

The Joys of Working with Data

David C. Anastasiu

Department of Computer Science and Engineering

Santa Clara University

Research summary

Data Analysis Methods and Applications

Information
Retrieval

Machine Learning
Data Mining

High
Performance
Computing

Improving Web search utility

[WWWJ'13, IEEE IC'2013, ACM CIKM'09, DMIN'09]

Clustering and pattern mining

[IEEE BDS'19, IEEE MC'19, StatsRef'17, SIP'14, CRC'13, IEEE CIKM'11, IEEE SIGIR'11, COLING'10]

Traffic analytics from video

[IEEE CVPR'22, IEEE CVPR'21, BDAT'20, IEEE CVPR'20, IEEE CVPR'19, IEEE SOSE'19, IEEE MC'19, IEEE CVPR'18, IEEE SmartWorld'17]

Applications of machine learning and data mining

[GHTC'22, ECTEL'22, Microbio'22, iDSC'19, IEEE CIKM'18, GHC'18, IEEE SCI'17, IEEE ICDE'15]

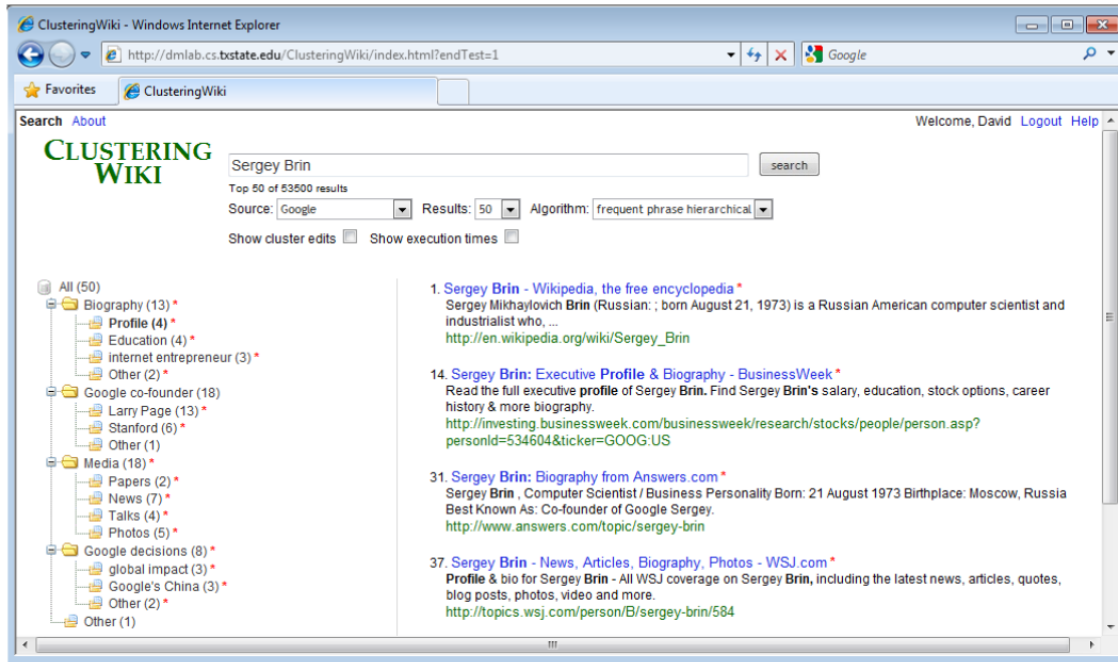
Fast nearest neighbors computation

[JPDC'17, JDSA'17, iDSC'17, IEEE DSAA'16, IA3'16, ACM CIKM'15, IA3'15, IEEE ICDE'14]

Improve Quality and Utility of Search Results

[CIKM'11] A Framework for Personalized and Collaborative Clustering of Search Results

- Developed a “label-first” hierarchical clustering technique.
- Path-based collaborative editing of cluster labels and assignments.
- Method incorporated in a document processing pipeline at LLNL.



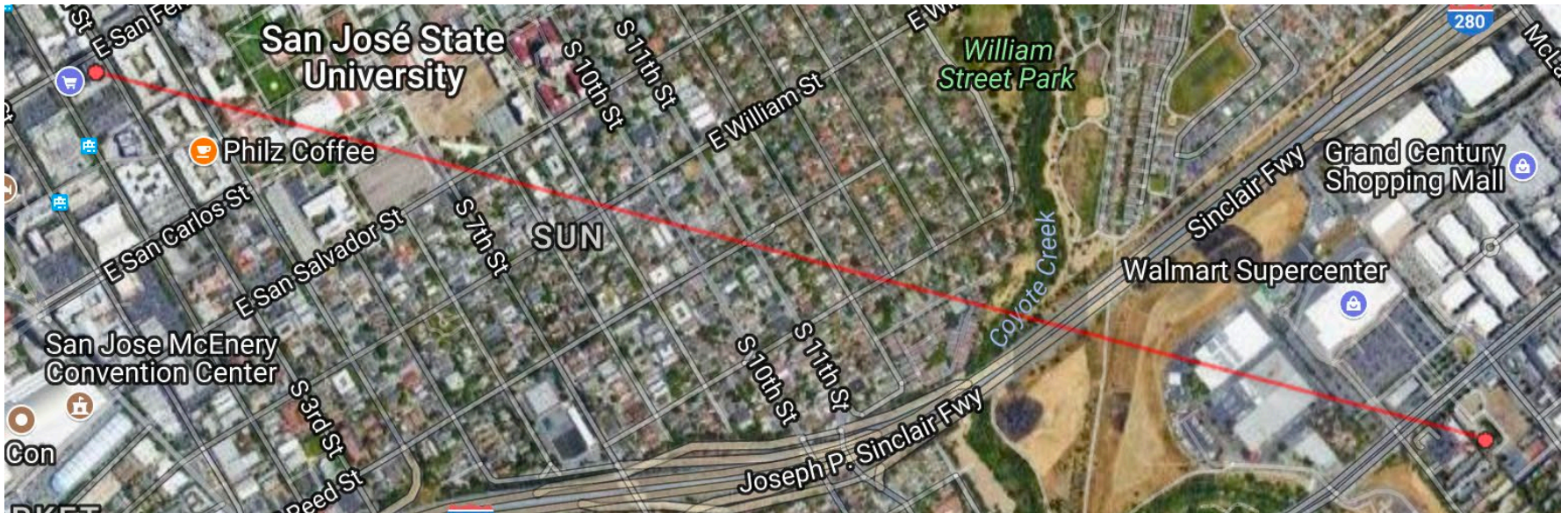
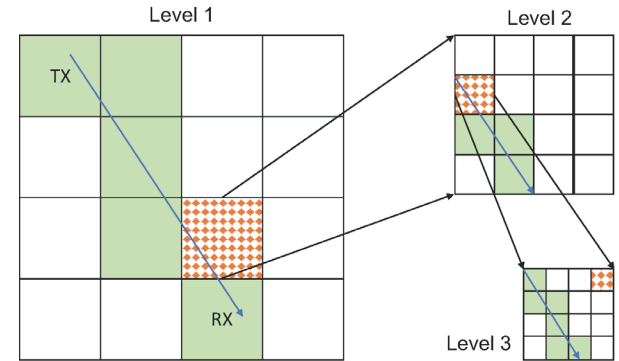
Lawrence Livermore National Laboratory

Smart City

[ACM CIKM'18] Data Structure for Efficient Line of Sight Queries

[IEEE SCI'18] Optimal Constrained Wireless Emergency Network Antenna Placement

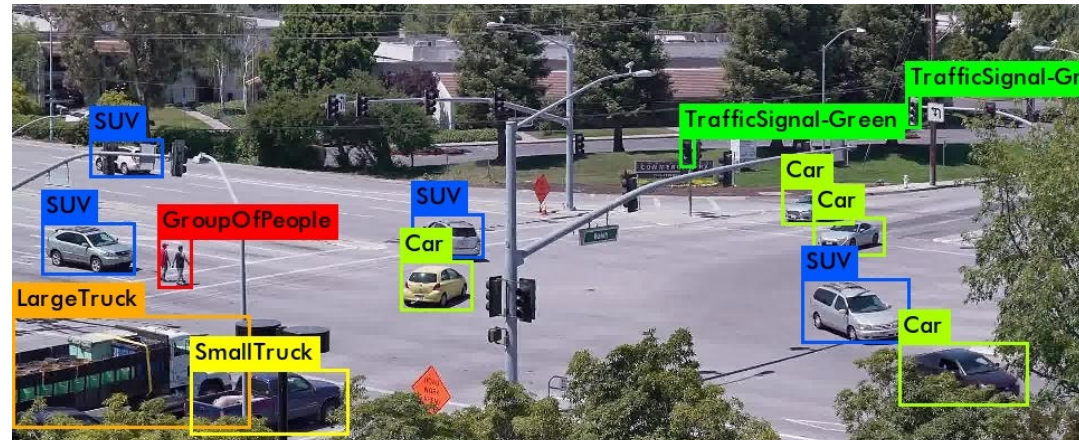
- Identify optimum locations of emergency wireless point-to-point wide area network nodes in a city.
- Multiple constraints, including:
 - Antenna physical limitations
 - Line of sight
 - Node priority
 - Minimum degree



[IEEE CVPR'19] CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification
[IEEE SOSE'19, IEEE SOSE'19, IEEE MC'19, IEEE CVPRW'18, IEEE SmartWorld'17]

w/ Milind Naphade, CTO of AI Cities, NVIDIA

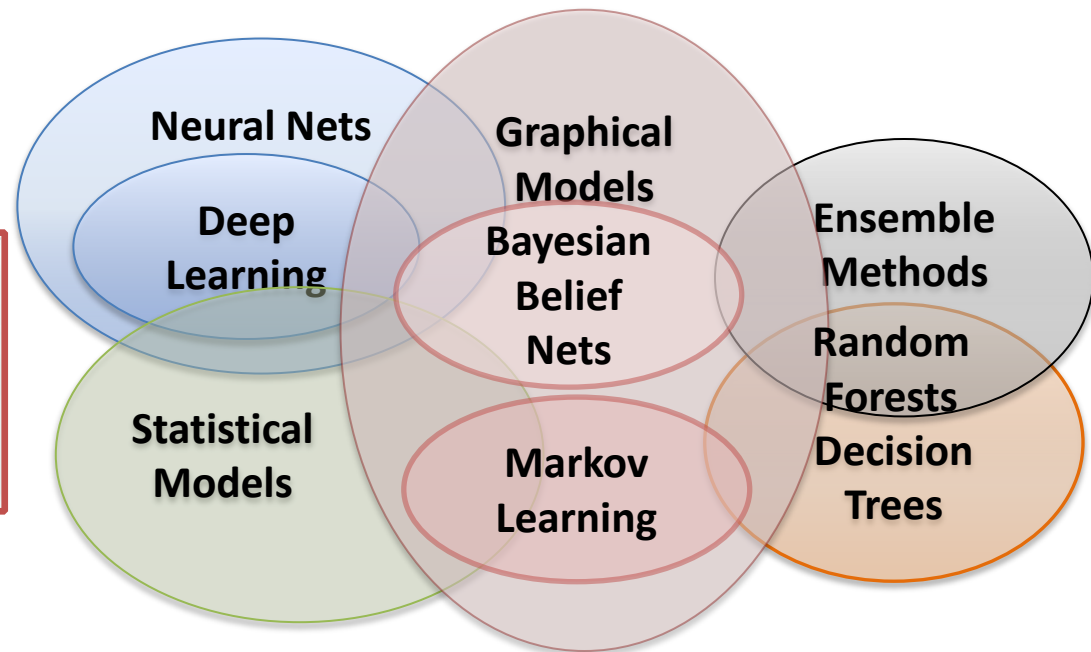
- Organizing member and Evaluation Chair for the AI City Challenge.
- Address challenges in traffic analysis from video, including:
 - Multi-camera vehicle tracking and multi-movement counting
 - Speed estimation from video
 - Anomaly detection



[Intel Labs Grant] DARPA HIVE project sub-contract.

Hierarchical Identify Verify Exploit (HIVE)

- Workload analysis for a new parallel processor
 - New hardware architecture optimized for sparse graphical models and decision trees
 - Hardware/software iterative co-design



Medical Analytics

[iDSC'19] A Data-Driven Approach for Detecting Autism Spectrum Disorders



Autism Prediction

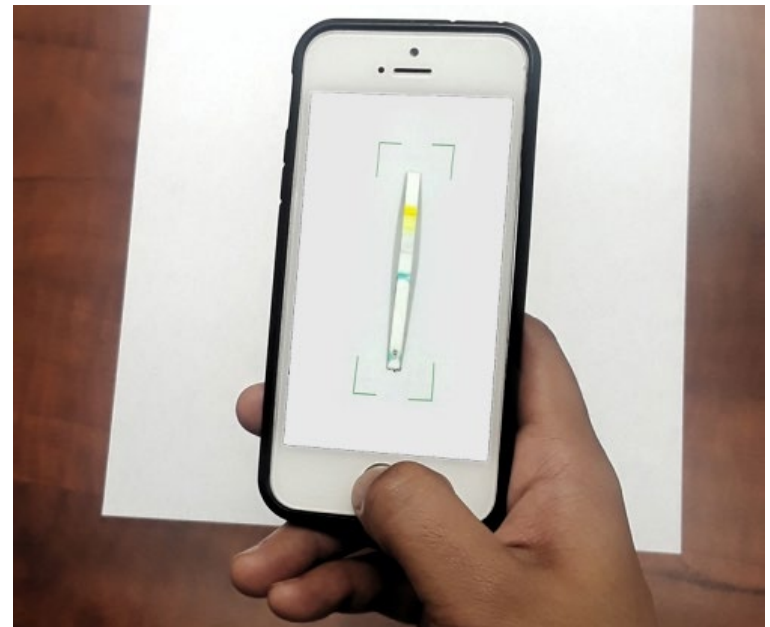
w/ Megan C. Chang, SJSU

- Derive if subject is autistic using sensor data
 - EKG and skin-conductance data collected during a sensory trial protocol.
 - Multivariate time series analysis with very long series (~4M points per subject).

Kidney Health Screening

w/ Alessandro Bellofiore, SJSU

- Decide severity of kidney disease by taking a picture of a test strip
 - Deep Learning localization models used to help ensure capture quality.
 - Machine Learning regression models used to translate picture to amount of creatinine in the blood.
 - Standard formulas used to classify based on regression output.





Open Modification Spectral Library Search

[Grant NSF 1850557] CRII: III: RUI: Effective Protein Characterization via Fast Exact Open Modification Searching

w/ William Stafford Noble, Genome Sciences, UW

- Methods for characterizing the protein composition of biological samples
 - Mass spectrometers output relative abundance histograms (spectra)
 - Massive databases exist for protein-associated spectra (spectral libraries)
 - Task is to match unknown spectra against nearest neighbor in library

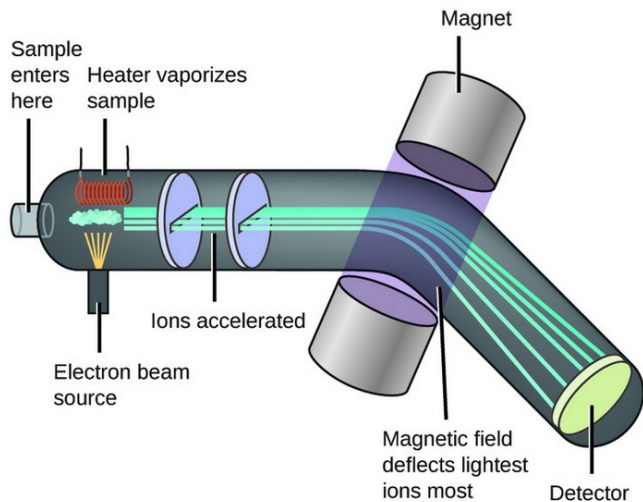
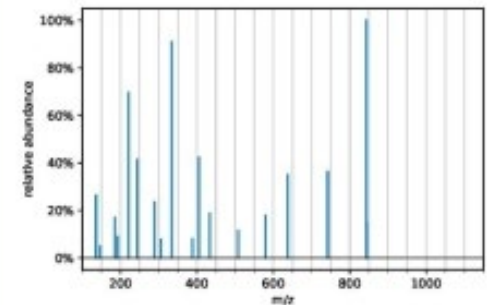
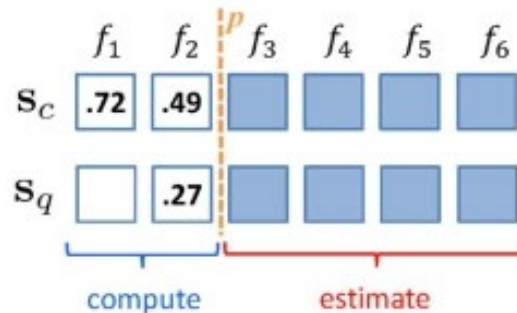


Image: <https://i.stack.imgur.com/iVYVY.png>

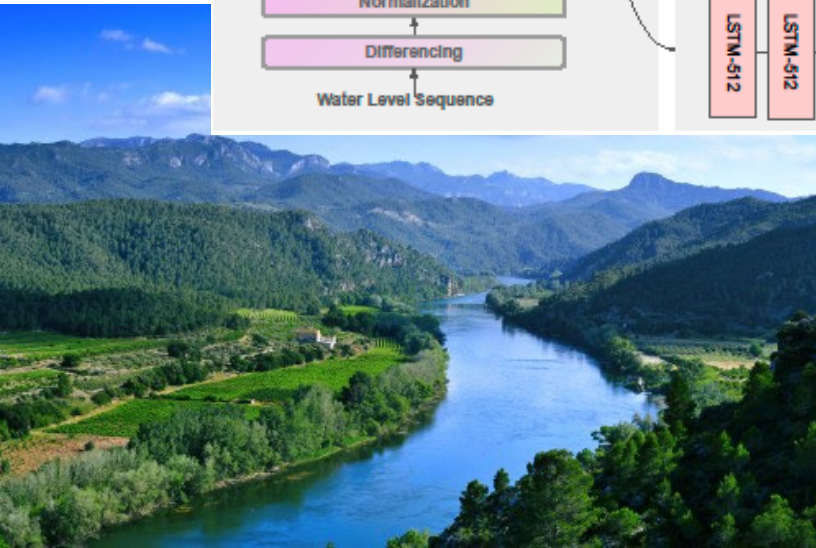
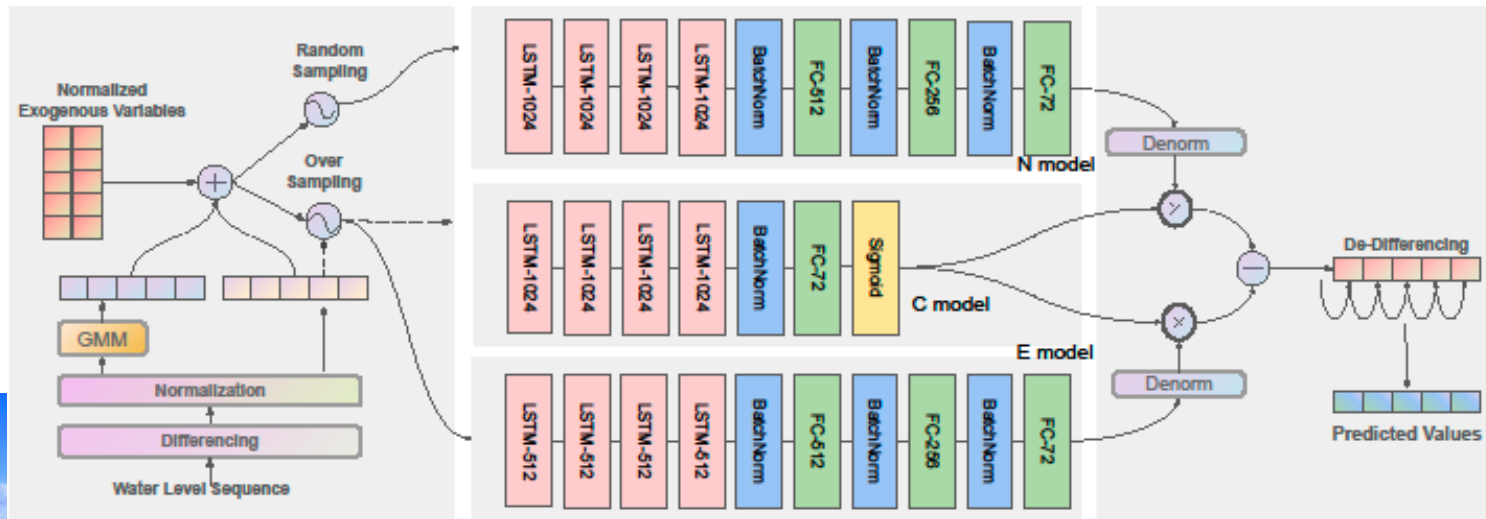
- Challenges
 - Imperfect ionization/spectrometry
 - Size of databases (10's to 100's or million)



AI Models for Hydrologic Flow Prediction

w/ Valley Water

- Methods for predicting flow volume of streams and rivers and water levels of reservoirs
 - Data have extreme events that are hard to distinguish from base levels
 - Proposed a probability-enhanced neural network model, NEC+, as a solution
 - Currently working on the next generation of NEC+



KDD Undergraduate Consortium

- Expand and enhance the participation of undergraduate students of diverse backgrounds in research pertaining to knowledge discovery from data.
- What to expect
 - Feedback on current research
 - Paper/poster presentations
 - Academic and industry mentors
 - Keynote talks about research careers and funding
 - Student panel on life as a researcher
 - Free conference attendance
- How to apply
 - QR code to the right or
 - <https://kdd.org/kdd2023/call-for-undergraduate-consortium/>



Questions?

References

- [AAAI'23] Yanhong Li, Jack Xi & David C. Anastasiu. An Extreme-Adaptive Time Series Prediction Model Based on Probability-Enhanced LSTM Neural Networks. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, AAAI Press, 2023.
- [GHTC'22] Alex Whelan, Soham Phadke & David C. Anastasiu. On-Device Prediction for Chronic Kidney Disease. In 2022 IEEE Global Humanitarian Technology Conference (GHTC) (GHTC 2022), 2022.
- [ECTEL'22] Arpita Vats, Gheorghi Guzun & David C. Anastasiu. CLP: A Platform for Competitive Learning. In Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption (EC-TEL 2022), pages 615-622, Springer International Publishing, 2022.
- [Microbio'22] Bipasa Bose, Taylor Downey, Anand K. Ramasubramanian & David C. Anastasiu. Identification of Distinct Characteristics of Antibiofilm Peptides and Prospection of Diverse Sources for Efficacious Sequences. *Frontiers in Microbiology*, 12, 2022.
- [CVPRW'22] Arpita Vats & David C. Anastasiu. Key Point-Based Driver Activity Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 3274-3281, 2022.
- [AIC'21] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E. Lopez, Anuj Sharma, Qi Feng, Vitaly Ablavsky & Stan Sclaroff. The 5th AI City Challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'21), pages 4263-4273, 2021.
- [AIC'20] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa & Pranamesh Chakraborty. The 4th AI City Challenge. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'20), 1:2665-2674, 2020.
- [BDAT'20] David C. Anastasiu, Jack Gaul, Maria Vazhaeparambil, Meha Gaba & Prajval Sharma. Efficient City-Wide Multi-Class Multi-Movement Vehicle Counting: A Survey. *Journal of Big Data Analytics in Transportation*, 2(3):235-250, 2020.
- [CVPR'19] Zheng Tang, Milind Naphade, Ming Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu & Jenq Neng Hwang. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In CVPR 2019: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), 2019.
- [AIC'19] Milind Naphade, Zheng Tang, Ming Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq Neng Hwang & Siwei Lyu. The 2019 AI City Challenge. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPRW'19), 1:452-460, 2019.
- [iDSC'19] Manika Kapoor & David C. Anastasiu. A Data-Driven Approach for Detecting Autism Spectrum Disorders. In Data Science -- Analytics and Applications (iDSC 2019), Springer Fachmedien Wiesbaden, 2019.

References

- [MC'19] Anupama Upadhayula, Avinash Ravilla, Ishwarya Varadarajan, Sowmya Viswanathan & David C. Anastasiu. Study Area Recommendation via Network Log Analytics. In The Seventh IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, IEEE, 2019.
- [SOSE'19] Shuai Hua & David C. Anastasiu. Effective Vehicle Tracking Algorithm for Smart Traffic Networks. In Thirteenth IEEE International Conference on Service-Oriented System Engineering (SOSE), IEEE, 2019.
- [CIKM '18] Swapnil Gaikwad, Melody Moh & David C. Anastasiu. Data Structure for Efficient Line of Sight Queries. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '18), ACM, 2018.
- [AIC'18] Milind Naphade, Ming Ching Chang, Anuj Sharma, David C. Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming Yu Liu, Rama Chellappa, Jenq Neng Hwang & Siwei Lyu. The 2018 NVIDIA AI City Challenge. In 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18), 1:53-60, 2018.
- [CVPRW'18] Shuai Hua, Manika Kapoor & David C. Anastasiu. Vehicle Tracking and Speed Estimation from Traffic Videos. In 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18), IEEE, 2018.
- [JDSA'17] David C. Anastasiu & George Karypis. Efficient identification of Tanimoto nearest neighbors; All Pairs Similarity Search Using the Extended Jaccard Coefficient. Springer International Journal of Data Science and Analytics, 4(3):153-172, 2017.
- [StatsRef'17] David C. Anastasiu & Andrea Tagarelli. Document Clustering. Wiley StatsRef: Statistics Reference Online, pages 1-11, American Cancer Society, 2017.
- [JPDC'17] David C. Anastasiu & George Karypis. Parallel cosine nearest neighbor graph construction. Elsevier Journal of Parallel and Distributed Computing, 2017.
- [SCI'17] Swapnil Gaikwad & David C. Anastasiu. Optimal Constrained Wireless Emergency Network Antenna Placement. In Proceedings of the IEEE Smart City Innovations 2017 Conference (IEEE SCI 2017), 2017.
- [iDSC'17] David C. Anastasiu. Cosine Approximate Nearest Neighbors. In Data Science -- Analytics and Applications (iDSC 2017), pages 45-50, Springer Fachmedien Wiesbaden, 2017.
- [SmartWorld'17] Niveditha Bhandary, Charles MacKay, Alex Richards, Ji Tong & David C. Anastasiu. Robust Classification of City Roadway Objects for Traffic Related Applications. In 2017 IEEE Smart World NVIDIA AI City Challenge (SmartWorld'17), IEEE, 2017.
- [DSAA'16] David C. Anastasiu & George Karypis. Efficient Identification of Tanimoto Nearest Neighbors. In 2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pages 156-165, 2016.

References

- [IA3'16] David C. Anastasiu & George Karypis. Fast Parallel Cosine K-Nearest Neighbor Graph Construction. In 2016 6th Workshop on Irregular Applications: Architecture and Algorithms (IA3) (IA3 2016), pages 50-53, 2016.
- [IA3'15] David C. Anastasiu and George Karypis. Pl2ap: Fast parallel cosine similarity search. IA3 2015. In conjunction with SC'15, IA3 2015, 2015.
- [CIKM'15] David C. Anastasiu and George Karypis. L2knng: Fast exact k-nearest neighbor graph construction with l2-norm pruning. CIKM '15, pages 791-800, New York, NY, USA, 2015. ACM.
- [ICDE'15] David C. Anastasiu, Al M. Rashid, Andrea Tagarelli, and George Karypis. Understanding computer usage evolution. ICDE 2015, pages 1549-1560, 2015.
- [ICDE'14] David C. Anastasiu and George Karypis. L2ap: Fast cosine similarity search with prefix l-2 norm bounds. ICDE 2014, pages 784-795, 2014.
- [SI'14] David C. Anastasiu, Jeremy Iverson, Shaden Smith, and George Karypis. Big data frequent pattern mining. In Frequent Pattern Mining, pages 225-260. Springer International Publishing, Switzerland, 2014.
- [CRC'13] David C. Anastasiu, Andrea Tagarelli, and George Karypis. Document clustering: The next frontier. In Data Clustering: Algorithms and Applications, pages 305-338. CRC Press, Boca Raton, FL, USA, 2013.
- [WWW'13] David C. Anastasiu, Byron J. Gao, Xing Jiang, and George Karypis. A novel two-box search paradigm for query disambiguation. World Wide Web, 16(1):1-29, 2013.
- [IC'13] Byron J. Gao, David Buttler, David C. Anastasiu, Shuaiqiang Wang, Peng Zhang, and Joey Jan. User-centric organization of search results. IEEE Internet Computing, 17(3):52-59, May 2013.
- [CIKM'11] David C. Anastasiu, Byron J. Gao, and David Buttler. A framework for personalized and collaborative clustering of search results. CIKM '11, pages 573-582, New York, NY, USA, 2011. ACM.
- [SIGIR'11] David C. Anastasiu, Byron J. Gao, and David Buttler. Clusteringwiki: personalized and collaborative clustering of search results. SIGIR 2011, pages 1263-1264, 2011.
- [COLING'10] Byron J. Gao, David C. Anastasiu, and Xing Jiang. Utilizing user-input contextual terms for query disambiguation. COLING '10, pages 329-337, Stroudsburg, PA, USA, 2010.
- [CIKM'09] Byron J. Gao, Mingji Xia, Walter Cai, and David C. Anastasiu. The gardener's problem for web information monitoring. CIKM '09, pages 1525-1528, New York, NY, USA, 2009. ACM.
- [DMIN'09] Walter Cai, David C. Anastasiu, Mingji Xia, and Byron J. Gao. Olap for multicriteria maintenance scheduling. DMIN '09, pages 35-41. CSREA Press, 2009.

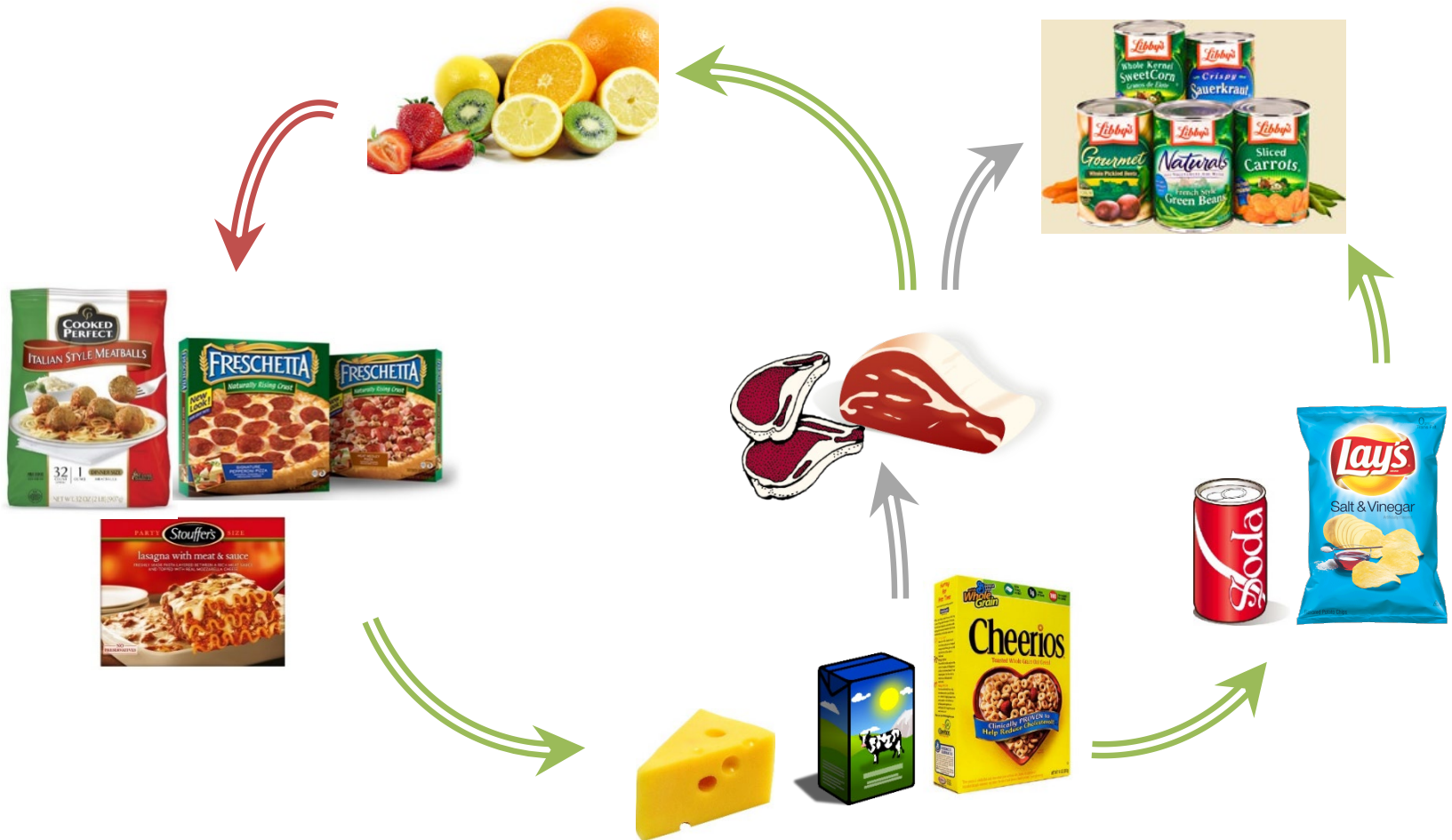
Other Projects



Purchasing Behavior Evolution

[ICDE'15] Understanding Computer Usage Evolution

- Developed a method to understand how people use their computers.
- In retail, we used it to understand dynamics of purchasing behavior.



[Flex Grant] Industry collaboration.

Detect fraud and/or savings opportunities in expense reports

- Receipt localization and classification (ResNet, YOLO-like models)
- Object character recognition (CNN + Bi-directional LSTM)
- Knowledge extraction (NER, heuristics)
- Report-receipt matching



Hard problem, complicated by the multinational and multilingual aspect of the Flex business

Near Duplicate Detection

[Elsevier JPDC'17] Parallel cosine nearest neighbor graph construction
[JDSA'17, iDSC'17, IEEE DSAA'16, IA3'16, ACM CIKM'15, IA3'15, IEEE ICDE'14]

- Developed efficient methods for nearest neighbor search and graph construction.
 - Can identify near-duplicates in large collections of millions of objects in mere seconds.

SpaceX rocket fails to land on barge

Company never expected to nail this landing, says SpaceX chief Elon Musk

The Associated Press | Posted: Mar 04, 2016 3:00 PM ET | Last Updated: Mar 04, 2016 8:46 PM ET



SpaceX has already succeeded in landing a Falcon rocket at an on-shore site near the Cape Canaveral pad where it launched, but it has failed in previous attempts to guide rockets back to ocean platforms. (SpaceX)

Related Stories

- SpaceX rocket launches satellite but botches ocean landing
- Why competition is good for the space race: Bob McDonald
- SpaceX rocket explosion debris likely found by UK Coast Guard

SpaceX has another launch under its belt, but not another rocket landing.

The leftover first-stage booster hit the floating platform hard Friday, said SpaceX chief Elon Musk. The company never expected to nail this landing, he said, because of the faster speed of the booster that was required to deliver the satellite to an extra-high orbit.

▪ SpaceX pushes satellite launch, rocket landing to Friday

SpaceX scored a rocket landing on the ground at Cape Canaveral in December, but has yet to nail a trickier barge landing at sea.

The good news, though, is that the unmanned Falcon 9 rocket successfully hoisted the broadcasting satellite for Luxembourg-based company SES.

It was the fifth launch attempt over the past 1½ weeks; Sunday's try ended with an engine shutdown a split second before liftoff. Friday's sunset launch provided a stunning treat along the coast.

SPACEX LAUNCHES SATELLITE, BUT FAILS TO LAND ROCKET ON BARGE



AP

Space-X's Falcon 9 rocket with the Jason-3 satellite aboard, stands ready for flight at Vandenberg Air Force Base, Calif. on Saturday, Jan. 16, 2016. (Matt Hartman)

[f Share](#) [G+](#) [T](#) Tweet

AP

Saturday, March 05, 2016 02:24PM

CAPE CANAVERAL, Fla. -- SpaceX has another launch under its belt, but not another rocket landing.

The leftover first-stage booster hit the floating platform hard Friday, said SpaceX chief Elon Musk. The company never expected to nail this landing, he said, because of the faster speed of the booster that was required to deliver the satellite to an extra-high orbit.