

Hubert Jin

Summary

A well-trained software engineer and applied statistician with 20+ years of experience, I am looking for a position in the fields of data science, data mining, machine learning, statistical analysis, natural language processing (NLP) and artificial intelligence (AI).

I have been dealing with data for most of my professional career, from creating and organizing data corpus to using data in statistical analysis, machine learning, applying data insights, and building advanced intelligent systems, as well as coaching people on how to analyze and interpret data. In addition, I have project management and business development experience for software outsourcing.

Education

| | | |
|---|--|------|
| M.S., Computer Science (Passed PhD Qualifying Exams), | University of Pennsylvania, Philadelphia | 2008 |
| M.S., Statistics (Passed PhD Qualifying Exams), | University of Washington, Seattle | 1995 |
| B.S., Computer Science, | Zhejiang University, China | 1986 |

Experience

2018-present Tabula Rasa Healthcare Moorestown, NJ

Data Scientist, Data Analytics Group

- Perform statistical analysis and data mining on real world data (RWD), including electronic healthcare claims, ICD10 codes, demographics, taxonomy, medication risk factors, HCC risk factors, SNOMED, prescription RxNorm, and physician NPI.
- Provide analytic insights and real world evidence (RWE) to support the core business operation of enhanced medication therapy management, as well as the Program of All-inclusive Care for the Elderly (PACE).
- Provide health economics analysis on Return-Of-Investment (ROI) of various pending business operations and potential contracts.

2014-2018 Princeton University Princeton, NJ

Senior Research Computing Specialist, Department of Politics (2014-2018)

Lecturer, Woodrow Wilson School (2016-2017)

- Administrated the daily operation of the Q-APS consulting service and personally provided consultation on statistical analysis, data science and high performance computing (HPC) to the students and faculty in social sciences across the campus. Developed R-Markdown to automatically create annual report with statistical analysis from the log data.
- Research and development (R&D), including education software development, web data scraping, statistical analysis, data visualization, big data, and high performance distributed/parallel computing.
- Managed a multi-TB data on the US voters and their demographics, including automatic data downloading, data ETL, database, access application and agreement process, and data coordination across the campus.
- Data collection for legislative bills from the 50 states in US, including data scraping and content extraction.
- Co-taught two workshops every year on Python web data scraping and advanced R statistical programming. Precepted the quantitative research seminar at the Woodrow Wilson School.

2013-2015 The CKCEST Project Chinese Academy of Engineering (CAE)

Technical Director of Data, Annotation and Analytics for Application System Platform (2013-2014)

Consultant, Annotation and Analytics for Application System Platform (2015)

- Co-authored the system platform proposal (4-million USD) that won a portion of the national big data project grant from the Chinese government (total of 16-million USD each year, and to be renewed for 9 years).
- Co-Led the initial design of the business and operation support system, from requirements analysis to feature design and solutions, including the use of big data technologies, i.e. Hadoop, MapReduce, cloud computing. Coordinated data integration and aggregation, both structured & unstructured from multiple sources.

- Developed medical image applications to estimate the 3D center surface of human left ventricle and the regional ventricular wall thickness between the surfaces representing the endocardium and epicardium. This innovative work was awarded a US patent.
- Implemented EM motivated algorithms for surface reconstruction from misregistered 3D scanner data. Developed programs for 3D rendering, curve and surface design, shape modeling and animation.

Consulting Provided

2012-2015 Client-Not-Be-Disclosed

- Provided probability based modeling to dynamically detect statistically significant ROI opportunities.
- Provided statistical analysis on risk management, as well as optimization on strategy.

Patents

- Co-inventor of the U.S. patent in information retrieval (No. 7062485): “Method and Apparatus for Score Normalization for Information Retrieval Applications”, Date of Patent: Jun 13, 2006.
- Co-inventor of the U.S. patent in 3D medical imaging (No. 5601084): “Determining cardiac wall thickness and motion by imaging and three-dimensional modeling”, Date of Patent: Feb 11, 1997

Competitions

- #1 in topic tracking and #2 in detection, 1999 DARPA TDT Evaluation, US. [NLP, Text Analytics]
- #1 in topic tracking and #2 in detection, 1998 DARPA TDT Evaluation, US. [NLP, Text Analytics]
- #3 in speech recognition for TV and broadcast news, 1997 DARPA Hub4 Evaluation, US. [NLP]
- Third place award, 1982 National Mathematics Competition, China.

Programming Skills

- Languages: C, C++, VC++, UNIX Shell Script, Perl, Python, R, S-plus, RStudio, SparkR, FORTRAN, PASCAL, SQL, QT, MATLAB, JavaScript, OpenGL, MPI, OpenMP, XML, Github, Gitlab, NLTK, Mallet, Numerical Recipes, Map/Reduce, and D3.
- Platforms: UNIX (including Solaris and etc.), Linux, Windows, High Performance Computing Clusters, AWS, Hadoop, Spark, and Shiny.

Publications

- Authored or co-authored 24 papers in computer graphics, image processing, speech recognition, information retrieval and extraction, text mining, natural language processing, and etc.
- Authored or co-authored 14 publications of data catalogues and corpora (ISBN registered).

Citizenship: U.S. Citizen.

Detailed List of Publications:

Papers

- 1) Hugues Hoppe, Tony DeRose, Tom Duchamp, Mark Halstead, **Hubert Jin**, John McDonald, Jean Schweitzer, Werner Stuetzle: Piecewise smooth surface reconstruction. Proceedings of the 21st annual conference on Computer graphics (SIGGRAPH 1994 Proceedings), Page 295, Orlando, FL, July 24 - 29, 1994.
- 2) **Hubert Jin**, Tom Duchamp, Hugues Hoppe, John McDonald, Kari Pulli, Werner Stuetzle: Surface reconstruction from misregistered data. Proceedings of SPIE, vol. 2573, pp. 324-8, 1995.
- 3) Edward Bolson, Florence Sheehan, Malcolm Legget, **Hubert Jin**, John McDonald, Paul Sampson, Ron Martin, Gerard Bashein, Catherine Otto: Applying the CenterSurface Model to 3-D Reconstruction of Left Ventricle for Regional Function Analysis. IEEE Computer in Cardiology, Vienna, Austria, 1995.
- 4) Francis Kubala, Tasos Anastasakos, **Hubert Jin**, Long Nguyen, Richard Schwartz: Transcribing Radio News. The Fourth International Conference on Spoken Language Processing, Wyndham Franklin Plaza Hotel, Philadelphia, PA, October 3-6, 1996.
- 5) Francis Kubala, **Hubert Jin**, Spyros Matsoukas, Long Nguyen, Rich Schwartz, John Makhoul: The 1996 BBN Byblos Hub-4 Transcription System. Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, Virginia, February 2-5, 1997.
- 6) **Hubert Jin**, Francis Kubala, Rich Schwartz: Automatic Speaker Clustering. Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, Virginia, February 2-5, 1997.
- 7) Richard Schwartz, **Hubert Jin**, Francis Kubala, Spyros Matsoukas: Modeling Those F-Conditions - Or Not. Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, Virginia, February 2-5, 1997.
- 8) Spyros Matsoukas, Rich Schwartz, **Hubert Jin**, Long Nguyen: Practical Implementations of Speaker-Adaptive Training. Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, Virginia, February 2-5, 1997.
- 9) Francis Kubala, **Hubert Jin**, Long Nguyen, Richard Schwartz, Spyros Matsoukas: Broadcast News Transcription. 1997 IEEE International Conference on ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Munich, Germany, April 21-24, 1997.
- 10) Francis Kubala, **Hubert Jin**, Spyros Matsoukas, Long Nguyen, Richard Schwartz, John Makhoul: ADVANCES IN TRANSCRIPTION OF BROADCAST NEWS. 5th European Conference on Speech Communication and Technology, Rhodes, Greece, Sept. 22-25, 1997.
- 11) Francis Kubala, Jason Davenport, **Hubert Jin**, Daben Liu, Tim Leek, Spyros Matsoukas, David Miller, Long Nguyen, Fred Richardson, Richard Schwartz, John Makhoul: The 1997 BBN Byblos System Applied to Broadcast News Transcription. Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne Conference Resort, Lansdowne, Virginia, February 8-11, 1998.
- 12) **Hubert Jin**, Spyros Matsoukas, Richard Schwartz, Francis Kubala: Fast Robust Inverse Transform SAT and Multi-stage Adaptation. Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne Conference Resort, Lansdowne, Virginia, February 8-11, 1998.
- 13) **Hubert Jin**, Spyros Matsoukas, Richard Schwartz, Francis Kubala: Fast Robust Inverse Transform Speaker Adapted Training Using Diagonal Transformations. 1998 IEEE International Conference on ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Seattle, WA, 1998.
- 14) Frederick G Walls, **Hubert Jin**, Sreenivasa Sista, Richard Schwartz: Probabilistic Models For Topic Detection And Tracking. 1999 IEEE International Conference on ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Phoenix, AZ, 1999.
- 15) **Hubert Jin**, Rich Schwartz, Sreenivasa Sista, Frederick Walls: Topic Tracking for Radio, TV Broadcast, and Newswire. 6th European Conference on Speech Communication and Technology, Budapest, Hungary, Sept. 5-9, 1999.
- 16) Frederick Walls, **Hubert Jin**, Sreenivasa Sista, Richard Schwartz: Topic Detection in Broadcast News. 6th European Conference on Speech Communication and Technology, Budapest, Hungary, Sept. 5-9, 1999.
- 17) **Hubert Jin**, Rich Schwartz, Sreenivasa Sista, Frederick Walls: Topic Tracking for Radio, TV Broadcast, and Newswire. DARPA Broadcast News Workshop, Herndon, Virginia, February 28-March 3, 1999.
- 18) Frederick Walls, **Hubert Jin**, Sreenivasa Sista, Richard Schwartz: Topic Detection in Broadcast News. DARPA Broadcast News Workshop, Herndon, Virginia, February 28-March 3, 1999.
- 19) Tim Leek, **Hubert Jin**, Sreenivasa Sista, Richard Schwartz: The BBN Crosslingual Topic Detection and Tracking System. 2000 Topic Detection and Tracking, Vienna, Virginia, February 28-March 1, 2000.
- 20) James Allan, Victor Lavrenko, **Hubert Jin**, D Malin: UMass and First Story Detection - including suggestions we can't do better. DARPA TDT Workshop at the Tysons Corner Sheraton Premiere, Vienna, Virginia, USA, February 28-March 1, 2000.

- 21) James Allan, Victor Lavrenko and **Hubert Jin**: First Story Detection in TDT is Hard. Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM), pp. 374-381, McLean, VA, Nov 6-11. 2000.
- 22) Mohamed Maamouri, David Graff, **Hubert Jin**: Christopher Cieri, and Tim Buckwalter: Dialectal Arabic Orthography-based Transcription and CTS Levantine Arabic Collection. Paper presented at the Parallel STT-NA Tracks Session of the EARS RT-04 Workshop, Palisades IBM Executive Center, New York, Nov. 10, 2004.
- 23) Jan Hajic, Otakar Smrz, Tim Buckwalter, and **Hubert Jin**: Feature-based tagger of approximations of functional Arabic morphology. Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, Spain, December 10, 2005.
- 24) **Hubert Jin** et al: Automatic Arabic POS tagging via Non-parametric Pseudo-likelihood estimation (in preparation).
- 25) David L. Bankes, **Hubert Jin**, Stephanie Finnel, Veronique Michaud, Calvin H. Knowlton, Jacques Turgeon and Alan Stein: Association of a Novel Medication Risk Score with Adverse Drug Events and Other Pertinent Outcomes Among Participants of the Programs of All-Inclusive Care for the Elderly, MDPI (Open Access), 2020.

Other ISBN Publications and Data Releases

- 1) Moussa Bamba, Mohamed Maamouri, **Hubert Jin**, Ann Bies, Xiaoyi M: Arabic Treebank: Part 1 - 10K-word English Translation. *ISBN: 1-58563-262-7*, Feb 25, 2003.
- 2) Mohamed Maamouri, Ann Bies, **Hubert Jin**, Tim Buckwalter: Arabic Treebank: Part 1 v 2.0. *ISBN: 1-58563-261-9*, Feb 03, 2003.
- 3) Mohamed Maamouri, Ann Bies, Tim Buckwalter, **Hubert Jin**: Arabic Treebank: Part 2 v 2.0. *ISBN: 1-58563-282-1*, Jan 30, 2004.
- 4) Mohamed Maamouri, Ann Bies, Tim Buckwalter, **Hubert Jin**: Arabic Treebank: Part 3 v 1.0. *ISBN: 1-58563-298-8*, May 21, 2004.
- 5) Mohamed Maamouri (Project head), Ann Bies, Tim Buckwalter, **Hubert Jin**: Arabic Treebank: Part 1 v 3.0 (POS with full vocal.+ syntactic analysis). *ISBN: 1-58563-330-5*, Feb 15, 2005.
- 6) Mohamed Maamouri, Tim Buckwalter, **Hubert Jin**: Arabic CTS Levantine Fisher Training Data Set 3, Transcripts. *ISBN: 1-58563-326-7*, Jan 15, 2005.
- 7) Mohamed Maamouri, Tim Buckwater, **Hubert Jin**: Arabic CTS Levantine Fisher Training Data Set 3, Speech. *ISBN: 1-58563-325-9*, Jan 15, 2005.
- 8) Mohamed Maamouri (project head), Ann Bies, Tim Buckwalter, **Hubert Jin**, and Wigdan Mekki: Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis). *ISBN: 1-58563-341-0*, Jun 15, 2005.
- 9) Mohamed Maamouri, Tim Buckwater, **Hubert Jin**: Levantine Arabic QT Training Data Set 4 (Speech + Transcripts). *ISBN: 1-58563-342-9*, Jun 15, 2005.
- 10) Mohamed Maamouri (Project head), Ann Bies, Tim Buckwalter, **Hubert Jin**, Wigdan Mekki: Arabic Treebank: Part 4 v 1.0 (MPG Annotation). *ISBN: 1-58563-343-7*, Oct 15, 2005.
- 11) Mohamed Maamouri, Tim Buckwater, Dave Graff, **Hubert Jin**: Levantine Arabic QT Training Data Set 5, Speech. *ISBN: 1-58563-377-1*, Feb 15, 2006.
- 12) Mohamed Maamouri, Tim Buckwater, Dave Graff, **Hubert Jin**: Levantine Arabic QT Training Data Set 5, Transcripts. *ISBN: 1-58563-378-X*, Feb 15, 2006
- 13) Mohamed Maamouri, Tim Buckwater, Dave Graff, **Hubert Jin**: Fisher Levantine Arabic Conversational Telephone Speech. *ISBN: 1-58563-438-7*, Mar 16, 2007.
- 14) Mohamed Maamouri, Tim Buckwater, Dave Graff, **Hubert Jin**: Fisher Levantine Arabic Conversational Telephone Speech, Transcripts. *ISBN: 1-58563-411-5*, Mar 16, 2007.