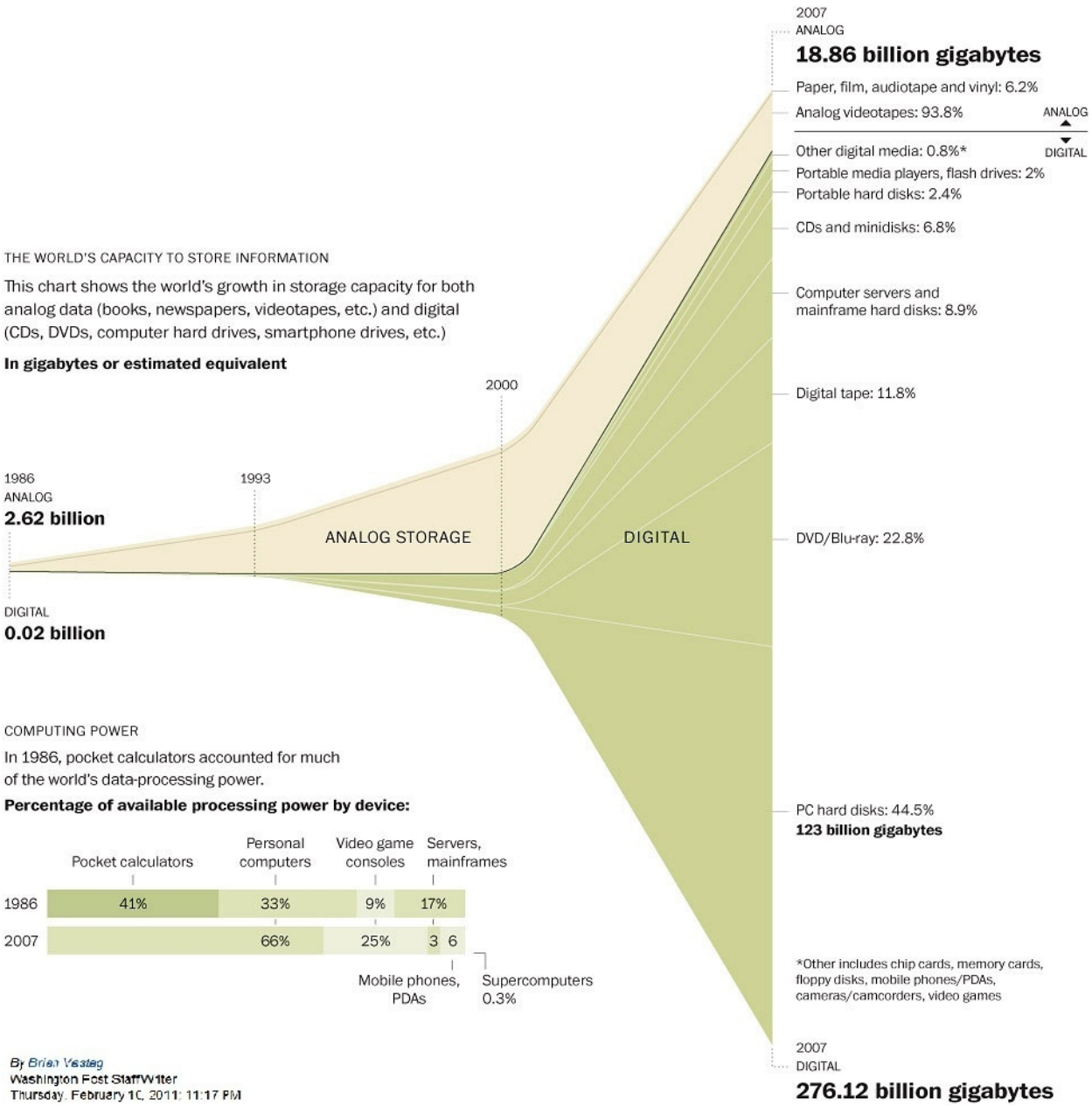
# Text Analysis in R

## Introduction

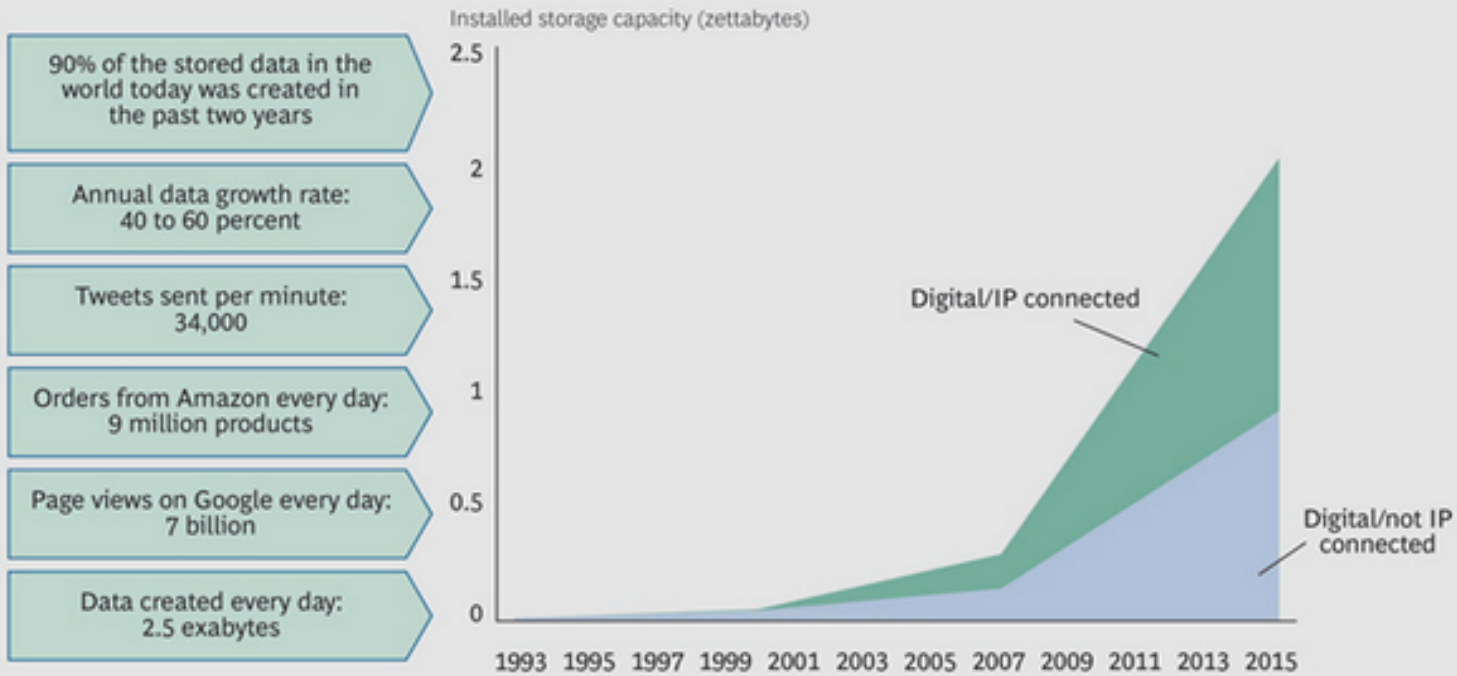
Helge Liebert

# Worldwide data storage capacity



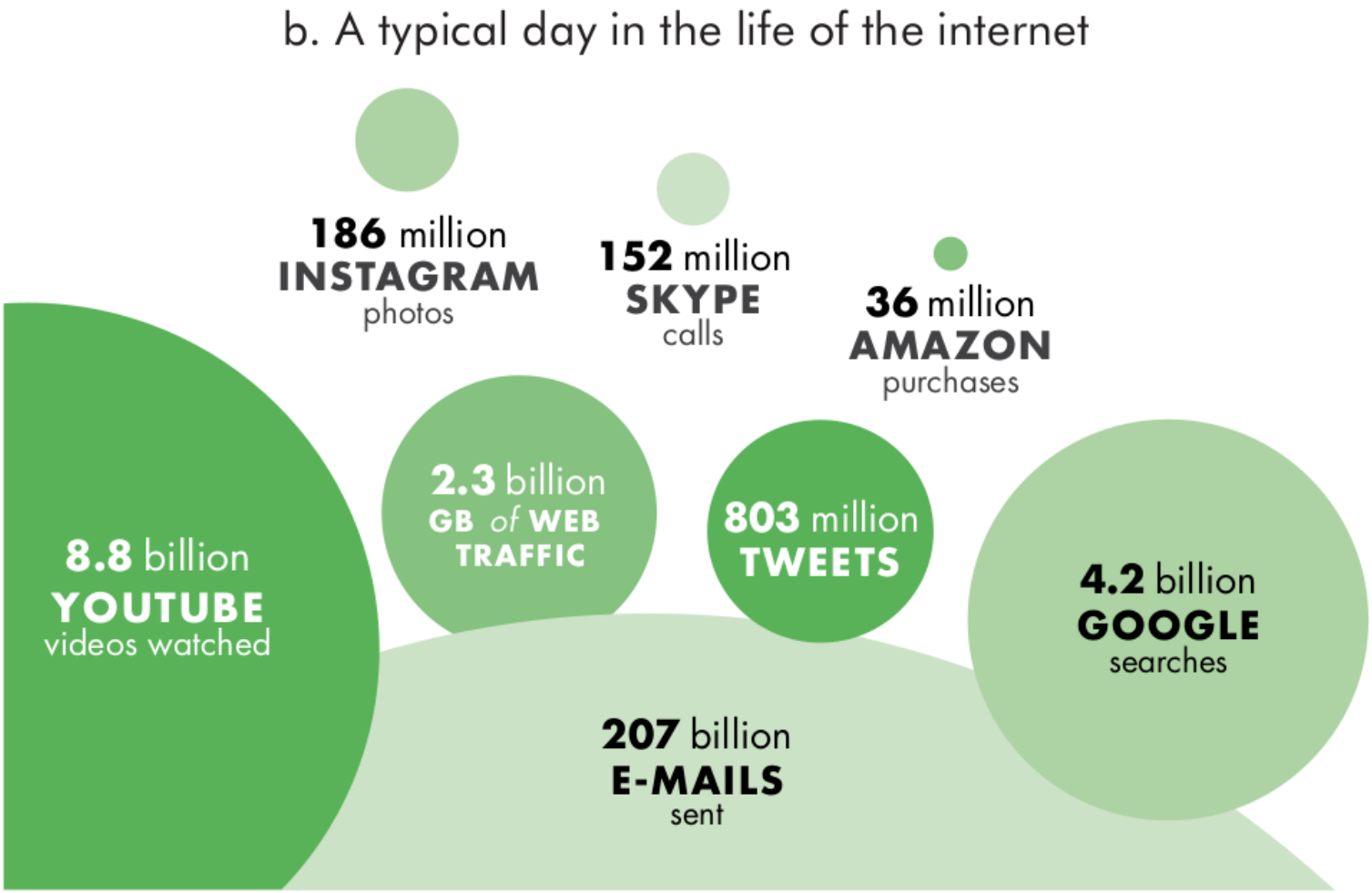
# Data, then and now

**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

**Representation**

2. Regular expressions and pattern matching
3. Representing text as data

**n-gram modeling approaches**

4. Supervised models for text data
5. Unsupervised models for text data

**Information retrieval and distributional language models**

6. Distributional models of meaning
7. Vector space representations

**Assignment**

# Dates

| | | | |
|---|---|---|---|
| Wednesday | 15.02.2022 | 08.30-16.30 | C130 |
| Thursday | 16.02.2022 | 08.30-16.30 | C130 |
| Friday | 17.02.2022 | 08.30-16.30 | C130 |

# Schedule

| | |
|---|---|
| green!1008.30-10.00 | Lecture |
| 10.00-10.30 | Break |
| green!1010.30-12.00 | Lecture |
| 12.00-13.30 | Lunch |
| green!1013.30-15.00 | Lecture |
| 15.00-15.30 | Break |
| green!1015.30-16.30 | Lecture |

# Technical requirements

- All class material is available online:
  https://github.com/hliebert/course-text-analysis-in-r.
- The lab materials can be accessed online:
  Jupyter notebooks
  Rstudio server
- Feel free to run the lab material locally on your own computer. Clone or download the course repository to get started.
- It will run on Windows, Mac or Linux (if the dependencies are satisfied).

# Programs

### Minimal

- A browser.

### Local

- R.
- Editor or 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.

### Optional

- Jupyter notebooks. Install Anaconda (or its smaller miniconda version). On Linux, you can also use python pip.

# Assignment

1. Independent project (100%)

- Deadline: 31.03.2022.
- 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.
- Textbooks are available online.
- Use latest 3rd edition draft of Jurafsky & Martin.
- For general ML/Stats, consult Hastie & Tibshirani or Murphy.



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.


Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press. URL: probml.ai.

# Secondary references

- Reference material, further reading, 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.


Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies 37. San Rafael: Morgan & Claypool Publishers.


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.


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

# Secondary references



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


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.