
Products
Think
Safety
First!

Can I See Some Hands Please?

- Who is familiar with
  - Semi-supervised learning?
  - Multiple instance learning?
  - Transfer learning / domain adaptation?
  - Active learning?

Some Response Please…?

- How many different classifiers are there around that optimize for classification error rate [or accuracy or 0-1 loss]?
  - < 3
  - = 3
  - = 4
  - > 4

Construction or Sketch Please...?

- Give me a classification problem $p(x, y)$ and a classifier such that expected error rate **grows** with **increasing** number of training samples
  - Or accuracy in expectation drops with more samples...

Dealing with Weakly and Partially Annotated Data

Concepts, Methodologies, and Safety First?

Marco Loog

Delft University of Technology
University of Copenhagen

*with here and there some stuff clearly inspired by work from Veronika Cheplygina and David M.J. Tax

Introduction

- Supervised leaning methods among best performing algorithms in biomedical image analysis
  - No distinction here between ML, PR, statistical learning, etc.
  - Note : in some sense, there is always room for supervised learning... unless one neglects proper validation

Introduction

- Obtaining annotations, however, remains bottleneck in biomedical image analysis
  - “Standard” tasks may have been solved but...
  - We seek solutions to evermore complicated and specific problems

Outline

- Focus is on
  - Classification / labeling / segmentation
  - General concepts, methodologies, ...
    - So yes, this applies to your deep net...

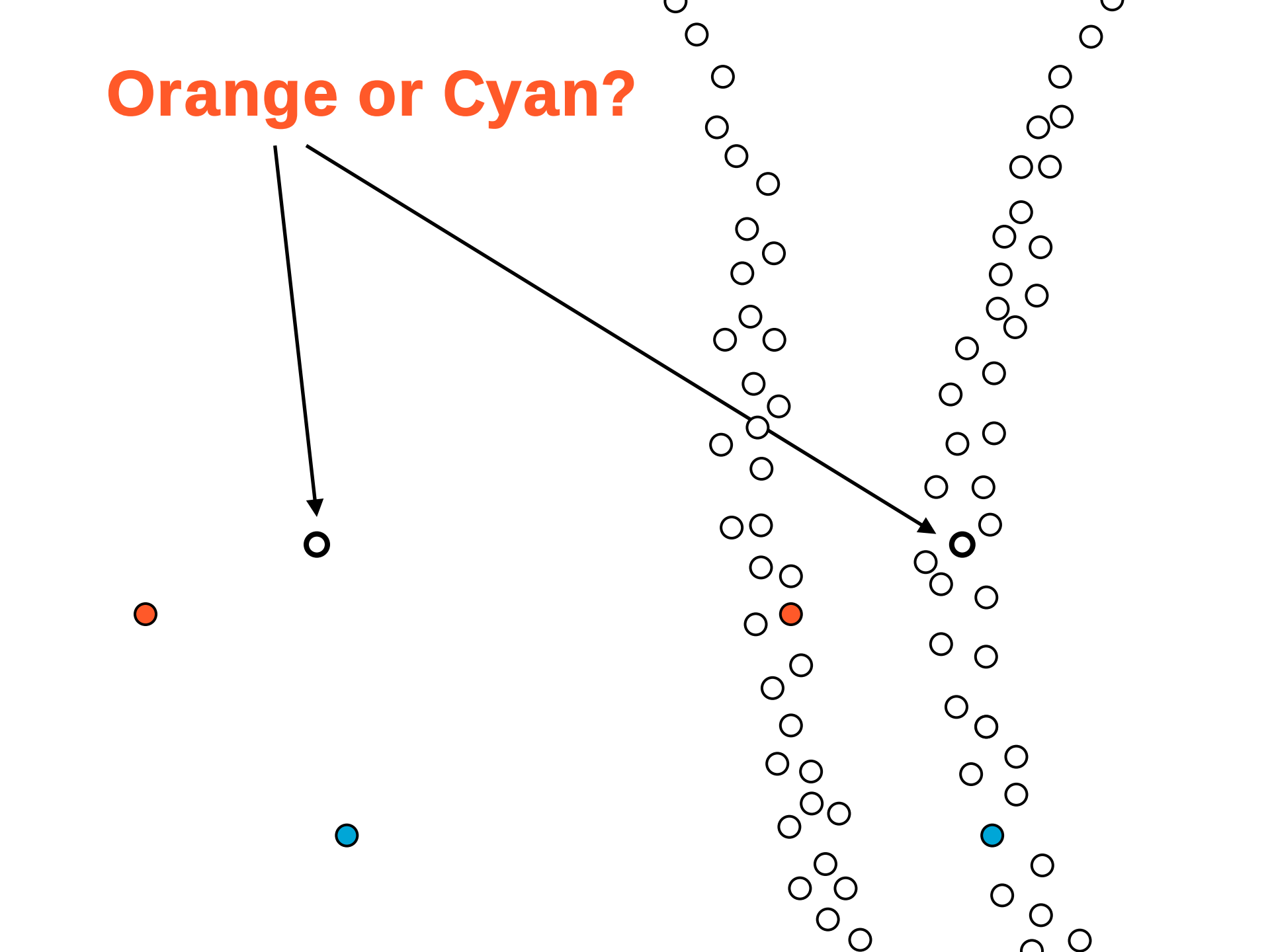
- What can we do?
- SSL, MIL, TL [or DA?], AL
- Cover general ideas and some challenges
- Discussion and conclusions

Reduction of Annotator Burden

- What can we do?
  - Crowdsource your way out!
  - Exploit unlabeled data
  - Handle coarse labels
  - Employ data already labeled
  - Human-in-the-loop
- Obviously, not necessarily mutually exclusive

Reduction of Annotator Burden

- What can we do?
  - Crowdsource your way out! This workshop?
  - Exploit unlabeled data : e.g. semi-supervision
  - Handle coarse labels : e.g. multiple instances
  - Employ data already labeled : e.g. transfer / DA
  - Human-in-the-loop : e.g. [inter]active learning
- Obviously, not necessarily mutually exclusive


Orange or Cyan?

Semi-Supervised Learning

- [Even] in biomedical domain unlabeled data is often abundant
- Natural question : can we exploit these to improve our classifiers? And if so, how?
  - With some labeled data at our disposal...

Semi-Supervision, How To

- Two somewhat distinct lines of thought
  - Make **additional** assumptions about problem
  - Use assumption **intrinsic** in choice of classifier

[1,2,3,4,5]

Additional Assumptions

- Means to link P_X and $P_{Y|X}$
  - X feature vector
  - Y label

- Smoothness assumption / local consistency
- Cluster assumption / global consistency
- Low-density separation
- Somehow exploit prior beliefs...

[1,2]

Alternatively

- Link P_X and $P_{Y|X}$ using constraints **intrinsic** to the classifier, acknowledging that the **choice of classifier** may already provide the necessary assumptions

[3,4,5]

Third Option?

- Self-learning and EM
  - Original idea can be traced back to 1930s
    - So natural, it gets reinvented every now and then [last case is from 2016] : “the PCA of SSL”
    - = self-training
      = self-taught learning
      = self-labeling
      = Yarowsky’s algorithm
      = pseudo-labeling
      = self-supervised learning
      = <your colleague’s method here?>
      = directly related to EM approach

Self-Learning

- Wrapper approach to exploit unlabeled data
- Basic loop
  - Train on labeled examples
  - Classify initially unlabeled examples
  - Include in training set and retrain
- Assumption underlying self-learning / EM?

Self-Learning

- All methods “in practical use” are variations on self-learning / EM
  - More generally, many a method has flavour of self-learning or can be recast in such terms
    - E.g., transductive SVMs, entropy-regularized logistic regression, label propagation, co-training, etc.

Some Theoretical Results in SSL

- Abney, Ben-David, Cozman, Cohen, Lafferty, Wasserman, Singh, Zhu, and many others

- Concerning performance : twofold message
  - Firstly, strong assumptions necessary to provide guarantees about classification performance
  - Secondly, not surprisingly, if assumption are wrong, one could be in trouble
  - N.B. self-learning and EM often do not work!

[6,7,8,9,10]

A Definition of Safety

- One may wonder to what extent SSL can provide techniques that, in some sense, always improve over supervised approaches
  - [Never worse and sometimes better is also OK]

- Such methods are often referred to as **safe** semi-supervised

  - If improvements can be guaranteed for every instantiation we refer to it as **strictly** safe

[3,11,12]

More Theoretical Results

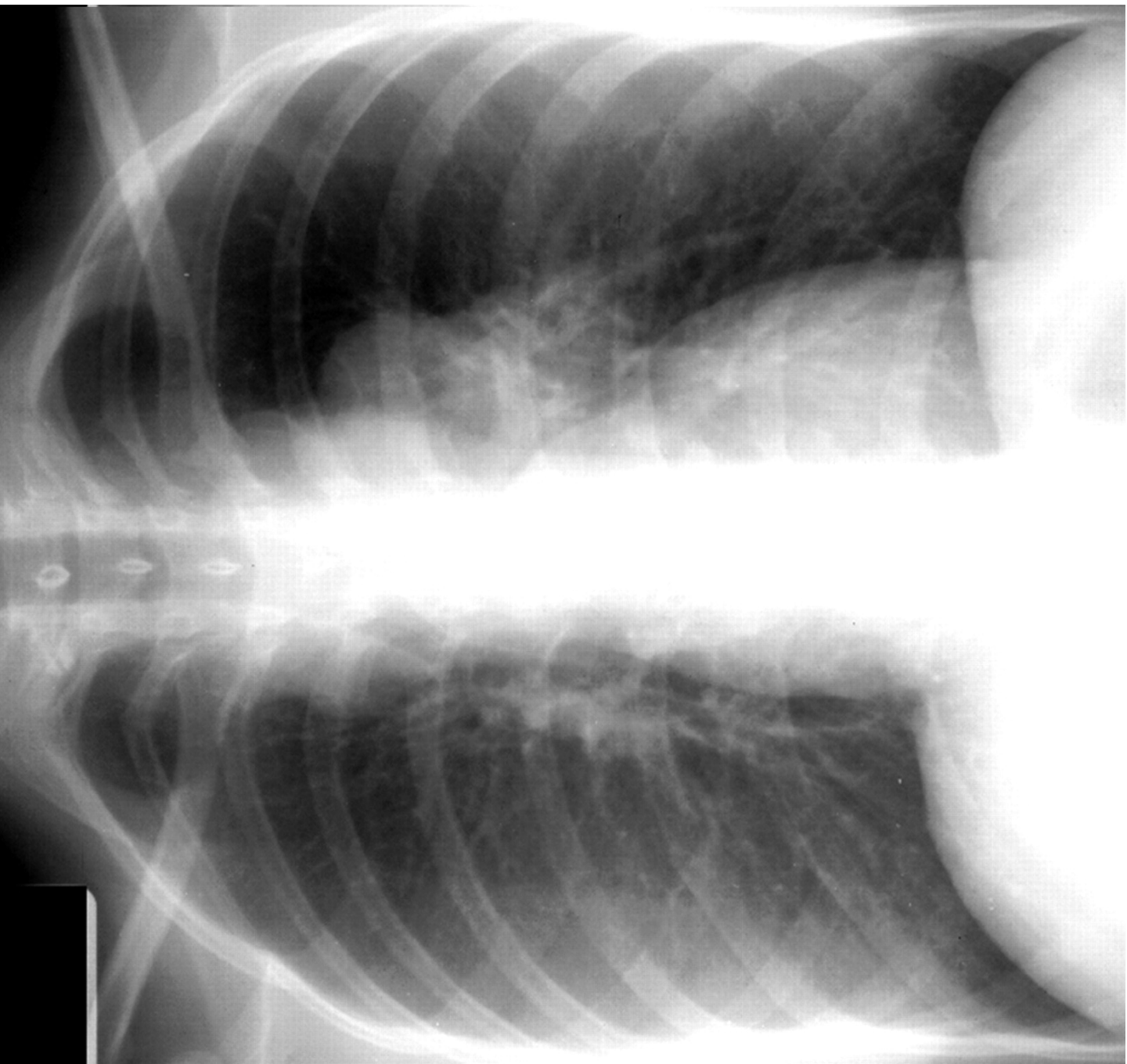
- Recently, strong results that provide guaranteed safety using intrinsic constraints
  - So-called contrastive pessimistic estimation for likelihood and projection estimators for least-squares classifier [or least-squares SVM]
  - More general theory developed, which further clarifies limitations of strictly safe SSL

[13,14,15]

Some Core Challenges

- To what extent is it at all possible to construct safe methods? For which classifiers?
- What kind of safety is practically acceptable? And can we construct methods that fulfil such safety requirements?
- [Additional] assumptions that lead to safe semi-supervision?
  - At least, say, within certain application areas?
  - Additional assumption => cannot always work

From Coarse to Fine?



From Coarse to Fine…

- Coarse labeling much easier than detailed
- Given coarse label, can we trace back its "cause" at finer resolution?
  - E.g. given MR images with and without diagnosis of particular brain disease, can we localize voxels in which the disease manifests itself?
  - Given X-rays with and without TB, can we segment regions with TB?

Multiple Instance Learning

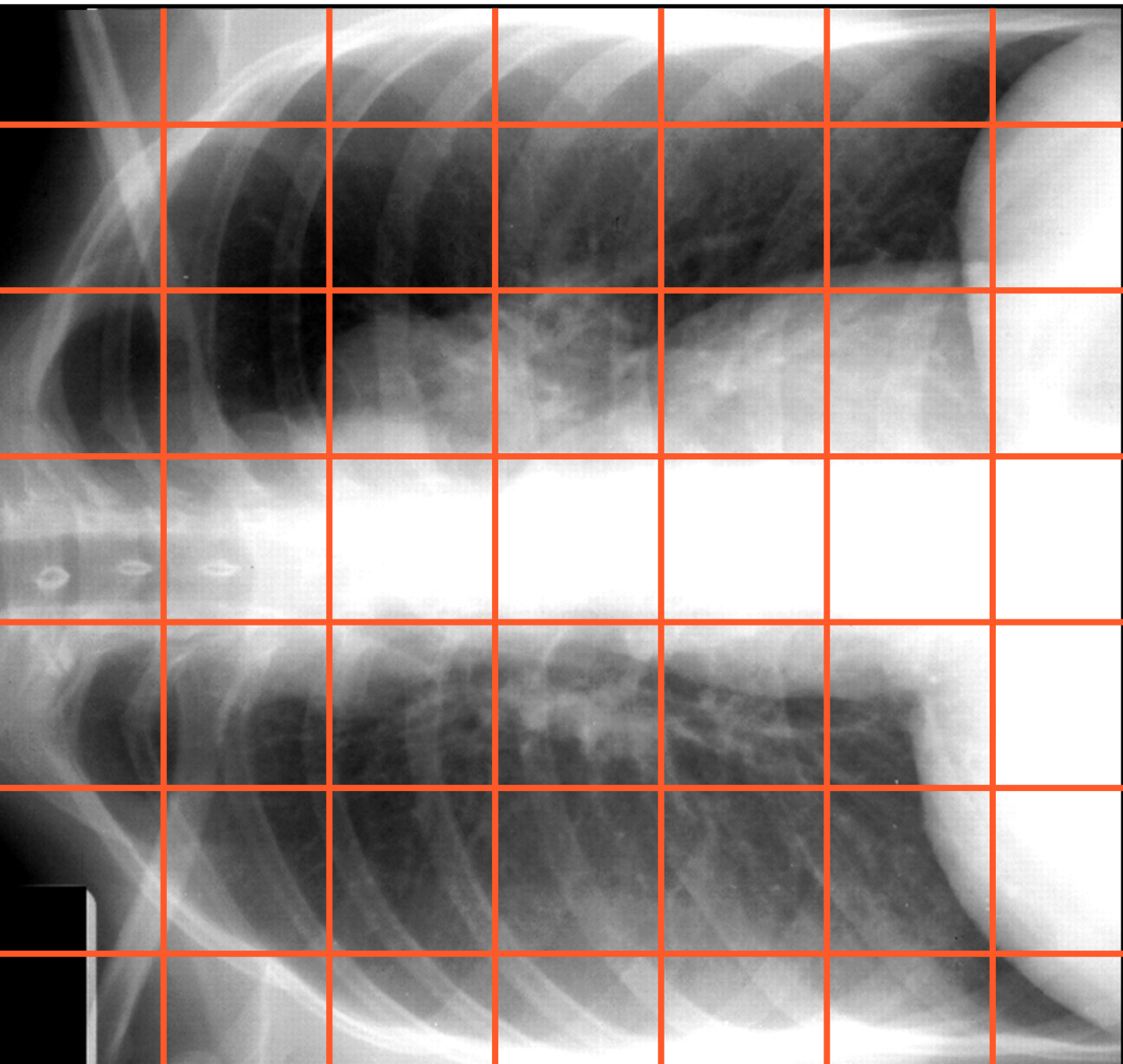
- Represent an object by a collection or [multi-]set of feature vectors, i.e., by means of **multiple instances**

- Sets or so-called bags [of instances] to represent every object, but still a **single label** per object

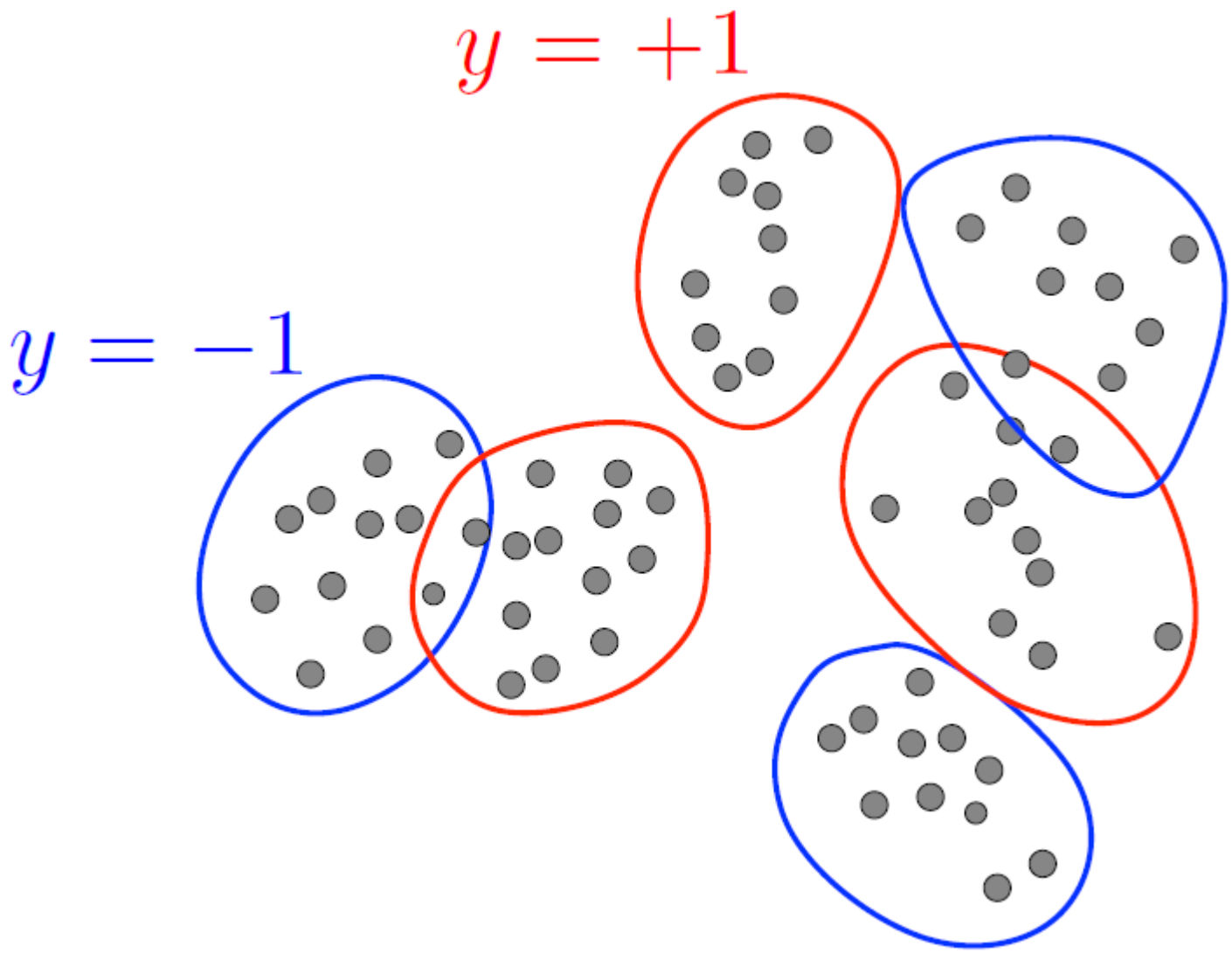
  - Vectors assumed to be in same feature space
  - Set sizes do not have to be equal

[16,17,18]

E.g. a Bag with Instances



Graphically Speaking



Original Goal is Twofold

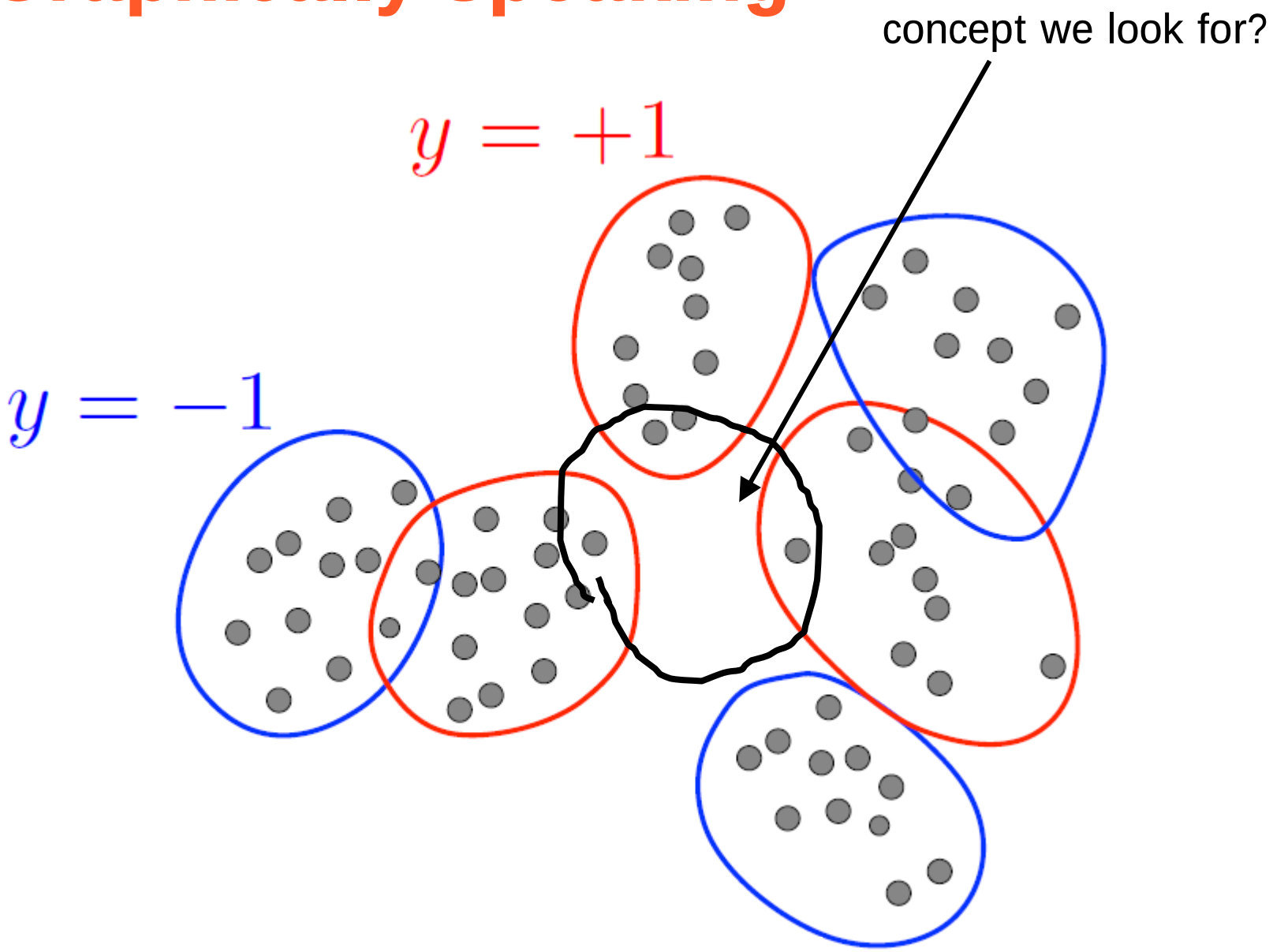
- MIL aims to classify new and previously unseen bags as accurate as possible

- But also : MIL tries to discover a **concept** that determines the positive class
  - Concepts are instances uniquely identifying a class

  - First more easy than second goal
  - Latter is often not considered in MIL literature
  - But latter is needed to go from coarse to fine...

[19,20]

Graphically Speaking



Most Promising MIL Approaches

- MILES and variations
  - Describe a bag by means of similarities to instances
- Dissimilarity-based approaches
  - Describe a bag by means of distances to other bags
- Good for classification, not for coarse to fine
  - Both do not discover something like a concept

[17,21,22]

Some Core Challenges

- Little known about relationship bag labels vs. instance labels : what assumptions are of use?
- To what extent is it at all possible to go to more details? What assumptions are necessary?
- How stable is coarse-to-fine process and how can we stabilize it?

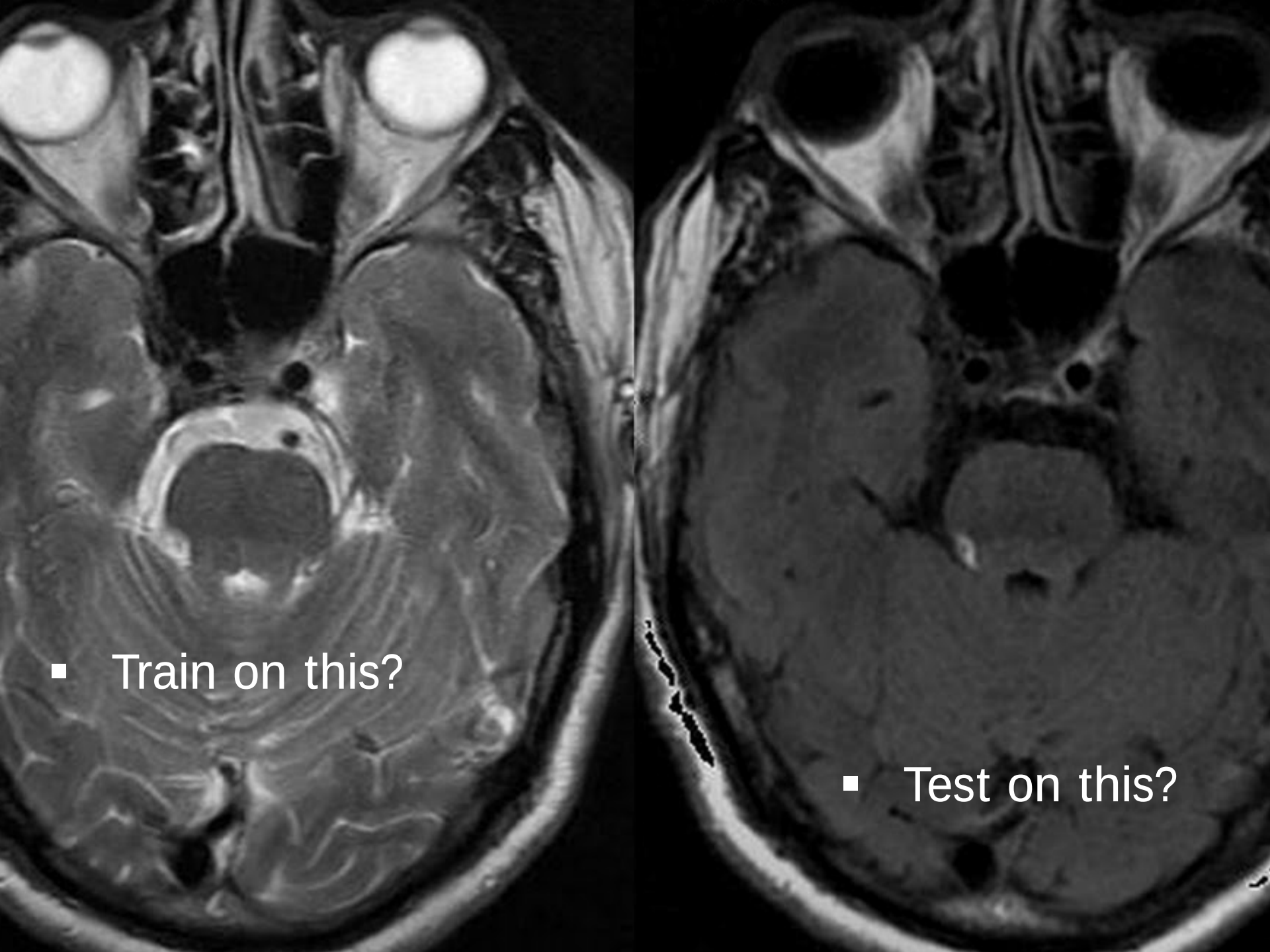
- Overall, little is known at theoretical level...

[20,23,24]

Some More Core Challenges

- Stronger methods needed that really can discover concepts or similarly
- Promising techniques based on convolutional networks [e.g. Somol, Rosmalen, Bazzani, Bency, …]
  - They use clever design of CNN
  - Relations to CRFs, MRFs?

[25,26,27]


Train on this?
Test on this?

Transfer Learning

- Same as / similar to domain adaptation...
- Can we transfer knowledge from one domain to another?
  - More specifically, use labeled data from the one domain to build a classifier in the other?
  - Most studied setting : labeled data from source domain / unlabeled data from a target domain

[28,29]

Transfer and Train and Test

- Would like to deal with...
  - Train and test from different sites
  - Train and test from different machines
  - Train and test from different protocols
  - Train and test from different modalities
  - Train and test from different organs
  - Train and test from different tasks even?
- N.B. Can also be used to undo sampling bias!

Core Challenges?

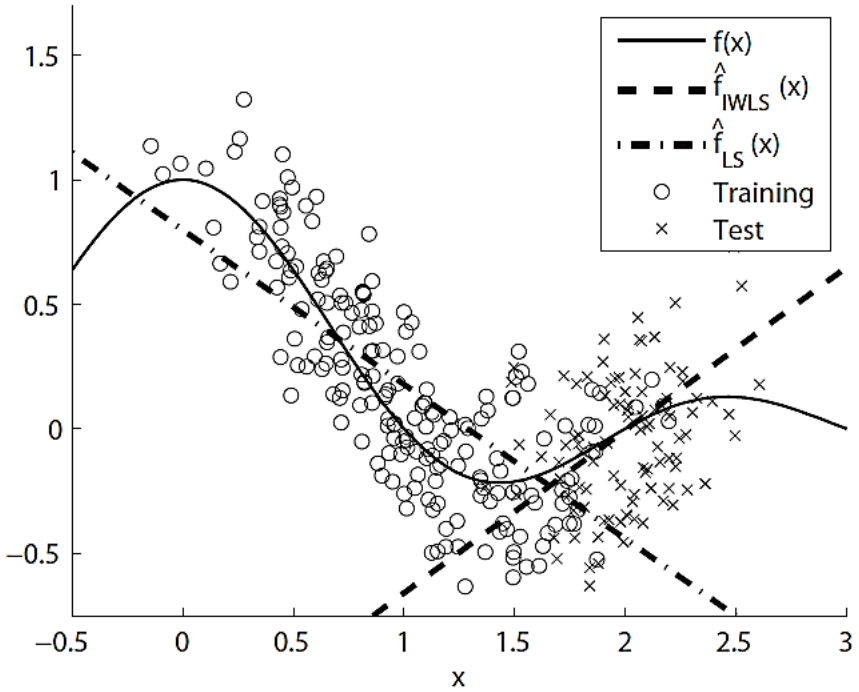
- Order in chaos
  - Many methods, but often unclear what really is solved / what underlying assumptions are
- How deviations from assumptions influence performance?
- Safe? Any guarantees of improvement of target classifier over, say, source classifier?

- General insight in and understanding of many procedures is lacking!

[30,31,32,33]

Reweighting Source Data

- One of the more well-studied TL / DA techniques
  - Especially applicable in learning under covariate shift : $p_S(x) \neq p_T(x)$ but $p_S(y|x) = p_T(y|x)$.
  - Canonical example :



[34,35,36]









Reweighting Source Data

- Especially applicable in learning under covariate shift, where $p_S(x) \neq p_T(x)$ but $p_S(y|x) = p_T(y|x)$
- In this case, reweighting is "justified" because source and target loss can be related through weighting

$$\int \ell(x, y|\theta) p_T(x, y) dx dy = \int \ell(x, y|\theta) w(x) p_S(x, y) dx dy$$

Reweighting Source Data

- Trick is to estimate **importance weights** w
  - Many ad hoc approaches exist
  - Key issue is to control additional variance introduced through [estimated] w
    - See Mohri, Cortes, Joachims, et al.
    - Related bounds and regularization terms derived

$$\int \ell(x, y|\theta) p_T(x, y) dx dy = \int \ell(x, y|\theta) w(x) p_S(x, y) dx dy$$

[37,38]

Main Challenges

- Application of current theory limited, also because of computational issues
- Like in SSL, safety is a concern
  - To what extent is it possible to construct safe methods? For which classifiers?
  - What kind of safety is practically acceptable? How to construct methods that fulfil such requirements?
- Currently no way to properly evaluate system and to set [hyper-]parameters

Active Learning

First : Human[s] in the Loop

- Instead of learning as static problem in which all data is gathered beforehand
  ...we can integrate gathering and learning
  - Now much more is possible [in principle]
  - But dynamics make getting to guarantees, bounds, and rules of thumb even more complicated
- Even simplest setting very difficult to analyse

Question in Active Learning

- Given trained initial classifier
- Say we want to improve its performance by adding one or more instances / feature vectors
- Can do better than pick next sample[s] at random?
  - Can we do it in a systematic way?

[39,40,41,42.43]

Why Active Learning?

- Like for many other human-in-the-loops
  - Suppress cost of annotation
  - Speed up learning
  - Keep amount of training data within bounds
  - Better data is often more useful than simply more data : quality over quantity

Approaches to Active Learning?

- Most solutions [if not all] rely on strong models or are heuristic in nature
- Two basic techniques
  - Use uncertainty
    - Uncertainty sampling
    - Sample points about which classifier is unsure
  - Use disagreement
    - Query by committee
    - Build ensemble and choose points with least agreement

Exploration vs. Exploitation

- Exploitation : make “optimal” decision based on current data available
- Exploration : decide to investigate “suboptimal” options in order to be more optimal in the long run
  - Well-known trade-off in reinforcement learning
  - [Most probably] crucial in active learning as well

Other Active Learning Heuristics

- Expected model change
- Variance reduction / maximization of Fisher information
- Compare labeled density with total density
- Expected error reduction

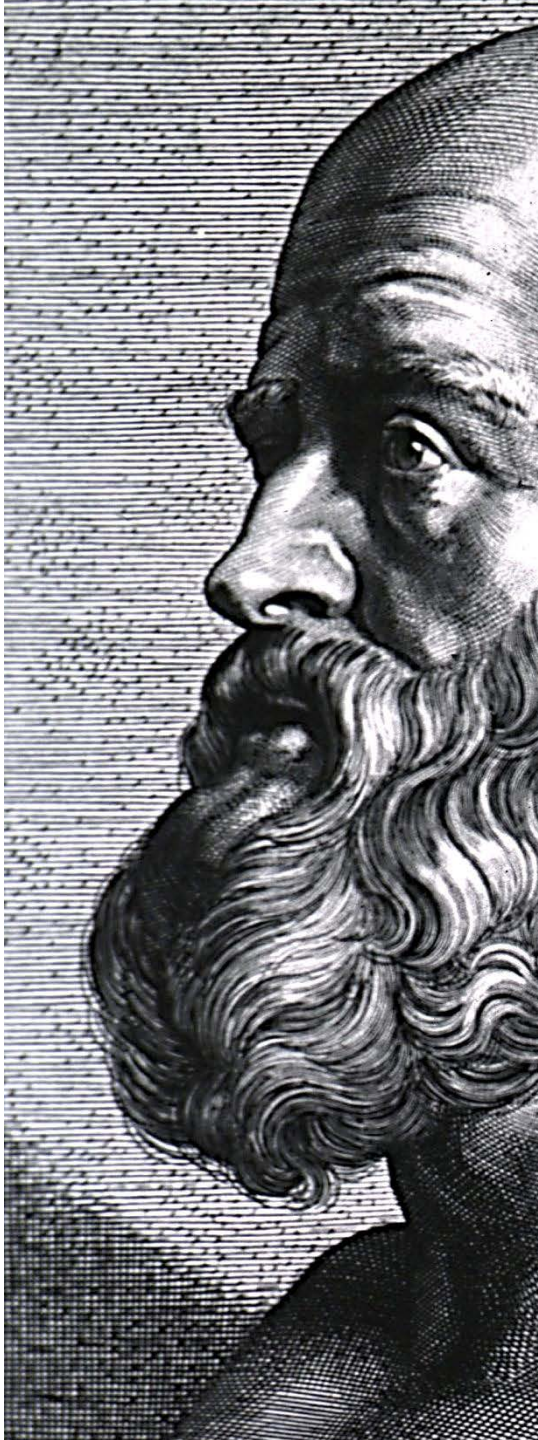
Some Core Challenges

- General insight and understanding is lacking
  - How to really trade off exploration vs. exploitation?
- Dealing with biased sample [transfer learning!]
  - No good procedures to evaluate method
  - No proper way to set hyper-parameters
- Safety is again an issue
  - How useful is AL if it sometimes dramatically fails?
  - Guarantees for minimal performance?
  - What is acceptable in real applications?

[41,44,45,46]

The Obligatory Quote?

- *”Practice two things*
  in your dealings with disease :
  either help
  or do not harm the patient”
  – Hippocrates



Overall Discussion & Conclusions

- Many techniques developed, but yet few in practical [industrial, clinical, etc.] use

- How come?

  - Unknown, unloved?
  - Like in much in PR, ML, CV, IP, and MedIA, most “applied” works are mere proofs of concept with weak validations
  - Little solid theory [experimental / empirical or mathematical] to go on

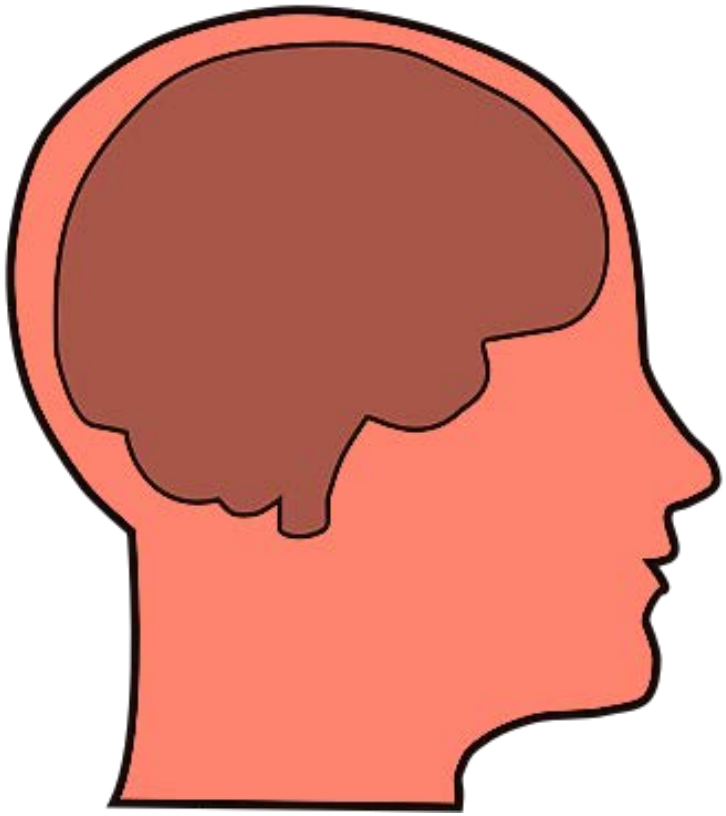
Overall Discussion & Conclusions

- Especially for AL and TL / DA, I would argue strongly for safety [of some sorts]
- Need for methods / theories that
  - Provide “blind” guarantees / plug and play
  - Can perform proper validation
  - Reliably tune hyper-parameters
  - Do not rely on uncheckable assumptions...
- But also good start : perform more extensive validations [or state less exorbitant claims]

Overall Discussion & Conclusions

- Similar views and issues apply to
  - Human in the loop
  - Interactive learning / segmentation
  - Crowdsourcing
- Safe to apply? How to validate and to tune?
  - Only now issues become even more complicated...
- Me pessimistic? Absolutely not!
  - Life is great : plenty of interesting stuff to be done!

SAFETY FIRST



Safety Starts Here

Think Safe...

Work Safe...

Be Safe

Slightly Biased Reference List

- [1] Chapelle, Schölkopf, Zien, “Semi-supervised learning”, MIT press, 2006
- [2] Zhu, “Semi-supervised learning literature survey”, University of Wisconsin, TR 1530, 2008
- [3] Loog, “Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier”, ECML, 2010
- [4] Loog, Jensen, "Constrained Log-Likelihood-Based Semi-supervised Linear Discriminant Analysis", S+SSPR, 2012
- [5] Loog, "Semi-supervised Linear Discriminant Analysis Using Moment Constraints", Partially Supervised Learning, 2012
- [6] Abney, “Understanding the Yarowsky algorithm,” Computational Linguistics, 2004
- [7] Ben-David, Lu, Pal, “Does unlabeled data provably help?”, COLT, 2008
- [8] Cozman, Cohen, “Risks of semi-supervised learning,” in Semi-Supervised Learning. MIT Press, 2006
- [9] Lafferty, Wasserman, “Statistical analysis of semi-supervised regression”, NIPS, 2007
- [10] Singh, Nowak, Zhu, “Unlabeled data: Now it helps, now it doesn't”, NIPS, 2008
- [11] Li, Zhou, "Towards making unlabeled data never hurt", ICML 2011
- [12] Krijthe, Loog, “Implicitly constrained semi-supervised least squares classification,” IDA, 2015
- [13] Loog, “Contrastive Pessimistic Likelihood Estimation for Semi-Supervised Classification”, IEEE TPAMI, 2016
- [14] Krijthe, Loog, “Robust semi-supervised least squares classification by implicit constraints,” PR, 2017
- [15] Krijthe, Loog, “Projected Estimators for Robust Semi-supervised Classification,” ML, 2017
- [16] Amores, “Multiple instance classification: Review, taxonomy and comparative study”, AI, 2013.
- [17] Cheplygina, Tax, Loog, “Multiple instancelearning with bag dissimilarities”, PR, 2015
- [18] Foulds, Frank, “A review of multi-instance learning assumptions”, KER, 2010
- [19] Dietterich, Lathrop, Lozano-Perez, “Solving the multiple instance problem with axis-parallel rectangles”, AI, 1997
- [20] Cheplygina, Sørensen, Tax, de Bruijne, Loog, “Label Stability in Multiple Instance Learning,” MICCAI, 2015
- [21] Chen, Bi, Wang, “MILES: Multiple-instance learning via embedded instance selection”, IEEE TPAMI, 2006
- [22] Tax, Loog, Duin, Cheplygina, Lee, “Bag dissimilarities for multiple instance learning”, SIMBAD, 2011
- [23] Ray, Craven, “Supervised versus multiple instance learning,” ICML, 2005
- [24] Alpydin, Cheplygina, Loog, Tax, “Single-vs. multiple-instance classification,” PR, 2015

Slightly Biased Reference List

- [25] Pevny, Somol, “Using Neural Network Formalism to Solve Multiple-Instance Problems”, arXiv, 2016
- [26] Rosmalen, “Localization Maps: A framework for weakly supervised positive class localization,” MSc thesis, TU Delft, 2016
- [27] Bergamo, et al. “Self-taught object localization with deep networks,” arXiv preprint, 2014
- [28] Pan, Yang, “A survey on transfer learning,”, IEEE TKDE, 2010
- [29] Quionero-Candela, Sugiyama, Schwaighofer, Lawrence, “Dataset shift in machine learning,” The MIT Press, 2009
- [30] Ben-David, Blitzer, Crammer, Kulesza, Pereira, Vaughan, “A theory of learning from different domains,” ML, 2010
- [31] Ben-David, Lu, Pál, “Impossibility theorems for domain adaptation,” AISTATS, 2010
- [32] Ben-David, Urner, “On the hardness of domain adaptation and the utility of unlabeled target samples,” ALT, 2012
- [33] Mansour, Mohri, Rostamizadeh, “Domain adaptation: Learning bounds and algorithms,” COLT, 2009
- [34] Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” J.St.Plan.Inf, 2000
- [35] Cortes, Mohri, “Domain adaptation and sample bias correction theory and algorithm for regression,” Theoretical CS, 2014
- [36] Sugiyama, Krauledat, Müller, “Covariate shift adaptation by importance weighted cross validation,” JMLR, 2007
- [37] Cortes, Mohri, “Domain adaptation and sample bias correction theory and algorithm for regression,” Th.Comp.Sc., 2014
- [38] Swaminathan, Joachims, “The self-normalized estimator for counterfactual learning,” NIPS, 2015
- [39] Cohn, Atlas, Ladner, “Improving generalization with active learning”, ML, 1994
- [40] Cohn, Ghahramani, Jordan, “Active learning with statistical models”, JAIR, 1996
- [41] Settles, “Active Learning Literature Survey”, TR 1648, University of Wisconsin–Madison, 2010
- [42] Tuia, Volpi, et al., “A survey of active learning algorithms for supervised remote sensing image classification”, J-STSP, 2011
- [43] Yang, Loog, “A benchmark and comparison of active learning methods for logistic regression,” arXiv, 2016
- [44] Loog, Yang, “An Empirical Investigation into the Inconsistency of Sequential Active Learning,” ICPR 2016
- [45] Hanneke, “Theory of disagreement-based active learning,” Foundations and Trends in Machine Learning, 2014
- [46] Yang, Hanneke, Carbonell, “A theory of transfer learning with applications to active learning,” ML, 2013
- [47] Kouw, Van der Maaten, Krijthe, Loog, “Feature-Level Domain Adaptation,” JMLR, 2016
- [48] De Bruijne, “Machine learning approaches in medical image analysis: From detection to diagnosis,” MedIA, 2016