Data Mining: How to pursue research, development and innovation together?

Wagner Meira Jr.

Universidade Federal de Minas Gerais
InWeb – National Institute of
Science and Technology for the Web

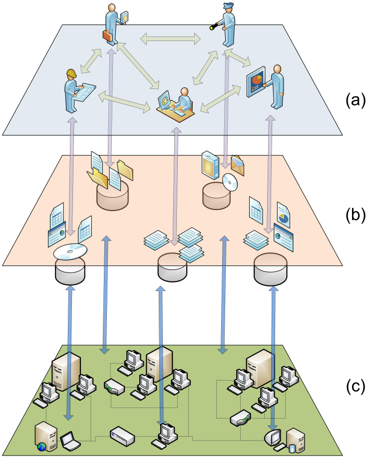
Apr 15, 2015

Mission

INWeb – Brazilian National Institute of Science and Technology for the Web

To develop models, algorithms and technologies to contribute to the integration of the Web with our society. As a result, we expect more effective and secure distribution of information, more efficient and useful applications, so that the Web can become a vector for social and economic changes in the country.

INWeb – National Institute of Science and Technology for the Web



Research Challenges

1. Identification, characterization and modeling of user interests and behavioral patterns on the web as well as of the social networks established among them.
2. Treatment of the information that circulates on the various networks of the web.
3. Delivery of information in a satisfying way regardless of time and place.

InWeb – Research Tracks

1. **Social Networks**
   (Coordinator: Virgilio Almeida)
2. **User Behavior and Interaction Modeling**
   (Coordinator: Jussara Almeida)
3. **Information Retrieval**
   (Coordinator: Nivio Ziviani)
4. **Web Data Management**
   (Coordinator: Alberto Laender)
5. **Parallel and Distributed Systems**
   (Coordinator: Dorgival Guedes)
6. **Knowledge Discovery**
   (Coordinator: Wagner Meira Jr.)

PIs

- Loic Cerf
- Wagner Meira Jr.
- Raquel Melo-Minardi
- Gisele Pappa
- Adriano Pereira
- Adriano Veloso

Researchers

- PhD students: 8
- MSc students: 11
- Undergrads: 10

Research, Development and Innovation

Research

Advance the state-of-the-art.

Development

Generate products.

Innovation

Evolve products by incorporating state-of-the-art results.

Data

- Tabular
  - categorical
  - numeric
- Text
- Graphs
- Sound
- Image
- Video

- Storage
- Accessing
- Engineering
  - Integration
  - Cleaning
  - Transformation
- Visualization

Curse of Dimensionality

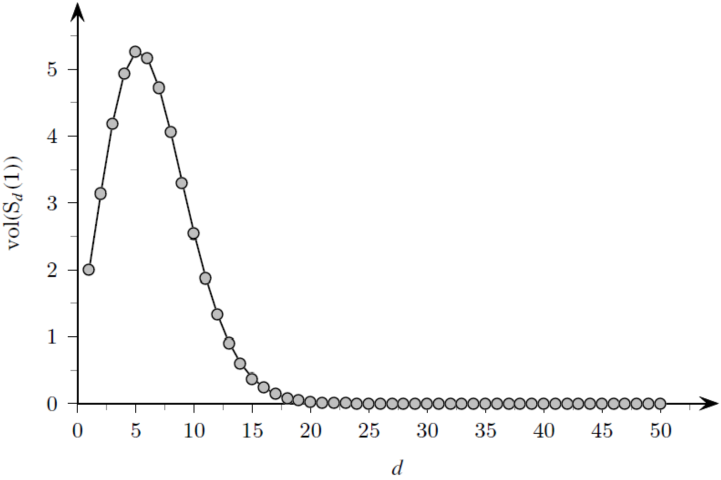
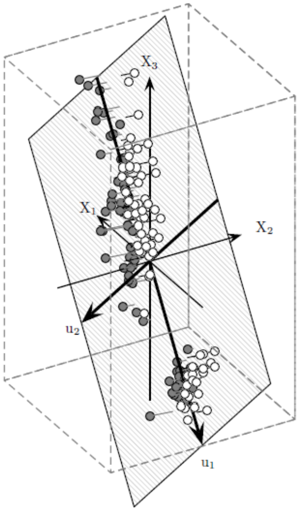
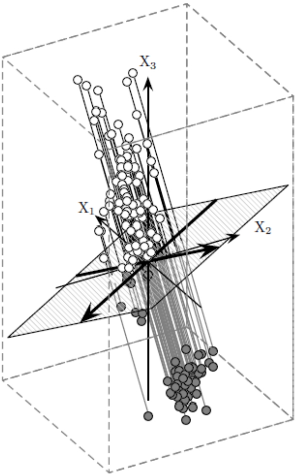


Figure 6.2. Volume of a unit hypersphere.

Curse of Dimensionality



(a) Optimal basis



(b) Nonoptimal basis

Data Mining

Concept

Automatic extraction of knowledge or patterns that are interesting (novel, useful, implicit, etc.) from large volumes of data.

Tasks

- Data engineering
- Characterization
- Prediction

Data Mining Models

Concept

A model aims to represent the nature or reality from a specific perspective. A model is an artificial construction where all extraneous details have been removed or abstracted, while keeping the key features necessary for analysis and understanding.

Paradigms

- Combinatorial
- Probabilistic
- Algebraic
- Graph-based

Combinatorial Models

Problem

Determine the sets of items that occur simultaneously in transactions.

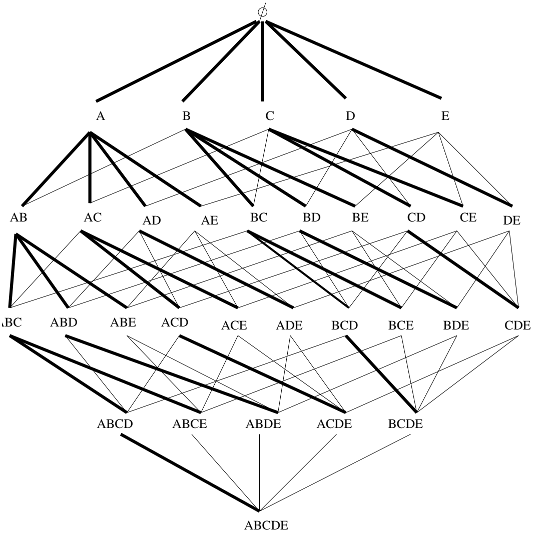
Strategy

Traverse the search space of sets of items determining whether they co-occur.

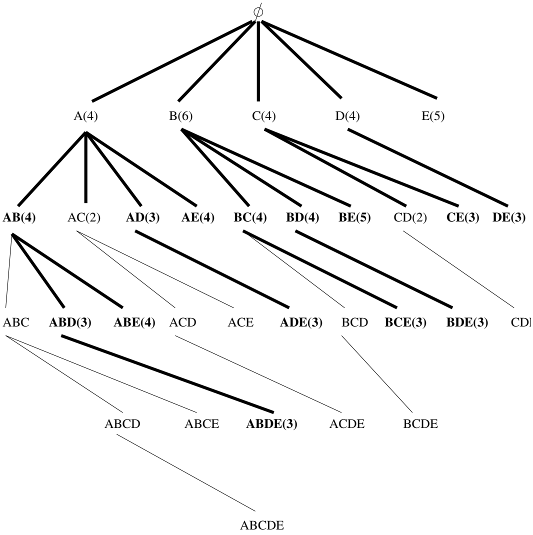
Challenge

There are $O(2^n)$ possible sets given n items.

Combinatorial Models



Combinatorial Models



Probabilistic Models

Problem

Determine the groups of entities that are similar and may be handled together.

Strategy

Model the likelihood ot belonging to a group (cluster) as a probabilistic function.

Challenge

We should determine an expressive yet simple to represent and manipulate model.

Probabilistic Models

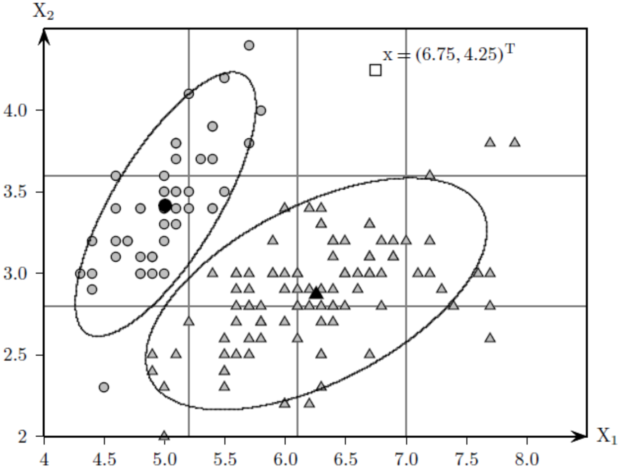


Figure 18.1. Iris data: X_1:`sepal length` versus X_2:`sepal width`. The class means are show in black; the density contours are also shown. The square represents a test point labeled x.

Probabilistic Models

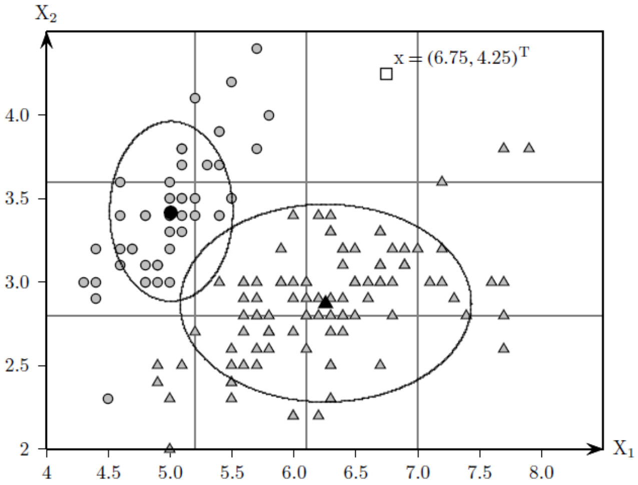
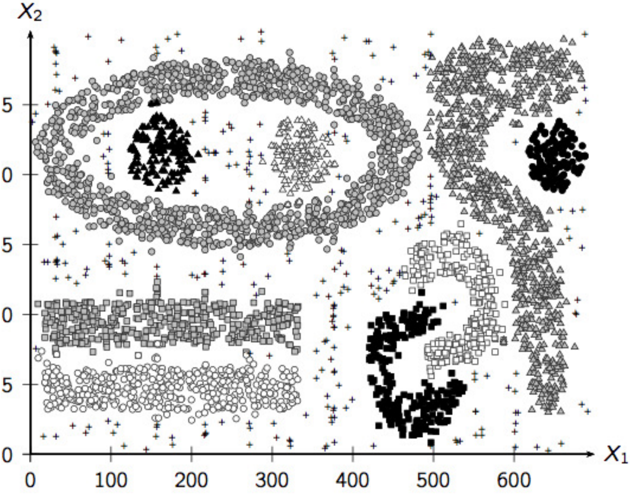
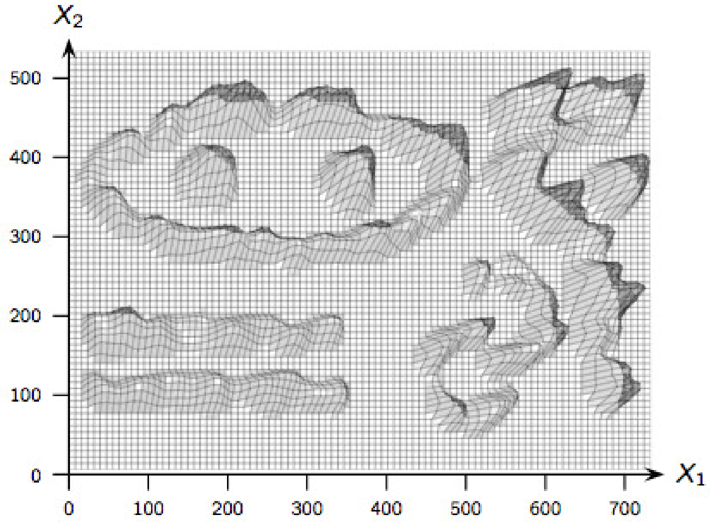


Figure 18.2. Naive Bayes: X_1:`sepal length` versus X_2:`sepal width`. The class means are shown in black; the density contours are also shown. The square represents a test point labeled x.

Probabilistic Models




X_2
500
400
300
200
100
0
X_1
0
100
200
300
400
500
600
700

Algebraic Models

Problem

Predict the class of an entity, given a set of known entities previously assessed.

Strategy

Create a prediction model that partitions the entities into classes and use the model to classify unknown samples.

Challenge

How to couple with bias and variance?

Algebraic Models

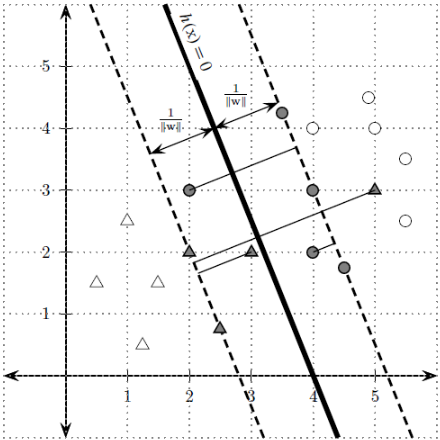


Figure 21.3. Soft margin hyperplane: the shaded points are the support vectors. The margin is $1/\|w\|$ as illustrated, and points with positive slack values are also shown (thin black line).

Graph-based Models

Problem

Determine the groups of entities that are similar and may be handled together.

Strategy

Model the relations among entities as a weighted graph and partition the graph looking for minimum cuts.

Challenge

Weight model.

Graph-based Models

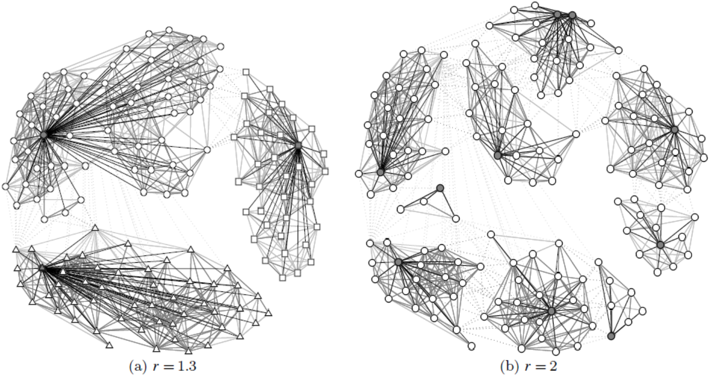
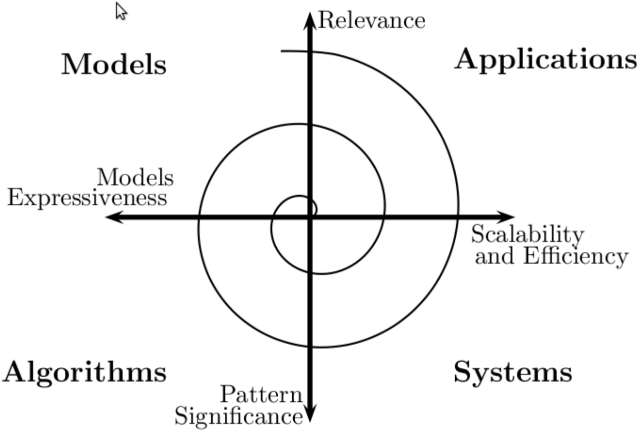


Figure 16.6. MCL on Iris graph.

- Huge number of relevant applications
- Broad spectrum of scenarios
- Data volume, nature and complexity variety
- Privacy, security and data quality issues
- Techniques demand data-dependent and manual parametrization

How may data mining models and algorithms account for:

- Social theories?
- Invariants?
- Premises?
- Dynamic behavior?


Relevance
Models
Applications
Models Expressiveness
Scalability and Efficiency
Algorithms
Systems
Pattern Significance

Social Network Mining Challenges

Fact

The evolution of the Internet and the Web makes them not only very popular, but also dynamic and diversified social media that may be used to sense and understand the society.

Mining social networks must deal with:

- Dynamic behavior
- Complex relationships
- Heterogeneous data
- Incomplete information
- Noisy data
- Lack of scalability

Social Network Mining Challenges

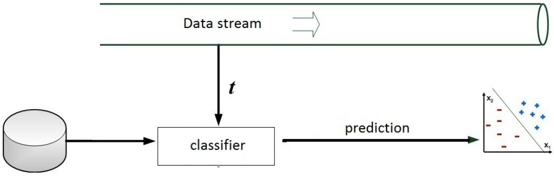
Mining social networks must deal with:

- **Dynamic behavior**
- Complex relationships
- Heterogeneous data
- Incomplete information
- Noisy data
- Lack of scalability

- Definition
  - Automatically extraction of opinions, sentiments, attitudes, and emotions expressed in text messages (i.e., Twitter).
- Motivation
  - It allows us to track products, brands and people to determine whether they are viewed positively or negatively.
- Problem
  - Content is created almost at the same time the event is happening in the real world.
    - Keeping track of sentiment streams is useful for advertising.

Classifying Sentiment Streams

- Effective classification requires:
  - Updating the training-set to mitigate drifts.
  - Updating the classifier accordingly.

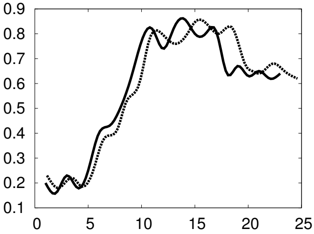


Research Questions

1. Effort:
   - How to reduce labeling effort?
2. Accuracy:
   - How to select messages to be kept and discarded?

Dealing with Drifts

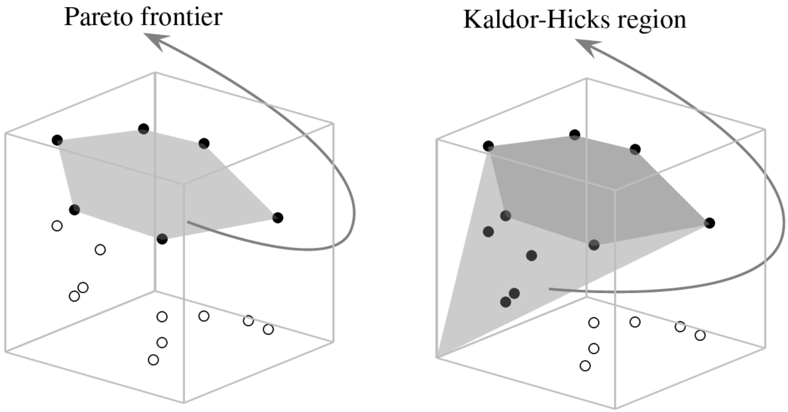
- Two properties are necessary in order to produce classifiers that are robust to drifts:
  - Adaptiveness:
    - The ability to adapt itself to drifts.
  - Memorability:
    - The ability to recover itself from drifts.



Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
  - Adaptiveness:
    - The ability to adapt itself to drifts.
    - The training-set must contain fresh messages.
  - Memorability:
    - The ability to recover itself from drifts.
    - The training-set must contain pre-drift messages.
- Improving both properties simultaneously may lead to a conflict-objective problem.
  - Improve adaptiveness may hurt memorability, and vice-versa.

Pareto and Kaldor-Hicks Principles

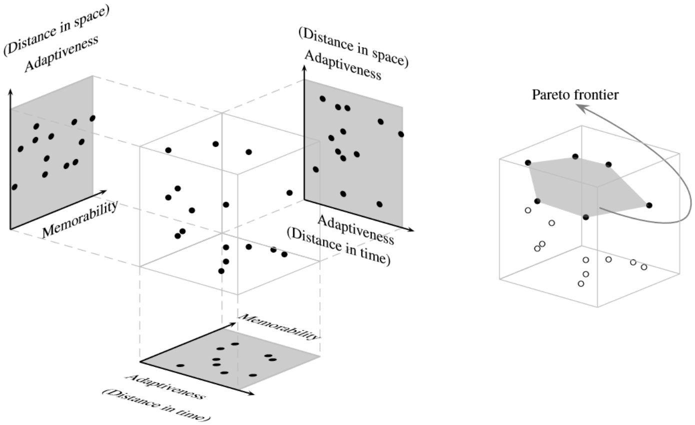


Utility Measures

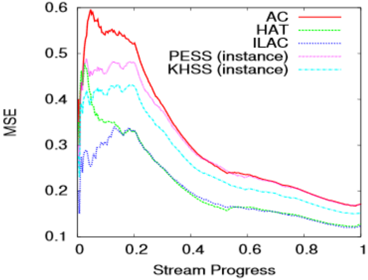
- Distance in space:
  - How similar message t_j is to the newest message t_n.
  - $U_s(t_j) = \frac{|\mathcal{R}(t_n) \cap \mathcal{R}(t_j)|}{|\mathcal{R}(t_n)|}$
- Distance in time:
  - How fresh is the message.
  - $U_t(t_j) = \frac{\gamma(t_j)}{\gamma(t_n)}$.
    - $\gamma(t_j)$ returns the time in which message t_j arrived.
- Random permutation of messages:
  - $U_r(t_j) = \frac{\alpha(t_j)}{|\mathcal{D}_n|}$
    - $\alpha(t_j)$ returns the position of t_j in the shuffle.
    - $\mathcal{D}_n$ is the training set at time step n.

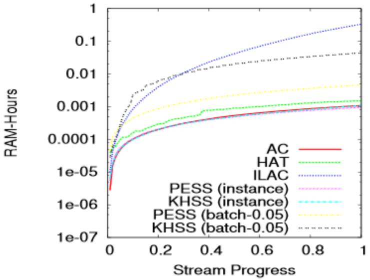
Utility Measures

1. At each time step n:
   1. Place candidate messages in the utility space.
   2. Select messages in the Pareto frontier.

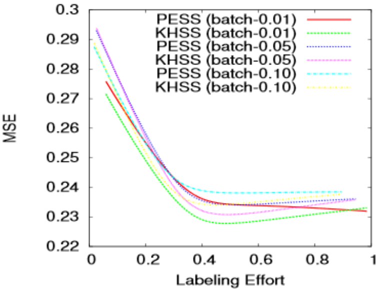


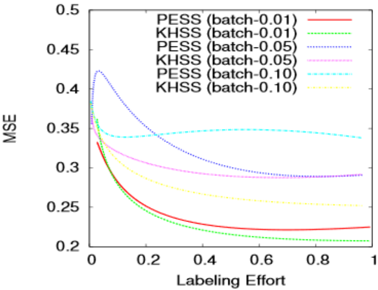
- MSE and RAM-Hours





- MSE and Labeling Effort





Social Network Mining Challenges

Mining social networks must deal with:

- Dynamic behavior
- **Complex relationships**
- Heterogeneous data
- Incomplete information
- Noisy data
- Lack of scalability

Motivation

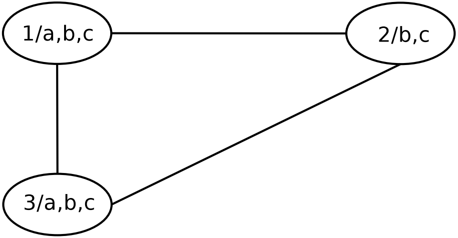
- Attribute patterns provide correlations in terms of the content
- Topological patterns provide correlations in terms of the network structure
- Both patterns refer to the same entities and information

How can we analyze them together?

Structural Correlation Patterns

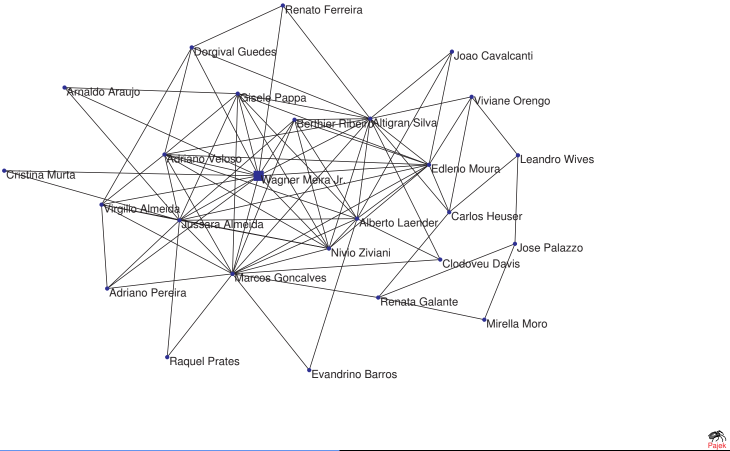
Problem

Determine attribute sets associated with the existence of dense connected subgraphs.

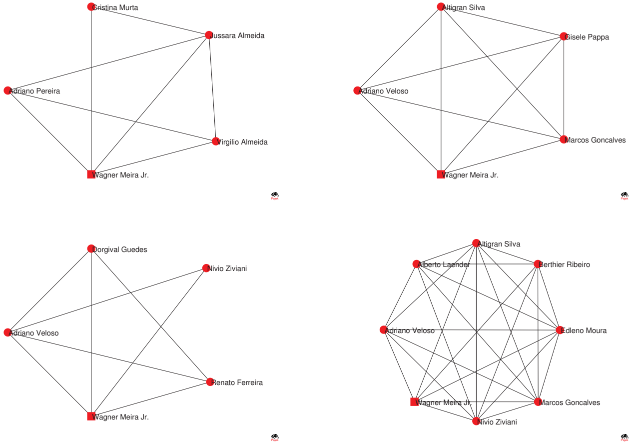


Structural Correlation Patterns

Co-authorship in InWeb



Structural Correlation Patterns
Co-authorship in InWeb



Structural Correlation Pattern Mining

What is the probability of a vertex that has an attribute set S be part of a correlated dense subgraph?

- An SCP is a pair (attribute set, dense subgraph)
- Dense subgraphs are defined as quasi-cliques

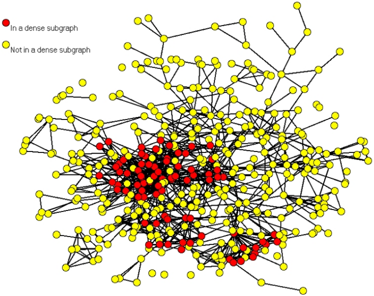
Problem:

Identifying attributes and their respective structural patterns (i.e., dense subgraphs) given a set of constraints:

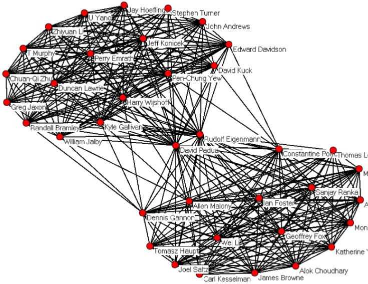
- Attribute set frequency, dense subgraph size and density, ϵ (structural coverage), statistical significance of ϵ.

Example: DBLP

attribute set	support	str. correlation	stat. significance
search rank	420	0.19	635,349
perform file	404	0.14	555,067
structur index	404	0.14	555,067
search mine	413	0.14	490,932
us xml	400	0.11	442,638



(a) $\{search, rank\}$



(b) $\{perform, system\}$

Social Network Mining Challenges

Mining social networks must deal with:

- Dynamic behavior
- Complex relationships
- **Heterogeneous data**
- Incomplete information
- Noisy data
- Lack of scalability

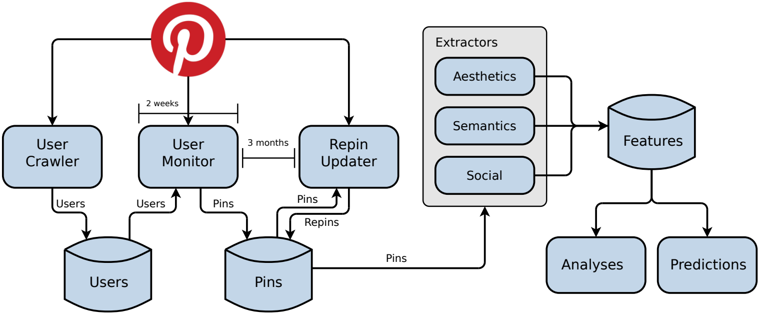
Impact of Visual Attributes on Diffusion

WebSci'14

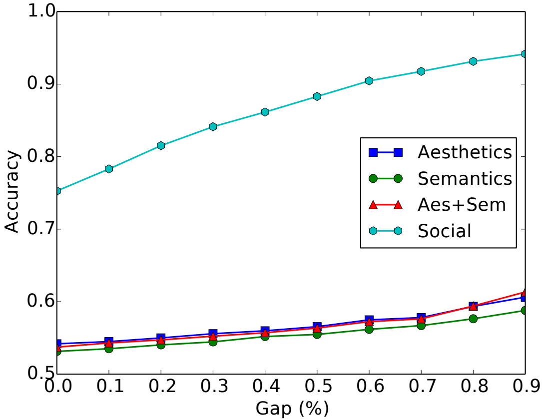
Can visual attributes explain the diffusion of images?

- Aesthetical: 12 properties (e.g., brightness, contrast, sharpness)
- Semantical: 85 concepts represented by image
- Social: 12 features derived from the network

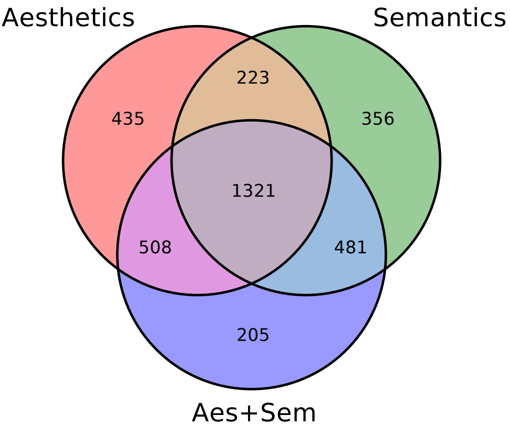
Framework


User Crawler
2 weeks
User Monitor
3 months
Repin Updater
Users
Users
Pins
Pins
Repins
Users
Pins
Pins
Extractors
Aesthetics
Semantics
Social
Features
Analyses
Predictions

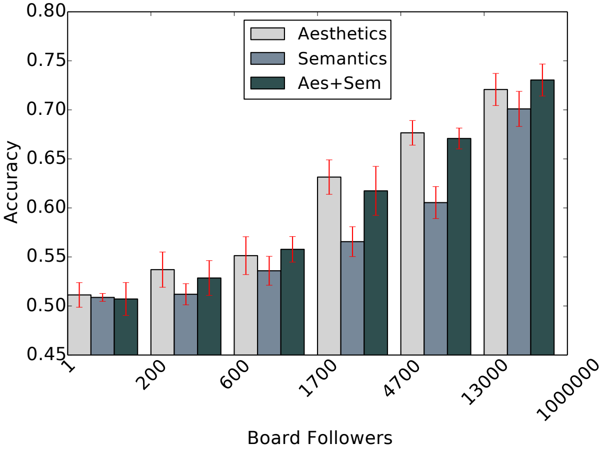
Accuracy in Predicting Popularity



Visual Attributes are Complementary



Popularity is a factor



Social Network Mining Challenges

Mining social networks must deal with:

- Dynamic behavior
- Complex relationships
- Heterogeneous data
- **Incomplete information**
- Noisy data
- Lack of scalability

Knowledge Transfer to Sentiment Analysis

KDD'11, JIDM'11, ICWSM'13, and WSDM'14

Sentiment Analysis

Sentiment Analysis (or opinion mining) aims to interpret text and predict polarity of the writer regarding a topic or entity.

Challenges

- Language ambiguity
- Dinamicity of discussions
- Lack of labeled textual data

Is it possible to analyze sentiment without assessing content?

Bias is inherent to most humans [Watson 1991], since they:

- take a particular position regarding a subject
- have a personal interest from the arguer in the outcome of the argument or discussion.
- lack proper balance and neutrality in argumentation
- lack proper critical doubt

Bias is inherent to most humans [Watson 1991], since they:

- take a particular position regarding a subject
- have a personal interest from the arguer in the outcome of the argument or discussion.
- lack proper balance and neutrality in argumentation
- lack proper critical doubt

On polarized networks, bias and opinions are dependent!

- Supporters of a candidate are likely to issue positive opinions on him/her
- Soccer team supporters act similarly

Social Media Endorsements as Evidence of User Bias

Endorsements: interactions through which a user implicitly **agrees** with another user w.r.t. a certain content:

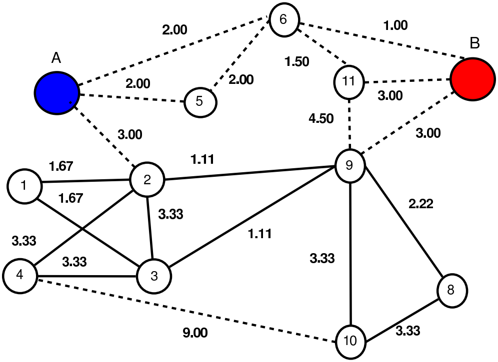
retweet
@OfficialMyTeamProfile, @CandidateX.

like
Democrats, Republicans, New York Giants

pin, repin
people, companies, causes

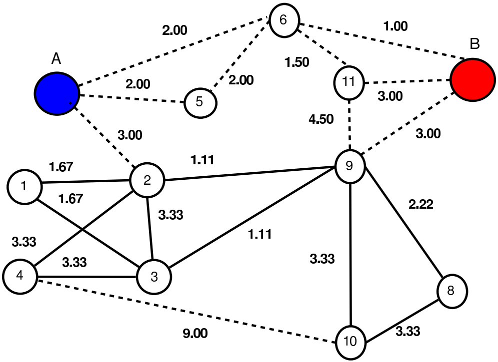
The Opinion Agreement Graph

- **Solid edge**: two users **endorse** the same users
- **Dashed edge**: two users **are endorsed** by the same users
- **Edge weight**: the lift of the size of both sets



The Opinion Agreement Graph

- Attractors: seeds that represent a polarized group



Transfer Knowledge to Sentiment Analysis

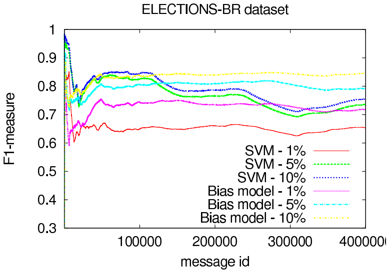
1. Collect data and identify attractors
2. Build the opinion agreement graph
3. Determine the bias of each user based on the attractor's messages endorsed by him/her
4. Analyze messages whose polarity is unknown through the bias vectors of the users who endorse them

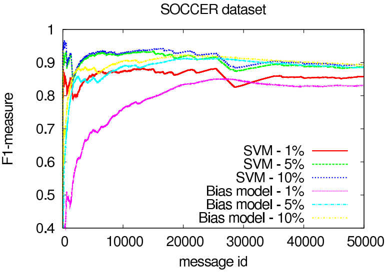
Comparison to SVM

Data Mining

InWeb
KD@InWeb
Data Mining
Graph Mining
Dynamic behavior
Complex relationships
Heterogeneous data
Noisy data
Incomplete information
Lack of scalability
Data Science
Summary

- Competitive to SVM, despite not using labeled textual data
- SVM performance decreases over time, bias-based does not





Latest developments
WSDM'14

- Users present self-report imbalances, that is, they
  - tend to report more positive emotions.
  - tend to report more extreme emotions.
- We exploit such imbalances by
  - considering positive emotions to label data.
  - considering terms used in spikes in social streams.
- Our social psychology-inspired framework produces accuracies up to 84% while analyzing live reactions.

Self-Report Imbalances

Self-Report Imbalances

- Social media: **self-reported** platforms [Rost et al., CSCW'13; Lin et al., WWW'13]

Self-Report Imbalances

- Social media: **self-reported** platforms [Rost et al., CSCW'13; Lin et al., WWW'13]
- Opinions seen on social media are **not** a random sample of the opinion population

Positive-negative Self-report Imbalance

1. People tend to express **positive** feelings more than **negative** feelings in social environments [Berger, 2013; Diener, 1985; Larson, 1982]

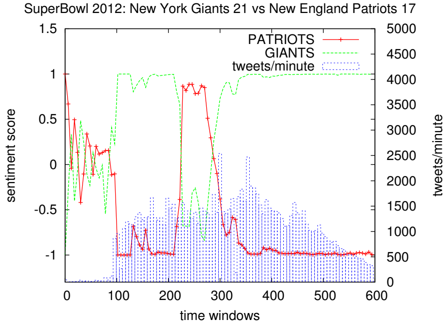


Figure:

Extreme-Average Self-Report Imbalance

1. People tend to express **extreme** feelings more than **average** feelings in social environments [Anderson, 1998; Dellaroccas, 2006; Kiciman, 2012]

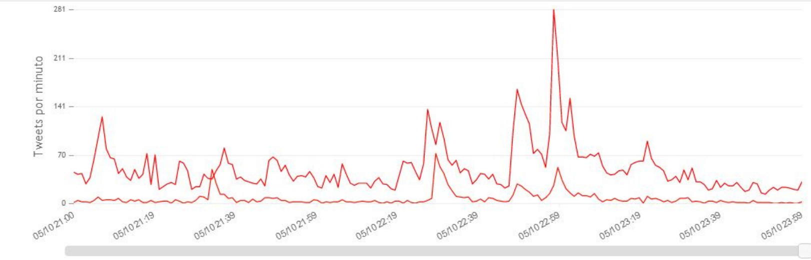


Figure: consequence: spikes tend to have meaningful, informative terms

Term arousal

- classical feature representation: TF, TF-IDF...
- problem: they are static and do not react quickly to new, discriminative sentiment terms

Term arousal

- classical feature representation: TF, TF-IDF...
- problem: they are static and do not react quickly to new, discriminative sentiment terms
- we propose a **term arousal** representation:

$$w_{t,term} = \frac{\overline{W_t, term}}{\overline{W_t}} \tag{1}$$

- intuition: informative "sentimental" terms should appear more frequently in spikes

Term arousal

Top 5 features according to TF-IDF...

- win!
- gol_from_team
- an_equalizer
- go!
- he_shoots!

Term arousal

Top 5 features according to TF-IDF...

- win!
- gol_from_team
- an_equalizer
- go!
- he_shoots!

... and according our new metric *term arousal*:

- great_goal (7.53)
- gooooooooool (6.80)
- he_scores (5.31)
- GOOOOL (5.00)
- penalty_for_team (3.34)

Social Network Mining Challenges

Mining social networks must deal with:

- Dynamic behavior
- Complex relationships
- Heterogeneous data
- Incomplete information
- **Noisy data**
- Lack of scalability

Motivation

- There is an increasing use of the Web in events of overall interest such as politics and sports.
- Major motivations are the lack of a central control and the fast information propagation.
- Recently, there has been an emphasis on "what you are doing" instead of "who you are".

Challenge

Qualify, quantify, and summarize the content being exchanged in the various Internet-related media on line and evaluate its impact on specific events.

Web Observatory

On line tool for capturing, analyzing and presenting the dynamics of a given scenario on the Web.

Scenarios

- Soccer World Cup
- Olympics
- Brazilian National Soccer League
- Brazilian Elections
- Public Safety
- Brand reputation
- **Dengue Epidemics**

Background on dengue

- Dengue is a mosquito-borne infection that causes a severe flu-like illness, and sometimes a potentially lethal complication
- Approximately 2 billion people from more than 100 countries are at risk of infection and about 50 million infections occur every year worldwide
- Outbreaks tend to occur every year during the rainy season but there is large variation of the degree of the epidemic in areas with similar rainfall

Background on dengue

- Current strategies for prediction of dengue epidemics are based on surveillance of insects, which provide only a rough estimate of cases
- Once disease outbreaks are detected in a certain area, efforts need to be concentrated to avoid further cases and to optimize treatment and staff - number of cases may reach several hundred thousands
- In Brazil, where there is a epidemics accounting system, detection of important outbreaks may take a few weeks, leading to loss of precious time to address the epidemy

- To analyze how dengue epidemics manifests in Twitter and to what extent that information can be used for surveillance.
- To design and implement an active surveillance framework that analyzes how social media reflects epidemics based on a combination of four dimensions: volume, location, time, and public perception.
- To exploit user generated content available in online social media to predict the dengue epidemics.

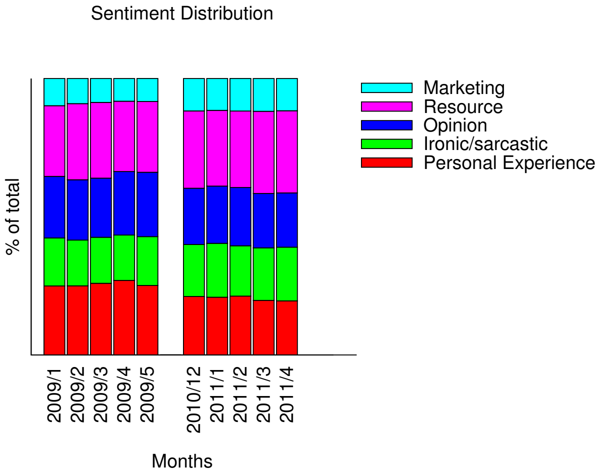
Methodology

- Active dengue surveillance based on four dimensions:
  - Public perception
  - Volume
  - Location
  - Time
- Methodology steps
  - Content analysis
  - Correlation analysis
  - Spatio-temporal analysis
  - Surveillance

Content analysis

- Determine the sentiment categories
  - **Personal experience:** "You know I have had dengue?"
  - **Ironic/sarcastic tweets:** "My life looks like a dengue-prone steady water"
  - **Opinion:** "The campaign against dengue is very cool"
  - **Resource:** "Dengue virus type 4 in circulation"
  - **Marketing:** "Everybody must fight dengue. Brazil relies on you"

Sentiment distribution over time

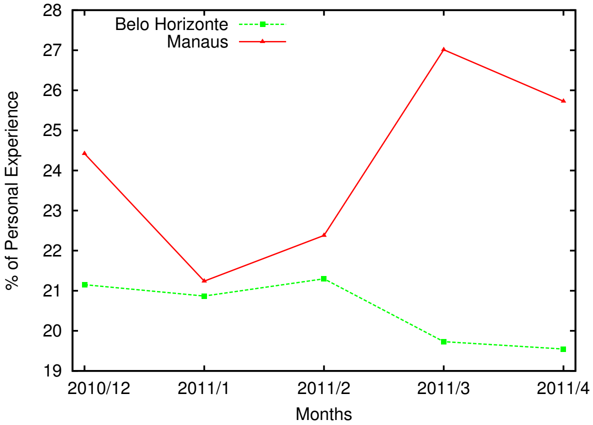


Content analysis

Data Mining

InWeb
KD@InWeb
Data Mining
Graph Mining
Dynamic behavior
Complex relationships
Heterogeneous data
Noisy data
Incomplete information
Lack of scalability
Data Science
Summary

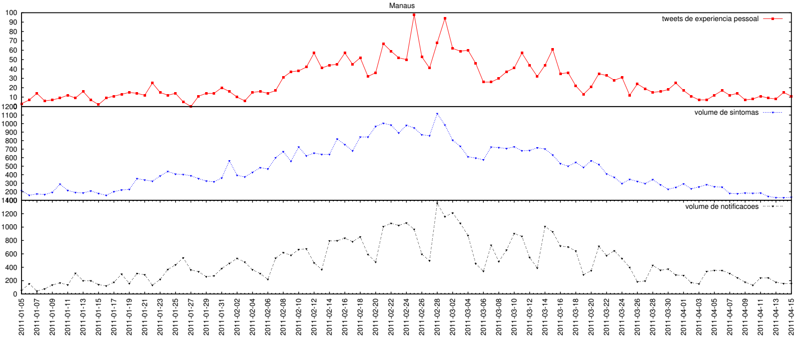
Is personal experience a good indicator of dengue's incidence?



Manaus

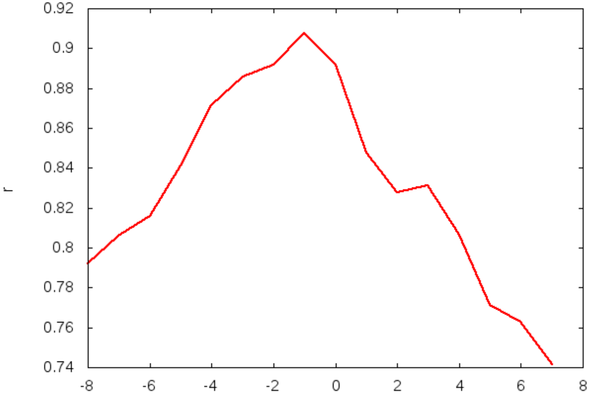
Personal experience, notifications and symptom perception

From November, 2010 to May, 2011



Manaus

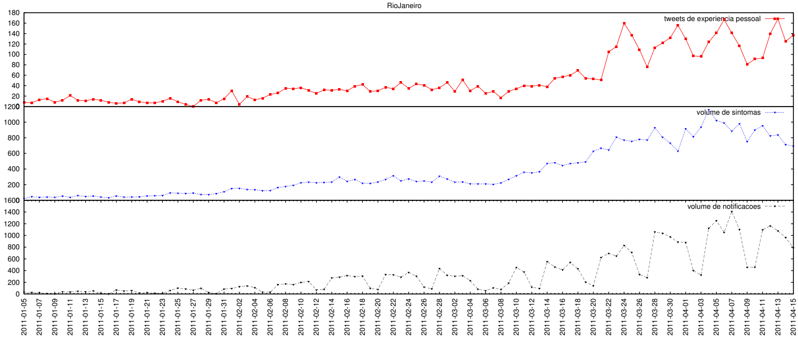
Cross-correlation between personal experience and symptom perception from November, 2010 to May, 2011



Rio de Janeiro

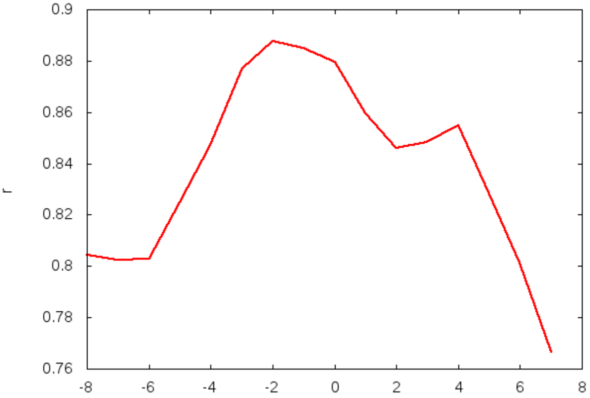
Personal experience, notifications and symptom perception

From November, 2010 to May, 2011



Rio de Janeiro

Cross-correlation between personal experience and symptom perception from November, 2010 to May, 2011



Spatio-temporal analysis

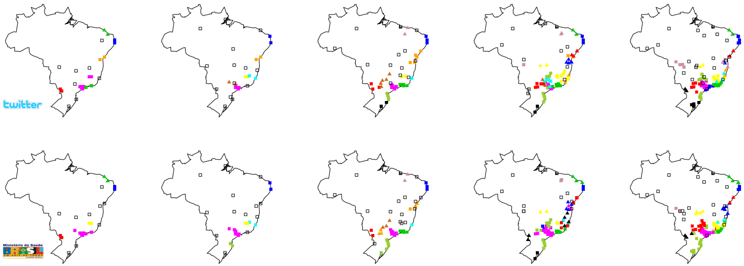
- Evaluated two metrics
  - the volume of tweets
  - the PTPE value

Spatio-temporal analysis

- Evaluated two metrics
  - the volume of tweets — *Rand Index* $= 0.8506$
  - the PTPE value — *Rand Index* $= 0.8914$

Spatio-temporal analysis

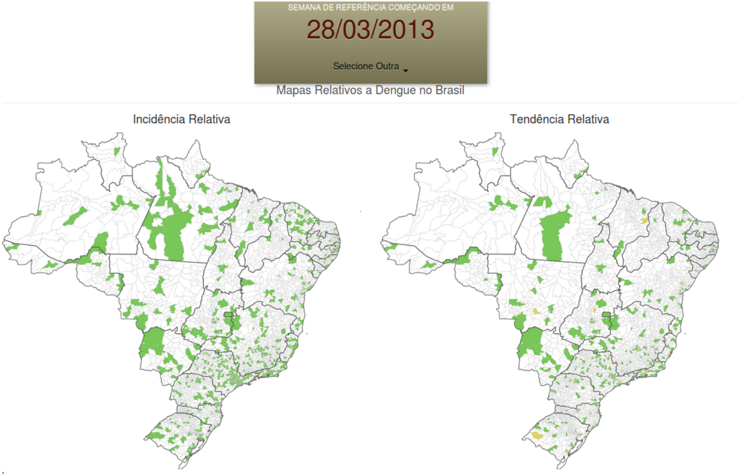
- Evaluated two metrics
  - the volume of tweets $\quad$ *Rand Index* $= 0.8506$
  - the PTPE value $\quad$ *Rand Index* $= 0.8914$



Surveillance

- Strategy: Analize the ratio of personal experience tweets weekely.
- Intuition: a sudden increase in this ratio indicates a surge
- Visual metaphors
  - maps
  - temporal graphs

Dengue Observatory



- Twitter data are useful for epidemics surveillance.
- Enablers:
  - Dengue is an urban disease, as it is the Internet usage in Brazil.
  - Dengue-related tweets are easy to collect.
  - People talk about dengue spontaneously.
- Tweets associated with "personal experience" present high correlation with dengue incidence.
- Simple alarm systems are effective to detect dengue surges.

Social Network Mining Challenges

Mining social networks must deal with:

- Dynamic behavior
- Complex relationships
- Heterogeneous data
- Incomplete information
- Noisy data
- **Lack of scalability**

Data mining algorithms are usually

- Irregular
- Intensive in terms of computing
- Intensive in terms of I/O

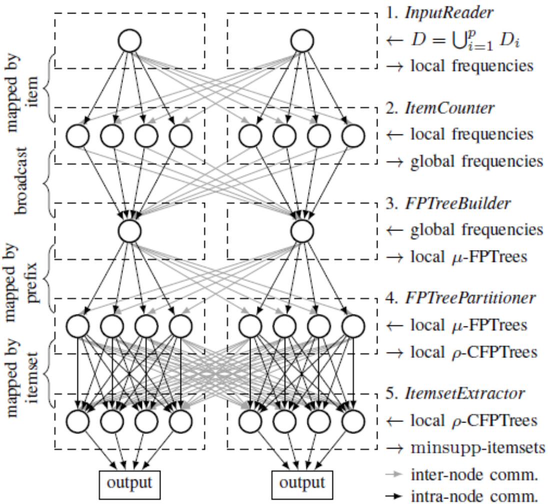
Hard to parallelize!

Twig: Adaptable, Scalable and Distributed FPM

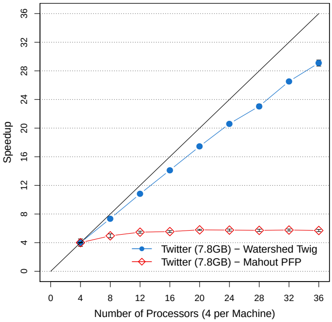
Strategy

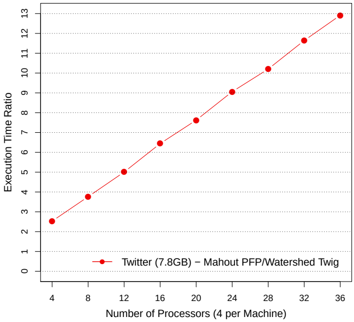
1. an adaptable and data-conscious partitioning scheme at the granularity of transactions which provides a complete and balanced distribution of the dataset, as well as of the tree that the algorithm builds and its associated projections, with a low communication overhead;
2. implementation in the filter-labeled stream paradigm, on top of the Watershed programming framework

Twig: Adaptable, Scalable and Distributed FPM



Twig X Mahout





How about Big Data?

Data Mining

- InWeb
- KD@InWeb
- Data Mining
- Graph Mining
- Dynamic behavior
- Complex relationships
- Heterogeneous data
- Noisy data
- Incomplete information
- Lack of scalability
- Data Science
- Summary

What's Big Data?

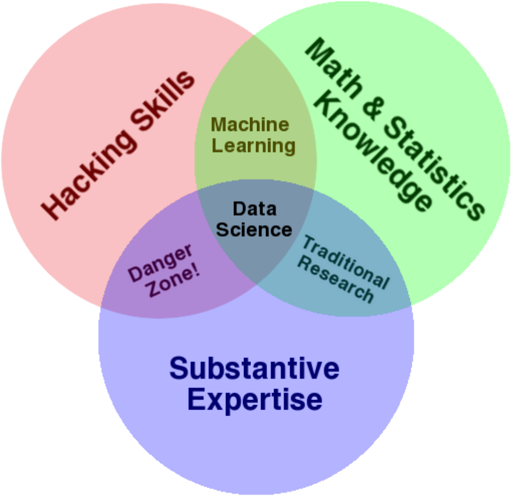
Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

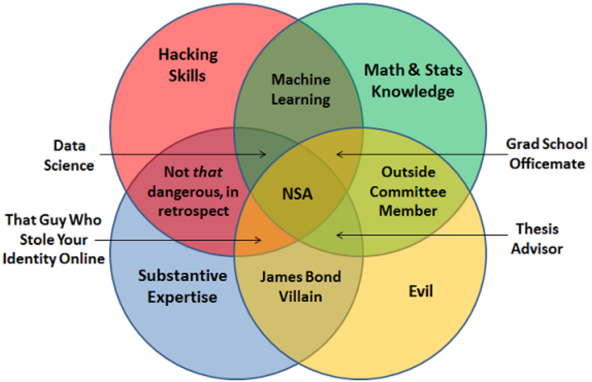
Dan Ariely

DATA

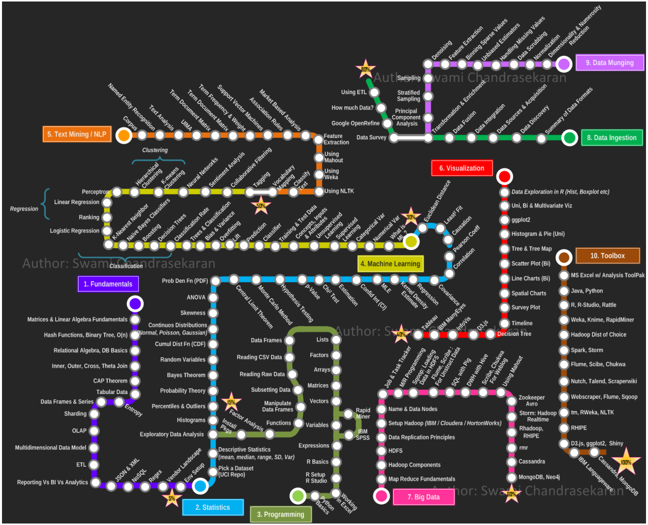

computer field
collection
information systems intelligence
statistical study
methods
recognition
based artificial
theory big
machine knowledge learning mining
terminology
programming
patterns
natural discovery ai machines
intelligent fields also
computing sets use
science
visualization
applications process
search
statistics large
analysis
data

What is Data Science?
Classical view




Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not that dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

Road to Data Science



Big Data or Big User?

Data Scientist

- Professional of the decade
- "Quants" from 80s, Software engineers from 90s e Web analysts from 00s

Profile

- Analytical ability
- Investigative capacity
- Entrepreneurship
- Business understanding
- **Programming skills**

Summary

- Problem demands evolve faster than we think.
- Maximizing quality and contributions is always a surviving strategy.
- Real problems help w.r.t. research relevance and enable innovation.
- Technically, big data has been here. The novelty is the big user.
- Data science formalizes the power shift to the big user.
- Data mining has plenty of room for research, development and innovation.

Thank you! Questions?

Wagner Meira Jr.
meira@dcc.ufmg.br