Big Data and Its Implication to Research Methodologies and Funding

Cornelia Caragea

TARDIS 2014

November 7, 2014

MLg Machine Learning Group
UNT Computer Science and Engineering

Data Everywhere

- Lots of data is being collected and warehoused
  - News articles and news comments
  - Weblogs, e-commerce data, customer reviews, forum threads
  - Scientific documents
    - PubMed currently has over 24 million documents
    - Google Scholar is estimated to have 160 million documents
  - Social network data
    - Facebook passes 1.23 billion monthly active users, 945 million mobile users, and 757 million daily users
    - Twitter usage: 284 million monthly active users, 500 million tweets sent per day, 80% of Twitter active users are on mobile

Big Data - What is it?

- A term used to describe the exponential growth and availability of data in almost all domains.
- Types of data:
  - Relational data (Tables / Transaction)
  - Text data (Web)
  - Semi-structured data (XML)
  - Graph data
    - Social networks, Semantic Web (RDF)
  - Streaming data

The Three Vs of Big Data

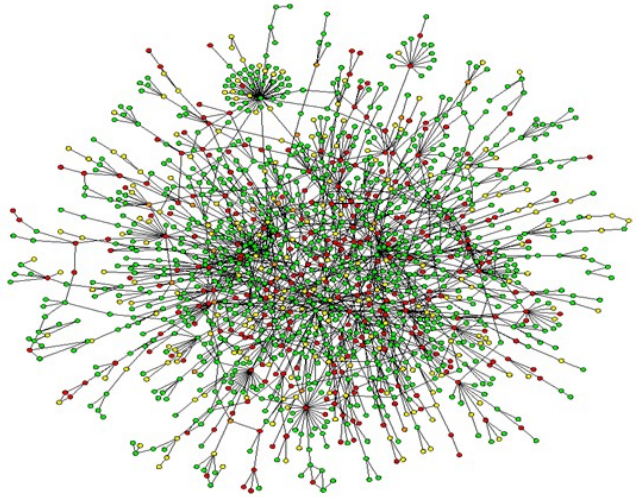
- Volume – Huge data volumes stored
  - Due, in part, to the decrease of storage costs.
- Velocity – Drink from a fire hose!
  - Data is now streaming in at unprecedented speed
    - It must be dealt with in a timely manner.
- Variety – Large number of diverse data sources to integrate
  - Data comes in all forms:
    - structured
    - numeric data such as weather data, sensor data
    - unstructured text
    - video
    - audio
    - …

Big Data - Why it Matters?

- Big Data has become as important as the Internet
  - A source of great benefits to discovery, learning, and staying informed
  - It may lead to more accurate analyses, leading to more confident decision making
    - Better decisions => reduced risk / improved revenue.
      - Quickly identify customers who matter the most.
      - Detect fraudulent behavior in real time.
  - It can make our lives easier and more productive, if we can infer the relationships of interest to us from the data.
  - It has the potential to save lives during disaster events.

Potential Pitfalls of Big Data

Defend Your Private Information from Unauthorized Access by Using Data Security Software



ENCRYPTSTICK™
When digital privacy matters

Big Data - Challenges

- To make effective use of these data
  - E.g., how to reduce false positives in medical data, which may lead to unnecessary surgeries.
  - Need to understand how to mine it effectively
    - Do we store it all?
    - Do we analyze it all?
    - How can we mine it to our best advantage?
    - What if the data volume is so large and varied that we do not know how to deal with it?
  - How to ensure sensitive information cannot be inferred

Implications of Big Data on Research Methodologies

- Move from simple (SQL) analytics to complex (non-SQL) analytics:
  - Knowledge Discovery
    - Discovery of useful, possibly unexpected, patterns in data
      - Data Mining
      - Machine Learning
- Incorporate massive data and modalities in analysis
  - Use high-performance technologies, e.g., Hadoop, MapReduce, etc.
- Determine upfront which data is relevant

Example Scenarios using Big Data

- Extracting Keyphrases from Document Networks
- Understanding Disaster Events through Social Media
- Analyzing Images' Privacy for the Modern Web

Extracting Keyphrases from Document Networks

Project funded by NSF

Why Keyphrase Extraction?

- Keyphrase extraction is the task of automatically extracting descriptive phrases or “concepts” from a document.

- Keyphrases:
  - Allow for *efficient processing of more information in less time*
  - Are useful in many applications:
    - topic tracking, information filtering and search, classification, clustering, and recommendation.

Previous Approaches to Keyphrase Extraction

- Use generally only the textual content of the target document [Mihalcea and Tarau, 2004], [Liu et al., 2010].
- Recently, models are proposed that incorporate a local neighborhood of a document [Wan and Xiao, 2008].
  - Obtained improvements over models that use only textual content.
  - However, their neighborhood is limited to textually-similar documents.

During these "Big Data" times – access to giant document networks

- *In addition to a document's textual content and textually-similar neighbors, are there other informative neighborhoods that exist in research document collections?*
- *Can these neighborhoods improve keyphrase extraction?*

From Data to Knowledge

- A typical scientific research paper:
  - Proposes new problems or extends the state-of-the-art for existing research problems
  - Cites relevant, previously-published papers in appropriate *contexts*

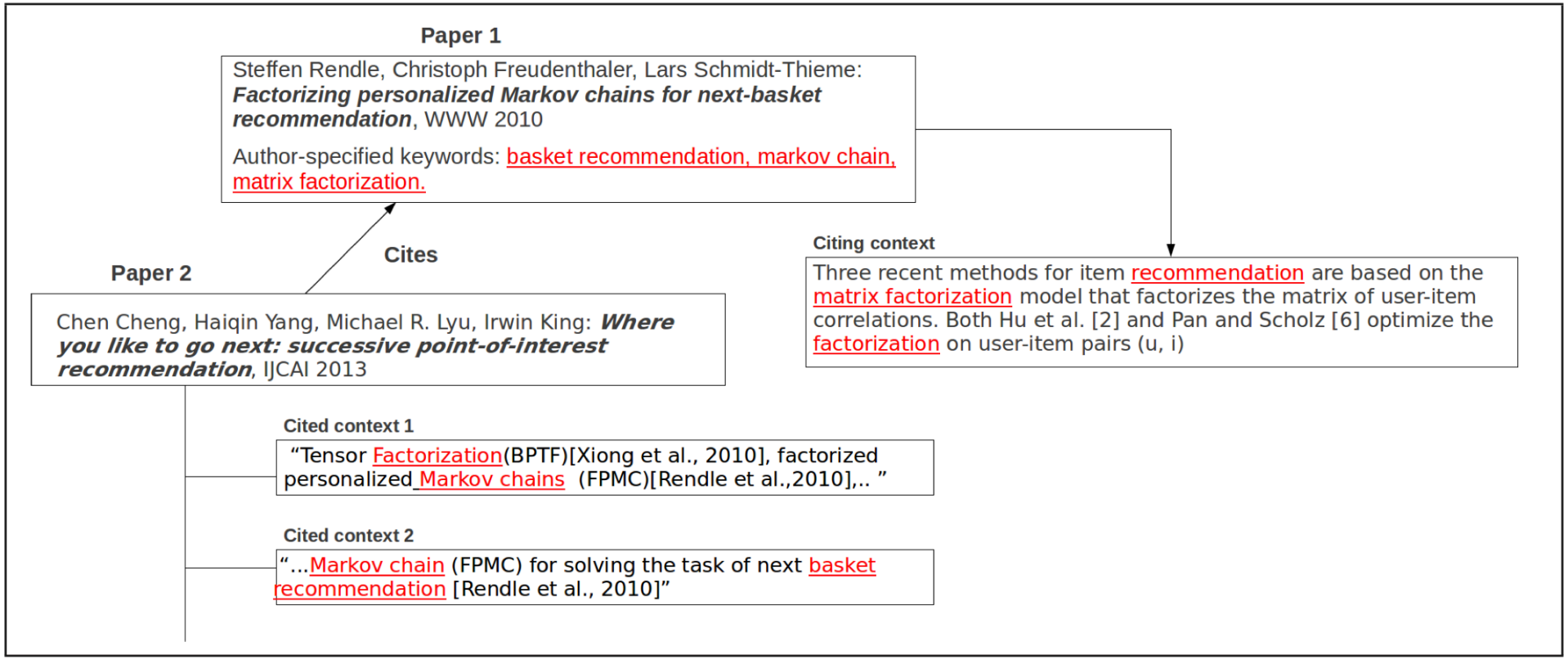
- The citations between research papers gives rise to an interlinked document network, commonly referred to as the *citation network*.

Citation Networks

- In a citation network, information flows from one paper to another via the citation relation [Shi et al., 2010].

- The influence of one paper on another as well as the flow of information are captured by means of citation contexts (short text segments surrounding a paper's mention)
  - They serve as "micro summaries" of a cited paper!

A Small Citation Network



- *Citation contexts* are very informative!

[Das G. and Caragea, 2014]; [Caragea et al., 2014]

Understanding Disaster Events through Social Media

Project funded by NSF

Social Media

- Social media is now part of our daily lives and everyday communication patterns.
- Scholars of disasters see hope in social media
  - Used around crises, it can produce accurate results, often in advance of official communications.
- However, social media data has not been incorporate much in emergency response systems.

Proof of Concept

- Using Twitter data from Hurricane Sandy, we identify the sentiment of tweets and then measure the distance of each categorized tweet from the epicenter of the hurricane.

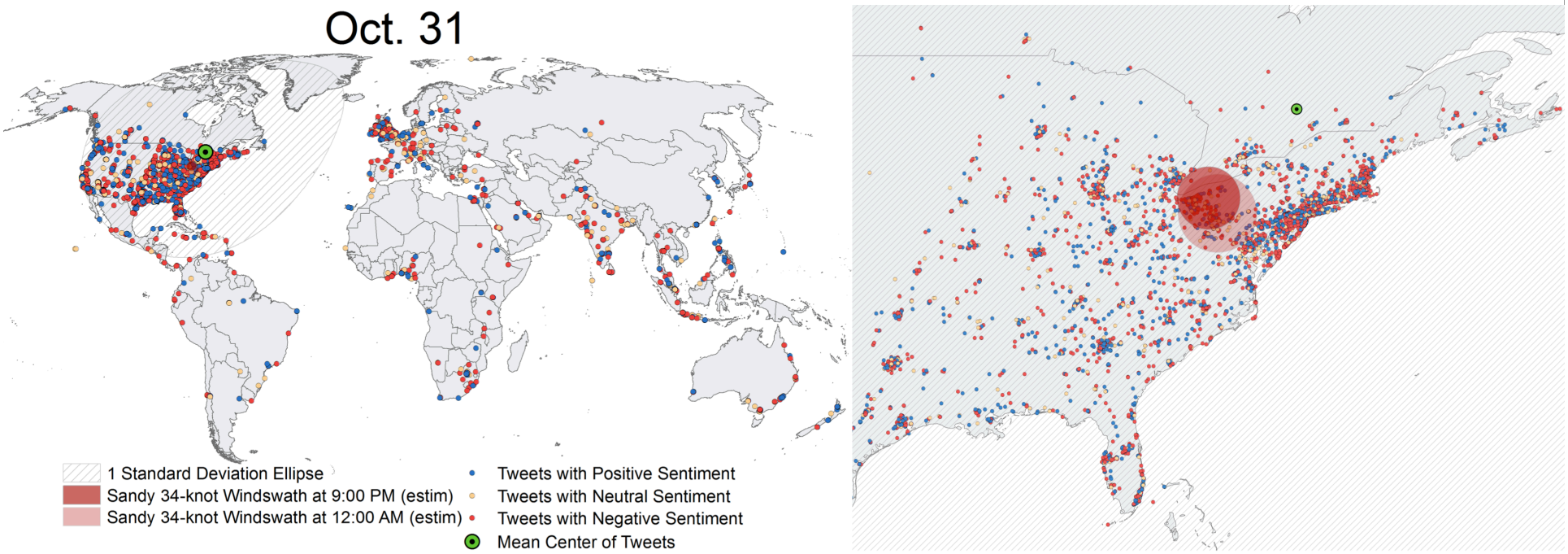
- Sandy Hurricane Twitter Data:
  - We crawled 12,933,053 tweets between 10-26-2012 and 11-12-2012.

Why Sentiment Analysis in Disaster Events?

- Can help understand the dynamics of the social network
  - The main users' concerns and panics
  - The emotional impacts of interactions among users.
- Can help obtain a holistic view about the general mood and the situation on the ground.

- Strong value to those experiencing the disaster and those seeking information about the disaster, as well as to the responder organizations.
  - Extracting sentiments during a disaster could help responders develop stronger situational awareness of the disaster zone itself.

Geo-Tagged Tweets Sentiment Analysis



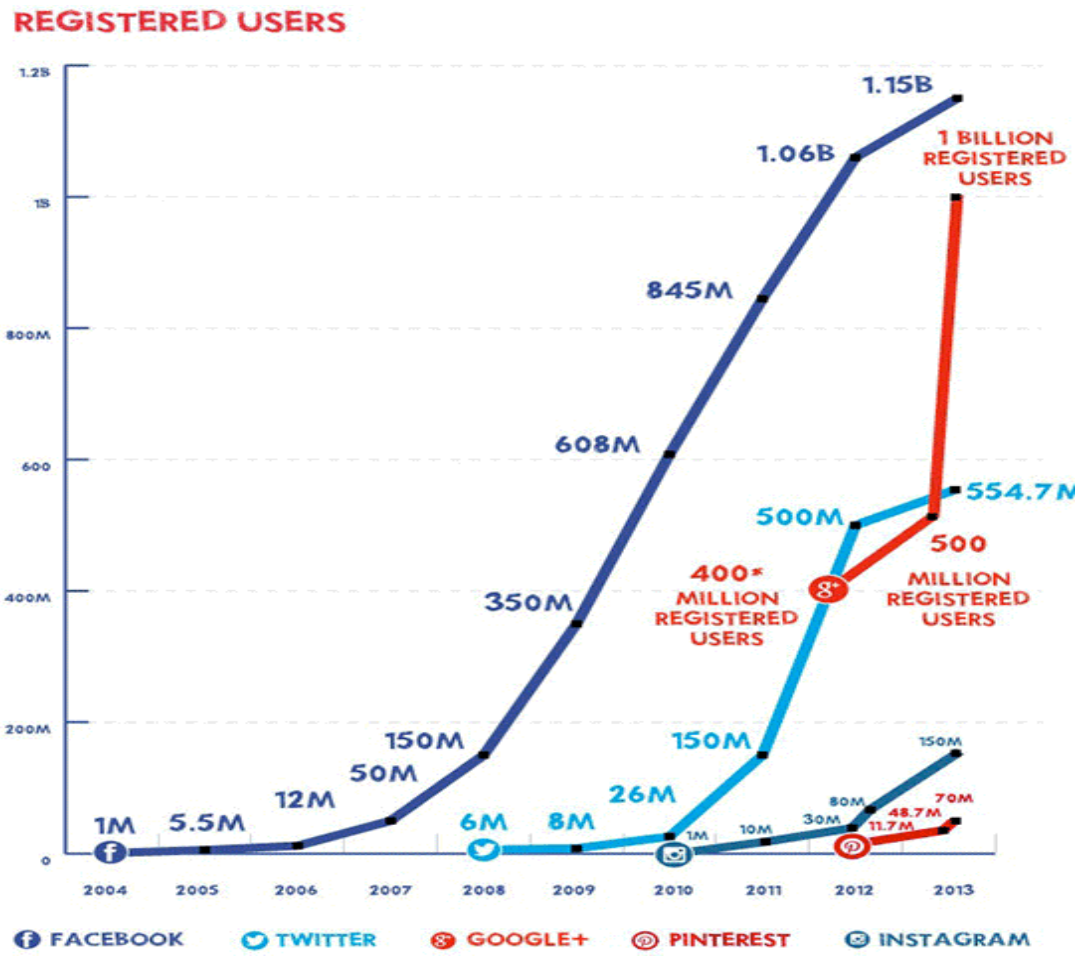
- Could be integrated into systems to help response organizations have a real time map to display the physical disaster and the spikes of intense emotional activity in its proximity.
- Using “Big Data”:
  - Automatically infer tweets geo-location
  - Automatically identifying trustworthy information spread around disaster events

[Caragea, Squiciarrini, Stehle, Neppalli, Tapia; 2014]

Analyzing Images' Privacy for the Modern Web

Project funded by NSF

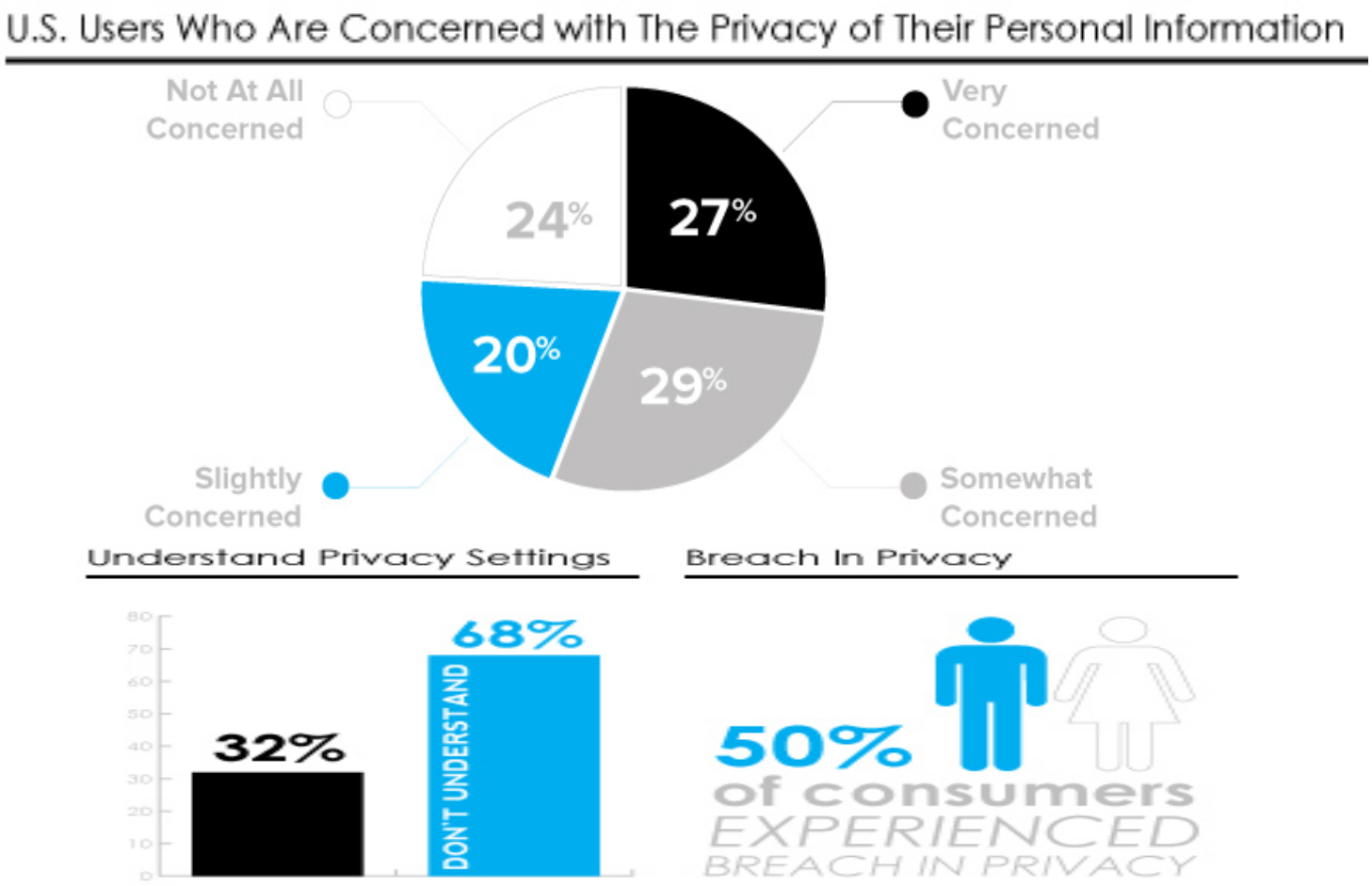
Why Online Image Privacy?



- Yahoo! Claims 880 billion images are shared in ‘14.
- 30K images per minute in Instagram.
- 200K images per minute in Facebook.
- Sharing sensitive images is also on a rise.

http://www.sociallystacked.com/2014/01/the-growth-of-social-media-in-2014-40-surprising-stats-infographic/

Why Online Image Privacy?



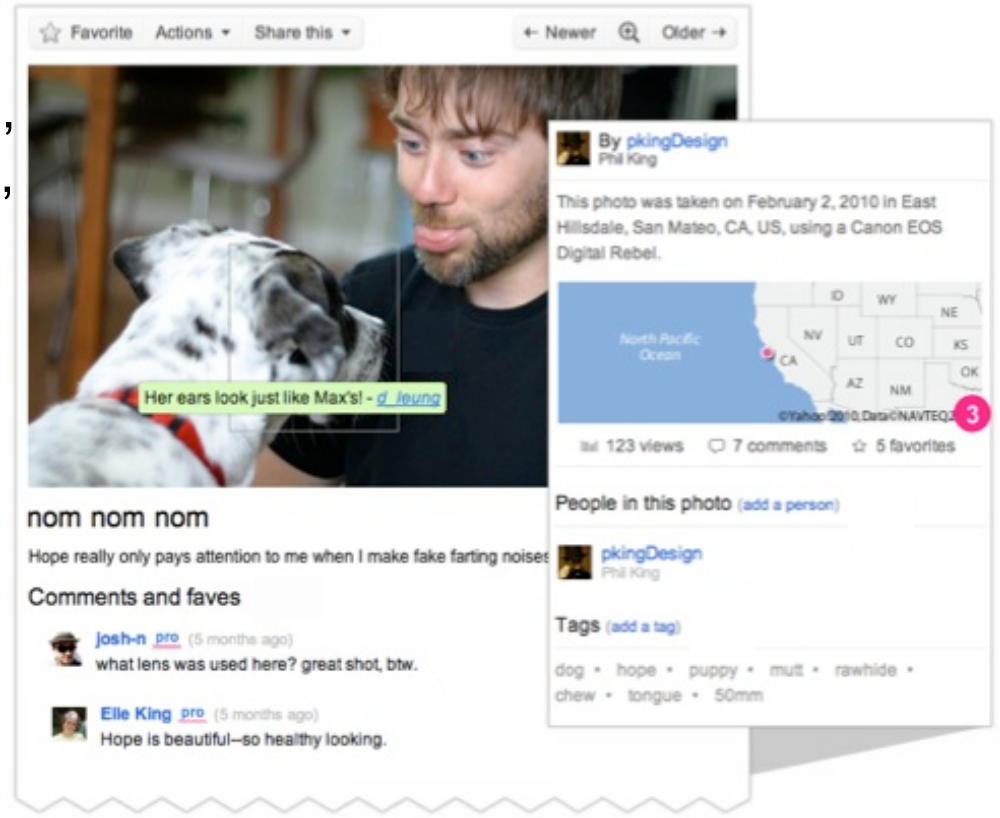
http://www.wordstream.com/articles/google-privacy-internet-privacy
SocialNetworkSecurityPrivacy. Barracudalabs.com

Why Online Image Privacy?

- With the advancements in mobile technology and Web 2.0
  - online image sharing is very easy.
- Many users are **ignorant** of privacy policies and risks of image sharing.
- Social network privacy policies are **complex**
  - Facebook explains 61 content privacy settings across 7 pages
  - Linkedin explains 52 content privacy settings across 18 pages
- Great need for methods to detect sensitivity of an image and recommend privacy policies.

Image Analysis for Privacy Setting

- Image features
  - Content and tag feature, e.g., RGB, SIFT, Edge direction, and Face detection.
    - Metadata types
      - Tags
      - Comments
      - People
      - Notes/Description
    - Contextual information
      - Type of objects
      - Names of people
      - Place of photo etc.



- Using thousands of Flickr images!!

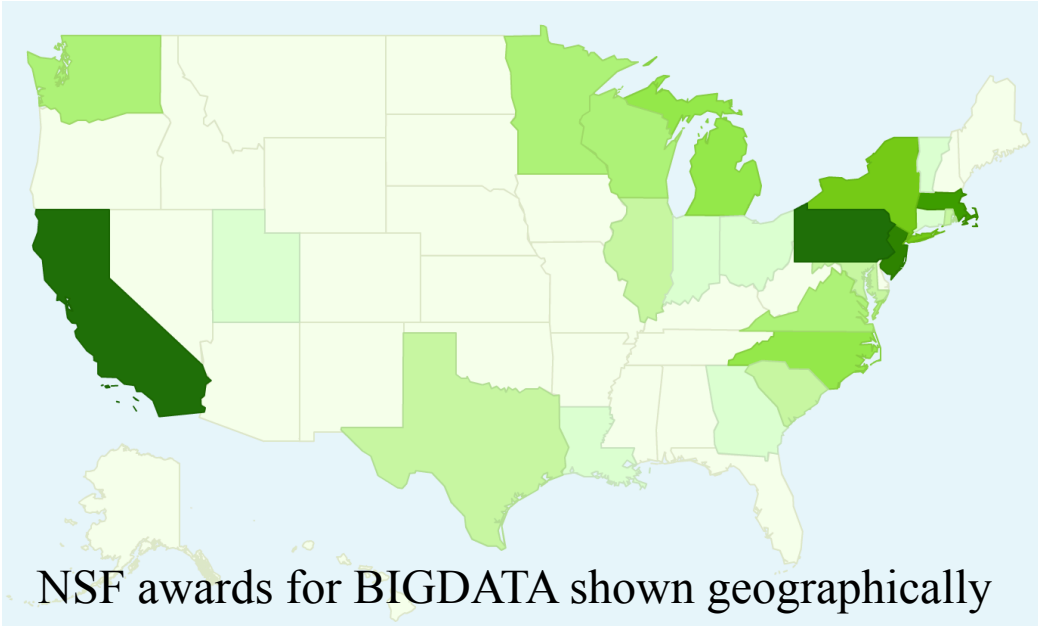
[Squiciarrini, Caragea, Balakavi; 2014]; [Godea, Caragea, Squiciarrini; 2014]

Implications of "Big Data"on Research Funding

- Great opportunities for complex data analytics
- Great funding opportunities

NIH Big Data to Knowledge (BD2K)

National Science Foundation
WHERE DISCOVERIES BEGIN



NSF awards for BIGDATA shown geographically

Conclusions

- Machine learning for “Big Data” is an exciting field of research with limitless practical application:
  - Finance, robotics, vision, machine translation, medicine, etc.
  - Open field, lots of room for new work

- 12 IT skills that employers cannot say “No” to
  - Machine Learning is #1

- “The beauty of machine learning? It never stops learning!”

References

- S. Das Gollapalli and C. Caragea (2014). Extracting Keyphrases from Research Papers using Citation Networks. In: Proceedings of the 28th American Association for Artificial Intelligence (AAAI 2014).
- C. Caragea, F. Bulgarov, A. Godea, and S. Das Gollapalli. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014).
- C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, A. H. Tapia. (2014). Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In: Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2014).
- A. Squicciarini, C. Caragea, and R. Balakavi. (2014). Analyzing Images' Privacy for the Modern Web. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT 2014).
- A. Godea, C. Caragea, A. Squicciarini. (2014). Analyzing Images to Improve Tag Recommendation. *Work in Progress.*

References

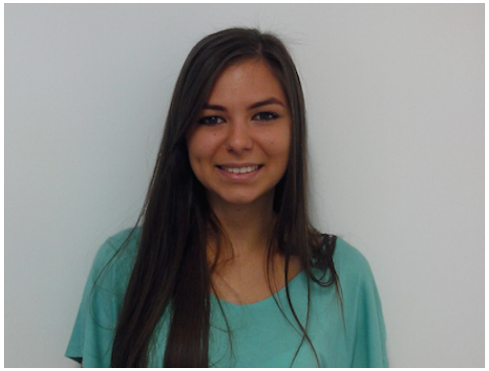
- Z. Liu, W. Huang, Y. Zheng, and M. Sun. (2010). Automatic keyphrase extraction via topic decomposition. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10).
- Mihalcea, R. & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04).
- Shi, X., Leskovec, J., & McFarland, D. A. (2010). Citing for high impact. In Proceedings of the Joint Conference on Digital Libraries (JCDL '10).
- Wan, X. & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI '08).

Thank you!

National Science Foundation
WHERE DISCOVERIES BEGIN

CiteSeerX =5M

Kishore Neppalli



Andreea Godea

Florin Bulgarov

UNT UNIVERSITY OF NORTH TEXAS

CSE Computer Science and Engineering