LEADING

LIS Education And Data Science Integrated Network Group

Data Integration and Quality w. OpenRefine

Richard **MARCIANO**
University of Maryland
AI-Collaboratory

Monday, June 14, 2021 – 1:00 to 2:30 p.m.
LEADING Bootcamp – Week 2

https://ai-collaboratory.net/cas/

contact@ai-collaboratory.net  http://ai-collaboratory.net  @aicollaboratory

Advanced Information Collaboratory

UNIVERSITY OF MARYLAND

OpenRefine

A free, open source, powerful tool for working with messy data

https://openrefine.org/download.html

OpenRefine 3.4.1

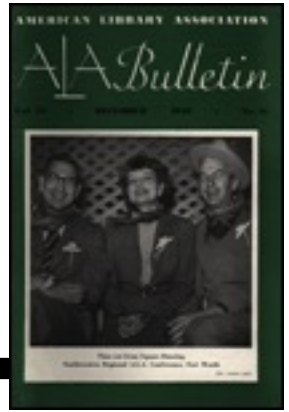
The latest stable release of OpenRefine 3.4.1, released on September 24, 2020.
Please backup your workspace directory before installing and report any problems that you encounter. A change log is provided on the release page.

- **Windows kit**, This requires Java to be installed on your computer. Download, unzip, and double-click on *openrefine.exe* or *refine.bat* if the former does not work.
- **Windows kit with embedded Java**, includes OpenJDK Java, available under the GPLv2+CE license. Download, unzip, and double-click on *openrefine.exe* or *refine.bat* if the former does not work.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. You do not need to install Java separately.
- **Linux kit**, Download, extract, then type *./refine* to start. This requires Java to be installed on your computer.

Data Processing in the Library School Curriculum

	Methods Analysis	Data Processing in the Library	Info. Systems Analysis	Research Emphasis
Drexel Institute of Technology	X	X		
University of Maryland	X	X	X	X
University of North Carolina			X	

Robert M. Haye



ALA Bulletin

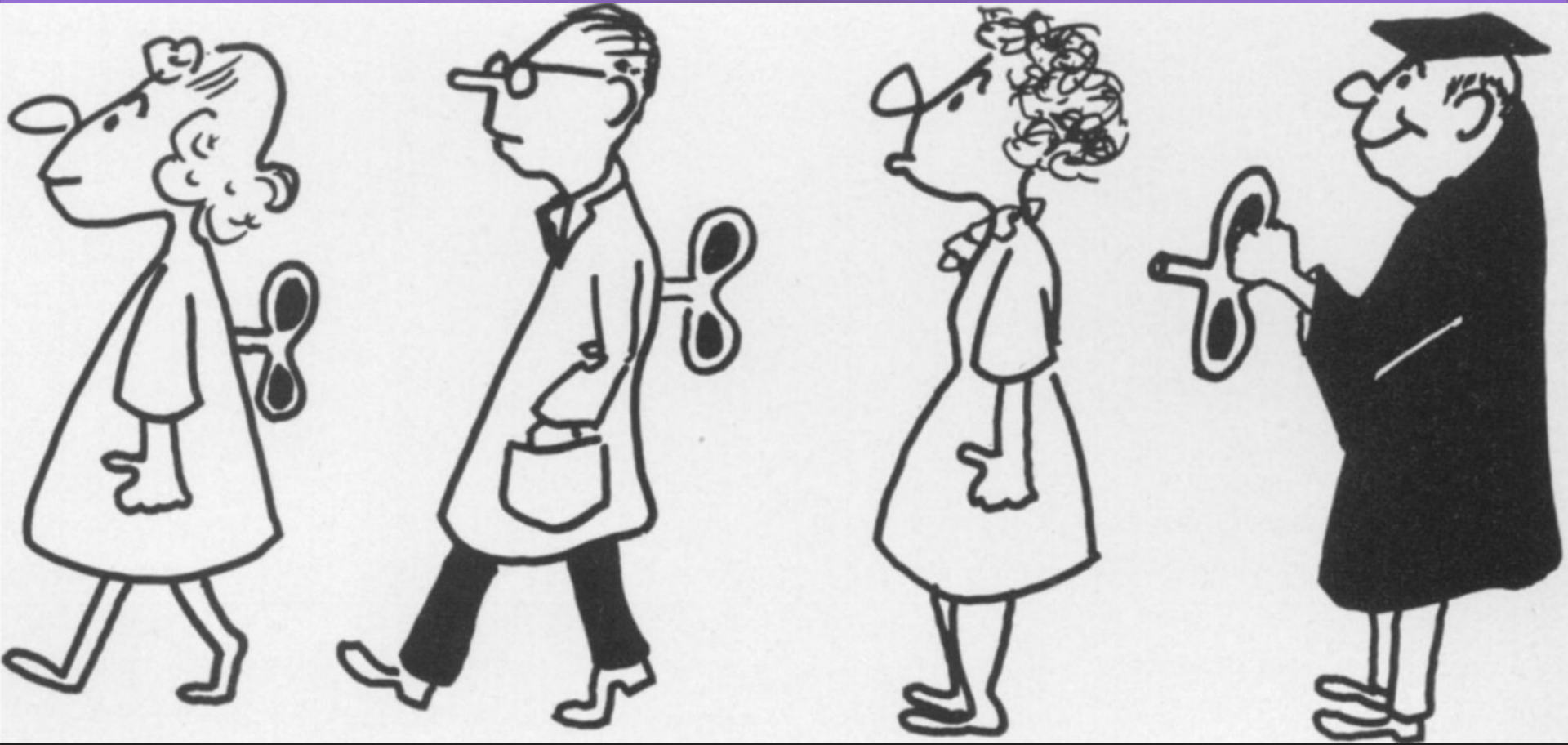
Vol. 61, No. 6 (June XXXX), pp. 662-669 (8 pages)

https://www.jstor.org/stable/25697657

Published by: American Library Association

1. Library school students generally lack the technical background necessary for data processing work as such, and yet, as library school graduates, they will find themselves in a world where that technology will play an increasingly important role. They must be given a sufficient orientation to be able to fit data processing into the context of library goals and purposes. The intent of the introductory course is to bridge the gap between their existing background and their future work.

2. There is a great and ever-growing need for library systems analysts, educated as librarians but with the technical tools for competent work with data processing technology.



Mapping CT to Library & Archival Science Education & Research

"To reading, writing, and arithmetic, we should add computational thinking to every child's analytical ability."

(Wing, 2006)

"… a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design", and scale.

"Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records.", CAS#5 Workshop
Richard Marciano, William Underwood et al.
Link: https://ai-collaboratory.net/wp-content/uploads/2020/03/2.Marciano.pdf

"Reframing Digital Curation Practices through a Computational Thinking Framework", CAS#5 Workshop
Richard Marciano et al.
Link: https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf

Viewpoint | Jeannette M. Wing

Computational Thinking

It represents a universally applicable attitude and skill set everyone, not just computer scientists, would be eager to learn and use.

Computational thinking builds on the power and limits of computing processes, whether they are executed by a human or by a machine. Computational methods and models give us the courage to solve problems and design systems that no one of us would be capable of tackling alone. Computational thinking confronts the riddle of machine intelligence: What can humans do better than computers? and What can computers do better than humans? Most fundamentally it addresses the question: What is computable? Today, we know only parts of the answers to such questions.

Computational thinking is a fundamental skill for everyone, not just for computer scientists. To reading, writing, and arithmetic, we should add computational thinking to every child's analytical ability. Just as the printing press facilitated the spread of the three Rs, what is appropriately incestuous about this vision is that computing and computers facilitate the spread of computational thinking.

Computational thinking involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science. Computational thinking includes a range of mental tools that reflect the breadth of the field of computer science.

Having to solve a particular problem, we might ask: How difficult is it to solve? and What's the best way to solve it? Computer science rests on solid theoretical underpinnings to answer such questions precisely. Stating the difficulty of a problem accounts for the underlying power of the machine—the computing device that will run the solution. We must consider the machine's instruction set, its resource constraints, and its operating environment.

In solving a problem efficiently, we might further ask whether an approximate solution is good enough, whether we can use randomization to our advantage, and whether false positives or false negatives are allowed. Computational thinking is reformulating a seemingly difficult problem into one we know how to solve, perhaps by reduction, embedding, transformation, or simulation.

Computational thinking is thinking recursively. It is parallel processing. It is interpreting code as data and data as code. It is type checking as the generalization of dimensional analysis. It is recognizing both the virtues and the dangers of aliasing, or giving someone or something more than one name. It is recognizing both the cost and power of indirect addressing and procedure call. It is judging a program not just for correctness and efficiency but for aesthetics, and a system's design for simplicity and elegance.

Computational thinking is using abstraction and decomposition when attacking a large complex task or designing a large complex system. It is separation of concerns. It is choosing an appropriate representation for a problem or modeling the relevant aspects of a problem to make it tractable. It is using invariants to describe a system's behavior succinctly and declaratively. It is having the confidence we can safely use, modify, and influence a large complex system without understanding its every detail. It is

Jeannette Wing: Communications of the ACM 2006:
http://www.cs.cmu.edu/~./15110-s13/Wing06-ct.pdf
CIRCL: Center for Innovative Research in Cyberlearning:
https://circlcenter.org/computational-thinking/

Mentimeter Survey conducted in Richard Marciano's MLIS class: Feb. 2021:

Mentimeter

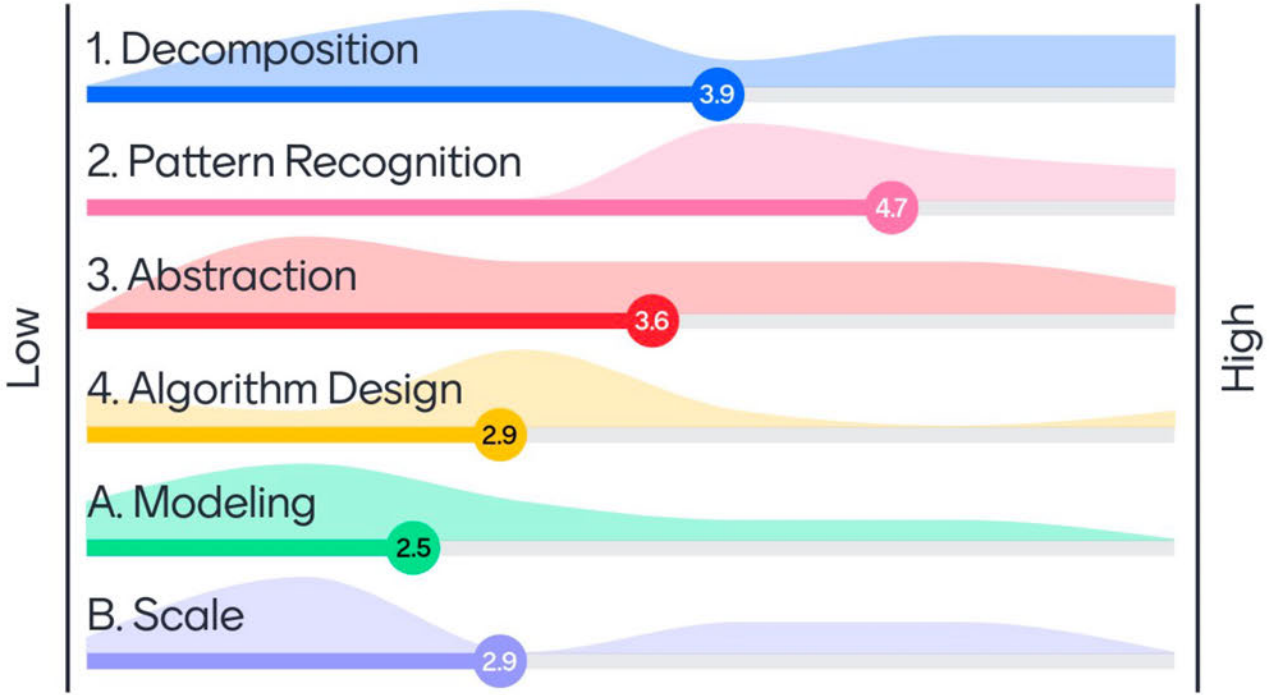
Computational Thinking Components

- 1. DECOMPOSITION
- .....Breaking a large problem into smaller more manageable parts
- 2. PATTERN RECOGNITION
- .....Recognizing which parts are the same and the various attributes we can use to define them
- 3. ABSTRACTION
- .....Filtering out the data you need and what you don't based on the attributes
- 4. ALGORITHM DESIGN
- .....Planning the step-by-step instructions that need to be carried out to achieve the goal
- A. Modeling
- Reformulating seemingly difficult problems into solvable forms using reduction, transformation, recursion, and simulation.
- B. Scale
- Solving problems that are large enough they require forms of automation

6

Rate your level of comfort [on a scale from 1-6]:

Mentimeter



10

This suggests that computational archival science is a blend of: (1) computational (2) archival thinking.

David Weintrop:
- CT-STEM Practices Taxonomy

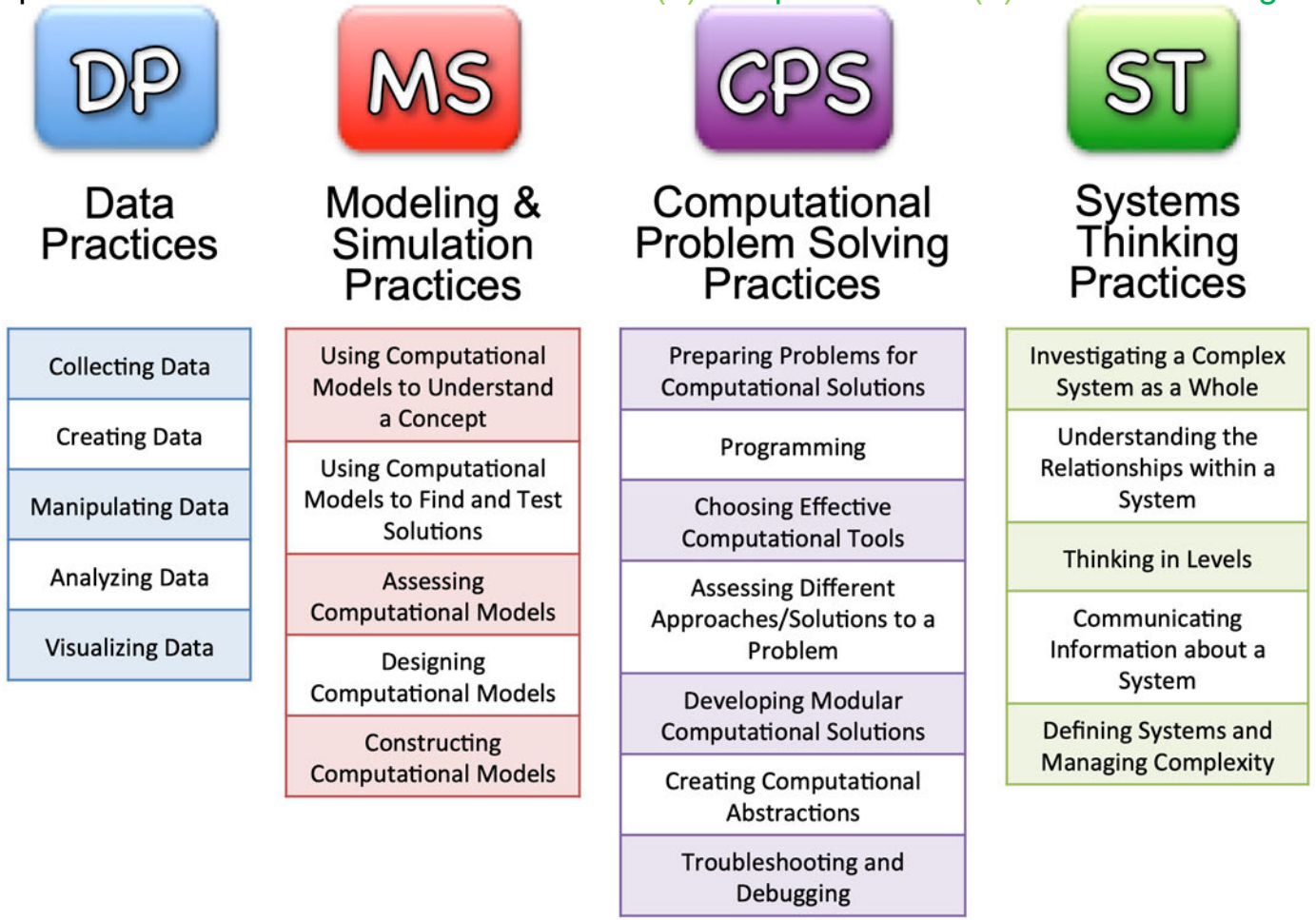
CITATION

Bill Underwood:
- CAS#4:

Analysis of the remaining eleven workshop papers indicates that the research that they report also involves CT.
https://ai-collaboratory.net/wp-content/uploads/2020/02/16_OpenMic_Bill-Underwood.pdf

- CT-LASER Practice:

Motivation for Integrating CT into UMD MLIS program in Library and Archival Studies, with examples of CT Practices being used in Archival Studies Research
https://ai-collaboratory.net/wp-content/uploads/2020/04/Underwood_CompThinkInArchResearch.pdf



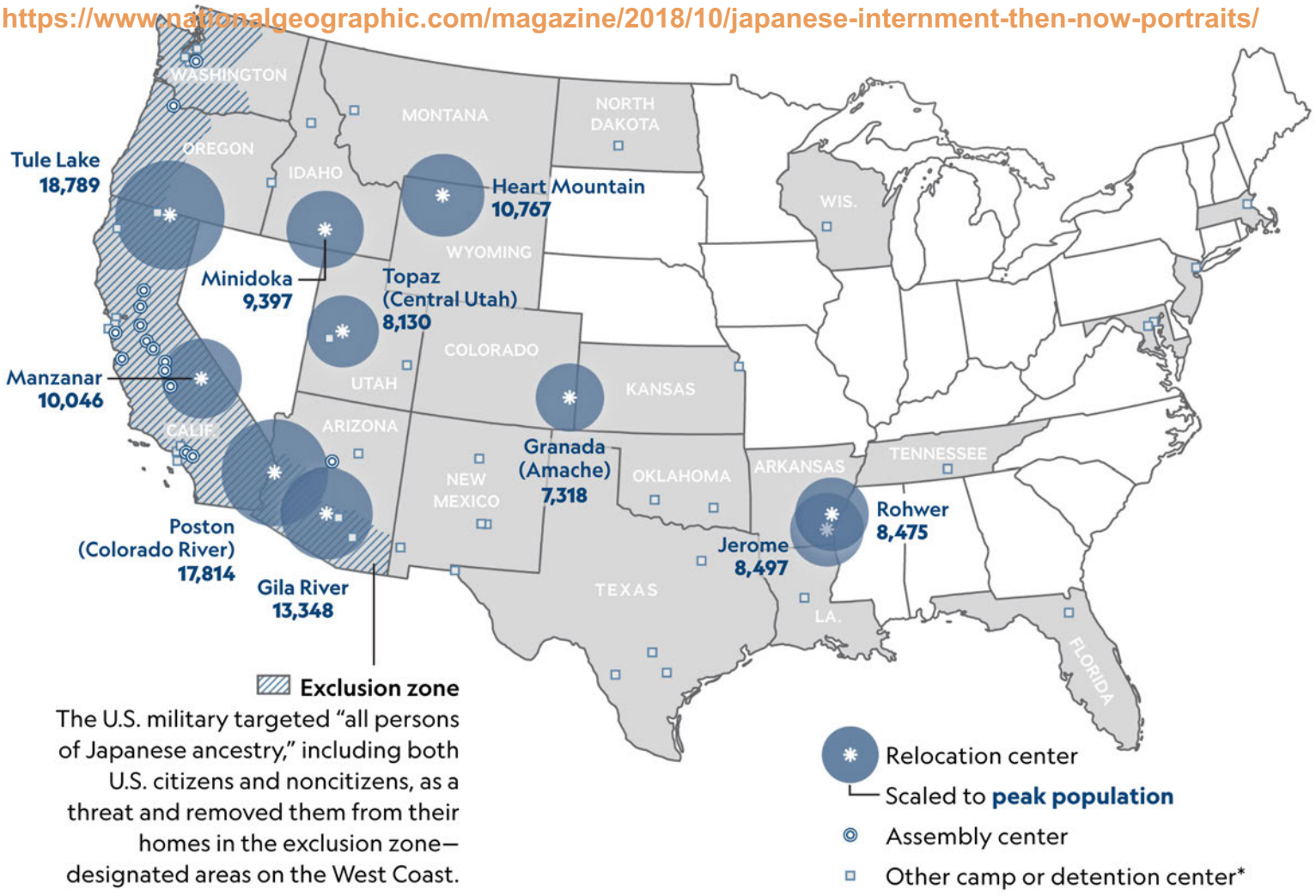
Example of CAS
with a CT-LASER mapping

Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camps

Proceedings of IEEE Big Data Conference 2018, CAS Workshop: Dec. 13, 2019, Seattle, WA.

- "Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records."
  Richard Marciano, William Underwood et al.
  Link: https://ai-collaboratory.net/wp-content/uploads/2020/03/2.Marciano.pdf

https://www.nationalgeographic.com/magazine/2018/10/japanese-internment-then-now-portraits/

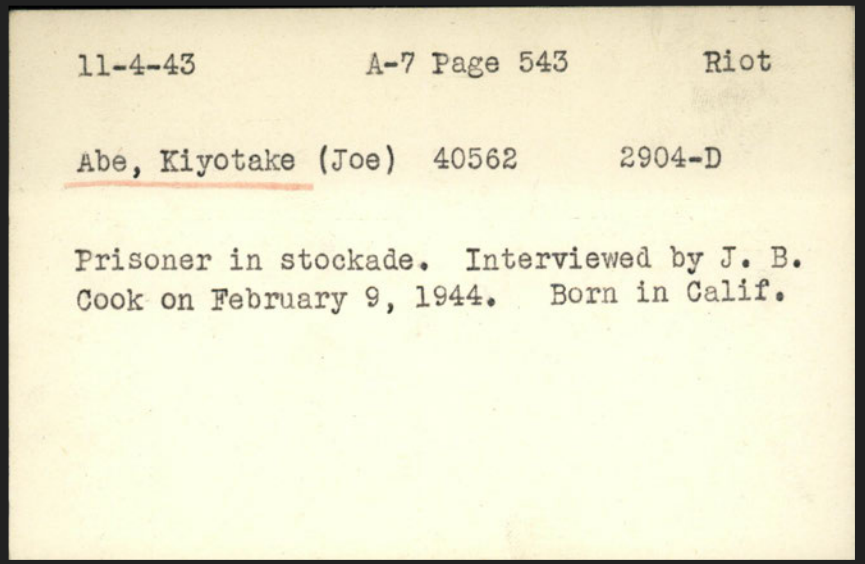


Automating the Detection of Personally Identifiable Information (PII) in Index Cards to Internal Security Case Reports

The records of the WRA (Record Group 210 from 1941-47) at the National Archives in Washington D.C. and Maryland, are comprised of over 100 series with motion picture films, drawings of incarceration centers, photos, maps, correspondence, yearbooks, rosters, etc.

Series 51 & 52 have immense value for survivors of the camps, their families, and historians, yet they are still not accessible.

Series 51, the “Internal Security Case Reports” from 1942 to 1946, comprises narrative reports prepared by camp investigators, police officers, and directors of internal security, relating cases of alleged “disorderly conduct, rioting, seditious behavior,” etc. at each of the 10 camps, with detailed information on the names and addresses in the camps of the persons involved, the time and place where the alleged incident occurred, an account of what happened, and a statement of action taken by the investigating officer.

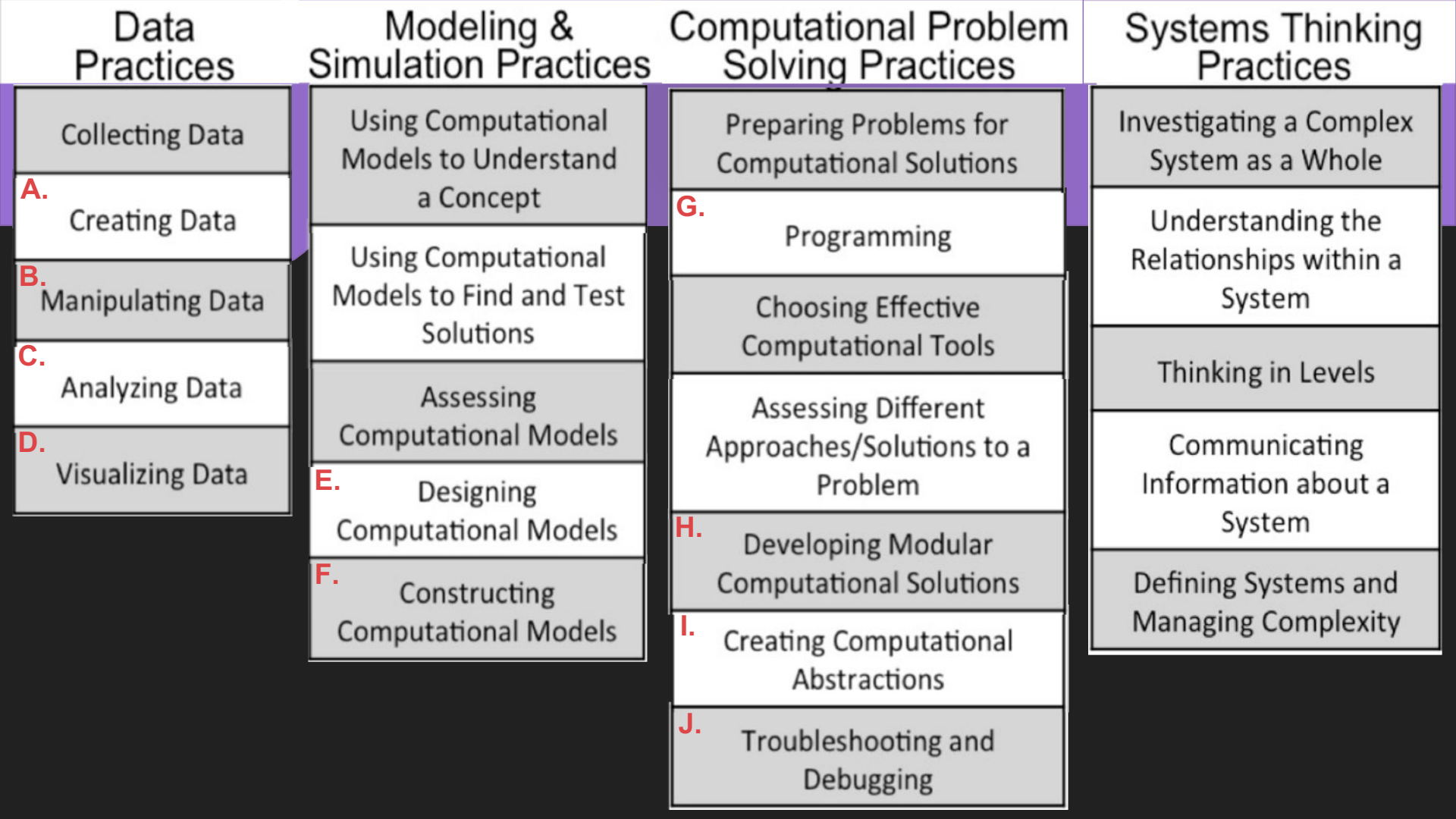


11-4-43 A-7 Page 543 Riot

Abe, Kiyotake (Joe) 40562 2904-D

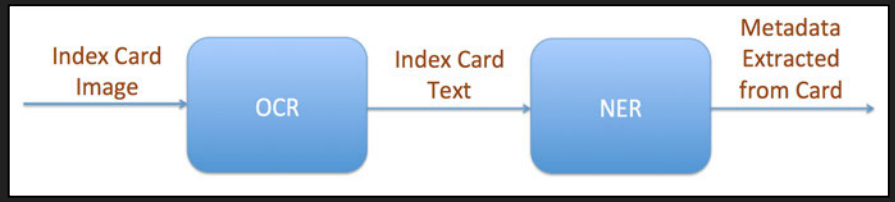
Prisoner in stockade. Interviewed by J. B. Cook on February 9, 1944. Born in Calif.

364:	Topaz UT (1%)
1,202:	Poston AZ (5%)
1,578:	Gila River AZ (6%)
763:	Granada CO (3%)
533:	Heart Mountain WY (2%)
2,146:	Manzanar CA (9%)
2,343:	Minidoka ID (9%)
15,648:	Tule Lake CA (63%)
468:	Rohwer/Jerome AK (2%)
25,045	


Data Practices
Collecting Data
A. Creating Data
B. Manipulating Data
C. Analyzing Data
D. Visualizing Data
Modeling & Simulation Practices
Using Computational Models to Understand a Concept
Using Computational Models to Find and Test Solutions
Assessing Computational Models
E. Designing Computational Models
F. Constructing Computational Models
Computational Problem Solving Practices
Preparing Problems for Computational Solutions
G. Programming
Choosing Effective Computational Tools
Assessing Different Approaches/Solutions to a Problem
H. Developing Modular Computational Solutions
I. Creating Computational Abstractions
J. Troubleshooting and Debugging
Systems Thinking Practices
Investigating a Complex System as a Whole
Understanding the Relationships within a System
Thinking in Levels
Communicating Information about a System
Defining Systems and Managing Complexity

A. Creating Data

"The increasingly computational nature of working with data in" **archival science** *"underscores the importance of developing computational thinking practices in the classroom." "Part of the challenge is teaching students that answers are drawn from the data available." "In many cases"* **archivists** *"use computational tools to generate data… at scales that would otherwise be impossible."*



11-4-43 A-7 Page 543 Riot

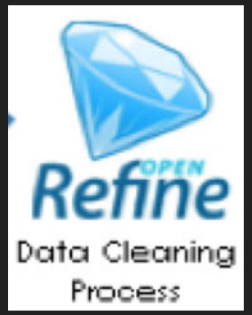
Abe, Kiyotake (Joe) 40562 2904-D

Prisoner in stockade. Interviewed by J. B. Cook on February 9, 1944. Born in Calif.

Japanese Name	Last Name	First Name	Anglo Name	Incident Date	Year	Age	Residence ID	Family Number
Y	Abe	Kiyotake	Joe	11-4-43	1943		2904-D	40562

B. Manipulating Data

"Computational tools make it possible to efficiently and reliably manipulate large and complex" **archival holdings**. *"Data manipulation includes sorting, filtering, cleaning, normalizing, and joining disparate datasets."*



Japanese Name	Last Name	First Name	Other Name	Incident Date	Year	Age	Residence ID	Family Number
Y	Abe		James	10/5/1942	1942	20	2803-A	
Y	Abe	Makoto		10/7/1942	1943	43	1206-A	
Y	Abe	Sakichi		11/2/1942	1942	62	5315-B	
Y	Abe	Shigeki		5/1/1943	1943	43		
Y	Abe	Kiyotake	Joe	11/4/43	1943		2904-D	40562
N	Jensen	Lloyd	H.	11/16/43	1943			

The Maryland State Archives Presents:

Legacy of Slavery in Maryland

An Archives of Maryland Electronic Publication

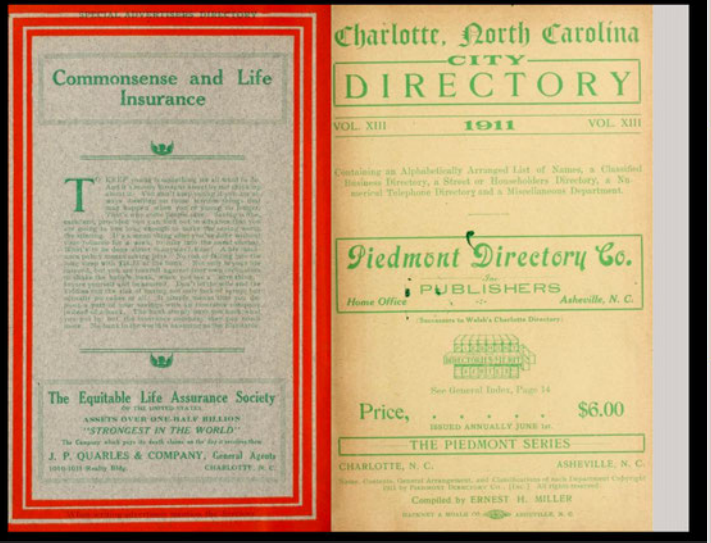
- http://slavery.msa.maryland.gov/
- http://slavery2.msa.maryland.gov/pages/Search.aspx
  - → Runaway Slave Ads: 12,112 records

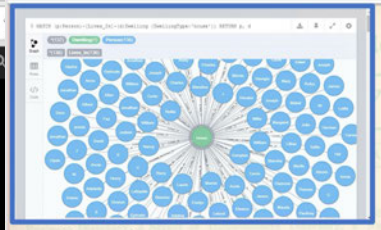
Certif. Freedom: 1. Archival Crawling → 2. Data Cleaning → 3. Data Analysis

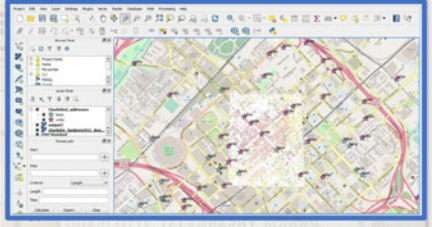
City Dirs.: 1. Internet Archives → 2. Data Cleaning → 3. Data Modeling → 4. Data Analysis

2 files:
- RunawaySlaveAds.xlsx → https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb
- Charlotte_1911_CITY-DIR.txt

https://archive.org/details/charlottenorthca1911pied







Graph Databases:
Alexis, Rosie

Data Visualization:
Emily, Phillip, Sara

Geospatial Transformations:
Cliff, Matt

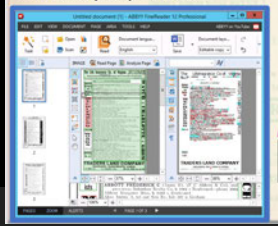
Spring 2021 AI Datathon Showcase

Computational Storytelling Using Jupyter Notebooks

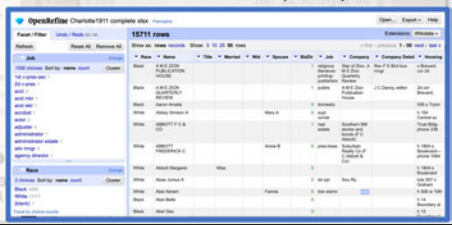
jupyter

Thursday, May 6, 2021:
6:00 – 7:00 PM

Digitization Management:
Annabel, Bo, Britton



Cleaning & Transforming:
Alan, Anna, Maureen



Advanced Information Collaboratory

UNIVERSITY OF MARYLAND 18 56

*Aaron Amelia, domestic 506 s Tryon
Abbey Simeon A (Mary A), supt constr Genl Fire Ext Co, h 104 Central av
ABBOTT F C & CO (F C Abbott), real estate and Southern Mill stocks and bonds, Trust Bldg—phone 238 (see side lines and Index for Adv)
ABBOTT FREDERICK C (Annie B), (F C Abbott & Co), and pres-treas Suburban Realty Co, h 1804 s Boulevard—phone 1064
Abbott Margaret Miss, h 1804 s Boulevard
Abee Junius A, tel opr Sou Ry, bds 507 n Graham
Abel Abram (Fannie), trav slsmn, h 506 w 10th
*Abel Belle, h 14 Boundary al
*Abel Geo, lab, h 12 Boundary al
Abernathy C L, h Lawyer's rd
Abernathy Clement E (Cora), emp Char Lea Belt Co, h 1117 s Tryon
Abernathy Cora Mrs, boarding 1117 s Tryon, h same
Abernathy David M (Enola), h 1310 e 4th ext
Abernathy Elizabeth Miss, h 430 Mint
*Abernathy Hannah, h 4 Bellinger
Abernathy Jno W (Nannie), carp, h Sunnyside
*Abernathy Jos, lab, h 422 w Hill
*Abernathy Lewis (Annie), carp, h 422 w Hill
Abernethy Connie Miss, stengr Burwell & Dunn Co, h 414 Templeton av
Abernethy E Glenn, slsmn Burwell & Dunn Co, h 409 w 11th
Abernethy Gertrude Miss, fitter Little-Long Co, h 306 w 7th
Abernethy J Lee (Ida), painter, h Seversville
Abernethy J S Dr, h Beatty's Ford rd
Abernethy Jas F (Alice), blksmith w Trade ext, h Seversville
Abernethy Lillian Miss, slsldy Efird's Dept Store, h Seversville
Abernethy Margaret K, wid Jas C, h 3 e 1st
Abernethy Mildred Miss, rms 603 n Davidson
Abernethy Nettie I Miss, bkkpr Pound & Moore Co, h 311 n College
Abernethy Thos J (Lucy), emp City, h 414 Templeton av
Abernethy W Leslie, mchst W S Abernethy, h 409 w 11th
Abernethy Wm S (Mamie), auto repr 29 w 4th, h 409 w 11th
Abraham Wm (Lula), h 516 s College
Academy of Music, 210-212 s Tryon; J S Crovo, mngr
Acme Barber Shop, 20½ s Tryon; E R Kirkman propr
ACME PLUMBING CO, 24-26 e 5th; M B Hunter pres, H P Hunter sec-treas
*Adams Anna, tchr, h 1021 s Church
*Adams Belle, maid Realty Bldg, h 419 w 2d
*Adams Berry A (Edith), porter W S Cramer, h Seversville
*Adams Bessie, cook 247 e Trade
Adams Beulah Miss, h 419 Elizabeth av
ADAMS CHAS C (Kate), with 1st Natl Bank, h 905 s Tryon
Adams Compton (Cora), mill hd, h Elizabeth Mills
Adams Dorothy Miss, stengr Oconee Mills Co, rms Y W C A
*Adams Edwd (Mittie), driver Rhyne Bros, h 305 w Palmer
Adams Geo, tailor Henry Miller Jr, bds 15 s Church
Adams Grain & Provision Co, 300 e Trade; J J Adams pres, G H Brockenbrough v-pres, H B Fowler sec-treas
*Adams Henry (Dicie), lab, h 308 Middle

Adams Henry L (Fannie S), route agt So Ry Co, h 327 n Tryon
Adams Jas (Gertrude), mngr, h 419 Elizabeth av
*Adams Jane, tchr, h 1021 s Church
*Adams Jno, lab, h 1031 s Church
Adams Jno J, pres Adams G & P Co, and Char Pepsi-Cola Co, h 309 e 6th
Adams Jno W (Cora), condr S A L Ry, h 21st nr Caldwell
*Adams Jos (Violet), cooper, h 1011 s Church
Adams Jos Q Rev (Leslie), h 1509 s Boulevard
Adams Julia M Miss, h 707 n Church
*Adams Kate, laund, h Groveton
Adams Lafayette N, clk Sou Ry, h 327 n Tryon
Adams Lawrence A, slsmn B S Moore & Co, rms 405 s Tryon
Adams Laurie A (Margaret N), mill hd, h Elizabeth Mills
*Adams Leland, porter J P Stowe & Co, h 403 s Myers
*Adams Lizzie, h 309½ w Morehead
Adams Lula Miss, clk Belk Bros, rms Y W C A
Adams M Grace, wid Geo O, h 601 n College
Adams M Luther (Mamie), h Groveton
*Adams Major, waiter Buford Hotel
*Adams Mattie, h 719 n Graham ext
Adams Pattie V Miss, stengr, bds 708 n Caldwell
*Adams Reuben (Belle), lab Y & B Co, h 419 w 2d
*Adams Rosa, h 714 s Caldwell
*Adams Rufus, lab, h Greenville
*Adams Rufus, driver Stand I & F Co, h Ross Town
Adams Sallie H Miss, asst Carnegie Library, h 707 n Church
ADAMS THADDEUS A (Emma), atty at law, 214-216 Law Bldg—phone 116 and v-pres R G Auten Elec Co, h Clement av, E H—phone 1403
*Adams Violet, servant 508 w Trade
Adams Wheeler F (Mamie), mldr, h 303 s Cedar
Adams Wm E, with The Chronicle, rms 300½ s Church
ADAMS WINSTON D, city editor Charlotte Daily Observer, rms Y M C A
Adcock Jno F, mill wkr, h 916 Calvine av
Adcock Millie M, wid Jas M, h 916 Calvine av
Adelsheimer Henry S (Lizzie), mill wkr, h 1216 Louise av
ADER CHAS E (Berta), cir mngr The Chronicle, h 704 s Church
*Adkins King, lab, 600 s Myers
Adkins Walter D (Leona E), lineman, h (r) 305 e 13th
Aetna Fire Insurance Co of Hartford Conn, 1012 Realty Bldg; F C Clarke, spl agt
*Afro-American Mutual Insurance Co, 412 e 2d; T L Tate pres-treas, T R Mack v-pres, J W Crockett sec-mngr
*Agers Nancy, cook, h 206 Wilson
*Agers Sallie, laund h 420 Jackson
Ahaus Herman (Frances E) tailor 203 w 4th, h 204 s Church
AHRENS FRED W (Laura), v-pres Mutual Bldg & Loan Assn, h 19 e Morehead
Aikel Jos, confr 317 e Trade, rms 225 w Trade
Aiken Geo W M (Barbara), supt Queen City M & G Wks, h 1120 s Caldwell
Aiken Henry, rms 9 e 3d
*Aiken Walter (Ella), lab, h 600 e 2d

*Aiken Wm (Ella), driver W I Henderson Gro Co, h 528 e 8th
Akers Jos J, chf clk Supt Sou Ry, h 908 s Tryon
Albea Jno C, pressmn News Ptg Co, h 310 s McDowell
Albea Lottie Miss, clk Floyd L Liles Co, h 10 w 11th
Albea S Emma, wid J F, h 310 s McDowell
Albrecht Gesha, wid Mathias, h 402 Elizabeth av
Albright Fay L, clk, h 15 s Caldwell
Albright Hal C, clk P O, h 15 s Caldwell
Albright Judson D (Belle C), U S int rev officer, h 15 s Caldwell
Alderman Elizabeth Miss, nurse Char Sanatorium, rms 12 w 7th
Aldred Curtis M (Victoria), mchst, h Seversville
Aldred Jas A (Mattie), clk, h Seversville
Aldred Wm L Rev (Marian), evangelist, h Seversville
Aldridge Millie S, wid Wm A, h Harrill st, Belmont Park
Alexain Nicholas, emp Met Cafe, rms 26 n Tryon
*Alexander Adelaide, h 910 (909) e 1st
*Alexander Adeline, h Lutheran st
Alexander Albert W (Alice), foreman Char Cordage Co, h 621 e 5th
*Alexander Alfred (Willie), driver, h 1016 e 3d
*Alexander Allen (Lizzie), lab, h 6 10th st al
*Alexander Alvey, lab, h Lutheran st
*Alexander Andrew (Bessie), gro 15½ e Boundary, h 514 e Stonewall
*Alexander Andrew W, rms 415 n Myers
Alexander Annie Miss, stengr C M Stieff, h 610 n Graham ext
Alexander Annie L Miss, phys 410 n Tryon, h same
Alexander Arnold E (Alice), laundryman, h 1424 e 5th
*Alexander Bernard, porter Young's Steam Baking Co
*Alexander Bessie, h 1006 e 1st
Alexander Blanche, mad, h 17 Spring
Alexander Blandina Miss, minr J B Ivey Co, h 610 n Graham ext
Alexander Bros (F B and J H), gros and millers, Graham cor w 14th
*Alexander Burdette, porter Colonial Club
Alexander C P, h Providence rd
*Alexander Calvin (Emeline), lab, h 213½ w 1st
Alexander Carl L, clk, h 200 e Oak
*Alexander Carrie, h Gaither's al
*Alexander Carson, chauffeur J E S Davidson, h 209 s Middle
*Alexander Chas, lab, h 410 n A
Alexander Chas F (Mary), clk P O, h 308 e Worthington av
Alexander Chas L (Sue), dentist 913-918 Realty Bldg, h 900 s Tryon
Alexander Chas Y, tchr King's Bus College, rms 305 s Church
Alexander Clarence W, emp J M Porter, h 1116 s Tryon
Alexander Clyde C, painter, h 810 n Brevard
*Alexander Daniel (Etta), lab, h 527 e 8th
*Alexander Decatur C (Alta), porter Colonial Club, h 515 e 1st
Alexander Edna Miss, nurse St Peters Hospital, rms same
*Alexander Edwd (Louvenia), lab, h 1012 e Stonewall
Alexander Eleanor Miss, h 900 s Tryon
*Alexander Eli (Jennie), gro Oil Town, h same
*Alexander Elizabeth, laundress, h (r) 414 s College
*Alexander Ellen, h 606 s McDowell
Alexander Emma V, wid Wm S, h 610 n Graham ext
*Alexander Ephraim (Alice), driver Coles & Smith, h Stonewall bet A & B

Cop Wants $25,000 for Misplaced Asterisk

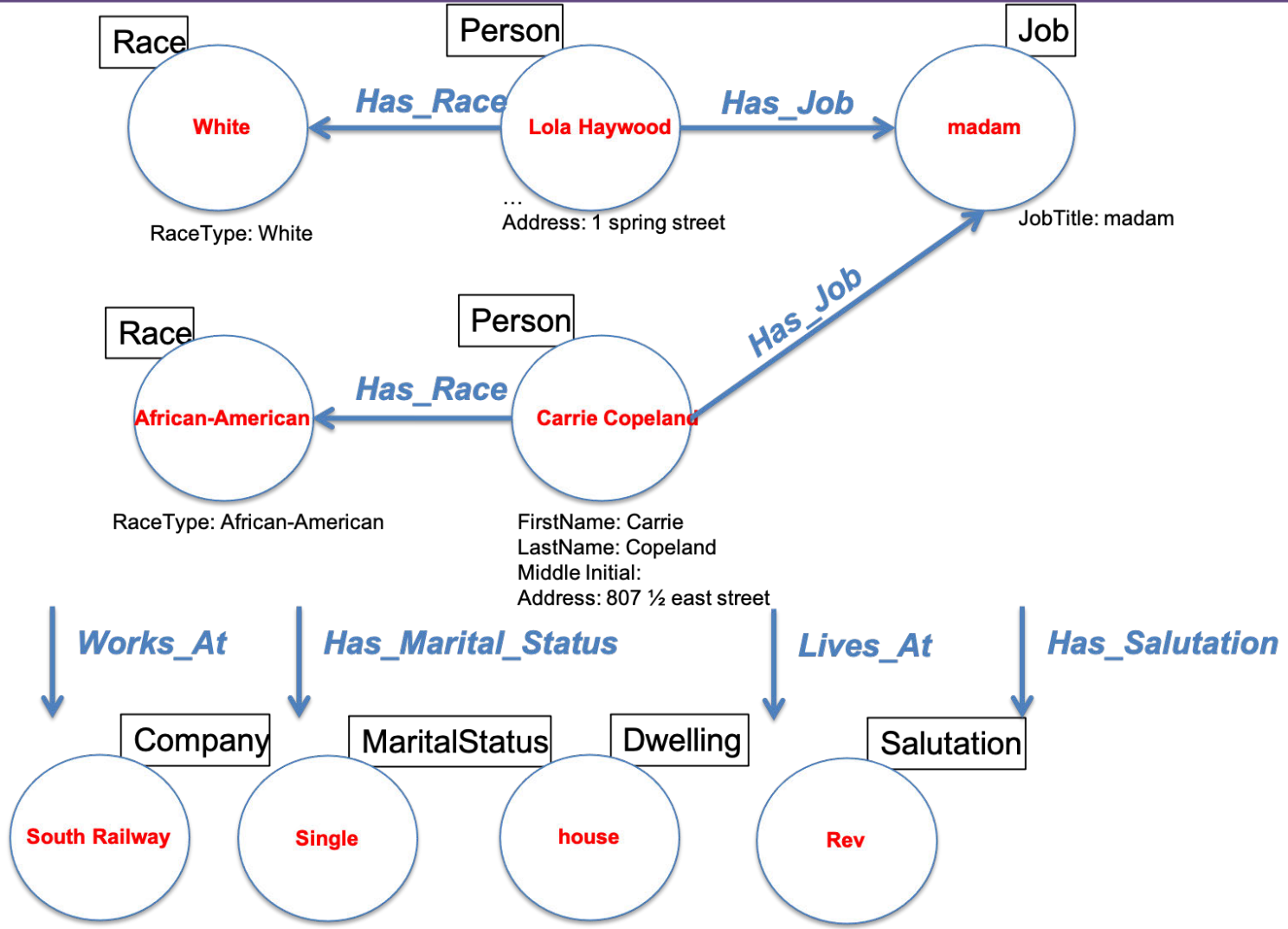
Afro-American (1893-1988); Apr 26, 1930; ProQuest Historical Newspapers: The Baltimore Afro-American pg. 1

Cop Wants $25,000 for Misplaced Asterisk

ASHEVILLE, N. C. —Lon Powers, white, of local police force, has filed suit in County Court against the Commercial Service, inc., the Piedmont Directory Company and the Miller Press, inc., promoters and publishers of the Asheville city directory, for $12,500 punlitve and $12,500 compensatory damages.

In the directory, the policeman said, his name appeared with an asetrisk, the symbol employed to denote citizens of the Negro race.

—o—


Race
Person
Job
White
Has_Race
Lola Haywood
Has_Job
madam
RaceType: White
…
Address: 1 spring street
JobTitle: madam
Race
Person
Has_Job
African-American
Has_Race
Carrie Copeland
RaceType: African-American
FirstName: Carrie
LastName: Copeland
Middle Initial:
Address: 807 ½ east street
Works_At
Has_Marital_Status
Lives_At
Has_Salutation
Company
MaritalStatus
Dwelling
Salutation
South Railway
Single
house
Rev