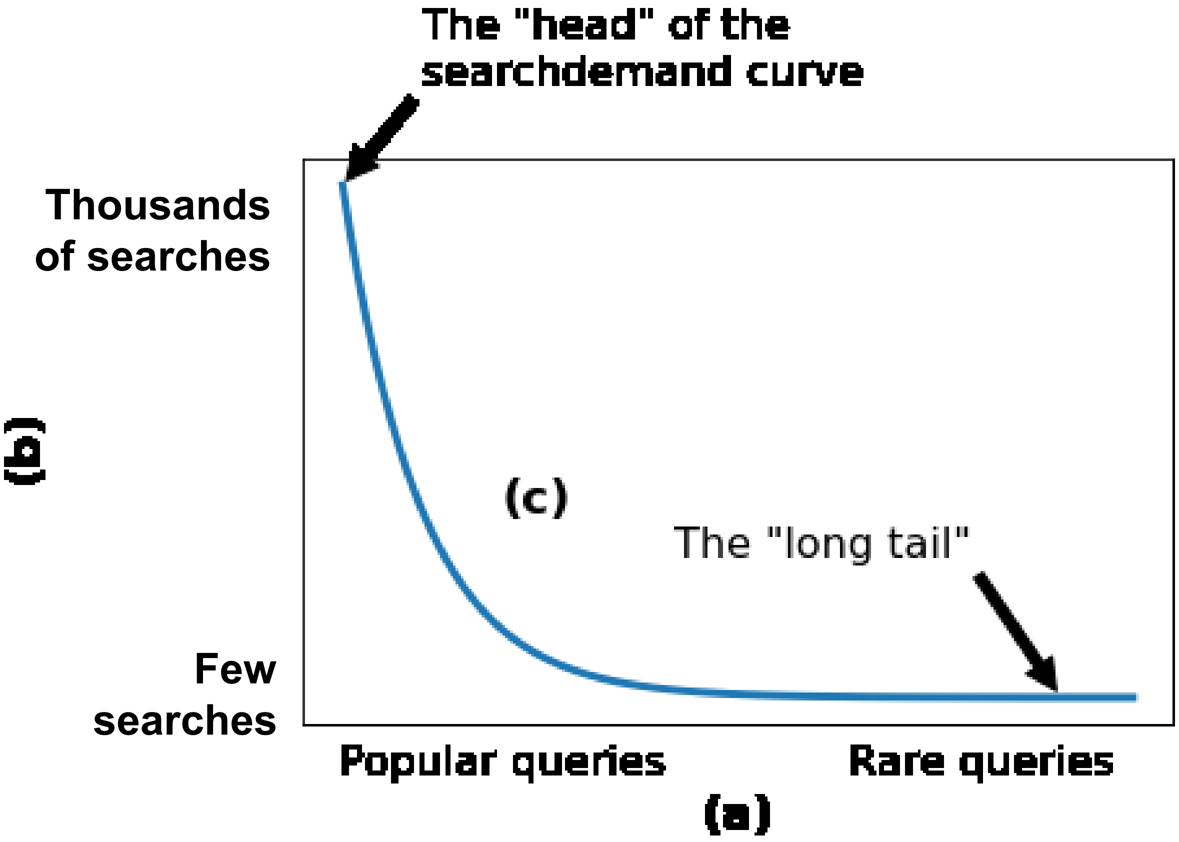
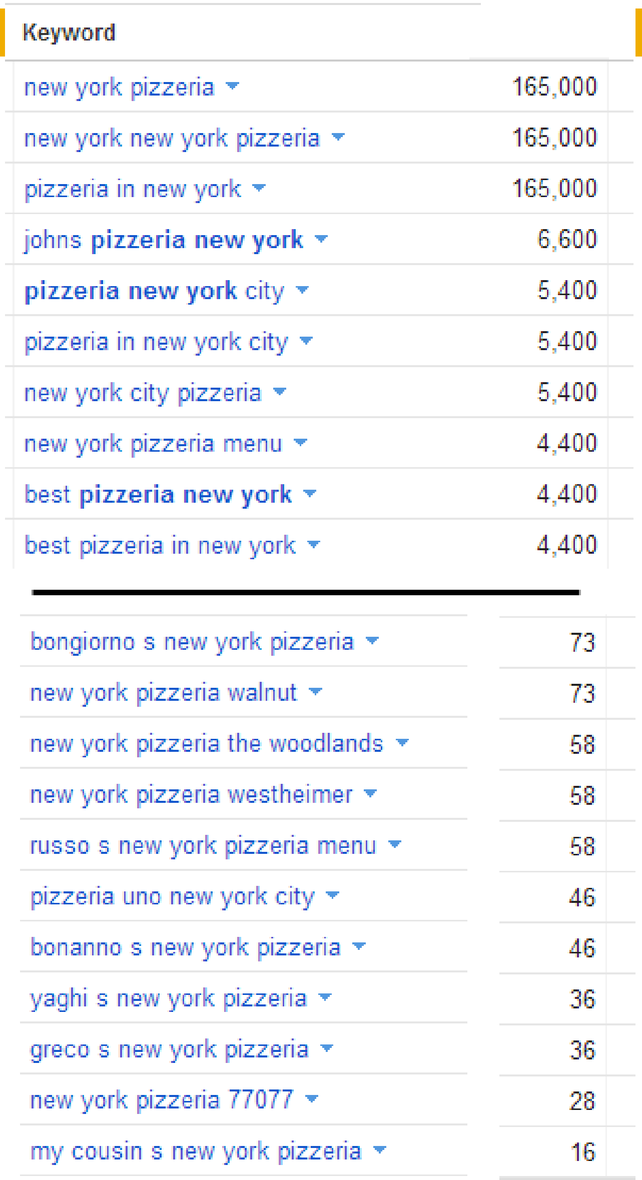
STATISTICAL METHODS FOR DATA SCIENCE

Project 2018/19:
Total Volume of Queries to Google

Data Science and Business Informatics Degree

The long-tail of query search





Keyword	
new york pizzeria	165,000
new york new york pizzeria	165,000
pizzeria in new york	165,000
johns **pizzeria new york**	6,600
pizzeria new york city	5,400
pizzeria in new york city	5,400
new york city pizzeria	5,400
new york pizzeria menu	4,400
best **pizzeria new york**	4,400
best pizzeria in new york	4,400

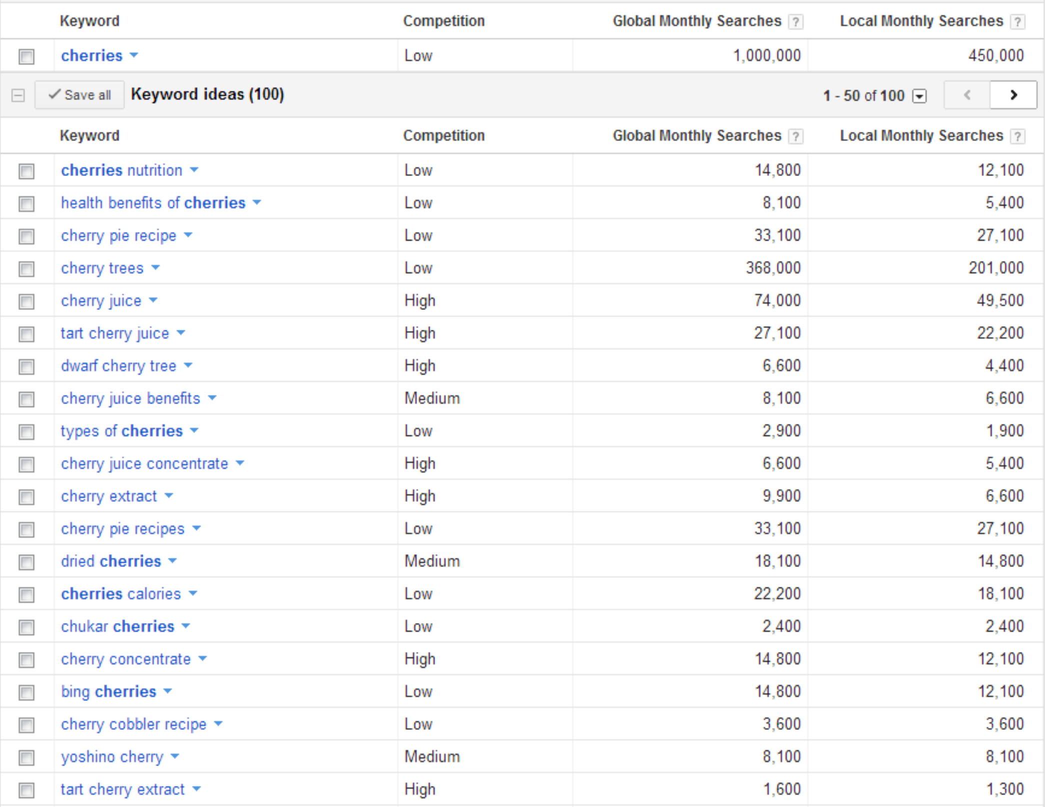
bongiorno s new york pizzeria	73
new york pizzeria walnut	73
new york pizzeria the woodlands	58
new york pizzeria westheimer	58
russo s new york pizzeria menu	58
pizzeria uno new york city	46
bonanno s new york pizzeria	46
yaghi s new york pizzeria	36
greco s new york pizzeria	36
new york pizzeria 77077	28
my cousin s new york pizzeria	16

Volume of a query

- Volume of a query
  - = frequency of searches = number of times the query is searched
- Who knows the TRUE volume of a query?
  - no one, not even Google (see later)
- Keyword research tools
  - They collect data on searches … from various sources
    - Actually, only Google/SE can do it.
    - Others use clickstream data or meta-search engines.
  - … and offer keyword research analytics
    - query search volume
    - related queries

Keyword volume/suggestion

- Google AdWords/Ads keyword planner
  - adwords.google.com/ko/KeywordPlanner



	Keyword	Competition	Global Monthly Searches ?	Local Monthly Searches ?
□	**cherries**	Low	1,000,000	450,000

⊟ ✓ Save all **Keyword ideas (100)** 1 - 50 of 100 ‹ ›

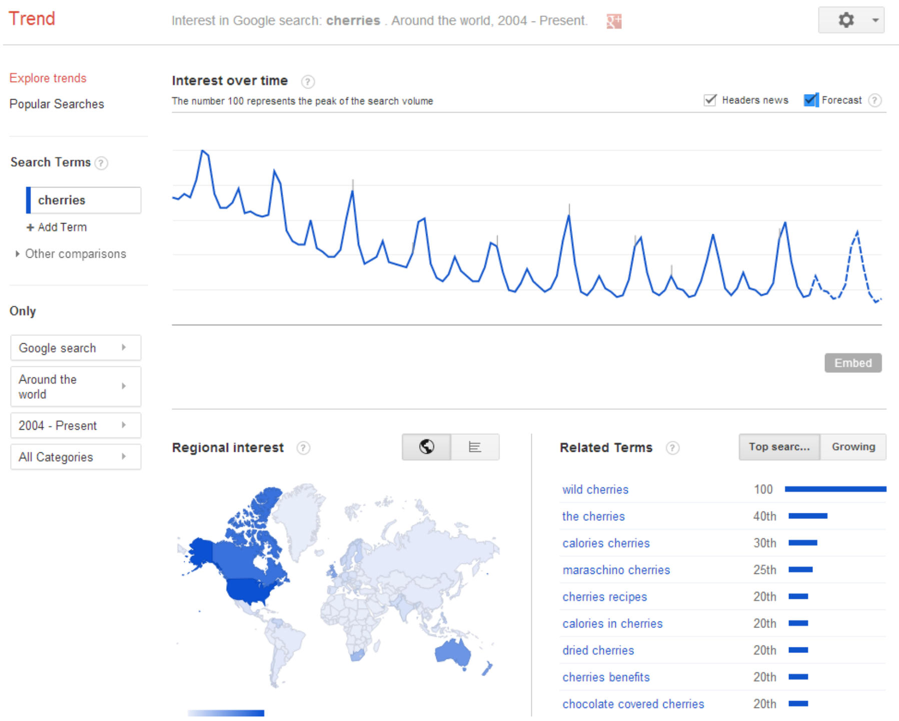
	Keyword	Competition	Global Monthly Searches ?	Local Monthly Searches ?
□	**cherries** nutrition	Low	14,800	12,100
□	health benefits of **cherries**	Low	8,100	5,400
□	cherry pie recipe	Low	33,100	27,100
□	cherry trees	Low	368,000	201,000
□	cherry juice	High	74,000	49,500
□	tart cherry juice	High	27,100	22,200
□	dwarf cherry tree	High	6,600	4,400
□	cherry juice benefits	Medium	8,100	6,600
□	types of **cherries**	Low	2,900	1,900
□	cherry juice concentrate	High	6,600	5,400
□	cherry extract	High	9,900	6,600
□	cherry pie recipes	Low	33,100	27,100
□	dried **cherries**	Medium	18,100	14,800
□	**cherries** calories	Low	22,200	18,100
□	chukar **cherries**	Low	2,400	2,400
□	cherry concentrate	High	14,800	12,100
□	bing **cherries**	Low	14,800	12,100
□	cherry cobbler recipe	Low	3,600	3,600
□	yoshino cherry	Medium	8,100	8,100
□	tart cherry extract	High	1,600	1,300

Keyword volume/suggestion

- Free version online:
  - KeywordKeg https://keywordkeg.com/
  - Ubersuggest https://neilpatel.com/ubersuggest
  - SearchVolume https://searchvolume.io/
  - Answerthepublic https://answerthepublic.com
  - Keyword shitter http://keywordshitter.com
- Commercial systems
  - Hrefs https://ahrefs.com
  - Semrush keyword research https://www.semrush.com/features/keyword-research
  - Keywordtool.io https://keywordtool.io

Keywords tools: trending searches

- Google trends
  - www.google.com/trends



Trend

Interest in Google search: **cherries** . Around the world, 2004 - Present.

Explore trends

Popular Searches

Interest over time

The number 100 represents the peak of the search volume

☑ Headers news ☑ Forecast

Search Terms

cherries

+ Add Term

Other comparisons

Only

Google search

Around the world

2004 - Present

All Categories

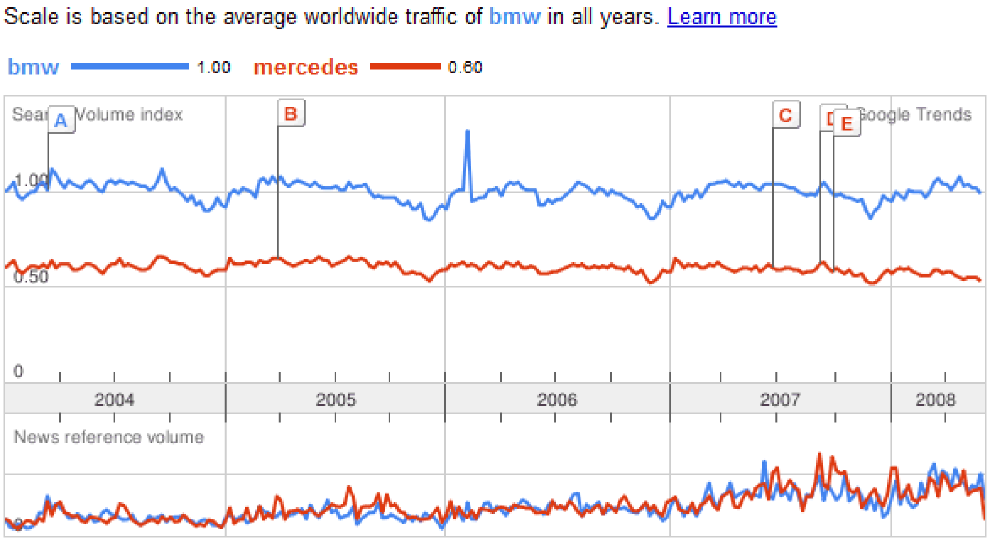
Embed

Regional interest

Related Terms

Top searc... | Growing

wild cherries	100
the cherries	40th
calories cherries	30th
maraschino cherries	25th
cherries recipes	20th
calories in cherries	20th
dried cherries	20th
cherries benefits	20th
chocolate covered cherries	20th



Scale is based on the average worldwide traffic of bmw in all years. Learn more

Volume of a query

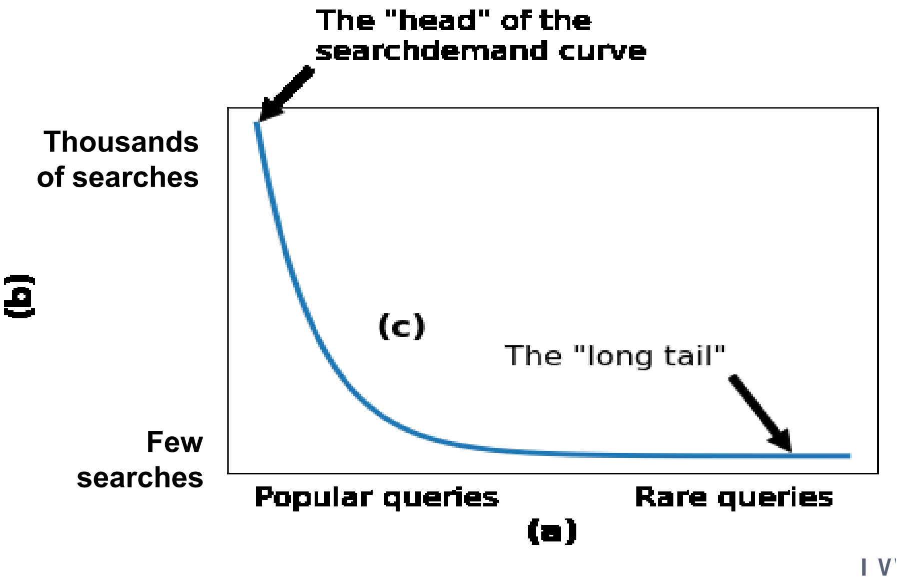
- Who knows the **TRUE** volume of a query?
  - no one, not even Google
- Why not?
  - Too complex to keep counts of all distinct queries
    - Approximated counting: count-min sketches
      - V = true volume V_0 = largest volume $\varepsilon \sim U(0, 10^{-3})$
      - approximated volume $\mathbf{X = V + \varepsilon V_0}$
  - Estimation by keyword research tools may also introduce approximations
    - How to model such an approximation?
  - Estimation by keyword research tools may produce discrete values (range), not continuous values
    - How to model such an setting?

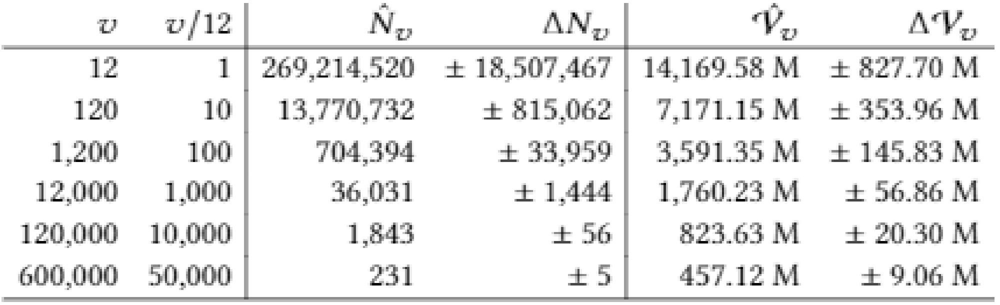
Collection of all queries

- Who knows the **TRUE** set of queries for a reference population?
  - Population = domain + time + geo
  - E.g., Italian searches of recipes in 2017
- No one
  - At best, we can construct a sample of queries of the population
  - Problem: how to do the sampling?

Project: Objective

- Given a sample of a population
  - recipes & cooking in Italian searches in 2017
  - Premier League in UK searches in last 12 months
- Estimate the number and total volume of queries in the population, for queries with at least v searches





v	$v/12$	$\hat{N}_v$	ΔN_v	$\hat{\mathcal{V}}_v$	$\Delta\mathcal{V}_v$
12	1	269,214,520	± 18,507,467	14,169.58 M	± 827.70 M
120	10	13,770,732	± 815,062	7,171.15 M	± 353.96 M
1,200	100	704,394	± 33,959	3,591.35 M	± 145.83 M
12,000	1,000	36,031	± 1,444	1,760.23 M	± 56.86 M
120,000	10,000	1,843	± 56	823.63 M	± 20.30 M
600,000	50,000	231	± 5	457.12 M	± 9.06 M

Available data

- Population: recipes & cooking in Italian searches in 2017
- Sample of 121K queries
- Volume estimation for a subset of them:
  - 10.7K using Ubersuggest, 1628 using KeywordKeg, 11K using SearchVolume, 1950 using SemRush
    - Absolute volume
    - Discrete volume: ranges
  - 121K using Google Trends (But only 18.5K are non-zero)
    - Relative volume: 1.0 is set arbitrarily
    - Question: can we estimate a scale from other datasets?

New data (all groups contribute)

- Population: Premier League in UK searches in last 12 months
- We have to generate sample queries
  - How to do?
  - See top web sites: https://serpstat.com/keywords/competitors/?search_type=keyword&query=premier%20league&se=g_uk
  - See lists of teams, players, coaches, ...
  - Combine the above
    - results of team A vs team B
    - referee errors in manchester vs liverpool
    - ...
- And then to collect query volume estimates
  - From Google Trends & others tools

Project: Steps

- Students work in non-competitive groups
- Collect new data – can be done in parallel with the following other steps
- Assume a statistical model of the distribution
  - Consider the continuous and/or the discrete case
- Check the model is consistent with the sample data
- Design estimators of parameters of the model
- Test estimators experimentally
- Use the model for answering the project objective

Shared Google Drive directory

- Data for recipes & cooking
- Relevant scientific papers