



智捷分析 洞鉴先机: SumUp Nucleus实时文本分析平台

Presented By



SUMUP ANALYTICS

吴永俊 (Jim Wu)
10月16日

智捷分析 洞鉴先机

SUMUP
ANALYTICS

THE PAIN POINT

Too Much To Read

Time Is Scarce

WHAT WE DO

High-granularity real-time analysis on Unstructured data

- x Time-consuming
- x Voluminous
- x Unstructured
- x Overwhelming

WHAT WE DO

High-granularity real-time analysis on Unstructured data



- x Time-consuming
- x Voluminous
- x Unstructured
- x Overwhelming

WHAT WE DO

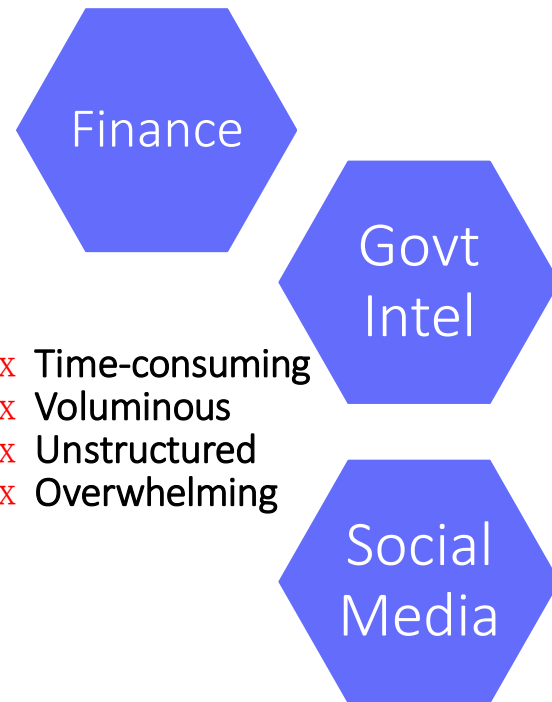
High-granularity real-time analysis on Unstructured data



- x Time-consuming
- x Voluminous
- x Unstructured
- x Overwhelming

WHAT WE DO

High-granularity real-time analysis on Unstructured data



WHAT WE DO

High-granularity real-time analysis on Unstructured data

Finance

Govt
Intel

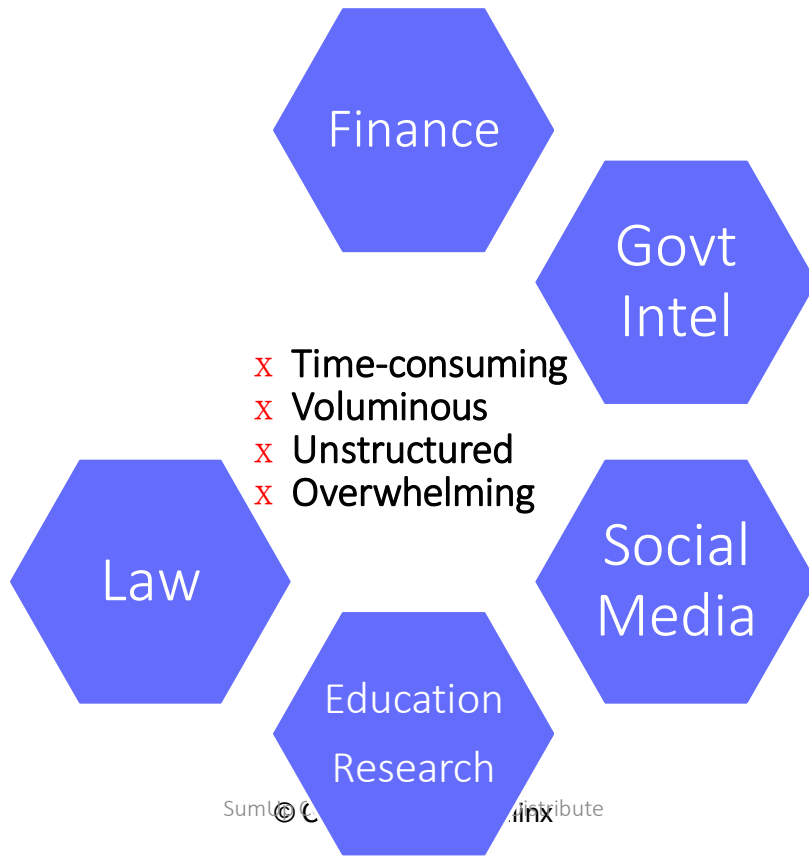
- x Time-consuming
- x Voluminous
- x Unstructured
- x Overwhelming

Social
Media

Education
Research

WHAT WE DO

High-granularity real-time analysis on Unstructured data



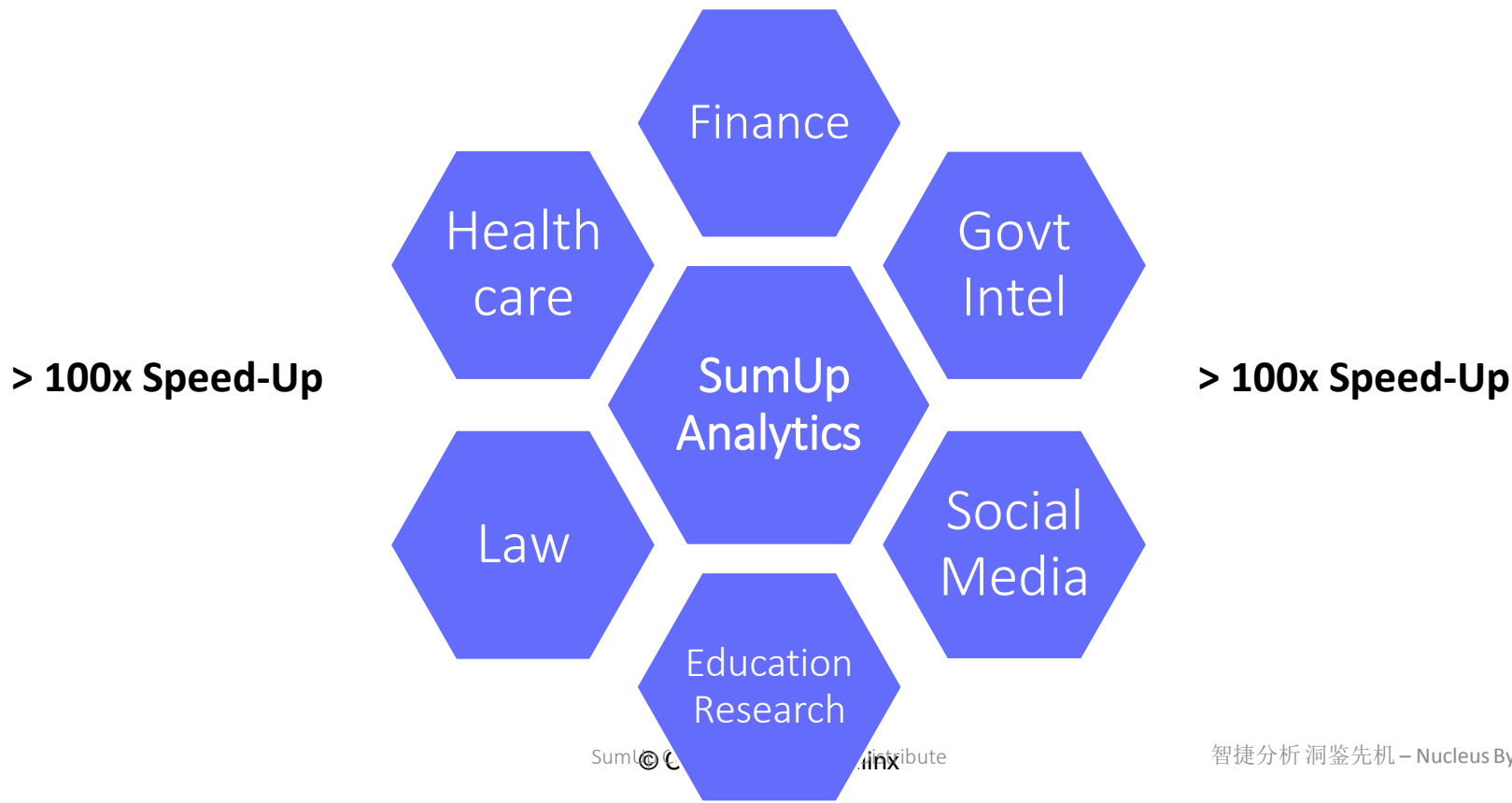
WHAT WE DO

High-granularity real-time analysis on Unstructured data



WHAT WE DO

High-granularity real-time analysis on Unstructured data



OUR SOLUTIONS

Nucleus SaaS

On-Demand Analytics APIs

Streaming Analytics APIs

Step 1: Upload | DataFeeds



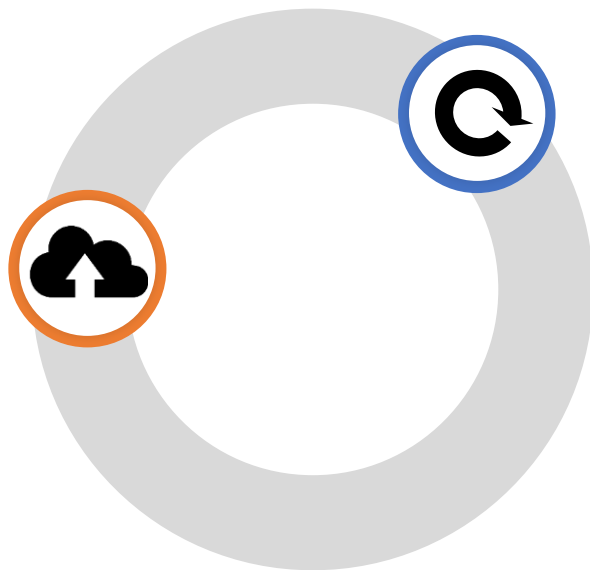
OUR SOLUTIONS

Nucleus SaaS

Streaming Analytics APIs

On-Demand Analytics APIs

Step 1: Upload | DataFeeds



Step 2: Iterative Analysis

- Identify Topics & Summarize
- Analyze Sentiment & Consensus
- Get document recommendations
- Deep-dive on-demand

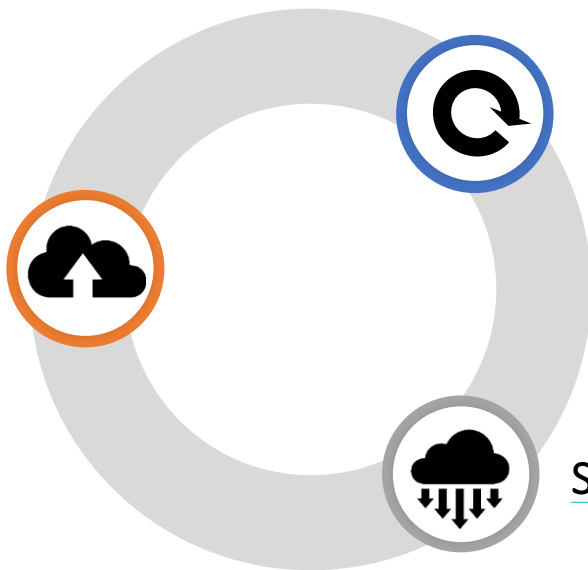
OUR SOLUTIONS

Nucleus SaaS

Streaming Analytics APIs

On-Demand Analytics APIs

Step 1: Upload | DataFeeds



Step 2: Iterative Analysis

- Identify Topics & Summarize
- Analyze Sentiment & Consensus
- Get document recommendations

Step 3: Export & Share

NEWS

News Feed

RECENT DATASETS

Test ordering

Test5

Chinese BIS

10K

Trump Twits

See All Datasets

Create New Dataset

DATASET OVERVIEW

DATASET NAME	10K
DATE CREATED	August 21, 2018
FILE COUNT	111 Files
SIZE	106.6 MB
LANGUAGES	English

TOPICS IN DATASET

TOPIC1	data—item
	financial—statements
	statements—supplementary
	supplementary—data
TOPIC2	income—loss
	cash—cash
	cash—equivalents
	income—taxes
TOPIC3	analysis—financial
	operations—item
	management—discussion
	item—management

FILTER DATASET

Run a deeper analysis on this dataset using specific keywords.

dividends ✕

+ ADD KEYWORD AND PRESS ENTER

Add Keyword

Analyze

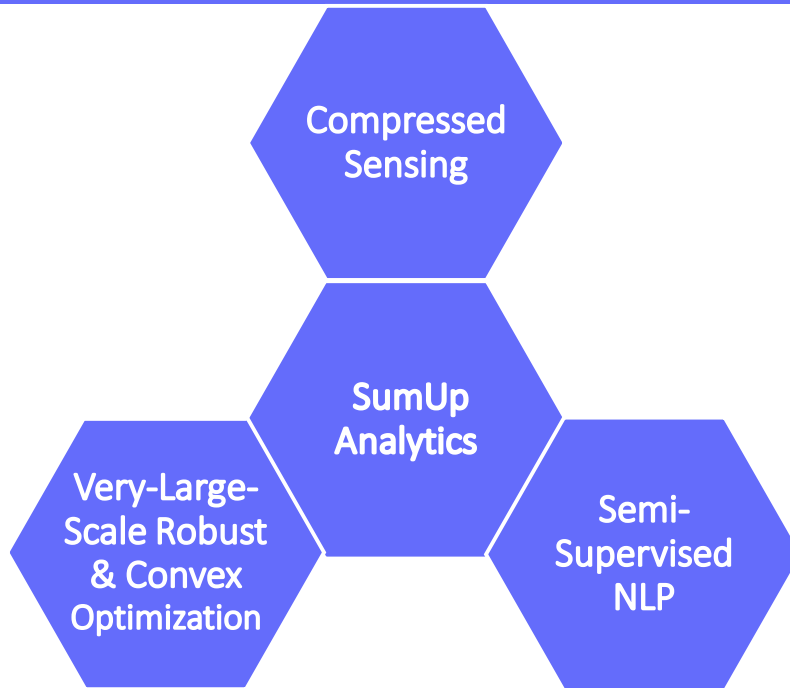
DATA—ITEM FINANCIAL—STATEMENTS STATEMENTS—SUPPLEMENTARY SUPPLEMENTARY—DATA

- SUMMARY
- Pronounced variability or delays in the construction of desalination plants or reductions in spending for desalination, could negatively impact our Water segment sales and revenue, which in turn could have an adverse effect on our entire business, financial condition, or results of operations and make it difficult for us to accurately forecast our future sales and revenue.
Source: [ERII10-K2018-03-08.txt](#)
 - MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS ", " Our fiscal year begins on April 1 and ends on March 31.
Source: [WMS10-K2016-09-15.txt](#)
 - ", "31", "Certain of our accounting estimates or assumptions constitute "critical accounting estimates" for us because:", "the nature of these estimates or assumptions is material due to the levels of subjectivity and judgment necessary to account for highly uncertain matters or the susceptibility of such matters to change; and", "the impact of the estimates and assumptions on financial condition and results of operations is material.
Source: [CWCO10-K2016-03-15.txt](#)



DIFFERENTIATED INTELLECTUAL PROPERTIES

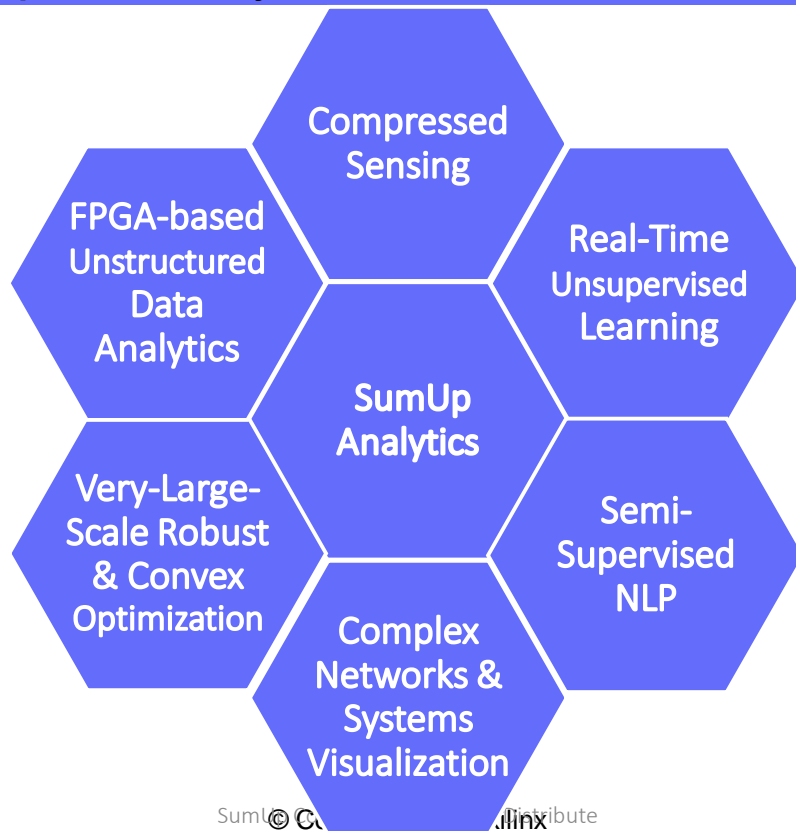
R&D in Analytics & Computational Systems



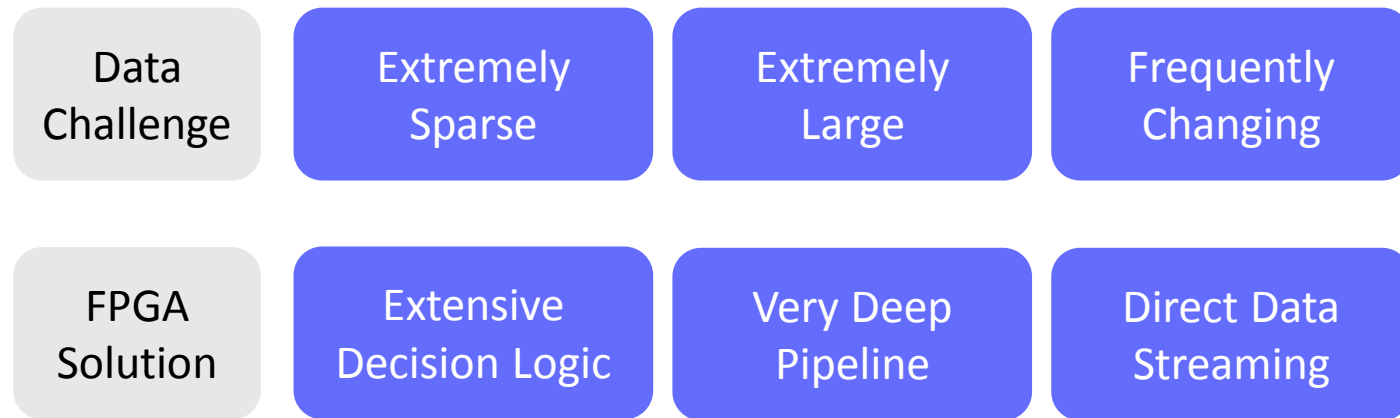
DIFFERENTIATED INTELLECTUAL PROPERTIES

R&D in Analytics & Computational Systems

Leverages our Xilinx partnership



FPGA: NATURAL FIT FOR UNSTRUCTURED DATA

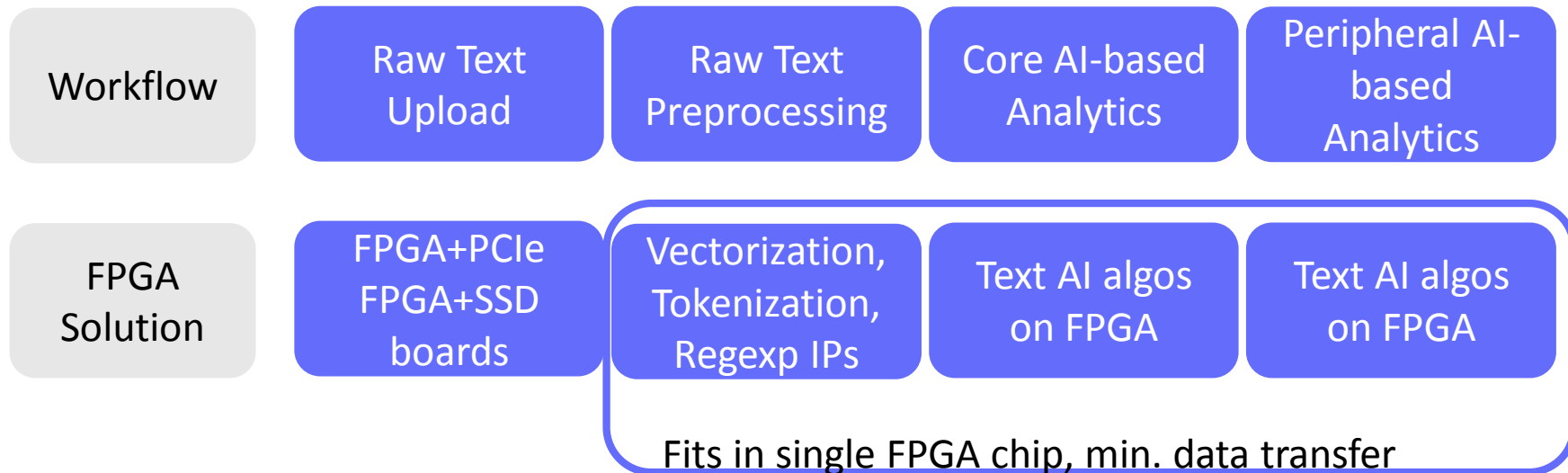


FPGA very appealing to implement AI algos for Unstructured Data

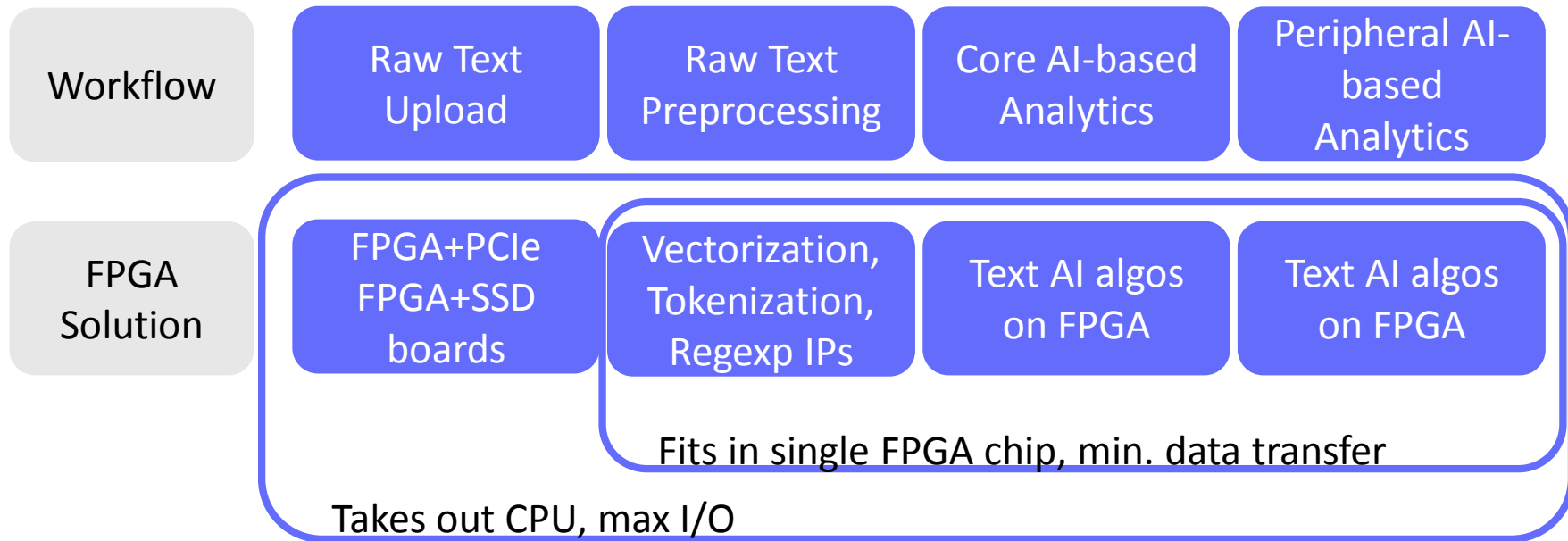
FPGA: POWERS NUCLEUS END TO END



FPGA: POWERS NUCLEUS END TO END



FPGA: POWERS NUCLEUS END TO END



ACCELERATED DEV & DEPLOY W/ XILINX

Xilinx Public IPs

SumUp Python/SDAccel Wrapper

AWS F1

Prototyping

03 / 2018

05 / 2018

Python
library

Core algo

Xilinx IPs

Standalone
FPGA-
based, v1

SDAccel

ACCELERATED DEV & DEPLOY W/ XILINX

Xilinx Public IPs

SumUp Python/SDAccel Wrapper

AWS F1

Prototyping

Integration

03 / 2018

05 / 2018

06 / 2018

07 / 2018

Python
library

Core algo

Core algo

Core algo

Xilinx IPs

Standalone
FPGA-
based, v1

Standalone
FPGA-
based, v2

F1-based,
v3

SDAccel

ACCELERATED DEV & DEPLOY W/ XILINX

Xilinx Public IPs

SumUp Python/SDAccel Wrapper

AWS F1

Prototyping

Integration

Deployment

03 / 2018

05 / 2018

06 / 2018

07 / 2018

08 / 2018

10 / 2018

Python
library

Xilinx IPs

SDAccel

Core algo

Standalone
FPGA-
based, v1

Core algo

Standalone
FPGA-
based, v2

Core algo

F1-based,
v3

Commercial
library

F1-deployed

Commercial
library

Alveo-ready

PRODUCT ROADMAP

Deployed on AWS

Powered by Xilinx FPGAs

Coming to VPC

Available Now

Front-End

Dynamic
Webpage

Compute

Peripheral
Analytics

Core Analytics

EC2

F1 | Alveo

Data
Pipeline

Preprocessing

EC2

PRODUCT ROADMAP

Deployed on AWS

Powered by Xilinx FPGAs

Coming to VPC

Available Now

Available 18Q4

Front-End

Dynamic
Webpage

On-Demand
APIs

API Gateway

Compute

Peripheral
Analytics

Core Analytics

EC2

F1 | Alveo

Data
Pipeline

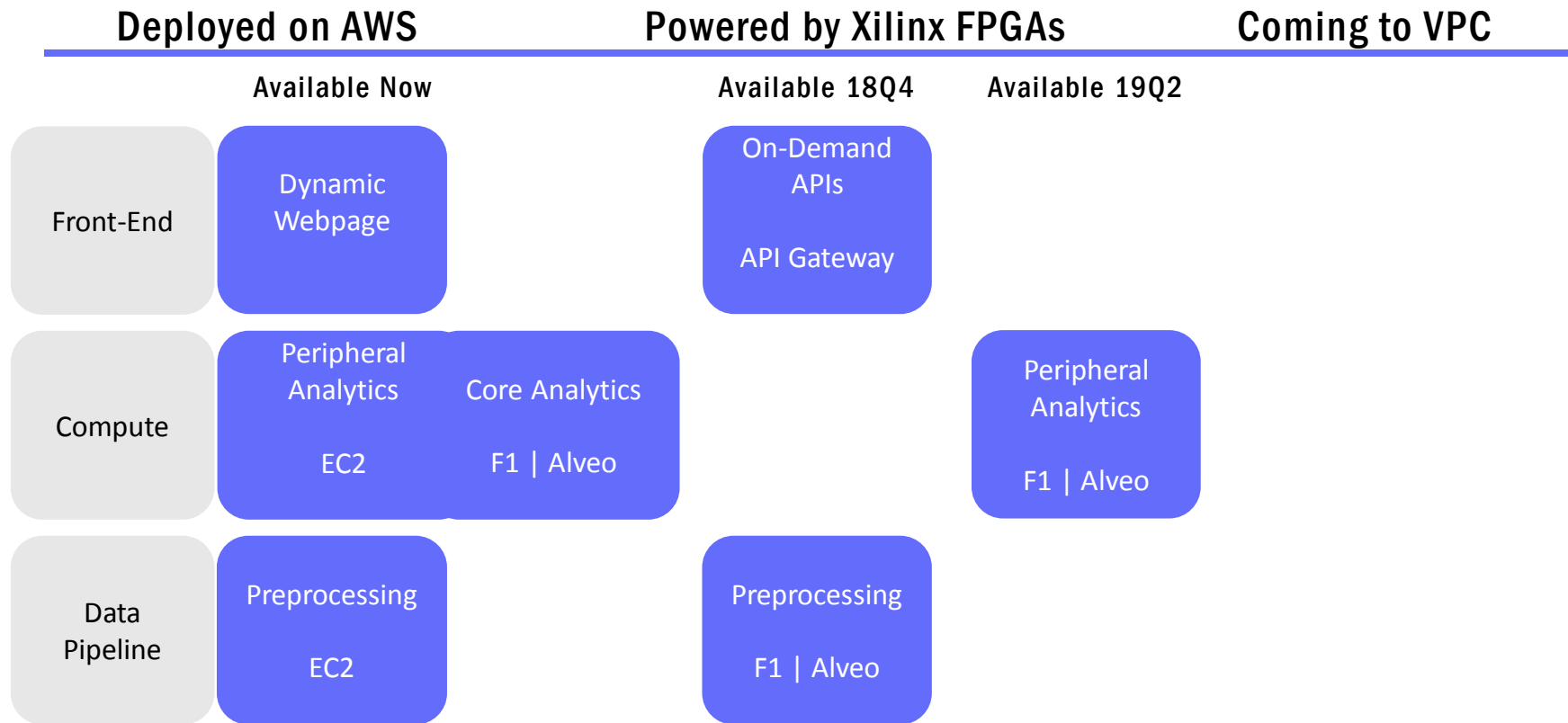
Preprocessing

EC2

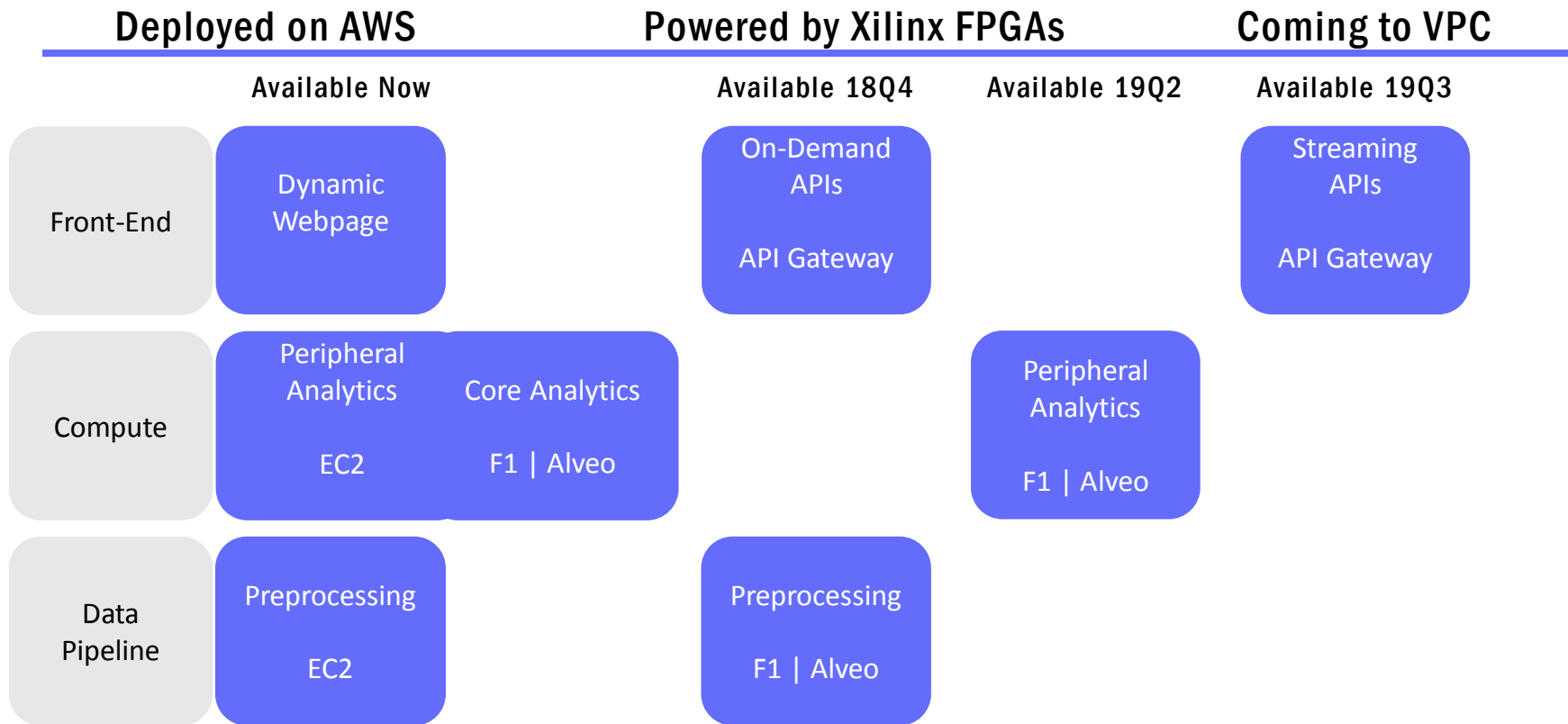
Preprocessing

F1 | Alveo

PRODUCT ROADMAP



PRODUCT ROADMAP



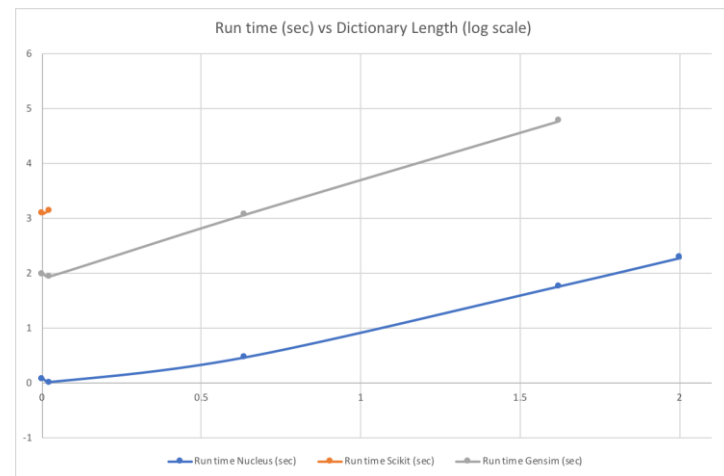
BENCHMARKS

> 100x Faster

Much higher topic quality

Nbr docs	Dictionary length	Nbr topics	Runtime in seconds			U_Mass Coherence		
			Nucleus	Scikit	Gensim	Nucleus	Scikit	Gensim
31,496	81,801	1	0.1	230	10	-9.7	-19.9	-19.9
		4	0.3	648	32	-9.4	-14.7	-15.5
		8	0.6	1157	51	-11.9	-11.8	-16.0
29,287	78,063	1	0.1	220	10	-4.9	-5.8	-18.2
		4	0.2	588	36	-9.8	-14.9	-16.9
		8	0.7	1023	57	-8.3	-15.8	-16.3
158,348	336,532	1	0.4	2817	98	-4.8	-14.0	-18.0
		4	1.0	11282	371	-6.8	-12.7	-17.6
		8	1.7	N/A	685	-6.9	N/A	-16.0
2,620,008	3,253,518	1	9	39600	8135	-10.0	-15.5	-15.4
		4	21	N/A	34850	-9.0	N/A	-16.8
		8	34	N/A	N/A	-6.3	N/A	N/A
7,539,661	7,817,625	1	20	N/A	N/A	-2.5	N/A	N/A
		4	54	N/A	N/A	-3.5	N/A	N/A
		8	113	N/A	N/A	-5.8	N/A	N/A

U_Mass: less negative is better



AWS c4.8xlarge, f1.2xlarge

N/A indicates run not completed within 36 hours

DEMO

Alpha Generation in Equities

1day 14h -> 3min 40s

- Portfolio managers,
financial analysts, quants

[See for yourself](#)

Social Media Monitoring

6days -> 55min

- Intelligence analysts,
compliance officers, data
scientists

IN SUMMARY

- Nucleus by SumUp offers the fast, flexible and transparent text analytics solution needed by professionals
- Nucleus typically is $> 100x$ faster than alternative solutions
- Further acceleration is achievable through complete integration of data pipeline and algos on FPGA board

