11-20-2016

Patron Activity Monitoring and Privacy Protection

Sam Kome
The Claremont Colleges

lita

Patron Activity Monitoring and Privacy Protection

Sam Kome @skome
Claremont Colleges Library

#litaforum

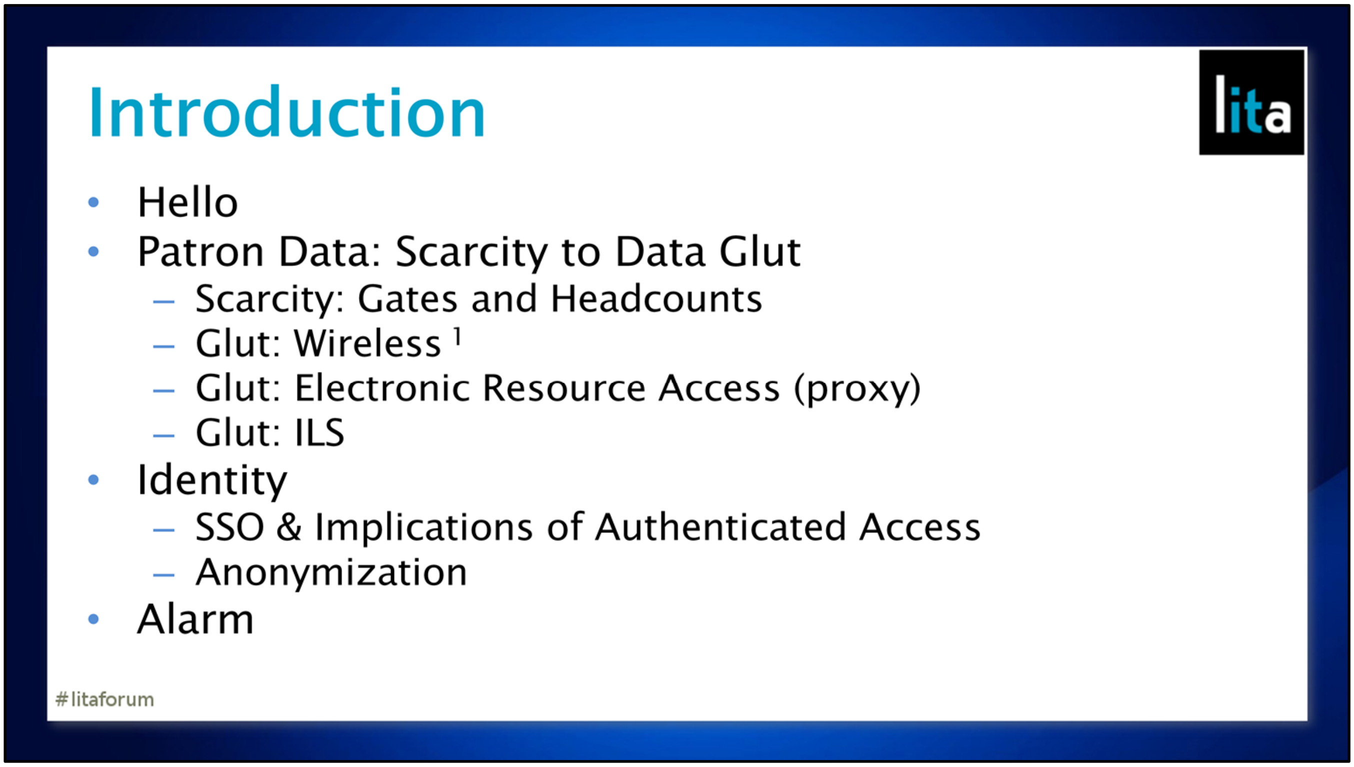
PUBLIC DOMAIN

Most if not all libraries collect and report building use, along with a the myriad other statistics.

Most if not all libraries collect and report Electronic resource usage data; we have sketchy COUNTER reports.

These data lack needed specificity and can’t be correlated to each other, e.g.; what resources do the early birds who sit in the north commons use? What if we knew to the minute and foot how many patrons occupied the building and where they sat? What if we also knew the electronic resources those same patrons used whether or not they were in the building? And what if we could perfectly protect their privacy? We can if combine and analyze Wireless, Proxy, and ILS data.

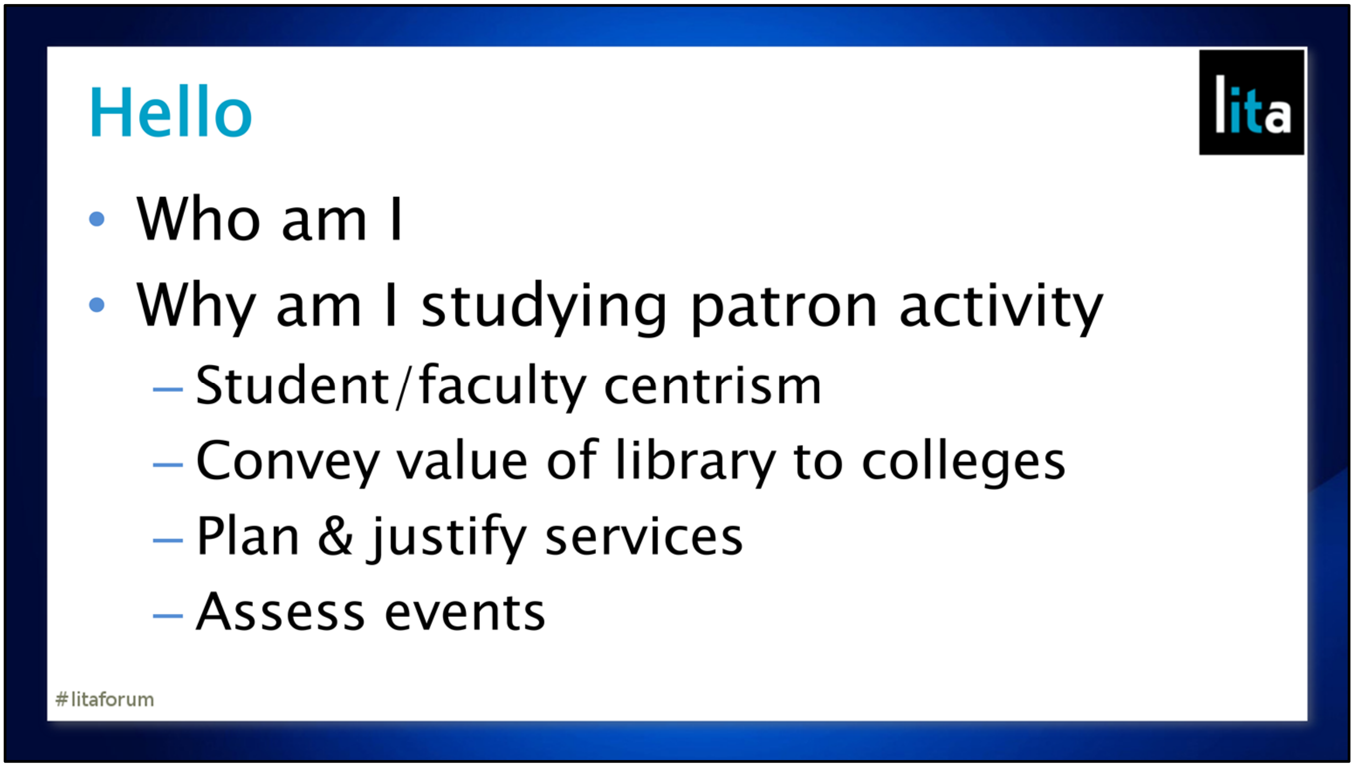
This presentation will be most relevant for libraries with: Centralized wireless (if not, this will provide some justification) Centralized authentication, e.g. CAS, LDAP, Shibboleth EZProxy or other web proxy to electronic resources Patron type (Faculty, Undergraduate, etc) information either in the ILS or in the CAS/LDAP/Shibboleth



Introduction

- Hello
- Patron Data: Scarcity to Data Glut
  - Scarcity: Gates and Headcounts
  - Glut: Wireless [1]
  - Glut: Electronic Resource Access (proxy)
  - Glut: ILS
- Identity
  - SSO & Implications of Authenticated Access
  - Anonymization
- Alarm

#litaforum



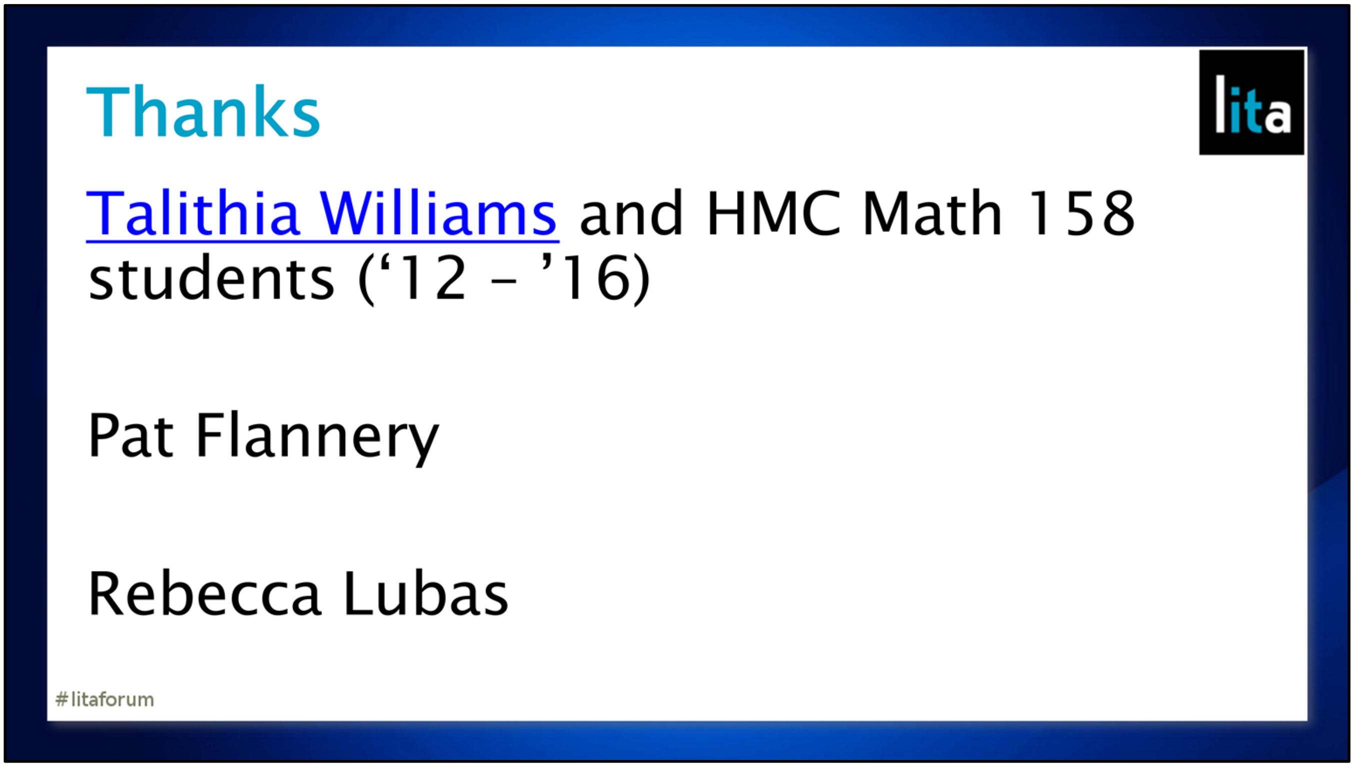
Hello

- Who am I
- Why am I studying patron activity
  - Student/faculty centrism
  - Convey value of library to colleges
  - Plan & justify services
  - Assess events

#litaforum

I have three roles. I work directly with patrons. I am the library IT liaison, and I sheepdog the strategic plan initiatives.
Our strategic plan is based on a student and faculty centric vision – so I want to know them in depth.
Our funding depends on our ability to explain its uses
The uses for funds depend on some objective data
And we reflect on those uses with data and other means.

So: I'm the lucky librarian who got to figure it out. The approach I took requires
Centralized wireless
Centralized authentication, e.g. CAS, LDAP, Shibboleth
EZProxy or other web proxy to electronic resources
Patron type (Faculty, Undergraduate, etc) information either in the ILS or in the CAS/LDAP/Shibboleth



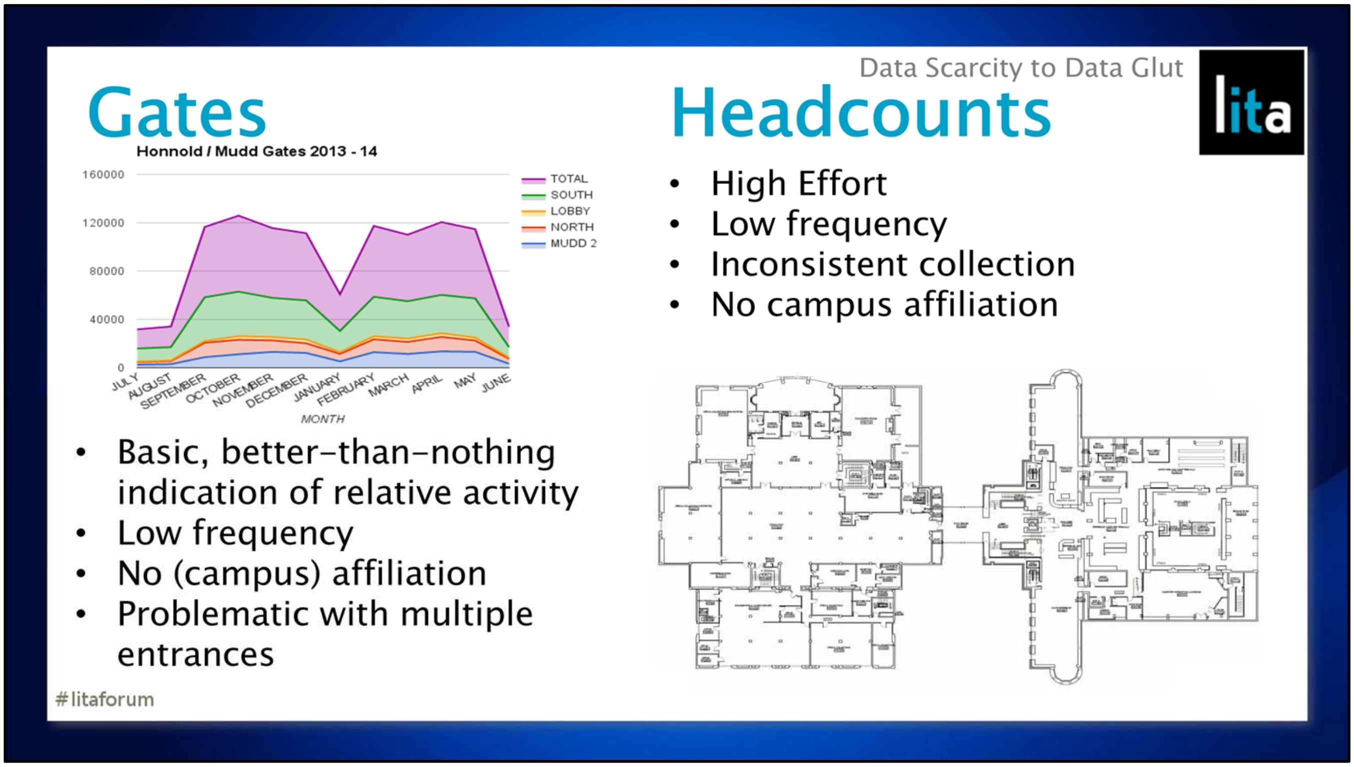
Thanks

Talithia Williams and HMC Math 158 students (‘12 – ’16)

Pat Flannery

Rebecca Lubas

#litaforum



Data Scarcity to Data Glut

Gates

Honnold / Mudd Gates 2013 - 14
160000
120000
80000
40000
0
TOTAL
SOUTH
LOBBY
NORTH
MUDD 2
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
MONTH

- Basic, better-than-nothing indication of relative activity
- Low frequency
- No (campus) affiliation
- Problematic with multiple entrances

Headcounts

- High Effort
- Low frequency
- Inconsistent collection
- No campus affiliation

#litaforum

I used to have a tough time getting data on building and electronic resource uses.
Electronic or mechanical gates that count ingress and egress provide a basic fact; a human (probably) moved past.

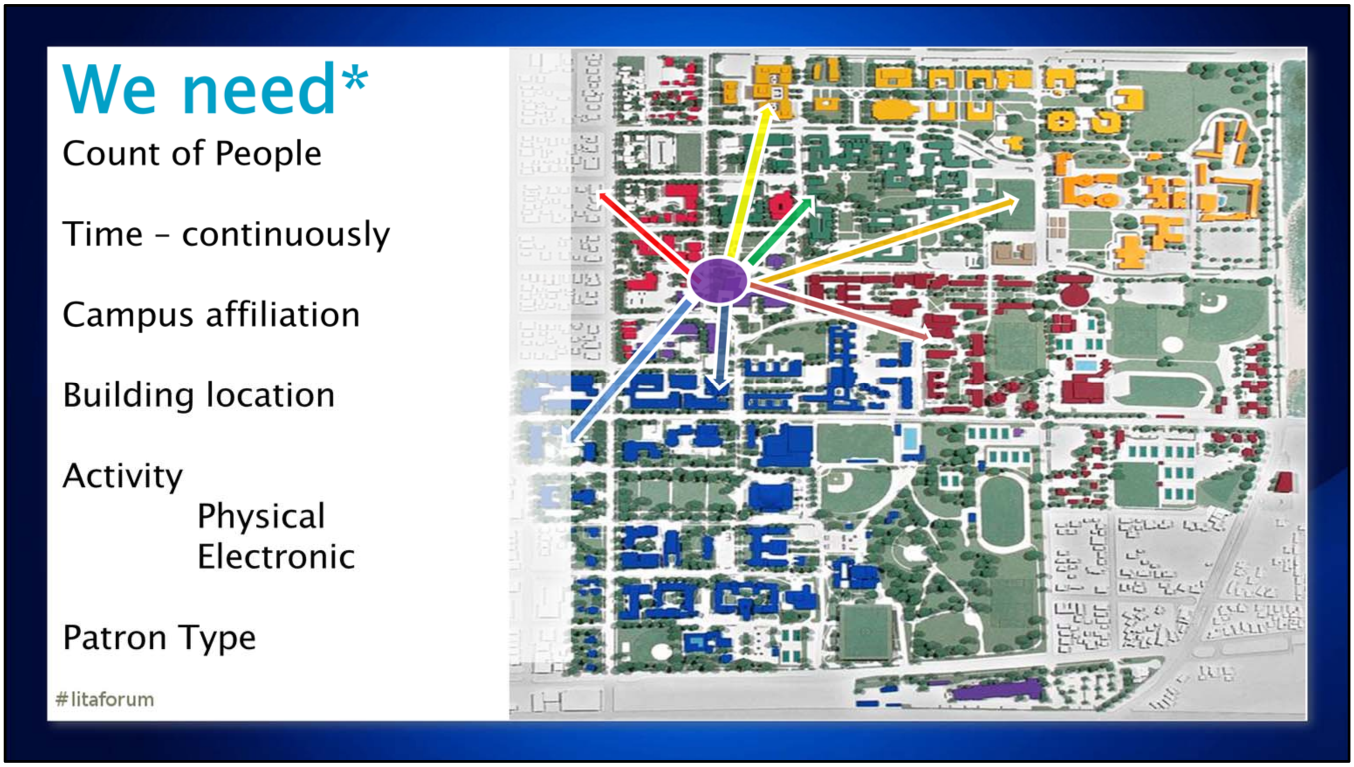
Many of these don’t provide timestamps so the data are recorded at some interval; in our case daily.
In the case of multiple entrances, these data are especially problematic. Any given patron may enter one door and exit another.
For example, we have three entrances each equipped with 3M gates. Each gate counts ingress and egress on one mechanical counter.
To get a daily ingress count we divide that number by 2, but we KNOW that patrons will enter the North entrance and exit via the café.
So we take these numbers with a big grain of salt.

Headcounts require one or more staff members to walk through our entire building with a collection instrument – we use paper seating charts, and would like to try the Suma tablet application from NCSU. Regardless it takes at least an hour to cover the buildings so this is a fairly expensive task that can only be performed a few times per day, at best.

The staff can’t always conform to an exact schedule, nor does everyone collect the data in exactly the same way. How for one example, should I count an empty seat at a table that has a bag and laptop on it? Where is that patron? Bathroom? Services Desk? Copiers? Know what I mean?

We don’t know if we’re counting campus-affiliated folks, and we really need to be able to know that and more – such as proportional campus representation.

So these data are not hugely reliable. And they don’t contain everything we need.



We need*

Count of People

Time – continuously

Campus affiliation

Building location

Activity
- Physical
- Electronic

Patron Type

#litaforum

So if we could design a perfect building census report, what would it include?
We do want an accurate total daily count, and actually let's have that **by hour** so we can be a bit smarter with desk staffing. Then we could know when it's best to allow noisy facilities maintenance projects.
Actually we often need to know the building population at arbitrary times, don't we? Perhaps we're planning study breaks, or want to pre and post assess an event. So let's count **every minute**.

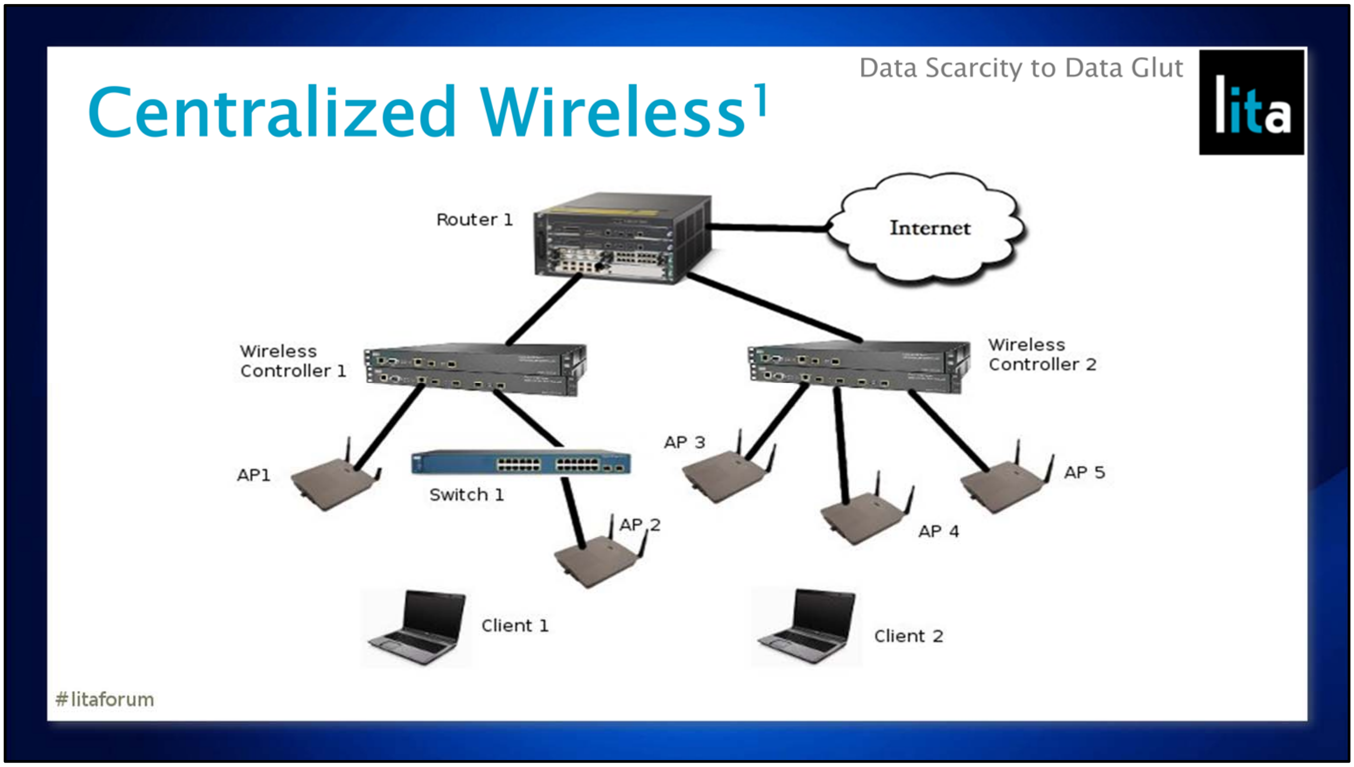
Sure would be great to have those data by building locations, like Honnold – First floor – East. Our library consists of three contiguous buildings 4 stories tall. At any given moment how many patrons are in the café? How many on the mezzanine, in the Green Room, the Special Collections Reading Room, the main conference room, etc.

Or do we need to get down to the table? Do we need that level of granularity?

Finally we in Claremont and perhaps at your institution, need to know Campus affiliation. How many Pitzer students use the quiet floor? Versus HMC? How long do they spend there, on average?

Oh hey, we'd like to know something about computing devices patrons are carrying. Smart phones? Tablets? Laptops?

But we can't walk around harassing patrons non-stop, and we don't even have the staff. We want all this and we want it to happen automatically and invisibly.



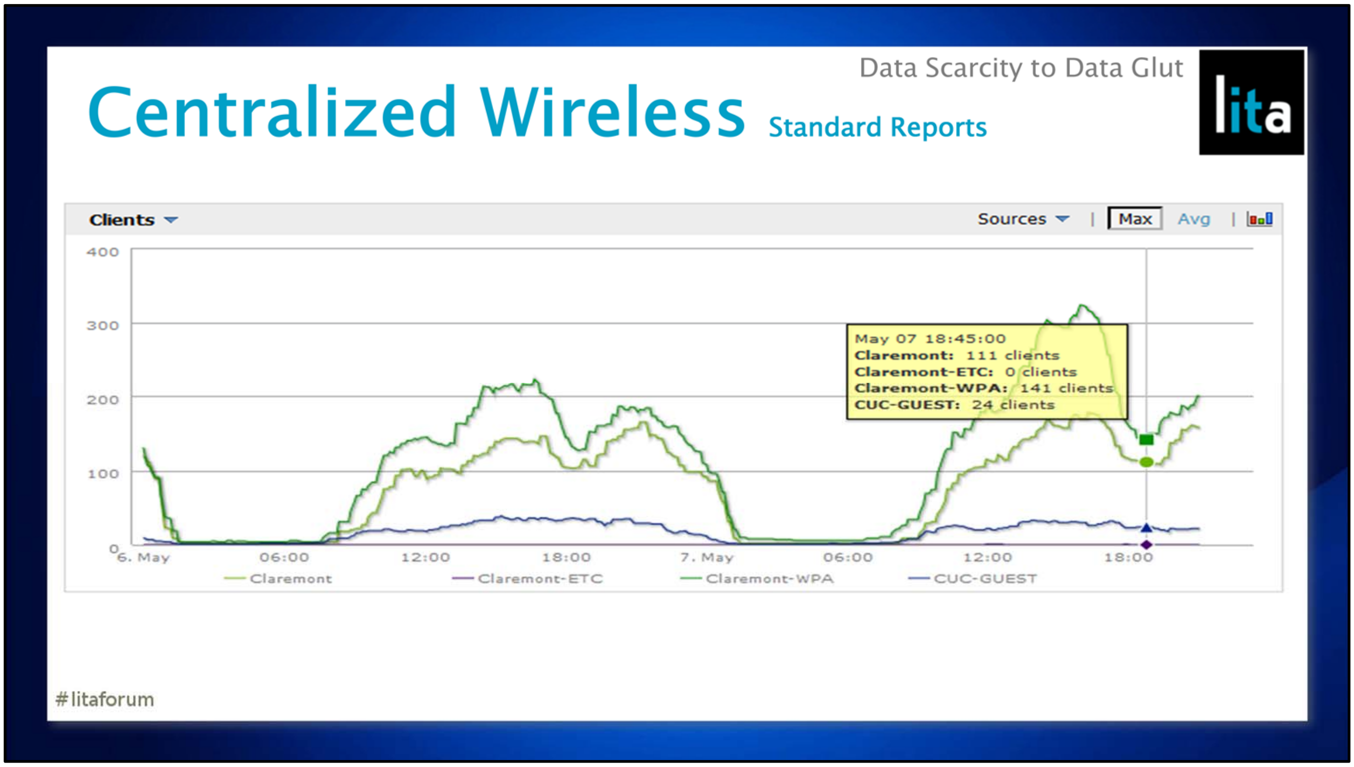
Data Scarcity to Data Glut

Centralized Wireless[1]

#litaforum

I wrote about this first part last year in Effortless Building Census

Centralized wireless automatically tells a lot of tales. We just have to ask.
Modern wireless is now commonly provisioned this way .



Data Scarcity to Data Glut

Centralized Wireless Standard Reports

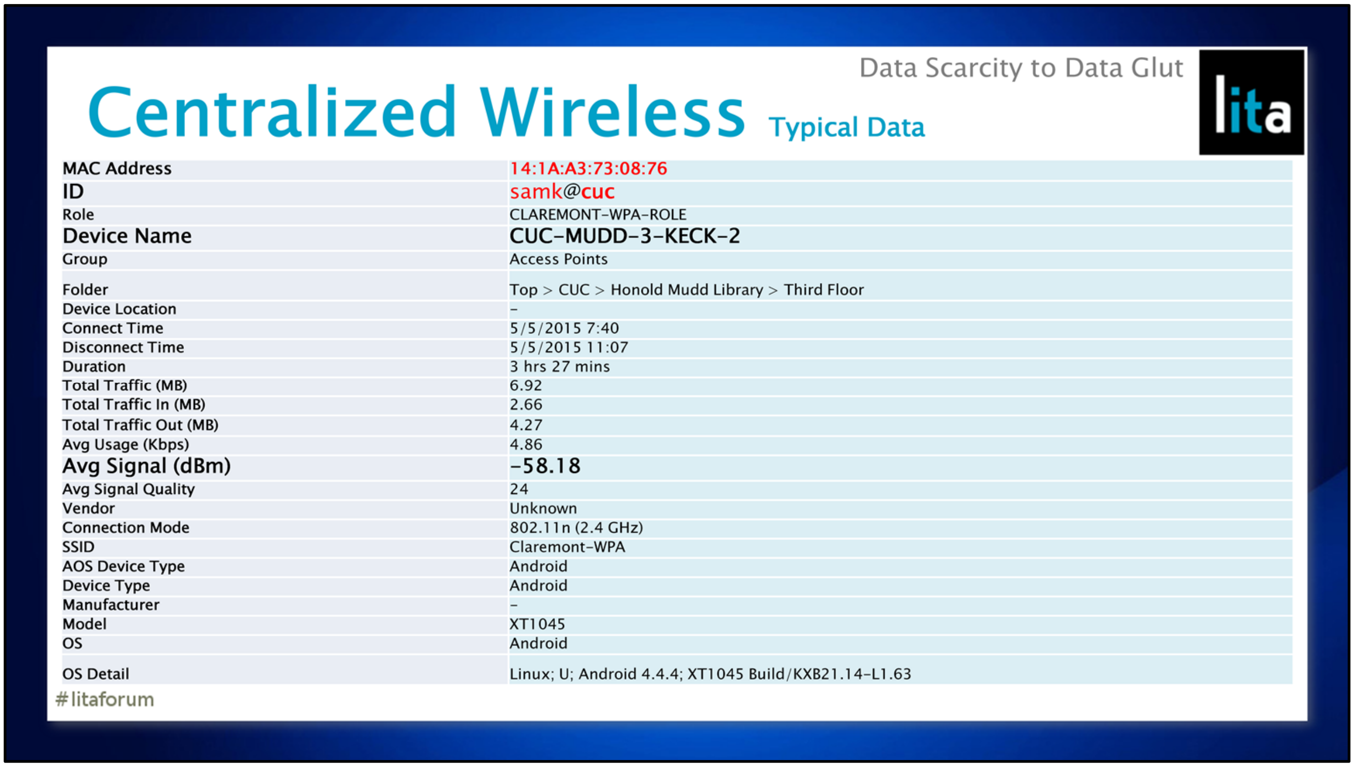
#litaforum

So that's the setup. A central controller provides wireless to clients via a fleet of access points. The controller also logs some aspects of the activity. Is anyone wondering what gets logged? It's important, right?

Your mileage will vary according to make and model. By logging onto the wireless controller we can see a dashboard view of the number of **devices**, aka clients, connected to the controller by arbitrary time periods.

Now if we assume that everyone carries one wireless device connected to the network, this is already roughly equivalent to a headcount, and can be fairly location-specific and is independent of time. Any time period can be examined.

Of course one device per person is a pretty big assumption. The 64k question is, how many patrons per device, or devices per patron. Hold onto that question.



Data Scarcity to Data Glut

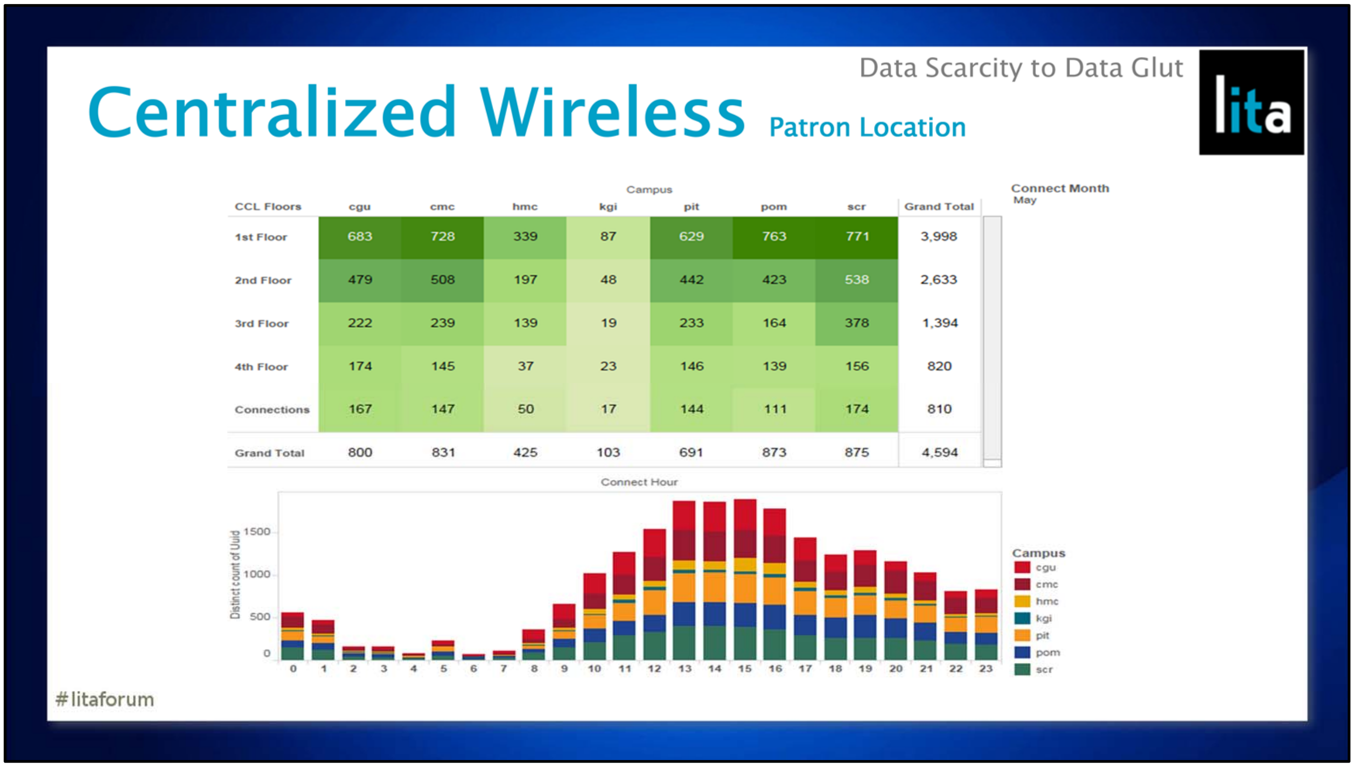
Centralized Wireless Typical Data

MAC Address	14:1A:A3:73:08:76
ID	samk@cuc
Role	CLAREMONT-WPA-ROLE
Device Name	CUC-MUDD-3-KECK-2
Group	Access Points
Folder	Top > CUC > Honold Mudd Library > Third Floor
Device Location	-
Connect Time	5/5/2015 7:40
Disconnect Time	5/5/2015 11:07
Duration	3 hrs 27 mins
Total Traffic (MB)	6.92
Total Traffic In (MB)	2.66
Total Traffic Out (MB)	4.27
Avg Usage (Kbps)	4.86
Avg Signal (dBm)	-58.18
Avg Signal Quality	24
Vendor	Unknown
Connection Mode	802.11n (2.4 GHz)
SSID	Claremont-WPA
AOS Device Type	Android
Device Type	Android
Manufacturer	-
Model	XT1045
OS	Android
OS Detail	Linux; U; Android 4.4.4; XT1045 Build/KXB21.14-L1.63

#litaforum

We parse off the campus affiliation from the network user id and then anonymize the id. In our case, which may be common for Shibboleth/CAS type environments, campus affiliation is part of the user network id.

That can be done with Open Refine, python, or your text mangling tool of choice.



Data Scarcity to Data Glut

Centralized Wireless Patron Location

Campus

CCL Floors	cgu	cmc	hmc	kgi	pit	pom	scr	Grand Total
1st Floor	683	728	339	87	629	763	771	3,998
2nd Floor	479	508	197	48	442	423	538	2,633
3rd Floor	222	239	139	19	233	164	378	1,394
4th Floor	174	145	37	23	146	139	156	820
Connections	167	147	50	17	144	111	174	810
Grand Total	800	831	425	103	691	873	875	4,594

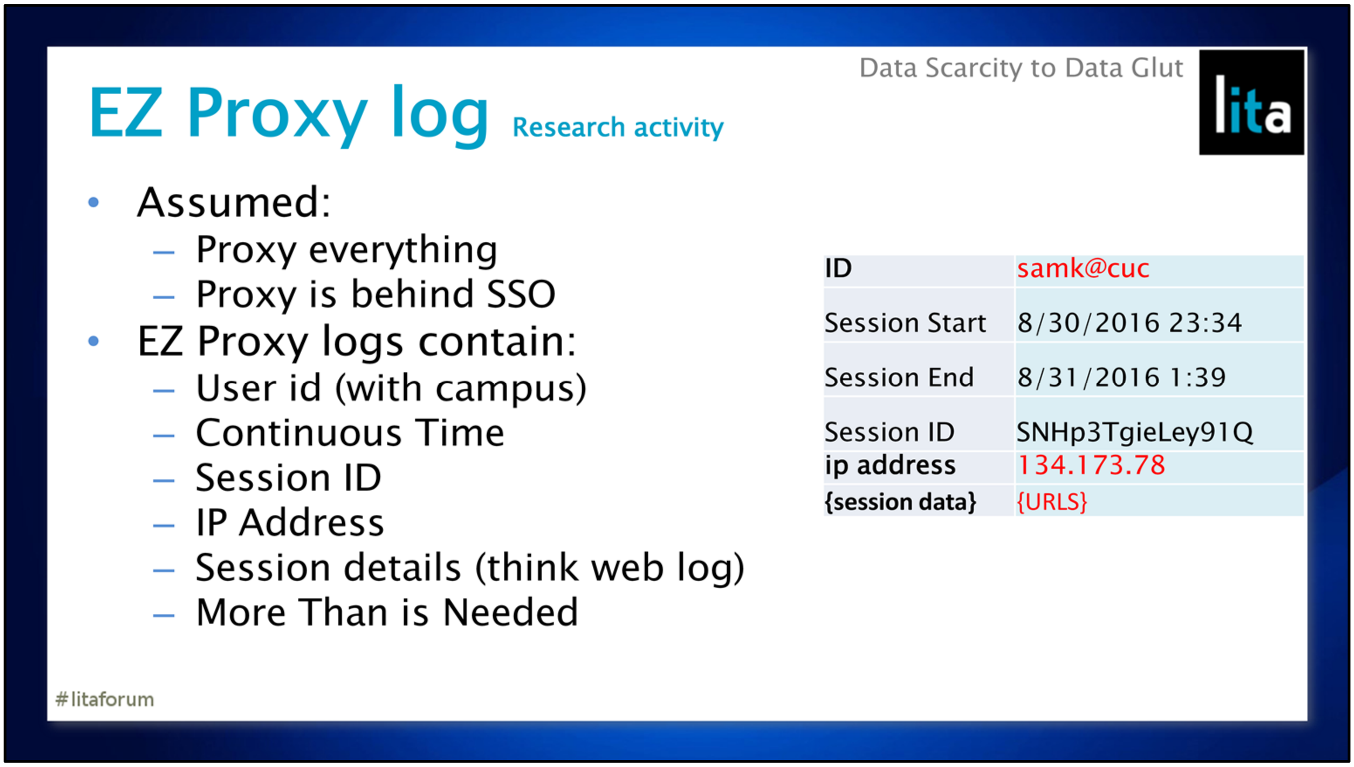
Connect Month
May

#litaforum

A graphic example of how signal strength can suggest patron location. Actual data, not actual locations, ids are obscured.   Let's look at the actual data.
[Image:
To figure out where patrons are in the building, it very much helps to name the access points for their buildilng locations. Let's look at that.

Consider that the access points provide wireless internet access to a discrete radius. Their locations are carefully chosen in order to provide full building coverage, so they are distributed on each and every floor, densest where higher populations are expected.

The signal strength of each client's connection is logged and can be reported. That signal strength number will helps us refine the client's location.
[Image: Heat map with APs overlaid]
Heat map with APs overlaid]



Data Scarcity to Data Glut

EZ Proxy log Research activity

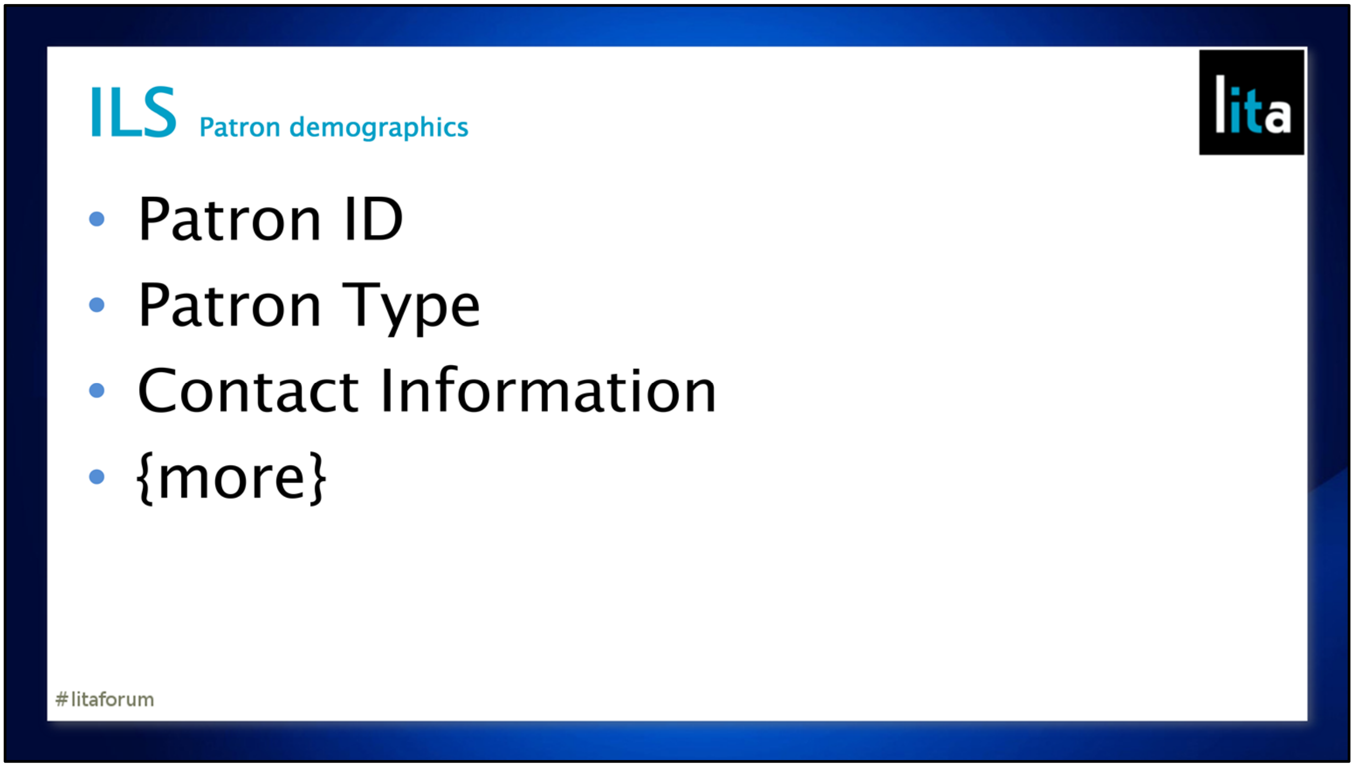
- Assumed:
  - Proxy everything
  - Proxy is behind SSO
- EZ Proxy logs contain:
  - User id (with campus)
  - Continuous Time
  - Session ID
  - IP Address
  - Session details (think web log)
  - More Than is Needed

ID	samk@cuc
Session Start	8/30/2016 23:34
Session End	8/31/2016 1:39
Session ID	SNHp3TgieLey91Q
ip address	134.173.78
{session data}	{URLS}

#litaforum

Proxy everything
Proxy is behind SSO/CAS

Because it’s behind SSO, the User ID here is the same as for wireless



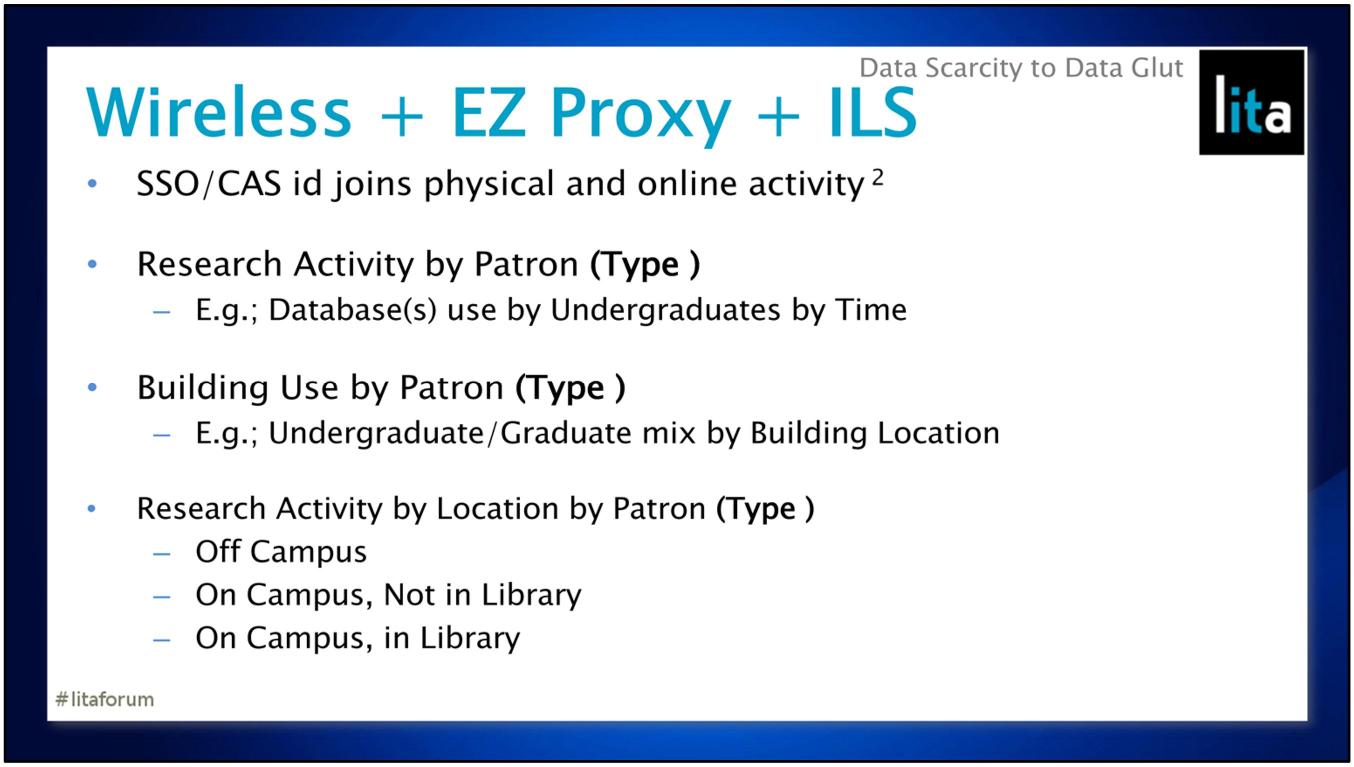
ILS Patron demographics

- Patron ID
- Patron Type
- Contact Information
- {more}

#litaforum

Authentication to ILS is via SSO/CAS

So the id is the same as for wireless and ezproxy



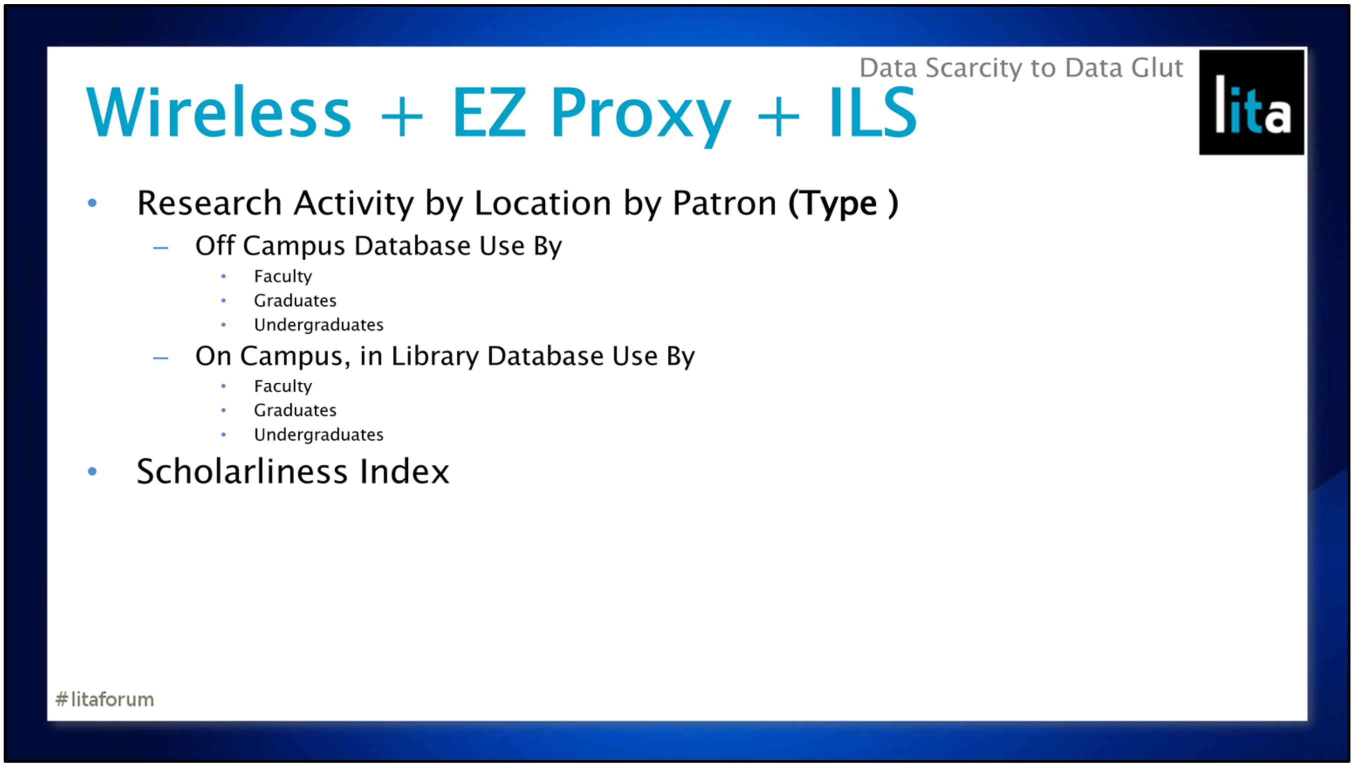
Data Scarcity to Data Glut

Wireless + EZ Proxy + ILS

- SSO/CAS id joins physical and online activity [2]
- Research Activity by Patron **(Type)**
  - E.g.; Database(s) use by Undergraduates by Time
- Building Use by Patron **(Type)**
  - E.g.; Undergraduate/Graduate mix by Building Location
- Research Activity by Location by Patron **(Type)**
  - Off Campus
  - On Campus, Not in Library
  - On Campus, in Library

#litaforum

CNI: Report on the CNI Authentication and Authorization Survey 2016
How about a ‘scholarlyness’ measure for students records or for APT committees? Hours in the library plus articles downloaded, pages turned, etc. Creepy yet?

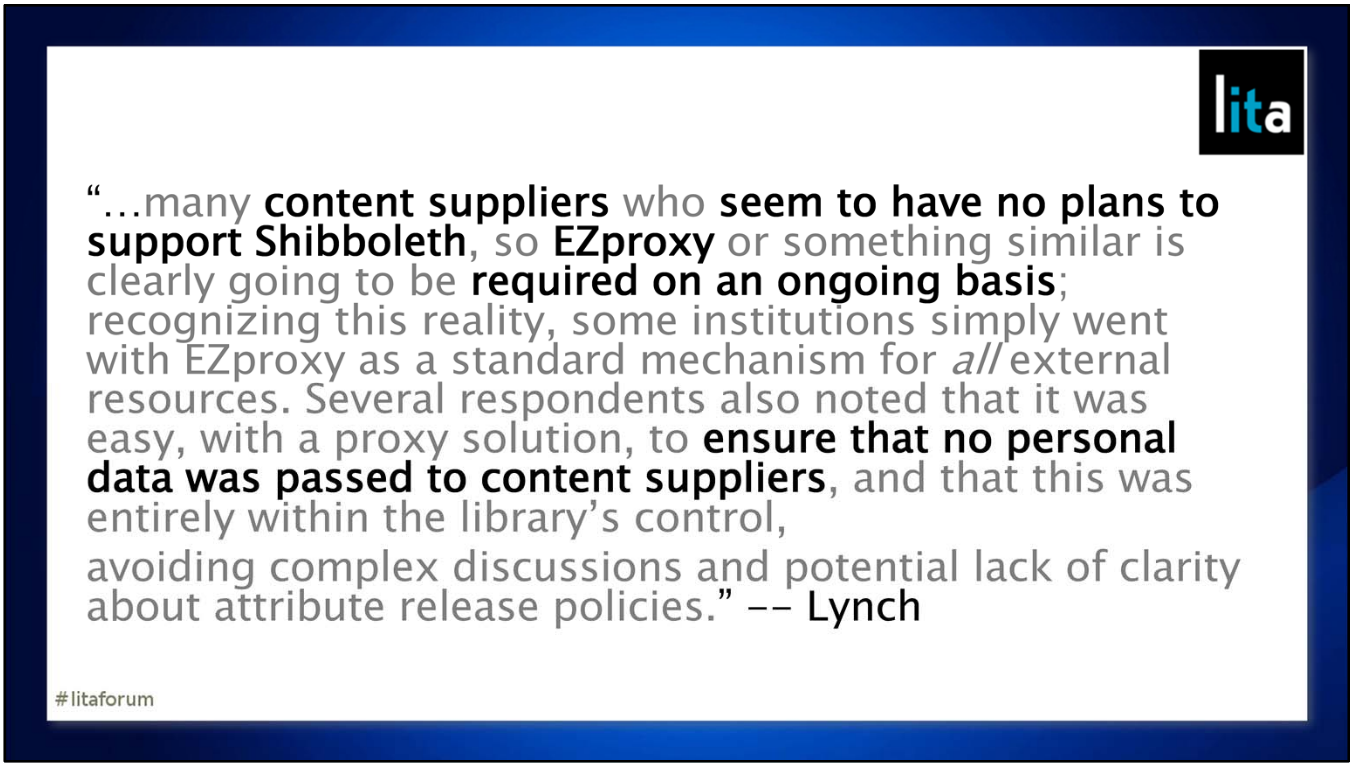


Data Scarcity to Data Glut

Wireless + EZ Proxy + ILS

- Research Activity by Location by Patron **(Type)**
  - Off Campus Database Use By
    - Faculty
    - Graduates
    - Undergraduates
  - On Campus, in Library Database Use By
    - Faculty
    - Graduates
    - Undergraduates
- Scholarliness Index

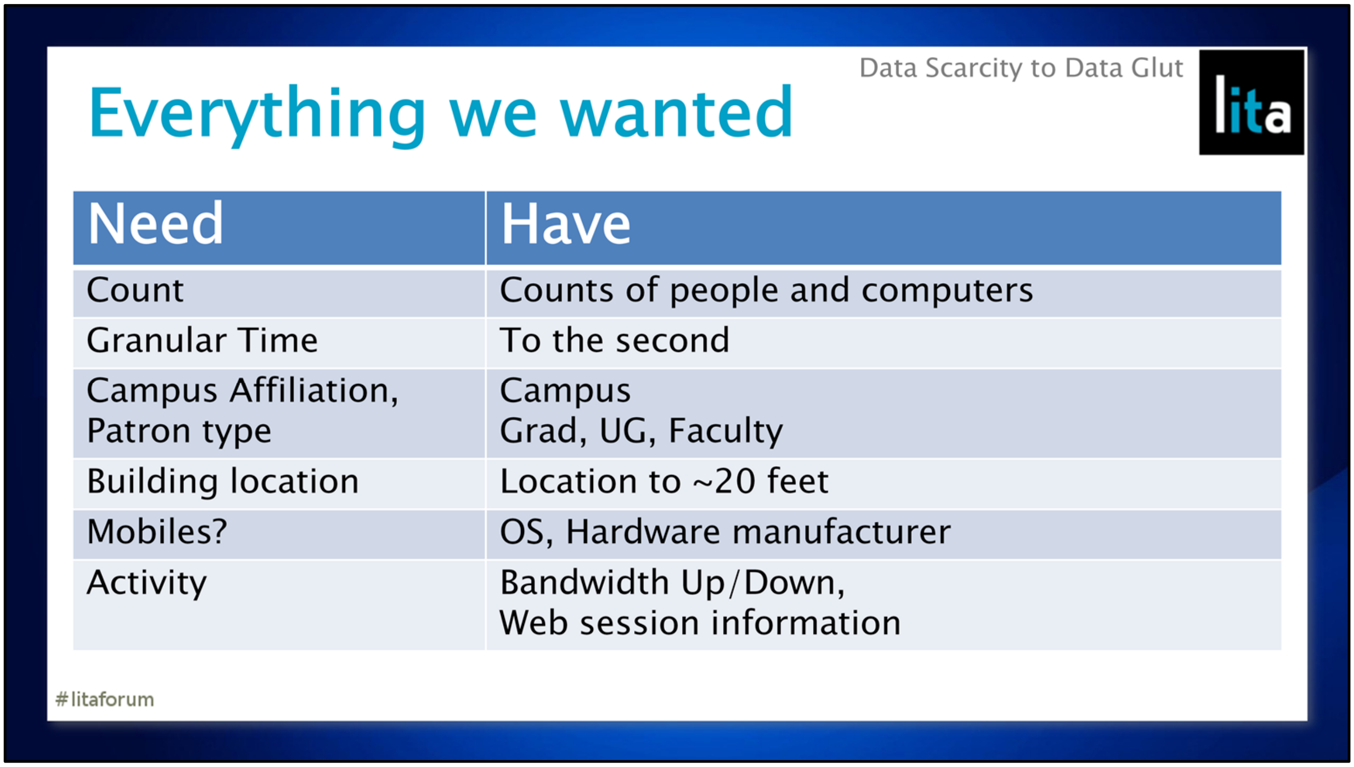
#litaforum



"…many **content suppliers** who **seem to have no plans to support Shibboleth**, so **EZproxy** or something similar is clearly going to be **required on an ongoing basis**; recognizing this reality, some institutions simply went with EZproxy as a standard mechanism for *all* external resources. Several respondents also noted that it was easy, with a proxy solution, to **ensure that no personal data was passed to content suppliers**, and that this was entirely within the library's control,

avoiding complex discussions and potential lack of clarity about attribute release policies." -- Lynch

EZProxy + SSO provides a good patron experience and protects patron privacy.
Passing everything through EZProxy enables nearly 100% report coverage.



Data Scarcity to Data Glut

Everything we wanted

Need	Have
Count	Counts of people and computers
Granular Time	To the second
Campus Affiliation, Patron type	Campus Grad, UG, Faculty
Building location	Location to ~20 feet
Mobiles?	OS, Hardware manufacturer
Activity	Bandwidth Up/Down, Web session information

#litaforum

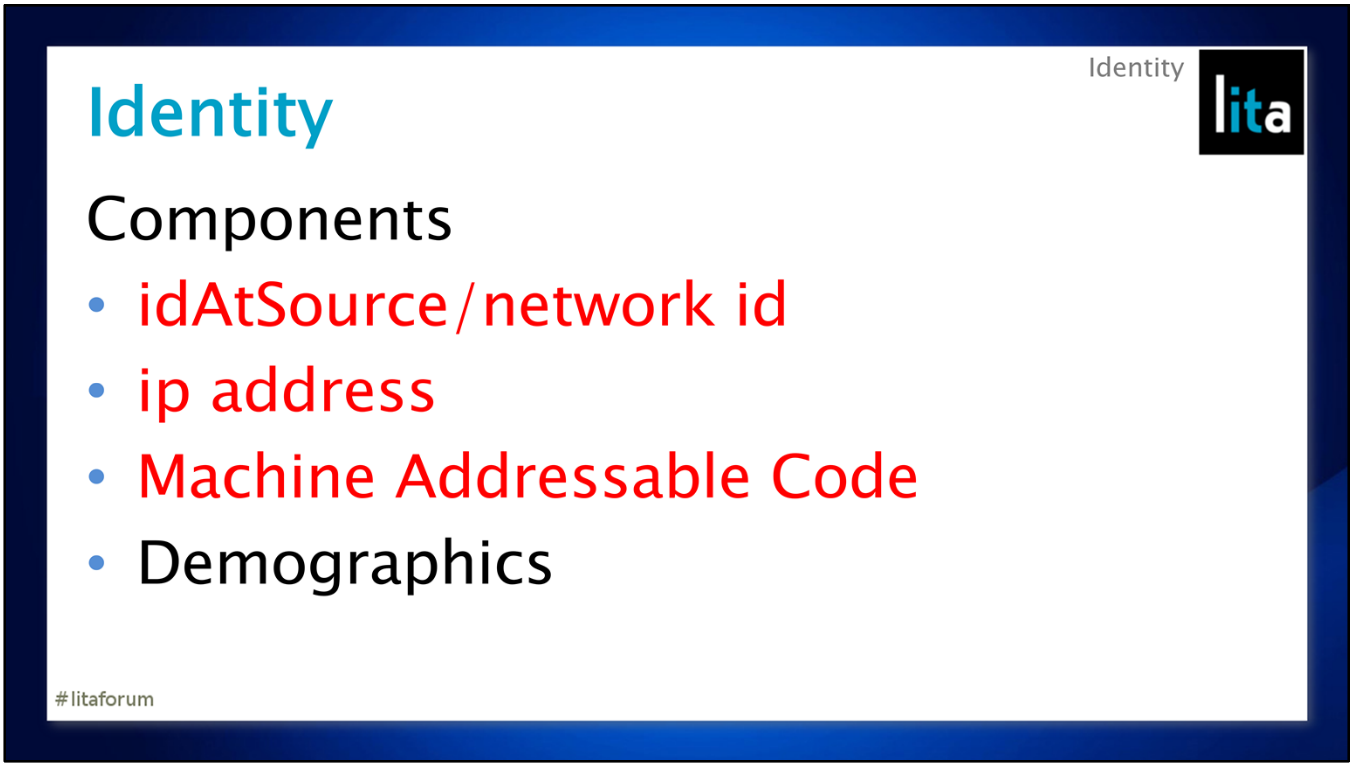
What do we get for our efforts?

First of all, anything we get, can have campus affiliation. That's just the best. Right away we can tell undergraduate from graduate uses. We can tell that engineering undergraduate students are in the library, on the quiet floor!

Our findings fall into the categories of People, Devices, and Time.
The people properties are anonymized unique id, so we can count distinct patrons.
We can consider Campus is one of their properties, and we can count the number of unique machines associated with a login.

We can tell quite a bit about devices; probably the most interesting are operating system and manufacturer, which reveal type: laptop, tablet, phone, watch, etc.

Time is given as a connect and disconnect time. For us, these settings are tricky to use – disconnects don't happen promptly, or ever. That's an IT thing; talk to them.
When we had a period of complaints about wireless, this was useful for diagnosis.
Otherwise we focus on connect time.



Identity

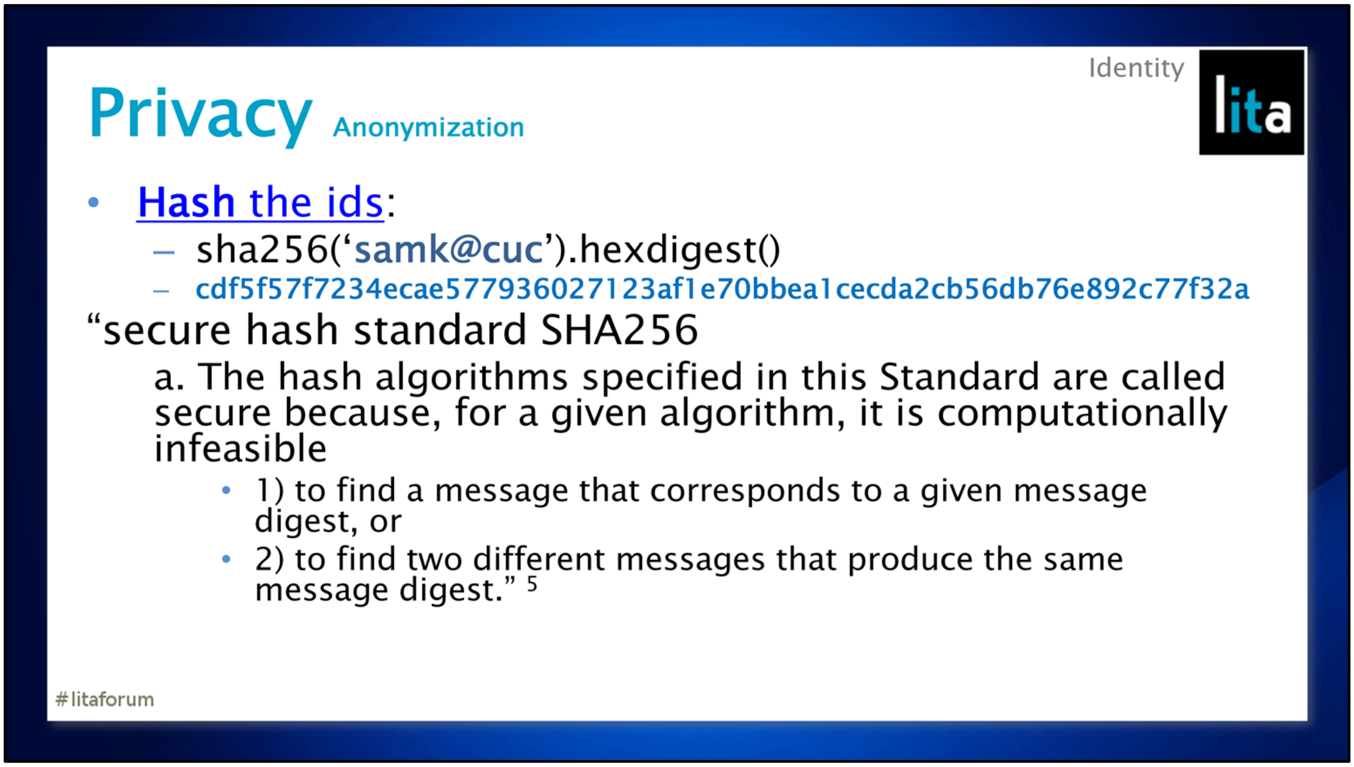
Identity

Components

- idAtSource/network id
- ip address
- Machine Addressable Code
- Demographics

#litaforum

Be aware
Anonymize everything in red.
PII and reverse engineered ID



Identity

Privacy Anonymization

- Hash the ids:
  - sha256(‘samk@cuc’).hexdigest()
  - cdf5f57f7234ecae577936027123af1e70bbea1cecda2cb56db76e892c77f32a

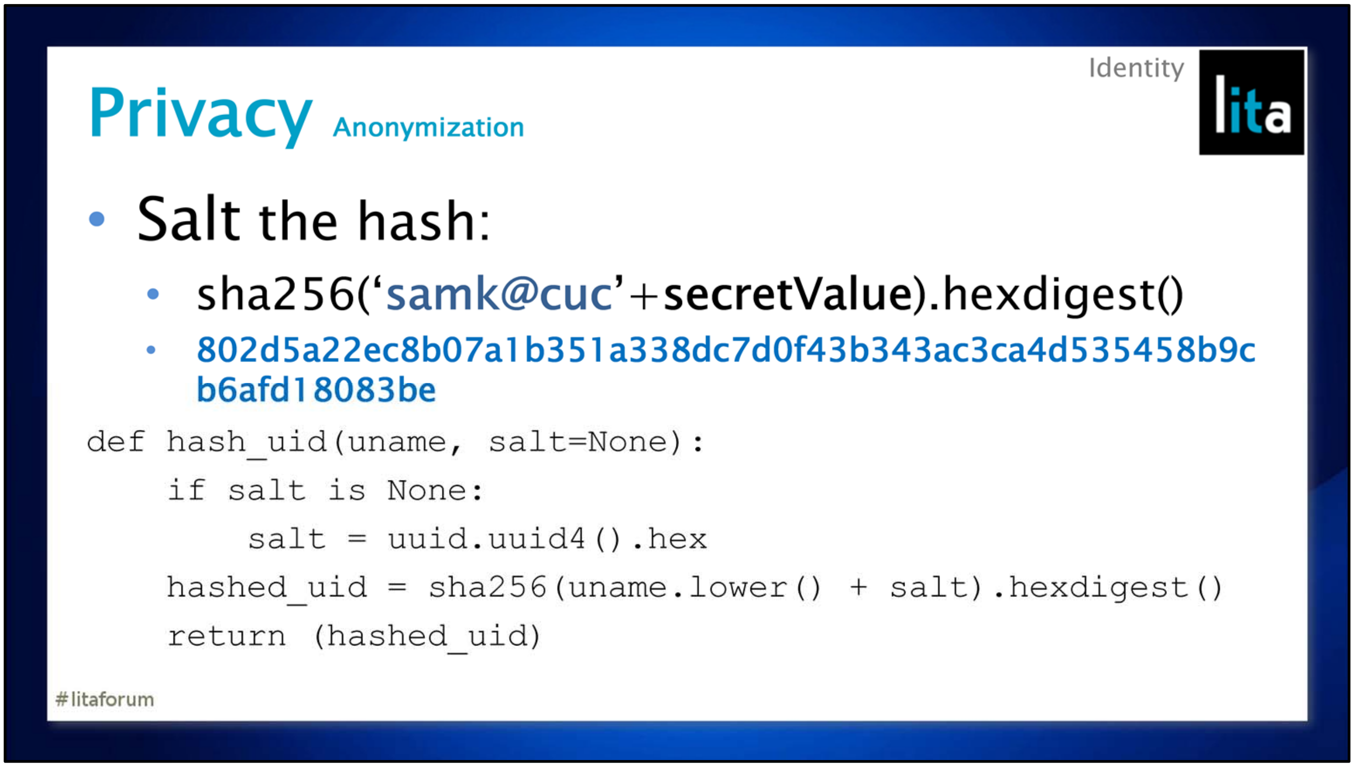
“secure hash standard SHA256

a. The hash algorithms specified in this Standard are called secure because, for a given algorithm, it is computationally infeasible

- 1) to find a message that corresponds to a given message digest, or
- 2) to find two different messages that produce the same message digest.” [5]

#litaforum

Go to https://duckduckgo.com/
A straight hash is NOT sufficient to anonymize user ids. An attacker can obtain the un-hashed values, and try standard hashes.



Identity

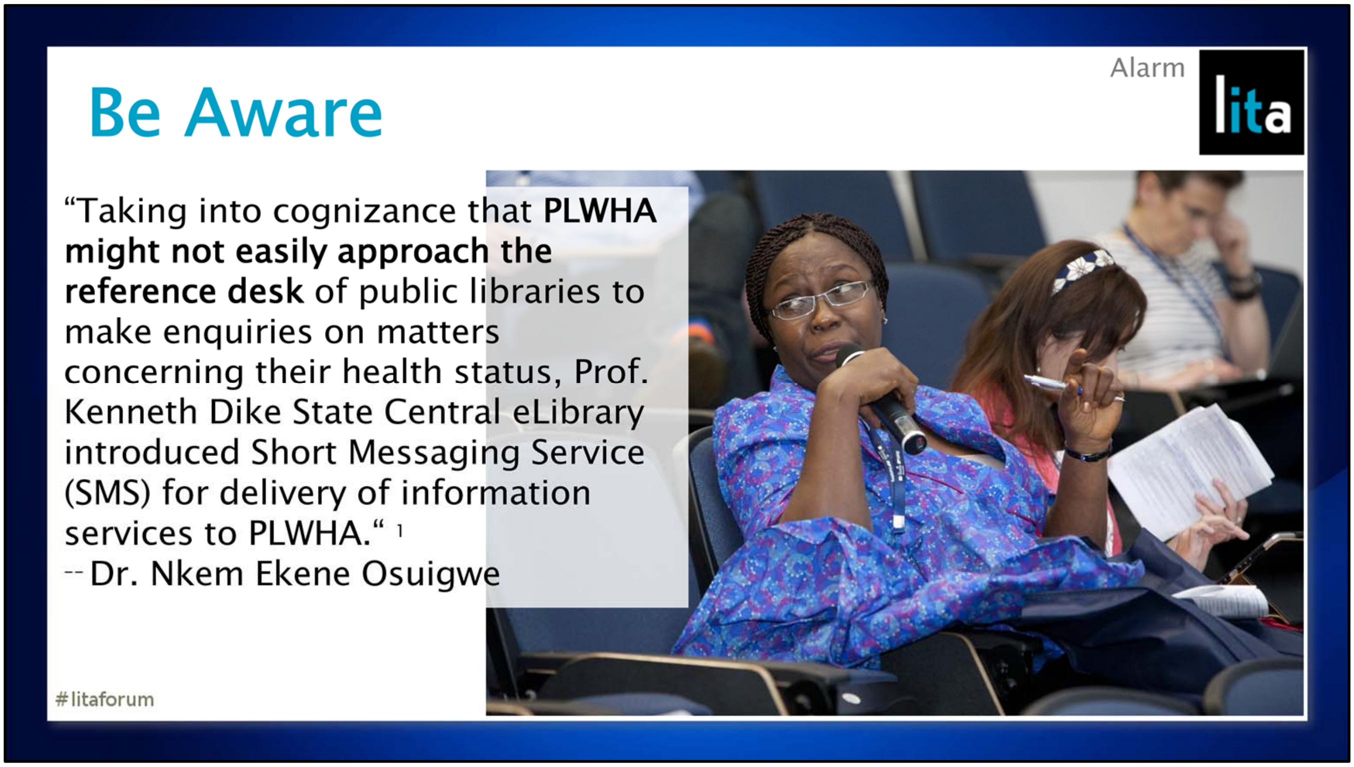
Privacy Anonymization

- Salt the hash:
  - sha256(‘samk@cuc’+secretValue).hexdigest()
  - 802d5a22ec8b07a1b351a338dc7d0f43b343ac3ca4d535458b9c b6afd18083be

```
def hash_uid(uname, salt=None):
    if salt is None:
        salt = uuid.uuid4().hex
    hashed_uid = sha256(uname.lower() + salt).hexdigest()
    return (hashed_uid)
```

#litaforum

It is not sufficient to simply hash the id because an attacker can obtain the unencrypted user ids and employ different hashes.
So add a secret value to the id strings.
Keep that value out of your git repo!
git ignore .cfg



Alarm

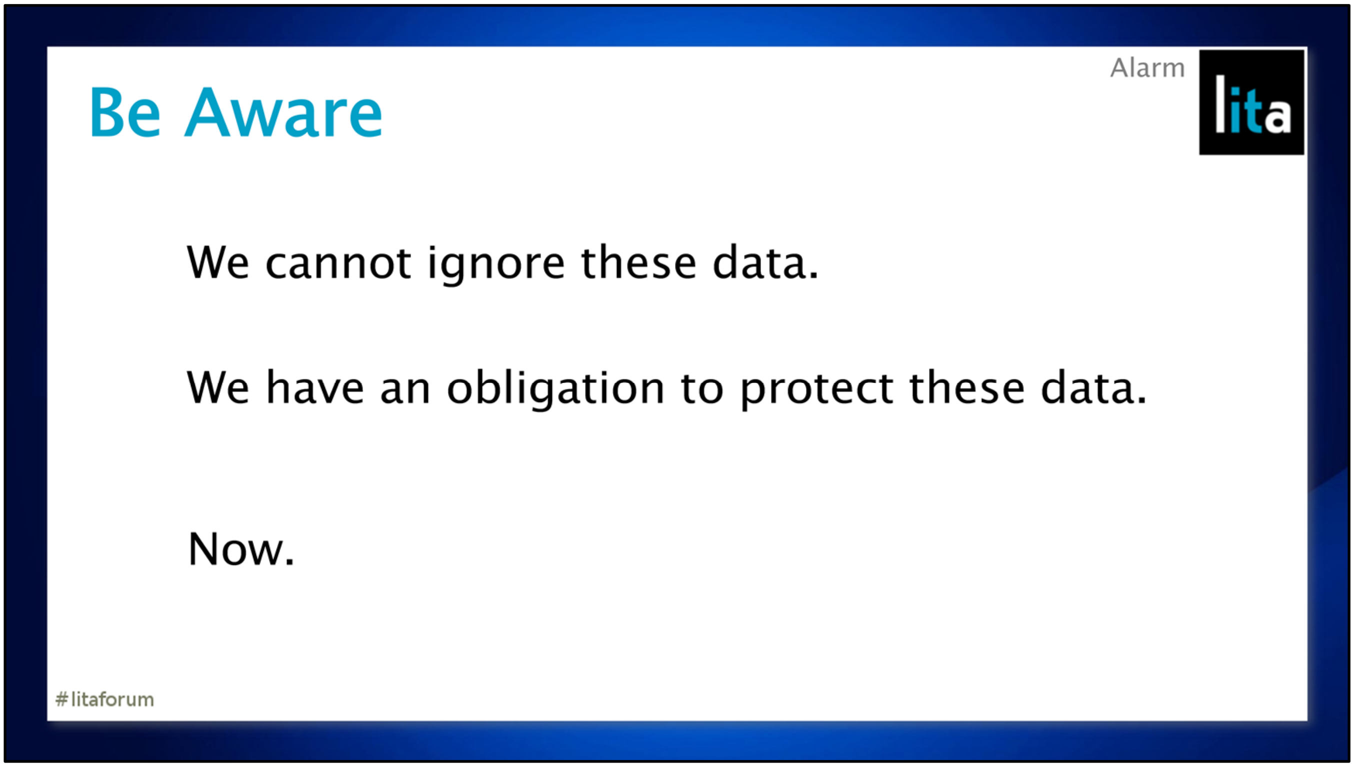
Be Aware

"Taking into cognizance that **PLWHA might not easily approach the reference desk** of public libraries to make enquiries on matters concerning their health status, Prof. Kenneth Dike State Central eLibrary introduced Short Messaging Service (SMS) for delivery of information services to PLWHA." [1]
-- Dr. Nkem Ekene Osuigwe

#litaforum

Dr. Osuigwe's quote understates the problem. She went on to describe an elaborate, blind information resource system for People Living With HIV/AIDS in Nigeria. The system is blind because it is not safe for people with HA or identified as particular races, classes, or religions to use a library. Not safe to ask questions. Not safe to seek to improve their lives.

Does this sound familiar? It didn't to me. Nigeria is far away and in America we celebrate free speech and value education, right?



Alarm

Be Aware

We cannot ignore these data.

We have an obligation to protect these data.

Now.

#litaforum

Which brings us to the crux. We have to do something about our data. They are too rich and too vulnerable.
This process I've described is in the wild; we LITA folks all over are building datasets with highly detailed, highly desirable patron behavior data.

Assume your networks are compromised – they almost certainly are. Cisco says so.
Assume your patrons are not safe – many are not.

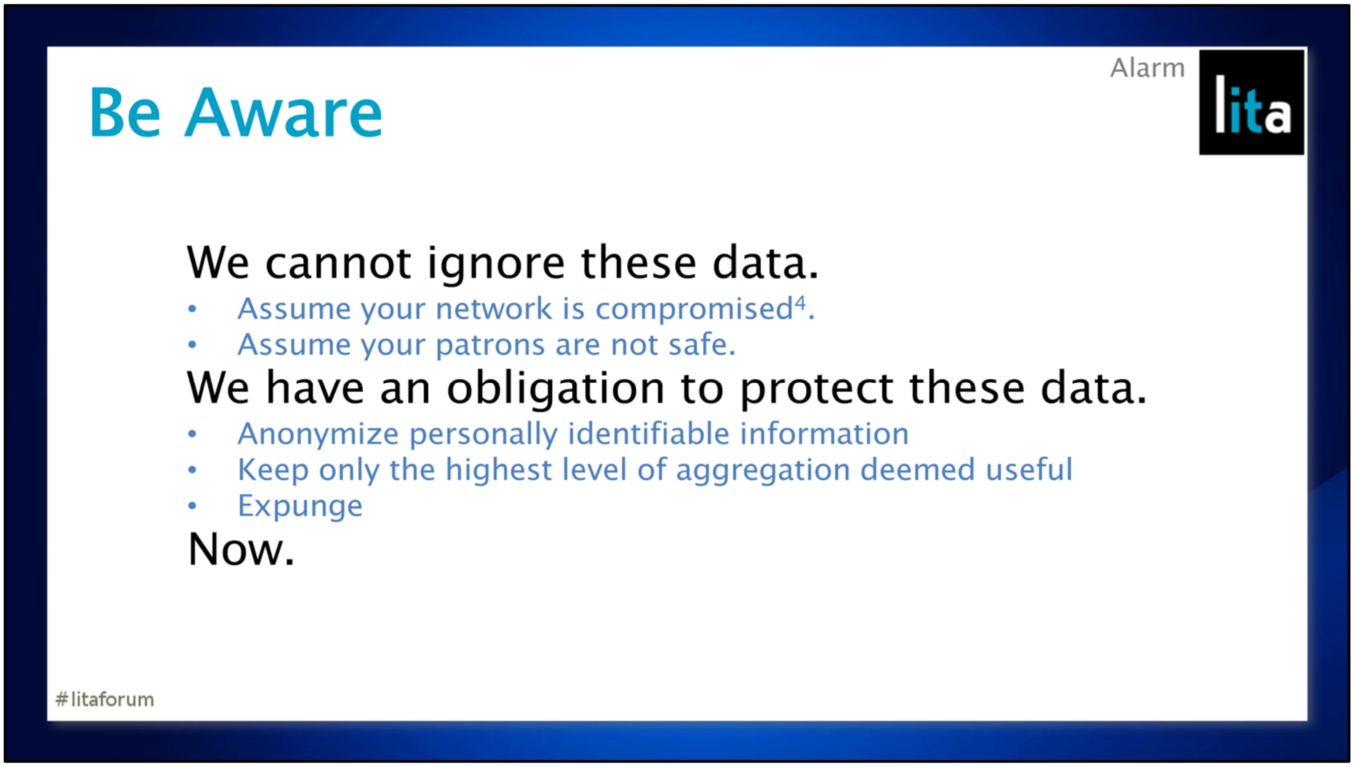
With the advent of the USA PATRIOT Act we found ourselves in an ethical conflict over patron privacy, between scholarship and political doctrine. Patron borrowing records were and are misguidedly sought with no regard to the Constitution nor to law enforcement efficacy.

That conflict is now worse than ever. We have far more data and we will soon have a federal administration objectively hostile to privacy, hostile to education, and quite possibly hostile to other freedoms we now take for granted, like the ability to walk into a library, or to freely access information online. The people who want to populate racist registries will be drawn to library data that right now today contain to-the-moment, 360 degree views of our patrons.

We have an obligation to our patrons to protect these data.
We also have an obligation to our administrations to use these data to good ends: to design services, to increase efficiency, to effectively budget.

These two issues must be reconciled.
Anonymize personally identifiable elements
Keep only the highest level of aggregation deemed useful
Expunge

Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism



Alarm

Be Aware

We cannot ignore these data.

- Assume your network is compromised[4].
- Assume your patrons are not safe.

We have an obligation to protect these data.

- Anonymize personally identifiable information
- Keep only the highest level of aggregation deemed useful
- Expunge

Now.

#litaforum

Which brings us to the crux. We have to do something about our data. They are too rich and too vulnerable.
Assume your networks are compromised – they almost certainly are. Cisco says so.
Assume your patrons are not safe – many are not.

With the advent of the USA PATRIOT Act we found ourselves in an ethical conflict over patron privacy, between scholarship and political doctrine. Patron borrowing records were and are misguidedly sought with no regard to the Constitution nor to law enforcement efficacy.

That conflict is now worse than ever. We have far more data and we will soon have a federal administration objectively hostile to privacy, hostile to education, and quite possibly hostile to other freedoms we now take for granted, like the ability to walk into a library, or to freely access information online. The people who want to populate racist registries will be drawn to library data that right now today contain to-the-moment, 360 degree views of our patrons.

We have an obligation to our patrons to protect these data.
We also have an obligation to our administrations to use these data to good ends: to design services, to increase efficiency, to effectively budget.
These two issues must be reconciled.
Anonymize personally identifiable elements
Keep only the highest level of aggregation deemed useful

Expunge

Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism

References

1. Kome, S. (2015). Effortless Building Census. *Library Staff Publications and Research*. Retrieved from http://scholarship.claremont.edu/library_staff/34

2. Report on the CNI Authentication and Authorization Survey 2016. (2016, August 3). Retrieved November 19, 2016, from https://www.cni.org/publications/cliffs-pubs/authentication-authorization-survey-2016

3. Osuigwe, N. (2016). Leveraging On Organizational Culture For Innovative Services: A Case Study Of Prof. Kenneth Dike State Central eLibrary, Awka. *Library Philosophy and Practice (E-Journal)*. Retrieved from http://digitalcommons.unl.edu/libphilprac/1447

4. Get the Latest Findings on Malware Threats. (n.d.). Retrieved November 19, 2016, from http://www.cisco.com/web/offers/lp/2014-annual-security-report/index.html

5. Federal Information Processing Standards Publication 180-4

Skome's wifi code @ github

#litaforum

Questions?

Thank You!

Feedback: https://www.surveymonkey.com/r/2016litaforum

#litaforum