

# Iterating Between the Data World and the Real World to Find Answers in Big Data

Hye-Chung Kum, Professor (kum@tamu.edu)  
Population Informatics Lab (<https://pinformatics.org/>)

Department of Health Policy and Management, School of Public Health  
Department of Computer Science and Engineering  
Department of Industrial and Systems Engineering  
The Center for Remote Health Technologies and Systems (CRHTS)  
Texas A&M University



10/16/2020


1

1

## Hye-Chung Kum: Background context

- PhD in computer science (datamining)
- Minor: MSW (policy & management)
- Primary appointment: School of Public Health, HPM (18+ years: CS, SW, HPM, ISEN)
- Joint appointments in Computer Science & Industrial Systems
- Population Informatics Lab, Texas A&M University
  - interdisciplinary: CS, HSR, health informatics, SW, sociology, ELSI
  - PCORI: Privacy Preserving Interactive Record Linkage (PPIRL)
  - NSF: A Benchmark Data Linkage Repository (DLRep)
  - TX HHSC: 1115 Medicaid Waiver Evaluation
  - NC DHHS: Management Assistance

2




## Target audience

- Junior graduate students starting out in data science
- Basic knowledge in
  - Modeling
  - Programming
- Learning objective: Critical thinking about data
  - Identify and communicate the data decisions
    - By learning to think about the data at hand, and the real problem
    - Data scientist can make data decisions or ask to get clarification
  - Understand all the details involved in a real data science research project
  - Describe how to do quality control on data science projects

10/16/2020 3

3



## A student

- “desire to learn SAS..



Before taking this class, I had limited experience with SAS and data science. I knew how to copy and paste basic SAS commands, but I didn’t really understand it, much less how to think about data. That changed throughout the 15 weeks of this class.

... But at some point in the semester, **it finally begins to “click.”** You begin to **understand how to troubleshoot and find new ways of thinking about problems.** After a while, you even start to feel confident in your ability to transform and analyze a dataset. I still have a lot to learn and improve upon, but I think now I have **a foundation to build upon throughout the rest of my career.**

learn how to think about and use data ... begin to **think like a data scientist.”**

10/16/2020 4


4



## Agenda

- What is Data Science ?
  - And what is my role as a data scientist?
- Applications: Case Studies
  - HCC (Liver Cancer) screening: measurement
  - Medicaid Waiver Evaluation: detecting change
  - Privacy Preference: common pitfalls
- Closing Thoughts

5



## What is Data Science?

Primary Methodology in Population Informatics: Data Science (KDD)



6

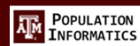
## New Era in Science : Data Science



- **Data** is the new raw material of business: an economic **input almost on par with capital and labor**. (Microsoft's Craig Mundie)
- **Those who can harness the power of data will lead the next century** and drive innovation in commerce, scientific discovery, healthcare, finance, energy, government, and countless other fields.
- Students who learn to be a data scientist will be in high demand.

7


## Knowledge Discovery & Data mining (KDD) = Data Science

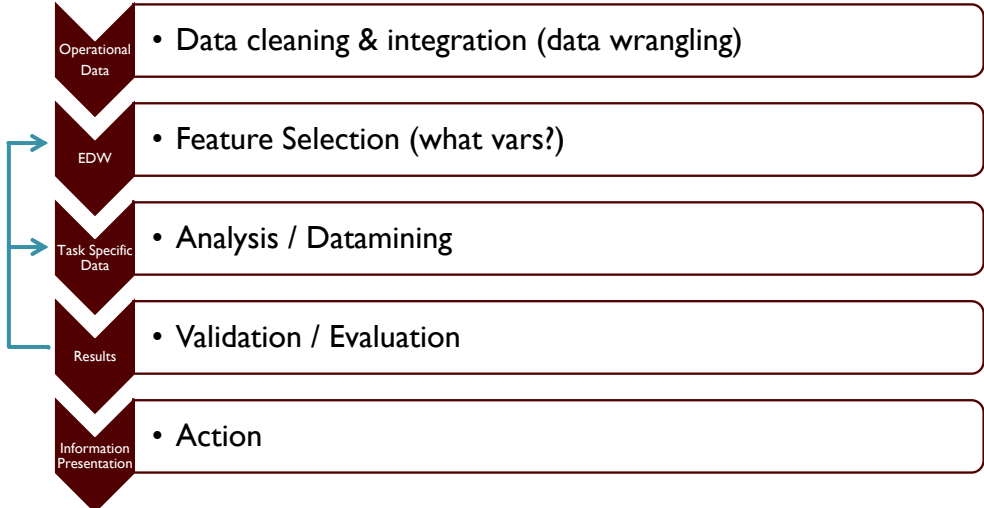


8


## Take Away I: Data Science 101


### KDD Process





9



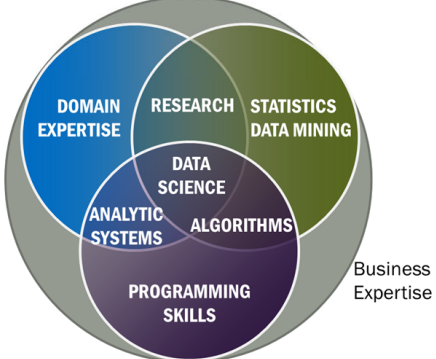


## Data Science Definition (Big Data less consensus)

---

- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.
- A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.

Big Data refers to digital data volume, velocity and/or variety whose management requires scalability across coupled horizontal resources



9/29/13

IEEE BigData Overview October 9 2013

8

10

## Bioinformatics

Apply Data Science to Human Genome Data

Biology

+

11

## Population informatics

Apply Data Science to Social Genome Data

Studies of society (groups of people)

- Social, Behavior, Economic sciences
- Health sciences (population health)

+

Kum, H.C., Krishnamurthy A., Machanavajjhala A., and Ahalt S. Social Genome: Putting Big Data to Work for Population Informatics. *IEEE Computer Special Outlook Issue*. pp 56-63. Jan 2014

12

ATM POPULATION INFORMATICS

## The Big Data Problem - Nutshelled

Michael Franklin (UC Berkley)

Something's gotta give:

13

ATM POPULATION INFORMATICS

## AMPLab: Integrating Three Key Resources

Michael Franklin (UC Berkley)

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines

- Clusters and Clouds
- Warehouse Scale Computing

People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts

14

## Thomas Davenport: Competing on Analytics

- Skill set for good data scientists
  - IT & Programming skills: Very basic programming concepts in SAS
    - <https://pinformatics.tamhsc.edu/phpm672/>
  - Statistical skills
  - Business skills:
    - Understand pros/cons of decisions & actions
    - Communication skills
    - Excel / PowerPoint
  - **Intense curiosity: the most important skill or trait.** “a desire to go beyond the surface of a problem, find the question at its heart, and distill them into a very clear set of hypothesis that can be tested”

15

## Data Science Team



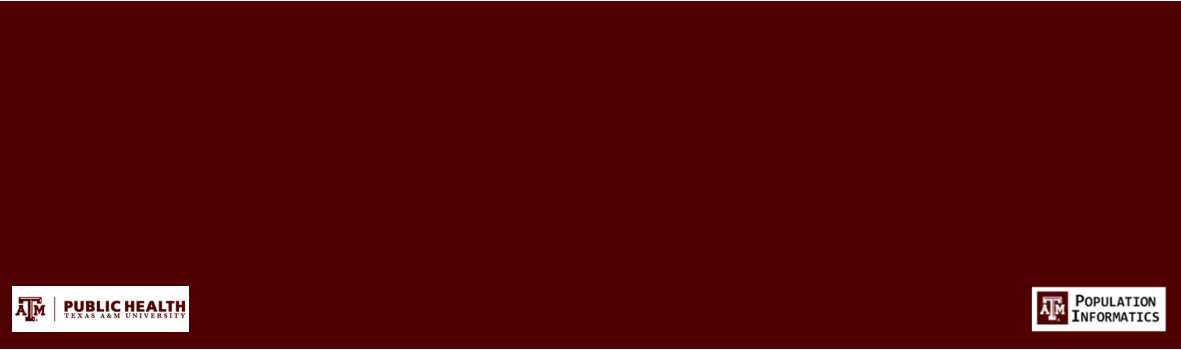
- Data science teams need people with the **skills and curiosity** to ask the big questions (oreilly)
  - **Technical expertise:** the best data scientists typically have deep expertise in some scientific discipline.
  - **Curiosity:** a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
  - **Storytelling:** the ability to use data to tell a story and to be able to communicate it effectively.
  - **Cleverness:** the ability to look at a problem in different, creative ways.
    - Team lead: good questions, good interpretation & implications
  - <http://radar.oreilly.com/2011/09/building-data-science-teams.html>

16



---

# Case Studies




17

---

# Case Studies: Prelude

## Clustering



18

Ice Breaker: Clustering

- cluster data into similar groups

10/16/2020
19

19

Participate

- [https://docs.google.com/document/d/1HCvpvlprmAyiYstyum9KeKdUsnM6\\_7cT3g\\_5-LwBkx0/edit](https://docs.google.com/document/d/1HCvpvlprmAyiYstyum9KeKdUsnM6_7cT3g_5-LwBkx0/edit)

MiniBatchKMeans	AffinityPropagation	MeanShift	SpectralClustering	Ward	DBSCAN
.01s	3.05s	.04s	2.08s	.11s	.35s
.01s	3.49s	.03s	6.60s	.13s	.36s
.01s	3.39s	.03s	.50s	.17s	.37s
.01s	2.93s	.06s	.77s	.11s	.35s

10/16/2020
20

20

## Case Studies I: Measurement

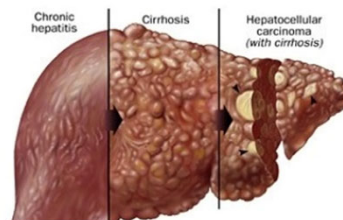
### HCC (Liver Cancer) Screening

21

## Background



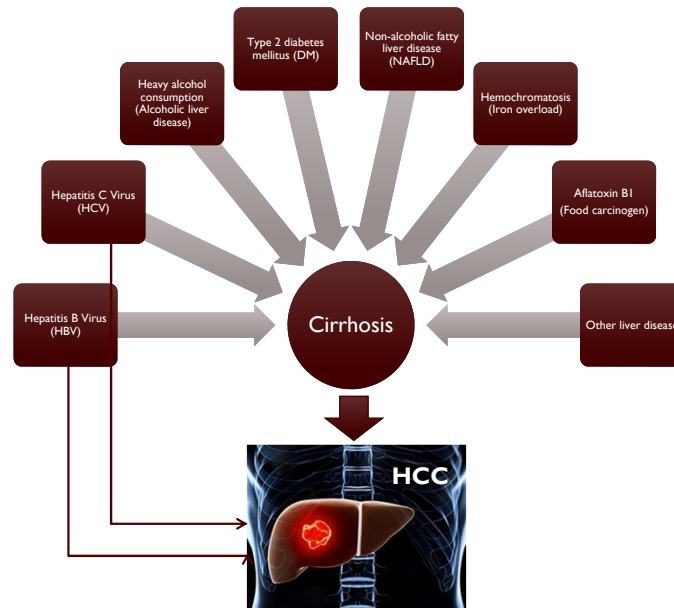
- Hepatocellular carcinoma (HCC) or primary liver cancer
  - 2<sup>nd</sup> leading cause of cancer-related death worldwide
  - 9<sup>th</sup> leading cause of cancer-related death in the U.S.
  - Leading cause of death among patients with cirrhosis (late stage of scarring (fibrosis) of the liver)
  - Projected to surpass breast and colorectal cancer to become the 3<sup>rd</sup> leading cause of cancer-related death by 2030 in the U.S.



22

## Background

- Risk factors for HCC include:



23

## Screening for HCC is very important



- Patients typically diagnosed at advanced stage
  - Asymptomatic
- Prognosis dependent on tumor stage at time of diagnosis
  - Curative treatment options are only available for patients diagnosed at early stage
- Survival for HCC patients is poor (5-year survival <5%)
- The National Comprehensive Cancer Network (NCCN) and American Association for the Study of Liver Diseases (AASLD) in 2018 recommends **screening for HCC every 6 months using ultrasound (US) with or without alpha-fetoprotein (AFP)**

Prior studies in literature found **less than 20%** of patients with cirrhosis receive HCC screening

24

## Objectives



- To design and test two improved alternative approaches to measure HCC screening using administrative data
- To study what patient and provider factors impact HCC screening
- To study the impact of HCC screening on early tumor detection and overall survival

25

## Data & Methods



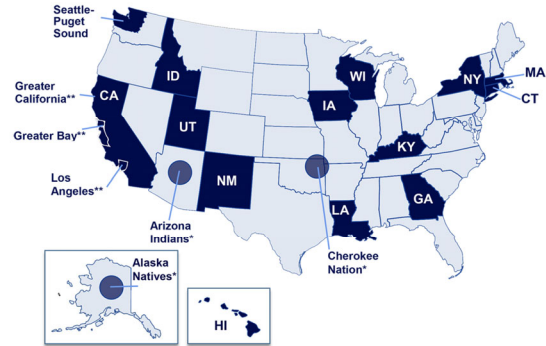
Objectives	1. Measurement of HCC Screening Using Administrative Data	2. Impact of Patient and Provider Factors on HCC Screening	3. Impact of HCC Screening on Early Tumor Detection and Overall Survival
<b>Data Sources</b>	1. Surveillance, Epidemiology and End Results (SEER)-Medicare 2. American Medical Association Master File (AMA)		
<b>Years</b>	HCC Diagnosis Years 2003-2013		
<b>Dependent Variables</b>	N/A – Descriptive characterization study	Proportion of time up-to-date (PUTD) with HCC screening	Early tumor detection (Milan criteria) and survival
<b>Methods</b>	N/A – Descriptive characterization study	2-part model (Tobit and generalized ordered logit for sensitivity analyses)	Logistic regression and Cox Proportional Hazards models
<b>Sample</b>	Main Sample: All HCC Patients= <b>13,714</b> Sub sample analysis: Known Cirrhosis Patients= <b>2,972</b>		
	Excluding: died at start of study All HCC Patients= <b>12,609</b> Known Cirrhosis Patients= <b>2,797</b>		

26

## CMS: SEER-Medicare



- Medicare claims: national data (5% national sample)
  - Denominator File - LDS
  - Standard Analytical Files (Medicare Claims) - LDS
    - Inpatient Data
    - Outpatient Data
    - Skilled Nursing Facility Data
    - Durable Medical Equipment Data (includes chemo)
    - Home Health Data
    - Hospice Data
- SEER: collects and publishes cancer incidence and survival data
  - 15ish states
  - population-based cancer registries covering approximately 34.6 percent of the U.S.
  - data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status
- Linked SEER\*Medicare claims
- Medicaid claims: state dependent (may get from CMS or state)



10/16/2020

27

27

## Sample: STUDY POPULATION

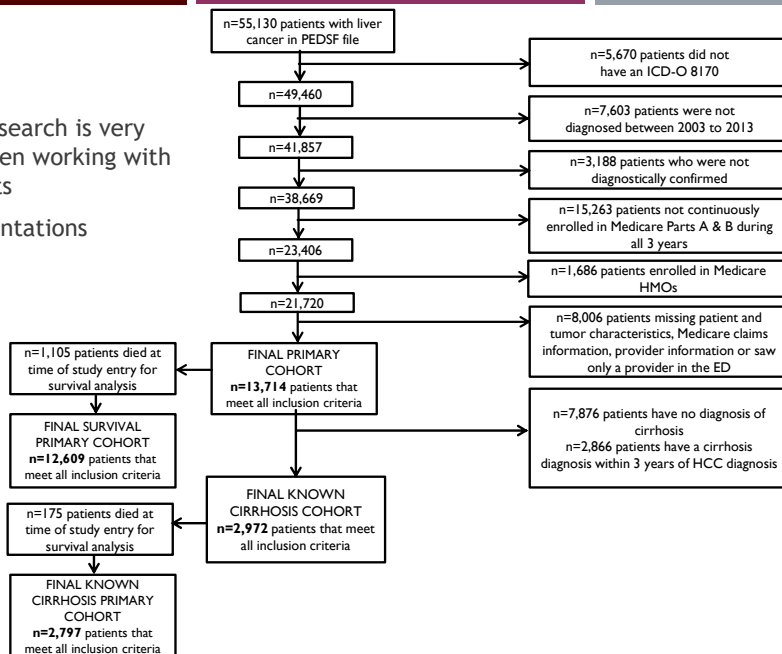


- INCLUSION:
  - All Medicare beneficiaries who have been diagnostically confirmed with HCC (ICD-O 8170) from 2003 to 2013
    - Positive histology, cytology, laboratory test/marker, positive radiology tests
    - Medicare Part A and B enrollment during entire 3 year screening period
- EXCLUSION:
  - Enrollment in Medicare HMOs
    - Approximately 13% of people were enrolled in a Medicare Advantage plan in 2003 and this increased to 28% in 2013
  - Missing patient and tumor characteristics
  - Patients who only saw emergency providers
    - Not reflective of preventative care for HCC
  - Patients who only saw providers 1 month prior to diagnosis
  - Patients who only saw providers with no specialty, practice arrangement or graduation information

28

## Sample

- Replicable Research is very important when working with large data sets
- Good documentations



29

## Descriptive on who (n=13,714)

Variable	Consistent screening* (n=937)	Inconsistent screening** (n=5,768)	No screening (n=7,009)	P-value
Age at HCC diagnosis	69.8 (9.8)	71.7 (9.9)	74.5 (9.2)	<.001
Sex (% male)	583 (62.2)	3,786 (65.6)	4,815 (68.7)	<.001
Race/ethnicity				<.001
Non-Hispanic White	436 (46.5)	3,390 (58.8)	4,624 (66.0)	
Black	83 (8.9)	624 (10.8)	713 (10.2)	
Hispanic	168 (17.9)	864 (15.0)	773 (11.0)	
Asian	177 (19.0)	584 (10.1)	515 (7.4)	
Other	73 (7.8)	306 (5.3)	384 (5.5)	
Metropolitan area (%)	884 (94.3)	5,360 (92.9)	6,419 (91.6)	.001
Census poverty level				.002
0% to <5%	168 (17.9)	1,095 (19.0)	1,406 (20.1)	
5% to 9%	204 (21.8)	1,392 (24.1)	1,683 (24.0)	
10% to 19%	315 (33.6)	1,739 (30.2)	2,240 (32.0)	
20% to 100%	250 (26.7)	1,542 (26.7)	1,680 (24.0)	
Year of HCC diagnosis				<.001
2003	47 (5.0)	358 (6.2)	463 (6.6)	
2004	50 (5.3)	367 (6.4)	492 (7.0)	
2005	47 (5.0)	435 (7.5)	495 (7.1)	
2006	62 (6.6)	452 (7.8)	558 (8.0)	
2007	80 (8.5)	504 (8.7)	607 (8.7)	
2008	66 (7.0)	562 (9.7)	706 (10.1)	
2009	71 (7.6)	607 (10.5)	706 (10.1)	
2010	113 (12.1)	578 (10.1)	711 (10.1)	
2011	107 (11.4)	612 (10.6)	714 (10.2)	
2012	122 (13.0)	690 (12.0)	801 (11.4)	
2013	172 (18.4)	603 (10.5)	756 (10.8)	

Variable	Consistent screening* (n=937)	Inconsistent screening** (n=5,768)	No screening (n=7,009)	P-value
Cirrhosis duration				<.001
No prior diagnosis	117 (12.5)	2,368 (41.1)	5,391 (76.9)	
< 3 years prior to HCC	270 (28.8)	1,820 (31.6)	776 (11.1)	
> 3 years prior to HCC	550 (58.7)	1,580 (27.4)	842 (12.0)	
Liver disease etiology				<.001
Hepatitis B	37 (4.0)	163 (2.8)	124 (1.8)	
Hepatitis C	132 (14.1)	918 (15.9)	848 (12.1)	
Alcohol-related	21 (2.2)	249 (4.3)	218 (3.1)	
Other liver disease	69 (7.4)	565 (9.8)	418 (6.0)	
>1 liver disease	637 (68.0)	2,061 (35.7)	668 (9.5)	
No known liver disease	41 (4.4)	1,812 (31.4)	4,733 (67.5)	
Milan criteria (% yes)	596 (63.6)	2,443 (42.4)	1,772 (25.3)	<.001
Ascites (%)	270 (28.8)	1,011 (17.5)	328 (4.7)	<.001
Hepatic encephalopathy (%)	287 (30.6)	796 (13.8)	235 (3.4)	<.001
NCI comorbidity index				<.001
None	5 (.53)	186 (3.2)	763 (10.9)	
Low (1-2)	85 (9.1)	975 (16.9)	2,100 (30.0)	
Moderate (3-4)	188 (20.1)	1,476 (25.6)	1,891 (27.0)	
High (5+)	659 (70.3)	3,131 (54.3)	2,255 (32.2)	

\*Receipt of ≥1 abdominal ultrasound per calendar year

\*\*Receipt of ≥1 abdominal ultrasound during study period but less than annually

30

30

## Q1: Measurement of HCC Screening Using Administrative Data

- Previous studies in literature defined having HCC screening as:
  - **Consistent screening:** Having had an annual AFP and/or ultrasound test at least 2 of the 3 years prior to HCC diagnosis
  - **Inconsistent screening:** Having had one or more AFP and/or ultrasound tests during the 3 years prior to HCC diagnosis
- This measure is outdated and not rigorous enough
  - Does not align with recommended NCCN and AASLD guidelines
- To propose and compare improved alternative measures for HCC screening

31

## Compare two methods

### Measure 1: categorical

- **Consistent screening:** Having  $\geq 1$  abdominal ultrasound (CPT 76700 and 76705) per calendar year
- **Inconsistent screening:** Having  $\geq 1$  abdominal ultrasound during the study period but less than annually
- **No screening**

### Method 2: continuous

- Construct a **proportion of time up-to-date with screening (PUTD)** measure (Used in pharmacy literature), defined as:
  - **Proportion of the 36-month screening period** in which patients had received screening, with each abdominal ultrasound providing 7 months of screening coverage
    - Example: If a patient received an abdominal ultrasound in January 2010 and then one more in July 2010 during the entire 36-month screening period, this patient was covered for 14/36 months or a proportion of 0.38
  - Some patients may have overlapping abdominal ultrasounds
    - Example: If a patient received an abdominal ultrasound in January 2010 and one more in March 2010 during the entire 36-month screening period, this patient was covered for only 9/36 months or a proportion of 0.25

10/16/2020

32

32



## Sensitivity analysis: Validated HCC screening algorithm With and Without Screening Intent



- To distinguish abdominal ultrasound tests for the purpose of HCC screening
  - Applied algorithm developed by Richardson et al. using a logistic regression model to obtain predicted probability of screening for each abdominal ultrasound claim
 

Log odds of surveillance for US=  
 $-0.9015 + -0.3943 * (\text{abdominal pain}) + -0.7932 * (\text{ascites}) + -0.4394 * (\text{drug dependence})$   
 $+ -1.0723 * (\text{HIV}) + 0.8223 * (\text{AFP test in the last 90 days})$
- Cutoff threshold of  $p=0.38$
- If predicted probability was  $\geq 0.38$  then imputed screening variable as:
  - 1 = Screening receipt
  - 0 = Otherwise

33

## 8 Set of Analysis = 2 samples\*2 methods\*2 measures



	Method 1 (categorical) Broken out over time		Method 2 (continuous) PUDF	
	W/O Intent	W/ Intent	W/O Intent	W/ Intent
Full Sample (13,714)	Results 1.1	Results 1.2	Results 1.3	Results 1.1
Subsample (2,972)	Results 2.1	Results 2.2	Results 2.3	Results 2.1

10/16/2020

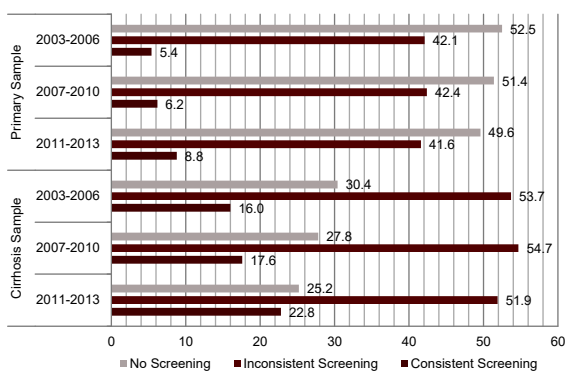
34

34

## Screening Rates: Method 1 - categorical over time Receipt of HCC screening over time



Percent (%) change in HCC screening receipt over time



- Most (51.1%) patients did not receive any screening in the 3 years prior to HCC diagnosis, while 42.1% underwent inconsistent screening, and only 6.8% underwent consistent screening.
- The proportion of patients receiving consistent screening steadily increased over time from 5.4% for patients diagnosed between 2003 and 2006 to 6.2% between 2007 and 2010, and 8.8% between 2011 and 2013. During this time period from 2003-2006 to 2011-2013, the number of patients with no screening decreased from 52.5% to 49.6%.
- Similarly, consistent screening increased from 16.4% to 21.2% over this time period in the subset of patients with known cirrhosis.

10/16/2020

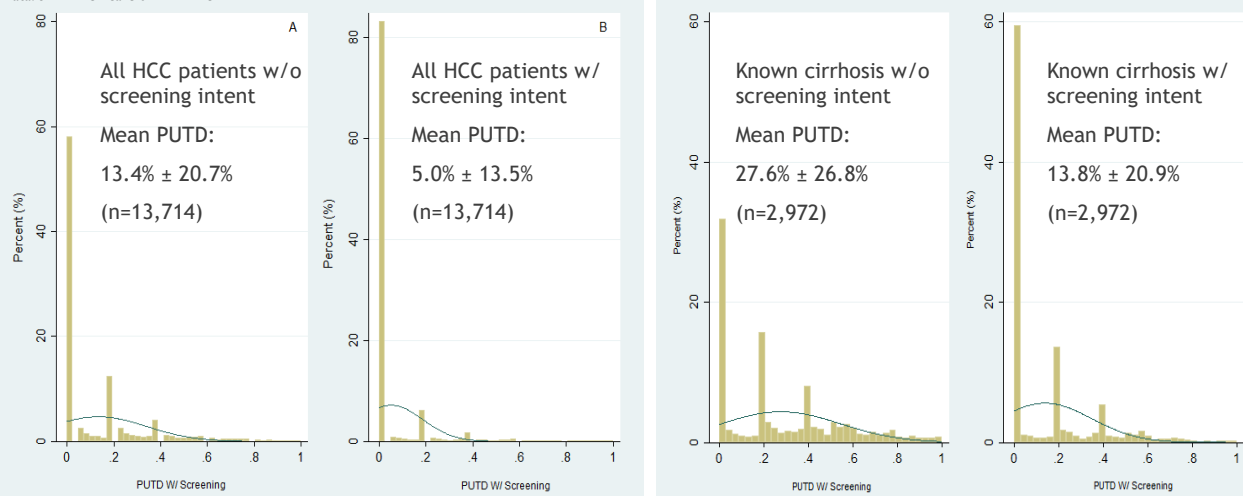
35

35

## Screening Rates: Method 2 - PUDF PUTD Distribution With & Without Intent



Data: SEER-Medicare & AMA File



36

## Q1: Measurement of HCC Screening Using Administrative Data

- Regardless of how you measure
- Screen rates are still low, despite steady increase over time

37

## Take Away Gap between real world and data world

- Real world
  - Would like to measure level of screening
  - With intent
  - For patients with cirrhosis
- Data world
  - No variable to capture intent: ML trained model to estimate
    - How valid is this? Still have question on threshold
    - Would you use it?
  - Not enough “events” of interest to model
    - model something that occurs more commonly, but may still introduce some useful information
    - Do you think modeling the full population, no intent adjusted, no cirrhosis would still generate useful results?

38

## Q2: Impact of Patient and Provider Factors on HCC Screening

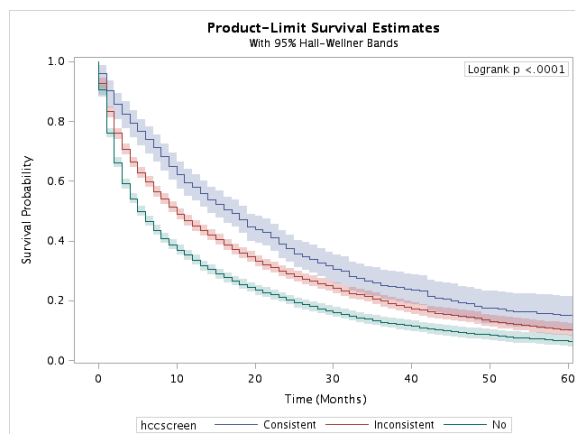
- Univariate Analysis: Correlates for PUDF covered by HCC screening
  - Among those with screening, female sex, Asian race, known cirrhosis, presence of a documented liver disease etiology, presence of decompensated cirrhosis, high comorbidity score and prior visit with a gastroenterologist/hepatologist were associated with higher PTC.
- Multivariable logistic regression model
  - Patients with consistent screening and inconsistent screening were associated with early tumor detection compared to no screening.

10/16/2020

39

39

## Q3: Impact of HCC Screening on Overall Survival Kaplan-Meier survival estimates by receipt of HCC screening




10/16/2020

40

40

- The median survival was 17 months for patients with consistent screening, 10 months for inconsistent screening, and 5 months for no screening estimated from Kaplan Meier curves.
- The 3-year survival rate was 25% (95% CI 22-28), 20% (95% CI 19-21), and 13% (95% CI 12-14) for patients with consistent, inconsistent, and no screening, respectively.
- Further **adjustment for length time bias** to screen-detected patients across all six assumptions had minimal impact on 1-, 3-, and 5-year survival rates (typically < 1% difference in survival rates compared to the estimators **adjusting for lead time bias alone**), so **inconsistent and consistent screening continued to be associated with a survival benefit relative to the no-screening group.**





## Conclusion

- In an analysis of the Surveillance, Epidemiology, and End Results Program (SEER) - Medicare database,
  - Q1: We found HCC screening to be underused for patients with cirrhosis
  - Q2: This contributes to detection of liver tumors at later stages and shorter times of survival
  - Q3: However, the proportion of patients screened for HCC has increased over time

10/16/202041

41





42

## Case Study: Medicaid Waiver Evaluation

### Detecting Change

10/16/2020

43

43


## Background: Medicaid 1115 Waiver Evaluation



- Medicaid: Insurance for the poor in the US
- Medicaid 1115 Waiver:
  - Billions of \$\$
  - Negotiated between the federal government (CMS) and Texas (TX-HHSC)
  - Many things... DSRIP (Delivery System Reform Incentive Payment) Program
- Evaluation: What was its impact?

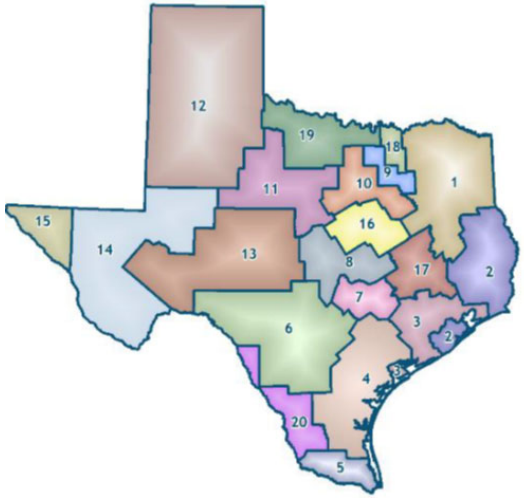
44

44




## Regional Healthcare Partnership (RHP)

- Research question
  - Did DSRIP change collaboration over time?
  - 2013 to 2020
  - After billions of \$\$



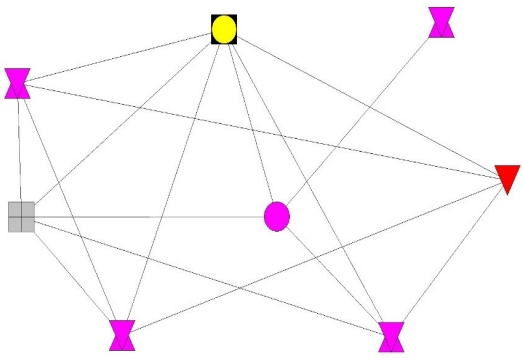
10/16/2020
45

45



## How do we measure collaboration?

- **Figure 1. Network Diagram, T<sub>0</sub>, RHP 15, All Collaboration**



**Organization Role in RHP (shape)**


- Anchor
- IGT only
- △ IGT + Performing Provider (Hospital)
- ▣ IGT + Performing Provider (CMHC)
- ▽ IGT + Performing Provider (Health Department)
- ◐ IGT + Performing Provider (HSC)
- ◊ IGT + Performing Provider (Health District)
- ⊘ Performing Provider only

**Organization Type (color)**

- Hospital
- Hospital / Health District or Hospital Authority
- County Government
- City Government
- School District
- EMS District
- CMHC
- Health Science Center
- Health Department
- Physician Practice
- Health District & Hospital Partnership

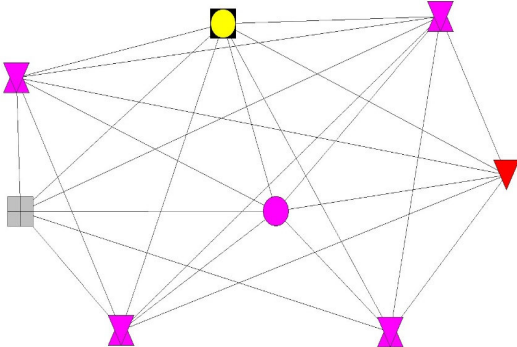
10/16/2020
46

46



## How do we measure collaboration?

- **Figure 1. Network Diagram, T<sub>1</sub>, RHP 15, All Collaboration**



**Organization Role in RHP (shape)**


- Anchor
- IGT only
- △ IGT + Performing Provider (Hospital)
- ▣ IGT + Performing Provider (CMHC)
- ▽ IGT + Performing Provider (Health Department)
- ⊙ IGT + Performing Provider (HSC)
- ◇ IGT + Performing Provider (Health District)
- ⊗ Performing Provider only

**Organization Type (color)**

- Hospital
- Hospital / Health District or Hospital Authority
- County Government
- City Government
- School District
- EMS District
- CMHC
- Health Science Center
- Health Department
- Physician Practice
- Health District & Hospital Partnership

10/16/2020 47

47



## Methods

- Social network analysis
- Graph algorithms
- Measures
  - # of ties (edges)
  - Type of ties (edges):
    - Any ties
    - Joint service delivery
    - Resource sharing
    - Data sharing
  - Density of network
  - Centrality of network

	Does your organization currently work with ___?			collaborate with ___ organization to deliver services?			Does your organization currently share tangible resources with ___ for the purpose of increasing access to services?			Does your organization currently have a data-sharing agreement with ___?		
	Yes	Not Sure	No	Yes	Not Sure	No	Yes	Not Sure	No	Yes	Not Sure	No
City of Laredo Health Department	○	○	○	○	○	○	○	○	○	○	○	○
Border Region MHMR Community Center	○	○	○	○	○	○	○	○	○	○	○	○
Doctors Hospital of Laredo	○	○	○	○	○	○	○	○	○	○	○	○
Laredo Medical Center	○	○	○	○	○	○	○	○	○	○	○	○

10/16/2020 48

48



## Methods: Density of network

- The proportion of ties that exist among the ties that are possible. If all organizations in a network share ties (indicate they work together) the density of ties in the network is 100%.

10/16/2020

49

49

Table 7. Network Density by RHP, All Collaboration

NETWORK DENSITY									
	T <sub>0</sub> (Pre-Waiver)	T <sub>1</sub> (2013)	T <sub>2</sub> (2015)	Change T <sub>0</sub> to T <sub>1</sub>		Change T <sub>1</sub> to T <sub>2</sub>		Overall Change T <sub>0</sub> to T <sub>2</sub>	
				Point Change*	% Change**	Point Change*	% Change**	Point Change*	% Change**
RHP 1	14%	22%	17%	8	54%	-5	-22%	3	21%
RHP 2	34%	38%	24%	4	11%	-14	-37%	-10	-30%
RHP 3	22%	24%	29%	3	12%	4	17%	7	31%
RHP 4	21%	26%	20%	5	25%	-6	-23%	-1	-3%
RHP 5	61%	75%	43%	14	24%	-32	-43%	-18	-29%
RHP 6	21%	28%	43%	7	36%	15	53%	22	108%
RHP 7	27%	27%	49%	0	0%	23	85%	23	85%
RHP 8	30%	30%	29%	0	0%	-1	-2%	-1	-2%
RHP 9	25%	28%	27%	4	15%	-1	-4%	3	11%
RHP 10	27%	27%	18%	0	-1%	-9	-34%	-10	-35%
RHP 11	43%	50%	18%	7	16%	-32	-65%	-25	-59%
RHP 12	29%	28%	21%	0	-1%	-8	-26%	-8	-27%
RHP 13	23%	43%	28%	20	87%	-15	-36%	5	21%
RHP 14	49%	56%	51%	8	16%	-5	-9%	3	6%
RHP 15	57%	89%	75%	32	56%	-14	-16%	18	31%
RHP 16	61%	83%	64%	22	36%	-19	-23%	3	5%
RHP 17	35%	37%	31%	2	5%	-6	-16%	-4	-12%
RHP 18	38%	69%	40%	31	82%	-29	-42%	2	6%
RHP 19	45%	56%	33%	12	26%	-23	-41%	-12	-26%
RHP 20	57%	61%	57%	4	6%	-4	-6%	0	0%
<sup>1</sup> Mean across RHPs	36%	45%	36%	9	25%	-9	-20%	0	0%

50

Data Science

- Task 1: Replicate the results?
- Task 2: Write a new report with new data
  - Are any modification needed?
  - If so, why?
  - And, what do I need to do differently?

10/16/2020
51

51

Raw data collected (20 RHP \* 3 time points = 60, n=4 to 38)

Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County								
Border Region Behavioral Health Center								
Camino Real Community Services								
City of Laredo Health Dept								
UTHSC-SA								
Maverick County Hospital District								
Driscoll Children's Hospital								
Laredo Medical Center								

10/16/2020
52

52

## Data Wrangling



The New York Times | <http://nyti.ms/1mZywnq>

TECHNOLOGY

### For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR | AUG. 17, 2014

- Data Wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning
  - reformatting data: MDY vs YMD
  - mapping data from one data model to another: ICD9 vs CPT code
  - and/or converting data into more consumable forms: to graphs
- 30-80% of the work in using big data
- Once raw data is “wrangled” into the correct analytic data
  - Running statistics models are fairly simple and similar to what you do traditionally
  - There are new methods but, usually requires a LOT of data

53

## Raw data collected (20 RHP \* 3 time points = 60, n=4 to 38)




Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X

10/16/2020


54

54

								
Raw data collected								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X


10/16/2020 55

55

								
Raw data collected (missing data)								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X


10/16/2020 56

56

								
Raw data collected (missing data)								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X		0			0		
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center				Missing				X


10/16/2020 57

57

								
Raw data collected (missing data)								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X


10/16/2020 58

58

								
Raw data collected								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center	1		1		1		1	X

10/16/2020 59

59

								
Raw data collected								
Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X

10/16/2020 60

60



Go to google docs in the chat box  
Do QI link

61

## What did they do?

- Reconstruction = use what you have from either respondent to form any tie (edge)
- Is this ok to do?
- Are there any problems with this?

62

**“Activities that make data more usable by changing their form, not their meaning”**



- Would you change the data format?
- Why? How?

Organization	Webb County	Border Region Behavioral Health Center	Camino Real Community SVC	City of Laredo Health Dept	UTHSC-SA	Maverick County Hospital District	Driscoll Children's Hosp.	Laredo Medical Center
Webb County	X							
Border Region Behavioral Health Center		X						
Camino Real Community Services			X					
City of Laredo Health Dept				X				
UTHSC-SA					X			
Maverick County Hospital District						X		
Driscoll Children's Hospital							X	
Laredo Medical Center								X

10/16/2020

63

63


Hospital A	Hospital B	Edge
Webb County	Border Region Behavioral Health Center	1
Webb County	Camino Real Community Services	0
Webb County	City of Laredo Health Dept	1
Webb County	UTHSC-SA	1
Webb County	Maverick County Hospital District	1
Webb County	Driscoll Children's Hospital	1
Webb County	Laredo Medical Center	1
Border Region Behavioral Health Center	Camino Real Community Services	1
Border Region Behavioral Health Center	City of Laredo Health Dept	1
Border Region Behavioral Health Center	UTHSC-SA	0
Border Region Behavioral Health Center	Maverick County Hospital District	1
Border Region Behavioral Health Center	Driscoll Children's Hospital	0
Border Region Behavioral Health Center	Laredo Medical Center	1
Camino Real Community Services	City of Laredo Health Dept	1
Camino Real Community Services	UTHSC-SA	0
Camino Real Community Services	Maverick County Hospital District	1
Camino Real Community Services	Driscoll Children's Hospital	0

64




Hospital A	Hospital B	Edge	RHP
Webb County	Border Region Behavioral Health Center	1	15
Webb County	Camino Real Community Services	0	15
Webb County	City of Laredo Health Dept	1	15
Webb County	UTHSC-SA	1	15
Webb County	Maverick County Hospital District	1	15
Webb County	Driscoll Children's Hospital	1	15
Webb County	Laredo Medical Center	1	15
Border Region Behavioral Health Center	Camino Real Community Services	1	15
Border Region Behavioral Health Center	City of Laredo Health Dept	1	15
Border Region Behavioral Health Center	UTHSC-SA	0	15
Border Region Behavioral Health Center	Maverick County Hospital District	1	15
Border Region Behavioral Health Center	Driscoll Children's Hospital	0	15
Border Region Behavioral Health Center	Laredo Medical Center	1	15
Camino Real Community Services	City of Laredo Health Dept	1	15
Camino Real Community Services	UTHSC-SA	0	15
Camino Real Community Services	Maverick County Hospital District	1	15
Camino Real Community Services	Driscoll Children's Hospital	0	15

65



Go to google docs in the chat box  
Do Q2 link

66



## Big data?

- 8 providers in that on RHP
  - $8*8=64-8=56$
  - If no direction:  $56/2=28$
- 20 RHPs...
- 300 hospitals
  - LOTS of rows.. Not by hand
- Big data: about complexity

10/16/2020 67

67

NETWORK DENSITY			
Density -T0 (Pre-Waiver)	Density-T1 -2013	Density-T2 -2015	Density-T3 -2020
14%	22%	17%	36%
34%	38%	24%	32%
22%	24%	29%	23%
21%	26%	20%	26%
61%	75%	43%	36%
21%	28%	43%	26%
27%	27%	49%	38%
30%	30%	29%	21%
25%	28%	27%	25%
27%	27%	18%	25%
43%	50%	18%	30%
29%	28%	21%	21%
23%	43%	28%	21%
49%	56%	51%	58%
57%	89%	75%	89%
61%	83%	64%	71%
35%	37%	31%	36%
38%	69%	40%	33%
45%	56%	33%	39%
57%	61%	57%	67%
36%	45%	36%	40%
30%	37%	30%	32%

10/16/2020 68

68

- Calculate the density change from T1 to T2 for the first row (17% to 36%)

NETWORK DENSITY			
Density -T0 (Pre-Waiver)	Density-T1 -2013	Density-T2 -2015	Density-T3 -2020
14%	22%	17%	36%
34%	38%	24%	32%
22%	24%	29%	23%
21%	26%	20%	26%
61%	75%	43%	36%
21%	28%	43%	26%
27%	27%	49%	38%
30%	30%	29%	21%
25%	28%	27%	25%
27%	27%	18%	25%
43%	50%	18%	30%
29%	28%	21%	21%
23%	43%	28%	21%
49%	56%	51%	58%
57%	89%	75%	89%
61%	83%	64%	71%
35%	37%	31%	36%
38%	69%	40%	33%
45%	56%	33%	39%
57%	61%	57%	67%
36%	45%	36%	40%
30%	37%	30%	32%

10/16/2020

69

69



Go to google docs in the chat box  
Do Q3 link

70

- Calculate the density change from T1 to T2 for the first row (17% to 36%)
- Point change
  - $36\% - 17\% = 0.36 - 0.17 = 0.19$
  - Is this 0.2 or 0.19?
- % change equation?

NETWORK DENSITY							
Density -T0 (Pre-Waiver)	Density-T1 -2013	Density-T2 -2015	Density-T3 -2020	Density-Overall T2 to T3		Density- Overall T0 to T3	
				Point Change*	% Change**	Point Change*	% Change**
14%	22%	17%	36%	0.2	111%	0.2	156%
34%	38%	24%	32%	0.1	35%	0.0	-5%
22%	24%	29%	23%	-0.1	-20%	0.0	6%
21%	26%	20%	26%	0.1	32%	0.1	26%
61%	75%	43%	36%	-0.1	-17%	-0.3	-42%
21%	28%	43%	26%	-0.2	-38%	0.1	26%
27%	27%	49%	38%	-0.1	-22%	0.1	41%
30%	30%	29%	21%	-0.1	-29%	-0.1	-32%
25%	28%	27%	25%	0.0	-6%	0.0	1%
27%	27%	18%	25%	0.1	37%	0.0	-9%
43%	50%	18%	30%	0.1	69%	-0.1	-29%
29%	28%	21%	21%	0.0	-2%	-0.1	-29%
23%	43%	28%	21%	-0.1	-27%	0.0	-11%
49%	56%	51%	58%	0.1	13%	0.1	18%
57%	89%	75%	89%	0.1	19%	0.3	57%
61%	83%	64%	71%	0.1	12%	0.1	17%
35%	37%	31%	36%	0.1	17%	0.0	4%
38%	69%	40%	33%	-0.1	-17%	0.0	-12%
45%	56%	33%	39%	0.1	19%	-0.1	-12%
57%	61%	57%	67%	0.1	17%	0.1	17%
36%	45%	36%	40%	0.0	10%	0.0	9%
30%	37%	30%	32%	0.0	5%	0.0	5%

10/16/2020

71

71

## Detecting change



- A=17% B=36%
- B-A=19%
- $(B-A)/A=111\%$
- $(B-A)/B=52.5\%$
- Which one?
- Depends: Devils in the details
- In this example,  $(B-A)/A=111\%$ 
  - Why?



What could still be wrong?

10/16/2020

72


72

	NUMBER OF PROVIDERS			NETWORK DENSITY							
	NPROV T0/T1	NPROVT2	NPROV T3	Density -T0 (Pre-Waiver)	Density-T1 -2013	Density-T2 -2015	Density-T3 -2020	Density-Overall T2 to T3		Density- Overall T0 to T3	
								Point Change*	% Change**	Point Change*	% Change**
RHP 1	38	40	20	14%	22%	17%	36%	0.2	111%	0.2	156%
RHP 2	17	17	15	34%	38%	24%	32%	0.1	35%	0.0	-5%
RHP 3	30	33	25	22%	24%	29%	23%	-0.1	-20%	0.0	6%
RHP 4	25	25	17	21%	26%	20%	26%	0.1	32%	0.1	26%
RHP 5	8	8	10	61%	75%	43%	36%	-0.1	-17%	-0.3	-42%
RHP 6	27	27	23	21%	28%	43%	26%	-0.2	-38%	0.1	26%
RHP 7	16	17	7	27%	27%	49%	38%	-0.1	-22%	0.1	41%
RHP 8	16	17	13	30%	30%	29%	21%	-0.1	-29%	-0.1	-32%
RHP 9	25	25	23	25%	28%	27%	25%	0.0	-6%	0.0	1%
RHP 10	30	33	24	27%	27%	18%	25%	0.1	37%	0.0	-9%
RHP 11	19	19	15	43%	50%	18%	30%	0.1	69%	-0.1	-29%
RHP 12	37	39	36	29%	28%	21%	21%	0.0	-2%	-0.1	-29%
RHP 13	21	21	13	23%	43%	28%	21%	-0.1	-27%	0.0	-11%
RHP 14	12	13	10	49%	56%	51%	58%	0.1	13%	0.1	18%
RHP 15	8	8	8	57%	89%	75%	89%	0.1	19%	0.3	57%
RHP 16	9	10	7	61%	83%	64%	71%	0.1	12%	0.1	17%
RHP 17	19	20	12	35%	37%	31%	36%	0.1	17%	0.0	4%
RHP 18	10	10	6	38%	69%	40%	33%	-0.1	-17%	0.0	-12%
RHP 19	13	15	12	45%	56%	33%	39%	0.1	19%	-0.1	-12%
RHP 20	8	8	4	57%	61%	57%	67%	0.1	17%	0.1	17%
Mean ad	-	-	-	36%	45%	36%	40%	0.0	10%	0.0	9%
Weighted average				30%	37%	30%	32%	0.0	5%	0.0	5%

10/16/2020

73

73



## Methods: Density of network

- The proportion of ties that exist among the **ties that are possible**. If all organizations in a network share ties (indicate they work together) the density of ties in the network is 100%.
- What about non-response? Missing data (veracity)
- In a network of 10 orgs, if you only have 5 responses, but each is asked about all 10 orgs, how do you measure ties?
  - Complete - case approach = only use full data (n=6; 15)
  - Mean imputation (56)
  - Reconstruction = use what you have from either respondent about the tie (28)
  - Complex multiple imputation (Not possible)

10/16/2020

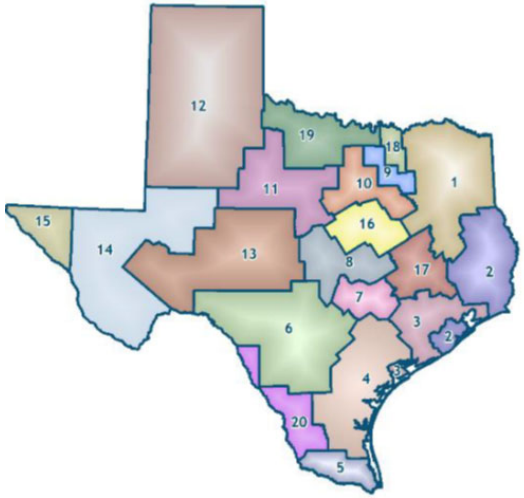
74

74

ATM POPULATION INFORMATICS

## Real World Regional Healthcare Partnership

- Research question
  - Did DSRIP change collaboration over time?



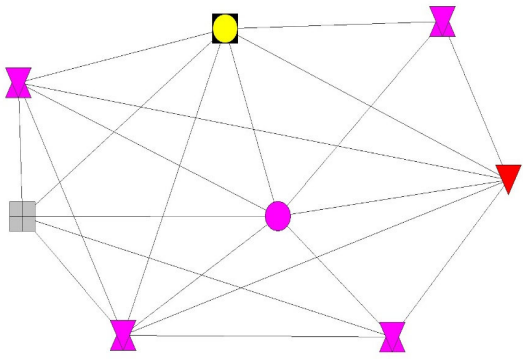
10/16/2020
75

75

ATM POPULATION INFORMATICS

## Data World How do we measure collaboration?

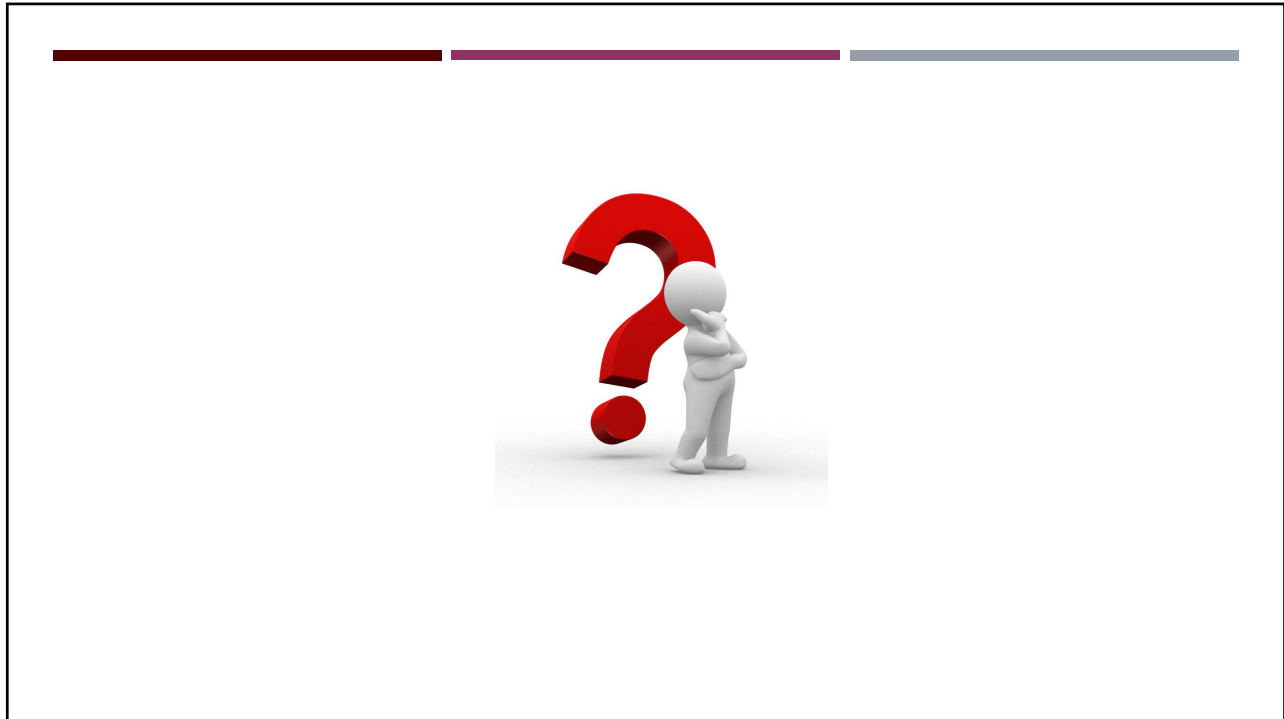
- **Figure 1. Network Diagram RHP 15, All Collaboration**



- Issues
  - Number of nodes changed
  - Incomplete data
  - Not sufficient data
- Potential Answers (gdocs Q4)
  - Mean imputation (Y/N)?
  - Compare as is?
  - Impossible
    - Any better than qualitative?

10/16/2020
76

76



77

The logo for ATM Population Informatics, featuring the letters 'ATM' in a stylized font next to the words 'POPULATION INFORMATICS'.

## Data cleaning: dropping for age

- If age  $\geq$  18
- Dropped 98 rows (of 504)

What is wrong? How will you troubleshoot?

10/16/2020

78

78

## Conditional to subset



```

1 library(tidyverse)
2
3 a <- read.csv("faq_data_v2.csv")
4
5 a$final <- 0
6 a[which(a$Q28 != "" & a$Q27 != "" & as.numeric(a$age2) >= 18), "final"] <- 1
7
8 b <- a[which(a$final == 0),] %>%
9   select(pid, Q27, Q28, age2)
10
11
12
13
14

```

Environment	History	Connections	Tutorial
<div style="display: flex; justify-content: space-between; align-items: center;"> <span>Import Dataset</span> <span>List</span> </div>			
Global Environment			
Data			
a	504 obs. of 28 variables		
b	98 obs. of 4 variables		

79


- The goal was delete rows with empty Q27 or Q28 or age < 18
- By opening b we can see that there are some rows with age > 18 that have value for Q27 and Q28!!!!
- Here only few of columns are showed. (dataset has more columns)

	pid	Q27	Q28	age2
3	1702445558	Extremely useful	Definitely prefer FAQ format	26
8	1540597975	Moderately useful	I have no preference	20
15	1632230547			65
17	1700550084	Very useful	Definitely prefer traditional privacy statement format	26
22	1505131424	Moderately useful	I have no preference	15
33	1674948158	Moderately useful	Somewhat prefer FAQ format	18
35	1702503056	Not useful	I have no preference	22
37	1500734419	Extremely useful	Definitely prefer traditional privacy statement format	14

Showing 1 to 9 of 98 entries, 4 total columns

80





## Special characters in CSV files

- Open data in excel:
 

3	1702445558	Female	White	Some Coll	Less than	South	Yes	Extremely	Definitely	26	26
---	------------	--------	-------	-----------	-----------	-------	-----	-----------	------------	----	----
  
- Open data in Notepad:
 

```
3,1702445558,Female,White,Some College completed,"Less than $30,000",South,\
,Yes,Extremely useful,Definitely prefer FAQ format,26,$26,
```
  
- Hidden special character may cause a bug in your program!

81




## What if my file is BIG?

- Sed
- Perl
- Vim
- Emacs
- Sublime

10/16/2020
82

82




## What you see is not what you get

- Change of decimals.. Why?

Point Change*	Point Change*
0.2	0.19

10/16/2020 83

83



## What you see is not what you get

- Change of decimals.. Why?


Point Change*	Point Change*
0.2	0.19

Binary Numbers

1001	0001	0101
8 4 2 1	8 4 2 1	8 4 2 1
$8^*1+1^*1=9$	$1^*1=1$	$4^*1+1^*1=5$

10/16/2020 84

84



## What you see is not what you get

- Change of decimals.. Why?

Point Change*
0.2


Point Change*
0.19

### Binary Numbers

1001	0 0
8 4 2 1	1/2 1/4 1/8 1/16
$8*1+1*1=9$	$0.5+0.0625=0.5625$

10/16/2020
85

85



## What you see is not what you get

- Change of decimals.. Why?

Point Change*
0.2

Point Change*
0.19

### Decimal fraction to Binary fraction Lose precision

$0.200000000000$   
 $= .00110011001100110011001$   
 $+ \text{remainder } 0.00000071526$

10/16/2020
86

86

## What is a Variable?

- A user defined name to represent a piece of memory for storing evaluated value(s). A variable consists of 5 items
- Name:
  - meaningful human readable name
  - How the user refers to variable
- Data Type: How to interpret variable for data representation
- Size:
  - How much storage memory is needed to store data value
  - Can be inferred from data type
- Value:
  - Actual value associated with variable
  - stored in memory
- Storage location:
  - Usually hidden from user by the interpreter or compiler
  - How the computer refers to a variable

10/16/2020

87

87

## Variable Types

Type	Stored value	Interpreted value	Label Interpreted Value
int	1000001 (65)	65	65 or older
Char/string (ASCII)	1000001 (65)	A	Asian
date	1000001 (65)	1960/3/6 (SAS)	


- 1 0 0 0 0 0 1 =64+1=65
- 64 32 16 8 4 2 1

88

0	<NUL>	32	<SPC>	64	@	96	`	128	À	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Á	161	°	193	¡	225	·
2	<STX>	34	"	66	B	98	b	130	Â	162	±	194	ª	226	,
3	<ETX>	35	#	67	C	99	c	131	Ã	163	£	195	»	227	"
4	<EOT>	36	\$	68	D	100	d	132	Ä	164	§	196	¼	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Å	165	•	197	½	229	À
6	<ACK>	38	&	70	F	102	f	134	Ä	166	¶	198	¾	230	Á
7	<BEL>	39	'	71	G	103	g	135	Å	167	ß	199	«	231	Â
8	<BS>	40	(	72	H	104	h	136	à	168	@	200	»	232	Ã
9	<TAB>	41	)	73	I	105	i	137	á	169	©	201	…	233	Ä
10	<LF>	42	*	74	J	106	j	138	â	170	™	202	—	234	Å
11	<VT>	43	+	75	K	107	k	139	ã	171	'	203	À	235	Ä
12	<FF>	44	,	76	L	108	l	140	ä	172	"	204	Á	236	Å
13	<CR>	45	-	77	M	109	m	141	å	173	#	205	Â	237	Ä
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	ø	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	-	240	•
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	◊
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	"	242	Ù
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	"	243	Ú
20	<DC4>	52	4	84	T	116	t	148	í	180	¥	212	'	244	Û
21	<NAK>	53	5	85	U	117	u	149	î	181	µ	213	'	245	Ü
22	<SYN>	54	6	86	V	118	v	150	ï	182	ð	214	÷	246	Ý
23	<ETB>	55	7	87	W	119	w	151	ó	183	Σ	215	◊	247	ÿ
24	<CAN>	56	8	88	X	120	x	152	ò	184	Π	216	ÿ	248	ÿ
25	<EM>	57	9	89	Y	121	y	153	ó	185	π	217	ÿ	249	ÿ
26	<SUB>	58	:	90	Z	122	z	154	ô	186	ƒ	218	/	250	ÿ
27	<ESC>	59	;	91	[	123	{	155	õ	187	á	219	€	251	ÿ
28	<FS>	60	<	92	\	124		156	ú	188	°	220	<	252	ÿ
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	>	253	ÿ
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	ÿ
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fi	255	ÿ

## ASCII: CHARACTER ENCODING

89



### Programming: talking to your computer

- OUTPUT : Know what you want
- INPUT : what you have
- Intermediate results: What you need
- Program: change what you have (INPUT) to what you need (intermediate results. Often more than one level) to what you want (OUTPUT)

```

graph LR
    INPUT[INPUT] --> PROGRAM[PROGRAM  
Intermediate Results]
    PROGRAM --> OUTPUT[OUTPUT]
  
```

90


Closing thoughts...

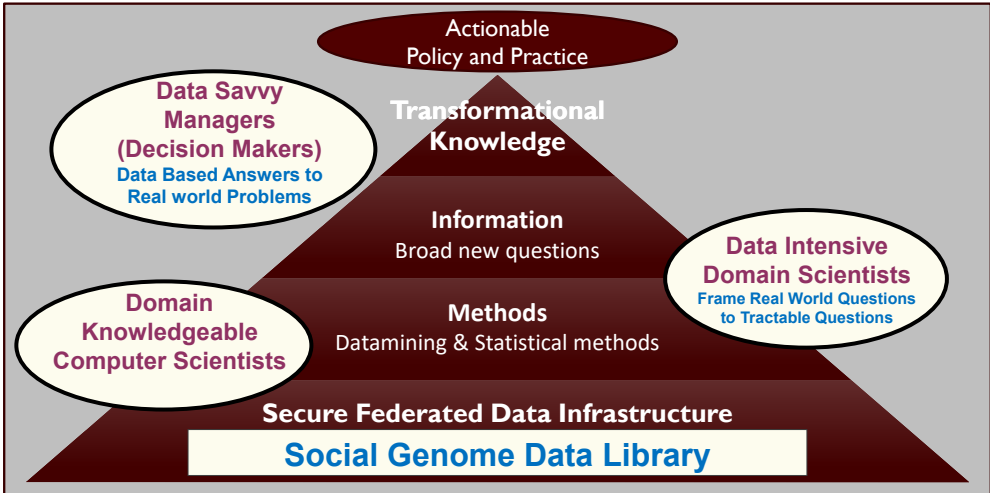
[Redacted]

10/16/2020 91

91

**Population Informatics: The systematic study of populations via secondary analysis of massive data collections (“big data”) about people**





Actionable Policy and Practice

Transformational Knowledge

Information  
Broad new questions

Methods  
Datamining & Statistical methods

Secure Federated Data Infrastructure

**Social Genome Data Library**

**Data Savvy Managers (Decision Makers)**  
Data Based Answers to Real world Problems

**Domain Knowledgeable Computer Scientists**

**Data Intensive Domain Scientists**  
Frame Real World Questions to Tractable Questions

92

## What is data science ?

### Hye-Chung Kum



- **Measurement (=features):** Smart/clever counting of real things (meaningful to people) in the digital data
- **Information generation:** Then modeling using those measures (features)
- **Delivery of information:** Storytelling with data
- **Develop agile data pipeline** for timely processing that can be iteratively updated to track the dynamic ever changing real world
- **Doing Data Science Right**
  - Devil is in the details!
  - Goldilocks principle: Not too hot, not too cold!
  - LOTS of critical thinking about
    - What exactly is the goal ?
    - What is real? Meaningful?

93

## Conclusion



- There is a lot you can do with digital data now
- **BUT,** lots of data is not the answer
  - You have to learn to use data properly
  - You have to learn to handle data if you want to do good research using massive secondary data
  - Massive secondary data requires as much or more preprocessing as primary data collection
    - Research design, data cleaning, data preparation
  - Nothing replaces common sense (critical thinking) and curiosity in research

94

## Building Capacity

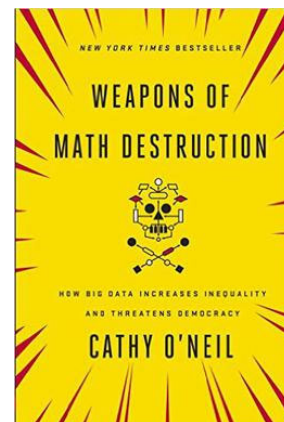
- For handling person level data
  - Know information privacy
  - Know basic IT and security
  - Know basic legal issues in person level data
  - Know how to handle/process raw data
    - Clean, Merge, Transform etc
  - Know how to build/detect meaningful features to use in modeling
  - Modeling
  - Be able to think critically and move between
    - Real world problem
    - Data problem



95

## Fairness in blackbox algorithms

- Some algorithms reinforce discrimination that exist in our real world
- She posits that these problematic mathematical tools
  - Are opaque
  - Unregulated and difficult to contest
  - And scalable
- Amplify any inherent biases to affect increasingly larger populations



10/16/2020

96

96



## Fairness in blackbox algorithms



- Examples
  - Google Photos mistakes in labeling
  - Facebook requires extra work for some native Americans to get an account
- Take away:
  - Human critical thinking and judgment is very important to using algorithms appropriately
  - There must be humans who will take on the responsibility for the decision



10/16/2020

97

97

## Thank You!!



Hye-Chung Kum (kum@tamu.edu)

Director of Population Informatics Lab (<https://pinformatics.org/>)

Director of ViDaL (<https://vidal.tamu.edu>)

IRB/DUA/Vidal

98

98