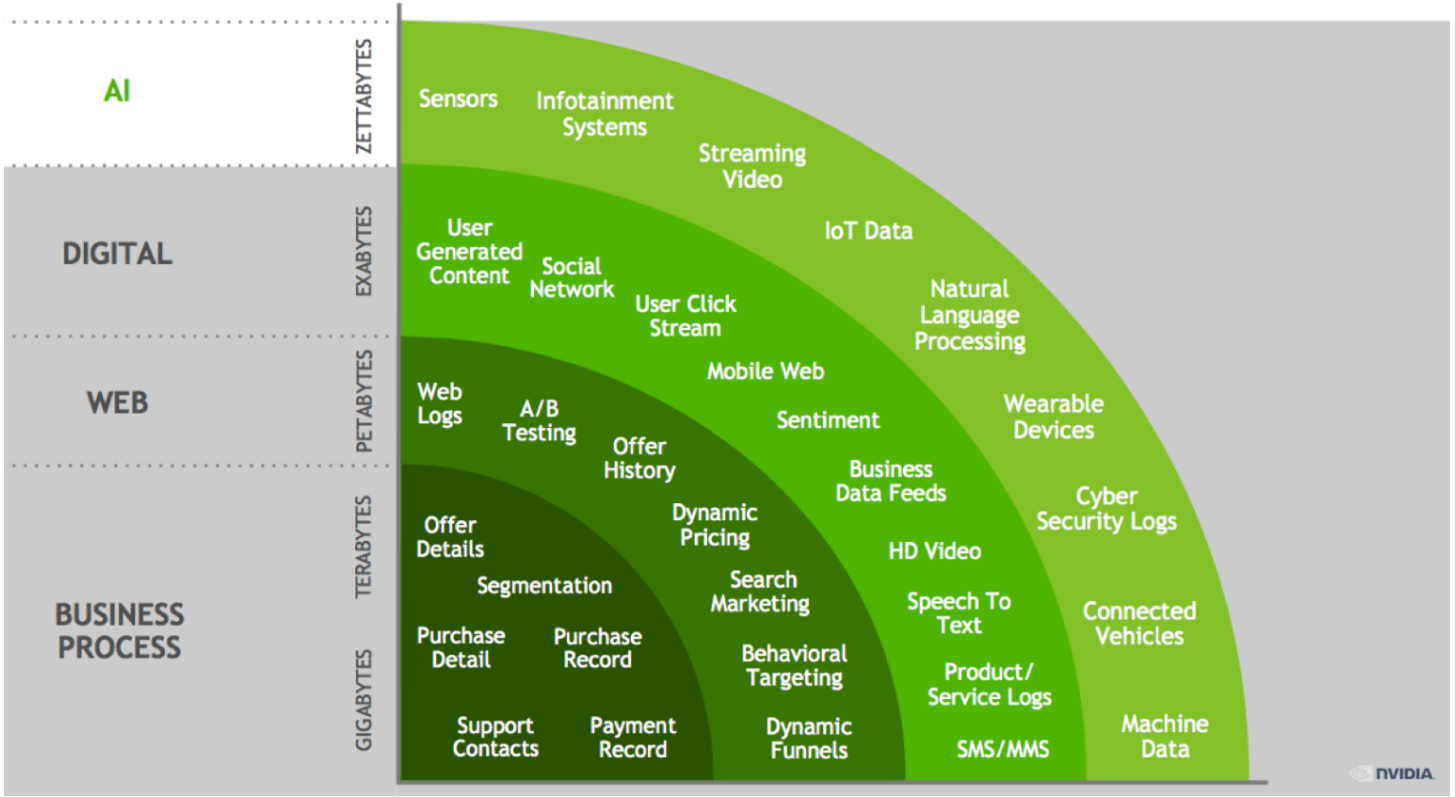
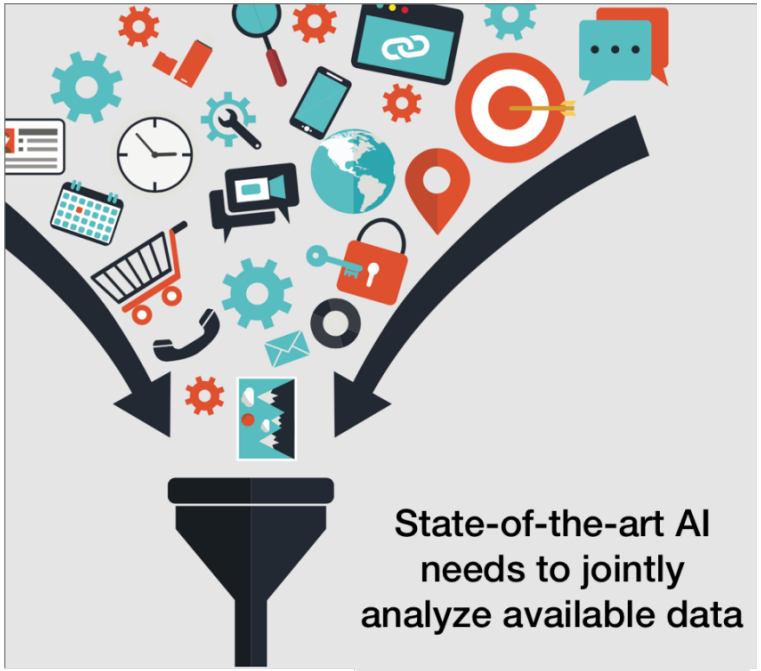
Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Training data creation
  - Data cleaning
- Part IV. Conclusions and research directions

ML is data-hungry



Successful ML requires Data Integration

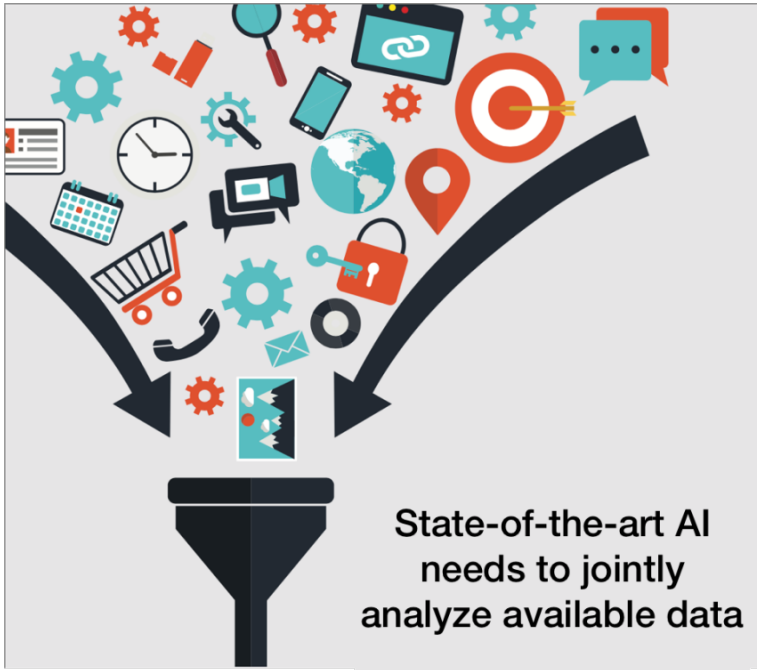


IMAGENET MovieLens

COCO is a large-scale object detection, segmentation, and captioning dataset.

Large collections of manually curated training data are necessary for progress in ML.

Successful ML requires Data Integration



IMAGENET MovieLens

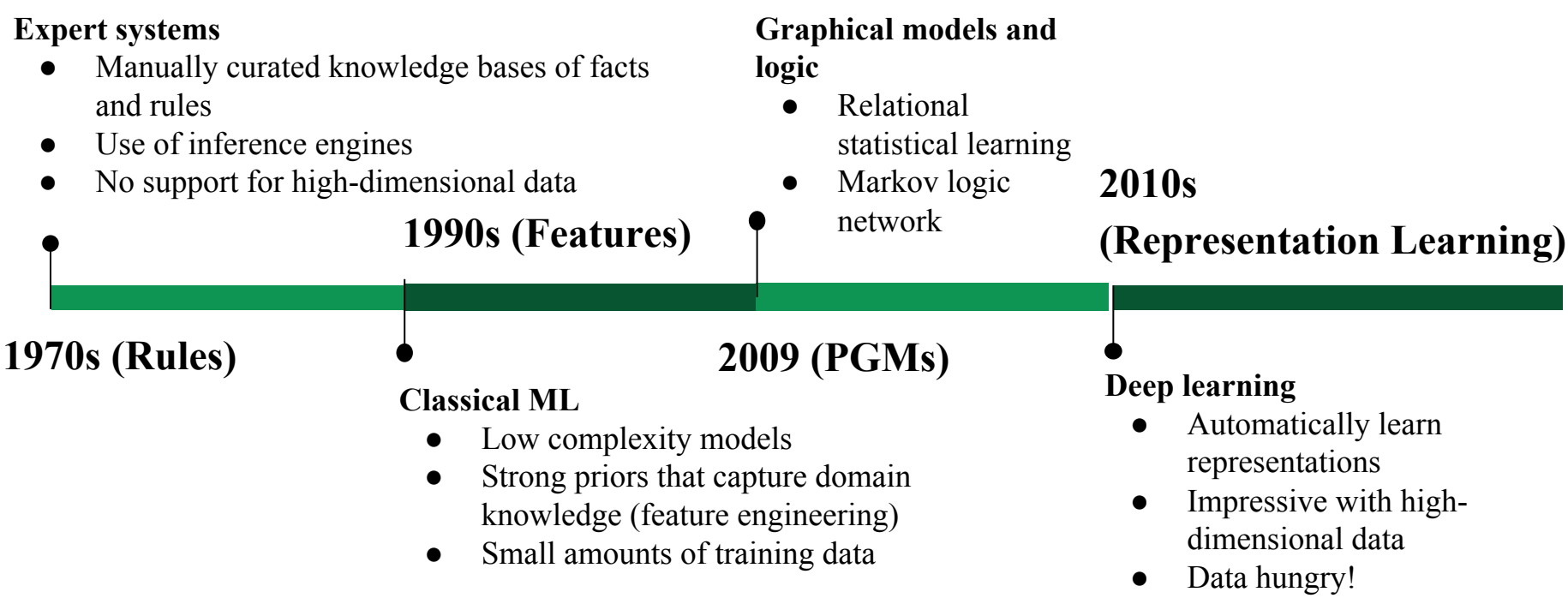
COCO is a large-scale object detection, segmentation, and captioning dataset.

Large collections of manually curated **training data** are necessary for progress in ML.

Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Training data creation
  - Data cleaning
- Part IV. Conclusions and research directions

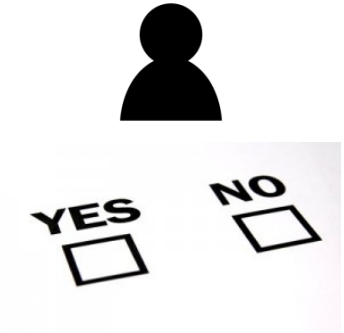
50 Years of Artificial Intelligence



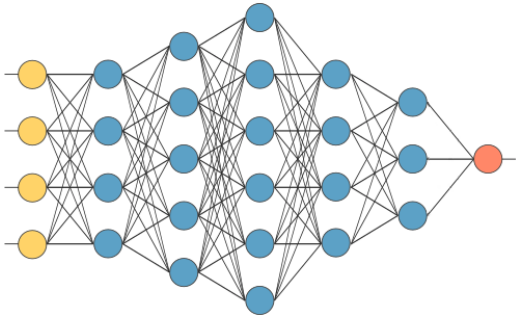
The ML Pipeline in the Deep Learning Era

Data Collection

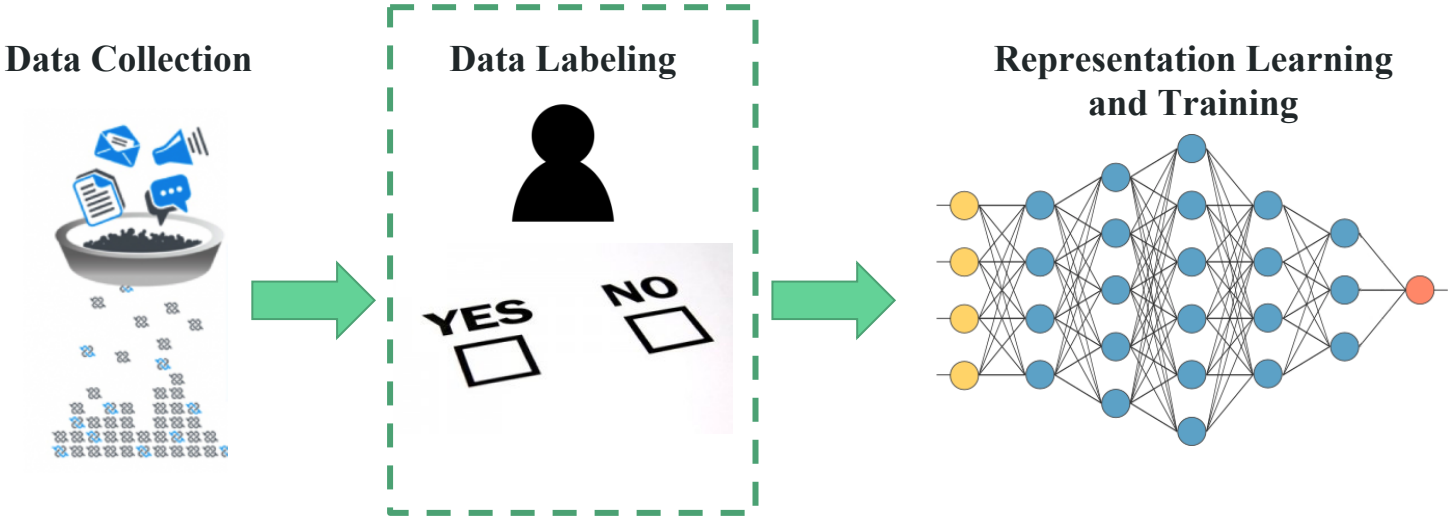
Data Labeling



Representation Learning and Training



The ML Pipeline in the Deep Learning Era



Main pain point today, most time spent in labeling data.

Training Data: Challenges and Opportunities

- Collecting training data is **expensive** and **slow.**
- We are overfitting to our training data. [Recht et al., 2018]
  - Hand-labeled training data does not change
- Training data is the point to inject domain knowledge
  - Modern ML is too complex to hand-tune features and priors

Training Data: Challenges and Opportunities

- Collecting training data is **expensive** and **slow.**
- We are overfitting to our training data. [Recht et al., 2018]
  - Hand-labeled training data does not change
- Training data is the point to inject domain knowledge
  - Modern ML is too complex to hand-tune features and priors

How do we get training data more effectively?

The Rise of Weak Supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Semi-supervised learning and ensemble learning

Examples:

- use of non-expert labelers (crowdsourcing),
- use of curated catalogs (distant supervision)
- use of heuristic rules (labeling functions)

NELL

DeepDive

HoloClean

snorkel

The Rise of Weak Supervision

- Alexa – Customer embrace of Alexa continues, with Alexa-enabled devices among the best-selling items across all of Amazon. We're seeing extremely strong adoption by other companies and developers that want to create their own experiences with Alexa. There are now more than 30,000 skills for Alexa from outside developers, and customers can control more than 4,000 smart home devices from 1,200 unique brands with Alexa. The foundations of Alexa continue to get smarter every day too. We've developed and implemented an on-device fingerprinting technique, which keeps your device from waking up when it hears an Alexa commercial on TV. (This technology ensured that our Alexa Super Bowl commercial didn't wake up millions of devices.) Far-field speech recognition (already very good) has improved by 15% over the last year; and in the U.S., U.K., and Germany, we've improved Alexa's spoken language understanding by more than 2[illegible]% over the last 12 months through enhancements in Alexa's machine learning components and the use of semi-supervised learning techniques. (These semi-supervised learning techniques reduced the amount of labeled data needed to achieve the same accuracy improvement by 40 times!) Finally, we've dramatically reduced the amount of time required to teach Alexa new languag[illegible] by using machine translation and transfer learning techniques, which allows us to serve customers in more countries (like India and Japan).

The Rise of Weak Supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Related to semi-supervised learning and ensemble learning

Examples: use of non-expert labelers (crowdsourcing), use of curated catalogs (distant supervision), use of heuristic rules (labeling functions)

Methods developed to tackle data integration problems are closely related to weak supervision.

Learning from Crowds [Raykar et al., JMLR'10]

Setup: Supervised learning but instead of gold groundtruth one has access to multiple annotators providing (possibly noisy) labels (no absolute gold standard).

Task: Learn a classifier from multiple noisy labels.

Closely related to Dawid-Skene!

Difference: Estimating the ground truth and the annotator performance is a byproduct here. Goal is to learn a classifier.

Learning from Crowds [Raykar et al., JMLR'10]

Example Task: Binary classification

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Learning from Crowds [Raykar et al., JMLR'10]

Example Task: Binary classification

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning from Crowds [Raykar et al., JMLR'10]

Example Task: Binary classification

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning: $\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} [a_i p_i + b_i(1 - p_i)]$

$$p_i := \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i).$$

$$a_i := \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1 - y_i^j}.$$

$$b_i := \prod_{j=1}^{R} [\beta^j]^{1 - y_i^j} [1 - \beta^j]^{y_i^j}.$$

Model parameters **{w, α, β}**

EM algorithm to obtain maximum-likelihood estimates. Difference with Dawid-Skene is the estimation of *w*.

Distant Supervision [Mintz et al., ACL'09]

Goal: Extracting structured knowledge from text.

Hypothesis: If two entities belong to a certain relation, any sentence containing those two entities is likely to express that relation.

Idea: Use a *database* of relations to gets lots of *noisy* training examples

- Instead of hand-creating seed tuples (bootstrapping)
- Instead of using hand-labeled corpus (supervised)

Benefits: has the advantages of supervised learning (leverage reliable hand-created knowledge), has the advantages of unsupervised learning (leverage unlimited amounts of text data).

Remember: Distant Supervision [Mintz et al., ACL'09]

Example task: Relation extraction.

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page ...

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]

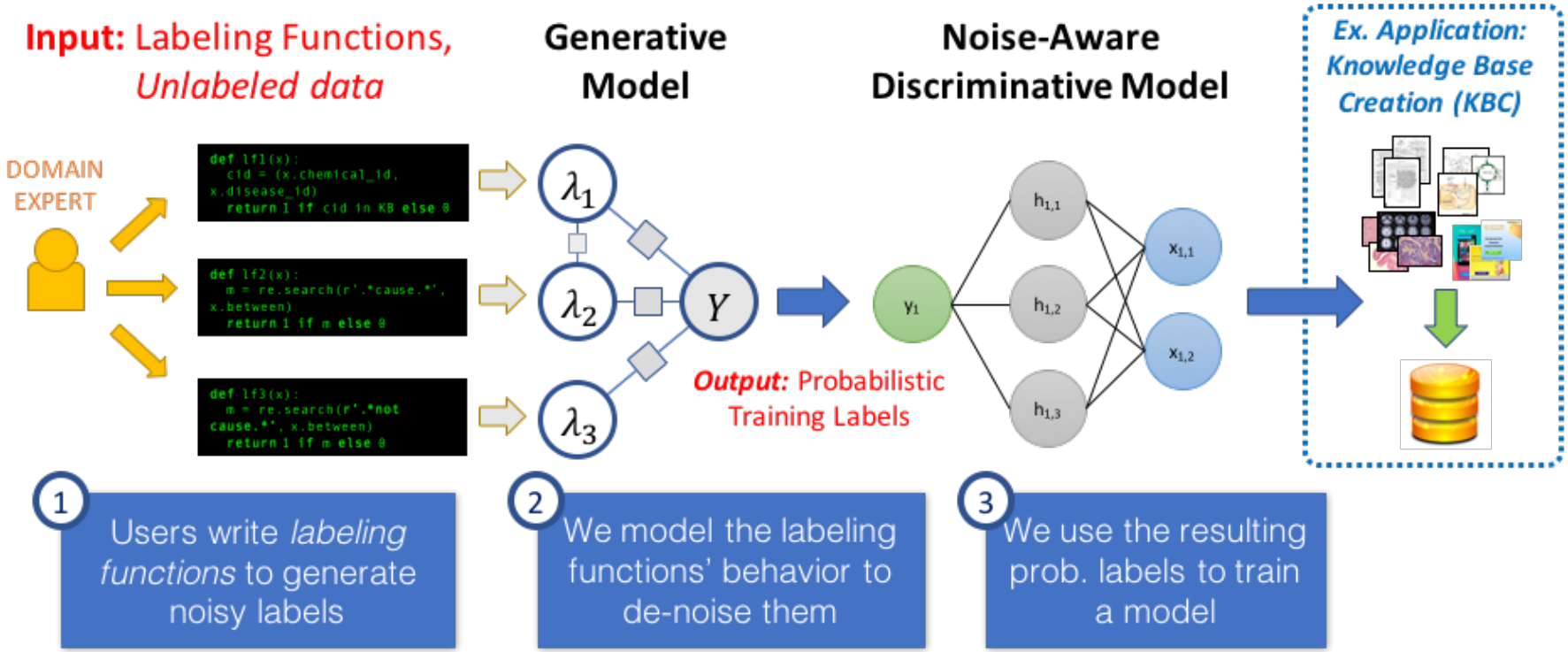
Distant Supervision [Mintz et al., ACL'09]

Entity Linking is an inherent problem in Distant Supervision.

The quality of matches can vary significantly and has a direct effect on extraction quality.

Relation	Freebase Matches	
	#sents	% true
/business/person/company	302	89.0
/people/person/place_lived	450	60.0
/location/location/contains	2793	51.0
/business/company/founders	95	48.4
/people/person/nationality	723	41.0
/location/neighborhood/neighborhood_of	68	39.7
/people/person/children	30	80.0
/people/deceased_person/place_of_death	68	22.1
/people/person/place_of_birth	162	12.0
/location/country/administrative_divisions	424	0.2

Snorkel: Code as Supervision [Ratner et al., NIPS'16, VLDB'18]



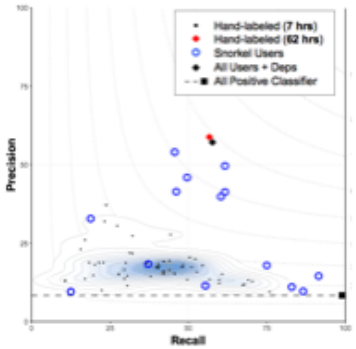
[Slide by Alex Ratner]

Snorkel: Code as Supervision [Ratner et al., NIPS'16, VLDB'18]

Snorkel biomedical workshop in collaboration with the NIH Mobilize Center

15 companies and research groups attended

How well did these new Snorkel users do?



71% New Snorkel users matched or beat 7 hours of hand-labeling

2.8x Faster than hand-labeling data

45.5% Average improvement in model performance

3rd Place Score

No machine learning experience

Beginner-level Python

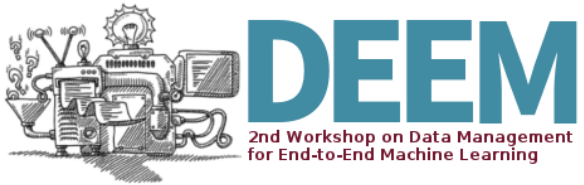
[Slide by Alex Ratner]

Alex (the creator of Snorkel) is on the market!

Alex Ratner

https://ajratner.github.io

Find out more about Snorkel MeTaL and weak supervision for Multi-task Learning at



Friday in Montgomery

Challenges in Creating Training Data

- Richly-formatted data is still a challenge. How can attack weak supervision when data includes images, text, tables, video, etc.?

- Combining weak supervision with other data enrichment techniques such as data augmentation is an exciting direction. How can reinforcement learning help here (http://goo.gl/K2qopQ)?
- How can we combine weak supervision with techniques from semi-supervised?
- Most work on weak supervision focuses on text or images. What about relational data? How can weak supervision be applied there?

Recipe for Creating Training Data

- Problem definition: **Go beyond gold labels to noisy training data.**
- Short answers
  - **Transition from “gold” labels to “high-confidence” labels.**
  - **Modeling error rates is key. The notion of *data source* is different.**
  - **Need for debugging tools, bias detection, and recommendations of weak supervision signals.**