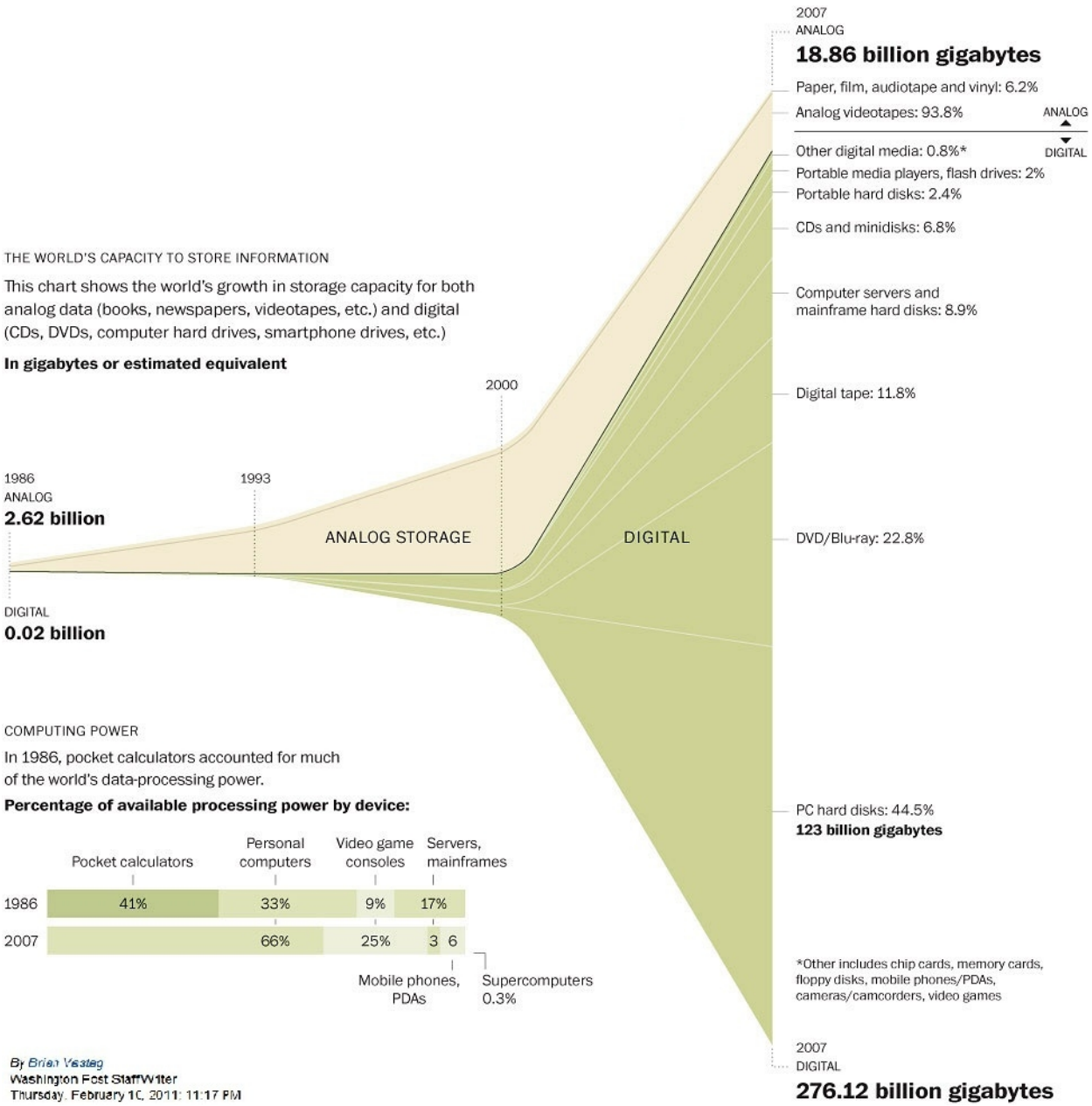
Methods for unstructured data

Introduction

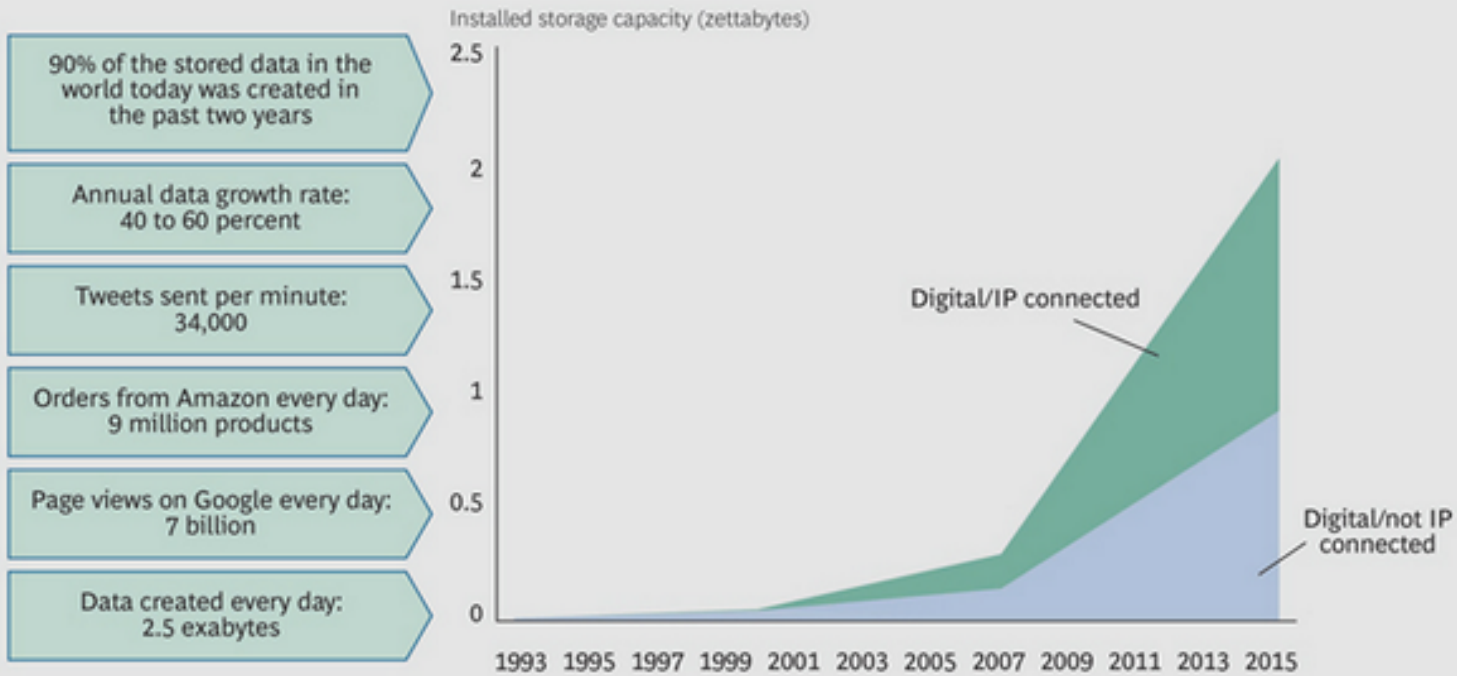
Helge Liebert

Worldwide data storage capacity



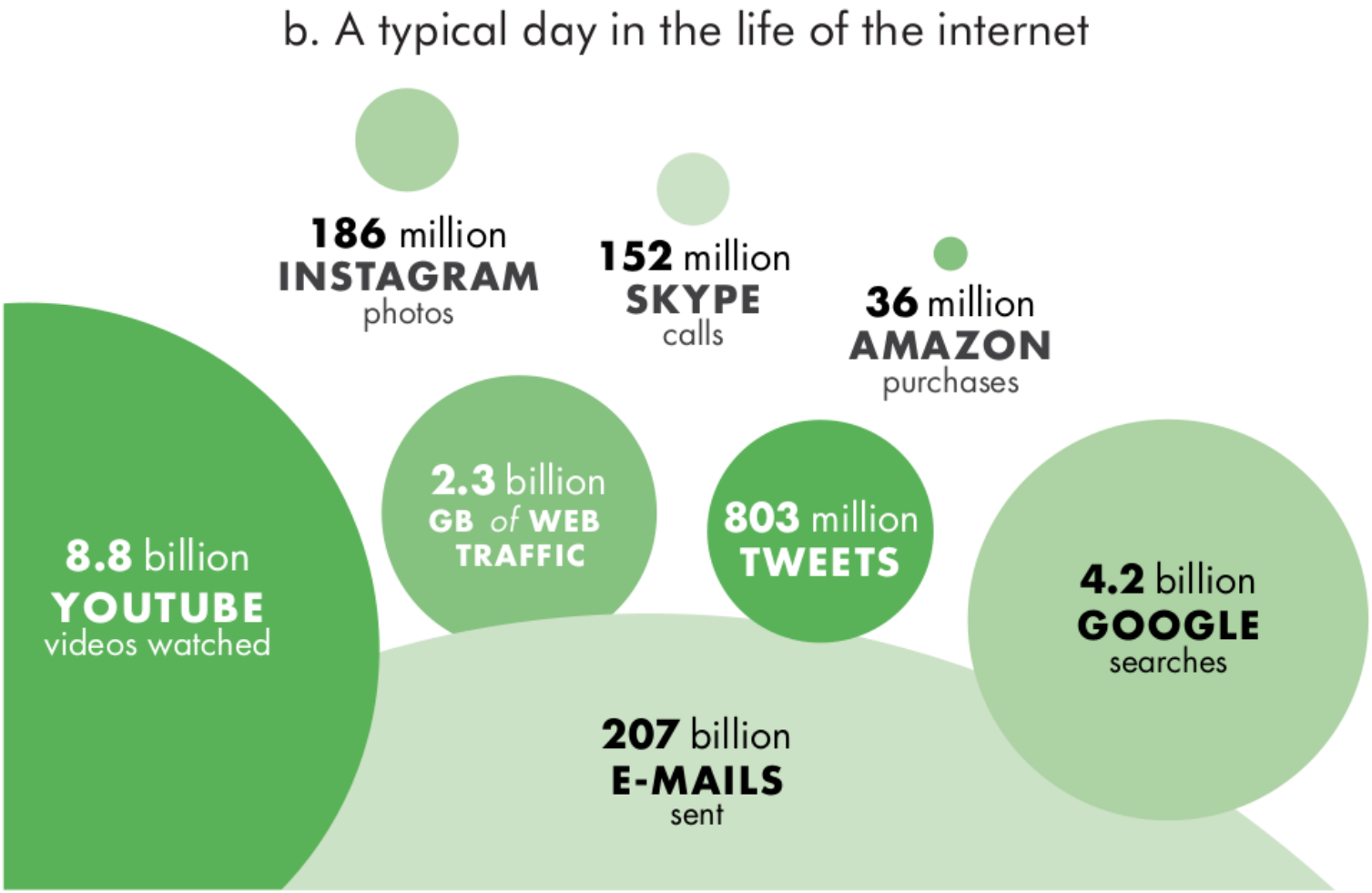
Data, then and now

EXHIBIT 2 | In 2015, More Than Half of All Data Will Have an IP Address



Sources: Martin Hilbert and Priscilla Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information," *Science*, February 2011; BCG estimates.

Data, then and now



Sources: World Development Indicators (World Bank, various years); WDR 2016 team; http://www.internetlivestats.com/one-second/ (as compiled on April 4, 2015). Data at http://bit.do/WDR2016-FigO_4.

Note: In panel a, for some years data for electricity are interpolated from available data. GB = gigabytes.

Introduction

- 90% of data today has been created in the last two years.
- 235 million emails sent per day.
- 3.3 million Facebook posts created every minute.
- 3.8 million Google searches performed each minute.
- 1.7 megabytes of new information created every second, per person.

Introduction

- 90% of data today has been created in the last two years.
- 235 million emails sent per day.
- 3.3 million Facebook posts created every minute.
- 3.8 million Google searches performed each minute.
- 1.7 megabytes of new information created every second, per person.

➥ An immense amount of data, new and old, is recorded as text.

➥ More generally, much of this data is unstructured.

Structured vs. unstructured

Structured data

- Adheres to a defined data model.
- Examples: Tables, spreadsheets, relational databases, ...

Unstructured data

- Does *not* adhere to a defined data model.
- Typically text-heavy.
- Examples: Text feeds, speech transcripts, audio, images ...

Structured vs. unstructured

Structured data

- Adheres to a defined data model.
- Examples: Tables, spreadsheets, relational databases, ...

Semi-structured data

- Does *not* adhere to a formal data model,
- ... *but* contains tags or semantic mark-up.
- Examples: JSON, XML, emails, tagged text, ...

Unstructured data

- Does *not* adhere to a defined data model.
- Typically text-heavy.
- Examples: Text feeds, speech transcripts, audio, images ...

Text as data

- Text differs from other, traditional forms of data.
- Text is inherently *unstructured* and *high-dimensional*.
- One of the major fields of application of machine learning methods.
- Fast-growing field. Many new techniques developed in industry.
- Recent applications in economics and other social sciences.

This lecture

This lecture covers techniques for unstructured data.

- Methods for wrangling data.

➥ When unstructured $\approx$ dirty (or differently structured).

This lecture

This lecture covers techniques for unstructured data.

- Methods for wrangling data.

➥ When unstructured $\approx$ dirty (or differently structured).

- Methods for analyzing data which are naturally unstructured.

➥ No rectangular (or graph) structure, no well-defined relations between data elements.

Focus points

Focus on three main points.

1. Processing and transforming un-/semi-structured data.
2. Representing inherently unstructured text data.
3. Analyzing text data and using models to discover structure. (Supervised and unsupervised learning.)

Outline

1. Introduction

Data management

2. Tools for scientific programming
3. Web scraping

Representation

4. Regular expressions and pattern matching
5. Representing text as data

n-gram modeling approaches

6. Supervised models for text data
7. Unsupervised models for text data

Information retrieval and distributional language models

8. Distributional models of meaning
9. Vector space representations

Assignment

Dates

Tuesday	11.10.2022	08.30-12.00	PC-Lab S18 HG.37
Wednesday	19.10.2022	08.15-12.00	PC-Lab S18 HG.37
Thursday	20.10.2022	12.15-18.00	PC-Lab S18 HG.37
Wednesday	26.10.2022	08.15-12.00	PC-Lab S18 HG.37
Thursday	27.10.2022	12.15-18.00	PC-Lab S18 HG.37
Thursday	03.11.2022	12.15-18.00	PC-Lab S18 HG.37

Technical requirements: Lab sessions

- All class material is available online:
  https://github.com/hliebert/course-unstructured-data.
- All material will run on the Windows computers in the lab.
- The lab materials can also be accessed online:
  Jupyter notebooks
  Rstudio server

- Feel free to set up your own computer.
- Installation of dependencies depends on OS (Windows, MacOS, Linux).
- Clone/download the course repository to get started.
- Please ask after class if you need help.

Programs

Minimal

- A browser.

Local: Core material

- R, plus Editor/GUI (RStudio, VScode with R plugin, Jupyter, Emacs+ESS, ...).
- Run the R install script provided with the class material to install the R package dependencies and the R Kernel for Jupyter notebooks.

Local: Additional material

- Jupyter notebooks. Install Anaconda (or its smaller miniconda version). You can also use pip to install Jupyter if you have Python installed.
- A shell (bash or zsh pre-installed on Linux or MacOS, bash via WSL or git bash on Windows).
- Git.

How to install them

- Linux: Use your distribution's package manager.
- Mac: Use installer packages or set up and use homebrew as a package manager (recommended).
- Windows: Use installer packages or look into scoop or chocolatey as native package managers for Windows. To get a Linux environment on Windows, install Windows Subsystem for Linux (WSL) (MS docs here).
- You can always use Linux in a Virtual Machine (VM). Use Virtual Box to run the VM. Build your own VM from a Linux install image (e.g. Ubuntu) or download a ready-to-use VM from osboxes.org.

Assignment

1. Web scraping assignment (20%)
2. Text analysis and prediction assignment (80%)

- Deadline: 19.12.2022.
- Course is graded.
- More details during the course of the lecture.

Primary references

- The course covers relatively broad and diverse topics, no single reference. Seminal references in the slides.
- Primary and secondary references below.
- Hastie et al. and Jurafsky & Martin books are available online (use newest 3rd edition draft of J&M).


Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as Data. *Journal of Economic Literature* 57(3), 535–574. DOI: 10/gf7rd5.


Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Ed. by R. Tibshirani and J. H. (H. Friedman. New York.


Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, N.J: Pearson Prentice Hall.


Shotts, W. E. (2019). *The Linux Command Line: A Complete Introduction*. Second edition. San Francisco: No Starch Press.

Secondary references

- Reference material, applied or introductory text books.


Baumer, B., D. Kaplan, and N. Horton (2017). *Modern Data Science with R*. CRC.


Casella, G. and R. L. Berger (2001). *Statistical Inference*. Second. Duxbury Press.


Chacon, S. and B. Straub (2014). *Pro Git*. Apress.


James, G., D. Witten, T. Hastie, and R. Tibshirani (2015). *An Introduction to Statistical Learning with Applications in R*. Springer.


Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press.


Mitchell, R. E. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web*. Second edition. Sebastopol, CA: O'Reilly Media.


Munzert, S. (2014). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Chichester, West Sussex, United Kingdom: Wiley.


Silge, J. and D. Robinson (2017). *Text Mining with R: A Tidy Approach*. First edition. Boston: O'Reilly.

Secondary references


Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.


Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York, NY: Springer.