L4:
Clinical Data Science

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

UC San Diego SCHOOL OF MEDICINE

Agenda

- Clinical Data Environments
  - US (HIPAA):
  - Europe (GDPR)
  - PII and Privacy
  - Developing Countries
- Clinical Machine Learning
  - Tabular data: Diabetes severity
  - Visualization: Scatterplots
  - Machine Learning: Linear Regression
  - Feature engineering

Clinical Data

- Data Quantity
  - Difficult in the US (HIPAA)
  - Easier in Europe (GDPR)
  - Easiest in developing economie
- Data Quality
  - Complete
  - Correct
  - Available

Centralized US Data Resource (CDC)

- CDC (Center for Disease Control): bit.ly/ucsd-cdc
- data.gov/health

Mortality
Nutrition
Ambulatory care
Insurance stats
Discharge stats
Tobacco use

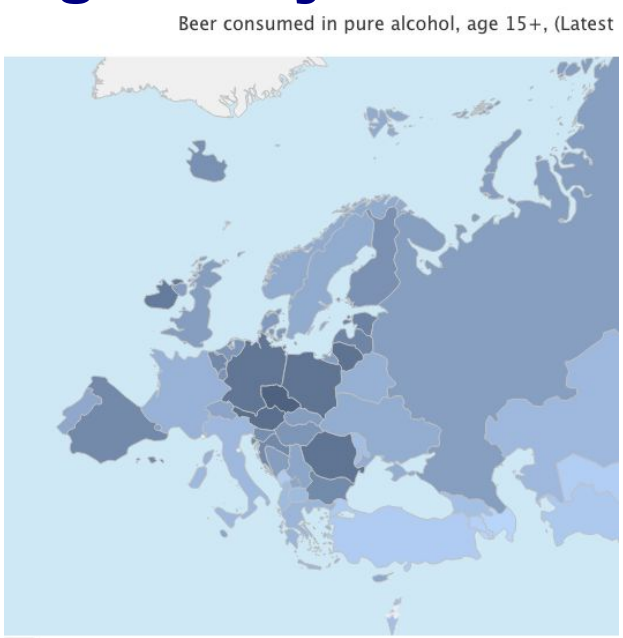
ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/

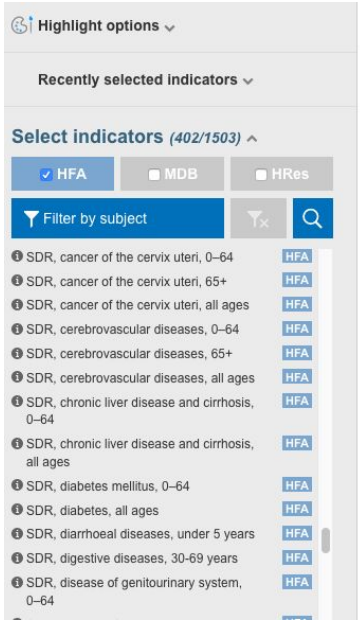
EU (European Union)

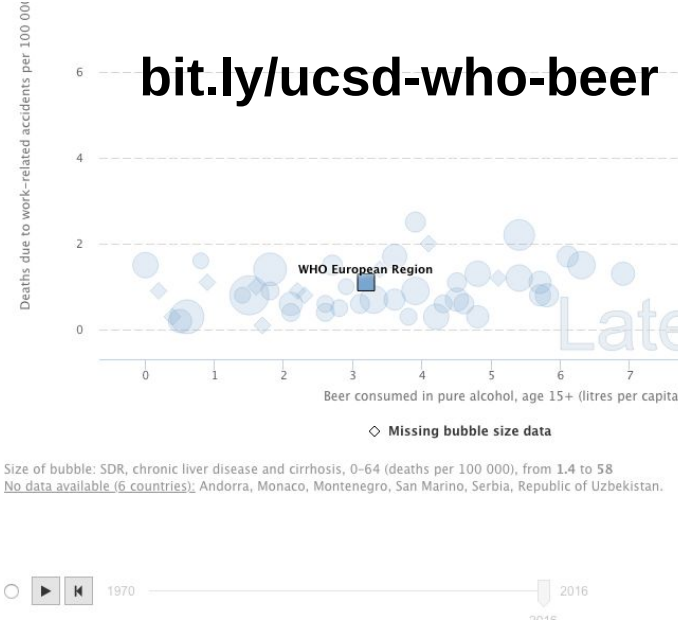
EEA (Economic Area)

Data in Europe

- WHO (World Healthcare Organization):
  gateway.euro.who.int



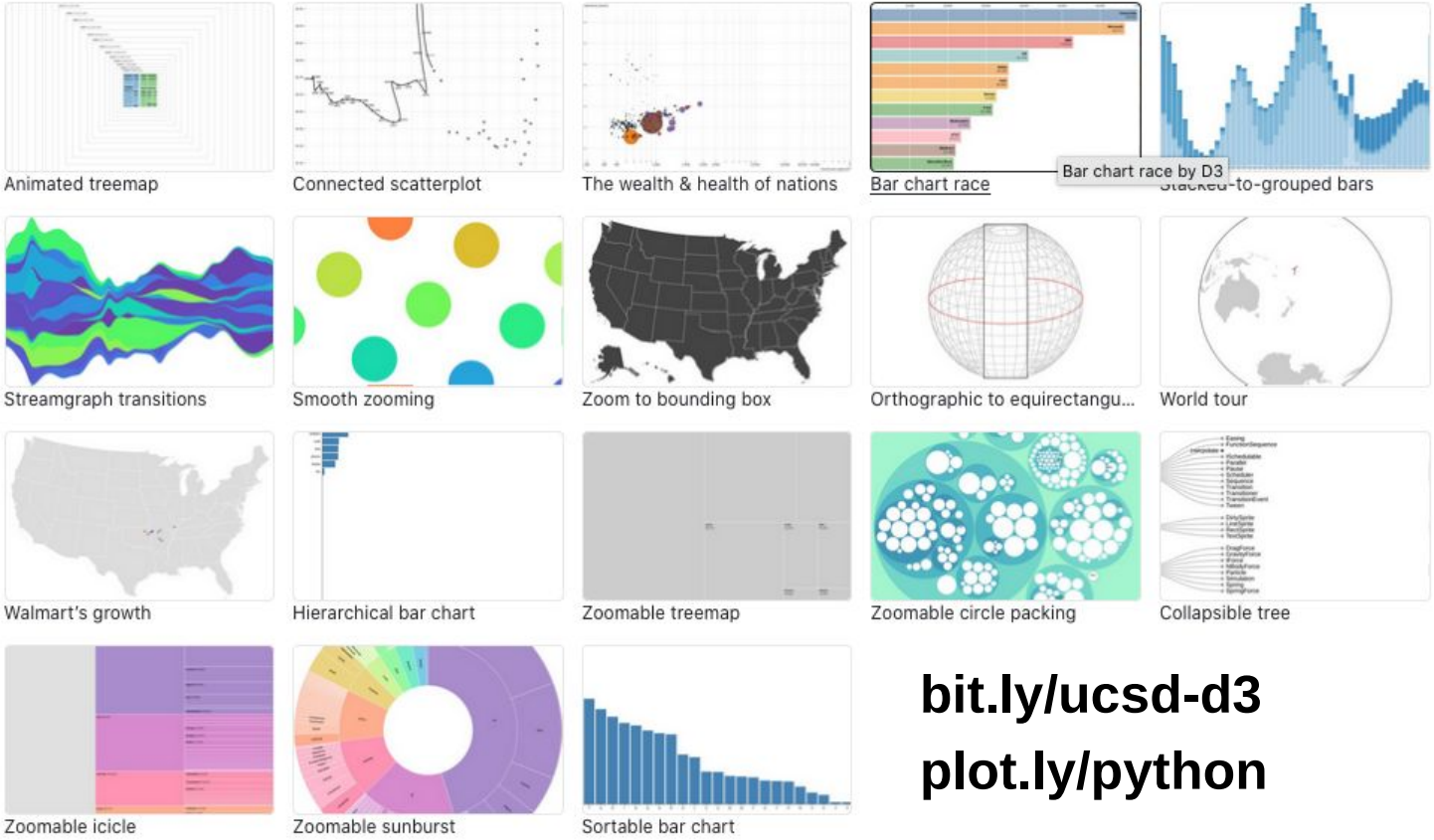




UK Hospital Data (bit.ly/ucsd-hes)

Hospital provider code and description†	Finished consultant episodes	Admissions	Male	Emergency	Waiting list	Mean time waited	Median time waited	Mean length of stay	Median length of stay	Mean age
East Midlands Strategic Health Authority	**1,305,107**	**1,121,792**	**570,951**	**387,834**	**394,779**	**45**	**31**	**5.7**	**1**	**52**
5N6 Derbyshire County PCT	10,469	10,452	4,334	1,284	5,601	44	36	26.5	18	66
5N8 Nottinghamshire County Teaching PCT	2,092	1,924	785	593	303	3	1	30.9	24	77
5PA Leicestershire County And Rutland PCT	14,529	14,469	6,350	1,015	9,403	31	23	27.1	20	67
5PC Leicester City PCT	421	*	139	*	*	-	-	19.0	16	80
5PD Northamptonshire Teaching PCT	1,748	1,740	857	224	513	-	-	30.4	11	69
NT322 Spire Leicester Hospital	38	38	*	-	-	-	-	1.5	1	56
NT407 BMI - Chatsworth Suite	127	127	65	-	127	34	34	1.2	1	51
NT427 BMI - The Park Hospital	124	124	63	-	124	22	22	1.5	1	51
NT441 BMI - Three Shires Hospital	431	431	254	-	431	-	-	2.1	1	52
NT450 BMI The Lincoln Hospital	*	*	*	-	*	156	96	-	-	52
NTA04 Nottingham NHS Treatment Centre(Nations Healthcare)	24,262	24,262	10,810	*	16,833	118	39	-	-	54
NVC23 Woodland Hospital	2,476	2,476	1,197	-	2,476	30	13	2.7	2	53
NVC26 Gainsborough NHS Treatment Centre	*	799	326	-	799	40	34	3.0	3	66
NVC27 Boston NHS Treatment Centre	2,409	2,409	1,120	-	2,409	31	23	4.3	2	62
NVC40 Nottingham Woodthorpe Hospital	2,630	2,630	1,209	-	2,630	42	14	2.4	1	53
RFS Chesterfield Royal Hospital NHS Foundation Trust	80,569	70,178	34,621	27,929	24,594	27	24	4.2	1	52
RHA Nottinghamshire Healthcare NHS Trust	4,049	3,168	2,196	977	402	1	1	71.0	17	46

D3.js and Plot.ly



bit.ly/ucsd-d3

plot.ly/python

PII (Personally Identifiable Information)

- Full name
- ID Number
- SSN, Driver's License
- Telephone, Credit Card
- Combinations
  - Birthdate + neighborhood
  - Hospital + age (if over 80)
- Why protected?
  - Prevent discrimination
  - Profit, bias, politics

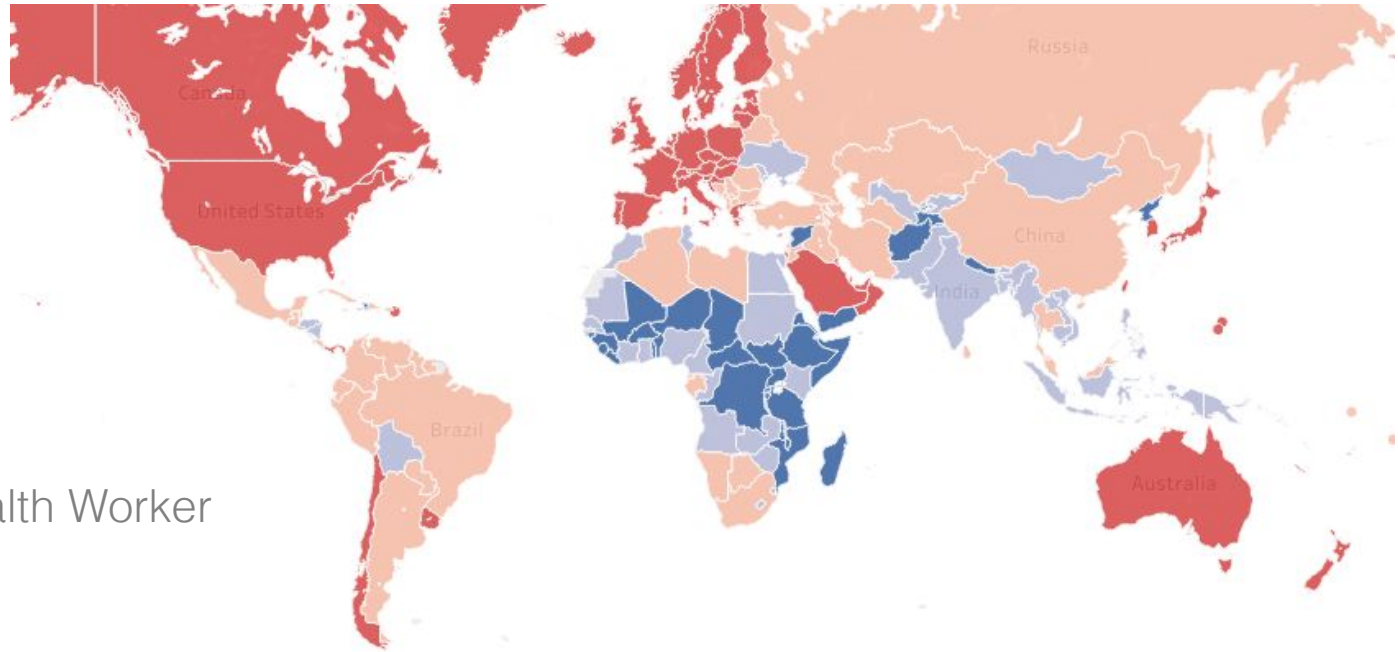
Anonymization

- Delete PII columns/fields
- Shuffle PII
- Add noise to PII
- Implement **“Differential Privacy”**
- Wait 50 years after individuals die
- **Wait until the data is made public**

bit.ly/ucsd-deanon
bit.ly/ucsd-netflix

Developing World is Different

- Hospital
- Clinic
- In the field
- Dr's office
- Home
- Community Health Worker
- Mobile app



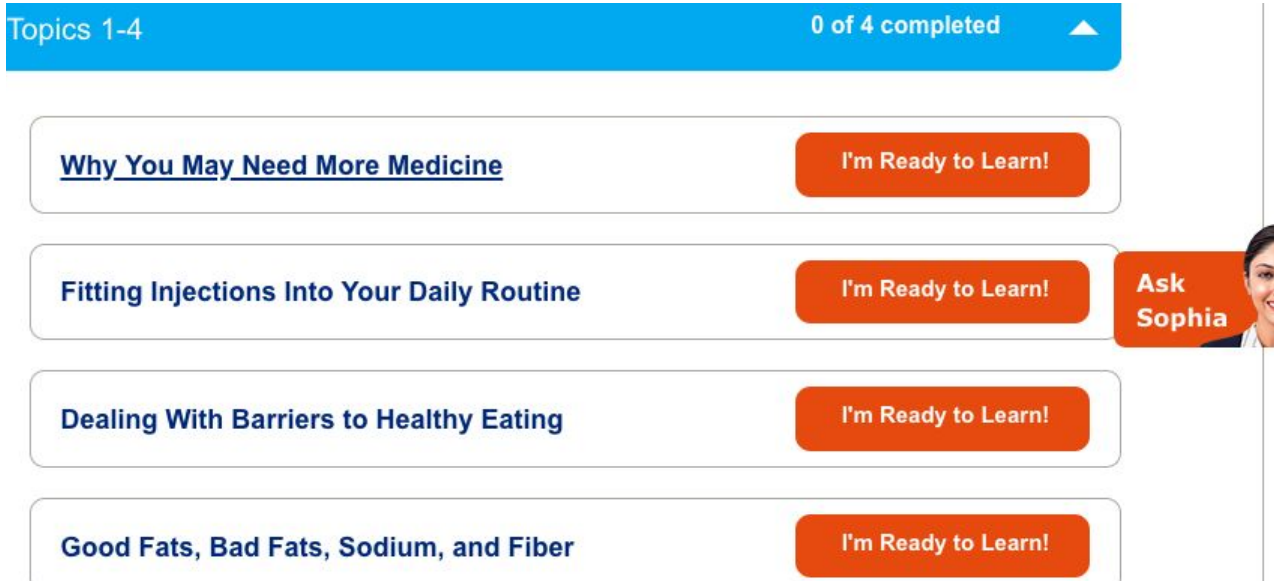
Clinical Data Science in Developing World

- Rural
- No clinic
- Intermittent Wireless
- Little Internet (WiFi)
- NGOs deliver care

Mobile Apps like **CommCare** (by Dimagi)

US Pharma Bots

- Diabetes coach
- Pharma **prior**:
  - insulin
  - medication

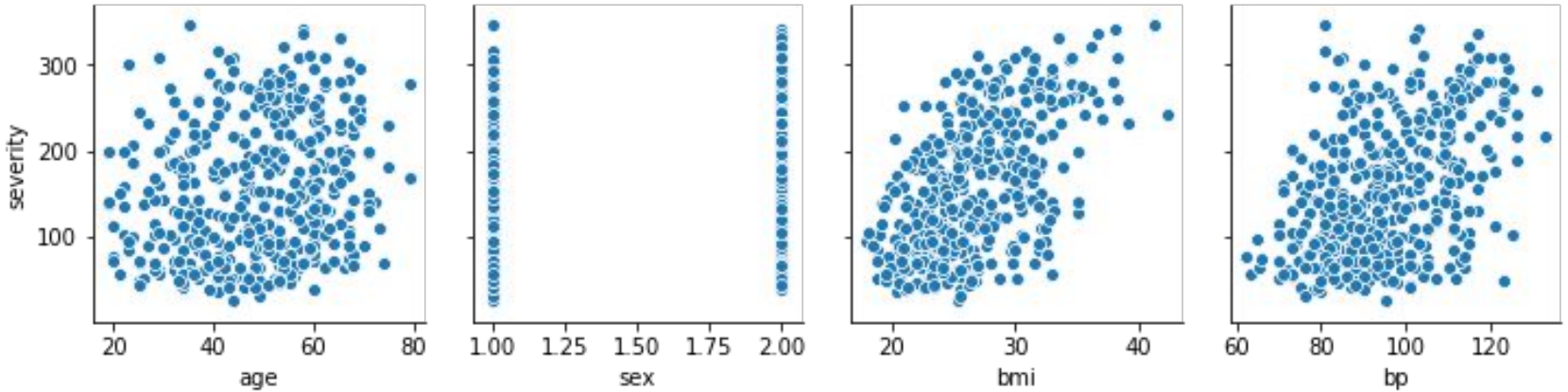


Clinical Dataset (Diabetes Severity)

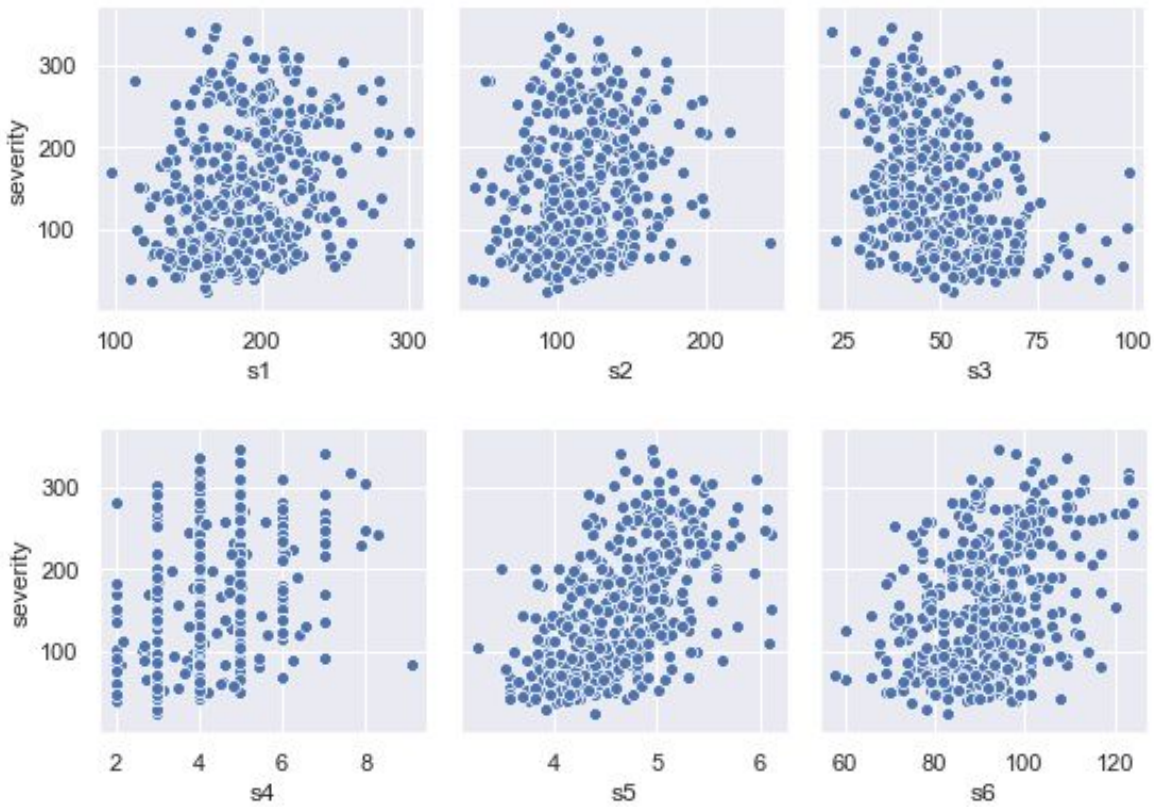
bit.ly/ucsd-diabetes

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	severity
0	59	2	32.1	101	157	93.2	38	4	4.9	87	**151**
1	48	1	21.6	87	183	103.2	70	3	3.9	69	**75**
2	72	2	30.5	93	156	93.6	41	4	4.7	85	**141**
3	24	1	25.3	84	198	131.4	40	5	4.9	89	**206**
4	50	1	23.0	101	192	125.4	52	4	4.3	80	**135**
5	23	1	22.6	89	139	64.8	61	2	4.2	68	**97**
6	36	2	22.0	90	160	99.6	50	3	4	82	**138**
7	66	2	26.2	114	255	185.0	56	4.6	4.2	92	**63**
8	60	2	32.1	83	179	119.4	42	4	4.5	94	**110**
9	29	1	30.0	85	180	93.4	43	4	5.4	88	**310**
10	22	1	18.6	97	114	57.6	46	2	4	83	**101**
11	56	2	28.0	85	184	144.8	32	6	3.6	77	**69**
12	53	1	23.7	92	186	109.2	62	3	4.3	81	**179**

Clinical Dataset (Age, Gender, BMI, BP)



Clinical Dataset (Blood Test Results)



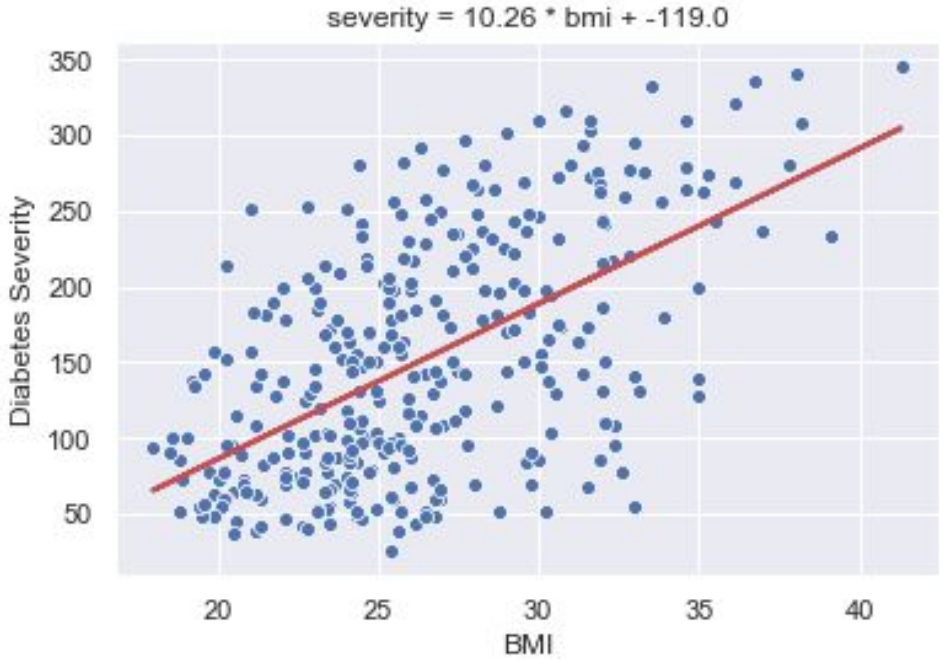
Single-variate Linear Regression (BMI)

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df[feature_names], df[target_names])
print(f'X_train.shape: {X_train.shape}; y_train.shape: {y_train.shape}')
print(f' X_test.shape: {X_test.shape};  y_test.shape: {y_test.shape}')
```

```
X_train.shape: (331, 10); y_train.shape: (331, 1)
 X_test.shape: (111, 10);  y_test.shape: (111, 1)
```

```
lr = LinearRegression()
lr = lr.fit(X_train[['bmi']],
            y_train)
```

```
y_train_pred.shape: (331,)
e_train.shape: (331,)
 mae_train: 51.0
rmse_train: 61.7
```



Test Set Error Larger than Training Set?

Training Set Error

```
mae_train: 51.0
rmse_train: 61.7
```

<

Unseen Test Set Error

```
mae_test: 53.3
rmse_test: 64.7
```

```
round((rmse_test - rmse_train) / rmse_train, 3)
```

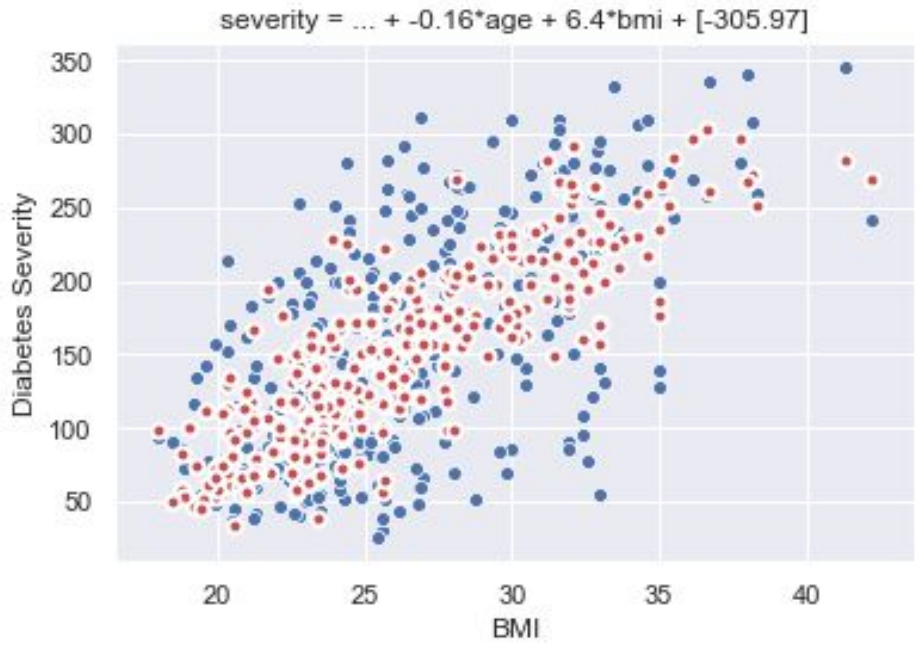
```
0.048
```

Test set error is **5%** larger than training set error for this simple model

Multivariate Linear Regression (Age, BMI, BP)

```
lr_multi = LinearRegression()
lr_multi = lr_multi.fit(X_train, y_train.reshape(-1,1))
print(f'lr_multi.intercept_: {lr_multi.intercept_.round(2)}')
print('lr_multi_coef:')
lr_multi_coef = pd.Series(lr_multi.coef_[0], index=feature_names)
print(lr_multi_coef_.round(2))
```

```
lr_multi.intercept_: [-305.97]
     lr_multi_coef_:
age     -0.16
sex    -19.89
bmi      6.40
bp       1.04
s1      -0.77
s2       0.41
s3      -0.03
s4       5.02
s5      64.10
s6       0.12
```



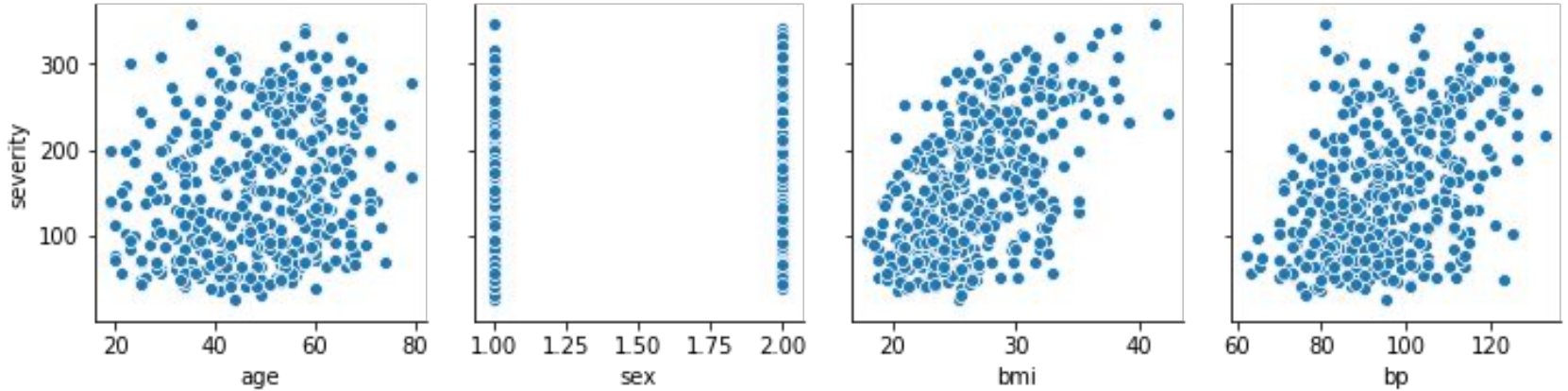
Coefficients of Multivariate Model

-0.16*age

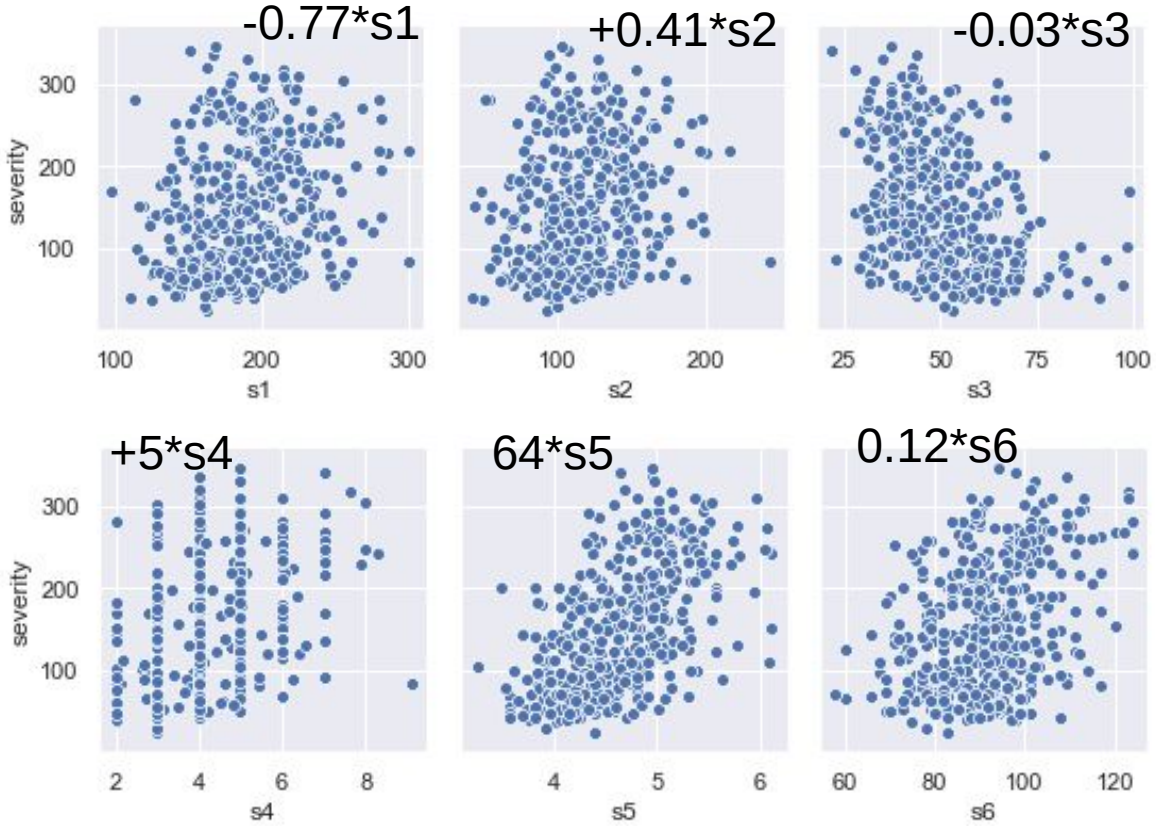
-20.89*sex

+6.4*bmi

+1.0*bp



Blood Test Coefficients



Too Many Features (Overfitting)?

Training Set Error

```
mae_train: 42.9
rmse_train: 52.7
```

<<

Unseen Test Set Error

```
mae_test: 45.1
rmse_test: 56.4
```

```
round((rmse_test - rmse_train) / rmse_train, 4)
```

```
0.0695
```

Test set error is **7%** larger than training set error for this simple model