Indexing Compressed Data for Fast Retrieval

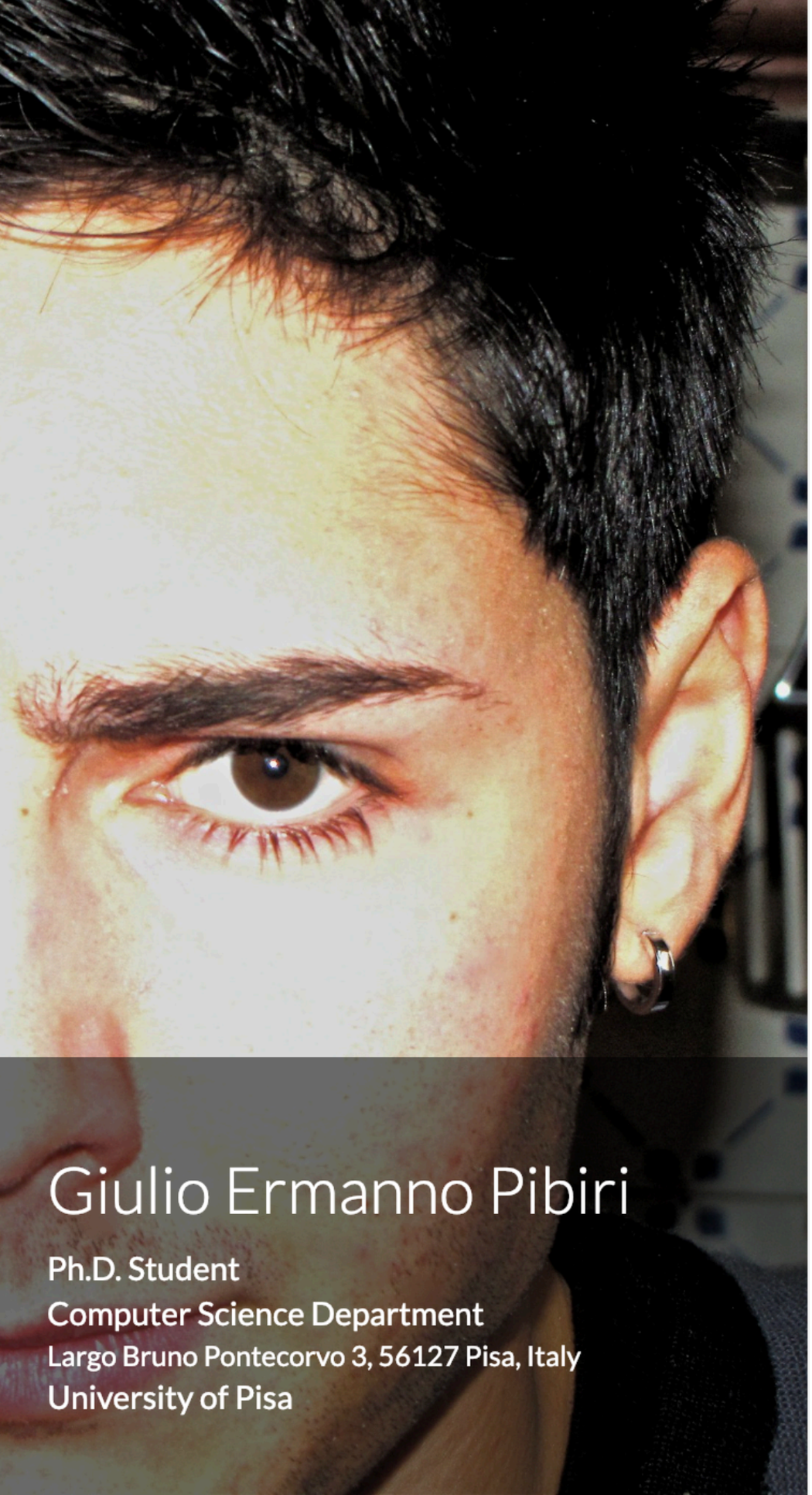
Giulio Ermanno Pibiri

giulio.pibiri@di.unipi.it

http://pages.di.unipi.it/pibiri

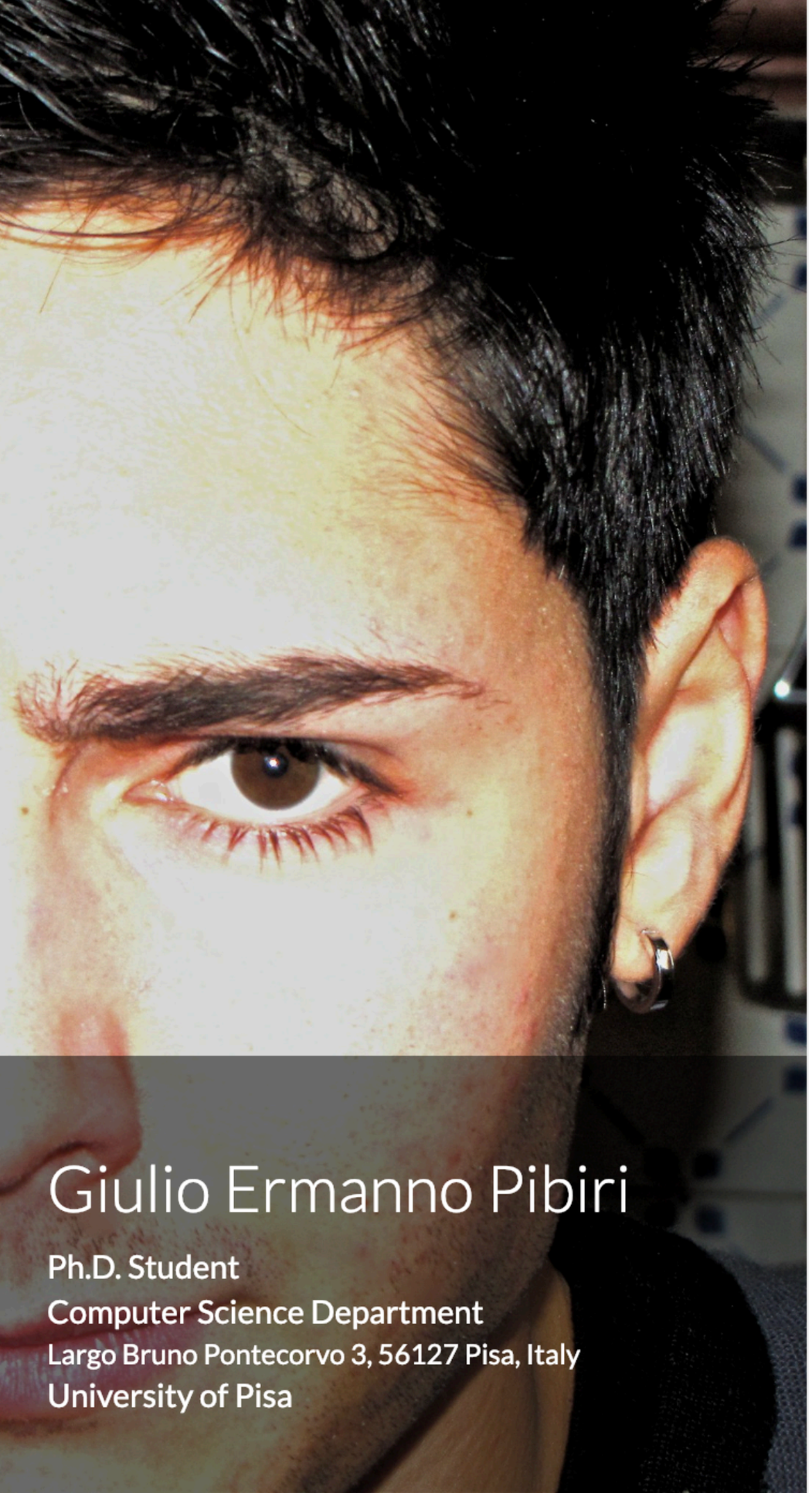
Department of Computer Science
University of Pisa

01/02/2019



Giulio Ermanno Pibiri

Ph.D. Student
Computer Science Department
Largo Bruno Pontecorvo 3, 56127 Pisa, Italy
University of Pisa



Giulio Ermanno Pibiri

Ph.D. Student
Computer Science Department
Largo Bruno Pontecorvo 3, 56127 Pisa, Italy
University of Pisa

Ph.D. Thesis

Space and Time-Efficient Data Structures for Massive Datasets

by

Giulio Ermanno Pibiri

Supervisor

Rossano Venturini

Referee

Daniel Lemire

Referee

Simon Gog

2018

Giulio Ermanno Pibiri

Ph.D. Student

Computer Science Department

Largo Bruno Pontecorvo 3, 56127 Pisa, Italy

University of Pisa

Ph.D. Thesis

Space and Time-Efficient Data Structures for Massive Datasets

by

Giulio Ermanno Pibiri

Supervisor

Rossano Venturini

Referee

Daniel Lemire

Referee

Simon Gog

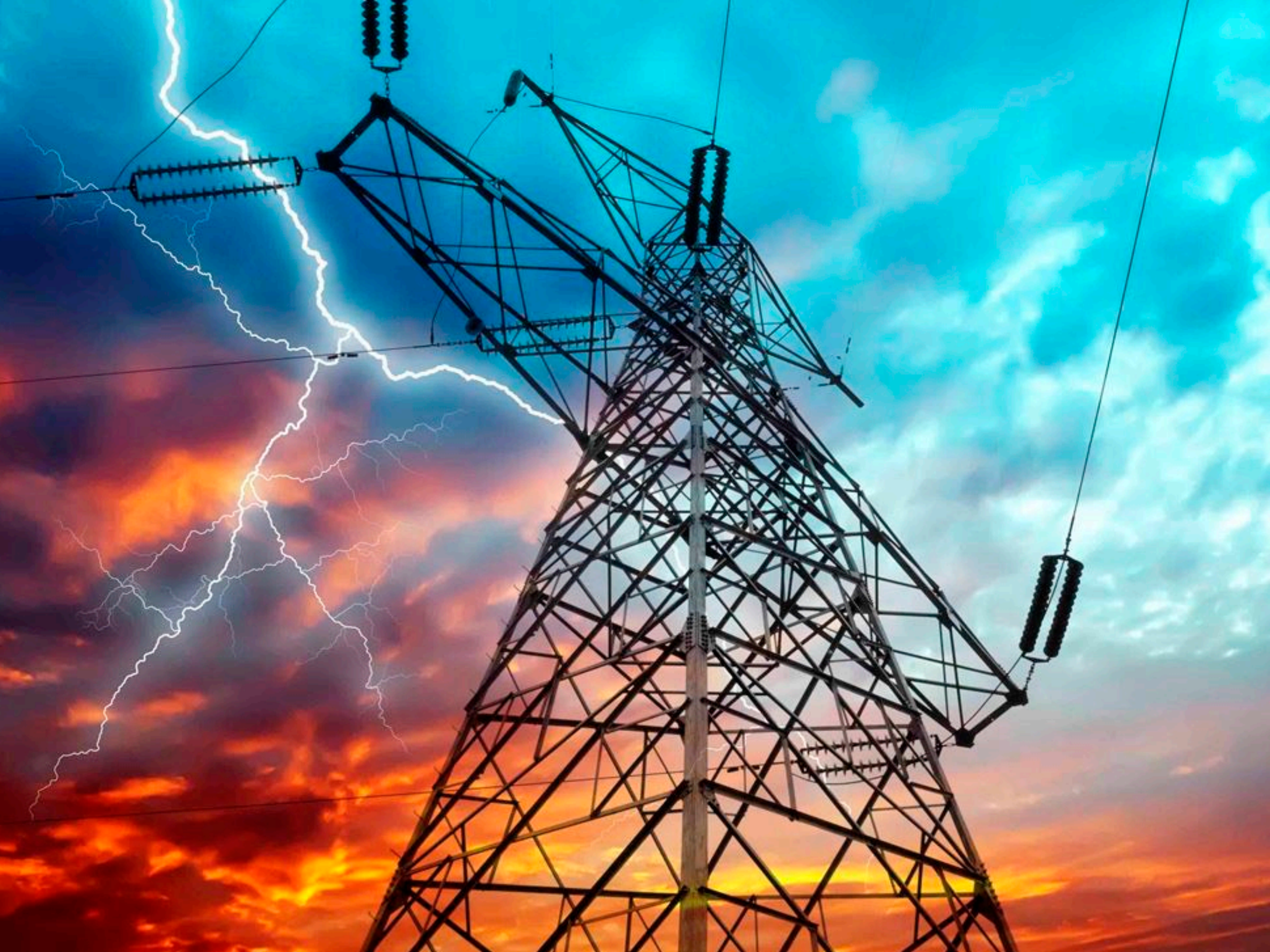
2018

HPC laboratory

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Raffele Perego


Statistics
Progress
Statistical
Analysis



tutorial c++

All Videos Images News Maps More Settings Tools

About 73,300,000 results (0.52 seconds)

C++ Language - C++ Tutorials - Cplusplus.com

www.cplusplus.com/doc/tutorial/

These **tutorials** explain the **C++** language from its basics up to the newest features introduced by C++11. Chapters have a practical orientation, with example ...

Compilers · Structure of a program · Variables and types · Classes

Google

tutorial c++

All Videos Images News Maps More Settings Tools

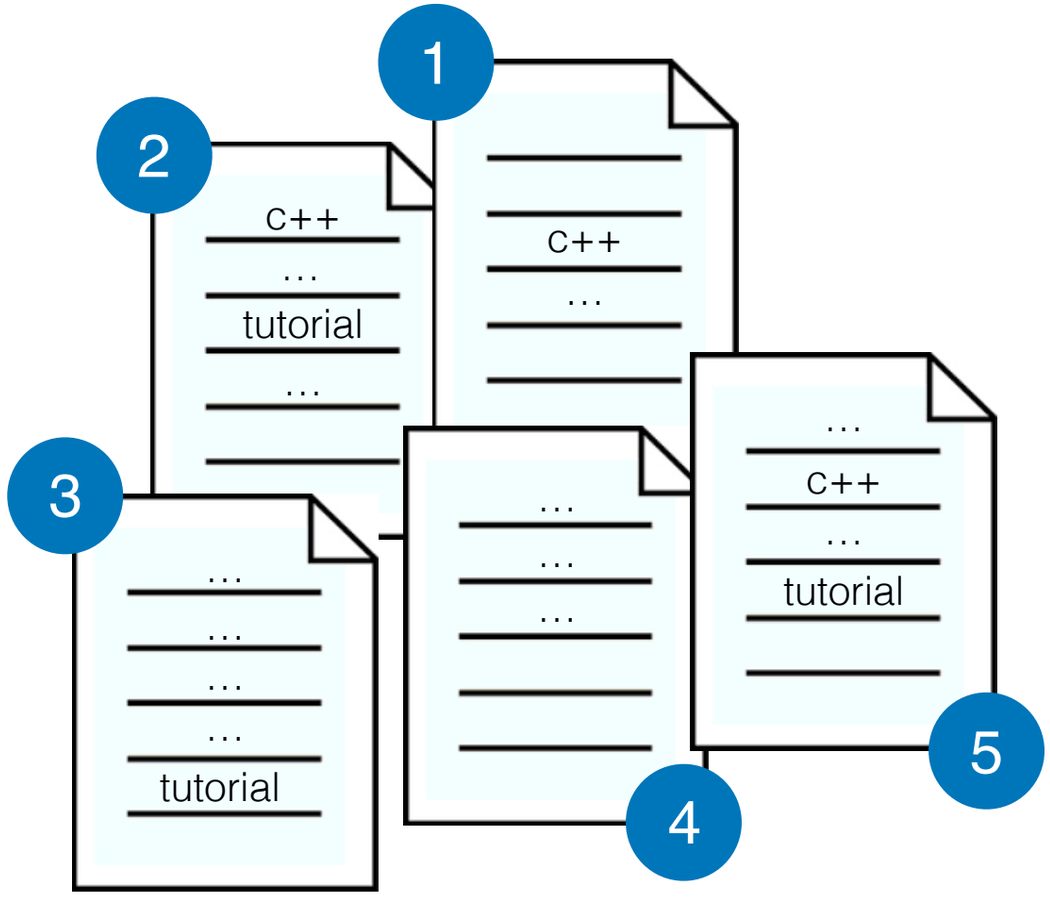
About 73,300,000 results (0.52 seconds)

C++ Language - C++ Tutorials - Cplusplus.com

www.cplusplus.com/doc/tutorial/

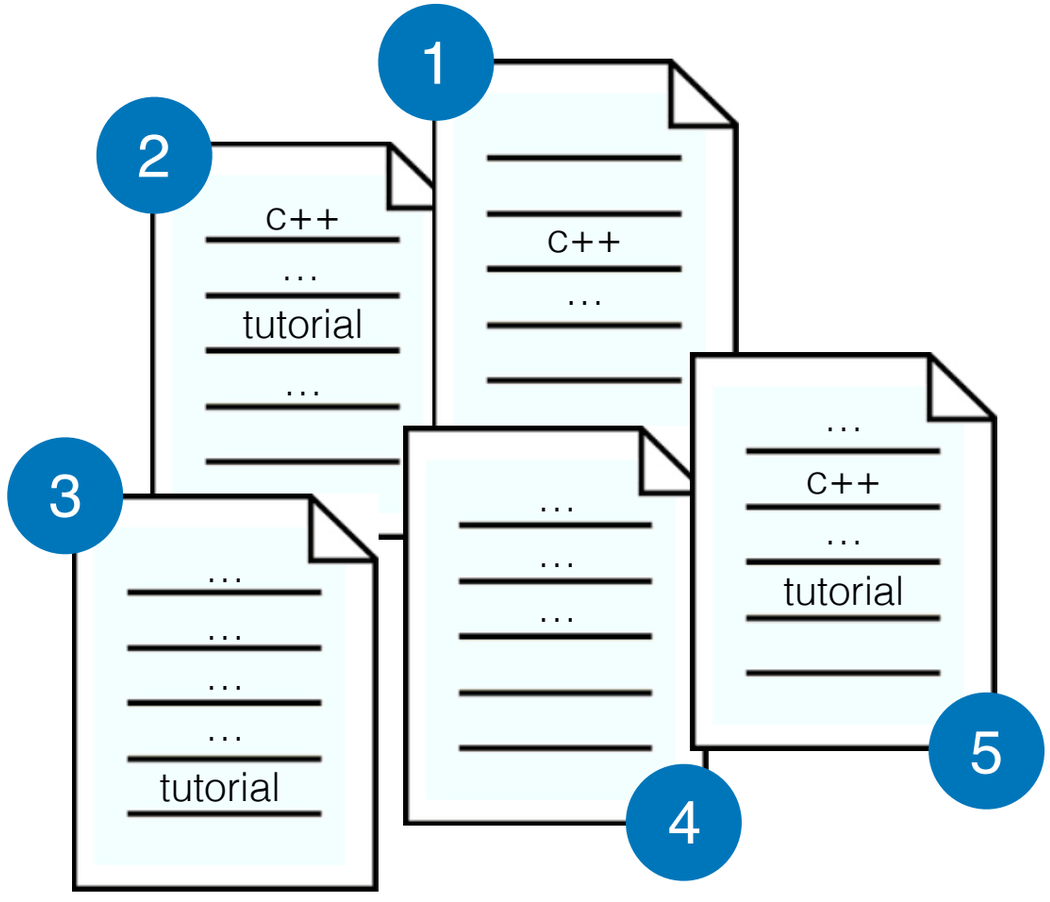
These **tutorials** explain the **C++** language from its basics up to the newest features introduced by C++11. Chapters have a practical orientation, with example ...

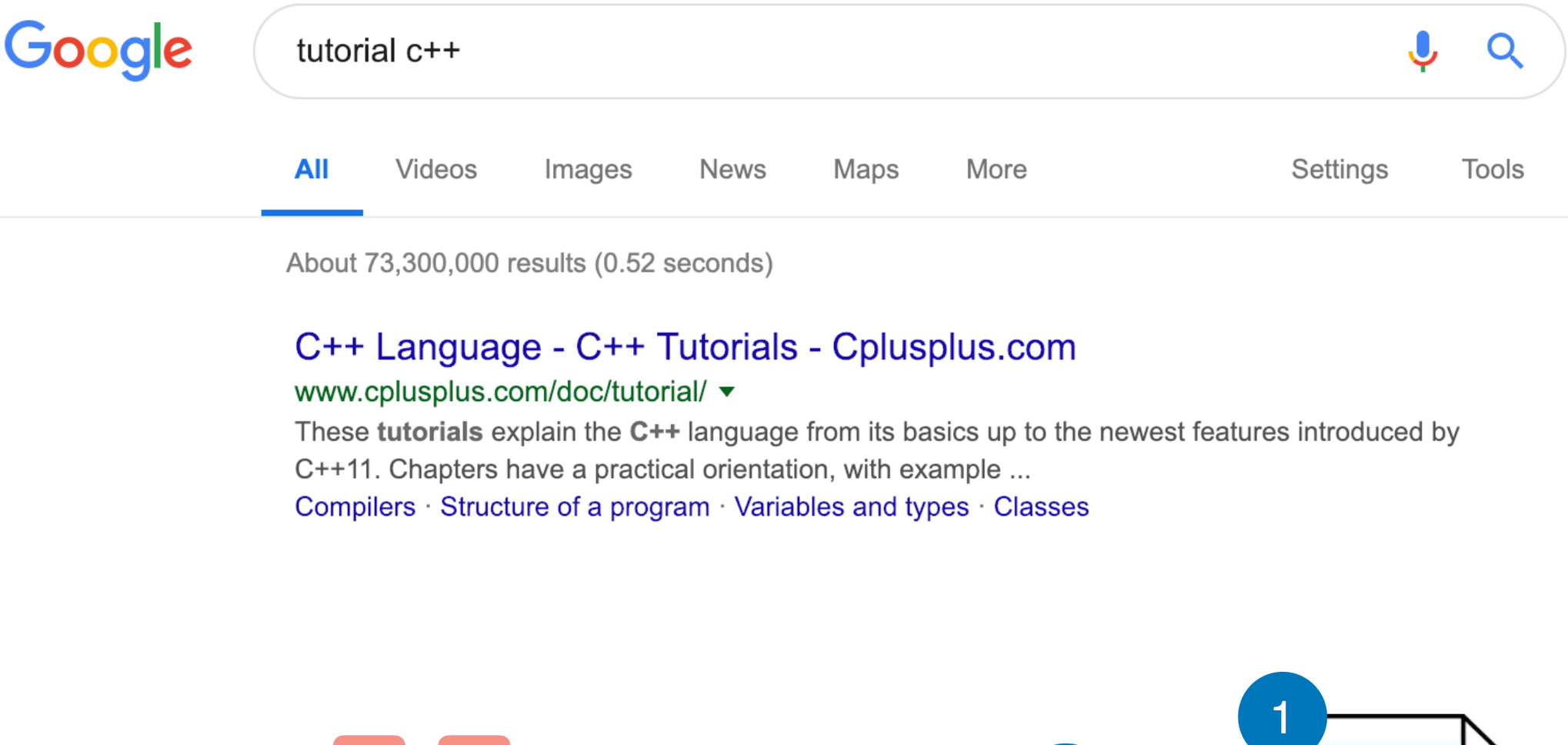
Compilers · Structure of a program · Variables and types · Classes



tutorial ⟶ **[2, 3, 5]**

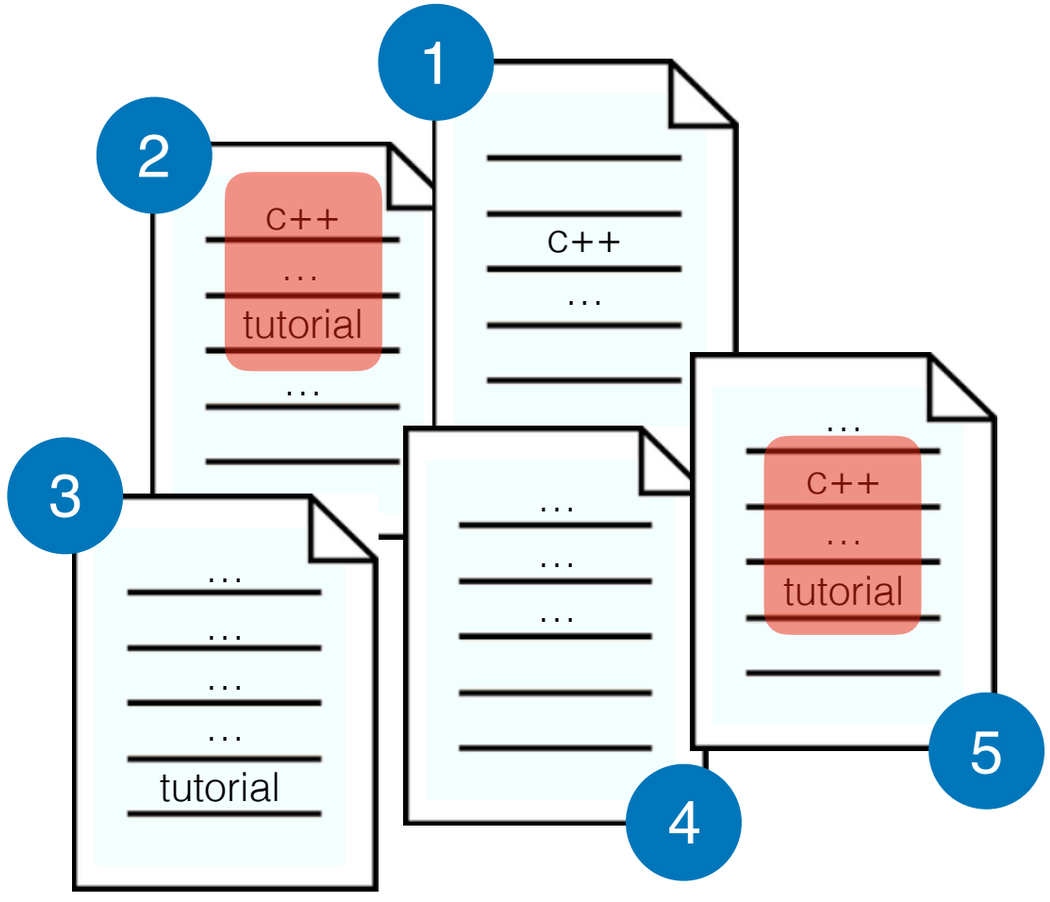
c++ ⟶ **[1, 2, 5]**

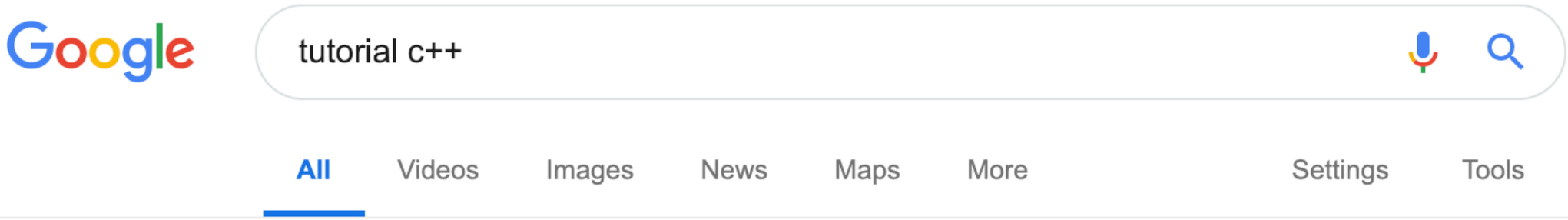




tutorial → **[2, 3, 5]**

c++ → **[1, 2, 5]**

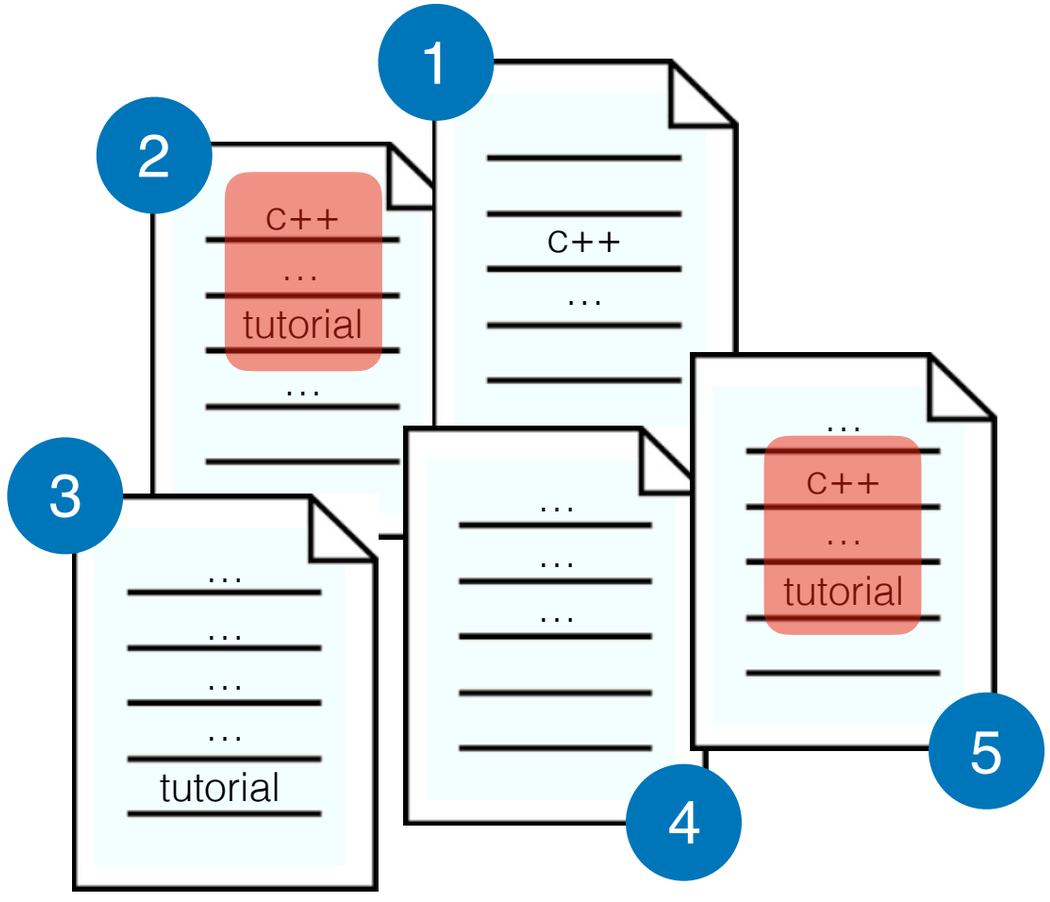


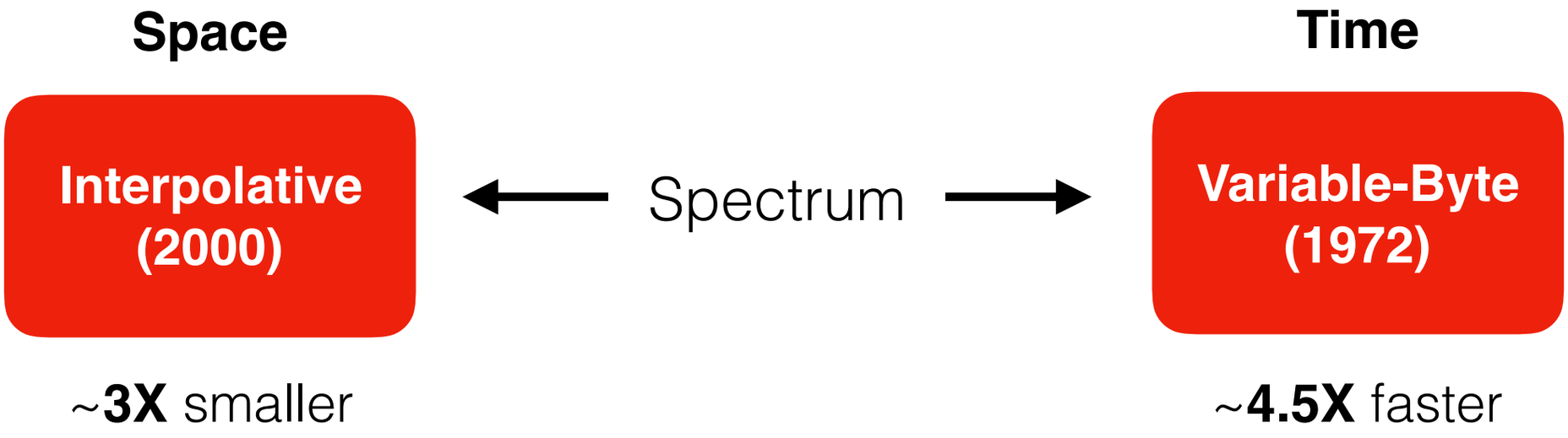


tutorial → [2, 3, 5]

c++ → [1, 2, 5]

Inverted Index




Space
Time
Interpolative (2000)
Spectrum
Variable-Byte (1972)
~3X smaller
~4.5X faster

Space

Time

Interpolative (2000) ⟵ Spectrum ⟶ **Variable-Byte (1972)**

~**3X** smaller

~**4.5X** faster

As small as Interpolative and **much faster**?

1

ACM TOIS 2017

Space

Time

Interpolative (2000) ⟵ Spectrum ⟶ **Variable-Byte (1972)**

~**3X** smaller

~**4.5X** faster

1

As small as Interpolative and **much faster**?

ACM TOIS 2017

2

As fast as Variable-Byte and **much smaller**?

IEEE TKDE 2019
(to appear)

Space

Interpolative (2000)

~**3X** smaller

Spectrum

Time

Variable-Byte (1972)

~**4.5X** faster

As small as Interpolative and **much faster**?

1

ACM TOIS 2017

As fast as Variable-Byte and **much smaller**?

2

IEEE TKDE 2019
(to appear)

What about **both** objectives at the same time?!

3

ACM WSDM 2019

Space

Interpolative (2000)

~**3X** smaller

Spectrum

Time

Variable-Byte (1972)

~**4.5X** faster

As small as Interpolative and **much faster**?

1

ACM TOIS 2017

As fast as Variable-Byte and **much smaller**?

2

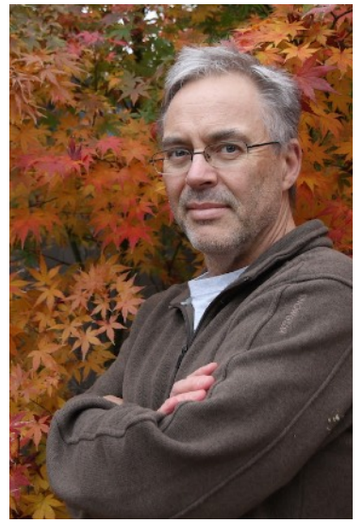
IEEE TKDE 2019
(to appear)

THE UNIVERSITY OF MELBOURNE

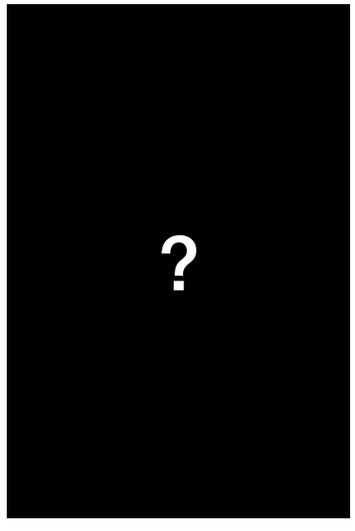
What about **both** objectives at the same time?!

3

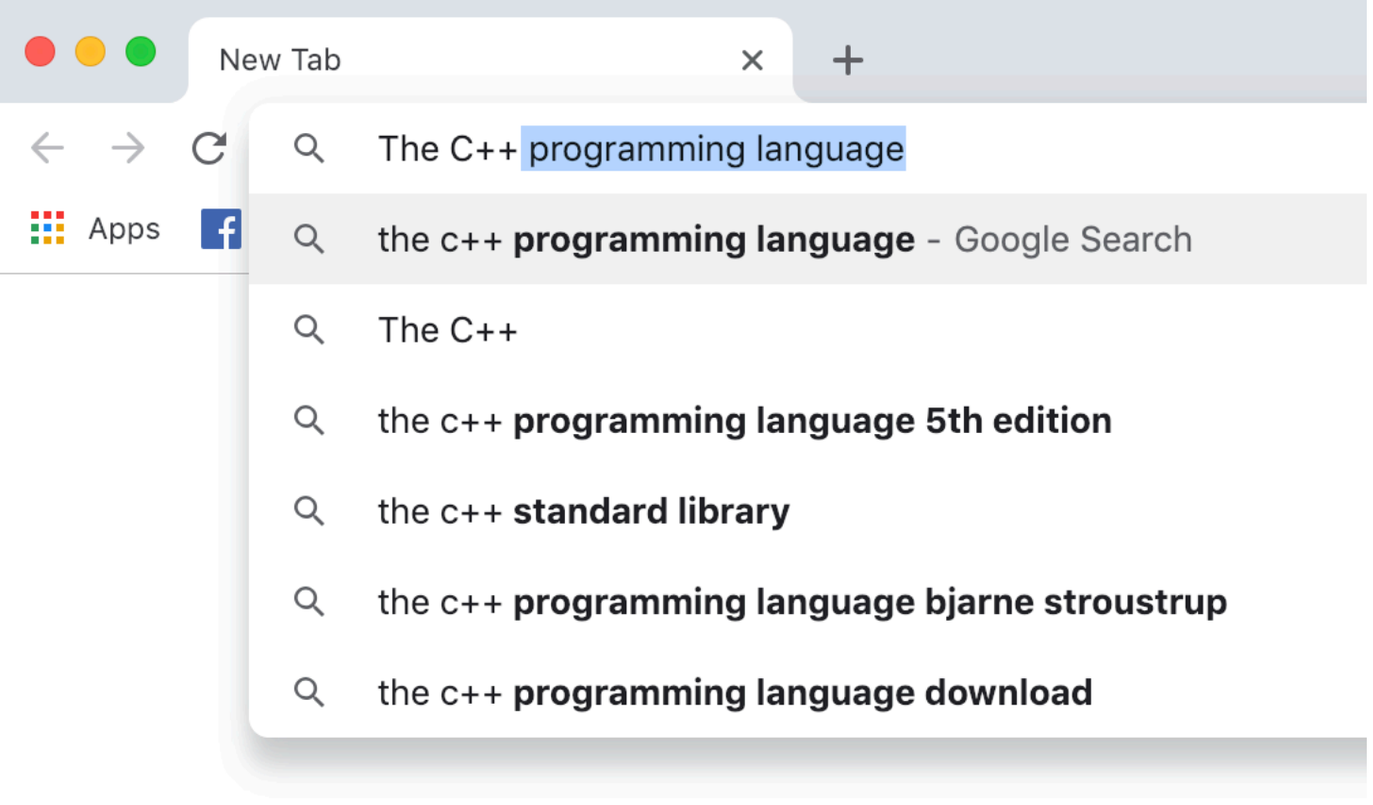
ACM WSDM 2019

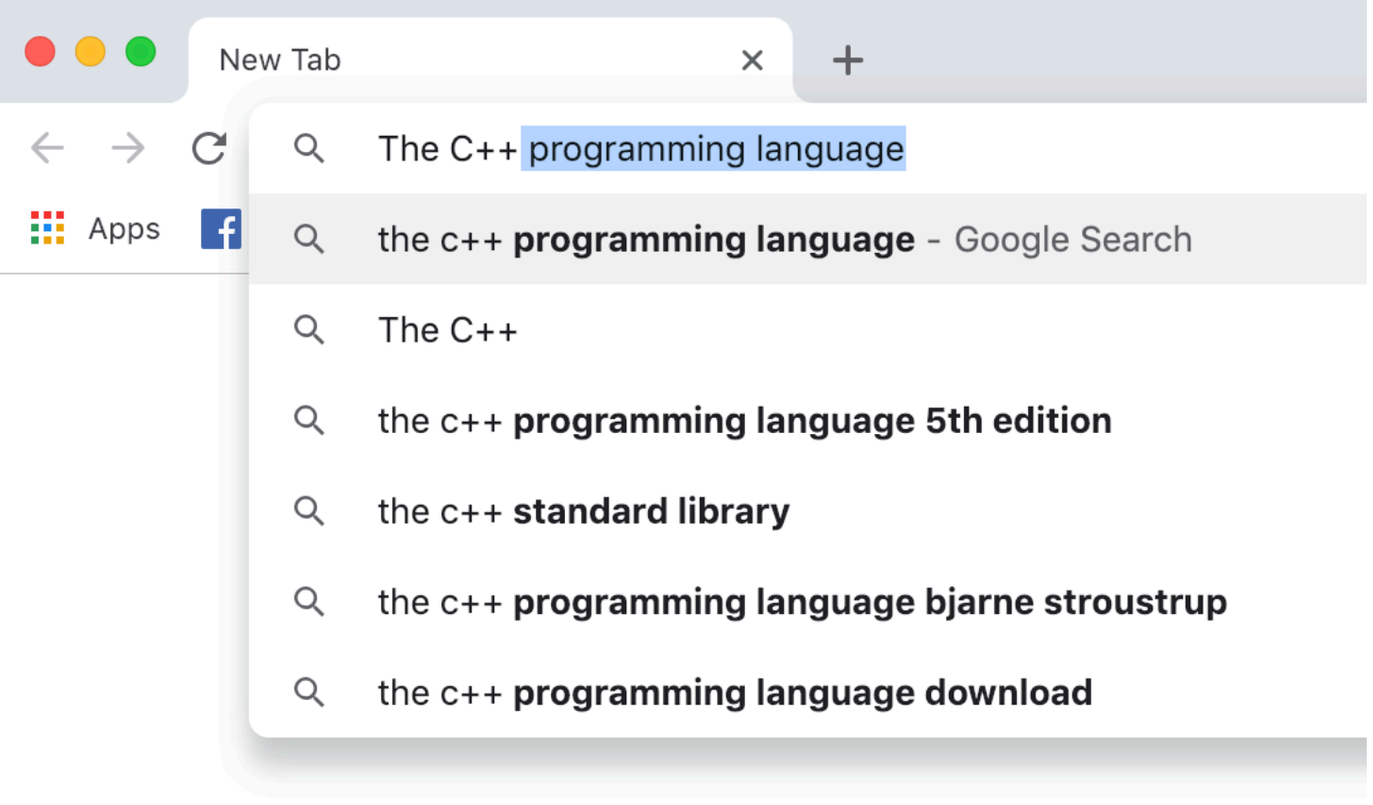


Alistair Moffat



Matthias Petri

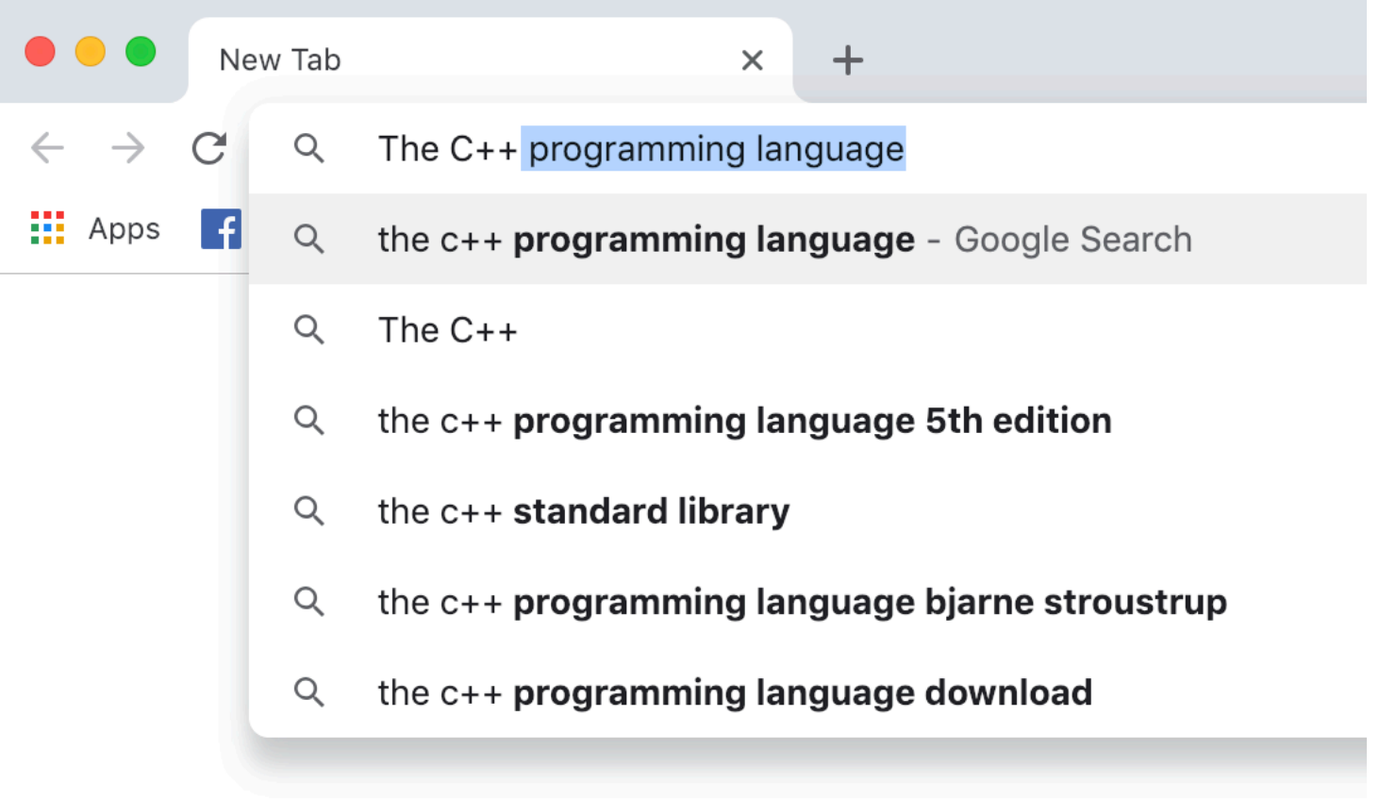

New Tab
Apps
The C++ programming language
the c++ programming language - Google Search
The C++
the c++ programming language 5th edition
the c++ standard library
the c++ programming language bjarne stroustrup
the c++ programming language download



N-gram lookup

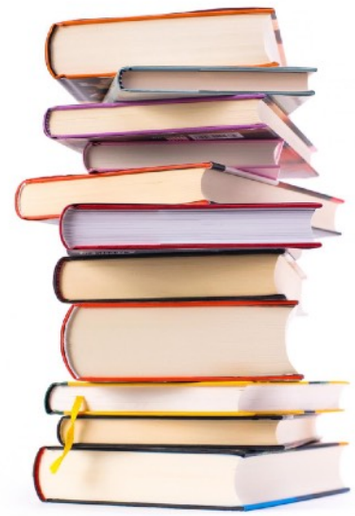
Hey Siri

Google Translate



Google Books

~6% of the books ever published
More than 11 billion N-grams!

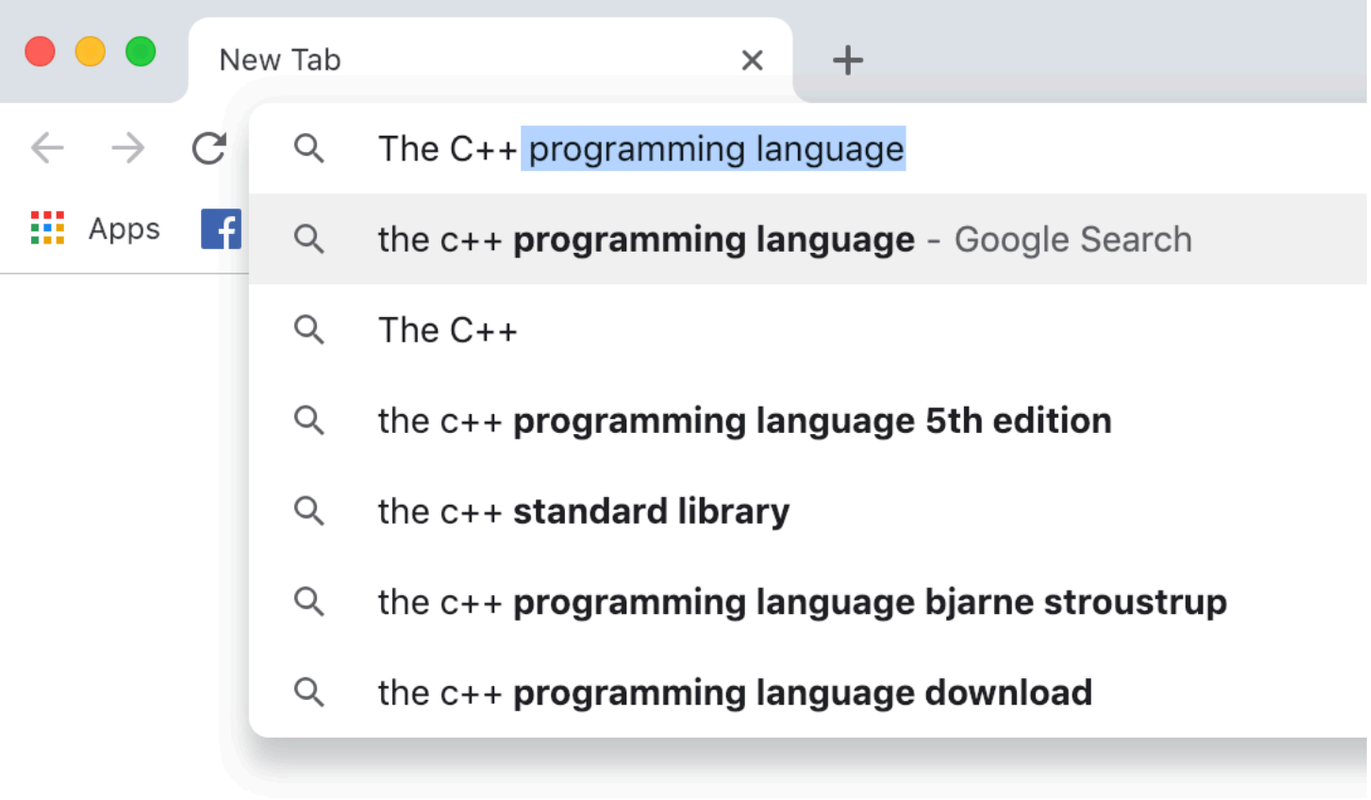


~58GB .gz

N-gram lookup

Hey Siri

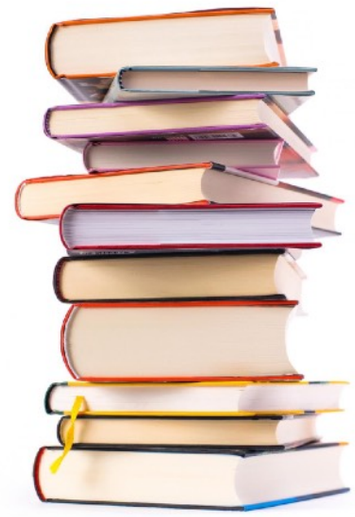
Google Translate



Google Books

~6% of the books ever published

More than 11 billion N-grams!



~58GB .gz

N-gram lookup

Hey Siri

Google Translate

ACM SIGIR 17

~29GB

Indexing?

1

New Tab

Apps

The C++ programming language

the c++ **programming language** - Google Search

The C++

the c++ **programming language 5th edition**

the c++ **standard library**

the c++ **programming language bjarne stroustrup**

the c++ **programming language download**

N-gram lookup

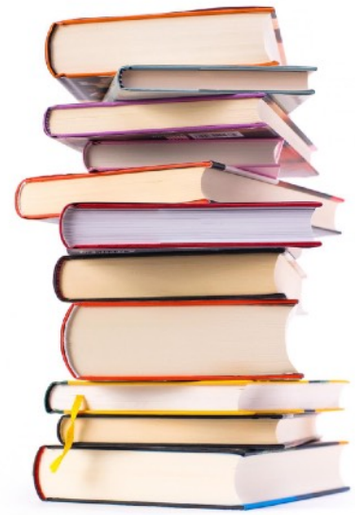
Hey Siri

Google Translate

Google Books

~6% of the books ever published

More than 11 billion N-grams!



~58GB .gz

ACM SIGIR 17

~29GB

Indexing? 1

ACM TOIS 19

(to appear)

Estimation? 2

Efficiency to deliver better services by using less resources.

Impact is far reaching and implies substantial economic gains.

Do not underestimate the impact of efficient software:
do not (just) rely on hardware.

My resources are publicly available:

http://pages.di.unipi.it/pibiri

https://github.com/jermp

Inverted indexes

Google YAHOO! bing

Databases

IBM Dropbox ORACLE

RDF indexing

ontotext

E-Commerce

ebay amazon

Geo-spatial data

Graph-compression

Thanks for your attention, time, patience!