Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Data & Inference

WEEK 2A

Course outline

Week	Description
1	Measurement, data types, data collection, data cleaning
2	Data manipulation, data exploration, visualization techniques
3	Probability, statistical distributions, descriptive statistics
4	Statistical hypothesis testing, quantifying confidence
5	Time series analysis, autoregression, moving averages
6	Linear regression, parameter estimation, model selection, evaluation

Today's Lecture

No.	Activity	Description
1	Challenge	Measuring human development
2	Discussion	How to access data
3	Case study	World Bank Indicators
4	Analysis	HDI, MDGs, SDGs
5	Q&A	Questions and feedback

Human Development Index

- The human development index (HDI) is a composite statistic of life expectancy, education, and income indices.
- It is used to rank countries into four tiers of human development.
- The HDI was created by Amartya Sen and Mahbub ul Haq in 1990 and is published by the UNDP.

Original HDI construction

- HDI (up to 2009) combined three dimensions:
- **Population health and longevity** as measured by life expectancy at birth.
- **Knowledge and education**, as measured by the adult literacy rate (with two-thirds weighting) and the combined primary, secondary, and tertiary gross enrollment ratio (with one-third weighting).
- **Standard of living**, as indicated by the natural logarithm of gross domestic product per capita at purchasing power parity.

Millennium Development Goals

The aim is to achieve the following MDGs by 2015:

1. To eradicate extreme poverty and hunger
2. To achieve universal primary education
3. To promote gender equality and empower women
4. To reduce child mortality
5. To improve maternal health
6. To combat HIV/AIDS, malaria, and other diseases
7. To ensure environmental sustainability
8. To develop a global partnership for development.

Tracking Global Goals

- **World Leaders have committed to 17 Global Goals (169 targets) to achieve 3 extraordinary things in the next 15 years.**
- **End extreme poverty. Fight inequality & injustice. Fix climate change.**

Class Poll 1

- What is the most important Global Goal?
- Entre a number 1 to 17.

www.slido.com event code **#15469**

Life expectancy

- Dashboards provide a dynamic means of interacting with data
- Without visuals and insights, the data does not have any impact
- Try out population.io to answer the following big questions:
  - What's my place in the world population?
  - How long will I live?

Histogram of life expectancy

- Use population.io to estimate your life expectancy in your country of birth

www.slido.com event code **#15469**

Population.io Dashboard

World Bank

- The World Bank Group has two ambitious goals:

- 1. End extreme poverty within a generation
- 2. Boost shared prosperity

Word Bank Indicators

Topics	
Agriculture & Rural Development	Health
Aid Effectiveness	Infrastructure
Climate Change	Poverty
Economy & Growth	Private Sector
Education	Public Sector
Energy & Mining	Science & Technology
Environment	Social Development
External Debt	Social Protection & Labor
Financial Sector	Trade
Gender	Urban Development

Accessing Data

- The Data API is a way for web sites and other tools to access data directly.
- This is the interface of choice for creating custom data visualizations.
- It is also possible to create live combinations with other data sources (mashups).
- data.worldbank.org is built using data through the Data API.
- Python links: World-Bank-Data & WBGAPI

World Bank API

- The Indicators API supports two basic ways to build queries: a url based structure and an argument based structure.
- For example, the following two requests will return the same data, a list of countries with income level classified as low income:
- **Argument based** > http://api.worldbank.org/countries?per_page=10&incomeLevel=LIC
- **URL based** > http://api.worldbank.org/incomeLevels/LIC/countries

Immediate Questions

- How many variables are there?
- How many years of data in total?
- Are there missing values?
- Are any of the variables redundant?
- What are the most important variables in your opinion?

Analysis

- What can we use the World Bank Indicators for?
- How does it relate to other variables calculated by different organizations?
- What conclusions can we reach from this analysis?
- What are the advantages of data?
- What are the possible shortcomings of such an analysis?

Class Poll 2

- Variables to consider:

  Education (E), Health (H), Wealth (W)

- Select from the following 7 options

  E, H, W

  E&H, E&W, H&W

  E&H&W

www.slido.com event code **#15469**

What can we measure?

- LE: Life expectancy at birth
- MYS: Mean years of schooling (Years that a 25-year-old person or older has spent in schools)
- EYS: Expected years of schooling (Years that a 5-year-old child will spend with his education in his whole life)
- GNIpc: Gross national income at purchasing power parity per capita

HDI composition

- HDI combines three variables:
- A long and healthy life: Life expectancy at birth
- Education index: Mean years of schooling and Expected years of schooling
- A decent standard of living: GNI per capita (PPP US$)

HDI formulae

- Life expectancy index:

  $LEI = (LE-20)/(85-20)$

- Education index: $EI = (MYSI+EYSI)/2$

  Mean Years of Schooling Index: $MYSI=MYS/15$

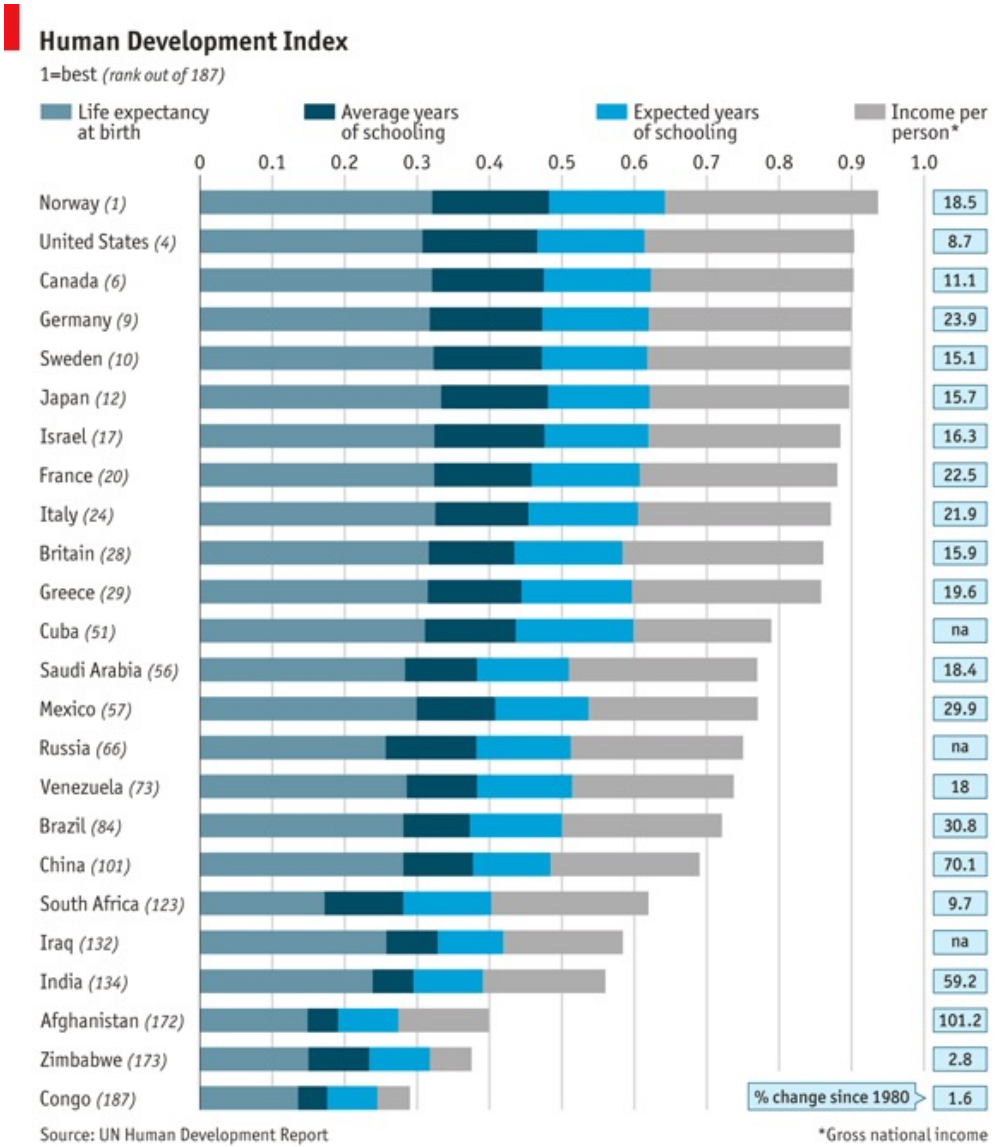
  Expected Years of Schooling Index: $EYSI = EYS/18$

- Income Index:

  $II = [\ln(GNIpc) - \ln(100)]/[\ln(75{,}000)-\ln(100)]$

- $HDI = (LEI \times EI \times II)^{1/3}$

Deconstructing HDI



Source: The Economist

Understanding HDI outcomes

- Norway is the highest.
- Australia is the second highest.
- Congo is the lowest.
- Niger is the second lowest
- You could construct your own HDI by changing the weights in the HDI

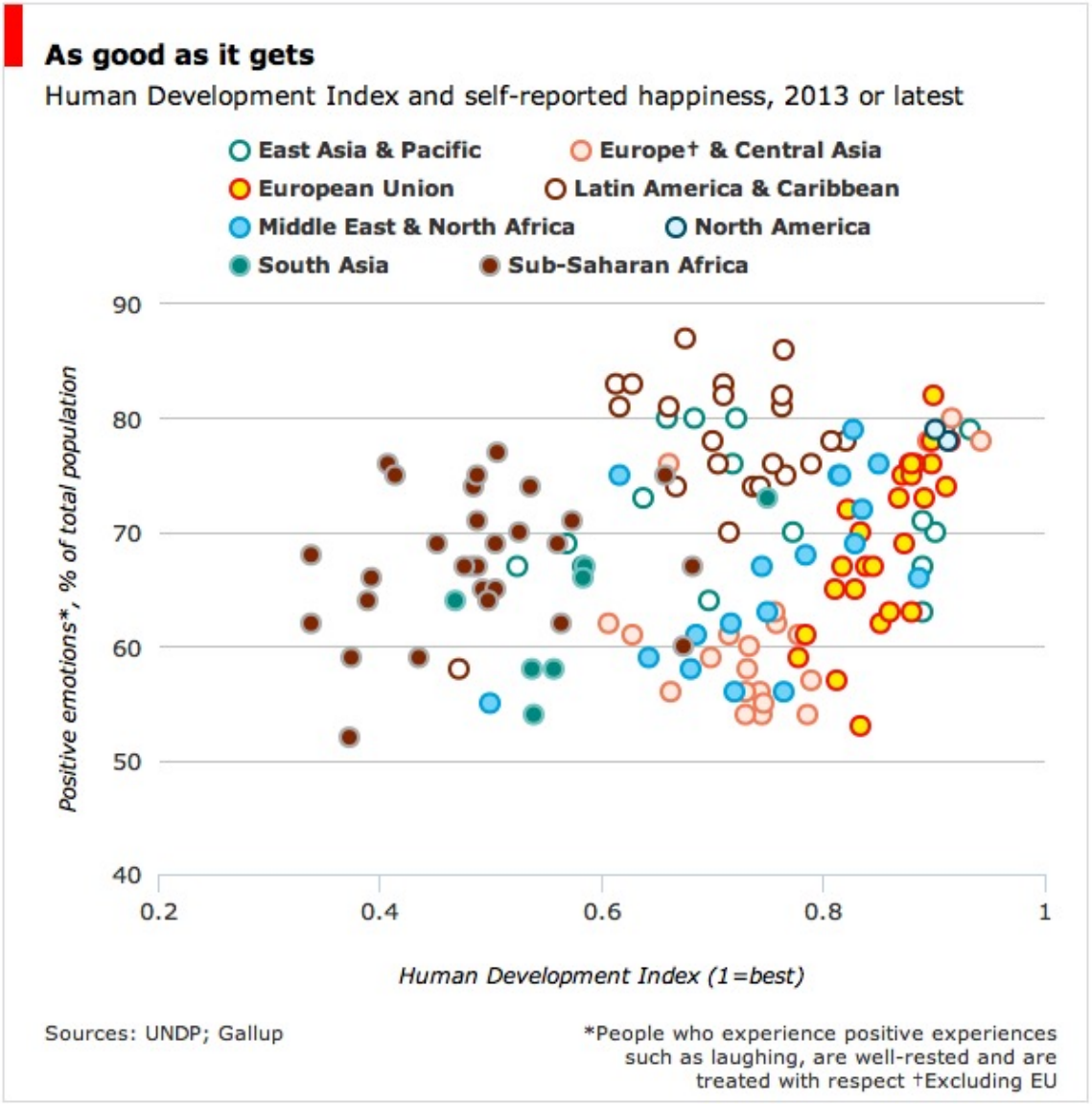
Problems with GDP

- GDP “measures everything,” quipped Bobby Kennedy, the American president’s brother, “except that which makes life worthwhile.”
- Bhutan focuses on gross national happiness (GNH) rather than gross domestic product (GDP).
- How does the HDI compare to day-to-day well-being?

Happiness

- Self-reported data on happiness from Gallup, an international polling company.
- It asks if people had been “laughing or smiling a lot, feeling well-rested, and being treated with respect” in the previous day.
- According to this measure Paraguay has been the happiest place on Earth for the past three years.
- Syria, locked in civil war, is lowest.

Happiness versus HDI



Source: Economist, 2014.

What do we learn?

- There is little correlation between the two measures (the correlation coefficient is 0.25).
- Lithuania has a happiness score of 53%. For its level of development one might expect happiness closer to 70%.
- Meanwhile Mali and Rwanda are much happier than their living standards might imply.

Other observations

- Note that development is generally clustered by region.
- But in terms of happiness, it runs the gamut from gloomy to chirpy within the same income group.
- And regional stereotypes reveal themselves:
  - People in far eastern Europe and central Asia are dour despite having reasonable living standards.
  - Those in Latin America at the same level of development tend to be cheery—around 20 basis points higher.

Class Poll 3

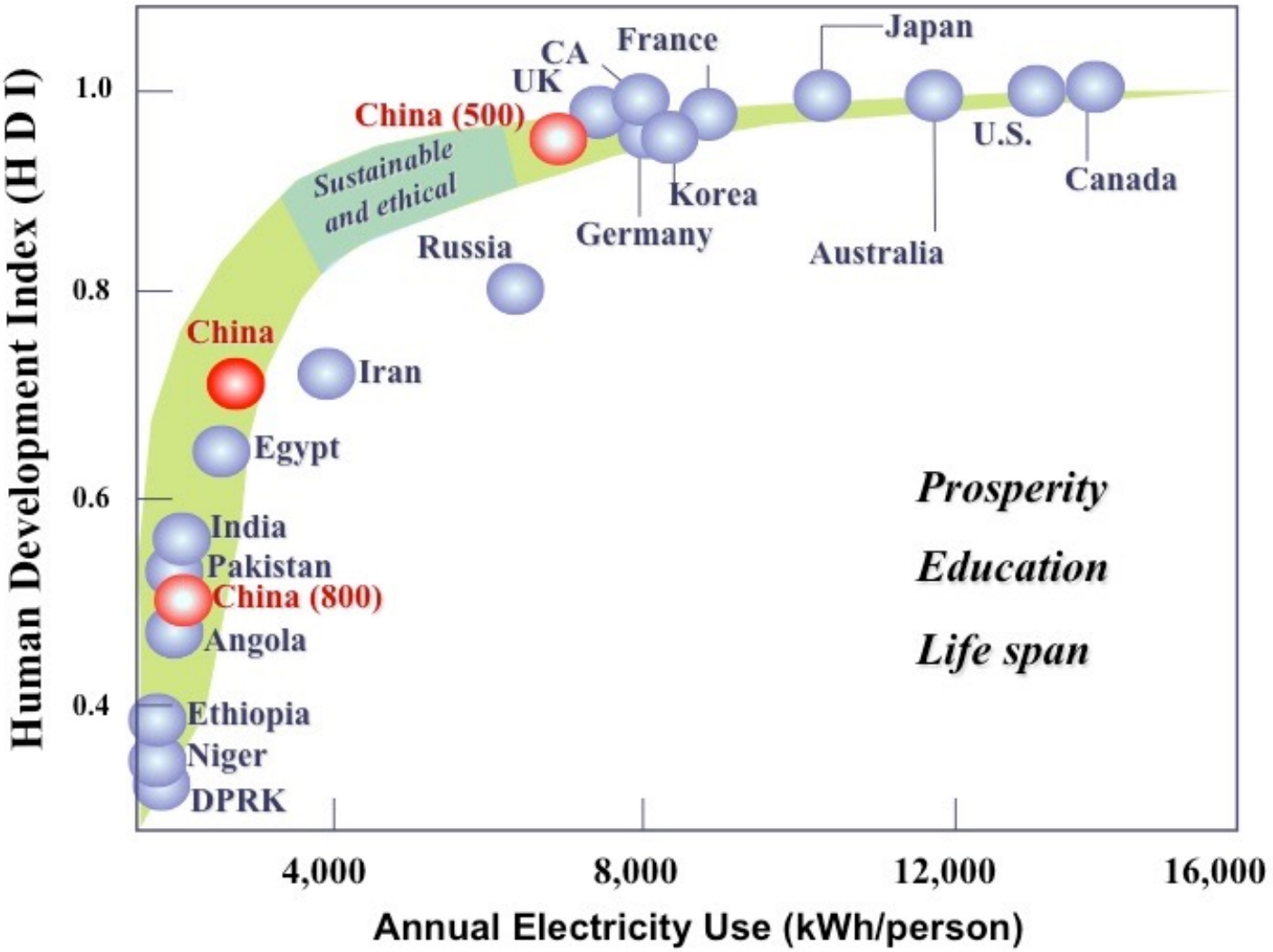
- Would you prioritise GDP (wealth) or GNH (happiness) if you were president?

- **www.slido.com** event code **#15469**

Middle Class

- The Economist defined middle class as the point where people have roughly a third of their income left for discretionary spending after paying for basic food and shelter.
- This allows people to buy consumer goods, improve their health care, and provide for their children's education and future.

HDI and Electricity



It takes about 3,000 kWh per person per year to be above 0.8 HDI, to have what we consider a good life.

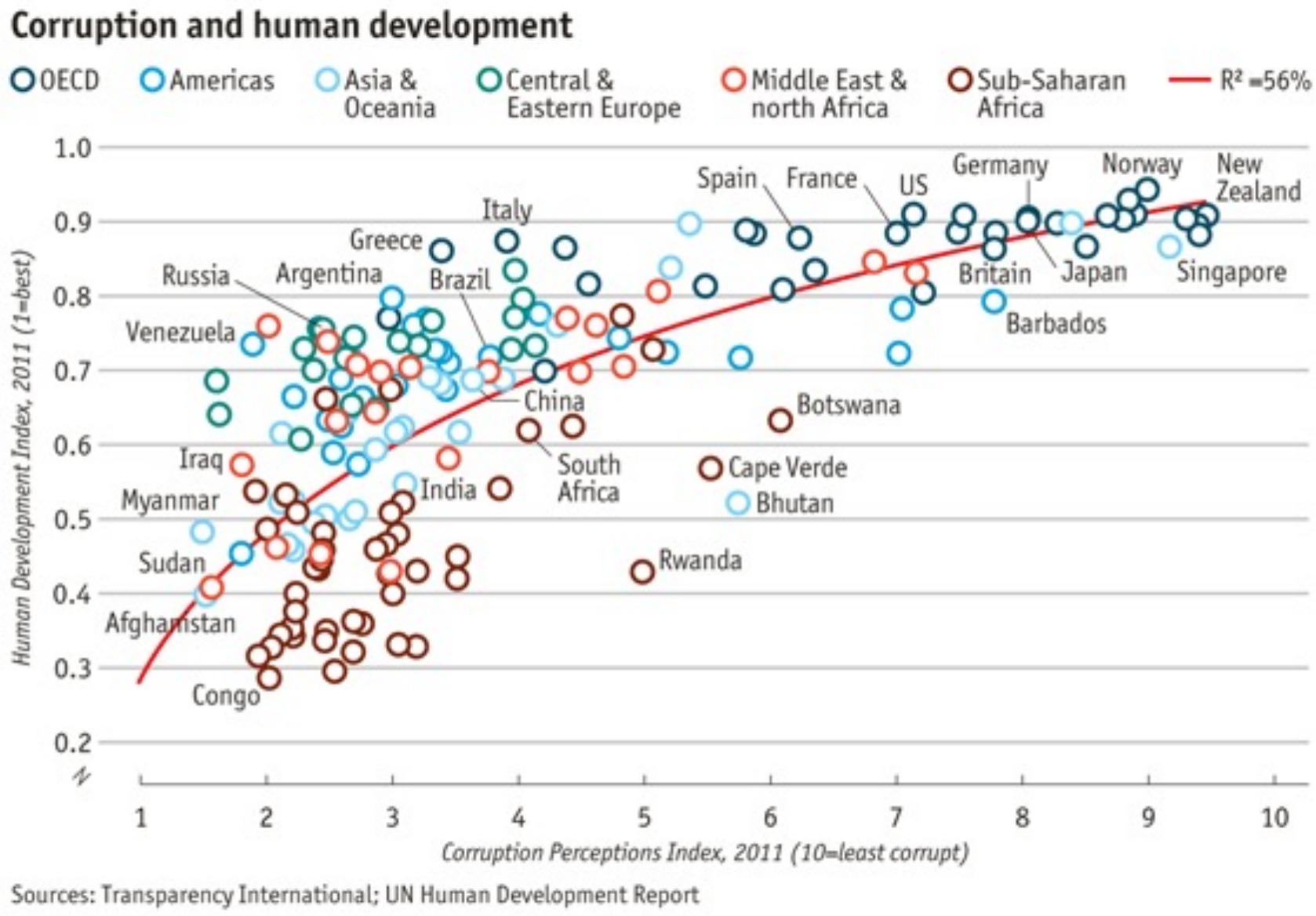
China's 0.70 HDI is a composite of 500 million people above 0.8 HDI and 800 million still below 0.6 HDI.

Source: Forbes (2012).

Impact of Energy?

- Middle class wealth is related to energy development, which relies on infrastructure investment.
- It takes between 3,000 and 6,000 kWhrs per person per year to have what we consider a good life, to get into the middle class (United Nations Human Development Index).

HDI and Corruption



Impact of Corruption?

- When the corruption index is between approximately 2.0 and 4.0 there appears to be little relationship with the human development index.
- As it rises beyond 4.0 a stronger connection can be seen.
- Outliers include small but well-run poorer countries such as Bhutan and Cape Verde.
- Greece and Italy stand out among the richer countries.

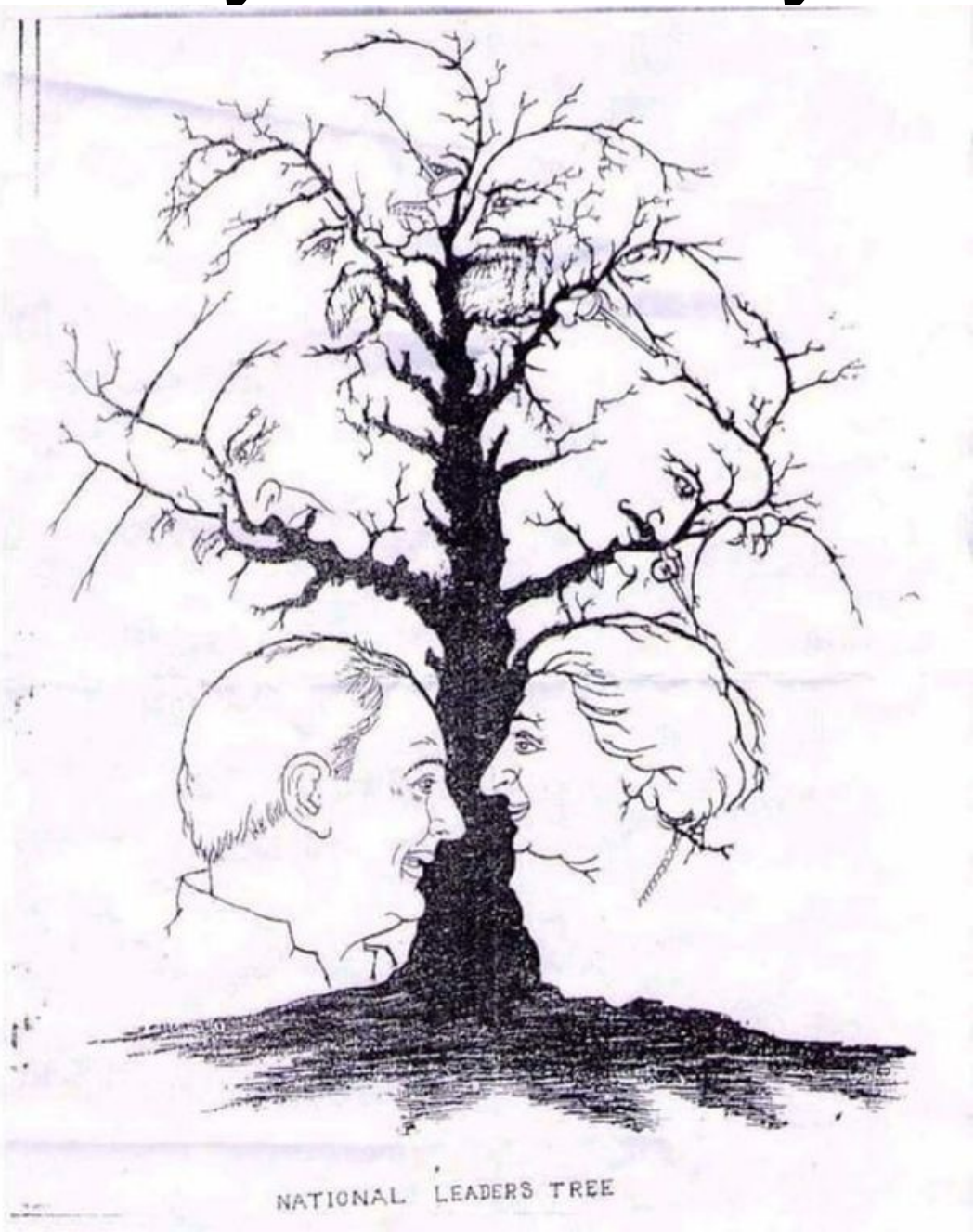
Data & Inference

WEEK 2B

Today's Lecture

No.	Activity	Description
1	Challenge	Accessing data in real-time
2	Discussion	How to explore data
3	Case study	QUANDL
4	Analysis	Interactive analysis, APIs
5	Demo	Graphs and visualization
6	Q&A	Questions and feedback

How many faces do you see?

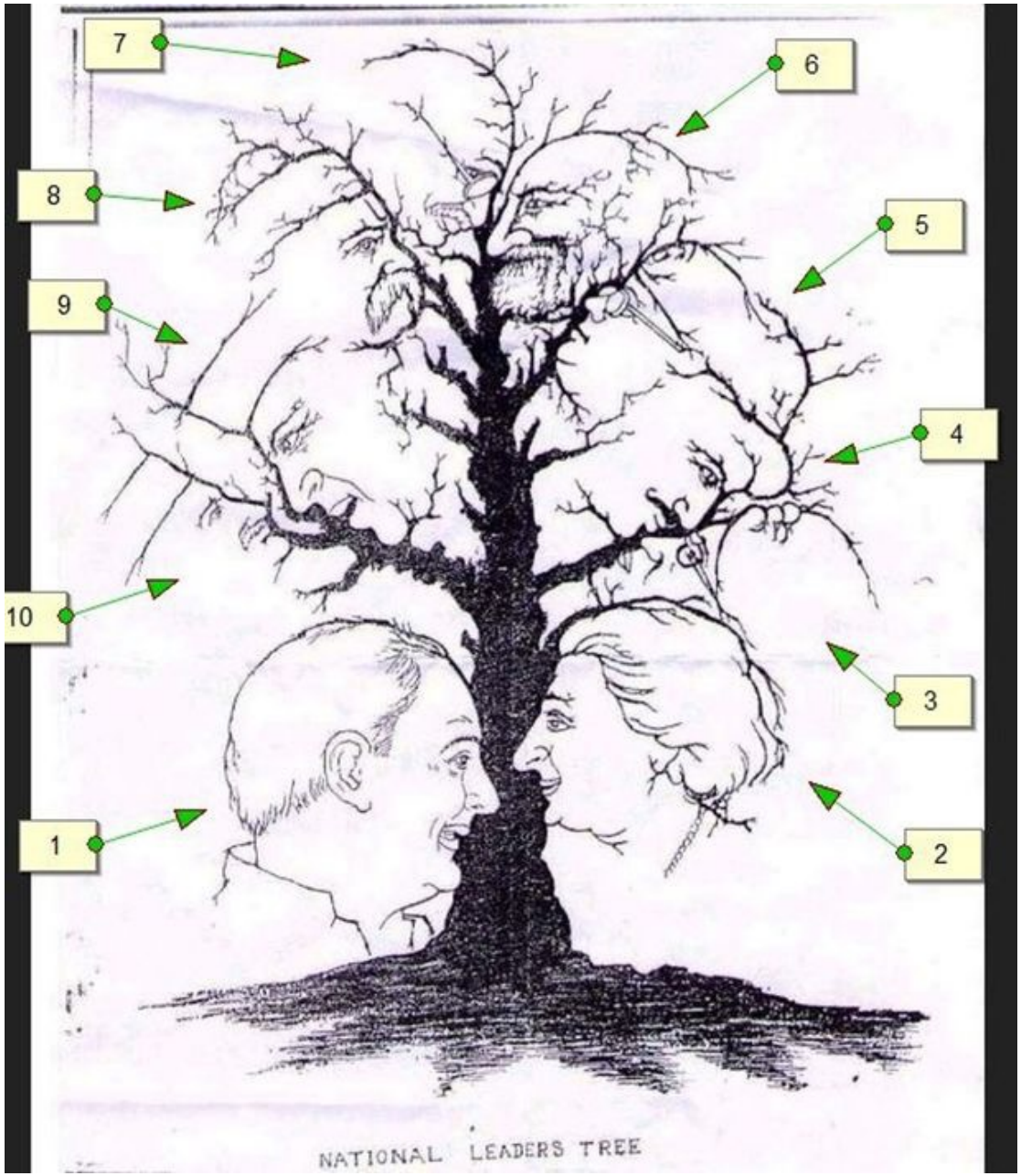


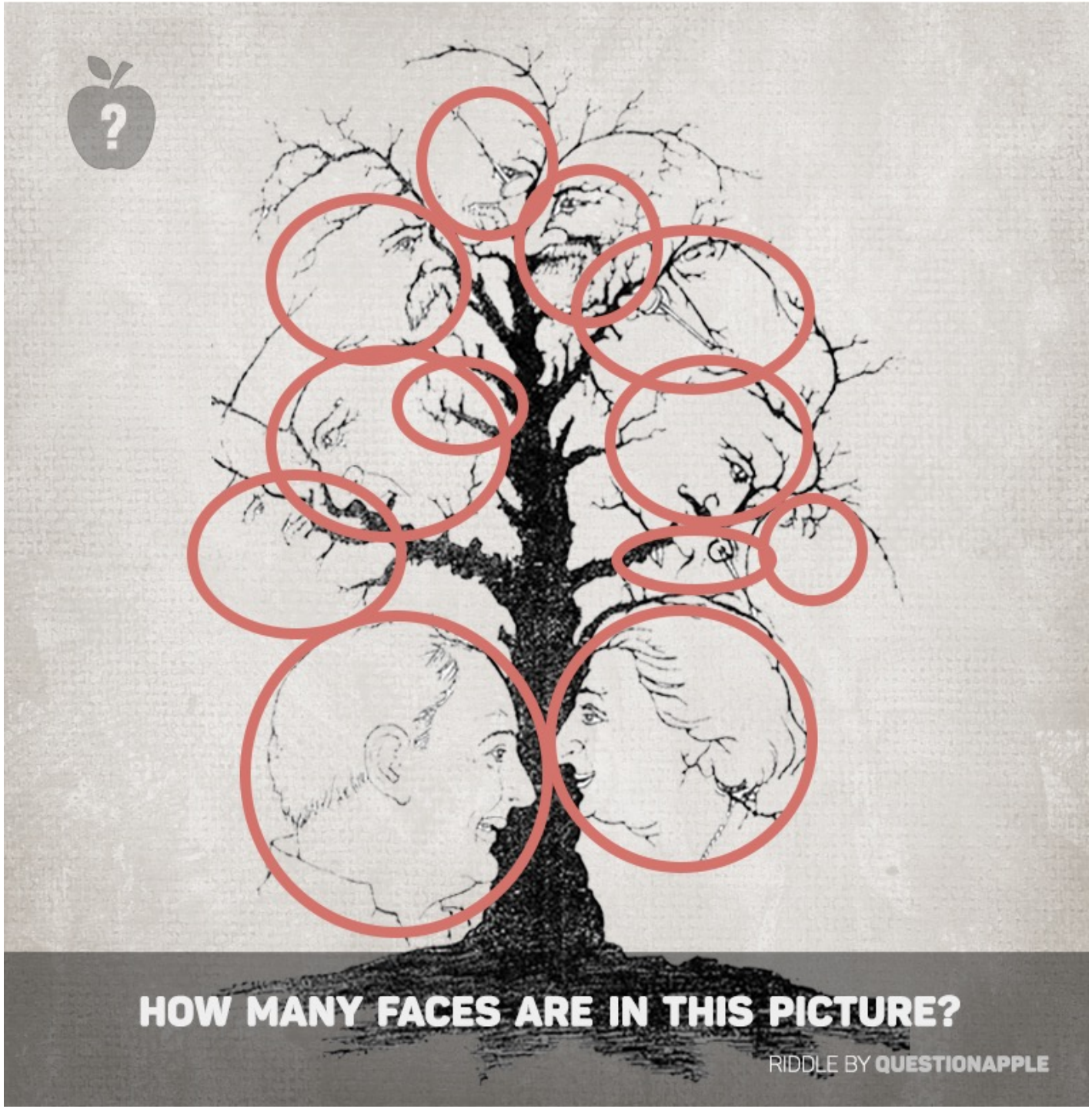
The *National Leaders Tree* was drawn in the 1880's by an unnamed illustrator for Harper's Bazaar.

Class Poll 1

slido.com
Event: #76755

Answer: 10




HOW MANY FACES ARE IN THIS PICTURE?
RIDDLE BY QUESTIONAPPLE

Static Analysis

- Static analysis refers to the traditional approach of collecting a dataset, taking it offline and analyzing it on a computer.
- The end result might be a model that we hope is relevant for making decisions in the future.
- If the system does not change much, then such an assumption is appropriate.

Data Streams

- Just because we stop looking does not mean that information is not available.
- Automated data capture and APIs allow us to access fresh data at all times.
- The process of having a continuous data stream entering a computer represents a paradigm shift.
- Real-time analytics benefit from having access to the most up-to-date data.

Dynamical Analysis

- Many complex systems evolve rapidly where dramatic changes can take place in the period of a few hours.
- Recent events may be extremely important for future decisions.
- Delay in data capture could be detrimental and lead to erroneous decisions involving extreme risk and substantial financial losses.

Data stream applications

- Sectors where real-time analytics is important:
- News services which increasingly rely on social media to detect breaking news events.
- Intelligence agencies tasked with the goal of preventing terrorism.
- Automated financial trading where algorithms rely on having immediate access to information.
- Weather dependent industries where reliable forecasts can be used to drive greater efficiencies.

Back-testing

- Demonstrating the value of data analytics is non-trivial.
- An evaluation protocol is required which ensures that the performance statistics are an appropriate estimate of what might be expected from an application.
- It is therefore important to use different data for estimation of parameters and evaluation of performance.

Walk-forward approach

- A first step is to walk forward in time and to evaluate the performance of the analytics in helping to support making decisions.
- This reduces the risk of using future data which would not be available in reality.
- Basically, we use a time-machine to go backwards and replay time.

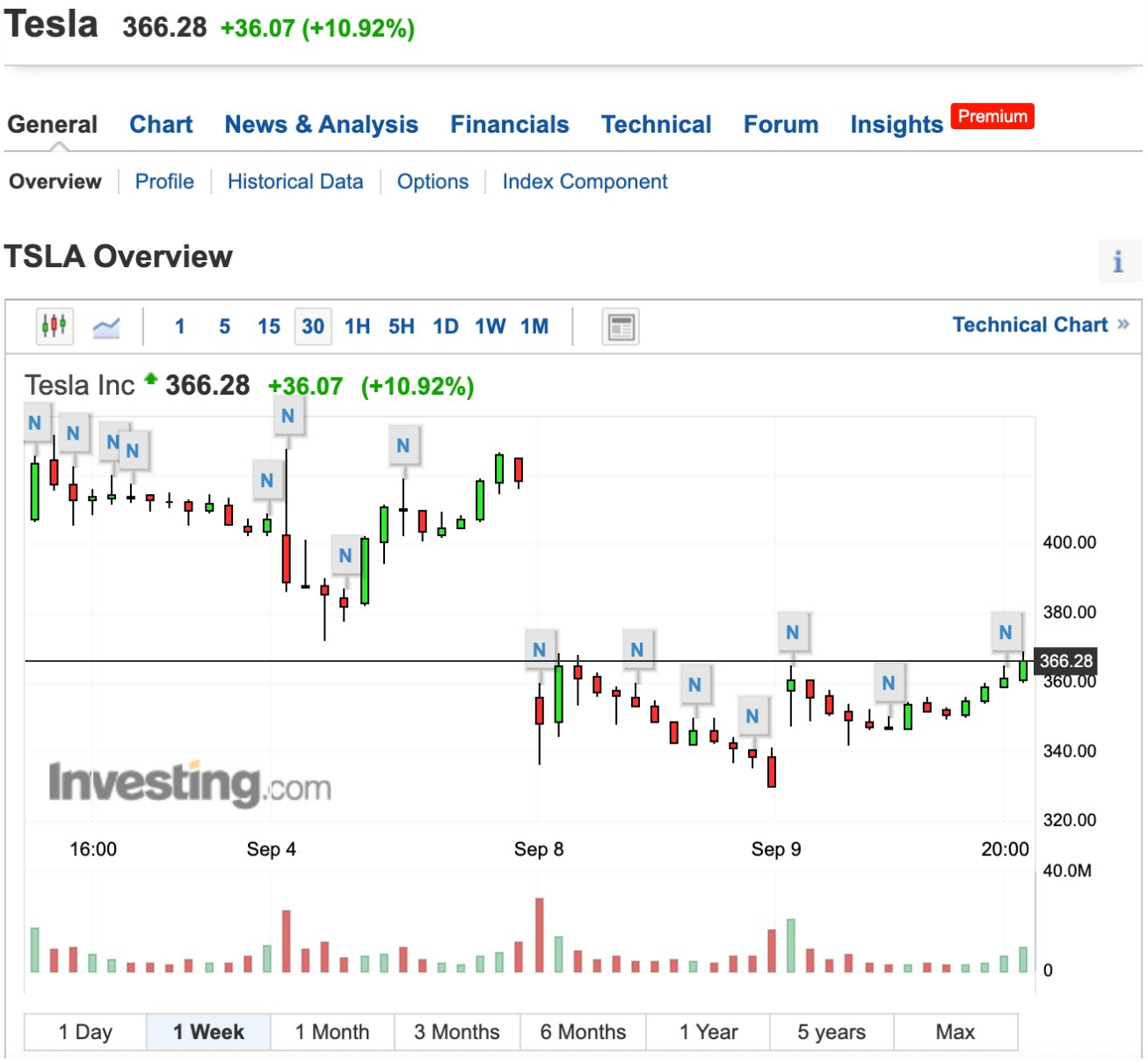
Time stamps

- Because of the importance of back-testing, we need to have accurate time stamps.
- Official statistics are extremely problematic in this sense as economic figures are often revised.
- Worse still, sometimes figures for a given month are not available until the following month.
- A common mistake in back-testing is to use data that would not have actually been available.
- The model estimated at time t can only use data that was available at time t.

Synchronizing Data

- When using multiple data sources, we need to ensure that the time stamps are properly aligned.
- Using data with daily time stamps could be dangerous if the application is sensitive to seeing information leaked from a few hours into the future.
- Similarly inaccuracies in monthly data could cause problems with estimates sensitive to daily timescales.

Observing and taking action



Developing Applications

- One quickly realizes that having access to multiple sources of data via an API with accurate time stamps is fundamental.
- The time spent in structuring data can be enormous.
- Worse still it is possible to waste a lot of time doing this in a static mode and to fail to address the need to make the entire process work for data streams.

Quandl

- Quandl hosts data from hundreds of publishers on a single easy-to-use website.
- Quandl makes it easy for data users to get the data they need in the format they want.
- In many ways, Quandl provides the necessary first step in undertaking scalable data analytics.

Quandl offering

- Quandl's core offering is free access to financial and economic data
- In addition, it offers actionable, profit-generating Insights from Alternative Data
- At Quandl, we source, evaluate and productize undiscovered data assets, transforming them into quantified, actionable intelligence for select institutional clients.

Quandl API

- www.quandl.com
- Quandl and Matlab:

https://github.com/quandl/Matlab

- Quandl and Python:

https://www.quandl.com/tools/python

- Quandl and R:

https://www.quandl.com/tools/r

Discussion on APIs

- What other innovations could be developed to make it easier to obtain datasets?
- What is your wish list for making APIs more usable?
- What would this require from developers ?
- What advice could we give policymakers?

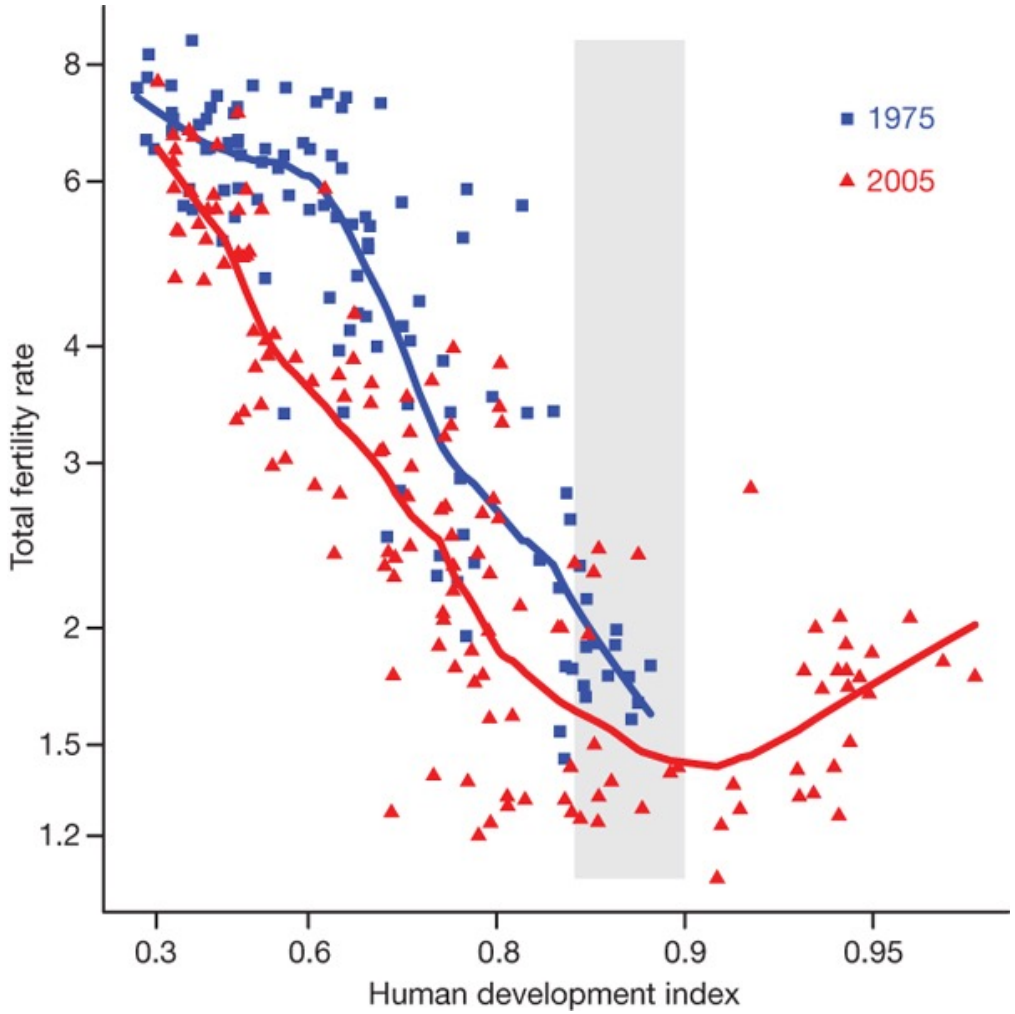
Discussion on Open Access

- How can we convince more institutions, governments and business of the advantages of open data policies?
- What are the rewards and how might they be quantified?
- What are the risks and how might these be minimized?

Class Poll 2

- Which of these would you prioritise?
  - Identifying AI opportunities
  - Mitigating AI risks
  - Buidling analytics capacity
  - Facilitating data access
  - Providing computing power
- **slido.com Event: #76755**

Fertility and HDI



Source: Myrskylä et al. (2009). Advances in development reverse fertility declines, Nature

Implications of J-shaped relationship

- During the twentieth century, the global population has gone through unprecedented increases in economic and social development that coincided with substantial declines in human fertility and population growth rates.
- The previously negative development–fertility relationship has become J-shaped, with the HDI being positively associated with fertility among highly developed countries.
- This reversal of fertility decline as a result of continued economic and social development has the potential to slow the rates of population ageing, thereby ameliorating the social and economic problems that have been associated with the emergence and persistence of very low fertility.

Quality of life

- Economist Intelligence Unit calculates a quality of life index (now called the “where to be born index”.
- This attempts to measure the best opportunities for a healthy, safe and prosperous life in the years ahead.
- The method links the results of subjective life-satisfaction surveys to the objective determinants of quality of life across countries.
- It also has a forward-looking element.

Linking survey results and data

- Survey to administer questionaire in order to measure the quality of life: y
- Surveys are expensive and take time so can we generate predictions instead?
- Find alternative data to use as explanatory variables $x_1, \ldots, x_n$
- Construct a predictive model:

$$y = w_1x_1 + \ldots + w_nx_n$$

Quality of Life

- Material well-being as measured by GDP per capita (in $, at 2006 constant PPPS)
- Life expectancy at birth
- The quality of family life based primarily on divorce rates
- The state of political freedoms
- Job security (measured by the unemployment rate)
- Climate (measured by two variables: the average deviation of minimum and maximum monthly temperatures from 14 degrees Celsius; and the number of months in the year with less than 30mm rainfall)
- Personal physical security ratings (based primarily on recorded homicide rates and ratings for risk from crime and terrorism)
- Quality of community life (based on membership in social organisations)
- Governance (measured by ratings for corruption)
- Gender equality (measured by the share of seats in parliament held by women)

Class Poll 3

- Quality of life viewed as best opportunities for a healthy, safe and prosperous life
- What would you use to measure quality of life?

- **slido.com Event: #76755**

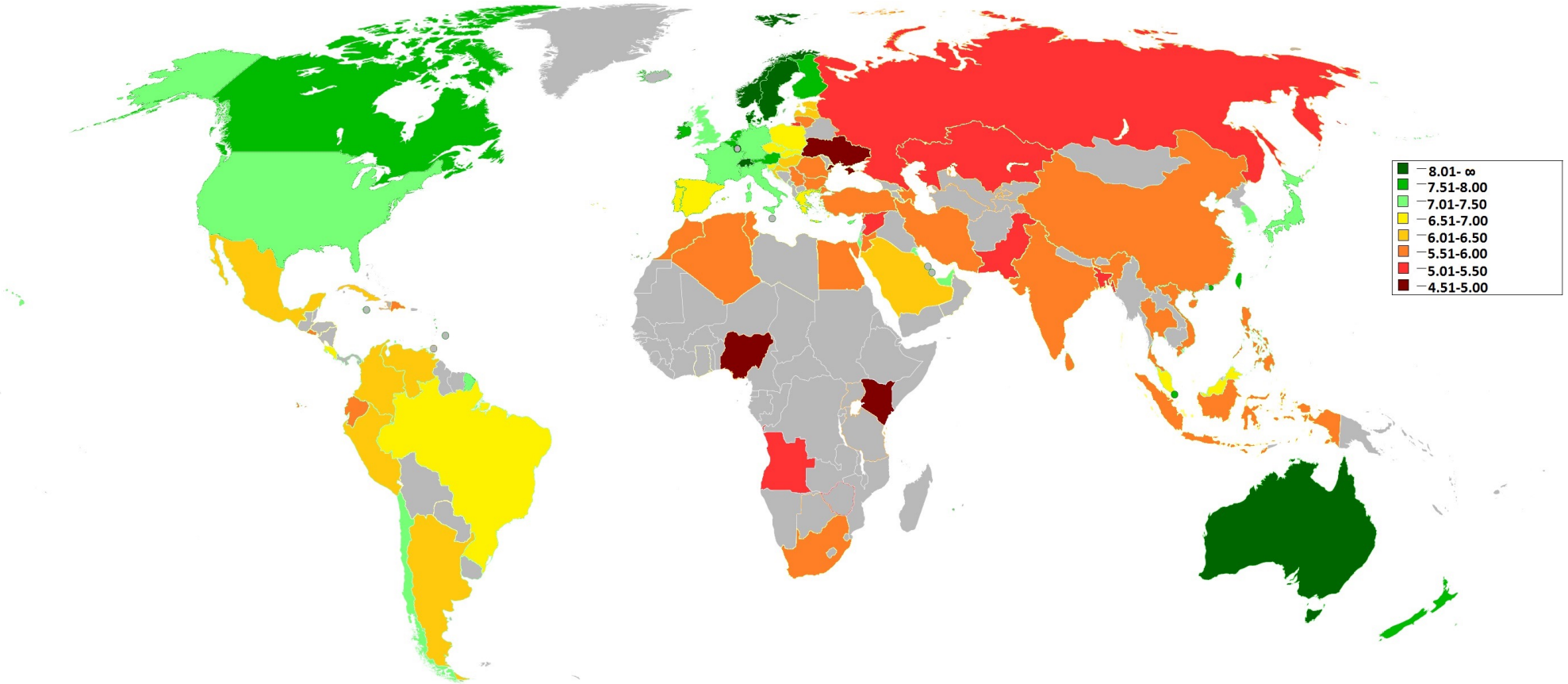
The where-to-be-born index, 2013

Rank	Country	Score*	Rank	Country	Score*	Rank	Country	Score*	Rank	Country	Score*
1	Switzerland	8.22	21	Italy	7.21	=40	Cuba	6.39	61	Bulgaria	5.73
2	Australia	8.12	22	Kuwait	7.18	42	Colombia	6.27	62	El Salvador	5.72
3	Norway	8.09	=23	Chile	7.10	43	Peru	6.24	=63	Philippines	5.71
4	Sweden	8.02	=23	Cyprus	7.10	=44	Estonia	6.07	=63	Sri Lanka	5.71
5	Denmark	8.01	25	Japan	7.08	=44	Venezuela	6.07	65	Ecuador	5.70
6	Singapore	8.00	26	France	7.04	=46	Croatia	6.06	=66	India	5.67
7	New Zealand	7.95	27	Britain	7.01	=46	Hungary	6.06	=66	Morocco	5.67
8	Netherlands	7.94	=28	Czech Rep.	6.96	48	Latvia	6.01	68	Vietnam	5.64
9	Canada	7.81	=28	Spain	6.96	49	China	5.99	69	Jordan	5.63
10	Hong Kong	7.80	=30	Costa Rica	6.92	50	Thailand	5.96	70	Azerbaijan	5.60
11	Finland	7.76	=30	Portugal	6.92	51	Turkey	5.95	71	Indonesia	5.54
12	Ireland	7.74	32	Slovenia	6.77	52	Dominican Rep.	5.93	72	Russia	5.31
13	Austria	7.73	33	Poland	6.66	53	South Africa	5.89	73	Syria	5.29
14	Taiwan	7.67	34	Greece	6.65	=54	Algeria	5.86	74	Kazakhstan	5.20
15	Belgium	7.51	35	Slovakia	6.64	=54	Serbia	5.86	75	Pakistan	5.17
=16	Germany	7.38	36	Malaysia	6.62	56	Romania	5.85	76	Angola	5.09
=16	United States	7.38	37	Brazil	6.52	57	Lithuania	5.82	77	Bangladesh	5.07
18	U.A.E.	7.33	38	Saudi Arabia	6.49	58	Iran	5.78	78	Ukraine	4.98
19	South Korea	7.25	39	Mexico	6.41	59	Tunisia	5.77	79	Kenya	4.91
20	Israel	7.23	=40	Argentina	6.39	60	Egypt	5.76	80	Nigeria	4.74

Source: Economist Intelligence Unit

*Score out of a maximum of 10

Where to be born



Discussion on Indices

- Where to be born
- Where to grow up
- Where to work
- Where to retire
- Any others?

Exploring data

- Graphing data in one dimension
- Scatter plots, Recurrence plots;
- Combining plots
- Multiple plots with lines, colors, markers
- Labels and presentation
- Printing figures

Matlab functions

- linspace
- plot, plotyy, scatter
- semilogx, semilogy, loglog,
- figure, subplot
- xlabel, ylabel,
- title, text, legend
- datetick

Q&A