

Data Visualization 1: Introduction

Stat 133 by Gaston Sanchez

Creative Commons Attribution Share-Alike 4.0 International CC BY-SA

Why Visualization?

GSW Per Game Statistics (2017-2018)

Rk		Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%
1	Klay Thompson	27	73	73	34.3	7.9	16.1	.488	3.1	7.1	.440	4.7	9.0	.526
2	Kevin Durant	29	68	68	34.2	9.3	18.0	.516	2.5	6.1	.419	6.7	11.9	.565
3	Draymond Green	27	70	70	32.7	4.0	8.8	.454	1.1	3.7	.301	2.9	5.2	.562
4	Stephen Curry	29	51	51	32.0	8.4	16.9	.495	4.2	9.8	.423	4.2	7.1	.595
5	Andre Iguodala	34	64	7	25.3	2.3	5.0	.463	0.5	1.8	.282	1.8	3.2	.567
6	Quinn Cook	24	33	18	22.4	3.7	7.6	.484	1.4	3.2	.442	2.3	4.5	.514
7	Nick Young	32	80	8	17.4	2.5	6.1	.412	1.5	4.1	.377	1.0	2.0	.481
8	Patrick McCaw	22	57	10	16.9	1.6	3.9	.409	0.3	1.4	.238	1.3	2.5	.503
9	Shaun Livingston	32	71	7	15.9	2.4	4.8	.501	0.0	0.1	.000	2.4	4.8	.509
10	Jordan Bell	23	57	13	14.2	2.0	3.2	.627	0.0	0.1	.000	2.0	3.2	.641
11	Zaza Pachulia	33	69	57	14.1	2.2	3.8	.564	0.0	0.0	.000	2.2	3.8	.567
12	Omri Casspi	29	53	7	14.0	2.3	3.9	.580	0.2	0.4	.455	2.1	3.5	.595
13	Kevon Looney	21	66	4	13.8	1.7	2.9	.580	0.0	0.1	.200	1.7	2.8	.590
14	David West	37	73	0	13.7	3.0	5.2	.571	0.0	0.1	.375	2.9	5.1	.576
15	JaVale McGee	30	65	17	9.5	2.1	3.4	.621	0.0	0.1	.000	2.1	3.3	.638
16	Damian Jones	22	15	0	5.9	0.7	1.5	.500	0.0	0.0		0.7	1.5	.500
17	Chris Boucher	25	1	0	1.0	0.0	1.0	.000	0.0	1.0	.000	0.0	0.0	

Quick questions

How many players in GSW roster?

Age of youngest player? (oldest player?)

Name of youngest player? (oldest player?)

Relationship between 3P and 2P?

Scored the most Field Goals (FG) per Minutes Played (MP)?

Paraphrasing the old saying

An **image** is worth a
thousand numbers

dataset 1		dataset 2		dataset 3		dataset 4	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

dataset `anscombe` (available in R)

What things would you like to calculate for each variable?

x1	x2	x3	x4
Min. : 4.0	Min. : 4.0	Min. : 4.0	Min. : 8
1st Qu.: 6.5	1st Qu.: 6.5	1st Qu.: 6.5	1st Qu.: 8
Median : 9.0	Median : 9.0	Median : 9.0	Median : 8
Mean : 9.0	Mean : 9.0	Mean : 9.0	Mean : 9
3rd Qu.:11.5	3rd Qu.:11.5	3rd Qu.:11.5	3rd Qu.: 8
Max. :14.0	Max. :14.0	Max. :14.0	Max. :19

y1	y2	y3	y4
Min. : 4.260	Min. :3.100	Min. : 5.39	Min. : 5.250
1st Qu.: 6.315	1st Qu.:6.695	1st Qu.: 6.25	1st Qu.: 6.170
Median : 7.580	Median :8.140	Median : 7.11	Median : 7.040
Mean : 7.501	Mean :7.501	Mean : 7.50	Mean : 7.501
3rd Qu.: 8.570	3rd Qu.:8.950	3rd Qu.: 7.98	3rd Qu.: 8.190
Max. :10.840	Max. :9.260	Max. :12.74	Max. :12.500

Motivation

Mean of **x** values: 9.0

Mean of **y** values: 7.5

Least Squares equation: **$y = 3 + 0.5x$**

Sum of squared errors: 110

Correlation coefficient: 0.816

Using only numerical
reduction methods in
data analysis is far too
limiting

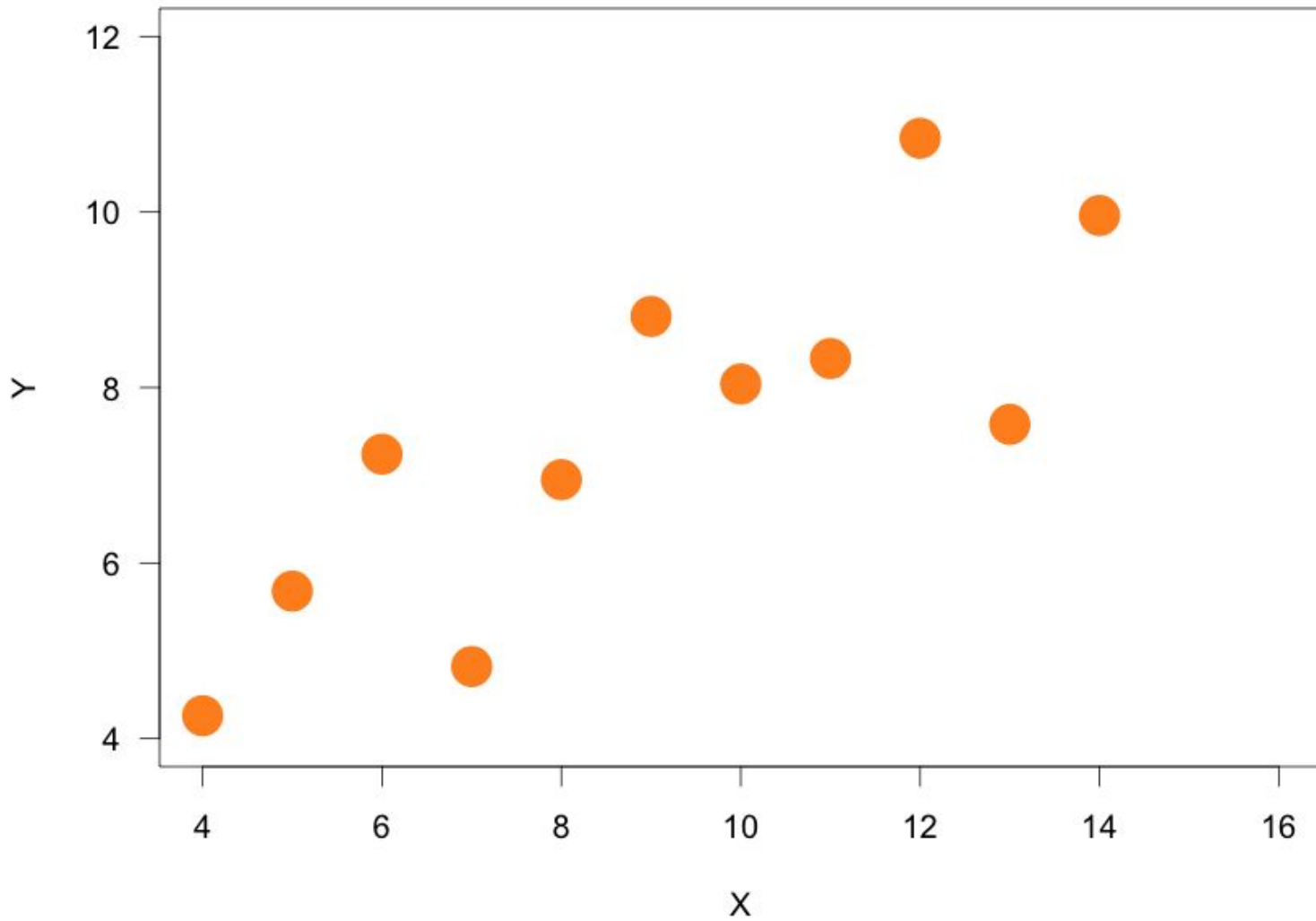
Are you able to see any patterns, relations?

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

X	Y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

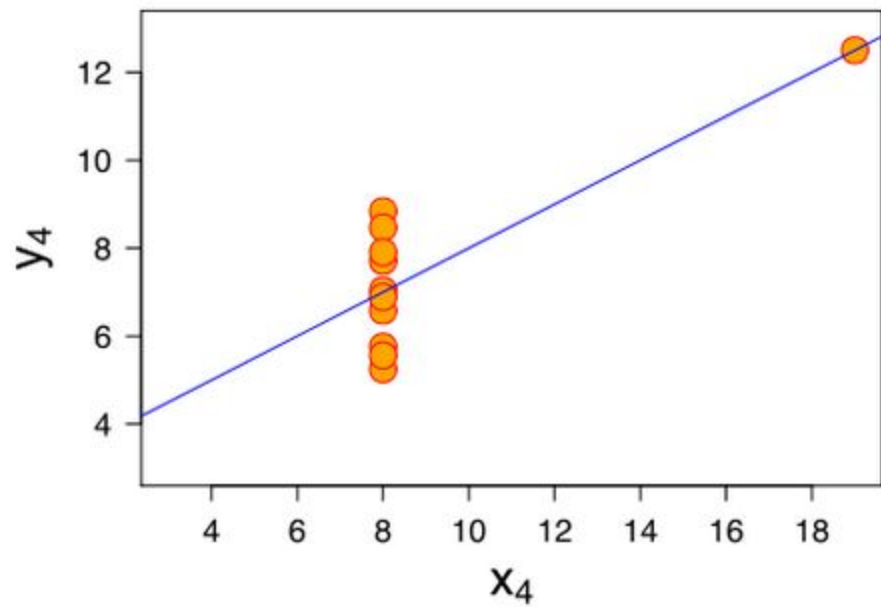
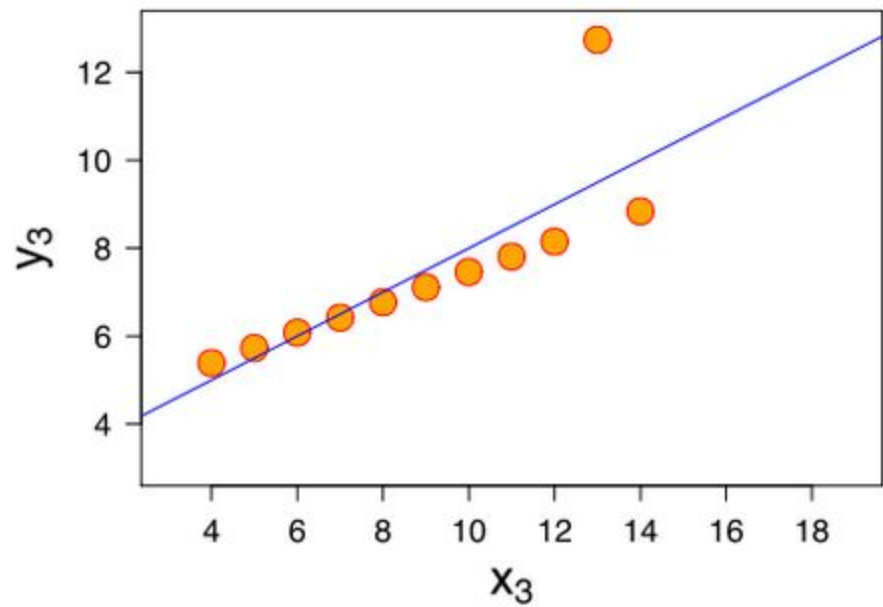
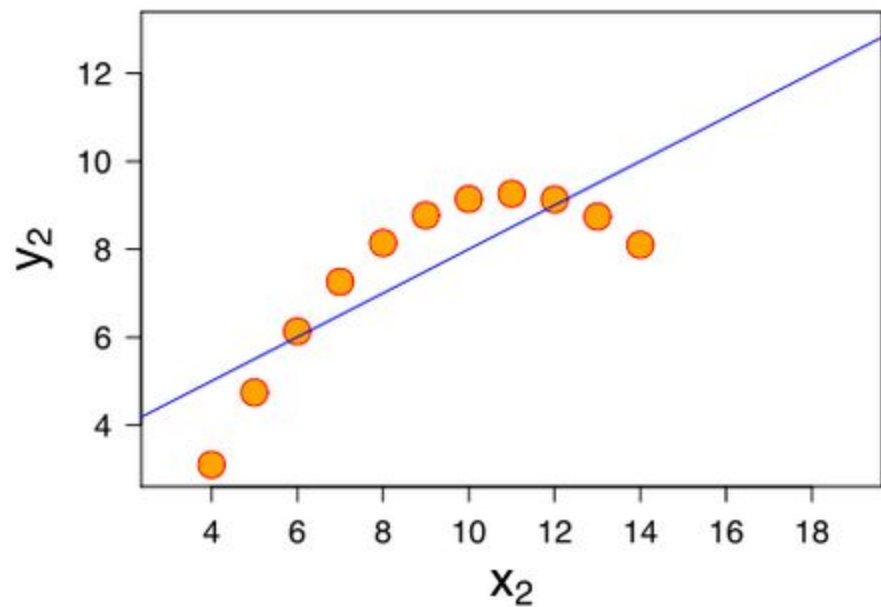
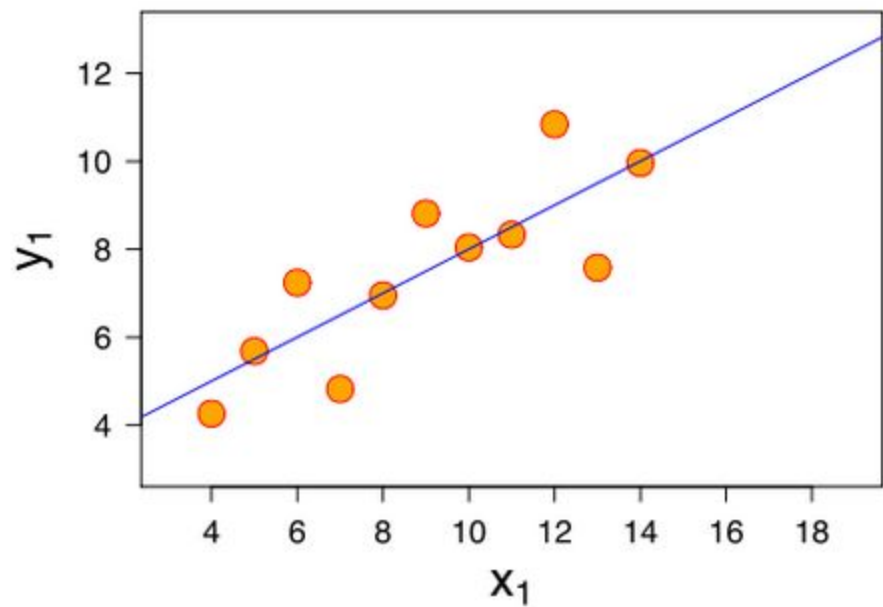
How **X** and **Y**
are related?

Scatter Diagram



Are you able to see any patterns, relations?

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Why Visualization?

Our brain is an exquisite change detector and pattern recognizing system.

But our brains evolved in a much simpler world with far less information coming at us.

Our brains have not been wired to perform mental calculations with “large” sets of numbers.

Why Visualization?

“Our modes of thinking and decision making evolved over the tens of thousands of years that humans lived as hunter-gathers.


Our genes haven't fully caught up with the demands of modern civilization, but fortunately human knowledge has.”

Daniel Levitin

Why Data Visualization?



A key component of
computing with data
consists of
data visualization



“Visualization provides insight that cannot be appreciated by any other approach to learning from data.”

William S. Cleveland

Why Visualization?

Data visualization, in the form of graphics, is mostly visual.

Understanding visual perception is fundamental to design better visual displays.

Vision, of our all senses, is the most powerful and efficient **channel for receiving information** from the physical world.



Around **70%** of the
body's receptors
reside in our **eyes**

Human Vision

Our eyes are not very good at making sense when looking at (many) numbers

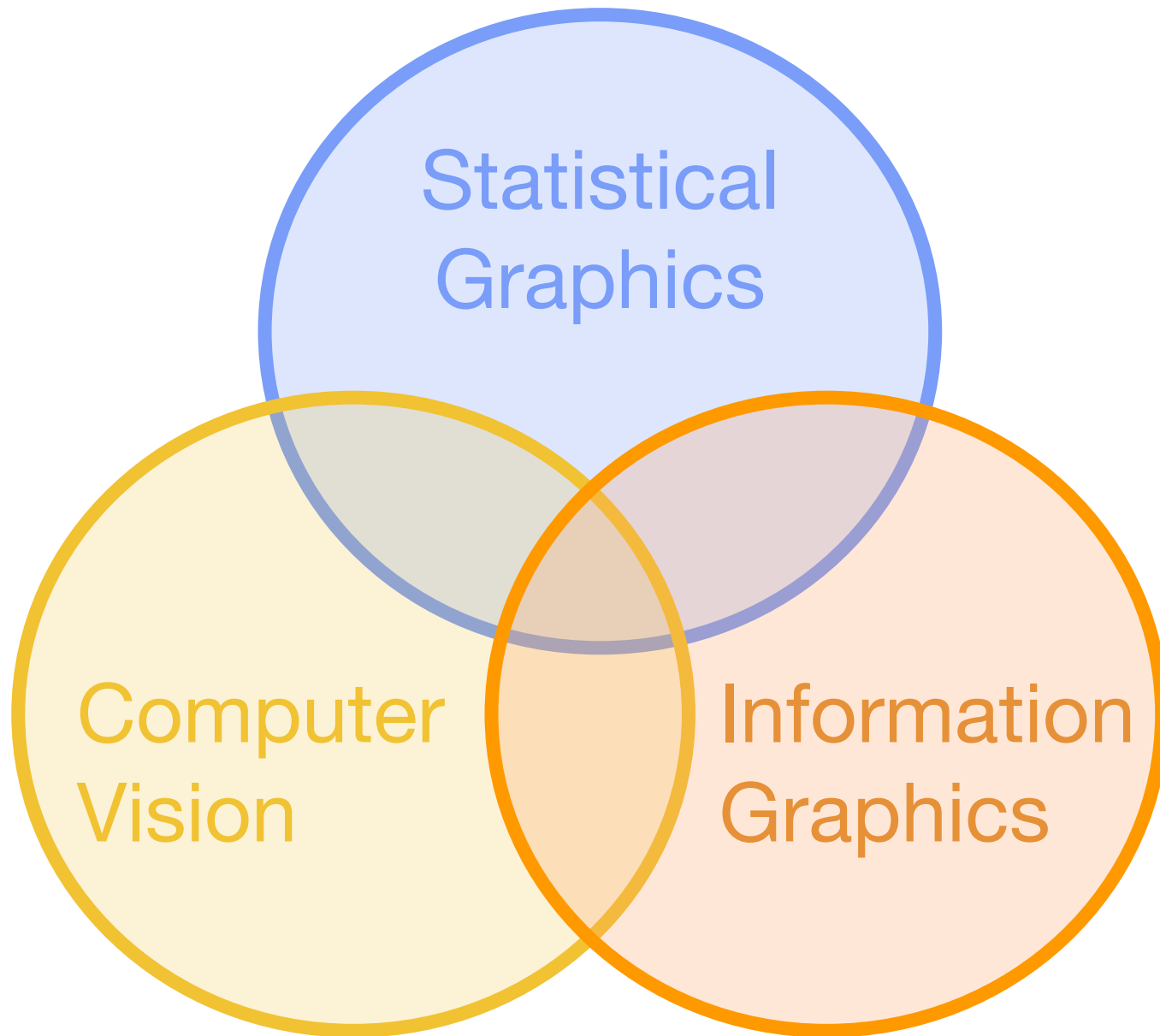
But they are great for looking at shapes and detecting patterns

About Data Visualization



data visualization





Data Visualization ...

Statistical Graphics?

Computer Graphics?

Computer Vision?

Infographics?

Data Art?

The Africa opportunity

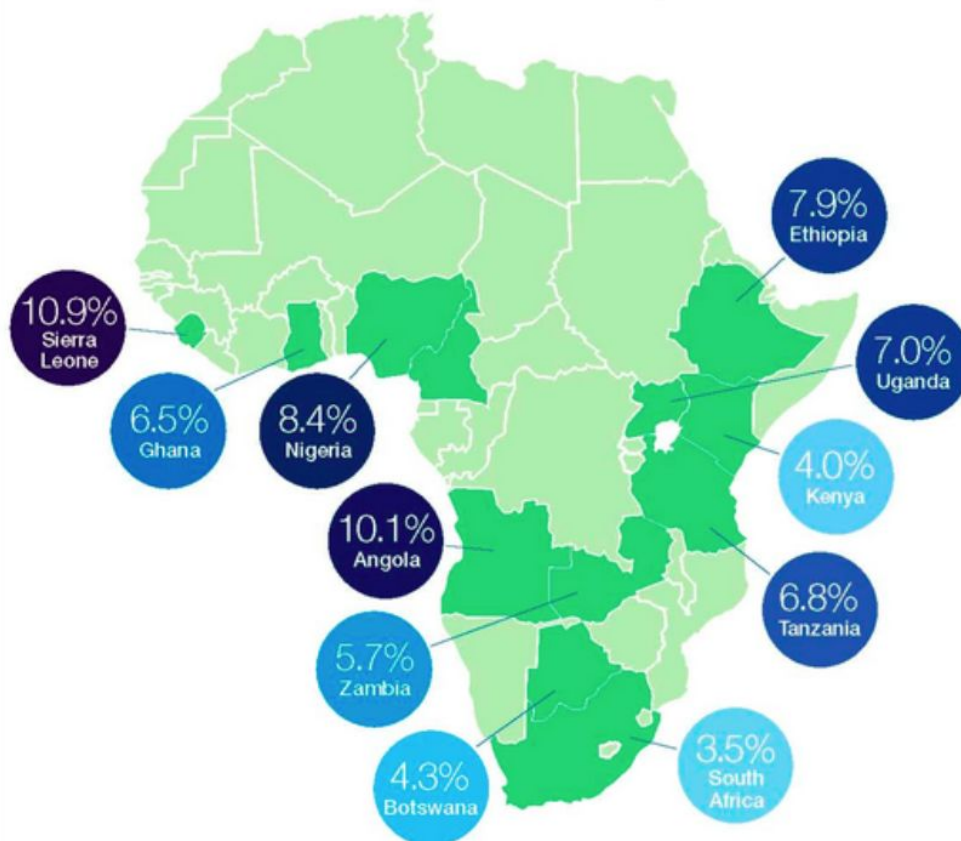


FLIGHTS TO
AFRICA
UP 85%
BETWEEN
2005-2011

MORE

MOBILE PHONE
SUBSCRIBERS
IN AFRICA
THAN EUROPE

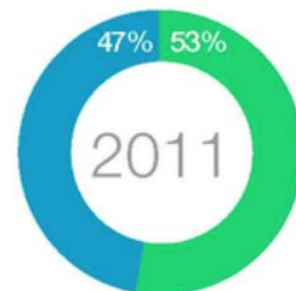
Sub-Saharan Africa average annual GDP growth, 2000-2012



Sub-Saharan Africa's trading partners

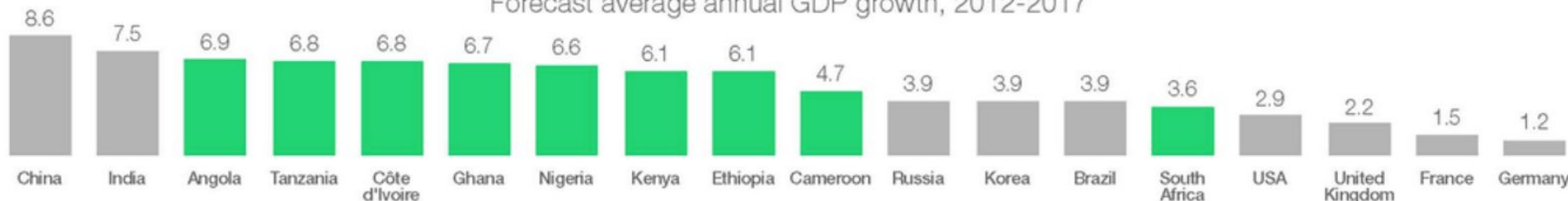


Sub-Saharan Africa's total world trade: US\$169bn

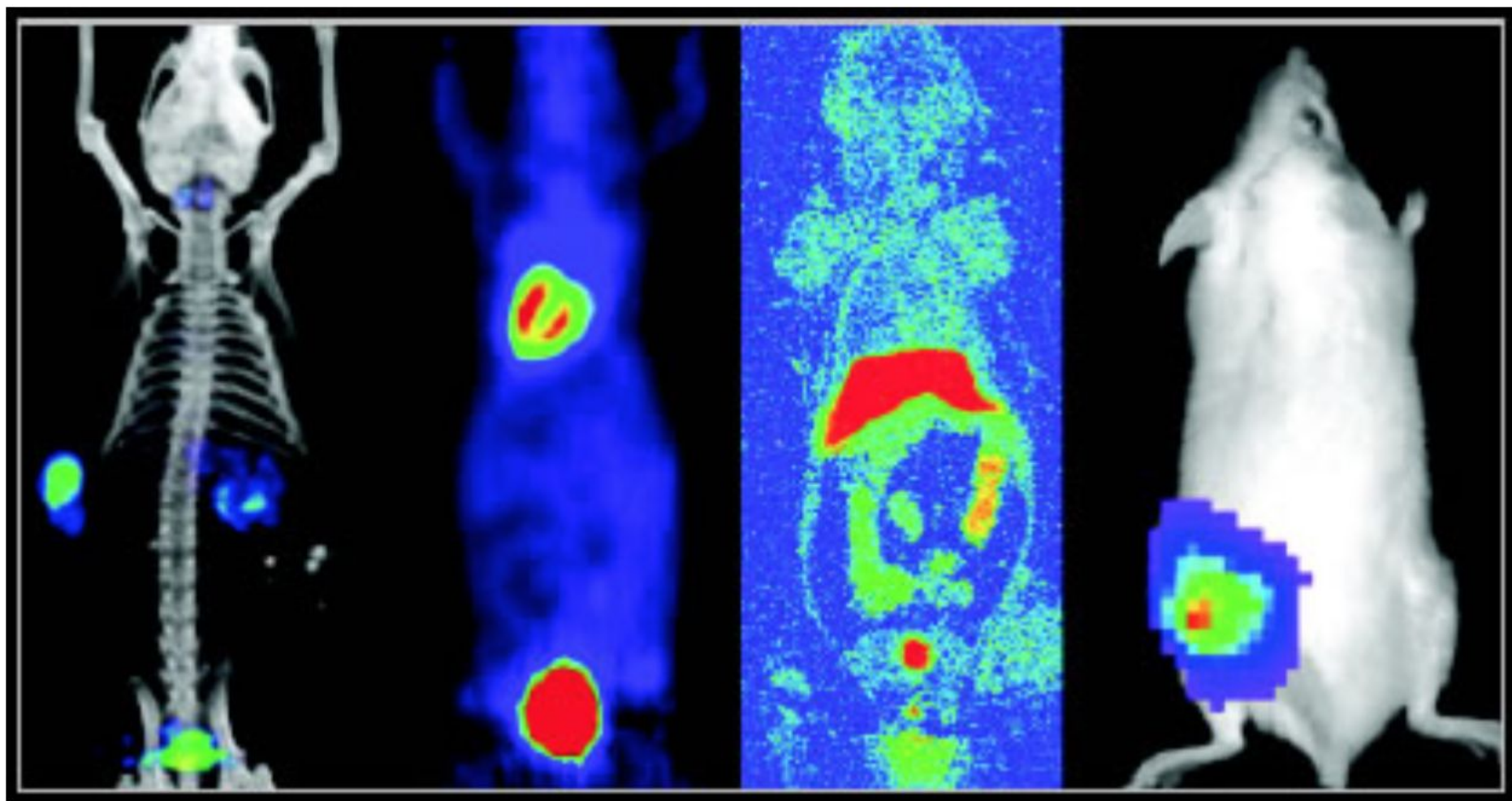


Sub-Saharan Africa's total world trade: US\$735bn

Forecast average annual GDP growth, 2012-2017



Sources: International Monetary Fund, World Economic Outlook Database, April 2012; International Monetary Fund Direction of Trade Statistics



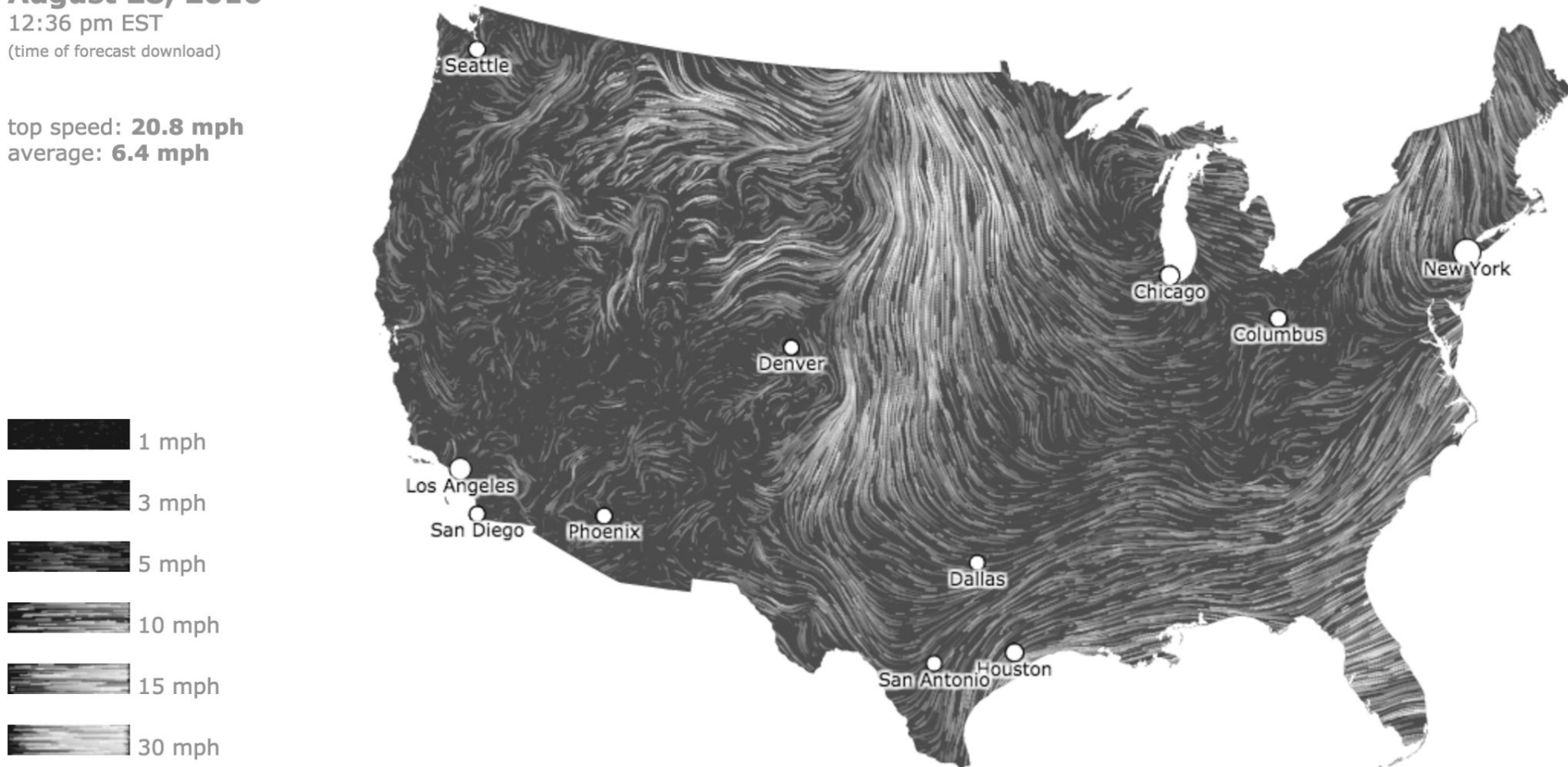
August 28, 2016

12:36 pm EST

(time of forecast download)

top speed: **20.8 mph**

average: **6.4 mph**



<http://hint.fm/wind/>


Visualization Continuum

Statistical
Graphics

Data
Art

Facts


Entertainment



“There's value in entertaining, putting a smile on someone's face, and making people feel something, as much as there is in optimized presentation.”

Nathan Yau, 2013

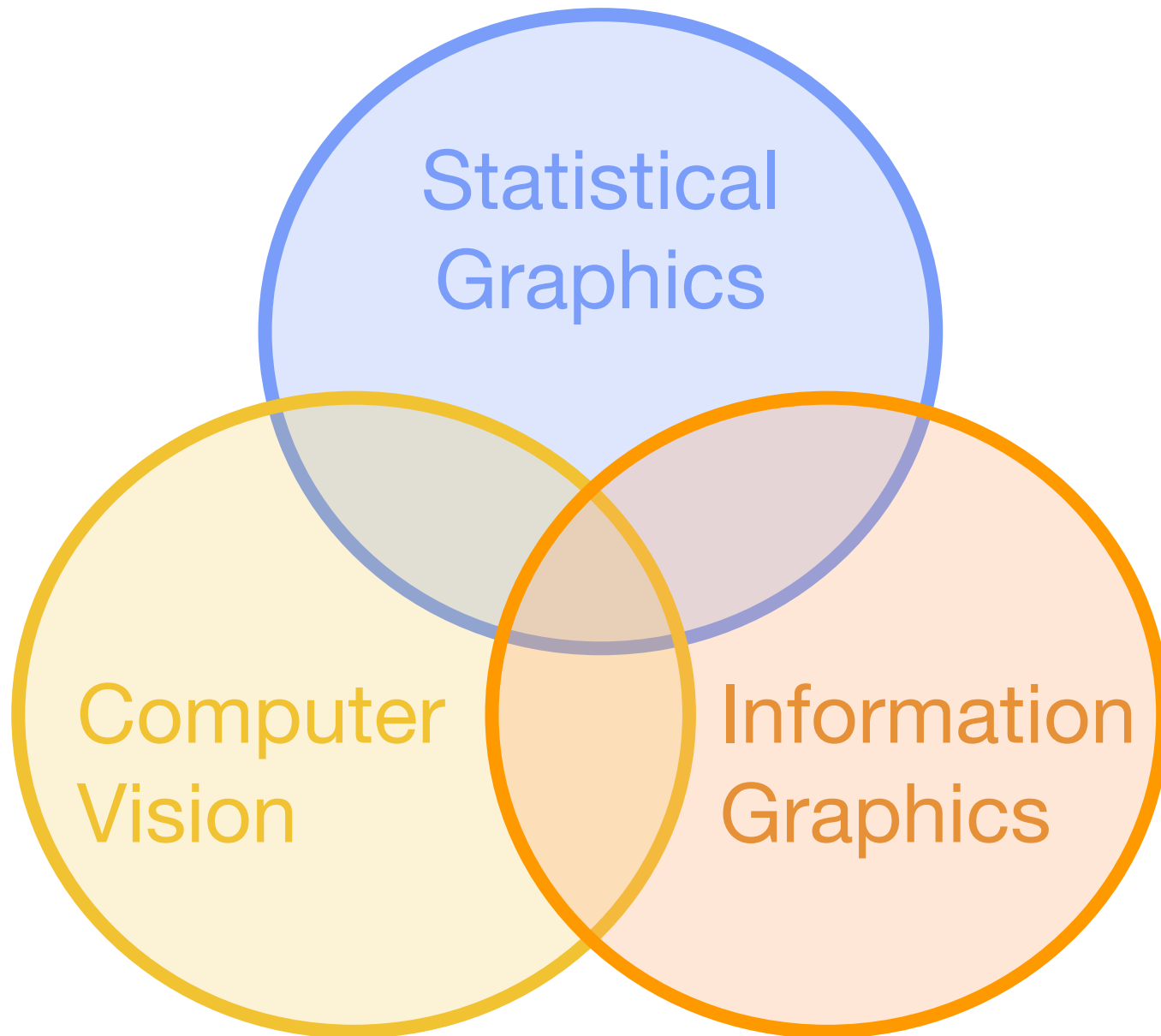
(Data Points, p 69)



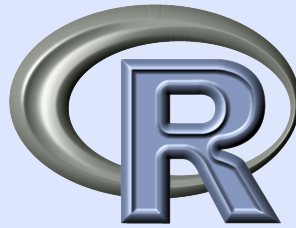
“Data Art: visualizations that strive to entertain or to create aesthetic experiences with little concern for informing.”

Stephen Few, 2012

Statistical Graphics



Statistical Graphics



Why Visualization?

Visualizing data is critical to data analysis

Graphs allow us to see overall patterns and to see detailed behavior

Graphs allow us to view complex mathematical models fitted to data

Things commonly said about statistical graphics

“The data should stand out”

“Story telling”

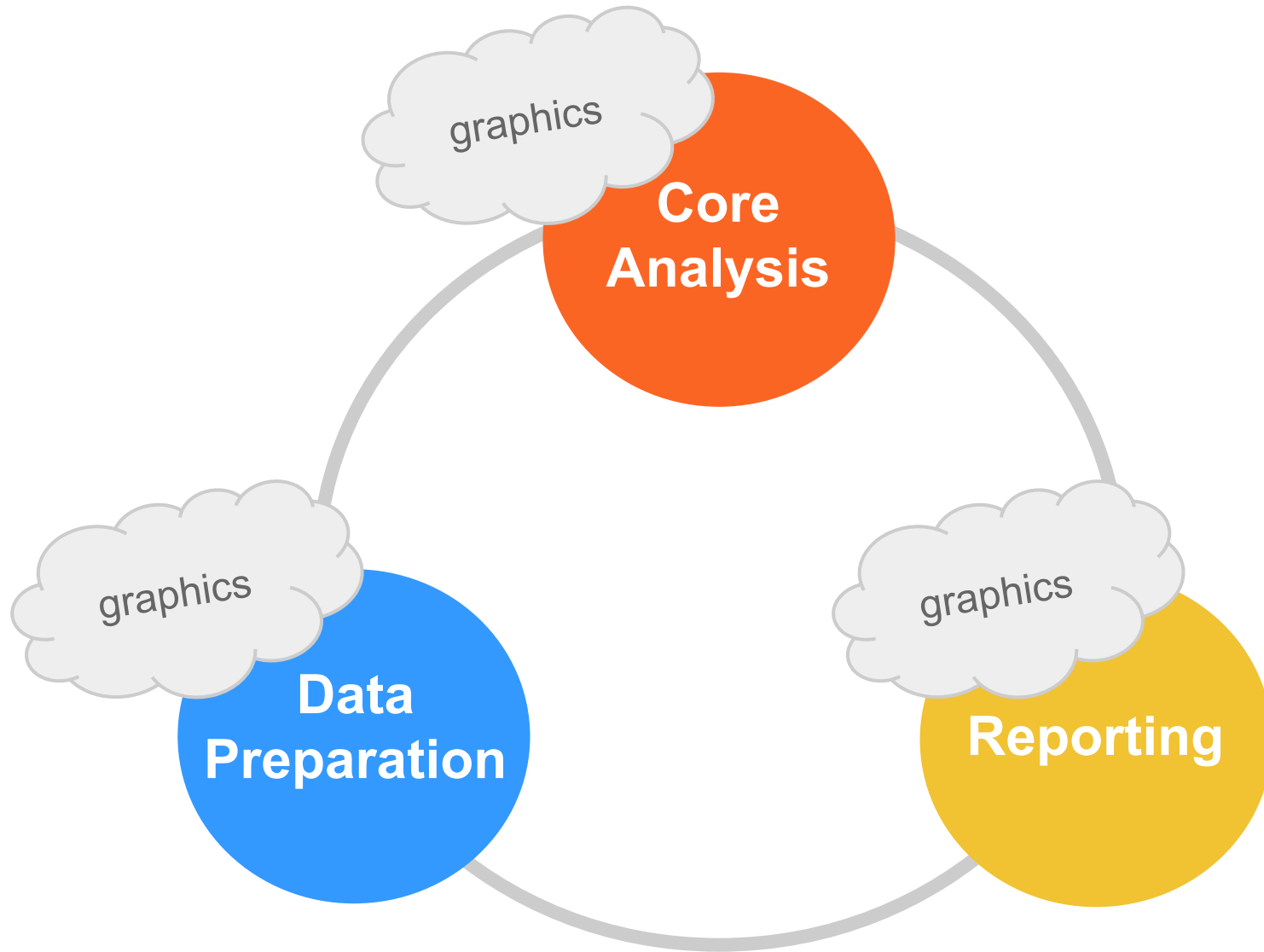
“Big picture”

“The purpose of visualization is insight, not pictures”

We'll focus on statistical
graphics and visual
displays of data in
science and technology

Graphics all over the DAC

Data Analysis Cycle (DAC)

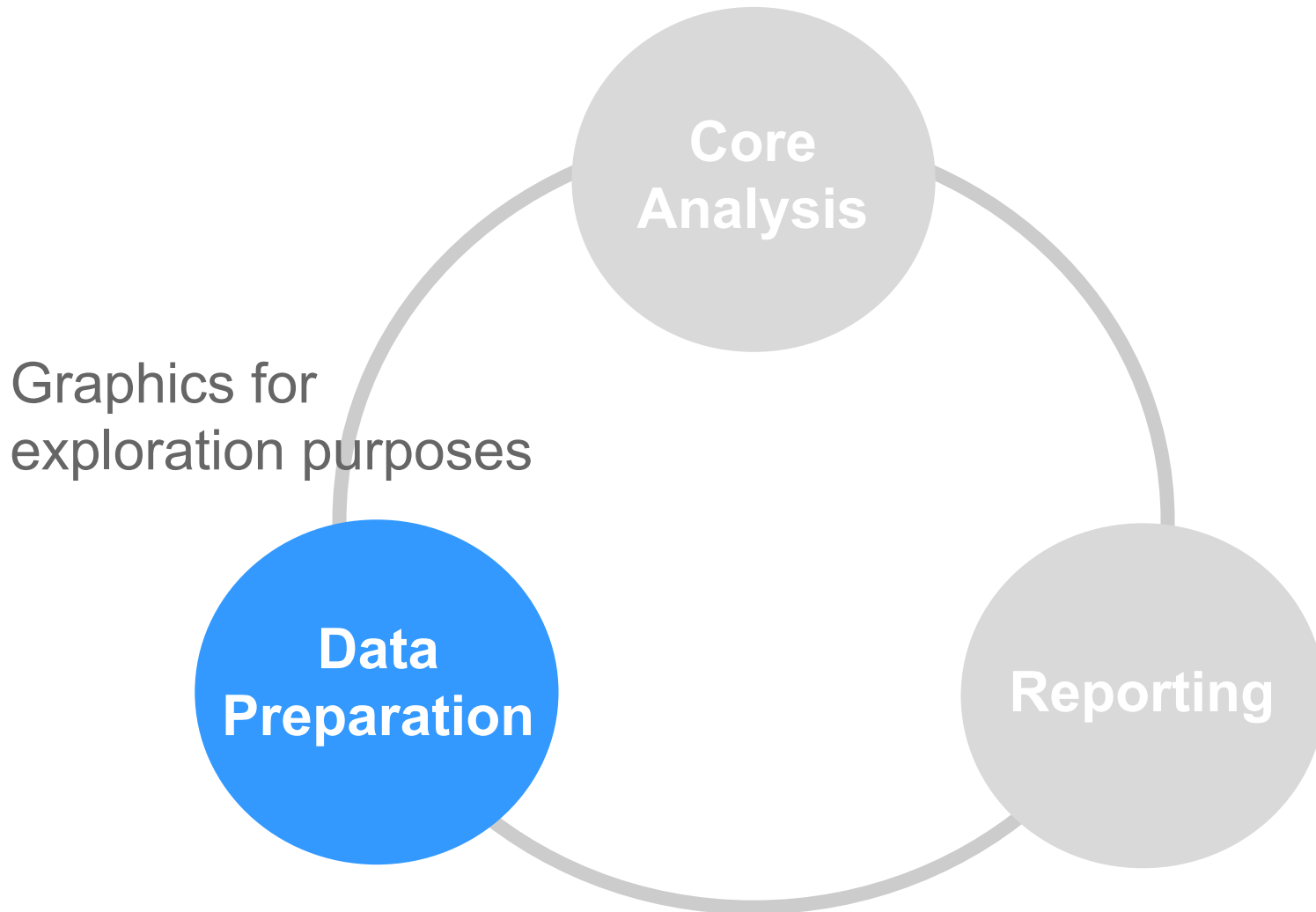


Graphics for

Exploration

Communication

Data Analysis Cycle (DAC)



Graphics for Exploration

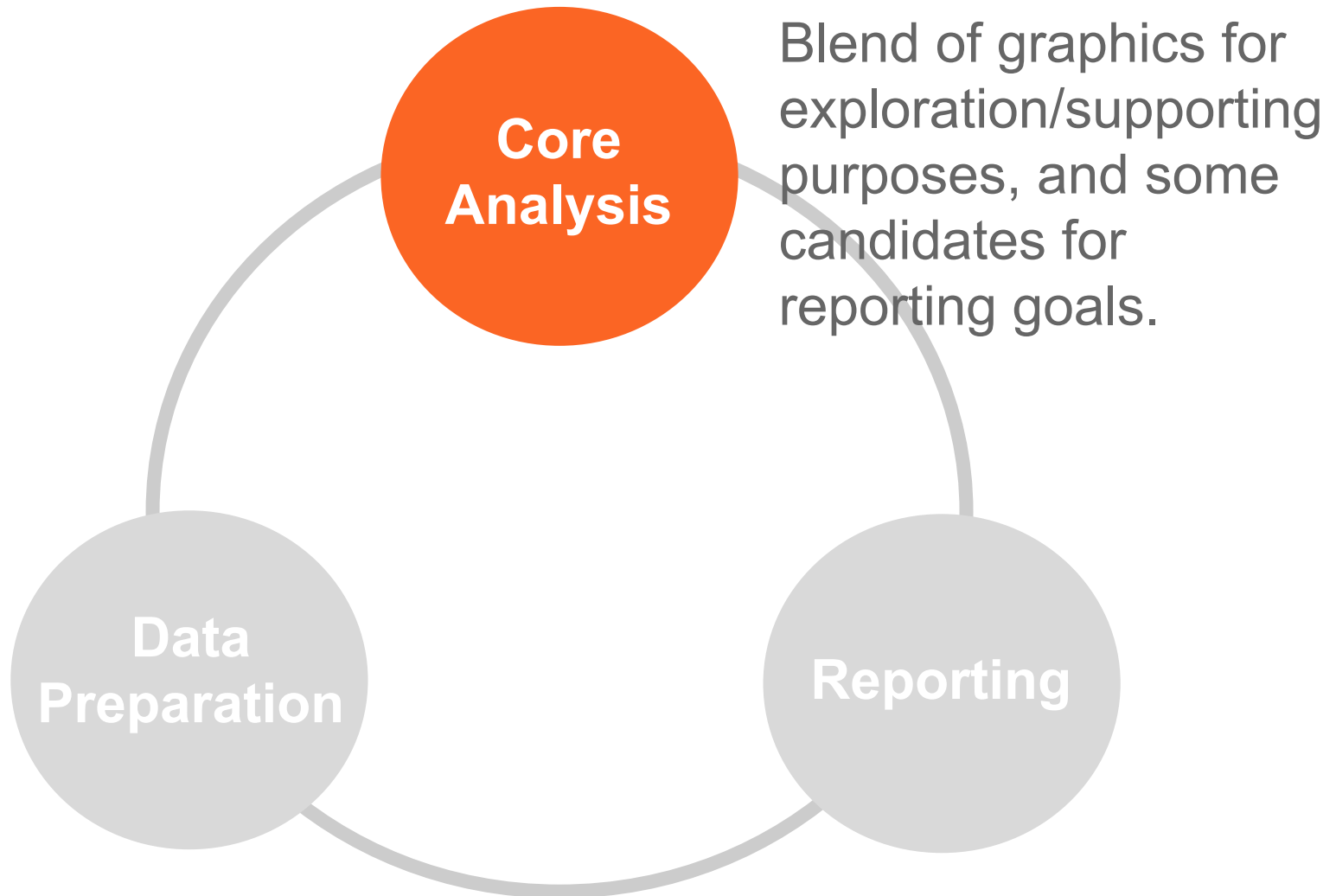
Graphics for verifying/understanding data

The analyst is the main (and usually only) consumer

Typically quick & dirty (less care about visual appearance and design principles)

Lifespan of a few seconds

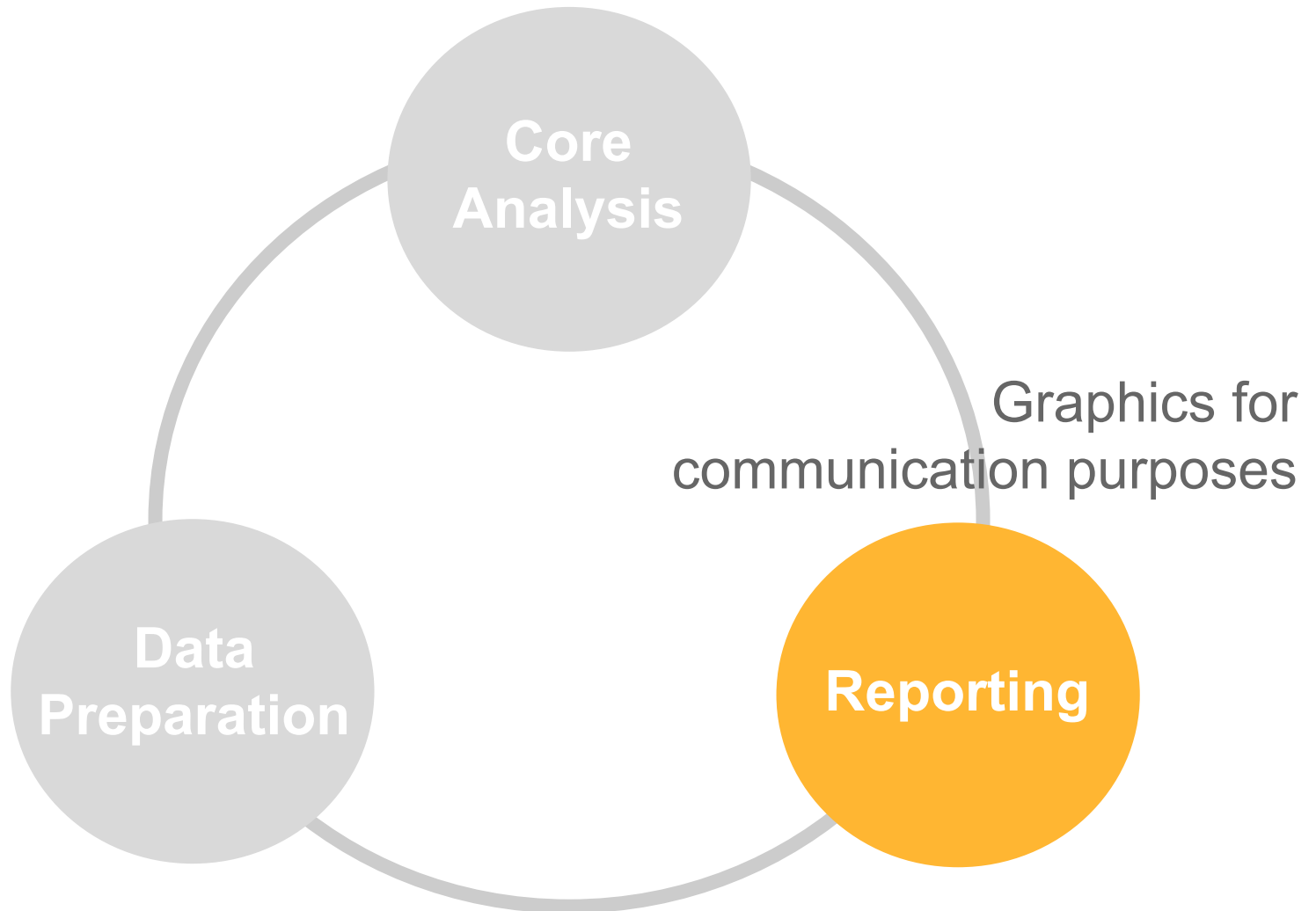
Data Analysis Cycle (DAC)



Graphics for Exploration/Supporting

Graphs produced in the “core analysis” stage will be a mix of exploration/supporting, and some candidates that can make it to the reporting phase.

Data Analysis Cycle (DAC)



Graphics for Communication

Graphics for presenting data

To be consumed by others

Must care about visual appearance and design principles

Require a lot of iterations in order to get the final version

What's the message? Who is the audience?



Stats

Stats Home

Players

Teams

Advanced

Scores

Schedule

Hustle Stats

SEASON
2016-17

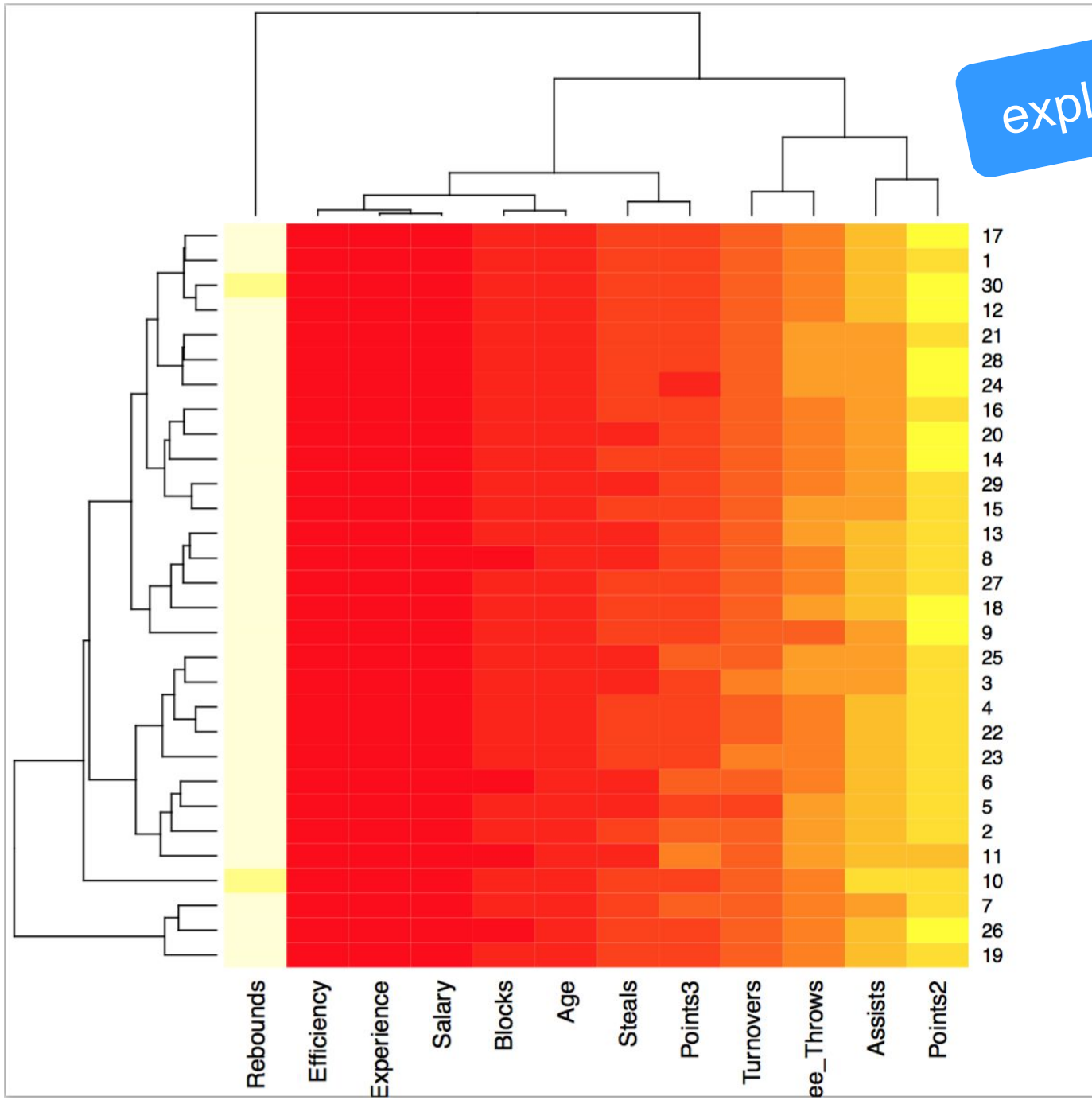
SEASON TYPE
Regular Season

PER MODE
Per Game

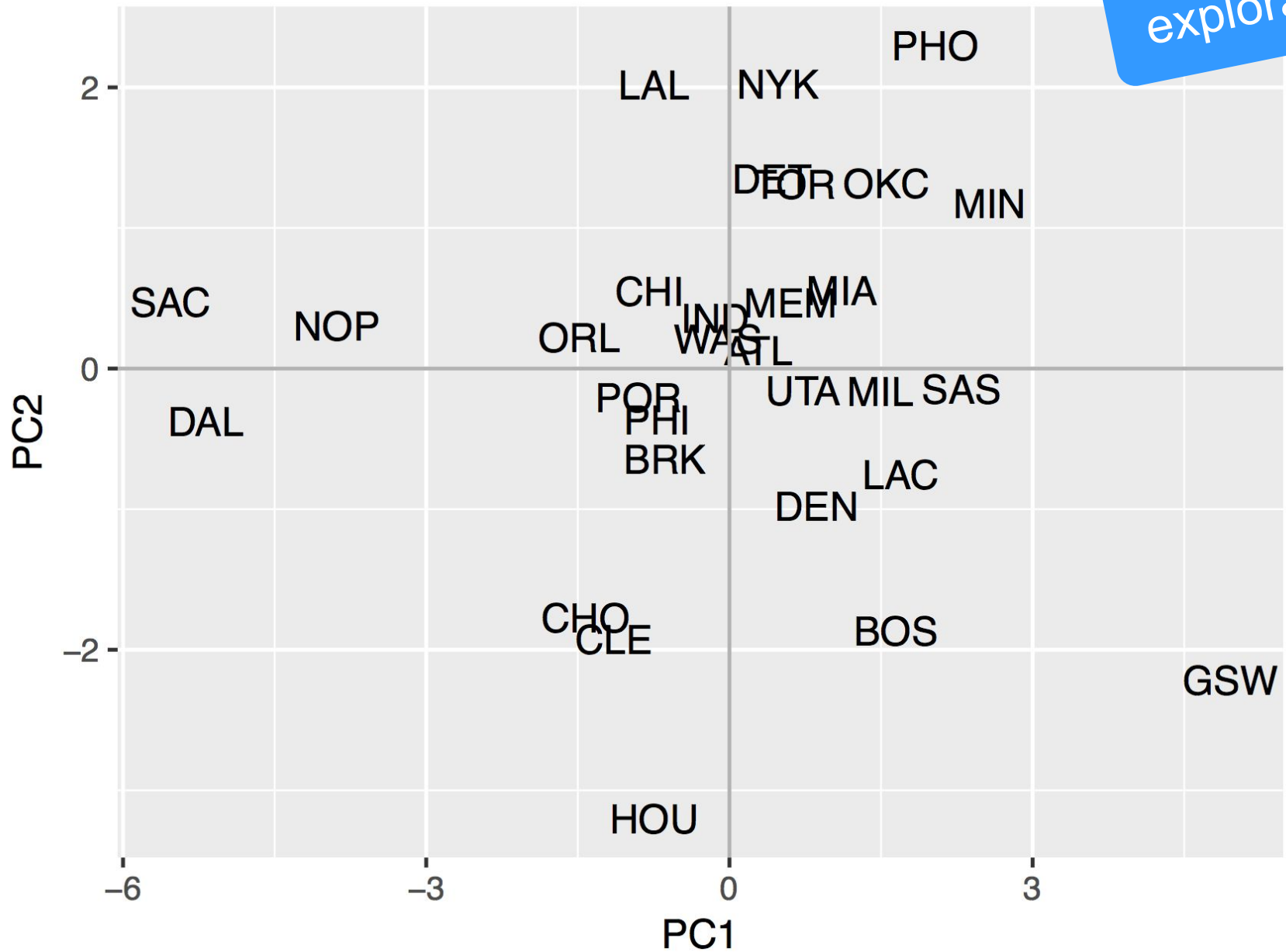
SEASON SEGMENT
All Games

TEAM	GP	W	L	WIN%	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB
1 Miami Heat	82	41	41	.500	48.2	103.2	39.0	85.8	45.5	9.9	27.0	36.5	15.2	21.6	70.6	10.6	33.0
1 Atlanta Hawks	82	43	39	.524	48.5	103.2	38.1	84.4	45.1	8.9	26.1	34.1	18.1	24.9	72.8	10.3	34.1
1 Brooklyn Nets	82	20	62	.244	48.2	105.8	37.8	85.2	44.4	10.7	31.6	33.8	19.4	24.6	78.8	8.8	35.1
1 Charlotte Hornets	82	36	46	.439	48.4	104.9	37.7	85.4	44.2	10.0	28.6	35.1	19.4	23.8	81.5	8.8	34.8
1 Chicago Bulls	82	41	41	.500	48.2	102.9	38.6	87.1	44.4	7.6	22.3	34.0	18.0	22.5	79.8	12.2	34.1
1 Cleveland Cavaliers	82	51	31	.622	48.5	110.3	39.9	84.9	47.0	13.0	33.9	38.4	17.5	23.3	74.8	9.3	34.4
1 Dallas Mavericks	82	33	49	.402	48.2	97.9	36.2	82.3	44.0	10.7	30.2	35.5	14.8	18.5	80.1	7.9	30.7
1 Denver Nuggets	82	40	42	.488	48.2	111.7	41.2	87.7	46.9	10.6	28.8	36.8	18.7	24.2	77.4	11.8	34.6
1 Detroit Pistons	82	37	45	.451	48.3	101.3	39.9	88.8	44.9	7.7	23.4	33.0	13.9	19.3	71.9	11.1	34.6
1 Golden State Warriors	82	67	15	.817	48.2	115.9	43.1	87.1	49.5	12.0	31.2	38.3	17.8	22.6	78.8	9.4	35.0

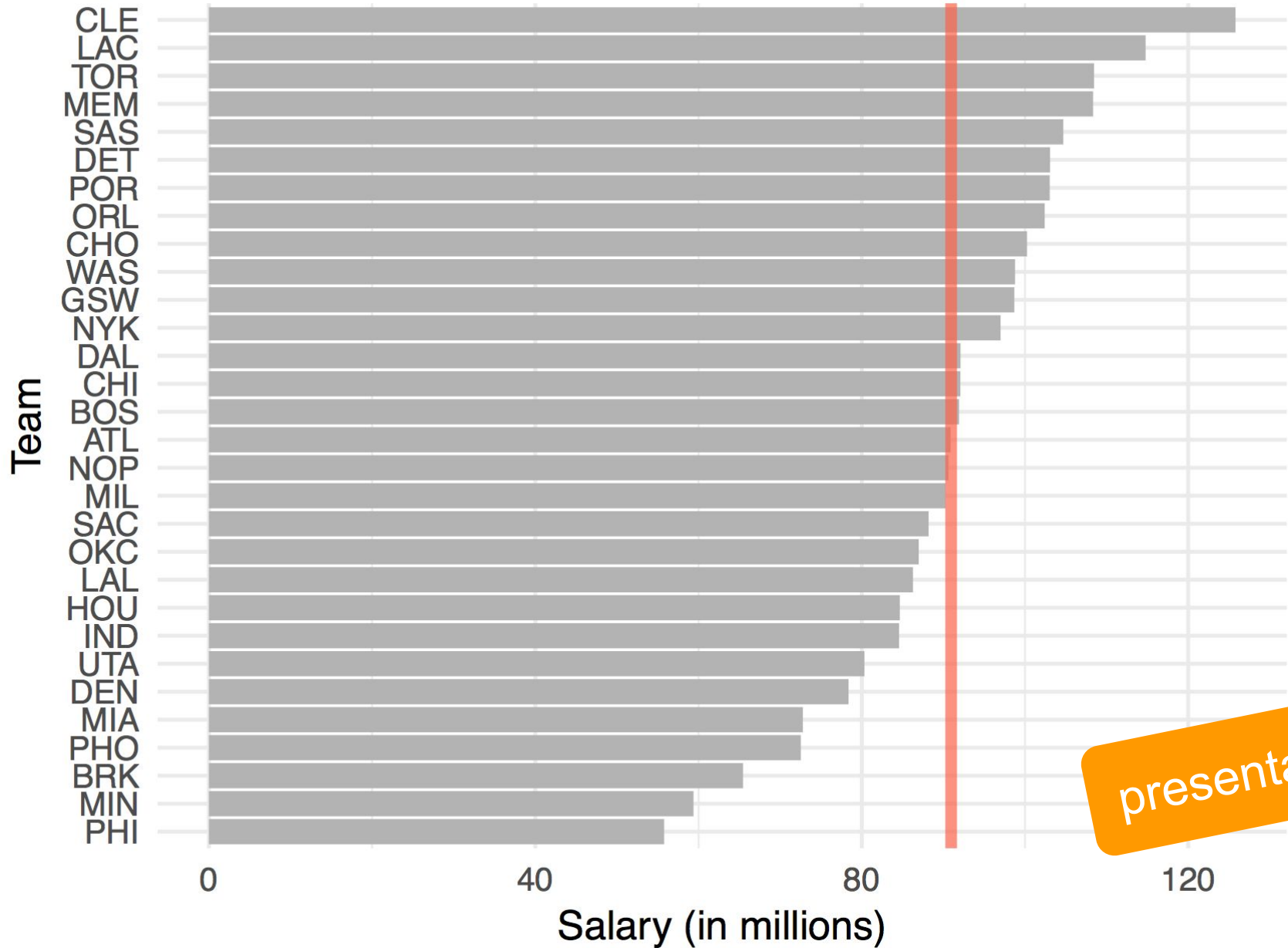
exploration



PCA plot (PC1 and PC2)



NBA Teams ranked by Total Salary



presentation