# Information Access Challenges in the Blogspace

**Gilad Mishne**

UNIVERSITEIT VAN AMSTERDAM

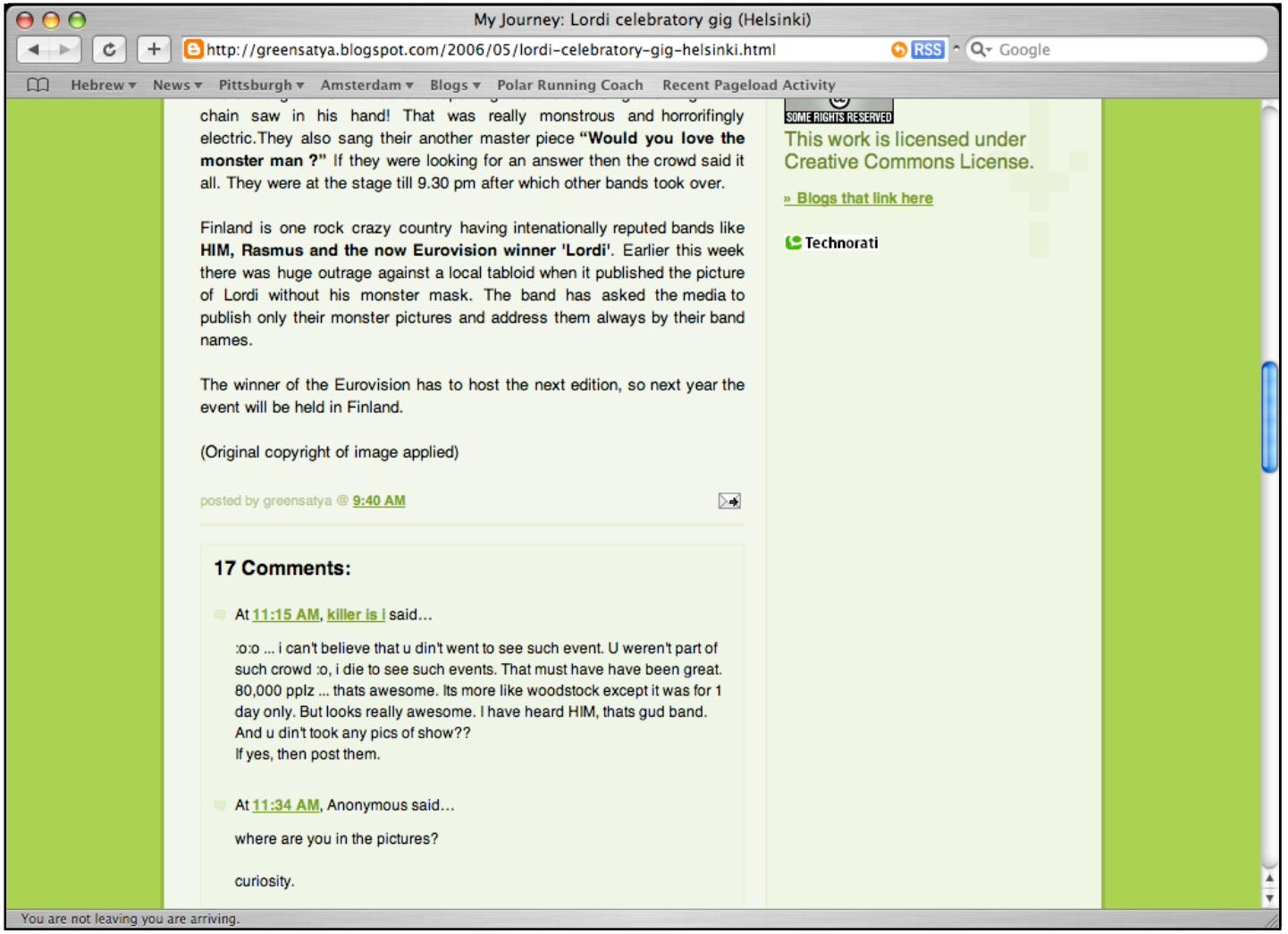
# Overview

- Blogs 101
- Properties of the blogspace
  - From an information access point of view
- Application test cases
  - http://moodviews.com
  - Profile-based product and ad matching
- Challenges and ongoing efforts

# Web logs

- In a nutshell: online diaries
- More formally, web pages whi
  - Include periodic, time-stamped entri
  - Are ordered in reverse timeline
  - Are published through a CMS
  - Are authored by a single person or a
  - Contain commentary about the blog web-related issues, …
  - Regularly updated
  - Allow visitor feedback
- Blogspace/blogosphere:
  - The totality of blogs

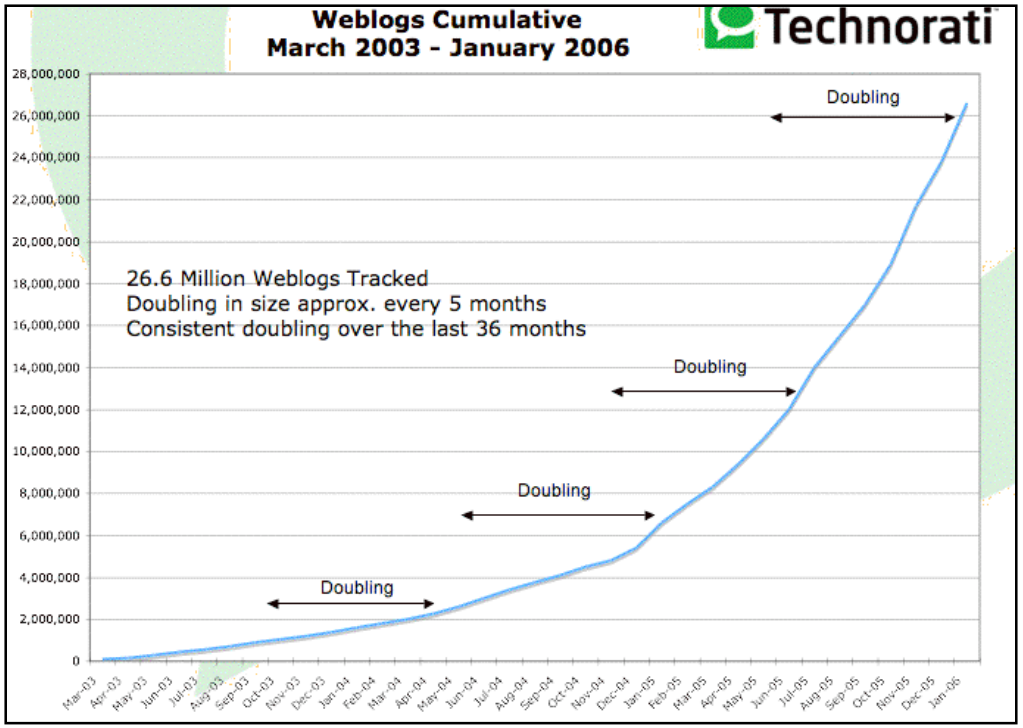
# Blog examples



# Blog examples

# The blogspace as a corpus

- Differences from other web corpora
  - Content
  - Structure
  - Timeline
  - Growth

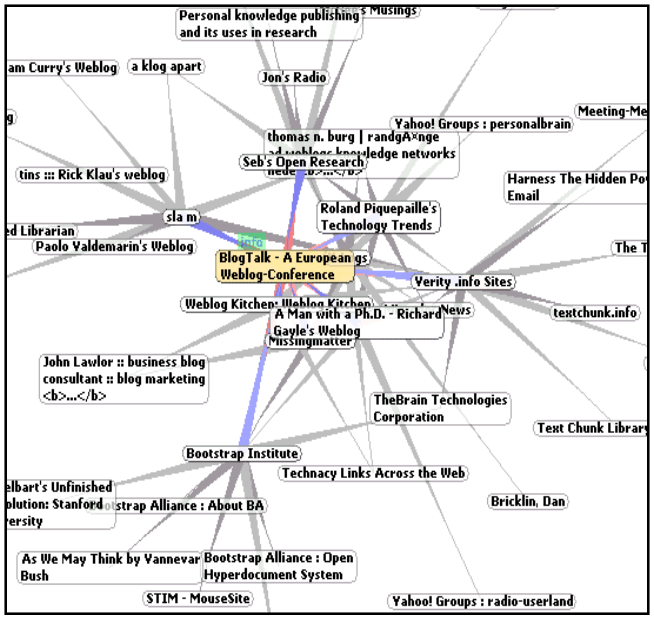


# The blogspace vs. the web

| | Blogs | Web pages |
|---|---|---|
| Content type | personal, diary-like, commentary, observations, sentiments, moods | Anything, really (but in most cases - low on sentiment content) |
| Content format | Template | Custom |
| Links | Dense; links change frequently; different types | Links usually static |
| Timeline | Frequently updated; reverse-chronologically-sorted | Usually static; time does not play major role |
| Represents | A person's life | Information |

# Structure of the blogspace

- The World Wide Web:
  - A scale-free network
- The blogspace:
  - A scale-free (scale-weak?)
  - and small-world network
- What does this mean?
  - There are a few influential blogs, and a “long tail”
  - Information diffuses quickly
  - There are lots of communities
  - A good model of a social network (better than the web)
    - Not surprising – this *is* a network of people

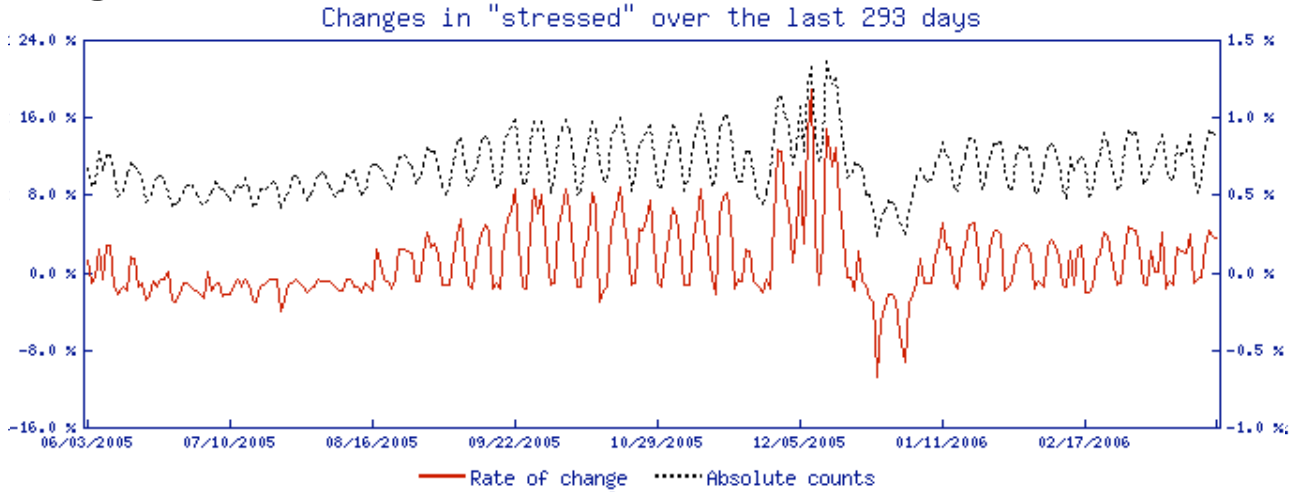


# Overview

- ✓ Blogs 101
- ✓ Properties of the blogspace
  - ✓ From an information access point of view
- **Application test cases**
  - **http://moodviews.com**
  - **Profile-based product and ad matching**
- **Challenges and ongoing activities**

# Timeline-oriented sentiments

- **Many bloggers indicate their mood at the time of writing**

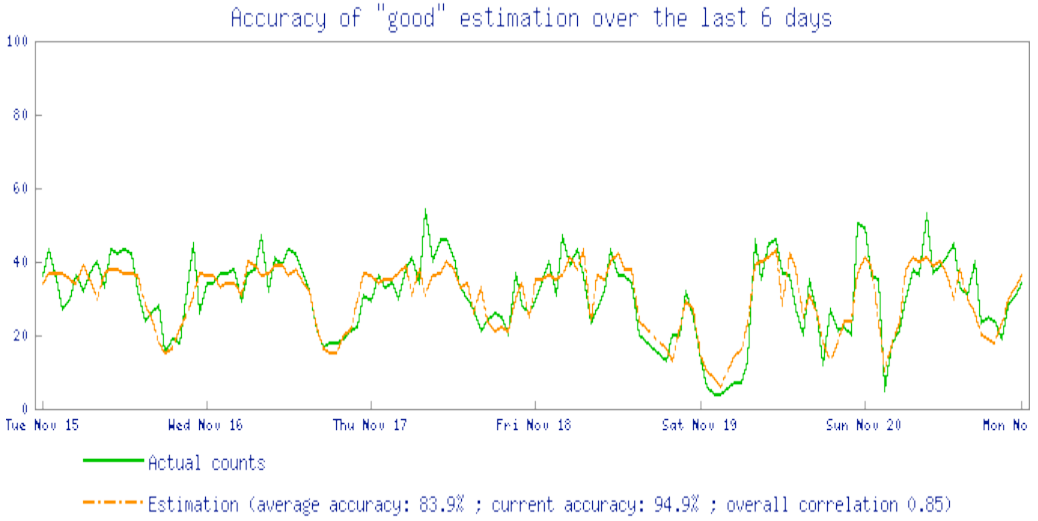



- **Questions:**
  - **Can we predict the mood using the bloggers' text?**
  - **Can we explain changes in the "global mood"?**
  - **…**

# http://www.moodviews.com

- **Mood level prediction:**
  - **Language model comparison identifies "distinctive mood-related words"**
  - **"Mood recipe" calculated for each mood**

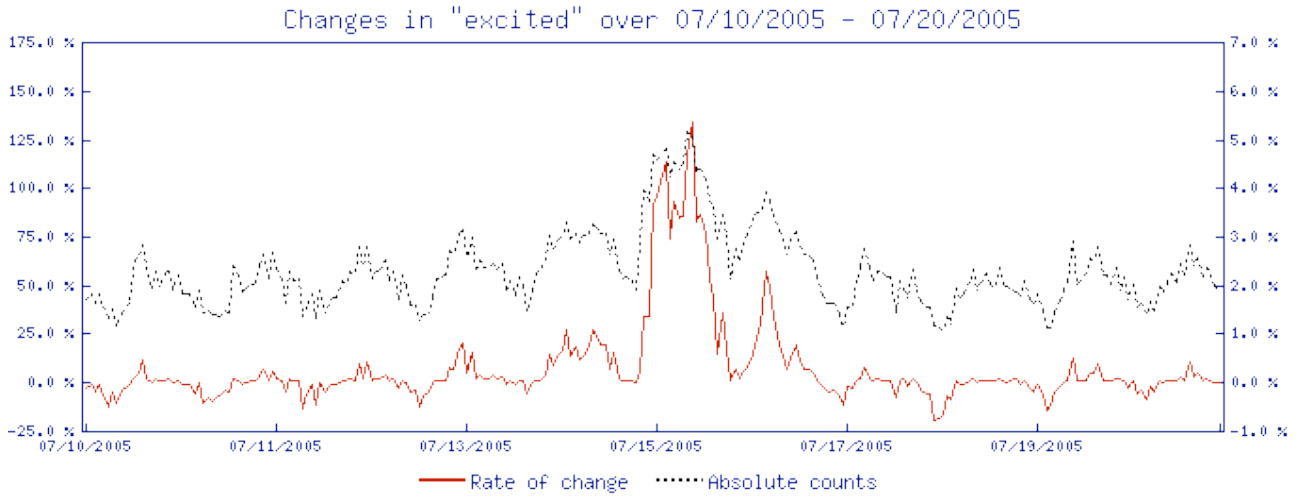
```
Happy =
      23.345 +
       0.0318 * total-posts +
      -2.4026 * count(always) +
    -114.9877 * count(day) +
      16.2727 * count(excited) +
      55.3942 * count(finally) +
     129.2576 * count(happy) +
     223.8079 * count(home) +
    -246.8737 * count(know) +
     506.9564 * count(lol) +
       5.7815 * count(thoughtful) +
     -88.1313 * count(will be)
```



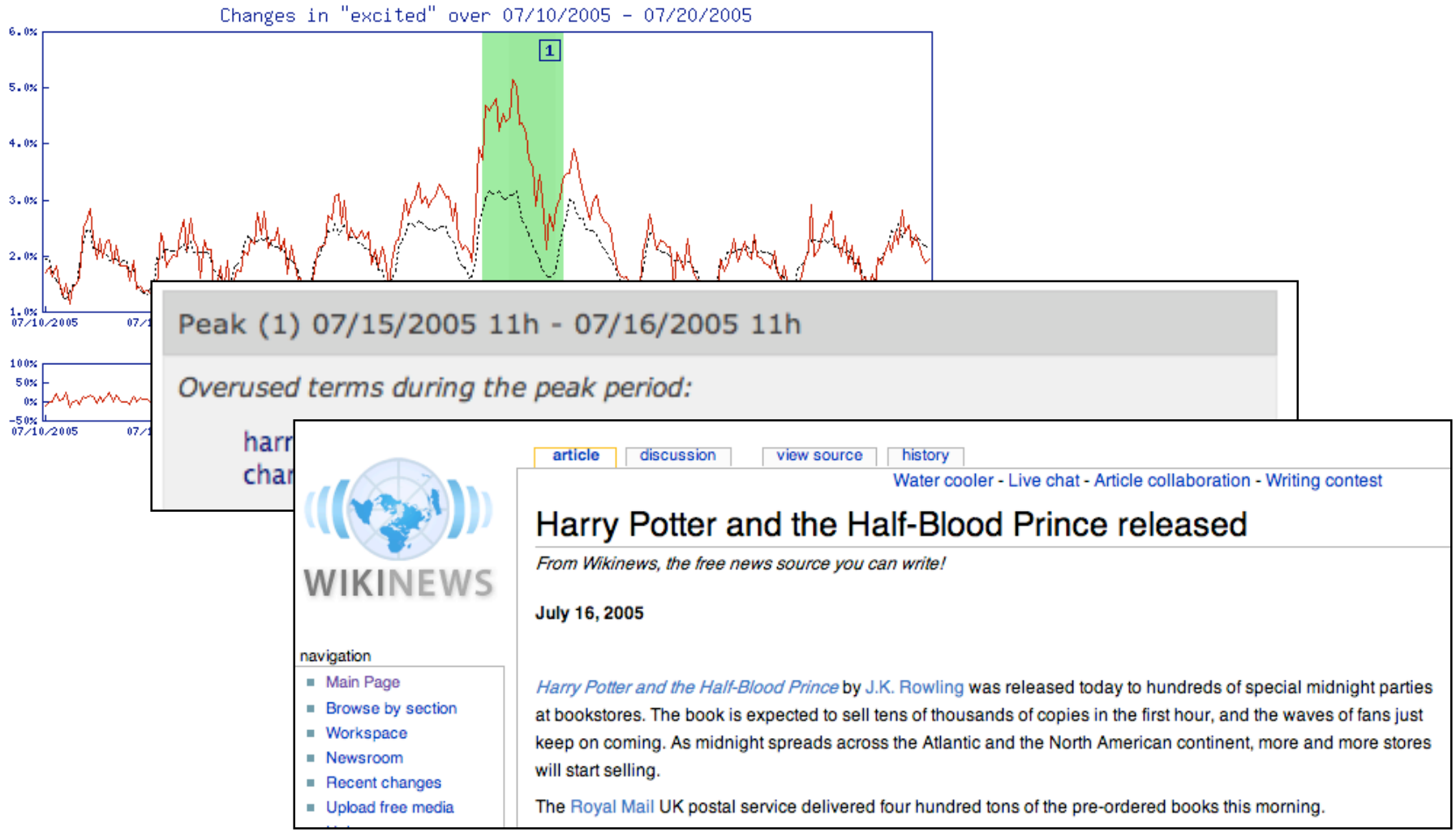

- **Prediction accuracy: 0.83 correlation**
  - **Up to 0.95 on some moods**

# http://www.moodviews.com

- **Explaining irregular moods:**
  - **Identify "spike"**
  - **Language model during spike compared to "expected" model**
  - **Distinctive words used as a query to an index of events**



# Moodsignals in action



# Blogger profiling

- Lots of work on collecting knowledge from the blogspace as a collective
  - Sales predictions, political reflections, …
- What can be mined from *individual* blogs?
- Intuition: a blog can be used to create a profile of the human behind it
- Textual profiles of blogs can be derived using
  - Keyword extraction
  - Summarization
  - your-favorite-method-here

# Blogger profiles in action 1

- Task: match blogger with products
  - In particular, books she is likely to appreciate

1. Create textual profile of blog
   - Use corpus-comparison techniques

2. Identify typical products related to profile
   - Use a large DB of products - Amazon

- Evaluation:
  - Compare the derived categories to those present in an explicit list of "desired products" by the blogger
  - Performs 121% better than simply trying to locate references to products in the blog

# Blogger profiles in action 2

- Task: contextual advertising in blog posts
  - More difficult than contextual advertising on other web pages, since blog posts can be very un-topical

# Blogger profiles in action 2

- Task: contextual advertising in blog posts
  - More difficult than contextual advertising on other web pages, since blog posts can be very un-topical

1. Extract textual profile of blog post components
   - Post, blog, community, comments, time, tags, similar posts

2. Combine models; distill single profile of post; match with ads by similarity

- Evaluation:
  - Used a corpus of blogs/ad used by the largest Dutch blogging platform
  - Human assessors
  - Outperform state-of-the-art in contextual advertising by 20%

# Main inf. access challenges

- In one sentence
  - Identify how the intrinsic properties of the blogspace can be used to access the information in it
    - Subjective, personal content
    - Temporal profile
    - Social-network structure

- In particular
  - Modeling bloggers and communities
  - Catering for different search requirements
  - Data quality
  - <the-next-big-thing>

# Questions?

UNIVERSITEIT VAN AMSTERDAM