

Structured Predictions: Practical Advancements and Applications in Natural Language Processing

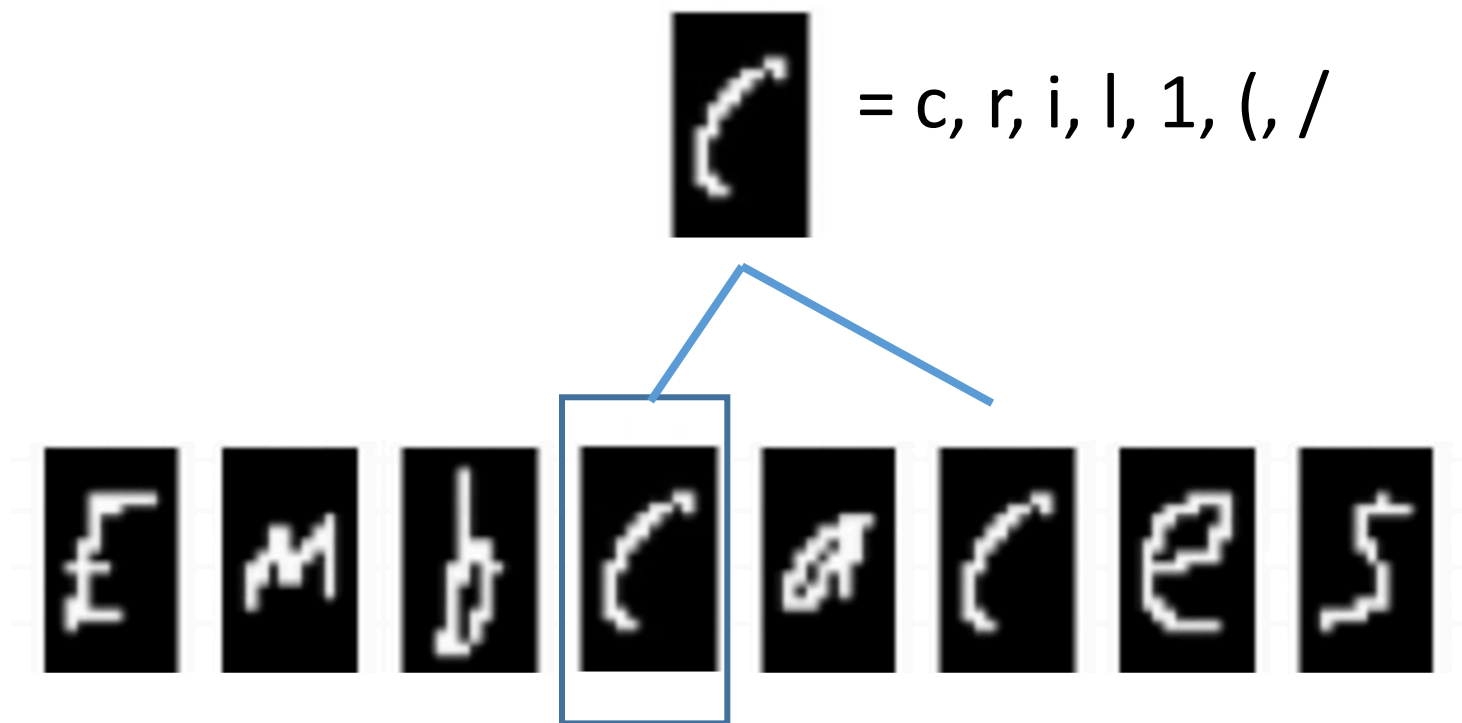
Kai-Wei Chang
UCLA



References: <http://kwchang.net/talk/sp.html>

Why is structure important? Hand written recognition example

❖ What is this English letter?

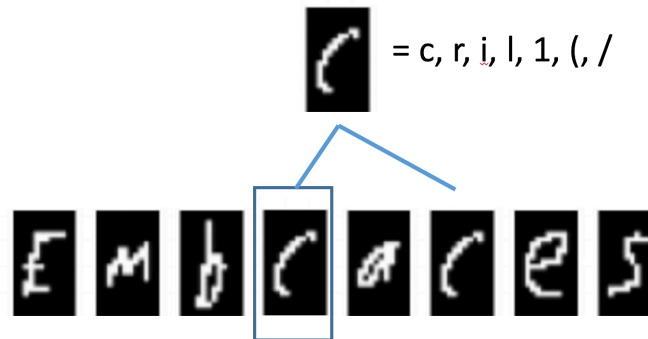


Credit: Ben Taskar

Why is structure important?

Hand written recognition example

❖ What is this English letter?



Credit: Ben Taskar

Part-of-speech (POS) tagging:

A	♯	♯	chases	a	mouse
Det	??	??	Verb	Det	Noun
		Noun			

Q: [Chris] = [Mr. Robin] ?

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

P.S. In fact, Alan Alexander Milne is the author.

Slide modified from Dan Roth

Complex Decision Structure

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

P.S. In fact, Alan Alexander Milne is the author.

Challenges--language is compositional



Carefully
Slide



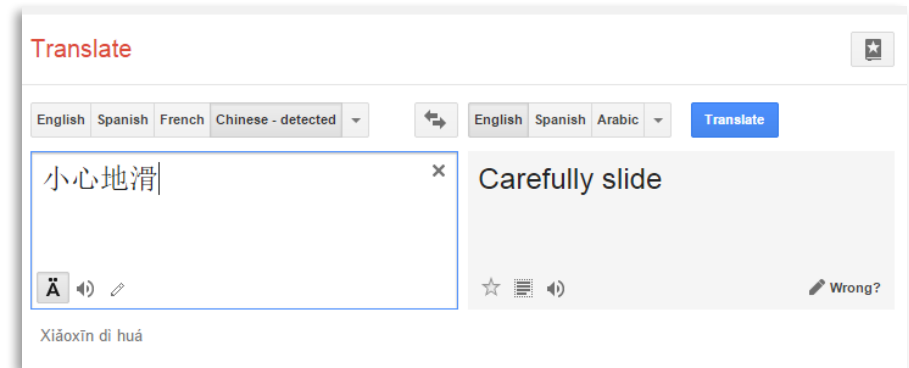
Challenges--language is compositional



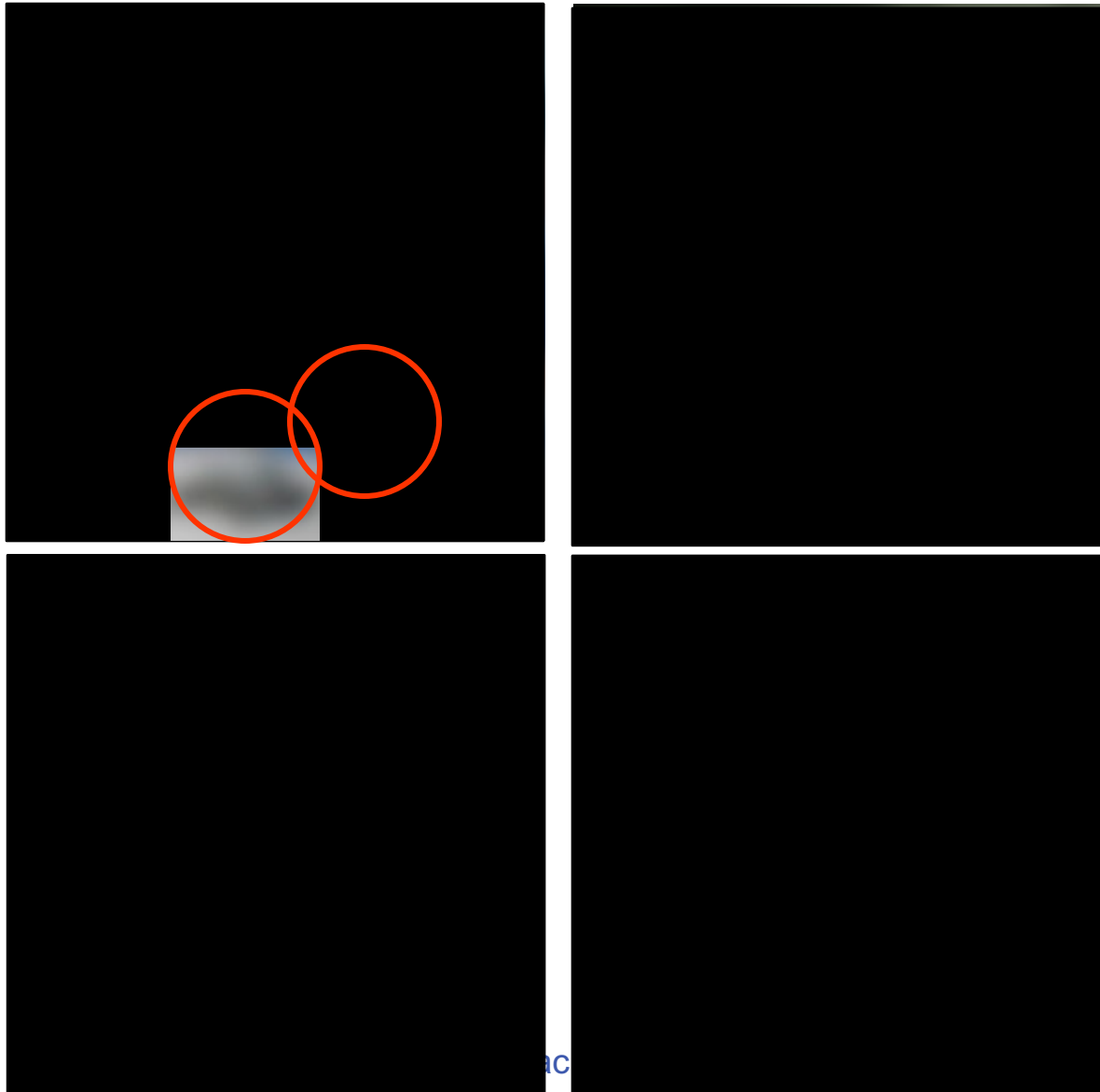
小心:
Carefully
Careful
Take
Care
Caution



地滑:
Slide
Landslip
Wet Floor
Smooth

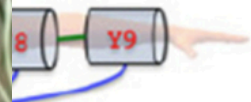
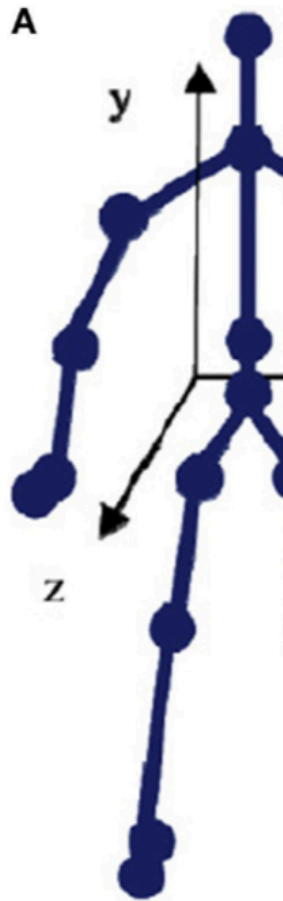


Visual recognition



Credit: Dhruv Batra

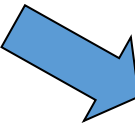
Human body recognition




Joint Inference with General Constraint Structure

[Roth&Yih'04,07,....]

Recognizing Entities and Relations



other	0.05
per	0.85
loc	0.10



other	0.10
per	0.60
loc	0.30



other	0.05
per	0.50
loc	0.45

Joint inference gives good improvement

Bernie's wife, Jane, is a native

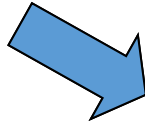
E_1

E_2


R_{12}

R_{23}

Key Questions:
How to learn the model(s)?
What is the source of the knowledge?
How to guide the global inference?



irrelevant	0.05
spouse_of	0.45
born_in	0.50

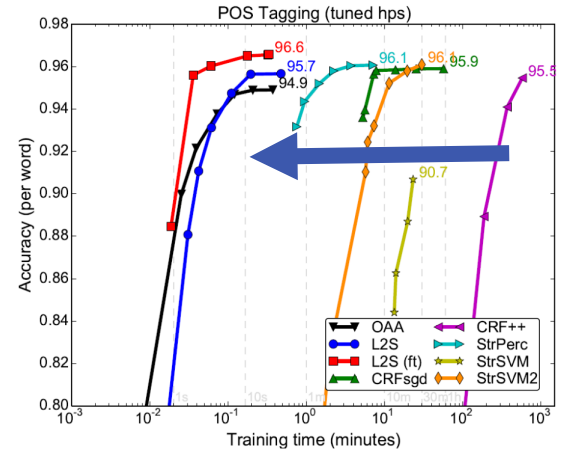
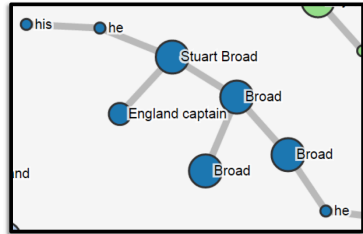


irrelevant	0.10
spouse_of	0.05
born_in	0.85

Models could be learned separately/jointly; constraints may come up only at decision time.

Structured Prediction Models

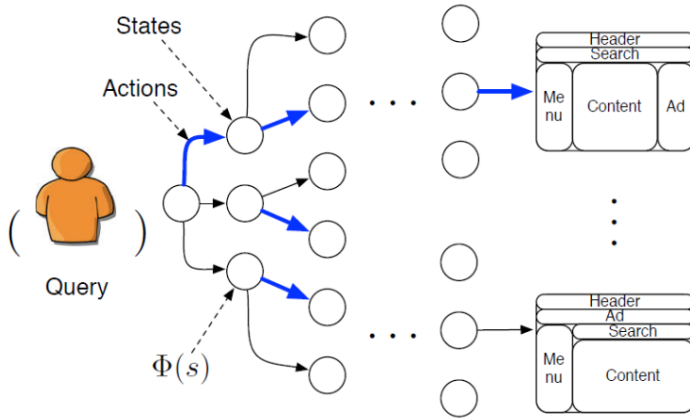
[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58] [59] [60] [61] [62] [63] [64] [65] [66] [67] [68] [69] [70] [71] [72] [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85] [86] [87] [88] [89] [90] [91] [92] [93] [94] [95] [96] [97] [98] [99] [100] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111] [112] [113] [114] [115] [116] [117] [118] [119] [120] [121] [122] [123] [124] [125] [126] [127] [128] [129] [130] [131] [132] [133] [134] [135] [136] [137] [138] [139] [140] [141] [142] [143] [144] [145] [146] [147] [148] [149] [150] [151] [152] [153] [154] [155] [156] [157] [158] [159] [160] [161] [162] [163] [164] [165] [166] [167] [168] [169] [170] [171] [172] [173] [174] [175] [176] [177] [178] [179] [180] [181] [182] [183] [184] [185] [186] [187] [188] [189] [190] [191] [192] [193] [194] [195] [196] [197] [198] [199] [200] [201] [202] [203] [204] [205] [206] [207] [208] [209] [210] [211] [212] [213] [214] [215] [216] [217] [218] [219] [220] [221] [222] [223] [224] [225] [226] [227] [228] [229] [230] [231] [232] [233] [234] [235] [236] [237] [238] [239] [240] [241] [242] [243] [244] [245] [246] [247] [248] [249] [250] [251] [252] [253] [254] [255] [256] [257] [258] [259] [260] [261] [262] [263] [264] [265] [266] [267] [268] [269] [270] [271] [272] [273] [274] [275] [276] [277] [278] [279] [280] [281] [282] [283] [284] [285] [286] [287] [288] [289] [290] [291] [292] [293] [294] [295] [296] [297] [298] [299] [300] [301] [302] [303] [304] [305] [306] [307] [308] [309] [310] [311] [312] [313] [314] [315] [316] [317] [318] [319] [320] [321] [322] [323] [324] [325] [326] [327] [328] [329] [330] [331] [332] [333] [334] [335] [336] [337] [338] [339] [340] [341] [342] [343] [344] [345] [346] [347] [348] [349] [350] [351] [352] [353] [354] [355] [356] [357] [358] [359] [360] [361] [362] [363] [364] [365] [366] [367] [368] [369] [370] [371] [372] [373] [374] [375] [376] [377] [378] [379] [380] [381] [382] [383] [384] [385] [386] [387] [388] [389] [390] [391] [392] [393] [394] [395] [396] [397] [398] [399] [400] [401] [402] [403] [404] [405] [406] [407] [408] [409] [410] [411] [412] [413] [414] [415] [416] [417] [418] [419] [420] [421] [422] [423] [424] [425] [426] [427] [428] [429] [430] [431] [432] [433] [434] [435] [436] [437] [438] [439] [440] [441] [442] [443] [444] [445] [446] [447] [448] [449] [450] [451] [452] [453] [454] [455] [456] [457] [458] [459] [460] [461] [462] [463] [464] [465] [466] [467] [468] [469] [470] [471] [472] [473] [474] [475] [476] [477] [478] [479] [480] [481] [482] [483] [484] [485] [486] [487] [488] [489] [490] [491] [492] [493] [494] [495] [496] [497] [498] [499] [500] [501] [502] [503] [504] [505] [506] [507] [508] [509] [510] [511] [512] [513] [514] [515] [516] [517] [518] [519] [520] [521] [522] [523] [524] [525] [526] [527] [528] [529] [530] [531] [532] [533] [534] [535] [536] [537] [538] [539] [540] [541] [542] [543] [544] [545] [546] [547] [548] [549] [550] [551] [552] [553] [554] [555] [556] [557] [558] [559] [560] [561] [562] [563] [564] [565] [566] [567] [568] [569] [570] [571] [572] [573] [574] [575] [576] [577] [578] [579] [580] [581] [582] [583] [584] [585] [586] [587] [588] [589] [590] [591] [592] [593] [594] [595] [596] [597] [598] [599] [600] [601] [602] [603] [604] [605] [606] [607] [608] [609] [610] [611] [612] [613] [614] [615] [616] [617] [618] [619] [620] [621] [622] [623] [624] [625] [626] [627] [628] [629] [630] [631] [632] [633] [634] [635] [636] [637] [638] [639] [640] [641] [642] [643] [644] [645] [646] [647] [648] [649] [650] [651] [652] [653] [654] [655] [656] [657] [658] [659] [660] [661] [662] [663] [664] [665] [666] [667] [668] [669] [670] [671] [672] [673] [674] [675] [676] [677] [678] [679] [680] [681] [682] [683] [684] [685] [686] [687] [688] [689] [690] [691] [692] [693] [694] [695] [696] [697] [698] [699] [700] [701] [702] [703] [704] [705] [706] [707] [708] [709] [710] [711] [712] [713] [714] [715] [716] [717] [718] [719] [720] [721] [722] [723] [724] [725] [726] [727] [728] [729] [730] [731] [732] [733] [734] [735] [736] [737] [738] [739] [740] [741] [742] [743] [744] [745] [746] [747] [748] [749] [750] [751] [752] [753] [754] [755] [756] [757] [758] [759] [760] [761] [762] [763] [764] [765] [766] [767] [768] [769] [770] [771] [772] [773] [774] [775] [776] [777] [778] [779] [780] [781] [782] [783] [784] [785] [786] [787] [788] [789] [790] [791] [792] [793] [794] [795] [796] [797] [798] [799] [800] [801] [802] [803] [804] [805] [806] [807] [808] [809] [810] [811] [812] [813] [814] [815] [816] [817] [818] [819] [820] [821] [822] [823] [824] [825] [826] [827] [828] [829] [830] [831] [832] [833] [834] [835] [836] [837] [838] [839] [840] [841] [842] [843] [844] [845] [846] [847] [848] [849] [850] [851] [852] [853] [854] [855] [856] [857] [858] [859] [860] [861] [862] [863] [864] [865] [866] [867] [868] [869] [870] [871] [872] [873] [874] [875] [876] [877] [878] [879] [880] [881] [882] [883] [884] [885] [886] [887] [888] [889] [890] [891] [892] [893] [894] [895] [896] [897] [898] [899] [900] [901] [902] [903] [904] [905] [906] [907] [908] [909] [910] [911] [912] [913] [914] [915] [916] [917] [918] [919] [920] [921] [922] [923] [924] [925] [926] [927] [928] [929] [930] [931] [932] [933] [934] [935] [936] [937] [938] [939] [940] [941] [942] [943] [944] [945] [946] [947] [948] [949] [950] [951] [952] [953] [954] [955] [956] [957] [958] [959] [960] [961] [962] [963] [964] [965] [966] [967] [968] [969] [970] [971] [972] [973] [974] [975] [976] [977] [978] [979] [980] [981] [982] [983] [984] [985] [986] [987] [988] [989] [990] [991] [992] [993] [994] [995] [996] [997] [998] [999] [1000]



How to model?

Training/test/dev speed

Query



activity	cooking
agent	woman
food	vegetable

Learning signals

Fairness (data biases)

Outline

❖ Part 1: 14:40-15:50

From binary to structured prediction

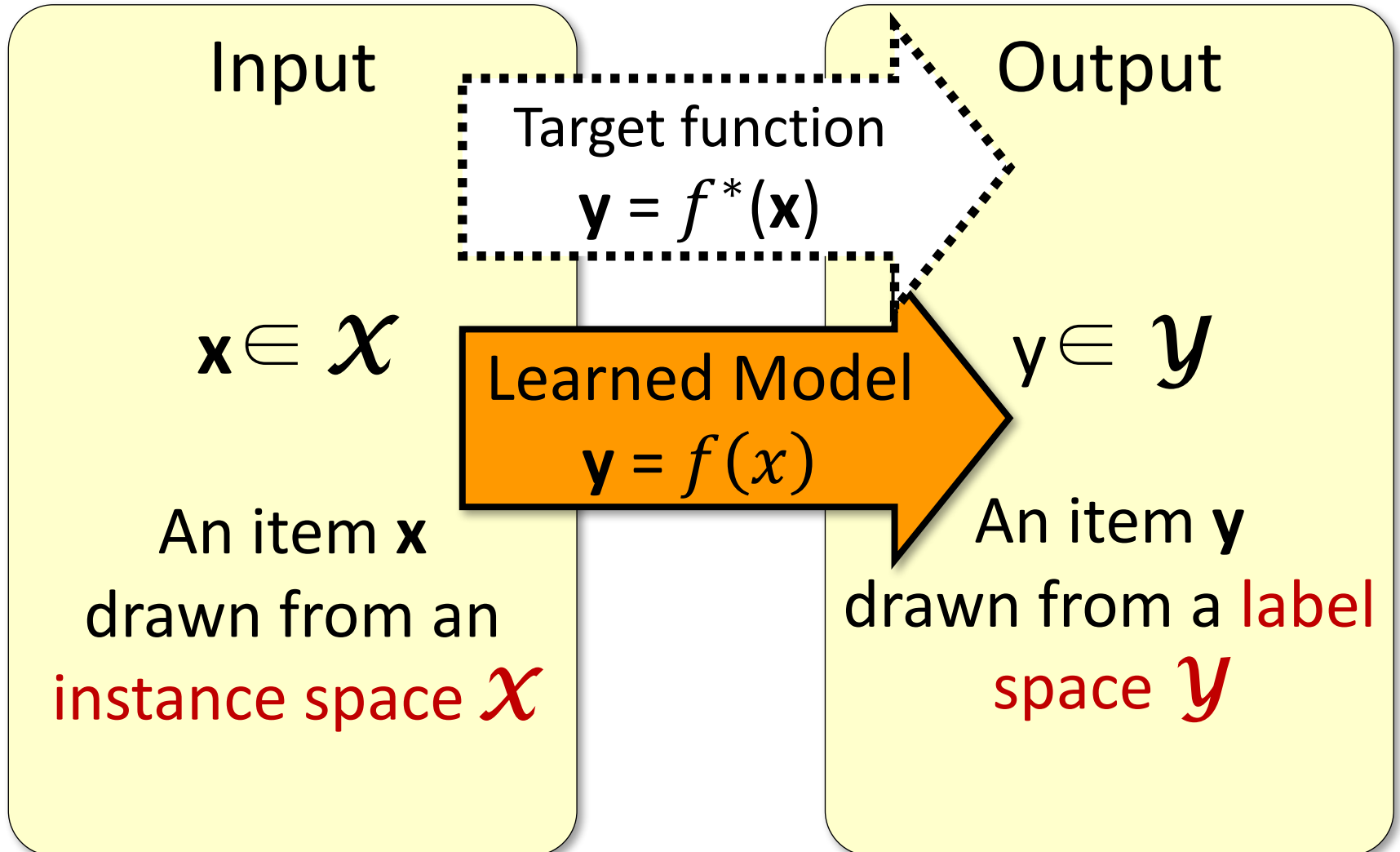
- Introduction to CRF
- Constrained conditional model

❖ Part 2: 16:00-17:30

Advanced topics

- Efficient inference/learning
- Learning from Indirect supervision signal
- Reducing human biases in structured models

Supervised learning



Supervised learning

y is represented in output space
(label space)

Different kinds of output:

- Binary classification:

$$y \in \{-1, 1\}$$

- Multiclass classification:

$$y \in \{1, 2, 3, \dots, K\}$$

- Regression:

$$y \in \mathbb{R}$$

- Structured output

$$y \in \{1, 2, 3, \dots, K\}^N$$

Output

$$y \in \mathcal{Y}$$

An item y
drawn from a label
space \mathcal{Y}

Combinatorial optimization problem

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} f(y; \mathbf{w}, \mathbf{x})$$

input

model parameters

output space

- ❖ Inference/Test: given \mathbf{w}, \mathbf{x} , solve argmax
- ❖ Learning/Training: find a good \mathbf{w}

Binary Linear Classifiers

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} f(y; \mathbf{w}, \mathbf{x})$$

$$\diamond \mathbf{x} \in \mathbb{R}^n, \mathcal{Y} = \{-1, 1\}$$

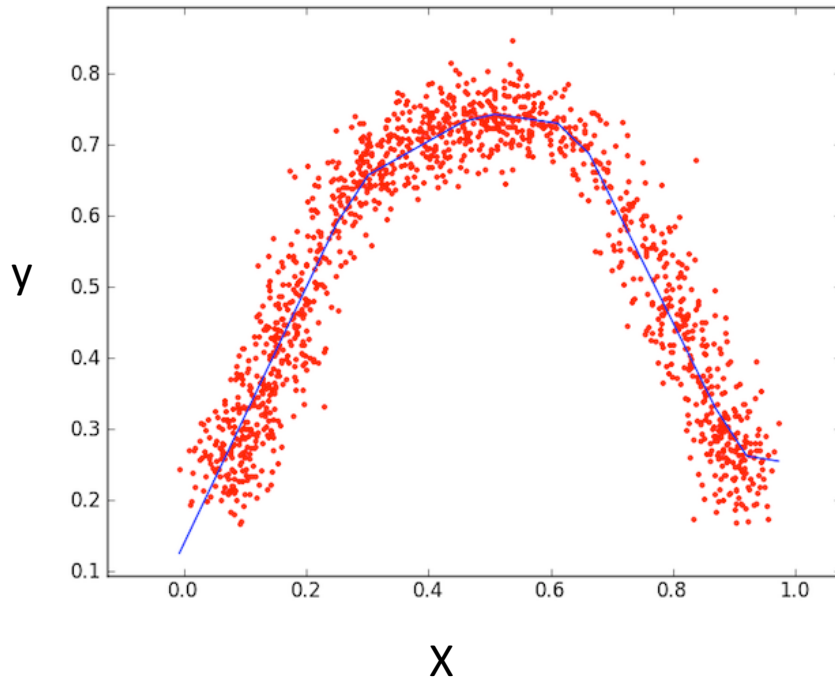
$$\diamond f(y; \mathbf{w}, \mathbf{x}) \stackrel{\text{def}}{=} y(\mathbf{w}^\top \mathbf{x} + \mathbf{b}) = y(\sum_i w_i x_i + \mathbf{b})$$

$$\diamond \operatorname{argmax}_{y \in \mathcal{Y}} f(y; \mathbf{w}, \mathbf{x}) = \begin{cases} 1, & \mathbf{w}^\top \mathbf{x} + b \geq 0 \\ -1, & \mathbf{w}^\top \mathbf{x} + b < 0 \end{cases} \\ = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + \mathbf{b})$$

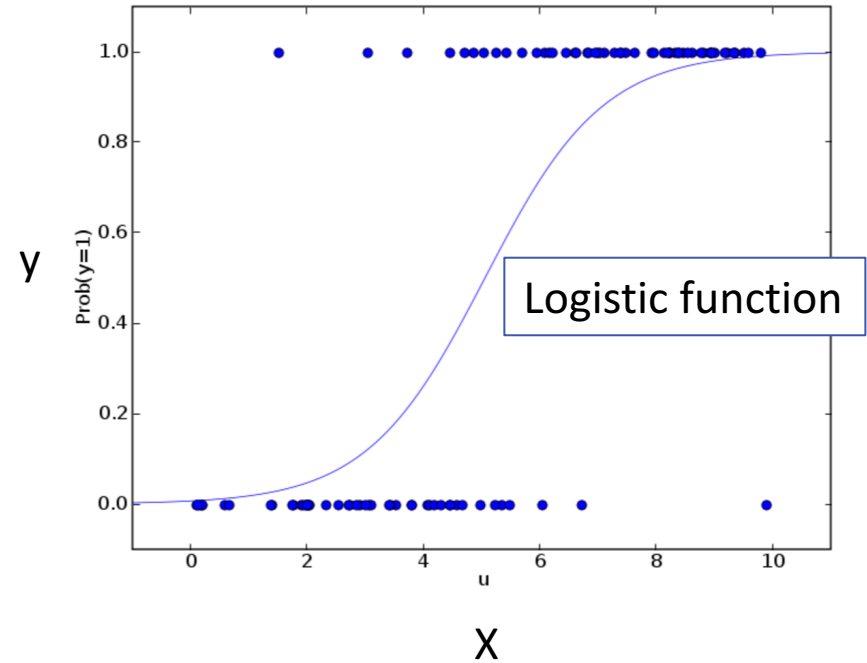
(break ties arbitrarily)

Recap: Logistic Regression

Regression



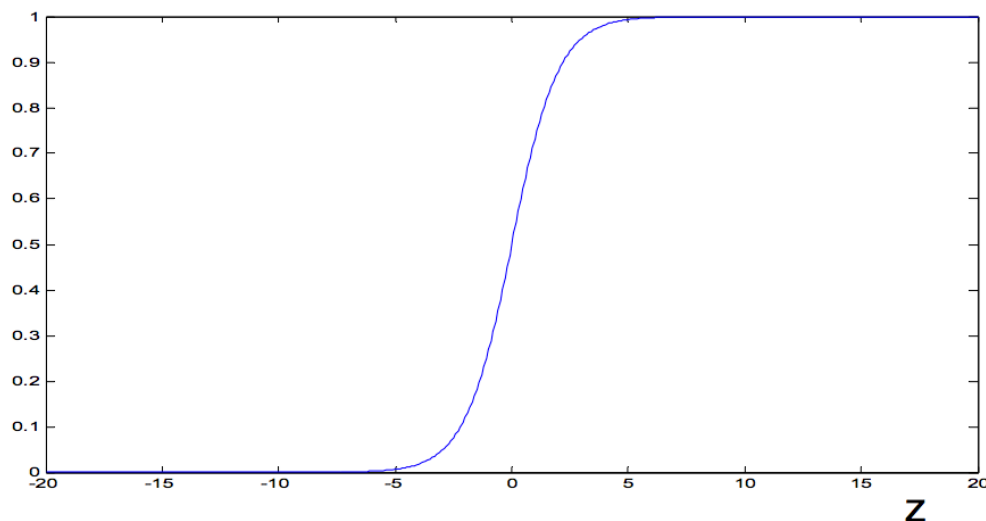
Logistic regression



logistic function or sigmoid function

- ❖ When $z \rightarrow \infty$ what is $\sigma(z)$?
- ❖ When $z \rightarrow -\infty$ what is $\sigma(z)$?
- ❖ When $z = 0$ what is $\sigma(z)$?

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Probabilistic Interpretation

Assume labels are generated using the following probability distribution:

$$P(y = 1 | \mathbf{x}, \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}}}{1 + e^{\mathbf{w}^T \mathbf{x}}} = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$
$$P(y = -1 | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}}}$$

$$P(y = 1 | \mathbf{x}, \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}}}{1 + e^{\mathbf{w}^T \mathbf{x}}} = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$

$$P(y = -1 | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}}}$$



$$P(y | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y \mathbf{w}^T \mathbf{x})}$$

$$P(y | x, w) = \sigma(y w^T x)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

How to make prediction?

Predict $y=1$ if $P(y=1 | x, w) > P(y= -1 | x, w)$

$$P(y = 1 | \mathbf{x}, \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}}}{1 + e^{\mathbf{w}^T \mathbf{x}}} = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$
$$P(y = -1 | \mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}}}$$

$$\log \frac{P(y = 1 | x, w)}{P(y = -1 | x, w)} = w^T x$$

❖ The decision boundary?

$$w^T x > 0$$

How to learn models

-- Maximum likelihood estimation

❖ Which bag of words more likely generate:

aaaDaaaKoaaaa



Maximum likelihood estimation

- ❖ Probabilistic model assumption:

$$P(y|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y\mathbf{w}^T \mathbf{x})}$$

- ❖ The log-likelihood of seeing a dataset $D = \{(x_i, y_i)\}$ if the true weight vector was w :

$$\log P(D|\mathbf{w}) = - \sum \log (1 + \exp(-y\mathbf{w}^T \mathbf{x}))$$

$$\begin{aligned} P(D|w) &= \prod_i P(y_i|x_i, w) \\ \Rightarrow \log P(D|w) &= \sum_i \log P(y_i|x_i, w) \end{aligned}$$

Minimizing negative log-likelihood

❖ Log likelihood

$$\log P(D|\mathbf{w}) = - \sum \log (1 + \exp(-y\mathbf{w}^T \mathbf{x}))$$

❖ Logistic regression

$$\min_{\mathbf{w}} \sum_i \log(1 + e^{-y_i(\mathbf{w}^T \mathbf{x}_i)})$$

❖ Let's add some prior (Gaussian Prior)

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_i \log(1 + e^{-y_i(\mathbf{w}^T \mathbf{x}_i)})$$

(multi-class) log-linear model

❖ Assumption:

$$P(y|x, w) = \frac{\exp(w_y^T x)}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x)}$$

Partition function



❖ This is a valid probability assumption. Why?

This often called soft-max

Softmax

- ❖ Softmax: let $s(y)$ be the score for output y
here $s(y) = w_y^T x$, but it can be computed by other metric.

$$P(y) = \frac{\exp(s(y))}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(s(y'))}$$

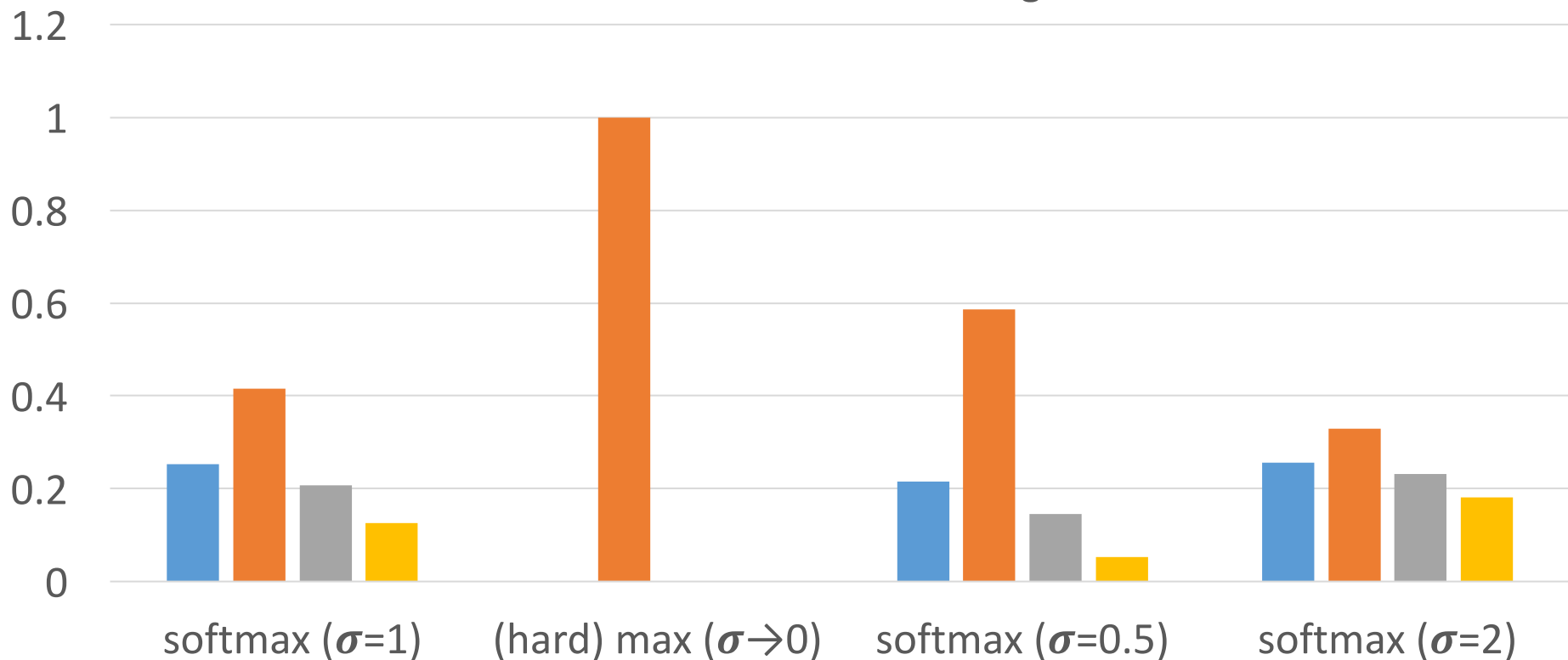
- ❖ We can control the peakedness of the distribution

$$P(y|\sigma) = \frac{\exp(s(y)/\sigma)}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(s(y')/\sigma)}$$

Temperature (example)

$$P(y|\sigma) = \frac{\exp(s(y)/\sigma)}{\sum_{y' \in \{1,2,\dots,K\}} \exp(s(y')/\sigma)}$$

■ Dog ■ Cat ■ Mouse ■ Duck



Maximum log-likelihood estimation

- ❖ Training can be done by maximum log-likelihood estimation i.e. $\max_w \log P(D|w)$

$$D = \{(x_i, y_i)\}$$

$$P(D|w) = \prod_i \frac{\exp(w_{y_i}^T x_i)}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x_i)}$$

$$\log P(D|w) = \sum_i [w_{y_i}^T x_i - \log \sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x_i)]$$

Maximum a posteriori

$$D = \{(x_i, y_i)\}$$

$$P(w|D) \propto P(w)P(D|w)$$

$$\max_w -\frac{1}{2} \sum_y w_y^T w_y + C \sum_i [w_{y_i}^T x_i - \log \sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x_i)]$$

or

$$\min_w \frac{1}{2} \sum_y w_y^T w_y + C \sum_i [\log \sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x_i) - w_{y_i}^T x_i]$$

(multi-class) log-linear model

❖ Assumption:

Partition function



$$P(y|x, w) = \frac{\exp(w_y^T x)}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w_{y'}^T x)}$$

❖ Another way to write this (with Kesler construction) is

$$P(y|x, w) = \frac{\exp(w^T \phi(x, y))}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w^T \phi(x, y'))}$$

Kesler construction

Assume we have a multi-class problem with K class and n features.

$$w_i^T x$$

models:

$$w_1, w_2, \dots, w_K,$$

$$w_k \in R^n$$

❖ Input:

$$x \in R^n$$

$$w^T \phi(x, i)$$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_y \\ \vdots \\ w_n \end{bmatrix}_{nK \times 1}$$

$$\phi(x, y) = \begin{bmatrix} 0_n \\ \vdots \\ x \\ \vdots \\ 0_n \end{bmatrix}_{nK \times 1}$$

x in y^{th} block;
Zeros elsewhere

$$w^T \phi(x, y) = w_y^T x$$

log-linear model

Partition function



❖ Assumption:

$$P(y|x, w) = \frac{\exp(w^T \phi(x, y))}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w^T \phi(x, y'))}$$

❖ Learning:

$$\min_w \frac{1}{2} w^T w + C \sum_i [\log \sum_{y' \in \{1, 2, \dots, K\}} \exp(w^T \phi(x_i, y')) - w^T \phi(x_i, y_i)]$$

How can we predict?

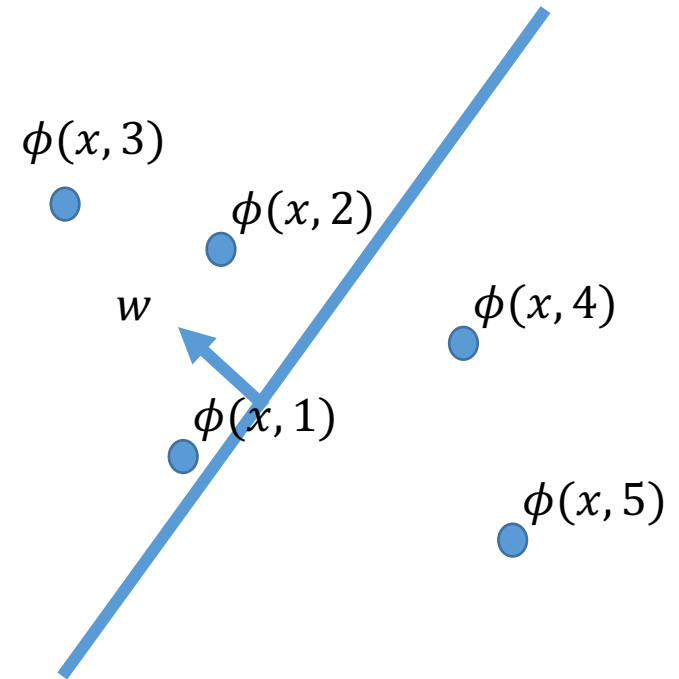
$$\operatorname{argmax}_y P(y | x, w)$$

$$\operatorname{argmax}_y w^T \phi(x, y)$$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_y \\ \vdots \\ w_n \end{bmatrix}_{nK \times 1} \quad \phi(x, y) = \begin{bmatrix} 0_n \\ \vdots \\ x \\ \vdots \\ 0_n \end{bmatrix}_{nK \times 1}$$

$$P(y|x, w) = \frac{\exp(w^T \phi(x, y))}{\sum_{y' \in \{1, 2, \dots, K\}} \exp(w^T \phi(x, y'))}$$

For input an input x ,
the model predict label is 3



This is equivalent to
 $\operatorname{argmax}_{y \in \{1, 2, \dots, K\}} w_y^T x$

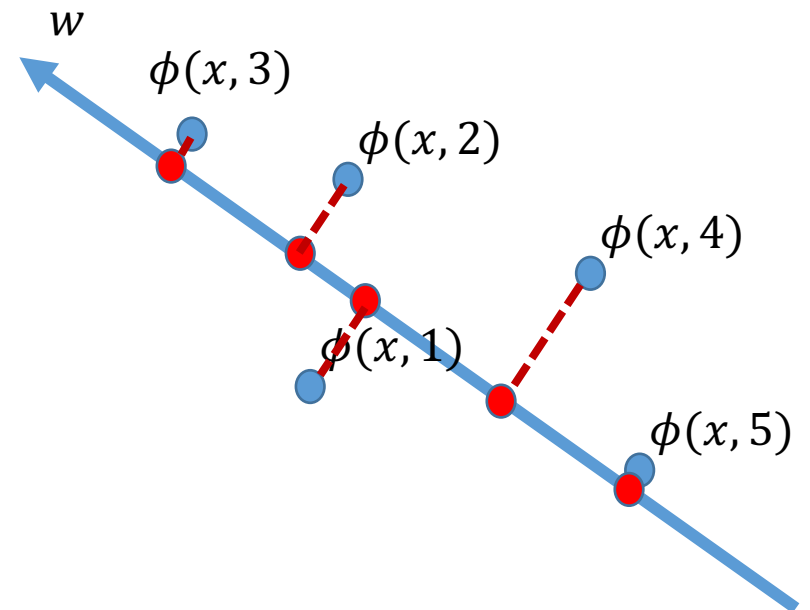
How can we predict?

$$\operatorname{argmax}_y w^T \phi(x, y)$$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_y \\ \vdots \\ w_n \end{bmatrix}_{nK \times 1}$$

$$\phi(x, y) = \begin{bmatrix} 0_n \\ \vdots \\ x \\ \vdots \\ 0_n \end{bmatrix}_{nK \times 1}$$

For input an input x ,
the model predict label is 3



Combinatorial optimization problem

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} f(y; \mathbf{w}, \mathbf{x})$$

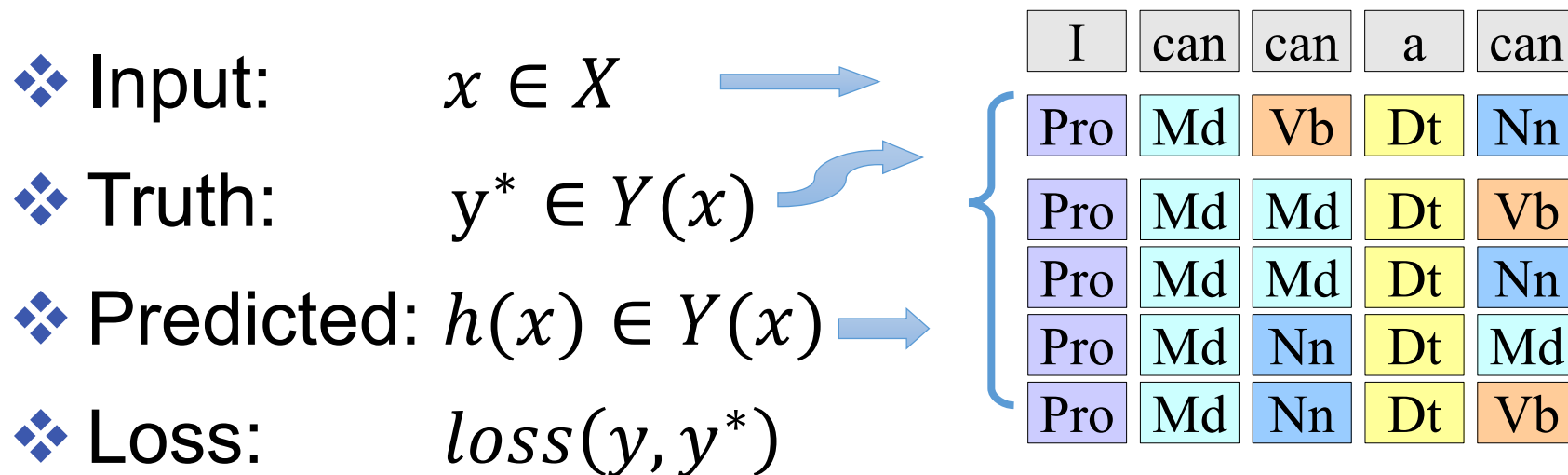
input

model parameters

output space

- ❖ Inference/Test: given \mathbf{w}, \mathbf{x} , solve argmax
- ❖ Learning/Training: find a good \mathbf{w}

Structured Prediction



Goal: make joint prediction to minimize a joint loss

find $h \in H$ such that $h(x) \in Y(x)$

minimizing $E_{(x,y) \sim D} [loss(y, h(x))]$ based on N samples $(x_n, y_n) \sim D$

General log-linear model

❖ Assumption:

$$P(y|x, w) = \frac{\exp(w^T \phi(x, y))}{\sum_{y' \in Y} \exp(w^T \phi(x, y'))}$$

Partition function

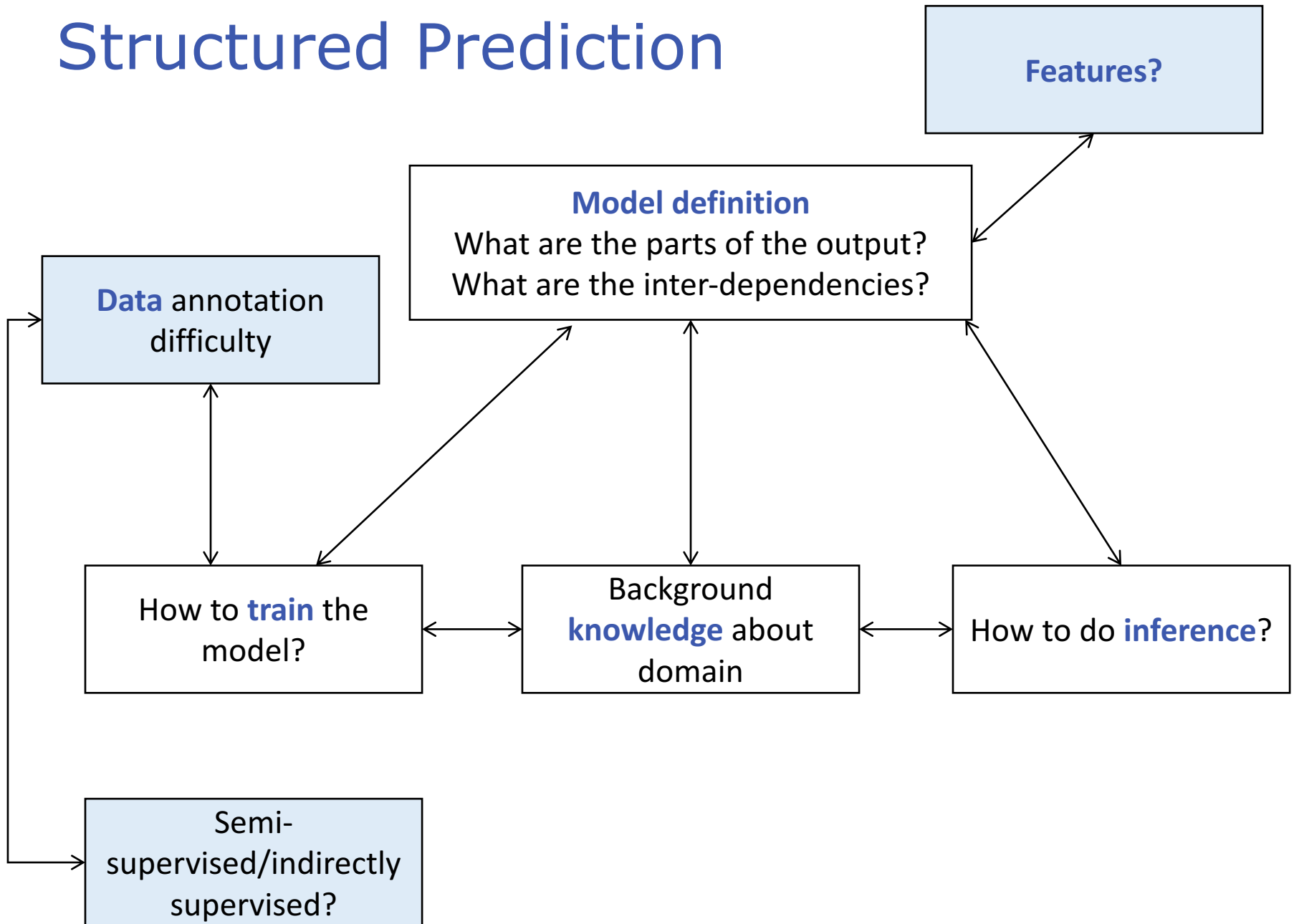


Summation over exponentially large output space

Challenges with structured output

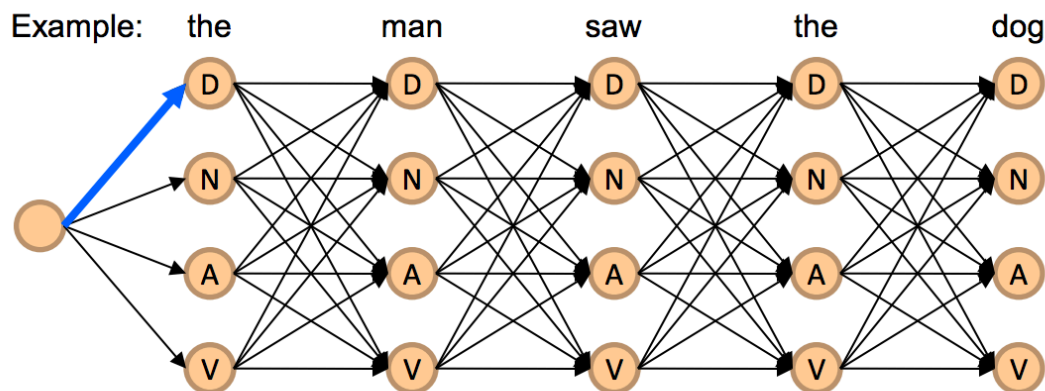
- ❖ We cannot train a separate weight vector for each possible inference outcome (why?)
 - ❖ For multi-class we train one weight vector for each class
- ❖ We cannot enumerate all possible structures for inference
 - ❖ Inference for multiclass was easy

Structured Prediction

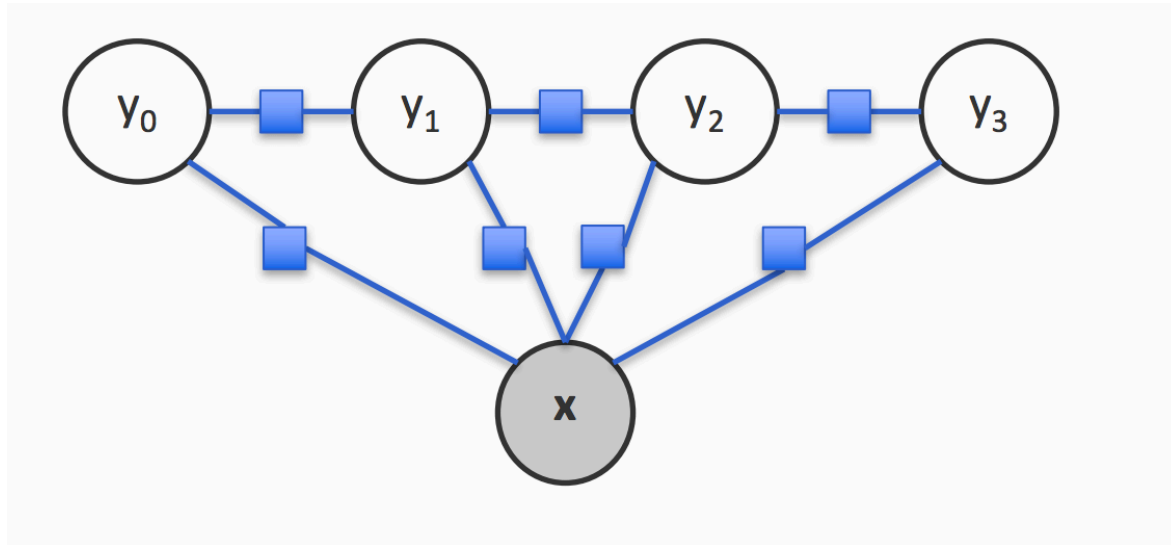


Deal with combinatorial output

- ❖ Decompose the output into **parts** that are labeled
- ❖ Define a **graph** to represent **(independent assumption)**
 - ❖ how the parts **interact** with each other
 - ❖ These labeled interacting parts are **scored**; the total score for the graph is the sum of scores of each part
 - ❖ an **inference algorithm** to assign labels to all the parts



Conditional Random Field: Factor graph

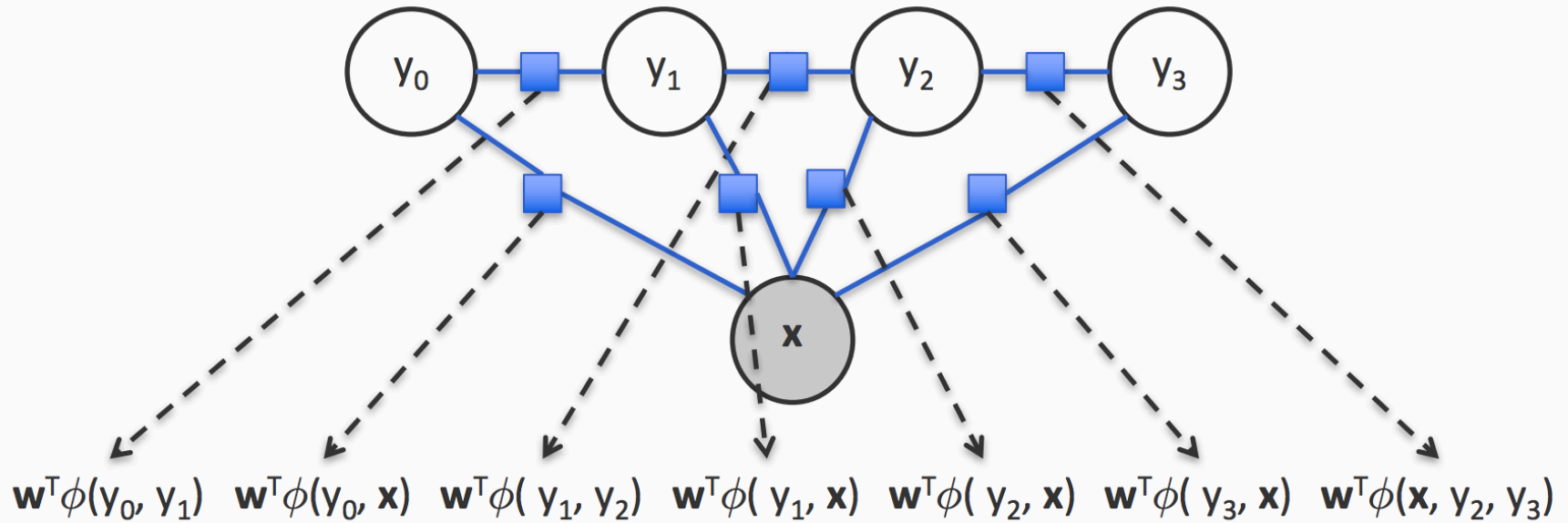


Each node is a random variable

We observe some nodes and need to assign the rest

Each factor is associated with a score

Conditional Random Field: Factor graph



Each node is a random variable

We observe some nodes and need to assign the rest

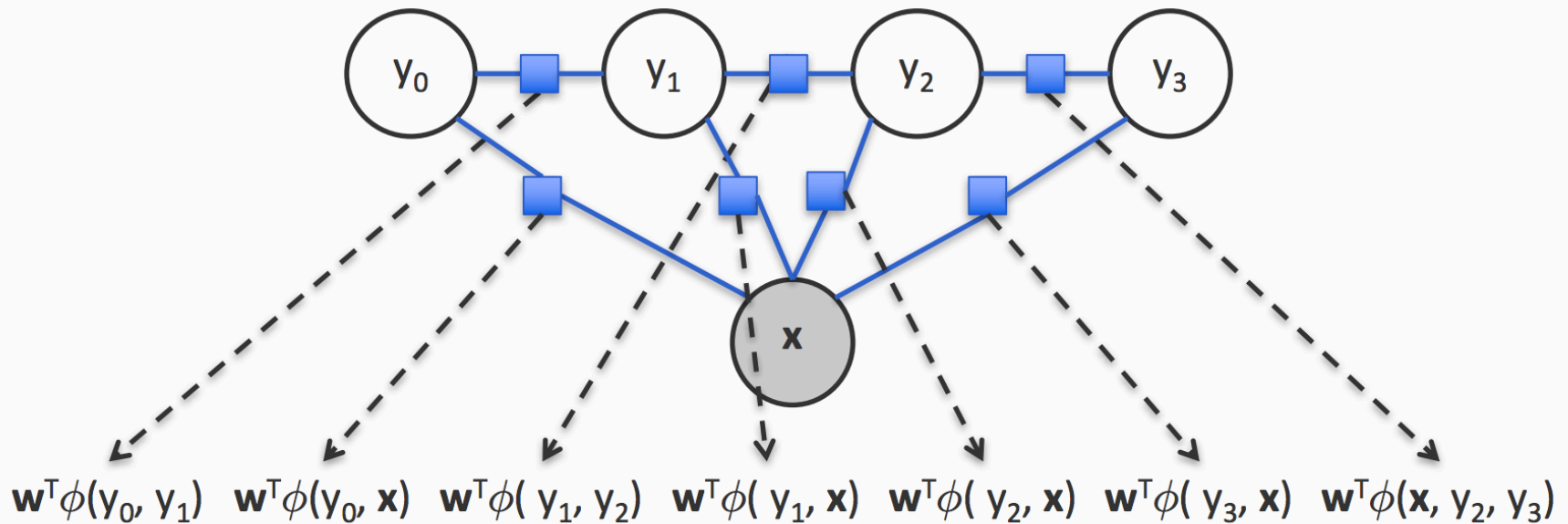
Each factor is associated with a score

Conditional Random Field for sequences

$$P(\mathbf{y} | \mathbf{x}) = \frac{1}{Z} w^T \phi(\mathbf{x}, y_0) \prod_i \exp(w^T \phi(y_i, y_{i-1}) + w^T \phi(\mathbf{x}, y_i))$$

Z: Normalizing constant,
sum over all sequences

$$Z = \sum_y w^T \phi(\mathbf{x}, y_0) \prod_i \exp(w^T \phi(y_i, y_{i-1}) + w^T \phi(\mathbf{x}, y_i))$$



Conditional Random Field for sequences

$$P(\mathbf{y} | \mathbf{x}) = \frac{1}{Z} w^T \phi(\mathbf{x}, \mathbf{y}_0) \prod_i \exp(w^T \phi(\mathbf{y}_i, \mathbf{y}_{i-1}) + w^T \phi(\mathbf{x}, \mathbf{y}_i))$$

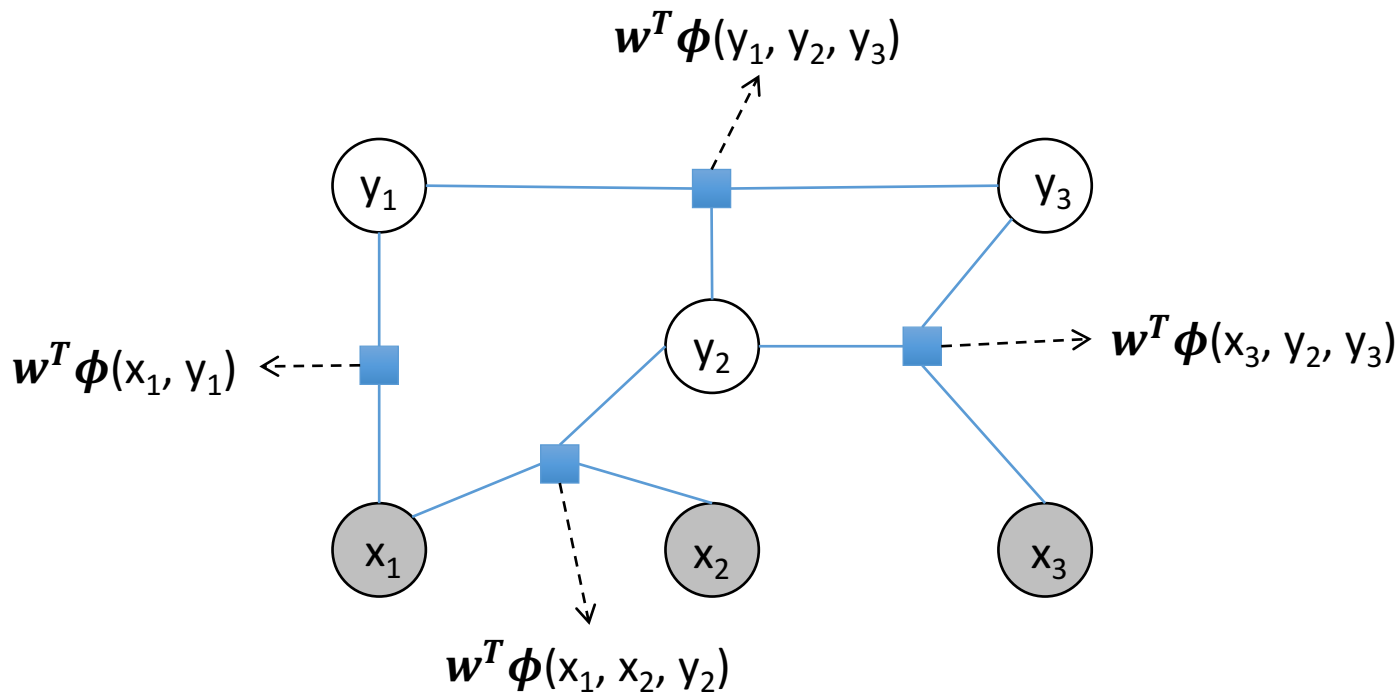
Z: Normalizing constant,
sum over all sequences

$$Z = \sum_{\mathbf{y}} w^T \phi(\mathbf{x}, \mathbf{y}_0) \prod_i \exp(w^T \phi(\mathbf{y}_i, \mathbf{y}_{i-1}) + w^T \phi(\mathbf{x}, \mathbf{y}_i))$$

With this strong independent assumption,
 z and $\operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{y}; \mathbf{w}, \mathbf{x})$ can be
estimated by a dynamic programming alg.

$w^T \phi(\mathbf{y}_0, \mathbf{y}_1)$ $w^T \phi(\mathbf{y}_0, \mathbf{x})$ $w^T \phi(\mathbf{y}_1, \mathbf{y}_2)$ $w^T \phi(\mathbf{y}_1, \mathbf{x})$ $w^T \phi(\mathbf{y}_2, \mathbf{x})$ $w^T \phi(\mathbf{y}_3, \mathbf{x})$ $w^T \phi(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_3)$

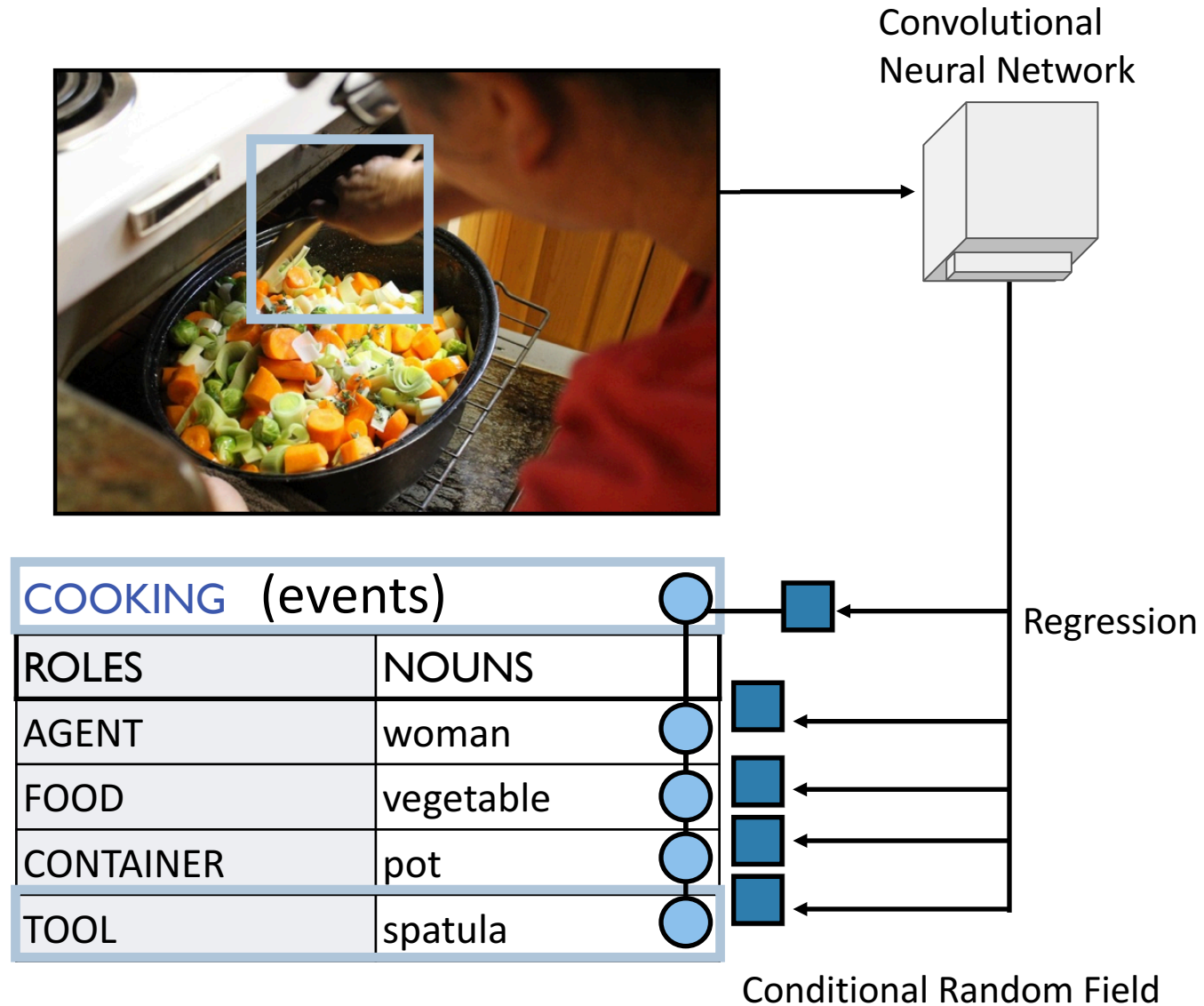
General CRFs



$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}))}{\sum_{\hat{\mathbf{y}}} \exp(\mathbf{w}^T \phi(\mathbf{x}, \hat{\mathbf{y}}))}$$

$$\phi(\mathbf{x}, \mathbf{y}) = \phi(x_1, y_1) + \phi(y_1, y_2, y_3) + \phi(x_3, y_2, y_3) + \phi(x_1, x_2, y_2)$$

CRF can be incorporated with deep learning



Kai-Wei Chang (kwchang.net/talks/sp.html)

Learning in log-linear model

❖ Assumption:

Partition function

$$P(y|x, w) = \frac{\exp(w^T \phi(x, y))}{\sum_{y' \in Y} \exp(w^T \phi(x, y'))}$$

❖ Learning:

Summation over exponentially large output space

$$\min_w \frac{1}{2} w^T w + C \sum_i [\log \sum_{y' \in Y} \exp(w^T \phi(x_i, y')) - w^T \phi(x_i, y_i)]$$

Computational questions

1. **Learning:** Given a training set $\{\langle \mathbf{x}_i, \mathbf{y}_i \rangle\}$

- ❖ Train via maximum likelihood (typically regularized)

$$\max_{\mathbf{w}} \sum_i \log P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{w}) = \max_{\mathbf{w}} \sum_i \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) - \log Z_{\mathbf{w}}(\mathbf{x}_i)$$

- ❖ Need to compute partition function during training

$$Z_{\mathbf{w}}(\mathbf{x}_i) = \sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}))$$

⋮

Computational questions

1. Learning: Given a training set $\{<\mathbf{x}_i, \mathbf{y}_i>\}$

- ❖ Train via maximum likelihood (typically regularized)

$$\max_{\mathbf{w}} \sum_i \log P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{w}) = \max_{\mathbf{w}} \sum_i \mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}_i) - \log Z_{\mathbf{w}}(\mathbf{x}_i)$$

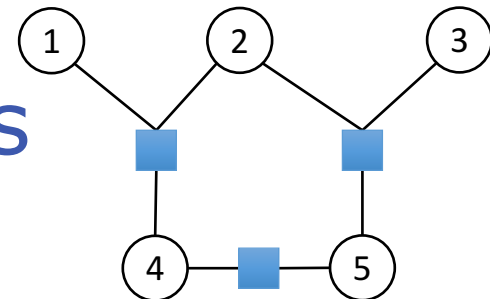
- ❖ Need to compute partition function during training

$$Z_{\mathbf{w}}(\mathbf{x}_i) = \sum_{\mathbf{y}} \exp(\mathbf{w}^T \phi(\mathbf{x}_i, \mathbf{y}))$$

2. Prediction: $\max_{\mathbf{y}} \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y})$

- ❖ Go over all possible assignments to the \mathbf{y} 's
- ❖ Find the one with the highest probability/score

Inference in graphical models



In general, compute probability of a subset of states

- ❖ $P(\mathbf{x}_A)$, for some subsets of random variables \mathbf{x}_A
- ❖ **Exact inference**
 - ❖ Variable elimination
 - ❖ Marginalize by summing out variables in a “good” order i
 - ❖ Belief propagation (exact only for graphs without loops)
 - ❖ Nodes pass messages to each other about their estimate of what the neighbor’s state should be
 - ❖ Generally efficient for trees, sequences (and maybe other graphs too)
- ❖ **“Approximate” inference**

Inference in graphical models

In general, compute probability of a subset of states

❖ $P(\mathbf{x}_A)$, for some subsets of random variables \mathbf{x}_A

❖ **Exact inference**

NP-hard in general, works for simple graphs

❖ **“Approximate” inference**

❖ **Markov Chain Monte Carlo**

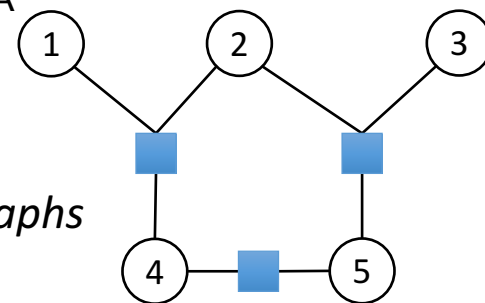
❖ Gibbs Sampling/Metropolis-Hastings

❖ Variational algorithms

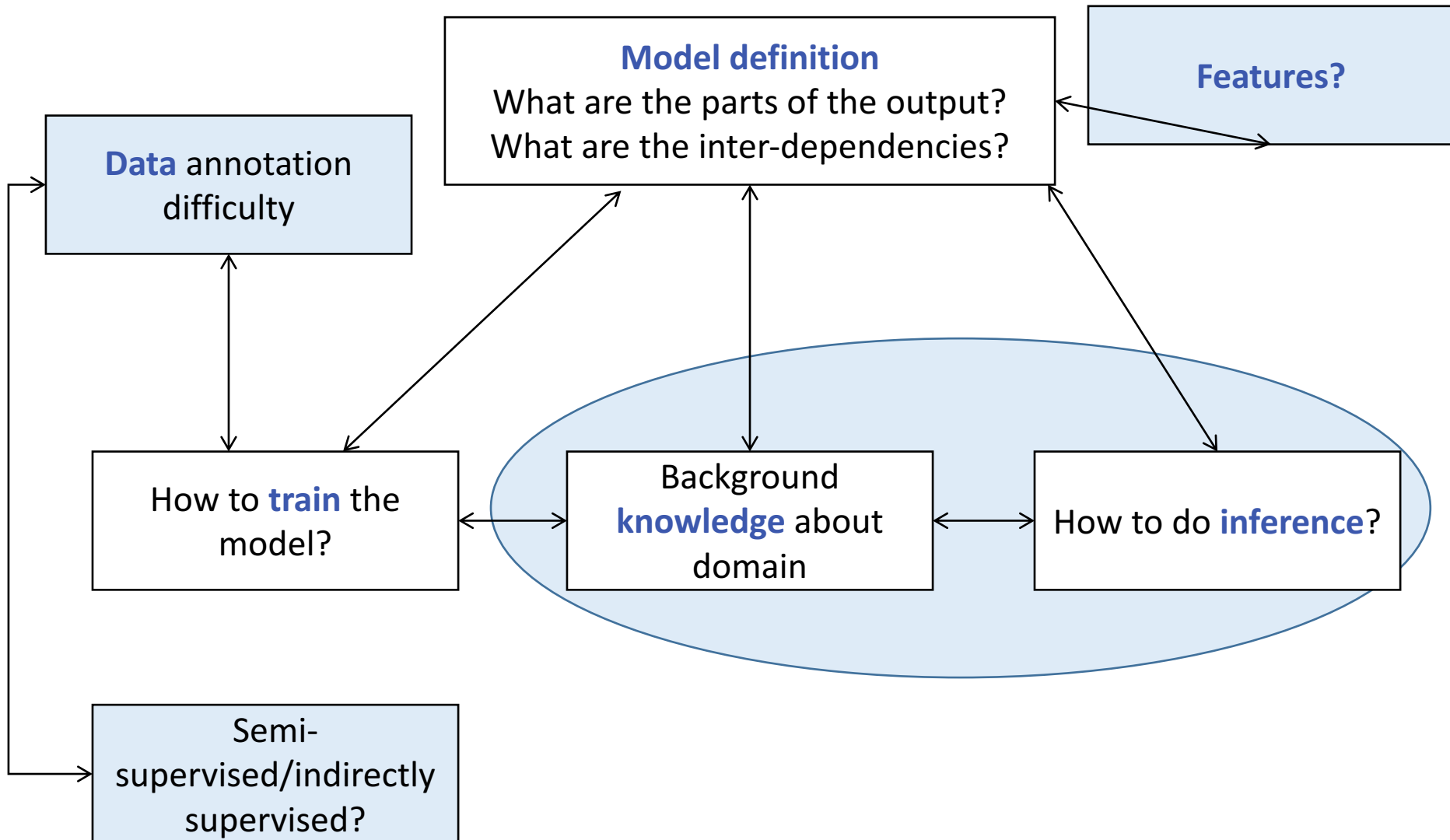
❖ Frame inference as an optimization problem, perturb it to an approximate one and solve the approximate problem

❖ Loopy Belief propagation

❖ Run BP and hope it works!



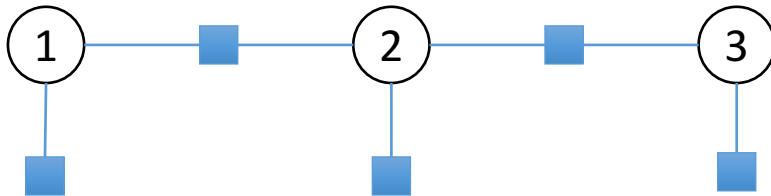
Constrained Conditional Model



Consistency of outputs

Or: How to introduce knowledge into prediction

Suppose we have a sequence labeling problem where the outputs can be one of A or B



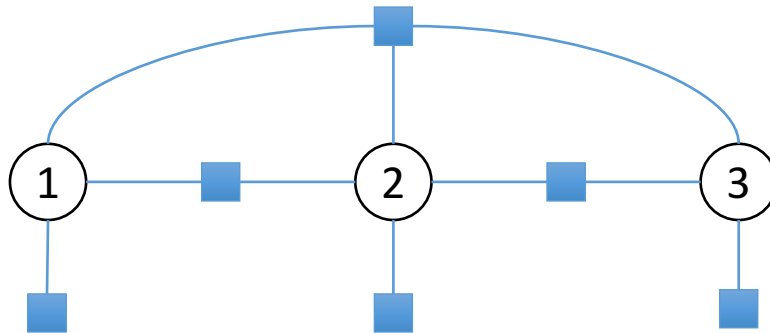
Consistency of outputs

Or: How to introduce knowledge into prediction

Suppose we have a sequence labeling problem where the outputs can be one of A or B

We want to add a condition:

There should be no more than one B in the output



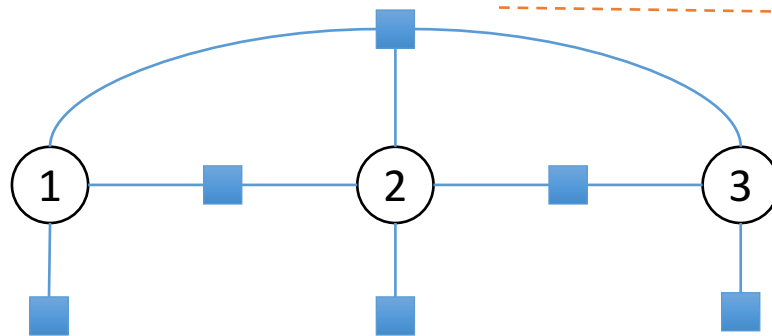
Consistency of outputs

Or: How to introduce knowledge into prediction

Suppose we have a sequence labeling problem where the outputs can be one of A or B

We want to add a condition:

There should be no more than one B in the output

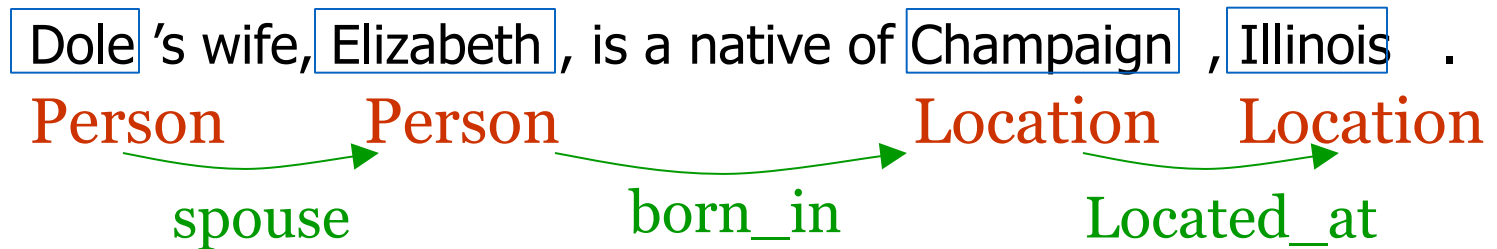


Potential function that ensures this condition

y1	y2	y3	f
A	A	A	0
A	A	B	0
A	B	A	0
A	B	B	-1
B	A	A	0
B	A	B	-1
B	B	A	-1
B	B	B	-1

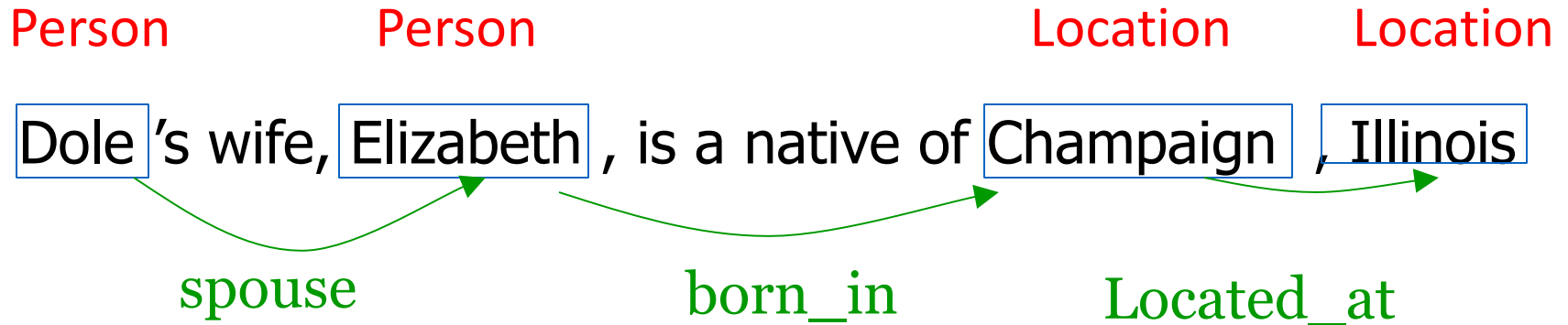
Should we learn what we can write down easily?
Especially for such large, computationally cumbersome factors

Entity Relation Extraction task



- ❖ Consistency constraint: A spouse relation can only hold between two person entities and cannot hold between two location entities

ILP Formulation for Entity Relation Task



Dole

PER	0.5
LOC	0.3
ORG	0.2

Elizabeth

PER	0.6
LOC	0.1
ORG	0.3

Dole-Elizabeth

spouse	0.7
born_in	0.1
Located_at	0.1
No-relation	0.1

ILP Formulation

Dole

PER	0.5	y_1
LOC	0.3	y_2
ORG	0.2	y_3

Elizabeth

PER	0.6	y_4
LOC	0.1	y_5
ORG	0.3	y_6

Dole-Elizabeth

spouse	0.7	y_7
born_in	0.1	y_8
Located_at	0.1	y_9
No-relation	0.1	y_{10}

$$\begin{aligned} & \text{maximize} \\ & 0.5y_1 + 0.3y_2 + 0.2y_3 + \\ & 0.6y_4 + 0.1y_5 + 0.3y_6 + \\ & 0.7y_7 + 0.1y_8 + 0.1y_9 + 0.1y_{10} \\ & \text{subj to } y_i \in \{0,1\} \\ & y_1 + y_2 + y_3 = 1 \\ & y_4 + y_5 + y_6 = 1 \\ & y_7 + y_8 + y_9 + y_{10} = 1 \\ & 2y_7 - y_1 - y_4 \leq 0 \end{aligned}$$

A spouse relation can only hold between two person entities

Amortized Inference for ILP

- ❖ We can write the ILP as

$$\max_{\mathbf{y}} \mathbf{c}\mathbf{y}$$

$$\mathbf{A}\mathbf{y} \leq \mathbf{b}$$

$$\mathbf{y}_i \in \{0, 1\}$$

- ❖ Inference problems discussed in previous sections can be represented as 0-1 ILPs.

maximize

$$0.5y_1 + 0.3y_2 + 0.2y_3 +$$

$$0.6y_4 + 0.1y_5 + 0.3y_6 +$$

$$0.7y_7 + 0.1y_8 + 0.1y_9 + 0.1y_{10}$$

subj to $y_i \in \{0, 1\}$

$$y_1 + y_2 + y_3 = 1$$

$$y_4 + y_5 + y_6 = 1$$

$$y_7 + y_8 + y_9 + y_{10} = 1$$

$$2y_7 - y_1 - y_4 \leq 0$$

Inference with constraints

- ❖ Combinatorial optimization problems can be often written as integer linear programs (ILP)
- ❖ The conversion is not always trivial
- ❖ Allows injection of “knowledge” in the form of constraints

- ❖ Different ways of solving ILPs
 - ❖ Commercial solvers: CPLEX, Gurobi, etc
 - ❖ Specialized solvers if you know something about your problem
 - ❖ Lagrangian relaxation, amortized inference, etc
 - ❖ Can approximate to linear programs and hope for the best

Integer linear programming

❖ In general

$$\begin{array}{ll} \max & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array}$$

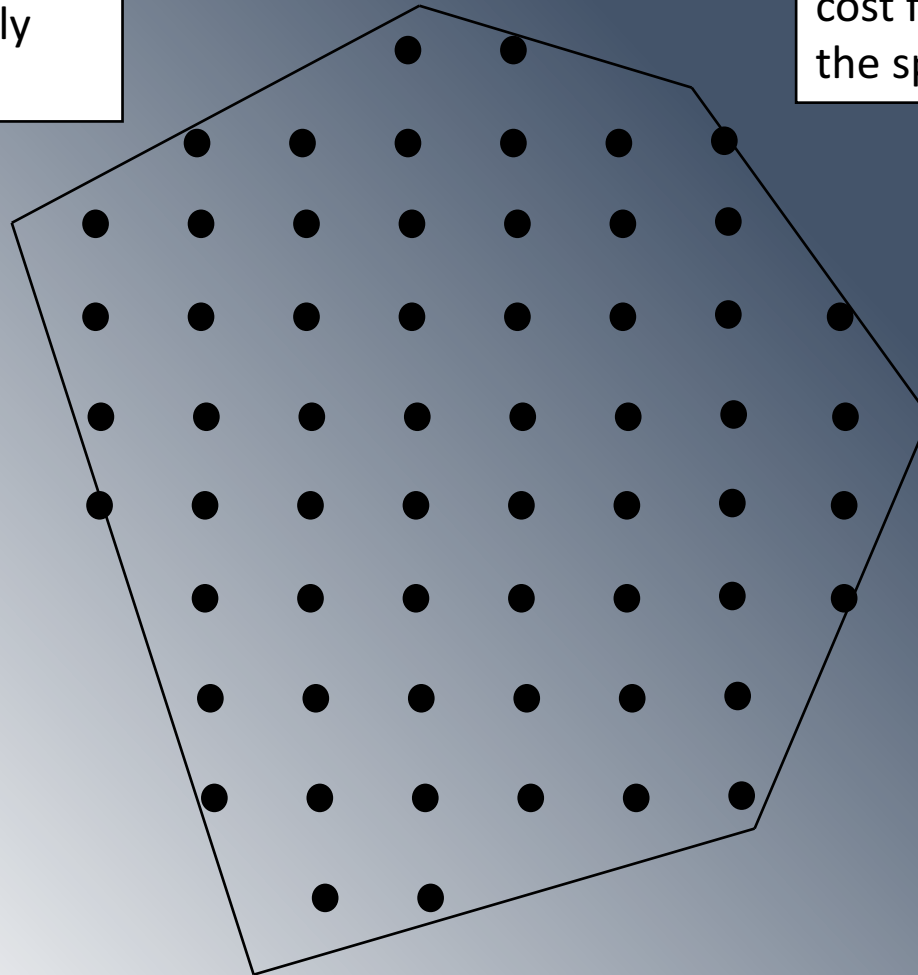
Each x_i is an integer.

Geometry of integer linear programming

The constraint matrix defines polytope that contains allowed solutions (possibly not closed)

The objective defines cost for every point in the space

Only integer points allowed



0-1 integer linear programming

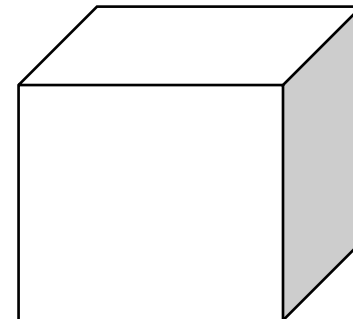
❖ In general

$$\begin{array}{ll} \max & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \\ & \mathbf{x} \in \{0, 1\}^n \end{array}$$

❖ An instance of integer linear programs

❖ Still NP-hard

❖ **Geometry:** We are only considering points that are vertices of the Boolean hypercube



Back to structured prediction

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} f(y; w, x)$$


output space
+ expert constraints

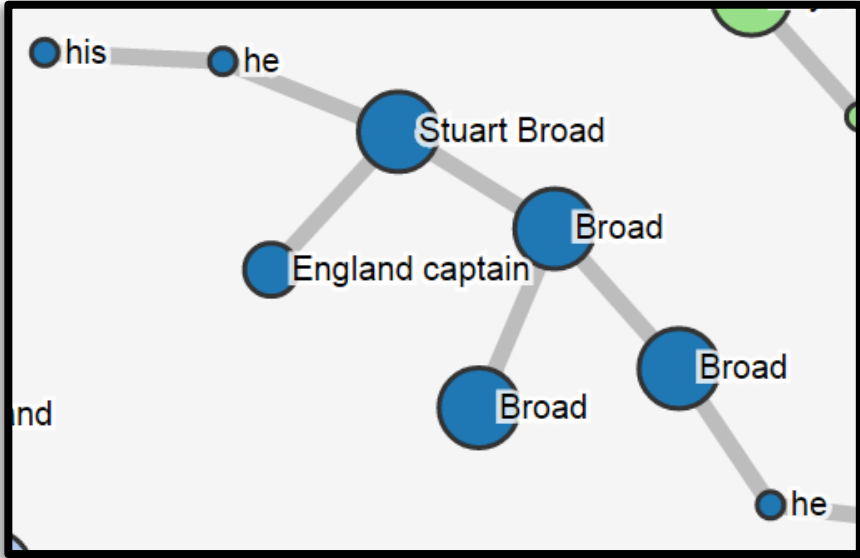
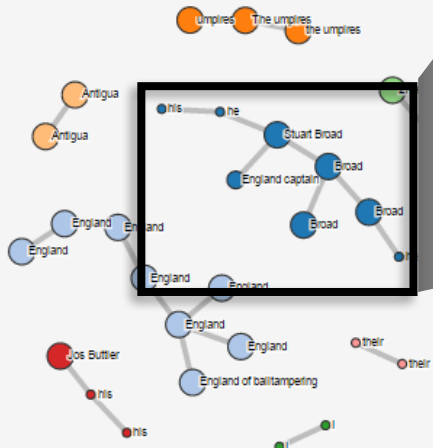
- ❖ General idea: Frame the argmax problem as a 0-1 integer linear program
- ❖ Allows addition of arbitrary constraints

Example application: Co-reference Resolution

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Co-reference Resolution Demo

[[England] captain] [Stuart Broad] says [he] is " baffled " by a decision to change the ball during [his] side ' s seriesclnching oneday win over West Indies . [The umpires] ruled it was showing " unnatural deterioration " as [England] sealed a 25run victory in the third and deciding match in [Antigua] . [I] saw no logic to it at all . [I] am baffled by it . " said [Broad] . " Throughout the three games , the ball was roughing up , little bits of leather were coming off it . " Former captain Bob Willis accused [England of balltampering] after [umpires] decided to change the ball during a Champions Trophy match against Sri Lanka in 2013 . [Broad] said that while [he] did not believe [the umpires] were making a similar allegation in [Antigua] , [he] was confused as to why they intervened . " It ' s not like the ball was reverseswinging , " said [27yearold allrounder] , [who] was leading [England] for the first time in a oneday series in place of the rested Alastair Cook . " I bowled three crossseamers with the ball they gave us and the same year was arriving on that ball . I ' m very confused as to why it was changed and made my confusion well known . " Before the ball controversy , Joe Root hit a maiden oneday international century and [Jos Buttler] fell one run short of [his] first ODI ton as [England] posted 3036 . [England] then reduced the hosts to 433 before Denesh Ramdin threatened to pull off a remarkable fightback with a maiden century of [his] own . With 40 needed from the final three overs , Ramdin took 14 from three Tim Bresnan deliveries before being bowled by a yorker which sealed [England] ' s victory in [their] first series since [their] dismal Ashes tour . " This is going to lift the boys , " said [Broad] . " The guys have held themselves brilliantly to come away with a series win . [England] now play three Twenty20s against the same opposition in Barbados in preparation for the World T20 in Bangladesh . Root could miss the first game on Sunday after suffering a thumb injury when struck by a ball from pace bowler Ravi Rampaul .



Co-reference Resolution

- ❖ Learn a pairwise similarity score function (local predictor)

Example features:

- ❖ same sub-string?
- ❖ positions in the paragraph
- ❖ other 30+ feature types

- ❖ Key components:

- ❖ Pairwise classification
- ❖ Clustering (jointly or not?)

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Approaches

Approach1: Decoupling Approach (CoNLL11)

Learn a pairwise similarity function



Cluster based on this function



Approach2: Latent Structured Learning (ICML 14, EMNLP 13)



Update the similarity function



Cluster based on this function.



Baseline: Decoupling Approach

A heuristic to learn the model [Soon+ 01, Bengtson+ 08, CoNLL11]

❖ **Decouple** learning and inference:

Learn a pairwise similarity function

Cluster based on this function

Decoupling Approach-Learning

As a boy, **Chris**₁ lived in a pretty home called Cotchfield Farm. When **Chris**₂ was three years old, **his father**₃ wrote a poem about **him**₄. The poem was printed in a magazine for others to read. **Mr. Robin**₅ then wrote a book


Positive Samples

(**Chris**₁, **him**₄)
(**Chris**₂, **him**₄)
(**Chris**₁, **Chris**₂)
(**his father**₃, **Mr. Robin**₅)


Negative Samples

(**Chris**₁, **his father**₃)
(**Chris**₂, **his father**₃)
(**him**₄, **his father**₃)
(**Chris**₁, **Mr. Robin**₅)
(**Chris**₂, **Mr. Robin**₅)
(**him**₄, **Mr. Robin**₅)


Greedy Best-Left-Link Clustering




[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].



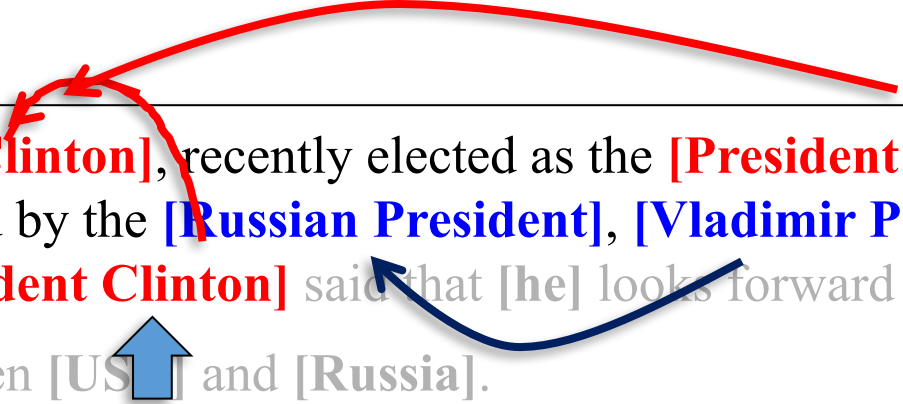
Greedy Best-Left-Link Clustering



[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, [Vladimir Putin], to visit [Russia]. [President Clinton]  that [he] looks forward to strengthening ties between [USA] and [Russia].

Greedy Best-Left-Link Clustering

[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, **[Vladimir Putin]**, to visit **[Russia]**. **[President Clinton]** said that **[he]** looks forward to strengthening ties between **[US]** and **[Russia]**.

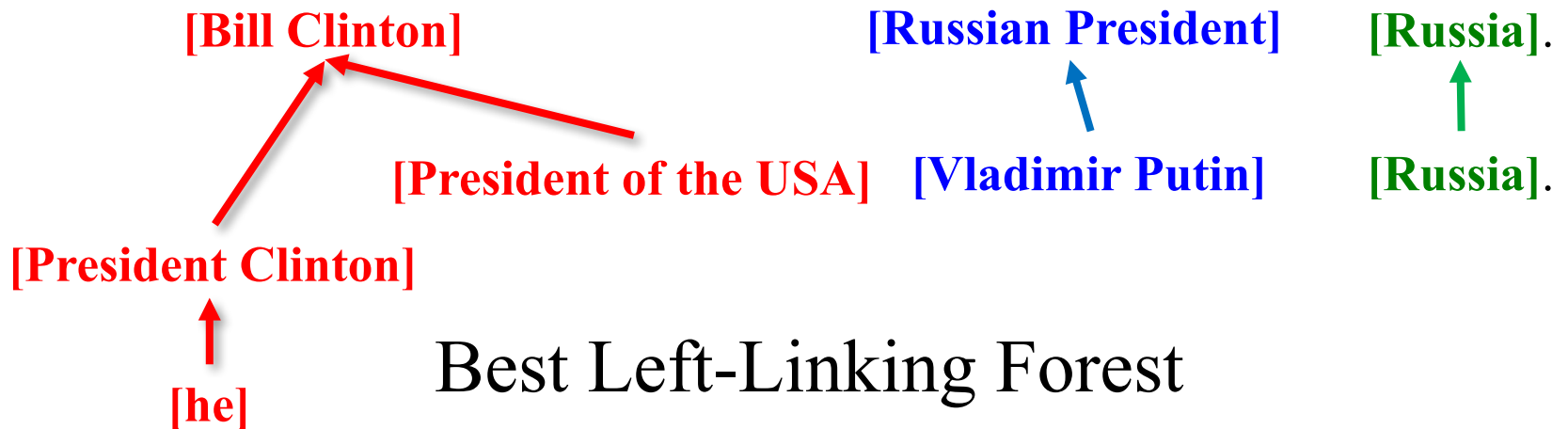


The diagram shows three arrows pointing to specific entity mentions in the text: a red arrow from the top points to "[Bill Clinton]"; a blue arrow from the right points to "[President Clinton]"; and a blue arrow from the bottom points to "[US]".

Greedy Best-Left-Link Clustering

[Soon+ 01, Bengtson+ 08, CoNLL11]

[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].

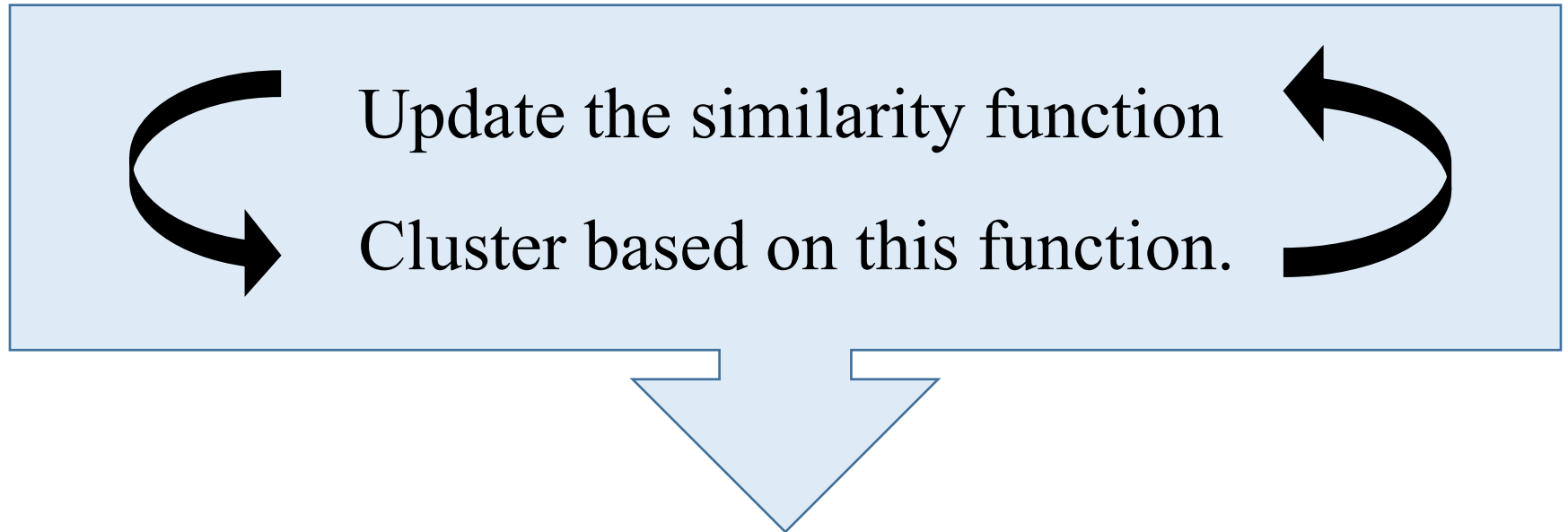


Can we do better?

❖ Decoupling may lose information

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Structured Learning Approach



Learn the similarity function while clustering

Attempt: All-Links Clustering

[Mccallum+ 04, CoNLL 11]

❖ Define a global scoring function:

Attempt: using all within-cluster pairs:

❖ Inference problem is too hard

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Latent Left-Linking Model (L3M)

[ICML 14, EMNLP 13]

Score (a clustering C)


= Score (the best left-linking forest that is consistent with C)

= \sum Score of edges in the forests

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Linguistic Constraints

- ❖ Must-link constraints:
 - ❖ E.g., **SameProperName**, ...
- ❖ Cannot-link constraints:
 - ❖ E.g., **ModifierMismatch**, ...



[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, **[Vladimir Putin]**, to visit **[Russia]**. **[President Clinton]** said that **[he]** looks forward to strengthening ties between **[USA]** and **[Russia]**.

- ❖ Clustering with constraints [(Basu+08, Zhi+14)]

Inference in L3M [ICML 14, EMNLP 13]

- ❖ Represented using an ILP formulation [Scott+ 2004/2007]
- ❖ Inference can be done using a greedy heuristics.

$y_{i,j} = 1 \Leftrightarrow i, j$ is an edge in the forest

$$\begin{array}{l} \arg \max_y \\ s. t \end{array} \quad \sum_c S_{i,j} y_{i,j} \quad \boxed{y_{i,j}}$$
$$\boxed{Ay \leq b; y_{i,j} \in \{0,1\}}$$

- Modeling constraints
- Linguistic constraints

Log linear model: Probabilistic L3M

[ICML 14, EMNLP 13a]

❖ Define a log-linear model

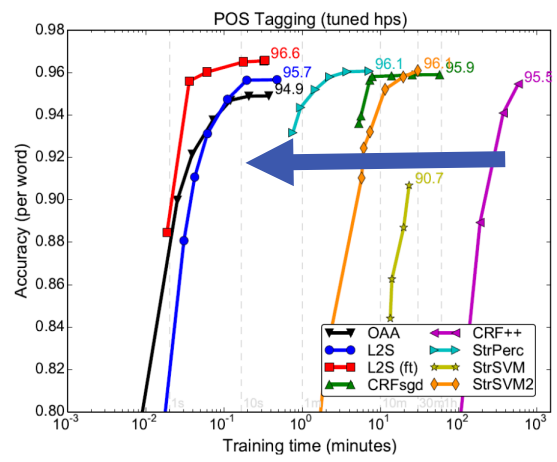
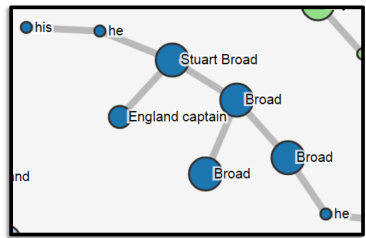
$$\begin{aligned} & \Pr [\text{a clustering } C] \\ &= \sum \Pr [\text{forests that are consistent with } C] \\ &= \sum \Pi \Pr [\text{edges in the forest}] \\ &= \Pi_i \sum_{j \in e(i)} \Pr [\text{edge}(j,i)] \\ & \Pr [\text{edge}(j,i)] \sim \exp(\mathbf{w} \cdot \phi(j,i) / \gamma) \quad (\gamma: \text{a parameter}) \end{aligned}$$

❖ Regularized Maximum Log-Likelihood Estimation:

$$\begin{aligned} \min_{\mathbf{w}} \text{LL}(\mathbf{w}) &= \beta \|\mathbf{w}\|^2 + \sum_d \log Z_d(\mathbf{w}) \\ &\quad - \sum_d \sum_i \log(\sum_{j < i} \exp(\mathbf{w} \cdot \phi(i,j) / \gamma) C_d(i,j)) \end{aligned}$$

Structured Prediction Models

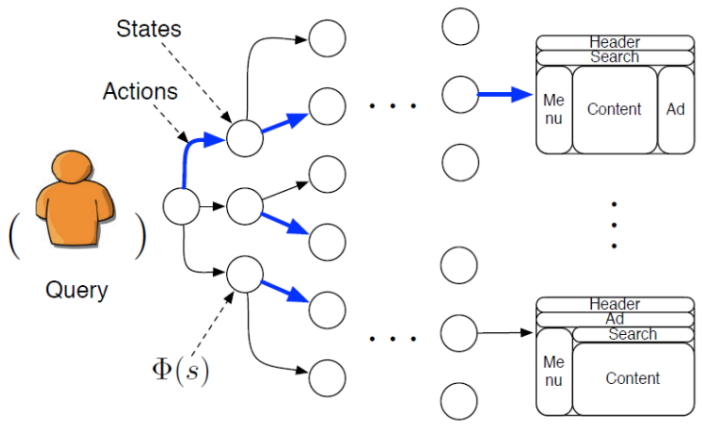
[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58] [59] [60] [61] [62] [63] [64] [65] [66] [67] [68] [69] [70] [71] [72] [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85] [86] [87] [88] [89] [90] [91] [92] [93] [94] [95] [96] [97] [98] [99] [100] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111] [112] [113] [114] [115] [116] [117] [118] [119] [120] [121] [122] [123] [124] [125] [126] [127] [128] [129] [130] [131] [132] [133] [134] [135] [136] [137] [138] [139] [140] [141] [142] [143] [144] [145] [146] [147] [148] [149] [150] [151] [152] [153] [154] [155] [156] [157] [158] [159] [160] [161] [162] [163] [164] [165] [166] [167] [168] [169] [170] [171] [172] [173] [174] [175] [176] [177] [178] [179] [180] [181] [182] [183] [184] [185] [186] [187] [188] [189] [190] [191] [192] [193] [194] [195] [196] [197] [198] [199] [200] [201] [202] [203] [204] [205] [206] [207] [208] [209] [210] [211] [212] [213] [214] [215] [216] [217] [218] [219] [220] [221] [222] [223] [224] [225] [226] [227] [228] [229] [230] [231] [232] [233] [234] [235] [236] [237] [238] [239] [240] [241] [242] [243] [244] [245] [246] [247] [248] [249] [250] [251] [252] [253] [254] [255] [256] [257] [258] [259] [260] [261] [262] [263] [264] [265] [266] [267] [268] [269] [270] [271] [272] [273] [274] [275] [276] [277] [278] [279] [280] [281] [282] [283] [284] [285] [286] [287] [288] [289] [290] [291] [292] [293] [294] [295] [296] [297] [298] [299] [300] [301] [302] [303] [304] [305] [306] [307] [308] [309] [310] [311] [312] [313] [314] [315] [316] [317] [318] [319] [320] [321] [322] [323] [324] [325] [326] [327] [328] [329] [330] [331] [332] [333] [334] [335] [336] [337] [338] [339] [340] [341] [342] [343] [344] [345] [346] [347] [348] [349] [350] [351] [352] [353] [354] [355] [356] [357] [358] [359] [360] [361] [362] [363] [364] [365] [366] [367] [368] [369] [370] [371] [372] [373] [374] [375] [376] [377] [378] [379] [380] [381] [382] [383] [384] [385] [386] [387] [388] [389] [390] [391] [392] [393] [394] [395] [396] [397] [398] [399] [400] [401] [402] [403] [404] [405] [406] [407] [408] [409] [410] [411] [412] [413] [414] [415] [416] [417] [418] [419] [420] [421] [422] [423] [424] [425] [426] [427] [428] [429] [430] [431] [432] [433] [434] [435] [436] [437] [438] [439] [440] [441] [442] [443] [444] [445] [446] [447] [448] [449] [450] [451] [452] [453] [454] [455] [456] [457] [458] [459] [460] [461] [462] [463] [464] [465] [466] [467] [468] [469] [470] [471] [472] [473] [474] [475] [476] [477] [478] [479] [480] [481] [482] [483] [484] [485] [486] [487] [488] [489] [490] [491] [492] [493] [494] [495] [496] [497] [498] [499] [500] [501] [502] [503] [504] [505] [506] [507] [508] [509] [510] [511] [512] [513] [514] [515] [516] [517] [518] [519] [520] [521] [522] [523] [524] [525] [526] [527] [528] [529] [530] [531] [532] [533] [534] [535] [536] [537] [538] [539] [540] [541] [542] [543] [544] [545] [546] [547] [548] [549] [550] [551] [552] [553] [554] [555] [556] [557] [558] [559] [560] [561] [562] [563] [564] [565] [566] [567] [568] [569] [570] [571] [572] [573] [574] [575] [576] [577] [578] [579] [580] [581] [582] [583] [584] [585] [586] [587] [588] [589] [590] [591] [592] [593] [594] [595] [596] [597] [598] [599] [600] [601] [602] [603] [604] [605] [606] [607] [608] [609] [610] [611] [612] [613] [614] [615] [616] [617] [618] [619] [620] [621] [622] [623] [624] [625] [626] [627] [628] [629] [630] [631] [632] [633] [634] [635] [636] [637] [638] [639] [640] [641] [642] [643] [644] [645] [646] [647] [648] [649] [650] [651] [652] [653] [654] [655] [656] [657] [658] [659] [660] [661] [662] [663] [664] [665] [666] [667] [668] [669] [670] [671] [672] [673] [674] [675] [676] [677] [678] [679] [680] [681] [682] [683] [684] [685] [686] [687] [688] [689] [690] [691] [692] [693] [694] [695] [696] [697] [698] [699] [700] [701] [702] [703] [704] [705] [706] [707] [708] [709] [710] [711] [712] [713] [714] [715] [716] [717] [718] [719] [720] [721] [722] [723] [724] [725] [726] [727] [728] [729] [730] [731] [732] [733] [734] [735] [736] [737] [738] [739] [740] [741] [742] [743] [744] [745] [746] [747] [748] [749] [750] [751] [752] [753] [754] [755] [756] [757] [758] [759] [760] [761] [762] [763] [764] [765] [766] [767] [768] [769] [770] [771] [772] [773] [774] [775] [776] [777] [778] [779] [780] [781] [782] [783] [784] [785] [786] [787] [788] [789] [790] [791] [792] [793] [794] [795] [796] [797] [798] [799] [800] [801] [802] [803] [804] [805] [806] [807] [808] [809] [810] [811] [812] [813] [814] [815] [816] [817] [818] [819] [820] [821] [822] [823] [824] [825] [826] [827] [828] [829] [830] [831] [832] [833] [834] [835] [836] [837] [838] [839] [840] [841] [842] [843] [844] [845] [846] [847] [848] [849] [850] [851] [852] [853] [854] [855] [856] [857] [858] [859] [860] [861] [862] [863] [864] [865] [866] [867] [868] [869] [870] [871] [872] [873] [874] [875] [876] [877] [878] [879] [880] [881] [882] [883] [884] [885] [886] [887] [888] [889] [890] [891] [892] [893] [894] [895] [896] [897] [898] [899] [900] [901] [902] [903] [904] [905] [906] [907] [908] [909] [910] [911] [912] [913] [914] [915] [916] [917] [918] [919] [920] [921] [922] [923] [924] [925] [926] [927] [928] [929] [930] [931] [932] [933] [934] [935] [936] [937] [938] [939] [940] [941] [942] [943] [944] [945] [946] [947] [948] [949] [950] [951] [952] [953] [954] [955] [956] [957] [958] [959] [960] [961] [962] [963] [964] [965] [966] [967] [968] [969] [970] [971] [972] [973] [974] [975] [976] [977] [978] [979] [980] [981] [982] [983] [984] [985] [986] [987] [988] [989] [990] [991] [992] [993] [994] [995] [996] [997] [998] [999] [1000]



How to model?

Training/test/dev speed

Query



activity	cooking
agent	woman
food	vegetable

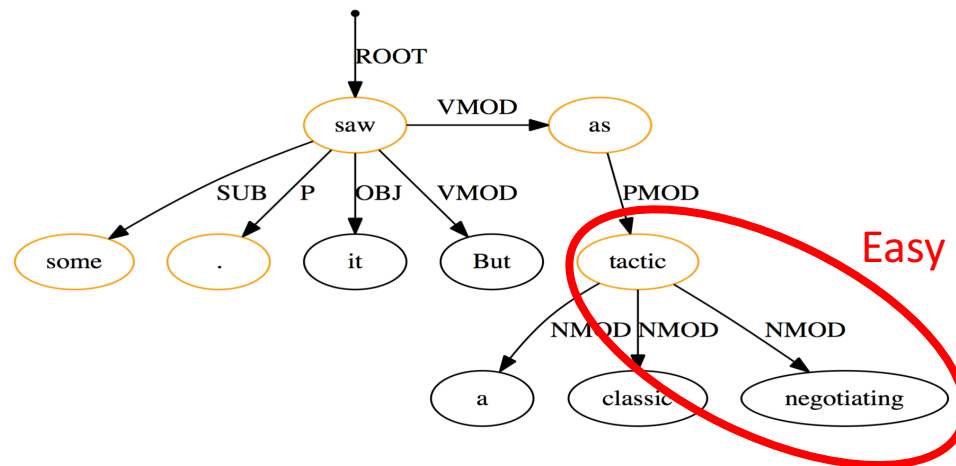
Learning signals

Fairness (data biases)

Idea 1: Adaptive feature selection [AAAI 17]

- ❖ **Observation:** some decisions are simpler than the others
- ❖ **Idea:** adaptively generate computationally costly features during test-time

But some saw it as a classic **negotiating tactic**



Idea 2: Amortized inference

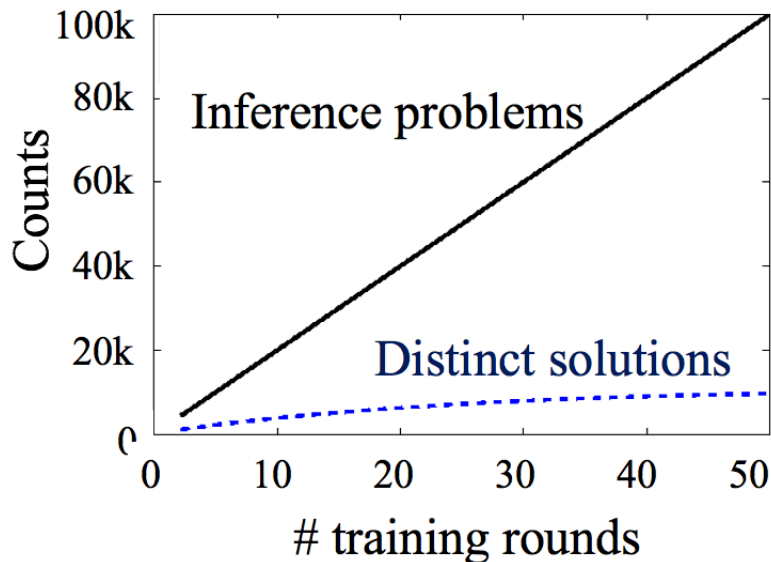
- ❖ Observation: Many inference problems share the same solution

S1	POS
He	Pronoun
is	VerbZ
reading	VerbG
a	Det
book	Noun

POS	S2
Pronoun	She
VerbZ	is
VerbG	watching
Det	a
Noun	movie

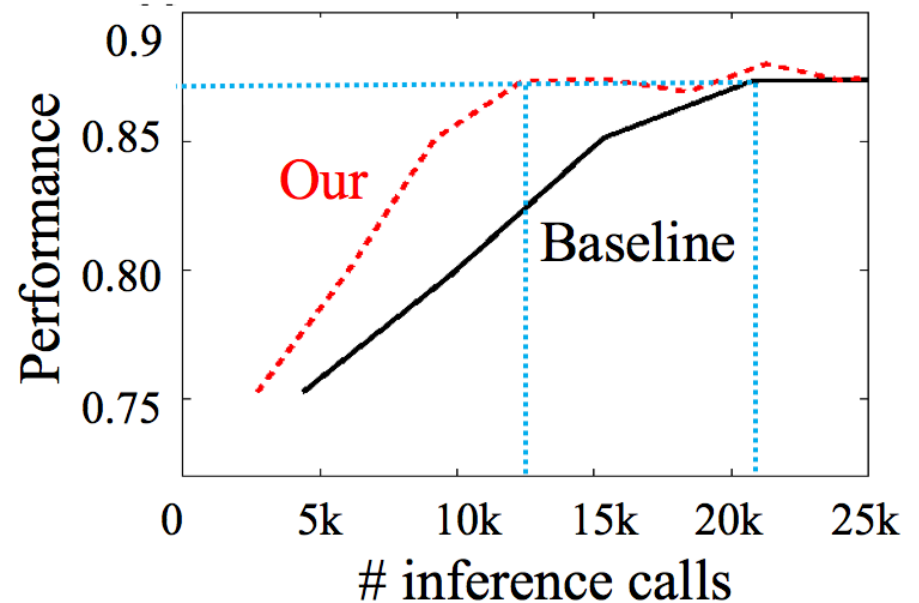
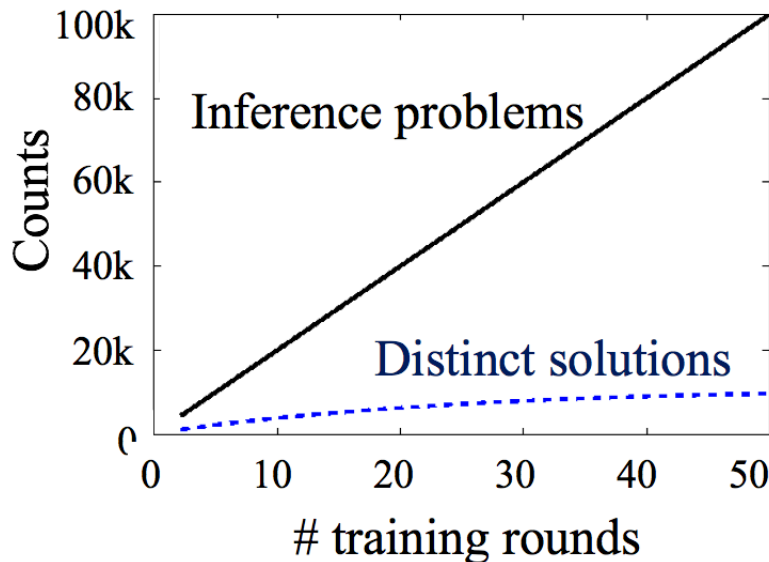
Idea 2: Amortized inference

- ❖ Idea: Exploit this redundancy by **caching old inference solutions** [AAAI 15]



Idea 2: Amortized inference

- ❖ Idea: Exploit this redundancy by **caching old inference solutions** [AAAI 15]





Amortized inference – key components

- A **general inference framework**
 - ... to represent inference problems
- A **condition**
 - ... to check if two problems have the same solution

```
If CONDITION (problem cache, new problem)
  then (no need to call the solver) 0.04 ms
    SOLUTION (new problem) = old solution
Else
  Call base solver and update cache
End 2 ms
```

Solution Methods

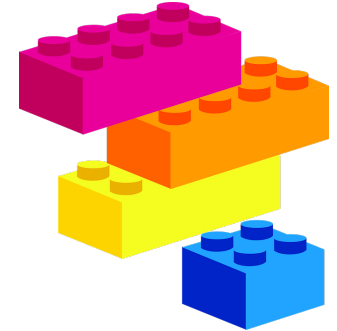
- ❖ Assume a graphical structure; optimize
 - ❖ Use within various structured predictions algorithms (e.g., CRF, Structured Perceptron, M3N, Structured SVM)
[Lafferty+ 01, Collins02, Taskar04]
 - ❖ See our AAIL16 tutorial (<https://goo.gl/TF7cGj>)
- ❖ Learning to search approaches
 - ❖ Assume the complex decision is incrementally constructed by a sequence of decisions
 - ❖ E.g., LASO, dagger, Searn, transition-based methods
 - ❖ See our NAACL15 tutorials (<http://hunch.net/~l2s>)

Libraries for Structured Predictions

- ❖ **Illinois-SL**: graph-based structured prediction
 - ❖ Support various algorithms; parallel \Rightarrow very fast
- ❖ **Vowpal-Wabbit**: credit assignment compiler
 - ❖ A general online learning library
 - ❖ Support search-based structured prediction

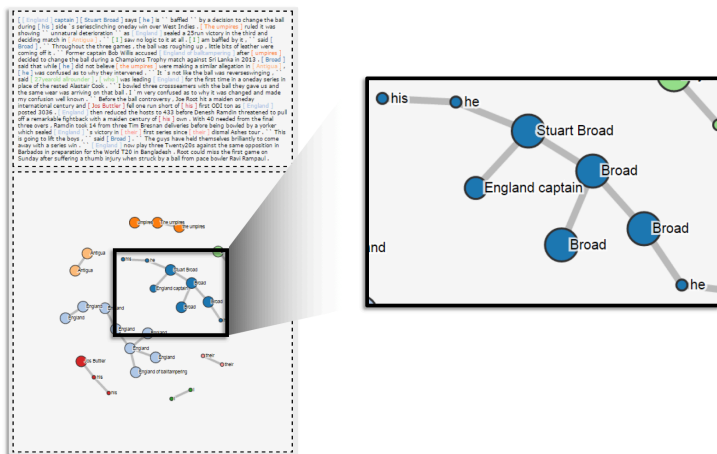
Learning to search approaches: Credit Assignment Compiler [NIPS16]

Sequential_RUN(*examples*)

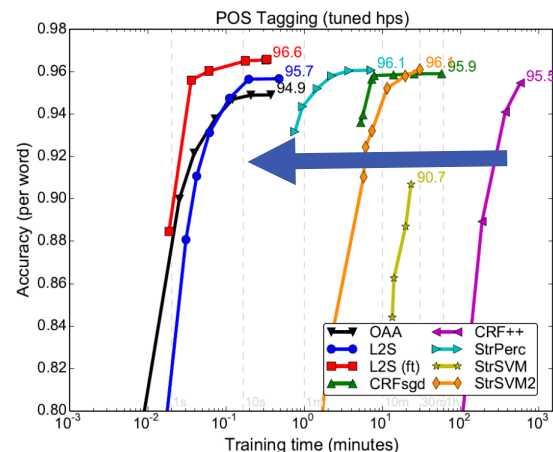


- 1: for $i = 1$ to $\text{len}(\text{examples})$ do
 - 2: $\text{prediction} \leftarrow \text{predict}(\text{examples}[i], \text{examples}[i].\text{label})$
 - 3: $\text{loss}(\text{prediction} \neq \text{examples}[i].\text{label})$
 - 4: end for
- ❖ Write the decoder, providing some side information for training
 - ❖ Library **translates** this piece of program with data to the update rules of model
 - ❖ Applied to dependency parsing, Name entity recognition, relation extraction, POS tagging...

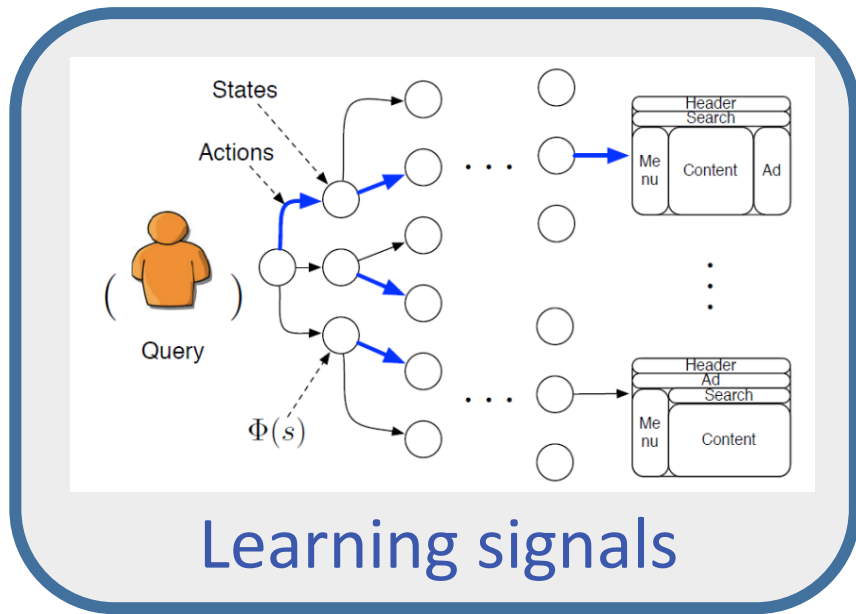
Outline



NLP Applications



Training/test speed



Learning signals

Query



activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Learning with Implicit/Partial Supervision

[ICML15, EMNLP16, AAI workshop17]

❖ Consider algebra word problem

Maria is now four times as old as Kate.
Four years ago, Maria was six times as
old as Kate. Find their ages now.

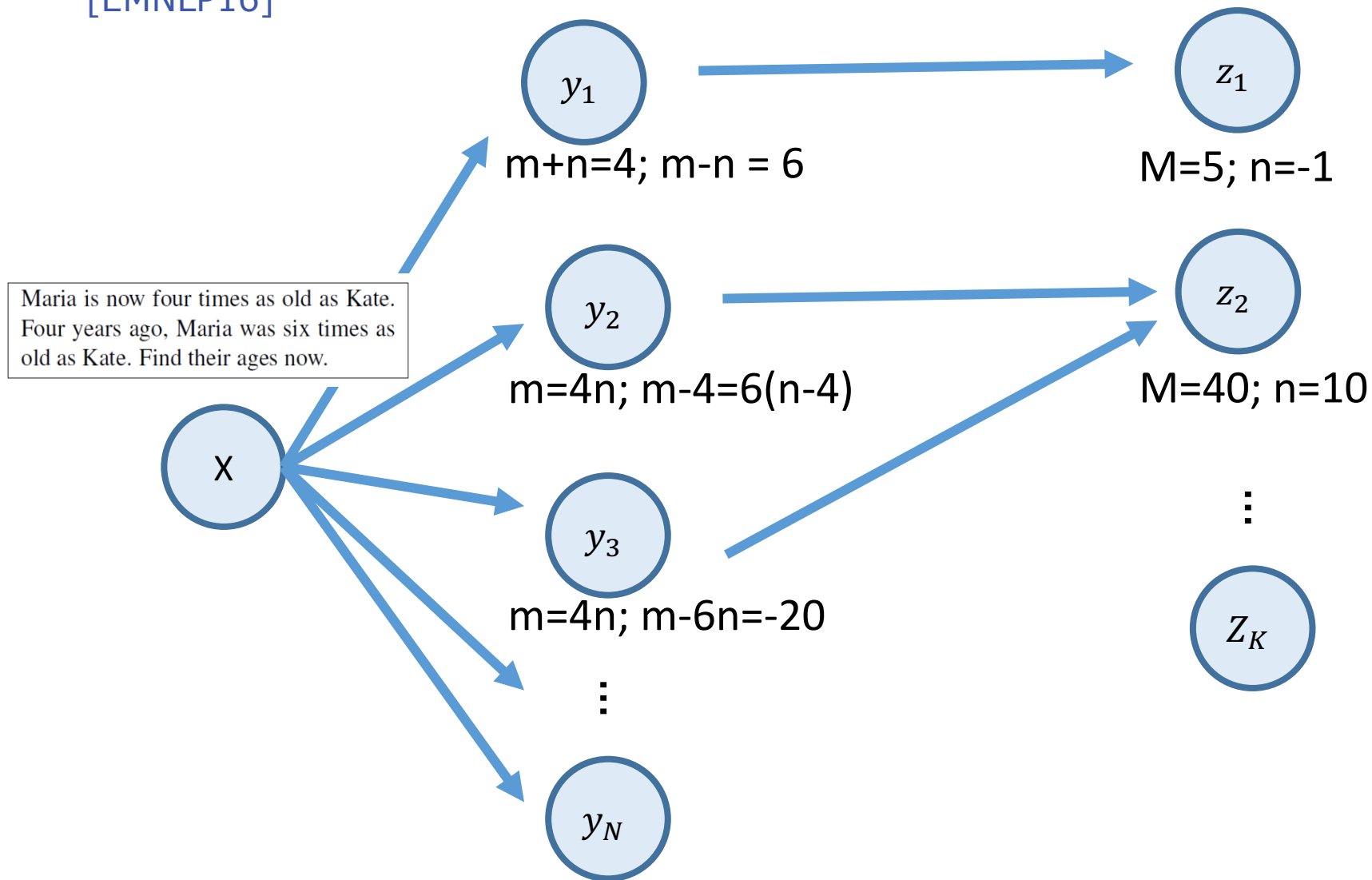
❖ Build semantic parser to translate question to an equation system

$$m = 4 \times n \text{ and } m - 4 = 6 \times (n - 4).$$

❖ Then answer can be derived: $m=40$, $n=10$

Learning with Implicit/Partial Supervision

[EMNLP16]

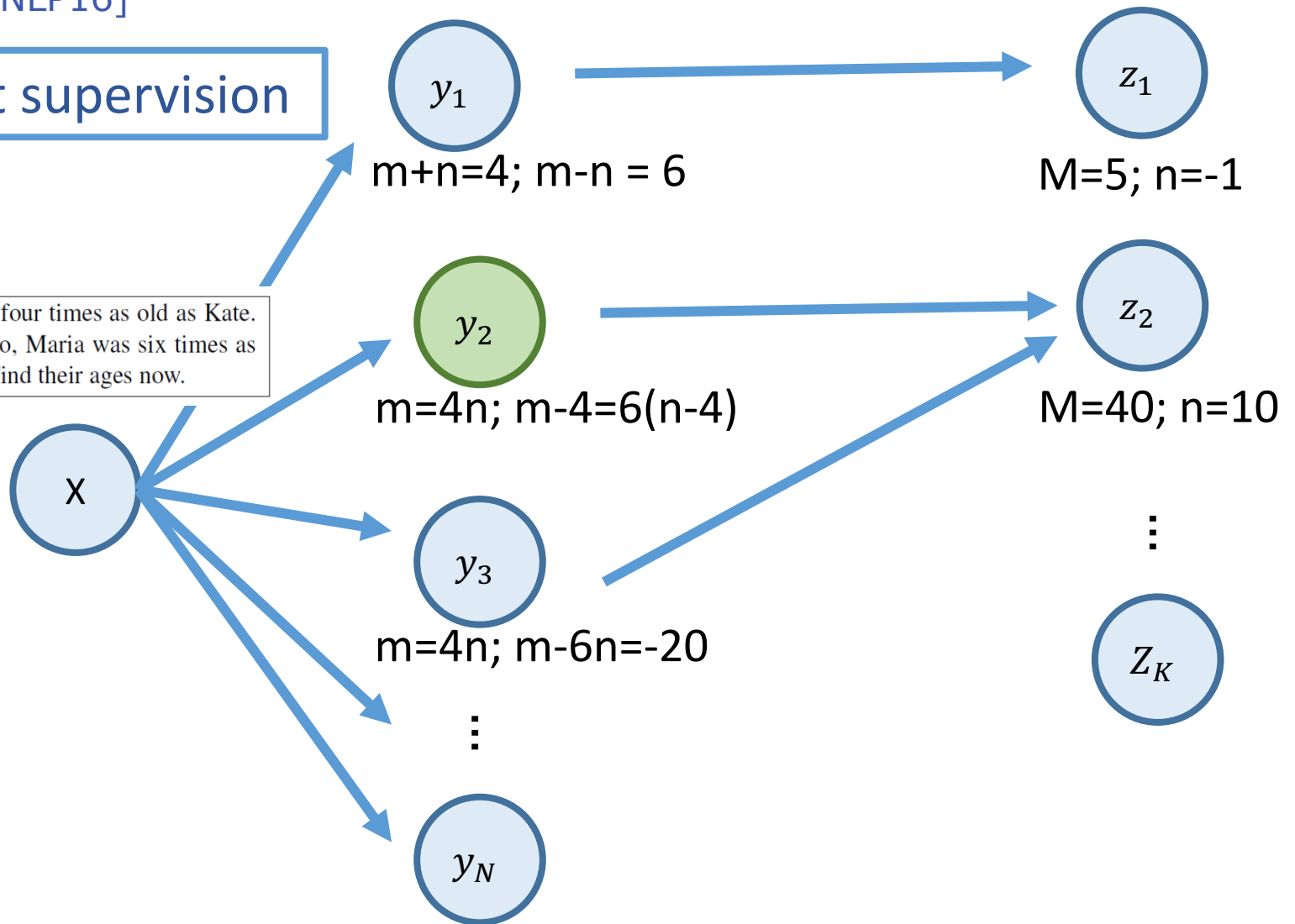


Learning with Implicit/Partial Supervision

[EMNLP16]

Explicit supervision

Maria is now four times as old as Kate.
Four years ago, Maria was six times as old as Kate. Find their ages now.

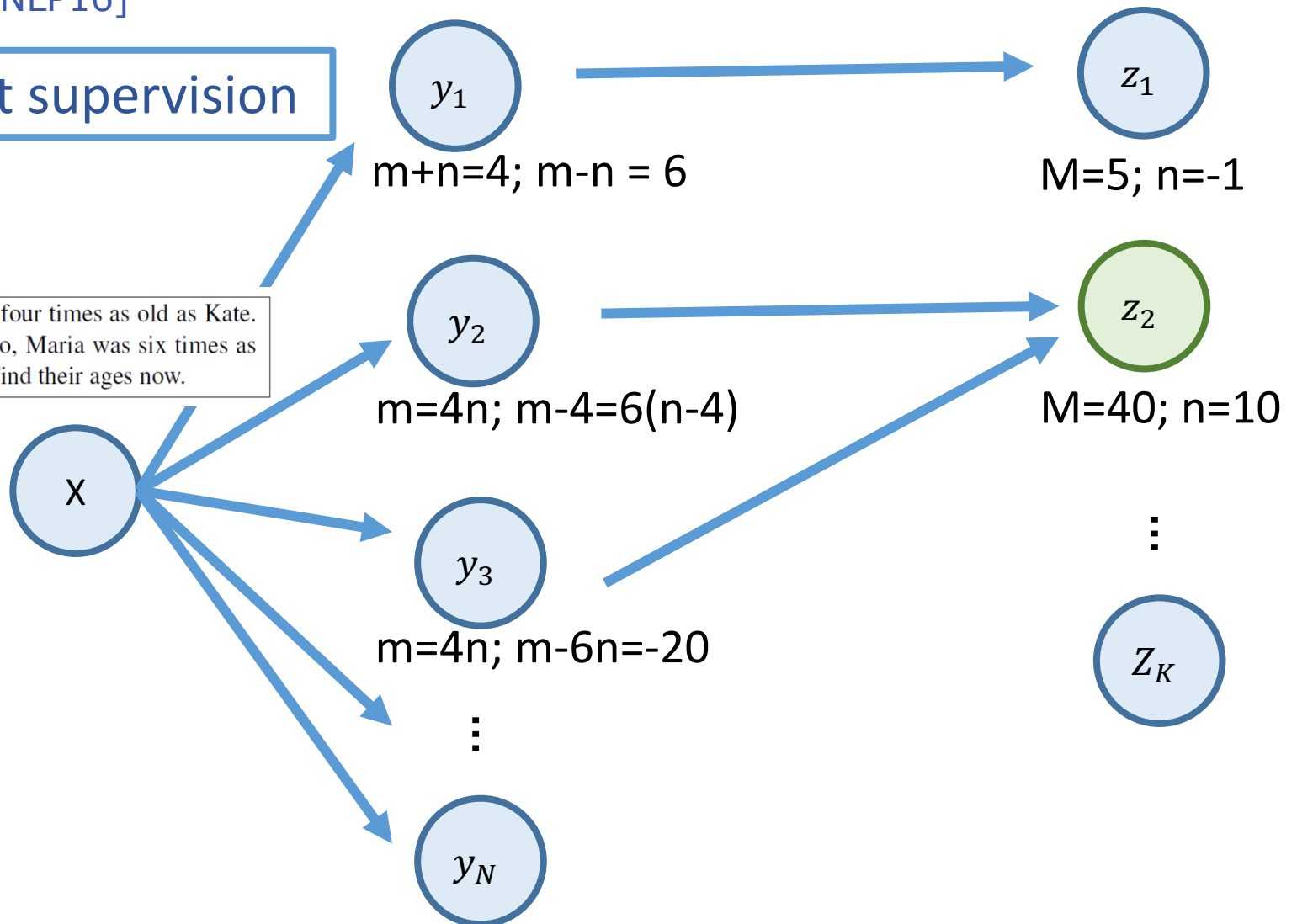


Learning with Implicit/Partial Supervision

[EMNLP16]

Implicit supervision

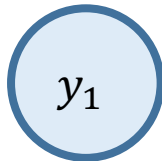
Maria is now four times as old as Kate.
Four years ago, Maria was six times as
old as Kate. Find their ages now.



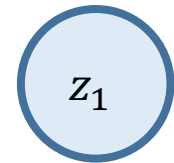
Learning with Implicit/Partial Supervision

[EMNLP16]

Implicit supervision

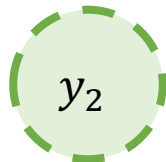


$$m+n=4; m-n=6$$

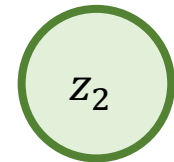


$$M=5; n=-1$$

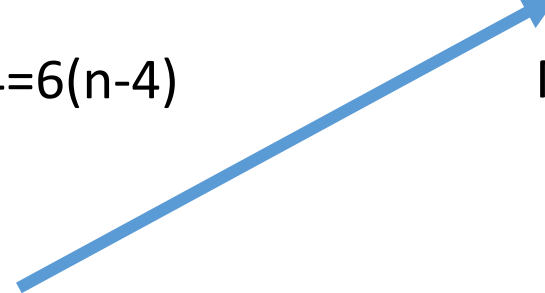
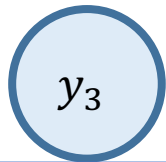
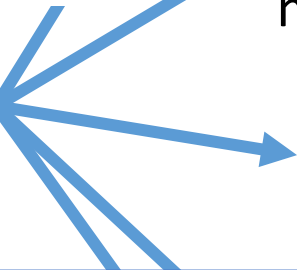
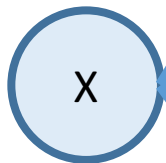
Maria is now four times as old as Kate. Four years ago, Maria was six times as old as Kate. Find their ages now.



$$m=4n; m-4=6(n-4)$$



$$M=40; n=10$$



⋮

Key idea: Use model to guide exploration on implicit supervision
⇒ Infer plausible equation systems

Our approach: Use the best equation system in the beam whose solution matches the implicit signal

Structured Contextual Bandit Setting

[ICML15, AAAI workshop17, IJCNLP 17]

w/ Akshay Krishnamurthy, Alekh Agarwal, Hal Daume; III, John Langford
w/ Kenneth Arnold, Adam Kalai

- ❖ Loss of only a single structured label can be observed

is my favorite restaurant in harvard square. i love italian food and this is one of the best places nearby for a variety of delicious italian dishes. i often get the same thing every time i go there i have the rigatoni with meat sauce and sometimes the veal. the lunch menu

is

very reasonably priced ,

was

great , the food is

has

something for everyone

Blue words preview the phrase that would be inserted with repeated taps

Prediction \neq Suggestion, especially in writing

- ❖ Baseline: an N-gram model doesn't work
- ❖ Accurate predictions may be poor suggestions
 - ❖ E.g., most frequent phrases in Yelp:
this place is great!; Love it, love it, love it!; I love this place
 - ❖ E.g., most frequent short email replies [Kannan+ 16]:
I love you; Thanks; sounds good
- ❖ Writers prefer suggestions occur less frequently but more creative/enthusiastic/informative:
 - ❖ e.g., this was truly a wonderful experience

Key ideas

1. Discriminative language model

- ❖ Allow to add features
 - ⇒ Tune the model to favor phrases with some properties

Key ideas

1. Discriminative language model

2. Counterfactual learning

❖ Cannot get full feedback

❖ Only candidates presented to users are annotated

❖ How can we estimate the model parameter unbiasedly?

is my favorite restaurant in harvard square. i love italian food and this is one of the best places nearby for a variety of delicious italian dishes. i often get the same thing every time i go there i have the rigatoni with meat sauce and sometimes the veal. the lunch menu

is	was	has
very reasonably priced ,	great , the food is	something for everyone

5%

3%

7%

is very delicious 2%

has a drink option 1%

is not worthwhile 6%

is good. 8%

...

...

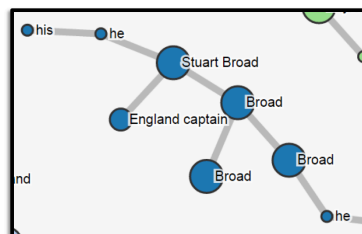
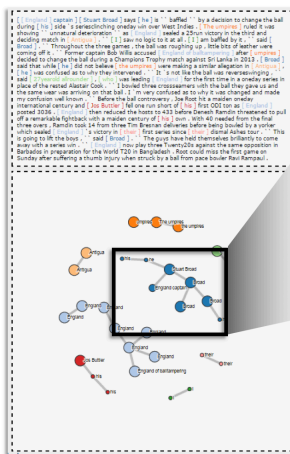
Counterfactual learning

- ❖ Collect behavior data using a reference model
(suggested phrase, reward, probability, context)
- ❖ Reference model: a tri-gram language model with Kneser-Ney smoothing & temperature parameter
- ❖ reward = # words accepted

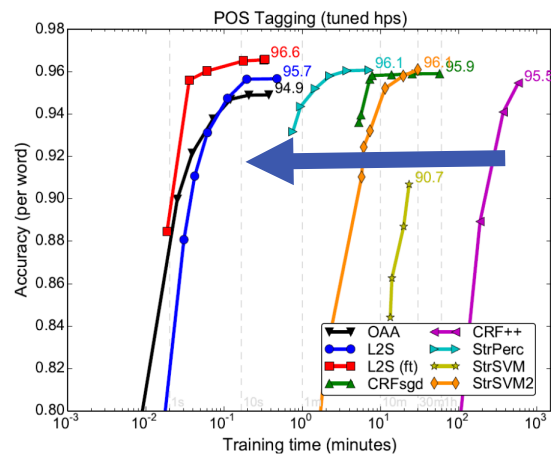
is my favorite restaurant in harvard square. i love italian food and this is one of the best places nearby for a variety of delicious italian dishes. i often get the same thing every time i go there i have the rigatoni with meat sauce and sometimes the veal. the lunch menu

is very reasonably priced ,	was great , the food is	has something for everyone
--------------------------------	----------------------------	-------------------------------

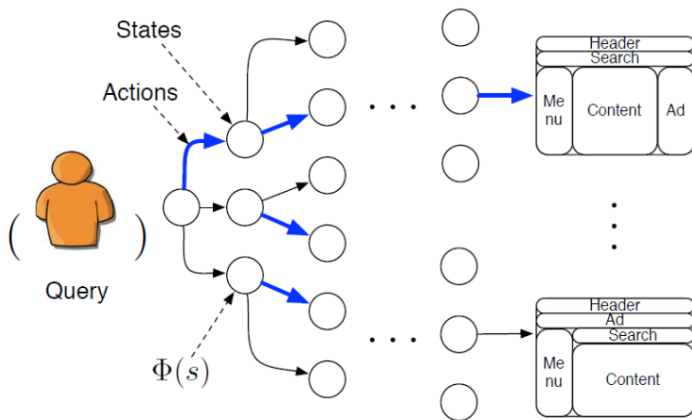
Outline



NLP Applications



Training/test speed



Learning signals

Kai-Wei Chang (kwchang.net/talks/sp.html)



activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Select photo



✗ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo [requirements](#).

Subject eyes closed

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to [contact us](#) about the photo, you must provide us with the reference number given above.

Please print this information for your records.

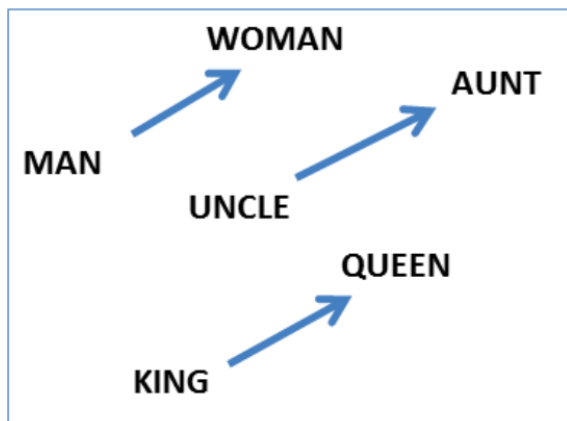


A screenshot of New Zealand man Richard Lee's passport photo rejection notice, supplied to Reuters December 7, 2016. Richard Lee/Handout via REUTERS

<https://www.reuters.com/article/us-newzealand-passport-error/new-zealand-passport-robot-tells-applicant-of-asian-descent-to-open-eyes-idUSKBN13WORLD>

Word Embeddings can be Dreadfully Sexist [nips16, **reported by NPR, MIT tech review**]

$$\diamond v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$$



he: ____	she: ____
uncle	aunt
lion	
surgeon	
architect	
beer	
professor	

We use Google w2v embedding trained from the news

Project occupations in gender direction

Extreme *he* occupations

- | | | |
|----------------|-------------------|----------------|
| 1. maestro | 2. skipper | 3. protege |
| 4. philosopher | 5. captain | 6. architect |
| 7. financier | 8. warrior | 9. broadcaster |
| 10. magician | 11. fighter pilot | 12. boss |

she

he

Extreme *she* occupations

- | | | |
|-----------------|-----------------------|------------------------|
| 1. homemaker | 2. nurse | 3. receptionist |
| 4. librarian | 5. socialite | 6. hairdresser |
| 7. nanny | 8. bookkeeper | 9. stylist |
| 10. housekeeper | 11. interior designer | 12. guidance counselor |

Natural Language Processing

Data

Machine Learning

Seeking Question



Where can I find some pre-trained word vectors for natural language processing/understanding?

 Answer

Request ▾

Follow 36

Comment 1

Share Downvote



4 Answers

**Volkan Cirik**, worked on word {vectors|embedding}

Written May 9, 2014

You can use Koc AI-Lab's repository [ai-ku/wvec](#) to induce word embeddings. That link should point you out pre-trained word embeddings for WSJ corpus. I already have pre-trained embeddings for RCV1 and Wikipedia corpora for different dimensions. They are not available for public yet, but I can provide word embeddings if you want.

Other word embeddings are :

Mikolov's Works : [word2vec](#) (the code is pretty fast)

Turian's Work : [Word representations for NLP](#)

Dhillon's Work : [Eigenwords](#)

Huang's Work : [Improving Word Representations Via Global Context And Multiple Word Prototypes](#)

16k Views · View Upvotes

Scholar

About 93 results (0.02 sec)

Articles

Case law

My library

Any time

Since 2016

Since 2015

Since 2012

Custom range...

Sort by relevance

Sort by date

 include patents include citations Create alert

Machine Learned **Resume**-Job Matching Solution

Y Lin, H Lei, PC Addo, X Li - arXiv preprint arXiv:1607.07657, 2016 - arxiv.org

... We use LDA to classify **resumes** into 32 and 64 topics respectively. ... each Chinese phrase as a word and each list of phrases as a sentence, after **word2vec** training, each ... In this paper, we have considered the **resume**-job matching problem and pro- posed a solution by using ...

Cite Save

[PDF] SKILL: A System for Skill Identification and Normalization.

[M Zhao](#), [F Javed](#), F Jacob, M McNair - AAAI, 2015 - pdfs.semanticscholar.org

... ThiS dictionary capacitateS 90% of noiSe exhibited in **reSUMe** SkillS SectionS. ... iS initiated firSt for the input queY ry (aka, Seed Skill phraSeS from **reSUMeS**) for proper ... implement and produce highly precise and relevant skills recognition system, we utilize **word2vec** (Mikolov et ...

Cited by 4 Related articles All 3 versions Cite Save More

Word2Vec vs DBnary ou comment (ré) concilier représentations distribuées et réseaux lexico-sémantiques? Le cas de l'évaluation en traduction automatique

[C Servan](#), [Z Elloumi](#), H Blanchon, [L Besacier](#) - TALN 2016, 2016 - hal.archives-ouvertes.fr

... Page 2. **Word2Vec** vs DBnary ou comment (ré)concilier représentations ... **RÉSUMÉ** Cet article présente une approche associant réseaux lexico-sémantiques et représentations distribuées de mots appliquée à l'évaluation de la traduction automatique. ...

Cite Save

Macau: Large-scale skill sense disambiguation in the online recruitment domain

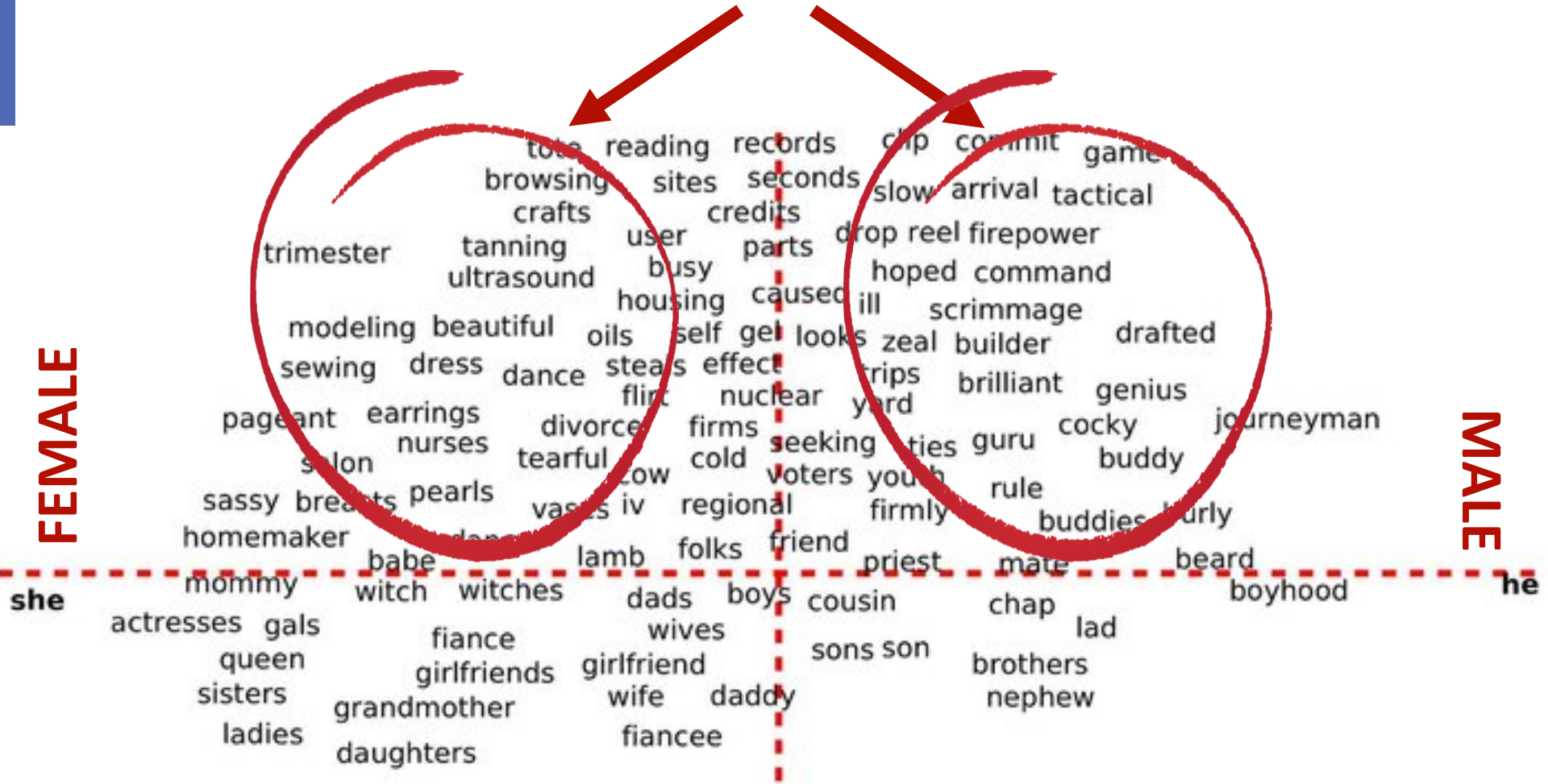
[Q Luo](#), [M Zhao](#), [F Javed](#), F Jacob - Big Data (Big Data), 2015 ..., 2015 - ieeexplore.ieee.org

... Contexts are extracted from either skill section(s) of **resumes** or requirement section(s) of job postings. We used a popular tool **word2vec** [12] with parameter

SEXIST

FEMALE

MALE



DEFINITIONAL

(related [Schmidt '15])

SEXIST

FEMALE

MALE

tote
browsing
tanning
scrimmage
dress
sewing
brilliant
nurse
cocky
genius
homemaker

she mommy witch witches dads boys cousin chap lad boyhood he
actresses gals fiance girlfriends wife daddy sons son brothers nephew
queen sisters ladies grandmother daughters fiancée

DEFINITIONAL

(related [Schmidt '15])

Debiasing algorithm (Soft version)

Find a linear transformation T of the gender-neutral words to reduce the gender component while not moving the words too much.

W = matrix of all word vectors.

N = matrix of neutral word vectors.

$$\min_T \left\| \underbrace{(TW)^T(TW) - W^T W}_{} \right\|_F^2 + \lambda \left\| \underbrace{(TN)^T(TB)}_{} \right\|_F^2$$

don't move too
much

minimize gender
component

Human Bias in Structured Prediction Models

[EMNLP 17*] w/ [Jieyu Zhao](#), Tianlu Wang, Mark Yatskar, Vicente Ordonez

What's the agent for this image?



Cooking	
Role	Object
agent	?
food	vegetable
container	bowl
tool	knife
place	kitchen

An example from a vSRL (visual Semantic Role Labeling) system

*Best Long Paper Award at EMNLP 17

Dataset Gender Bias

33%



Male

66%



Female

Model Bias After Training

16%

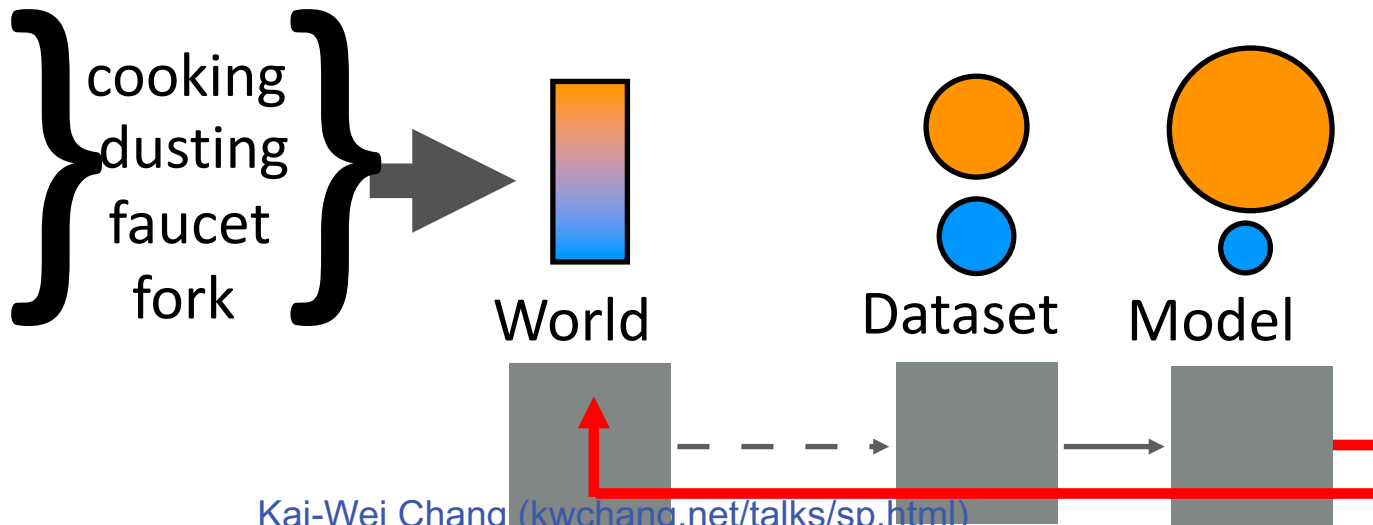
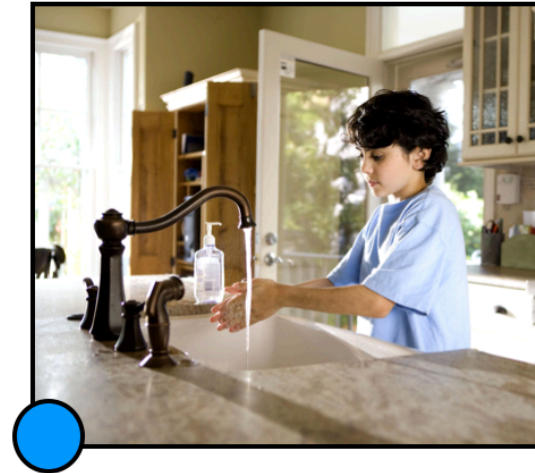
84%



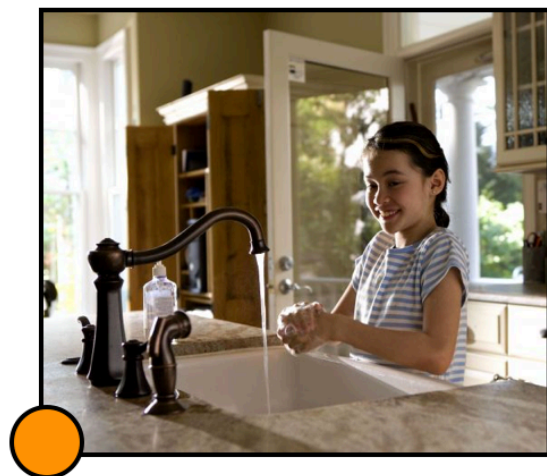
Male

Female

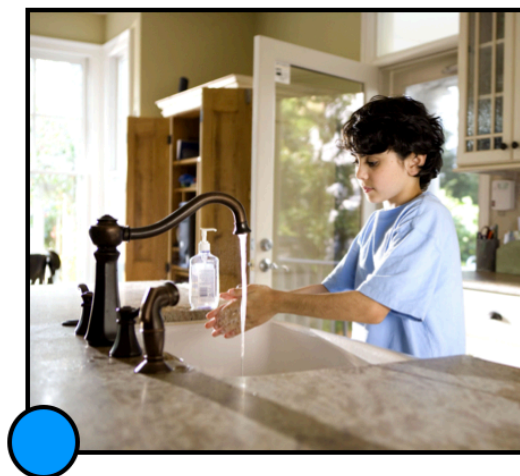
Algorithmic Bias in Grounded Setting



Algorithmic Bias in Grounded Setting

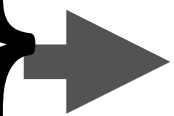


woman cooking

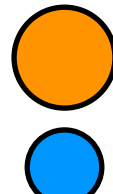


man fixing faucet

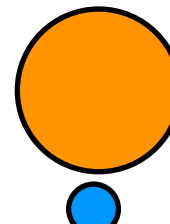
} cooking
} dusting
} faucet
} fork



World



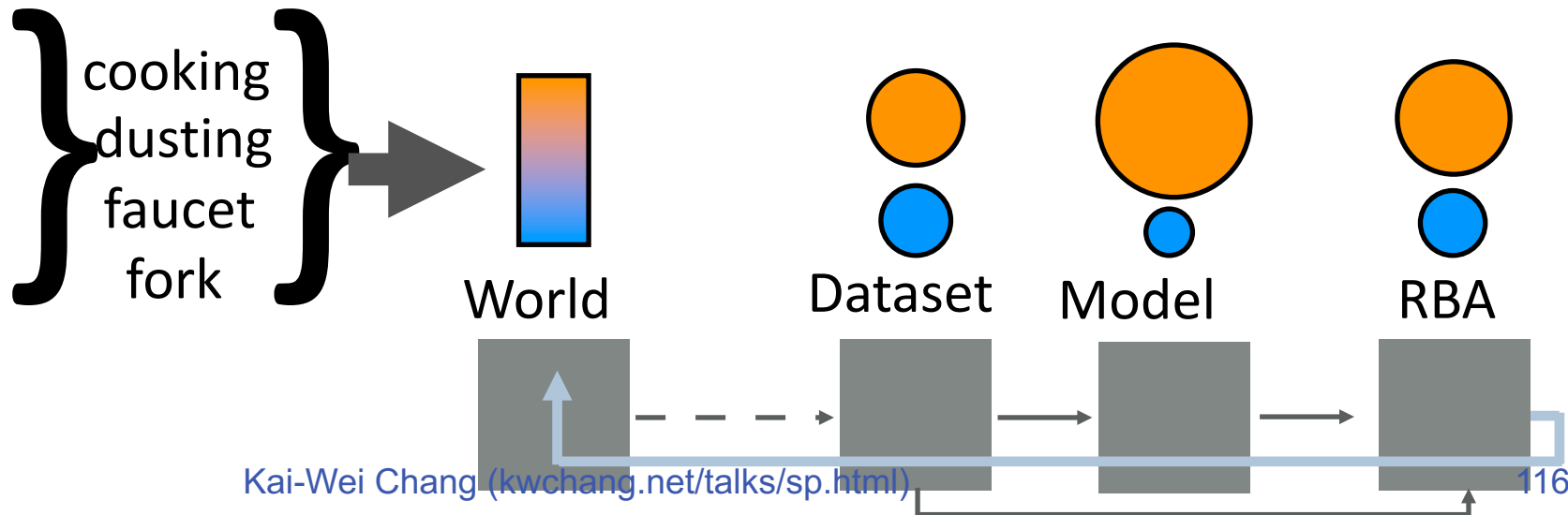
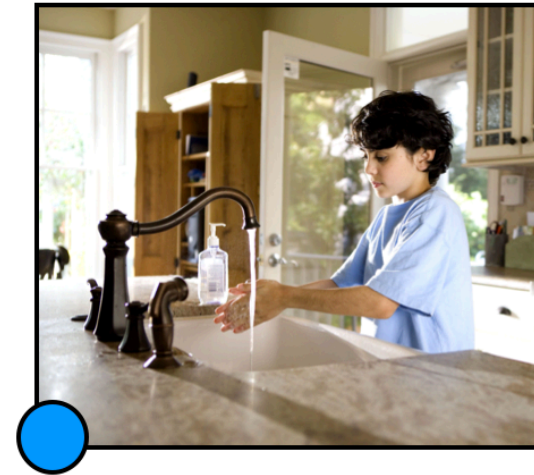
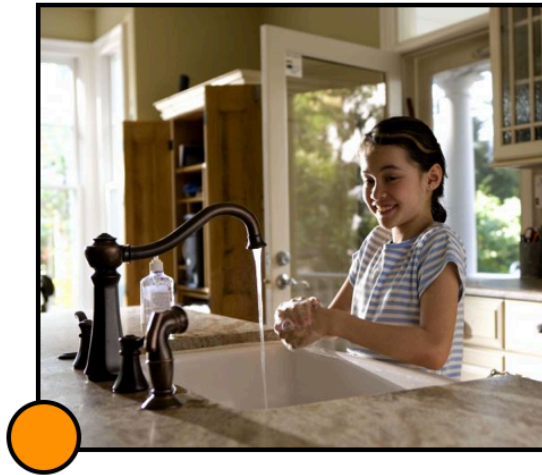
Dataset



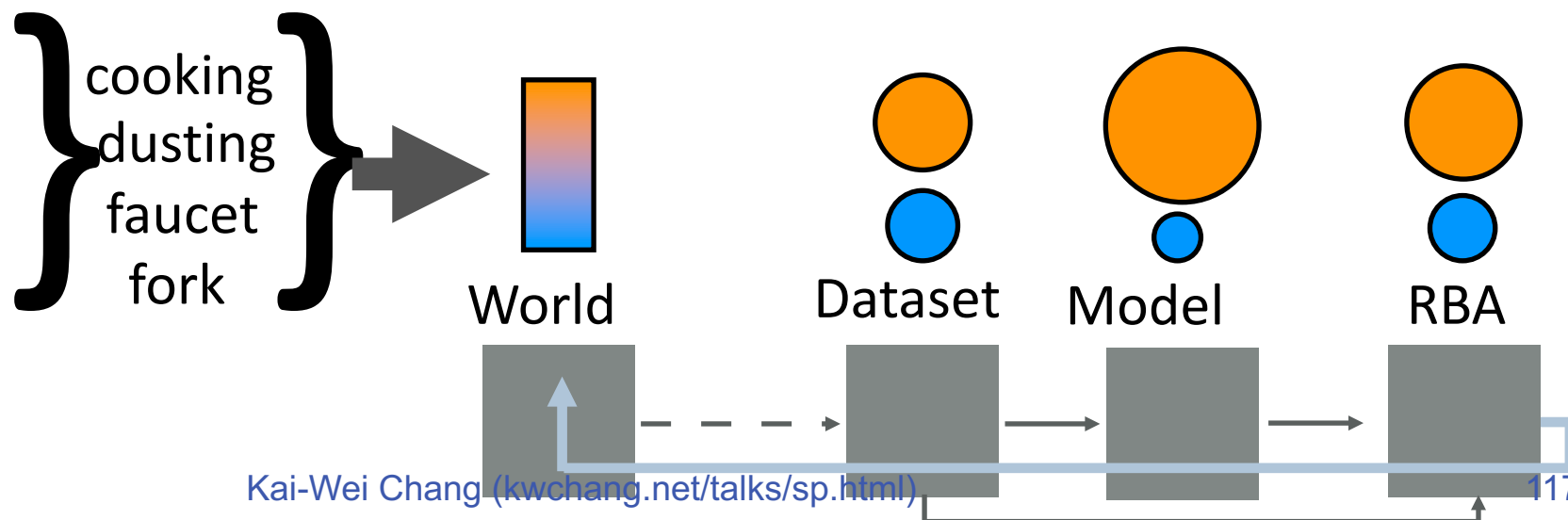
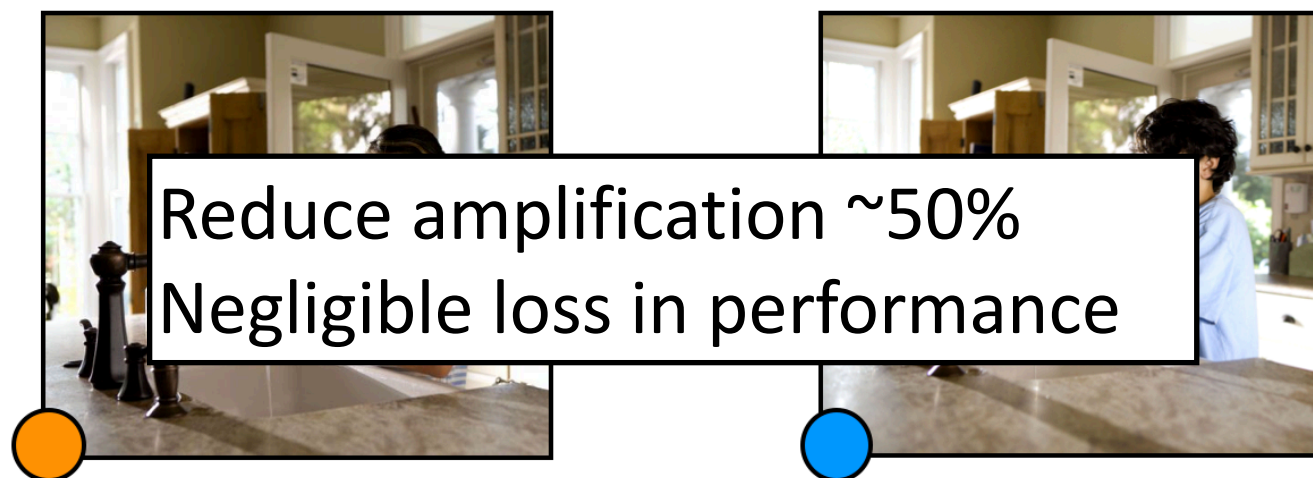
Model



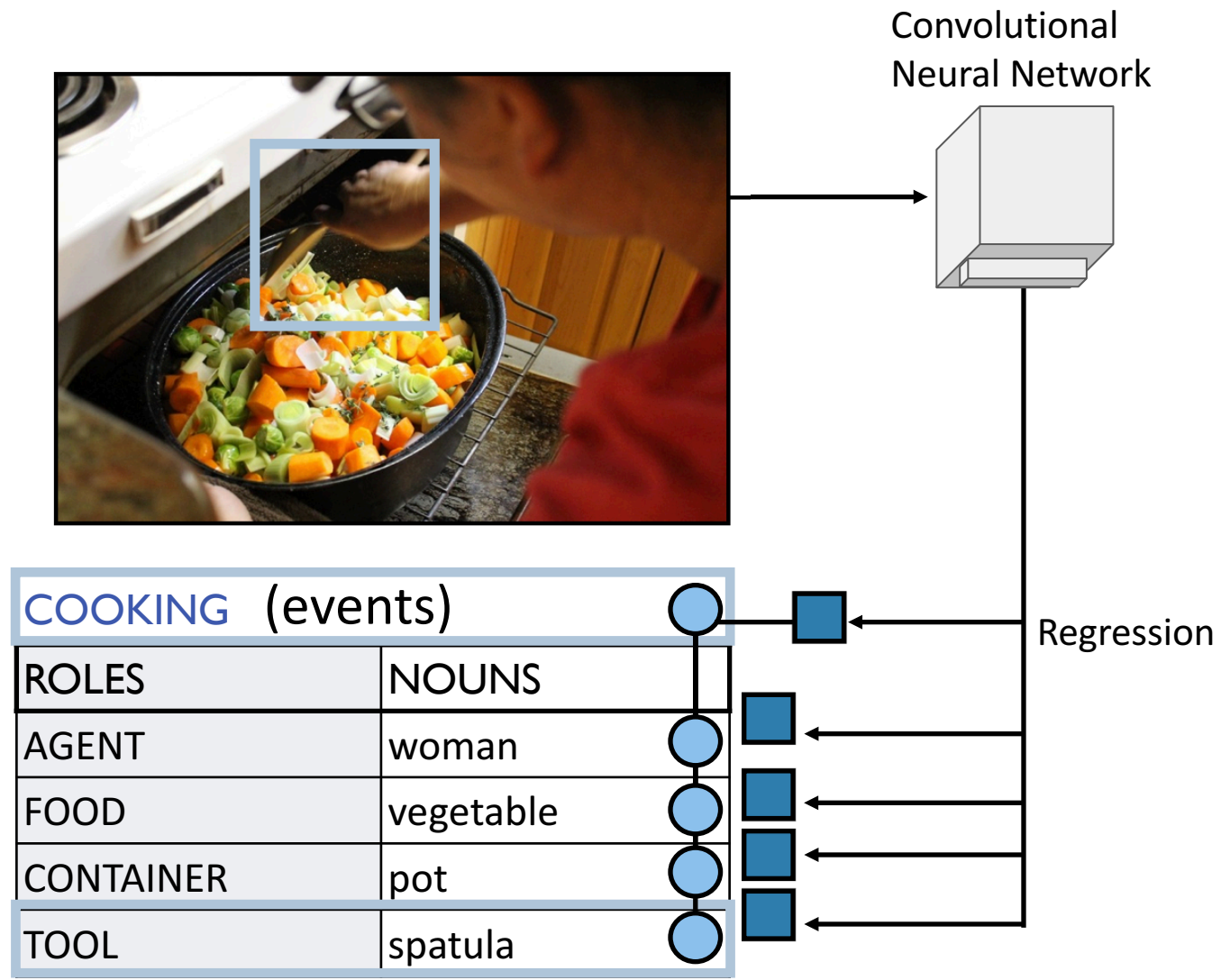
Algorithmic Bias in Grounded Setting



Algorithmic Bias in Grounded Setting



imSitu Visual Semantic Role Labeling (vSRL)

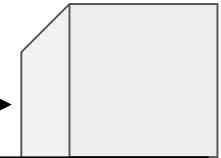


Kai-Wei Chang (kwchang.net/talks/sp.html)

imSitu Visual Semantic Role Labeling (vSRL)

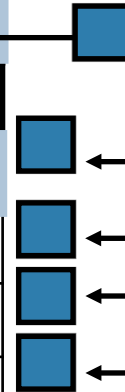


Convolutional
Neural Network



Need to model correlation between variables
Model can use that machinery to amplify gender bias

COOKING (events)	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula



Regression

Conditional Random Field

COCO Multi-Label Classification (MLC)



a woman is smiling in a kitchen near a pizza on a stove

COCO
Objects



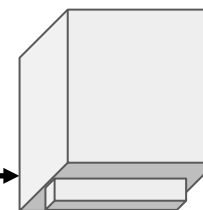
WOMAN	(objects)
PIZZA	yes
ZEBRA	no
FRIDGE	yes
CAR	no
...	...

← Caption Inferred
Label

COCO Multi-Label Classification (MLC)



Convolutional
Neural Network



WOMAN		●	■
PIZZA	yes	●	■
ZEBRA	no	●	■
FRIDGE	yes	●	■
CAR	no	●	■
...	...	●	■

Regression

Conditional Random Field

Defining Dataset Bias (events)

Training Gender Ratio (◆ verb)

Training Set

- ◆ cooking
- woman
- man



◆ COOKING	
ROLES	NOUNS
● AGENT	woman
FOOD	stir-fry



◆ COOKING	
ROLES	NOUNS
● AGENT	man
FOOD	noodle

$$\frac{\#(\text{◆ cooking}, \text{● man})}{\#(\text{◆ cooking}, \text{● man}) + \#(\text{◆ cooking}, \text{● woman})} = 1/3$$

Defining Dataset Bias (objects)

Training Gender Ratio (▲ noun)

Training Set

- ▲ snowboard
- woman
- man



● man	MAN	
▲ snowboard	snowboard	yes
	refrigerator	no
	bowl	no



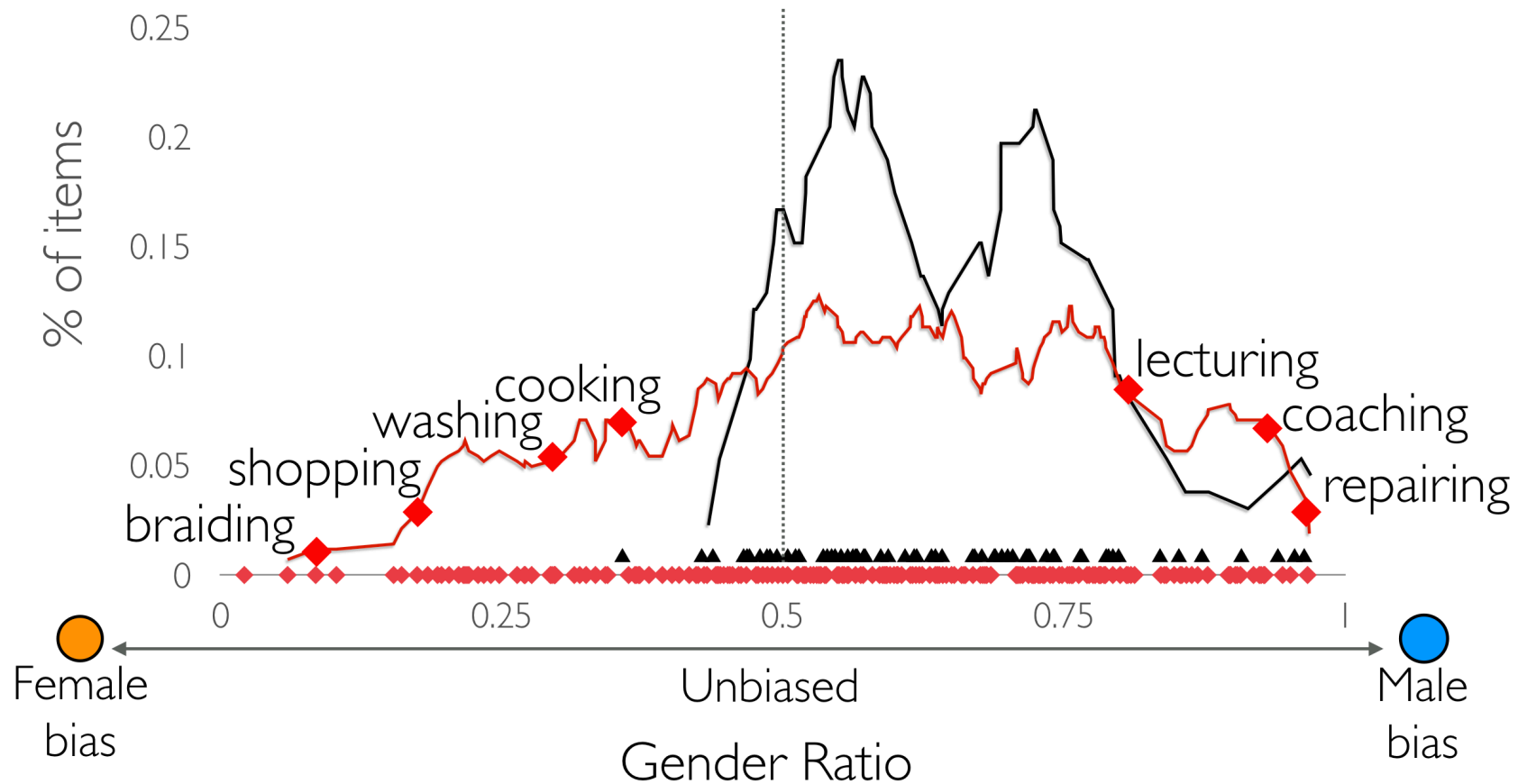
● woman	WOMAN	
▲ snowboard	snowboard	yes
	refrigerator	no
	bowl	no

$$\frac{\#(\blacktriangle \text{ snowboard}, \bullet \text{ man})}{\#(\blacktriangle \text{ snowboard}, \bullet \text{ man}) + \#(\blacktriangle \text{ snowboard}, \bullet \text{ woman})} = 2/3$$

Gender Dataset Bias

◆ imSitu Verb

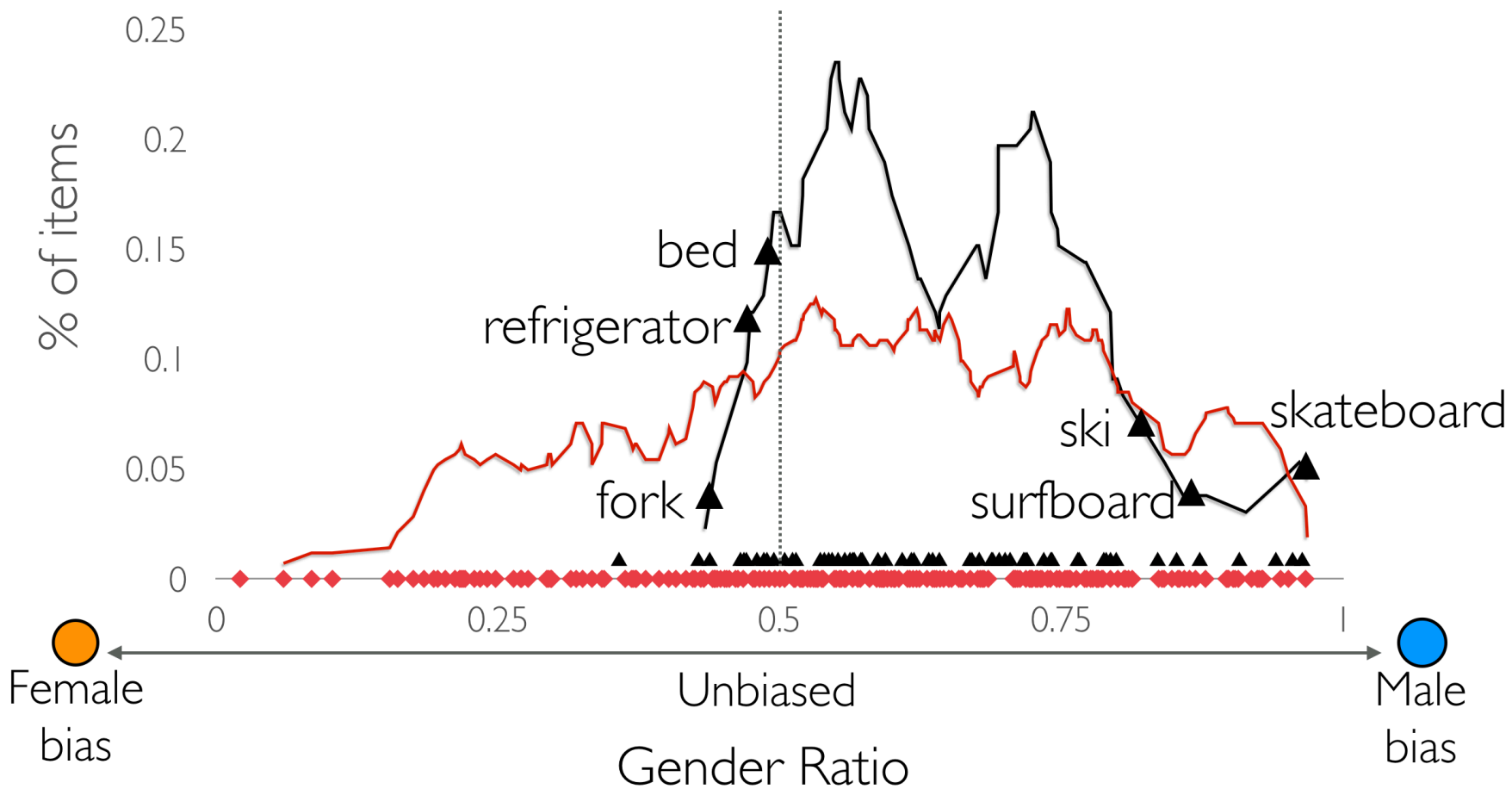
▲ COCO Noun



Gender Dataset Bias

◆ imSitu Verb

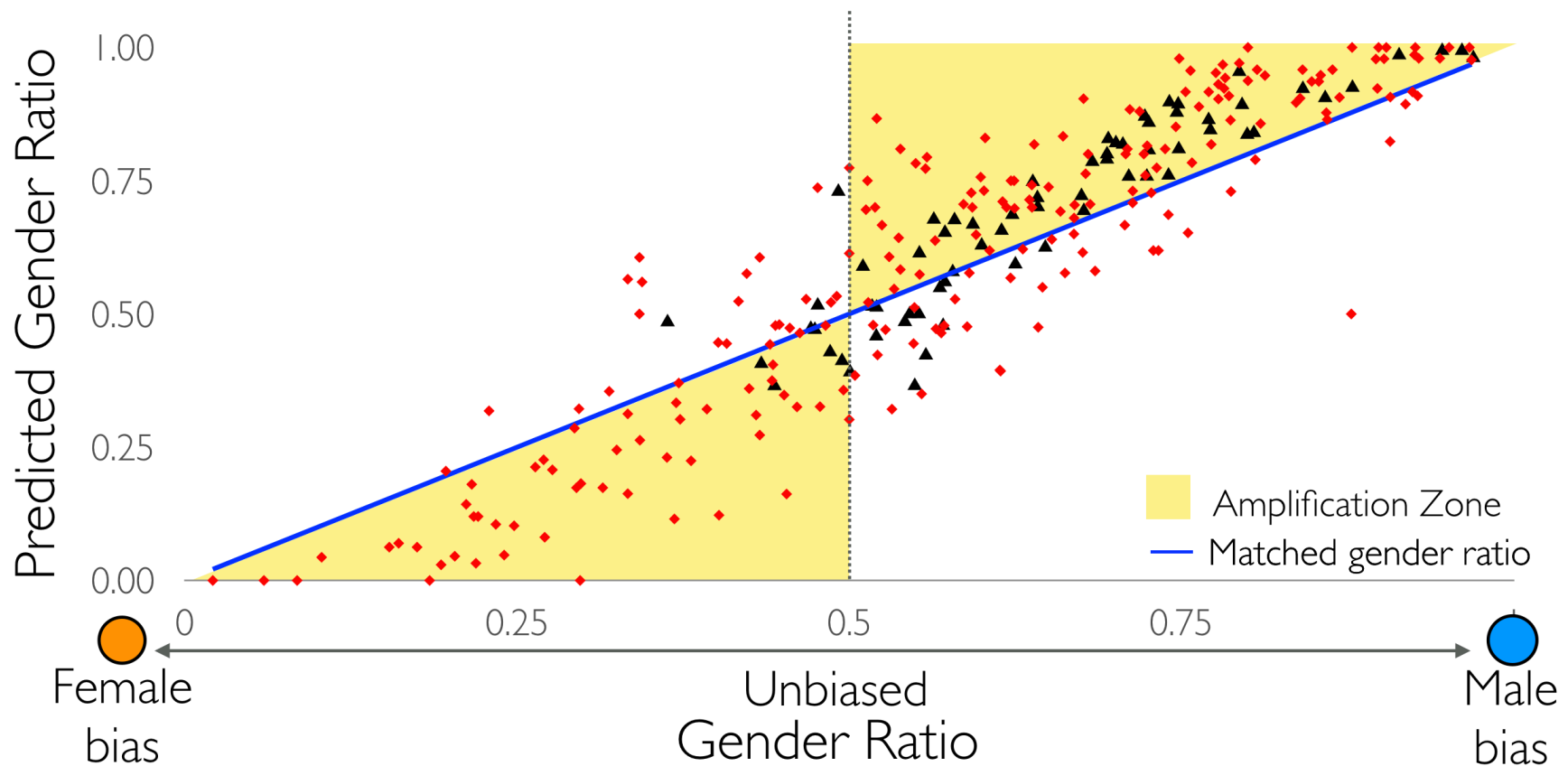
▲ COCO Noun



Model Bias Amplification

◆ imSitu Verb

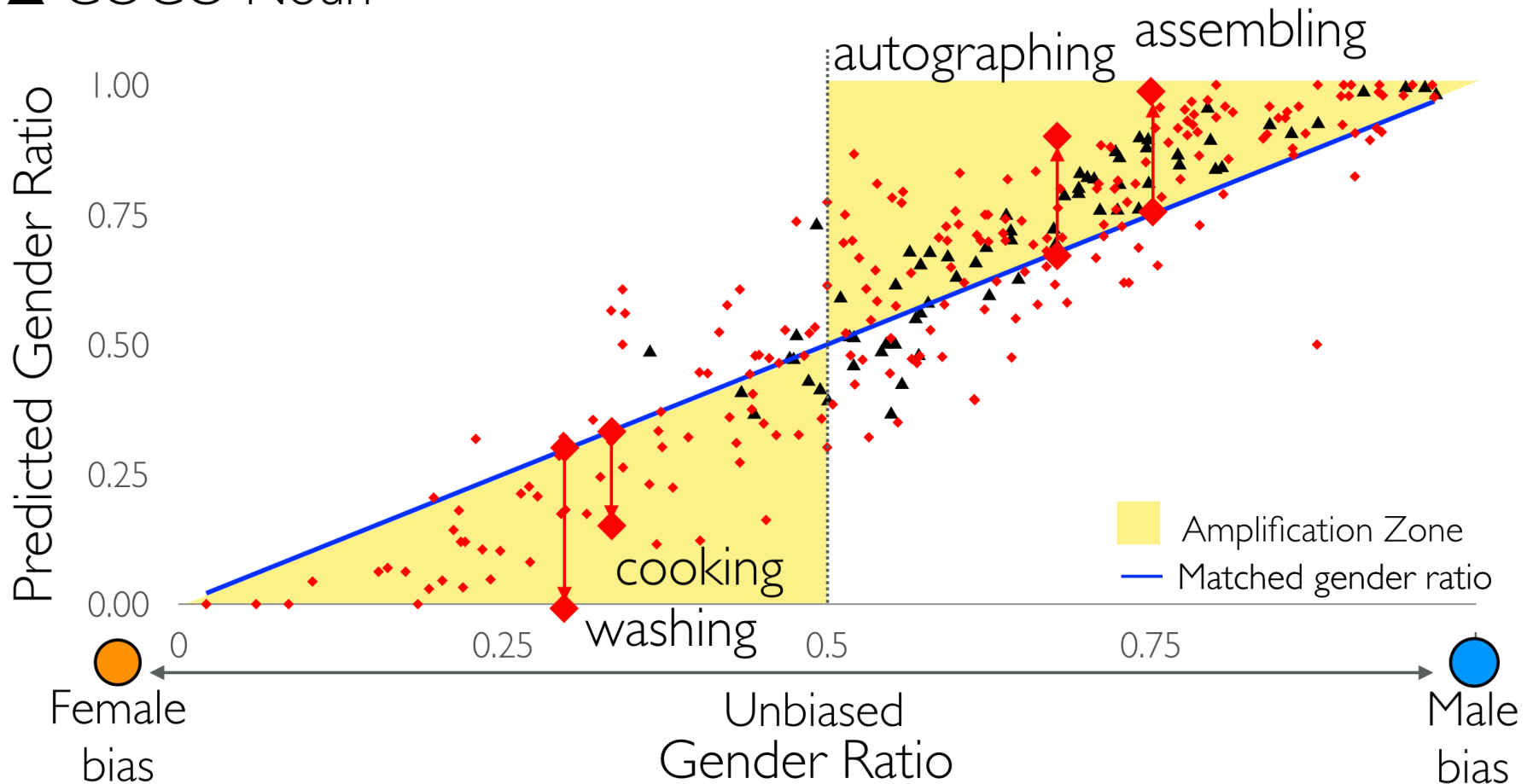
▲ COCO Noun



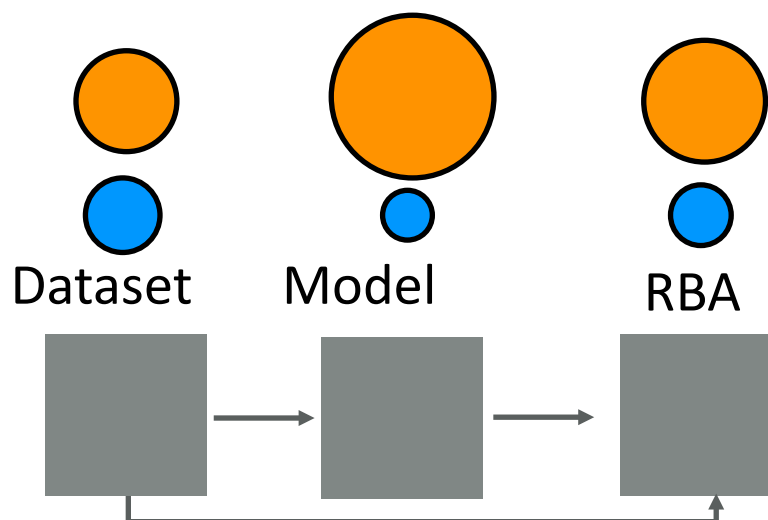
Model Bias Amplification

◆ imSitu Verb

▲ COCO Noun



Reducing Bias Amplification (RBA)



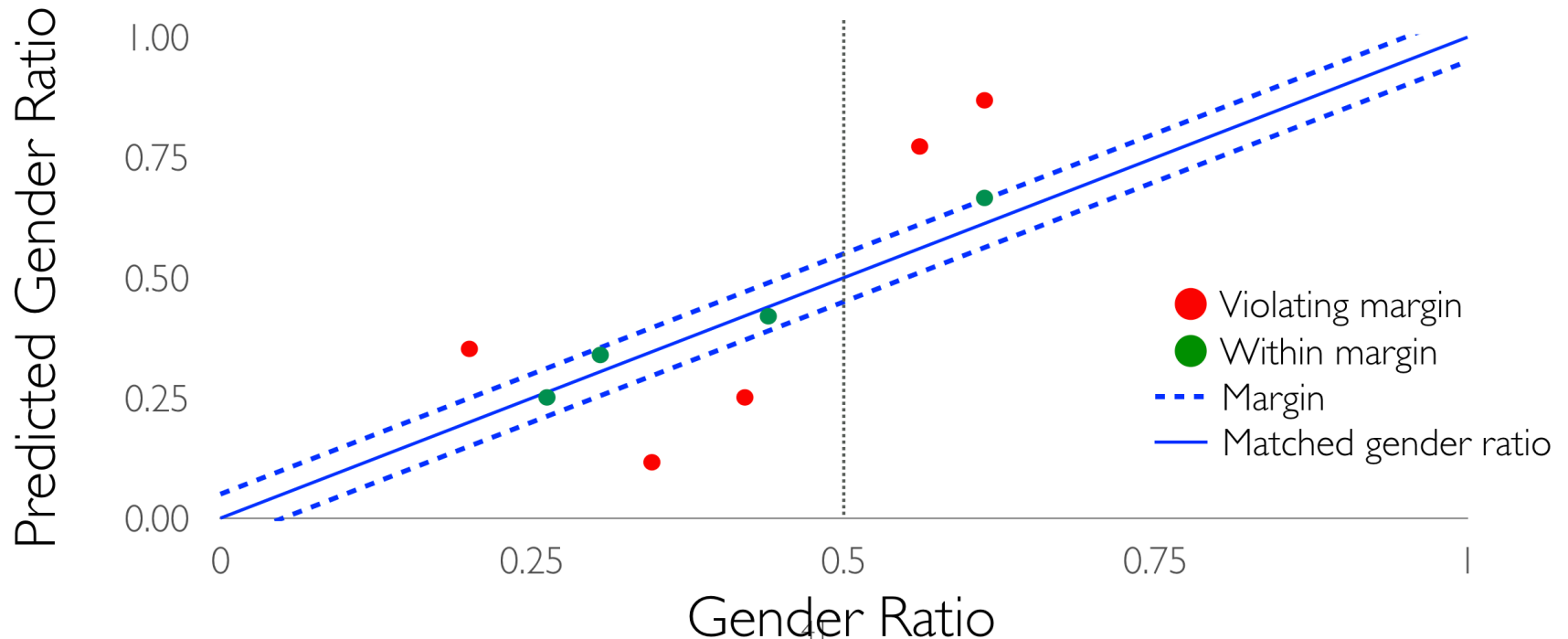
- ❖ Corpus level constraints on model output (ILP)
- ❖ Doesn't require model retraining
- ❖ Reuse model inference through Lagrangian relaxation
- ❖ Can be applied to any structured model

Reducing Bias Amplification (RBA)

Integer Linear Program

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points } \left| \text{Training Ratio} - \text{Predicted Ratio} \right|_{f(y_1 \dots y_n)} \leq \text{margin}$$

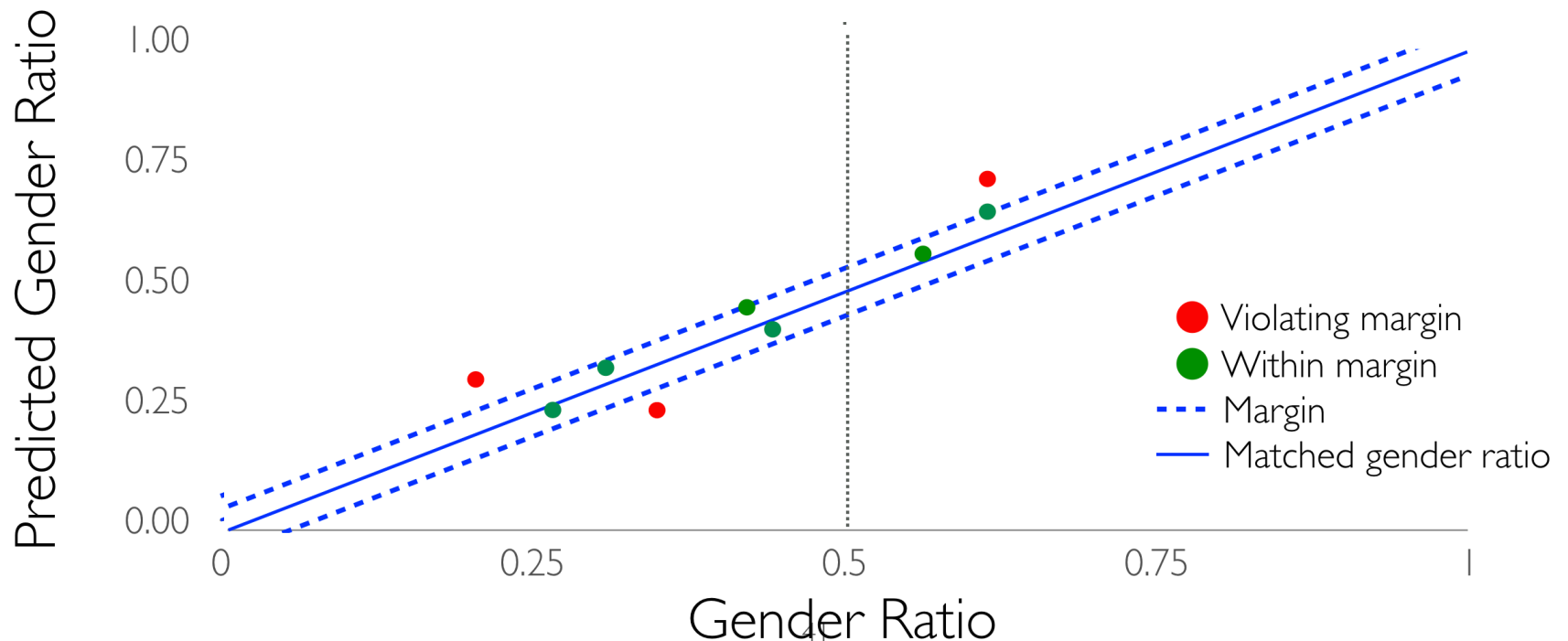


Reducing Bias Amplification (RBA)

Integer Linear Program

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points } \left| \text{Training Ratio} - \text{Predicted Ratio} \right|_{f(y_1 \dots y_n)} \leq \text{margin}$$

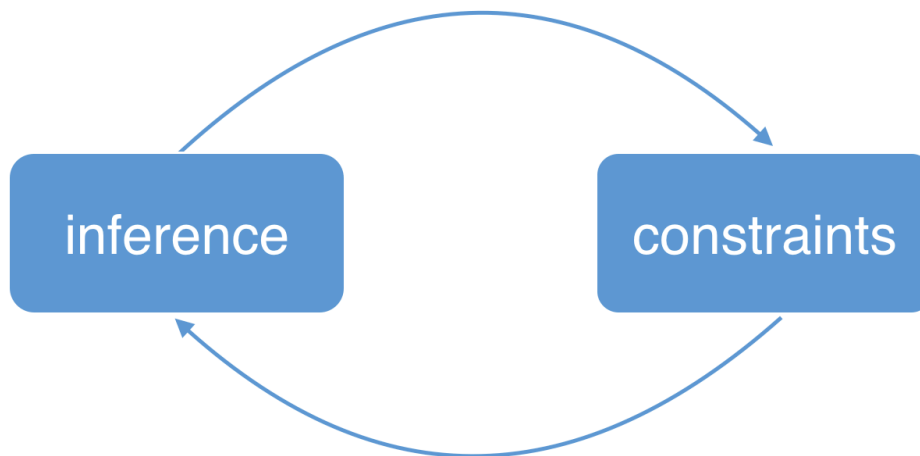


Reducing Bias Amplification (RBA)

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

\forall points $\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$
 $f(y_1 \dots y_n)$

Lagrangian Relaxation



Reducing Bias Amplification (RBA)

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points} \quad \left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$$

$$\max_{\{y^i\} \in \{Y^i\}} \sum_i f_{\theta}(y^i, i), \quad \text{s.t.} \quad A \sum_i y^i - b \leq 0$$

Lagrangian : $\sum_i f_{\theta}(y^i) - \sum_{j=1}^l \lambda_j (A_j \sum_i y^i - b_j) \quad \lambda_j \geq 0$

Lagrangian Relaxation



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Lagrangian Relaxation



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin} \quad (1/2)$$

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Lagrangian Relaxation



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

$$\lambda^{(t)} = \max \left(0, \lambda^{(t-1)} + \sum_i \eta (A y^{i,(t)} - b) \right)$$

Lagrangian Relaxation



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Lagrangian Relaxation



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake



COOKING	
ROLES	NOUNS
AGENT	man
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| \leq \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

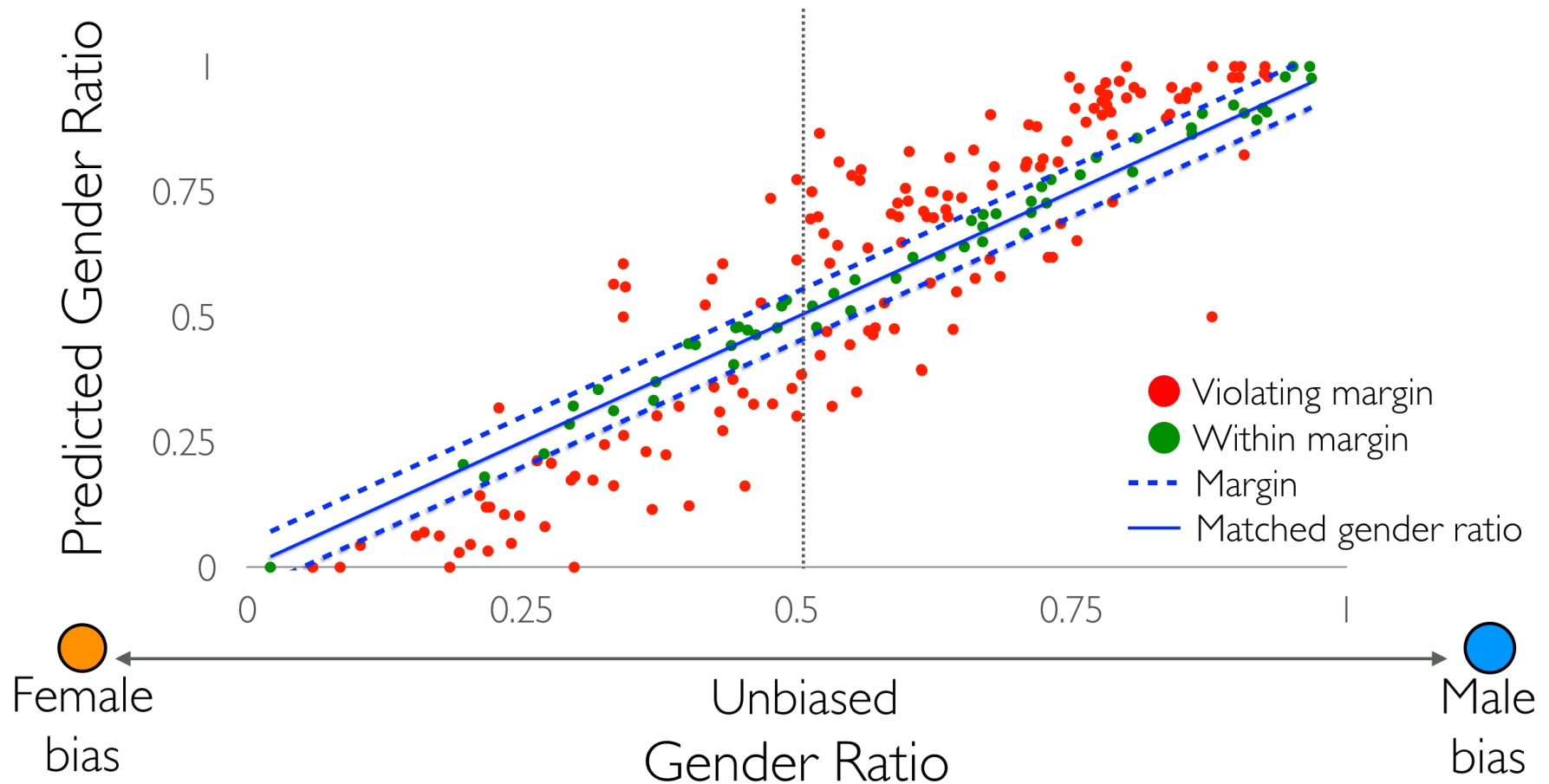
Gender Bias De-amplification in imSitu

imSitu Verb

Violation: 72.6%

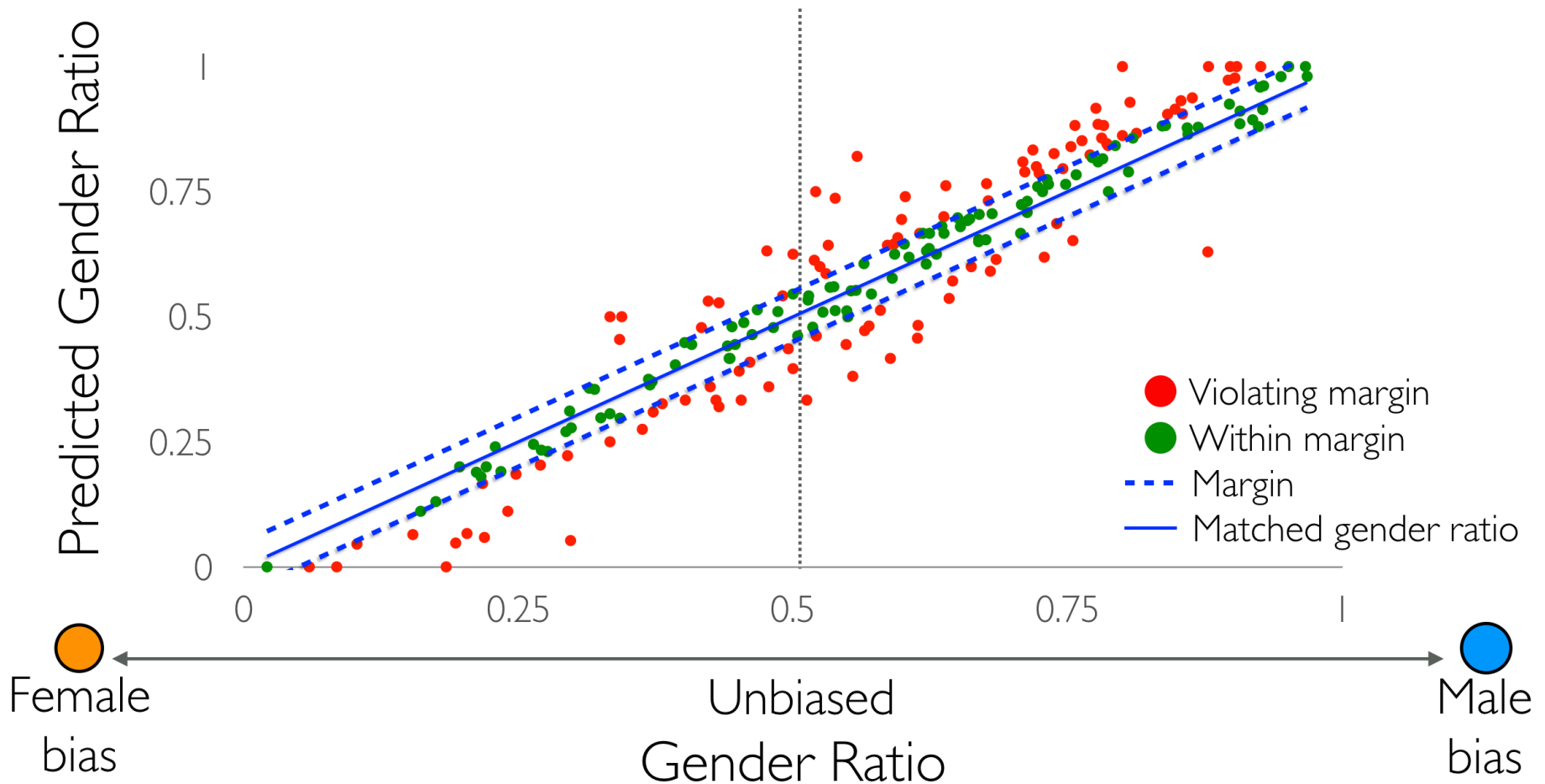
.050 |bias↑|

24.07 acc.



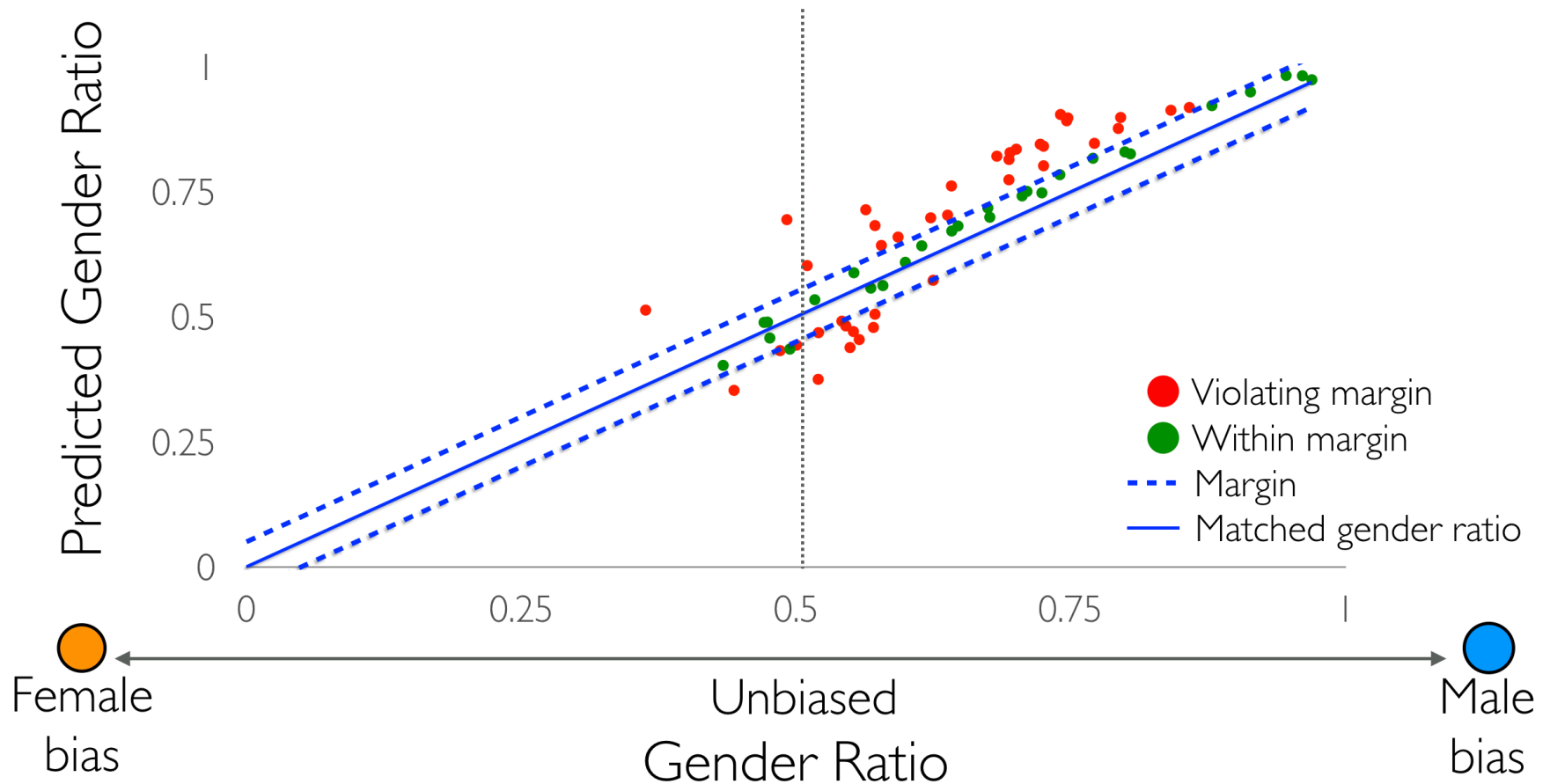
Gender Bias De-amplification in imSitu

imSitu Verb	Violation: 72.6%	.050 bias↑	24.07 acc.
w/ RBA	Violation: 50.5%	.024 bias↑	23.97 acc.



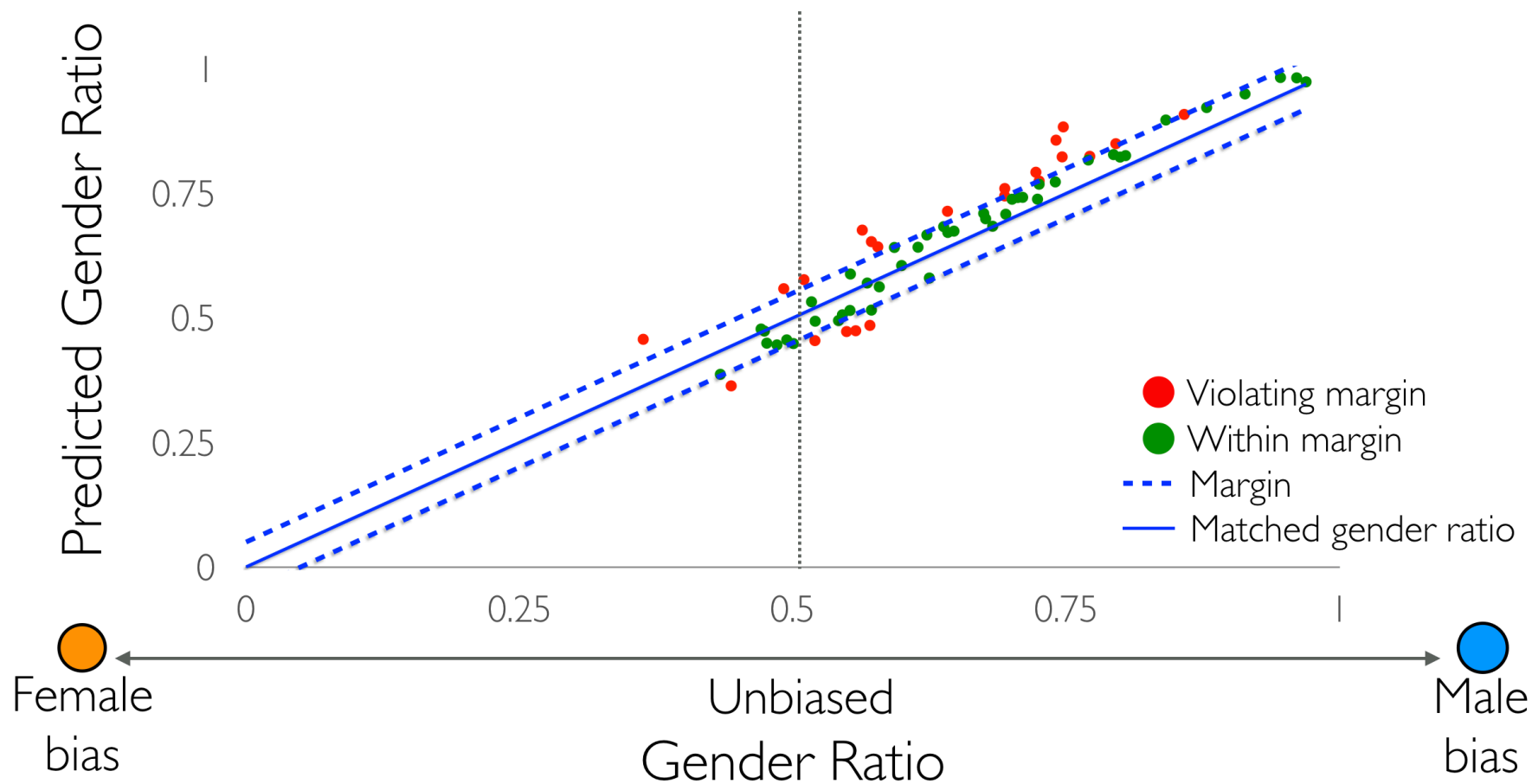
Gender Bias De-amplification in COCO

COCO Noun Violation: 60.6% .032 |bias↑| 45.27 mAP

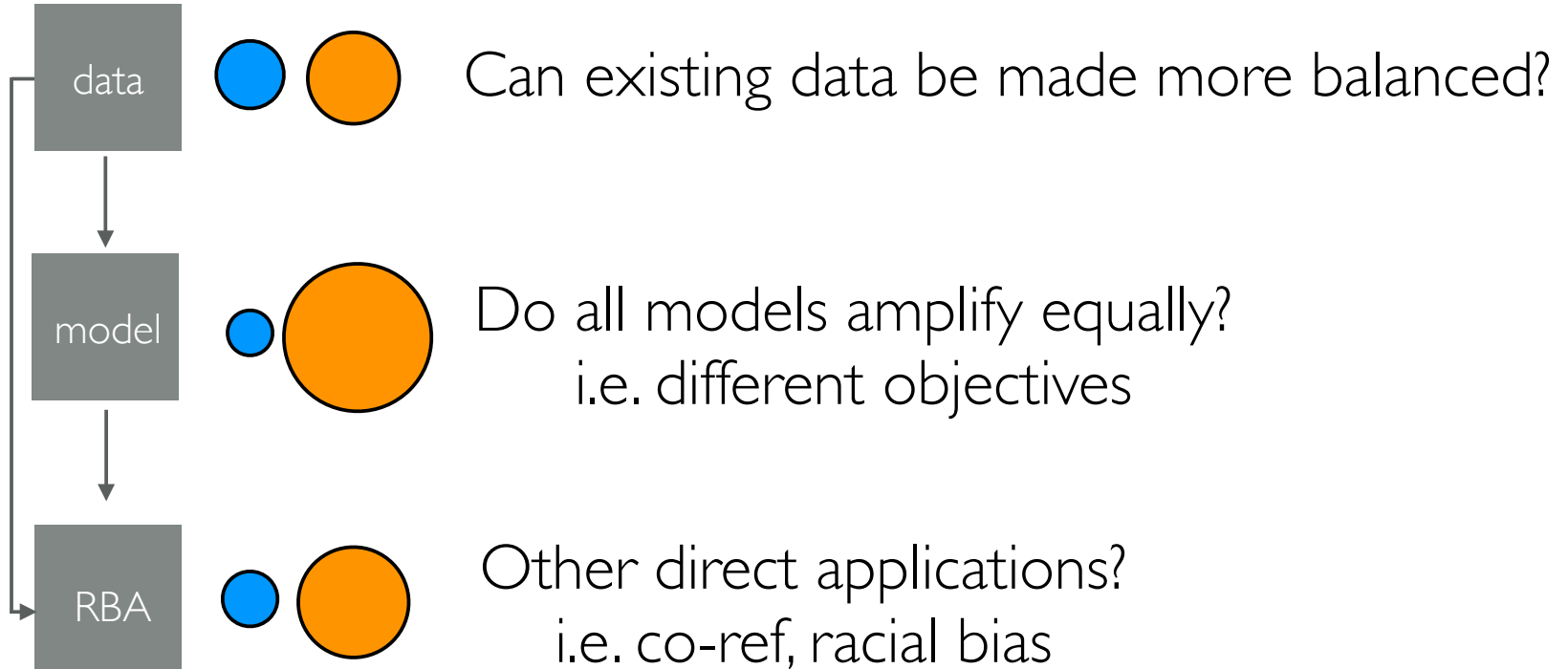


Gender Bias De-amplification in COCO

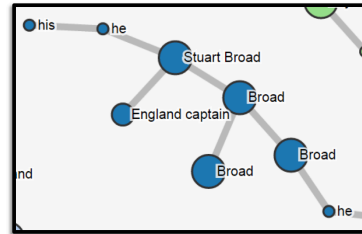
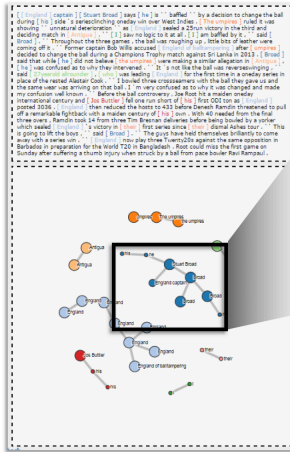
COCO Noun	Violation: 60.6%	.032 bias↑	45.27	mAP
w/ RBA	Violation: 36.4%	.022 bias↑	45.19	mAP



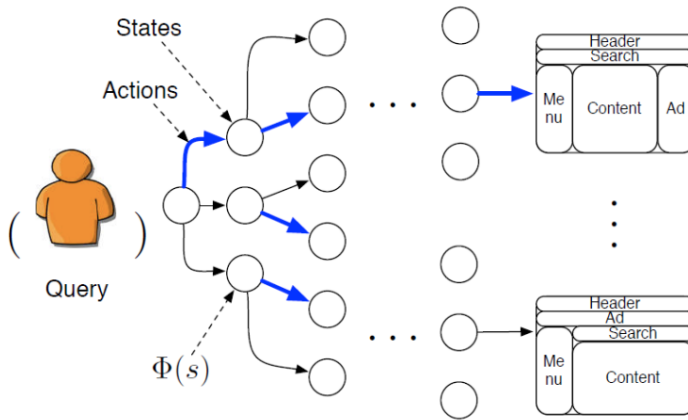
Future Work



What We Care about

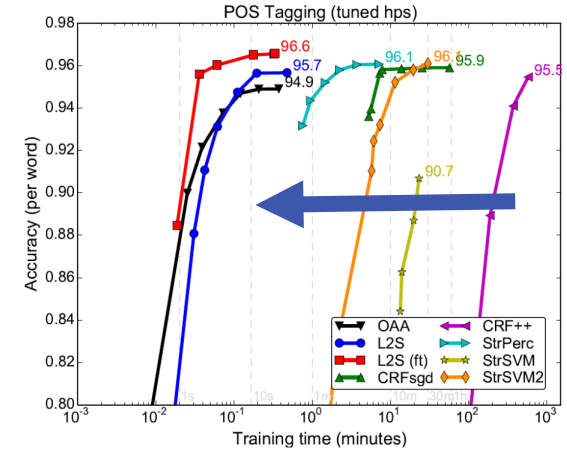


NLP Applications



Learning signals

Kai-Wei Chang (kwchang.net/talks/sp.html)



Training/test/dev speed

Query



activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Collaborators

❖ Academic

- ❖ Hal Daume (UMD)
- ❖ Chih-Jen Lin (NTU)
- ❖ Dan Roth (Upenn)
- ❖ Venkatesh Saligrama (BU)
- ❖ Alexander Rush (Harvard)
- ❖ Cho-Jui Hsieh (UCDavis)
- ❖ Vivek Srikumar (Utah)
- ❖ Hongning Wang (UVa)
- ❖ Vicente Ordonez (UVa)

❖ Industrial

- ❖ John Langford (MSR)
- ❖ Adam Kalai (MSR)
- ❖ Ming-Wei Chang (Google)
- ❖ Scott Yih (AI2)
- ❖ Hoifung Poon (MSR)
- ❖ Mark Yatskar (AI2)

❖ Students

- ❖ Wasi Ahmad (UCLA)
- ❖ Jieyu Zhao (UCLA)
- ❖ Rizwan Parvez (UCLA)
- ❖ Tianlu Wang (UVa)
- ❖ Ken Arnold (Harvard)
- ❖ Shyam Upaphaya (UIUC)

Conclusions

Goal: Practical Structured Prediction Approaches

Tutorials/Workshops:

1. AAI-16: Learning and Inference in SP Models
2. NAACL15: Hands-on Learning to Search for SP
3. EMNLP 16, 17: workshop SP for NLP

References/Code/Demos:

<http://kwchang.net>

Illinois-SL: a structured learning package

Vowpal Wabbit: an online learning library