Goal:

Digital fraud is unavoidable product of the digit era.

This course will survey the digital fraud schemes in many industries. Students will learn how to set up a fraud risk management program. Students will have hands-on programming experiences in data analysis. The techniques include (not limited to) all machine learning methods, neural network, textual detection.

* This course aims to provide a general overview of the state-of-the-art methods for anomaly detection. We will cover techniques for unsupervised, semi-supervised and supervised approaches. We stress the importance of feature engineering as regard to the root cause of the detected anomalies. We also present the programming in real-world fraud detection applications such as XGBoost, H2O autoencoders, etc.

Goal: to train

Table of contents

1. Types of fraud. Introduction to data mining. Supervised vs. unsupervised learning. Industry standards. Benford’s Law.
2. Best practices for fraud risk management
   1. Fraud types
   2. Fraud prevention
3. Feature engineering
   1. Credit card fraud examples
   2. Healthcare fraud examples
4. Extracting features from text
   1. <https://www.researchgate.net/post/How_can_I_extract_features_from_text_data>
   2. <https://code.google.com/archive/p/word2vec/>
   3. <https://www.casact.org/education/specsem/f2008/handouts/ellingsworth.pdf>
5. Detecting fraud using supervised techniques
   1. Sampling methods to combat with extremely imbalanced data
   2. GBM & XGboost:
   3. <http://dk.archive.ubuntu.com/pub/pub/cran/web/packages/dismo/vignettes/brt.pdf>
   4. <https://www.kaggle.com/aniruddhachakraborty/lasso-gbm-xgboost-top-20-0-12039-using-r>
   5. <https://www.youtube.com/watch?v=sRktKszFmSk&list=PL6c2aKbLXM7lViGtsJr40MB-1qSB_Exzk>
   6. <http://xgboost.readthedocs.io/en/latest/model.html>
   7. <http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/>
   8. <https://jessesw.com/XG-Boost/>
   9. https://www.kaggle.com/nschneider/gbm-vs-xgboost-vs-lightgbm
   10. Deep learning: Tensorflow, Keras
   11. Autoencoder (H2O)
6. Detecting fraud using unsupervised techniques
   1. Z-Score, MAD,
   2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
   3. MeanShift
   4. Gaussian Mixture Model (GMM)
   5. One-class SVM
   6. Hierarchical
   7. Hidden Markov Model (HMM)
   8. Self-organizing map (SOM)
7. Detecting fraud using semi-supervised techniques
8. Detecting fraud using graph-based techniques
   1. Introduction ß
   2. NoSQL
   3. Neo4j: https://www.youtube.com/watch?v=ujimD6MP87I
   4. <https://www.youtube.com/watch?v=AeNufTq1W5I>
   5. Whiplash for cash: <http://www.after-car-accidents.com/whiplash-settlements.html>
   6. <https://linkurio.us/blog/stolen-credit-cards-and-fraud-detection-with-neo4j/>
   7. <https://neo4j.com/blog/financial-services-neo4j-fraud-detection/>
   8. First party fraud: <https://www.scmagazine.com/first-party-fraud/article/554398/>
   9. <http://www.infoglide.com/blog/first-party-fraud/>
   10. <http://www.infoglide.com/blog/first-party-fraud-assessing-the-damage/>
   11. <https://github.com/yudong-94/Realtime-Unsupervised-Fraud-Detection-Algorithm/blob/master/Project%202%20Report%20-%20Team%204.pdf>
   12. <https://www.youtube.com/watch?v=tbyf7_LmLTM>
   13. Using cypher procedures and functions: <https://github.com/neo4j-contrib/gists/blob/master/other/BankFraudDetection.adoc>
   14. Helpful: <https://www.youtube.com/watch?v=Eh_79goBRUk&t=644s>
   15. Basic query: <https://neo4j.com/docs/developer-manual/current/cypher/query-tuning/basic-query-tuning-example/>
   16. Basic query: <https://neo4j.com/graphgist/importing-csv-files-with-cypher>
   17. Interesting <https://www.quackit.com/neo4j/tutorial/neo4j_query_language_cypher.cfm>
9. Systems for real-time detection

1. Fraud detection and prevention

2. Understand fraud schemes:

· Banking fraud schemes

· Internal audit for fraudulent disbursements (mischaracterized expenses, overstated expenses, fictitious expenses, check tampering, forged maker, altered payees, etc.)

· Inventory: larceny and misuse

· Financial statement fraud

· Medical billing fraud, legal billing fraud, etc.

3. Setting up a fraud risk management program in the digital era

· Fraud awareness metrics & dashboard

· Fraud reporting

· Internal control review & policy

4. Data analysis techniques for fraud detection

· Detection methods for numerical data

· Detection methods for textual data

. Deep learning methods

10 use cases in Python: Machine learning methods for credit card fraud detection, insurance fraud, billing frauds, etc.

Machine learning for unsupervised fraud detection:

* Very interesting <http://www.diva-portal.org/smash/get/diva2:897808/FULLTEXT01.pdf>
* https://www.youtube.com/watch?v=AhFESJpiCa0
* GMM outlier detection scikit learn: <http://scikit-learn.org/stable/auto_examples/applications/plot_outlier_detection_housing.html>
* GMM outlier detection blog: <https://codefying.com/2016/09/03/gaussian-mixture-model-with-application-to-anomaly-detection/>
* GMM github: <https://github.com/LocalGroupAstrostatistics2015/MachineLearning/blob/master/indepth_GMM.ipynb>
* GMM on MNIST dataset: <http://www.datasciencecourse.org/anomaly_detection.pdf>
* GMM codes: <https://aqibsaeed.github.io/2016-07-17-anomaly-detection/>

Imbalanced data:

* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3946903/>
* <https://www.r-bloggers.com/dealing-with-unbalanced-data-in-machine-learning/>
* <https://www.kaggle.com/qianchao/smote-with-imbalance-data>

**Tools to make my life easier:**

* How to convert Jupyter notebook to Wordpress: <https://pythonandr.com/2016/07/18/sharing-ipython-jupyter-notebooks-via-wordpress/>
* Converting Jyputer notebook to Wordpress: <http://www.mianchen.com/wordpress-blogging-with-jupyter-notebook-in-five-simple-steps/>
* <https://www.datacamp.com/community/blog/jupyter-notebook-r>
* R markdown to ipynb: <https://github.com/aaren/notedown>
* Publisher: http://authors.packtpub.com/

**Machine learning courses and grading:**

* <http://www2.cs.uh.edu/~ceick/ML/ML.html>
* <http://www.utstat.utoronto.ca/~radford/sta414.S06/>
* <http://www.timvanerven.nl/teaching/statlearn2014/>
* <https://ece521.github.io/>
* <https://www.onlc.com/outline.asp?ccode=A20774>
* <https://www.cs.cmu.edu/~tom/10601_fall2012/exams/midterm_solutions.pdf>
* <https://classes.soe.ucsc.edu/cmps142/Winter10/handouts/NewSampleMLExam.pdf>
* Deep learning tutorial: <https://www.datacamp.com/community/tutorials/deep-learning-python>
* A practical introduction to deep learning with Caffe and Python: <http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/>
* Not useful Fraud detection on-line course: <http://www.deanza.edu/online-ed/syllabi/ACCT73_W17-Mello.pdf>
* <http://www.cs.cornell.edu/courses/cs4780/2015fa/styled/index.html>
* Not useful Fraud detection and data mining: <https://is.njit.edu/sites/is/files/lcms/academics/courses/Fall_2014_IS_687_Syllabus_v2.pdf>
* Machine learning/data mining books: <https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md>

**General fraud types and prevention:**

* Interesting: Fraud risk management: a guide to good practice: <https://www.cgma.org/Resources/Reports/DownloadableDocuments/fraudriskmanagement.pdf>
* Why fraud detection in banking: <https://www.romexsoft.com/blog/credit-card-fraud-detection-in-banking/>
* Suspicious id: <https://www.equifax.com/assets/IFS/Fraud/suspicious_id_wp.pdf>
  + Fraud prevention methods

### 

Supervised and unsupervised learning:

<https://thisdata.com/blog/unsupervised-machine-learning-with-one-class-support-vector-machines/>

* Unsupervised machine learning is machine learning without labelled data (where data hasn't been labelled beforehand to say what it is -- in our case, whether a network access is an attack or not). Most programmers are familiar, at least in some way, with [supervised ML](https://en.wikipedia.org/wiki/Supervised_learning). This is where a model is trained (learns) from labelled data and then uses that model to make predictions or give some other output. In unsupervised learning the model is trained without labels, and a trained model picks novel or anomalous observations from a dataset based on one or more measures of similarity to "normal" data.

Unsupervised machine learning can be useful in information security problems because a) we don't always have accurately labelled data, and b) we want to be able to identify novel (or anomalous) observations in data without having necessarily seen an example of that behaviour in the past.

**Anomaly detection literature:**

* Data mining anomaly detection: <https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap10_anomaly_detection.pdf>
  + What are anomalies or outliers? The very first definition of an outlier dates back to 1980, and is given by Douglas M. Hawkins [Hawkins, 1980]: Definition 1 (Hawkins’ Definition of Outlier, 1980) “An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.”
  + They are the data points that are far away from the remainder of the data. The challenge is how an outlier is defined and how many there are.
* Graph based anomaly detection: <http://www3.cs.stonybrook.edu/~leman/pubs/14-dami-graphanomalysurvey.pdf>
  + Examples include detecting network intrusion or network failure [Ding et al., 2012, Ide ́ and Kashima, 2004, Sun et al., 2008], credit card fraud [Bolton and Hand, 2001], calling card and telecommunications fraud [Cortes et al., 2002, Taniguchi et al., 1998], auto insurance fraud [Phua et al., 2004], health insurance claim er- rors [Kumar et al., 2010], accounting inefficiencies [McGlohon et al., 2009], email and Web spam [Castillo et al., 2007], opinion deception and reviews spam [Ott et al., 2012], auction fraud [Pandit et al., 2007], tax evasion [Abe et al., 2010, Wu et al., 2012], customer activity monitoring and user profiling [Fawcett and Provost, 1996, 1999], click fraud [Jansen, 2008, Kshetri, 2010], securities fraud [Neville et al., 2005], malicious cargo shipments [Das and Schneider, 2007, Eberle and Holder, 2007] malware/spyware detection [Invernizzi and Comparetti, 2012, Ma et al., 2009, Provos et al., 2007], false advertising [Lee et al., 2010], data-center monitoring [Li et al., 2011b], insider threat [Eberle and Holder, 2009], image/video surveillance [Damnjanovic et al., 2008, Krausz and Herpers, 2010], and many others.
  + Why graphs? Graph-based approaches can reveal inter-dependent nature of the data
  + The graph-based approach presents a plain graph database to find the nodes, edges or substructures in the graph that differ significantly from the majority of the reference objectives.
  + The challenges: (1) lack and noise of labels, (2) extreme rare events or class imbalance, (3) new novelty, the fraudsters realize the detecting system and find new ways,
  + The graph-based detection can cover (a) a static graph, (b) time-series or dynamic graphs
  + Graph-based fraud detection: <http://www3.cs.stonybrook.edu/~leman/icdm12/ICDM12-Tutorial%20-%20PartI.pdf>

Semi-supervised learning:

* <http://scikit-learn.org/stable/modules/label_propagation.html>
* <https://github.com/kimiyoung/planetoid>
* <https://github.com/tmadl/semisup-learn>
* Anomaly detection: <http://courses.washington.edu/css581/lecture_slides/18_anomaly_detection.pdf>
  + Anomaly and outliers are essentially the same thing. Historically the field of statistics tried to find and remove outliers to improve analyses. In some applications, outliers are even called the noise and tend to be dropped in the data cleaning step. But now many fields are interested in the outliers and anomalies. Those rare events may be the ones with the most negative impact.
  + Types of fraud include credit card fraud, insurance claim fraud, mobile/cell phone fraud, insider trading, failure in a complex industrial system, etc. The challenges to a failure in a complex industrial system are: (a) data is extremely large and noisy, (b) most of the applications exhibit temporal behavior, and (c) detected anomalies usually require immediate intervention.
  + The use of data labels in anomaly detection includes three types: The first type is *supervised*, where labels are available for both normal data and anomalies. This is similar to classification with high class imbalance. The second type is semi-supervised, where labels are available only for normal data. The third type is unsupervised, where there is no label at all with very rare events.
* Interesting: Anomaly detection video <https://www.oreilly.com/ideas/what-are-the-challenges-in-building-an-anomaly-detection-system-for-streaming-and-live-data>
* Survey of outlier detection: <http://eprints.whiterose.ac.uk/767/1/hodgevj4.pdf>
* Fraud detection github code: <https://github.com/kskk02/Fraud_Detector/blob/master/Fraud%20Detection.ipynb>
* Install Jupyter: <https://lectures.quantecon.org/py/getting_started.html>

**Payment fraud literature:**

* <http://raw.rutgers.edu/MiklosVasarhelyi/Resume%20Articles/MAJOR%20REFEREED%20ARTICLES/M60.JETA%20MAV%202012%20for%20resume.pdf>
* A Comparative Analysis of Decision Trees Vis-a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. Gepp, Wilson, Kumar, Bhattacharya. Journal of Data Science 10(2012), 537-561.
* <http://didawikinf.di.unipi.it/lib/exe/fetch.php/dm/ar-creditcard-fraudedetection.pdf>

**Credit card fraud literature:**

* Interesting: “Transaction aggregation as a strategy for credit card fraud detection”: <https://pdfs.semanticscholar.org/a2d2/b7bb3496c51acecdf3e3574278dfbf17174b.pdf>
* Feature engineering: “Data mining for credit card fraud: a comparative study”: <https://pdfs.semanticscholar.org/9d26/f0ba02ee5efe9b9c7bdcb5f528c8b8253cf7.pdf>
* Small value: “Learning lessons in credit card fraud detection from a practitioner perspective”: <http://www.ulb.ac.be/di/map/adalpozz/pdf/FraudDetectionPaper_8.pdf>
* Interesting: “Credit card fraud and detection techniques: a review”: <http://eprints.hud.ac.uk/id/eprint/19069/1/AbdouCredit.pdf>
* Interesting: “Feature engineering strategies for credit card fraud detection”: <http://albahnsen.com/files/Feature%20Engineering%20Strategies%20for%20Credit%20Card%20Fraud%20Detection_published.pdf>
* Interesting: Real time credit card fraud detection with Apache Spark and Event streaming: <https://mapr.com/blog/real-time-credit-card-fraud-detection-apache-spark-and-event-streaming/>
* Interesting: feature engineering: <http://www.cs.rochester.edu/~kshen/csc296-fall2013/lectures/Mizes_CardFraud.pdf>
* Interesting: Features: <http://www.lusispayments.com/uploads/4/4/8/2/44826195/a_comparison_of_machine_learning_techniques_for_credit_card_fraud_detection.pdf>

**Kaggle Credit card fraud detection:**

* <https://www.kaggle.com/dalpozz/creditcardfraud>
* <https://www.kaggle.com/qianchao/smote-with-imbalance-data>
* Predicting fraud with tensorflow: <https://www.kaggle.com/currie32/predicting-fraud-with-tensorflow>
* Predicting fraud with GBM: <https://www.kaggle.com/nschneider/gbm-vs-xgboost-vs-lightgbm>
* Approaching almost any machine learning problem: <http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/>
* Using XGboost: <https://github.com/dvdbisong/Credit-Card-Fraud-Detection-using-Extreme-Gradient-Boosting/blob/master/FraudDetection.ipynb>
* Using Autoencoders: <https://shiring.github.io/machine_learning/2017/05/01/fraud>
* Using Autoencoders: <https://medium.com/@curiousily/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd>
* Good paper list: <https://github.com/dalpozz/AMLFD>
* Fraud analytics: <http://www.dataminingapps.com/wp-content/uploads/2015/08/68614_excerpt-1.pdf>
  + Have purchased
* Using PCA and autoencoder: <https://github.com/yudong-94/DSO-562-Project-2-Realtime-Unsupervised-Fraud-Algorithm/blob/master/Project%202%20Report%20-%20Team%204.pdf>
  + Product data: Dataset available
  + Unsupervised data
* Credit fraud prevention with spark and graph analysis: <https://www.youtube.com/watch?v=0VO-ts0dsbI>
* Interesting (Kaggle): “Synthetic Financial Datasets For Fraud Detection”: <https://www.kaggle.com/ntnu-testimon/paysim1>
  + Dataset available
  + Labeled data available
  + Codes available

**Healthcare fraud literature:**

* Maybe: “Using data mining to detect health care fraud and abuse: a review of literature”: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4796421/pdf/GJHS-7-194.pdf>
  + The literature review is useful.
  + When fraudsters become aware of a particular detection method, they will adapt their strategies to avoid detection (Sparrow, 1996). As we noted above, supervised methods are useful in detecting previously known patterns of fraud and abuse. In theory, we can apply unsupervised approaches to identify new types of fraud or abuse.
  + In another study, association rules mining were applied to examine claims of specialist physicians (Shan et al., 2008). The data was organized in transactions which were defined as all the items claimed or billed for one patient on one day by one specialist. Association rules are statements of the form if antecedent (s) then consequent (s). For example, if a physician prescribed drug A and drug B then he will prescribe drug C with a likelihood of 98%. They identified 215 association rules.
* Maybe: “Fraud types”: <https://legaldictionary.net/fraud/>
* Interesting: Big data analytics in healthcare: <https://www.biomedcentral.com/track/pdf/10.1186/2047-2501-2-3?site=hissjournal.biomedcentral.com>
  + Big data in healthcare is overwhelming not only be- cause of its volume but also because of the diversity of data types and the speed at which it must be managed [7].
  + The totality of data related to patient healthcare and well- being make up “big data” in the healthcare industry. It includes clinical data from CPOE and clinical decision support systems (physician’s written notes and prescrip- tions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine generated/sensor data, such as from monitoring vital signs; social media posts, in- cluding Twitter feeds (so-called tweets) [8], blogs [9], status updates on Facebook and other platforms, and web pages; and less patient-specific information, including emergency care data, news feeds, and articles in medical journals.
  + Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific in- dividual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Certain devel- opments or outcomes may be predicted and/or esti- mated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications; patients at risk for sepsis, MRSA, C. difficile, or other hospital-acquired illness; illness/disease progression; pa- tients at risk for advancement in disease states; causal factors of illness/disease progression; and possible co- morbid conditions (EMC Consulting).
  + Future applications of real-time data, such as detecting infections as early as possible, identifying them swiftly and applying the right treatments (not just broad-spectrum antibiotics) could reduce patient morbidity and mortality and even prevent hospital outbreaks. Already, real-time streaming data monitors neonates in the ICU, catching life-threatening infections sooner [6]. The ability to per- form real-time analytics against such high-volume data in
  + The 4 “Vs” of big data analytics in healthcare: volume, velocity, variety, veracity
* Interesting: “Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection”: <https://pdfs.semanticscholar.org/b03c/b34e9268055ceee20a780e79c519de0a2aa3.pdf>
  + The table “Medical environment overview” is helpful
  + According to Sparrow [6] there are two different types of fraud: “hit-and-run” and “steal a little, all the time”. “Hit-and-run” perpetrators simply submit many fraudulent claims, receive payment, and disappear. “Steal a little, all the time” perpetrators work to ensure fraud goes unnoticed and bill fraudulently over a long period of time. The provider may hide false claims within large batches of valid claims and, when caught, will claim it an error, repay the money, and continue the behavior. The FBI [7] highlights and categorizes some of the most prevalent known Medicaid fraud schemes:
  + • Phantom Billing – Submitting claims for services not provided. • Duplicate Billing – Submitting similar claims more than once. • Bill Padding – Submitting claims for unneeded ancillary services to Medicaid. • Upcoding – Billing for a service with a higher reimbursement rate than the service provided. • Unbundling – Submitting several claims for various services that should only be billed as one service. • Excessive or Unnecessary Services – Provides medically excessive or unnecessary services to a patient. • Kickbacks – A kickback is a form of negotiated bribery in which a commission is paid to the bribe-taker (provider or patient) as a quid pro quo for services rendered [8]. Sparrow [6] proposes that for effective fraud detection one has to look at the data beyond the transaction level, defining seven levels of healthcare fraud control (see Table 2).
* Interesting: “Improving fraud and abuse detection in general physician claims: a data mining study”: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770922/>
  + Look at the features in Table 3: abuse detection and fraud detection

**Healthcare fraud github:**

* Data from Centers for Medicare and medicaid services: <https://www.r-bloggers.com/exploring-us-healthcare-data/>
  + Only dataset is interesting
* Detecting medicare fraud: <https://blog.dataiku.com/2015/08/12/medicare-fraud>
  + Dataset available
  + Labeled data available
* Neo4j Graphgists to detect fraud ring: <https://neo4j.com/graphgist/credit-card-fraud-detection>

<< NOT SO INTERESTING >>

* Background information: <http://www.acl.com/pdfs/DP_Fraud_detection_BANKING.pdf>
* Background information: <https://www.fbi.gov/scams-and-safety/common-fraud-schemes>
* Background information: <http://www.acfe.com/rttn-occupational.aspx>
* Not useful: “Hybrid association rule learning and process mining for fraud detection”: <http://www.iaeng.org/IJCS/issues_v42/issue_2/IJCS_42_2_01.pdf>
* Not useful: “Fraud detection method and system”: <https://www.google.com/patents/US8725524>
* Not useful: “Survey of Insurance Fraud Detection Using Data Mining Techniques”: <https://arxiv.org/pdf/1309.0806.pdf>
* Not good enough: “A survey on statistical methods for health care fraud detection”: <https://www.researchgate.net/profile/Jianjun_Shi/publication/23290716_A_survey_on_statistical_methods_for_health_care_fraud_detection/links/553e4b5b0cf210c0bda937cf/A-survey-on-statistical-methods-for-health-care-fraud-detection.pdf>
* Only TOC: “Principles of fraud examinations”: <http://www.cpestore.com/pdf_courses/AA144404/AA144404_toc.pdf>
* Maybe the cited papers: “Survey of clustering based financial fraud detection research”: <https://www.researchgate.net/profile/Andrei_Sabau/publication/267686837_Survey_of_Clustering_based_Financial_Fraud_Detection_Research/links/57597dd308ae9a9c954eff50/Survey-of-Clustering-based-Financial-Fraud-Detection-Research.pdf>
* Not useful: <http://www.dot.state.sc.us/inside/pdfs/ocia_fraud_risk_management_rprt_111209.pdf>
* <https://www.slideshare.net/just2luku/analysis-ofcreditcardfaultdetection?next_slideshow=1>
* <http://cs.uwc.ac.za/~abagula/cos730/docs/ImportantSlides-Outlier-Detection.pdf>
* The challenges of healthcare fraud: <https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx>
* “Graph analysis for detecting fraud, waste and abuse in healthcare data”: <https://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/viewFile/9705/9888>
  + Fraud, waste and abuse (FWA)
* Maybe: “Health data analytics”: <https://vtechworks.lib.vt.edu/bitstream/handle/10919/75176/HDA-TOC.pdf?sequence=1&isAllowed=y>
* Healthcare fraud github: <http://sujitpal.blogspot.com/2014/05/outlier-detection-on-medical-claims.html>

# 