# USC Viterbi School of Engineering

## CSCI 648: Information Integration on the Web, Spring 2019

**Units: 4**

Class: Monday, Wednesday, 5:00-6:50pm, GFS113
Final Exam: Friday, May 1, 4:30-6:30 p.m, GFS113

**Instructor:** Jose Luis Ambite
**Office Hours:** Monday & Wednesday after class, or by appointment
**Contact Info**: ambite@isi.edu, 310-448-8472.

**Instructor:** Shobeir Fakhraei
**Office Hours:** Monday & Wednesday after class, or by appointment
**Contact Info:** fakhraei@usc.edu, 310-448-9341.

**Teaching Assistant:** Roohy Shemirani
**Office Hours:** TBD
**Contact Info:** rshemira@usc.edu

**Prerequisite(s):** CSCI 561
**Co-Requisite (s):** none
**Concurrent Enrollment:** none
**Recommended Preparation**: CSCI 585 and programming experience

### Course Description

This course covers foundations and advanced techniques for information extraction, data modeling and integration. Topics include logical data integration, entity linkage, schema mappings, source modeling, data cleaning, information extraction, knowledge representation, Semantic Web (RDF, OWL, SPARQL) and linked data.

The class emphasizes automating the data integration process as much as possible, which generally requires the use of machine learning methods, including techniques such as probabilistic graphical models and deep learning.

The class will be run as a mixture of lectures by the instructors and reading/presentations of advance materials by student participants. As an integral part of the course each student will do a project using the research and tools covered in the class. Ideally, students will apply the methods learned in the course to their area of research or interest.

*In a nutshell, this course deals with everything you need to do to prepare your data for meaningful analysis/machine learning, which often takes 80% of the time of a data science project.*

### Learning Objectives

- Understand theory and techniques of data integration, including logical data integration and entity linkage.
- Understand approaches for information extraction from unstructured (e.g., natural language text, or ungrammatical text like classified ads or tweets) and semi-structured (e.g., HTML) data.
- Understand the theory and application of the state-of-the-art software and tools for data cleaning and data normalization.
- Understand approaches for automatic source modeling and learning logical schema mappings.
- Understand the foundations and techniques of the Semantic Web, including knowledge representation languages like RDF and OWL; query languages like SPARQL; and data modeling approaches such as Linked Data.
- For any given integration problem, be able to select and apply the most relevant information integration techniques to solve that problem

## Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in a language such as Java, C++, or Python. Students are also expected to have their own laptop or desktop computer where they can install and run software to do the homework assignments.

## Readings and Supplementary Materials

Textbook: Principles of Data Integration by Doan, Halevy, & Ives, Morgan Kaufmann, 2012
The book is available online at no cost from the USC library at:
http://www.sciencedirect.com/science/book/9780124160446
and is also available for purchase.

Required and optional readings are listed in the course schedule. We also list some readings to be presented by students, though these may be changed with novel developments as needed.

## Description and Assessment of Assignments

**Class readings/presentations**: Students will read recent research papers and present in class. The number of presentations would depend on enrollment, but we expect at least 2 presentations per student.

**No Homeworks:** The class will not have graded homeworks. We will make the homeworks of CSCI548 available to students that want to practice the techniques in the course in a guided fashion. We expect the work in the project to provide in-depth practical experience in the topics in the course.

**Midterm**: There is no mid-term for this class.

**Final Exam:** There is a final exam at the end of the semester covering all of the material covered in the class.

**Course Project:** An integral part of this course is the course project, which builds on the topics and techniques covered in the class. Students can work individually or in teams of up to two people on this project. They will present a project proposal in class, conduct the project, and present the project in class. Ideally the project contributes significantly to the student's research area and leads to a publication submission in a leading conference. (Paper acceptance is not required for full grade).

## Grading Breakdown

| | |
|---|---|
| Class Presentations | 30% |
| Final | 20% |
| Class Project | 50% |
| Total | 100% |

Course grades will range from A through F. The following is the breakdown for grading:

| | |
|---|---|
| 94 - 100 = A | 74 – 76.9 = C |
| 90 – 93.9 = A- | 70 – 73.9 = C- |
| 87 – 89.9 = B+ | 67 – 69.9 = D+ |
| 84 – 86.9 = B | 64 – 66.9 = D |
| 80 – 83.9 = B- | 60 – 63.9 = D- |
| 77 – 79.9 =C+ | Below 60 is an F |

## Course Schedule

| | Topics | Readings | Homeworks | Instructor |
|---|---|---|---|---|
| Jan 7 | Course Introduction | | | Prof. Ambite |
| Jan 9 | Wrapper Generation | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 9. Morgan Kaufmann, 2012. http://www.sciencedirect.com/science/book/9780124160446<br><br>Optional:<br><br>T Furche, G Gottlob, G Grasso, X Guo, G Orsi, C Schallhart, C Wang. DIADEM: thousands of websites to a single database. Proceedings of the VLDB Endowment 7 (14), 1845-1856. www.vldb.org/pvldb/vol7/p1845-furche.pdf<br><br>T Furche, J Guo, S Maneth, C Schallhart. Robust and noise resistant wrapper induction. Proceedings of the 2016 International Conference on Management of Data, 773-784 http://christian.schallhart.net/publications/2016--sigmod--robust-and-noise-resistant-wrapper-induction.pdf<br><br>Ion Muslea, Steve Minton, and Craig A. Knoblock. A hierarchical approach to wrapper induction. In Proceedings of the 3rd International Conference on Autonomous Agents, Seattle, WA, 1999. http://www.isi.edu/integration/papers/muslea99-agents.pdf.<br><br>W. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner. Towards automatic data extraction from large web sites. 2001. http://www.vldb.org/conf/2001/P109.pdf.<br><br>B. Cenk Gazen and Steven Minton. Overview of autofeed: An unsupervised learning system for generating webfeeds. In Proceedings of AAAI, 2006. http://www.isi.edu/integration/courses/csci548/Papers/gazen06-aaai.pdf<br><br>T Furche, G Gottlob, G Grasso, X Guo, G Orsi, C Schallhart, C Wang. DIADEM: thousands of websites to a single database. Proceedings of the VLDB Endowment 7 (14), 1845-1856. www.vldb.org/pvldb/vol7/p1845-furche.pdf<br><br>T Furche, J Guo, S Maneth, C Schallhart. Robust and noise resistant wrapper induction. Proceedings of the 2016 International Conference on Management of Data, 773-784 http://christian.schallhart.net/publications/2016--sigmod--robust-and-noise-resistant-wrapper-induction.pdf | Optional HW1: Wrappers | Prof. Ambite |
| Jan 14 | Information Extraction 1: (Introduction, CRF, and Name Entity Extraction) | Andrew McCallum. Information Extraction: Distilling Structured Data from Unstructured Text. ACM Queue, volume 3, Number 9, November 2005. http://people.cs.umass.edu/~mccallum/papers/acm-queue-ie.pdf<br><br>Hanna M. Wallach, Conditional Random Fields: An Introduction, 2004. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&co | | Dr. Fakhraei |

| | | | | |
|---|---|---|---|---|
| | | ntext=cis_reports<br><br>Optional:<br><br>Sarawagi, Sunita. "Information extraction." Foundations and Trends in Databases 1.3 (2008): 261-377 (Parts 1-3). http://pages.cs.wisc.edu/~anhai/courses/784-fall13/ieSurvey.pdf<br><br>Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." Foundations and Trends® in Machine Learning 4.4 (2012): 267-373. http://faculty.cse.tamu.edu/huangrh/Fall16/crf_tut.pdf<br><br>Ratinov, Lev, and Dan Roth. "Design challenges and misconceptions in named entity recognition." Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009. http://www.aclweb.org/anthology/W09-1119 | | |
| Jan 16 | Information Extraction 2 (Name Entity Extraction, Relation Extraction) | Sarawagi, Sunita. "Information extraction." Foundations and Trends in Databases 1.3 (2008): 261-377 (Part 4). http://pages.cs.wisc.edu/~anhai/courses/784-fall13/ieSurvey.pdf<br><br>Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, Open Information Extraction from the Web, 2007. https://homes.cs.washington.edu/~soderlan/OpenIE-ijcai07.pdf<br><br>Student Presentations:<br><br>Suwisa Kaewphan, Farrokh Mehryary, Kai Hakala, Tapio Salakoski and Filip Ginter. TurkuNLP Entry for Interactive Bio-ID Assignment. Proceedings of BioCreative VI, 2017.<br><br>Fader, Anthony, Stephen Soderland, and Oren Etzioni. "Identifying relations for open information extraction." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011. http://reverb.cs.washington.edu/emnlp11.pdf<br><br>T. Mitchell, W. Cohen, et. al. Never-Ending Learning, In Proceedings of the Conference on Artificial Intelligence (AAAI), 2015. http://www.cs.cmu.edu/~tom/pubs/NELL_aaai15.pdf | Optional HW2: Information Extraction | Dr. Fakhraei |
| Jan 21 | NO CLASS | Martin Luther King Day | | |
| Jan 23 | Information Extraction 3 (Ungrammatical, Unstructered) | Matthew Michelson and Craig A. Knoblock. Semantic Annotation of Unstructured and Ungrammatical Text. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-2005), Edinburgh, Scotland, 2005. http://www.isi.edu/integration/papers/michelson05-ijcai.pdf<br><br>Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011. | | Dr. Fakhraei<br><br>Prof. Ambite |

| | | | | |
|---|---|---|---|---|
| | | http://www.aclweb.org/anthology/D11-1141<br><br>Ritter, Alan, Oren Etzioni, and Sam Clark. "Open domain event extraction from twitter." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.<br>https://homes.cs.washington.edu/~mausam/papers/kdd12.pdf | | |
| Jan 28 | Information Extraction 4 (Deep Learning Methods) | Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional LSTM-CNN-CRF." arXiv preprint arXiv:1603.01354 (2016).<br>https://arxiv.org/pdf/1603.01354.pdf<br><br>Student Presentations:<br><br>Li, Jing, et al. "A Survey on Deep Learning for Named Entity Recognition." arXiv preprint arXiv:1812.09449 (2018).<br>https://arxiv.org/pdf/1812.09449.pdf<br><br>Yadav, Vikas, and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models." Proceedings of the 27th International Conference on Computational Linguistics. 2018.<br>http://aclweb.org/anthology/C18-1182<br><br>Optional:<br><br>Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." Transactions of the Association for Computational Linguistics 4 (2016): 357-370.<br>https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00104<br><br>Understanding LSTM Networks<br>http://colah.github.io/posts/2015-08-Understanding-LSTMs/<br><br>3BLUE1BROWN Deep Learning Video Series:<br>https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi | | Dr. Fakhraei |
| Jan 30 | Entity Linkage 1: (String Matching) | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, Chapter 4 (String Matching). Morgan Kaufmann, 2012.<br>http://www.sciencedirect.com/science/book/9780124160446 | | Dr. Fakhraei |
| Feb 4 | Entity Linkage 2: (Overview) | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, Chapter 7 (Data Matching). Morgan Kaufmann, 2012.<br>http://www.sciencedirect.com/science/book/9780124160446 | Optional HW3: Entity Linkage | Dr. Fakhraei |
| Feb 6 | Entity Linkage 3: (Advanced Topics) | Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets, Chapter 3 (Finding Similar Items).<br>http://infolab.stanford.edu/~ullman/mmds/ch3.pdf<br><br>Singla, Parag, and Pedro Domingos. "Entity resolution with markov logic." Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006.<br>https://alchemy.cs.washington.edu/papers/singla06b/singla06b.pdf | | Dr. Fakhraei |

| | | | | |
|---|---|---|---|---|
| | | Optional:<br><br>Datar, Mayur, et al. "Locality-sensitive hashing scheme based on p-stable distributions." Proceedings of the twentieth annual symposium on Computational geometry. ACM, 2004. http://www.cs.princeton.edu/courses/archive/spring05/cos598E/bib/p253-datar.pdf | | |
| Feb 11 | Entity Linkage 4: (Deep Learning Methods) | Fakhraei, Shobeir, and Jose Luis Ambite. "NSEEN: Neural Semantic Embedding for Entity Normalization." arXiv preprint arXiv:1811.07514 (2018). https://arxiv.org/pdf/1811.07514.pdf<br><br>Ebraheem, Muhammad, et al. "Distributed representations of tuples for entity resolution." Proceedings of the VLDB Endowment 11.11 (2018): 1454-1467. http://da.qcri.org/ntang/pubs/vldb18-deeper.pdf<br><br>Student Presentation:<br><br>Mudgal, Sidharth, et al. "Deep Learning for Entity Matching: A Design Space Exploration." Proceedings of the 2018 International Conference on Management of Data. ACM, 2018. http://pages.cs.wisc.edu/~anhai/papers1/deepmatcher-tr.pdf | | Dr. Fakhraei |
| Feb 13 | Data Cleaning (openrefine, Wrangler, trifecta, tamr) Normalization | Erhard Rahm, Hong Hai Do. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin, 2000. https://dbs.uni-leipzig.de/file/TBDE2000.pdf<br><br>Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. http://vis.stanford.edu/papers/wrangler.<br><br>Open Refine, Explore data. http://youtu.be/B70J_H_zAWM.<br><br>Open Refine, Clean and transform data. http://youtu.be/cO8NVCs_Ba0.<br><br>Open Refine, Reconcile and match data. http://youtu.be/5tsyz3ibYzk.<br><br>Student Presentations:<br><br>Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Nan Tang. The Data Civilizer System. CIDR 2017<br><br>Essam Mansour, Dong Deng, Raul Castro Fernandez, Abdulhakim Ali Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang. Building Data Civilizer Pipelines with an Advanced Workflow Engine. ICDE 2018: 1593-1596 | Optional HW4: Data Cleaning | Prof. Ambite |

| | | | | |
|---|---|---|---|---|
| | | Optional:<br><br>Bo Wu, Pedro Szekely, and Craig A. Knoblock. Minimizing user effort in transforming data by example. In Proceedings of the International Conference on Intelligent User Interface, 2014. http://www.isi.edu/integration/papers/wu14-iui.pdf<br><br>Wu, B.; and Knoblock, C. A. An Iterative Approach to Synthesize Data Transformation Programs. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015. http://usc-isi-i2.github.io/papers/wu15-ijcai.pdf | | |
| Feb 18 | NO CLASS | President's Day | | |
| Feb 20 | Database Theory Basics | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 2.1, 2.2, 2.3 and 2.4. Morgan Kaufmann, 2012. http://www.sciencedirect.com/science/book/9780124160446<br><br>Student Presentations:<br><br>Michael Stonebraker, Ihab F. Ilyas. Data Integration: The Current Status and the Way Forward. IEEE Data Eng. Bull. 41(2): 3-9 (2018) http://sites.computer.org/debull/A18june/p3.pdf | | Prof. Ambite |
| Feb 25 | Logical Data Integration 1 (Query Rewriting) | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 2.4, 3.1, 3.2, 3.3, 3.4. Morgan Kaufmann, 2012. http://www.sciencedirect.com/science/book/9780124160446<br><br>Student Presentations:<br><br>Alin Deutsch, Lucian Popa, Val Tannen "Query reformulation with constraints". SIGMOD Record (SIGMOD) 35(1):65-73 (2006).<br><br>Lucian Popa, Alin Deutsch, Arnaud Sahuguet, and Val Tannen. "A chase too far?" In ACM SIGMOD, 2000.<br><br>Meier, Michael. "The backchase revisited." The VLDB Journal (2013):1-22. | Optional HW5: Logical Data Integration | Prof. Ambite |
| Feb 27 | Logical Data Integration 2 (Scalability) | Alon Halevy and Rachel Pottinger. A scalable algorithm for answering queries using views. The VLDB Journal The International Journal on Very Large Data Bases, 2001. http://www.vldb.org/conf/2000/P484.pdf.<br><br>Scalable query rewriting: a graph-based approach, 2001. http://www.isi.edu/~ambite/konstantinidis2011-sigmod.pdf<br><br>Student Presentations<br><br>M. König, M. Leclère, M.-L. Mugnier, M. Thomazo, Sound, Complete and Minimal UCQ-Rewritings for Existential Rules. Semantic Web Journal, online (2014), 6(5): 451-475 (2015). http://www.semantic-web-journal.net/content/sound-complete-and-minimal-ucq-rewriting-existential-rules-0<br><br>J.-F. Baget, Michel Leclère, M.-L. Mugnier, Swan Rocher, Clément | | Prof. Ambite |

| | | | | |
|---|---|---|---|---|
| | | Sipieter, Graal: A Toolkit for Query Answering with Existential Rules RuleML 2015: 328-344<br>http://graphik-team.github.io/graal/publications | | |
| Mar 4 | Project Proposals | | | Prof. Ambite<br>Dr. Fakhraei |
| Mar 6 | Data Warehousing: Logical Approaches | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 10. Morgan Kaufmann, 2012.<br>http://www.sciencedirect.com/science/book/9780124160446<br><br>Student Presentations:<br><br>R. Fagin, P.G. Kolaitis, R.J. Miller, and L. Popa. "Data exchange: semantics and query answering". Theoretical Computer Science, 2005.<br><br>R. Fagin, P.G. Kolaitis, and L. Popa. "Data exchange: getting to the core". ACM Transactions on Database Systems (TODS), 2005.<br><br>Giansalvatore Mecca, Paolo Papotti, and Salvatore Raunich. 2012. Core schema mappings: Scalable core computations in data exchange. Information Systems: 37 (7). November 2012.<br><br>Optional:<br><br>Deutsch, Alin, Alan Nash, and Jeff Remmel. "The chase revisited." Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2008.<br><br>Georg Gottlob, Alan Nash: Efficient core computation in data exchange. J. ACM 55(2) (2008) | | Prof. Ambite<br>Dr. Fakhraei |
| Mar 10-17 | NO CLASS | Spring Recess | | Prof. Ambite |
| Mar 18 | Data Warehousing: Kafka/Spark, And PolyStores | Student Presentations<br><br>Vijay Gadepally, Kyle O'Brien, Adam Dziedzic, Aaron J. Elmore, Jeremy Kepner, Samuel Madden, Tim Mattson, Jennie Rogers, Zuohao She, Michael Stonebraker. BigDAWG version 0.1. HPEC 2017: 1-7<br><br>Katherine Yu, Vijay Gadepally, Michael Stonebraker. Database engine integration and performance analysis of the BigDAWG polystore system. HPEC 2017: 1-7<br><br>Kyle O'Brien, Vijay Gadepally, Jennie Duggan, Adam Dziedzic, Aaron J. Elmore, Jeremy Kepner, Samuel Madden, Tim Mattson, Zuohao She, Michael Stonebraker. BigDAWG Polystore Release and Demonstration. CoRR abs/1701.05799 (2017)<br><br>Vijay Gadepally, Jennie Duggan, Aaron J. Elmore, Jeremy Kepner, Samuel Madden, Tim Mattson, Michael Stonebraker. The BigDAWG Architecture. CoRR abs/1602.08791 (2016) | | Prof. Ambite<br>Dimitri Stripelis |

| | | | | |
|---|---|---|---|---|
| | | Optional:<br><br>Dimitris Stripelis, Jose Luis Ambite, Yao-Yi Chiang, Sandrah P. Eckel, and Rima Habre. A Scalable Data Integration and Analysis Architecture for Sensor Data of Pediatric Asthma. IEEE International Conference on Data Engineering, San Diego CA, April 2017.<br>http://www-scf.usc.edu/~stripeli/documents/publications/ICDE2017.pdf<br><br>Dimitris Stripelis, Chrysovalantis Anastasiou, José Luis Ambite. Extending Apache Spark with a Mediation Layer. International Semantic Big Data Workshop, SIGMOD, Houston TX, April 2018<br>http://www-scf.usc.edu/~stripeli/documents/publications/SIGMODSBD2017.pdf | | |
| Mar 20 | Schema Mapping | AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration, chapter 5. Morgan Kaufmann, 2012.<br>http://www.sciencedirect.com/science/book/9780124160446<br><br>Reconciling schemas of disparate data sources: a machine-learning approach, 2001.<br>http://homes.cs.washington.edu/~pedrod/papers/sigmod01.pdf<br><br>Student Presentations:<br><br>Ronald Fagin, Laura M. Haas, Mauricio Hernández, Renée J. Miller, Lucian Popa, Yannis Velegrakis. Clio: Schema Mapping Creation and Data Exchange. Conceptual Modeling: Foundations and Applications. Lecture Notes in Computer Science Volume 5600, 2009.<br><br>A Collective, Probabilistic Approach to Schema Mapping. Kimmig, A.; Memory, A.; Miller, R. J.; and Getoor, L. In ICDE, pages 921-932, 2017. | | Prof. Ambite |
| Mar 25 | Semi-Automatic Source Modeling (Karma) | Craig Knoblock, Pedro Szekely, Jose Luis Ambite, Aman Goel, Shubham Gupta, Kristina Lerman, Parag Mallick, Maria Muslea and Mohsen Taheriyan. Semi-Automatically Mapping Structured Sources into the Semantic Web. Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012), Heraklion, Crete, Greece, 2012.<br>http://www.isi.edu/~ambite/eswc-karma.pdf<br><br>Student Presentations:<br><br>Ziawasch Abedjan, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker. DataXFormer: A robust transformation discovery system. ICDE 2016: 1134-1145<br><br>John Morcos, Ziawasch Abedjan, Ihab Francis Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker. DataXFormer: An Interactive Data Transformation Tool. SIGMOD Conference 2015: 883-888<br><br>Optional:<br>Craig A. Knoblock, Pedro Szekely. Exploiting Semantics for Big Data | | Prof. Ambite |

| | | | | |
|---|---|---|---|---|
| | | Integration. AI Magazine, 2015. http://usc-isi-i2.github.io/papers/knoblock15-aimagazine.pdf<br><br>Mohsen Taheriyan, Craig A. Knoblock, Pedro Szekely, José Luis Ambite. Learning the Semantics of Structured Data Sources. Journal of Web Semantics Special Issue on Knowledge Graphs, 2015. https://www.sciencedirect.com/science/article/pii/S1570826815001444 | | |
| Mar 27 | Automatic Source Modeling | Mark James Carman and Craig A. Knoblock. Learning semantic descriptions of web information sources. In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), January 2007. http://www.isi.edu/integration/papers/carman07-ijcai.pdf.<br><br>José Luis Ambite, Sirish Darbha, Aman Goel, Craig A. Knoblock, Kristina Lerman, Rahul Parundekar, and Thomas Russ. Automatically constructing semantic web services from online sources. In Proceedings of the 8th International Semantic Web Conference (ISWC 2009), 2009. http://www.isi.edu/integration/papers/ambite09-iswc.pdf | | Prof. Ambite |
| Apr 1 | RDF | Frank Manola and Eric Miller. RDF primer. Technical report, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.<br><br>Tim Berners-Lee. Why rdf model is different from the xml model. Technical report, W3C, 1998. http://www.w3.org/DesignIssues/RDF-XML.html. | | Prof. Ambite |
| Apr 3 | RDF Schema / Inference | RDF Schema 1.1: W3C Recommendation 25 February 2014. https://www.w3.org/TR/rdf-schema/ | | Prof. Ambite |
| Apr 8 | SPARQL Query Language | SPARQL 1.1 Query Language: W3C Recommendation 21 March 2013. http://www.w3.org/TR/sparql11-query/ | | Prof. Ambite |
| Apr 10 | OWL2 and Ontology-based Data Integration | Krtzsch Markus, Simancik Frantisek, and Horrocks Ian. A description logic primer. 2012. http://arxiv.org/pdf/1201.4089.pdf.<br><br>Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. DL-lite: tractable description logics for ontologies. In Proc. of the 20th National Conference on Artificial Intelligence, 2005. http://www.aaai.org/Papers/AAAI/2005/AAAI05-094.pdf.<br><br>Hector Perez-Urbina, Ian Horrocks, and Boris Motik. Efficient query answering for OWL 2. In International Semantic Web Conference, 2009. Efficient Query Answering for OWL 2. https://www.cs.ox.ac.uk/boris.motik/pubs/puhm09query-OWL2.pdf<br><br>Student presentations:<br><br>Ontology-based data access: A survey<br>Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, and Michael Zakharyaschev. In Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI 2018), pages 5511--5519. Int. Joint Conf. on Artificial Intelligence Organization, 2018. | Optional HW6: RDF/S SPARQL OWL | Prof. Ambite |

| | | | | |
|---|---|---|---|---|
| | | Diego Calvanese and Enrico Franconi. First-order ontology mediated database querying via query reformulation. In Sergio Flesca, Sergio Greco, Elio Masciari, and Domenico Saccà, editors, A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, volume 31 of Studies in Big Data, pages 169--185. Springer, 2018.<br>http://www.inf.unibz.it/~calvanese/papers-html/SEBD25-2018.html | | |
| Apr 15 | Knowledge Graphs | Singhal, Amit. "Introducing the Knowledge Graph: Things, Not Strings". Google Official Blog. May 16, 2012.<br>https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html<br><br>Introducing the Knowledge Graph<br>https://www.youtube.com/watch?v=mmQl6VGvX-c<br><br>Szekely, P., Knoblock, C. A., Slepicka, J., Philpot, A., Singh, A., Yin, C., ... & Ferreira, L. (2015). Building and Using a Knowledge Graph to Combat Human Trafficking. In *The Semantic Web-ISWC 2015* (pp. 205-221). Springer International Publishing.<br>http://iswc2015.semanticweb.org/sites/iswc2015.semanticweb.org/files/93670175.pdf<br><br>Student Presentations:<br><br>Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, Michael Stonebraker:<br>Aurum: A Data Discovery System. ICDE 2018: 1001-1012<br><br>Nickel, Maximilian, et al. "A review of relational machine learning for knowledge graphs." Proceedings of the IEEE 104.1 (2016): 11-33.<br>https://arxiv.org/pdf/1503.00759.pdf<br><br>Wang, Quan, et al. "Knowledge graph embedding: A survey of approaches and applications." IEEE Transactions on Knowledge and Data Engineering 29.12 (2017): 2724-2743.<br>https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8047276 | | Guest Speaker |
| Apr 17 | Network Extraction | Namata, Galileo Mark, Ben London, and Lise Getoor. "Collective graph identification." ACM Transactions on Knowledge Discovery from Data (TKDD) 10.3 (2016): 25.<br>https://linqs.soe.ucsc.edu/sites/default/files/papers/namata-tkdd.pdf<br><br>Tang, Jie, et al. "Arnetminer: extraction and mining of academic social networks." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.<br>http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.3839&rep=rep1&type=pdf | | Dr. Fakhraei |
| Apr 22 | Geospatial Data Integration | Fonseca, F. T., Egenhofer, M. J., Agouris, P., & Câmara, G. (2002). Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS, 6(3), 231–257.<br><br>Zhang, Y., Chiang, Y.-Y., Szekely, P., & Knoblock, C. A. (2013). A Semantic Approach to Retrieving, Linking, and Integrating Heterogeneous Geospatial Data. In Joint Proceedings of the | | Prof. Chiang |

| | | | | |
|---|---|---|---|---|
| | | Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities (pp. 31–37). New York, NY, USA: ACM.<br><br>Optional:<br><br>Güting, R. H. (1994). An Introduction to Spatial Database Systems. The International Journal on Very Large Data Bases, 3(4), 357–399.<br><br>Church, R. L. (2002/5). Geographical information systems and location science. Computers & Operations Research, 29(6), 541–562.<br><br>Chiang, Y.-Y., Leyk, S., & Knoblock, C. A. (2014). A Survey of Digital Map Processing Techniques. ACM Computing Surveys (CSUR), 47(1), 1. | | |
| Apr 22 | Project Presentations | | | Prof. Ambite Dr. Fakhraei |
| Apr 24 | Project Presentations | | | Prof. Ambite Dr. Fakhraei |

## Statement on Academic Conduct and Support Systems

### Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

### Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information http://emergency.usc.edu* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology