Lecture 11: The Database Becomes the Organization

Case Study: Educational Technology

- **Homework 1** provided a case study on the performance of a (fictionalized) Physics I class that was using an **online learning application**.
- The learning application **recorded the clicks** of the students as they used the system.
- The result was some interesting information that could be quantified about the **students' behaviors** in studying the lessons of the class.

The Database

- The instrutor provided the **students' grades** in a spreadsheet file.
- The **university's Registrar** provided information on the students' prior knowledge.
- The **online application** provided data on the usage of the application over time.

All of the data provided here is simulated with fictional names.

The Grades

```
the.grades <- fread(input = "Course Grades.csv")
datatable(data = the.grades, rownames = FALSE)
```

Show 10 entries Search:

Last Name	First Name	Homework	Midterm	Final	Grade
Fuller	Mersiha	93.5	97	90	93.5
Davis	Cody	99.1	88	89	92.74
Lee	Andrew	98.6	99	100	99.14
Worley	Travis	92.1	86	78	86.04
Lauer	Katheryn	88.5	89	89	88.8
Pham	Gabriel	91.8	92	95	92.82
Yang	William	97.4	97	89	94.76
Lindquist	Taylor	89	99	90	92.3
Dove	Jenna	100	91	85	92.8
Schoneman	Kirsten	96.2	93	83	91.28

Showing 1 to 10 of 237 entries

Previous 1 2 3 4 5 … 24 Next

The Registrar's Data

```
the.registrar <- fread(input = "Registrar Data.csv")
datatable(the.registrar, rownames = FALSE)
```

Show 10 entries   Search:

ID	Last Name	First Name	SAT Math	HS GPA
XTj6dQzT	Manzanares-Scisney	Barbara	660	3.5
5FCtbOcY	Kim	Man	620	3.6
DAOOPnmh	Wheeler	Isiah	580	3.6
HnDCl2Qb	Gutierrez	Kenia	670	3.7
jJajnCYo	Kim	Connor	620	3.5
63YW3iYv	Torres	Denise	700	3.3
UfMr1Q94	Thao	Kimberly	680	3.3
eUTXb6Qy	Skaggs-Godino	Brenda	690	3.6
iROLjDig	Sandoval	Emily	690	3.4
LDCLV1vQ	Sexton	Jose	650	3.7

Showing 1 to 10 of 194 entries   Previous 1 2 3 4 5 … 20 Next

Before Merging

```
the.grades[, .N]
```

```
[1] 237
```

```
the.registrar[, .N]
```

```
[1] 194
```

```
names(the.grades)[names(the.grades) %in% names(the.registrar)]
```

```
[1] "Last Name"  "First Name"
```

Considerations for Merging

- There appear to be **fewer records** from the Registrar.
- Any analysis of the Registrar's fields (**SAT Math and HS GPA**) will have some degree of missing data.
- The only matching columns in the two data sets are the **first and last names** of the students.

Merging the Grades and the Registrar's Data

```r
id.name <- "ID"
first.name <- "First Name"
last.name <- "Last Name"
homework.name <- "Homework"
midterm.name <- "Midterm"
final.name <- "Final"
grade.name <- "Grade"
sat.name <- "SAT Math"
gpa.name <- "HS GPA"

grades_and_registrar <- merge(x = the.grades, y = the.registrar,
    by = c(first.name, last.name), all = TRUE)

grades_and_registrar[, .N]
```

```
[1] 279
```

New Rows

- The combined table has **more rows** than either of its components.
- This was a result of setting **all = TRUE** in the merge. Mismatching cases were included as additional rows.
- Understanding the **mismatching cases** will require some investigation.

A Mismatching Example

```
rows.with.mismatches <- grades_and_registrar[is.na(get(grade.name)) |
    is.na(get(sat.name)), ]
mismatching.case <- rows.with.mismatches[get(last.name) ==
    get(last.name)[1], ]
datatable(data = mismatching.case, rownames = FALSE)
```

Show 10 entries

Search:

First Name	Last Name	Homework	Midterm	Final	Grade	ID	SAT Math	HS GPA
Aaron	Martinez Chacon	85.7	80	95	86.78			
Andres	Martinez Chacon					Ox8QNltk	670	3.7

Showing 1 to 2 of 2 entries

Previous 1 Next

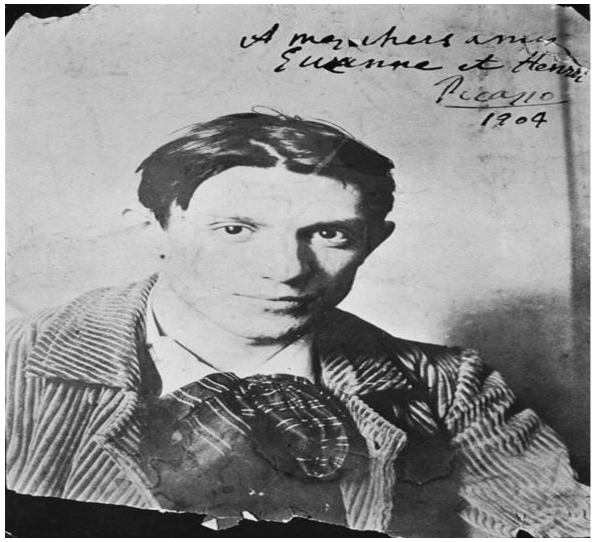
Mismatching Names

- This case looks like it **might be the same student**.
- If so, the student was using a **different first name** in class than what was officially in Registrar's files.
- This turns out to be **quite common**.

Names and Variations

- **First names** have abbreviations, nicknames, and cultural substitutions.
- Some people include their **middle names**, initials, or variations in the spelling.
- Last names can be hyphenated, include a suffix, or be placed ahead of a first name.

Diego José Francisco



Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y **Picasso**.

https://en.wikipedia.org/wiki/Pablo_Picasso#/media/File:Pablo_Picasso,_1904,_Paris,_photograph_b

Duplicates are Indubitable in Databases

- Any small variation in a case can lead to a mismatched records.
- A **duplicate** in a database is any record that should be linked to one account but generates a second account.
- If the **account** is a student, then it will appear as if two or more students all have incomplete records.

Identifying Duplicates

- Matching on **multiple criteria**: name, address, phone number, etc.
- Some degree of inspection and **logical deduction** can often help.
- Using **approximate matching** with the **agrep** function:

```
possible.match.indices <- agrep(pattern = "Aaron Martinez",
    x = grades_and_registrar[, sprintf("%s %s", get(first.name),
        get(last.name))], max.distance = 0.3)
```

The Possible Matches

```r
datatable(data = grades_and_registrar[possible.match.indices,
    ], rownames = FALSE)
```

Show 10 entries

Search:

First Name	Last Name	Homework	Midterm	Final	Grade	ID	SAT Math	HS GPA
Aaron	Martinez Chacon	85.7	80	95	86.78			
Andres	Martinez Chacon					Ox8QNltk	670	3.7
Brandon	Martinez Mcgraw	88.7	86	89	87.98	UZU6aqZO	680	3.6

Showing 1 to 3 of 3 entries

Previous 1 Next

In this case, the compound last name (**Martinez Chacon**) plus the complementary records might lead us to conclude that **Aaron and Andres** are the same person.

More Complex Duplicates

- Two distinct people have the **same name**.
- A prior patient is referred to a medical practice through an electronic record, but the **address and insurance have changed**.
- A patient's **name has changed** since the last visit.

Dual Identities

- Sometimes a database include **deliberate duplicates**.
- Many users of social media have **multiple accounts** and **multiple devices** for any single account.
- Other users create a new account when **rejoining** the service after a period of time away.

Multiple Journeys

- When users return after an extended absence, it raises questions for how to analyze their records.

- **One journey**: The user's entire history is part of one coherent chain of events, even if that includes extended absences.

- **Multiple journeys**: Each return after an extended absence is the beginning of a new round of engagement with the application. Different journeys should be treated separately for analytical purposes.

Linking Devices and Identities

- Technical applications have many ways to **identify and link** multiple devices and accounts to a single user.
- This is especially useful for tracking **long-term usage** and **lifetime value** models.
- It also means that creating a new account **does not truly create a fresh start.**

Creating Better Links

- Much of the trouble we found with merging **might have been avoided**.
- The **Course Grades** did not include the student's ID. This would have provided a clear link to the Registrar's data and to the Application's records.
- Unfortunately, it's **not uncommon** to see suboptimal designs. Even after we identified the problem and a good solution, the **same issue** occurred again the next semester!

Imperfect Links

- Having a **unique identifier** is not always sufficient.
- For instance, an **email address** is likely unique to an individual student, but many students have **multiple addresses** or aliases.
- Names, addresses, telephone numbers, titles: many features are **good but imperfect** identifiers.

A Gold Standard for Identifiers

- A good database will attach a **unique and unambiguous** identifier to every record that pertains to any related object.

- Usually a long string of randomly generated numbers and characters is preferred.

- Unfortunately, even some large providers can have trouble generating good identifiers. Some systems I've used will create identifiers that are only unique in a **case-sensitive** sense.

- Other programs (e.g. Excel) cannot easily distinguish between case-sensitive identifers like **ABCDE** and **abcde**.

Attribution

- A soundly designed database will include **clear links** to every relevant piece of information.

- Clear attribution in every table will enable us to track which users joined from **specific advertising campaigns**, paid for subscriptions through **certain promotions**, etc.

- With good attribution, the records in the database can be easily used for any purpose (billing, analysis, new designs).

De-Duplication

- Sound engineering designs can reduce – **but not eliminate** – the issues of duplicates.
- Because new records are constantly being created, **de-duplication requires ongoing efforts**.
- Investing in a database also requires investments in **quality assurance**.

The Learning Application's Data

- Let's take a look at the data related to the students' usage of the online application:

```
the.clicks <- fread(input = "Application Clicks.csv")
datatable(data = the.clicks[1:100, ], rownames = FALSE)
```

Show 10 entries Search:

ID	Subject	Date
y8PdLaNs	Mechanics	2018-01-22T05:02:43Z
QODfctPn	Mechanics	2018-01-22T05:09:00Z
Cbyw3agj	Mechanics	2018-01-22T05:09:04Z
N1NS6ODI	Mechanics	2018-01-22T05:09:12Z
5FCtbOcY	Mechanics	2018-01-22T05:15:02Z
aU3i4Hpq	Mechanics	2018-01-22T05:16:41Z
CVLnR2qV	Mechanics	2018-01-22T05:17:28Z
Za3W9QY3	Mechanics	2018-01-22T05:18:45Z
BsaUI01G	Mechanics	2018-01-22T05:23:51Z
vjbu9RNS	Mechanics	2018-01-22T05:29:37Z

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

Some Exploration

```
subject.name <- "Subject"
mechanics.value <- "Mechanics"
date.name <- "Date"
the.clicks[, .N]
```

```
[1] 118804
```

```
the.clicks[, length(unique(get(id.name)))]
```

```
[1] 162
```

```
the.clicks[, unique(get(subject.name))]
```

```
[1] "Mechanics"    "Momentum"     "Gravity"      "Electricity" "Magnetism"
[6] "Relativity"
```

```
the.clicks[, .(Min_Date = min(get(date.name)), Max_Date = max(get(date.name)))]
```

```
               Min_Date             Max_Date
1: 2018-01-22T05:02:43Z 2018-05-08T03:57:19Z
```

Some Early Observations

- Not all of the students used the system.
- The records include clicks on a variety of subjects within Physics I.
- The data collected roughly cover the clicks over the course of a typical spring semester.

One Student's Clicks

- As an example, let's consider the clicks undertaken by **one student**:

```
setorderv(x = the.clicks, cols = c(id.name, subject.name, date.name), order = 1)
calendar.date.name <- "Calendar Date"
the.id <- the.clicks[1, get(id.name)]
the.clicks[, eval(calendar.date.name) := as.Date(get(date.name))]
daily.counts.one.student <- the.clicks[get(id.name) == the.id, .N, keyby = c(id.name, subject.name, calendar.date.name)]
datatable(data = daily.counts.one.student, rownames = FALSE)
```

Show 10 entries

Search:

ID	Subject	Calendar Date	N
101uRffy	Electricity	2018-03-01	1
101uRffy	Electricity	2018-03-02	1
101uRffy	Electricity	2018-03-07	1
101uRffy	Electricity	2018-03-11	1
101uRffy	Electricity	2018-03-13	1
101uRffy	Electricity	2018-03-14	1
101uRffy	Electricity	2018-03-15	2
101uRffy	Electricity	2018-03-17	2
101uRffy	Electricity	2018-03-21	1
101uRffy	Electricity	2018-03-22	1

Showing 1 to 10 of 47 entries

Previous 1 2 3 4 5 Next

Sporadic Usage

- This student **did not use the site every day**.
- Some lessons (e.g. **Mechanics**) include almost no activity, while other subjects had more regular usage.
- The usage for different subjects occurred in **different parts of the semester**. This is consistent with our expectations for the online application.

What Isn't Counted

- Counting with **.N** is a fast way to count the **records that exist** in a data.table.
- However, with sporadic usage, some students have **large gaps** in between their sessions with the application.
- We are not counting the **days with zero clicks**.

Do Zeros Count?

- There is room for a **range of opinions**:
- **Of course!** A student's **average daily usage** should factor in the zeros.
- **Of course not!** A student's interaction with the application is better measured as the average number of clicks **when the student signs in at all**.

Preparing for Either Case

- **Zeros In**: We can either **count the total clicks** over a time period or **insert the zeros** into the table.
- **Zeros Out**: Performing the **counting is easy**. However, **data visualizations** such as bar graphs would **not include the full range of dates**.
- Even if you opt for Zeros Out, it can still help to place the zeros into the table. They can always be excluded with a filtering step later.

Counting with Zeros

```
category.counts <- function(dat, by.names, count.name = NA,
    include.zeros = TRUE) {
    require(data.table)
    dat <- setDT(x = dat)

    if (is.na(by.names[1])) {
        measured.counts <- dat[, .N]
    }
    if (!is.na(by.names[1])) {
        measured.counts <- dat[, .N, by = by.names]
    }
    if (is.na(count.name)) {
        count.name <- "N"
    }
    if (!is.na(count.name)) {
        setnames(x = measured.counts, old = "N", new = count.name)
    }

    if (include.zeros == TRUE & !is.na(by.names[1])) {
        the.unique.values <- list()
        for (i in 1:length(by.names)) {
            the.unique.values[[i]] <- dat[, unique(get(by.names[i]))]
        }
        unmeasured.counts <- setDT(expand.grid(the.unique.values))
        setnames(x = unmeasured.counts, old = names(unmeasured.counts),
            new = by.names)
        unmeasured.counts[, `:=`(eval(count.name), 0)]
    } else {
        unmeasured.counts <- data.table()
    }

    all.counts <- rbindlist(l = list(measured.counts, unmeasured.counts),
        fill = TRUE)
    the.counts <- all.counts[, .SD[1], by = by.names]

    setorderv(x = the.counts, cols = by.names, order = 1)
    return(the.counts)
}
```

Updated Counts

```
the.counts.with.zeros <- category.counts(dat = the.clicks,
    by.names = c(id.name, subject.name, calendar.date.name),
    include.zeros = TRUE)

datatable(the.counts.with.zeros[get(id.name) == the.id &
    get(subject.name) == mechanics.value, ], rownames = FALSE)
```

Show 10 entries

Search:

ID	Subject	Calendar Date	N
101uRffy	Mechanics	2018-01-22	0
101uRffy	Mechanics	2018-01-23	0
101uRffy	Mechanics	2018-01-24	0
101uRffy	Mechanics	2018-01-25	0
101uRffy	Mechanics	2018-01-26	0
101uRffy	Mechanics	2018-01-27	0
101uRffy	Mechanics	2018-01-28	0
101uRffy	Mechanics	2018-01-29	0
101uRffy	Mechanics	2018-01-30	0
101uRffy	Mechanics	2018-01-31	0

Showing 1 to 10 of 107 entries

Previous 1 2 3 4 5 … 11 Next

What We Can Investigate

Now that we have a database with detailed information about the students' studying behaviors, what can we answer?

- **Patterns of Usage**
- Effectiveness of the system in **increasing grades**.
- Areas for **improvement**.

Let's take a look at each of these questions.

Patterns of Usage

- Who **uses the app**, and who does not?
- With measures of **prior knowledge** (e.g. earlier course grades, SAT scores, etc.), can we see any **patterns with usage**?
- We ultimately chose to partition the students into **subgroups based on their prior knowledge**, with values of Low, Medium, and High.

What We Found: Patterns of Usage

- High performers were the **least likely** to use the app. However, those who did had **high usage**.
- **Medium performers** tended to have more moderate usage. Nearly everyone in this group used the app, but not necessarily at a high volume.
- Low performers had the **highest usage** of any group.

What We Found: Variation in Usage by Subject

- **More challenging** subjects tended to correspond to **higher usage** for the class and in the groups.
- Some subjects had **significantly more content** than others, and the period of time was also not uniform. This made closer comparisons of the subjects more challenging.
- The instructors were considering different modes of using the system, such as A-B testing **mandatory versus voluntarily usage**. Because this system was used as a **voluntary pilot program**, we could not say how much the students might use the system if they were required to do so.

Effectiveness of the System

- We could examine the effect of **using the system** on the students' grades (on quizzes, exams, and overall).
- We could also investigate the impact of **increased usage** on these outcomes.
- This work could be performed in a **linear regression model** or also with **t-tests** within each subgroup.
- However, this was **not performed in a randomized controlled trial**.

Effectiveness: What We Found

- **Using the system** had a positive impact, although this may have been a **selection bias** in terms of who chose to use it.

- **Most of the effect** of using the system came from those with the **lowest prior knowledge**. They improved enough to pass the class, but did not turn into the highest performers overnight.

- Among those who used the system, **increased usage** was not especially helpful. Most students seemed to use the system as much as they needed to, and extra usage was often **negatively selected** among those who were struggling the most. Our results here were inconclusive.

Areas for Improvement

- Creating **mandatory assignments** on the application would provide a boost to every low-performing student.
- **Adding more content** to some subjects would give us a better sense of how they impact the outcomes.
- **Finding avenues to help the medium-level students** was considered a priority. To do this, we investigated more granular successes on specific topics to look for clues on how to help these students improve.

General Conclusions:

- The system **helps at-risk students** to study and pass the class.
- **B-level** students had difficulties becoming A-level students, and their usage of the system only led to small improvements.
- **A-level** students were doing just fine with or without the system, but some of them found the application useful.

Your Work Goes Beyond the Data

- As a data scientist, you can help an organization better utilize information to understand problems and create solutions.
- No matter how skillfully you analyze the data, your work is only valuable **if improvements are made** as a result of it.
- At some point, **executing the strategy** becomes the priority.

Beyond Programming

- For the remainder of the class, we will be stepping back from data sets and programming code.
- Instead, we'll focus on how you can **work with an organization** to **create improved processes and results**.
- Your technical skills will help you **identify problems**, **devise solutions**, and **take leadership** over the changes you make.

Goals for the Organization's New Data Scientist

Within a relatively short period of time on a new project, you should be able to:

- **Understand** most of what is measured in the organization's existing databases.
- **Identify the opportunities** to use this information to help the organization.
- **Have foresight about the challenges** that may prevent you from achieving these results.

Assessing the Information

- Getting **set up with the technical system** can take some time.
- **Exploration of the data** can be a **meandering process** that may or may not give you a full appreciation for what you're working with.
- Your knowledge will be enhanced by getting started on some project – any project – that starts to answer questions.

A Monitoring Report

- For the **social media company**, I began by putting together a series of **monitoring reports**.
- Each report would track an important metric – **daily active users**, the **volume of clicks** on a specific page, or the **number of paid subscribers**.
- These reports would also look at these features in **different markets**, track the results **over time**, and look more closely at **associated A-B tests** to improve the results.

Scorecards

- In Lecture 4 (Panel Data), we discussed a **diabetes intervention** for newly diagnosed patients.
- The program presented me with its **scorecard** of metrics:

Show 10 entries Search:

Metric	Baseline	6 Months	P Value
Takes Medication	57%	72%	< 0.001
Regularly Checks Blood Sugar	64%	71%	< 0.001
Smoker	22%	18%	< 0.001
A1C	8.3	8.2	0.32
Weight	245	241	0.04
Daily Physical Activity	10	15	< 0.001
Recent Hospitalizations	16%	12%	< 0.001

Showing 1 to 7 of 7 entries

Previous 1 Next

Initial Reactions

- The program is making improvements in **most of the metrics**.
- **A1C Scores** tend to lag behind the other measures, but an **improved process** of healthy behaviors should lead to **improved results** over time.
- For a modest intervention, this looked like a reasonably **successful program**.

Questions Arise

With an opportunity to examine the database, I also had a chance to re-evaluate **what we measured** and how to **present the results**. My questions focused on:

- Were we **collecting the right data** to answer the question?
- Were we **analyzing the data** in an appropriate manner?
- How might we **change the program** to make improvements?

Going through each of these metrics turned into a **more expansive project** than we initially suspected.

Medication Adherence

- Not every patient was **prescribed** the same treatments.
- The results were **not split out by medication**. Perhaps **blood pressure medications** had a different rate of adherence than **diabetes medications**.
- Data were gathered with a **simple yes/no** question about whether the patients regularly took their medicines.

Medication-Specific Measures

- Our group eventually **redesigned the survey** to track adherence for **each prescribed medication**.
- We **updated the metrics** so that we could track adherence on medicines, **individually or in categories**.
- We also **encouraged the patients** to **obtain prescriptions** on medicines that it looked like they should be taking.

Measuring Adherence

- Asking a **yes/no question** at baseline and 6 months likely ignores the **variation in behaviors** over time.
- We ultimately switched to a medically validated **adherence questionnaire** that could better assess the **level of adherence** instead of using an absolute measure.
- This would provide better information **at baseline and 6 months**.

Additional Follow-Up

- We also decided that **improving adherence requires more regular follow-up** with the patients.
- The program was modified to **contact the patients more frequently** and to obtain data on medication adherence at each interaction.
- Any range of options – e.g. **patients' logs** or **technological applications with reminders** – were then on the table.

Modifications to the Database – Adherence

To accommmodate these changes to the intervention, we had to make corresponding changes to the database:

- **New formats for the tables** allowed us to track **multiple medications** at **variable time points**.
- The **interface for the questionnaire** and entering the **medical case notes** had to be updated.
- The **associated tasks of outreach** also had to be modified in the database. This ensured that our team would **contact the patients at the right times**.

Checking Blood Sugar

Our investigation here turned out to be **more straightforward**:

- We were **satisfied with the questions** that we asked to **assess the patient's activity**.
- We did opt to **increase the frequency** at which we checked in with the patients. These questions were added to the **follow-up calls** related to medication adherence.
- With relatively **small modifications to the database**, we were set up to record this information.

Smoking

- The original questionnaire included **multiple questions** related to smoking: whether the **patient is a smoker**, whether **other members of the household were smokers**, the **number of cigarettes smoked per day**, etc.

- However, **the scorecard only tracked individual smoking** as a **yes/no** question.

- Some of the patients with Type II Diabetes **were children**. Some of these children were smokers.

Additional Metrics for Smoking

- The **Percentage Regularly Exposed to Household Smoke**.
- The **Average Daily Number of Cigarettes Smoked**.
- Segmenting **all of the smoking metrics** for **children** and for **adults**.

Greater Specificity, Greater Insight

- The **expanded range of metrics** allowed us to **report on a wider range of circumstances**.
- We effectively created a **menu of options** for highlighting **different aspects of the program**.
- Moreover, we found ways to identify **specific groups** that were in need of **enhanced support**.

But Did We Solve the Problem?

- Smoking rates and volumes had **improved in a statistically significant way**.
- At the same time, the **overall rate of individual smoking** was **still 18% of the patients** at 6 months.
- While the progress was good, it was also clear that **too many patients** were not changing their behaviors.

Changing the Program

- Smoking is one of the **largest risk factors** for adverse events among patients with diabetes.

- In some sense, all of the program's good work on diets, exercise, and medication adherence **would not amount to much** if the patients continued to smoke cigarettes.

- We ultimately decided to **place a greater emphasis** on smoking cessation, with **referrals to secondary programs** and **additional outreach** to help smokers quit.

A1C Scores

- This measure had shown **very little improvement** in the original results.
- However, it turned out that **most of the patients** had not **received a new laboratory test of A1C** during their follow-up period.
- This is **not that surprising**. Most patients with diabetes only have the test 2-4 times per year.

A Small Sample Size

- The comparison was only performed on patients with **both a baseline and a 6-month A1C score**.
- This turned out to be only a modest **fraction of the cohort**.
- Moreover, there may be a **selection bias** in terms of who takes the test and who does not. The results of the measured comparison **may not represent** the results of the broader cohort.

What is a 6-Month Measure, Anyway?

- It is very rare to obtain follow-up information at **exactly 6 months** after the baseline.

- We therefore have to decide what a 6-month measure **really means**.

- This could be defined **in a number of ways**:

  ++ The reading obtained **closest to 6 months** within a time frame (e.g. plus or minus 30) days.

  ++ The first reading obtained **after 6 months has passed**.

  ++ The **average** of all readings **within a window of time** (e.g. plus or minus 30 days).

Loss of Follow-Up

- Measures over a longer period of time will have **fewer patients** due to **administrative censoring** or a **loss of follow-up**.

- Even many of the patients who **remained in the program** were effectively lost to follow-up on this measure.

- **Long-term effectiveness** will necessarily have to be balanced with obtaining a **reasonable sample size** to answer the question.

Changes to the Program

- We modified the intervention to **emphasize obtaining A1C scores**.
- However, this follow-up was **less frequent** than our new protocols for follow-up on medication adherence and blood sugars.
- This step wasn't only for the purpose of data collection. Instead, it was more in line with the medical guidelines for helping the patients to manage their conditions.

From Questionnaire to Interactive Support

- All of these changes shifted the focus of the program.
- The intervention evolved from **initial classes** into a program of **regular support**.
- The follow-up data became more than a questionnaire. Instead, it was used to provide ongoing guidance to the patients.

Weight and Daily Physical Activity

- After some investigation, we remained satisfied with our manner of gathering the data.
- The **frequency of the measurements** increased, placing these measurements as part of the regular follow-up on medication adherence and blood sugar.
- We also aimed for **more ambitious targets** on weight loss and physical activity.

Recent Hospitalizations

- The questionnaire asked the patients: **Have you had any hospitalizations in the past 30 days?**
- The reported numbers were then **annualized** to be the average number of hospitalizations per patient per year.

Questions Designed for Research

- For hospitalizations, the 6-month outcomes would provide a simple benchmark for writing an **academic article**.

- The one-month **look-back period** was mirrored in the **baseline** and the **6-month** outcomes.

- This would give us a clear and simple picture of the **general effect** of the program.

Limitations of the Hospitalization Question

- The **first 5 months** of follow-up would not necessarily be included in the data for the 6-month outcomes.
- The question only asked **whether a patient was hospitalized** rather than **how many times** this had occurred.
- There was no information about the **number of days** that a patient stayed in the hospital.
- The **dates, locations, and causes** of the hospitalizations were also not recorded.

Creating a New Table

- Based on these observations, we began recording **more detailed information** about each reported hospitalization over **the entire duration** of follow-up.
- This table would allow us to build **survival curves** for keeping the patients out of the hospital.
- With information about the number of visits and the length of the stays, we could also perform **economic calculations of the cost** of hospitalizations.

The Granularity of the Information

- The original version of how we recorded hospitalizations was **missing information** and **lacked specificity**.
- Our initial database was **designed for one question**, but it did not have the foresight to consider **additional uses** of that information.
- We could only fully answer the other questions by **collecting more granular information**.

Granular Information in Social Media

- A social media app offered **paid subscriptions** that allowed access to some of its features.
- Information about the subscriptions was **recorded in a table** on the database.
- We wanted to create a **monitoring report** that provided detailed information about the **volume and revenue** of the subscriptions over time.

Subscriptions: What Was Recorded

- Unfortunately, the table in the database only included the following information:

  ++ User ID

  ++ Date of First Subscription

  ++ Date of Most Recent Payment

  ++ Price of Most Recent Payment

  ++ Expiration Date of Most Recent Subscription

Subscriptions: What Was Missing

- The previous table did not provide a **full history** of a user's subscriptions:

  ++ Times Subscribed

  ++ Times Not Subscribed

  ++ Prices Paid

  ++ Discounts Applied

  ++ Gift Subscriptions Sent or Received

- Without this information, we could not answer a variety of questions:

  ++ The **duration** of continuous initial subscriptions

  ++ The **lifetime value**

  ++ The likelihood of **re-subscribing** after a lapse

Subscriptions: A More Granular Approach

- I noticed this issue, but the team had already made an appropriate change a few months earlier.
- The new table recorded:

  ++ The dates of each new payment

  ++ The beginning and expiration dates associated with the payment.

  ++ The price paid

  ++ Discounts and gifts

After the Change is Implemented

- With a new design, it will become possible to **prospectively evaluate** questions based on the information you're collecting.

- This will require **additional time** to enroll more subjects and collect longitudinal outcomes.

- But what about the **old information**? Can anything be done with the less granular records?

Backfilling Tables

- In some cases, you may be able to **estimate** what the old records **would have looked like** if you had recorded them according to the new protocol.
- When this is possible, you may be able to **backfill** the new table with your estimates from the old records.
- In this case, it is important to note in the granular table **which records are backfilled** and which are not.

Backfilling: A Partial Success

- In the case of the **hospitalizations**, we were able to backfill some of the records through **imputation** on the data we later collected, with some **logical deduction** of the records we had, and with a **targeted campaign** to contact some of the previous patients.

- In the case of the **subscriptions**, we felt that the **product and the business had changed** enough that a historical view would not be helpful. For descriptions of the past, we felt that the **revenue data** were sufficient, and any questions about future subscriptions could be **addressed later** after gathering data with the more granular table.

The Volume of Metrics

- Our work **created a broader menu** of quantities to measure.
- However, having more expansive options **is not necessarily a good thing** if the quantity of metrics **detracts from the focus** on the most important indicators.
- **Measuring anything and everything** is no substitute for defining the priorities of your work in terms of the most valuable opportunities.

Better Data

- The original data on hospitalizations was **self-reported** by the patients.
- Later on, we were able to collect some limited data from **electronic medical records**.
- Recording more precise data is likely to lead to greater precision over time. Obtaining **higher quality data** will always be helpful.

Assessing One Analysis Cultivated Multiple Projects

- Because of this work, the program **changed the frequency with which it collects data**.

- The program also **identified new ways to help the patients**.

- We were also better able to demonstrate our **medical effectiveness and economic value** by collecting more detailed data on the most important outcomes.

It Takes a Team

- In the case of these programs (which are all fictionalized versions of some real projects), we were ultimately able to make a variety of improvements that **helped the patients/customers** and **served the business**.

- **Data Science played an important role** in identifying some of the issues, but the **ideas came from many sources** throughout the team.

- Ultimately these improvements only came about because the team was **committed to using data** and willing to consider proactive changes.

The Database Becomes the Organization

- Better designs for the database **enable better investigations**.
- Ultimately, the **goals of the organization** become intertwined with its ability to **collect and analyze** the right information.
- Increasingly, the effectiveness of the organization can be greatly driven by its database systems and the team's approach to using it.

Your Role as a Data Scientist

- Your efforts can play a major role in **creating new initiatives** for the organization.
- This can require developing your **capacity for leadership**.
- It is **no longer sufficient** to merely perform the technical work. As a person who understands the data, the analyses, and the implications for the project, no one is more qualified than you to have a say in how your findings can lead to changes.

Ambiguities are Everywhere

- Much of today's lecture involves **identifying issues** and **defining your own tasks** to resolve them.
- This process is **quite different** than what we as students have been trained to do for much of our lives: solve problems with **clear definitions** and **one sound approach**.
- Your growth as a creative problem solver will require you to become **comfortable with ambiguities, uncertainties, and limitations**.

Opportunities are Also Everywhere

- Your investigations will enable you to uncover the issues and play a role in generating improvements.
- Better yet, your **productivity and multi-faceted skills** will help your organizations understand and resolve these issues much faster and easily than they otherwise might.