ROMIP:
one step forward,
one step aside

http://romip.ru/en/

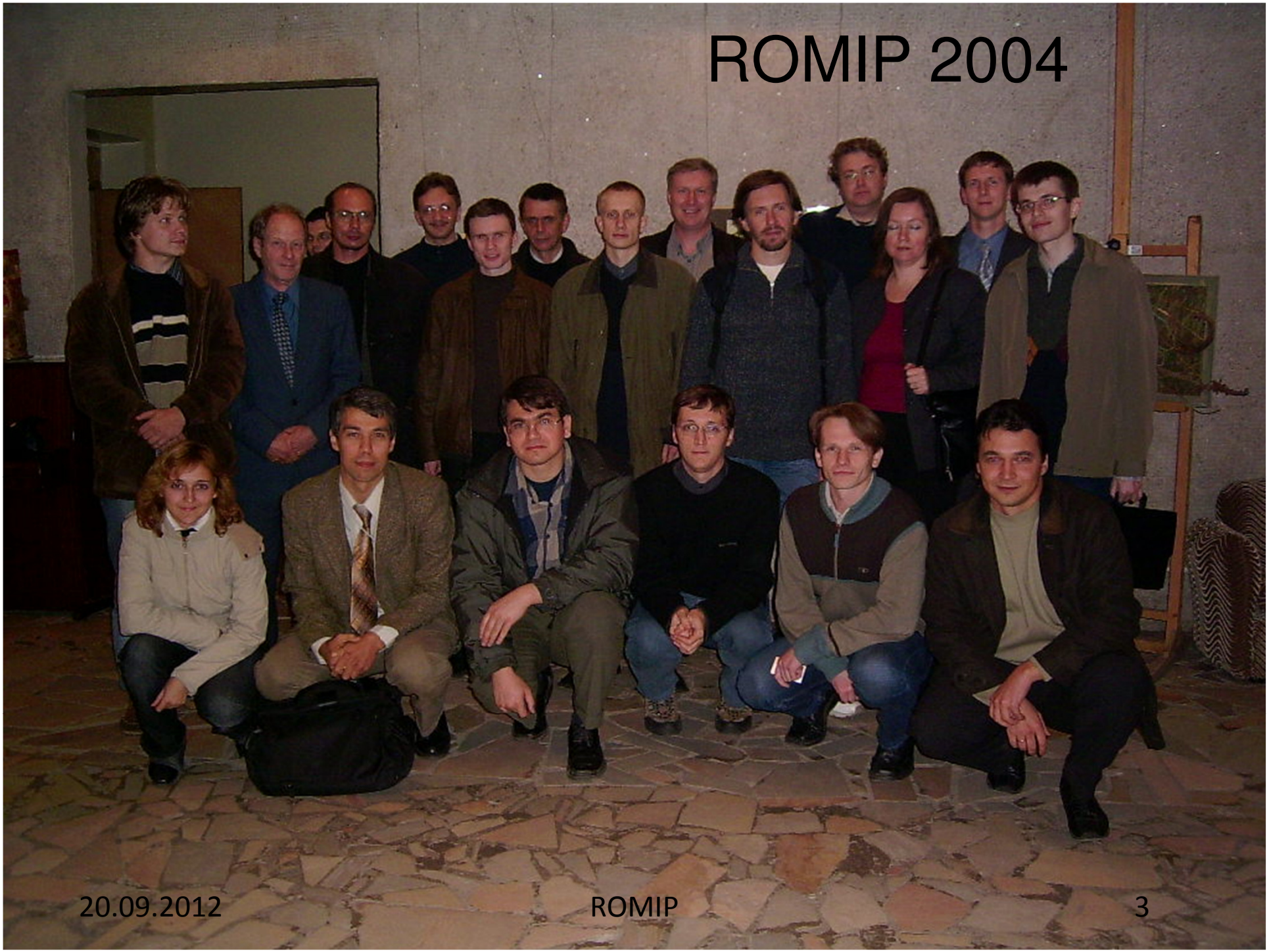
Pavel Braslavski, Ilia Chetviorkin, Maxim Gubin,
Natalia Lukashevich, Igor Nekrestyanov,
Marina Nekrestyanova, Natalia Vassileva

CLEF 2012

ROMIP at a glance

- TREC-like Russian initiative
- Started 2002
- Several **freely available** text and image collections
- 10-15 participating teams each year
- Remote participation + live meeting
- Popular testbed for IR research in Russia
- Related activities: RuSSIR

ROMIP 2004



Largest text collections

Collection	Documents	Size (compressed)	Topics	Evaluated within ad-hoc search track
Legal	~300,000	2 Gb	14,794	220
By.Web	1,524,676	8 Gb	~ 60,000	1 500+
KM.RU	3,010,455	13 Gb	~ 60,000	~250

(Retired) text document tracks

- Ad-hoc text retrieval
- Text categorization
- Snippet generation
- QA and fact extraction
- News clustering
- Search by sample document

Image collections

- Photo collection: 20,000 images from **Flickr**
- Dups collection: 15 hrs video → 37 800 frames
- Panoramic series: 55,000 images (data recycled from Internet Math 2011)

Image tracks

- Content based image retrieval
- Near-duplicate detection
- Image annotation
- Finding panoramic series

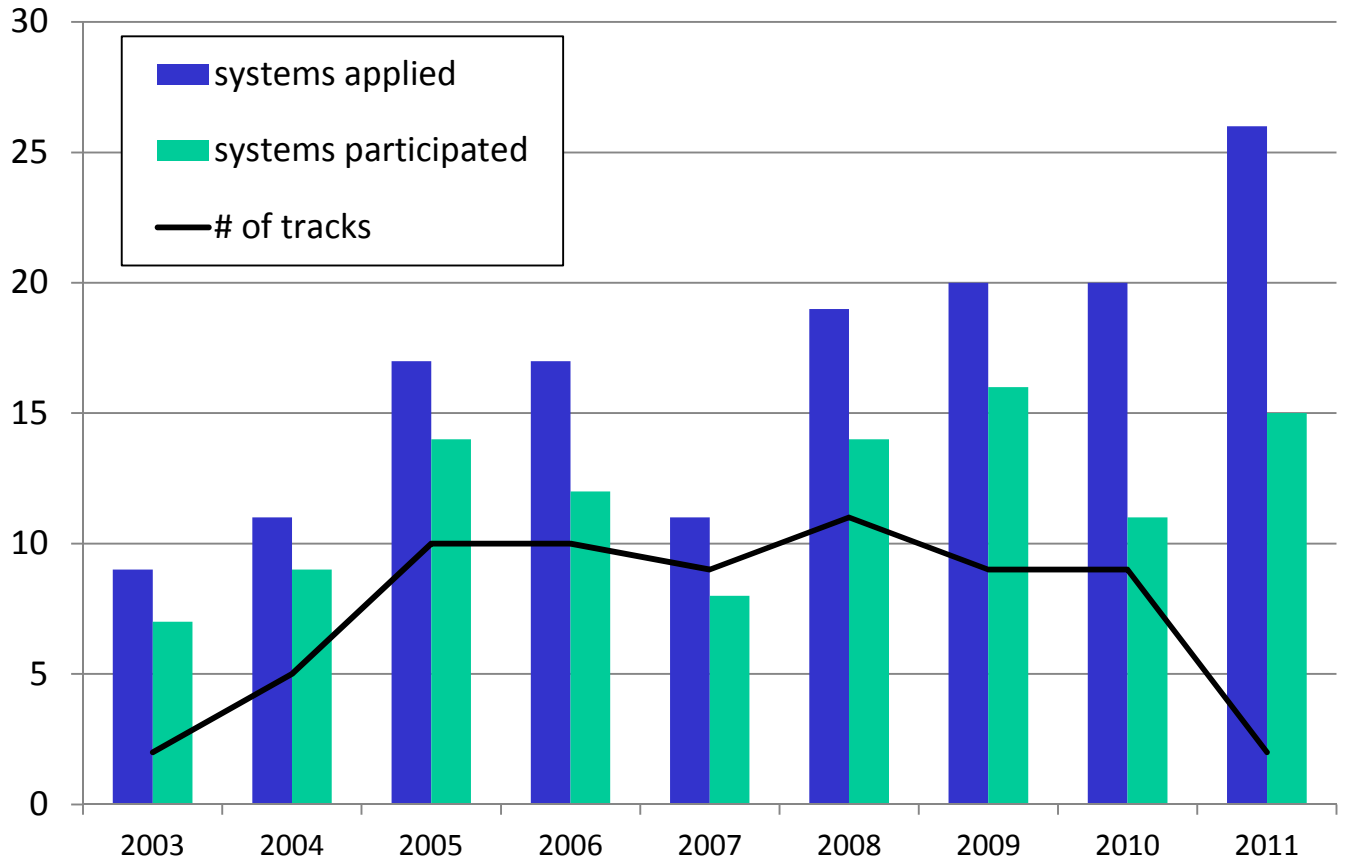
ROMIP by 2011

- Low participation from academia
- Fatigue of classical IR tasks
  - available relevance tables – no need to participate
  - overfitting on available datasets;
  - hard to model realistic settings and data;
  - well-studied tasks – new results are hard to expect.
- Limited resources
- ML challenges (e.g. www.kaggle.com)

ROMIP *light* 2011

- Sentiment analysis
- Search by query image (low participation ☹)
- Schedule shifted to fall

ROMIP timeline



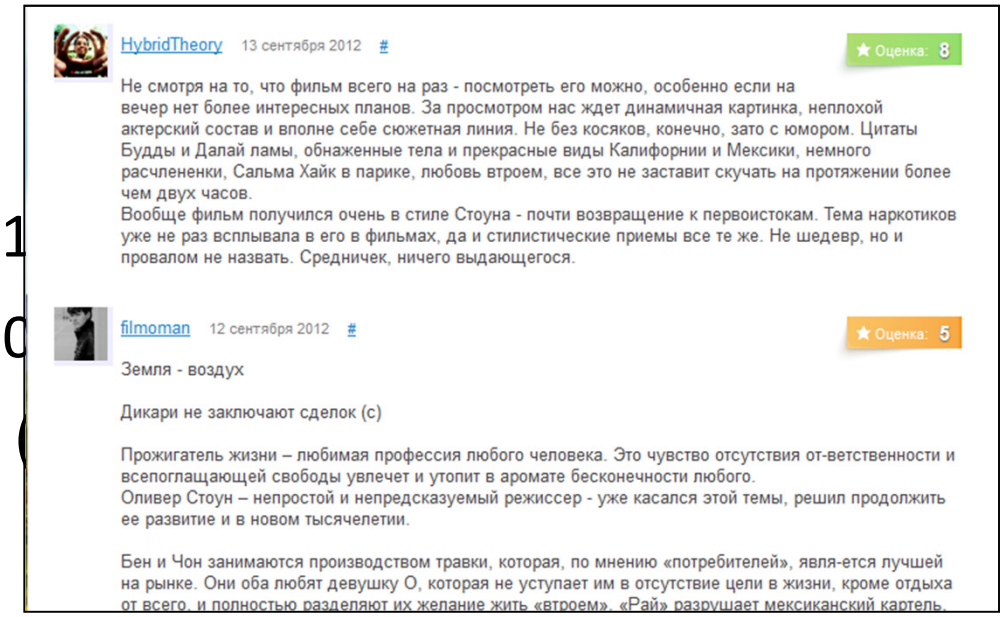
Sentiment analysis (SA)

- Three domains: *movies*, *books*, and *digital cameras*
- 'Transfer learning' (data from different sources)
- Classification into 2, 3, and 5 classes
- 23 teams registered→ 12 submitted results → 6 reports published
- 2-class: 105 runs, 3-class: 81 runs, 5-class: 30 runs

SA: data

Training set

- 15,000+ movie reviews (1
- 24,000+ book reviews (10
- 10,000+ camera reviews (



Test set

blog posts collected via blog search w. subsequent filtering

- 275 posts on movies
- 329 posts on books
- 270 posts on digital cameras

Plans

- New edition of SA track
  - Finer granularity (sentence)
  - Opinions for a given entity
- Re-launch of image tracks (in cooperation with Graphicon conference)
- Machine translation track

MT evaluation track (2012)

- Strong industrial players
- 1M parallel sentences (Ru-En) to release
- Collaboration with TAUS Labs
- Metrics
  - BLEU
  - human assessment

Thank you!
Questions?

Pavel Braslavski

pb@kontur.ru