NETFLIX

The missing MatPlotLib for Scala/Spark

@NetflixResearch
@aishfenton @datamusing

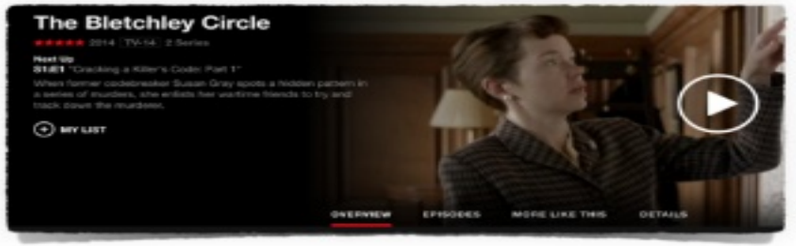
NETFLIX

NETFLIX

AT NETFLIX, WE USE ML EVERYWHERE

NETFLIX

NETFLIX ORIGINAL
NARCOS
Season 2 Coming September 2
06 DAYS
07 HOURS
14 MINS
28 SECS
REWATCH SERIES
TRAILER


The Bletchley Circle
MY LIST
OVERVIEW
EPISODES
MORE LIKE THIS
DETAILS

NETFLIX
HOUSE of CARDS

Trending Now
MIDSOMER MURDERS
QUANTICO
JANE GOT A GUN
POIROT
VINCENT
AMERICAN CRIME

Vintage Crime ›
NO COUNTRY FOR OLD MEN
AQUARIUS
LIZZIE BORDEN
DARYBA DIARIES
THE ICEMAN

STRANGER THINGS
RIVER
HOUSE of CARDS
Grace and Frankie
FEARLESS
NARCOS

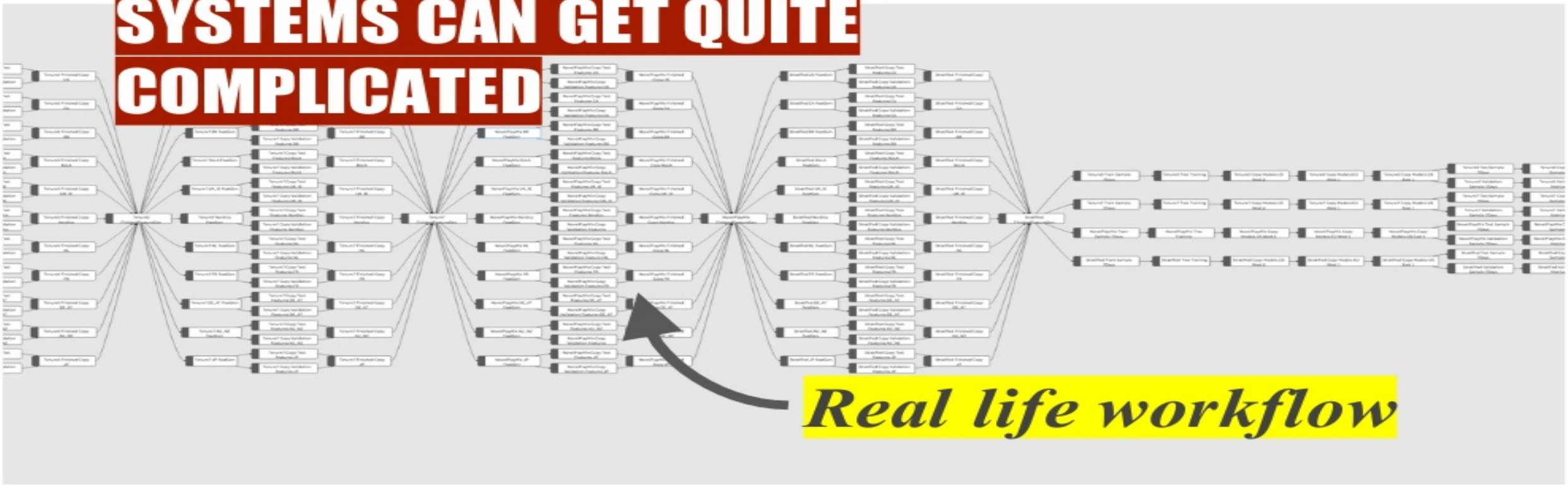
#netflixeverywhere
Jan 6th, 2016

NETFLIX

CONSTANTLY INNOVATING

. . . .

MACHINE LEARNING SYSTEMS CAN GET QUITE COMPLICATED



NETFLIX

VISUALIZATIONS CAN HELP WITH

GUIDANCE
INTROSPECTION
EVALUATION

DURING THE INNOVATION CYCLE

NETFLIX

STATISTICAL VISUALIZATIONS CAN BE PAINFUL

Consider a researcher at Netflix who who has raw data in a spark dataframe with columns:

show_id, num_of_views, country, timestamp, video_age

The researcher wants to make the following plot:

Plot the *five most popular* titles,

according to total number of views,

in the *last 5 hours*,

as bar charts *faceted by country*,

where the bars are *color coded* by video_age.

Sorting

Aggregating after filtering by timestamp

Grouping data by a categorical value (Country)

Mapping a quantitative column to a color

One could perform all these operations on the DF first and then create one bar plot per country via a loop.

Painful indeed!

NETFLIX

VEGAS

DECLARATIVE STATISTICAL VISUALIZATION GRAMMAR

IN SCALA

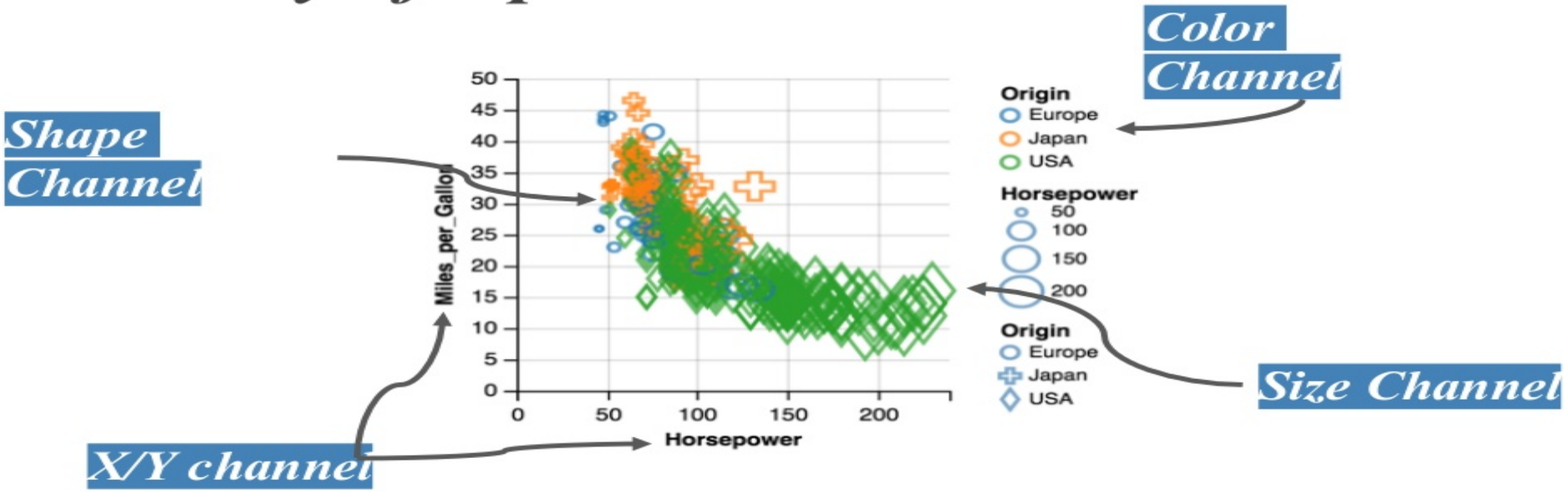
You tell is ***WHAT*** should be done with the data, and it knows ***HOW*** to do it!

Operations such as ***filtering***, ***aggregation***, ***faceting*** are built into the visualization, rather than putting the burden on the user to massage the data into shape.

Complex visualizations can be built with a few high level abstractions:

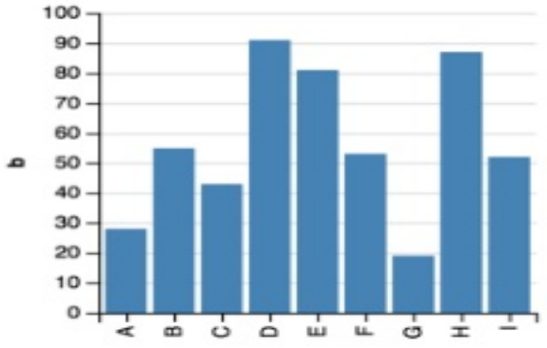
cf : Altair Talk by Brian Granger in PyData 2016 https://youtu.be/v5mrwq7yJc4

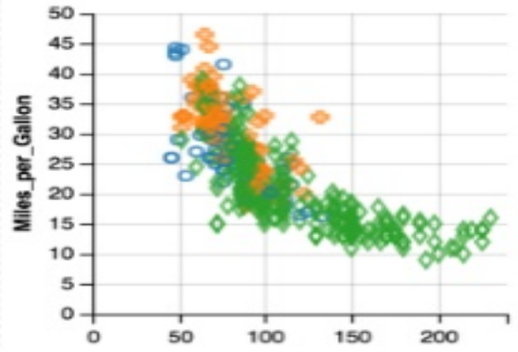
Anatomy of a plot

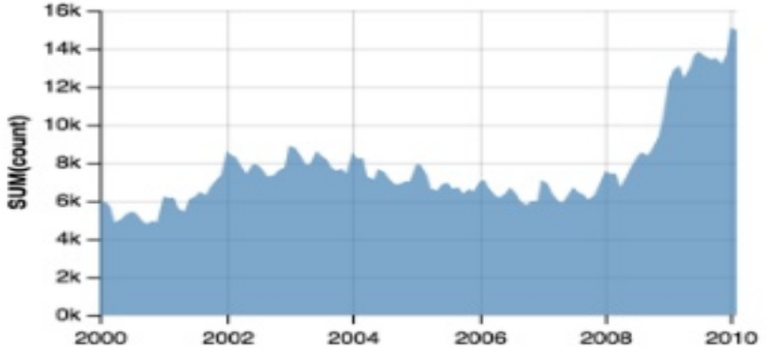


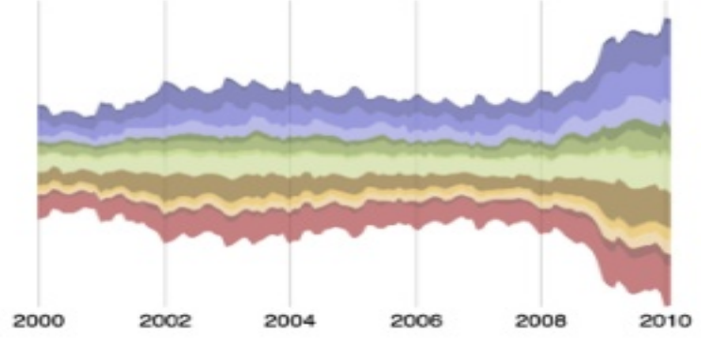
Features...

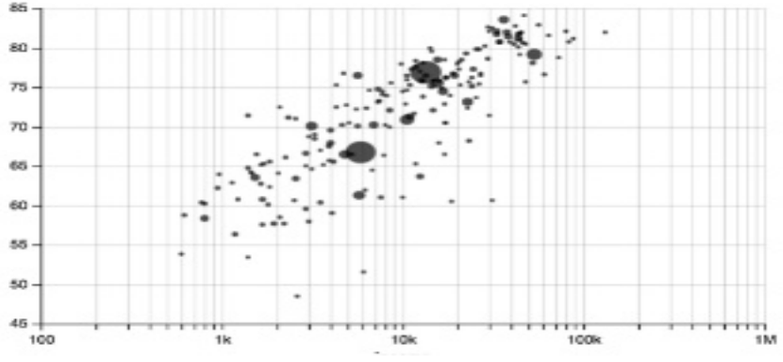
1. Supports most plot types



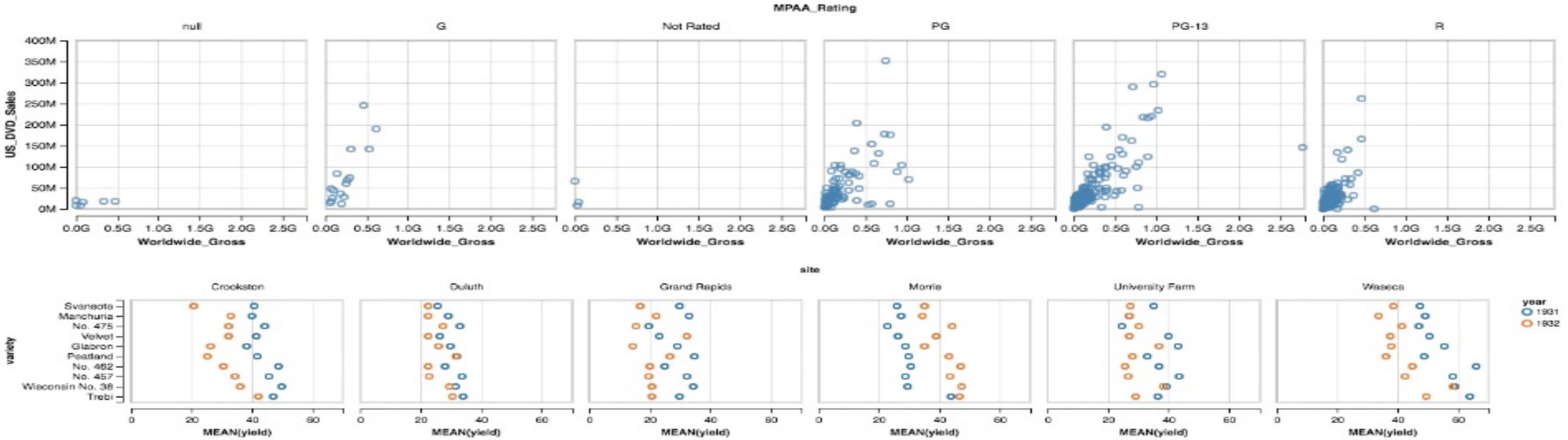








2. Trellis plots

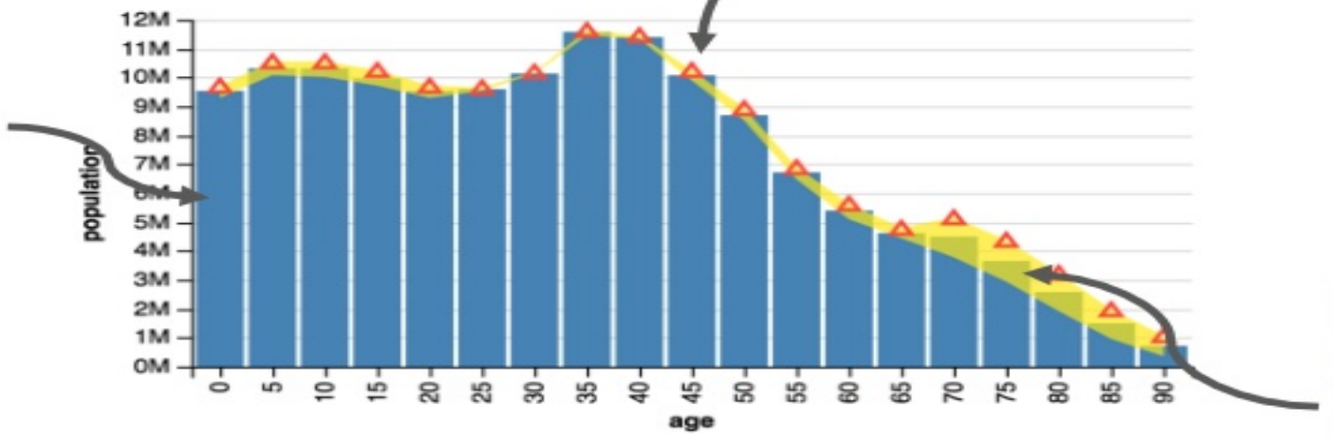


3. Layers

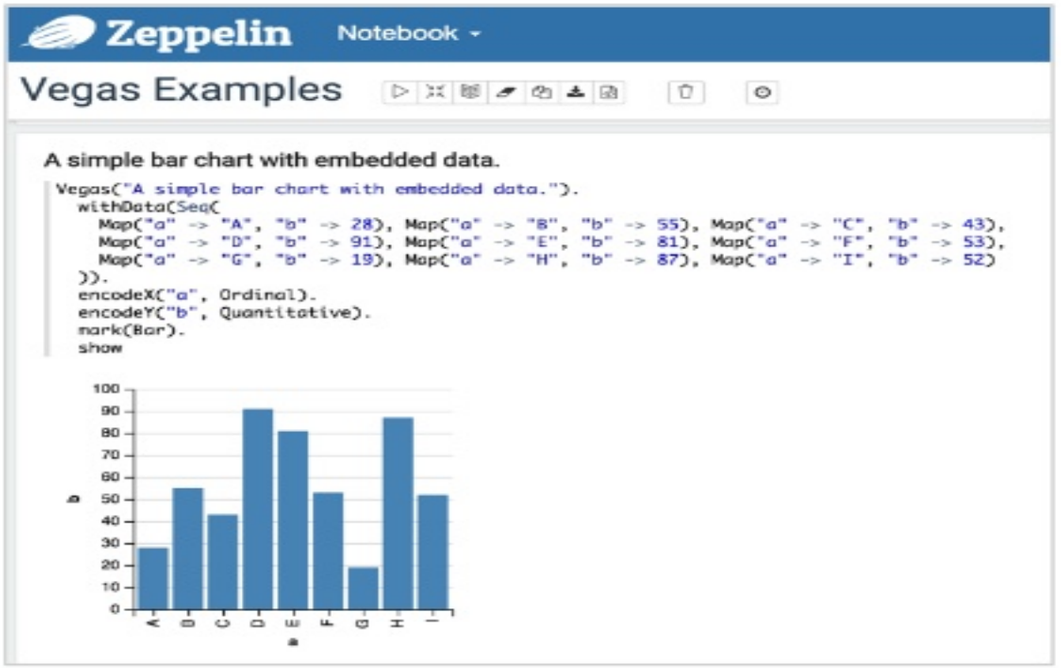
Layer 1.

Layer 2.

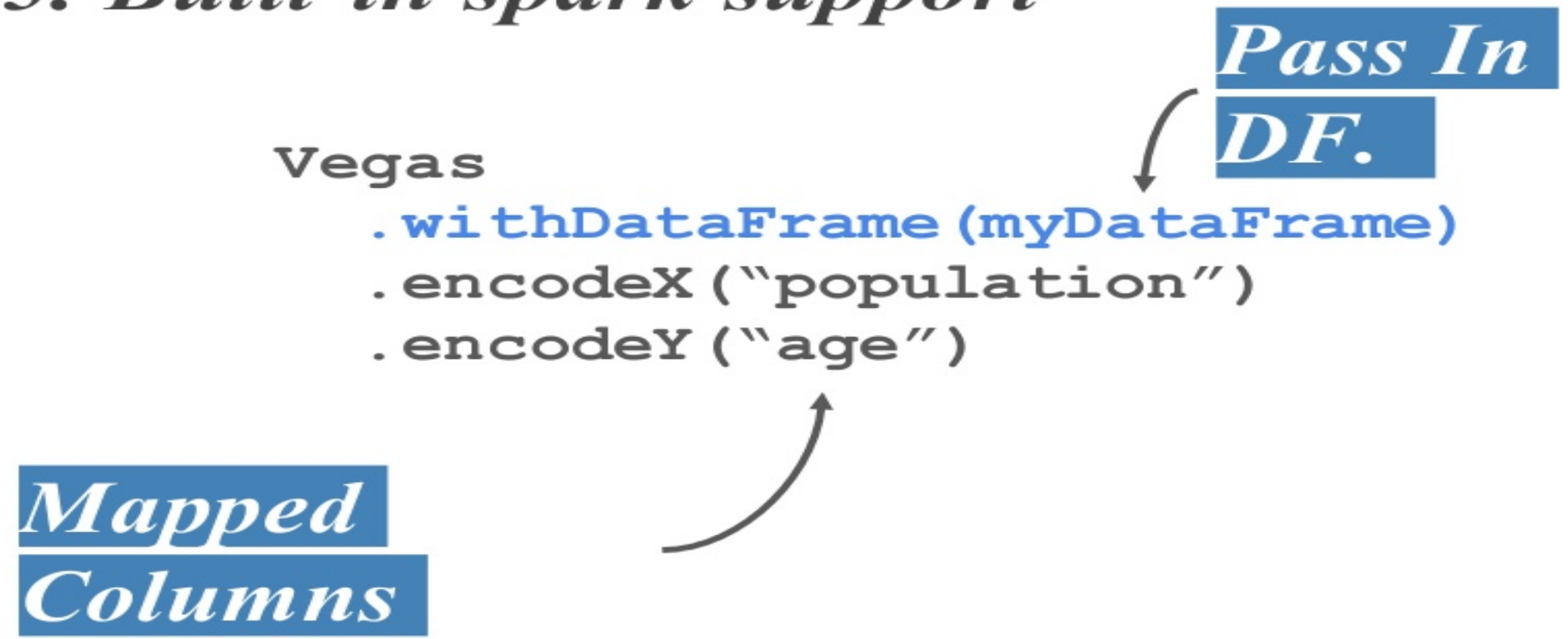
Layer 3.



4. Notebook and Consoles



5. Built-in spark support



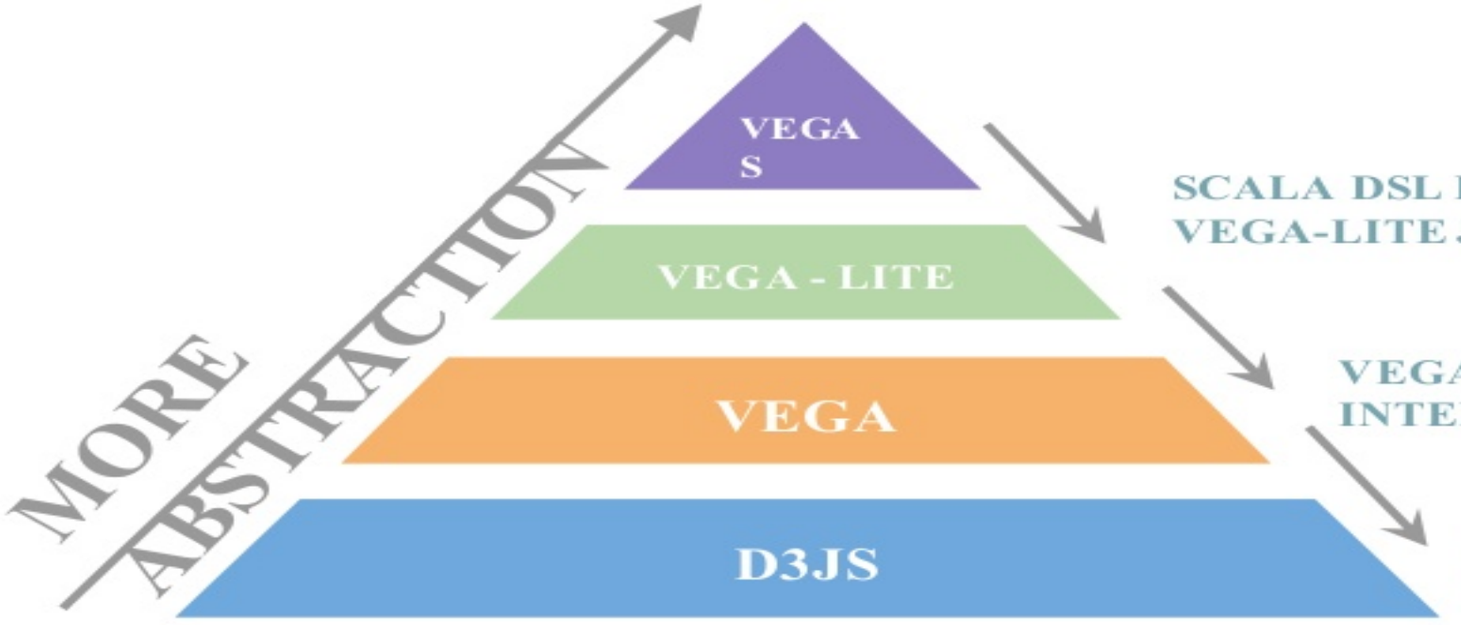
6. Visual statistics

- Advanced Binning
- Sorting
- Scaling
- Custom Transforms
- Time Series
- Aggregation
- Filtering
- Math functions (log, etc)
- Missing data support
- Descriptive Statistics

How It Works !

NETFLIX

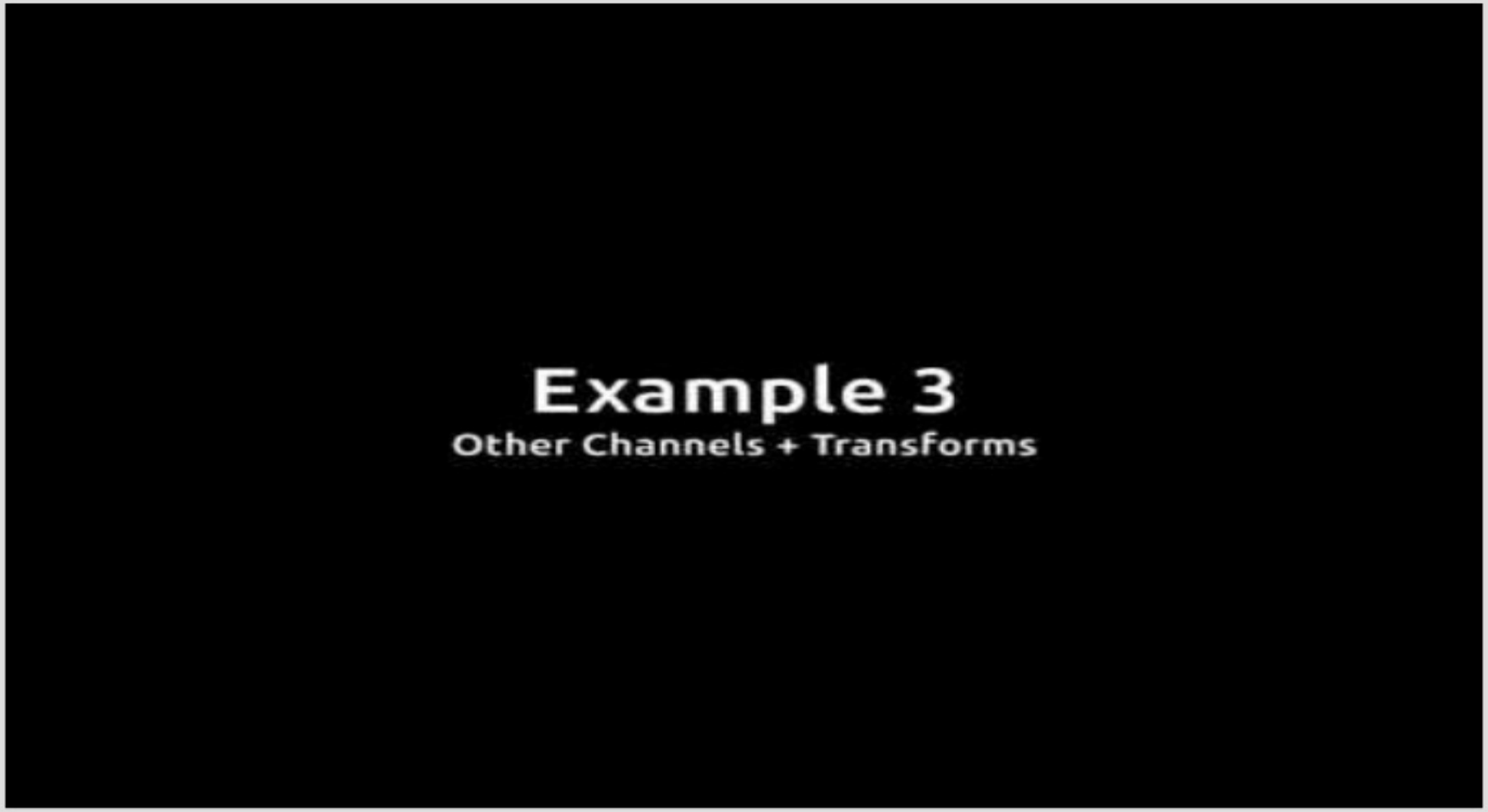
A SCALA DSL FOR VEGA-LITE



SCALA DSL EMITS TYPE-CHECKED VEGA-LITE JSON

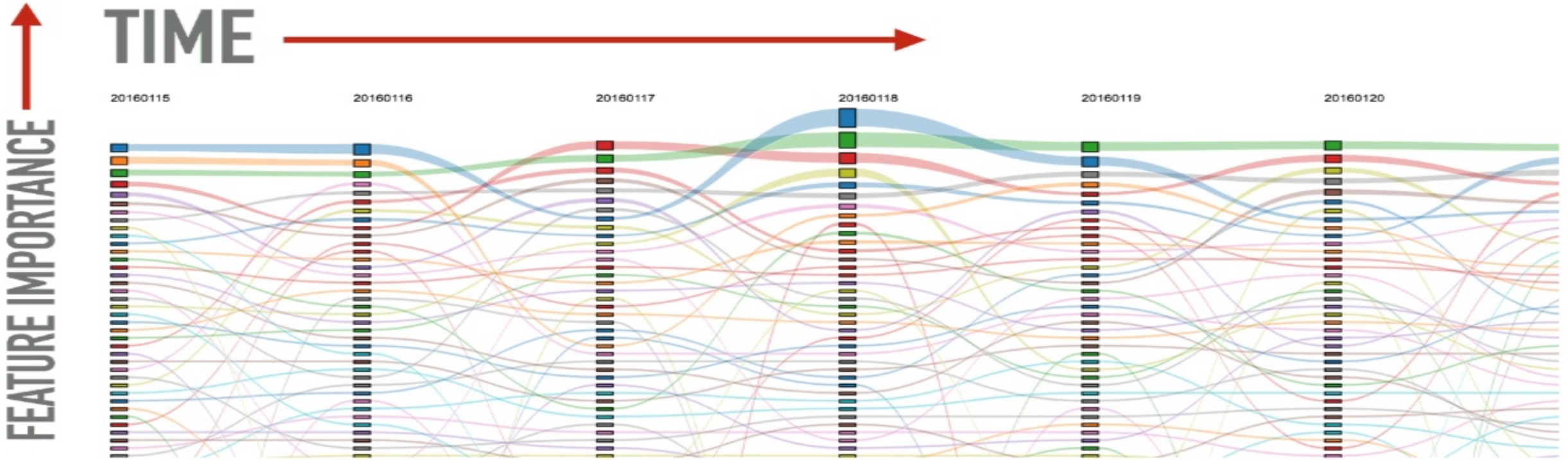
VEGA-LITE CONVERTS INTERNALLY TO VEGA JSON SPEC

VEGA TRANSLATES JSON TO D3JS CODE THAT CAN BE VERY VERBOSE


Example 3
Other Channels + Transforms

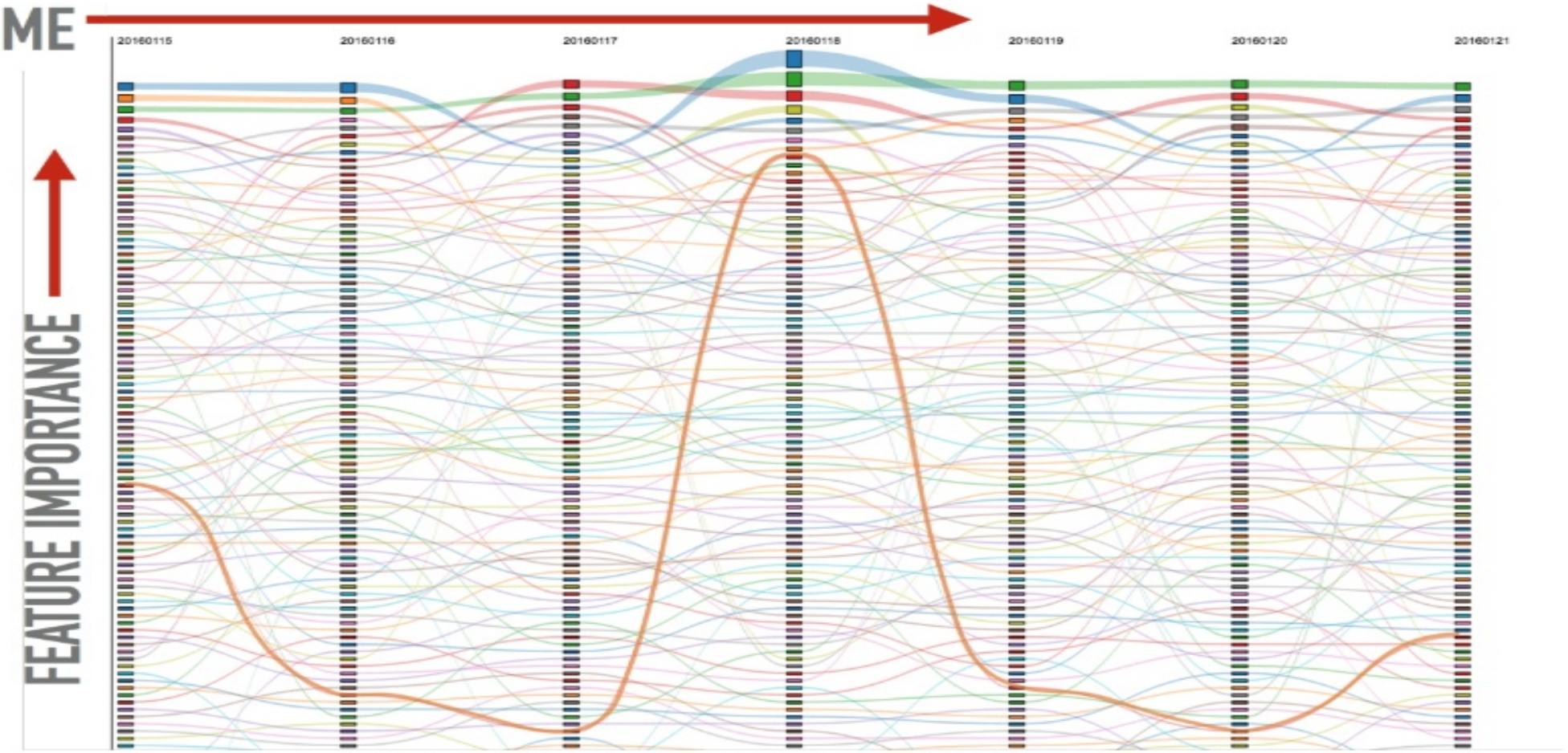
ANY
QUESTIONS

NETFLIX


TIME
FEATURE IMPORTANCE
20160115
20160116
20160117
20160118
20160119
20160120

NETFLIX

WHAT HAPPENED WITH THAT FEATURE?



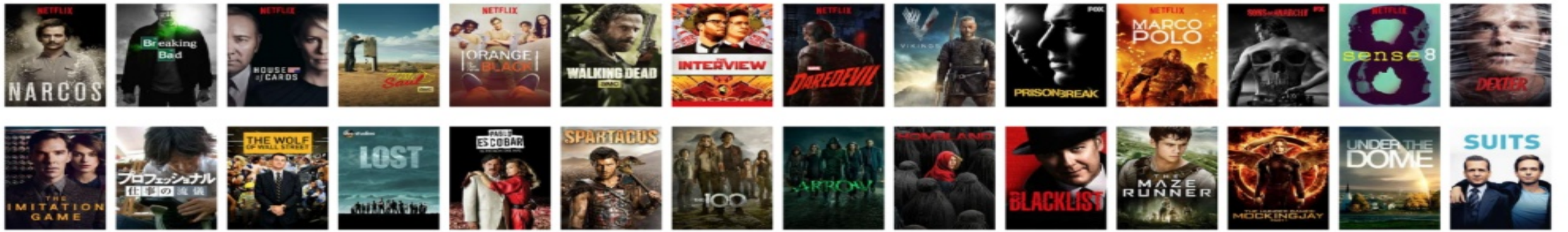
NETFLIX

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99

NETFLIX

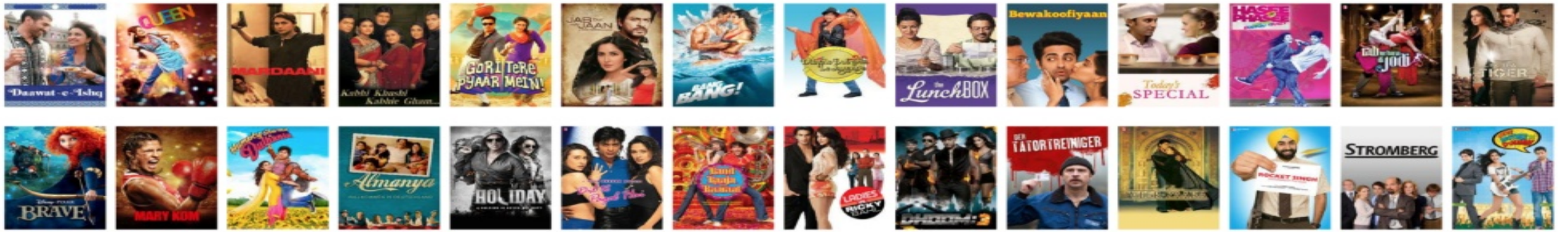
Data-Driven Documents

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99



NETFLIX

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99

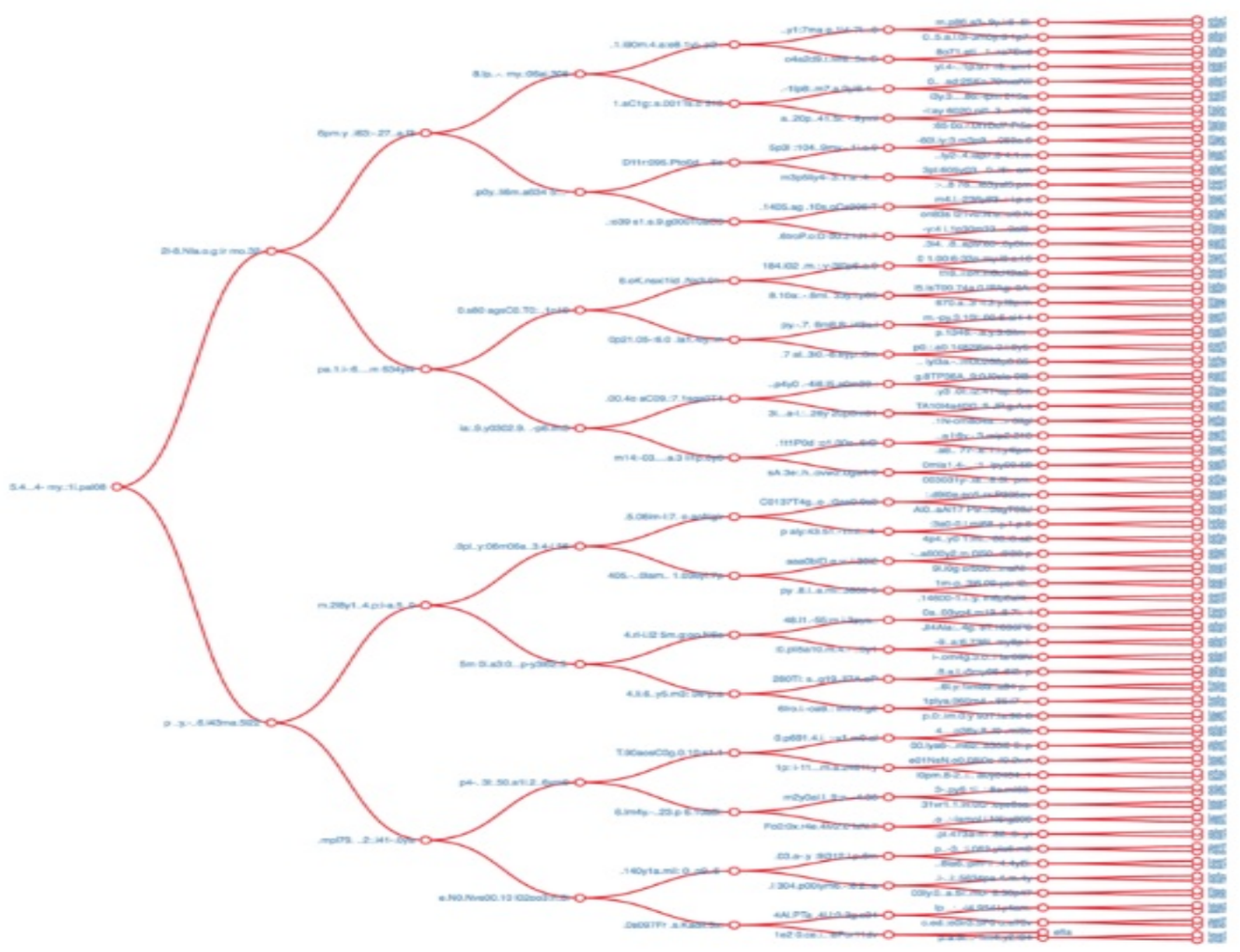
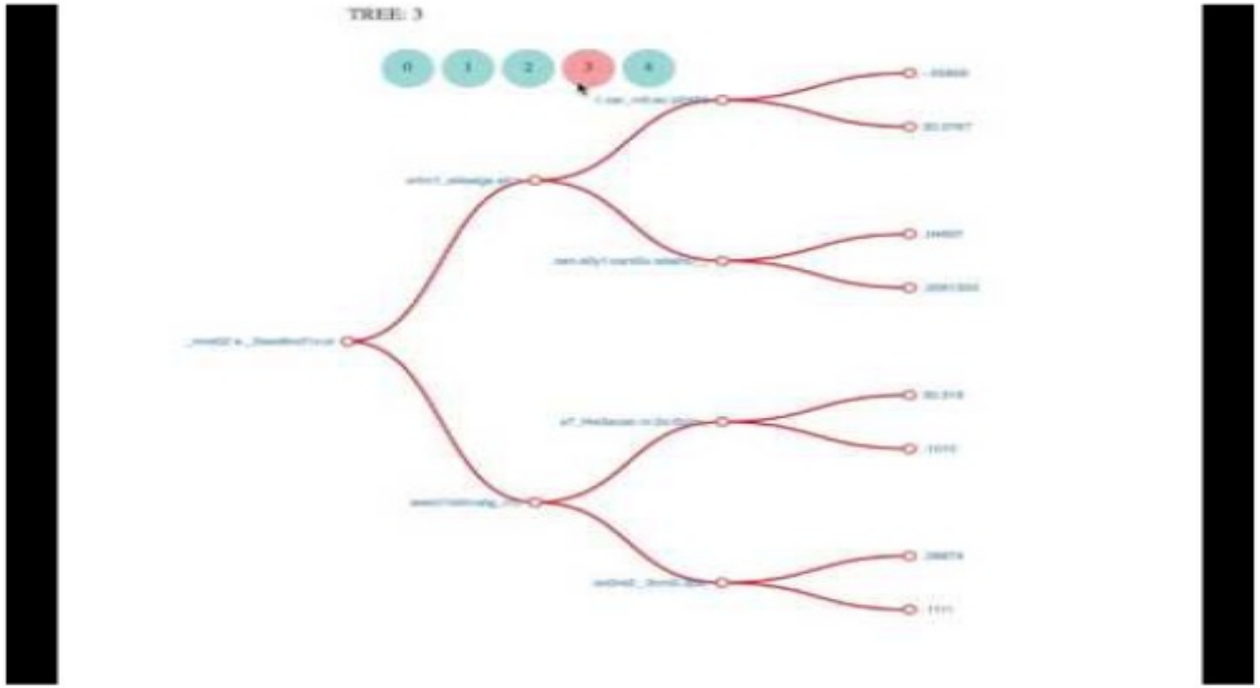


NETFLIX

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99

NETFLIX

SPLITTING HAIRS!



NETFLIX