Transfer learning methods for the discovery of host-pathogen protein-protein interactions

Meghana Kshirsagar*, Judith Klein-Seetharaman, Jaime Carbonell

Language Technologies Institute, Carnegie Mellon University

Highlights

Propose techniques for discovering protein-protein interactions (PPIs) in new hosts or pathogens using interactions in known hosts/pathogens

We use known Salmonella-Human PPIs to predict interactions between Salmonella-Arabidopsis proteins

Salmonella   labeled PPI   **Source** host

transfer

Salmonella   **Target** host
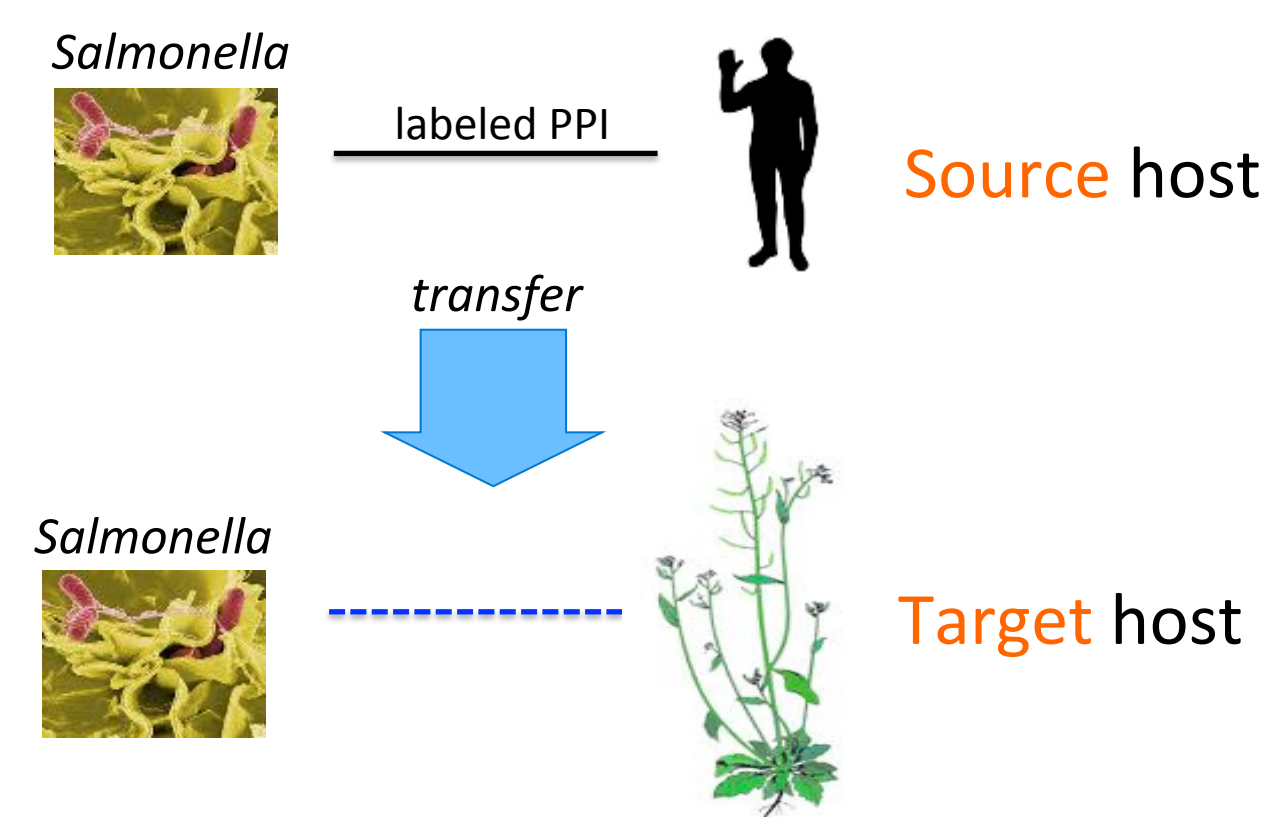
Two approaches are used and predictions from each are combined:
1. Infer interactions using orthologs of the host proteins, filter using intra-host PPI network alignment
2. Apply Transductive Support Vector Machines to label interactions in new host using known interactions as labeled data
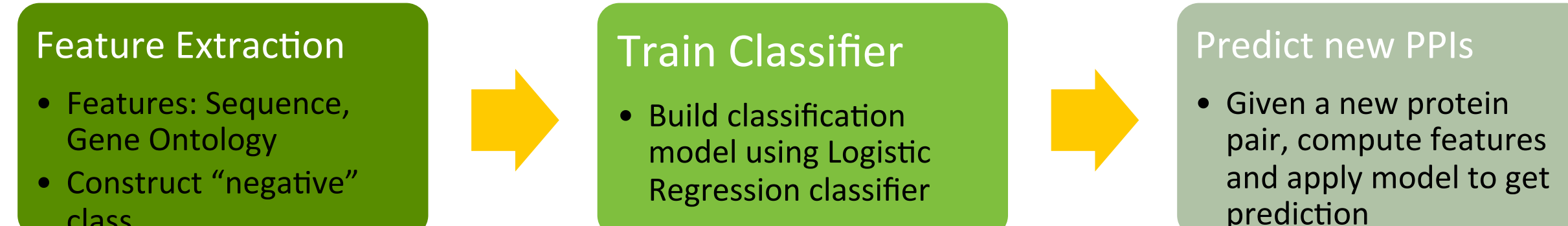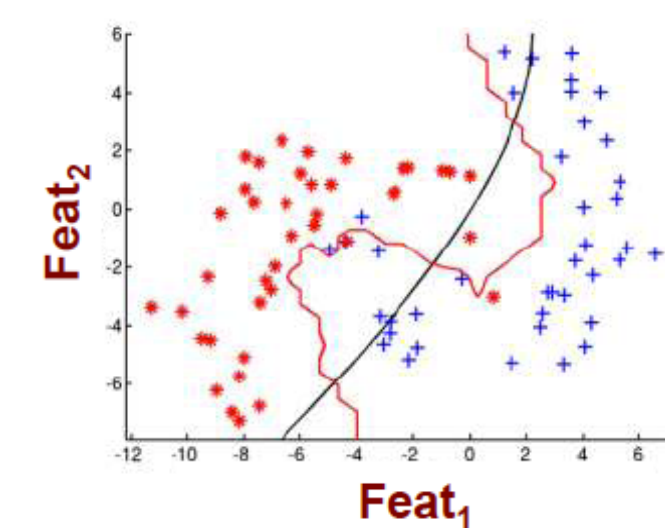
Challenges:

(a) No labeled data available for Arabidopsis, and a very small labeled dataset available for Human
(b) The two hosts namely Human and Plant have very different features. Building a joint model is tricky!
(c) Difficulty in evaluation of predicted interactions

Supervised learning models (Background)

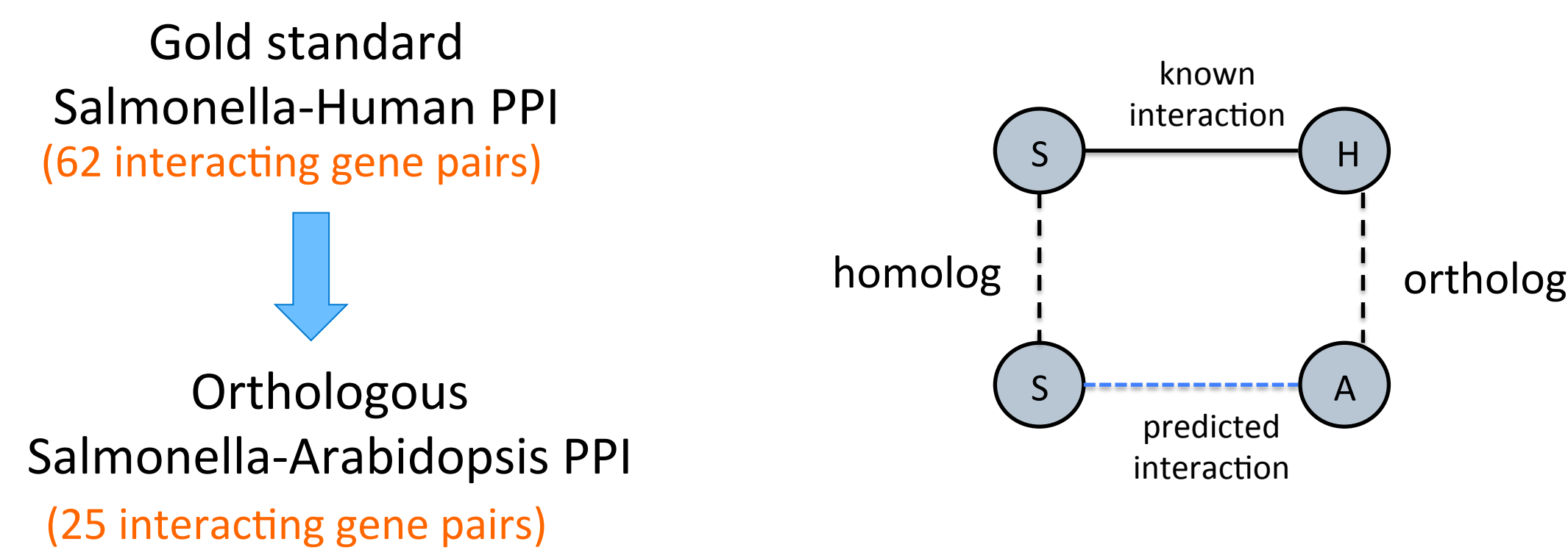Protein pair : (🟠 , 🟣) ➡ [feat$_1$, feat$_2$ .... feat$_N$]

Positive class * : known interacting pairs
Negative class + : other random pairs

Feat$_2$

Feat$_1$

Feature Extraction
• Features: Sequence, Gene Ontology
• Construct "negative" class

Train Classifier
• Build classification model using Logistic Regression classifier

Predict new PPIs
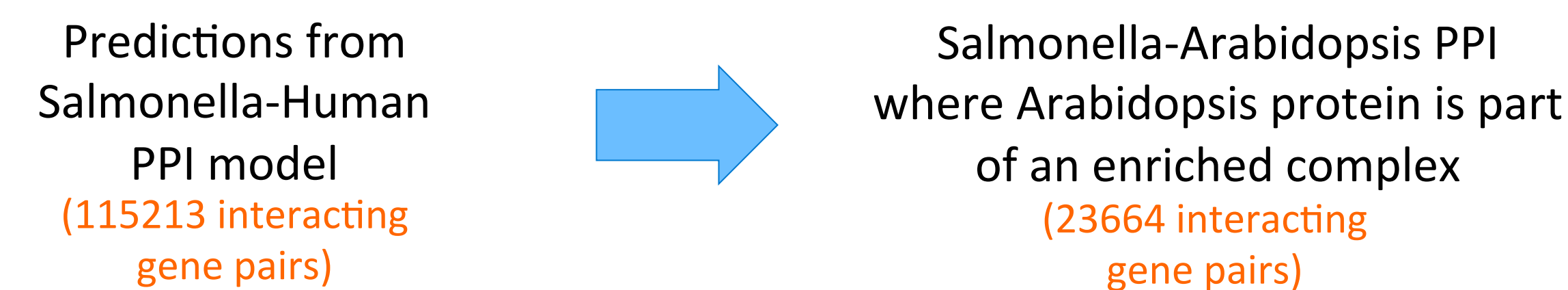• Given a new protein pair, compute features and apply model to get prediction

Approach - 1

(A) Using gold standard PPIs from source host (*Human*)

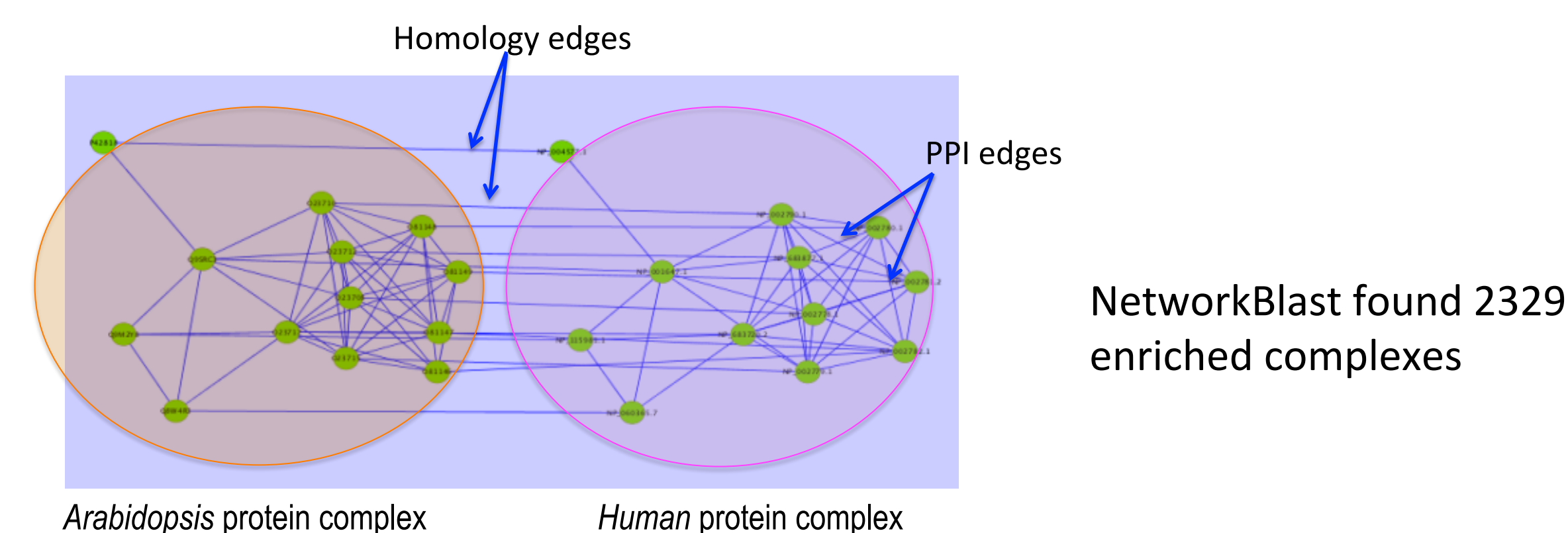Transfer using "ortholog" relationship between source and target host proteins

Gold standard Salmonella-Human PPI
(62 interacting gene pairs)

known interaction

S — H

homolog   ortholog

S - - A

predicted interaction

Orthologous Salmonella-Arabidopsis PPI
(25 interacting gene pairs)

(B) Using predicted interactions from source host (*Human*)

Transfer using "network alignment" relationship between the host PPI networks

Predictions from Salmonella-Human PPI model
(115213 interacting gene pairs)

➡

Salmonella-Arabidopsis PPI where Arabidopsis protein is part of an enriched complex
(23664 interacting gene pairs)

Enriched protein complexes computed using NetworkBlast[1] algorithm which aligns intra-host PPI network of source with target

Homology edges

PPI edges

NetworkBlast found 2329 enriched complexes

Arabidopsis protein complex   *Human* protein complex

Transductive SVM (TSVM)

+ −

SVM
Labeled data only

Unlabeled data

Transductive SVM

Optimizes over label assignments
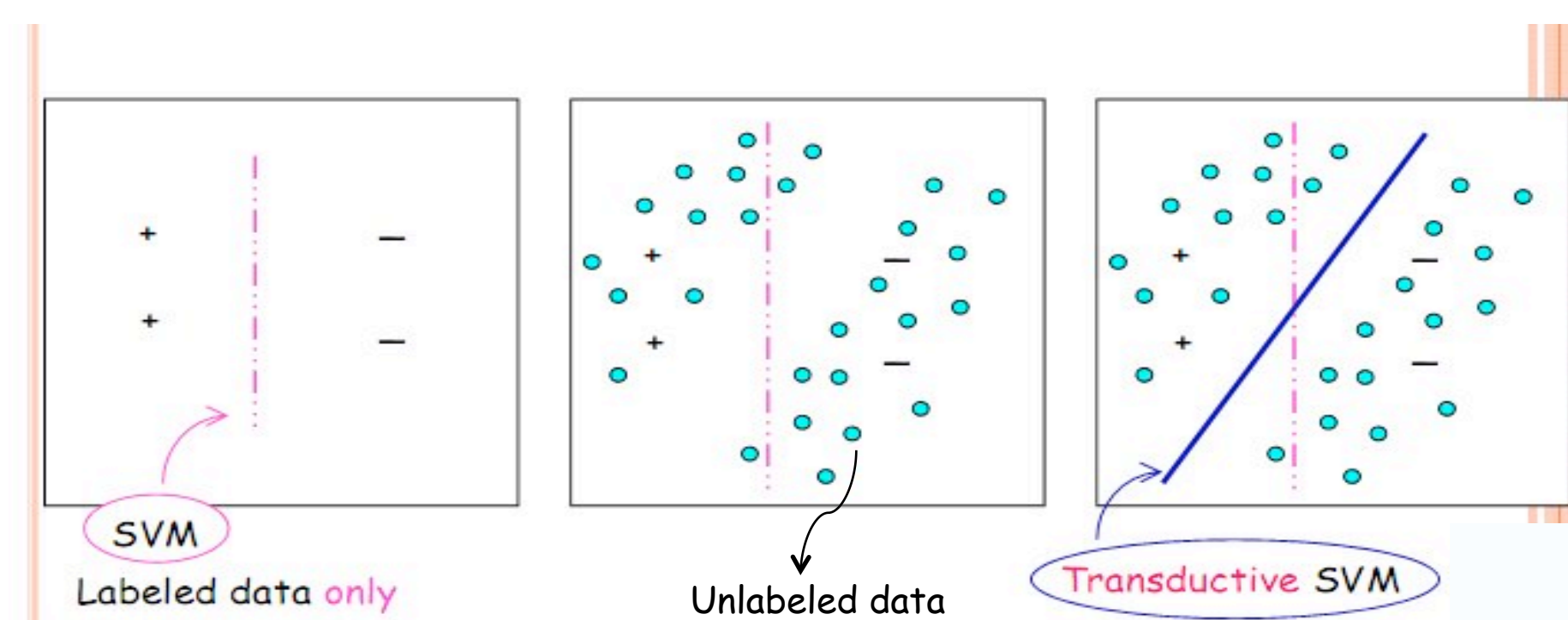
Transductive SVM Optimization function

$Minimize\ over\ (y_1^*, ..., y_n^*, \vec{w}, b):$

$$\frac{1}{2}\|\vec{w}\|^2$$

$subject\ to:$
$\forall_{i=1}^n : y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1$
$\forall_{j=1}^k : y_j^*[\vec{w} \cdot \vec{x}_j^* + b] \geq 1$

Approach - 2

(C) Uses unlabeled PPIs from target host (*Arabidopsis*) in addition to gold standard PPIs from source host (*Human*)

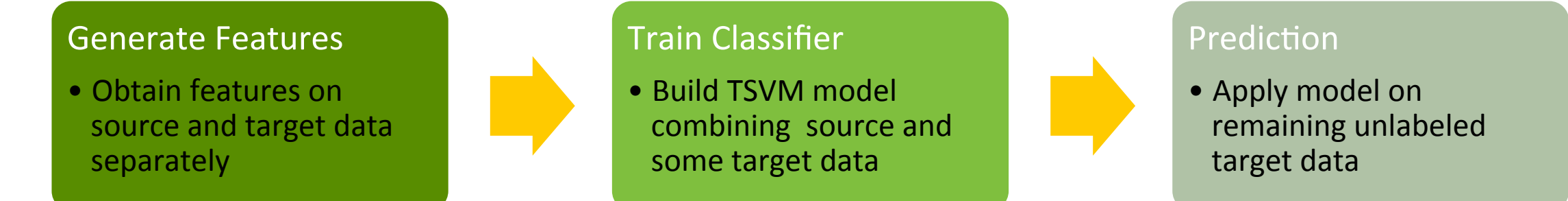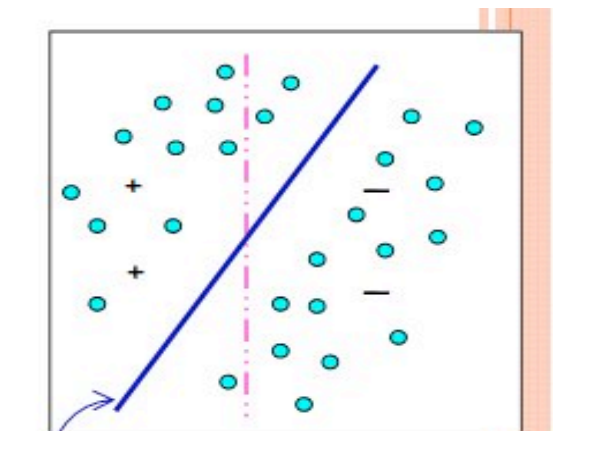Transfer using "similarity" between the source and target host proteins

Measure of similarity: defined using host protein properties like Gene-Ontology, Gene expression

$k(x_s^i, x_t^j) = sim(s_s, s_t) + sim(h_s, a_t)$
$k(x_s^i, x_s^j) = dot\ (x_s^i, x_s^j)$
$k(x_t^i, x_t^j) = dot\ (x_t^i, x_t^j)$

s_s/s_t − salmonella protein in source/target
h_s/a_t − human/ arabidopsis proteins in source/target resp.

+ positive *Salmonella-Human* PPI
− negative *Salmonella-Human* PPI
🔵 unlabeled *Arabidopsis-Human* PPI

Generate Features
• Obtain features on source and target data separately

Train Classifier
• Build TSVM model combining source and some target data

Prediction
• Apply model on remaining unlabeled target data

Experiments and Results

Source data (*Salmonella-Human PPIs*)
62 positives, 6200 negatives

Target data (Unlabeled *Salmonella-Arabidopsis*)
150584 interactions

Training: 3 fold CV using all source data + 2000 target examples

Best model applied to remaining unlabeled target examples to get predictions on target
Obtained 1087 interacting gene-pairs

Performance on source data using 3-fold Cross-validation		
Precision	Recall	F1
82.63	58.33	68.14

GO enrichment analysis

Applied FuncAssociate[2] for GO term enrichment analysis on predictions from both approaches. Some top terms are

Mass-spectrometry:

Binding studies on some predicted plant partners of Salmonella protein *spvC* show positive results.

abscisic acid transport	brassinosteroid mediated signaling pathway
defense response to bacterium	cellular response to hypoxia
response to karrikin	phospholipase activator activity
histone kinase activity	tubulin complex
basipetal auxin transport	calmodulin-dependent protein kinase activity

Conclusion

Our approaches to build a cross-species model shows very promising results. The techniques can be applied for any new host or pathogen.

Disadvantages:
TSVM solving combinatorial optimization using an approximation, no guarantees on optimality

Future Work:
1. Better experimental validation of predictions.
2. Other ways to transfer knowledge between organisms