

Utilizing the Microbiome

A Data Scientist Trek to better enlightenment
or *a mentorship* discussion

Presenters:
Ken Lassesen, M.Sc.
Bhoumik Mishra, AI/ML



The Basic Journeys

› Determine which taxa/bacteria to address

› Determine how to modify the selected taxa

› Validate the intervention worked

› Questions And Answers





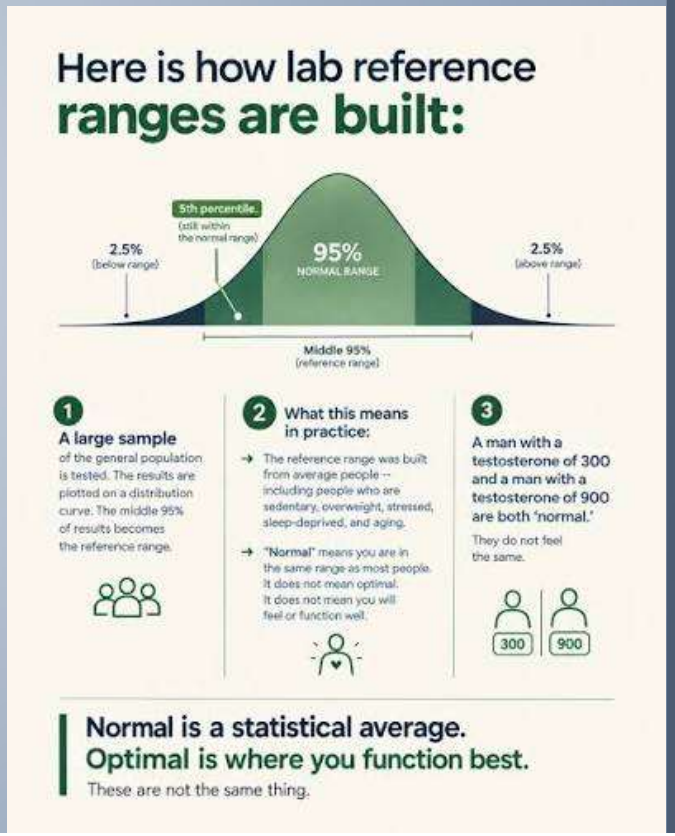
Taxa Selection

- Single bacteria associated – trivial
- Multiple associations or no known associations
  - Naïve: Assume a bell curve, pick taxa > +/- 1.6 Std Dev
  - Non-parametric: Kaltoft-Moldrup curve
    - Where values are dissimilar to its neighbours
  - Regression models to symptoms
  - Weak Associations via means (studies based)
- Odds Ratio for a condition or for “healthy”
  - Incidence / presence–absence analysis
  - Ranges

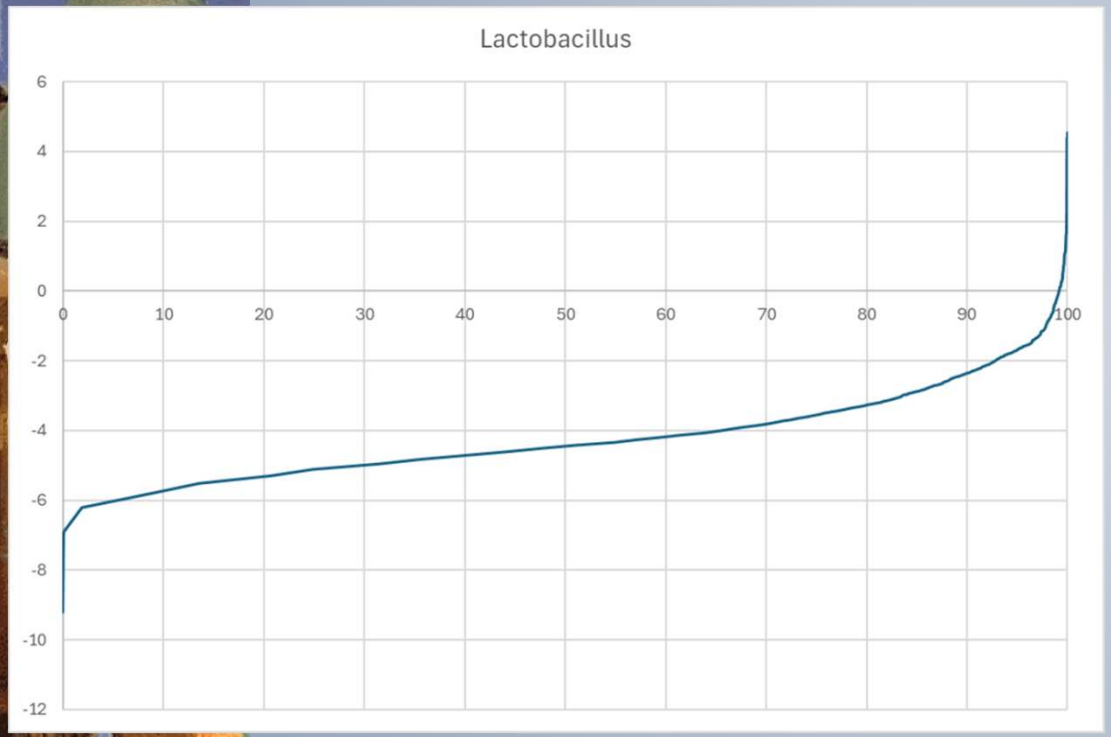


Bell Curve And Variants

- Follows the usual practice for testing labs (“tradition”).
- ***Variant***: Use the top/bottom 2%ile and 5%ile (observed values)
- Business Motivations
  - Cheap: small sample to get range
  - Easy to explain to customers / staff
  - Sufficient for most customers



Using Behavior: Kaltoft-Moldrup



- From stock trading
- Looks for points outside of typical
- Process:
  - Apply monotonic Increasing transform and Order
  - Most of data is on a flat line
  - When 2nd derivative deviates from near zero, that is a behavior change point.
    - < 3%ile
    - > 75%ile



Study Based: Weak Associations

- Requires patient to have diagnosis matching study
  - Must use the *identical microbiome processing flow as study*
  - Shotgun results can not be used with 16s results
  - Often “average below/above control” is net report
- Example: https://pubmed.ncbi.nlm.nih.gov/40971415/
  - *Therapy-Naïve de Novo Pediatric Inflammatory Bowel Disease*

Results: Alpha diversity progressively decreased from HC to CGI ($P < .001$) and decreased even further in IBD patients ($P = .0056$). Beta diversity analysis showed significant clustering differences ($P < .001$, $R2 = 0.045$). Differential abundance analysis revealed 116 species differing between HC and IBD, and 30 species between CGI and IBD. Prediction models based on microbiome features accurately distinguished IBD from HC (area under the curve [AUC] = 0.96) and from CGI (AUC = 0.71). However, these models were outperformed by clinical features, such as fecal calprotectin. Microbiome-based prediction of response to induction therapy in general showed limited accuracy (AUC = 0.63), as well as for response to nutritional induction therapy (AUC = 0.67).



Every study is “fuzzy”

- The taxa named in a study may be:
  - Not reported on with your processing
  - May be identified as a different taxa with your processing
  - May depend on age, gender and existing conditions
- For Interventions, we will assume “central limit theorem”

> “There are currently 97 different ways to analyze the same raw data, and they will give you 97 different answers.”
>
> — **Scott Jackson,** *leader, NIST Complex Microbial Systems Group*



Build Regression / Forecast model

- You need annotated samples
  - Typically a lot of samples
  - 75 - 90% should be accurately forecast
- Linear Model is simplest:
  - Increased risk of symptom, reduce
  - Decreased risk of symptom, increase
- Downstream benefit is ‘objective’ measurement of change

Symptom Pattern Matching

Show 10 entries Search:

Symptom	New Sample	Change	Old Sample
Virus: Human Herpesvirus 6 positive (HHV6)	77.2	↓	89.8
DePaul University Fatigue Questionnaire : Weight change	71.4	↓	77.2
DePaul University Fatigue Questionnaire : Blurred Vision	65.8	↓	79.2
Neurocognitive: Difficulty paying attention for a long period of time	65.6	↓	72.3
Neurological: Impairment of concentration	65.4	↓	70.9
Physical: Amalgam fillings	65.2	↓	72.3
Condition: Acne	64.5	↓	68.9
DePaul University Fatigue Questionnaire : Pain in Multiple Joints without Swelling or Redness	64	↓	86.3
Neurological: emotional overload	61.8	↓	75.7
Immune Manifestations: general malaise	61.6	↓	70.6
Better	103	Worse	69



Presence–Absence Analysis

- Using Odds Ratios, good prediction results may be obtained
  - Presence that is harmful is easy to reduce
  - Absence that is harmful is untreatable
  - Presence that is beneficial is unclear if encouragement is best
  - Absence that is beneficial is moot
- Useful for classification,
  Questionable for intervention



Odds Ratio Ranges

- Often provides good and bad ranges per Taxa
  - Also has no **significant impact** at P < 0.01 or 0.001
- If in good range, is there a better range to move to?
- If in bad range:
  - If a good range exists, move towards that
  - If no good range exists, move to no **significant impact** range
  - If neither, try to eliminate it
- If in not significant
  - If a good range exists, move towards that

Taxa Examples

This is the result for Ruminococcus (NCBI Taxon 1263)

Start	End	Chi2	Odds
20	45	8.44	0.21
50	75	15.13	3.02
55	80	8.06	2.30
55	75	8.17	2.37
60	80	12.27	2.78
65	75	10.69	3.02
70	80	6.75	2.51
70	75	10.23	3.59



Step 1 Completed: Taxa Selected with Direction

- How do you codify an intervention studies as a number?
  - Positive, increases taxa
  - Negative, decrease taxa
- Handling Contra Results
  - You have 8 studies, 4 reports increase and 4 reports decrease
  - You have 31 studies, 29 reports increase and 2 reports decrease
- Options
  - Use a formula factoring in date of study, testing methodology, number of studies, etc?
  - Naïve (ignore number of studies):
    - Do a simple average: 0 and 0.93
    - Do a winner takes all: 0 and 1 (average < 0 ➔ -1, average > 0 ➔+1)



Contra Interventions and Avoids

- With a set of Taxa and a set of interventions (IV), we may have:
  - IV moves all taxa in right direction when shift known
    - Safest: minimum risk of harm given knowledge available ➔ Take
  - IV moves all taxa in wrong direction when shift known
    - Danger: high risk of harm ➔ Avoid
  - IV moves some in right direction and some in wrong direction
    - More complex situation to evaluate, some harm risk.



Contradictory Interventions

- Like Taxa, interventions often have a relationship structure.
  - Compounded by fuzzy definitions for some interventions
- Diets are particularly challenging
  - Diets: for example, Mediterranean dietary pattern
  - Whole grains, gluten free and explicit grain may conflict
    - Barley is reported to increase(5 studies) or decrease(6 studies) Bacteroidaceae
    - Oats is reported to increase(1 studies) or decrease(2 studies) Bacteroidaceae



Mediterranean dietary pattern

- Diet definitions can have significant differences
- Greece often include fast days

Feature	Greece	Italy	Spain	U.S. Mediterranean-style
Overall emphasis	Vegetables, greens, legumes, olive oil, and fermented dairy such as feta and strained yogurt. onlinelibrary.wiley +1	Pasta, bread, vegetables, beans, and olive oil; the classic pattern is particularly associated with southern Italy. onlinelibrary.wiley +1	Grains, legumes, vegetables, olive oil, and a strong seafood component. onlinelibrary.wiley +1	The shared nutritional principles translated into accessible foods: vegetables, beans, whole grains, nuts, olive oil, and fish. clevelandclinic
Grains and starches	Bread and barley rusks are characteristic examples; grains accompany vegetables and legumes. oldwayspt	Pasta has a prominent role, including combinations with beans and vegetables—not simply large pasta portions with meat or cream. onlinelibrary.wiley +1	Bread and rice-based dishes are characteristic, alongside other grains. onlinelibrary.wiley +1	Often explicitly emphasizes whole-wheat bread and pasta, oats, brown rice, or quinoa rather than requiring traditional Mediterranean grains. clevelandclinic
Dairy	Feta and strained yogurt are particularly characteristic. onlinelibrary.wiley	Regional cheeses appear, for example mozzarella with tomatoes and bread; amounts matter more than whether a cheese is "Italian." oldwayspt +1	Dairy is not the defining feature; grains, seafood, and plant foods are more prominent in descriptions of the traditional pattern. onlinelibrary.wiley	Plain yogurt and moderate cheese are commonly included; some U.S. heart-health guidance emphasizes low-fat or fat-free dairy. clevelandclinic +1



The Double and Triple Count Challenge

- Bacteria/taxa has a hierarchy
  - Associations may be found at various levels on a branch
  - You may have interventions for each level
- Are you doing Forecasts?
  - With Odds Ratio you may be inflating/deflating Odds
- Do you do all interventions?
  - Potentially *double counting* the impact?
  - Do you go with the highest or lowest level
- Each taxa level has a percentage that do not have children identified
  - Sum of Bifidobacterium Species may be just 40% of the Genus



Fuzzy data is every where!

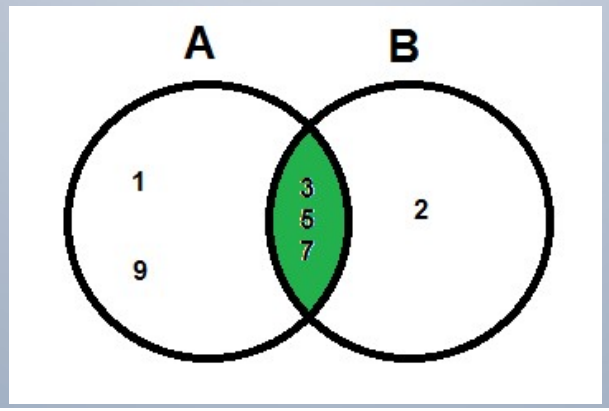
- Which taxa was identified is fuzzy
- The impact of an intervention is double fuzzy!
  - Taxa impacted is fuzzy
  - Intervention may be fuzzy
- Is intervention study applicable?
  - On Chinese peasants?
  - On Vegan Indians?
  - On Texan Males?
  - On mice?
  - On pigs?
  - On zebra fish?

- **Type-1:** Each input gets one definite membership value—for example, “hot = 0.7.” This represents vagueness using a precisely specified scale
- **Type-2:** The membership value itself is uncertain. In the common *interval type-2* version, “hot” might have membership between 0.6 and 0.8, rather than one fixed value
- **Type-3:** extends type-2 by adding another layer of uncertainty: the membership assigned to an input is itself a type-2 fuzzy set, rather than a number or a type-1 fuzzy set.



Too many masters! (Symptoms)

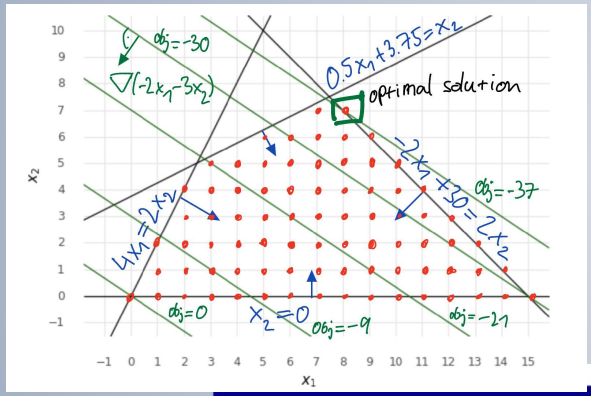
- It is common for a person to have multiple symptom
- Each symptom may be associated to different taxa, and thus different interventions
- Resolving by interventions (and not taxa) is elegant solution
  - Ideally, interventions that works for everything with no adverse.
  - Do a ***Union*** of taxa-to-intervention and use that for other interventions





Academic versus Clinical Presentation

- Academic may be presented as mixed-integer programming program using probability density functions



- Clinical needs to be easy for a patient to understand

Positive Impacts	Negative Impacts	Modifier
4	13	bifidobacterium
2	11	Malus domestica {apple}
7	15	bacillus,lactobacillus,streptococcus,saccharomyces probiotic
6	14	Saccharomyces cerevisiae var boulardii {S. boulardii}
6	14	resistant starch
6	14	Lacticaseibacillus rhamnosus {l. rhamnosus}

Positive Impacts	Negative Impacts	Modifier
4	0	laminaria hyperborea {Cuvie}
4	0	Carica papaya {papaya}
4	0	Hibiscus {Rose mallow}
3	0	Rhus chinensis {galla rhois}
2	0	Eucalyptus {Gum Tree}
2	0	Tudca
2	0	Cymbopogon citratus {Lemongrass}



Validating the intervention

- Subjective “I feel better” is not reliable
- Patient captured data (number of stool movement per day) is suspect
- If we can accurately forecast, the change of forecast is not subject to the human factor
  - Regression models
  - Odds Ratios

Symptom Pattern Matching

Show 10 entries Search:

Symptom	New Sample	Change	Old Sample
Virus: Human Herpesvirus 6 positive (HHV6)	77.2	↓	89.8
DePaul University Fatigue Questionnaire : Weight change	71.4	↓	77.2
DePaul University Fatigue Questionnaire : Blurred Vision	65.8	↓	79.2
Neurocognitive: Difficulty paying attention for a long period of time	65.6	↓	72.3
Neurological: Impairment of concentration	65.4	↓	70.9
Physical: Amalgam fillings	65.2	↓	72.3
Condition: Acne	64.5	↓	68.9
DePaul University Fatigue Questionnaire : Pain in Multiple Joints without Swelling or Redness	64	↓	86.3
Neurological: emotional overload	61.8	↓	75.7
Immune Manifestations: general malaise	61.6	↓	70.6
Better	103	Worse	69



Questions and Answers

- Who am I?
- Microbiome Attributes
- LLM and current hot AI
- The scale of data…



Who am I?

- M.Sc. in Statistics and Operations Research,
  - Some PhD. Courses
  - University Instructor of Computer Science, including AI
  - High School Science Teacher, College Instructor of Physics and Chemistry
  - Many decades of experience doing performance statistics
  - Programming since 1969
- Became interested in the microbiome due to health issues in family
  - Saw too much speculation not supported by reasonable evidence



› Microbiome Attributes

- › The more attributes used, the more samples needed
- › Do not target sample sizes suitable for a normal distribution
- › Design UI for ease of finding
  - Hierarchy is recommended
  - Use *common language* not IDC names

Autism

- ☐ High Functioning
- ☐ Less Abnormal Body Posturing or Facial Expressions
- ☐ Less Abnormal Tone of Voice
- ☐ Less Avoidance of Eye Contact or Poor Eye Contact
- ☐ Less Behavioral Disturbances
- ☐ Less Decrease in Speaking
- ☐ Less deficits in Language Comprehension
- ☐ Less Flat or Monotonous Speech
- ☐ Less Inappropriate Social Interaction
- ☐ Less Intense Focus on One Topic
- ☐ Less Lack of Empathy
- ☐ Less Repeating Words or Phrases
- ☐ Less Repetitive Movements
- ☐ Less Self-Abusive Behaviors
- ☐ Less Sleep Disturbances
- ☐ Less Unusual Responses to an Unexpected Event
- ☐ More Abnormal Body Posturing or Facial Expressions
- ☐ More Abnormal Tone of Voice
- ☐ More Avoidance of Eye Contact or Poor Eye Contact
- ☐ More Behavioral Disturbances
- ☐ More Decrease in Speaking
- ☐ More deficits in Language Comprehension
- ☐ More Flat or Monotonous Speech
- ☐ More Inappropriate Social Interaction
- ☐ More Intense Focus on One Topic
- ☐ More Lack of Empathy
- ☐ More Repeating Words or Phrases
- ☐ More Repetitive Movements
- ☐ More Self-Abusive Behaviors
- ☐ More Sleep Disturbances
- ☐ More Unusual Responses to an Unexpected Event
- ☐ Spectrum (not full)



LLM and current hot AI

- Does it handle type-1,2,3 fuzzy logic?
  - Most AI use quasi-binary logic
  - Can you train it on the issues?
  - Prone to over-generalization
- Useful for identifying studies *for human review*

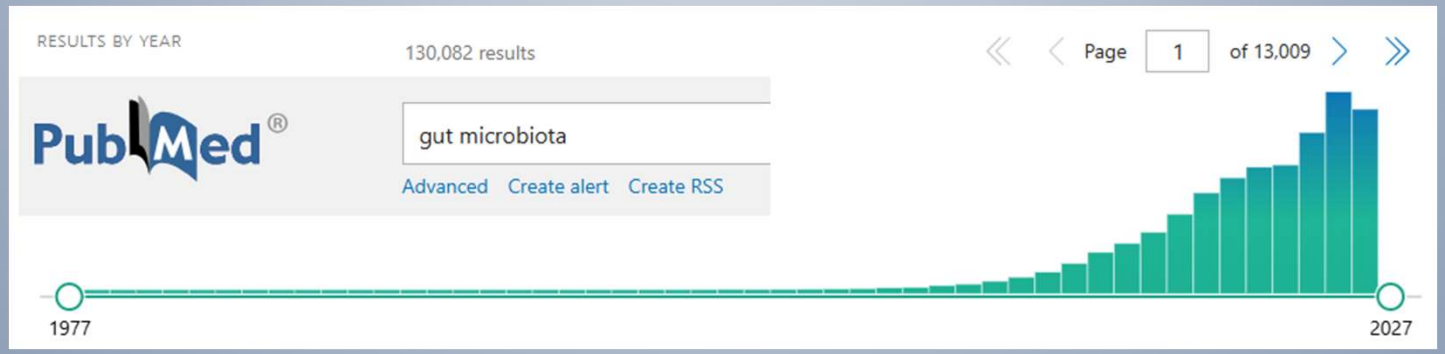
Answer from Perplexity

My practical assessment: use an LLM to propose linguistic categories, draft rules, and explain results; execute the membership calculations and rule aggregation in explicit code.



The scale of data involved

- Taxa: over 3,016,439 different ones on NCBI
- **KEGG: Kyoto Encyclopedia of Genes and Genomes**
  - Enzymes: over 8,300
  - Compounds: almost 20,000
  - *Better associations often seen from these*
- Possible Interventions: over 2,000
- Pub Med studies with *useful* data: over 22,000 studies





Final Take Away

- Avoid excluding because:
  - Too complex – learn skills needed!
  - Too resource intensive – do the trek!
  - Many paths lead to better gut health
    - Most are narrow grueling paths
    - Easy paths rarely help people
- There is no ideal model
  - Be prepared to explain your model
  - Consider dual models
    - One that is academic and best attempt
    - One that is simple for a client to understand
    - Present the simple model after checking for reasonable agreement with academic model



“primum non nocere”

- “Do no harm” means: before acting, consider whether your action could cause unnecessary injury, suffering, or damage—and avoid it when the risks outweigh the benefits. In medicine, it is the ethical principle of **non-maleficence**

A clinician applies it by asking:

- Will this intervention help more than it hurts?
- Are there safer or less invasive alternatives?
- Are the risks clearly communicated and consented to?
- Could inaction itself cause avoidable harm?
- Is the treatment still appropriate as circumstances change?