Supplementary material

1 Normalized Levenshtein Distance

The Levenshtein distance is a distance metric between two DNA or protein sequences. It measures the minimum number of single character (e.g., nucleotide) edits require to change one sequence to another. Each edit is either an insertion, deletion or substitution. Computing the Levenshtein distance, requires finding the optimal alignment that requires the minimal number of edits, and therefore is considerably more expensive computationally compared with the Hamming distance. In Fig. S1, we illustrate the Levenshtein distance computation.

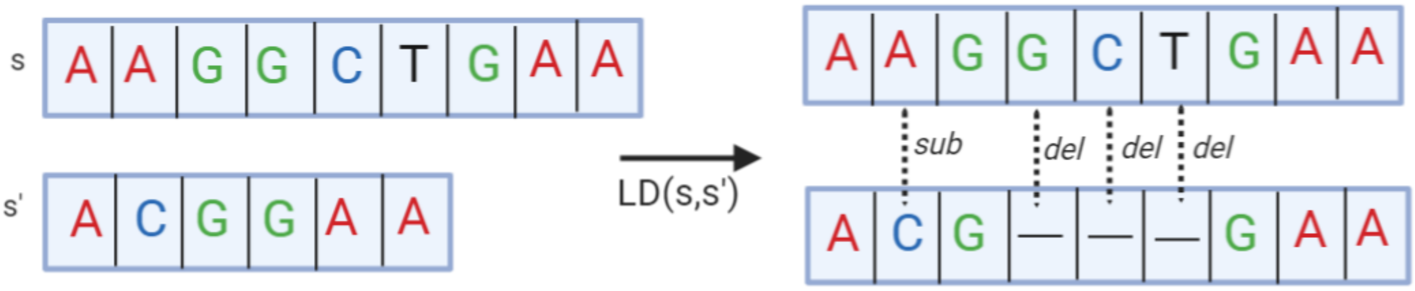


Figure S1: The Levensthein distance between sequence s and s' is determined by the minimum number of edits required to transition form one sequence to the other. In this example three deletion (d) and one substitution (s) are the minimal number of edits required, thus the Levenshtein distance is 4.

The Levenshtein distance allows us to compare two sequences with different lengths. To reduce bias caused by length differences, [Yujian and Bo, 2007] propose a normalized Levenshtein distance that incorporates the length of both sequences. The normalized Levenshtein distance is defined as

$$NLD(s,s') = \frac{2 \cdot LD(s,s')}{|s| + |s'| + LD(s,s')},$$

where $LD(s,s')$ denotes the Levenshtein distance and $|s|, |s'|$ are the lengths of sequence s and s' respectively.

2 A deeper look into clones with multiple V, J or junction length

In the main text, we have analyzed clones identified using the *alignment-free* method with non unique V or J gene assignments, or with multiple junction lengths. We have computed the histogram of the distance-to-nearest between sequences with such non unique properties. A schematic description of the distance-to-nearest evaluation of such clones appears in Fig. S2.

The histograms presented in Fig. 11 (main text) are based on the normalized Levenshtein distance metric applied to pairs of full sequences. This includes the V and J segments and the junction. Next, we repeat the procedure presented in Fig. S2, but apply the normalized Levenshtein distance only to the junction part of the sequence. In Fig. S3, we present the distributions of distance-to-nearest within clones with non unique characteristics and background groups. The distributions of clones with non unique V and J genes (3(a) and 3(b)) provide yet another support that the majority of these predicted clones are enriched with true clonal expansions.

Finally, in Fig. S4, we present a multi-sequence alignment of all the sequences in the clones which are presented in Fig. 12 (main text). The shared mutations in these alignment demonstrate that sequences with multiple junction length can belong to the same clone and likely result from true clonal expansions.

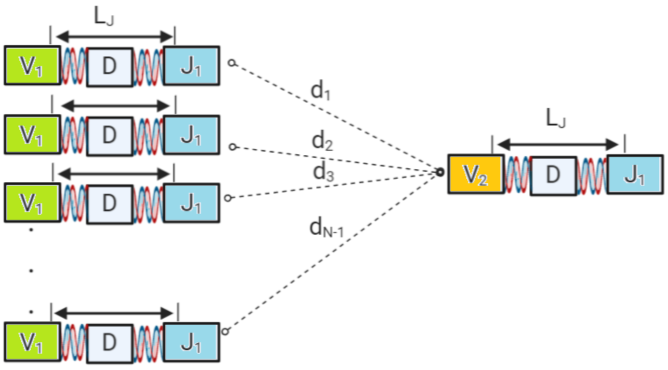


Figure S2: Here we illustrate the distance-to-nearest computation used in Fig. 11 (main text). The group of $N-1$ sequences (majority group) share the same V and J genes assignments and have the same junction length L_J. In this example, one of the sequences has a different V gene assignment (minority group). We compute all distances between the minority and majority groups $d_1, ..., d_{N-1}$ and compute the histogram of the minimal normalized Levenshtein distances $\min(d_1, ..., d_{N-1})$ over all clones. As a background distribution, we generate artificial groups of sequences that share the same V gene, J gene and junction characteristics, and include an additional sequence that differs by one of these characteristics. The number of sequences in each such artificial group is set using the actual sizes of the clones.

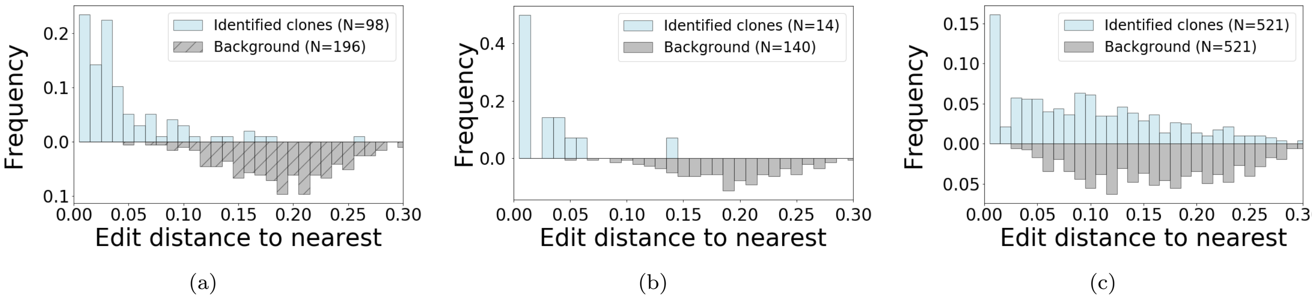


Figure S3: (a) Distribution of distance-to-nearest within clones containing sequences with non unique J genes. (b) Distribution of distance-to-nearest within clones containing sequences with non unique V genes. (c) Distribution of distance-to-nearest within clones containing sequences with non unique junction lengths.

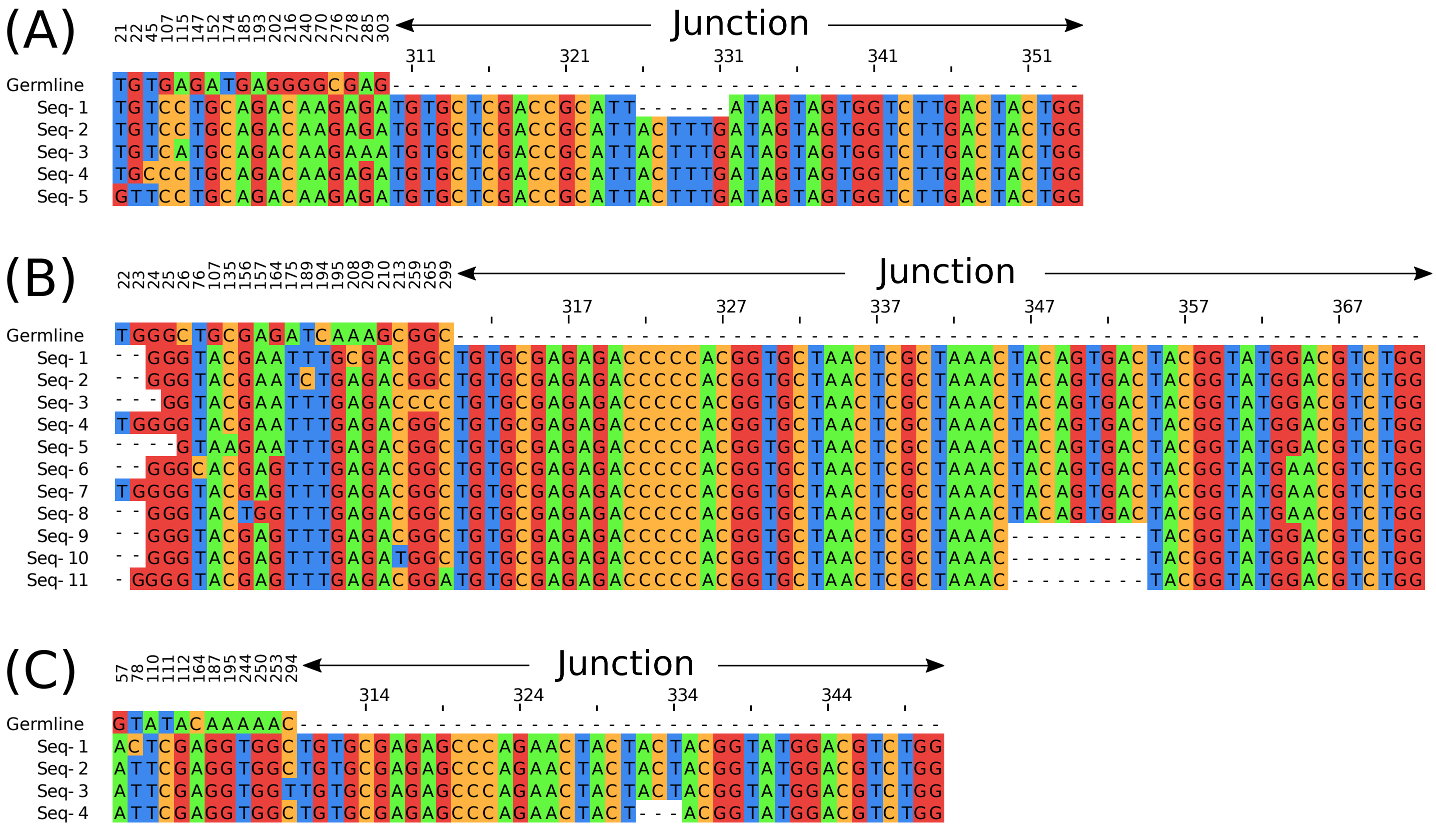

(A)
21 22 45 107 115 147 152 174 185 193 202 216 240 270 276 278 285 303
Junction
311
321
331
341
351
Germline TGTGAGATGAGGGGCGAG
Seq- 1 TGTCCTGCAGACAAGAGATGTGCTCGACCGCATT------ATAGTAGTGGTCTTGACTACTGG
Seq- 2 TGTCCTGCAGACAAGAGATGTGCTCGACCGCATTACTTTGATAGTAGTGGTCTTGACTACTGG
Seq- 3 TGTCATGCAGACAAGAAATGTGCTCGACCGCATTACTTTGATAGTAGTGGTCTTGACTACTGG
Seq- 4 TGCCCTGCAGACAAGAGATGTGCTCGACCGCATTACTTTGATAGTAGTGGTCTTGACTACTGG
Seq- 5 GTTCCTGCAGACAAGAGATGTGCTCGACCGCATTACTTTGATAGTAGTGGTCTTGACTACTGG
(B)
22 23 24 25 26 76 107 135 156 157 164 175 189 194 195 208 209 210 213 259 265 299
Junction
317
327
337
347
357
367
Germline TGGGCTGCGAGATCAAAGCGGC
Seq- 1 --GGGTACGAATTTGCGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGGACGTCTGG
Seq- 2 --GGGTACGAATCTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGGACGTCTGG
Seq- 3 ---GGTACGAATTTGAGACCCCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGGACGTCTGG
Seq- 4 TGGGGTACGAATTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGGACGTCTGG
Seq- 5 ----GTAAGAATTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGGACGTCTGG
Seq- 6 --GGGCACGAGTTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGAACGTCTGG
Seq- 7 TGGGGTACGAGTTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGAACGTCTGG
Seq- 8 --GGGTACTGGTTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAACTACAGTGACTACGGTATGAACGTCTGG
Seq- 9 --GGGTACGAGTTTGAGACGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAAC---------TACGGTATGGACGTCTGG
Seq- 10 --GGGTACGAGTTTGAGATGGCTGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAAC---------TACGGTATGGACGTCTGG
Seq- 11 -GGGGTACGAGTTTGAGACGGATGTGCGAGAGACCCCCACGGTGCTAACTCGCTAAAC---------TACGGTATGGACGTCTGG
(C)
57 78 110 111 112 164 187 195 244 250 253 294
Junction
314
324
334
344
Germline GTATACAAAAAC
Seq- 1 ACTCGAGGTGGCTGTGCGAGAGCCCAGAACTACTACTACGGTATGGACGTCTGG
Seq- 2 ATTCGAGGTGGCTGTGCGAGAGCCCAGAACTACTACTACGGTATGGACGTCTGG
Seq- 3 ATTCGAGGTGGTTGTGCGAGAGCCCAGAACTACTACTACGGTATGGACGTCTGG
Seq- 4 ATTCGAGGTGGCTGTGCGAGAGCCCAGAACTACT---ACGGTATGGACGTCTGG

Figure S4: Multi-sequence alignment of the clones presented in the main text (Fig. 12). Nucleotides are colored to highlight shared and unique mutations in the V segment and junction region. Top row indicates the germline sequence. While the entire junction is shown, only nucleotides with mutations are shown for the V segment.

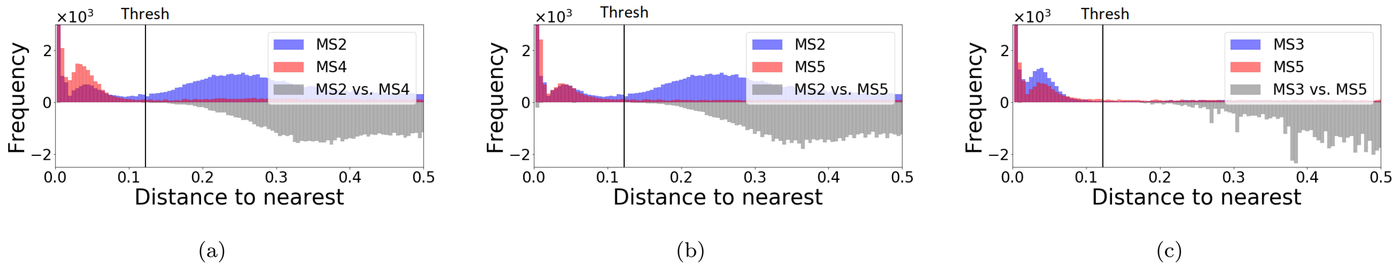
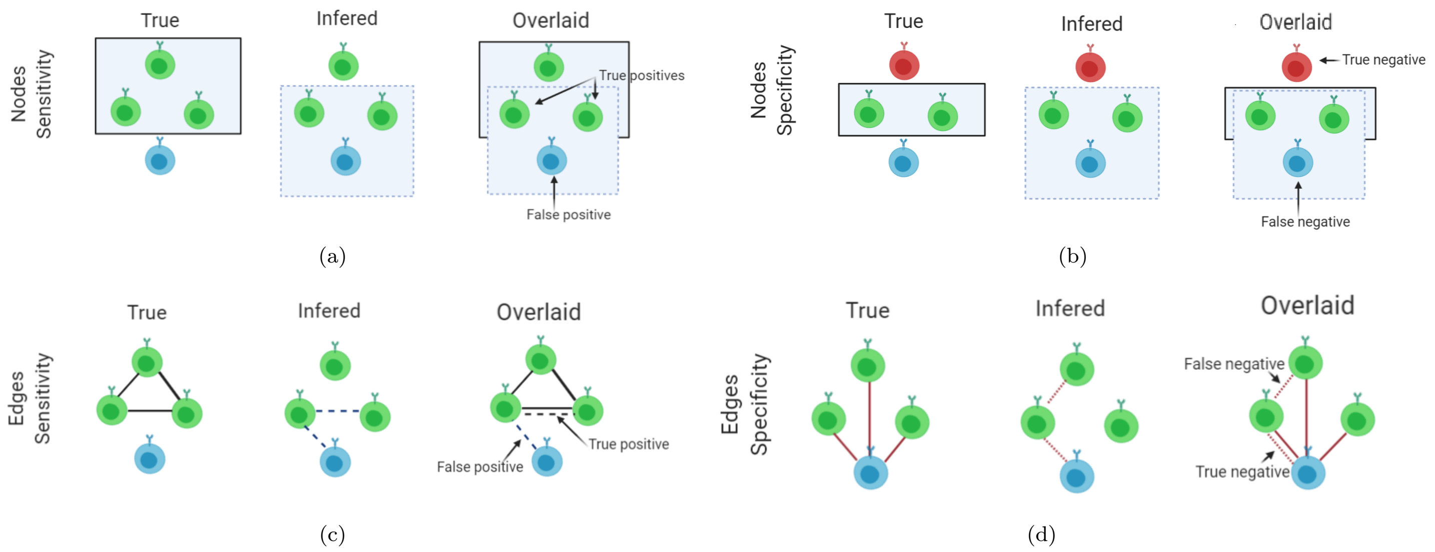


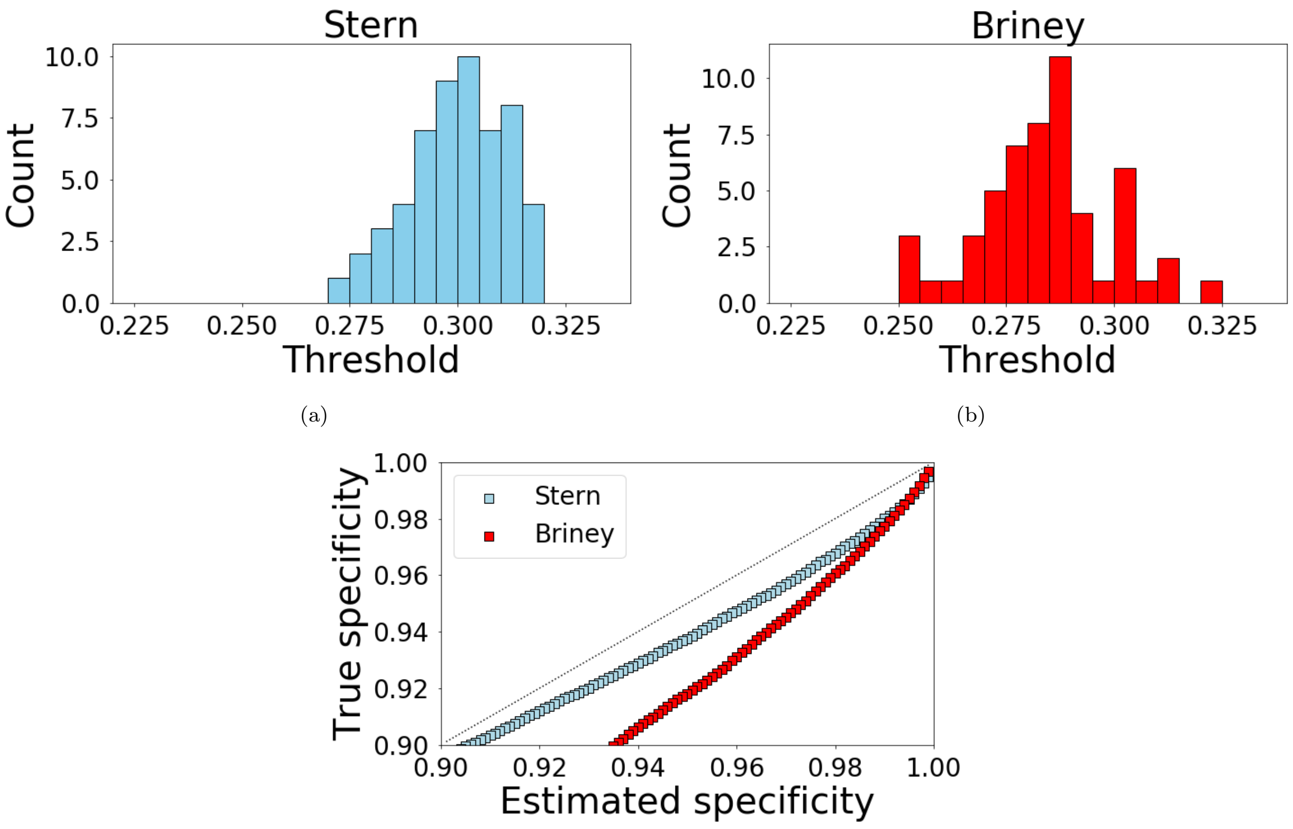
Figure S5: Distribution of distance-to-nearest in the *tf-idf* space. Distances were computed between sequences within one subject (above the x-axis) and between pairs of subjects (below the x-axis), for 3 pairs of subjects in the MS dataset.



(a) (b)

(c) (d)

Figure S6: Illustration of *node* and *edge* sensitivity and specificity. Each sequence is represented by a node, green cells belong to expanded multi-sequence clones, while blue cells represent singletons. Top row: from left to right, true clone membership indicated using a solid box, inferred clone membership indicated using a dashed box and the overlap between true and inferred node memberships is indicated by the intersection between both boxes. (a) The *node* sensitivity is defined as the ratio between nodes in the intersection box and all cells which belong to the expanded multi-sequence clone (green). *Node* PPV is the number of true node assignments (nodes in intersection box) divided by the number nodes assigned as members of expanded multi-sequence clones (nodes in dashed box). (b) The *node* specificity is defined as the ratio between cells outside the union of both boxes and the total number of singletons (blue). (c) True edge clonal relationships are indicated using a solid black lines (edges), inferred clonal relationships are indicated using dashed black lines. The *edge* sensitivity is defined by the ratio between true positive edges (solid black dashed double lines) and all positive edges (solid lines). *Edge* PPV is the number of true positive edges (solid dashed double lines) divided by the number of inferred edges (all dashed lines). (d) True non-edge clonal (negative) relationships are indicated using a solid red lines (edges), inferred non-edge clonal relationships are indicated using dashed red lines. The *edge* specificity is defined by the ratio between true negative edges (solid red dashed double lines) and all negative edges (solid lines).



(a) (b)

(c)

Figure S7: (a) Distribution of the 54 estimated thresholds using data from lymph node repertoires [Stern et al., 2014] (Stern) as negation sequences. Each threshold corresponds to one artificially generated repertoire used as a foreground distribution. (b) Distribution of the 54 estimated thresholds using data from blood repertoires [Briney et al., 2016] (Briney) as negation sequences. Here, the negation sequences are defined by sub sampling 30, 000 sequences from repertoires 316188, 326650 and 326651. (c) Comparing the estimated specificity, tuned based on the negation repertoire (MS or Healthy) and the true *node* specificity computed based on the average specificity over the 54 simulated repertoires. This result suggests that the negation method provides a slightly biased estimate of the true specificity of the alignment-free method.

References

[Briney et al., 2016] Briney, B., Le, K., Zhu, J., and Burton, D. R. (2016). Clonify: unseeded antibody lineage assignment from next-generation sequencing data. *Scientific Reports*, 6(1):1–10.

[Stern et al., 2014] Stern, J. N., Yaari, G., Vander Heiden, J. A., Church, G., Donahue, W. F., Hintzen, R. Q., Huttner, A. J., Laman, J. D., Nagra, R. M., Nylander, A., et al. (2014). B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine*, 6(248):248ra107–248ra107.

[Yujian and Bo, 2007] Yujian, L. and Bo, L. (2007). A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.