

Data Preprocessing

A decorative graphic at the bottom of the slide consists of two curved bands of small blue dots. One band arches upwards from the bottom left, and the other arches downwards from the top right, meeting in the center.

DATA
SCIENCE
INDONESIA

Raymond Christopher Sitorus

Chief Data Officer of **delman.io**

I love data and I love to share about
what I know about data



2007



2012



2015



2017



2018

delman.io

What & When

Apa itu *data preprocessing* dan kapan proses tersebut dilakukan?

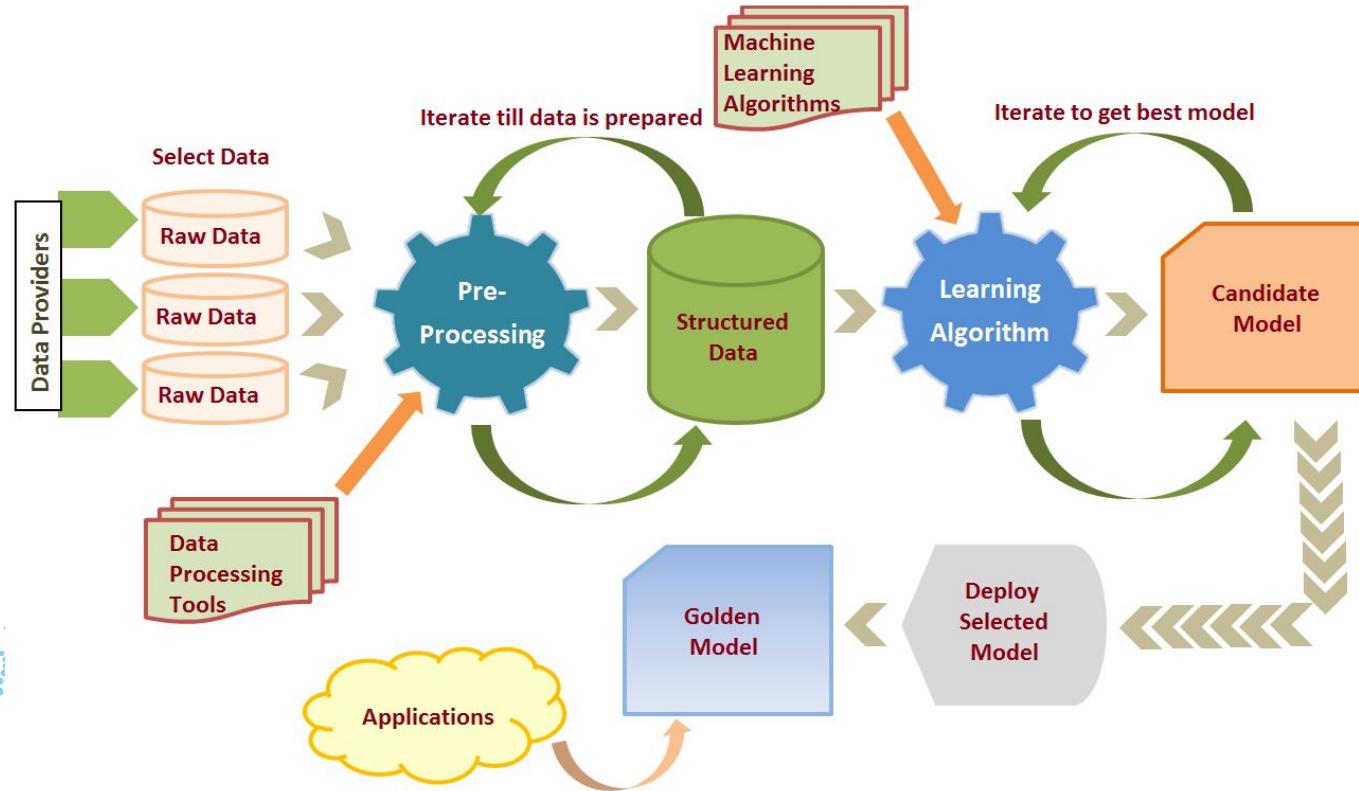
“

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

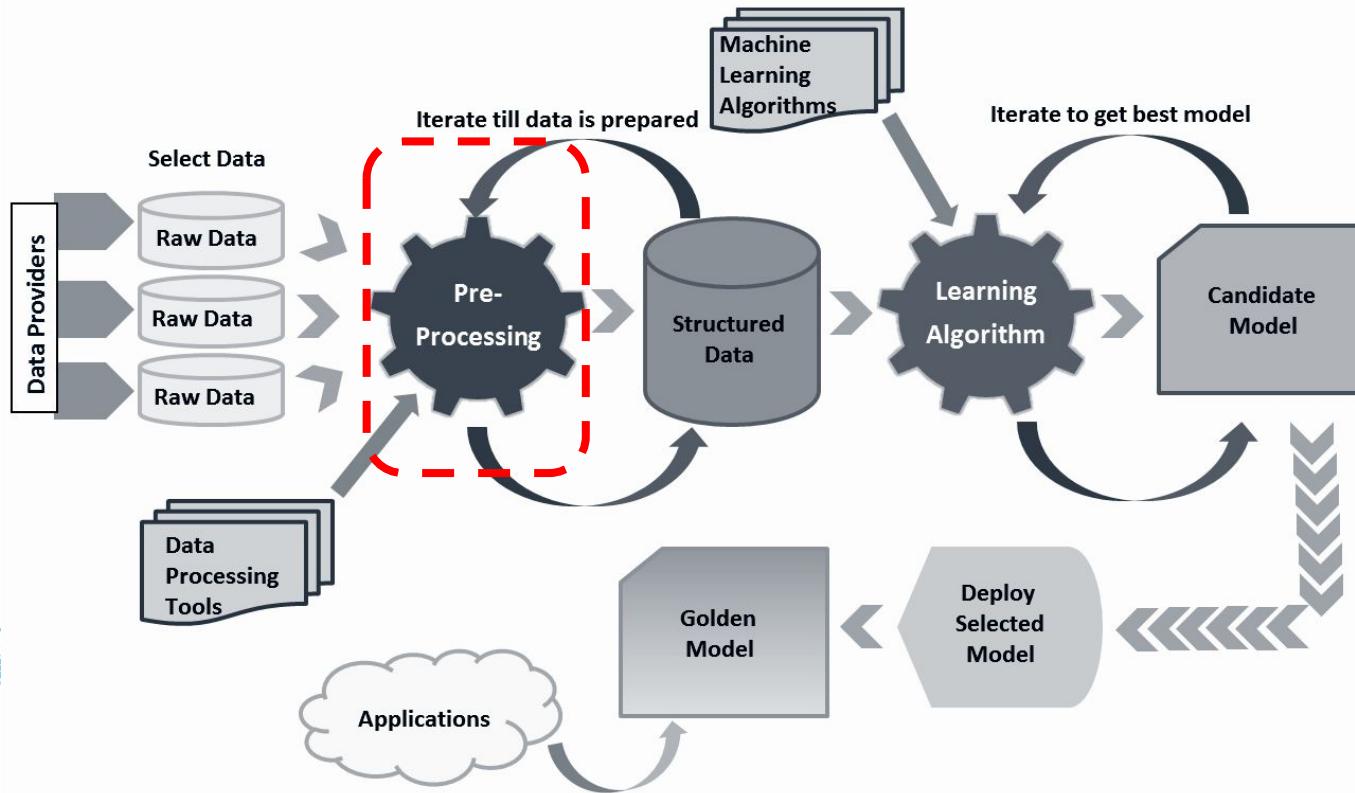
Apa itu data preprocessing?

- Sering disebut juga **data preparation** atau **data cleaning**
- *pre* = sebelum → kegiatan yang dilakukan sebelum suatu data diproses lebih lanjut

Kapan preprocessing dilakukan?



Kapan preprocessing dilakukan?



Why

Mengapa kita perlu melakukan *data preprocessing*?



DIMITRI OTIS IMAGES

Alasan #1 dari data preprocessing

Your data can worth much more
after being **cleaned!**

The truth about real life data

What you think about your data



What your data actually looks like



Apa masalah dengan raw data ini?

color	director_name	duration	gross	movie_title	language	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall Street	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful Eight	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	Margaret	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of Smaug	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of Justice	English	USA	250000000	2016	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected Journey	English	USA	180000000	2012	7.9
Color	Edward Hall	180		Restless	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud Atlas	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon Tattoo	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of God	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Desolation of Smaug	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6

Sample data dari IMDB

“Garbage in, garbage out”

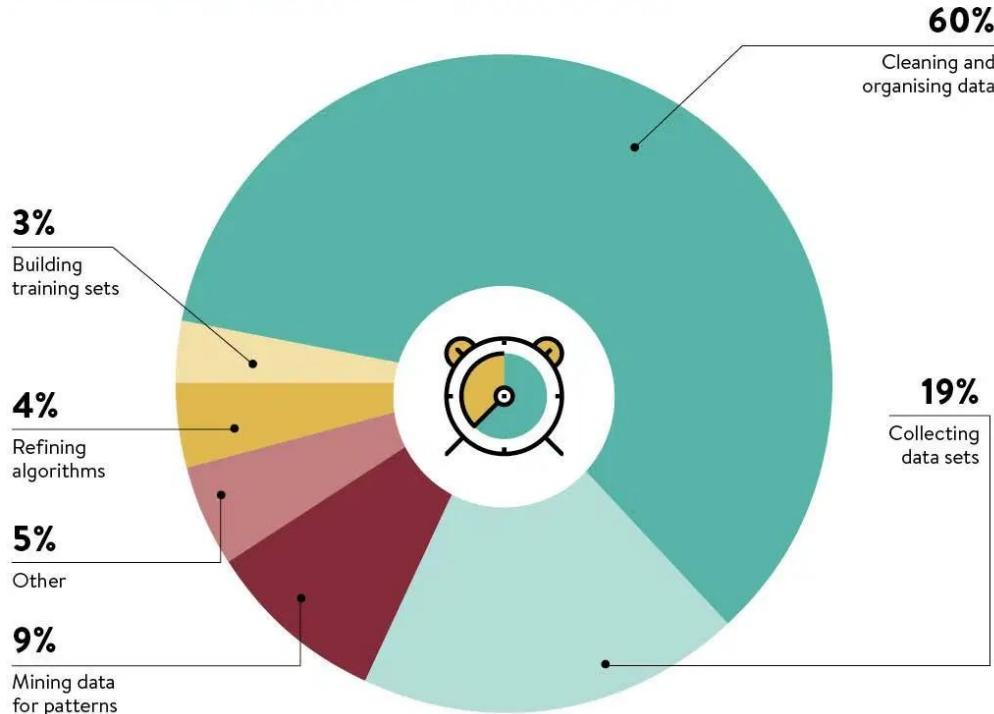


Problems

Apakah masalah yang kita mungkin temukan ketika melakukan *data preprocessing* ?

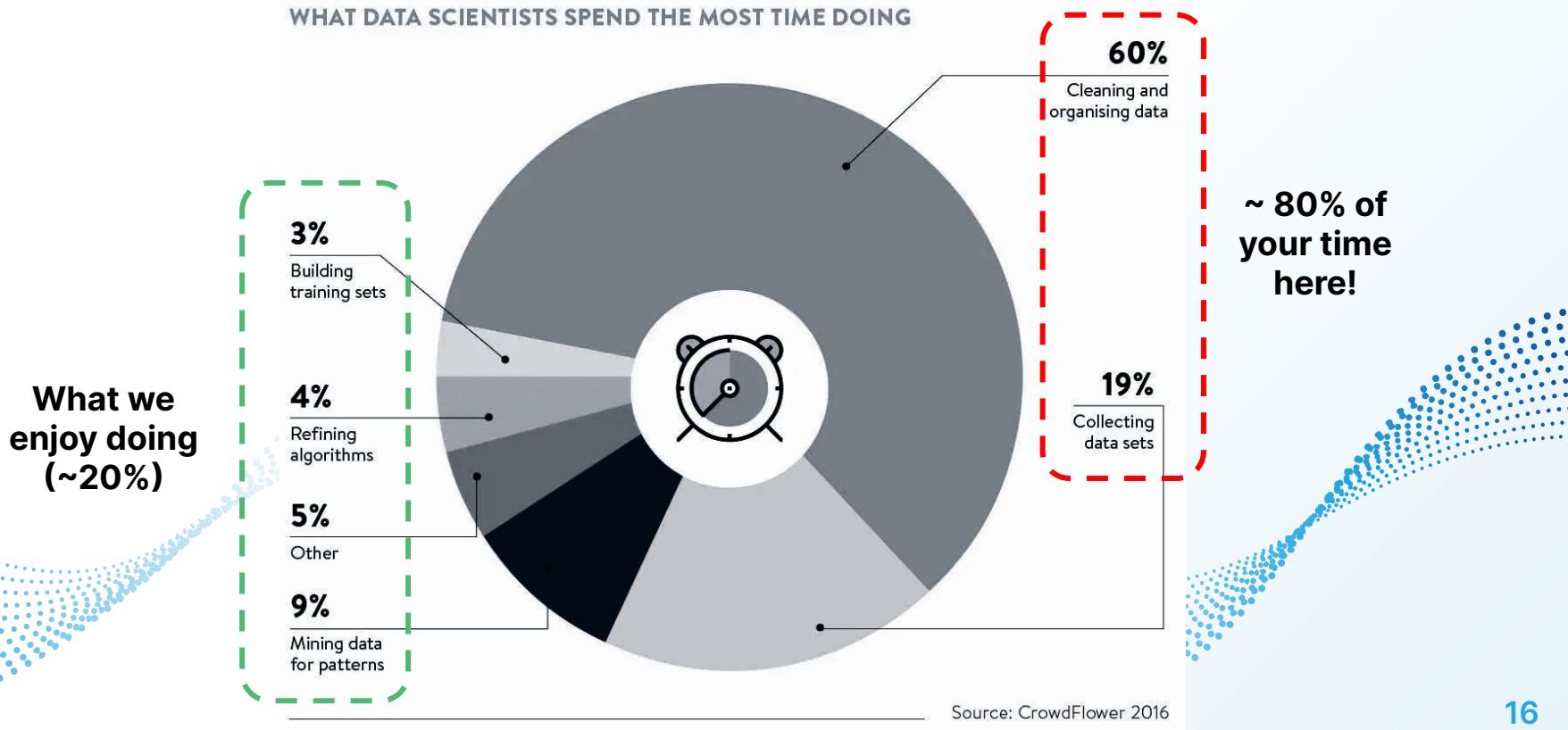
Data cleaning memakan waktu

WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



Source: CrowdFlower 2016

Data cleaning memakan waktu



Jenis masalah yang bisa ditemukan

color	director_name	duration	gross	movie_title	language	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall Street	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful Eight	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	Margaret	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of Smaug	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of Justice	English	USA	250000000	2016	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected Journey	English	USA	180000000	2012	7.9
Color	Edward Hall	180		Restless	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud Atlas	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon Tattoo	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of God	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Desolation of Smaug	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6

Sample data dari IMDB

Jenis masalah yang bisa ditemukan

- Input yang salah atau *typos*
- *Duplicate/redundant* data
- *Missing* data
- *Outliers* atau anomali
- Data format yang beragam
- Skala data yang beragam

Sangat peka dengan konteks data



How

Apa saja *steps* yang biasa dilakukan dalam konteks *data preprocessing*?

Tools & libraries for demo

colab

pandas

NumPy

PANDAS
PROFILING

scikit
learn

Demo Time

You can access the Google Collab on the following link

bit.ly/34X6JEc

Resources

Hal-hal apa yang dapat membantu saya melakukan *data preprocessing* dengan lebih baik?

Learning resources



Cassie Kozyrkov

Following ▾

Head of Decision Intelligence, Google. ❤️ Stats, ML/AI, data, puns, art, theatre, decision science. All views are my own. twitter.com/quaesita

Medium member since September 2019 · Top writer in Technology

35K Followers ·



towardsdatascience.com

The screenshot shows the homepage of towardsdatascience.com. The header features the "towards data science" logo and a subtitle "Sharing concepts, ideas, and codes". Below the header is a navigation bar with links for DATA SCIENCE, MACHINE LEARNING, PROGRAMMING, VISUALIZATION, VIDEO, ABOUT, CONTRIBUTE, and social media icons. Two article thumbnails are visible: one for "Probabilistic Reasoning on Knowledge Graphs" featuring a photo of a smiling person in a yellow hat, and another for "An Overview of Model Compression Techniques for Deep Learning in Space" featuring a photo of Earth from space with a satellite.

Learning resources

coursera



Browse > Data Science > Data Analysis

Offered By JOHNS HOPKINS UNIVERSITY

Getting and Cleaning Data

★★★★★ 4.6 7,361 ratings | 90% complete

Jeff Leek, PhD +2 more instructors

www.coursera.org/specializations/jhu-data-science

Browse > Data Science > Data Analysis

Offered By UNIVERSITY OF MICHIGAN

Applied Data Science with Python Specialization

Gain new insights into your data. Learn to apply data science methods and techniques, and acquire analysis skills.

★★★★★ 4.5 38,042 ratings

Christopher Brooks +3 more instructors

kaggle™

www.kaggle.com



≡

THE SCHOOL OF

Data Science

Build expertise in data manipulation, visualization, predictive analytics, machine learning, and data science. With the skills you learn in a Nanodegree program, you can launch or advance a successful data career. Start acquiring valuable skills right away, create a project portfolio to demonstrate your abilities, and get support from mentors, peers, and experts in the field. We offer five unique programs to support your career goals in the data science field.

www.udacity.com/school-of-data-science

Data preprocessing tools



TRIFACTA
cloud.trifacta.com

colab

colab.research.google.com

alteryx

www.alteryx.com/designer-trial/free-trial

Thank you!

Any questions?

You can find me at **raymond@delman.io**

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.

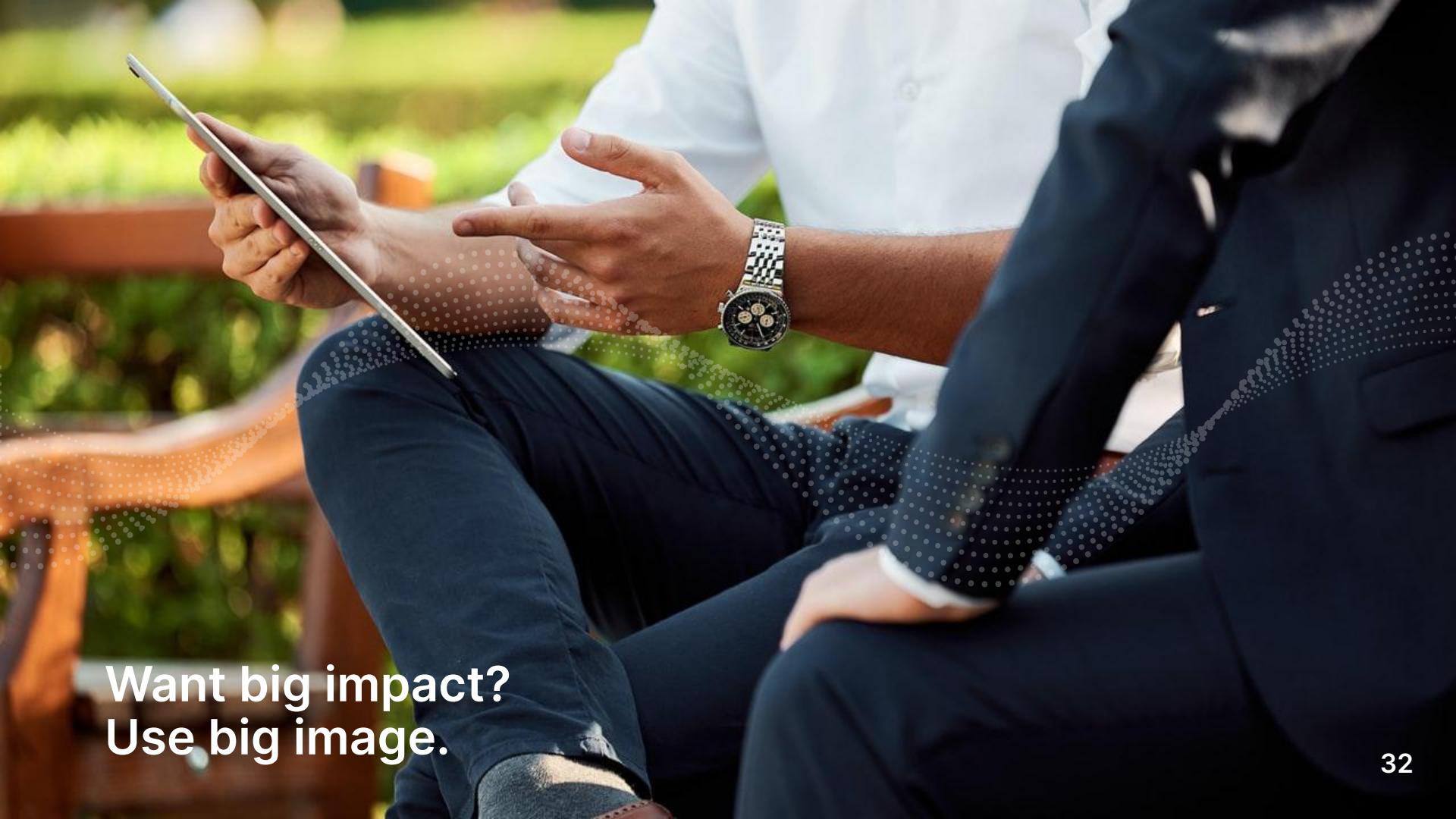
Red

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

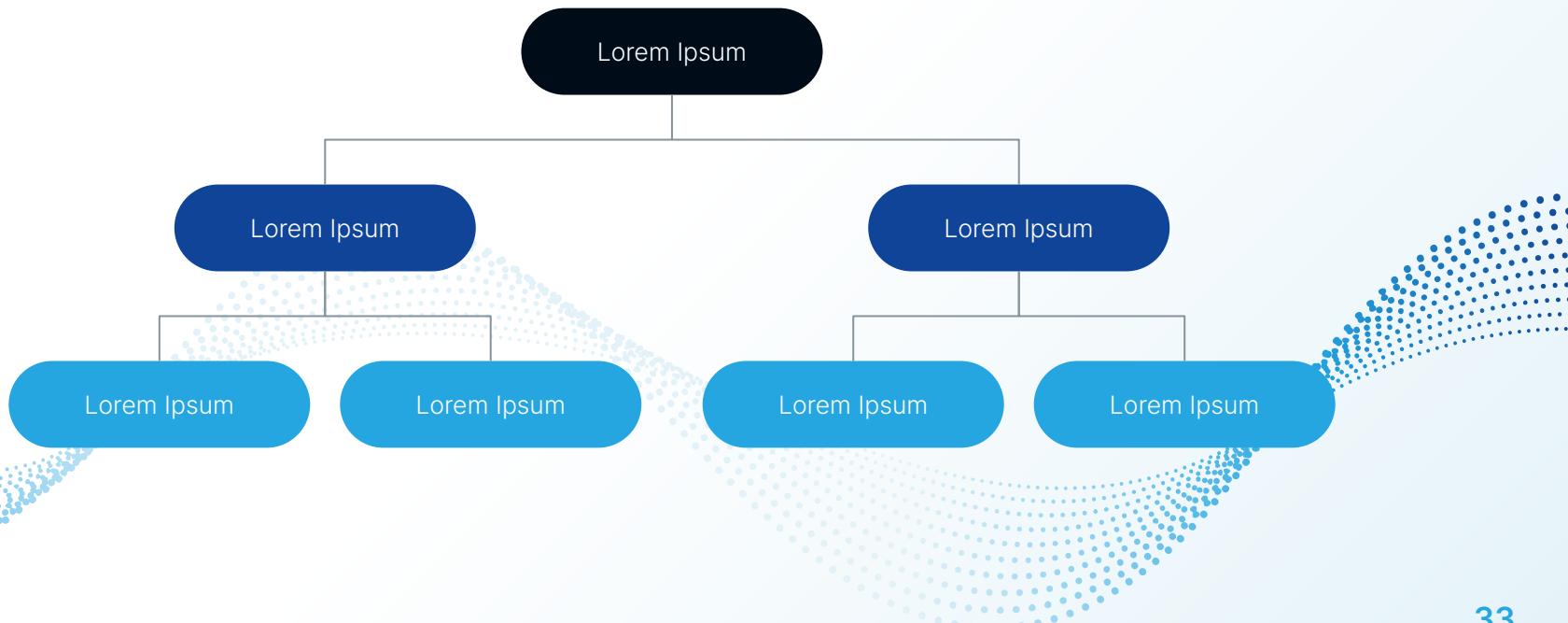
A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.





Want big impact?
Use big image.

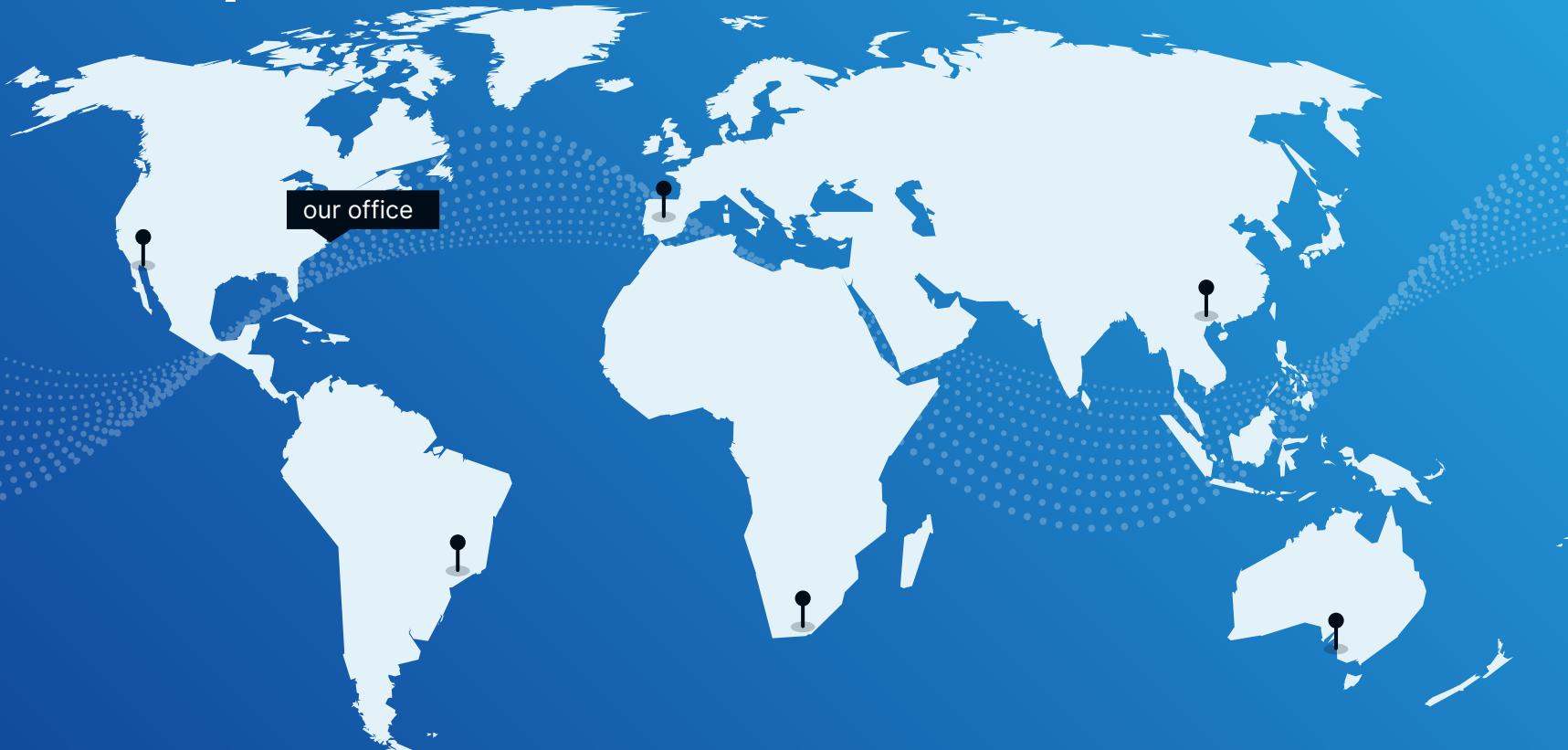
Use diagrams to explain your ideas



And tables to compare data

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

Maps



89,526,124

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

89,526,124\$

That's a lot of money

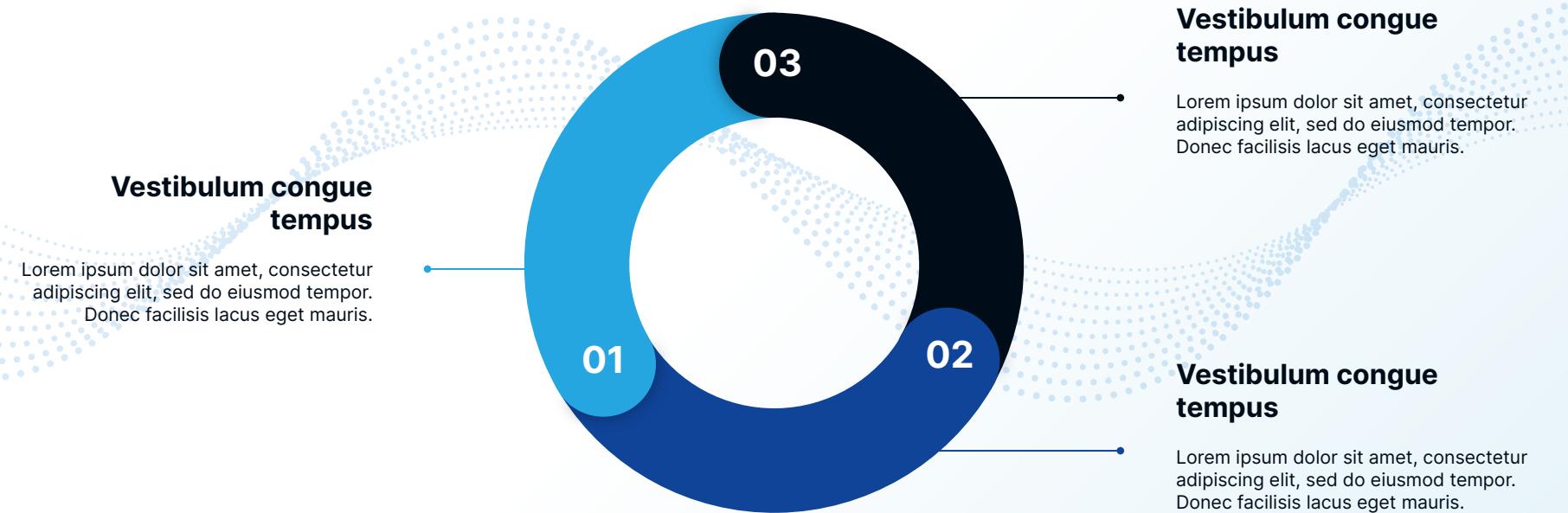
185,244 users

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

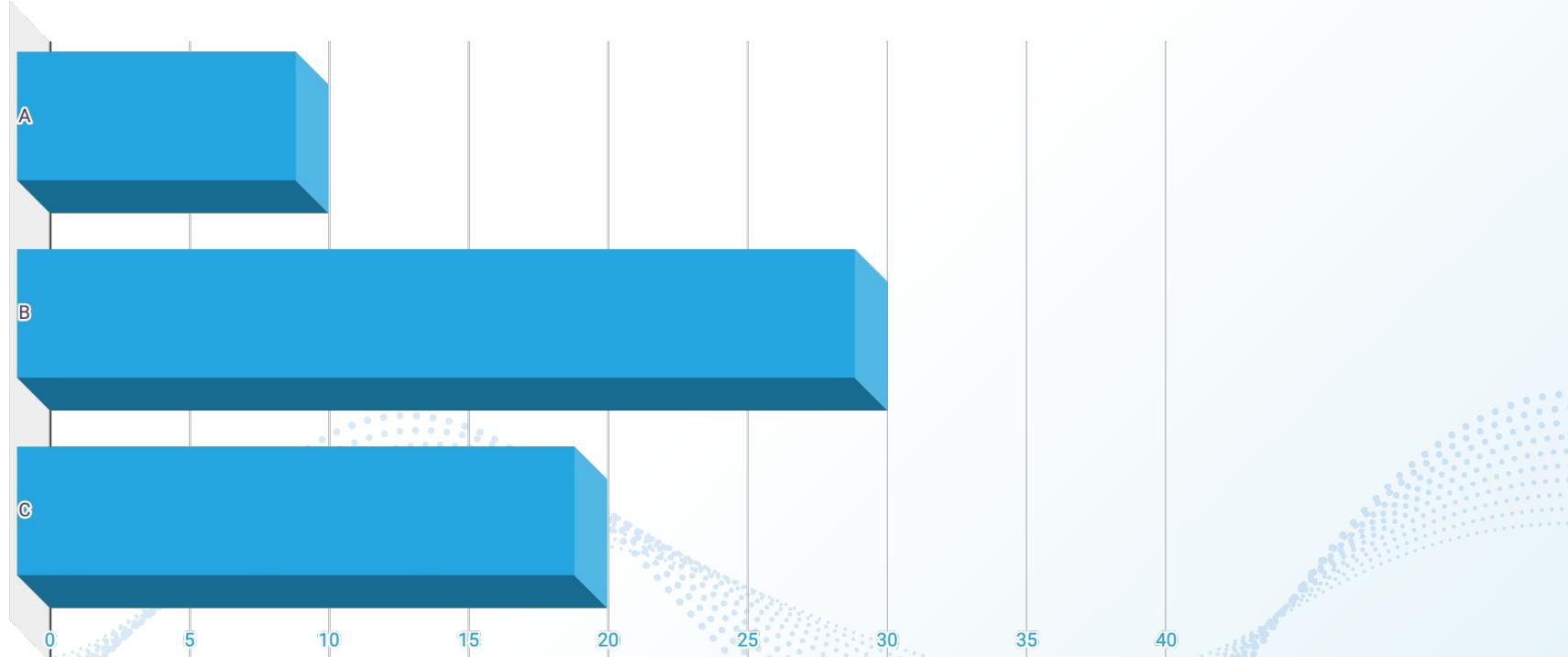
Is the colour of the clear sky and the deep sea. It is located between violet and green on the optical spectrum.

Red

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

Red

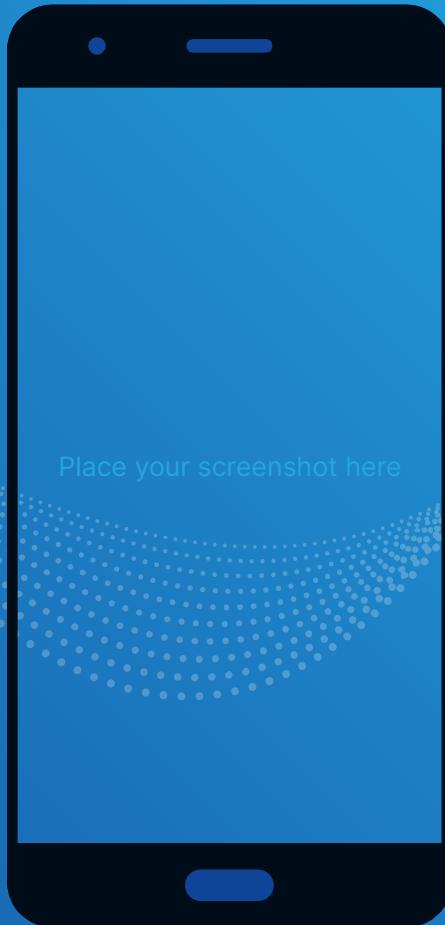
Is the color of blood, and because of this it has historically been associated with sacrifice, danger and courage.



You can insert graphs from Excel or Google Sheets

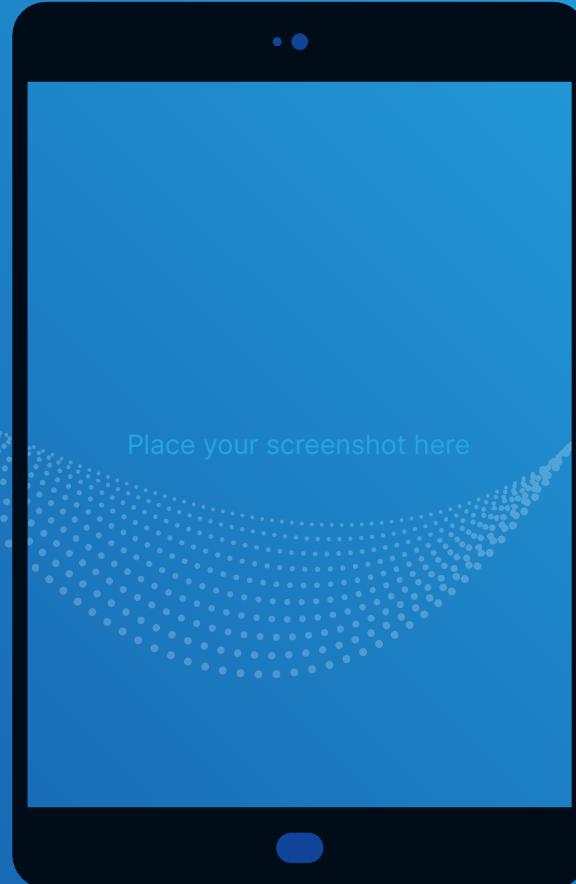
Mobile project

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



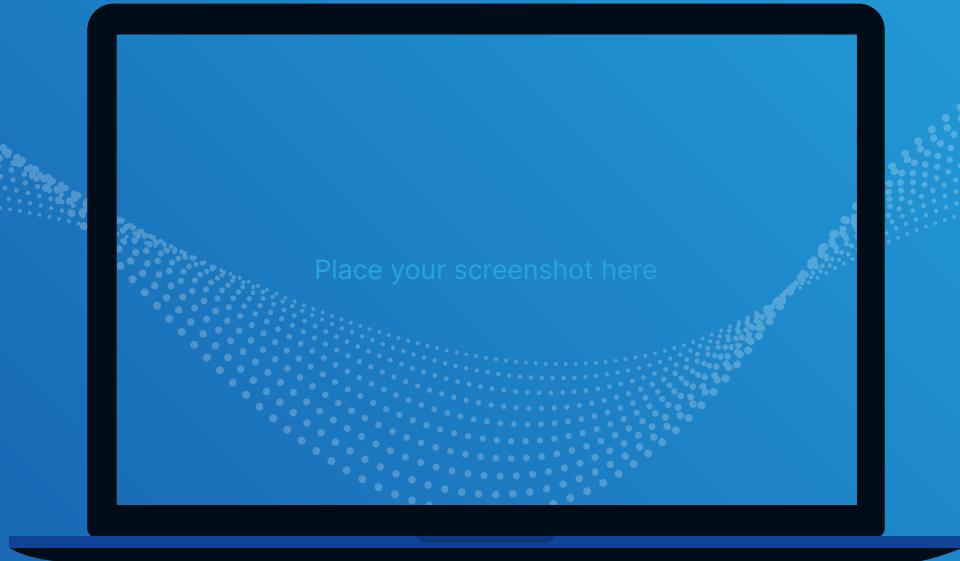
Tablet project

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



Presentation design

This presentation uses the following typographies:

- Titles: Inter Semibold
- Body copy: Inter Light

Download for free at:

<https://www.fontsquirrel.com/fonts/inter>

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®



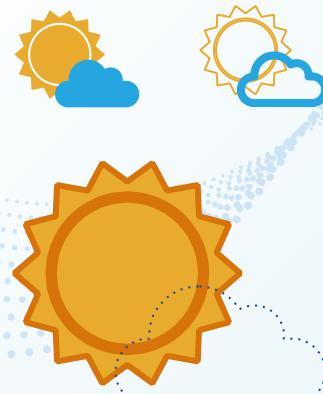
SlidesCarnival icons are editable shapes.

This means that you can:

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

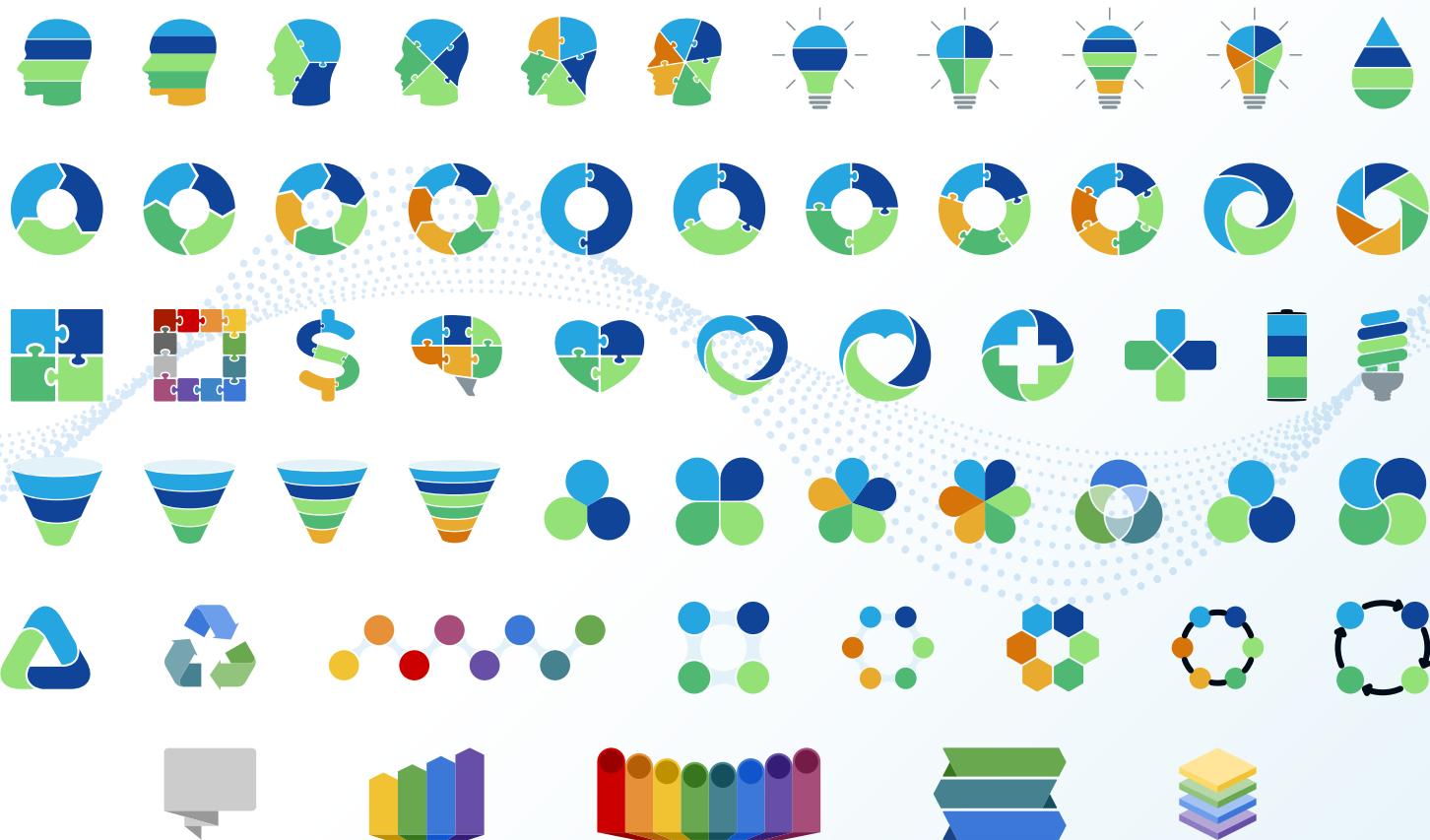
Isn't that nice? :)

Examples:



Find more icons at
slidescarnival.com/extr-free-resources-icons-and-maps

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>



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