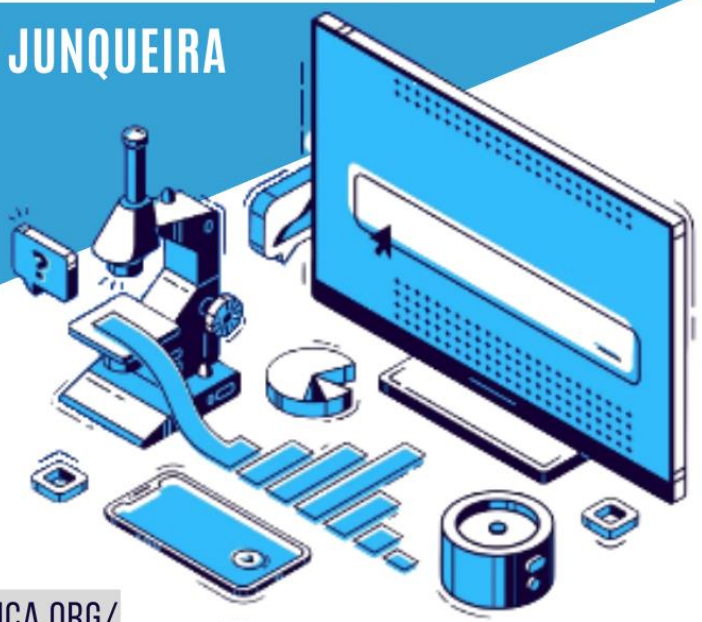
OBTENDO DADOS NA INTERNET

COM ÊNFASE EM WEBSCRAPING E TWITTER

28 DE NOVEMBRO A 01 DE DEZEMBRO, 19H ÀS 21H

PROF. DR. MURILO JUNQUEIRA

CURSO ONLINE E GRATUITO



INSCRIÇÕES:

HTTPS://ESCOLAAMAZONICA.ORG/

APOIO:

ESCOLA AMAZÔNICA DE FORMAÇÃO METODOLÓGICA

IFCH

FACS Faculdade de Ciências Sociais

APRESENTAÇÃO

Apresentação do Professor

Murilo de Oliveira Junqueira

- Professor de ciências sociais e ciência política da UFPA
- Cientista social e cientista político, com graduação, mestrado e doutorado pela USP-SP.
- Ex-Professor do Ensino Técnico (gestão pública e políticas públicas).
- Gestor público por 8 anos.
- Apaixonado por análise de dados, tecnologia, computação, história, filosofia, economia (e outras coisas...)

Apresentação do Professor

Murilo de Oliveira Junqueira

- Atualmente Coordena dois grupos de estudos abertos:

- **CSCC**: Grupo de Estudos em Ciências Sociais Computacionais.

- **GEPRS**: Grupo de Estudos em Política e Redes Sociais.

Interessados são bem vindos!

Dinâmica do Curso

- Aulas expositivas (8 horas)
- Exercícios teste (em geral bem fáceis)
- Exercícios com o R (Swril)

Roteiro de Viagem:

- Aula 01:
  - Apresentação da dinâmica do curso e das ferramentas colaborativas.
  - Fazendo pesquisa na era digital (análise de traços digitais e *big data*).
  - Introdução ao *webscraping*.

Roteiro de Viagem:

- Aula 02:
  - Introdução ao API Twitter
  - Criação da conta de desenvolvedor*
  - Pacote Rtweet
  - Dados de usuários dados de timeline
  - buscas por assunto.
  - Manejando o ratelimit

  - *** Adiantado para a aula 01**

Roteiro de Viagem:

- Aula 03:
  - Mais exemplos de webscraping e uso do Twitter API.
  - Armazenando os dados em bancos de dados eficiente.

Roteiro de Viagem:

- Aula 04:
  - Autohotkey: uma alternativa para sites “difíceis”.
  - Análise de dados das redes sociais
  - Produção de gráficos e relatórios.

FERRAMENTAS COLABORATIVAS

Apresentação dos alunos

- Se apresente na planilha Cantinho da Colaboração!

Onde acessar as informações das aulas

- Vou colocar as anotações importantes do curso no neste documento:
  - Tutoriais
  - Exercícios
  - Curiosidades

Anote o link!

FAZENDO PESQUISA NA ERA DIGITAL

Crescente Digitalização da Vida

A digitalização social permitiu uma abundância de dados observacionais sobre uma série de fatos socialmente relevantes:

- Interações em mídias sociais
- Vendas online,
- Aplicativos eletrônicos,
- Registros administrativos
- Arquivos históricos digitalizados

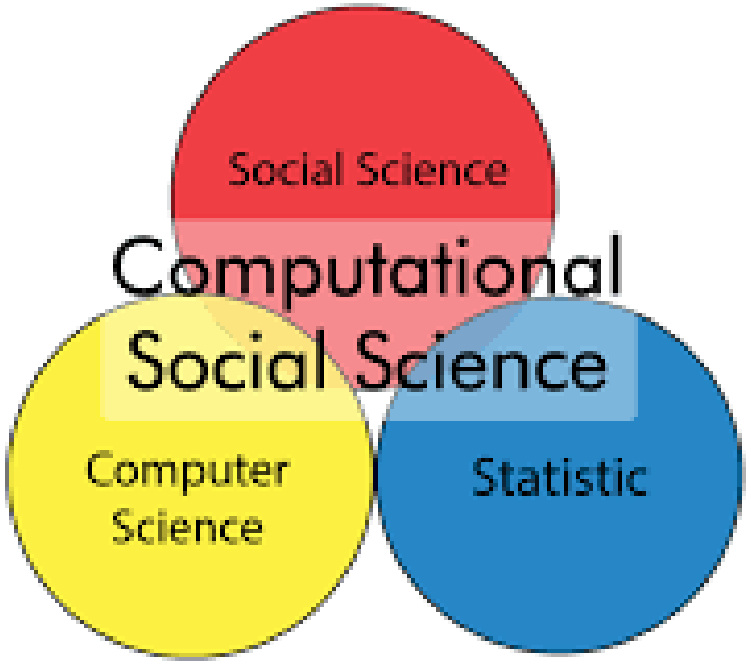
Dois mal-entendidos da nova era:

- “Embora a pesquisa social na era digital ainda não tenha produzido contribuições intelectuais maciças e grandes mudanças de paradigma, a taxa de melhoria na pesquisa da idade digital é incrivelmente rápida. É essa taxa de mudança - mais do que o nível atual - que torna a pesquisa da idade digital tão emocionante para mim. ”

Matthew J. Salganik,
Sociólogo,
Universidade de Princeton

Dois mal-entendidos da nova era:

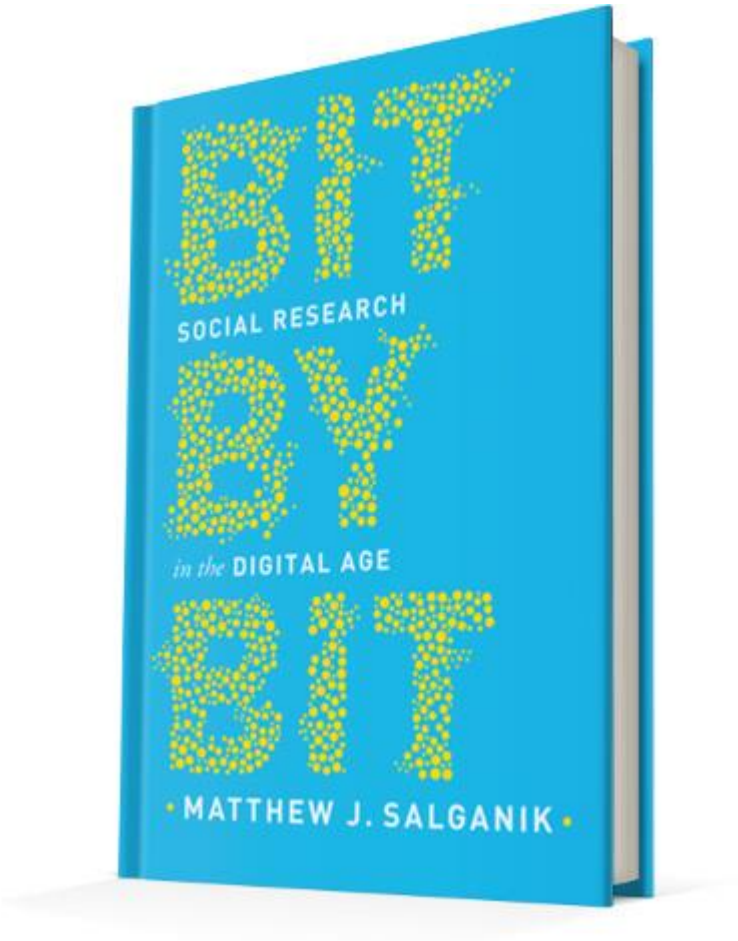
- Mais dados resolvem automaticamente problemas.
- Ciências sociais é apenas um monte de conversa esnobe.



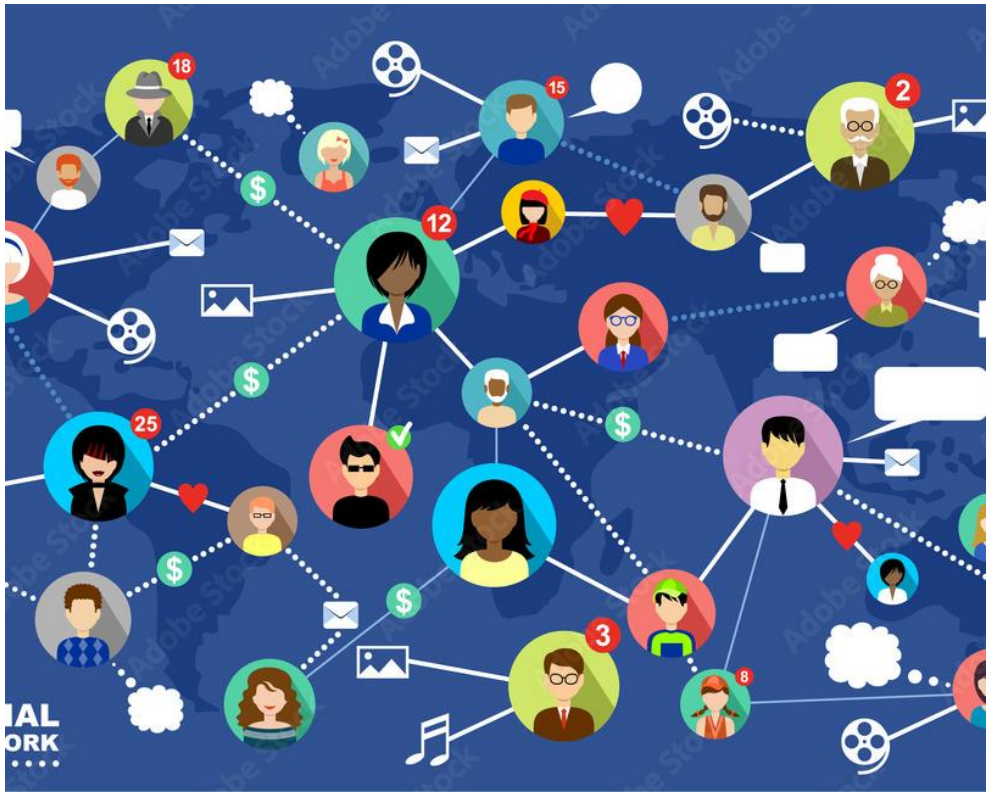
Diferença entre CS e CD.

- **Cientistas sociais** que têm treinamento e experiência no estudo do comportamento social, mas que estão menos familiarizados com as oportunidades criadas pela era digital.

- **Cientistas de dados** que se sentem muito confortáveis usando as ferramentas da era digital, mas que são novas no estudo do comportamento social.



A era do Big Data

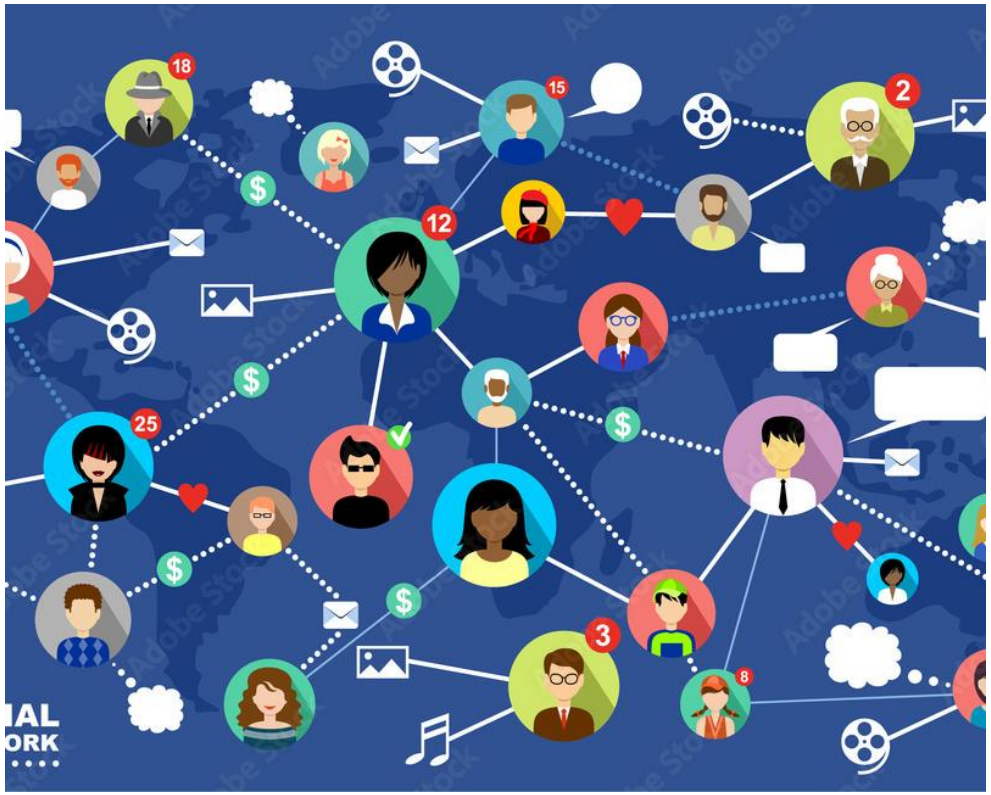


Salganik:

- Muitas coisas em sua vida que costumavam ser analógicas agora são digitais.
  - Câmeras agora são digitais (junto com smartphones).
  - Jornais agora são digitais Leia um jornal online.
  - Pagamentos agora são digitais (PIX)

- A mudança de analógica para digital significa que mais dados sobre você estão sendo capturados e armazenados digitalmente.

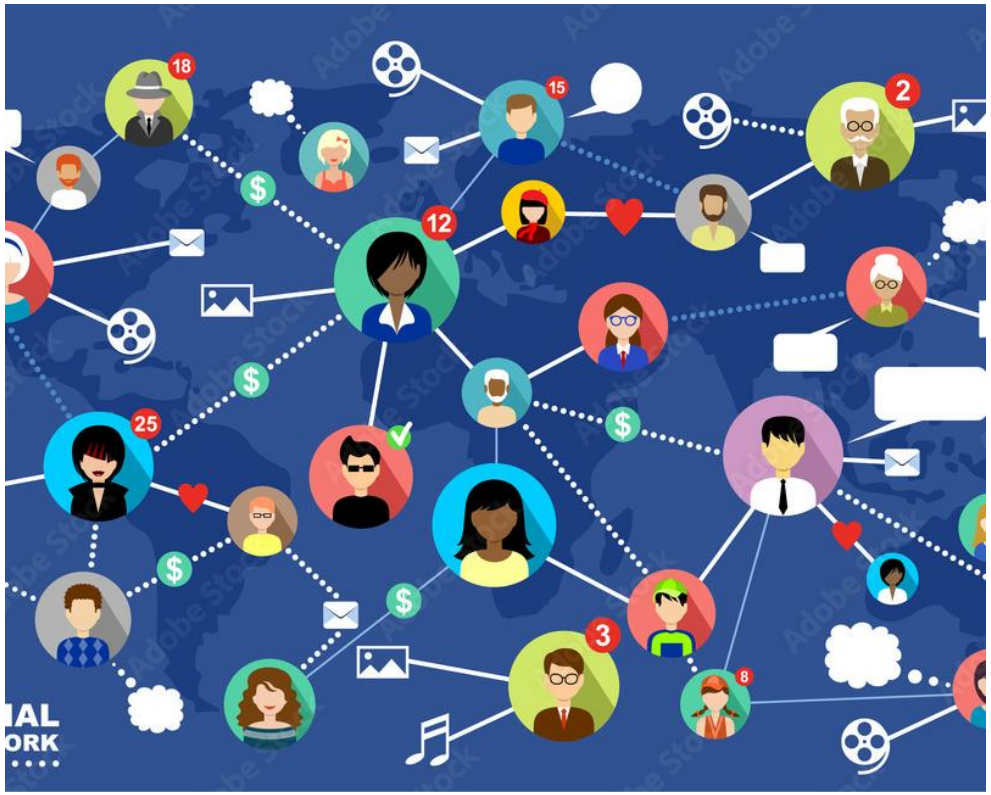
A era do Big Data



Salganik:

- "A quantidade de informações no mundo está aumentando rapidamente e mais dessas informações são armazenadas digitalmente, o que facilita sua análise, transmissão e integração (...) Todas essas informações digitais passaram a ser chamadas de '**big data**'".

A era do Big Data



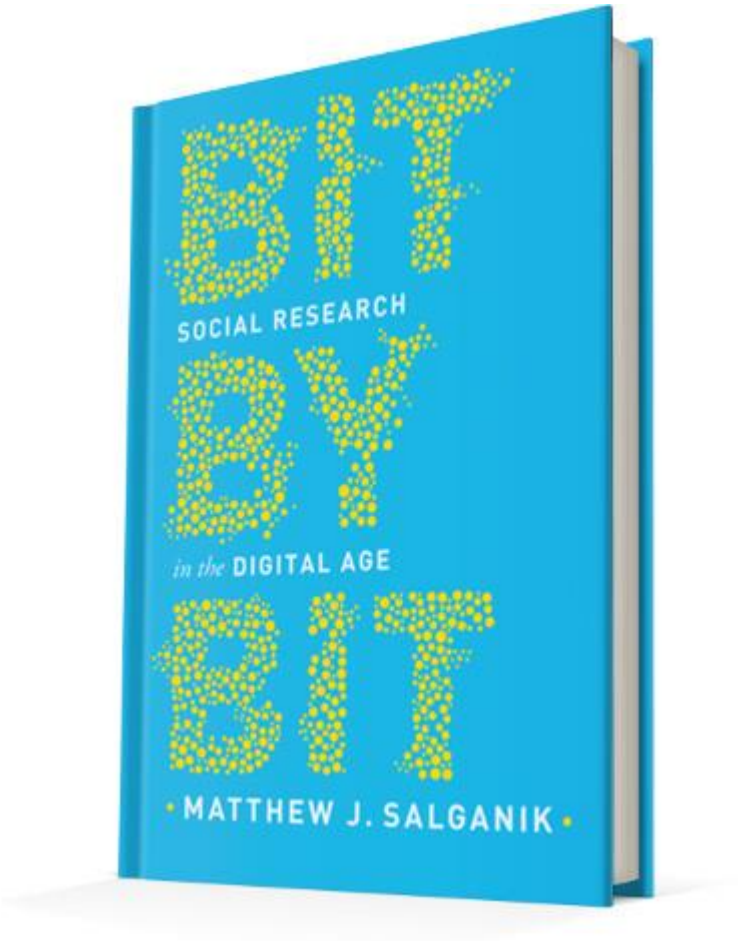
Salganik:

- Explosão em nosso acesso ao poder de computação.
- É provável que essas tendências continuem no futuro próximo.
- Transição da idade analógica para a era digital ainda não está completa

Os traços digitais

- "Na era analógica, coletando dados sobre comportamento - quem faz o que e quando - era caro e, portanto, relativamente raro. Agora, na era digital, os comportamentos de bilhões de pessoas são registrados, armazenados e analisáveis. "

- O registro digital do seu comportamento que é criado e armazenado por uma empresa é chamado de **traços digitais**.



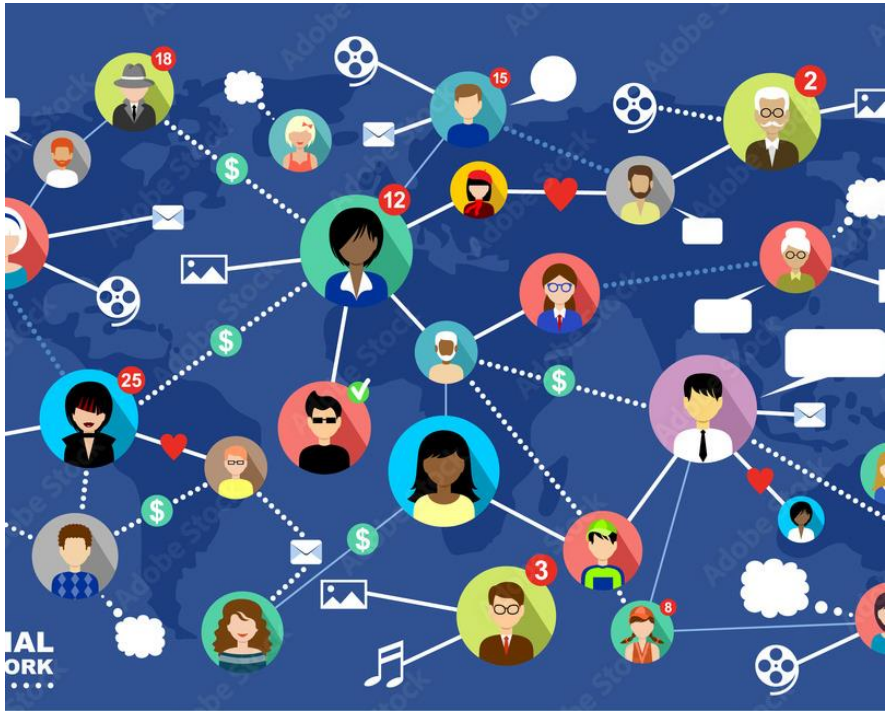
Twitter vs Censo

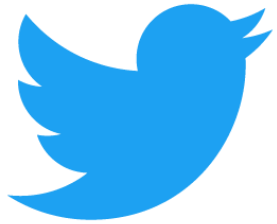
- O Twitter opera em uma escala e velocidade que o censo não pode corresponder,
- O censo trabalha cuidadosamente as amostras e mantém uma comparabilidade ao longo do tempo que o Twitter não pode atingir.
- *- Não faz sentido dizer que um tipo de dado é melhor que o outro.

Características dos traços digitais

Dez características comuns de big data:

- Big
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically confounded
- Dirty
- Sensitive





CRIAÇÃO DE CONTA DE DESENVOLVEDOR NO TWITTER

Um recado importante:

- Já criem uma conta de desenvolvedor no Twitter!

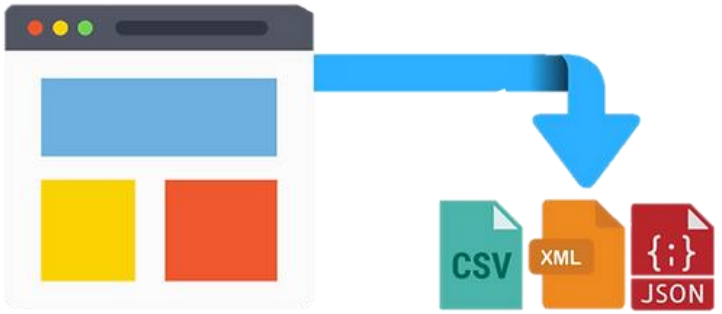
- Às vezes o Twitter demora um pouco até aceitar a conta de desenvolvedor, então, já iniciem o processo logo!

INTRODUÇÃO AO *WEBSCRAPING*

(raspagem de tela)

Introdução ao WebScraping

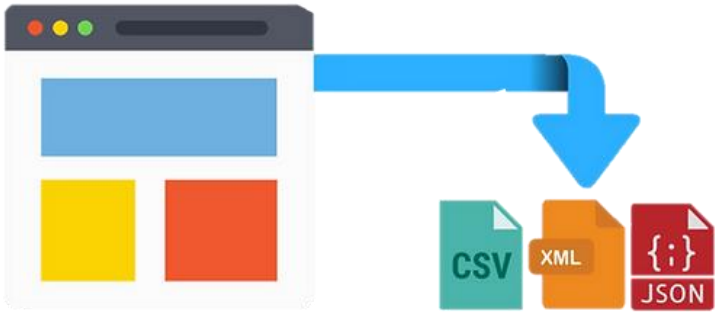
- *WebScraping* (ou raspagem de tela) é o processo de algoritmicamente extrair informações de paginas da internet (*sites*) e organiza-las em bancos de dados.



- É um processo extremamente útil e poderoso de obter informações na internet.

Estruturas das páginas na internet

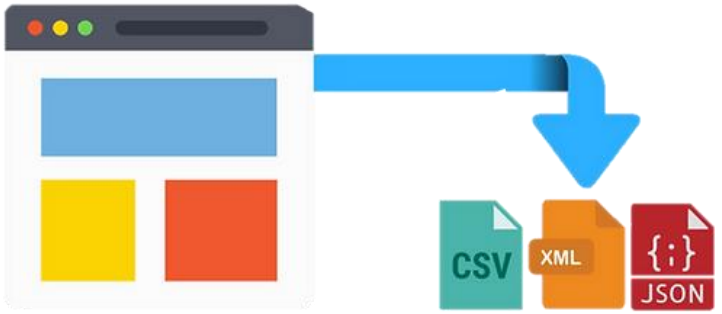
- As páginas da internet são escrita com uma lingagem chamada de HTML (*HyperText Markup Language*, que significa: "Linguagem de Marcação de Hipertexto“)



- Basicamente, os textos são formatados com sinais de “<>” no início da marcação e “< / >” no final da marcação.
  - Exemplo: Murilo => **Murilo**
    - b é a marca de « bold » (negrito)

Estruturas das páginas na internet

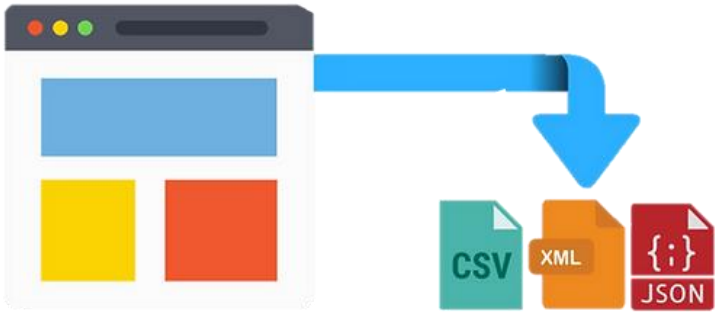
- Modernamente, a linguagem HTML se mistura com os bancos de dados em XML, criando o princípio XHTML (XML + HTML).



- A linguagem XML (*Extensible Markup Language*) marca as variáveis dos bancos de dados também com “<>” e “< / >”:
  - <nome>Murilo</nome>
  - <profissão>Professor</profissão>

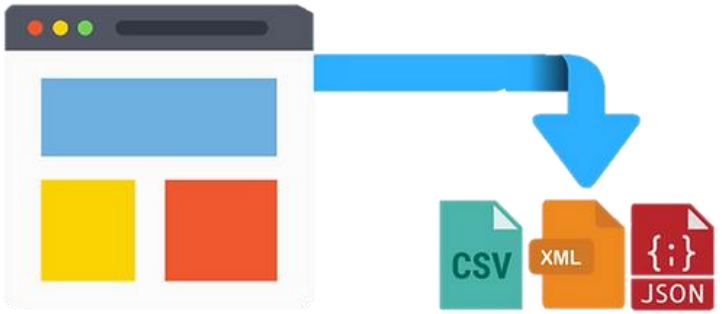
Estruturas das páginas na internet

- Assim, as páginas atualmente são (mais ou menos) bancos de dados em XML. Então, basta entender sua estrutura para extrairmos dados!



Agora um exemplo prático!

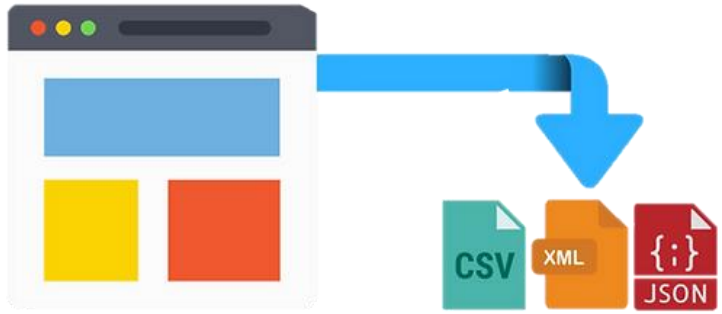
- Vamos extrair as informações de uma única notícia de jornal do site “Congresso em Foco”.



- *-Programar ao vivo é muito difícil, por isso sejam generosos ao me verem, ok! 😁

Agora um exemplo prático!

- Agora vamos usar uma lista de busca para busar várias páginas ao mesmo tempo!



Muito obrigado!

m.Junqueira@yahoo.com.br