# Commercial Privacy

# Privacy: It's Not a New Issue

- In 1890, Samuel Warren and Louis Brandeis published "The Right to Privacy" in the Harvard Law Review
- Recognizing that computers changed things, the Committee on Science and Law of the New York City Bar Association started its formal privacy study in 1962
  - This led to Alan Westin's 1967 book "Privacy and Freedom", a report on the committee's work
- The US Congress held hearings on technology and privacy throughout the 1960s
- Legal academics wrote extensively on the topic
- (Jewish writings circa 200 CE mention a right to physical privacy, and derive it from a Biblical text)

# Notice and Consent

- **Westin**: “A central aspect of privacy is that individuals and organizations can determine for themselves which matters they want to keep private and which they are willing—or need—to reveal.”
- This has been the basis for virtually *all* privacy regulation since then

# Privacy Regulation

**1973:** A US government committee came up with the “Fair Information Practice Principles”

**1974:** The US government passes the *Privacy Act of 1974*, implementing them—but only for the Federal government

**1980:** The OECD guidelines suggested more or less the same thing

**1994:** The *Data Protection Directive* is enacted

**2012:** The GDPR is adopted

From 10,000 meters, all of these are more or less the same: notice and consent

CS@CU

# Notice and Consent

- Sites tell you what they'll collect, and what they'll do with it
- By using the site, you are deemed to have consented to this policy
- The big differences in the GDPR: it's mandatory, you have to be more precise about uses, and subjects have a right to see their information

# Where Are We?

- We have not solved the technical problems identified more than 50 years ago
- But we still have notice and consent
- Does it work?
- Nope…

# Problems with Notice and Consent

- Amount of data collected, and by whom
- Privacy policies
- Location data is collected, often without folks' knowledge
- Many governments

# Overcollection

- Data brokers—outside parties with whom consumers have no association, and to whom they have never consented—collect, buy, and sell a tremendous amount of data
- Websites track users
- Ads are from outside brokers, who use HTTP redirection to gather even more data
- Also: third-party “like” buttons (e.g., Facebook and Twitter) and third-party authentication (e.g., Facebook and Google)

# Analytics Platforms

“To those first-party profiles, Rubicon typically adds details from third-party data aggregators, like BlueKai or eXelate, such as users’ sex and age, interests, estimated income range and past purchases. Finally, Rubicon applies its own analytics to estimate the fair market value of site visitors and the ad spaces they are available to see.”

*(New York Times)*

# Privacy Policies

- No one reads them
  - Cranor estimated the opportunity cost at US$3500/year to read them all
- They're deliberately vague and expansive
  - "We may collect personal information and other information about you from business partners, contractors and other third parties." (Reidenberg et al)
- "Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent." (PCAST report)

# Governments

- If data exists, it’s available to governments
- Some governments have a complex, restricted, and somewhat painful process required to gain access to data
- Other governments don’t care very much about such niceties
- Some governments collect data via espionage, technical and otherwise

# It’s Not Just PII

- Virtually all privacy laws are based on protecting PII: Personally Identifiable Information
  - N.B.: The definition of PII varies
- But: you don’t need PII to invade folks’ privacy
  - Amazon doesn’t need your identity to recommend products
  - Netflix doesn’t need your identity to recommend movies
- PII is actually a database key!

# Deidentification

- Many people have tried “deidentifying” data: removing the PII
- Reidentification is often possible, based on other fields
- Good anonymization can destroy the utility of the data, e.g., for some medical research

# Machine Learning

- Today's ML algorithms can infer things not directly observed, e.g., sexual orientation
- This is much harder to control: it is *not* based on data collection, which is usually what's regulated
- Even when some inputs are disallowed by law, there are often proxy variables that are strong correlates

# The Trouble with Notice and Consent

- No one knows who collects data
- No one knows what they'll do with it
- No one knows where it's stored
- And some of the most sensitive stuff, e.g., location, is dual-use: it's used for your benefit (map programs) *and* it's part of your "data shadow"

But what should replace it?

# The Web

# Dawn of the Web

- In 1990, Tim Berners-Lee invented the web as a way to distribute documentation
  - Crucial notion: *hypertext,* a way for documents to contain links to other documents
  - (Hypertext in quasi-modern form dates to the 1960s)
- Others wanted to enable e-commerce
- The original design couldn't quite accommodate this in a clean fashion

# The Web: Design

- Two primary components, *HTML* and *HTTP*
  - **HTML:** HyperText Markup Language; describes how a page should be formatted
  - **HTTP:** HyperText Transport Protocol; used to transmit web pages over the Internet
- Stateless design
  - Open a connection, download a page, close the connection
  - No link between downloads—and hence no way to have a *session*

# Sample HTTP

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.16; rv:85.0) Gecko/20100101 Firefox/85.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Upgrade-Insecure-Requests: 1
```

# Sample HTML

```
<html>
<title> weblog </title>
<body> <h1> I heard you say </h1> <font size+=4>
<pre>GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.16; rv:85.0) Gecko/20100101 Firefox/85.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Cookie: WhoYouAre=1804289383; ID-Age=1612720694; Last-Seen=1612721125; Flavor="Chocolate-chip"; Size="Large"
Upgrade-Insecure-Requests: 1
</font>
</pre>

</body>
</html>
```

# Session Needs

- Ability to log in
- *No requirement for a login name*
- Persistent preferences, e.g., language
- Shopping cart
- The accepted answer: *cookies*

# Cookies

- Cookies are arbitrary text strings sent to a browser by a web site
- They’re retained by the browser in non-volatile storage and returned when the site is next visited
- *Cookies can be persistent identifiers*
- They can hold anything else a site wants, too

# Cookies

**I heard you say**

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.16; rv:85.0) Gecko/20100101 Firefox/85.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Upgrade-Insecure-Requests: 1
```

**from 24.194.9.206:47607**

I just sent you, #1804289383, a cookie; reload this page to see it coming back to me.

# Cookies Reloaded

**I heard you say**

```
GET / HTTP/1.1
Host: greylock.cs.columbia.edu
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.16; rv:85.0) Gecko/20100101 Firefox/85.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
DNT: 1
Connection: keep-alive
Referer: http://greylock.cs.columbia.edu/
Cookie: Size="Large"; WhoYouAre=1804289383; ID-Age=1612724999; Last-Seen=1612725000
Upgrade-Insecure-Requests: 1
```

**from 24.194.9.206:37450**

I just sent you, #1804289383, a cookie; reload this page to see it coming back to me.
ID Age: Sun Feb 7 14:09:59 2021
Last visit: Sun Feb 7 14:10:00 2021

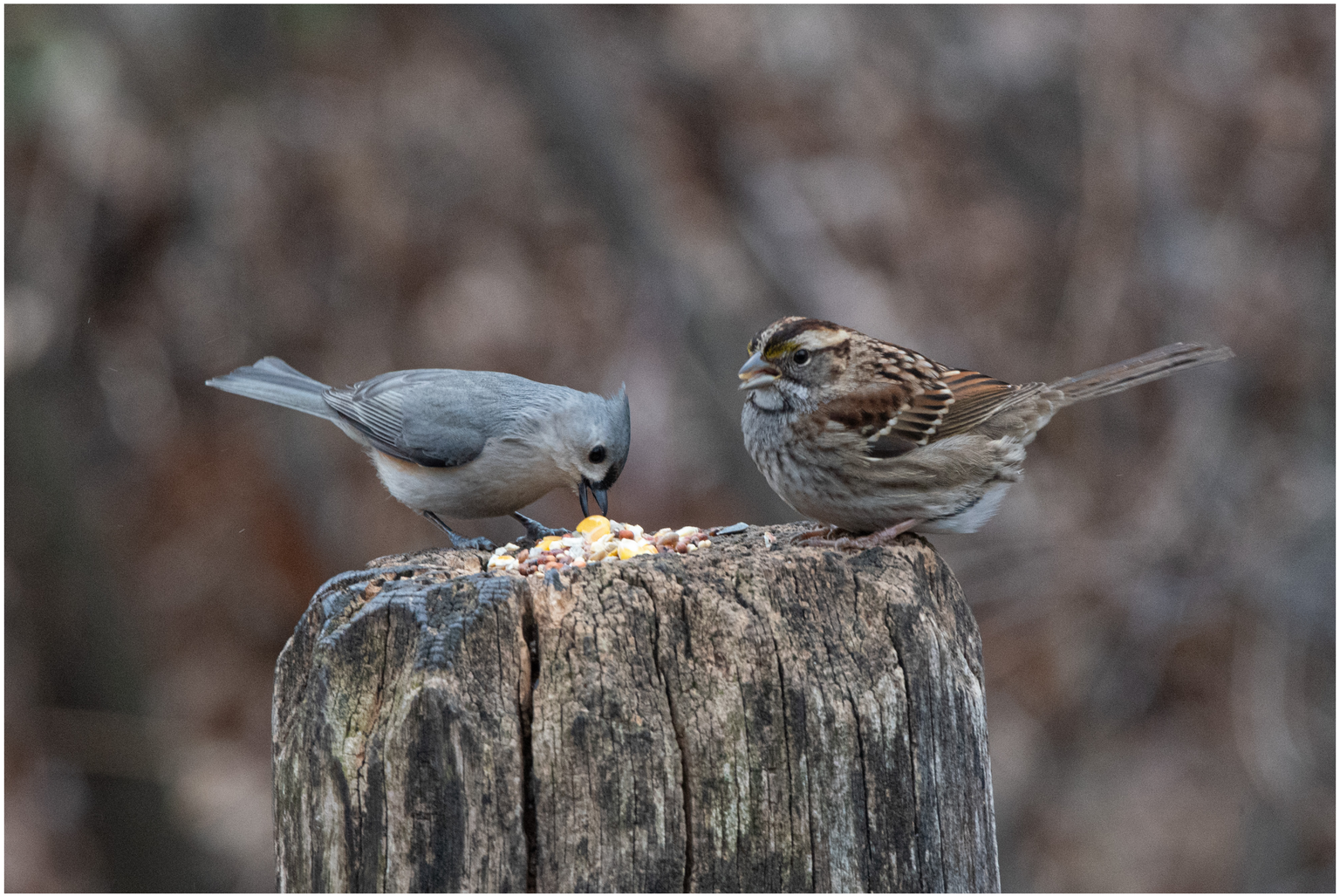
# Third-Party Cookies

- Images and IFRAMES—embedded web pages—are loaded via separate URLs
- These URLs can point to a different site—and each such site can send and receive its own cookies
- HTTP requests for embedded content contain "`Referer`" lines that identify the parent page
- *Most ads are in IFRAMEs pointing to third-party ad brokers*
- *Consequence: third parties can track you around the web*

# Internet Advertising

- Dominated by Google and Facebook
- They observe the content of the embedding pages to learn your interests
  - Other features—Facebook's embedded "like" buttons, Google Analytics, either's single sign-on—also contain embedded content and track you around the web
- Cookies are not PII per se—but many web sites know real names, etc.
  - Besides, gmail addresses and Facebook logins are PII

# Daily Bird



(Tufted titmouse and white-throated sparrow, Central Park, January 31, 2021)

CS@CU

# 1960s Warnings

# There Are Problems

**Westin**: “It should be recognized that consent to reveal information to a particular person or agency, *for a particular purpose*, is not consent for that information to be circulate to all or used for other purposes.”

**Michael**: “The would-be invader who knows about these centralized or clustered inventories *need not search for sources*, and therefore he may be much more inclined to examine the records than if a major search for the sources of information were necessary.”

**Michael**: “We can expect a great deal of information about the social, personal, and economic characteristics of individuals to be *supplied voluntarily*—often eagerly—in order that, wherever they are, they may have access to the benefits of the economy and the government.”

# Security Considerations

- **Miller:** “Another important security function that a privacy-oriented monitor program must perform is the identification of all users and terminals attempting to gain access to the files”
- He went on to suggest call-backs, token identification, and biometrics

# Passwords

**Dr. PIORE.** The user then keys in his six-character password.

**Senator LONG.** Doctor

**Dr. PIORE.** Yes?

**Senator LONG.** But he could give that password to someone else, could he not?

**Dr. PIORE.** He can, and you find that some people do not protect their own password, and this is a human characteristic. You have a friend, and he says, “Let me use your program.” It is like saying “Come up to my apartment, and here is the key to my apartment.”

*(US Senate Committee hearing)*

# Encryption

- **Miller:** “In the case of remote-access systems, protection against wiretapping can be achieved by using ‘scramblers’ to garble the data before transmission, and installing complementary devices in the authorized terminals to reconstitute the signal.”
- “Coding has a number of tangential advantages from the privacy perspective, including verifying the source of an inquiry or input”

# Hackers and Insiders!

- **Miller:** “Even the most sophisticated set of safeguards can be undermined by the people who gain access to the system in one fashion or another. The reports of college students at MIT and elsewhere defeating the monitor protections in time-sharing projects emphasize the reality of this threat.”
- “There is a danger that these people will become so entranced with operating sophisticated machine systems and manipulating large masses of data that they will not be sufficiently sensitive to the question of privacy.”

# Metadata

**Miller:** “One of the simplest of the present generation of snooping devices is the pen register, which, when attached to a telephone line, records on paper a series of dashes representing all numbers dialed from the selected telephone. But this snooping capability would be increased by several orders of magnitude if a few pen registers were attached to suspects' telephone lines and the information drawn in by these devices fed into a central computer. This technique could quickly provide a revealing analysis of patterns of acquaintances and dealings among a substantial group of people.”

“Optical scanners designed to decipher license numerals and send them directly to the computer obviously would make the process more efficient—and, as a by-product, might enable the compilation of comprehensive records of the movements of a person’s automobile, perhaps for later inferential relational analysis.”

# Jewish Sources

# Numbers 24:2, 5

וַיִּשָּׂא בִלְעָם אֶת-עֵינָיו , וַיַּרְא אֶת-יִשְׂרָאֵל, שֹׁכֵן, לִשְׁבָטָיו; וַתְּהִי עָלָיו, רוּחַ אֱלֹהִים.

…

מַה-טֹּבוּ אֹהָלֶיךָ, יַעֲקֹב; מִשְׁכְּנֹתֶיךָ, יִשְׂרָאֵל.

As Balaam looked up and saw Israel encamped tribe by tribe, the spirit of God came upon him…  How fair are your tents, O Jacob, your dwellings, O Israel!

# Mishnah Bava Batra 3

לֹא יִפְתַּח אָדָם לַחֲצַר הַשֻּׁתָּפִין פֶּתַח כְּנֶגֶד פֶּתַח וְחַלּוֹן כְּנֶגֶד חַלּוֹן

In a jointly held courtyard a man may not build a door directly opposite another’s door, or a window directly opposite another’s window.

# Talmud Bava Batra 60a

אמר רבי יוחנן דאמר קרא וישא בלעם את עיניו וירא את ישראל שוכן לשבטיו מה ראה שאין פתחי אהליהם מכוונין זה לזה

Rabbi Yoḥanan says that verse (Numbers 24:2) states: “And Balaam lifted up his eyes, and he saw Israel dwelling tribe by tribe”; What did he see? He saw that the entrances of their tents were not aligned with each other.