University of California
San Francisco

# Database Research

## *An Opportunity Worth Exploring*

M. Shoaib Khan, MD, MSPH, AAHIVS, FAAFP
Department of Family and Community Medicine, UCSF Fresno

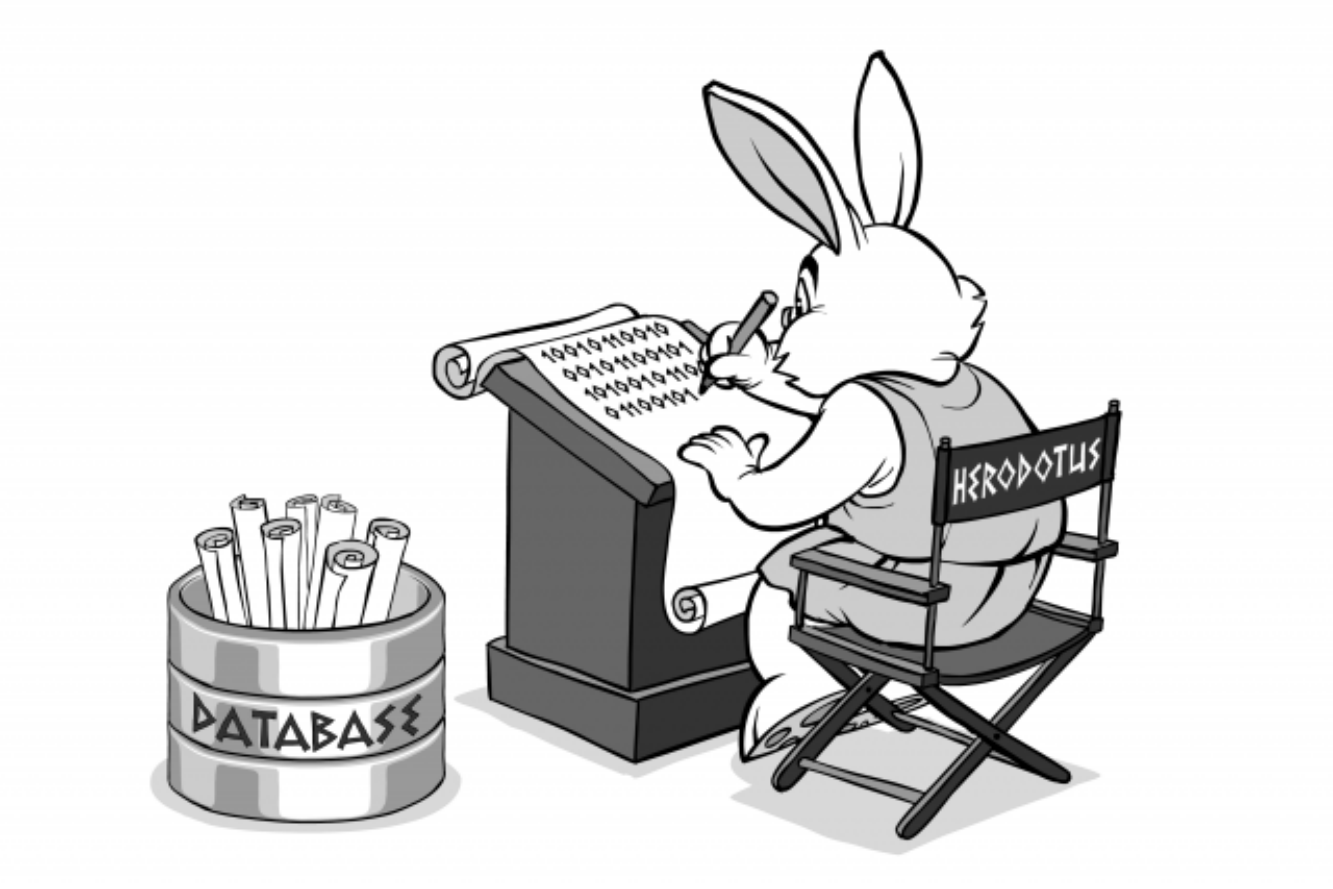
# Disclosures

- None

# Objectives

- To provide an overview of healthcare-related databases
- To compare the benefits and limitations of database research
- To discuss the skills and resources needed for research using big data
- To provide an example of a database research project: inception to publication

# Historical Journey of Big Data



# History of Big Data Collection in the US

- **1790**: 1st census in the US started and is conducted every 10 years since then

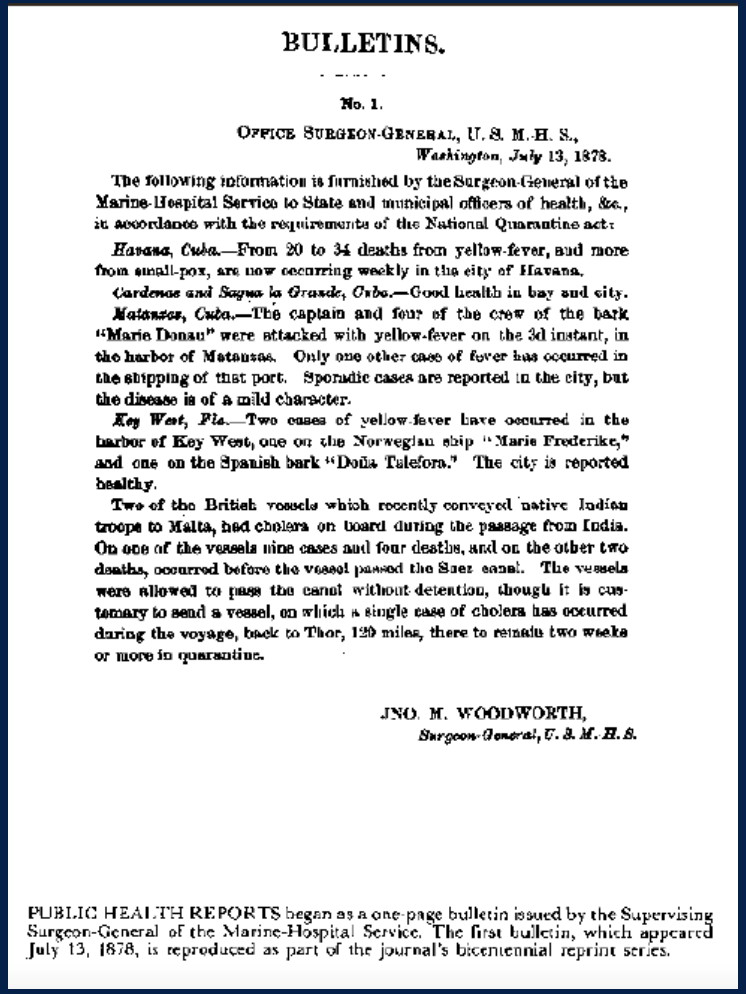
- **1937**: 1st major data gathering project awarded to a IBM as part of the Social Security Act to create punch card-reading machines

- **1965**: 1st data center created by the US to store tax returns and fingerprint data on magnetic tape

# History of Healthcare Data Collection in the US

- **1878**: National Quarantine Act

  - Required the Surgeon General of the U.S. Marine-Hospital Service (later the U.S. Public Health Service) to collect reports from U.S. consular officers on the sanitary condition of vessels departing for the US and to notify federal and state officers through weekly abstracts



BULLETINS.

No. 1.

OFFICE SURGEON-GENERAL, U. S. M.-H. S.,
*Washington, July 13, 1878.*

The following information is furnished by the Surgeon-General of the Marine-Hospital Service to State and municipal officers of health, &c., in accordance with the requirements of the National Quarantine act:

*Havana, Cuba.*—From 20 to 34 deaths from yellow-fever, and more from small-pox, are now occurring weekly in the city of Havana.

*Cardenas and Sagua la Grande, Cuba.*—Good health in bay and city.

*Matanzas, Cuba.*—The captain and four of the crew of the bark "Marie Donau" were attacked with yellow-fever on the 3d instant, in the harbor of Matanzas. Only one other case of fever has occurred in the shipping of that port. Sporadic cases are reported in the city, but the disease is of a mild character.

*Key West, Fla.*—Two cases of yellow-fever have occurred in the harbor of Key West, one on the Norwegian ship "Marie Frederike," and one on the Spanish bark "Doña Talefora." The city is reported healthy.

Two of the British vessels which recently conveyed native Indian troops to Malta, had cholera on board during the passage from India. On one of the vessels nine cases and four deaths, and on the other two deaths, occurred before the vessel passed the Suez canal. The vessels were allowed to pass the canal without detention, though it is customary to send a vessel, on which a single case of cholera has occurred during the voyage, back to Thor, 120 miles, there to remain two weeks or more in quarantine.

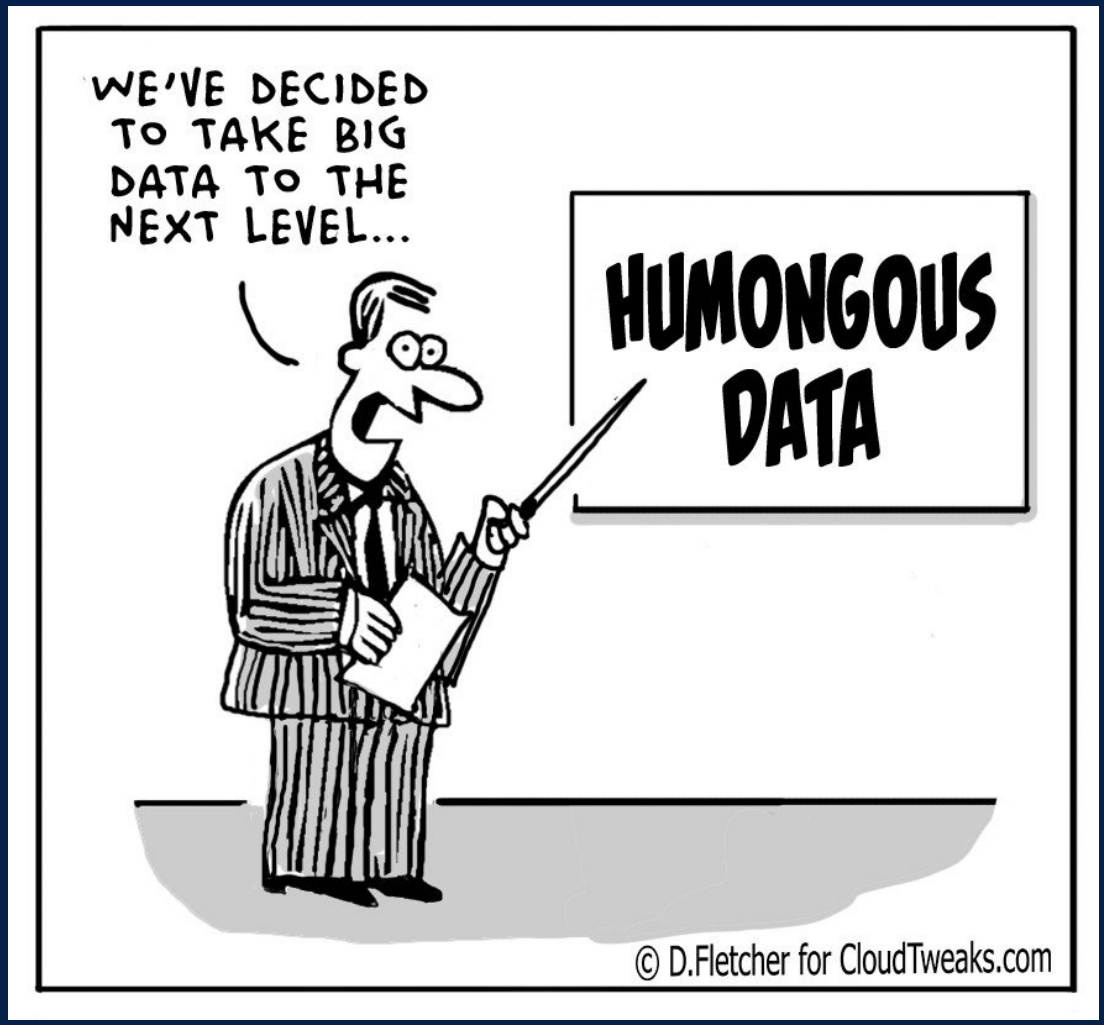
JNO. M. WOODWORTH,
*Surgeon-General, U. S. M. H. S.*

PUBLIC HEALTH REPORTS began as a one-page bulletin issued by the Supervising Surgeon-General of the Marine-Hospital Service. The first bulletin, which appeared July 13, 1878, is reproduced as part of the journal's bicentennial reprint series.

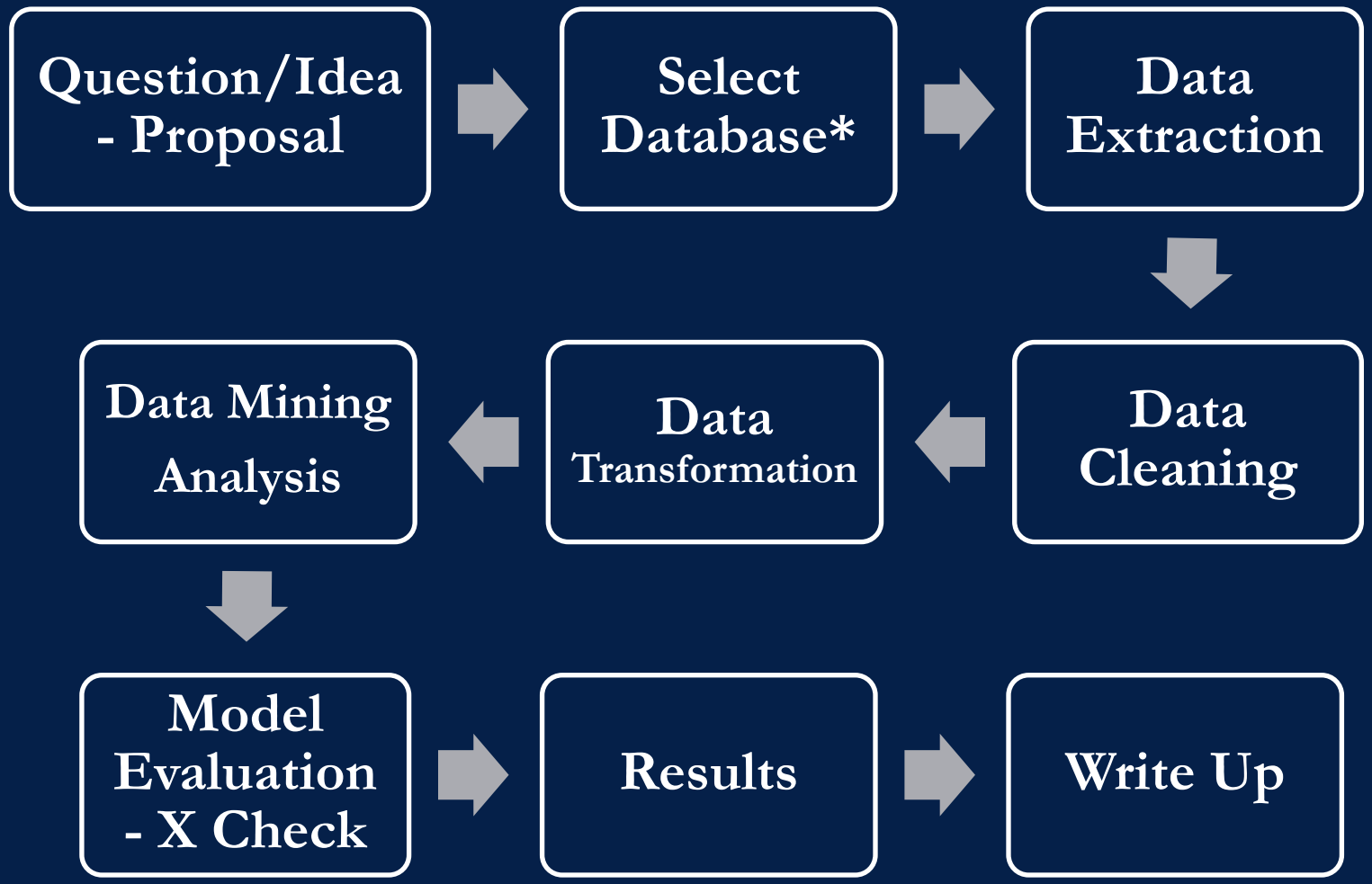
# History of Healthcare Data Collection in the US

| Year | Major Event |
|---|---|
| 1878 | National Quarantine Act allows the 1st issue of *The Bulletin of the Public Health*, the first ancestor of *MMWR*, is published. |
| 1887 | The 1st *Weekly Abstract of the Sanitary Reports* is published. |
| 1896 | *The Weekly Abstract of Sanitary Reports* becomes *Public Health Reports,* the official journal of the U.S. Public Health Service |
| 1925 | All states started reporting morbidity information but was not used for disease control |
| 1950 | National Surveillance Program started. Dissemination of federal morbidity and mortality statistics is transferred from *Public Health Reports* to the *Weekly Morbidity Report,* a new publication of the federal National Office of Vital Statistics (NOVS). |
| 1952 | NOVS changes the name of the *Weekly Morbidity Report to* the *Morbidity and Mortality Weekly Report (MMWR)*. |
| 1961 | CDC publishes its first issue of *MMWR.* |

# Healthcare Databases for Researchers



# Steps from Inception to Publication

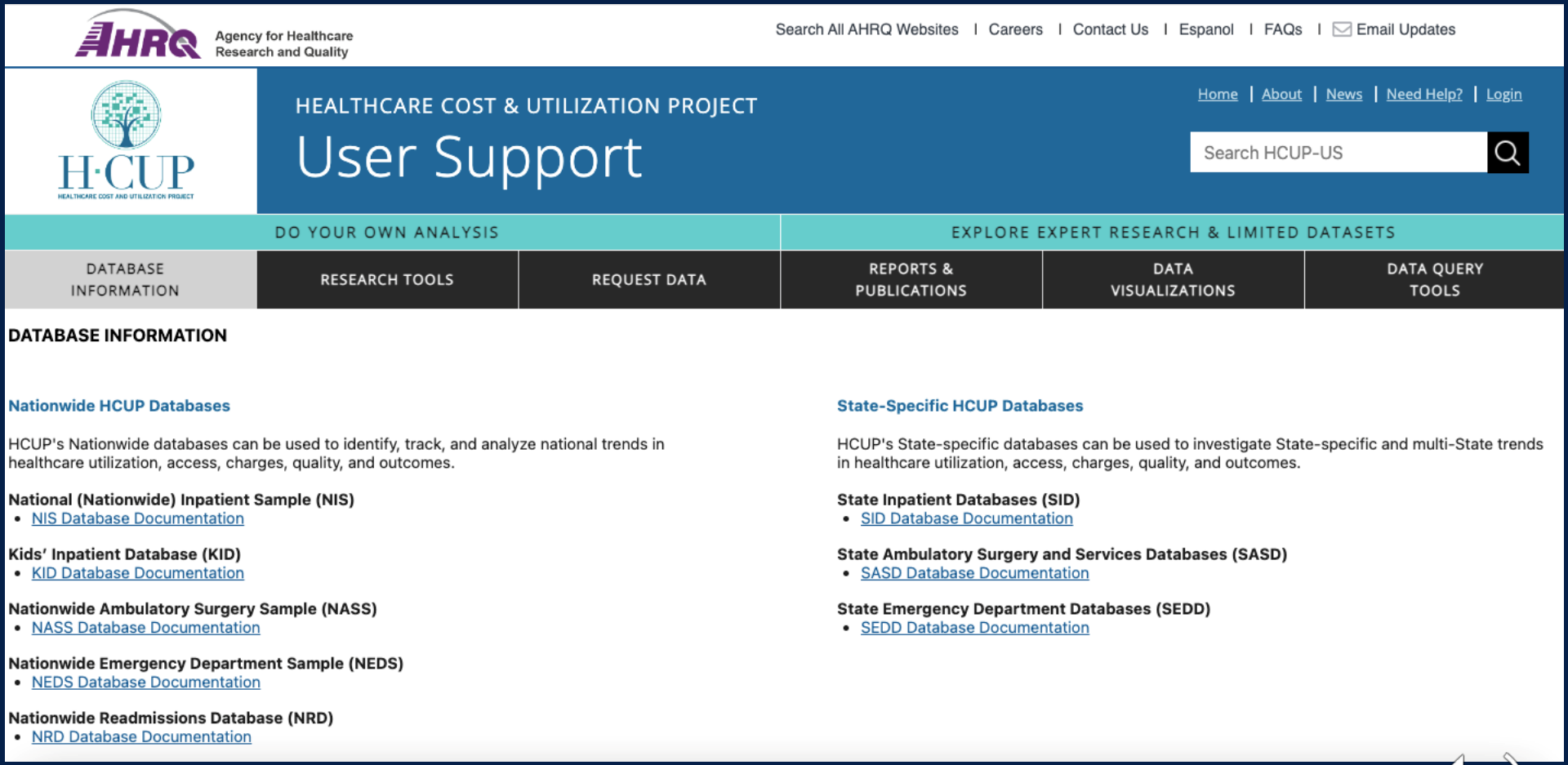


# US Healthcare Data available to researchers

- AHRQ: Healthcare Cost and Utilization Project (HCUP)
  - https://hcup-us.ahrq.gov/databases.jsp
  - Cost: Free

HCUP provides reliable, comprehensive information that can be used to answer questions about healthcare use, access, outcomes, and costs related to hospital inpatient stays, ambulatory surgery and services, emergency department visits, and readmissions.

# US Healthcare Data available to researchers



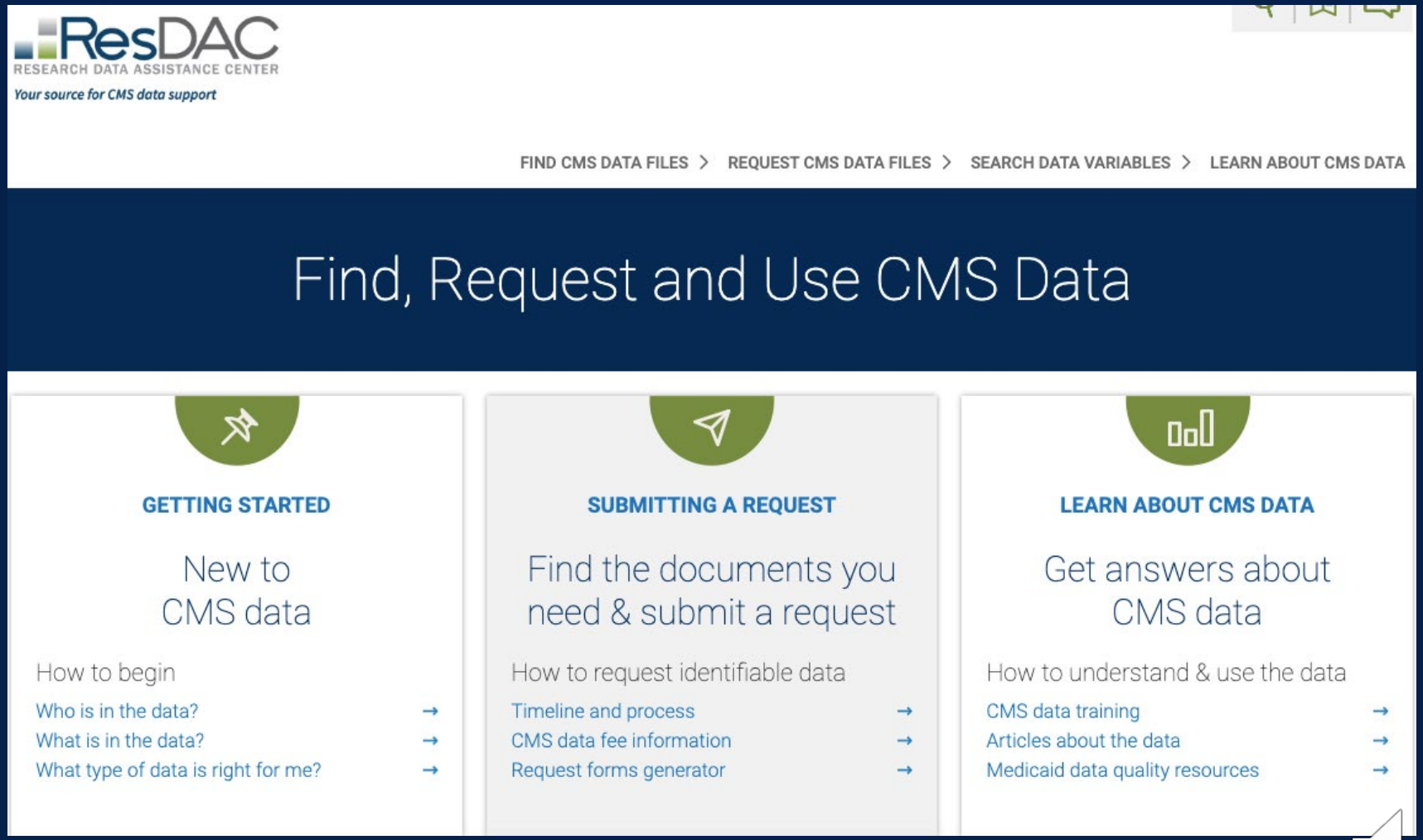

UCSF

# US Healthcare Data available to researchers

- CMS: Centers for Medicare & Medicaid Services
- https://resdac.org/
- Cost: $2000 for researchers; fee waived for students

The Research Data Assistance Center (ResDAC) provides free assistance to academic and non-profit researchers interested in using Medicare, Medicaid, SCHIP, and Medicare Current Beneficiary Survey (MCBS) data for research. Primary funding for ResDAC comes from a CMS research contract.

# US Healthcare Data available to researchers



# Benefits and Limitations of Databases Research

# Big Data Research: Benefits & Limitations

| Considerations | Pros | Cons |
|---|---|---|
| Time Spent in Data Collection | None. Efficient way to answer a question fast | Not able to answer all questions |
| Quality of Data | Consistent and vetted by organizations | Dependent on participating organizations. Limited by scope. |
| Type of Data | Depends on the database. Quantitative. | Lacks depth and granularity, especially for clinical projects |
| Types of Analysis | Depends on the database | User dependent and improved with experience |
| Cost of Use | Depends on the database | Participating institutions have an upper hand, so it's not equitable |
| Skills needed | Mostly statistical and data management skills | Can and sometimes required to use database personnel that will cost |
| Equipment/ Software needed | Data able to be analyzed with common software | Larger databases require bigger RAMs, faster processing computers |

# Skills and Resources Needed to Use Databases

# Skill #1: Learn to use the database

- Data files may be coded as numeric data; learn to accurately decode using “Key” file
  - E.g., gender coded as 1, 2, 3, 4; key will indicate 1= male, 2= female, 3= non-binary, 4= UNK
- Be careful copying data and keep logs of your actions with time stamps
- Be consistent in the way you use the database, so that you can trace your steps
  - Develop standardized operating procedures and divide roles clearly when working in a team to avoid duplication of efforts

# Skill #2: Statistics

- Learn basic level skills, work on them continuously to expand them
  - See reference file for training resources
- Ask for help from trained statisticians when performing complex analysis
  - Have them verify your numbers
- Write a clear statistical plan in your proposal prior to starting the analysis
  - You can always change your plan, but make sure to update the proposal

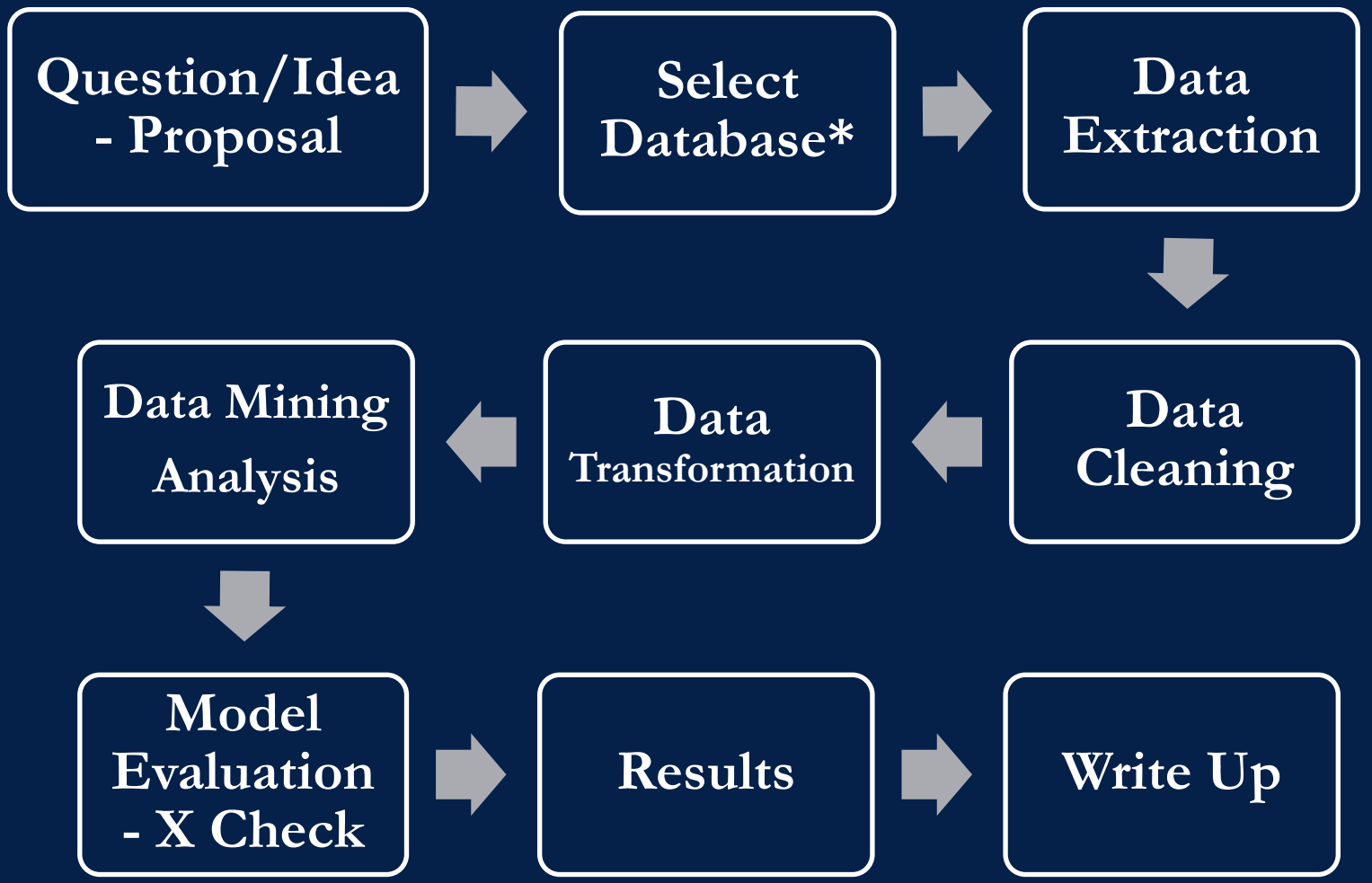
# Resource #1: Computer

- Large data files can include data points in the millions that will require faster processing, RAM computers
- Some database files are downloadable, but some will come on discs; may need to obtain DVD players
- Save your datafiles often and on secure servers on the cloud; independent of computer malfunctioning
  - Create standardized procedure to save your data files, so you can identify what file was used for a specific analysis

# Resource #2: Analytical Software

- Most databases allow for common analytical software to open data files
  - STATA, SAS, SPSS
- Learn one of these analysis software, if possible
  - Helps speeds up the analysis, at least at the preliminary stages
  - UCSF Epi teaches STATA
  - SPSS is the most user friendly for beginners and can be learnt through online tutorials for free

# Steps from Inception to Publication



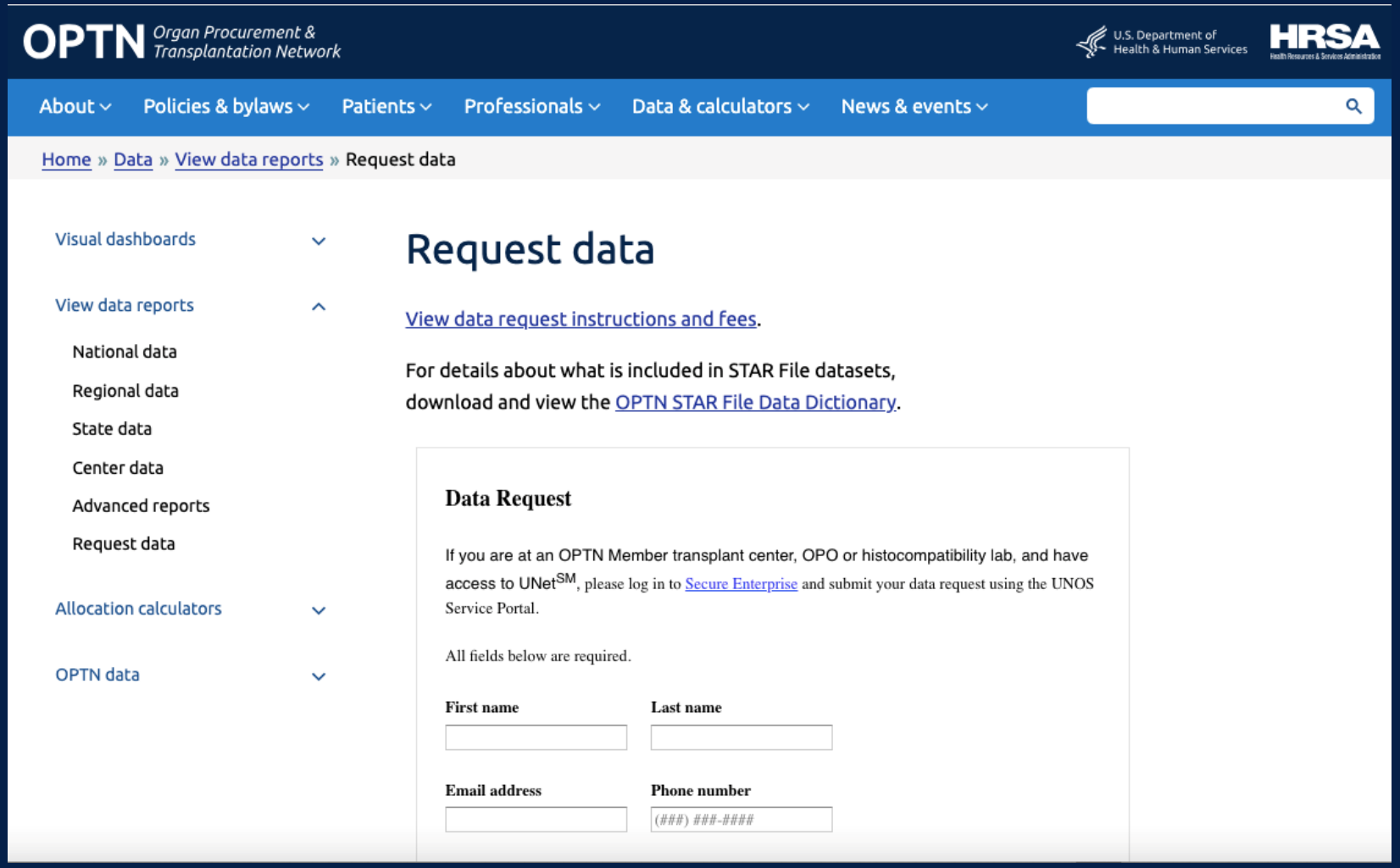
# Example of a Project Based on a Database

- OPTN: Organ Procurement & Transplantation Network
  - https://optn.transplant.hrsa.gov/data/request-data/
  - Cost: Free - $1500

The OPTN is a unique public-private partnership linking all professionals involved in the U.S. donation and transplantation system. Also, crucial to the partnership are individuals who sign organ donor cards, people who comment on policy proposals and countless volunteers who support donation and transplantation, among many others.

UCSF

# Example of a Project Based on a Database



# Example of a Project Based on a Database

- Observation:
  - Outcomes of adult lung transplantation (LTx) have been consistently reported to be better at high volume centers
  - Unclear if true for pediatric LTx
  - Pediatrics encompasses multiple age groups
    - Very small, small children, teenagers
  - Pediatrics LTx are performed in adult hospitals, large and small pediatrics hospital

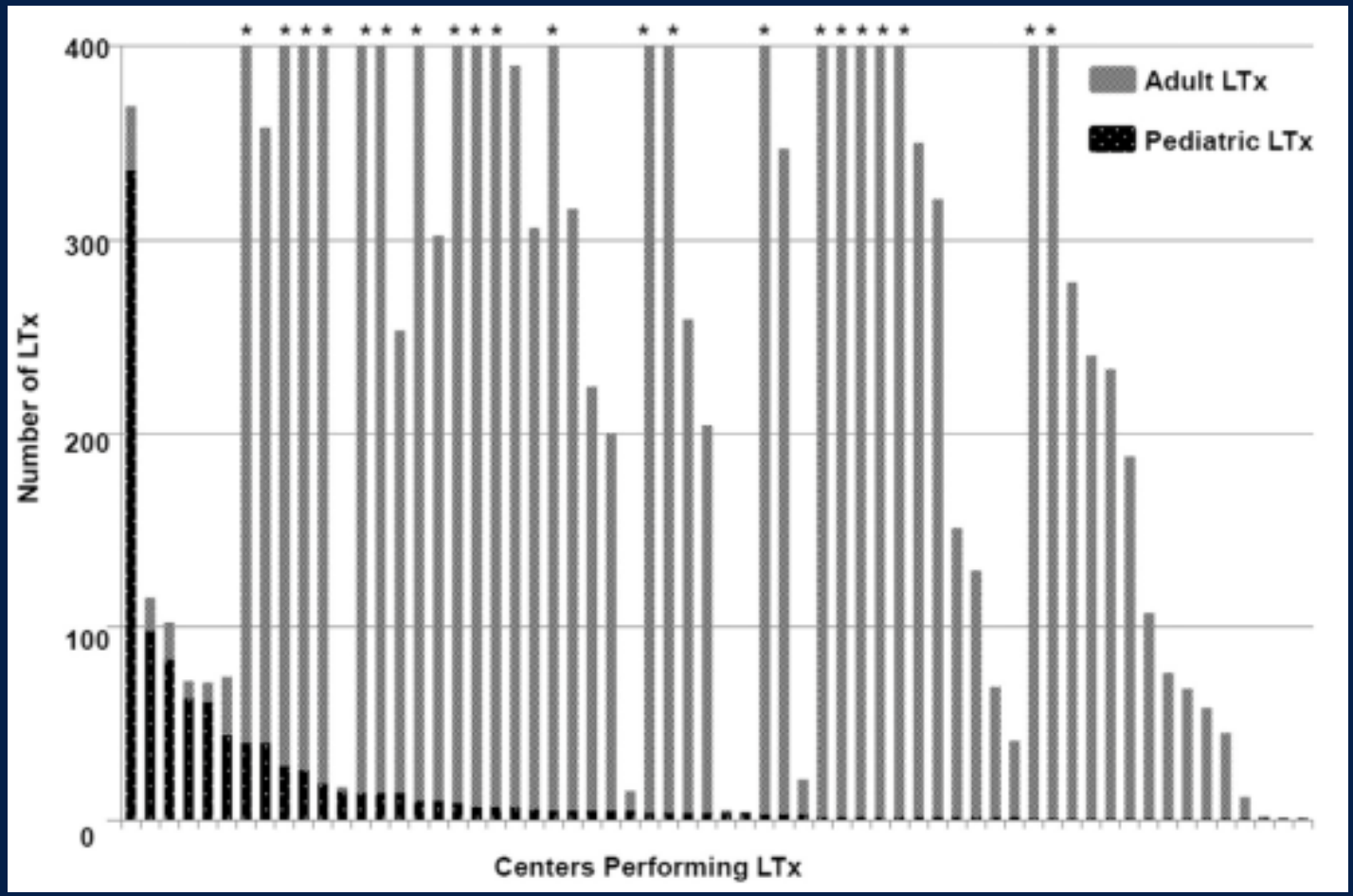
# Example of a Project Database

- Hypothesis:
  - High-volume adult centers (that perform pediatric LTx) will provide better outcomes than pediatric LTx centers, especially in older children and adolescents
  - In younger children, pediatric centers will provide comparable outcomes to adult centers, because pediatric-specific experience may be more important than surgical volume

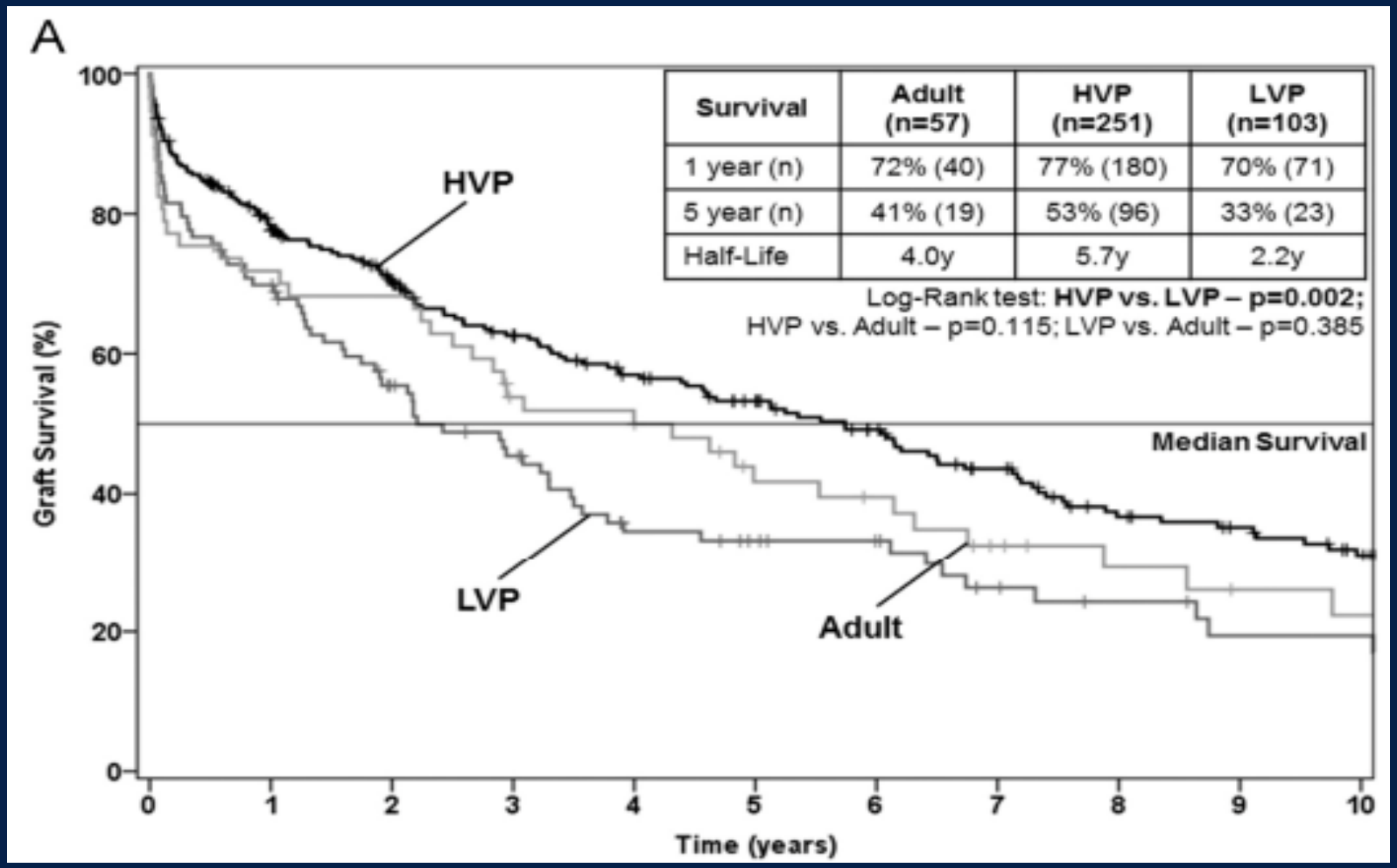
# Example of a Project Based on a Database

- Obtained the entire lung transplantation (LTx) database
- Removed unwanted data, i.e., adult LTx
- Identified isolated pediatric LTx, i.e., no multiorgan Tx
- Defined what is an adult LTx centers in the database
- Defined what is a high volume, low volume ped hospital
- Planned the type of analysis best suited to answer these questions

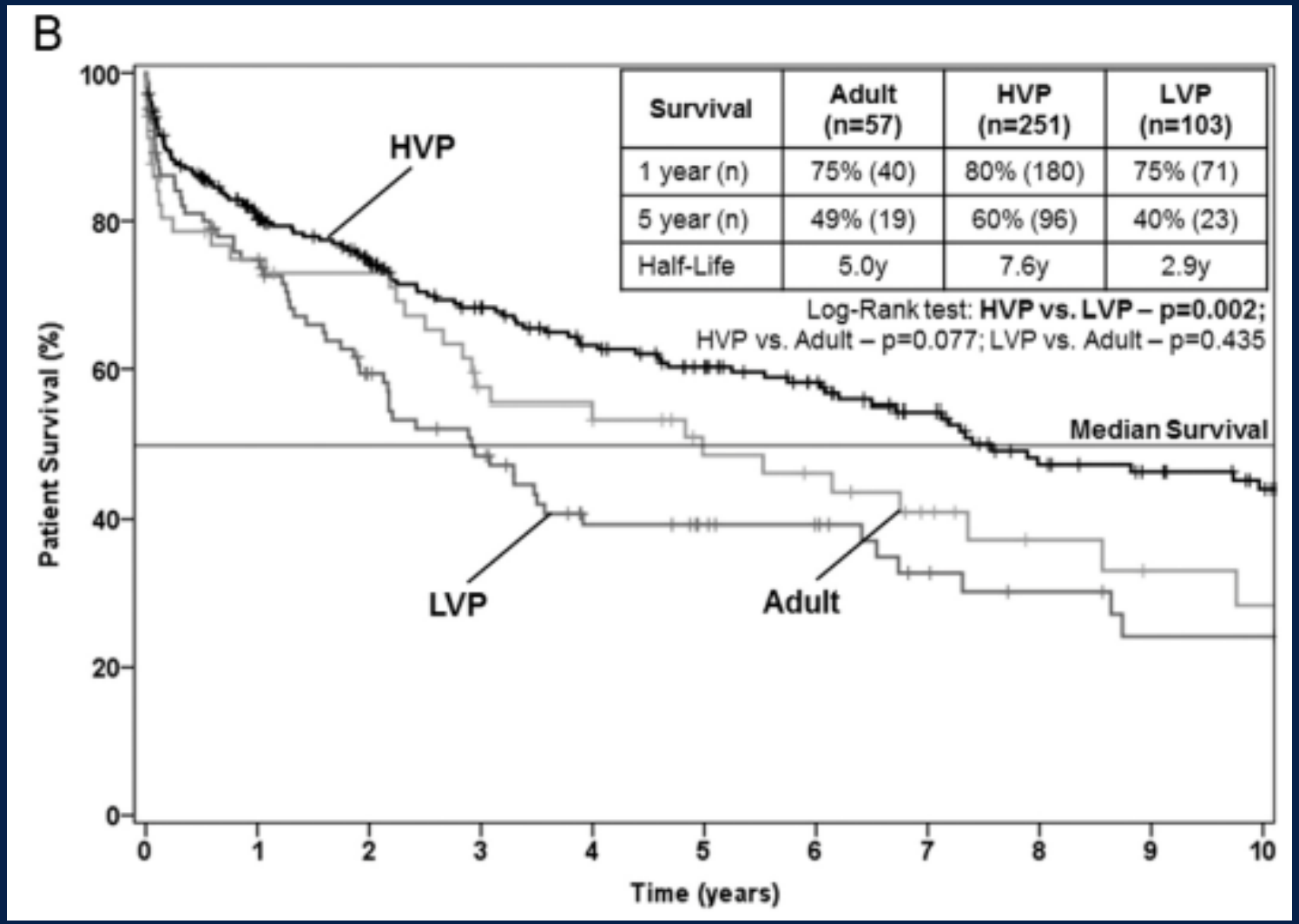
# LTx volume graphed with the type of hospital



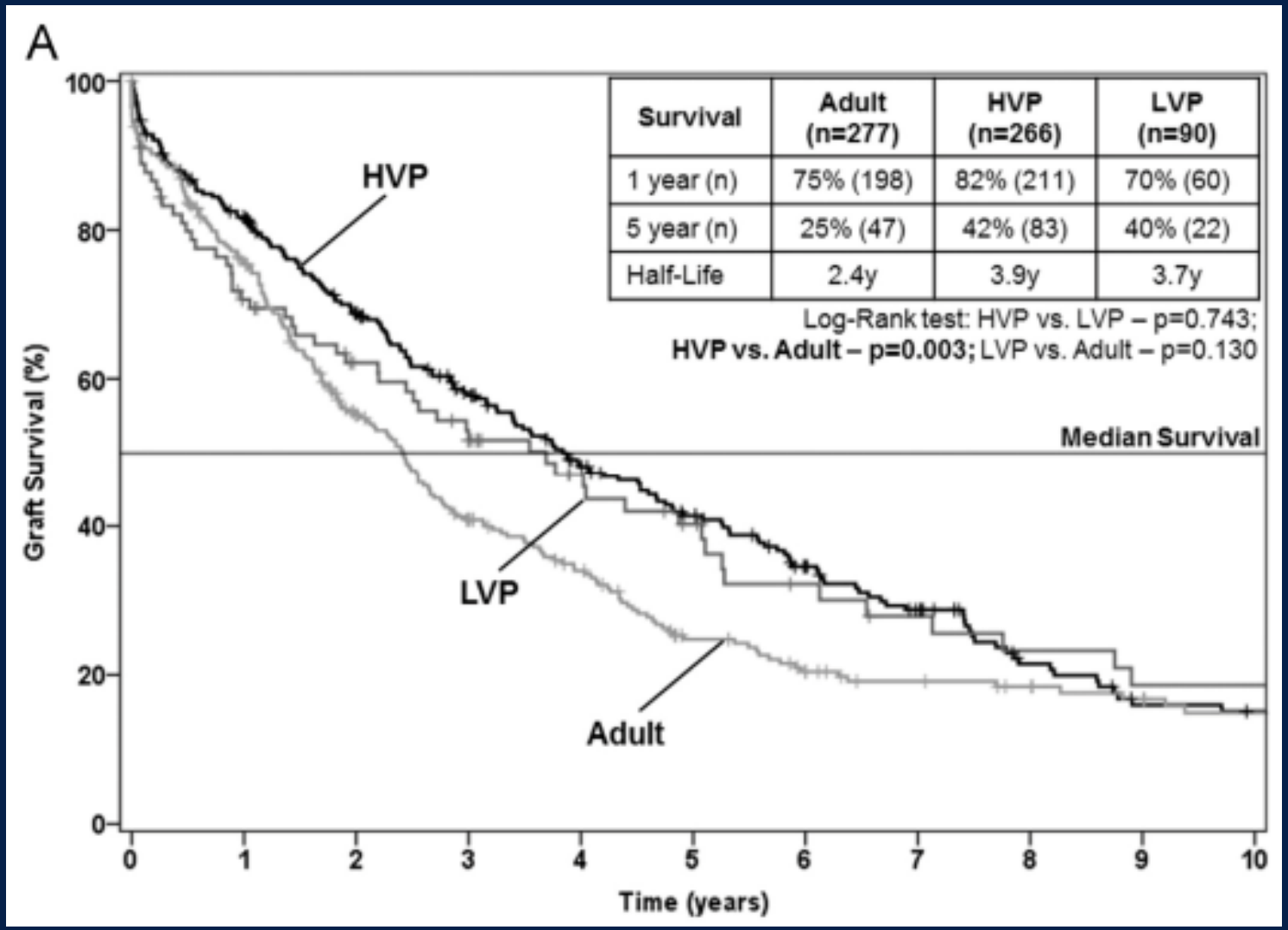
# Graft survival for patients <12 years old



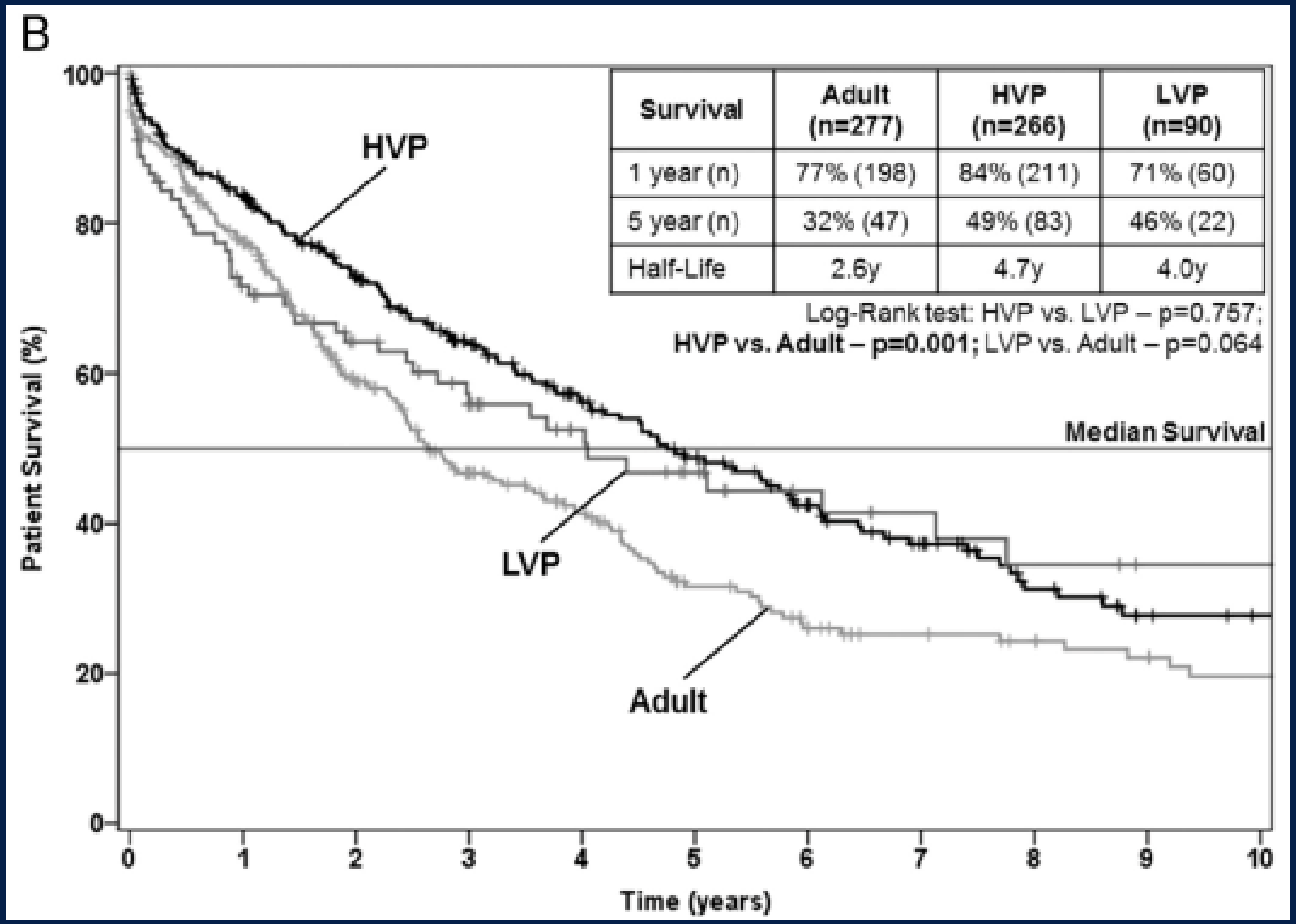
# Patient survival for patients <12 years old



# Graft survival for patients >12 years old



# Patient survival for patients >12 years old




| Survival | Adult (n=277) | HVP (n=266) | LVP (n=90) |
|---|---|---|---|
| 1 year (n) | 77% (198) | 84% (211) | 71% (60) |
| 5 year (n) | 32% (47) | 49% (83) | 46% (22) |
| Half-Life | 2.6y | 4.7y | 4.0y |

# Multivariate model for predictors of LTx graft survival

**Table 3** Multivariate Cox Regression Model for Predictors of Pediatric LTx Graft Failure Across the USA

| Variables | *p*-value | HR | 95% CI for HR Lower | Upper |
|---|---|---|---|---|
| **Adult center** | **<0.001** | **1.6** | **1.231** | **1.977** |
| **Low volume pediatric center** | **0.002** | **1.5** | **1.169** | **2.010** |
| Recipient age in years | 0.762 | 1.0 | 0.950 | 1.072 |
| Donor age in years | 0.119 | 1.0 | 0.983 | 1.002 |
| **Non-Caucasian recipient** | **0.014** | **1.4** | **1.067** | **1.760** |
| Female recipient | 0.149 | 1.1 | 0.950 | 1.401 |
| Recipient weight | 0.104 | 1.0 | 0.998 | 1.023 |
| Recipient height | 0.979 | 1.0 | 0.989 | 1.011 |
| **Cystic fibrosis** | **0.002** | **1.5** | **1.168** | **2.041** |
| Primary pulmonary hypertension | 0.942 | 1.0 | 0.697 | 1.475 |
| Retransplantation | 0.051 | 1.5 | 0.998 | 2.298 |
| **Pre-LTx mechanical ventilator use** | **<0.001** | **2.0** | **1.415** | **2.783** |
| Pre-LTx long-term steroid use | 0.199 | 1.1 | 0.927 | 1.436 |
| Pre-LTx ECMO support | 0.367 | 1.5 | 0.603 | 3.920 |
| Pre-LTx creatinine levels | 0.078 | 1.1 | 0.993 | 1.137 |
| Lung surgery pre-LTx | 0.155 | 1.3 | 0.905 | 1.878 |
| Unilateral LTx | 0.072 | 1.4 | 0.970 | 2.013 |
| Early era (1987–2000) | 0.431 | 1.1 | 0.878 | 1.357 |

Boldface highlights variables predicting pediatric LTx graft failure. CI, confidence interval; ECMO, extracorporeal membrane oxygenation; HR, hazard ratio; LTx, lung transplantation.

# Conclusion

"… outcomes of pediatric LTx in the data set used in this study proved superior in HVP centers compared with LVP centers or adult centers. Despite much larger experience in LTx as whole, outcomes in adult centers in older children and adolescents are inferior to those in pediatric centers. Therefore, not only center volume but pediatric-specific experience should be considered to have a major impact on outcomes of pediatric LTx."

# Impact factor: 13.569 (2021)

ELSEVIER

The Journal of Heart and Lung Transplantation

http://www.jhltonline.org

## Survival in pediatric lung transplantation: The effect of center volume and expertise

Muhammad S. Khan, MD,[a] Wei Zhang, PhD,[b] Rachel A. Taylor, BA,[b] E. Dean McKenzie, MD,[b] George B. Mallory, MD,[c] Marc G. Schecter, MD,[d] David L.S. Morales, MD,[a] Jeffrey S. Heinle, MD,[b] and Iki Adachi, MD[b]

*From the [a]Pediatric Cardiothoracic Surgery, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; [b]Congenital Heart Surgery, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas; [c]Department of Pediatrics, Section of Pulmonology, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas; and the [d]Department of Pediatrics, Section of Pulmonology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio.*

J Heart Lung Transplant 2015;34:1073–1081

UCSF

# Summary

- Databases/big data provide an efficient and important way to answer important clinical questions
- By mindfully developing a research question and obtaining the database best suited to answer that question, you can create a high-quality product in a relatively short amount of time
- Crucial to understand the limitations of all databases, and your own skill sets
- Skills grow and process improves with repeated projects and working with experienced mentors