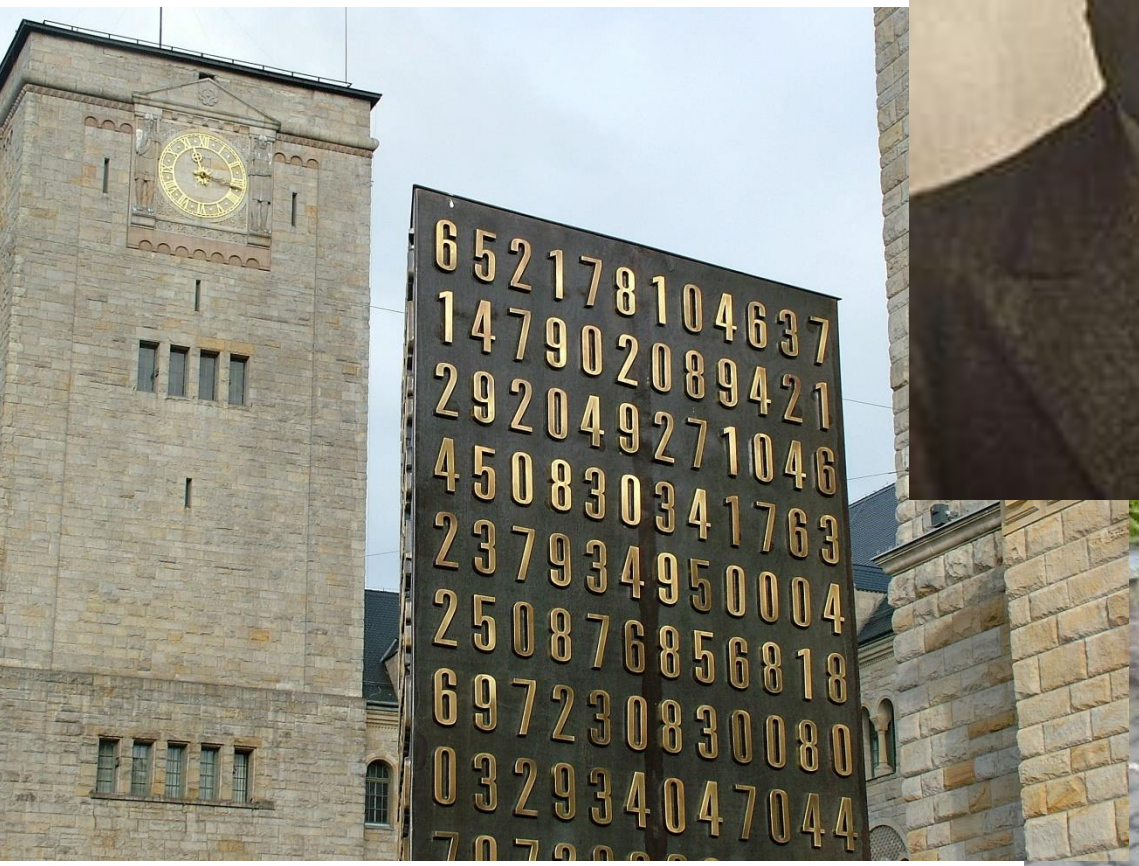
Microsoft

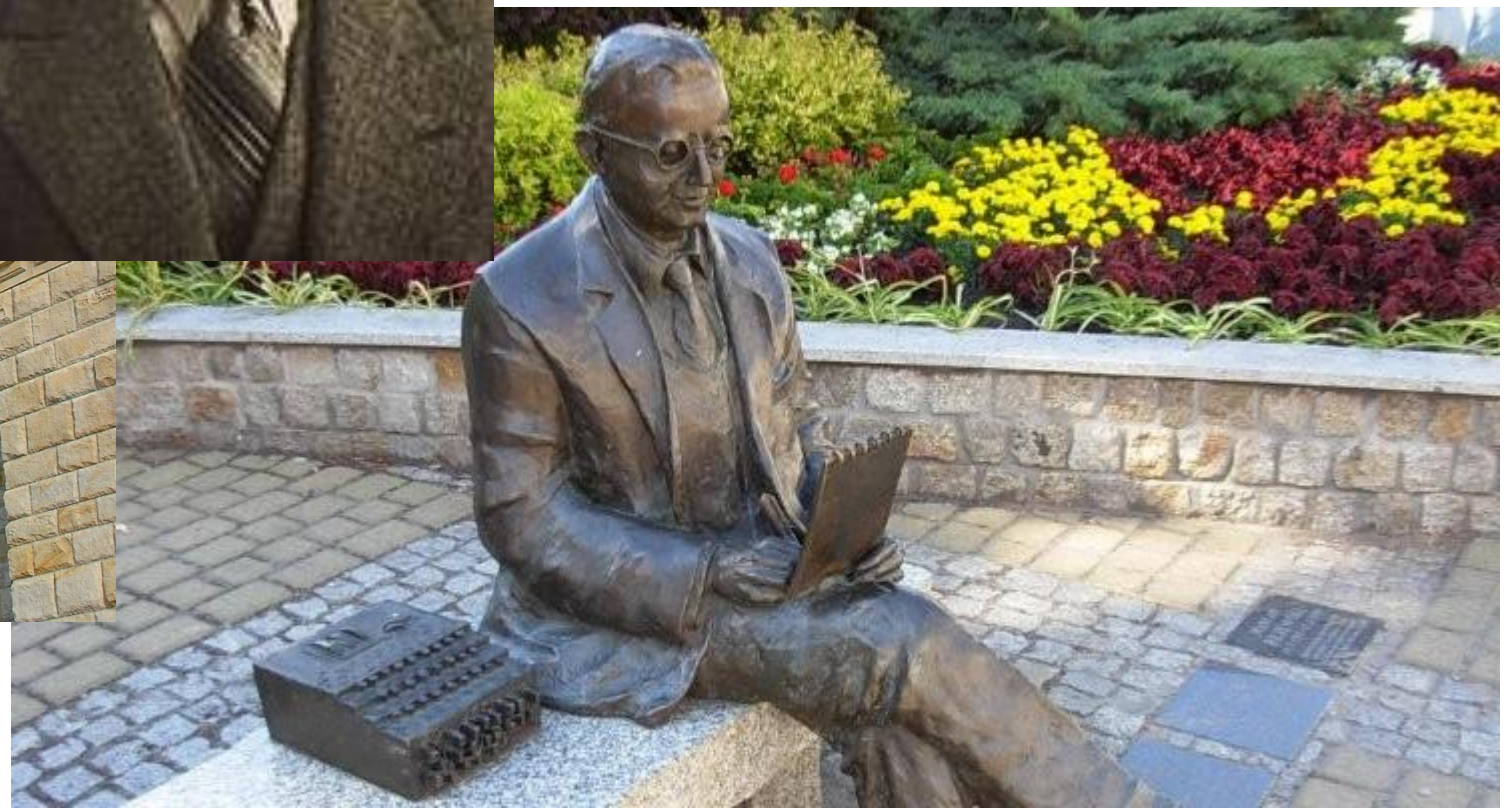
Marian: Homecoming

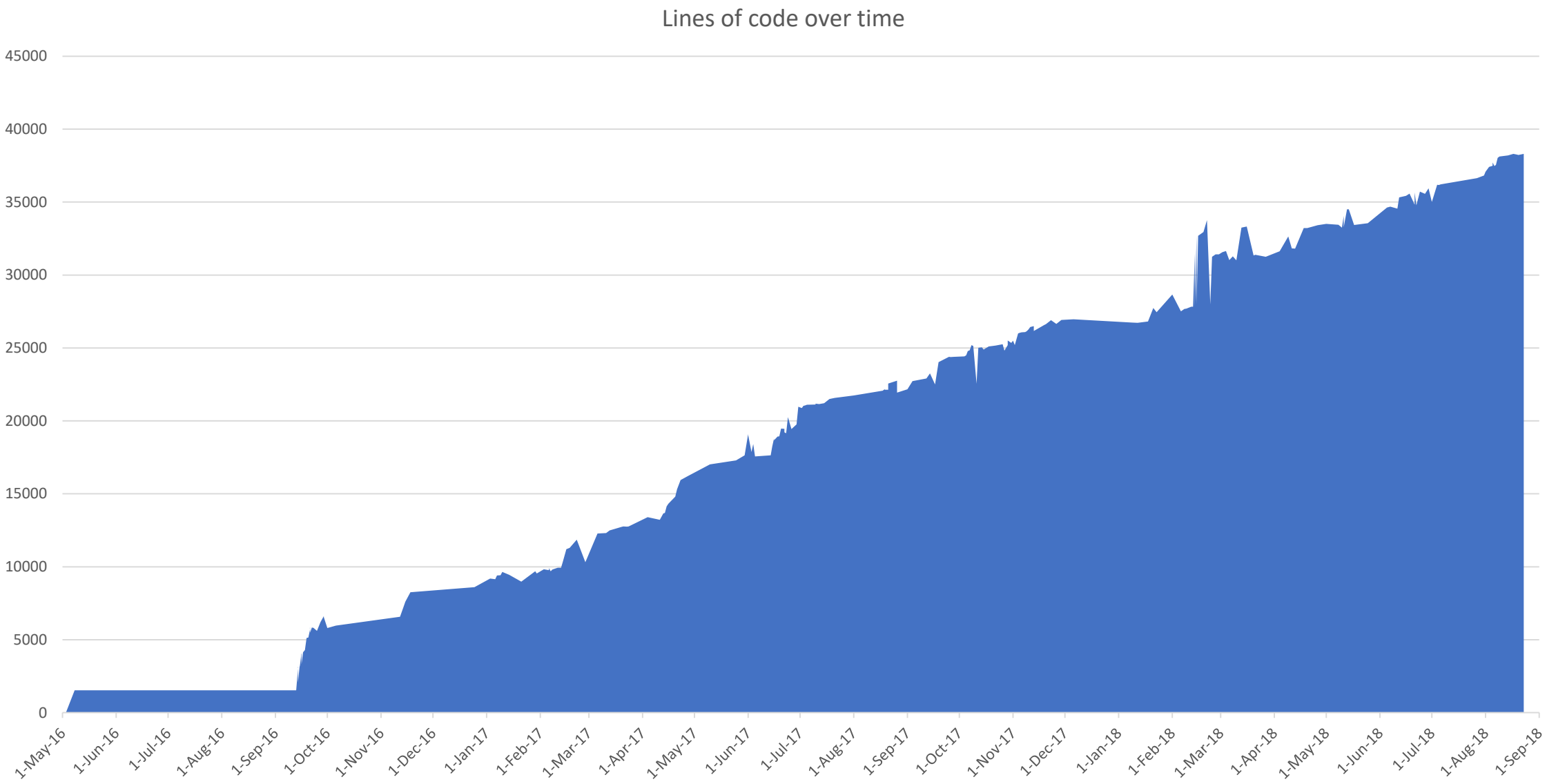
Marcin Junczys-Dowmunt

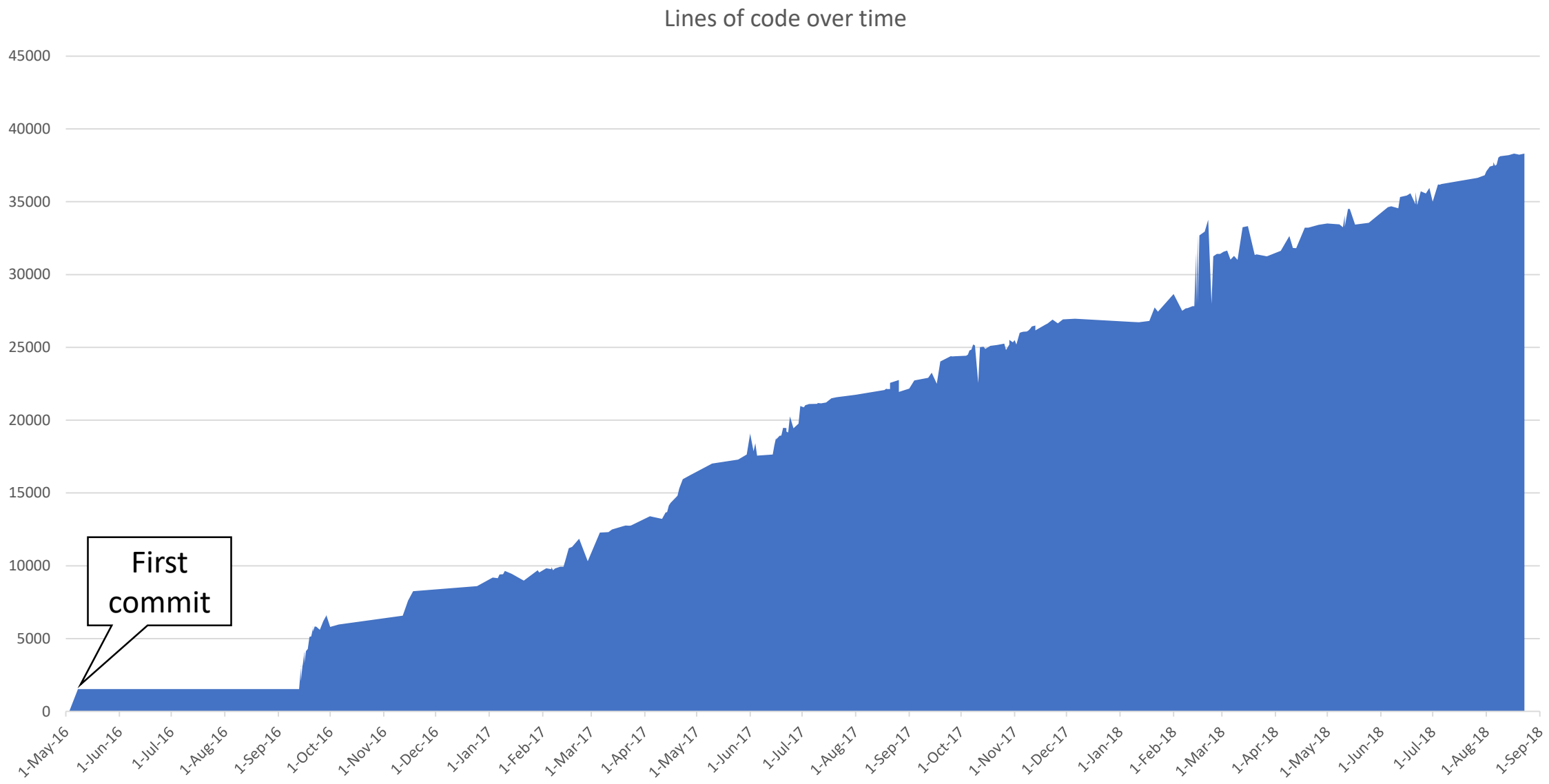
Translator

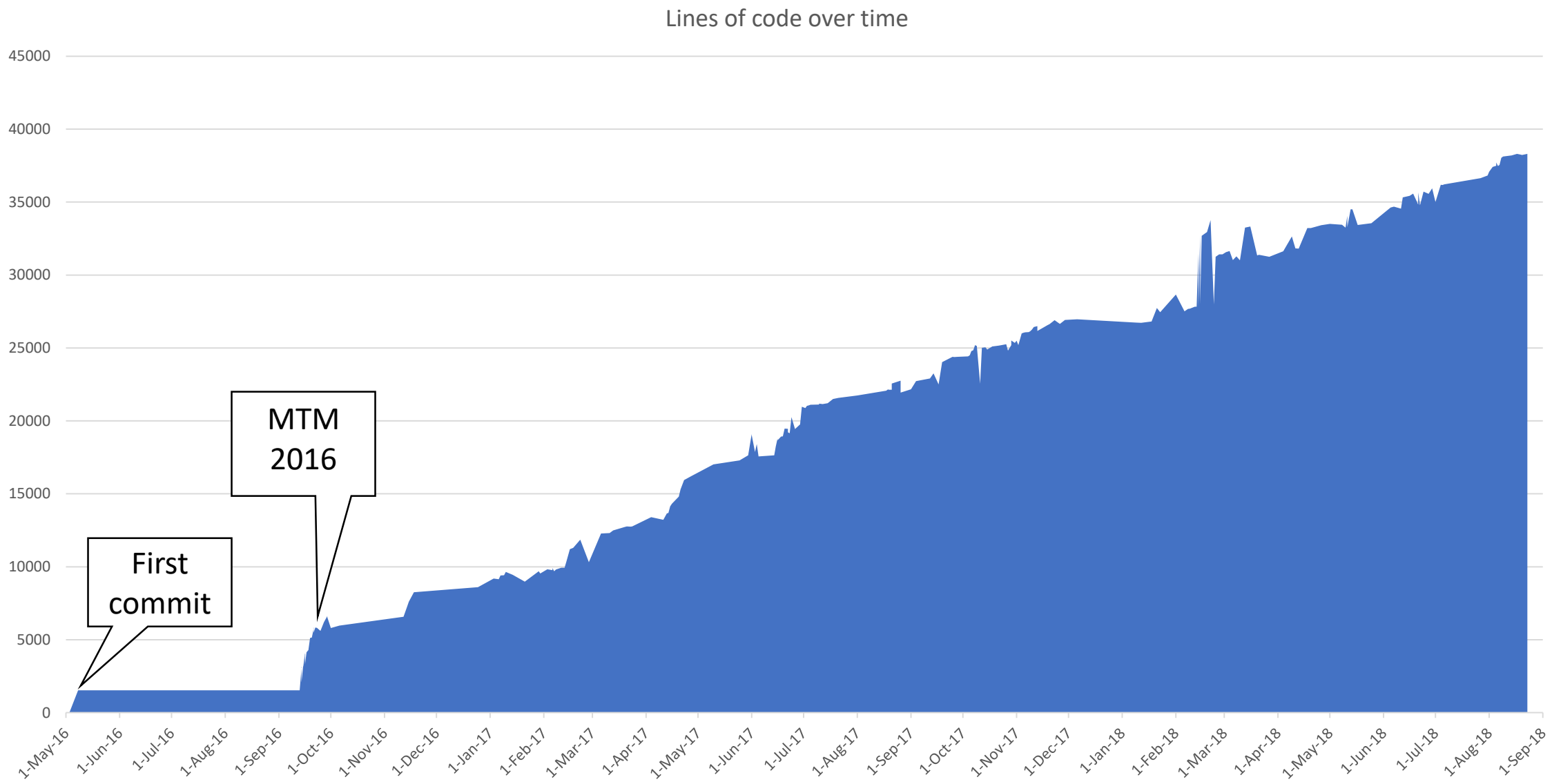
Marian Rejewski

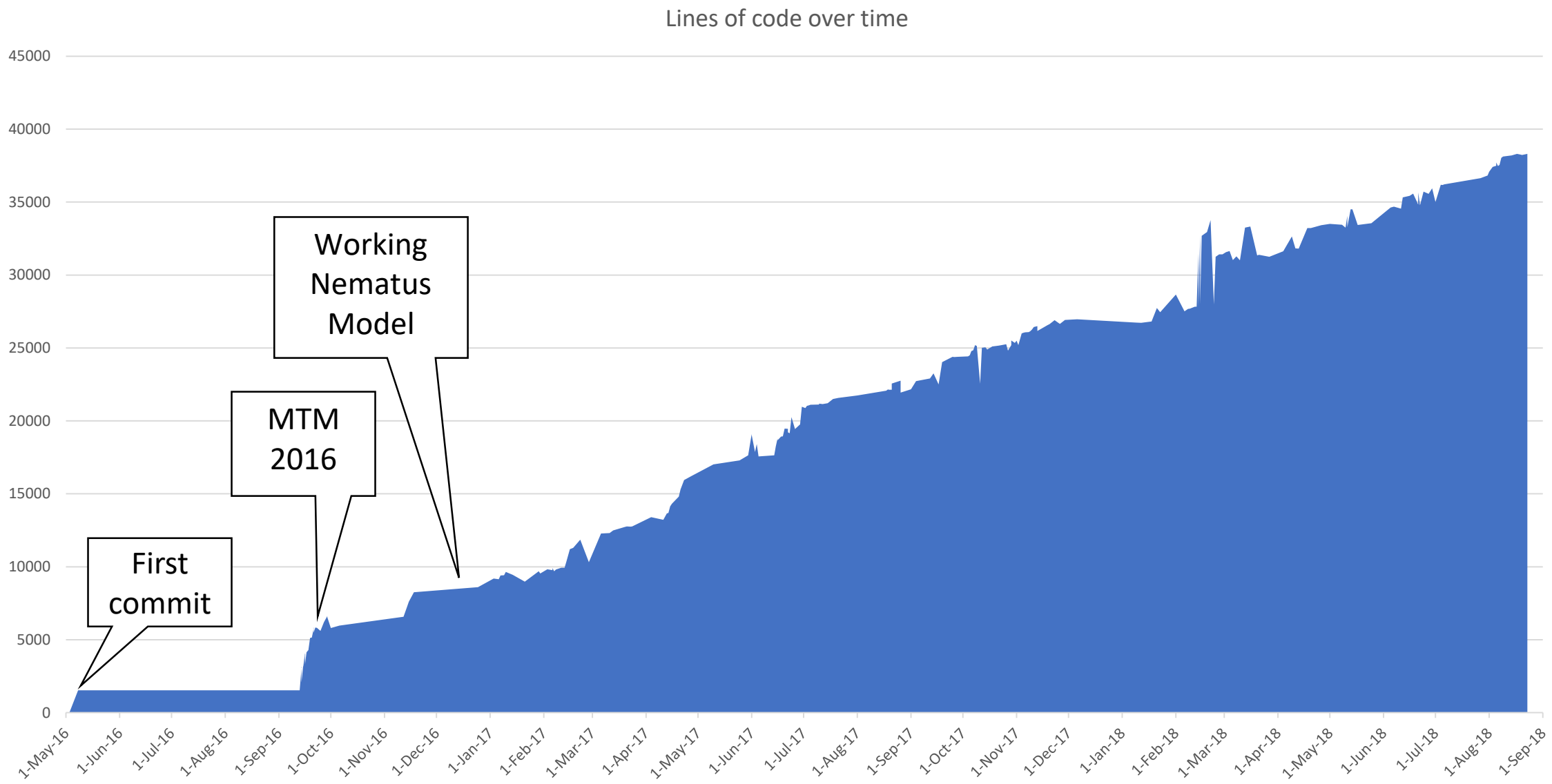


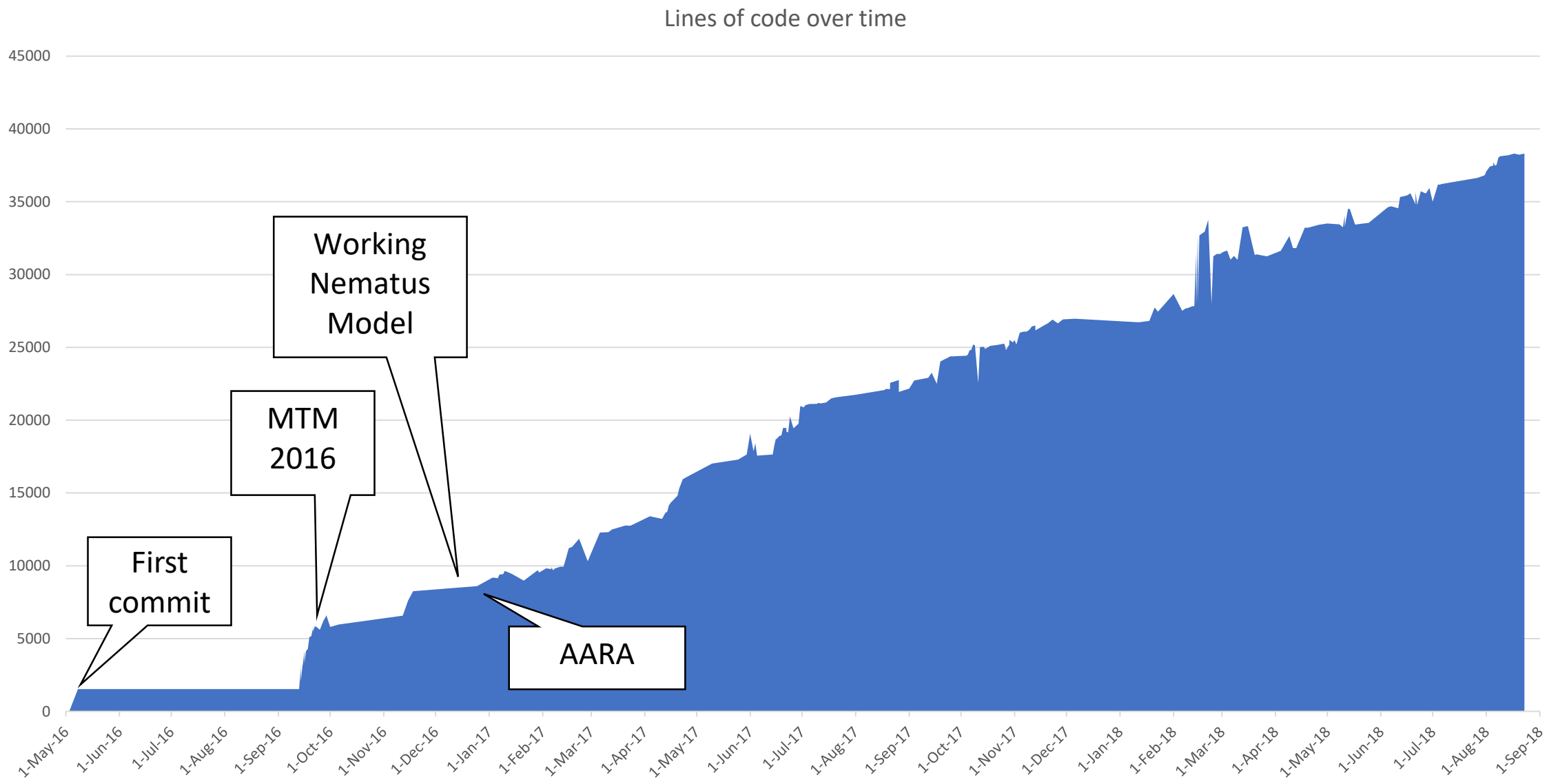


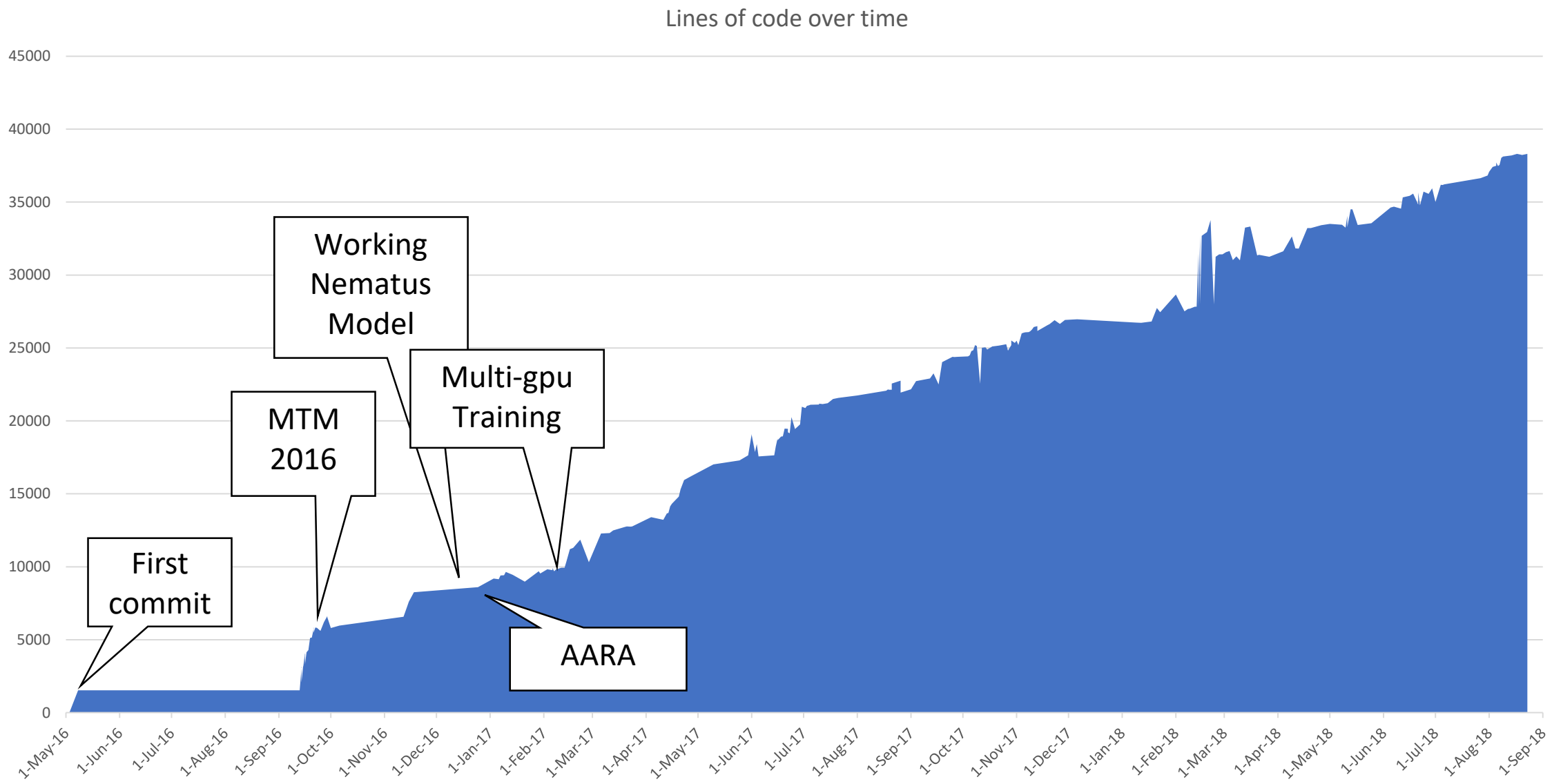

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

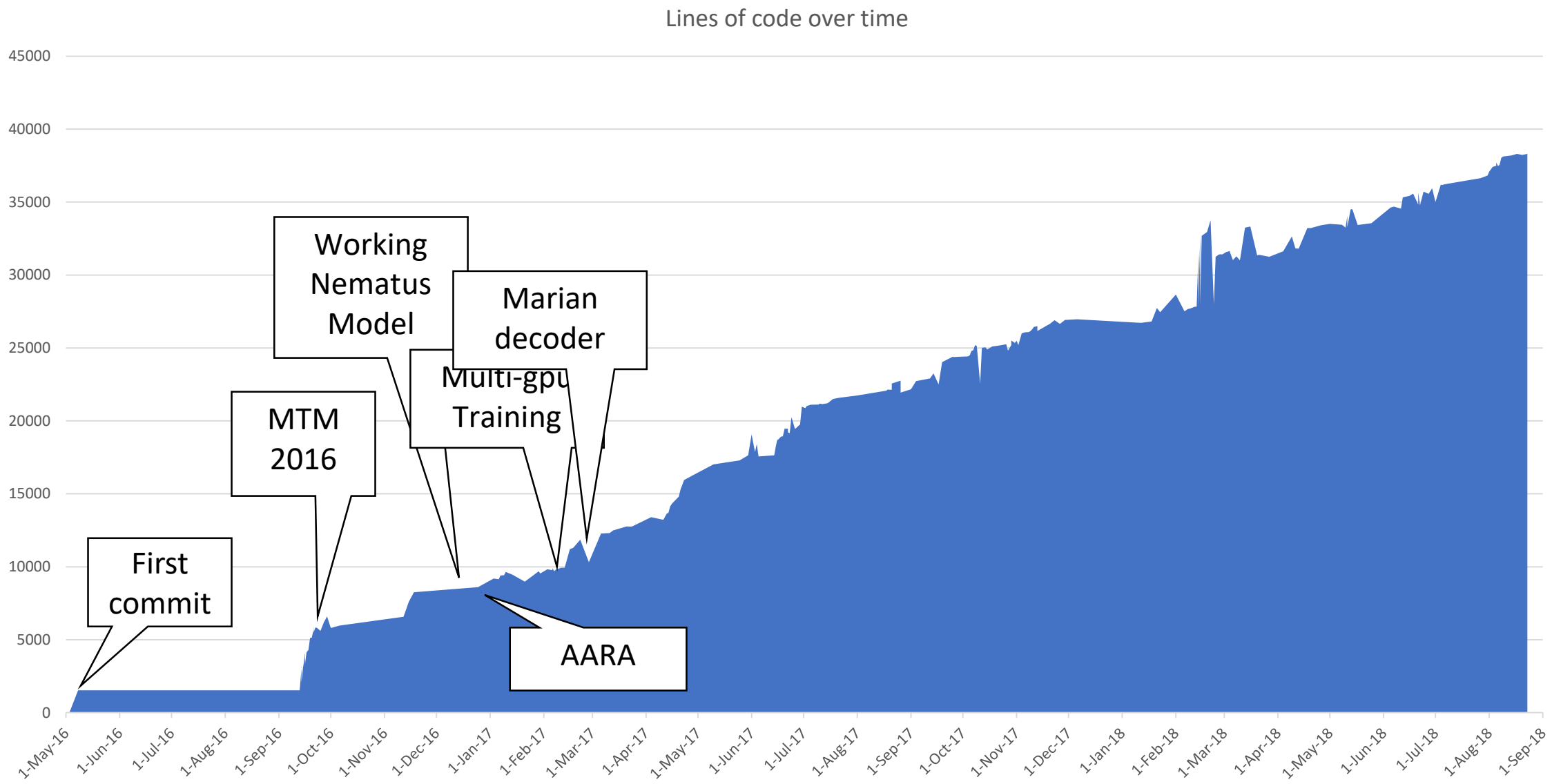

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
First commit
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

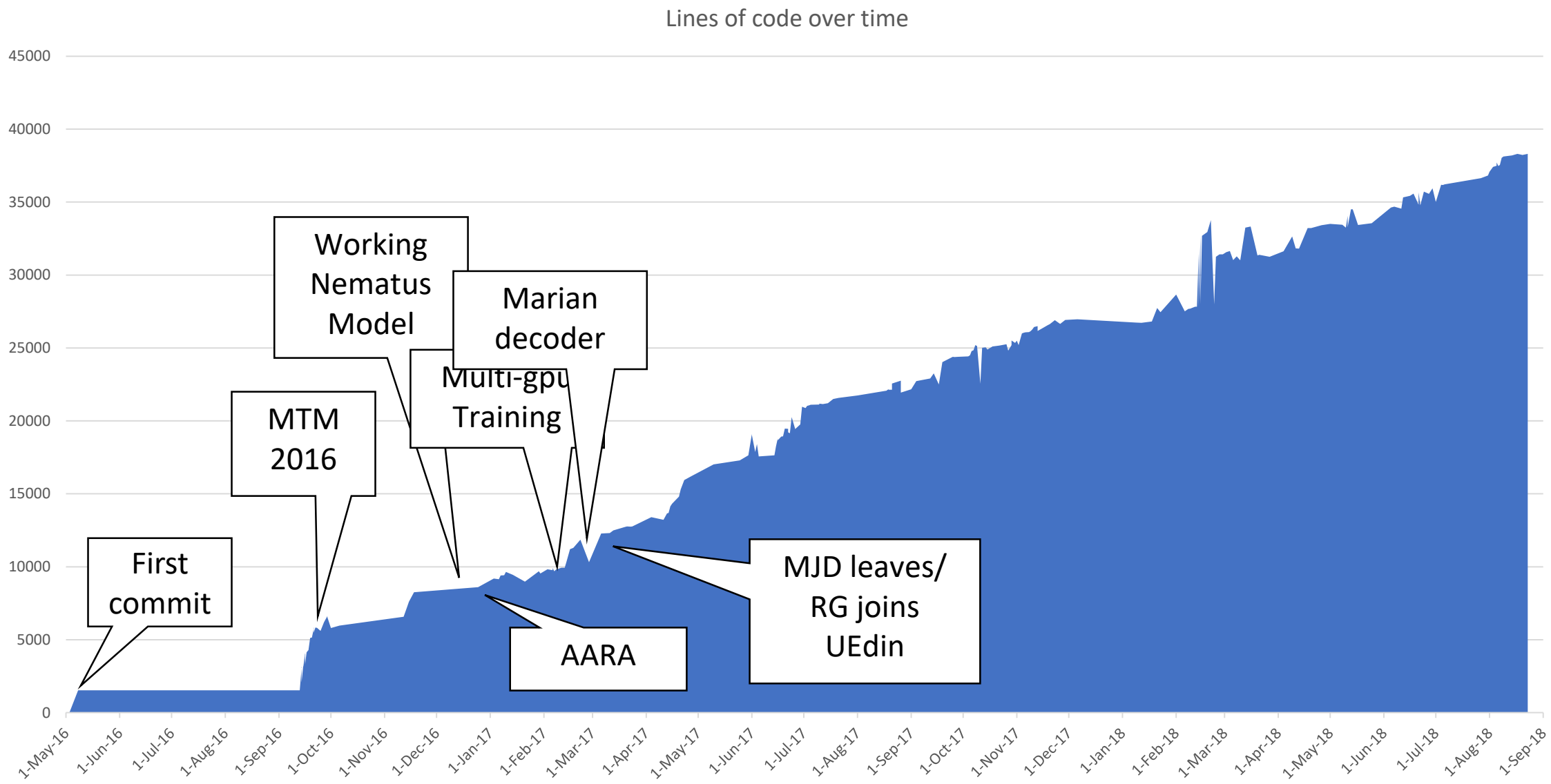

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
First commit
MTM 2016
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

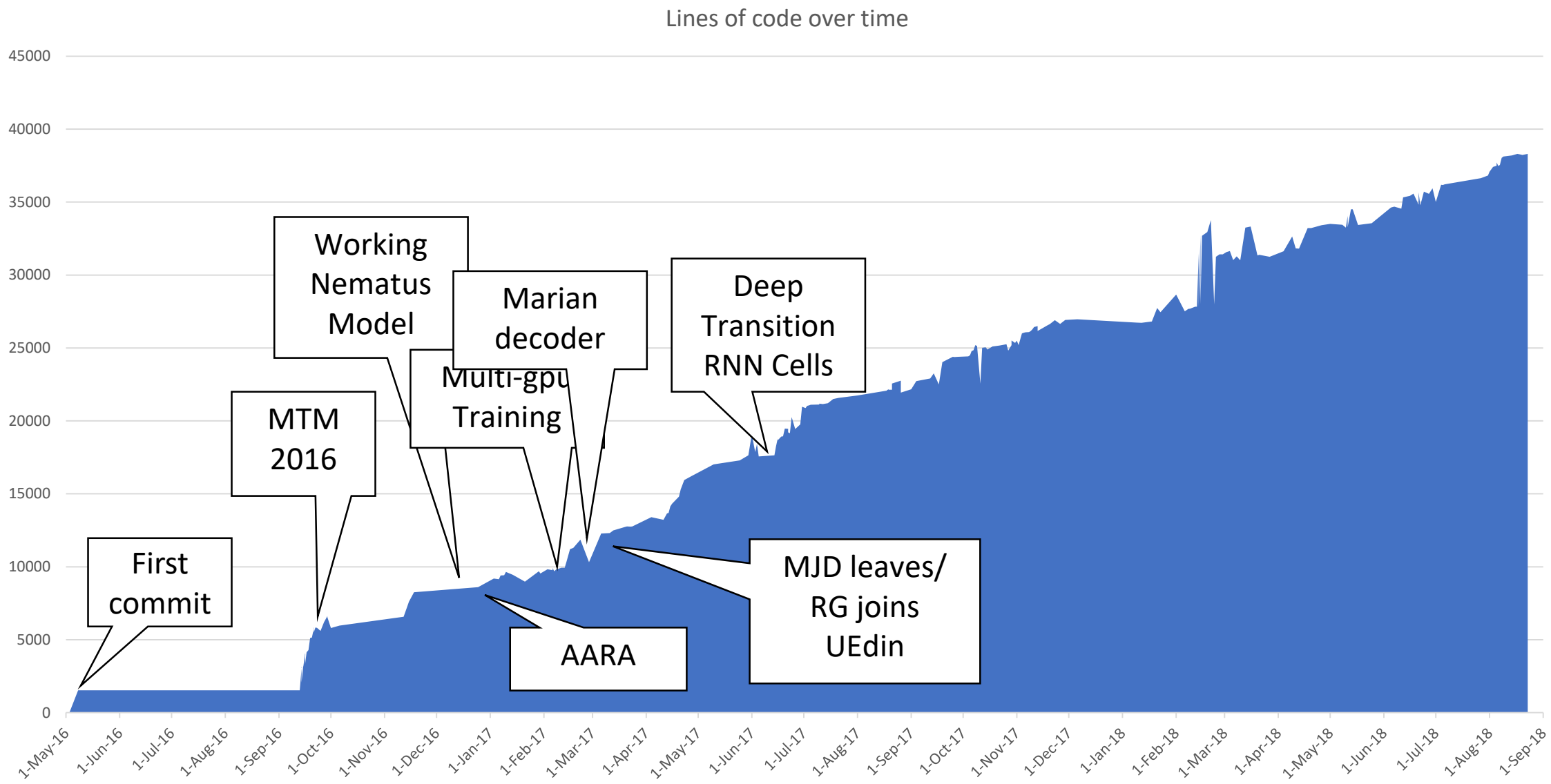

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
First commit
MTM 2016
Working Nematus Model
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

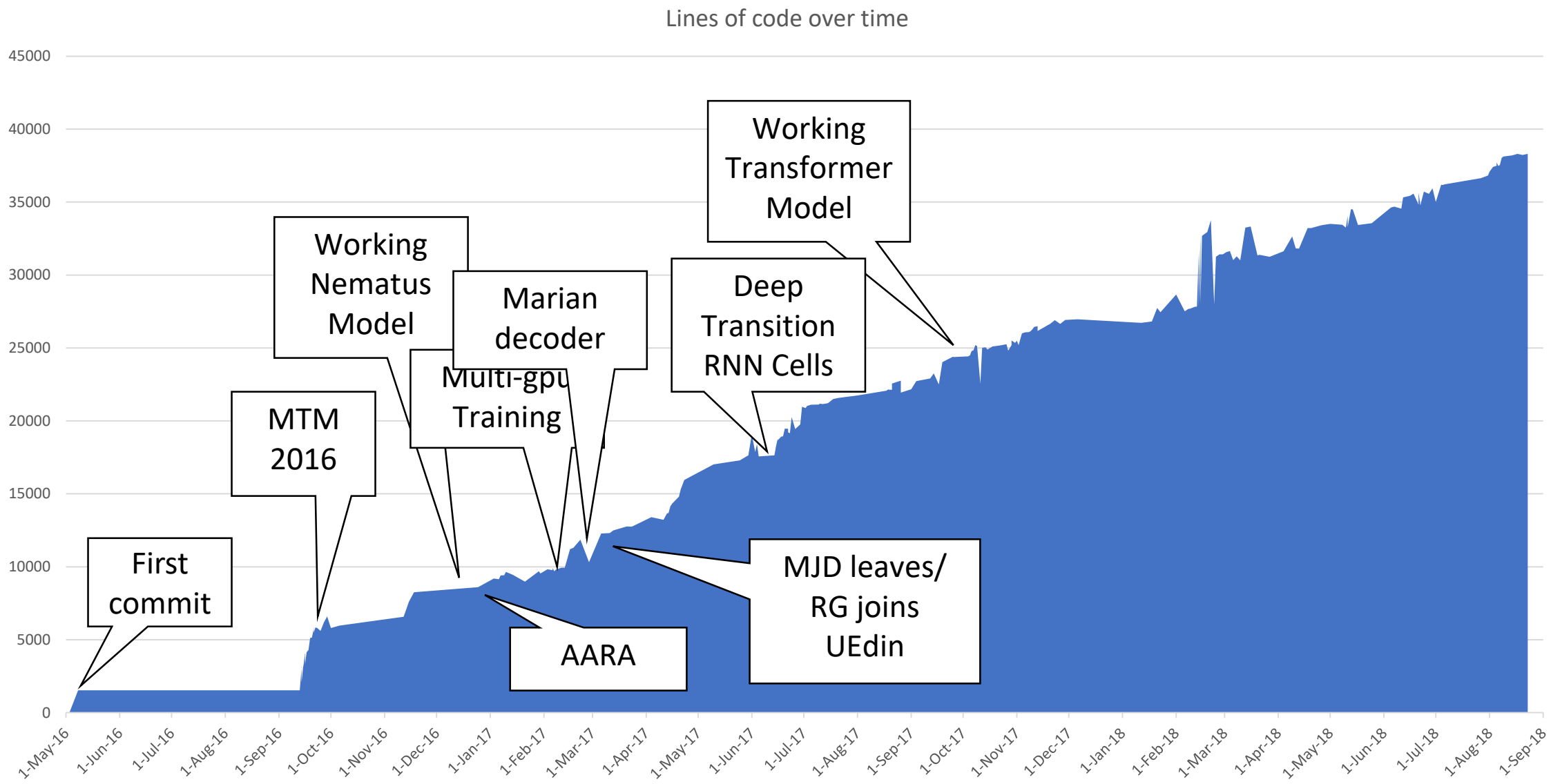

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
First commit
MTM 2016
Working Nematus Model
AARA
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

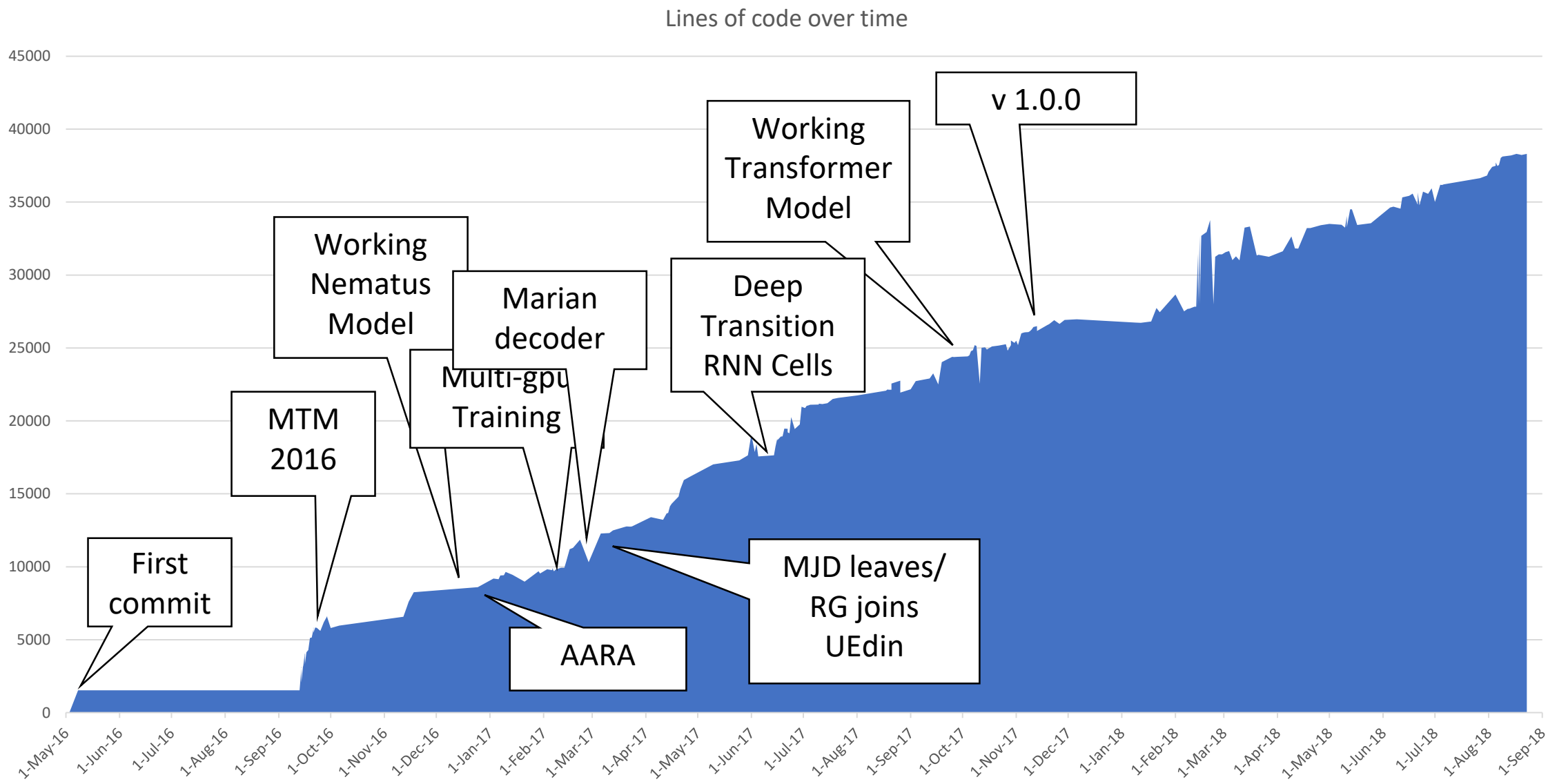

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
AARA
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18

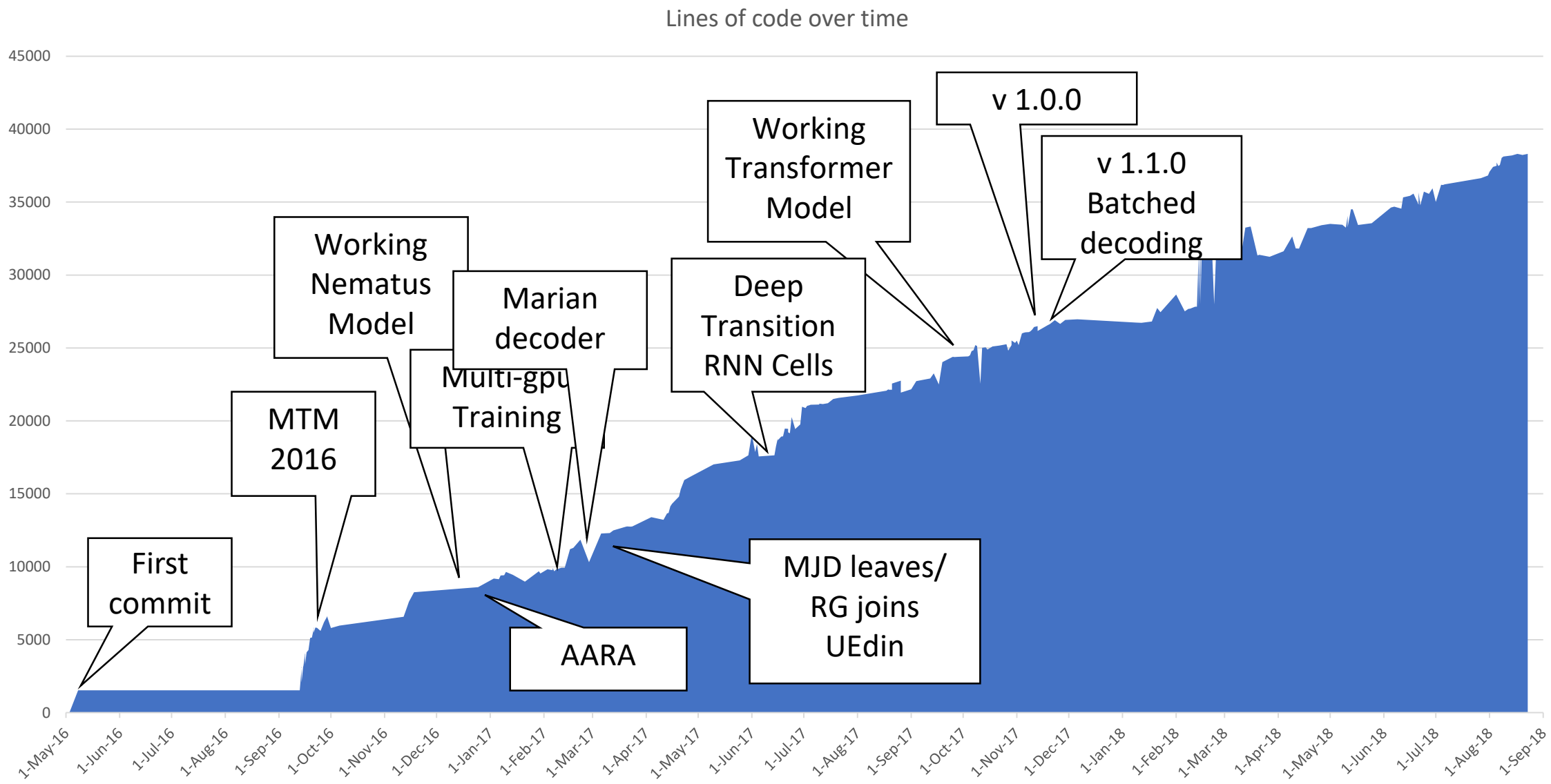

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA

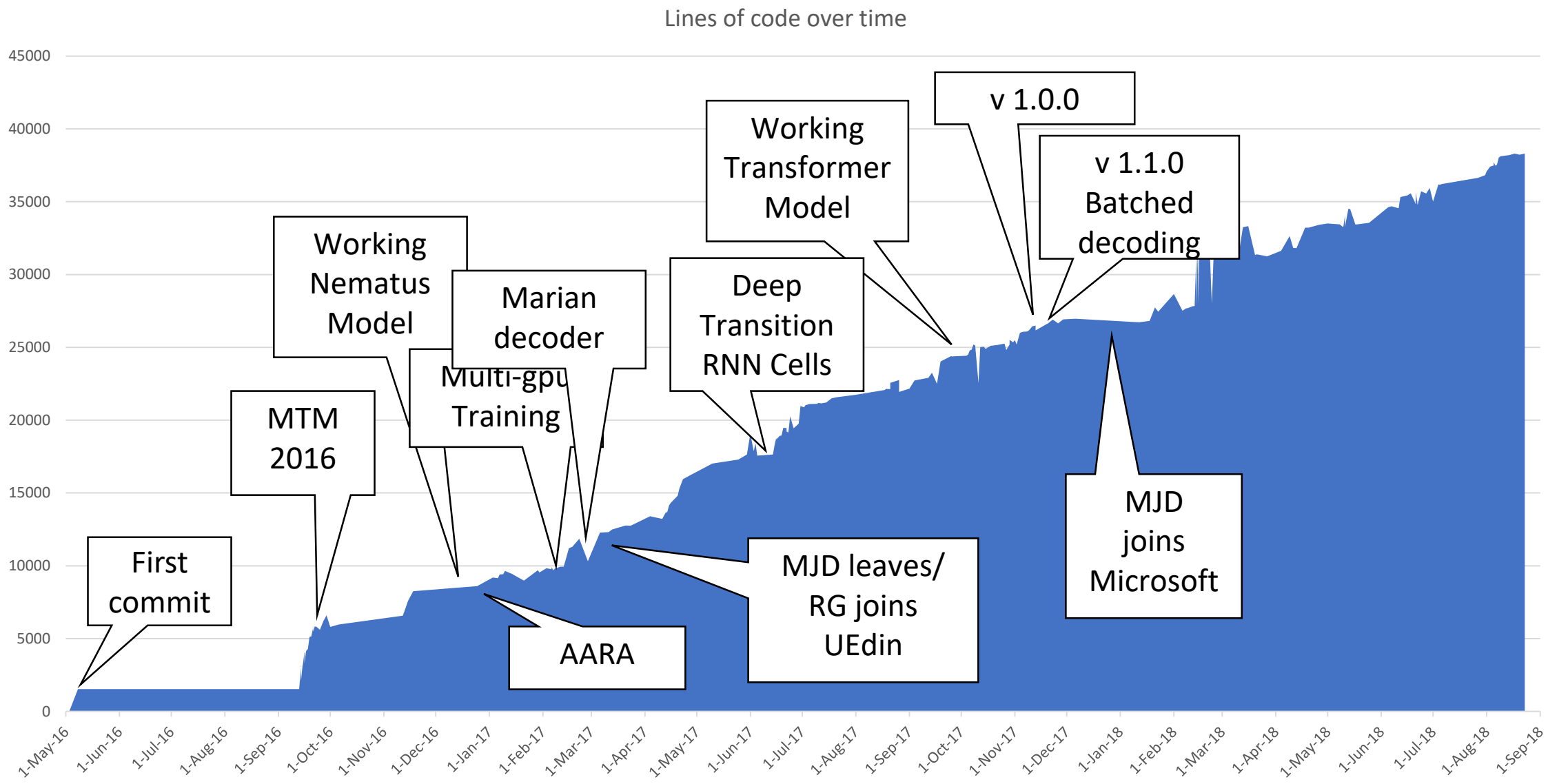

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
MJD leaves/ RG joins UEdin

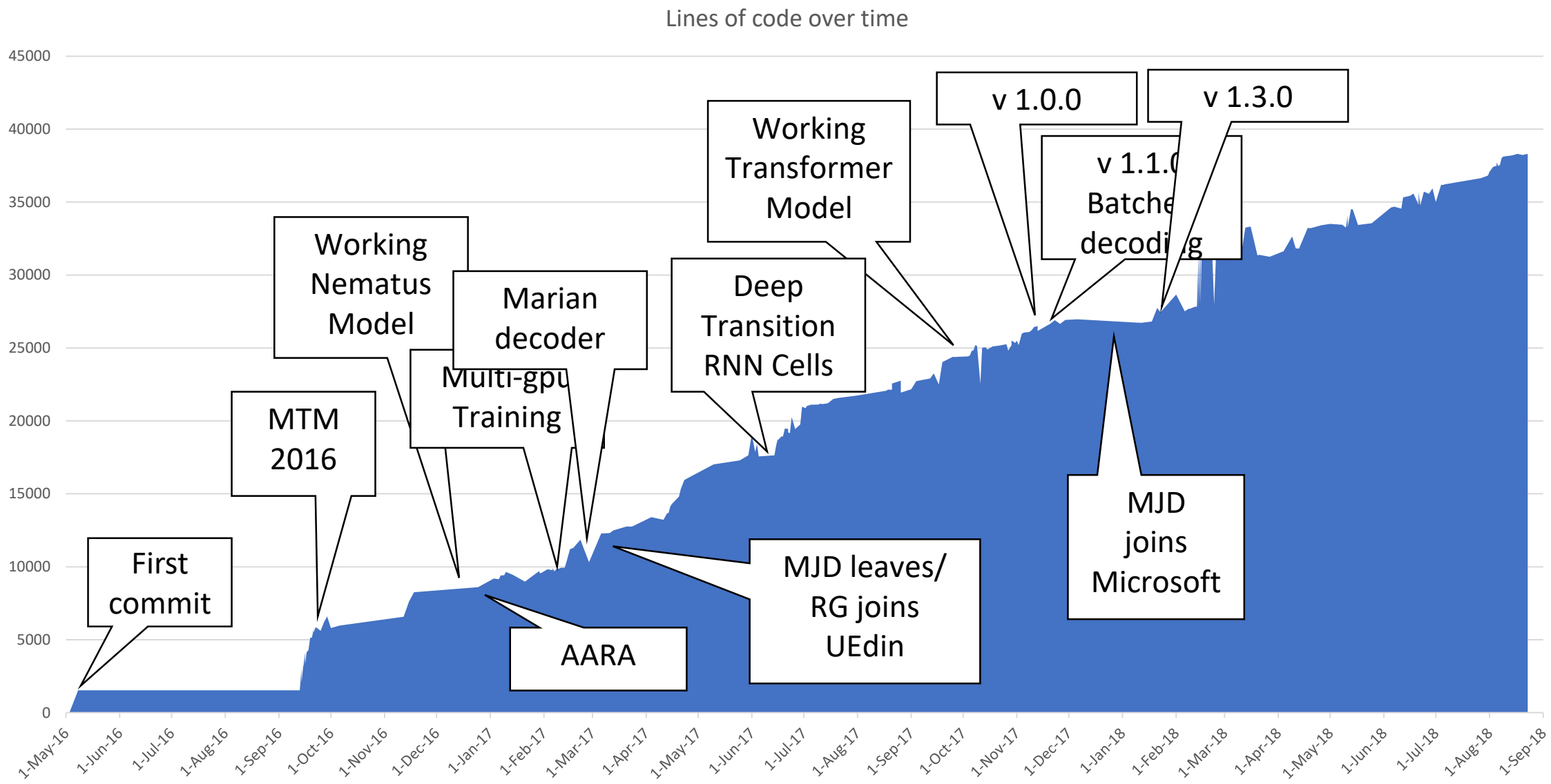

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
MJD leaves/ RG joins UEdin

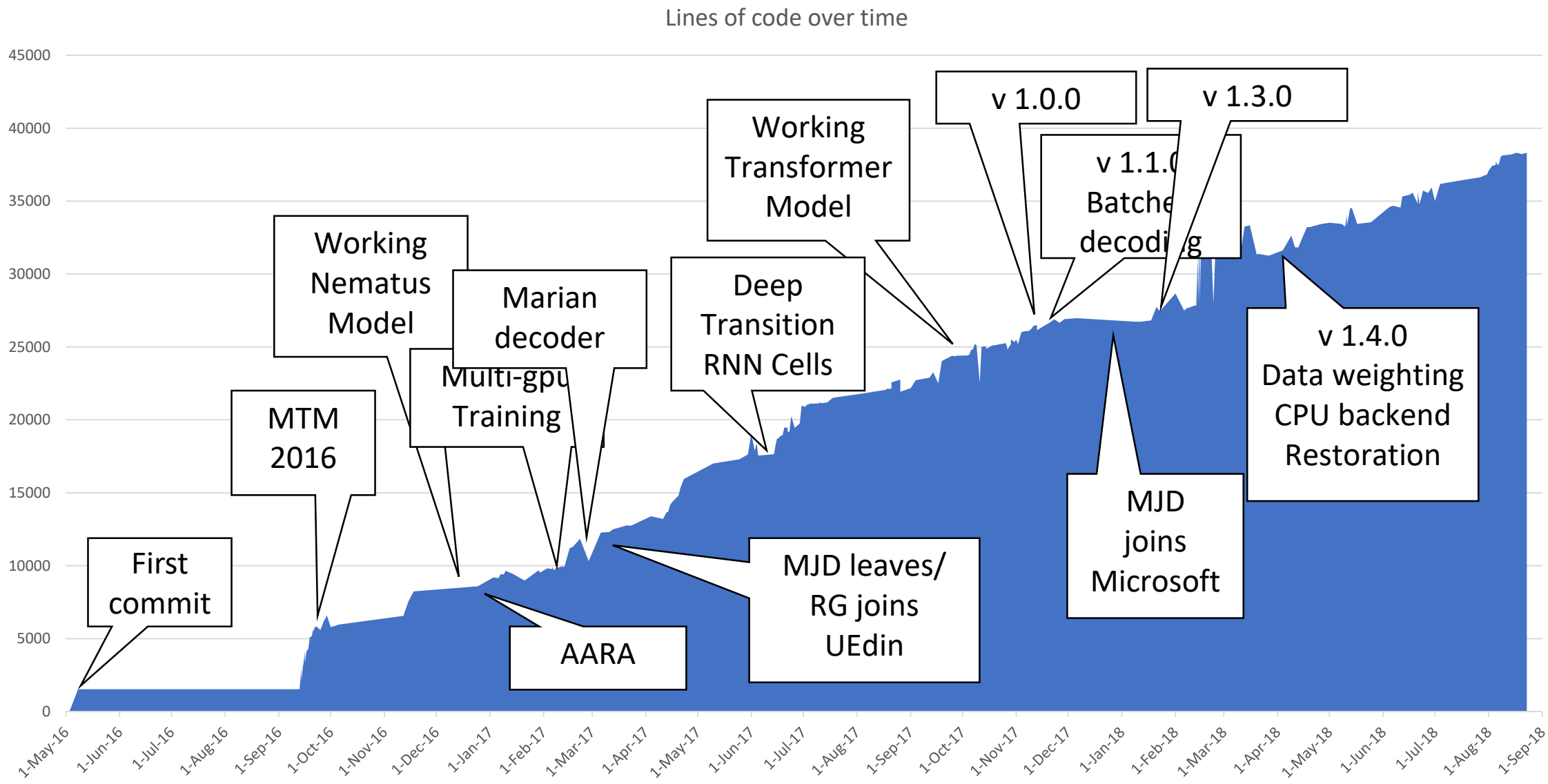

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin

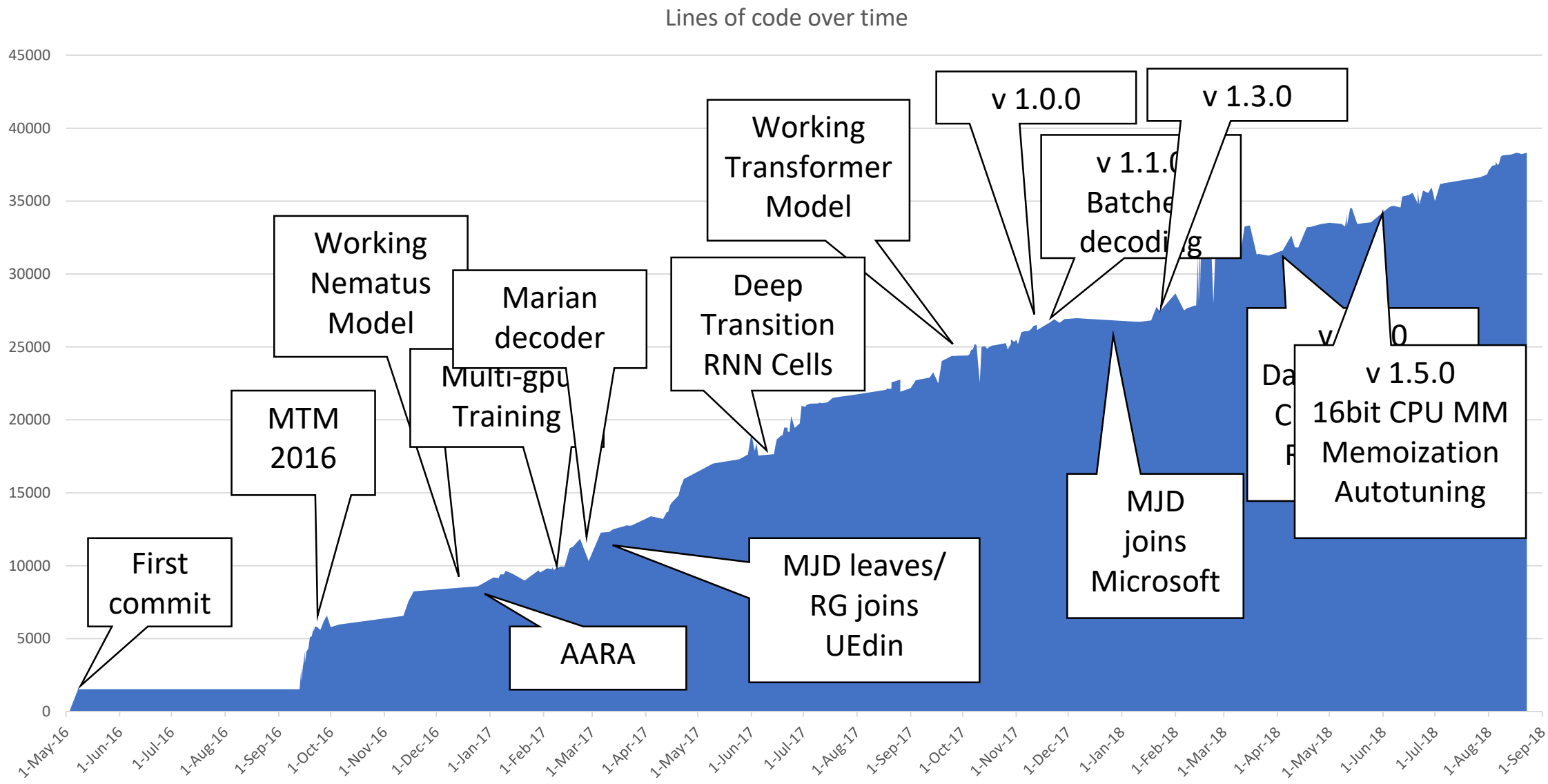

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
v 1.0.0
MJD leaves/ RG joins UEdin

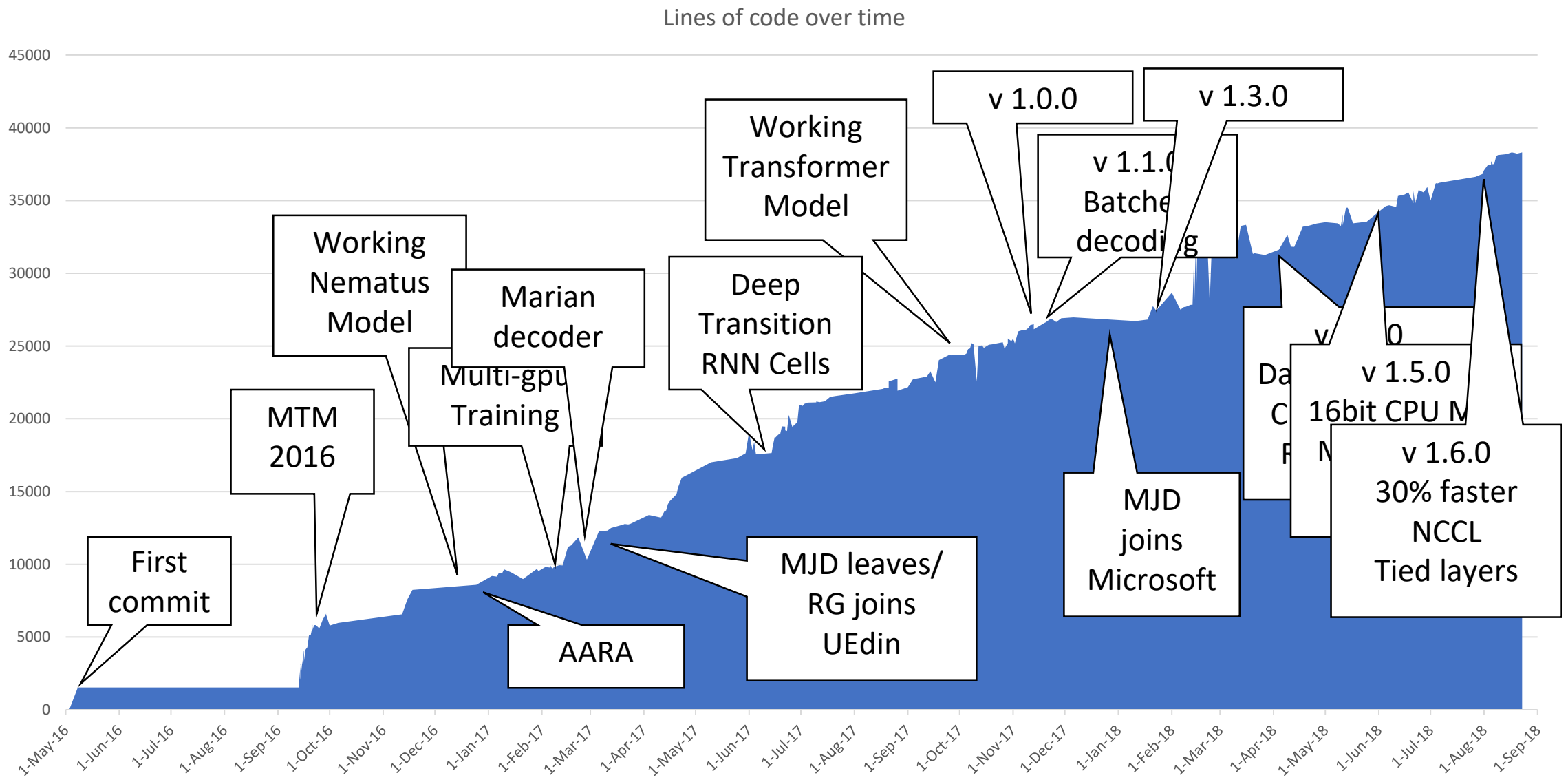

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0 Batched decoding

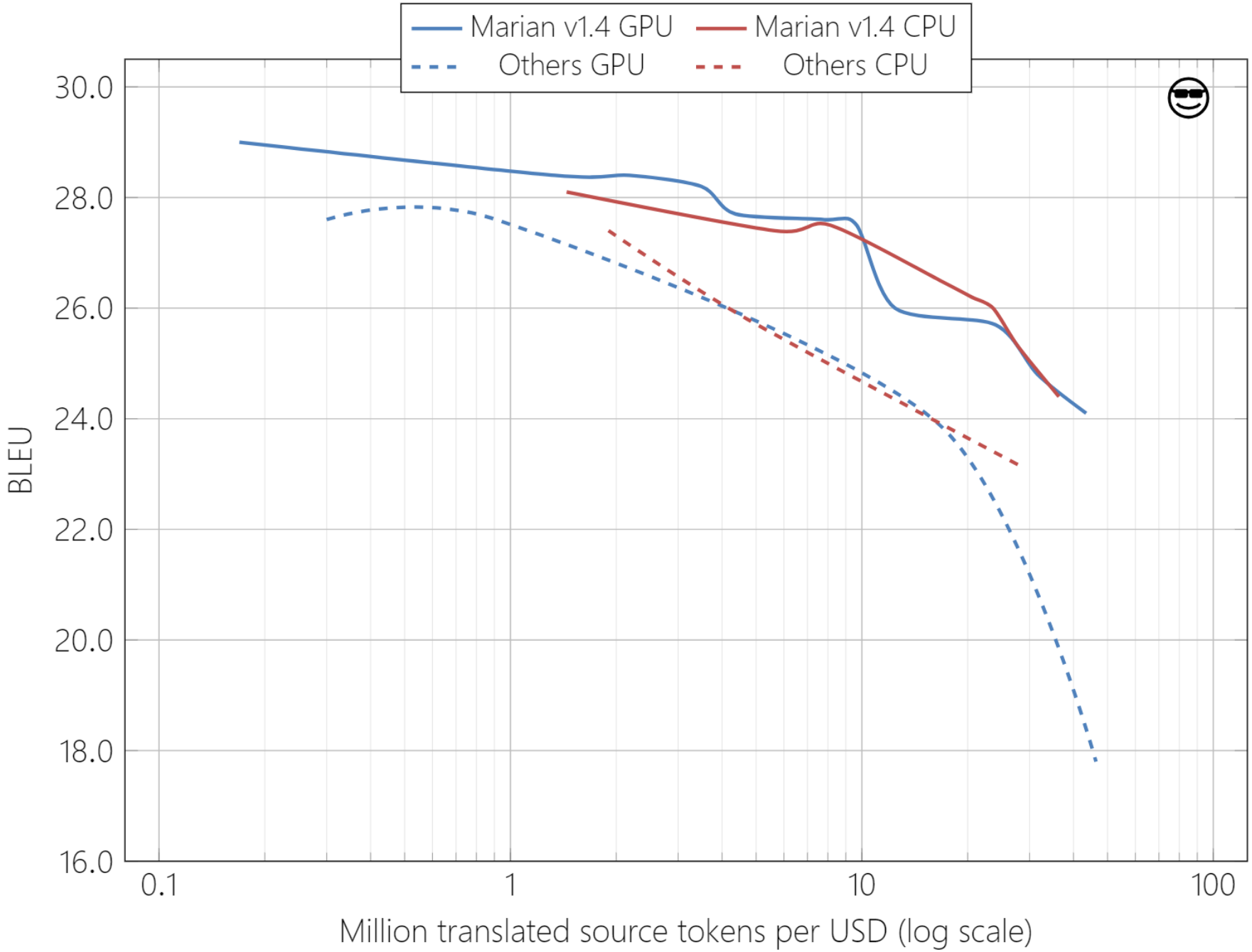

Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0 Batched decoding
MJD joins Microsoft


Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0 Batched decoding
v 1.3.0
MJD joins Microsoft


Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0 Batched decoding
v 1.3.0
MJD joins Microsoft
v 1.4.0 Data weighting CPU backend Restoration


Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Multi-gpu Training
Marian decoder
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0 Batched decoding
v 1.3.0
MJD joins Microsoft
v 1.5.0 16bit CPU MM Memoization Autotuning


Lines of code over time
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1-May-16
1-Jun-16
1-Jul-16
1-Aug-16
1-Sep-16
1-Oct-16
1-Nov-16
1-Dec-16
1-Jan-17
1-Feb-17
1-Mar-17
1-Apr-17
1-May-17
1-Jun-17
1-Jul-17
1-Aug-17
1-Sep-17
1-Oct-17
1-Nov-17
1-Dec-17
1-Jan-18
1-Feb-18
1-Mar-18
1-Apr-18
1-May-18
1-Jun-18
1-Jul-18
1-Aug-18
1-Sep-18
First commit
MTM 2016
Working Nematus Model
Marian decoder
Multi-gpu Training
AARA
Deep Transition RNN Cells
Working Transformer Model
MJD leaves/ RG joins UEdin
v 1.0.0
v 1.1.0
Batched decoding
v 1.3.0
MJD joins Microsoft
v 1.5.0
16bit CPU M
v 1.6.0
30% faster
NCCL
Tied layers


Marian v1.4 GPU
Marian v1.4 CPU
Others GPU
Others CPU
30.0
28.0
26.0
24.0
22.0
20.0
18.0
16.0
BLEU
0.1
1
10
100
Million translated source tokens per USD (log scale)

System Output List

English-German newstest2018

Interested in Contributing?

- Check out available resources.
- Create an account and start submitting your own systems.

Scored Systems

System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
Marian-Transformer *(Details)*	marcinjd *Microsoft*	Marian Transformer-Big	yes	Transformer-big ensemble x4. With back-translation, data-filtering on Paracrawl with domain-weighting. Decoder-time ensembling with transformer-LM, right-to-left decoding.	48.9	**48.3**	0.407	0.697	0.362
NMT-SMT Hybrid *(Details)*	fstahlberg *University of Cambridge*	MBR-based combination of neural models and SMT	yes		47.1	46.6	0.415	0.691	0.369
NTT Transformer-based System *(Details)*	makoto-mr *NTT*	Based on Transformer Big model. Trained with filtered version of CommonCrawl, ParaCrawl and synthetic corpus of newscrawl2017. R2L reranking.	yes		47.0	46.5	0.426	0.688	0.370
Primary Submission *(Details)*	pianist *Karlsruhe Institute of Technology*	Primary Submission	yes		46.9	46.3	0.428	0.687	0.382
MMT-unconstraint *(Details)*	nicolabertoldi *MMT srl*		no		46.7	46.2	0.432	0.682	0.387
JHU *(Details)*	jhu-nmt *Johns Hopkins University*	Marian Deep RNN	yes	Marian deep model, ensemble of 4 runs using base data (without Paracrawl), re-back-translated news 2016. Not final system yet.	43.6	43.0	0.453	0.670	0.394
MMT-constraint-dec *(Details)*	nicolabertoldi *MMT srl*		yes		42.9	42.5	0.463	0.667	0.411
NJUNMT *(Details)*	ZhaoChengqi *Nanjing University*	transformer base without back translation	yes	transformer base without back translation	40.6	40.0	0.496	0.647	0.436
LMU-unsupervised-nmt-wmt18-en-de *(Details)*	Matthias.Huck *LMU Munich*	Unsupervised NMT (no parallel training corpora)	yes		15.8	15.5	0.762	0.500	*failed*
RWTH Unsupervised NMT Ensemble *(Details)*	yunsukim *RWTH Aachen University*	(Unsupervised) Transformer with shared encoder/decoder, separate top-50k word	yes		15.9	14.8	0.753	0.514	0.607

System Output List

English-German newstest2018

Interested in Contributing?

- Check out available resources.
- Create an account and start submitting your own systems.

Scored Systems

System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
Marian-Transformer *(Details)*	marcinjd *Microsoft*	Marian Transformer-Big	yes	Transformer-big ensemble x4. With back-translation, data-filtering on Paracrawl with domain-weighting. Decoder-time ensembling with transformer-LM, right-to-left decoding.	48.9	**48.3**	0.407	0.697	0.362
NMT-SMT Hybrid *(Details)*	fstahlberg *University of Cambridge*	MBR-based combination of neural models and SMT	yes		47.1	46.6	0.415	0.691	0.369
NTT Transformer-based System *(Details)*	makoto-mr *NTT*	Based on Transformer Big model. Trained with filtered version of CommonCrawl, ParaCrawl and synthetic corpus of newscrawl2017. R2L reranking.	yes		47.0	46.5	0.426	0.688	0.370
Primary Submission *(Details)*	pianist *Karlsruhe Institute of Technology*	Primary Submission	yes		46.9	46.3	0.428	0.687	0.382
MMT-unconstraint *(Details)*	nicolabertoldi *MMT srl*		no		46.7	46.2	0.432	0.682	0.387
JHU *(Details)*	jhu-nmt *Johns Hopkins University*	Marian Deep RNN	yes	Marian deep model, ensemble of 4 runs using base data (without Paracrawl), re-back-translated news 2016. Not final system yet.	43.6	43.0	0.453	0.670	0.394
MMT-constraint-dec *(Details)*	nicolabertoldi *MMT srl*		yes		42.9	42.5	0.463	0.667	0.411
NJUNMT *(Details)*	ZhaoChengqi *Nanjing University*	transformer base without back translation	yes	transformer base without back translation	40.6	40.0	0.496	0.647	0.436
LMU-unsupervised-nmt-wmt18-en-de *(Details)*	Matthias.Huck *LMU Munich*	Unsupervised NMT (no parallel training corpora)	yes		15.8	15.5	0.762	0.500	*failed*
RWTH Unsupervised NMT Ensemble *(Details)*	yunsukim *RWTH Aachen University*	(Unsupervised) Transformer with shared encoder/decoder, separate top-50k word	yes		15.9	14.8	0.753	0.514	0.607

James Bradbury
@jekbradbury

Following

Facebook's fairseq MT engine is really, really fast... Like, 50% faster than @marian_nmt (which is itself way faster than Sockeye/OpenNMT/Tensor2Tensor/xnmt/Nematus/etc) at generating from the same Transformer model facebook.com/61013326/posts ...

2:24 PM - 15 Jun 2018

37 Retweets **139** Likes

2 37 139

Tweet your reply

Marian NMT @marian_nmt · Jun 15

Replying to @jekbradbury

Hold my beer ;)

1 13

James Bradbury
@jekbradbury

Following

Facebook's fairseq MT engine is really, really fast... Like, 50% faster than @marian_nmt (which is itself way faster than Sockeye/OpenNMT/Tensor2Tensor/xnmt/Nematus/etc) at generating from the same Transformer model
facebook.com/61013326/posts ...

2:24 PM - 15 Jun 2018

37 Retweets **139** Likes

2 37 139

Tweet your reply

Marian NMT @marian_nmt · Jun 15
Replying to @jekbradbury
Hold my beer ;)

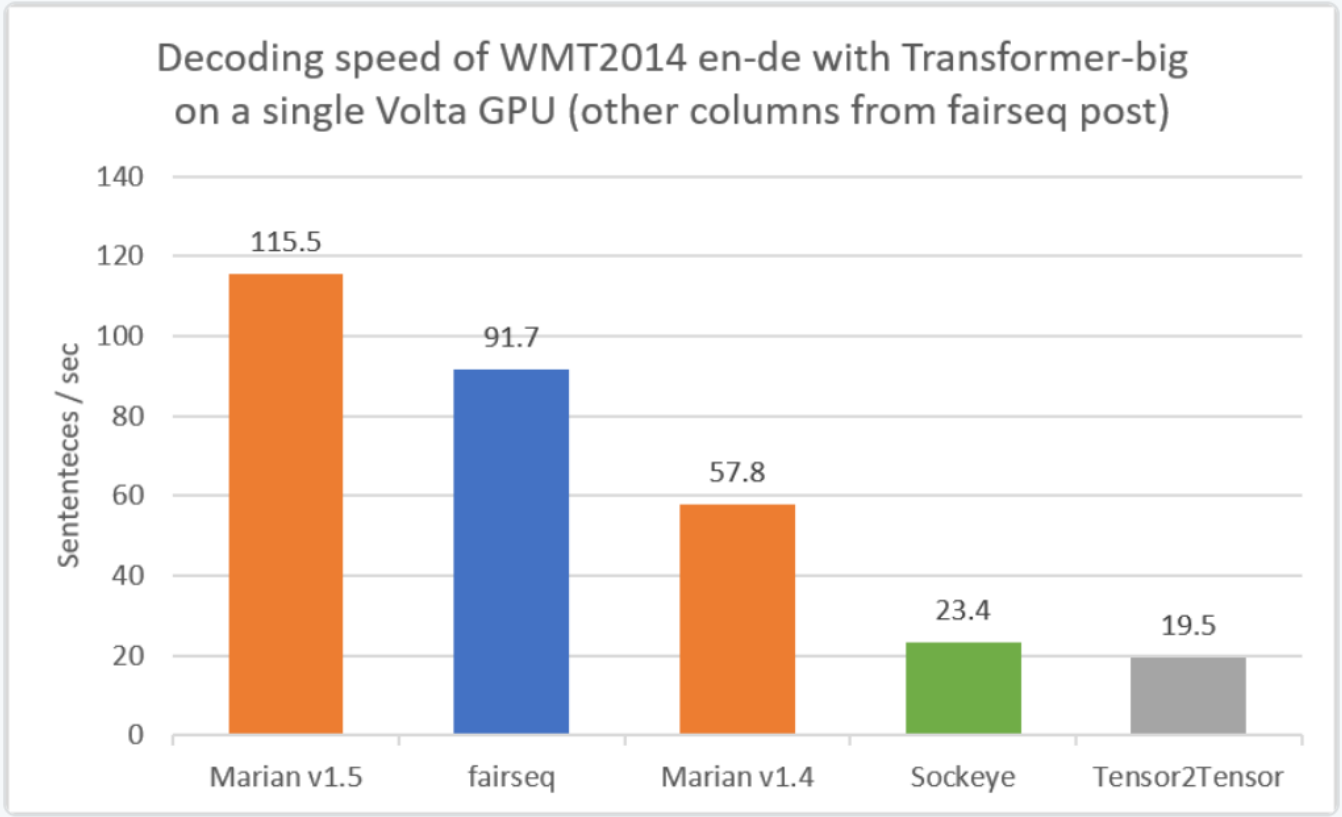
1 13

Marian NMT @marian_nmt · Jun 17

Boom! Marian v1.5.0 released.

Includes:
- Extensions from the WNMT shared task on efficiency arxiv.org/abs/1805.12096
- Optimized GPU-decoding for Transformer models.

See chart below for speed comparison to v1.4.0 (based on FAIR's post)
@jekbradbury @alvations



4 33 108