UNIVERSITAS
GADJAH MADA

Metode akuisisi data (crawling)

Lukman Heryawan, PhD

29 Agustus 2022

Perkuliahan

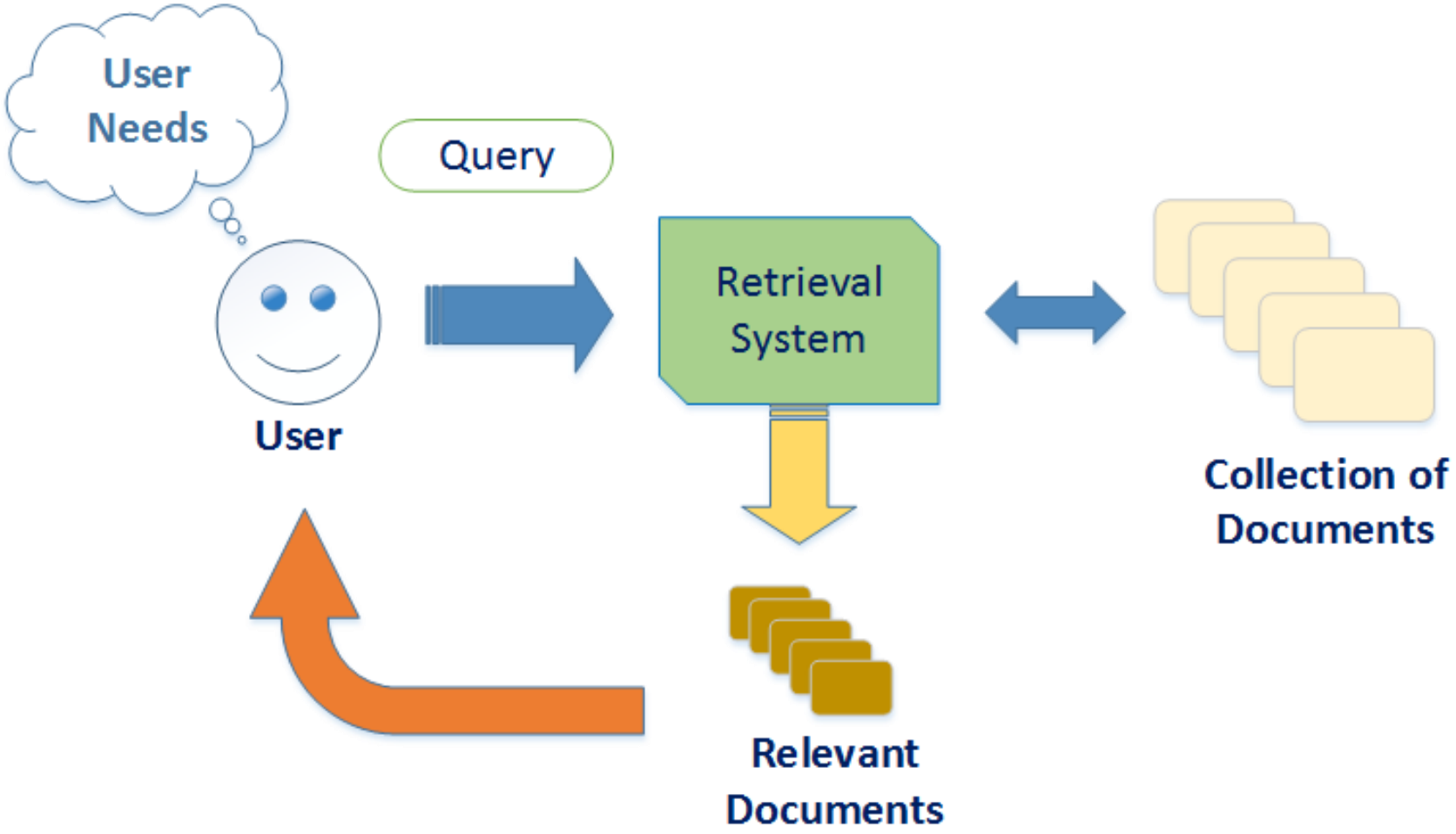
- Link elok: https://elok.ugm.ac.id/course/view.php?id=10789 (Sistem temu balik informasi)

- 7 kali pertemuan:
  - Definisi dan kategori Sistem Temu Kembali Informasi
  - Arsitektur dan komponen mesin pencari
  - Metode akuisisi data (crawling)
  - Metode transformasi data
  - Metode pembuatan indeks
  - Metode kompresi indeks
  - Rekap (demo+presentasi)
  - UTS
- Penilaian:
  - Tugas - 30%
  - UTS - 20%

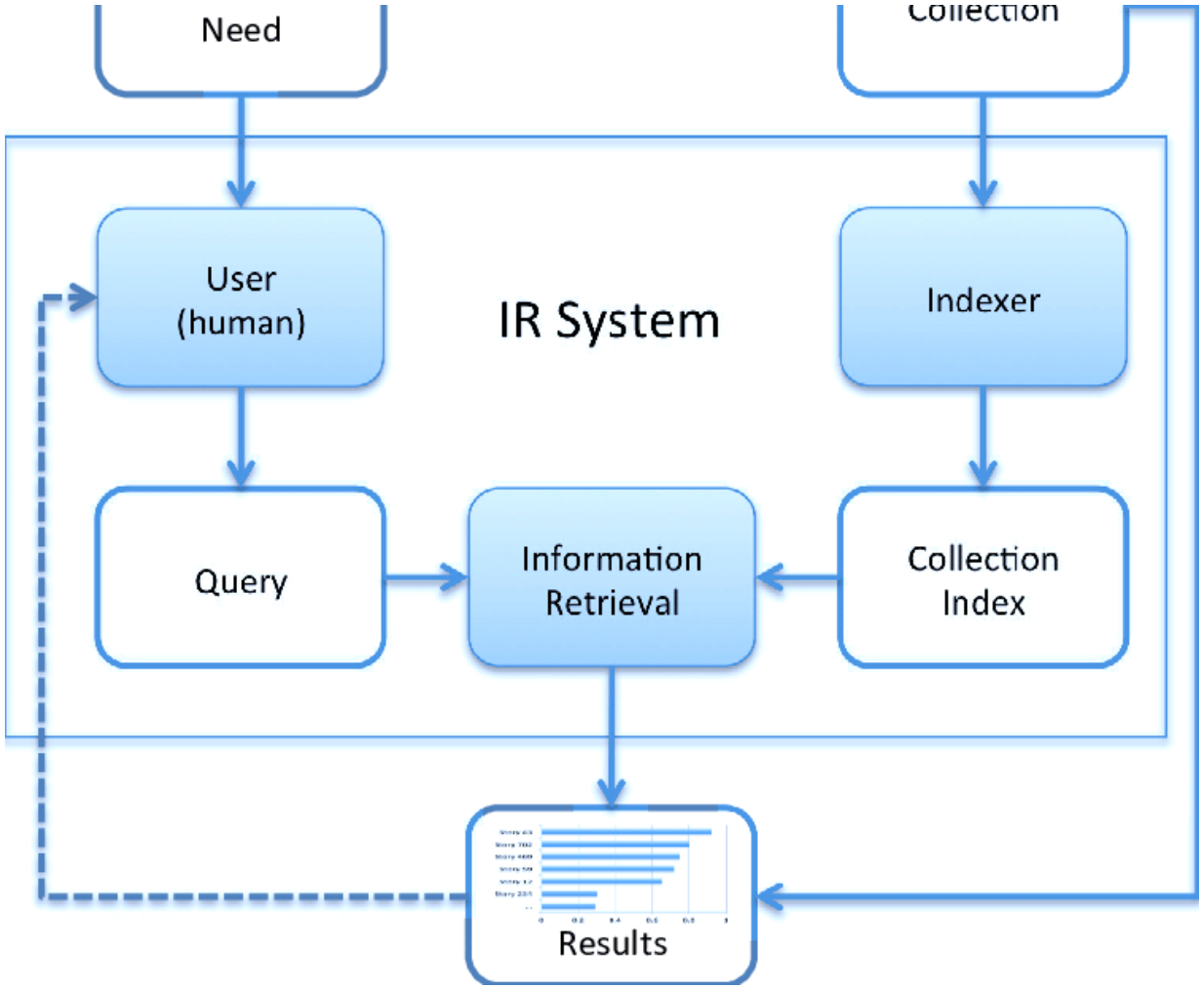
UNIVERSITAS
GADJAH MADA

Materi

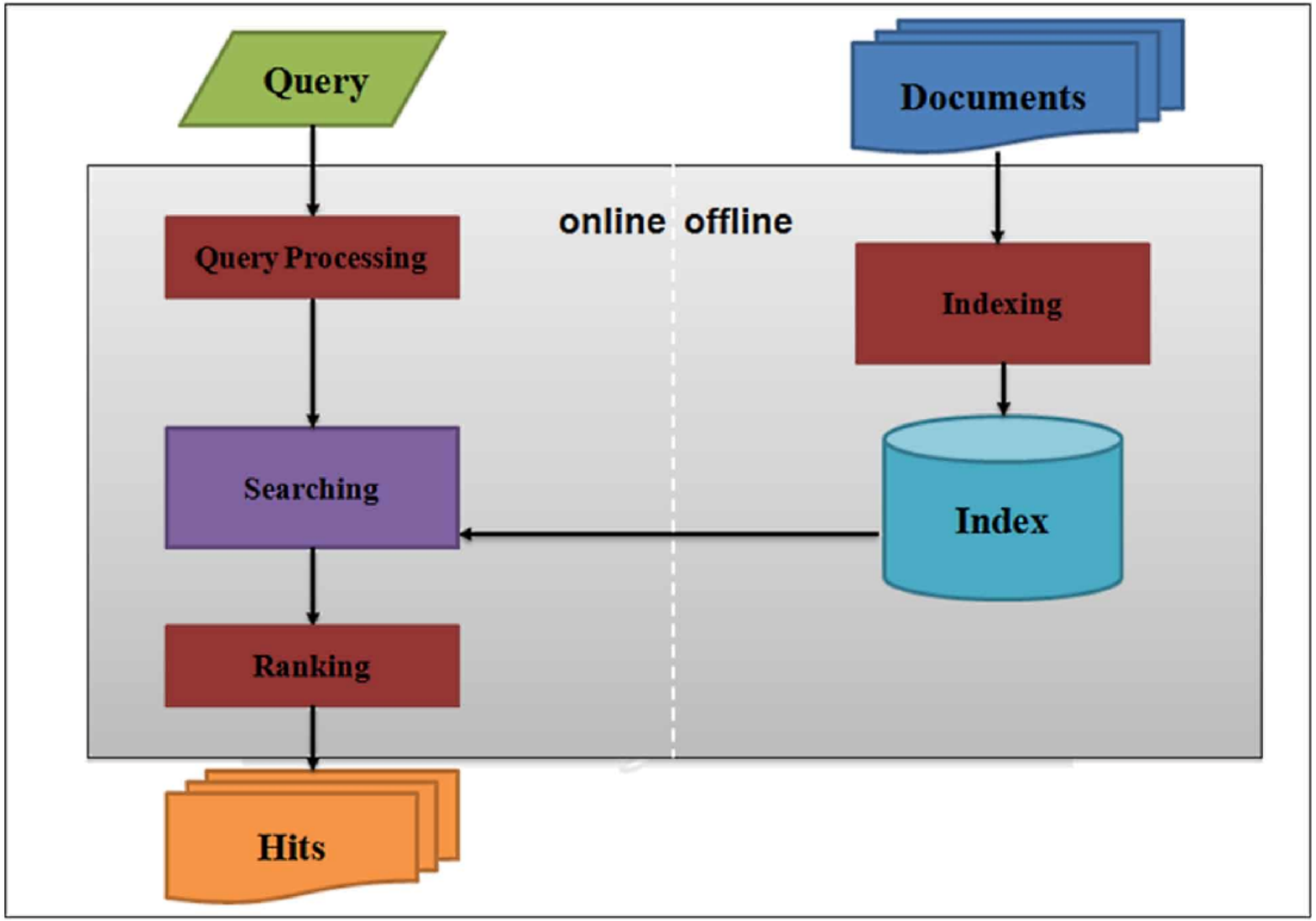
Information retrieval



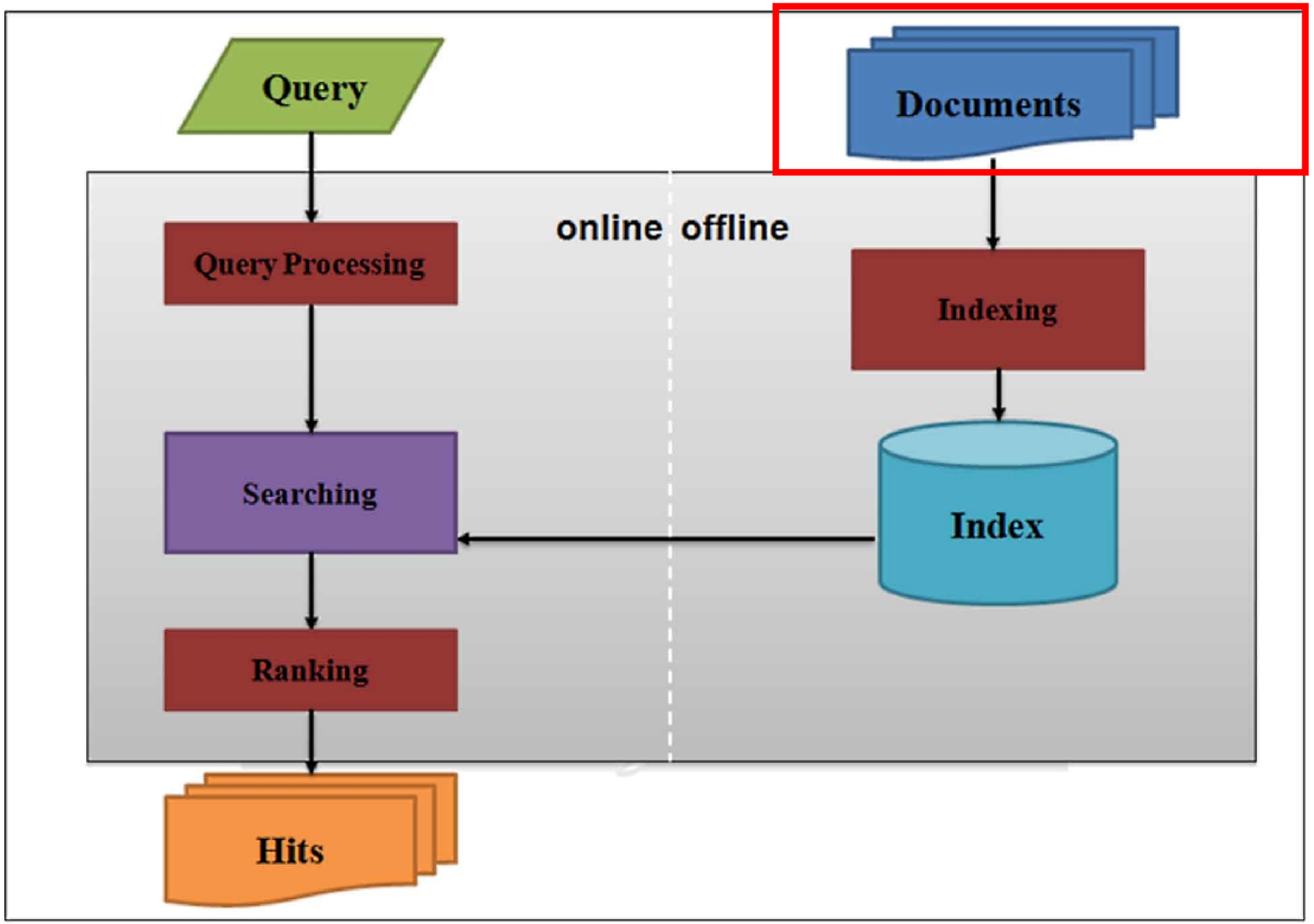
Architecture



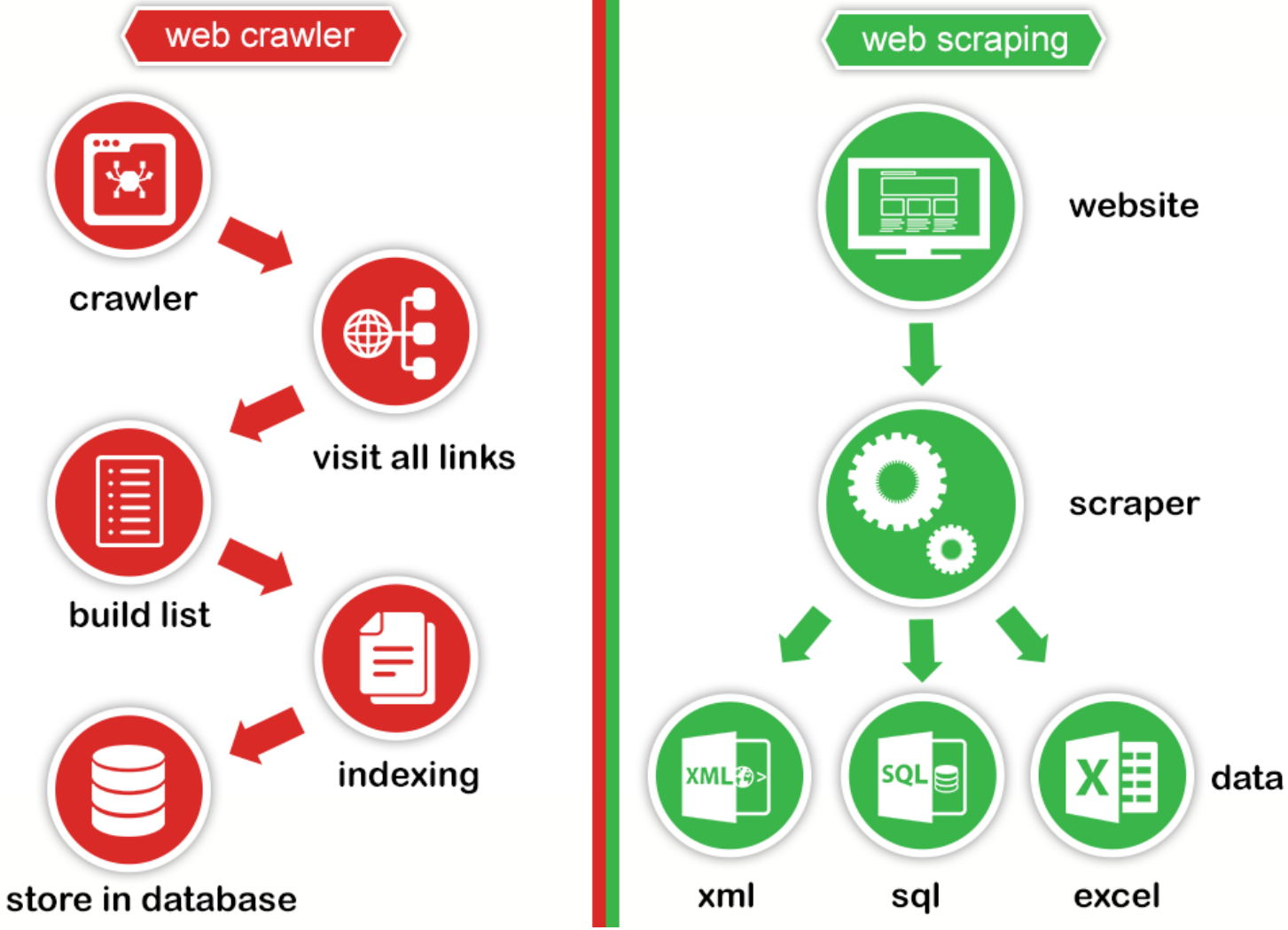
Architecture (cont.)



Architecture (cont.)

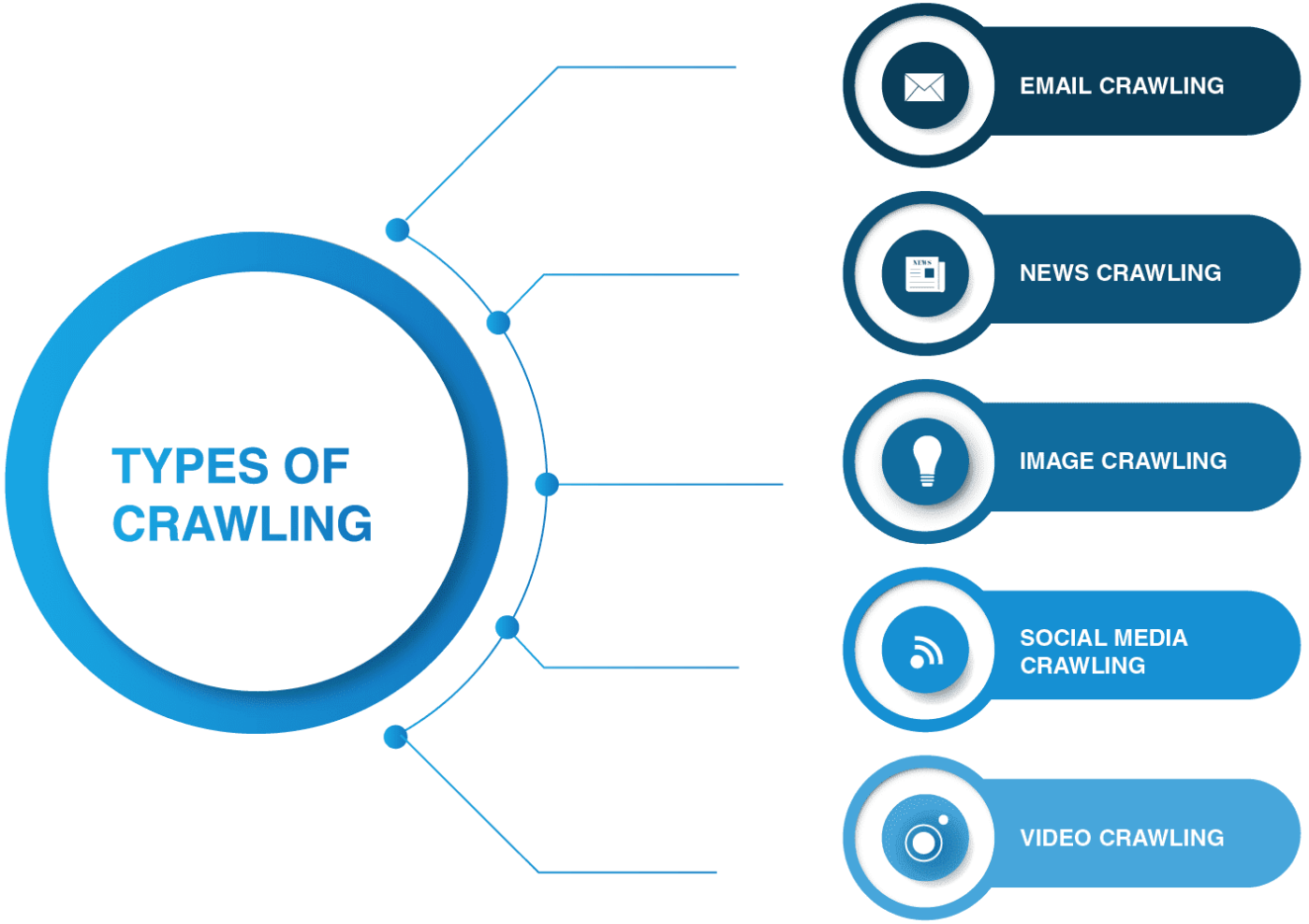


Collecting the documents (web crawling vs web scraping)



Web crawling

- https://litslink.com/blog/what-is-a-web-crawler-and-how-does-it-work


TYPES OF CRAWLING
EMAIL CRAWLING
NEWS CRAWLING
IMAGE CRAWLING
SOCIAL MEDIA CRAWLING
VIDEO CRAWLING

UNIVERSITAS
GADJAH MADA

Case study

Tugas minggu ini
(due date: 5 September 2022, class presentation)

- Membuat **document spesifikasi STKI** yang akan dikembangkan (tugas 1)
- Membuat **document arsitektur dan komponen STKI** (tugas 2)
- Membuat **Document** metode dan hasil crawling - dataset (tugas 3)
- Tugas berikutnya adalah:
  - **Document** metode dan hasil transformasi dataset (tugas 4)
  - **Document** metode dan hasil pembuatan indeks (tugas 5)
  - **Document** metode dan hasil kompresi indeks (tugas 6)
  - Paper yang menggabungkan tugas 1,2,3,4,5,6 + Presentasi (tugas 7)

- Tugas dapat dikerjakan secara perorangan atau kelompok (maksimal 3 orang)
  - Jika kelompok, maka semua anggota harus memiliki kontribusi dan melakukan presentasi+demo
  - Jika ada salah satu atau lebih anggota yang tidak berkontribusi dan tidak melakukan presentasi+demo, maka nilai kelompok tersebut dipotong 50%
- Document dibuat dalam file doc/docx atau editable file
- Paper ditulis menggunakan format ACM atau IEEE dan diharapkan menggunakan editor overleaf (https://www.overleaf.com/)
- Presentasi dibuat dalam file ppt/pptx/google slides/prezi
- Dokumentasi dikirim ke eLOK pada tugas yang bersesuaian
- Document 1,2,3,4,5,6, dataset, dan source code software diupload di github (https://github.com/)
- Dataset juga dapat diupload di https://www.kaggle.com/labrpldugm (dapat menghubungi PIC Lab –Lukman)

UNIVERSITAS
GADJAH MADA

Tanya jawab

- Email: lukmanh@ugm.ac.id
- Scholar profile: https://scholar.google.co.id/citations?user=V_iMAWYAAAAJ&hl=en