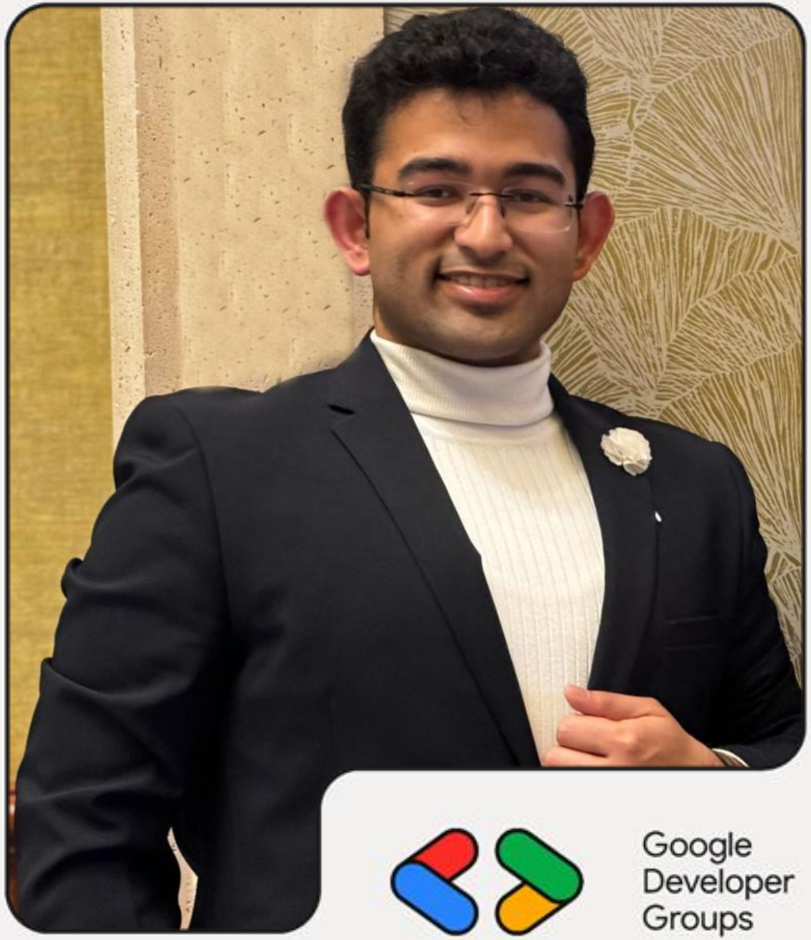
Google Nvidia Meetup

Dubai, UAE

# Model Distillation Learning

Rohan Mitra

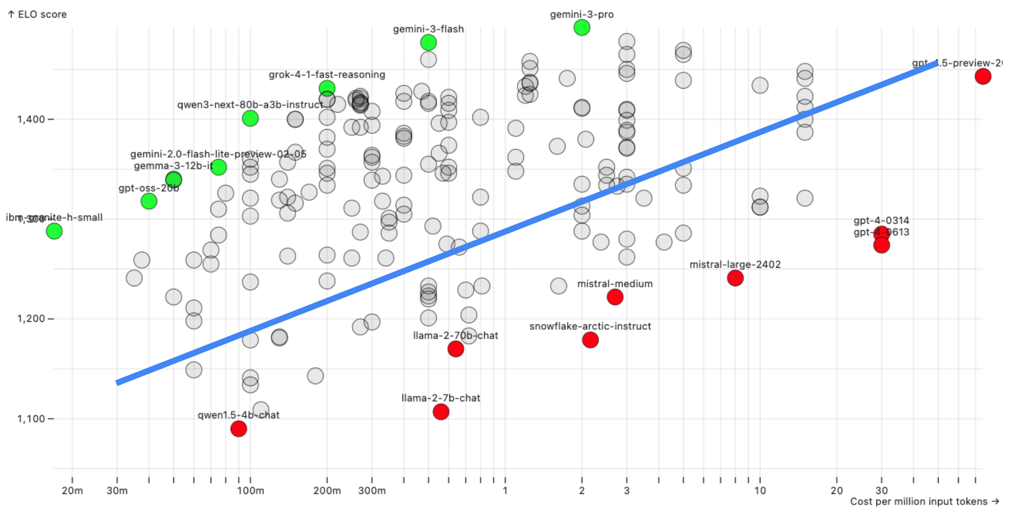
ML Engineer @ Bayut | Dubizzle
Research Assistant for ML @ CUD
Organizer & Speaker GDG Sharjah



Google Developer Groups

# The Inference Trade Offs

- 70B models are smart but too expensive for high-volume tasks
- We need the intelligence of Gemini 3 Pro at the cost of a 10 cent CPU



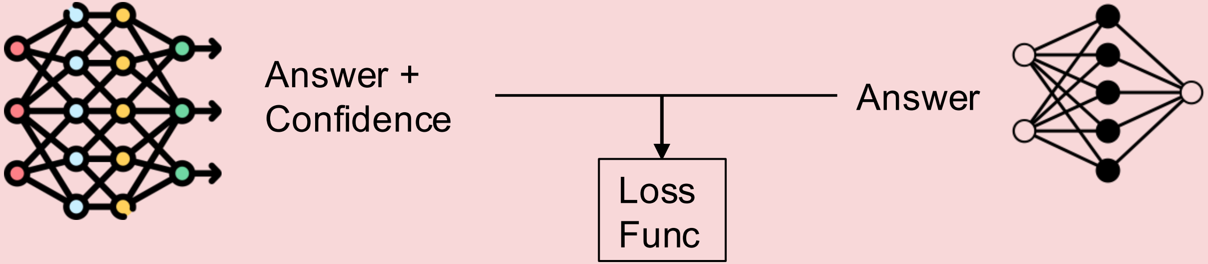
https://sanand0.github.io/llmpricing/

# Distillation vs Fine Tuning

- **Fine-tuning:** Learning new information from raw data
  - Based on explicit “correct answers”
- **Distillation:** Learning the behavior / reasoning process from a smarter teacher
  - Don’t look only at the final result, look at the behavior to get there
- 3 types of distillation:
  - **White Box (Soft Labels):** Full access to Teacher’s brain
  - **Black Box (Hard Labels):** Teacher is an API
  - **Synthetic Data:** Teacher generates training set

# Black Box Distillation

- The teacher’s final output is the ”Gold Standard”
- We don’t have probability vector, but have the final answer and its overall confidence
- Cant learn to behave like teacher, but can learn what the teacher was confident about
  - Eg: Weigh each row in the dataset by teacher’s confidence
- **Condition:** Teacher is closed-source, but we can get the final answer and maybe its overall confidence



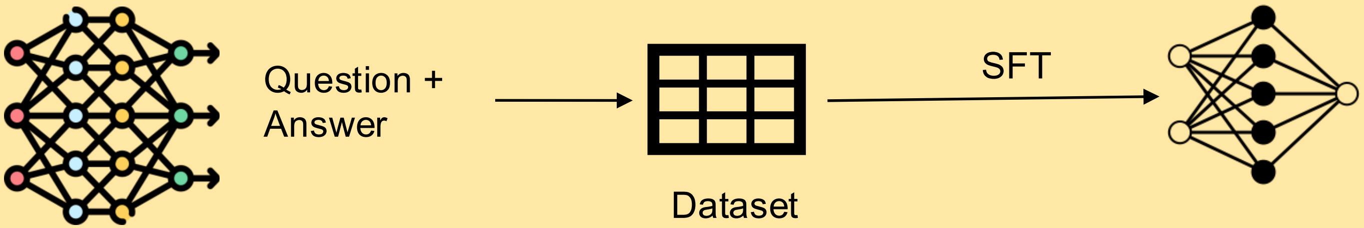
Answer + Confidence

Answer

Loss Func

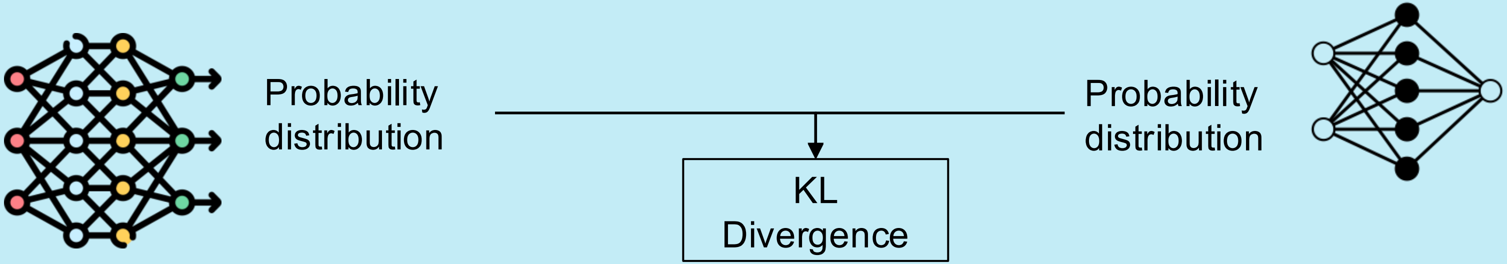
# Synthetic Data

- Ask Gemini 3 Pro to generate diverse inputs, its chain of reasoning, and the final answer
- Can treat it as supervised fine-tuning problem once synthetic data is generated
- Teach the student how to think by mimicking the teacher's reasoning
- **Condition:** Teacher is closed-source and/or we don't have enough data



# White Box Distillation

- Use “Dark Knowledge”
  - Train not just to predict the same final class/token as the teacher, but match the probability distribution of the teacher
  - Ground truth tells you it's a Dog. The Teacher tells you “It's a Dog, but it looks 5% like a Wolf”
- **Condition:** Teacher & Student share the tokenizer

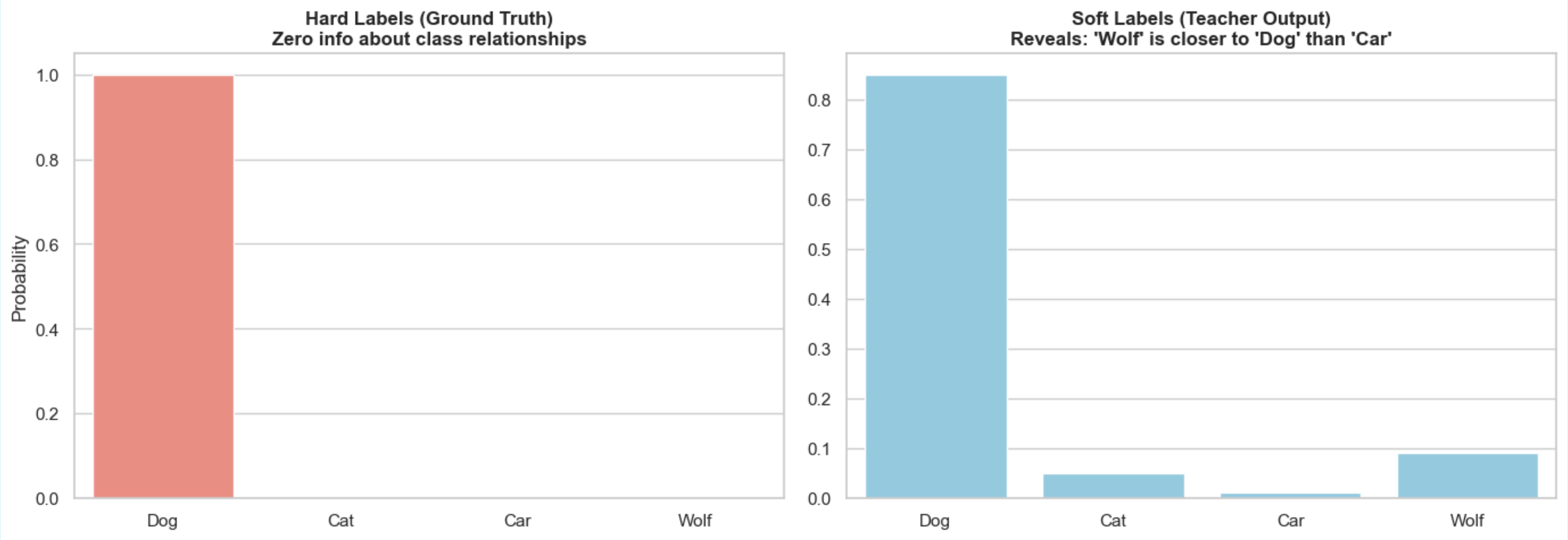


Probability distribution

Probability distribution

KL Divergence

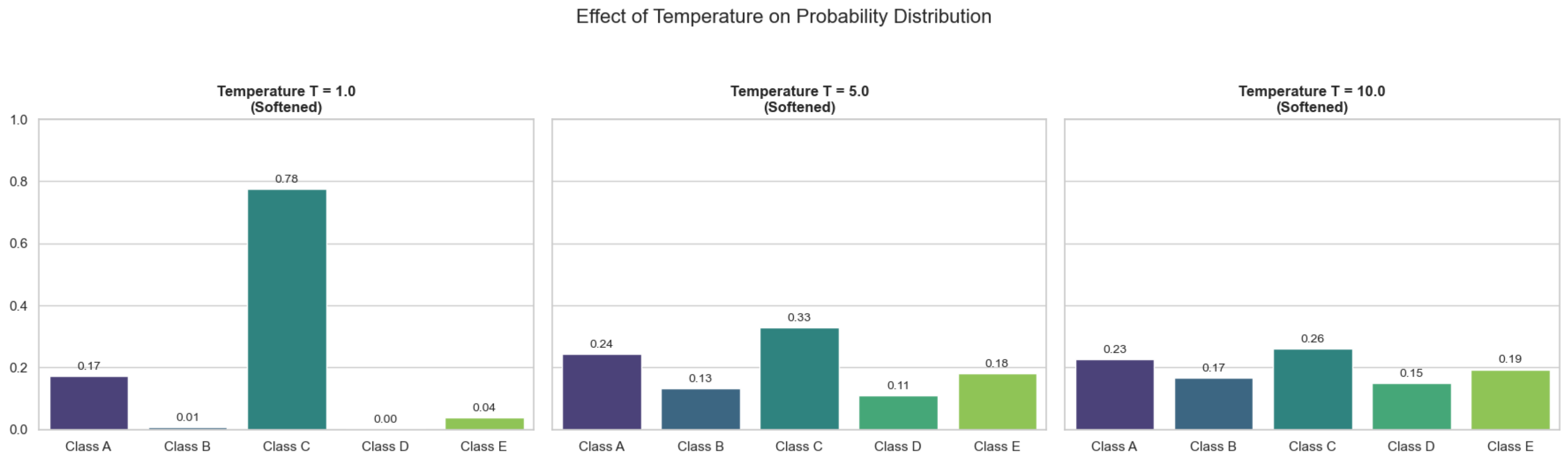
# White Box Distillation



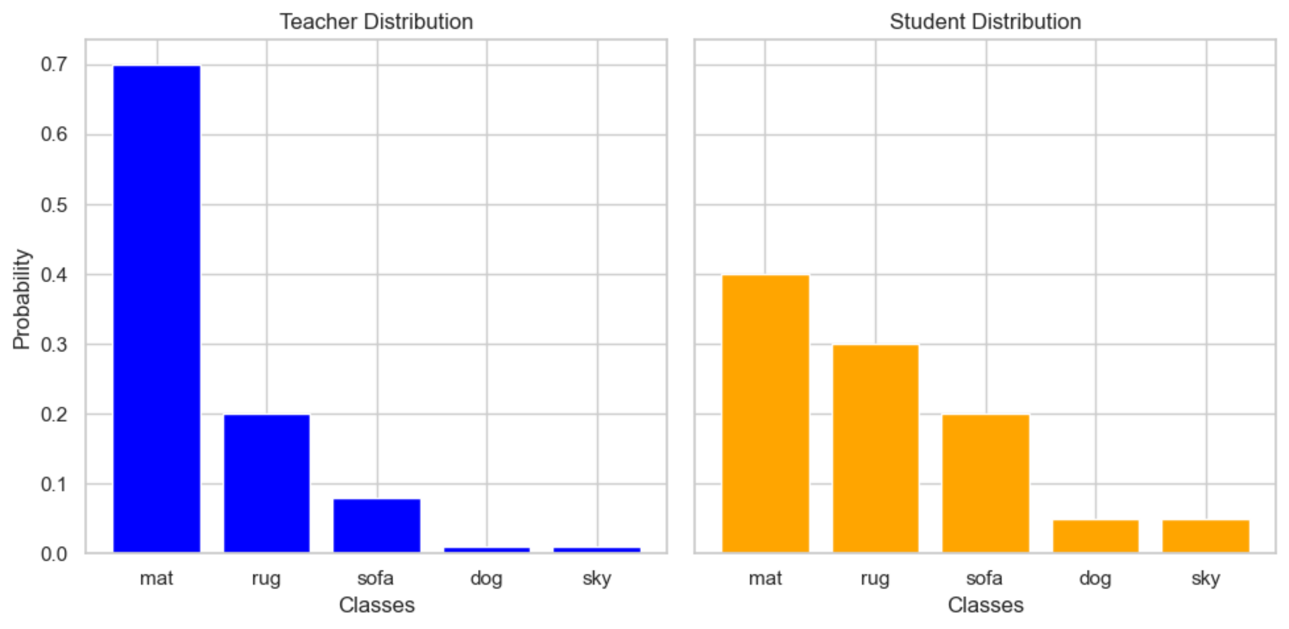
# Temperature

- Need to amplify the effect of the Dark Knowledge for the student to learn

$$Softmax(x_i) = \frac{e^{\frac{x_i}{T}}}{\sum e^{\frac{x_j}{T}}}$$

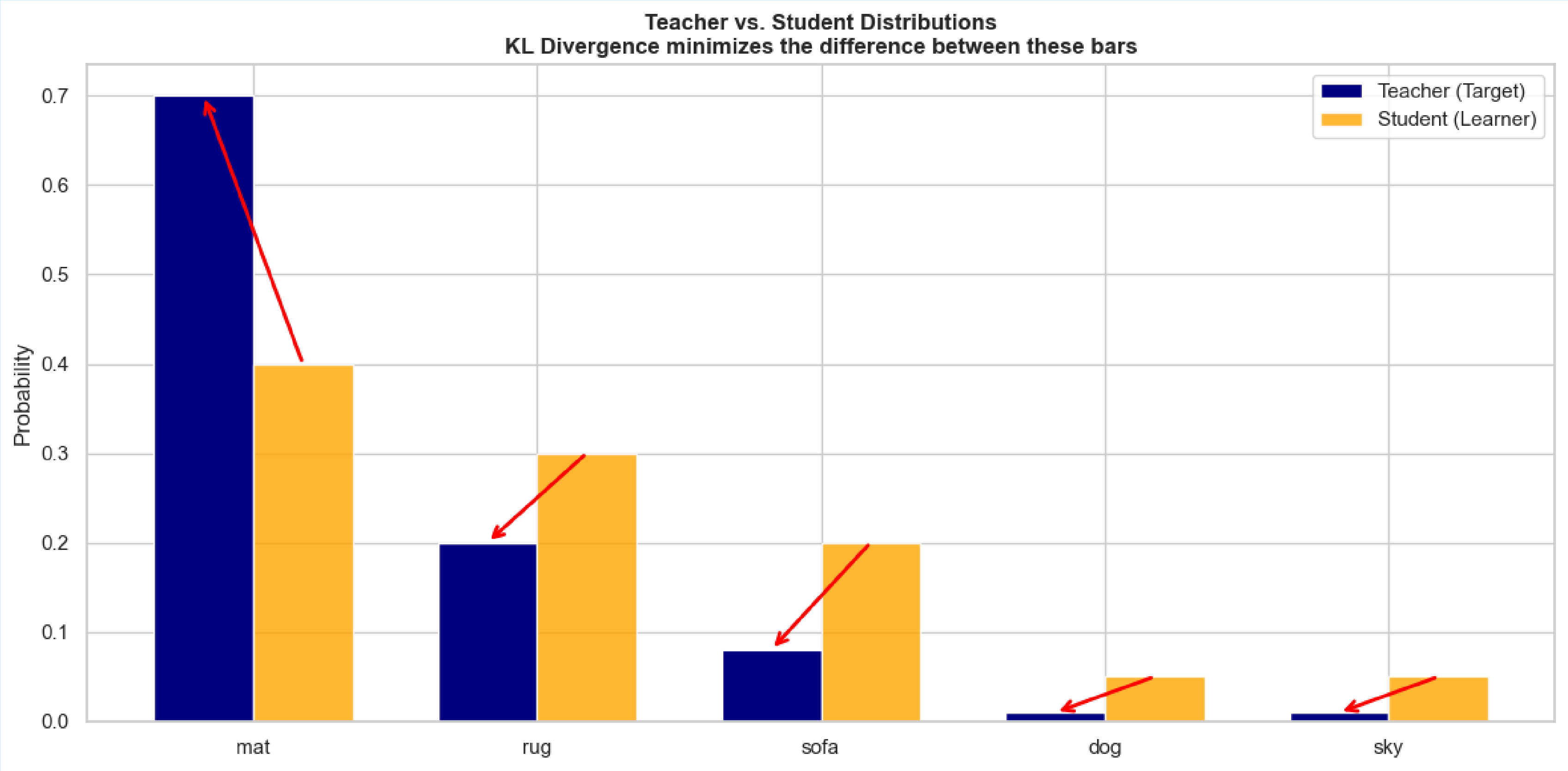


# White Box Distillation



- Compare probability distributions to make the student probabilities closer to the teacher's probabilities
- KL Divergence!

# White Box Distillation



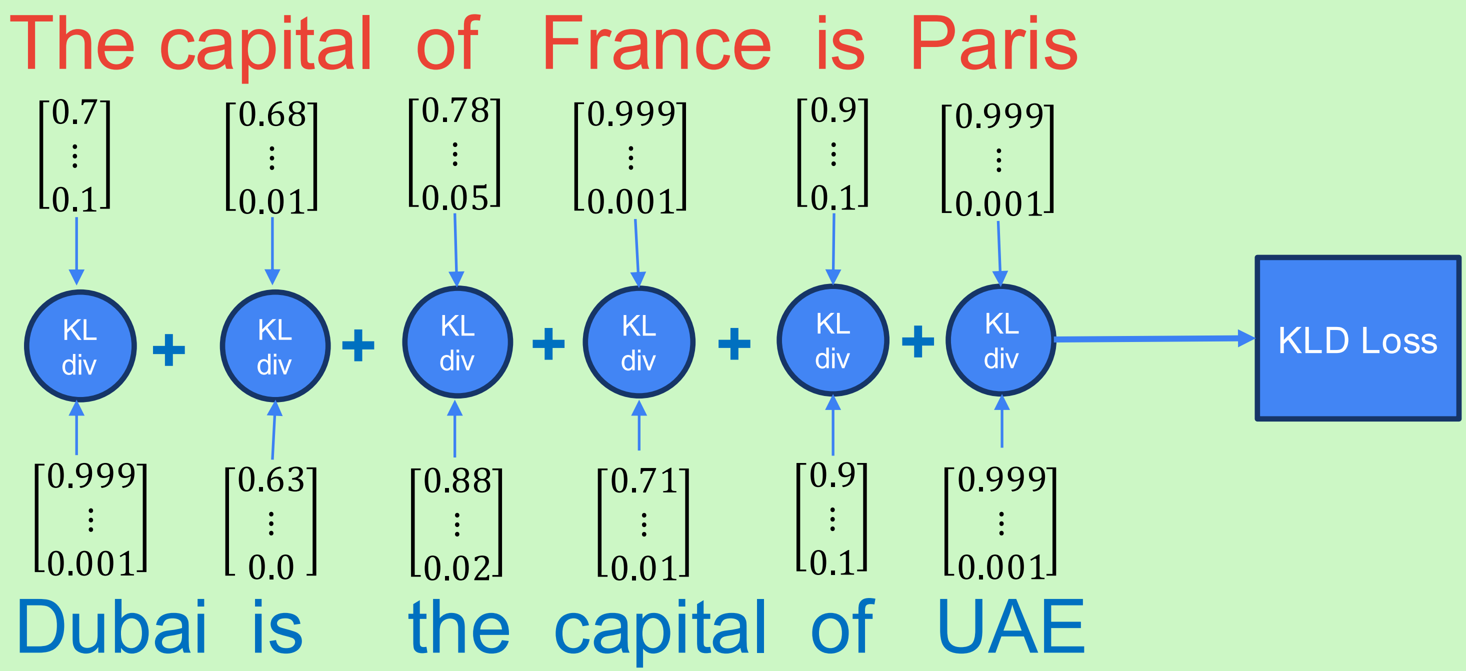
# KL Divergence

Probability of token $x$
by teacher model

$$D_{KL}(P||Q) = \sum_{x=1}^{vocab\_size} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Probability of token $x$
by student model

# KL Divergence



The capital of France is Paris

$\begin{bmatrix}0.7\\ \vdots\\ 0.1\end{bmatrix}$ $\begin{bmatrix}0.68\\ \vdots\\ 0.01\end{bmatrix}$ $\begin{bmatrix}0.78\\ \vdots\\ 0.05\end{bmatrix}$ $\begin{bmatrix}0.999\\ \vdots\\ 0.001\end{bmatrix}$ $\begin{bmatrix}0.9\\ \vdots\\ 0.1\end{bmatrix}$ $\begin{bmatrix}0.999\\ \vdots\\ 0.001\end{bmatrix}$

$\begin{bmatrix}0.999\\ \vdots\\ 0.001\end{bmatrix}$ $\begin{bmatrix}0.63\\ \vdots\\ 0.0\end{bmatrix}$ $\begin{bmatrix}0.88\\ \vdots\\ 0.02\end{bmatrix}$ $\begin{bmatrix}0.71\\ \vdots\\ 0.01\end{bmatrix}$ $\begin{bmatrix}0.9\\ \vdots\\ 0.1\end{bmatrix}$ $\begin{bmatrix}0.999\\ \vdots\\ 0.001\end{bmatrix}$

Dubai is the capital of UAE

# Let’s Code!

- Distill a fine-tuned BERT sentiment analysis classifier into a small DistillBERT model
  - Fast to run on Colab for this workshop!
  - The same logic follows for token generators like Gemma, Llama, etc

- Evaluate the model
  - Focus not just on accuracy, but on alignment with teacher model

- Deploy the smaller distilled model onto a **CPU** instance for inference on GCP
  - Once distilled, we no longer need GPU for fast inference!

# Google Nvidia Meetup

Dubai, UAE

github.com/ro1406/model-distillation

trygcp.dev/claim/gdg-google-nvidia-meetup

Google Developer Groups

# Thank You!

rohanmitra.dev

linkedin.com/in/rohan-mitra14/

**Rohan Mitra**

Machine Learning Engineer @Bayut & Dubizzle (OLX Group) | Research Assistant for Machine Lea...

Google Nvidia Meetup

Dubai

# Thank You!

rohanmitra.dev

My Website

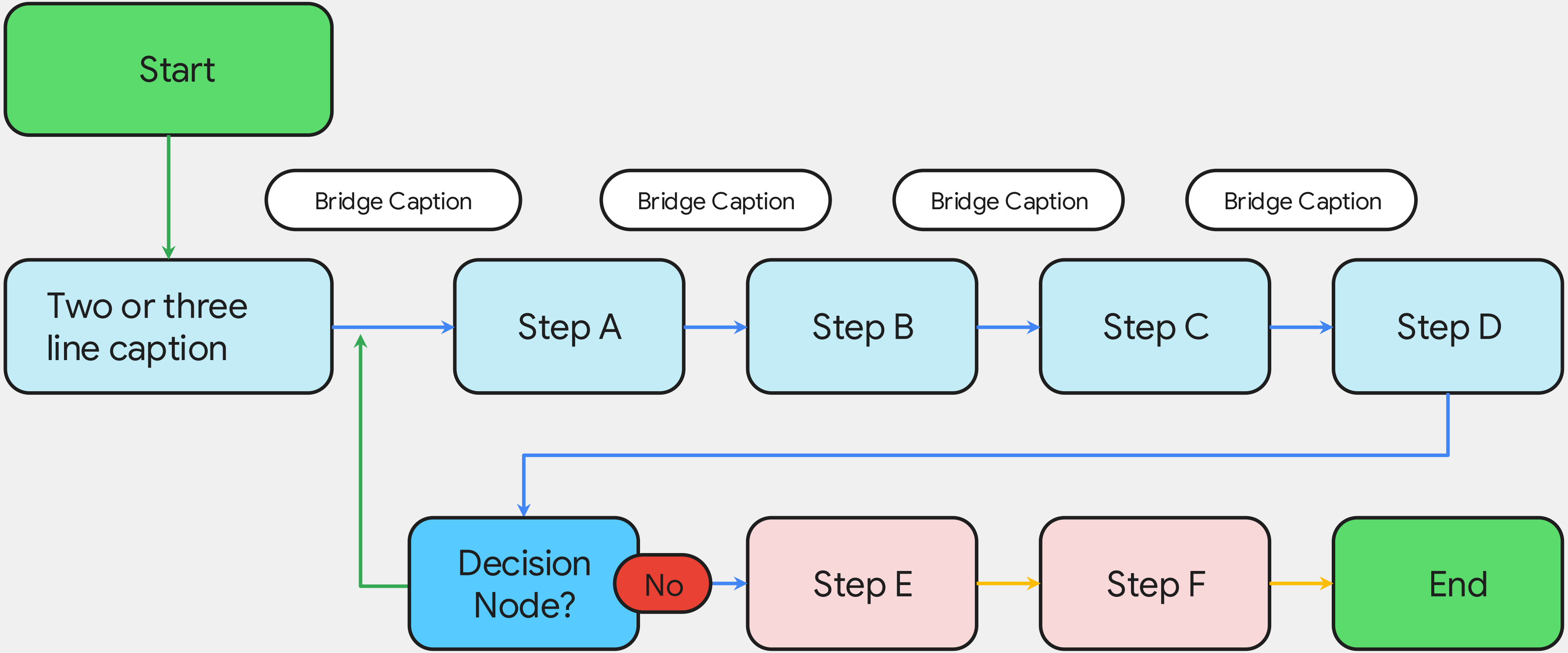
linkedin.com/in/rohan-mitra14/

**Rohan Mitra**

Machine Learning Engineer @Bayut & Dubizzle (OLX Group) | Research Assistant for Machine Lea...

Google Developer Groups

# Our Architecture



Start

Bridge Caption

Bridge Caption

Bridge Caption

Bridge Caption

Two or three line caption

Step A

Step B

Step C

Step D

Decision Node?

No

Step E

Step F

End

# Two Column Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vehicula nulla a leo placerat, in convallis justo molestie. Ut id maximus mauris, vitae pharetra justo.

1. Bullet one
2. Bullet two
3. Bullet three
4. Bullet four
5. Bullet five

# Two Column Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vehicula nulla a leo placerat, in convallis justo molestie. Ut id maximus mauris, vitae pharetra justo.

1. Bullet one
2. Bullet two
3. Bullet three
4. Bullet four
5. Bullet five

# Two Column Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vehicula nulla a leo placerat, in convallis justo molestie. Ut id maximus mauris, vitae pharetra justo.

1. Bullet one
2. Bullet two
3. Bullet three
4. Bullet four
5. Bullet five

# Two Column Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vehicula nulla a leo placerat, in convallis justo molestie. Ut id maximus mauris, vitae pharetra justo.

1. Bullet one
2. Bullet two
3. Bullet three
4. Bullet four
5. Bullet five

Cloud
@DevFest

# This is an example of a Cloud-specific section title slide.

Google Developer Groups

Web
@DevFest

# This is an example of a Web-specific section title slide.

Google
Developer
Groups

Mobile
@DevFest

# This is an example of a Mobile-specific section title slide.

Google
Developer
Groups

AI
@DevFest

# This is an example of an AI-specific section title slide.

Google Developer Groups

# Color & Typography

|  |  |  |  |  |
|---|---|---|---|---|
| **Google Blue 500**<br>#4285f4 | **Google Green 500**<br>#34a853 | **Google Yellow 500**<br>#fbbc04 | **Google Red 500**<br>#ea4335 |  |
| **Halftone Blue**<br>#57caff | **Halftone Green**<br>#5cdb6d | **Halftone Yellow**<br>#ffd427 | **Halftone Red**<br>#ff7daf | **Off-White**<br>#f0f0f0 |
| **Pastel Blue**<br>##c3ecf6 | **Pastel Green**<br>#ccf6c5 | **Pastel Yellow**<br>#ffe7a5 | **Pastel Red**<br>#f8d8d8 | **Black 02**<br>#1e1e1e |

# Headlines

**Google Sans Bold**

## Subheads

Roboto Mono Light

## Body

Google Sans Normal

Google Developer Groups

# Slide Formats

# Full screen slides

Use full-screen slides whenever it makes the most sense for the content you're creating. The extra slide real estate for full-screen allows for larger graphics and larger text.

If you're doing a VO or designing a full-screen animated video, you would use this layout for every slide in your deck.

DevFest

Mauritius

Simple statement, URL or quote goes here. Limit text to four lines or less.

Google Developer Groups

Cloud-specific simple statement, URL or quote goes here. Limit text to four lines or less.

Google
Developer
Groups

Cloud
@DevFest

Web-specific simple statement, URL or quote goes here. Limit text to four lines or less.

Google Developer Groups

Web @DevFest

Google Developer Groups

Mobile @DevFest

Google
Developer
Groups

AI
@DevFest

# Left Aligned Title

1. Bullet one
2. Bullet two, adjust size to match length of content
3. Bullet three
4. Bullet four, a little longer example
5. A final bullet point

# Left Aligned Title

1. Bullet one
2. Bullet two, adjust size to match length of content
3. Bullet three
4. Bullet four, a little longer example
5. A final bullet point

# Left Aligned Title

1. Bullet one
2. Bullet two, adjust size to match length of content
3. Bullet three
4. Bullet four, a little longer example
5. A final bullet point

# Left Aligned Title

1. Bullet one
2. Bullet two, adjust size to match length of content
3. Bullet three
4. Bullet four, a little longer example
5. A final bullet point

# Left Aligned Title

1. Bullet one
2. Bullet two, adjust size to match length of content
3. Bullet three
4. Bullet four, a little longer example
5. A final bullet point

# Left Aligned Title

( 1 )

**Bullet Point One:**
Can add further details

( 2 )

**Bullet Point Two:**
Can add further details

( 3 )

**Bullet Point Three:**
Can add further details

Use more than

# 50%

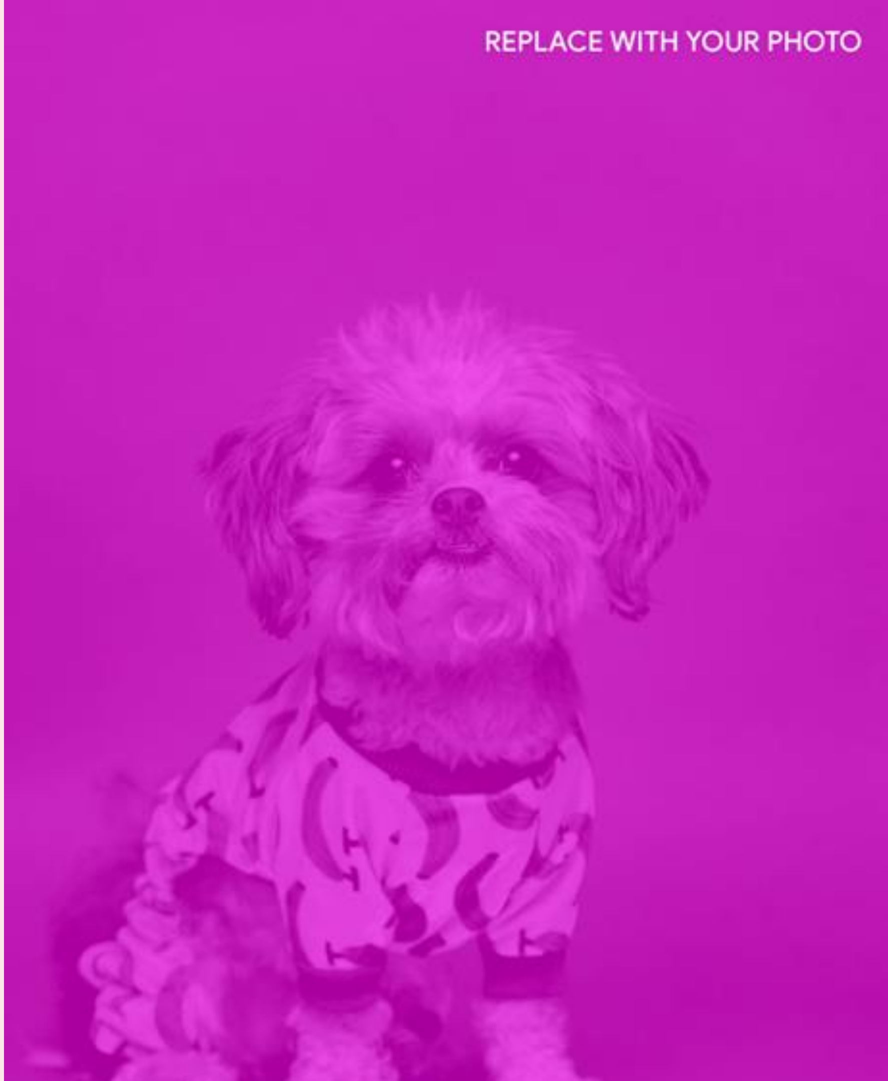
of this large number slide

# Half screen speaker

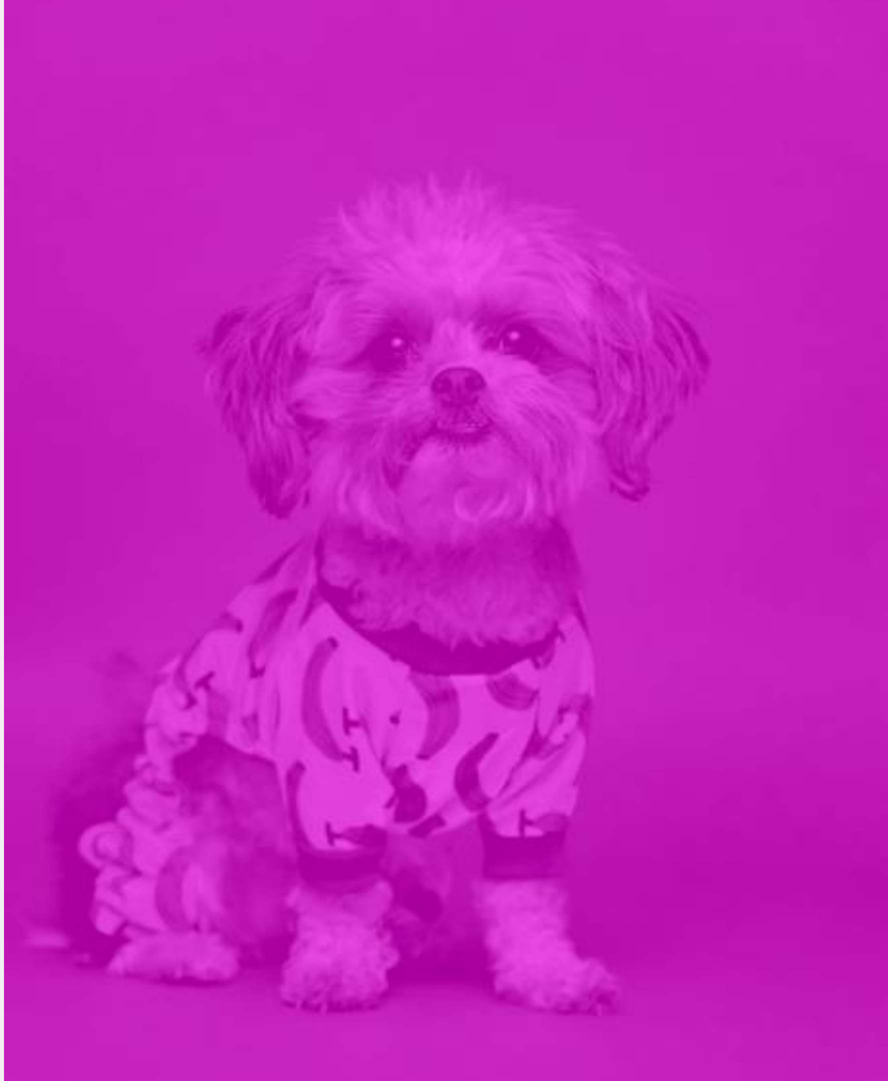
Presenter slides are a great way to engage your audience by showing the presenter on on the right half of the screen and graphics on the left.

When designing presenter slides make sure your graphics never cross the centerline. This will make sure there is enough space to fit both the presenter and the graphic.

If your content is too large to fit on a presenter slide or is better for the learner with more space, please use a full-screen slide instead.



Simple quote or statement goes here. Ideally limit to four or five lines max.



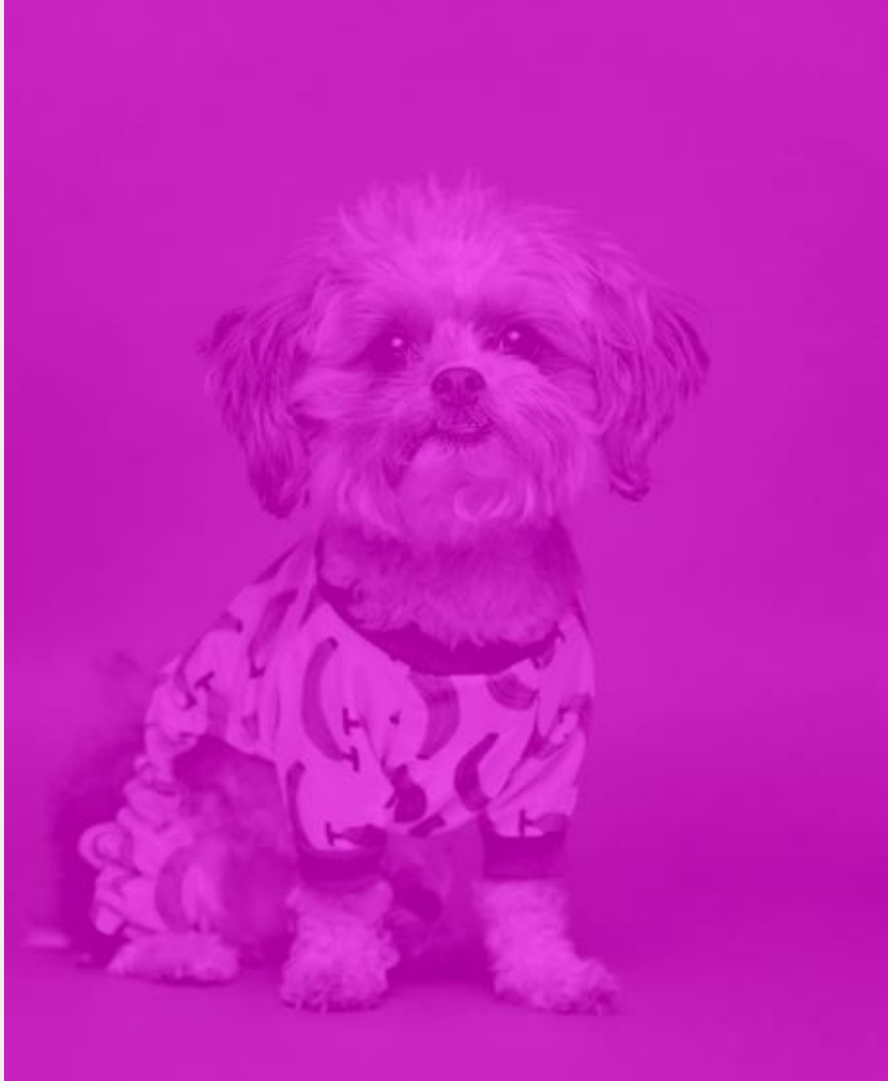
Google Developer Groups

Simple quote or statement goes here. Ideally limit to four or five lines max.

Use more than

# 50%

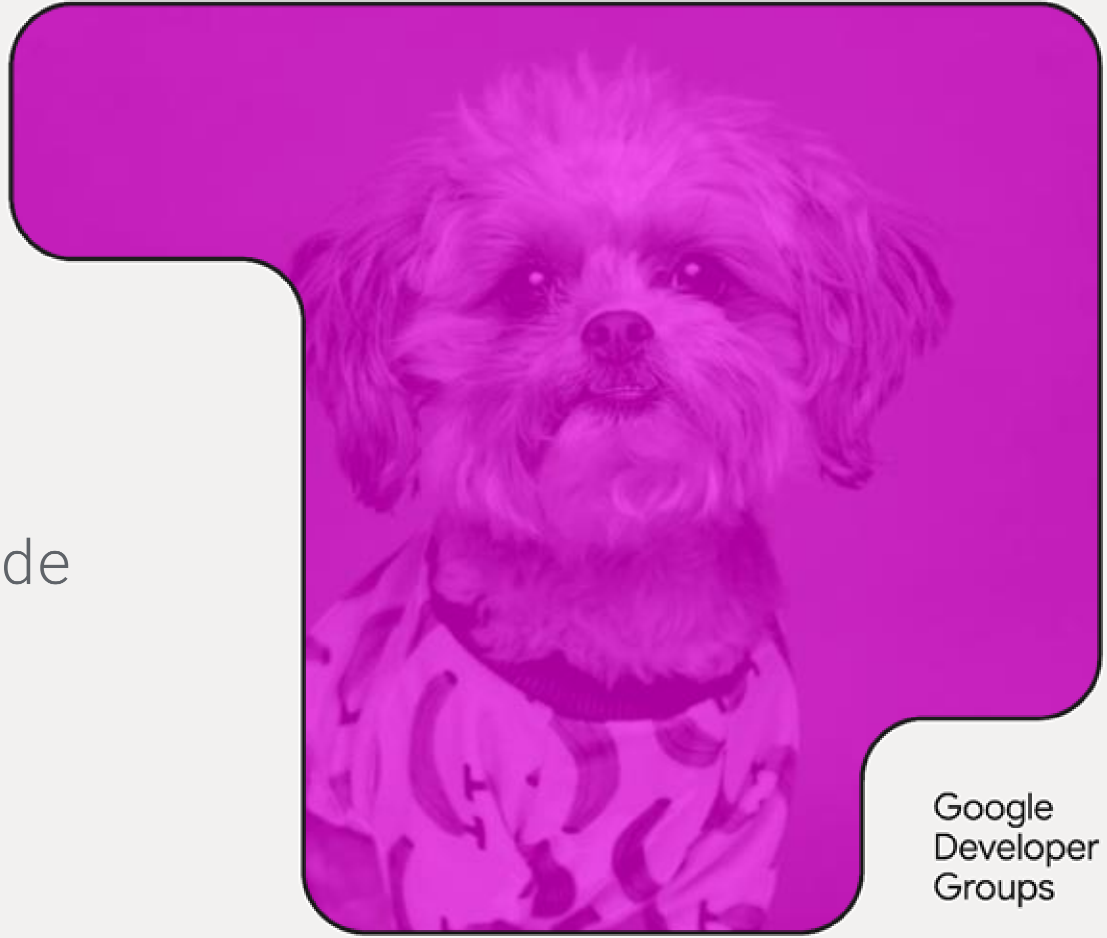
Of this number slide



Use more than

# 50%

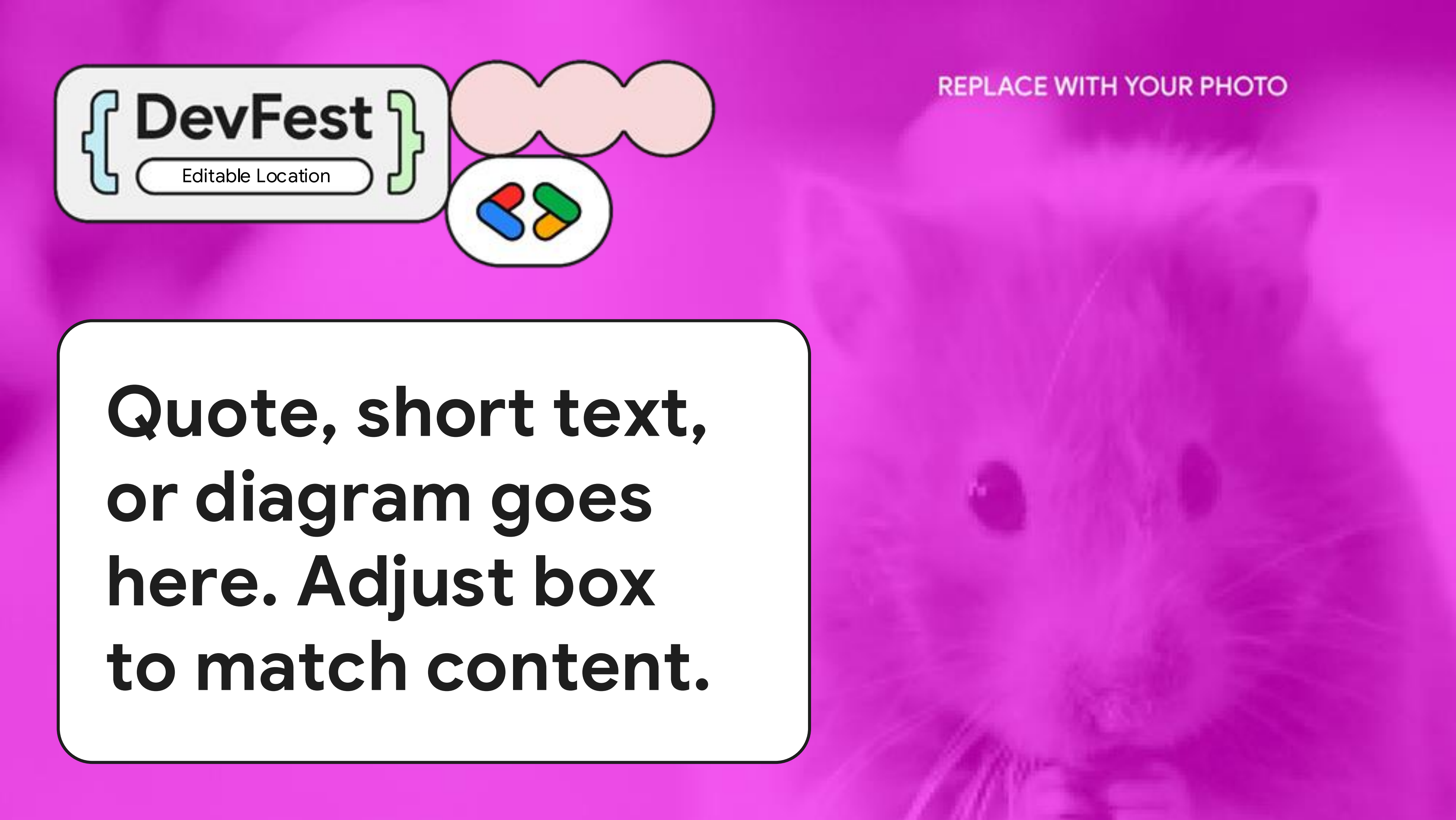
Of this number slide



Google Developer Groups

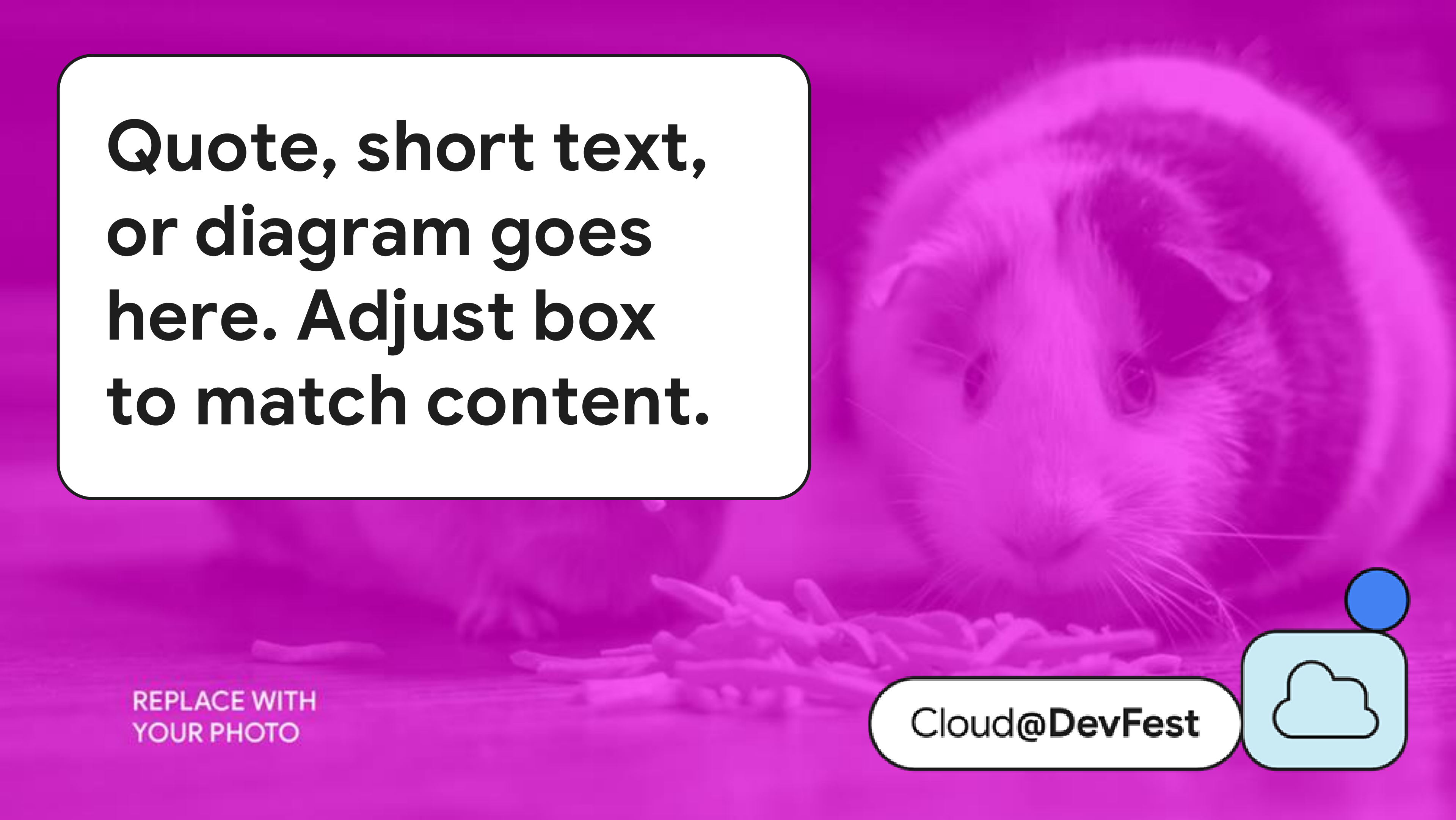
DevFest

Editable Location


REPLACE WITH YOUR PHOTO

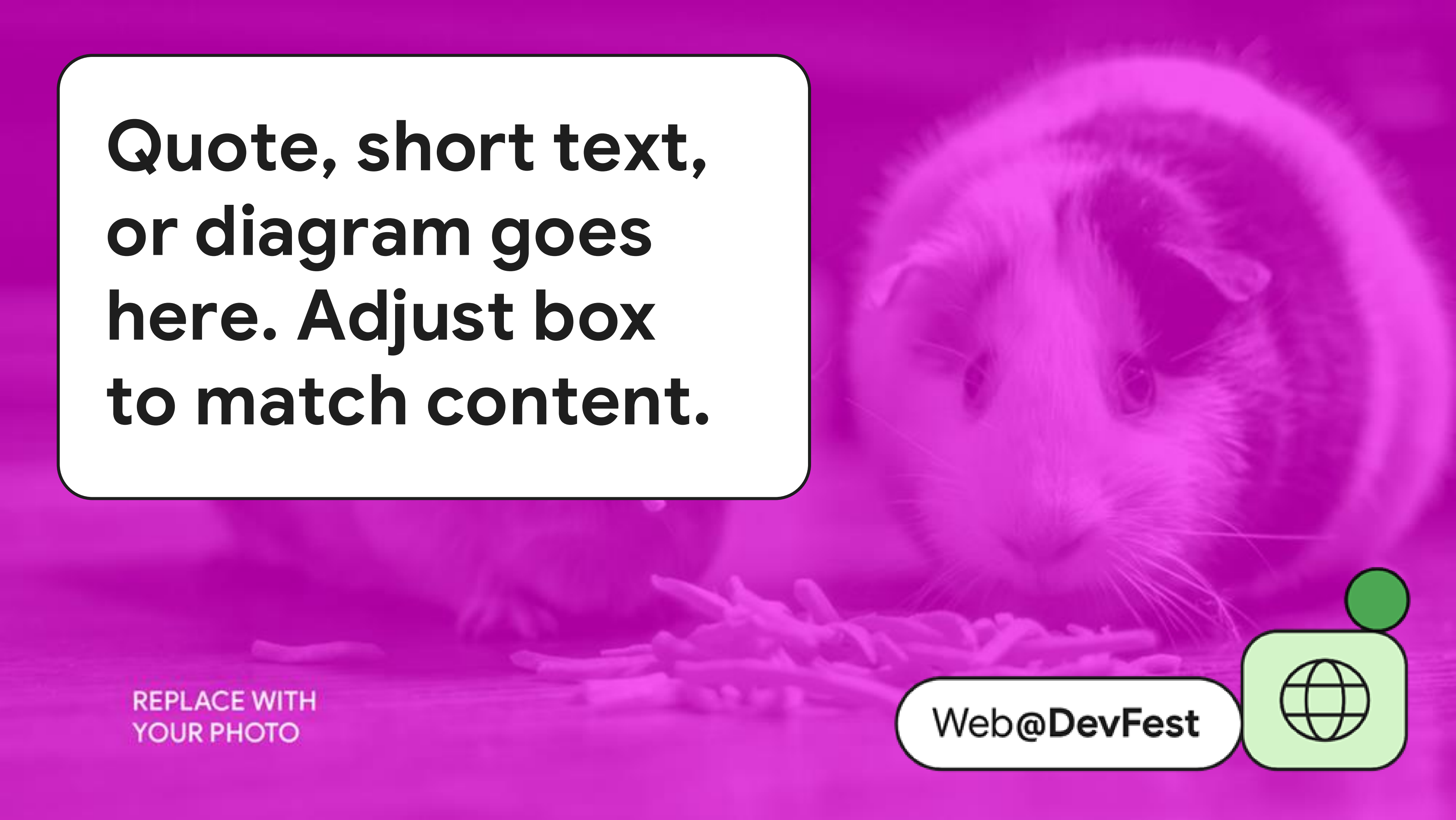
Quote, short text, or diagram goes here. Adjust box to match content.

Quote, short text, or diagram goes here. Adjust box to match content.



REPLACE WITH YOUR PHOTO

Cloud@**DevFest**

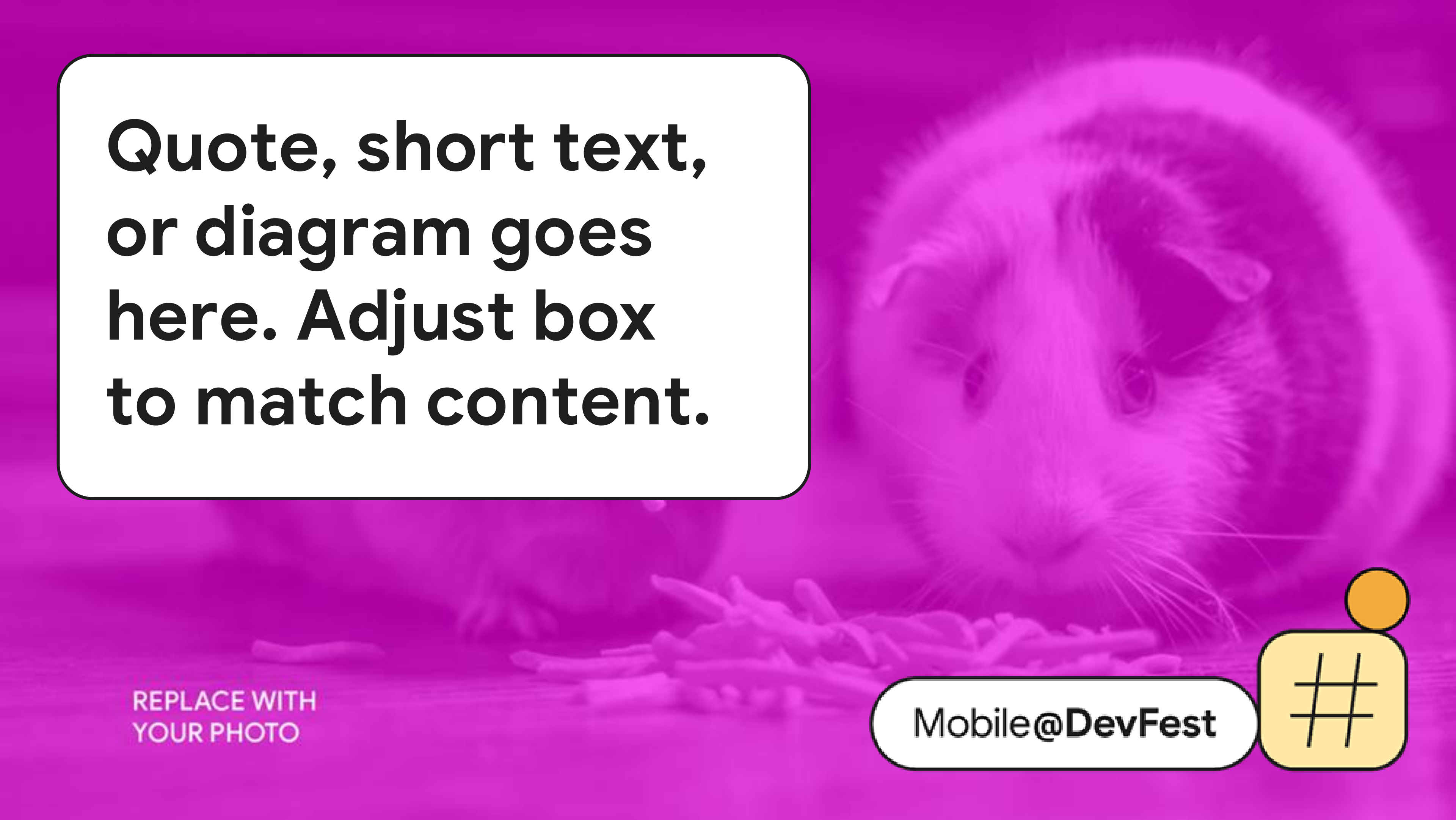


Quote, short text, or diagram goes here. Adjust box to match content.

REPLACE WITH YOUR PHOTO

Web@DevFest

Quote, short text, or diagram goes here. Adjust box to match content.



REPLACE WITH
YOUR PHOTO

Mobile@**DevFest**

Quote, short text, or diagram goes here. Adjust box to match content.

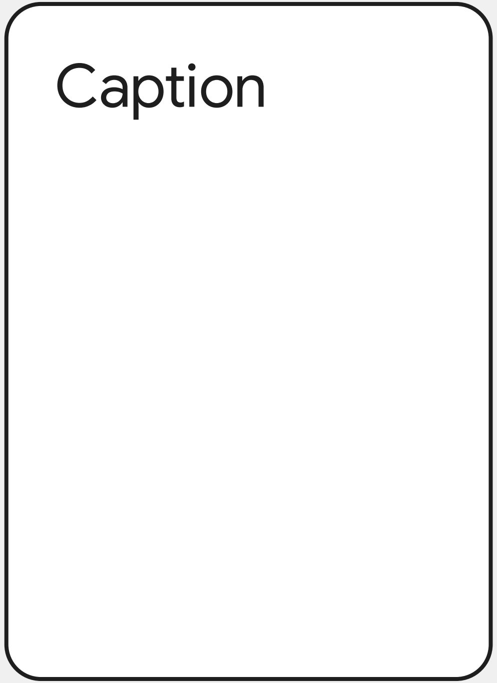
REPLACE WITH
YOUR PHOTO

AI@DevFest

# Charts

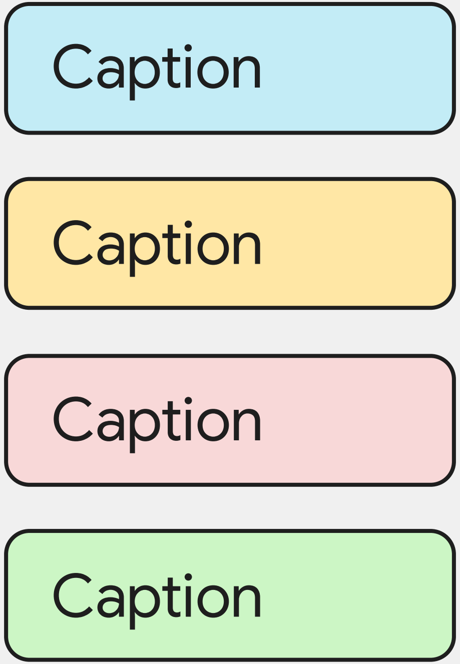
# Chart Elements

## Large Box



Caption

## Medium Box



Caption

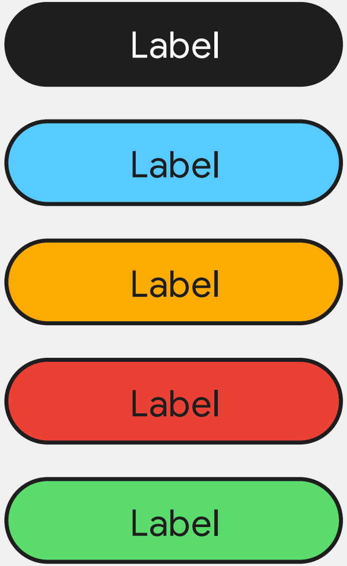
Caption

Caption

Caption

## Labels

Only use colored labels on corresponding medium box color (blue with pastel blue etc.)



Label

Label

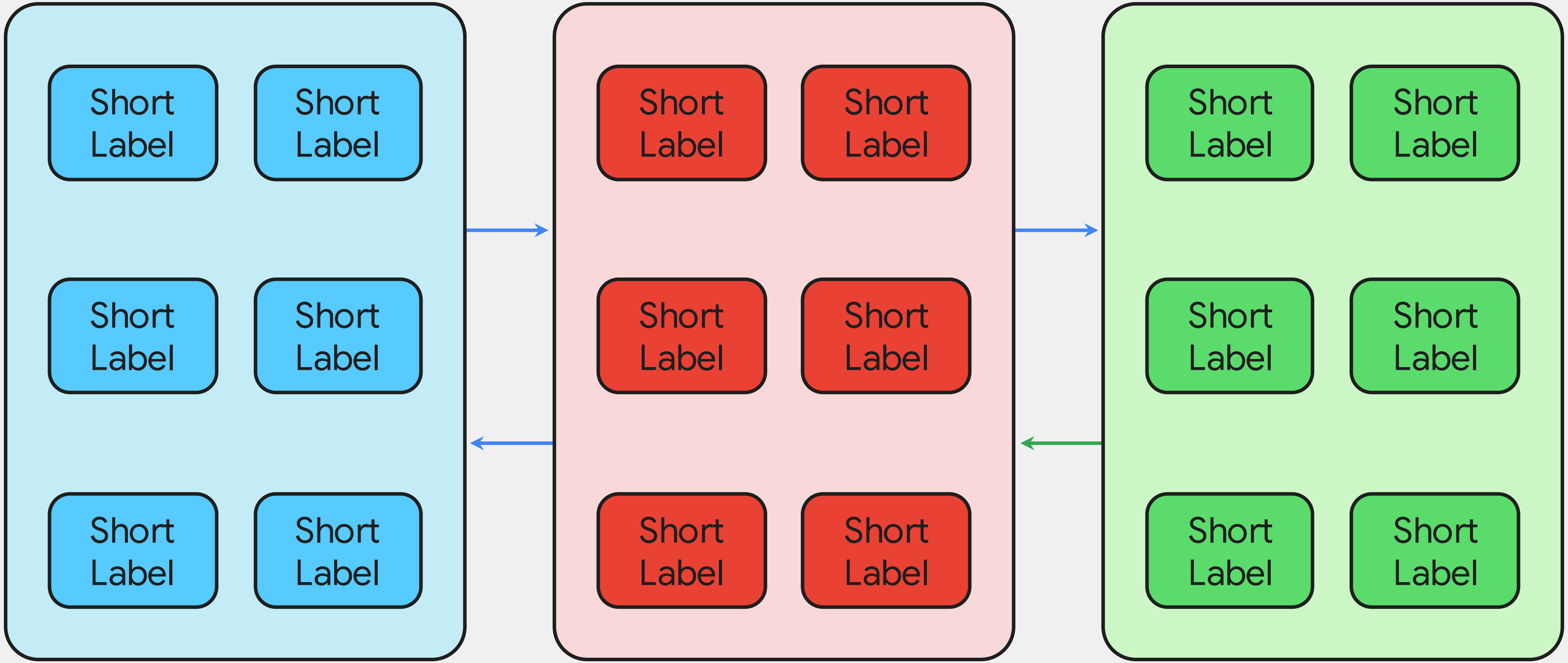
Label

Label

Label

## Arrows

# Simple Chart



Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

Short Label

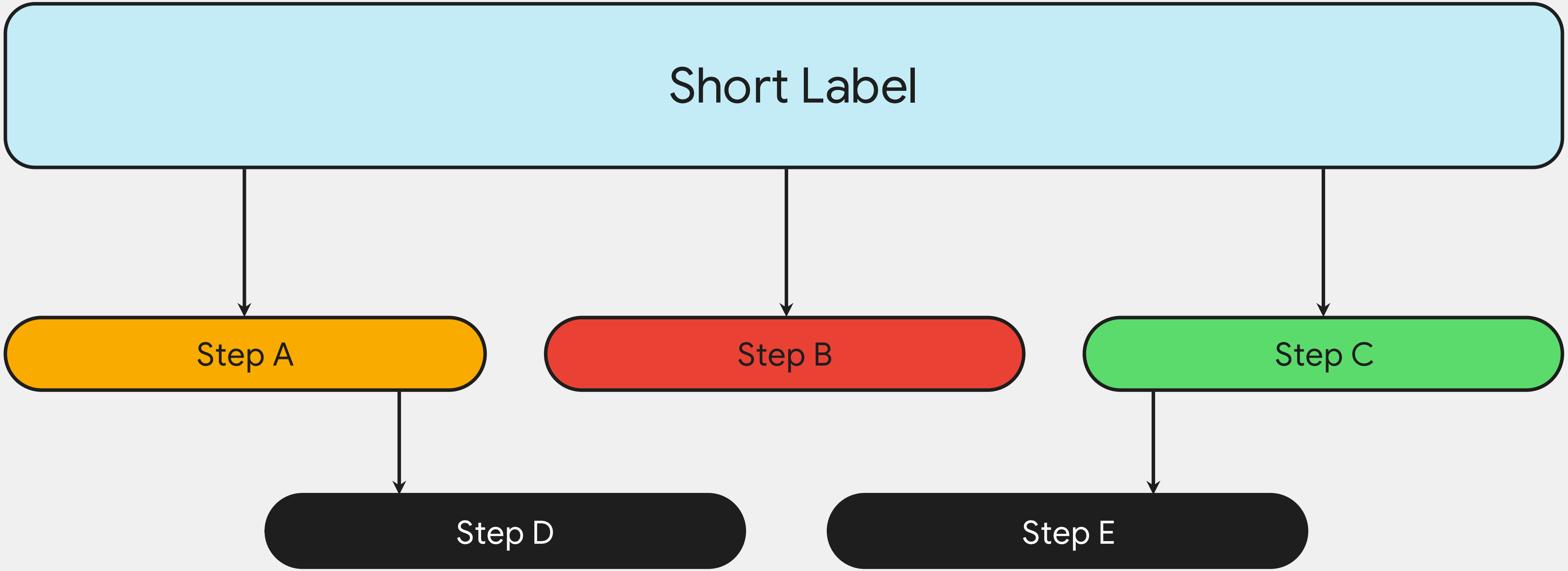
Short Label

Short Label

Short Label

Short Label

# All-Purpose Chart



Short Label

Step A

Step B

Step C

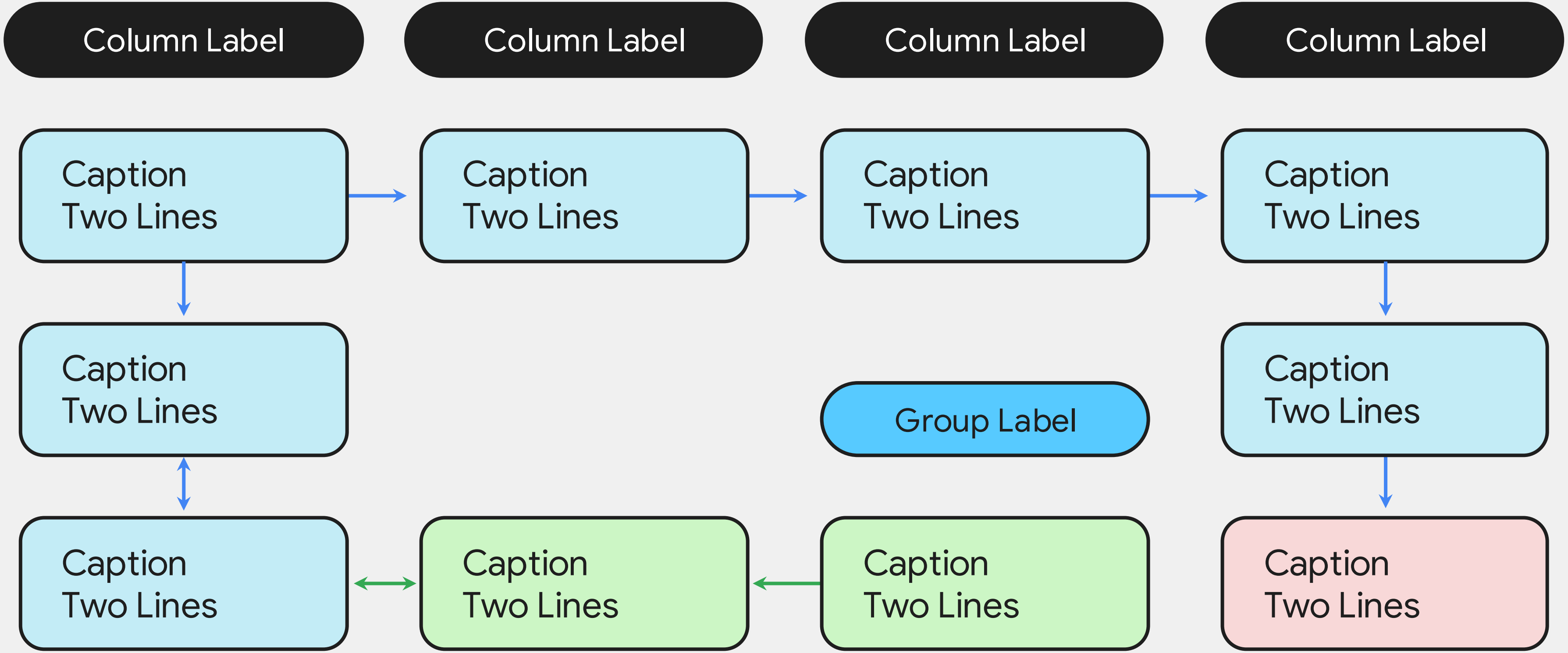
Step D

Step E

# Big Portfolio Chart

| Label One | Label Two | Label Three | Label Four |
|---|---|---|---|
| Short Label | Short Label | Short Label | Short Label |
| Short Label | Short Label | Short Label | Short Label |
| Short Label | Short Label | Short Label | Short Label |

# Process Chart



Column Label

Column Label

Column Label

Column Label

Caption
Two Lines

Caption
Two Lines

Caption
Two Lines

Caption
Two Lines

Caption
Two Lines

Group Label

Caption
Two Lines

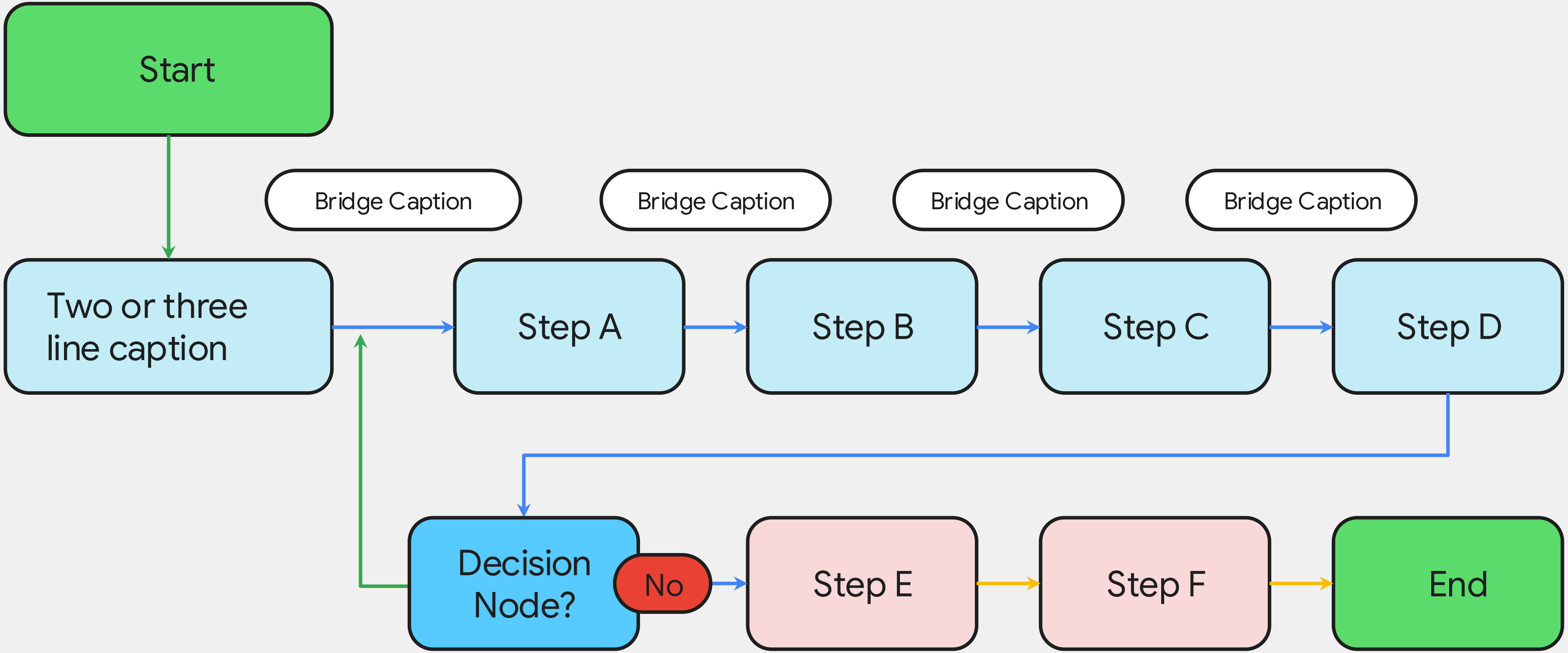
Caption
Two Lines

Caption
Two Lines

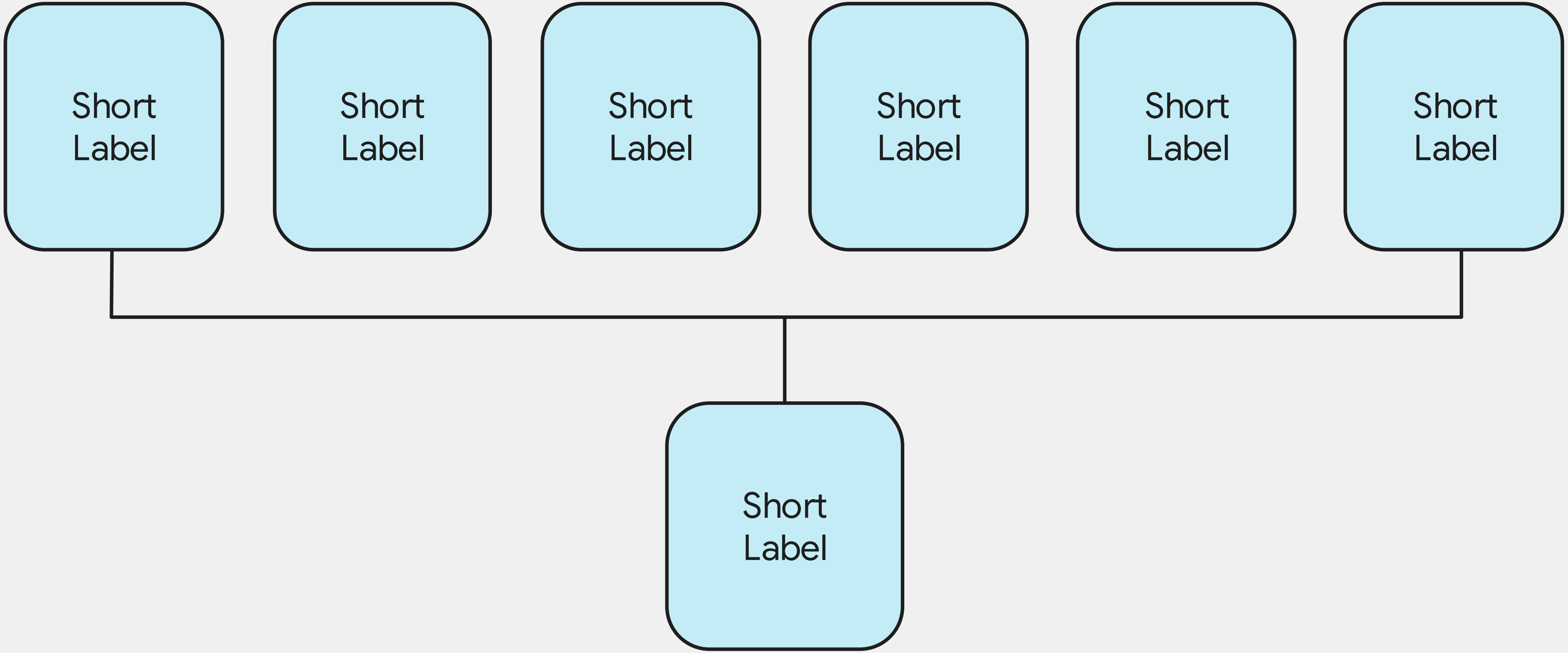
Caption
Two Lines

Caption
Two Lines

# Flow-Style Chart



# Group Chart



Short Label

Short Label

Short Label

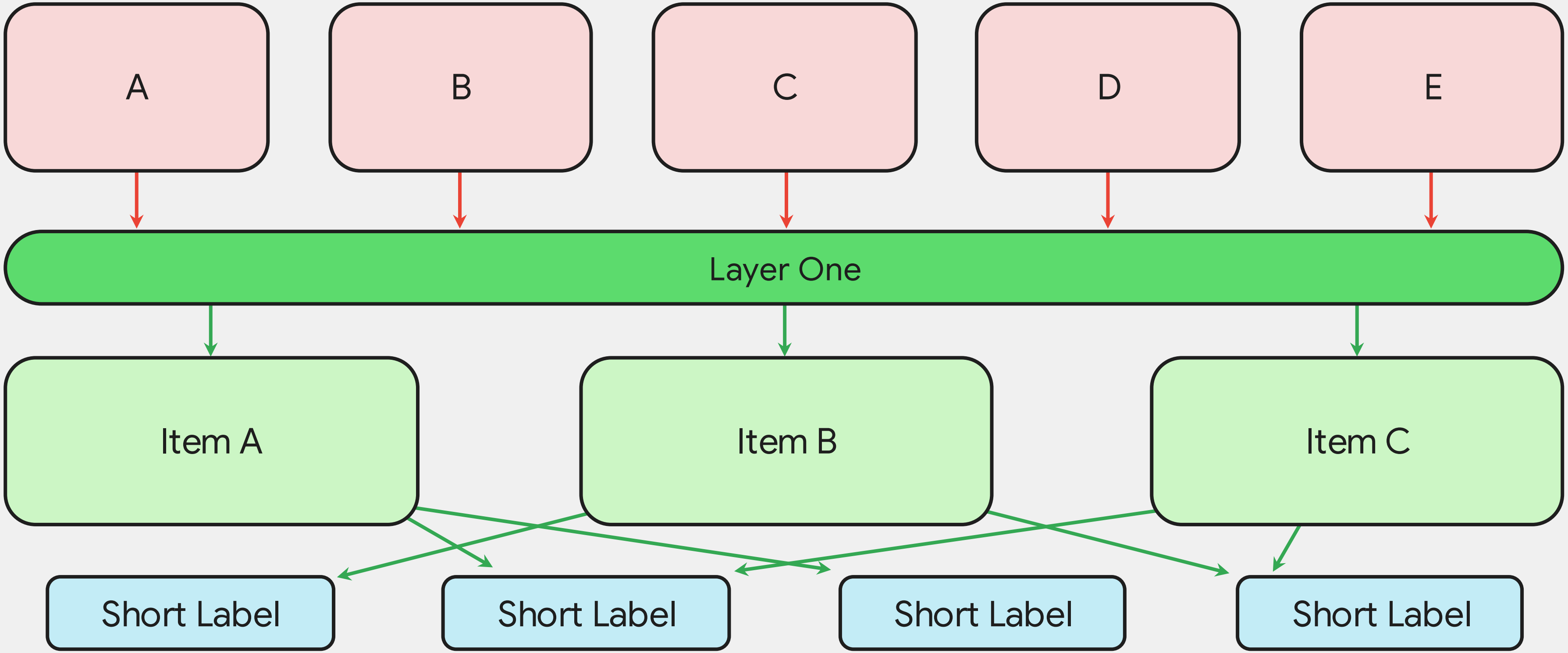
Short Label

Short Label

Short Label

Short Label

# Process Chart



A

B

C

D

E

Layer One

Item A

Item B

Item C

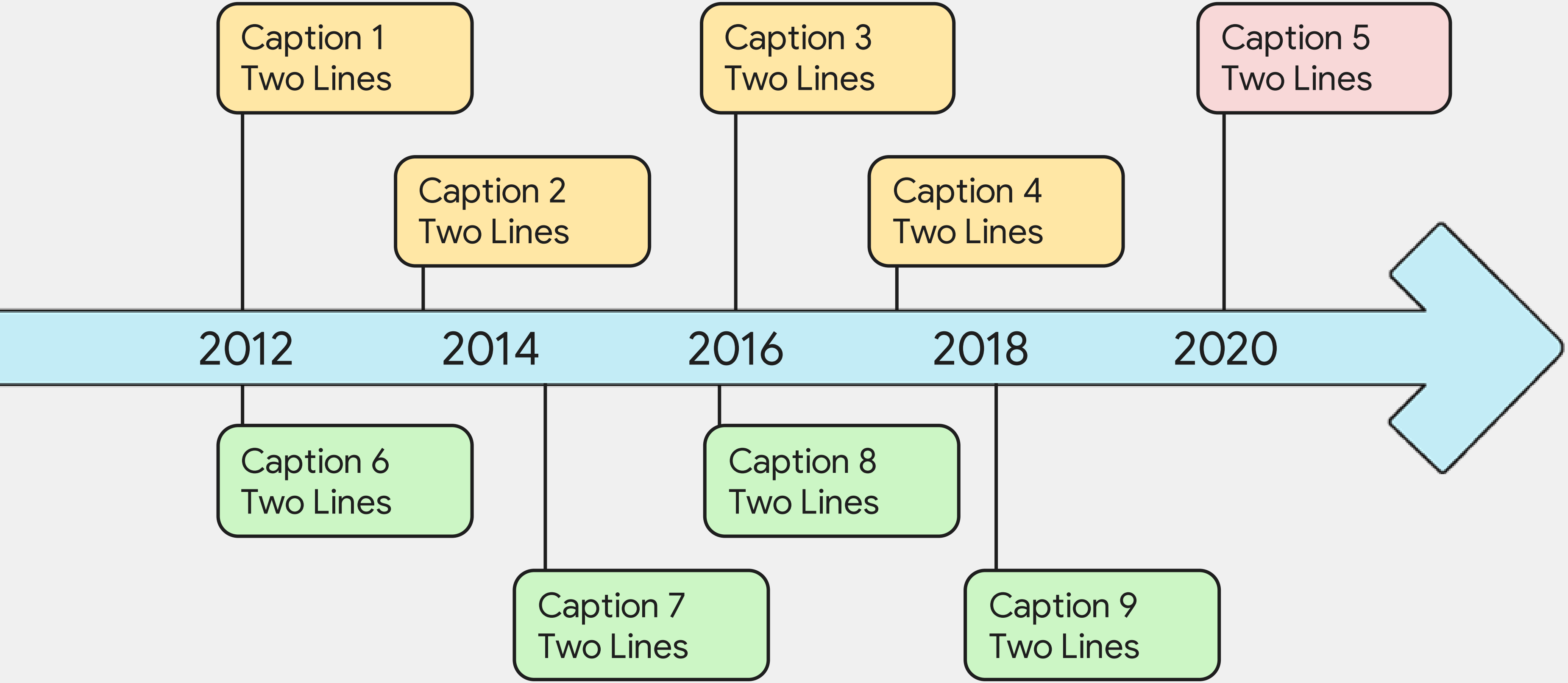
Short Label

Short Label

Short Label

Short Label

# Chart Title



Caption 1
Two Lines

Caption 2
Two Lines

Caption 3
Two Lines

Caption 4
Two Lines

Caption 5
Two Lines

2012

2014

2016

2018

2020

Caption 6
Two Lines

Caption 7
Two Lines

Caption 8
Two Lines

Caption 9
Two Lines

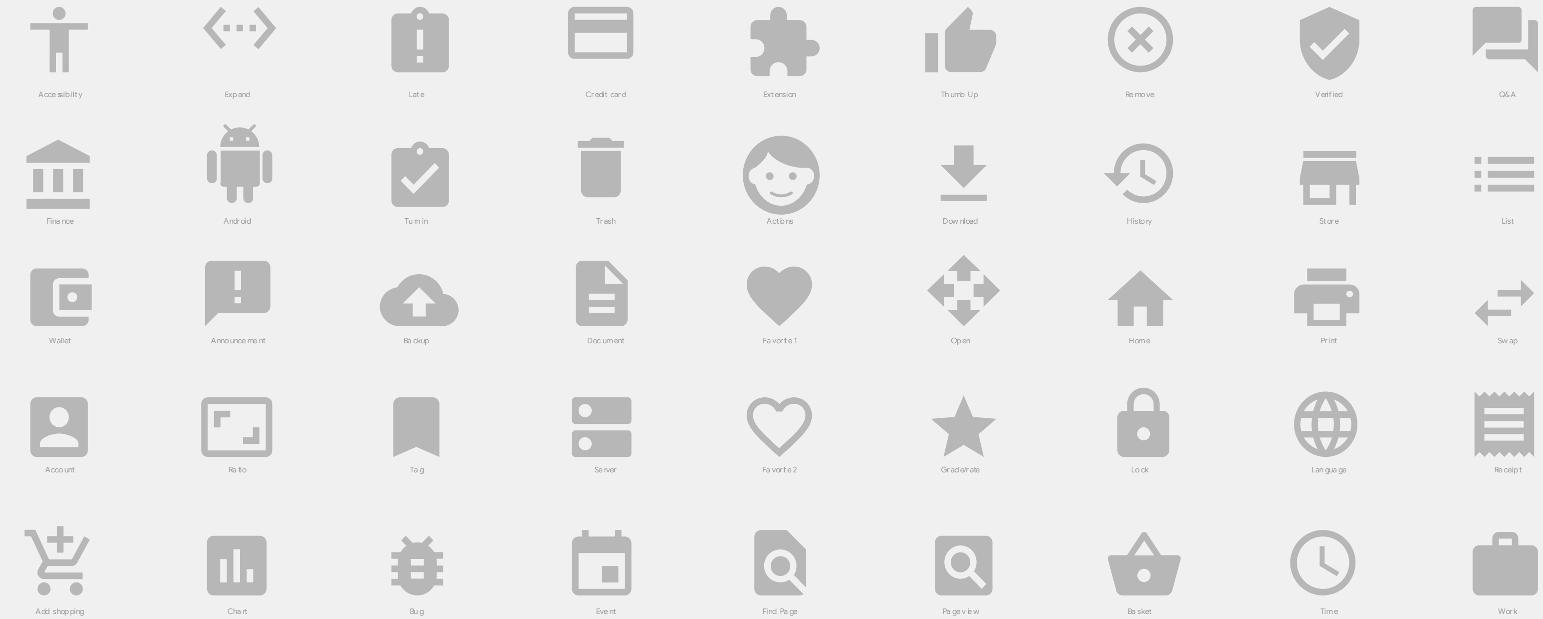
# Icons

# Icons

All icons are vector objects and can be recolored using the fill menu.

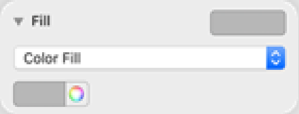
Fill

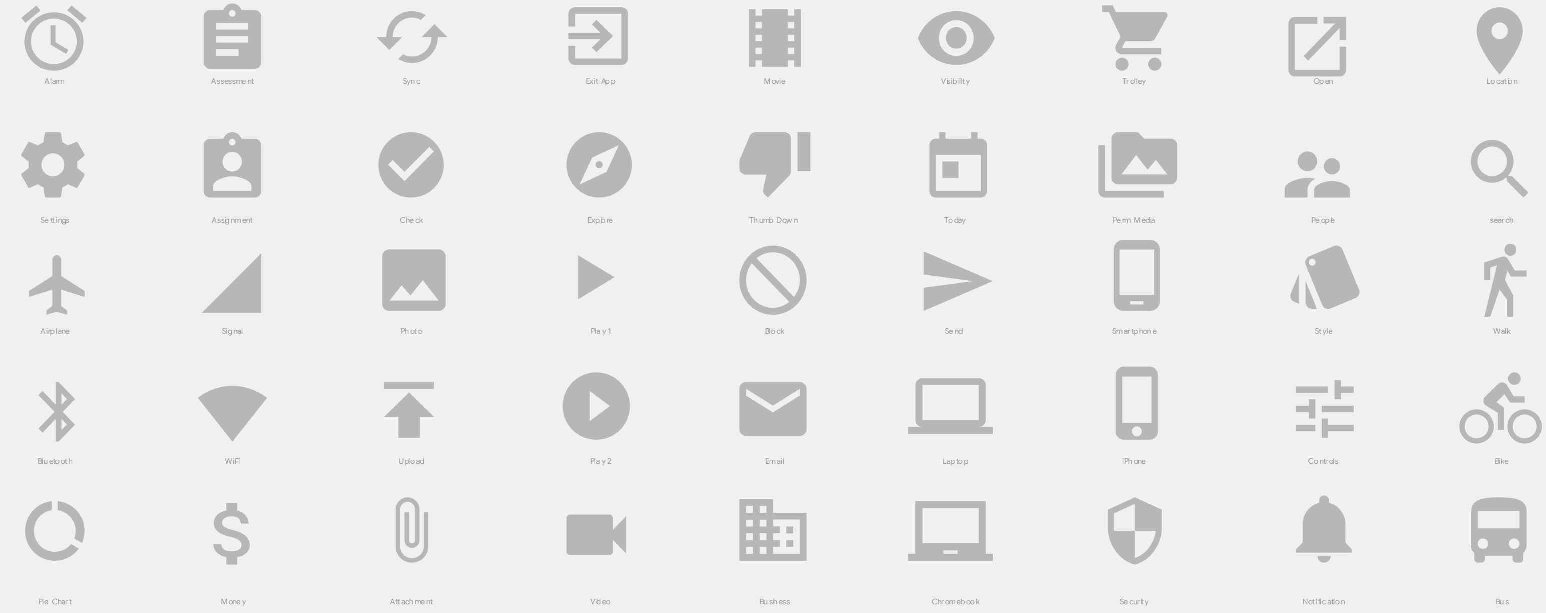
Color Fill



# Icons

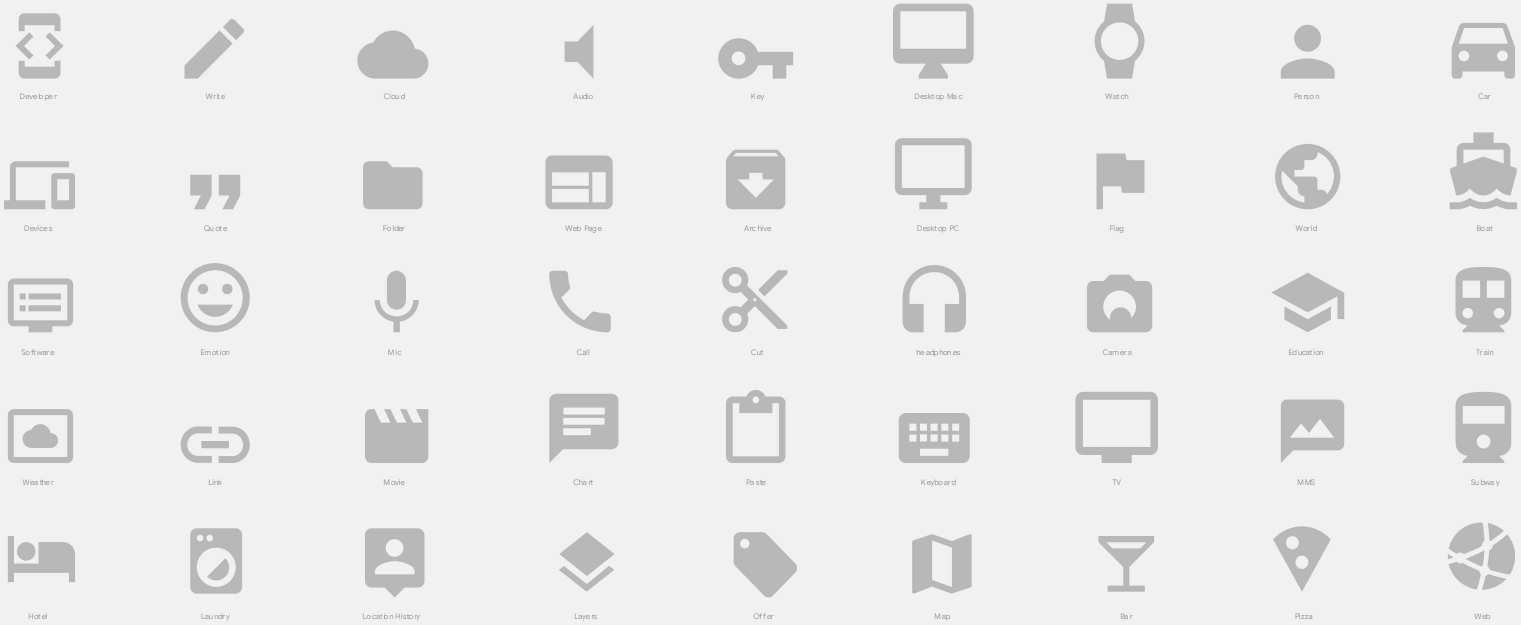
All icons are vector objects and can be recolored using the fill menu.





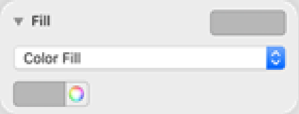
# Icons

All icons are vector objects and can be recolored using the fill menu.

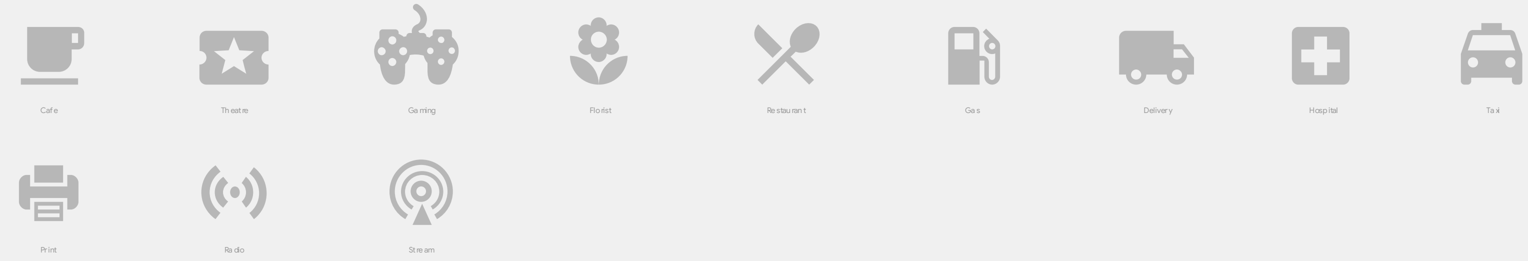


# Icons

All icons are vector objects and can be recolored using the fill menu.

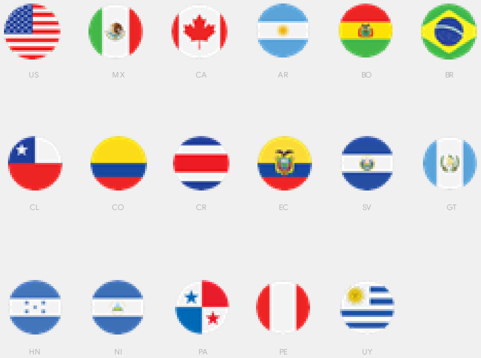


Color Fill

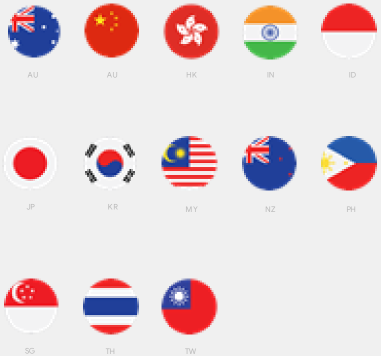


# Flags

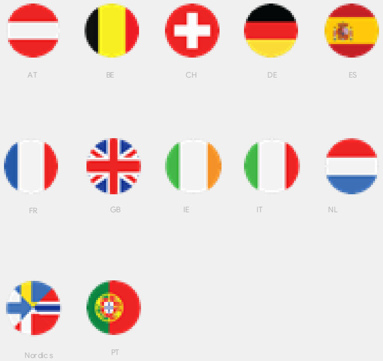
Americas



APAC



EMEA

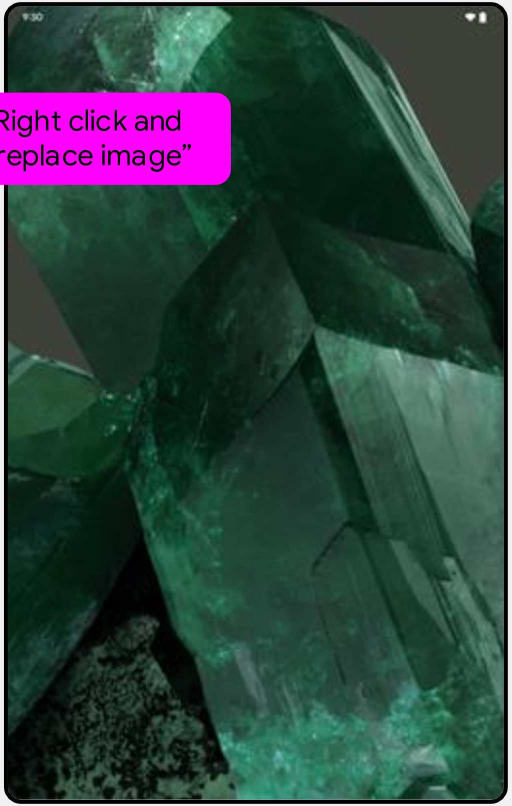


WIREFRAME - MOBILE

Right click and “replace image”

Right click and “replace image”

WIREFRAME - TABLET



Right click and
“replace image”

Right click and
“replace image”

WIREFRAME - WEAR OS

Right click and “replace image”

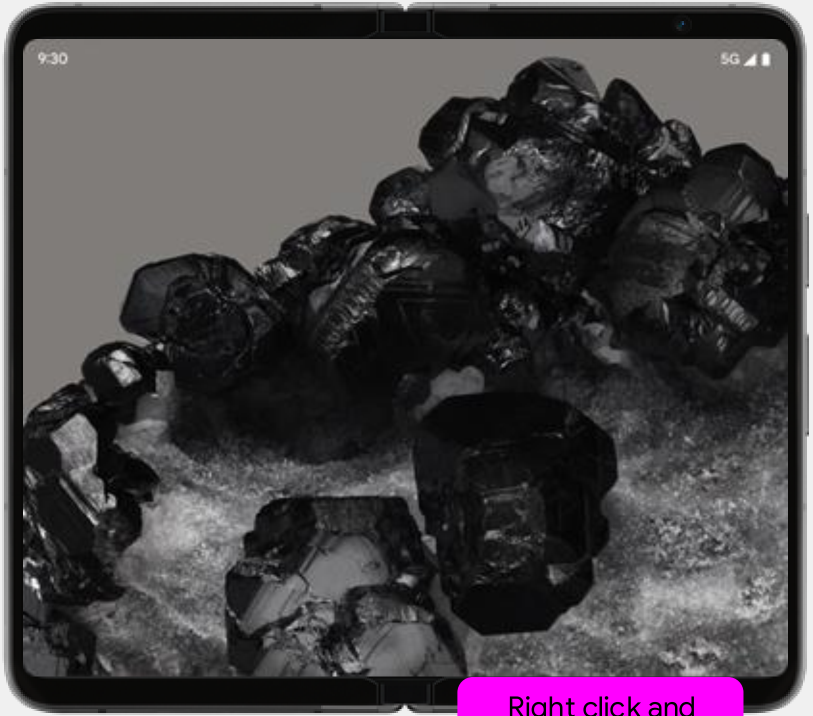
DEVICE FRAME - PIXEL 8

Right click and
“replace image”

Right click and
“replace image”

DEVICE FRAME - PIXEL FOLD

Right click and “replace image”



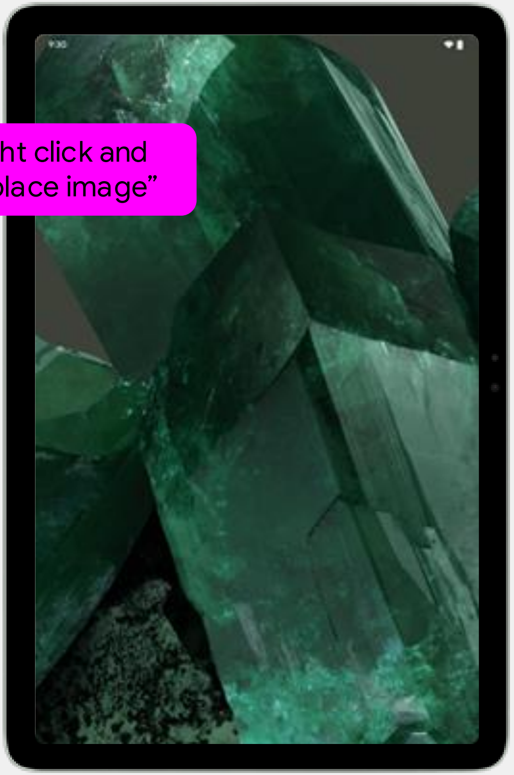
Right click and “replace image”

DEVICE FRAME - IPHONE 14

Right click and
“replace image”

Right click and
“replace image”

DEVICE FRAME - TABLET



Right click and “replace image”

Right click and “replace image”

DEVICE FRAME - WEAR OS

Right click and "replace image"

DEVICE FRAME - CHROMEBOOK PLUS

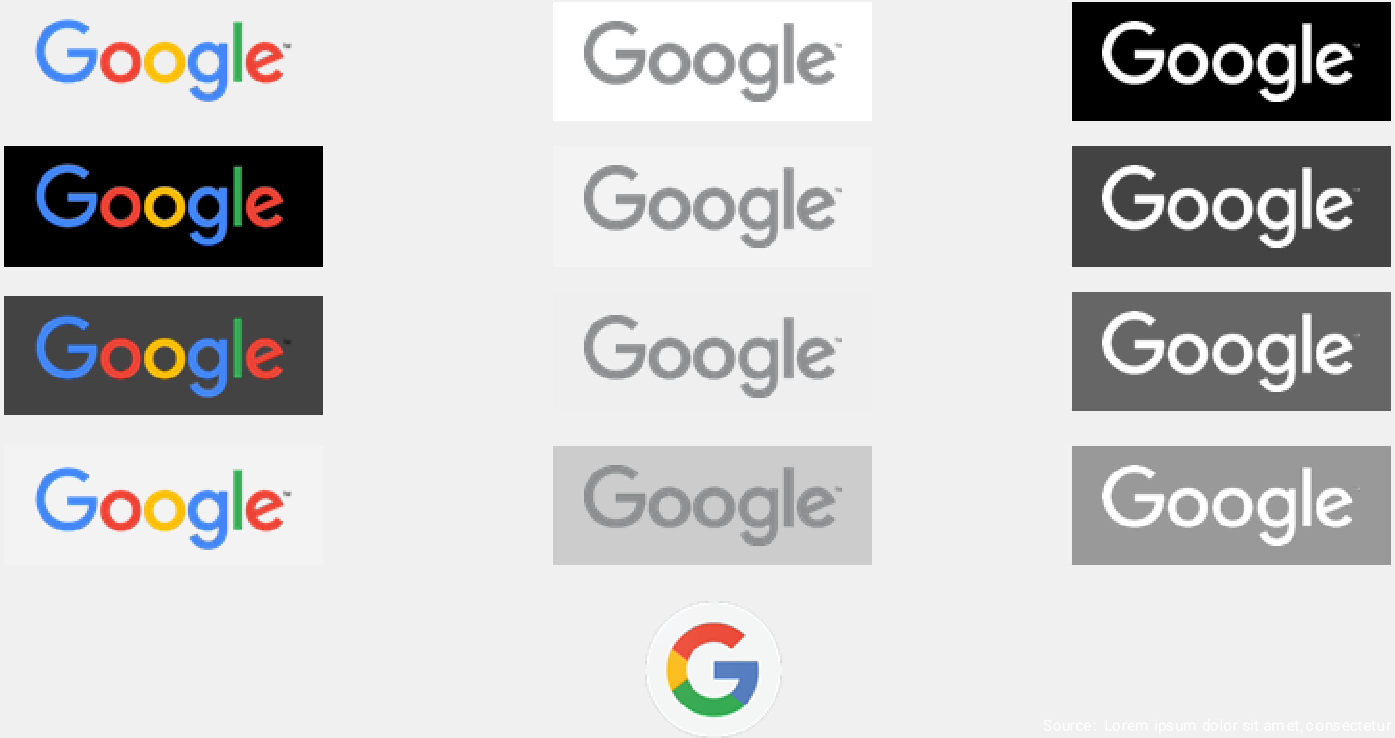
Right click and “replace image”

DEVICE FRAME - PIXELBOOK GO

Right click and
“replace image”

# Logo Library

Logos can be scaled to any size



Source: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis non erat sem