Indexing and Filtering

Paweł Kordek
SOFTWARE ENGINEER

@pawel_kordek https://kordek.github.io

Overview

Typical questions

Selecting the right data

Dos and don’ts

The Task

We Would Like to Know…

How many distinct artists are there in the dataset?

How many artworks by Francis Bacon are there?

What is the artwork with the biggest dimensions?

The Basics

df[‘artist’]

	artist	artistId	...
1023	Blake, William	39	...
1045	Bacon, Francis	682	...
1087	Bacon, Francis	682	...
1099	Bacon, Francis	682	...
1011	Blake, William	39	...

	artist
1023	Blake, William
1045	Bacon, Francis
1087	Bacon, Francis
1099	Bacon, Francis
1011	Blake, William

df[[‘artist’, ‘artistId’]]

	artist	artistId	...
1023	Blake, William	39	...
1045	Bacon, Francis	682	...
1087	Bacon, Francis	682	...
1099	Bacon, Francis	682	...
1011	Blake, William	39	...

	artist	artistId
1023	Blake, William	39
1045	Bacon, Francis	682
1087	Bacon, Francis	682
1099	Bacon, Francis	682
1011	Blake, William	39

Avoid This

df[‘artist’]

~~df.artist~~

Demo

We will:

- Count the number of distinct artists in the dataset

Filtering

```
df[‘artist’] == ‘Bacon, Francis’
```

	artist	artistId	...
1023	Blake, William	39	...
1045	Bacon, Francis	682	...
1087	Bacon, Francis	682	...
1099	Bacon, Francis	682	...
1011	Blake, William	39	...

	artist
1023	False
1045	True
1087	True
1099	True
1011	False

Demo

We will:

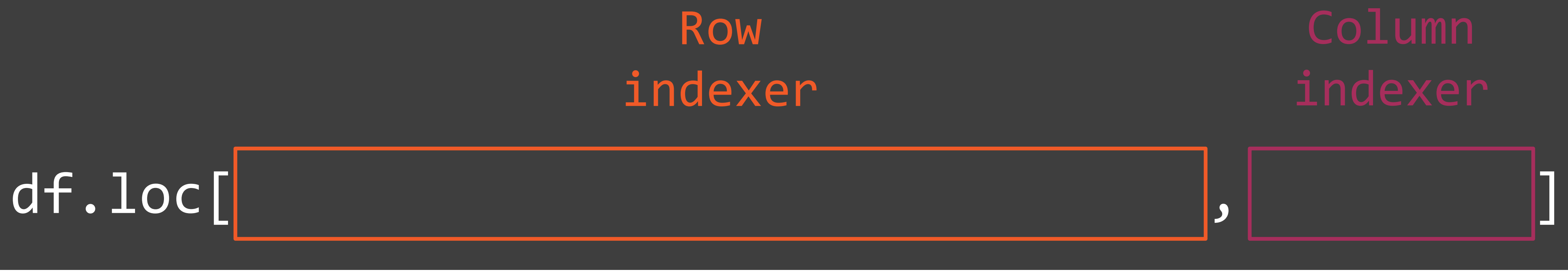
- Count the number of artworks created by Francis Bacon

Indexing Done the Right Way

“loc” and “iloc”

Labels vs Positions

		0	1	...
		artist	artistId	...
0	1023	Blake, William	39	...
1	1045	Bacon, Francis	682	...
2	1087	Bacon, Francis	682	...
3	1099	Bacon, Francis	682	...
4	1011	Blake, William	39	...



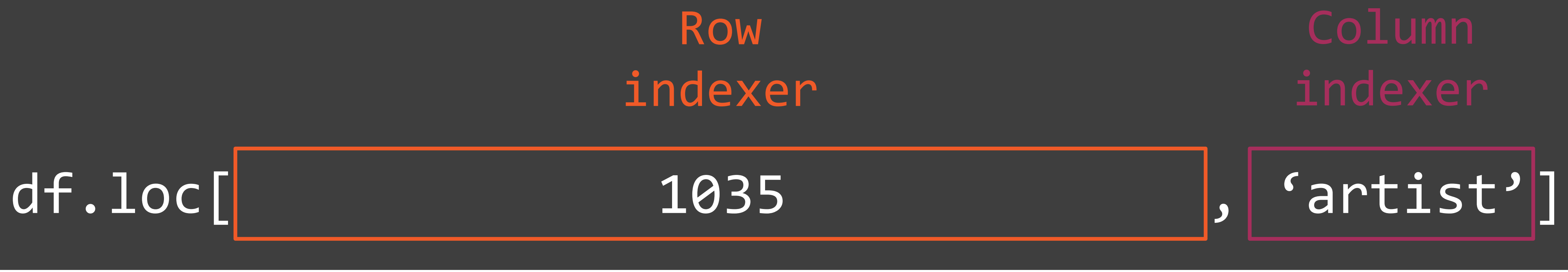
Row indexer

Column indexer

```
df.loc[ , ]
```

Selecting by Label

Reliable approach



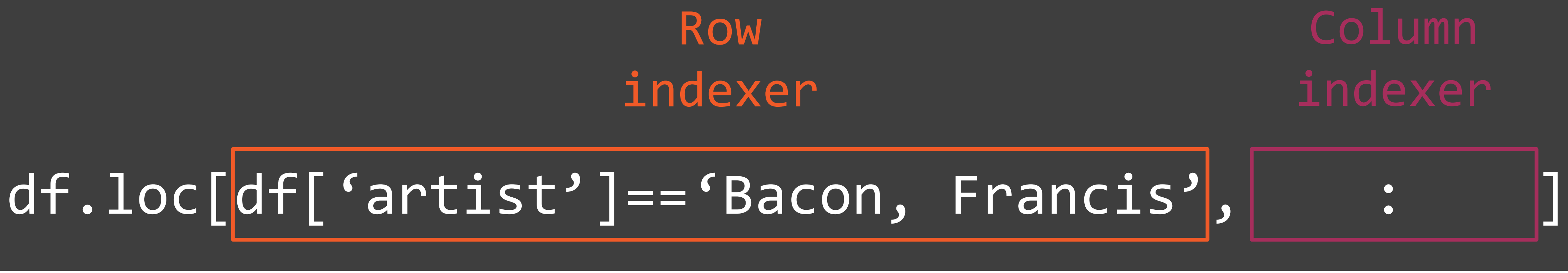
Row indexer

Column indexer

```
df.loc[1035, ‘artist’]
```

Selecting by Label

Reliable approach



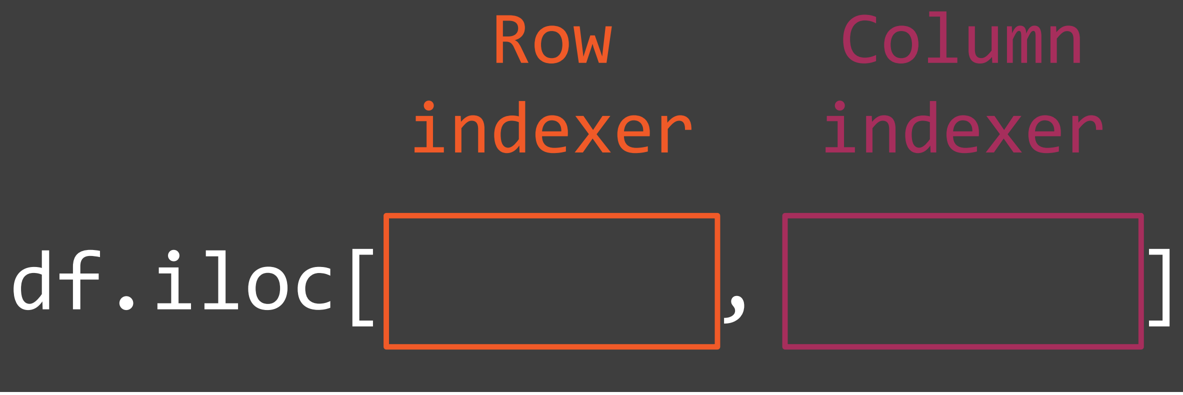
Row indexer

Column indexer

```
df.loc[df[‘artist’]==‘Bacon, Francis’, : ]
```

Selecting by Label

Reliable approach



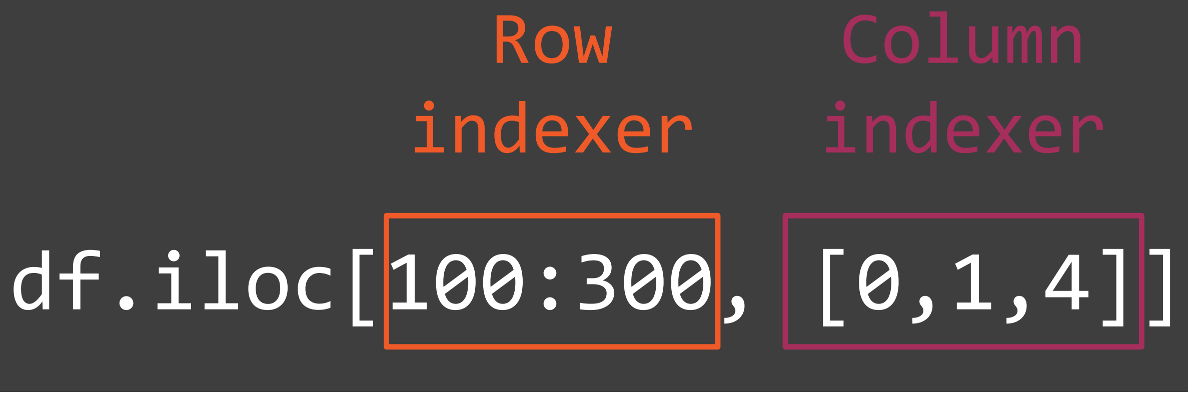
Row indexer

Column indexer

```
df.iloc[          ,          ]
```

Selecting by Position

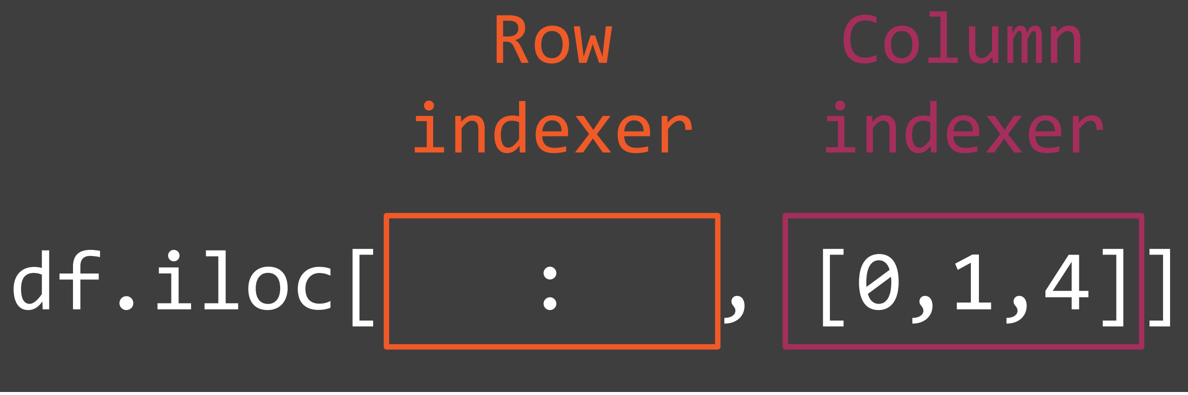
Reliable approach



```
df.iloc[100:300, [0,1,4]]
```

Selecting by Position

Reliable approach



Row indexer

Column indexer

```
df.iloc[ : , [0,1,4]]
```

Selecting by Position

Reliable approach

Demo

We will:

- Practice using loc and iloc
- Find the biggest artwork in the collection
- Learn how to deal with common problems in data analysis

Dos and Don’ts

Always use iloc and loc!

```
df[‘artist’]

df[[‘artist’, ‘title’]]

df.loc[df[‘artist’] == ‘Blake, William’, ‘title’]
```

Rare Exceptions

OK to use shorthand methods