Computing in Context: Fall 2024

Lecture 1 | Overview

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Dr Sarah Ashcroft-Jones | Mailman School of Public Health

Second Quarter – Context-focus

- 2 lectures – more theoretical
- 1 lab – more practical
- Assignments – introduced and explained next week
- Disclaimer

Health Policy

Health Data

Evidence based decision making

HEALTH POLICY

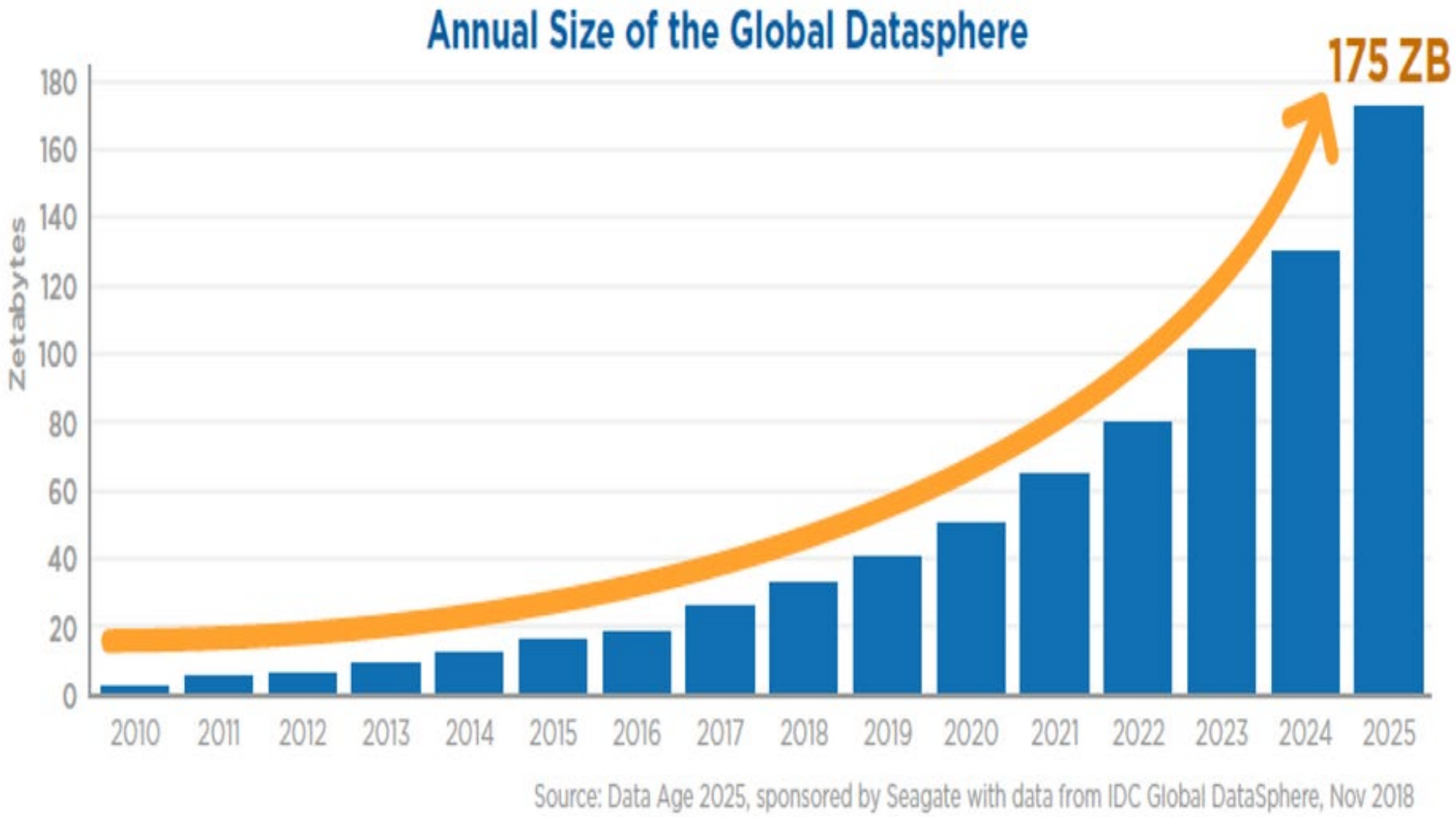
HEALTH DATA

EVIDENCE BASED HEALTH POLICY

Data and Computing

An important aside

Amount of Data Created Each Year



Scale of the Global Data Growth

- 1 zettabyte = a trillion gigabytes
- 1 gigabyte = 1000 megabyte,
- 1 megabyte = 1000 kilobytes
- 1 kilobyte = 1000 bytes
- Each byte has 8 bits of information
- A bit is the smallest unit of digital information

Downloading global datasphere would take 1.8 billion years

Why does it matter?

1. Data is out there – do you need more?
2. Finding signal in the noise – how?
3. Limit may not be your data but your compute

Hardware Components of Computing

- **Processor (i.e. CPU):** The more powerful the processor is, the faster it can perform tasks assigned to it

- **Memory (i.e. RAM):** Akin to your brain's "working memory" – larger memory means the processor can operate on more info

- **Storage (i.e. HDD or SDD):** Akin to your brain's "long-term memory" – data that is not currently being used, but is kept for retrieval/editing

“The Big Data Revolution”

In the past 50 years - explosion in the processor speed, memory size, and storage density

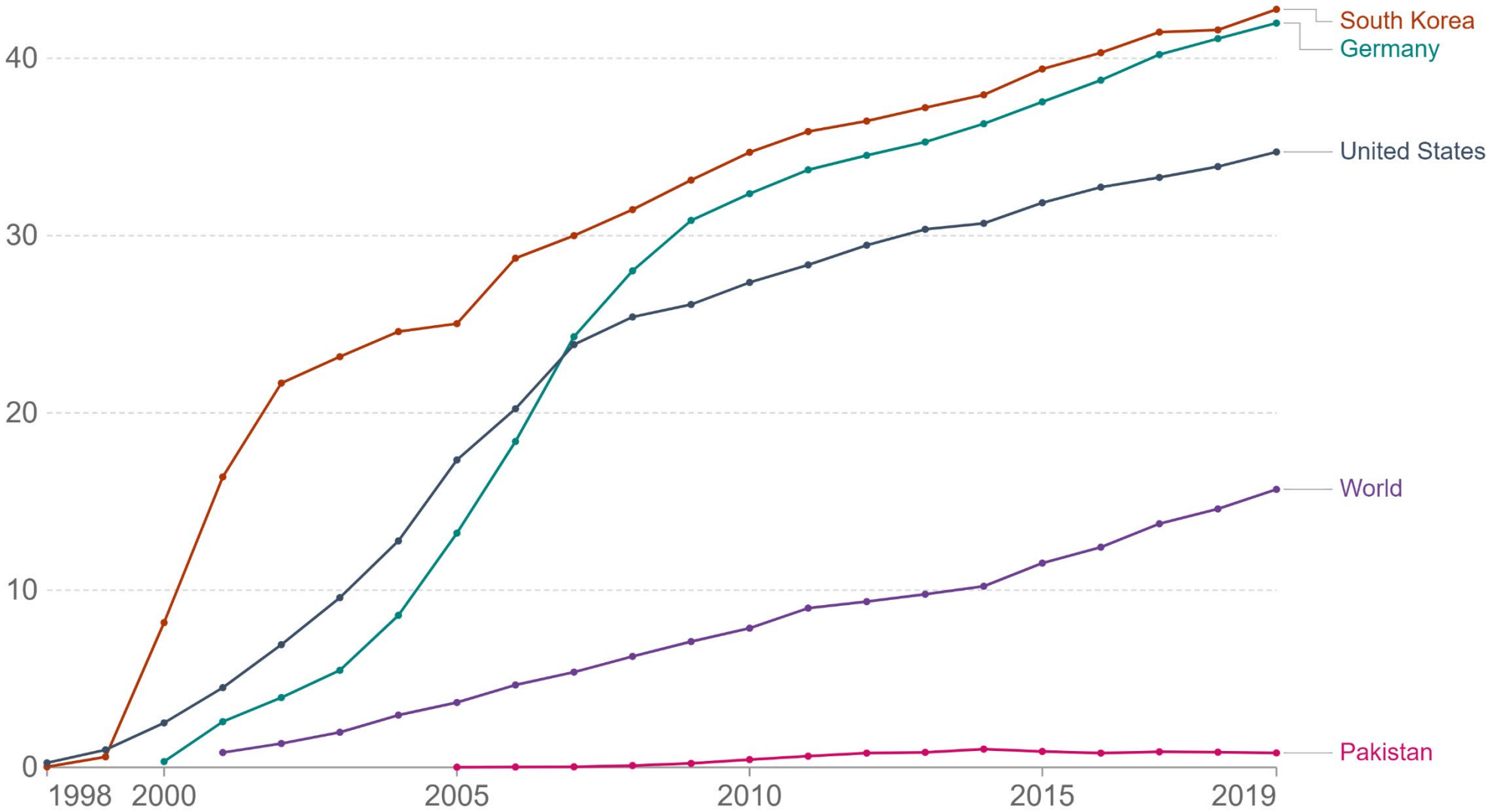
In the past 20 years - broadband (high-speed) internet networks have grown, rapid transmission of large quantities of data *between computers* now feasible

- This makes data more *useful*, as it can be communicated to others

Broadband subscriptions per 100 people, 1998 to 2019

Broadband subscriptions refer to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s.

Our World in Data



Source: International Telecommunication Union (via World Bank)
Note: For more details on the definition see the sources tab.

Health data

Where does it come from?

Traditional Data Source: Surveys

Historically, health data not readily available (or recorded)

- Records weren't digitized (in many cases they still aren't!)

So, quantitative HP research data was collected explicitly:

- Example: surveys administered by government agencies
- e.g. the National Health Interview Survey (NHIS), administered by the Census Bureau since 1957

Upsides to survey data:

- Survey designer can choose what information is collected
- Data is generally well-structured → “ready” to analyze

Downsides to survey data:

- Expensive to collect → limits sample size & question #
- Respondents may provide inaccurate information
- Can only know what the respondent can readily recall

Administrative Data

Information on/for the operations of large organizations:

- Since Mesopotamians in 7500 BC – clay tablets for bookkeeping
- E.g. Land deeds, government proceedings, tax rolls, records of births/deaths/marriages.

From hard copy data: unreliable, errors, data loss, sharing …

To digitization in 20th and 21st centuries with growing compute …

Modern Administrative Data

- Accounting records
- Inventory tracking
- Customer service
- Human resources
- Tax records
- Health records
- Insurance claims

Goal of administrative data is to keep detailed records for future reference, so this data is inherently “big”:

- Both in the number of entries recorded, and the activities recorded

(Dis)Advantages of Administrative Data

Typical **advantages** of admin data relative to survey data:

- Larger sample sizes
- More precise measurement
- Less bias in responses
- Larger number of variables

Typical **disadvantages** of admin data relative to survey data:

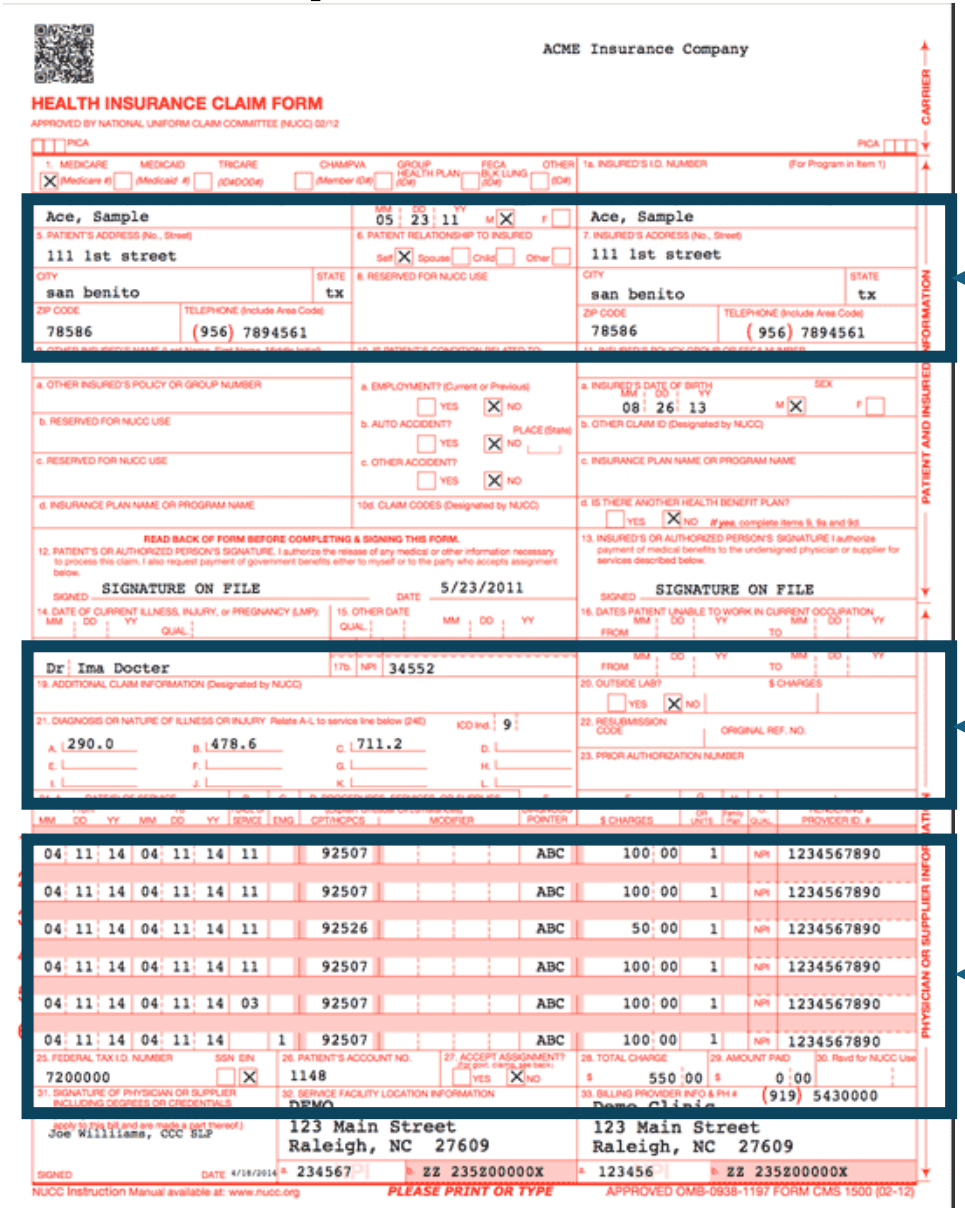
- Not be crafted for analyst's intended purpose
- Unlikely to capture important measures of interest

Administrative Data in Healthcare

The healthcare system produces two unique, massive forms of admin data used by researchers, managers, and policymakers:

- **Health Insurance Claims**
- **Electronic Health Records (EHR)**

Sample Health Insurance Claim Form



ACME Insurance Company

HEALTH INSURANCE CLAIM FORM

APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE (NUCC) 02/12

PICA

1. MEDICARE (X) (Medicare #) | MEDICAID ( ) (Medicaid #) | TRICARE ( ) (ID#/DOD#) | CHAMPVA ( ) (Member ID#) | GROUP HEALTH PLAN ( ) (ID#) | FECA BLK LUNG ( ) (ID#) | OTHER ( ) (ID#)

1a. INSURED'S I.D. NUMBER (For Program in Item 1)

2. PATIENT'S NAME: Ace, Sample
3. PATIENT'S BIRTH DATE: 05 23 11 M (X) F ( )
4. INSURED'S NAME: Ace, Sample
5. PATIENT'S ADDRESS: 111 1st street
6. PATIENT RELATIONSHIP TO INSURED: Self (X) Spouse ( ) Child ( ) Other ( )
7. INSURED'S ADDRESS: 111 1st street
CITY: san benito | STATE: tx
8. RESERVED FOR NUCC USE
CITY: san benito | STATE: tx
ZIP CODE: 78586 | TELEPHONE: (956) 7894561
ZIP CODE: 78586 | TELEPHONE: (956) 7894561

9. OTHER INSURED'S POLICY OR GROUP NUMBER
a. EMPLOYMENT? (Current or Previous) YES ( ) NO (X)
a. INSURED'S DATE OF BIRTH: 08 26 13 SEX: M (X) F ( )
b. RESERVED FOR NUCC USE
b. AUTO ACCIDENT? YES ( ) NO (X) PLACE (State)
b. OTHER CLAIM ID (Designated by NUCC)
c. RESERVED FOR NUCC USE
c. OTHER ACCIDENT? YES ( ) NO (X)
c. INSURANCE PLAN NAME OR PROGRAM NAME
d. INSURANCE PLAN NAME OR PROGRAM NAME
10d. CLAIM CODES (Designated by NUCC)
d. IS THERE ANOTHER HEALTH BENEFIT PLAN? YES ( ) NO (X) If yes, complete items 9, 9a and 9d.

READ BACK OF FORM BEFORE COMPLETING & SIGNING THIS FORM.

12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE. I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below.
SIGNED: SIGNATURE ON FILE | DATE: 5/23/2011

13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below.
SIGNED: SIGNATURE ON FILE

14. DATE OF CURRENT ILLNESS, INJURY, or PREGNANCY (LMP): MM DD YY | QUAL
15. OTHER DATE: QUAL | MM DD YY
16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION: MM DD YY FROM | MM DD YY TO

17. NAME OF REFERRING PROVIDER OR OTHER SOURCE: Dr Ima Docter
17b. NPI: 34552
18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES: MM DD YY FROM | MM DD YY TO
19. ADDITIONAL CLAIM INFORMATION (Designated by NUCC)
20. OUTSIDE LAB? YES ( ) NO (X) $ CHARGES
21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY. Relate A-L to service line below (24E) ICD Ind. 9
A. 290.0 B. 478.6 C. 711.2 D. E. F. G. H. I. J. K. L.
22. RESUBMISSION CODE | ORIGINAL REF. NO.
23. PRIOR AUTHORIZATION NUMBER

24A DATE(S) OF SERVICE From MM DD YY	To MM DD YY	PLACE OF SERVICE	EMG	24B CPT/HCPCS	MODIFIER	24C DIAGNOSIS POINTER	24D $ CHARGES	24E DAYS OR UNITS	24F Family Plan	24G EMG	24H EPSDT Family Plan	24I ID. QUAL	24J RENDERING PROVIDER ID. #
04 11 14	04 11 14	11		92507		ABC	100 00	1				NPI	1234567890
04 11 14	04 11 14	11		92507		ABC	100 00	1				NPI	1234567890
04 11 14	04 11 14	11		92526		ABC	50 00	1				NPI	1234567890
04 11 14	04 11 14	11		92507		ABC	100 00	1				NPI	1234567890
04 11 14	04 11 14	03		92507		ABC	100 00	1				NPI	1234567890
04 11 14	04 11 14		1	92507		ABC	100 00	1				NPI	1234567890

25. FEDERAL TAX I.D. NUMBER: 7200000 SSN ( ) EIN (X)
26. PATIENT'S ACCOUNT NO.: 1148
27. ACCEPT ASSIGNMENT? (For govt. claims, see back) YES ( ) NO (X)
28. TOTAL CHARGE: $ 550 00
29. AMOUNT PAID: $ 0 00
30. Rsvd for NUCC Use
31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS: Joe Williams, CCC SLP
32. SERVICE FACILITY LOCATION INFORMATION: DEMO
123 Main Street
Raleigh, NC 27609
33. BILLING PROVIDER INFO & PH #: Demo Clinic (919) 5430000
123 Main Street
Raleigh, NC 27609

SIGNED | DATE 4/18/2014
a. 234567 b. ZZ 235200000X
a. 123456 b. ZZ 235200000X

NUCC Instruction Manual available at: www.nucc.org
PLEASE PRINT OR TYPE
APPROVED OMB-0938-1197 FORM CMS 1500 (02-12)

CARRIER

PATIENT AND INSURED INFORMATION

PHYSICIAN OR SUPPLIER INFORMATION

Patient Identifiers, DOB, Sex

Provider ID, Diagnosis Codes

Procedure Codes, Date, Provider IDs

Administrative Data in Healthcare

An individual claim can’t tell you much about the patient

- Basic demographics
- Diagnosis codes given by the provider
- Provider IDs, and the services they rendered

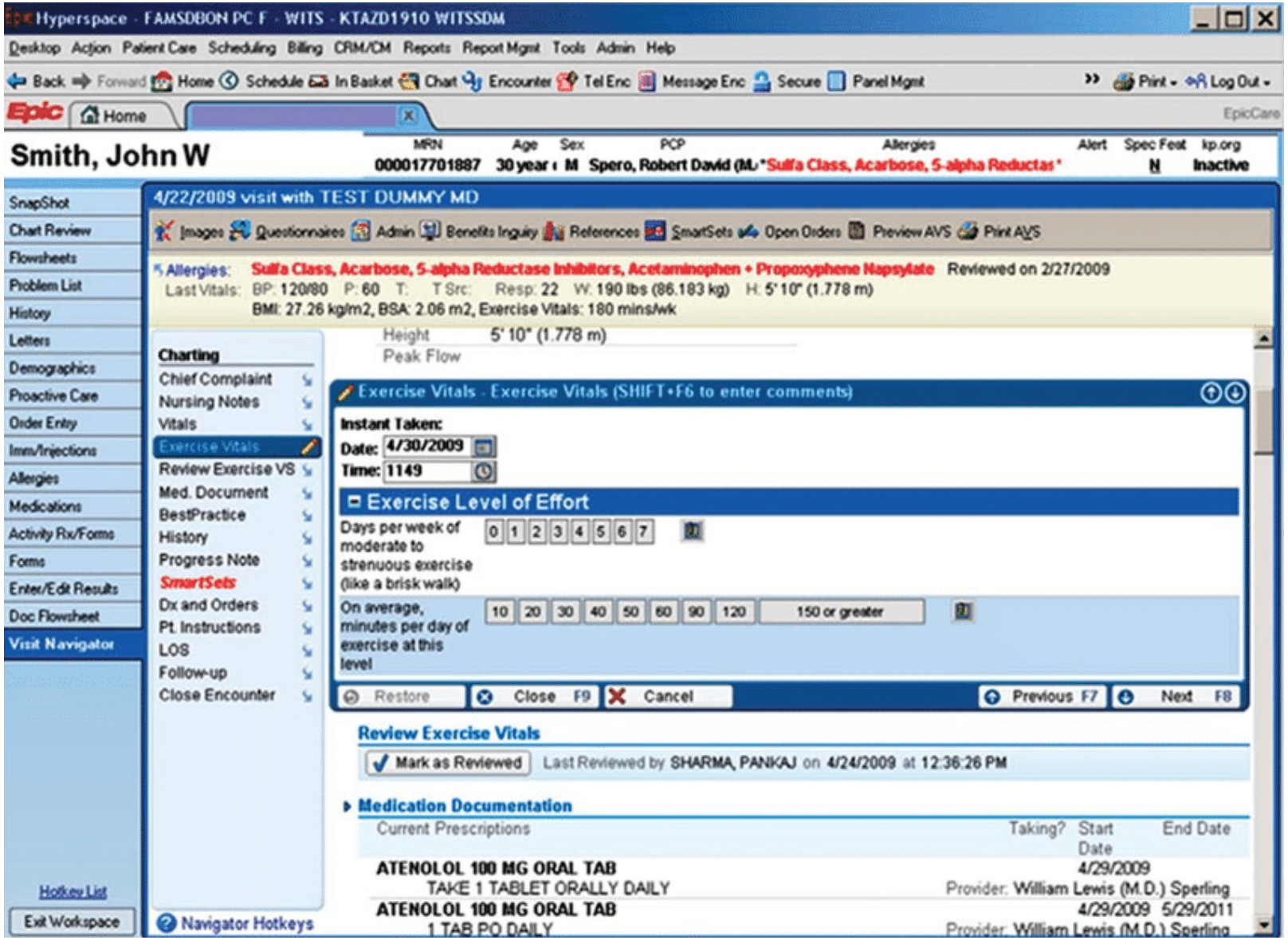
However, observing all claims for an individual over an extended period can give the analyst (you!) sense as to their health status

Electronic Health Records

On the other hand, electronic health records (EHR) can give a much clearer picture of an individual's health

- Detailed date/time data
- Patient vitals (heart rate, blood pressure, BMI, etc.), lab values
- Physician notes, medical history
- Diagnoses, Procedures, imaging (MRI, X-ray, etc.)
- And more…

Electronic Health Records (EHR)



Sallis et al. (2016)

Data Details

Structure, content, and “big” data

UNSTRUCTURED DATA

VS

STRUCTURED DATA

Data: “Structured” vs “Unstructured”

In your courses, it’s possible you’ve only worked with “structured” data – clearly defined and categorized easily

But most real world data is unstructured.

Intuition for “structured” vs. “unstructured”:

“Could this type of data be readily analyzed in a spreadsheet?”

“yes” = structured.

Example: Structured Data

	person_id Scrambled individual identifier	sample_ed Individual residing in a zip code included in the ED study	any_visit_pre_ed Any ED visit, pre-randomization	any_visit_ed Any ED visit in the study period
1	60562	1	0	0
2	51142	1	0	1
3	60314	1	0	1
4	50902	1	0	0
5	70733	1	0	0
6	56758	1	0	0
7	52926	1	1	1
8	4692	1	0	1
9	24115	1	0	1
10	63106	1	0	0
11	49727	1	0	0
12	44409	1	0	0
13	16262	1	0	0
14	60563	1	0	1
15	19345	1	1	0
16	69404	1	1	1
17	52156	1	1	1
18	1614	1	0	1

Healthcare Data: Structured and Unstructured

“Less than 15 percent of health data in EHRs are entered in structured data fields.” (Roski et al., 2014)

Traditionally, structure was a precondition for analysis and retrieval

But “big data” approaches enable the efficient linkage and analysis of unstructured data to answer operational or research questions

Transforming Data

A large part of what folks call “data science” consists of “ETL” (**Extract, Transform, Load)** tasks

- “How do I extract data of interest in its current (often raw, unstructured) form, and transform it so that it can be more easily used?”

Example: suppose you’d like to use the unstructured text from physicians’ EHR notes to better predict future adverse health outcomes for patients. *How might you do this?*

Example: Structuring Physician Notes

Perhaps you know a keyword that would be helpful to identify:

- Generate flag (i.e., “dummy”) variables for each patient encounter that indicate whether the physician used that word in her notes.

Maybe not and they all could be important?

- Generate a flag for each word/phrase in the physician notes?
- … thousands of variables (think: Pandas Dataframe columns)!

Data Dimensions: “Long” vs. “Wide”

cardiac	pain	stroke
1	0	0
0	1	0
0	1	0
0	0	0
1	0	0
1	0	0
0	1	1
0	0	0
0	1	0
0	1	0
0	0	0
1	1	0
0	0	0

N

cardiac	pain	stroke	male	leg	head	sharp	fever	blood
1	0	0	0	0	0	0	0	1
0	1	0	1	1	0	1	0	0
0	1	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	1
0	1	1	1	0	1	1	0	0

K

N: # of observations in dataset (rows)
K: # of variables in dataset (columns)

- When K is large: “wide” data
- When N is large: “long” data
- **“wide” and “long” are NOT mutually exclusive terms!**

Defining “Big Data”

“Big Data” can be used to refer to data that is wide or long, but is typically both, and often contains unstructured data fields.

Generally, we’re talking about data that’s so “big” that it cannot be contained in the memory of a typical computer.

However, most “Big Data” analysis tools (like machine learning algorithms) are uniquely suited to handling **wide** data – finding patterns across large combinations of variables.

Big Data: Linking Across Sources

Another important aspect that contributes to the ‘width’ of Big Data is that datasets from separate administrative sources are often “merged” together to address questions.

Examples:

- A health system linking SSA death records to its EHR, enabling it to better measure patient (past and present) outcomes
- Merging IRS income tax records with data from a randomized policy evaluation, to determine its effect on subsequent employment

Real-Time Digitization of Health

The advances in computing capacity have lead to proliferation of interconnected mobile devices (the “internet of things”)

This is not restricted to telecommunication and consumer electronics – consider recent developments in health

- Mobile imaging
- “Wearable” technology (heart rate and blood pressure sensors, glucose monitors, exercise trackers, health apps)

This data is all stored somewhere…

Data Misuse Concerns

Given the sensitive nature of health data, there are concerns regarding the security of personally identifiable information (PII).

It can be surprisingly easy to trace even "de-identified / anonymized" data back to a specific an individual especially with the data is "wide" or the sample is specific

In addition to privacy, also concern that such data could be misused to illegally discriminate against individuals/groups

Questions?