

Data and Society

Introduction – Lecture 1

1/25/21

What to Expect

CSCI 6370 (Grads) / 4370 (Undergrads)

ITWS 6960 (Grads) / 4960 (Undergrads)

- Class time: 12:20 – 2:10 Mondays and Thursdays
- All on-line
- A lot of discussion, everyone participates
- Lecture/Discussion first hour (ish), 2 student presentations second hour (ish)
- Reading, writing, speaking
- Cool famous guest speakers 😊
- Attendance is mandatory and part of your grade
- Please keep your camera on during class and mic muted when you're not speaking

Data and Society

PEW

TOPICS PROJECTS FEATURES ABOUT GET INVOLVED SEARCH

Western States Use Science and Data to Safeguard Migrating Wildlife

New science is helping define better ways to manage big game species

ARTICLE August 6, 2020 By: Matt Skroch & Nic Callero Topics: Land Conservation Projects: U.S. Public Lands and Rivers Conservation Tags: Wilderness, Public lands & Habitat protection Read time: 2 min

Western States Use Science and Data to Safeguard Migrating Wildlife Share Read Mode

THE VERGE

TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

Capital One ordered to pay \$80 million penalty for its role in a 2019 data breach

More than 100 million customers' personal info was exposed in the hack

By Kim Lyman | Aug 8, 2020, 10:12am EDT

f t SHARE

Find a job Sign in Search

on Sport Culture Lifestyle More

The Guardian

US edition

US midterms 2018 Business Tech Science

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Advertisement

Become a Guardian digital subscriber

BROOKINGS

POLICY 2020 CITIES & REGIONS GLOBAL DEV INTL AFFAIRS U.S. ECONOMY U.S. POLITICS & GOV

ID: 110222198001011010	ID: 110222198001011010	ID: 110222198001011010
88.96%	53.25%	48.30%

What are the proper limits on police use of facial recognition?

Nile Bala and Caleb Watney | Thursday, June 20, 2019

TECHTANK

BuzzFeed News

REPORTING TO YOU

ABOUT US GOT A TIP? SUPP

Trending Larry King Biden Vaccine Rollout Texas Deportations QAnon

A Home Security Tech Hacked Into Cameras To Watch People Undressing And Having Sex, Prosecutors Say

Telefonos Aviles admitted he took notes of homes where attractive women lived and hacked into more than 200 accounts over several years.

Salvador Hernandez
BuzzFeed News Reporter

Last updated on January 21, 2021, at 5:54 p.m. ET
Posted on January 21, 2021, at 4:30 p.m. ET

Tweet Share Copy



BUSINESS INSIDER

TECH | FINANCE | POLITICS | STRATEGY | LIFE | ALL

PRIME | INTELLIGENCE

A couple says that Amazon's Alexa recorded a private conversation and randomly sent it to a friend

Fran Berman, Data and Society, CSCI 4370/6370

Who am I?

- Professor: Dr. Fran Berman
- Office Hours: By appointment (send email to bermaf@rpi.edu). Let's talk!
- Course website (linked off Fran's RPI web page): <https://www.cs.rpi.edu/~bermaf/Data2021.html>
- My research interests (FYI): Public Interest Technology, data policy and cyberinfrastructure, Internet of Things, social impacts of technology

Today (1/25/21)

- Course logistics
 - Syllabus
 - Expectations and grading
 - Learning objectives, grads and undergrads
- Lecture 1
- Reading for Thursday 1/28
- Introductions

Class is a safe space for you and your opinions

Zoom etiquette for all of us:

- Keep your camera on during class. No recording.
- Keep your mic muted when you're not speaking
- Everything we can see you wearing should be what you would wear to an on-site class
- Backgrounds and behavior should not be distracting ...
- Be respectful of everyone
- Use chat only for class-related stuff and only in "everyone" or host mode
- Your Zoom presence should be something your Parents/Grandparents/Advisor/Kids/etc. would be proud of ...

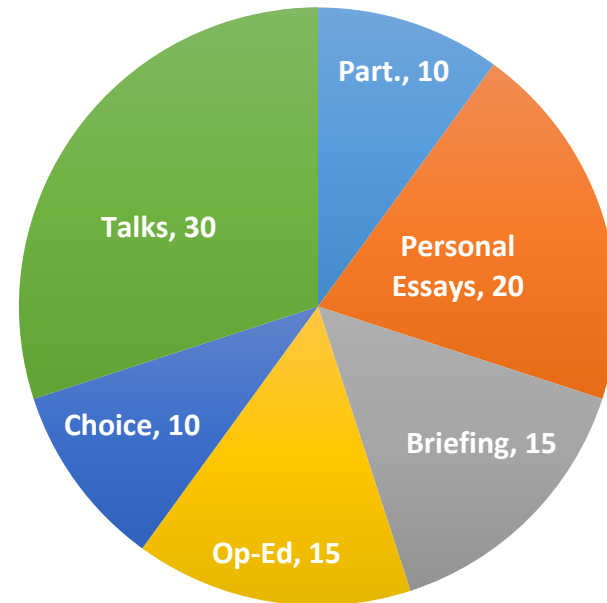
Date	Topic	Speaker	Date	Topic	Speaker
1-25	Introduction	Fran	1-28	The Data-driven World	Fran
2-1	Data and COVID-19	Fran	2-4	Data and Privacy -- Intro	Fran
2-8	Data and Privacy – Differential Privacy	Fran	2-11	Data and Privacy – Anonymity	Fran
2-15	NO CLASS / PRESIDENT’S DAY		2-18	Data and Privacy – Law	Ben Wizner
2-22	Digital rights in the EU and China	Fran	2-25	Data and Discrimination 1	Fran
3-1	Data and Discrimination 2	Fran	3-4	Data and Elections 1	Fran
3-8	Data and Elections 2	Fran	3-11	NO CLASS / WRITING DAY	
3-15	Data and Astronomy	Alyssa Goodman	3-18	Data Science	Fran
3-22	Digital Humanities	Brett Bobley	3-25	Data Stewardship and Preservation	Fran
3-29	Data and the IoT	Fran	4-1	Data and Smart Farms	Rich Wolski
4-5	Data and Self-Driving Cars	Fran	4-8	Data and Ethics 1	Fran
4-12	Data and Ethics 2	Fran	4-15	Cybersecurity	Fran
4-19	Data and Dating	Fran	4-22	Data and Social Media	Fran
4-26	Tech in the News	Fran	4-29	Wrap-up / Discussion	Fran
5-3	NO CLASS				

Grading

- **Class participation (1):** 10 points (includes attendance)
- **Personal Essays (2)** / storytelling: $2 \times 10 = 20$
 - *January, February/March*
- **Briefing (1)** / collaboration, informative writing: 15
 - *February*
- **Op-ed (1)** / persuasive writing: 15
 - *March*
- **Choice: Personal Essay or Op-ed (1)** / your choice: 10 or 15 (if op-ed, you may get up to 5 points extra credit)
 - *April*
- **Talks (2)** / oral communication: $2 \times 15 = 30$
 - *You select*

You are responsible for managing your assignments and participating in class discussion.
No late work.

Grade Distribution



Attendance and writing assignments

- **Attendance**

- I will take attendance at the **beginning of each class (12:20)**
- If you decide to drop the class, please let me know (bermaf@rpi.edu) and I will let in someone on the waiting list.

- **Writing assignments and presentations.**

- **Send them to bermaf@rpi.edu before the date/time they are due.**
- **Send writing assignments in MS .docx.** I will give you comments in track changes.
- **Send presentations in .pdf.** I will use these for reference when I grade your presentations.
- None of these will go on the web unless otherwise specified.

Course Materials

- Course website (<https://www.cs.rpi.edu/~bermaf/Data2021.html>) will have all up-to-date information and materials
 - Trajectory of lectures may evolve slightly – web page and previous lecture will always be the “state-of-the-art”
 - Fran’s lectures will be on the web after they are given
 - All readings will be on the web
- Presentation articles will be on the web for the day they are to be presented
- No student-authored materials or presentations will go on the web unless otherwise indicated

Learning Objectives and Outcomes, Grads/Ugrads

Learning Objective	Outcome
Develop greater data literacy	Be able to understand and explain the role that data plays as well as its limitations in various areas of research, commerce and modern life.
Develop critical thinking skills around data	Be able to read, understand, assess, and discuss data-oriented professional and popular publications and articles.
Develop communication skills around data	Be able to advance an evidence-based argument about data, data cyberinfrastructure and data-oriented efforts to both knowledgeable specialists within the field as well as non-specialists.

- **There will be slightly different expectations for grad students and undergrads:** Students will be assessed at a level appropriate to *their* educational level (undergrad or grad)

Academic Integrity

- Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these.
- In this class, **all assignments that are turned in for a grade must represent the student's or group's own work.** In cases where help outside project expectations was received, a notation on the assignment should indicate your collaboration. **If references or other materials are used, they should be cited.** **Submission of any assignment that is in violation of this policy will result in a penalty.**
- If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: warning, probation, suspension, expulsion, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities. **If you have any question concerning this policy before submitting an assignment, please ask for clarification.**

Class Participation

- **Class participation (10 points based on attendance and engagement):**
 - Students are expected to attend **at least 23 out of 26** class meetings for full Attendance credit.
 - **Attendance taken at the beginning of class**
 - **Students are expected stay until the end of class**
 - You are expected to participate in class discussions and be an attentive audience (and question asker) for speakers

You get out of the course what you put into it

- **Spend time on the readings, writings, and presentations.** Don't do this at the last minute.
 - Do more than one draft of all writings and presentations
 - Practice your presentations before you give them
 - Have someone “red team” the assignments for feedback
 - Talk to Fran during office hours (by appt.)
 - Read all the readings and be prepared to discuss / answer questions
- Be prepared to **engage in class.**
- **You make the effort**, Fran = scorekeeper. Focus on building skills as an outcome of this class.

Skills you can improve if you spend time on the assignments and participating in class:

- Persuasive writing
- Storytelling
- Informative writing
- Presentation style
- Communication skills
- Collaboration
- Thinking on your feet (questions, discussion)
- Critical thinking / evidence-based reasoning

Lecture 1

Questions for today:

- What is data?
- Where is data?
- Why does data matter?
- What happens when data is wrong?
- How do we know when data is correct?
- How accurate is the data on you?

What is data?

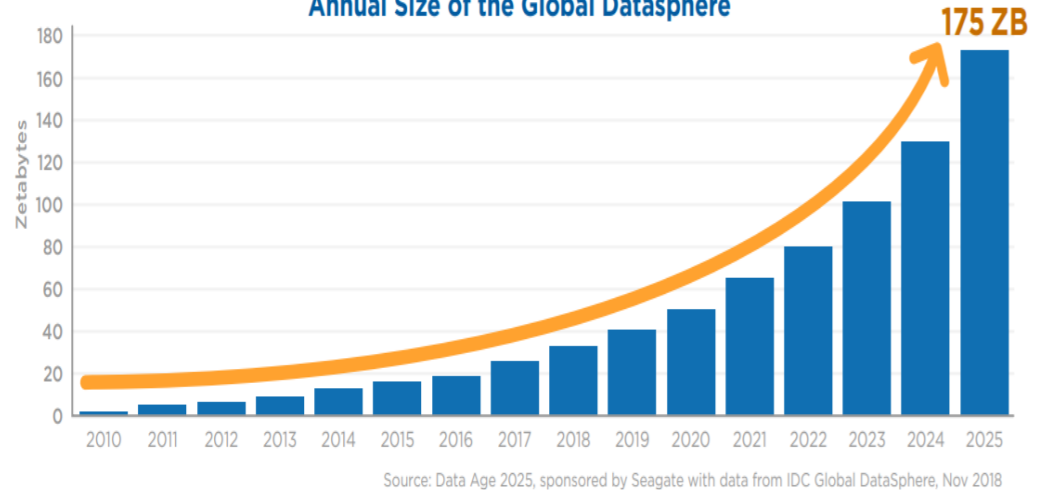
Data is information that has been translated into a form that is efficient for movement or processing.

Relative to today's computers and transmission media, data is information converted into binary digital form – 0's and 1's.

- Google hosts more than 3.5 billion searches a day
- Internet users generate 2.5 petabytes of data each day
- In 2019, internet users spent 1.2B years online
- Social media accounts for 33% of total time spent on-line (2016)
- 90% of all data has been created in the last two years (2017)
- Job listings for data science and analytics expected to be 2.7M in 2020.

<https://techjury.net/blog/big-data-statistics/#gref>

Annual Size of the Global Datasphere



<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}
Zetta	10^{21}
Yotta	10^{24}

Metadata: the “Digital Shadow”

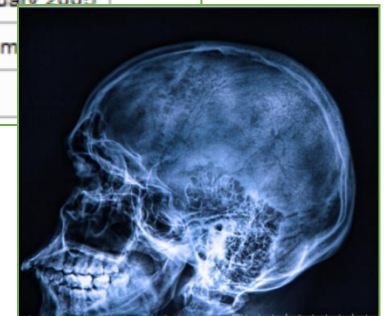
- Less than half of your digital footprint is related to individual actions – taking pictures, making VoIP calls, uploading files, etc.
- The rest of your digital footprint is “ambient” content and metadata related *to you*: surveillance images, banking records, medical records, information about your web searches and behavior in social networks, etc.



Metadata

This file contains additional information, probably added from the digital camera or scanner used to create or digitize it. If the file has been modified from its original state, some details may not fully reflect the modified image.

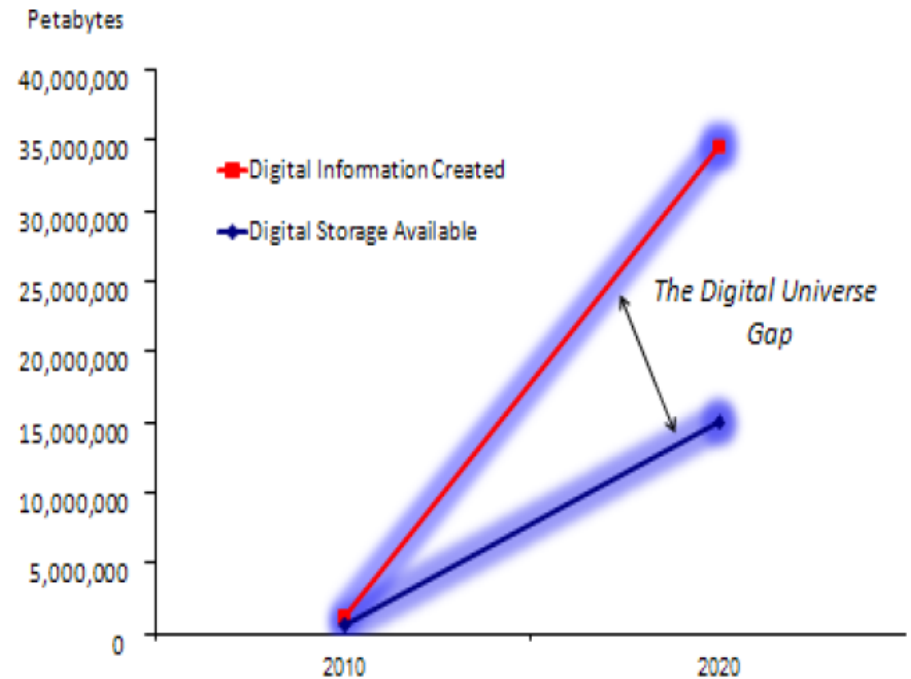
Camera manufacturer	Canon
Camera model	Canon EOS 30D
Author	unknown
Exposure time	1/160 sec (0.00625)
F Number	f/5.6
Date and time of data generation	20:21, 20 January 2005
Lens focal length	400 mm
Show extended details	



All data cannot be stored

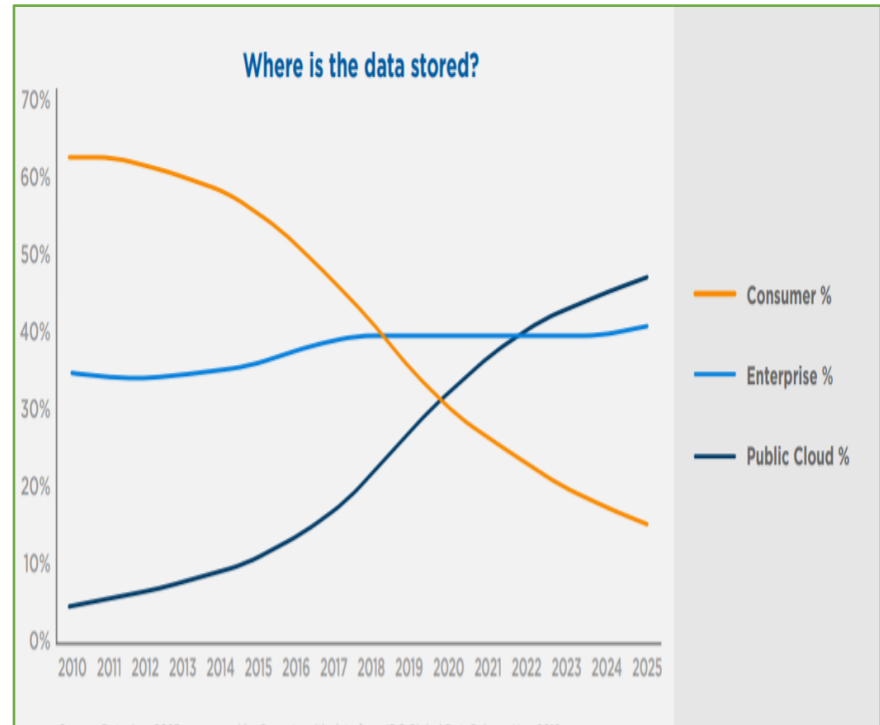
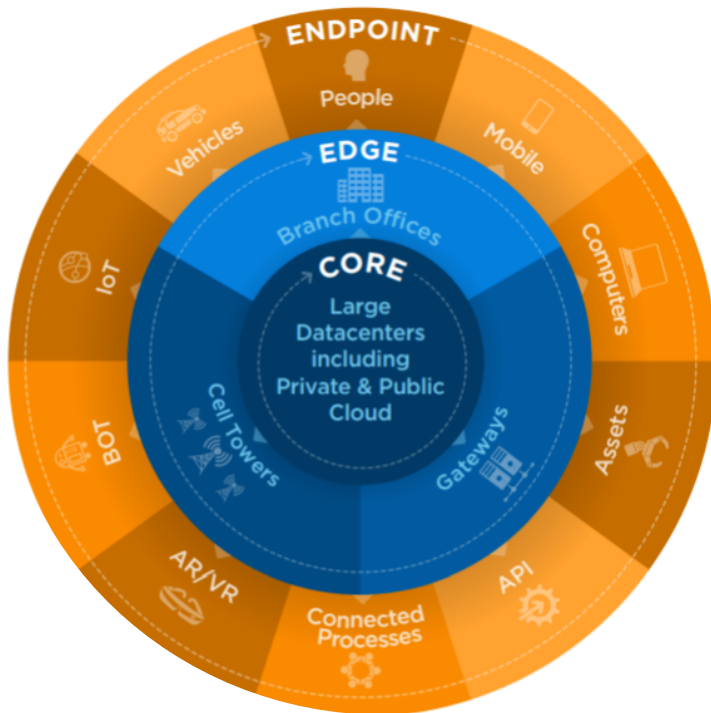
- 2007 was the “crossover year”: Began to generate more digital data than storage to keep it
- In 2020, more than twice as much information will be created as storage available

Figure 5: The Emerging Gap
Information Creation > Storage Available



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Where is the data?



<https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

Why does data matter to organizations?

- [**Competitive advantage**] “A company’s ability to compete is now measured by how successfully it applies analytics to vast, unstructured data sets across disparate sources to drive product innovation. ... smart data scientists can make or break a product.” Sequoia Capital
- [**Facilitate better outcomes**] “Data is valuable because it enables better, more connected services, improved policies and decision making, and the development of new, innovative products.” Australian Government
- [**Deeper understanding**] “Data inspires progress and galvanizes change. To know where we need to go, we need to know what we’ve achieved – where progress is being made and where major challenges remain.” Maura Pally (Clinton Foundation)

When data is “wrong”

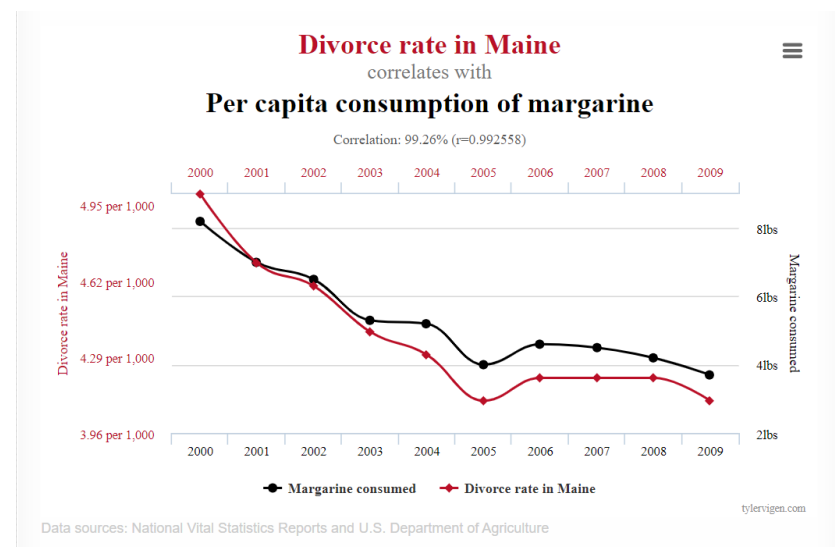
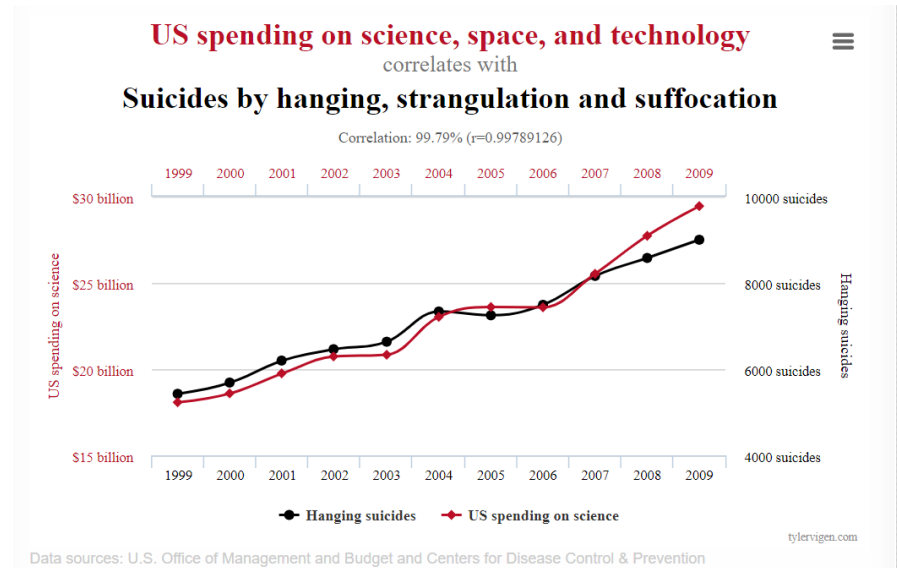
- How data can be “wrong”:

- Not representative
- Incorrect / low integrity
- Not interpreted correctly
- Misleading
- Not consistent
- Fabricated, etc.

- Data is just “the math” – it’s the **human’s responsibility** to make sure that the data, model, algorithms, interpretation is appropriate and correct for the situation.

- Beware of overfitting
- Beware problems with **correlation vs. causation!**

<https://www.tylervigen.com/spurious-correlations>



Data about Fran




Fran's Twitter Data

Date	Gender	Age	Languages
10/19	Female	21-54, >65	English, Portuguese
8/20	Female	35-54	English
1/21	Female	13-54	English, French

https://biographynetworth.com/francine-berman-net-worth/

https://vpn.net.rpi.edu... Junior candidates - Le... Warning

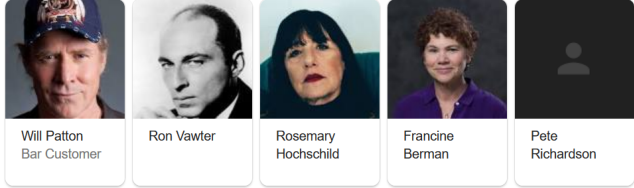


Francine Berman Net Worth is **\$12 Million**

Google King Blank Cast

All Shopping Videos News Images More Settings Tools

King Blank / Cast



King Blank (1983) - IMDb
https://www.imdb.com/title/tt0231908/
★ ★ ★ ★ Rating: 4.6/10 - 10 votes
With Rosemary Hochschild, Ron Vawter, Will Patton, Francine Berman. Set in a motel room at ...
King Blank Poster Set in a motel Cast overview: Rosemary

Promoting “correct” data (and outcomes)

To get data you can trust...

Improving quality is not easy, and nobody sets out to waste time or sabotage their efforts by not collecting good data. But as these reminders show, it's all too easy to get problem data even when you're being careful!

When you collect data, remember to:



1. Create a data collection plan.



2. Assess your measurement system.



3. Make sure your collection methods do not create bias or other problems.



4. Make sure your team understands how gathering good data benefits them.



5. Perform a preliminary review of the data before in-depth analysis begins.

The screenshot shows the Scientific American website interface. At the top, there are navigation links for 'Subscribe', 'Latest Issues', and 'Cart'. The main navigation bar includes 'CORONAVIRUS', 'THE SCIENCES', 'MIND', 'HEALTH', 'TECH', 'SUSTAINABILITY', 'VIDEO', 'PODCASTS', 'OPINION', and 'PUBLICATIONS'. Below this is a sign-up prompt for newsletters. The main content area features the 'E&E NEWS EARTH' logo and the article title '2020 on Track to Rank in the Top 5 Hottest Years on Record'. A sub-headline reads 'The first three months of the year were the second warmest in 141 years of record keeping'. The author is listed as 'By Thomas Frank, E&E News on April 18, 2020'. To the left of the article is a social media sharing icon. To the right is a 'READ THIS NEXT' section with sponsored content and other articles.

Verify predicted outcomes against observed outcomes

Next time

- Lecture 2: the Data-driven world
- Fran will give a model presentation and instructions for you; first sign-up for presentation assignments
- First personal essay assignment will be given (including instructions)

Read this before the next class -- Thursday, January 28

- “What happens when you click ‘agree’?”, The New York Times (link on the class website)
- <https://www.nytimes.com/2021/01/23/opinion/sunday/online-terms-of-service.html?referringSource=articleShare>



Class introductions (take 1-2 min)

1. Name, major, school, location / time zone
2. What is the most interesting thing you've recently heard / read about data?
3. Best thing you did last fall (that you're willing to share)