

Improving the online shopping experience with ONNX

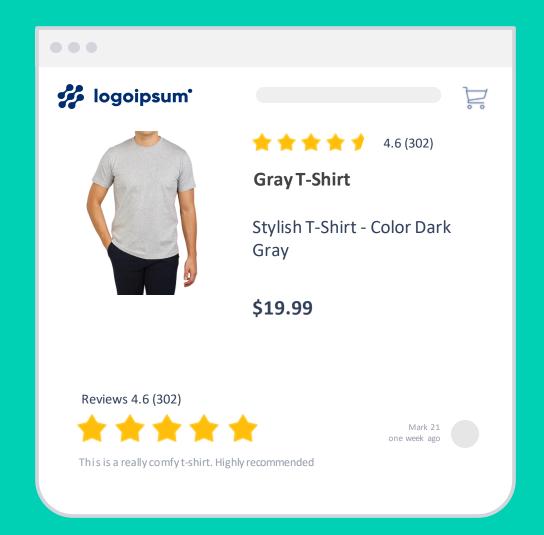
Matthew Leyburn

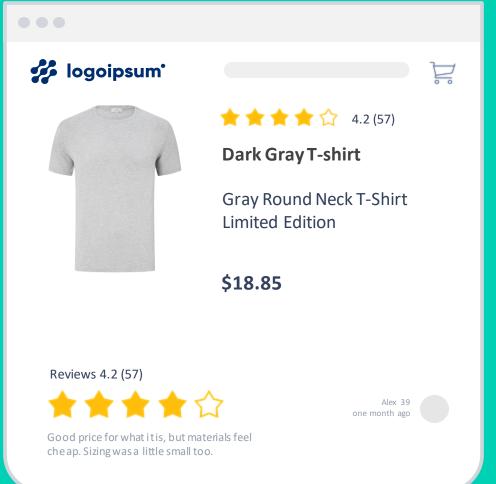
Software Engineer, Bazaarvoice



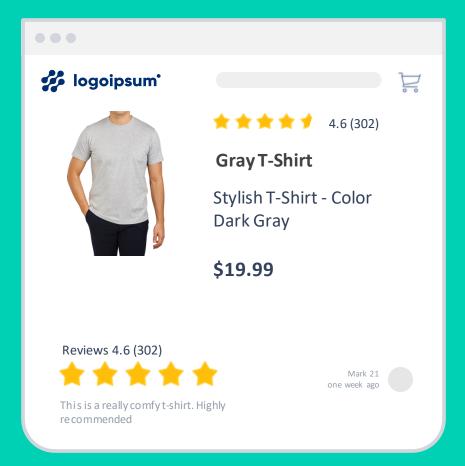


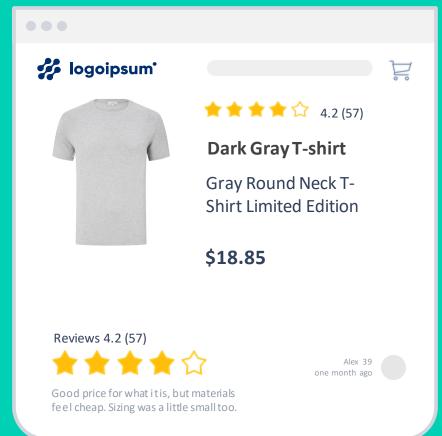
Product matching





Product matching





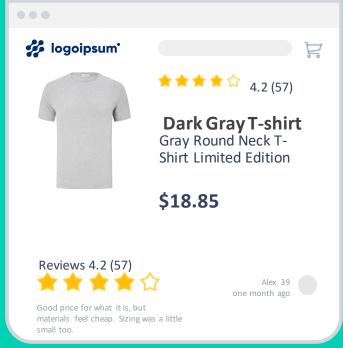
Numerical processing

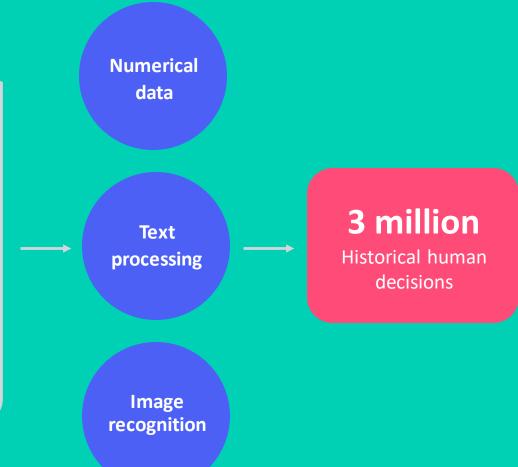
Text processing

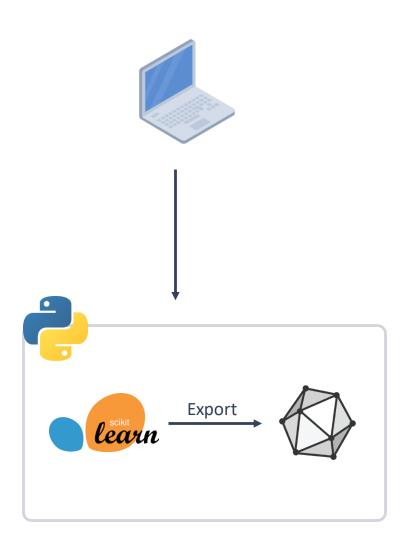
Image recognition

Product matching



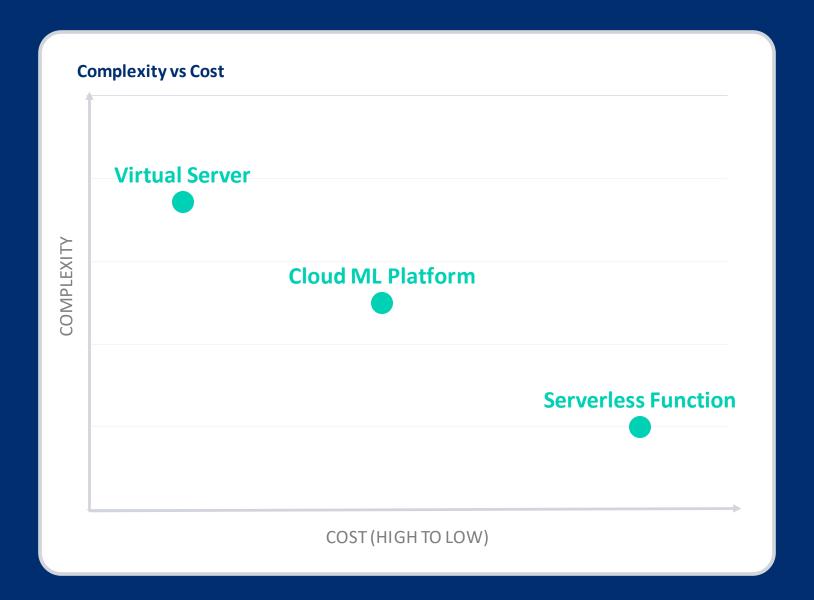


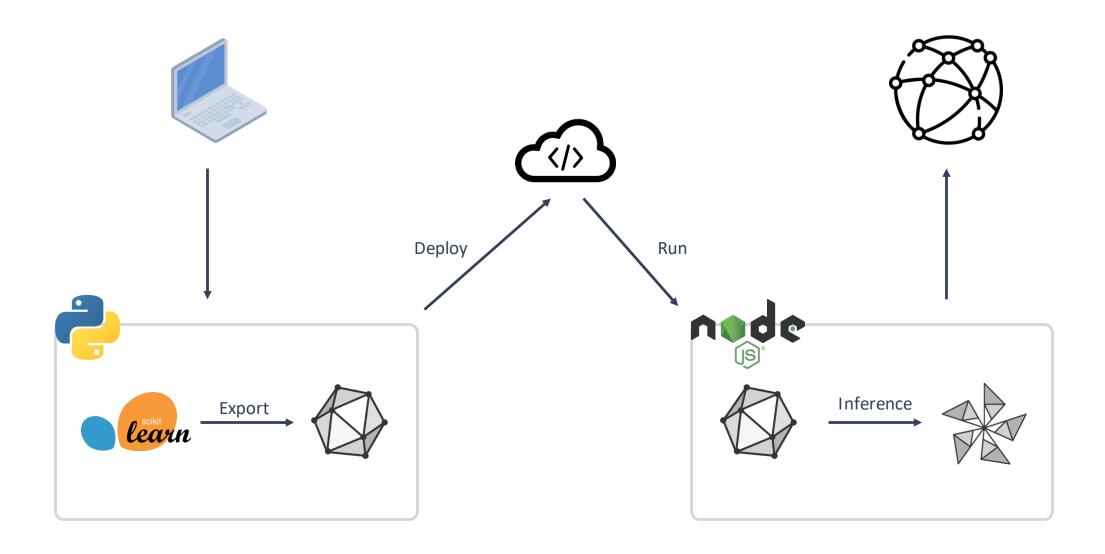




Implementation

Approaches to ML in the cloud





Scaling serverless functions

Performance and cost

Allocated memory (MB)	Memory used (MB)	Execution time (s)	Invocations per month	Cost (USD)
128	114	900 (TIMEOUT)	1,000,000	1875.38
256	232	871	1,000,000	3629.89
512	470	439	1,000,000	3659.07
1024	901	207	1,000,000	3450.69
2048	816	86	1,000,000	2867.24
4096	821	85	1,000,000	5417.15

Scaling serverless functions

Performance and cost

Allocated memory (MB)	Memory used (MB)	Execution time (s)	Invocations per month	Cost (USD)
128	114	900 (TIMEOUT)	1,000,000	1875.38
256	232	871	1,000,000	3629.89
512	470	439	1,000,000	3659.07
1024	901	207	1,000,000	3450.69
2048	816	86	1,000,000	2867.24
4096	821	85	1,000,000	5417.15

Scaling serverless functions

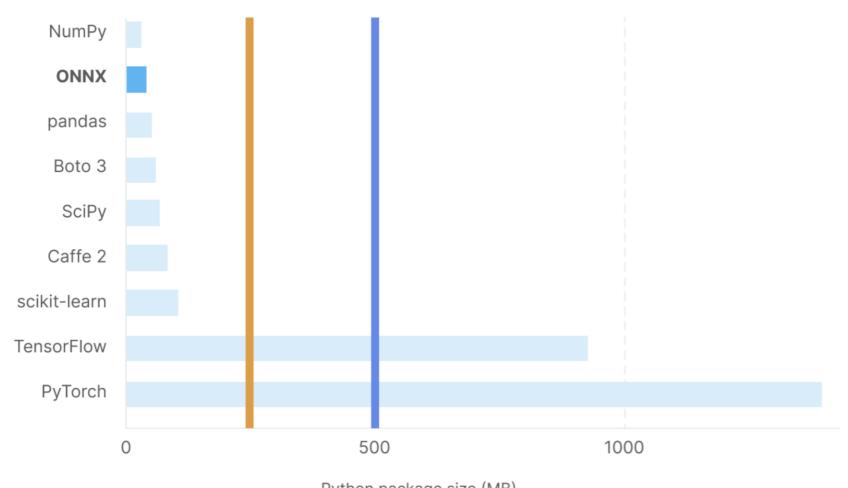
Performance and cost

Allocated memory (MB)	Memory used (MB)	Execution time (s)	Invocations per month	Cost (USD)
128	114	900 (TIMEOUT)	1,000,000	1875.38
256	232	871	1,000,000	3629.89
512	470	439	1,000,000	3659.07
1024	901	207	1,000,000	3450.69
2048	816	86	1,000,000	2867.24
4096	821	85	1,000,000	5417.15

Cost comparison

	Estimated Monthly Cost
Serverless function	\$2,800
Cloud ML Platform	\$4,900
Virtual server	\$14,300

Deployment size limits (MB)



Python package size (MB)





Improving the online shopping experience with ONNX