Social Behavior and Trends

Foundations of Computational Social Science

Lecturer: **Lisette Espín-Noboa**

espin@csh.ac.at | @lespin

Postdoc at Complexity Science Hub Vienna

Postdoc at Central European University

October, 31, 2023

TU Graz

Social Behavior and Trends

Foundations of Computational Social Science

https://github.com/**lisette-espin/TeachingMaterials**

Lecturer: **Lisette Espín-Noboa**
espin@csh.ac.at | @lespin
Postdoc at Complexity Science Hub Vienna
Postdoc at Central European University

October, 31, 2023
TU Graz

What is behavior?
What is a trend?

Social behavior & trends

Differences

Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.

A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.

Social behavior & trends

Differences

Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.

A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.

Social behavior & trends

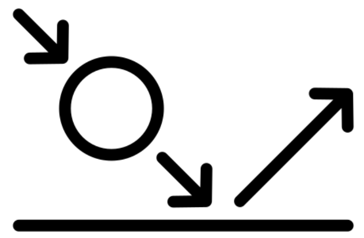
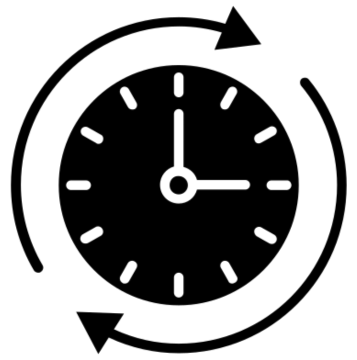
Differences

Focus

Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.	Interactions of and among individuals
A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.	Changes in behavior or attitudes

Social behavior & trends

Differences

	Focus	Timeframe
Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.	Interactions of and among individuals	 Immediate action/reaction
A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.	Changes in behavior or attitudes	 Long-term development

Social behavior & trends

Differences

	Focus	Timeframe	Scope
Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.	Interactions of and among individuals	Immediate action/reaction	Individuals or groups
A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.	Changes in behavior or attitudes	Long-term development	General population

Social behavior & trends

Differences

	Focus	Timeframe	Scope	Analysis
Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.	Interactions of and among individuals	Immediate action/reaction	Individuals or groups	Why? Intercept with social theory? What triggered it?
A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.	Changes in behavior or attitudes	Long-term development	General population	Who started it? Who has adopted it? For how long will it last?

Social behavior & trends

Similarities

	Focus	Timeframe	Scope	Analysis
Behavior refers to the actions, reactions, or conduct of individuals or groups in response to a particular situation.	Both involve the interaction of individuals (patterns)	Both are influenced by societal norms, cultural practices, and environmental factors		Both are within the interests of computational social scientists
A **Trend** is a pattern, fashion or tendency that persists over time. Also, a direction in which something is developing or changing.				

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior | Trend

- How users navigate the Web, the city, etc.
- The use of ChatGPT in education & science
- Online consumer decision-making
- The increasing use of ML in CSS research
- The growing interest in understanding the ethical implications of CSS research
- The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.
- Modeling social networks (how edges form)
- Strategies in online gaming
- How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Examples

Behavior

Trend

How users navigate the Web, the city, etc.

The use of ChatGPT in education & science

Online consumer decision-making

The increasing use of ML in CSS research

The growing interest in understanding the ethical implications of CSS research

The increasing focus on interdisciplinary collaboration between computer science, social sciences, and statistics.

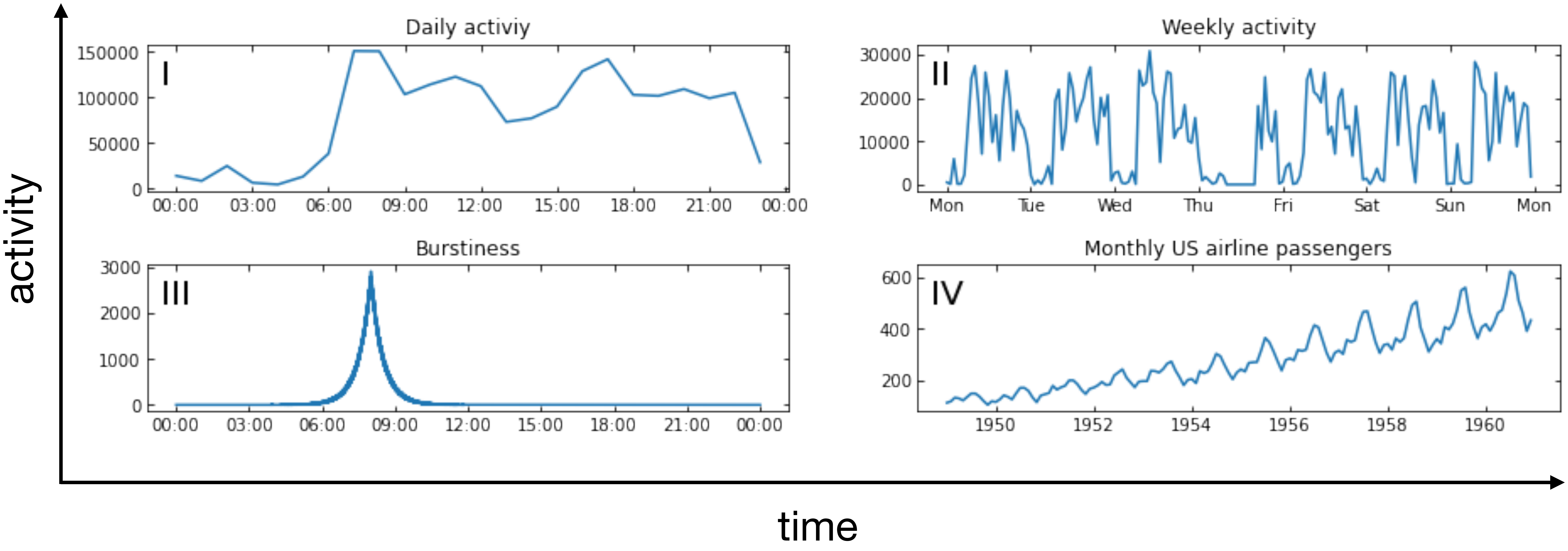
Modeling social networks (how edges form)

Strategies in online gaming

How scientists collaborate in academia

Social behavior & trends

Temporal patterns (time vs. social activity)



Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics

1. Social Science
2. CSS
3. Digital Traces
4. Examples

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy		

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

Social Behavior

Social behavior

and the social sciences

Social behavior

and the social sciences

- **Social Psychology** focuses on understanding how individuals' thoughts, feelings, and behaviors are influenced by the presence of others.
    - Key concepts: social influence, attitudes, and group dynamics.

Social behavior

and the social sciences

- **Social Psychology** focuses on understanding how individuals' thoughts, feelings, and behaviors are influenced by the presence of others.
  - Key concepts: social influence, attitudes, and group dynamics.
- **Cognitive Science** focuses on understanding how people perceive, think, and remember information.
  - Key concepts: attention, memory, and decision-making.

Social behavior

and the social sciences

- **Social Psychology** focuses on understanding how individuals' thoughts, feelings, and behaviors are influenced by the presence of others.
  - Key concepts: social influence, attitudes, and group dynamics.
- **Cognitive Science** focuses on understanding how people perceive, think, and remember information.
  - Key concepts: attention, memory, and decision-making.
- **Behavioral Economics** combines insights from psychology and economics to understand how people make decisions.
  - Key concepts: heuristics and biases, and how they can be applied to understand social behavior.

Social behavior

and computational social science

Social behavior

and computational social science

- The digital revolution (BigData & AI) has affected the social sciences. Moving from data scarcity and local to large-scale, complex, and global [Veltri 2023].

Social behavior

and computational social science

- The digital revolution (BigData & AI) has affected the social sciences. Moving from data scarcity and local to large-scale, complex, and global [Veltri 2023].
    - **More data:** amount, types or modes, complexity

Social behavior

and computational social science

- The digital revolution (BigData & AI) has affected the social sciences. Moving from data scarcity and local to large-scale, complex, and global [Veltri 2023].
    - **More data:** amount, types or modes, complexity
    - **New computational methods:** scalable

Social behavior

and computational social science

- The digital revolution (BigData & AI) has affected the social sciences. Moving from data scarcity and local to large-scale, complex, and global [Veltri 2023].
  - **More data:** amount, types or modes, complexity
  - **New computational methods:** scalable
  - **Existing theories:** need to be revised (using more and new kinds of data)

Social behavior

and computational social science

Social behavior

and computational social science

- This new paradigm has enabled the study of **human behavior** from three different perspectives:

Social behavior

and computational social science

- This new paradigm has enabled the study of **human behavior** from three different perspectives:
  - **Online population-based experiments** (e.g., survey, randomized controlled trials)

Social behavior

and computational social science

- This new paradigm has enabled the study of **human behavior** from three different perspectives:
    - **Online population-based experiments** (e.g., survey, randomized controlled trials)
    - **Observational studies** or large-scale <u>natural experiments</u> (e.g., non-randomized trials, logs, historical data)

Social behavior

and computational social science

- This new paradigm has enabled the study of **human behavior** from three different perspectives:
    - **Online population-based experiments** (e.g., survey, randomized controlled trials)
    - **Observational studies** or large-scale natural experiments (e.g., non-randomized trials, logs, historical data)
    - **Digital traces** in online platforms can provide new insights on human mobility, opinion and communication dynamics, human-human and human-computer interaction, mental health, disease or information spreading, political polarization, voter behavior, shopping/music/entertainment preferences, learning behavior, crime patterns and potential threats, etc. [Keusch and Kreuter 2021]

Social behavior

and computational social science

- This new paradigm has enabled the study of **human behavior** from three different perspectives:
  - **Online population-based experiments** (e.g., survey, randomized controlled trials)
  - **Observational studies** or large-scale natural experiments (e.g., non-randomized trials, logs, historical data)
  - **Digital traces** in online platforms can provide new insights on human mobility, opinion and communication dynamics, human-human and human-computer interaction, mental health, disease or information spreading, political polarization, voter behavior, shopping/music/entertainment preferences, learning behavior, crime patterns and potential threats, etc. [Keusch and Kreuter 2021]
    - Including **ethical research** to ensure transparency, data privacy, and fairness.

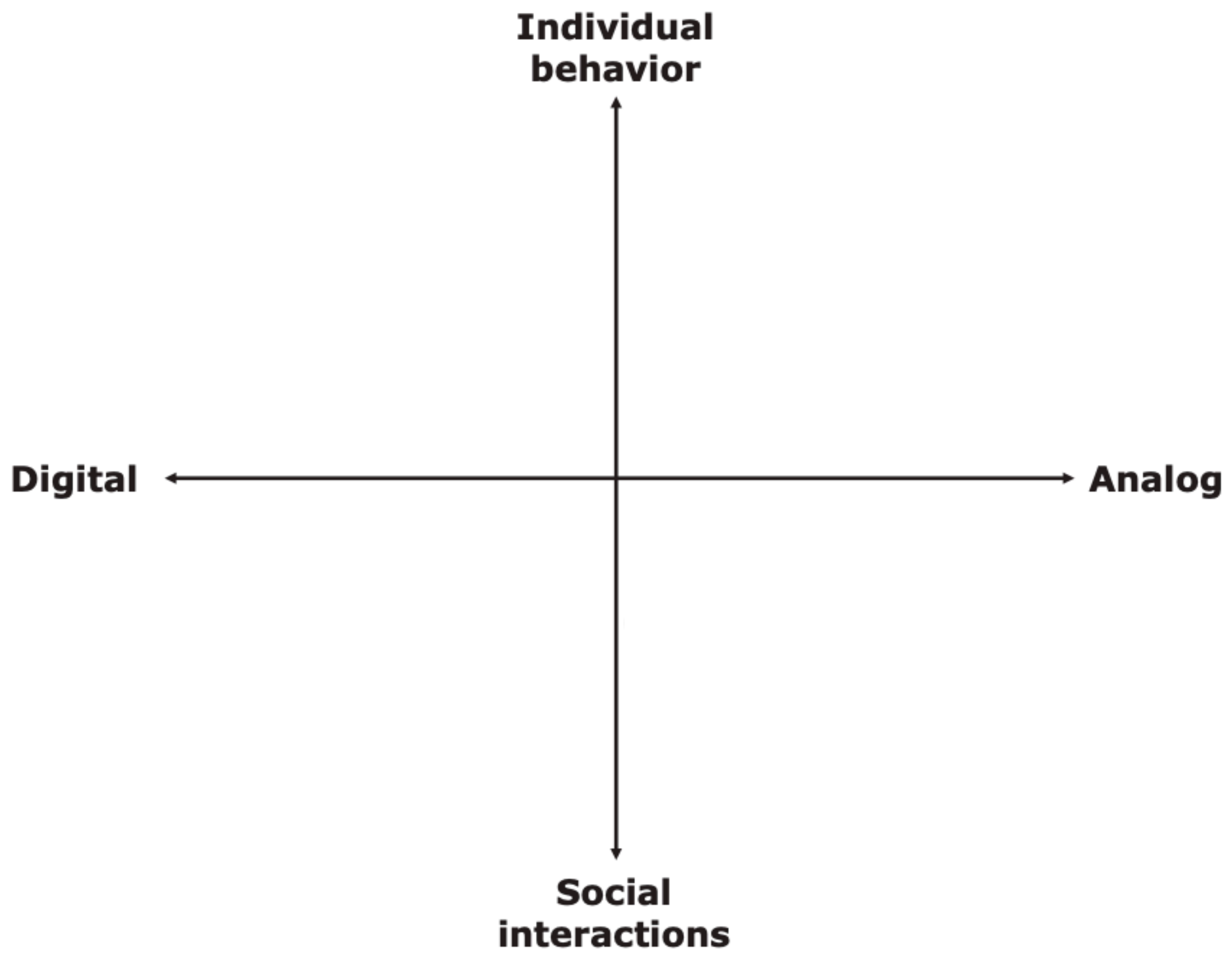
Social behavior

that can be studied using digital trace data

[Keusch and Kreuter 2021]

Social behavior

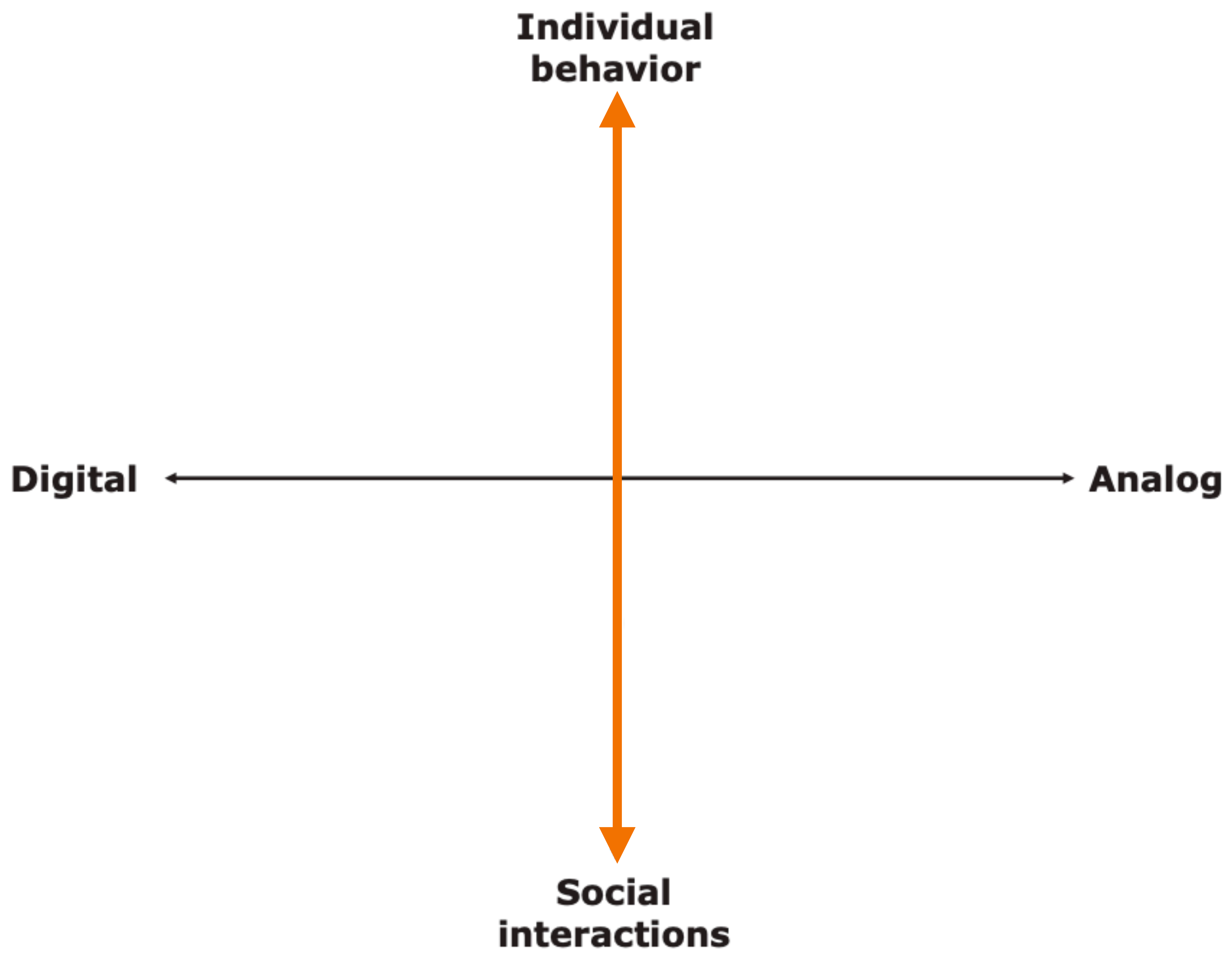
that can be studied using digital trace data



[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data

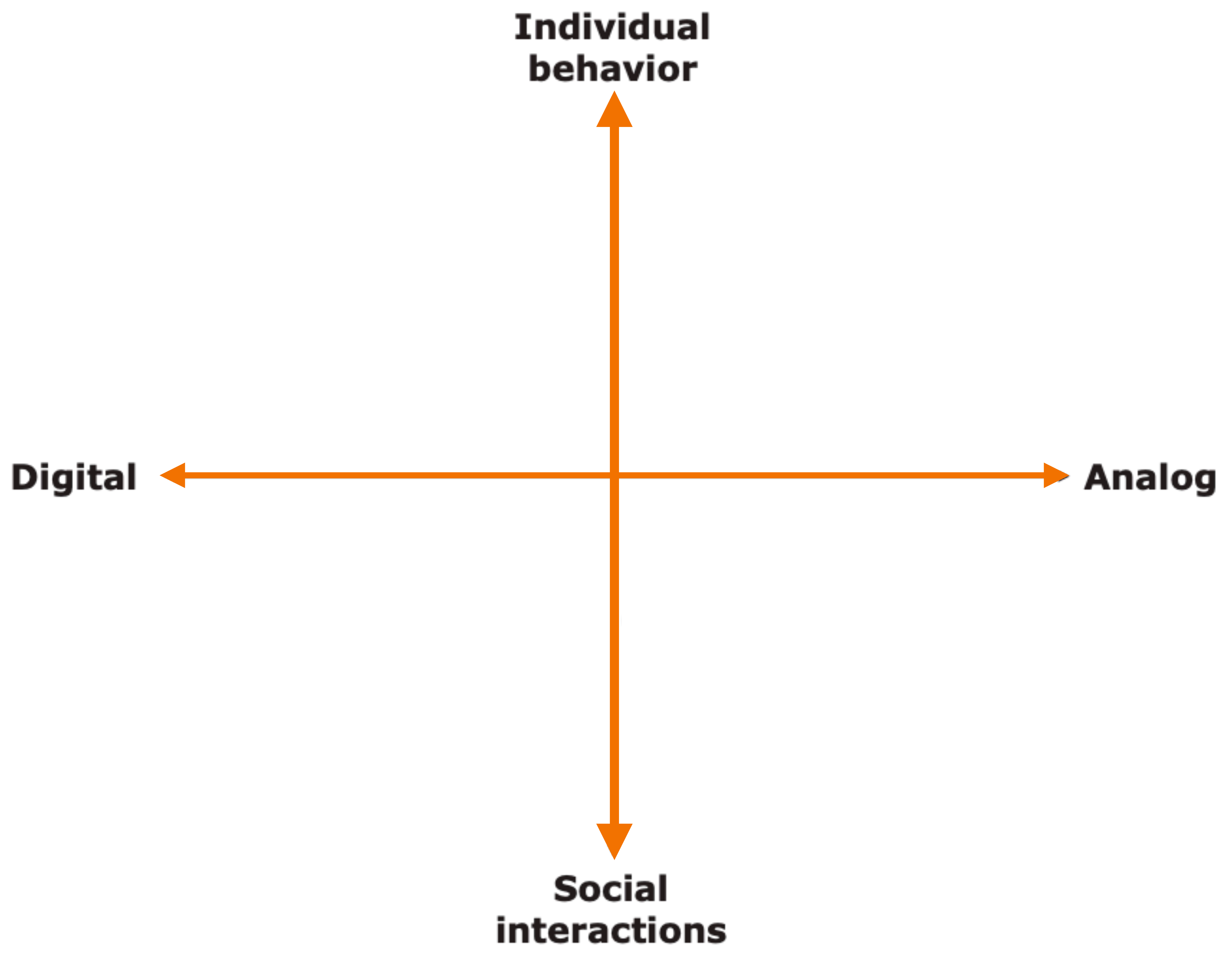


human

[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data



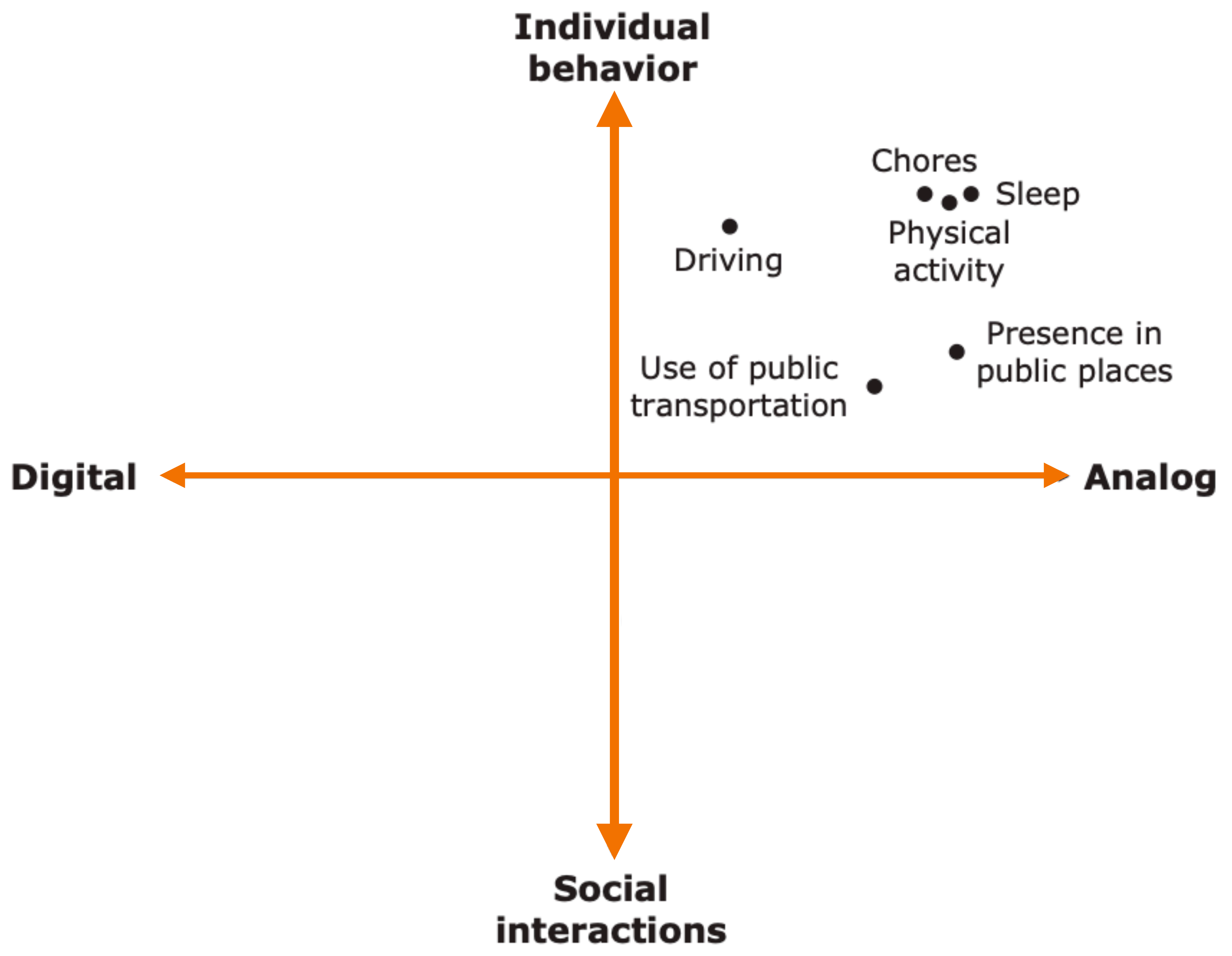
technology

human

[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data



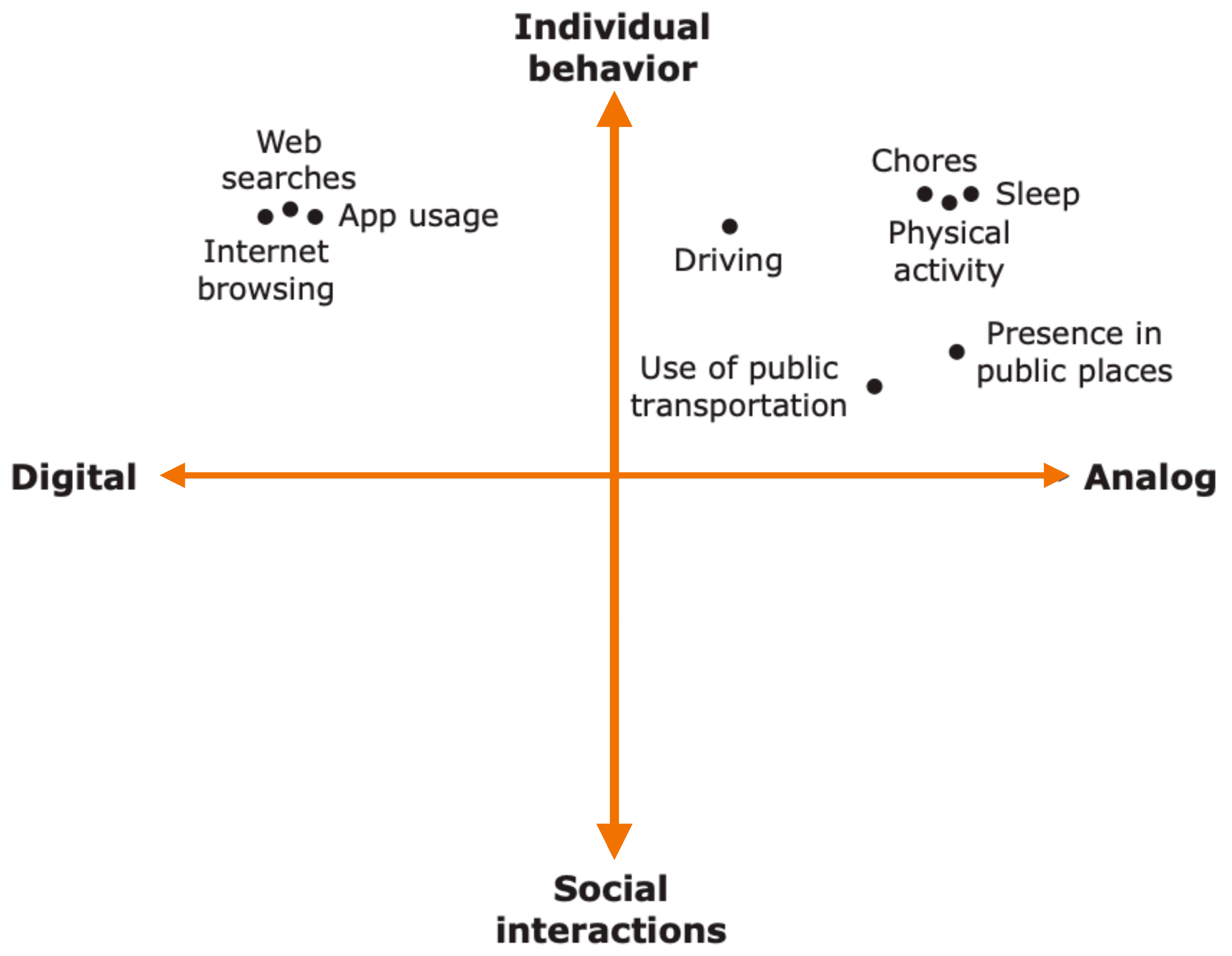
technology

human

[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data



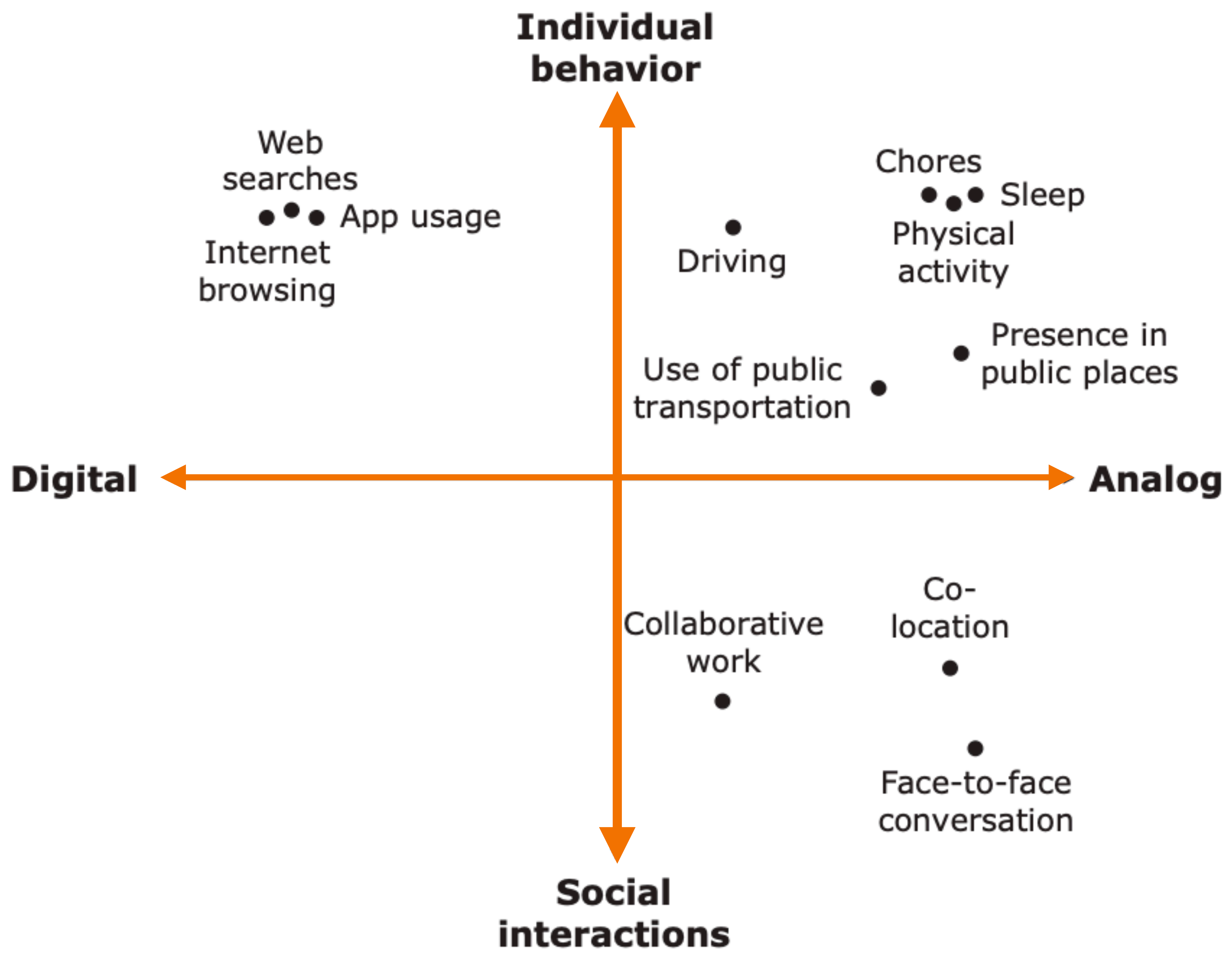
technology

human

[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data



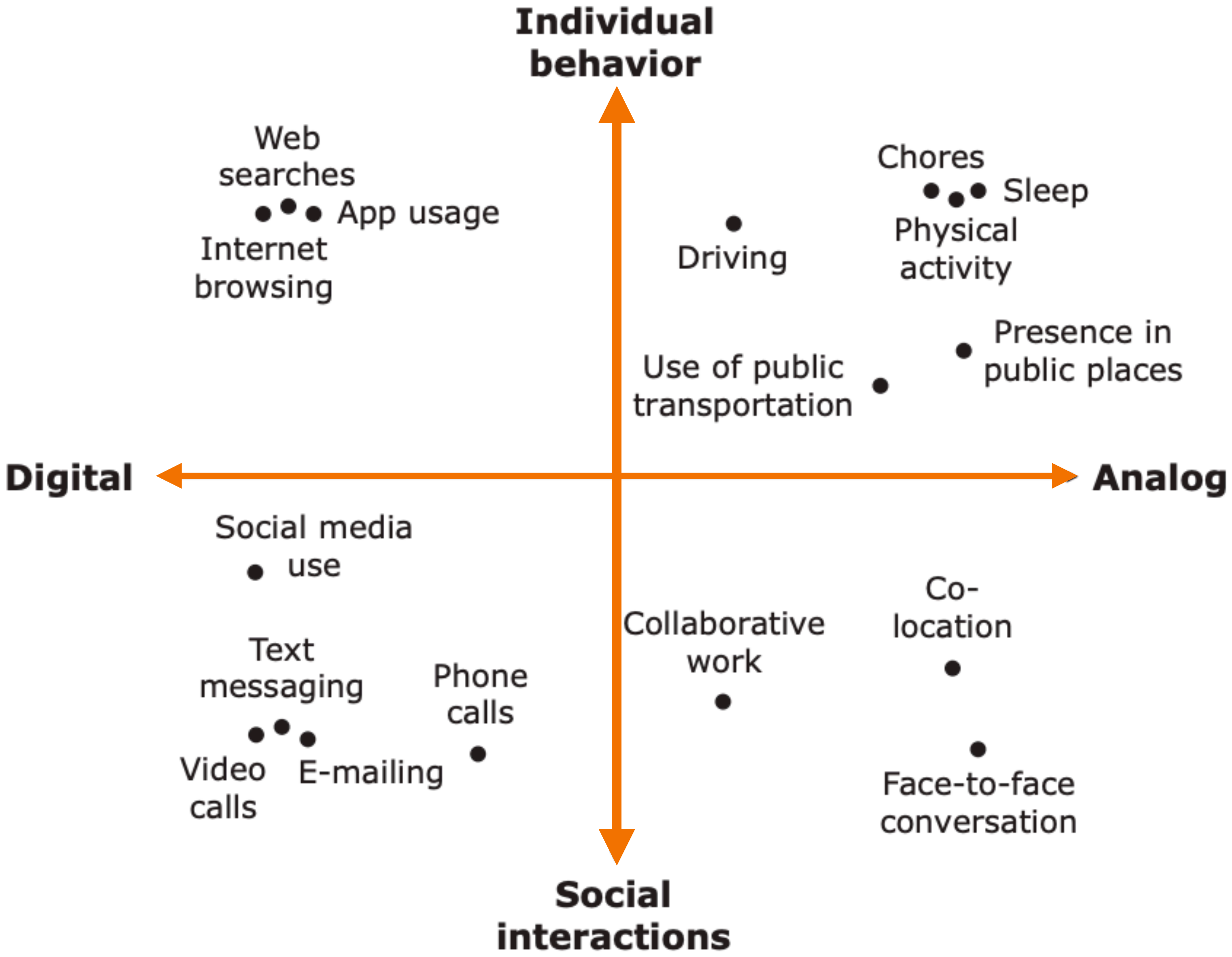
technology

human

[Keusch and Kreuter 2021]

Social behavior

that can be studied using digital trace data



technology

human

[Keusch and Kreuter 2021]

Examples

Human mobility using taxi, census, and Foursquare data

Examples

Human mobility using taxi, census, and Foursquare data

Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data

Authors: Lisette Espín Noboa, Florian Lemmerich, Philipp Singer, Markus Strohmaier (2016)

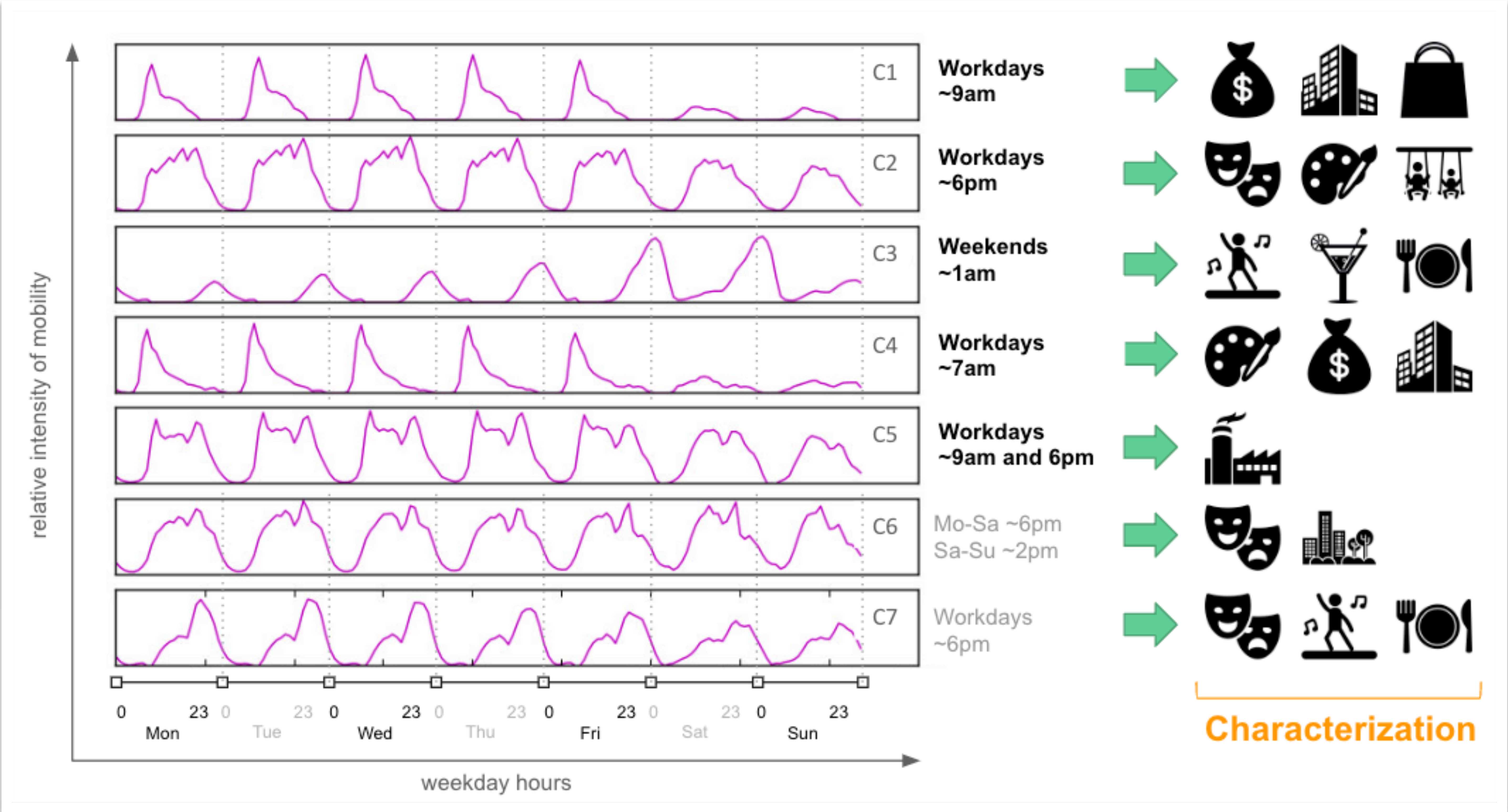
Examples

Human mobility using taxi, census, and Foursquare data

Discovering and Characterizing Mobility Patterns in Urban Spaces: A Study of Manhattan Taxi Data

Authors: Lisette Espín Noboa, Florian Lemmerich, Philipp Singer, Markus Strohmaier

(2016)



Examples

Human (online) navigation using Wikipedia

Examples

Human (online) navigation using Wikipedia

What Makes a Link Successful on Wikipedia?

(2017)

Dimitar Dimitrov*
GESIS – Leibniz Institute for the Social Sciences
dimitar.dimitrov@gesis.org

Philipp Singer*
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
philipp.singer@gesis.org

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
florian.lemmerich@gesis.org

Markus Strohmaier
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
markus.strohmaier@gesis.org

Examples

Human (online) navigation using Wikipedia

What Makes a Link Successful on Wikipedia?

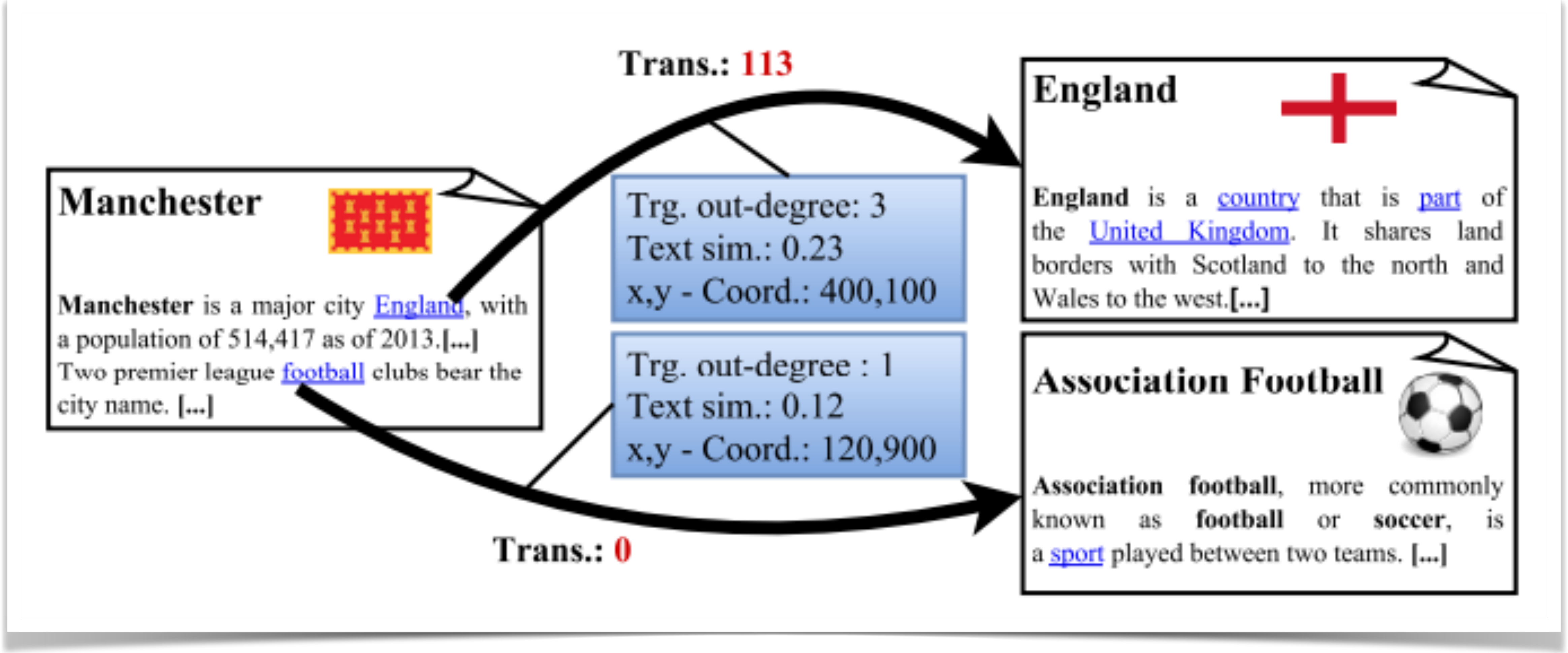
(2017)

Dimitar Dimitrov*
GESIS – Leibniz Institute for the Social Sciences
dimitar.dimitrov@gesis.org

Philipp Singer*
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
philipp.singer@gesis.org

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
florian.lemmerich@gesis.org

Markus Strohmaier
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
markus.strohmaier@gesis.org



Examples

Human (online) navigation using Wikipedia

What Makes a Link Successful on Wikipedia?

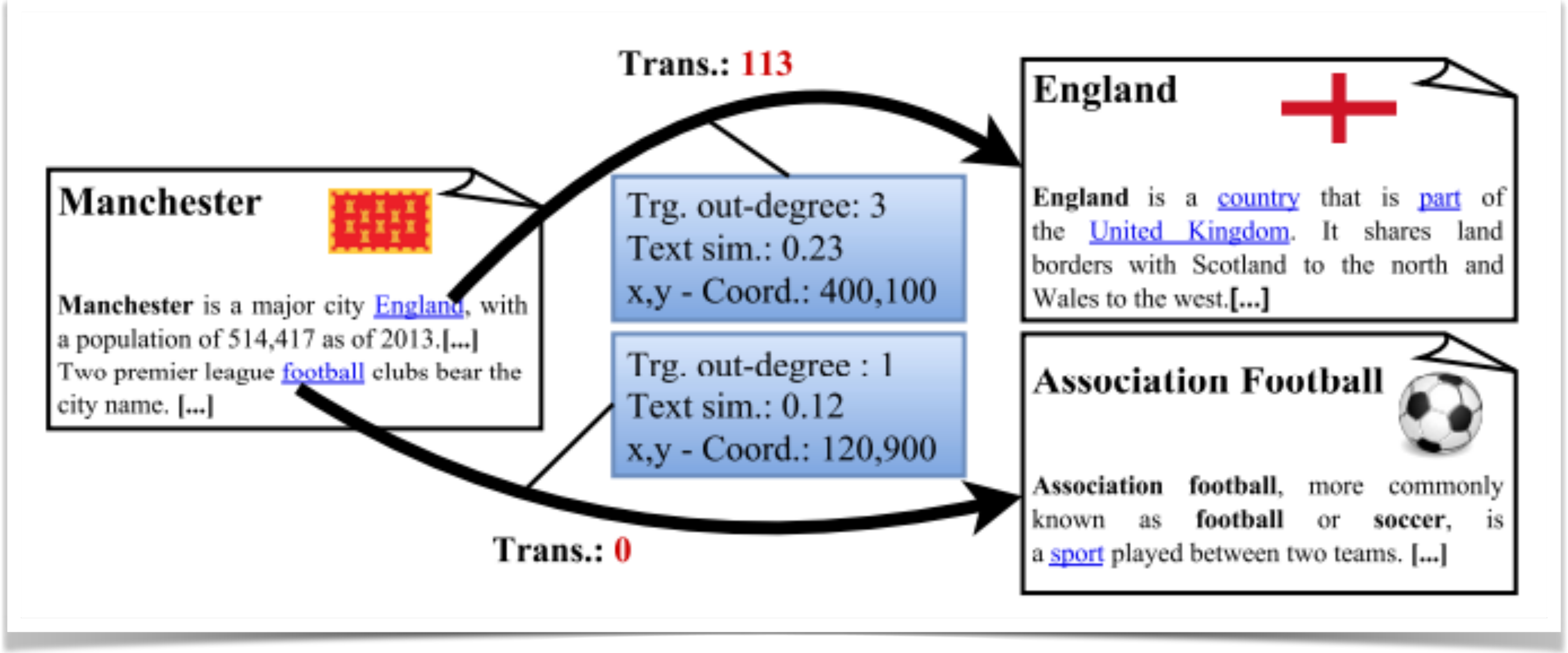
(2017)

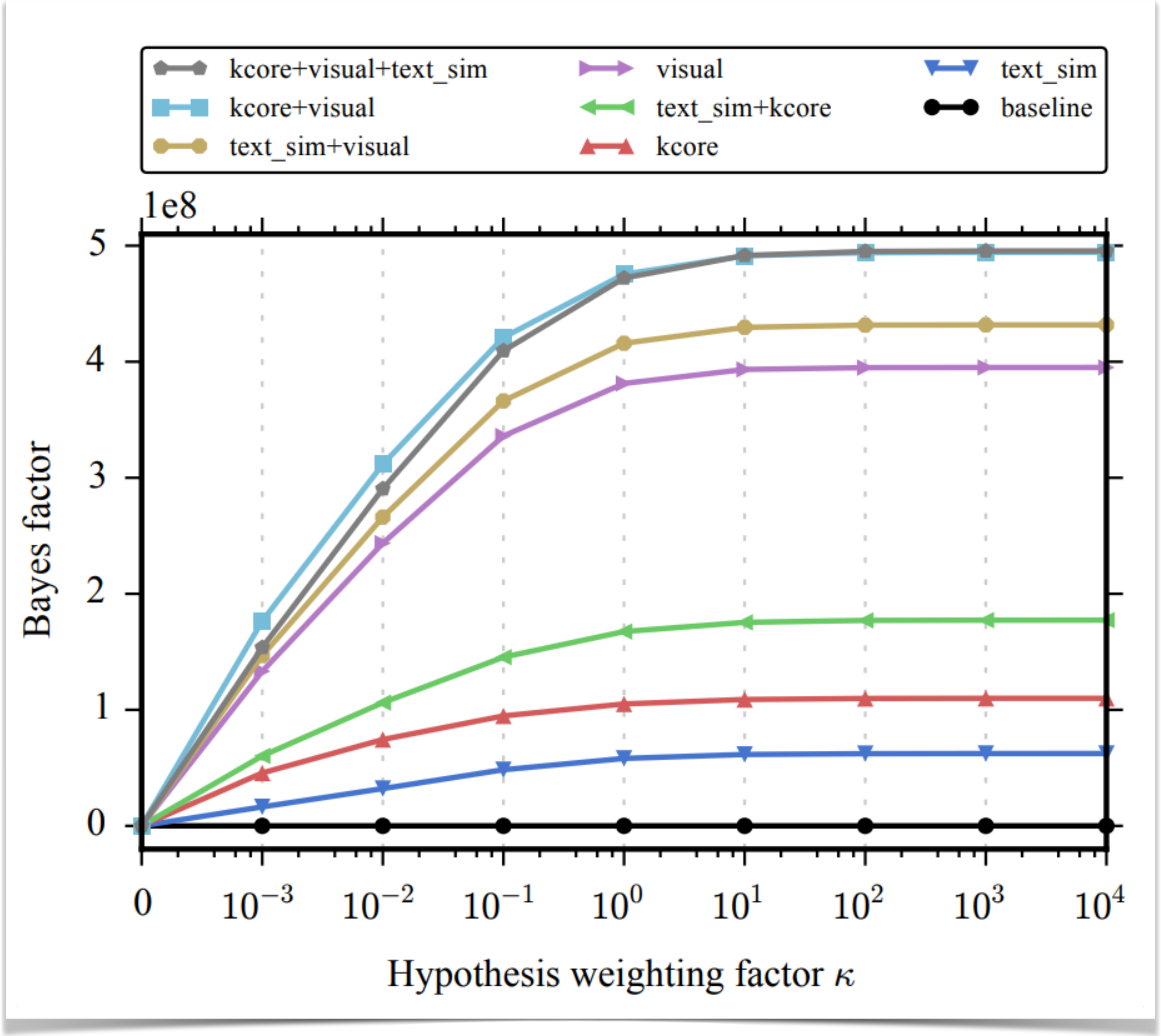
Dimitar Dimitrov*
GESIS – Leibniz Institute for the Social Sciences
dimitar.dimitrov@gesis.org

Philipp Singer*
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
philipp.singer@gesis.org

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
florian.lemmerich@gesis.org

Markus Strohmaier
GESIS – Leibniz Institute for the Social Sciences
& University of Koblenz-Landau
markus.strohmaier@gesis.org





Examples

Human (online) navigation using Wikipedia

Examples

Human (online) navigation using Wikipedia

Why We Read Wikipedia

(2017)

Philipp Singer[*1], Florian Lemmerich[*1], Robert West[†2],
Leila Zia[3], Ellery Wulczyn[3], Markus Strohmaier[1], Jure Leskovec[4]

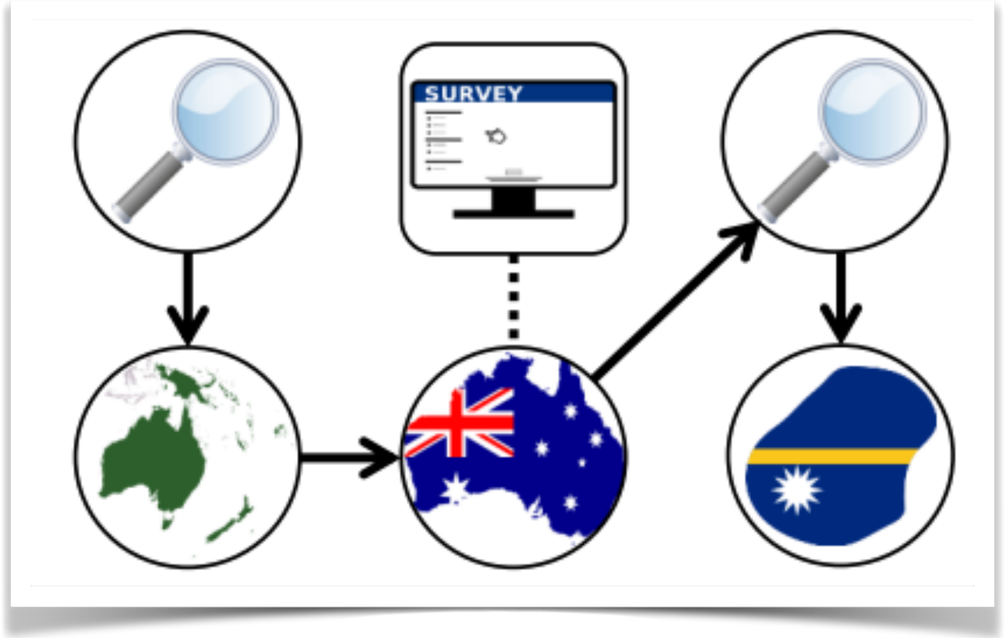
Examples

Human (online) navigation using Wikipedia

Why We Read Wikipedia

(2017)

Philipp Singer*[1], Florian Lemmerich*[1], Robert West†[2],
Leila Zia[3], Ellery Wulczyn[3], Markus Strohmaier[1], Jure Leskovec[4]



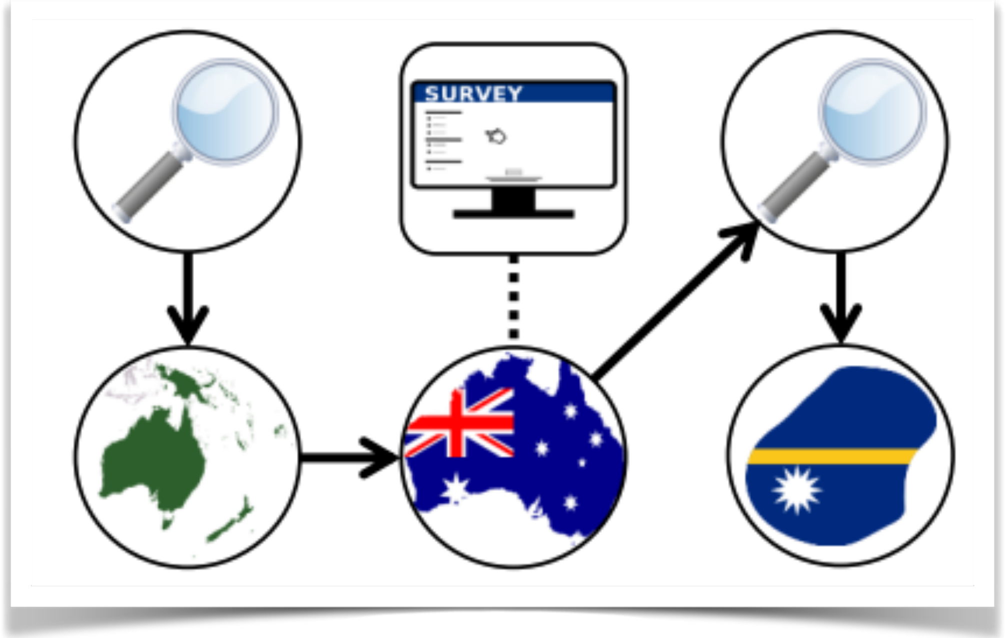
Examples

Human (online) navigation using Wikipedia

Why We Read Wikipedia

(2017)

Philipp Singer[*1], Florian Lemmerich[*1], Robert West[†2],
Leila Zia[3], Ellery Wulczyn[3], Markus Strohmaier[1], Jure Leskovec[4]



This paper uses both survey responses and web request log (navigation)

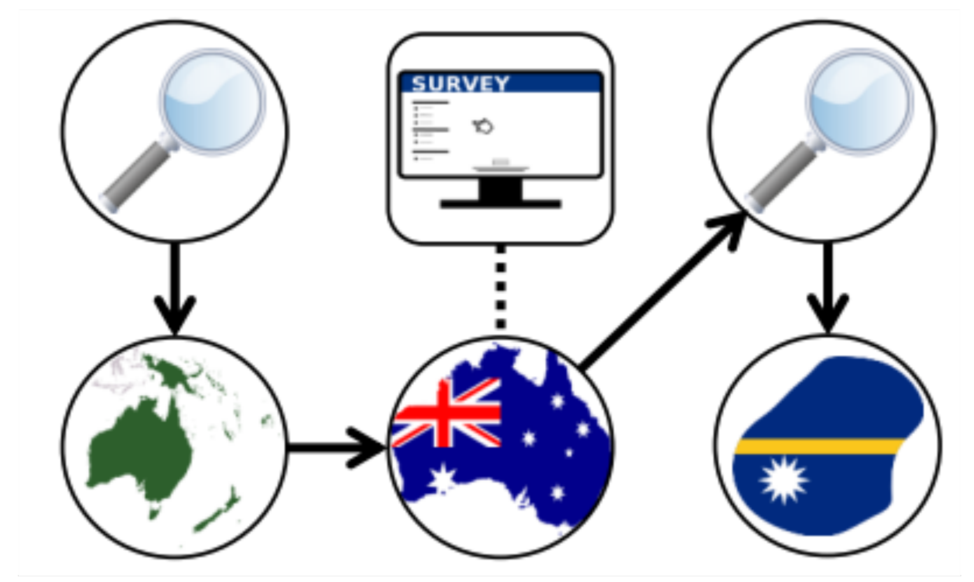
Examples

Human (online) navigation using Wikipedia

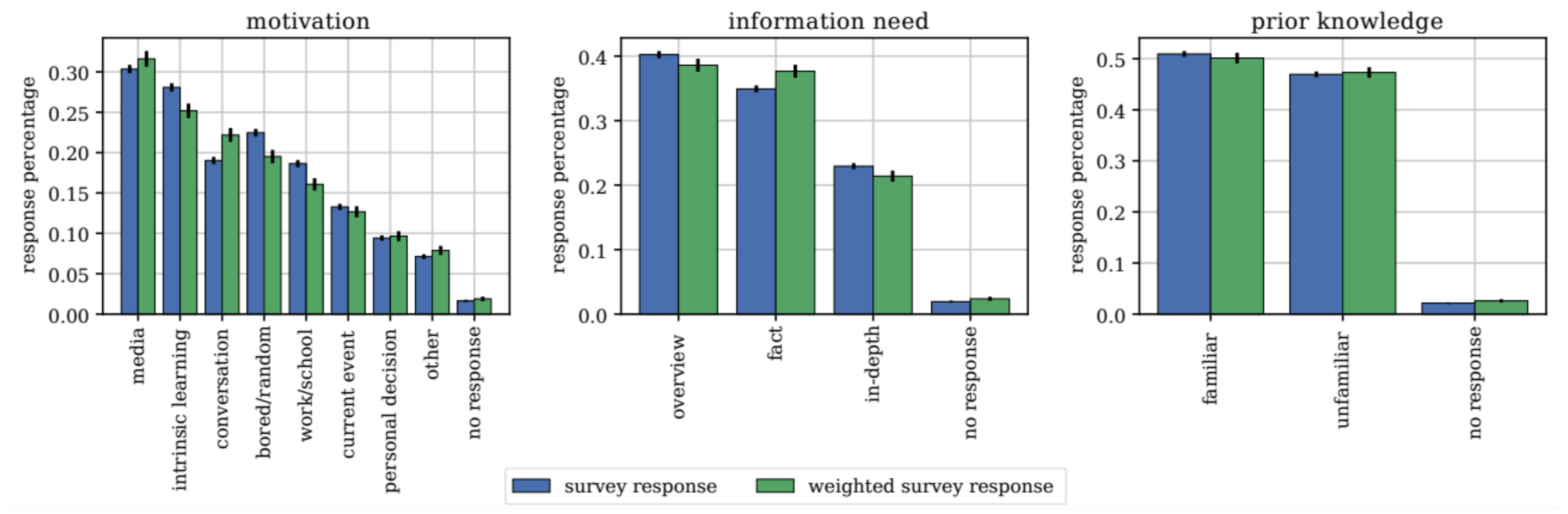
Why We Read Wikipedia

(2017)

Philipp Singer[*1], Florian Lemmerich[*1], Robert West[†2],
Leila Zia[3], Ellery Wulczyn[3], Markus Strohmaier[1], Jure Leskovec[4]



This paper uses both survey responses and web request log (navigation)



Examples

Migration patterns from online data

Examples

Migration patterns from online data

Using Facebook and LinkedIn Data to Study International Mobility

(2023)

Carolina Coimbra Vieira[1], Masoomali Fatehkia[2], Kiran Garimella[3], Ingmar Weber[2] and Emilio Zagheni[1]

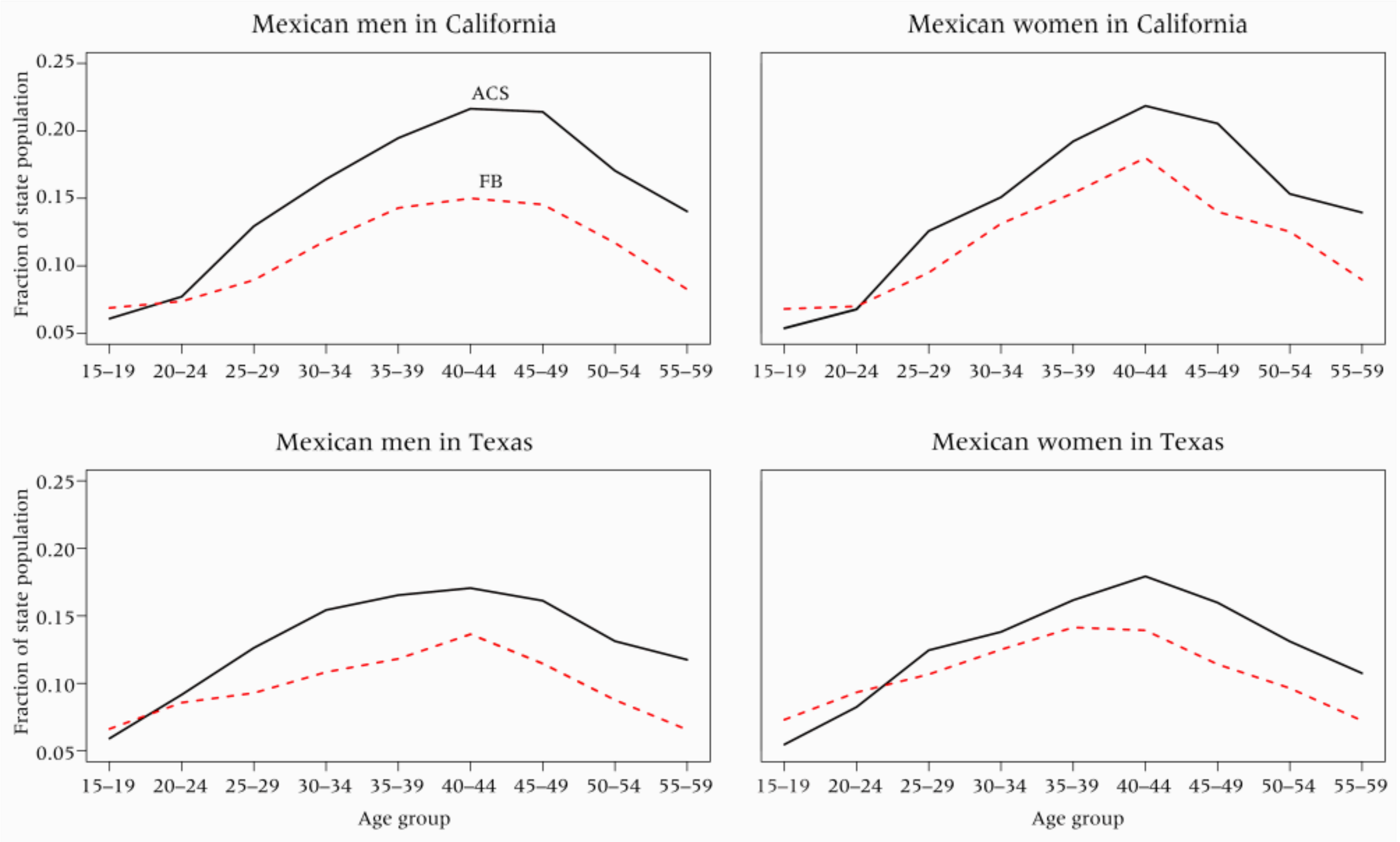
Examples

Migration patterns from online data

Using Facebook and LinkedIn Data to Study International Mobility

(2023)

Carolina Coimbra Vieira[1], Masoomali Fatehkia[2], Kiran Garimella[3], Ingmar Weber[2] and Emilio Zagheni[1]



SOURCES: American Community Survey (ACS 2014); Facebook Adverts Manager.

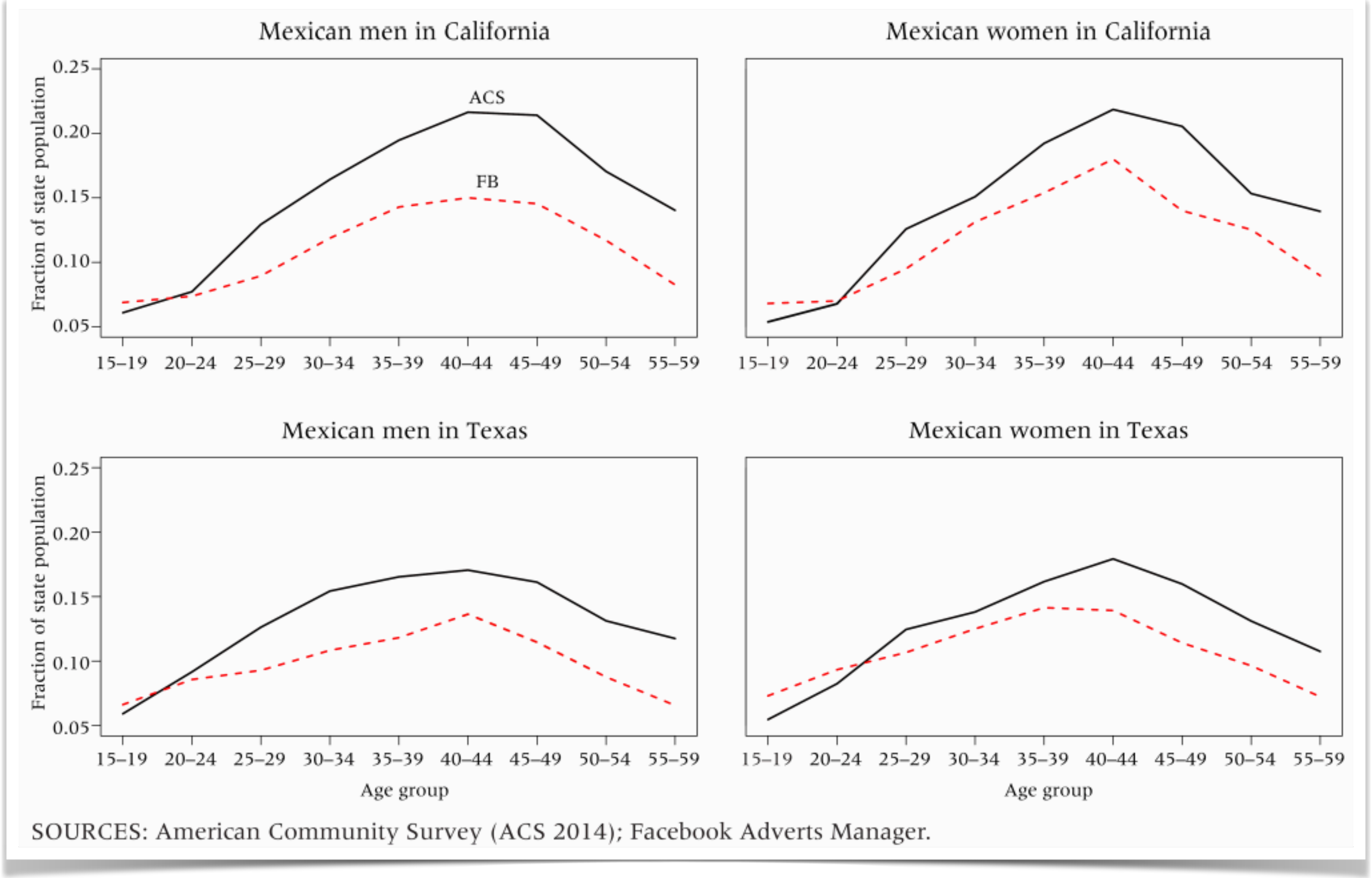
Fraction of men or women based on survey and Facebook

Examples

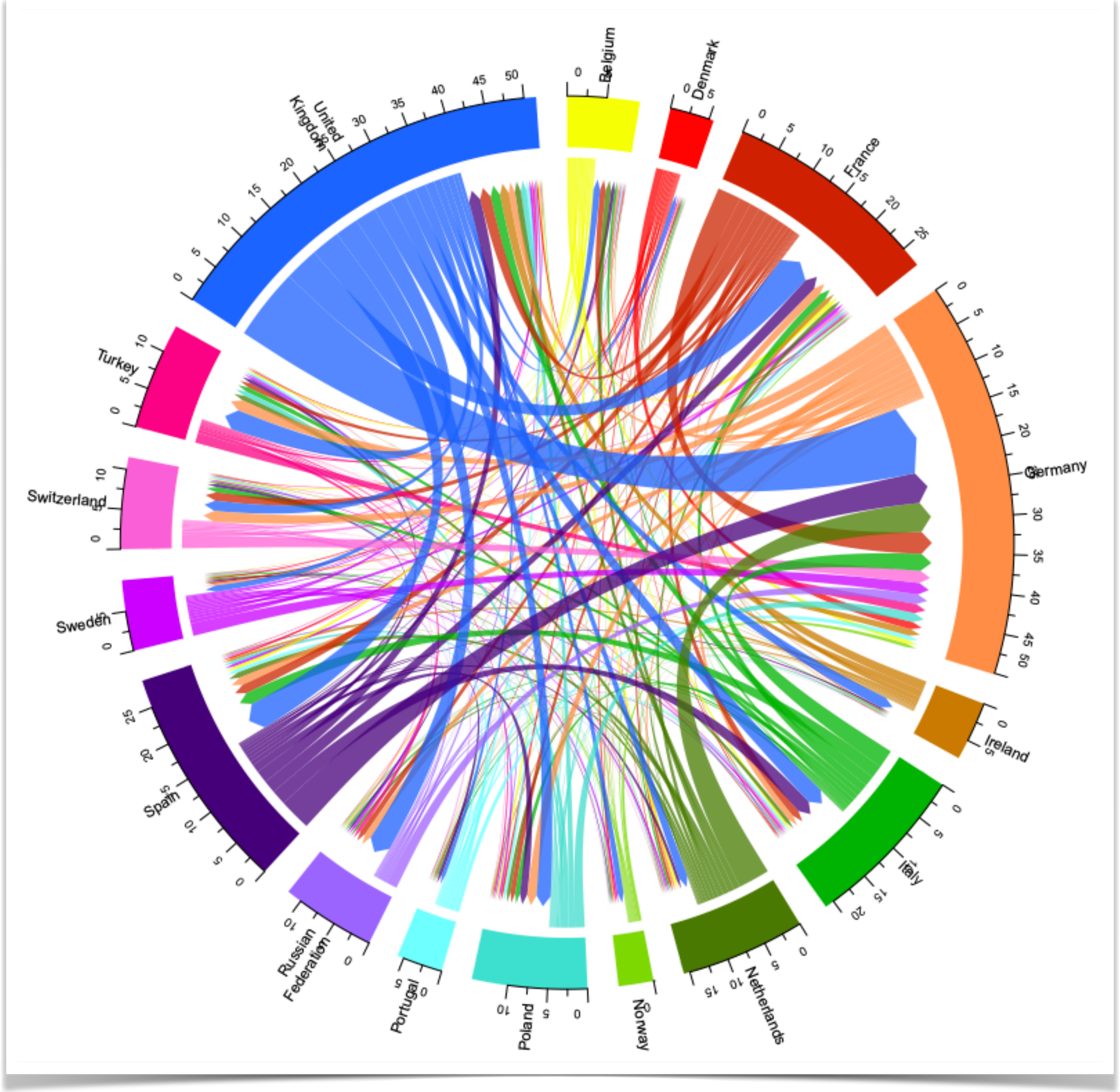
Migration patterns from online data

Using Facebook and LinkedIn Data to Study International Mobility
(2023)

Carolina Coimbra Vieira[1], Masoomali Fatehkia[2], Kiran Garimella[3], Ingmar Weber[2] and Emilio Zagheni[1]



Fraction of men or women based on survey and Facebook



Migrants in 1000's (who studied in country x and live in country y)

Examples

Inferring poverty from the sky and the Web

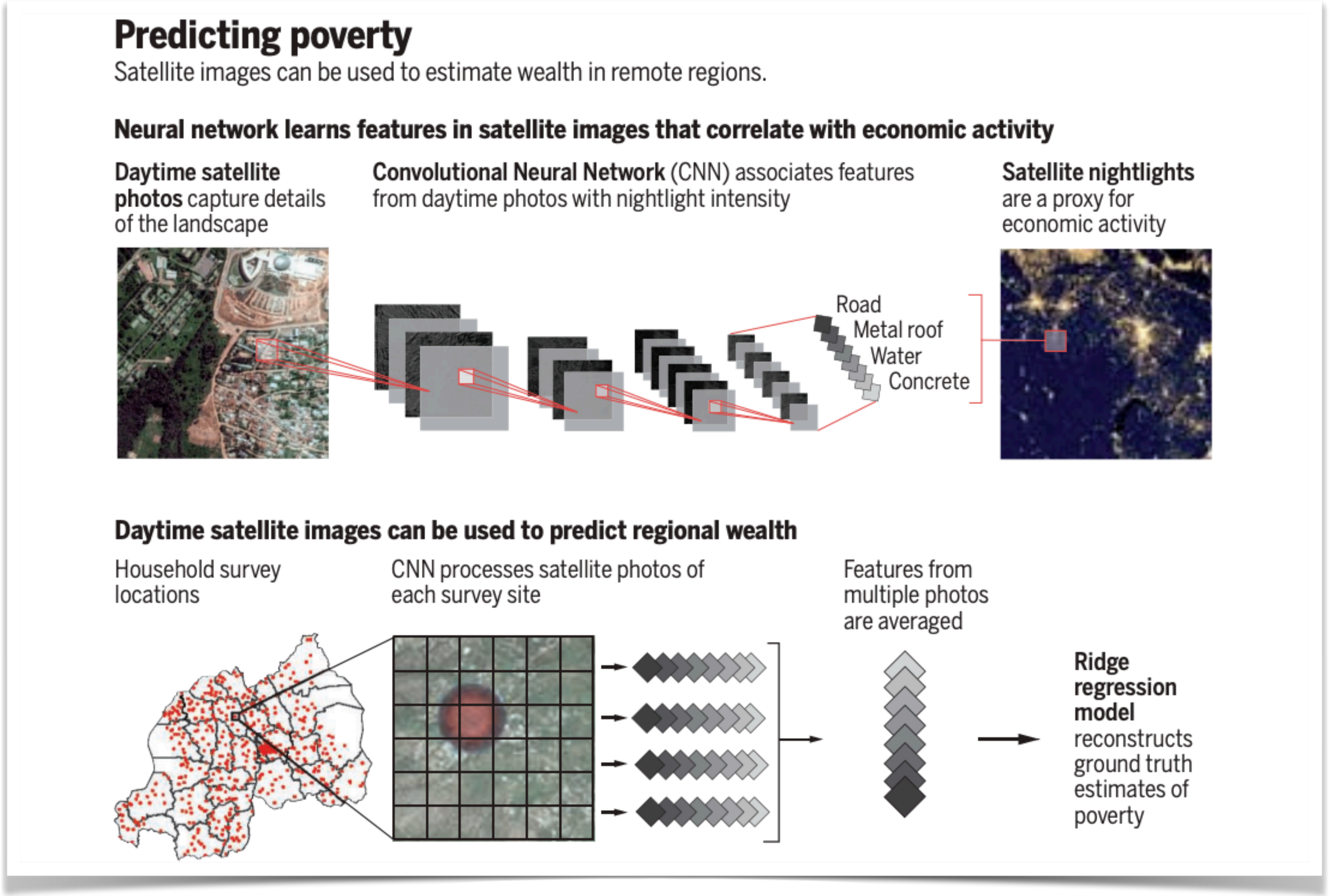
Examples

Inferring poverty from the sky and the Web

Fighting poverty with data

Machine learning algorithms measure and target poverty

By **Joshua Evan Blumenstock** (2016)



Examples

Inferring poverty from the sky and the Web

Fighting poverty with data

Machine learning algorithms measure and target poverty

By **Joshua Evan Blumenstock**

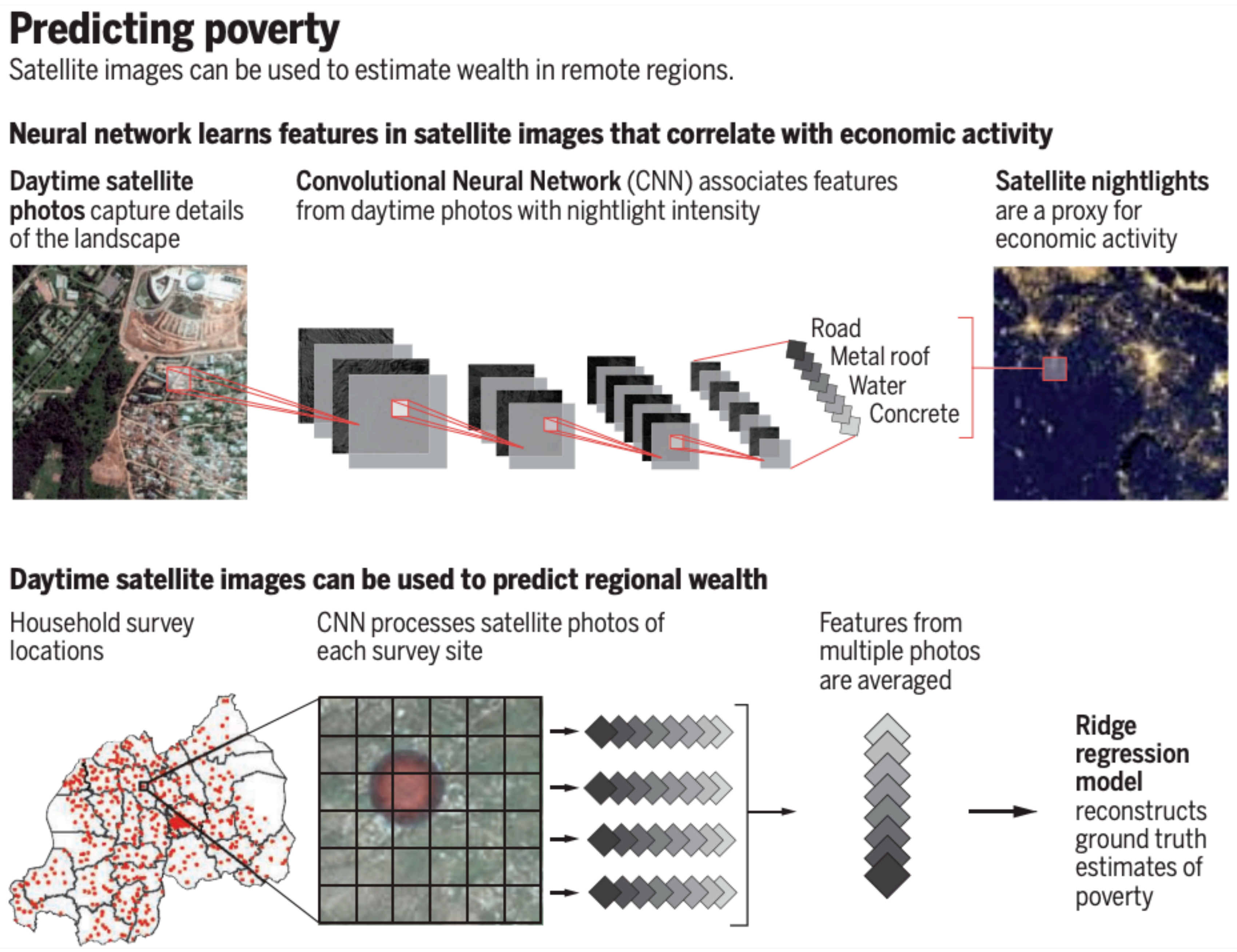
(2016)

Interpreting wealth distribution via poverty map inference using multimodal data

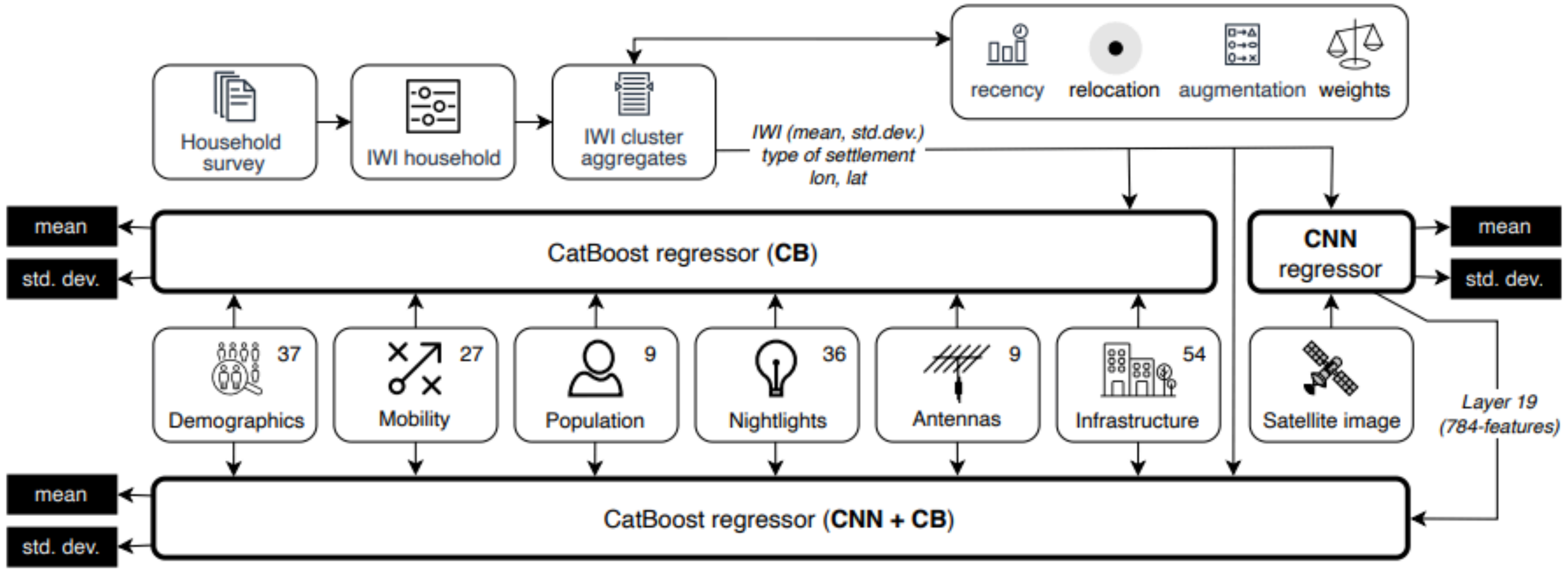
Lisette Espín-Noboa
EspinL@ceu.edu
Central European University
Complexity Science Hub Vienna

János Kertész
KerteszJ@ceu.edu
Central European University
Complexity Science Hub Vienna

Márton Karsai
KarsaiM@ceu.edu
Central European University
Rènyi Institute of Mathematics



https://vis.csh.ac.at/poverty-maps



(2023)

Examples

Health and social media

Examples

Health and social media

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**

Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

(2013)

Examples

Health and social media

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

(2013)

Having a job again makes me happy. Less time to be depressed and eat all day while watching sad movies.
“Are you okay?” Yes…. I understand that I am upset and hopeless and nothing can help me… I’m okay… but I am not alright
“empty” feelings I WAS JUST TALKING ABOUT HOW I I HAVE EMOTION OH MY GOODNESS I FEEL AWFUL
I want someone to hold me and be there for me when I’m sad.
Reloading twitter till I pass out. *lonely* *anxious* *butthurt* *frustrated* *dead*

Table 2: Example posts from users in the depression class.

Examples

Health and social media

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**

Microsoft Research, Redmond WA 98052

{munmund, mgamon, counts, horvitz}@microsoft.com

(2013)

Having a job again makes me happy. Less time to be depressed and eat all day while watching sad m
"Are you okay?" Yes…. I understan less and nothing can help me… I'm
"empty" feelings I WAS JUST TAL HAVE EMOTION OH MY GOOD
I want someone to hold me and be th
Reloading twitter till I pass out. *lon *frustrated* *dead*

Table 2: Example posts from user

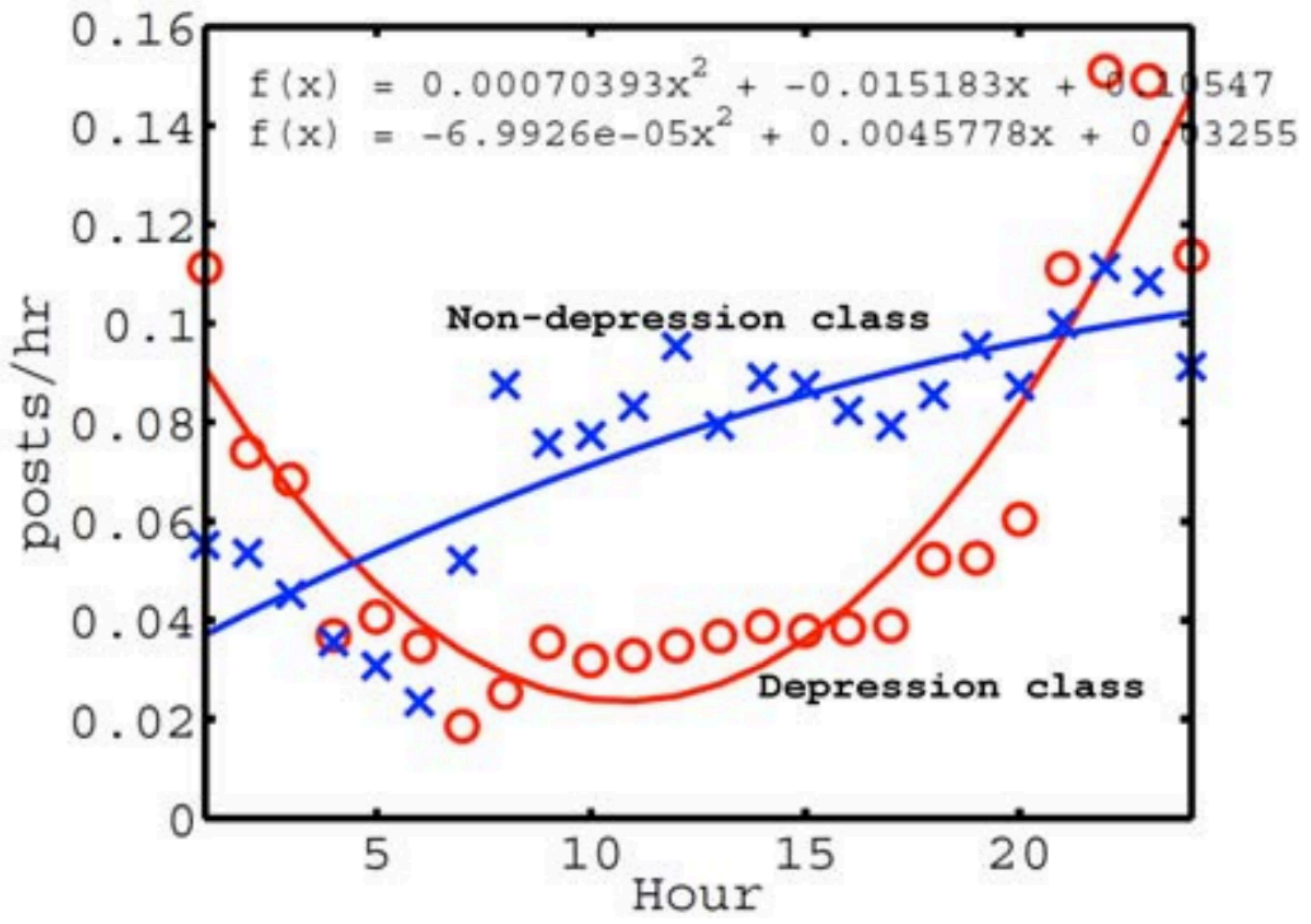


Figure 2: Diurnal trends (i.e. mean number of posts made hourly throughout a day) for the two classes. The line plots correspond to least squares fit of the trends.

Examples

Health and social media

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

(2013)

Having a job again makes me happy. Less time to be depressed and eat all day while watching sad m
"Are you okay?" Yes…. I understan less and nothing can help me… I'm
"empty" feelings I WAS JUST TAL HAVE EMOTION OH MY GOODN
I want someone to hold me and be th
Reloading twitter till I pass out. *lon *frustrated* *dead*

Table 2: Example posts from user

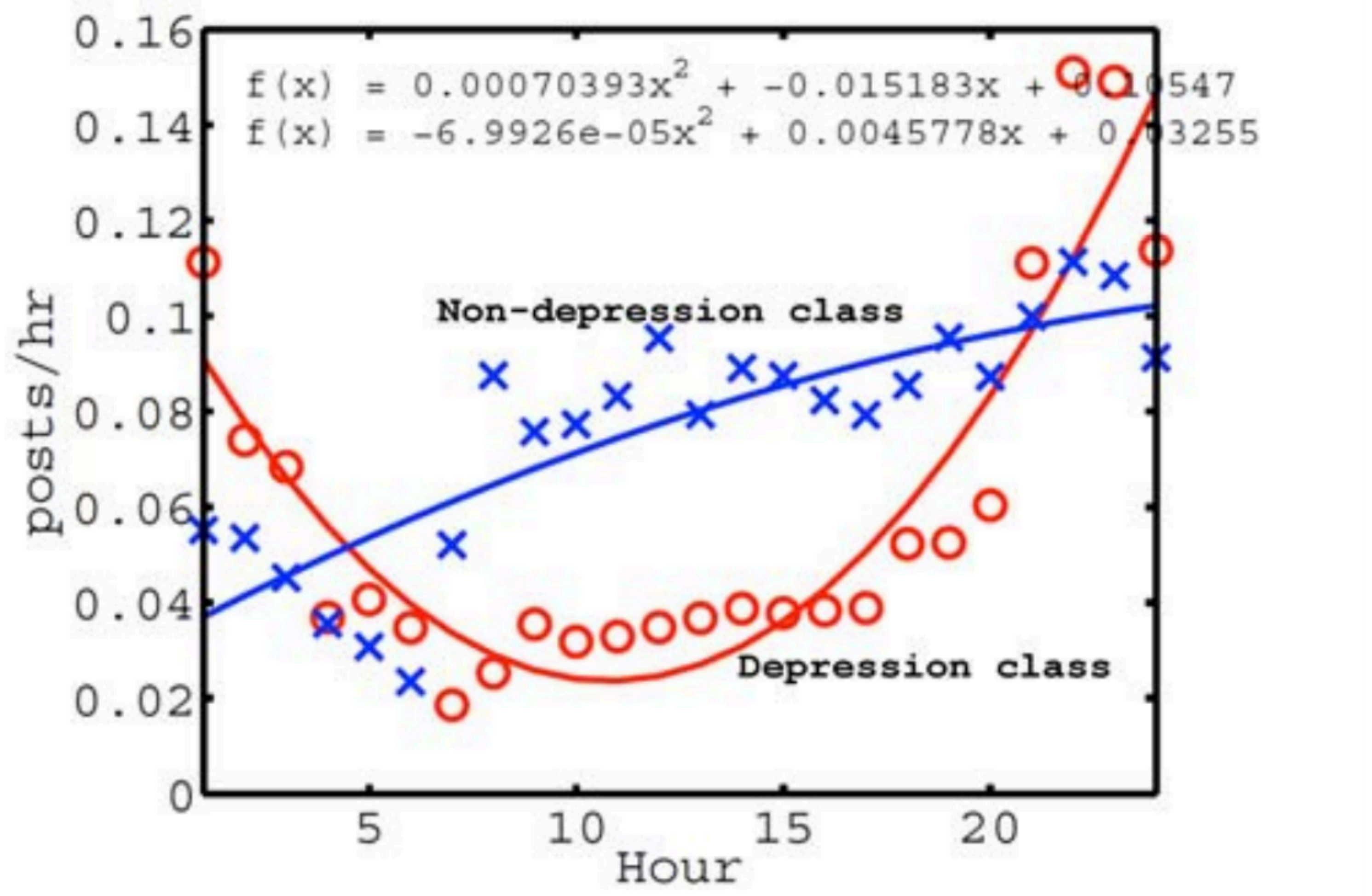


Figure 2: Diurnal trends (i.e. mean number of posts made hourly throughout a day) for the two classes. The line plots correspond to least squares fit of the trends.

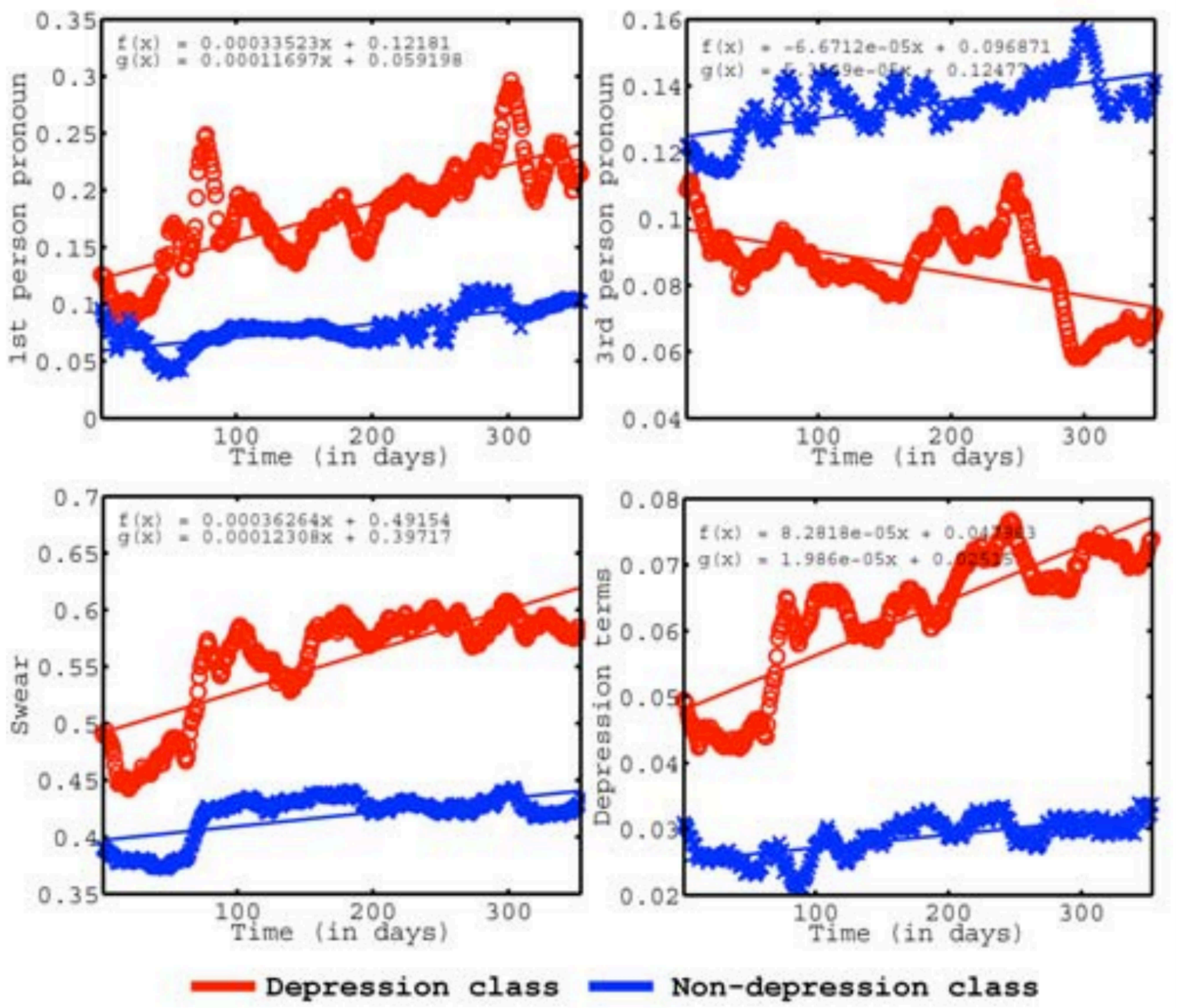


Figure 3. Trends for various features corresponding to the depression and non-depression classes. Line plots correspond to least squares fit.

Examples

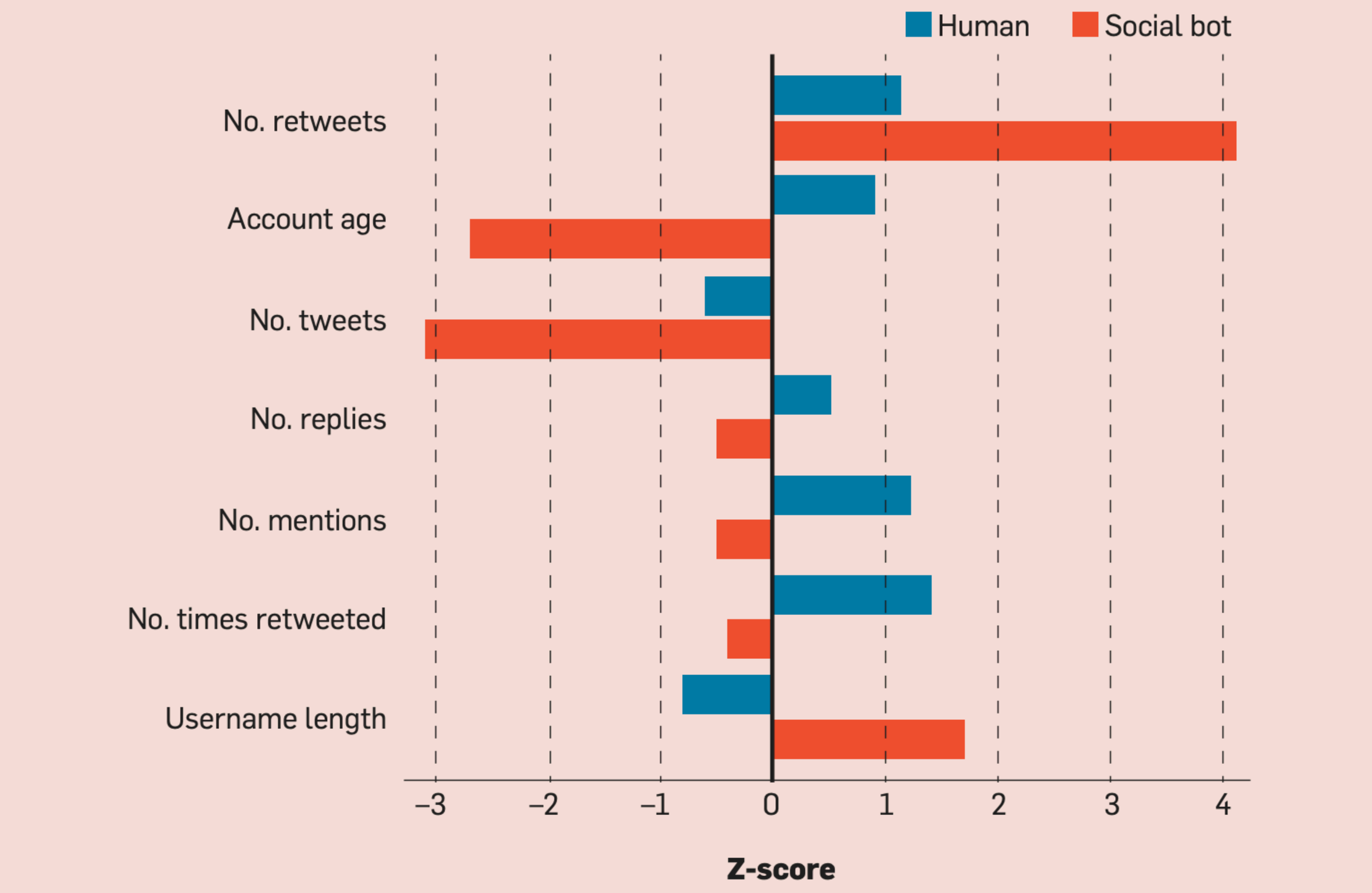
Bots and fake news in social media

BY EMILIO FERRARA, ONUR VAROL, CLAYTON DAVIS, FILIPPO MENCZER, AND ALESSANDRO FLAMMINI (2016)

The Rise of Social Bots

Figure 2. User behaviors that best discriminate social bots from humans.

Social bots retweet more than humans and have longer user names, while they produce fewer tweets, replies and mentions, and they are retweeted less than humans. Bot accounts also tend to be more recent.



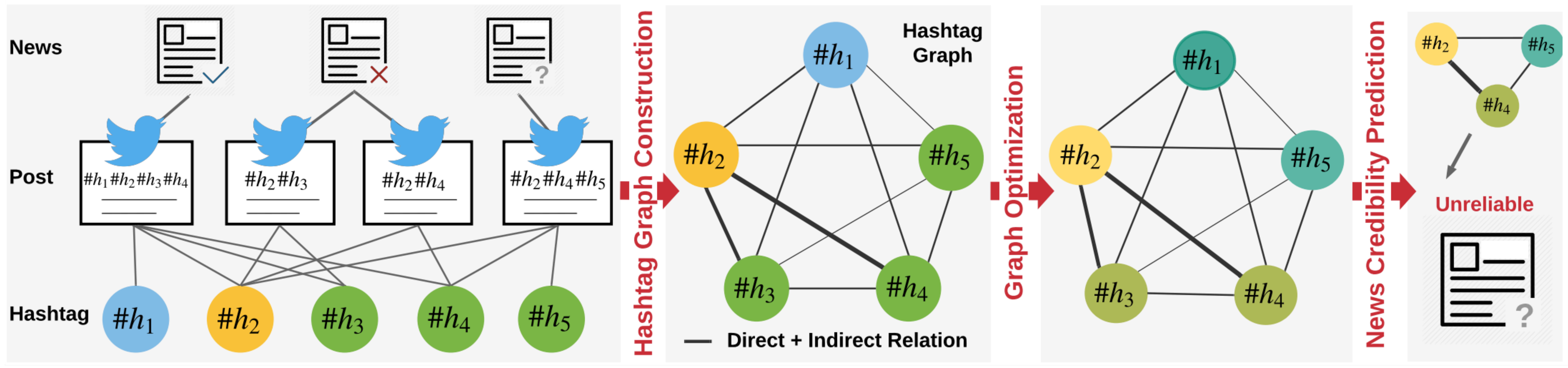
From Fake News to #FakeNews: Mining Direct and Indirect Relationships among Hashtags for Fake News Detection

Xinyi Zhou, zhouxinyi@data.syr.edu, Syracuse University, U.S.A

Reza Zafarani, reza@data.syr.edu, Syracuse University, U.S.A

Emilio Ferrara, emiliofe@usc.edu, University of Southern California, U.S.A

(2022)



Examples

Polarization

Examples

Polarization

The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

Lada Adamic
HP Labs
1501 Page Mill Road
Palo Alto, CA 94304
lada.adamic@hp.com

Natalie Glance
Intelliseek Applied Research Center
5001 Baum Blvd.
Pittsburgh, PA 15217
nglance@intelliseek.com

4 March 2005

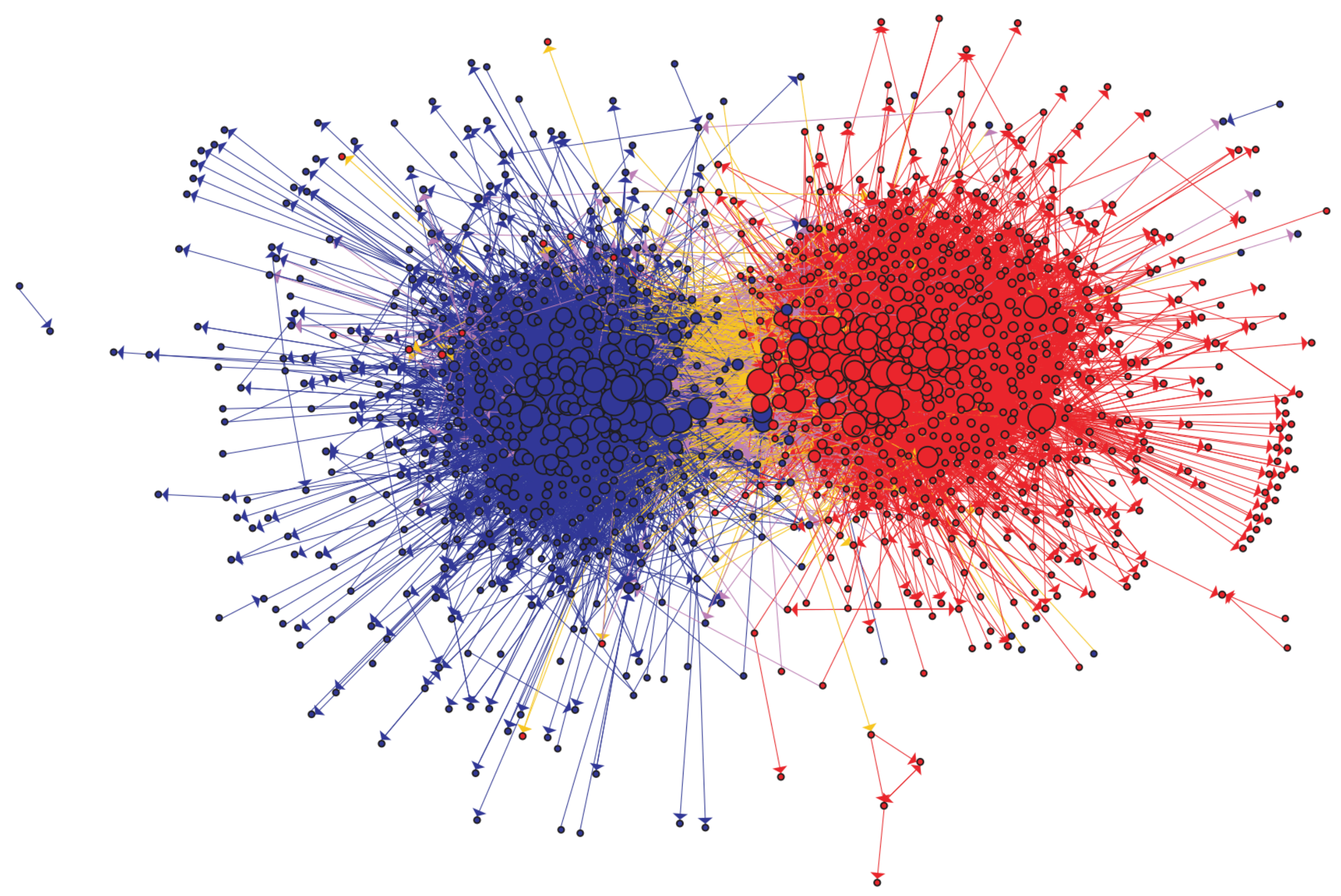


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Examples

Polarization

The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

Lada Adamic
HP Labs
1501 Page Mill Road
Palo Alto, CA 94304
lada.adamic@hp.com

Natalie Glance
Intelliseek Applied Research Center
5001 Baum Blvd.
Pittsburgh, PA 15217
nglance@intelliseek.com

4 March 2005

Aalto University publication series
DOCTORAL DISSERTATIONS 20/2018

Polarization on Social Media

Kiran Garimella

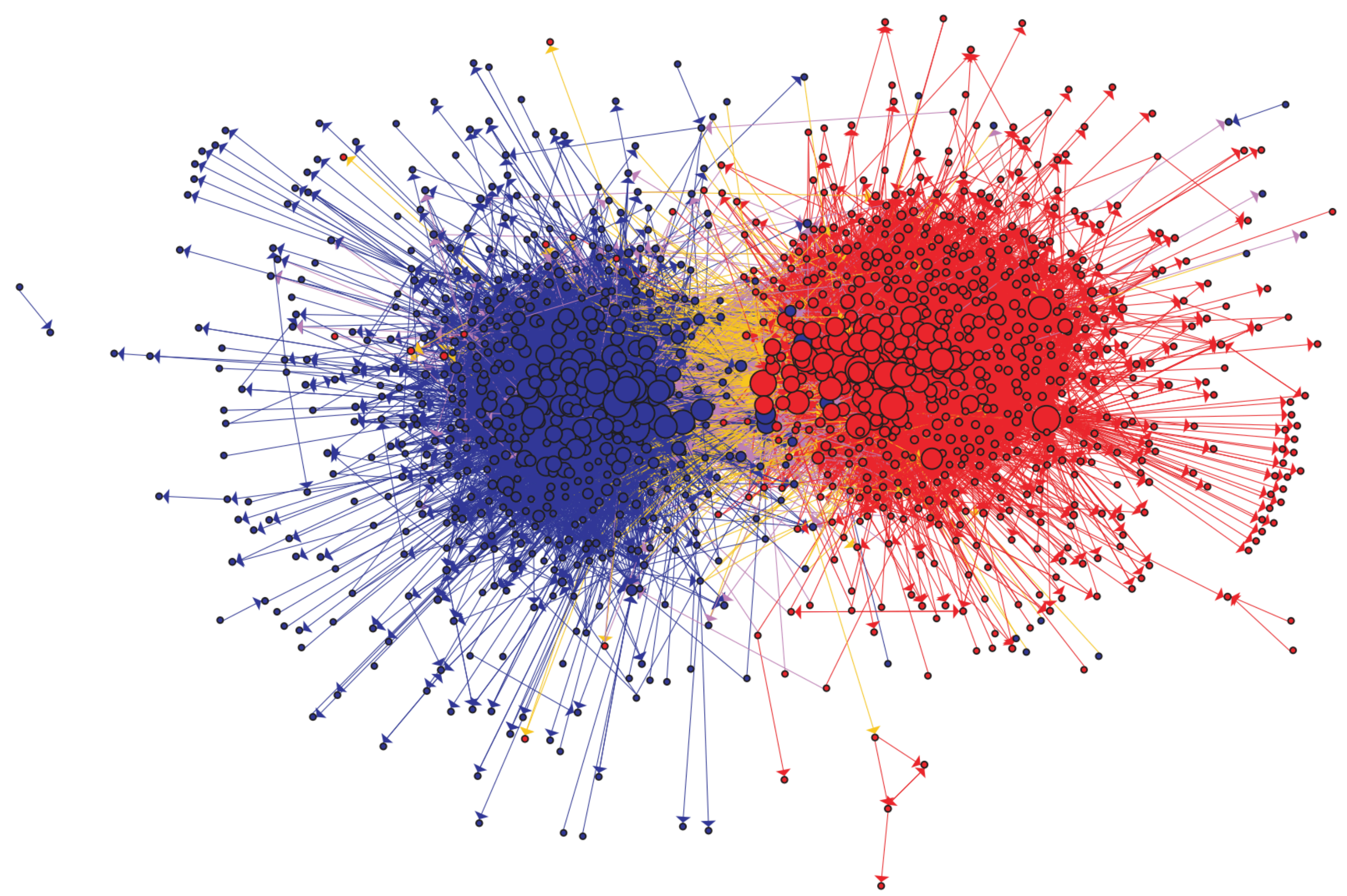
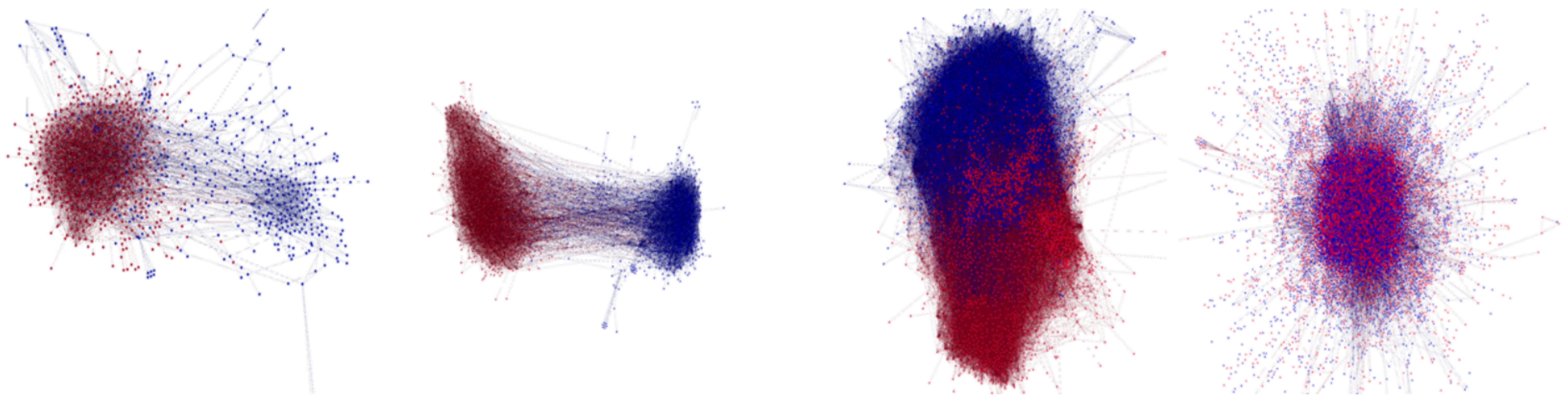


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.



(a) **(b)** **(c)** **(d)**

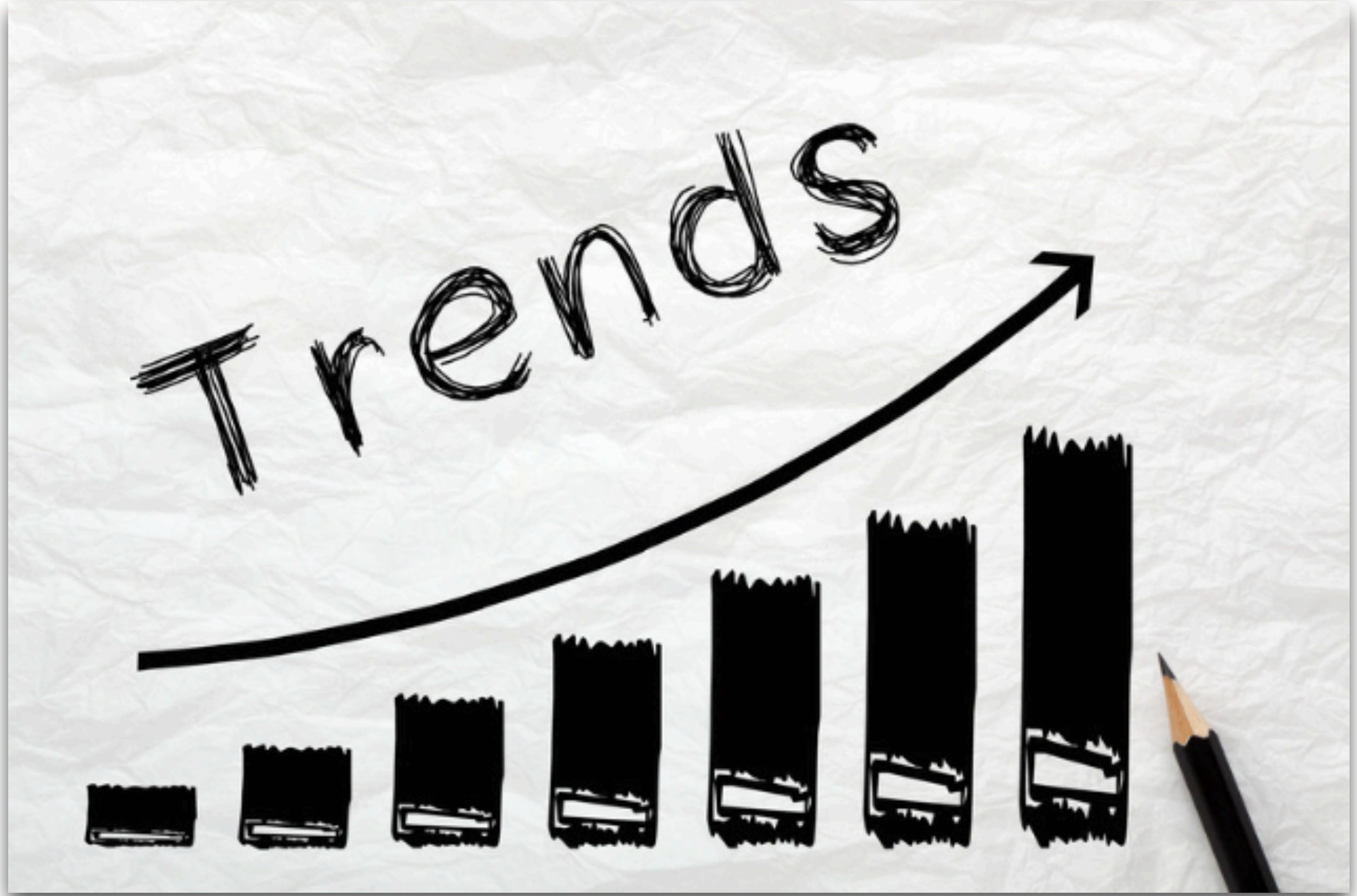
Figure 4.4. Sample follow graphs for polarized topics, (a) #beefban, (b) #russia_march, and non-polarized topics, (c) #sxsw, (d) #germanwings.

Outline

Today's class

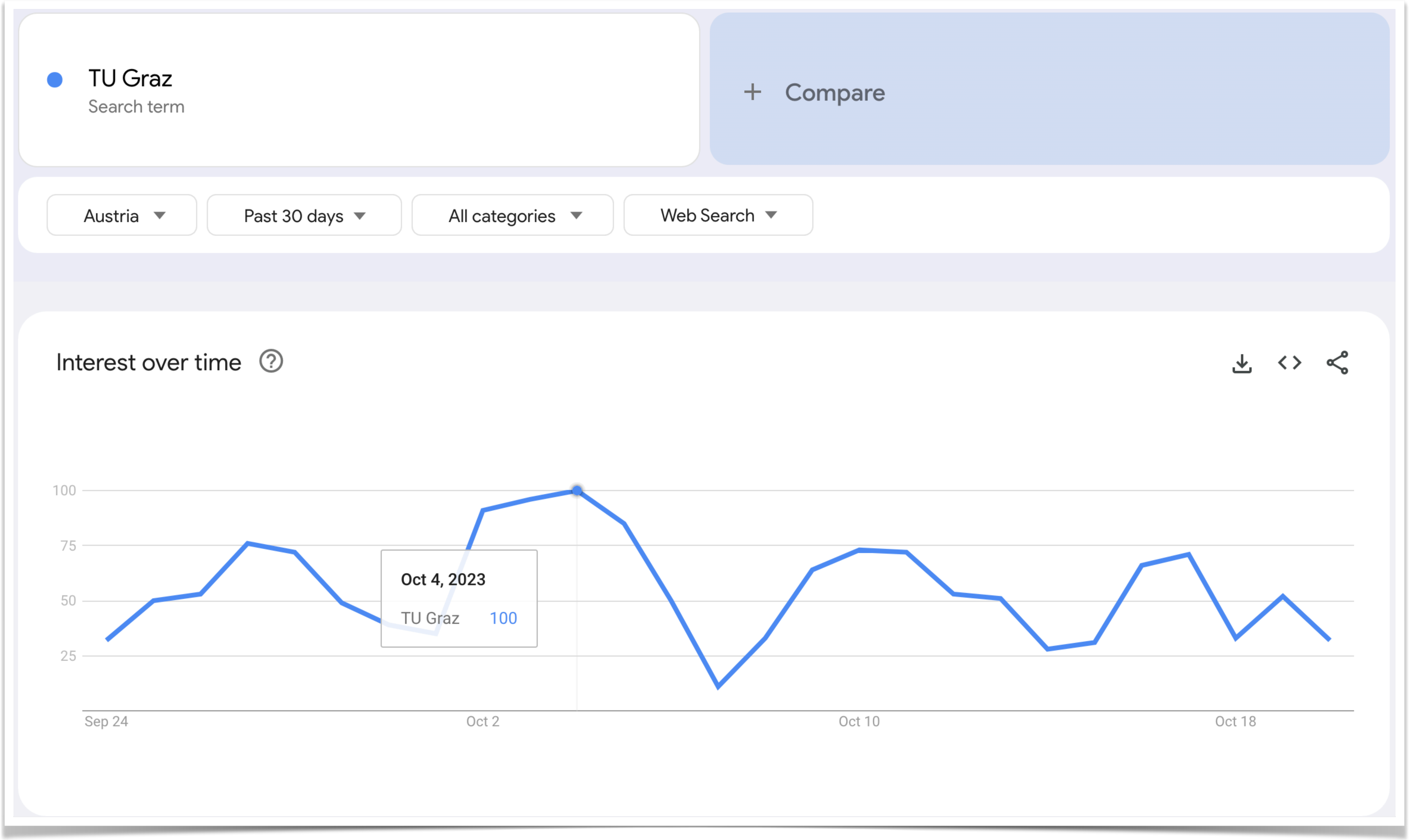
BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

Social Trends



Google Trends

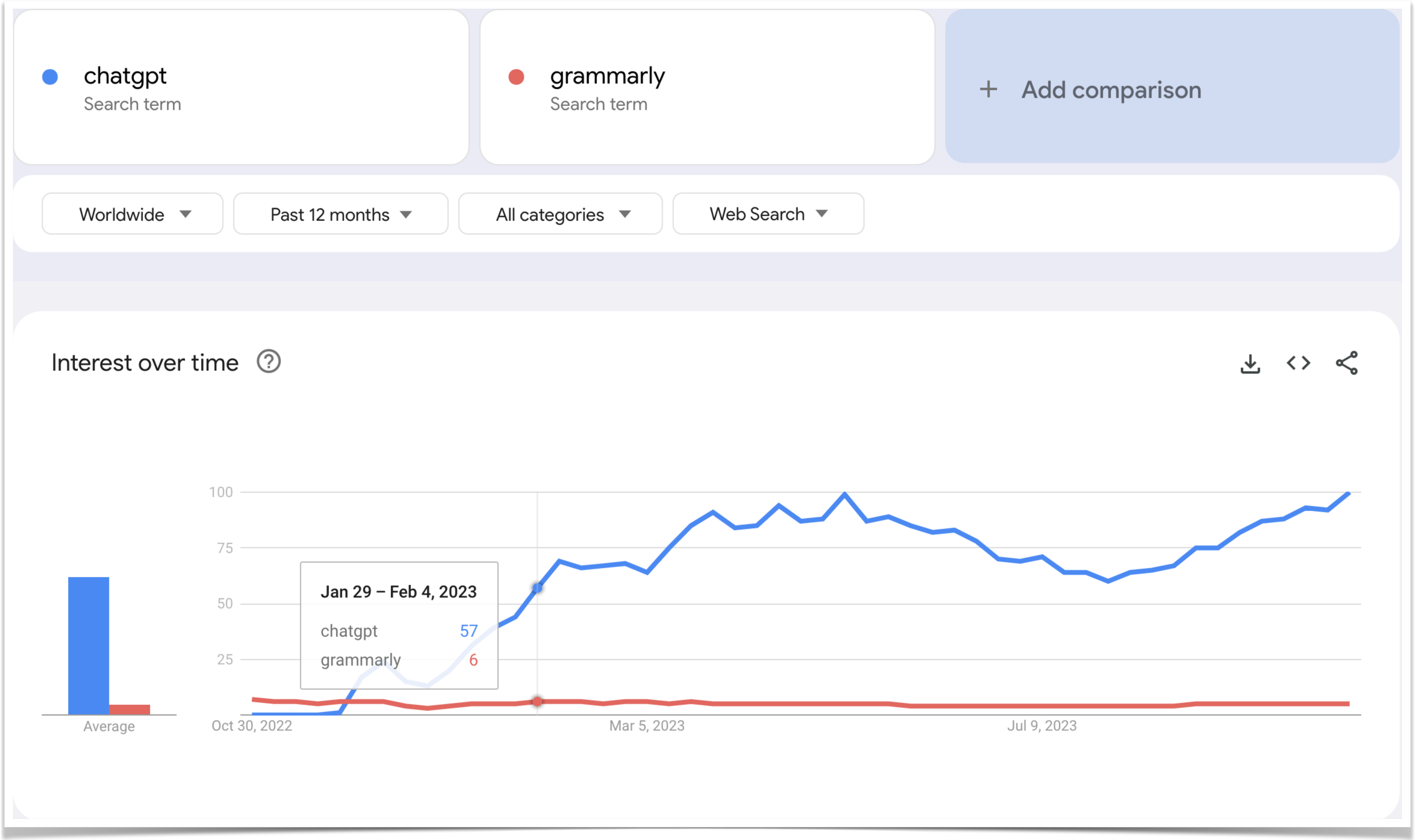
Shows the Google search volume of a term within a given time interval



trends.google.com

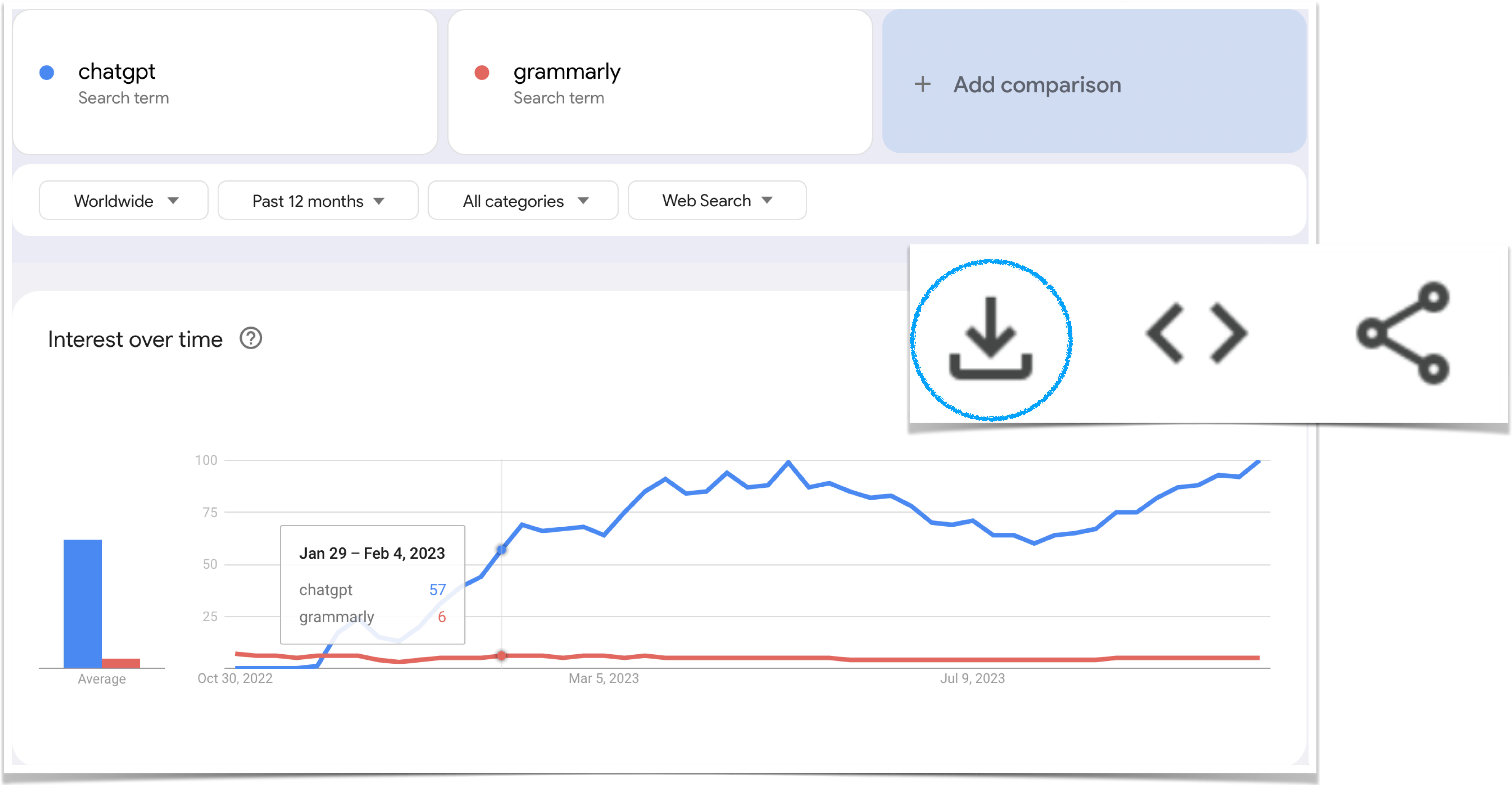
Google Trends

Searching for various trends



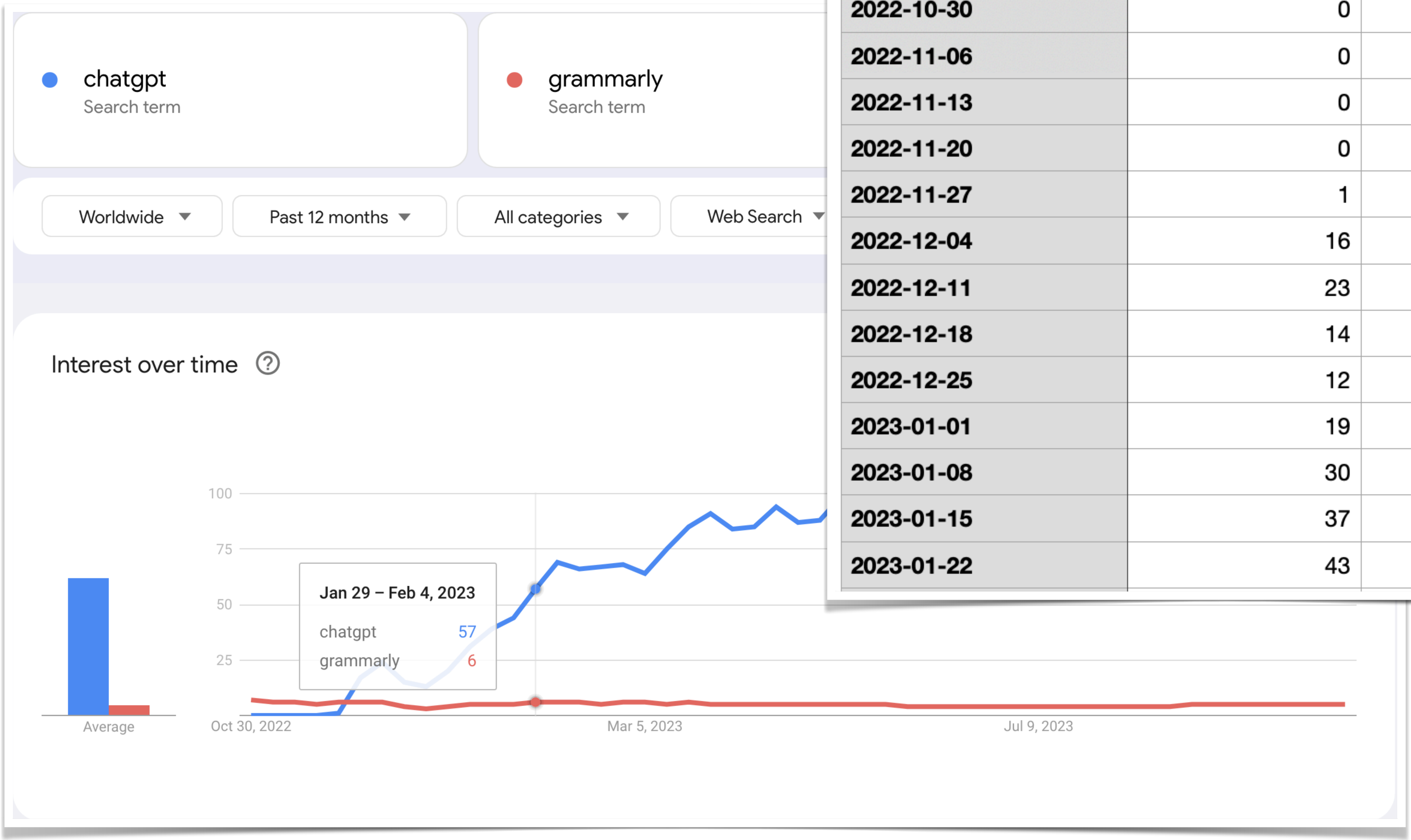
Google Trends

Exporting data



Google Trends

Export file format

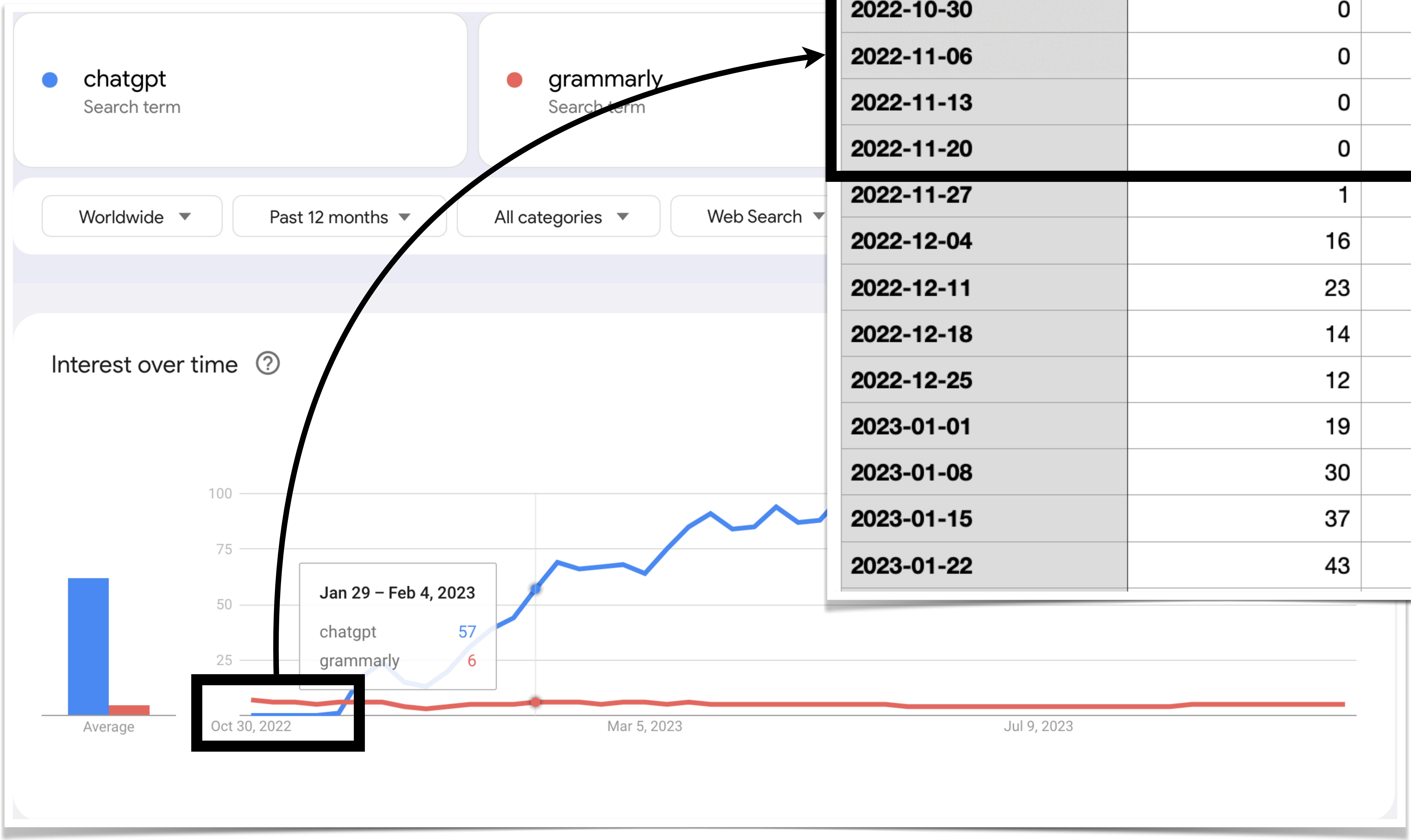


multiTimeline

Category: All categories		
Week	chatgpt: (Worldwide)	grammarly: (Worldwide)
2022-10-30	0	6
2022-11-06	0	6
2022-11-13	0	6
2022-11-20	0	5
2022-11-27	1	6
2022-12-04	16	6
2022-12-11	23	6
2022-12-18	14	4
2022-12-25	12	3
2023-01-01	19	4
2023-01-08	30	5
2023-01-15	37	5
2023-01-22	43	5

Google Trends

Export file format



multiTimeline

Category: All categories		
Week	chatgpt: (Worldwide)	grammarly: (Worldwide)
2022-10-30	0	6
2022-11-06	0	6
2022-11-13	0	6
2022-11-20	0	5
2022-11-27	1	6
2022-12-04	16	6
2022-12-11	23	6
2022-12-18	14	4
2022-12-25	12	3
2023-01-01	19	4
2023-01-08	30	5
2023-01-15	37	5
2023-01-22	43	5

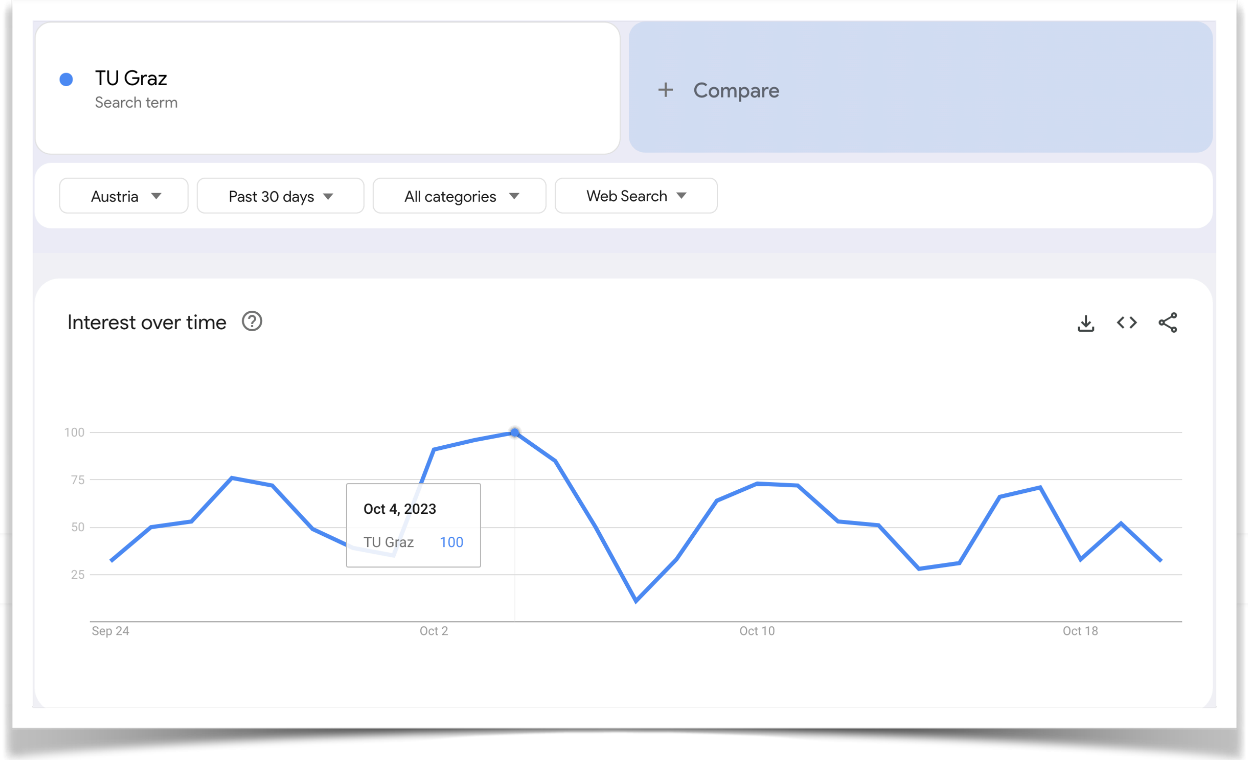
Google Trends

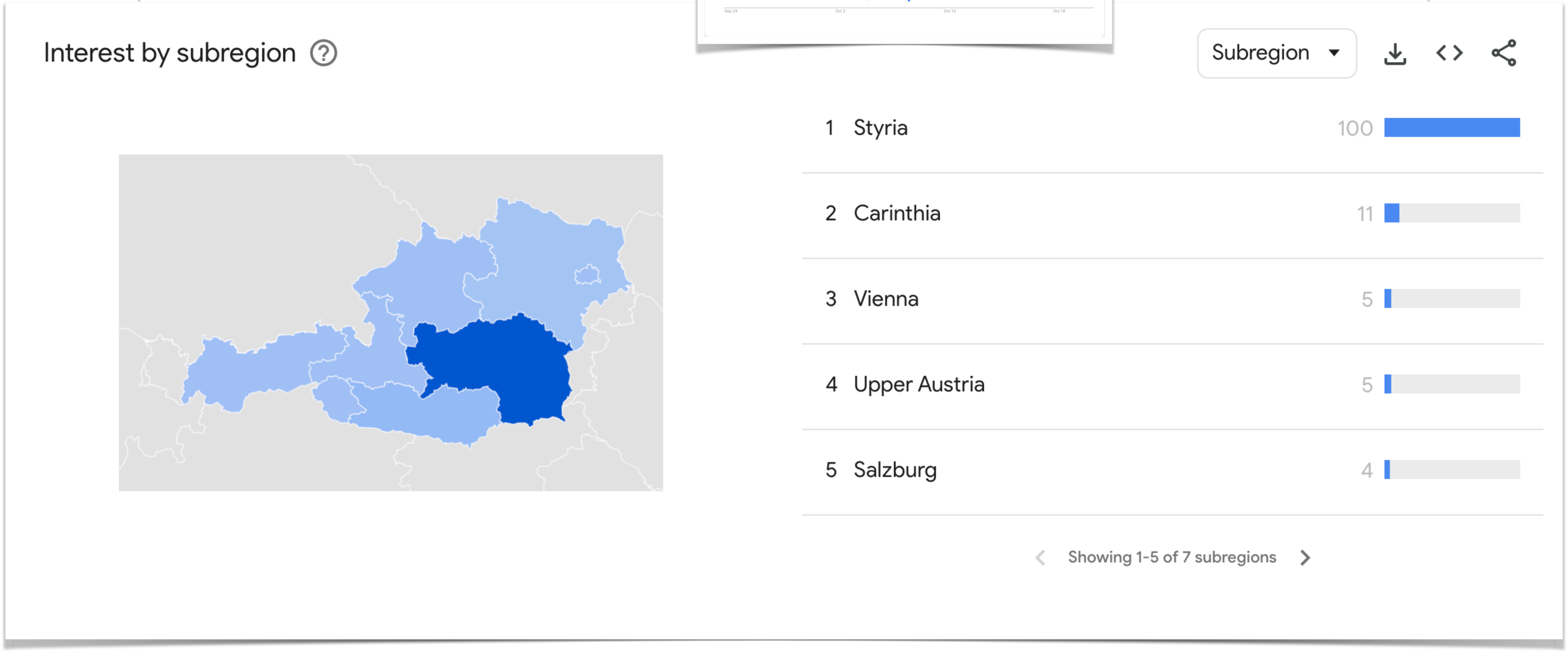
Comparing regions

Google Trends

Comparing regions

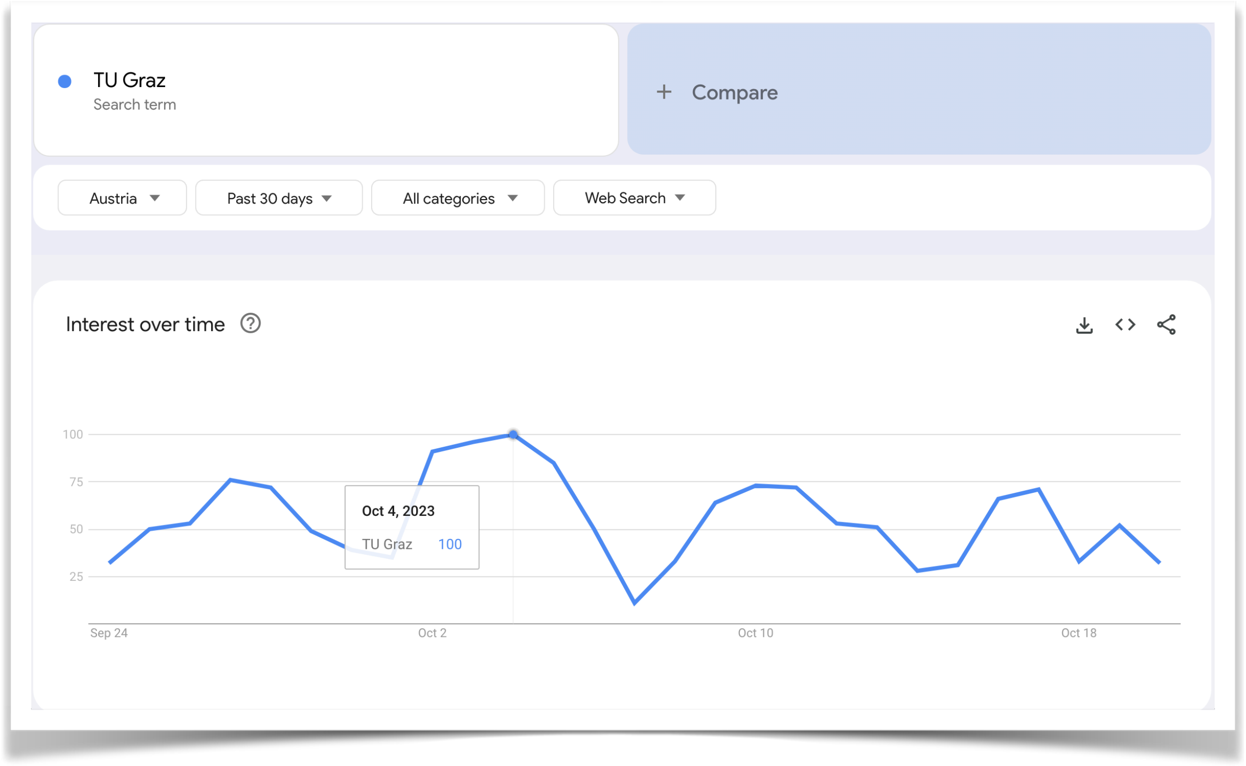
search term: **TU Graz**





Google Trends

Related topics and queries



search term: **TU Graz**

Related topics — Rising

1	Professor - Job title	Breakout
2	Institut für angewandte Informationsverarb...	Breakout
3	university cafeteria - Topic	Breakout
4	Virtual private network - Topic	Breakout
5	Research - Organization type	Breakout

Showing 1-5 of 7 topics

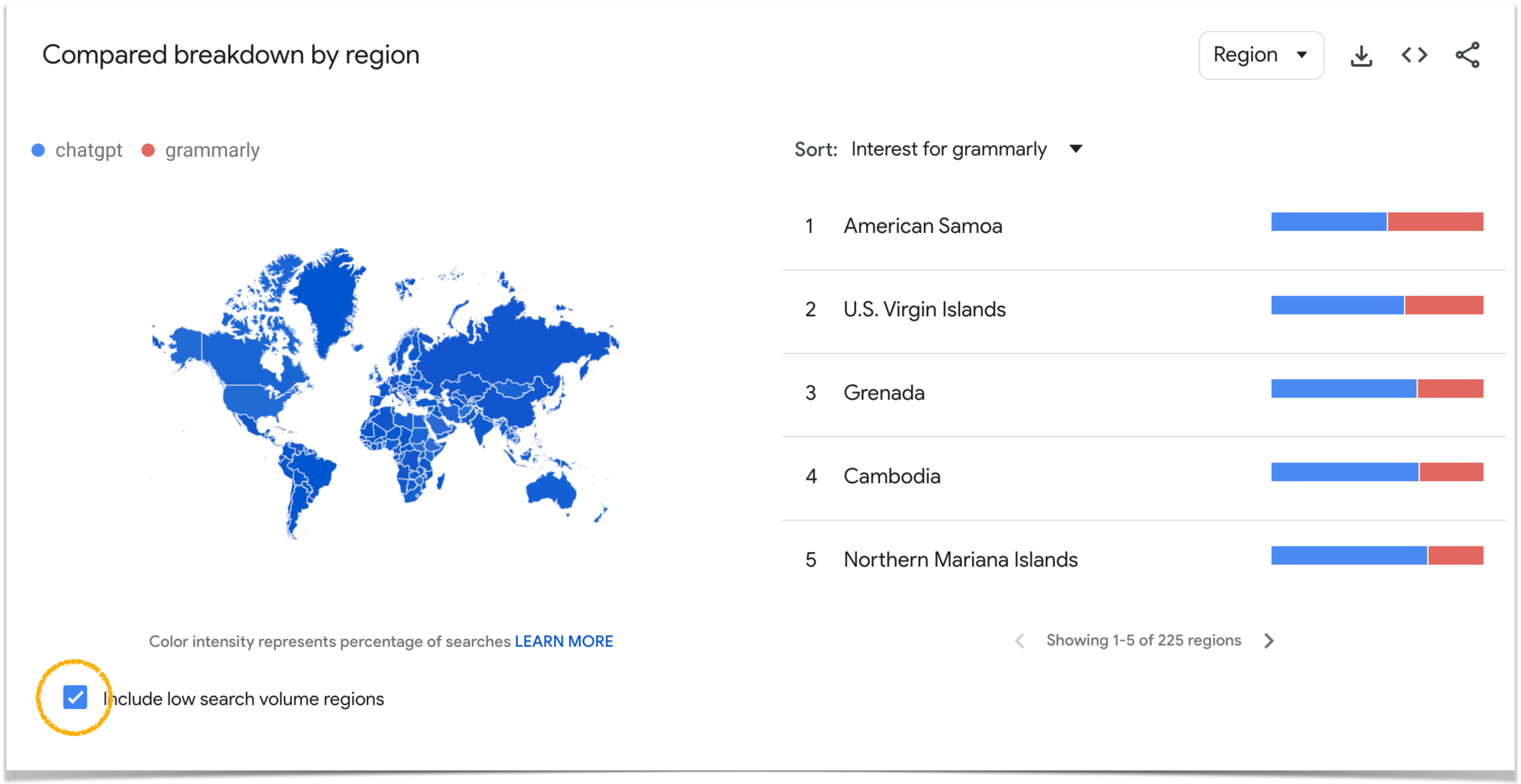
Related queries — Rising

1	tu fest graz	Breakout
2	tu graz welcome days	Breakout
3	teach center tu graz	Breakout
4	tu fest graz 2023	Breakout
5	tc tu graz	Breakout

Showing 1-5 of 9 queries

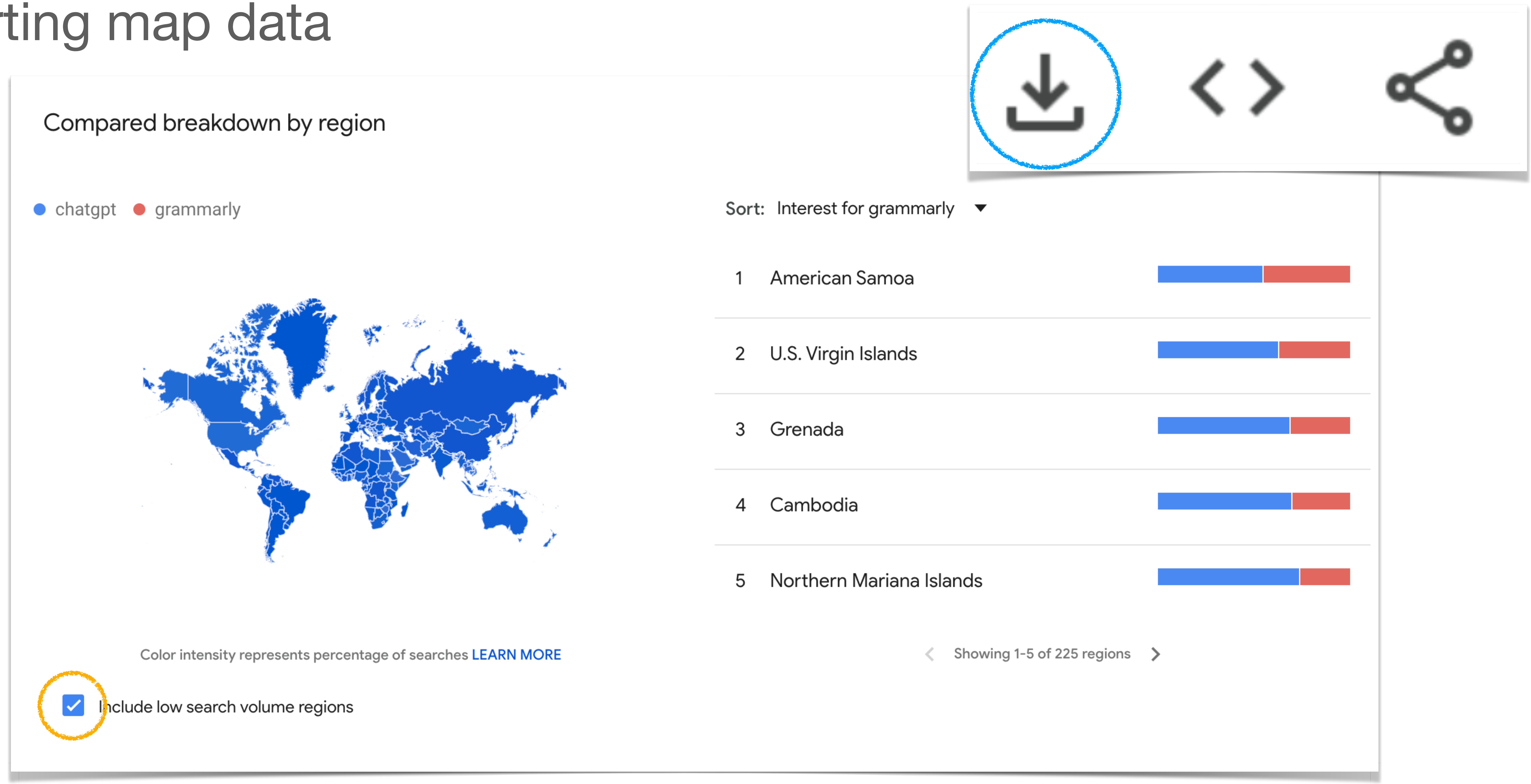
Google Trends

Exporting map data



Google Trends

Exporting map data



Google Trends

Export file format for maps

Compared breakdown by region

chatgpt grammarly

Color intensity represents percentage of searches LEARN MORE

Showing 1-5 of 225 regions

Include low search volume regions

geoMap

Category: All categories		
Country	chatgpt: (10/24/22 - 10/24/23)	grammarly: (10/24/22 - 10/24/23)
China	97 %	3 %
Bhutan	97 %	3 %
Malawi	98 %	2 %
Solomon Islands	100 %	
Nepal	97 %	3 %
Madagascar	99 %	1 %
Djibouti	100 %	
Philippines	84 %	16 %
Singapore	95 %	5 %
Sri Lanka	93 %	7 %
Pakistan	93 %	7 %
Vanuatu	100 %	

What can we do with **Google Trends** data?

The Future Orientation Index (FOI)

Preis et al. 2012

The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past**.

The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past**.

- The FOI for a country c on year y is calculated as: $FOI_{c,y} = \dfrac{G_c(y+1)}{G_c(y-1)}$

The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past**.
- The FOI for a country c on year y is calculated as: $FOI_{c,y} = \dfrac{G_c(y+1)}{G_c(y-1)}$
- Where $G_c(y_a)$ is the Google Trends volume of searches for year y_a in country c

The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past.**
- The FOI for a country c on year y is calculated as: $FOI_{c,y} = \dfrac{G_c(y+1)}{G_c(y-1)} \left\{ \begin{array}{l} >1 \text{ seek the future +} \\ =1 \text{ seek f\&p equally} \\ <1 \text{ seek the past +} \end{array} \right.$
- Where $G_c(y_a)$ is the Google Trends volume of searches for year y_a in country c

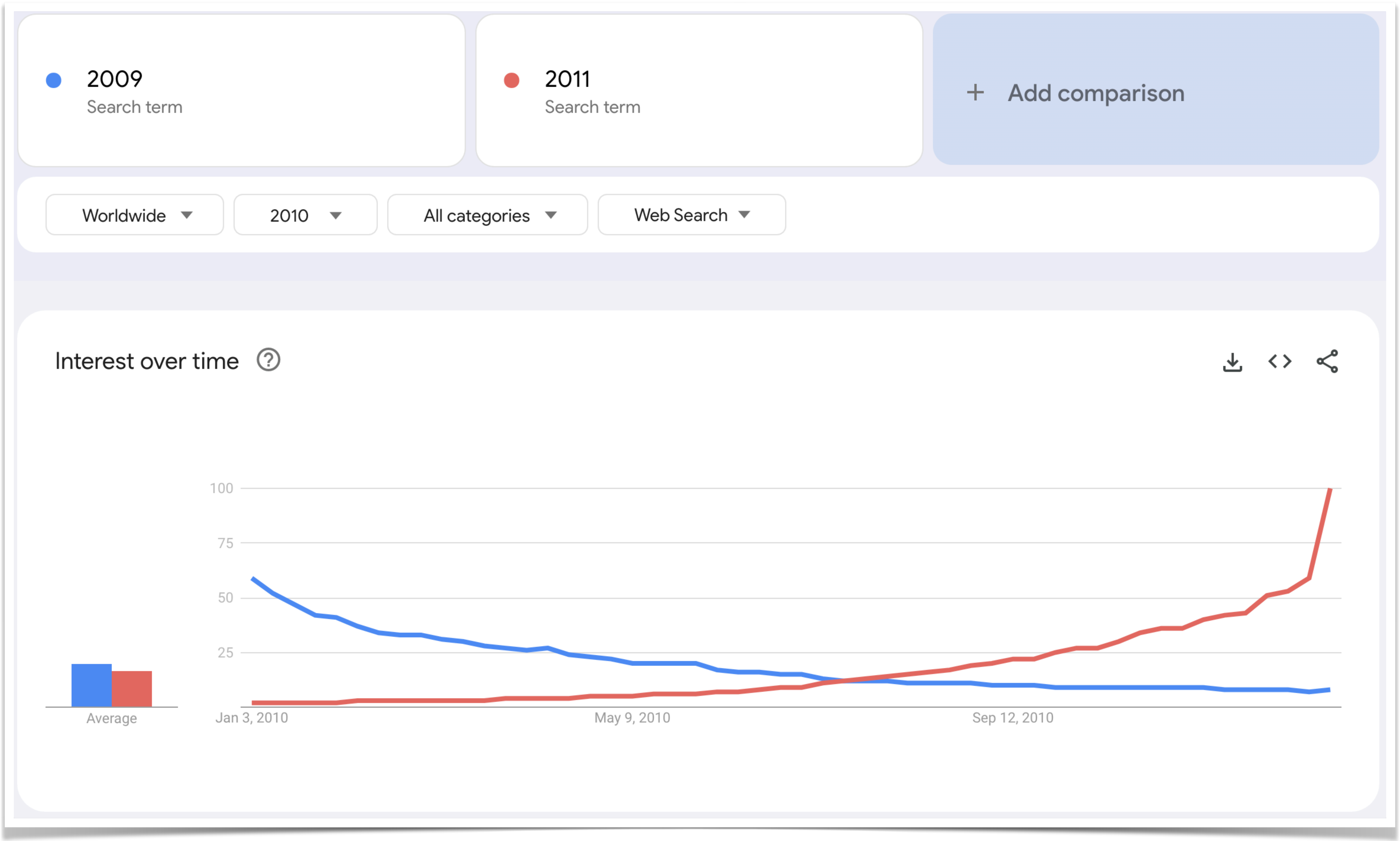
The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past**.
- The FOI for a country c on year y is calculated as: $FOI_{c,y} = \dfrac{G_c(y+1)}{G_c(y-1)}$ $\begin{cases} \text{>1 seek the future +} \\ \text{=1 seek f\&p equally} \\ \text{<1 seek the past +} \end{cases}$
- Where $G_c(y_a)$ is the Google Trends volume of searches for year y_a in country c
- In other words: The FOI measures the **ratio** of search volume within a country for the next year (**future**) divided by the search volume of the previous year (**past**) in the same country.

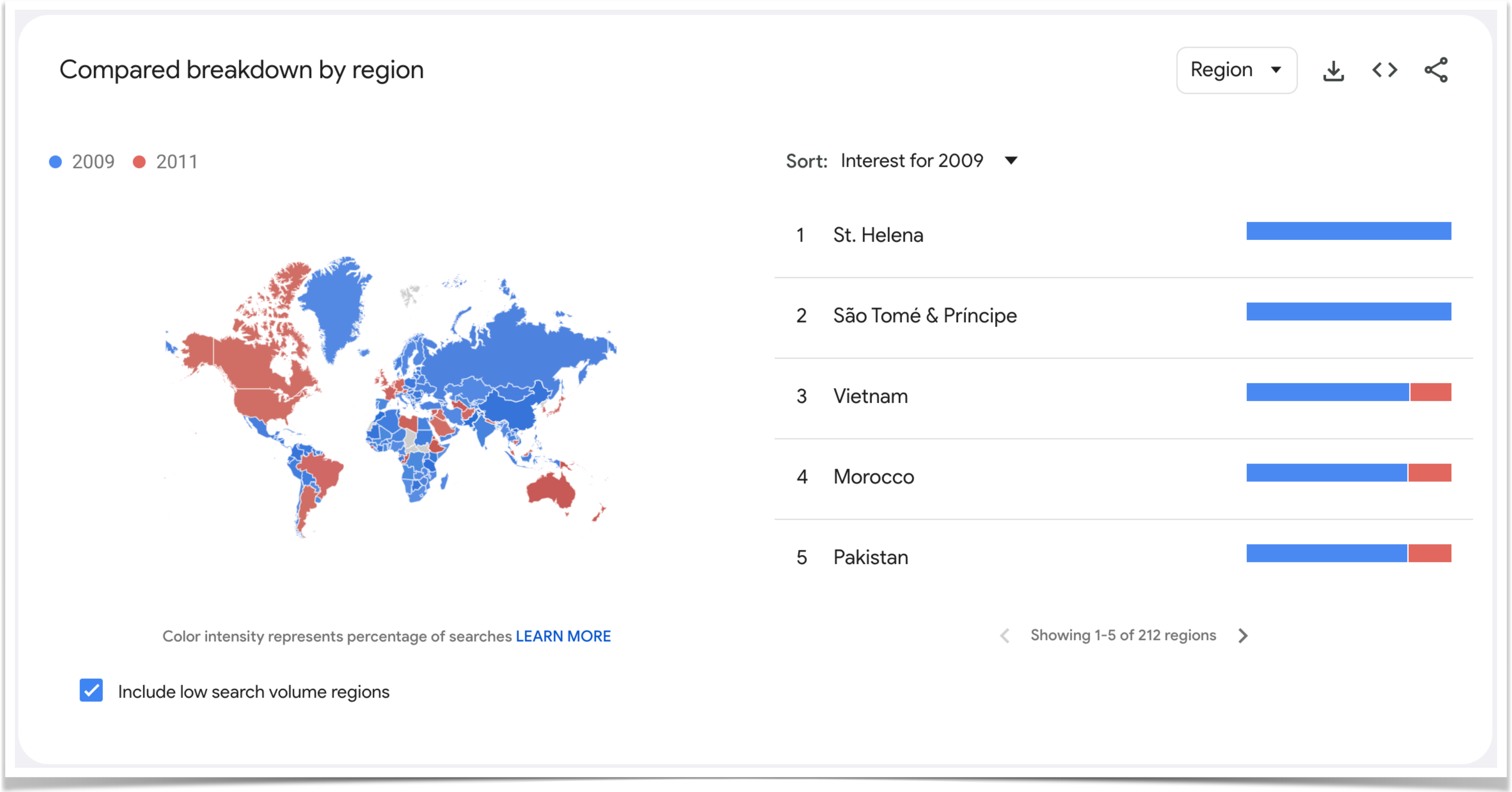
Examples of FOI

using Google Trends (per region, map data)



Examples of FOI

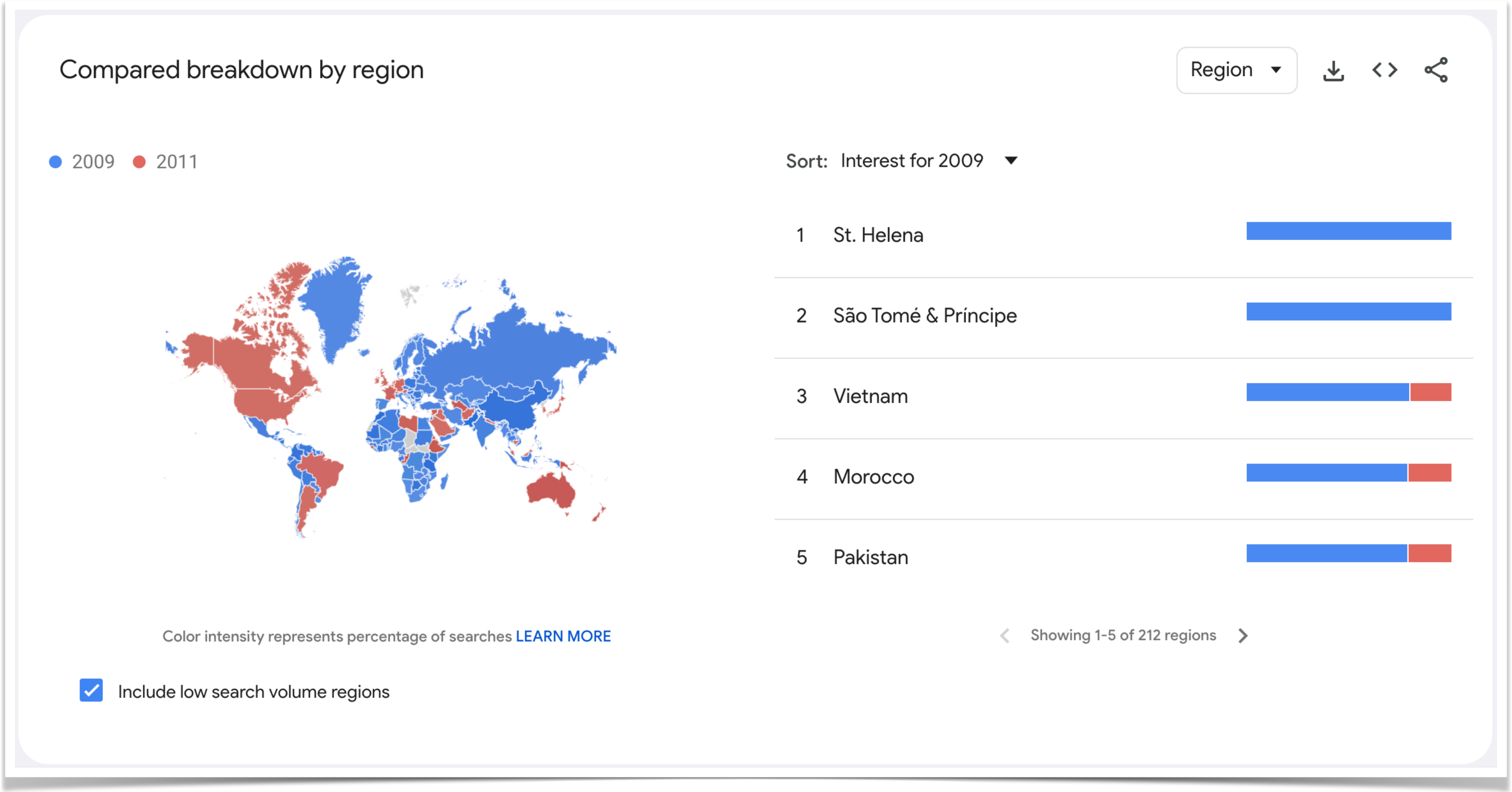
using Google Trends (per region, map data)



Examples of FOI

using Google Trends (per region, map data)

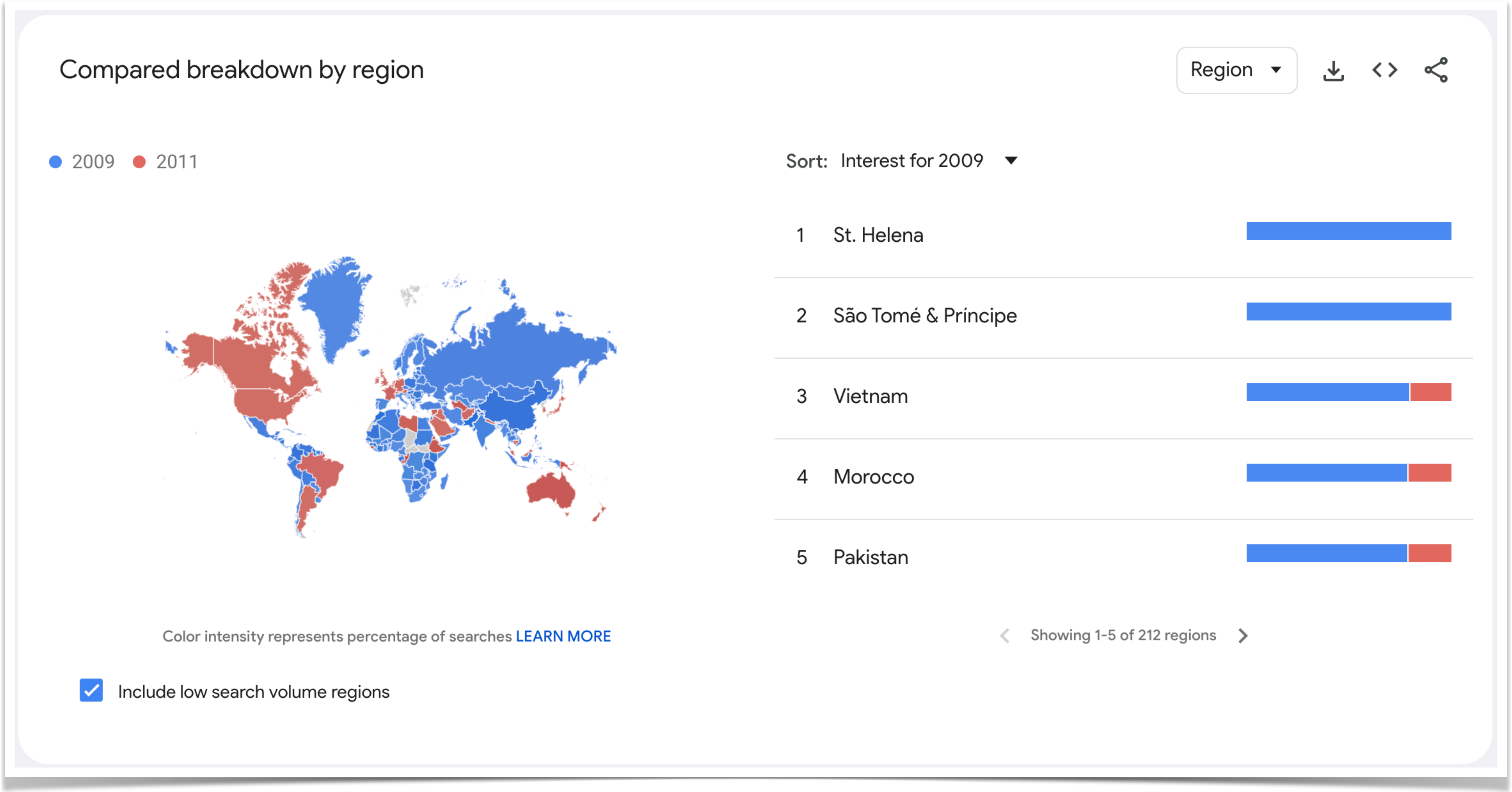
$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)}$$



Examples of FOI

using Google Trends (per region, map data)

$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)}$$

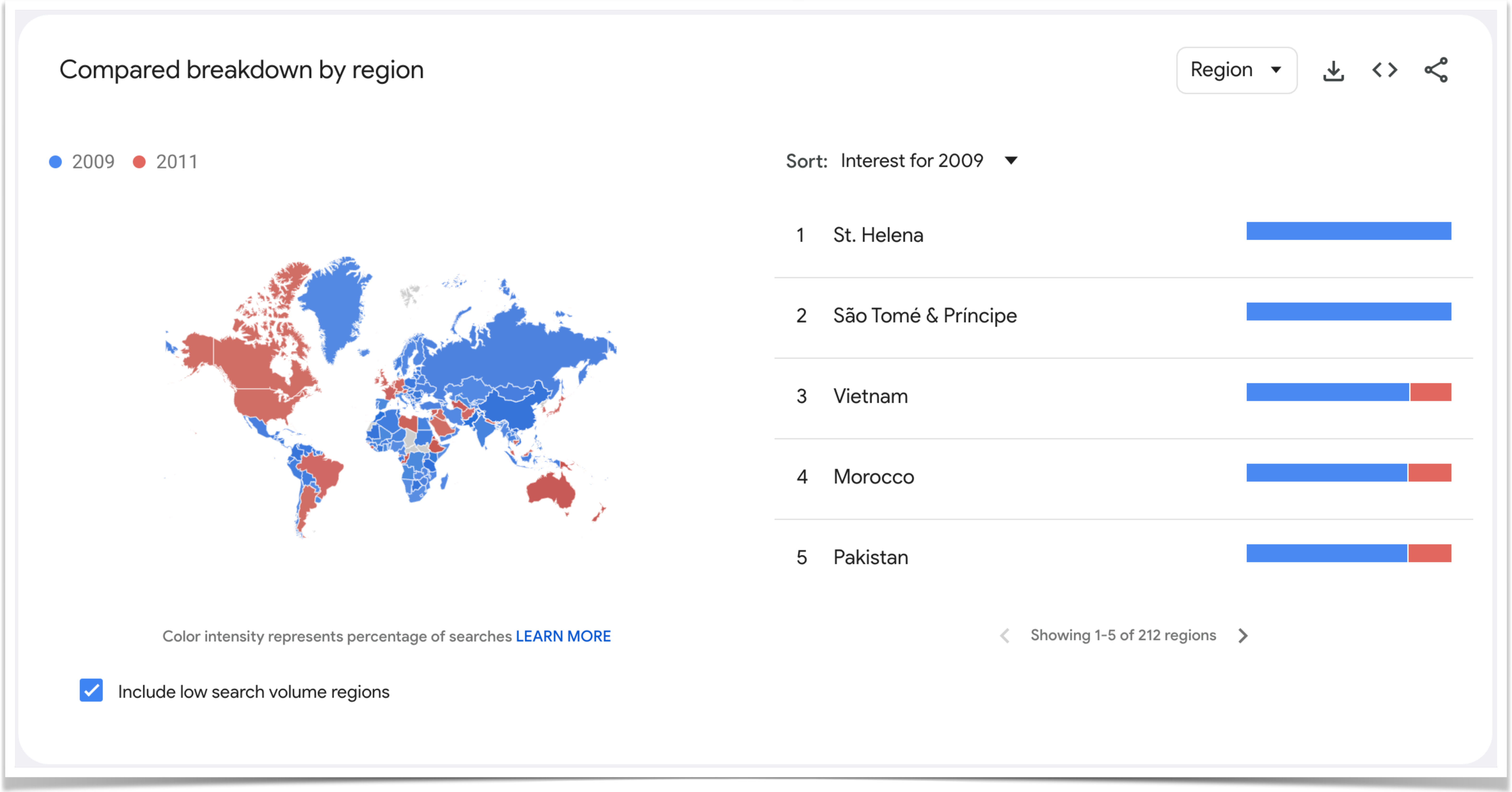


	2009	2011
Country		
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

Examples of FOI

using Google Trends (per region, map data)

$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)}$$



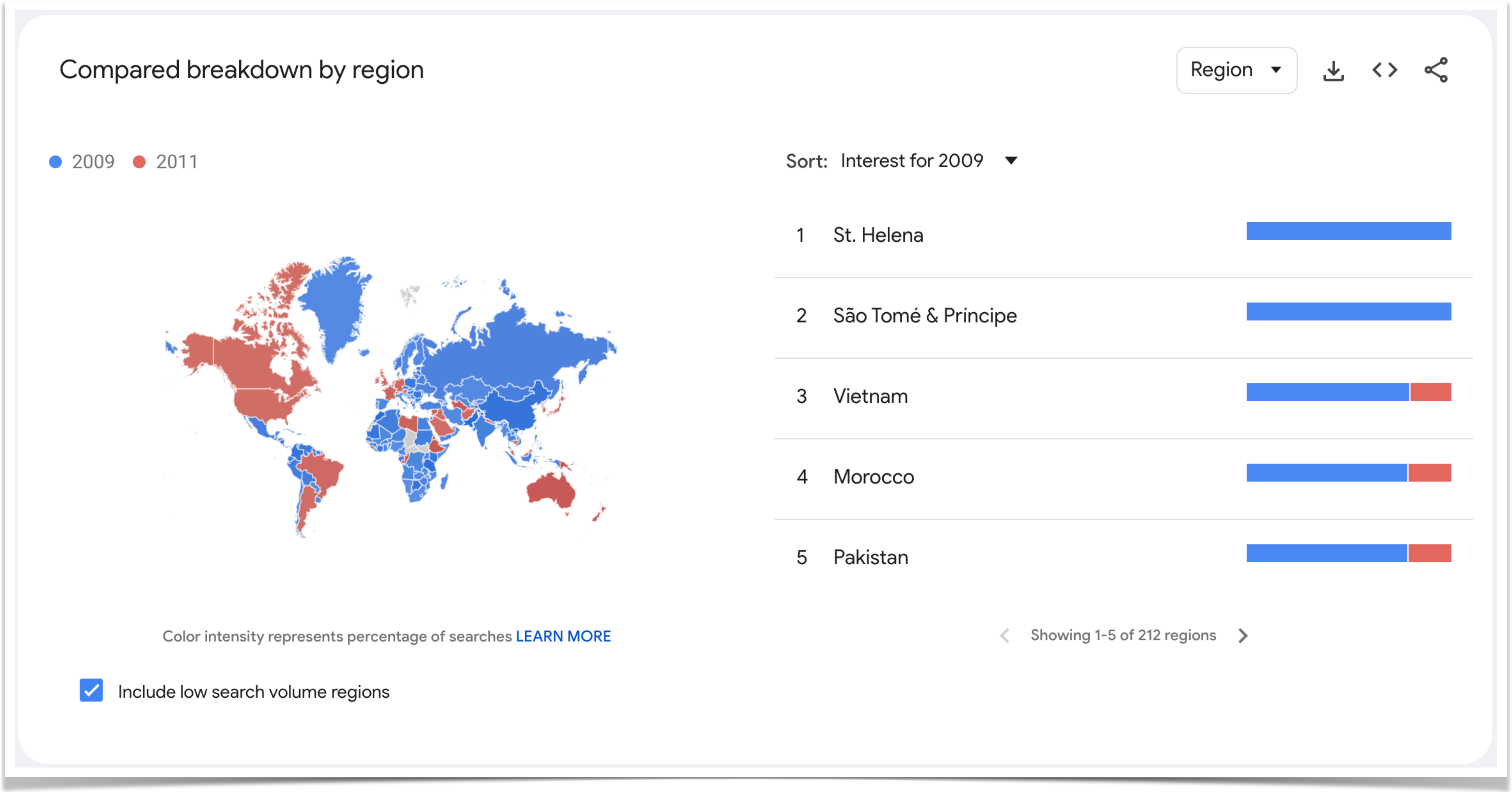
Country	2009	2011
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

$$FOI_{DE,2010} = \frac{54}{46} = 1.17$$

Examples of FOI

using Google Trends (per region, map data)

$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)}$$



Country	2009	2011
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

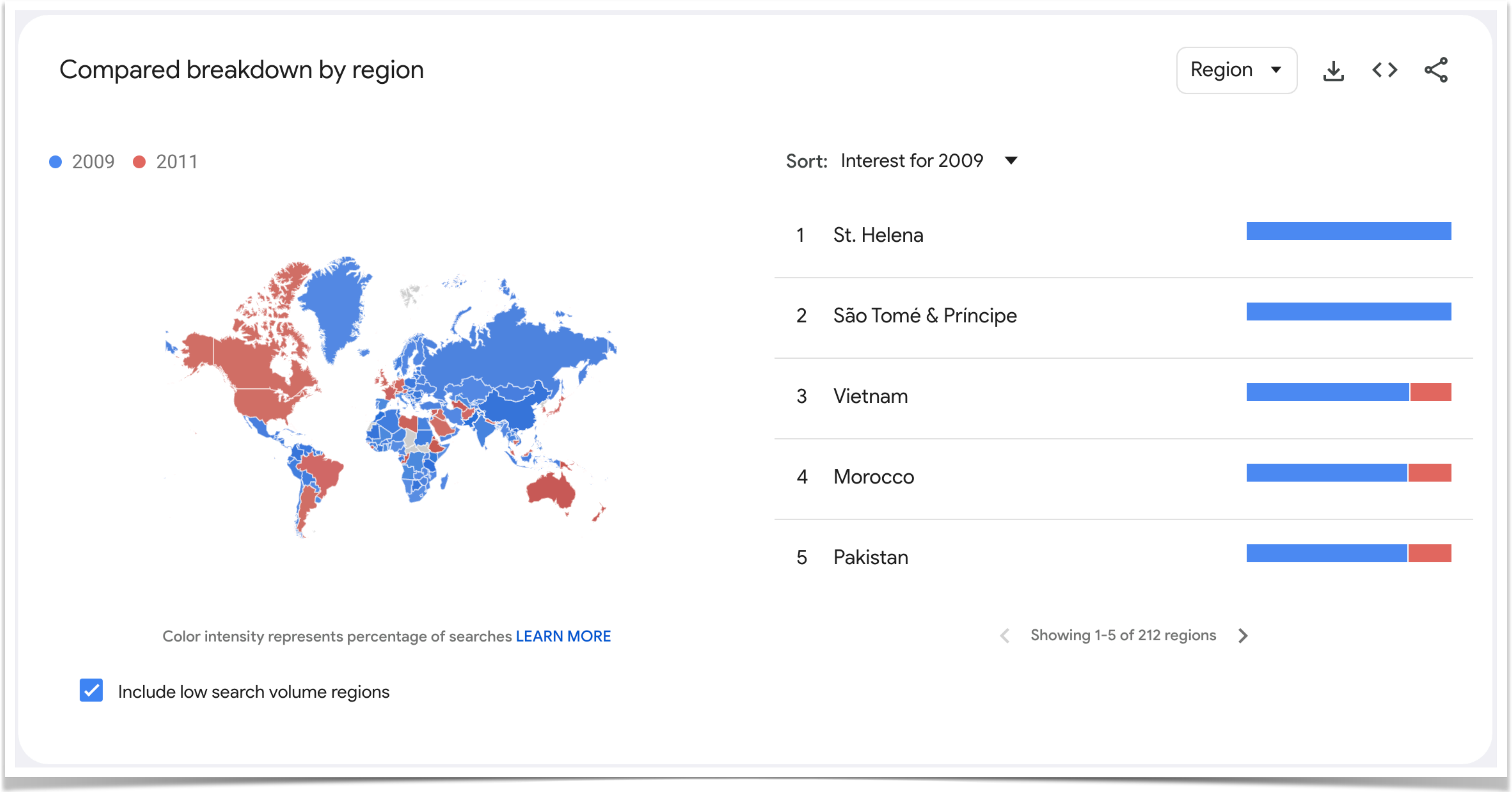
$$FOI_{DE,2010} = \frac{54}{46} = 1.17$$

$$FOI_{AT,2010} = \frac{54}{46} = 1.17$$

Examples of FOI

using Google Trends (per region, map data)

$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)}$$



	2009	2011
Country		
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

$$FOI_{DE,2010} = \frac{54}{46} = 1.17$$

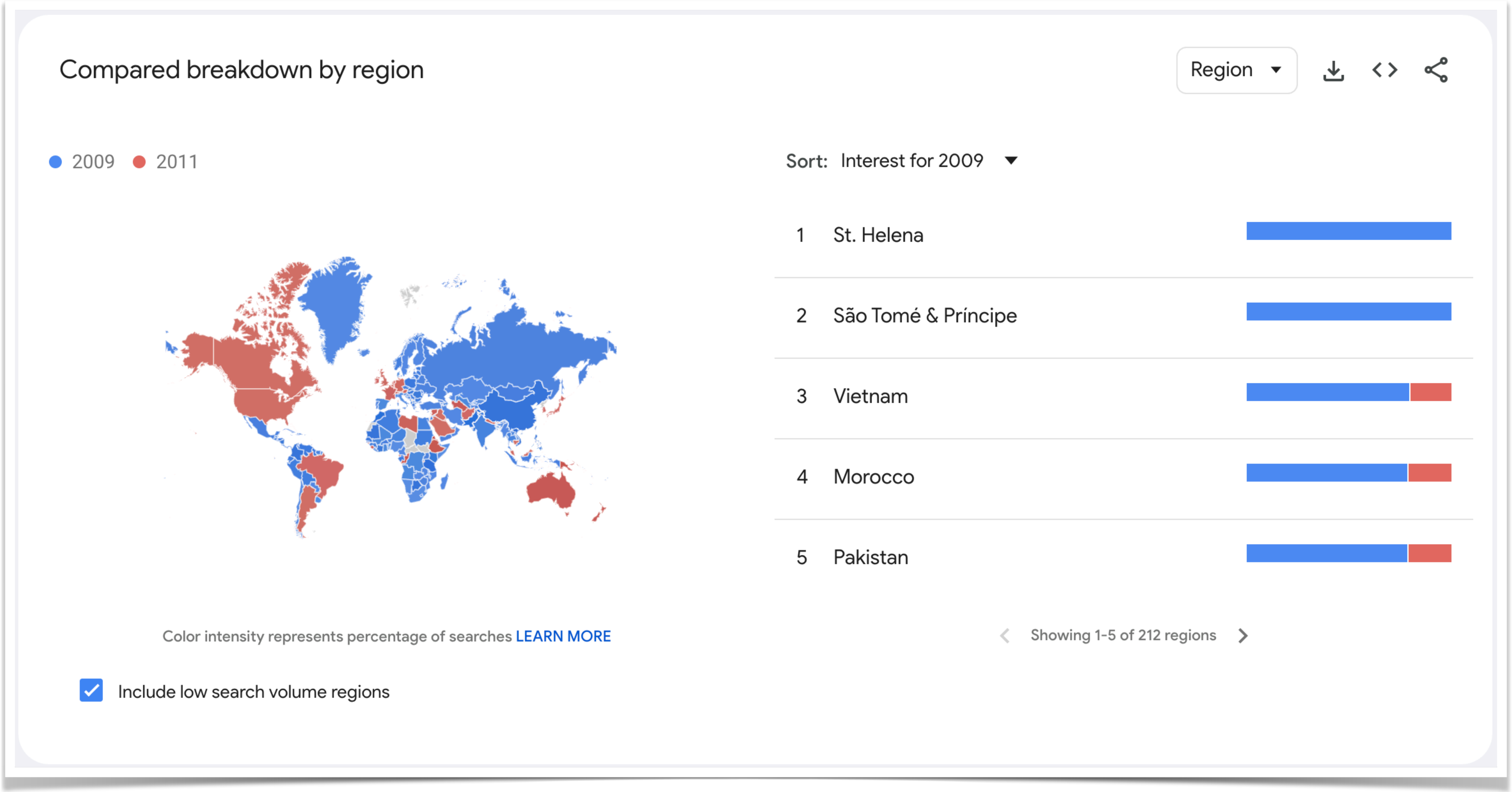
$$FOI_{AT,2010} = \frac{54}{46} = 1.17$$

$$FOI_{EC,2010} = \frac{37}{63} = 0.59$$

Examples of FOI

using Google Trends (per region, map data)

$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)} \begin{cases} >1 \text{ seek the future +} \\ =1 \text{ seek f\&p equally} \\ <1 \text{ seek the past +} \end{cases}$$



Country	2009	2011
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

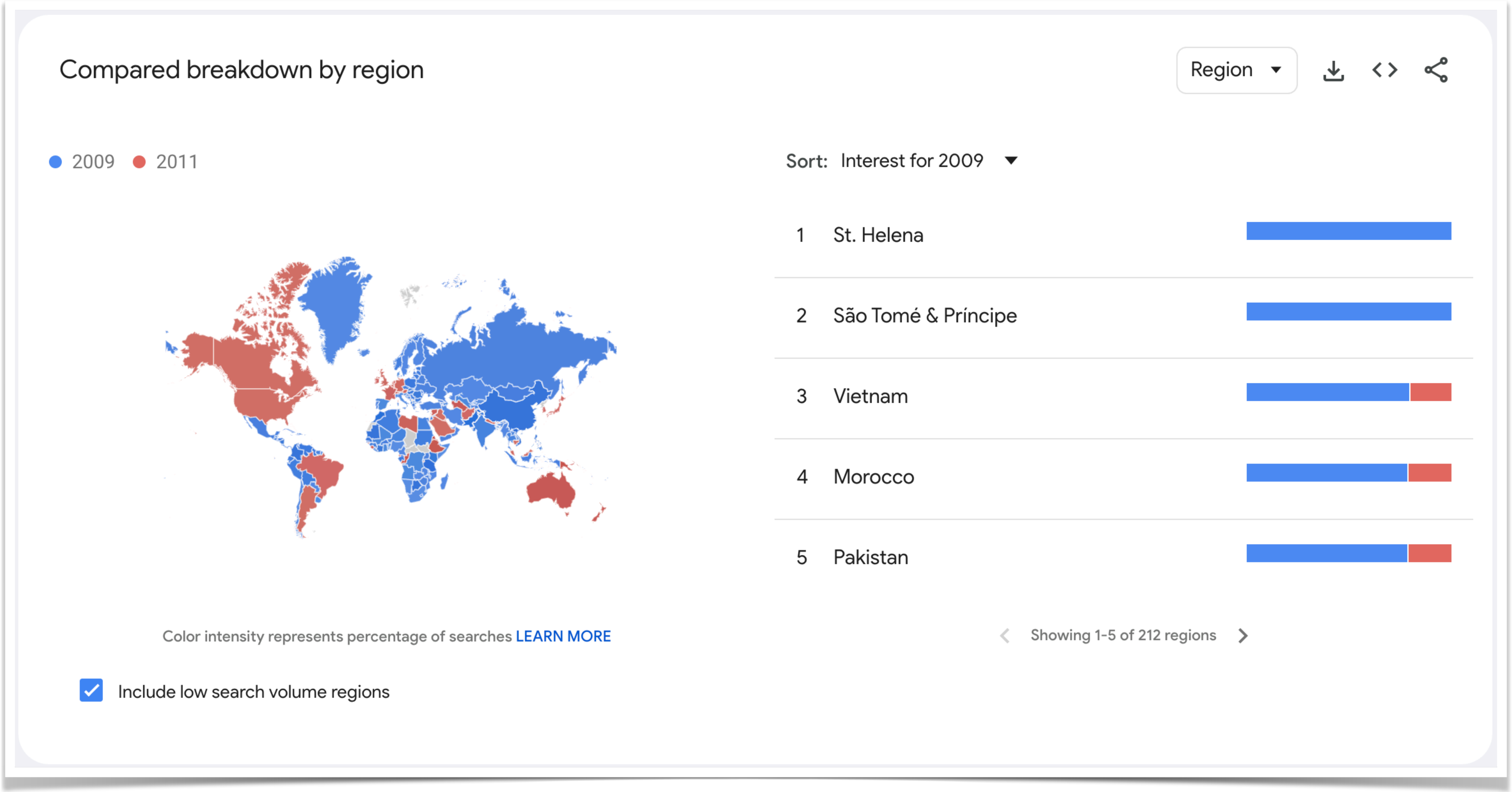
$$FOI_{DE,2010} = \frac{54}{46} = 1.17$$

$$FOI_{AT,2010} = \frac{54}{46} = 1.17$$

$$FOI_{EC,2010} = \frac{37}{63} = 0.59$$

Examples of FOI

using Google Trends (per region, map data)



$$FOI_{c,y} = \frac{G_c(y+1)}{G_c(y-1)} \begin{cases} >1 \text{ seek the future +} \\ =1 \text{ seek f\&p equally} \\ <1 \text{ seek the past +} \end{cases}$$

Country	2009	2011
Germany	46.0	54.0
Austria	46.0	54.0
Ecuador	63.0	37.0

$$FOI_{DE,2010} = \frac{54}{46} = 1.17$$

$$FOI_{AT,2010} = \frac{54}{46} = 1.17$$

$$FOI_{EC,2010} = \frac{37}{63} = 0.59$$

So what?

Culture vs. Economy

Long-term orientation by Geert Hofstede

Culture vs. Economy

Long-term orientation by Geert Hofstede

- Long term orientation refers to whether the society looks more towards the future or leans towards its past.

Culture vs. Economy

Long-term orientation by Geert Hofstede

- Long term orientation refers to whether the society looks more towards the future or leans towards its past.
- Long-term oriented societies believe that the most important events in life will occur in the future; short-term oriented societies believe that those events occurred in the past or take place now.

Culture vs. Economy

Long-term orientation by Geert Hofstede

- Long term orientation refers to whether the society looks more towards the future or leans towards its past.
- Long-term oriented societies believe that the most important events in life will occur in the future; short-term oriented societies believe that those events occurred in the past or take place now.

Research Question:

Are long-term oriented societies wealthier?

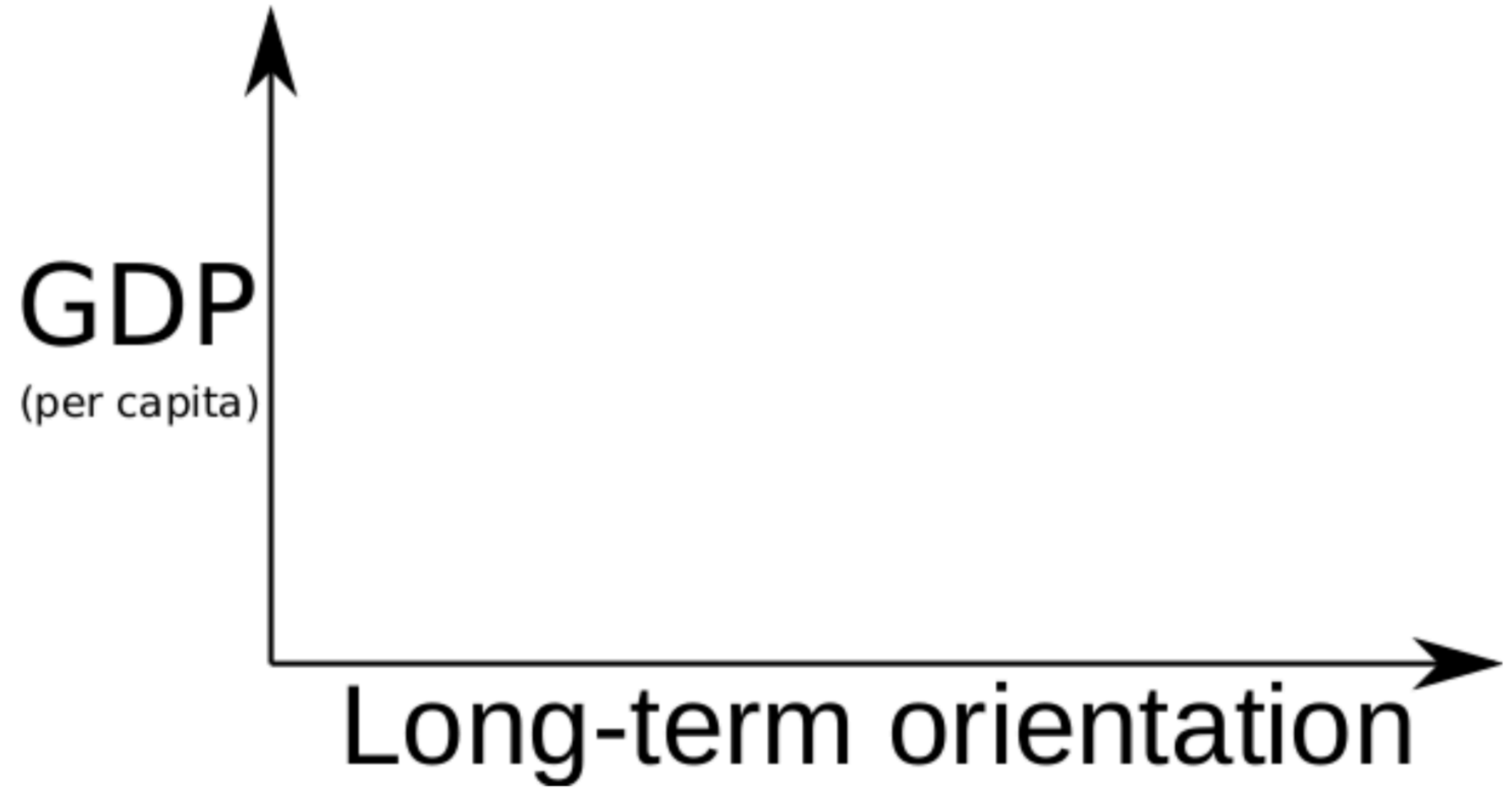
Culture vs. Economy

Long-term orientation by Geert Hofstede

- Long term orientation refers to whether the society looks more towards the future or leans towards its past.
- Long-term oriented societies believe that the most important events in life will occur in the future; short-term oriented societies believe that those events occurred in the past or take place now.

Research Question:

Are long-term oriented societies wealthier?



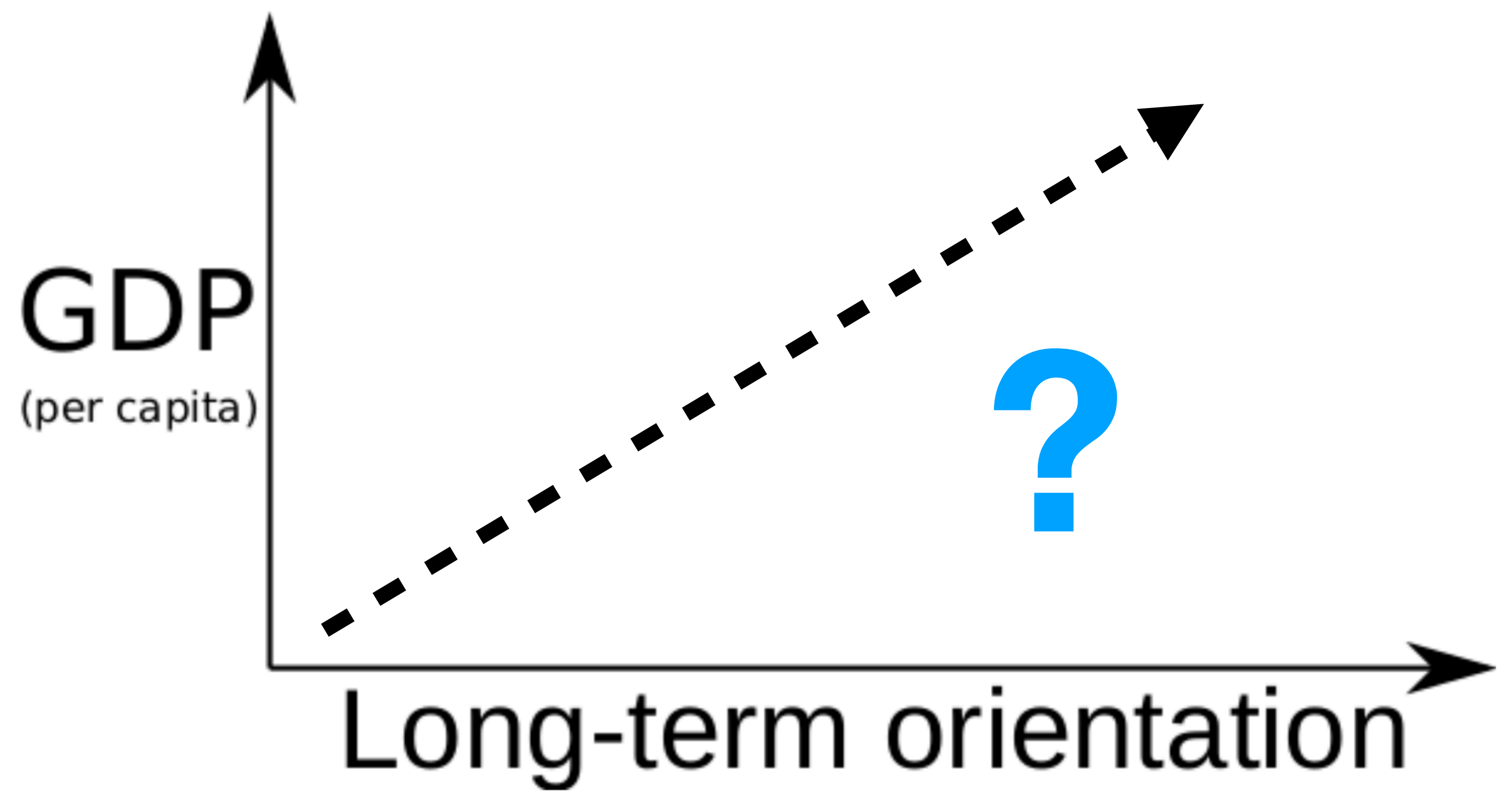
Culture vs. Economy

Long-term orientation by Geert Hofstede

- Long term orientation refers to whether the society looks more towards the future or leans towards its past.
- Long-term oriented societies believe that the most important events in life will occur in the future; short-term oriented societies believe that those events occurred in the past or take place now.

Research Question:

Are long-term oriented societies wealthier?



The Future Orientation Index (FOI)

Preis et al. 2012

- This metric quantifies the degree to which Internet users worldwide seek more information about years in the **future** than years in the **past**.
- The FOI for a country c on year y is calculated as: $FOI_{c,y} = \dfrac{G_c(y+1)}{G_c(y-1)}$
- Where $G_c(y_a)$ is the Google Trends volume of searches for year y_a in country c
- In other words: The FOI measures the ratio of search volume within a country for the next year divided by the search volume of the previous year in the same country.
- Using Google Trends, Preis et al. found that **users from countries with a higher per capita GDP are more likely to search for information about the future** than information about the past.

Examples of FOI

using Google Trends (per region, map data)

Examples of FOI

using Google Trends (per region, map data)

Country Name	GDP_2010	FOI_2010
Germany	37760.91	1.17
Austria	43334.51	1.17
Ecuador	5331.38	0.59

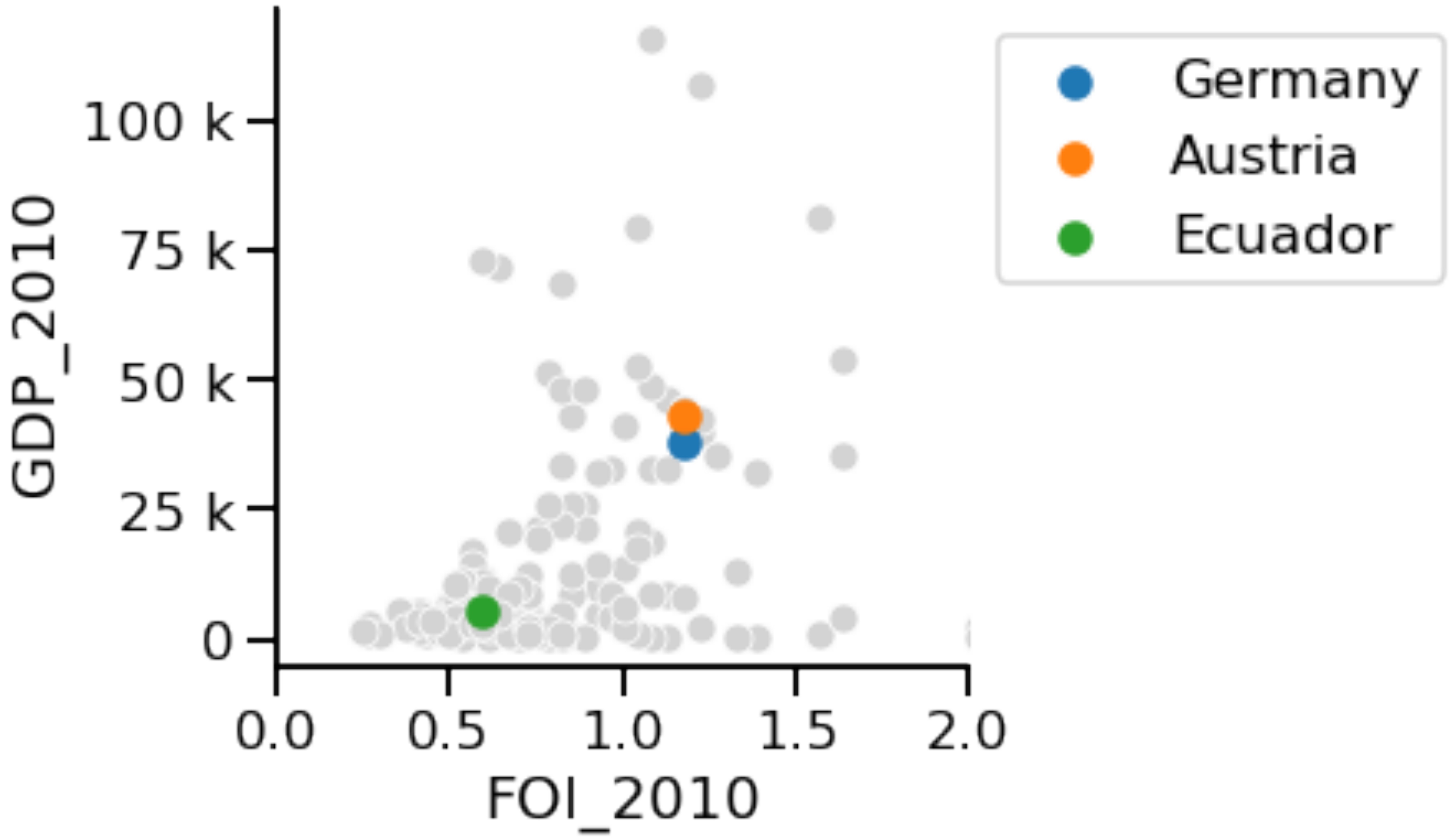
GDP: https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/LTST/endyear/LTST/indicator/NY-GDP-PCAP-KD

Examples of FOI

using Google Trends (per region, map data)

Country Name	GDP_2010	FOI_2010
Germany	37760.91	1.17
Austria	43334.51	1.17
Ecuador	5331.38	0.59

GDP: https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/LTST/endyear/LTST/indicator/NY-GDP-PCAP-KD

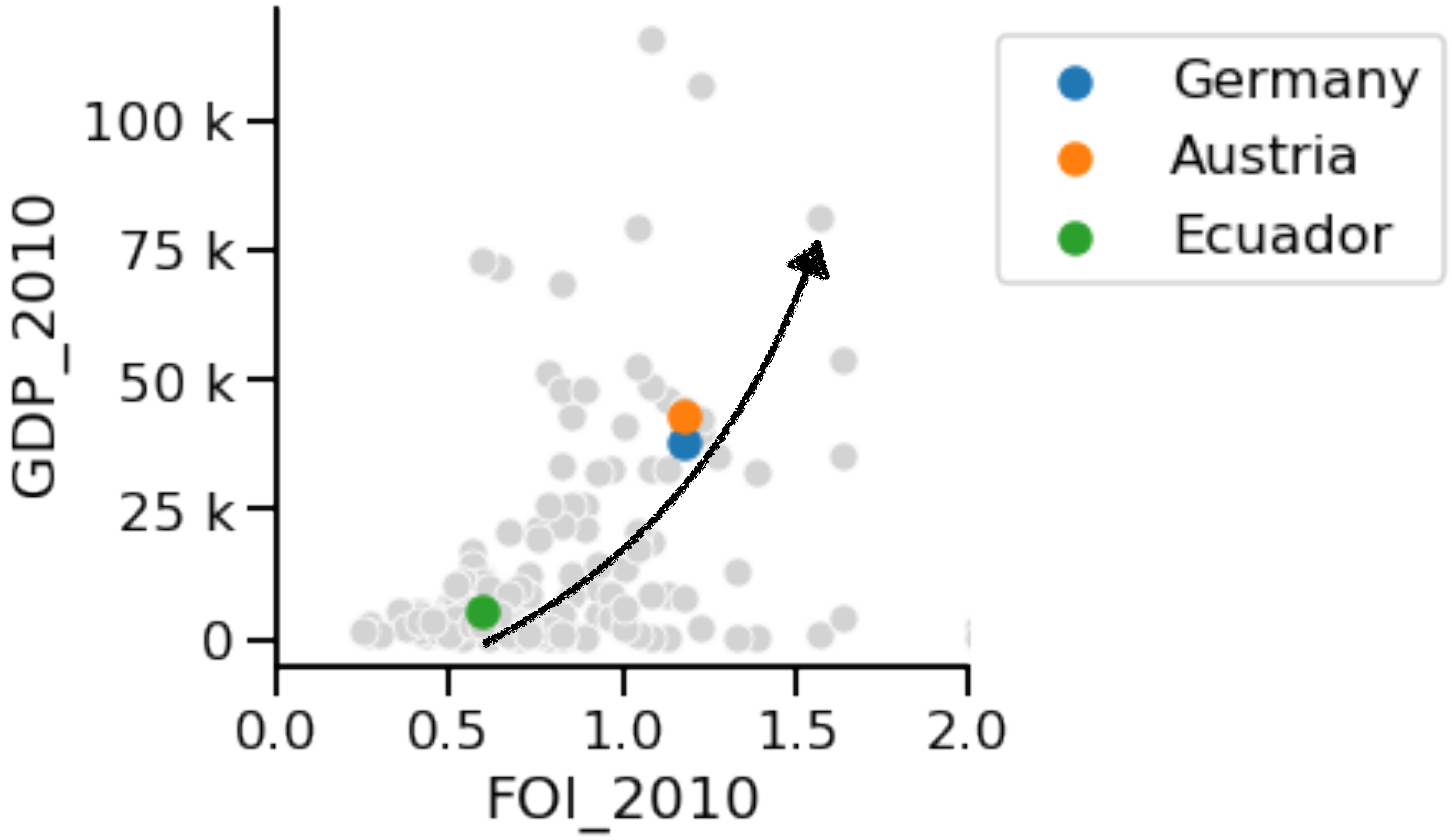


Examples of FOI

using Google Trends (per region, map data)

Country Name	GDP_2010	FOI_2010
Germany	37760.91	1.17
Austria	43334.51	1.17
Ecuador	5331.38	0.59

GDP: https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/LTST/endyear/LTST/indicator/NY-GDP-PCAP-KD

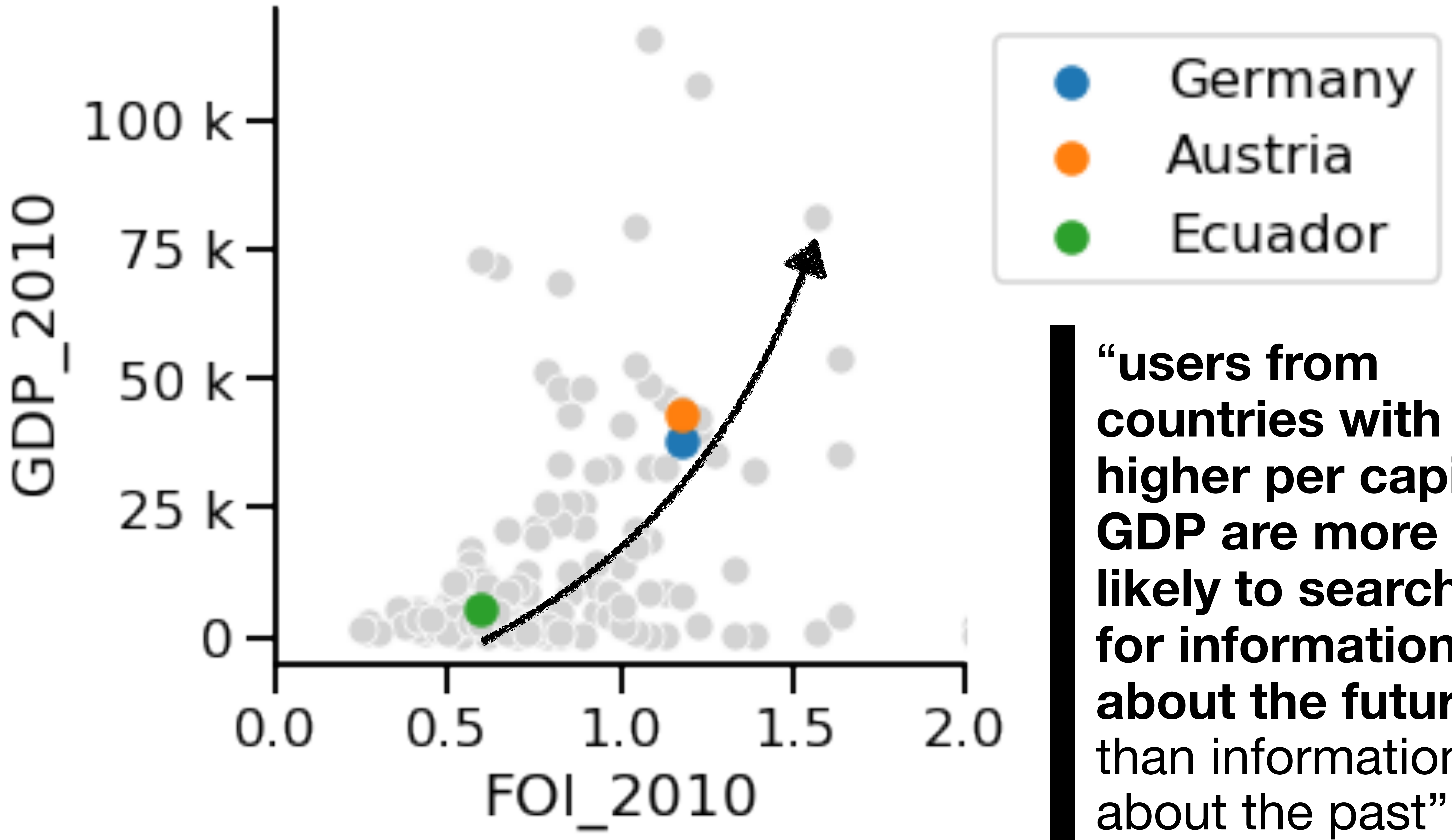


Examples of FOI

using Google Trends (per region, map data)

Country Name	GDP_2010	FOI_2010
Germany	37760.91	1.17
Austria	43334.51	1.17
Ecuador	5331.38	0.59

GDP: https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/LTST/endyear/LTST/indicator/NY-GDP-PCAP-KD



"users from countries with a higher per capita GDP are more likely to search for information about the future than information about the past"
Preis et al. 2012

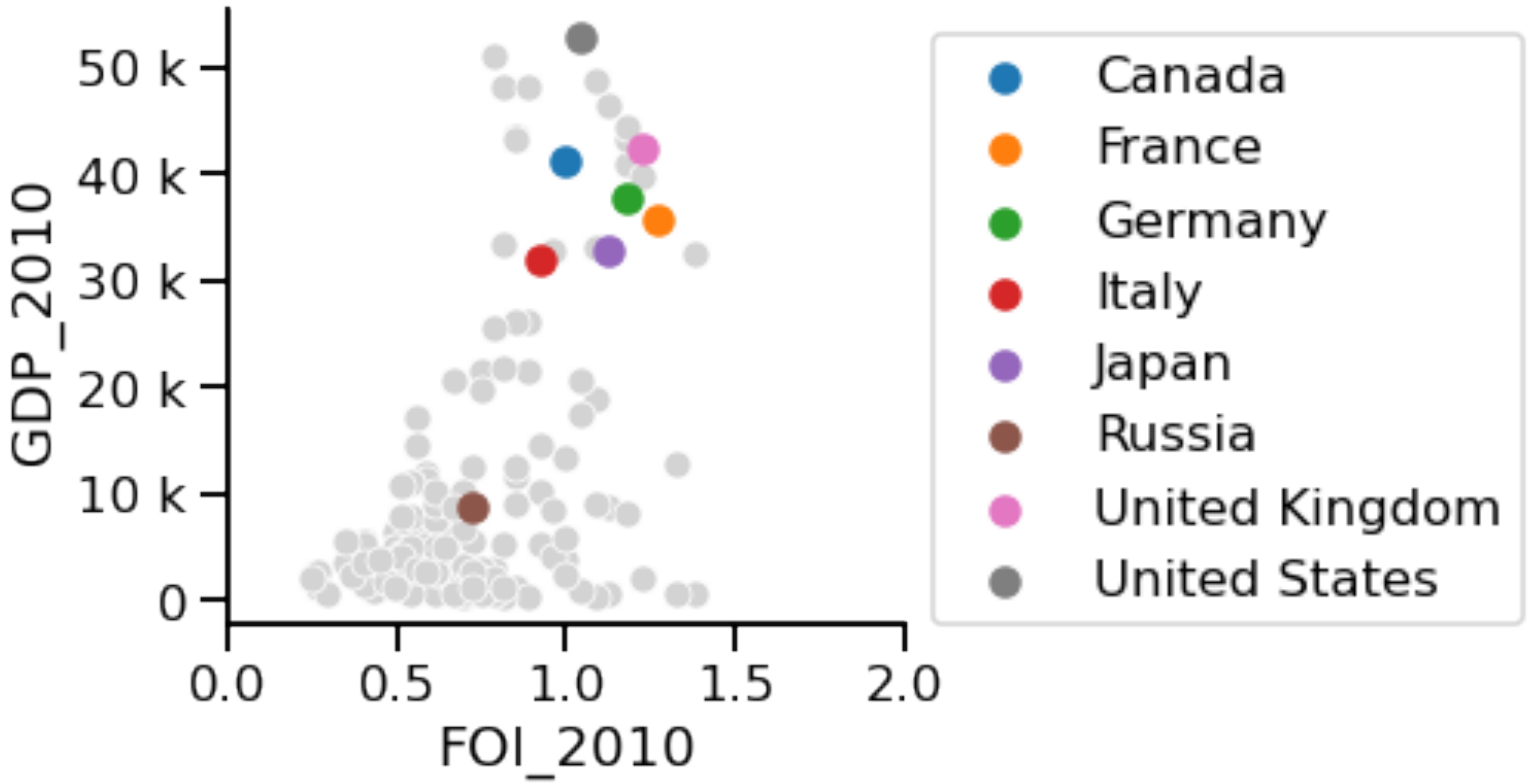
Examples of FOI

Replicating results from Preis et al. 2012

Examples of FOI

Replicating results from Preis et al. 2012

a)

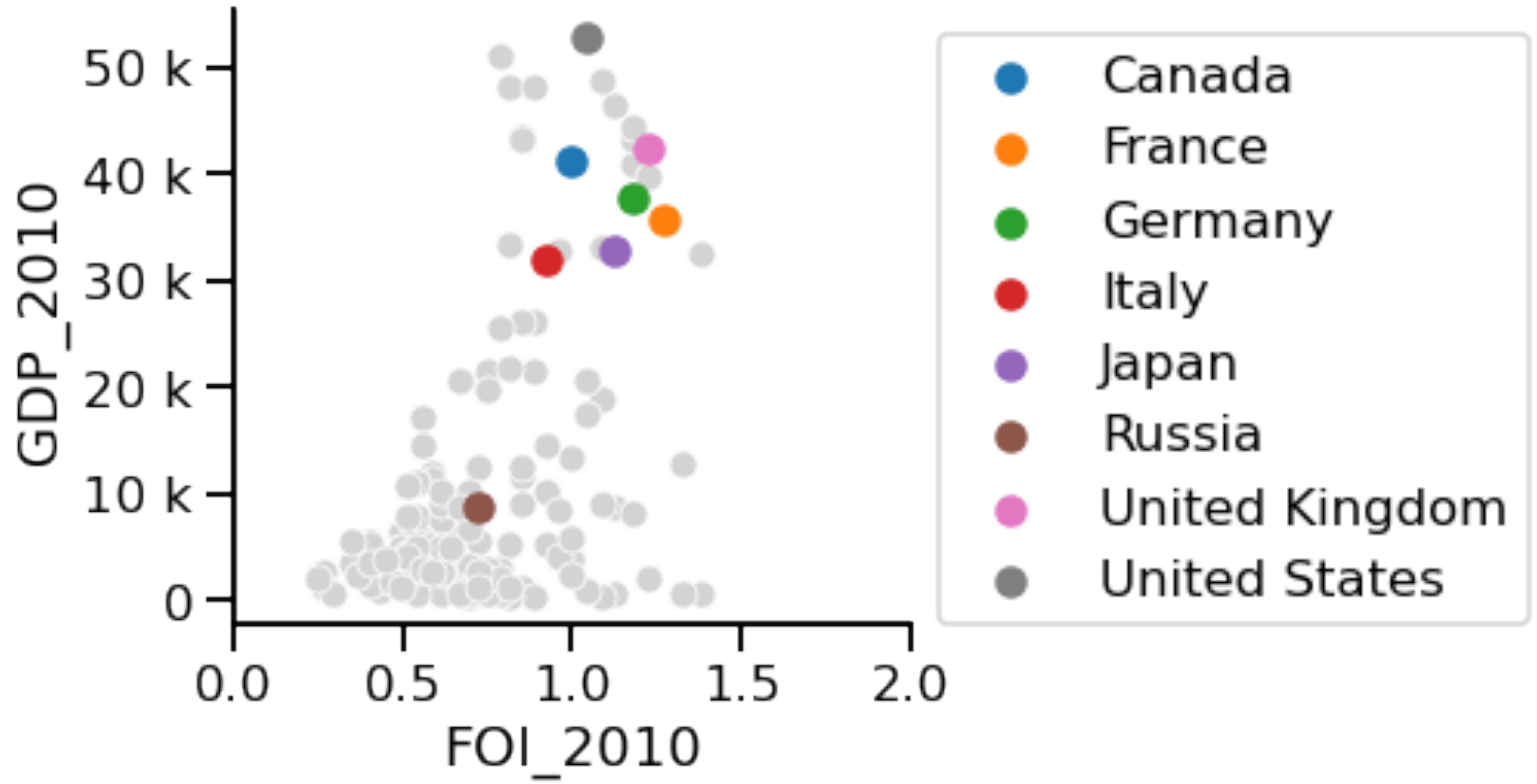


My replication
140 countries

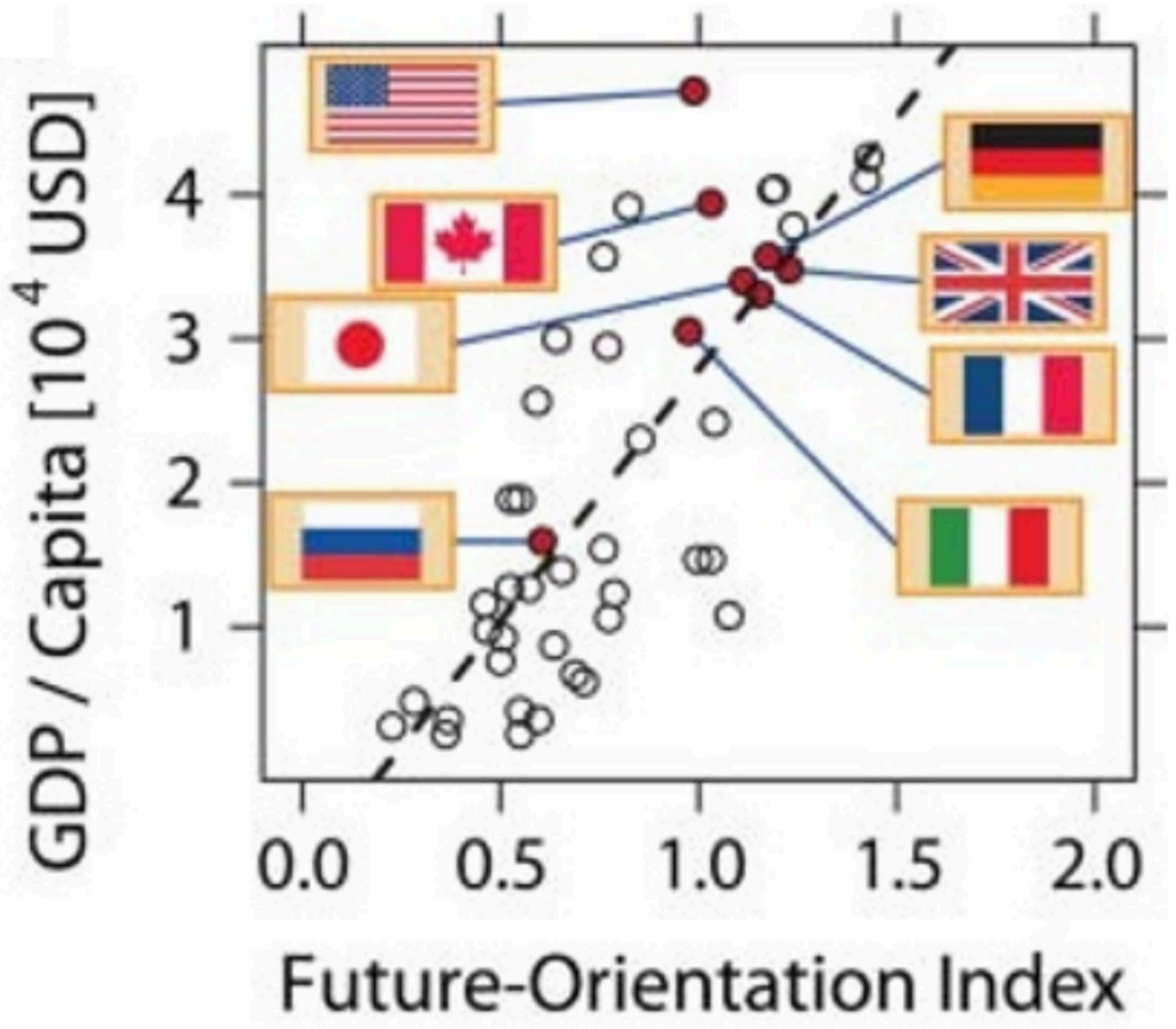
Examples of FOI

Replicating results from Preis et al. 2012

a)



b)



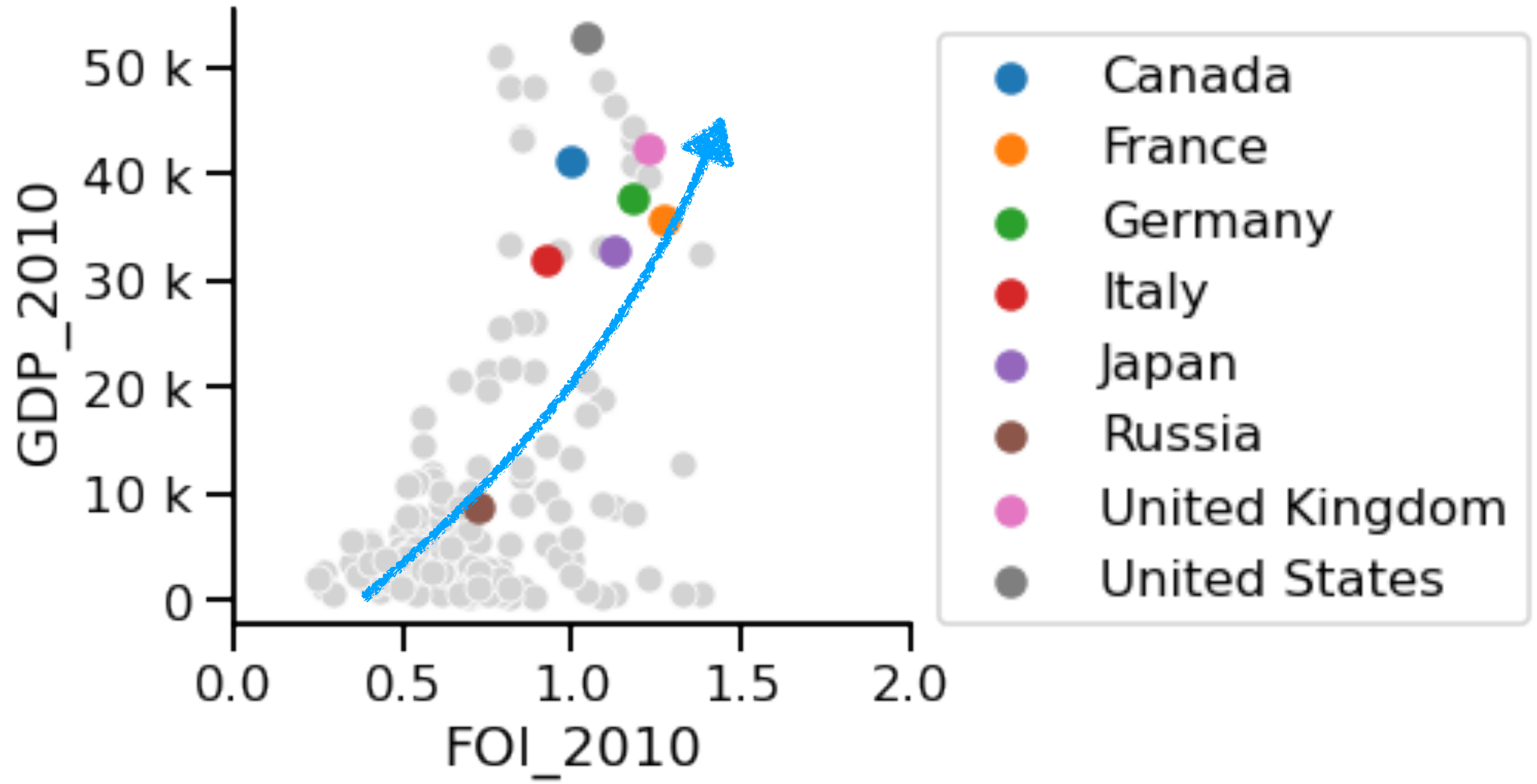
My replication
140 countries

Preis et al. 2012
45 countries

Examples of FOI

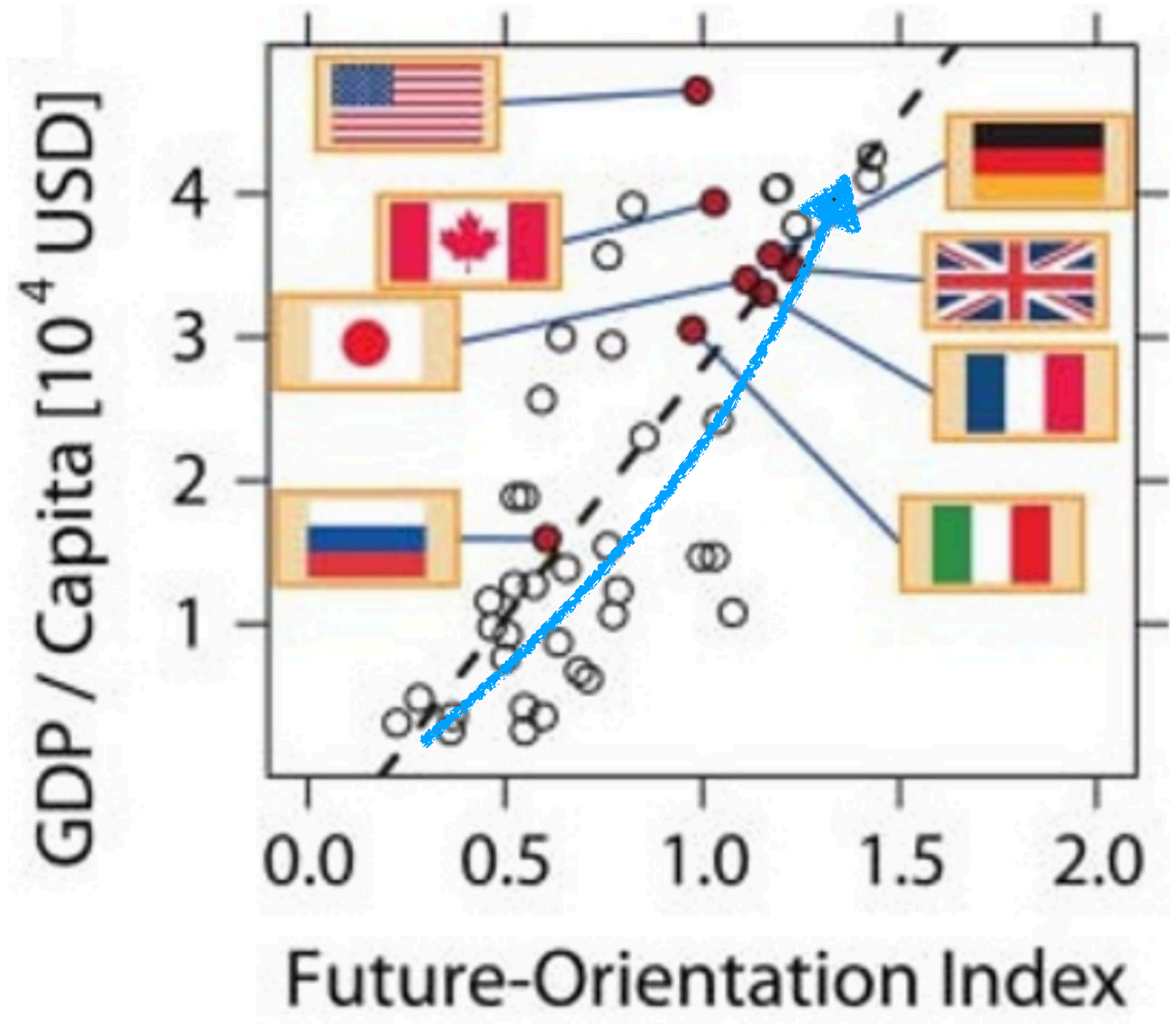
Replicating results from Preis et al. 2012

a)



My replication
140 countries

b)

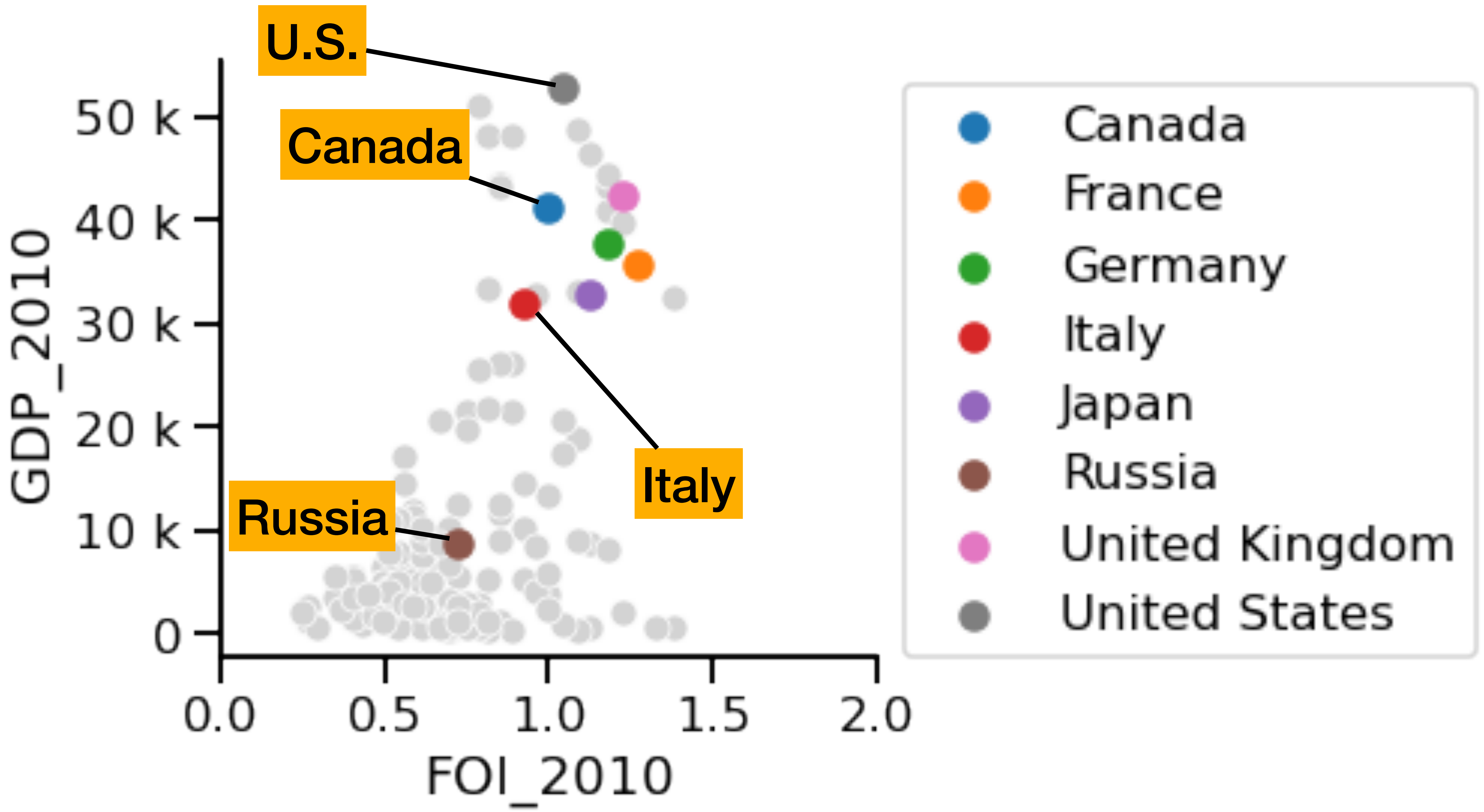


Preis et al. 2012
45 countries

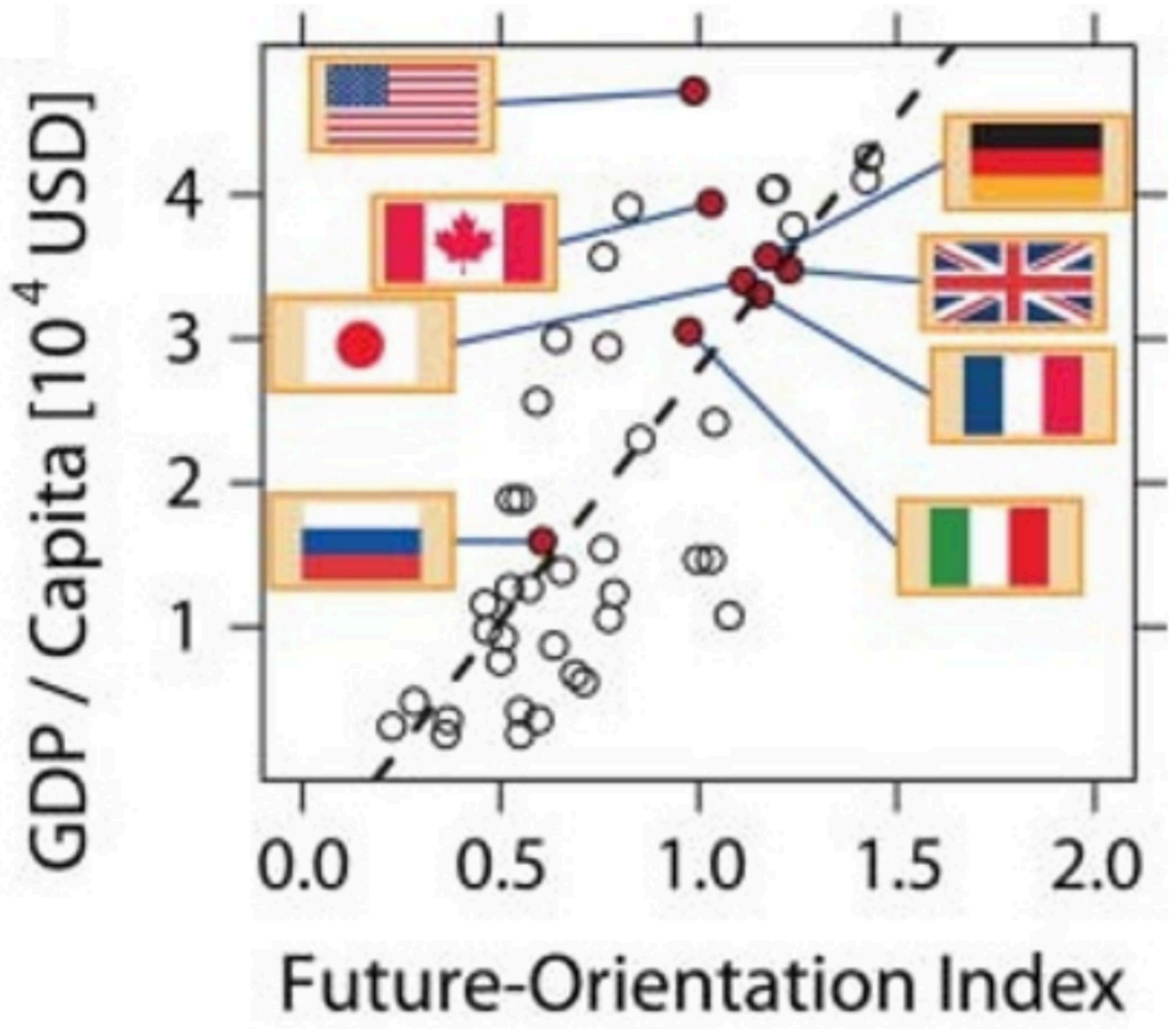
Examples of FOI

Replicating results from Preis et al. 2012

a)



b)



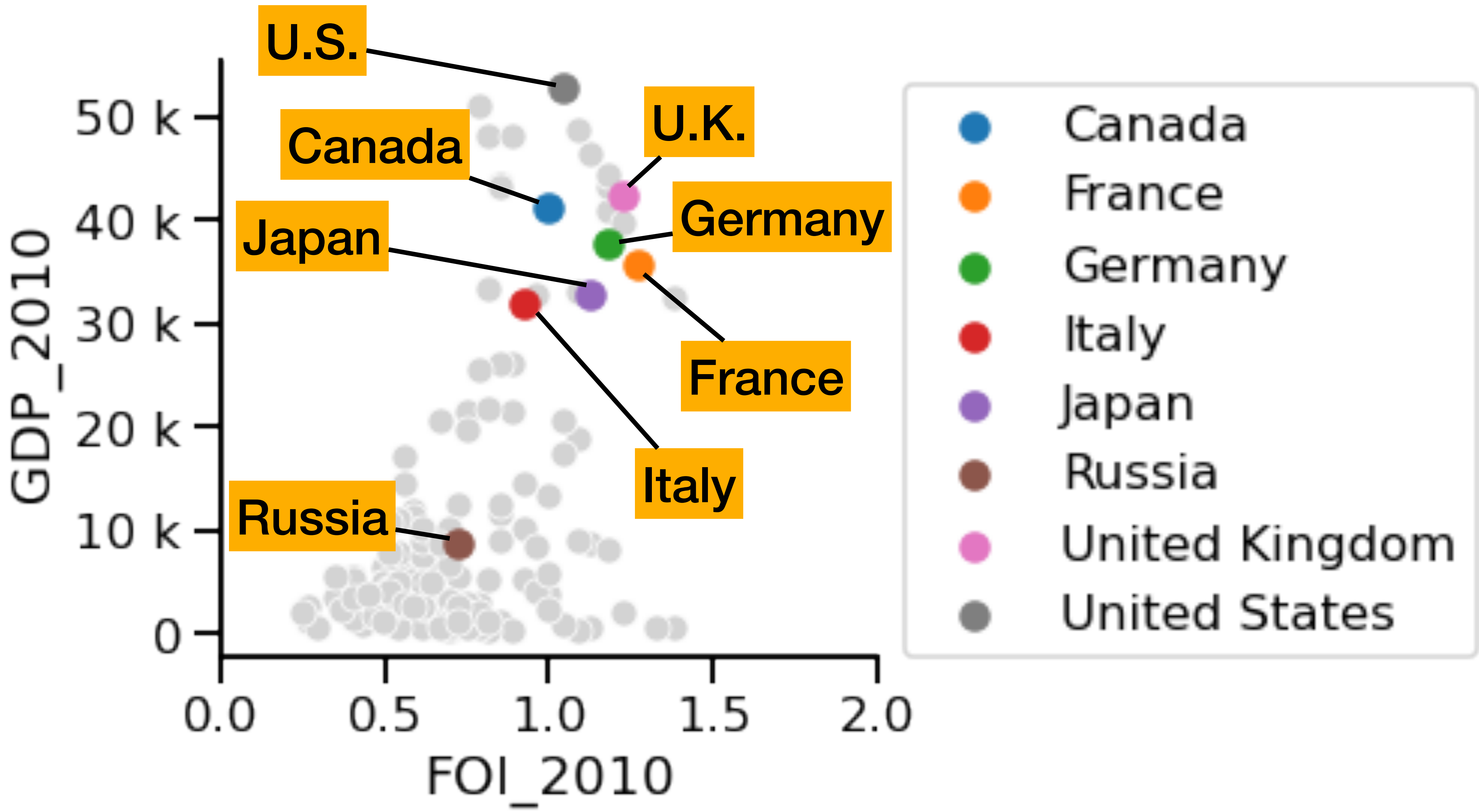
My replication
140 countries

Preis et al. 2012
45 countries

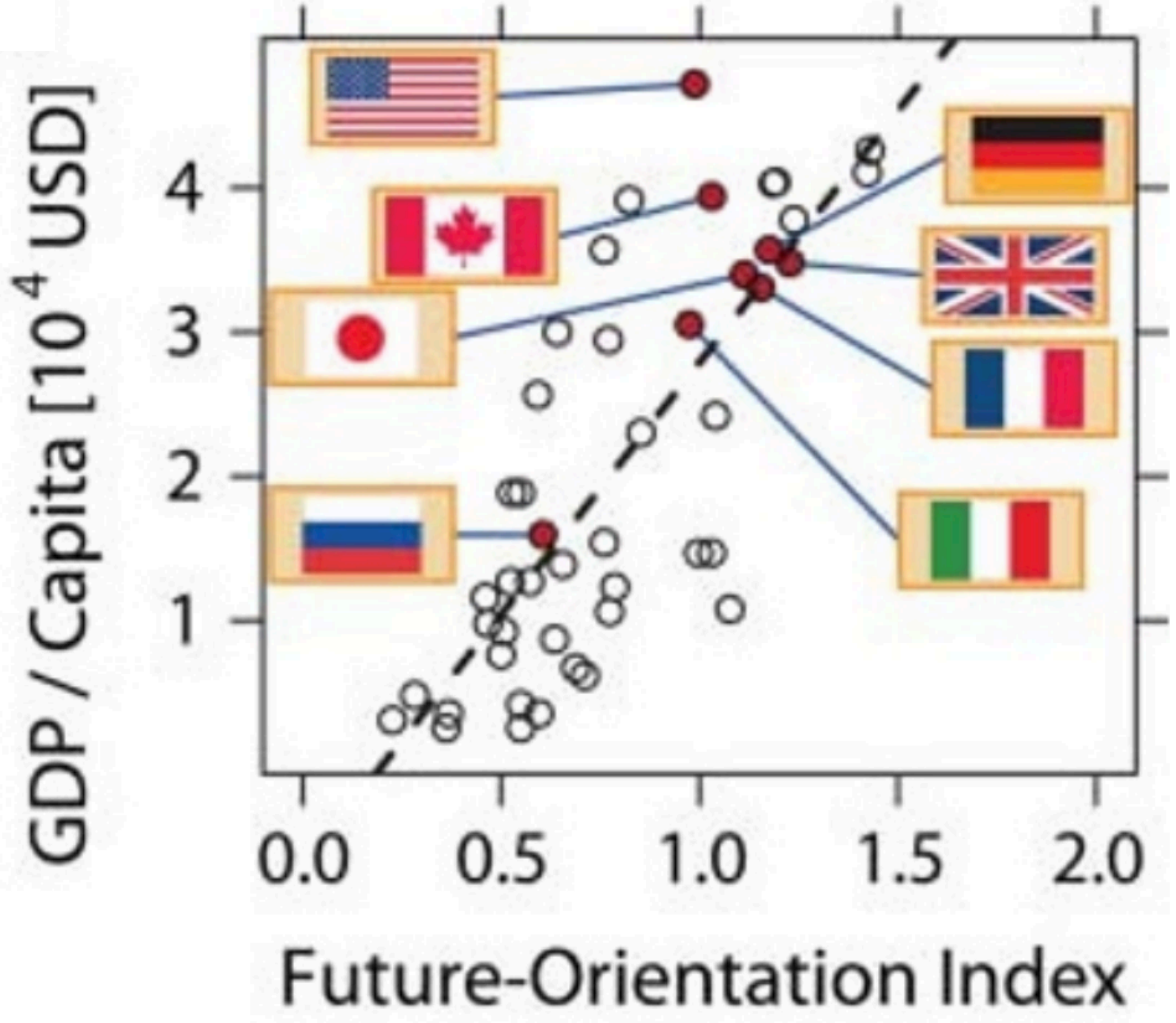
Examples of FOI

Replicating results from Preis et al. 2012

a)



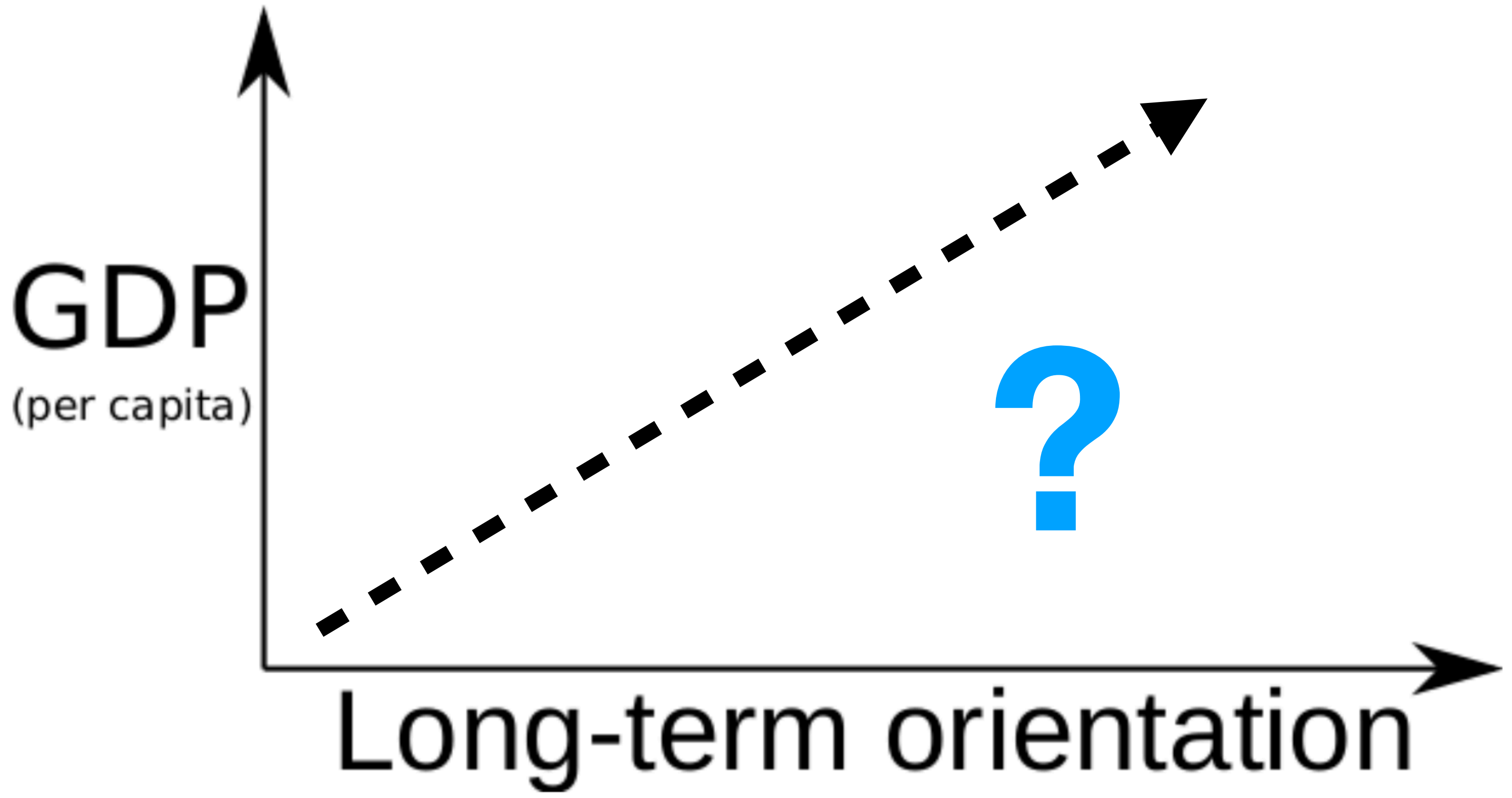
b)



My replication
140 countries

Preis et al. 2012
45 countries

Can this **relationship** (pattern) be numerically measured?



Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Search Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

Measuring Correlation

What is correlation?

Definition

What is correlation?

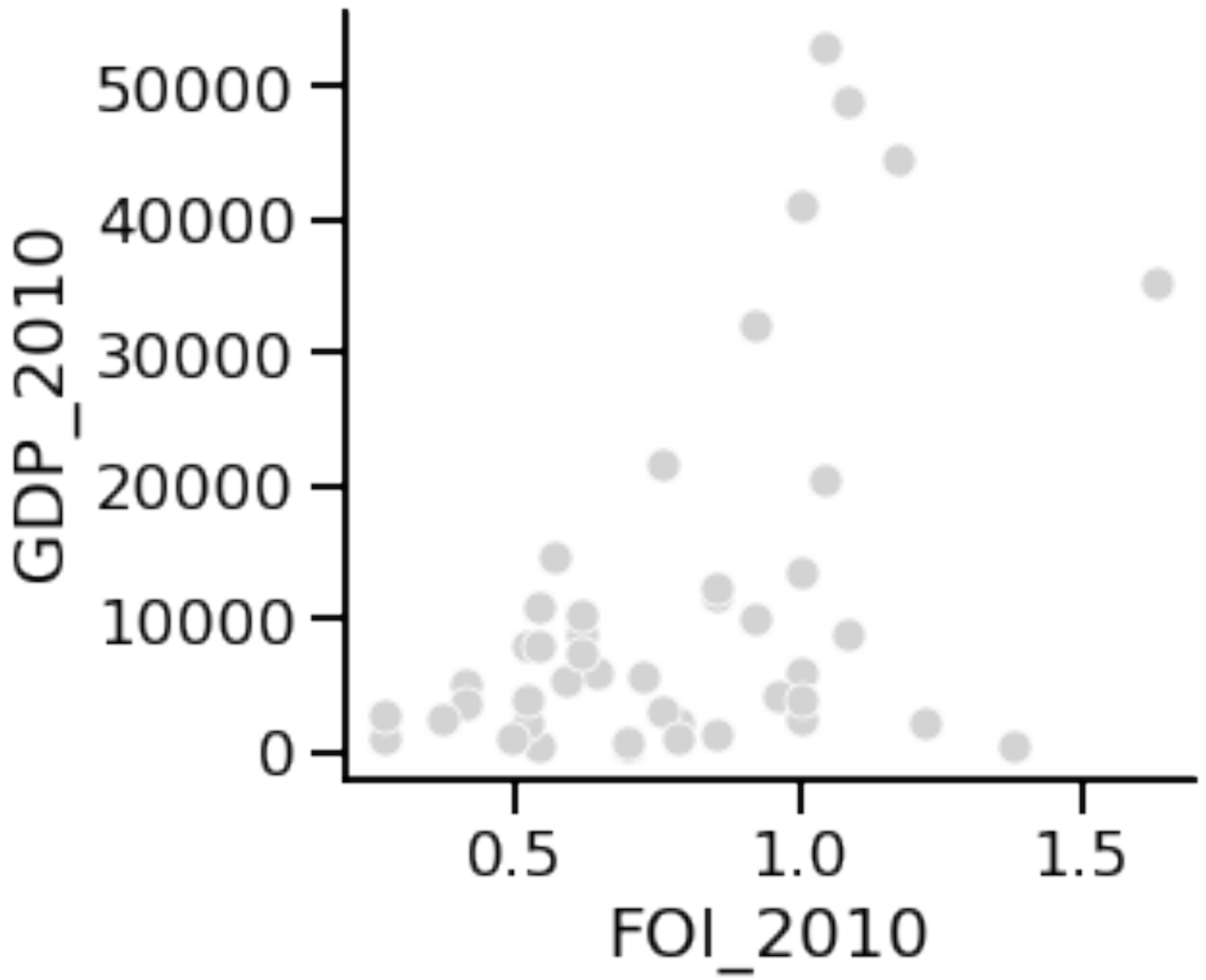
Definition

- Correlation means there is a relationship or pattern between the values of two variables.

What is correlation?

Definition

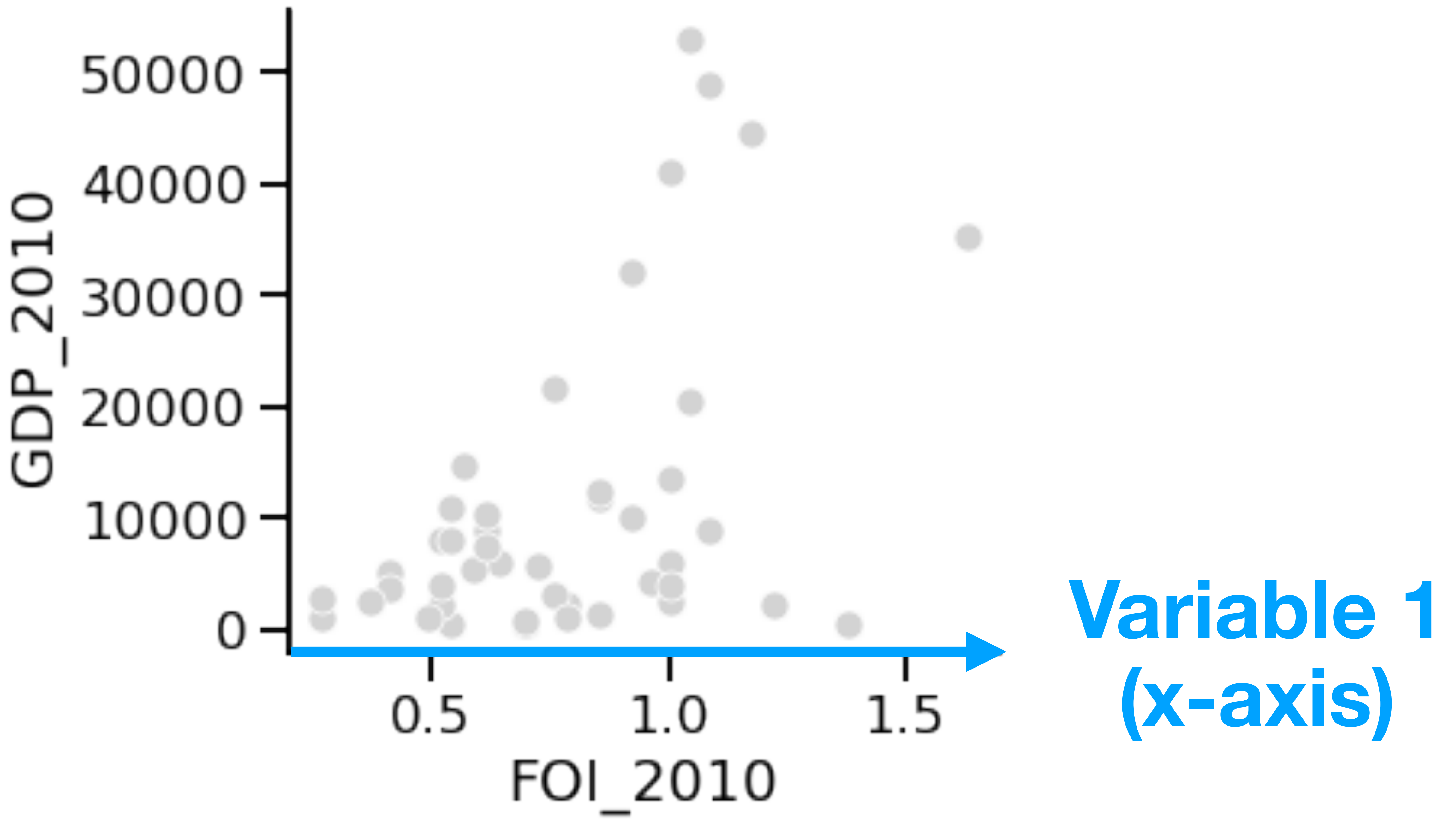
- Correlation means there is a relationship or pattern between the values of two variables.



What is correlation?

Definition

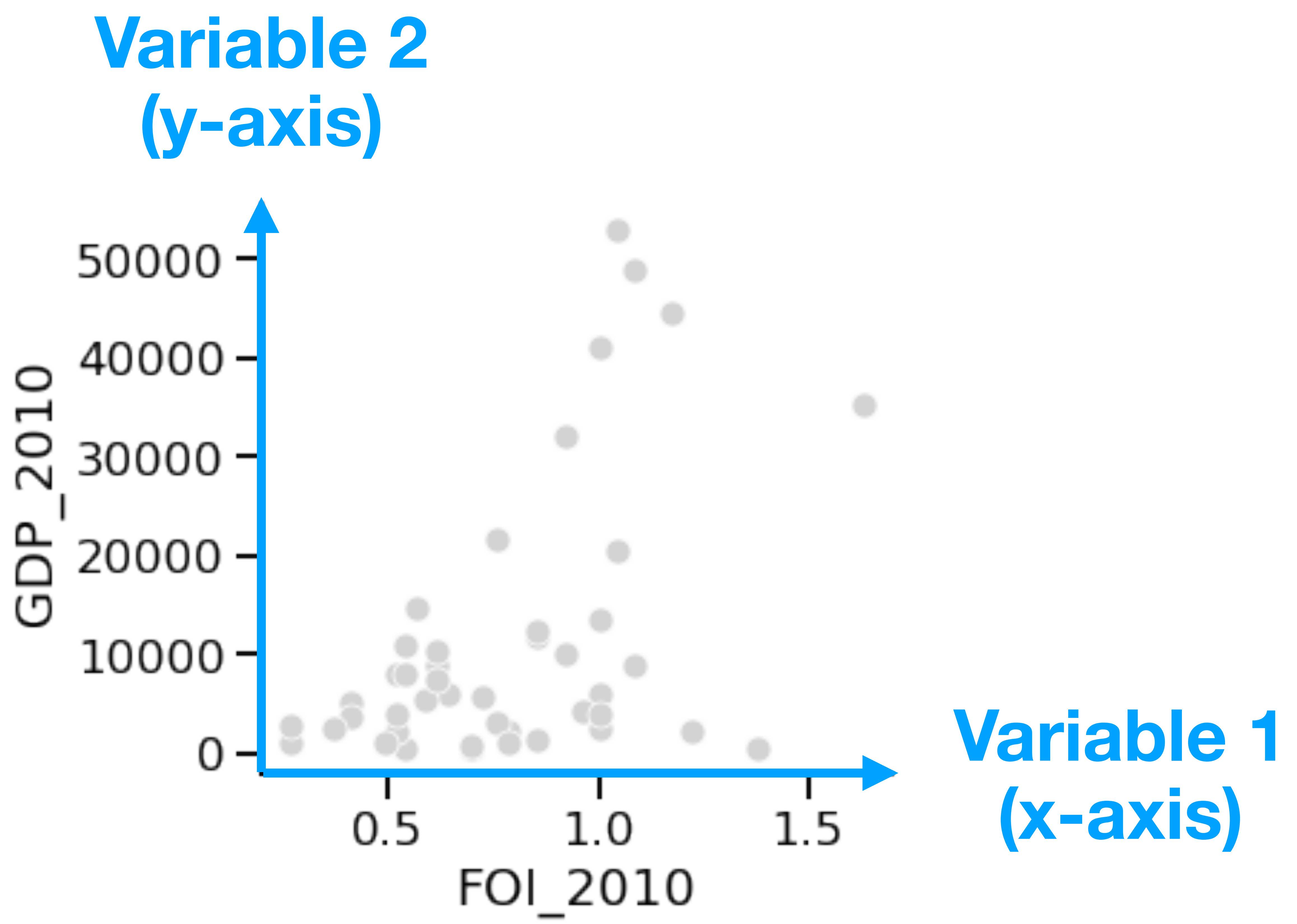
- Correlation means there is a relationship or pattern between the values of two variables.



What is correlation?

Definition

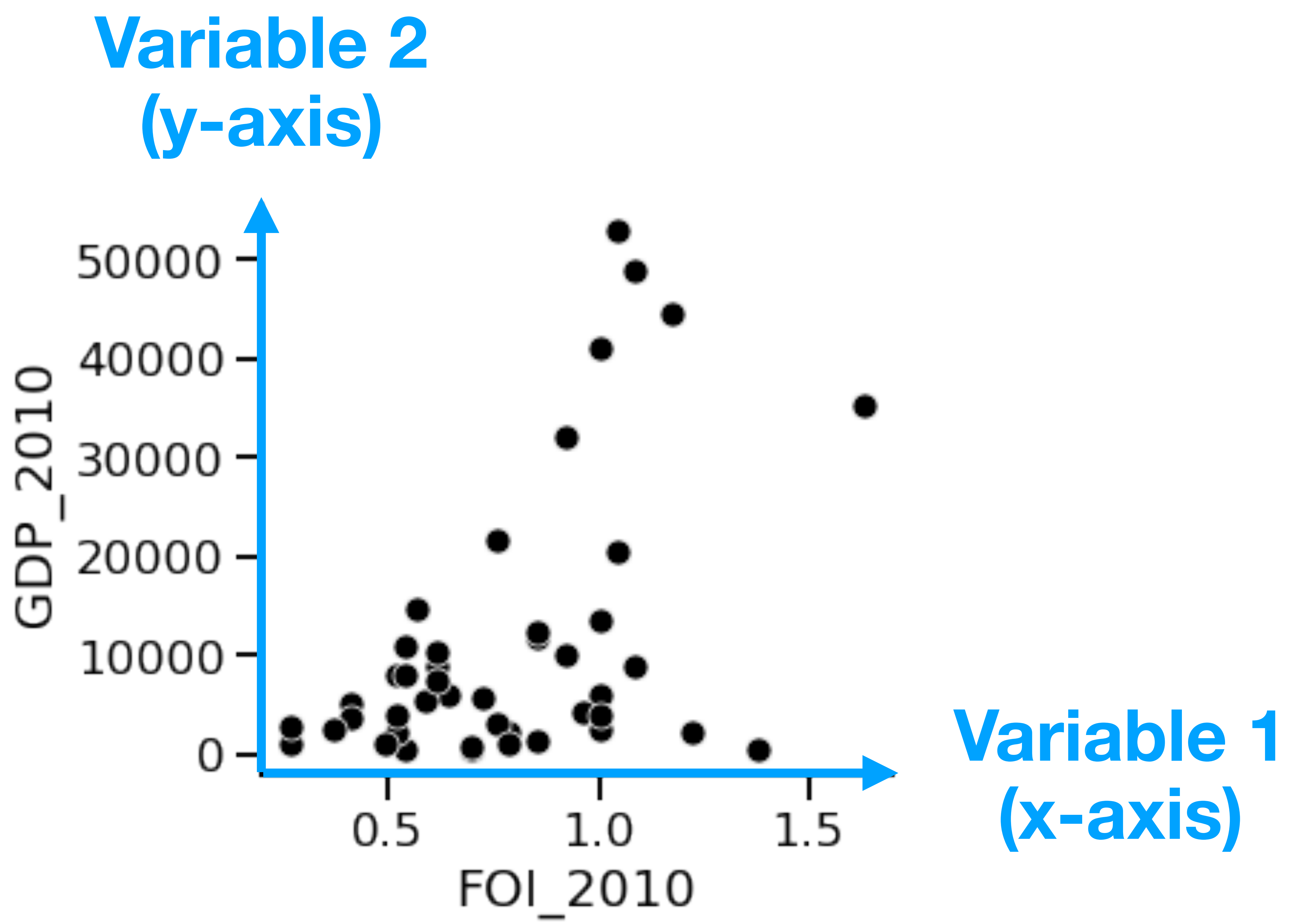
- Correlation means there is a relationship or pattern between the values of two variables.



What is correlation?

Definition

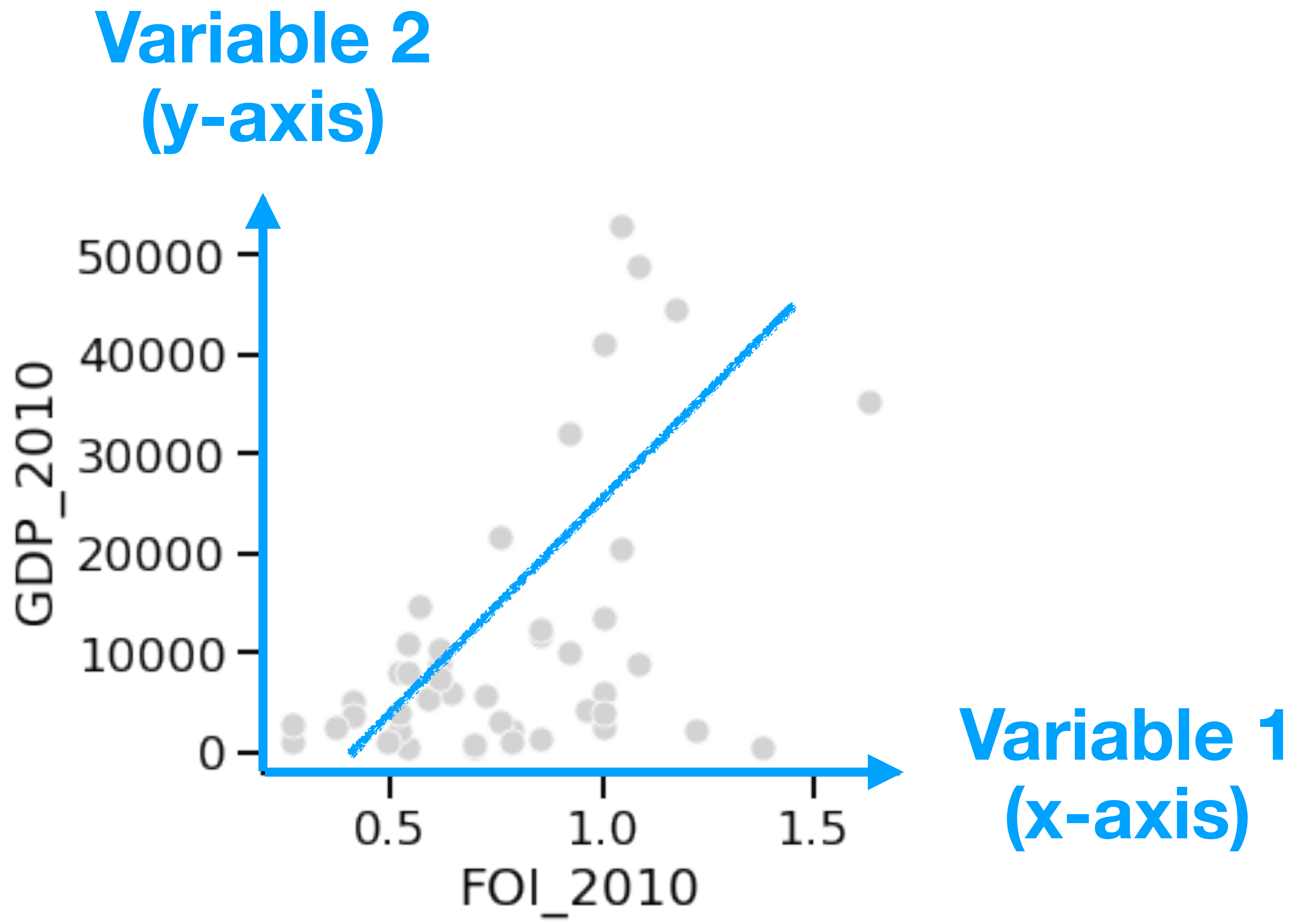
- Correlation means there is a relationship or pattern between the values of two variables.
- Correlation is often visually represented using a scatter plot (dots).



What is correlation?

Definition

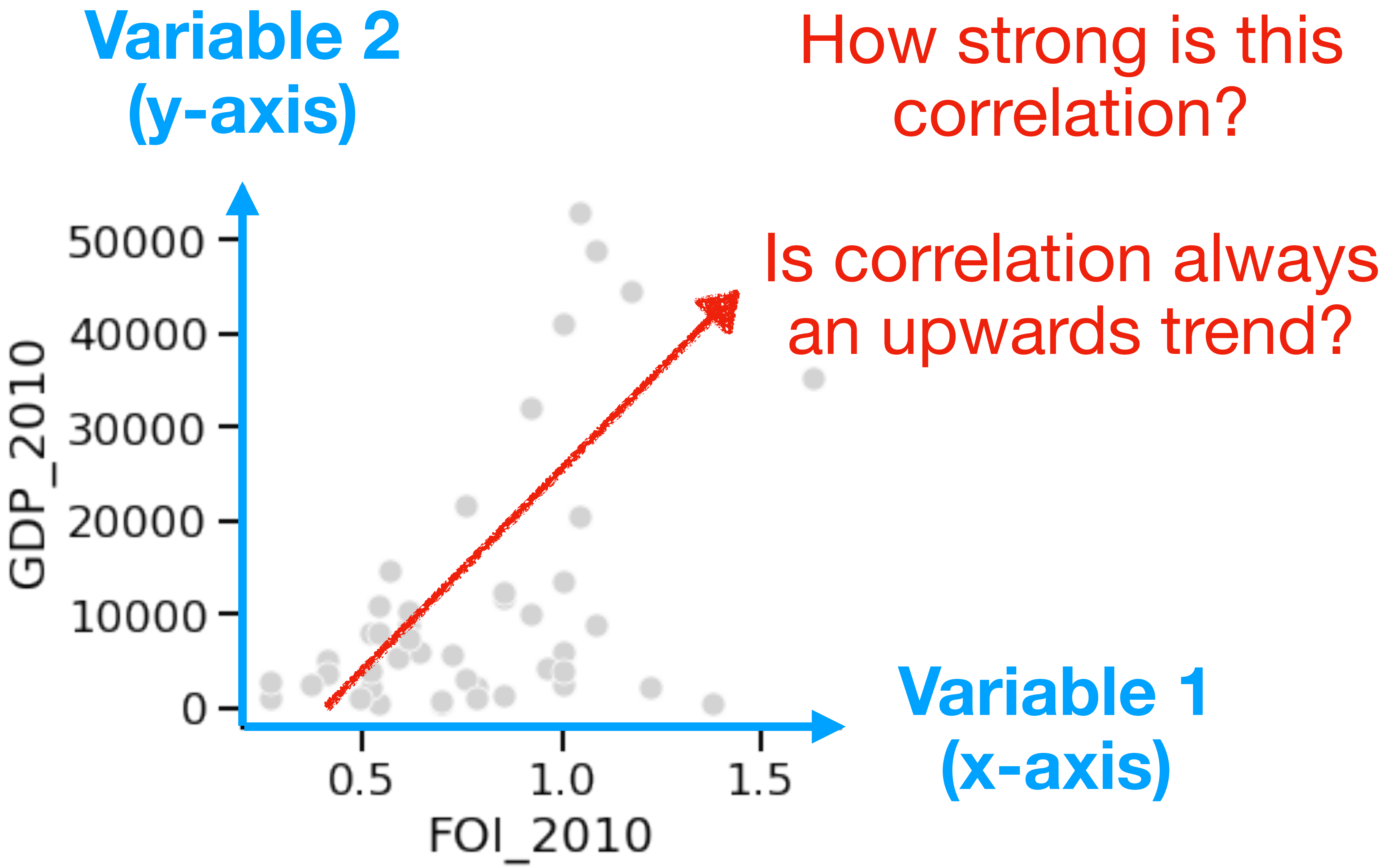
- Correlation means there is a relationship or pattern between the values of two variables.
- Correlation is often visually represented using a scatter plot (dots).
- Correlation measures the **strength** of the relationship. The stronger the correlation, the closer the data points are to a straight line.



What is correlation?

Definition

- Correlation means there is a relationship or pattern between the values of two variables.
- Correlation is often visually represented using a scatter plot (dots).
- Correlation measures the **strength** of the relationship. The stronger the correlation, the closer the data points are to a straight line.



Types of correlation

Direction of the trend

Types of correlation

Direction of the trend

Correlation also measures the **direction** of the relationship:

Types of correlation

Direction of the trend

Correlation also measures the **direction** of the relationship:

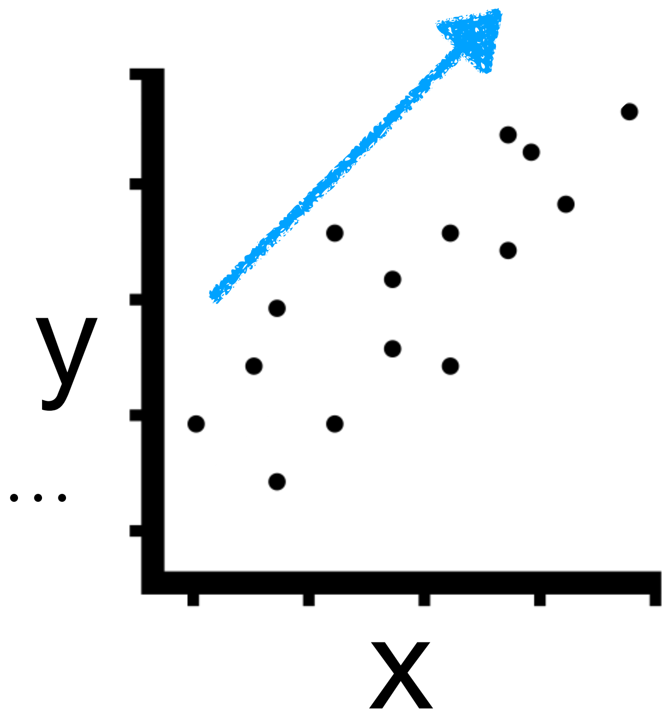
- **Positive (direct) correlation**: as one variable <u>increases</u>, the other also <u>increases</u> (and vice versa).

Types of correlation

Direction of the trend

Correlation also measures the **direction** of the relationship:

- **Positive (direct) correlation**: as one variable increases, the other also increases (and vice versa).

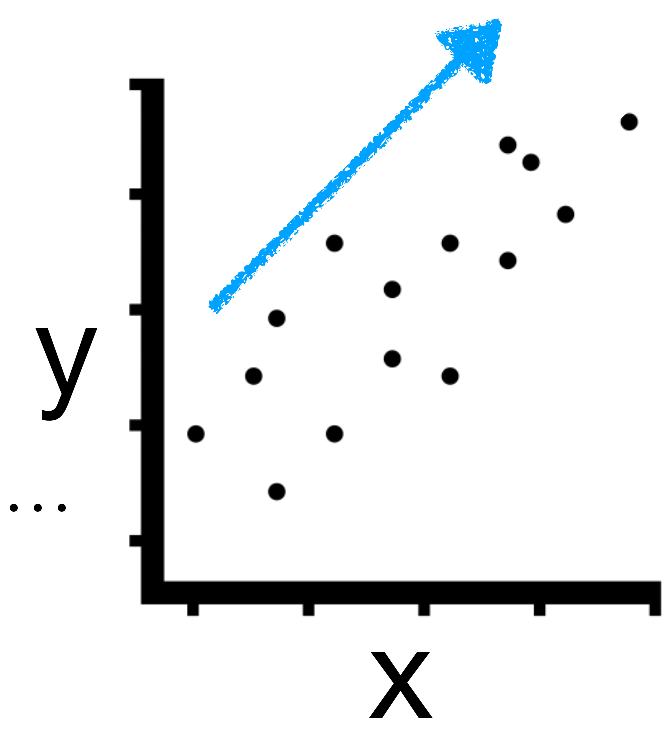


Types of correlation

Direction of the trend

Correlation also measures the **direction** of the relationship:

- **Positive (direct) correlation**: as one variable <u>increases</u>, the other also <u>increases</u> (and vice versa).



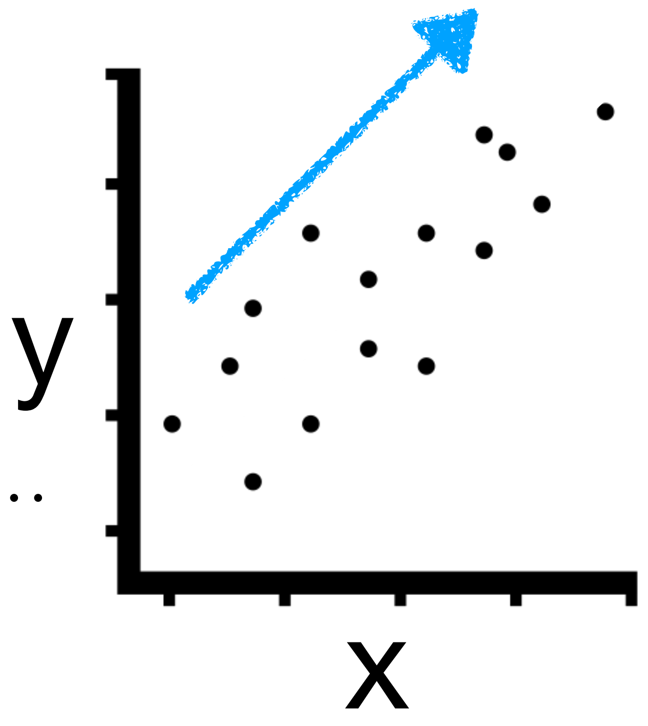
- **Negative (inverse) correlation**: as one variable <u>increases</u>, the other <u>decreases</u> (and vice versa).

Types of correlation

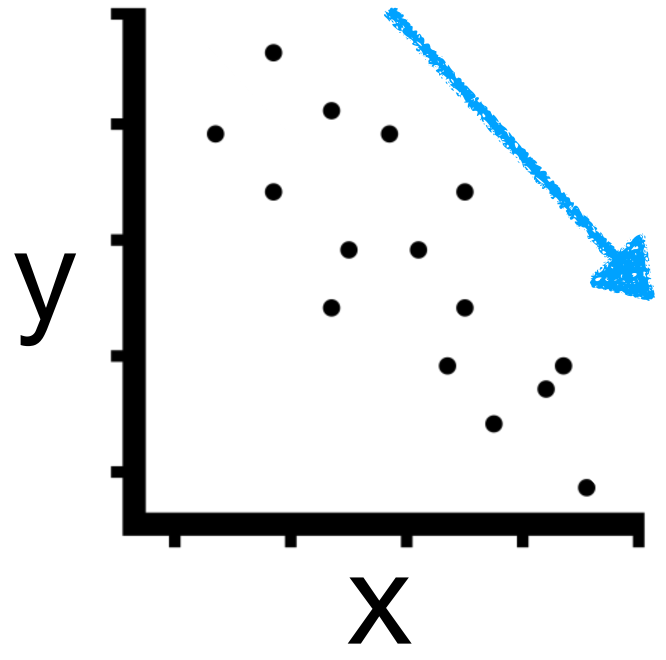
Direction of the trend

Correlation also measures the **direction** of the relationship:

- **Positive (direct) correlation**: as one variable <u>increases</u>, the other also <u>increases</u> (and vice versa).



- **Negative (inverse) correlation**: as one variable <u>increases</u>, the other <u>decreases</u> (and vice versa).

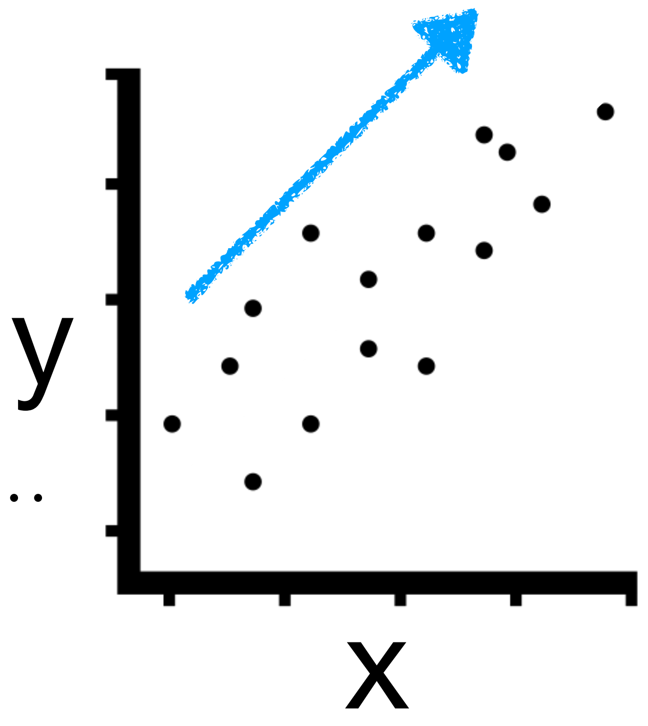


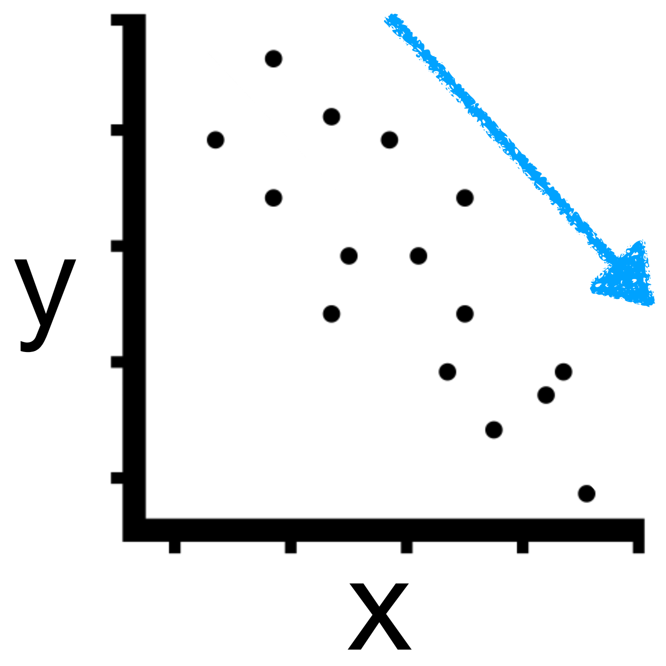
Types of correlation

Direction of the trend

Correlation also measures the **direction** of the relationship:

- **Positive (direct) correlation**: as one variable increases, the other also increases (and vice versa).
- **Negative (inverse) correlation**: as one variable increases, the other decreases (and vice versa).
- **No correlation**: no apparent relationship between the variables.



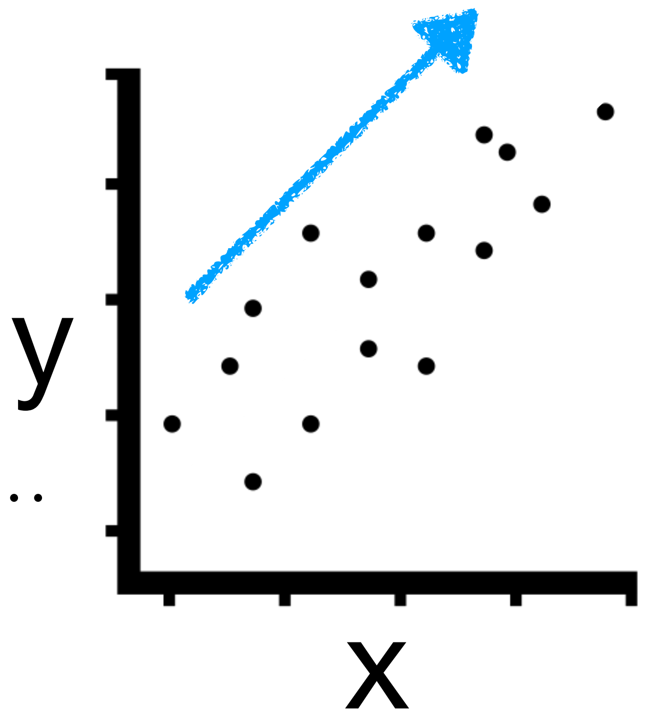


Types of correlation

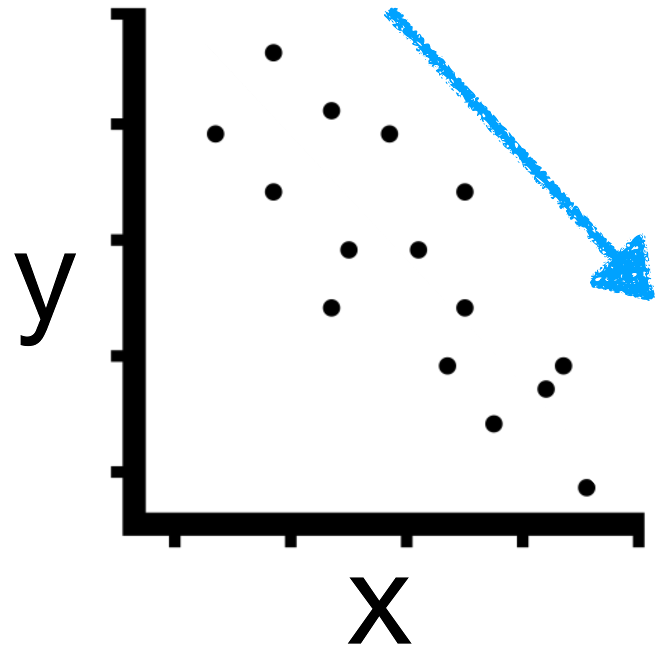
Direction of the trend

Correlation also measures the **direction** of the relationship:

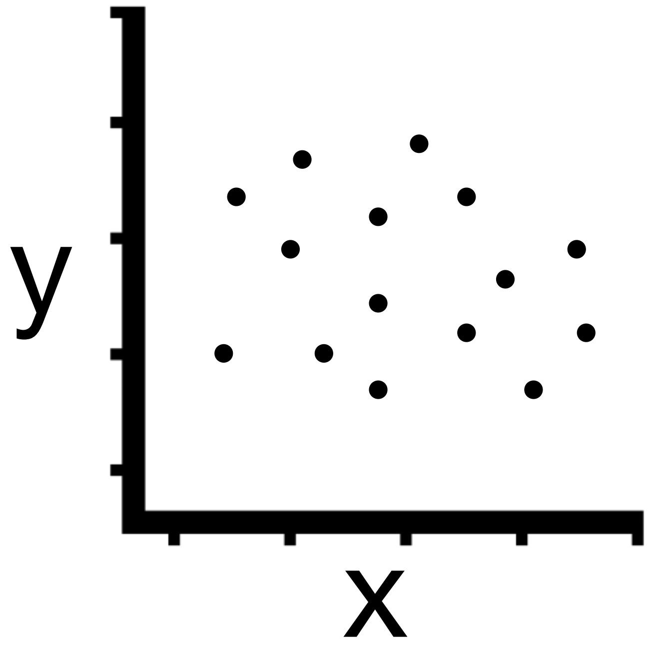
- **Positive (direct) correlation**: as one variable increases, the other also increases (and vice versa).



- **Negative (inverse) correlation**: as one variable increases, the other decreases (and vice versa).

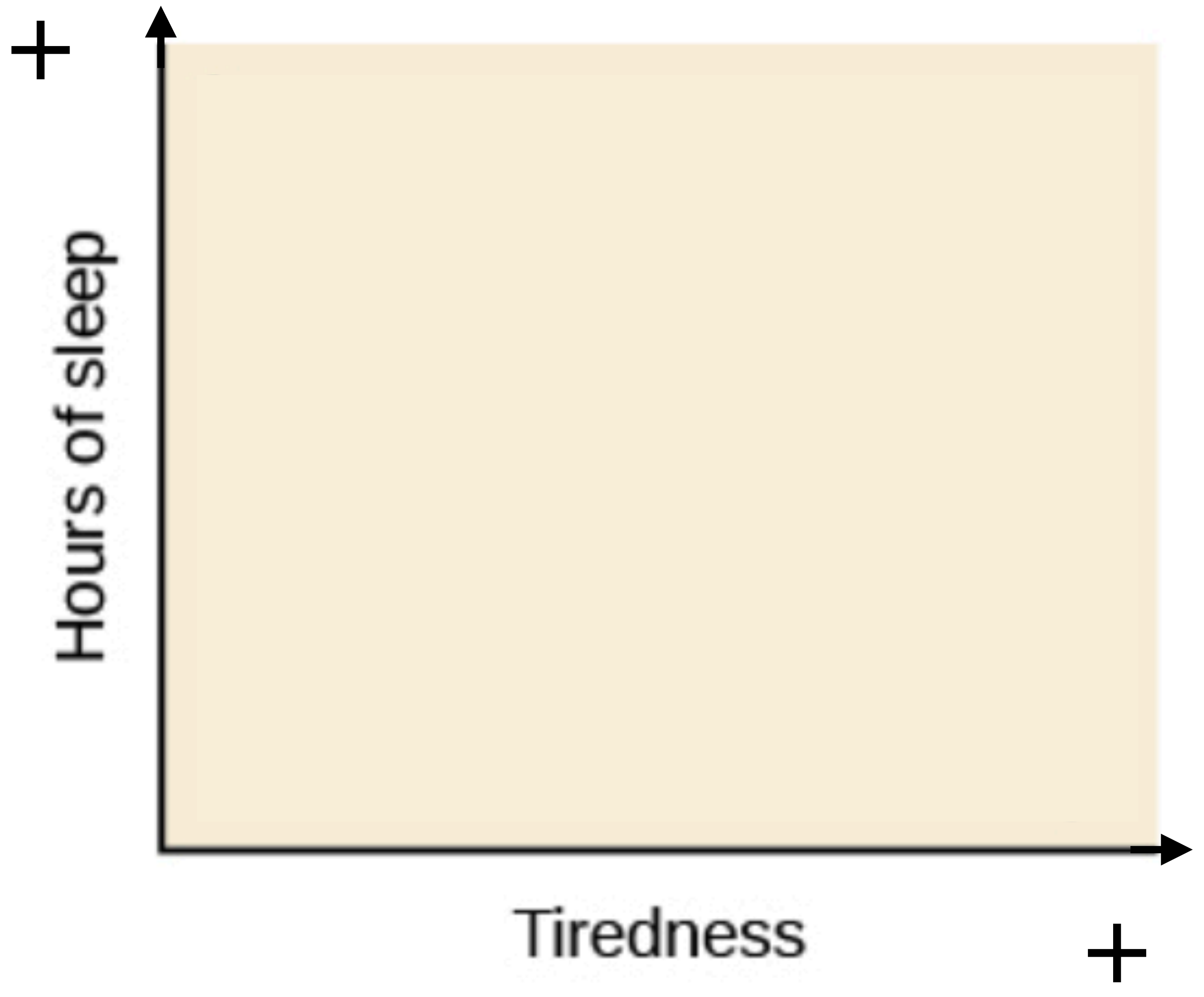


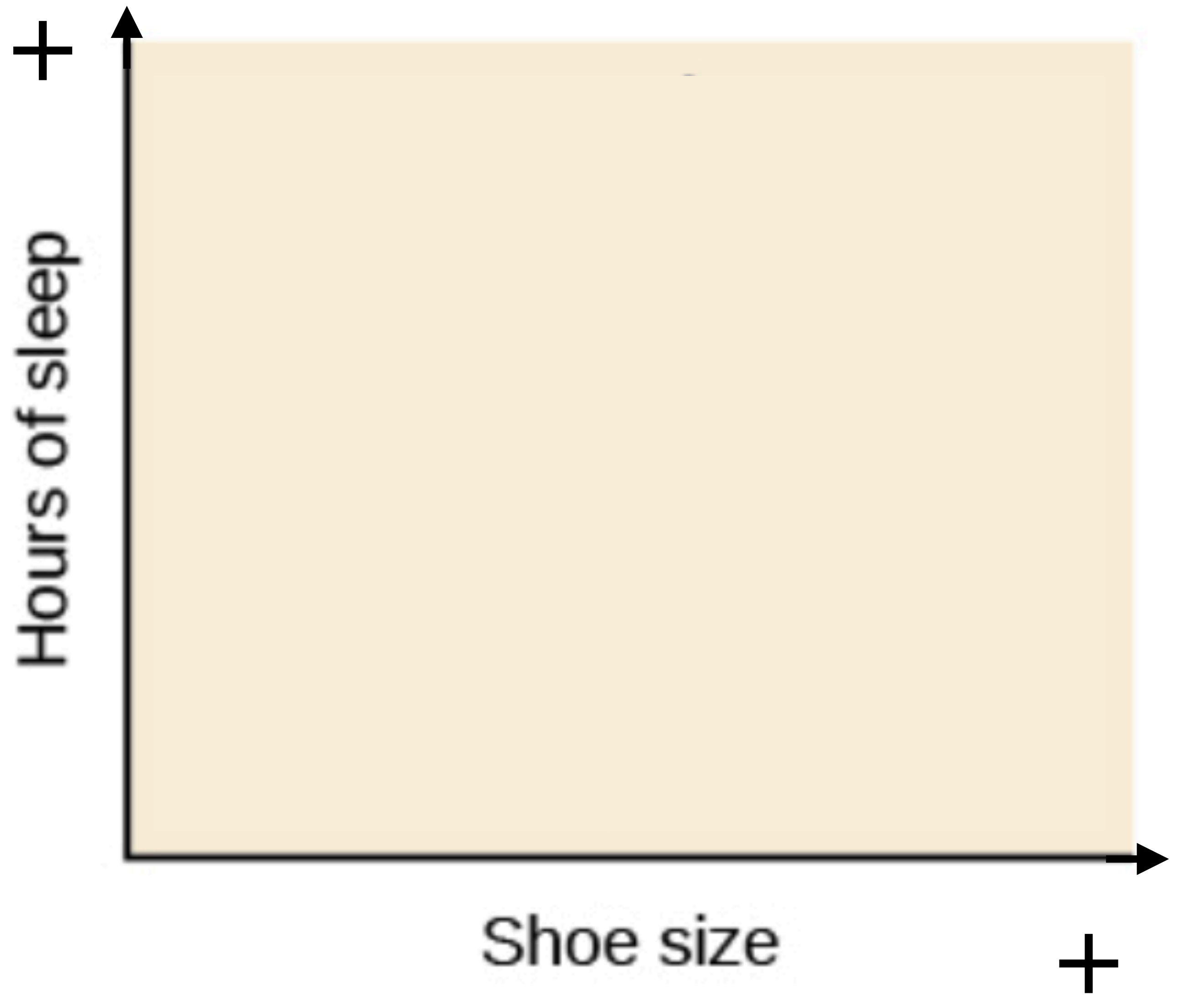
- **No correlation**: no apparent relationship between the variables.

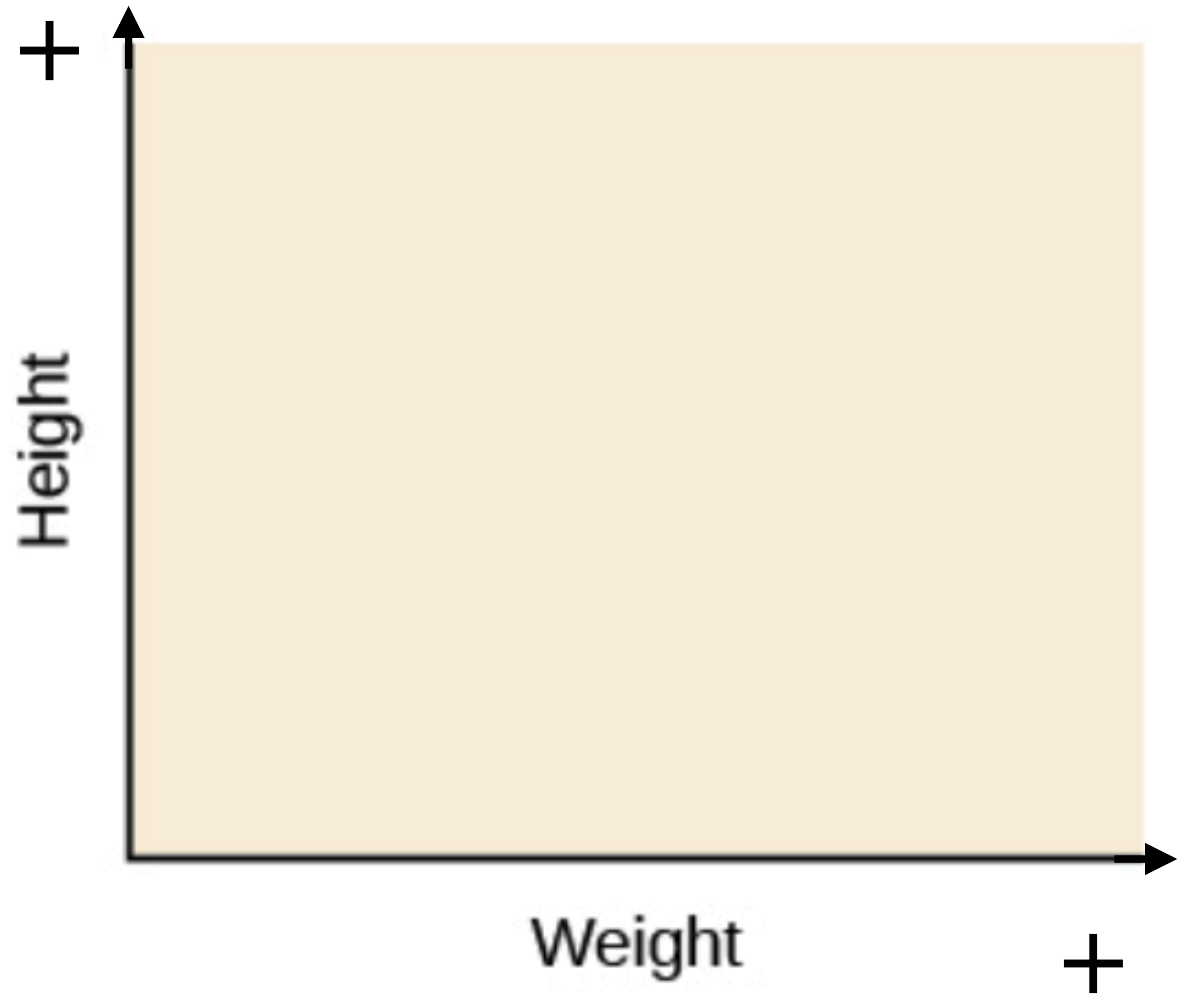


Examples of correlation

Positive? Negative? None?

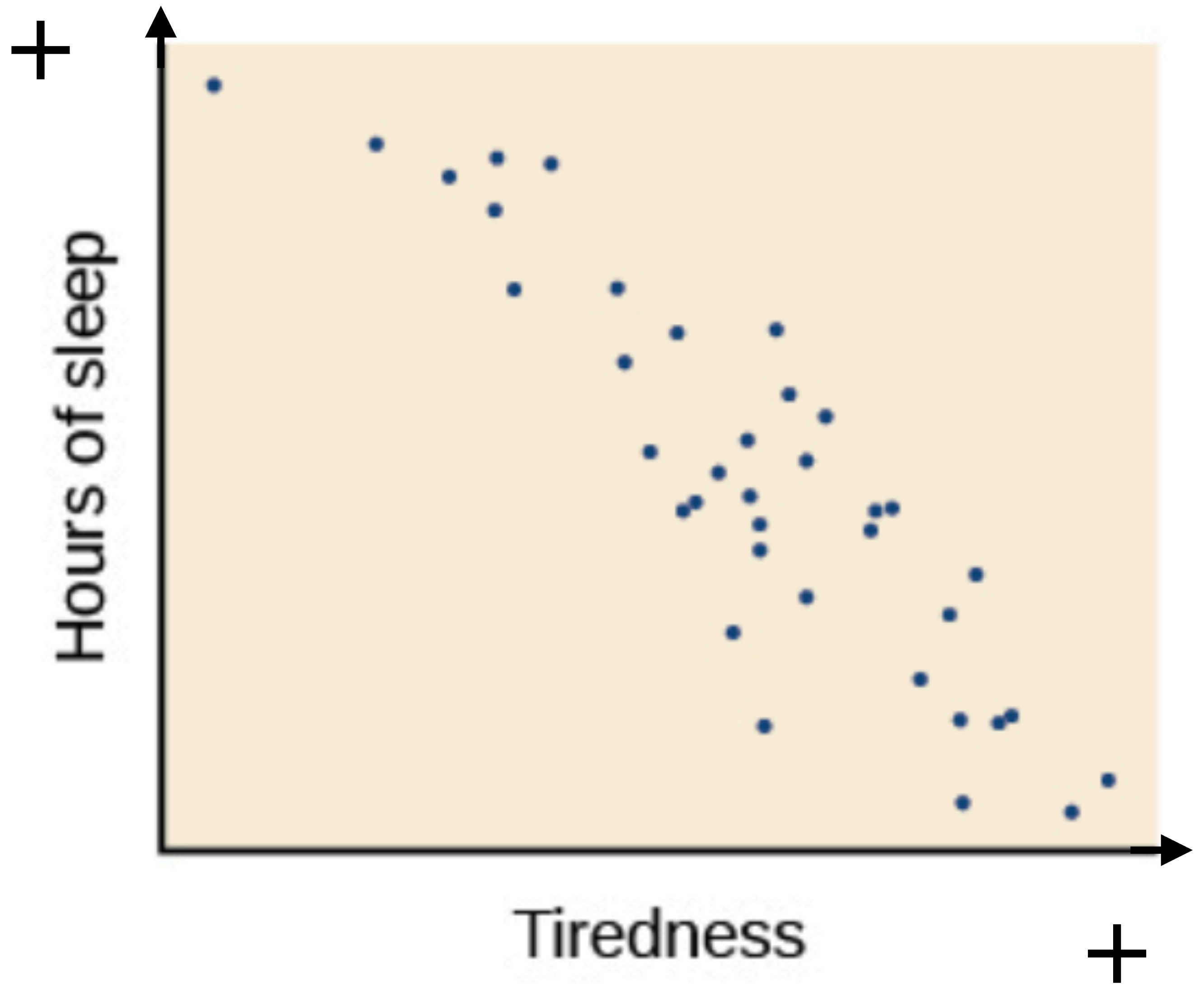


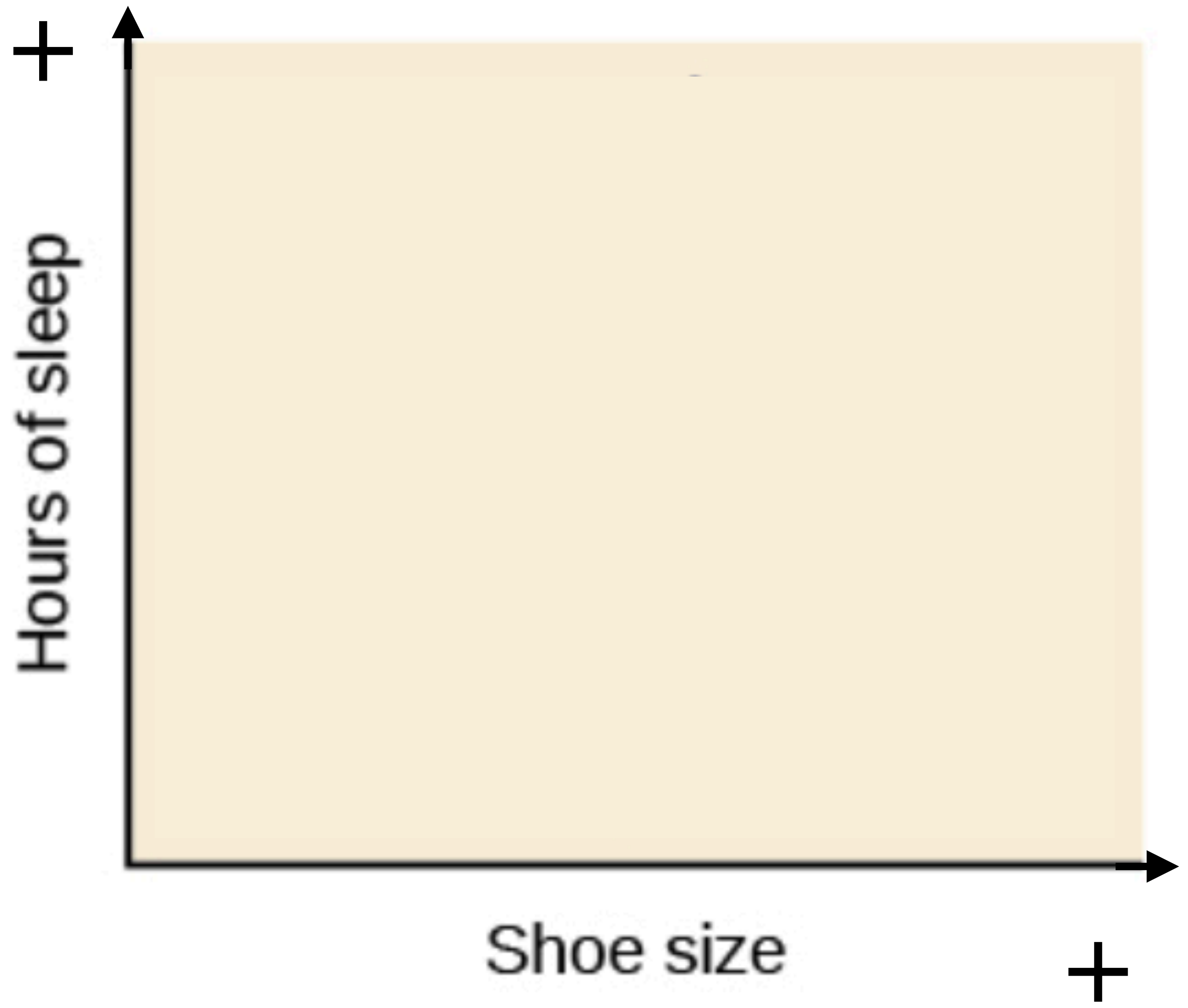


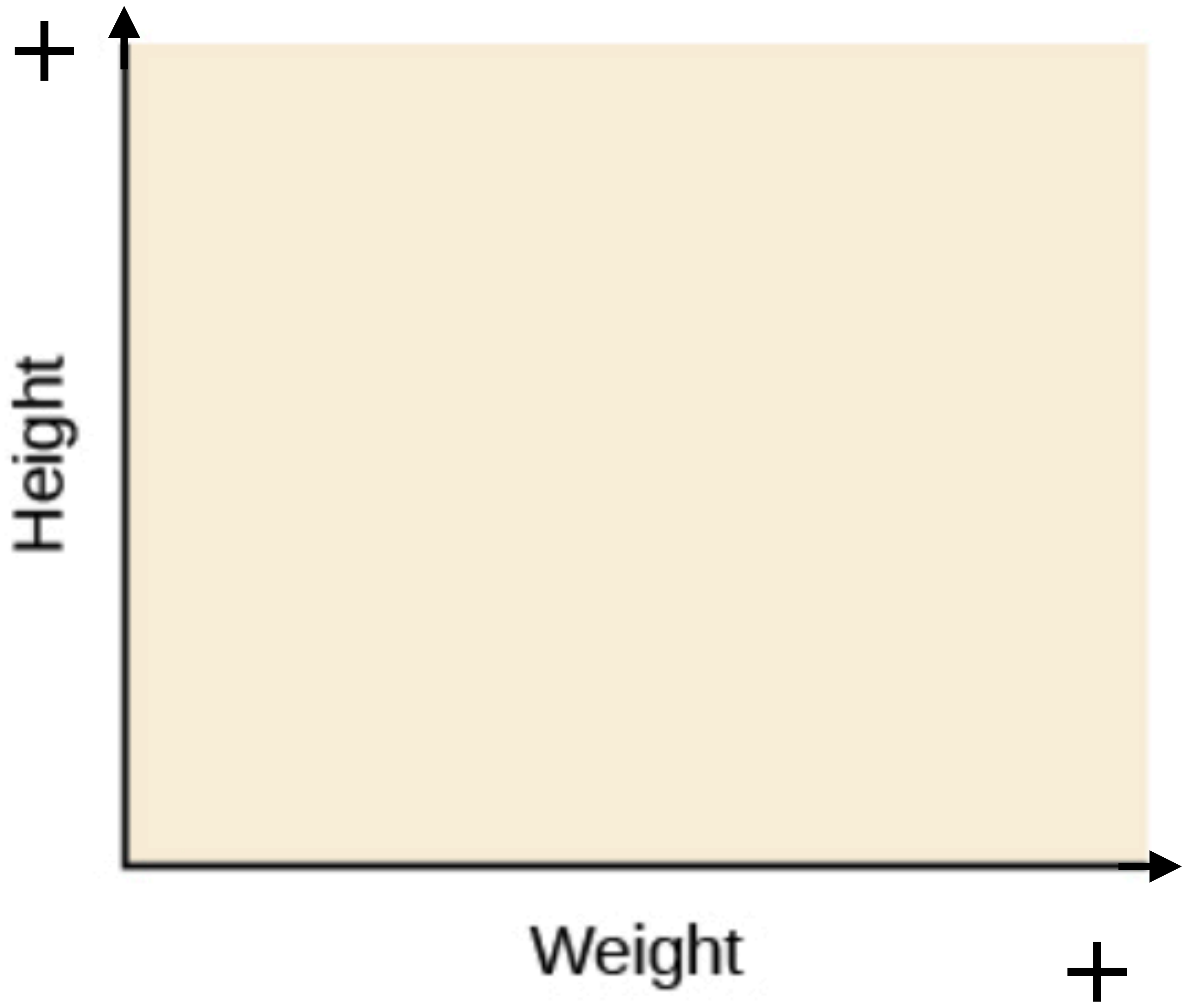


Examples of correlation

Positive? Negative? None?



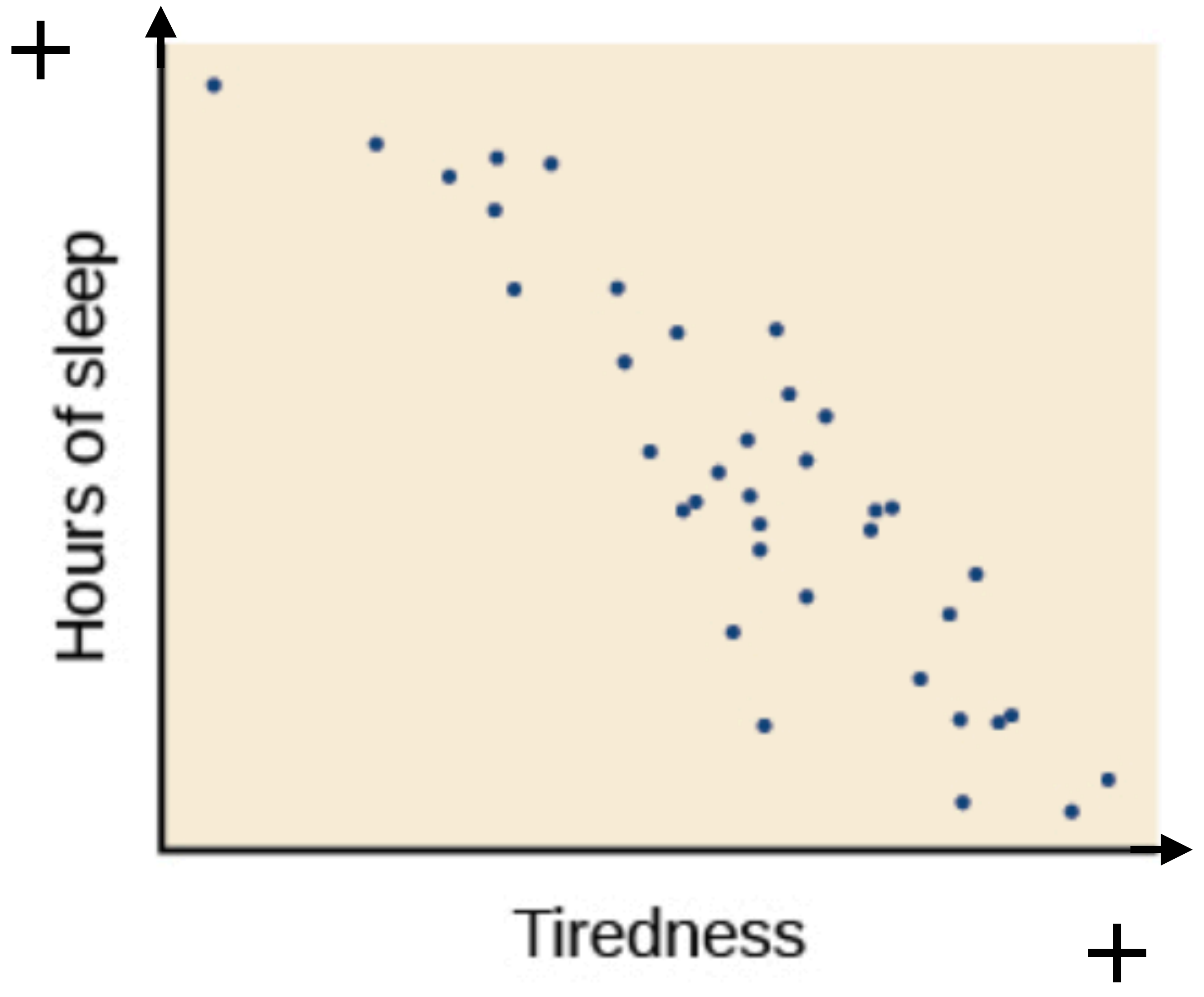




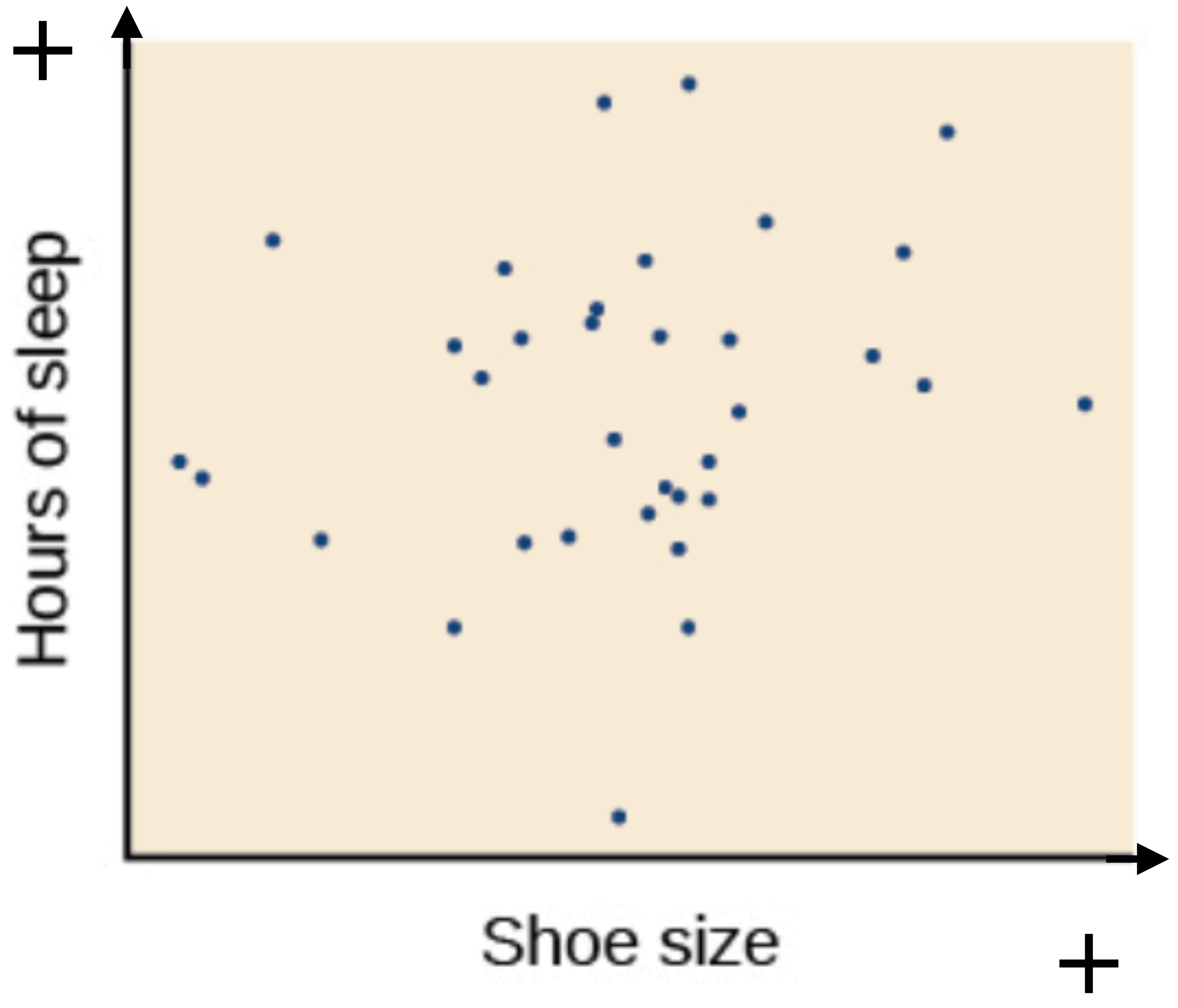
Negative
correlation

Examples of correlation

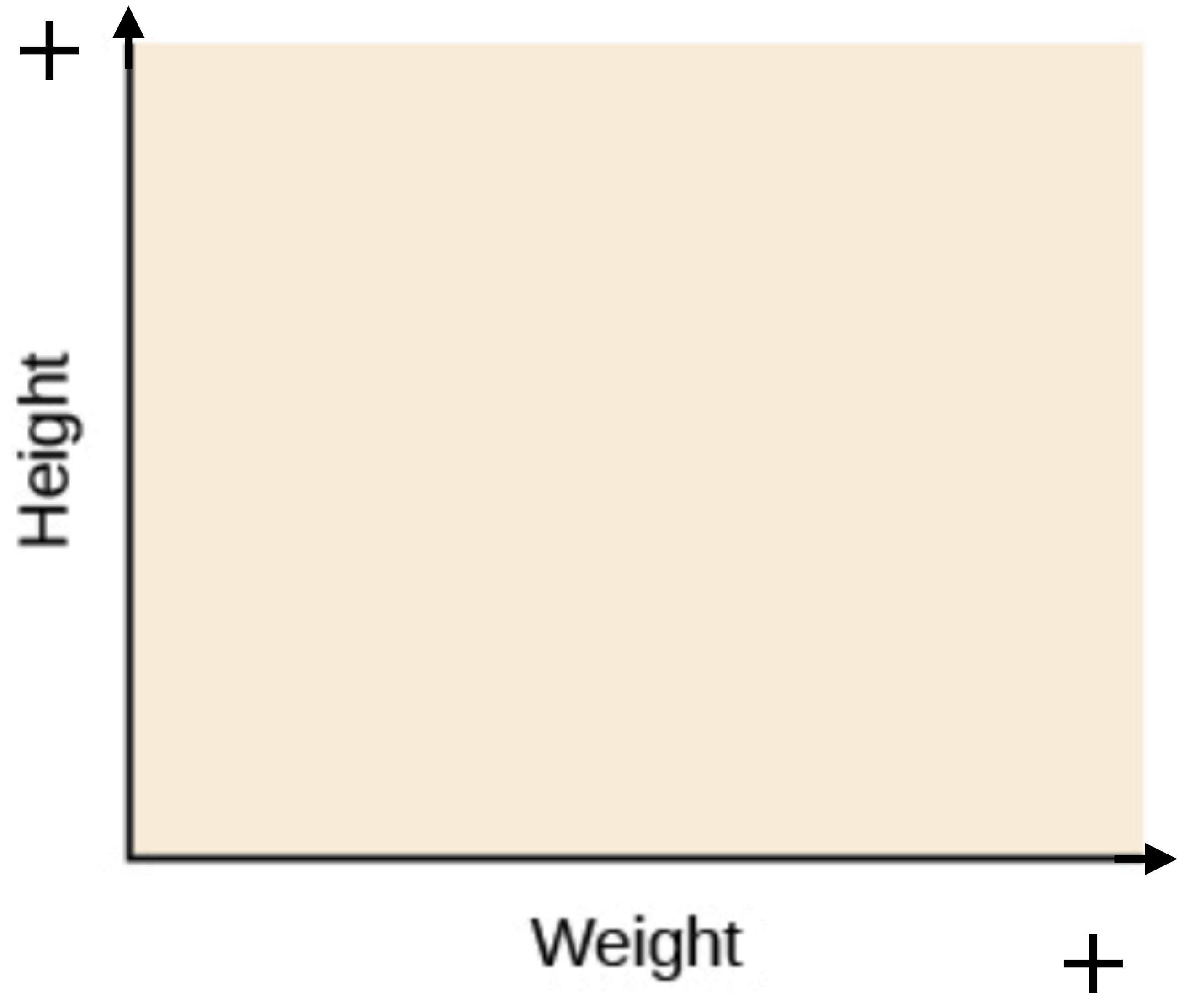
Positive? Negative? None?



Negative
correlation

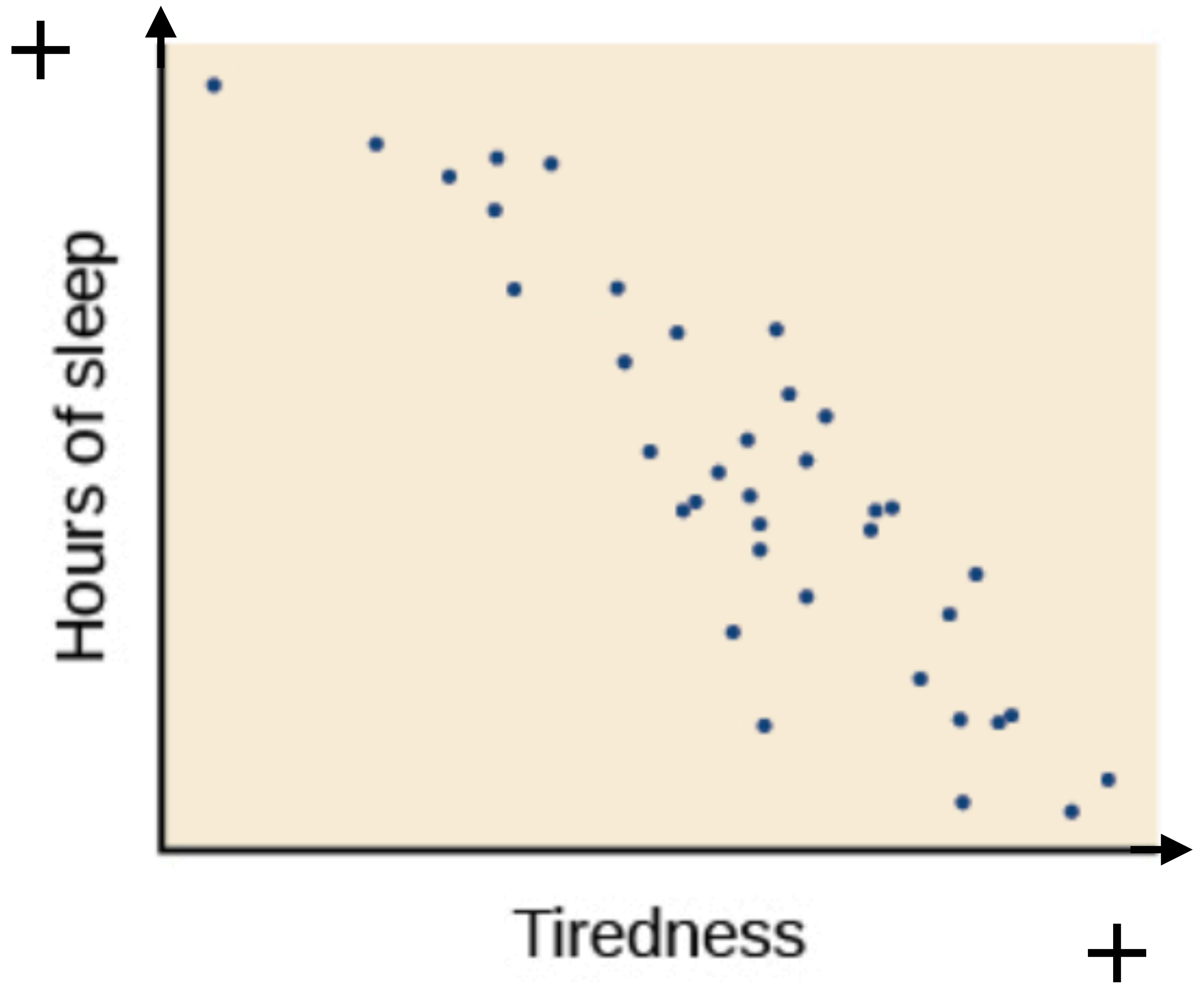


No
correlation

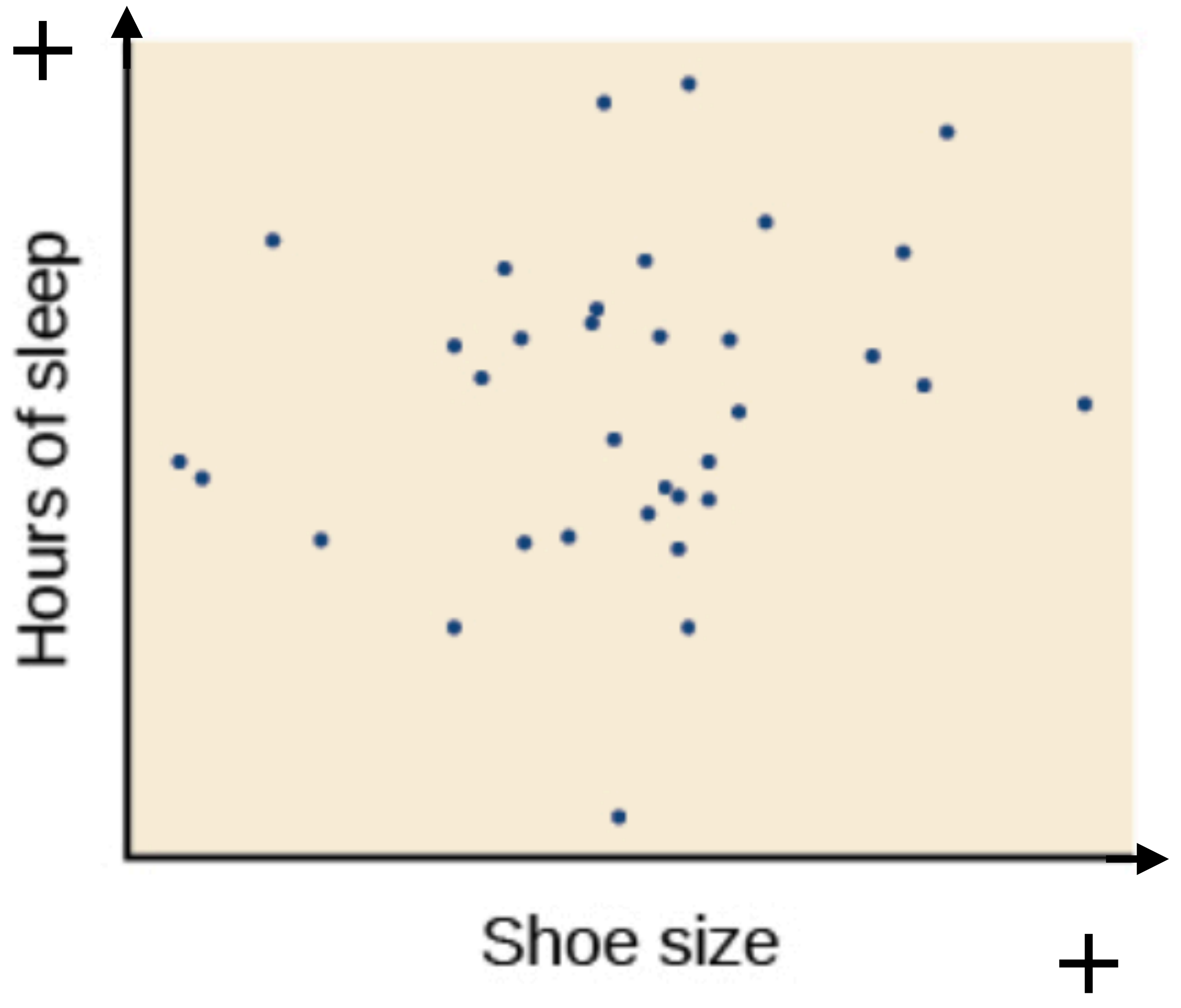


Examples of correlation

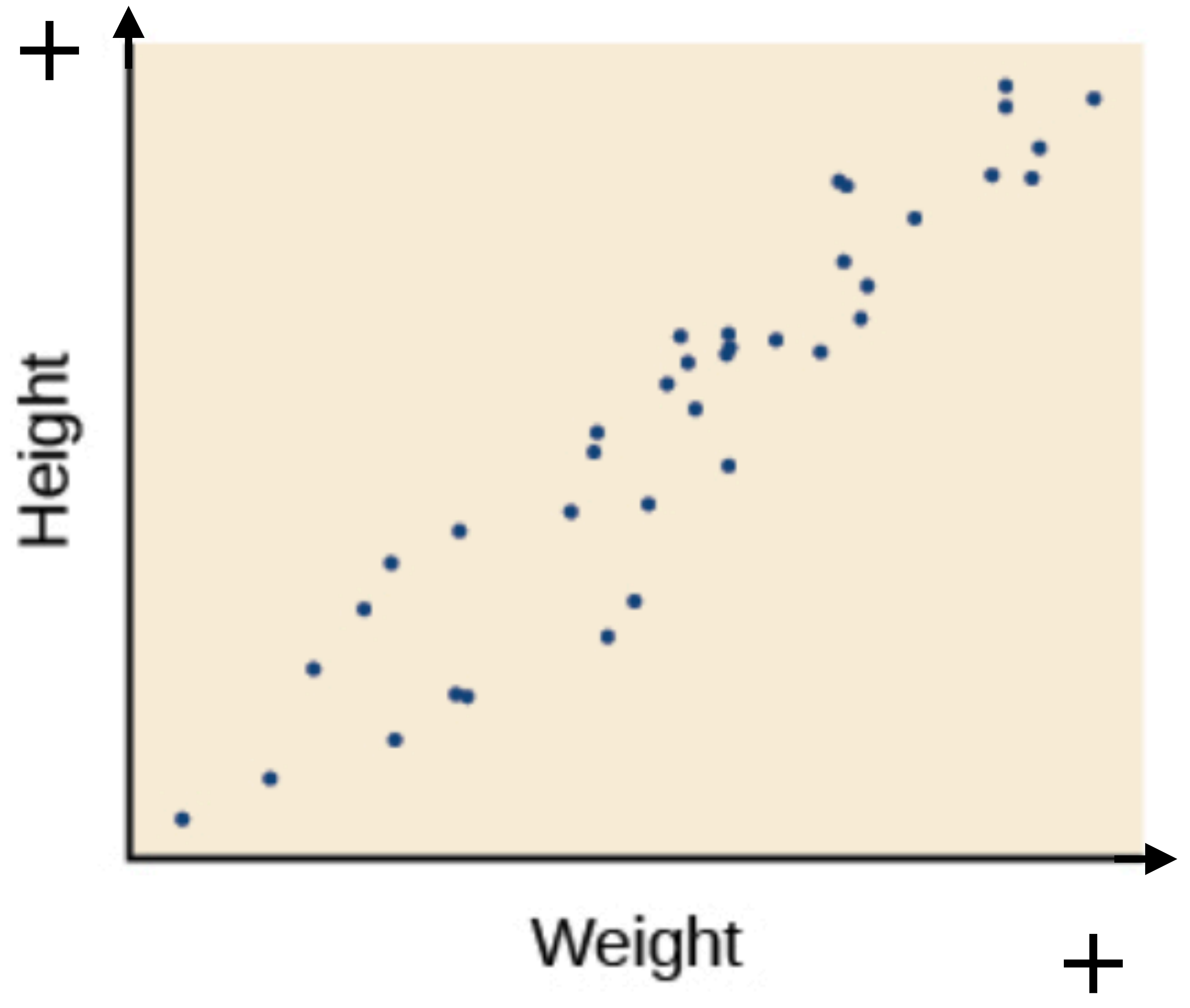
Positive? Negative? None?



Negative
correlation



No
correlation



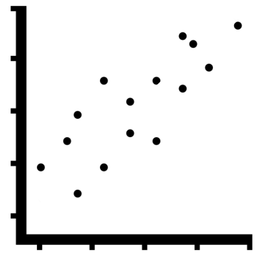
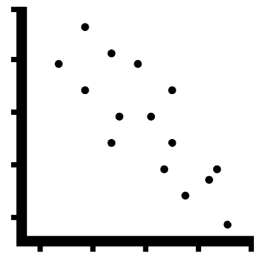
Positive
correlation

Measuring correlation

Strength and direction

Measuring correlation

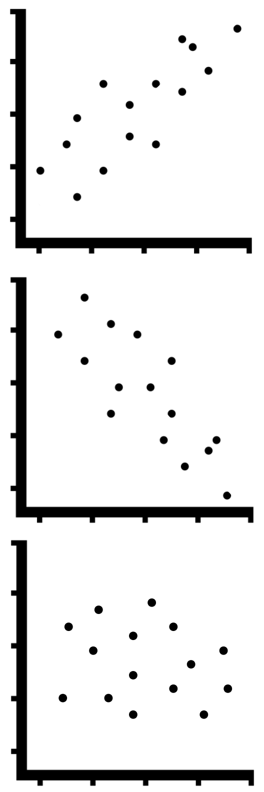
Strength and direction

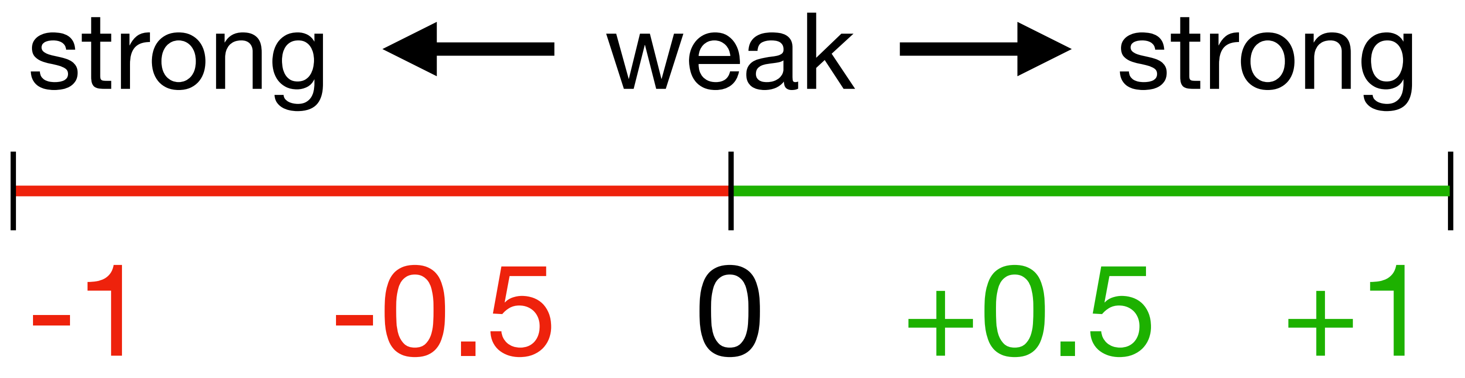
- The strength and direction of a correlation can be measured by a **correlation coefficient**, which ranges from -1 to 1
  - $\rho > 0$ positive correlation 
  - $\rho < 0$ negative correlation 
  - $\rho \approx 0$ no correlation 

Measuring correlation

Strength and direction

- The strength and direction of a correlation can be measured by a **correlation coefficient**, which ranges from -1 to 1
  - $\rho > 0$ positive correlation
  - $\rho < 0$ negative correlation
  - $\rho \approx 0$ no correlation



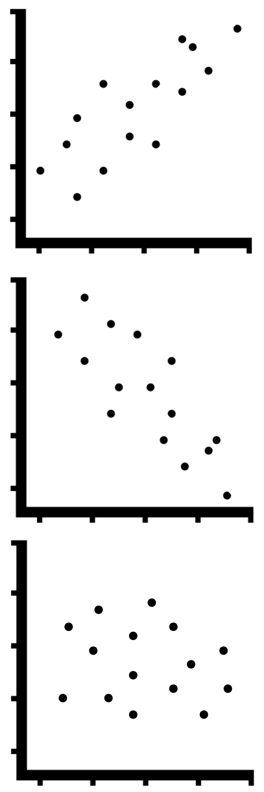


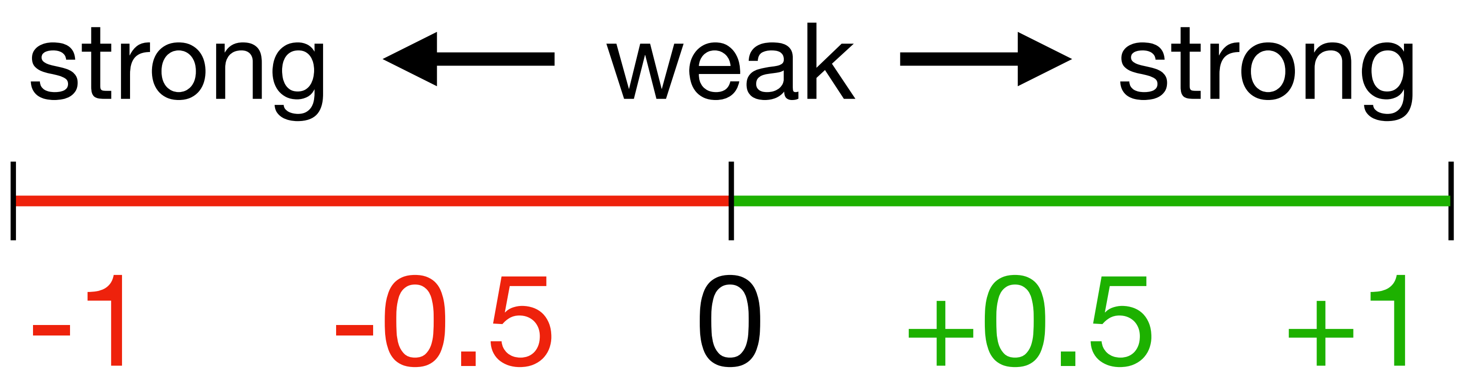
- A correlation coefficient close to the extremes (-1 or 1) indicates a **strong** correlation, while a coefficient close to 0 indicates a **weak** correlation.

Measuring correlation

Strength and direction

- The strength and direction of a correlation can be measured by a **correlation coefficient**, which ranges from -1 to 1
  - $\rho > 0$ positive correlation
  - $\rho < 0$ negative correlation
  - $\rho \approx 0$ no correlation





- A correlation coefficient close to the extremes (-1 or 1) indicates a **strong** correlation, while a coefficient close to 0 indicates a **weak** correlation.

- The correlation coefficient is usually represented by the letter r or greek letter ρ (rho).

Pearson's correlation coefficient

$\rho(X, Y)$

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.
- It is based on the method of **covariance**.

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.
- It is based on the method of **covariance**.

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.

- It is based on the method of **covariance**.

Pearson's correlation between X and Y

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.

- It is based on the method of **covariance**.

Pearson's correlation between X and Y

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

The product of the standard deviations of X and Y

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.
- It is based on the method of **covariance**.

Expected values of the product

Pearson's correlation between X and Y

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

The product of the standard deviations of X and Y

Pearson's correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.

- It is based on the method of **covariance**.

Expected values of the product

The product of the expected values

Pearson's correlation between X and Y

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

The product of the standard deviations of X and Y

Pearson’s correlation coefficient

$\rho(X, Y)$

- Pearson correlation coefficient is a measure of the strength and direction of a **linear association** between two variables X and Y.
- It is based on the method of **covariance**.

Expected values of the product

The product of the expected values

Pearson’s correlation between X and Y

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

The product of the standard deviations of X and Y

Some univariate statistics notation

to understand the Pearson's correlation coefficient

Some univariate statistics notation

to understand the Pearson's correlation coefficient

- X is a random variable
  - In data: x_i is the i^{th} element in X (or the value of the variable for entry i)
  - For example, the GDP of a country

Some univariate statistics notation

to understand the Pearson's correlation coefficient

- X is a random variable
  - In data: x_i is the i^{th} element in X (or the value of the variable for entry i)
  - For example, the GDP of a country
- $E[X]$ is the expected value of X
  - We estimate the expected value as the mean of X:

$$E[X] = \mu_X = \bar{x} = \frac{1}{N} \sum_i x_i$$

  - N is the number of data points, for example the number of countries

Some univariate statistics notation

to understand the Pearson's correlation coefficient

- X is a random variable **lower-case x**
    - In data: x_i is the i^{th} element in X (or the value of the variable for entry i)
    - For example, the GDP of a country **upper-case (capital) X**
- $E[X]$ is the expected value of X
    - We estimate the expected value as the mean of X:

$$E[X] = \mu_X = \bar{x} = \frac{1}{N}\sum_i x_i$$

- N is the number of data points, for example the number of countries

Some univariate statistics notation (ii)

to understand the Pearson's correlation coefficient

Some univariate statistics notation (ii)

to understand the Pearson's correlation coefficient

- $V[X]$ is the variance of X
  - We calculate it as the expected squared difference to the mean X:

$$V[X] = \frac{1}{N} \sum_i (x_i - \mu_X)^2$$

  - It is measured in squared units of X

Some univariate statistics notation (ii)

to understand the Pearson's correlation coefficient

- $V[X]$ is the variance of X
  - We calculate it as the expected squared difference to the mean X:

$$V[X] = \frac{1}{N}\sum_i (x_i - \mu_X)^2$$

  - It is measured in squared units of X
- σ_X is the standard deviation of X
  - $\sigma_X = \sqrt{V[X]}$,which is convenient because it measures dispersion in the same units as X
  - You can calculate it with the functions $std()$ from **`numpy`** or $stdev()$ from **`statistics`**

Some univariate statistics notation (iii)

to understand the Pearson's correlation coefficient

Some univariate statistics notation (iii)

to understand the Pearson's correlation coefficient

- If X and Y are **independent**, then they satisfy that the expectation of the product equals the product of expectations:

$$E[XY] = E[X]E[Y]$$

Some univariate statistics notation (iii)

to understand the Pearson's correlation coefficient

- If X and Y are $\overbrace{\textbf{independent}}^{\text{No correlated}}$, then they satisfy that the expectation of the product equals the product of expectations:

$$E[XY] = E[X]E[Y]$$

Some univariate statistics notation (iii)

to understand the Pearson's correlation coefficient

- If X and Y are **independent**, then they satisfy that the expectation of the product equals the product of expectations:

$$E[XY] = E[X]E[Y]$$

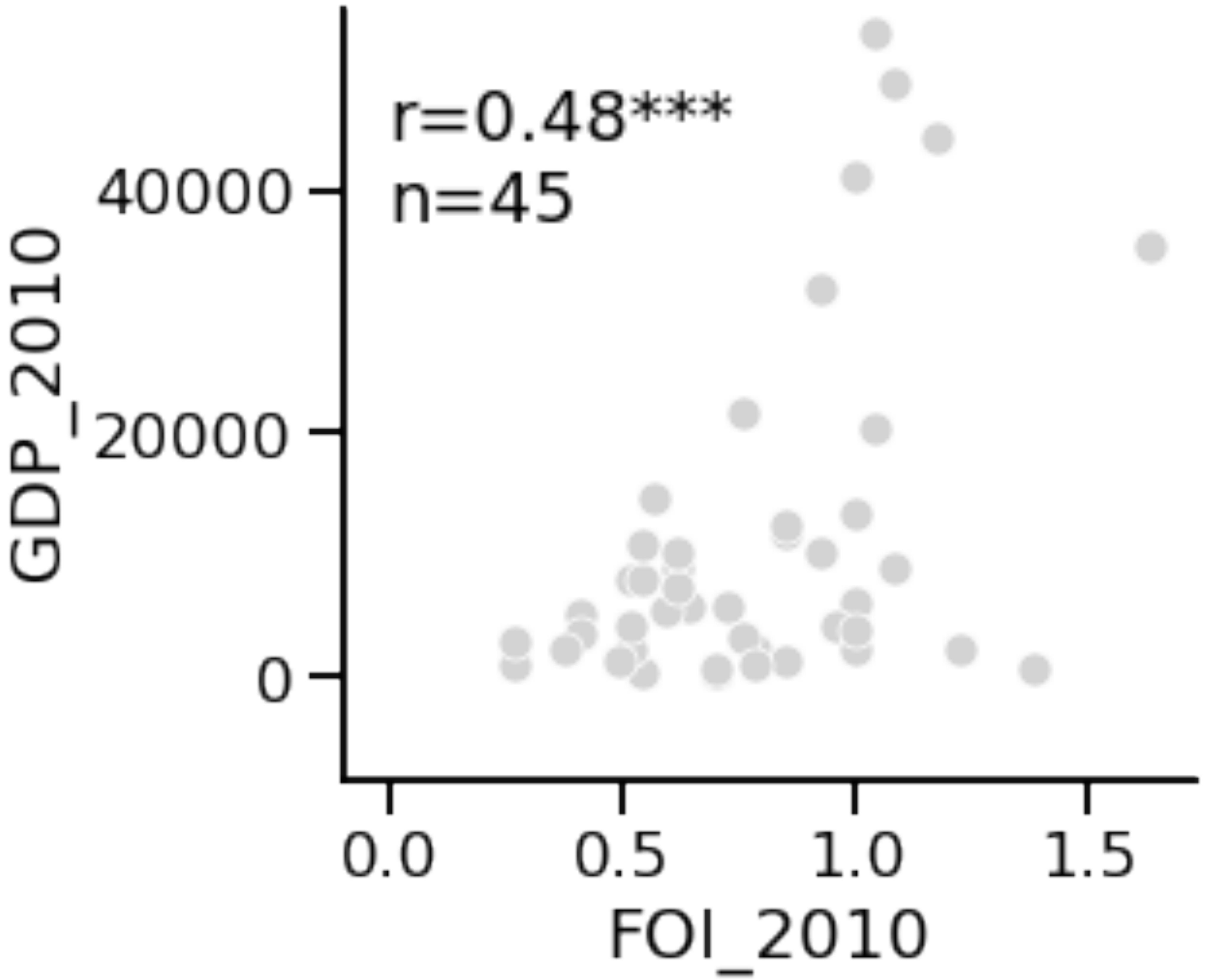
- The principle: correlation as the deviation from $E[XY] - E[X]E[Y] = 0$ *(no correlation)*
- The absolute value of this difference can be at most $\sigma_X \sigma_Y$ *(normalizing factor)*
- Thus, $\sigma_X \sigma_Y$ rescales the difference to be between –1 and 1

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$$

Pearson's correlation between X and Y

Correlation between FOI and GDP

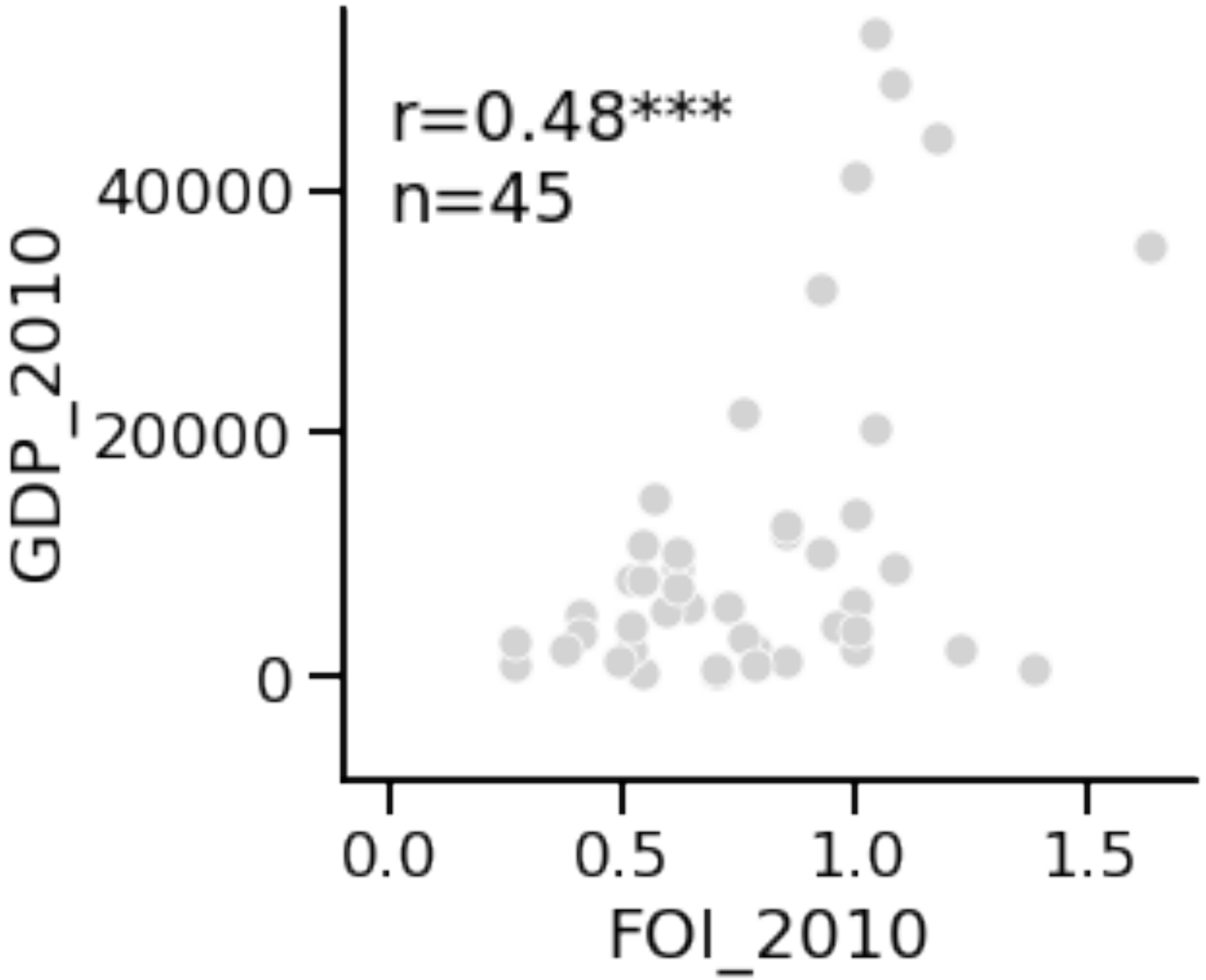
Pearson's correlation coefficient



Correlation between FOI and GDP

Pearson's correlation coefficient

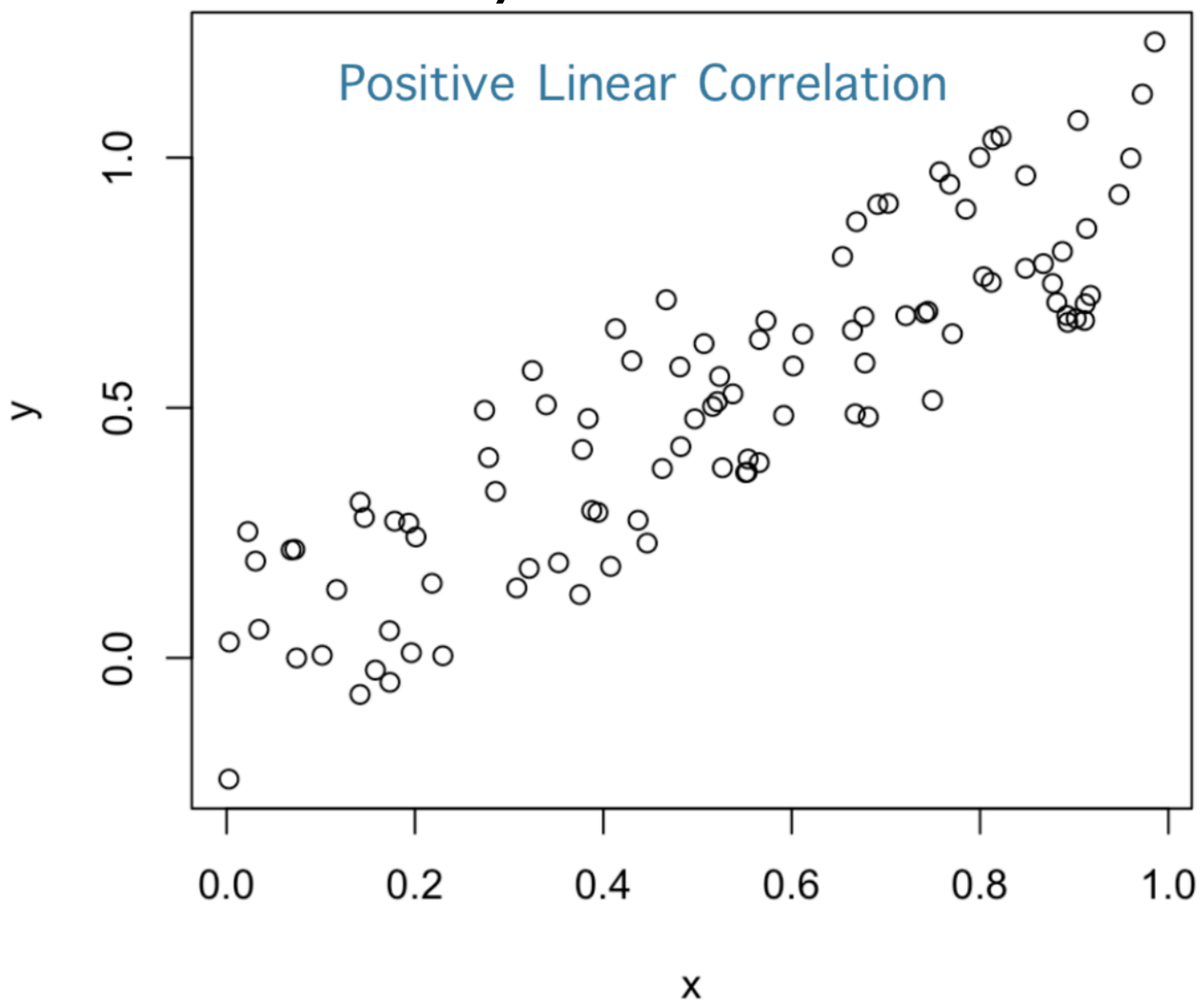
Positive and moderate correlation



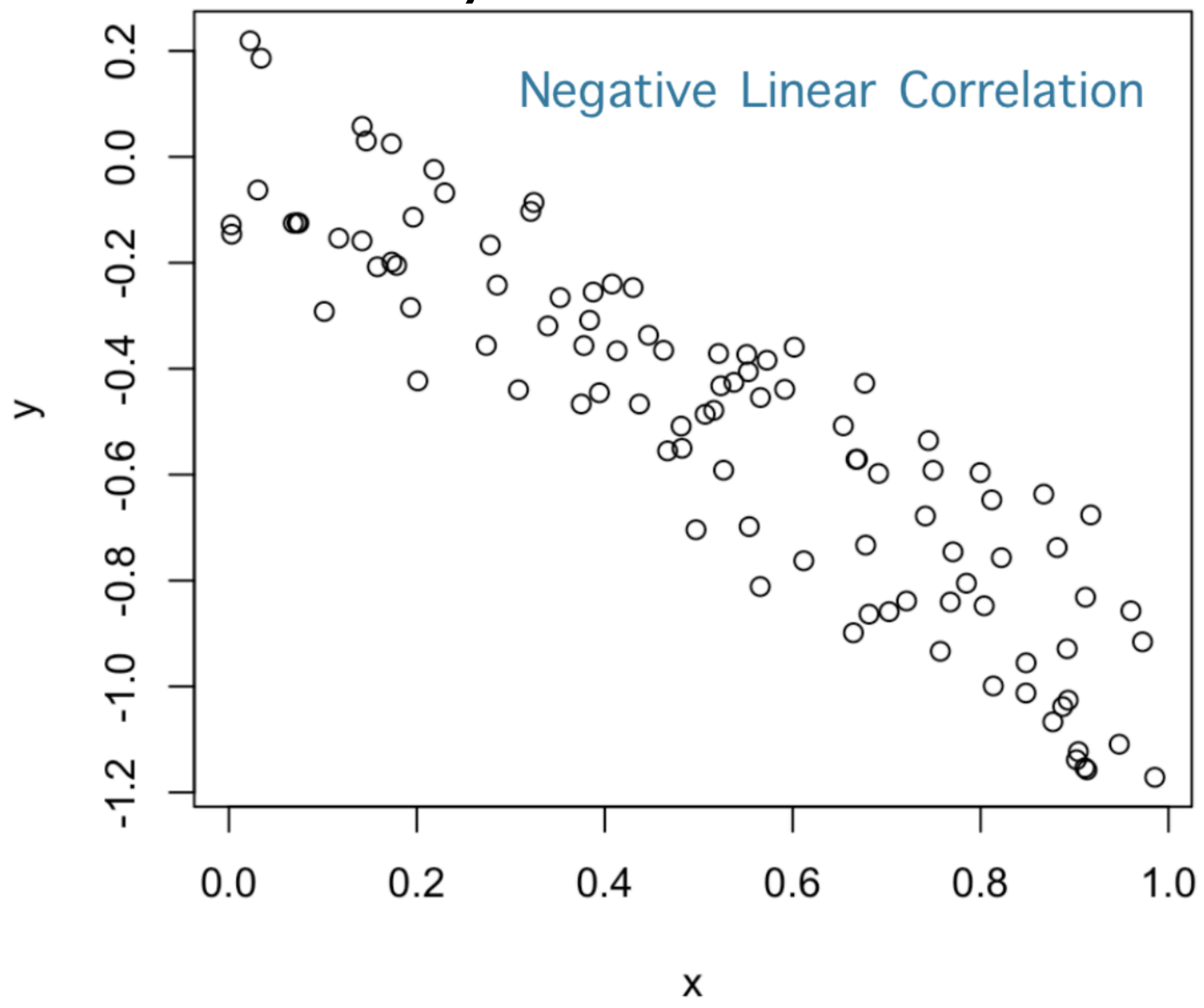
More examples

Pearson's correlation coefficient

$\rho = 0.8765$



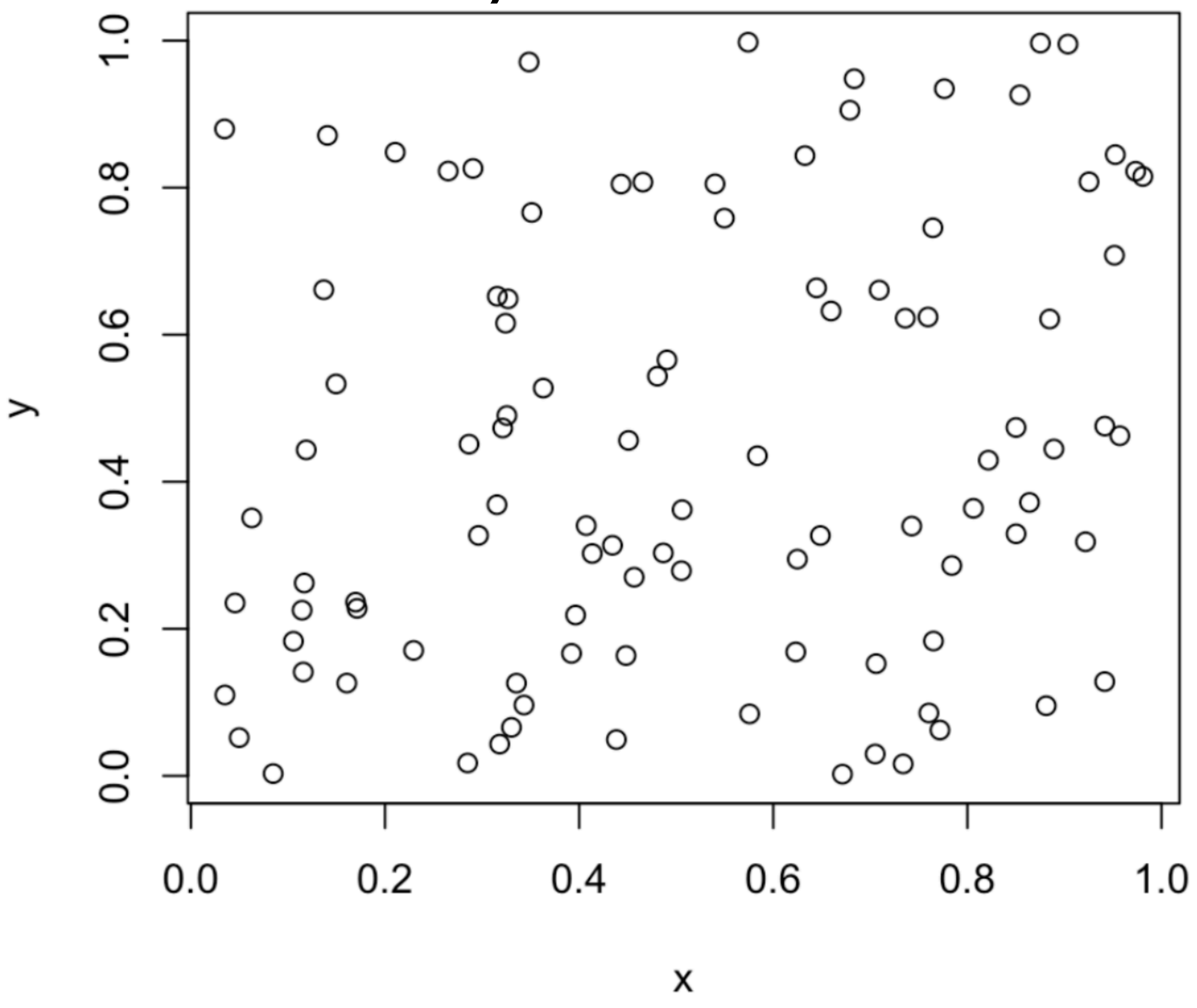
$\rho = -0.9046$



More examples

Pearson's correlation coefficient

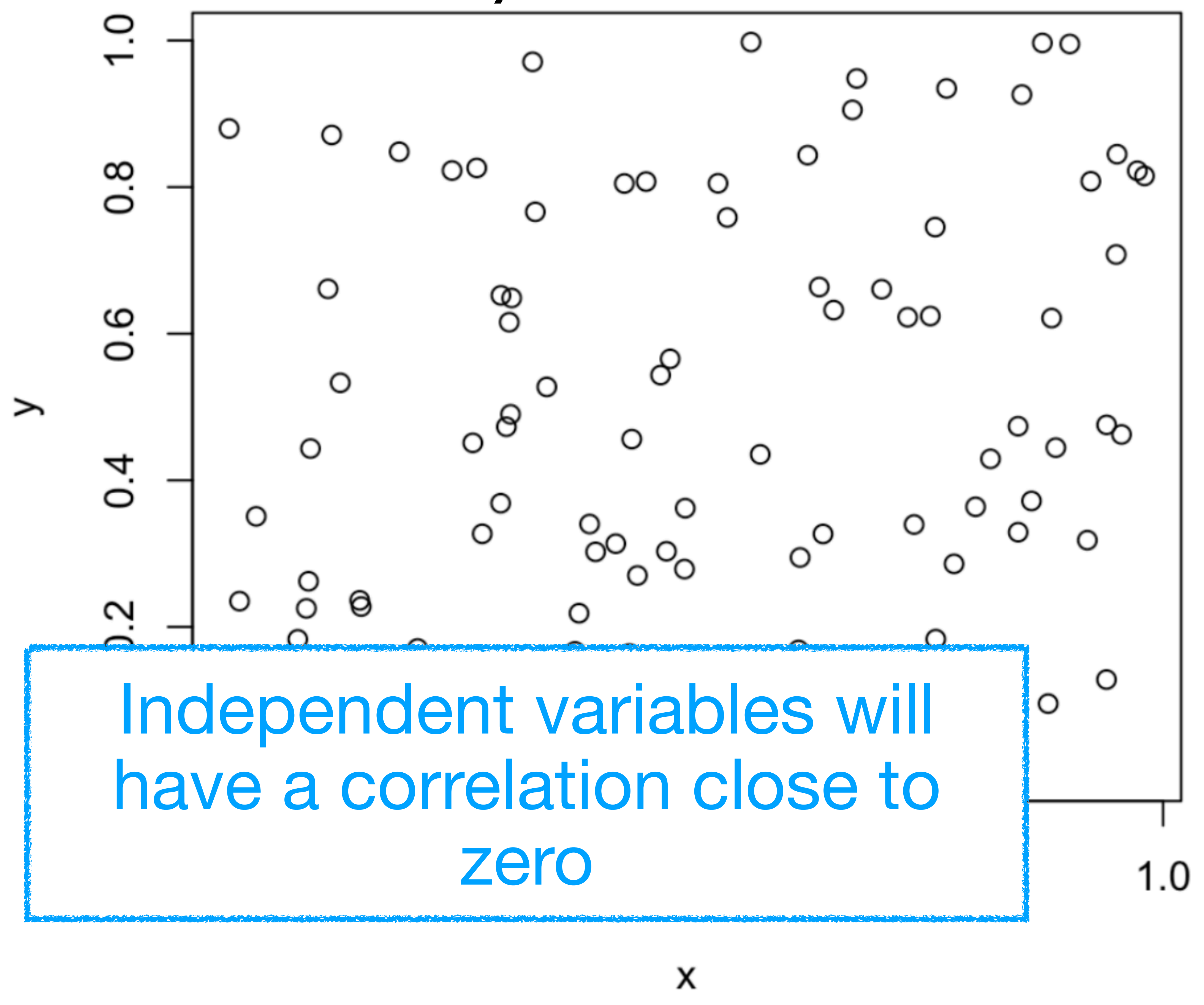
$$\rho = 0.2253$$



More examples

Pearson's correlation coefficient

$$\rho = 0.2253$$

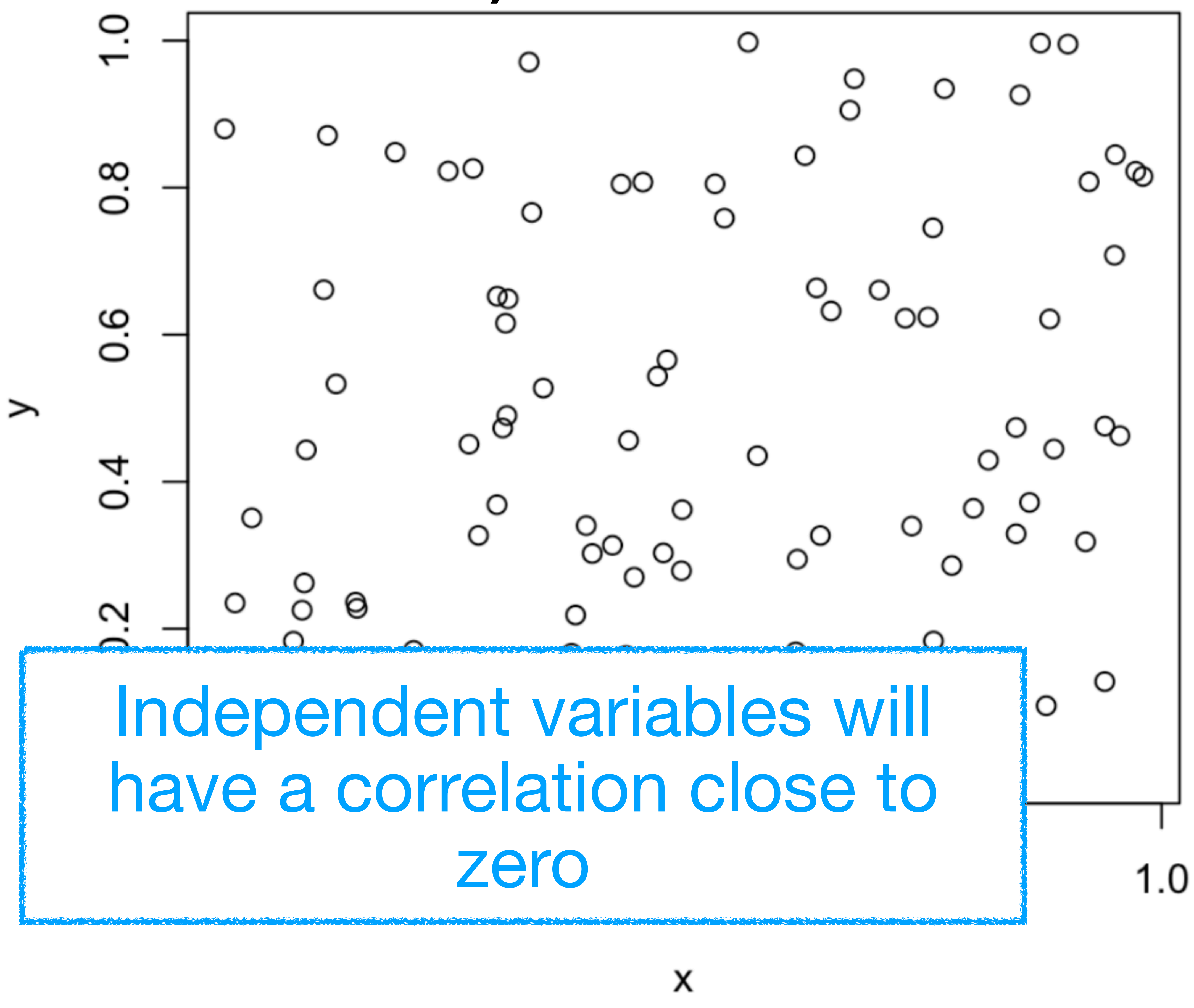


Independent variables will have a correlation close to zero

More examples

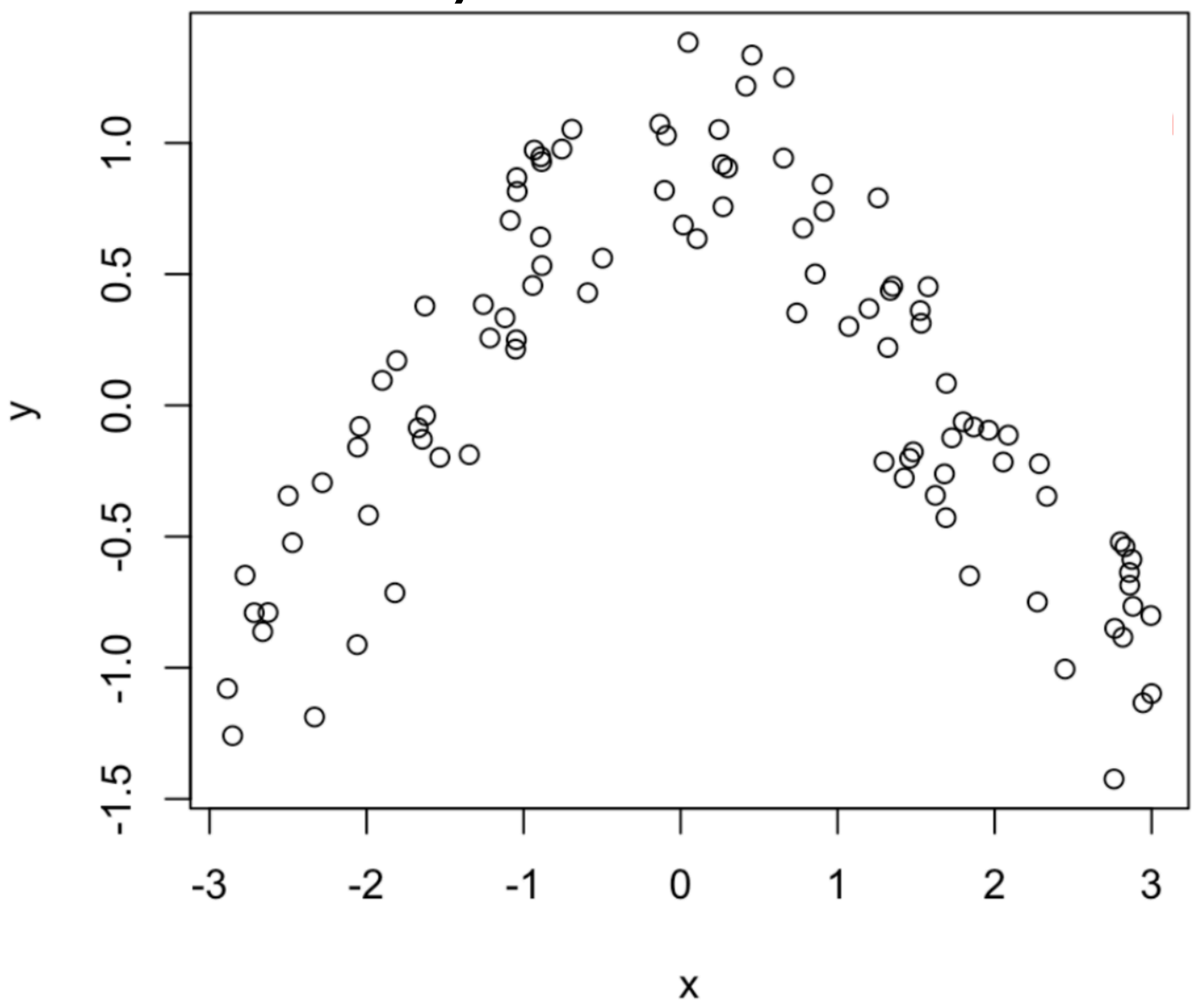
Pearson's correlation coefficient

$\rho = 0.2253$



Independent variables will have a correlation close to zero

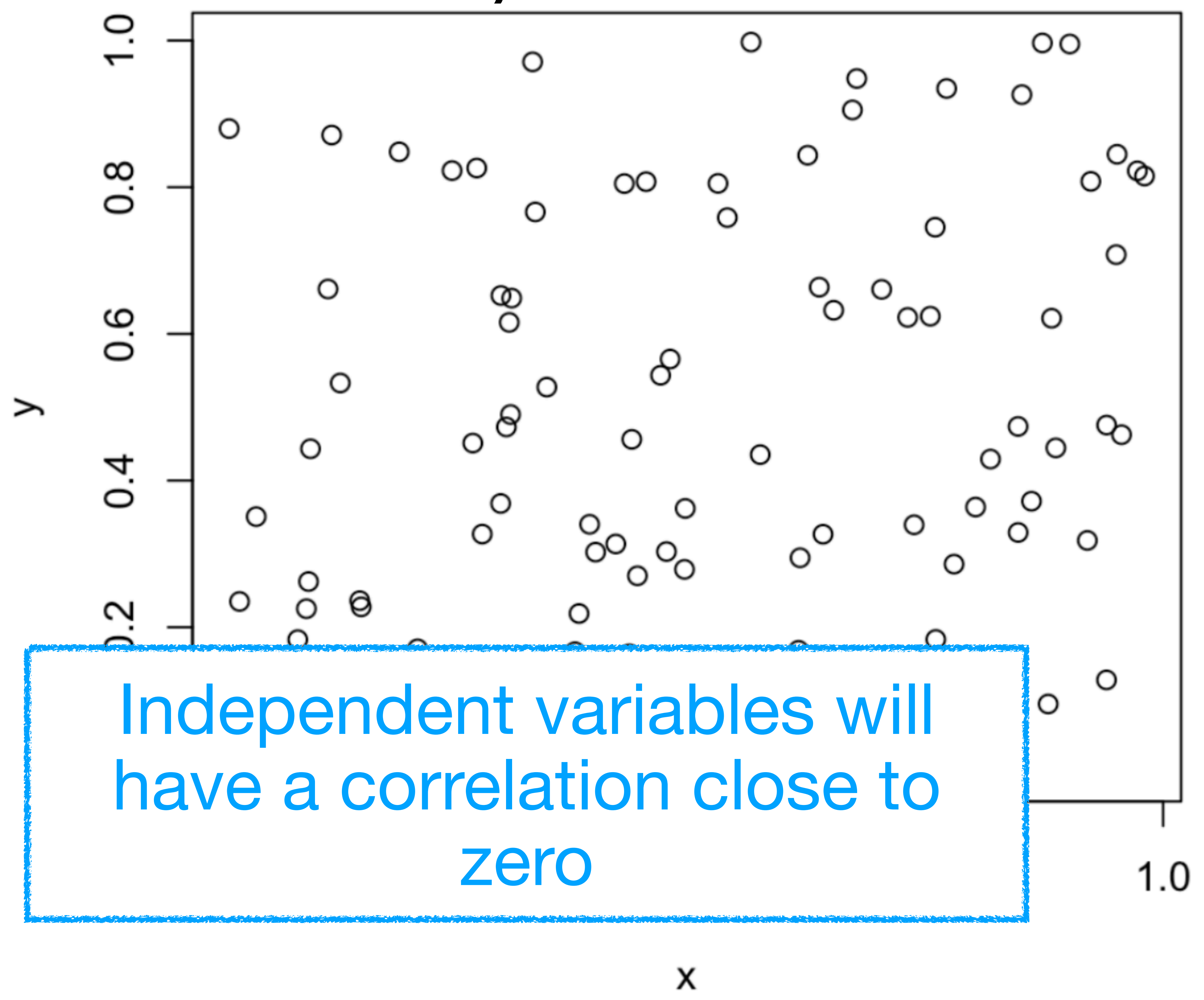
$\rho = -0.1245$



More examples

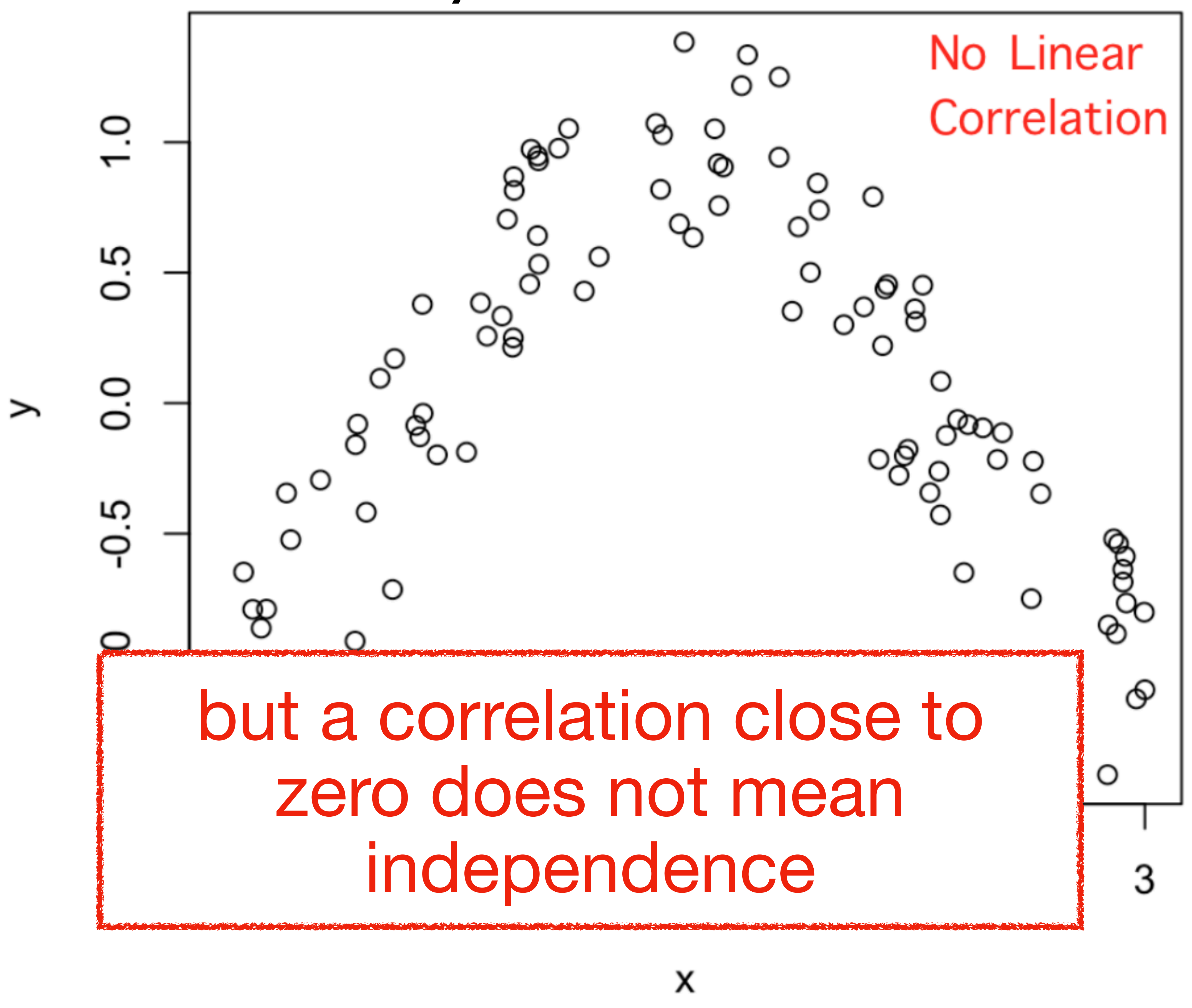
Pearson's correlation coefficient

$\rho = 0.2253$



Independent variables will have a correlation close to zero

$\rho = -0.1245$



but a correlation close to zero does not mean independence

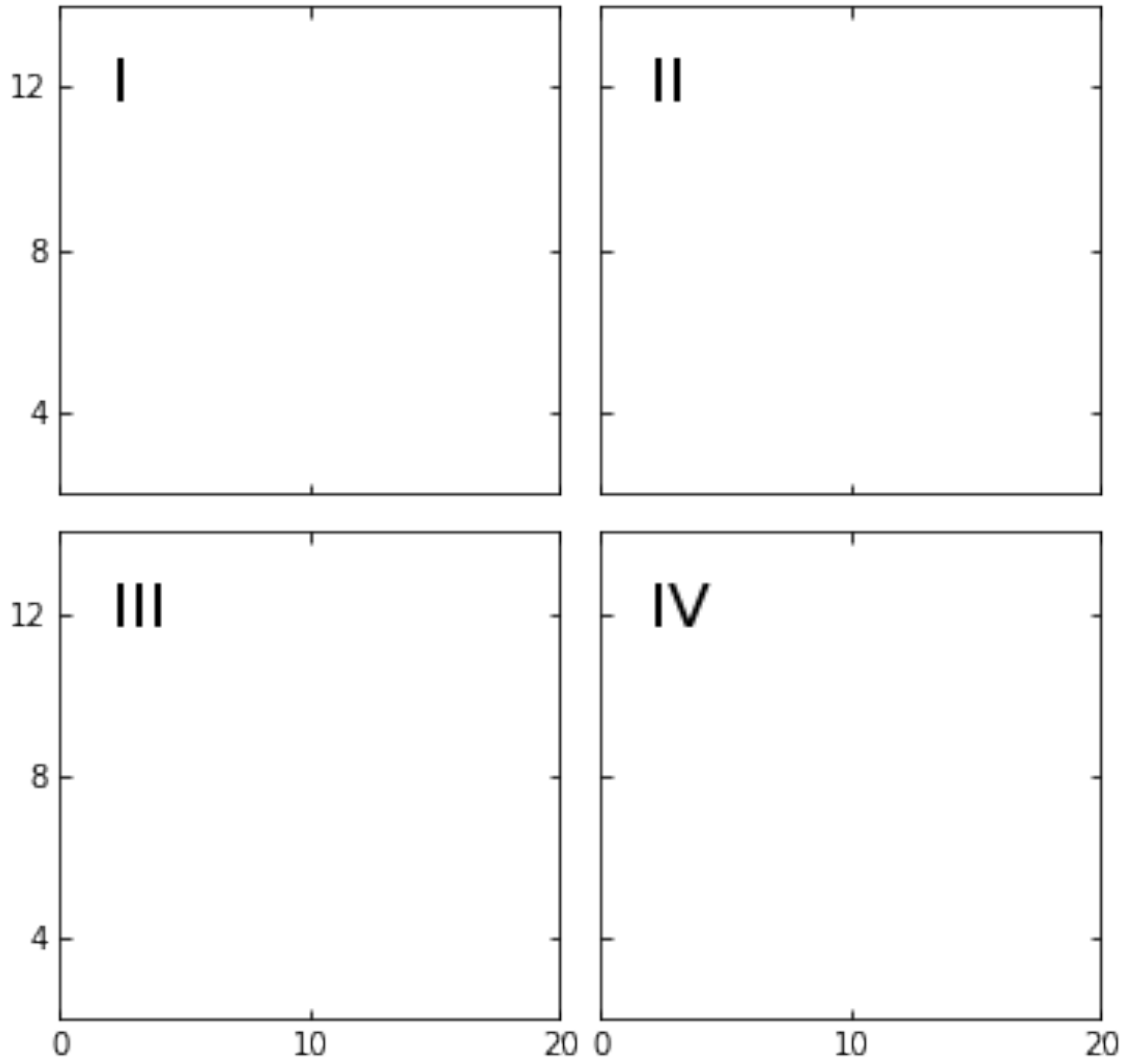
Anscombe's quartet

"numerical calculations are exact, but graphs are rough"

Anscombe's quartet

"numerical calculations are exact, but graphs are rough"

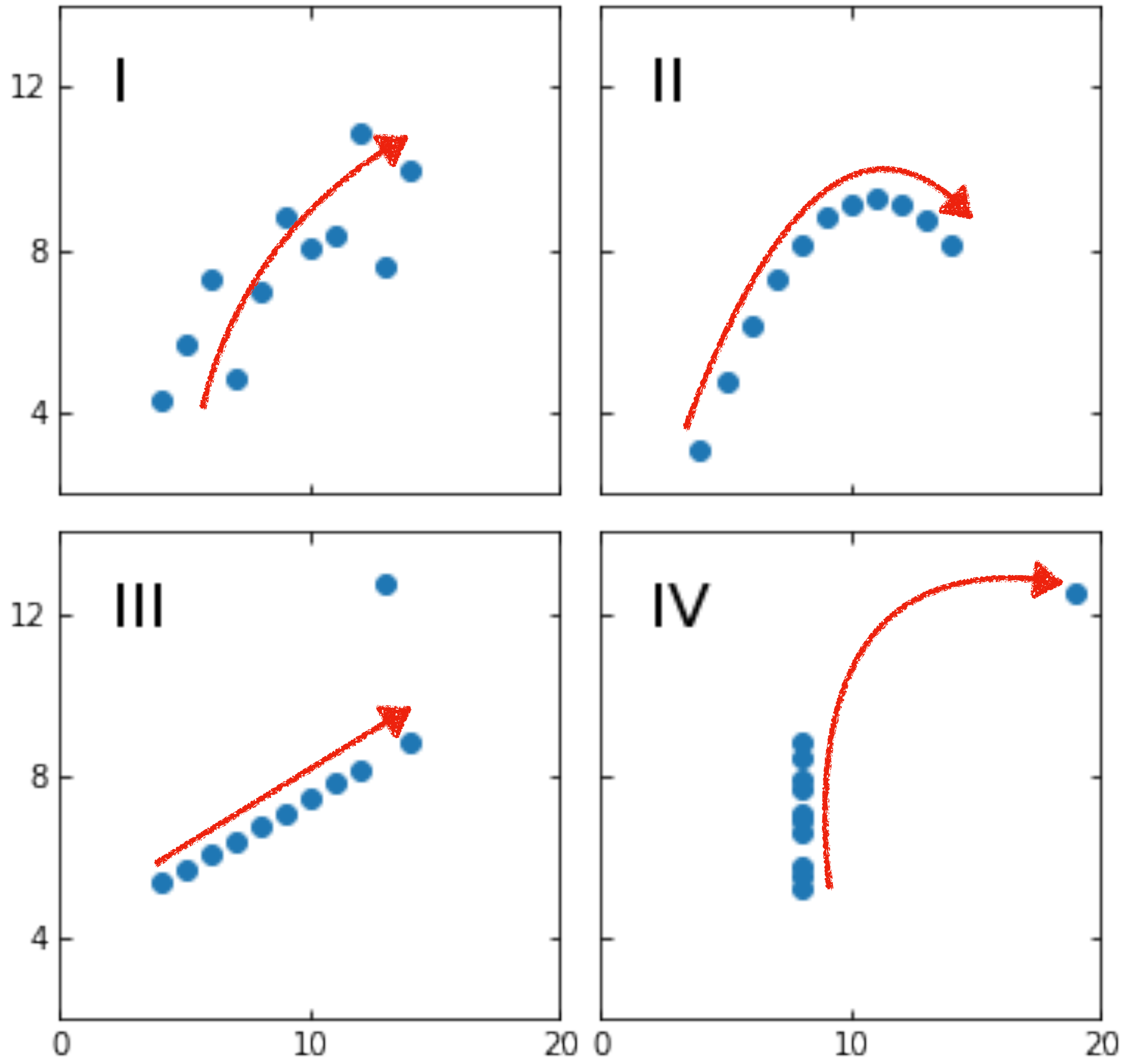
- These are 4 different datasets (X, Y)



Anscombe's quartet

"numerical calculations are exact, but graphs are rough"

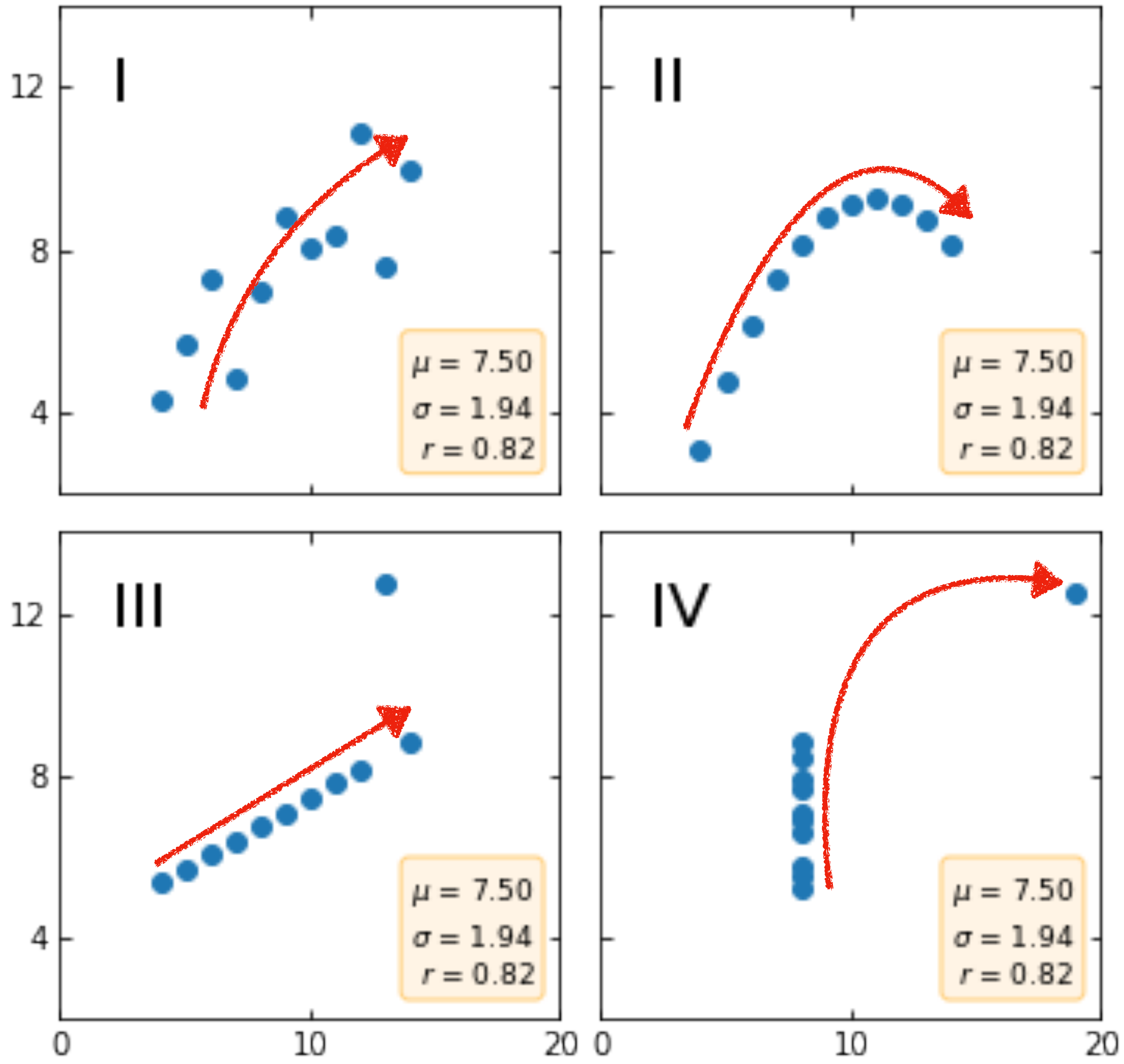
- These are 4 different datasets (X, Y)
- Qualitatively, they are very different



Anscombe's quartet

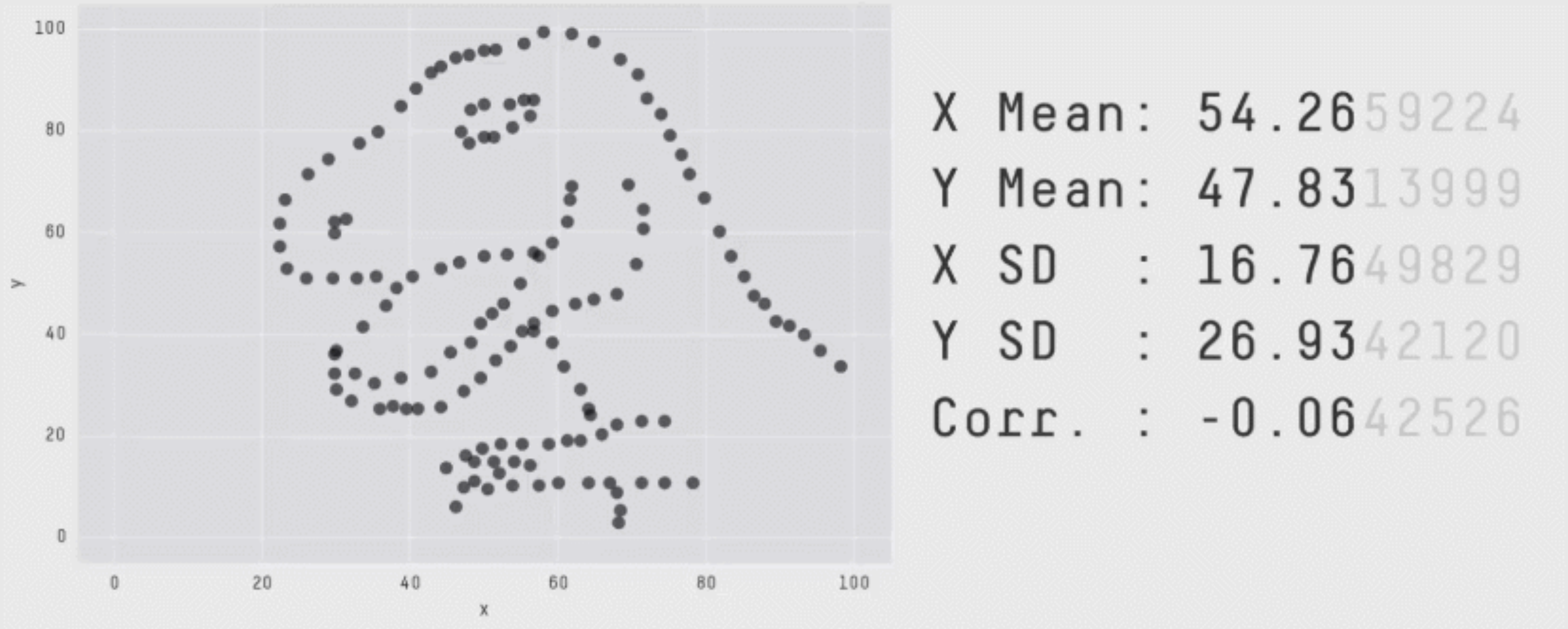
"numerical calculations are exact, but graphs are rough"

- These are 4 different datasets (X, Y)
- Qualitatively, they are very different
- Quantitative, they are the same: they have the same mean, standard deviation, and Pearson correlation



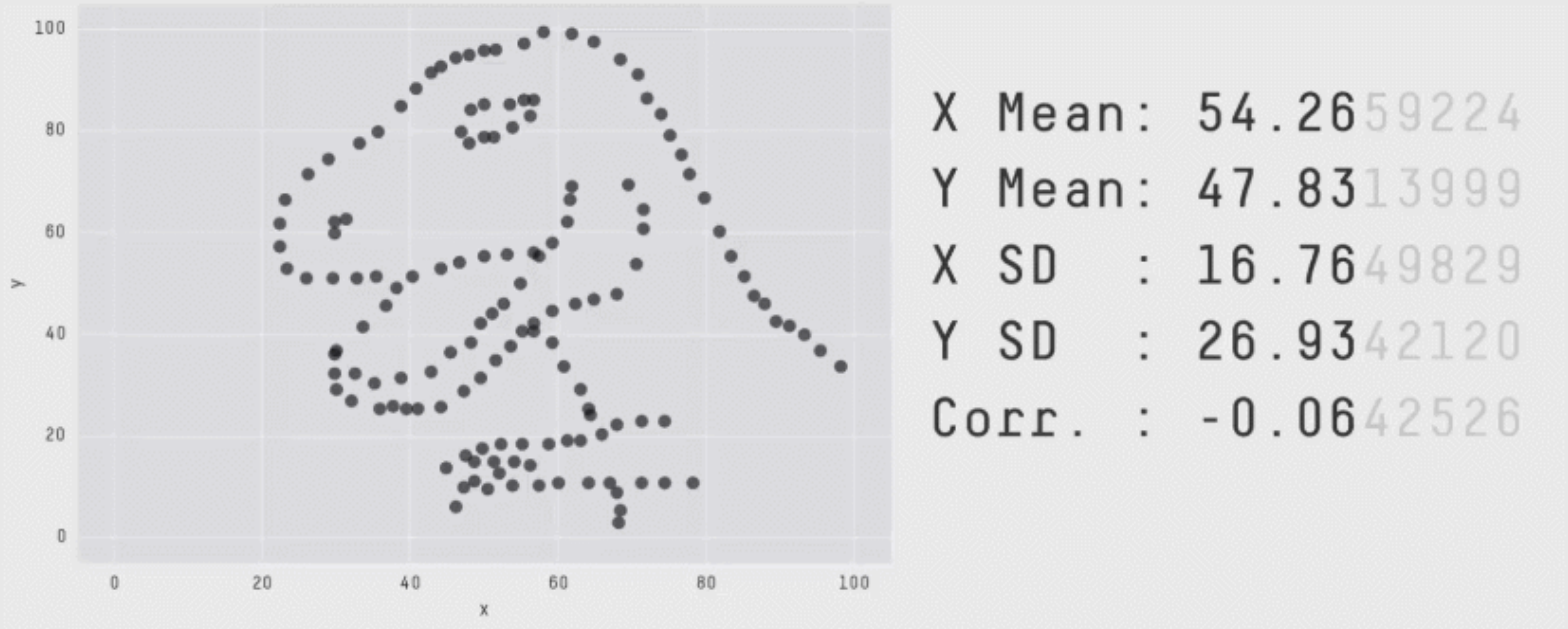
The Datasaurus dozen

Same Stats, different Graphs



The Datasaurus dozen

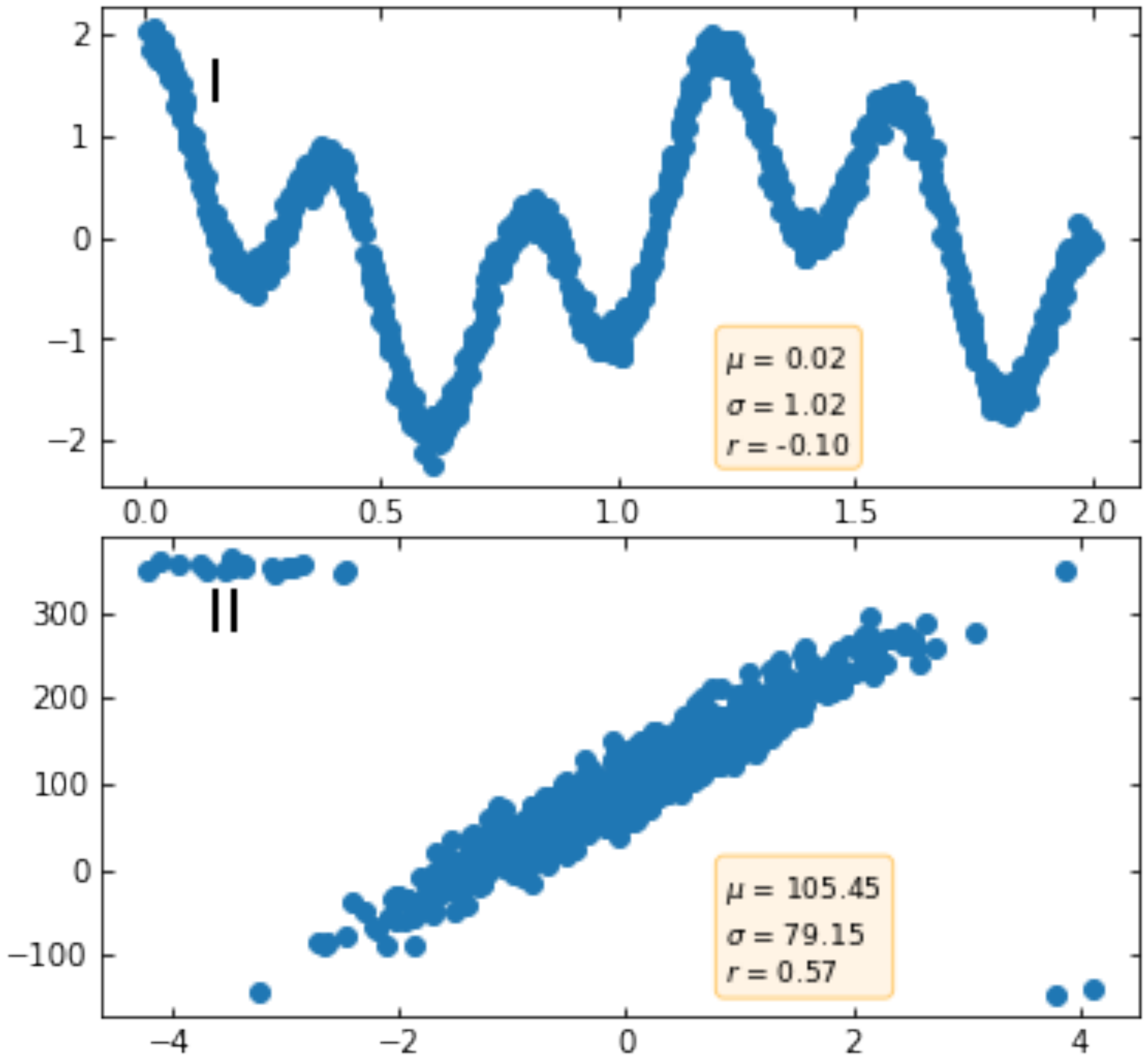
Same Stats, different Graphs



Limitations of Correlation

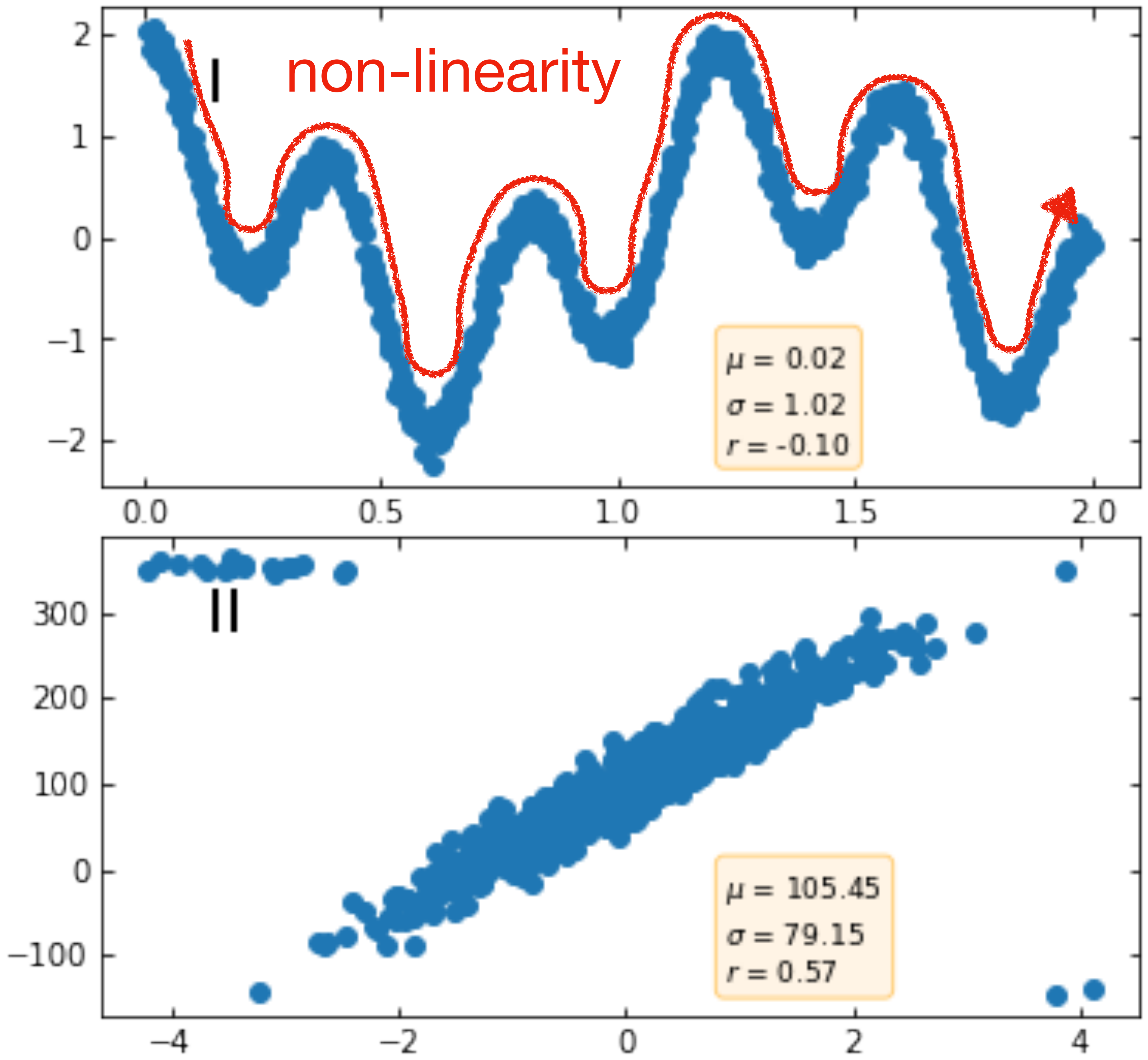
Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**



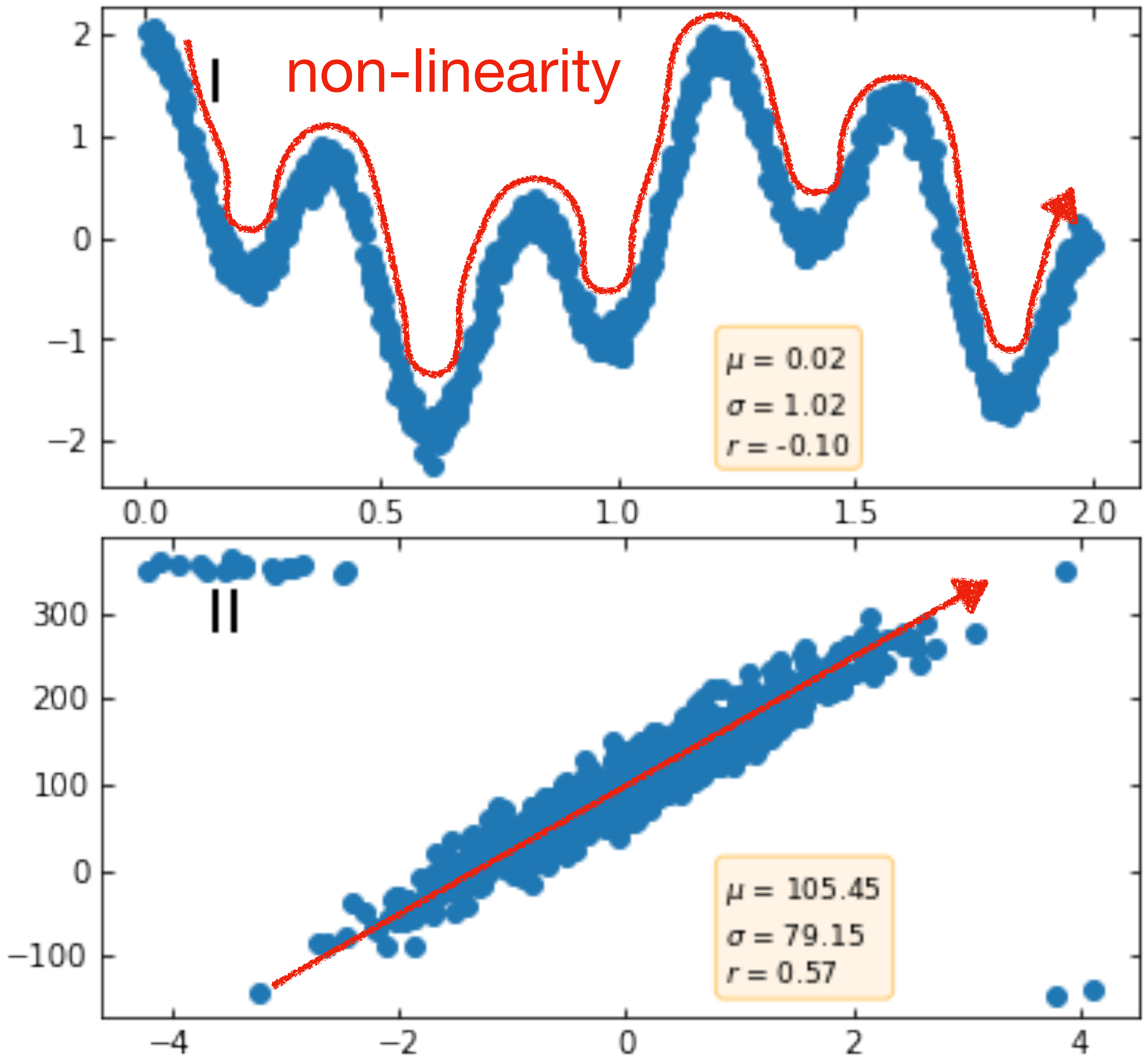
Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**



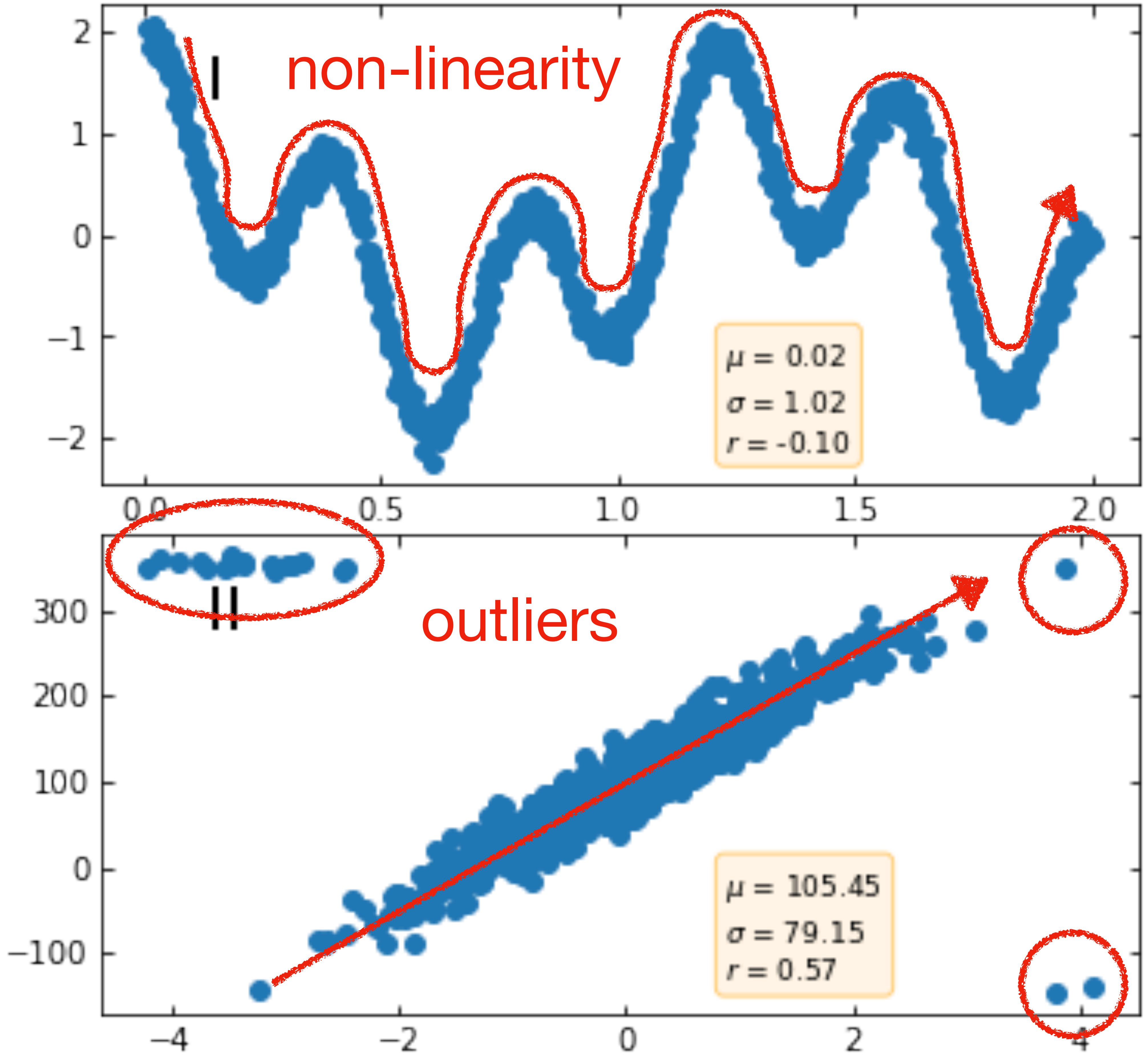
Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**



Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**



Limitations of Correlation

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

Did you know that as ice cream sales increase, so does violent crime. When ice cream sales decrease, so does violent crime.

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

Did you know that as ice cream sales increase, so does violent crime. When ice cream sales decrease, so does violent crime.

So it seems pretty clear that eating ice cream causes people to be violent, right?

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

Did you know that as ice cream sales increase, so does violent crime. When ice cream sales decrease, so does violent crime.

So it seems pretty clear that eating ice cream causes people to be violent, right?

Did you know that the divorce rate in Oregon is positively correlated with per capita consumption of whole milk?

$$r = 0.9$$

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

Did you know that as ice cream sales increase, so does violent crime. When ice cream sales decrease, so does violent crime.

So it seems pretty clear that eating ice cream causes people to be violent, right?

Did you know that the divorce rate in Oregon is positively correlated with per capita consumption of whole milk?

$$r = 0.9$$

If you want to divorce, then drink whole milk!

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

These are just spurious correlations

Did yo **that** as ice cream rease,
so c crime. When i ales
c does vi

So it seems p ng ice cream
causes peo ent, right?

Did you kn in Oregon
is po elated pita
tion of wh
$r = 0.9$

If you w nt to divorce, then drink w ole milk!

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

- Correlation does not provide information about the presence of **confounding** variables

These are just spurious correlations

Both ice cream sales and violent crime are associated with a **third variable**: seasonality (summer)

Did you know that the divorce rate in Oregon is positively correlated with per capita consumption of whole milk?

$$r = 0.9$$

Limitations of Correlation

- Correlation can be influenced by **non-linear** relationships or **outliers**

- Correlation does not provide information about the presence of **confounding** variables

- Correlation does not imply **causation**

These are just spurious correlations

Both ice cream sales and violent crime are associated with a **third variable**: seasonality (summer)

Whole milk and divorces are only related by **coincidence**.

Studying Causality

(brief overview)

What is causality

Definition

What is causality

Definition

- Causality is the relationship between **cause** and **effect**, where a change in one variable (the cause) leads to a change in another variable (the effect).

What is causality

Definition

- Causality is the relationship between **cause** and **effect**, where a change in one variable (the cause) leads to a change in another variable (the effect).
- Establishing causality requires three essential conditions:
  - **Temporal** sequencing: X must come before Y in time.
  - **Non-chance** relationship: The observed relationship between X and Y did not happen by chance alone.
  - **No alternative** explanation: There is nothing else that accounts for the $X \to Y$ relationship.

How is causality measured?

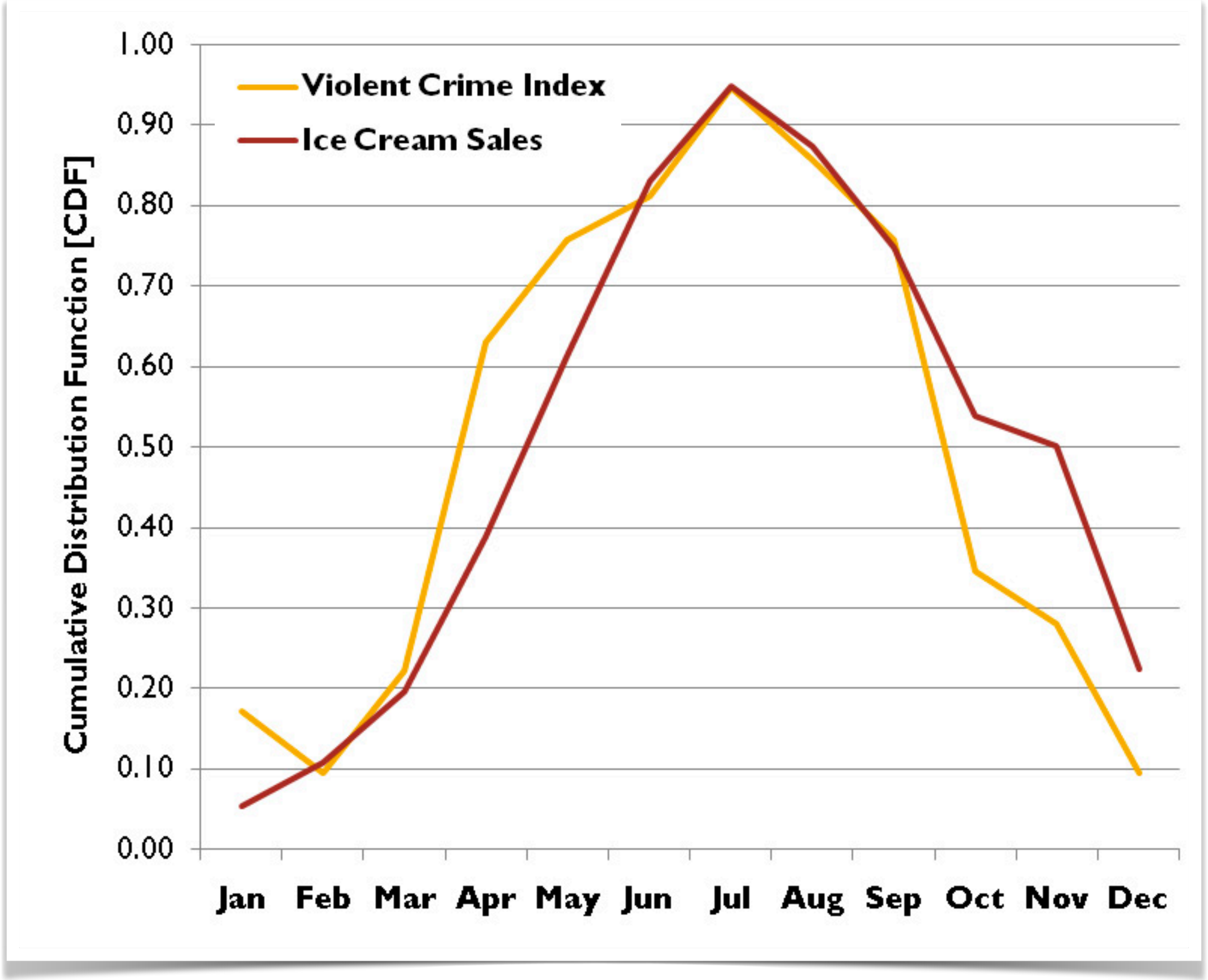
- **Experimentation:** A controlled experiment, where the independent variable X is manipulated (controlled), and the effect on the dependent variable Y is measured.
    - Randomized controlled trials (RCTs) are often used in clinical medical research to demonstrate causality.
        - Also known as “counterfactuals” in computer science (AI/ML & fairness)

Examples

of causality

Examples

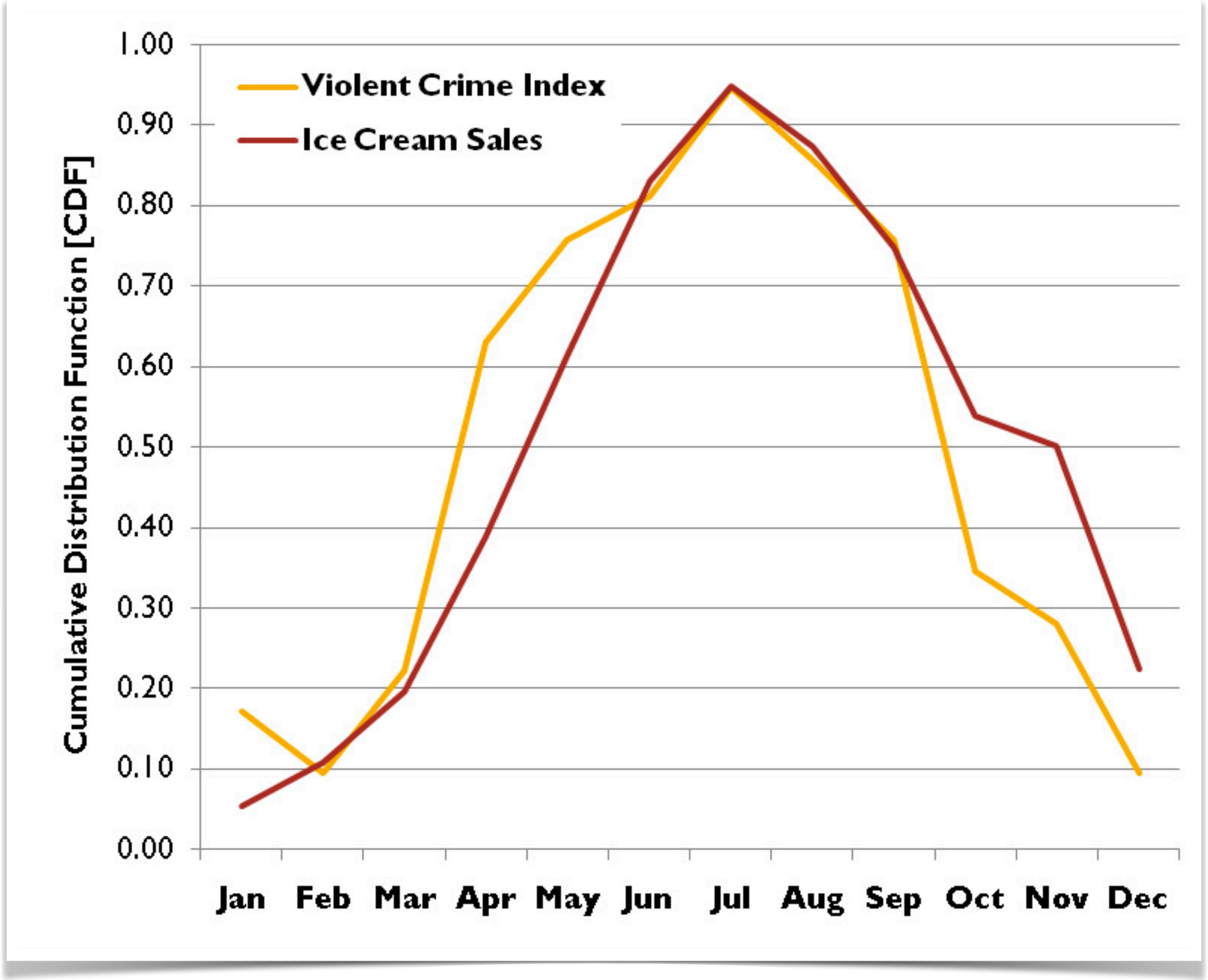
of causality



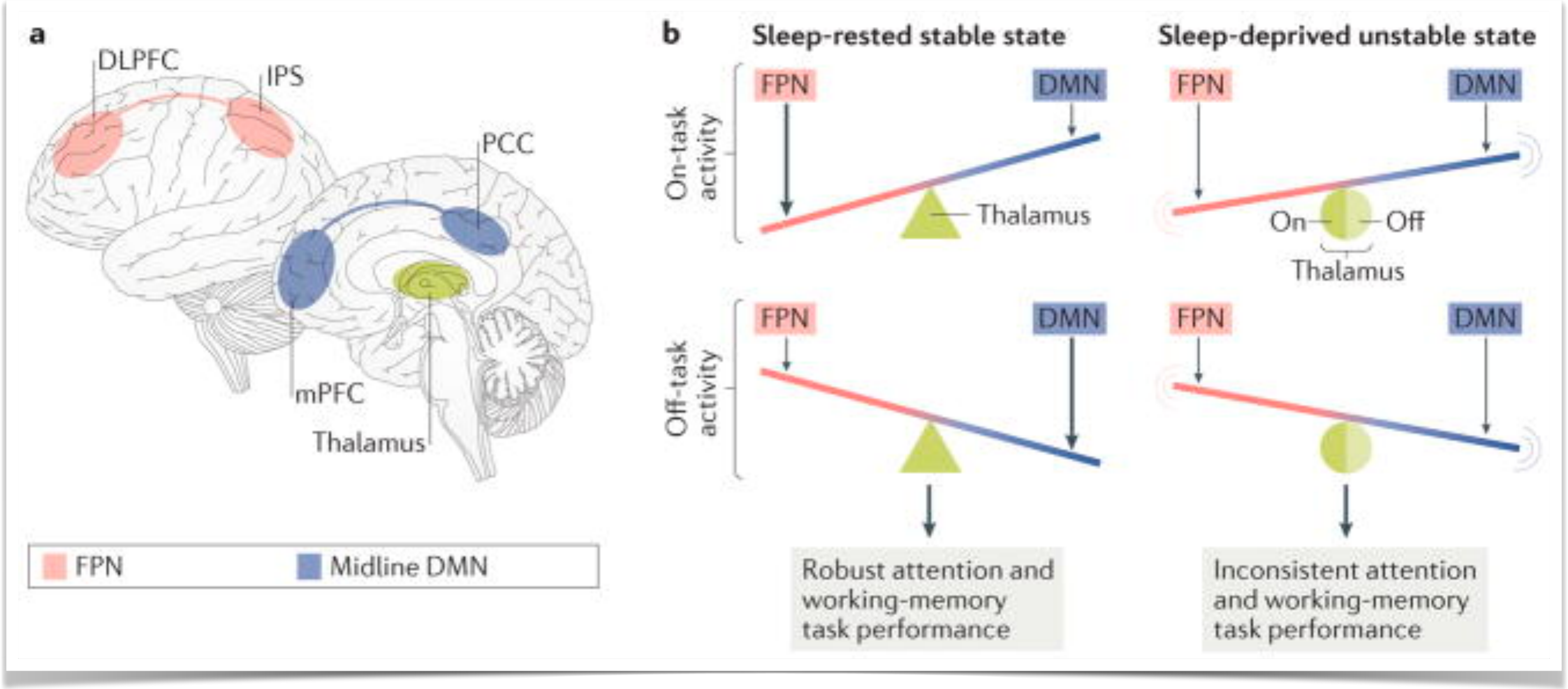
High temperature increases ice cream consumption and crime [Drescher 2014]

Examples

of causality

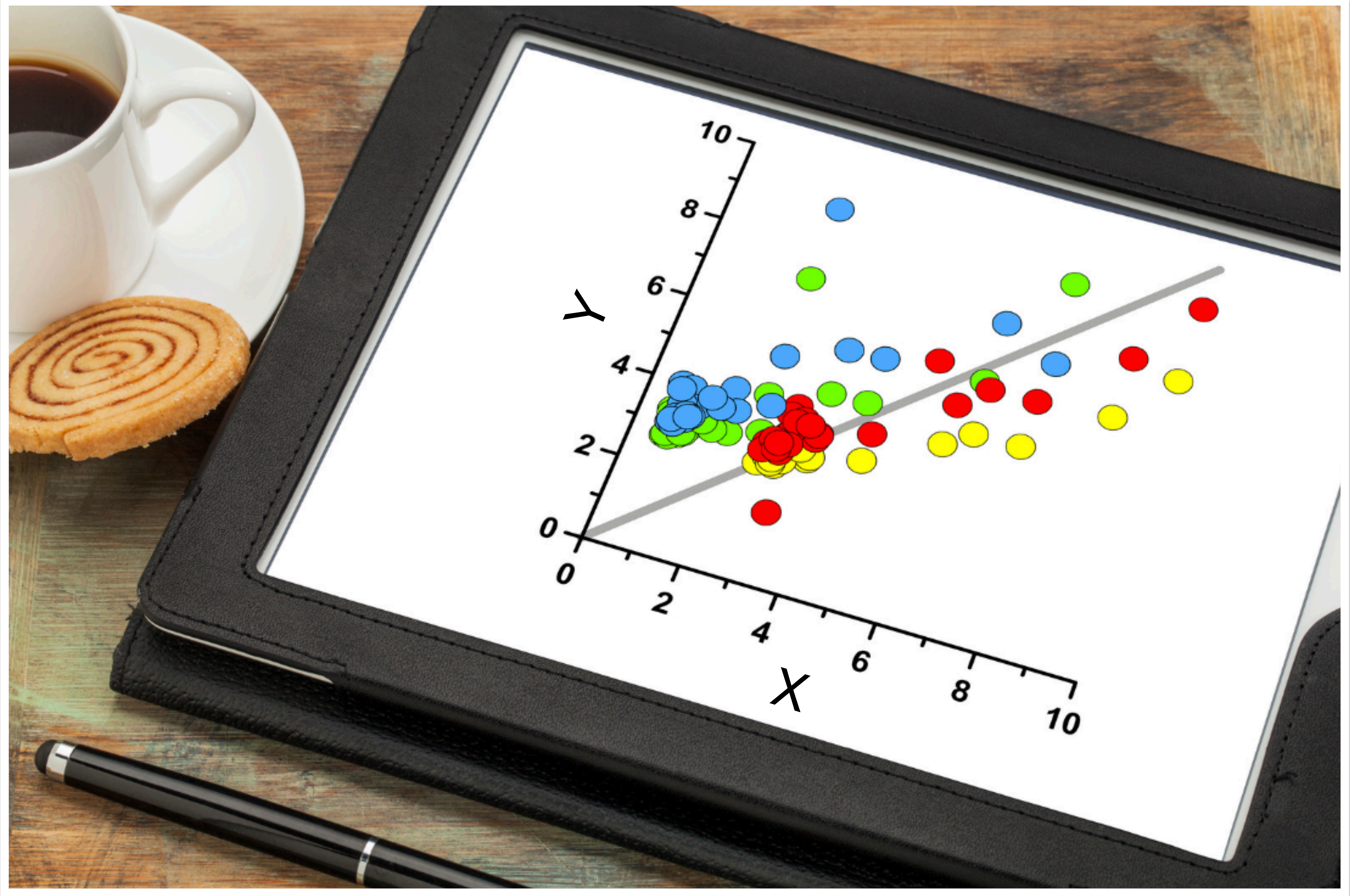


High temperature increases ice cream consumption and crime [Drescher 2014]



Sleep deprivation causes deficit in attention and working memory [Krause et al. 2017]

Linear Regression



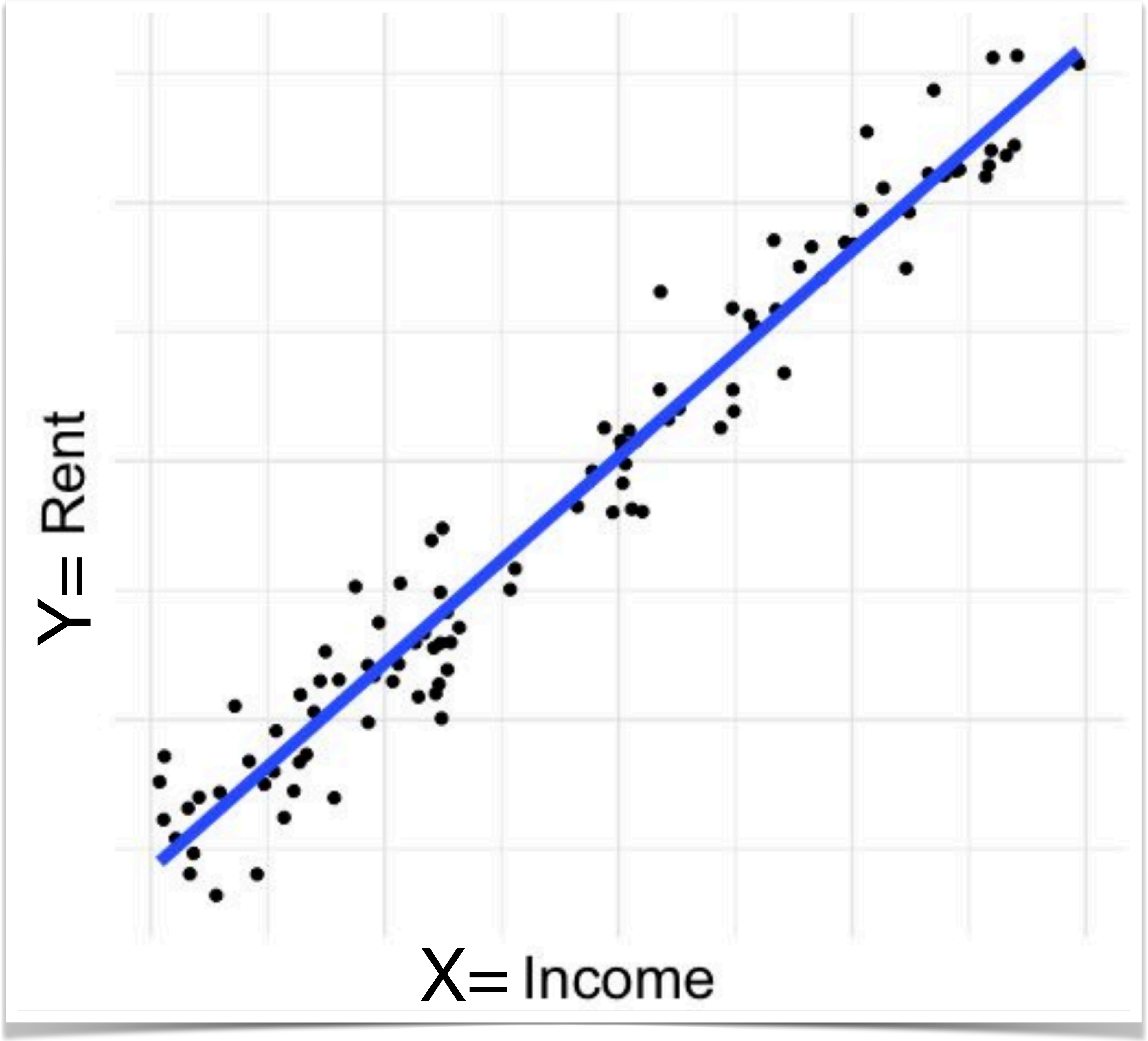
Linear regression

Basics

Linear regression

Basics

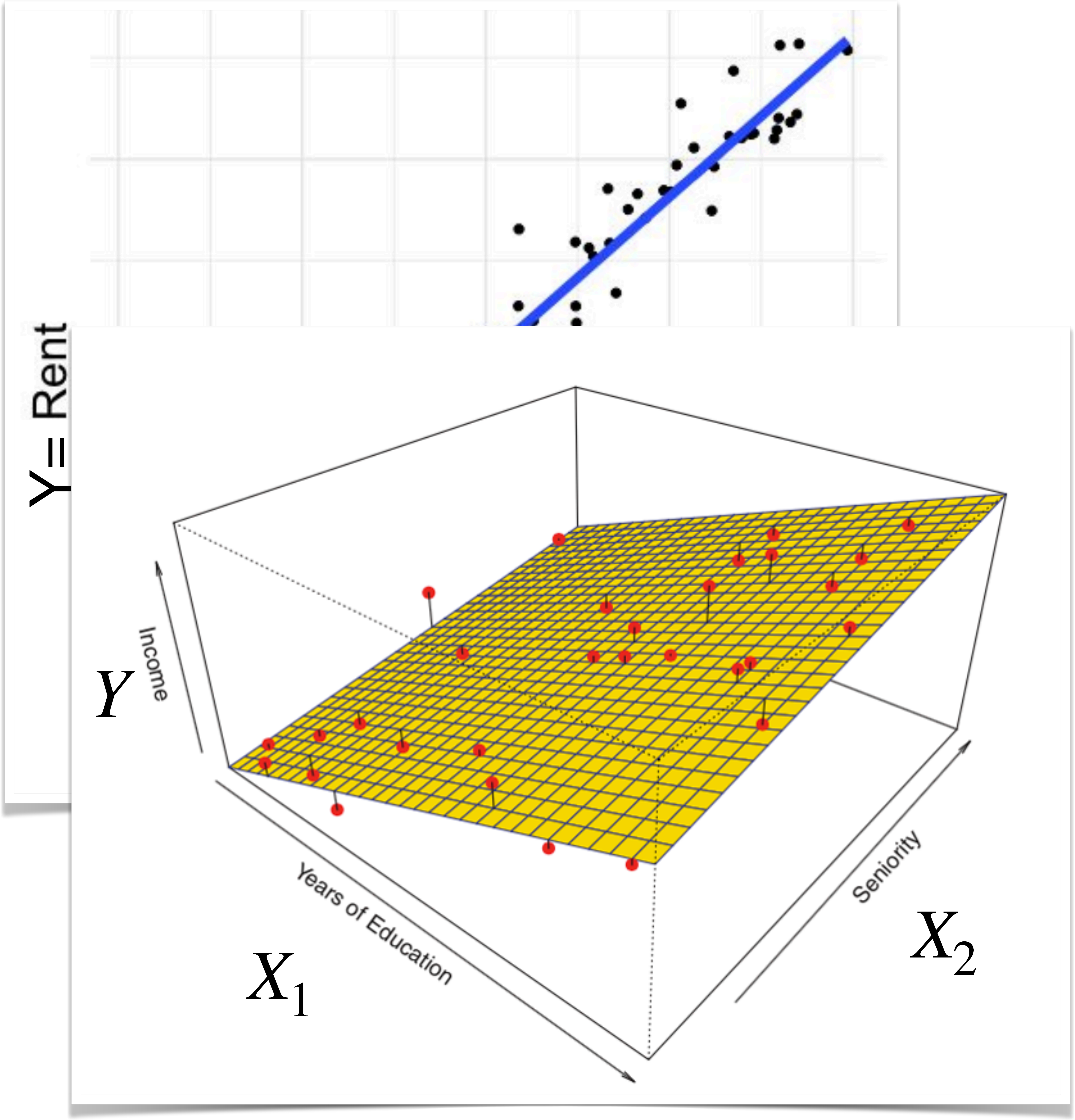
- It predicts a continuous variable Y using one or multiple variables X's (can be continuous, categorical, or both).



Linear regression

Basics

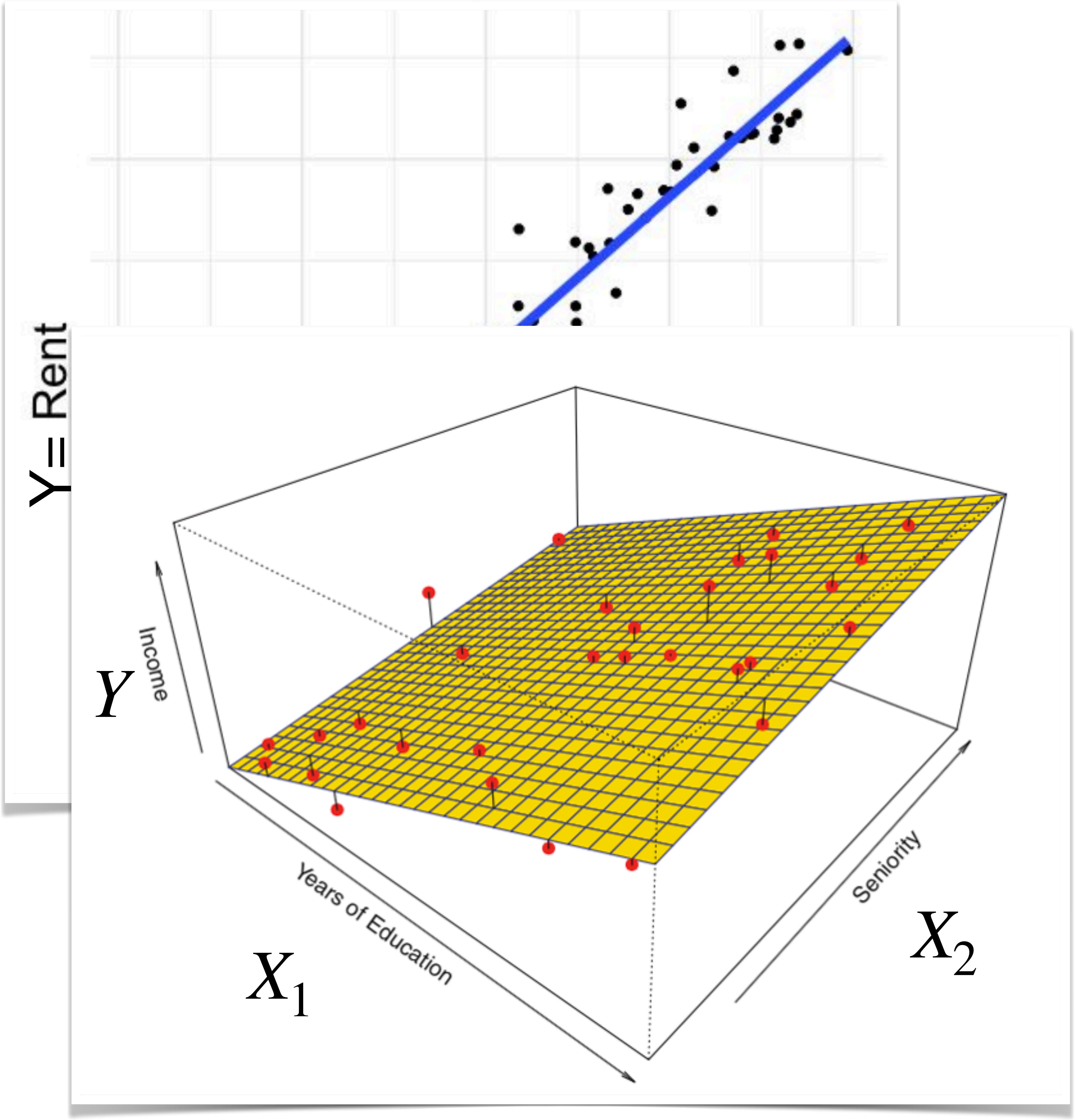
- It predicts a continuous variable Y using one or multiple variables X's (can be continuous, categorical, or both).
- It describes the relationship between these variables using a line (2D) or a plane (3D)



Linear regression

Basics

- It predicts a continuous variable Y using one or multiple variables X's (can be continuous, categorical, or both).
- It describes the relationship between these variables using a line (2D) or a plane (3D)
    - Similar to correlation, but **correlation** measures the strength and direction of a linear relationship between two variables, and **regression** measures how those variables affect each other using an equation (it estimates the best straight line that summarizes the relation).



Linear regression

Definition

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Y is called the dependent (output) variable

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Y is called the dependent (output) variable
- $X_1, X_2, \ldots, X_k$ are called the independent (predictor, feature) variables

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Y is called the dependent (output) variable
- $X_1, X_2, \ldots, X_k$ are called the independent (predictor, feature) variables
- a is the intercept, which measures the expected value of Y that does not depend on the dependent variables

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Y is called the dependent (output) variable
- $X_1, X_2, \ldots, X_k$ are called the independent (predictor, feature) variables
- a is the intercept, which measures the expected value of Y that does not depend on the dependent variables
- $\beta_1, \beta_2, \ldots, \beta_k$ are called the slopes or the coefficients

Linear regression

Definition

- Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables, e.g. $X_1, X_2, \ldots, X_k$

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Y is called the dependent (output) variable
- $X_1, X_2, \ldots, X_k$ are called the independent (predictor, feature) variables
- a is the intercept, which measures the expected value of Y that does not depend on the dependent variables
- $\beta_1, \beta_2, \ldots, \beta_k$ are called the slopes or the coefficients
- ϵ are the residuals, the errors of the equation in the data

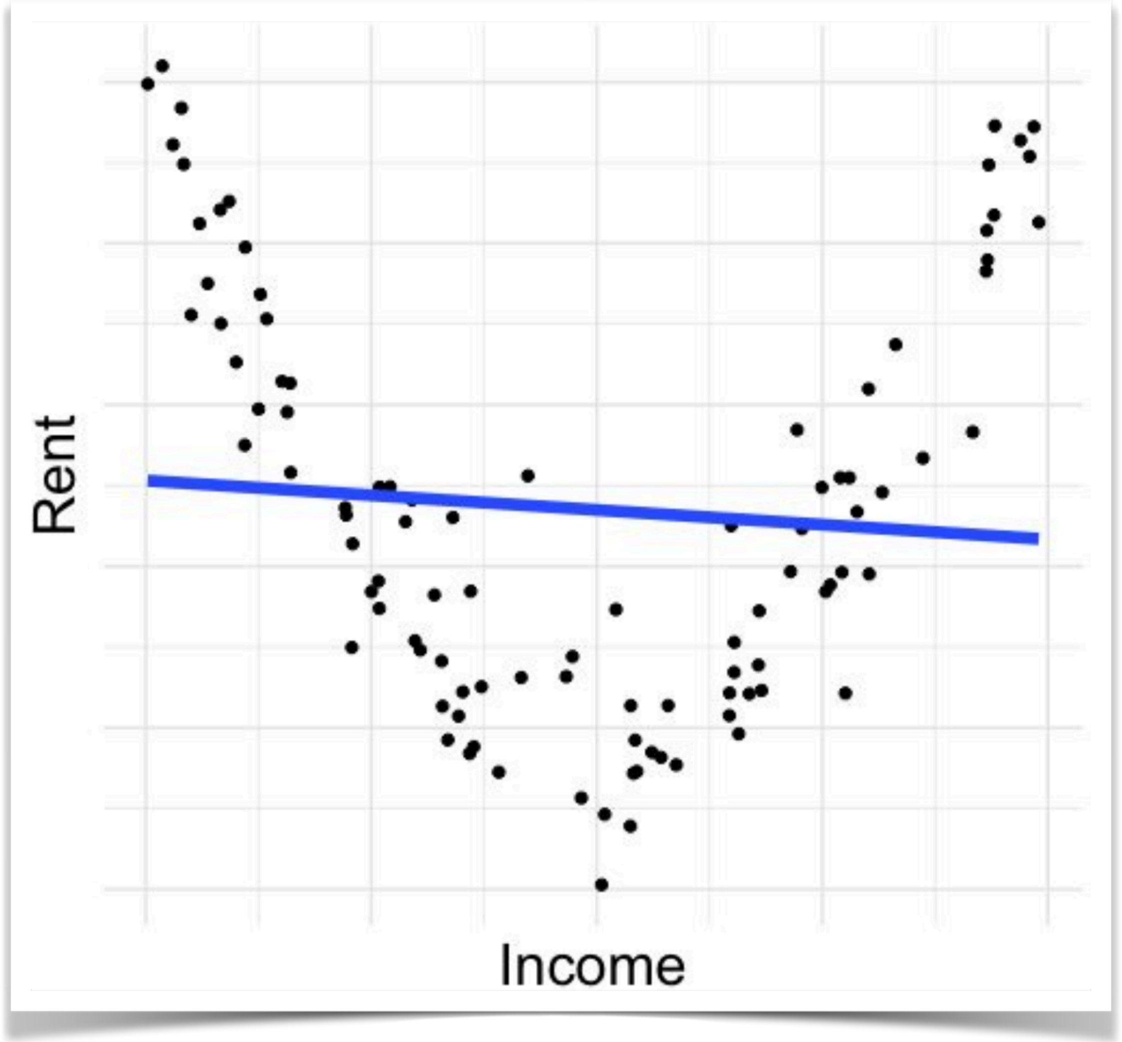
Linear regression

Assumptions

Linear regression

Assumptions

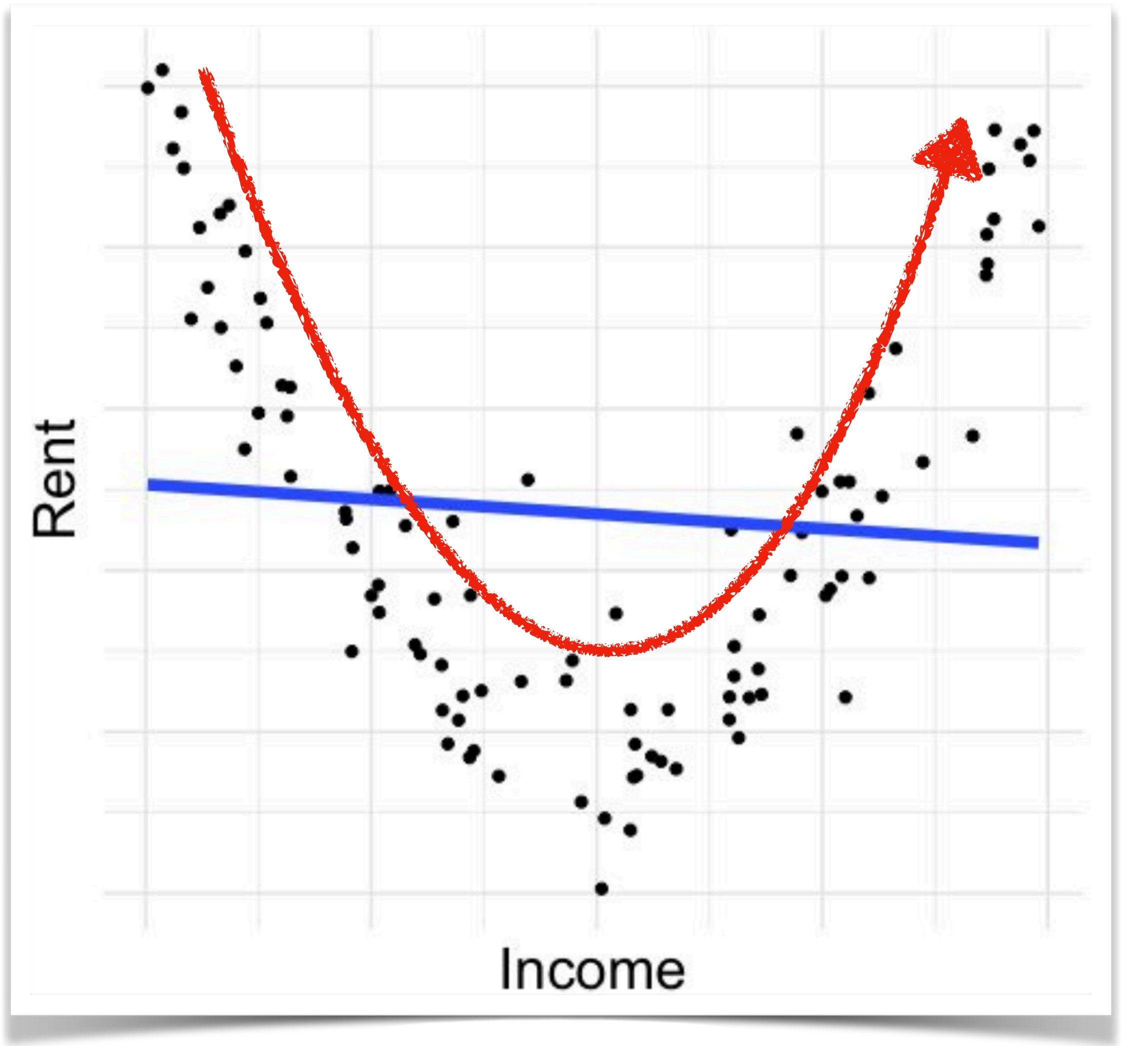
- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.



Linear regression

Assumptions

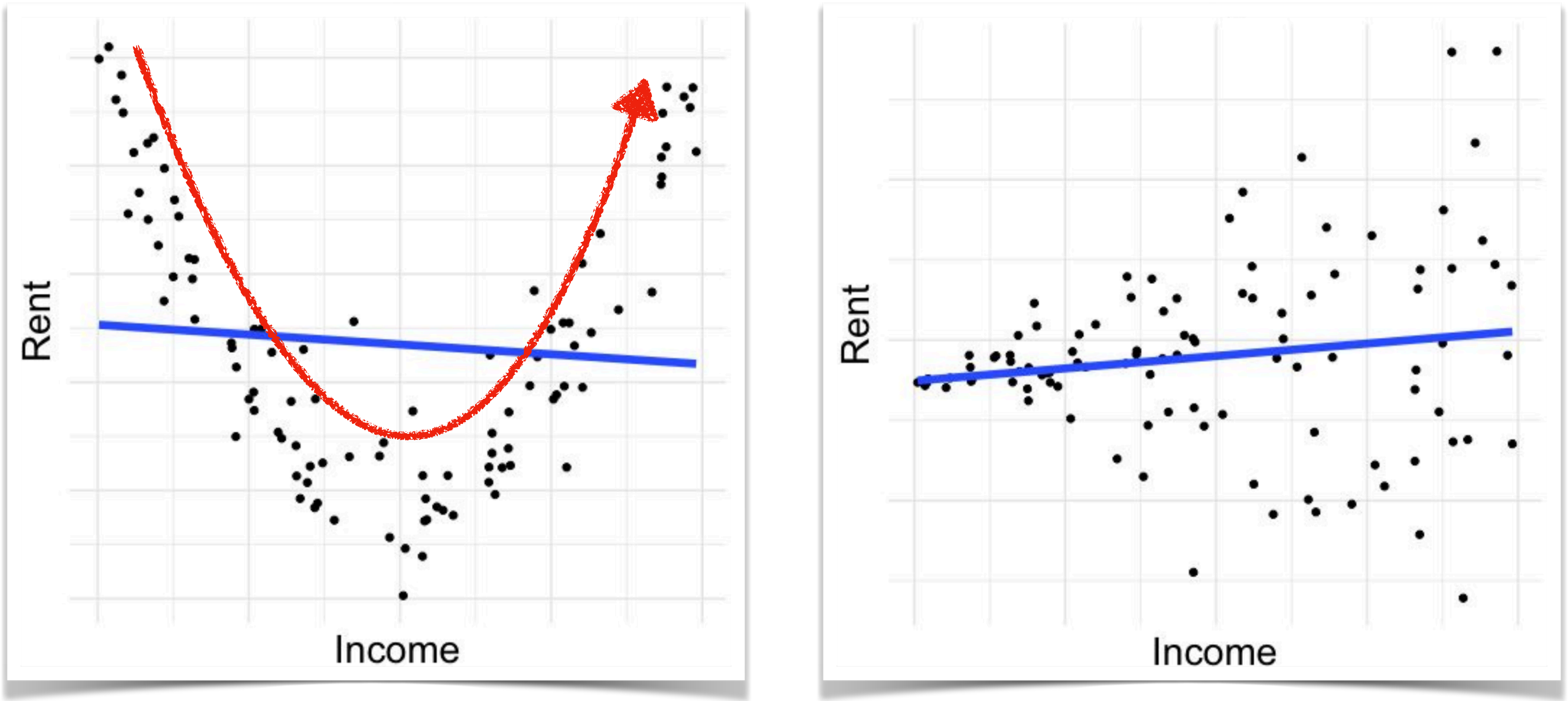
- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.



Linear regression

Assumptions

- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.
- **Homoscedasticity:** The variance of ϵ (the residuals) is the same for any value of X.
  - The variance of the data points is roughly the same for all data points.

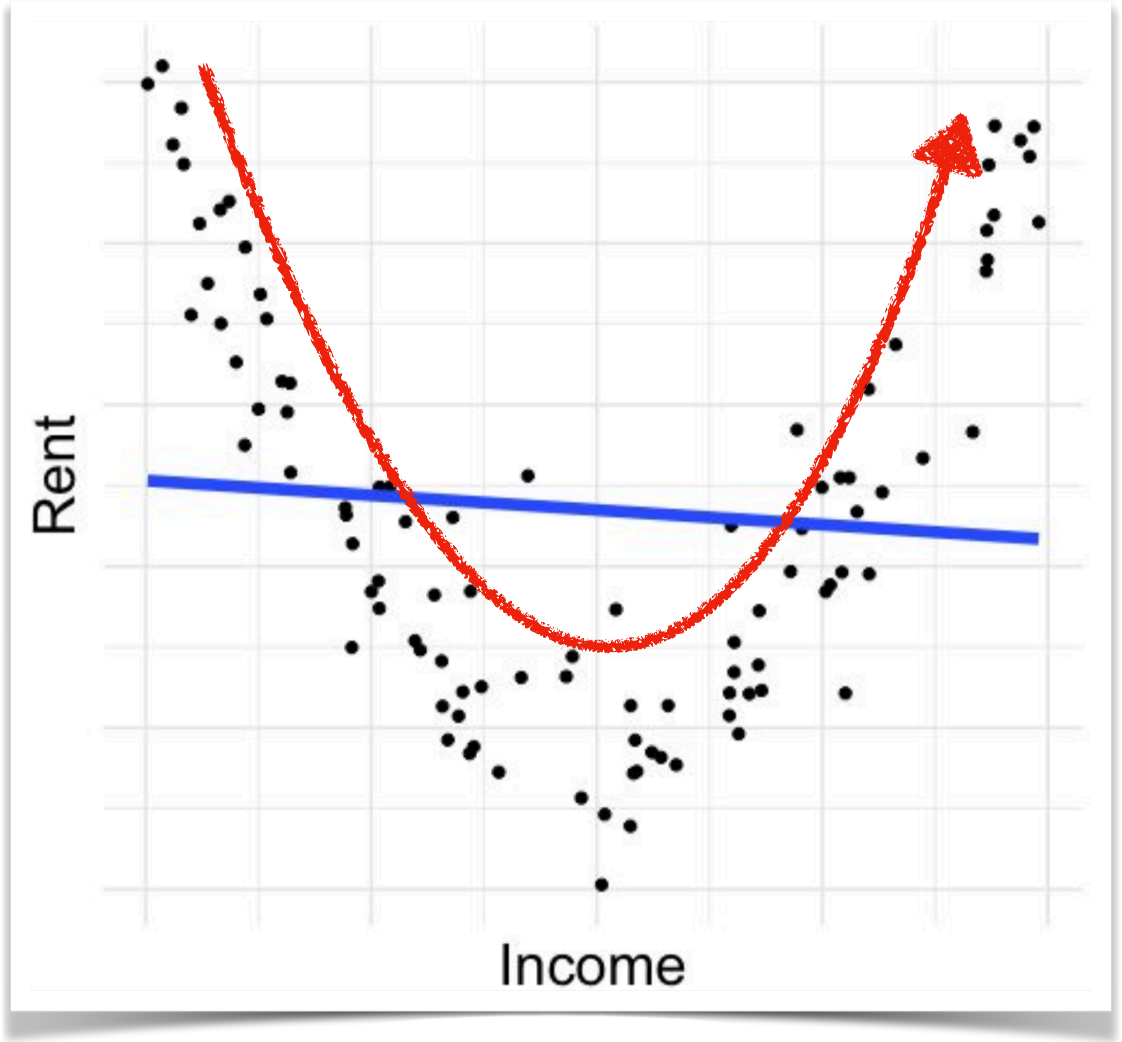


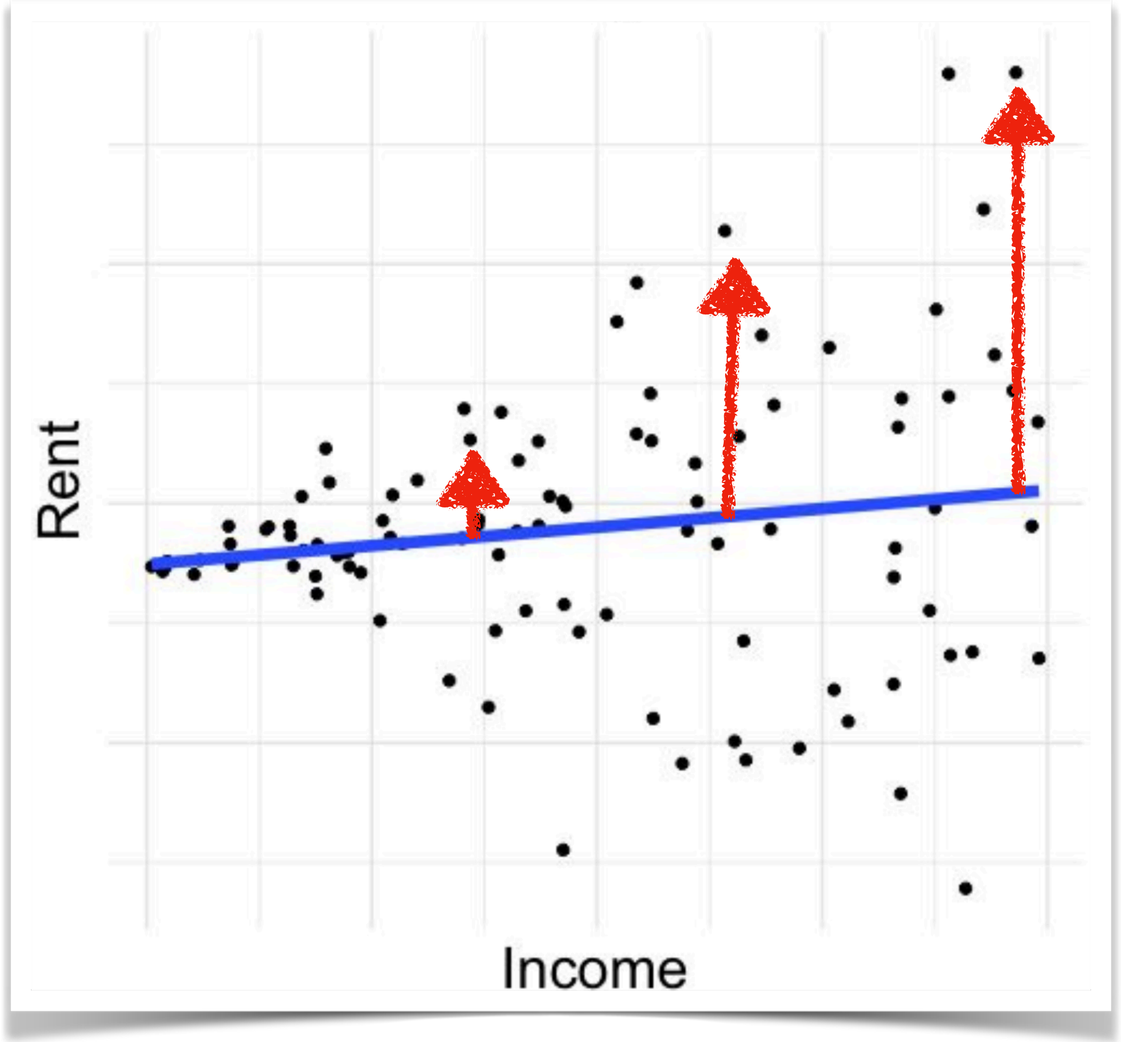
Linear regression

Assumptions

- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.

- **Homoscedasticity:** The variance of ϵ (the residuals) is the same for any value of X.
  - The variance of the data points is roughly the same for all data points.

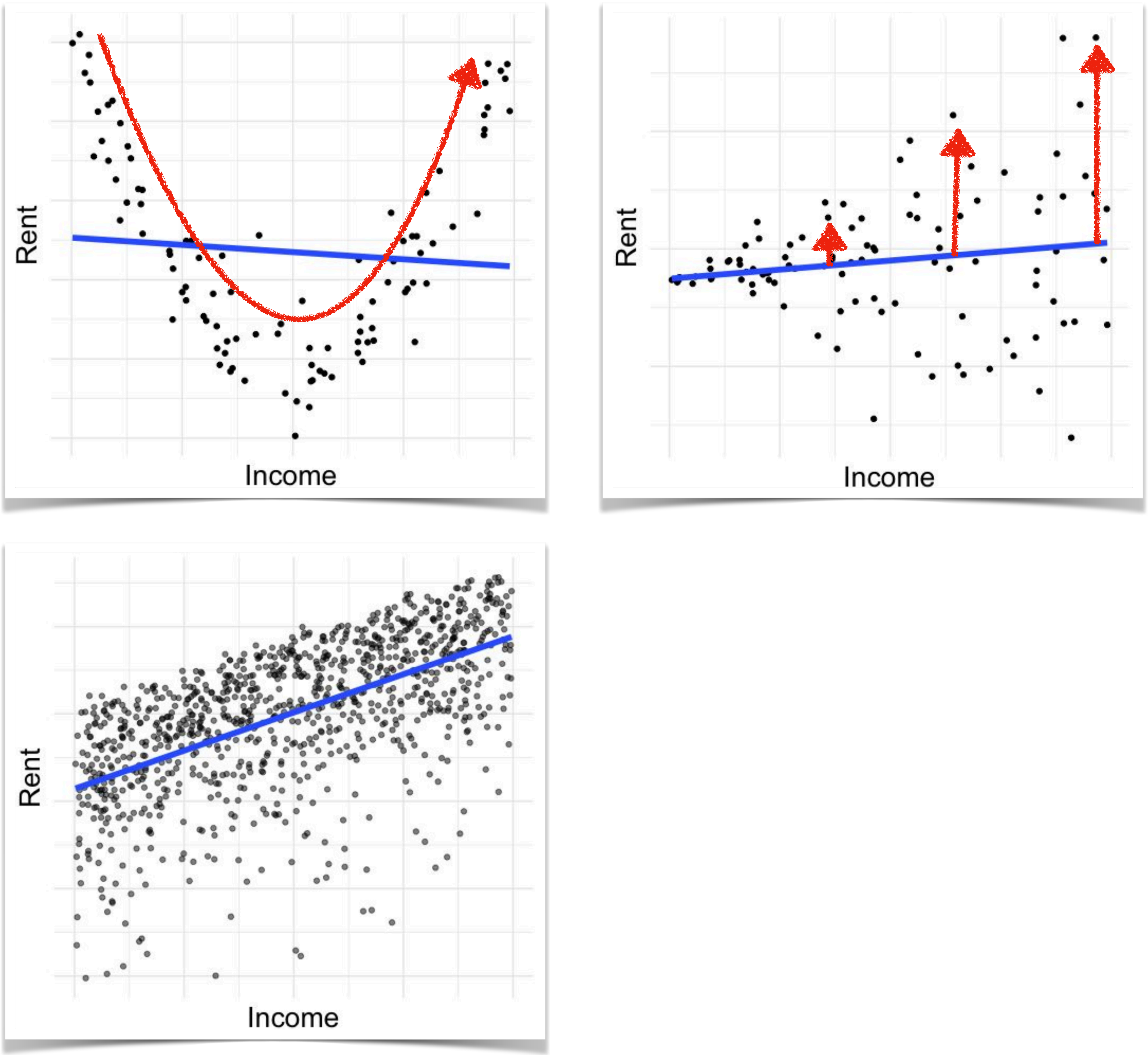




Linear regression

Assumptions

- **Linearity:** The relationship between X and the mean of Y is linear.
- **Homoscedasticity:** The variance of ϵ (the residuals) is the same for any value of X.
    - The variance of the data points is roughly the same for all data points.
- **Normality:** For any fixed value of X, Y is normally distributed.



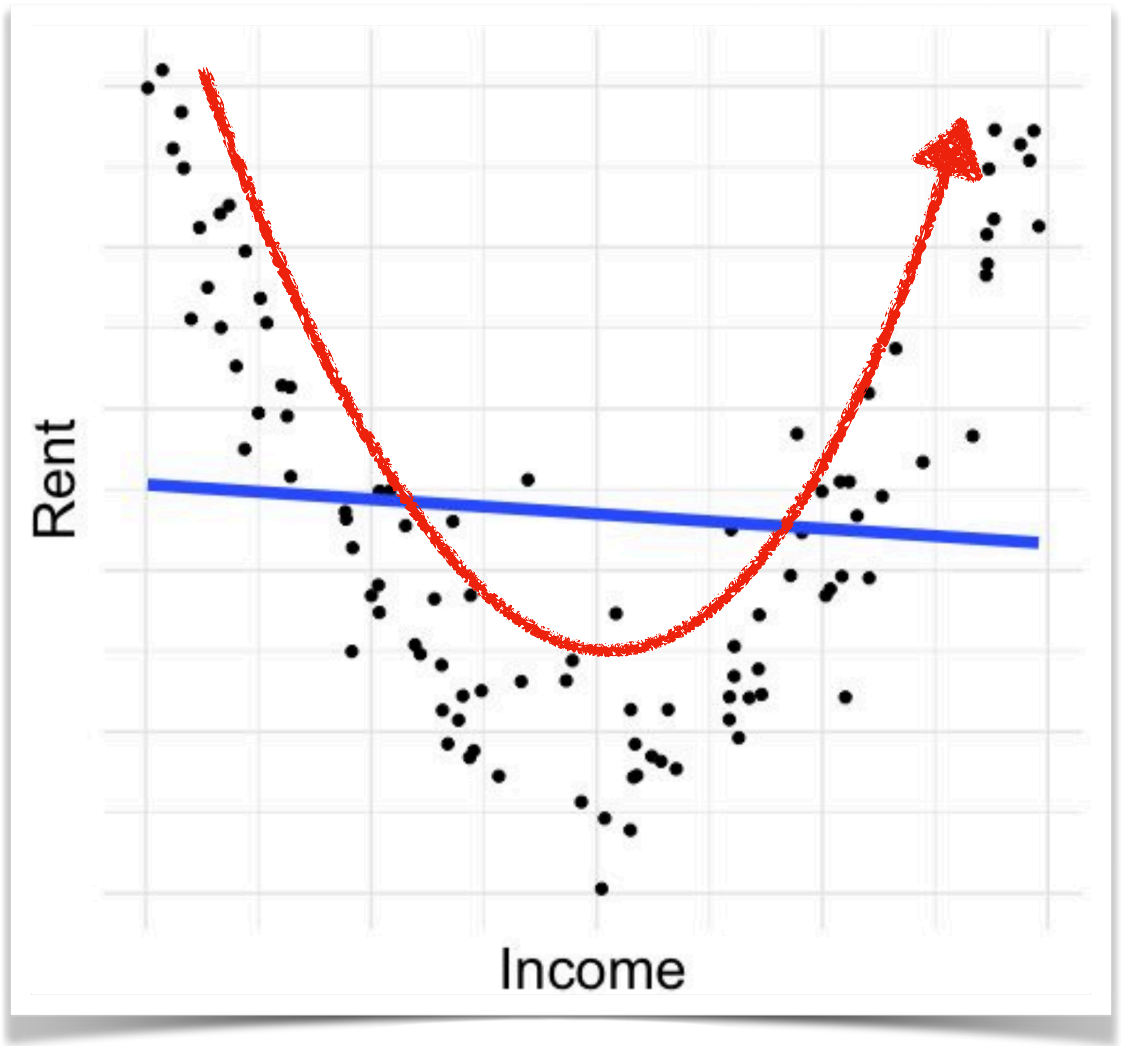
Linear regression

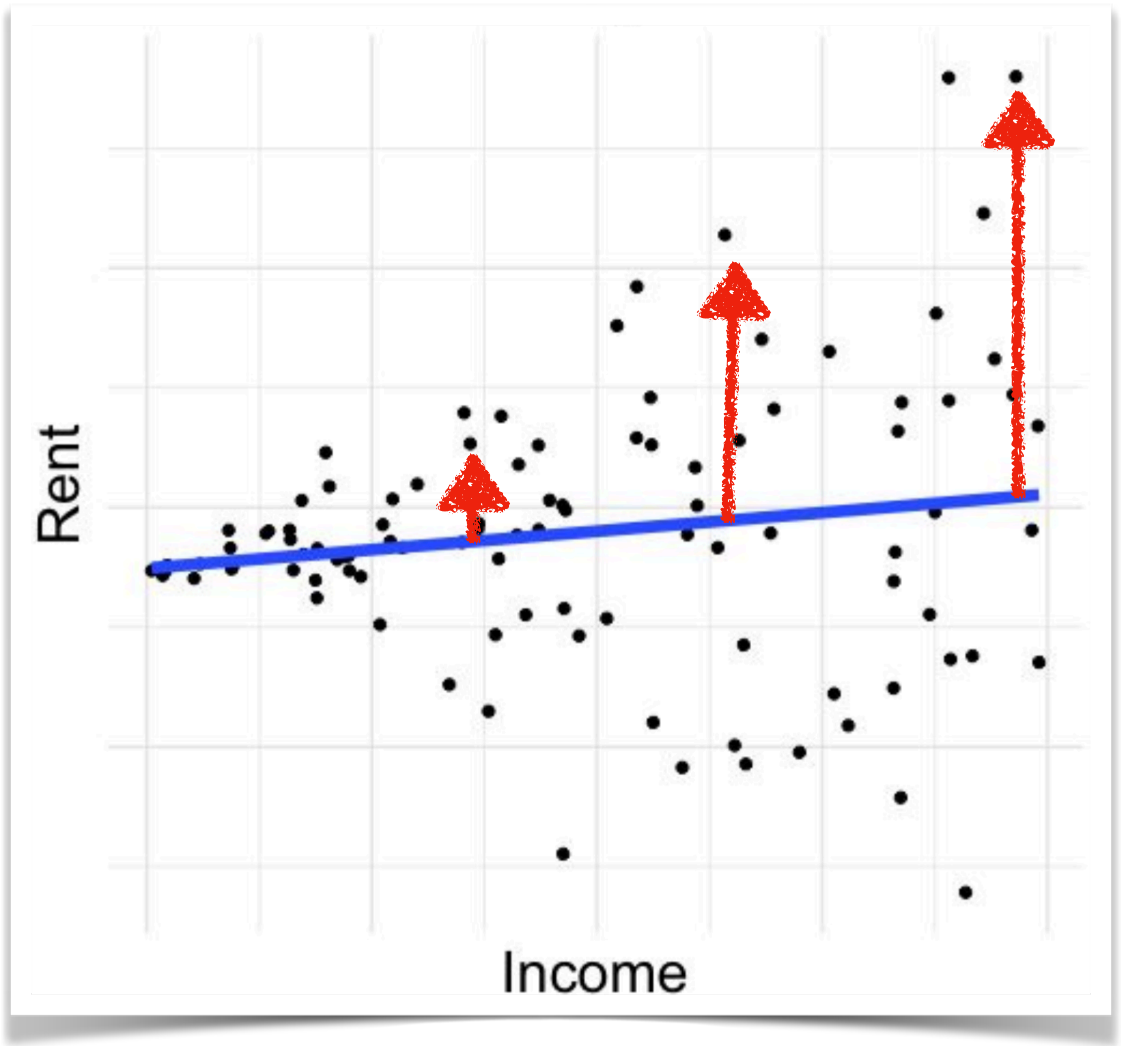
Assumptions

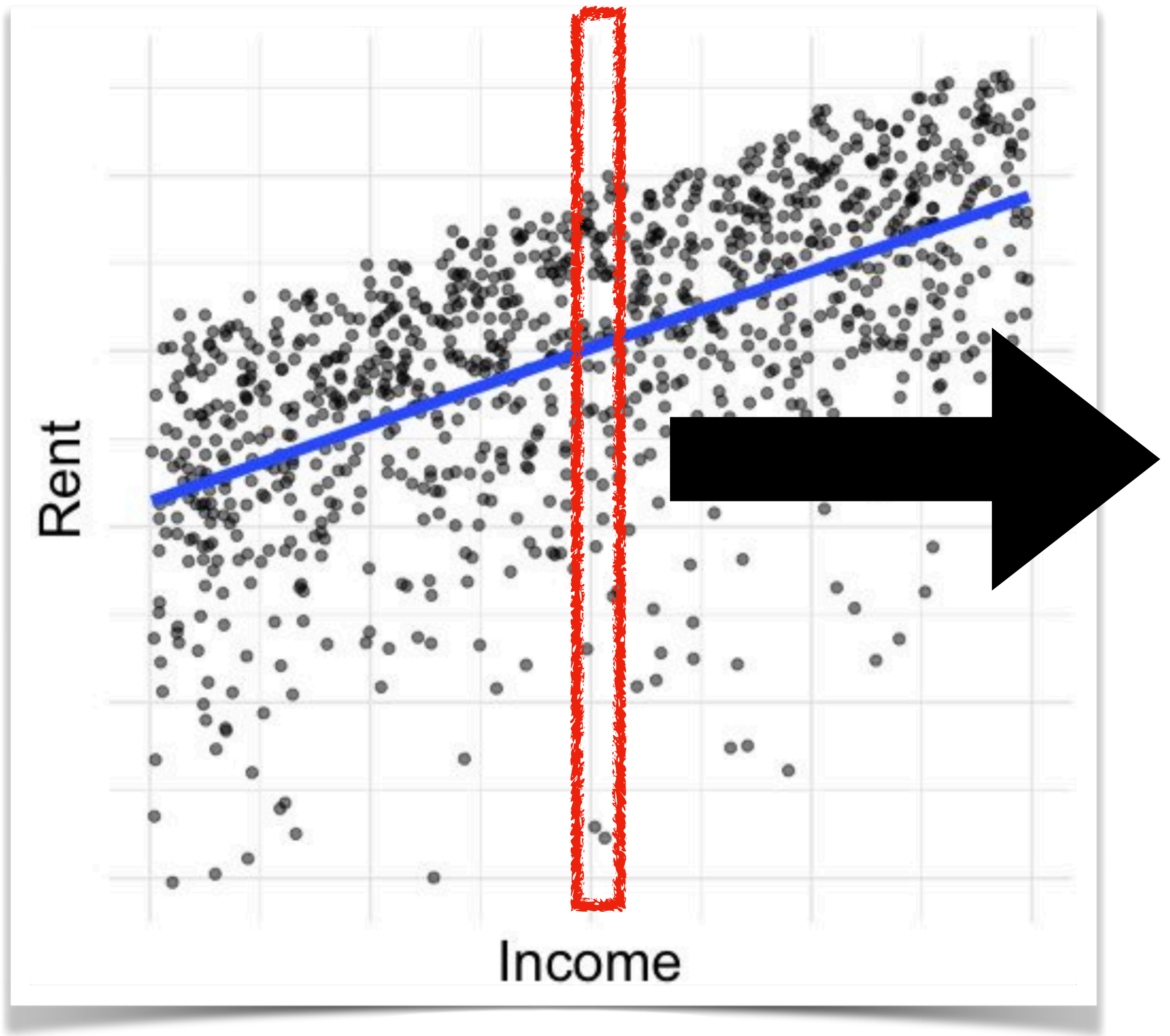
- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.

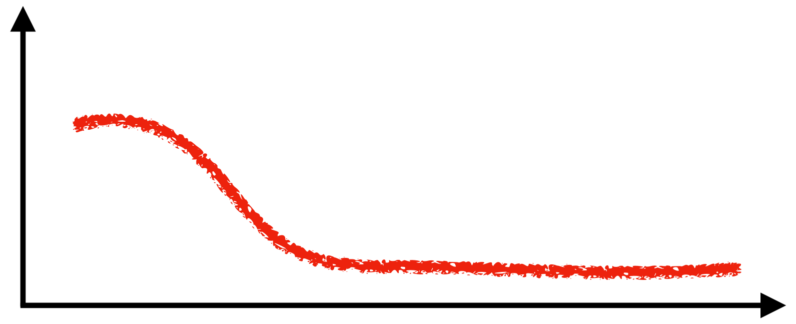
- **Homoscedasticity:** The variance of ϵ (the residuals) is the same for any value of X.
  - The variance of the data points is roughly the same for all data points.

- **Normality:** For any fixed value of X, Y is normally distributed.







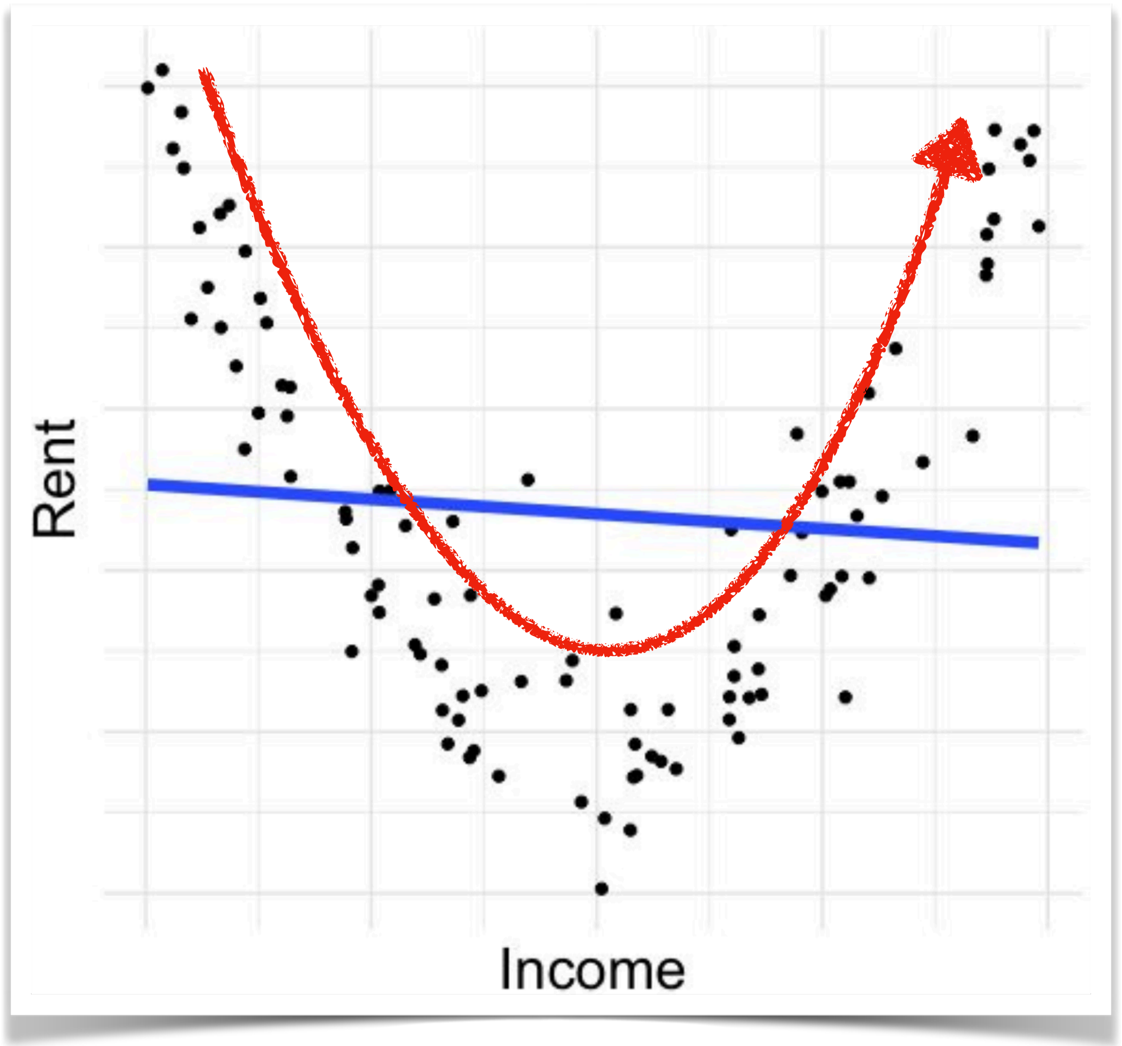


(not a bell-shape)

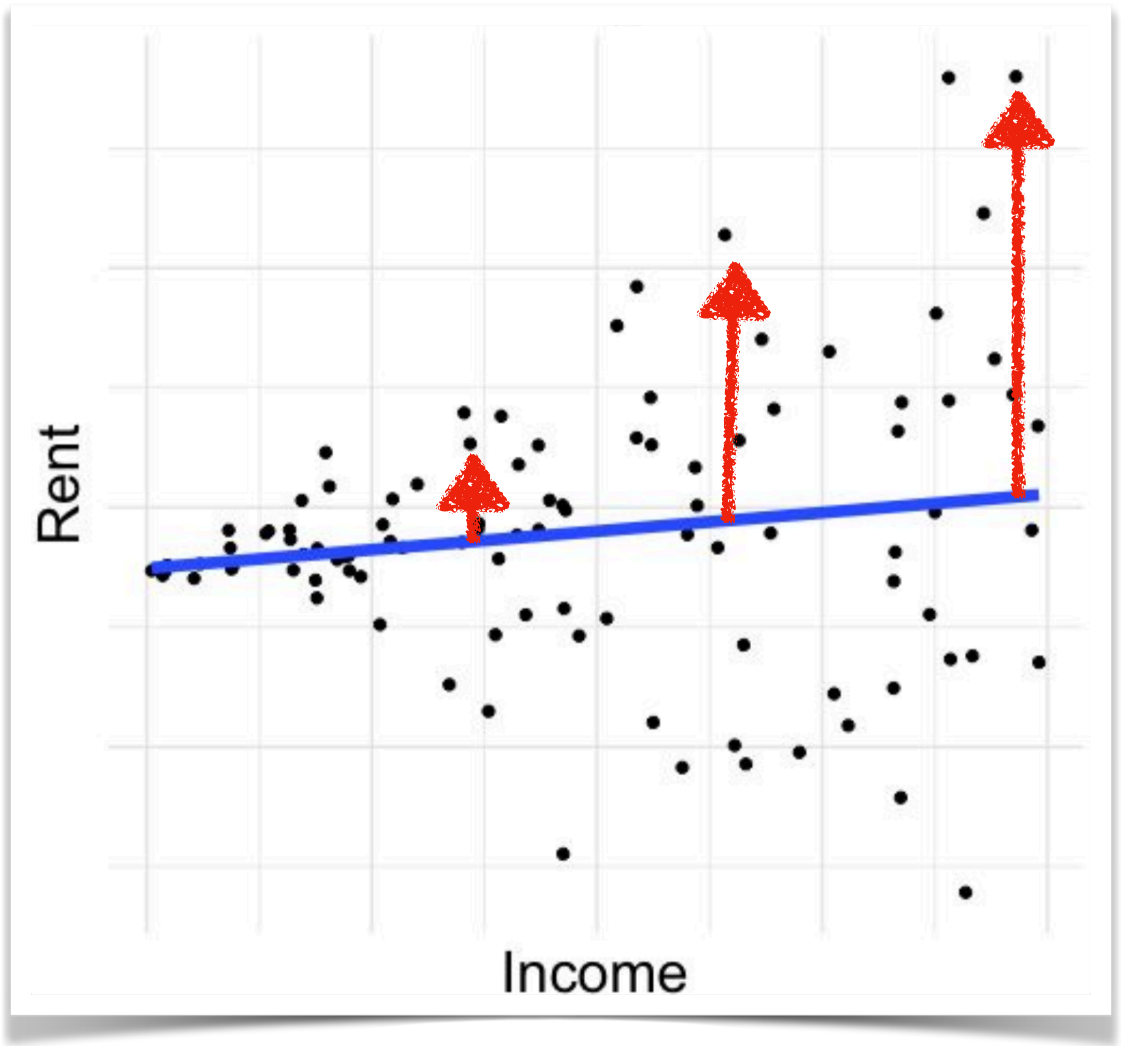
Linear regression

Assumptions

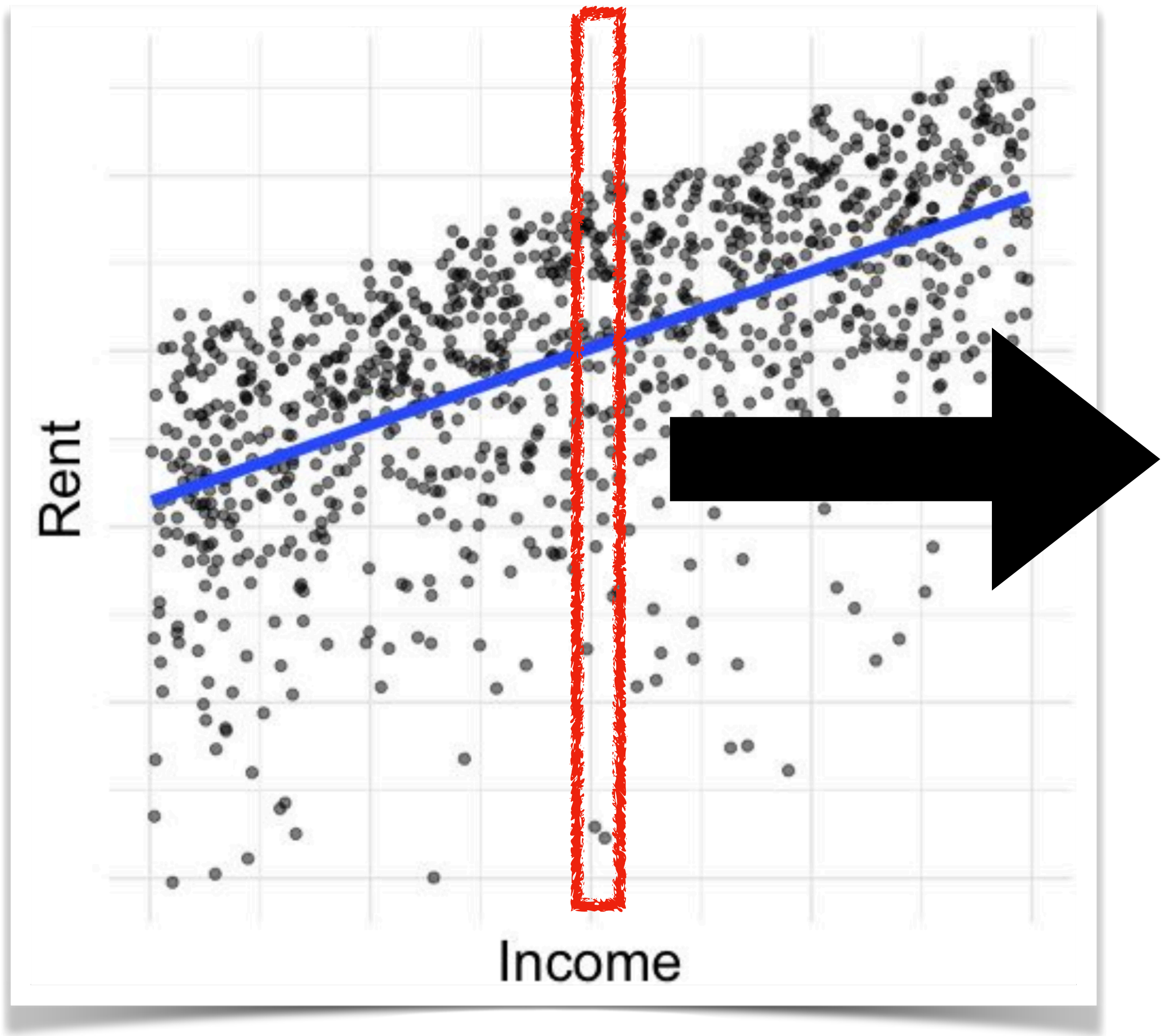
- **Linearity:** The relationship between X and the <u>mean</u> of Y is linear.

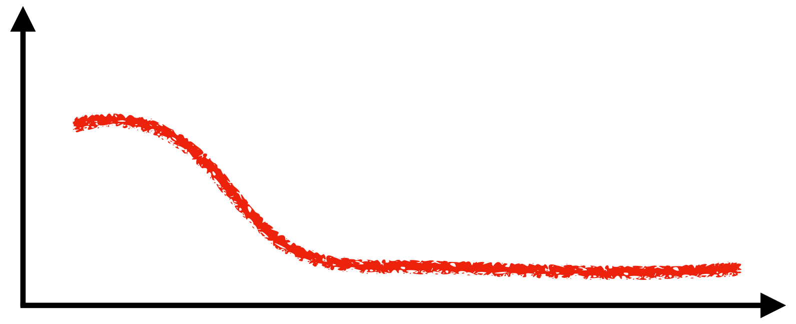


- **Homoscedasticity:** The variance of ϵ (the residuals) is the same for any value of X.
    - The variance of the data points is roughly the same for all data points.



- **Normality:** For any fixed value of X, Y is normally distributed.



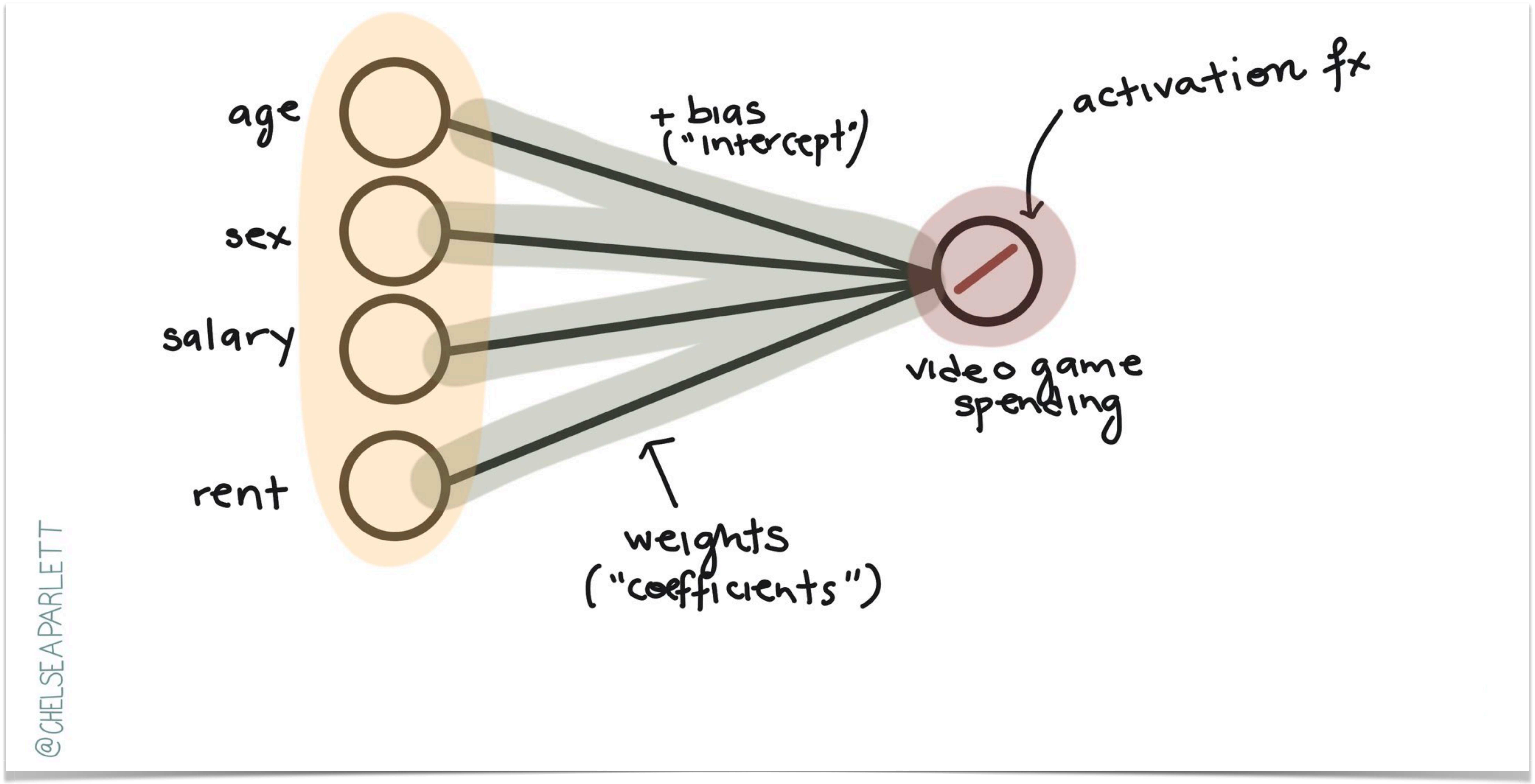


(not a bell-shape)

- **Independence:** Observations are independent of each other.

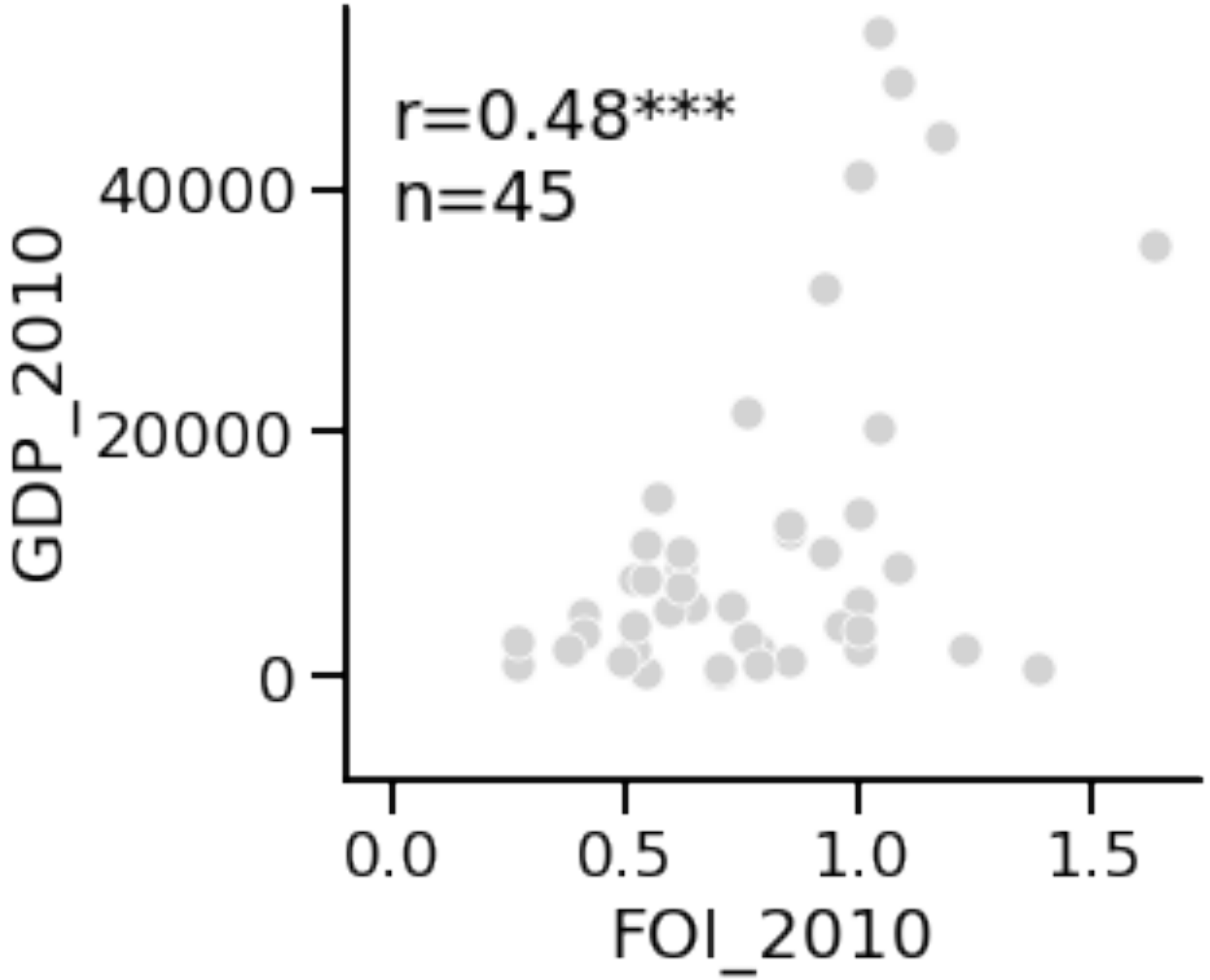
Linear regression

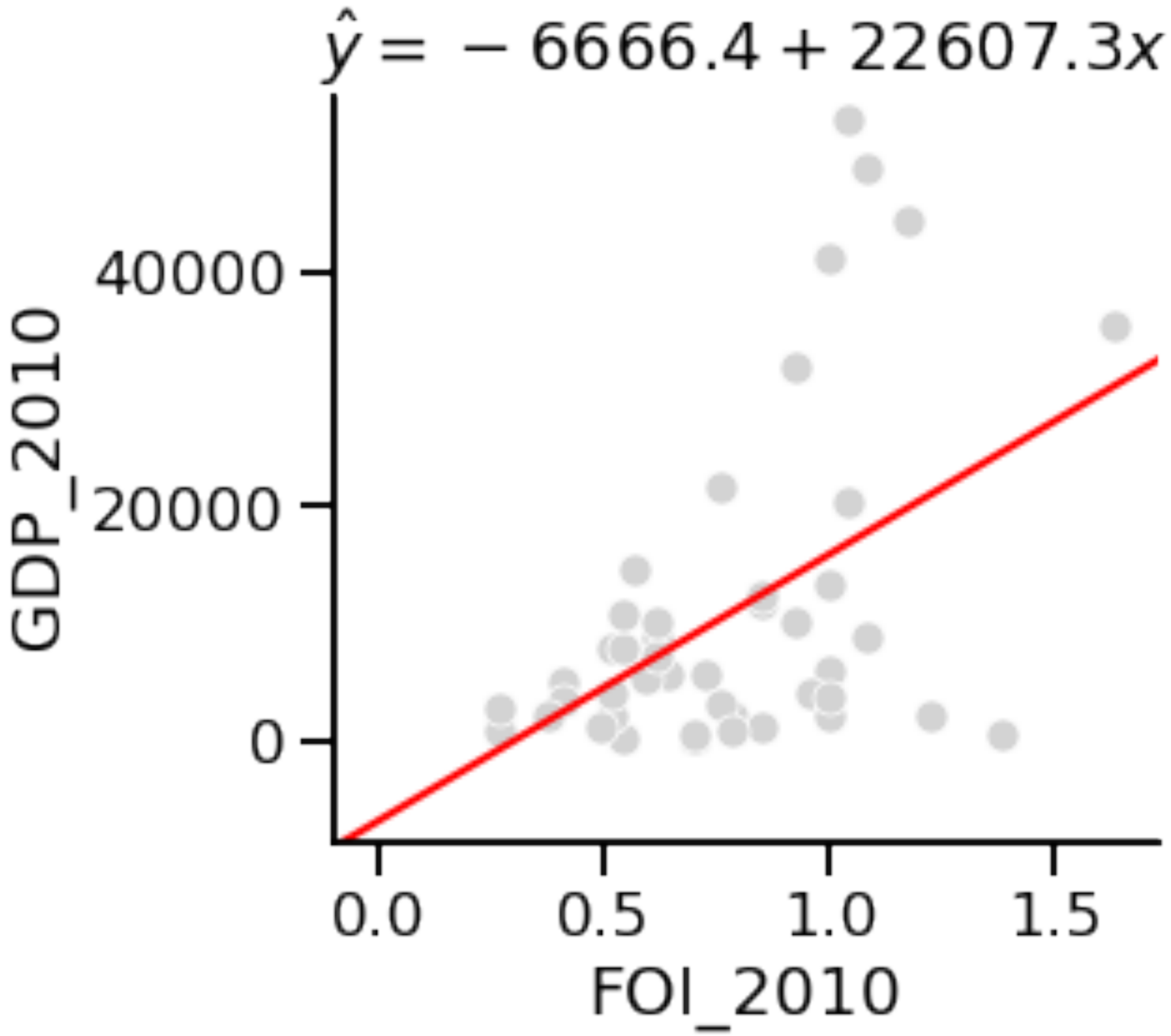
as a neural network



Example FOI and GDP

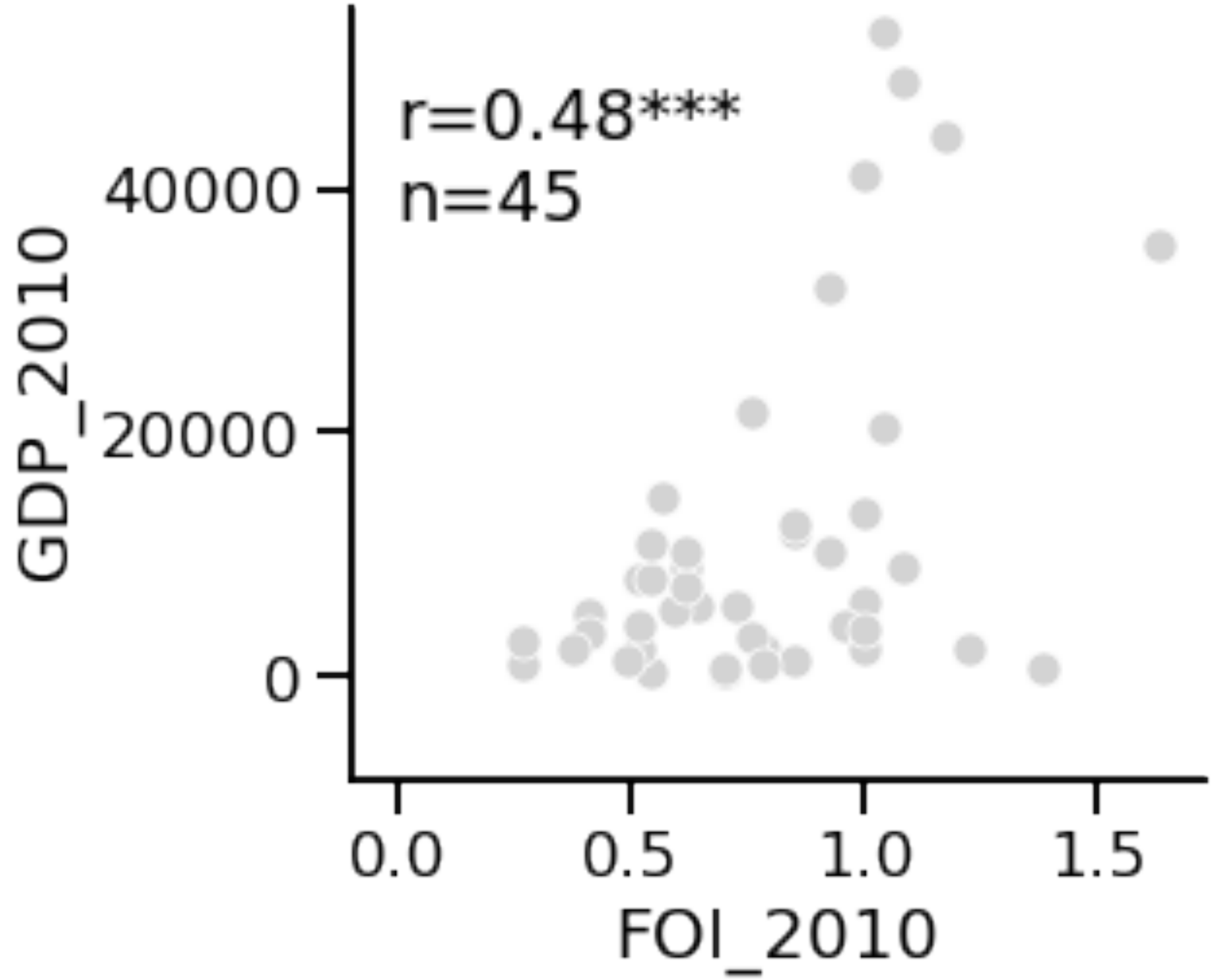
Correlation vs. Linear regression





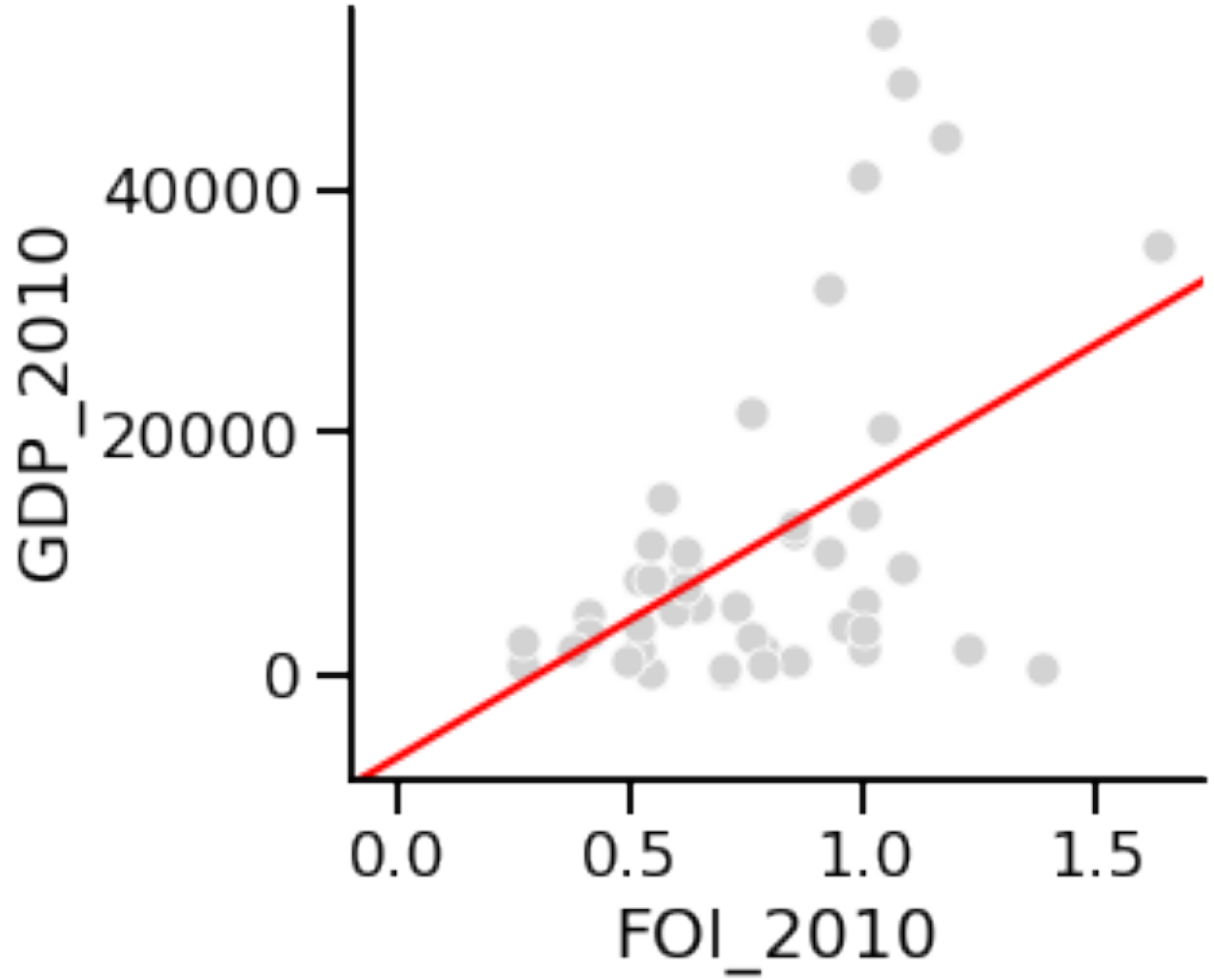
Example FOI and GDP

Correlation vs. Linear regression



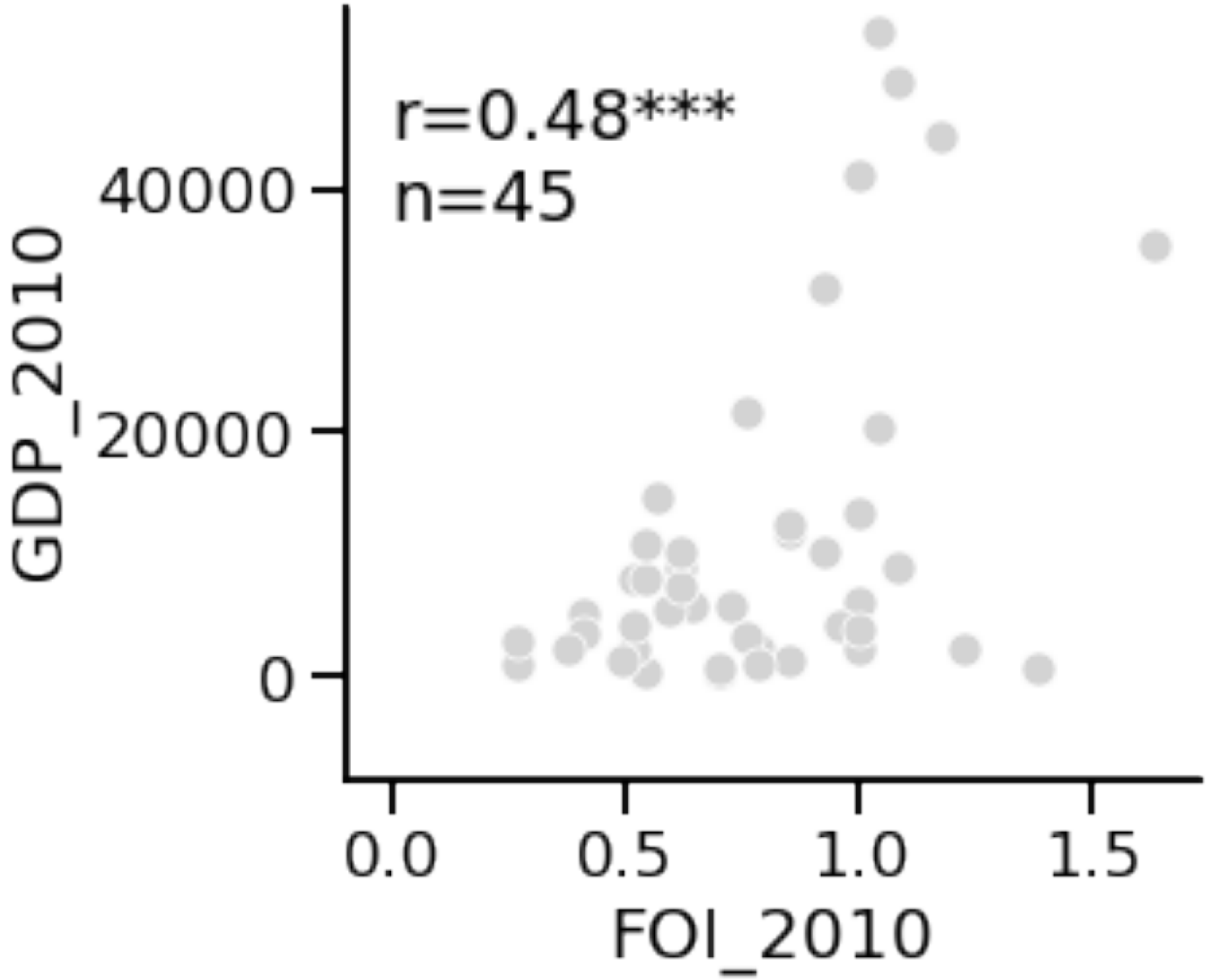
Intercept a

$\hat{y} = -6666.4 + 22607.3x$

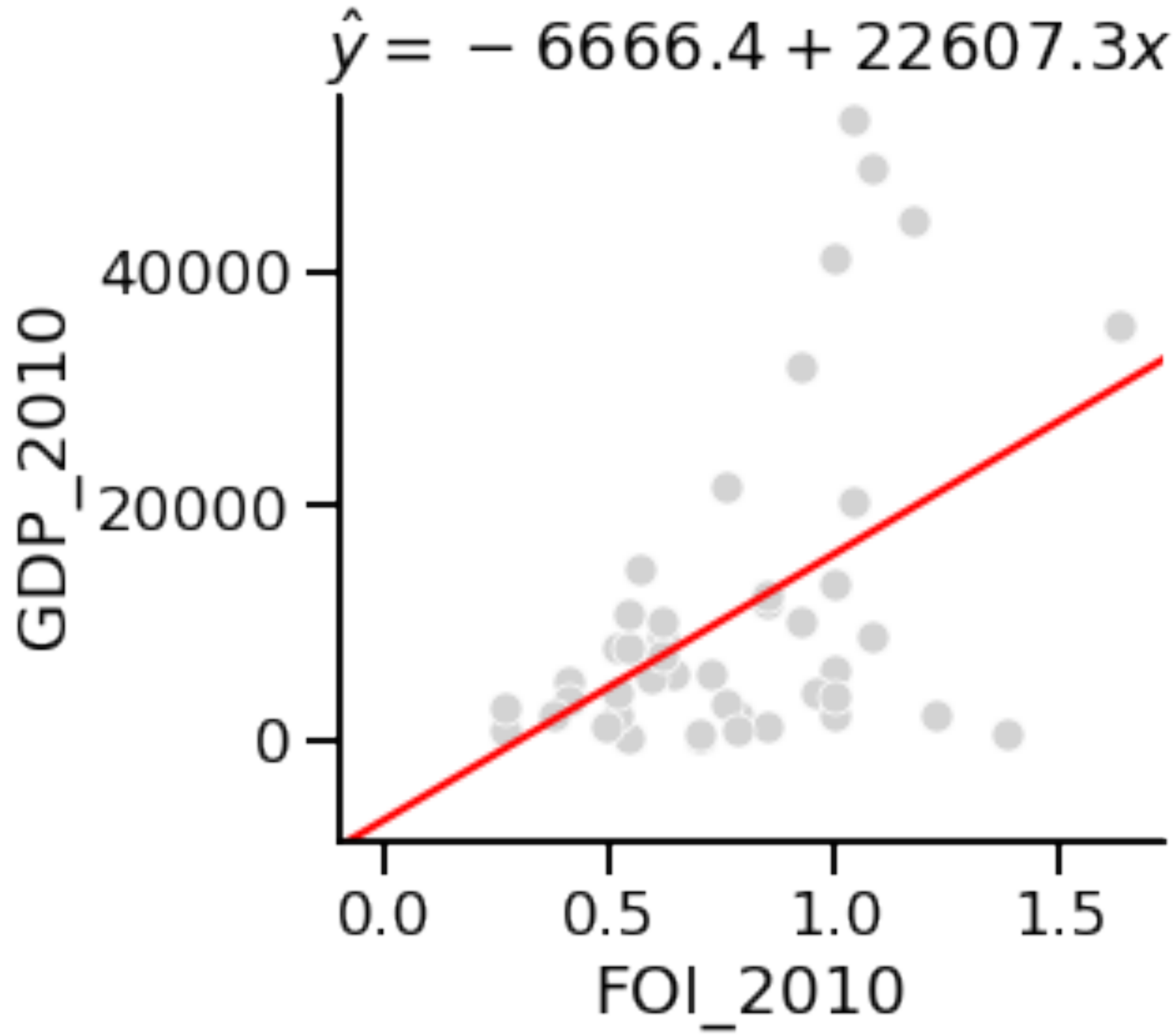


Example FOI and GDP

Correlation vs. Linear regression

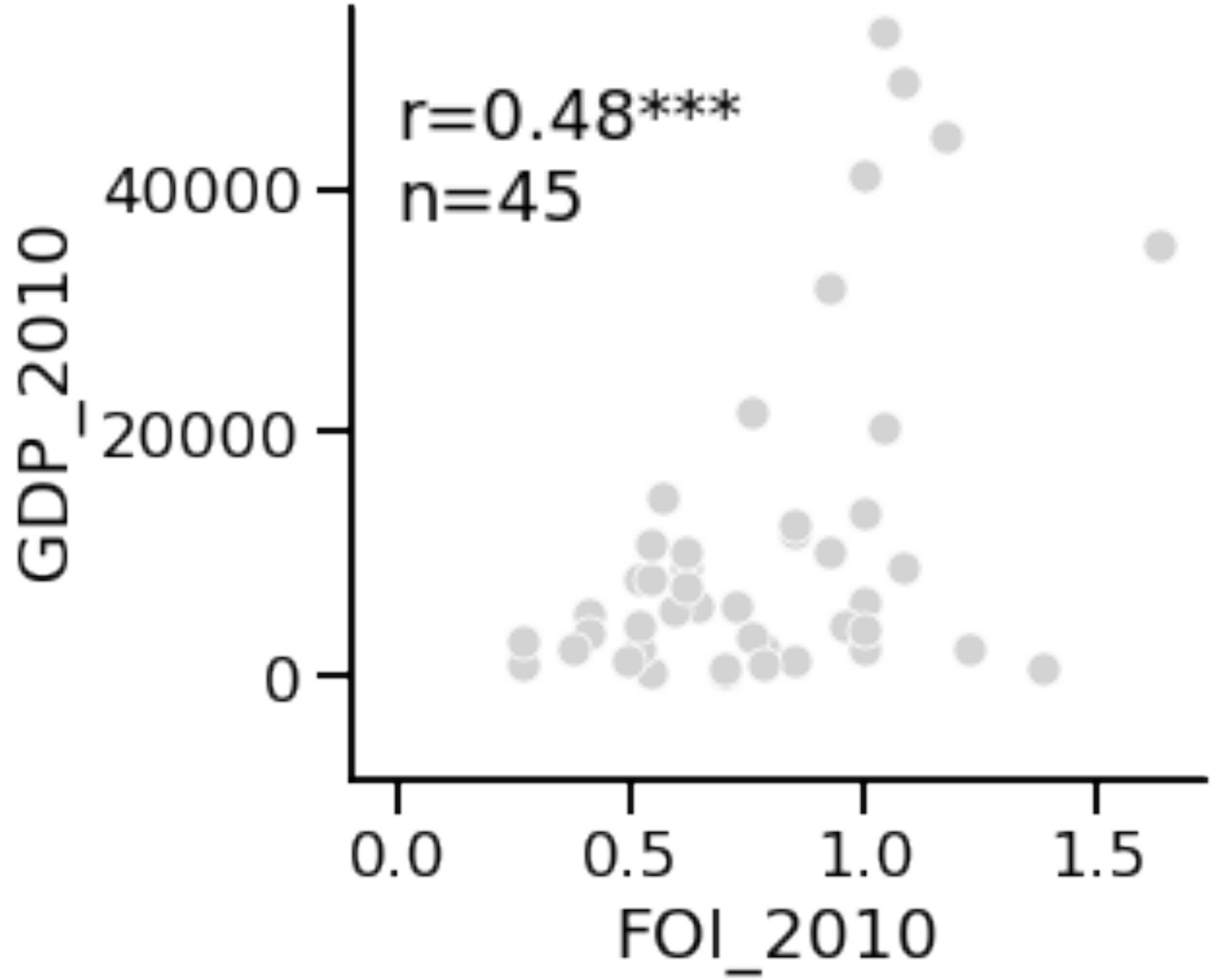


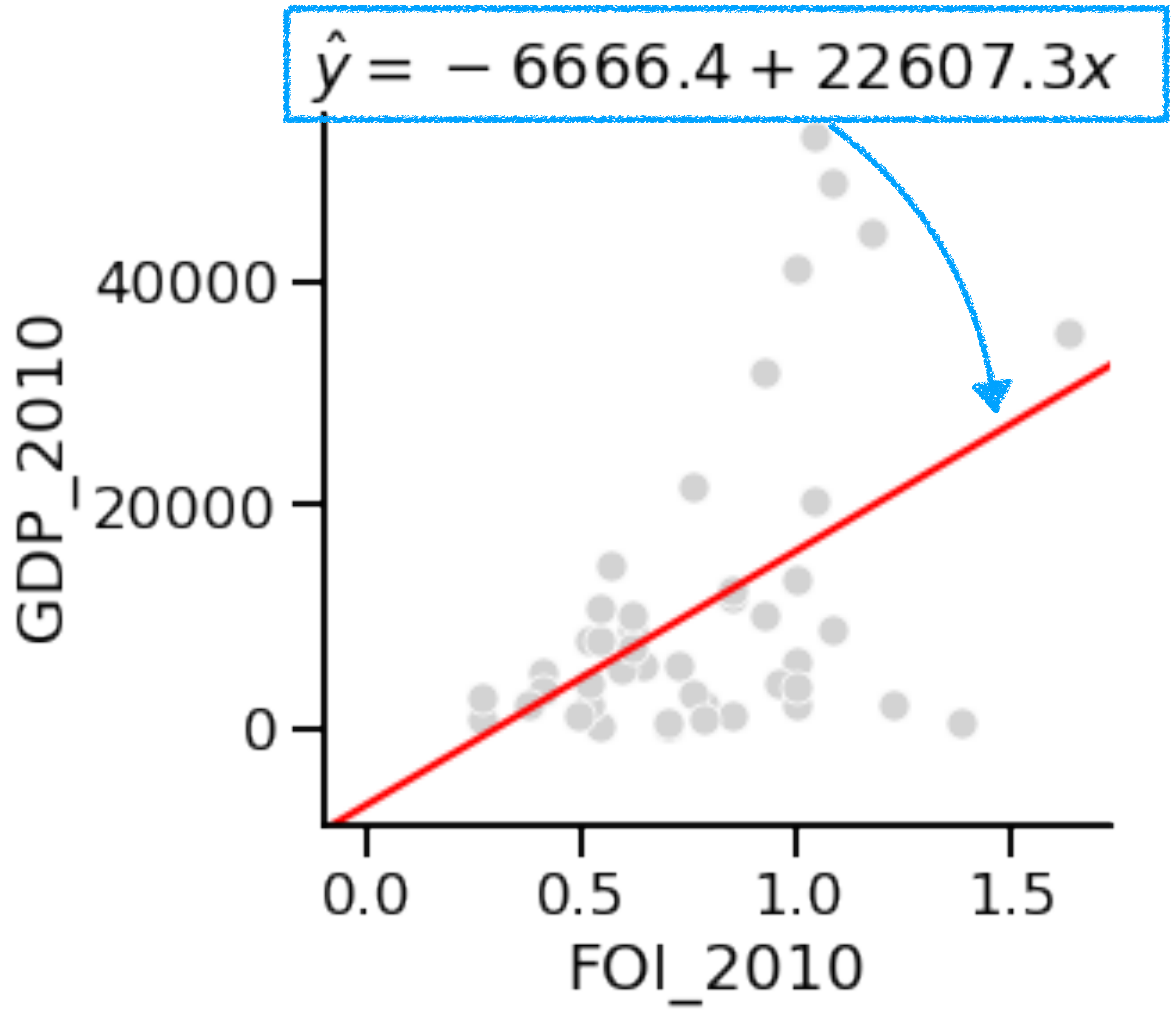
Intercept a Slope β_{FOI}



Example FOI and GDP

Correlation vs. Linear regression





Regression residuals

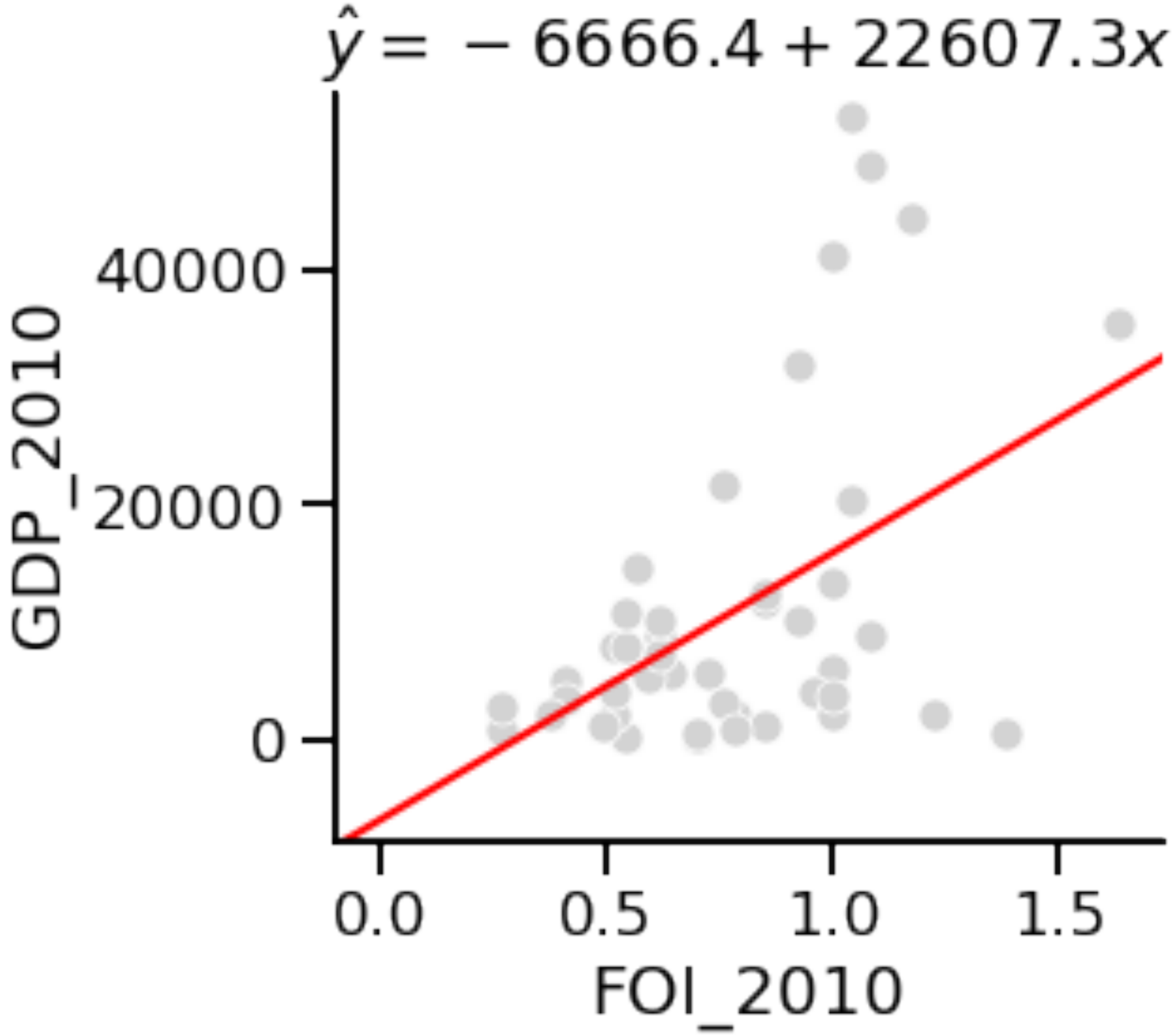
Error term: ϵ

- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$

Regression residuals

Error term: ϵ

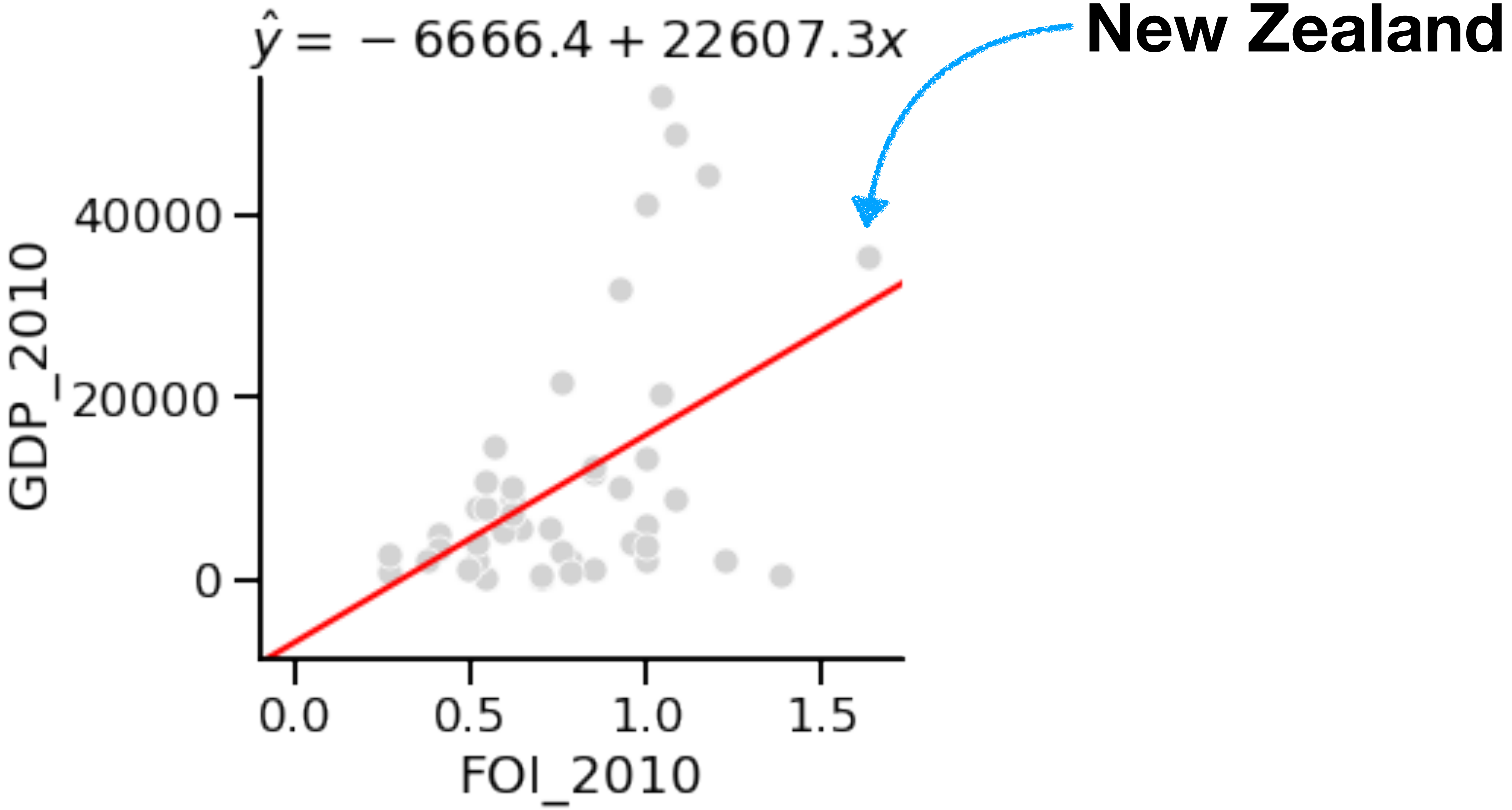
- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$



Regression residuals

Error term: ϵ

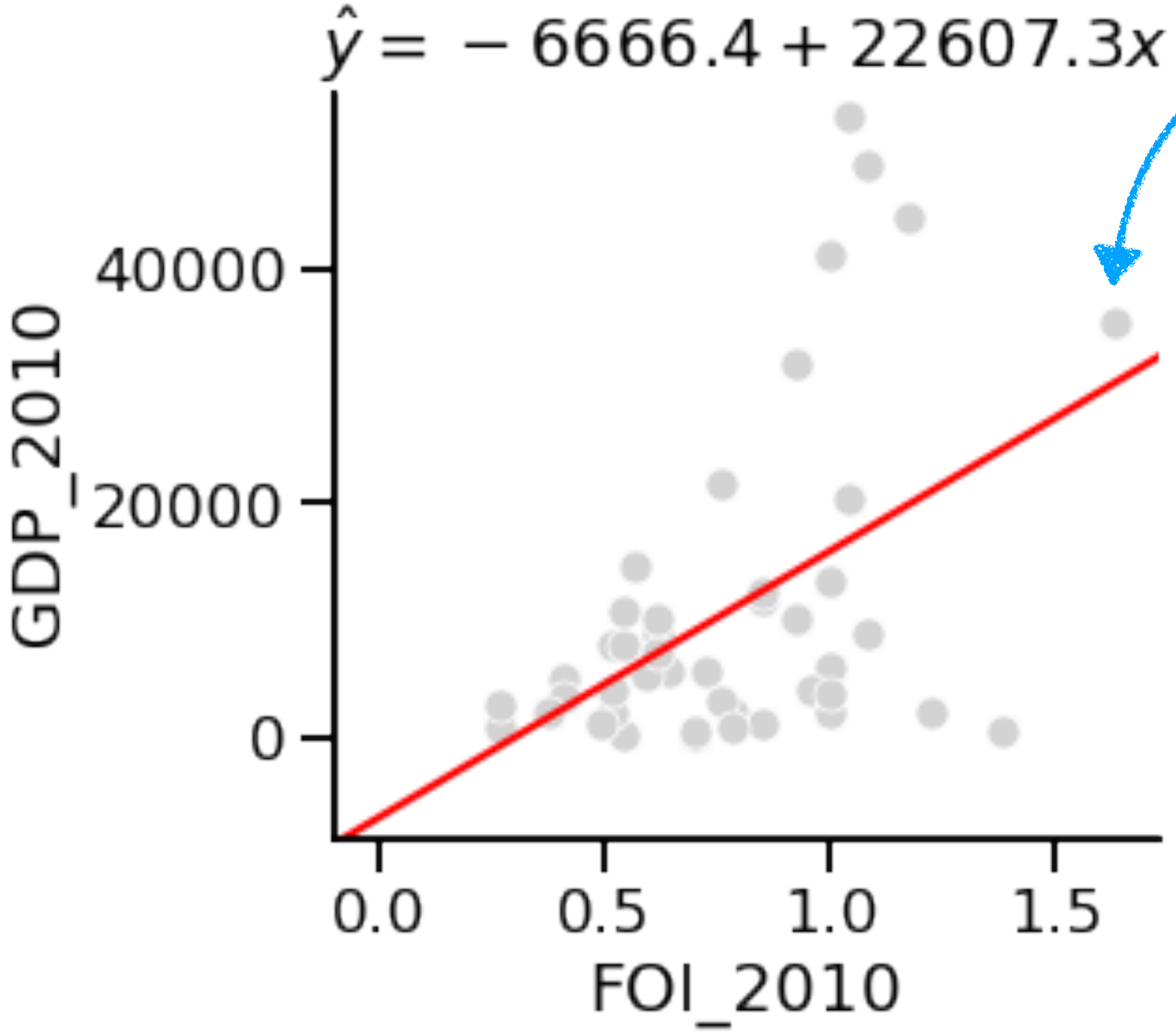
- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$



Regression residuals

Error term: ϵ

- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$



New Zealand

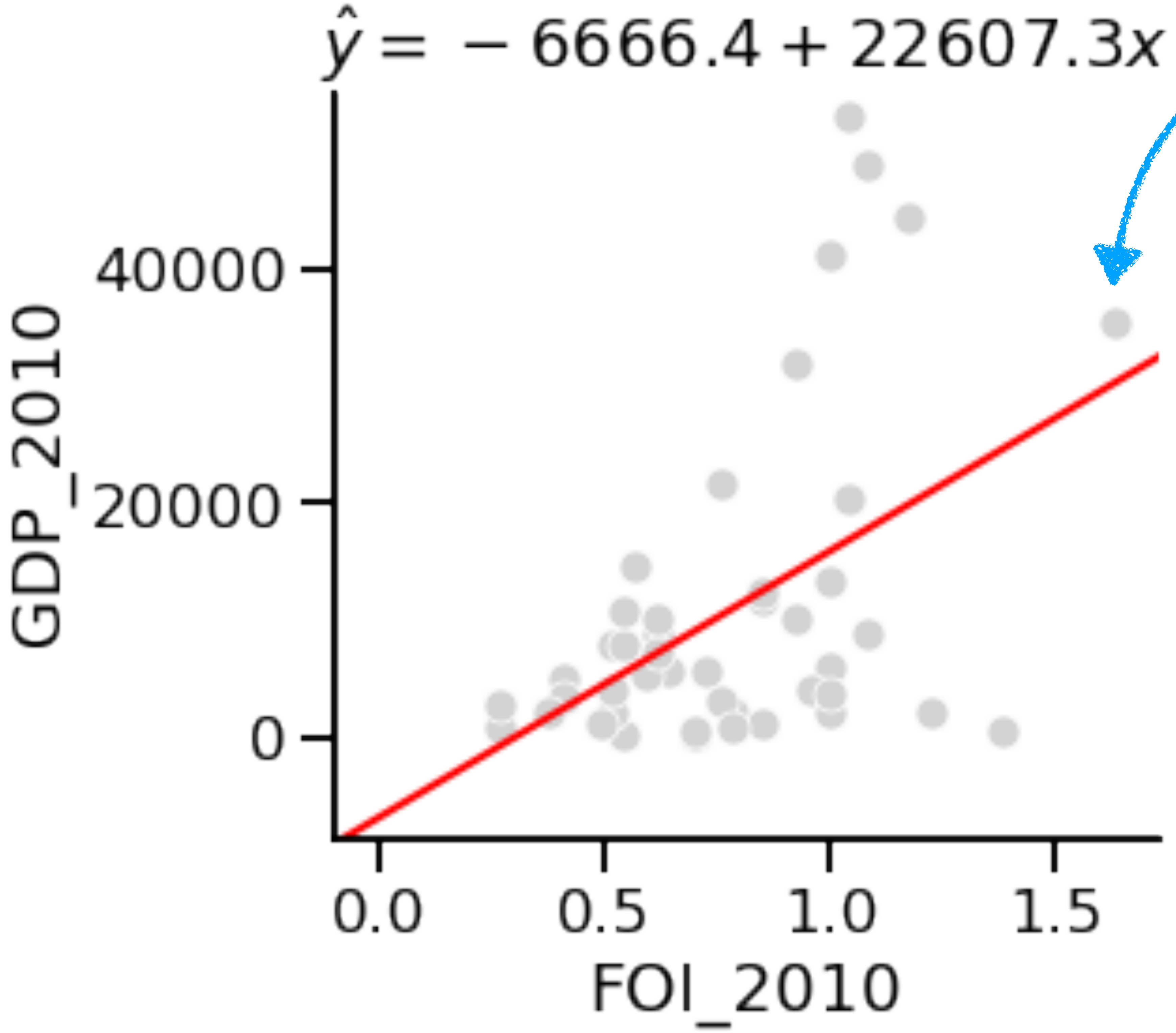
$x = 1{,}631579$

$y = 35.374{,}93$

Regression residuals

Error term: ϵ

- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$



New Zealand

$x = 1{,}631579$

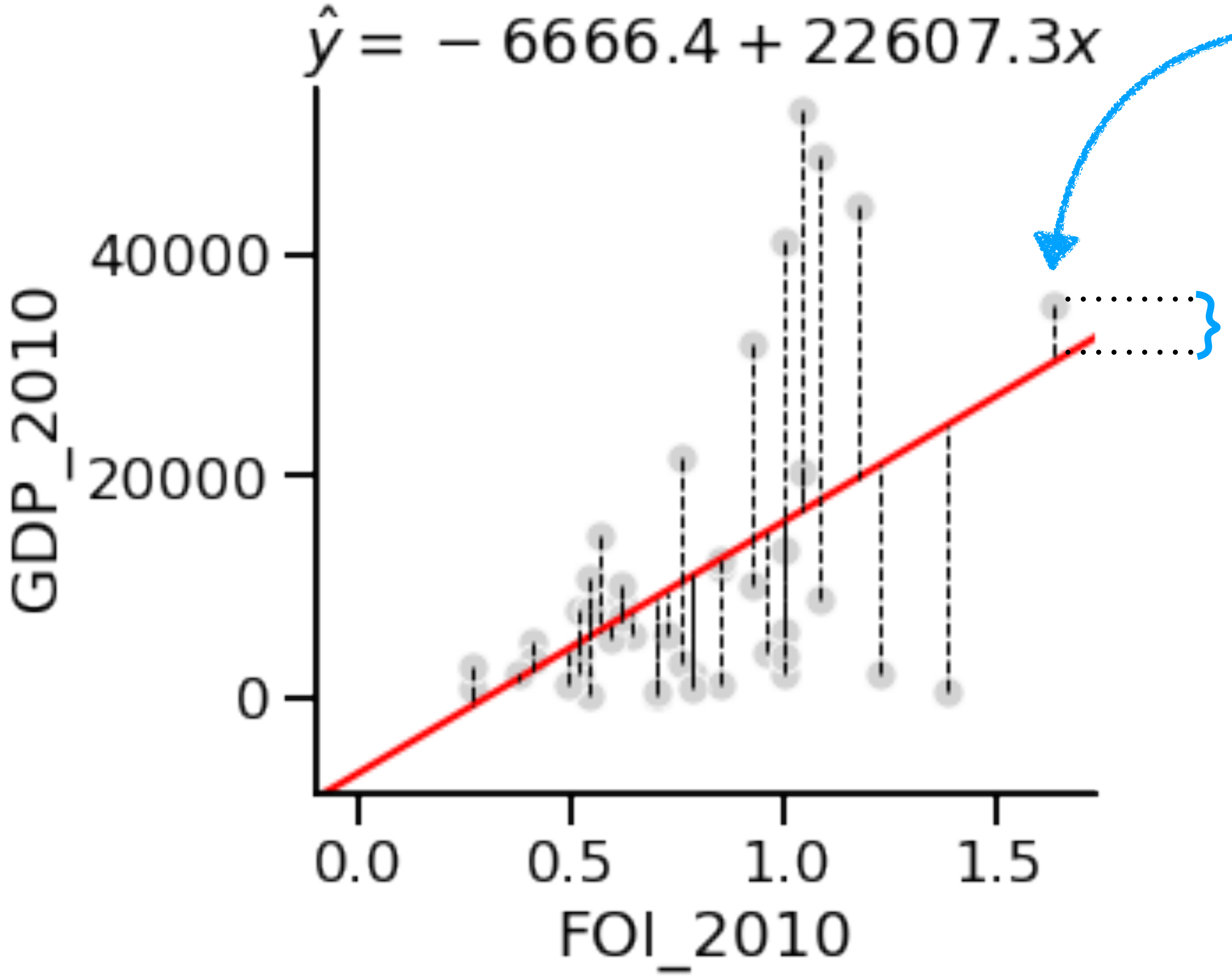
$y = 35.374{,}93$

$\hat{y} = 30.219{,}20$

Regression residuals

Error term: ϵ

- Residuals (ϵ) are the differences in between the empirical values y_i and their fitted values $\hat{y}_i$



New Zealand

$$x = 1{,}631579$$

$$y = 35.374{,}93$$

$$\hat{y} = 30.219{,}20$$

$$\epsilon = y - \hat{y} = 5.155{,}73$$

Ordinary Least Squares (OLS)

Model fitting

Ordinary Least Squares (OLS)

Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

Ordinary Least Squares (OLS)

Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.
- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

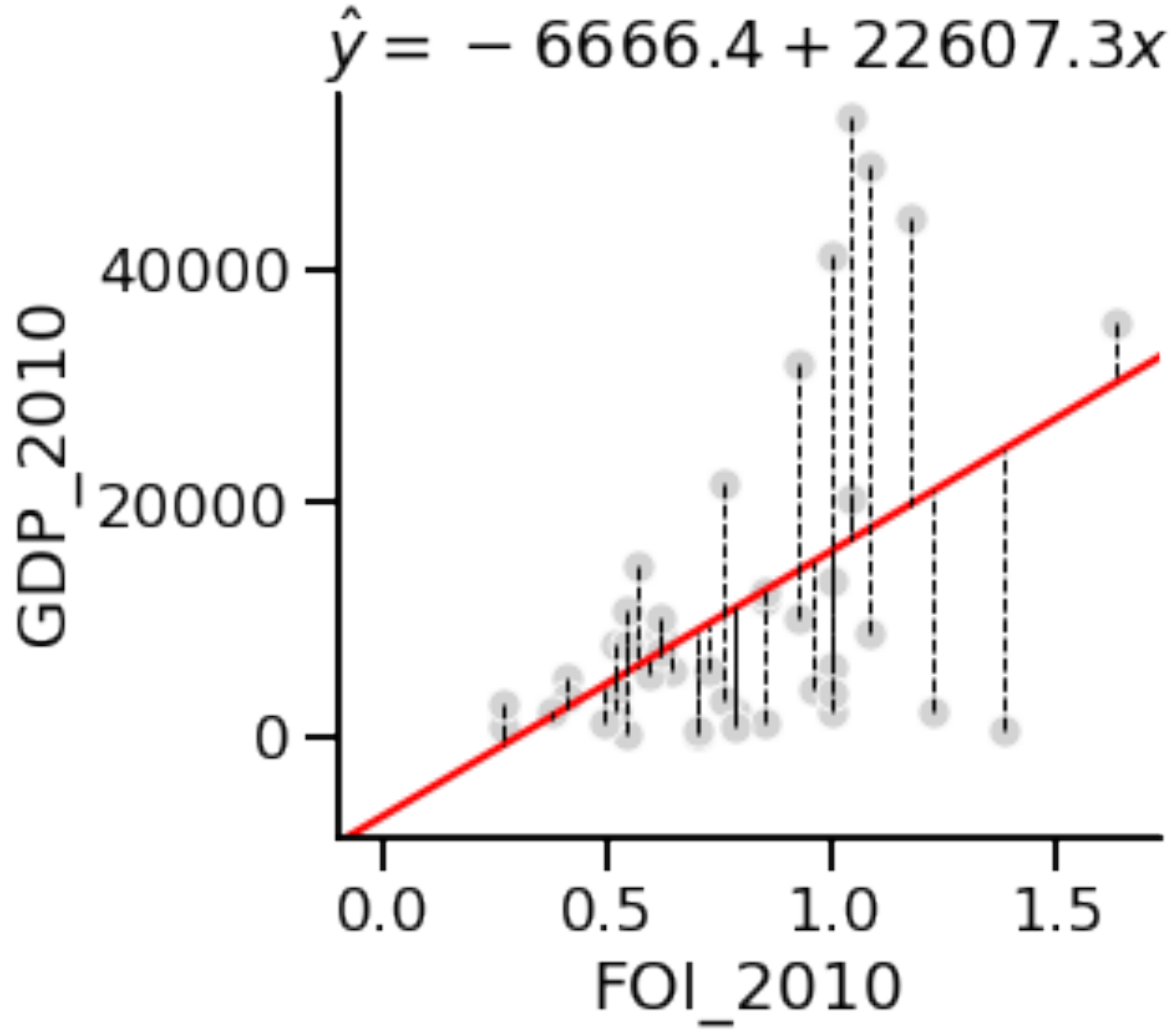
Ordinary Least Squares (OLS)

Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$



Ordinary Least Squares (OLS)

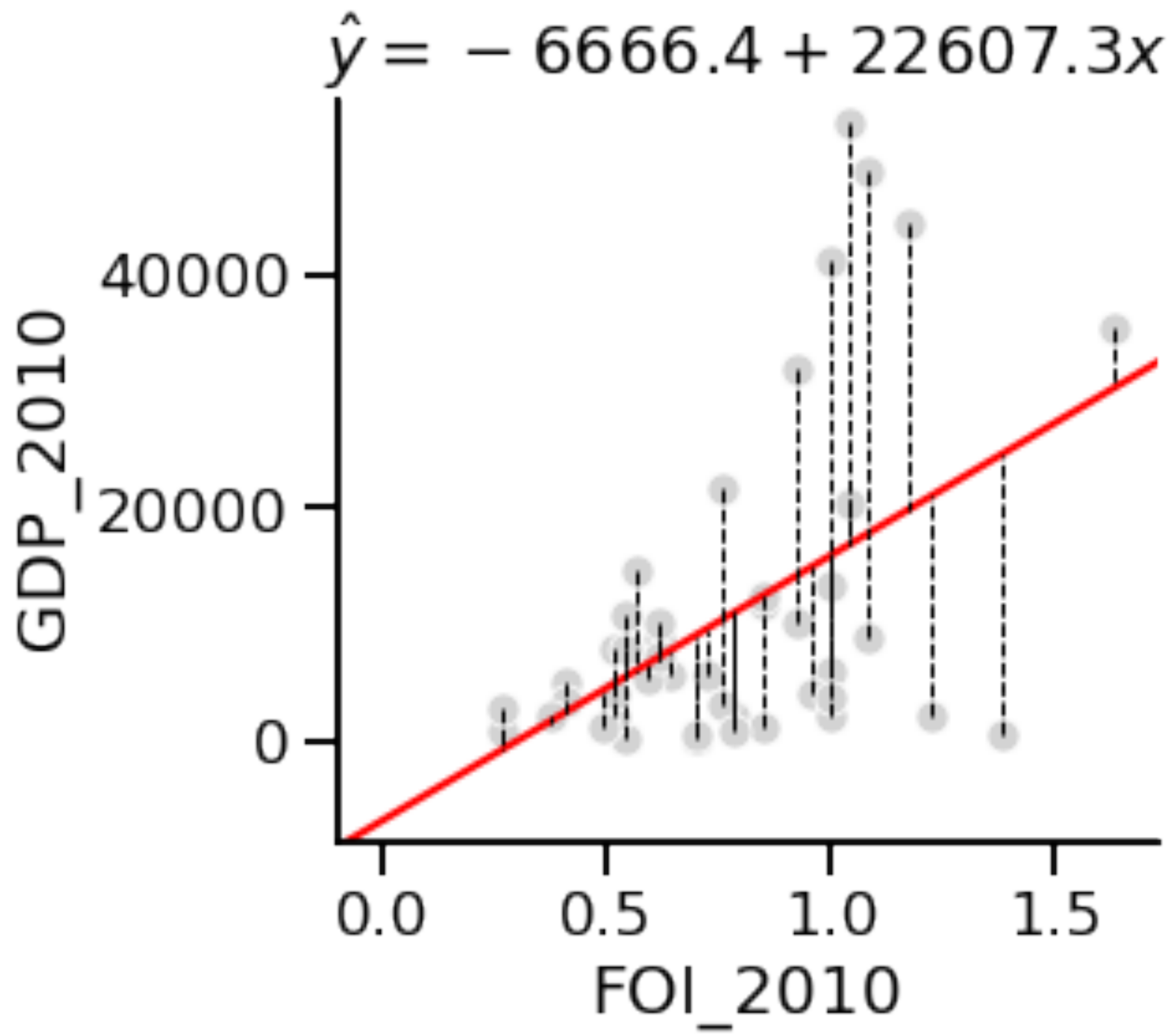
Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

Fitted model



Ordinary Least Squares (OLS)

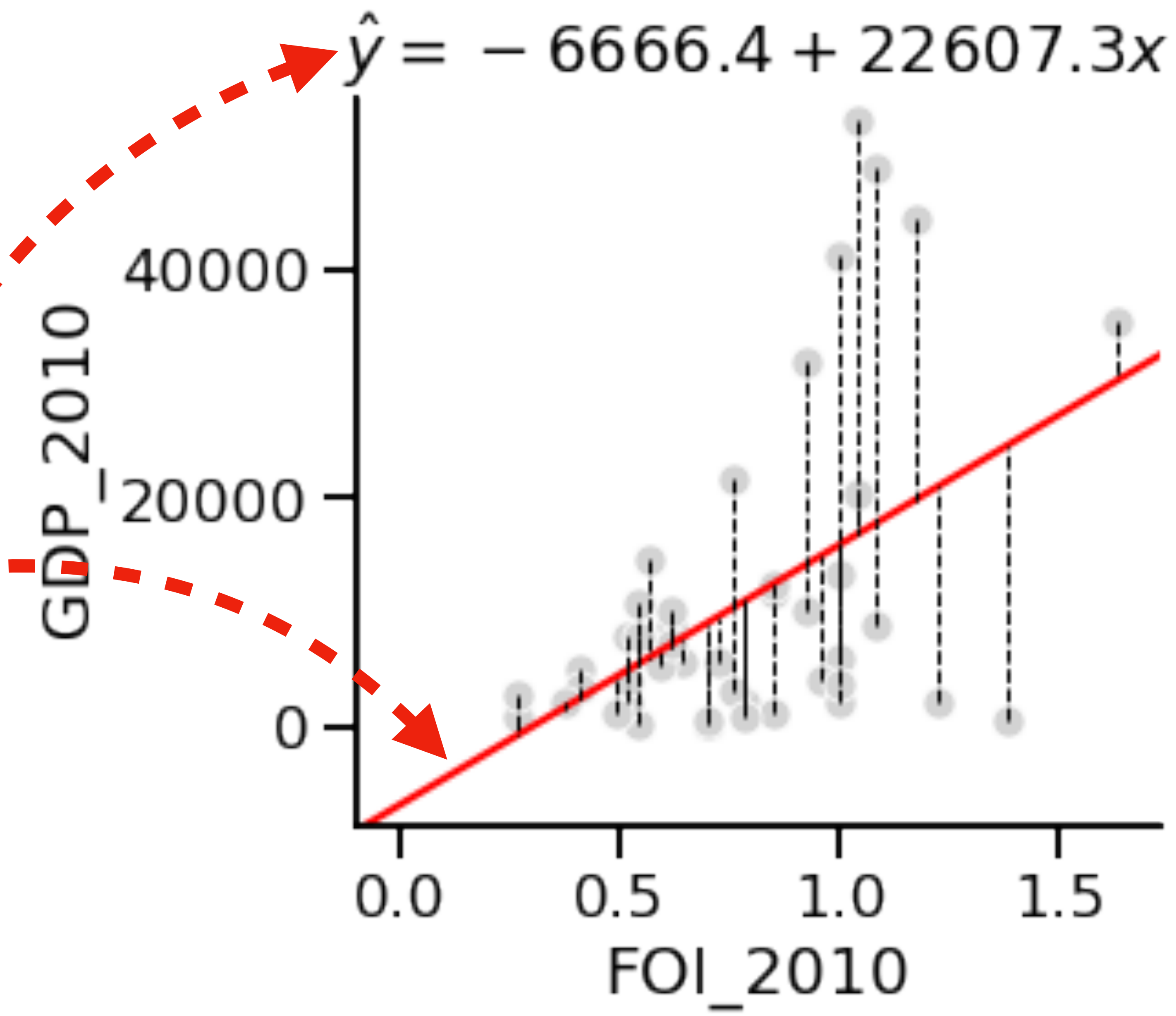
Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

Fitted model



Ordinary Least Squares (OLS)

Model fitting

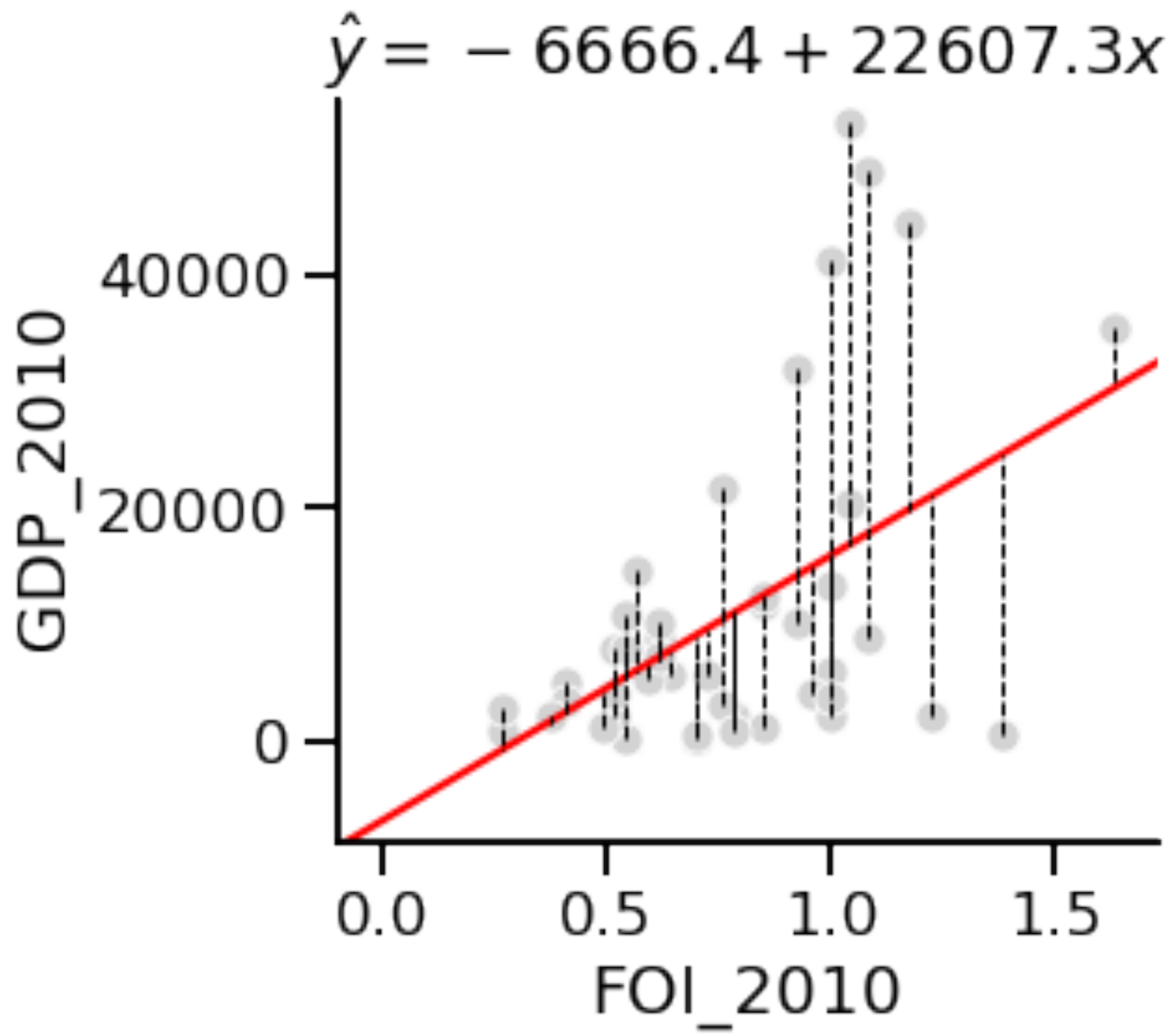
- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

Fitted model

Observation



Ordinary Least Squares (OLS)

Model fitting

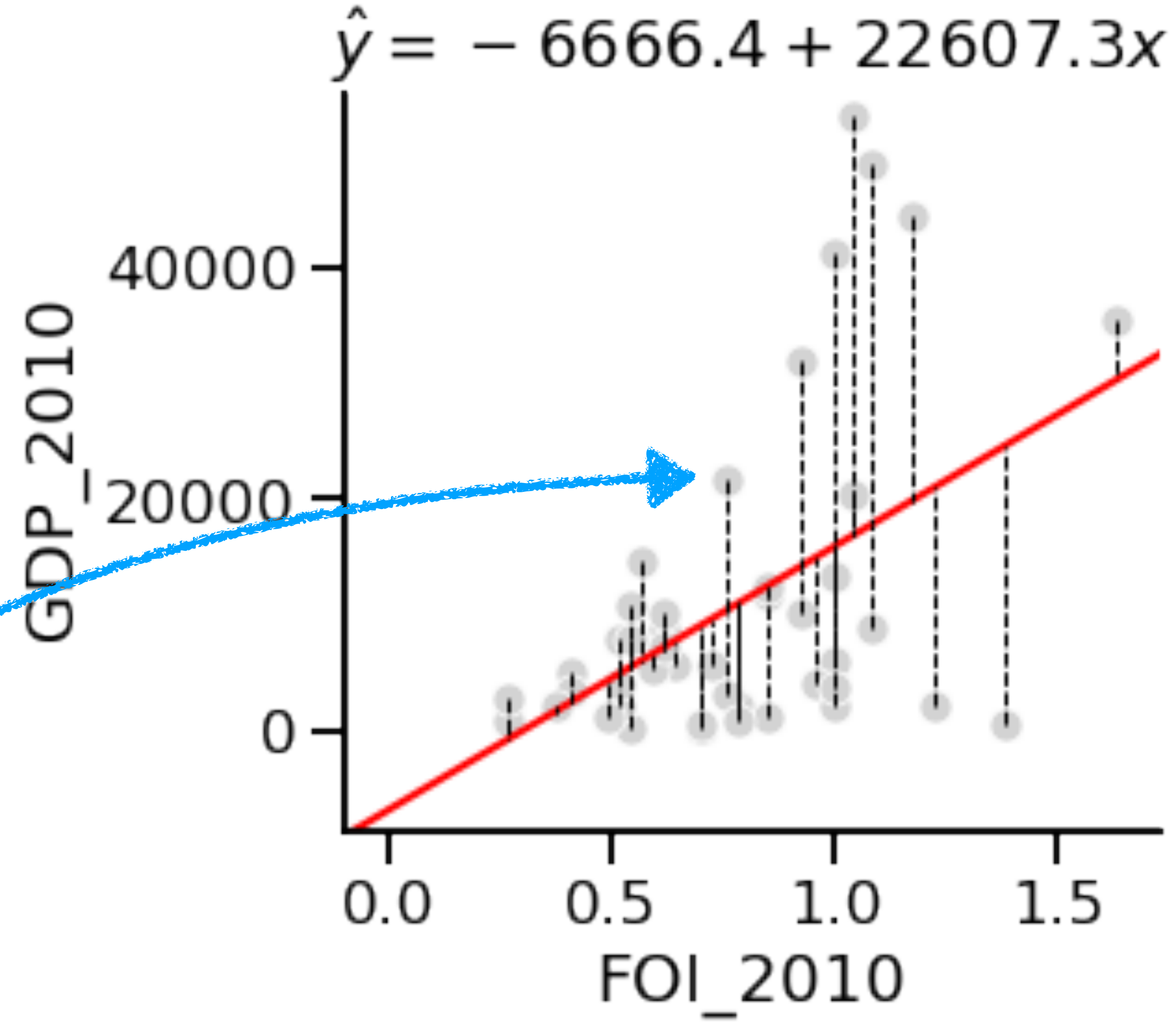
- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

Fitted model

Observation



Ordinary Least Squares (OLS)

Model fitting

- Fitting a regression model is the task of finding the values of the coefficients $(a, \beta_1, \beta_2, \ldots, \beta_k)$ in a way that minimizes the sum of residuals of the model.

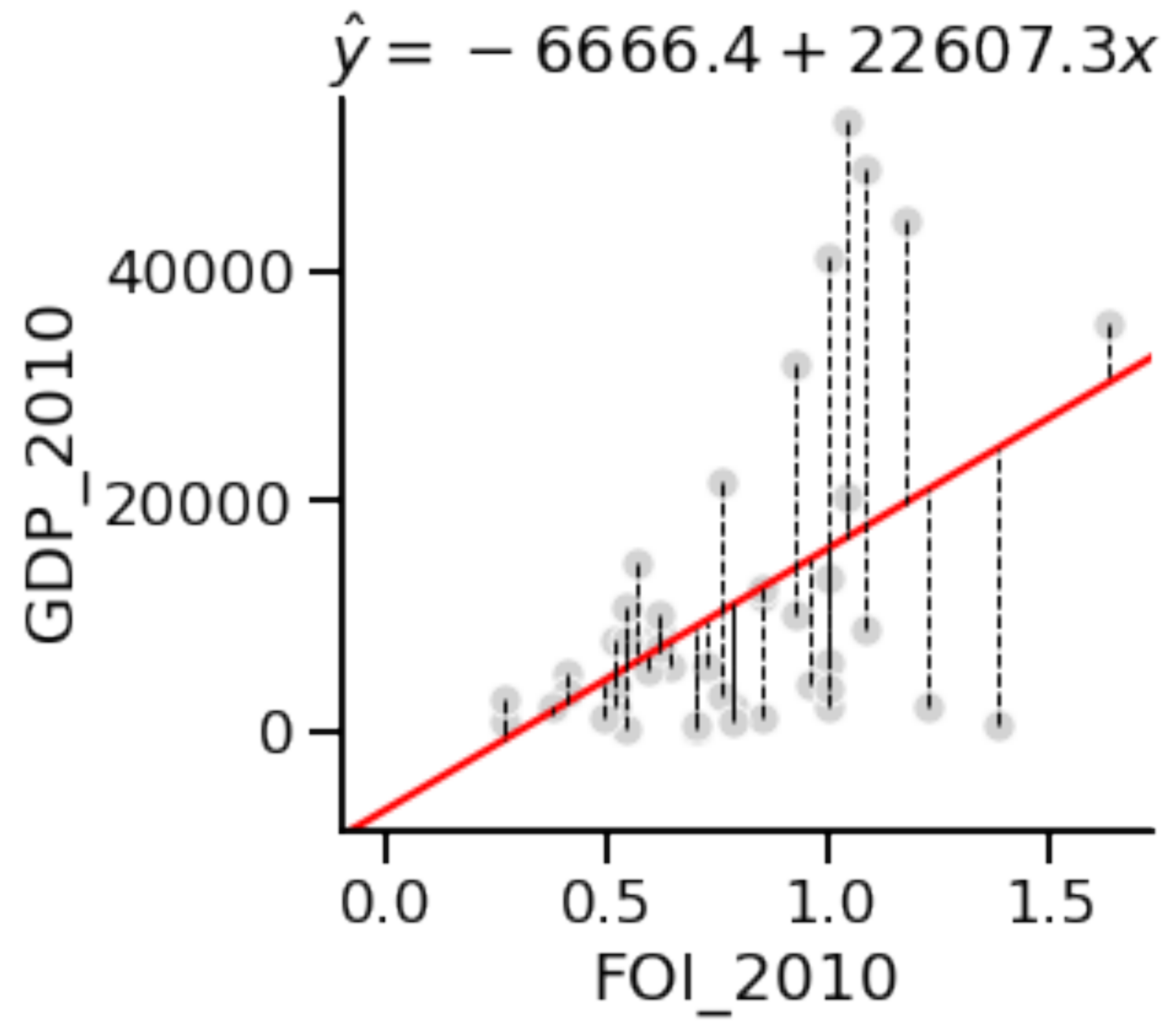
- One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

Fitted model

Observation

- The Ordinary Least Squares method (OLS) looks for the values of coefficients that minimize the RSS. This way, you can think about the OLS result as the line that minimizes the sum of squared lengths of the vertical lines in the figure.



Goodness of fit

Model fitting

- A way to measure the quality of a model fit is to calculate the proportion of variance of the dependent variable ($V[Y]$) that is explained by the model.
- We can do this by comparing the variance of residuals ($V[\epsilon]$) to the variance of Y.
- This is captured by the coefficient of determination, also known as R^2:

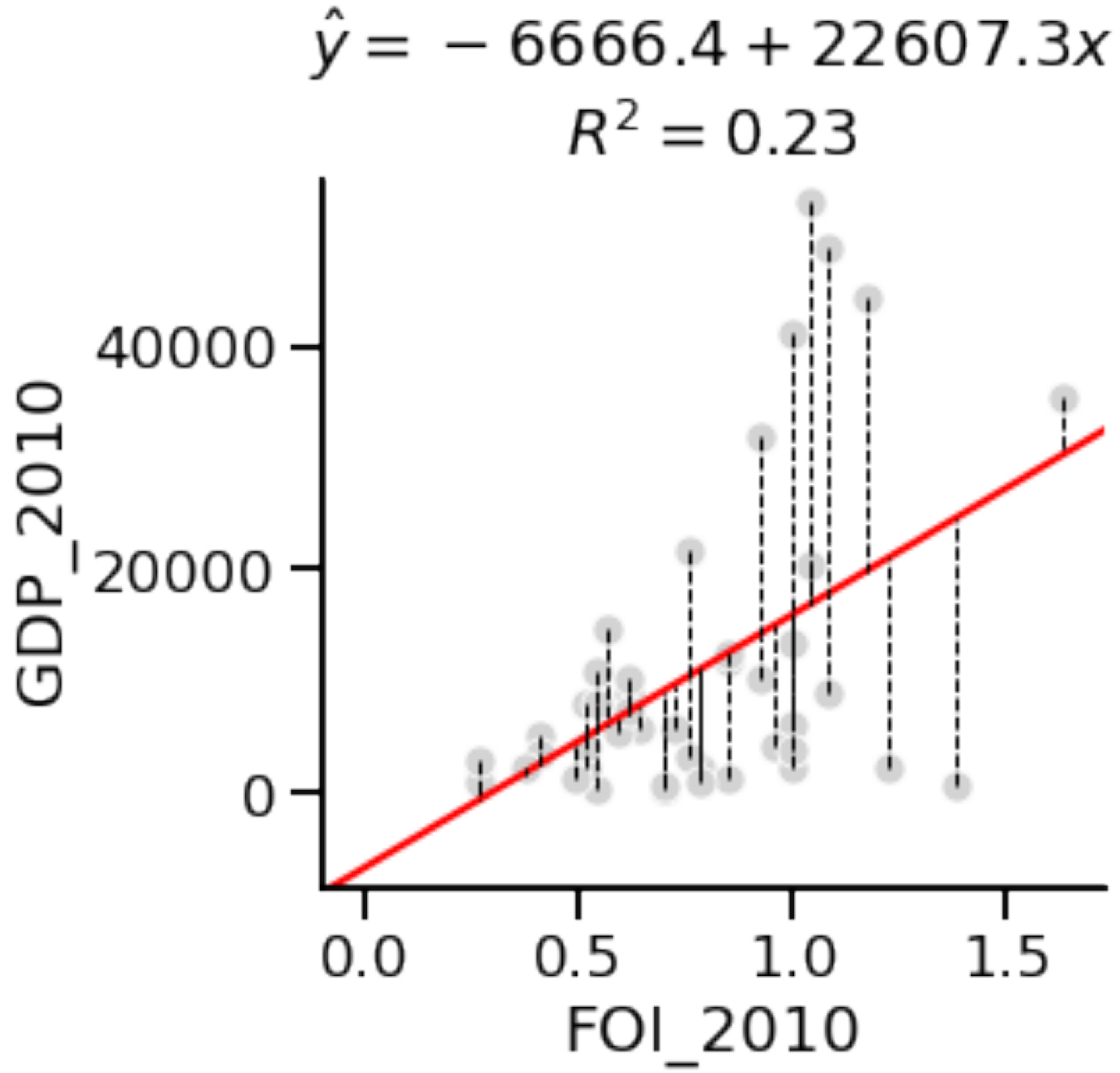
$$R^2 = 1 - \frac{V[\epsilon]}{V[Y]} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Goodness of fit

Model fitting

- A way to measure the quality of a model fit is to calculate the proportion of variance of the dependent variable ($V[Y]$) that is explained by the model.
- We can do this by comparing the variance of residuals ($V[\epsilon]$) to the variance of Y.
- This is captured by the coefficient of determination, also known as R^2:

$$R^2 = 1 - \frac{V[\epsilon]}{V[Y]} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$



Other metrics of “model success”

Model fitting

- Adjusted R squared: $R^2_{adj} = 1 - \left[\frac{(1-R^2)(N-1)}{(N-k-1)}\right]$
- Mean average error: $MAE = \frac{1}{N}\sum_i y_i - \hat{y}_i$
- Mean square error: $MSE = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$
- Root mean square error: $RMSE = \sqrt{\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2}$
- Mean absolute percentage error: $MAPE = \frac{1}{N}\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$

In the formulas N is the number of observations and k is the number of the independent variables in the data.

Other metrics of “model success”

Model fitting

- Adjusted R squared: $R^2_{adj} = 1 - \left[\frac{(1 - R^2)(N - 1)}{(N - k - 1)}\right]$
- Mean average error: $MAE = \frac{1}{N}\sum_i y_i - \hat{y}_i$
- Mean square error: $MSE = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$
- Root mean square error: $RMSE = \sqrt{\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2}$
- Mean absolute percentage error: $MAPE = \frac{1}{N}\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$

In the formulas N is the number of observations and k is the number of the independent variables in the data.

For comparing the accuracy among different linear regression models, $RMSE$ is a better choice than R^2.

Outline

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

How does social behavior spread in society?

The Simmel effect

Theory of fashion (1895)

The Simmel effect

Theory of fashion (1895)

- The Simmel effect refers to the dynamics of **status symbols** in hierarchically ordered societies.
    - Status symbols are externally displayed traits or cultural features associated with high social class, e.g. surnames, clothing, sport, food, etc.

The Simmel effect

Theory of fashion (1895)

- The Simmel effect refers to the dynamics of **status symbols** in hierarchically ordered societies.
    - Status symbols are externally displayed traits or cultural features associated with high social class, e.g. surnames, clothing, sport, food, etc.
- This phenomenon was highlighted by the sociologist Georg Simmel, when he attempted to explain the rapid **diffusion** and **decline** of **fashion**.
    - Simmel noticed that *fashions come and go*, but fashion is always present. When something becomes popular, it is bound to lose its popularity.

The Simmel effect

Theory of fashion (1895)

- The Simmel effect refers to the dynamics of **status symbols** in hierarchically ordered societies.
  - Status symbols are externally displayed traits or cultural features associated with high social class, e.g. surnames, clothing, sport, food, etc.
- This phenomenon was highlighted by the sociologist Georg Simmel, when he attempted to explain the rapid **diffusion** and **decline** of **fashion**.
  - Simmel noticed that *fashions come and go*, but fashion is always present. When something becomes popular, it is bound to lose its popularity.
- Simmel hypothesized that the *instability* of fashion results from the combined action of **imitation** and **distinction**.
  - Simmel hypothesized that status symbols spread through the population downwards, from the highest to the lowest status. As they spread, old symbols are replaced with new ones. Thereby, social differentiation persists under the instability of status symbols.

The mechanisms of Simmel's theory

Imitation and distinctiveness

The mechanisms of Simmel's theory

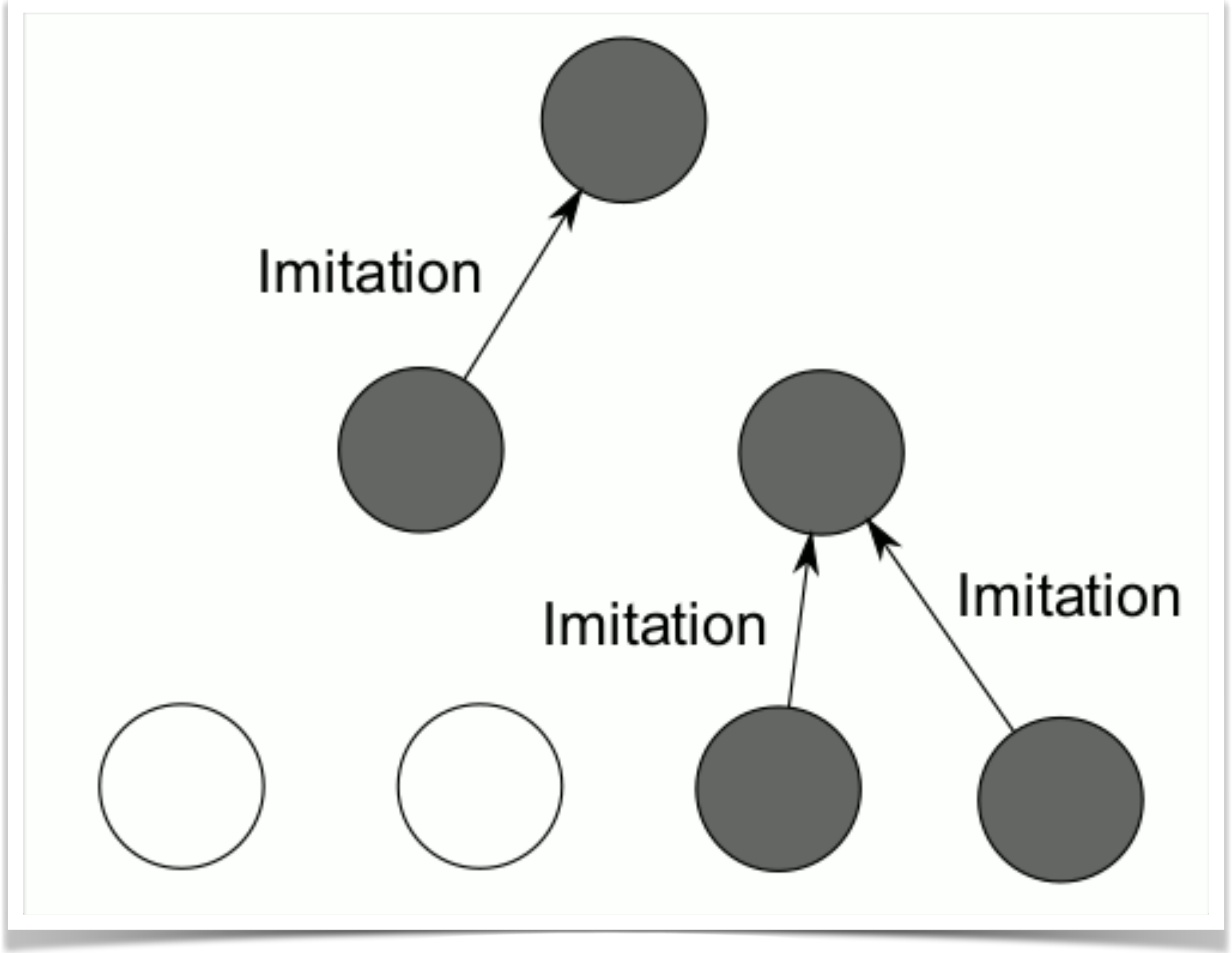
Imitation and distinctiveness

- On one hand, each of us has tendency to **imitate others**. On the other, we also have a tendency to **distinguish ourselves from others**.
  - Fashion's flux needs both of these contradictory tendencies in order to work.

The mechanisms of Simmel’s theory

Imitation and distinctiveness

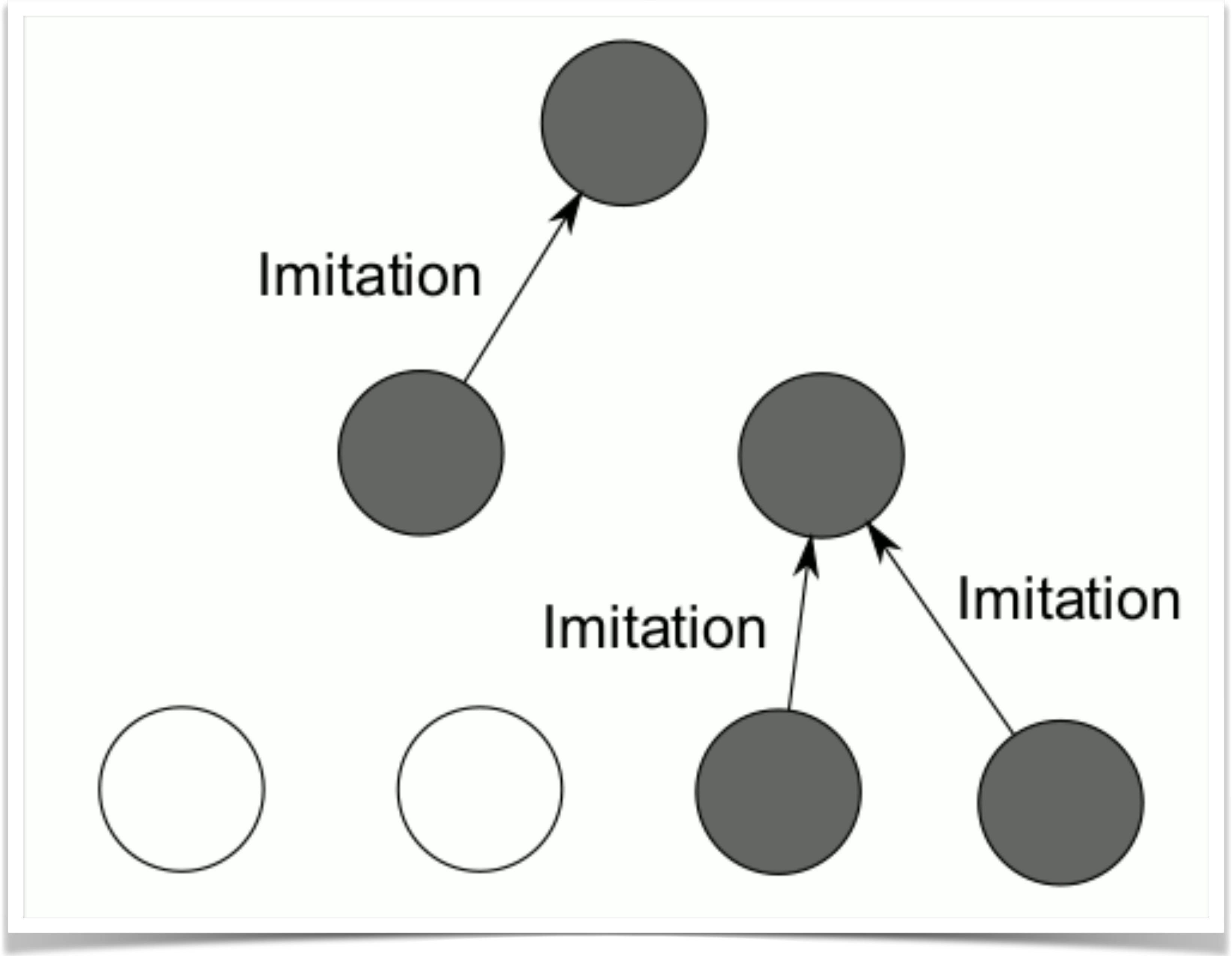
- On one hand, each of us has tendency to **imitate others**. On the other, we also have a tendency to **distinguish ourselves from others**.
  - Fashion's flux needs both of these contradictory tendencies in order to work.



The mechanisms of Simmel’s theory

Imitation and distinctiveness

- On one hand, each of us has tendency to **imitate others.** On the other, we also have a tendency to **distinguish ourselves from others.**
  - Fashion's flux needs both of these contradictory tendencies in order to work.



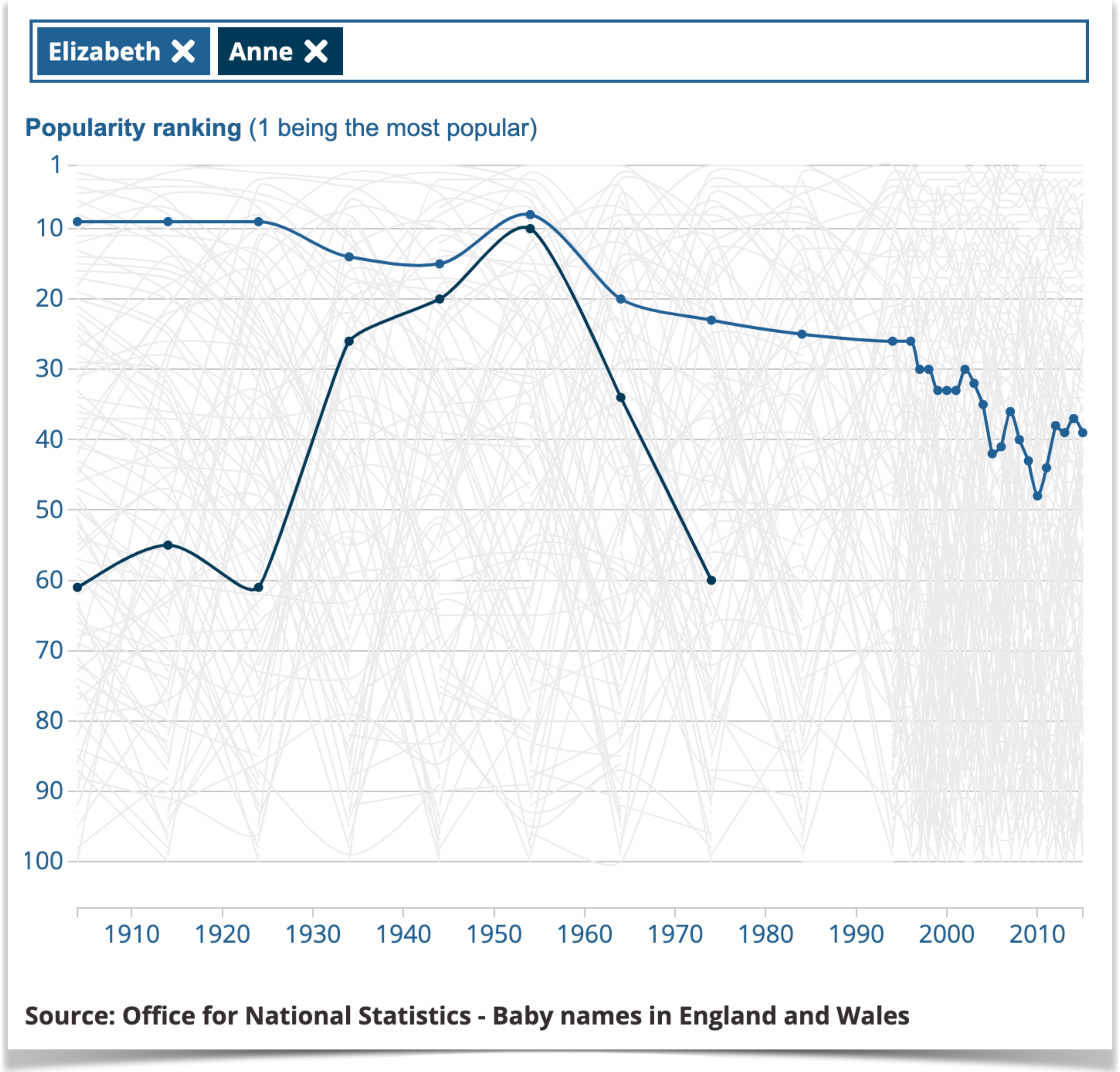
- Jeans were invented in 1871
  - Originally created for miners.
  - Popularized in 1950 in films.
  - 1960 widespread adoption.
- White sneakers were invented in 1916
  - Made popular by Adidas in 1970?
- Music challenge videos on TikTok

The case of baby names

- First names can be status symbols and carry subjective and social values.
- Copying the name of your baby from someone else is an example of **imitation**.

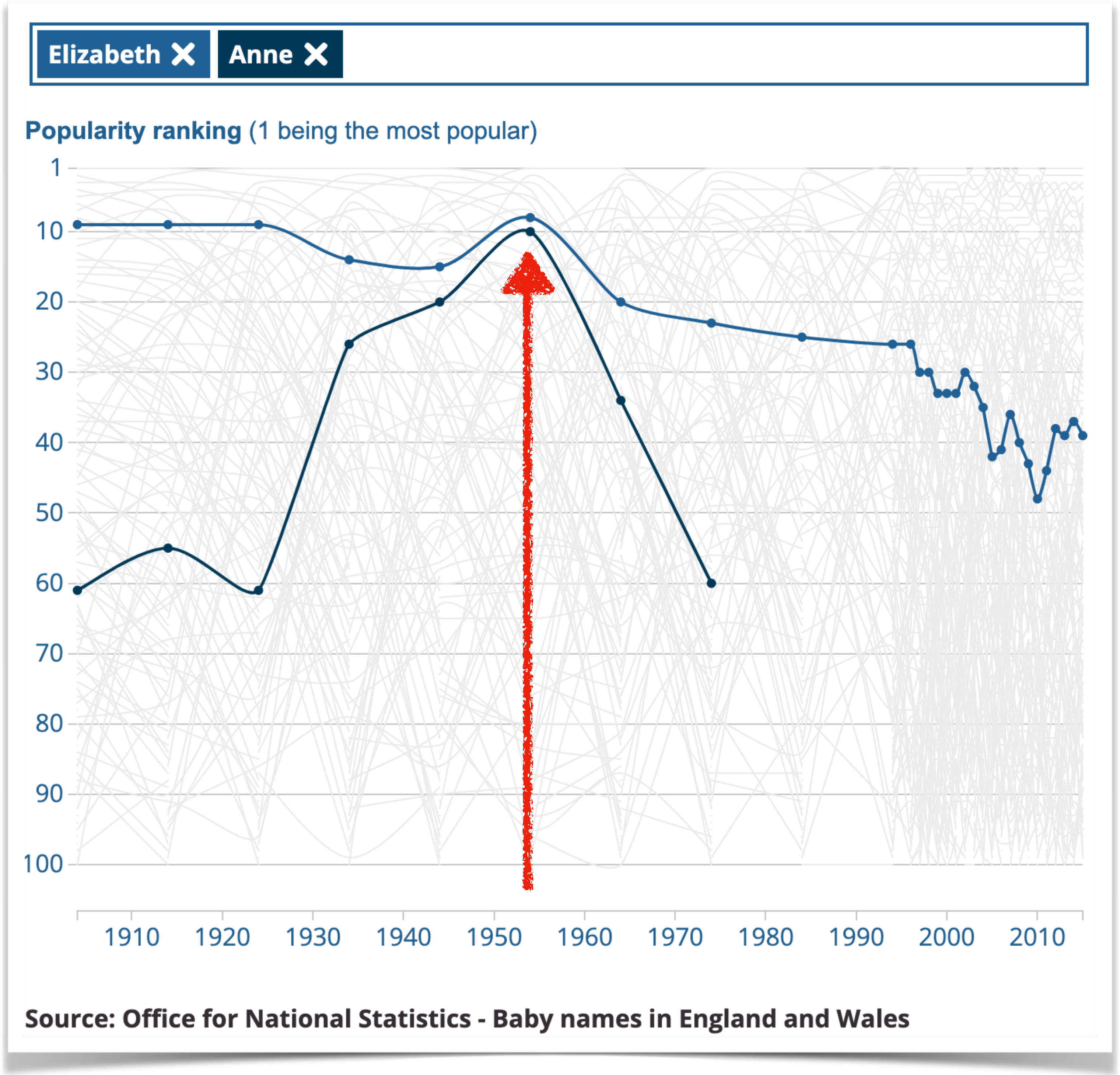
The case of baby names

- First names can be status symbols and carry subjective and social values.
- Copying the name of your baby from someone else is an example of **imitation**.



The case of baby names

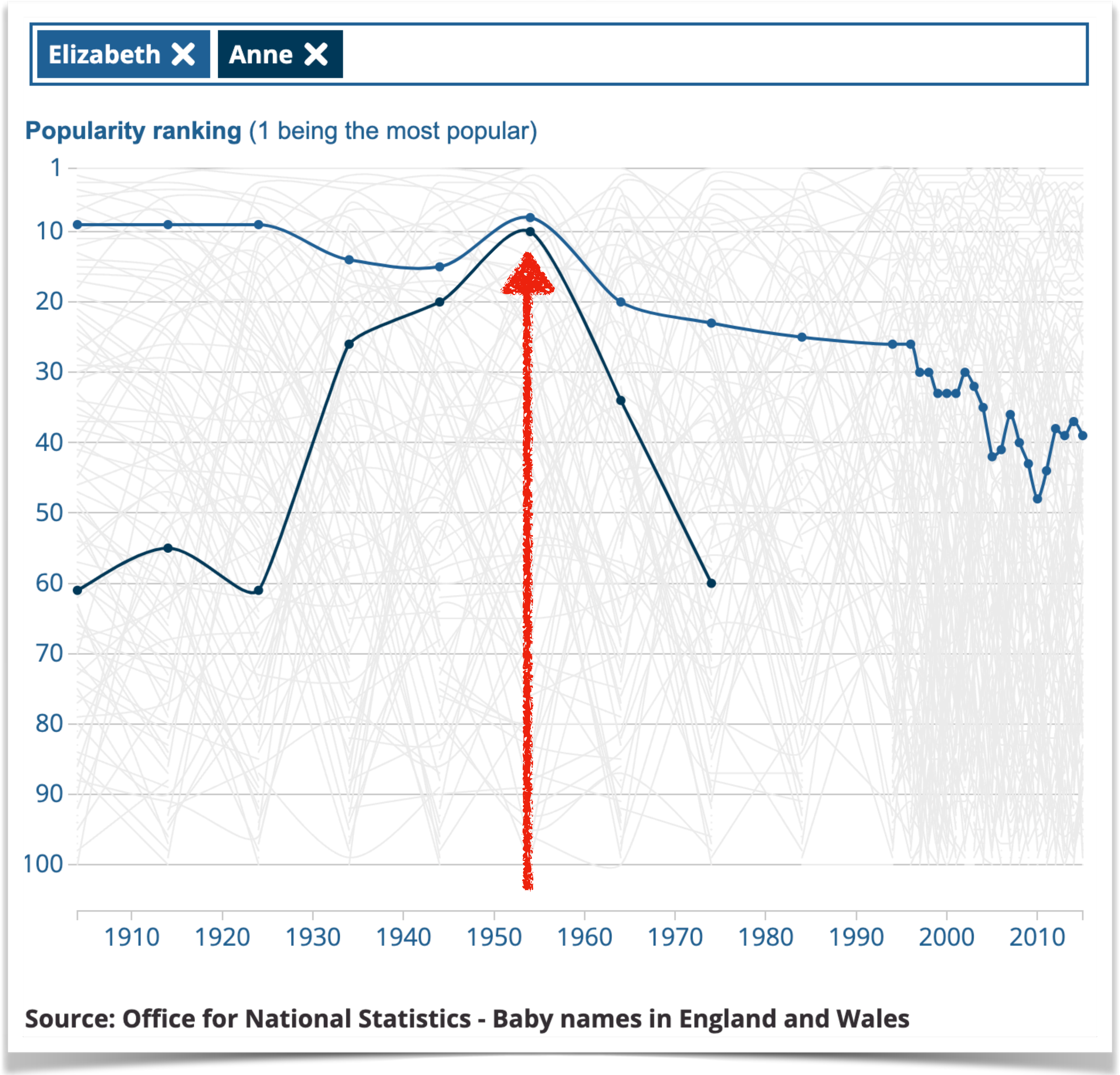
- First names can be status symbols and carry subjective and social values.
- Copying the name of your baby from someone else is an example of **imitation**.



What happened around 1950 in the UK?

The case of baby names

- First names can be status symbols and carry subjective and social values.
- Copying the name of your baby from someone else is an example of **imitation**.



What happened around 1950 in the UK?

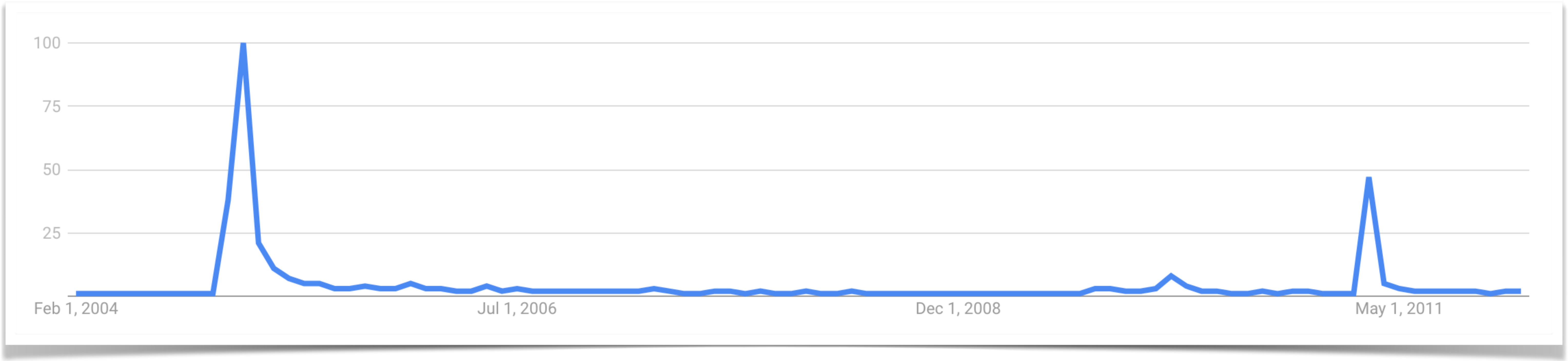
- 1950 Princess Anne was born
- 1953 Queen Elizabeth II was crowned

Social trends in online platforms

Google Search Trends

Social trends in online platforms

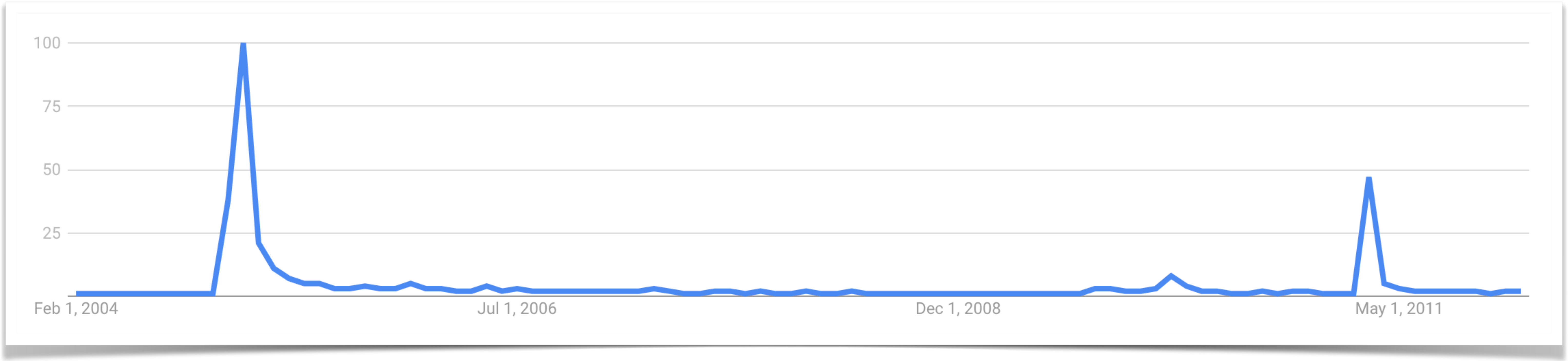
Google Search Trends



“Tsunami”

Social trends in online platforms

Google Search Trends

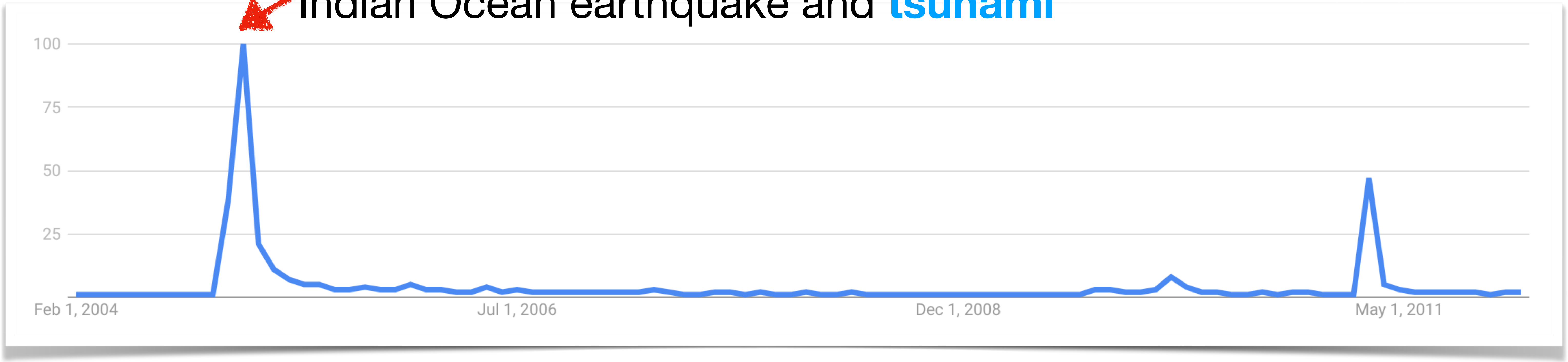


"Tsunami"
An **exogenously** (sudden) triggered search volume

Social trends in online platforms

Google Search Trends

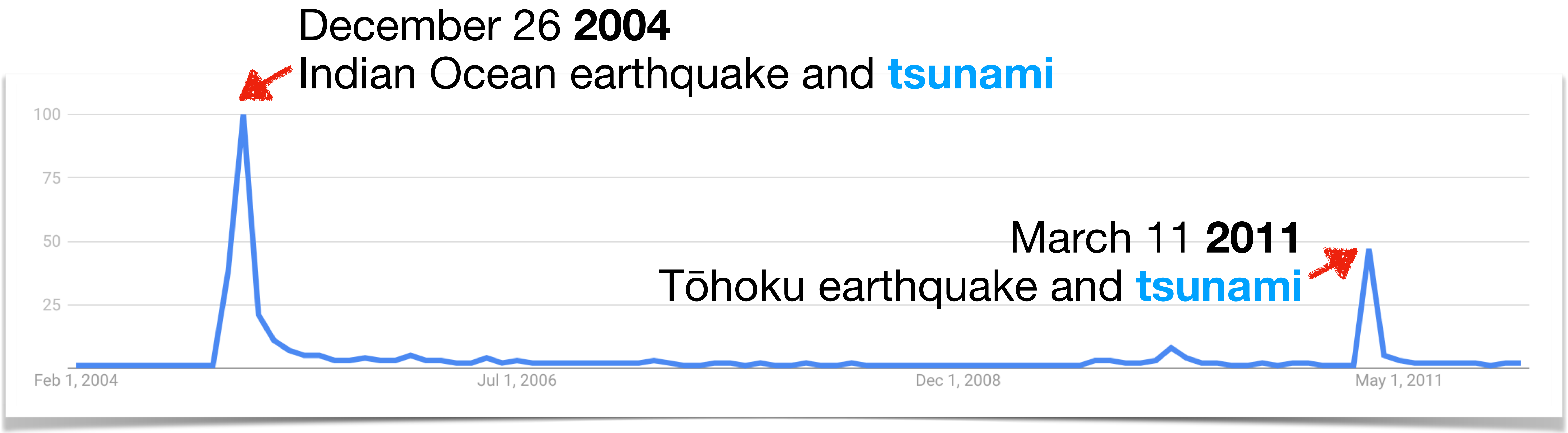
December 26 **2004**
Indian Ocean earthquake and **tsunami**



“Tsunami”
An **exogenously** (sudden) triggered search volume

Social trends in online platforms

Google Search Trends



"Tsunami"
An **exogenously** (sudden) triggered search volume

Social trends in online platforms

Google Search Trends

December 26 **2004**
Indian Ocean earthquake and **tsunami**

March 11 **2011**
Tōhoku earthquake and **tsunami**

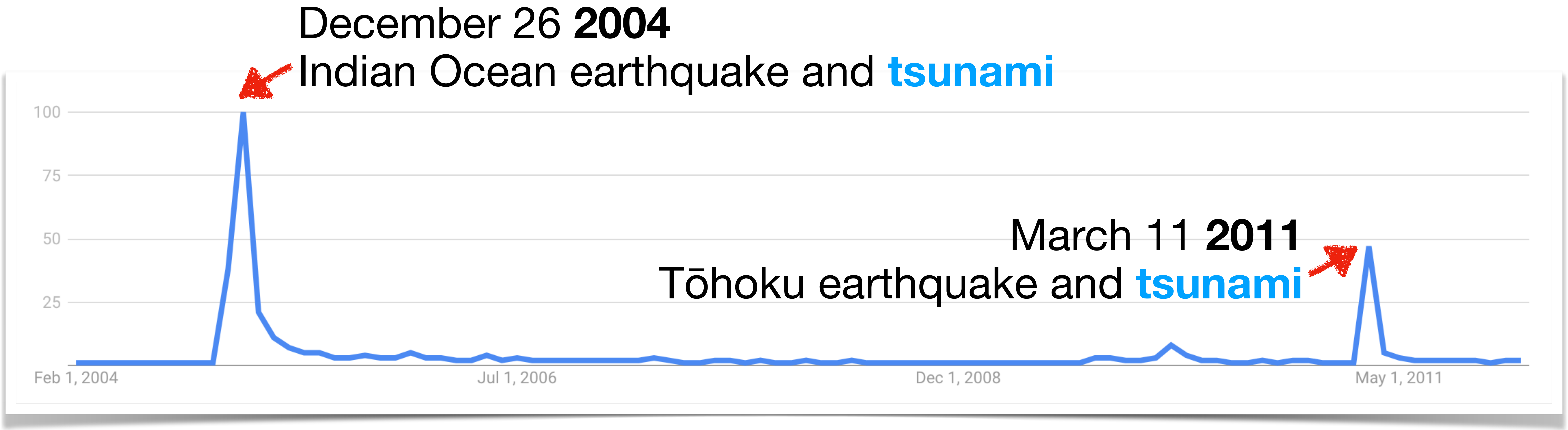
"Tsunami"
An **exogenously** (sudden) triggered search volume

Harry Potter and the Order of the Phoenix was released in summer **2007** worldwide (movie #5).

"Harry Potter"

Social trends in online platforms

Google Search Trends



"Tsunami"
An **exogenously** (sudden) triggered search volume

"Harry Potter"
An **endogenously** (gradual) driven search

...

Harry Potter was first introduced in the novel Harry Potter and the Philosopher's Stone in 1997, and released four movies prior to 2007: in 2001, 2002, 2004, 2005.

The endo-exo model

[Crane and Sornette 2008]

The endo-exo model

[Crane and Sornette 2008]

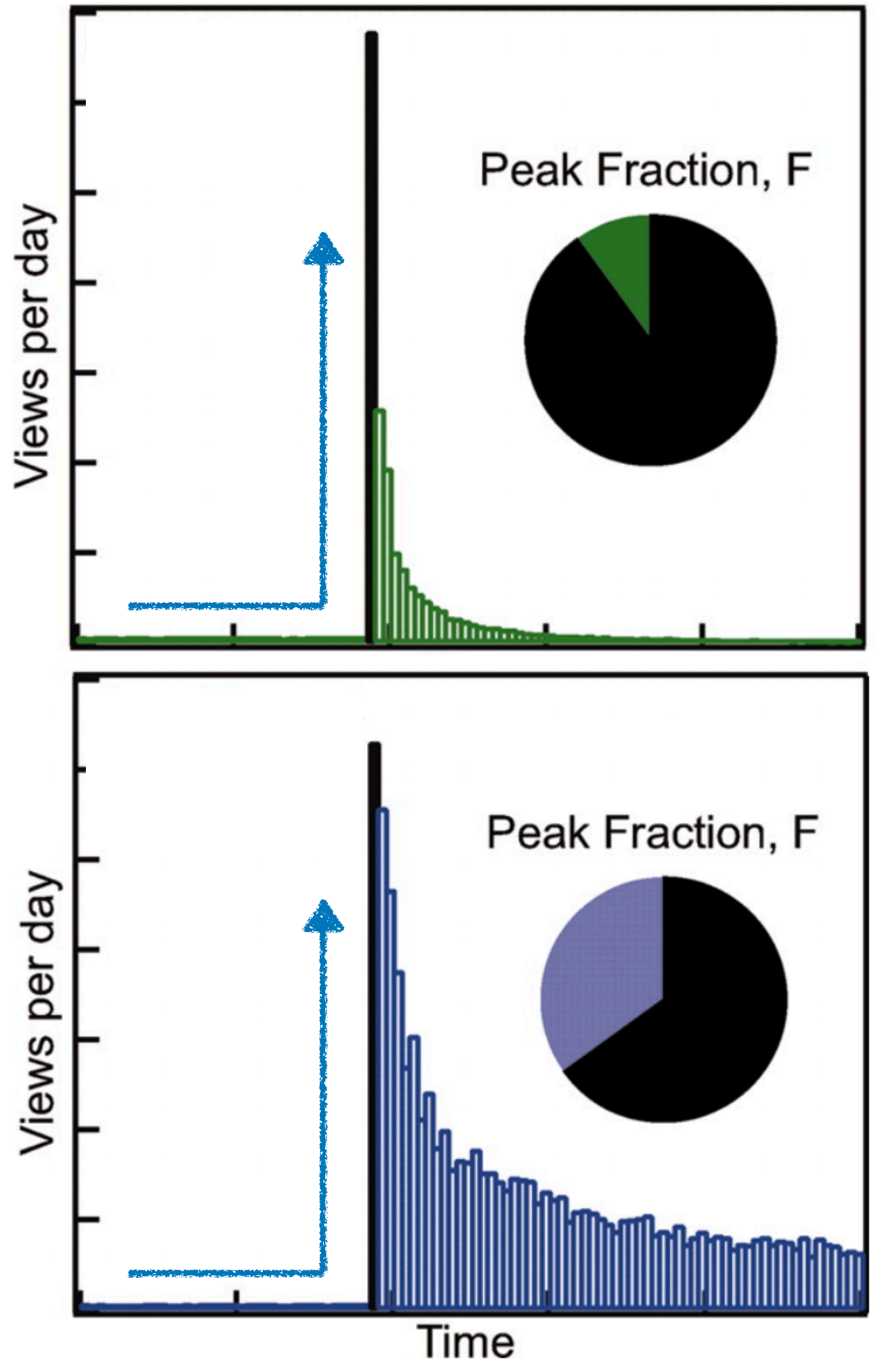
- According to this model, the aggregated social dynamics can be classified by a combination of the type of disturbance (**endo/exo**) and the ability of individuals to influence others to action (**critical/sub-critical**)

The endo-exo model

[Crane and Sornette 2008]

- According to this model, the aggregated social dynamics can be classified by a combination of the type of disturbance (**endo/exo**) and the ability of individuals to influence others to action (**critical/sub-critical**)
    - **Exogenous trigger:** when a central event influences lots of people at the same time, as in the *tsunami* example.

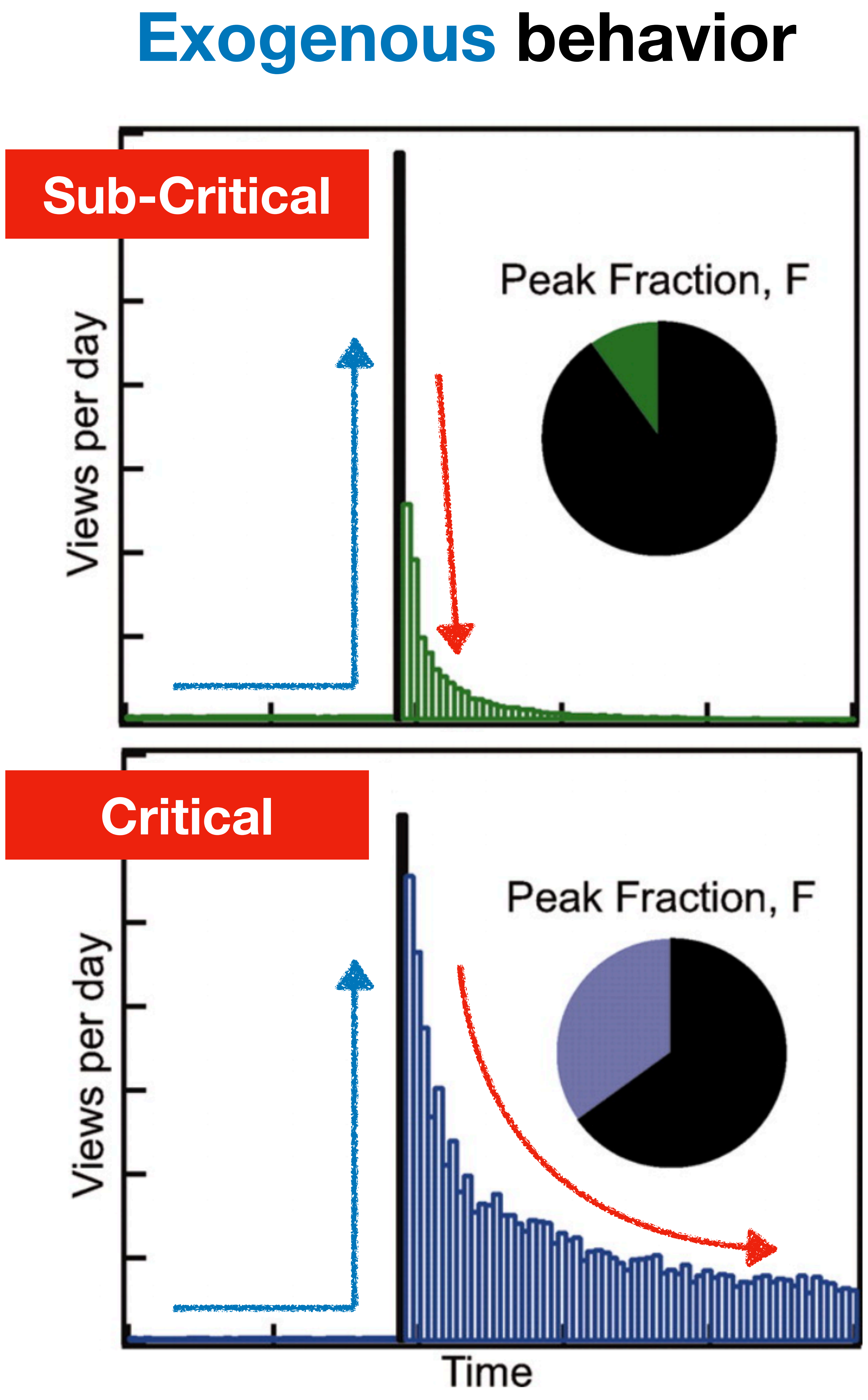
Exogenous behavior



The endo-exo model

[Crane and Sornette 2008]

- According to this model, the aggregated social dynamics can be classified by a combination of the type of disturbance (**endo/exo**) and the ability of individuals to influence others to action (**critical/sub-critical**)
    - **Exogenous trigger:** when a central event influences lots of people at the same time, as in the *tsunami* example.
    - **Critical:** when the social interaction between individuals leads to further responses and it is stronger than the rate of losing interest.



The endo-exo model

[Crane and Sornette 2008]

The endo-exo model

[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

The endo-exo model

[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

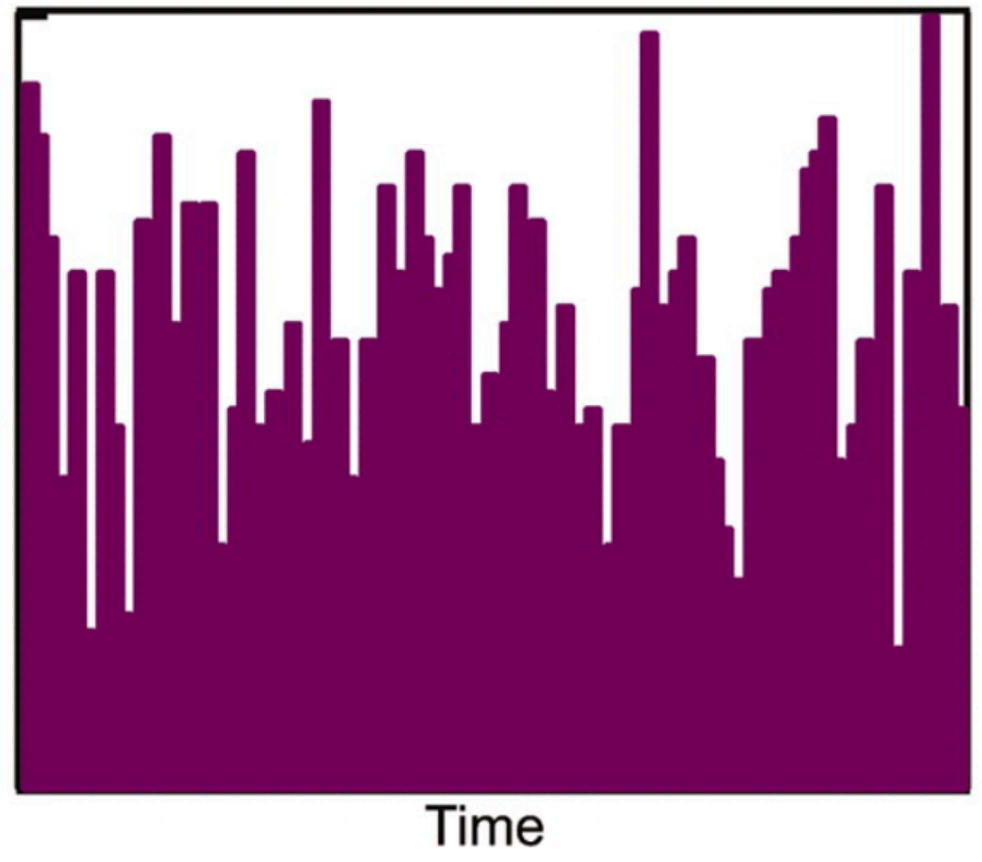
	Exogenous	Endogenous
Sub-Critical		
Critical		

The endo-exo model

[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

  1. **Endogenous sub-critical:** no clear peak, absence of trend.

	Exogenous	Endogenous
Sub-Critical		
Critical		

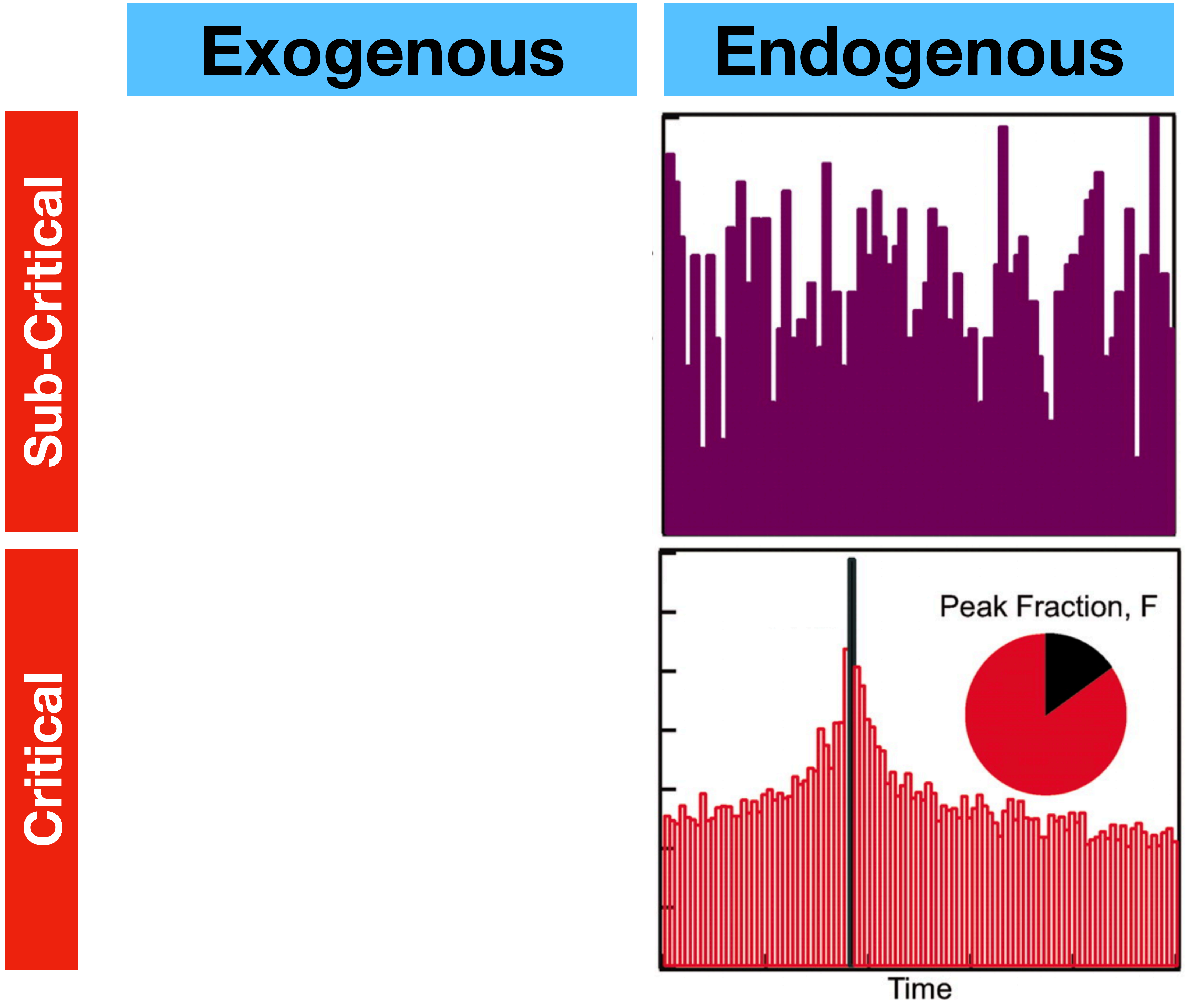
The endo-exo model

[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

  1. **Endogenous sub-critical:** no clear peak, absence of trend.

  2. **Endogenous critical:** "viral" peak driven by word of mouth.



The endo-exo model

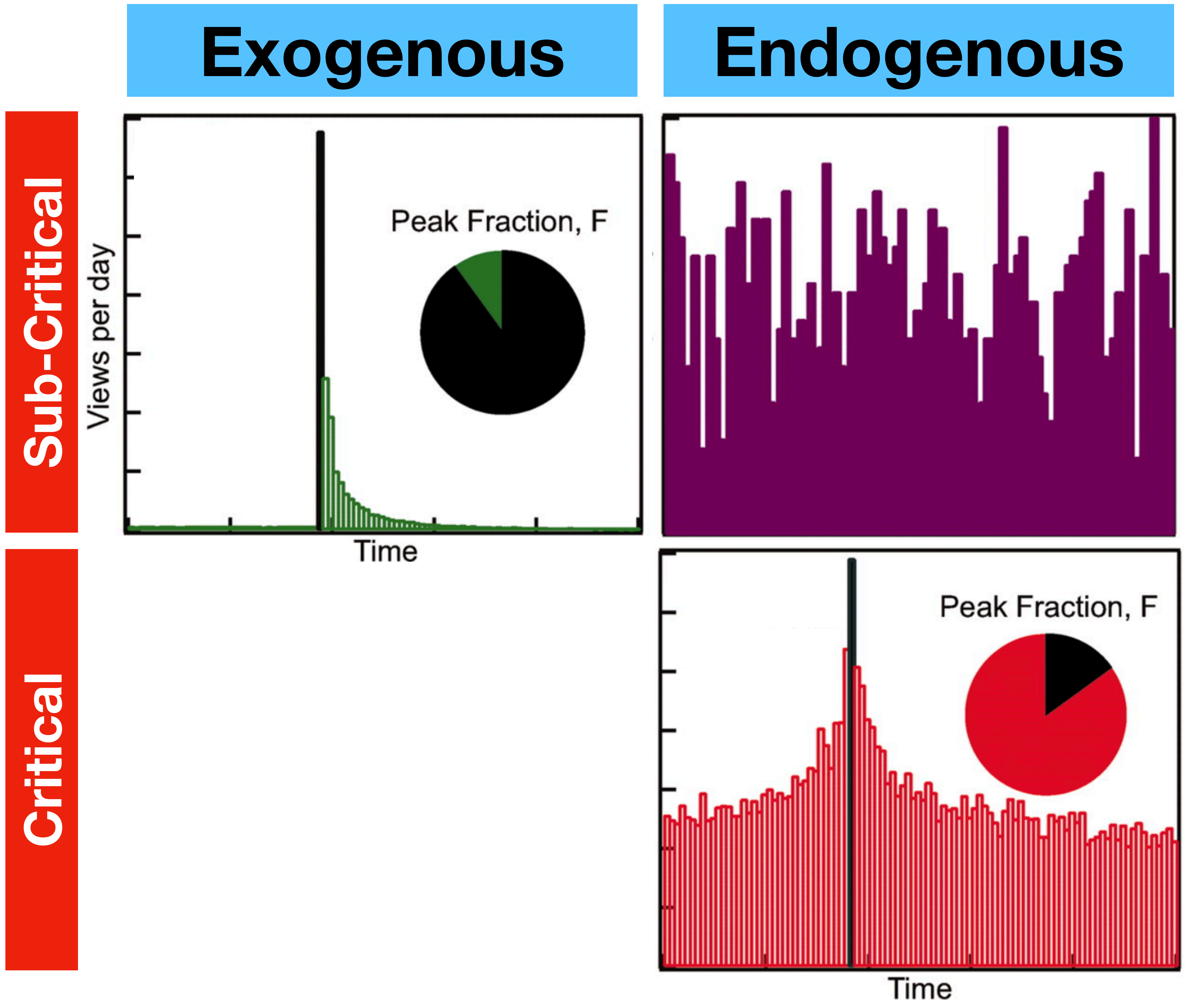
[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

 1. **Endogenous sub-critical:** no clear peak, absence of trend.

 2. **Endogenous critical:** "viral" peak driven by word of mouth.

 3. **Exogenous sub-critical:** sharp peak but fast decay due to lack of strong social interaction.

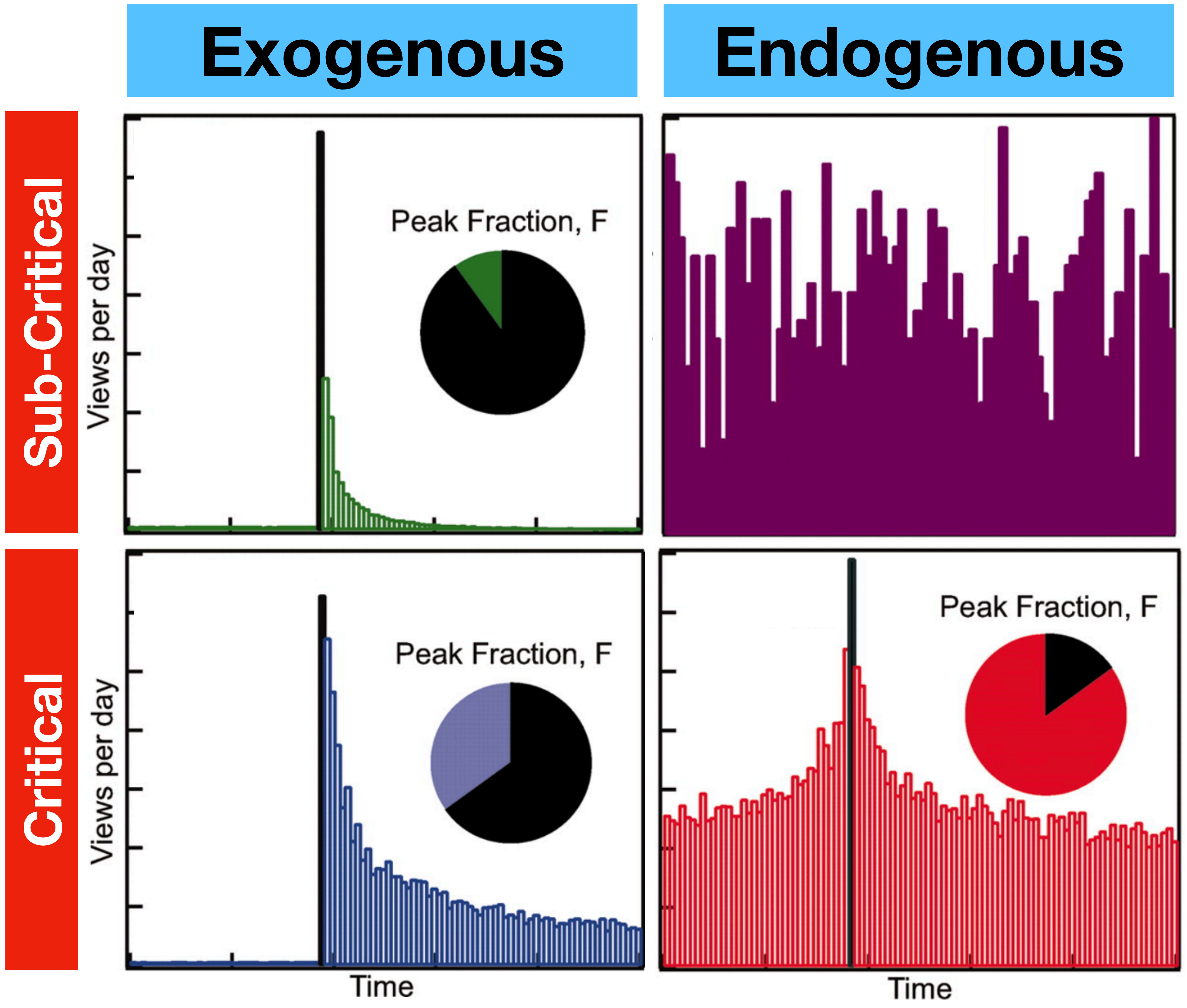


The endo-exo model

[Crane and Sornette 2008]

- These two properties (exogenous and critical) are not exclusive, leading to four types of responses:

  1. **Endogenous sub-critical:** no clear peak, absence of trend.
  2. **Endogenous critical:** "viral" peak driven by word of mouth.
  3. **Exogenous sub-critical:** sharp peak but fast decay due to lack of strong social interaction.
  4. **Exogenous critical:** sharp peak but slow decay due to strong interaction after shock.



Trends on Twitter

#hashtags

Trends on Twitter

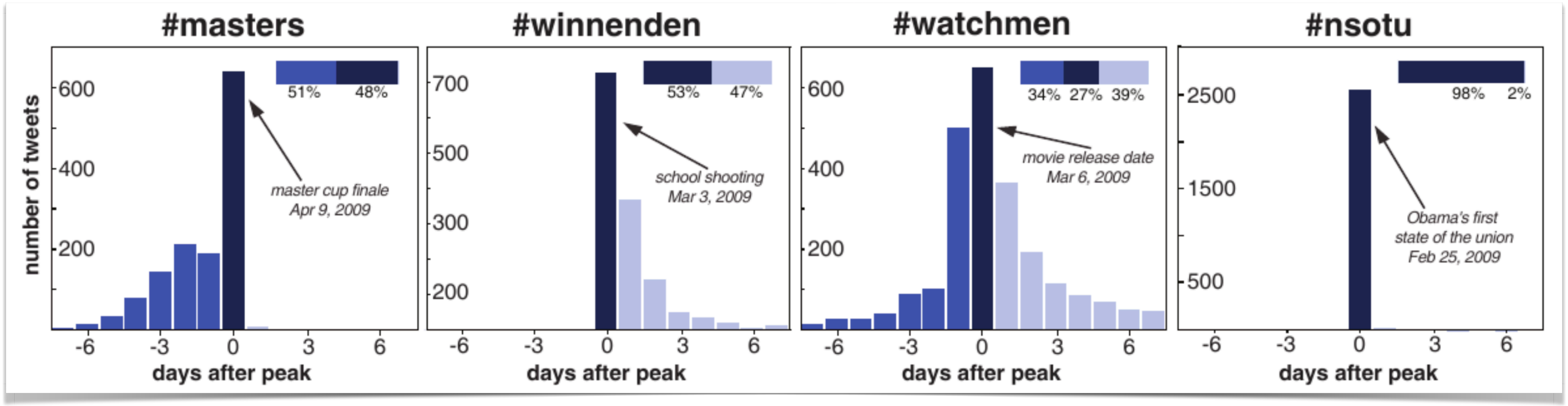
#hashtags

- This model has been applied to classify Twitter hashtag trends by Lehmann et al.

Trends on Twitter

#hashtags

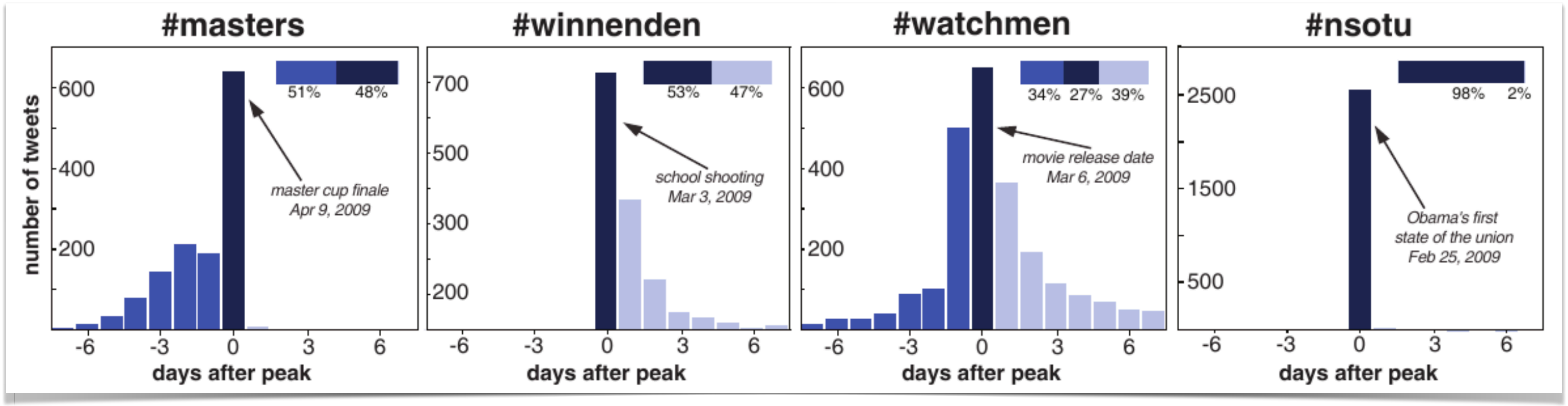
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

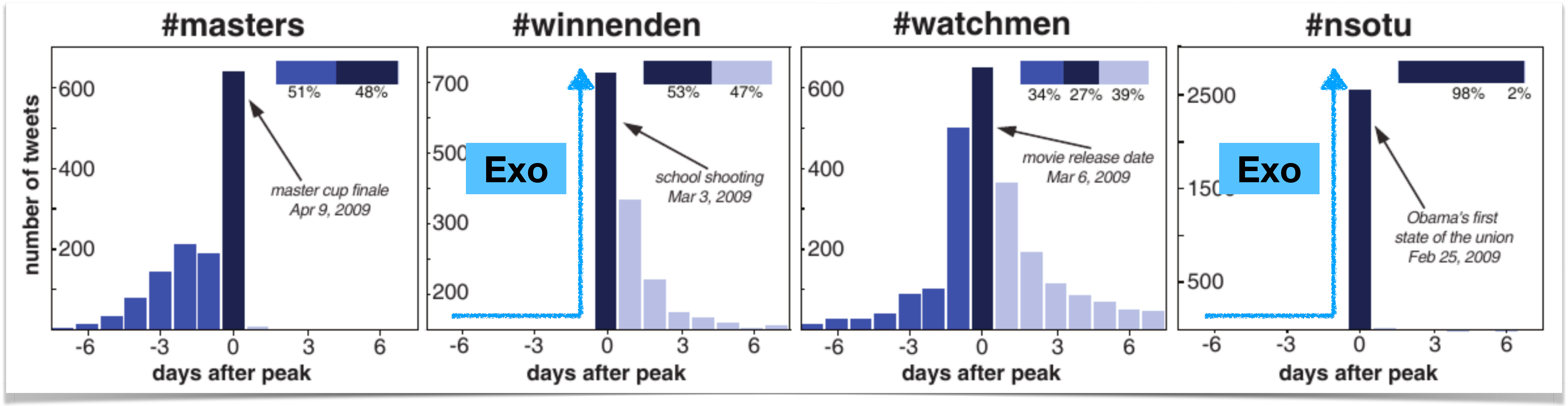
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

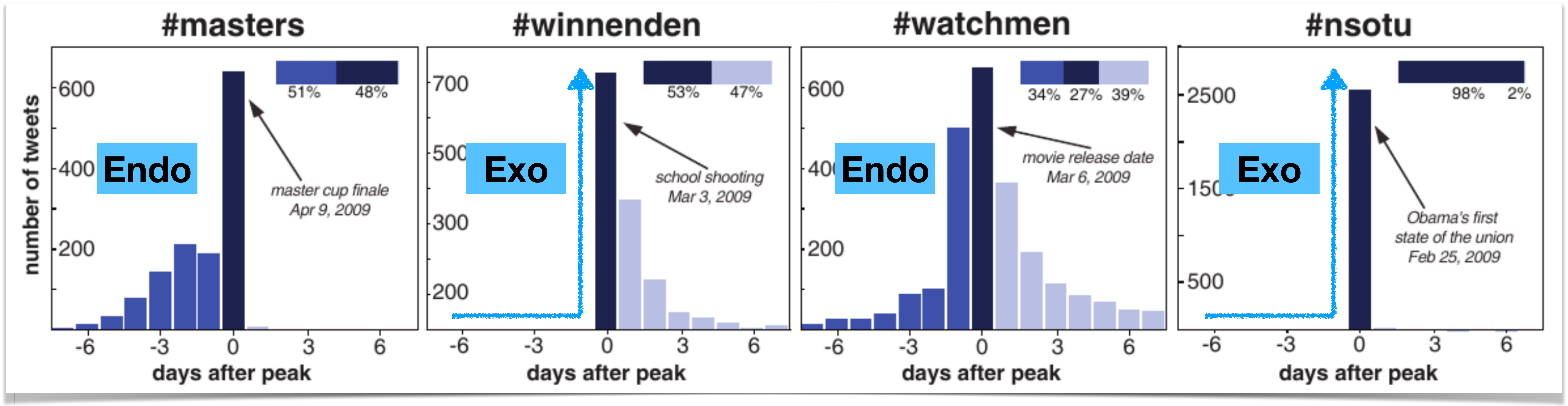
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

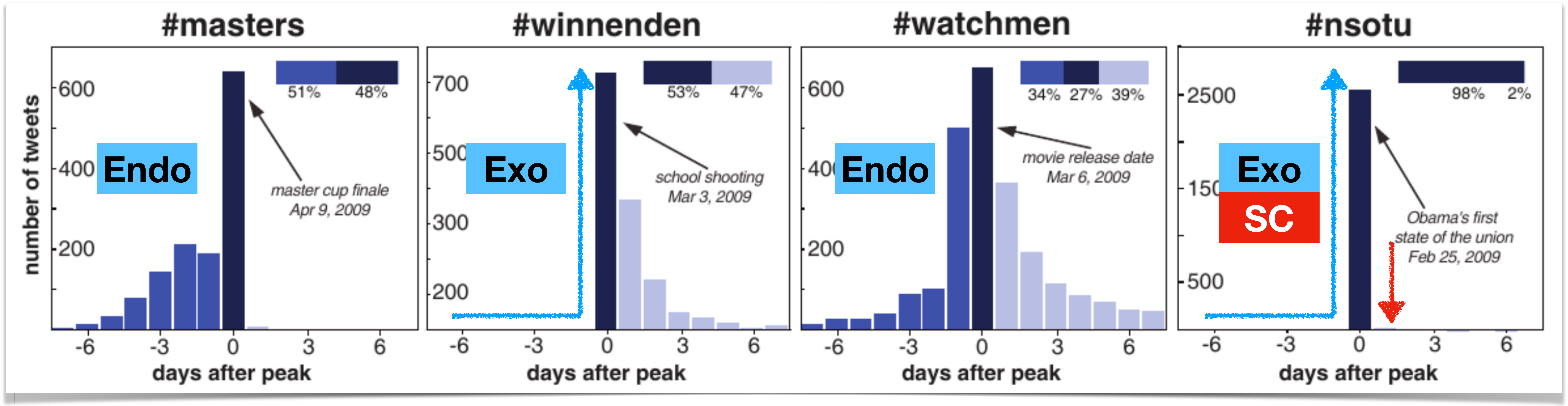
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

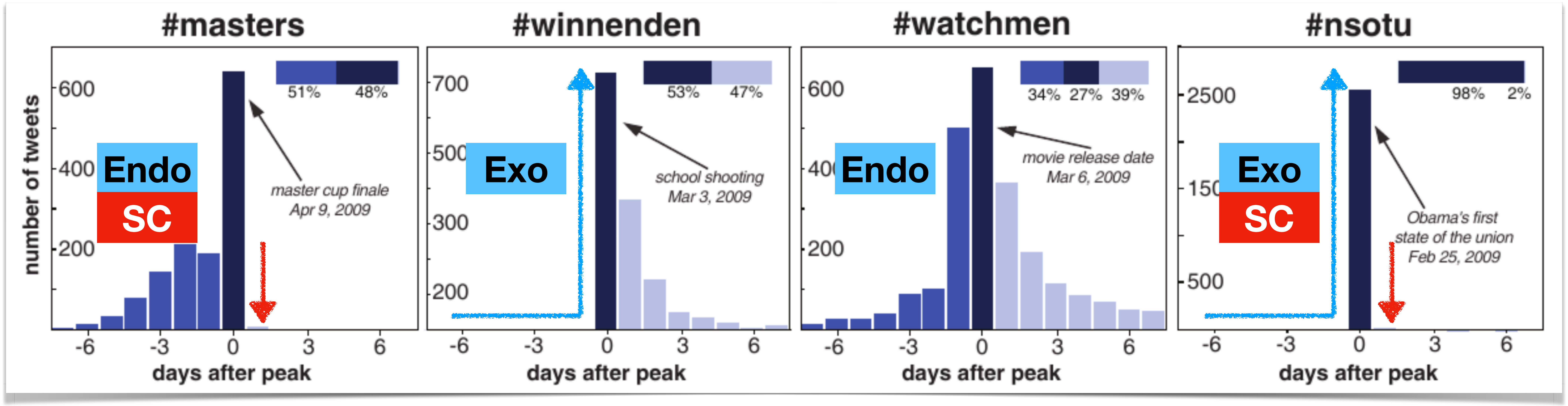
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

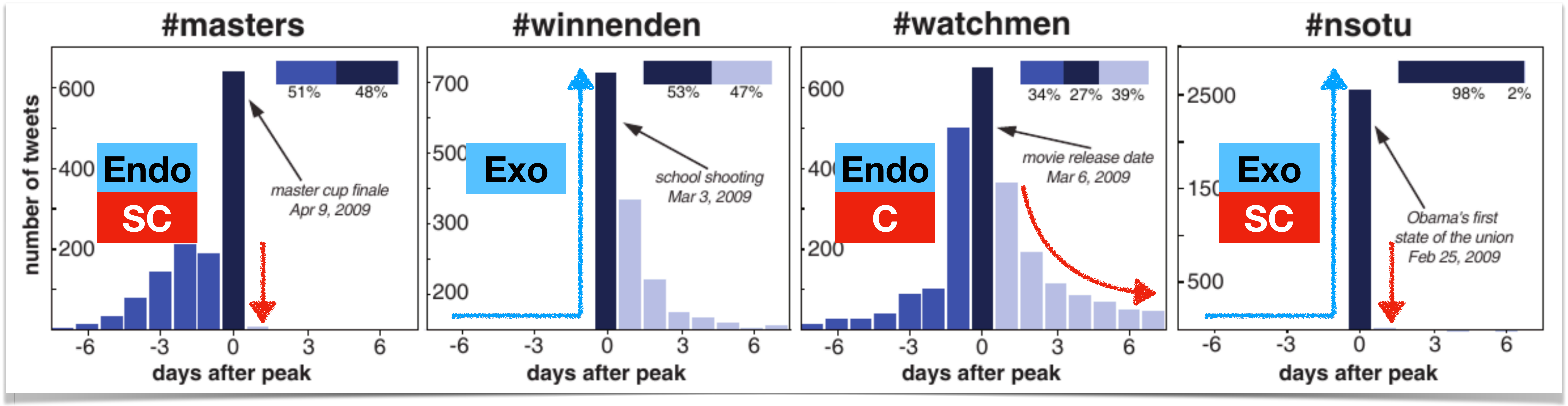
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

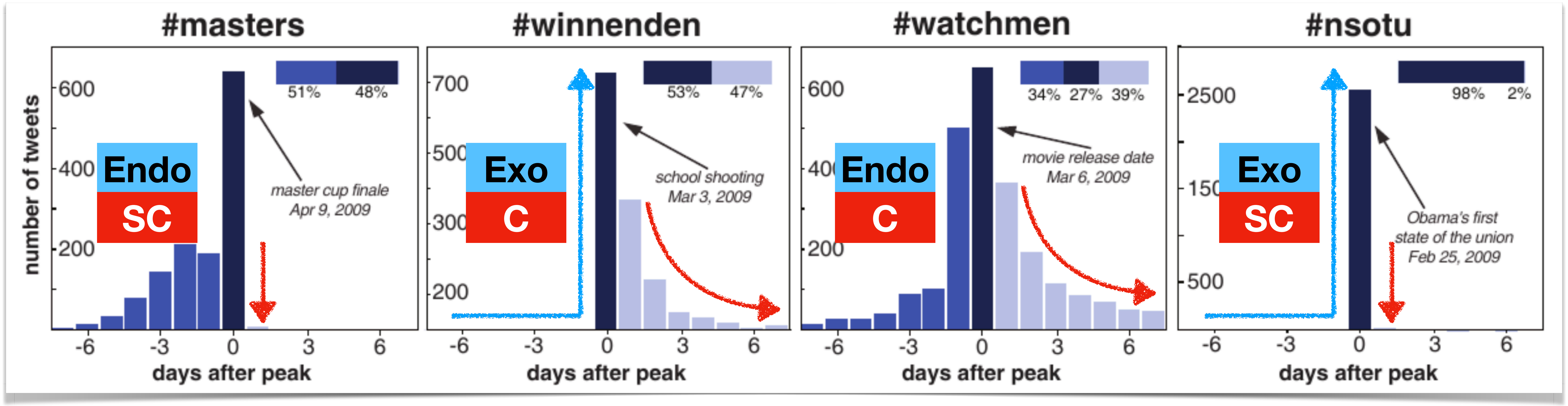
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

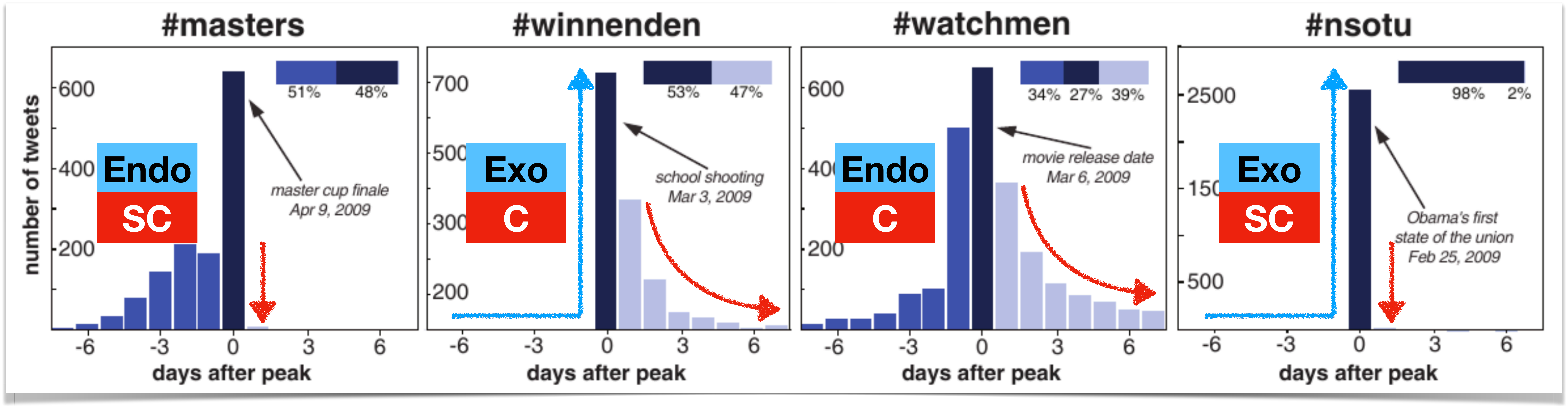
- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Trends on Twitter

#hashtags

- This model has been applied to classify Twitter hashtag trends by Lehmann et al.



Gathering this kind of volume data is best done by using the Twitter API v2.

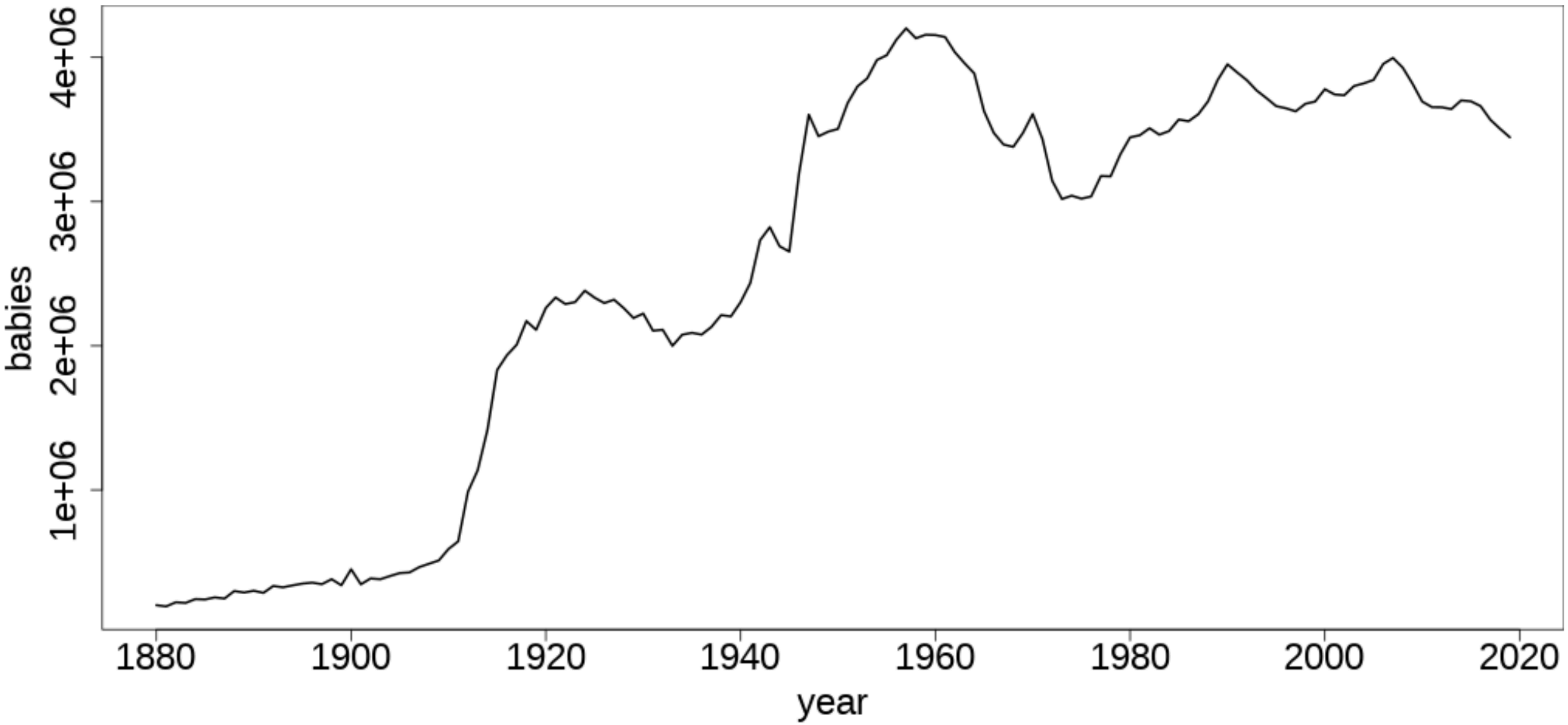
- Free access: up to 1500 tweets per month
- Basic (100 USD/month): up to 10K tweets per month
- Pro (5000 USD/month): up to 1M tweets per month

Old BigData

Baby name trends

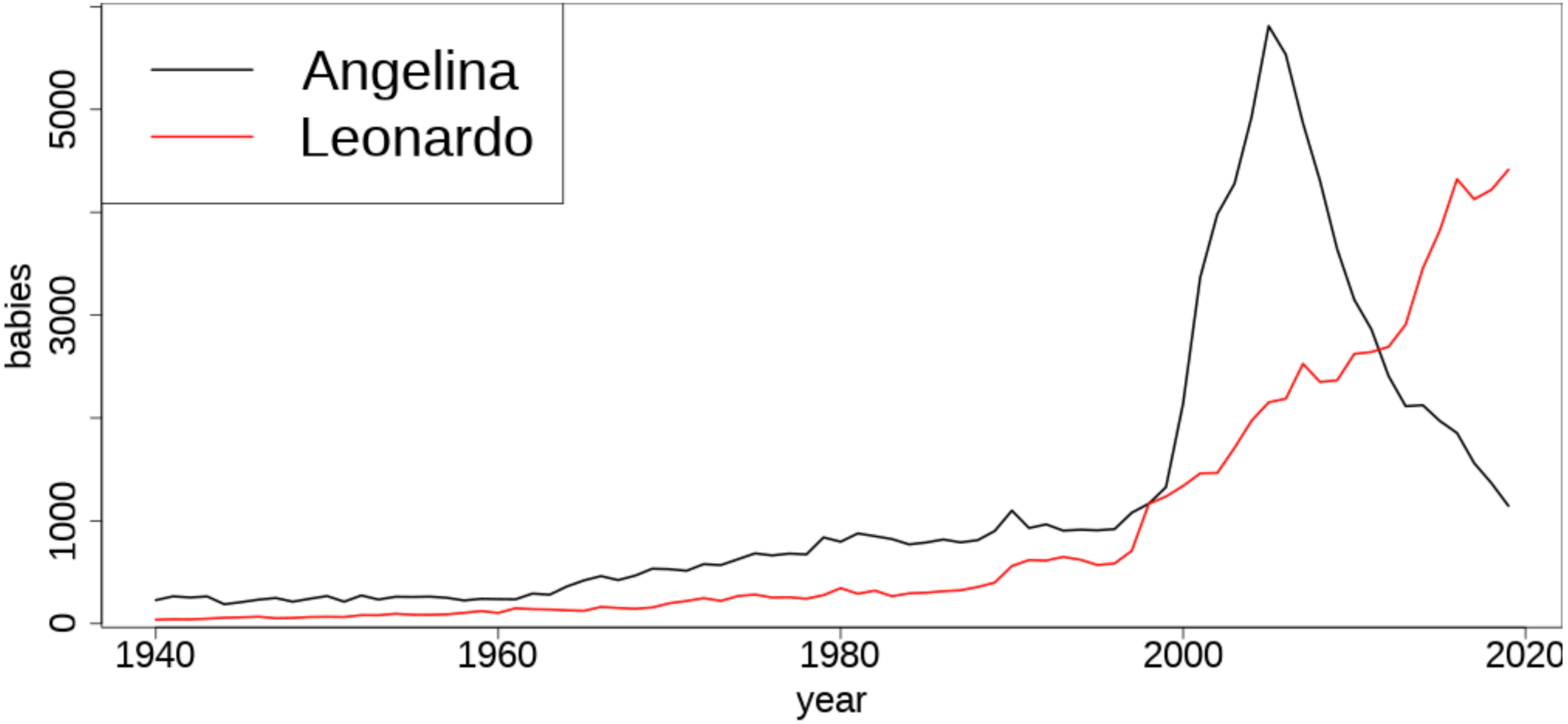
USA SSA baby name data

(Social Security Administration)



Baby name trends

Examples

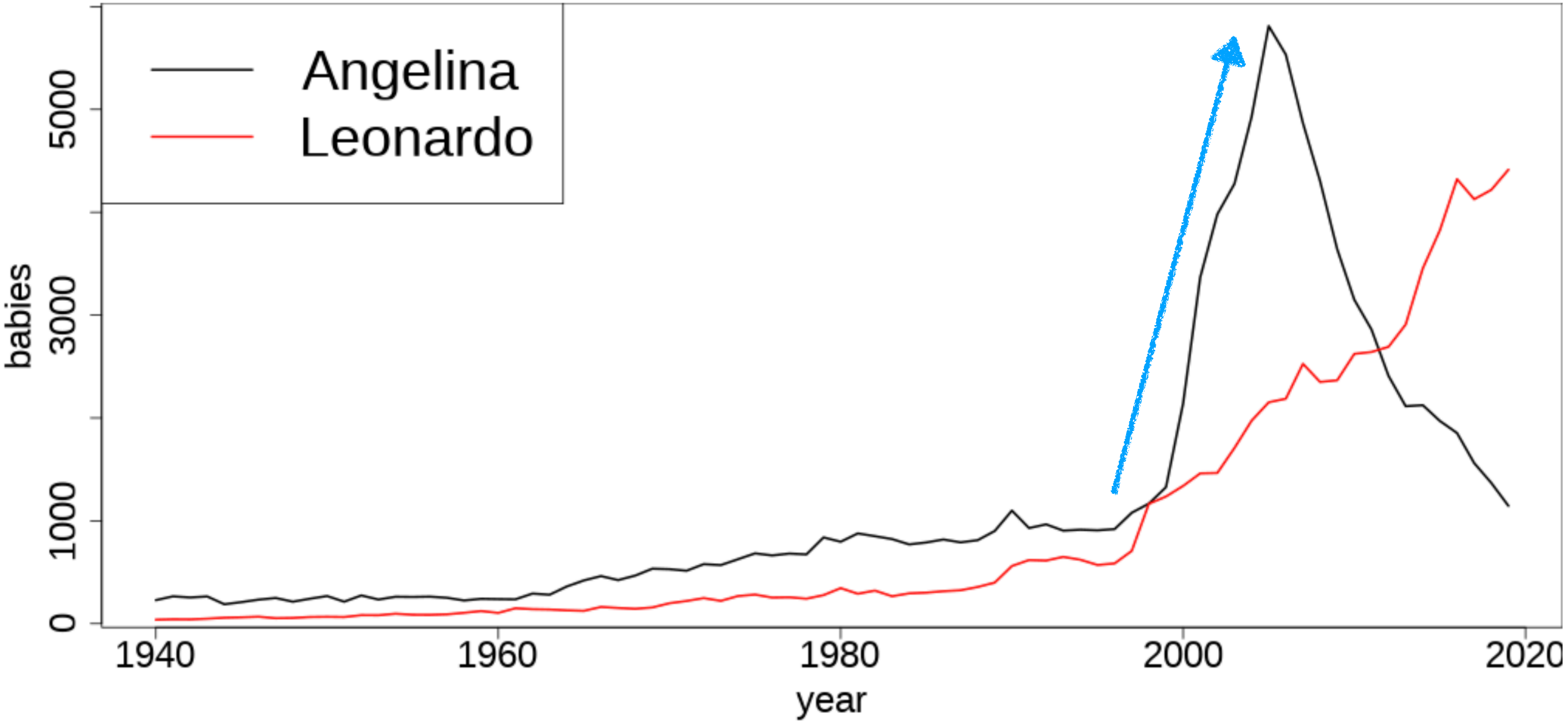


Baby name trends

Examples

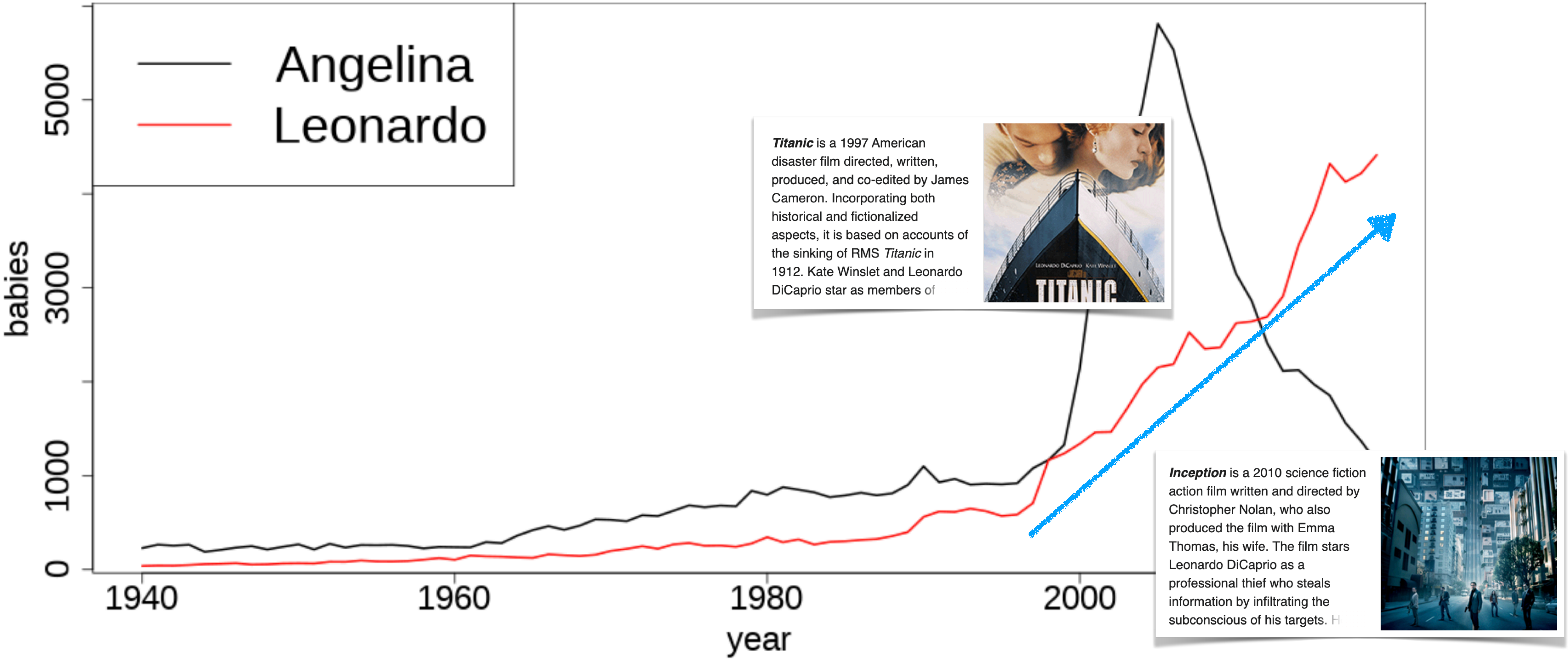
The Bone Collector is a 1999 American crime thriller film directed by Phillip Noyce and starring Denzel Washington and Angelina Jolie. The film is based on the 1997 crime novel of the same name written by Jeffery Deaver, concerning the tetraplegic detective Lincoln Rhyme.

Lara Croft: Tomb Raider – The Cradle of Life is a 2003 action adventure film directed by Jan de Bont and based on the *Tomb Raider* video game series. Angelina Jolie stars as the titular character Lara Croft with supporting performances from Gerard Butler, Ciarán Hinds, C



Baby name trends

Examples



The QWERTY effect in baby names

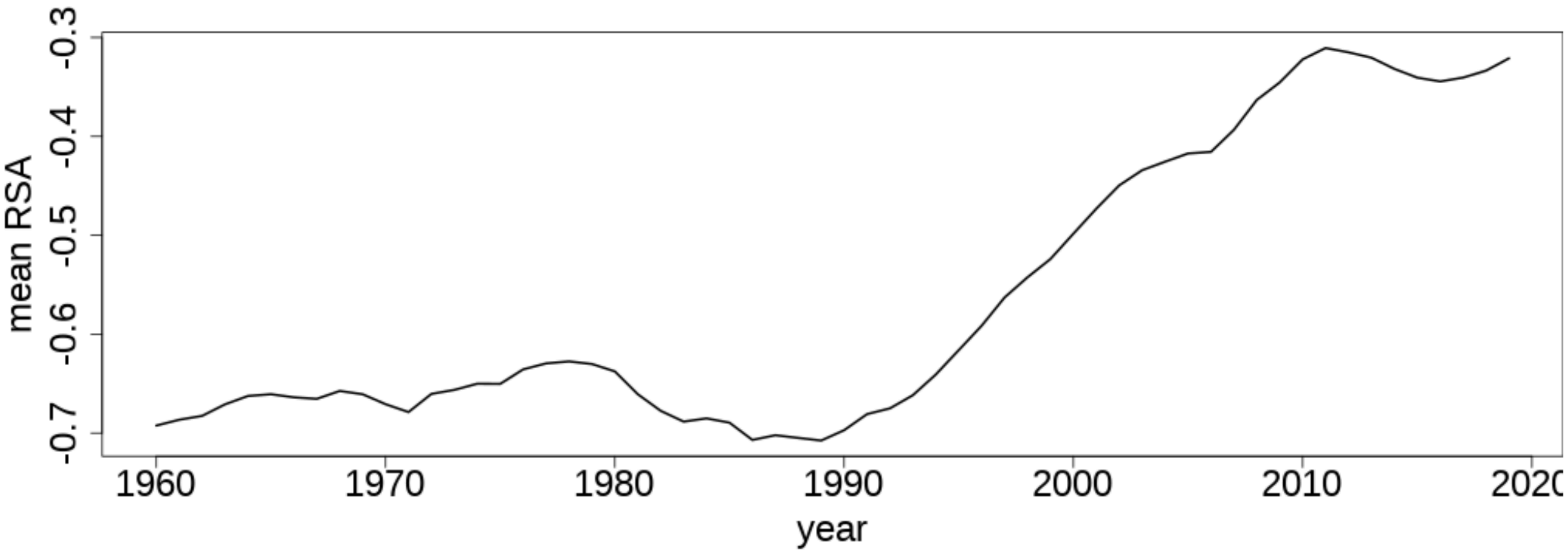
[Casasanto et al. 2014]

- The QWERTY effect is a hypothesis in Psychology that postulates that words that are written with more **right-hand** letters of the keyboard are, on average, **more positive** than words that are written with more left-hand letters of the keyboard. The fraction of right-hand letters in US baby names has been increasing:

The QWERTY effect in baby names

[Casasanto et al. 2014]

- The QWERTY effect is a hypothesis in Psychology that postulates that words that are written with more **right-hand** letters of the keyboard are, on average, **more positive** than words that are written with more left-hand letters of the keyboard. The fraction of right-hand letters in US baby names has been increasing:

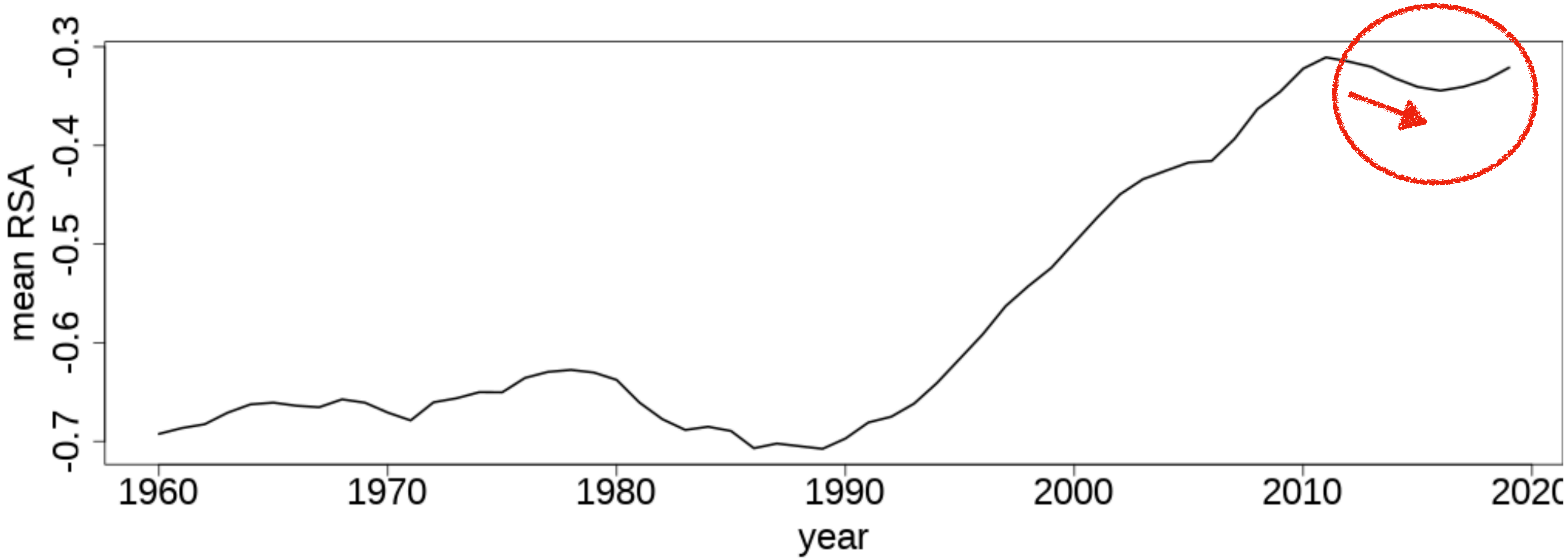


* RSA (right-side advantage) = (# right-side letters) - (# left-side letters)

The QWERTY effect in baby names

[Casasanto et al. 2014]

- The QWERTY effect is a hypothesis in Psychology that postulates that words that are written with more **right-hand** letters of the keyboard are, on average, **more positive** than words that are written with more left-hand letters of the keyboard. The fraction of right-hand letters in US baby names has been increasing:

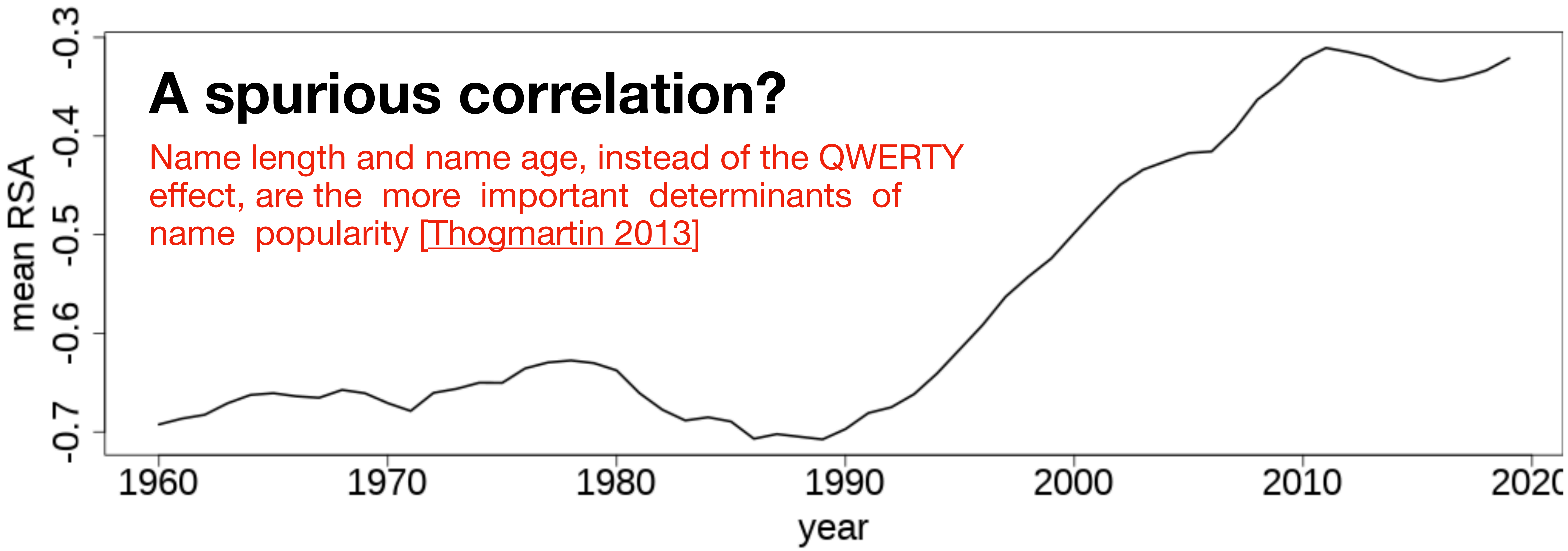


* RSA (right-side advantage) = (# right-side letters) - (# left-side letters)

The QWERTY effect in baby names

[Casasanto et al. 2014]

- The QWERTY effect is a hypothesis in Psychology that postulates that words that are written with more **right-hand** letters of the keyboard are, on average, **more positive** than words that are written with more left-hand letters of the keyboard. The fraction of right-hand letters in US baby names has been increasing:

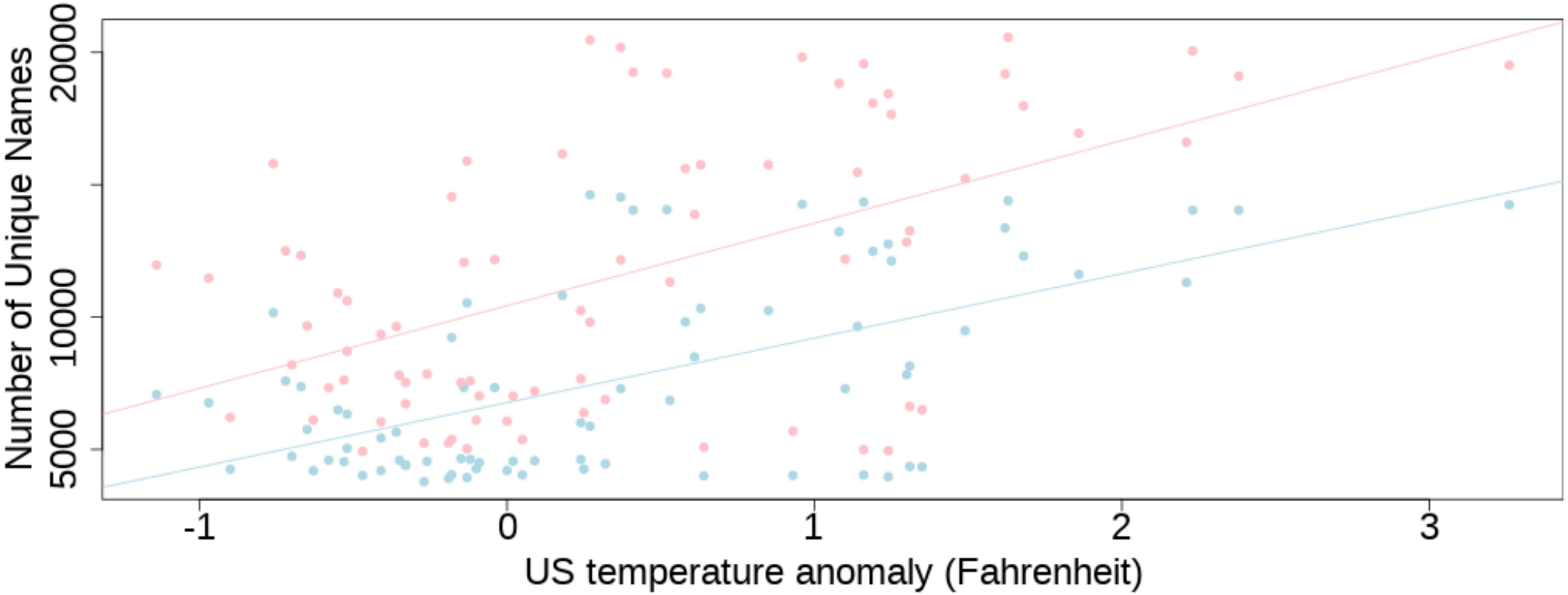


* RSA (right-side advantage) = (# right-side letters) - (# left-side letters)

Wacky baby name research

Proceedings of the Natural Institute of Science (a real parody)

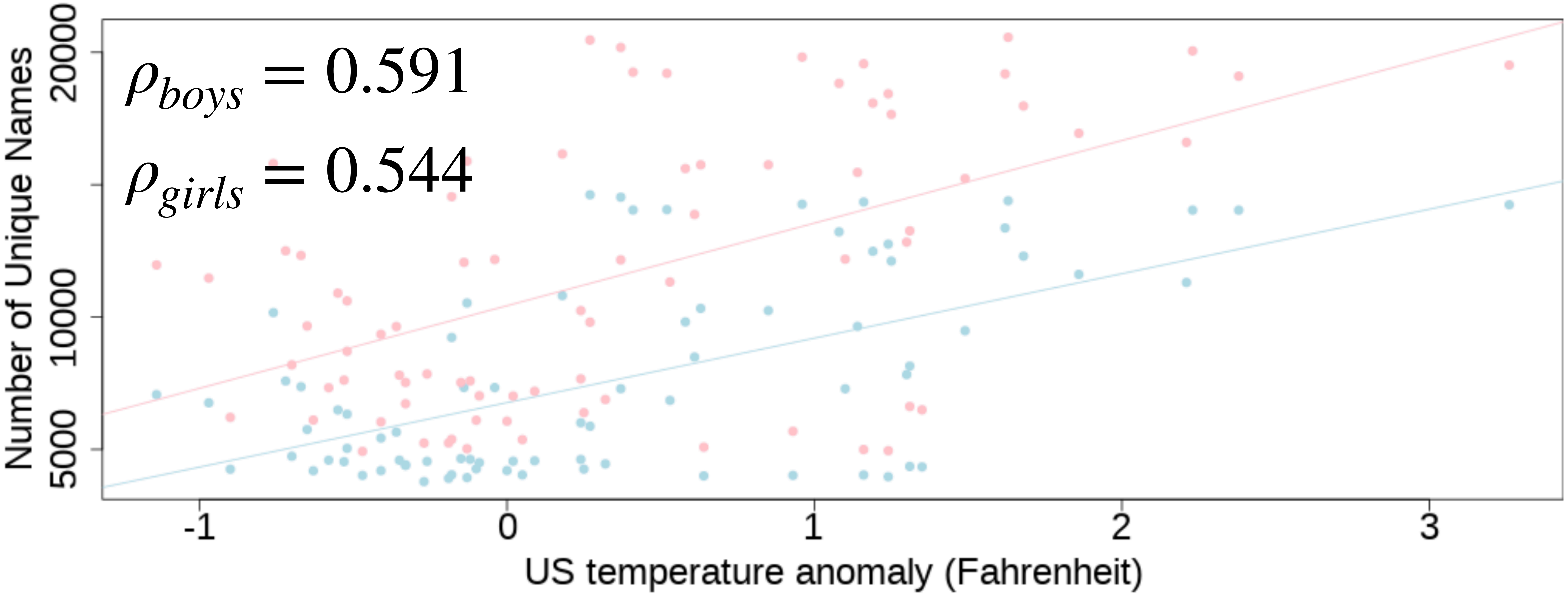
- A parody paper titled "We are entering an unprecedented age in baby name flux" reported; "baby name diversity also seems to have risen with the increasing annual temperature of the US (i.e., climate change)."



Wacky baby name research

Proceedings of the Natural Institute of Science (a real parody)

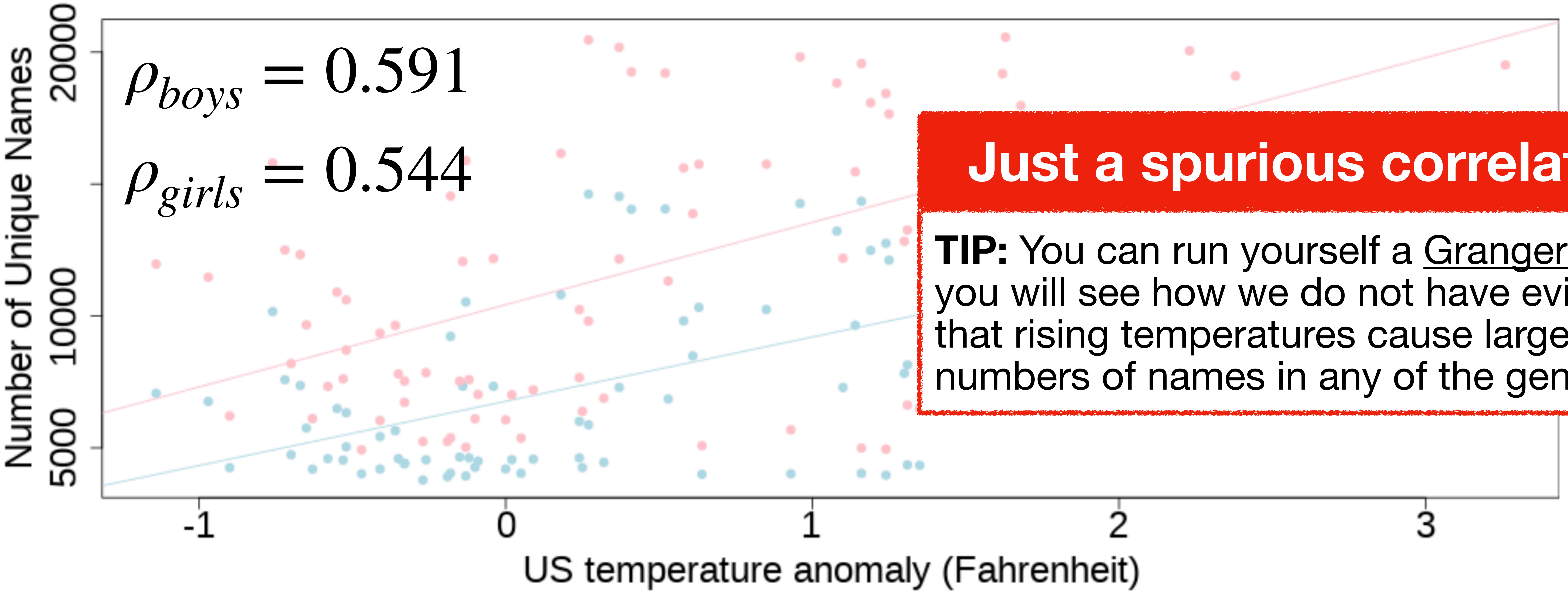
- A parody paper titled "We are entering an unprecedented age in baby name flux" reported; "baby name diversity also seems to have risen with the increasing annual temperature of the US (i.e., climate change)."



Wacky baby name research

Proceedings of the Natural Institute of Science (a real parody)

- A parody paper titled "We are entering an unprecedented age in baby name flux" reported; "baby name diversity also seems to have risen with the increasing annual temperature of the US (i.e., climate change)."



Just a spurious correlation!!!

TIP: You can run yourself a Granger test and you will see how we do not have evidence that rising temperatures cause larger numbers of names in any of the genders.

The limits of baby name predictability

Simmel effect in baby name popularity

- The book Freakonomics (2004) explains the **imitation** part of the **Simmel effect** and explains how **people imitate their richer neighbors when naming their babies.** The book goes as far as making a prediction of what will be the top US baby names in 2015, based on a data analysis exercise that is never explained in detail in the article. Here is the prediction:

The limits of baby name predictability

Simmel effect in baby name popularity

- The book Freakonomics (2004) explains the **imitation** part of the **Simmel effect** and explains how **people imitate their richer neighbors when naming their babies.** The book goes as far as making a prediction of what will be the top US baby names in 2015, based on a data analysis exercise that is never explained in detail in the article. Here is the prediction:

MOST POPULAR GIRLS' NAMES OF 2015?

Annika	Eleanora	Isabel	Maya
Ansley	Ella	Kate	Philippa
Ava	Emma	Lara	Phoebe
Avery	Fiona	Linden	Quinn
Aviva	Flannery	Maeve	Sophie
Clementine	Grace	Marie-Claire	Waverly

MOST POPULAR BOYS' NAMES OF 2015?

Aidan	Bennett	Johan	Reagan
Aldo	Carter	Keyon	Sander
Anderson	Cooper	Liam	Sumner
Ansel	Finnegan	Maximilian	Will
Asher	Harper	McGregor	
Beckett	Jackson	Oliver	

The limits of baby name predictability

Most popular **girl** names in 2015 (and the prediction)

Prediction

Annika	Clementine	Flannery	Linden	Phoebe
Ansley	Eleanora	**Grace**	Maeve	Quinn
Ava	**Ella**	**Isabel**	Marie-Claire	**Sophie**
Avery	**Emma**	Kate	Maya	Waverly
Aviva	Fiona	Lara	Philippa	

Real

Abigail	Avery	Emily	Isabella	Sofia
Addison	Charlotte	Emma	Madison	Sophia
Amelia	Chloe	Evelyn	Mia	Victoria
Aubrey	Elizabeth	Grace	Olivia	Zoey
Ava	Ella	Harper	Scarlett	

The limits of baby name predictability

Most popular **girl** names in 2015 (and the prediction)

Prediction

Annika	Clementine	Flannery	Linden	Phoebe
Ansley	Eleanora	**Grace**	Maeve	Quinn
Ava	**Ella**	**Isabel**	Marie-Claire	**Sophie**
Avery	**Emma**	Kate	Maya	Waverly
Aviva	Fiona	Lara	Philippa	

$$acc = \frac{7}{24} = 0{,}29$$

Real

Abigail	Avery	Emily	Isabella	Sofia
Addison	Charlotte	Emma	Madison	Sophia
Amelia	Chloe	Evelyn	Mia	Victoria
Aubrey	Elizabeth	Grace	Olivia	Zoey
Ava	Ella	Harper	Scarlett	

The limits of baby name predictability

Most popular **boy** names in 2015 (and the prediction)

Prediction

Aidan	Beckett	Harper	Maximilian	Summer
Aldo	Bennett	**Jackson**	McGregor	**Will**
Anderson	**Carter**	Johan	**Oliver**	
Ansel	Cooper	Keyon	Reagan	
Asher	Finnegan	**Liam**	Sander	

Real

Aiden	David	Jacob	Logan	Noah
Alexander	Elijah	James	Lucas	Oliver
Benjamin	Ethan	Jayden	Mason	Samuel
Carter	Gabriel	Joseph	Matthew	William
Daniel	Jackson	Liam	Michael	

The limits of baby name predictability

Most popular **boy** names in 2015 (and the prediction)

Prediction

Aidan	Beckett	Harper	Maximilian	Summer
Aldo	Bennett	**Jackson**	McGregor	**Will**
Anderson	**Carter**	Johan	**Oliver**	
Ansel	Cooper	Keyon	Reagan	
Asher	Finnegan	**Liam**	Sander	

$$acc = \frac{6}{22} = 0{,}27$$

Real

Aiden	David	Jacob	Logan	Noah
Alexander	Elijah	James	Lucas	Oliver
Benjamin	Ethan	Jayden	Mason	Samuel
Carter	Gabriel	Joseph	Matthew	William
Daniel	Jackson	Liam	Michael	

Predicting is hard

Prediction vs. Explanation

Predicting is hard

Prediction vs. Explanation

- There is not much overlap between the prediction and the results for 2015.

Predicting is hard

Prediction vs. Explanation

- There is not much overlap between the prediction and the results for 2015.
- What you see is that predicting which names in particular will be the most popular is a very difficult task. **The Simmel effect describes forces that create observable patterns, but that does not mean that the model is predictive to tell us which of all names will become popular ten years from now**, even if we had data of the social status of parents.

Predicting is hard

Prediction vs. Explanation

- There is not much overlap between the prediction and the results for 2015.
- What you see is that predicting which names in particular will be the most popular is a very difficult task. **The Simmel effect describes forces that create observable patterns, but that does not mean that the model is predictive to tell us which of all names will become popular ten years from now**, even if we had data of the social status of parents.
- This is the difference between **explanatory** and **predictive** power of a model. A model can explain phenomena without being useful to make predictions, as in this case, but can also be predictive without giving explanations, like in the case of deep learning or other black-box approaches.

Predicting is hard

Prediction vs. Explanation

- There is not much overlap between the prediction and the results for 2015.
- What you see is that predicting which names in particular will be the most popular is a very difficult task. **The Simmel effect describes forces that create observable patterns, but that does not mean that the model is predictive to tell us which of all names will become popular ten years from now**, even if we had data of the social status of parents.
- This is the difference between **explanatory** and **predictive** power of a model. A model can explain phenomena without being useful to make predictions, as in this case, but can also be predictive without giving explanations, like in the case of deep learning or other black-box approaches.

Take home message: understanding does not imply predictive power and vice versa

To recap…

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

Summary

Part 1 & 2: Behavior and trends

Social behavior and trends are both important aspects of human behavior that involve the interactions of individuals, and are influenced by societal and environmental factors.

Social behavior focuses on the interactions of individuals (between them, or between external factors such as a website, a technology, etc.)

Social trends are more broad and can be observed at the group or societal level, focusing on the larger patterns and changes in behavior or attitudes.

Summary

Part 3: Correlation, causation, and linear regression

The Future Orientation Index (FOI) measures the relationship between culture (Google Search Trends) and the economy (GDP)

Correlation measures the strength and direction of the relationship between two variables, but it does not explain “why” (correlation is not causation)

A regression model formalizes how one quantity depends on a linear combination of others. We can evaluate its “goodness-of-fit”.

Summary

Part 4: Simmel effect and baby names

Fashion always changes but there is always a fashion.
It is explained by imitation and distinctiveness.

The endo-exo model to explain social trends in online platforms.

Trends are hard to predict but show patterns of behavior.

Summary

Today's class

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4
Social Behavior	Social Trends	Quantifying Trends	Behavior & Trend Dynamics
1. Social Science 2. CSS 3. Digital Traces 4. Examples	1. Google Trends 2. The Future Orientation Index 3. Culture and Economy	1. Correlation 2. Causation 3. Regression	1. The Theory of Fashion 2. The Endo-Exo model 3. Examples

References

Bibliography

Papers used to prepared these slides

- Keusch, F., & Kreuter, F. (2021). Digital trace data: Modes of data collection, applications, and errors at a glance. [Taylor & Francis]
- Veltri, G. A. (2023). Describing Human Behaviour Through Computational Social Science. In Handbook of Computational Social Science for Policy (pp. 163-176). [Springer]
- Pedone, R., & Conte, R. (2001). Dynamics of status symbols and social complexity. Social science computer review. [Sage]
- Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R. (2012). Quantifying the advantage of looking forward. [Scientific Reports]
- Krause, A. J., Simon, E. B., Mander, B. A., Greer, S. M., Saletin, J. M., Goldstein-Piekarski, A. N., & Walker, M. P. (2017). The sleep-deprived human brain. [Nature reviews]
- Vizcaíno-Verdú, A., & Abidin, C. (2022). Music challenge memes on TikTok: Understanding in-group storytelling videos. [International Journal of Communication]
- Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. [PNAS]
- Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012, April). Dynamical classes of collective attention in twitter. [WWW
- Casasanto, D., Jasmin, K., Brookshire, G., & Gijssels, T. (2014). The QWERTY Effect: How typing shapes word meanings and baby names. [Cognitive Science Society]
- Thogmartin, W. E. (2013). The qwerty effect does not extend to birth names. [Names]

Resources

Other materials used to prepared these slides

- Digital traces [GESIS]
- Correlation and Causation [KhanAcademy] [icecream-crime]
- Establishing Causality by Brian Anderson [Blog]
- Foundations of CSS by David Garcia [GitHub]
- Correlational research [University of Central Florida]
- Graphs in Statistical Analysis by J. S. Anscombe [article] [matplotlib]
- The datasaurus R package [website]
- Fashion trends [white sneakers] [jeans] [UK baby names]
- Linear regression [ChelseaParlett] [R Tutorial]