Suzanna Kangas
Daniel Leblanc
Lucas Berge

Fredrik Fostvedt
Joseph Schutz
Eric Smith

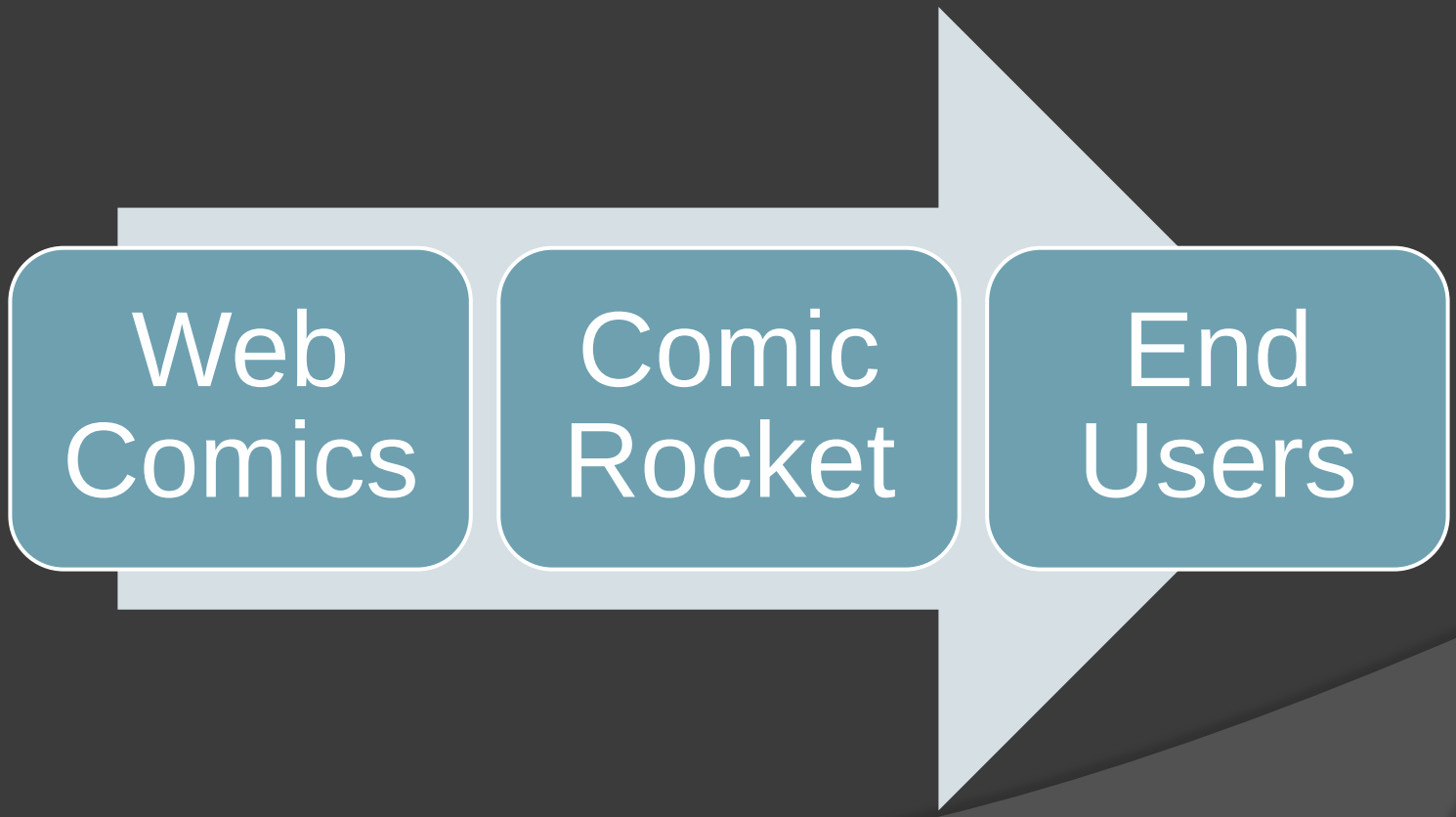
COMIC ROCKET™


COMIC ROCKET
ARTIFACT
PENNY ARCADE
girls with slingshots
THE ABADDON
SHE'S GETTING EXISTENTIAL AGAIN.
IT'S OKAY, I HAVE A SUPER SOAKER.
ROCK PAPER CYNIC
A WEBCOMIC FOR THE UNIMPRESSED
Legend of Bill
by David Reddick
WASTED TALEN
Questionable Content
Indie rock, romance, and little robots.
Gunnerkrigg Court
Amya
Evil inc.
PACK
S*P
THE ADVENTURES OF DR MCNINJA
DERELICT
girl genius
©www.studiofoglio.com

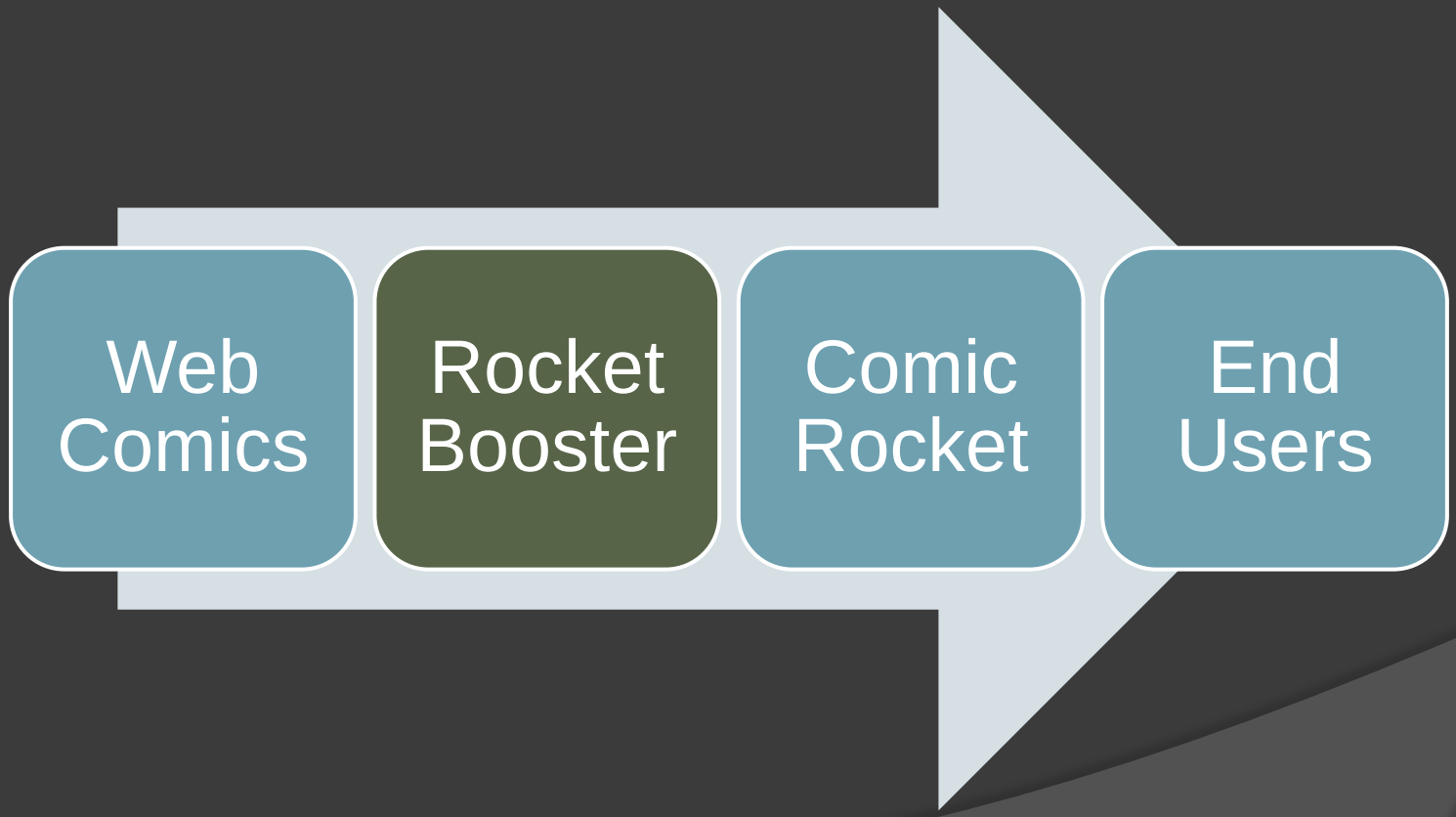
Comic Rocket

- Over 12,000 titles indexed
  - 4.1% more titles per month
- More than 1.5 million pages
  - 5.3% more pages per month
- Unique visitors climbing by 30% per month

Comic Rocket



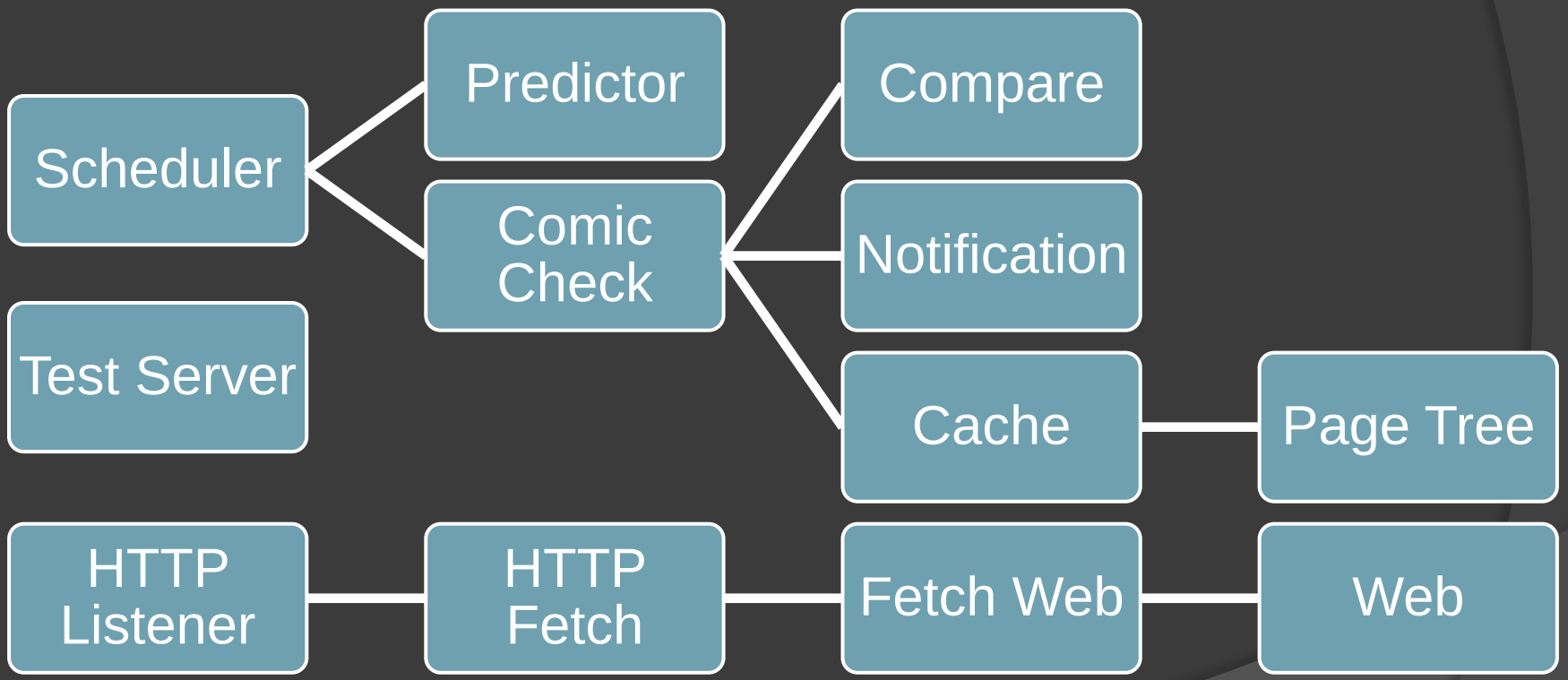
Rocket Booster



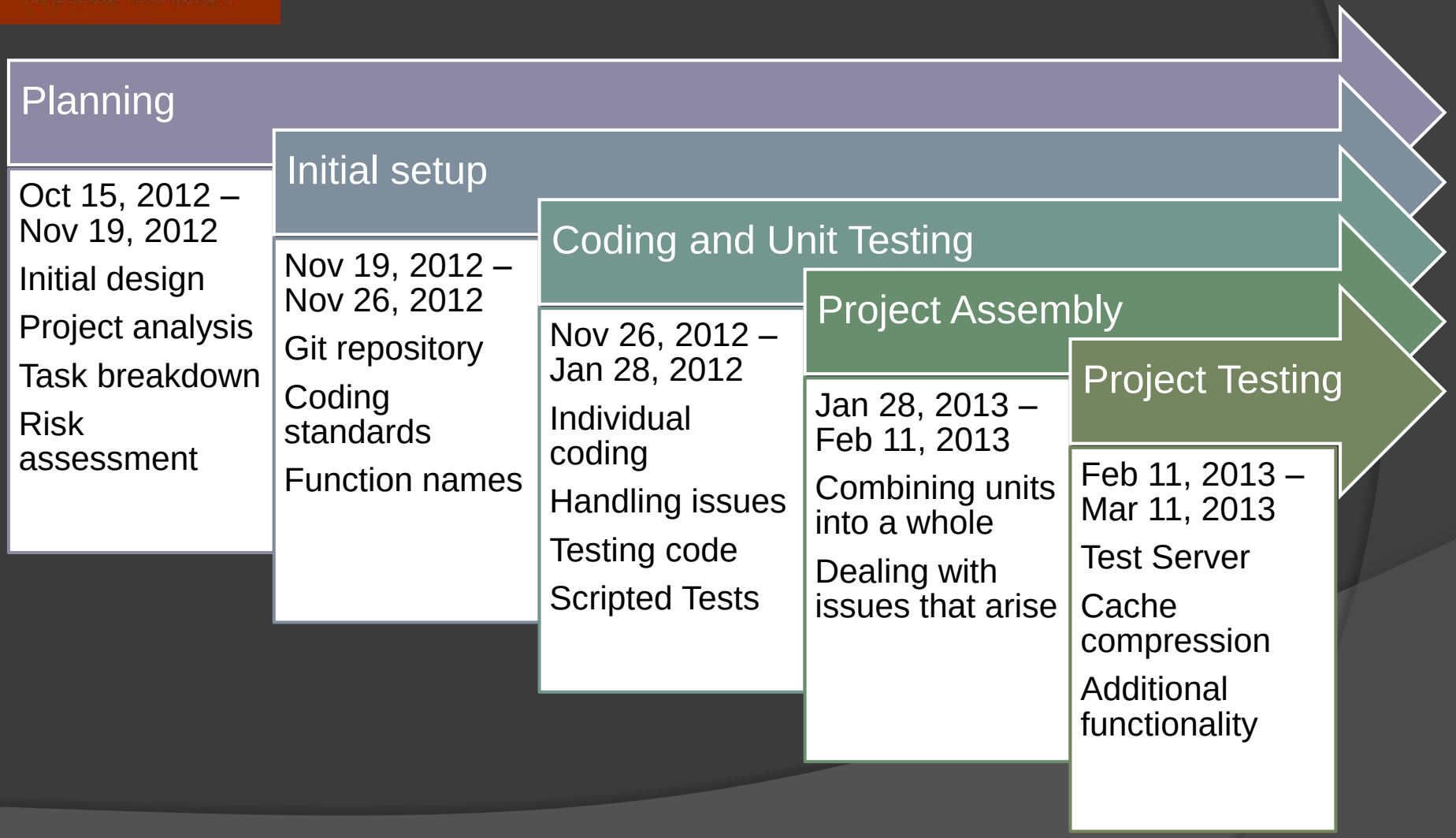
Rocket Booster

- Manage Comic Rocket's interaction with the web
- Cache all pages
- Check for new comics
- Check for updates to old comics
- Limit number of checks

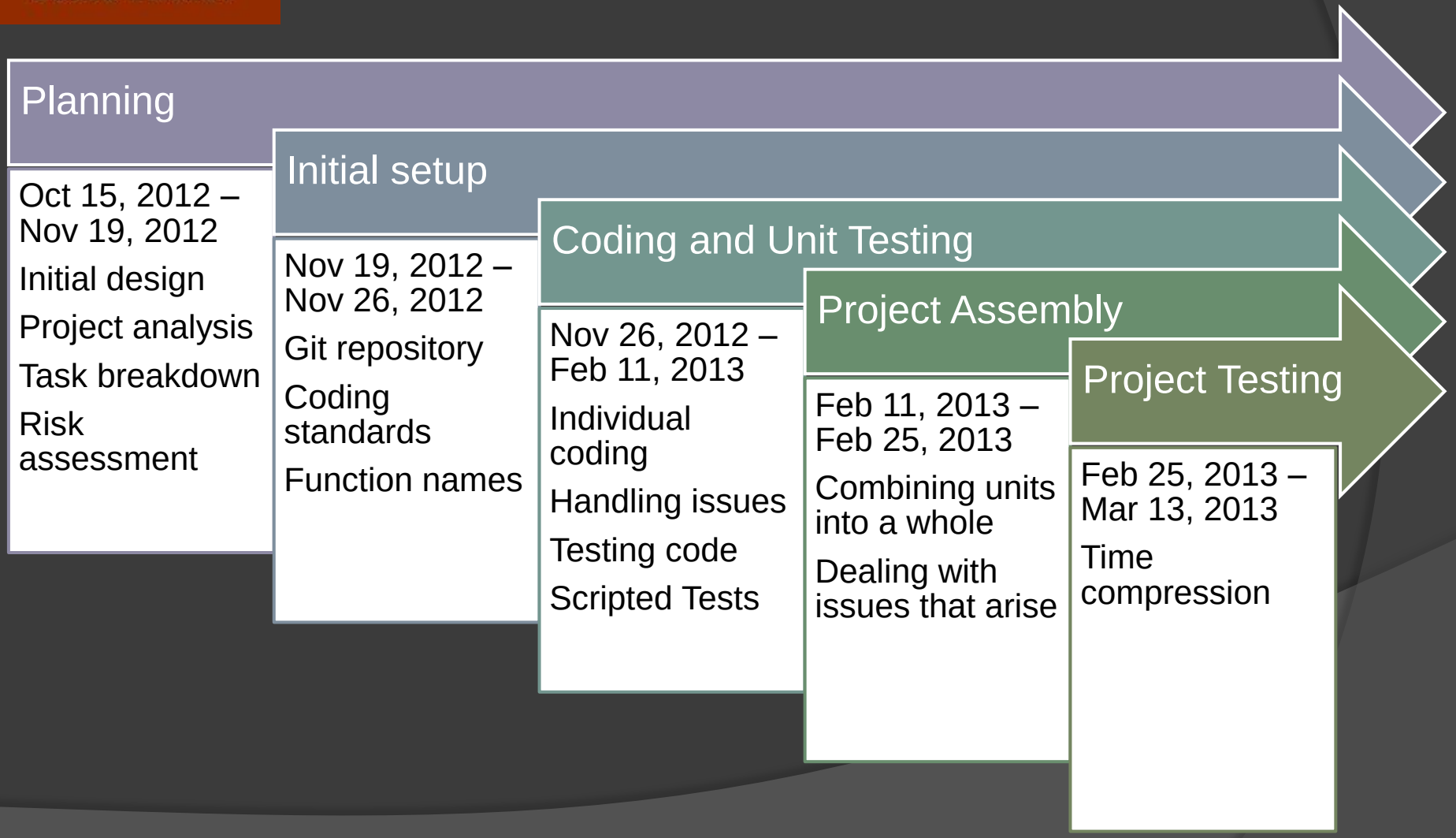
Rocket Booster Architecture



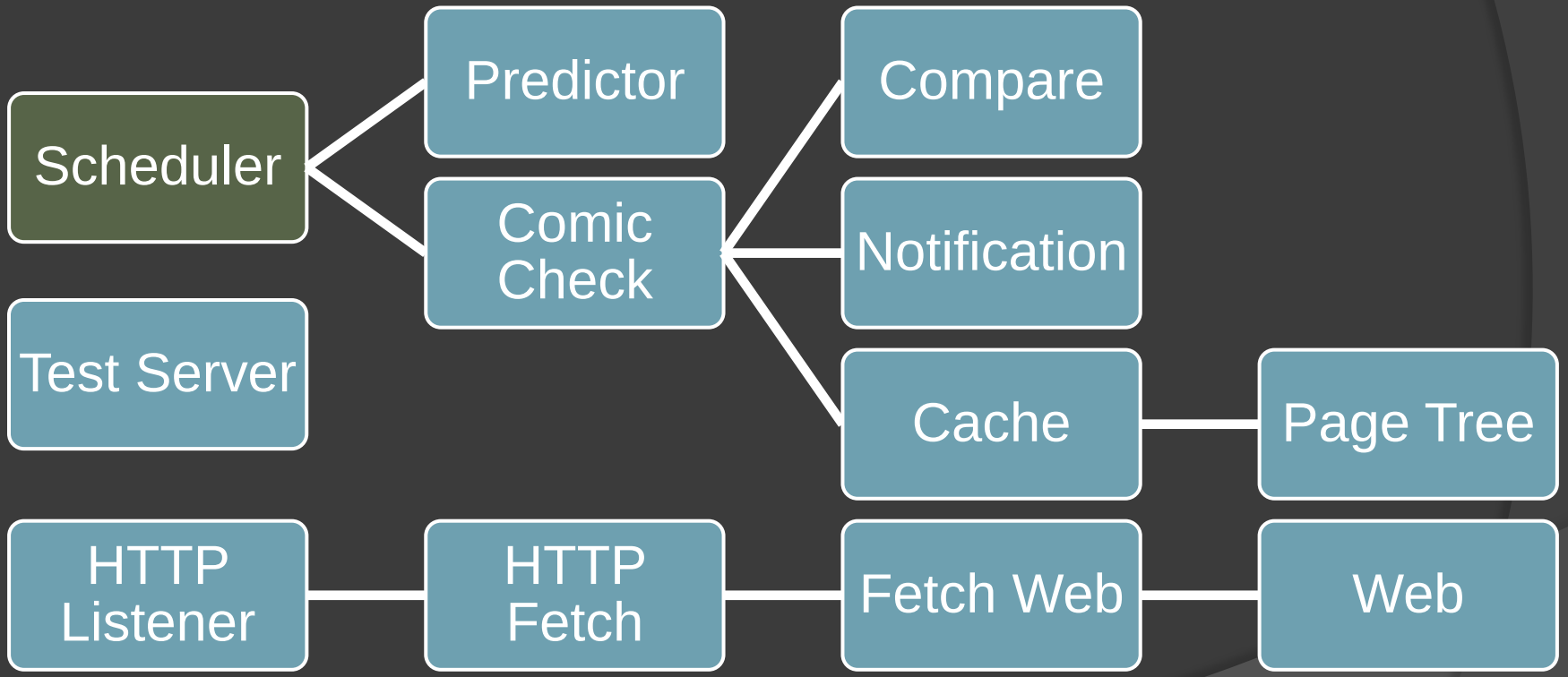
Planned Timeline



Actual Timeline



Daniel Leblanc



Scheduler

- Keeps track of history checking
  - Continuous loop of all comics
  - Checks at most 20 per hour per domain
- Checks the list of comics provided by the Predictor every hour
- Getting it to terminate was a challenge

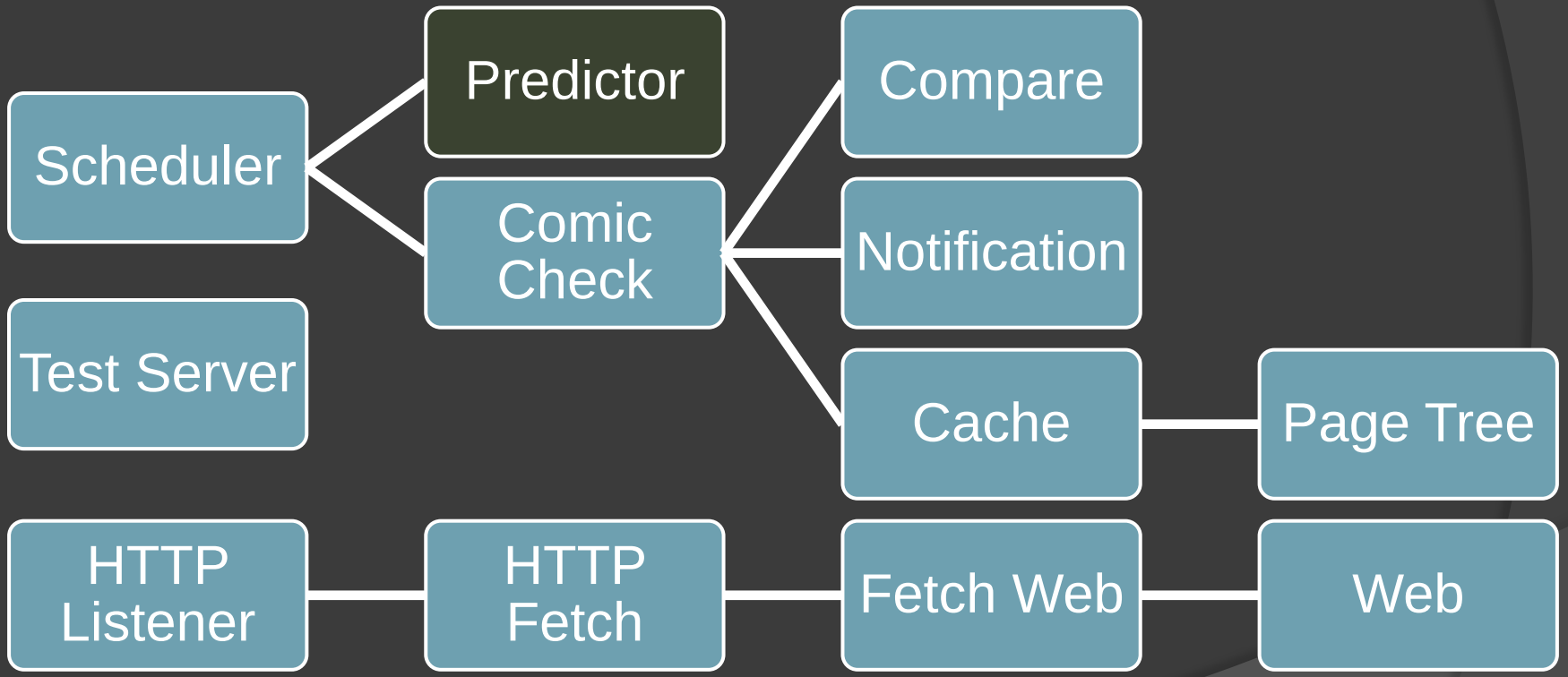
HTTP Requester

- Pretends to be Jamey
- Receives notifications and responds with HTTP requests
- Keeps track of bad notifications and missed updates

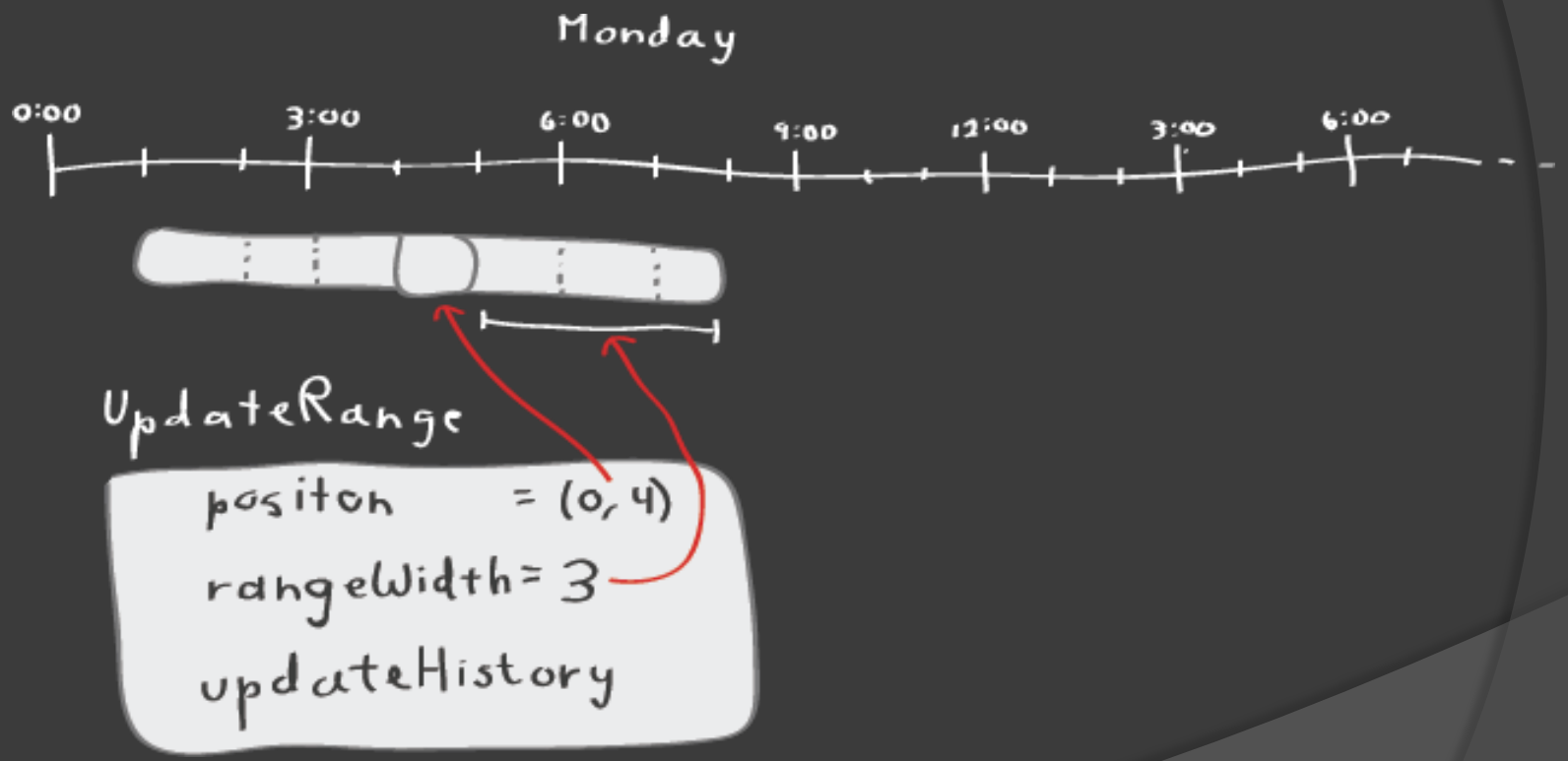
What I Learned

- Mitigating risks
- Explaining a technical idea can be challenging

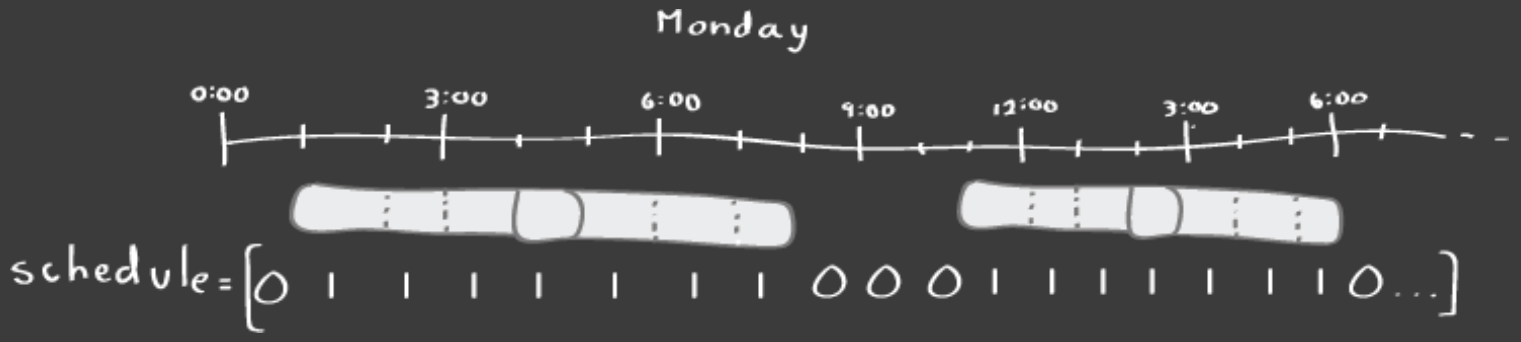
Fredrik Fostvedt



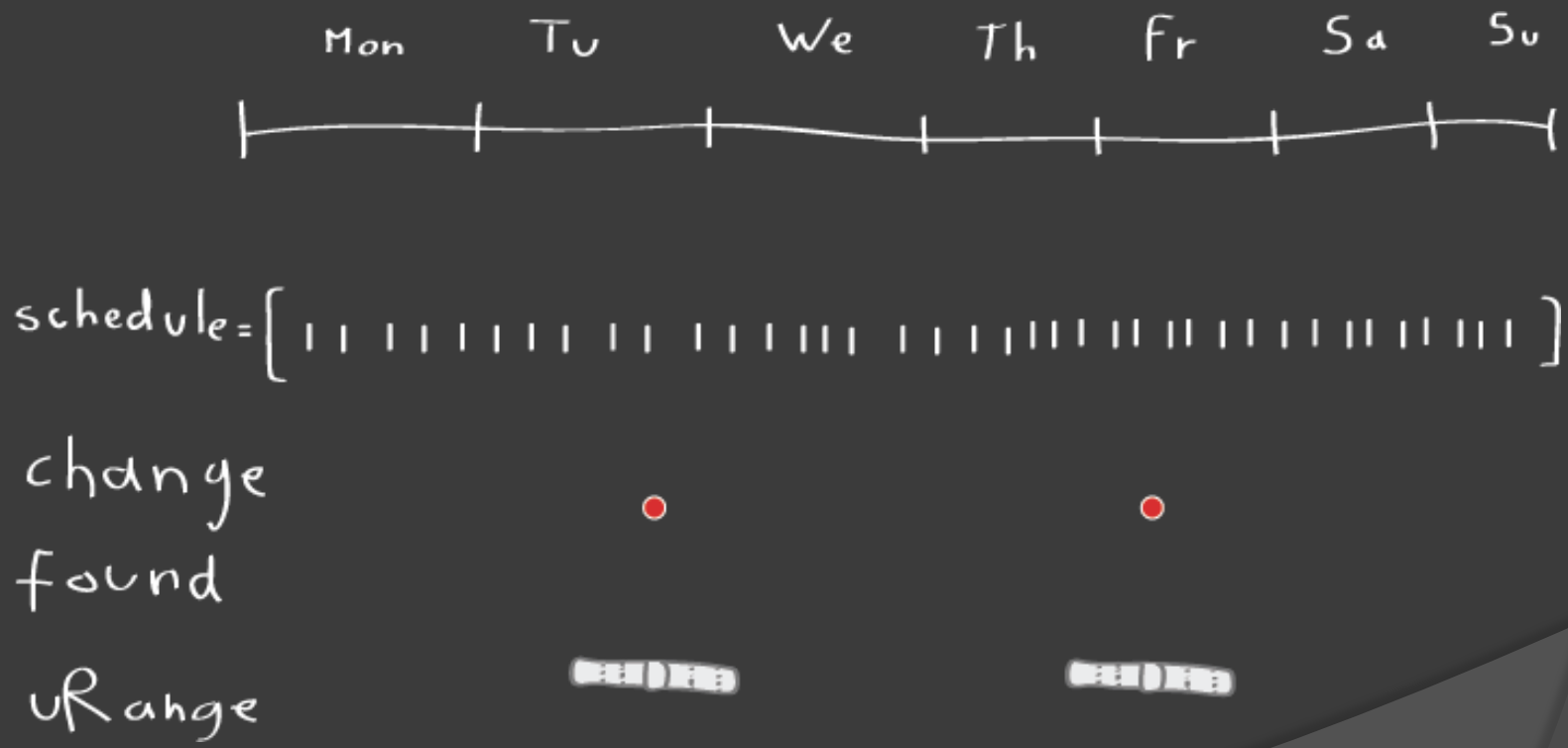
Predictor Structure



Predictor Structure



Predictor Building

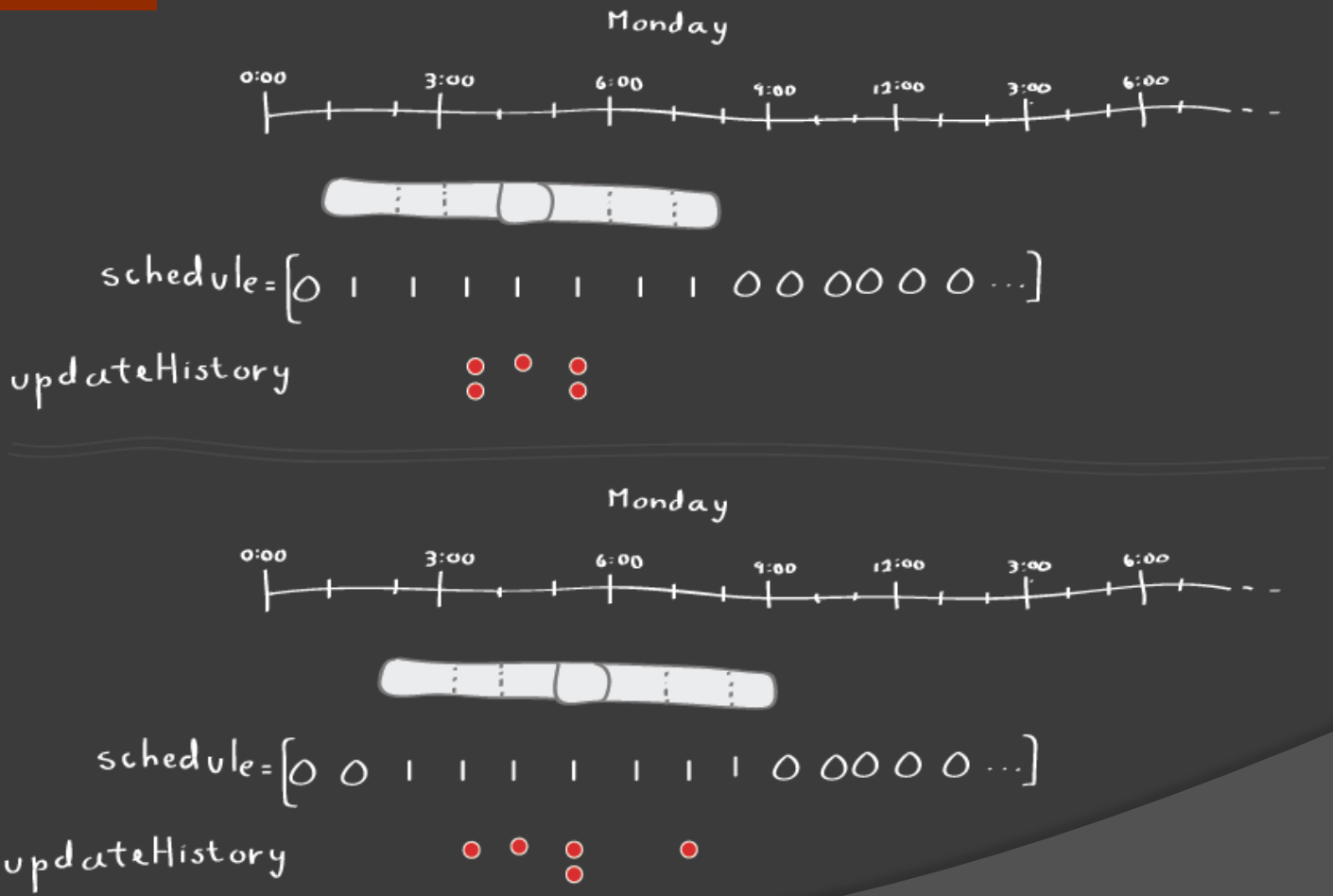


Predictor Learning

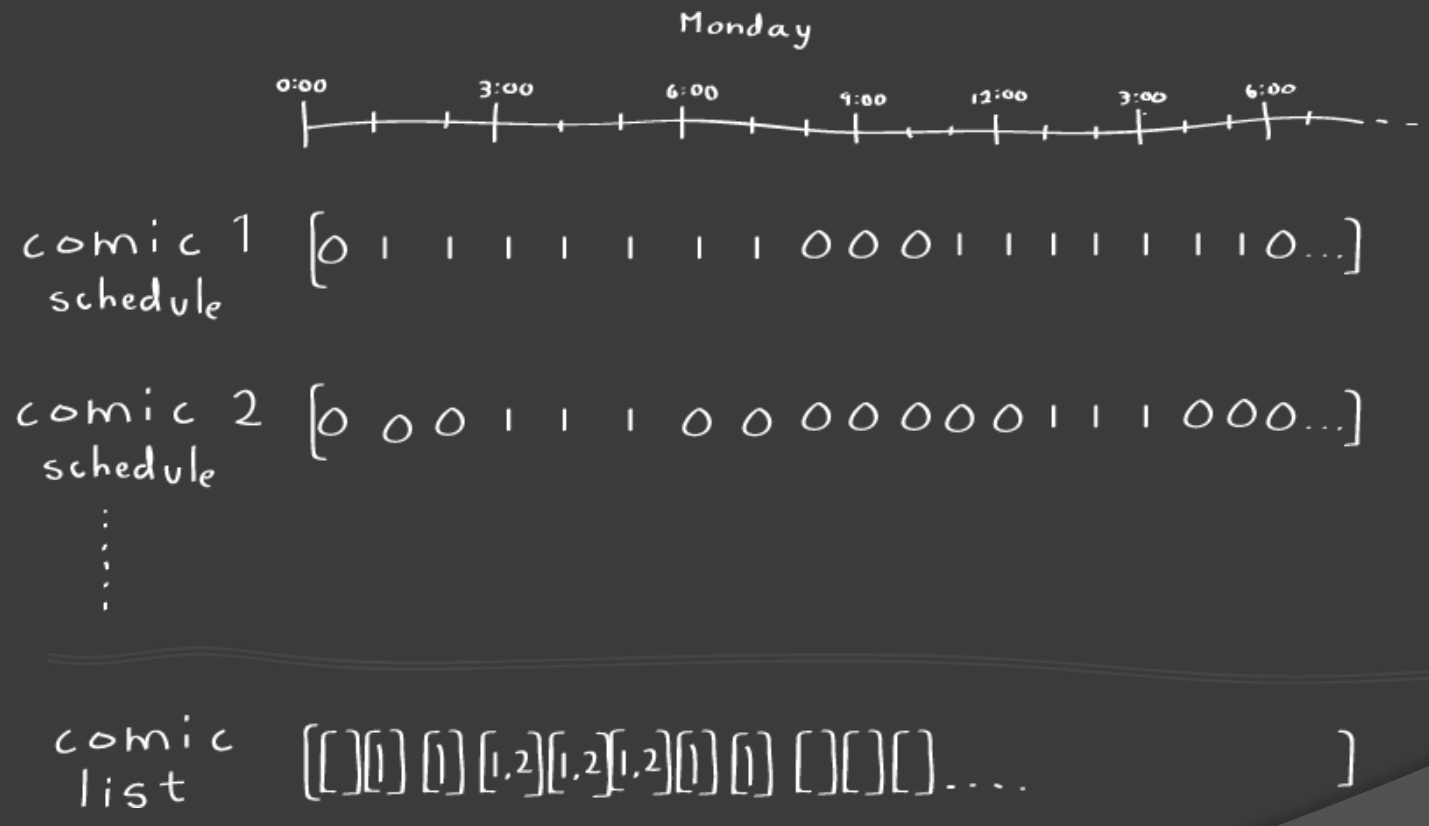
UpdateRange

```
updateHistory = [ (day, hour), ... ]
position      = averge(updateHistory)
```

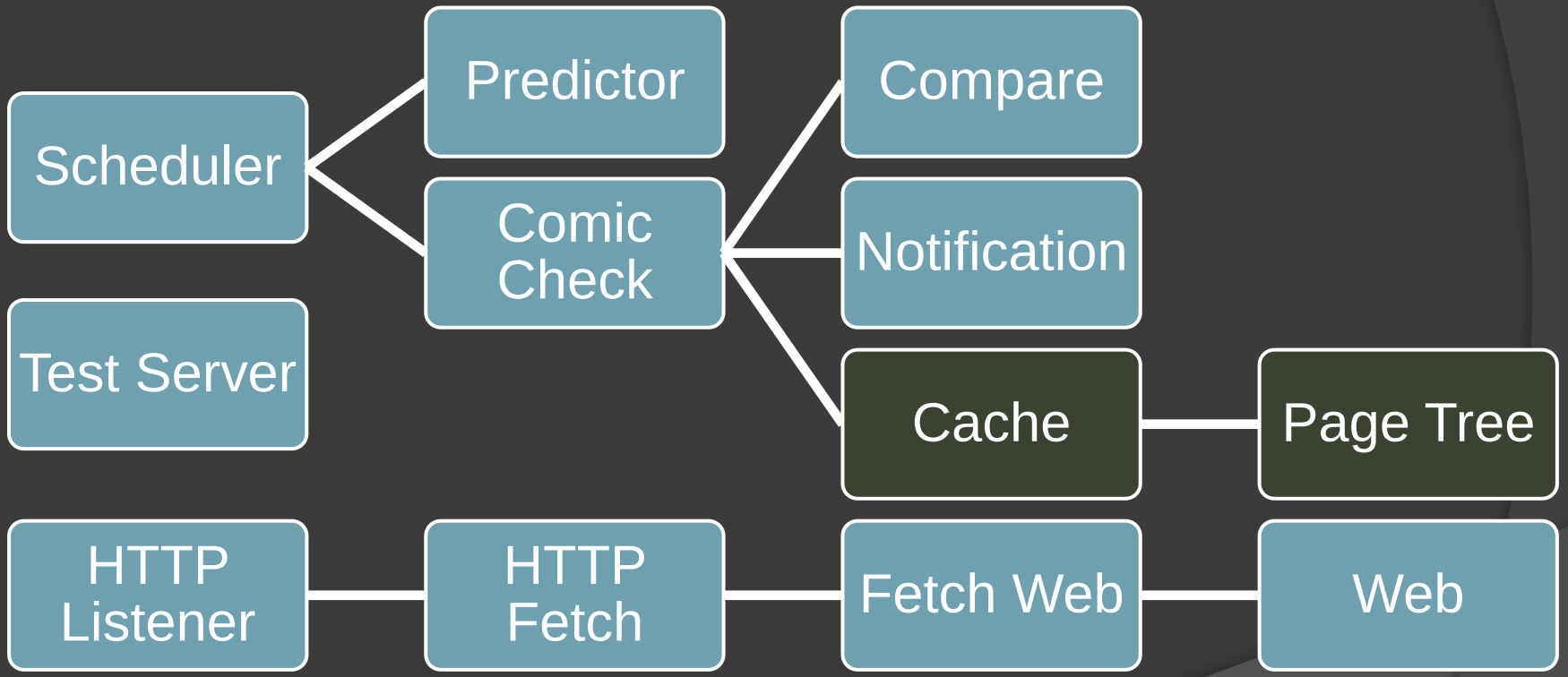
Predictor Learning



Predictor Scheduling



Suzanna Kangas



Cache

- Stores page versions for comparison
- Retrieve any past version
- Revision pushing
- File structure of storage
- No compression currently

Page Tree

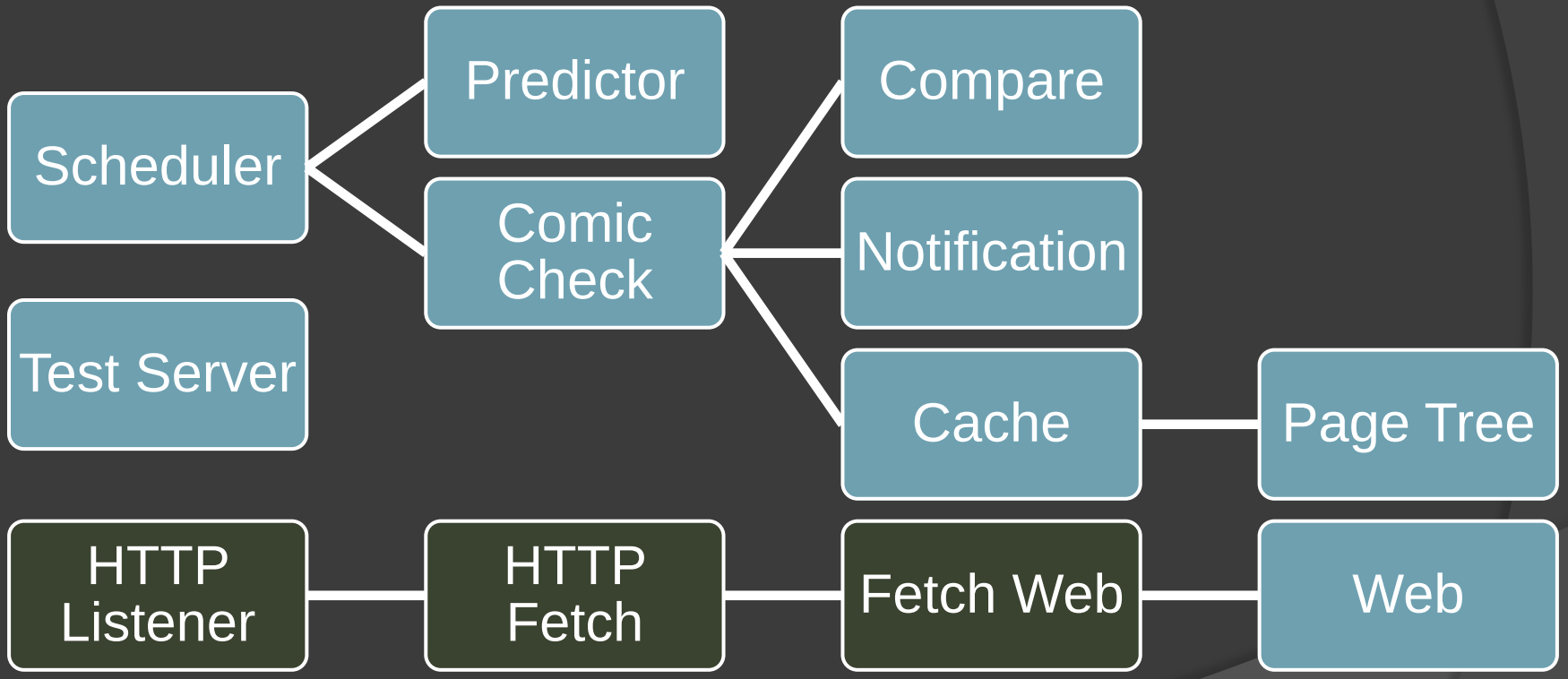
- Internal representation of a web page
- Readily extensible

What I Learned

- Python
- Working with a team
- Got a glimpse of how important testing can be


COMIC ROCKET

Lucas Berge


Scheduler
Predictor
Comic Check
Compare
Notification
Test Server
Cache
Page Tree
HTTP Listener
HTTP Fetch
Fetch Web
Web

HTTP Listener

- What it is
  - Modified SimpleHTTPRequestHandler
- What it does
  - Runs on a separate thread waiting for requests
  - Searches Cache for comicID, gets from Web if not found
  - Allows a user to send commands to predictor
  - Internal notifications

Fetch HTTP

- What it is
  - Checks cache for a page, requests from Web if not found
- What it does
  - Asks for a pointer to the object, if none call FetchWeb

Fetch Web

- What it is
  - HTTP Client that constructs pageTree objects and stores them in the cache
- What it does
  - Requests objects with urllib2
  - Parses out all links with BS4
  - Constructs a pageTree object and caches it with ComicID key
  - Internal notifications

What I Learned

- GitHub basics
  - Not the most user friendly
- Python
  - Great language
- Libraries
  - More useful when you know how they work

GitHub

- Commit Often!
  - Even if you don’t like what you have
  - Conflicts
- Stash and pull
- Use the shell
- Don’t be afraid to break everything

Python

- No prior experience
  - Easy to understand syntax
  - Lots of support and documentation
  - Human readable libraries!
- 2.7 vs 3.0
  - Urllibs
  - Whitespace
- My new language of choice

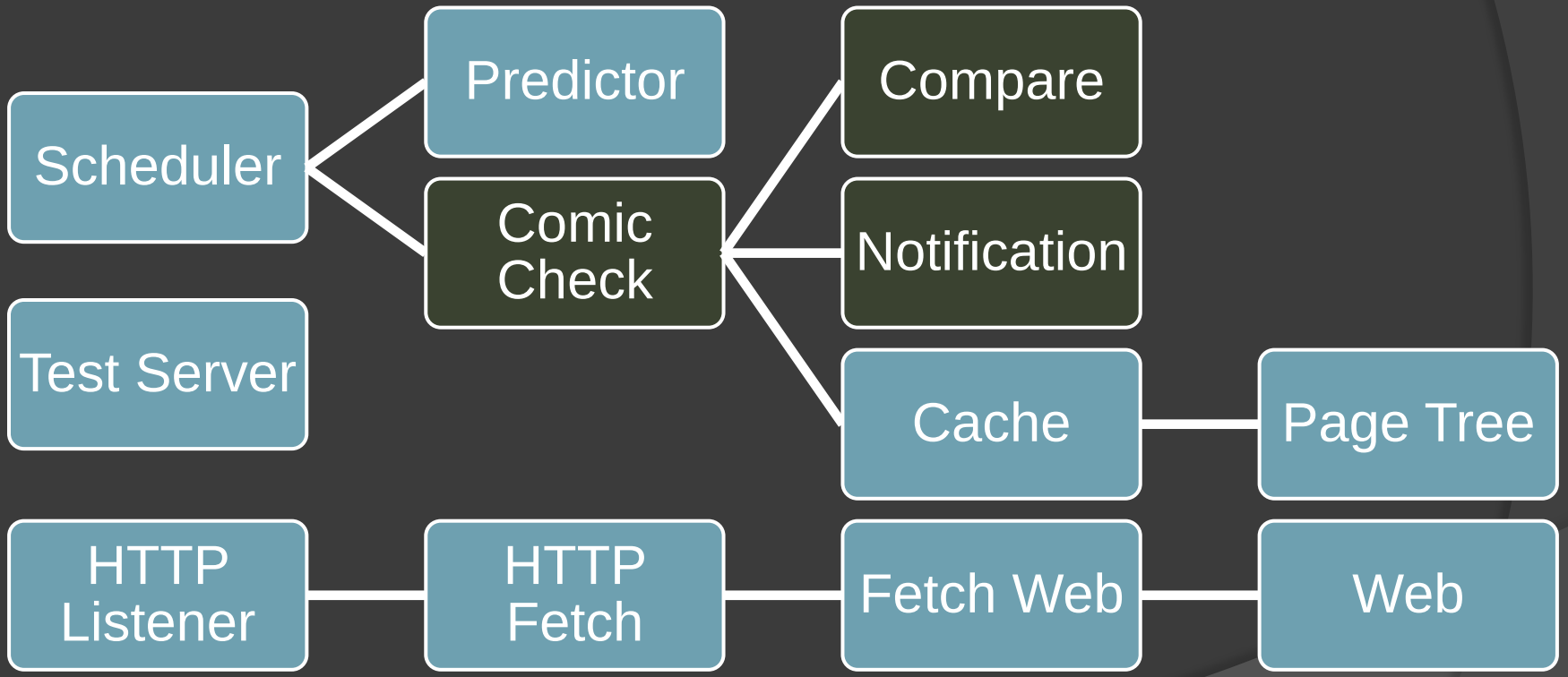
Libraries

- Urllib2
- BaseHTTPServer
- BeautifulSoup
- Threading

- Read Manuals, Look at examples, debug objects, delve into library code


COMIC ROCKET

Eric Smith


Scheduler
Predictor
Comic Check
Compare
Notification
Test Server
Cache
Page Tree
HTTP Listener
HTTP Fetch
Fetch Web
Web

Compare

- Compares two pages and reports if they are different
- Three iterations:
  - Page Data
  - Hash value
  - Links

Notification

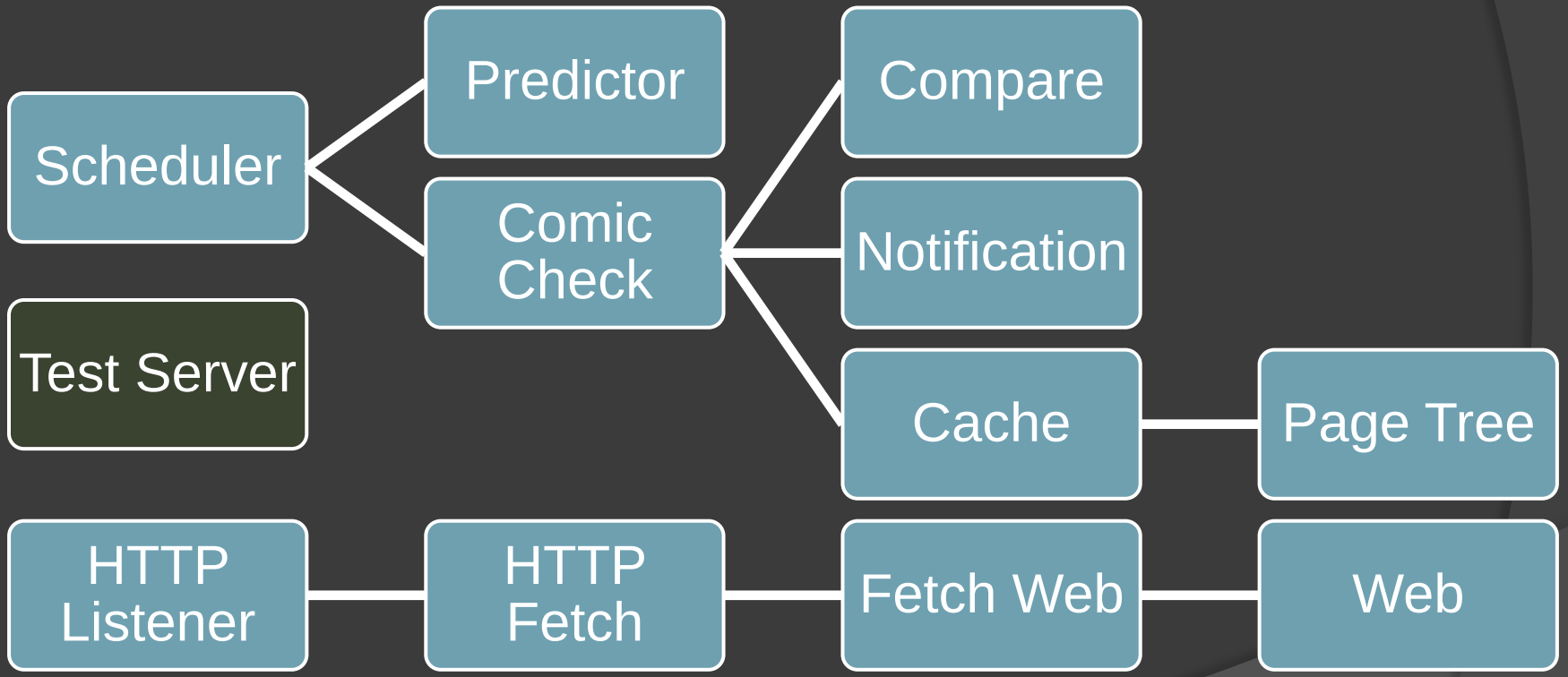
- Sends a message to Jamey via Rabbitmq
- Two iterations:
  - New comic notifications only, list of notifications sent
  - General notifications (new comic, error, etc…), no list

Comic Check

- newComic & histComic
- Wrapper functions that call the other functions

- Three iterations:
  - Latest three
  - One by one
  - List based

Joseph Schutz



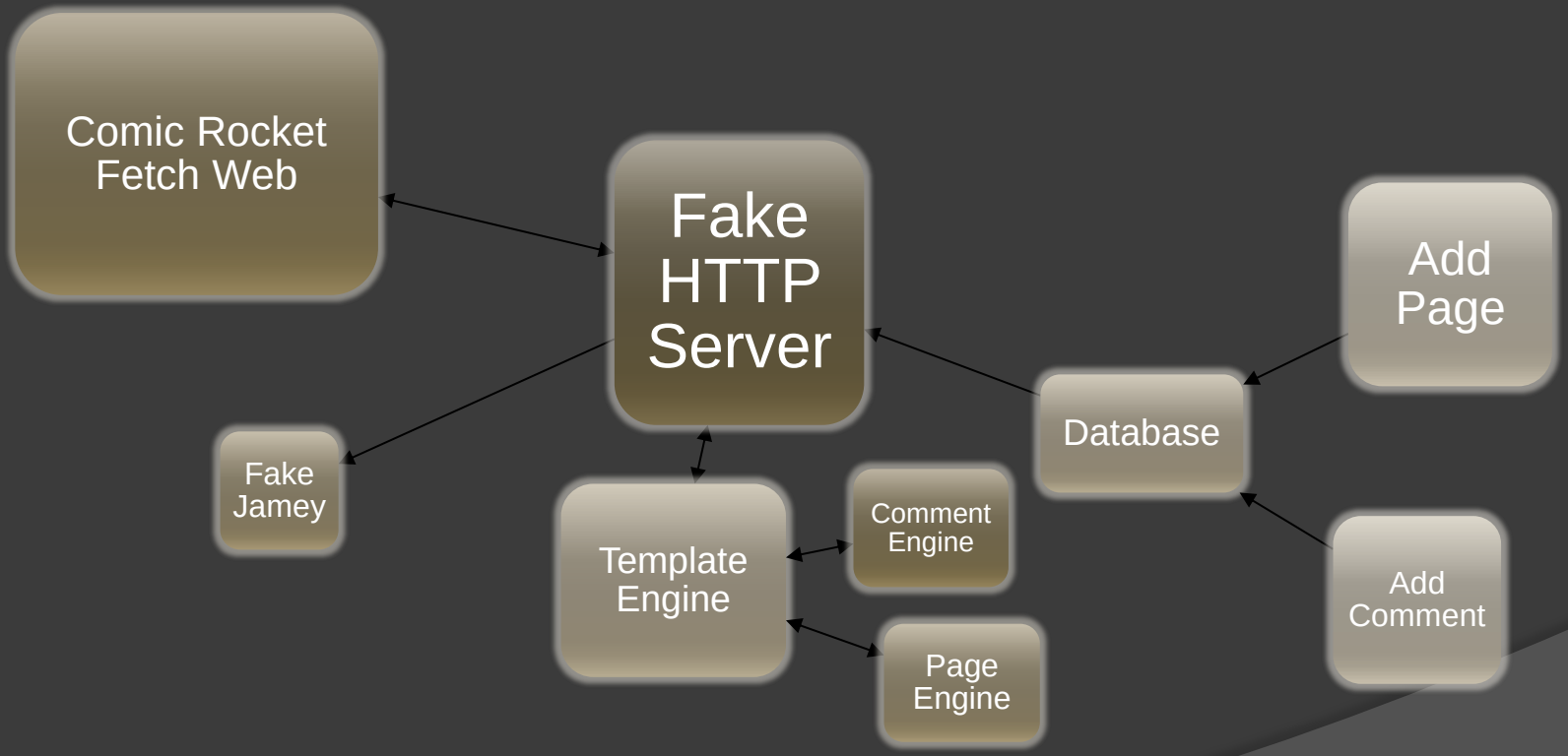
Challenges

Web Comics updated slowly.

Web Comics update in uncontrollable ways

Most Web Comics don’t appreciate extra bandwidth from “testing”

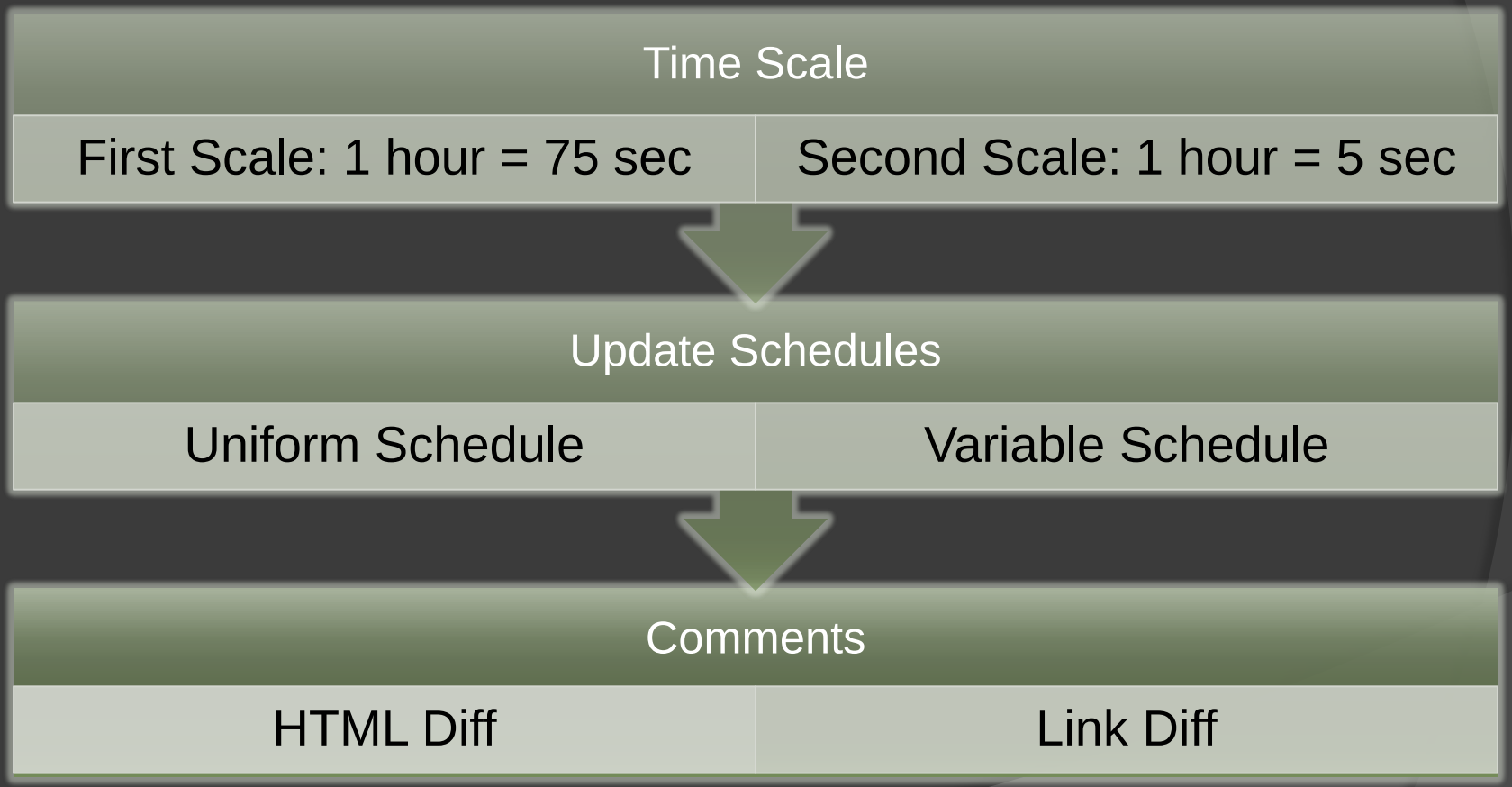
Solution


Comic Rocket Fetch Web
Fake HTTP Server
Add Page
Database
Fake Jamey
Template Engine
Comment Engine
Page Engine
Add Comment

Solution

Http Server	• Parses specific domains, and calls the template engine to build the page
Template Engine	• Builds a page, with {key = value} pairs • Builds comments in a similar manner
Crontab	• Scheduler to run scripts at specific times, ie updates for comics or comment insertion
Hosts File	• Redirects specific domains to 127.0.0.1 (or localhost)

Testing the Rocket Booster



Future Enhancements

- Should be easily extensible to store more page information
- Compression of the cache
- Filter out unimportant changes