CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

INTRODUCTION TO DATA SCIENCE AND AI

DAT405, 2019-2020, READING PERIOD 1

Course organisation

- Part I: Introduction to data science (3 weeks)
  - Graham Kemp (kemp@chalmers.se), Computer Science and Engineering
- Part II: Statistical methods in data science and AI (2 weeks)
  - Marina Axelson-Fisk (marinaa@chalmers.se), Mathematical Sciences
- Part III: Introduction to AI (3 weeks)
  - Ashkan Panahi (ashkanp@chalmers.se), Computer Science and Engineering

Teaching assistant:

- Emilio Jorge (emilio.jorge@chalmers.se), Computer Science and Engineering

Examination form

- The examination is through **weekly assignments**, executed in student pairs.
- All assignments need to be passed in order to pass the course.
- Some exercises will only have a pass/fail grade, while others will be graded 3, 4, 5 (or fail).
- The final course grade will be an aggregate of the combined efforts.
- Deadline for each week's assignment will be on **Monday at noon** (12:00) the week after.

Student representatives

- Abdullah Awad
- Marcus Forsberg
- Emma Petersson Svensson
- Mattias Westerberg

CHALMERS
UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG

MODULE 1:
INTRODUCTION TO DATA SCIENCE AND PYTHON

DAT405, 2019-2020, READING PERIOD 1

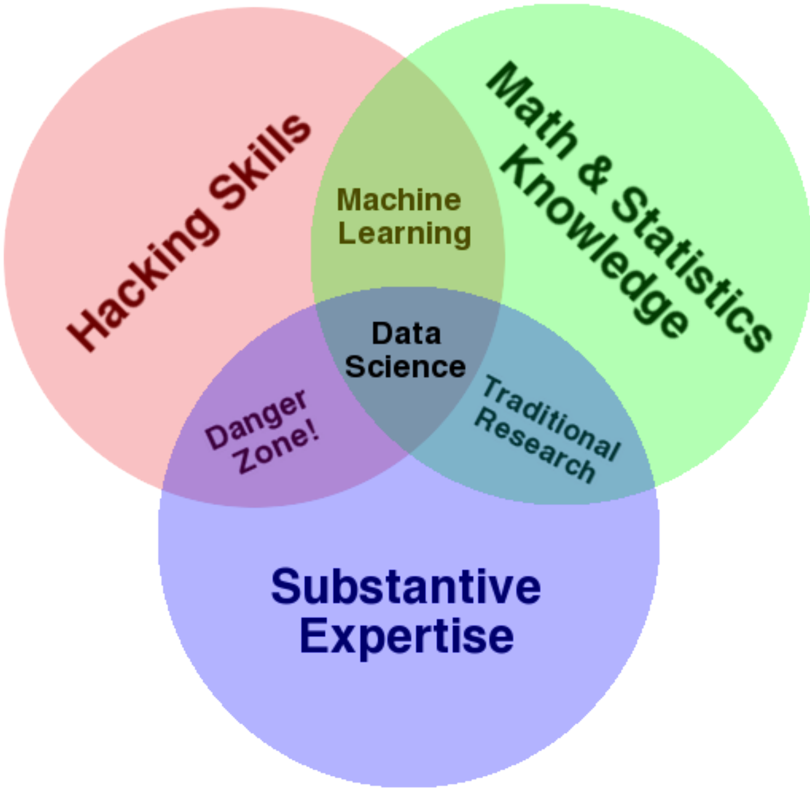
Introducing Data Science

Data Science is concerned with extracting meaning from (big) data.

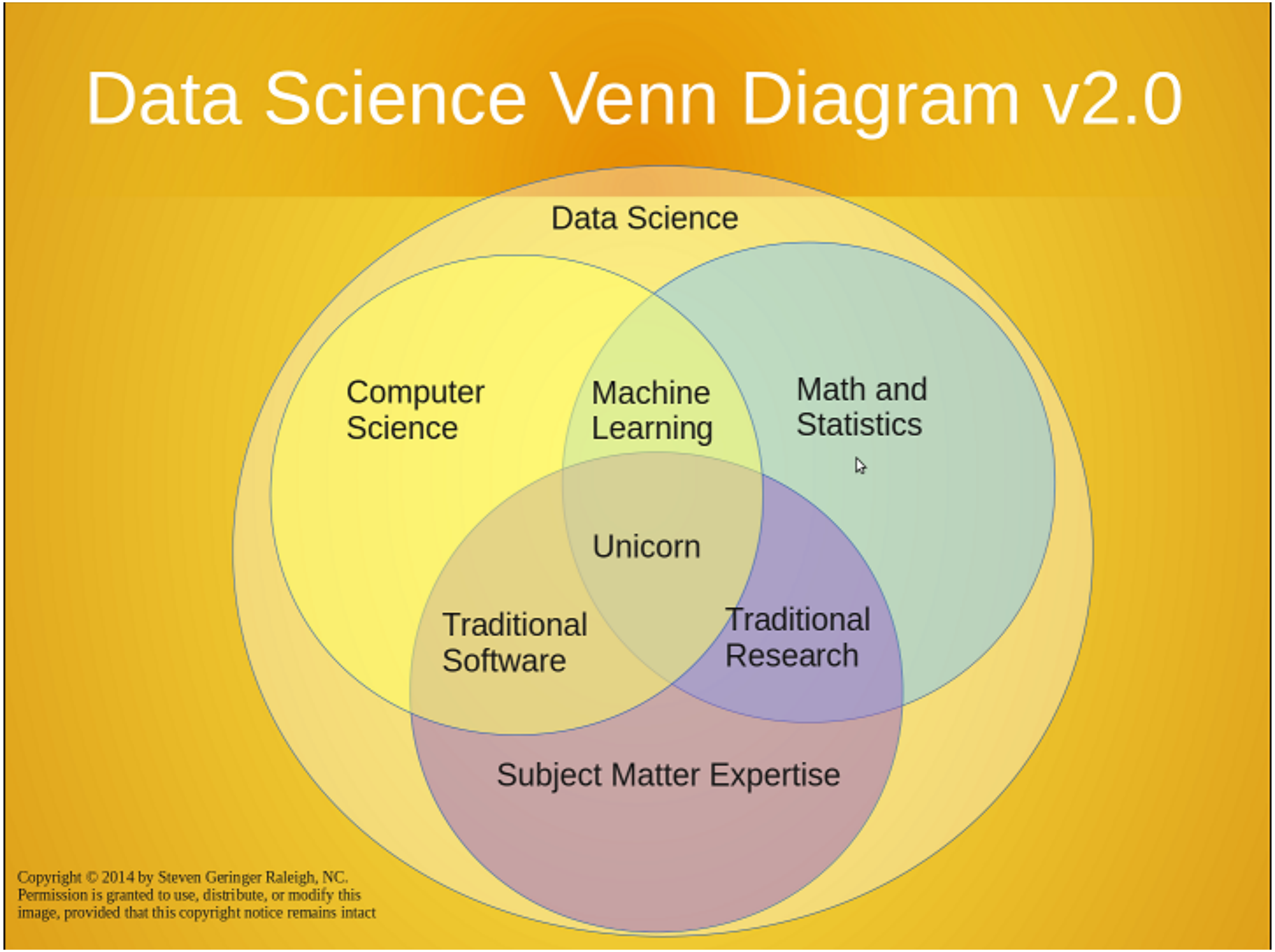
Central topics within Data Science include:

- data mining
- machine learning
- databases
- the application of data science methods in natural sciences, life sciences, humanities and social sciences, as well as in industry and society.

The Data Science Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

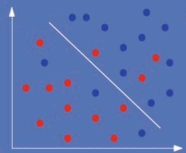
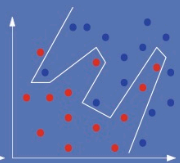
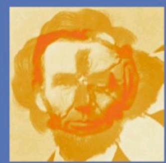
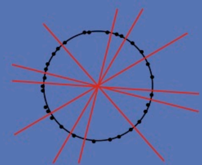
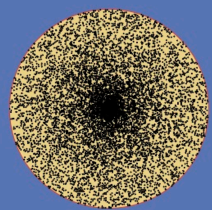
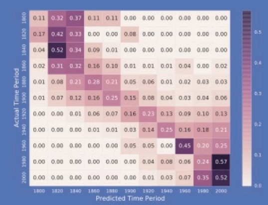
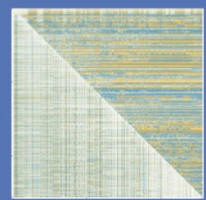

Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise
Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this
image, provided that this copyright notice remains intact

Applied Data Science

- Translating between problem and method domain for stakeholders
- Understanding data collection process and implications on results
- Understanding consequences of method choices
- *Communication is a central aspect*
- Expertise in Application Domain necessary
  - interpreting results
  - avoiding wrong conclusions

TEXTS IN COMPUTER SCIENCE

THE Data Science Design MANUAL

Steven S. Skiena

Springer

Application success stories

Case Study:

Influences in English Literature

Large-scale literature analysis

- 4357 novels
- 150 Years (average of 29 books per year)
- British (73%), Irish (5%), and American (22%)
- Male (55%), Female (36%), and Anonymous (9%)
- 1875 unique authors (2.32 books per author)

http://www.matthewjockers.net/slides-etc/ Matthew Jockers

Author	*Title*	Distance
Dickens, Charles	*A Tale of Two Cities*	0.000000
Kirkland, Caroline Matilda	*The Fountain and the Bottle*	1.361071
Milman, Edward Augustus	*Arthur Conway; or, Scenes in the Tropics*	1.385395
Liddell Charles Francis	*Hi*	
Dickens, Charles	*Th*	
Armstrong, Francis Claudius	*Th*	
Spofford, Harriet Elizabeth Prescott	*Sir*	
Fay, Theodore Sedgwick	*No Tin*	
Shillaber Benjamin Penhallow	*Kn Tex*	
Dickens, Charles	*Ba*	
Paulding, James Kirke	*Ch*	

Author	*Title*	Distance
Austen, Jane	*Pride and Prejudice*	0.000000
Austen, Jane	*Emma*	1.260236
Austen, Jane	*Sense and Sensibility*	1.268725
Austen, Jane	*Mansfield Park*	1.421373
Austen, Jane	*Northanger Abbey*	1.600394
Austen, Jane	*Persuasion*	1.673071
Gaskell, Elizabeth	*Ruth*	1.716687
Craik, Dinah Maria	*Olive*	1.745832
Church A. B. Mrs.	*Greymore a Story of Country Life*	1.747513
Grant, Louisa	*Charles Stanley*	1.765758
Tainsh, Edward Campbell	*One Maiden Only*	1.767951

http://www.matthewjockers.net/slides-etc/

Matthew Jockers

Q: what to do with it?

Q: identify influental writers?

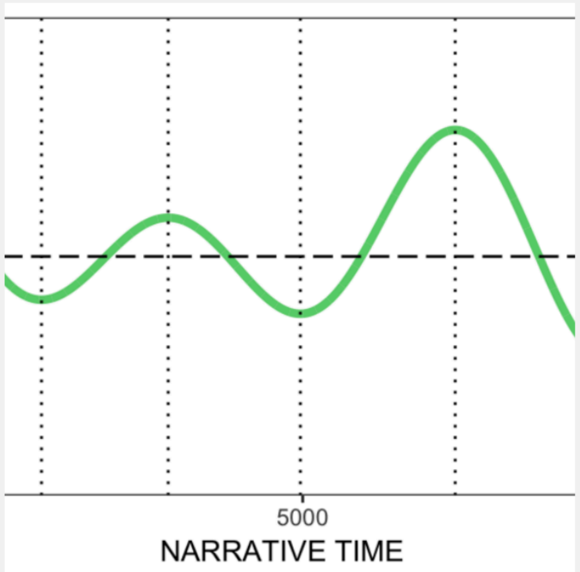
Author	*Title*	Distance
Dickens, Charles	*A Tale of Two Cities*	0.000000
Kirkland, Caroline Matilda	*The Fountain and the Bottle*	1.361071
Milman, Edward Augustus	*Arthur Conway; or, Scenes in the Tropics*	1.385395
Liddell Charles Francis	*Hic*	
Dickens, Charles	*The*	
Armstrong, Francis Claudius	*The*	
Spofford, Harriet Elizabeth Prescott	*Sir*	
Fay, Theodore Sedgwick	*No Tin*	
Shillaber Benjamin Penhallow	*Kn Tex*	
Dickens, Charles	*Ba*	
Paulding, James Kirke	*Ch*	

Author	*Title*	Distance	
Austen, Jane	*Pride and Prejudice*	0.000000	1813
Austen, Jane	*Emma*	1.260236	
Austen, Jane	*Sense and Sensibility*	1.268725	
Austen, Jane	*Mansfield Park*	1.421373	
Austen, Jane	*Northanger Abbey*	1.600394	
Austen, Jane	*Persuasion*	1.673071	
Gaskell, Elizabeth	*Ruth*	1.716687	1853
Craik, Dinah Maria	*Olive*	1.745832	1850
Church A. B. Mrs.	*Greymore a Story of Country Life*	1.747513	1860
Grant, Louisa	*Charles Stanley*	1.765758	1854
Tainsh, Edward Campbell	*One Maiden Only*	1.767951	1870

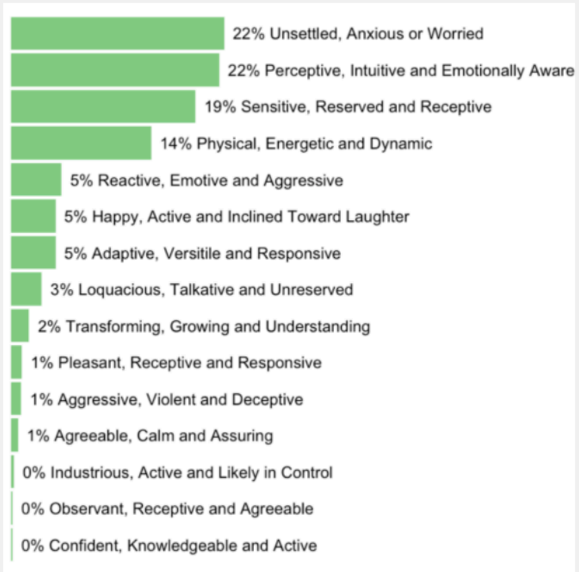
http://www.matthewjockers.net/slides-etc/

Matthew Jockers

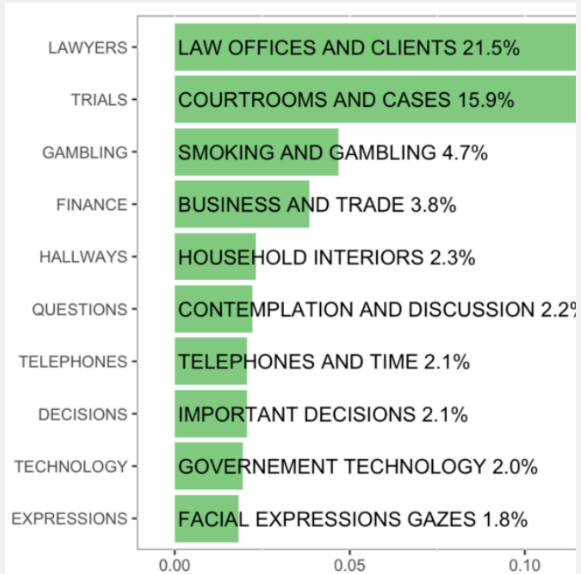
Manuscript analysis – over 3000 data points



Plot Shape



Character Personalities



Major Themes

http://www.archerjockers.com/home/

Case Study:

Society and policy

GLOBAL PULSE

ABOUT
PROJECTS
LABS
NEWS
CHALLENGES
PRIVACY
PARTNERSHIPS
RESOURCES
CONTACT
HOME

Projects

Welcome to the repository of Global Pulse's projects. Find out more about collaborative research, prototypes and experiments analyzing digital data to support global development and humanitarian action.

Using Call Detail Records To Understand Refugee Integration In Turkey

Exploring The Effects Of Extremist Violence On Online Hate Speech

Qatalog - An Analysis Tool For Insights Into The SDGs

Understanding Perceptions Of Migrants And Refugees With Social Media

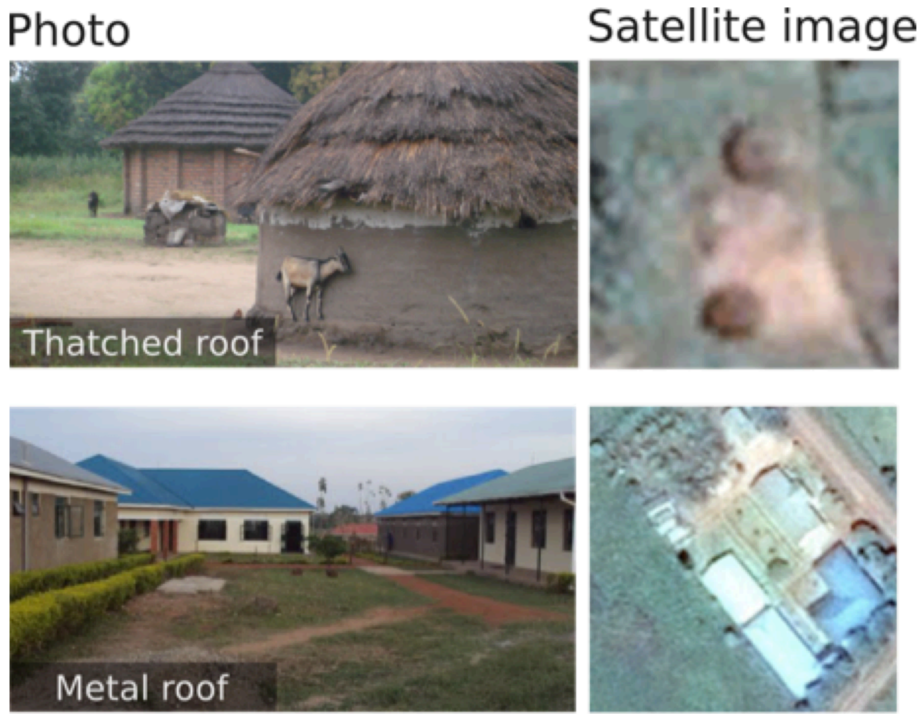
BROWSE BY LAB

Jakarta
Kampala
New York

BROWSE BY PROGRAMME

Climate & Resilience
Data Privacy & Protection
Economic Well-being
Food & Agriculture
Gender
Humanitarian Action
Public Health
Real-time Evaluation
The Sustainable Development Goals (SDGs)

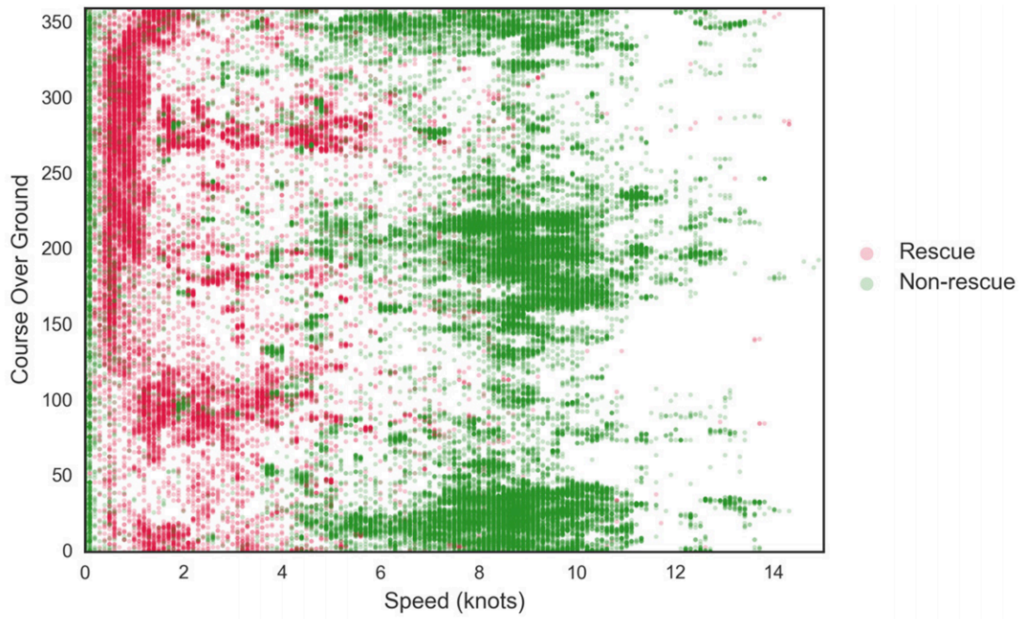
Measuring poverty with roof counting



https://www.unglobalpulse.org/projects/measuring-poverty-machine-roof-counting

Rescue patterns in the Mediterranean

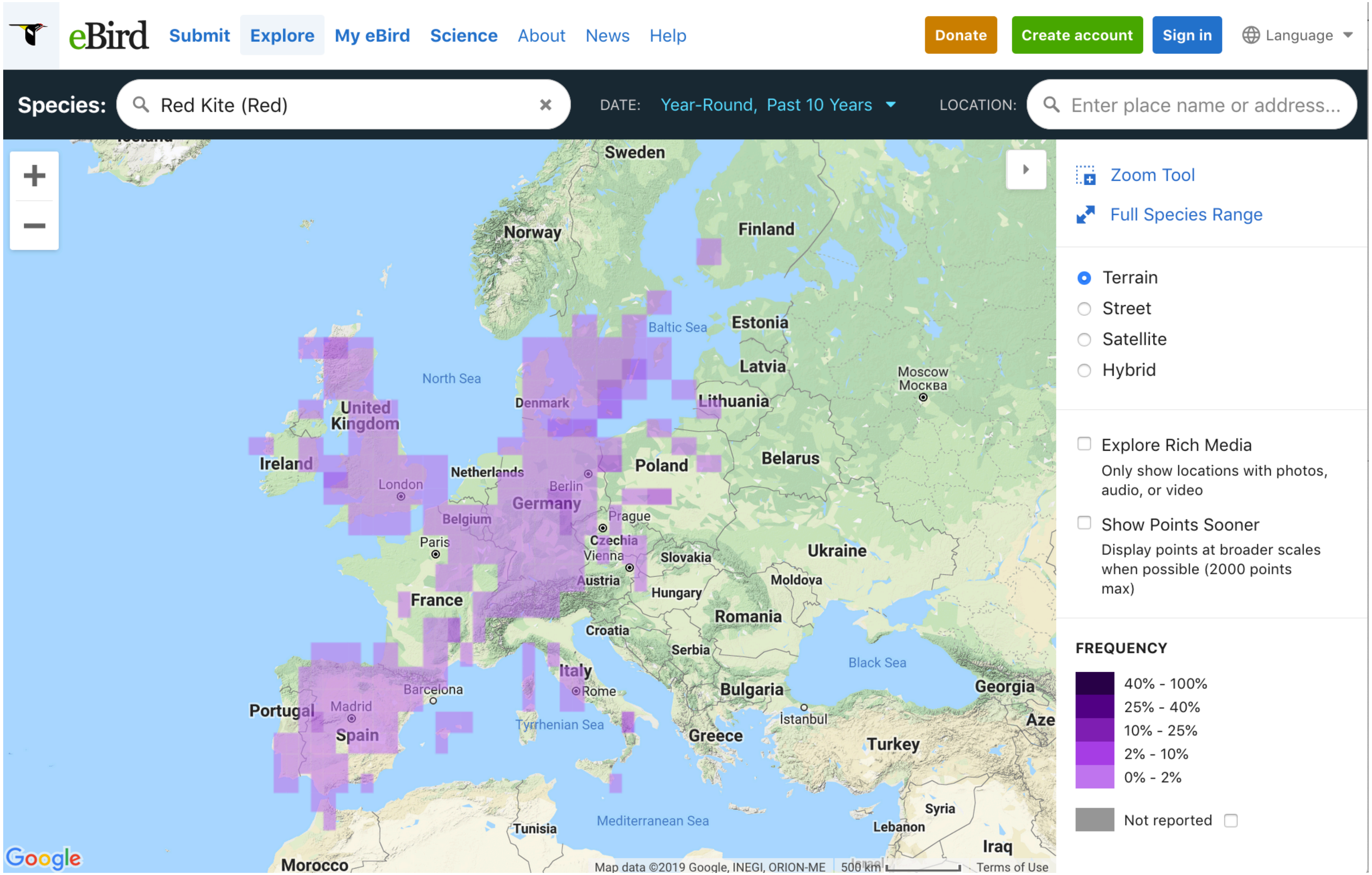
- Automatic Information System (AIS) - vessels regularly broadcast information, including their identifier, vessel type, latitude and longitude, speed, course and destination.
- speed and course appear to be good predictors of whether a vessel is conducting a rescue operation



https://www.unglobalpulse.org/projects/using-big-data-study-rescue-patterns-mediterranean

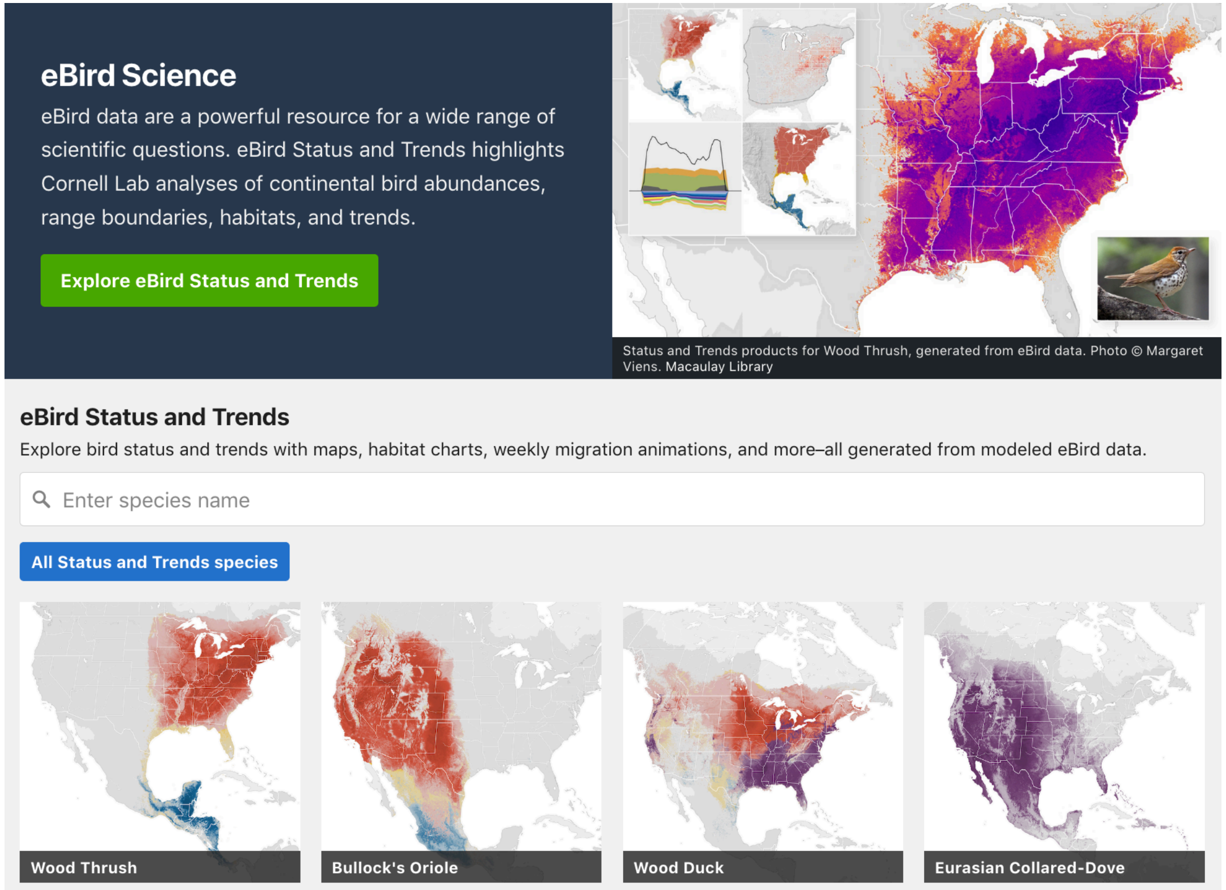
Case Study:

Ecology


eBird
Submit
Explore
My eBird
Science
About
News
Help
Donate
Create account
Sign in
Language
Species:
Red Kite (Red)
DATE:
Year-Round, Past 10 Years
LOCATION:
Enter place name or address...
Zoom Tool
Full Species Range
Terrain
Street
Satellite
Hybrid
Explore Rich Media
Only show locations with photos, audio, or video
Show Points Sooner
Display points at broader scales when possible (2000 points max)
FREQUENCY
40% - 100%
25% - 40%
10% - 25%
2% - 10%
0% - 2%
Not reported
Sweden
Norway
Finland
Estonia
Latvia
Lithuania
Baltic Sea
North Sea
Denmark
United Kingdom
Ireland
London
Netherlands
Berlin
Germany
Poland
Belarus
Moscow
Москва
Belgium
Prague
Czechia
Paris
Vienna
Slovakia
Ukraine
Austria
Moldova
Hungary
France
Romania
Croatia
Serbia
Black Sea
Georgia
Italy
Rome
Barcelona
Madrid
Portugal
Spain
Bulgaria
Istanbul
Tyrrhenian Sea
Greece
Turkey
Syria
Lebanon
Iraq
Tunisia
Mediterranean Sea
Morocco
Map data ©2019 Google, INEGI, ORION-ME
500 km
Terms of Use

eBird

- Quantified Bird Watching
- Bird watcher as "sensors"
- Citizen Science



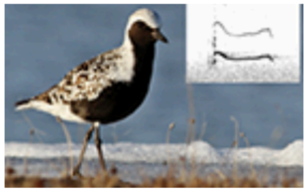
eBird visualisation

Species Maps

Explore interactive range maps by species or subspecies — zoom in for details

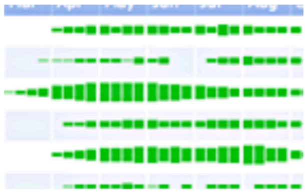
Explore Hotspots

Discover the best places for birding nearby or around the world.



Search Photos and Sounds

Explore media through the Macaulay Library



Bar Charts

Find out what birds to expect throughout the year in a region or location

https://ebird.org/explore

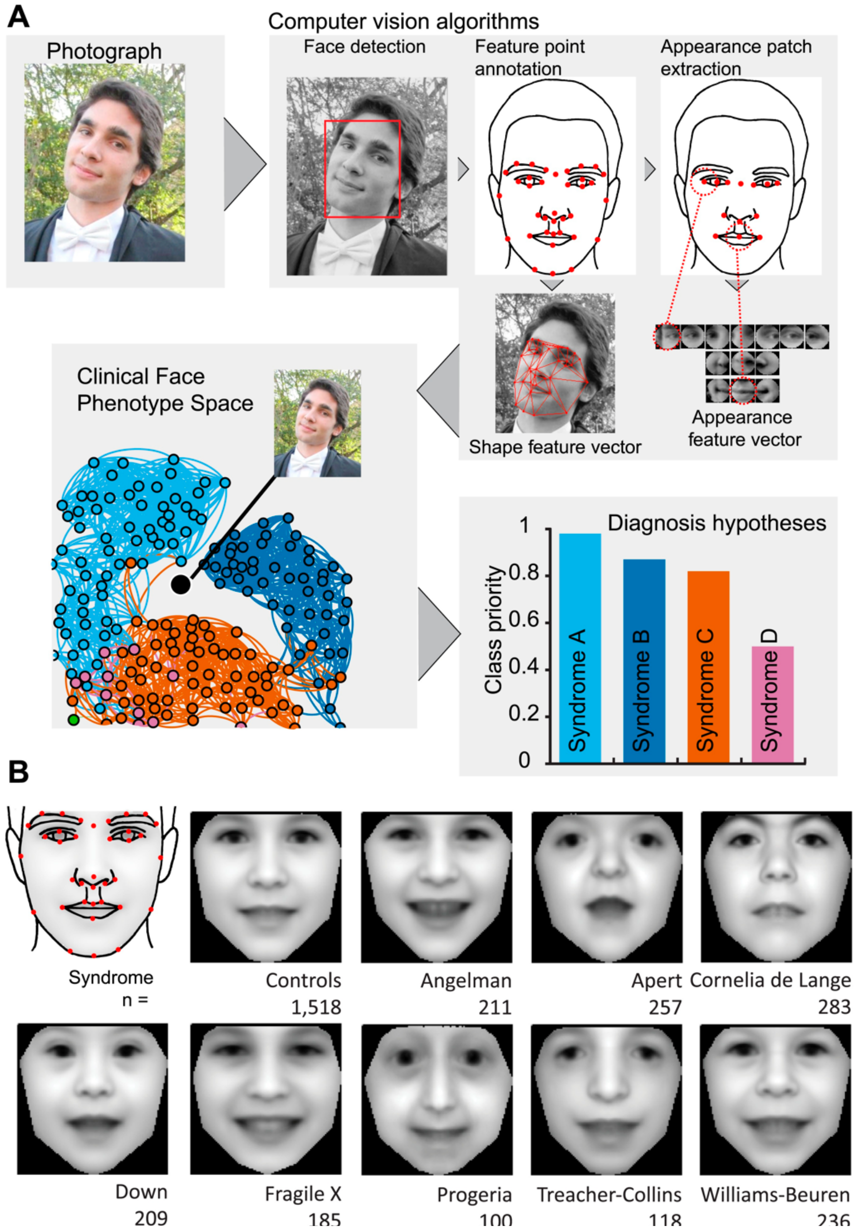
Case Study:

Diagnosing rare genetic diseases from photographs

Overview

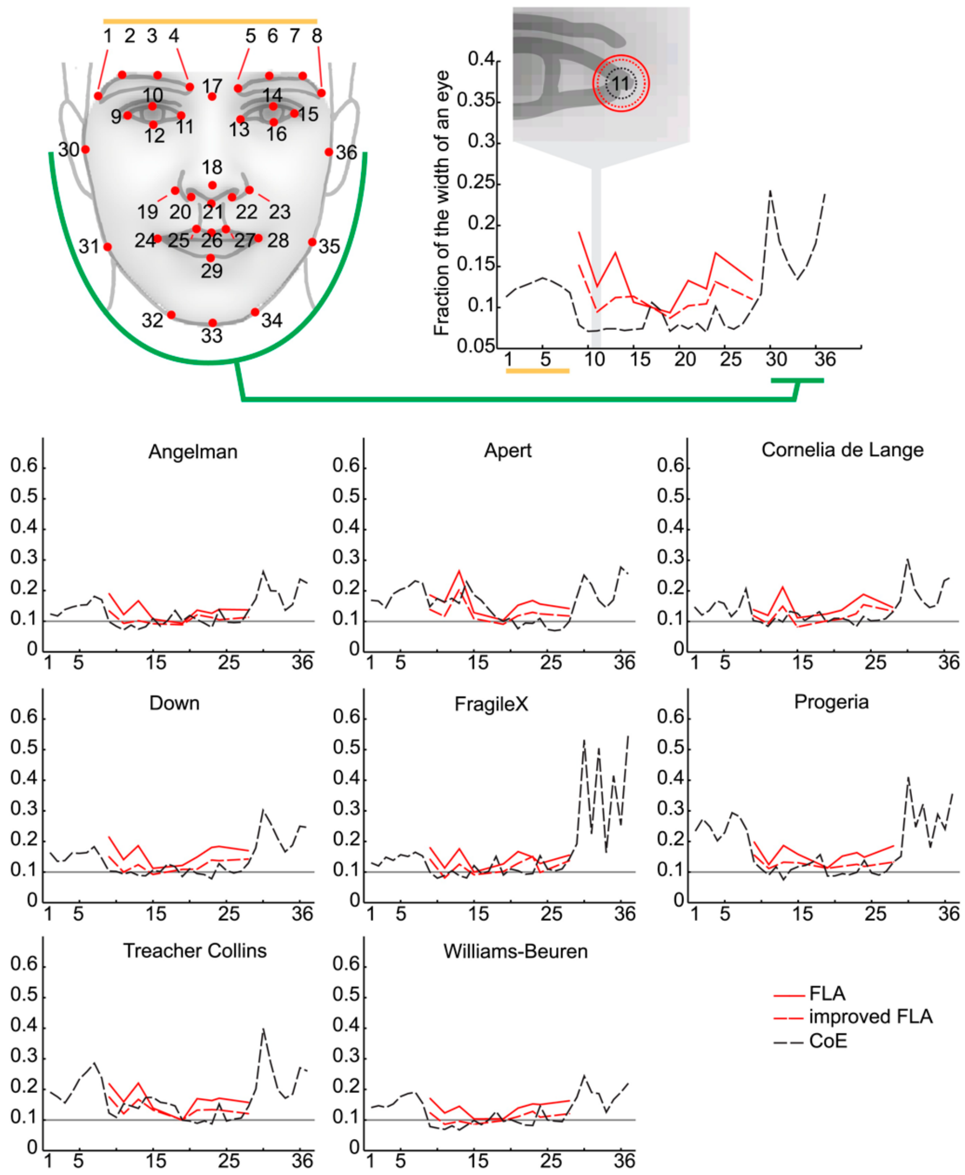
- (A) Overview of computational approach
- (B) Average faces

Ferry et al.. eLife (2014)



Facial feature points

- Accuracy of automatic image annotation, relative to manually annotated ground truth

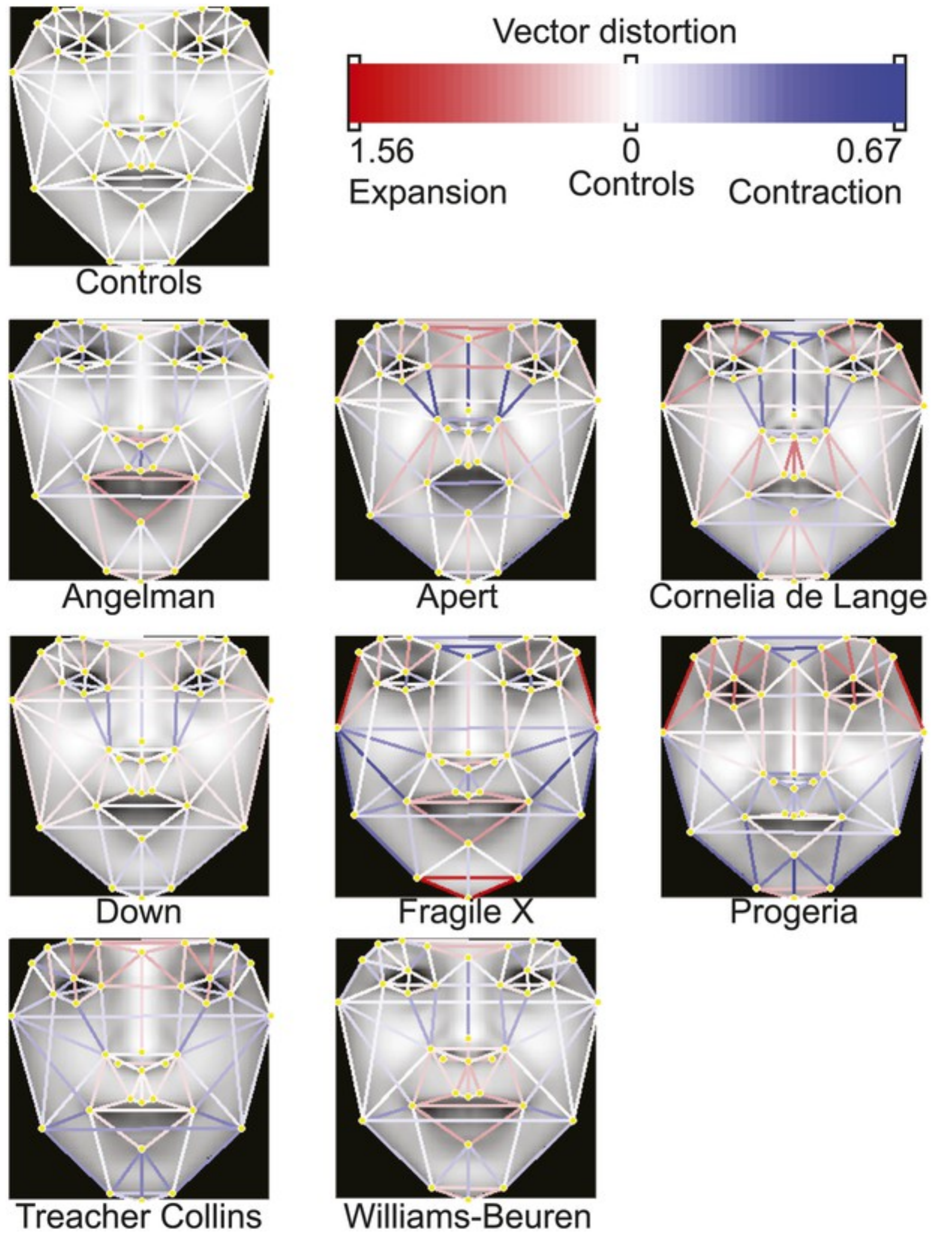


Ferry et al.. eLife (2014)

Distortion graphs

- Represent the characteristic deformation of syndrome faces relative to the average control face.

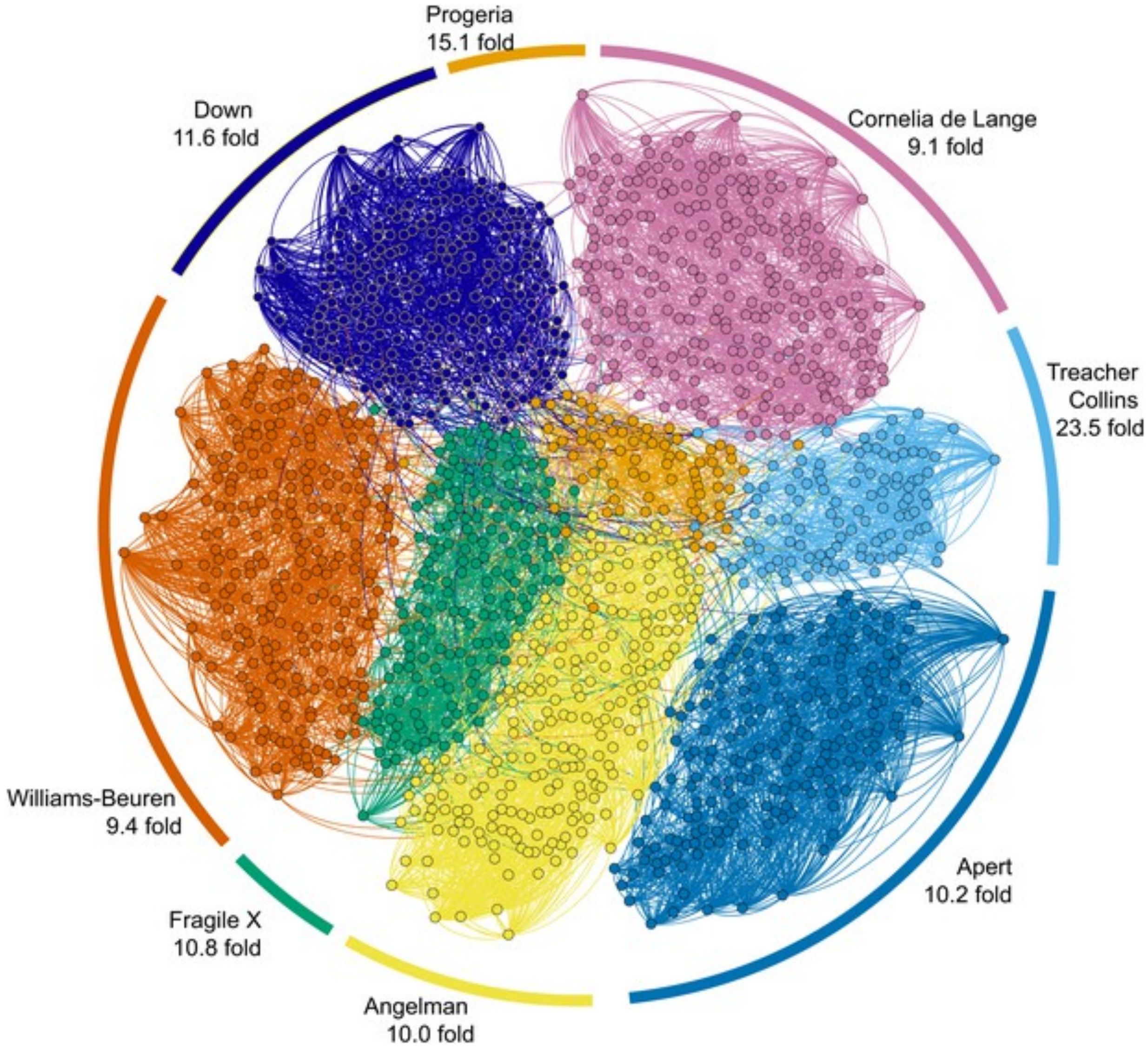
- Shorter (blue), extended (red) and similar (white) distances.



Ferry et al.. eLife (2014)

Clinical face phenotype space

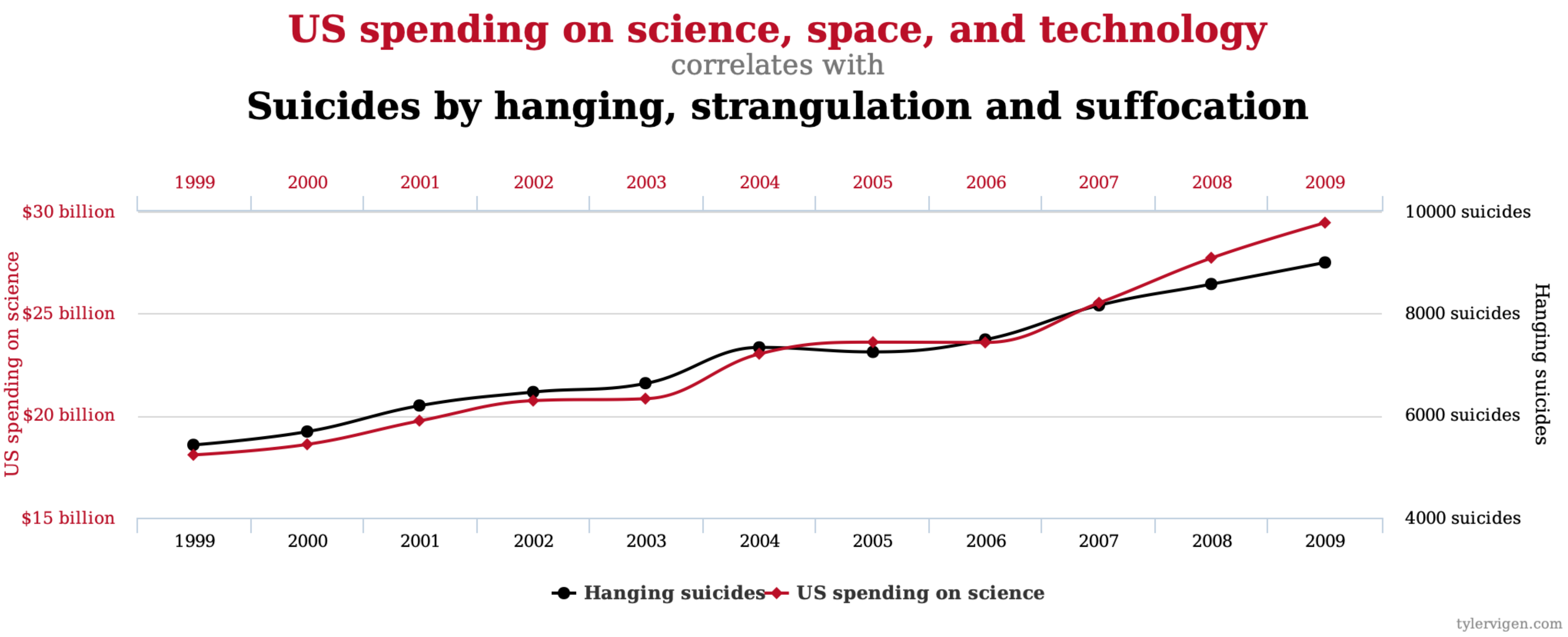
- X fold better clustering than random expectation
- Links to the 10 nearest neighbours of each photo

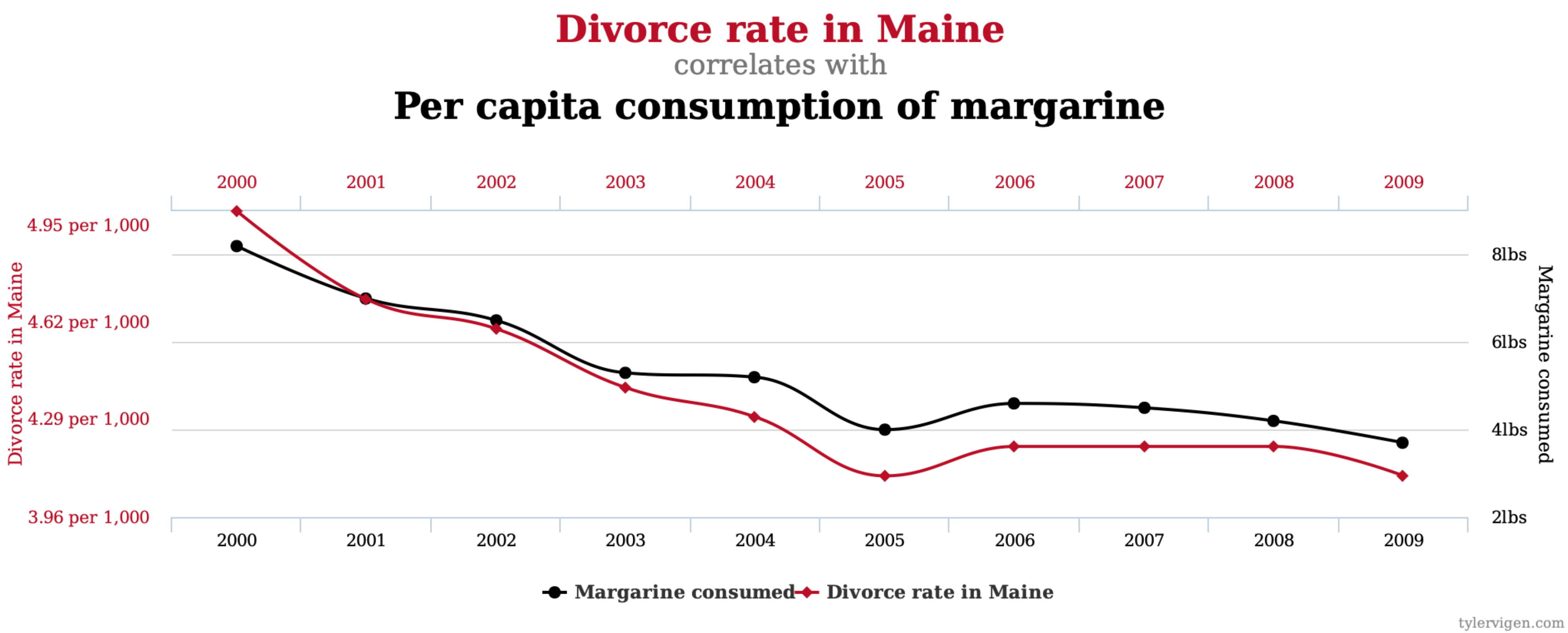


Ferry et al.. eLife (2014)

Arrange these tasks into groups:

A. Predict whether a manuscript will be a bestseller novel
B. Find texts in a corpus that probably have the same author
C. Predict what will be a company’s share price tomorrow
D. Predict which companies’ shares will go up tomorrow
E. Find evolutionary relationships among a set of species
F. Determine whether a news article is fake
G. Find ”communities” of users of a music service who have similar tastes
H. Predict the population of Gothenburg in 2030
I. Identify sets of genes that are “switched on” in similar conditions
J. Predict a patient’s blood pressure one hour from now
K. Diagnose a genetic disorder based on facial shape
L. Identify whether a picture is of a cat or a dog
M. Predict how long your journey home will take today
N. Arrange a set of data science tasks into groups


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Hanging suicides
US spending on science
tylervigen.com


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com

https://ourworldindata.org/

Case Study:

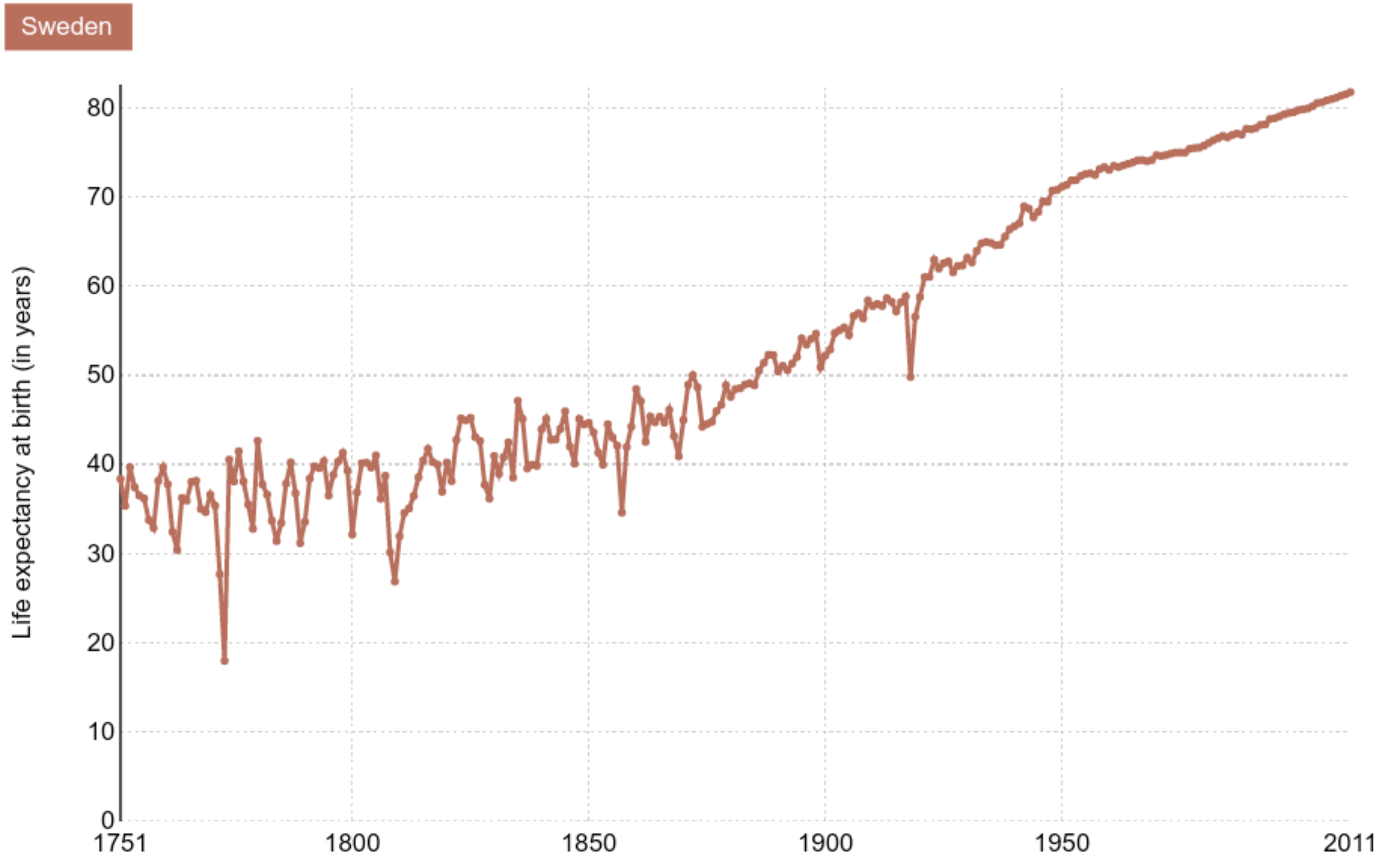
Human Longevity

Human Longevity

- Questions:
  - What lead to increases in human longevity?
  - How should we spend resources to increase longevity?
  - Which metric should we use?

Life expectancy

Shown is period life expectancy at birth. This corresponds to an estimate of the average number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life

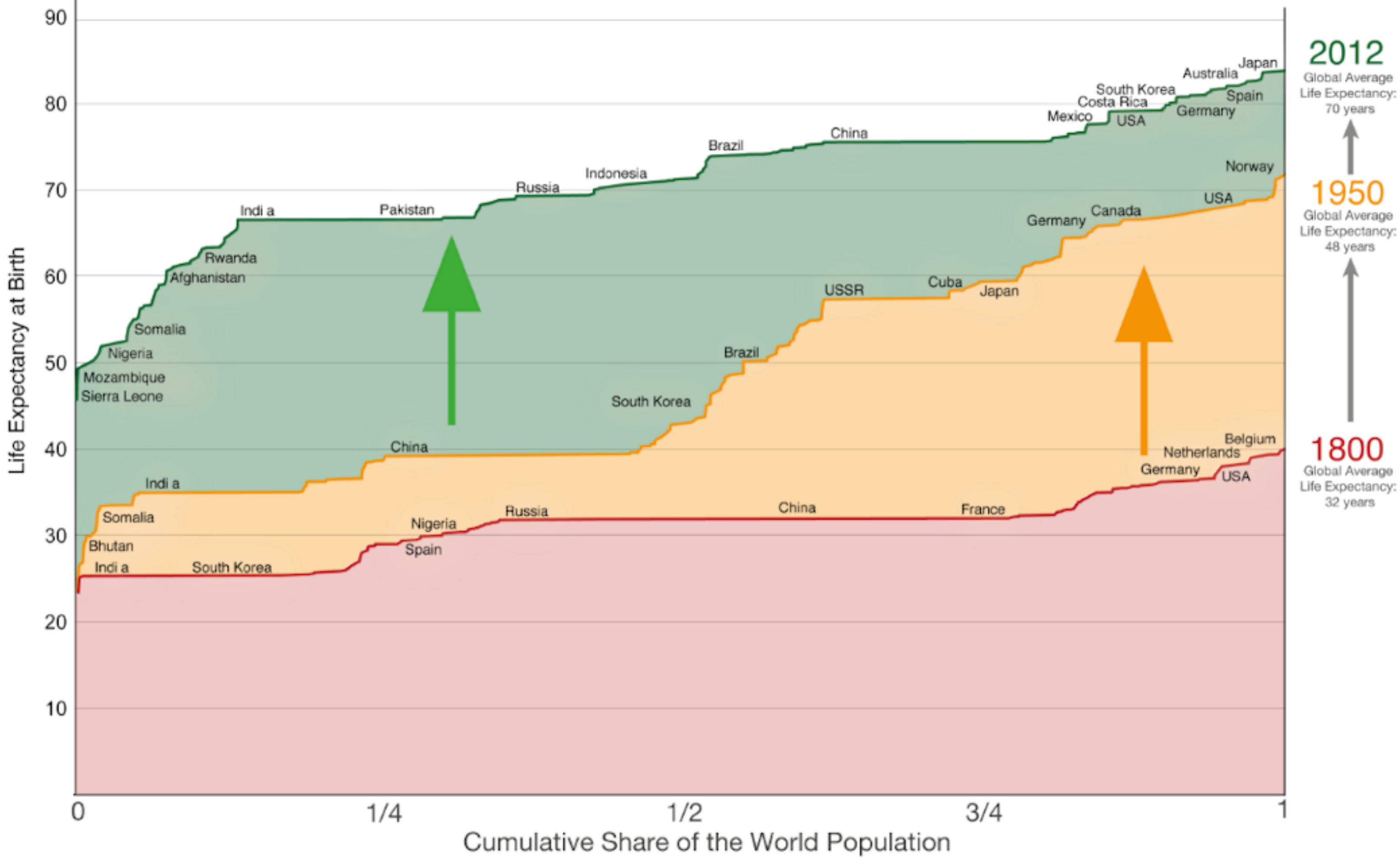


Source: Clio Infra (life expectancy, both genders)

Our World in Data

Life Expectancy of the World Population in 1800, 1950 and 2012

Countries are ordered along the x-axis ascending by the life expectancy of the population. Data for almost all countries is shown in this chart, but not all data points are labelled with the country name.

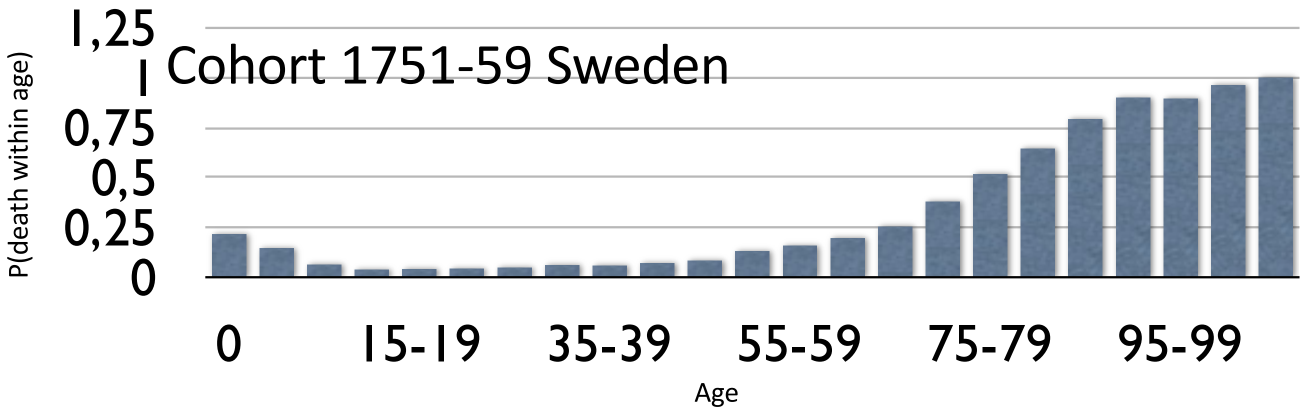


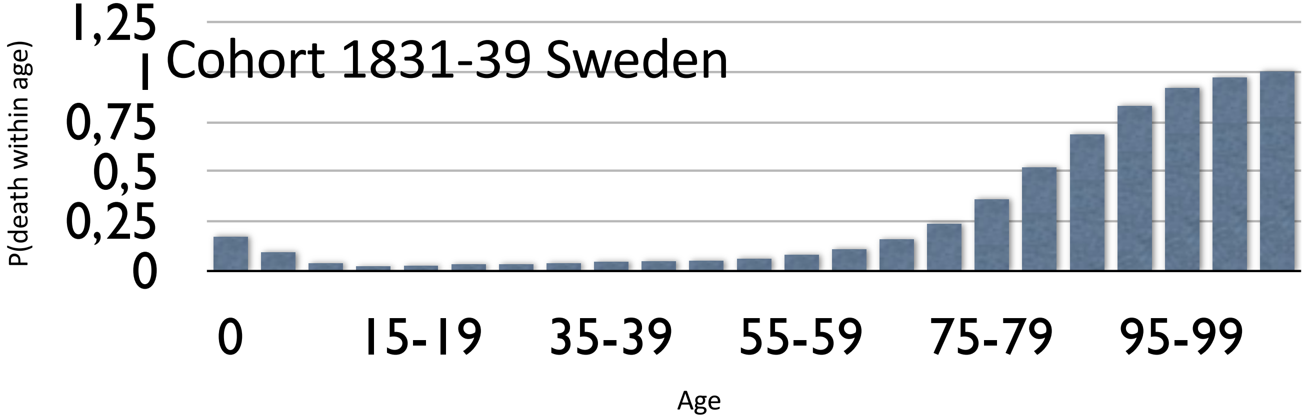
Data source: The data on life expectancy by country and population by country are taken from Gapminder.org.

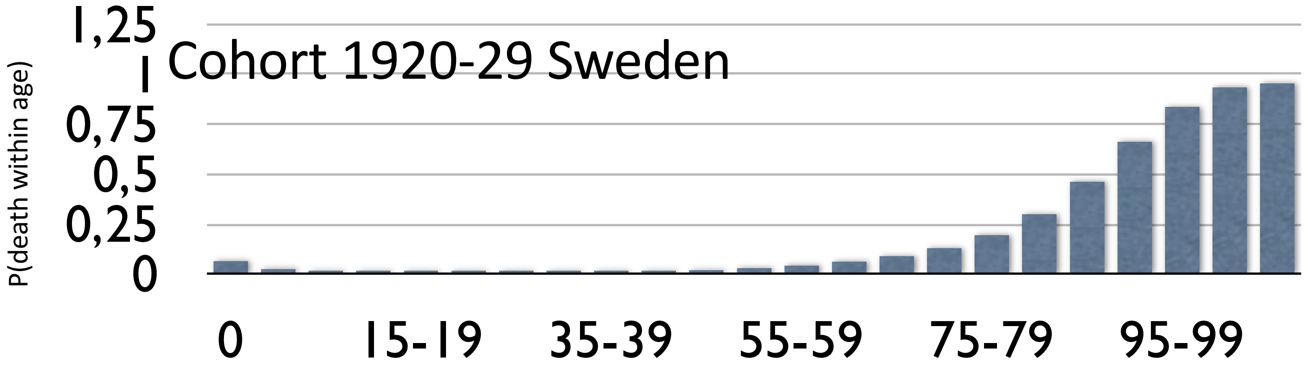
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

Average age at death

- Why did it increase?
    - Did the average person just live longer?
    - Other factors?
- What policies would be appropriate if the first hypothesis is true?


Cohort 1751-59 Sweden
P(death within age)
1,25
1
0,75
0,5
0,25
0
0
15-19
35-39
55-59
75-79
95-99
Age


Cohort 1831-39 Sweden
P(death within age)
1,25
1
0,75
0,5
0,25
0
0
15-19
35-39
55-59
75-79
95-99
Age

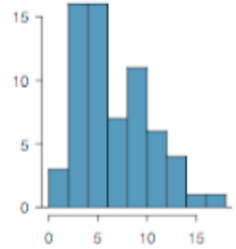

Cohort 1920-29 Sweden
P(death within age)
1,25
1
0,75
0,5
0,25
0
0
15-19
35-39
55-59
75-79
95-99

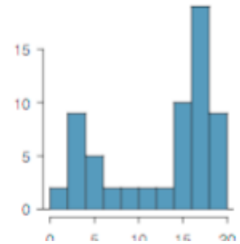
Insight

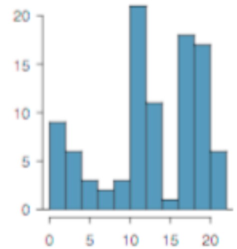
- Life expectancy is impacted by
  - dying of “old age”
  - dying before reaching adulthood
- Difference in life expectancy between 1720-1920 largely dependent on reduction of child mortality.
- *Two groups of individuals and their proportions changed*

Modes of distributions

- A mode of a distribution is a significant peak. A distribution with
  - one peak is called *unimodal*
  - two peaks is called *bimodal*
  - and two or more peaks is called *multi-modal*



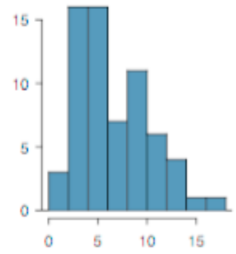


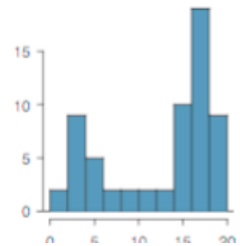


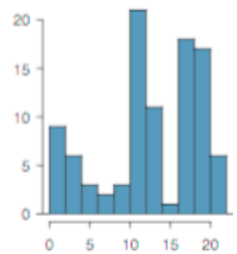
Averages & Modes

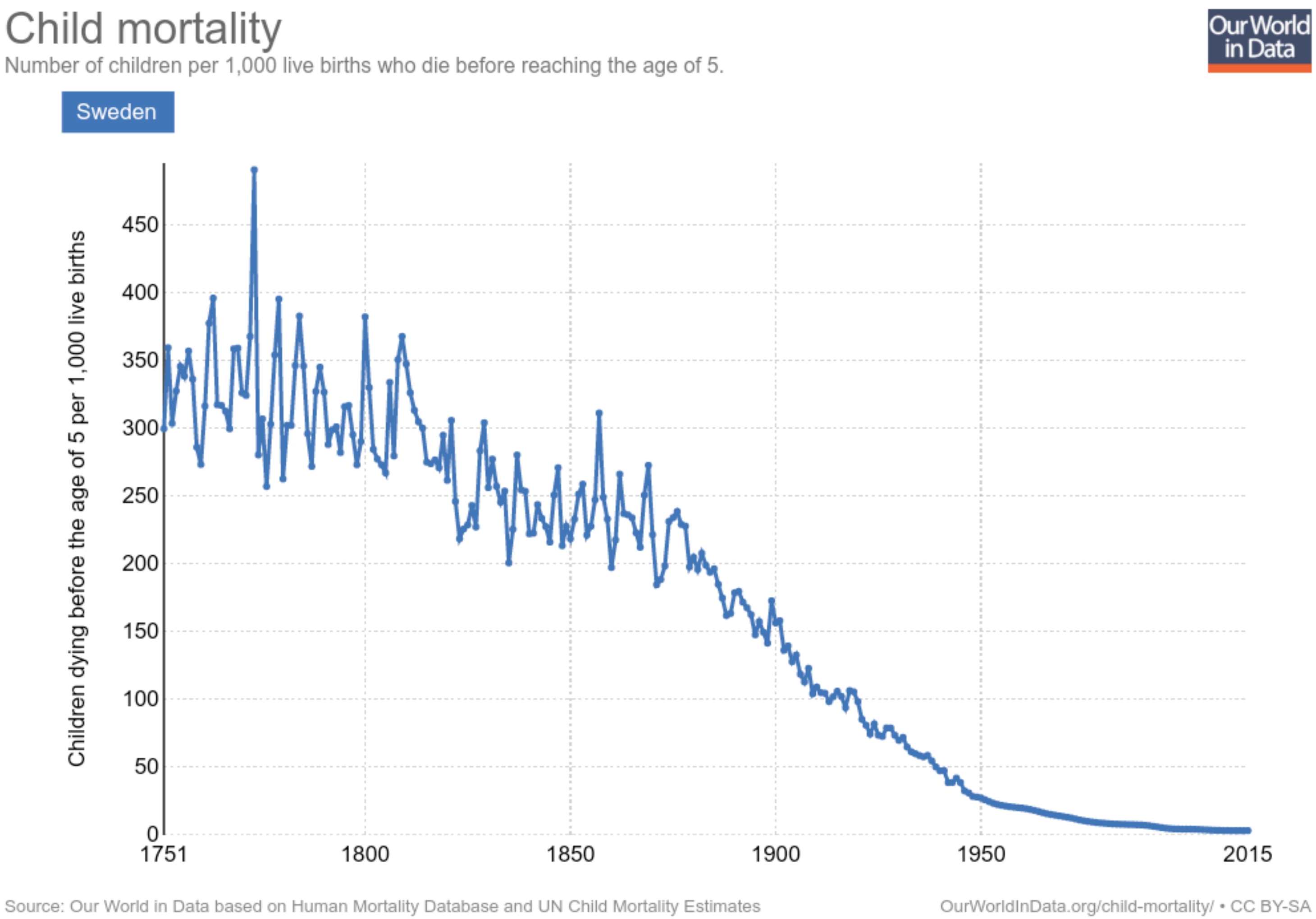
- Averages are appropriate for *unimodal* distributions
- For *bimodal* and *multimodal* distributions the average might be where there are no observations

⇒ Plot distributions



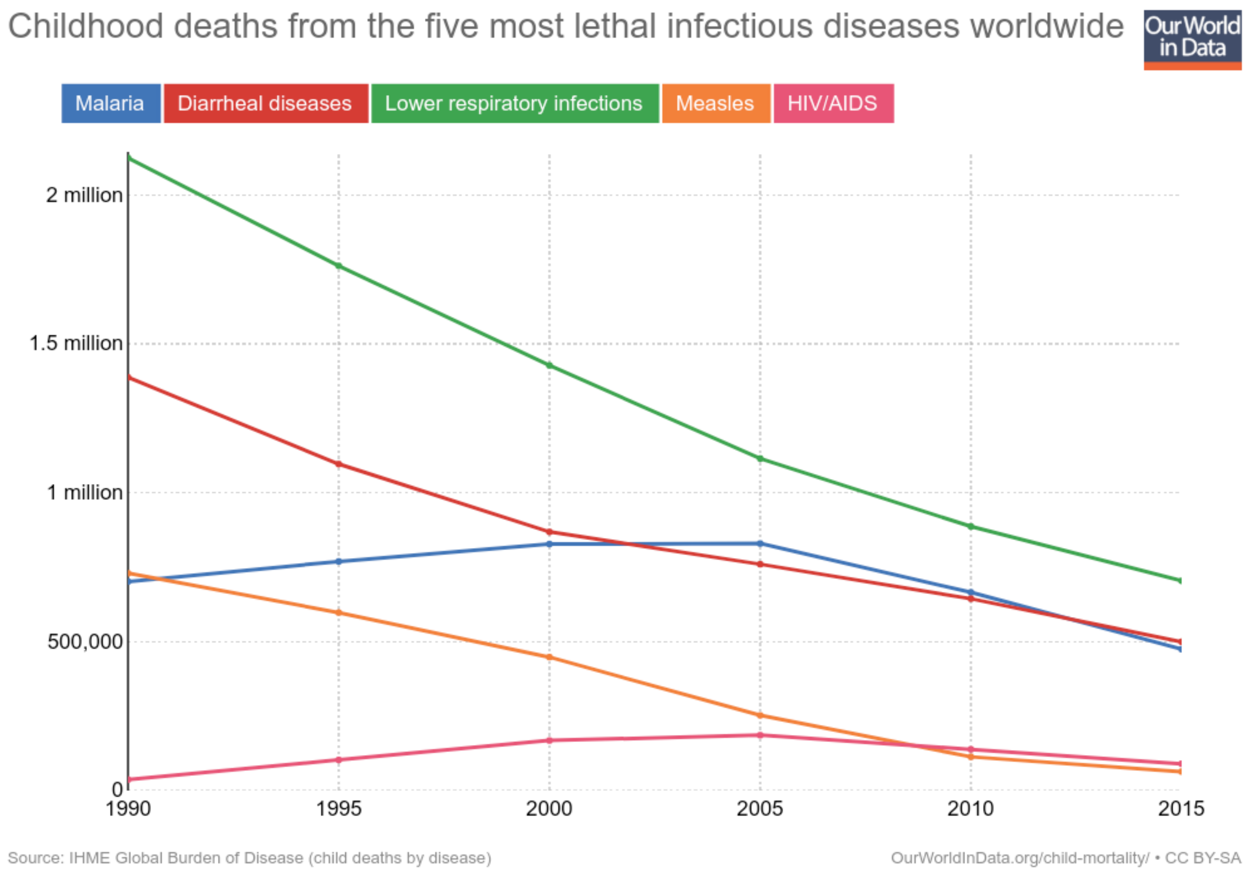


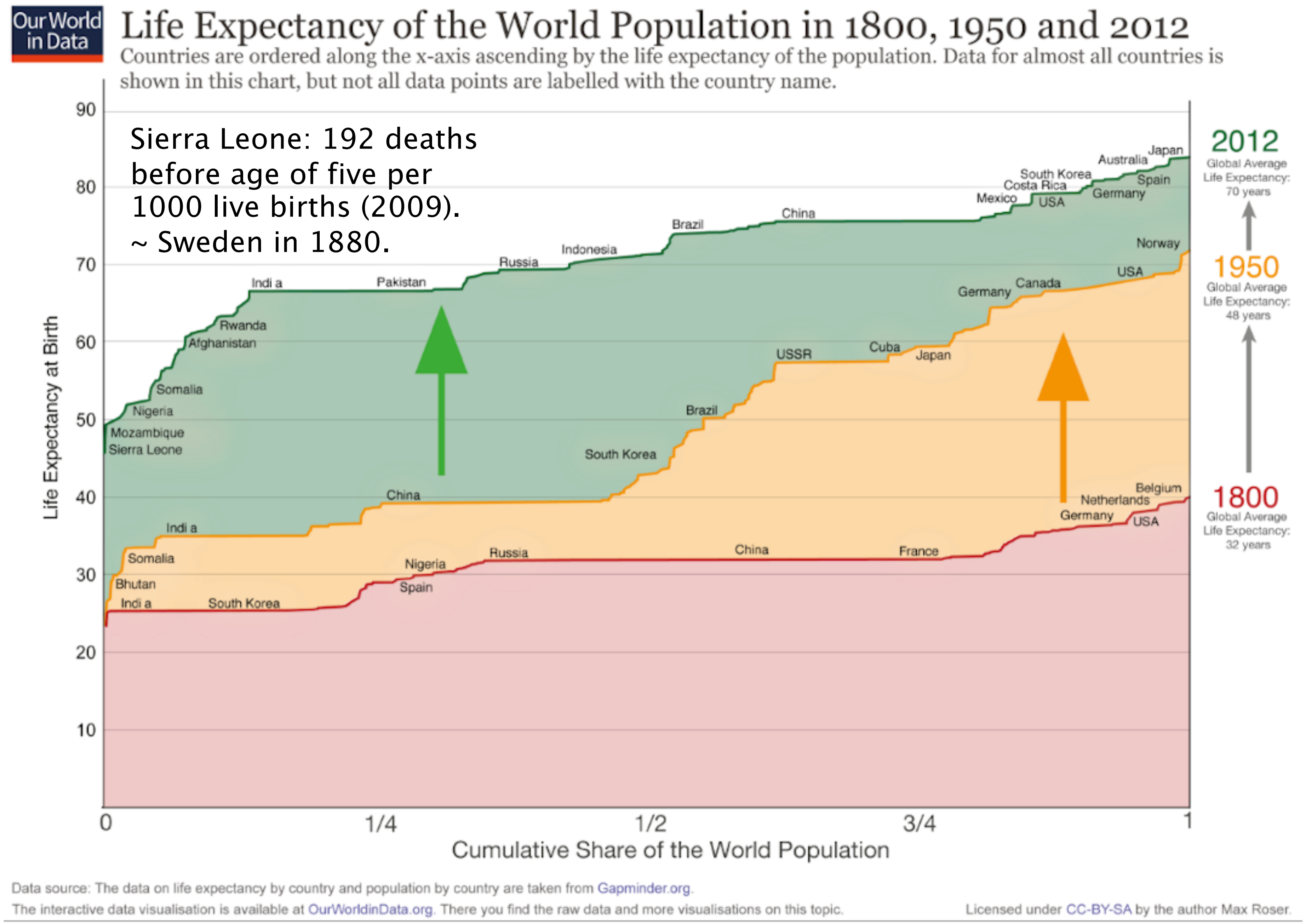




Child mortality
Number of children per 1,000 live births who die before reaching the age of 5.
Our World in Data
Sweden
Children dying before the age of 5 per 1,000 live births
450
400
350
300
250
200
150
100
50
0
1751
1800
1850
1900
1950
2015
Source: Our World in Data based on Human Mortality Database and UN Child Mortality Estimates
OurWorldInData.org/child-mortality/ • CC BY-SA

Possible policy



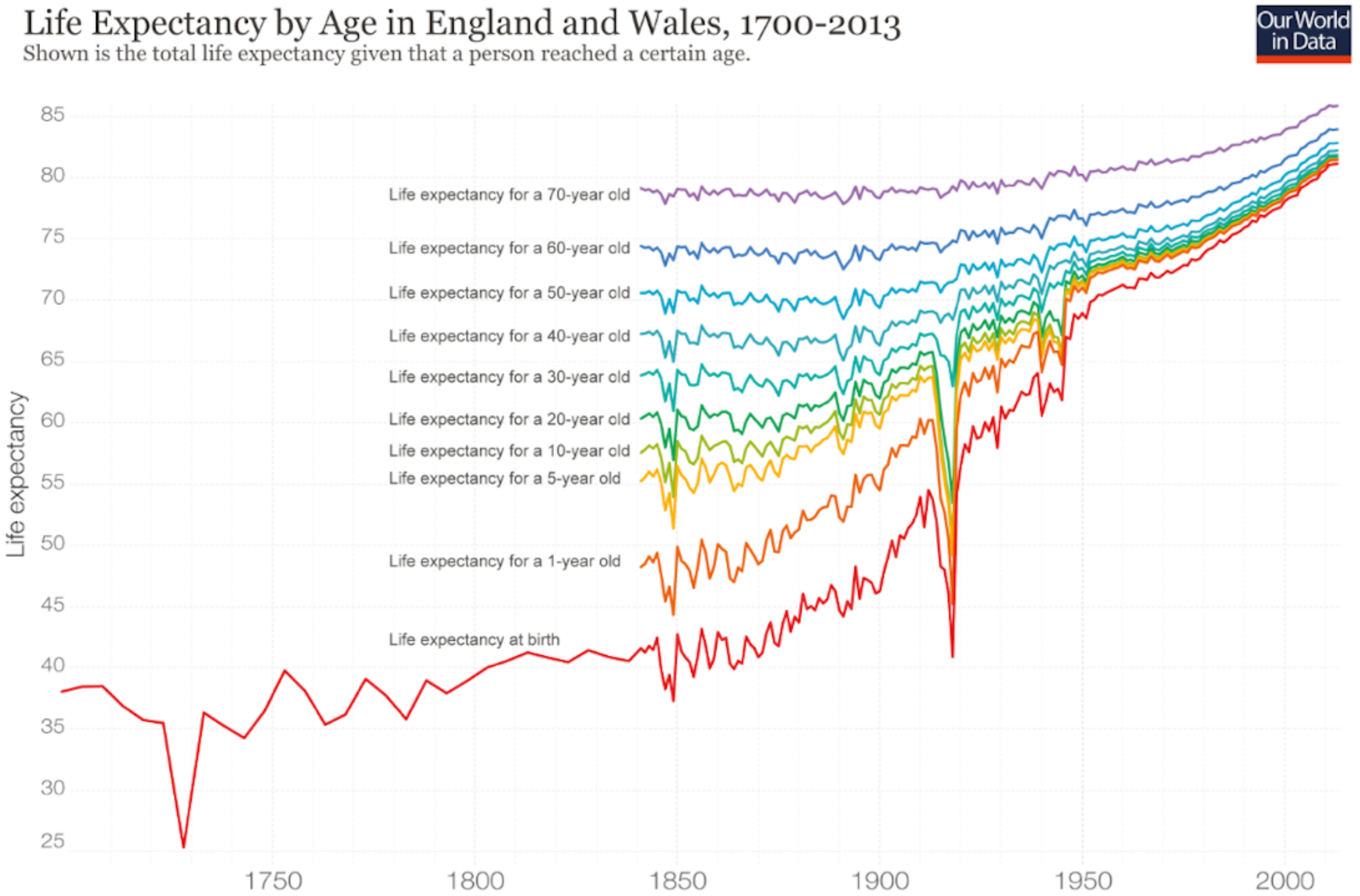

Our World in Data
Life Expectancy of the World Population in 1800, 1950 and 2012
Countries are ordered along the x-axis ascending by the life expectancy of the population. Data for almost all countries is shown in this chart, but not all data points are labelled with the country name.
Sierra Leone: 192 deaths before age of five per 1000 live births (2009).
~ Sweden in 1880.
Life Expectancy at Birth
90
80
70
60
50
40
30
20
10
0
1/4
1/2
3/4
1
Cumulative Share of the World Population
2012
Global Average Life Expectancy: 70 years
1950
Global Average Life Expectancy: 48 years
1800
Global Average Life Expectancy: 32 years
Japan
Australia
South Korea
Costa Rica
Spain
Mexico
USA
Germany
China
Brazil
Indonesia
Russia
Norway
Pakistan
Indi a
USA
Canada
Germany
Rwanda
Afghanistan
USSR
Cuba
Japan
Somalia
Nigeria
Brazil
Mozambique
Sierra Leone
South Korea
China
Belgium
Netherlands
Germany
USA
Indi a
Somalia
Nigeria
Russia
China
France
Bhutan
Spain
Indi a
South Korea
Data source: The data on life expectancy by country and population by country are taken from Gapminder.org.
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.
Licensed under CC-BY-SA by the author Max Roser.

Is this all?

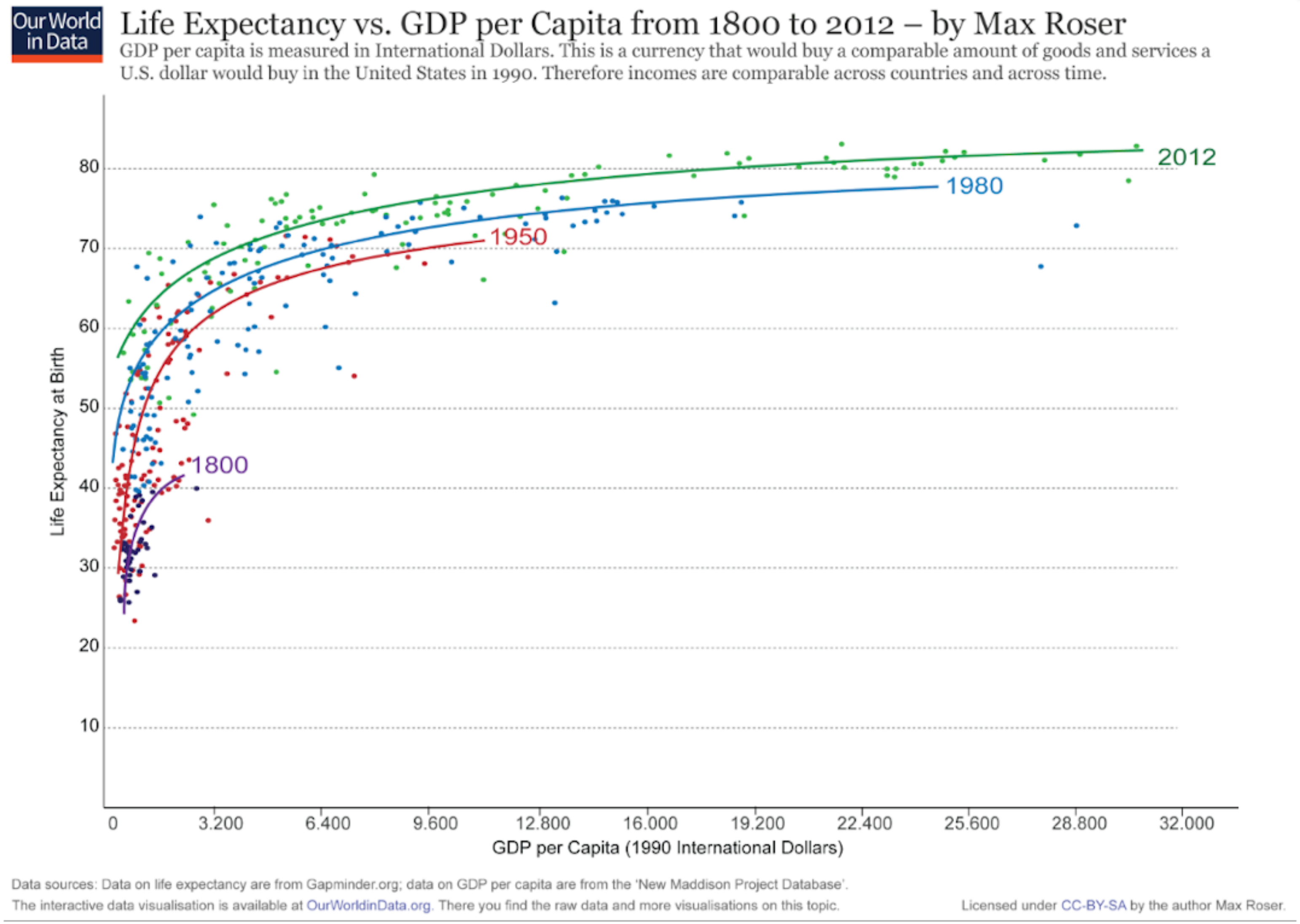
- Which data should we inspect to get an insight whether reducing child mortality is all that matters?

Life Expectancy by Age in England and Wales, 1700-2013

Shown is the total life expectancy given that a person reached a certain age.



Data source: Life expectancy at birth Clio-Infra. Data on life expectany at age 1 and older from the Human Mortality Database (www.mortality.org).
The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic.



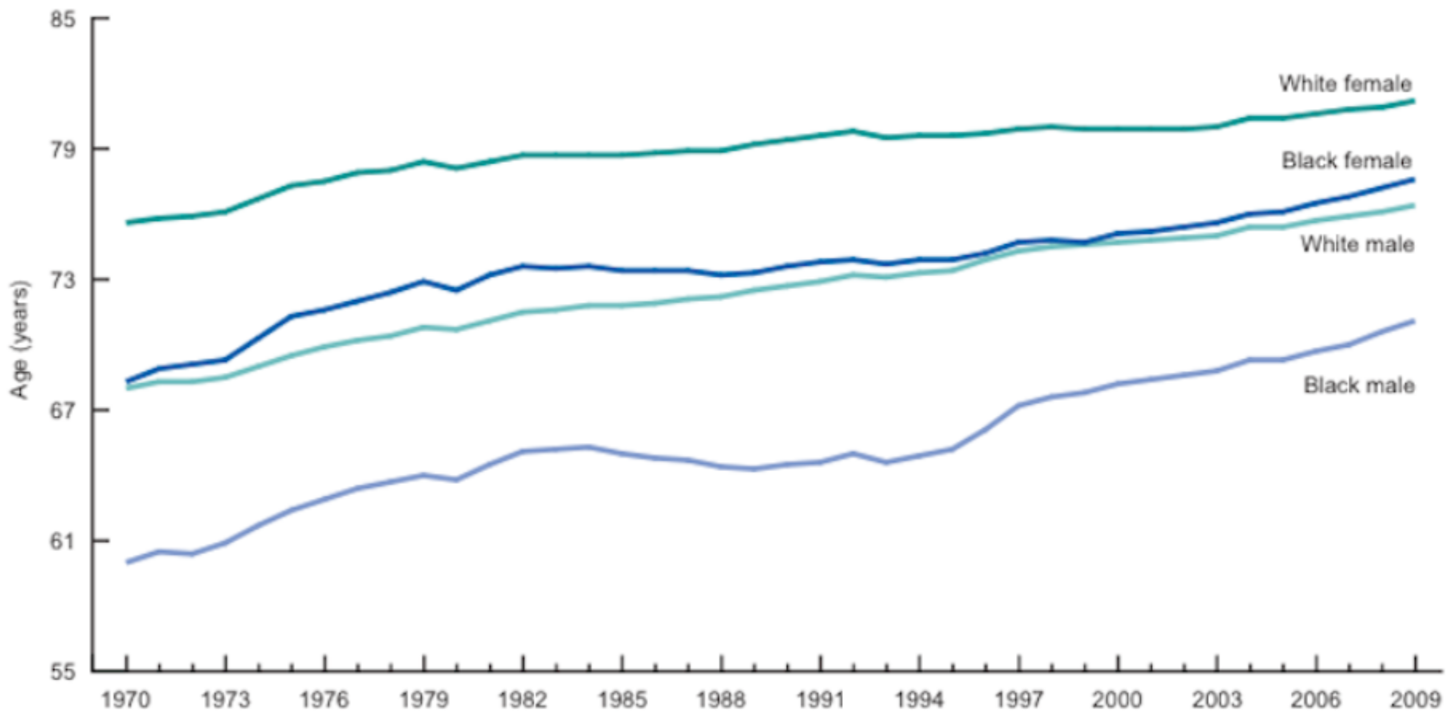
Our World in Data
Life Expectancy vs. GDP per Capita from 1800 to 2012 – by Max Roser
GDP per capita is measured in International Dollars. This is a currency that would buy a comparable amount of goods and services a U.S. dollar would buy in the United States in 1990. Therefore incomes are comparable across countries and across time.
Life Expectancy at Birth
GDP per Capita (1990 International Dollars)
2012
1980
1950
1800
80
70
60
50
40
30
20
10
0
3.200
6.400
9.600
12.800
16.000
19.200
22.400
25.600
28.800
32.000
Data sources: Data on life expectancy are from Gapminder.org; data on GDP per capita are from the 'New Maddison Project Database'.
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.
Licensed under CC-BY-SA by the author Max Roser.

Stratification

- Divide data into homogeneous subpopulations before analysis
- Possible variables: gender, age, socio-economic status, ...

- Can result in further insights

Race and sex vs. life expectancy

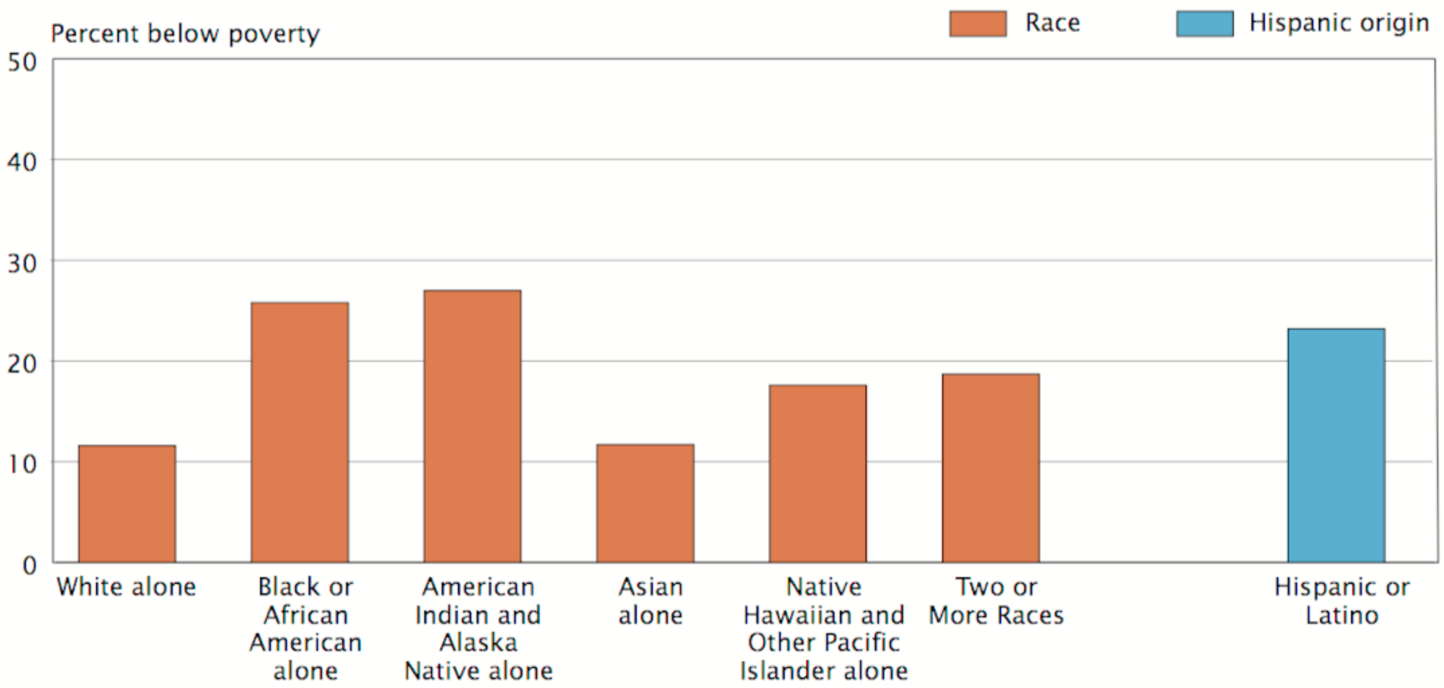


Life expectancy at birth by race and sex, United States CDC/NCHS, National Vital Statistics System

Figure 1.

U.S. Poverty Rates by Race and Hispanic or Latino Origin: 2007–2011

(For information on confidentiality protection, sampling error, nonsampling error, and definitions, see *www.census.gov/acs/www/*)



Note: Persons who report only one race among the six defined categories are referred to as the race-alone population, while persons who report more than one race category are referred to as the Two or More Races population. This figure shows data using the race-alone approach. Use of the single-race population does not imply that it is the preferred method of presenting or analyzing data. The Census Bureau uses a variety of approaches. Because Hispanics may be of any race, data in this figure for Hispanics overlap with data for race groups.

Source: U.S. Census Bureau, 2007–2011 American Community Survey.

Uninsured Population by Household Income, Percent

State of Iowa

Less than $15,000	21.5%
$15,000-$24,999	22.2%
$25,000-$34,999	14.2%
$35,000-$49,999	8.8%
$35,000-$49,999	3.1%
$75,000+	1.7%

Uninsured Population by Education, Percent

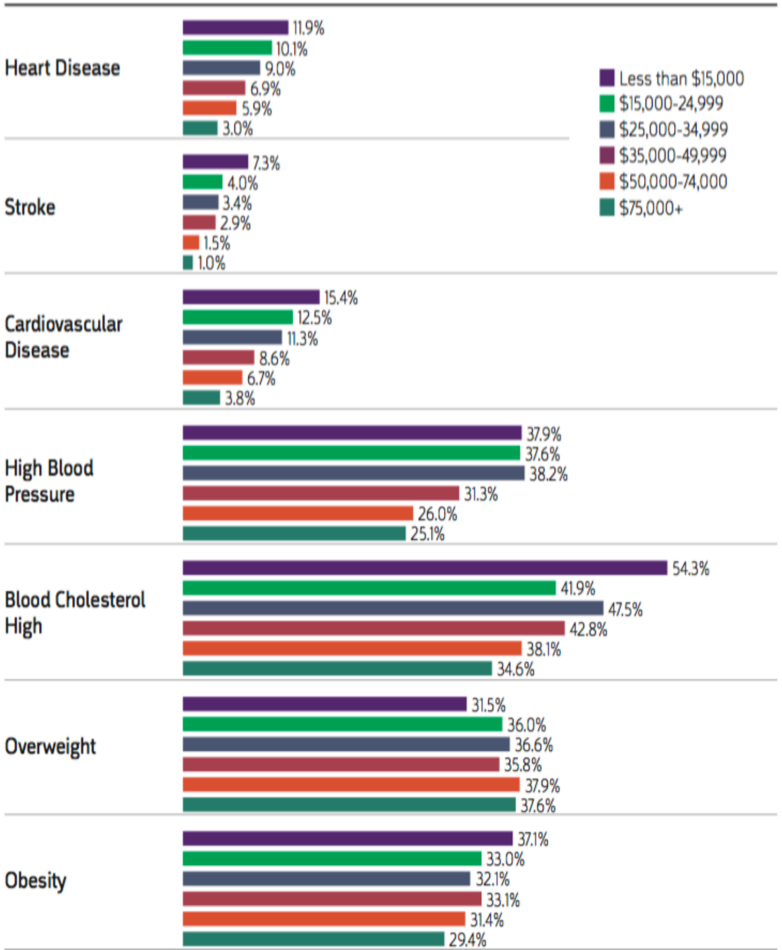
State of Iowa

Less than high school graduate	25.3%
High school or G.E.D.	11.4%
Some post-high school	9.7%
College graduate	3.6%

Source: Health in Iowa Annual Report from the Behavioral Risk Factor Surveillance System, Iowa 2013

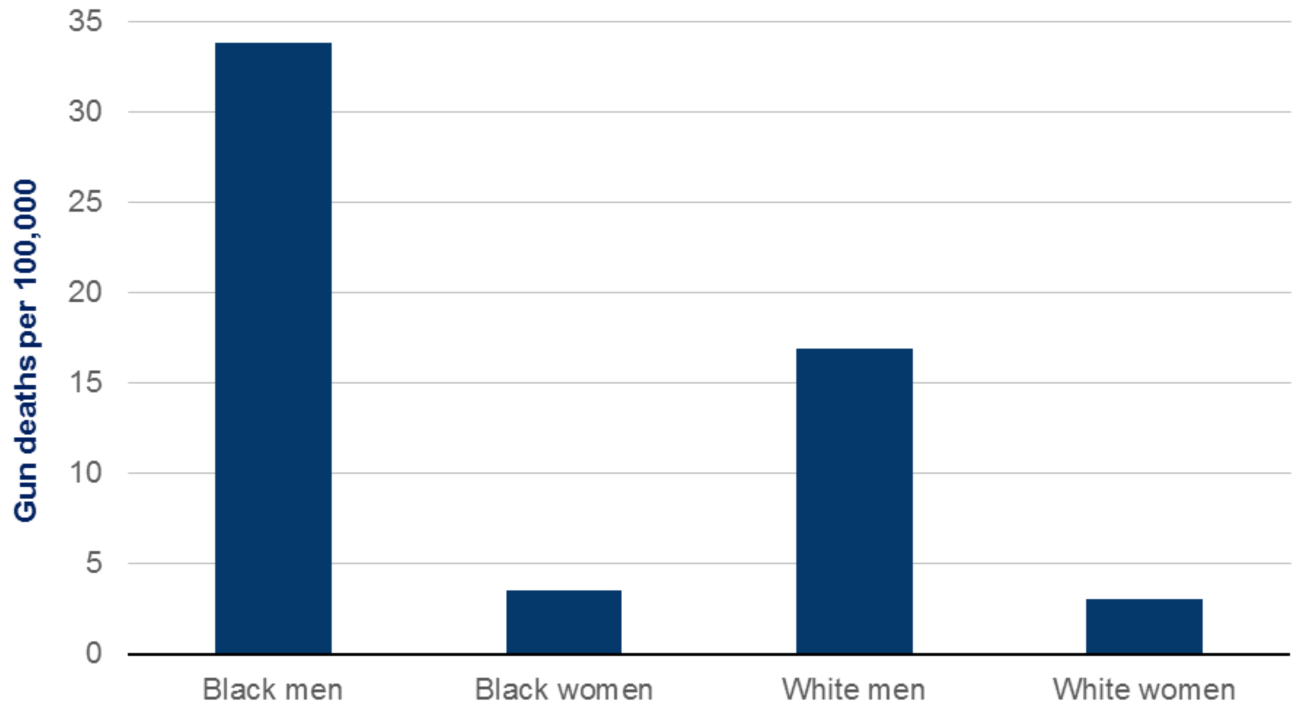
Disease Prevalence by Household Income

State of Iowa



Source: Health in Iowa Annual Report from the Behavioral Risk Factor Surveillance System, Iowa 2013

U.S. gun deaths by race and gender, 2011-2013

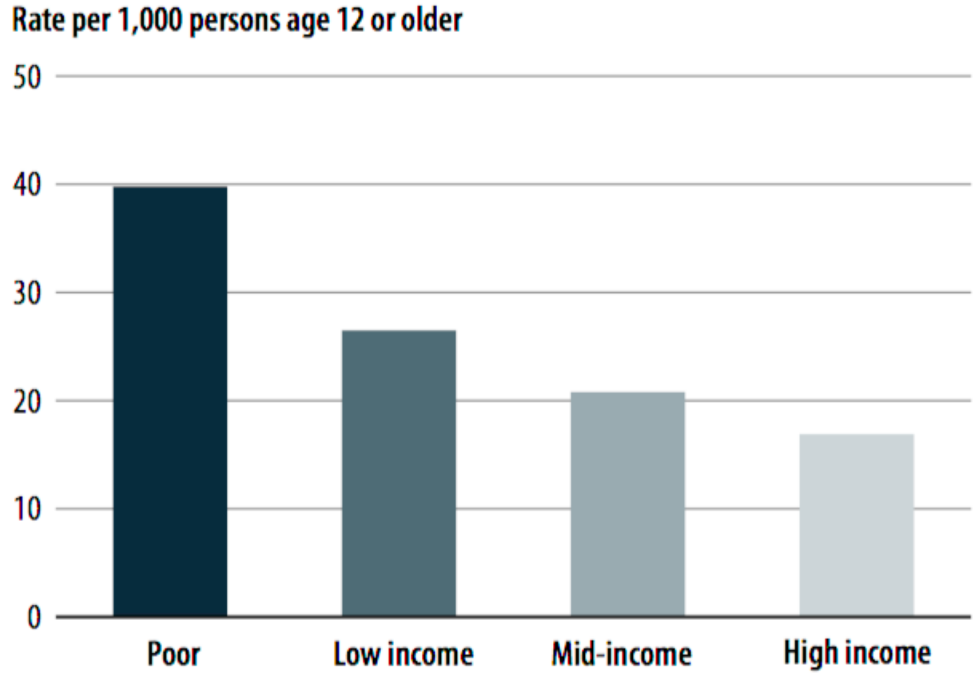


Note: These figures have all been calculated using a 2011-2013 average to smooth single-year fluctuations.
Source: CDC Injury Prevention & Control database.

BROOKINGS

FIGURE 1

Rate of violent victimization, by household poverty level, 2008–2012



Note: Poor refers to households at 0% to 100% of the Federal Poverty Level (FPL). Low income refers to households at 101% to 200% of the FPL. Mid-income refers to households at 201% to 400% of the FPL. High income refers to households at 401% or higher than the FPL. See table 1 for estimates and appendix table 1 for standard errors.

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2008–2012.

Reading

- 1.2 “Asking Interesting Questions from Data”
- 6.1 “Exploratory Data Analysis”
- 6.2 “Developing a Visual Aesthetic”