VISUAL ANALYTICS INTRODUCTION

LECTURE 1

Petra Isenberg

informatics mathematics
Inria

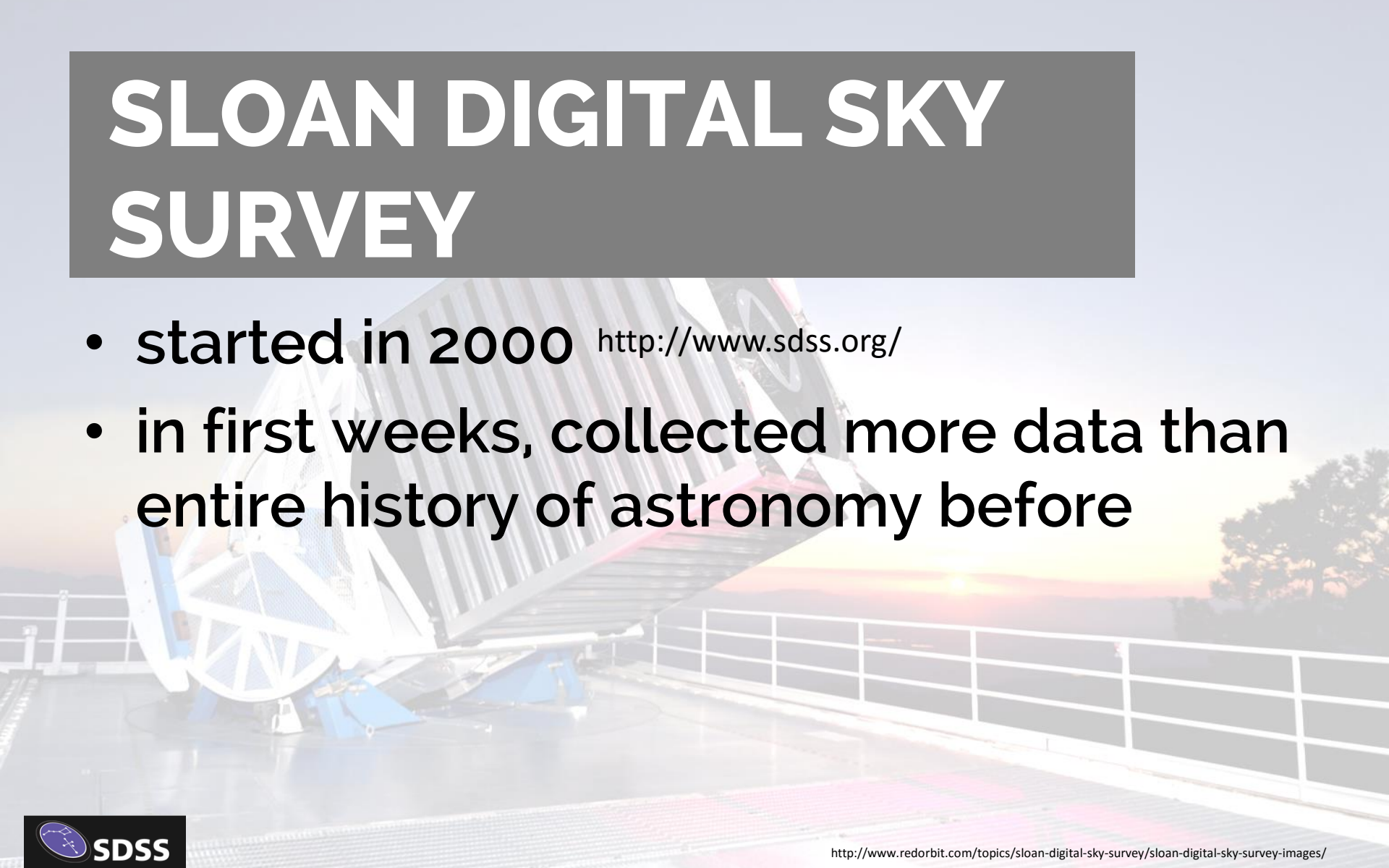
The Economist

Special Report | Data, data everywhere

Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier (interviewed here)—but also big headaches

SLOAN DIGITAL SKY SURVEY

- started in 2000 http://www.sdss.org/
- in first weeks, collected more data than entire history of astronomy before



http://www.redorbit.com/topics/sloan-digital-sky-survey/sloan-digital-sky-survey-images/

WALMART

- 1 million customer transactions per hour
- likely has information on >145 million Americans [1]

Image http://harryhammer.wordpress.com/2010/03/20/wal-mart-a-true-champion/

...AND MORE

- YouTube users upload 300 hours of new video every minute of the day
  http://expandedramblings.com/index.php/youtube-statistics/
- Facebook has currently on average 1.04 billion active users daily http://newsroom.fb.com/company-info/
- the Library of Congress adds 12,000 items to their collection every day
  http://www.loc.gov/about/fascinating-facts/

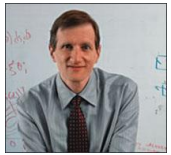
WHAT IS USEFUL?

- data != useful information
- you want insight

→ analysis is needed

ANALYSIS IS NOT SIMPLE

- research project: predict U.S. unemployment rate
- method: Twitter & social media analysis → sentiment analysis by word count


Gary King, Harvard

unemployment

jobs

classifieds

→

Look for counts of those words & correlate to monthly unemployment rate

http://www.infoworld.com/article/2611729/big-data/big-data-without-good-analytics-can-lead-to-bad-decisions.html

ANALYSIS IS NOT SIMPLE

- spike in people looking for jobs?
- lots of people going to get laid off?

HUMAN-IN-THE LOOP

- it is sometimes dangerous to rely on purely automated analyses
- human judgment and intervention often needed
  - for: background information, flexible analysis (unintended directions), creativity
  - because: data can be incomplete, inconsistent, or deceptive

COURSE OBJECTIVES

- learn about data, its properties, and its problems
- learn how to analyze (& visualize) data
  - Getting data
  - Cleaning data
  - Analyzing data
  - Visualizing data (with existing tools)
    - Normally: information visualization course directly follows this course

INSTRUCTORS

- Petra Isenberg petra.isenberg@inria.fr

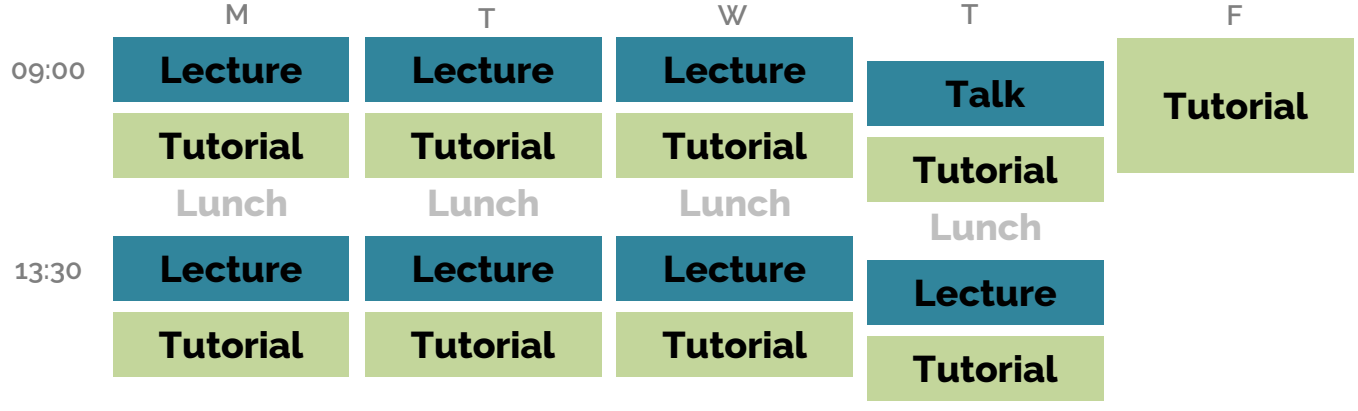
Acknowledgements

- Wesley Willett co-designed the course and made many of the slide decks
- Pierre Dragicevic is the creator of the stats lecture and assignment

OFFICE HOURS

- office: at Université Paris Sud / Bâtiment 660 (plateau de Saclay)
- Will be @ Dresden until Friday 12:00h

COURSE INFO



February 22 - 26

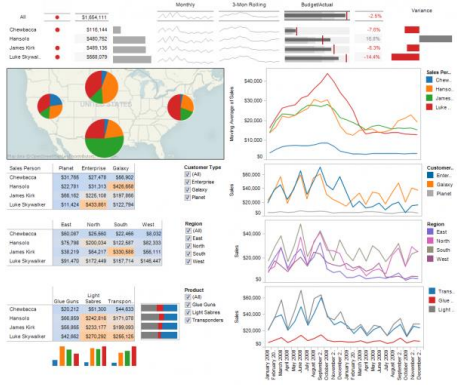
Class website:
http://tinyurl.com/VADresden

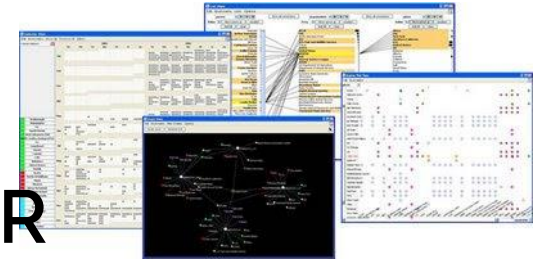
LESSON PLAN

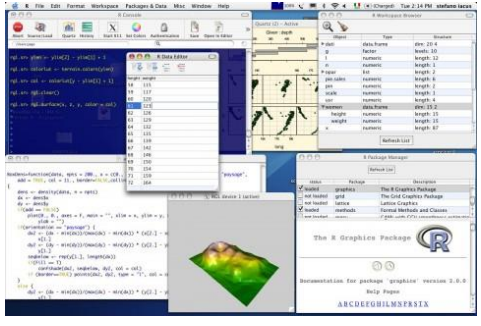
- Lecture 1: Introduction
- Lecture 2: Data Collection + Data and Ethics
- Lecture 3: Data Cleaning / Wrangling
- Lecture 4: Sensemaking
- Lecture 5: Basic Statistics
- Lecture 6: Reproducible Research
- Lecture 7: Analysis at Scale

TUTORIALS

- You will learn about:
  - Data scraping
  - Data cleaning
  - Simple statistical analysis with R
  - Analysis with Tableau
  - Making reports







GRADING SCHEME

- Assignments: 30%
  - check the website for due dates of assignments and how to submit them
- Project: 70%

READINGS

- I will announce readings on a per-lecture basis
- they will mostly be meant as additional information

QUESTIONS

WHAT IS VISUAL ANALYTICS

And where does it come from?

WHAT IS DATA ANALYSIS?

- traditionally: data analysis = statistics
- generally: data analysis = careful thinking about evidence (data)
- data analysis now covers a range of activities and skills
  - defining your problem
  - disassembling problems and data into analyzable pieces
  - evaluate the data & draw conclusions
  - make or recommend a decision

reference [3]

DATA ANALYSIS EXAMPLE

What might we be interested in analyzing?

What do you notice in the data?

	September	October	November	December	January	February
Gross sales	$5,280,000	$5,501,000	$5,469,000	$5,480,000	$5,533,000	$5,554,000
Target sales	$5,280,000	$5,500,000	$5,729,000	$5,968,000	$6,217,000	$6,476,000
Ad costs	$1,056,000	$950,400	$739,200	$528,000	$316,800	$316,800
Social network costs	$0	$105,600	$316,800	$528,000	$739,200	$739,200
Unit prices	$2.00	$2.00	$2.00	$1.90	$1.90	$1.90

reference [3]

What has been happening during the last six months with sales?

How do their gross sales figures compare to their target sales figures?

	September	October	November	December	January	February
Gross sales	$5,280,000	$5,501,000	$5,469,000	$5,480,000	$5,533,000	$5,554,000
Target sales	$5,280,000	$5,500,000	$5,729,000	$5,968,000	$6,217,000	$6,476,000
Ad costs	$1,056,000	$950,400	$739,200	$528,000	$316,800	$316,800
Social network costs	$0	$105,600	$316,800	$528,000	$739,200	$739,200
Unit prices (per oz.)	$2.00	$2.00	$2.00	$1.90	$1.90	$1.90

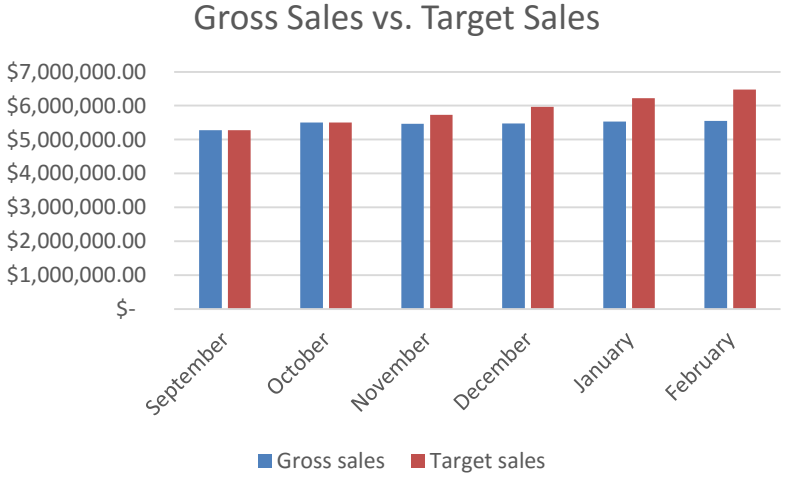
Do you see a pattern in Acme's expenses?

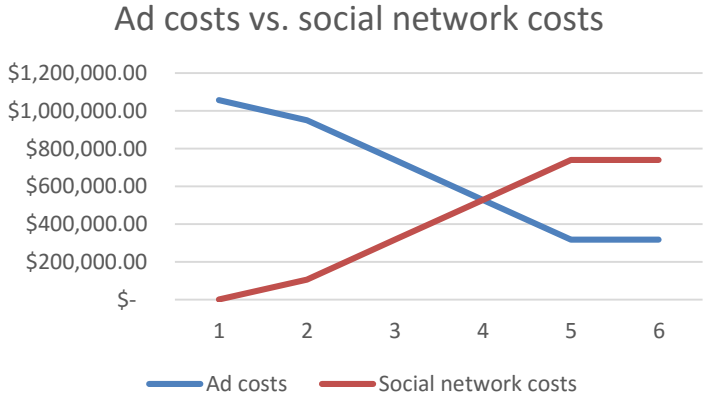
What do you think is going on with these unit prices? Why are they going down?

reference [3]

VISUAL ANALYTICS

"the science of analytical reasoning facilitated by interactive visual interfaces" [1]





VISUAL ANALYTICS

Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [5].

GRAND CHALLENGE

Enable profound insight

– allow an analyst to examine

- massive, multi-dimensional, multi-source, time-varying information
- to make the right decisions (in time-critical manner)

METHOD

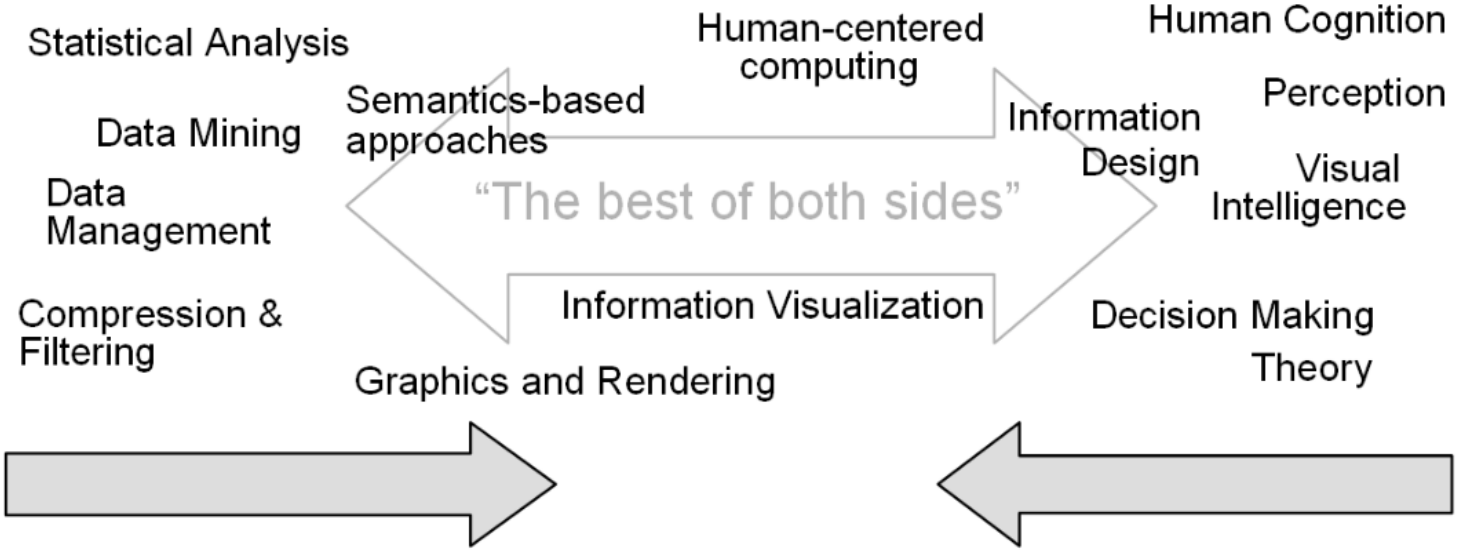
- combine automated analysis with human intervention
- represent data visually to
  - allow interaction
  - insight generation
  - drawing of conclusions
  - make better decisions

Reference [2]

SCOPE

automated analysis

human analysis



Reference [5]

CONFIRM VS. EXPLORE

confirmatory analysis

- start with a hypothesis about the data
- confirm that it is true

focus of fully automated analysis methods

exploratory analysis

- likely no a-priori information about the data
- not sure about patterns and information present
- explore to create hypotheses & confirm later

focus of visual analytics

SCOPE

visual analytics = an iterative process that involves

- information gathering
- data preprocessing
- knowledge representation
- interaction
- decision making.

Reference [2]

EXAMPLES

EXAMPLES

https://www.youtube.com/watch?v=K9PvskathGI

EXAMPLES

For this and the following videos, see:
http://ieeevis.org/year/2014/info/overview-amp-topics/paper-sessions

EXAMPLES

Integrating Predictive Analytics and Social Media

Yafeng Lu, Robert Krüger, Dennis Thom, Feng Wang,
Steffen Koch, Thomas Ertl, Ross Maciejewski

ASU VADER USTUTT VIS

online demo: https://www.youtube.com/watch?v=Zwjg8w8Xigo

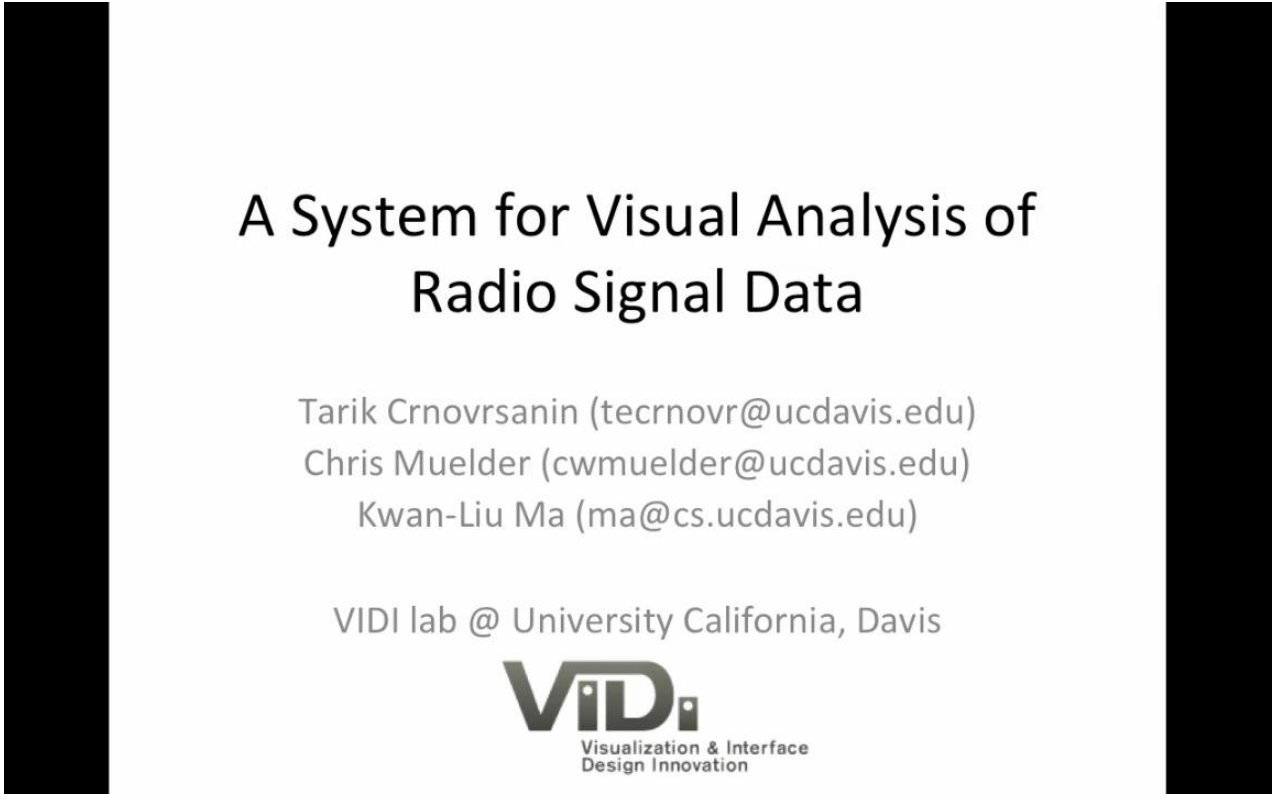
EXAMPLES


LoyalTracker: Visualizing Loyalty Dynamics in Search Engines
Conglei Shi, Yingcai Wu, Shixia Liu, Hong Zhou and Huamin Qu

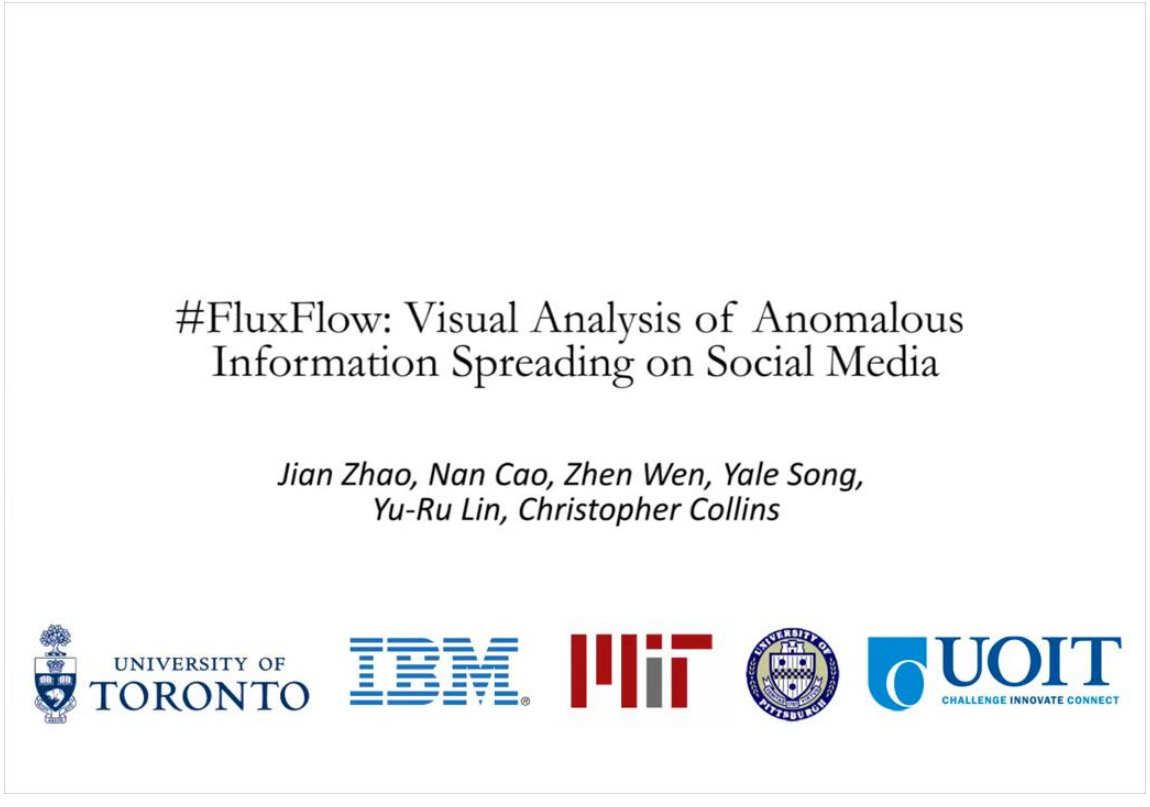
EXAMPLES


PEARL: An Interactive Visual Analytic Tool for Understanding Personal Emotional Style Derived from Social Media
Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou
University of Toronto
IBM Research

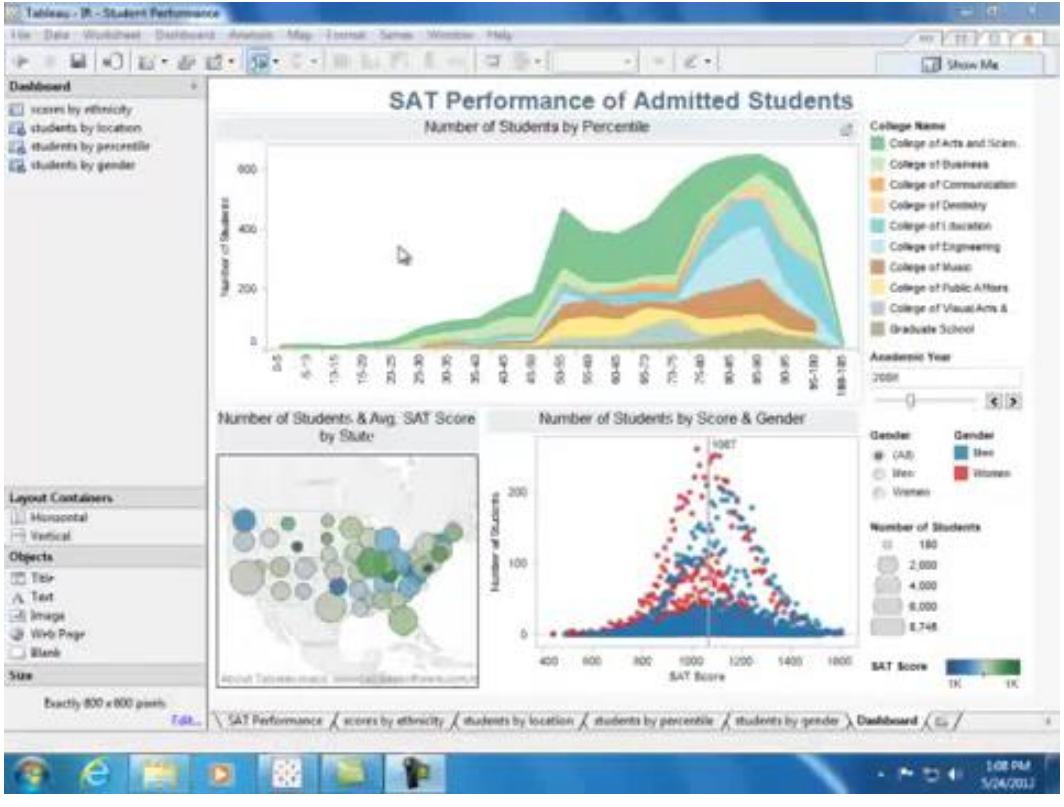
EXAMPLES


A System for Visual Analysis of Radio Signal Data
Tarik Crnovrsanin (tecrnovr@ucdavis.edu)
Chris Muelder (cwmuelder@ucdavis.edu)
Kwan-Liu Ma (ma@cs.ucdavis.edu)
VIDI lab @ University California, Davis
VIDI
Visualization & Interface Design Innovation

EXAMPLES


#FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media
Jian Zhao, Nan Cao, Zhen Wen, Yale Song,
Yu-Ru Lin, Christopher Collins
UNIVERSITY OF TORONTO
IBM
UOIT
CHALLENGE INNOVATE CONNECT

EXAMPLES



https://www.youtube.com/watch?v=_Ytz8op5Iig&list=PL722C2D5AE0BF7E99

REQUIREMENTS

development & understanding of

- data transformations & analysis algorithms
- analytical reasoning techniques
- visual representations and interactions
- techniques for production, presentation, and dissemination

CHALLENGES

human reasoning & decision making

- understanding and supporting how humans reason about data
- support convergent & divergent thinking
- create interfaces that are meaningful, clear, effective, and efficient

CHALLENGES

adoption

- communicate benefits of developed tools to drive frequent use
- make tools accepted by users

CHALLENGES

evaluation

- develop methods to compare novel tools to existing ones
- assess how good (effective, efficient, etc.) a tool is
  - very difficult for measures other than time & error, e.g. how many insights a tool generates

CHALLENGES

data

- help machines understand semantics
- quality of data is often low
- dealing with uncertainty in the data
- understanding the history or trustworthiness of data
- quantity (e.g. large and streaming data)

CHALLENGES

scalability

- data quantity (e.g. large and streaming data)
- visualization of data
- complexity and urgency of tasks
- collaboration

MOORE'S LAW...

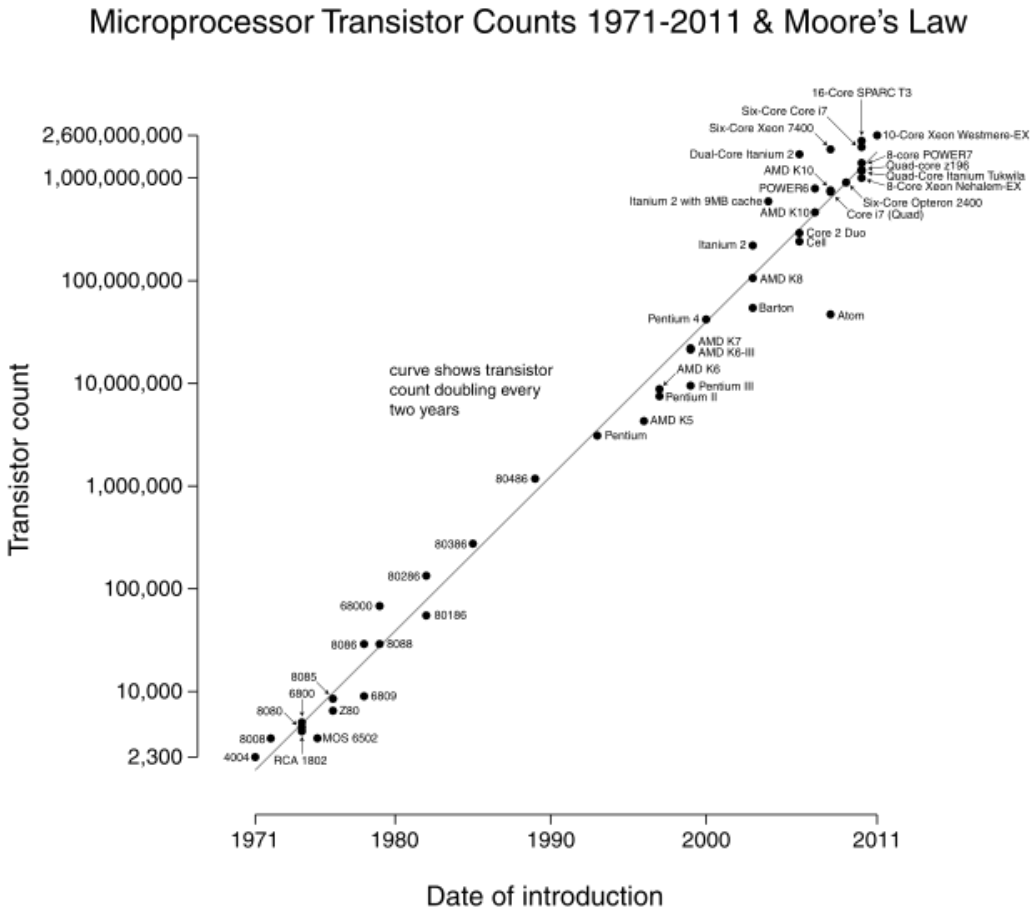
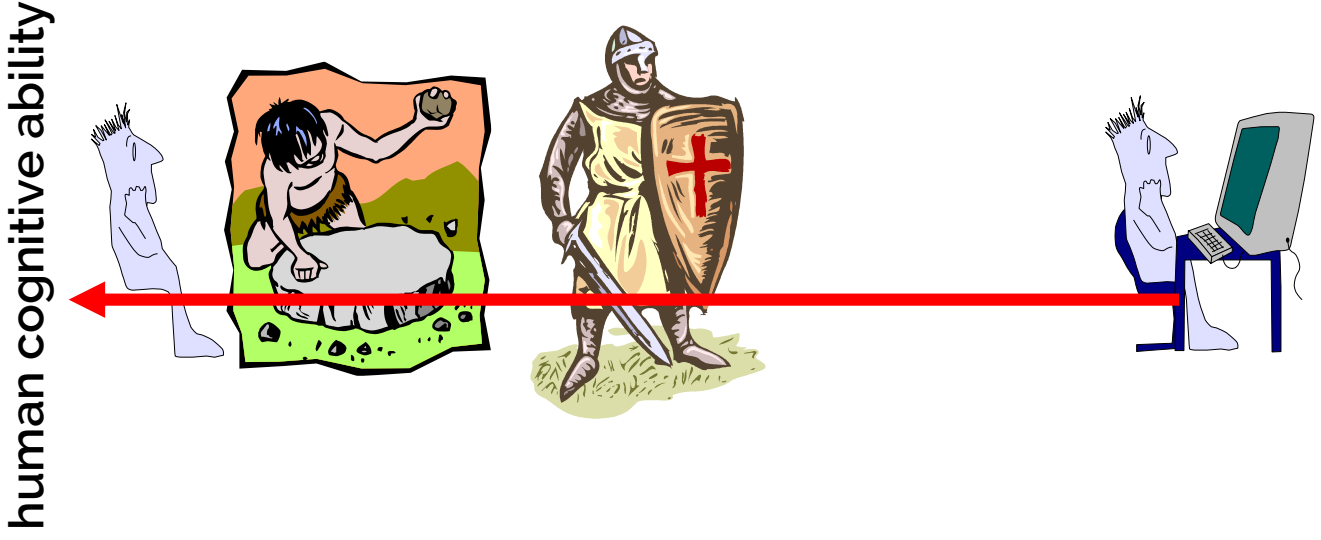


image source wikipedia

PEOPLE STAY ~THE SAME ...



information glut = we can access more information than we can process

SCALABILITY TYPES

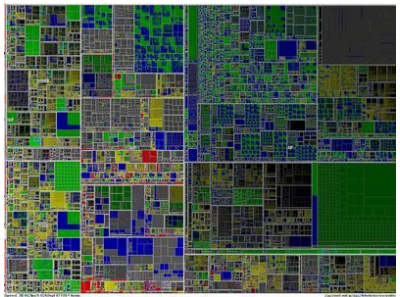
information scalability

- capability to extract relevant information from massive (possibly dynamically changing) data streams
- methods: abstract data sets, filter & reduce data, represent data in multi-resolution

SCALABILITY TYPES

visual scalability

- capability to of visualizations to effectively display massive data sets in terms of number of data items or dimensions
- depends on quality of layout, interaction techniques, perceptual capabilities

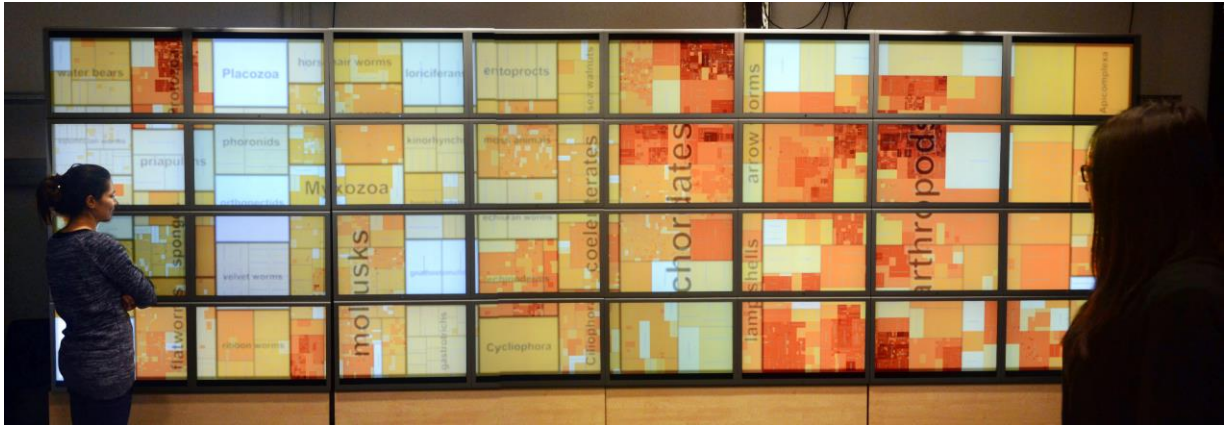


Treemap of a million items
http://www.cs.umd.edu/hcil/millionvis/

SCALABILITY TYPES

display scalability

– capability to of visualizations and tools to scale to different types of displays



Sony SmartWatch

SCALABILITY TYPES

human scalability

- human skills don't scale but numbers of humans involved in analysis can
- techniques must scale from a single to multiple users

SCALABILITY TYPES

- software scalability
  - software systems and algorithms must scale to larger data & different data
- others
  - privacy and security in multi-user settings
  - collaboration across languages and borders

CHALLENGES

problem interdependence

- analysis in the "real world" often does not consist of isolated problems or questions
- problems are often correlated and how one is solved influences how one should approach another
- synthesis of analyses is needed

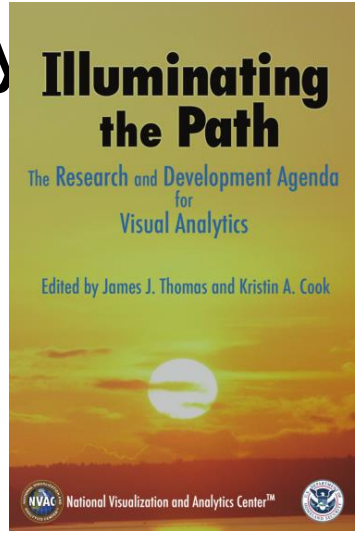
CHALLENGES

integration of analysis methods

- it is simple to do many isolated analyses
- it is hard to integrate them well into one tool, interface for human analysis

HISTORY

- outgrowth of the Scientific & Information Visualization community
- started with US National Visualization and Analytics Center (NVAC) at PNNL in 2004
- developed the first research and development agenda "Illuminating the Path"
- sponsored initially by DHS (US Department of Homeland Security)



ORIGINAL GOALS

- analyzing terrorist threats
- safeguarding boarders and ports
- preparing for and responding to emergencies

→ now only part of the larger research goals

HISTORY

- VAST symposium → conference
  - visual analytics, science, and technology
- part of the IEEE Visualization conference
- started Visual Analytics as its own research area in 2006

HISTORY

- 2008 EU funds VisMaster, a Coordination Action to join European academic and industrial R&D
- in Europe initial focus not on "homeland" security, rather broad applicability
  - physics, astronomy, climate monitoring, weather, etc.

HISTORY

- many centers in Europe
- In France mainly Inria
- In Germany mainly: Konstanz, Fraunhofer, Rostock, Stuttgart
- web: visual-analytics.eu
- book: Mastering the information age – solving problems with visual analytics
- YouTube: you saw it already

FUTURE

The Sexiest Job of the 21st Century: Data Analyst

Chris Morris, Special to CNBC.com
Wednesday, 5 Jun 2013 | 1:00 PM ET

CNBC

Photo: Biddlboo | Getty Images

In tech jobs market, data analysis is tops

Jon Swartz, USA TODAY 10:20 a.m. EDT October 5, 2012

Second of five reports this week on the job outlook in key industries.

(Photo: Elaine Thompson, AP)

256 CONNECT | 215 TWEET | 47 LINKEDIN | 3 COMMENT | EMAIL | MORE

SAN FRANCISCO -- Like a coveted free agent in sports, Kelly Halfin had a multitude of choices when she decided to take a job in tech in the U.S.

The Belgian had five American companies lined up, eager to sign her on to lead their data analysis

http://www.cnbc.com/id/100792215#.

READINGS

1. Illuminating the Path: The Research and Development Agenda for Visual Analytics Paperback – January 1, 2005 by James J. Thomas (Editor), Kristin A. Cook (Editor)
2. Daniel A. Keim and Florian Mansmann and Jörn Schneidewind and Hartmut Ziegler and Jim Thomas, *Visual Analytics: Scope and Challenges*, 2008, Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, Springer, Lecture Notes In Computer Science (lncs)
3. Michael Milton. Head First Data Analysis: A learner's guide to big numbers, statistics, and good decisions.
4. Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges (pp. 154-175). Springer Berlin Heidelberg.