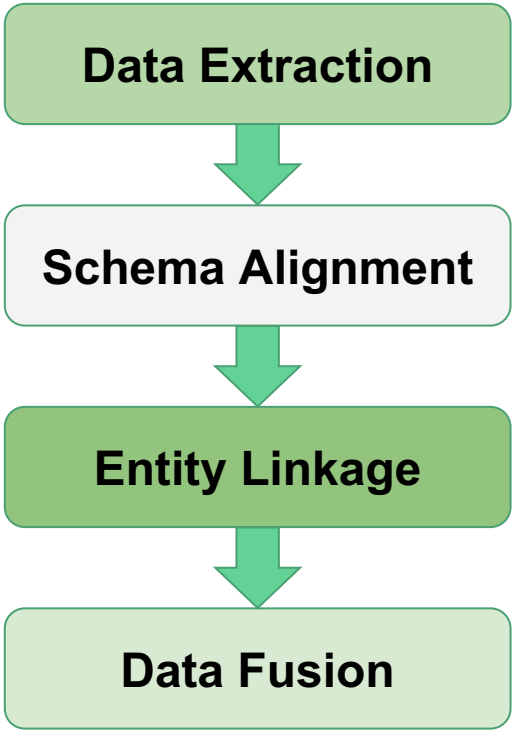
Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
- Part IV. Conclusions and research directions

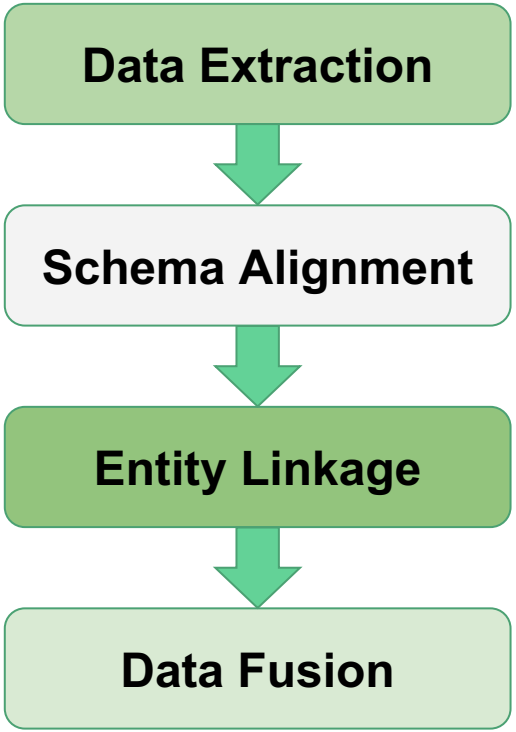
Data Integration Overview

- Entity linkage: linking records to entities; indispensable when different sources exist
- Data extraction: extracting structured data; important when non-relational data exist
- Data fusion: resolving conflicts; necessary in presence of erroneous data
- Schema alignment: aligning types and attributes; helpful when different relational schemas exist



Recipe

- Problem definition
- Brief history
- State-of-the-art ML solutions
- Summary w. a short answer

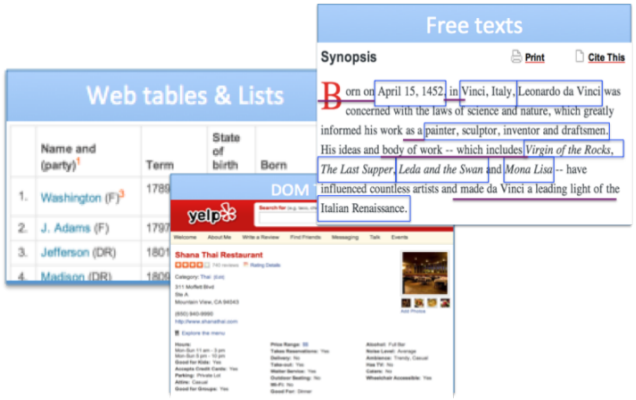


Theme I. Which ML Model Works Best?

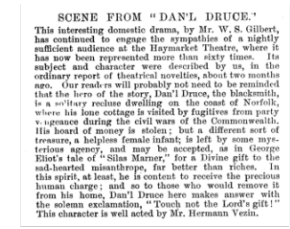
Which ML Model Works Best?

ID	NAME	CLASS	MARK	SEX
1	John Deo	Four	75	female
2	Max Ruin	Three	85	male
3	Arnold	Three	55	male
4	Krish Star	Four	60	female
5	John Mike	Four	60	female
6	Alex John	Four	55	male
7	My John Rob	Fifth	78	male
8	Asruid	Five	85	male
9	Tes Qry	Six	78	male
10	Big John	Four	55	female

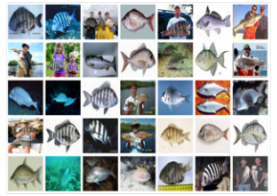
Tree-based models



??


SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant; is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.



Neural network

Theme II. Does Supervised Learning Apply to DI?

- **Supervised learning has made a big splash recently in many fields**

- **However, it is hard to bluntly apply supervised learning to DI tasks**
  - Our goal is to integrate data from many different data sources in different domains
  - The different sources present different data features and distributions
  - Collecting training labels for each source is a huge cost