



Introduction to Web Science

1. Introduction

Jun Sun

`junsun@uni-koblenz.de`

Slides by Matthias Thimm, Steffen Staab



What is Web Science?

Computer
Science

Astronomy

Web
Science

Science
about
Computers

Telescope
Science

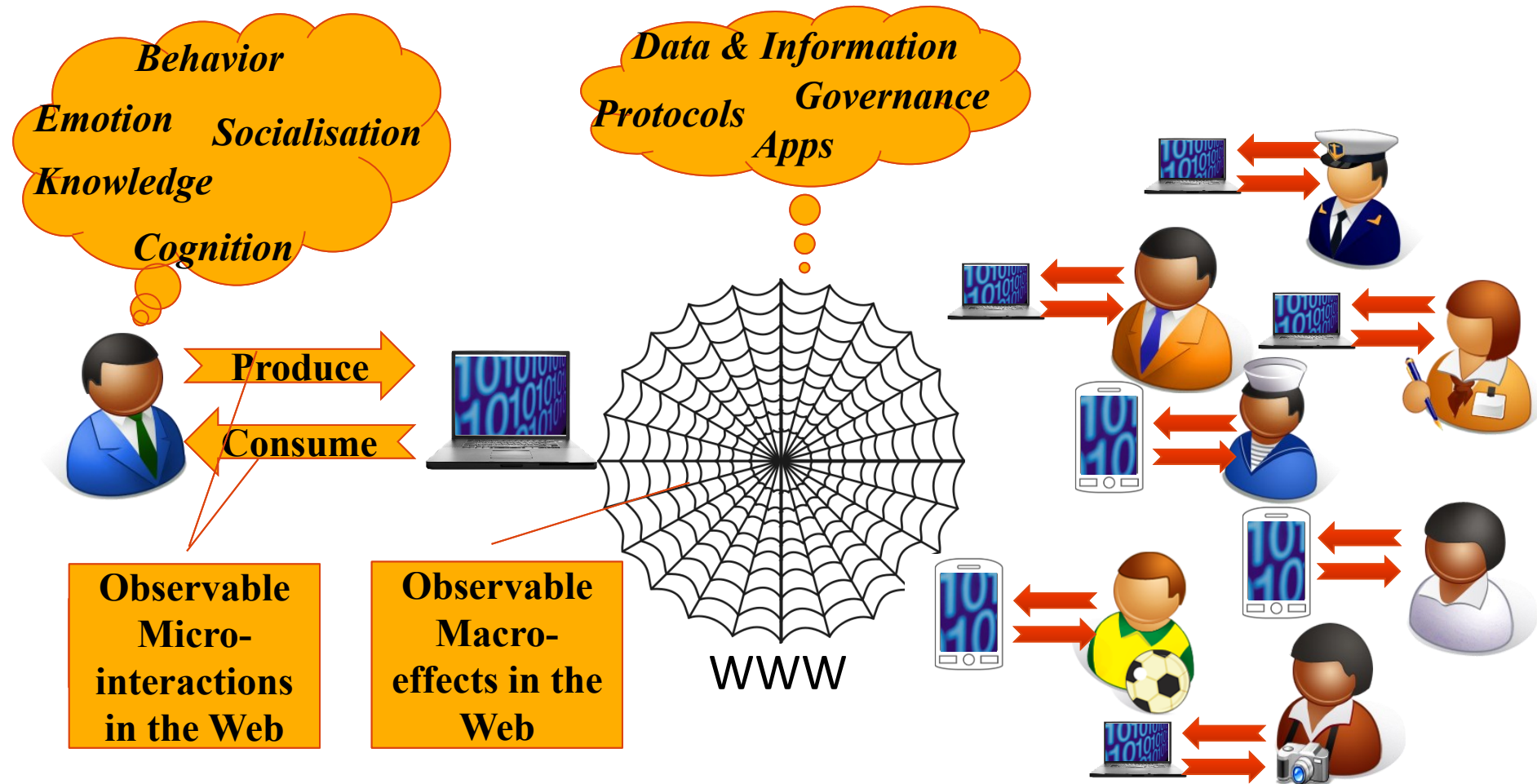
Science
about the
Web

Definition of Web Science

Web science is an emerging interdisciplinary field concerned with the study of **large-scale socio-technical systems**, particularly the **World Wide Web**. It considers the relationship between **people** and **technology**, the ways that society and technology co-constitute one another and the impact of this co-constitution on broader society. Web Science combines research from disciplines as diverse as **sociology**, **computer science**, **economics**, and **mathematics**.

Wikipedia, 2020-10-28

Web Science



Agenda

- What is Web Science?
- What is the Web?
 - Aspects of the Web at Large
- How to investigate the Web?
 - Observing the Web
 - An example using the architecture: bias in the Web
 - How to model aspects of the Web
- What is the past and the future of the Web?

What is the Web?

The Web as a Device

Software

- Browsers
 - IE, Firefox, Chrome,...
- Web Servers
 - Apache, Tomcat,...
- Content Management and Data Delivery
 - Wordpress, drupal, databases...
- Search Engines
- ...

Standards

- Uniform Resource Locator (URL)
- HyperText Transfer Protocol (http)
- HyperText Markup Language (html)
- Domain name service
- ...+ many more

The Web as Content

For human consumption
(primarily)

Text, Hypertext

Images

Video

Audio

Multimedia

Interaction (Games...)

Braille

Mathematics

For machine consumption
(primarily)

Metadata

Data

Ontologies

The Web and its Stakeholders

People

- Citizens
- Customers
- Leisure seekers
- Workers
- Software developers

Internet providers

- Landline
- Mobile
- Nested providers
(internet cafe...)

Platform operators

- Shops
- News
- Web 2.0
- Payment
- Advertisement networks
- Trust centers

Government

- Police
- Military
- Secret service
- Law
- Citizen services
- Administration
- Politics

The Web as a Process

Governing

- Standards processes
 - W3C
 - Internet Engineering Task Force (IETF)
 - RFC
- Domain name registration
- Internet routing
 - E.g. „great Chinese firewall“

Regulation

- Legal
 - copyright
 - where enforced
 - hate speech
 - ...
- Private
 - E.g. Facebook, Instagram ... Pictures

Observing the Web as a Medium and Mirror of (anti-)social Practices

- (Self-)Expression
- Dark Web
 - Crime
 - Gold farming
 - Violence
 - Pornography
 - Identity theft
- Sex lifes
 - Fetishes
 - prostitution
- Relationships
 - Breakups
 - Mobbing
 - Stalking
 - Advising
 - Counseling
 - Democracy
- Politics
 - Agenda setting
 - Discussions
 - Evaluation

The Web and Artificial Intelligence

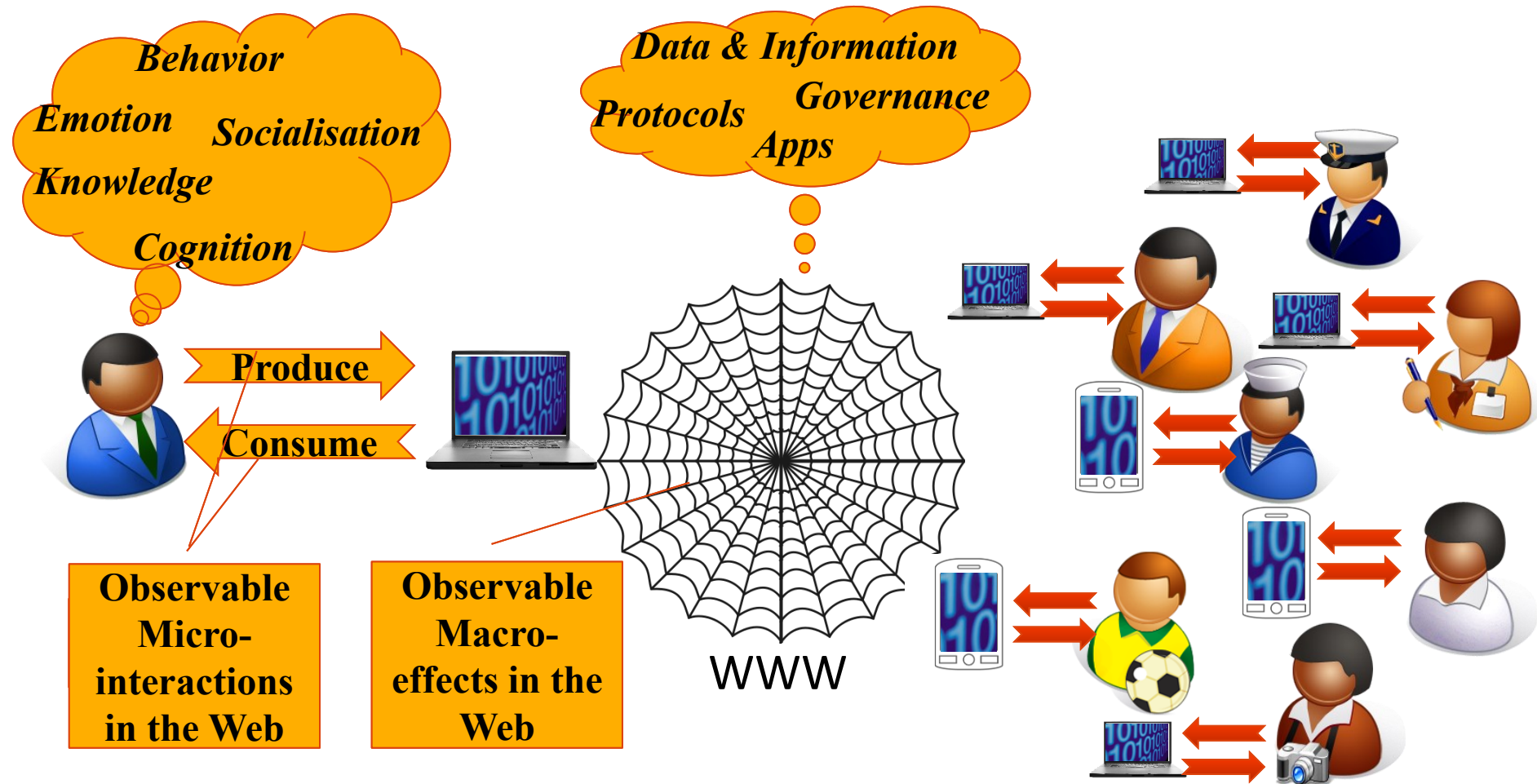
The Web as a precondition for Artificial Intelligence

- Human like recognition based on Web data
- Crowdsourcing for AI

Artificial Intelligence on the Web

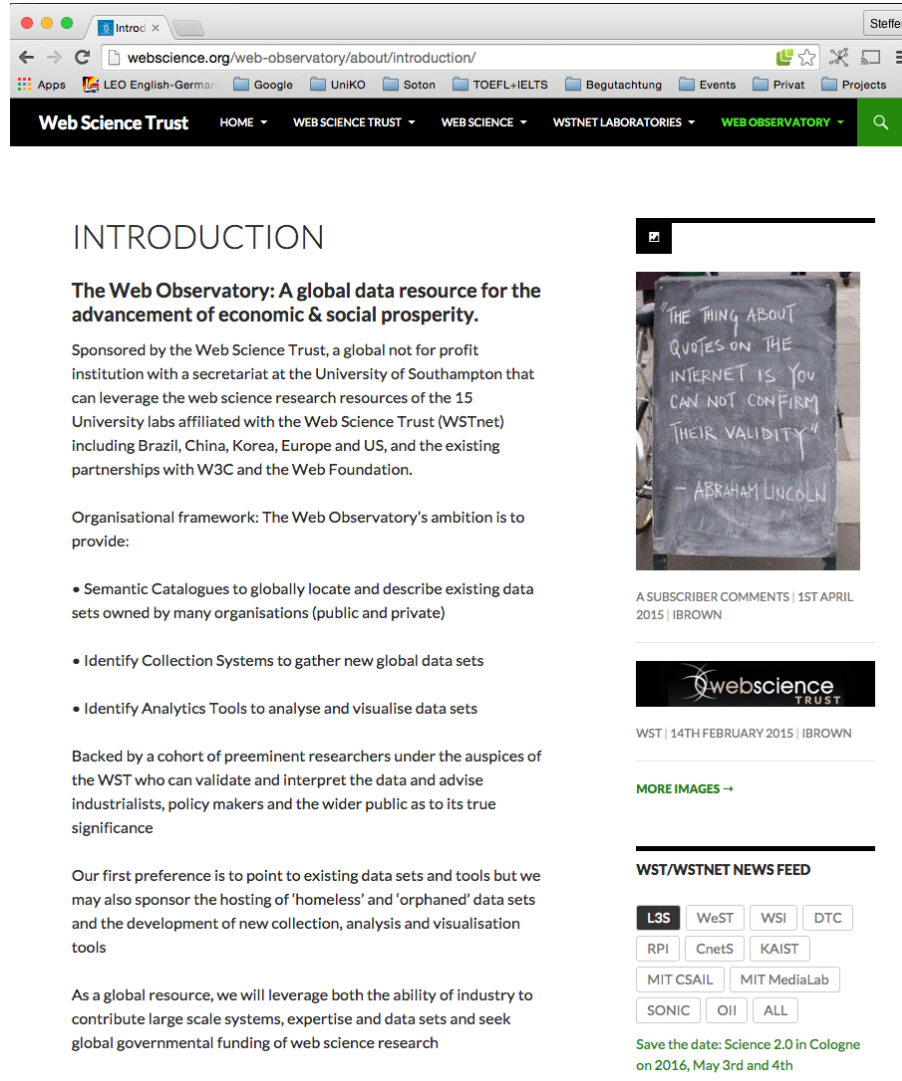
- Chatbots
- Social bots
- Intelligent search
- ...

Web Science: Discipline or Transdisciplinary Endeavour?



How to investigate the Web?

Web Observatories



The screenshot shows a web browser window with the URL `webscience.org/web-observatory/about/introduction/`. The page title is "Web Science Trust" and the navigation bar includes links for HOME, WEB SCIENCE TRUST, WEB SCIENCE, WSTNET LABORATORIES, and WEB OBSERVATORY. The main content area is titled "INTRODUCTION" and describes the Web Observatory as a global data resource for the advancement of economic and social prosperity. It mentions sponsorship by the Web Science Trust, a global not-for-profit institution at the University of Southampton, and lists 15 affiliated University labs. The page also outlines the organisational framework and provides a list of services: Semantic Catalogues, Identify Collection Systems, and Identify Analytics Tools. A quote by Abraham Lincoln is featured: "THE THING ABOUT QUOTES ON THE INTERNET IS YOU CAN NOT CONFIRM THEIR VALIDITY". The page includes a subscriber comment from 1st April 2015, a logo for the Web Science Trust, and a news feed section with filters for L3S, WeST, WSI, DTC, RPI, CnetS, KAIST, MIT CSAIL, MIT MediaLab, SONIC, OII, and ALL. The footer mentions a date: Science 2.0 in Cologne on 2016, May 3rd and 4th.

INTRODUCTION

The Web Observatory: A global data resource for the advancement of economic & social prosperity.

Sponsored by the Web Science Trust, a global not for profit institution with a secretariat at the University of Southampton that can leverage the web science research resources of the 15 University labs affiliated with the Web Science Trust (WSTnet) including Brazil, China, Korea, Europe and US, and the existing partnerships with W3C and the Web Foundation.

Organisational framework: The Web Observatory's ambition is to provide:

- Semantic Catalogues to globally locate and describe existing data sets owned by many organisations (public and private)
- Identify Collection Systems to gather new global data sets
- Identify Analytics Tools to analyse and visualise data sets

Backed by a cohort of preeminent researchers under the auspices of the WST who can validate and interpret the data and advise industrialists, policy makers and the wider public as to its true significance

Our first preference is to point to existing data sets and tools but we may also sponsor the hosting of 'homeless' and 'orphaned' data sets and the development of new collection, analysis and visualisation tools

As a global resource, we will leverage both the ability of industry to contribute large scale systems, expertise and data sets and seek global governmental funding of web science research

THE THING ABOUT QUOTES ON THE INTERNET IS YOU CAN NOT CONFIRM THEIR VALIDITY"
- ABRAHAM LINCOLN

A SUBSCRIBER COMMENTS | 1ST APRIL 2015 | IBROWN

webscience TRUST

WST | 14TH FEBRUARY 2015 | IBROWN

MORE IMAGES →

WST/WSTNET NEWS FEED

L3S WeST WSI DTC
RPI CnetS KAIST
MIT CSAIL MIT MediaLab
SONIC OII ALL

Save the date: Science 2.0 in Cologne on 2016, May 3rd and 4th

Why to observe?

- Understanding
 - Collecting
 - Describing
 - Analyzing
 - Modeling
 - Predicting

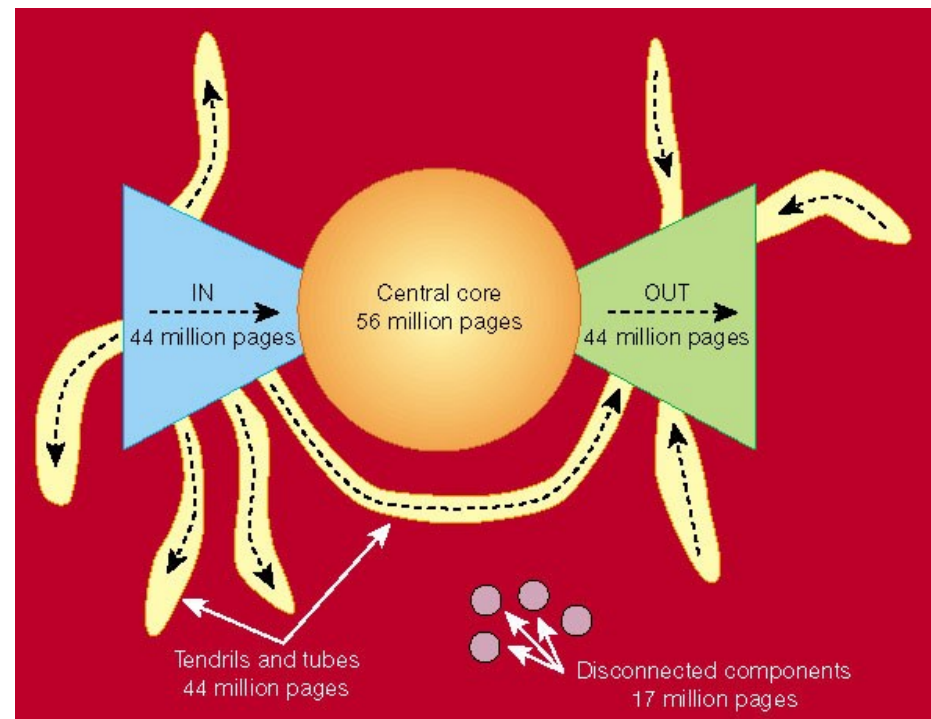
Challenges – Data Collection Issues

Legal and/or Ethical

- Crawling
 - May be disallowed by provider
- Usage logging
 - Privacy of individuals
- Reproducibility

Challenges – Data Collection Issues

- Crawling
 - What does it mean to crawl a heavily interactive site?
 - Incomplete data
 - Unreachability
 - Time outs



The web is a bow tie. *Nature* **405**, 113 (2000). <https://doi.org/10.1038/35012155>

Challenges – Data Collection Issues

- Crawling
 - What does it mean to crawl a heavily interactive site?
 - Incomplete data
 - **Where to start?**
 - We cannot observe everything!
 - Even just for data size!
 - What appear to be most fruitful starting points?

Challenges – Data Collection Issues

- Crawling
 - What does it mean to crawl a heavily interactive site?
 - Incomplete data
 - Where to start?
 - **Where to stop?**
 - Each crawl is a view
 - Twitter
 - » Tweet
 - URL
 - Web Page
 - Subweb
 - » Followers
 - Followers' Followers
 - ...

Challenges – Data Publishing Issues

Legal and/or Ethical Example Issues

- AOL query log (2006)
- Netflix challenge (2009)
- Twitter
 - Collecting, but no sharing
 - SocialSensor project

Challenges – Data Publishing Issues

Technical/Modelling issues

- Generic format, e.g. RDF
- Format ready for digestion by a certain software, e.g. for Matlab processing
- Openness to other data
 - E.g. references to DBPedia/Wikipedia
- Accuracy of publishing
 - <http://me.org> showed „...“
 - <http://me.org> showed „...“@2013-05-01:0900CEST
 - <http://me.org> showed „...“@2013-05-01:0900CEST called from IP 193.99.144.85 using browser...version...history...

Sharing Software

- Software
 - For crawling or usage logging
 - Rather than sharing the data, share the code for observing
- Example:
 - code for crawling Twitter in a certain way
- Issues
 - Limited repeatability
 - Disturbance liability („Störerhaftung“) – at least in DE
 - If you provide source code for crawling, e.g., Facebook, even if you do not crawl FB, FB can sue you

Example Topic: Bias

Bias in the Software

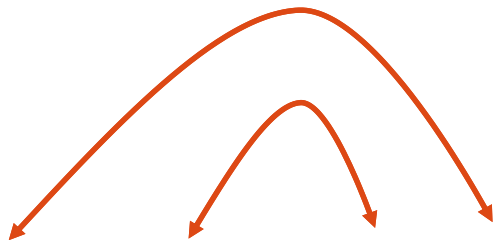
Search engines

- Categorizing people and animals
 - White vs black
 - http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0
- Job advertisements
 - Well-paid job not offered to females

Bias in Content/Data

Data protection laws suggest not to process sensitive data attributes like „sex“ or „ethnic“

correlated

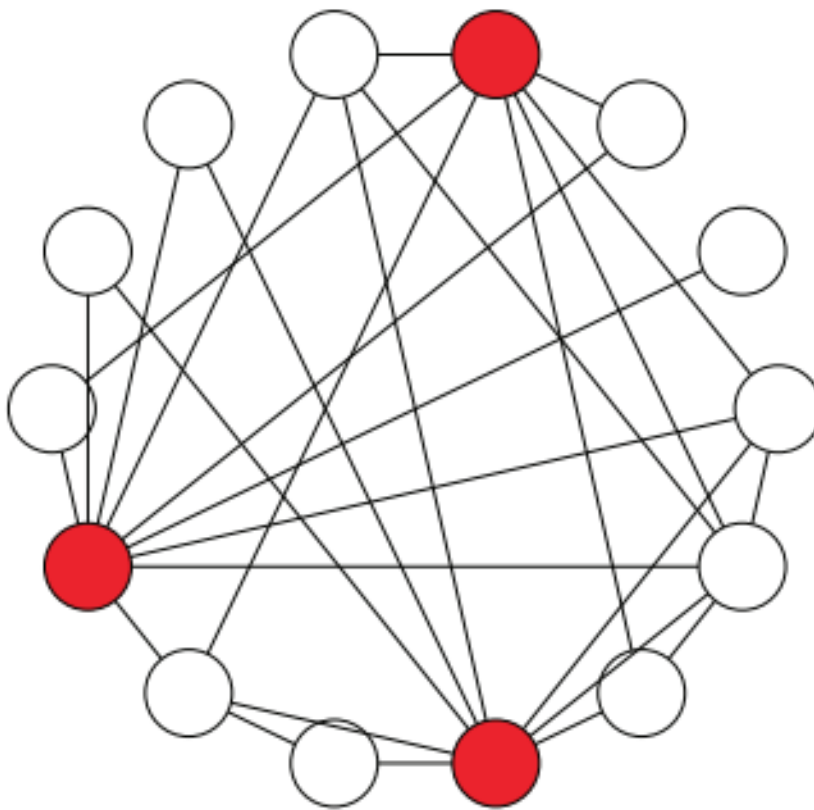


The diagram shows two red curved arrows originating from the 'Sex' and 'Ethnic' headers. One arrow points to the 'Credit' header, and the other points to the 'Hire' header, indicating a correlation between these sensitive attributes and the outcome variables.

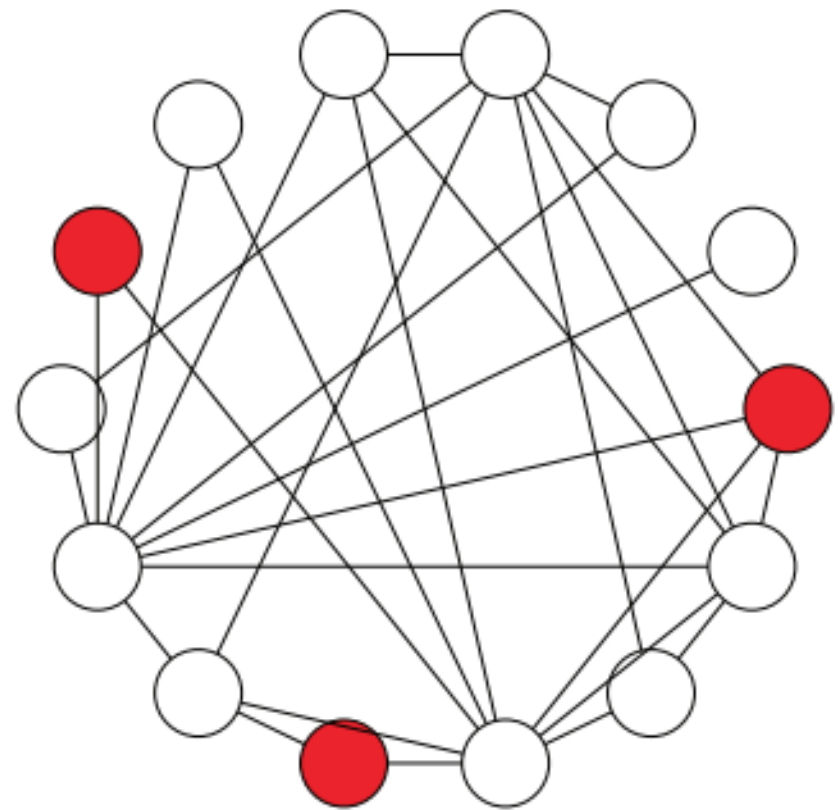
Credit	Hire	Sex	Ethnic	Zip	Height	
+	+							
+	-							
-	+							
+	+							
-	-							

Bias in Content: Social Networks

(Lerman et al 15)



(a)



(b)

Bias and Processes

Wikipedia

- Efforts to counter bias

Law

- E.g. UK equality act
- Protected characteristics:
 - Age, disability, gender, marriage, religion,...

Web Models

Web Models

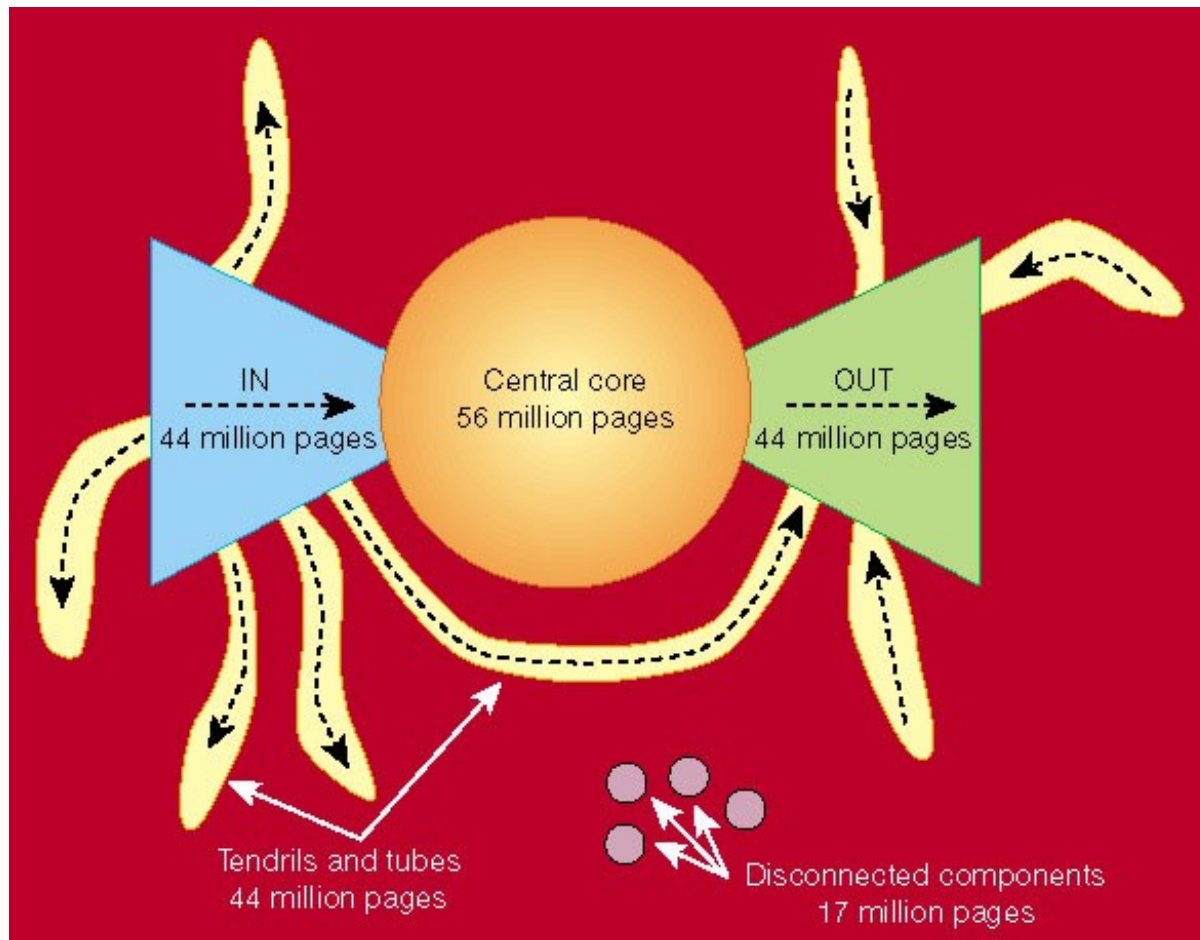
- Descriptive
 - Qualitative
 - Statistical
- Predictive
 - Modeling deterministic regularities
- Generative
 - Modeling non-deterministic principles
 - Liking a song
 - Creating a link

Descriptive Models

Example:

Bow Tie Structure of the Web

Bow-tie structure of the Web



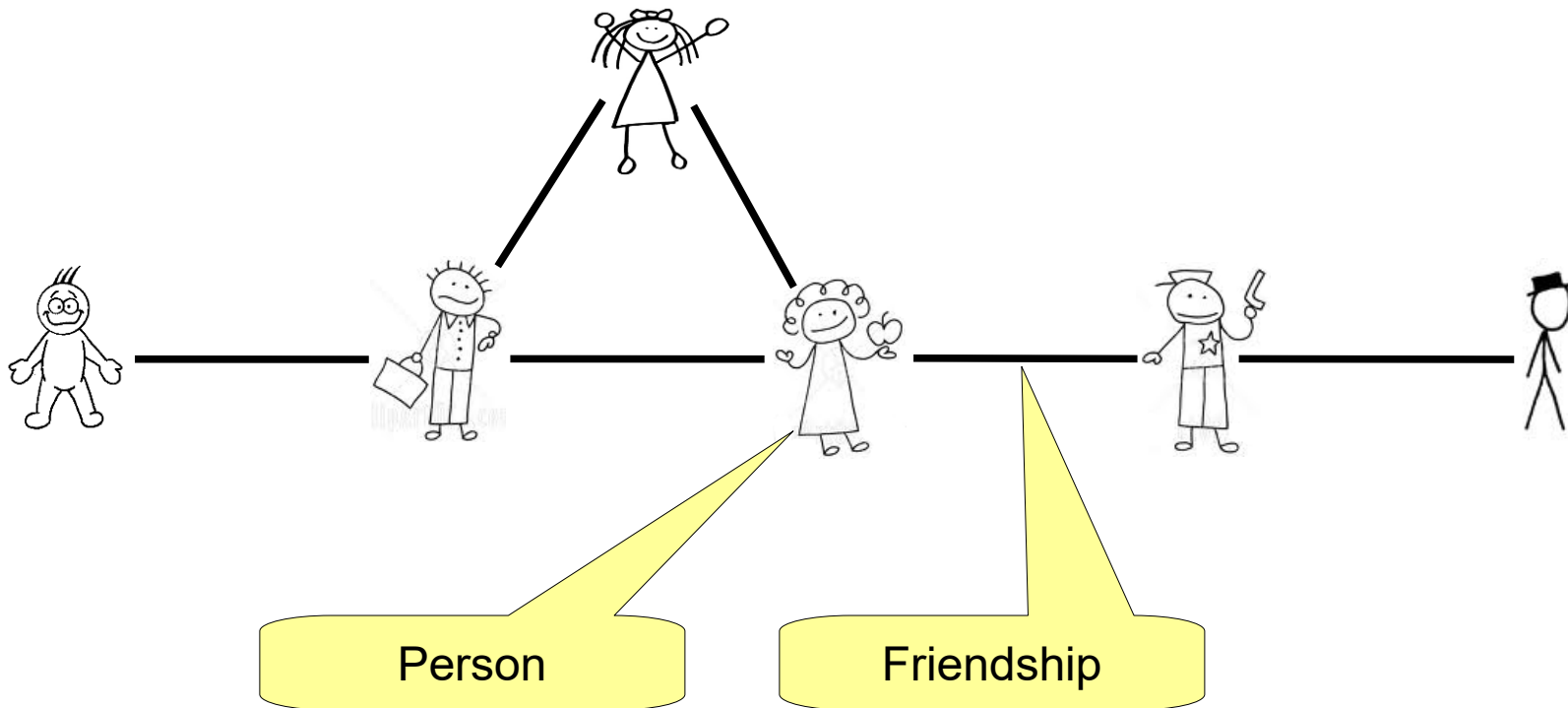
The web is a bow tie. *Nature* **405**, 113 (2000). <https://doi.org/10.1038/35012155>

Predictive Models

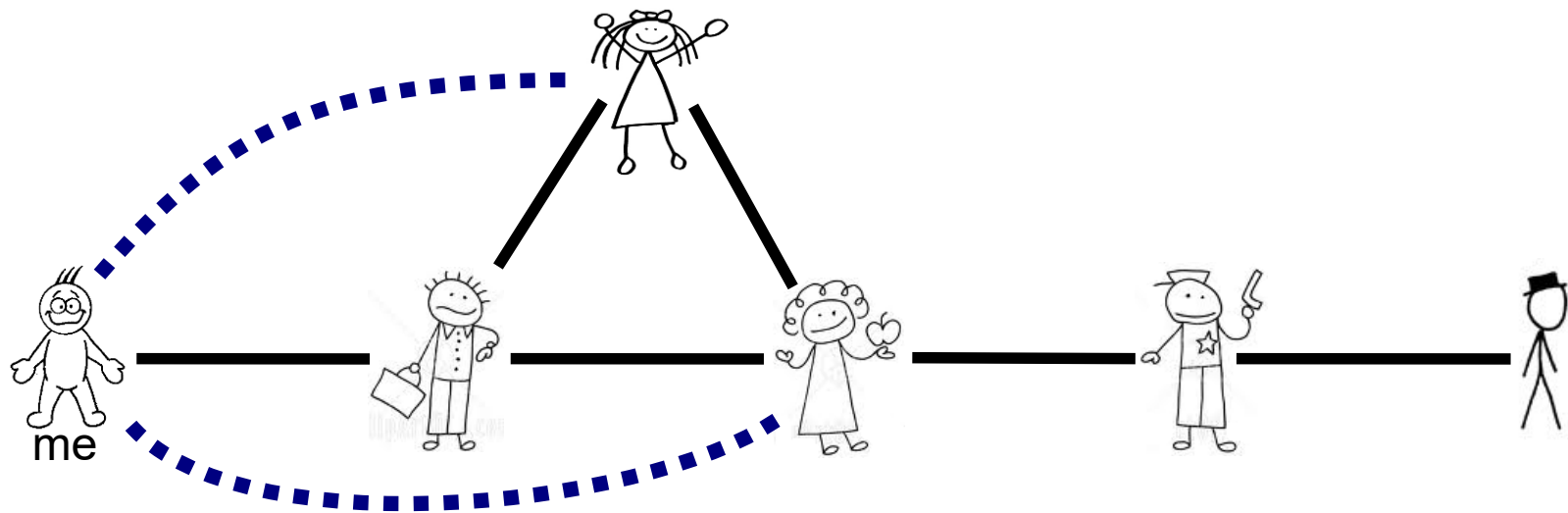
Example:

Link Prediction by Triangle Closing

Social Network



Recommender Systems

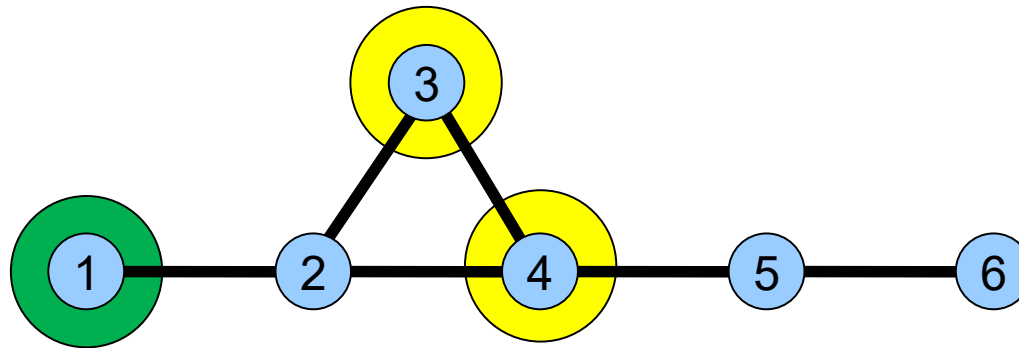


Predict who I will add as friend next

Standard algorithm: find friends-of-friends

Friend of a Friend

Count the number of ways a person can be found as the friend of a friend.



Generative Models

Example:

Link creation by the Barabasi-Albert model

Barabasi-Albert model (for social networks)

- We model the creation of a network as a dynamic process with the simplifying assumptions
 - at each time step a new user u enters the network
 - at the very same time step u creates links to a fixed number of already present users
 - existing users receive a link from u proportionally to the links they already have (rich-get-richer phenomenon)
- General question: does a generative model explain the structure of the whole network?

The Past and the Future Web

Pre-Web

- 1945 Vannevar Bush, „As we may think“, Memex
- 1962 Ted Nelson, Hypertext
- 1965 Wide area network
- 1968 Doug Engelbart, The mother of all demos
- 1972 Public Arpanet, E-mail
- 1974-82 Internet protocol TCP/IP
- 1978 Consumer information services and E-mail
- 1983 AOL, online service for games, communities...
- 1984 Domain name service

The World Wide Web

- 1989 Concept drafted by Tim Berners-Lee
- 1993 National Center for SuperComputing Applications launched Mosaic X
- 1994 First WWW conference
- 1994 W3C started at MIT
- Commercial websites began their proliferation
- Followed by local school/club/family sites
- The web exploded
 - 1994 – 3,2 million hosts and 3,000 websites
 - 1995 – 6,4 million hosts and 25,000 websites
 - 1997 – 19,5 million hosts and 1,2 million websites
 - January 2001 – 110 million hosts and 30 million websites

The World Wide Web

- 1994/1995 Amazon
- 1994/1995 Wiki
- 1995 AltaVista Search Engine
- 1995 Internet Explorer
- 1997-2001 Browser wars
- 1996-1998 XML recommendation
- 1998 Google
- 1999 First W3C recommendation on RDF (Semantic Web)
- 2001 Dot.com bubble bursts
- 2001 Wikipedia
- 2003/2004 Facebook
- 2004 Flickr
- 2005 YouTube

Concepts	Example Applications	
Web of Intelligences	Siri (2011), Echo, Alexa, Google Assistant	2011
Web of People	Physical transport service (Uber (2009), Lyft), accommodation service (AirBnB, Couchsurfing), online dating service	2009
Web of Things	Smart city, ambient intelligence, personal and public health information, personal and public transport information	2007
Web of Services	Cloud services, Digital transformation, Programmable Web (2005)	2005
Web of Data	User generated content applications (Facebook, Wikipedia (2001) and Wikidata,...), Linked open gov data	2001
Web of Documents	HTTP, HTML, XML, Browser (Mosaic 1993)	1993
Computer Networks/Internet	Document delivery (internet 1982), VOIP, Streaming	1982

Concepts	Delivered technical capabilities
Web of Intelligences	Cognitive capabilities, intelligent communication
Web of People	Identification and rating by/of people
Web of Things	Identification, linking, aggregation, monitoring and controlling of things
Web of Services	Identification, composition and calling of services
Web of Data	Identification, linking and retrieval of data
Web of Documents	Identification, linking and retrieval of documents
Computer Networks/Internet	Identification of and communication between computers

Concepts	Standards
Web of Intelligences	<i>No standards yet</i>
Web of People	<i>No mature standards yet</i>
Web of Things	<i>No mature standards yet</i>
Web of Services	REST, JSON, JSONLD
Web of Data	RDF, SPARQL
Web of Documents	HTTP, HTML, XML, AJAX
Computer networks/Internet	Internet, TCP/IP, Optical fibre, 5G

Conclusions

Summary

Accomplishments

- Web of Services
- Web of Data
- Web of Documents
- Computer networks/Internet

Future in the making

- Web of People
- Web of Things

How to identify and use?

How to observe?

How to resolve issues?

Thank you for your attention