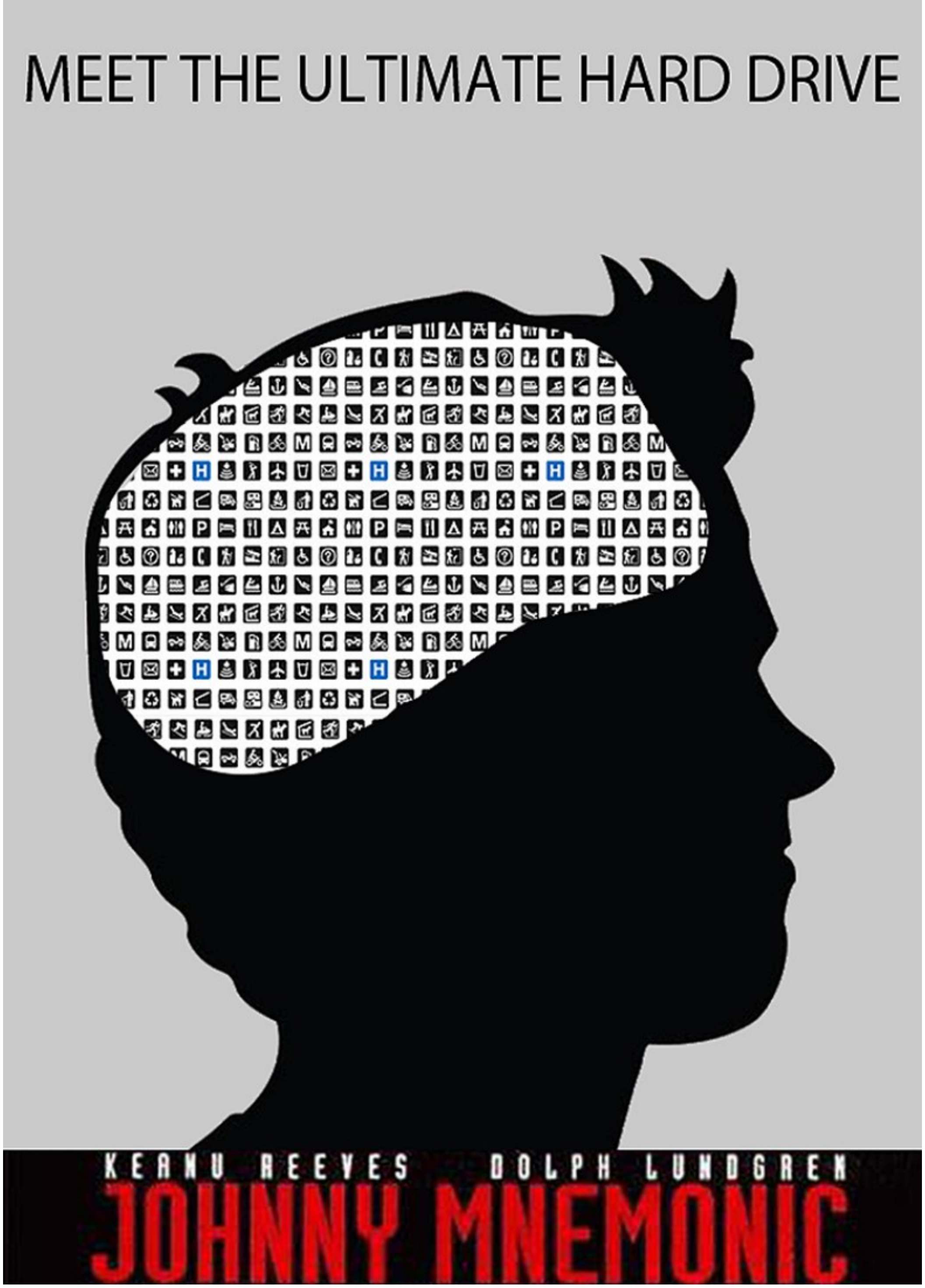
# Mining the Social Web

Prof. Emilio Ferrara / emiliofe@usc.edu / www.emilio.ferrara.name
AI: TBD
Office hours: Friday, 1h after class; or by appointment



Johnny Mnemonic (1995) — © TriStar Pictures

## Course description and learning objectives

Learn how to unleash the full power and potential of Social Web data for research and business application purposes!

The Social Web pervades all aspects of our lives: we connect and share with friends, search for jobs and opportunities, rate products and write reviews, establish collaborations and projects, all by using online social platforms like Facebook, LinkedIn, Yelp and GitHub. We express our personality and creativity through social platforms for visual discovery, collection and bookmarking like Tumblr and Pinterest. We keep up-to-date, communicate and discuss news and topics of our interest on Twitter and Reddit.

In this course we will explore the opportunities provided by the wealth of social data available from these platforms. You will learn how to acquire, process, analyze and visualize data related to social networks and media activity, users and their behaviors, trends and information spreading. This journey will bring through the lands of data mining and machine learning methods: supervised and unsupervised learning will be applied to practical problems like social link analysis, opinion mining, and building smart recommender systems. We will explore open-source tools to understand how to extract meaning from human language, use network analysis to study how human connect, and discover affinities among people's interests and tastes by building interest graphs.

Taking this course, you should expect to learn about:

- Supervised learning: Crush course on Data Classification.
  - Eager vs. Lazy learning: Decision Tree and k-Nearest Neighbors.
  - Probabilistic models: Näive Bayes classifier.
  - Ensemble methods, bagging and boosting: Random Forest and AdaBoost.
  - Classification performance evaluation: Precision/Recall/F1, Accuracy and ROC Curves.
- Unsupervised learning: Crush course on Clustering Data.
  - Distance and similarity measures & K-means clustering.
  - Hierarchical Clustering and Dendrograms.
  - Density-based clustering.
  - Clustering performance evaluation.
- Applications of texts and documents analysis.
  - Natural Language Processing and Part-of-speech tagging.
  - Sentiment Analysis.
  - Topic Modeling.
- Networks:
  - Statistical descriptors of networks: link analysis, centrality, and prestige.
  - Network clustering: modularity and community detection.
  - Dynamics of information and epidemics spreading: threshold and information cascade models.
  - Network visualization algorithms: spring-like layouts, multidimensional scaling, Gephi.
- Collective intelligence:
  - Recommender systems & Collaborative filtering
  - Matrix factorization

All topics will be explored from an applied, practical, computational perspective. This will allow the interested student to deepen the rigorous theoretical implications of the methods in other courses offered by USC (for example, CSCI-567 Machine Learning). Throughout this course we will deliver several "hands-on" sessions with live coding, data analysis, and problem solving!

## Prerequisites

A basic understanding of programming that will allow you to manipulate data and implement basic algorithms, using any programming language, is recommended. Python will be the "official" programming language used during the hands-on sessions and for learning purposes. We will use IPython Notebook as environment. However, feel free to use the language you prefer for your assignments and class project. A basic understanding of statistics and algebra will help too.

## Books and learning material

Required textbooks (total Amazon price [new/used]: $100/$60)

1. Web Data Mining (2nd Ed.) —by Bing Liu (Amazon price [new/used]: $48/$35)
2. Mining the Social Web (2nd Ed.) —by Matthew A. Russell (Amazon price [new/used]: $27/$15)
3. Programming Collective Intelligence —by Toby Segaran (Amazon price [new/used]: $25/$10)
4. Network Science Book —by Laśzló Barabási (FREE: `http://barabasilab.neu.edu/networksciencebook/`)
5. Dive into Python —by (FREE: `http://www.diveintopython.net/`)

Some details: (1) will provide insights on methods and approaches studied throughout the course from a machine learning perspective; (2) and (3) will serve as recipe books to effectively design and make those methods work with Social Web data; (4) and (5) are free resources we will exploit to gather additional material on networks and Python programming.

Technical, recommended (non-required) Python "cookbooks":

- Python Data Visualization Cookbook —by Igor Milovanović (ebook: $14)
- Learning IPython for Interactive Computing and Data Visualization —by Cyrille Rossant (ebook: $10)
- Learning scikit-learn: Machine Learning in Python —by Raúl Garreta and Guillermo Moncecchi (ebook: $10)

## Policy & Grading

Class participation and engagement are essential ingredients for success in your academic career, therefore during class turn off cellphones and ringers (no vibrate mode), laptops and tablets. The only exception to use laptops during class is to take notes. In this case, please sit in the front rows of the classroom: no email, social media, games, or other distractions will be accepted. Students will be expected to do all readings and assignments, and to attend all meetings unless excused, in writing, at least 24 hours prior. This is the (tentative) system that will be employed for grading:

| Component | Weight | Description |
|---|---|---|
| Participation | 20% | Class participation, weekly presentation, and engagement. Attendance is mandatory. |
| Assignments | 20% | Five assignments on social Web data analysis and modeling. |
| Midterm exam | 30% | Mid-term Hackathon (grading will be informed by a peer-review system). |
| Final exam | 30% | Final project paper. |

The following misconducts will automatically result in a zero weight for that component of the grade: (1) failing to attend class on the day of your presentation; (2) failing to turn in the assignments by the expected dates; (3) failing to attend meetings of your group's Hackathon and/or final presentation; (4) failing to submit your final paper by the expected date. Extenuating circumstances will normally include only serious emergencies or illnesses documented with a doctor's note.

## Readings & discussion

At the beginning of each lecture (starting lecture 2), one student will hold a 10m presentation on one daily reading and moderate a 5m discussion about it. The list of required readings is available at the end of the syllabus and at `http://www.emilio.ferrara.name/i400-590-mining-the-social-web/`

## Assignments

Throughout the course there will be five assignments to be carried out independently by each student. The goal of these assignments is to allow you to track your own progresses and understand whether you are grasping the essential concepts of the course. They will occur tentatively at the end of each of the five parts the course plan (see Syllabus). The assignments will consist of part "theory" (including material from the mandatory readings) and part coding tasks. They will be based on topics, problems and questions discussed during class each week.

## Mid-Term Hackathon

The mid-term exam is in the form of a collaborative hackathon.[1] The goal is to develop crucial abilities such as:

- Intellectual development: leveraging expertise and multidisciplinary backgrounds, sharing ideas and knowledge.
- Team work skills: effective brainstorming, communication and presentation, and group problem solving.
- Project management skills: ability to set goals, map progress, prototyping-delivery, and matching deadlines.

We expect participants to form groups of 3 or 4 members with the goal of solving a single problem. Graduate students are encouraged to form groups of three. Each group will receive a different problem.

We will propose several problems of interest for the course, as well as receive your *explicit solicitations*, that should be submitted by 3 days from the beginning of the mid-term week, in the form of a short one-page proposal clearly stating:

- What is the problem.
- Why it is deemed relevant.
- How the group plans to solve the problem.
- Bibliographic references to at least one relevant related paper.

All project proposals will be subject to our approval. Groups will be assigned an approved project, either selected among those proposed by the Instructor, or by the group itself. The rules of the hackathon will be released the week before the mid-term. Each group will receive a 15m slot for presentation of their results, in which each member of the group is expect to discuss at least one critical task of the project. The grading of the projects will be in part based on crowd-sourced ratings attributed by other fellow students and submitted in anonymous form at the end of each presentation day.

## Final Paper

A serious final paper will be expected. The manuscript will be at least 2,500 words, and will include appropriate figures and tables. The work should cover the following points:

- Statement of the problem & Why the problem is important.
- How the problem was faced —including a description of methodology and dataset(s).
- Discussion of results, findings and limitations of the study.
- Related literature & Final remarks/conclusions.

Ideally, the final paper is based on the student's mid-term hackathon project. Text with other group members cannot be shared, figures/tables can be shared when appropriate with proper credit attribution. Grading will be based on soundness (both quality and quantity of original work). Groups composed by three graduate students will be allowed (upon written permission) to turn in a single joint-authored manuscript, in the format of a submission for an appropriate peer-reviewed journal or conference. Each author must contribute sufficient material to justify his/her "equal contribution" in the work.

[1] http://en.wikipedia.org/wiki/Hackathon

## Academic integrity

The principles of academic honesty and professional ethics will be vigorously enforced in this course, following the USC Code of Student Rights, Responsibilities, and Conduct, the Informatics Academic Regulations, and the CS Program Statement on Academic Integrity.

This includes the usual standards on acknowledgment of help, contributions and joint work, even when you are encouraged to build on libraries and other software written by other people. Any code or other assignment you turn in for grading and credit must be your individual work (except for group projects). Even if you work with a study group (which is encouraged), the work you turn in must be exclusively your own. If you turn in work done together with, or with the assistance of, anyone else other than the instructors, this is an instance of cheating.

Cases of academic misconduct (e.g., cheating, fabrication, plagiarism, interference, or facilitating academic dishonesty) will be reported to the Office of the Dean of Students. Typical consequences include an automatic F grade in the course.

Your submission of work to be graded in this class implies acknowledgement of this policy. If you need clarification or have any questions, please see the instructor during office hours.

## Final remarks

We would like to hear from anyone who has a disability or other issues that may require some modifications or class adjustments to be made. The offices of Disability Services and Psychological Services are available for assistance to students. Please see the instructor after class or during office hours.

We welcome feedback on the class organization, material, lectures, assignments and exams. You can provide us with constructive criticism. Please share your comments and suggestions so that we can improve the class.

## On the cover: Johnny Mnemonic

Johnny Mnemonic is a cyberpunk sci-fi cult movie from the early nineteens, adapted from the homonymous short story by William Gibson. It tells the story of Johnny, a data courier, nicely interpreted by a Keanu Reeves in his early career, struggling with a huge load of data stored in his head. I find it a nice metaphor of our current data-overloaded society, and, coincidentally, is one of the defining movies I watched as a teenager that brought me to love Computer Science.

**[version 0.2: Jun. 25, 2016]**

# Syllabus

## Part 1—Supervised Learning

http://sctr7.com/2013/08/17/data-science-as-an-experimental-process-unsupervised-and-supervised-learning/

### Week One

- Introduction of the course & Crash intro to Supervised learning.
  Readings: Papers [16] and [11] — Chapters: WDM:3.1
- Eager vs. Lazy learning—Decision Tree & k-Nearest Neighbors.
  Readings: Papers [20] — Chapters: WDM:3.2 and WDM:3.9

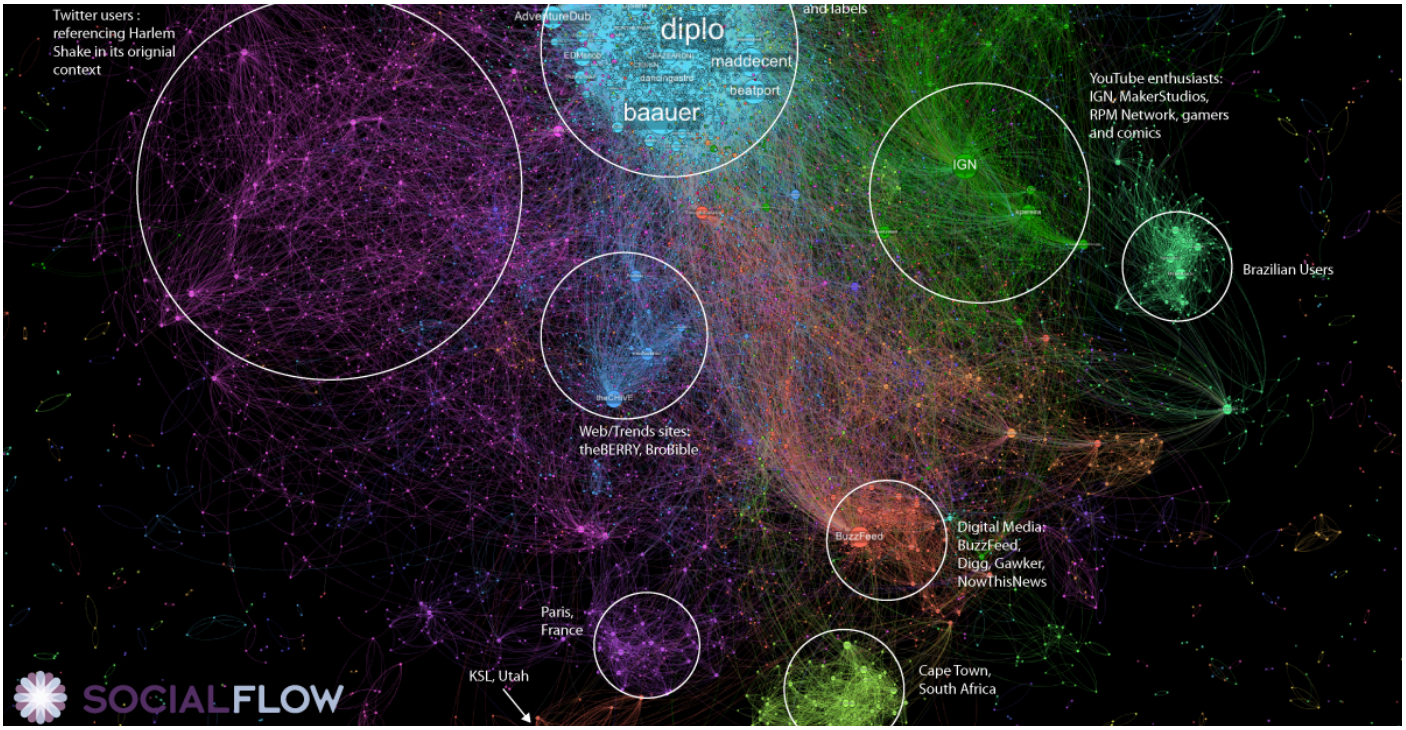
### Week Two

- Ensemble methods, bagging and boosting & Classification performance evaluation.
  Readings: Papers [9] — Chapters: WDM:3.3 and WDM:3.10
- *hands-on session*: mining Twitter.
  Readings: Papers [12] — Chapters: MtSW:1[pp.5-26]
  Documentation: Twitter API (`https://dev.twitter.com/`)

### Week Three

- *hands-on session*: mining Twitter.
  Readings: Papers [28] — Chapters: MtSW:1[pp.26-44]
- *hands-on session*: mining Twitter.
  Readings: Papers [33] — Chapters: PCI:7[pp.142–165]

# Part 2—Unsupervised Learning


Twitter users : referencing Harlem Shake in its orignial context
AdventureDub
diplo
EDMsnob
maddecent
dancingastro
beatport
baauer
and labels
YouTube enthusiasts: IGN, MakerStudios, RPM Network, gamers and comics
IGN
Brazilian Users
theCHIVE
Web/Trends sites: theBERRY, BroBible
BuzzFeed
Digital Media: BuzzFeed, Digg, Gawker, NowThisNews
Paris, France
KSL, Utah
Cape Town, South Africa
SOCIALFLOW

http://giladlotan.com/2013/03/harlem-shake-anatomy-of-a-viral-meme/

## Week Four

- Crash introduction to Unsupervised learning—Distance measures & K-means clustering.
  Readings: Papers [34] and [35] — Chapters: WDM:4.1–4.3[pp.133–147]
- *hands-on session*: mining Twitter.
  Readings: Papers [18] — Chapters: 9[pp.351–396]

## Week Five

- Hierarchical clustering & Dendrograms.
  Readings: Papers [23] — WDM:4.3–4.5[pp.147–155]
- *hands-on session*: mining LinkedIn.
  Readings: Papers [32] — Chapters: MtSW:3[pp.89–132]
  Documentation: LinkedIn API (`https://developer.linkedin.com/apis`)

## Week Six

- Density-based clustering & Clustering performance evaluation
  Readings: Papers [29] — Chapters: WDM:4.6–4.10[pp.155–165]
- *hands-on session*: mining LinkedIn.
  Readings: Papers [13] — PCI:3[pp.29–53]

# Part 3—Text and Documents

*Image Credit: carlos castilla / Shutterstock*

## Week Seven

- Natural Language Processing & Part-of-Speech Tagging.
  Readings: Papers [15] — Chapters: WDM:6.5 and MtSW:5.3–5.5[pp.190–222]
- Sentiment Analysis & *hands-on session*: mining Google+.
  Readings: Papers [14] — Chapters: MtSW:4[pp.135–180]
  Documentation: Google+ API (`https://developers.google.com/+/api`)

## Week Eight

- Topic modeling.
  Readings: Papers [2] — Chapters: WDM:6.7
- *hands-on session*: mining Instagram.
  Readings: Papers [5] and [6]
  Documentation: Instagram API (`http://instagram.com/developer/`)

## Week Nine: Mid-term Hackathon week

- Mid-term Hackathon presentations
- Mid-term Hackathon presentations

Fall break: TBD

# Part 4—Networks

https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

## Week Ten

- Crash introduction to Networks—Statistical descriptors of networks.
  Readings: Papers [21] and [4] — Chapters: NSB:1 and NSB:2
- *hands-on session*: mining Facebook.
  Readings: Papers [7] — Chapters: WDM:7.1 and WDM:7.3–7.4 and MtSW:7[pp.279–320]
  Documentation: Facebook API (`https://developers.facebook.com/`)

## Week Eleven

- Network clustering.
  Readings: Papers [26] and [30] — Chapters: NBS:9 and WDM:7.5
- *hands-on session*: mining Facebook.
  Readings: Papers [10] — Chapters: MtSW:2[pp.45–86]

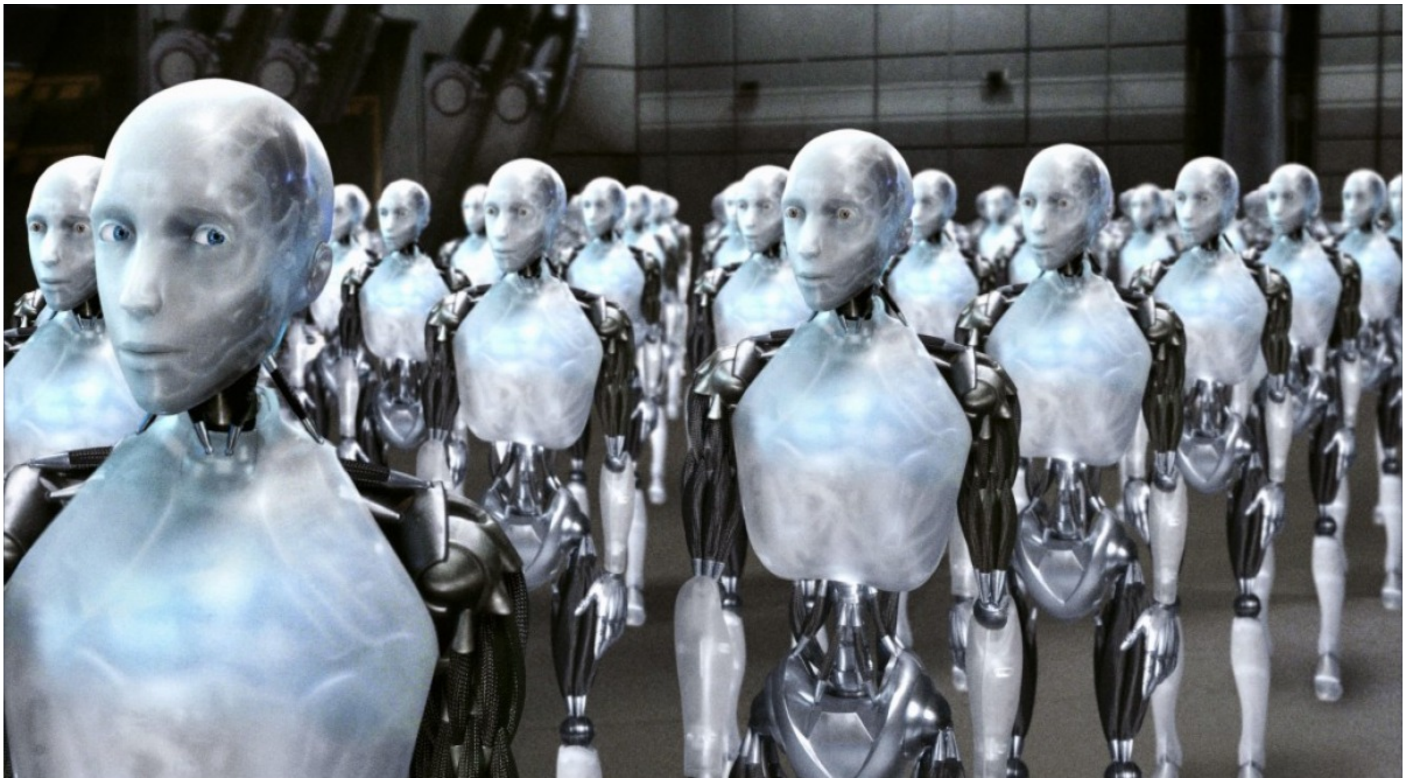
## Week Twelve

- Dynamics of information and epidemics spreading.
  Readings: Papers [25] — Chapters: NSB:10.1–10.3[pp.11–29]
- *hands-on session*: tutorial on Gephi.
  Readings: Papers [3] and [19] — Chapters: NSB:10.4–10.7[pp.30–58]
  Documentation: Gephi Wiki (`https://wiki.gephi.org/index.php/Main_Page`)

## Week Thirteen

- Network visualization algorithms.
  Readings: Papers [1] and [27] — Chapters: PCI:12[pp.300–302(MDS)]
- *hands-on session*: tutorial on Gephi.
  Readings: Papers [24] and [8]

# Part 5—Collective Intelligence



I, robot (2004) — © Twentieth Century Fox Film Corporation

## Week Fourteen

- Recommender systems: Collaborative filtering algorithm.
  Readings: Papers [31] and [17] — Chapters: WDM:12.4
- Recommender systems: Non-negative Matrix Factorization algorithm.
  Readings: Papers [22] — Chapters: PCI:10[pp.226–249]

## Week Fifteen

- Project presentations.
- Project presentations.

## Week Sixteen: Finals Week

- Final project poster presentations (to be confirmed).

## Reading list

All papers are linked at `http://www.emilio.ferrara.name/i400-590-mining-the-social-web/`

[1] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[2] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[4] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[5] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[6] D. Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[7] A. Cho. Ourselves and our interactions: the ultimate physics problem? *Science*, 325(5939):406, 2009.

[8] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[9] V. Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.

[10] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.

[11] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[12] W. Fan and M. D. Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.

[13] M. T. Gastner and M. E. Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499–7504, 2004.

[14] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[15] R. F. i Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

[16] N. Jones. Computer science: The learning machines. *Nature*, 505(7482):146, 2014.

[17] Y. Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[18] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[19] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, page 201320040, 2014.

[20] D. Lazer, R. Kennedy, G. King, and A. Vespignani. Big data. the parable of google flu: traps in big data analysis. *Science*, 343(6176):1203, 2014.

[21] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

[22] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[23] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[24] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[25] P. T. Metaxas and E. Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

[26] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[27] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[28] R. Nuzzo. Scientific method: statistical errors. *Nature*, 506(7487):150–152, 2014.

[29] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[30] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[31] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[32] M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing. A network framework of cultural history. *Science*, 345(6196):558–562, 2014.

[33] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[34] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425, 2009.

[35] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, 2012.