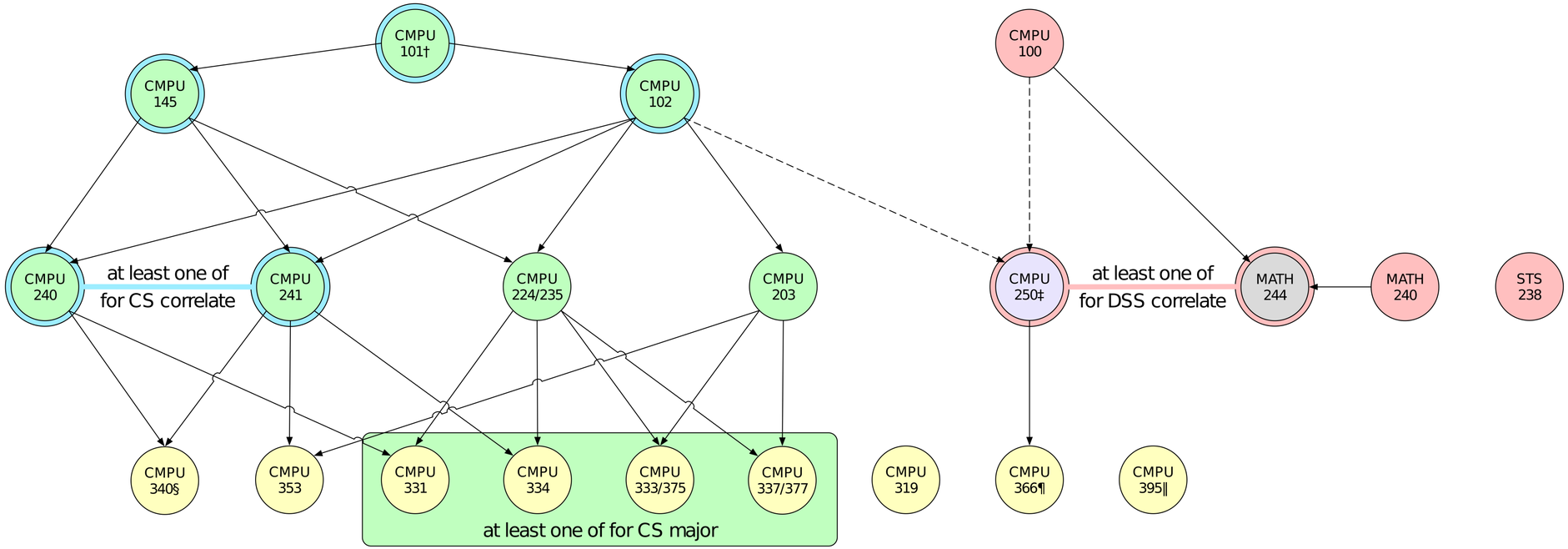
CMPU 100 · Programming with Data

Case Study: Privacy and Data Linking

Class 17



Preregistration suggestion



Computer Science major-required courses (11)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 203 - Software Design and Implementation
CMPU 235 - Computer Systems (previously CMPU 224)
CMPU 240 - Theory of Computation
CMPU 241 - Analysis of Algorithms
CMPU 33X - One course in the 330 range
CMPU 3XX - Two additional 300-level units
CMPU 2/3XXX - One additional elective unit, at the 200 or 300 level

Computer Science correlate-required courses (6)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 240 - Theory of Computation or CMPU 241 - Analysis of Algorithms
CMPU 2xx - Any other 200-level course
CMPU 3xx - Any 300-level course

200-level electives

CMPU 250 - Data and its Discontents

300-level electives

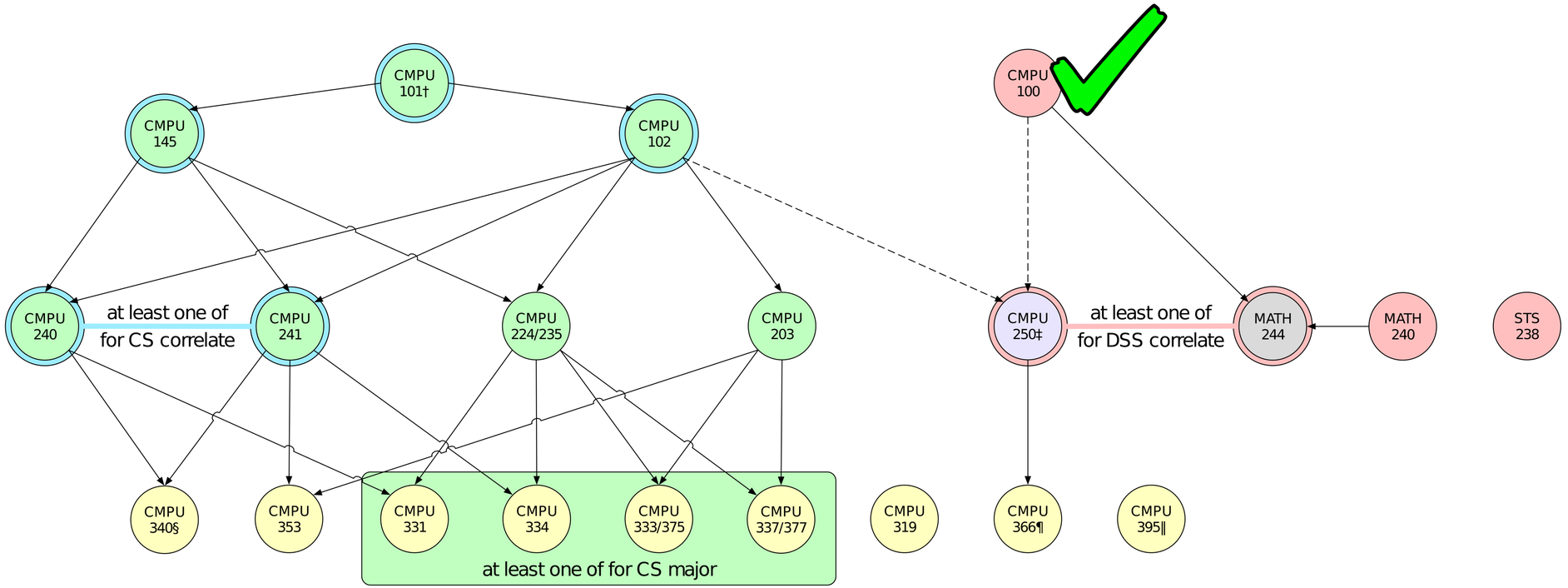
CMPU 319 - Modeling Minds, Brains, and Behavior
CMPU 331 - Compilers
CMPU 333 - Computer Networks (previously CMPU 375)
CMPU 334 - Operating Systems
CMPU 337 - Parallel Computing (previously CMPU 377)
CMPU 340 - Discrete Optimization
CMPU 353 - Bioinformatics
CMPU 366 - Natural Language Processing
CMPU 395 - Advanced Special Topics

Data Science & Society correlate-required courses (5)

CMPU 100 - Programming with Data
MATH 240 - Introduction to Statistics
STS 238 - Data and Society
CMPU 250 - Data and its Discontents or MATH 244 - Intermediate Data Science
Approved 300-level capstone

Notes

* At least two CMPU 200-level courses required for any CMPU 300-level course.
† Students can place out of CMPU 101 by taking an exam - contact department for details.
‡ CMPU 250 requires CMPU 100 or 102.
§ MATH 263 can replace CMPU 241 prerequisite, MATH 241 or 261 can replace CMPU 240.
¶ MATH 244 can replace CMPU 250 prerequisite.
‖ Prerequisites vary by semester-specif c topic, see Course Catalogue.



Computer Science major-required courses (11)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 203 - Software Design and Implementation
CMPU 235 - Computer Systems (previously CMPU 224)
CMPU 240 - Theory of Computation
CMPU 241 - Analysis of Algorithms
CMPU 33X - One course in the 330 range
CMPU 3XX - Two additional 300-level units
CMPU 2/3XXX - One additional elective unit, at the 200 or 300 level

Computer Science correlate-required courses (6)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 240 - Theory of Computation or CMPU 241 - Analysis of Algorithms
CMPU 2xx - Any other 200-level course
CMPU 3xx - Any 300-level course

200-level electives

CMPU 250 - Data and its Discontents

300-level electives

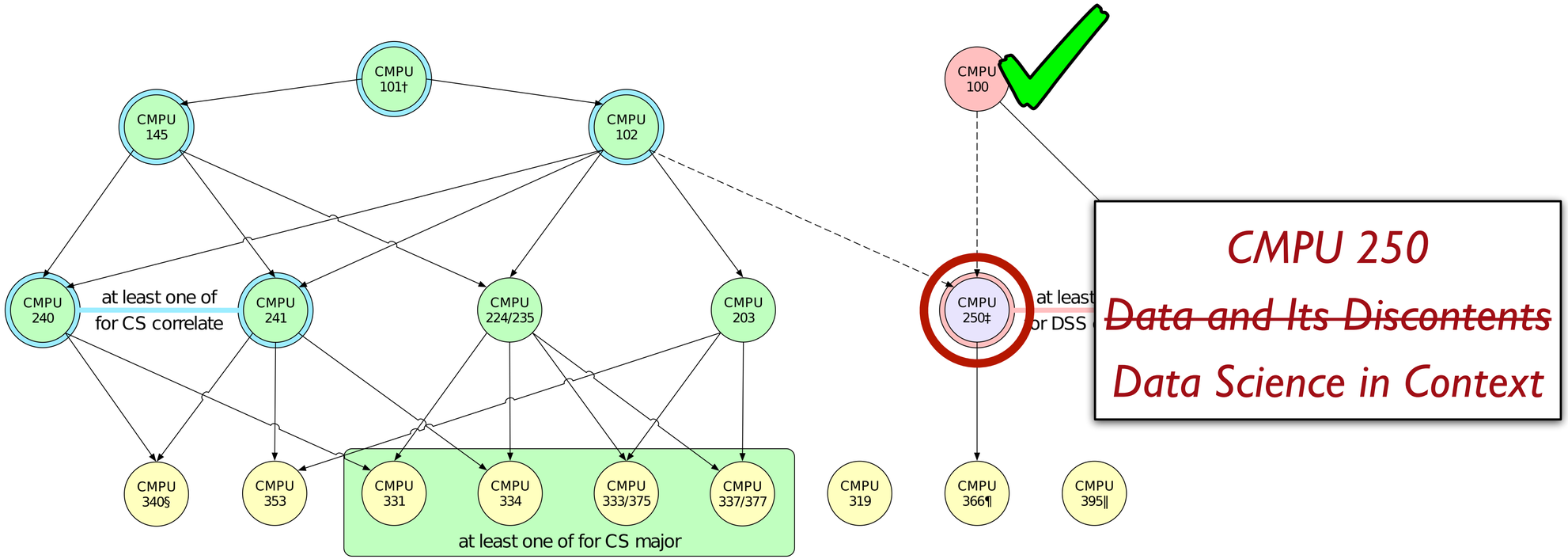
CMPU 319 - Modeling Minds, Brains, and Behavior
CMPU 331 - Compilers
CMPU 333 - Computer Networks (previously CMPU 375)
CMPU 334 - Operating Systems
CMPU 337 - Parallel Computing (previously CMPU 377)
CMPU 340 - Discrete Optimization
CMPU 353 - Bioinformatics
CMPU 366 - Natural Language Processing
CMPU 395 - Advanced Special Topics

Data Science & Society correlate-required courses (5)

CMPU 100 - Programming with Data
MATH 240 - Introduction to Statistics
STS 238 - Data and Society
CMPU 250 - Data and its Discontents or MATH 244 - Intermediate Data Science
Approved 300-level capstone

Notes

* At least two CMPU 200-level courses required for any CMPU 300-level course.
† Students can place out of CMPU 101 by taking an exam - contact department for details.
‡ CMPU 250 requires CMPU 100 or 102.
§ MATH 263 can replace CMPU 241 prerequisite, MATH 241 or 261 can replace CMPU 240.
¶ MATH 244 can replace CMPU 250 prerequisite.
‖ Prerequisites vary by semester-specif c topic, see Course Catalogue.



Computer Science major-required courses (11)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 203 - Software Design and Implementation
CMPU 235 - Computer Systems (previously CMPU 224)
CMPU 240 - Theory of Computation
CMPU 241 - Analysis of Algorithms
CMPU 33X - One course in the 330 range
CMPU 3XX - Two additional 300-level units
CMPU 2/3XXX - One additional elective unit, at the 200 or 300 level

Computer Science correlate-required courses (6)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 240 - Theory of Computation or CMPU 241 - Analysis of Algorithms
CMPU 2xx - Any other 200-level course
CMPU 3xx - Any 300-level course

200-level electives

CMPU 250 - Data and its Discontents

300-level electives

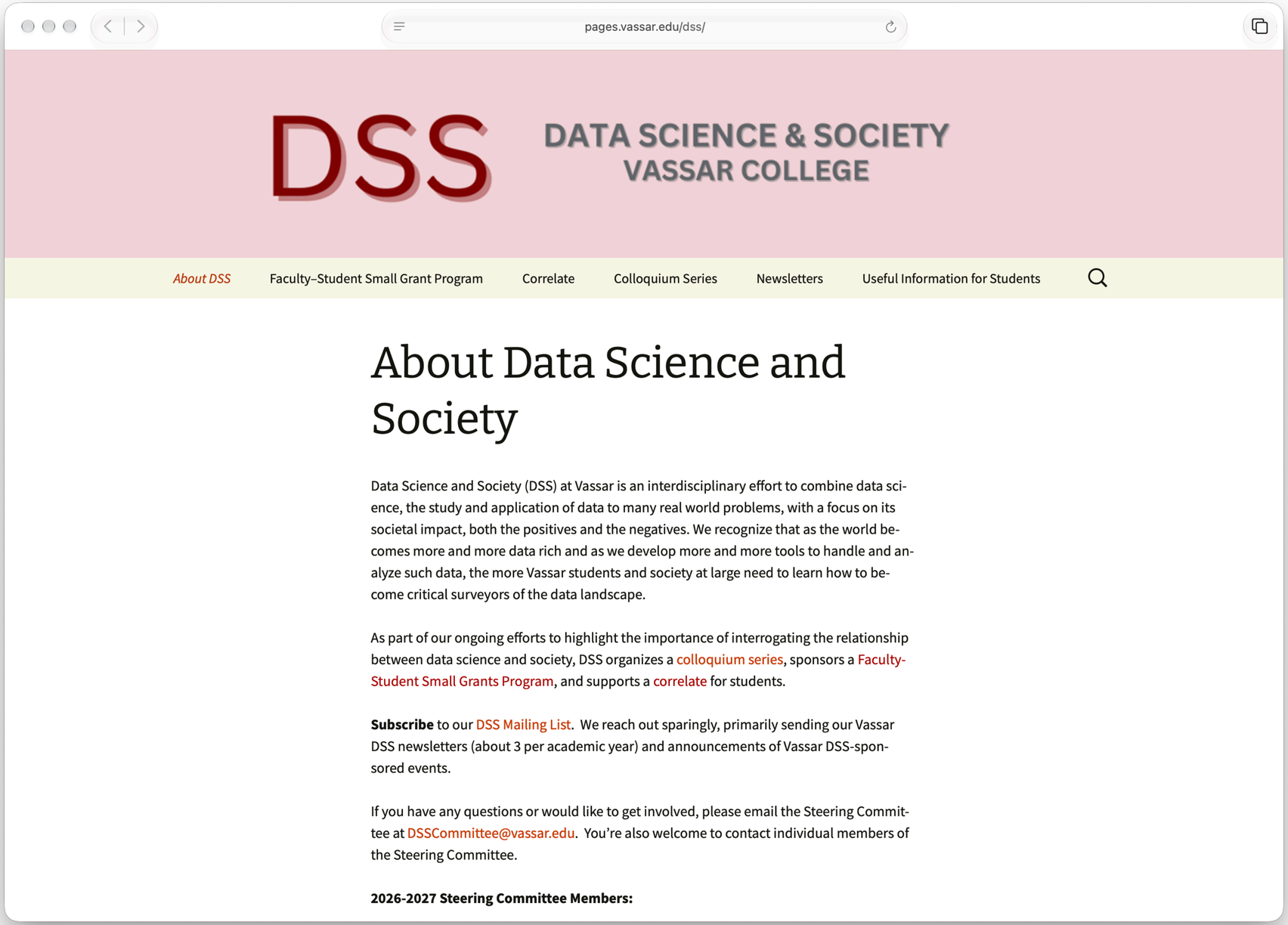
CMPU 319 - Modeling Minds, Brains, and Behavior
CMPU 331 - Compilers
CMPU 333 - Computer Networks (previously CMPU 375)
CMPU 334 - Operating Systems
CMPU 337 - Parallel Computing (previously CMPU 377)
CMPU 340 - Discrete Optimization
CMPU 353 - Bioinformatics
CMPU 366 - Natural Language Processing
CMPU 395 - Advanced Special Topics

Data Science & Society correlate-required courses (5)

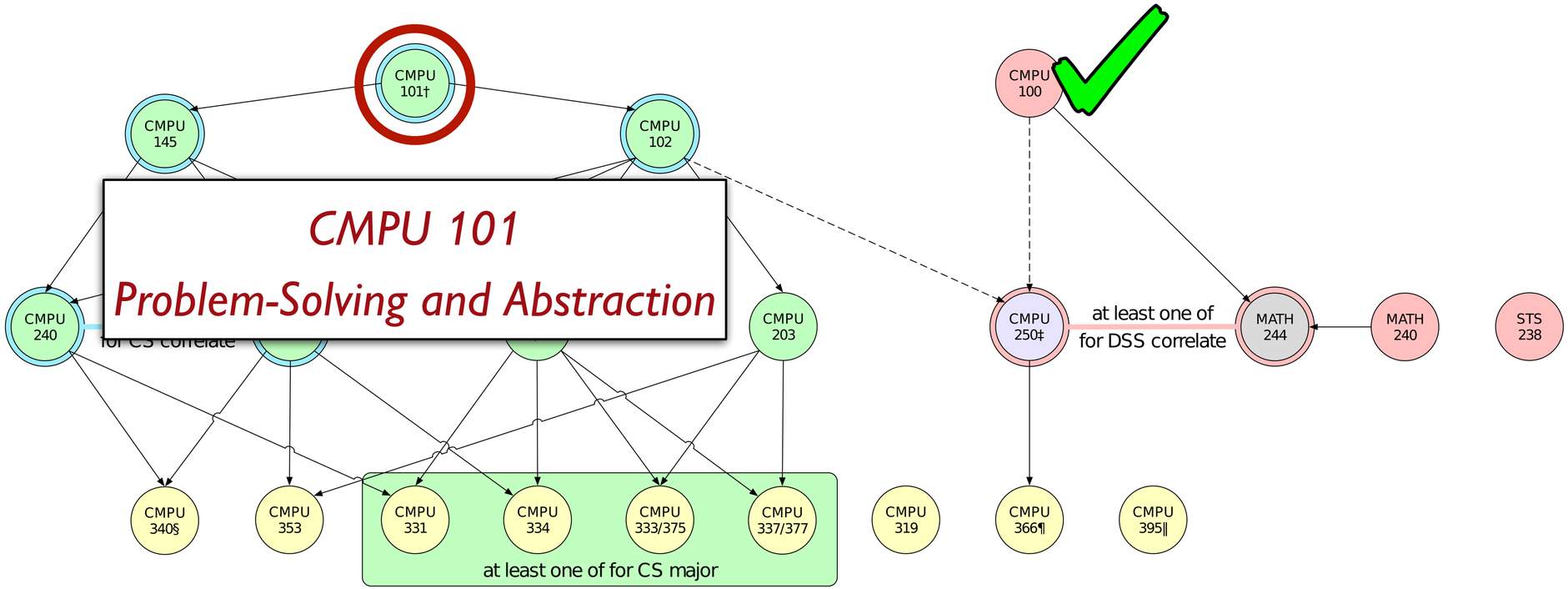
CMPU 100 - Programming with Data
MATH 240 - Introduction to Statistics
STS 238 - Data and Society
CMPU 250 - Data and its Discontents or MATH 244 - Intermediate Data Science
Approved 300-level capstone

Notes

* At least two CMPU 200-level courses required for any CMPU 300-level course.
† Students can place out of CMPU 101 by taking an exam - contact department for details.
‡ CMPU 250 requires CMPU 100 or 102.
§ MATH 263 can replace CMPU 241 prerequisite, MATH 241 or 261 can replace CMPU 240.
¶ MATH 244 can replace CMPU 250 prerequisite.
‖ Prerequisites vary by semester-specif c topic, see Course Catalogue.



pages.vassar.edu/dss



Computer Science major-required courses (11)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 203 - Software Design and Implementation
CMPU 235 - Computer Systems (previously CMPU 224)
CMPU 240 - Theory of Computation
CMPU 241 - Analysis of Algorithms
CMPU 33X - One course in the 330 range
CMPU 3XX - Two additional 300-level units
CMPU 2/3XXX - One additional elective unit, at the 200 or 300 level

Computer Science correlate-required courses (6)

CMPU 101 - Problem-Solving and Abstraction
CMPU 102 - Data Structures and Algorithms
CMPU 145 - Foundations of Computer Science
CMPU 240 - Theory of Computation or CMPU 241 - Analysis of Algorithms
CMPU 2xx - Any other 200-level course
CMPU 3xx - Any 300-level course

200-level electives

CMPU 250 - Data and its Discontents

300-level electives

CMPU 319 - Modeling Minds, Brains, and Behavior
CMPU 331 - Compilers
CMPU 333 - Computer Networks (previously CMPU 375)
CMPU 334 - Operating Systems
CMPU 337 - Parallel Computing (previously CMPU 377)
CMPU 340 - Discrete Optimization
CMPU 353 - Bioinformatics
CMPU 366 - Natural Language Processing
CMPU 395 - Advanced Special Topics

Data Science & Society correlate-required courses (5)

CMPU 100 - Programming with Data
MATH 240 - Introduction to Statistics
STS 238 - Data and Society
CMPU 250 - Data and its Discontents or MATH 244 - Intermediate Data Science
Approved 300-level capstone

Notes

* At least two CMPU 200-level courses required for any CMPU 300-level course.
† Students can place out of CMPU 101 by taking an exam - contact department for details.
‡ CMPU 250 requires CMPU 100 or 102.
§ MATH 263 can replace CMPU 241 prerequisite, MATH 241 or 261 can replace CMPU 240.
¶ MATH 244 can replace CMPU 250 prerequisite.
‖ Prerequisites vary by semester-specif c topic, see Course Catalogue.

Data and privacy

“If I’m not doing anything wrong, I don’t need to hide anything from other people”

SYSTEM ERROR



WHERE BIG TECH WENT WRONG AND HOW WE CAN REBOOT

ROB REICH
MEHRAN SAHAMI
JEREMY M. WEINSTEIN



Case study: Taylor Swift

ENTERTAINMENT + CELEBRITIES +

Man Arrested Three Times in Five Days for Allegedly Stalking, Harassing Taylor Swift

ADD TIME ON GOOGLE

by Anna Gordon and Koh Ewe

Updated: JAN 19, 2026 5:46 AM ET • Published: JAN 23, 2024 8:16 AM ET



Taylor Swift: A man has been arrested 3 times in 5 days for stalking near the singer's New York City home. *GC Images/Gotham/Getty Images*

It was only a couple hours after leaving court with pending stalking and harassment charges that David Crowe, a 33-year-old man from Seattle, was arrested on Wednesday for a third time in five days for suspiciously lurking near Taylor Swift's New York City home.

The court had just granted Crowe supervised release before he promptly allegedly violated an order of protection the court had also granted, NBC reported. Prosecutors already alleged that Crowe has shown up to Swift's building 30 times in the past two months despite being repeatedly asked to leave.



Private www.instagram.com/p/BoCAq2wFgGO

Instagram Log In Sign Up



ismconnect • Follow

ismconnect 423w
We only have a couple shows left to shake it off, come out and say 👋

ismconnect 423w
.
.
.
.
#EventLife #WinWithISM #EventPros #Activate #Experiential #OnTour #LiveEvent #SummerConcert #Concert #LiveMusic #DigitalOutOfHome #Advertising #DOOH #Marketing #MarketingAnalytics #MeasuredResults

Like Reply

44 1

September 22, 2018

Log in to like or comment.

Private

Camera

www.instagram.com/p/BoCAq2wFgGO

Instagram

Log In Sign Up



ismconnect • Follow

ismconnect 423w
We only have a couple shows left to shake it off, come out and say 👋

ismconnect 423w
.
.
.
.
#EventLife #WinWithISM #EventPros #Activate #Experiential #OnTour #LiveEvent #SummerConcert #Concert #LiveMusic #DigitalOutOfHome #Advertising #DOOH #Marketing #MarketingAnalytics #MeasuredResults

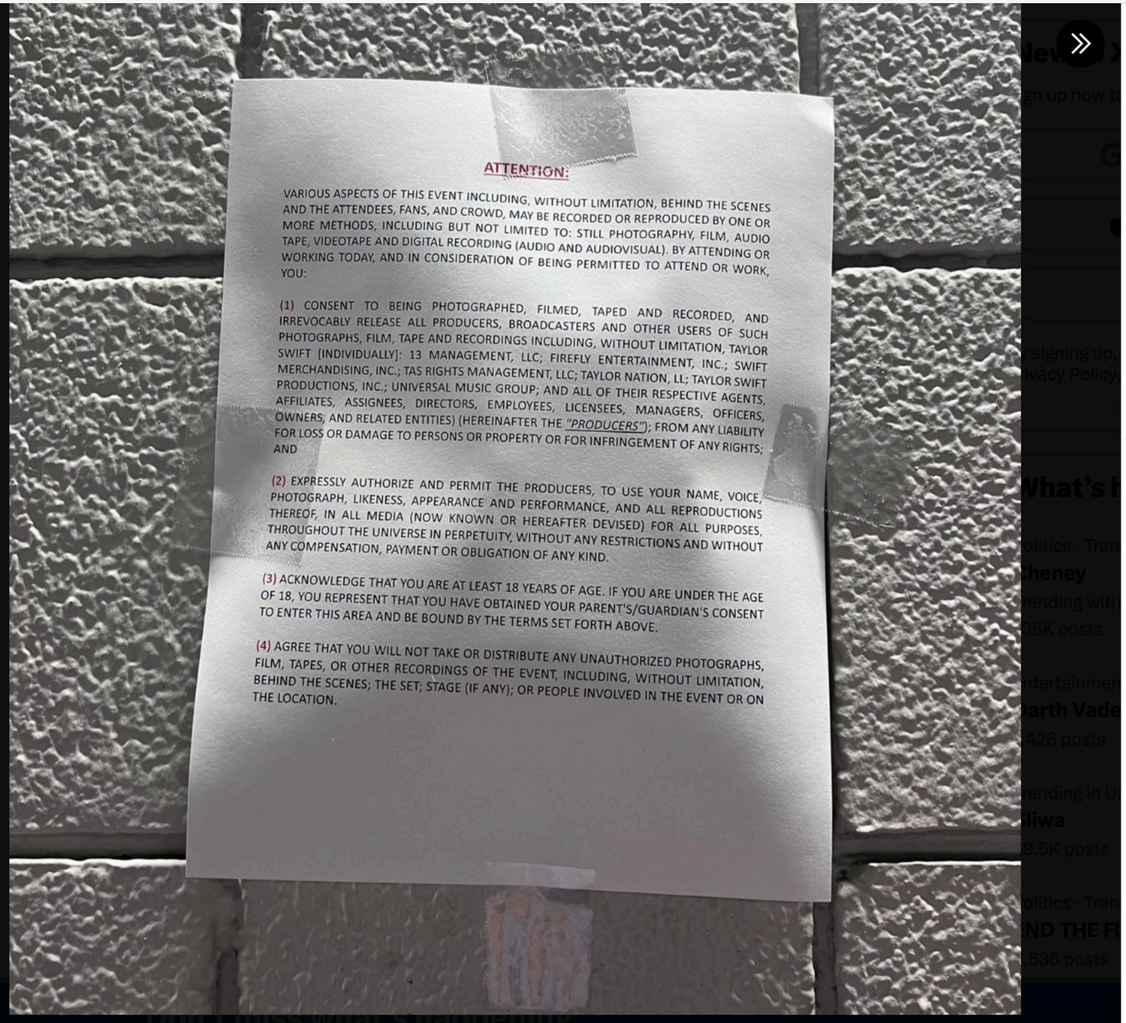
Like Reply

44 1

September 22, 2018

Log in to like or comment.

x.com/TSwiftEdits_13/status/1847409992715927576/photo/2



ATTENTION:

VARIOUS ASPECTS OF THIS EVENT INCLUDING, WITHOUT LIMITATION, BEHIND THE SCENES AND THE ATTENDEES, FANS, AND CROWD, MAY BE RECORDED OR REPRODUCED BY ONE OR MORE METHODS, INCLUDING BUT NOT LIMITED TO: STILL PHOTOGRAPHY, FILM, AUDIO TAPE, VIDEOTAPE AND DIGITAL RECORDING (AUDIO AND AUDIOVISUAL). BY ATTENDING OR WORKING TODAY, AND IN CONSIDERATION OF BEING PERMITTED TO ATTEND OR WORK, YOU:

(1) CONSENT TO BEING PHOTOGRAPHED, FILMED, TAPED AND RECORDED, AND IRREVOCABLY RELEASE ALL PRODUCERS, BROADCASTERS AND OTHER USERS OF SUCH PHOTOGRAPHS, FILM, TAPE AND RECORDINGS INCLUDING, WITHOUT LIMITATION, TAYLOR SWIFT (INDIVIDUALLY): 13 MANAGEMENT, LLC; FIREFLY ENTERTAINMENT, INC.; SWIFT MERCHANDISING, INC.; TAS RIGHTS MANAGEMENT, LLC; TAYLOR NATION, LL; TAYLOR SWIFT PRODUCTIONS, INC.; UNIVERSAL MUSIC GROUP; AND ALL OF THEIR RESPECTIVE AGENTS, AFFILIATES, ASSIGNEES, DIRECTORS, EMPLOYEES, LICENSEES, MANAGERS, OFFICERS, OWNERS, AND RELATED ENTITIES) (HEREINAFTER THE "PRODUCERS"); FROM ANY LIABILITY FOR LOSS OR DAMAGE TO PERSONS OR PROPERTY OR FOR INFRINGEMENT OF ANY RIGHTS; AND

(2) EXPRESSLY AUTHORIZE AND PERMIT THE PRODUCERS, TO USE YOUR NAME, VOICE, PHOTOGRAPH, LIKENESS, APPEARANCE AND PERFORMANCE, AND ALL REPRODUCTIONS THEREOF, IN ALL MEDIA (NOW KNOWN OR HEREAFTER DEVISED) FOR ALL PURPOSES, THROUGHOUT THE UNIVERSE IN PERPETUITY, WITHOUT ANY RESTRICTIONS AND WITHOUT ANY COMPENSATION, PAYMENT OR OBLIGATION OF ANY KIND.

(3) ACKNOWLEDGE THAT YOU ARE AT LEAST 18 YEARS OF AGE. IF YOU ARE UNDER THE AGE OF 18, YOU REPRESENT THAT YOU HAVE OBTAINED YOUR PARENT'S/GUARDIAN'S CONSENT TO ENTER THIS AREA AND BE BOUND BY THE TERMS SET FORTH ABOVE.

(4) AGREE THAT YOU WILL NOT TAKE OR DISTRIBUTE ANY UNAUTHORIZED PHOTOGRAPHS, FILM, TAPES, OR OTHER RECORDINGS OF THE EVENT, INCLUDING, WITHOUT LIMITATION, BEHIND THE SCENES; THE SET; STAGE (IF ANY); OR PEOPLE INVOLVED IN THE EVENT OR ON THE LOCATION.

17 379 5.2K 150K

Taylor Swift Edits
@TSwiftEdits_13

🚨The filming notice for tonight show in Miami in the same as the one used to film 'The Eras Tour Film' in LA last year!!!#MiamiTSTheErasTour

6:50 PM · Oct 18, 2024 · **150.4K** Views

17 379 5.2K 191

Read 17 replies



Case study:
White House protestor

"[Her] mother had won a prize for her work as a schoolteacher and was invited to the White House with other teachers for an award ceremony. Upon presenting her credentials at the White House gate, she had been denied entry. When she asked why, she learned that her face had come up in a database of people who had participated in public protests against President Trump…"

Weinstein, Sahami, and Reich, *System Error: Where Big Tech Went Wrong and How We Can Reboot*, 2021



"*She had done nothing wrong. She had nothing to hide.* Her face had been recorded while she was behaving lawfully in public… It's easy to see that with surveillance comes a potential loss of essential liberties in a democratic society, such as a freedom to protest and freedom of expression."

Weinstein, Sahami, and Reich, *System Error: Where Big Tech Went Wrong and How We Can Reboot*, 2021

“But the data was anonymized…”

Anonymized ≠ anonymous



Latanya Sweeney

Photograph by Kayana-Szmczak

In the 1990s, the state of Massachusetts – through the Group Insurance Commission (GIC) – released records of state employees' hospital visits, to use for research.

Removed: names, addresses, SSNs

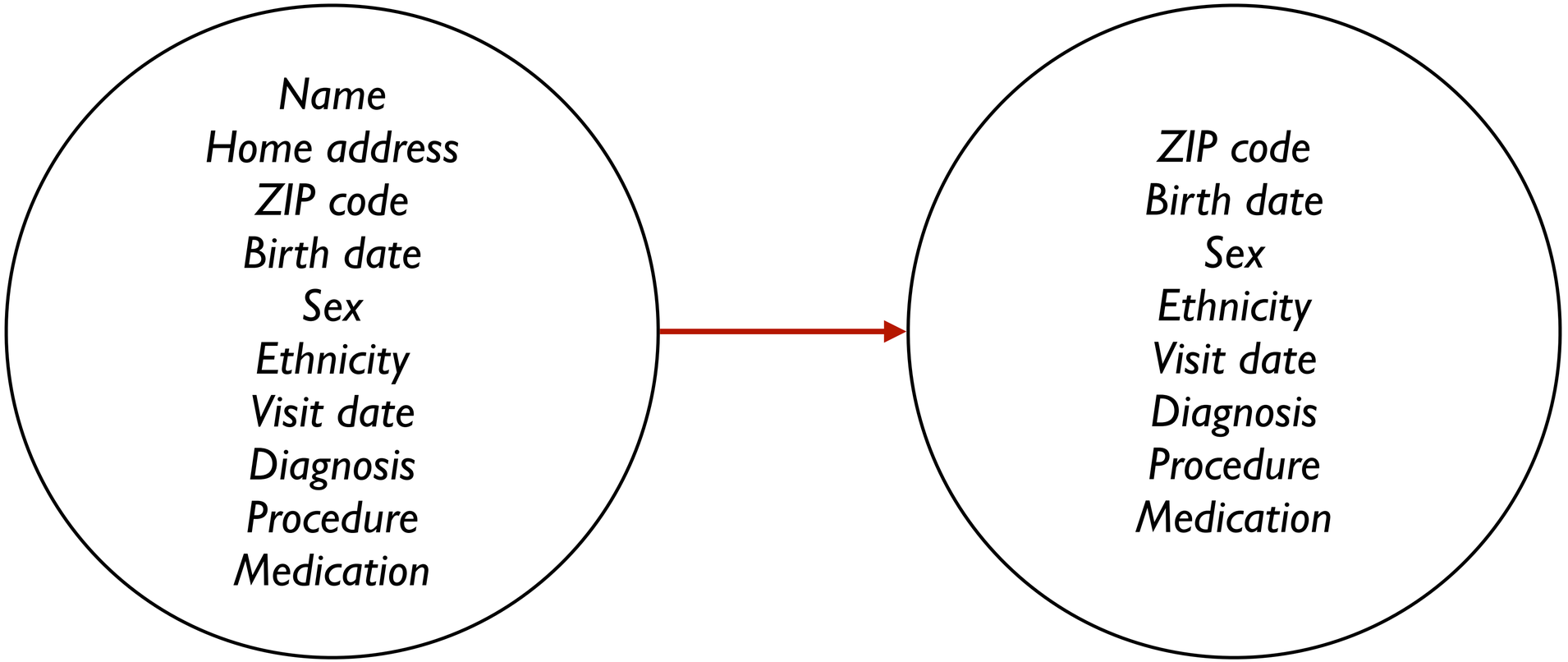
Kept: ZIP code, birth date, sex, diagnoses, procedures, medications

In the 1990s, the state of Massachusetts – through the Group Insurance Commission (GIC) – released records of state employees' hospital visits, to use for research.

Removed: names, addresses, SSNs

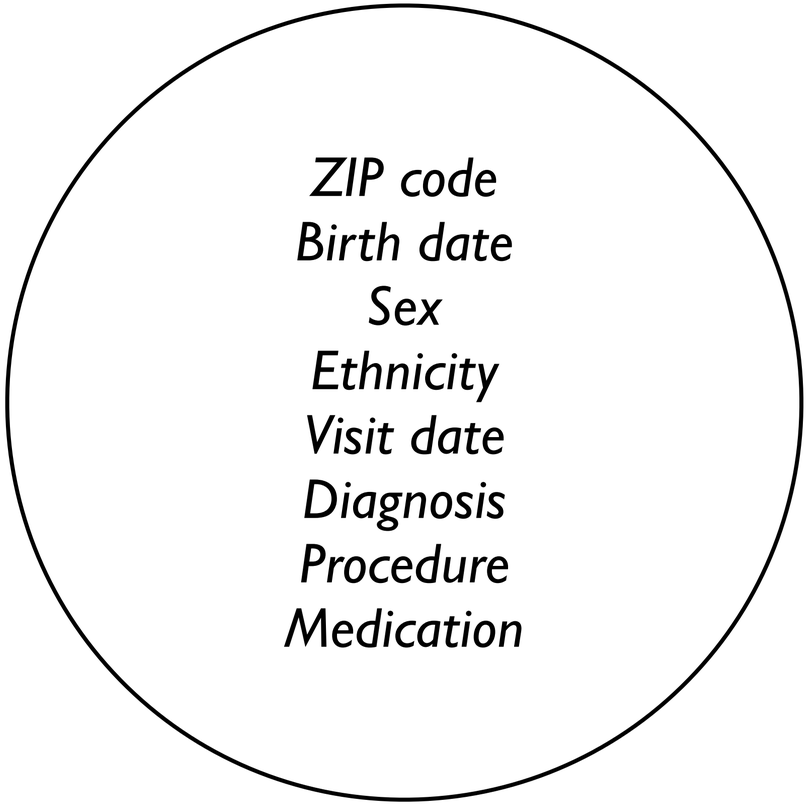
Kept: ZIP code, birth date, sex, diagnoses, procedures, medications

Why?



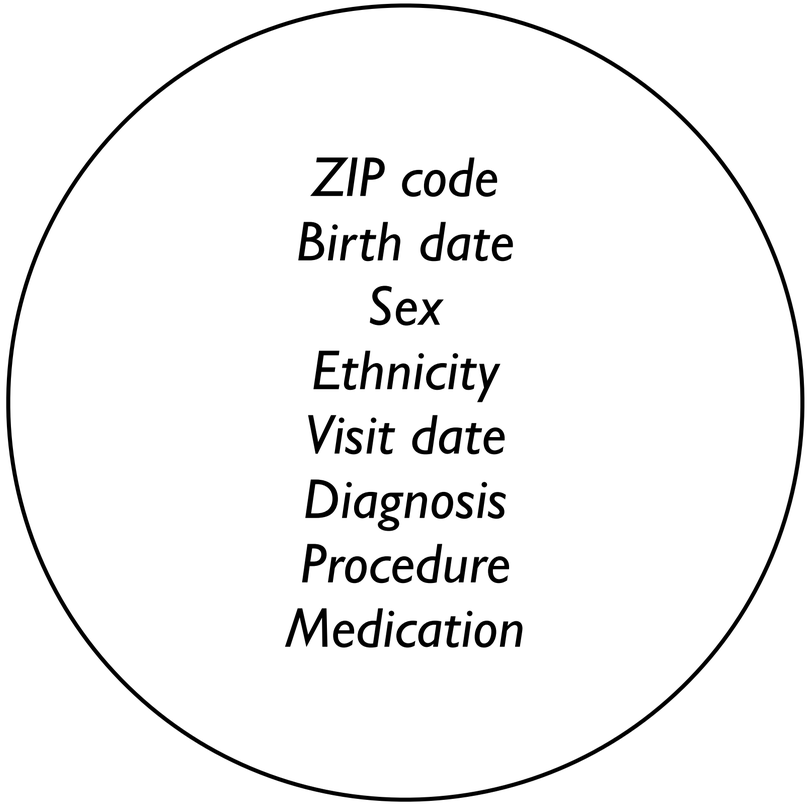
Medical data

De-identified medical data



De-identified medical data

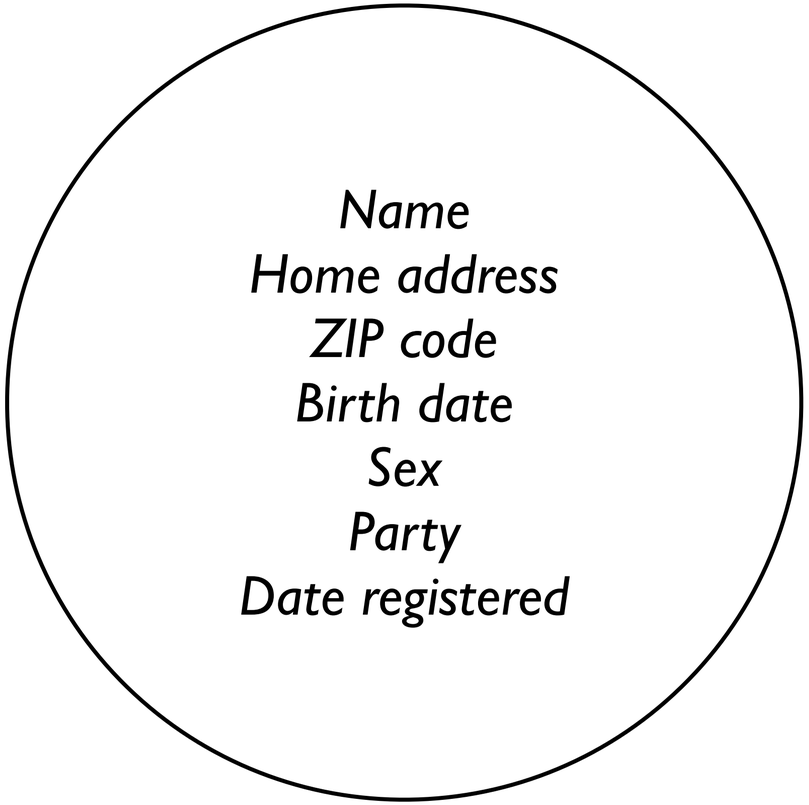
Public!

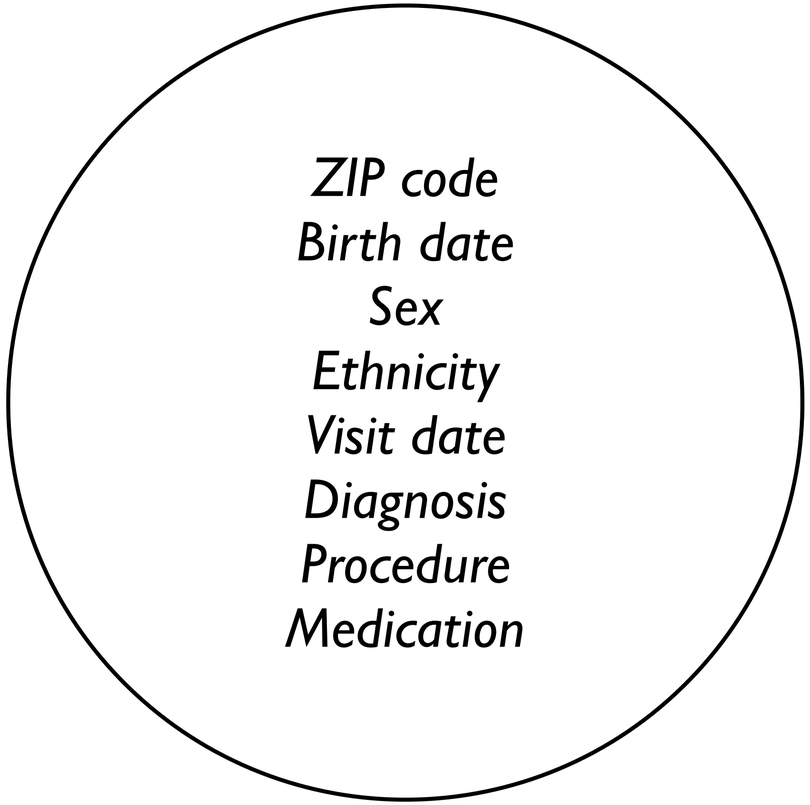


De-identified medical data

Bought for $20!

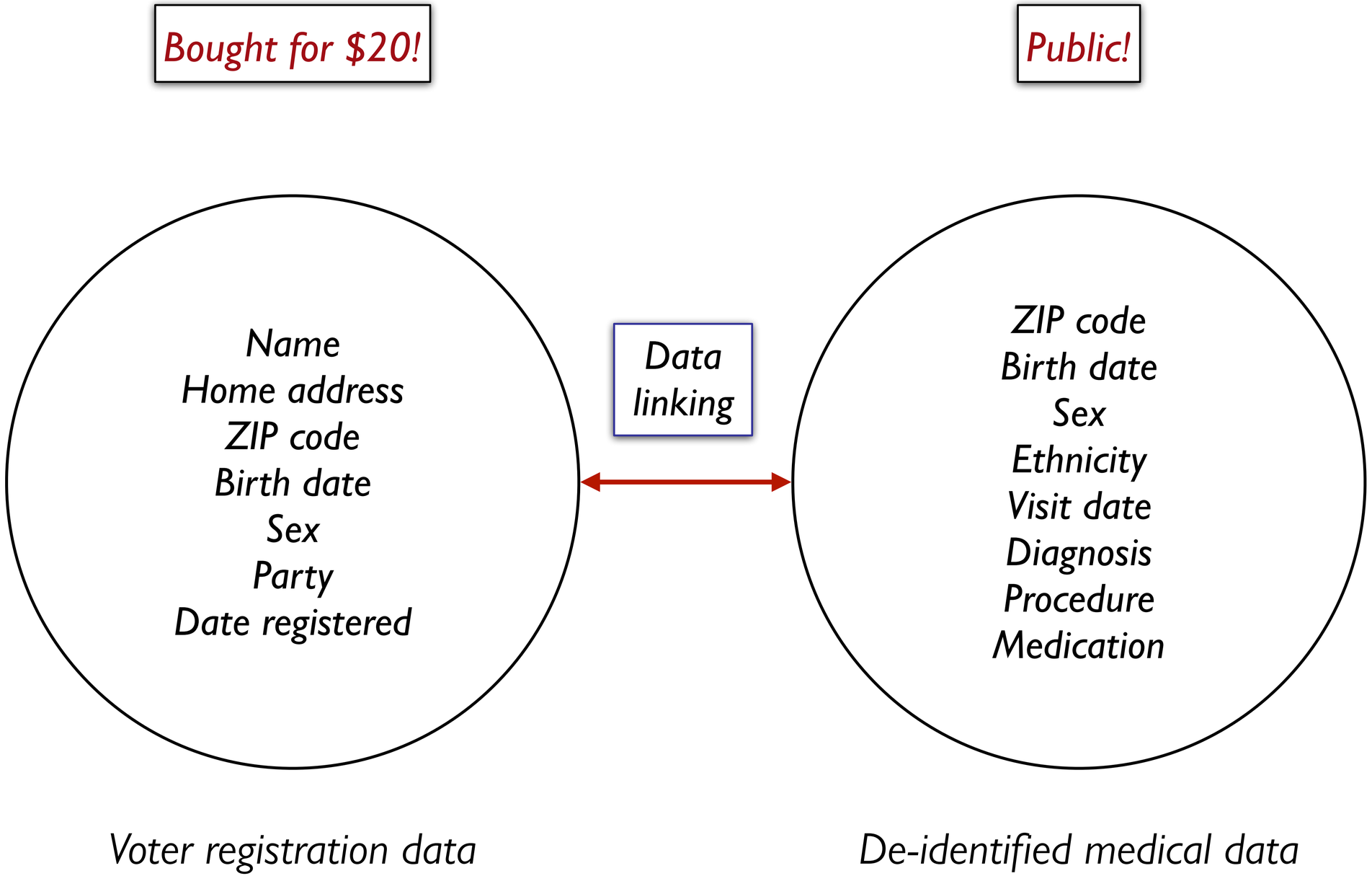
Public!

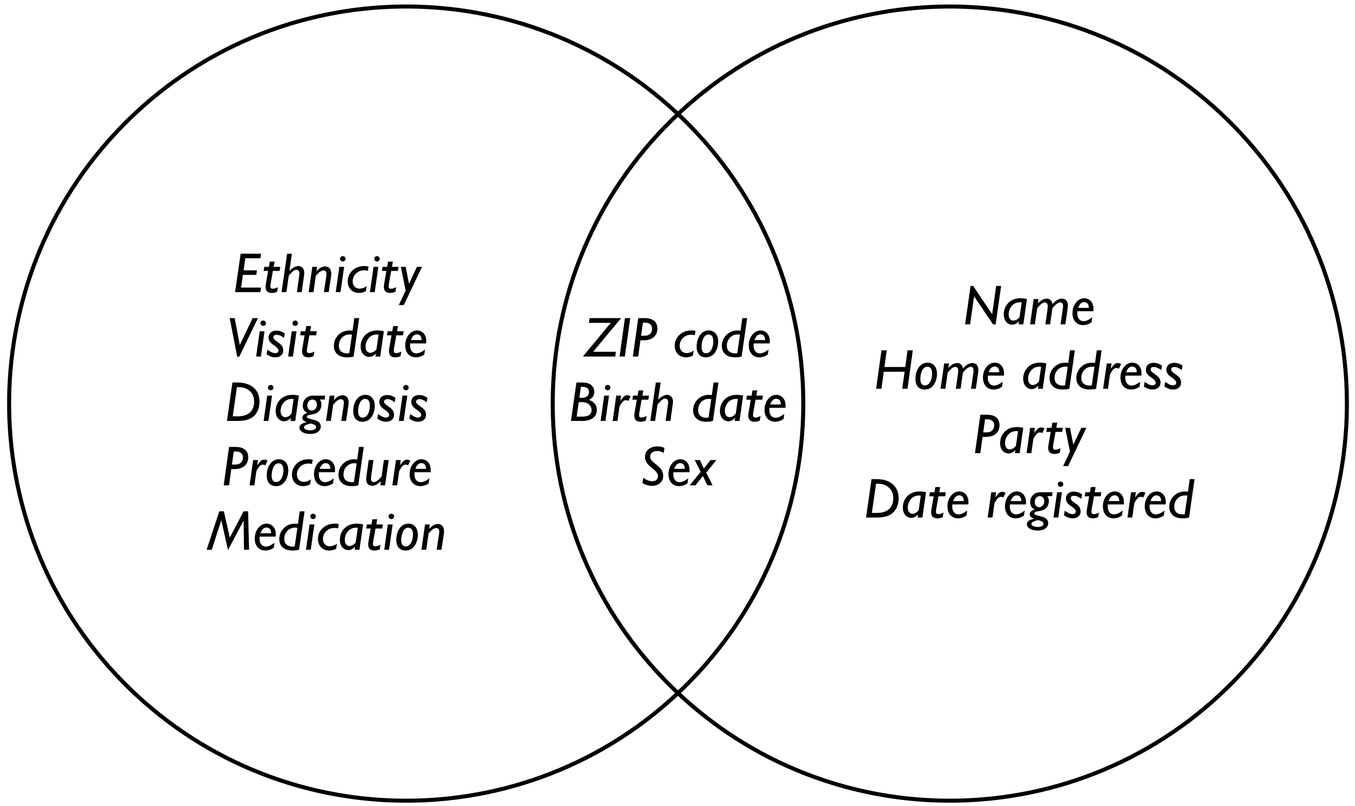




Voter registration data

De-identified medical data


Bought for $20!
Public!
Name
Home address
ZIP code
Birth date
Sex
Party
Date registered
Data linking
ZIP code
Birth date
Sex
Ethnicity
Visit date
Diagnosis
Procedure
Medication
Voter registration data
De-identified medical data



Re-*identified medical data*

Governor William Weld lived in Cambridge and, as a state employee, his medical records were included in the anonymized GIC dataset.

According to the voter list:

- 6 people shared his birthdate
- Only 3 of them were men
- Only 1 of those men lived in his ZIP code.

Sweeney mailed the governor his own medical records.



Photograph by Brian Snyder / Reuters

How unique are you?

Sweeney (2000): 87% of Americans are the only person with their ZIP code + birth date + sex.

> A later re-analysis with 2000 Census data (Golle, 2006) found 63% – still a majority

These are *quasi-identifiers*: harmless alone, identifying together

Your turn: be the attacker

→ Notebook: *Part 1: Be the attacker*

Predict

What percentage of voters are unique on ZIP + birth date + sex?

A. Under 25%

B. 25–50%

C. 50–75%

D. Over 75%

→ Notebook: *Part 2: How unique are we?*

Your turn: be the Census Bureau

This summer's policy change

June 4, 2026: Commerce Department order DAO 216-26

Bans "noise infusion" in Census Bureau and BEA data products

- Including differential privacy, used for the 2020 Census

Preferred method instead: coarsening (rounding, ranges, aggregation)

- Suppression (withholding data) as a last resort

The 2025 American Community Survey release slipped from September to late October

Coarsening isn't new

HIPAA's "Safe Harbor" rule (2000) for sharing medical data:

- Dates: keep only the year
- ZIP codes: keep only the first 3 digits (000 if that area has ≤ 20,000 people)
- Ages over 89: lump into "90+"

Partly a response to attacks like Sweeney's!

→ Notebook: *Part 3: Be the Census Bureau*

Another way: add noise

Why the Census Bureau added noise

Census tables only show counts – no names, no individual rows

But billions of published counts are a giant system of equations

The Bureau attacked its own 2010 data:

- Reconstructed 46% of people's records exactly (block, age, sex, race, ethnicity)
- Linked those to commercial data and re-identified about 52 million people

So, for 2020: differential privacy – add carefully calibrated random noise

Activity: A private survey

Question:

Have you ever turned in an assignment where you used AI when it wasn't allowed?

Coin Flipper

This form allows you to flip virtual coins. The randomness comes from atmospheric noise, which for many purposes is better than the pseudo-random number algorithms typically used in computer programs.

Flip 1 virtual coin(s) of type US 1¢ Penny

Flip Coin(s) Reset Form

It is not always easy to decide what is heads and tails on a given coin. Numismatics (the scientific study of money) defines the obverse and reverse of a coin rather than heads and tails. The obverse (principal side) of a coin typically features a symbol intended to be evocative of stately power, such as the head of a monarch or well-known state representative. In the case of coins that do not have royalty or state representatives on them, the side that features the name of the country is usually considered the obverse.

Thanks to lots of helpful people for donating coin pictures!

If you don't see your currency listed here and you have good quality coin pictures that you would like to donate, feel free to email us. (Remember that you must own the copyright to the pictures, or they must be in the public domain.)

© 1998-2026 RANDOM.ORG

random.org/coins

Flip two coins and don't show anyone.

Question:

Have you ever turned in an assignment where you used AI when it wasn't allowed?

First coin is heads? Answer truthfully.

First coin is tails? Look at the second coin. If it's heads, say yes. If it's tails, say "no".

Estimating from noisy answers

Half of you answered truthfully; a quarter of you were forced to say “yes”.

$P(\text{yes}) = \frac{1}{2} \cdot (\text{true rate}) + \frac{1}{4}$

So the true rate $= 2 \cdot P(\text{yes}) - \frac{1}{2}$

Everyone gets plausible deniability, and we still learn about the group

Randomized response: Warner (1965) – the seed of differential privacy.

→ Notebook: *Sneak preview*

You're running the 2030 Census

A. Exact counts

+ Accurate and detailed

− Easy to link and re-identify

B. Coarsen and suppress

+ Exact numbers that are easy to explain

− Small places and groups lose detail or vanish

C. Add noise

+ Full detail, provable privacy

− Small-area numbers can be off; hard to explain

“If I’m not doing anything wrong, I don’t need to hide anything from other people”

Acknowledgments

This class incorporates material from:

Katie Keith and Stephen Freund, CSCI 104, Williams College

Edwin Vargas Navarro, Data 6, University of California, Berkeley (CC BY-NC-SA)

Weinstein, Sahami, and Reich, *System Error: Where Big Tech Went Wrong and How We Can Reboot*, 2021

