is more descriptive/specific than the other. Including if one variable contains more information than the other.
Map variables where the measurements are the same, but one variable is restricted to a specific context.
Map variables that are the same measurement, but a different method.
Map age to age variables when there are no qualifiers.
Map gender and sex variables to each other unless they are relevant to the protocol topic (gender identify measure, etc.)
Map weight to other weight variables and height to other height variables unless they are not referring to standard body height or weight (i.e., don't map knee height or change in weight).
In the case of longitudinal measurements repeated at different timepoints, keep all measurements.
In the case of repeat measurements for the same time point, map the average of the measurements (if available), otherwise map the first measurement. The rest are redundant and should be N/A.
Do not map variables that ask similar questions, but the concept is different.
Do not map variables with a sibling relationship.
Do not map telehealth and medical care/treatment variables (telehealth and medical care are not equivalent).
Do not map administrative variables. Mark these as NA.
Do not map open ended questions to more specific questions.
Do not map age and date variables unless they are mappings of age to age without any qualifiers. Mark these as NA.
Do not map "other" variables in most cases. Mark these as NA.
Do not map "comment", "specify", "remarks" variables. Mark these as NA.
Do not map "score" variables. Mark these as NA
Do not map medical history to current medical conditions

AI can be retrained in an ‘active learning’ process with humans developing, training, and reviewing results.



Linking Angelman and Dup15q Data for Expanded Research (LADDER)

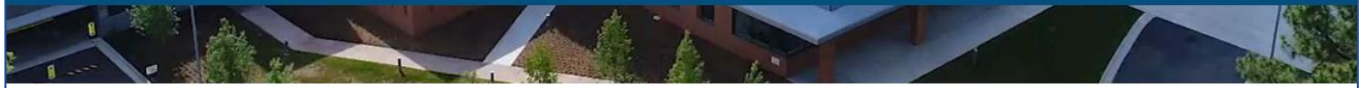
angelman SYNDROME FOUNDATION | FAMILIES. RESEARCH. CLINICS. COMMUNITY. WITH YOU FOR THE JOURNEY.

WALK | CONTACT REGISTRY | NEWS | BLOG

Search ...

DONATE

WHAT IS ANGELMAN SYNDROME | FOR PARENTS | RESEARCH | MAKE AN IMPACT | CLINICS | RESOURCES | EVENTS | ABOUT



ARTICLES AND ANNOUNCEMENTS

Ladder Database

- ASF TO HOST SCREENING OF CAREGIVER DOCUMENTARY, UNSEEN
- ANGELMAN SYNDROME FOUNDATION AND DUP15Q ALLIANCE PARTNER TO LAUNCH THE LADDER LEARNING NETWORK, SHOWING CONTINUED DEDICATION TO RESEARCH
- NEWBORN SCREENING GRANT FOR EARLY CHECK
- ASF FUNDED RESEARCH ON GENE THERAPY IN AS PUBLISHED IN JCI INSIGHT

LADDER
Linking Angelman and Dup15q Data for Expanded Research

NOVEMBER 25, 2019

LADDER DATABASE

A new network platform, Linking Angelman and Dup15q Data for Expanded Research (LADDER), will offer caregivers access to information about a larger population of patients living with Angelman or Dup15 syndromes to better inform their decisions regarding treatments and interventions.

RTI International (RTI), a nonprofit research institute, together with Angelman Syndrome and Dup15q Patient Advocacy Groups and the Clinic Network will develop and host the clinic network database to be shared across multiple national and international sites. This collaborative effort is funded by the Angelman Syndrome Foundation (ASF) and the Dup15q Alliance.

"The database will establish a global network of patient-powered data that will be used to improve care for people living with Angelman or Dup15q syndrome," said Anne Wheeler, PhD

M3:Link identifies the best match

source term	target term 1	target score 1	target term 2	target score 2	target term 3	target score 3
RepetitivePickRubOfObjAtAge Repetitive picking/rubbing of objects, At what age	Bathes or showers assist Age Unk Bathes or showers assist Age Unk bathes_showers_assist_ageU	0.98	Age Age PLS5_age	0.98	Gest age microceph detected Gest age microceph detected gest_age_microceph_detected	0.98
BirthFOCCentile Birth FOC centile (%)	Birth HC centile Birth HC centile birth_HC_centile	1.00	Birth History Birth History birth_hx_ttl1	0.99	Sibling Year of Birth Sibling Year of Birth sibling_year_of_birth_8	0.96
SleepDifficulties Sleep History Sleep difficulties	Sleep History_header Sleep History_header sleep_history_header	0.99	Sleep difficulties Sleep difficulties sleep_difficulties	0.99	Sleep Latency (minutes) Sleep Latency (minutes) sleep_latency_min	0.97
BPdiastolic BP Diastolic	Diastolic Blood Pressure Diastolic Blood Pressure DBP	1.00	Murmur diastolic Murmur diastolic cardiac_murmur_diastolic	0.97		
FollowSimpleCommWOGestureAge Receptive Language Follows simple command without gesture Age (Months)	Follows one step command without gesture age (Months) Follows one step command without gesture age (Months) follows_wo_gesture_age_mths	0.98	Puts on any clothing without support age (Months) Puts on any clothing without support age (Months) puts_on_clothing_age_mths	0.98	Balances without support age (calculated months) Balances without support age (calculated months) balances_1foot_wo_sup_age_calc	0.98

Mapping multiple and/or large datasets onto each other.

- The PhenX Toolkit provides well-established, high-quality, low-burden measurement protocols intended for use in biomedical, epidemiological, clinical, translational, and genomic research.
  - Currently 900 measurements with 9,148 items.
  - There is redundancy in the measurements.
- Mapping PhenX terms to dbGAP metadata.
  - dbGAP is a database of genomic data with some clinical variables. 70,945 non genetic terms submitted since June 1, 2016.
  - De-duplication of meta-data terms in one database/set of dataset is important.

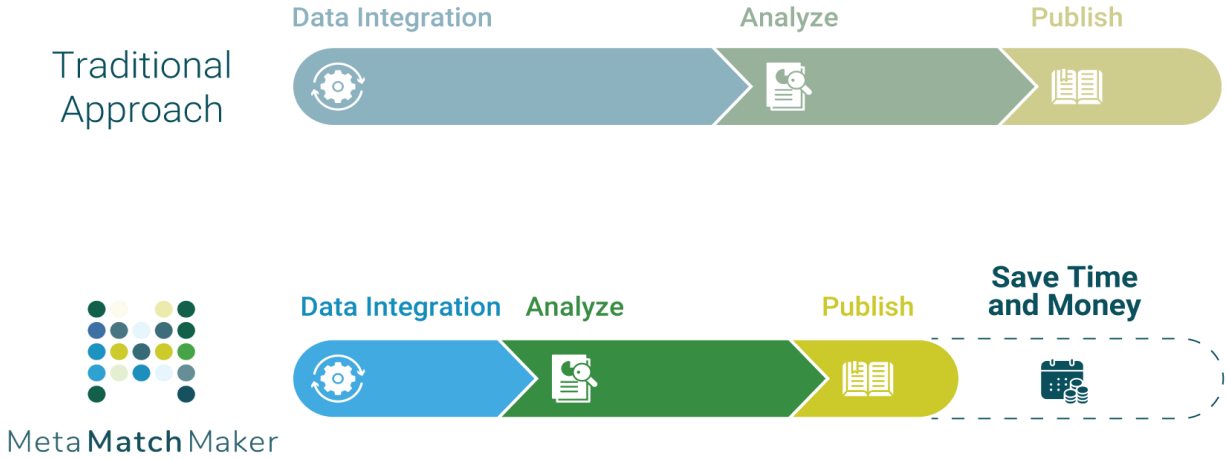
Example – dbGAP –PhenX Mapping

Jan 2022 M3: Run	# after initial filter
PhenX variables	9,148
dbGaP variables	70,945
Full results for similarity score >= 0.9	6,491,435
filter NHLBI none included in training model	1,756,816
Remove admin and other variables	1,270,037
Full results for similarity score >= 0.95	437,872
99-100	32,167
98-99	62,790
97-98	88,125
96-97	119,016
95-96	135,774

- Time to run < 1 hours
- 32,167 links had a score > 0.99.
- Manual review revealed 79% true positive
- Use the negatives to improve the model
- Working on time estimate of savings, at least 50% reduction in effort

Summary

- M3:Find can be used to make finding data easier.
- M3:Link can be used to reduce the time and effort to link metadata. M3:Link can be easily retrained to expand the vocabulary and for different decisions on what is considered matching.



Tech Talk

Questions and Discussion

Grier Page
gpage@rti.org
205-873-0669

Eric Earley
eearley@rti.org
919-541-5601

RTI INTERNATIONAL

MetaMatchMaker
Science. Faster.