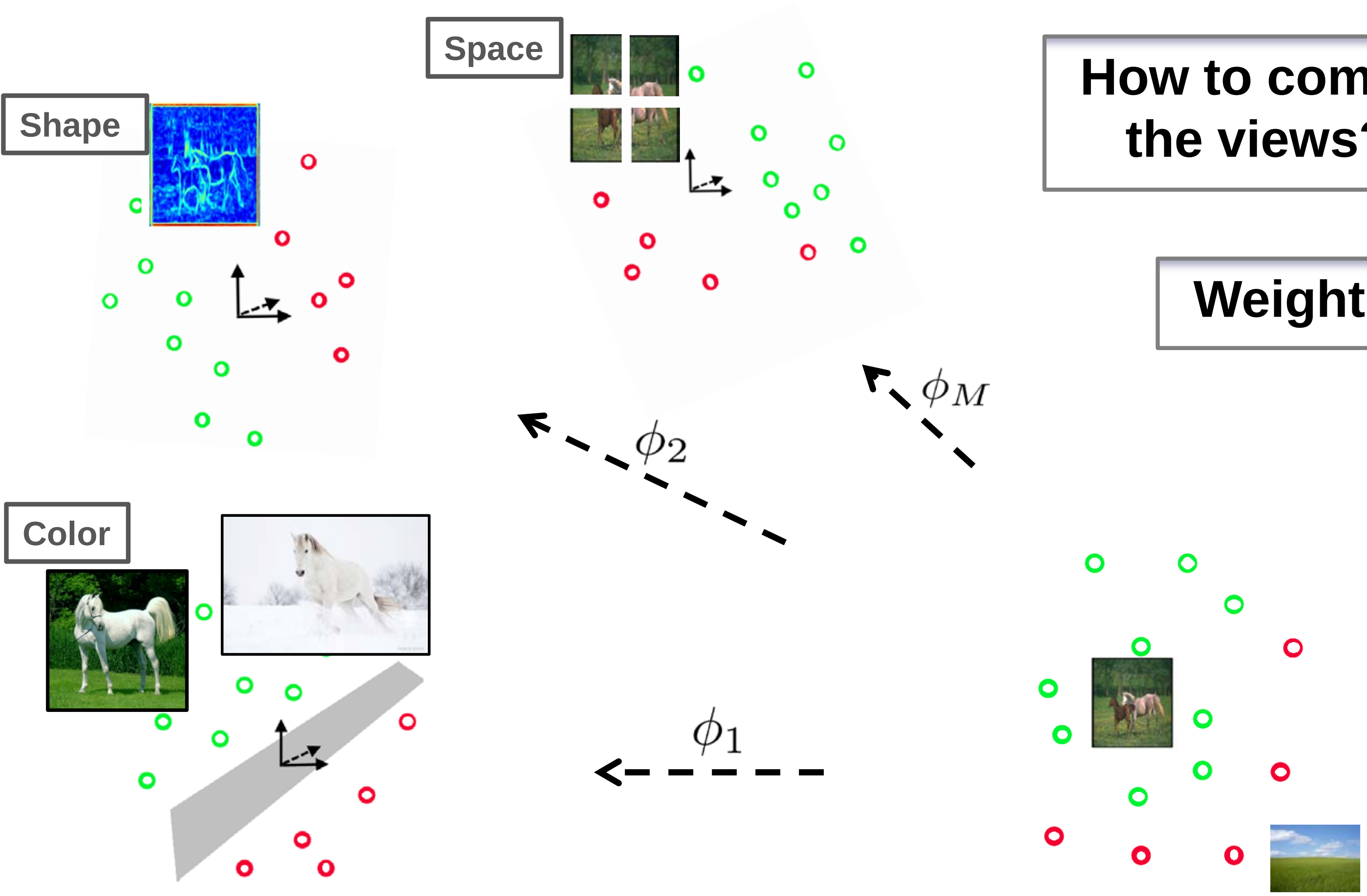
A Quick Glance on Multiple Kernel Learning

Marius Kloft

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Multiple Views / Kernels

(Lanckriet, 2004)

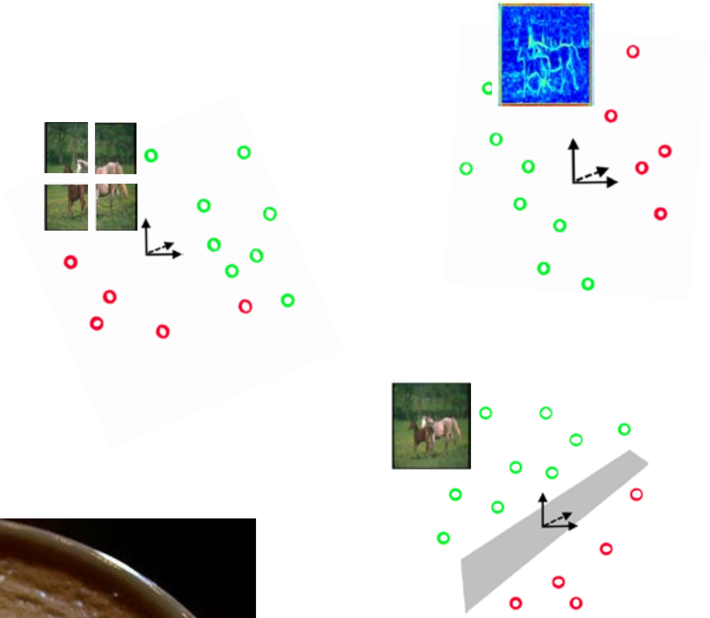


How to combine the views?

Weightings.

Computation of Weights?

- **State of the art** (Bach, 2008)
  - Sparse weights
    - Kernels / views are completely discarded
      - But why discard information?



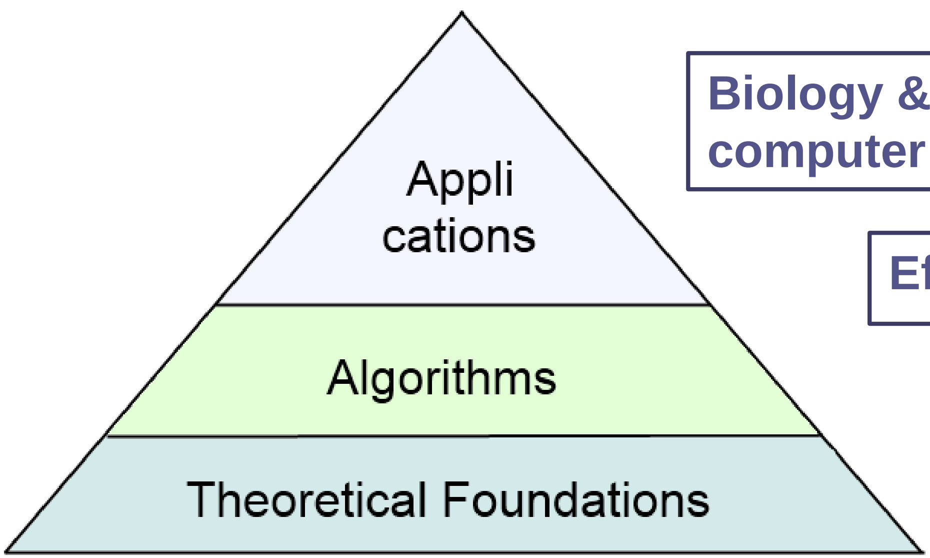
From Vision to Reality?

- **State of the art: sparse method**
  - empirically ineffective

(Gehler et al., Noble et al., Shawe-Taylor et al., *NIPS* 2008, Cortes et al., *ICML* 2009)

- **New methodology**
  - established as a standard

(K., 2011,2012,2013; K. et al., 2009a/b, 2010, 2011, 2012, 2013)



Biology & computer vision

Efficient algorithms

Breakthrough at lear-ning bounds: O(M/n)

Methodology

(K. et al., JMLR 2011)

- **Computation of weights?**
  - Model $f_{\boldsymbol{w},\boxed{\theta}}(x) = \langle \boldsymbol{w}, \phi_{\boxed{k_\theta}(x)} \rangle$
    - Kernel $\boxed{k_{\boldsymbol{\theta}} = \theta_1 k_1 + \cdots + \theta_M k_M}$
  - Mathematical program

$$\inf_{\boldsymbol{w},\boxed{\boldsymbol{\theta}}} \|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} L(f_{\boldsymbol{w},\boxed{\boldsymbol{\theta}}}(x_i), y_i)$$

$$\text{s.t.} \quad \|\boldsymbol{\theta}\|_\psi \leq 1\,, \quad \boldsymbol{\theta} \geq 0$$

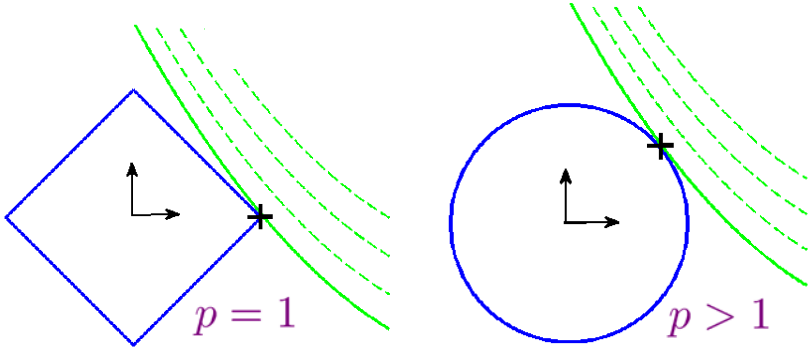
Optimization over weights

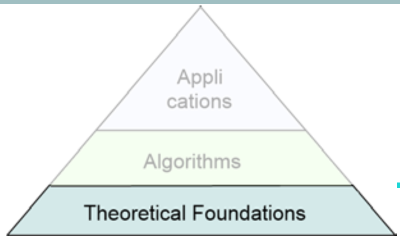
Convex problem.

- **Generalized formulation**
  - arbitrary loss L
  - arbitrary norms $\|\cdot\|_\psi$
    - e.g. ℓ_p-norms:

$$\|\boldsymbol{\theta}\|_p = \left(\sum_{m=1}^{M} |\theta|^p\right)^{\frac{1}{p}}, \; p > 1$$

    - 1-norm leads to sparsity:





Theoretical Analysis

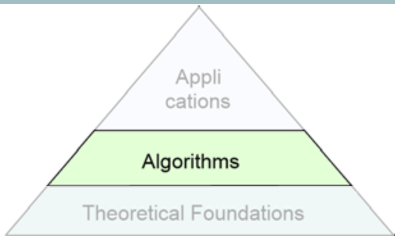
- **Theoretical foundations**
  - Active research topic
    - NIPS workshop 2010
  - We show:
    - **Theorem** (Kloft & Blanchard)**.** The local Rademacher complexity of MKL is bounded by:

- **Corollaries (Learning Bounds)**
  - Upper bound with rate $O(Mn^{-1})$
    - best known rate: $O(\sqrt{Mn^{-1}})$

      (Cortes et al., ICML 2010)
    - Generally $n >> M$
      - for $n = 100\,000\,, \;\; M = 10\,,$ improvement of two orders of magnitude

$$R_r(H_p) \leq \min_{t\in[p,2]} \sqrt{\frac{16}{n}\left\| \left(\sum_{j=1}^{\infty} \min\left(rM^{1-\frac{2}{t^*}}, ceD^2 t^{*2} \lambda_j^{(m)} \right) \right)_{m=1}^{M} \right\|_{\frac{t^*}{2}}} + \frac{\sqrt{BeDM^{\frac{1}{t^*}}}\, t^*}{n}$$

(Kloft & Blanchard, NIPS 2011, JMLR 2012)



Optimization

- **Algorithms** (Kloft et al., JMLR 2011)
  1. Newton method
  2. sequential, quadratically constrained programming with level set projections
  3. block-coordinate descent alg.
     - **Alternate** (Sketch)
       - solve (P) w.r.t. *w*
       - solve (P) w.r.t. $\boldsymbol{\theta}$:

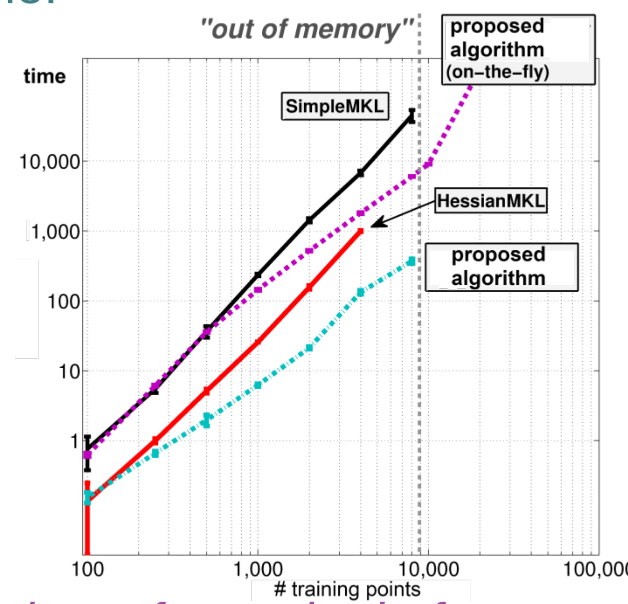
$$\theta_m^* = \frac{\|\boldsymbol{w}_m\|^{\frac{2}{p+1}}}{\sqrt[p]{\sum_i \|\boldsymbol{w}_i\|^{\frac{2p}{p+1}}}}$$

     - **Until** convergence

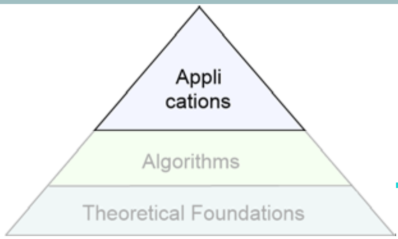
analytical

(proved)

- **Implementation**
  - In C++ (SHOGUN Toolbox)
  - Runtime:



~ 1-2 orders of magnitude faster



Application Domain: **Computer Vision**

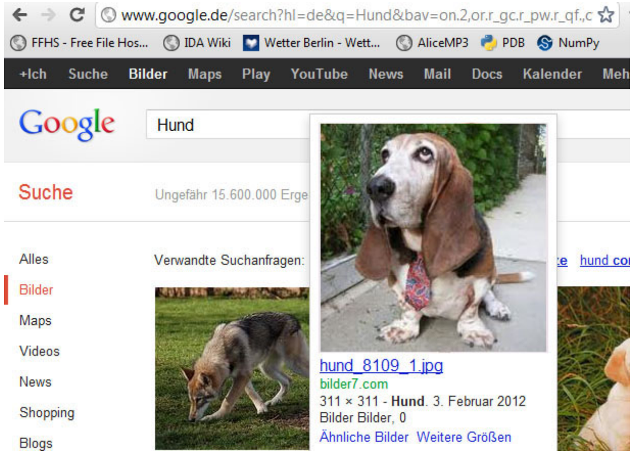
- **Visual object recognition**
  - Aim: annotation of visual media (e.g., images)

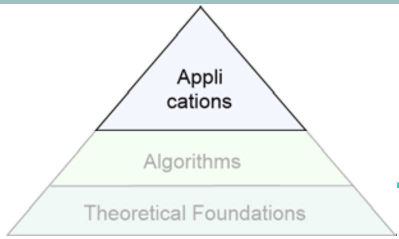
aeroplane

bicycle

bird

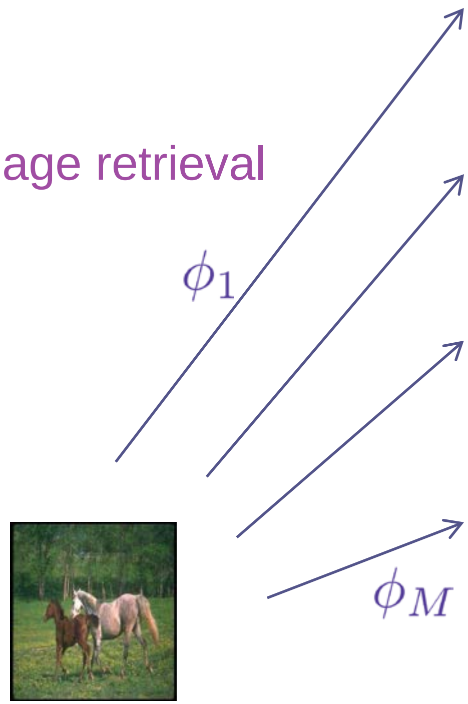
  - Motivation:
    - content-based image retrieval



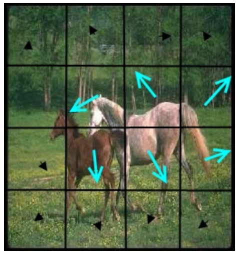


Application Domain: **Computer Vision**

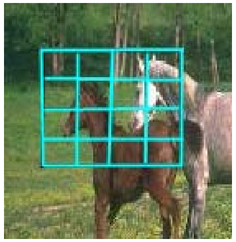
- **Visual object recognition**
  - Aim: annotation of visual media (e.g., images)
  - Motivation:
    - content-based image retrieval



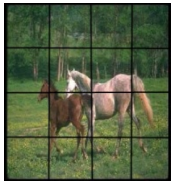
- **Multiple kernels**
  - based on
    - Color histograms
    - shapes (gradients)

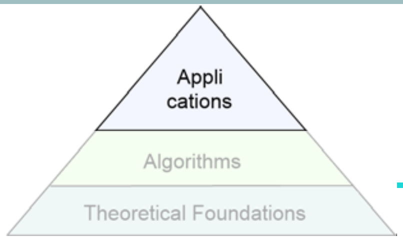


- local features (SIFT words)



- spatial features

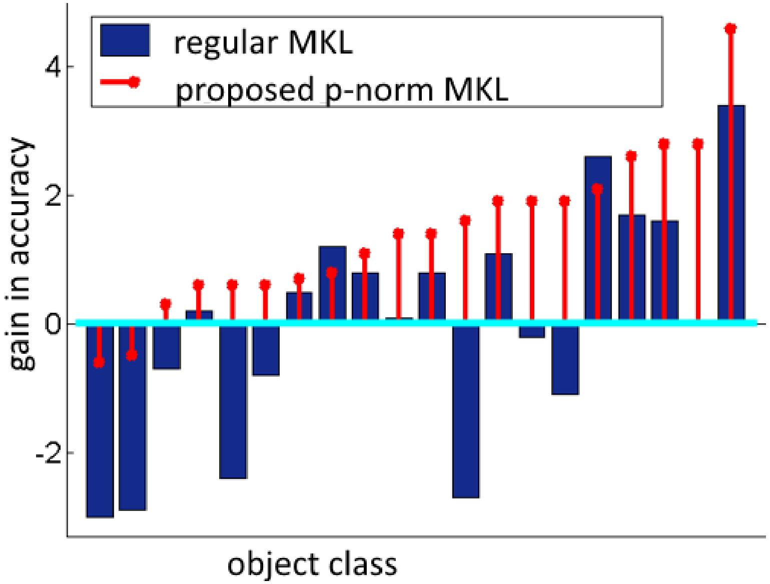


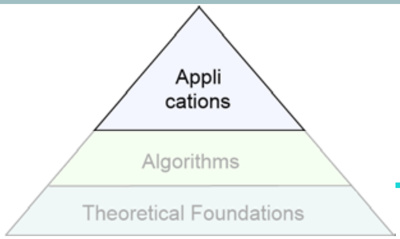


Application Domain: **Computer Vision**

- **Empirical Analysis**
  - PASCAL VOC'08 challenge data
  - Experiments using SHOGUN

Winner: *ImageCLEF 2011 Photo Annotation challenge!*

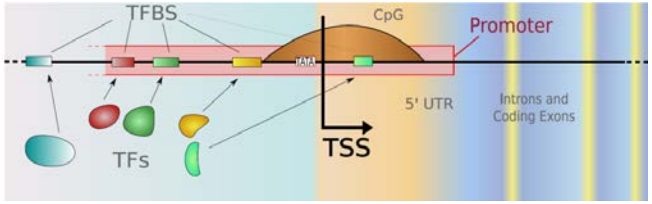




Application Domain: **Genetics**

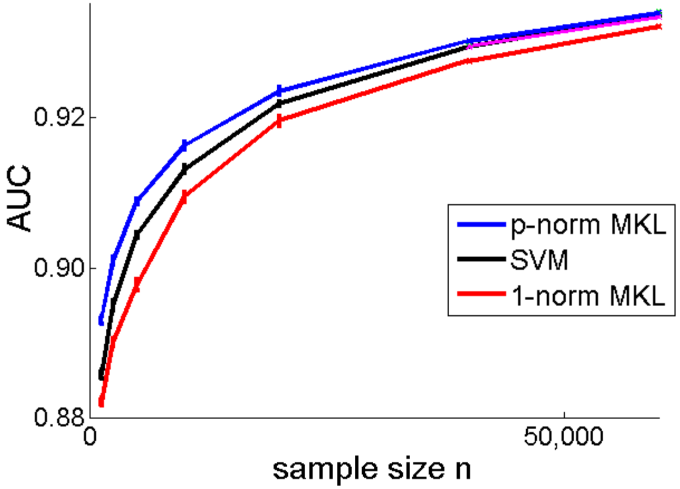
(K. et al., NIPS 2009, JMLR 2011)

- **Detection of**
  - transcription start sites:



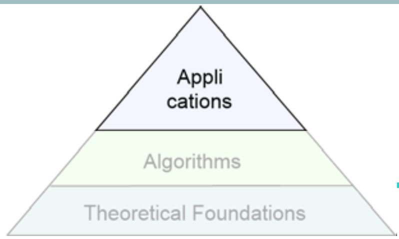
- **by means of kernels based on:**
  - sequence alignments
  - distribution of nukleotides
    - downstream, upstream
  - folding properties
    - binding energies and angles

- **Empirical analysis (SHOGUN)**
  - detection accuracy (AUC):

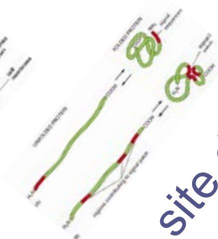


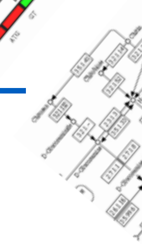
  - higher accuracies than sparse MKL and ARTS
    - ARTS winner of international comparison of 19 models

(Abeel et al., 2009)



Further Applications





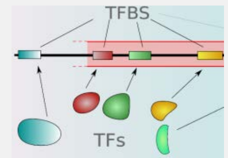
- Computer vision
  - Visual object recognition
- Bioinformatics
  - Subcellular localization
  - Protein fold prediction
  - DNA Transcription splice site detection
  - Metabolic network reconstruction

non-sparsity

$p = 1.11$ $p = 1$ $p = 1.14$ $p = 1.33$ $p = 2$

Conclusion: Non-sparse Multiple Kernel Learning

Computational Biology



More accurate gene detector than winner of int. comparison

Visual Object Recognition

established standard: winner of ImageCLEF 2011 Challenge

Appli-cations

SHOGUN Implementation

Much more efficient than other MKL solvers

Training with > 100,000 data points and > 1 000 Kernels

Sharp learning bounds

Thank you for your attention.

I will be pleased to answer any additional questions.

References

- **Abeel, Van de Peer, Saeys (2009).** *Toward a gold standard for promoter prediction evaluation.* Bioinformatics.
- **Bach (2008).** *Consistency of the Group Lasso and Multiple Kernel Learning.* Journal of Machine Learning Research (JMLR).
- **Kloft, Brefeld, Laskov, and Sonnenburg (2008).** *Non-sparse Multiple Kernel Learning.* NIPS Workshop on Kernel Learning.
- **Kloft, Brefeld, Sonnenburg, Laskov, Müller, and Zien (2009).** *Efficient and Accurate L_p-norm Multiple Kernel Learning.* Advances in Neural Information Processing Systems (NIPS 2009).
- **Kloft, Rückert, and Bartlett (2010).** *A Unifying View of Multiple Kernel Learning.* ECML.
- **Kloft, Blanchard (2011).** *The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning.* Advances in Neural Information Processing Systems (NIPS 2011).
- **Kloft, Brefeld, Sonnenburg, and Zien (2011).** *Lp-Norm Multiple Kernel Learning.* Journal of Machine Learning Research (JMLR), 12(Mar):953-997.
- **Kloft and Blanchard (2012).** *On the Convergence Rate of Lp-norm Multiple Kernel Learning.* Journal of Machine Learning Research (JMLR), 13(Aug):2465-2502.
- **Lanckriet, Cristianini, Bartlett, El Ghaoui, Jordan (2004).** *Learning the Kernel Matrix with Semidefinite Programming.* Journal of Machine Learning Research (JMLR).