

GALIASdoc: Sistema para la extracción automática de información estructurada a partir de documentos con elementos de maquetación comunes

Diego Trabazo Sardón

Director: José Carlos Dafonte Vázquez

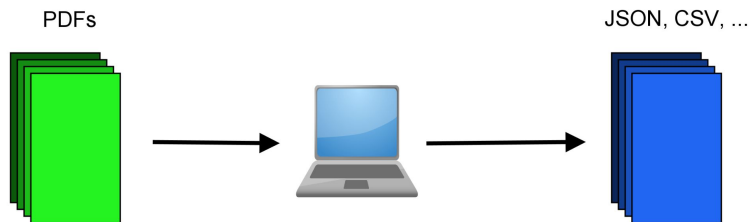
¿En qué consiste?

- ⦿ Propuesta inicial de Odeene Ingeniería para la **extracción automática** de información de facturas y albaranes
- ⦿ Estas tareas se continúan realizando manualmente
- ⦿ Especialmente indicado cuando hay proveedores con un **gran número de documentos**



Oportunidades del proyecto

- ⊙ Aplicable siempre y cuando existan documentos que compartan un modelo de **maquetación común**
- ⊙ Desde PDF a formatos estructurados con contenido **personalizado** (JSON, CSV, ...)
- ⊙ Procesamiento automático: rompe la relación entre el volumen de trabajo y las horas/hombre necesarias



Información manejada

Investigación inicial →
contenido + **coordenadas**
+ plantillas

Dos tipos de PDF según el
tipo de contenido: basados
en texto o imagen

- ③ Las plantillas son descripciones de las regiones de interés en formato JSON
- ③ Una única plantilla por cada tipo de documento
- ③ Indican las coordenadas y tipo de las regiones

Regiones

Factura

Contenido secuencial

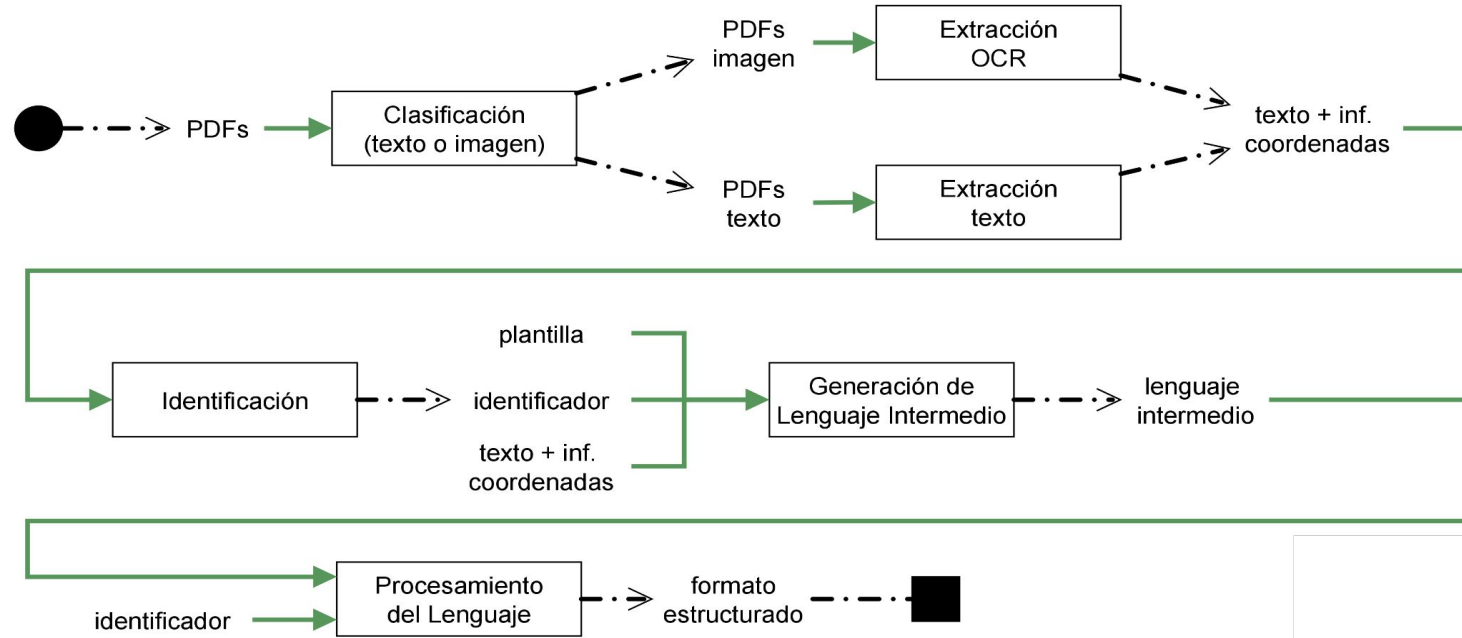
1 fila, n columnas

Factura	Fecha	Referencia	Fecha de vencimiento	NIF/CIF
5F 201910161	31/12/2019		01/01/2020	B70264312

Fecha	Descripción	Importe
31/12/2019	1 us Office 365 Empresa Essentials de 1/1 al 31/1	4,00
31/12/2019	3 us Quiosco de Exchange Online de 4/12 al 31/12	4,31
31/12/2019	62 us Quiosco de Exchange Online de 1/1 al 31/1	98,58
31/12/2019	51 us Exchange Online (plan 1) de 1/1 al 31/1	163,20
31/12/2019	2 us Office 365 Empresa de 1/1 al 31/1	17,60
31/12/2019	1 us Power BI Pro de 1/1 al 31/1	8,40
31/12/2019	1 us Office 365 Empresa Premium de 13/12 al 19/12	2,30
31/12/2019	2 us Office 365 Empresa Premium de 20/12 al 31/12	7,90
31/12/2019	28 us Office 365 Empresa Premium de 1/1 al 31/1	285,60
31/12/2019	2 us Exchange Online (plan 2) de 1/1 al 31/1	13,00
31/12/2019	2 us Office 365 E3 de 1/1 al 31/1	37,80

Tabla de datos (T1)

Flujo de información



Herramientas utilizadas

El proyecto está desarrollado empleando principalmente software libre: Ubuntu, Bison, OpenCV...

- ◎ Barreras de entrada más bajas: no necesito negociar licenciamiento por uso, acudir a reuniones, ...
- ◎ Apoyo a estándares existentes como hOCR
- ◎ La funcionalidad más utilizada de pdftotext la aportó una persona ajena al proyecto Poppler

Liberación del proyecto

La liberación del proyecto fue un proceso:

- ⦿ Permiso a Odeene, la empresa licenciada
- ⦿ Usos de librerías de terceros: lista, strbuf y cJSON
- ⦿ Selección de la licencia: GPL v.3
- ⦿ Creación de un repositorio específico
- ⦿ Adición de la licencia al proyecto




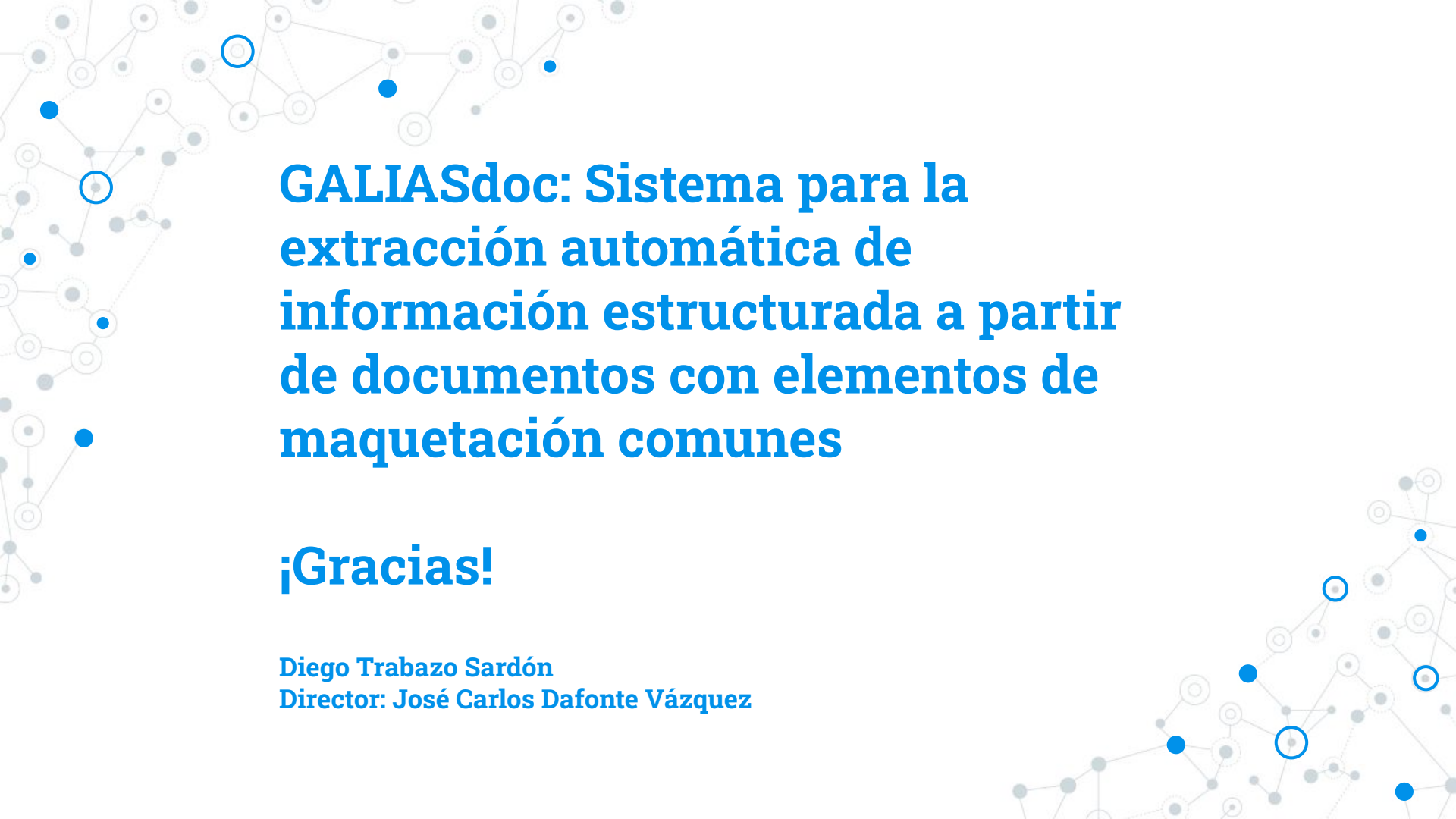


Para finalizar

No existen soluciones libres en el ámbito de este trabajo.
Supone una gran satisfacción poder compartirlo, me parece la mejor oportunidad para que el esfuerzo realizado no se pierda

Existen mejoras que se pueden aportar, como evolucionar la salida actual del programa pdftotext al microformato hOCR





GALIASdoc: Sistema para la extracción automática de información estructurada a partir de documentos con elementos de maquetación comunes

¡Gracias!

Diego Trabazo Sardón
Director: José Carlos Dafonte Vázquez

Instructions for use

EDIT IN GOOGLE SLIDES

Click on the button under the presentation preview that says "Use as Google Slides Theme".

You will get a copy of this document on your Google Drive and will be able to edit, add or delete slides.

You have to be signed in to your Google account.

EDIT IN POWERPOINT®

Click on the button under the presentation preview that says "Download as PowerPoint template". You will get a .pptx file that you can edit in PowerPoint.

Remember to download and install the fonts used in this presentation (you'll find the links to the font files needed in the [Presentation design slide](#))

More info on how to use this template at
www.slidescarnival.com/help-use-presentation-template

This template is free to use under [Creative Commons Attribution license](#). You can keep the Credits slide or mention SlidesCarnival and other resources used in a slide footer.

Hello!

I am Jayden Smith

I am here because I love to
give presentations.

You can find me at:

@username



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue.

1.

Transition headline

Let's start with the first set of slides

A decorative graphic at the top of the slide featuring a network of interconnected nodes and lines. A central node is highlighted with a dashed circle and contains a large blue quotation mark.

“


*Quotations are commonly printed
as a **means of inspiration** and to
invoke philosophical thoughts from
the reader.*



This is a slide title

- ◎ Here you have a list of items
- ◎ And some text
- ◎ But remember not to overload your slides with content

Your audience will listen to you or read the content, but won't do both.



Big concept

Bring the attention of your audience over a key concept using icons or illustrations



You can also split your content

White

Is the color of milk and fresh snow, the color produced by the combination of all the colors of the visible spectrum.

Black

Is the color of ebony and of outer space. It has been the symbolic color of elegance, solemnity and authority.

In two or three columns

Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.

Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.

A picture is worth a thousand words

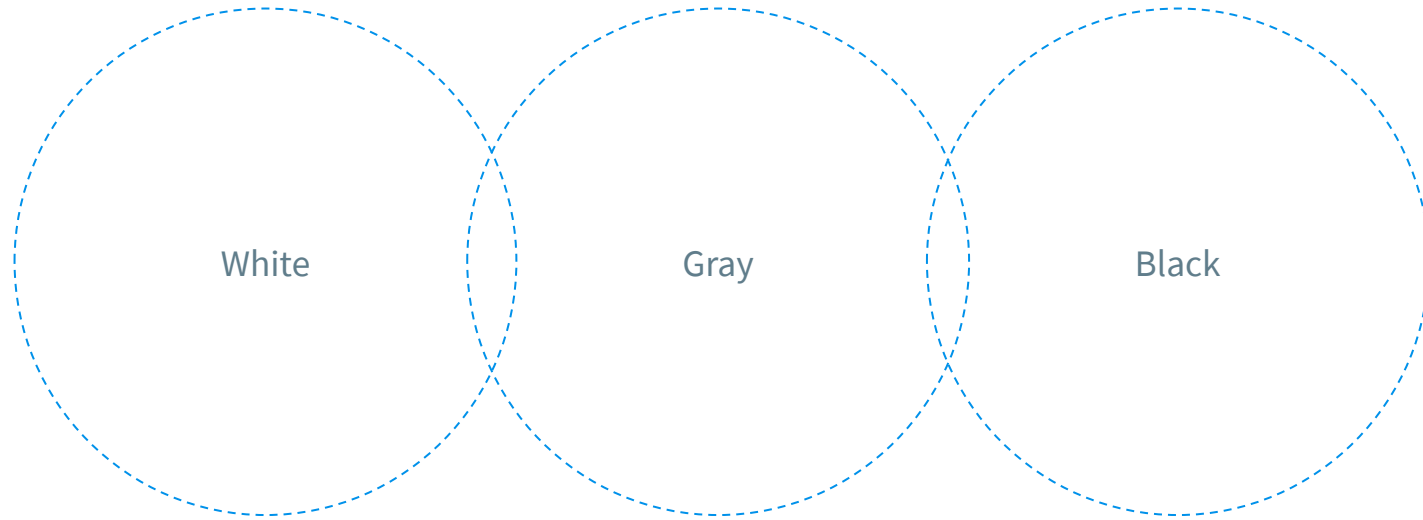
A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.



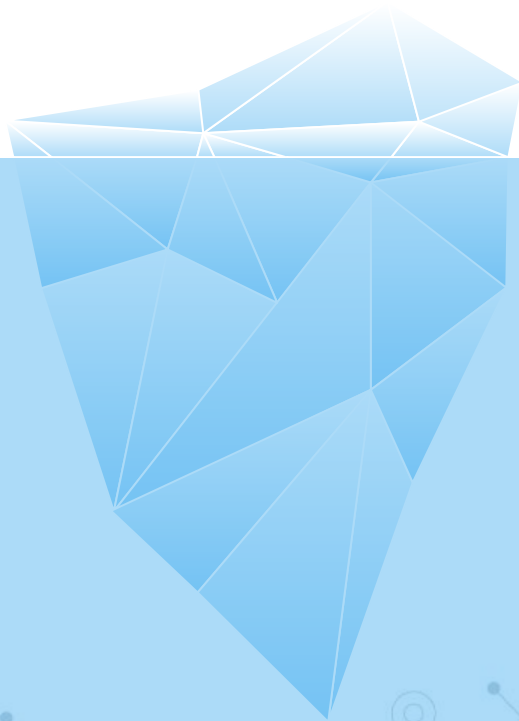


**Want big
impact?**
Use big image.

Use charts to explain your ideas



Or diagrams to explain complex ideas



Example text.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam venenatis nisi at nisl tempor, et luctus diam lobortis. Nulla sit amet metus consequat velit iaculis tempor.

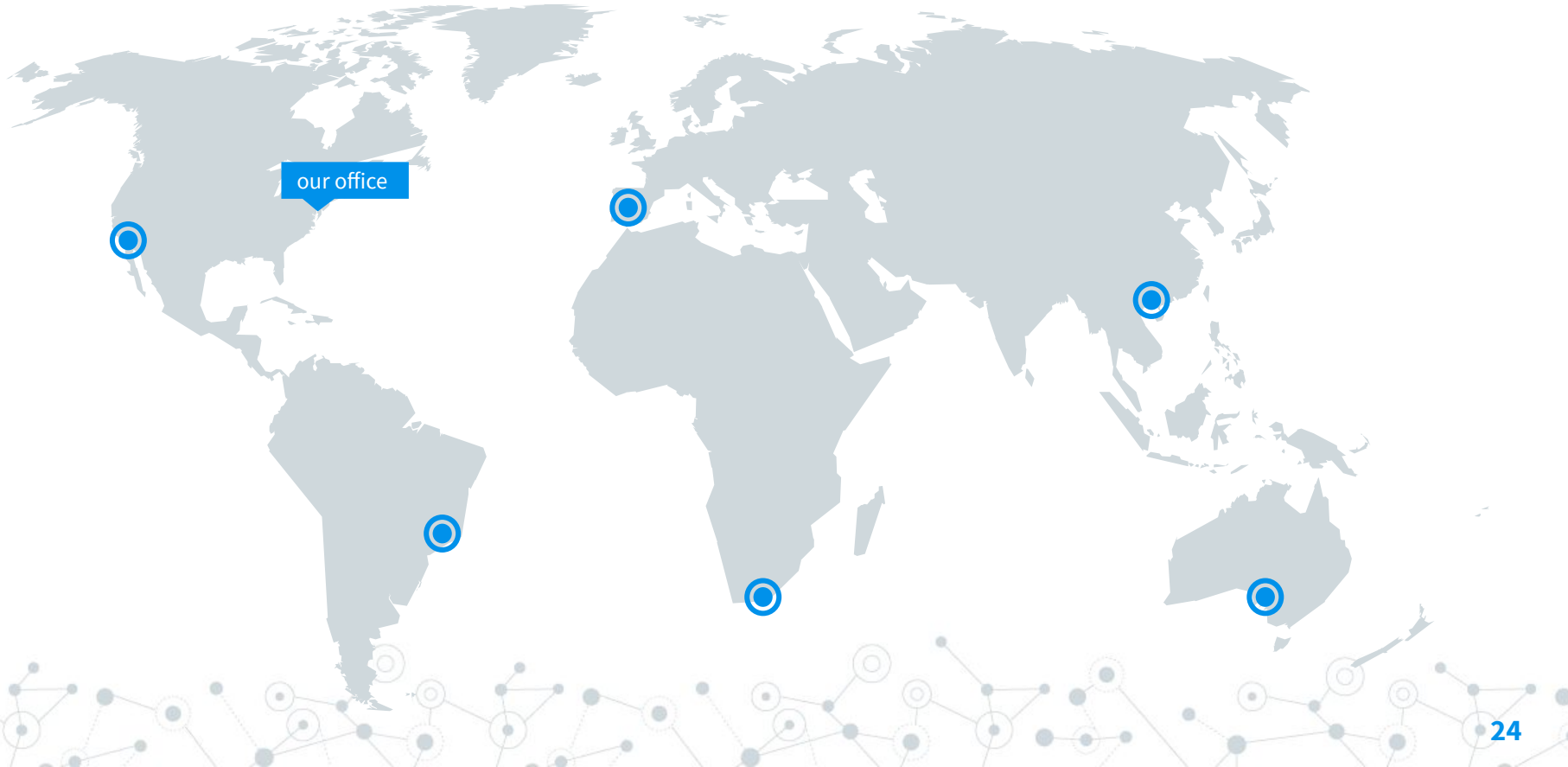
Example text.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam venenatis nisi at nisl tempor, et luctus diam lobortis. Nulla sit amet metus consequat velit iaculis tempor.

And tables to compare data

	A	B	C
Yellow	10	20	7
Blue	30	15	10
Orange	5	24	16

Maps



The background of the slide features a light gray network pattern. It consists of numerous small circles, some of which are double-lined, connected by thin, light gray lines. These connections form a complex, web-like structure that fills the entire background.

89,526,124

Whoa! That's a big number, aren't you proud?

Presentation design

This presentations uses the following typographies and colors:

- Titles: **Roboto Slab**
- Body copy: **Source Sans Pro**

Download for free at:

<https://www.fontsquirrel.com/fonts/roboto-slab>

<https://www.fontsquirrel.com/fonts/source-sans-pro>

You don't need to keep this slide in your presentation. It's only here to serve you as a design guide if you need to create new slides or download the fonts to edit the presentation in PowerPoint®

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue.

89,526,124\$

That's a lot of money

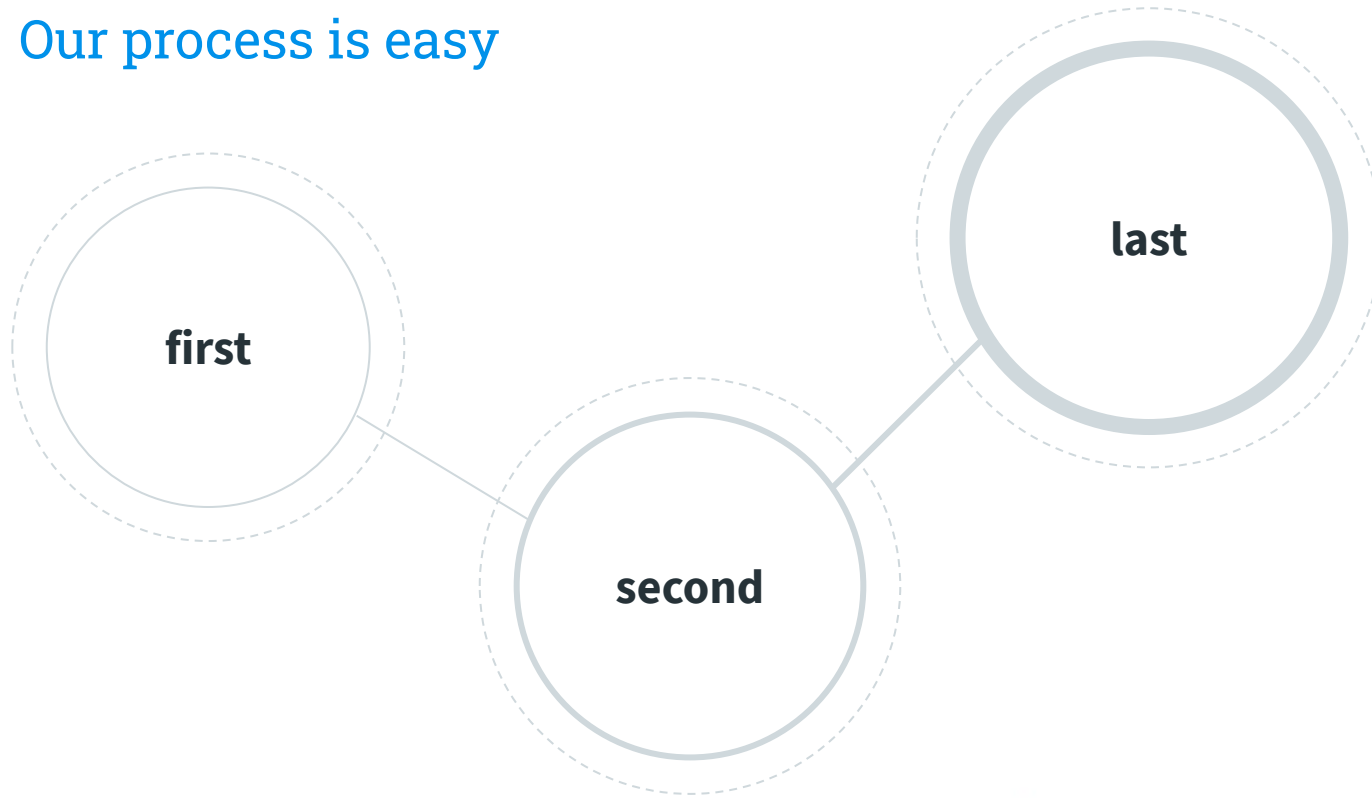
185,244 users

And a lot of users

100%

Total success!

Our process is easy



Let's review some concepts



Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.



Yellow

Is the color of gold, butter and ripe lemons. In the spectrum of visible light, yellow is found between green and orange.



Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.



Blue

Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.



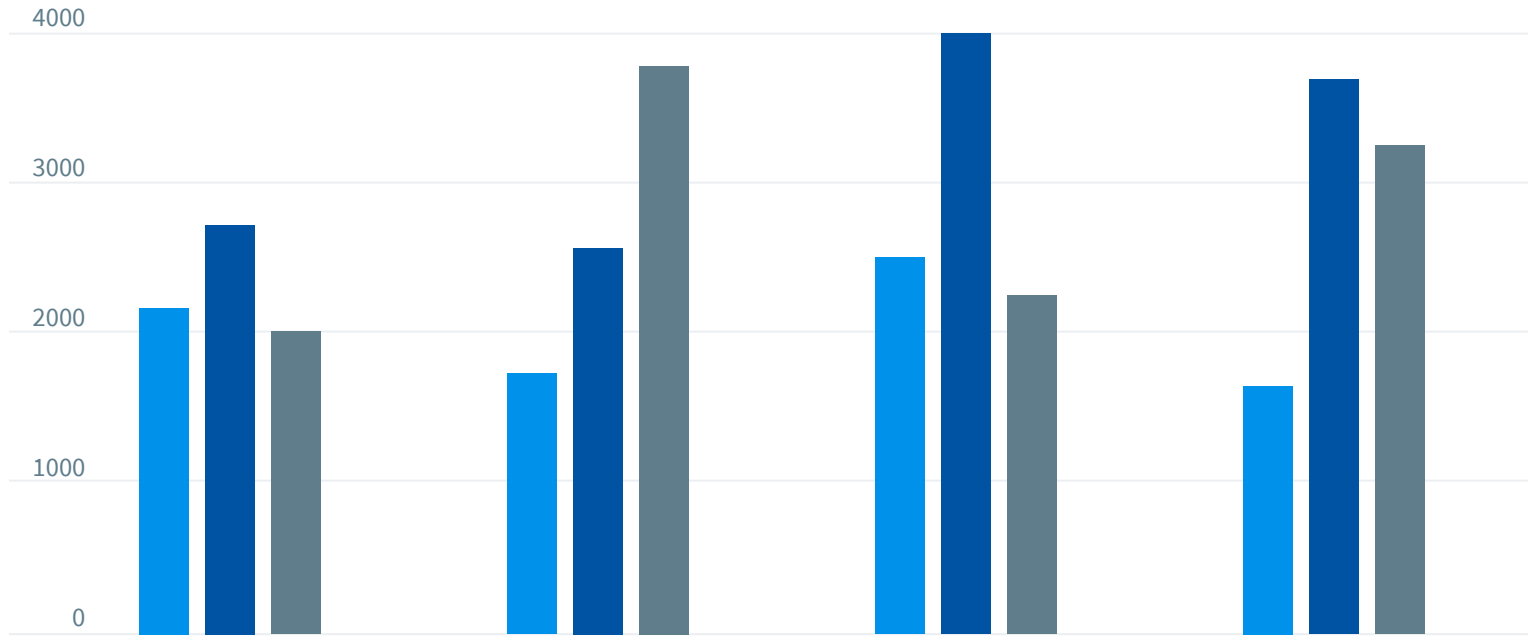
Red

Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.



Red

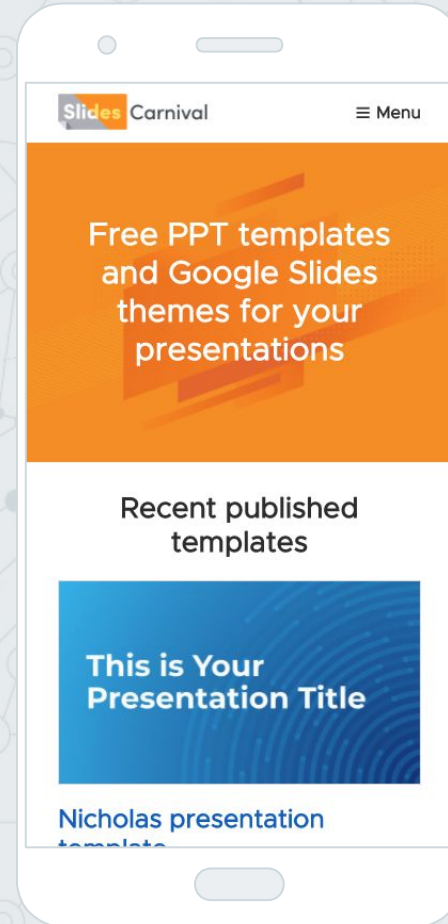
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.



You can insert graphs from Excel or Google Sheets

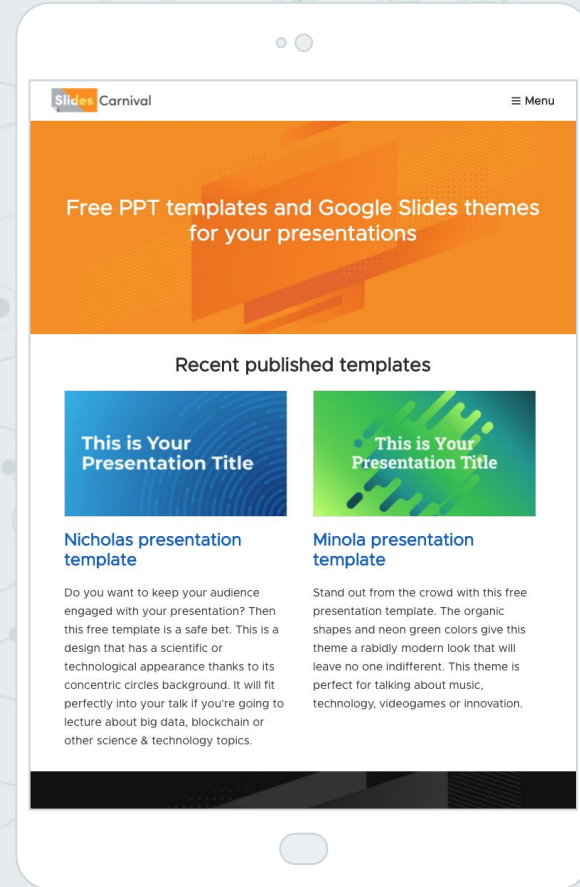
Mobile project

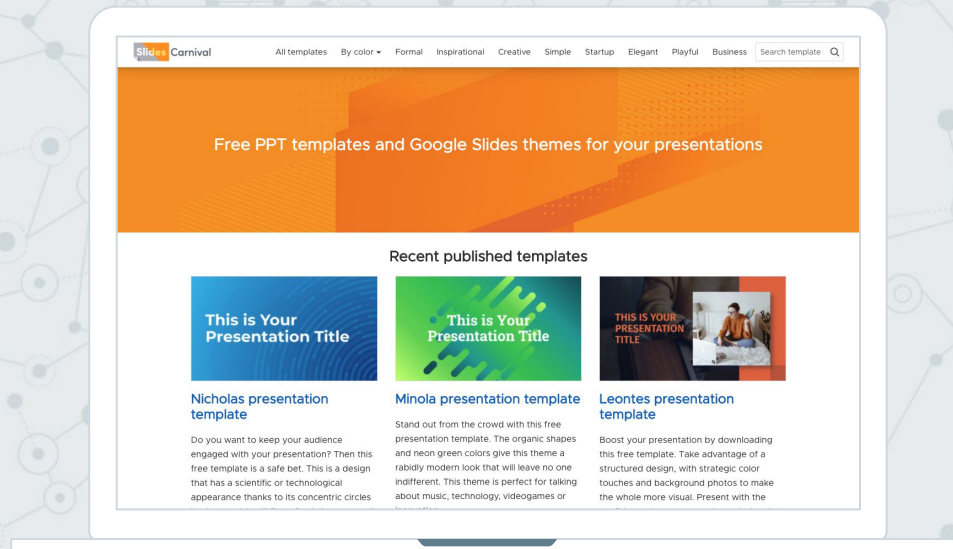
Show and explain your web, app or software projects using these gadget templates.



Tablet project

Show and explain your web, app or software projects using these gadget templates.





Desktop project

Show and explain your web, app or software projects using these gadget templates.



Thanks!

Any questions?

You can find me at:

@username & user@mail.me



Credits

Special thanks to all the people who made and released these awesome resources for free:

- ◎ Presentation template by SlidesCarnival
- ◎ Photographs by Unsplash

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

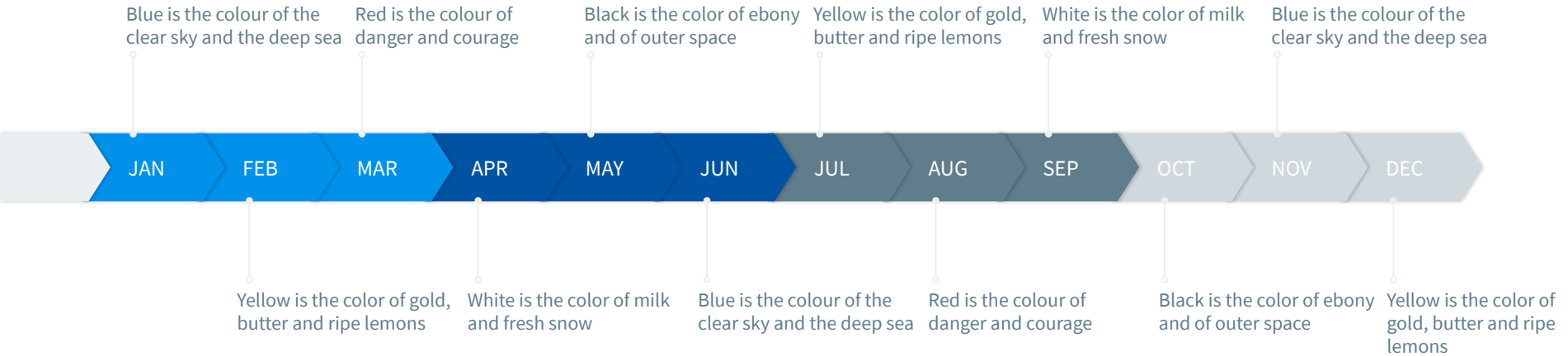
2.

Extra Resources

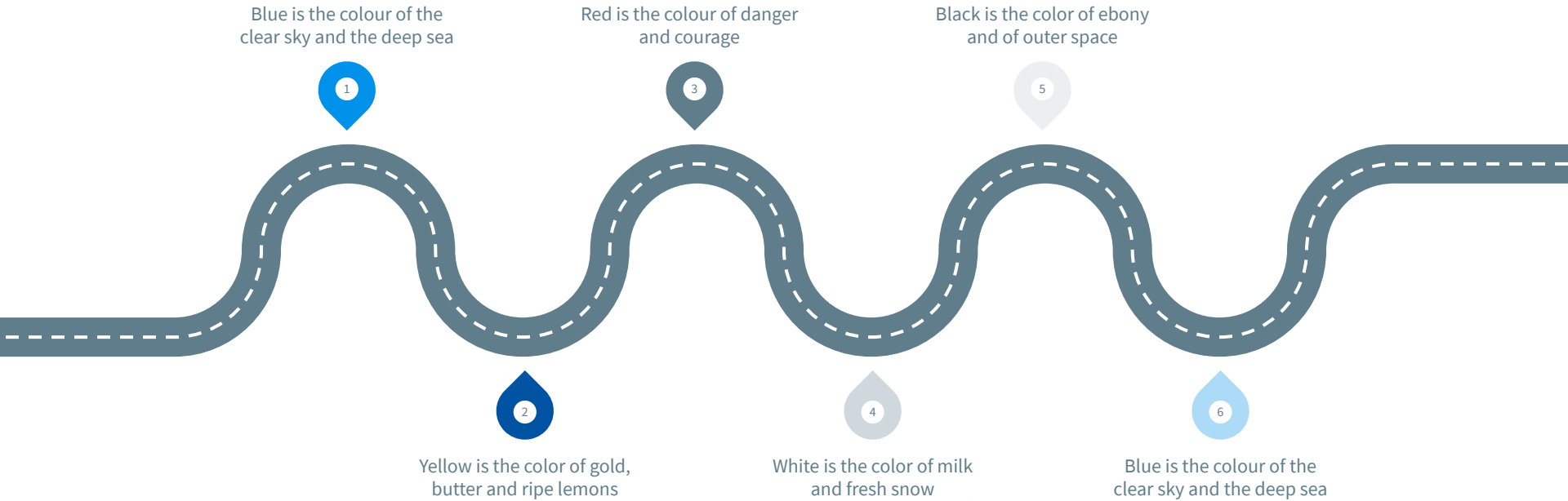
For Business Plans, Marketing Plans,
Project Proposals, Lessons, etc

A decorative network diagram in the bottom-right corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

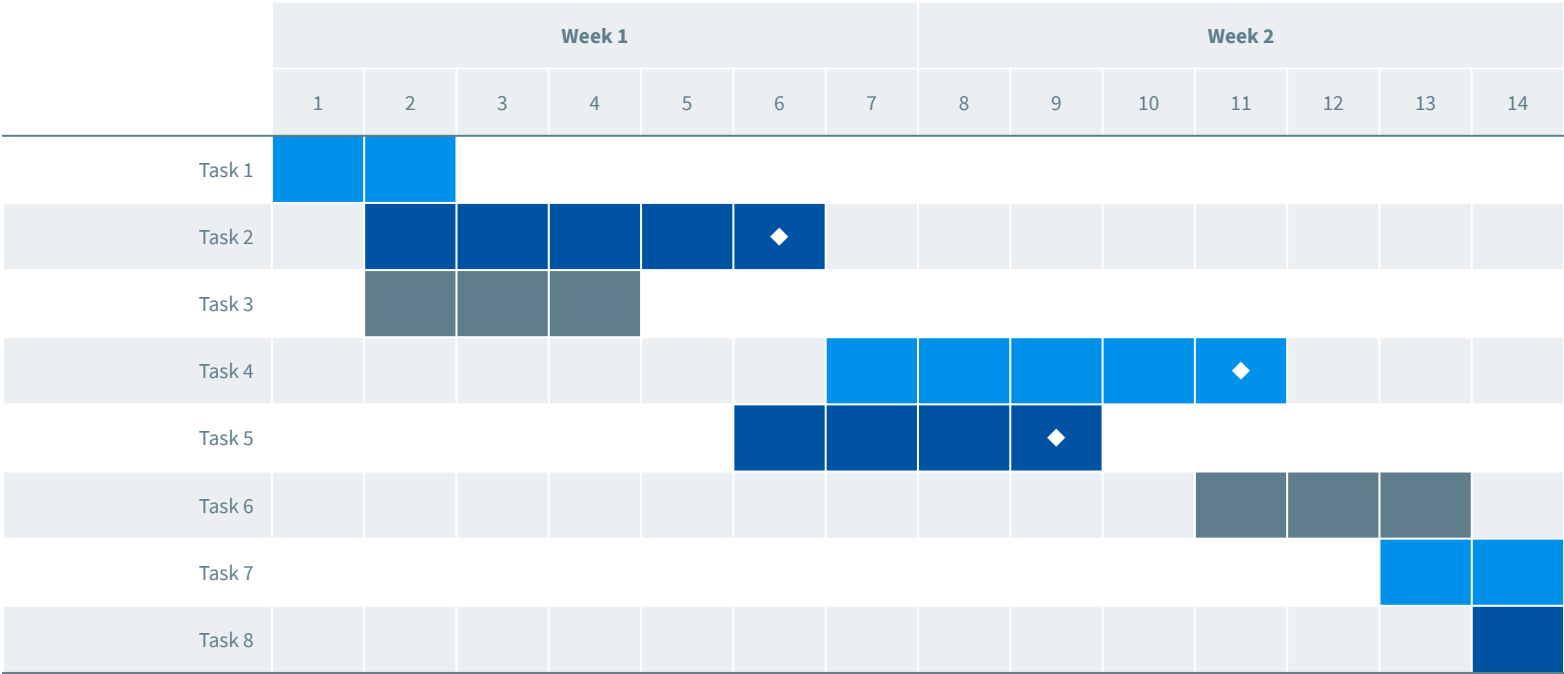
Timeline



Roadmap



Gantt chart



SWOT Analysis

STRENGTHS

Blue is the colour of the clear sky and the deep sea

S

WEAKNESSES

Yellow is the color of gold, butter and ripe lemons

W

O

Black is the color of ebony and of outer space










OPPORTUNITIES

T

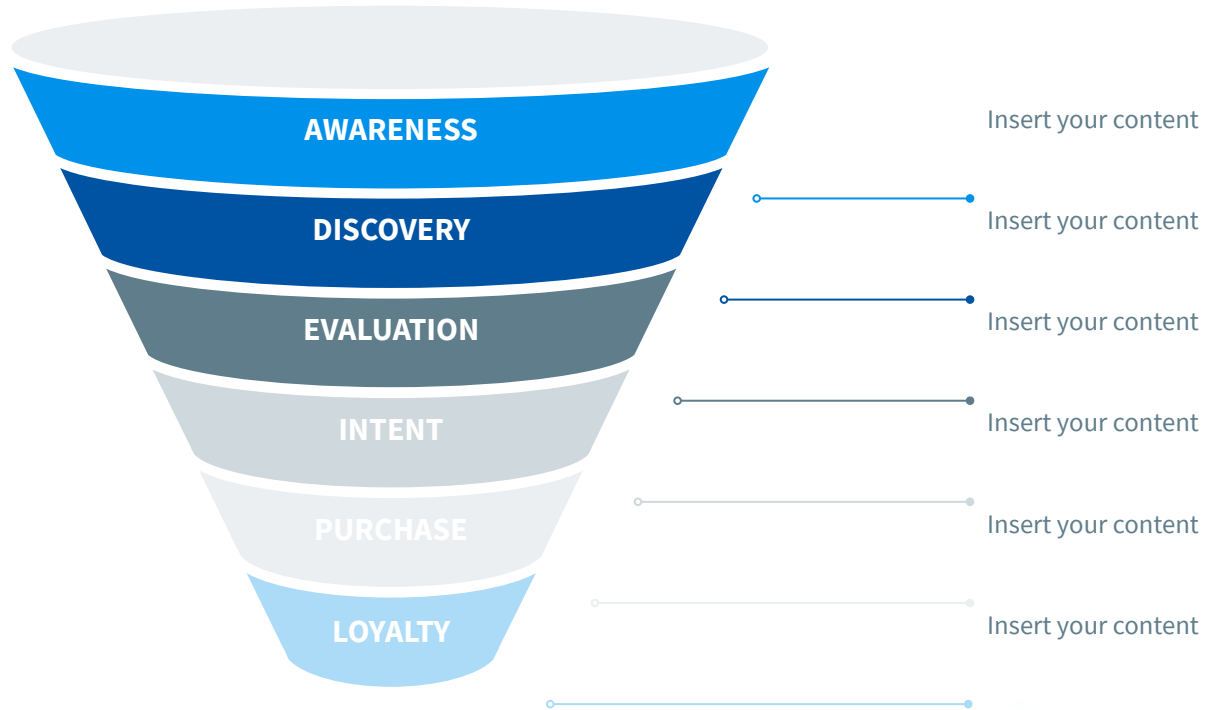
White is the color of milk and fresh snow

THREATS

Business Model Canvas

Key Partners Insert your content 	Key Activities Insert your content 	Value Propositions Insert your content 	Customer Relationships Insert your content 	Customer Segments Insert your content 
	Key Resources Insert your content 		Channels Insert your content 	
Cost Structure Insert your content 			Revenue Streams Insert your content 	

Funnel



Team Presentation



Imani Jackson

JOB TITLE

Blue is the colour of the clear
sky and the deep sea



Marcos Galán

JOB TITLE

Blue is the colour of the clear
sky and the deep sea



Ixchel Valdía

JOB TITLE

Blue is the colour of the clear
sky and the deep sea

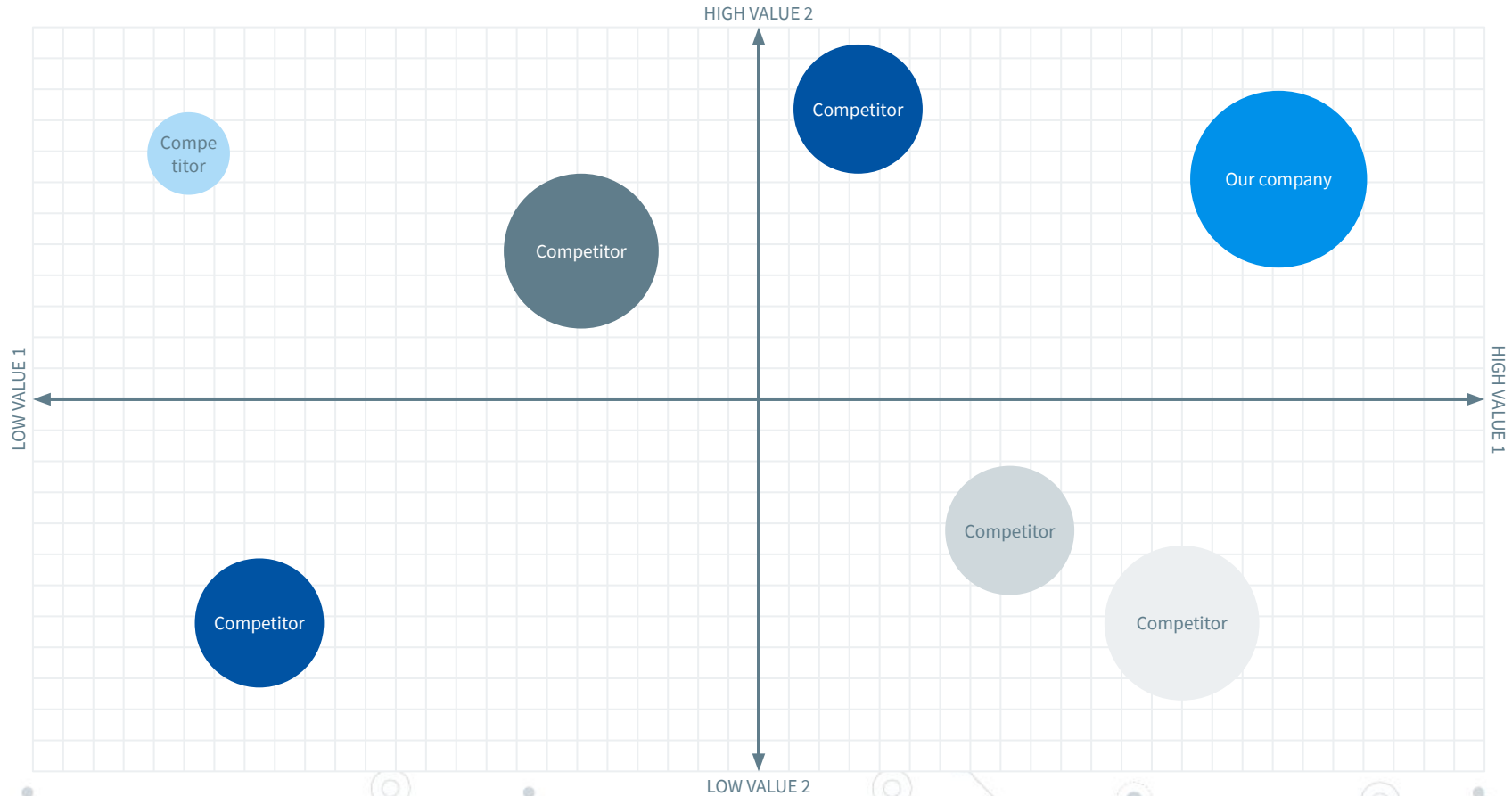


Nils Årud

JOB TITLE

Blue is the colour of the clear
sky and the deep sea

Competitor Matrix



Weekly Planner

	SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
09:00 - 09:45	Task	Task	Task	Task	Task	Task	Task
10:00 - 10:45	Task	Task	Task	Task	Task	Task	Task
11:00 - 11:45	Task	Task	Task	Task	Task	Task	Task
12:00 - 13:15	✓ Free time	✓ Free time	✓ Free time	✓ Free time	✓ Free time	✓ Free time	✓ Free time
13:30 - 14:15	Task	Task	Task	Task	Task	Task	Task
14:30 - 15:15	Task	Task	Task	Task	Task	Task	Task
15:30 - 16:15	Task	Task	Task	Task	Task	Task	Task



SlidesCarnival icons are editable shapes.

This means that you can:

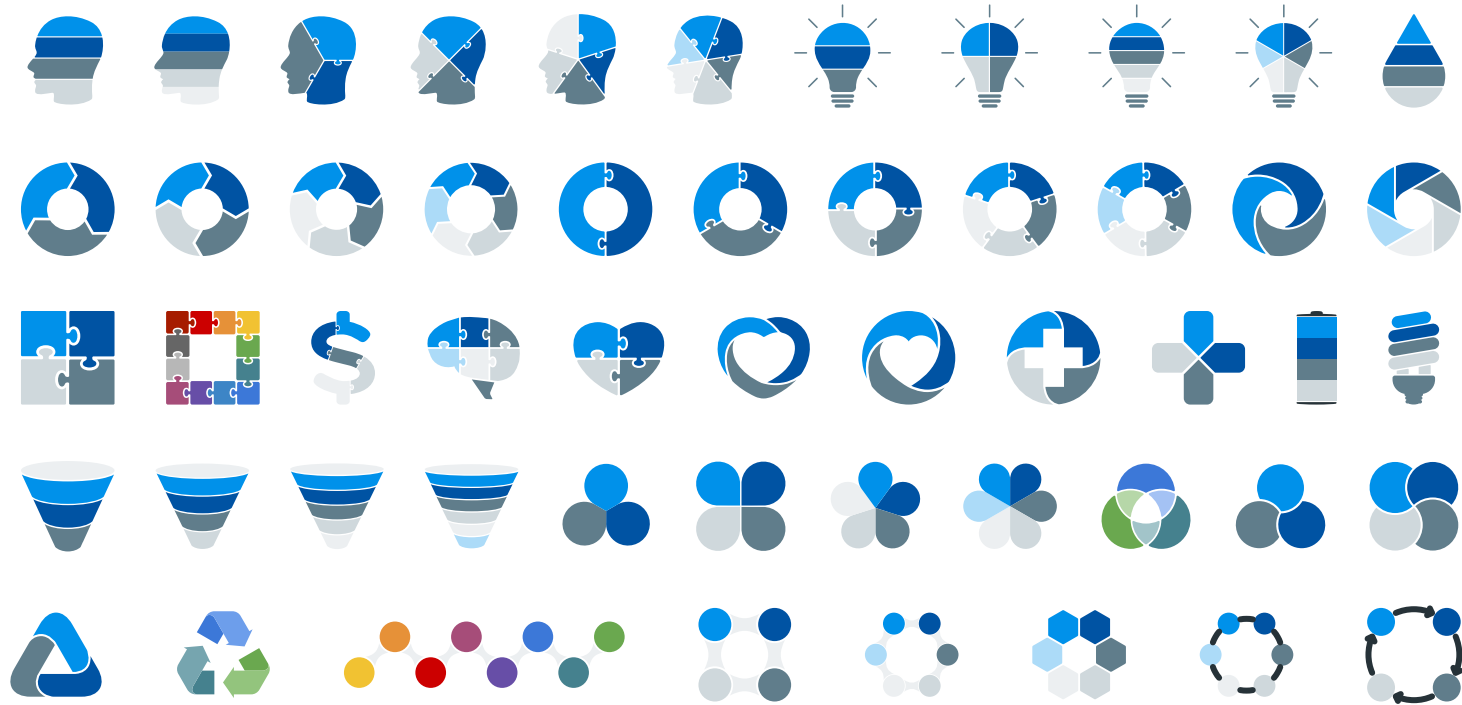
- Resize them without losing quality.
- Change line color, width and style.

Isn't that nice? :)

Examples:



Diagrams and infographics



You can also use any emoji as an icon!

And of course it resizes without losing quality.

How? Follow Google instructions <https://twitter.com/googledocs/status/730087240156643328>



and many more...



Free templates for all your presentation needs



For PowerPoint and
Google Slides



100% free for personal
or commercial use



Ready to use,
professional and
customizable



Blow your audience
away with attractive
visuals