Seeing is Believing

Analytics of Visual Data

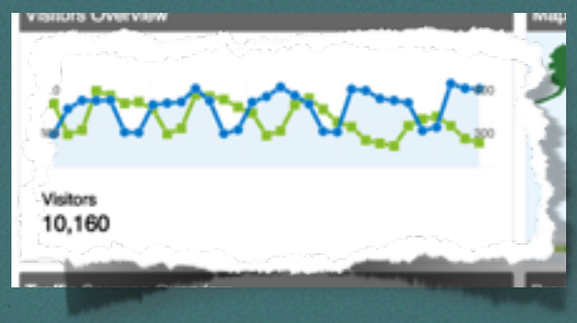
Lou Kratz
lou@curalate.com

CURALATE

People Communicate Visually

Brain recognizes
objects in 100ms

Increasingly popular on
social networks



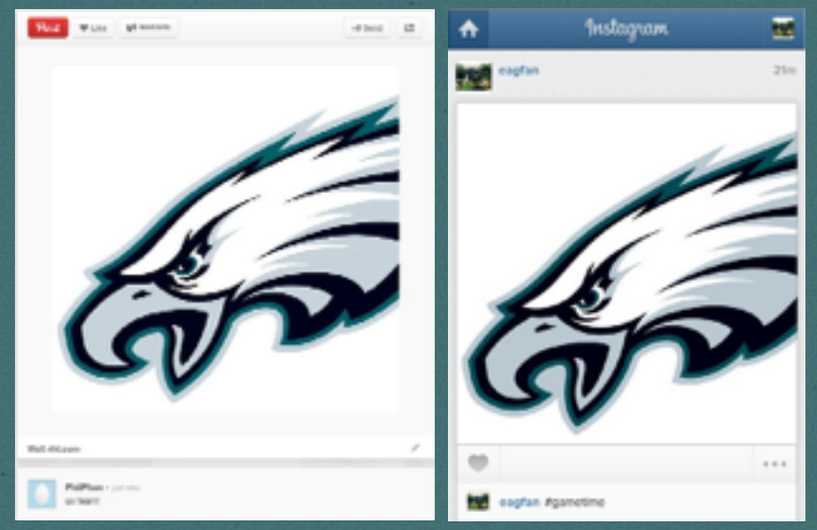
De facto choice
for analytics

Images are an efficient and
effective way to communicate.

Keyword Analytics

Visual Analytics

Data



Aggregate `count("eagles")`

?

Analyze

?

The image is the conversation.

CURALATE

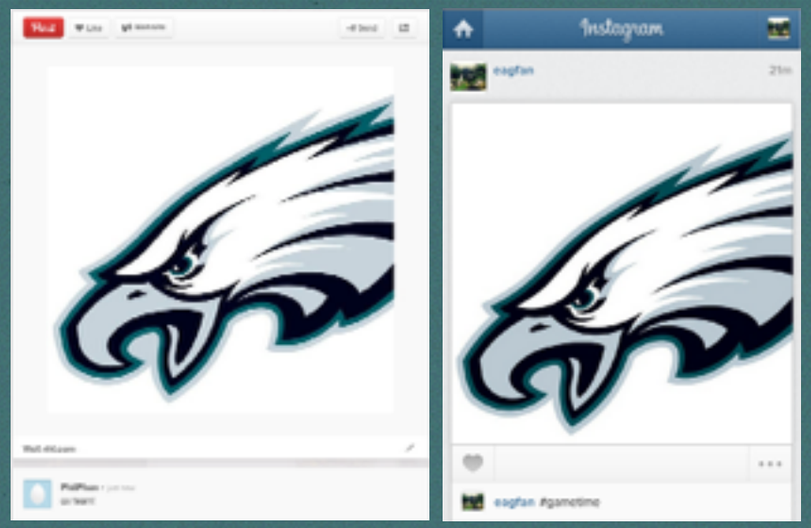
- Philadelphia Startup
- Marketing suite for visual web
- Visual analytics for social networks

Myself

- **lou@curalate.com**
- Software Engineer
- Machine Learning, Computer Vision, Tech Transfer

Visual Analytics

Data



Aggregate

Image Matching?

- Accurate
- Efficient
- Scalable

Image Matching

Are these the same image?

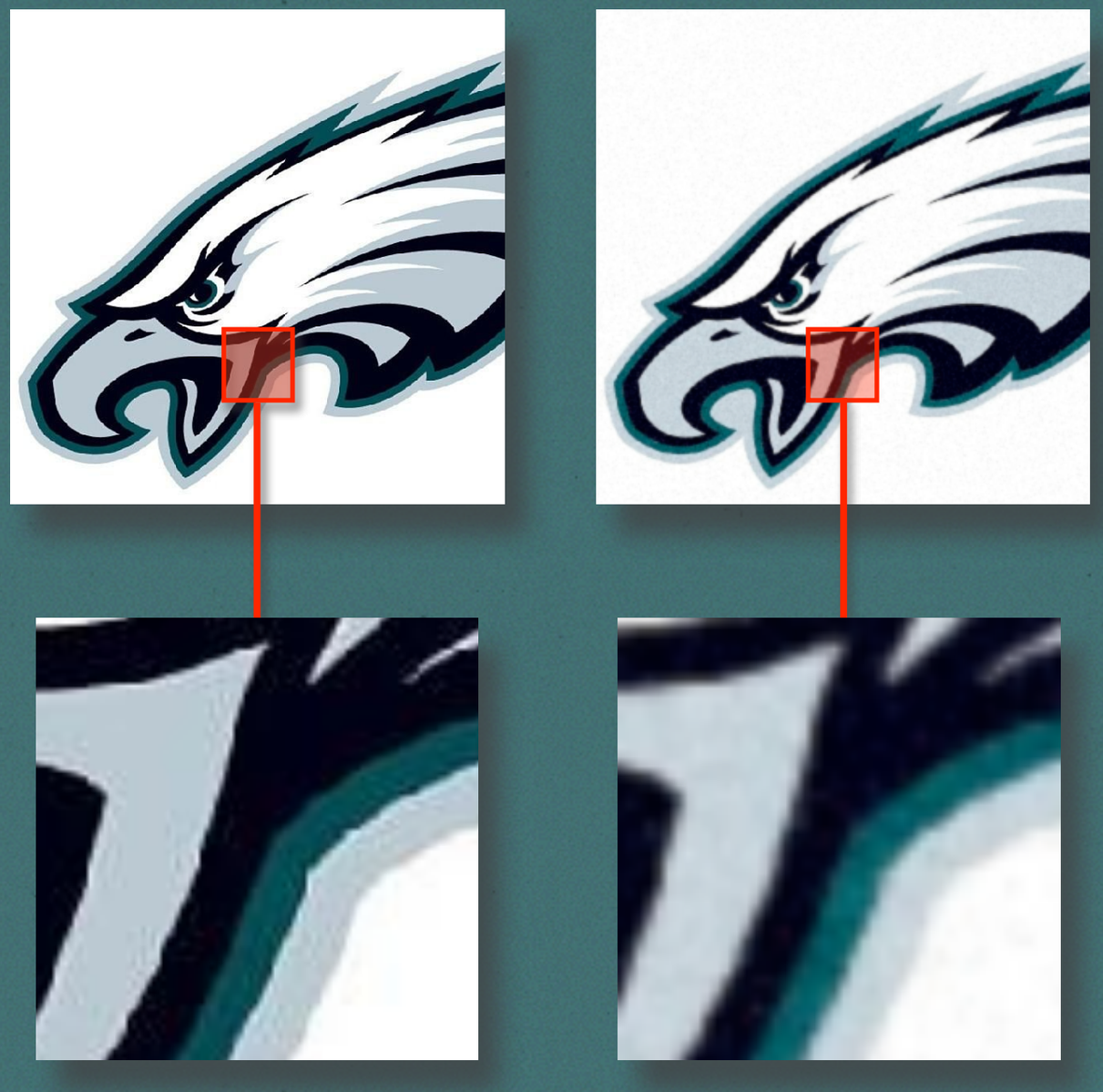


Image Matching

Are these the same image?

Difference

- Compression
- Resolution
- Processing Time*

Efficient Matching

- **Image Fingerprinting**
  - Small numeric representation
  - Fast to compute
  - Represent as much information as possible

Image

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right]$$

Discrete Cosine Transform

Binary Encoding

64 bit integer

Fingerprint

Key Idea: Identical images have identical fingerprints

Relaxing Constraint

Bits can differ due to noise or compression

64 bits

fp1

64 bits

fp2

Threshold on hamming distance

```
BIT_COUNT( XOR(fp1, fp2) ) <= 2
```

Image Matching

Image

Fingerprint

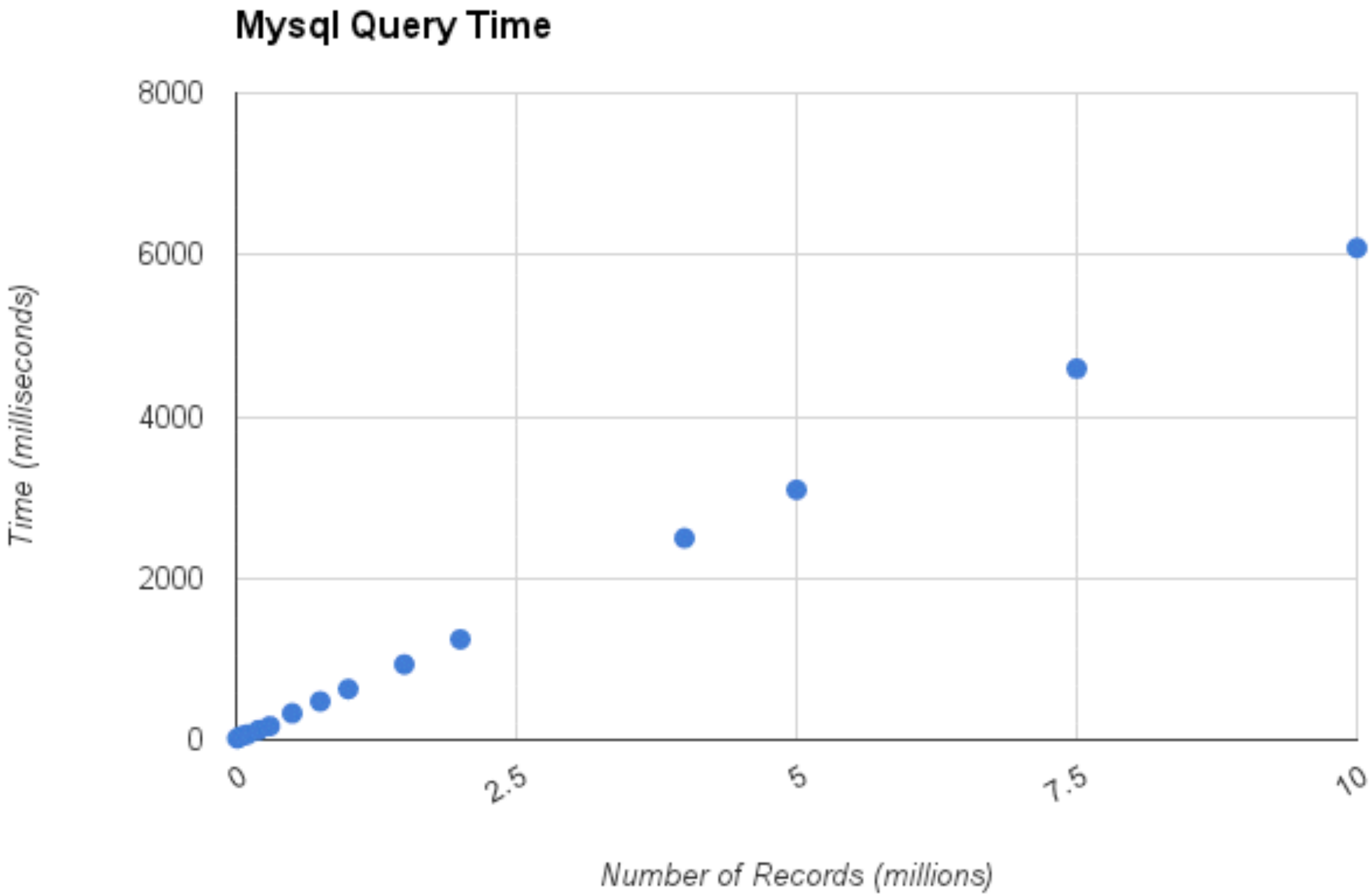
Search

Database

Fingerprint	File
32432521355	Eagles-Logo.jpg
745907134	NFL-Philly.jpg

Matching is now a search problem

Mysql Query Time



Option 2: NoSQL

Fingerprint

64 bits

Number of reads $= \sum_{k=0}^{r} \binom{b}{k}$

- b: number of bits in fingerprint (64)
- r: distance threshold (2)

For 64 bits and r=2, *2081* reads!

NoSql

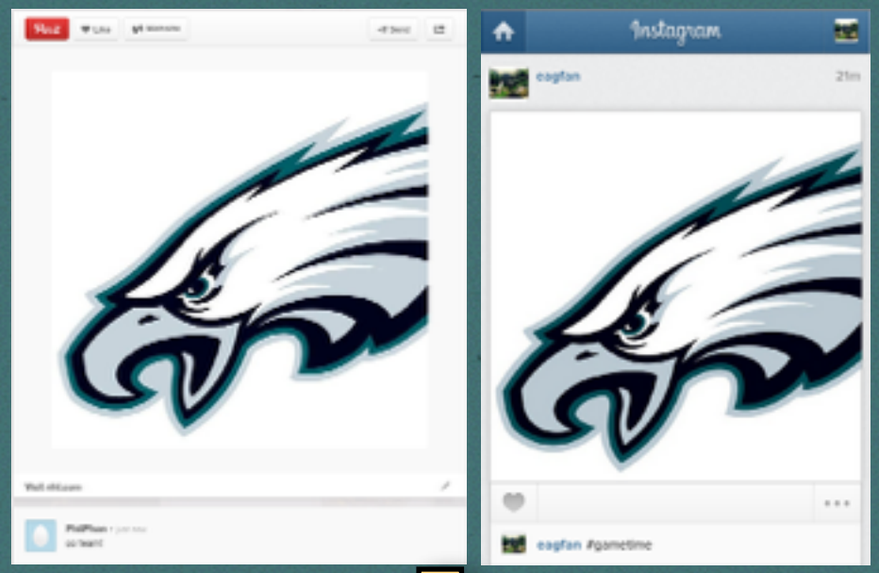
- 2081 reads for an image lookup
  - substantially reduced with optimization
- Fast (Reads~10ms)
- Scalable: Hosted solutions are distributed
- Number of images only minutely affects search speed

Our System

- Deployed in August of 2013
- Webservice using Amazon Elasticbeanstalk
- Currently:
  - 126 million URLS
  - 116 million Files
  - 86 million unique images
  - 600 images matched per second

Visual Analytics

Data



Aggregate

Image Matching

Analyze

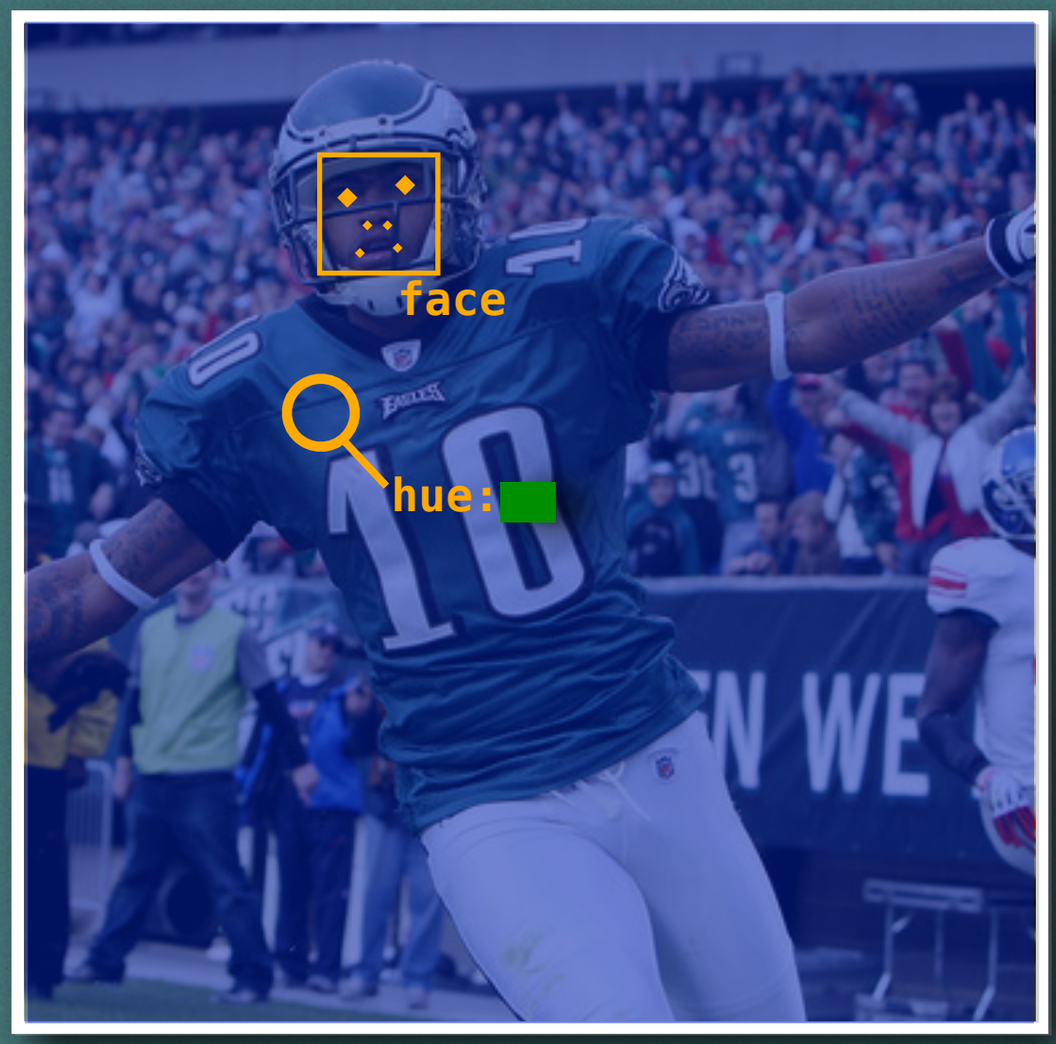
?

Identical Images

Different Images

Image Fetures

- Dominant Hue (color)
- Face Detection
- Texture
- Many more!



Color

BLUEISH IMAGE

vs.

REDDISH-ORANGE IMAGE

Images that are reddish-orange get roughly **TWICE AS MANY** repins than images that are blue.

Number of Colors

VS.

MULTIPLE DOMINANT COLORS

SINGLE DOMINANT COLOR

Images with multiple dominant colors have **3.25 TIMES MORE** repins than single dominant color images.

Lightness

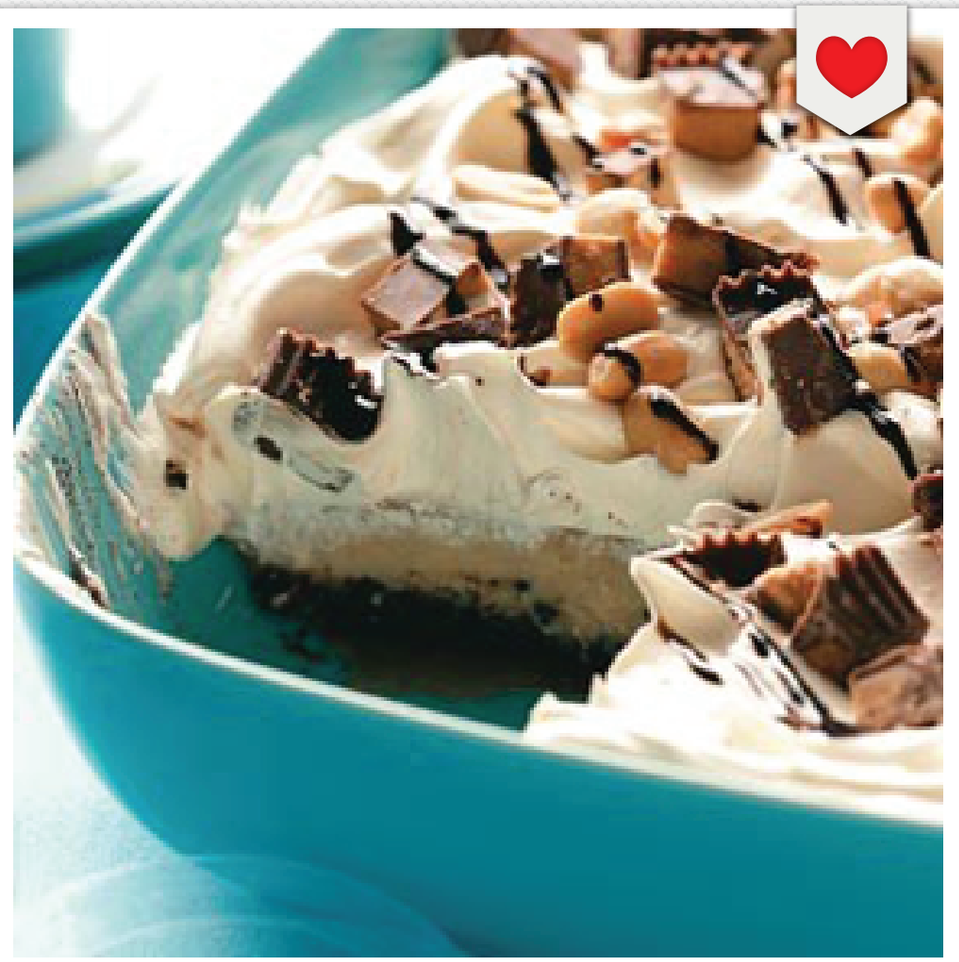
VERY DARK

VS.

MEDIUM LIGHTNESS

Images with medium lightness are repinned **20 TIMES MORE** than very dark images.

Saturation



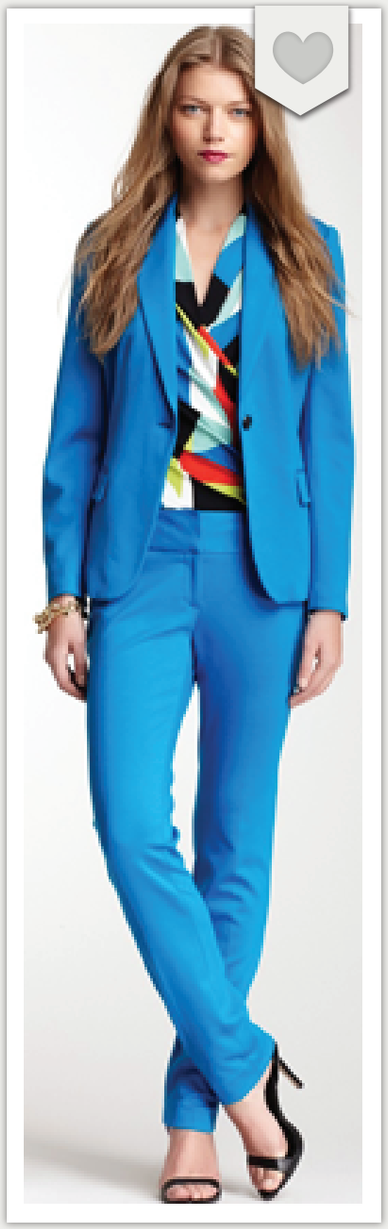
50% SATURATION

VS.

VERY DESATURATED

Images that are 50% saturated have **10 TIMES MORE** repins than very desaturated images.

Aspect Ratio



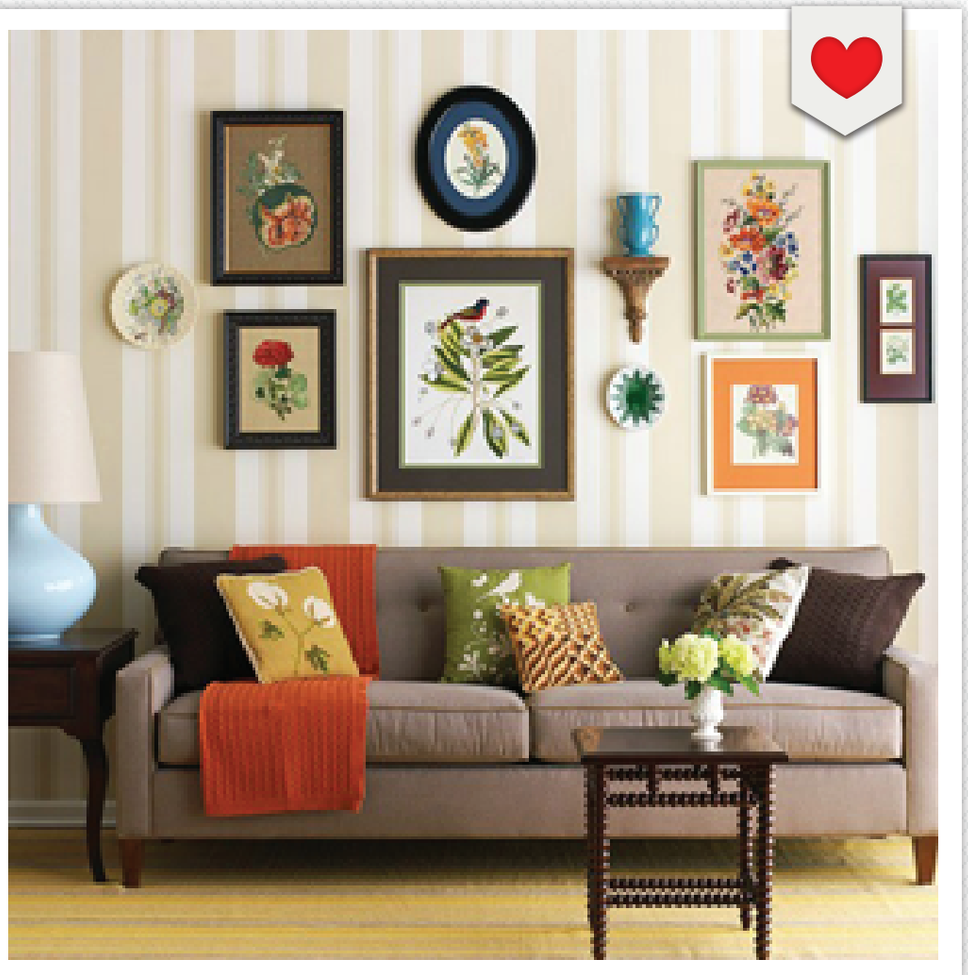
VERY TALL

vs.

2:3 ASPECT RATIO

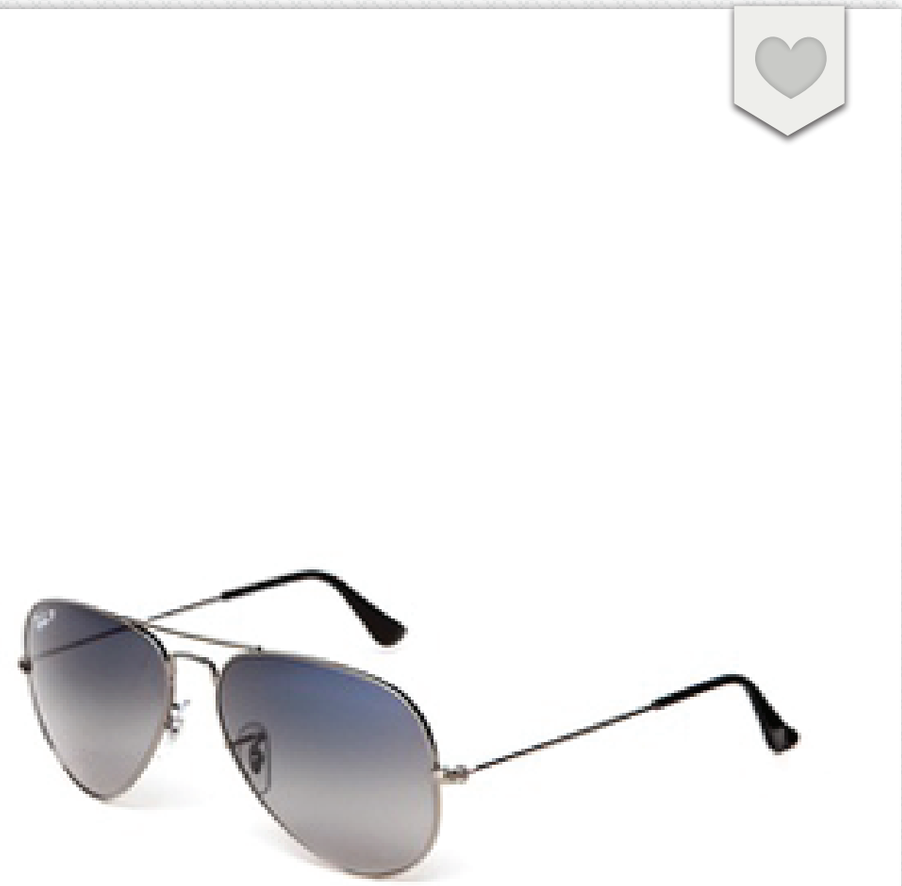
Vertical images with an aspect ratio between 2:3 and 4:5 get **60% MORE** repins than very tall images.

Background



LESS THAN 10% BACKGROUND

VS.



MORE THAN 90% BACKGROUND

Images with <10% background receive **2–4 TIMES MORE** repins than images with >40% background.

Texture

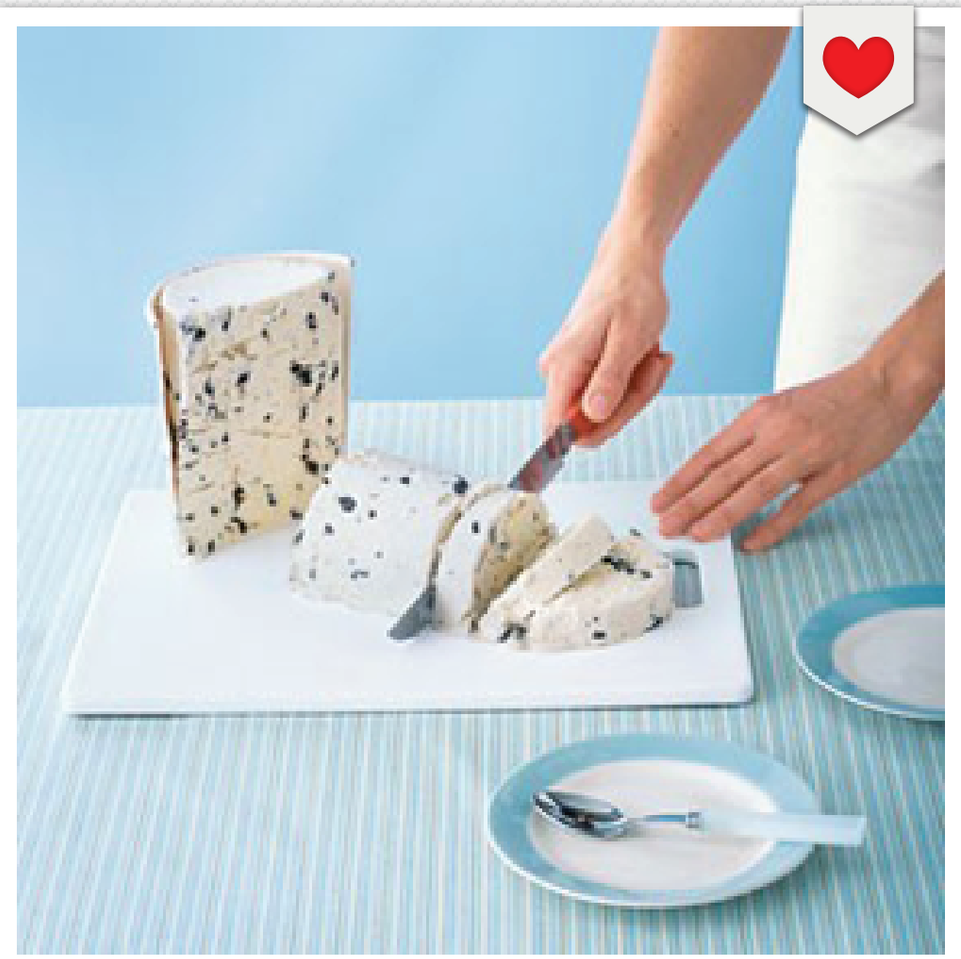
“ROUGH” IMAGE

VS.

“SMOOTH” IMAGE

Images with a smooth texture are repinned **17 TIMES MORE** than images with a rough texture.

Faces



DOESN'T HAVE A FACE

VS.

HAS A FACE

Brand images without faces receive **23% MORE** repins than images with faces.

Conclusions

- Visual communications are important
- Visual Analytics are awesome
- Each step requires novel techniques – new paradigm for analytics

lou@curalate.com

CURALATE