DSC190: Machine Learning with Few Labels

Weak/distant supervision
A "standardized" view of ML

Zhiting Hu

Lecture 8, October 19, 2021

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Distant supervision
- A “standardized” view of ML

The difficulty with supervised learning

- Annotated data is expensive and costs increase when...
  - *A task requires specialized expertise*

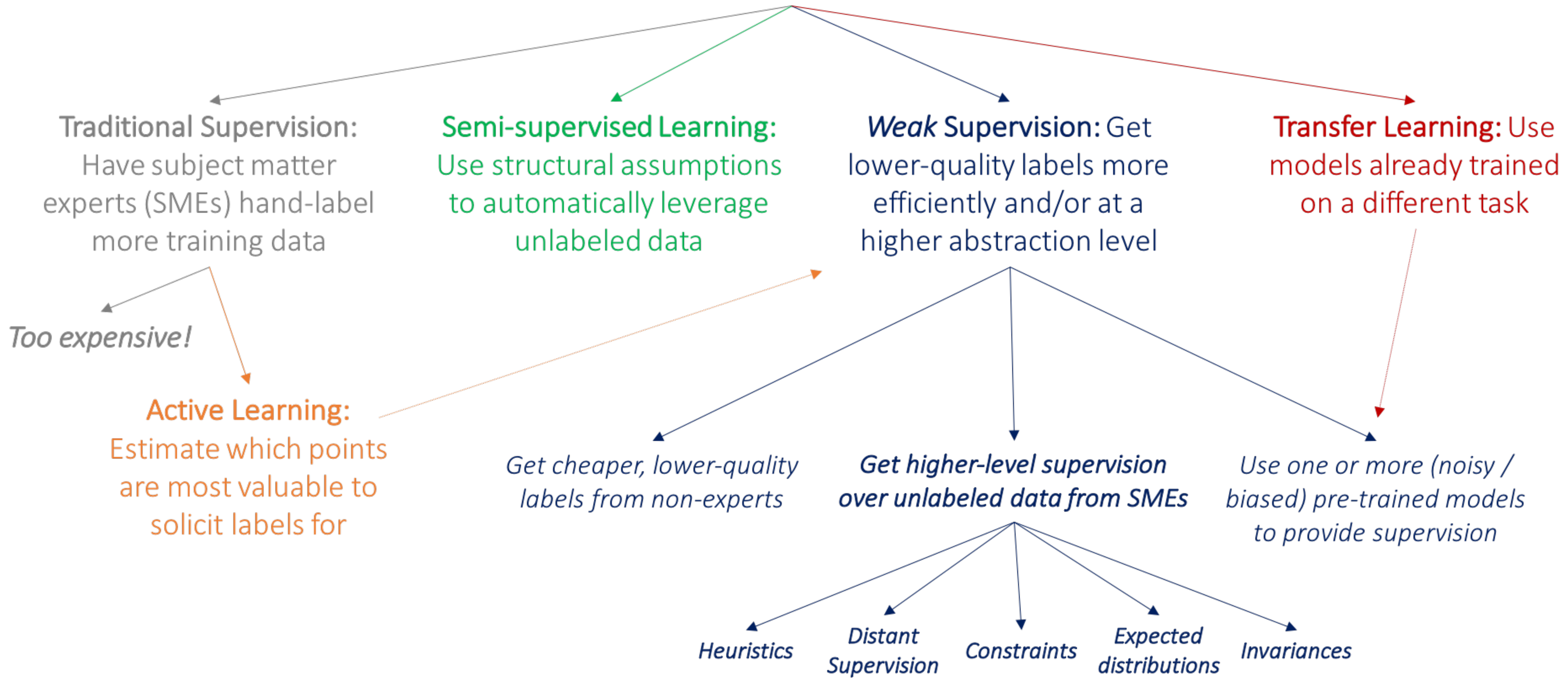
    E.g. "Only a trained linguist or a board certified radiologist can label my data"
  - *Labeling examples involves making multiple decisions*

    E.g. "Annotate this sentence with a parse tree"

    (instead of a single binary decision)

Credit: https://svivek.com/teaching/lectures/slides/weak-supervision/weak-supervision.pdf

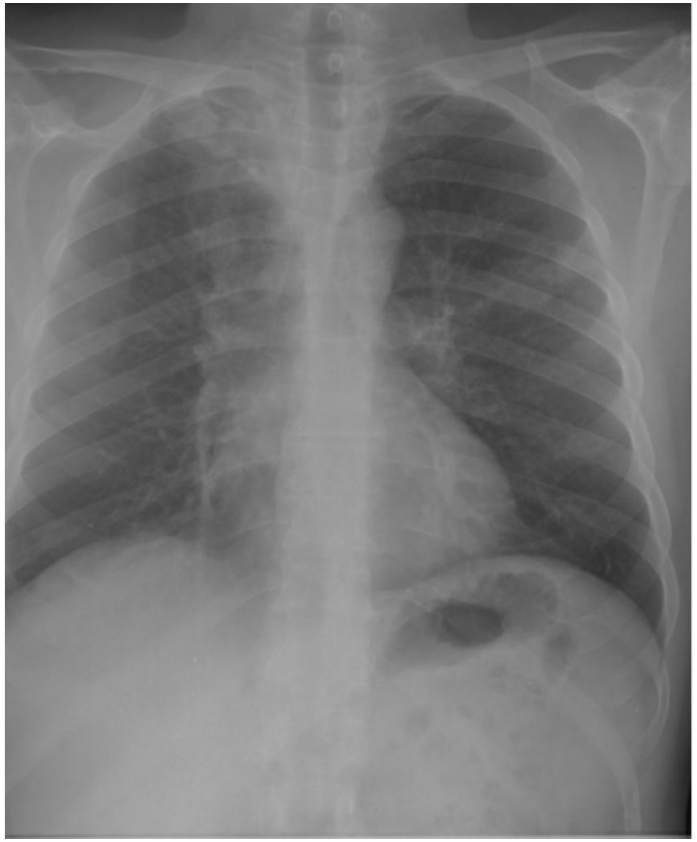
How to get more labeled training data?



Credit: https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/

Example (I): labeling with heuristics

Task: Build a chest x-ray classifier
(normal/abnormal)

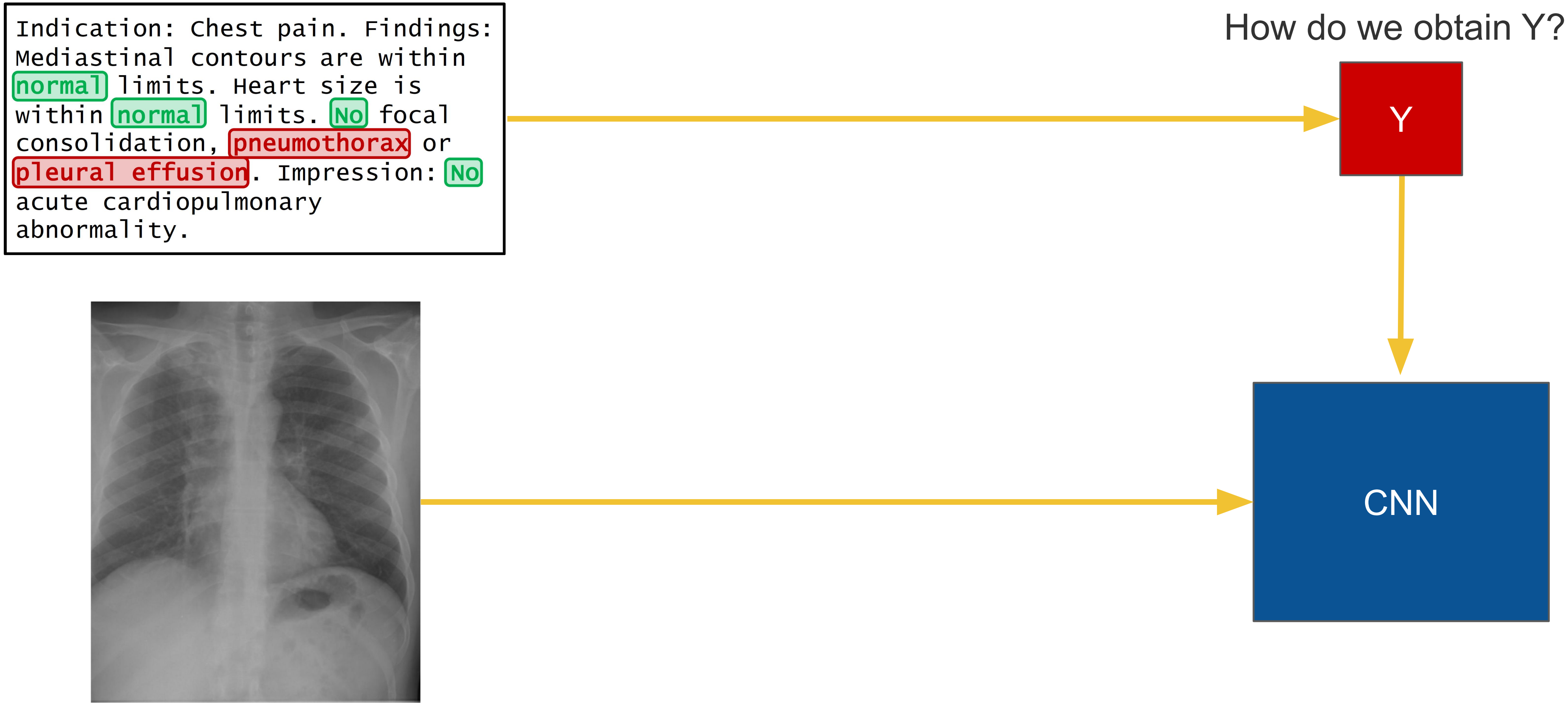


Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

Can you use the accompanying medical report (text modality) to label the x-ray (image modality)?

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (I): labeling with heuristics



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (I): labeling with heuristics

```
Indication: Chest pain. Findings:
Mediastinal contours are within
normal limits. Heart size is
within normal limits. No focal
consolidation, pneumothorax or
pleural effusion. Impression: No
acute cardiopulmonary
abnormality.
```

Normal Report

```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
    return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
    return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
     report.words)) > thresh:
      return "NORMAL"
```

LFs
(labeling functions)

Source: Khandwala et. al 2017, Cross Modal Data Programming for Medical Images

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Example (II): Labeling with knowledge bases

Task: relation extraction from text

- Hypothesis: If two entities belong to a certain relation, any sentence containing those two entities is likely to express that relation
- Key idea: use a *knowledge base* of relations to get lots of *noisy* training examples

Adapted from https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Frequent Freebase relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Adapted from https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

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

Training data

Freebase

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

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

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

Training data

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

Freebase

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

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Corpus text

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

Training data

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

Freebase

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

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

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

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

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

(Larry Page, Google)
Label: Founder
Feature: Y was founded by X

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Example (II): Labeling with knowledge bases

Negative training data

Can't train a classifier with only positive data!
Need negative training data too!

Solution?
Sample 1% of unrelated pairs of entities.

Corpus text

Larry Page took a swipe at Microsoft...
...after Harvard invited Larry Page to...
Google is Bill Gates' worst fear ...

Training data

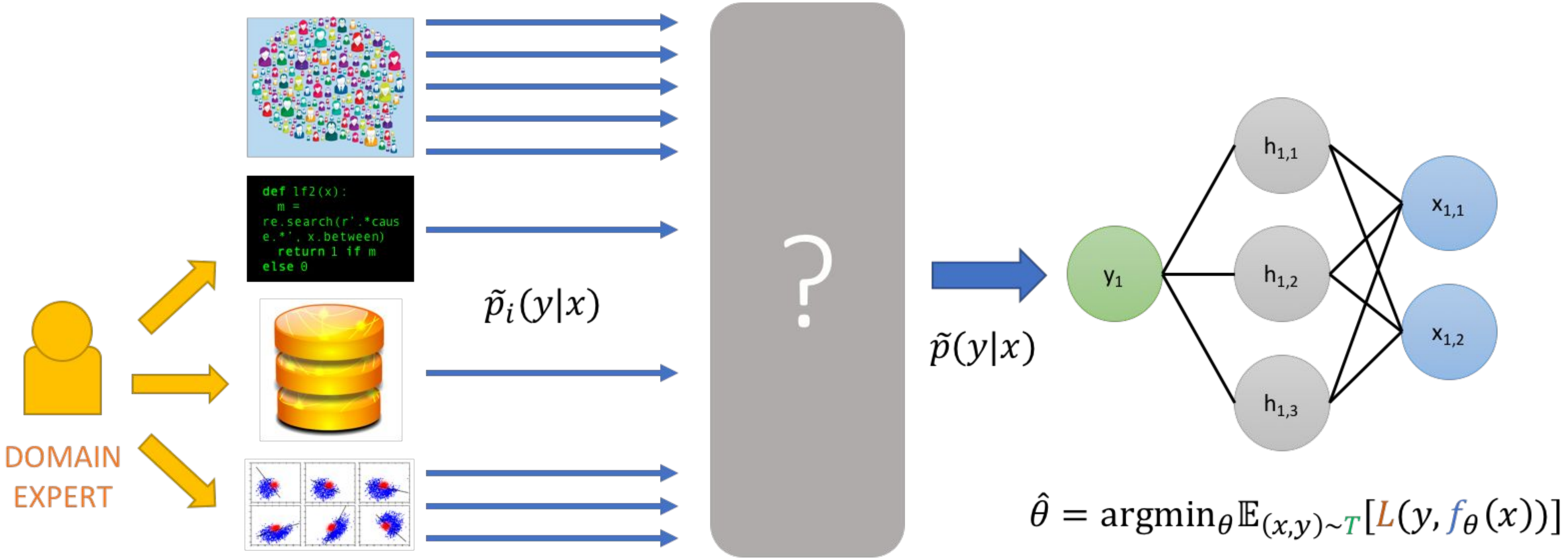
(Larry Page, Microsoft)
Label: NO_RELATION
Feature: X took a swipe at Y

(Larry Page, Harvard)
Label: NO_RELATION
Feature: Y invited X

(Bill Gates, Google)
Label: NO_RELATION
Feature: Y is X's worst fear

Credit: https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtractionII.pdf

Integrating multiple noisy labels



Example Weak
Supervision Sources

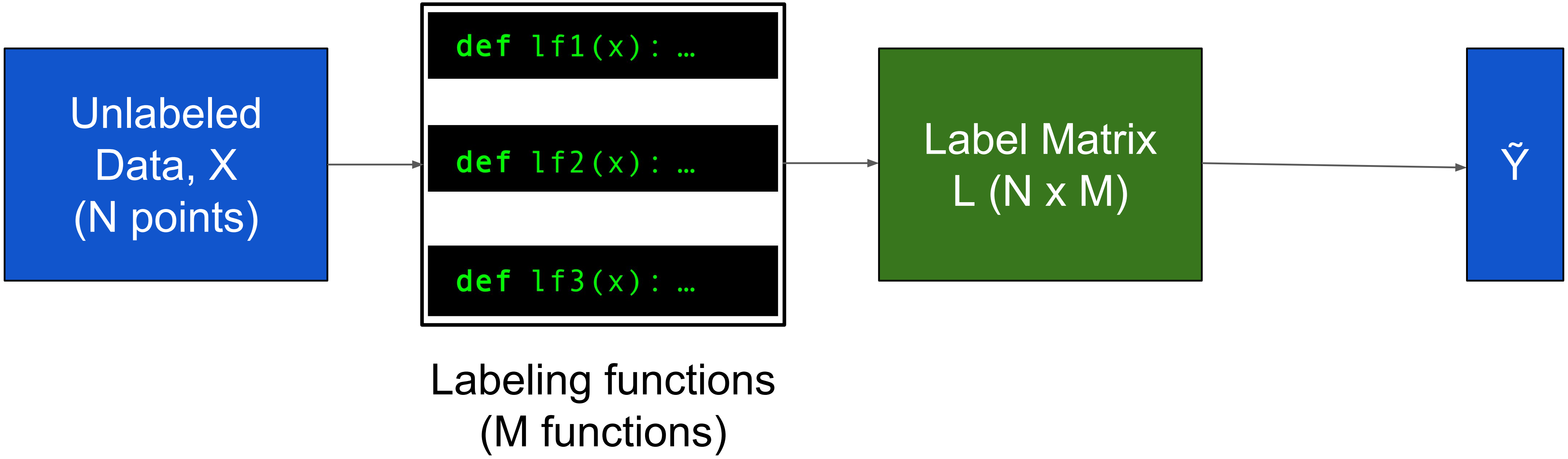
**Technical Challenge:
Integrating & Modeling
Diverse Sources**

Use Weak Supervision
to Train End Model

Source: A. Ratner et. al https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/

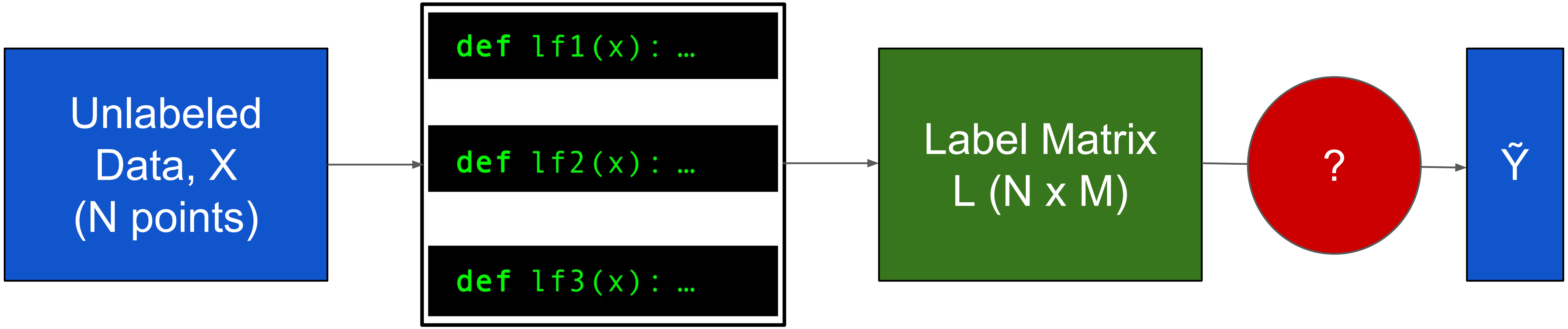
[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels



Labeling functions
(M functions)

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$**,** from the label matrix, **L**?

Approach 1 - Majority Vote

Take the majority vote of the labelling functions (LFs).

Let's say **L** = [[0, 1, 0, 1, 0]; [1, 1, 1, 1, 0]].

$$\tilde{\mathbf{Y}} = [0, 1]$$

[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$, from the label matrix, **L**?

Approach 1 - Majority Vote

Majority vote fails:

```
Indication: Chest pain. Findings:
Mediastinal contours are within
normal limits. Heart size is
within normal limits. No focal
consolidation, pneumothorax or
pleural effusion. Impression: No
acute cardiopulmonary
abnormality.
```

Normal Report

```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
      return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
      return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
    report.words)) > thresh:
      return "NORMAL"
```

LFs

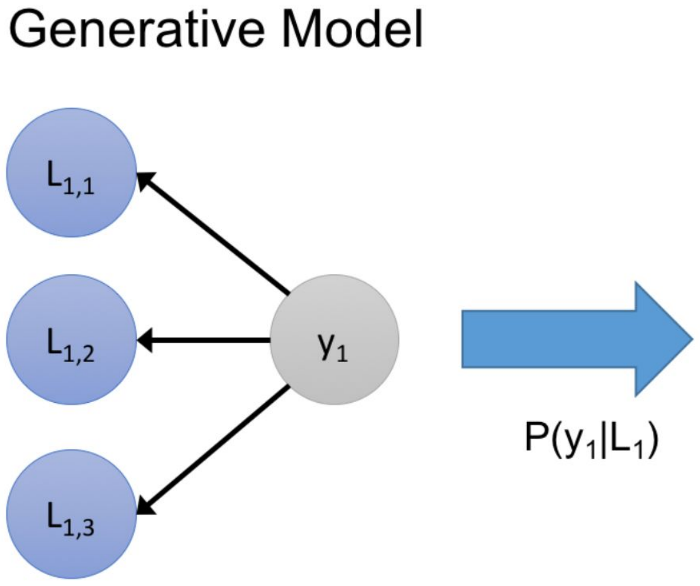
[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Integrating multiple noisy labels

How do we obtain probabilistic labels, $\tilde{\mathbf{Y}}$, from the label matrix, **L**?

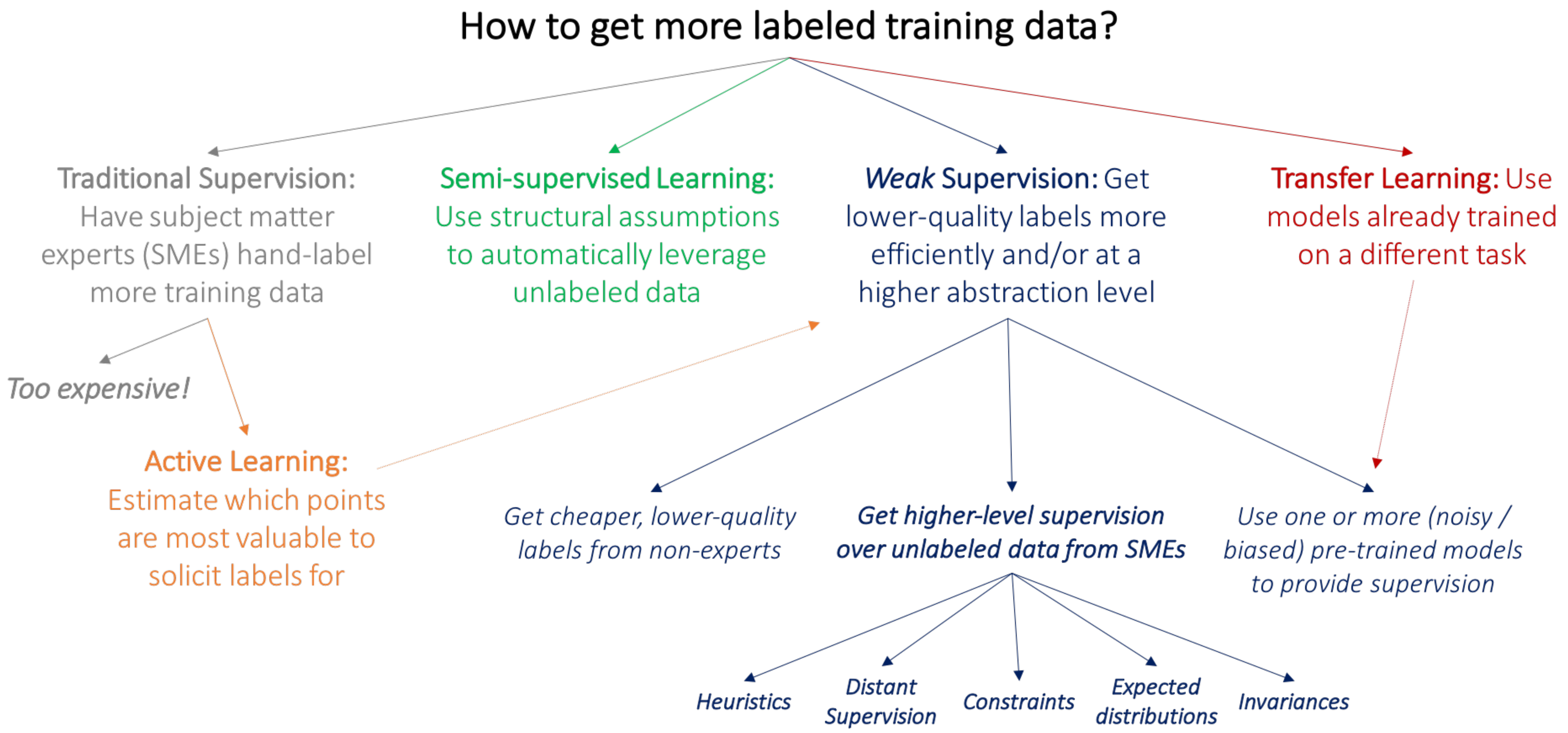
Approach 2

Train a generative model over **P(L, Y)** where **Y** are the **(unknown)** true labels



[Credit: http://cs231n.stanford.edu/slides/2018/cs231n_2018_ds07.pdf]

Summary: Weak/distant supervision



Summary: Weak/distant supervision

- Noisy labels from heuristics, knowledge bases, constraints, …
- Integrating multiple noisy labels
  - Majority vote
  - Generative modeling
  - …

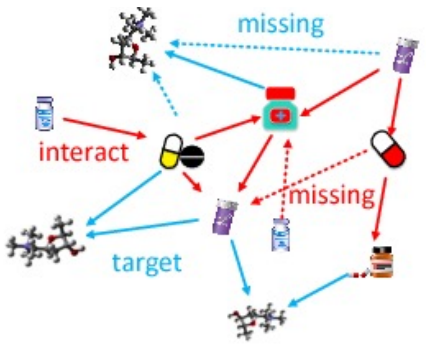
- Not all information/experiences can easily be converted into labels
  - "Every part of speech sequence should have a verb"
  - "In a sentence with word 'but', the sentiment of text after 'but' dominates"
  - "Every image patch that is recognized as a bicycle should have at least one patch that is recognized as a wheel"
  - I have a "discriminator" model that can tell me whether a model-generated image is good or not
- Need a more flexible framework to incorporate all forms of experiences

Experiences of all kinds

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

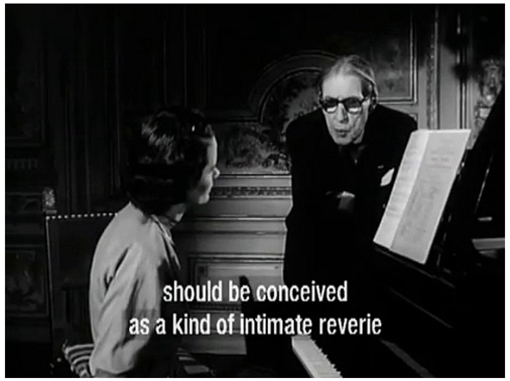


Knowledge graphs

Rewards

Auxiliary agents

Adversaries

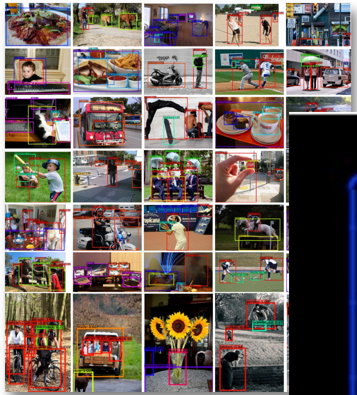


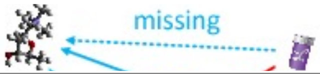
Master classes

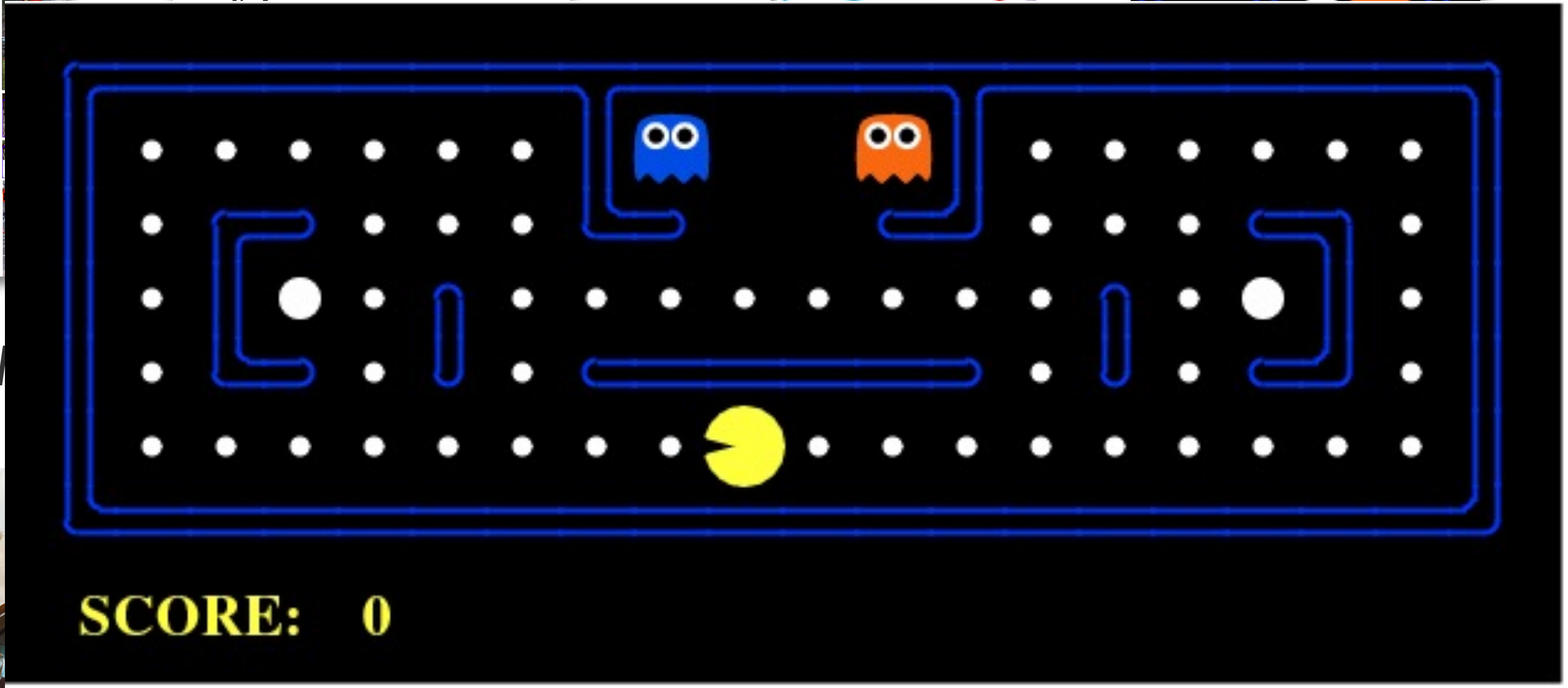
…

And all combinations thereof

Experiences of all kinds



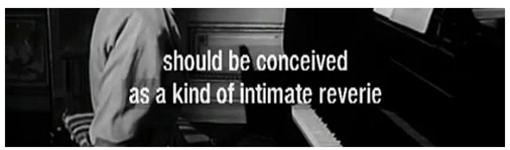




Data exampl

Auxiliary agents

ations thereof



Adversaries

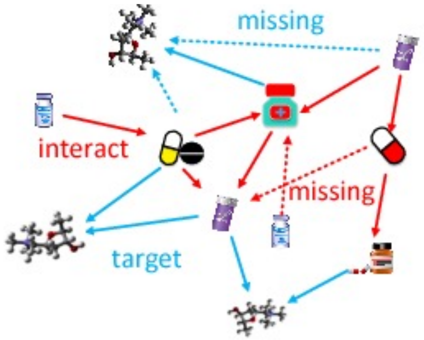
Master classes

Experiences of all kinds

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

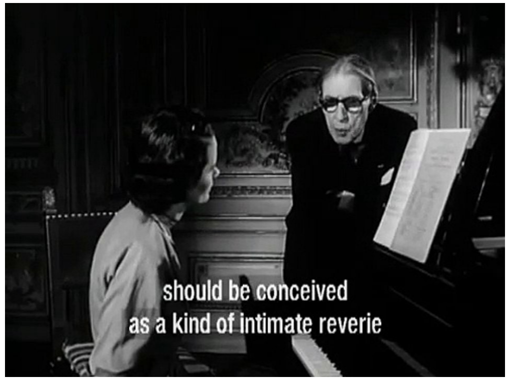


Knowledge graphs

Rewards

Auxiliary agents

Adversaries



Master classes

…

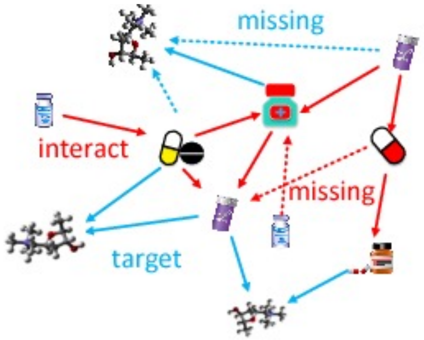
And all combinations thereof

Can we incorporate all types of experiences in learning?

Data examples

Type-2 diabetes is 90% more common than type-1

Rules/Constraints

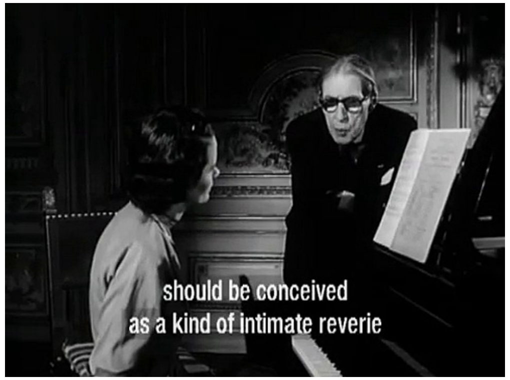


Knowledge graphs

Rewards

Auxiliary agents

Adversaries



Master classes

…

And all combinations thereof

Algorithm marketplace

Designs Driven by: experience, task, loss function, training procedure …

maximum likelihood estimation
reinforcement learning as inference
data re-weighting
inverse RL
active learning
policy optimization
data augmentation
reward-augmented maximum likelihood
label smoothing
imitation learning
softmax policy gradient
actor-critic
adversarial domain adaptation
GANs
posterior regularization
knowledge distillation
intrinsic reward
constraint-driven learning
prediction minimization
generalized expectation
regularized Bayes
learning from measurements
energy-based GANs
weak/distant supervision

Algorithm marketplace

Designs Driven by: experience, task, loss function, training procedure …

ence

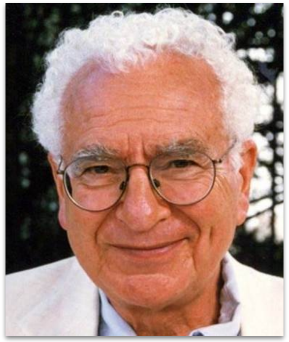
g

ood

ion

ing

Need a unifying perspective



You don't need something more in order to get something more.

-- Murray Gell-Mann (1929–2019), Physicist, Nobel laureate

"Standard equations" in Physics

Diverse electro-magnetic theories

Maxwell's Eqns: original form

$e+\frac{df}{dx}+\frac{dg}{dy}+\frac{dh}{dz}=0$	(1)	Gauss' Law
$\mu\alpha=\frac{dH}{dy}-\frac{dG}{dz}$ $\mu\beta=\frac{dF}{dz}-\frac{dH}{dx}$ $\mu\gamma=\frac{dG}{dx}-\frac{dF}{dy}$	(2)	Equivalent to Gauss' Law for magnetism
$P=\mu\left(\gamma\frac{dy}{dt}-\beta\frac{dz}{dt}\right)-\frac{dF}{dt}-\frac{d\Psi}{dx}$ $Q=\mu\left(\alpha\frac{dz}{dt}-\gamma\frac{dx}{dt}\right)-\frac{dG}{dt}-\frac{d\Psi}{dy}$ $R=\mu\left(\beta\frac{dx}{dt}-\alpha\frac{dy}{dt}\right)-\frac{dH}{dt}-\frac{d\Psi}{dz}$	(3)	Faraday's Law (with the Lorentz Force and Poisson's Law)
$\frac{d\gamma}{dy}-\frac{d\beta}{dz}=4\pi p'$ $\quad p'=p+\frac{df}{dt}$ $\frac{d\alpha}{dz}-\frac{d\gamma}{dx}=4\pi q'$ $\quad q'=q+\frac{dg}{dt}$ $\frac{d\beta}{dx}-\frac{d\alpha}{dy}=4\pi r'$ $\quad r'=r+\frac{dh}{dt}$	(4)	Ampère-Maxwell Law
$P=-\xi p \quad Q=-\xi q \quad R=-\xi r$		Ohm's Law
$P=kf \quad Q=kg \quad R=kh$		The electric elasticity equation (**E** = **D**/ε)
$\frac{de}{dt}+\frac{dp}{dx}+\frac{dq}{dy}+\frac{dr}{dz}=0$		Continuity of charge

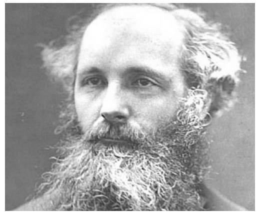
Maxwell's Eqns simplified w/ rotational symmetry

$$\nabla\cdot\mathbf{D}=\rho_V$$

$$\nabla\cdot\mathbf{B}=0$$

$$\nabla\times\mathbf{E}=-\frac{\partial\mathbf{B}}{\partial t}$$

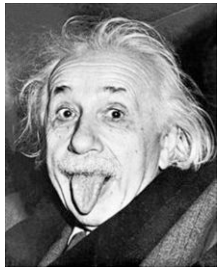
$$\nabla\times\mathbf{H}=\frac{\partial\mathbf{D}}{\partial t}+\mathbf{J}$$



Maxwell's Eqns further simplified w/ symmetry of special relativity

$$\varepsilon^{uvk\lambda}\partial_v F_{k\lambda}=0$$

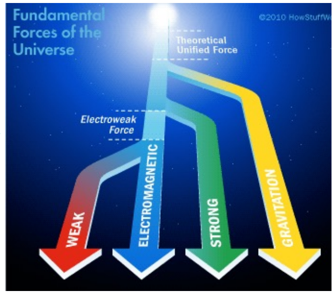
$$\partial_v F^{uV}=\frac{4\pi}{c}j^u$$



Standard Model w/ Yang-Mills theory and US(3) symmetry

$$\mathcal{L}_{\mathrm{gf}}=-\frac{1}{2}\mathrm{Tr}(F^2)=-\frac{1}{4}F^{a\mu\nu}F^{a}_{\mu\nu}$$

Unification of fundamental forces?



1861

1910s

1970s

?

A “Standardized” View of ML

Recap: MLE

- The most classical learning algorithm

- Supervised:
  - Observe data $\mathcal{D} = \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$
  - Solve with SGD

$$\min_\theta - \mathbb{E}_{(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathcal{D}} \Big[\log p_\theta(\boldsymbol{y}^* | \boldsymbol{x}^*) \Big]$$

- Unsupervised:
  - Observe $\mathcal{D} = \{(\boldsymbol{x}^*)\}$, $\boldsymbol{y}$ is latent variable
  - Posterior $p_\theta(\boldsymbol{y}|\boldsymbol{x})$
  - Solve with EM, etc

$$\min_\theta - \mathbb{E}_{\boldsymbol{x}^* \sim \mathcal{D}} \Big[\log \int_{\boldsymbol{y}} p_\theta(\boldsymbol{x}^*, \boldsymbol{y}) \Big]$$

Recap: MLE as entropy maximization

- Duality between Supervised MLE and maximum entropy, when p is exponential family

Shannon entropy H

$$\min_{p(\boldsymbol{x},\boldsymbol{y})} \; H(p)$$

features $T(\boldsymbol{x},\boldsymbol{y})$

$$s.t.\;\; \mathbb{E}_p[T(\boldsymbol{x},\boldsymbol{y})] = \mathbb{E}_{(x^*,y^*)\sim\mathcal{D}}[T(\boldsymbol{x},\boldsymbol{y})]$$

data as constraints

** Proof with Lagrangian method*

Recap: MLE as entropy maximization

- Unsupervised MLE can be achieved by maximizing the negative free energy:
  - Introduce auxiliary distribution $q(\boldsymbol{y}|\boldsymbol{x})$ (and then play with its entropy and cross entropy, etc.)

$$\log \int_{\boldsymbol{y}} p_\theta(\boldsymbol{x}^*, \boldsymbol{y}) = \mathbb{E}_{q(\boldsymbol{y}|\boldsymbol{x}^*)}\left[\log \frac{p_\theta(\boldsymbol{x}^*, \boldsymbol{y})}{q(\boldsymbol{y}|\boldsymbol{x}^*)}\right] + \mathrm{KL}\big(q(\boldsymbol{y}|\boldsymbol{x}^*) \,||\, p_\theta(\boldsymbol{y}|\boldsymbol{x}^*)\big)$$

$$\geq H\big(q(\boldsymbol{y}|\boldsymbol{x}^*)\big) + \mathbb{E}_{q(\boldsymbol{y}|\boldsymbol{x}^*)}[\log p_\theta(\boldsymbol{x}^*, \boldsymbol{y})]$$

Bayesian Inference

- Posterior

$$p(\boldsymbol{z}|\mathcal{D}) = \frac{\pi(\boldsymbol{z}) \prod_{\boldsymbol{x}^* \in \mathcal{D}} p(\boldsymbol{x}^*|\boldsymbol{z})}{p(\mathcal{D})}$$

- Connecting to maximum entropy, as an optimization problem [Zellner, 1988]:

$$\min_{q(\boldsymbol{z})} \; -\mathrm{H}(q(\boldsymbol{z})) + \log p(\mathcal{D}) - \mathbb{E}_{q(\boldsymbol{z})}\left[\log \pi(\boldsymbol{z}) + \sum_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\right]$$

$$s.t. \; q(\boldsymbol{z}) \in \mathcal{P}$$

(the normality constraint of a probability distribution)

Posterior Regularization

- Under the optimization viewpoint of Bayesian inference, it's natural to consider other types of constraints that encode richer problem structures and domain knowledge
- Posterior regularization [Ganchev et al., 2010], or regularized Bayes (Zhu et al., 2014)

$$\begin{aligned} \min_{q,\boldsymbol{\xi}} \quad & -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\sum_{\boldsymbol{x}^*\in\mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right] + U(\boldsymbol{\xi}) \\ s.t. \quad & q(\boldsymbol{z}) \in \mathcal{Q}(\boldsymbol{\xi}) \\ & \boldsymbol{\xi} \geq 0, \end{aligned}$$

  - $\boldsymbol{\xi}$: slack variables
  - $U(\boldsymbol{\xi})$: a penalty function (e.g., L1 norm of $\boldsymbol{\xi}$)
  - $Q(\boldsymbol{\xi})$: a subset of valid distributions over $\boldsymbol{z}$ that satisfy the constraints determined by $\boldsymbol{\xi}$

Posterior Regularization

$$\min_{q,\boldsymbol{\xi}} \quad -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\sum\nolimits_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right] + U(\boldsymbol{\xi})$$

$$s.t. \;\; q(\boldsymbol{z}) \in \mathcal{Q}(\boldsymbol{\xi})$$

$$\boldsymbol{\xi} \geq 0,$$

- Ex: let $T(\boldsymbol{x}^*, \boldsymbol{z})$ be a feature vector of data instance $\boldsymbol{x}^*$, a constrained posterior set $Q(\boldsymbol{\xi})$ with "feature expectation" constraints can be defined as

$$Q(\boldsymbol{\xi}) := \{q(\boldsymbol{z}) \; : \; \mathbb{E}_q\left[T(\boldsymbol{x}^*; \boldsymbol{z})\right] \leq \boldsymbol{\xi}\}$$

  - i.e., bounds the feature expectations with $\boldsymbol{\xi}$
- Assuming $U(\boldsymbol{\xi}) = \sum \xi_i$, rewrite without slack variables

$$\min_{q,\boldsymbol{\xi}} \quad -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\sum\nolimits_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^*|\boldsymbol{z})\pi(\boldsymbol{z})\right] - c \cdot \mathbb{E}_{q(\boldsymbol{z})}\left[T(\boldsymbol{x}^*; \boldsymbol{z})\right]$$

Posterior Regularization

$$\min_{q,\boldsymbol{\xi}} \quad -\mathrm{H}\left(q(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z})}\left[\sum\nolimits_{\boldsymbol{x}^* \in \mathcal{D}} \log p(\boldsymbol{x}^* | \boldsymbol{z}) \pi(\boldsymbol{z})\right] + U(\boldsymbol{\xi})$$

$$s.t. \;\; q(\boldsymbol{z}) \in \mathcal{Q}(\boldsymbol{\xi}) := \{q(\boldsymbol{z}) \; : \; \mathbb{E}_q\left[T(\boldsymbol{x}^*; \boldsymbol{z})\right] \leq \boldsymbol{\xi}\}$$

$$\boldsymbol{\xi} \geq 0,$$

- $U(\boldsymbol{\xi}) = \sum \xi_i$
- solution for $q(\boldsymbol{z})$: $q(\boldsymbol{z}) = \exp\{\log p(\boldsymbol{x}, \boldsymbol{z}) + T(\boldsymbol{x}^*, \boldsymbol{z})\} \,/\, Z$

$$= p(\boldsymbol{x}, \boldsymbol{z}) \exp\{T(\boldsymbol{x}^*, \boldsymbol{z})\}/Z$$

** Proof with Lagrangian method*

Key Takeaways

- Recap: Supervised MLE and maximum entropy
- Recap: Unsupervised MLE and maximum entropy
- Bayesian inference and maximum entropy
    - Bayesian inference as optimization
- Posterior regularization:
    - Constrained Bayesian inference => constrained optimization

Questions?