Introduction To Web Scraping

With David Teather

slidesmania.com

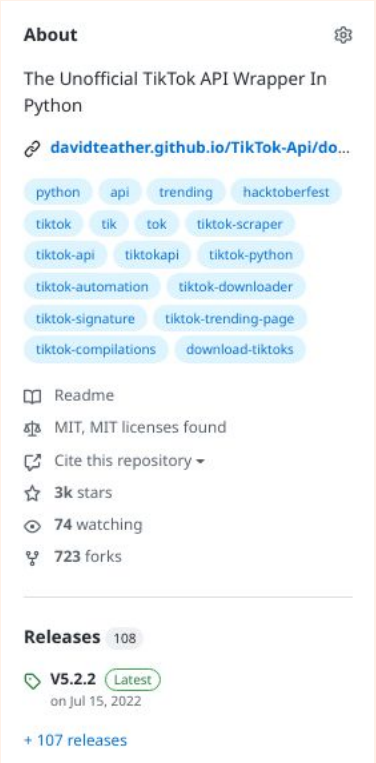
Who am I? 🤔

- CS & SWE
- A general clown
- Uses emojis excessively

slidesmania.com

TikTokAPI

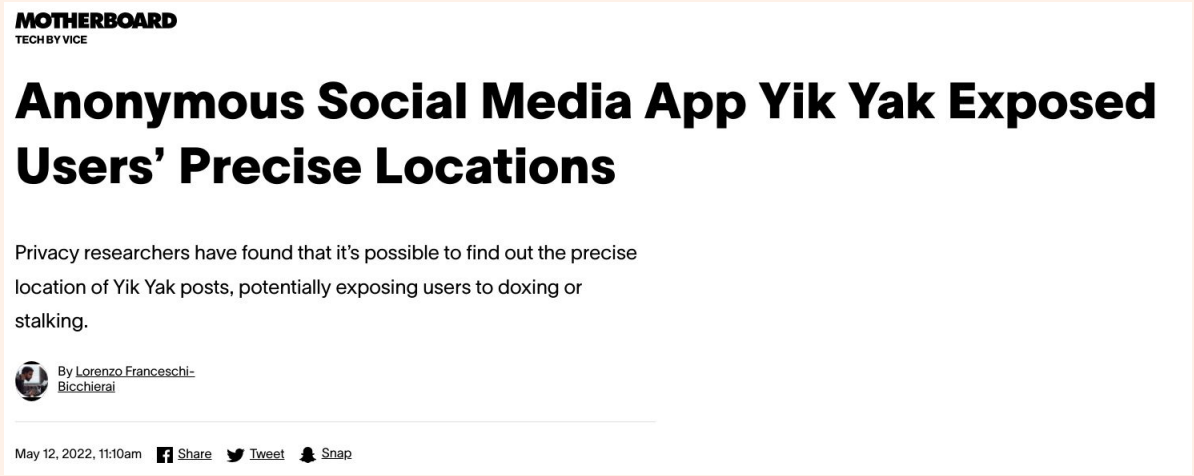
- Programmatically extract TikTok data from Python
- 800k+ downloads
- 3k+ stars



slidesmania.com

YikYak “Security Analysis”

- YikYak was exposing GPS coords of posts within 10-15 feet
- TheResponseTimes.com



MOTHERBOARD
TECH BY VICE

Anonymous Social Media App Yik Yak Exposed Users’ Precise Locations

Privacy researchers have found that it’s possible to find out the precise location of Yik Yak posts, potentially exposing users to doxing or stalking.

By Lorenzo Franceschi-Bicchierai

May 12, 2022, 11:10am Share Tweet Snap

slidesmania.com

GitHub Campus Expert

- Please take more stickers 😭

slidesmania.com

Why should you learn web scraping?

slidesmania.com

Why you should learn

- Data is (almost) priceless
- Improves decision making
- Not all companies have their partner’s data

slidesmania.com

Record Labels

Imagine you're a record label and you want to find artists that will be successful in the future.

Q: What data would be valuable to have?

slidesmania.com

Record Labels

- TikTok
- Spotify
- YouTube
- Twitter
- Instagram
- Etc

Especially smaller companies might not have deals with these companies for their artists data

slidesmania.com

Record Labels

(Before 2018)

- Hardly any data utilized in music
- Talent relation agents went to multiple concerts a day to scout
- Listened to hundreds of irrelevant hours music

slidesmania.com

Record Labels

(After 2018)

- Still listen to music 24/7, but more strategic in choosing what to listen to
- Better predictions of who will be the next big star
- Critics claim the numbers ruin the art in music
- All major labels utilize data heavily

WARNER MUSIC GROUP

Sodatone.

slidesmania.com

Tracking.Exposed

- Trying to research algorithms as a 3rd party
  - Maybe for independent auditing
- Usually don't have research APIs available
  - And some accidentally mislead researchers
- Need to collect tons of data yourself

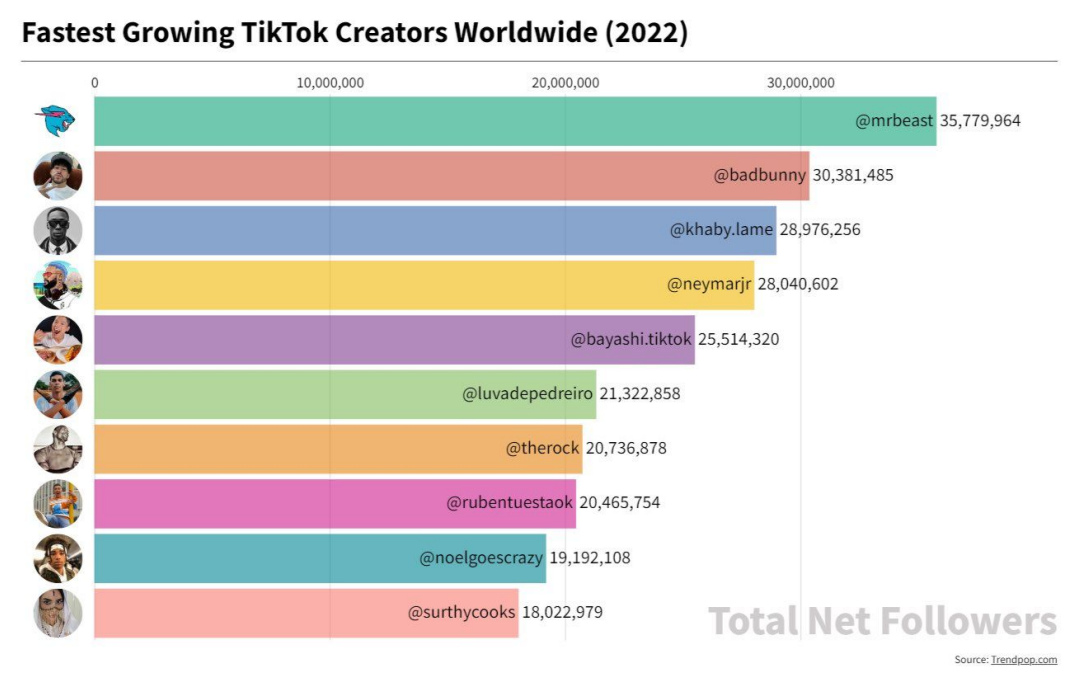
slidesmania.com

Trendpop/Collab

- Collab
  - Signing content creators
- Trendpop
  - Analyzing social media primarily for marketers
    - Talent & Media Agencies
    - Music Streaming Platforms
    - Content Creators

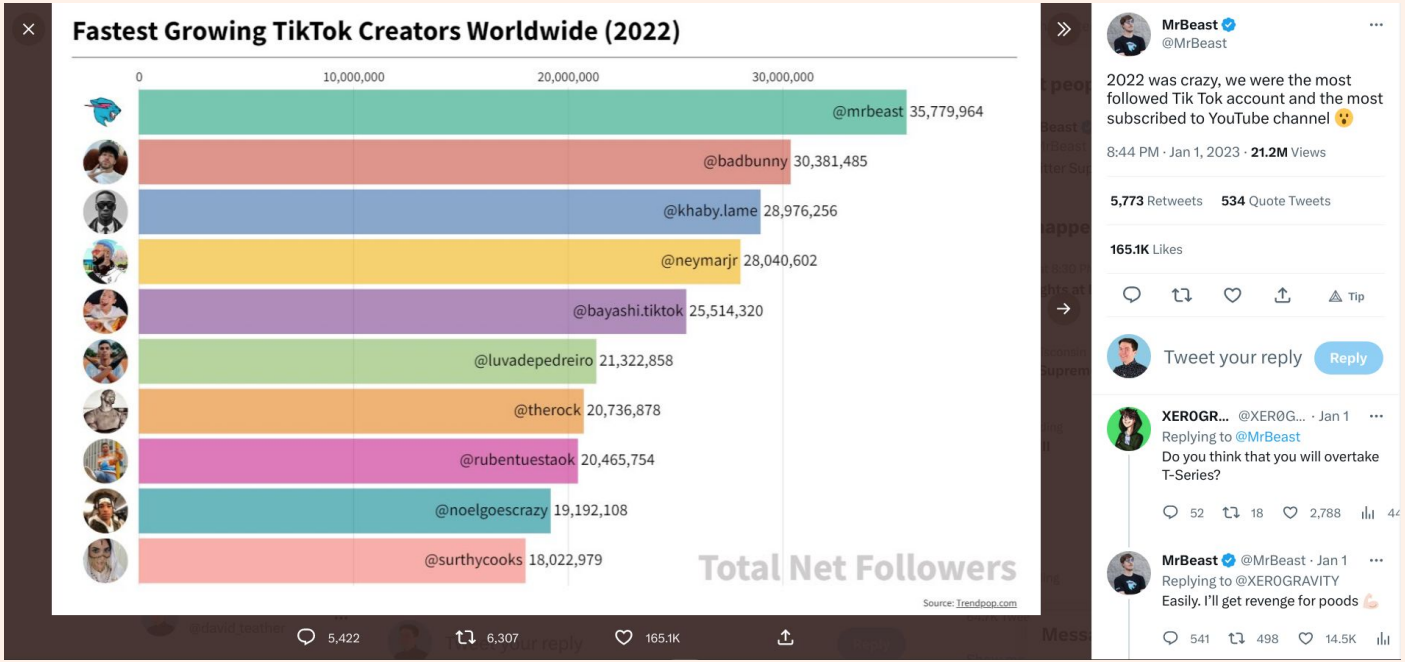
slidesmania.com

Trendpop/Collab



slidesmania.com

Trendpop/Collab



slidesmania.com

Possibilities are Endless

(Currently)

- Predominantly hyper-competitive industries
  - Social Media
  - Investment Firms
  - More “Tech” Companies

slidesmania.com

Possibilities are Endless

(Future)

- Everyone utilizing data
- Managers in companies will leverage highly accurate data
  - Even data that they might not necessarily own
  - Leverage web scraped data for competing with others
  - etc

slidesmania.com

Let’s get started!

slidesmania.com

Websites Send Files

- Adds the Structure
- The “Base” language

slidesmania.com

Websites Send Files

- Adds the Structure
- The “Base” language

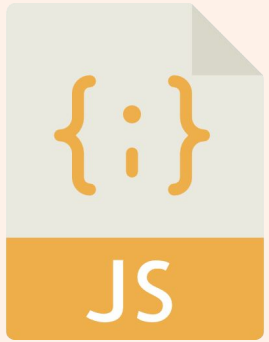
- Adds styling
- What makes a website appealing

slidesmania.com

Websites Send Files

- Adds the Structure
- The “Base” language

- Adds styling
- What makes a website appealing



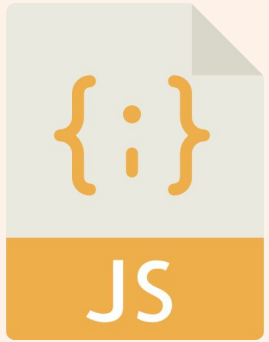
- Interactive
- “Responsive”

slidesmania.com

Websites Send Files

- Adds the Structure
- The “Base” language

- Adds styling
- What makes a website appealing



- Interactive
- “Responsive”

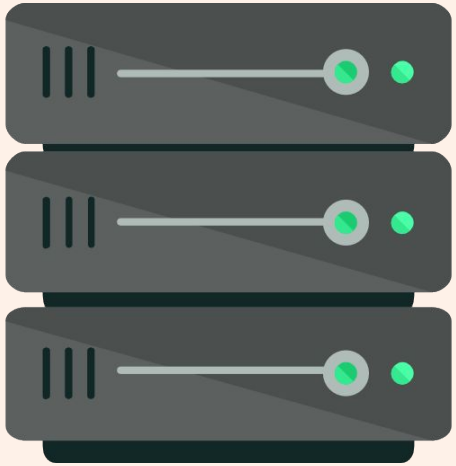
slidesmania.com

How Websites Send Data

1. Static Site/Server Side Rendering (SSR)
2. Dynamic / AJAX

slidesmania.com

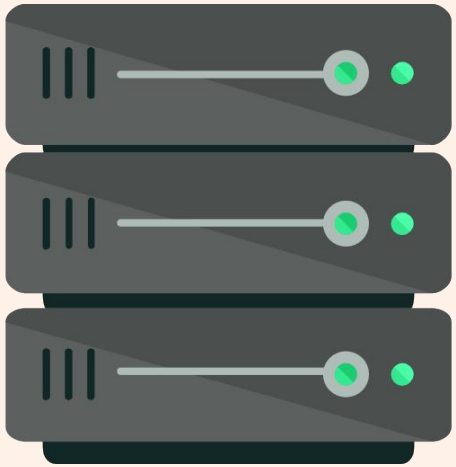
Static Site / SSR



slidesmania.com

Static Site / SSR

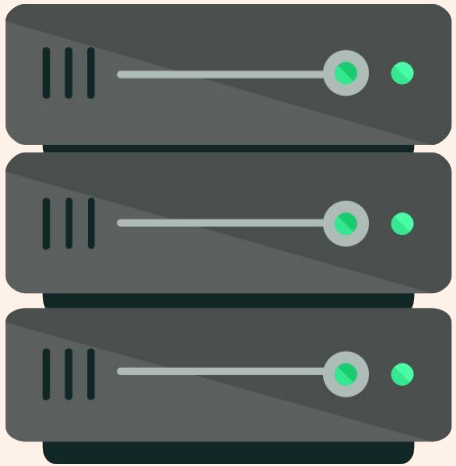
GET example.com



slidesmania.com

Static Site / SSR

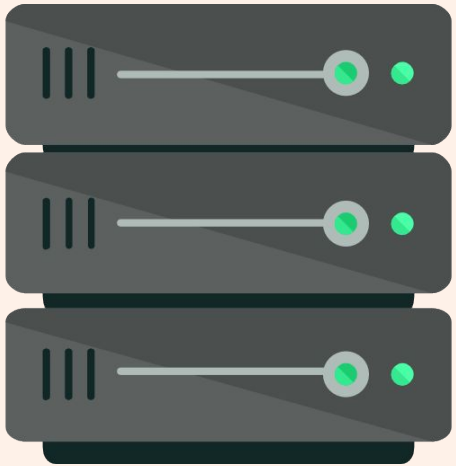
GET example.com



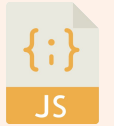
slidesmania.com

Static Site / SSR

GET example.com

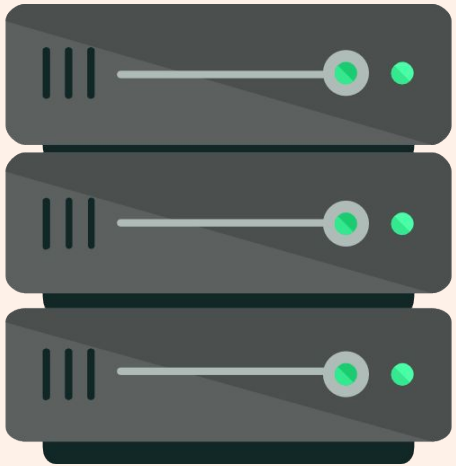


Directly has data



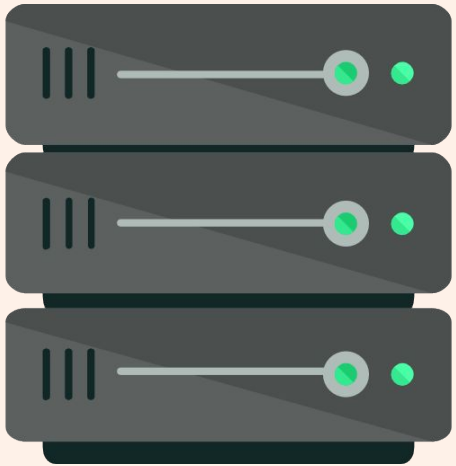
slidesmania.com

Dynamic / AJAX



slidesmania.com

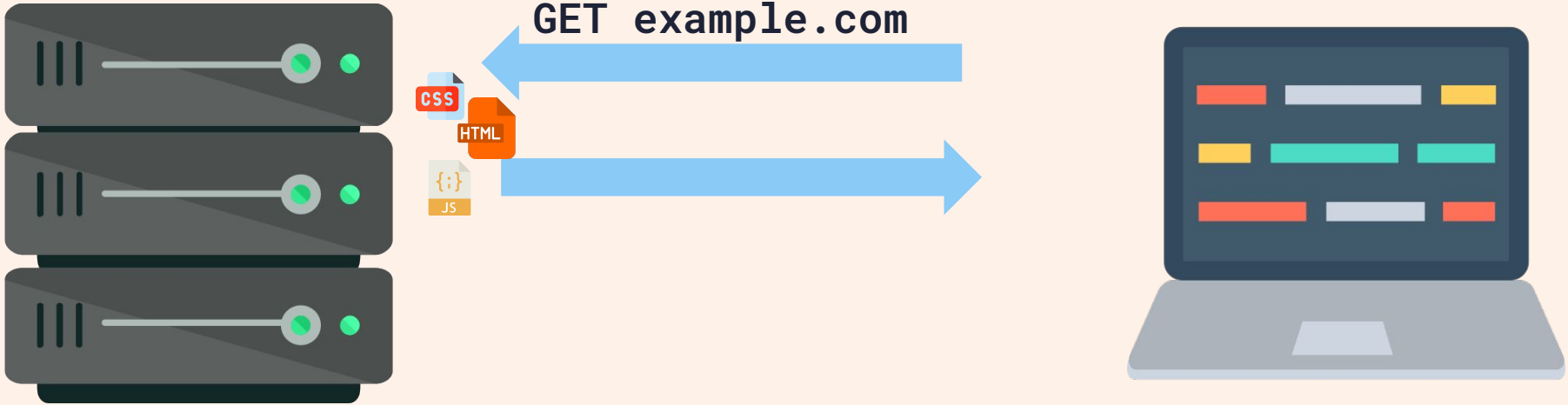
Dynamic / AJAX



GET example.com

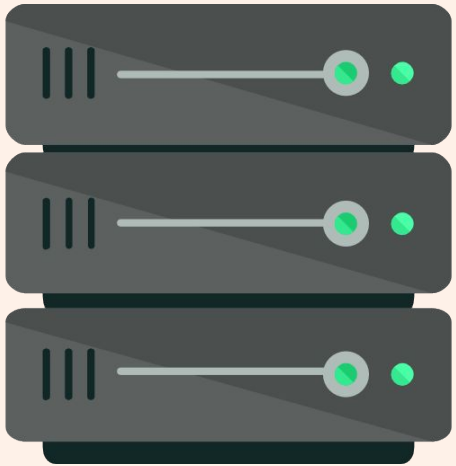
slidesmania.com

Dynamic / AJAX



slidesmania.com

Dynamic / AJAX

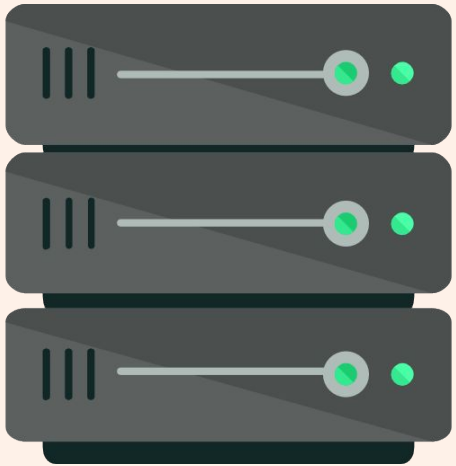


GET example.com

No relevant data

slidesmania.com

Dynamic / AJAX



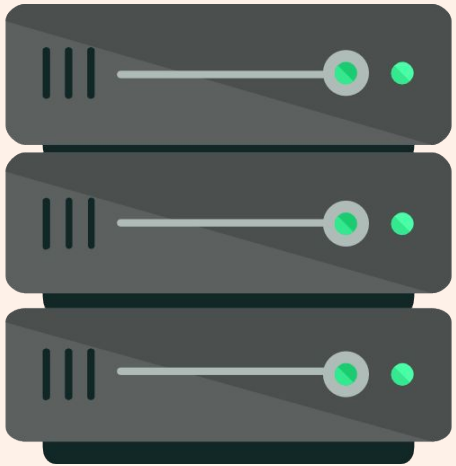
GET example.com

No relevant data

GET /api/feed

slidesmania.com

Dynamic / AJAX



GET example.com

No relevant data

GET /api/feed

Returns data

slidesmania.com

Some Websites Will Do Both

1. Initially send some data in the HTML
    - For faster load times
2. Then use AJAX
    - For responsive feel

Ex: TikTok, YouTube, etc

slidesmania.com

Workshop's Focus

- Extracting data from the HTML
- Dynamic / AJAX sites not covered here
  - Everything Web Scraping
  - Additional resources at end of slides
  - Talk to me :)

slidesmania.com

DEMO TIME!

bit.ly/webscraping-madhacks

slidesmania.com

BONUS:

Advanced Techniques

slidesmania.com

Websites Blocking Requests

- Websites don't want us to web scrape 
- Will block requests that look like a bot
  - But can't block real users or they can't access the site
- Goal: Look identical to real users

slidesmania.com

How Bots are Identified

- User Agents
- IPs
- Headers
- Cookies
- Behavior
- Browser Fingerprinting
- Much more

slidesmania.com

User Agents

- Identify what is making the request

Browser: Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/110.0.0.0 Safari/537.36

Python: python-requests/2.25.1

slidesmania.com

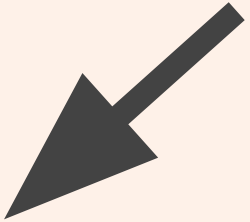
User Agents

- Identify what is making the request

Browser: Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/110.0.0.0 Safari/537.36

Python: python-requests/2.25.1

Easy to detect python requests module making requests



slidesmania.com

User Agents

- Easy to change user-agent

```
requests.get(url, headers={‘user-agent’:
‘whatever’})
```

slidesmania.com

IPs

- Rate limiting
  - Only allow X requests per minute per IP
- Bypassed with proxies
- Everything Web Scraping Lesson

slidesmania.com

Headers

- Metadata passed on each request
- Can identify browser
- User agent is a header

slidesmania.com

Cookies

- Another type of header
- Can store session data
  - Random marketing IDs
  - “Your account”
  - Also behavior

slidesmania.com

Behavior

- Less applicable to static site web scraping
- Websites could tie your IP/cookies to your history on the site
- Ex: If you're jumping around to pages not accessible from the page you're on could look suspicious

slidesmania.com

Browser Fingerprinting

- Uses a combination of all browser attributes to uniquely identify users
- Again, your goal is appear as average as possible
- Cool/Scary Resource: AmIUnique.org
  - Even if you don't give it any explicit permissions on your browser usually uniquely identifies me :(

slidesmania.com

Browser Automation

- Programmatically control a browser
- Can emulate human behavior
- Can still extract HTML or scrape through that
- Good solution for dynamically generated sites
  - Although Forging API Requests has some advantages over this

slidesmania.com

Connect With Me!

David Teather

LinkedIn

GitHub

Feedback

bit.ly/david-workshop-feedback

(also on the GitHub Repo)

please do this it helps me improve :)

slidesmania.com