



Aalto University  
School of Science  
and Technology

# Multilabel classification through structured output learning

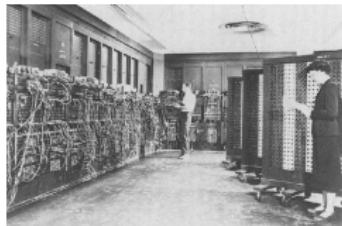
Hongyu Su

Department of Computer Science  
School of Science, Aalto University  
[hongyu.su@aalto.fi](mailto:hongyu.su@aalto.fi)

March 27, 2015

# Machine learning

- ▶ In 1946, the first fully electronic computer was built, known as ENIAC.



- ▶ In 1957, the perceptron algorithm was invented (Rosenblatt, 1958).
- ▶ In 1958, New York Times wrote perceptron as “the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence”.
- ▶ In 1959, Arthur Samuel defined machine learning as a “Field of study that gives computers the ability to learn without being explicitly programmed”.

# Main scope of this dissertation

- ▶ The dissertation focuses on classification learning, and multilabel classification in particular.

## Example: dog vs. cat?

- ▶ We have 5000 pictures of dog and 5000 pictures of cat.



- ▶ Computer digitalize each picture into  $100 \times 100$  pixels.
- ▶ Given a new picture, we want to answer: is it a dog or a cat?
- ▶ Simple task for human, dog, or cat.
- ▶ Golle (2008) claimed this is a difficult task for machines with only 82.7% accuracy (probability of getting a right answer).
- ▶ In 2013, 98.5% accuracy was reported in a Kaggle competition (<https://www.kaggle.com/c/dogs-vs-cats>).
- ▶ Why is this useful?

# In human verification system

- ▶ Human verification system is a program that protects website from robots by generating and grading test that human can pass but machine cannot.
- ▶ CAPTCHA system (Ahn et al., 2003) uses distorted text.



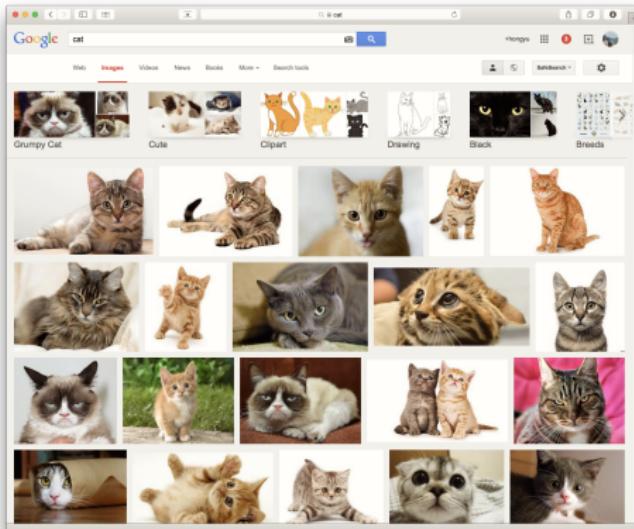
- ▶ ASIRRA system (Elson et al., 2007) uses images.



- ▶ To test if the ASIRRA system is safe from machine learning attack.
  - ▶ One should get all 12 pictures right!
  - ▶ Accuracy for machine is  $(98.5\%)^{12} \approx 83.4\%$ .

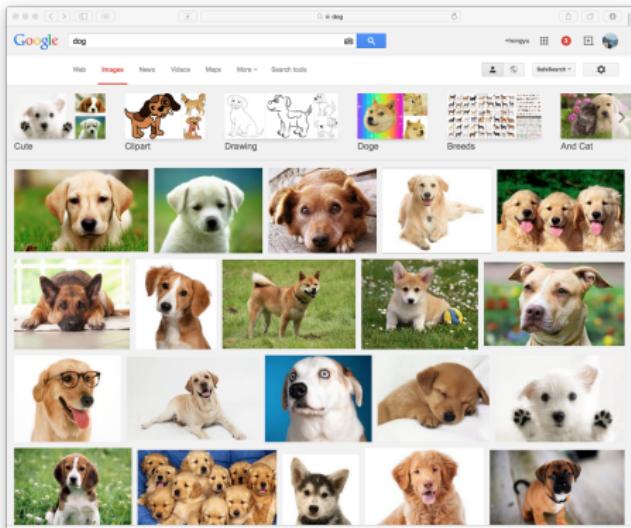
# In search engine

- ▶ If machine can assign cat/dog to all pictures correctly, we can search pictures with keywords.
- ▶ Search all **cat** pictures.



# In search engine

- ▶ If machine can assign cat/dog to all pictures correctly, we can search pictures with keywords.
- ▶ Search all **dog** pictures.



# Single label classification

- ▶ In machine learning, the problem is known as *single label classification*.
  - ▶ Input is an object  $\mathbf{x}$  (e.g., a picture).
  - ▶ Output is an attribute  $y$  called *label* (e.g.,  $y = +1$ :dog,  $y = -1$ :cat).
  - ▶ Explore a set of known object and label pairs called *training data*

$$\underbrace{\{(\mathbf{x}_1, +1), \dots, (\mathbf{x}_{5000}, +1)\}}_{\text{dog pictures}}, \underbrace{\{(\mathbf{x}_{5001}, -1), \dots, (\mathbf{x}_{10000}, -1)\}}_{\text{cat pictures}}.$$

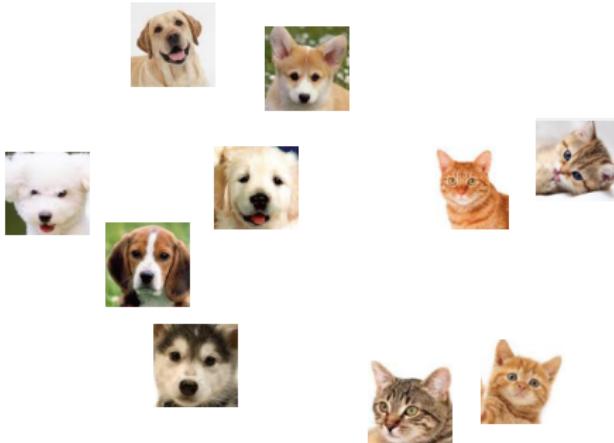
- ▶ Learn a *mapping function*  $f$  that predicts the label of a new object.

$$\mathbf{x} \xrightarrow{f} y, y \in \{+1, -1\}.$$

- ▶ Many algorithms are available to tackle single label classification problems, e.g., support vector machines (Cortes and Vapnik, 1995), logistic regression (Chen and Rosenfeld, 1999).

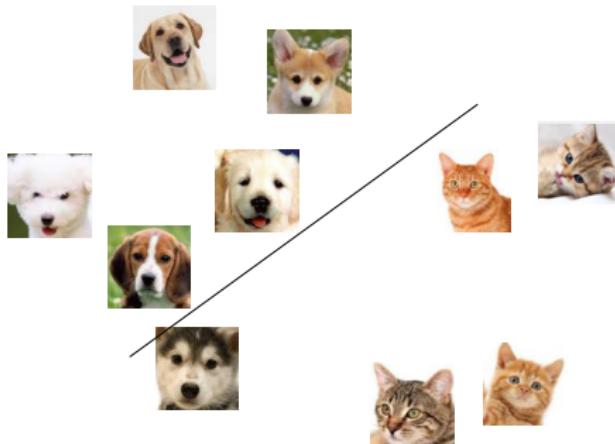
# Support vector machine

- ▶ Represent objects into a feature space (e.g., points in 2D space.)
- ▶ A feature space is a high dimensional space made by *kernel functions* (Shawe-Taylor and Cristianini, 2004).



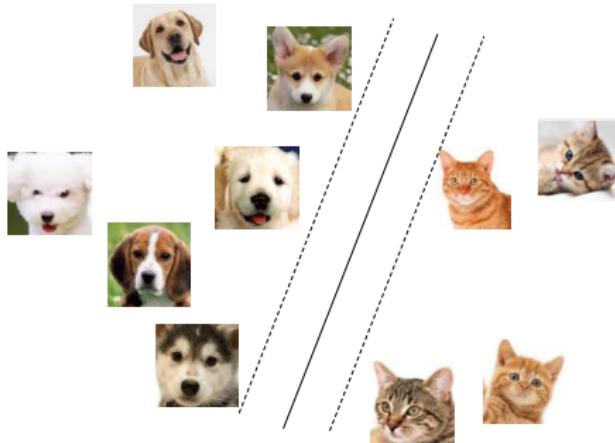
# Support vector machine

- ▶ Find a *hyperplane (classifier)* to separate objects of two classes.
- ▶ Minimize the number of mistakes made by the classifier. This is known as *empirical risk minimization* (Vapnik, 1992).



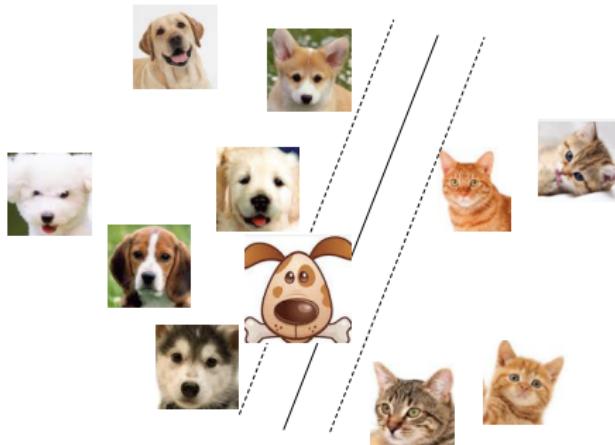
# Support vector machine

- ▶ We want the hyperplane to separate two classes with a big “gap”.
- ▶ “Gap” is known as *margin* which gives us enough confidence to deal with new objects (Evgeniou et al., 1999, 2002).



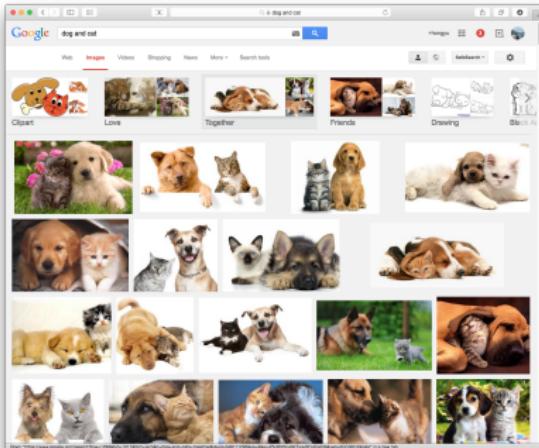
# Support vector machine

- ▶ Represent the new object into the same feature space.
- ▶ The classifier will generate the label of the new object according to its side.



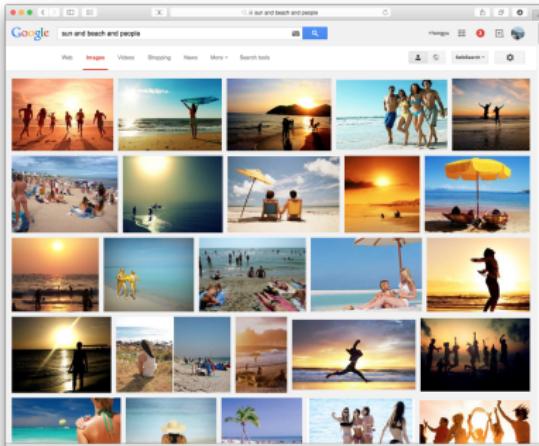
# Image annotation task

- ▶ We are often interested in multiple attributes of a single picture.
- ▶ For example, we want to assign multiple tags to one picture.  
    {boat, sea, sun, beach, people, dog, cat}
- ▶ Correct annotations will allow us to search with multiple attributes.
- ▶ Search with **dog & cat**.



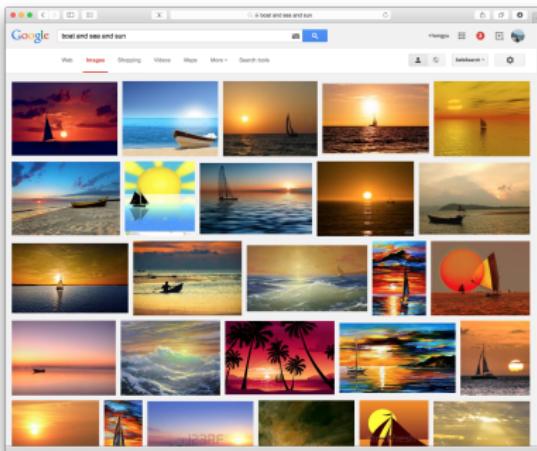
# Image annotation task

- ▶ We are often interested in multiple attributes of a single picture.
- ▶ For example, we want to assign multiple tags to one picture.  
    {boat, sea, sun, beach, people, dog, cat}
- ▶ Correct annotations will allow us to search with multiple attributes.
- ▶ Search with **sun & beach & people**.



# Image annotation task

- ▶ We are often interested in multiple attributes of a single picture.
- ▶ For example, we want to assign multiple tags to one picture.  
    {boat, sea, sun, beach, people, dog, cat}
- ▶ Correct annotations will allow us to search with multiple attributes.
- ▶ Search with **boat & sea & sun**.



# Multilabel classification

- ▶ The problem is known as *multilabel classification*, which is a natural extension to single label classification.
  - ▶ Input  $\mathbf{x}$  is an object (e.g., a picture).
  - ▶ Output  $\mathbf{y}$  are multiple attributes called *multilabel*

$$\mathbf{y} = (+1, +1, -1, -1, +1, -1, -1).$$

boat    sea    sun    beach    people    dog    cat

- ▶ Explore a set of known object and label pairs called training data.
- ▶ Learn a *mapping function* that predicts the best multilabel of a new object.

$$\mathbf{x} \xrightarrow{f} \mathbf{y} = (y_1, \dots, y_k).$$

- ▶ Multilabel classification is an active research field in machine learning.

# Applications

- ▶ Pictures can associate with multiple tags.


$$(+1, +1, -1, -1, -1, +1, +1)$$

boat sea sun beach people ice land

- ▶ News articles can be assigned to multiple categories.


$$(+1, +1, -1, -1, -1, -1, -1)$$

news economics sports politics movie science art

- ▶ Drugs can be effective for multiple symptoms.


$$(+1, +1, +1, +1, -1, -1, +1)$$

heart stroke blood fever digest liver swelling

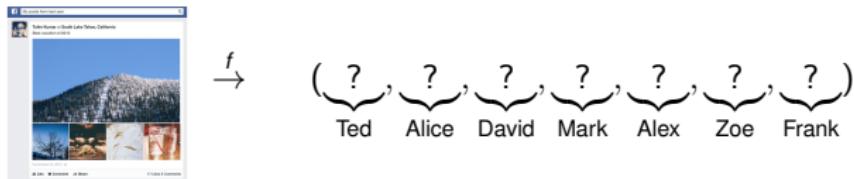
- ▶ Information can spread through multiple users in social networks.


$$(+1, -1, +1, -1, +1, -1, -1)$$

Ted Alice David Mark Alex Zoe Frank

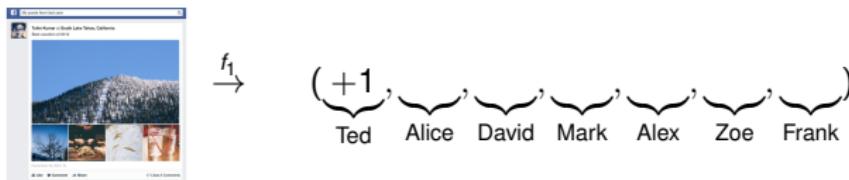
# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



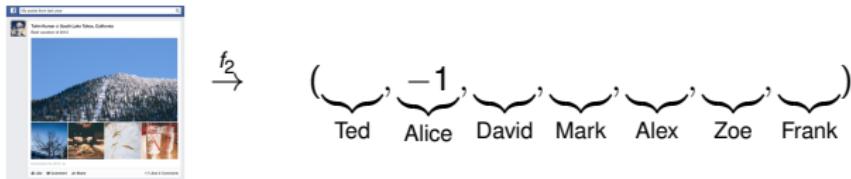
# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



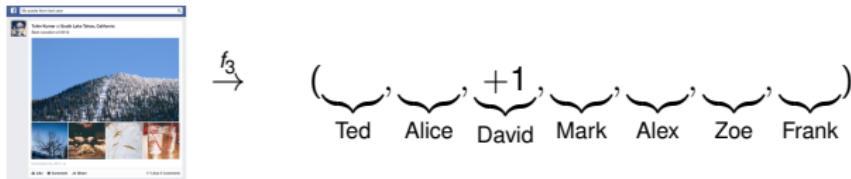
# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



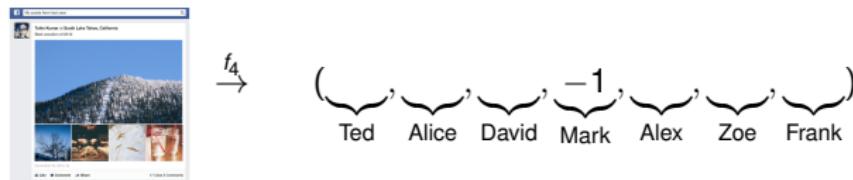
# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



# How to solve multilabel classification?

- ▶ Reduce the multilabel classification problem as a collection of single label classification problems.
- ▶ Solve each individual problem independently.
- ▶ Concatenate the predictions.



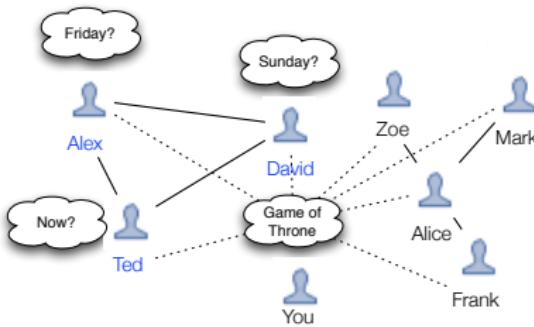
$$f_1, \dots, f_7 \rightarrow (+1, -1, +1, -1, +1, -1, -1)$$

Ted   Alice   David   Mark   Alex   Zoe   Frank

# Label correlations

- ▶ Multiple attributes are often closely related. Similar attributes will have similar responses to an input.
- ▶ Social network analysis: friends have similar hobbies.

( $\underbrace{+1}_{\text{Ted}}$ ,  $\underbrace{-1}_{\text{Alice}}$ ,  $\underbrace{+1}_{\text{David}}$ ,  $\underbrace{-1}_{\text{Mark}}$ ,  $\underbrace{+1}_{\text{Alex}}$ ,  $\underbrace{-1}_{\text{Zoe}}$ ,  $\underbrace{-1}_{\text{Frank}}$ )

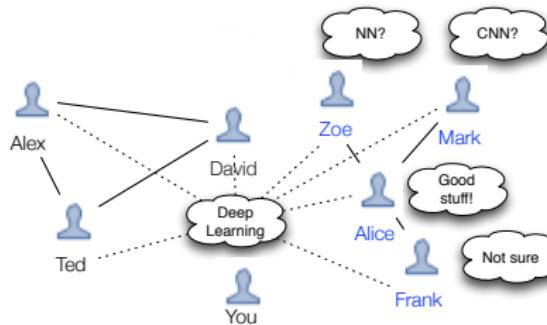


- ▶ Document classification: A news about politics may be also about economics.
- ▶ Label correlations can help make better predictions.

# Label correlations

- ▶ Multiple attributes are often closely related. Similar attributes will have similar responses to an input.
- ▶ Social network analysis: friends have similar hobbies.

( $\underbrace{-1, +1}_{\text{Ted}}$ ,  $\underbrace{-1, +1}_{\text{Alice}}$ ,  $\underbrace{-1, +1}_{\text{David}}$ ,  $\underbrace{+1, -1}_{\text{Mark}}$ ,  $\underbrace{-1, +1}_{\text{Alex}}$ ,  $\underbrace{+1, +1}_{\text{Zoe}}$ ,  $\underbrace{+1, +1}_{\text{Frank}}$ )



- ▶ Document classification: A news about politics may be also about economics.
- ▶ Label correlations can help make better predictions.

# Structured output prediction

- ▶ We want to explore the correlations of attributes to improve the performance on multilabel classification problems.
- ▶ In statistics, graph is a natural way to model correlations. *Output graph* is defined by
  - ▶ Nodes correspond to multiple attributes.
  - ▶ Edges correspond to correlations of attributes.
- ▶ *Structured output prediction* method
  - ▶ predicts multiple attributes of an object at the same time.
  - ▶ explores the correlations described by an output graph.

# Contributions of this dissertation

- ▶ The main contributions are several structured output learning algorithms that improve the performance on multilabel classification problems.
- ▶ In addition, it also contributes to theoretical studies on the performance of the proposed learning algorithms.
- ▶ For the multilabel classification problems where the output graph is given *a priori*.
  - ▶ Improve the performance on drug sensitivity prediction problems (Su et al., 2010).
  - ▶ Predict reliably the spread of a content in social networks (Su et al., 2014).
- ▶ For general multilabel classification problems without predefined output graph
  - ▶ Several ensemble methods that combine a collection of structured output learners (Su and Rousu, 2011, 2013, 2014).
  - ▶ A joint learning and inference framework with theoretical guarantee on the performance (Marchand et al., 2014).

# Future work

- ▶ Proposed algorithms can be applied to many real world multilabel classification tasks.
  - ▶ Image annotation, document classification, drug activity prediction, social network analysis.
  - ▶ Sentiment analysis, music categorization, protein function prediction, etc.
- ▶ Algorithm developments:
  - ▶ Select and combine a collection of random output graphs to discover latent structure.
  - ▶ Develop new and fast inference algorithm that allows learning on large scale datasets.

# Image source

- ▶ Eniac pictures are from Wikipedia.
- ▶ Animal pictures are from Google.
- ▶ CAPTCHA system picture is from <http://www.captcha.net>.
- ▶ ASIRRA system picture is from  
<http://research.microsoft.com/en-us/um/redmond/projects/asirra/>.
- ▶ Application pictures are from Google.
- ▶ Social network pictures are from Facebook.

# Bibliography

- Ahn, L. V., Blum, M., Hopper, N. J., and Langford, J. (2003). Captcha: Using hard ai problems for security. In *Proceedings of the 22Nd International Conference on Theory and Applications of Cryptographic Techniques*, EUROCRYPT'03, pages 294–311, Berlin, Heidelberg. Springer-Verlag.
- Chen, S. and Rosenfeld, R. (1999). *A Gaussian Prior for Smoothing Maximum Entropy Models*. PhD thesis, Computer Science Department, Carnegie Mellon University. Technical Report CMU-CS-99-108.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Elson, J., Douceur, J. R., Howell, J., and Saul, J. (2007). Asirra: A captcha that exploits interest-aligned manual image categorization. In *Proceedings of 14th ACM Conference on Computer and Communications Security (CCS)*. Association for Computing Machinery, Inc.
- Evgeniou, T., Poggio, T., Pontil, M., and Verri, A. (2002). Regularization and statistical learning theory for data analysis. *Comput. Stat. Data Anal.*, 38(4):421–432.

## Bibliography (cont.)

- Evgeniou, T., Pontil, M., and Poggio, T. (1999). A unified framework for regularization networks and support vector machines. Technical report, Cambridge, MA, USA.
- Golle, P. (2008). Machine learning attacks against the asirra captcha. In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, CCS '08, pages 535–542, New York, NY, USA. ACM.
- Marchand, M., Su, H., Morvant, E., Rousu, J., and Shawe-Taylor, J. (2014). Multilabel structured output learning with random spanning trees of max-margin markov networks. In *Advances in Neural Information Processing System NIPS2014*, page to appear.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.
- Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

## Bibliography (cont.)

- Su, H., Gionis, A., and Rousu, J. (2014). Structured prediction of network response. In *Proceedings, 31th International Conference on Machine Learning ICML2014*, volume 32 of *Journal of Machine Learning Research WCP*, pages 442–450.
- Su, H., Heinonen, M., and Rousu, J. (2010). Structured output prediction of anti-cancer drug activity. In *Proceedings, 5th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB2010)*, volume 6282 of *Lecture Note in Computer Science*, pages 38–49.
- Su, H. and Rousu, J. (2011). Multi-task drug bioactivity classification with graph labeling ensembles. In *Proceedings, 6th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB2011)*, volume 7035 of *Lecture Note in Computer Science*, pages 157–167.
- Su, H. and Rousu, J. (2013). Multilabel classification through random graph ensembles. In *Proceedings, 5th Asian Conference on Machine Learning (ACML2013)*, volume 29 of *Journal of Machine Learning Research WCP*, pages 404–418.

# Bibliography (cont.)

- Su, H. and Rousu, J. (2014). Multilabel classification through random graph ensembles. *Machine Learning*, 1(1).
- Vapnik, V. (1992). Principles of risk minimization for learning theory. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. Morgan-Kaufmann.

Thank you!