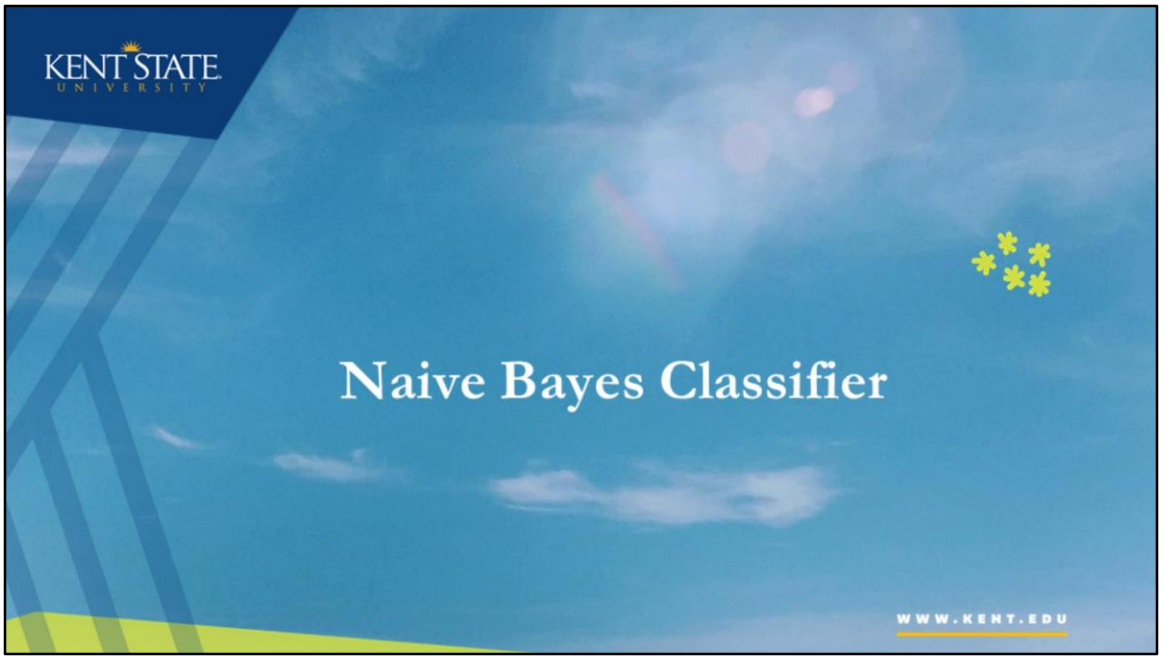
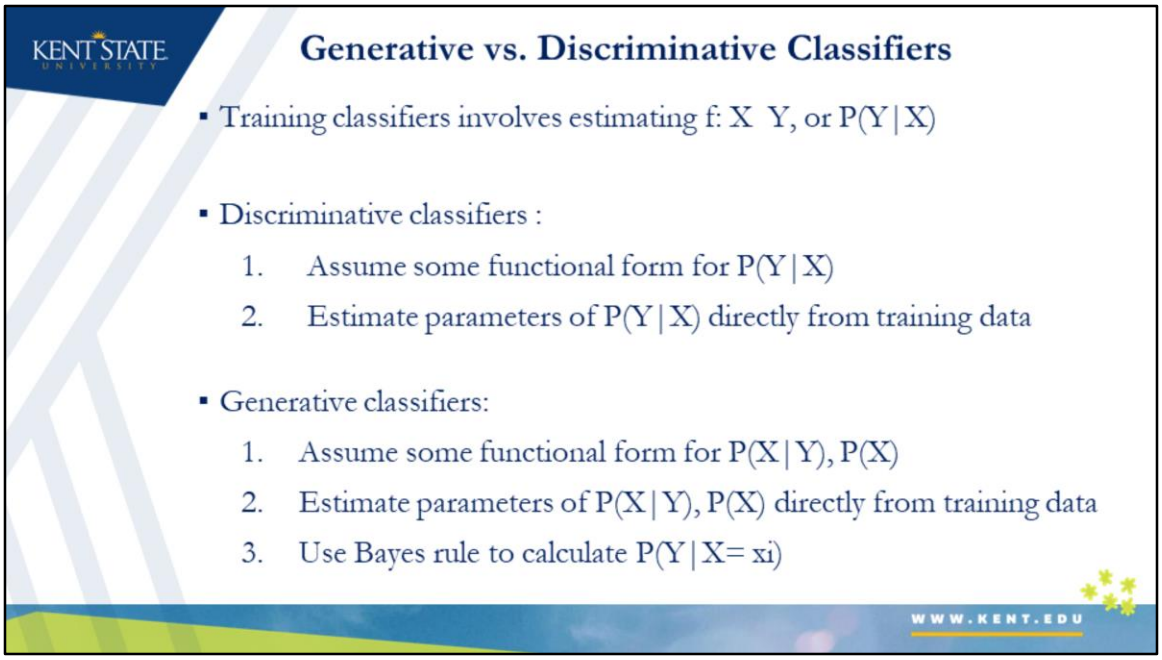
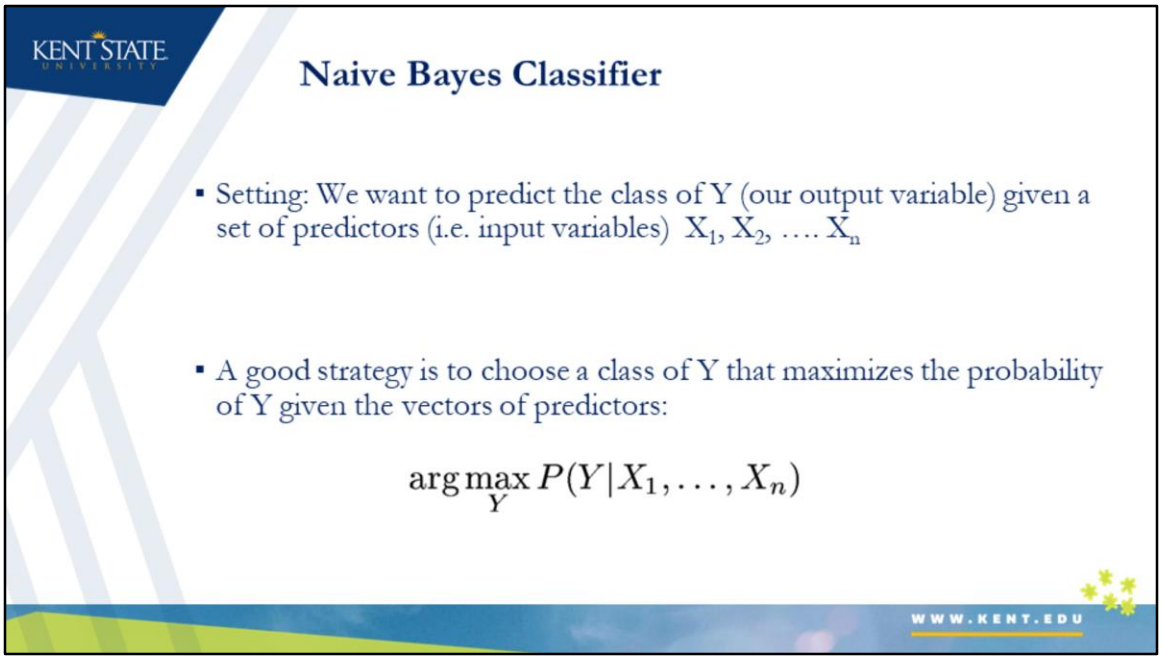


In this module, we discuss the Naive Bayes Classifier. The naive Bayes method, and a branch of statistics called Bayesian Statistics, is named after the Reverend Thomas Bayes (1702–1761).



Consider an example of classifying an observation, say whether a disease is malignant or benign, based on factors like gender, age, etc. Here, Y, the label, is either benign or malignant, and x, is the set of predictor variables, age, gender, etc. The difference between generative and discriminative classifiers is then as follows:

Generative classifiers learn a model of the joint probability, p(x,y), of the inputs x and the label y, and then make their predictions using the Bayes rule to estimate the probability of y (malignant or benign) given the values of the predictor variables, p (y | x) in notation, and then picking the most likely label. Discriminative classifiers, on the other hand, directly estimate p(y | x). An example of the latter method is logistic regression. An example of the former method is the Naive Bayes approach, which we study in this module.



(no audio)
The idea behind NB classifier is simple. We are interested in predicting the class of Y (the label) based on values of a set of predictors. One strategy is to assign the record to the class Y that maximizes the probability of seeing Y given the values seen for the set X in that observation. This approach can be easily implemented as follows for a given observation to be predicted:

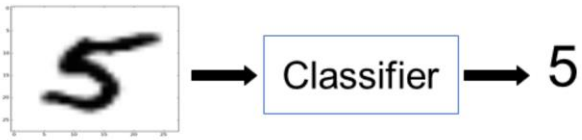
1. Find all the other records with the same predictor profile (i.e., where the predictor values are the same).
2. Determine the probability that those records belong to the class of interest.

Let us see a simple example to illustrate this.

KENT STATE UNIVERSITY

Example

- Let assume the input variables are the intensity (or brightness) of each of the pixels of a image. The task is to classify a handwritten digit as 5 or 6.

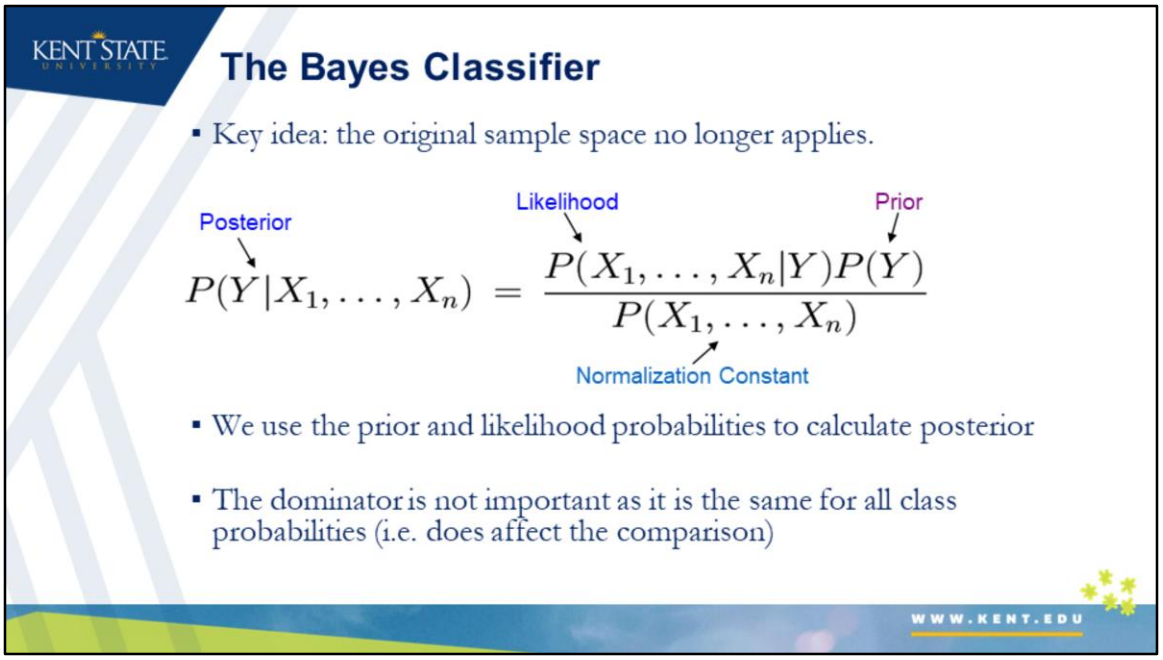


- We can calculate the probability of the output being 5 given the input, and also calculate to calculate probability of the output being 6 given the input.
- We then compare the probabilities and decide whether the input was a 5 or 6 based on those probabilities.

WWW.KENT.EDU

The objective of the classifier is to determine the probability, or propensity, for an observation to belong to a certain class. In this example, we can calculate the probability that the image is a 5 or a 6 given the specific values of the image intensity. Once we know the probabilities, or the relative values of the probabilities, we can assign the observation to the label with the highest probability.

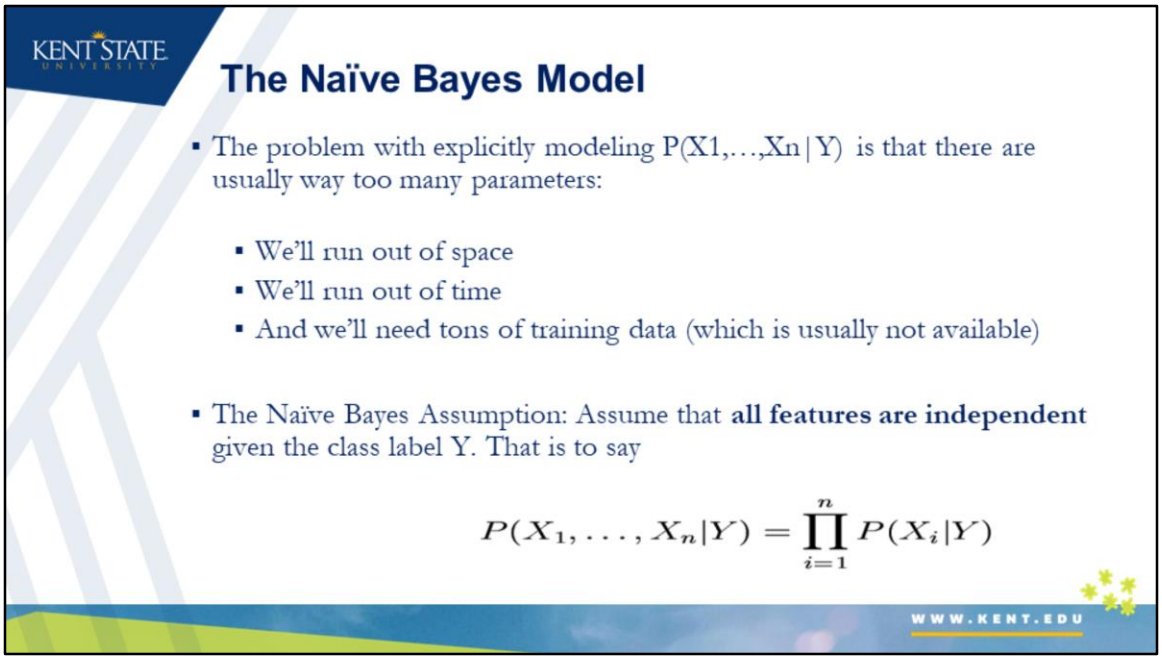
The probability calculations rely on the concept of conditional probabilities, which we study next.



The NB classifier uses the concept of conditional probability, or the probability of event A given that event B has occurred [denoted P(A|B)]. In this case, we will be looking at the probability of the record belonging to class Yi given that its predictor values are x1, x2, . . . , xn. In general, for a response with m classes Y1, Y2,...,Ym, and the predictor values x1,x2,...,xn, we want to compute P (Yi|x1, . . . , xn). This is called the Posterior Probability, and is shown in the equation above.

To classify a record, we compute its probability of belonging to each of the classes in this way, then classify the record to the class that has the highest probability or use the cutoff probability to decide whether it should be assigned to the class of interest.

From this definition, we see that the Bayesian classifier works only with categorical predictors. If we use a set of numerical predictors, then it is highly unlikely that multiple records will have identical values on these numerical predictors. Therefore, numerical predictors must be converted to categorical predictors.



The approach outlined above amounts to finding all the records in the sample that are exactly like the new record to be classified in the sense that all the predictor values are all identical. This may work well for samples with few predictors, but is impractical when there are multiple predictors. The NB model modifies the above approach by making a key assumption, that of conditional independence.

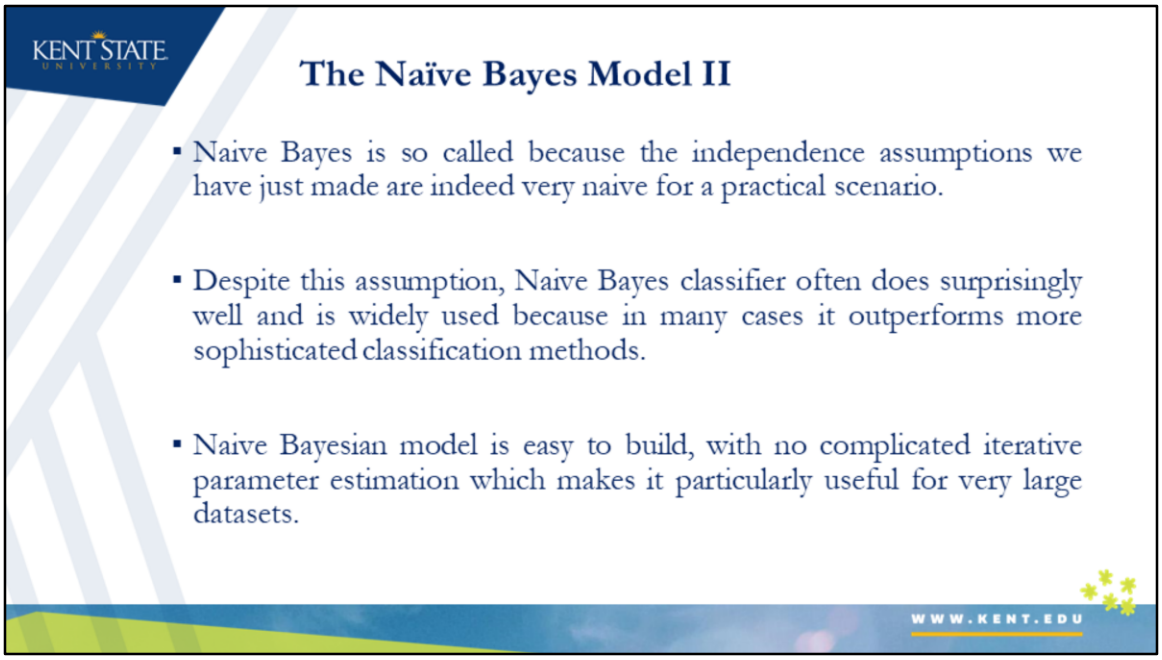
As such, In the naive Bayes solution, we no longer restrict the probability calculation to those records that match the record to be classified. Instead we use the entire dataset.

Returning to our original basic classification procedure outlined before, recall that the procedure for classifying a new record was:
1. Find all the other records with the same predictor profile (i.e., where the predictor values are the same).
2. Determine the probability that those records belong to the class of interest.

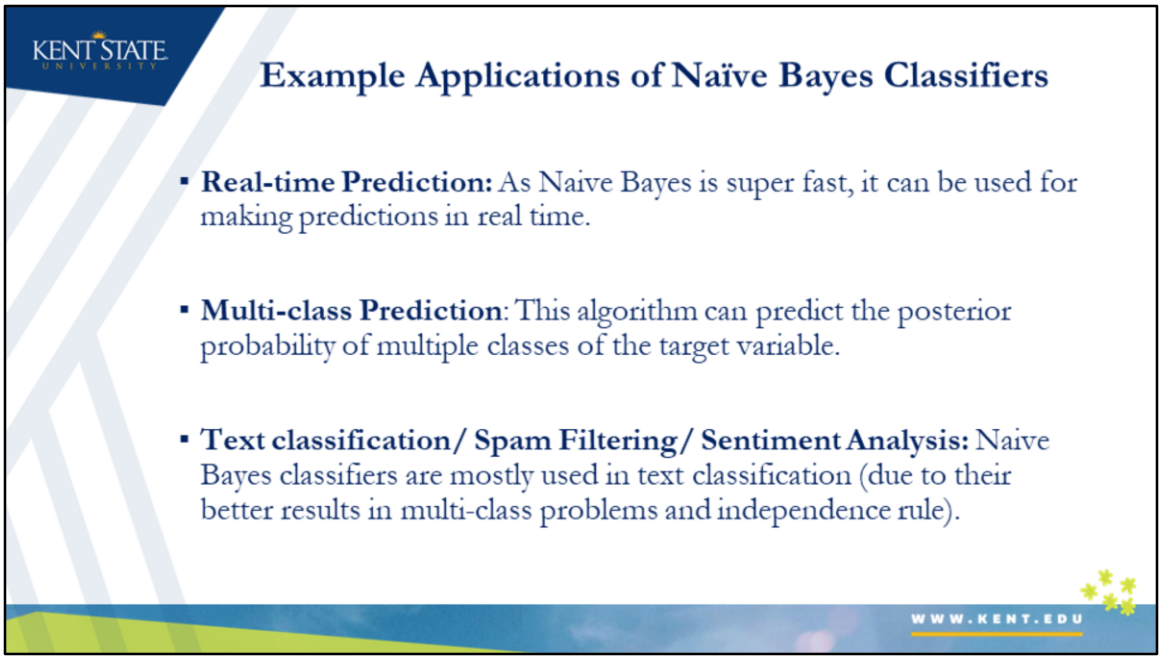
The naive Bayes modification (for the basic classification procedure) is as follows:

1. For class Y1, estimate the individual conditional probabilities for each predictor

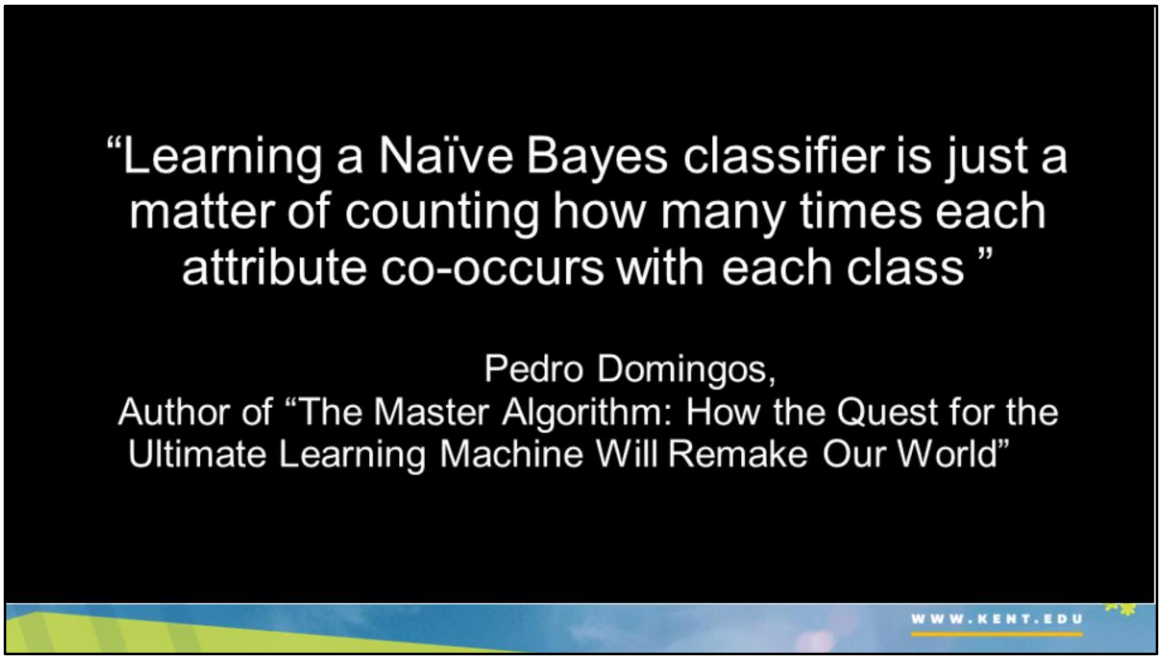
P(xj|Y1)—these are the probabilities that the predictor value in the record to be classified occurs in class Y1. For example, for X1 this probability is estimated by the proportion of x1 values among the Y1 records in the training set.
2. Multiply these probabilities by each other, then by the proportion of records belonging to class Y1. This gives the probability in the numerator for the equation on the previous page.
3. Repeat Steps 1 and 2 for all the classes. This provides the numerator for all classes Yi in the equation on the previous page. We can further calculate the actual probability by doing the following:
4. Estimate a probability for class Yi by taking the value calculated in Step 2 for class Yi and dividing it by the sum of such values for all classes.



A key assumption in the NB model is that of conditional independence. This is unlikely to be true in real-world settings as predictors are often correlated, but surprisingly, the NB model does well from a practical standpoint. While the posterior probabilities calculated using this approach do not usually match the exact probability, the values calculated are nevertheless ranked similarly to the real probability values. Thus the classification rule to assign a record to the highest probability still makes the correct classification, even if the calculated probability is different from the true probability. In other words, the rank order of a record to a class is correct, even if the exact values used are not.



The NB algorithm is fast, and does well. But, as mentioned earlier, it is typically used only when all predictor variables are categorical. Further, since the probabilities are not the exact probability, the NB is not used for example in credit scoring, where we require exact values, rather than rank order. Nevertheless, NB is an efficient algorithm to use, and in the next module, we will see how this model can be implemented, and then adapted to handle some of the above limitations.



This concludes our module on NB.