Data Science

Hye-Chung Kum

Population Informatics Research Group

http://research.tamhsc.edu/pinformatics/

http://pinformatics.web.unc.edu/

Course URL:

http://pinformatics.tamhsc.edu/phpm672

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Properties of BIG DATA : 4V

- Volume : constantly generating
- Velocity : constantly changing
- Variety : expressed in many ways
- Veracity : lots of errors

EXAMPLE: the INTERNET!
What do you do to find information/knowledge on the Internet?

The Big Data Problem – Nutshelled Michael Franklin (UC Berkley)

Something's gotta give:

AMPLab: Integrating Three Key Resources

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines

- Clusters and Clouds
- Warehouse Scale Computing

People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts

POPULATION INFORMATICS RESEARCH GROUP

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

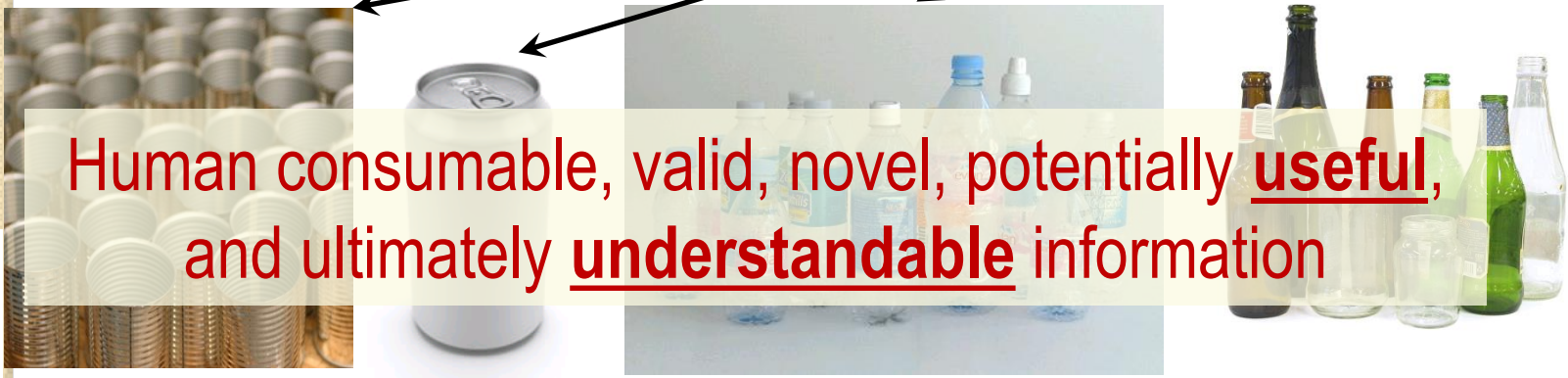
What is Data Science?

- Other words
  - Knowledge Discovery & Data mining (KDD)
  - Business Intelligence / Business Analytics
- Collecting and refining information from many sources
- Analyzing and presenting the information in useful ways
- So people can make better business decisions

Data Science
Knowledge Discovery & Data mining (KDD)

Big Data :
operational data

KDD
Clean, Merge, Reprocess

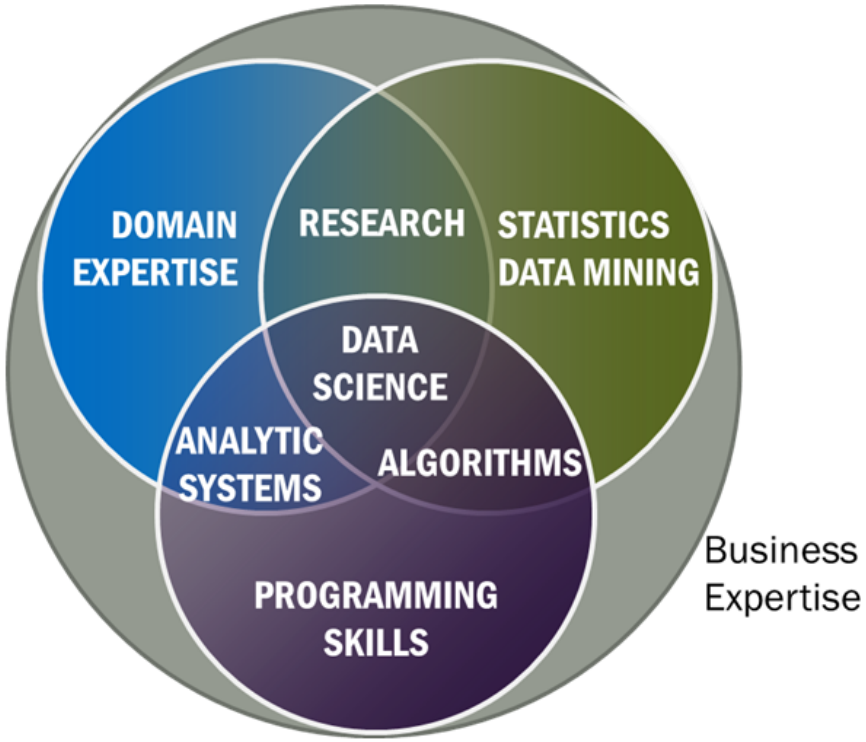


Human consumable, valid, novel, potentially **useful**, and ultimately **understandable** information

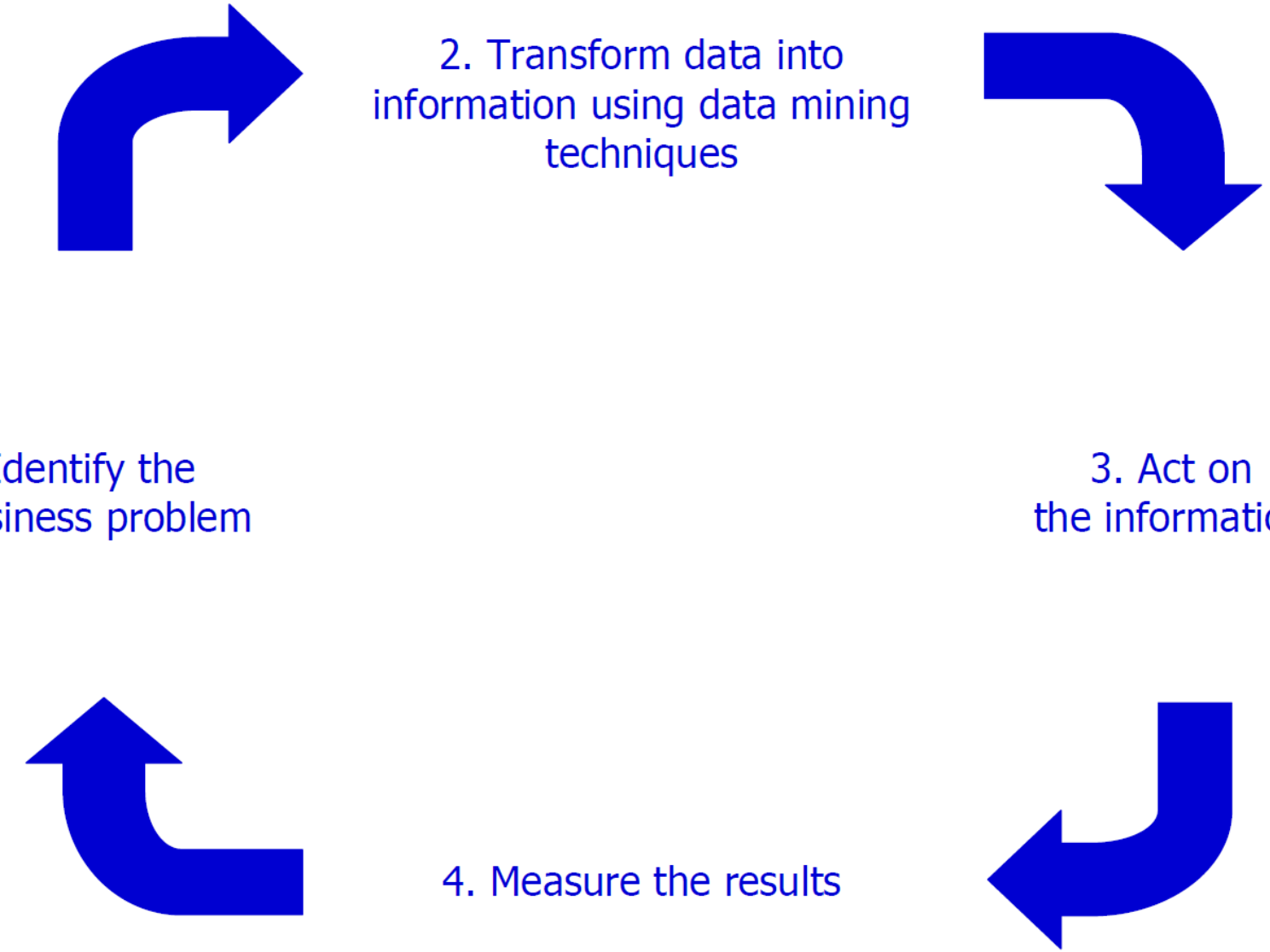
Data Science Definition (Big Data less consensus)

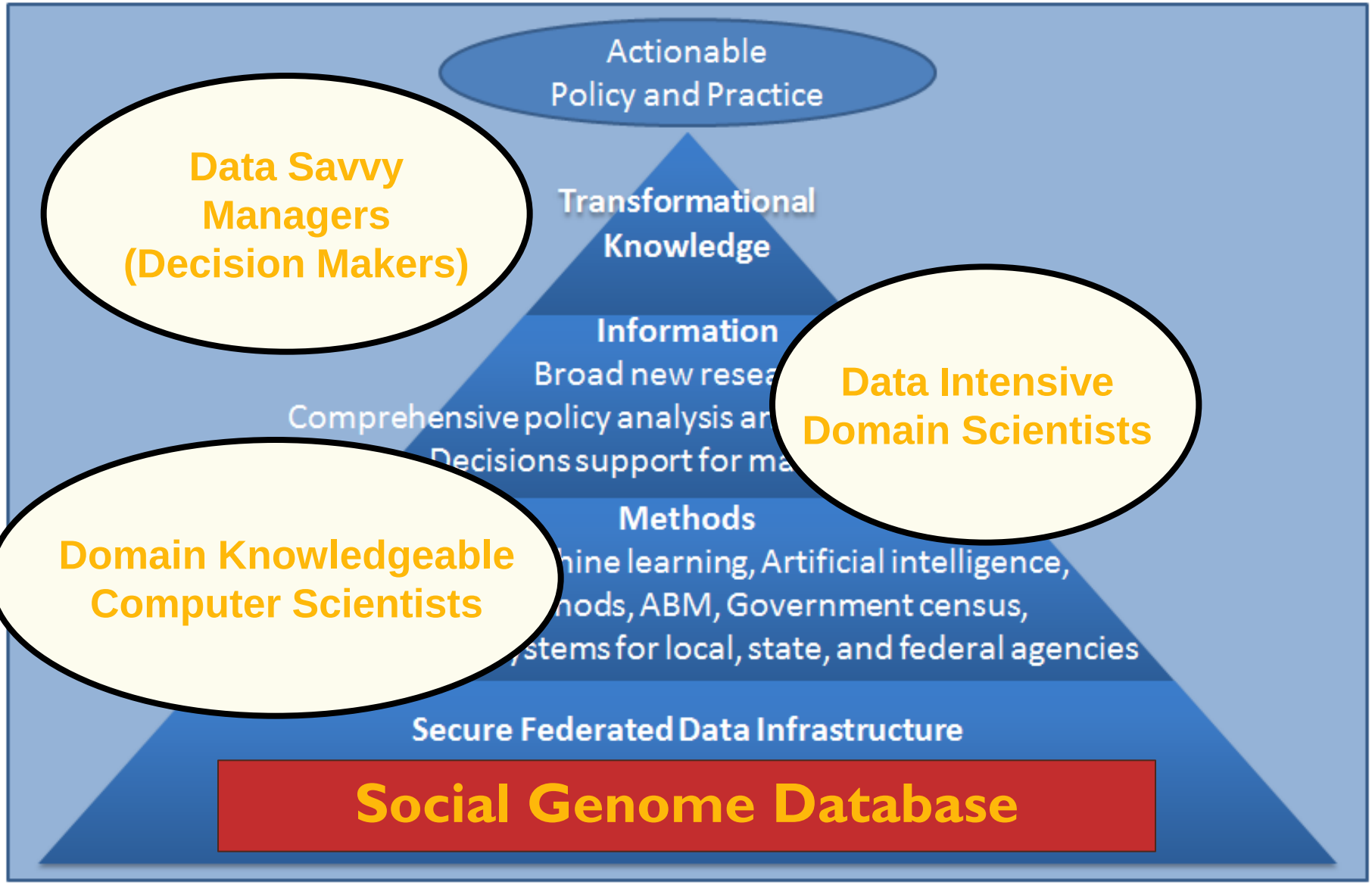
- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.
- A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, 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



The Virtuous Cycle of Data to Decision & Action





UNC POPULATION INFORMATICS RESEARCH GROUP

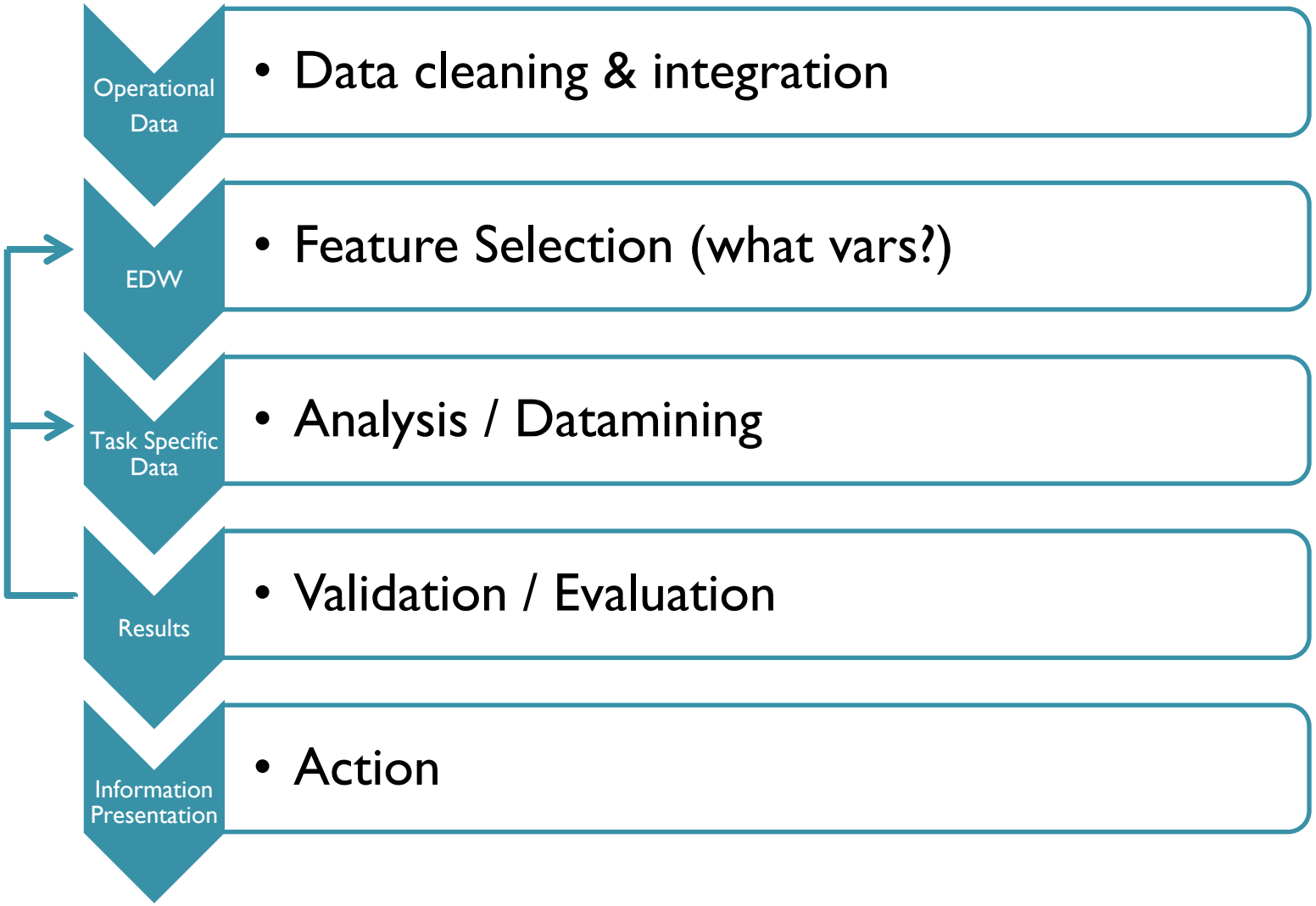
Population Informatics

Hye-Chung Kum, Population Informatics Research Group
Dept. of Computer Science, UNC-CH, http://pinformatics.web.unc.edu/

Social Genome

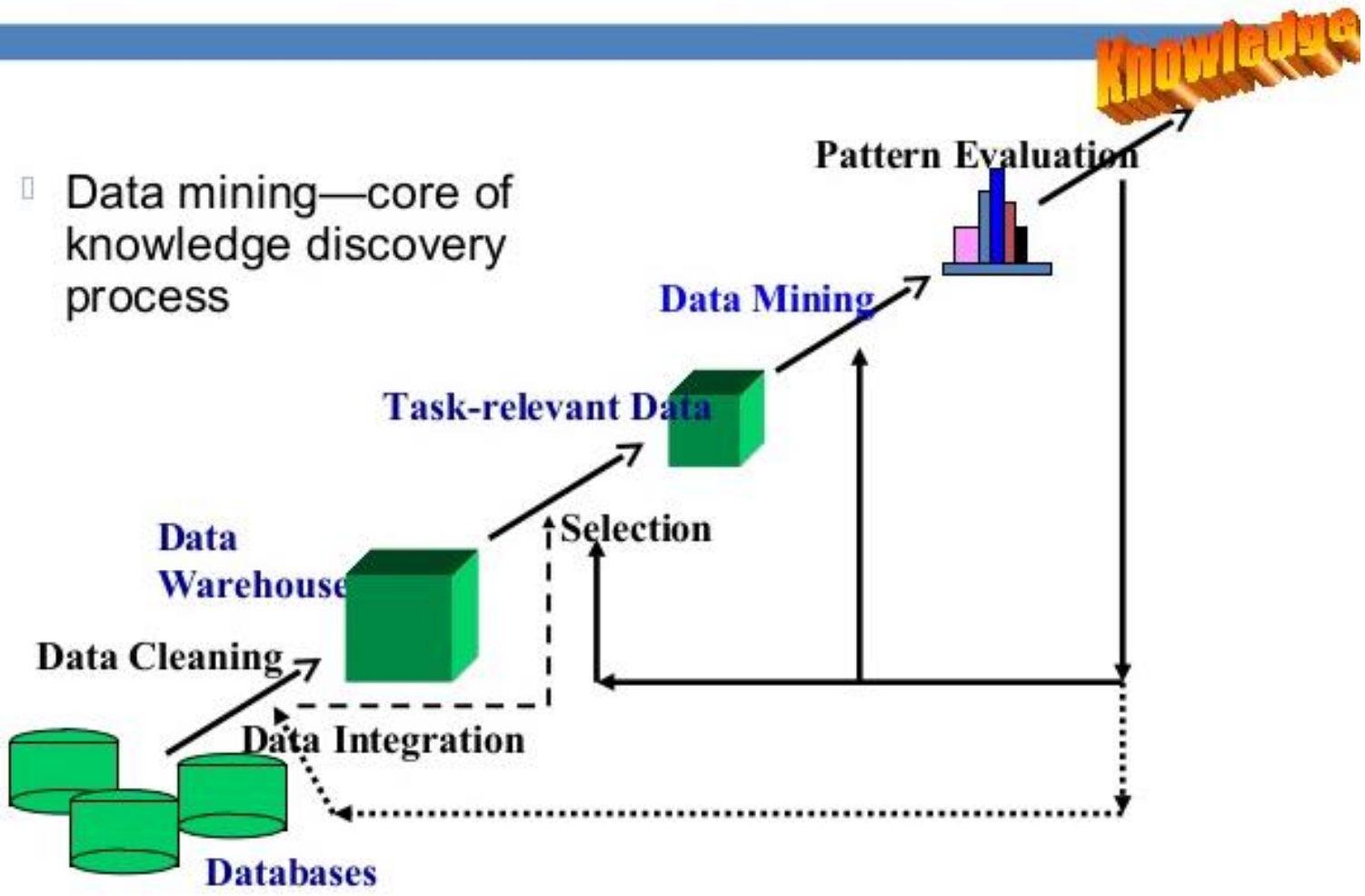
- Data-intensive research using distributed, federated, person-level datasets in near real time has the potential to transform social, behavioral, economic, and health sciences—but issues around privacy, confidentiality, access, and data integration have slowed progress in this area. When technology is properly used to manage both privacy concerns and uncertainty, big data will help move the growing field of population informatics forward.

KDD Process



KDD Process

- Data mining—core of knowledge discovery process



Thomas Davenport Competing on Analytics

- Skill set for good data scientists
  - IT & Programming skills
  - 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"

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.
- Health is a very important domain
  - Team lead: good questions, good interpretation & implications
- http://radar.oreilly.com/2011/09/building-data-science-teams.html

Public Health Informatics

- Yasnoff et al. (2000): Public health informatics is defined as the systematic application of information and computer science and technology to public health practice, research, and learning

Public Health Informatics Competencies

- Public Health Informatics Competencies Working Group, CDC
- Report: Informatics competencies for public health professionals
  - *Class 1: Effective use of information*
  - *Class 2: Effective use of information technology*
  - *Class 3: Effective management of information technology projects*
  - http://www.nwcphp.org/docs/phi/comps/phic_web.pdf

Public Health Informatics Competencies

- Describe at a basic level the fundamentals of a computer network
- Describe at a basic level the Internet and World Wide Web
- Describe at a basic level technologies employed to ensure computer systems' security
- Public Health Informatics Competencies Working Group

- Cunningham et al. 2007. Baseline Assessment of Public Health Informatics Competencies in Two Hudson Valley Health Departments. Public Health Reports. May-Jun 2007. V122

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

- "All models are wrong but some are useful"
- The Petabyte Age is different because more is different
  - Google translate
- Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. Correlation is enough.

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

- But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.
- Biology: Hutchinson disease
- Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.
- What can science learn from Google?

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

There are ways to do analytics right Value in Big Data

- 63 percent of healthcare executives in the federal government believe that big data will improve population health management.
- The Mckinsey report on big data valued integrated data on patient services at $300 billion.
- But so few have proper goals and strategies for their data

There are ways to do analytics right It’s not the data, but the people

- It's not about the data, but how you're going to manage it.
- The focus, according to Hughes, should be on information management, including data governance, stewardship and quality.
  - If you are just about grabbing data, you will be on a data grab forever.
- It may sound impressive to say that your organization has access to terabytes of patient information, but without robust technology and smart people to manipulate it, that data is simply words and numbers without context
- A severe shortage of analytics pros makes navigating this landscape all the more difficult
- “It's also a mistake to think you can staff up on this easily"
- lack of qualified data engineers

There are ways to do analytics right Data Management is Key

- Raw data from claims or from an EMR database are not suitable for analysis.
- Turning raw data into usable information requires preparation, including normalization and validation.
- Only then can an organization gain trustworthy insights from the information and put it to use in maximizing patient care, reducing risk and strengthening a business's bottom line

There are ways to do analytics right EDW: Organized Data Library

- Hughes says that organizations have been spending too much time and money on enterprise data warehouses, which he sometimes refers to as "data landfills."
- An EDW isn't where data goes to die. An EDW is a staging point for analytics.
- An EDW needs to be easy for clinicians to understand and interpret, and also needs to interoperate with and push data back out to other systems
- Sometimes this is done in too fragmented a fashion
- My thoughts: Art
  - Appropriate size task, data
  - Balance organizing data with actual using data

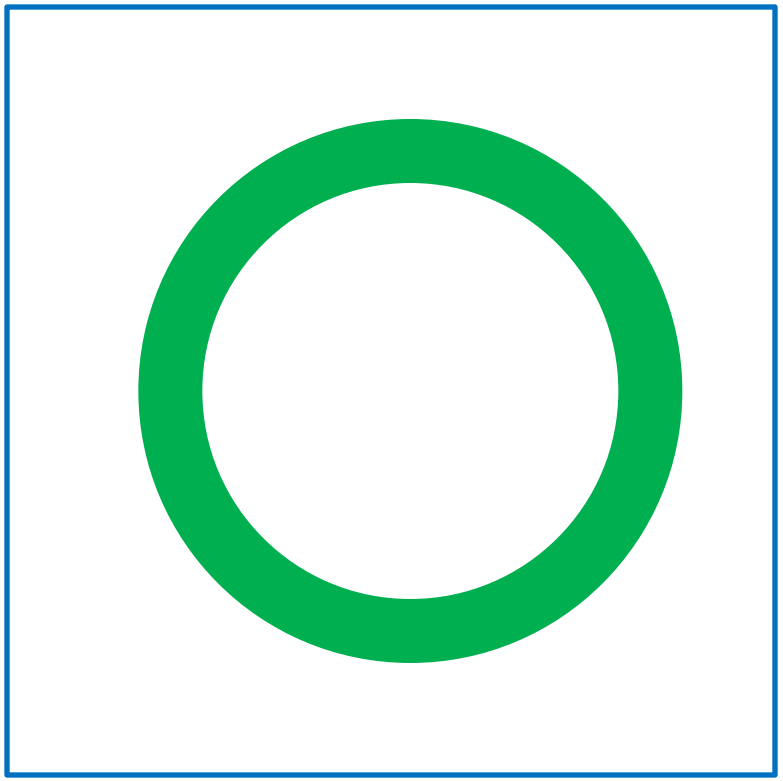
Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

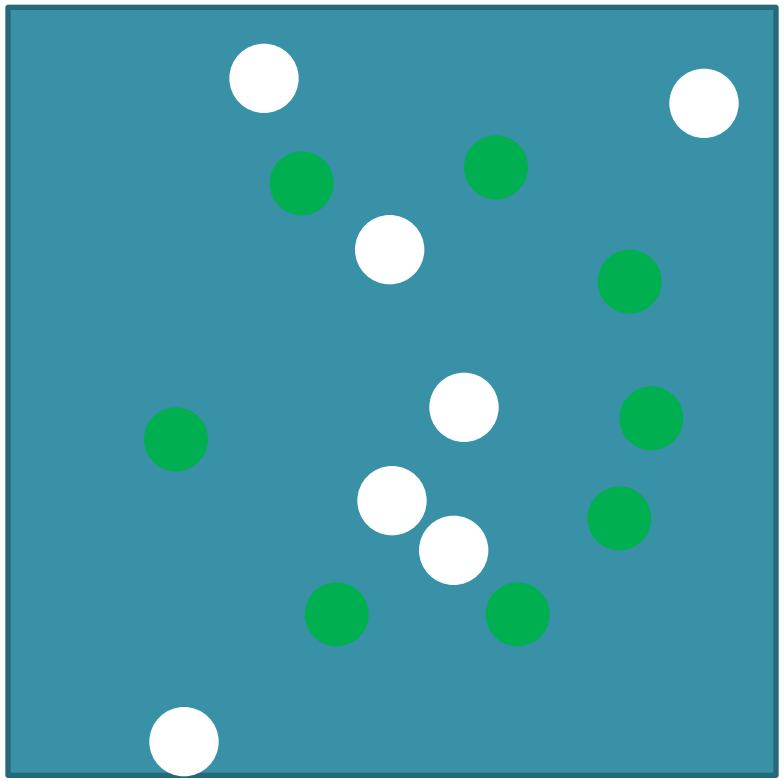
Inflation:
Traditional social science vs Data Science

- Consumer Price Index (CPI)
  - Representative basket of goods and services purchased for consumption by urban households (monthly)
  - This index value has been calculated every year since 1913
  - Bureau of Labor Statistics
- Billion Prices Project : MIT
  - The Billion Prices Project is an academic initiative that uses prices collected from hundreds of online retailers around the world on a daily basis to conduct economic research.
  - Pricing Behavior, Daily Inflation and Asset Prices, Pass-Through (price and exchange rate and international rate), Green Markups (premium for green prod.)

What is the shape of the green line?



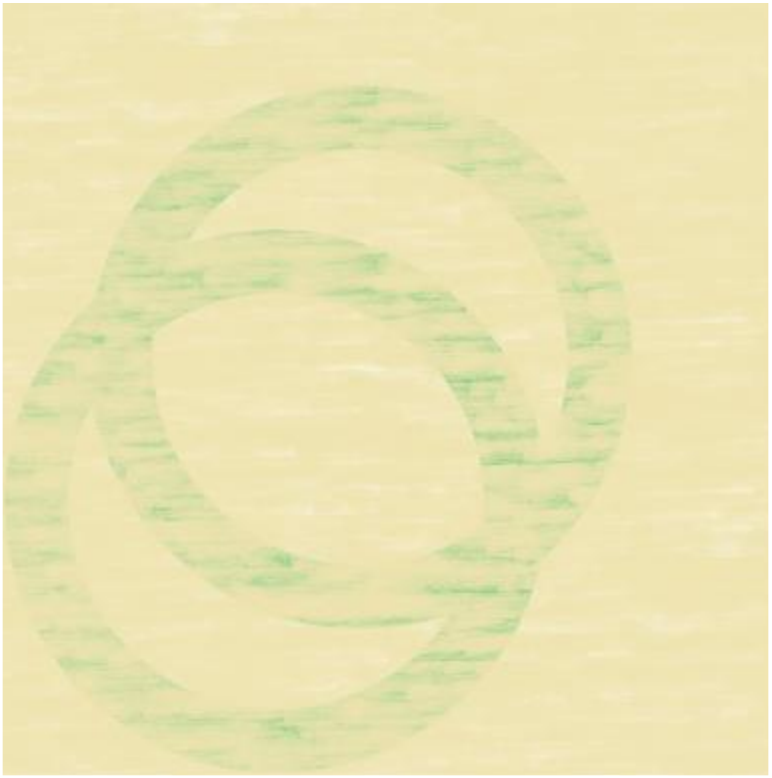
Traditional Science :
Start with nothing – collect data well



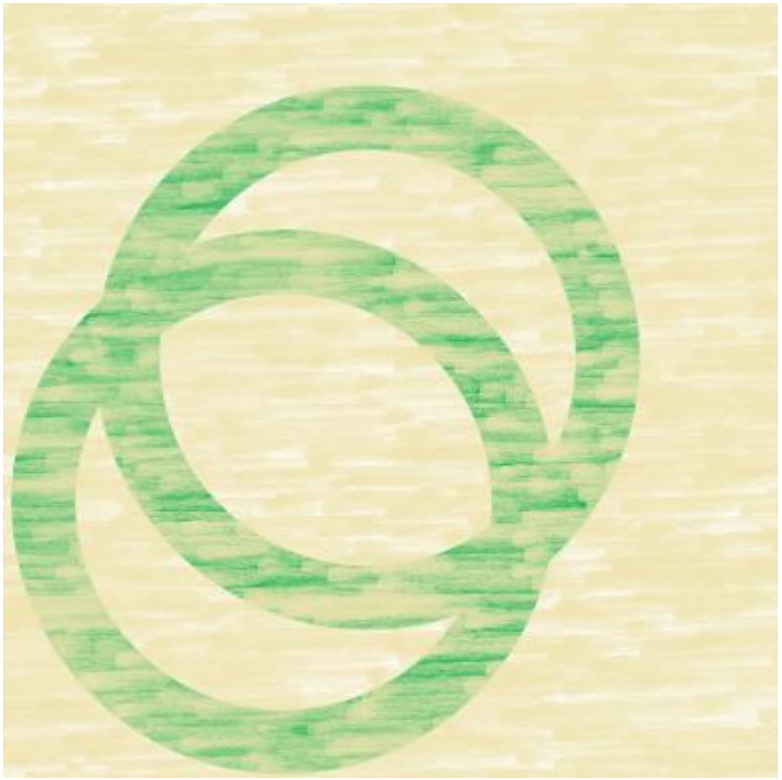
Data Science: EVERYTHING

POPULATION INFORMATICS RESEARCH GROUP

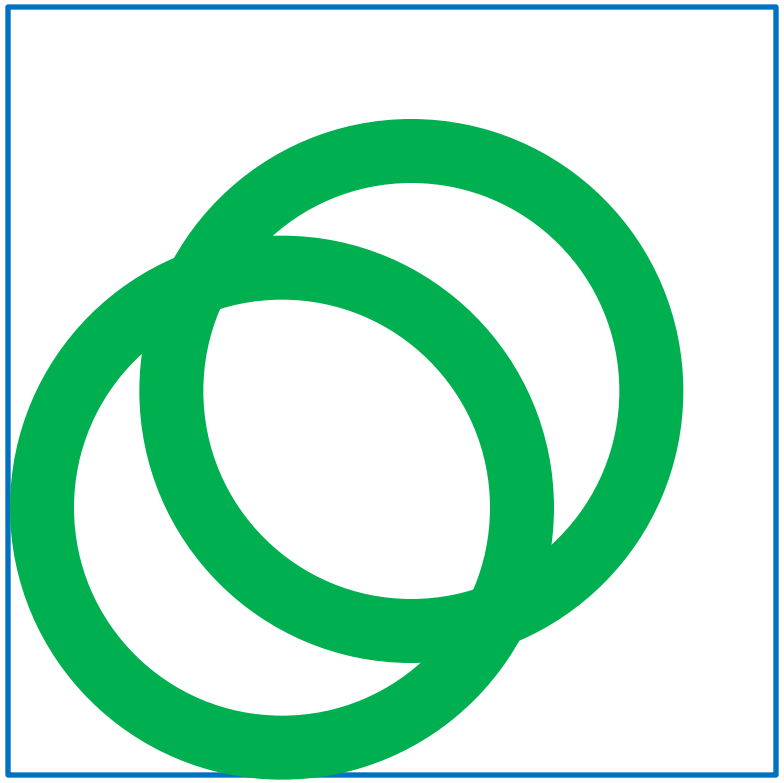
First, separate out only the relevant data



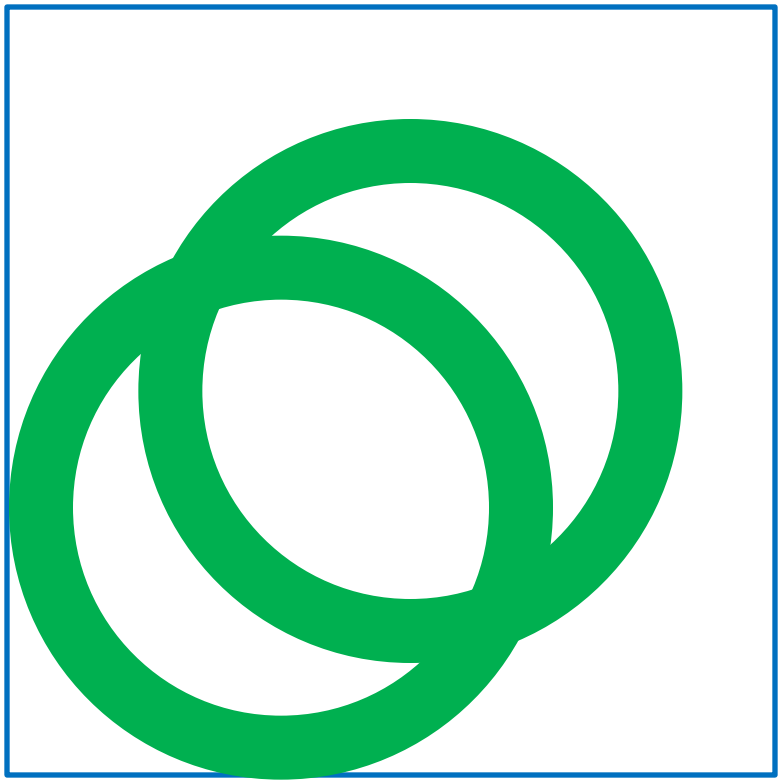
Second, clean noise as much as possible



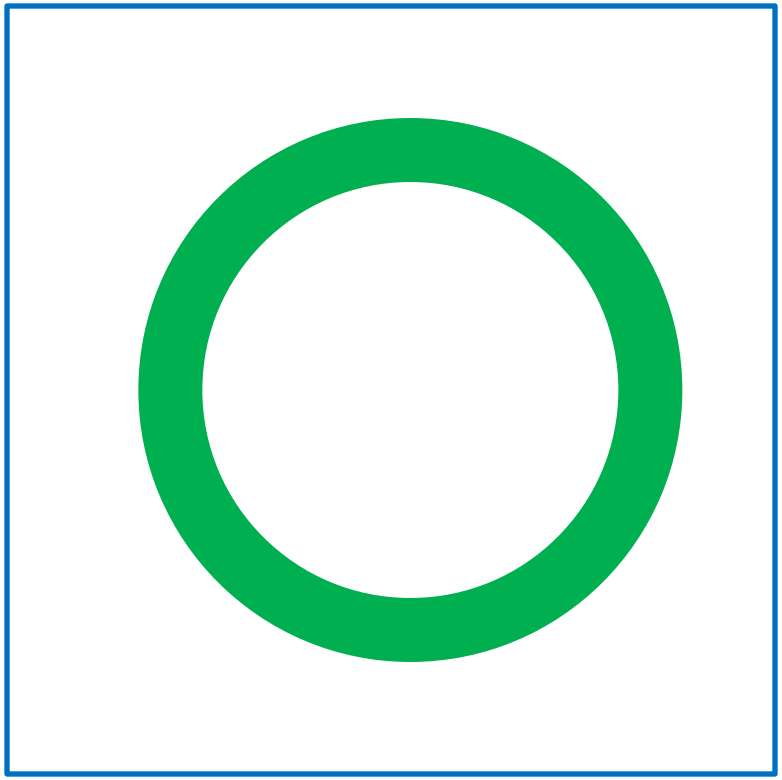
Third: Model



Fourth: Validate to avoid overfitting



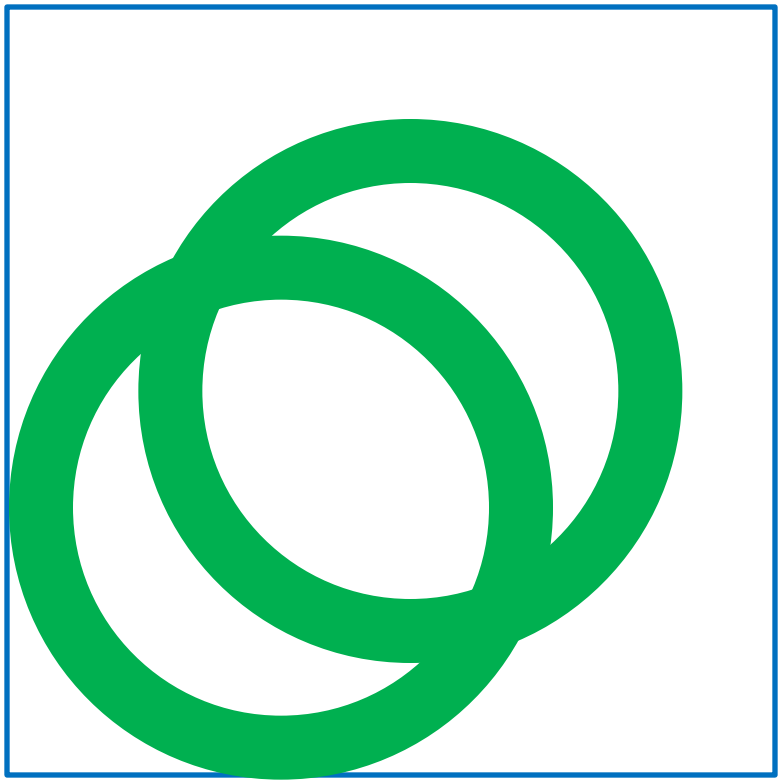
Model



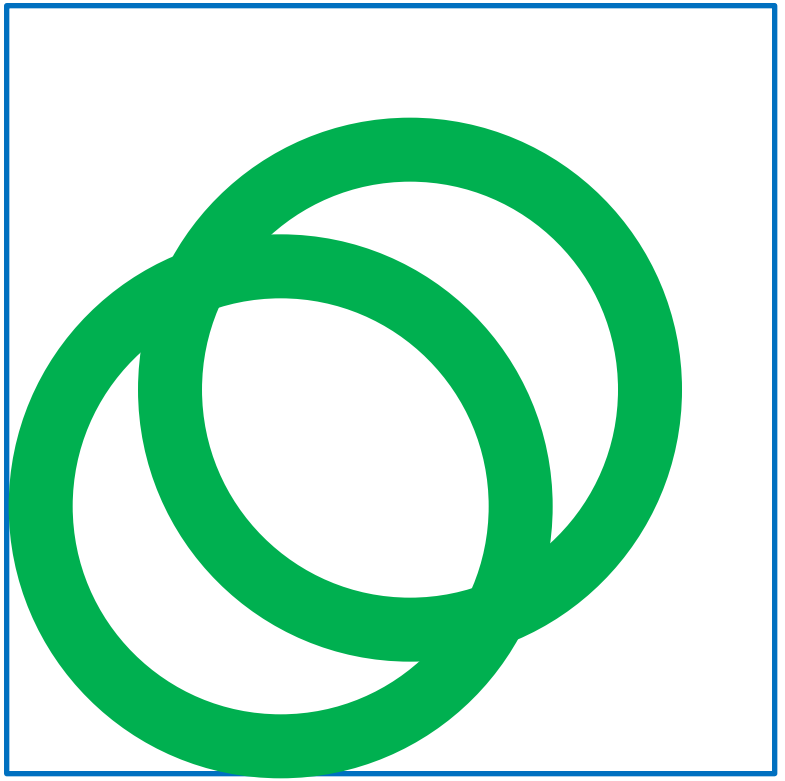
Validate

Population Informatics Research Group

Sometimes models differ between the two approaches. Why ?



Model



Validate

Points to look out for

Traditional Approach

- Use statistics to model from the data points
- Measurement
  - With out seeing the other colors
  - Based on theory decide to measure green only
- Measurement Error
  - Reduce by designing well
- Bias : Random Points
- Are there enough points?

Data Science

- Use statistics to model from the data points
- Measurement
  - You can see the other colors and compare
    - Is this fishing for results?
- Measurement Error
  - Know what it is, adjust for it
- Bias & Over fitting: validation, know the bias
- Is the data clean enough?
  - Sensitivity analysis

Video

Tips

- Autoexec.sas
- Config.sas
- Lib.sas

Bonus: Feedback Email TA

- Have you learned much in this class?
- If not, why do you think?
- If so what? What is the most helpful ?
- Any other thoughts for improving the class you want to share?
- Your name, if you are willing to be open
  - Because I know your level of programming, it helps to put your comments into level of programming skills

Where are we now

- Assignment 1
  - Setup work environment
  - Use the SAS software
  - SAS programming basics
    - data step & proc step
    - Libname (where is the folder)
    - Writing code & Reading logs
- Assignment 2
  - Understand variables (names, types, labels)
  - To write conditional logic codes
  - Subset columns (variables) from a table
  - Subset rows (observations) from a table
  - Recode, rename variables and calculate new variables
  - Label variables and values
- Assignment 3
  - use for loops (iterative loops)
  - use while loops (conditional loops)
  - SAS: use one dimensional arrays
- Assignment 4 & 5
  - Concatenate multiple tables (more rows)
    - using set
  - Link up multiple tables using a shared key (more columns)
    - using merge
    - What is a 1-to-1 link
    - What is a 1-to-N link
    - What is a N-to-N link
  - Combine multiple rows into one row
    - by group processing
    - proc summary
  - Reshape table to flip rows & columns
    - proc transpose
- Assignment 6
  - Read Macros

Grades

- Midterm Average: 16.5/20=83%
- Overall (56%)
- Assignment 5: 4 pts
- Assignment 6: 8 pts
- Final Project: 32 pts

Final Project

- Based on assignment 4
  - More variables
  - If you are not sure about SAS
  - 1 person
  - www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm
  - The Public Use Data File (PUDF) for 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006 and 2007
  - User Manual: note could be different per year
- Own project: 1 or 2
  - Talk to me before next week

Final Project

- Milestone 1: Due in One week
- List of datasets
  - List of questions: Need to be enough (will approve)
- Milestone 2: Due in One week
  - Program (does not need to be completed)
- Milestone 3: Due in One week
  - Presentation (12 minutes + 3 minutes for Q)
  - Write up: Not formal paper
    - Hypothesis
    - Describe Data & Method
    - Questions & Answers
  - Celebrate!
    - Suggestions for catering lunch (Jason's)

Final Project: Milestone 1

Readme file: Answer the following questions in a text file called readme.txt

<County Year Analysis>

How many total observations do you have when the merge is complete?

How many observations are missing patient data?

How many counties in your dataset are missing patient data for AT LEAST 1 year? (notice, we want the number of counties, not the number of observations that are missing patient data) ?

What is the average size (in terms of the population variable) of all observations (county/years) that have missing patient data?

Which observations (which county/years) have more deaths than inpatient discharges? How many observations have more deaths than physicians ?

How does the Ratio of DPC per 100,000 Population variable differ by border and non-border counties; that is, which has a higher average Ratio of DPC per 100,000 Population?

<PHR Year Analysis>

Compare the average number of physicians and the average number of patients across all PHRs for each year. Which PHR has the highest average physician to average patients ratio in each year? (Hint: you will need to calculate this ratio before running a command to compare across PHRs)

<MSA Border Analysis>

How many MSAs are in the border region?`

Which MSA had the largest increase in patient discharges in 2011? What about 2012? (Hint: one way to do this would be to sort your data before printing)

Final Project PHPM677

- Milestone 1 (drafts): 4
- Milestone 2 (drafts): 4
- Milestone 3: 24
  - Program (final): 12
  - Presentation: 12
    - Hypothesis
    - Describe Data & Method
    - Questions & Answers

- 4 point scale
  - Really good
  - Good (75%)
  - OK (50%)
  - Bad

Assignment 6

- Objective
  - Read and write SAS macro variables
  - Read, use, and modify SAS macro functions
- Change
  - %concatI(dc, 20101, 20124, 2)
  - %concatM(dc, 201001, 201204)
  - %concatMC(dc, 201001, 201204)

Review

- Lab 6: TA?
- Macros in Assignment 4
- What would be useful?

Macros

Agenda

- Data Science
  - What is Big Data
  - What is Data Science/Population Informatics?
  - Data vs Theory
  - Doing Analytics Right
- Traditional Statistics vs Data Science
- Examples of population informatics
  - Management Decision Support
  - Evaluation
  - Research

Data to Information for …

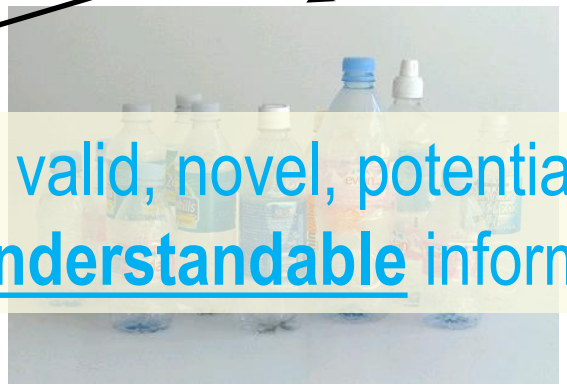
- Management decision support
  - communicate with stakeholders
  - State / counties
- Evaluation
- Research

Data Science
Knowledge Discovery & Data mining (KDD)

Big Data

KDD

Clean, Merge, Reprocess



Human consumable, valid, novel, potentially **useful**, and ultimately **understandable** information

KDD Architecture for SW Administrative Data

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

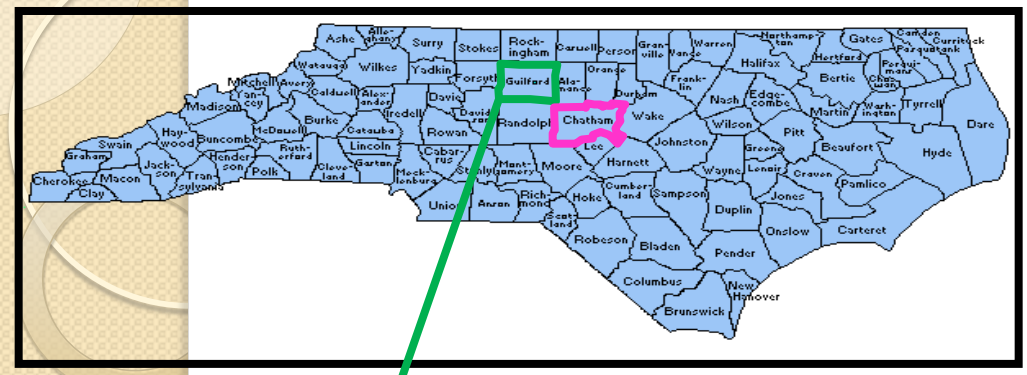
UNC-CH School of Social Work & NC DSS

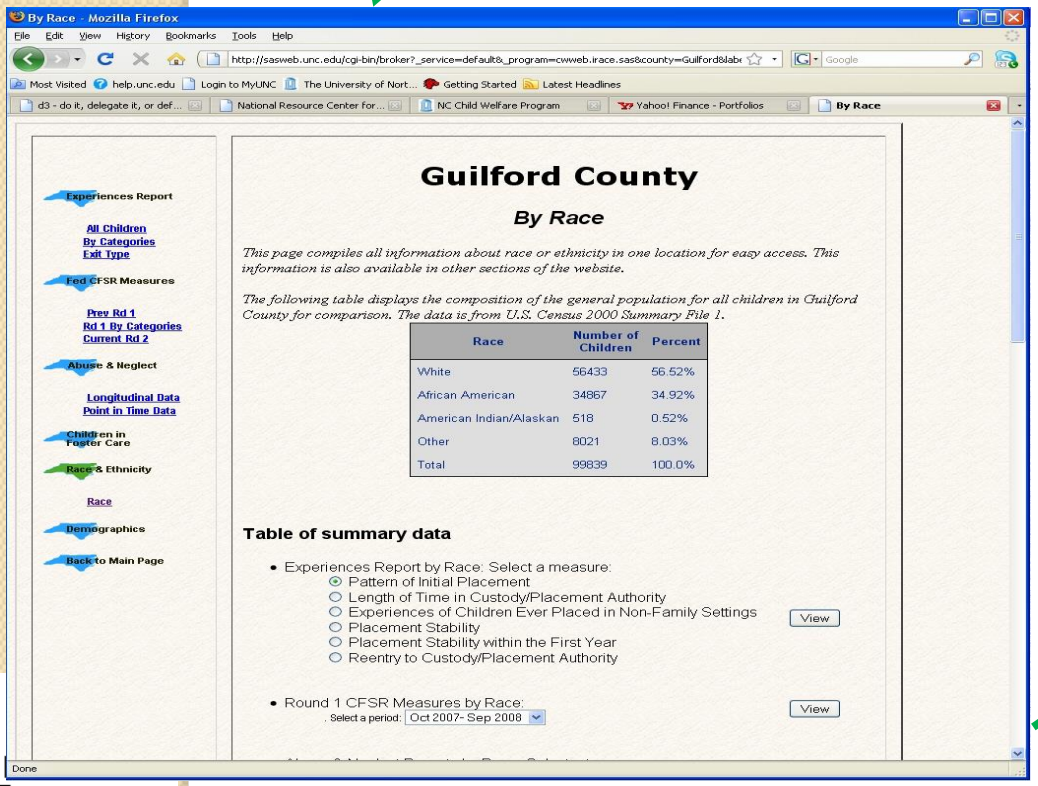
Jordan Institute for Families, http://ssw.unc.edu/ma

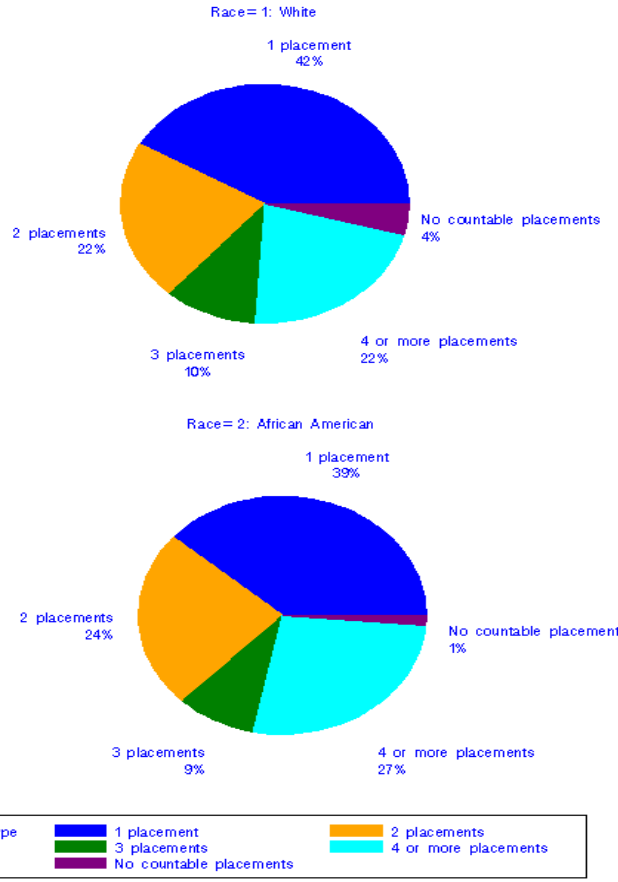
Duncan, D., Kum, H., Flair, K., & Stewart, C.J. (2009). Management Assistance for Child Welfare, Work First, and Food & Nutrition Services in North Carolina. Jordan Institute for Families website: http://ssw.unc.edu/ma

KDD system

- Easy to use and maintain
- Timely information
- Difficulties :counting!
  - What to measure? How to measure?
  - Measure child maltreatment in county ?
  - When you have a solid count of important things, easy to apply sophisticated statistics on the count
- IT Difficulties
  - Project management : testing, help pages
  - Data management





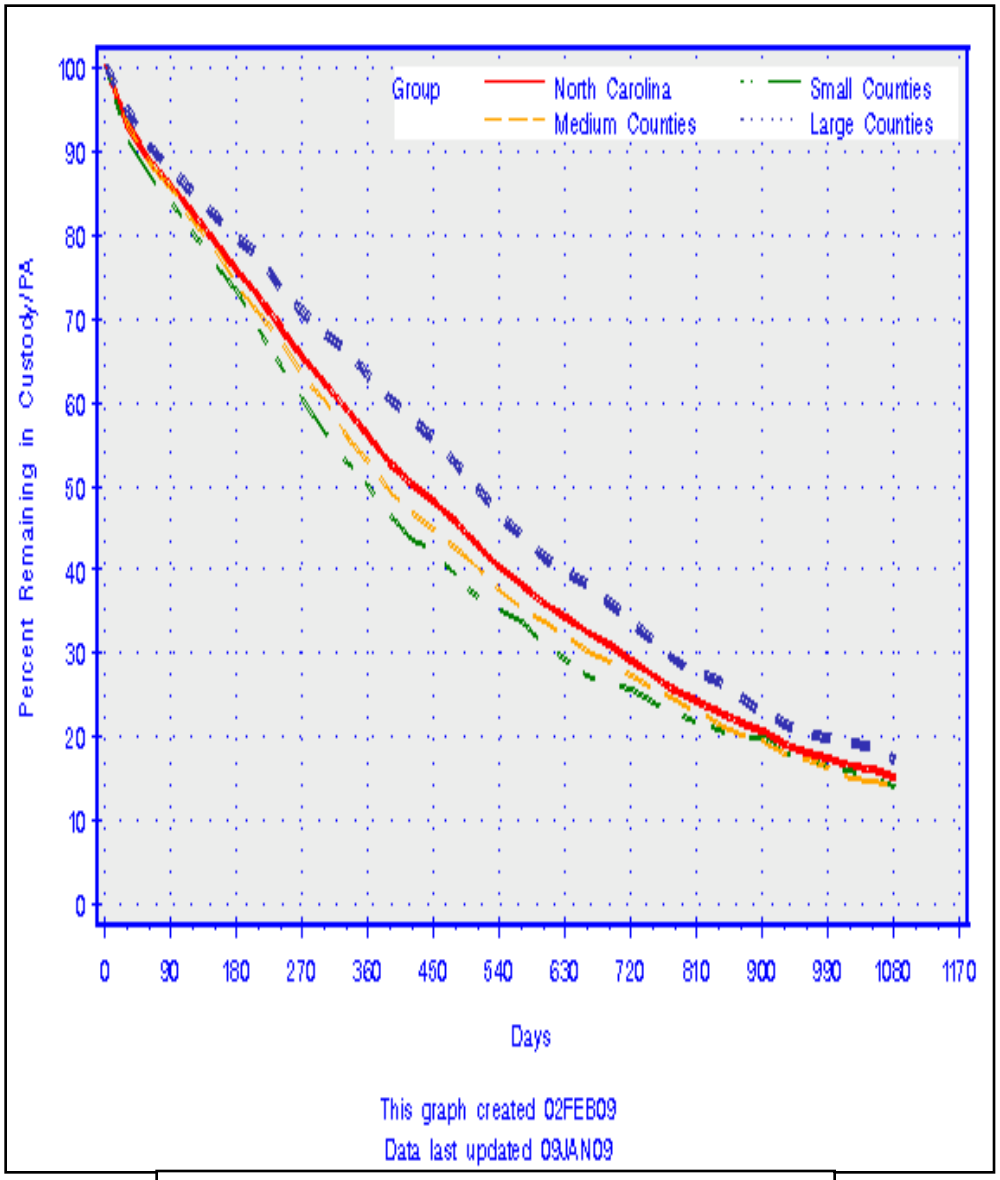


•Placement Stability for SFY03_04 Cohort
•In Guilford County by Race

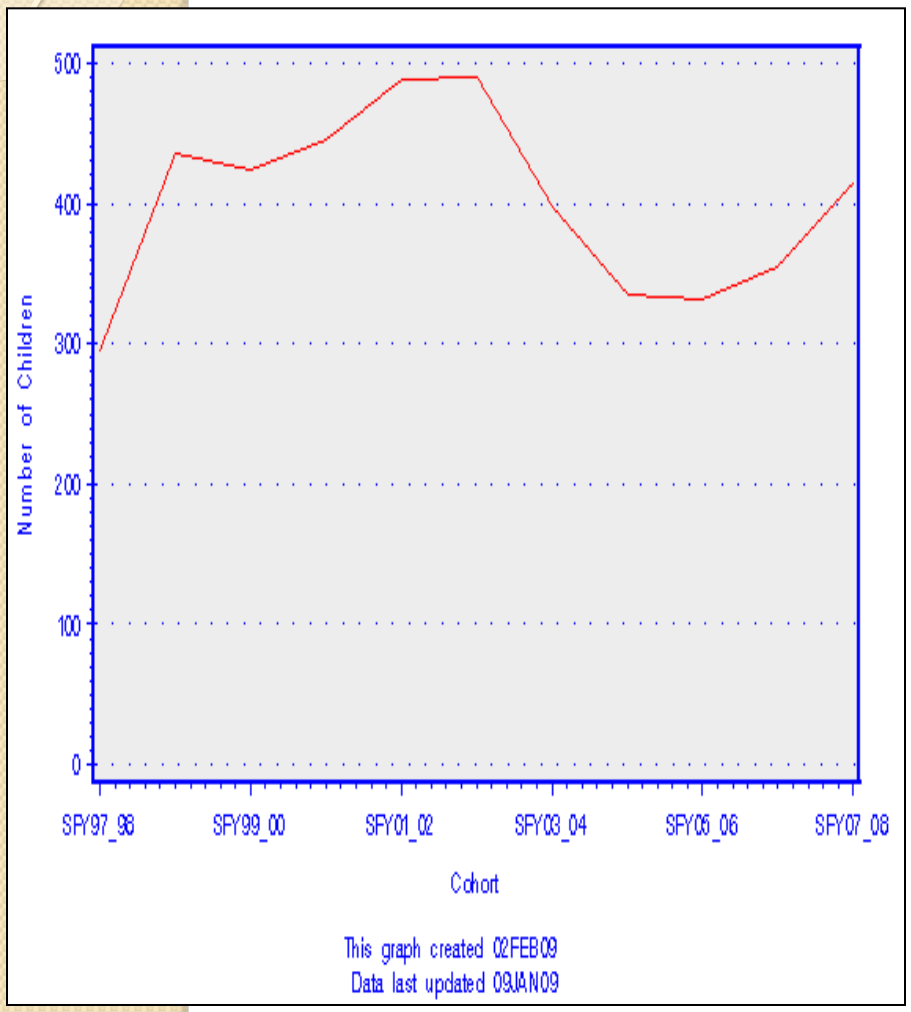
INFORMATICS RESEARCH GROUP

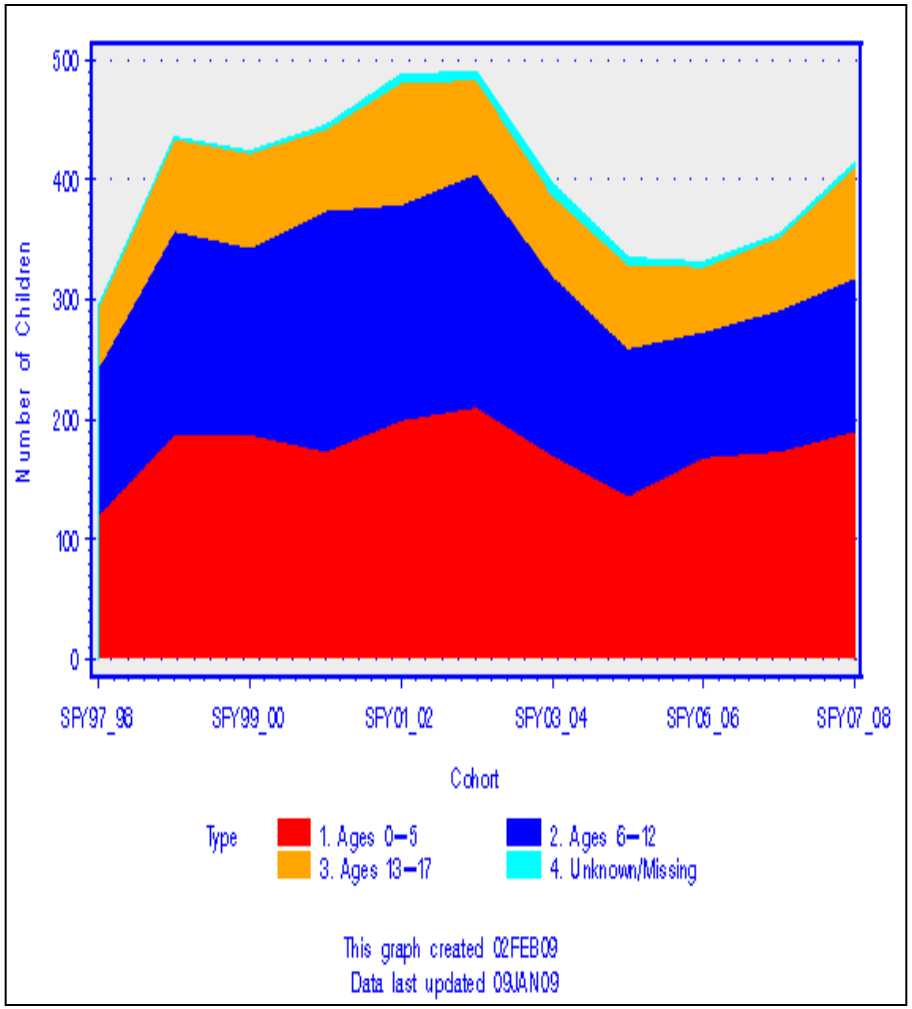
•**Explanation for this Chart**

•This chart depicts how many days children spent in their first Custody/Placement Authority. Where the line crosses the 50 percent line, half of the children in that cohort are no longer in custody. Length of stay is a longstanding concern regarding children's experiences in out-of-home care …



•**Rate of Leavaing Custody for the Children in SFY05_06 Cohort for North Carolina**





- Reports of Abuse and Neglect
- In Chatham County
- Unique Number of Children
- by First Ever Report Cohort

- Reports of Abuse and Neglect by Age
- In Chatham County
- Unique Number of Children
- by First Ever Report Cohort

Child Welfare in North Carolina

- Child Welfare
  - Experiences Report
    - All Children
    - By Categories
    - Summary Data
  - CFSR Measures
    - Prev Rd 1
    - Rd 1 By Categories
    - Current Rd 2
  - Abuse & Neglect
    - Longitudinal Data
    - Point in Time Data
  - Children in Foster Care
    - All Children
    - Age Out
  - Race & Ethnicity
- Work First
- Food & Nutrition Services
- Papers & Reports
- Additional Information
- Help

Stanly County
Stokes County
Surry County
Swain County
Transylvania County
Tyrrell County
Union County
Vance County
Wake County
Warren County
Washington County
Watauga County
Wayne County
Wilkes County
Wilson County
Yadkin County
Yancey County
Judicial District 1
Judicial District 2

[North Carolina] : Reports of Abuse and Neglect Type of Finding on Most Severe Report by Categories

Unique Number of Children by First Ever Report: Longitudinal Data

State Fiscal Year=SFY1997_1998

Type Found	Total	White	African American	American Indian/Alaskan	Other Races	Hispanic	Non_Hispanic	Male	Female	Ages 0 to 5	Ages 6 to 12	Ages 13 to 17	Missing Age Information
Abuse and Neglect	587	370	184	10	23	55	532	233	354	226	209	148	4
Abuse	1145	712	370	11	52	78	1067	434	711	352	465	328	0
Neglect	13587	7525	5297	258	507	1062	12525	6896	6691	7358	4480	1653	96
Dependency	248	111	118	3	16	39	209	121	127	134	52	59	3
Services Recommended	3	0	3	0	0	0	3	2	1	2	1	0	0
Unsubstantiated	43215	26297	14725	878	1315	3256	39959	21880	21334	20243	15658	7079	235
Services Not Recommended	3	0	3	0	0	0	3	2	1	1	0	0	2

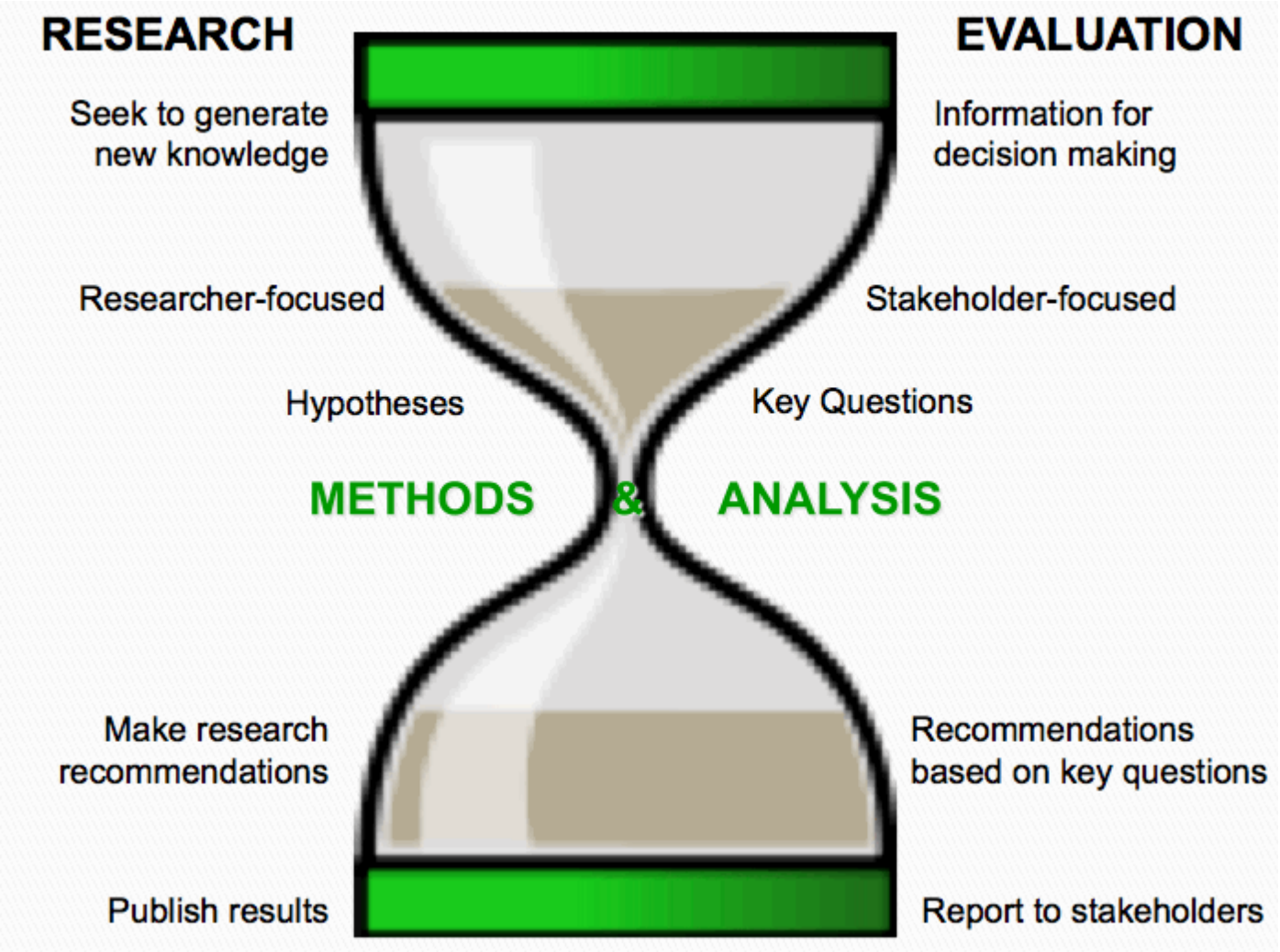
State Fiscal Year=SFY1998_1999

Type Found	Total	White	African American	American Indian/Alaskan	Other Races	Hispanic	Non_Hispanic	Male	Female	Ages 0 to 5	Ages 6 to 12	Ages 13 to 17	Missing Age Information
Abuse and Neglect	606	394	191	4	17	56	550	237	369	228	240	137	1
Abuse	988	627	304	12	45	78	910	349	639	319	388	275	6
Neglect	13776	7689	5250	250	587	1126	12650	7219	6557	7270	4628	1779	99
Dependency	242	129	101	4	8	30	212	118	124	100	63	76	3
Unsubstantiated	45463	27246	15720	924	1573	3421	42042	23043	22420	21199	16415	7619	230
Services Not Recommended	2	0	2	0	0	0	2	2	0	1	0	0	1

Data to Information for …

- Management decision support
- Evaluation
  - Self evaluation
- Research

What is Public Health Evaluation



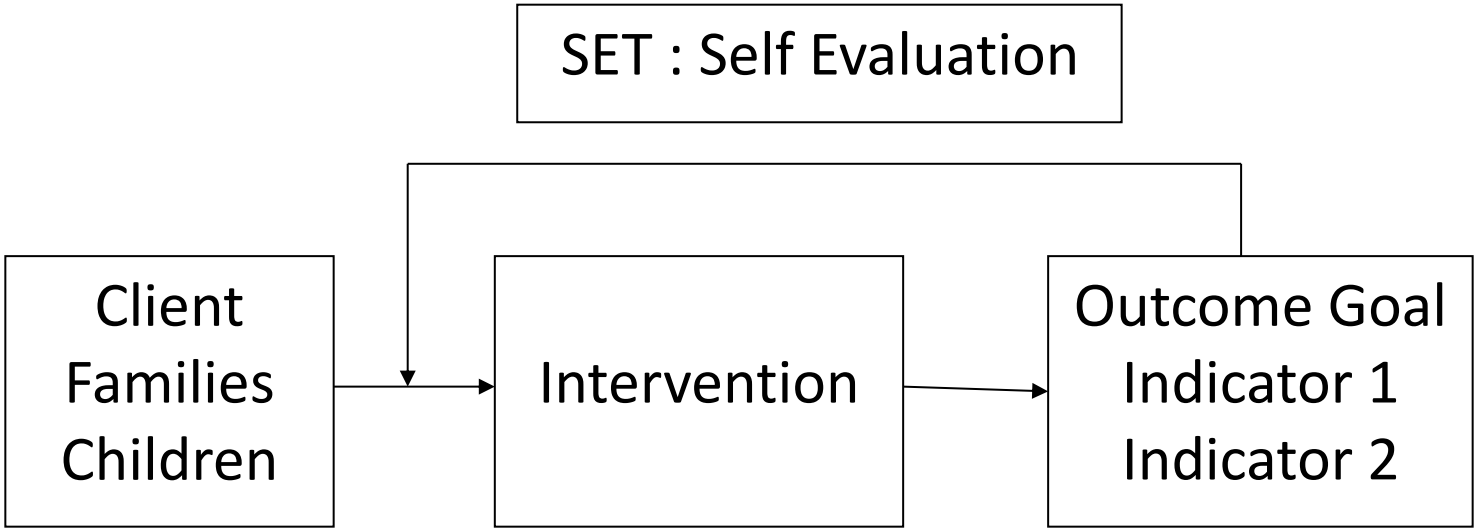
Papers

- **Kum, H.C.**, Duncan, D.F., & Stewart, C. J., **Supporting Self-Evaluation in Local Government via KDD**, *Government Information Quarterly: Building the Next-Generation Digital Government Infrastructures,* 26(2):pp 295-304, April 2009.
- Stewart, C.J., **Kum, H.-C.,** Barth, R.P., Duncan, D.F. **Former foster youth: Employment outcomes up to age 30.** Children and Youth Services Review, 2014. 36(0): p. 220-229.
- Barth, R. P., Duncan, D. F., Hodorowicz, M.T., and **Kum, H.C.**, **Felonious Arrests of Former Foster Care and TANF-Involved Youth**, *Journal of the Society for Social Work and Research,* 1:pp 104-123, 2010.
- Cilenti, D., **Kum, H.-C.,** Wells, R., Whitmire, T., Goyal, R., Hillemeier, M. **Trends in North Carolina Maternal Health Service Use and Outcomes During State Budget Cuts**. Journal of Public Health Management & Practice. Submitted

Guilford County : Leading By Results

- Goal : At-risk children and families should be safe and healthy in stable environments
  - Indicator 1: We will decrease the rate of children placed away from their home from 7% in FY 2003 to 5% by the end of FY 2005.
  - Indicator 2: We will decrease the rate of children entering foster care who are initially placed in shelter or group home care from 13% to 11% by the end of FY 2005.
  - Indicator 3: We will reduce the rate of children re-entering care from 10% in FY 2003 to 8% by the end of FY 2005.
  - Indicator 4: We will reduce the number of children in care with four or more placement moves from 14% in FY 2003 to 12% by the end of FY 2005.
  - Indicator 5: We will maintain the percent of children substantiated/in need of services that are not repeat victims of substantiated maltreatment at 92% by the end of FY 2005.
  - Indicator 6: We will continue to work on addressing disparities associated with race/ethnicity as evidenced by a decrease in the percentage of African American children in care from 57% by the end of FY 2005.

Alternative : Self Evaluation



- is a form of empowerment evaluation
- that is collaborative and participatory
- an ongoing process
- as long at it is technically strong, a viable alternative

Key ingredients for self-evaluation

- Outcomes for clients are clearly defined and disseminated throughout the agency
- It is a collaborative process that brings together individuals with different kinds of expertise to discuss the data used to measure outcomes and agency processes
- It is an ongoing process with regular meetings to discuss agency status on outcomes
- **It requires timely and accessible data that appropriately measured outcomes and other indicators of interest**
- It should include ongoing attention to implementation progress of the core strategies to improve the outcomes.
- **It requires technical expertise to ensure defensibility and adaptability**

→ **KDD Technology**

Data to Information for …

- Management decision support
- Evaluation
- Research

Research

- Management/evaluation [information/knowledge]
  - Management assistance website
- Policy [methods/information]
  - Evaluate the Medicaid reimbursement policy change in maternal care coordination (RWJ)
  - Employment outcomes of age out youth
    - ASPE (US-DHHS), Urban Institute, UC Berkeley
  - Felonious outcomes for youth (CW, TANF)
    - NC-DHHS (Department of Health and Human Services)
- Computer science [data/methods]
  - Sequential pattern mining
  - Digital government
  - Privacy preserving data integration

Research

- Management/evaluation [information/knowledge]
  - Management assistance website
- Policy [methods/information]
  - Evaluate the Medicaid reimbursement policy change in maternal care coordination (RWJ)
  - Employment outcomes of age out youth
    - ASPE (US-DHHS), Urban Institute, UC Berkeley
  - Felonious outcomes for youth (CW, TANF)
    - NC-DHHS (Department of Health and Human Services)
- Computer science [data/methods]
  - Sequential pattern mining
  - Digital government
  - Privacy preserving data integration

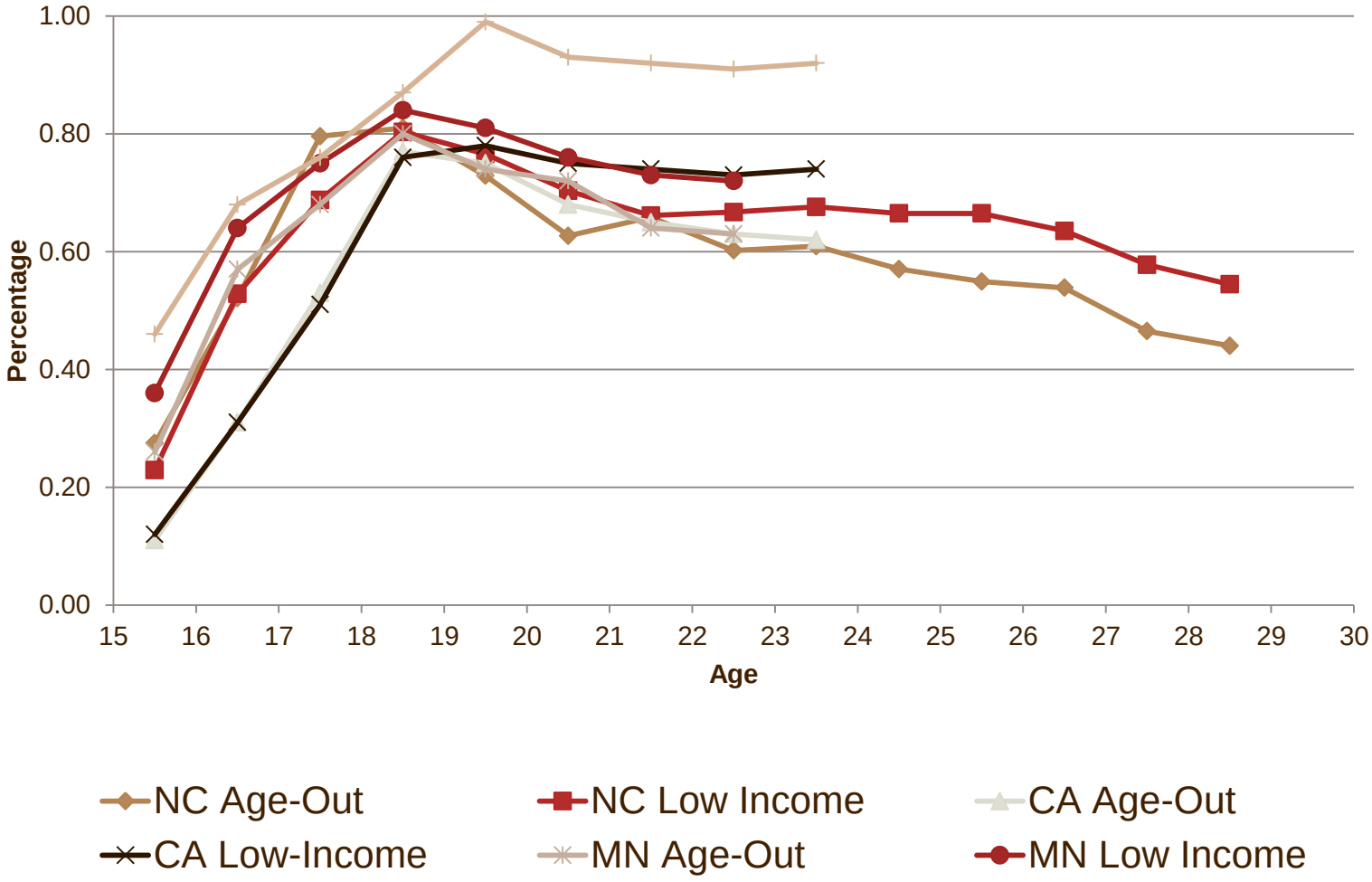
Papers

- **Kum, H.C.**, Duncan, D.F., & Stewart, C. J., **Supporting Self-Evaluation in Local Government via KDD**, *Government Information Quarterly: Building the Next-Generation Digital Government Infrastructures,* 26(2):pp 295-304, April 2009.
- Stewart, C.J., **Kum, H.-C.,** Barth, R.P., Duncan, D.F. **Former foster youth: Employment outcomes up to age 30.** Children and Youth Services Review, 2014. 36(0): p. 220-229.
- Barth, R. P., Duncan, D. F., Hodorowicz, M.T., and **Kum, H.C.**, **Felonious Arrests of Former Foster Care and TANF-Involved Youth**, *Journal of the Society for Social Work and Research,* 1:pp 104-123, 2010.
- Cilenti, D., **Kum, H.-C.,** Wells, R., Whitmire, T., Goyal, R., Hillemeier, M. **Trends in North Carolina Maternal Health Service Use and Outcomes During State Budget Cuts**. Journal of Public Health Management & Practice. Submitted

Employment outcomes for age out youth

- Data: Child welfare, TANF, UI
- Prior research : former foster youth who age out of foster care find that these youth generally experience high unemployment, unstable employment patterns, and earn very low incomes in the period between ages 18 and 21.
- Method: OLS, logistic, Cox proportional hazard, trajectory analysis
- Q1: What are the employment outcomes for youth who age out of foster care through their middle/late twenties—compared to comparable sample of low-income youth
- Q2: How do they fair in their twenties after they have made the initial transition into adulthood ?
- Q3: Do the unstable patterns of employment stabilize when youth reach their mid-twenties?

Results : Percent of Working Youth Employed Less are working

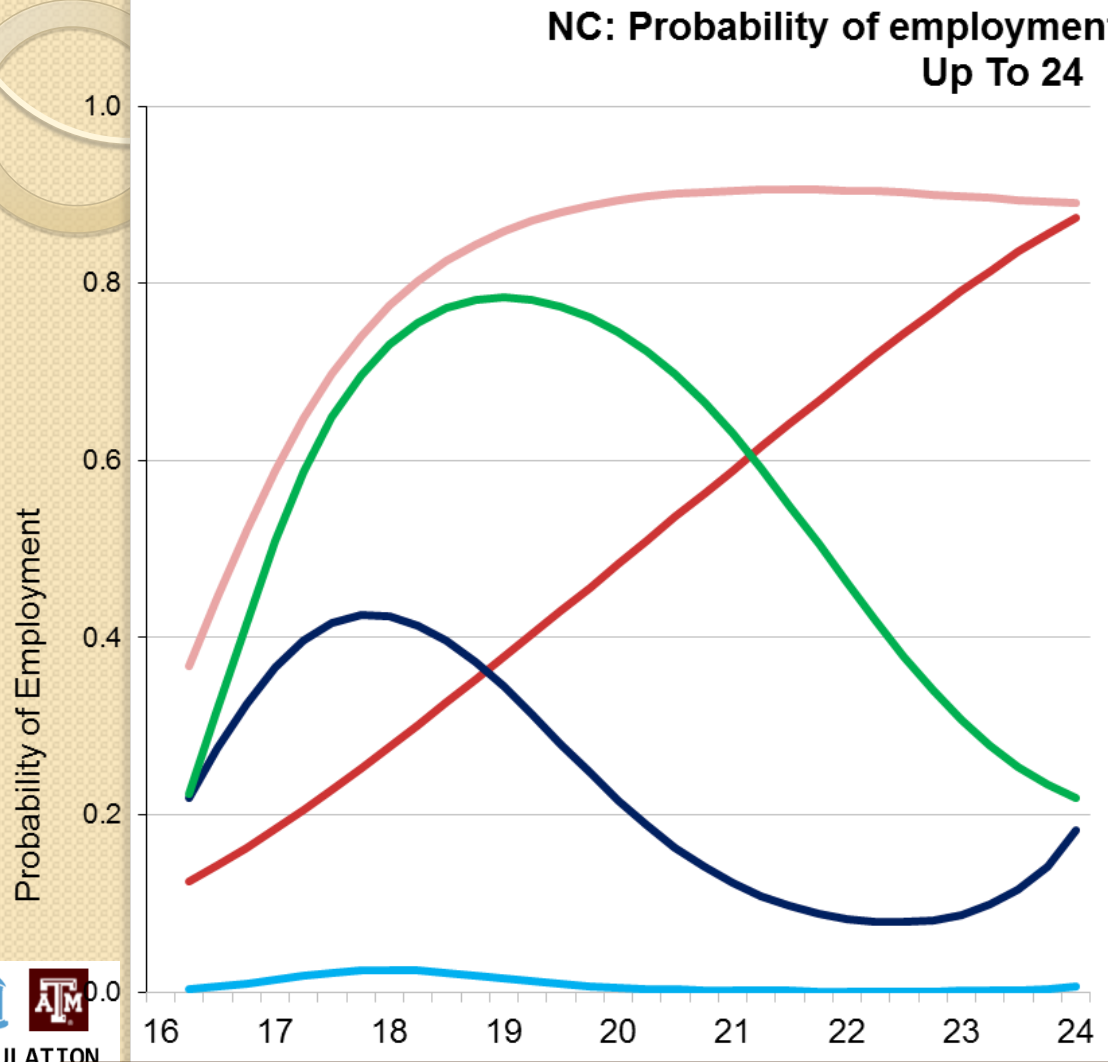


Results : Mean earnings of working youth of those working, make less

- At age 24, “working youth” who aged out of care had poorer outcomes than TANF and national comparison
  - Earned less (no significant difference in NC), although TANF receipt was lower
  - Less stable employment

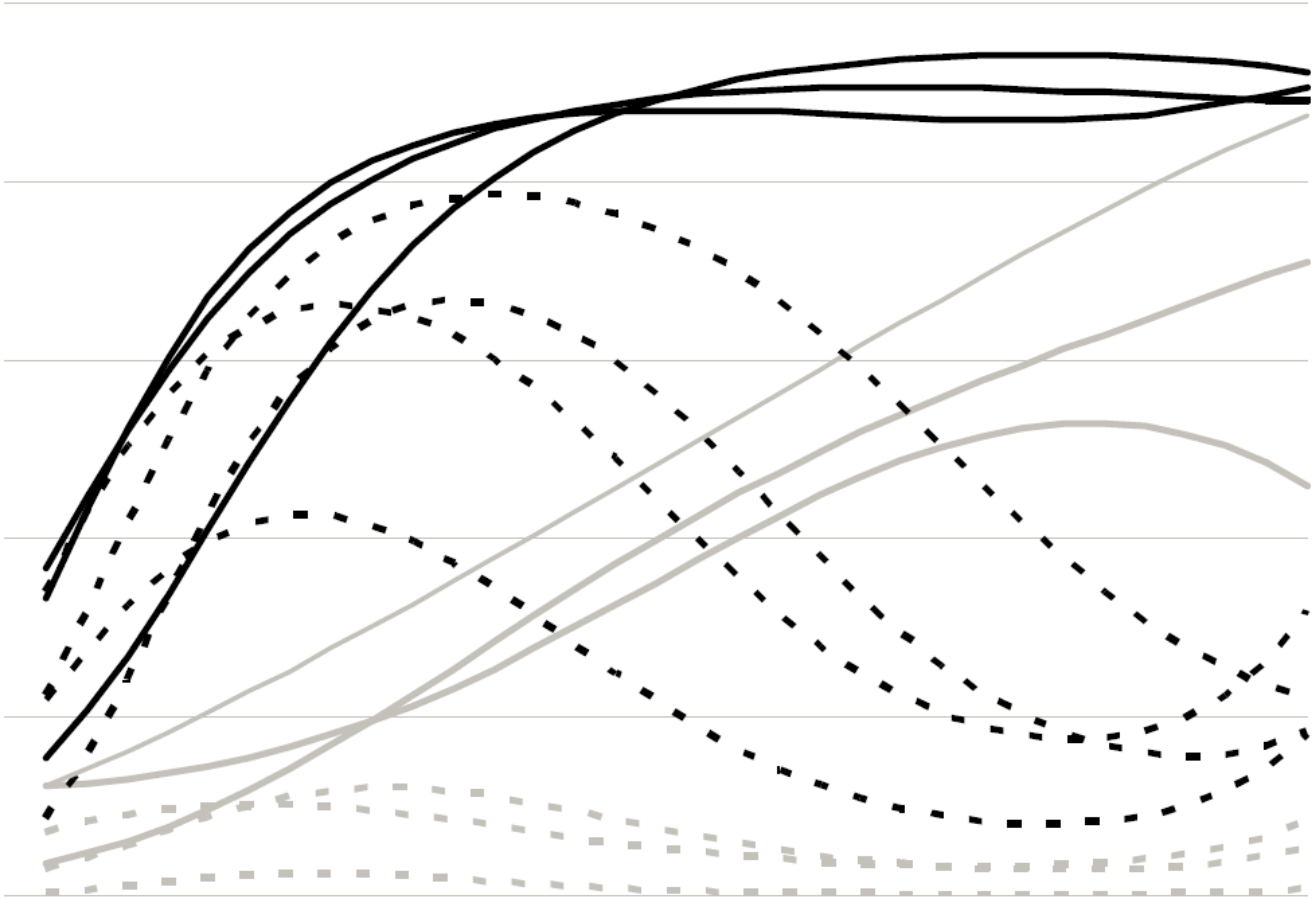
	Aged Out	TANF	National
CA	$690	$970	
MN	$575	$865	
NC	$450	$570	
National			$1,535

NC : Connectedness to the workforce



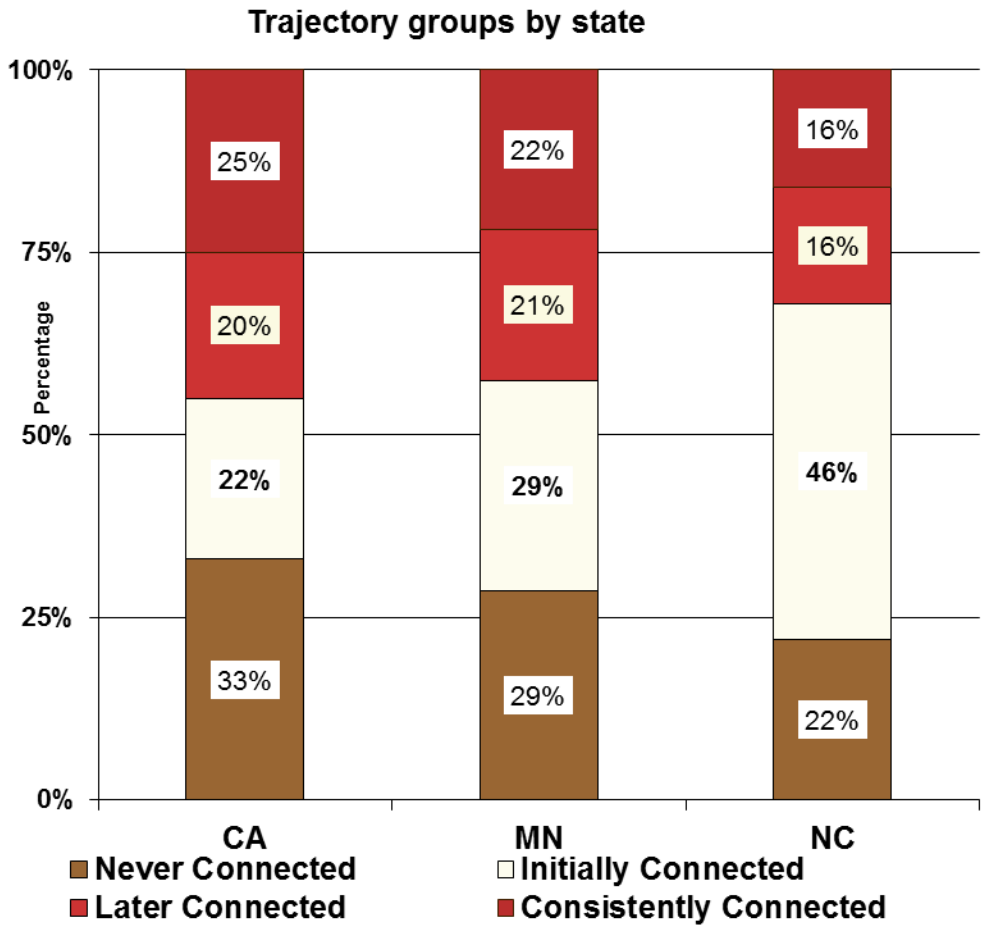
- ✓ **Trajectory analyses** were conducted on the sample of youth who age out of foster care in each state to identify distinct employment patterns over time
- ✓ This method groups youth with similar employment patterns and tracks the probability of employment at each age
- ✓ Semi-parametric group-based approach by Nagin (1999)

NC, CA, MN : consistent



Connectedness to the workforce

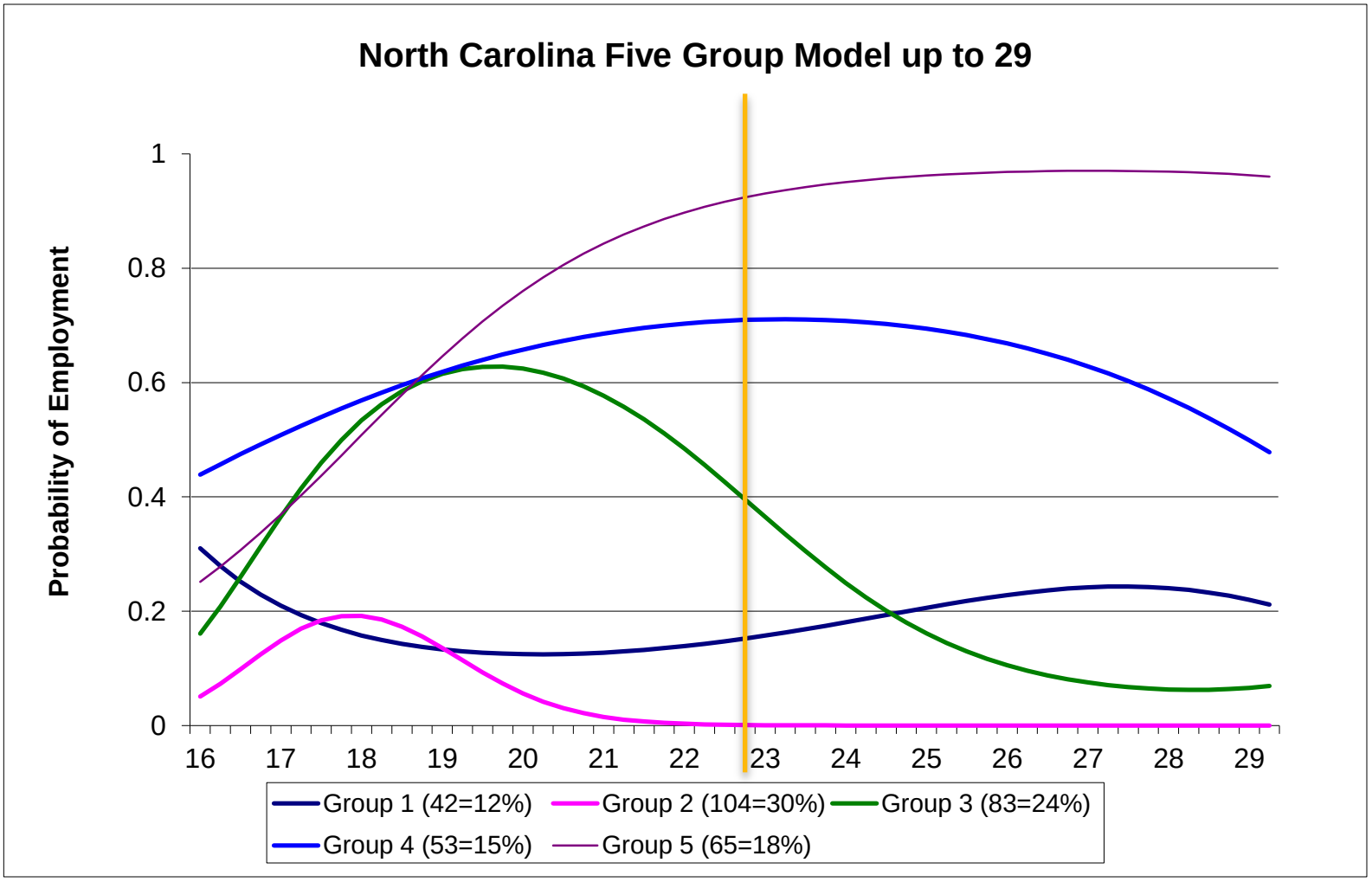
- Consistently connected
  - Consistent connections as adults
- Later connected
  - Not connected initially, but begin working more in early twenties
- Initially connected
  - Connected in late teens, but then employment drops
- Never connected
  - Never or minimally connected between 16 and 24



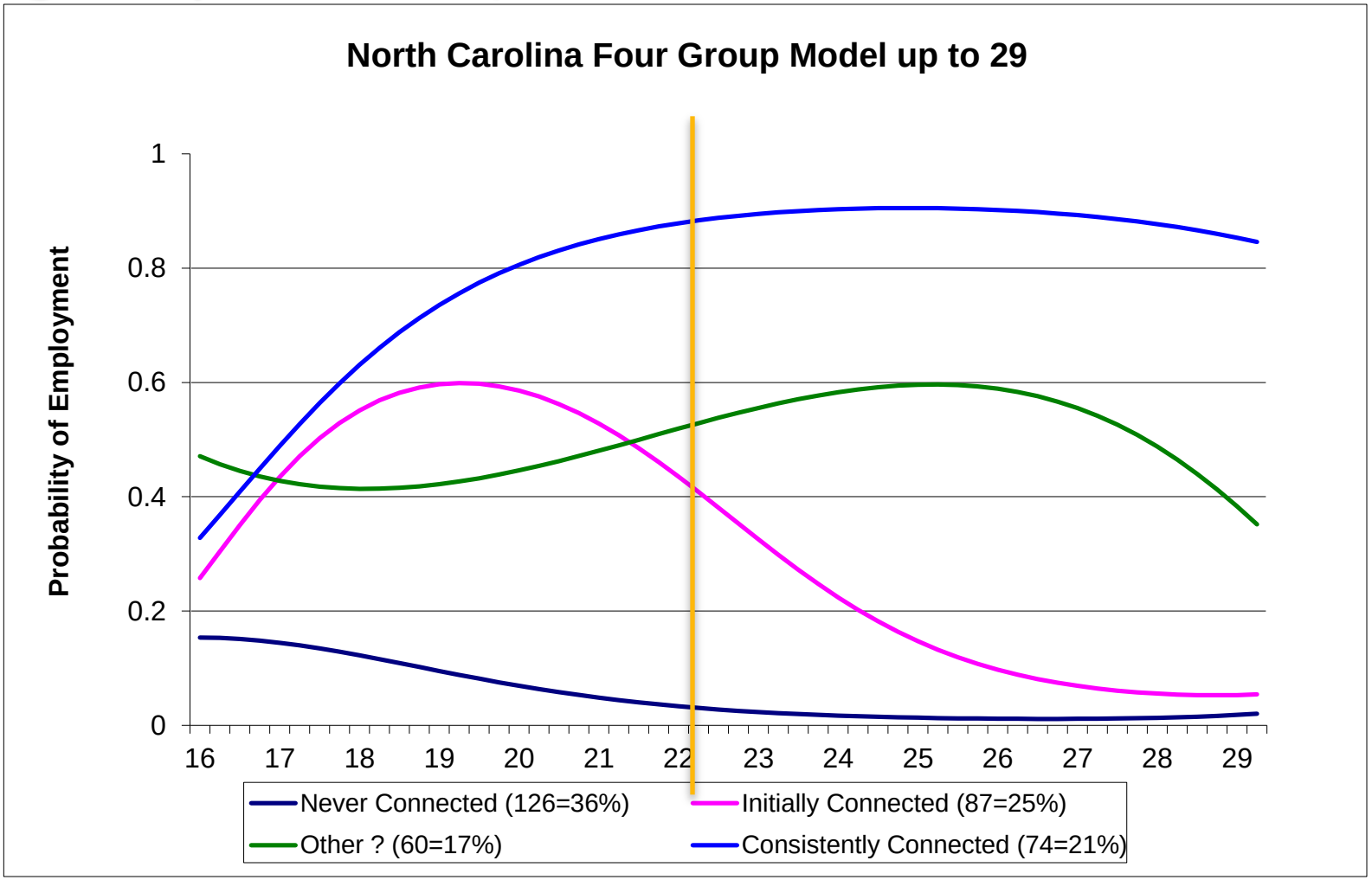
Employment outcomes for age out youth

- Data: Child welfare, TANF, UI
- Method: OLS, logistic, Cox proportional hazard, trajectory analysis
- Q1 : What are the employment outcomes for youth who age out of foster care through their middle twenties—compared to comparable sample of low-income youth
- Follow-up (Analysis extended to the early thirties for NC) Q2: Do youth who age out of FC catch up or continue to experience less employment and significantly lower earnings than their peers even into their late twenties.

Connectedness to the workforce : Extend to 29



Connectedness : Extend to 29 (4 group model)



Results

- 19% ($p<0.01$) more likely to be employed compared to those who did not work before age 18

Implications : Youth Aging Out

- Helping youth connect to the workforce prior to adulthood may have benefits later
  - employment prior to age 18 is associated with positive employment at age 24 in all three states, and at age 30 for NC.
  - Further explore later exit and early work experience findings
- Findings support related research about poor employment outcomes for youth who aged out of care
- Employment trajectories suggest different needs

Limitations : Youth Aging Out

- UI data excludes military, out-of-state, and “off the books” employment
  - Quarterly data
- Youth may be in school or incarcerated
- Omitted variables (e. g., family income, social support, mental and physical health, education)

Lessons : common sense counting measurement

- What to count?
- What is the context (denominator)?
- Understand exactly what you have
  - What is the unit of rows
  - What is being counted in variables (columns)
- ALWAYS check for face validity of counts
  - Data that is not used goes stale
- Cross check whenever you can
- Sensitivity analysis
- Look at your data (don't be scared of big data)

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

Papers

- **Kum, H.C.**, Duncan, D.F., & Stewart, C. J., **Supporting Self-Evaluation in Local Government via KDD**, *Government Information Quarterly: Building the Next-Generation Digital Government Infrastructures,* 26(2):pp 295-304, April 2009.
- Stewart, C.J., **Kum, H.-C.,** Barth, R.P., Duncan, D.F. **Former foster youth: Employment outcomes up to age 30.** Children and Youth Services Review, 2014. 36(0): p. 220-229.
- Barth, R. P., Duncan, D. F., Hodorowicz, M.T., and **Kum, H.C.**, **Felonious Arrests of Former Foster Care and TANF-Involved Youth**, *Journal of the Society for Social Work and Research,* 1:pp 104-123, 2010.
- Cilenti, D., **Kum, H.-C.,** Wells, R., Whitmire, T., Goyal, R., Hillemeier, M. **Trends in North Carolina Maternal Health Service Use and Outcomes During State Budget Cuts**. Journal of Public Health Management & Practice. Submitted

Felonious outcomes for youth

- DATA: Child welfare, TANF, arrest records
- Method: OLS, logistic, Cox proportional hazard
- Q1 : What are the adverse criminal justice-related outcomes for former foster youth

Felonious Arrest Study

- This cross-sectional study compares the felonious arrests as adults of youth
  - who were involved with TANF (*n* = 6,596)
  - who were involved with CWS (*n* = 1316). I
- n the CWS population, we compare youth
  - who emancipated from foster care (*n* = 841),
  - those who reunified (*n* =278),
  - and those with other exits (*n* = 197).
- Outcomes
  - the severity of the felony an individual was charged with as an adult
  - the hazard of being arrested after age 16 years.

Results: Felonious Arrest Study

- Findings show that as adults, TANF-involved youth are less likely than former foster youth to be charged with a felony;
- When former TANF youth are charged with felonies as adults, the felonies are less severe than felony charges incurred by former foster youth.
- Overall, as young adults, males and African American youth are not only more likely to be charged with felonies than other race youth or females but also are more likely to be charged with felonies of higher severity ratings than other race youth or females.

Results: Felonious Arrest Study

- Entering foster care as an adolescent and having multiple placements in care are associated with poorer outcomes in adulthood.
- However, youth who remain in foster care longer are less likely to be charged with a felony.
- The study confirms earlier work on adverse criminal justice-related outcomes for former foster youth and clarifies the need for specific interventions to reduce subsequent criminal involvement.

Papers

- **Kum, H.C.**, Duncan, D.F., & Stewart, C. J., **Supporting Self-Evaluation in Local Government via KDD**, *Government Information Quarterly: Building the Next-Generation Digital Government Infrastructures,* 26(2):pp 295-304, April 2009.
- Stewart, C.J., **Kum, H.-C.,** Barth, R.P., Duncan, D.F. **Former foster youth: Employment outcomes up to age 30.** Children and Youth Services Review, 2014. 36(0): p. 220-229.
- Barth, R. P., Duncan, D. F., Hodorowicz, M.T., and **Kum, H.C.**, **Felonious Arrests of Former Foster Care and TANF-Involved Youth**, *Journal of the Society for Social Work and Research,* 1:pp 104-123, 2010.
- Cilenti, D., **Kum, H.-C.,** Wells, R., Whitmire, T., Goyal, R., Hillemeier, M. **Trends in North Carolina Maternal Health Service Use and Outcomes During State Budget Cuts**. Journal of Public Health Management & Practice. Submitted

Evaluation on medicaid reimbursement policy change

- Data: Birth certificate, medicaid claims, WIC
- Method: logistic, regression
- Q1: What happened to health care access for low income pregnant women as North Carolina's recession deepened?

Data & Methods

- Random sample ~8,000 women
- Delivered between 10/1/2008-9/30/2010
- Covered through Medicaid
- Data from Medicaid eligibility files, claims, WIC files, and birth certificates

Multiple Regression

- Logistic for 1/0 outcomes; negative binomial for count variables
- Month of recession measured through month of baby's birth
- Controlling for whether or not mother was enrolled in full Medicaid, Pregnancy or Family Planning Waiver, maternal age, new mother, African American, or Hispanic

Results : Medicaid reimbursement policy

- Overall: Medicaid paid for significantly more deliveries over time
- Medicaid covered a greater percentage of the pregnancy period
- Service utilization stayed about the same over time, with the following exceptions:
  - Number of OB visits decreased over time for all three trimesters and postpartum
  - Service utilization at local health departments declined somewhat, particularly among Whites and African Americans
  - Family planning service utilization during the first 3 months postpartum declined, especially for Whites

Results : Medicaid reimbursement policy

- Outcomes
  - Maternal weight gain: women in general, especially white women, were slightly less likely to gain excessive amount of weight
  - Preterm /low birth weight: no significant changes except for Hispanic women who were less likely to deliver a baby at low birth weight over time. There is weak evidence indicating they were less likely to deliver preterm babies.
  -

Implications : Medicaid reimbursement policy

- Broadening the safety net for low-income pregnant women may have resulted in some respects less access for insured women, perhaps as a result of less active outreach, busier staff, longer wait times, and bottlenecks in service delivery
- Preparing to provide health services to increasing numbers of individuals under the Affordable Care Act will require meaningful investment in workforce and health care system infrastructure

Limitations : Medicaid reimbursement policy

- Excludes any services not paid through Medicaid
- May fail to account for other rival hypotheses (e.g., changing attitudes)
- There were some major changes to the birth certificate data in 2010, which resulted in not being able to control for some variables like mother's education and prenatal care indicators

Findings: Service Utilization, Overall

	LHD Care n=7,911		Family Planning 3 mos postpartum n=7,911	
	OR	p-value	OR	p-value
Month of birth	**0.991**	**0.0139**	**0.99**	**0.0069**
Full Medicaid	1.643	<.0001	1.366	<.0001
Maternal age	0.927	<.0001	0.982	0.0006
Parity < 1	1.585	<.0001	0.827	0.0009
African American	1.206	0.001	0.908	0.0711
Hispanic	0.857	0.018	0.151	<.0001

Findings: Service Utilization subgroups

	LHD Care n=2,577		Family Planning 3 mos postpartum n=2,577	
African American	OR	p-value	OR	p-value
Month of birth	**0.983**	**0.0089**	**0.994**	**0.3279**
Full Medicaid	1.468	<.0001	1.178	0.0536
Maternal age	0.92	<.0001	0.994	0.4902
Parity < 1	2.383	<.0001	0.862	0.1137

	LHD Care n=3,617		Family Planning 3 mos postpartum n=3,617	
White	OR	p-value	OR	p-value
Month of birth	**0.988**	**0.0289**	**0.987**	**0.0117**
Full Medicaid	1.714	<.0001	1.287	0.0003
Maternal age	0.909	<.0001	0.978	0.0014
Parity < 1	1.538	<.0001	0.737	<.0001

Findings: Count of OB Visits

	1st Trimester n=5,098		2nd Trimester n=6,488		3rd Trimester n=6,953		Postpartum n=7,911	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Month of birth	**-0.023**	**<.0001**	**-0.014**	**<.0001**	**-0.012**	**<.0001**	**-0.036**	**<.0001**
Full Medicaid	0.390	<.0001	0.243	<.0001	0.221	<.0001	0.728	<.0001
Maternal age	-0.003	0.348	-0.004	0.070	-0.008	0.000	-0.020	<.0001
Parity < 1	0.009	0.768	0.009	0.703	0.034	0.167	0.008	0.755
African American	-0.061	0.026	0.014	0.531	0.005	0.815	0.101	<.0001
Hispanic	-0.286	<.0001	-0.498	<.0001	-0.520	<.0001	-0.888	<.0001

Findings: Count of Mental Health and Substance Abuse Visits

	Mental Health (n=1,375)		Substance Abuse (n=589)	
Overall	Estimate	p-value	Estimate	p-value
Month of birth	**0.017**	**0.085**	**0.036**	**0.072**
Full Medicaid	0.738	<.0001	0.311	0.295
Maternal depression	0.698	<.0001	0.362	0.272
Maternal bipolar	1.070	<.0001	0.567	0.107
Maternal schizophrenic	0.789	0.054	0.146	0.865
Maternal anxiety	-0.254	0.062	-0.414	0.236
Maternal trauma	0.913	<.0001	0.123	0.798
Maternal substance abuse	0.725	<.0001	0.080	0.021
Maternal age	0.015	0.194	-0.532	0.082
Parity < 1	0.337	0.025	-0.274	0.310
African American	0.448	0.003	-1.877	0.082
Hispanic	-0.280	0.405	0.036	0.072

Findings: Count of Mental Health and Substance Abuse Visits

	Mental Health (n=58)		Substance Abuse (n=11)	
Hispanic	Estimate	p-value	Estimate	p-value
Month of birth	**0.1434**	**0.0747**	**-1.2769**	**0.4744**
Full Medicaid	2.0981	0.0977	21.1217	0.5321
Maternal depression	2.5397	0.0844	-10.4475	0.8009
Maternal bipolar	0.8571	0.4632	0	--
Maternal schizophrenic	0	--	0	--
Maternal anxiety	1.4173	0.1569	3.4301	0.9302
Maternal trauma	2.6274	0.013	0	--
Maternal substance abuse	2.608	0.2516	--	--
Maternal age	0.098	0.1554	0.4834	0.61
Parity < 1	0.4461	0.581	17.6225	0.4772

Findings: Count of Mental Health and Substance Abuse Visits

	Mental Health (n=912)		Substance Abuse (n=380)	
White	Estimate	p-value	Estimate	p-value
Month of birth	**0.022**	**0.084**	**0.044**	**0.100**
Full Medicaid	0.529	0.006	0.454	0.215
Maternal depression	0.569	0.004	0.140	0.758
Maternal bipolar	1.093	<.0001	0.421	0.381
Maternal schizophrenic	-0.008	0.991	0.098	0.930
Maternal anxiety	-0.294	0.089	-0.227	0.598
Maternal trauma	0.844	0.002	-0.049	0.941
Maternal substance abuse	0.945	<.0001	0.118	0.012
Maternal age	0.037	0.021	-0.556	0.209
Parity < 1	0.445	0.020	0.044	0.100

Findings: Maternal Outcomes

Overall	Maternal Smoking (n=4,499)		Excessive Weight Gain (n=7,136)	
	OR	p-value	OR	p-value
Month of birth	**0.99**	**0.2815**	**0.992**	**0.029**
Full Medicaid	1.775	<.0001	1.056	0.3429
Maternal age	1.005	0.5629	1.014	0.0132
Parity < 1	0.673	<.0001	1.771	<.0001
African American	0.278	<.0001	1.088	0.1347
Hispanic	0.077	<.0001	0.523	<.0001

White	Maternal Smoking (n=2,262)		Weight Gain (n=3,548)	
	OR	p-value	OR	p-value
Month of birth	**0.99**	**0.3723**	**0.988**	**0.0286**
Full Medicaid	1.716	<.0001	1.047	0.5659
Maternal age	0.993	0.4701	1.011	0.169
Parity < 1	0.685	0.0006	1.835	<.0001

Findings: Infant Outcomes

Overall	Preterm Delivery (n=7,136)		Low Birthweight (n=7,136)	
	OR	p-value	OR	p-value
Month of birth	**0.999**	**0.887**	**0.995**	**0.450**
Full Medicaid	1.228	0.033	1.406	0.001
Maternal age	1.016	0.070	1.021	0.021
Parity < 1	0.998	0.981	1.237	0.046
African American	1.205	0.048	1.499	<.0001
Hispanic	0.763	0.071	0.680	0.024

Hispanic	Preterm Delivery (n=1,039)		Low Birthweight (n=1,039)	
	OR	p-value	OR	p-value
Month of birth	**0.962**	**0.061**	**0.944**	**0.019**
Full Medicaid	1.056	0.906	1.925	0.131
Maternal age	1.058	0.018	1.051	0.074
Parity < 1	1.550	0.178	2.089	0.044