BIKE RIDES AND DATA MINING

"Big data" as it relates to cycling activities; and the value of statistical analysis

Osman Isvan; April, 2020

Over the years I have uploaded 1,348 bike rides to Strava. Each record contains my cycling data at 1-second resolution, chronicling my location, power, heart rate, elevation, speed, cadence and more. The Veloviewer.com web site allows me to fetch my data from Strava and look at it from different angles. Its online analysis and visualization tools are great, but the feature I appreciate the most is its ability to export aggregated ride data as a text file so that I can perform my own analyses off line.

For example, let's say I wanted to know how, on average, my power output affects my speed. Of course, there are trusted, experimentally validated theoretical models that would give me the answer, provided that all external conditions (wind speed, wind direction, hill slope, road surface etc.) are known, and user-specific and equipment related factors (body weight, bike weight, drag coefficients, etc.) are accurately estimated or specified. But if I want to know instead, my *actual* power and my *actual* speed, averaged over a large data set representing *my* bike rides, in places where *I* ride most often, over a long time in prevailing circumstances, rather than hypothetical scenarios under predetermined test conditions, I can turn to statistics where troves of information are embedded in the data. This article is about tools for uncovering this information by statistically analyzing the available ride data.

Figure 1 shows two screenshots from the Veloviewer analysis tool. In each frame, every dot is a ride.

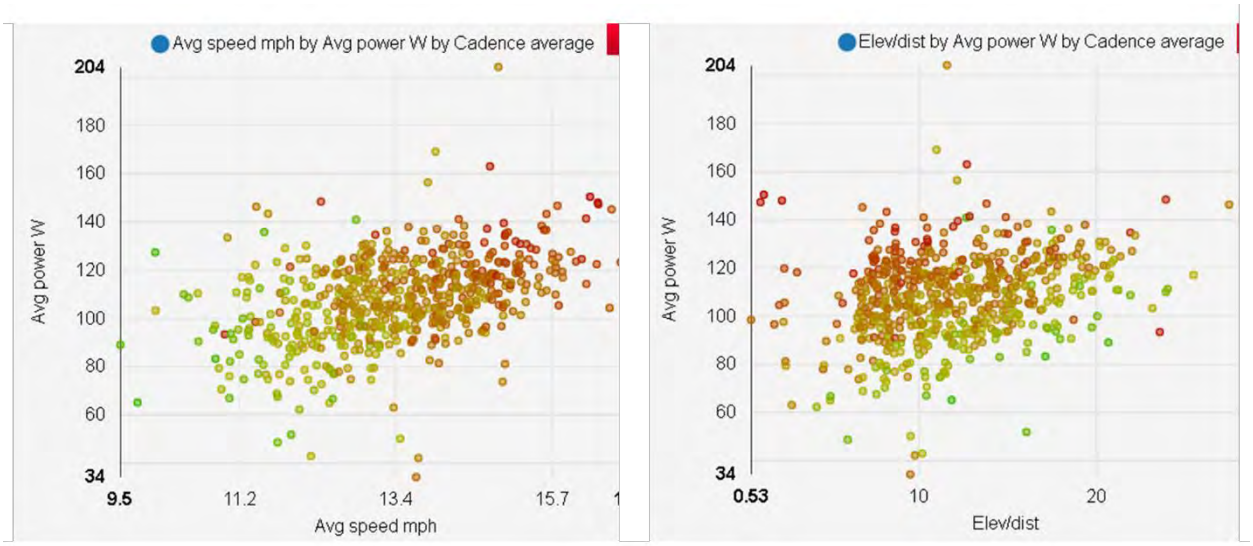

Figure 1 –Average power vs. average speed (left) and vs. elevation gain per distance (in meters/km –right).

Average cadence is color-coded (light green ≈60rpm; dark brown ≈80rpm). Considered are 600 rides longer than 15 miles with the same bike with a power meter. Scatter plots were generated using the Veloviewer visualization tool (elevation unit and conversion factor: 10m/km = 52.8 ft/mile).

TEST RIDES and FIELD DATA: The scatter graphs of Figure 1 reveal that when the hilliness of the route is disregarded, power and cadence both increase with speed, and when speed is disregarded, power and cadence both increase with hilliness. Unlike test rides under controlled conditions, in this analysis

disregarding a variable does not mean preventing it from changing. To the contrary, it means allowing the variable to change naturally, which has some advantages over conventional test protocols. Interestingly, my average power on a route depends on the hilliness of the route, and I decide on how much power to apply according to some criteria yet to be determined.

PATTERN RECOGNITION: Figure 1 is a good illustration of the value of visual pattern recognition in data analysis: with regard to cadence, I notice that the color fades from left to right in the speed chart, and somewhat from right to left but more significantly from bottom to top in the hill chart. This means that in practice, cadence is (for me anyway) a strong function of *vertical speed*. The flatter the route is, the higher is my average cadence.

To draw conclusions from this kind of data I rely on the "rule of averages", so in general I want there to be as many data points as possible, to make the conclusion accurate. But to isolate specific conditions of interest I apply filters, which reduce the quantity but increase the quality of the data regarding the condition in question. For example, to create Figure 1, in order to minimize the effect of measurement errors, I asked the software to exclude the rides that I have done without a power meter. Their inclusion would have contaminated the higher-quality data acquired with a power meter. Over the years, I have used 4 different power meters on this particular bike.

Mathematical analysis can yield precise numbers when needed. But even visual inspection of the left frame of the graph (Fig. 1a) is sufficient to conclude that the more power I apply, the faster I go. This, of course, was a foregone conclusion. But the curious thing is that my average speed and average power appear to remain more-or-less in direct proportion. This is an unexpected finding because under steady state conditions, the relationship between power and speed is known to be highly nonlinear, at least on flat terrain without wind. One plausible explanation is that on hillier routes I apply more power than I do on flatter routes, as evidenced by the right frame (Fig. 1b); and as a result, the terrain-induced increase in power and the corresponding decrease in speed would tend to flatten the growth function of the power/speed ratio. Still, it is far from obvious that the end result should be a straight-line (linear) relationship between average power and average speed on varied terrain, as Fig. 1a appears to indicate. This finding is also depicted in Figure 2.

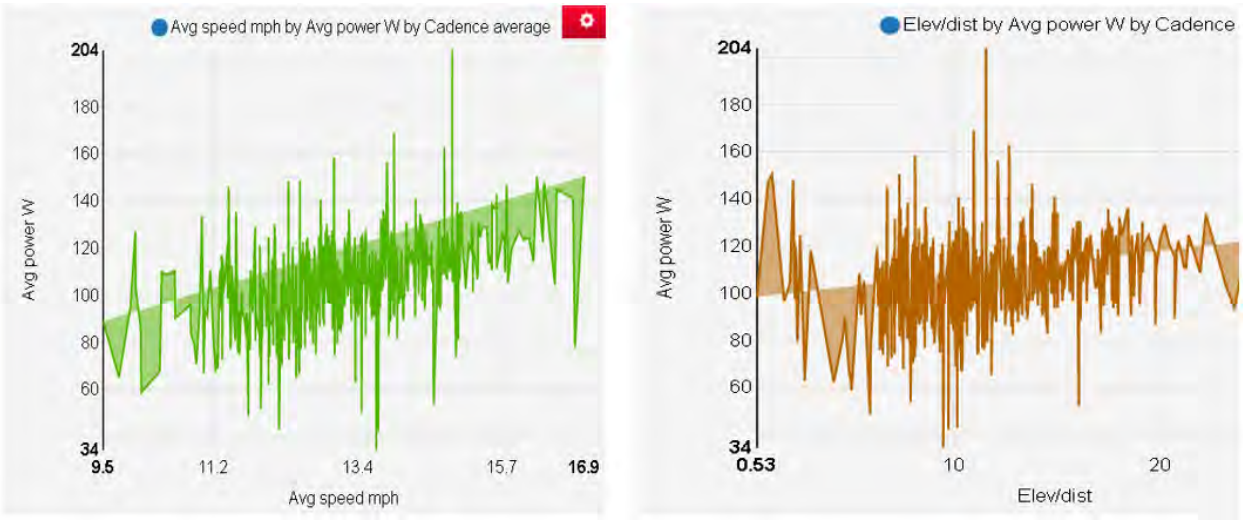

Figure 2 –A line graph presentation of the data in Figure 1

Filtering the data by elevation, duration, energy and other variables reveals the same kind of relationship (i.e., the relationship remains linear). On hilly routes, the relative share of aerodynamic power is reduced, explaining why the average power is not increasing with the *cube* of the average speed. After all, my average speeds are fairly low to begin with, and so, on average, I am not particularly burdened by air drag. But when we look at all of my routes including flat ones, with average speeds of up to 17 miles per hour, why do we not see even a hint of increase in the power / speed ratio? Even when leave out the hilly routes (not shown here), I still don't see any evidence of that.

To investigate questions of this kind, I find the analysis tools provided by Veloviewer.com to be very versatile and useful; but for more specific and detailed investigations, I export the data and analyze it off line. Figure 3 depicts two representations of the same data, plotted using a spreadsheet. The scatter graph in the left frame (Fig. 3a) shows the speed and power of every ride individually (same format as in Figure 1). In the line graph on the right frame (Fig. 3b), individual rides are hidden, and in return, power and speed data are processed as follows: the horizontal axis is divided into speed bins that are (in this case) 1 mph wide. For example, the bin centered at 15mph (blue slice in Fig. 3a) is comprised of only the rides where my average speed was between 14.5mph and 15.5mph. The average (arithmetic mean) of the average powers in each speed bin is then calculated as the power value corresponding to the bin center, and the line graph (Fig. 3b) is made from these values. To convey additional information, the number of rides contained in each speed bin is plotted as a dashed line and projected onto the secondary vertical axis to the right of the plot area. For example, the graph shows that with this particular bike, I have uploaded 110 rides with a power meter where my average speed falls within the 14.5 to 15.5 mph tolerance window, and on this subset of rides my average power was 118 watts. The scatter and line formats shown in Figure 3a and 3b respectively are two visualizations of the same underlying data.

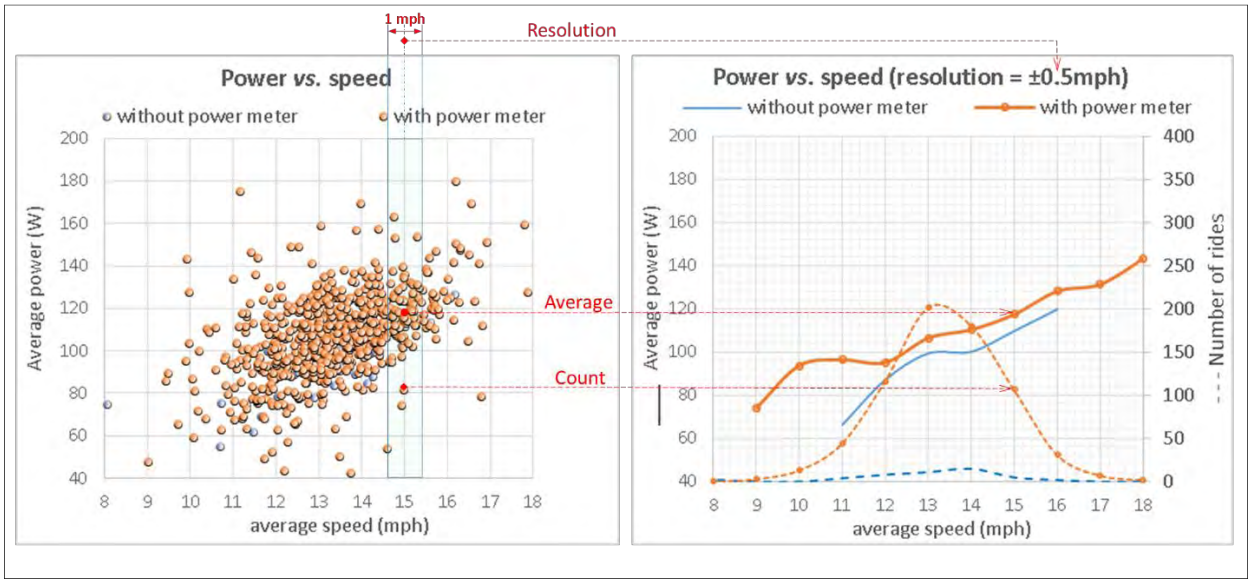

Figure 3 –Scatter and line graph presentations of the same data (751 rides; 704 with, 47 without a power meter)

It is not feasible to count the dots and visually estimate the average power at each speed by looking at the scatter graph; the line graph is easier to read. On the other hand, the scatter graph makes it easier to visualize the *distribution* of the power and speed data across the speed and power axes, respectively.

For the rest of this article I shall use scatter and line graphs depending on which format is more convenient or vivid to illustrate the underlying information.

I should point out that "average speed" is not as simple a concept as it may appear at first. For example, if the first half of a route is covered at 10mph and the second half at 20mph, the average speed for the whole route will not be 15mph. It will be 13.33mph. This means that, for example, if I had uploaded my morning commutes and evening commutes as separate activities, this particular analysis would yield a lower average speed than if I had uploaded each round trip commute as a single activity. The case for average power is even more complicated, and in the context of human power, it may be more revealing to pay attention to a different quantity called "weighted average power" (more on that later) instead of average power.

Even with these caveats, I think that much useful information can be inferred from this kind of analysis. As we can see, Strava's virtual power meter (blue solid line) is consistently estimating less power than my power meters do, but where there is enough data, the general trend is similar. Perhaps I could use this information to "calibrate" Strava's virtual power meter against my physical power meters by editing some of the items in my Strava user profile, for example body weight and bike weight. When a physical power meter is present, Strava records its output *in lieu of* estimating my power; so, the physical and virtual power meter readings shown in these graphs (brown and blue dots and curves) come from entirely separate sets of rides. They are not related other than having the same average speed.

I am relying on "the rule of averages" to reduce the influence of random factors, but it looks like I just haven't uploaded enough rides on this bicycle without a power meter to extract useful information by averaging. Even with a power meter, only between 12 and 15mph did I upload more than 100 rides with ±0.5mph tolerance for each average speed.

GRAPHICS RESOLUTION: The width (i.e., ± tolerance) of the averaging window determines the graph resolution. A wide window results in a smooth curve as shown in Figure 3 above. Figure 4 below demonstrates the tradeoff between choosing moderately wide ±0.25mph and very narrow ±0.05mph tolerance windows.

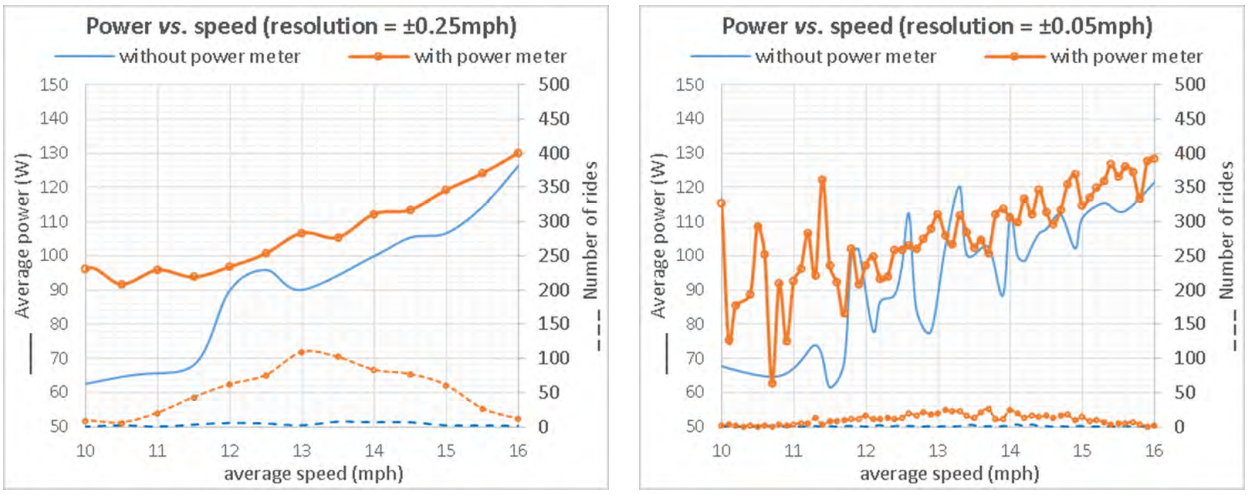

Figure 4 –The consequences of changing the number of speed bins (graph resolution)

When the tolerance windows (i.e., speed bins) are wider, there are fewer points on the graph, but more rides are averaged at each point. On the other hand, using narrower bins reveals more detail, at the expense of locally greater uncertainty due to having fewer items to average. As seen in Fig. 4b with ±0.05mph resolution, even at the most populous bins (13 – 14 mph) the average power is computed from fewer than 30 rides, as opposed to ≈100 rides with ±0.25mph resolution as seen in Figure 4a, and from ≈200 rides with ±0.5mph resolution as seen in Figure 3. When converting a scatter graph into a line graph, making the averaging bins wide in order to produce a smooth curve is not the same thing as producing a rough curve first and then smoothing it with a moving average low-pass filter, because with the former process each scatter value is averaged only once.

In the line graphs, the curves recorded with a power meter (brown) are smoother than the curves recorded without a power meter (blue) for two reasons. 1 –They are averaged from a greater number of rides; and 2 –Physical power meters are inherently more reliable than virtual power meters.

The ±0.25mph tolerance appears to yield a near-optimum graph resolution with the available data, as Fig. 4a reveals a hint of the expected nonlinearity without the curves' looking to be possibly "overdamped" as in Fig. 3b or "too noisy" as in Fig. 4b. This optimization is highly subjective. Nonlinearity refers to the slope of the curve leveling off at low speeds and steepening towards higher speeds. "Noise" refers to the random peaks and dips in Fig. 4b that can obscure the underlying trend.

POWER, SPEED and HILLINESS: So far, we have seen that my power output increases with speed. But at the same time, my speed must be decreasing with hill slope, and each ride has an elevation profile with a unique distribution of hill slopes. So it would be interesting to look at the mutual interactions among 3 variables:

- Hilliness (measured in units of elevation gain per distance)
- Average power
- Average speed

Clearly, hilliness (i.e., average climb rate) is the only independent variable here. Average power is a variable of my own choosing, but I expect my choice to be influenced by the hilliness of the route. Average speed is the most dependent of the three variables, since it is the *outcome* of the application of a *chosen* power to a route with a *given* hilliness.

According to the available data, as the hilliness of the route increases, my average power increases and my average speed decreases (Figure 5).

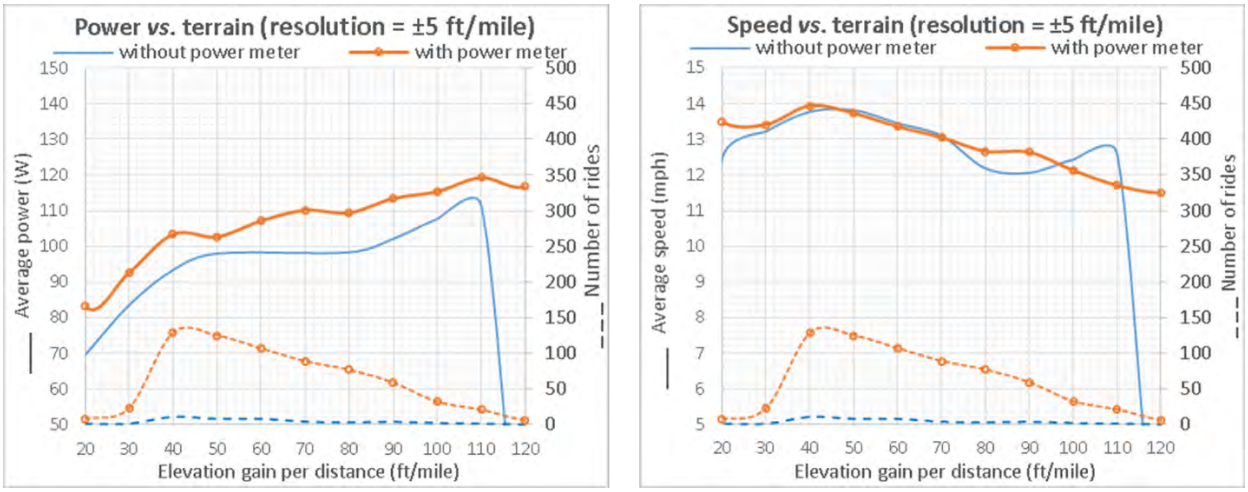

Figure 5 –Effect of the hilliness of the route on average power and average speed. Increased hilliness (i.e., average climbing grade) results in a steady increase in average power and a steady decrease in average speed.

Note that for climb rates greater than 40 ft/mi where most of the data are found, the percentage rate of the increase in power is approximately equal to the percentage rate of the *decrease* in speed (i.e., my average *pace* remains proportional to my average power). This finding is consistent with a supposition that on hilly routes most of the work is done against gravity. The frequency distribution of the number of rides across the hilliness axis is peculiar, with a singularity around 40 ft/mi followed by an almost linear decline up to 120 ft/mi. I was expecting to see a normal (i.e., approximately bell-shaped) distribution similar to those for average speed (Figures 3 and 4) and average power (Figures 7 and 8).

Data analysis alone cannot reveal the criteria by which the average power on a cycling route is related to the hilliness of the route, but seeing *how* the two variables are related (at least on my rides) may reveal some information. Of course I slow down for hills, and in the process I may increase my power output, perhaps in an attempt to partially compensate for lost time. But I have a choice in the matter, and that choice alone defines the shapes of the curves in Figure 5. The gradient of a hill constrains my speed, but it does not constrain my power output. When the road turns from flat to uphill, I have no obligation to work harder I could have chosen to slow down just enough to maintain the same power regardless of the hill. In that case, the curve of Figure 5a would have been flat, and the negative slope of the curve of Figure 5b would have been steeper. Typically, I *do* work harder on hills (as we shall discuss later), but, when I encounter a hill, how do I decide *how much* harder I should work? Looking for an answer to this question is a big part of my incentive to conduct this analysis.

Even if the analysis doesn't answer *why* or *how* I adjust my power according to the terrain, the information revealed here does help me predict how fast I am likely to ride on a route with a given elevation profile. If nothing else, this information is useful for route planning. I can imagine routing software that will update my estimated time anywhere on the route as I am drawing a tentative route on an online map. The information for this customized time-prediction feature would be based on the statistics of my personal ride data. Also, the "virtual partner" features found on GPS-enabled cycling computers would benefit from this custom information.

THE COST FUNCTION TO BE MINIMIZED: When I ride my bike, I must be adjusting my power output to maintain some quantity at a constant level, or minimize or maximize it. This "cost function" may be a

physical quantity or a perceived quantity. It doesn't have to be the same for all cyclists, but a quantity can be parametrically defined such that the parameters can be determined for each individual. In another article, I have shown theoretical evidence that if a cyclist's goal were to complete a given route at a given average speed with the least amount of energy, then power output should be increased on hills. But that's a big "if". First of all, it is not clear that saving energy is a common pursuit among cyclists; it certainly isn't for me (mainstream use of e-bikes for sporting purposes may change this in the future). Secondly, that strategy would have me striving to ride at *constant speed regardless of hills*, except when I reach my maximum power limit, but in Figure 5 I don't see any evidence of this. Also, if the minimum-energy hypothesis is based on the idea of using energy judiciously for fear of getting tired towards the end, I would expect my shorter rides to have higher average speeds than my longer rides, but the evidence of average speed vs. distance (not shown here), does not support that theory.

Since my power goes up and my speed goes down with increasing hilliness of my routes, it is natural to ask whether my data might be consistent with my trying to make my power / speed ratio remain constant on hills. The relationship between this ratio and the hilliness variable is shown in Figure 6.

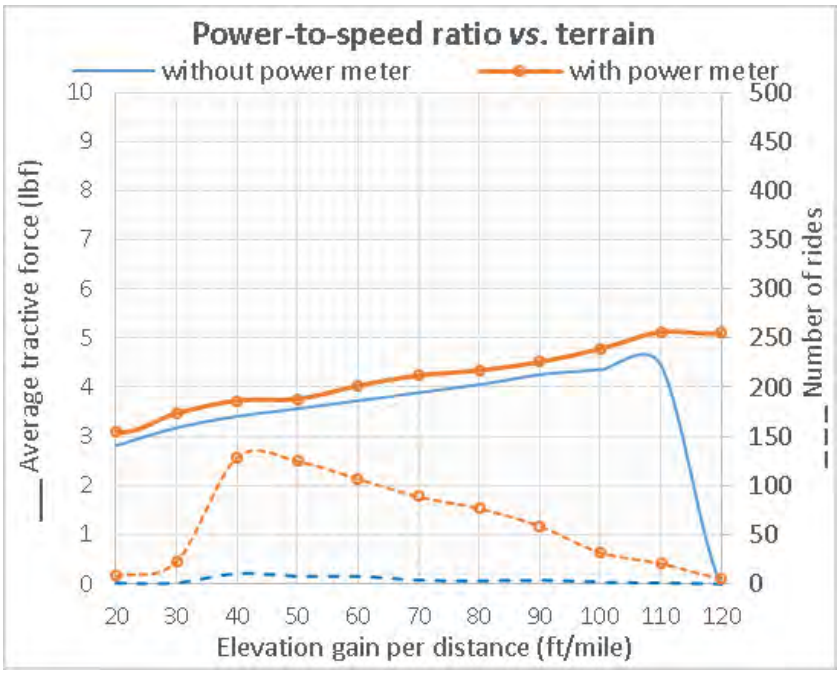

Figure 6 –The ratio between average power and average speed (tractive force) steadily increases with hilliness.

Incidentally, as the axis label of Figure 6 indicates, the ratio between power and speed is *tractive force*. For clarity, I make the distinction between *propulsive* force and *tractive* force as follows: tractive force is the forward component of the force at the rear tire's contact patch with the ground, whereas propulsive force is the resultant sum of all forward-facing forces acting on the bicycle, including the tractive force. For example when pedaling on a descent, the bike is being propelled by two power sources at once: gravity power and human power. Gravity power acts at the combined center of gravity, and human power acts at the force center of the tire-contact patch. In this example, the propulsive force is the *sum* of tractive force and gravity force. It is equal in magnitude (and opposite in direction), to the sum of the opposing forces that include air drag and tire rolling resistance. On the flats, since gravity force is zero, without acceleration tractive force equals propulsive force. On climbs, gravity force becomes negative,

and without acceleration, propulsive force equals the *difference* between tractive force and gravity force. During a bike ride, as the opposing forces change all the time, the cyclist has a choice in adjusting the tractive force, by increasing or decreasing the pedal force and cadence and by shifting gears. Evidently, according to Figure 6, I make these adjustments in such a way that my average tractive force remains approximately proportional to the hilliness of the route, across a wide range of topography from very flat terrain (20ft/mi) to very hilly terrain (120ft/mi).

RULE OF AVERAGES: Of course, hills are not the only variable for cyclists to adjust their tractive force or power output. But when I statistically process a sufficiently large body of data to evaluate my response to hills, I can reasonably assume that all other factors will tend to average out. For example, sometimes I may be taking it easy not because it is flat but because I am riding with slower people. On other occasions, I would be riding with faster people and working extra hard not to get dropped. So, if I compute my average power in a data set containing *enough* rides, the variability of my power due to the dynamics of group rides will tend to disappear. The rule of averages will force my average power to converge to a value that reflects my own typical condition, group or solo rides, strong or calm winds and other contributing factors, and the result will reflect a statistically valid combination of all of the rides in the data set.

THE WIND: Likewise, the wind is another obvious factor, but if I have done enough rides, wind conditions, too, will have affected my statistics in the appropriate ways I would deem to be "typical for me", even if they may be not typical for other cyclists. Unfortunately, the data that I upload to Strava does not include wind measurements. I do record my airspeed with a wind sensor on my bike, in sync with the rest of my cycling data that I upload to Strava, but that record, which is saved for each ride in a separate file, cannot be uploaded to Strava due to proprietary and incompatible file formats, and I don't (yet) have the technology to reconcile the two sets of data for multiple rides. This is unfortunate because having wind data (i.e., my airspeed) in sync with the rest of the data that I am analyzing here would have made a huge difference in the usefulness of the results. It would benefit Garmin, Wahoo Fitness and Strava to partner (or at least cooperate) with companies specializing in airspeed measurements on bicycles, such as the manufacturer of the wind sensor that I use, Velocomp. This instrument includes other sensors besides a wind-pressure sensor, and can be described as an "aero lab on wheels", but is marketed - shortsightedly in my opinion - as a low-budget power meter. Other bicycle-specific wind measuring devices have been developed, and some of them are capable of measuring wind yaw in addition to airspeed.

BIG DATA: This article is more about exploring the use of analytical tools and data-mining opportunities for cyclists than an evaluation of the underlying data that is specific to my rides. I have access only to my own data, and I realize that the rule of averages can only get me so far. But my motivation for analyzing my data is the potential application of the methods to cycling databases available from global aggregators such as Garmin, Strava, RidewithGPS, Apple, Google and others, including platform-independent mobile apps. These crowd-sourced databases are not limited to storing data on any one person. They store data on millions of people. If the methods I describe in this article are useful for uncovering relevant information about my own rides, then certainly they can be used for expanding our collective knowledge and understanding of the sport of cycling in general.

ROLLING HILLS: So far I tried to account for the effects of climbing by analyzing elevation gain, but elevation gain is not a sufficient metric to represent all relevant aspects of the elevation profile of a

cycling route. For example, rolling hills would affect average speed and average power very differently than sustained climbs with the same elevation gain. Also, steep gradients are known to affect cyclists very differently than moderate slopes with the same elevation gain. But firstly, although the analysis I am presenting in this article does not differentiate short hills from long hills, or steep slopes from moderate slopes, the underlying database does contain this information, which needs only to be accessed and analyzed. (This will be the subject of a future article.) Secondly, the value of this approach to analysis is that the result reflects the terrain that I ride often. In other words, my "average" data is, in fact, a weighted average where the appropriate weighting coefficients are automatically applied and statistically validated for me. In that sense, analyzing ride statistics may be more informative than controlled experiments and special test rides designed to isolate and study each contributing factor separately. At least, there is a case for collecting both sets of data and comparing the results.

NON-RECIPROCITY: That my average speed on rides with 70ft/mi climbing is 13mph does not mean that the average climbing on rides where I average 13mph must be 70ft/mi. The same is true for average power. This can be seen, for example, by comparing Figure 7 with Figure 5a. These two graphs are based on the same data; the only difference is that dependent and independent variables are transposed. Clearly, the relationship between these two variables is not reciprocal. The profound difference in the frequency distribution of the number of rides across the horizontal axes is also meaningful.

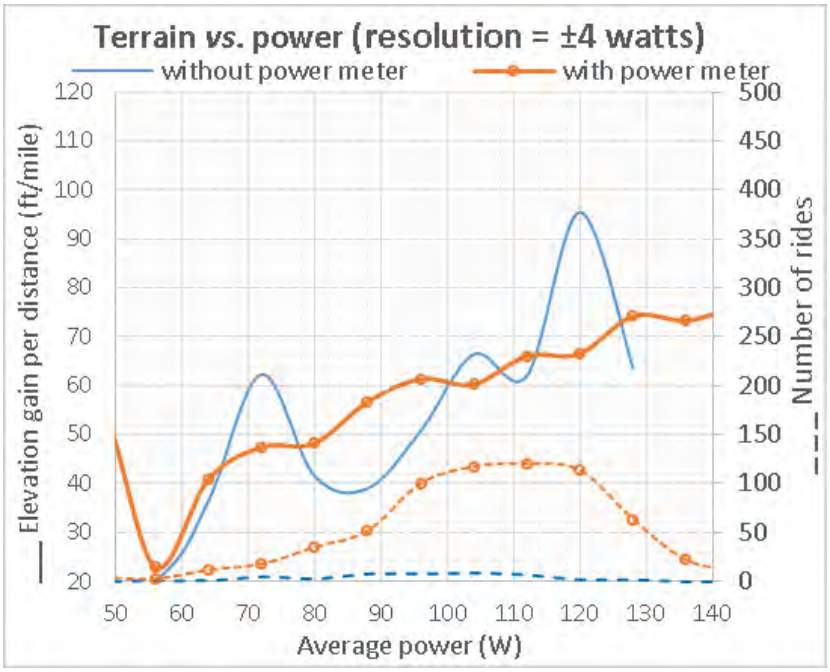

Figure 7 –Hilliness vs. power (compare with Fig. 5a).

In Figure 5 we can see that the range of powers between 90W and 120W corresponds to almost the entire range of hilliness (≈30 - 110 ft/mi), whereas in Figure 7 the same power range corresponds to a much narrower range of hilliness (≈55 - 70 ft/mi). If this seems odd at first, keep in mind that the hilliness of the terrain is a given, but my power output is an entirely voluntary choice. If I had made a conscious decision to ride hilly routes exclusively at high power, but do flat rides at high power and low power, external circumstances would not have prevented me from doing just that, and my ride statistics would reflect that particular choice. Of course I am not following explicit rules that dictate how hard I

must ride under which conditions; but apparently *something* is causing me to ration my power according to a "secret formula". With statistical analysis, I am trying to uncover this formula, and I am hoping to find its relation to known laws of physics and human physiology.

WEIGHTED AVERAGE POWER: Performing at a constant power is not as taxing on the human body as performing at variable power that yields the same average. A training parameter called weighted average power (or normalized power) has therefore been defined, which quantifies a cyclist's power and its variation during a ride. Weighted average power is considered to be a metric that better represents the cardiac load and therefore the training level of a given bike ride, than average power. Figure 8 shows how my simple average power and weighted average power change with terrain.

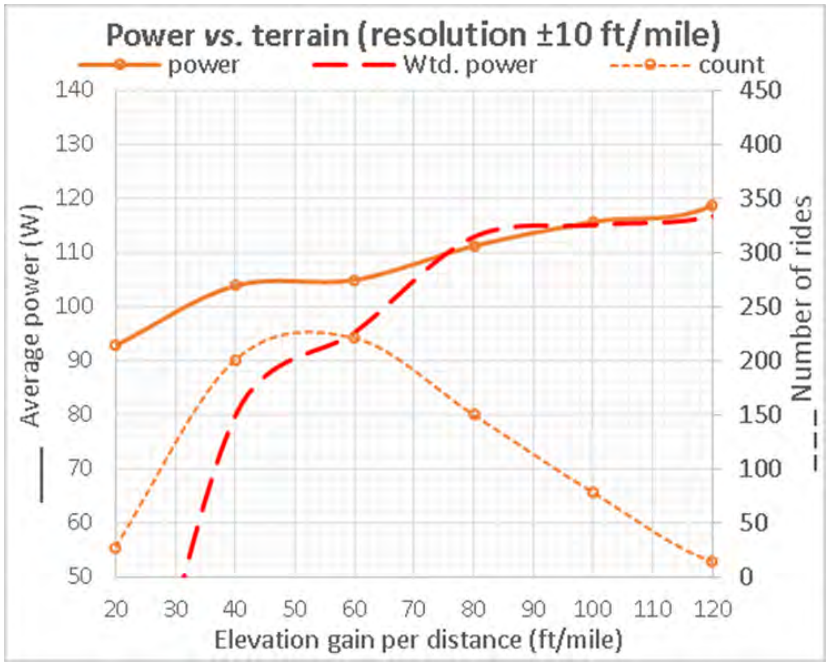

Figure 8 –Average power and weighted average power

On hilly rides, the two curves converge. I was expecting the opposite. The metric is normalized to yield the same value as average power if the whole ride was done at a constant power. Apparently my power output is more nearly constant on routes with more than 80 ft/mile of climbing than on flat routes. A possible explanation might be that on the particular terrain where I do most of my rides, flat routes are full of short hills where I don't have to pace myself, whereas on hilly rides my power is less variable because I operate closer to my functional threshold power (FTP). However, this is not a satisfactory explanation because even on the hilliest rides on this chart, my average power is much less than my FTP.

POWER and HEART RATE: There are plenty of theories, supported by hard data, on how cyclists and other athletes perform on treadmills, ergometers and on the road, under a wide range of imposed conditions and test protocols. The body of knowledge gained from these controlled scientific studies guides medical professionals as well as sporting teams, coaches, trainers and athletes. But in the digitally connected world, large quantities of actual ride data recorded under naturally occurring conditions on recreational bike rides, previously never available to even the most advanced research teams, has suddenly become easily accessible to anyone. In my opinion, the information potential of statistical

analysis of such data is invaluable. Studies will not simply verify or duplicate the information gained from lab tests, but will uncover new information, which will guide the lab scientists to investigate new and different questions in their labs. When we "go for a bike ride," we are not performing at a specified intensity level relative to a given target or heart rate threshold. Well, maybe some of us do, at least some of the time; and when we analyze our ride data we should include training rides, maximal efforts, hill repeats, races and the like. But in the end, I want to know what is statistically relevant under "normal" circumstances, where *normal* is defined as whatever actually happens in the daily routines of the individual (or group of individuals) whose data are being analyzed.

So, here are my data: Figure 9 shows that as I age, my average heart rate and average power are consistently declining year after year. Most interesting is the periodic seasonal variation in my average heart rate. The *timing and period* of these variations are revealing: the vertical major gridlines in Figure 9 mark my birthday, which is in January. Because I live and ride in the Northern Hemisphere, the periodic variations indicate that my average heart rate is generally lower in summer rides than in winter rides.

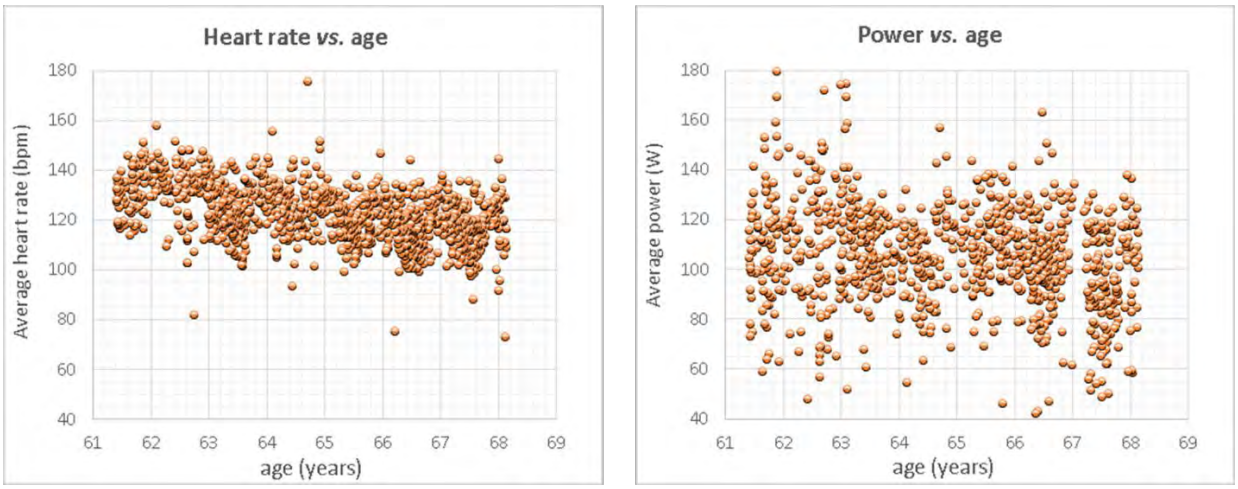

Figure 9 –Heart rate and power data on all rides I uploaded to Strava in the last 7 years

One logical explanation of this finding is the training effect: in the summertime, my increased fitness level allows a lower heart rate. Other possible explanations can be found. For example, one could argue that my winter rides are more intense, because they are shorter in duration. However, data correlation between average heart rate and ride duration suggests the opposite: I find that my cycling heart rate is actually higher on *longer* rides, which are by the way, more common in summer months. Furthermore, the fluctuations with a one-year period, which clearly stand out in the heart rate graph, are barely noticeable in the power graph. So, it looks like during the summer months I get more work done with fewer heartbeats. And as I get older I do nearly the same amount of work per heartbeat. Can this all be true? I can't think of a more rational interpretation of the data, but to be more confident, I decided to find out how my cardiac efficiency (the ratio between power and heart rate) has changed over time. This is shown in Figure 10 along with my average speed. Because speed and efficiency values are an order of magnitude apart, I am plotting them on a logarithmic scale to make the time rate of change easier to evaluate by visual pattern recognition (with a logarithmic vertical axis, a vertical difference in position corresponds to a *% difference* -rather than a *difference*- in value). By visual inspection of Fig. 9 and Fig. 10 we can make four important, interrelated observations:

1. The average heart rate and average power steadily decline with age.
2. The effect of aging on cardiac efficiency appears to be very small, if any.
3. Periodic variation in heart rate, which may be attributed to seasonal changes in fitness level, is hardly noticeable in power, speed or cardiac efficiency.
4. It appears as though my speed declines with age at a lower time rate than my power output.

These findings are quite significant, and the fact that they are uncovered merely by looking at general-purpose cycling data from actual recreational rides, without the need for formal scientific studies conforming to academic standards and prescribed test protocols, is remarkable.

It is important to keep facts and interpretations separate. The finding that my heart rate and power have declined over the last 7 years is a fact. It is not *proven* to be the result of aging, but that is the most reasonable interpretation of the available data, given that I am 68 years old. The finding that in spite of the age-related declines in power and heart rate, there is no noticeable change in cardiac efficiency is also a fact, and it is quite educational. The seasonal modulation in my average heart rate during my bike rides is also a fact, and that these changes are noticeable by visual inspection of the raw data is also remarkable. But the attribution of this fact to changes in fitness levels is only a theory at this point. An alternate explanation could be that when I ride in cold weather I generate more *thermal power*, and my preferred workout intensity depends on the mechanical power rather than the sum of mechanical and thermal powers.

There are still unanswered questions in need of further investigation: If over the years I am losing power but not speed, where does the extra energy come from? Well, the answer can be found by simply looking at the scatter of Figure 11. Apparently as I get older I seem to prefer flatter routes more often.

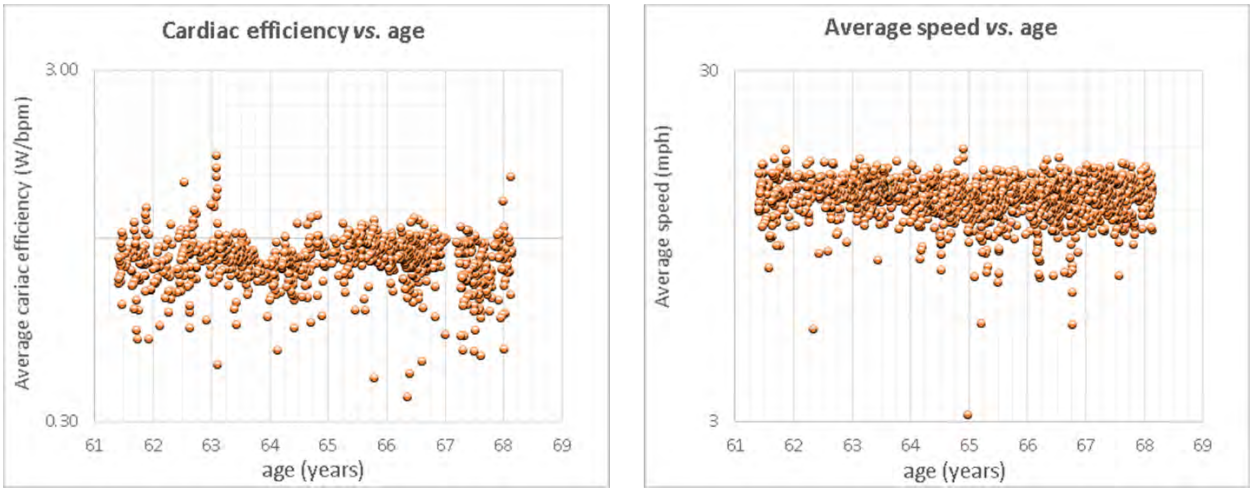

Figure 10 –Cardiac efficiency and average speed as a function of my age at the time of each ride

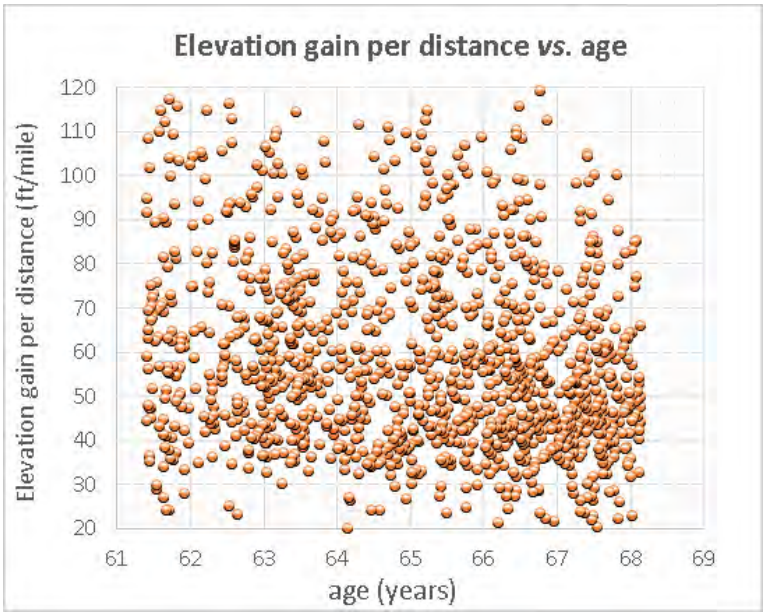

Figure 11 –climbing preference vs. age

E-BIKES: The majority of e-bikes on the mainstream market today are designed with the notion that people who would buy them are incapable of making or unwilling to make their own decision about how much electric assistance is appropriate under dynamically changing conditions. As a result, with a few exceptions, most e-bikes don't even have a thumb lever, twist grip or similar input device to let the user control the amount of electric power. Instead, an automaton decides how much power the user needs, depending on feedback from cadence and torque sensors. The user is given some choice in the matter, but it is limited to a "mode" function, which is often not even located on the handlebar, that is basically a multi-position switch with a poor user interface (a touch sensitive button coupled to some LEDs to virtually simulate a physical selector) that is meant to be operated in a "set and forget" mentality rather than for active engagement. As a result, firstly, these e-bikes serve only a self-selecting market of cyclists to whom this mentality is appealing; and secondly, the artificial intelligence that decides how much electric power is appropriate in a given situation is crucial for the satisfaction of the user and hence the success of the product. Therefore, I think that any knowledge about the statistics of what unassisted cyclists are doing under which set of conditions regarding hill slope, wind, cadence, crank torque and wheel torque, is of great importance for e-bike manufacturers.

SECURITY AND PRIVACY OF CYCLING DATA: According to Yahoo Finance, as of February 2020 Strava had 50 million users (adding about 1 million each month), and the total number of uploaded activities had reached 3 thousand million. So, I am one of 50 million users, and I have uploaded 1,679 of the 3 thousand million activities. Ok, I have uploaded more than my share of data (the average Strava user has uploaded 60 activities apparently), but to put it into perspective, the database that is available to Strava is 1.8 million times bigger than what you see here. However we estimate it, the commercial value of this data must be staggering. This not only raises security concerns, but also leads to philosophical and ethical questions, including: "whose property is it?" My account is password protected, ostensibly implying that I own the contents. Of course I am sharing my data in excruciating detail here; but nevertheless, if I were notified of a security breach where unauthorized persons have accessed my data, I would not be happy about it. What would have upset me in that hypothetical situation isn't only the

risk of malicious use of the information. For cycling data, that risk is fairly low compared with, for example, credit card transactions, confidential e-mails or embarrassing photos. But I would feel violated because I consider "my data" personal. Perhaps this mindset is a remnant from a bygone era: we tend to liken digital storage to a depository of possessions that belongs to *us*, like a safe deposit box where we keep our jewelry. But the analogy is not quite accurate because unlike jewelry, data can be duplicated at virtually no cost. When someone "steals" our data, we still have it. Yes, information is power, and the question is who has the authority to deny access to it, and why. Regardless of our views on this ethics question, to decide how strictly we want to protect our cycling data from others, we should appropriately weigh the value of the accessibility of our data to researchers. Anonymity of the data is tricky to define and control, but imagine the added benefit if each numerical value in this article were averaged from 1.8 million times as many data points! Imagine if I could increase the resolution of any of the graphs here, and/or filter the underlying data by geographic regions, age and gender categories, weather conditions, cycling experience, fitness level, proficiency level, wind speed and a myriad of other variables, and still have statistically significant numbers of rides to average! So, as a general rule, when I ride my bike I almost always upload my data, and I make it available to everyone.

SEGMENTS: So far I have presented my cycling activities with a data resolution where each data point is a bike ride. But the data files that I (and apparently 50 Million other cyclists) upload to Strava (not to mention Garmin Connect, Ride with GPS and others) contain much more information than that. Within each bike ride there are dozens (and often hundreds) of "segments". A Strava segment can be any section of a road. It can also be a combination of roads that comprises a portion of a route, or it can even be an entire route. Climbing segments are automatically defined and categorized, but any user can define custom segments that will appear in their ride data as well as in those of other users who have ridden on that segment. Because I ride the same roads again and again, many of the 1,348 bike rides which I have analyzed for this article have common segments. When I statistically analyze my cycling data at *segment resolution*, I am accessing orders of magnitude more data because there are many more segments than rides, and because I have ridden each segment multiple times. There are some segments near my home which I have done, for example, more than 300 times. Figure 12 is given as an example to show the potential for analyzing cycling data at the segment level.

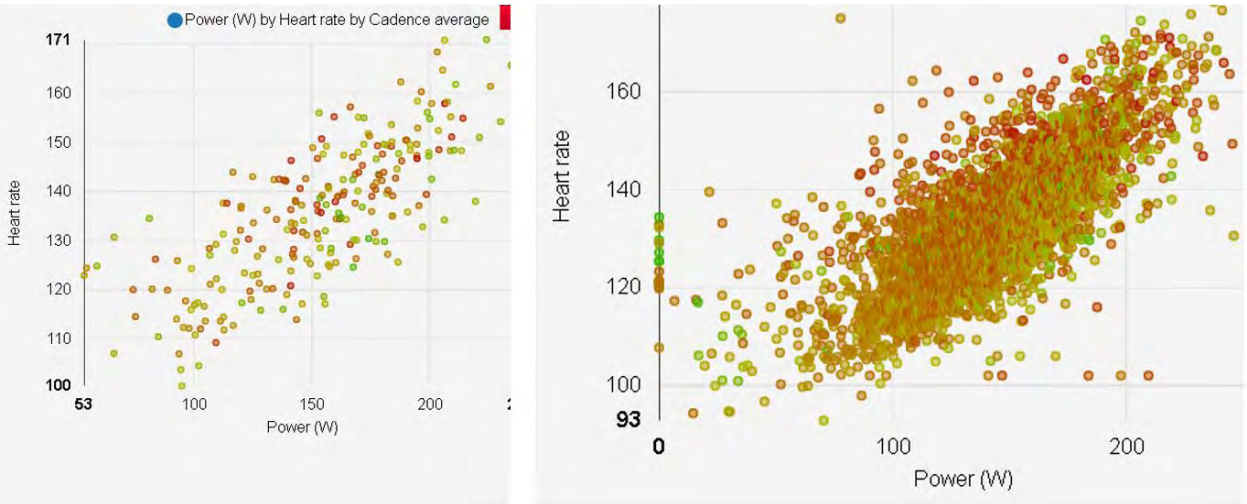

Figure 12 – Example of heart rate vs. power data on selected segments. Left: 300-320 feet of elevation gain; Right: 300-600 feet of elevation gain. Screenshot from Veloviewer. Color represents cadence.

Here, the scattered dots are not rides; they are *climbs.* More specifically, these are a small subset of the climbs within a specific range of elevation gain that I did with the same bicycle and with a power meter. There are several dozen climbs on almost every one of my rides. The scatter graph on the left represents 233 of those climbs that have 300-320 feet of elevation gain; the right one represents 3,204 climbs with 300-600 feet of elevation gain. Some of my climbing segments have more than 1,000 feet of elevation gain. Some segments are flat and some are downhill.

Average cadence is color coded and varies from light green to dark brown. Data for all segments would have been too numerous to plot.

The analysis of ride data at *segment resolution* may be the subject of a future article.