Social and Information Networks

Models and Machine Learning Methods

SEMINAR, UNI KL

NOV 28, 2015 – FEB 10, 2016

What? When? Where? Who?

What? You will apply machine learning to an interesting real-world large dataset

When? From Nov 28 to Feb 10

Where? MPI-SWS 112
networks-seminar@mpi-sws.org
http://learning.mpi-sws.org/networks-seminar

Who?

Manuel Gomez Rodriguez

Isabel Valera

Utkarsh Upadhyay

Project in Machine Learning

Machine learning (research) project on real-world large datasets

How does it work?

We give you access to:

1. **Large real-world datasets**
2. **Large machine(s)**
   (many cores, 512GB memory)
3. **Project ideas & mentoring**

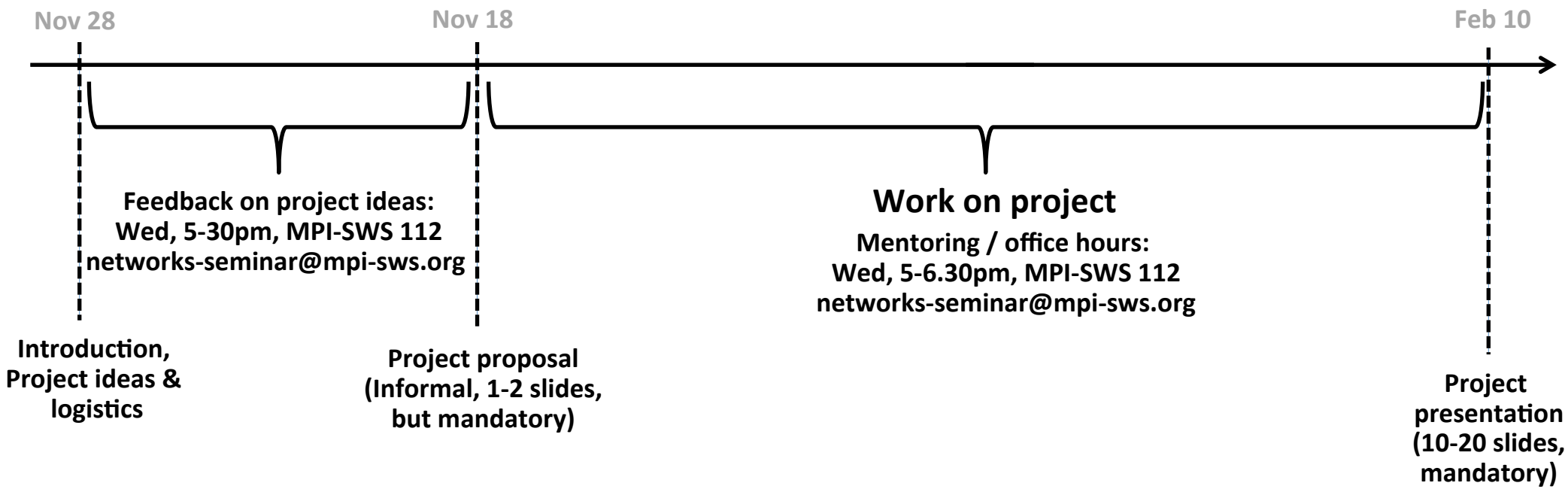
You provide:

1. **Project proposal**

Work

2. **Project presentation**

Timeline



more info at:

http://learning.mpi-sws.org/networks-seminar

Project Types (I)

This seminar is particularly focused in **research questions arising at the intersection of**

NETWORKS

INFORMATION

SOCIETY

with an emphasis on

THE WEB & SOCIAL MEDIA

Project Types (II)

Choose between two types of projects:

1. Data analysis and modeling

Discover insights about a real-world phenomena

2. Methods to solve an algorithmic problem

(a) Extend or improve some existing algorithm

(b) Develop new algorithm for a previously studied problem

(c) Define a new problem and develop an algorithm to solve it

Research Proposal

Project Idea

Dataset(s) you plan to use

Methodology you plan to use

We'll be happy to help you
(Office hours, Wed 5-6.30pm, MPI-SWS 112; networks-seminar@mpi-sws.org)

Datasets & Project Ideas

How to access the datasets?

Utkarsh will explain later

Multiple Datasets

News

MemeTracker

Product reviews

amazon

Knowledge

WIKIPEDIA

StackExchange

Mobility

NYC Taxi & Limousine Commission

Opinions

reddit

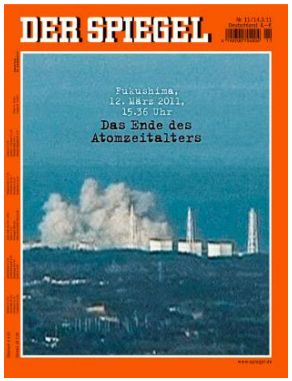
News dataset

MemeTracker

Memetracker: the dataset

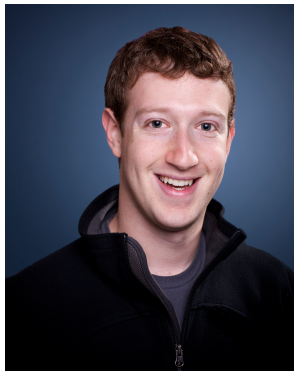
A **meme** is an **idea**, **behavior**, or **style** that spreads from person to person within a culture [Merriam-Webster].

The MemeTracker dataset contains short textual phrases that travel through the Web, which act as **tracers for memes** [Leskovec et al., KDD '09].



"Fukushima, March 12, 15.36: the end of the nuclear age"

"I just killed a pig and a goat"

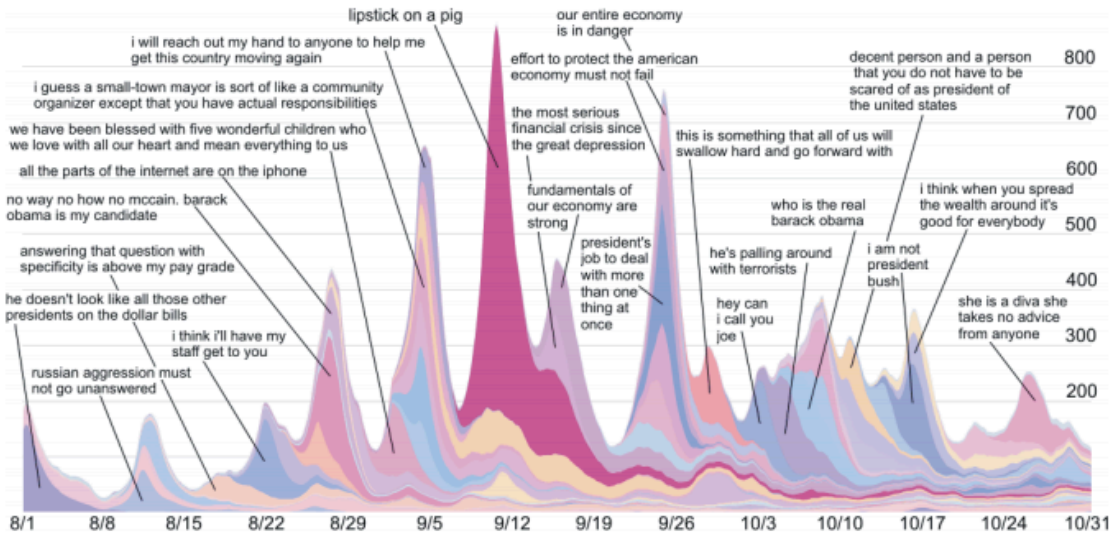


"You can put lipstick on a pig, but it's still a pig"

Memetracker: previous work examples

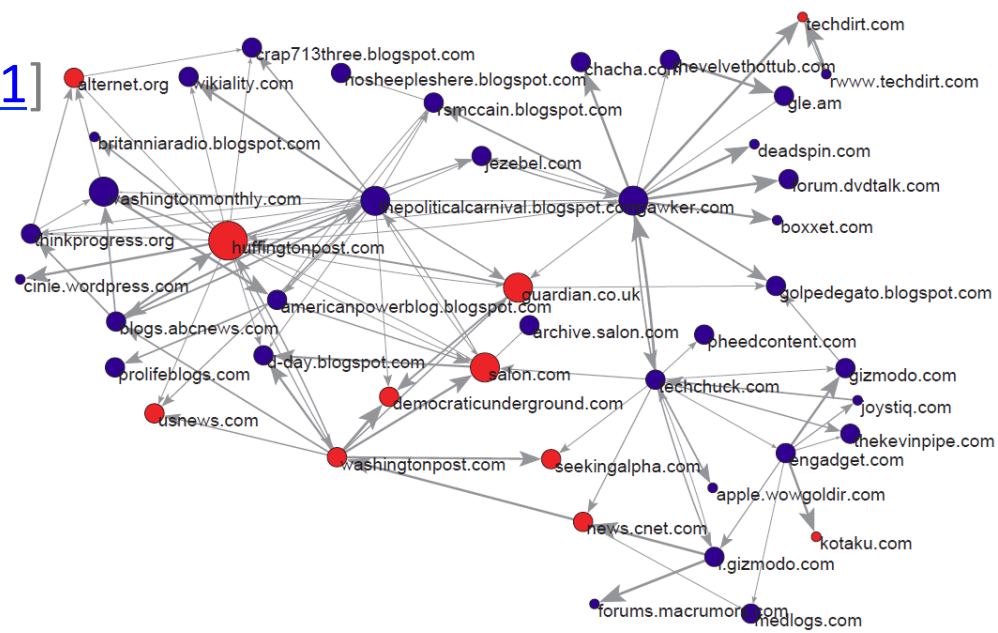
[Leskovec et al., KDD '09]

Insights on the news lifecycle



[Gomez-Rodriguez et al., ICML '11]

Inferring information networks from textual traces



Memetracker: project ideas (I)

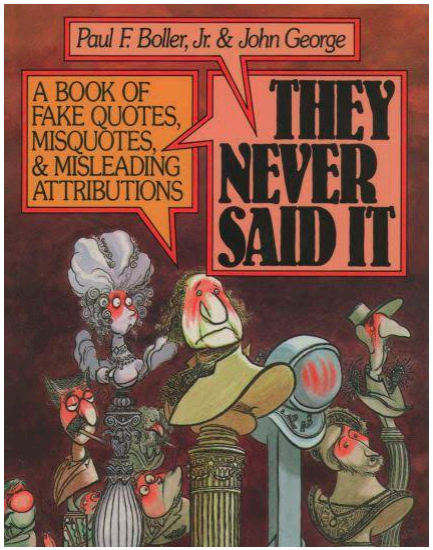
Investigate how memes **change and mutate** over time:

CBSNEWS

Mitt Romney attack ad misleadingly quotes Obama

Slate

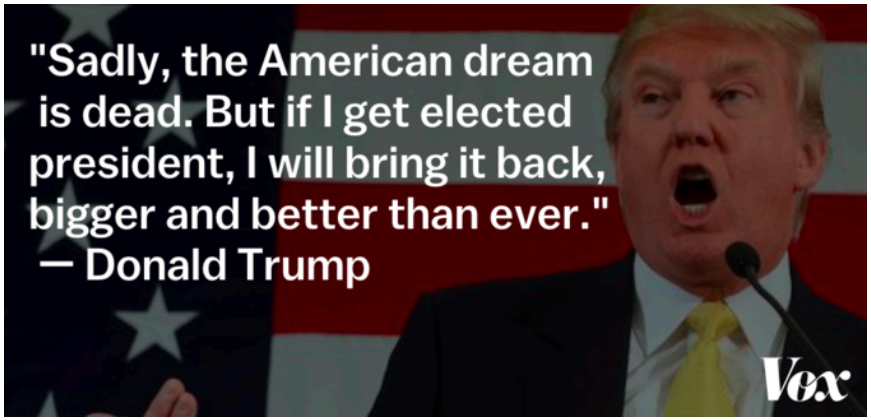
Journalists: Stop Passing Off a Paraphrase as a Direct Quotation



1. **Cluster mutations** of the same meme
2. **Mutation model, which predicts when a quote will mutate**
3. Distinguish **intentional from unintentional** mutations
4. Find misleading **memes used out of context**
5. Find **websites that systematically change memes**

Memetracker: project ideas (II)

Memes are often **parts of a story that unfolds over time** (e.g., news or event)



1. **Story timeline reconstruction** based on memes
2. Investigate **partial views of a story provided by a website(s)**
3. Measure **bias & polarization** in partial views
4. **Find websites that provide bias & polarized partial view**
5. **Find influential memes** in a story line?
6. Connecting memes, even if a site reproduce all memes, does it provide a **balance view**?

Knowledge datasets

WIKIPEDIA

StackExchange

Wikipedia: the dataset

WIKIPEDIA is a **free encyclopedia** built **collaboratively** using wiki software [Wikipedia].

Big data

From Wikipedia, the free encyclopedia

Tim Berners-Lee

From Wikipedia, the free encyclopedia

Germanwings Flight 9525

From Wikipedia, the free encyclopedia

Germany at the FIFA World Cup

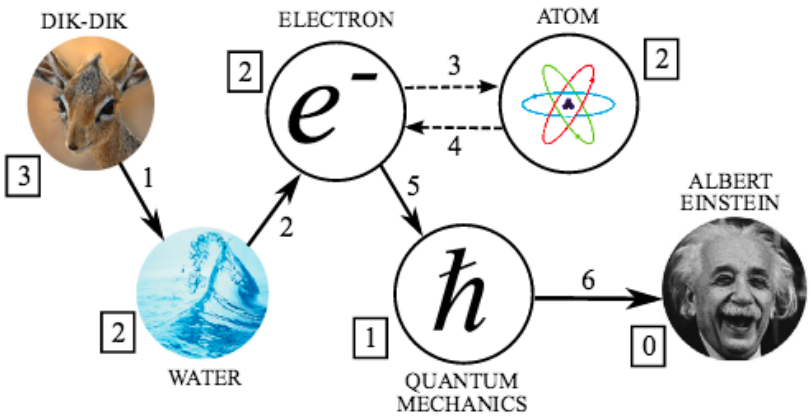
From Wikipedia, the free encyclopedia

The Wikipedia dataset contains who edits **what** and **when** [G. Kossinets].

- (cur | prev) 22:33, 17 October 2015 2602:304:28aa:d5f0:64d0:527d:38b6:cff (talk) . . (80,130 bytes) (+1) . . (undo)
- (cur | prev) 22:32, 17 October 2015 2602:304:28aa:d5f0:64d0:527d:38b6:cff (talk) . . (80,129 bytes) (-7) . . (→Retail) (undo)
- (cur | prev) 14:00, 16 October 2015 McGeddon (talk | contribs) . . (80,136 bytes) (-236) . . (Reverted 1 edit by Dharnett21 (talk): WP:PRIMARY-sourced / promotion. (TW)) (undo)
- (cur | prev) 13:44, 16 October 2015 Dharnett21 (talk | contribs) . . (80,372 bytes) (+236) . . (→Technology: add a new example of how Big Data can be used in mobile games) (undo) (Tag: VisualEditor)
- (cur | prev) 16:43, 13 October 2015 McGeddon (talk | contribs) . . (80,136 bytes) (-106) . . (→Critique: cut a Wikipedia editor's joke cartoon) (undo)
- (cur | prev) 05:07, 13 October 2015 BG19bot (talk | contribs) m . . (80,242 bytes) (+27) . . (WP:CHECKWIKI error fix for #61. Punctuation goes before References. Do general fixes if a problem exists. - using AWB (11700)) (undo)
- (cur | prev) 01:28, 13 October 2015 Kuru (talk | contribs) . . (80,215 bytes) (-1,784) . . (this seems very redundant with the rest of the article) (undo)
- (cur | prev) 21:34, 12 October 2015 Preyansh07 (talk | contribs) m . . (81,999 bytes) (+1,784) . . (Added a new sub section on Big Data Analytics, as this is very relevant to this article on Big Data and there is no separate article on the former topic.) (undo) (Tag: VisualEditor)
- (cur | prev) 11:20, 12 October 2015 Kuru (talk | contribs) . . (80,215 bytes) (-184) . . (rmv promotional external link) (undo)
- (cur | prev) 10:07, 12 October 2015 121.244.36.65 (talk) . . (80,399 bytes) (+1) . . (undo)
- (cur | prev) 10:06, 12 October 2015 121.244.36.65 (talk) . . (80,398 bytes) (-57) . . (undo)
- (cur | prev) 10:03, 12 October 2015 121.244.36.65 (talk) . . (80,455 bytes) (+240) . . (undo)
- (cur | prev) 09:59, 12 October 2015 121.244.36.65 (talk) . . (80,215 bytes) (-362) . . (undo)

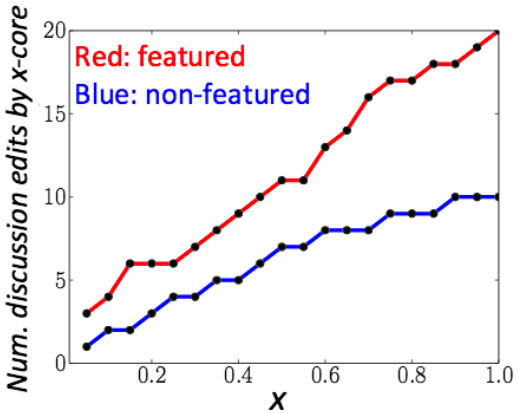
Wikipedia: previous work examples

[West et al., WWW '12]
Human wayfinding in information networks



One of the optimal shortest path:
<Dik-Dik, Water, Germany, Albert Einstein>

[Romero et al., ICWSM '15]
Coordination and efficiency in decentralized collaboration



More coordination between editors leads to higher quality pages

Wikipedia: project ideas

Who contributes **what** in Wikipedia? Investigate knowledge sharing by ***wikipedians***

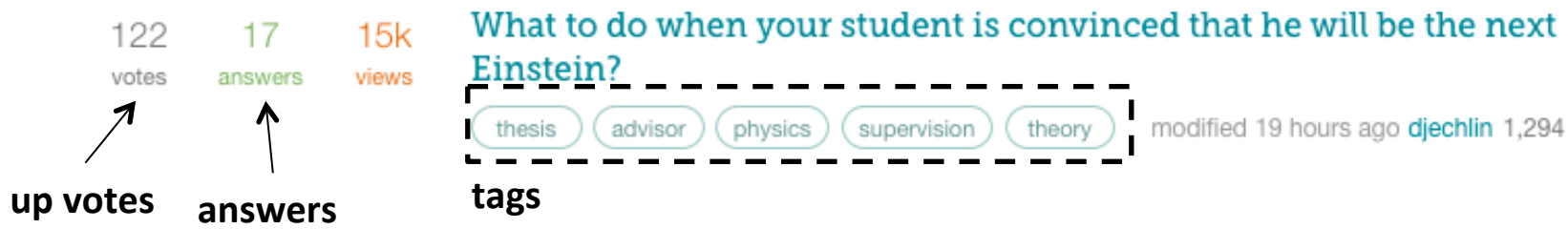
1. **How much & how diverse** is the knowledge contributed by the **typical Wikipedian**? How frequently does he contribute?

2. Does knowledge typically **survive**? Reliability, controversy and relevancy

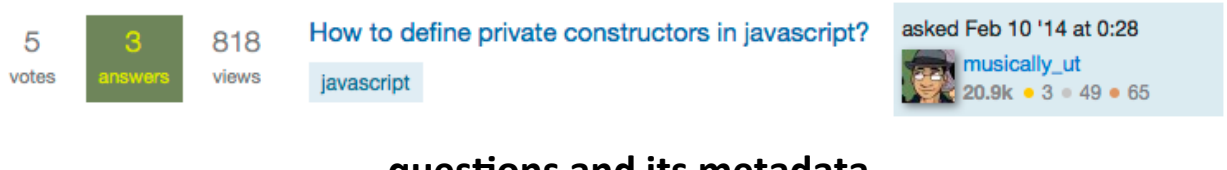
3. Many **wikipedians include themselves** in several **categories (e.g., art lover)**. Do they contribute about them? Are they **faithful**?

StackExchange: the dataset

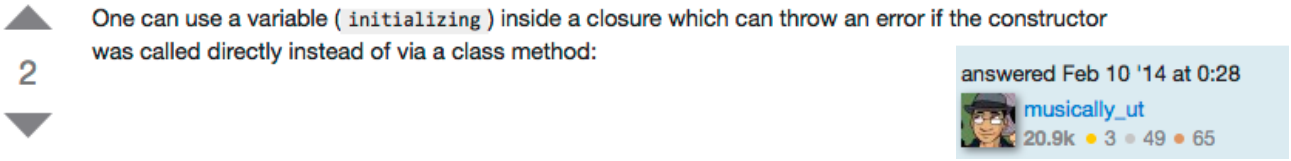
StackExchange is a **question answering site**, where questions, answers, and users are subject to a **reputation award process** [Wikipedia].



The StackExchange dataset contains:



questions and its metadata



answers and their metadata

user badges

StackExchange: previous work examples

[Posnett et al., SocInf '12]

Mining stack exchange: expertise is evident from initial contributions

Jon Skeet 401 5712 6847 member for: 7 years	#5 year rank	-1 change	815,954 total reputation	76,877 year reputation
Darin Dimitrov 119 2158 2159 member for: 7 years	#18 year rank	-4 change	618,710 total reputation	57,990 year reputation
BalusC 158 2034 2323 member for: 6 years, 2 months	#9 year rank	+1 change	603,009 total reputation	72,168 year reputation
Hans Passant 66 735 1337 member for: 7 years, 1 month	#10 year rank	-4 change	582,562 total reputation	71,638 year reputation
Marc Gravell 125 1549 2043 member for: 7 years	#31 year rank	-15 change	564,987 total reputation	52,012 year reputation
VonC 140 1433 1524 member for: 7 years, 1 month	#6 year rank	-1 change	531,625 total reputation	75,923 year reputation

[Anderson et al., WWW '13]

Steering user behavior with badges

Nice Question	Question score of 10 or more	285.2k awarded
Good Question	Question score of 25 or more	81.7k awarded
Great Question	Question score of 100 or more	11.7k awarded
Enthusiast	Visit the site each day for 30 consecutive days. (Days are counted in UTC.)	107.1k awarded
Fanatic	Visit the site each day for 100 consecutive days. (Days are counted in UTC.)	16.7k awarded
Tenacious	Zero score accepted answers: more than 5 and 20% of total	31.8k awarded
Unsung Hero	Zero score accepted answers: more than 10 and 25% of total	12.1k awarded
Self-Learner	Answer your own question with score of 3 or more	73.8k awarded

StackExchange: project ideas

Understand the **dynamics of question answering** on StackExchange

Image by : opensource.com

See a question that's not getting any good answers?

Place a bounty on it!

1. Can we identify (and predict) when **a question** is ***fully answered***? How do **answers complement each other**?

2. Investigate to which extent **users provide answers to unanswered questions** or **complete other answers**?

3. Are **badges and bounties really encourage knowledge sharing**? Which are more valuable?

Opinions (and more) dataset

reddit

Reddit: the dataset

reddit is an entertainment, social networking, and news website where **community members can submit content**; it is an **online bulletin board system**. [Wikipedia]

2328

TIL Buzz Aldrin's mother's maiden name was, coincidentally, Moon. (buzzaldrin.com)
submitted 12 hours ago by namebar115
66 comments share

770

Politics Americans show "a surprising willingness" to accept internet surveillance (dailydot.com)
submitted 1 day ago by dogdays12304
179 comments share

The Reddit Dataset contains:

subreddit threads and its metadata

32

TMAF about the universe self.TellMeAFact
Submitted 16 hours ago by arainbowpony
19 comments share

[–] bl0bfish 10 points 10 hours ago*
The further the galaxy is away from us that we see, the further back in time we are actually seeing.

[–] halocupcake 6 points 7 hours ago
Stars can be above a billion times the size of our sun. Source: https://en.m.wikipedia.org/wiki/UY_Scuti

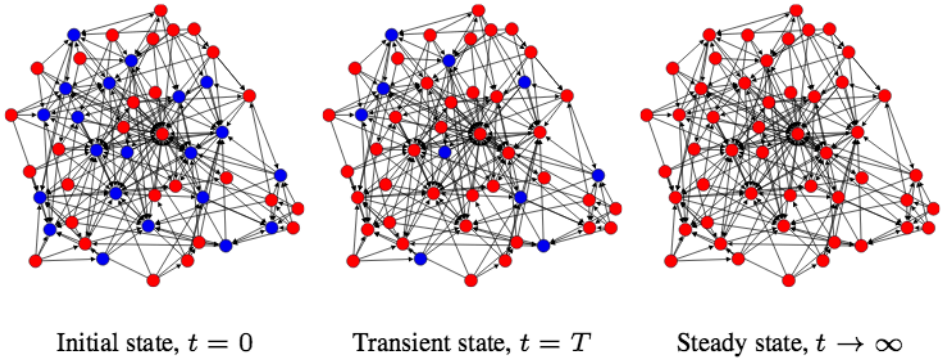
Reddit: previous work examples

[Lakkaraju et al., ICWSM '13]
What's in a name? Understanding the interplay between titles, content, and communities in social media

BUSINESS INSIDER

Computer Scientists Figured Out How To Execute The Perfect Reddit Submission

[De et al., Arxiv '15]
Learning opinion dynamics in social networks



Understanding how Reddit users influence each other's opinion

Reddit: project ideas

Understand **opinion and sentiment dynamics through Reddit**

1. **Temporal evolution of sentiment and opinions** subreddits and comments threads? **Polarization, consensus and controversy**.

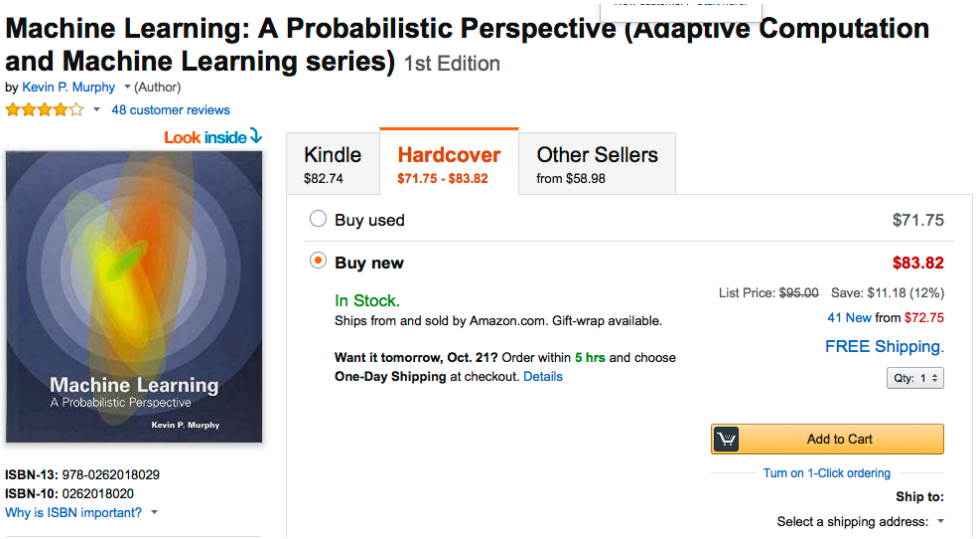
2. Are people more willing to share **controversial content using anonymous accounts**? Where is the **troll boundary**?

3. In reddit, **moderators** are not chosen by reddit, but **chosen by the *crowd***. Investigate the emergence of moderators, characteristic patterns & subreddit governance.

Product reviews dataset amazon

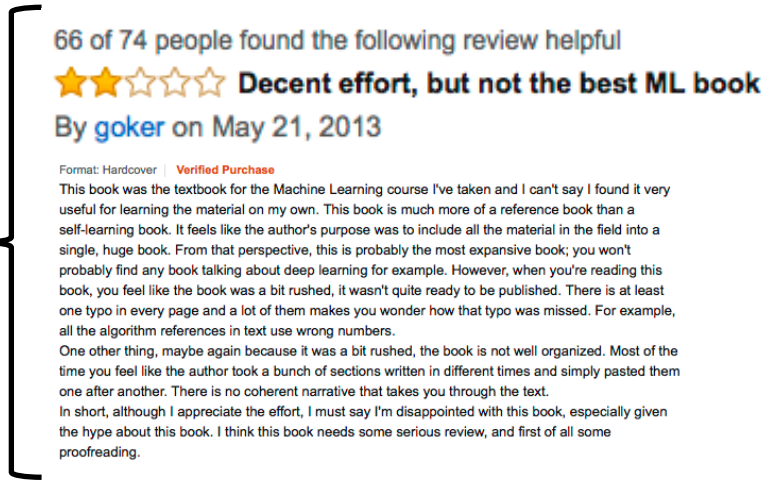
Amazon: the dataset

amazon is an American electronic commerce selling a wide variety of products, from books and DVDs to jewelry [Wikipedia]



The Amazon dataset contains:

reviews



66 of 74 people found the following review helpful

Decent effort, but not the best ML book

By goker on May 21, 2013

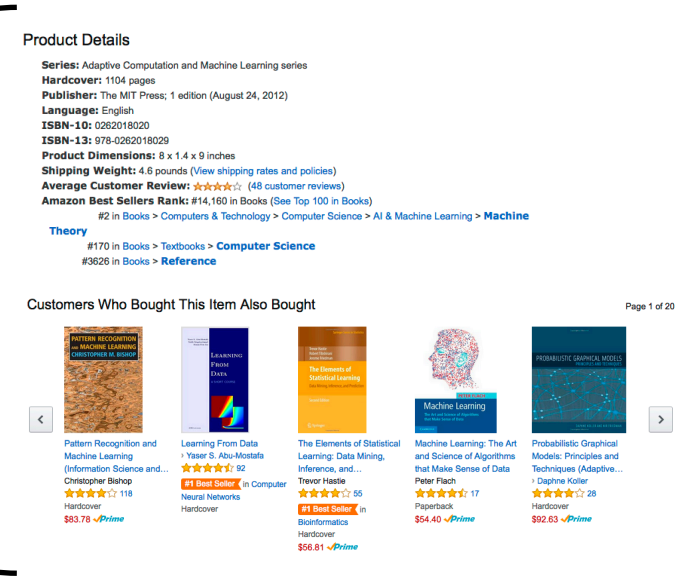
Format: Hardcover | Verified Purchase

This book was the textbook for the Machine Learning course I've taken and I can't say I found it very useful for learning the material on my own. This book is much more of a reference book than a self-learning book. It feels like the author's purpose was to include all the material in the field into a single, huge book. From that perspective, this is probably the most expansive book; you won't probably find any book talking about deep learning for example. However, when you're reading this book, you feel like the book was a bit rushed, it wasn't quite ready to be published. There is at least one typo in every page and a lot of them makes you wonder how that typo was missed. For example, all the algorithm references in text use wrong numbers.

One other thing, maybe again because it was a bit rushed, the book is not well organized. Most of the time you feel like the author took a bunch of sections written in different times and simply pasted them one after another. There is no coherent narrative that takes you through the text.

In short, although I appreciate the effort, I must say I'm disappointed with this book, especially given the hype about this book. I think this book needs some serious review, and first of all some proofreading.

metadata



Product Details

Series: Adaptive Computation and Machine Learning series
Hardcover: 1104 pages
Publisher: The MIT Press; 1 edition (August 24, 2012)
Language: English
ISBN-10: 0262018020
ISBN-13: 978-0262018029
Product Dimensions: 8 x 1.4 x 9 inches
Shipping Weight: 4.6 pounds (View shipping rates and policies)
Average Customer Review: (48 customer reviews)
Amazon Best Sellers Rank: #14,160 in Books (See Top 100 in Books)
#2 in Books > Computers & Technology > Computer Science > AI & Machine Learning > **Machine Theory**
#170 in Books > Textbooks > **Computer Science**
#3626 in Books > **Reference**

Customers Who Bought This Item Also Bought

Page 1 of 20

Pattern Recognition and Machine Learning (Information Science and... Christopher Bishop 118 Hardcover $83.78 Prime

Learning From Data › Yaser S. Abu-Mostafa 92 #1 Best Seller in Computer Neural Networks Hardcover

The Elements of Statistical Learning: Data Mining, Inference, and... Trevor Hastie 55 #1 Best Seller in Bioinformatics Hardcover $56.81 Prime

Machine Learning: The Art and Science of Algorithms that Make Sense of Data Peter Flach 17 Paperback $54.40 Prime

Probabilistic Graphical Models: Principles and Techniques (Adaptive... › Daphne Koller 28 Hardcover $92.63 Prime

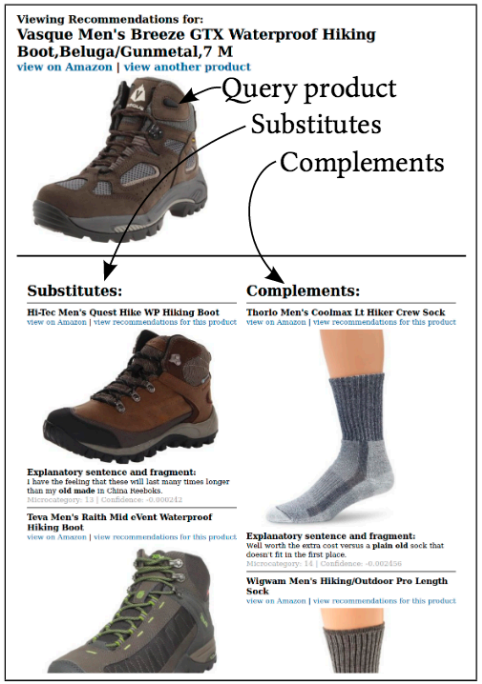
Amazon: previous work examples

[McAuley et al., SIGIR ’15]

Image-based recommendations on styles and substitutes

[McAuley et al., KDD ’15]

Inferring networks of substitute and complementary products



Amazon: project ideas

Estimate **brand reputation and business intelligence from analyzing Amazon reviews**

"Almost two-thirds of companies say that social media is a significant or critical risk to their brand reputation"
altimetergroup.com

1. **Brand reputation and lack of choices** may boost sales of products with not particularly positive reviews. Can we quantify to which extent?

2. Can we **identify the products a brand got its reputation from**? How is this **reputation spilling to other products** within the brand?

3. Products are often bought together. Can we identify which **products a brand should expand on,** by identifying products from other brands bought together with products of the brand?

Mobility dataset

NYC Taxi & Limousine Commission

NY Taxi: dataset

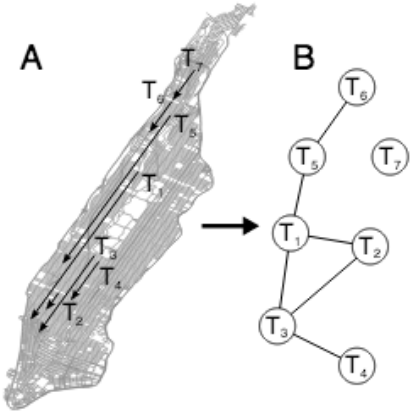
NYC Taxi & Limousine Commission serves **600,000 passengers per day (236 Million/year)** with >50,000 drivers. A typical taxi travels **70,000 miles per year** [NYC TLC]

The TCL Trip Record dataset contains, for each trip:

(1) pick-up and drop-off dates/times
(2) pick-up and drop-off locations
(3) trip distances
(4) itemized fares, rate types, payment types
(5) driver-reported passenger counts

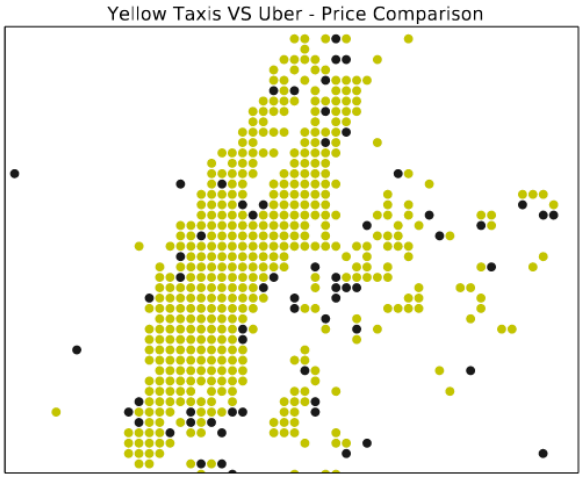
NY Taxi: previous work examples

[Santi et al., PNAS '14]
Algorithm to optimally share taxis across people to increase sustainability



Weighted maximum graph matching

[Salnikov et al., NetMob '15]
OpenStreetCab: exploiting taxi mobility patterns in NYC to reduce commuter costs



Most of the time, taxis are cheaper than Uber

NY taxi: project ideas

Estimate **optimal taxi routing from analyzing fine-grained GPS taxi data**

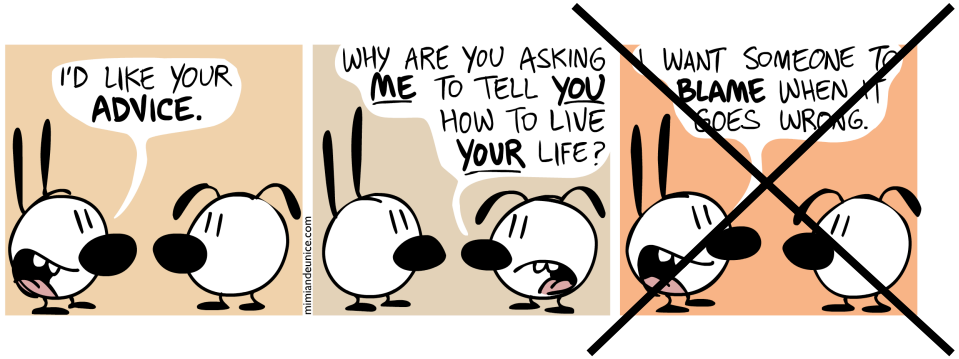
1. **Characterizing taxi drivers' abilities** based on fine-grained GPS data. What is the strategy of the wealthiest taxi drivers?

2. Develop an **algorithm that provides smart routing for taxi drivers** to maximize profit, low stress, or sustainability?

3. What about an algorithm to do **smart global routing of a taxi fleet**?

What do you need to learn?

It **depends** largely on the **type of project you choose!**



Choose a project idea and dataset → **We can advise you afterwards**

Next steps

Find a project idea:	**From Oct 28 to Nov 18**
Receive feedback:	**Nov 4, 11 & 18 (In class)** **From Oct 28 to Nov 18 (e-mail)**
Project proposal: **(Informal, 1-2 slides)**	**Nov 18 (In class)**

reach us at:

networks-seminar@mpi-sws.org

http://learning.mpi-sws.org/networks-seminar