Structured Prediction

Final words

CS 6355: Structured Prediction

THE UNIVERSITY OF UTAH

A look back

- What is a structure?
- The machine learning of interdependent variables

Recall: A working definition of a structure

A structure is a concept that can be applied to any complex thing, whether it be a bicycle, a commercial company, or a carbon molecule. By *complex*, we mean:

1. It is divisible into parts,
2. There are different kinds of parts,
3. The parts are arranged in a specifiable way, and,
4. Each part has a specifiable function in the structure of the thing as a whole

From the book Analysing Sentences: An Introduction to English Syntax by Noel Burton-Roberts, 1986.

An example task: Semantic Parsing

Find the largest state in the US

SELECT expression FROM table WHERE condition

MAX (numeric list)

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

$\boxed{\text{Find}}$ the largest state in the US

SELECT expression FROM table WHERE condition

$\boxed{\text{SELECT expression FROM table WHERE condition}}$

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

US_STATES

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

name US_STATES

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

`name` `US_STATES` Expression 1 = Expression 2

SELECT expression FROM table

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

`name` — `US_STATES` — Expression 1 = Expression 2

SELECT expression FROM table

MAX numeric list

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest $\boxed{\text{state}}$ in the US

SELECT expression FROM table WHERE condition

name

US_STATES

Expression 1 = Expression 2

SELECT expression FROM table

MAX numeric list

US_STATES

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

A plausible strategy to build the query

Find the largest state in the US

SELECT expression FROM table WHERE condition

`name` `US_STATES` Expression 1 = Expression 2

`size`

Or perhaps population?

SELECT expression FROM table

MAX numeric list `US_STATES`

`size`

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

```
US_CITIES      US_STATES
  name           name
population     population
  state          size
                capital
```

A plausible strategy to build the query

Find the largest state in the US

- At each step many, many decisions to make
- Some decisions are simply not allowed
  - A query has to be well formed!
- Even so, many possible options
  - Why does "Find" map to SELECT?
  - Largest by size/population/population of capital?

M table

US_STATES

size

SELECT expression FROM table WHERE condition

MAX numeric list

ORDERBY predicate

DELETE FROM table WHERE condition

SELECT expression FROM table

Expression 1 = Expression 2

US_CITIES	US_STATES
name	name
population	population
state	size
	capital

Standard classification tools can't predict structures

X: *"Find the largest state in the US."*

Y:

```
SELECT name
  FROM us_states
  WHERE size = (SELECT MAX(size) FROM us_states)
```

Classification is about making one decision

- Spam or not spam, or predict one label, etc

We need to make *multiple decisions*

- Each part needs a label
  - Should "*US*" be mapped to us_states or us_cities?
  - Should "*Find*" be mapped to SELECT or DELETE?
- The decisions interact with each other
  - If the outer FROM clause talks about the table us_states, then the inner FROM clause should not talk about utah_counties
- How to compose the fragments together to create the whole structure?
  - Should the output consist of a WHERE clause? What should go in it?

How did we get here?

Binary classification

- Learning algorithms
- Prediction is easy – Threshold
- Features (???)

Multiclass classification

- Different strategies
  - One-vs-all, all-vs-all
- Global learning algorithms
- One feature vector per outcome
  - Each outcome scored
- Prediction = highest scoring outcome

Structured classification

- Global models or local models
- Each outcome scored
- Prediction = highest scoring outcome
- Inference is no longer easy!
  - Makes all the difference

Structured output is…

- A graph, possibly labeled and/or directed
  - Possibly from a restricted family, such as chains, trees, etc.
  - A discrete representation of input
  - Eg. A table, the SRL frame output, a sequence of labels etc

Representation

- A collection of inter-dependent decisions
  - Eg: The sequence of decisions used to construct the output

Procedural

- The result of a combinatorial optimization problem
  - $\text{argmax}_{\mathbf{y} \in \text{all outputs}} \text{score}(\mathbf{x}, \mathbf{y})$

Formally

Challenges with structured output

- Two challenges
  1. We cannot train a separate weight vector for each possible inference outcome
     - For multiclass, we could train one weight vector for each label
  1. We cannot enumerate all possible structures for inference
     - Inference for binary/multiclass is easy

- Solution
  – Decompose the output into parts that are labeled
  – Define
    - how the parts interact with each other
    - how labels are scored for each part
    - an inference algorithm to assign labels to all the parts

Multiclass as a structured output

- A structure is…
  - A graph (in general, hypergraph), possibly labeled and/or directed
  - A collection of inter-dependent decisions
  - The output of a combinatorial optimization problem

    $\mathrm{argmax}_{\mathbf{y} \in \text{all outputs}} \mathrm{score}(\mathbf{x}, \mathbf{y})$

- Multiclass
  - A graph with one node and no edges
    - Node label is the output
  - Can be composed via multiple decisions
  - Winner-take-all

    $\mathrm{argmax}_i \, \mathbf{w}^T \phi(\mathbf{x}, i)$

Multiclass is a structure: Implications

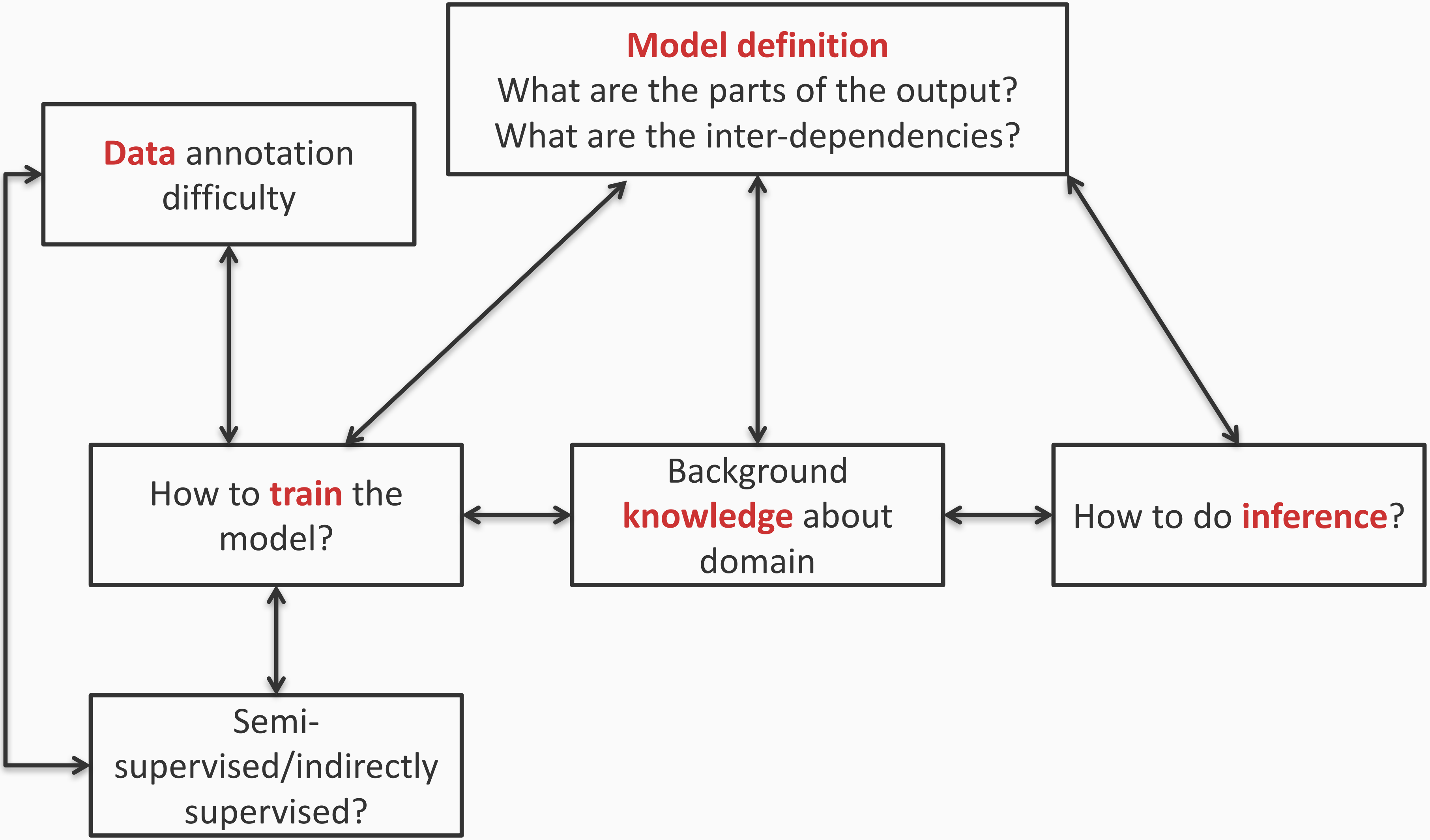
1. A lot of the ideas from multiclass *may* be generalized to structures
   – Not always trivial, but useful to keep in mind

2. Broad statements about structured learning must apply to multiclass classification
   – Useful for sanity check, also for understanding

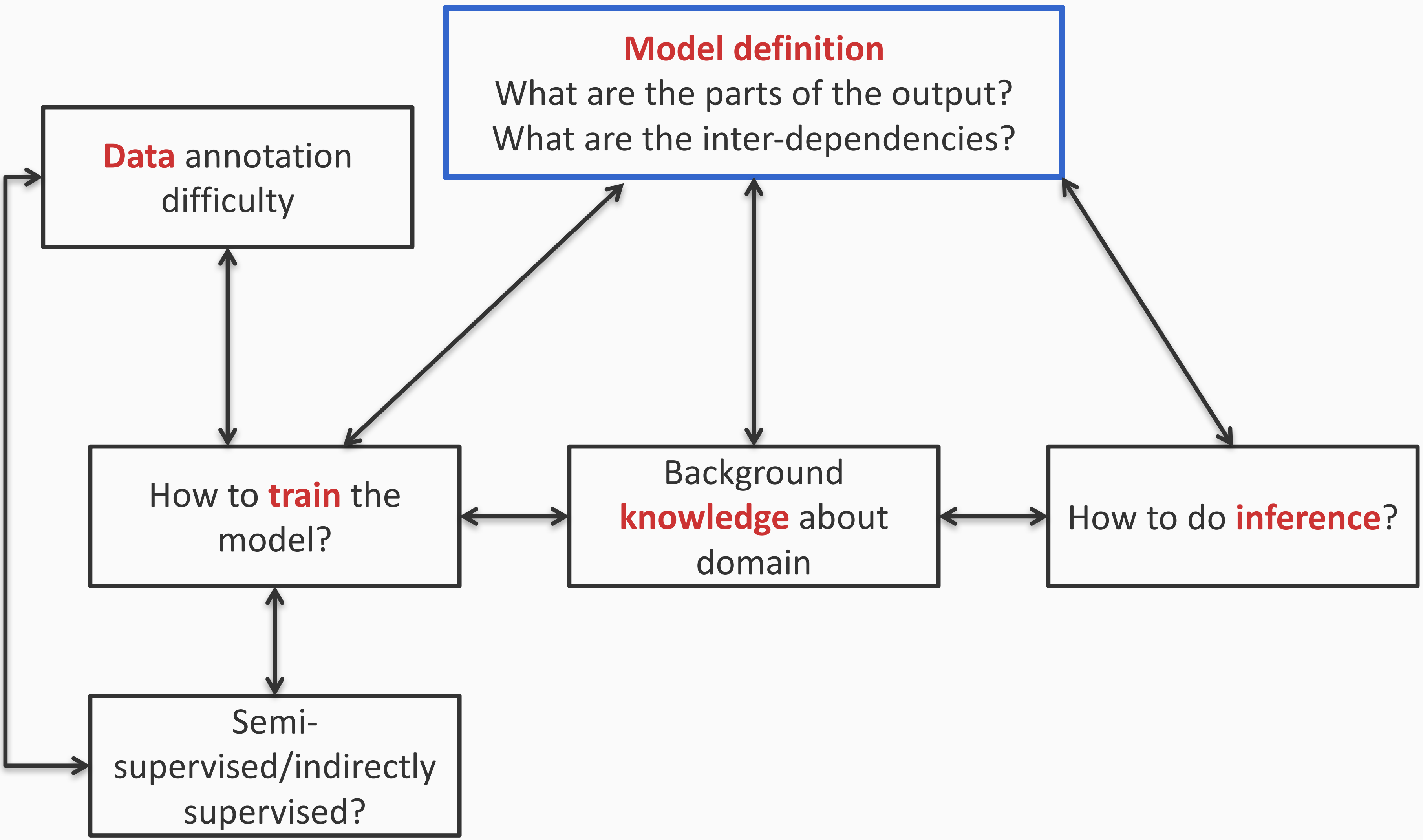
3. Binary classification is the most “trivial” form of structured classification
   – Multiclass with two classes

Structured Prediction
The machine learning of interdependent variables

Computational issues

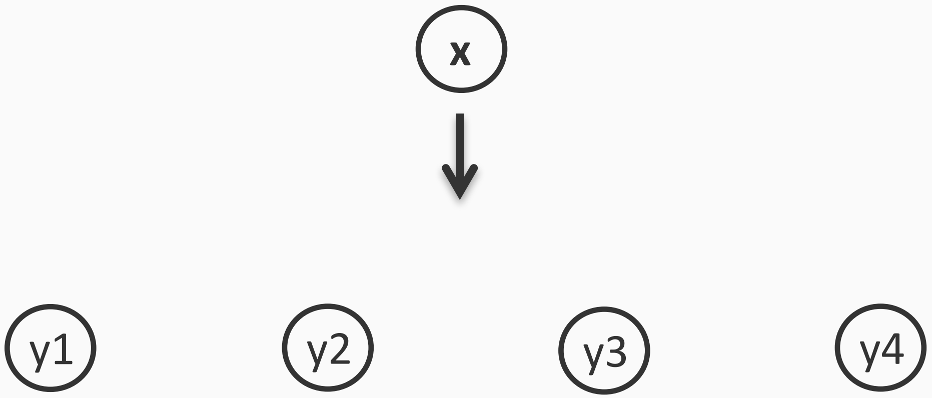


Computational issues



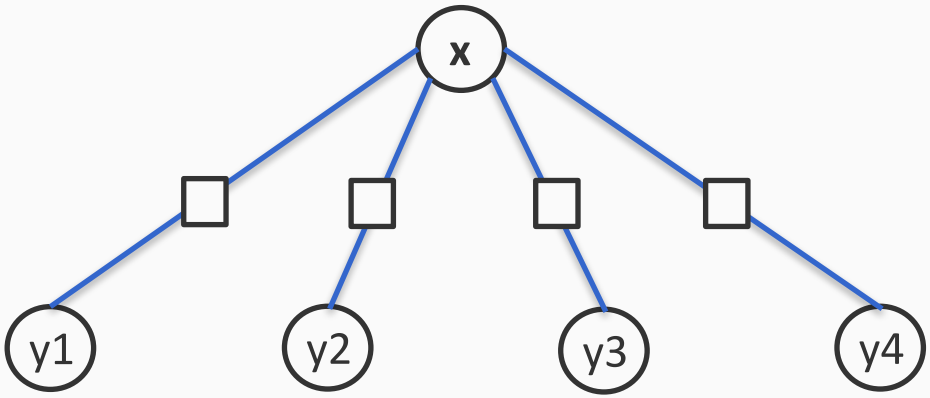
What does it mean to define the model?

Say we want to predict four output variables from some input



What does it mean to define the model?

Say we want to predict four output variables from some input



Recall: Each factor is a local expert about all the random variables connected to it

i.e. A factor can assign a score to assignments of variables connected to it

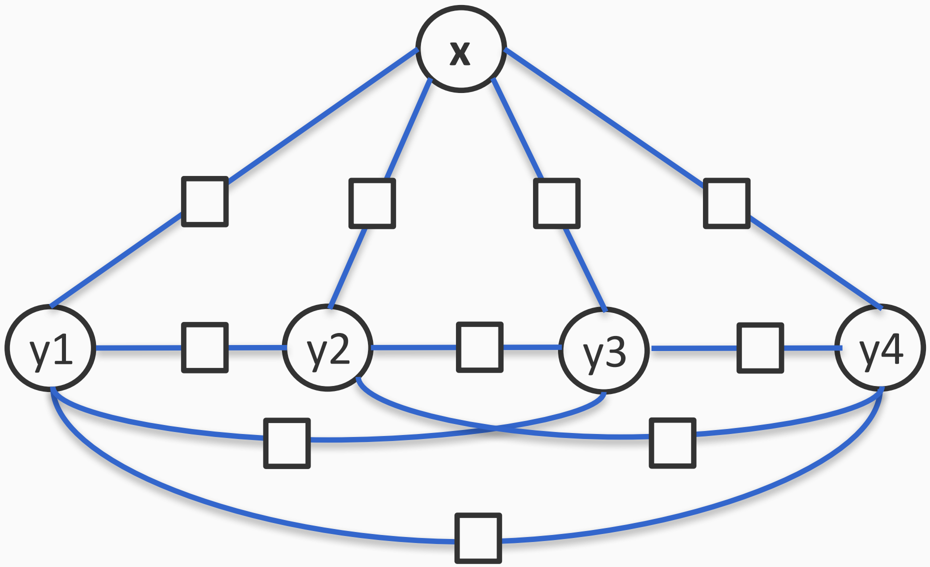
Option 1: Score each decision separately

Pro: Prediction is easy, each y independent

Con: No consideration of interactions

What does it mean to define the model?

Say we want to predict four output variables from some input



Option 2: Add pairwise factors

Recall: Each factor is a local expert about all the random variables connected to it

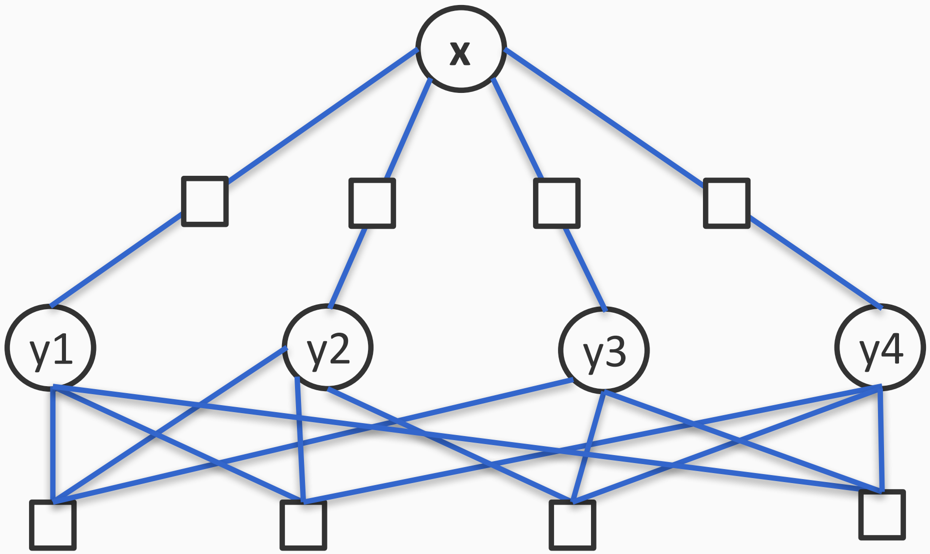
i.e. A factor can assign a score to assignments of variables connected to it

Pro: Accounts for pairwise dependencies

Cons: Makes prediction harder, ignores third and higher order dependencies

What does it mean to define the model?

Say we want to predict four output variables from some input



Recall: Each factor is a local expert about all the random variables connected to it

i.e. A factor can assign a score to assignments of variables connected to it

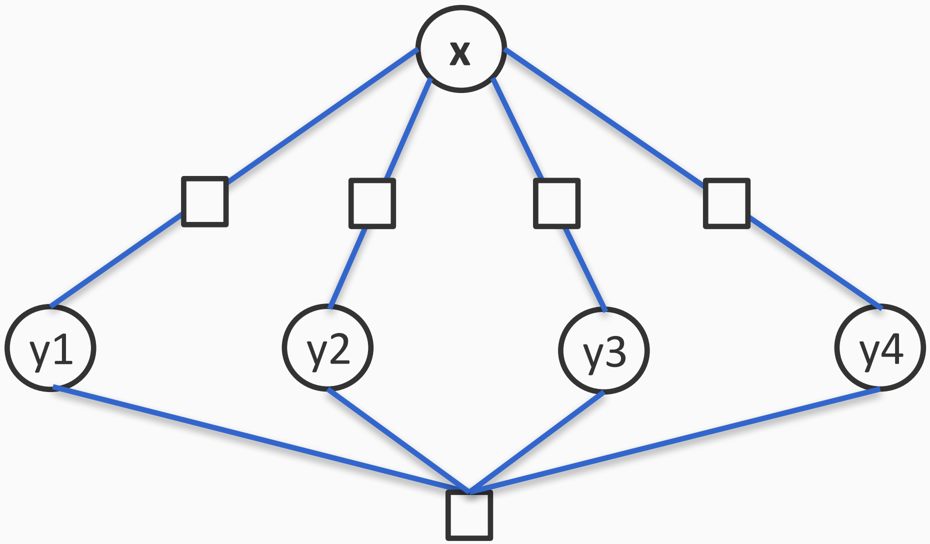
Option 3: Use only order 3 factors

Pro: Accounts for order 3 dependencies

Cons: Prediction even harder. Inference should consider all triples of labels now

What does it mean to define the model?

Say we want to predict four output variables from some input



Recall: Each factor is a local expert about all the random variables connected to it

i.e. A factor can assign a score to assignments of variables connected to it

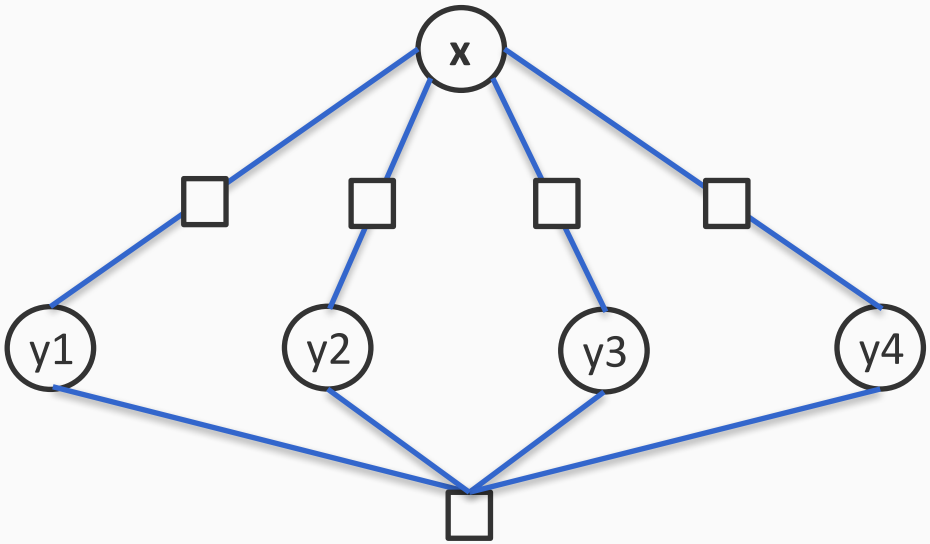
Option 4: Use order 4 factors

Pro: Accounts for order 4 dependencies

Cons: Basically no decomposition over the labels!

What does it mean to define the model?

Say we want to predict four output variables from some input



Recall: Each factor is a local expert about all the random variables connected to it

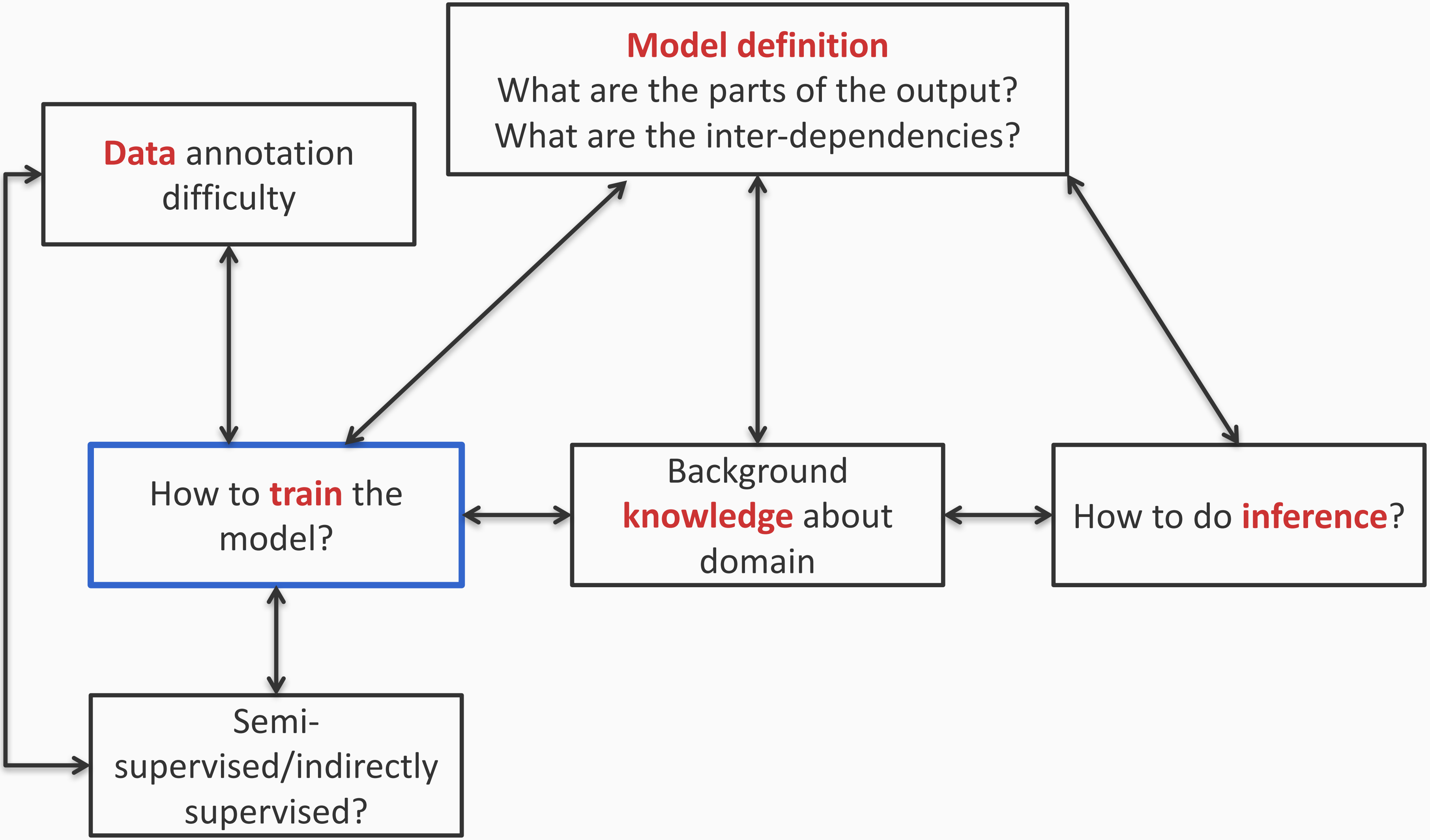
i.e. A factor can assign a score to assignments of variables connected to it

How do we decide what to do?

Some aspects to consider

- Availability of supervision
  - Supervised algorithms are well studied; supervision is hard (or expensive) to obtain
- Complexity of model
  - More complex models encode complex dependencies between parts; complex models make learning and inference harder
- Features
  - Most of the time we will assume that we have a good feature set to model our problem. But do we?
- Domain knowledge
  - Incorporating background knowledge into learning and inference in a mathematically sound way

Computational issues



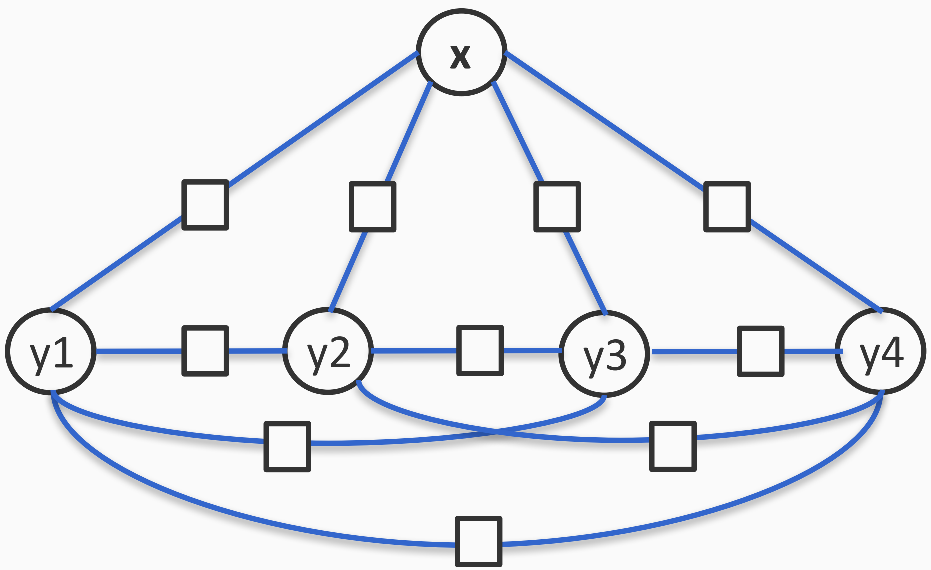
Training structured models

- Inference in training makes all the difference from multiclass/binary classification
- Empirical risk minimization principle
  - Minimize loss over the training data
  - Regularize the parameters to prevent overfitting
- We have seen different training strategies falling under this umbrella
  - Conditional Random Fields
  - Structural Support Vector Machines
  - Structured Perceptron (doesn’t have regularization)
- Different algorithms exist
  - We saw stochastic gradient descent in some detail

Training considerations

- Train globally vs train locally

Global: Train according to your final model

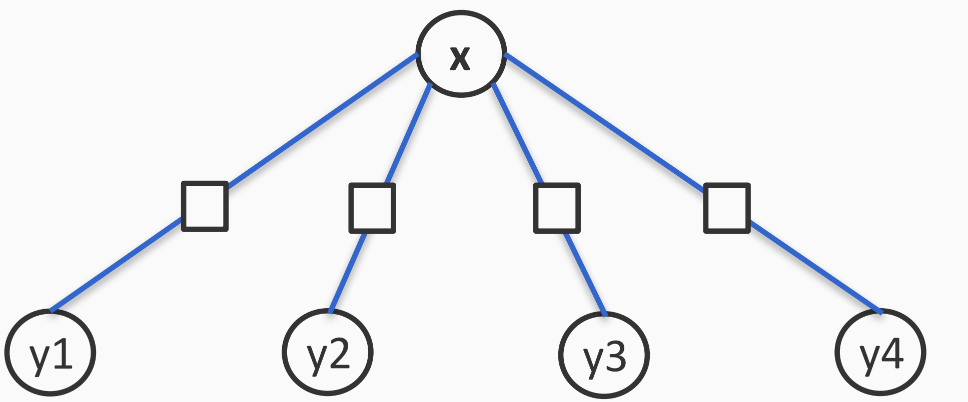


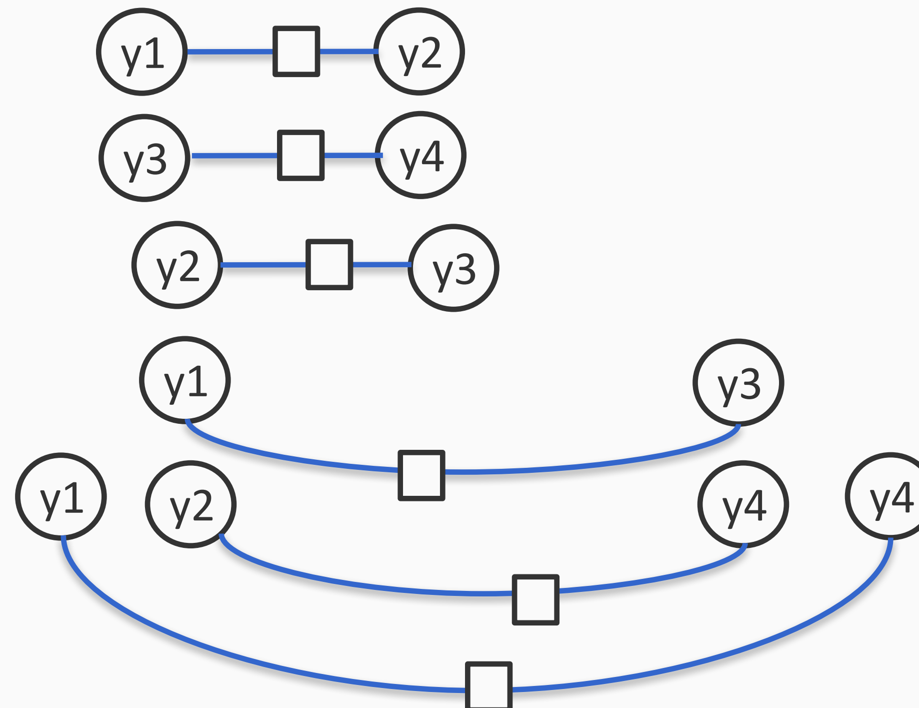
Pro: Learning uses all the available information
Con: Computationally expensive

Training considerations

- Train globally vs train locally

Local: Decompose your model into smaller ones and train each one separately
Full model still used at prediction time





Pro: Easier to train
Con: May not capture global dependencies

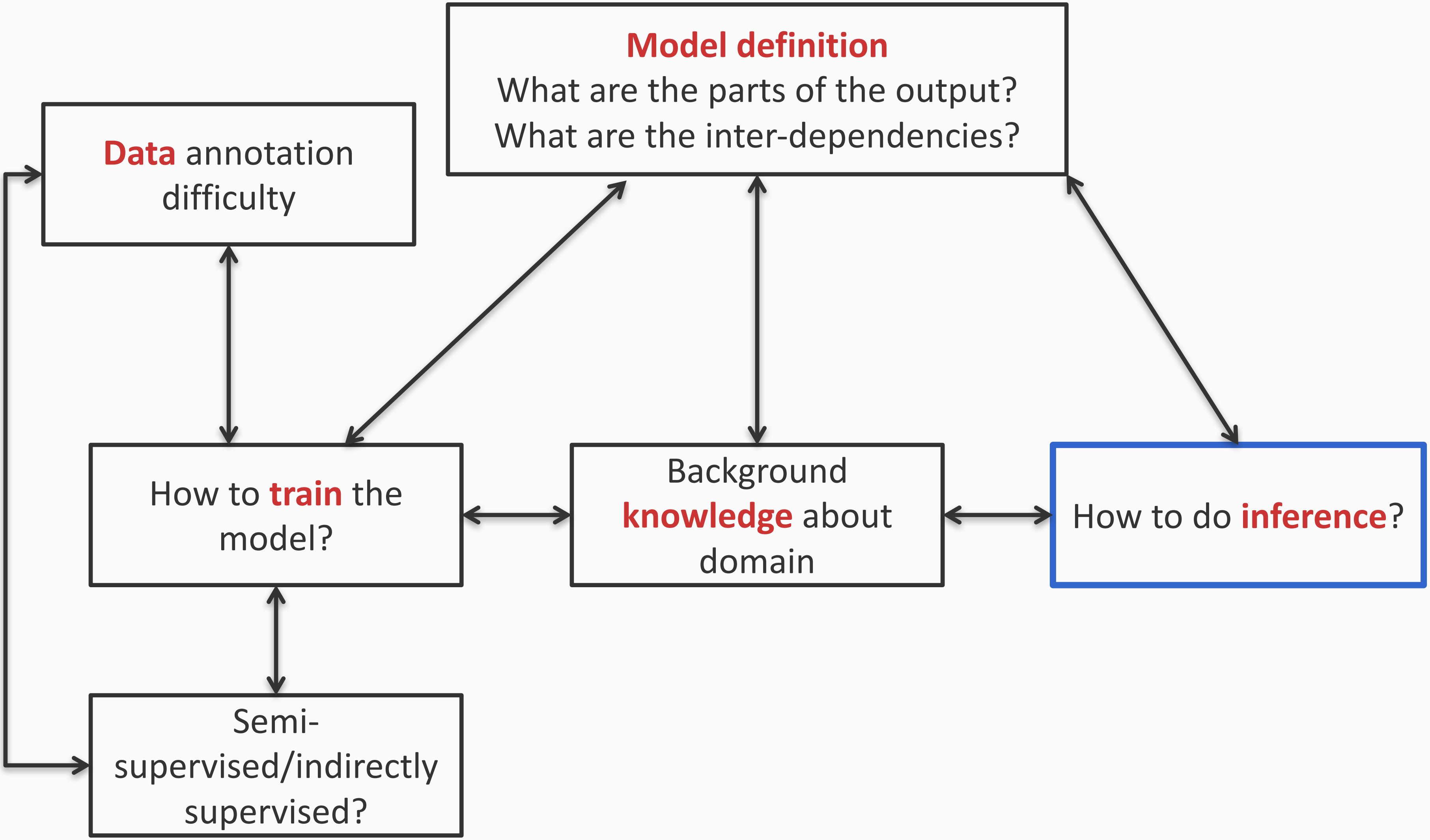
Training considerations

- Local vs global
  - Local learning
    - Learn parameters for individual components independently
    - Learning algorithm not aware of the full structure
  - Global learning
    - Learn parameters for the full structure
    - Learning algorithm “knows” about the full structure

How do we choose?

  - Depends on inference complexity
  - Jury still out on which one is better
  - Depends on size of available data too

Computational issues



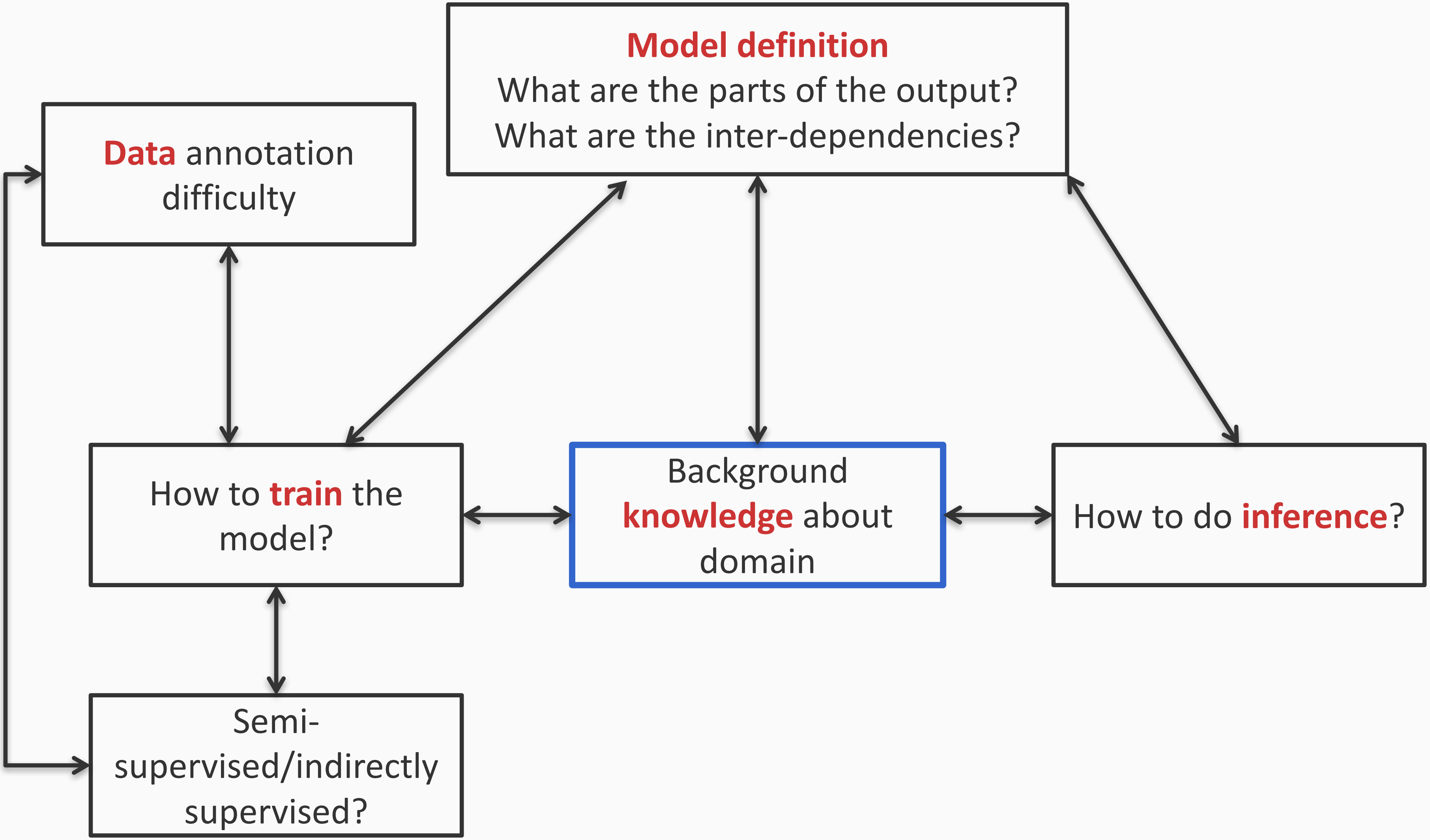
Inference

- What is inference? The prediction step
  - More broadly, an aggregation operation on the space of outputs for an example: max, expectation, sample, sum
  - Different flavors: MAP, marginal, loss augmented.
- Many algorithms, solution strategies
  - Combinatorial optimization, one size doesn't fit all
  - Graph algorithms, integer linear programming, heuristics, Monte Carlo methods, ...

How do we choose?

- Some tradeoffs
  - Programming effort
  - Exact vs inexact
  - Is the problem solvable with a known algorithm?
  - Do we care about the exact answer?

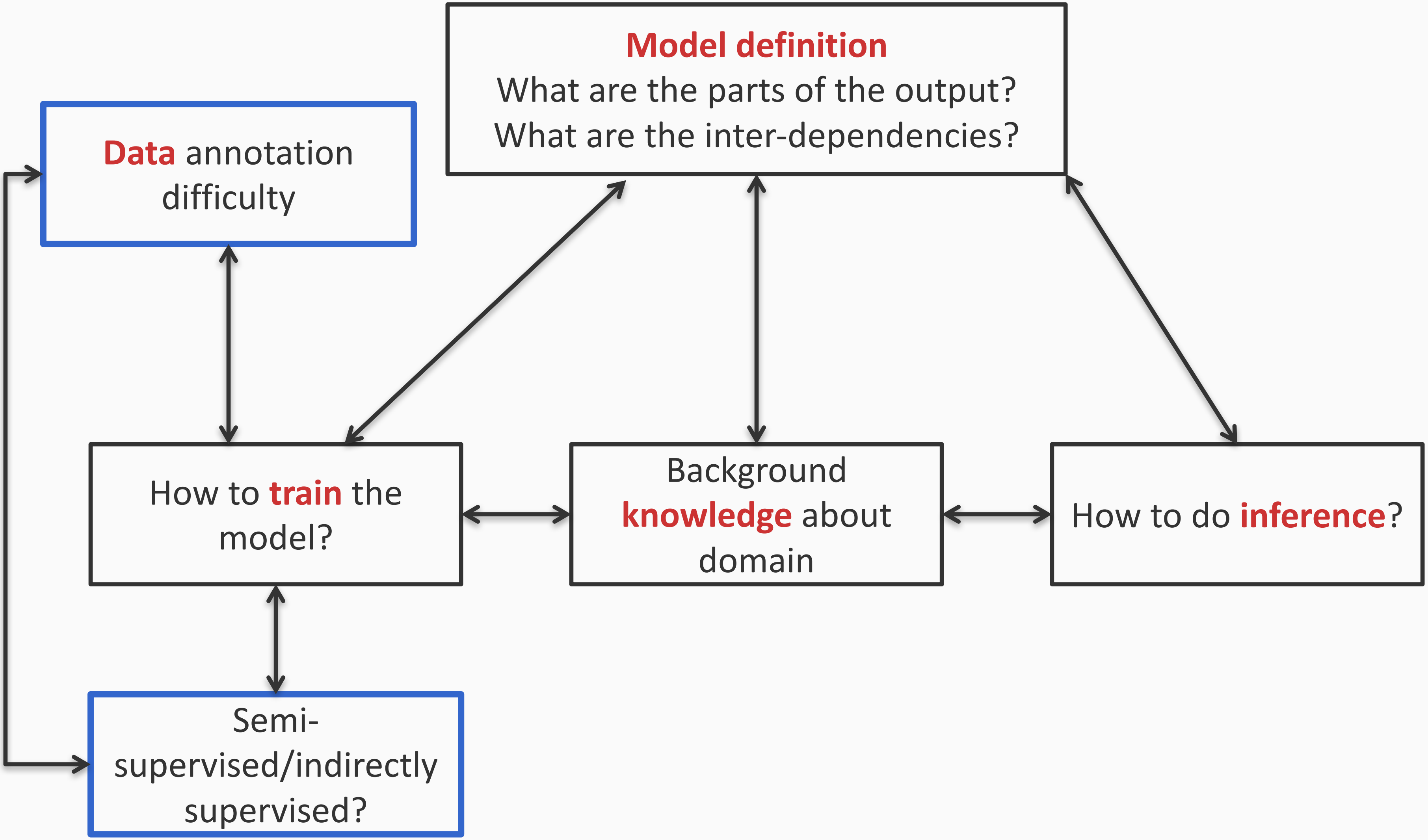
Computational issues



How does background knowledge affect your choices?

- Background knowledge biases your predictor in several ways
  - What is the model?
    - Maybe third order factors are not needed… etc
  - Your choices for learning and inference algorithms
  - Feature functions
  - Constraints that prohibit certain inference outcomes

Computational issues



Data and how it influences your model

- ***Annotated data is a precious resource***
  - Takes specialized expertise to generate
  - Or: very clever tricks (like online games that make data as a side effect)

- Important directions
  - Learning with latent representations, indirect supervision, partial supervision
  - In all these cases
    - Learning is rarely a convex problem
    - Modeling choices become very important! Bad model ***will*** hurt

Looking ahead

- Big questions (a very limited and biased set)
  - Representations
    - Can we learn the factorization?
    - Can we learn feature functions?
  - Dealing with the data problem for new applications
    - Clever tricks to get data
    - Taming latent variable learning
  - Applications
    - How does structured prediction help you?
    - Gathering importance as computer programs have to deal with uncertain, noisy inputs and make complex decisions