L2: Spreadsheet Data Science

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

UC San Diego SCHOOL OF MEDICINE

Agenda

Topic	Concepts
Advantages	When to use a spreadsheet
Applications	DIY, Healthcare
Data	Data types
Database	Relational database Graph database (nosql)
Spreadsheet statistics	Sum, mean, standard deviation
Spreadsheet visualization	Scatter plots & histograms
Spreadsheet modeling	Linear regression Classification

Advantages

- Approachable
- Universal
- Transparent
- Automatic variables (“A1”, “B3”)
- Data + Code + Graphics
- Killer feature: automatic filters

E2 | f_x $\sum$ = | =C2*1.9 + IF(B2="Male", 40, 5)

	A	B	C	D	E	F
1	PatientID	Gender	Height (in)	Weight (lb)	Weight Est. (lb)	Weight Error (%)
2	0	Male	67.04	167.91	167.37	-0.32%
3	1	Female	63.09	124.83	124.86	0.03%
4	2	Male	64.37	162.50	162.30	-0.12%
5	3	Female	66.41	161.99	131.18	-19.02%
6	4	Female	64.80	152.83	128.12	-16.17%
7	5	Female	66.21	140.51	130.80	-6.91%
8	6	Male	66.33	156.61	166.02	6.01%

DIY Spreadsheet Data Science

bit.ly/ucsd-lewis

- Fitness: step count, heart rate
- Psychology: Emotional support
- Sleep: CPAP data, smart watch
- Heart disease: wearable EKG
- DIY CGM + insulin pump
- DIY Parkinson's pharma

bit.ly/ucsd-sturr

Healthcare Applications

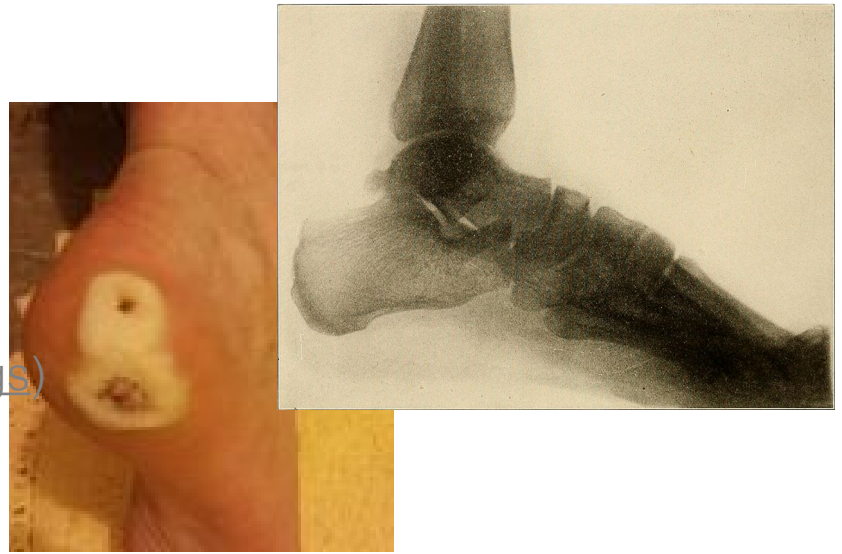
- Patient “intake”
- Operations management
- Insurance actuarial tables
- Presentation (visualization)
- Cooperative analysis

A	B	C	D
PatientID	Gender	Height (in)	Weight (lb)
0	Male	67.04	167.91
1	Female	**unknown**	124.83
2	Male	64.37	**unknown**
3	**unknown**	66.41	161.99

Kinds of Datasets

- Tabular data
- Relational databases (SQL)
  - Hospital, insurance, pharma, etc
- Images
  - X-Rays, MRI, CAT, dermatology
- Time series
  - Hospital records, log files, audio
- Image time series
  - Video of patient interviews (Awakenings)
- Unstructured data (natural language):
  - Doctor & nurse notes

A	B	C	D
PatientID	Gender	Height (in)	Weight (lb)
0	Male	67.04	167.91
1	Female	**unknown**	124.83
2	Male	64.37	**unknown**
3	**unknown**	66.41	161.99



Spreadsheet Tabs vs Relational Database Tables

- **Relational** Database captures **relation**ships
- Structured Query Language (SQL):

```
SELECT IQ.Age FROM Patient WHERE Patient.Height > 65
```

IQ

PrimaryKey

PatientID	Gender	Age	FIQ	VFT1_T0	VFT2_T1	VFT3_T0	VFT1_T6	VFT2_T6	VFT3_T6
A	Girl	14	92	3	4	4	3	4	4
B	Boy	14	83	3	3	4	3	3	3
C	Girl	10	94	3	4	4	4	3	5
D	Girl	12	92	4	2	3	4	4	4
E	Girl	14	101	3	2	3	3	3	4
F	Girl	13	89	3	4	4	3	5	4
G	Girl	10	99	2	4	4	2	4	5

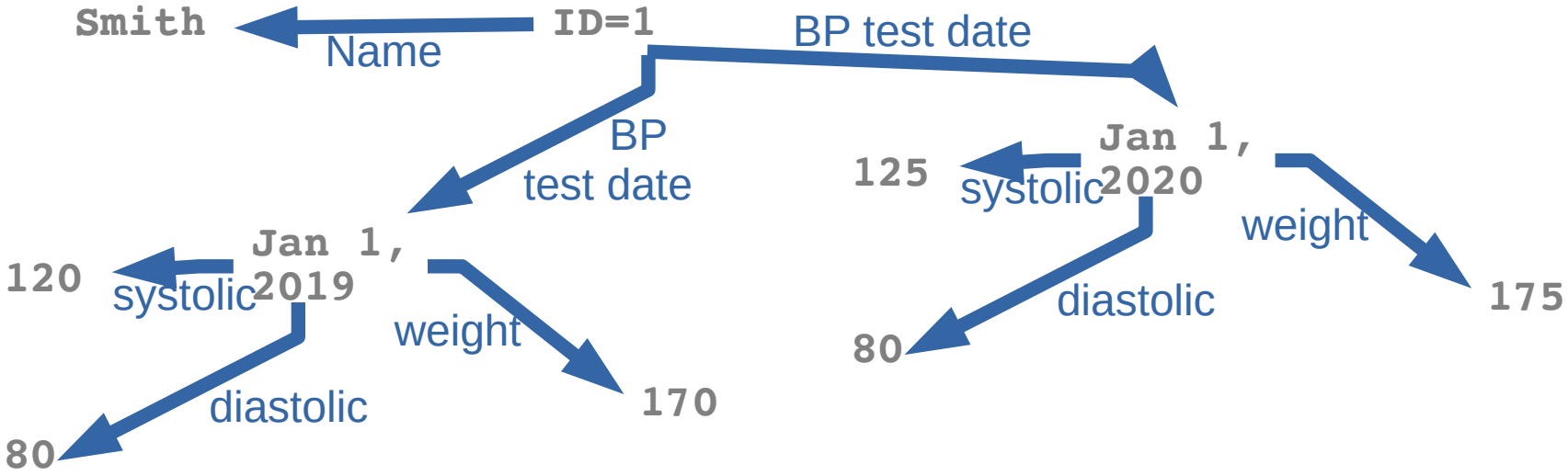
Patient

ForeignKey

PatientID	Gender	Height (in)	Weight (lb)
A	Girl	67.04	167.91
B	Boy	63.09	124.83
C	Girl	64.37	162.50
D	Girl	66.41	161.99
E	Girl	64.80	152.83
F	Girl	66.21	140.51

NOSQL Database (Graph Database)

- **Deep, dynamic relationships**



Data Dichotomies

- Continuous vs discrete
- Categorical vs numerical
- High-D, Low-D
- Structured (Labeled) vs unstructured
- Deterministic vs Chaotic/Random (Noise)
- Big (“out of core”) vs small

Data File Formats

- Text files (CSV, TSV, JSON, TXT)
- Compressed (ZIP, GZ)
- Binary files (XLS, PDF)
- Images (PNG, JPG, TIF)
- Web pages (HTML)
- Databases (SQL, NOSQL, HDF)

- Hint: check out “Pandas” and the `pandas.read_csv()` function

Data types

- **Continuous:** numerical values like height, weight, blood pressure, temperature
- **Categorical:** gender, eye color, disease names
- **Natural language:** symptom descriptions, medical procedure descriptions
- **Sequence:** genome, DNA, RNA, protein, chemical pathways
- **Time series:** treatment timelines, hospital records, EKG/EEG recordings)
- **Geographic:** epidemiology, maps of clinic locations
- **Imagery:** X-rays, MRI slices, CAT scan slices, photos of skin abnormalities

Geographic Data

- Pop Health (Population Health)
- Epi (Epidemiology)
  - Annual flu vaccines around the world
  - Hepatitis outbreak in southern US
  - Ebola epidemic in Africa
- Examples
  - Latitude and Longitude
  - State
  - Zip Code

Sequence Data

- Genomics
  - Self-service genetic testing (23andme)
  - Prenatal screening (Counsyl)
  - Pre-exposure allergy prediction
  - Asthma anticipation
  - Resistance to disease (AIDS, Malaria, West Nile, Ebola)

Weight Guesser Example

- Weight guessing game

- … useful in healthcare for:
  - Data cleaning
  - Detecting anomalies (disease)
  - Body Mass Index (BMI)

Weight Guesser Dataset

- Height
- Weight
- Gender

- Do you notice anything wrong?
- Other uses (targets) for this data?

PatientID	Gender	Height (in)	Weight (lb)
0	Male	67.04	167.91
1	Female	63.09	124.83
2	Male	64.37	162.50
3	Female	66.41	161.99
4	Female	64.80	152.83
5	Female	66.21	140.51
6	Male	66.33	156.61
7	Male	72.39	199.45
8	Female	60.15	103.98
9	Female	56.07	89.57
⋮	⋮	⋮	⋮
9998	Female	66.21	156.54
9999	Male	71.41	216.92

Artificial Pancreas Timeline

1973: Wearable insulin pump (Dean Kaman)

2003: Wearable CGM (glucose monitoring)

2013: Dana Lewis adds #DIYPS to #WeAreNotWaiting

2015: Dana Lewis DIYs an artificial pancreas

2017: Medtronic integrates Dana's technology

2018: Medtronic notifed of security vulnerability

2019: FDA recalls MiniMed insulin pumps