Data and Data Studies

Aleksi Aaltonen
Stevens Institute of Technology

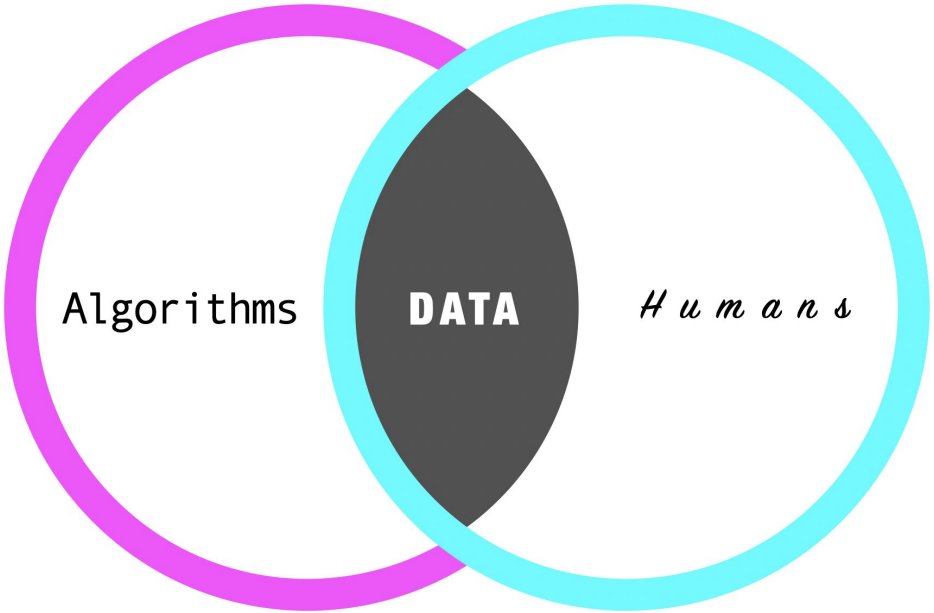
Marta Stelmaszak Rosa
University of Massachusetts Amherst

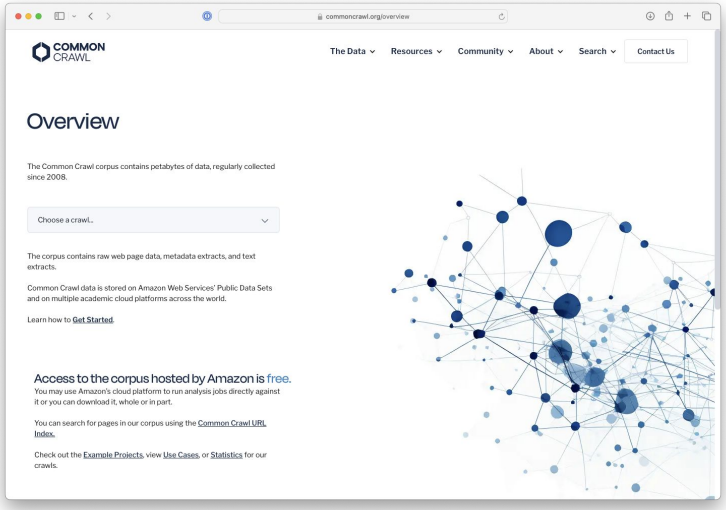
SIG DITE PhD Research Academy 17 April 2026

AGENDA

Why study data?	1
What are data?	2
What are digital data?	3
Data studies	4
Q&A	5

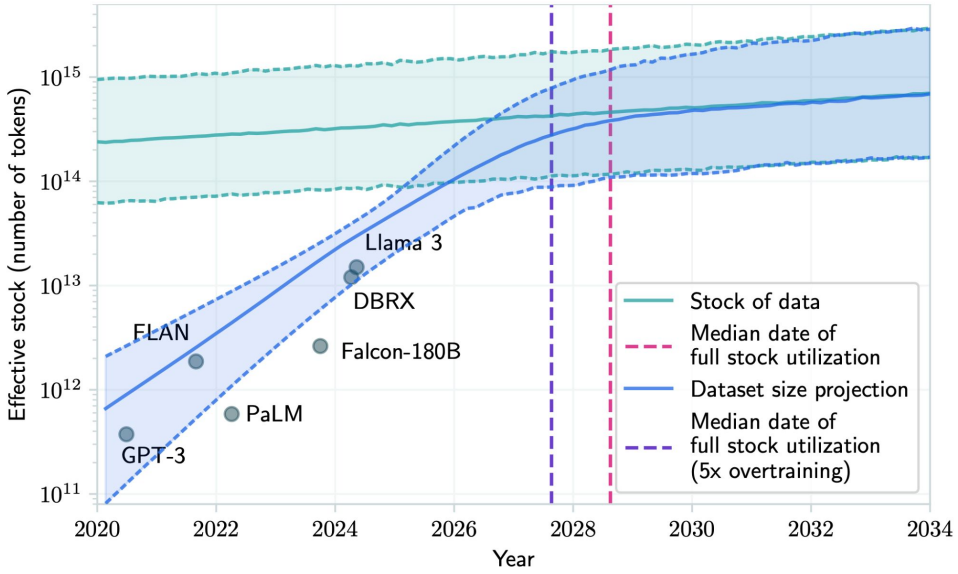
Why study data?



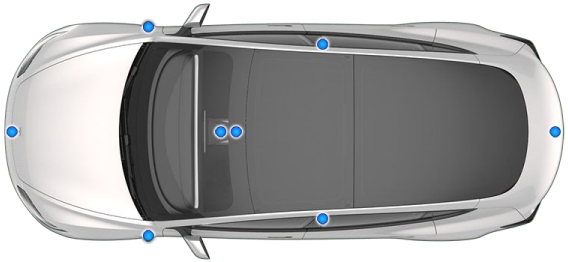


https://commoncrawl.org

We are running out of human-generated training data!



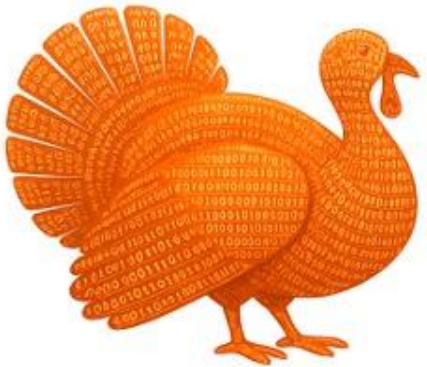
Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2022). Will we run out of data? Limits of LLM scaling based on human-generated data (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2211.04325



What if all these would record everything all the time?

Still, however much and however detailed data we have,

ALL data are wrong, but some ~~are~~ (may be) useful.



Tsagbey, S., De Carvalho, M., & Page, G. L. (2017). All Data are Wrong, but Some are Useful? Advocating the Need for Data Auditing. The American Statistician, 71(3), 231–235.

The study of data is more relevant than ever!

1. Data mediates between algorithmic systems, organization, and their environment (e.g., Baskerville et al. 2020; Recker et al. 2021).
2. AI turns everything into potential training data → yet, we have currently tapped just the fraction of potential training data sources.
3. Data are never perfect representations of reality (but they may increasingly become the reality that matters).
4. Data, their production and use are not neutral but touches upon the most intimate and sensitive issues of our lives (Markus 2026, Ngwenyama et al. 2024, Zuboff 2015).

What are data?

WHAT ARE DATA?

1 Physical data

Characteristics

- Recorded when physical form created
- Change requires destruction
- Tied to a specific purpose

Example

Clay tokens
(Alaimo & Kallinikos, 2024)

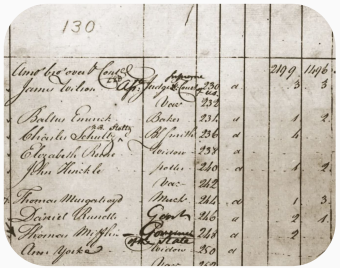
2 Analog data

Characteristics

- Recorded when needed but rarely
- Changes very difficult
- Created for and tied to specific purposes

Example

Census record
(Hirschheim et al., 1995)



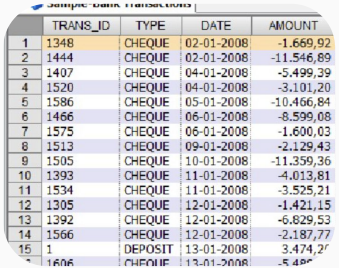
3 Computerized data

Characteristics

- Recorded at the point of transaction
- Changes not impossible but laborious
- Other systems can enable reuse

Example

Transaction record
(Wand & Weber, 1995)



	TRANS_ID	TYPE	DATE	AMOUNT
1	1348	CHEQUE	02-01-2008	-1.669,92
2	1444	CHEQUE	02-01-2008	-11.546,89
3	1407	CHEQUE	04-01-2008	-5.499,39
4	1520	CHEQUE	04-01-2008	-3.101,20
5	1586	CHEQUE	05-01-2008	-10.466,84
6	1466	CHEQUE	06-01-2008	-8.599,08
7	1575	CHEQUE	06-01-2008	-1.600,03
8	1513	CHEQUE	09-01-2008	-2.129,43
9	1505	CHEQUE	10-01-2008	-11.359,36
10	1393	CHEQUE	11-01-2008	-4.013,81
11	1534	CHEQUE	11-01-2008	-3.525,21
12	1305	CHEQUE	12-01-2008	-1.421,15
13	1392	CHEQUE	12-01-2008	-6.829,53
14	1566	CHEQUE	12-01-2008	-2.187,77
15	1	DEPOSIT	13-01-2008	3.474,2
	1606	CHEQUE	13-01-2008	-5.48

4 Digital data

Characteristics

- Recorded whenever an activity takes place
- Very easy to edit
- Collected for unspecified purposes and can be easily reused

Example

Online behavioral traces
(Aaltonen & Stelmaszak, 2024)

Data as Records

Core view: Objective, factual inscriptions of real-world entities or events; neutral inputs reflecting reality.

Characteristics: Intrinsic, measurable properties (e.g., accuracy, completeness, consistency); stable across contexts.

Representative papers: Ballou and Tayi (1985); Wang and Strong (1996); Lee et al. (2002); Goodhue et al. (1992); Hackathorn & Karimi (1988); Wixom & Watson (2001); Dey & Kumar (2013)

Data as Representations

Core view: Symbolic abstractions shaped by modeling grammars, human cognition, and communicative intent.

Characteristics: Semantic and interpretive; dependent on modeling context, user understanding, and design intent.

Representative papers: Wand and Weber (1995); Lyytinen (1985); Storey and Goldstein (1993); Parsons (1996); Chan et al. (1993); Batra & Zanakis (1994); Byrd et al. (1992)

Data as artifacts

Core view: Constructed, performative, and infrastructurally embedded entities that co-constitute the realities they describe.

Characteristics: Syntactic, semantic, and infrastructural; mutable, referentially unclear, and designed for performativity.

Representative papers: Alaimo & Kallinikos (2022); Eriksson & Ågerfalk (2022); Stelmaszak et al. (2024); Parmiggiani et al. (2022); Aaltonen & Tempini (2014); Günther et al. (2022); Monteiro & Parmiggiani (2019)

dsb data studies bibliography

WHAT ARE DIGITA DATA?

“Records of events and entities (in the physical, digital, or artificial world) encoded into a computational medium” (Stelmaszak et al., 2026)

- **Updatability** — Constantiou & Kallinikos, 2015
- **Structure** — Aaltonen & Penttinen, 2020
- **Editability** — Aaltonen et al., 2021
- **Granularity** — Yoo, 2015
- **Heterogeneity** — Orlikowski & Scott, 2014
- **Functional coupling** — Kallinikos et al., 2021

WHAT ARE DIGITAL DATA?

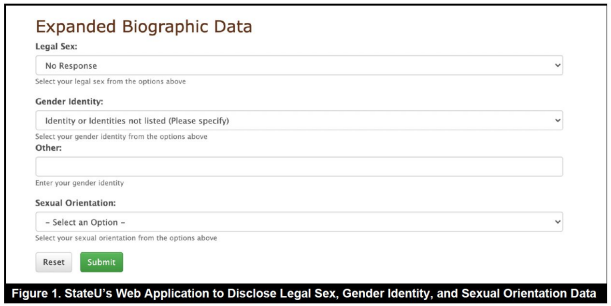
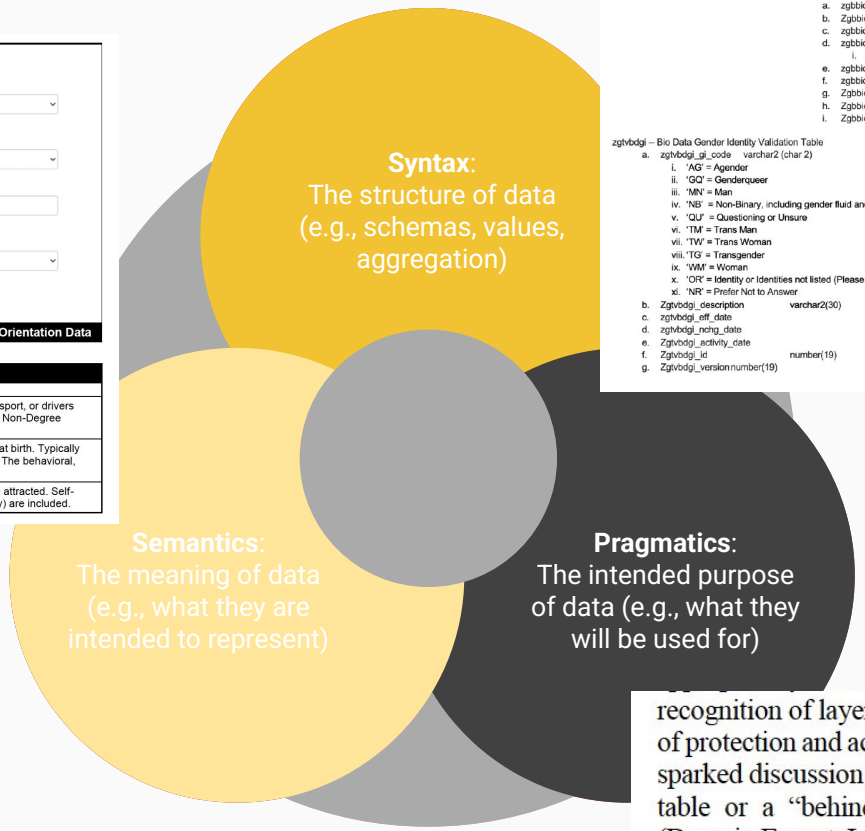


Figure 1. StateU's Web Application to Disclose Legal Sex, Gender Identity, and Sexual Orientation Data

Table 2. Functional Definitions of SOGI Data at StateU (Internal Document)

Term	Functional definition
Legal sex designation	The sex marker that is on government-issued identification such as SS card, passport, or drivers license. Self-reported on admissions applications for students (all Admission and Non-Degree applications). Designation is verified for employees.
Gender identity	The internal sense of one's gender that may be different than their sex assigned at birth. Typically used with reference to social and cultural differences rather than biological ones. The behavioral, cultural, or psychological traits typically associated with one sex.
Sexual orientation	An individual's sexual identity, which may include the gender(s) to which they are attracted. Self-reported on all forms where biographical data (including gender, race, or ethnicity) are included.



zgbbiod – Biographic Data table
- a. zgbbiod_pidm number (8)
- b. Zgbbiod_aidm number (8)
- c. zgbbiod_gi_code varchar2 (char 1)
- d. zgbbiod_gi_other varchar2 (char 30)
  - i. populated if zgbbiod_gi_code = 'OR' ("Identity or identities not listed")
- e. zgbbiod_so_code varchar2 (char 2)
- f. zgbbiod_so_other varchar2 (char 30)
- g. Zgbbiod_activity_date date
- h. Zgbbiod_id number(19)
- i. Zgbbiod_version number(19)

zgtvbdgi – Bio Data Gender Identity Validation Table
- a. zgtvbdgi_gi_code varchar2 (char 2)
  - i. 'AG' = Agender
  - ii. 'GQ' = Genderqueer
  - iii. 'MN' = Man
  - iv. 'NB' = Non-Binary, including gender fluid and gender non-conforming
  - v. 'QU' = Questioning or Unsure
  - vi. 'TM' = Trans Man
  - vii. 'TW' = Trans Woman
  - viii. 'TG' = Transgender
  - ix. 'WM' = Woman
  - x. 'OR' = Identity or Identities not listed (Please specify)
  - xi. 'NR' = Prefer Not to Answer
- b. Zgtvbdgi_description varchar2(30)
- c. zgtvbdgi_eff_date
- d. zgtvbdgi_nchg_date
- e. Zgtvbdgi_activity_date
- f. Zgtvbdgi_id number(19)
- g. Zgtvbdgi_version number(19)

zgtvbdso – Bio Data Sexual Orientation Validation Table
- a. zgtvbdso_so_code varchar2 (char 2)
  - i. 'AS' = Asexual
  - ii. 'BS' = Bisexual
  - iii. 'GA' = Gay
  - iv. 'HS' = Heterosexual/Straight
  - v. 'LE' = Lesbian
  - vi. 'PS' = Pansexual
  - vii. 'QR' = Queer
  - viii. 'QU' = Questioning or Unsure
  - ix. 'SG' = Same Gender Loving
  - x. 'OR' = Identity or Identities not listed (Please specify)
  - xi. 'NR' = Prefer Not to Answer
- b. zgtvbdso_description varchar2(30)
- c. zgtvbdso_eff_date
- d. zgtvbdso_nchg_date
- e. zgtvbdso_activity_date
- f. Zgtvbdso_id number(19)
- g. Zgtvbdso_version number(19)

recognition of layered identities). On the other hand, the level of protection and access limitations imposed on the new tables sparked discussion among the project members about a secret table or a "behind-the-scenes table that no one can see" (Domain Expert, Interview), with frequent references made to the SOGI data being "locked down" (Domain Expert, Interview), hampering the recognition of layered identities in wider StateU systems. Data semantics for recognition were, in

Stelmaszak, M., Wagner, E. L., & DuPont, N. N. (2024). Recognition in personal data: data warping, recognition concessions, and social justice. *MIS Quarterly*, 48(4), 1611–1636.

“The characteristics of digital data we discussed above make data like glass fragments in a kaleidoscope: they can be always recomposed to form a different rendering of a phenomenon. The kaleidoscopic nature of data is not only about the perspectives that we as observers bring to bear on the world, but also about refraction. Digital data allow to create specific versions of reality while obscuring or diminishing others”

Stelmaszak, M., Aaltonen, A., & Lyytinen, K. (2026). Looking through the digital data kaleidoscope: Introduction to the Research Handbook on Digital Data. In: Aaltonen, A., Stelmaszak, M., & Lyytinen, K. (2026). (Eds.) R*esearch Handbook on Digital Data: Interdisciplinary Perspectives*. Edward Elgar Publishing.

Data studies

≈ research that takes data as its object of inquiry

Waht is data studies? Some provisional answers…

Data studies is not a i) theory, ii) methodology, or a iii) standalone research program.

Currently, data studies is perhaps best described as an emerging narrative, group of scholars with similar interests, series of loosely connected events… **with more to come!**

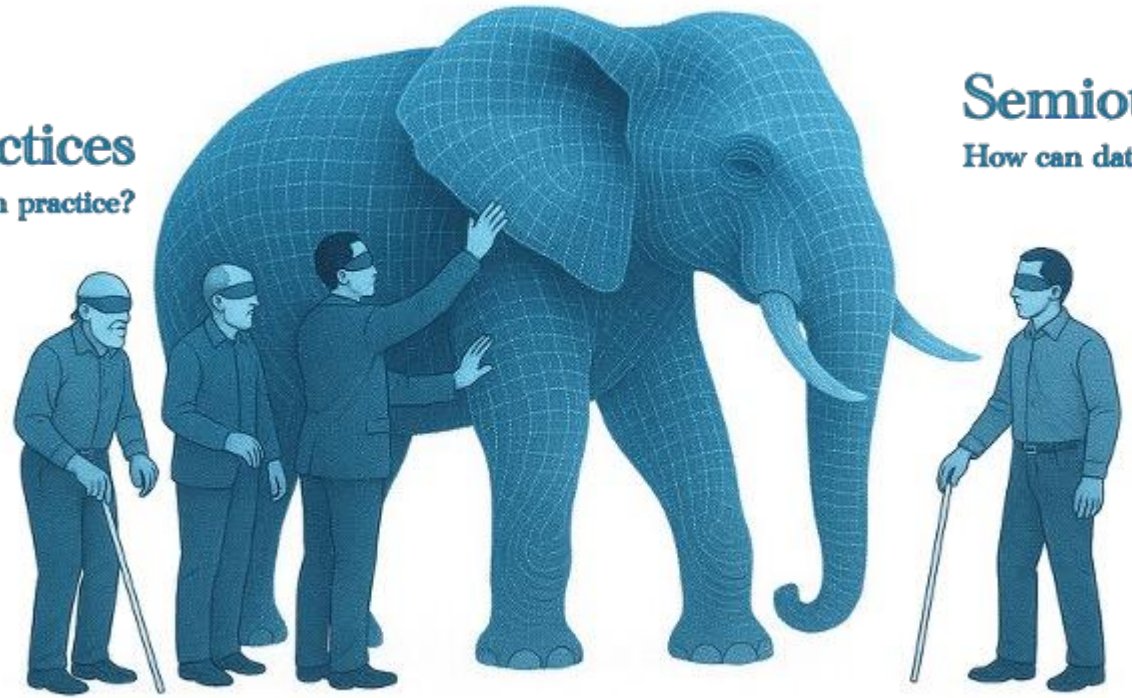
But, is data studies really needed?

Practices

What makes data 'data' in practice?

Semiotics

How can data convey meanings?



Engineering

How data work as technology?

The "standard view"

Records are data
…taken as 'ready at hand' facts

Implicitly assumed, for instance, in the use of data analytics and quantitative analysis…

Data are representations
"How can we model more truthful data?"

Most of the traditional conceptual modeling and data quality research make this assumption.

Data as artifacts
"What are data and how do they work?"

Focus on the human-made aspects of data artifacts and their implications: data design, data-related innovation, semiotic analyzes of data…

What makes data possible?

'Conditions of possibility' for something to appear as data: relational perspectives, data work, data infrastructures and infrastructuring…

Where is the value of data studies?

Oppotunities—we will not run out of things to study

Data and innovation. Data-driven innovation, data-based innovation, data innovation…

Data design and new kinds of data. We do not know well enough how arrangements that are used to produce certain data are configured in practice.

Data governance (ownership, sharing, privacy, etc.). There are significant opportunities in data sharing in different domains, but much of the data remain locked in organizations because we do not have the (governance) solutions for sharing them.

Data justice and fairness. The more important data becomes in our private and professional lives, the more they will reflect and potentially amplify biases and unfairness of society.

Data poisoning. When important decision are made by or influenced by data (e.g., through AI), there will be attempts to poison the training data.

Data and energy consumption. The consumption of electricity is becoming a limiting factor in digital and data-based innovation.

Temporalities in data. What were once data, may or should not be data anymore. How to sustain and expire data?

De-datification. Once something has been datified, it will be difficult to de-datify. What should not be datified? When do we collect too much data? How can we de-datify phenomena?

Semantic ambivalence. The increasingly ambivalent referential nature of data—a world made in synthetic data?

Sensitivity to ontologically different data. E.g., personal data should not be designed similarly to data about inanimate things, or data in quantum computing.

…and many others. Data in quantum computing, data as both rival and non-rival asset…

data studies bibliography

Over 200 unique visitors per week!

→ 151 journal papers, 127 conference papers, 27 books, 21 teaching cases

→ Blog with regular guest posts

→ scholars@datastudies.net mailing list (over 300 subscribers)

Aaltonen, A., Alaimo, C., Parmiggiani, E., Stelmaszak, M., Jarvenpaa, S. L., Kallinikos, J., & Monteiro, E. (2023). What is missing from research on data in information systems? Insights from the Inaugural Workshop on Data Research. *Communications of the Association for Information Systems,* 53(1), 475–490. https://doi.org/10.17705/1CAIS.05320

Xu, D., Stelmaszak, M., & Aaltonen, A. (2025). What is changing the game in data research? Insights from the "Innovating in Data-based Reality" professional development workshop. *Communications of the Association for Information Systems,* 56, 194–208. https://doi.org/10.17705/1CAIS.05608

SIG Data coming!

Aaltonen, A., Stelmaszak, M., & Lyytinen, K. Eds. (2026). *Research Handbook on Digital Data: Interdisciplinary Perspectives.* Edward Elgar.

Thank you!
Q&A

Let us know your name and what you're currently working on in the chat!

Aleksi: aaaltone@stevens.edu
Marta: mstelmaszak@isenberg.umass.edu