# Big Data, Visualization, and Digital Humanities

**for april 24:**
**students meeting list**

**Spring 2013**
**The Graduate Center, CUNY (City University of New York)**
365 5th Avenue, New York City.
**Instructor:** **Lev Manovich**

Course numbers: IDS 81650 / MALS 78500.
Format: graduate seminar open to PhD and MA students.
Classes meet on Mondays, 2pm-4pm. Room: 3309.

First class meeting: Monday, January 28, 2013.

---

## Course description and requirements

---

# Resources:

my **list of resources and tools for visualizing cultural data**

other resources from my course Data visualization and computational art history (2012)
other resources from my course Digital Humanities++ (2011)

**Online courses in data visualization and analysis**

**List of tools for collecting, analyzing and visualizing data** from occupydatanyc.org (2012)

links to DH tools **from MALS 78100 – The Digital Humanities in Research and Teaching** (Spring 2012)

**The CUNY Digital Humanities Resource Guide**

**data journalism handbook**

For people just starting in digital humanities, I recommend that you go through

**Tooling Up for Digital Humanities**

---

# Classes schedule:

This schedule shows ***the topics we covered in class, and the plans for the next classes***. It will be updated during the semester. The topics are drawn from the larger *Relevant course topics* list below.

**class 1 / overview / january 28**
Course overview. Concepts: big data, visualization, digital humanities.
Lecture **How and why study big cultural data** (introduction to some of the course topics).

**class 2 / analog vs. born digital data; capturing physical world/ february 4**
Types of cultural data: analog - born digital. Examples of techniques for capturing physical and biological world: 19th - 21st centuries.
**lecture 2 and additional notes**

**class 3 / cultural data types; typical interactive interfaces for cultural collections / february 11**
Examples of artistic projects which use recently developed techniques for capturing the physical world. Common types of cultural data: quantitative data - text - network. Typical interactive interfaces (=visualization techniques) for cultural collections: list, gallery, frequency graph, map, network diagram.
Software: short manyeyes demo (text visualization) and introduction to Mondrian software.
**lecture 3 and additional notes**

**class 4 / data preparation and summarization / february 20** (instead of feb. 18 - holiday)
Organizing cultural data: tables vs. databases. Typical format for cultural data tables: objects (rows), and their attributes / metadata / variables / features (columns).
Practical skills: data cleanup, data summarization, sampling, basic descriptive statistics.
Software: Excel.
**lecture notes**

**class 5 / current work with big cultural data across the disciplines / february 25**
Examples of open accesss cultural data sets and social media data. Discussion of recent papers analyzing big cultural and social data (from social computing, linguistics, cultural sociology, literary studies, film studies).
**class notes**
**links to papers**

**class 6 / visualizing numerical and categorical data / march 4**
Practical skills: classical and recent visualization techniques for one dimensional, two dimensional and multidimensional data; data transformations (e.g., log scale).
Software: R, Mondrian, Excel, Google Charts, manyeyes
(If you use Windows, you can also use Tableau Public).
**class notes - basic commands and visualization in R**

**class 7 / attend NYU conference / march 11**
Surface Reading/Machine Reading: New Approaches to Texts and Text Data
20 Cooper Square, Floor 5

homework for class 8-9

**class 8 / representing space and time in art; visualizing spatial and temporal data; "science of cities" / march 18**
Representing space and time in the arts (for example: linear perspective - parallel projection, panorama - film montage, etc.) Visualization vs. mapping. Examples of innovative visualizations of spatial and temporal data. "Science of cities."
Practical skills: techniques for visualization geo-spatial and time-based data. Brief introduction to linear regression, curve fitting, time series analysis.
Software: R, www.wessa.net, Google Fusion Table
**class notes - curve fitting in R**

homework for class 8-9 (prepare to discuss projects highlighted in RED)

**class 9 / visualizing spatial and temporal data** (continued from class 8) **/ april 8**

homework for class 10

**class 10 / exploring large media collections / april 15**
Exploratory media analysis ("media visualization") using existing metadata. Media visualization techniques for exploring image and video collections. Creative sampling. Media art. analytics and visualization.
Practical skills: exploring image collections with media visualization techniques.
Software: ImageJ and software developed by softwarestudies.com.
**class notes**

**class 11 / data semiotics and modern art; features extraction / april 22**
Making art into a "language": traditions in modern art, cinema, and architecture. Counter movements (such as "blog architecture.") Concepts of features. Examples of features for text, sound, image, video. Extracting basic global features from images and video. Digital image processing and computer vision. "Media analytics" as the new stage of *media society*.
Practical skills: feature extraction from images.
Software: ImageJ and software developed by softwarestudies.com.

homework for class 12

**class 12 / analysis of multi-dimensional data: feature space, distance matrix, PCA, MDS / april 29**
Feature space. Distance in feature space. Distance (similarity) matrix. Dimension reduction and PCA. Using PCA for data exploration and visualization. Introduction to supervised machine learning and "data mining." Machine learning as the main technique for knowledge generation and decision-making in *data society*.
Practical skills: PCA, MDS.
Software: R
**class notes**

**homework for class 13**

**class 13 / cultural and social categories; Bruno Latour on visualization and sociology /** unfinished topics from lecture 3 (social networks and social media) **/ May 6**

http://en.wikipedia.org/wiki/Social_media_measurement

---

# Relevant course topics

Below is the list of *all relevant topics for this course* which I am qualified to teach (based on my research - and which I would have covered if we had enough time). The list showing which topics we actually covered is above (*classes schedule*).

**SOCIAL NETWOKS / SOCIAL MEDIA MONITORING AND ANALYTICS**
Social networks, social media. Web sources of cultural data. API. Differences between archive, database, and the web. Web analytics, social media control centers. Humanities (close-reading) vs. digital humanities (distant reading) vs. computational social science vs. social computing vs. industry analysis of web data ("Learning from Google.")
*software demo: Google Analytics.*

**TYPES OF CULTURAL DATA**
**lecture notes**
The types of "cultural data" **:** analog - born digital; quantitative data - text - network representation; scales of measurements from psychology; media types and their semiotics; static artifact - interactive process. Industry methods for the analysis of interactive experiences: game analytics; UI usability; web analytics.
*software demo: manyeyes, Mondrian (interface, basic visualization, using coordinated views).*

**DATA AND METADATA**
**lecture notes**
Data - metadata (description - meta-description). Content analysis in social sciences. Analyzing data - analyzing metadata: advantages and drawbacks. Franco Moretti: from the analysis of texts to the analysis of metadata about literature. Taxonomy - folksonomy. Data aggregation and data fusion. Generating new information and knowledge not explicitly contained in input data; recovering information from data which was not originally visible.

**DATA CLEANING AND SUMMARIZATION**
**lecture notes**
*software and techniques training: data cleanup, preparation for computer analysis and visualization, and data summarization (Excel, Google spreadsheets).*

**BASIC VISUALIZATION TECHNIQUES**
History of statistics. Descriptive statistics, calculating machines, the rise of modern mass societies. Exploratory data analysis. Classical visualization techniques for one dimensional, two dimensional and multi-dimensional data.
*software demo: R (interface, reading data, data summaries, plots), Mondrian (visualization of multi-dimensional data)*

**VISUALIZATION CULTURES**
information design, scientific visualization, data visualization, artistic visualization. Visualization and the history of visual media. Visualization vs. mapping. 3D visualization. Media art as a form data analysis.
*software demo: geo-spatial data and mapping (various mapping tools).*

**MEDIA VISUALIZATION**
Exploratory media analysis ("media visualization") using existing metadata. Media visualization techniques for exploring image and video collections. Creative sampling. Artistic projects in

"media visualization." Work on visualization of image collections in CS.
*software demo: ImageJ (interface, using commands, media visualization)*

**VISUALIZING TEMPORAL DATA**
Visualizing temporal processes. Linear regression, curve fitting. Cinemetrics studies. Timeline vs. style space. Database vs data stream interface in social media.
*software demo: R (linear regression, traditional and radial timelines).*

**EXTRACTING FEATURES FROM DATA**
Concepts of features and feature space. Examples of features for text, sound, image, video. Extracting basic global features from images and video, and (optional) text. Digital image processing and computer vision. "Media analytics" as the new stage of media society.
*software demo: ImageJ (feature extraction); After Effects (motion tracking).*

**ANALYSIS of MULTI-DIMENSIONAL DATA/ INTRODUCTION TO MACHINE LEARNING**
Feature space concept. Distance in feature space. Dimension reduction and PCA. Using PCA for visualization. Introduction to supervised machine learning and "data mining." Search engines, recommendation systems. Machine learning as the main echnique for knowledge generation and decision making in data society.
*software demo: R (PCA, MDS, optionally: categorical data analysis and cluster analysis), ImagePlot.*
*Google Predict service.*

**NETWORK ANALYSIS and VISUALIZATION**
*software demo: R (web analytics - downloading and analyzing Twitter data; network analysis and visualization)*
MOMA Inventing Abstraction 2013 exhibition: network visualization

**TEXT ANALYSIS AND VISUALIZATION**
Computer analysis of texts: concepts and possibilities.
*software demo: text analytics with R and Mallet*
http://www.fredgibbs.net/clio3workspace/blog/reading-with-r/

**CULTURAL CATEGORIES and DATA ANALYSIS**
Key philosophical theories of concepts. Recent theories from cognitive psychology. Classification, data exploration, and cultural categories. How to use data mining and visualization to "unlearn" existing cultural categories. "Style space."

**INTERFACES TO CULTURAL COLLECTIONS**
Artistic techniques as interfaces. History of computer interfaces. Interactive media installations. Examples of innovative interfaces to cultural collections.

**MEDIA AFTER SOFTWARE**
Software studies: How does the use of software changes "media"? Do separate mediums still

exist in software society? How to conceptualize interactive cultural software "artifacts" and "processes"? Algorithms and "software epistemology" (examples from digital photography and motion tracking).
*software demo: analysis of Photoshop and After Effects interfaces*

**HOW MUCH DATA?**
Information society, network society, social media, internet of things, data society. Computers and the growth of information. "How much information?" and the politics of such measurements. The dream of semantic web and perfect metadata.

**REPRESENTING DATA SOCIETY**
How do artists represent society through data? Data, representation, symbol. Modernity - modernism, information society - info-aesthetics.

---

# Homework:

The **homework** to be done before each class is **listed below** (after lecture notes).

Because our class meets only once a week, I will only be able to go over some of the topics each week. Other topics are described in lecture notes with the links to additional materials.

Therefore, In addition to the assigned readings and sites to view in homeworks, you should also go through **lecture notes and the links for each class**, researching any subjects which interest in more detail. You can do that before or after the class.

If you are already familiar with any of the readings, projects, concepts, or data analysis/visualization techniques covered in any of the homework, skip them.

If you don't have computer science background to understand details some of the readings, try at least to get the key ideas - read an introduction and a conclusion.

---

# LECTURE NOTES AND LINKS:

---

## class 1 / overview / january 28

Course overview.

Lecture **How and why study big cultural data** - introduction to some of the course topics.

---

## homework for class 2:

1) **Install on your laptop the following software - make sure to bring your laptop to every class:**

- Mondrian software

- Excel (if you don't have it already on your laptop, download 60 day trial version).

2) Start following the following online sources - **check them every week**:

infosthetics.com

digitalhumanitiesnow.org

**Read / view:**

**rise of visualization:**

1) Manuel Lima. Outburst of Visualization (2010). In Lev Manovich, Jeremy Douglass, William Huber, *Mapping Time*. gallery@calit2, forthcoming 2013.
**Visit the web sites and projects referred in the article.**

**(**Manuel Lima is the edtor of the influential visualcomplexity gallery of network visualizations, and the author of Visual Complexity: Mapping Patterns of Information,

2011.)

**big cultural data + visualization: example:**

2) In 500 Billion Words, New Window on Culture. NYT, 12/16/2010.
Play with http://ngrams.googlelabs.com.
More information: http://www.culturomics.org/

> For more details and many usage exampls, see the original article in *Science*: Quantitative Analysis of Culture Using Millions of Digitized Books, December 16 2010, Science.

**the work of my lab - analysis and visualization of large cultural data:**

3) http://lab.softwarestudies.com/2008/09/cultural-analytics.html

4) Go through the slides of class 1 lecture: How and why study big cultural data
(Note: the concepts and examples in the lectures will be discussed in more details in later classes - this is just a preview to give you an idea about what will be coming.)

**sources of cultural data:**

5) Lev Manovich. "Cultural Analytics: Visualizing Cultural Patterns in the Era of 'More Media.' *Domus*, Spring 2009.

**Some of the methods for collecting data about cultural experiences - eye tracking, brain recordings:**

6) Tobii, Advertising research and eye tracking.

> if you are interested to play with already recorded eye movement data, see this project:
>
> The DIEM project is an investigation of how people look and see. DIEM has so far collected data from over 250 participants watching 85 different videos. All of our data is freely available for research and non-commercial use as restricted by a CC-NC-SA 3.0 Creative Commons license.

7) rise-neurocinema-how-hollywood-studios-harness-your-brainwaves-win-oscars, Fast

Company, Feb, 25, 2011. If you want more details, see the original article:
Uri Hasson, Neurocinematics: The Neuroscience of Film, 2008.

**Example of using digital footprints -**
**web analytics software, twitter analytics, social media analytics:**

8) If you don't use Google Analytics on your site/blog, or have not heard of it, see this overview:

http://www.google.com/analytics/features/index.html

9) Try these Twitter analytics tools with your own account and others:

http://twtrland.com/
http://www.twitonomy.com/

**Concepts:**

if you are not familiar with any of the concepts below, read these linked Wikpedia page. (And if you are familiar with them, I suggest you take a look at these Wikipedia pages anyway since you will most likely find new things you did not know.)

read - you can skip the technical parts:

API
scale types
digital repository
digital curation
born digital
digital footprint

you all know this, but just in case (and to see latest trends in these areas):

web 2.0
social media
social network service
the long tail
user-generated content
mass_collaboration

# class 2 / DATA / february 4 - lecture notes and links:

## The types of "cultural data"

Conceptual distinctions crucial for the collection, analysis and visualization of cultural data:

**1) analog - born digital**

digitization (From Stanford Digital Humanities guide)

book scanner example (video)

advantages of working with born digital cultural data:

often has detailed metadata recorded by capture devices or computers (and therefore accurate) - human-entered metadata often has mistakes, incomplete, and subjective. Example of metadata captured by devices from iPhoto.

with originally-analog images and video, the digital files available online (from users and institutions like arstor.org) are often not accurate. Example: Comparing set of image of Gauguin paintings we downloaded from two different sources.

**examples of techniques for capturing physical and biological world and processes:**

examples of movement recording techniques from 19th and first part of 20th century (see my 2013 article Visualizing Vertov for discussion):

Étienne-Jules Marey

online exhibition of his work

visualizations of recordings of bird's wing movement

Eadweard Muybridge

Frank and Lillian Gilbreth: time and motion study

Taylor: "time studies involved breaking down each job into component parts, timing each part and rearranging the parts into the most efficient method of working."

"The Gilbreths made use of scientific insights to develop a study method based upon the analysis of **work motions**, consisting in part of filming the details of a worker's activities while recording the time.The films

served two main purposes. One was the visual record of how work had been done, emphasising areas for improvement. Secondly, the films also served the purpose of training workers about the best way to perform their work."

contemporary method: motion capture (Wikipedia article)

example of using motion and face capture in film production (*Avatar*, 2009)
visual effects in *Avatar* (Wikipedia)

virtual cinematography and universal capture methods (The *Matrix Reloaded*, 2003)
articles about "virtual cinematography" and universal capture:
articles by George Borshukov
Debevec, Paul (2006). "Virtual Cinematography: Relighting through Computation" *Computer*(IEEE) **39** (8): 57–65.
Manovich, Lev. (2006). Image Future.

Capturing archeological data:
work at Calit2 (CISA3 group) - exploring 3D model in CAVE (video)

Eye tracking:
Tobii, Advertising research and eye tracking
Use of eye tracking in usability research (usability article on Wikipedia)

Analyzing face to infer emotions:
NYT article about the research of Rosalind Picard, director of the affective computing research group at the MIT Media Lab.
affective computing research lab - projects
their sample data sets available for download
company span by Picard - web demo

use of fMRI, EEG (brain activity recording) for the stuty of cultural experiences
fMRI (Wikipedia article)
use of fMRI to study film cognition:
neurocinema video - using fMRI to study emotional reactions to commercials (YouTube)
rise-neurocinema-how-hollywood-studios-harness-your-brainwaves-win-oscars, Fast Company, Feb, 25, 2011.
If you want more details on neurocinema, see the original article: Uri Hasson, Neurocinematics: The Neuroscience of Film, 2008.

use of fMRI and EEG to study music cognition:
research articles (Google Scholar)

(The other topics which originally were listed here have been moved to *lecture 3* notes).

---

# homework for class 3:

In class 3 we will finish the topics from lecture notes for class 2.

I will show how to use Mondrian software to explore data sets.

We will also start topics listed in the class schedule (above) for class 3.

Here is your homework:

1) Go through the links in my lecture notes for class 2. If any topics interest you in particular, research them further on your own using web resources (usually Wikipedia pages references is a good start).

2) if you are interested to do practical work with data in this course and don't know Excel, please **familiarize yourself with Excel** interface, using Fill command, and using Formulas.

new readings:

3) Life in the network: the coming age of computational social science. Science. 2009 February 6; 323(5915): 721–723.

4) Computational social science: Making the links. Nature, 22 August 2012.
Learn more about the research of people mentioned in the article.

5) Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2009 February 6; 323(5915): 721–723.

6) Manovich, Lev. Against Search (a part of a longer article). 2012.

# class 3 / DATA / february 11 - lecture notes and links:

## Software demos:

ManyEyes (on the Mac, use Firefox browser)

Mondrian software (interface, basic visualization, using coordinated views)

Google Analytics

Digitization of cultural "artifacts" and recording of "digital traces" are just latest episodes in the longer history of recording the world and the humans as "data," and later as "media" (19th - 20th centuries).

Today we have two parallel processes: 1) recording of the physical world and humans as media and also directly as digital data; 2) translation of older analog recordings into digital data.

## New *data capture technologies* have major effects on society, economy, media, culture, and lead to new forms of art (when they become widely available):

19th-20th century:
photography, film, audio and video recording, radar, remote sensing technologies

Examples of art projects done with the recently developed new data captures technologies in the 21st century:

**Lidar scanners** (Wikipedia article)
Aaron Koblin, House of Cards video for Radiohead (interactive version). Non-interactive s video version (YouTube). Making Of (YouTube). For other Aaron's

projects, see http://www.aaronkoblin.com/work.html. Newest: Light Echoes.

**GPS receivers**

Masaki Fujihata, Field-Works@Alsace (2002)
milkproject.net (2004) (one of the earlist art projects which uses GPS)
Since location capture is now built into mobile phones and many popular social media apps, lots of projects explore this - for example: Eric Fisher, locals and tourists (set on Flickr) (2010)

**social media data**

listening post (2002)
wefeelfine.org (2006)
Pulse of the Nation (2010)

**motion capture**

Ghostcatching, a digital art installation by Paul Kaiser and Shelley Eshkar uses motion capture of the dance performances by Bill T. Jones (1999).
Hacking and creative coding around Kinect platform.

**High resolution digital photography**

http://9-eyes.com/ (Jon Rafman)
http://www.wired.com/design/2013/08/22-beautiful-photographs-hidden-in-this-insane-150-gigapixel-image-of-tokyo/

## The types of "cultural data" (continued from class 2):

Conceptual distinctions crucial for the collection, analysis and visualization of cultural data:

**2) media types and their semiotics** (for computational analysis and visualization).

Human language have discrete units (morpheme, word, clause, etc. ) and also is characterized by double articulation.
This makes it easy to extract these unites and analyze their patterns quantitatively. Because the unites such as morphemes and words have stable meaning in a particular language, this also allows for the automatic computational content analysis (at least in some cases), machine translation, and other projects.
Semiotics project of the 1960s wanted to inderstand all forms of cultutal communication using the model of natural languages. The classical example is Roland

Barthes, Elements of Semiology, 1964.

However, sound, images, spaces, shapes, and smells in the natural world usually not have such discrete meaningful unites. This was also the problem 1960s semiotics could not solve. Over the decades, computer scientists made progress in automatic analysis of the content of this "analog" data (i.e., data types which do not have a priori discrete units) - however this problem is not yet solved (and may never being solved completely.)

Many non-linguistic cultural artifacts and their systems created by humans do have discrete unites with clearly defined (at least in theory ) meanings and/or emotional effects. Examples: traffic signs, 3D shapes modelled with computer graphics. Modern artists and art theorists proposed various artificial "languages" consisting of units - for example:

modernist geometric abstraction

see current MOMA exhibition Inventing Abstraction

Kandinsky, Point and Line to Plane

Russian montage school in cinema (Kuleshov, Eisenstein, Vertov, etc).

Eisenstein, Methods of Montage.

See our visualizations of Vertov films on Flickr (for my article Visualizing Vertov).

In later classes we will learn about digital image processing and computer vision - the fields of computer science which develops concepts about what are the useful unites for analysis and algorithms for extracting these units at different scales from any images and video. (This process is called "feature extraction.")

**3) static artifact - interactive process**

Almost all digital humanities research so was only dealt with digitized artifacts from the past (static files) - but we leave in the era interactive digital media. How do we represent interactive experiences as "data"?

(My article about some ways to repesent cultural interactive experiences (in the case of a virtual world) - we will come back to this in more detail in a later class:

Manovich, Lev (2012). How to Follow Software Users? (Digital Humanities, Software Studies, Big Data), forthcoming in Digital Humanities Quarterly.

## Some of the standard ways to categorize data in statistics, data visualization, and the sciences (study on your own):

**4) numbers - text - network representation - geo-spatial data**

While there are many other data types, from the point of visualizing data, these four are the most common.

Demo: examples of visualizing text and network data using ManyEyes. (On the Mac, use Firefox browser.)

Demo: Examples of visualizing numerical data using Mondrian software.

> data set for visualization: images and metadata for 778 Vincent van Gogh (distributed with our ImagePlot software) - the data file `van_gogh_data.txt` available in Dropbox folder.

ManyEyes descriptions of data formats for common visualization techniques (note that the techniques are classified by the same types: numbers, text, network, map).

(We will discuss this point in a latter class:)

Data visualizations vs. maps - are these fundamentally different representational methods, or can we think of maps as a subset of visualizations? (Examples: most software listed on Datavisualizatio.ch list are mapping tools; a big proportion of projects on infosthetics.com are also mappng projects).

An example of mapping + visualization of social media activity: http://tweetping.net/

**5) ordinal - interval - ratio** scales (from psychology)

scale types theory (Wikipedia article - explanations below are from this source)

**ordinal scale**:
*Ordinal measurements describe order, but not relative size or degree of difference between the items measured.*
Example: 'completely agree', 'mostly agree', 'mostly disagree', 'completely disagree' in measuring opinion.

**interval scale:**
*Interval scales tell us about the order of data points, and the size of the intervals in between data points.*
Quantitative attributes are all measurable on interval scales. A highly familiar example of interval scale measurement is temperature with the Celsius scale.
The "zero point" on an interval scale is arbitrary; and negative values can be used.

**ratio scale:**
*A ratio scale is an interval scale with a true zero point.* Most measurement in the physical sciences and engineering is done on ratio scales. Mass, length, duration,plane angle, energy and electric charge are examples of physical measures that are ratio scales. Examples of ratio scale measurement in the behavioral sciences are all but non-existent.

### 6) statistical data types

depending on the data types, some statistical measures are meaningful and some are not.

**categorical data** (explanation from categorical variable Wikipedia article):
In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values. Examples of values that might be represented in a categorical variable:

The blood type of a person: A, B, AB or O.
The state that a resident of the United States lives in.
The political party that a voter in a European country might vote for: Christian Democrat, Social Democrat, Green Party, etc.

### Sources of cultural data, APIs

Big cultural data sources from Digging Into Data NEH/NSF competition repository.

List of some of the museums that already put images and data of their collections online and/or offer API to their collections.

Example of using API to download social media data: our code to harvest photos and metadata from Flickr.

### Web and social media analytics

Social media control centers - for example, hootsuite command center

Google Analytics is one of numerous applications and services designed for the analysis of users' interactions with the web sites, services, and apps. (See Wikipedia article web analytics). While it only analyzes the data from a single web site/blog, other products analyze all social media activity around a single brand, or perform other analysis which uses large social data.

Demo: Google Analytics software.

For an overview of the social media monitoring concepts, see http://www.slideshare.net/StefanBetzold/social-media-monitoring-tools-an-overview

For an overview of the social media "command centers" being now created by many companies, see NASA-style mission control centers for social media are taking off, CNN Money, October 25, 2012.

Example of a simple social media "command center" web interface (HootSuite): http://hootsuite.com/super-bowl-XLVII

Example of a company building software for the analysis of large scale consumer social media data:

Blue Fin Labs - see product overviews.
Article about the company from MIT Technology Review.

Different industry approach to analyze social media: web site centric, user centric, brand centric. How these different approaches can be applied for the cultural and social analysi outside of media industry (turning them into methodologies to research both past and present)?
For example:

web site centric - collecting and analyzing all available data about a particular cultural artifact;

user centric - following cultural lives of people; expanding usual biographical analysis of important cultural figures to include and visualize more data;

brand centric - collecting and analyzing all data about a cultural movement, school, etc.

**Differences between archive, database, and the web**:

If cultural objects and data are usually stored in archives, modern society since the 1970s uses databases to store and access most of its data.

Theories in modern cultural theory:
Manoff, Marlene. Theories of the Archive from Across the Disciplines:

> "Many scholars (whether or not they describe themselves as postmodernists) have come to understand the historical record, whether it consists of books in libraries or records in archives, not as an objective representation of the past, but rather as a selection of objects that have been preserved for a variety of reasons (which may include sheer luck). These objects cannot provide direct and unmediated access to the past. Historian Dominick LaCapra has
> described the dangers of the "archive as fetish," of believing that the

> archive "is a literal substitute for the 'reality' of the past which is 'always already' lost for the historian." Whatever the archive contains is already a reconstruction—a recording of history from a particular perspective; it thus cannot provide transparent access to the events themselves. But regardless of what historians may have once believed, there is currently a widespread sense that even government records that appear to be mere collections of numbers are, in fact, already reconstructions and interpretations. Someone decided what was worth counting and how to count it."

With databases, you can use software to perform various operations on the data (search, filter, combine data sources, summarize, etc.) software can also access databases on the remote servers, update data, and also fetch data and display in the web app (all social networks, social media services and web stores operate in this way.

There are many types of database architecture. Most commonly used type today is relational model (Wikipedia article - see illustration showing the relational database consisting from a number of connected tables). Its most widely used implementation today is mySQL. Examples of using mySQL language.

After mySQL: recently developed data storage systems to handle big data (used by Google, Facebook, etc.)

Big data definition from IT industry.

How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day (techcrunch.com, 2012/08/22).

Can the World Wide Web be considered a database?

(Social network services - the full "database" only accessible to the company vs. data provided via API vs. user view, i.e. a datastreams - we will discuss this in detail in a later class.)

Google search engine (Wikipedia article).

How big is the web (one of the projects to measure the size of the web).

If you want to work with the massive web data (alternative to Google web data which is not publically accessible), you can use Common Crawl collection - currently 5 billion captured web pages (81 TB).

**What is the "data"?**

one answer: *data* a collection of information in the format which allows it to be analyzed with computers.

---

# homework for classes 4-5:

1) Start going through these papers to prepare for their class discussion on **Feb. 25:**

(You don't need to read every paper in detail. Just read enough to understand the key ideas, methods, and findings.)

Examples of the analysis of big cultural data from different disciplines:

literary studies:

Matthew Jockers. Computing and Visualizing the 19th-Century Literary Genome. DH 2012 conference.

Ted Underwood and Jordan Sellers, The Emergence of Literary Diction. Journal of Digital Humanities, Vol. 1, No. 2 Spring 2012.

(Highly recommended: Ted Underwood's blog.)

linguistics and "cultunomics":

Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2009 February 6; 323(5915): 721–723.

quantitative film studies:

Cutting, J. E., Brunick, K. L, DeLong, J. E., Iricinschi, C., & Candan, A. (2011). Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception, 2*, 569-

576.

(Other articles by Cutting on quantitative film analysis.)

history of technology:

Andrew Buchanan, Norman Packard, and Mark Bedau. Darwinian evolution of culture as reflected in patent records. Artificial Life, 2012.

social networks and social media analysis (computer science / social computing):

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moo.
What is Twitter, a Social Network or a News Media? WWW 2010 conference.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. Proc. of International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010.

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. Proc. of Usenix/ACM SIGCOMM Internet Measurement Conference (IMC), October 2007.

Justin Cranshaw, Raz Schwartz, Jason I. Hong, Norman Sadeh.
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. The 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland, June 2012. See the project web site: http://livehoods.org/.

---

# class 4 / DATA / february 20 - lecture notes and links:

If you have solid experience with the topics below, you don't have to come to this class meeting.

If you want me to cover any additional topics, email me before the class.

# Organizing data using tables, cleaning data, preparing data for analysis and visualization, creating data summaries:

**Sample data sets for this lecture** (archive file 3.8 MB.)

**Organizing data using a table:**

The most common structure for organizing data: a table.

Typical **table organization**: each row contains data for one object/actor/event; each column contains a separate attribute for this object.

(Alternative "square table structure" - representing relations between actors/objects).

Depending on the disciplines, attribute columns have different names:
"variables" (statistics)
"features" (computer science)
"metadata" (digital humanities, libraries, museums)

Examples of data tables:

NYC colleages and universities (select "view as table" in the toolbar in the upper right corner).

Google Scholar list of publications

sample data for 1000 images from deviantArt (category: Traditional Art).

user sessions data from Scalable City installation

Examples of historical data tables - US Census sample pages 1790-1930

Databases such as mySQL combine multiple tables together, allowing for many more powerful ways of working with big data - we will discuss this in a later class.

Example of a database structure from live sciences

Tables can be stored in files or spreadsheets. Most common file formats for data file are comma delimited (.csv) and tab delimited (.txt). These types of files are often called "text files," even though they usually contain numbers, or numbers and text strings.

Common spreadsheets software: Excel, Numbers (Mac), Google Spreadheets, Open Office.

Spreadsheet software typically includes graphing functions, so you can use a spreadsheet both to organize and explore data using visualization techniques.

Spreadsheets can't deal with really large data (typically over a million rows). Databases don't have this limit. However, traditional databases (such as mySQL) are too slow for real-time processing of massive social network data; companies (Facebook, Google, Yahoo) created new technologies to handle such data. Parts of these technologies are available as open source.

Further info (if interested, explore on your own):

How Big Is Facebook's Data? 2.5 Billion Pieces Of Content And 500+ Terabytes Ingested Every Day (August 2012 data)

Apache Cassandra

Apache Hadoop

Google BigTable

Popular text editors for **opening very large text files**: TextWrangler, BBEdit

Most common **data types** used in tables: integers, floating point numbers, text strings. How Excel displays these data types.

**Practical principles for organizing data in tables**:

1) each column should contain a single attribute. If you column contains multiple attributes (e.g., last name and first name), you will not be able to explore each of the attributes separately using visualization software.

2) If any of the cells starts with space, delete these spaces.

3) don't mix data types in the same column. I.e., each column should contain integers, or floating point numbers, or text strings. If your column contains even one cell with text string, Excel will interpet the whole column as text.

4) Unless the column contains text string and you want space(s) as part of this string (for instance, the name of a movie), don't leave any empty spaces in any cells.

5) Be careful with using special characters in text strings (i.e., commas, quotes, etc) - these may confuse many visualization programs.

**Web tools for preparing and working with data:**

Data Wrangler

Google Refine

Google Fusion Table

Most common operations for **data cleaning, preparing data for computer analysis and visualization, and data summarisation (using Excel)**:

**- cleaning and organizing data - common operations:**

- using functions for working with data:
  Excel functions list
  Google spreadsheet functions list

-copying the column containing formula calculation into a new column: Edit > Paste Special > Values

- converting between data types (int, float, char, etc.): Format > Cells
  convert between text and numbers

- concatenating data from multiple columns (CONCATENATE)

- splitting data into multiple columns: Data > Text to Columns

- operations on text strings (LEFT, RIGHT, LEN, MATCH)

- adding sequential numbers to rows: Edit > Fill > Series

- how Excel treats dates and times, converting between dates / times and numbers

How to use dates and times in Excel

How to Convert Text to Dates

- creating **random samples**

Creating a random data sample in Excel:
add a new column to your spreadsheet;
fill it with rand formula:
=rand()
copy this column into a new column using “paste special” (select “copy values” option)
sort spreadsheet by the new column
select first N rows

- **data summaries** in Excel: using Filter, formulas (countif, sumif) and Subtotals command.

Using Subtotals command in Excel to summarize the data:

1) sort the column you want to summarize
2) use Subtotal command on this column
3) copy the results (how to copy subtotals results)

Additional data cleanup and summarization resources:

Guide to data cleaning in Excel

Guide to data cleaning in Excel

5 methods to summarize data in Excel

How to consolidate data in Excel and graph it

techniques to add and count Excel data

Add numbers based on multiple conditions in Excel

How to consolidate data in Excel and graph it

- calculating **basic descriptive statistics** in Excel: mean, median, variance, standard deviation.

with formula:
mean: AVERAGE
median: MEDIAN
variance: VAR
difference variance functions
standard deviation: STDEV
different standard deviation functions

display averages using Filter and Excel interface

- **using Google spreadsheets to read external data:**

Google specific spreadsheet functions

using Google spreadsheets to scrape tables from the web:

converting web page tables into Google spreadsheets

example:
a list of van Gogh paintings from Wikipedia page
cleaned data using Google spreadsheets functions
the titles (with periods added) data uploaded to manyeyes

- **converting between file formats** which all can be used to store tables:

- .xls, .txt, and .csv formats;

- converting between Google Spreadsheets, Excel, and text files;

- converting from XML, and from JSON:

Reading XML files into Excel

Using JSON data in Excel

- “data liberation” is the term used to describe getting data from proprietary sources (twitter, facebook, etc). or complex formats.

---

# class 5 / BIG CULTURAL DATA ANALYSIS ACROSS THE DISCIPLINES / february 25 - lecture notes and links:

**Examples / sources for cultural and social data sets** (research on your own):

NYC data (with build-in visualization tools):
https://data.cityofnewyork.us/

social networks data sets:
http://snap.stanford.edu/data/

lists of public data sets:
http://www.kdnuggets.com/2011/02/free-public-datasets.html

Museums and other cultural institutions which offer API:
more museums put their collections online (posted on March 7, 2012).

collaboratively produced big data example: OpenStreetMap database (90 GB). Info on OpenStreetMap (Wikipedia)

**Using web as a data source** (research on your own):

Getting data from the web (general description of various strategies - from Data Journalism Handbook)

Web scraping: we used SiteSucker for some of our projects to scrape pretty big data and image sets.

Using sites which offer API - see great tutorials by Jer Thorp.

Internet archive from archive.org - 240,000,000,000 URLs, 5 TB.

Web pages data from Common Crawl Foundation - 2012 crawl corpus, 5 billion pages, 81 TB.
Common Crawl: going after Google on a non-profit budget. The Verge, 03.01.2013.

**Big Data: examples of how this topic is described and discussed:** (research on your own:

http://en.wikipedia.org/wiki/Big_data

http://www.diggingintodata.org/

Sasha Issenberg. How President Obama's campaign used big data to rally individual voters. December 19, 2012, http://www.technologyreview.com. (very detailed discussion of using big data analytics in 2012 US Elections by Obama team).

articles about "big data" in New York Times.

## Analysis of big cultural and social data in different disciplines:

**Prepare to discuss the following questions:**

are there some general research themes across these papers?

main method of analysis in each paper ? (time graphs of features; similarity/dissimilarity between data objects using features; spatial mapping and analysis; etc.)

What are the differences in data sets, research questions & methods used in these papers ?

Can we cluster the papers into a few groups based on commonality of data sets, research questions, or methods?

**Other questions to consider when reading the papers:**

size of the data? what is “big data” for different fields?
do researchers provide data sets?
do they explain the analytical procedures with enough detail that the work can be duplicated?
how is the selection of particular data is justified?
does researchers analyze only metadata or also data? (each approach has pluses and minuses)
the use and limitations of line and curve fitting to temporal data
do researches clearly discuss the limitations and possible biases of their data and analysis

**My notes on the papers (summaries and questions)**

**literary studies:**

a) Matthew Jockers. Computing and Visualizing the 19th-Century Literary Genome. DH 2012 conference.

(You can also watch a video of his talk based on this paper).

b) Ted Underwood and Jordan Sellers, The Emergence of Literary Diction. Journal of Digital Humanities, Vol. 1, No. 2 Spring 2012.

(Highly recommended: Ted Underwood’s blog.)

**linguistics and “cultunomics”:**

c) Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. 12/6/2010. (The link above is to the copy of the article in Dropbox; you can access original web version with full size illustrations if your school library has *Science*.)

Articles about the project:

New York Times. In 500 Billion Words, New Window on Culture. December 16, 2010.

New York Times. Avalanches of Words, Sifted and Sorted. March 24, 2012.

**quantitative film studies:**

d) Cutting, J. E., Brunick, K. L, DeLong, J. E., Iricinschi, C., & Candan, A. (2011). Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception, 2*, 569-576.

(Other articles by Cutting on quantitative film analysis.)

Article about Cutting:

New York Times. Bringing New Understanding to the Director's Cut. March 1, 2010.

**Sociology of culture:**

e) Currid, E. and Williams, S. The geography of buzz: art, culture and the social milieu in Los Angeles and New York. J. Econ. Geogr. (2010) 10 (3): 423-451.

Article about the project: New York Times. Mapping the Cultural Buzz: How Cool Is That? April 6, 2009.

The Spatial Information Design Lab at Columbia which developed this and many other city data projects.

Compare the patterns reported in the paper with the patterns of tourists' photos: Eric Fisher. Locals and Tourists. See also Fisher's project See something or say something.

**social computing** (social networks and social media analysis):

F) Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moo. What is Twitter, a Social Network or a News Media? WWW 2010 conference.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. Proc. of International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010.

Visit project data page and explore interactive visualization apps.

E) Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User

Generated Content Video System. Proc. of Usenix/ACM SIGCOMM Internet Measurement Conference (IMC), October 2007.

G) Justin Cranshaw, Raz Schwartz, Jason I. Hong, Norman Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. The 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland, June 2012.

See the project web site for the complete city clusters maps for a number of cities and research summary: http://livehoods.org/.

**maps of knowledge / bibliometrics**

E) Johan Bollen. Clickstream Data Yields High-Resolution Maps of Science. PLOS ONE, March 11, 2009. Large image of the visualization.

This paper is a part of the MESUR (MEtrics from Scholarly Usage of Resources) project.

Bibliometrics (Wikipedia article)

History of citation indexing (from Thomson Reuters Web of Science service)

---

# homework for class 6:

1) For students who are completely new to working with data and data visualization, read the following (and if you are familiar, browse these anyway - many good links):

Shawn Allen:

A BRIEF HISTORY OF VISUALIZATION.

CHARTED TERRITORY.

Also, read the following explanations of classic visualization techniques as they are implemented on manyeyes:

Bar Chart
Line Graph
Pie Chart
Scatterplot

Using these explanations, upload simple data sets to manyeyes and practice these visualization techniques (and others if you like).

2) Look at Annual Reports by Nicholas Felton at feltron.com - outstanding examples of how most basic visualization techniques can be used together creatively.

3) Recommended - read Michael Friendly. A Brief History of Data Visualization.

Optional: If you want to explore history of visualization further, browse through Milestones in the History of Visualization, with the particular focus on 1700s and 1800-1850 periods.

4) Watch: Hans Rosling. Stats that reshape your worldview (2006).

Explore Gapminder World.

Optional: watch other videos from TED series Making sense from too much data.

## 5) If you want to learn R and follow my R demos in classes:

a) download R to your computer

b) familiarize yourself with R interface, R data types, and basic commands:

**for people without any previous experience working with data and programing or scripting:** read and do all examples in the first few chapters of Introduction to Data Science ( explains in detail R interface and R basics, slowly step by step)

**for people with little experience,** I recommend R tutorial, the sections on Basic Input, Data Types, Basic operations.

**for people with previous experience with data and programming**, I recommend that you go through official Introduction to R

c) Once you understand R interface and basic concepts, look through and try examples of commands in **chapter 3 on Data Exploration** in R and data mining: Examples and Case Studies

---

## R - Recommended resources:

Once you understand the basics, this is the most useful resource to quickly look up how do different things with R:

**Quick R**

There are also numerous R tutorials and free books on the web you can consult. Many of these books have chapters explaining R interface and basic concepts - for example:

R programming

Introduction to Data Science

**Good introductions to making graphs in R:**

**Getting started with charts in r**

**http://www.statmethods.net/graphs/**

Here is a long list of online resources for R.

---

# class 6: visualizing numerical and categorical data / march 4 / lecture notes and links: /

classical and recent visualization techniques for one dimensional, two dimensional and multi-dimensional data; data transformations (e.g., log scale).

**Popular software for exploring data with graphs and data analysis:**
desktop: R, Mondrian, Excel.
web apps: Google Charts, manyeyes, wessa.net *(*If you are on Windows, you can also use Tableau Public*)*.

**Using basic visualization techniques to create effective infographics:**

Nicholas Felton: Annual Reports
GE Powering the Kitchen (fathom.info - Ben Fry)

## Graphic Design Principles for Information Visualization

**Standard techniques for data visualization:**

There are various ways to categorize visualization techniques. We will categorize them in this way:

**visualizing one dimensional data (single variable):**

**pie chart, bar chart**

Excel: van_gogh_summary.xlsx
Mondrian: van_gogh_data.txt

**histogram**

Mondrian: van_gogh_data.txt
explain differences between a bar chart and a histogram

**visualizing time series - line graph**

Excel: van_gogh_data.txt

**visualizing two dimensional data (two variables):**

**scatter plot**

explain differences between line graph and scatter plot
Excel: van_gogh_summary.xlsx, van_gogh_data.txt

data transformations (log, etc. - Excel graphs axis options; Mondrian: Calc > Transform)

Excel: dA.traditional_art.images_metadata.subtotals_countries.Graph.xlsx

(you can also scale the data directly using formulas before making graphs)

**visualizing multi-dimensional data (multiple variables):**

**radar plot**

Excel: van_gogh_summary.xlsx

**parallel plot**

Mondrian: van_gogh_data.txt

**mosaic plots**

(I will demo in a later class)

**plot matrix**

Mondrian - scatterplot matrix: van_gogh_additional_measurements.txt

Gauguin - radar plot matrix

**small multiple**

not supported by current versions of Excel, Mondrian, etc - but we will learn how to do it in R.

if you want to learn more:

Multi-dimensional visualization lecture

**questions:**

which graph types used to show summarized data?
which graph types are used to show complete data?
what is lost when data is summarized in a graph? What is gained?

**class R demo: loading data file, examining its properties, and creating basic visualizations:**

**recommended - R graphs references:**

Getting started with charts in r

http://www.statmethods.net/graphs/
Producing simple graphs in R

**useful general commands:**

navigating through history: upper/lower keys
getting help on any command: help(command_name)

**Reading a data file into R:**

Two methods:

1) You can read a file from any directory by specifying its complete path - for example:

>manga_data < read.csv("/Users/manovich/manga_data.csv")

2) Alternatively, you can set a ’working directory” first using RT menu”

Misc > Change Working Directory...

Once you did, you can read a tab-delimited data file using read.delim command - for example:

> vangogh <- read.delim("van_gogh_data.txt")

If your file is big, often explictly specifying options is a good idea - for instance:

>da_users <- read.table("da_users.txt", header=TRUE, sep = "\t")

**examine data objects in your R workspace using the Workspace Browser:**

Workspace > Workspace Browser

**examine the data objects in your R workspace using commands:**

>class(vangogh)
> str(vangogh)

basic summary statistics:

> summary(vangogh)

**basic scatter plot - specifying parameters:**

>plot(vangogh$Year_Month, vangogh$Saturation_Median)

// remove axis lables

```
>plot(vangogh$Year_Month, vangogh$Saturation_Median, ann=FALSE)

// change the size of dots
>plot(vangogh$Year_Month, vangogh$Saturation_Median, ann=FALSE, cex=.6)

// make a plot with one pixel points
plot(vangogh$Year_Month, vangogh$Saturation_Median, pch=".")
```

**saving a plot as PNG file:**

```
>png(file="graph_test.png", width = 1200, height = 600)
>plot(vangogh$Year_Month, van_Gogh_data$Saturation_Median, ann=FALSE,
cex=.6)
>dev.off()
```

// other techniques for saving plots from R:
// http://www.stat.berkeley.edu/users/spector/s133/saving.html

## R - additional basic techniques to work through on your own (not shown in class):

**Working with data objects and workspace:**

copy the data object into the object with a new name:

```
> vangogh_2 <- vangogh
```

examine all objects in the workspace:

```
> objects()
```

remove a particular object from the workspace:

```
>remove(vangogh_2)
```

to save current workspace with all objects:

Workspace > Save Workspace file...

to load the saved workspace (containing objects you saved):

Workspace > Load Workspace file...

examining the first few lines of the data object:

```
> head(vangogh)
```

examining the last few lines of the data object:

```
> tail(vangogh)
```

**Selecting parts of data:**

examining the first few lines of the data object:

```
> head(vangogh)
```

examining the last few lines of the data object:

```
> tail(vangogh)
```

how to refer to particular rows? This command will print rows 1-10:

```
> vangogh[1:10,]
```

how to refer to particular columns? This command will print all values in column 3:

```
> vangogh[,3]
```

create a new object which contains only 1st and 3rd - 5th columns:

```
> newdata <- vangogh[c(1,3:5)]
```

create a new object which contains only first 100 rows:

```
> newdata <- vangogh[1:100,]
```

selecting part of the data based on a single condition:

```
> vangogh_paris <- vangogh[ which(vangogh$Label_Place=='2_Paris'),]
```

**Data summaries:**

Summarize a numerical column (here, its column 3):

```
> summary(vangogh[,3])
```

Summarize integer column - here, we summarize by Year to obtain the number of rows for each year):

```
> table(vangogh$Year)
```

Summarize integer column - here, we summarize by Label_Place
to obtain the number of rows for each label:

```
> table(vangogh$Label_Place)
```

**Making plots of selected numerical data columns:**

scatter plot of 3rd and 4th columns:

```
>plot(vangogh[,3], vangogh[,4])
```

scatter plot of 3rd and 4th columns, limiting x range to 1885-1890:

```
> plot(vangogh[,2], vangogh[,3], xlim=c(1885,1890))
```

scatter plot of two columns which are next to each other:

```
> plot(vangogh[2:3])
```

scatter plot matrix (R version) - plot matrix of columns 2 through 4:

```
> pairs(vangogh[2:4])
```

histogram of a single column:

```
> hist(vangogh$Brightness_Median)
```

change the number of bins to 50:

```
> hist(vangogh$Brightness_Median, breaks = 50)
```

**Making bar plots of data summaries for factors (i.e., categorical variables):**

summarize data by Year category and make a bar plot:

```
> barplot(table(vangogh$Year))
```

customizing barplot - force horizontal layout, force printing of labels for all:

```
> barplot(table(vangogh$Year), horiz=TRUE, las=2)
```

customizing barplot - additional options: change the size of labels, add main label:

```
> barplot(table(vangogh$Year), horiz=TRUE, las=2, cex.names=0.8,
main="number of paintings per year")
```

---

# **homework for class 8-9** (march 18, april 8):

**1) if you are new to data visualization** (and recommended for all)**:**

How visualization designers work? Review descriptions of some of the projects from datavisualization.ch (feel free to look at other projects as well):

How We Visualized 23 Years of Geo Bee Contests
Visualizing The World's Well-Being
How We Visualized America's Food and Drink Spending
Visualizing The Health Care Reform

Look at:
http://flowingdata.com/2010/01/07/11-ways-to-visualize-changes-over-time-a-guide/

P.S. Also look through their other tutorials that interest you:
http://flowingdata.com/category/tutorials/

**2) if you want to master standard visualization techniques** (shown in class 6):

practice them with the same data sets used in class, or your own data sets, using any of the app(s) you prefer - Tableau, Excel, Google Charts, Numbers (on the Mac), manyeyes, Mondrian, etc. (Note that not every technique is available in every app.)

**3) If you want to continue learning R in the class:**

3.1. Work through examples of basic R commands we covered in class 8, using class notes.

3.2. Work through the examples of additional commands listed in class 6 under "**R - additional basic techniques to work through on your own"**
Use van_gogh_data.txt data set for practicing all commands in this section.

3.3. Using the reference below (or any alternative R reference listed above) become familiar with the R commands in these areas:

http://en.wikibooks.org/wiki/R_Programming/Manage_your_workspace
http://en.wikibooks.org/wiki/R_Programming/Data_types
http://en.wikibooks.org/wiki/R_Programming/Working_with_data_frames

3.4. If you have not done this yet, look through and try examples of commands in **chapter 3 on Data Exploration** in R and data mining: Examples and Case Studies

3.5. I recommend that you also work through any other commands (and their options)

for data visualization which look interesting to you in any of these references:

http://flowingdata.com/2012/12/17/getting-started-with-charts-in-r/
http://www.statmethods.net/graphs/
Producing simple graphs in R

**4) For all - visualizing time:**

**look at and prepare to discuss the**
**examples of innovative visualizations of temporal processes**
(wait for the link inside the webpage to load)

**5) For all - visualizing space:**

view as many maps as you can:
**http://pinterest.com/janwillemtulp/maps/**

view particular examples of recent maps using social media data:

atNight
Twitter NYC
Global Twitter Heartbeat
How Obama Won Re-election
Movement In Manhattan

**Think about and prepare to discuss the following questions:**
What are the common features of recent maps driven by big data and social media data?
Which maps stand out from the rest and why?
What is missing?
Do we have more techniques for visualizations of temporal processes, or more techniques in spatial data mapping? What aspects of temporal cultural processes can't be visualized with current techniques?

6) Recommended:

browse through these references/articles about "science of cities":

http://www.nytimes.com/2010/12/19/magazine/19Urban_West-t.html
http://www.nytimes.com/2013/02/24/technology/nyu-center-develops-a-science-of-cities.html?pagewanted=all
http://www.complexcity.info/

7) Recommended - historical timelines

http://www.datavis.ca/gallery/timelines.php

---

# class 8: visualizing spatial and temporal data / curve fitting with R

## Resources:

**time-based data:**

3 million time-based open data sets:
http://blog.revolutionanalytics.com/2013/02/quandl-a-wikipedia-for-time-series-data.html

R functions to use these data sets:
http://blog.revolutionanalytics.com/2013/03/quandl-package-released-to-cran.html

Help: http://www.quandl.com/help/r

**spatial data:**

https://openpaths.cc/about
http://en.wikipedia.org/wiki/OpenStreetMap

**list of visualization and mapping software:** http://selection.datavisualization.ch/

**examples of visualization software which can create maps, timelines and all other basic vis techniques:**

https://developers.google.com/chart/interactive/docs/gallery
http://d3js.org/ (currently most popular for web vis)

**example of software for creating web timelines:**

http://www.simile-widgets.org/timeline/

**examples of software for creating interactive web maps:**

http://mapbox.com/
example of maps created with mapbox (and other tools from this company:
http://mapbox.com/reinventgreen/
https://foursquare.com/

http://cartodb.com/
example of maps created with cartodb:
http://cartodb.github.com/torque/examples/uspo.html
http://mwcimpact.com/

# Practical topics:

### Fitting a line / curve to a data:

Theory:
Linear Least Square Regression (Engineering Statistics Textbook)
Trend estimation (Wikipedia)

Terms:
Regression

Least square fitting

> "The linear least squares fitting technique is the simplest and most commonly applied form of linear regression and provides a solution to the problem of finding the best fitting straight line through a set of points.
>
> In practice, the vertical offsets from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the perpendicular offsets."

The importance of data exploration using graphics as opposed to only statistical summaries:
EDA/Graphics example from Engineering Statistics Textbook

Curve fitting in Excel - how to:
Linear regression in Excel

## **Tutorial: fitting curves in R** ( and using graphic parameters )

read the data file into a data object *vangogh:*

```
>vangogh <- read.delim("van_gogh_data.txt")
```

examine the data object *vangogh*:

```
>str(vangogh)
```

plot columns 2 and 3 in *vangogh* (using smaller circles)

```
>plot(vangogh[2:3], cex=.5)
```

change the dimensions of the following graphs:

```
>dev.new(width=12, height=6)
```

plot the columns 2 and 3 of *vangogh* again in a new window:

```
>plot(vangogh[2:3], cex=.5)
```

create **fit curve** data:

```
>curve_fit_1 <- lowess(vangogh[,2], vangogh[,3], f=0.8)
```

plot the fitted curve over the previous plot:

```
> lines(curve_fit_1)
```

generated another fit curve:

```
>curve_fit_2<- lowess(vangogh[,2], vangogh[,3], f=0.5)
```

plot the fit curve, using blue color, and thick lines:

```
> lines(curve_fit_2, col=”blue”, lwd=3)
```

generate more tight fit curve:

```
>curve_fit_3 <- lowess(vangogh[,2], vangogh[,3], f=0.4)
```

plot the fit curve, using red color, and normal line width:

```
>lines(curve_fit_3, col=”red”)
```

create **linear fit** data:

```
>linear_fit <- lm(vangogh[,3] ~ vangogh[,2])
```

plot linear fit over the previous plot:

>abline (linear_fit)

examine the details of the lnear fit data:

>summary(linear_fit)

how to interpret LM output

in particular, look at *Multiple R-squared* - "the proportion of variability in a data set that is accounted for by the statistical model"

http://en.wikipedia.org/wiki/Coefficient_of_determination

**Time series analysis:**

Introduction to time series analysis from Engineering Statistics Textbook

Autocorrelation Plot

examples of autocorrelation plots for different types of data:

random data

Moderate Autocorrelation

Strong Autocorrelation

Practical Time series analysis - wessa.net

---

# homework for class 10

1) Required: Download **ImagePlot 1.1 software package**: ImagePlot macro, ImageJ application (Windows / Mac / Linux), sample data sets, theory and methodology articles.

ImagePlot_v1.1.zip

(We will learn ImagePlot in class 11; in class 10 we will use ImageJ program itself. Also, in class 10 we will only work with collections still images as opposed to video files)

After you download software, follow these instructions to increase memory setting for ImageJ:

2) Required: **familiarize yourself with basic ImageJ concepts and interface**. You can use any of these guides:

http://rsb.info.nih.gov/ij/docs/concepts.html
http://rsb.info.nih.gov/ij/docs/examples/index.html

3) Look through the examples (and descriptions) of our projects which use media visualization techniques:

http://www.slideshare.net/formalist/visualizing-image-and-video-collections-examples

4) Optional: If you want to start studying the techniques which I will be showing in classes 10 and 11, work through the guide I wrote:

Visualizing Image Sequences with ImageJ

5) Required: **Read:**

Lev Manovich. "What is Visualization?" In *Visual Studies*. Routledge, 2011.

Lev Manovich. "Media Visualization: Visual Techniques for Exploring Large Media Collections." In *Media Studies Futures*, ed. Kelly Gates. Blackwell, 2012.

---

# class 10: exploring large image collections

# (media visualization)

Sample data: Time covers, 1950-1976

Examples of historical work and artistic projects which use “media visualization” techniques:

Francis Galton

Marey

Jason Salavon

Cinema Redux

Media visualization with ImageJ:

ImageJ for media visualization - increase default memory setting

using ImageJ batch commands to measure, resize and convert image files

working with stacks in ImageJ using built-in commands:

import a file sequence into a stack
save stack as an image sequence
virtual stacks
scale stack
crop stack
montage (“montage” visualization techniques)
orthphogonal views (“slice” visualization technique)
Z project (“average” visualization technique)

using our custom ImageJ montage and slice macros:

Image-montage

Image-slice

# homework for class 12

**1) Required:**
http://www.slideshare.net/formalist/visualizing-image-and-video-collections-examples?ref=http://lab.softwarestudies.com/2008/09/cultural-analytics.html

**2) Required:**
Lev Manovich. "How to Compare One Million Images?" In David Berry, ed., *Understanding Digital Humanities* (Palgrave, 2012).

**3) Recommended:**
Lev Manovich. Style Space: How to compare image sets and follow their evolution. 2011.

**4) Recommended:** Wikipedia articles about some of the fields which use CS to analyze large media data sets:

http://en.wikipedia.org/wiki/Content-based_image_retrieval

http://en.wikipedia.org/wiki/Video_fingerprinting

http://en.wikipedia.org/wiki/Text_analytics

http://en.wikipedia.org/wiki/Music_information_retrieval

http://en.wikipedia.org/wiki/Recommendation_system

**5) Recommended:** articles about and web sites of companies doing large scale media analytics:

http://the.echonest.com/

www.bluefinlabs.com

https://de5w14y12gh72.cloudfront.net/website/bluefin_mit-tech-review.pdf

We will be using **R** in class, so if you have not touched it in a while, get ready.

---

# class 12: exploratory analysis of multi-dimensional data: feature space, distance matrix, PCA, MDS

**Concept of features and feature space**

http://stats.stackexchange.com/questions/46425/what-is-feature-space

http://pages.cs.wisc.edu/~bsettles/cs540/lectures/16_feature_spaces.pdf

examples of features and their use in text analytics:
http://www.innovation-mining.net/?q=en/taxonomy/term/15

example of features and their use in the analysis of art images:
http://imageemotion.org/machajdik_hanbury_affective_image_classification.pdf

**Problem: how to explore multi-dimensional feature space?**

**Exploratory data analysis**

http://en.wikipedia.org/wiki/Exploratory_data_analysis

**Principal Component Analysis (PCA):**

introduction to PCA for data visualization

Lecture about PCA

http://en.wikipedia.org/wiki/Principal_component_analysis

To calculate **PCA in Mondrian** directly (without going into R each time) you need to install R on your computer, and follow instructions in Mondrian software Help for Starting Rserve

When doing PCA in Mondrian, select Standardize Data (the default).

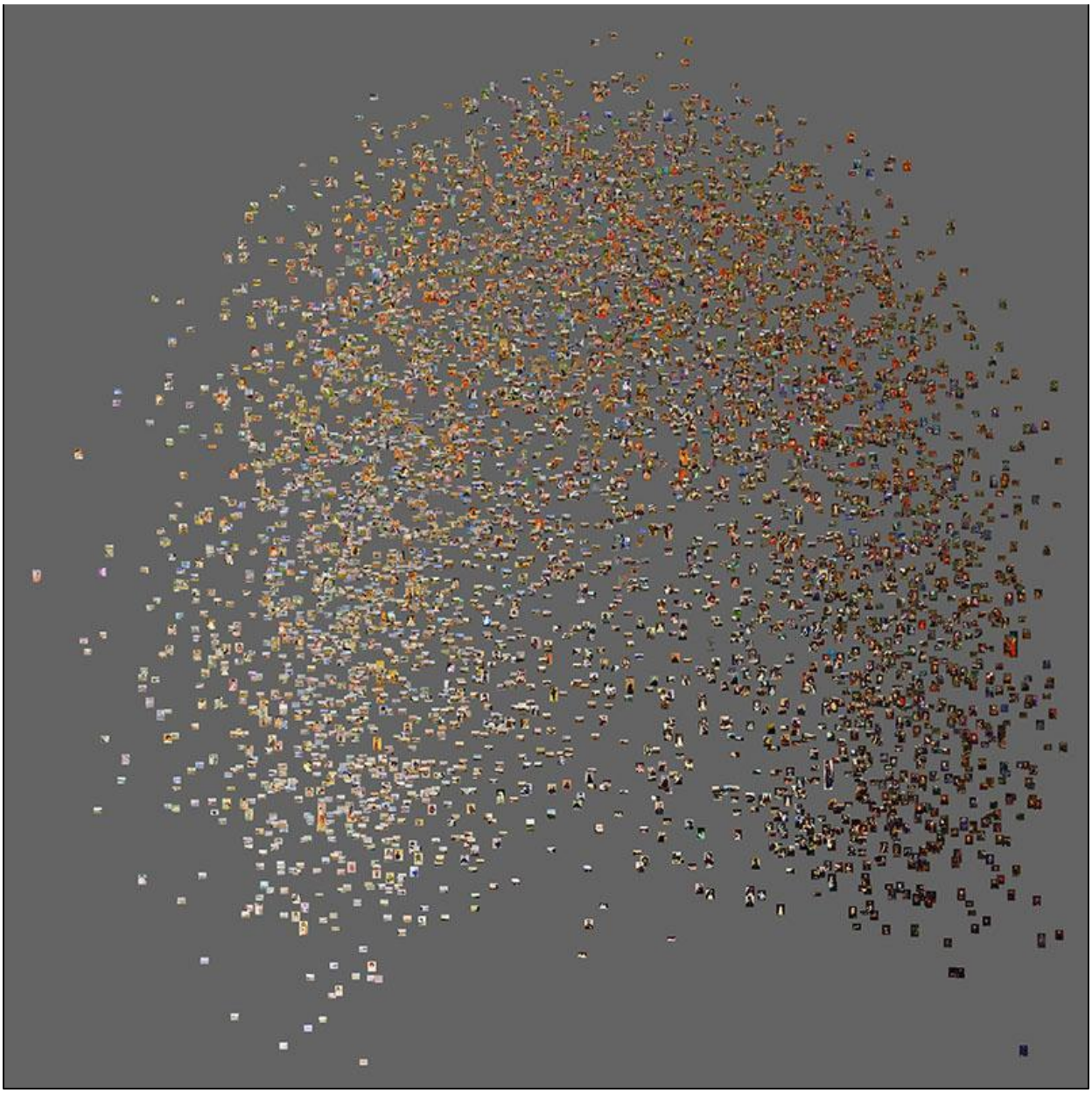
basic commands for doing PCA in R:
http://www.statmethods.net/advstats/factor.html

Example of doing PCA with image features in R:

```
impr_with_filename <- read.delim("Impressionism_light.tab.txt")
impr_filename_excluded <- impr_filename[,-1]
impr_pca_cor <- princomp(impr_filename_excluded, cor=TRUE)
```

examine results:

```
summary(impr_pca_cor) // examine the results
plot(impr_pca_cor) // show variance explained by components
plot(impr_pca_cor, type="lines") // same as line graph
biplot(impr_pca_cor) // plots 1st and 2nd components by default
loadings(impr_pca_cor) // show percentage of variables used by each component
```



Visualization of 5433 Impressionist paintings using PCA of 47 lightness features.
X-axis = PCA.1
Y-axis = PCA.2

Example of using PCA in stylometry:
http://en.wikipedia.org/wiki/Stylometry

**Multi-dimensional scaling (MDS):**

http://en.wikipedia.org/wiki/Multidimensional_scaling

doing MDS in R:
http://www.statmethods.net/advstats/mds.html

### distance matrix / similarity matrix / dissimilarity matrix

Representation of objects as points in a feature space allows us to quantify their difference.

http://en.wikipedia.org/wiki/Distance_matrix

### calculating and visualizing distance matrix in R

`dist(x, diag = TRUE)`
This function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.

`image(as.matrix(dist(x))`
plot distance matrix

`heatmap(as.matrix(dist(x))`
A heat map is a false color image (basically `image(t(x))` ) with a dendrogram added to the left side and to the top.

colors:
http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap

explanation of distance matrix in R with an example:
http://www.r-tutor.com/gpu-computing/clustering/distance-matrix

---

# homework for class 13

**Topic 1: Social science, big data, and visualization:**

Bruno Latour. TARDE'S IDEA OF QUANTIFICATION.
In Mattei Candea (editor) The Social After Gabriel Tarde: Debates and Assessments, Routledge, London, pp. 145-162, 2010.

Recommended background:
1) Victoria Coven. A History of Statistics in the Social Sciences, 2003.

2) Wikipedia article on Durkheim: Establishing sociology, Methodology, Social Facts

**Topic 2: Linguistic and cultural categories, and visualization:**

Wikipedia article on Prototype theory (lingustics):
http://en.wikipedia.org/wiki/Prototype_(linguistics)