# ARC v2

Leonard D’Avolio PhD & Thien Nguyen

MAVERIC

# Introduction

Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC)

About 140 people focused on:

- Large scale clinical trials (CSP CC)
- Epidemiology (CSP ERIC)
- Biospecimen Repository
- Informatics

Projects include:

- Million Veteran Program & GenISIS
- Point of Care Clinical Trials

# Background

Growing need for access to EMR data for secondary uses

- Quality measurement & improvement
- Comparative effectiveness
- Evidence-based medicine
- Bio-surveillance
- Cohort & registry building
- Personalized medicine & genomic analyses

# The Challenge

- EMRs were designed for 1-on-1 interactions
- As a result:
  - Few widely implemented standards
  - Questionable quality of these few
  - Large amounts of free text
    - Estimated 70% of information in free text

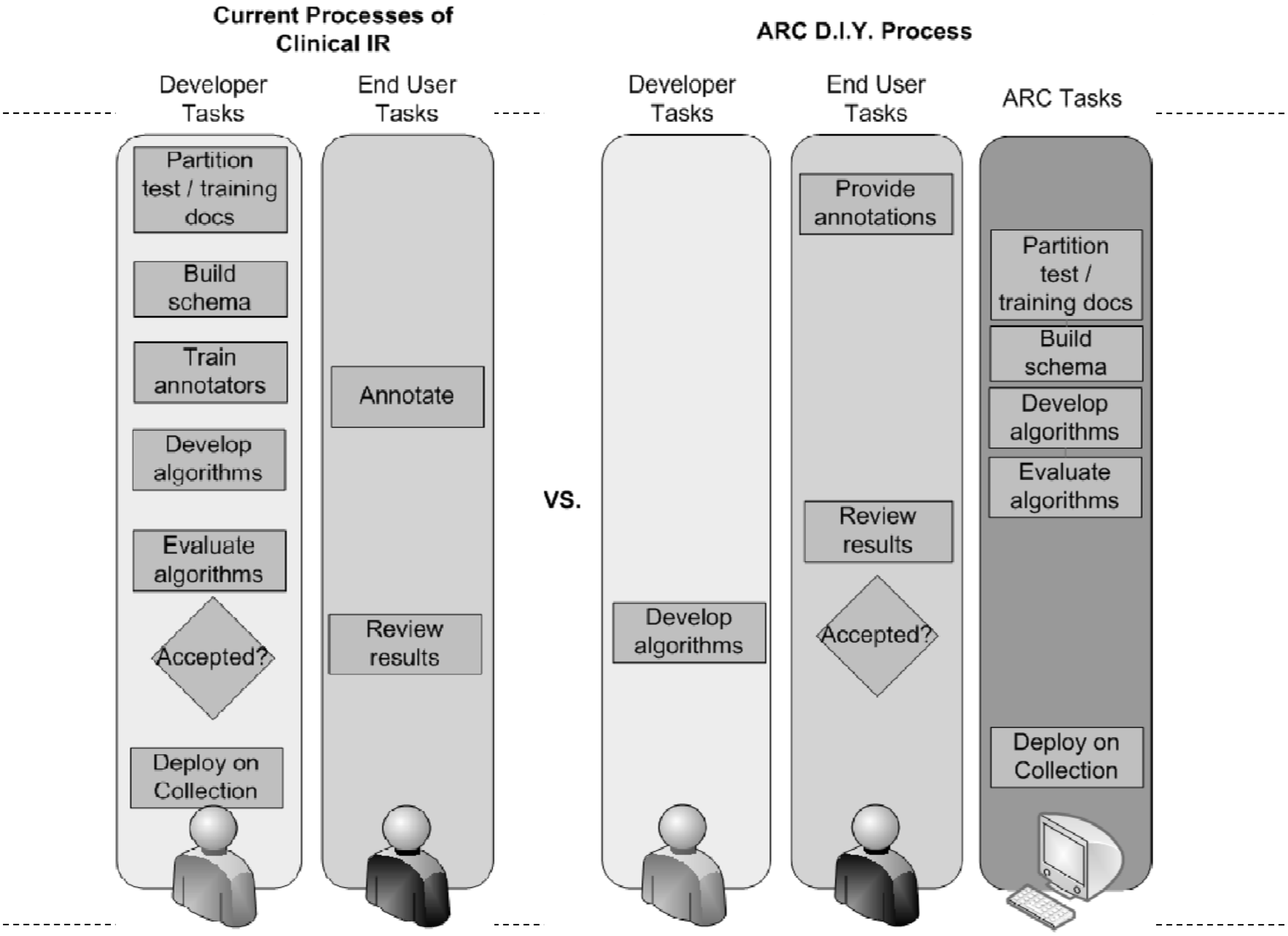
- In response – 40+ years of 1-off NLP solutions
  - 113 studies = no systems nor results are generalizable
    - Stanfill et al. 2010. A systematic literature review of automated clinical coding and classification systems. JAMIA 17: 646-651

# The Challenge (cont.)

- Of the few implemented systems, most are:
  - Applied for specialized applications
  - At institutions of origin
  - Heavily reliant on the systems’ developers
- Why?
  - Complex nature of the challenge
  - Many processes involved
  - Lots of customization
  - Economics of research

# The Idea

- Change the workflow so that researchers don’t have a service-oriented dependency on software developers
- Move away from rule-based systems to do information retrieval
- Take advantage of 20+ years of empirical evidence and open source code
- Use existing NLP frameworks to allow for other developers to extend and modify the tool
- 90/90 goal using generalized approach



## Current Processes of Clinical IR

Developer Tasks

- Partition test / training docs
- Build schema
- Train annotators
- Develop algorithms
- Evaluate algorithms
- Accepted?
- Deploy on Collection

End User Tasks

- Annotate
- Review results

VS.

## ARC D.I.Y. Process

Developer Tasks

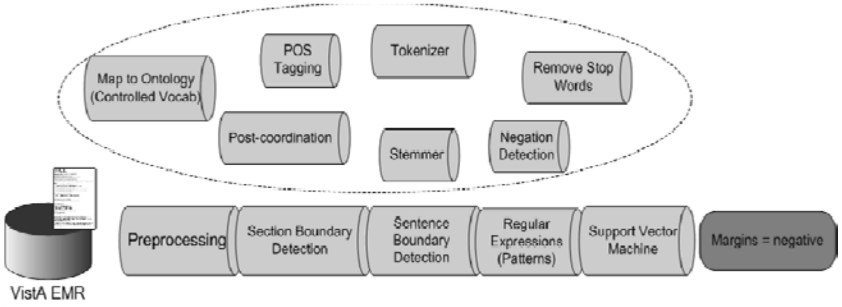
- Develop algorithms

End User Tasks

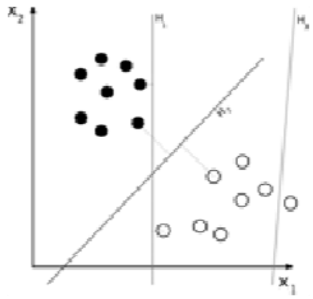
- Provide annotations
- Review results
- Accepted?

ARC Tasks

- Partition test / training docs
- Build schema
- Develop algorithms
- Evaluate algorithms
- Deploy on Collection



+



+

- Import annotations (eHOST and Knowtator)
- Use existing NLP pipelines (cTAKES)
- Combine with machine learning (MALLET)
- Provide a simple workflow taking advantage of powerful tools

## Document Retrieval

| | Recall | Precision | F-Measure |
|---|---|---|---|
| **Prostate Cancer Path Reports** | 0.97 | 0.95 | 0.94 |
| **Colorectal Cancer Path Reports** | 0.90 | 0.92 | 0.89 |
| **Lung Cancer Imaging Reports** | 0.76 | 0.80 | 0.75 |
| **PTSD Psychotherapy Notes (seeking PE)** | 0.72 | 0.93 | 0.81 |
| **PTSD Psychotherapy Notes (seeking CBT)** | 0.86 | 0.98 | 0.91 |
| **Breast Cancer Path Reports (from 130 hospitals)** | 0.95 | 0.95 | 0.94 |
| **Breast Cancer Clinic Notes** | 0.96 | 0.88 | 0.92 |
| **Breast Cancer Operative Reports** | 0.91 | 0.85 | 0.87 |
| **Pneumonia Imaging Reports** | 0.80 | 0.81 | 0.80 |

| Concept Retrieval (inexact span matching) | | | |
|---|---|---|---|
| | **Recall** | **Precision** | **F-Measure** |
| **2010 i2b2/VA Medical Problems** | 0.75 | 0.93 | 0.83 |
| **2010 i2b2/VA Medical Treatments** | 0.76 | 0.89 | 0.82 |
| **2010 i2b2/VA Medical Tests** | 0.76 | 0.90 | 0.83 |

- 2009 cTAKES with minor changes and a couple of additional annotators (bumper and sentence)
- 2008 LVG
- 2008 UMLS
- Out-of-the box we were on par with average entrants’ performance
    - With less than 5 minutes of human time per task

# What we learned

- Works well on document classification tasks
    - “Find more cases like these”
- Use existing structured data to narrow sets
    - Known ICD-9’s + RXs then use ARC
- You need to understand how the process works
    - Training set, test set, performance metrics
- Know what is “acceptable” & estimated prevalence before you start

# Demo

Do-It-Yourself Interface

- >150 downloads
- Google group, video tutorials, dummy data sets, etc
- http://maveric.org/mig/arc.html
- http://code.google.com/p/mavericarc/
- http://groups.google.com/group/ClinicalNLP
- Available on VINCI
- Installer package via iDASH
- iDASH NLP VM also has ARC
- Thanks to HSR&D Consortium for Healthcare Informatics Research (CHIR)