Ubiquitous Machine Learning

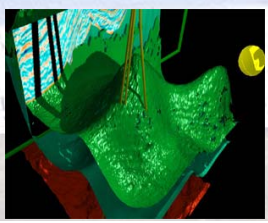
Fraunhofer IAIS
Institut
Intelligente Analyse- und
Informationssysteme

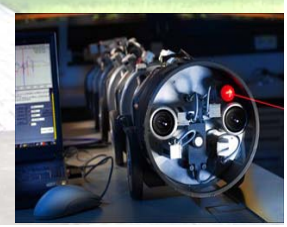
universitätbonn
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT

Prof. Dr. Stefan Wrobel

Fraunhofer IAIS: Intelligent Analysis and Information Systems

- 230 people: scientists, project engineers, technical and administrative staff, students
- Located on Fraunhofer Campus Schloss Birlinghoven/Bonn
- Joint research groups and cooperation with





Core research areas:

- Machine learning and adaptive systems
- Data Mining and Business Intelligence
- Automated media analysis
- Interactive access and exploration
- Autonomous systems

Directors: T. Christaller, S. Wrobel (exec.)

Brainyquote.com

Learning is not attained by chance, it must be sought for with ardor and diligence.

Abigail Adams

Abigail Adams (November 11, 1744 – October 28, 1818)

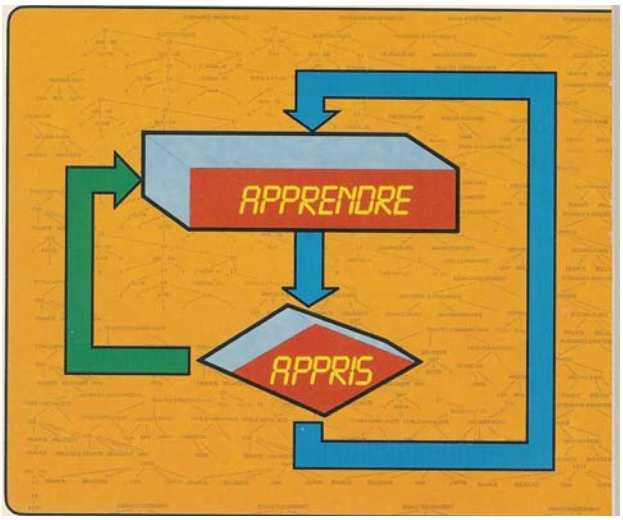
First Lady, wife of John Adams, 2nd President of the United States

Wikipedia

Outline

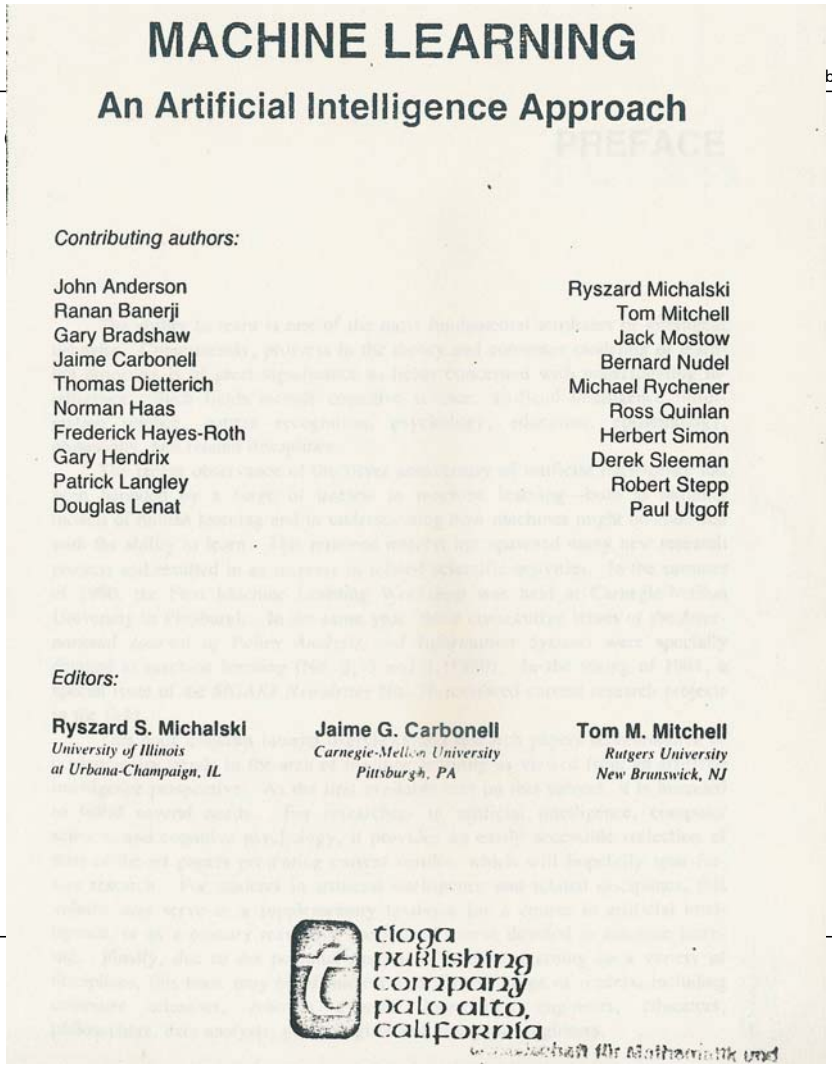
- The beginnings
- Important Trends
- The Need for Machine Learning
- Ubiquitous Learning
- Conclusion

1986: machine learning is starting



Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

We also had



MACHINE LEARNING

An Artificial Intelligence Approach

PREFACE

Contributing authors:

John Anderson
Ranan Banerji
Gary Bradshaw
Jaime Carbonell
Thomas Dietterich
Norman Haas
Frederick Hayes-Roth
Gary Hendrix
Patrick Langley
Douglas Lenat
Ryszard Michalski
Tom Mitchell
Jack Mostow
Bernard Nudel
Michael Rychener
Ross Quinlan
Herbert Simon
Derek Sleeman
Robert Stepp
Paul Utgoff

Editors:

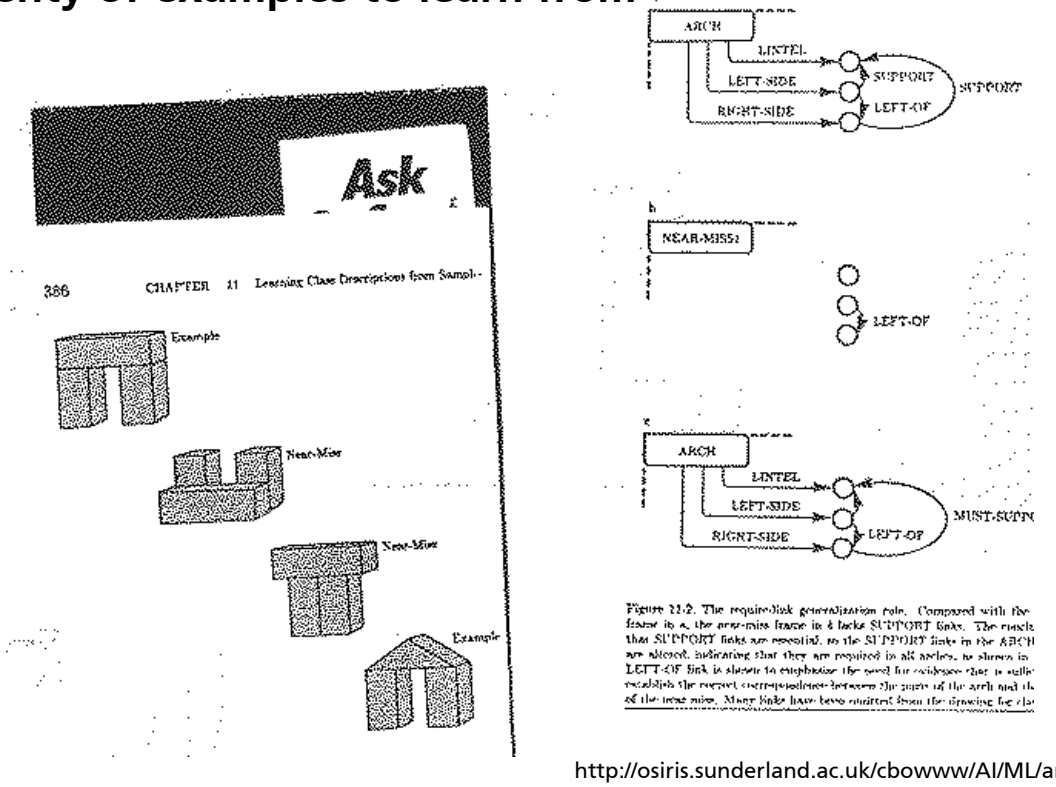
Ryszard S. Michalski
University of Illinois at Urbana-Champaign, IL

Jaime G. Carbonell
Carnegie-Mellon University Pittsburgh, PA

Tom M. Mitchell
Rutgers University New Brunswick, NJ

tioga publishing company palo alto, california

And plenty of examples to learn from



http://osiris.sunderland.ac.uk/cbowww/AI/ML/arch1.html

Even more, actually …

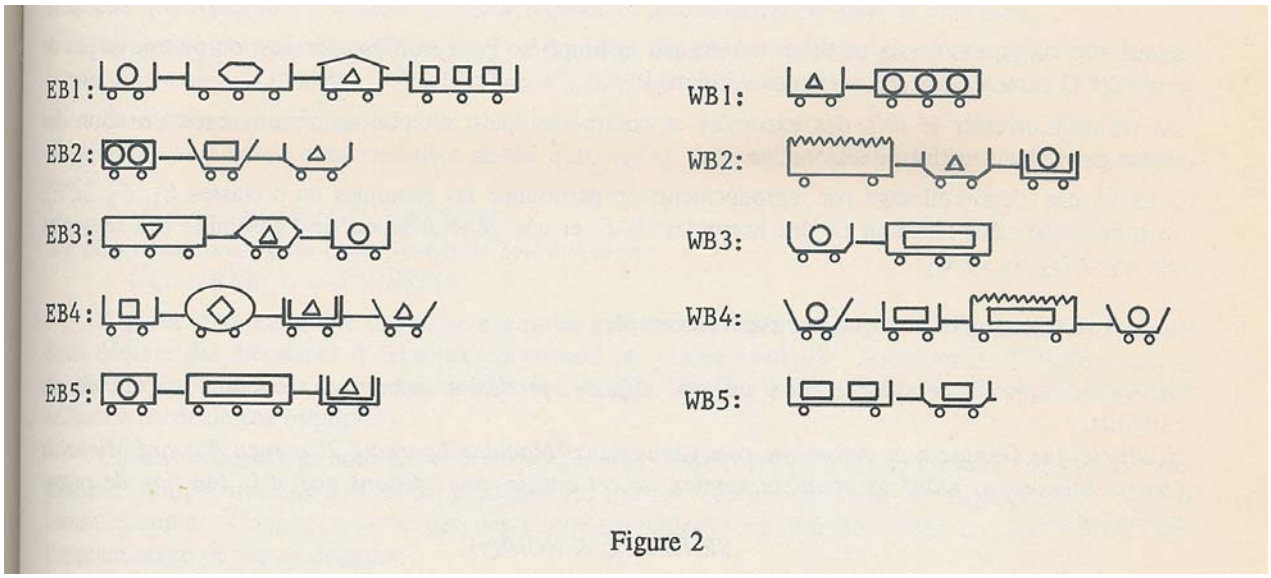


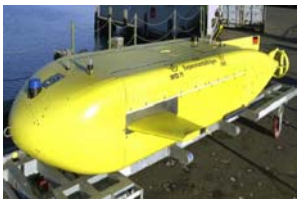
Figure 2

2008: Four Trends

Convergence

Ubiquitous intelligent systems

Users as producers



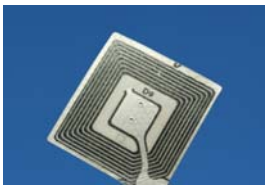
Networked autonomy

Convergence

- Universal digital representation of any media content
  - Web, MP3, digital cameras, Video
- Internet formats replace traditional delivery channels
  - Online Magazines, Blogs, Podcasts, Webradio, IPTV, Video on Demand
- Explosive growth of accessible media assets
  - digitalisation, crosslinking, swapping
- Enabling new business models
  - Flatrate models, individual access, niche content
- Search and management and interactivity are of central relevance

Ubiquitous intelligent systems





- Personal devices, integrated processors (Factor 20 – 30 above PCs)
- Interactivity, Sensors, Actuators
- Enormous production of data
- Physical and virtual worlds merge




Users as producers

- Web 2.0, Social Web, Crowdsourcing
- Exploding growth of content
- Media providers transform from content to confidence providers, competing with social communities
- Users expect full interactivity and control

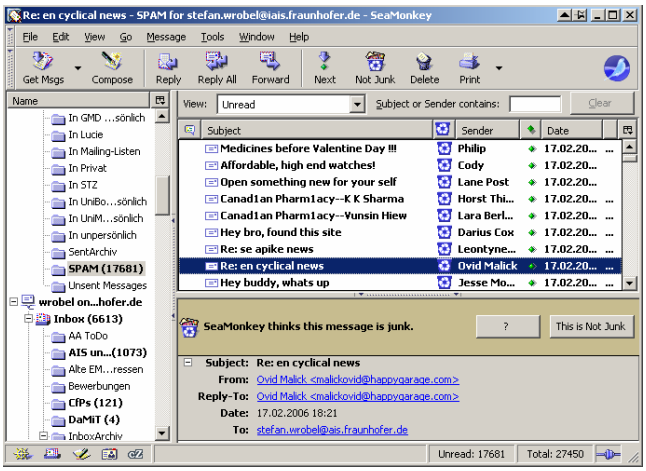
- Quality control, confidence, choice and searching are becoming central

myspace.com a place for friends
flickr
YouTube Broadcast Yourself™
HumanGrid™
studiVZ BETA 2.0 DAS STUDIVERZEICHNIS
XING
joost™
del.icio.us social bookmar

Networked Autonomy

- Growing readiness to use loosely controlled systems (autonomous agents)
- Loosely coupled company structures
- Service orientation (SOA) in IT systems
- First mobile autonomous systems

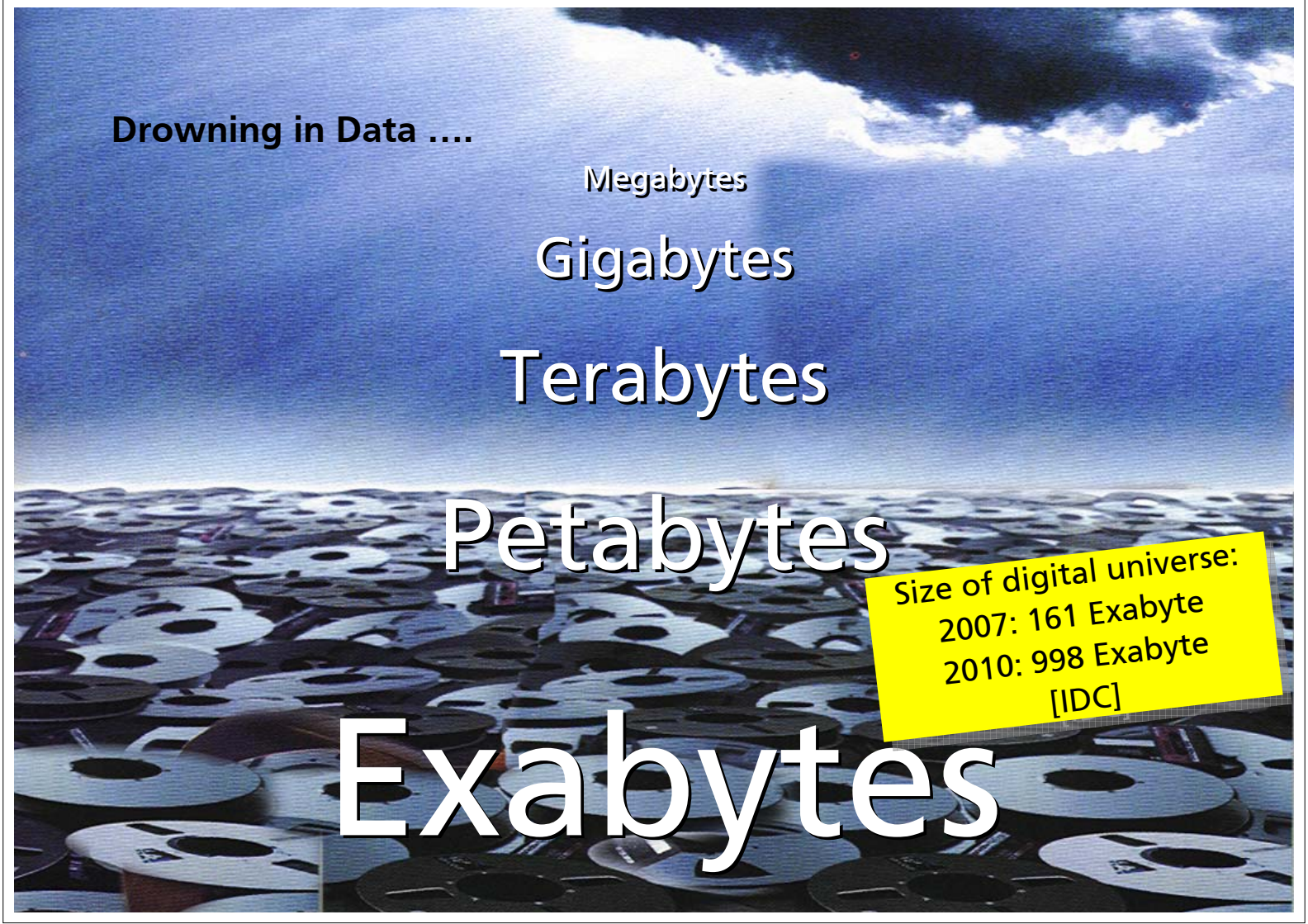
- Flexibility and capability for autonomous decisions on the basis of observations and goals is becoming central



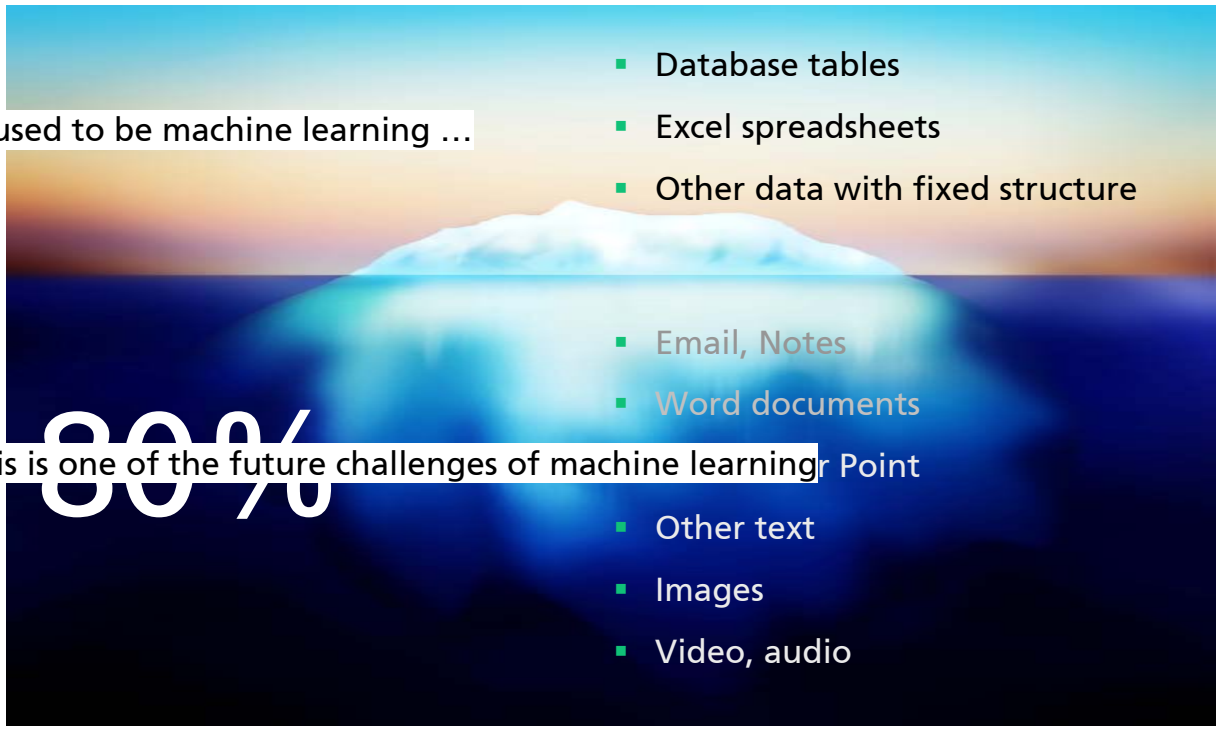
iRobot

ROBOWATCH

Fraunhofer Institut Intelligente Analyse- und Informationssysteme



The data iceberg

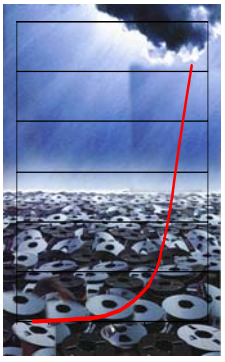


This used to be machine learning …

… this is one of the future challenges of machine learning

Challenges and research opportunities

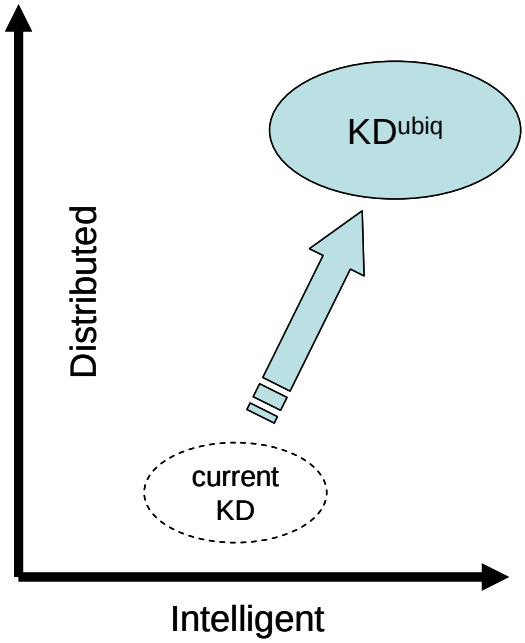
- Amount and variety of available data is growing with enormous dynamics
- Systems, people and organizations cannot handle them
- Yet using the knowledge hidden in those data is crucial for making the right decisions!



⇨ We need machine learning! More than ever.

⇨ Machine learning needs to become ubiquitous

Ubiquitous knowledge discovery and learning



Knowledge discovery process inside mobile, distributed, dynamic environments, in presence of massive amounts of data

=

Ubiquitous Knowledge Discovery

Project example: Outdoor Advertising Reach - Frequency Atlas

Customer:

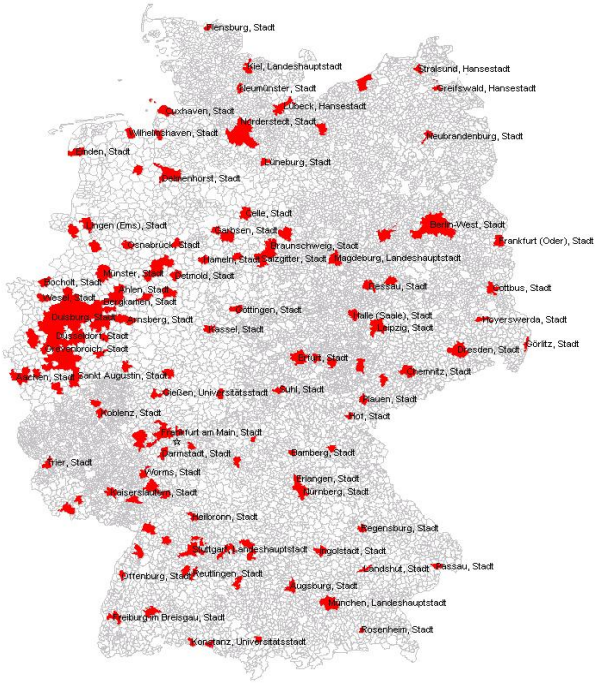
- Fachverband für Außenwerbung (FAW; Outdoor Advertising Association)

Task:

- Performance value assessment of advertising media
- Traffic volume forecast
- separate for private cars, public transport, pedestrians
- Spatial data mining, active learning procedures

First approach: a model based on stationary measurements

- Complete model for all German cities with more than 50.000 inhabitants (192 cities) = ca 1.000.000 street segments!
- Complete model includes, for each segment,
  item
  - car frequency
  - pedestrian frequency
  - public transport frequency
- The model is presently beeing extended to to all cities with between 20.000 and 50.000 inhabitants
- Official model for entire German outdoor advertising industry since May 2007



Fraunhofer Institut Intelligente Analyse- und Informationssysteme

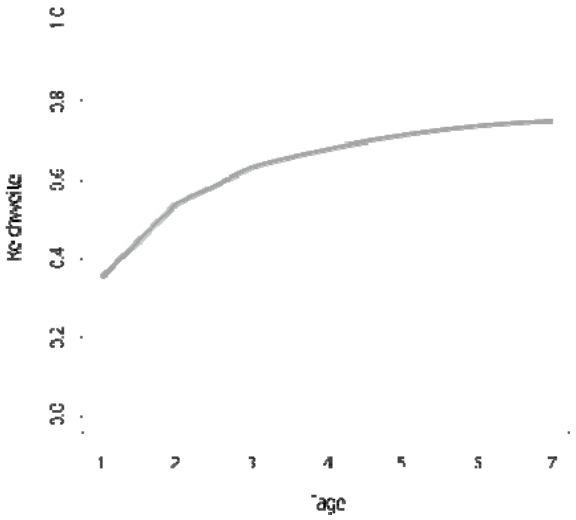
Ubiquitous approach: Mobility analysis based on GPS-tracks

- introduction of new pricing model for poster sites based on GPS tracks
- registration of contact frequencies with poster sites
- contact extrapolation for target groups:
  - socio-demographic characteristics
  - residential areas

Fraunhofer Institut Intelligente Analyse- und Informationssysteme

Time patterns

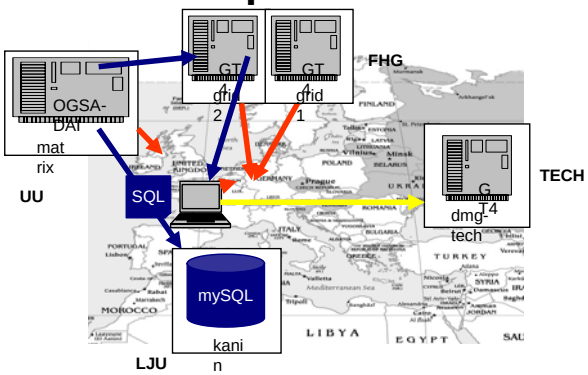
- **Patterns / Questions**
  - How long (days) does it take till x% of objects visit all locations?
  - How long does it take till x% of objects visit at least one location twice?
- **Applications**
  - determine mobility of a group of people
  - reach of poster networks
  - find popularity of locations (theatres, supermarkets, hospitals)



Fraunhofer Institut Intelligente Analyse- und Informationssysteme

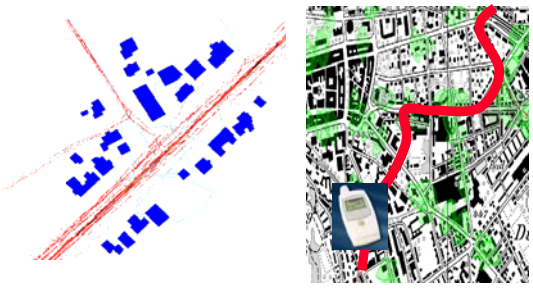
universität bonn

More examples …

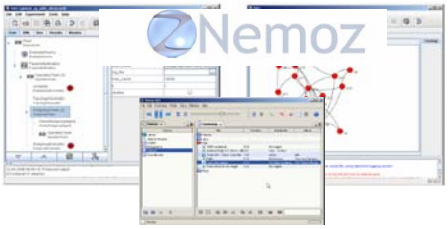


Grid-based Data Mining & Data Mining Based Grid Monitoring (Technion, Fraunhofer, Daimler)

Data Stream Mining (Univ. Porto)



Mobility Mining from GPS-Tracks (Fraunhofer, Univ. Pisa, Univ. Sabanci)



P2P/Web 2.0 Music Mining

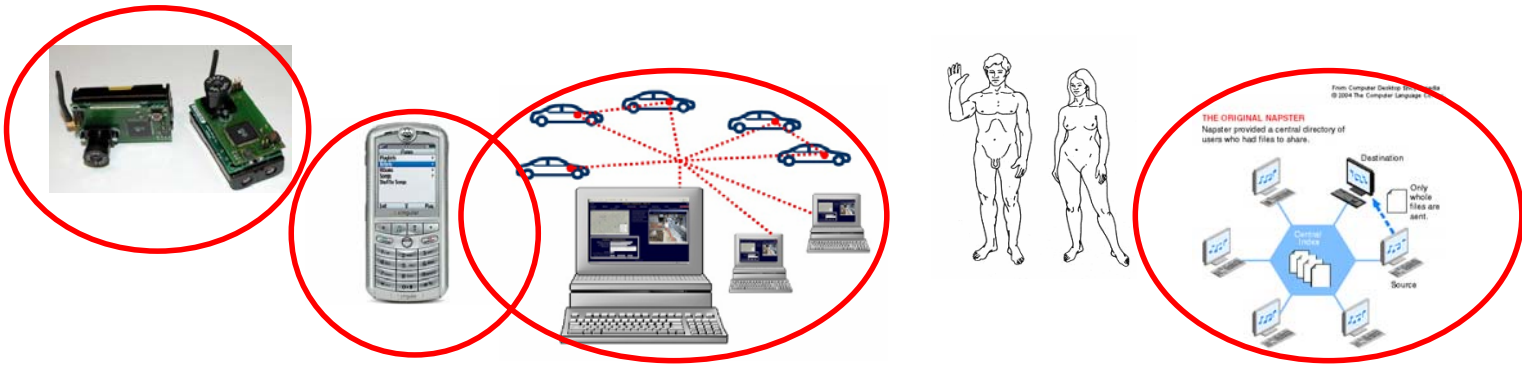
(Univ. Dortmund)

Key characteristics

1. **Time and space**. The objects of analysis exist in time and space. Often they are able to move.
2. **Dynamic environment**. These objects might not be stable over the life-time of an application. Instead they might appear or disappear.
3. **Information processing capability**. The objects themselves have information processing capabilities
4. **Locality**. The objects never see the global picture - they know only their local spatio-temporal environment.
5. **Real-Time**. They often have to take decisions or act upon their environment - analysis and inference has to be done in real-time.
6. **Distributed**. In many cases the object will be able to exchange information with other objects, thus forming a truly distributed environment

Objects of Study

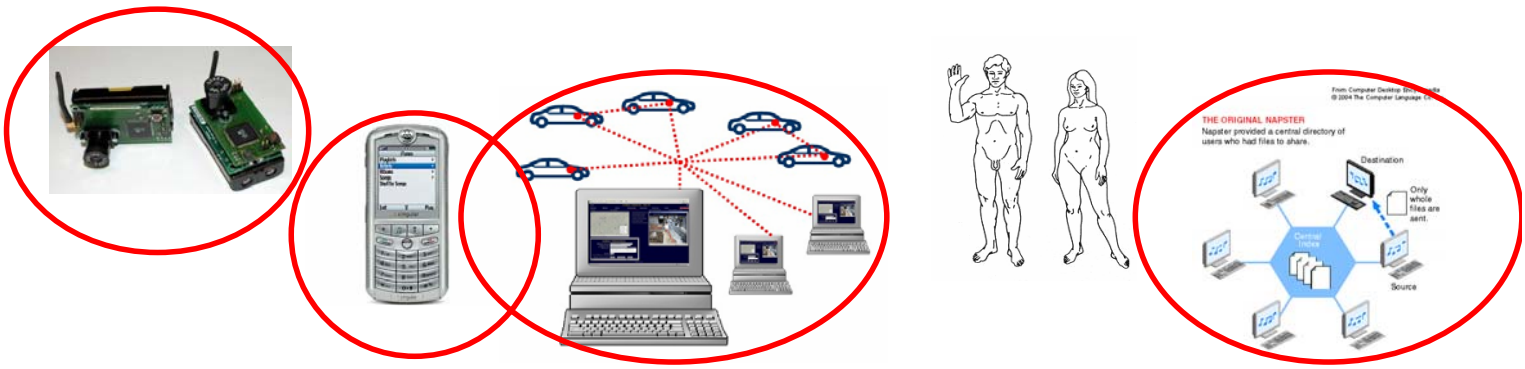
- Systems that have these properties are humans, animals, and increasingly, computing devices
- KDUbiq investigates artificial systems
  - The machine learning or data mining is not applied to data *about* the system,
  - it is rather *part* of the information processing capabilities of the system
- *This is a large departure from the current mainstream in machine learning and data mining!*



Characterization

- Ubiquitous knowledge discovery investigates learning *in situ*, inside distributed interacting artificial devices and under real-time constraints.

- Traditional machine learning and data mining collect data and analyze them offline at al later stage

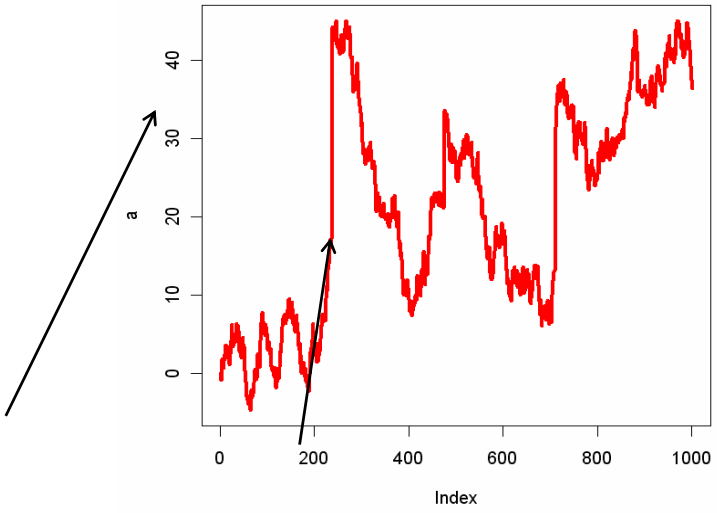


Resource Constraints

- Devices are resource constrained in terms of battery power, bandwidth, memory, …
  - This leads to a data streaming setting and to algorithms that may have to trade-off accuracy and effciency by using sampling, windowing, approximate inference etc.
  - In a traditional setting, data is processed in batch mode

Locality

- Inference is both temporally and spatially local.
  - This leads to focus on inference for non-stationary, non-independent data.
  - The distribution may be both temporally and spatially varying, and it may change both slowly or abruptly.
- A traditional setting assumes a random sample from a fixed distribution



Abrupt change Slow change

Spatial Locality

- Spatial locality (combined with resource constraints) leads to algorithms that are tailored for specifc network topologies and that make use of graph theoretic or geometric properties.
- Example: local majority voting for association rule mining (Wolff & Schuster 2003)
- A traditional setting assumes global availability of information

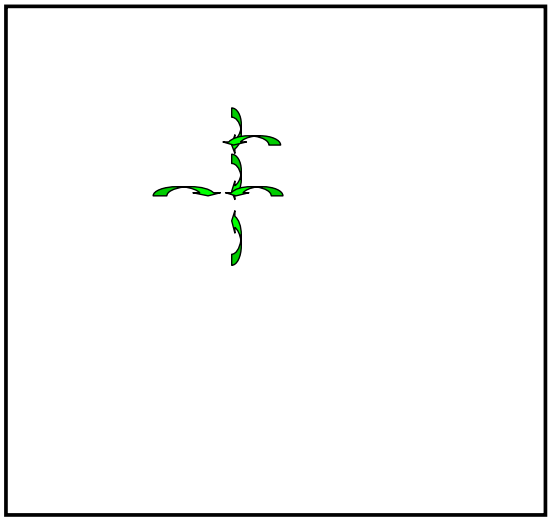
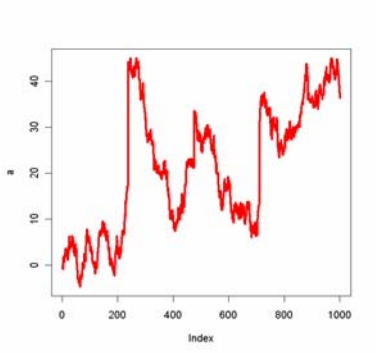


Image: Sc

Temporal Locality

- Temporal locality combined with real-time properties leads to online algorithms and to a shift from prediction to
  - monitoring,
  - change detection,
  - filtering or
  - short-term forecasts.



- Global forecasting (as in a traditional setting) is oft situation!

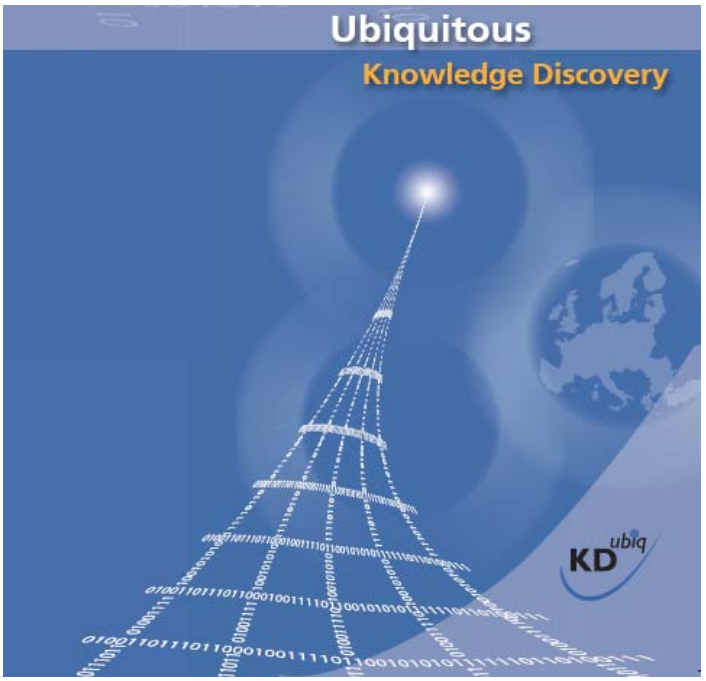
Further challenges

- Integrating results from
  - distributed data mining
  - privacy preserving data mining
  - spatio-temporal learning
  - Learning from data streams
  - collaborative data mining

  in Ubiquitous Learning systems

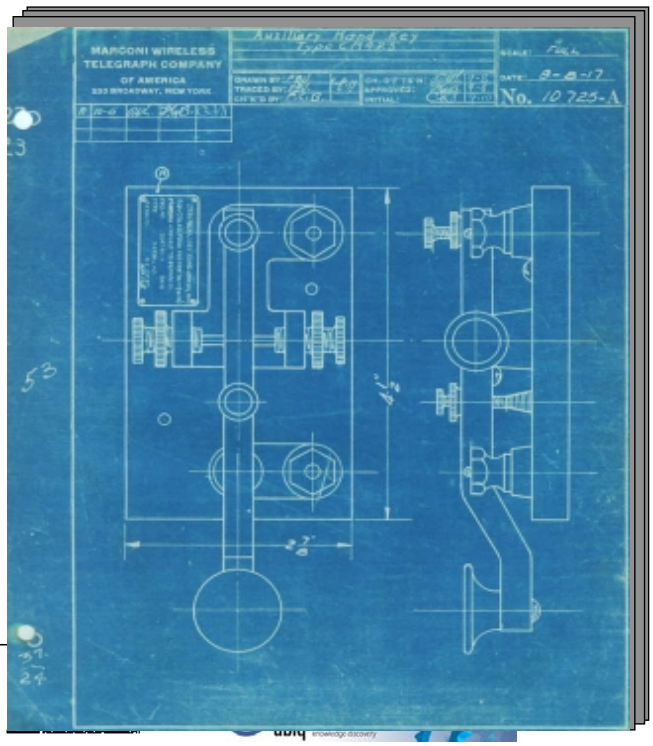
KD^{ubiq} Coordination Action

- To stimulate research, to define the field, and to shape the community in Europe, the KDUbiq research network was launched in 2006.
- It is funded by the European Commission
- Currently it has more than 50 members from research and industry
- Not a research project, it's about shaping a community
- Buget 1.2 Mio $, 2006-2008
- www.kdubiq.org
- **KDubiq IST-FP6-021321**
- Coordinator: Fraunhofer IAIS



Blueprint – collaborative book editing

- A collaborative effort to define the research challenges
- Six working groups corresponding to six main chapters
- 30 partners actively contributing
- Will result in a joint book in 2008



Summary

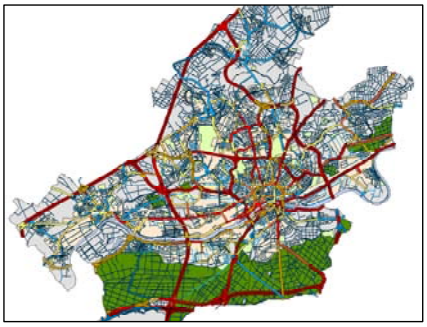
- After 35 years, machine learning is more up-to-date than ever
- We have gone from very few examples/data to more than we can handle:
  - Convergence
  - Ubiquitous intelligent devices
  - Users as producers
  - Networked autonomy
- Systems and applications will not work optimally if they do not learn
- Learning will be distributed and ubiquitous
  - Embedded in devices
  - Employing spatial context
  - Creating entirely new resource-aware abstractions of learning settings

Most work hasn't been done yet – what a wonderful future!
(Ingvar Kamprad)

Determining reach of a poster board

Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

Frequency



+

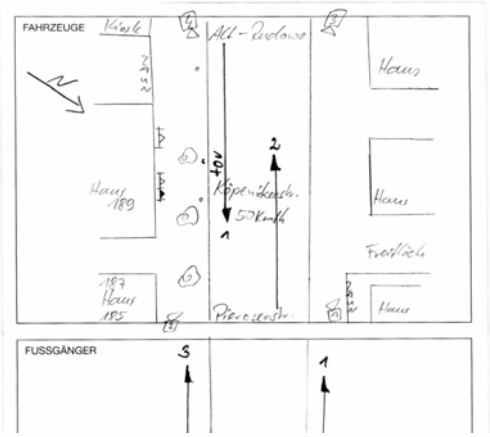
GfK Gesellschaft für Konsumforschung

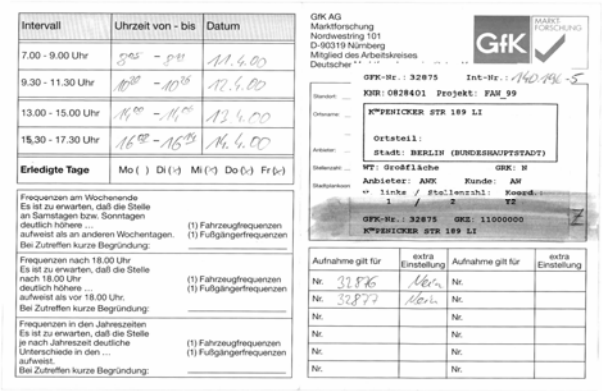
Media factories = poster reach

Basic Data: traffic measurements

Manual traffic measurement at selected poster locations
- 4 times 6 minutes at four days of the week at four times of day

- **Additional empirical model of day totals**
- Properties
  - Well defined measurements
  - Distribution of measurements tries to avoid systematic bias
  - Extended measurement period, so conceptdrift can not be excluded
- Total of 96.000 manual measurements

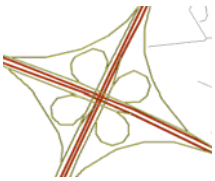




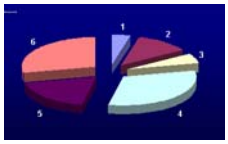
Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

universität bonn

Secondary data



Street network

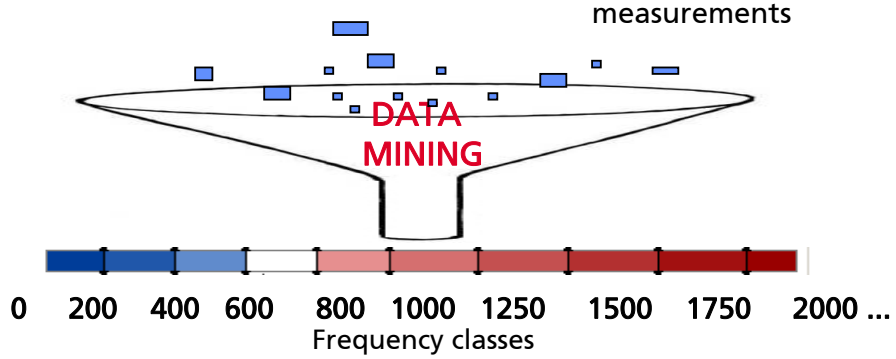


Soxiodemographics + Socioeconomics

Points of Interest (POI)

Frequency measurements

Public transport network



Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

universität bonn

Simple neighborhood model

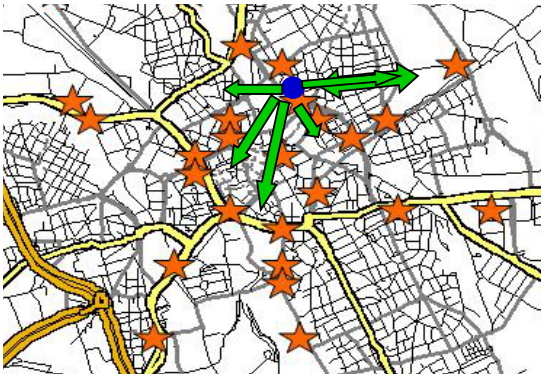
- **Attributes** of street segments:
  - Name, type, …. class
  - Points of Interest
  - Spatial coordinates
- **Locations with measurement values** ★
- **Distance beetween two segments** x_a, x_b

$$d(x_a, x_b) = \sum_{m=1}^{M} |x_{am} - x_{bm}|$$

- **Selection of the** k closest $x_1, \ldots, x_k$
- **Prediction for new segment** x_q

$$\hat{y}_q = \sum_{i=1}^{k} w_i y_i \Big/ \sum_{i=1}^{k} w_i \qquad \text{with } w_i = \frac{1}{d(x_q, x_i)}$$

- (Project has actually used specially adapted distance measure)

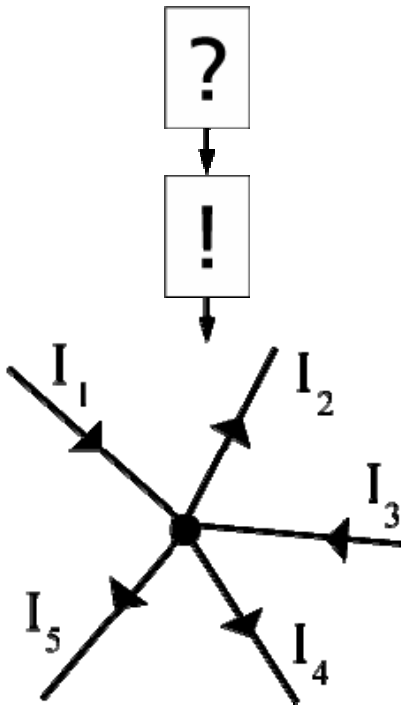


Smoothing based on flow constraints

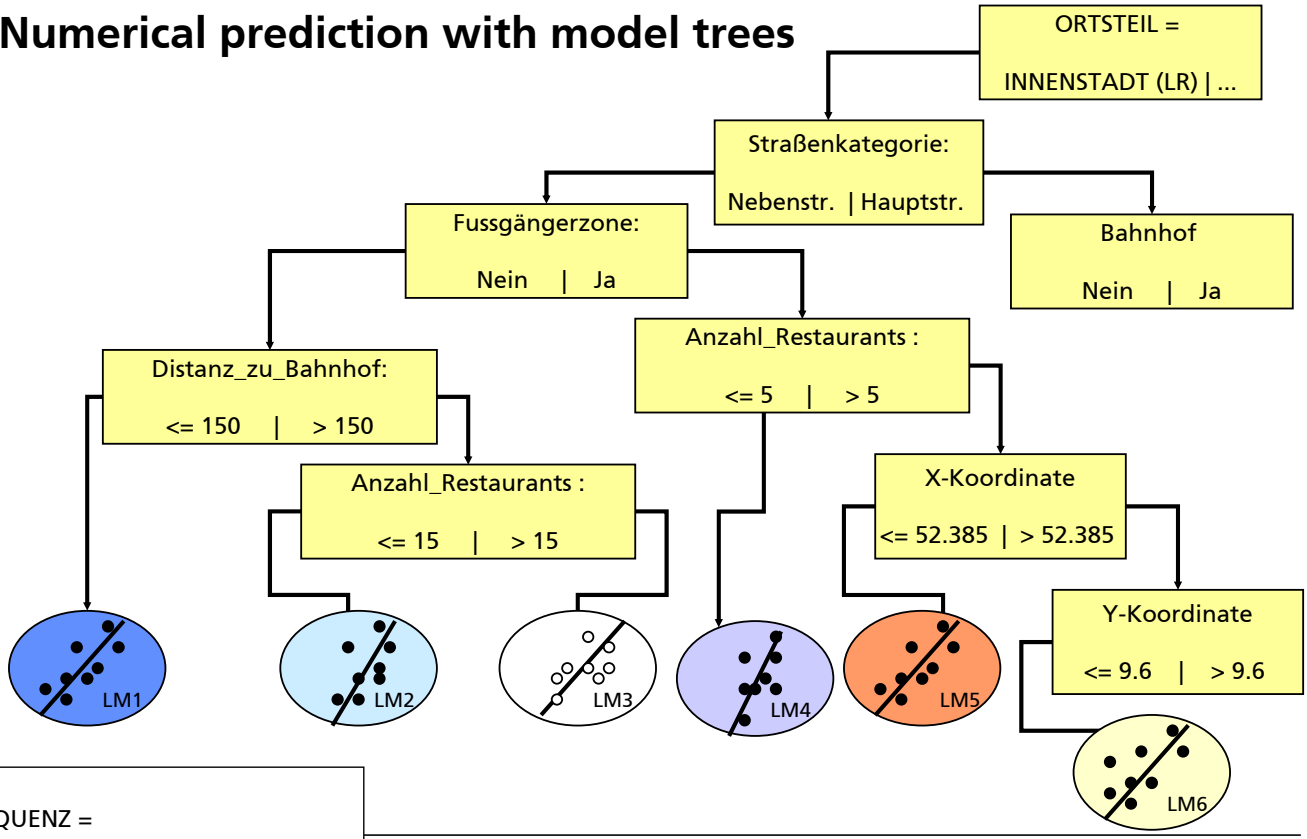
- Measurement errors lead to inconsistencies
- Need plausible assignment of frequencies

Solution:

- Use Kirchhoff's law as constraint
  - Sum of inputs = sum of outputs
- Smoothing algorithm finds locally optimal solution using constraint relaxation



Numerical prediction with model trees

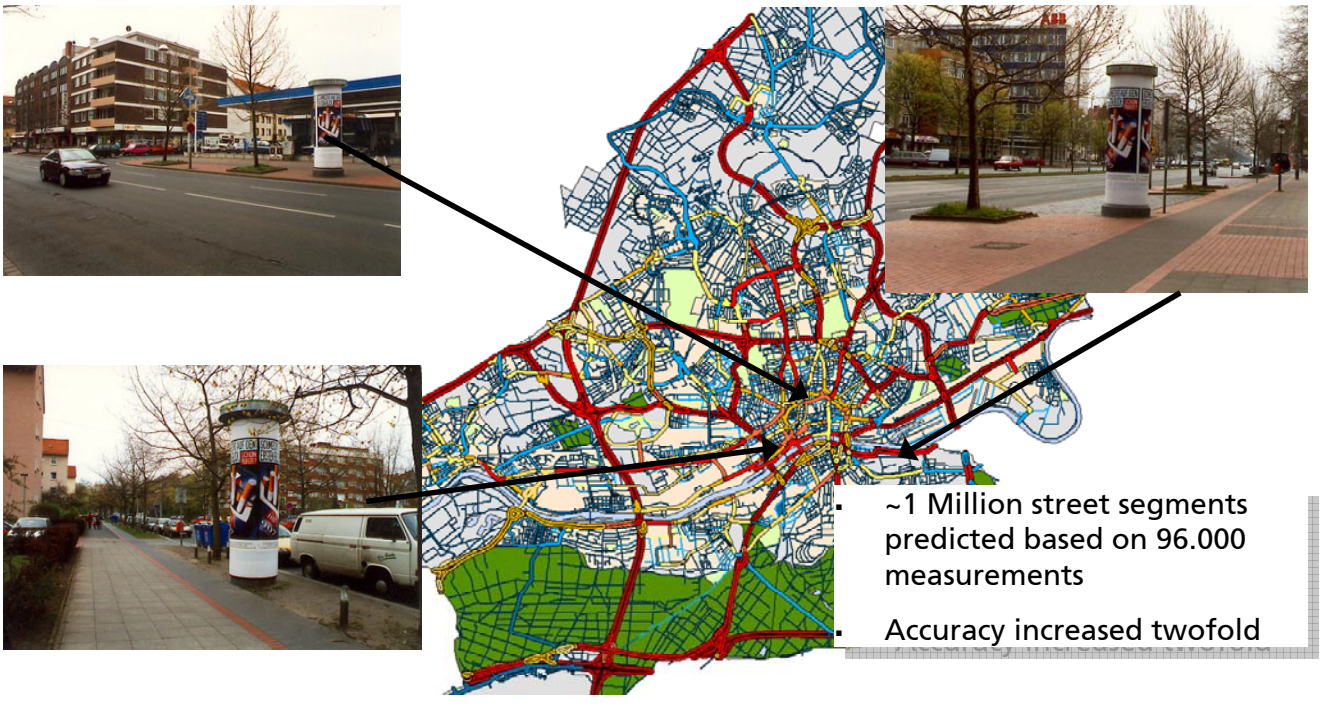


LM1
FREQUENZ =
2277.3186 * X +
75.4087 * ANZAHL_EINKAUF +
-142.4217 * MESSE +
-21221.8497

Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

universität bonn

Final result: frequency atlas (cars, public transport, pedestrians)



- ~1 Million street segments predicted based on 96.000 measurements
- Accuracy increased twofold

Fraunhofer IAIS Institut Intelligente Analyse- und Informationssysteme

universität bonn