



2021

# VTK Hackathon

# Intro to OTA Insight

# OTA Insight - Who are we?



Founded 2012 in  
London



200+ stellar employees



London, Ghent, Dallas,  
Denver, Singapore



Scale up in the  
hospitality industry



Help the hospitality industry  
to visualise their data

# Our company history

Where it all started

SEPT  
2012

A billion-dollar idea  
(So they thought)



The London Olympics

MARCH  
2013

Pivot



First customer visits  
First of many pivots  
Pooling money

SEPT  
2013

Incubator



Smart money  
Learning the industry  
16/24 – 7/7

JAN  
2014

MVP



First employees  
MVP  
Early adopter

MARCH  
2014

Product - Market fit



Strategic ground work



# Our company history



Where it all started

2016



Product team  
Sales team grows  
OTA Insight goes to the US



2018



100 employees!

OTA Insight keeps growing  
Launch RI  
30,000 partners



2019



Continuous growth

Scaling up  
#200 employees



2020



Getting ready for the upturn

Market insight launch  
Tripadvisor Partnership  
Getting ready for upturn



2021



Road to recovery



# Where will you find our teams?

With more than 30 nationalities on the team, we truly are a global company



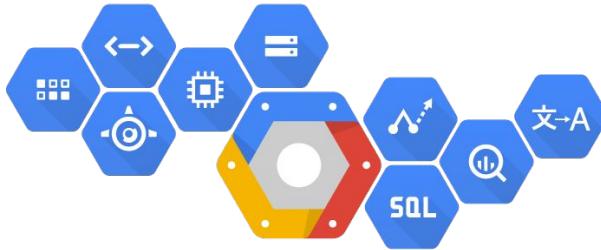
## Facts & Figures

Signed hotels	55,000
Signed locations	185
Employees	200+
Data collection	1.5B
Products	4

# Facts & Figures



- We crunch (actual) big data
  - 100 TB of data processed per day (= binge-watching Netflix for 4 years)
  - $3 \times 10^{12}$  of rates data points in database (= amount of fish in the sea)
  - 2 million reservations processed per day
- We are passionate about technology



Google Cloud Platform



Grafana



python



pandas  
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



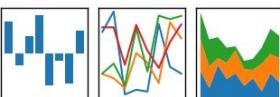
Stackdriver



Technology company  
at heart



kubernetes



docker



GitLab

# we're building the #bestteamever

## Frontend engineers

You'll be creating complex **single-page MVC applications**, visualizing large amounts of data in an actionable way.

## Data engineers

You'll build **high-performance data processing pipelines** which can guarantee industry's best data integrity and quality.

## Backend engineers

You'll architect, code and maintain our **API backend and microservices infrastructure**.

*More info @ [www.careers.otainsight.com](http://www.careers.otainsight.com)*



# **Workshop content**

# Content of tonight:

We in the company do a lot of data gathering and analysis of that gathered data, and tonight you'll do the same (on a very limited level)

- First part: web scraping of a website to gather data
- Second part: data science on existing dataset

Both parts can/should be done in parallel

There are questions for you to answer to gain points



# Content of tonight:

<https://github.com/OTA-Insight/vtk-hackathon-2021>

For the website crawling part: read through to see the questions you need to answer! You can answer questions partially to us!

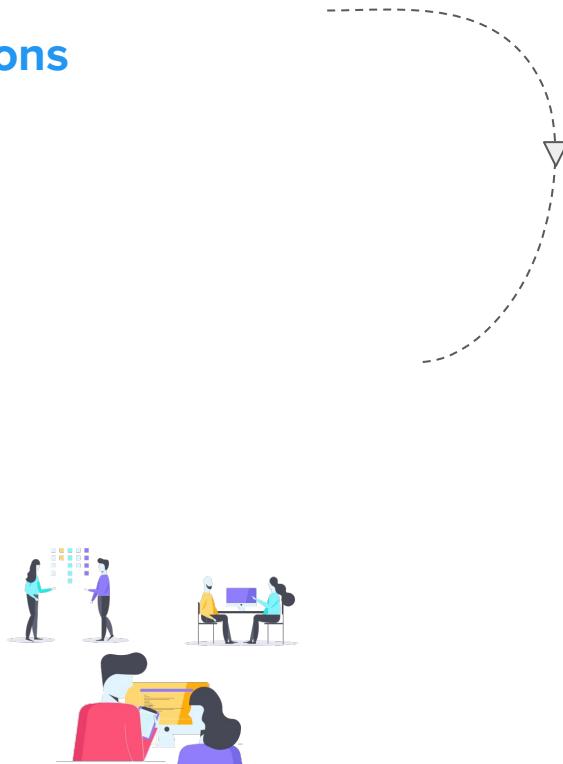
First to answer a particular question gets more points than the next one.



# Content of tonight:

<https://github.com/OTA-Insight/vtk-hackathon-2021>

- Contains a readme with the workshop instructions
- Contains skeleton code to start with



# Content of tonight:

<https://github.com/OTA-Insight/vtk-hackathon-2021>

We are available for questions and help

At the end of the night (let's see how everybody progresses for an end time) we'll shortly go over the solutions

Feedback welcome! [mathieu@otainsight.com](mailto:mathieu@otainsight.com) / [lana@otainsight.com](mailto:lana@otainsight.com)



# Solutions

# Web crawling: countermeasures



- /robots.txt

A screenshot of a web browser window displaying a robots.txt file. The address bar shows the URL `35.233.25.116/robots.txt` and indicates it is "Not Secure". The page content is a plain text file with the following rules:

```
User-agent: *
Allow: /sitemap/hotels/Amsterdam/
Allow: /sitemap/hotels/Brussels/
Allow: /sitemap/hotels/Paris/
Disallow: /sitemap/hotels/
Allow: /rates/Amsterdam/
Allow: /rates/Brussels/
Allow: /rates/Paris/
Disallow: /rates/
```

- <https://moz.com/learn/seo/robotstxt>

# Web crawling: countermeasures



- User Agent

```
--  
19  
20     class UserAgentDownloaderMiddleware:  
21         user_agent_choices = [  
22             "Edg/93", "Edg/94", "Edg/95", "Edg/96",  
23             "Chrome/93", "Chrome/94", "Chrome/95", "Chrome/96",  
24             "Firefox/93", "Firefox/94", "Firefox/95", "Firefox/96",  
25             "Safari/603", "Safari/604", "Safari/605", "Safari/606",  
26             "Trident/7",  
27         ]  
28  
29         def process_request(self, request, spider):  
30             # user_agent = request.headers.get('User-Agent')  
31             # Look at what user_agent is by default ...  
32  
33             # Replace by a random choice from a good list  
34             request.headers['User-Agent'] = random.choice(self.user_agent_choices)  
35  
36             return None  
37
```

- Default = 'Scrapy/\*\*\*\*'

# Web crawling: countermeasures



- Rate limiting

```
23 // rateLimit will apply a ratelimit for certain hotels + rates
24 // rate limit consists of the destination - ip address - user agent
25 func rateLimit(DestinationID string) adapter {
26     return func(h http.Handler) http.Handler {
27         return http.HandlerFunc(func(w http.ResponseWriter, r *http.Request) {
28
29
30         // Call the getVisitor function to retrieve the rate limiter for the current key.
31         limiter := getVisitor(DestinationID + r.RemoteAddr + toUserAgentType(r.UserAgent()))
32         if !limiter.Allow() {
33             WriteJsonResponse(w, http.StatusTooManyRequests, nil)
34             return
35         }
36
37         h.ServeHTTP(w, r)
38     })
39 }
40 }
41 }
```

- <https://newbedev.com/how-to-handle-a-429-too-many-requests-response-in-scrapy>

# Web crawling: countermeasures



- Rate limiting

```
38
39     class RateLimitRetryMiddleware(RetryMiddleware):
40         def process_response(self, request, response, spider):
41             if request.meta.get('dont_retry', False):
42                 return response
43             elif response.status == 429:
44                 self.crawler.engine.pause()
45                 time.sleep(2)
46                 self.crawler.engine.unpause()
47                 reason = response_status_message(response.status)
48                 return self._retry(request, reason, spider) or response
49             elif response.status in self.retry_http_codes:
50                 reason = response_status_message(response.status)
51                 return self._retry(request, reason, spider) or response
52             return response
53
54
```

- <https://newbedev.com/how-to-handle-a-429-too-many-requests-response-in-scrapy>

# Web crawling: countermeasures



- Content differences: room count & star rating

Inntel Hotels Amsterdam Centre [239 rooms]

Nieuwezijds Kolk 19  
1012 PV Amsterdam  
Netherlands

Coordinates: Lat 52.376186, Long 4.894449

This hotel has 4 stars

Newhotel Charlemagne\*\*\*\*

Boulevard Charlemagne 25-27  
1000 Brussel  
Belgium

Coordinates: Lat 50.845184, Long 4.3826704

There are 68 rooms in this hotel

Riu Plaza Berlin

Martin-Luther-Straße 1  
10777 Berlin  
Germany

Coordinates: Lat 52.500294, Long 13.3467455

There are 357 rooms in this hotel

This hotel has 4 stars

# Web crawling: countermeasures

- Content differences: refundable & breakfast

No free breakfast included

Refundable

Breakfast at additional cost

Can be cancelled free of charge

No breakfast

Free cancellation

Breakfast at additional cost

No fee upon cancellation

Breakfast included: No

Refundable: Yes

# Web crawling: countermeasures



- Broken html with similar look but different



Room:

Roomname: Deluxe Double Room

Price: EUR 123.24

Breakfast at additional cost

Free cancellation

Number of guests: 2



Room:

Roomname: Deluxe Room

Price: GBP 215

Breakfast included: No

Can be cancelled free of charge

Number of guests: 2

# Web crawling: countermeasures



- Header required: base64 encoded path

The screenshot shows a browser developer tools interface with the Network tab selected. A request to 'app.js' is highlighted. The request headers pane shows the following:

```
Request Headers View source
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9
Accept-Encoding: gzip, deflate
Accept-Language: en-GB,en-US;q=0.9,en;q=0.8,nl;q=0.7
Cache-Control: no-cache
Connection: keep-alive
Cookie: controlId=L3JhdGVzL0Jlcmbi8xMzc3MDczLz8mZGVzdGluYXRpb2490mVybGluJmFycml2YWxEXRlPTIwMjEtMTEtMjUmZGVwYXJ0dXJlRGF0ZT0yMDIxLTExLTI2Jm51bVBlcnNvbN9Mg==
Host: 35.233.25.116
Pragma: no-cache
Referer: http://35.233.25.116/results/?destination=Berlin&arrivalDate=2021-11-25&departureDate=2021-11-26&numPersons=2
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/95.0.4638.69 Safari/537.36
```

# Web crawling: countermeasures



- Header required: base64 encoded path

```
54
55  class ControlIDCookiesMiddleware:
56      def process_request(self, request, spider):
57          if '/Berlin' in request.url:
58              path = request.url.removeprefix(server_location)
59              encoded_path = base64.b64encode(path.encode()).decode('utf-8')
60              request.cookies['controlid'] = encoded_path
61
62          return None
```

# Data science: exceptional deals



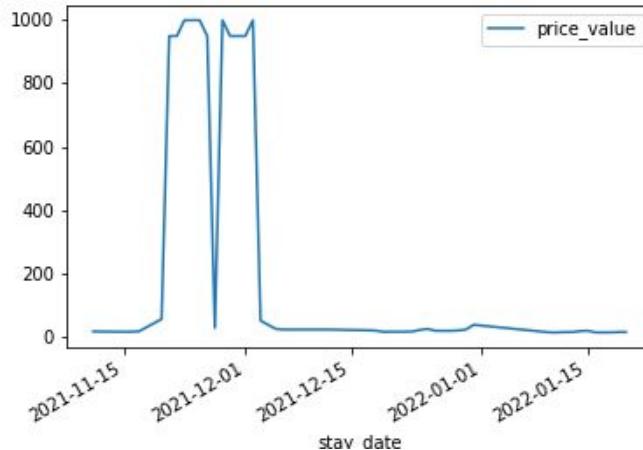
- Normalize prices to price index
- Sort by price index to find most exceptional prices
- Normalizer should be robust

# Data science: exceptional deals



Normalizing by mean price value

- Mean is sensitive to outliers
- Lowest price index rate:
  - price value: 14 euro
  - normalizer: 222 euro
  - index: 0.06

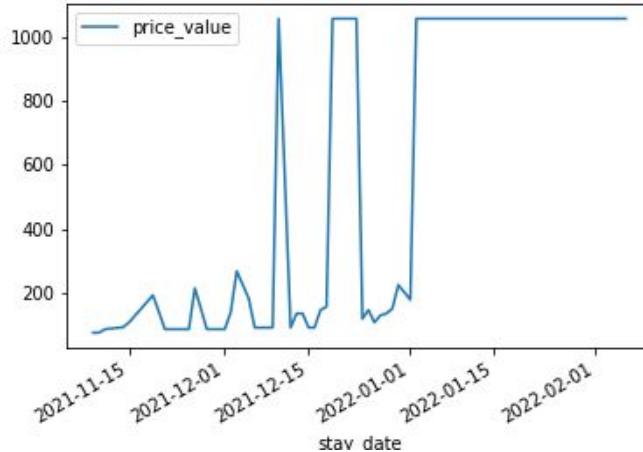


# Data science: exceptional deals



Normalizing by median price value

- Issue when more than 50% is unreal price
- Lowest price index rate:
  - price value: 76 euro
  - normalizer: 1056 euro
  - index: 0.07

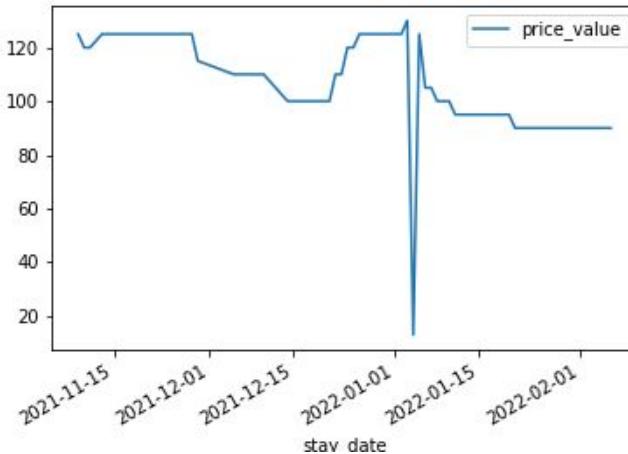


# Data science: exceptional deals



Normalizing by 10th percentile price value

- Robust method of describing usual price value
- Lowest price index rate:
  - price value: 13 euro
  - normalizer: 90 euro
  - index: 0.14

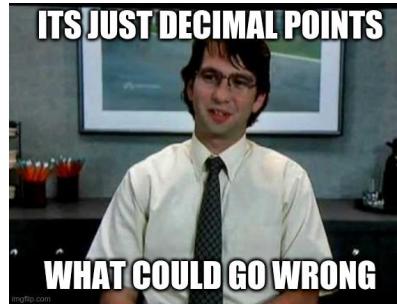




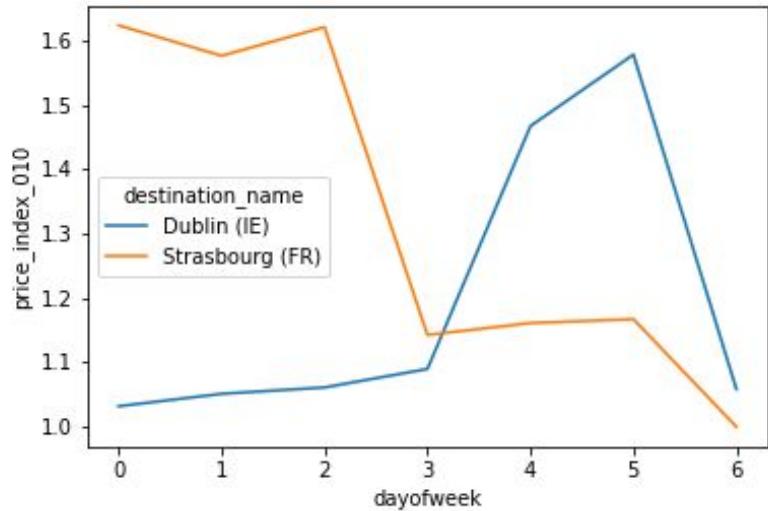
# Data science: exceptional deals

<b>Suite with Terrace 40 m<sup>2</sup></b>	<b>2</b>	<b>€ 135</b>	<b>Includes taxes and charges</b>	<b>Superb breakfast included</b>	<b>0</b>
<b>Only 2 left on our site</b>					
1 sofa bed  and 1 large double bed					
Balcony  City view					
Air conditioning  Patio					
Ensuite bathroom  Dishwasher					
Flat-screen TV  Soundproofing					
Terrace  Coffee machine  Free WiFi					
✓ Free toiletries ✓ Kitchen					
✓ Safety deposit box ✓ Washing machine					
✓ Streaming service (like Netflix) ✓ Toilet					
✓ Sofa ✓ Bath or shower					
✓ Hardwood or parquet floors ✓ Towels					
✓ Linen ✓ Socket near the bed					
✓ Cleaning products ✓ Hypoallergenic					
✓ Tile/marble floor ✓ Desk ✓ Seating Area					
✓ TV ✓ Slippers ✓ Refrigerator					
✓ Telephone ✓ Ironing facilities					
✓ Satellite channels ✓ Tea/Coffee maker					
✓ Iron ✓ Microwave ✓ Heating					
✓ Hairdryer ✓ Kitchenware ✓ Kitchenette					
✓ Extra long beds (> 2 metres)					
✓ Dressing room					
✓ Wake up service/Alarm clock					
✓ Electric kettle ✓ Outdoor furniture					
✓ Outdoor dining area ✓ Cable channels					
✓ Tumble dryer ✓ Wardrobe or closet					
✓ Oven ✓ Stovetop ✓ Toaster					
✓ Dining area ✓ Dining table					
✓ Clothes rack ✓ Fold-up bed					
✓ Drying rack for clothing ✓ Toilet paper					
✓ Sofa bed ✓ Hand sanitiser					
<a href="#">Less</a>					
<b>January 5</b>					

<b>Suite with Terrace 40 m<sup>2</sup></b>	<b>2</b>	<b>€ 23</b>	<b>Includes taxes and charges</b>	<b>Superb breakfast included</b>	<b>0</b>
<b>Only 2 left on our site</b>					
1 sofa bed  and 1 large double bed					
Balcony  City view					
Air conditioning  Patio					
Ensuite bathroom  Dishwasher					
Flat-screen TV  Soundproofing					
Terrace  Coffee machine  Free WiFi					
✓ Free toiletries ✓ Kitchen					
✓ Safety deposit box ✓ Washing machine					
✓ Streaming service (like Netflix) ✓ Toilet					
✓ Sofa ✓ Bath or shower					
✓ Hardwood or parquet floors ✓ Towels					
✓ Linen ✓ Socket near the bed					
✓ Cleaning products ✓ Hypoallergenic					
✓ Tile/marble floor ✓ Desk ✓ Seating Area					
✓ TV ✓ Slippers ✓ Refrigerator					
✓ Telephone ✓ Ironing facilities					
✓ Satellite channels ✓ Tea/Coffee maker					
✓ Iron ✓ Microwave ✓ Heating					
✓ Hairdryer ✓ Kitchenware ✓ Kitchenette					
✓ Extra long beds (> 2 metres)					
✓ Dressing room					
✓ Wake up service/Alarm clock					
✓ Electric kettle ✓ Outdoor furniture					
✓ Outdoor dining area ✓ Cable channels					
✓ Tumble dryer ✓ Wardrobe or closet					
✓ Oven ✓ Stovetop ✓ Toaster					
✓ Dining area ✓ Dining table					
✓ Clothes rack ✓ Fold-up bed					
✓ Drying rack for clothing ✓ Toilet paper					
✓ Sofa bed ✓ Hand sanitiser					
<b>January 4</b>					

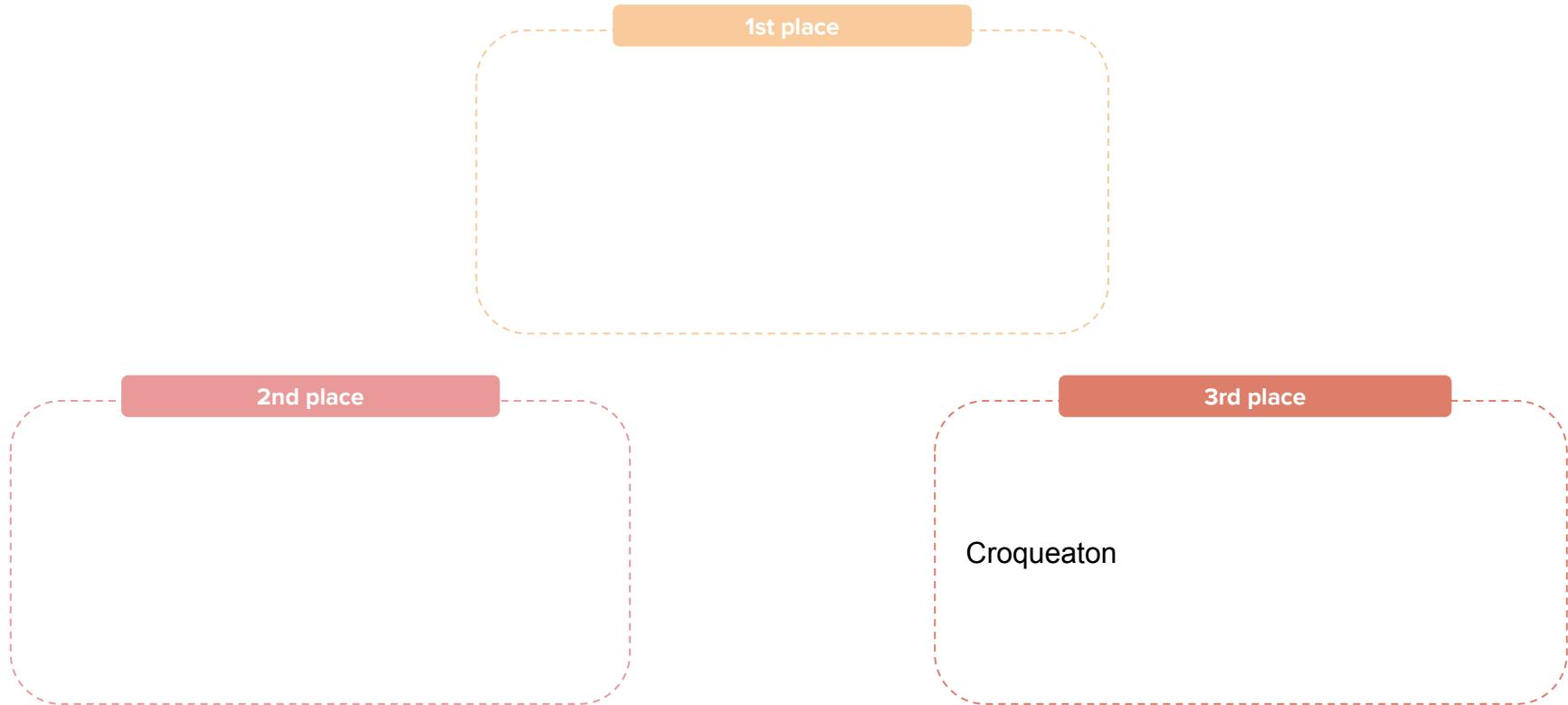


# Data science: destination patterns



**\*Winners!\***

# And the winners are...



# And the winners are...



1st place

2nd place

3rd place

The Brogrammers

Croqueaton

# And the winners are...



1st place

The Dream Team

2nd place

The Brogrammers

3rd place

Croqueaton



# Q&A



ota-insight



@otainsight



/ otainsight

[www.otainsight.com](http://www.otainsight.com)