

Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models

**Juan M. Banda, Martin Seneviratne, Tina Hernandez-Boussard,
and Nigam H. Shah**

Jianhui Gao presented at EHR reading group in August, 2021

Electronic Phenotyping

Process of identifying patients with certain characteristics of interest

Table 1 Applications of electronic phenotyping across study types

Study type	Use cases
Cross-sectional	Epidemiological research
	Hospital administration/resource allocation
	Adherence to diagnostic/treatment guidelines
	Quality measurement
Association (case-control/cohort)	Genome-wide association studies
	Pharmacovigilance
	Identifying clinical risk factors and protective factors
	Clinical decision support
	Clinical effectiveness research
	Predictive modeling
Experimental	Clinical trial recruitment
	Pragmatic trials
	Adaptive/randomized, embedded, multifactorial, adaptive platform trials

Challenges

- Structured data and unstructured data in EHR
- Practices of EHR input differ by site
- Accuracy of content entered by clinicians

Phenotyping is far more challenging than a simple code search and requires sophisticated methods that can account for heterogeneity

Paper Selection

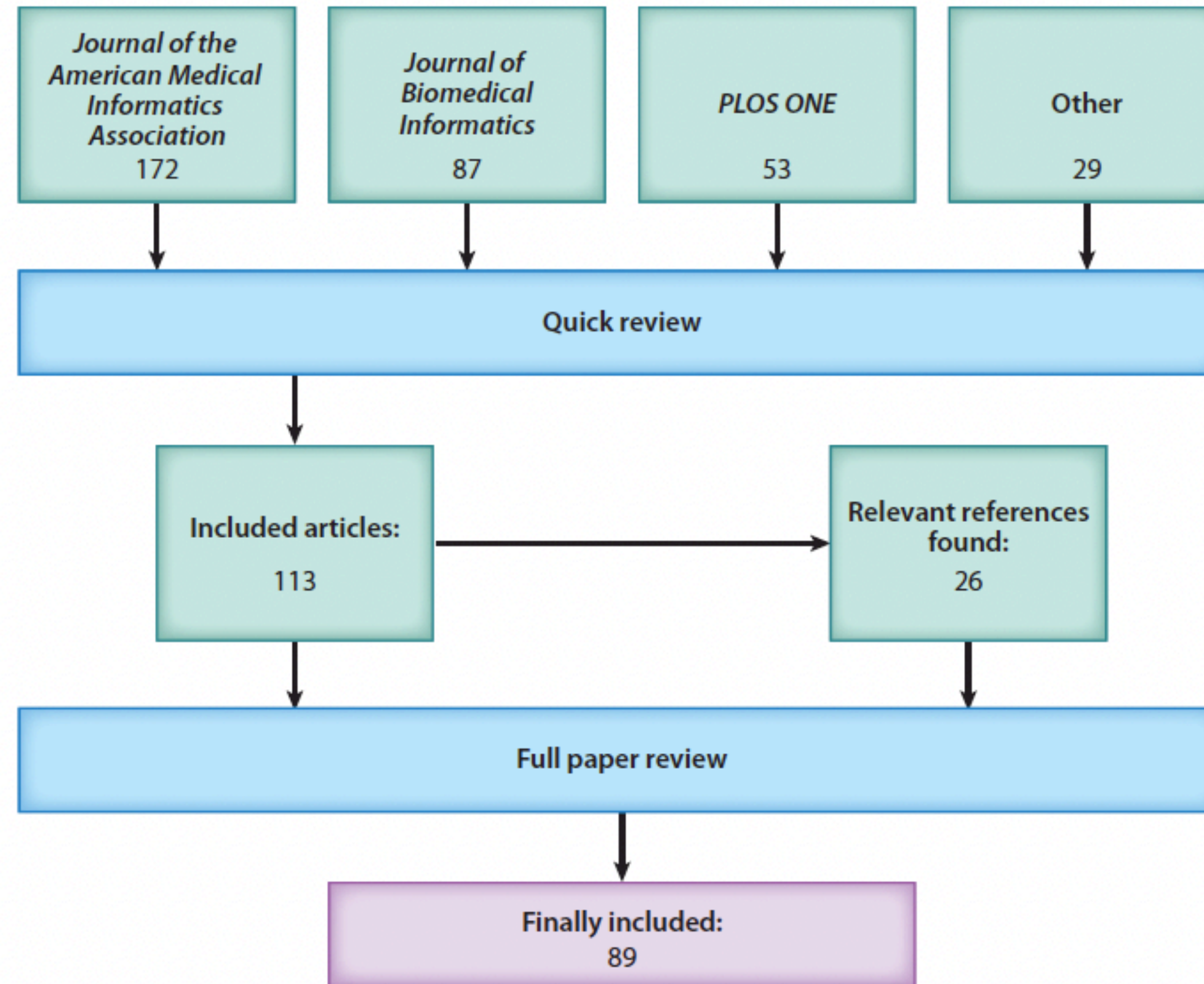



Figure 1

Paper selection process outline. Combining a PubMed query from major journals with Google Scholar alerts and other recommendations, we narrowed down the papers we reviewed in a two-step process. First, we reduced the initial number of papers (numbers inside each box in row 1) via a quick paper review. We then conducted a more thorough paper review before settling on 89 papers to discuss in this review.

Rule-based methods

- Inclusion and exclusion criteria based on **structured data**
- Example: type 2 diabetes
 - Mention of the diagnosis code
 - At least one hypoglycemic medication
 - HbA1c above certain threshold
- work well for phenotypes that have clear diagnosis and procedure codes.

Phenotype Knowledgebase (PheKB)

Public Phenotypes

Public

Collaboration

Public phenotypes are believed to be complete and final by their authors. When you are logged in you can view and edit phenotypes in your groups that are non public and in various stages of development.

Login To View Private Group Phenotypes

Institution

Type of Phenotype






Owner Phenotyping Groups

View Phenotyping Groups

Data Model

- Any -

Apply

Title	Institution	Phenotype Attributes	Owner Phenotyping Groups	View Groups	Has new content	Status	Type
 Abdominal Aortic Aneurysm (AAA)	Geisinger	CPT Codes, ICD 9 Codes, Vital Signs	eMERGE Geisinger Group	eMERGE Geisinger Group, eMERGE Phenotype WG		Final	Disease or Syndrome
 ACE Inhibitor (ACE-I) induced cough	Vanderbilt University	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE Vanderbilt Group	eMERGE Phenotype WG		Final	Drug Response - adverse effect or efficacy
 ADHD phenotype algorithm	CHOP	ICD 9 Codes, Medications, Natural Language Processing	eMERGE CHOP Group	eMERGE Phenotype WG		Final	Disease or Syndrome
 Anxiety algorithm	CHOP	CPT Codes, ICD 10 Codes, ICD 9 Codes, Medications	eMERGE CHOP Group	eMERGE CHOP Group, eMERGE Phenotype WG		Final	Disease or Syndrome
 Appendicitis	Cincinnati Children's Hospital Medical Center	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE CCHMC/BCH Group	eMERGE Phenotype WG		Final	Disease or Syndrome

The majority of which are rule-based and use structured data

Short Discussion

Why rule-based method is likely to be sub-optimal?

- lack of portability between phenotypes and health systems
- Complex diseases
- The quality of structured data
- What else?

Unstructured Data

- clinical notes
- discharge summaries
- radiology
- pathology reports
- contain a wealth of phenotypic information
- represent approximately 80% of data captured in EHRs

Nature Language Processing (NLP)

- Predominantly statistical: can extract contextual cues such as causality and temporality
- Example use of NLP in clinical data
 - extract coded concepts from radiology reports
 - extract respiratory diagnoses and smoking statuses from discharge summaries
 - extract adverse outcomes following drug exposures

Machine Learning(ML) Methods

Standard ML Overview

FSSMC (feature selection via supervised model construction)

Huang et al., 2007

a cohort of diabetic patients and controls

Manually selected 47 features from 410 structured variables

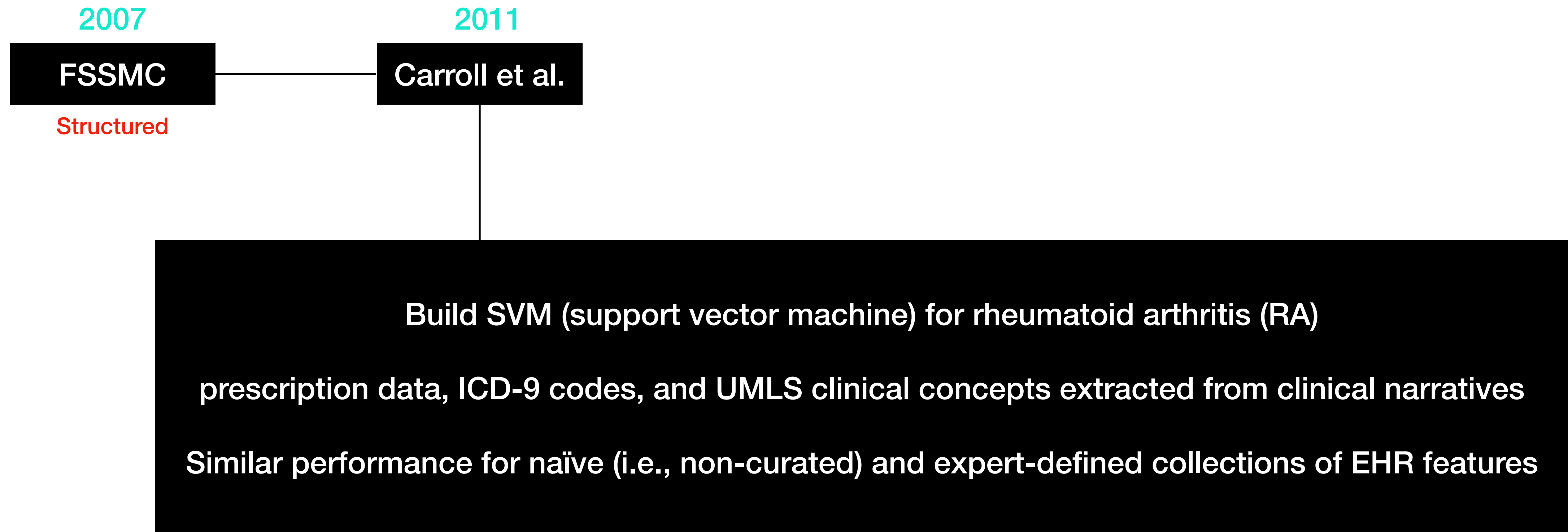
FSSMC was used to rank the top 15 features

3 classifiers: naive Bayes, C4.5, IB1

A best predictive accuracy of 95% and sensitivity of 98% was achieved

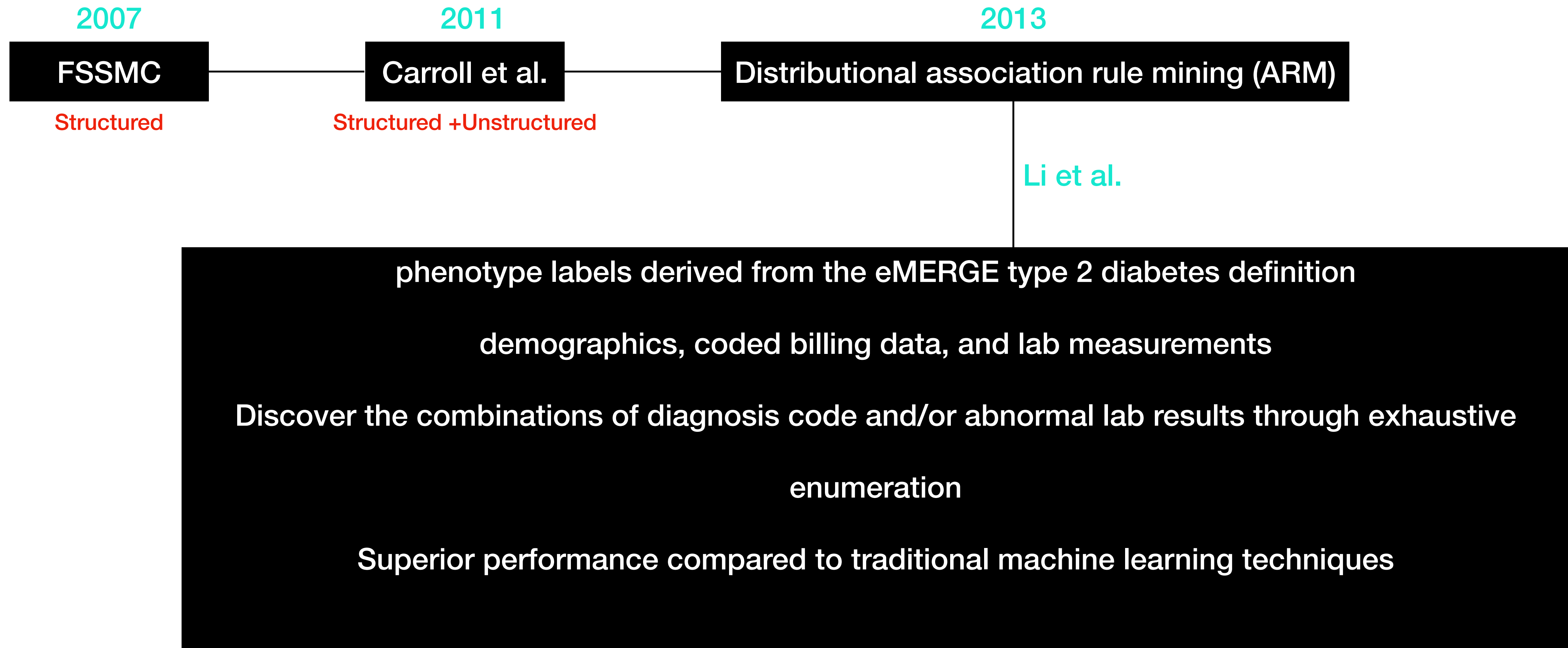
Machine Learning(ML) Methods

Standard ML Overview



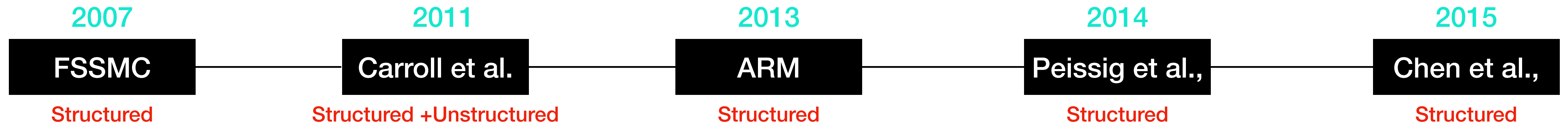
Machine Learning(ML) Methods

Standard ML Overview



Machine Learning(ML) Methods

Standard ML Overview



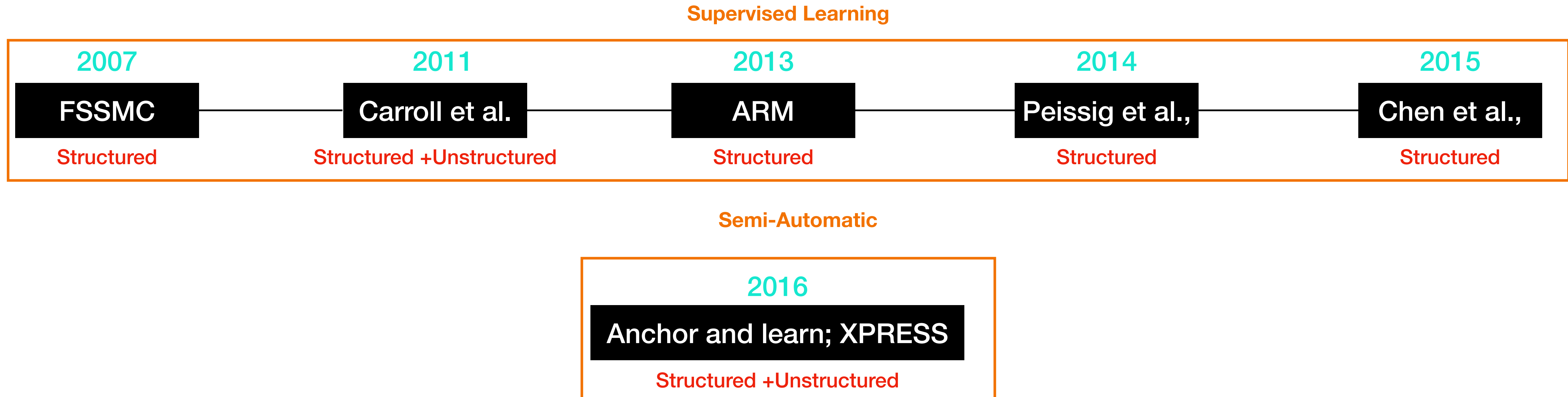
All require manually labeled gold standard training and test data sets for model building and validation

Machine Learning(ML) Methods

Learning with Noisy Data

- Assumption: A large volume of training data should compensate for inaccuracies in the labels
- Anchor variables: highly informative, clinically relevant features for a specific phenotype.
- Examples:
 - Anchor-based learning (Halpern et al., 2014; Halpern et al., 2016)
 - XPRESS (extraction of phenotypes from records using silver standards) (Agarwal et al., 2016)

Machine Learning(ML) Methods



Initial set of anchors or noisy labels is identified by an expert in order for the phenotype model to be relevant.

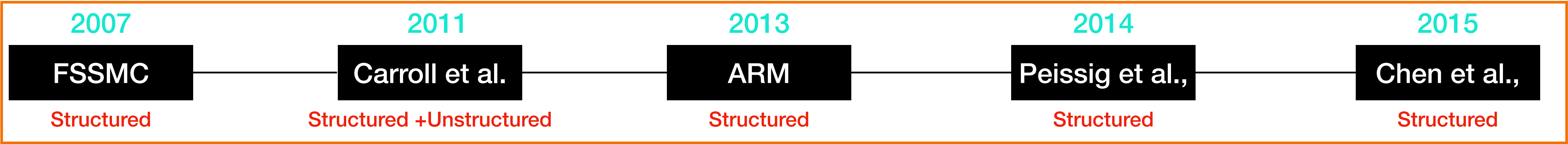
Machine Learning(ML) Methods

Unsupervised Phenotype Discovery

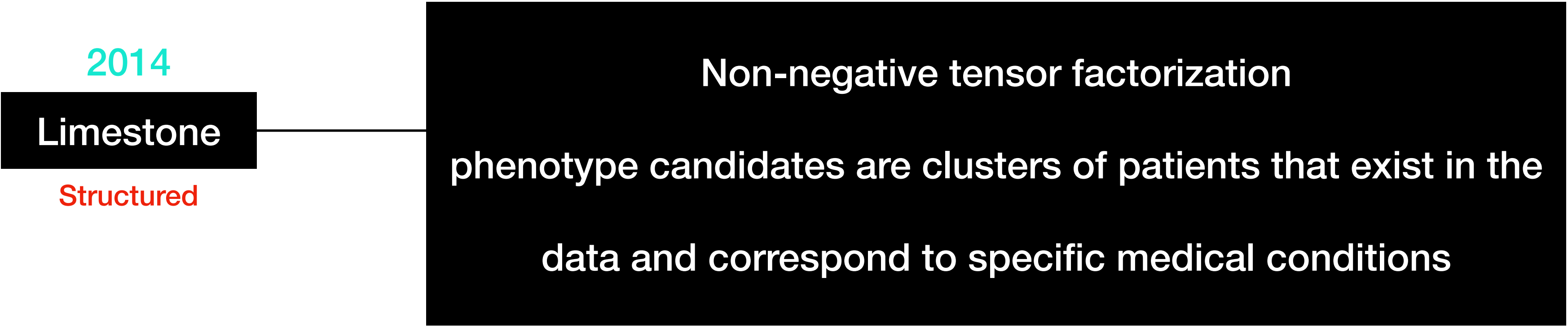
- Generate phenotypes with minimal human supervision
- Ho et al. defined “ideal” phenotype:
 - (a) A phenotype represents complex interactions between several features (e.g. diagnosis and medication)
 - (b) the definition should be **concise and understandable** by a medical professional
 - (c) the definition can be translated into new domain knowledge.

Machine Learning(ML) Methods

Supervised Learning

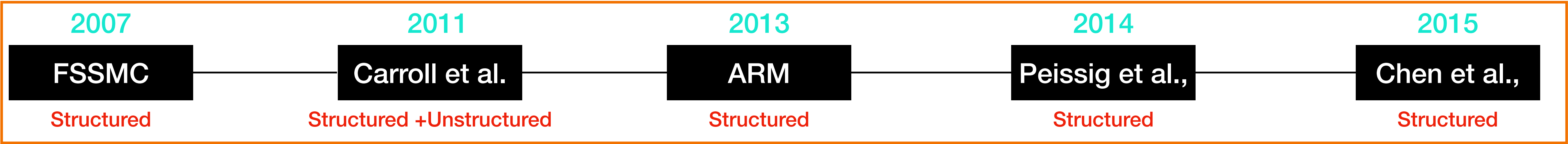


Semi-Automatic



Machine Learning(ML) Methods

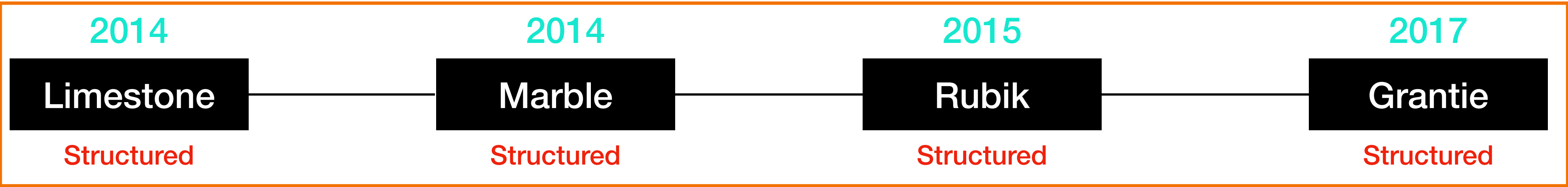
Supervised Learning



Semi-Automatic



Unsupervised learning



Machine Learning(ML) Methods

Hybrid

- AFEP (automated feature extraction for phenotyping)
 - leveraging publicly available data sources of medical knowledge like Medscape and Wikipedia
 - produces a list of UMLS concepts that are proposed as features to use
 - further refined by only keeping the concepts found in clinical notes of EHRs
- SAFE (surrogate-assisted feature extraction)
 - Five data source
 - incorporating the idea of imperfect labeling

Future Directions

- Accurate phenotyping using EHRs is necessary
- Complementing EHRs with other types of data, including registries, wearable feeds, multiomic data, imaging, and patient-reported outcomes
- Validating unsupervised phenotype definitions
- explore representation learning to reduce the burden of feature engineering, as well as to investigate portable methods for feature engineering across phenotypes
- expansion of these collaborative networks, as phenotyping tools stand to benefit from shared data and diverse test sites.