Capstone Project - Healthcare Fraud Dataset
Project proposal

1. Group description
1.1. Group name

Medical Detectives (MDs)

1.2. Students names, background and target industry if any

Deborah Leong (Finance, open to diverse opportunities)
Doug Devens (Engineering, open to diverse opportunities)
Sam Nuzbrokh (Engineering/Physics, open to investigation/research)

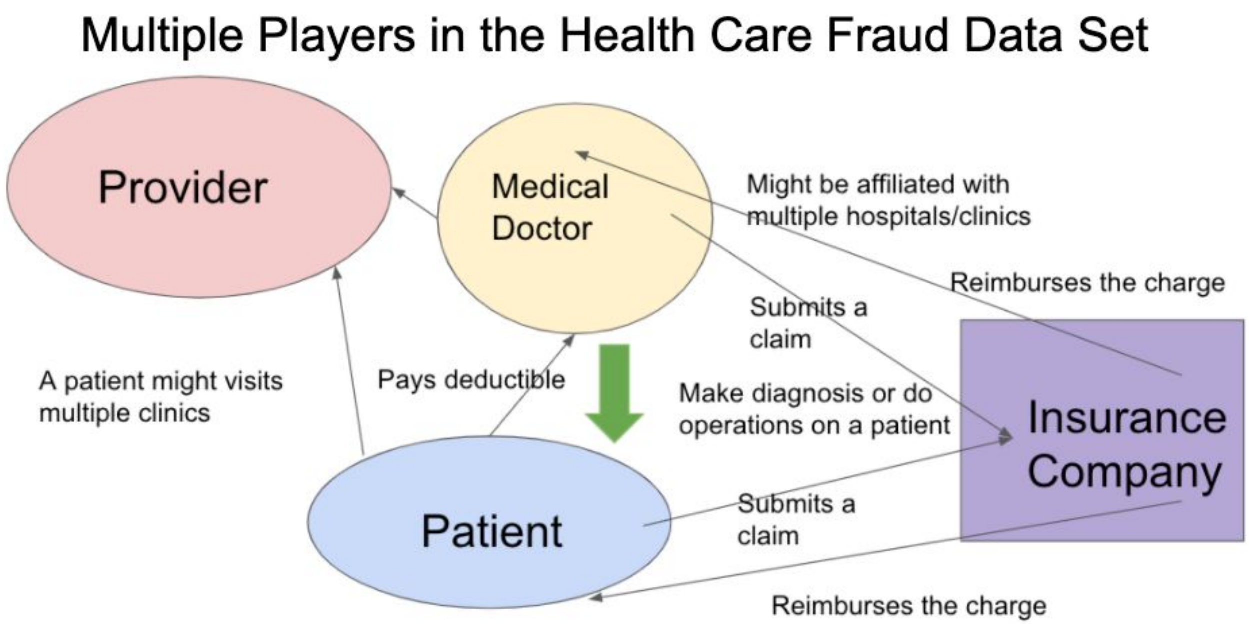
1.3. Group structure: roles and responsibilities

Student	Data science	Project team
Deborah	- beneficiary dataset exploration - business insights - machine learning	- project timeline - presentation
Doug	- SQL Database - inpatient dataset exploration - machine learning	- time management - technical lead
Sam	- NetworkX Visualization & EDA - outpatient dataset exploration - R Shiny Mapping - machine learning	- execute project plan - presentation

2. Why do we want to develop a data science project?

2.1 Objective:

Context:



There are several types of healthcare frauds.
For example:

- Billing for services and covered service not provided
- Duplicate claims
- Misrepresentation in service provided
- Overstating service provided for higher claim amounts, etc.

Identification of potential insurance claim fraud is done manually using data analytics. This project aims to create a model that automates this process.

Objective:

Detect frauds rings among medical practitioners
Provide fraud prediction with confidence intervals
Figure out scale of fraud rings

Measure of success:
- More robust assessment of potential fraud

- Fraud assessment on the level of doctors or on per record basis

- What problem do you want to solve?
    - Detect frauds rings among medical practitioners
    - Provide fraud prediction with confidence intervals?
    - On what scale are fraud rings?
        - County? State? Inter-state? Continental? International? Interplanetary??
- What questions are you trying to answer?
    - Can we find relations between doctors/patients based on shared patients/doctors/hospitals/insurance providers
    - Can we predict with 90% precision actual fraud?
        - Potential fraud is only 9% in data
    - Can we use the knowledge of network connections to adjust our precision threshold?
- How will you **measure the success** of your analysis from a business/user perspective?
    - Confirm or deny potential fraud that was given to us in the data set
        - Given data is of potential fraud - NOT actual
        - Use integrated prediction matrix (use network insights, etc)

2.2. Scope of application: what population and timeframe will your analysis/model be applied to or used for?

Population: Nationwide claims for patients between 26 to 100 years old
Timeframe: November 2008 - December 2009
Target variable: potential fraud

3. How do you translate the objective and scope in terms of data?

3.1. What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

	Train	**Test**	**Total**
Number of records	138,556	63,968	202,524
% Records	68.4%	31.6%	100%
Column Summary	BeneID: 138,556 unique values DOB: from 1905 to 1983 DOD (mostly N/As) Gender: category 1 and 2 Race: category 1 to 5 Renal Disease Indicator: 14% Y State: category 1 to 54 County: category 0 to 999 NoOfMonths_PartACov (mostly 12s rest 0s) NoOfMonths_PartBCov (mostly 12s rest 0s) Category 1 and 2: ChronicCond_Alzheimer 1 ChronicCond_Heartfailure ChronicCond_KidneyDisease ChronicCond_Cancer ChronicCond_ObstrPulmonary ChronicCond_Depression ChronicCond_Diabetes ChronicCond_IschemicHeart ChronicCond_Osteoporasis ChronicCond_rheumatoidarthritis ChronicCond_stroke IPAnnualReimbursementAmt: -$8,000 to $161k IPAnnualDeductibleAmtL: $0 to $38k OPAnnualReimbursementAmt: -$70 to $103k OPAnnualDeductibleAmt: $0 to $14k	BeneID: 63,968 unique values DOB: from 1909 to 1983 DOD (mostly N/As) Gender: category 1 and 2 Race: category 1 to 5 Renal Disease Indicator: 17% Y State: category 1 to 54 County: category 0 to 999 NoOfMonths_PartACov (mostly 12s rest 0s) NoOfMonths_PartBCov (mostly 12s rest 0s) Category 1 and 2: ChronicCond_Alzheimer 1 ChronicCond_Heartfailure ChronicCond_KidneyDisease ChronicCond_Cancer ChronicCond_ObstrPulmonary ChronicCond_Depression ChronicCond_Diabetes ChronicCond_IschemicHeart ChronicCond_Osteoporasis ChronicCond_rheumatoidarthritis ChronicCond_stroke IPAnnualReimbursementAmt: -$1,000 to $156k IPAnnualDeductibleAmtL: $0 to $38k OPAnnualReimbursementAmt: -$60 to $98k OPAnnualDeductibleAmt: $0 to $14k	
IP + OP reimbursement	24%	30%	
Test/Train Observations		IP and OP deductibles are highly concentrated below $1,900 and $700 respectively. IP and OP annual reimbursement are highly concentrated below $68k and $48k respectively.	
General Observations	DOB left skewed, huge drop in DOBs from 1942 onwards Gender: skewed towards 1 Race: skewed towards 5 State and County: unevenly distributed Reimbursement amounts and deductibles vs number of pre-conditions (total 11): - normal distribution, reimbursement max at patients with 5 - 6 pre-conditions - no obvious correlation between reimbursement and deductible amounts vs less pre-conditions		

Train:

ChronicCount1	ChronicCount2	Sum of IPAnnualReimbursementAmt	Sum of IPAnnualDeductibleAmt	Sum of OPAnnualReimbursementAmt	Sum of OPAnnualDeductibleAmt
0	11	3,076,070	377,892	5,473,300	1,648,608
1	10	13,025,870	1,500,148	11,070,600	3,346,700
2	9	28,709,950	3,158,104	16,438,580	4,913,036
3	8	48,087,420	5,099,640	20,933,270	6,207,953
4	7	65,161,050	7,096,838	26,124,030	7,715,715
5	6	79,220,570	8,536,866	28,786,130	8,243,424
6	5	88,927,560	9,646,704	27,921,340	7,955,606
7	4	81,324,080	8,982,720	21,560,240	6,163,582
8	3	58,039,840	6,373,462	13,572,140	3,898,124
9	2	31,311,800	3,433,688	6,326,260	1,768,335
10	1	8,773,080	1,003,248	1,515,510	432,348
11	0	1,505,680	191,932	154,680	41,700
Grand Total		**507,162,970**	**55,401,242**	**179,876,080**	**52,335,131**

Test:

ChronicCount1	ChronicCount2	Sum of IPAnnualReimbursementAmt	Sum of OPAnnualReimbursementAmt	Sum of IPAnnualDeductibleAmt	Sum of OPAnnualDeductibleAmt
0	11	1,278,270	2,205,900	147,384	644,310
1	10	5,935,360	4,969,280	655,284	1,521,250
2	9	13,320,560	8,361,820	1,452,280	2,428,024
3	8	23,086,780	11,195,030	2,470,952	3,333,090
4	7	34,149,710	15,270,750	3,677,976	4,377,079
5	6	44,219,690	17,663,000	4,853,798	5,047,849
6	5	53,337,940	18,103,330	5,714,304	5,141,462
7	4	49,957,000	13,986,350	5,567,454	3,965,462
8	3	37,611,850	9,081,910	4,089,692	2,622,020
9	2	20,927,870	4,463,700	2,288,208	1,223,275
10	1	6,492,580	1,055,030	735,416	297,838
11	0	1,220,360	132,090	157,756	36,720
Grand Total		**291,537,970**	**106,488,190**	**31,810,504**	**30,638,379**

Dataset 2: Outpatient data Assigned to Sam

	Total	Train	Test
Number of Records	643,578	517,737	125,841
Proportion	100%	76%	24%
Summary	● Outpatient care: ○ Hospital or medical facility care received without being admitted ○ Stays < 24 hours (even if this stay occurs overnight) in hospital. ○ ER visits initially considered outpatient ■ If doctor orders you to be formally admitted, then your status becomes inpatient ■ Hospital care you receive is considered inpatient until discharge ○ Despite hospital stay, care may be considered outpatient if you receive care on the same day as discharge - even if you spend the night in the hospital ○ If doctor orders observation of condition or tests to help diagnose condition, you remain classified as outpatient until doctor orders readmission ● Loans originating during the last 5 years ● Internal system data		
Unique Distributions	● Columns not present in outpatient that are in inpatient ○ AdmissionDt ■ Patient is not admitted to hospital by definition ○ DischargeDt ■ Patient was not discharged from hospital by definition ○ DiagnosisGroupCode	● Deductibles Paid: ○ 0., 10., 20., 30., 40., 50., 60., 70., 80., 90., 100., 200., 865., 876., 886., 897., 1068. ●	

Dataset 3: Inpatient data

	Column Description	Train	Test
Number of claim records (claimIDs are unique)		40,474	9,551
	'BeneID': beneficiary (patient) identification	31,289 unique values	8351 unique
	'ClaimID': record # for insurer of reimbursement claims	40,474 (key) unique values	9551 unique
	'ClaimStartDt': start date of claimed services	11/27/08 - 12/31/09	11/27/08 - 12/31/09
	'ClaimEndDt':	1/1/09 - 12/31/09	1/1/09 - 12/31/09
	'Provider': clinic or hospital providing services	2092 unique values	520 unique
	'InscClaimAmtReimbursed': amount paid by insurance co	From $0 to $125,000	From $0 to $125,000
	'AttendingPhysician': physician responsible for patient's care	11,605 unique values, 112 NA	2658 unique, 31 NA
	'OperatingPhysician': physician performing operation/svcs	8,288 unique values, 16,644 NA	1871 unique, 3962 NA
	'OtherPhysician', other physician providing consult/help	2,878 unique values, 35,784 NA	659 unique, 8538 NA
	'AdmissionDt': date admitted to hospital	11/27/08 - 12/31/09	11/27/08 - 12/31/09
	'ClmAdmitDiagnosisCode': ICD code for admission	1,928 unique values	1113 unique
	'DeductibleAmtPaid': amount paid as part of insurance cvrge	Only 1 value, $1,068, 899 NA	1 value, $1,068, 196 NA
	'DischargeDt': date let go from hospital	1/1/09 - 12/31/09	1/1/09 - 12/31/09
	'DiagnosisGroupCode': ?	736 unique values	712 unique,
	'ClmDiagnosisCode_1',: ICD code for diagnosis performed	2,554 unique values	1298 unique
	'ClmDiagnosisCode_2': ‘’	2,440 unique values, 226 NA	1383 unique, 54 NA
	'ClmDiagnosisCode_3': ‘’	2,428 unique values, 676 NA	1387 unique, 169 NA
	'ClmDiagnosisCode_4': ‘’	2442 unique values, 1534 NA	1344 unique, 404 NA
	'ClmDiagnosisCode_5': ‘’	2375 unique values, 2894 NA	1323 unique, 719 NA
	'ClmDiagnosisCode_6': ‘’	2359 unique values, 4838 NA	1313 unique, 1197 NA
	'ClmDiagnosisCode_7',: ‘’	2311 unique values, 7258 NA	1248 unique, 1736 NA
	'ClmDiagnosisCode_8',: ‘’	2244 unique values, 9942 NA	1290 unique, 2360 NA
	'ClmDiagnosisCode_9': ‘’	2095 unique values, 13497 NA	1159 unique, 3238 NA
	'ClmDiagnosisCode_10': ‘’	953 unique values, 36547 NA	384 unique, 8664 NA
	'ClmProcedureCode_1',: ICD code for therapy performed	1118 unique values, 17326 NA	658 unique, 4118 NA
	'ClmProcedureCode_2': “	298 unique values, 35020 NA	171 unique, 8297 NA
	'ClmProcedureCode_3',: “	155 unique values, 39509 NA	68 unique, 9328 NA
	'ClmProcedureCode_4': “	49 unique values, 40358 NA	21 unique, 9522 NA
	'ClmProcedureCode_5': “	7 unique values, 40465 NA	3 unique, 9549 NA
	'ClmProcedureCode_6': “	40474 NA	9551 NA

Observations:

- Most patients were in once or twice, with a few in as many as 8 times in both train and test
- Most providers (clinics/hospitals) had 0-50 claims/visits, but some had 400-600 in train and test
- Most attending physicians had 0-25 claims/visits, but some had 300-400 claims/visits in train and test

- Most ‘other’ physicians were included on only 0-5 claims but some had as many as 30-70 claims in train and test
- Most operating physicians had 0-25 claims, but some had 150-200 claims in train and test
- Insurance claim reimbursement was usually 0-10,000 but had some as high as 120,000. In a plot of Length of Stay vs reimbursement there was an anomaly at 60,000 reimbursement where it didn’t matter how long the patient stayed
- Most of the claims in train and test had 9 diagnostic codes out of 10 possible codes for diagnosis
- Most of the claims had 0-1 procedural ICD codes, with a few out at 4 codes, in both train and test
- It looks like claim duration and length of stay were capped at 35 days; not sure if that’s real or not.

3.2. What **data treatment and analysis** do you plan? Data aggregation, target variable definition, tools, analysis/machine learning, ...

Data preparation
- merge datasets at BeneID → validate IDs and keys
- merge datasets at diagnosisCode to ICD dataset (external) → validate IDs and keys
- handle missingness

Target variable
- potential fraud at provider level (provided in train dataset)
- experimentation: potential fraud at physician/claim level (not provided at all -> unsupervised)
    - any relationships between physicians and providers?

Tools
- Data preparation in Python
- Network Analysis in R/Shiny [Data extraction in SQL]
    - Degrees of Separation between Actors
    - Claim Network Growth over Time (Time indexed by Claim Date)
    - Minimum Path between Major Actors
    - Provider Level Network Clusters
    - Reimbursement Weighted Directed Graphs (follow the money)
- Supervised model development in Python
- Unsupervised model development in Python

Analysis
- Exploratory data analysis:
    - univariate and bivariate analyses → initial insights to share with stakeholders
- Exploratory data analysis - network analysis
    - Unsupervised learning
    - Shiny Visualizations
    - Clustering

- Prediction model:
    - Logistic regression/Gradient Boost/SVM/Random Forest for fraud classification
    - Clustering for physician classification based on association with potentially fraudulent providers

4. Project plan

** Daily check-in as a team for discussion around findings, questions and progress**

			June						
			Mo	Tu	We	Th	Fr	Sa	Su
Lead	**Jupyter Notebook Cross Reference**	**Focus Area: Kickoff**						6	7
Doug, Deb, Sam		Project Scoping							
Doug		Data Extraction (SQL databse setup)							
Deb		Github Repo setup							
		Focus Area: Data Analysis	8	9	10	11	12	13	14
Doug, Deb, Sam	[Capstone Project Proposal Example]()	Project Declaration with Aiko							
Doug, Deb, Sam	Warmup Questions (1 - 5)	Exploratory Data Analysis							
Doug, Deb, Sam	Unsupervised Market Basket Analysis	Extracting Information from the Patients' Chronic Conditions							
	Milestone 1:	**Complete Data Analysis, Set direction for Machine Learning**							
		Focus Area: Machine Learning	15	16	17	18	19	20	21
Doug, Deb	Unsupervised Clustering	Unsupervised Clinic							
Sam, Deb	Unsupervised Clustering	Doctor Network Analysis							
Doug, Sam	Unsupervised Clustering	Weighted Graphs and NetworkX							
Doug, Sam	Supervised Learning and Anomaly Detection	Explore SVM, RandomForests							
	Milestone 2:	**Conclusion on Achievement of Objectives**							
		Focus Area: Delivery	22	23	24	25	26	27	28
Doug, Deb, Sam		Powerpoint preparation (Note: Final Exam 22nd)							
		Capstone Project due							
		Presentation							