# Machine Learning and Prediction of Antimicrobial Resistance (AMR)

Sam Zhu
4th year Ph.D. student
Supervisor: Professor Margaret IP

Joint Graduate Seminar

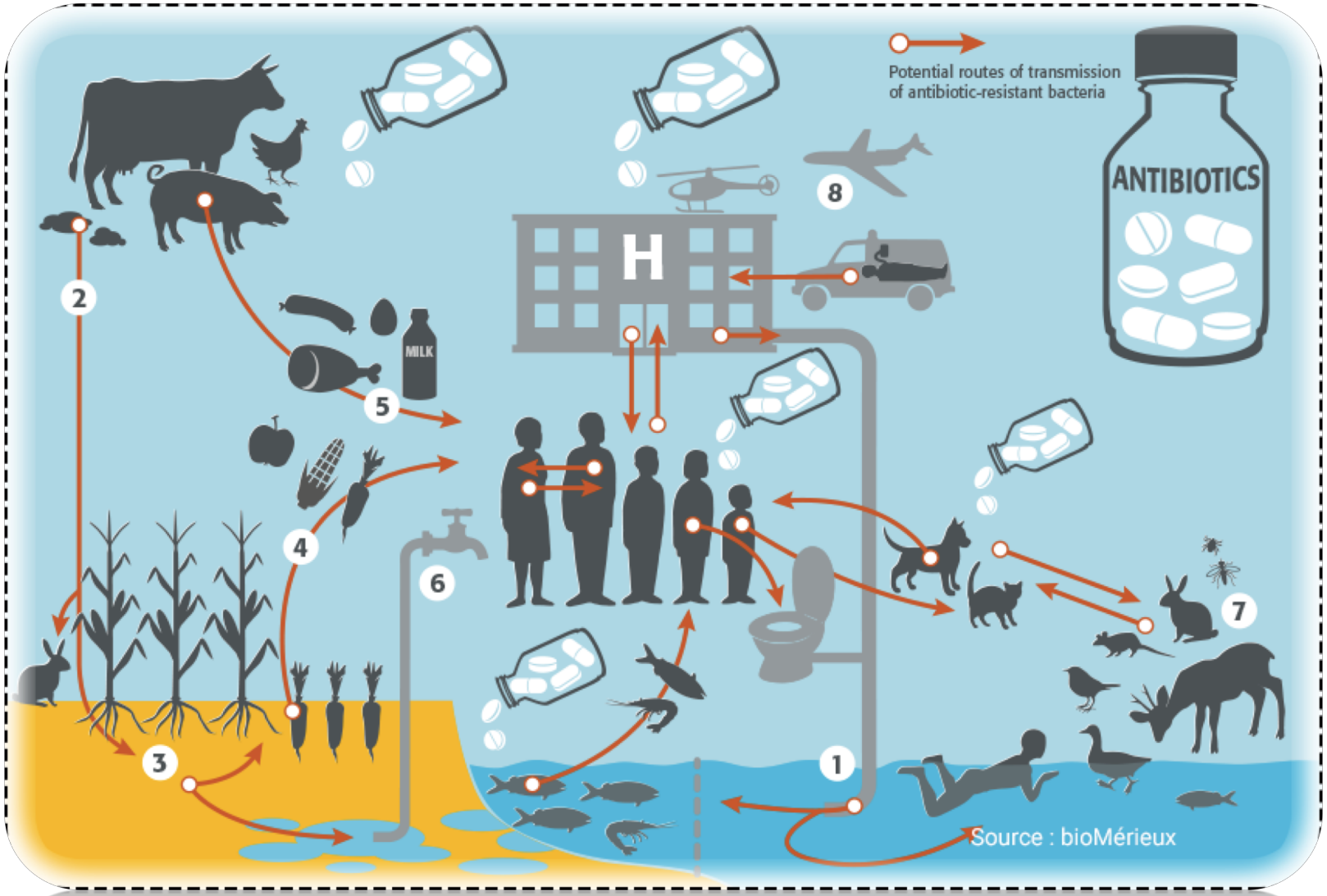
Department of Microbiology

13th December 2018

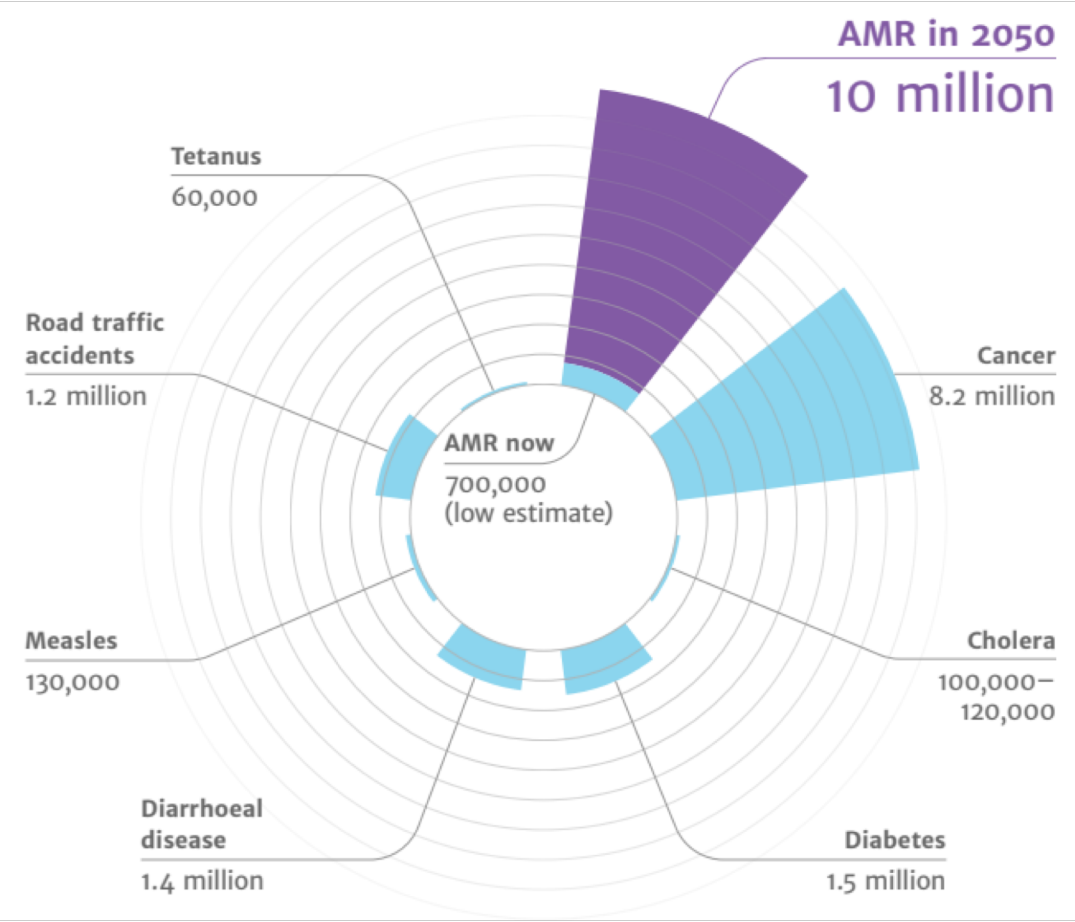
# Outline

- Burden and introduction of AMR
- From genotype to phenotype
- Application of machine learning

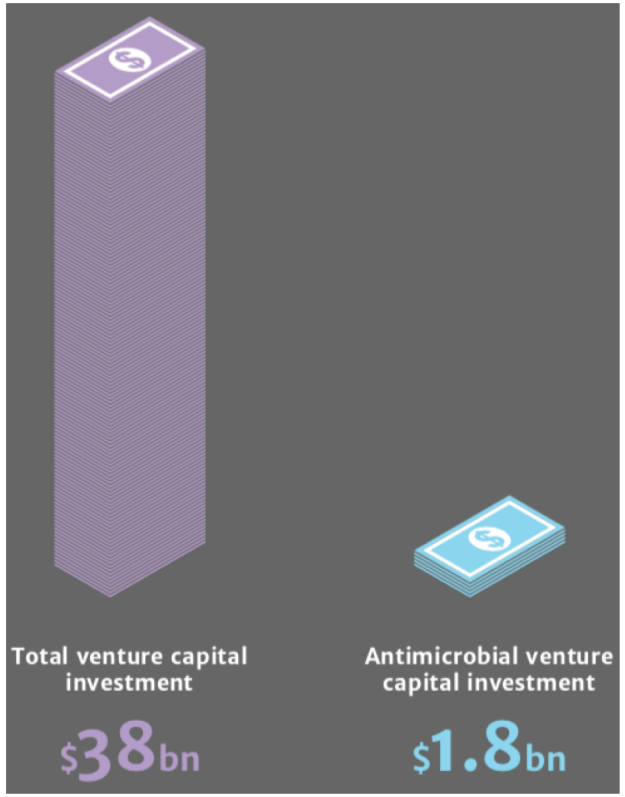
# Potential routes of transmission of AMR bacteria

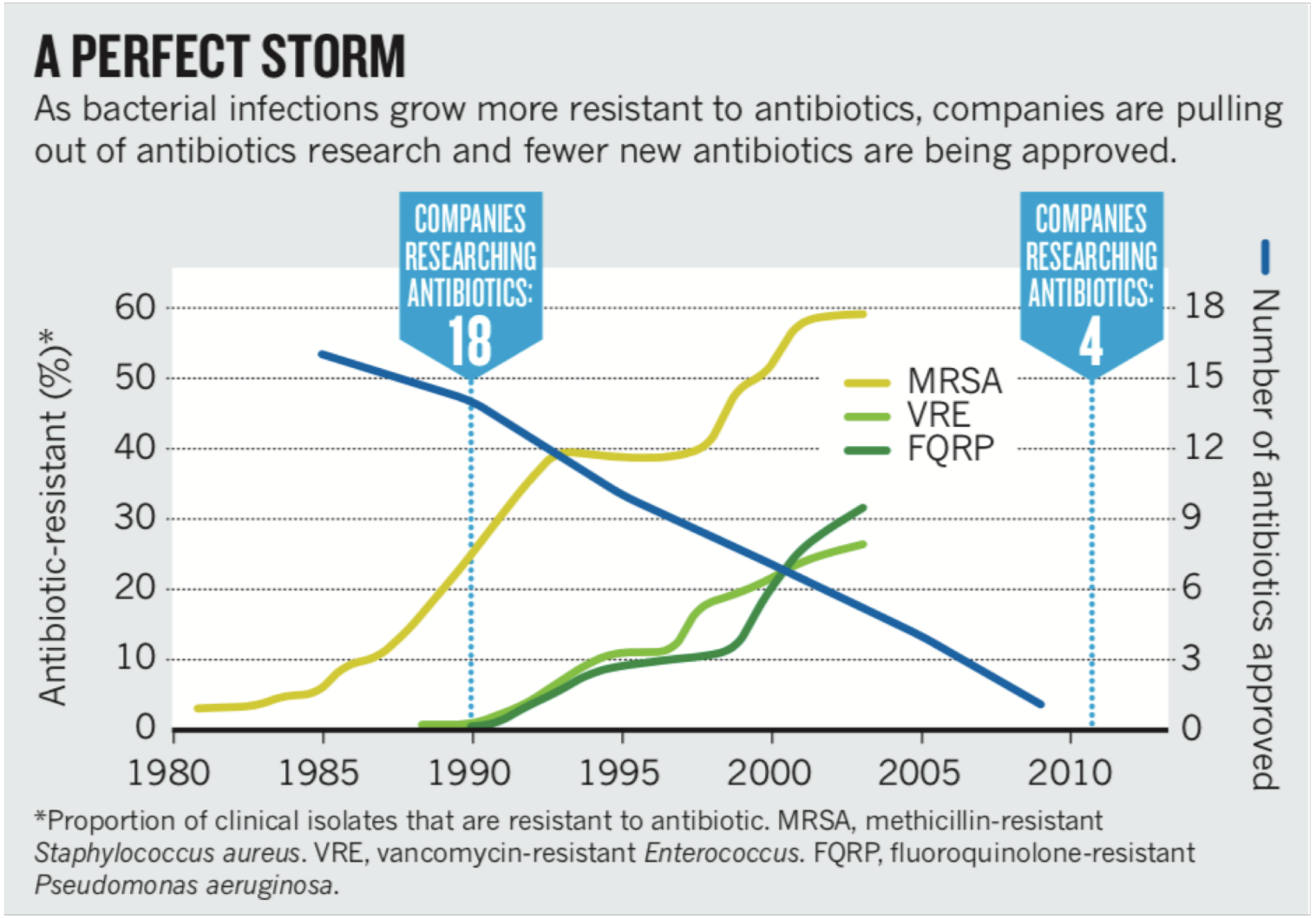


# Death attribute to AMR

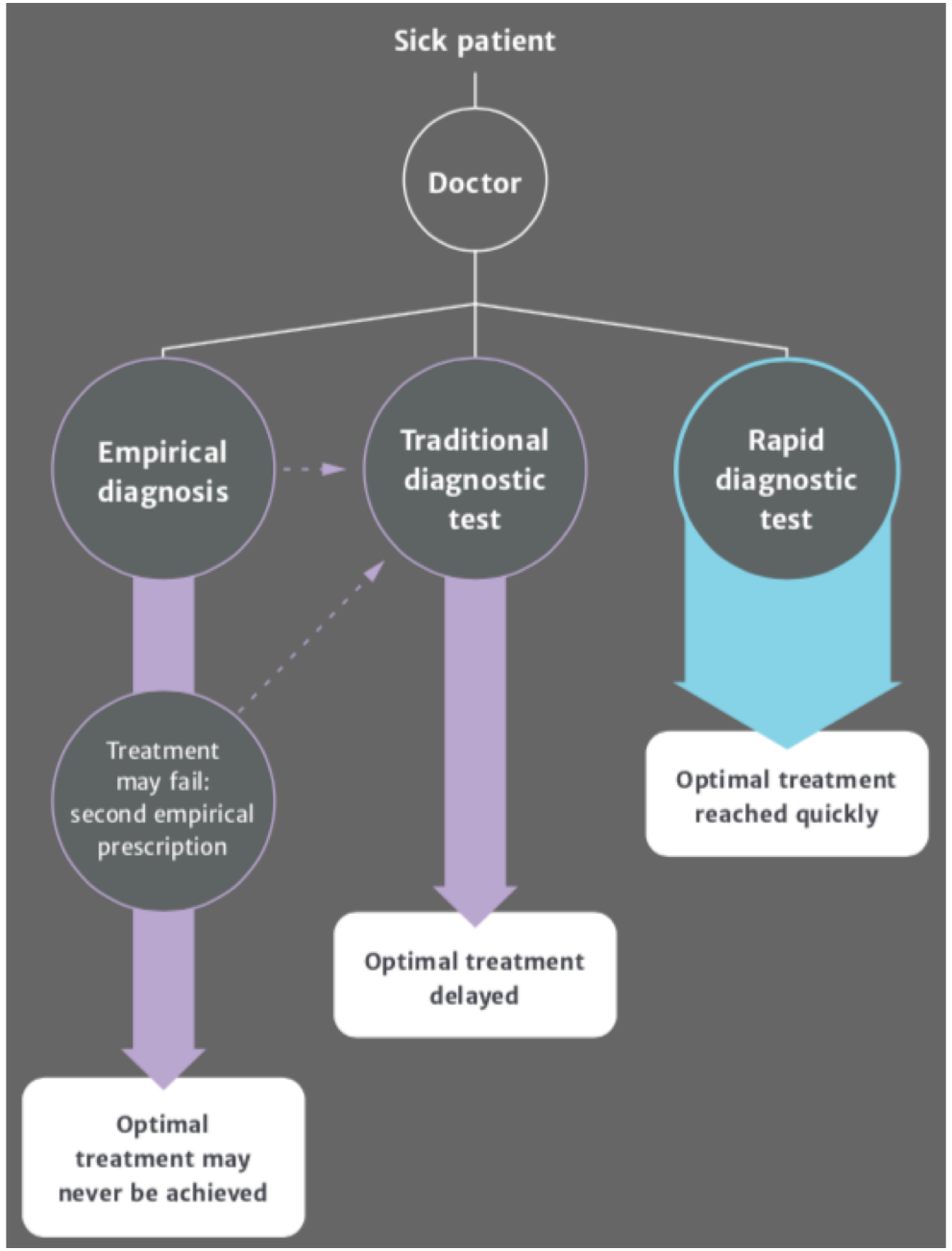


# Research & Development in antibiotics



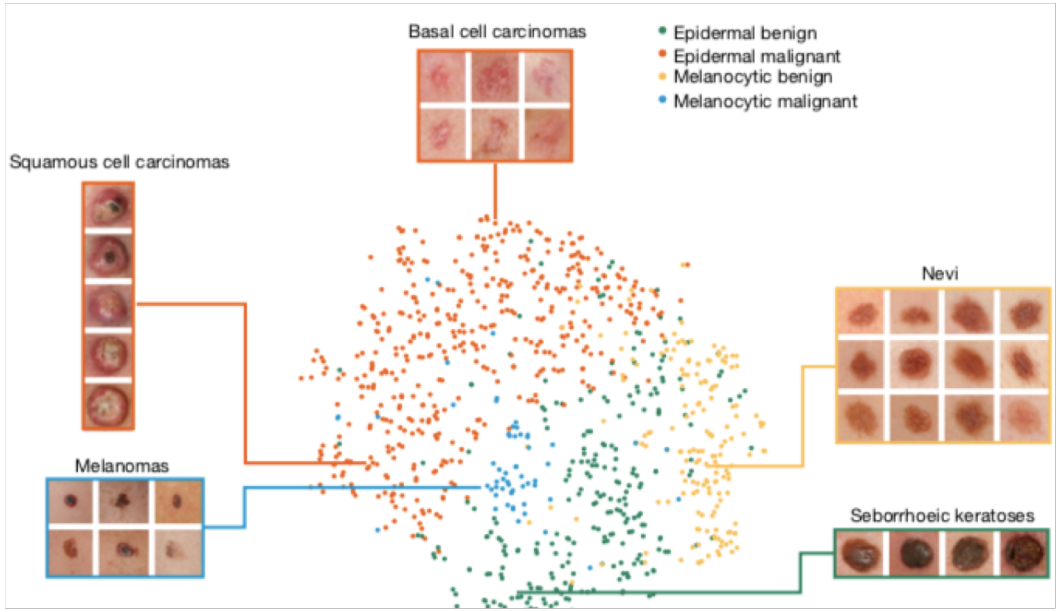




O'Neill 2016, Cooper 2011




# Rapid diagnostics to optimise treatment

**Machine learning application in dermatology, histopathology images**




O'Neill 2016, Andre Esteva 2017, Dmitrii Bychkov 2017

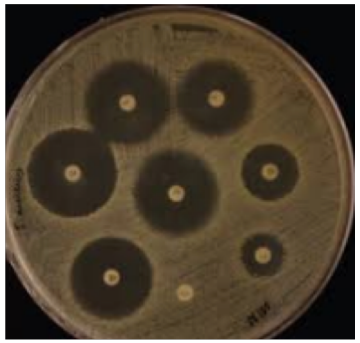
# Antimicrobial Susceptibility Test: Methods

1. Broth Dilution

0.125|0.25|0.5|1| 2 | 4 | 8 | 16 |32|64|

Minimum Inhibitory concentration (MIC)

2. Antimicrobial gradient diffusion, disk diffusion test



3. Automated instrument systems

Phoenix

Vitek 2

Sensititre ARIS 2X

# Antimicrobial Susceptibility Test: Criteria



Interpretive criteria by CLSI/EUCAST based on:
(1) microbiologic data
(2) pharmacokinetic and pharmacodynamic data (PK/PD)
(3) Clinical study results

Sensitive(S) Intermediate(I) Resistant(R)

# Class II Special Controls Guidance Document: Antimicrobial Susceptibility Test (AST) Systems

FDA U.S. FOOD & DRUG ADMINISTRATION

Performance:

1. Essential Agreement (EA): > 90%

   exact agreement or within ± one two - fold dilution of the reference method

2. Discrepancy – major: ≤ 3%

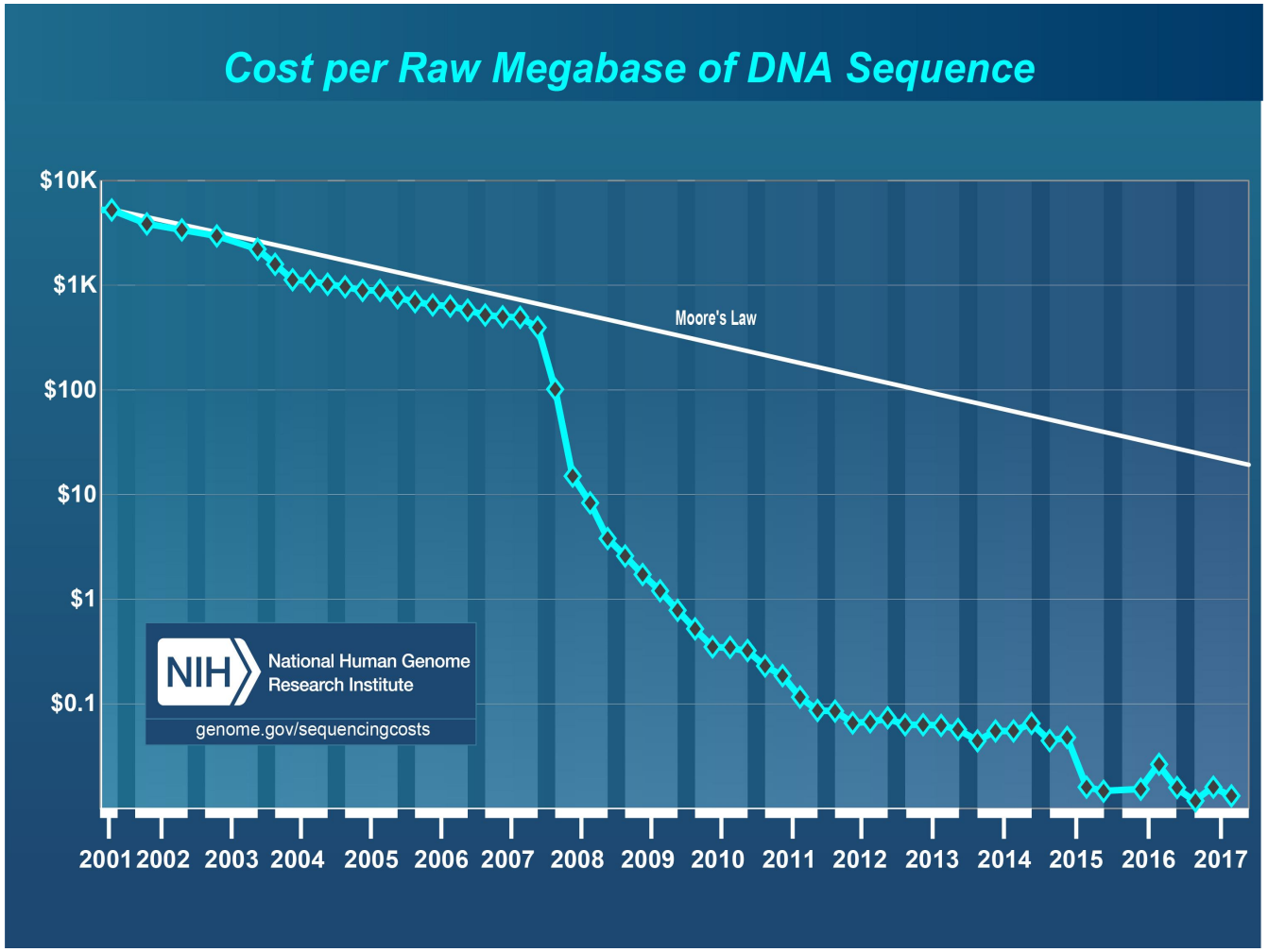
   The reference category result is S and the new device result is R.

3. Discrepancy – very major: based on population

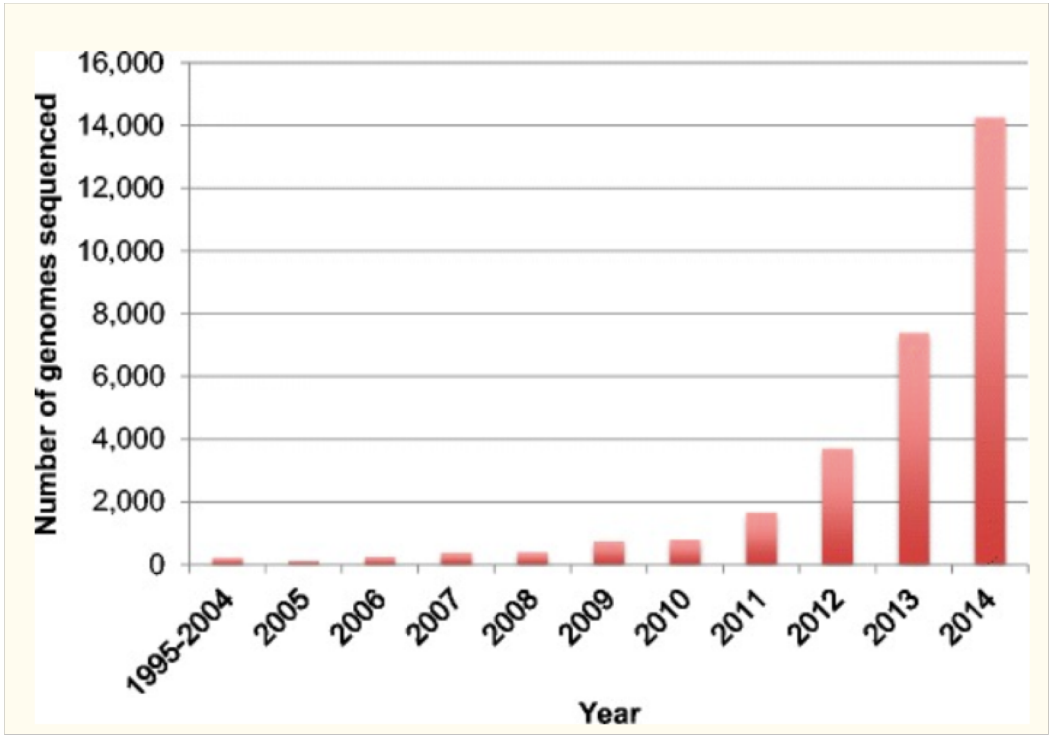
   The reference category result is R and the new device result is S.

| Number of Resistant Organisms | Acceptable Number of Discrepancies | Estimated Rate[a] | 95% Confidence Interval[b] for True VMJ Rate |
|---|---|---|---|
| 48 | 0 | 0.00 | (0.00, 7.40) |
| 50 | 0 | 0.00 | (0.00, 7.11) |
| 60 | 0 | 0.00 | (0.00, 5.96) |
| 70 | 0 | 0.00 | (0.00, 5.13) |
| 72 | 1 | 1.39 | (0.04, 7.50) |
| 80 | 1 | 1.25 | (0.03, 6.77) |
| 90 | 1 | 1.11 | (0.03, 6.04) |
| 94 | 2 | 2.13 | (0.26, 7.48) |
| 100 | 2 | 2.00 | (0.24, 7.04) |
| 110 | 2 | 1.82 | (0.22, 6.41) |
| 120 | 3 | 2.50 | (0.52, 7.13) |
| 130 | 3 | 2.31 | (0.48, 6.60) |
| 140 | 4 | 2.86 | (0.78, 7.15) |

FDA, 2009




Number of bacterial and archaeal genomes sequenced each year and submitted to NCBI.




Land, 2015

# Susceptible or Resistant based on presence and absence of AMR genes

| Year | Bacteria | Model | Antibiotics | EA | ME | VME |
|---|---|---|---|---|---|---|
| 2013 | 74 *E. coli*<br>69 *Kpne* | - | 7 | 96% | 2.1% | 1.2% |

| Species | Antibiotics (ME) | Genotype | Phenotype |
|---|---|---|---|
| *E. coli* | Ceftazidime (15%) | blaTEM, blaCTX-M | S |
| *Kpne* | Amoxicillin (4%) | blaSHV | S |

| Species | Antibiotics (VME) | Genotype | Phenotype |
|---|---|---|---|
| *Kpne* | Ciprofloxacin (6%) | - | R |

EA: essential agreement >90%
ME: major error S->R ≤ 3%
VME: very major error R->S < 3%

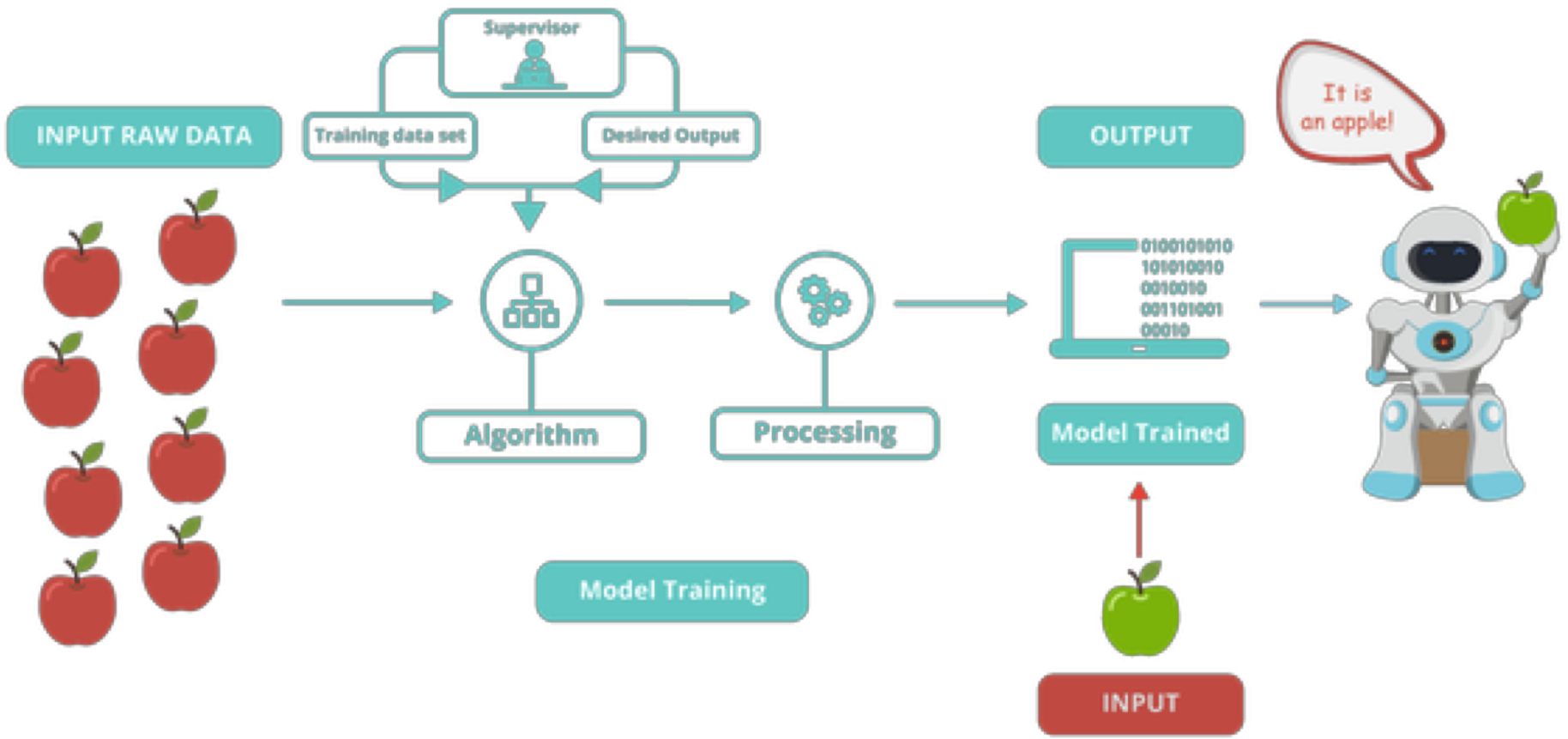
Others: *S. aureus*, MTB, Nontyphoidal *Salmonella*…

Stoesser 2013, Bradley 2015, Gordon 2014, McDermott 2016

| Minimum Inhibitory concentration (MIC) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.12 | 0.25 | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 |

*How to predict exact value?*

# Supervised leaning in Machine Learning



# Minimum Inhibitory concentration (MIC) Prediction by machine learning methods

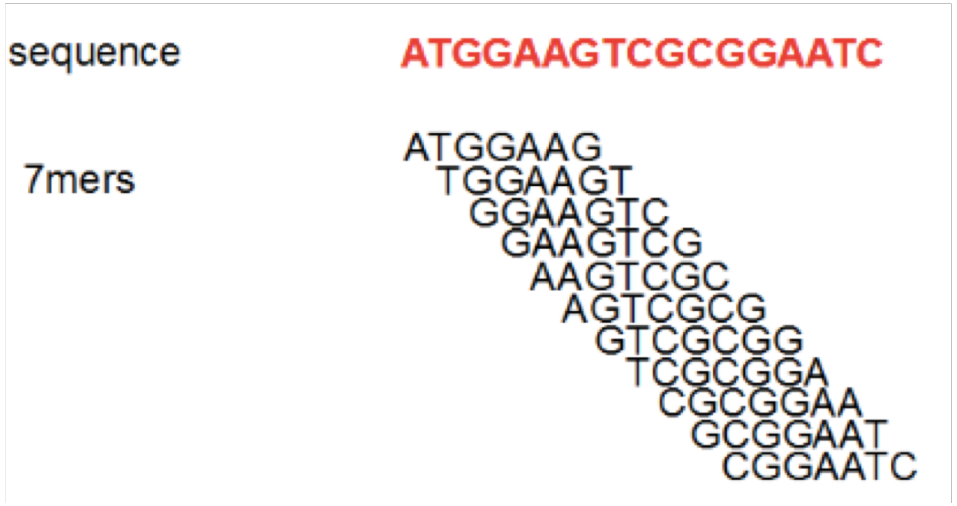
| Year | Species | Model | Target | Total | EA | ME | VME |
|---|---|---|---|---|---|---|---|
| 2017 | pneumococcus | Random forest | 3 PBP types | 4309 | 97% | 1.2% | 1.4% |
| 2017 | *Neisseria gonorrhoeae* | Multivariate linear regression models | ~10 AMR genes | 681 | 93% | 1.3% | 1.7% |

EA: essential agreement >90%
ME: major error S->R ≤ 3%
VME: very major error R->S < 3%

Li 2017, Eyre 2017

*Can We Use Whole Genome Data Without A Priori Information?*

# K-mer based modelling




| MATRIX | MIC | AAAATTTTCC | AAAATTTTCG | … |
|---|---|---|---|---|
| sample1 | 4 | 20 | 30 | … |
| sample2 | 8 | 10 | 20 | … |
| … | … | … | … | … |

***XGBoost (Extreme Gradient Boosting) model***

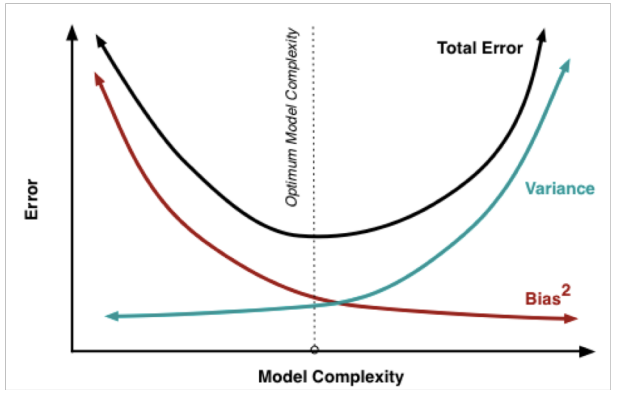
Ensemble Learning

Bagging | Boosting | Stacking

Random forest

Gradient Boosting

***XGBoost (Extreme Gradient Boosting) model***




Chen, 2016

# K-mer based modelling (1)

| Year | Species | Model | Target | Total | EA | ME | VME |
|---|---|---|---|---|---|---|---|
| 2017 | *Klebsiella pneumoniae* | XGBoost | Whole genome | 1668 | 92% | 3.7% | 3.1% |

Requires no *a priori* knowledge

EA is largely dependent on the number of resistant isolates that were sampled for each antibiotic

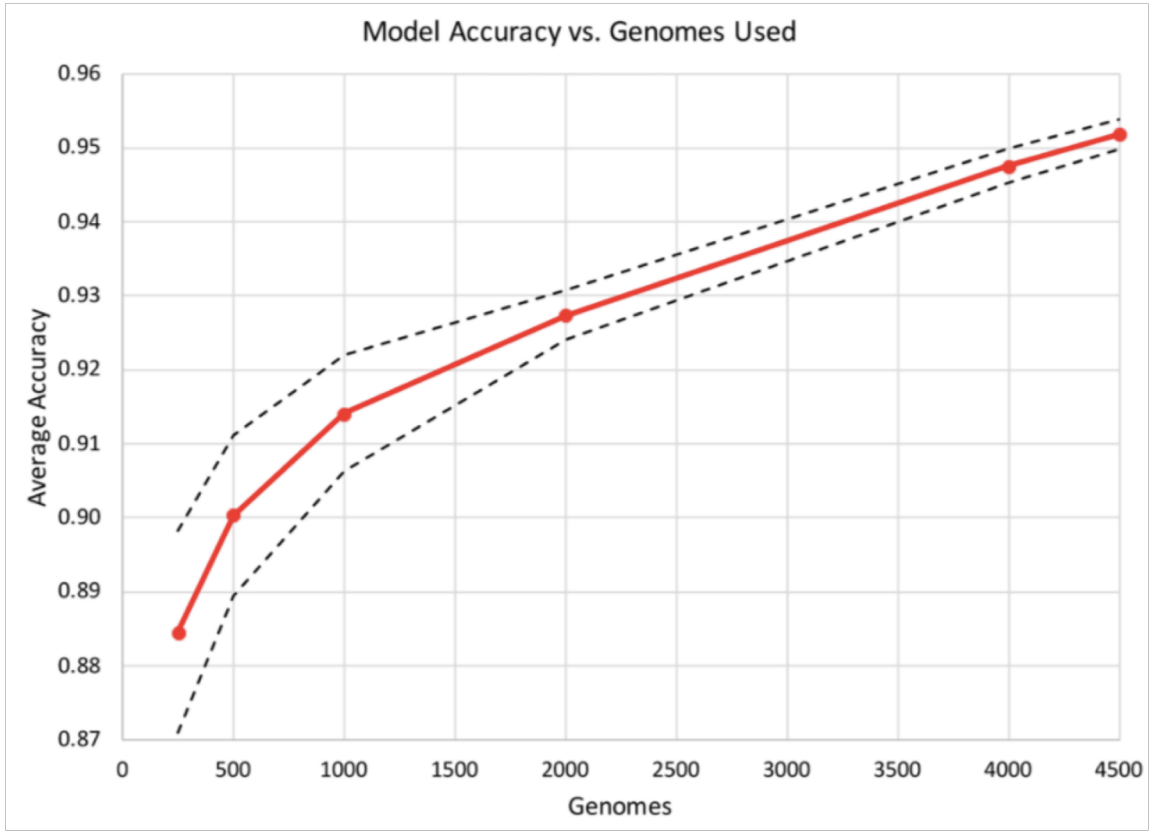
EA is similar (92%) of 3 models: Whole genome data & AMR genes  & Non-AMR genes

Nguyen, 2017

# K-mer based modelling (2)

| Year | Species | Model | Target | Total | EA | ME | VME |
|---|---|---|---|---|---|---|---|
| 2018 | Nontyphoidal *Salmonella* | XGBoost | Whole genome | 5278 | 95% | 2.7% | 0.1% |

Only trained 4500 genomes due to memory limit (1.5TB)

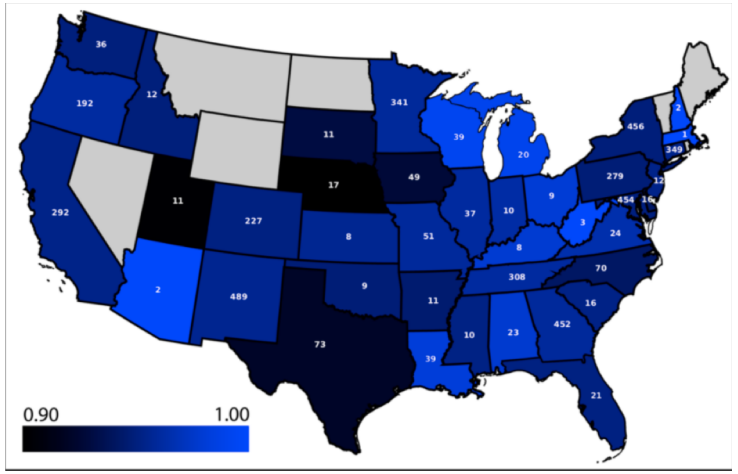



Nguyen, 2018

# EA is stable by year, source, states

| Collection Date | Accuracy |
|---|---|
| 2002 | 0.97 |
| 2003 | 0.95 |
| 2004 | 0.96 |
| 2005 | 0.95 |
| 2006 | 0.95 |
| 2007 | 0.94 |
| 2008 | 0.95 |
| 2009 | 0.95 |
| 2010 | 0.94 |
| 2011 | 0.95 |
| 2012 | 0.96 |
| 2013 | 0.97 |
| 2014 | 0.95 |
| 2015 | 0.95 |
| 2016 | 0.96 |

| Source | Accuracy |
|---|---|
| Chicken | 0.96 |
| Cow/Beef | 0.94 |
| Pig/Pork | 0.95 |
| Turkey | 0.94 |




| Training set years | Test set years | Accuracy |
|---|---|---|
| 2002-2008 | 2009-2016 | 0.88 |
| 2002-2009 | 2010-2016 | 0.88 |
| 2002-2010 | 2011-2016 | 0.88 |
| 2002-2011 | 2012-2016 | 0.88 |
| 2002-2012 | 2013-2016 | 0.88 |
| 2002-2013 | 2014-2016 | 0.86 |
| 2002-2014 | 2015-2016 | 0.92 |

# XGBoost assigns important k-mers predict MIC change

**Table 6.** The highest-ranking AMR-related protein function (or genomic region) with a matching k-mer from the XGBoost models.

| Antibiotic | K-mer Rank | Distance between k-mer and AMR gene[1] | k-mer | PATRIC Annotation(s) |
|---|---|---|---|---|
| AMP | 1 | Direct match | CTTAATCAGTGAGGC | Class A beta-lactamase (EC 3.5.2.6) => TEM family |
| AUG | 1 | Direct match | AAACGTCTTACTAAC | Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family |
| AXO[2] | 1 | 566.0 ± 39.7 | AAAGAGAAAGAAAGG | Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family |
| AZI | 8 | Direct match | CCCATTTCCGCCGCC | Macrolide 2'-phosphotransferase => Mph(A) family |
| CHL[2] | 1 | 611.8 ± 5.1 | AGACAAGTAAGCCGC | Chloramphenicol/florfenicol resistance, MFS efflux pump => FloR family |
| CIP | 1 | 313.5 ± 70.5 | ACAGTCCATCCAGGA | Pentapeptide repeat protein QnrB family => Quinolone resistance protein QnrB10 |

**Table 7.** Important k-mers used by the individual antibiotic models for predicting susceptible MICs.

| Antibiotic | k-mer | Sus[1] | Res[1] | Frac Sus[2] | Frac Res[2] | Genomic region[3] | PATRIC annotation or genomic region |
|---|---|---|---|---|---|---|---|
| NAL | ATTCCGCAGTGTATG | 5233 | 45 | 1.00 | 0.38 | PEG | DNA gyrase subunit A (EC 5.99.1.3) |
| AXO | TGGTATTCGCATCAA | 4508 | 769 | 0.78 | 0.48 | PEG | Phosphoethanolamine transferase EptA |
| KAN | CTGGCTTTTTTTTT | 837 | 84 | 0.30 | 0.00 | RNA | RyhB RNA |
| STR | CCCTTATCCAACACG | 872 | 1919 | 0.85 | 0.55 | PEG | Respiratory nitrate reductase delta chain (EC 1.7.99.4) |
| AXO | CAGAACCAGAATTTG | 4508 | 769 | 0.74 | 0.46 | PEGs | Formate-dependent nitrite reductase complex subunit NrfF, and Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly |

# Conclusions

- Whole genome sequencing offers the potential in predicting AMR
- Machine learning algorithms demonstrate value in MIC prediction with acceptable accuracy in clinical diagnosis
- XGBoost is readily to be applied to other important human pathogens even without *a priori* AMR information

# Limitation

- Training set: large, balanced database with metadata
- Interpretation: Machine learning models exhibit a trade-off between accuracy and intelligibility

# Reference (1)


- Jim O'Neill . *Tackling drug-resistant infections globally: final report and recommendations*. Review on Antimicrobial Resistance, 2016.
- Cooper, Matthew A., and David Shlaes. "Fix the antibiotics pipeline." *Nature* 472.7341 (2011): 32.
- Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115.
- Bychkov, Dmitrii, et al. "Deep learning based tissue analysis predicts outcome in colorectal cancer." *Scientific reports* 8.1 (2018): 3395.
- US Food and Drug Administration. "Class II special controls guidance document: antimicrobial susceptibility test (AST) systems." *Food and Drug Administration, Silver Spring, MD*(2009).
- Land, Miriam, et al. "Insights from 20 years of bacterial genome sequencing." *Functional & integrative genomics* 15.2 (2015): 141-161.
- Stoesser, N., et al. "Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data." *Journal of Antimicrobial Chemotherapy* 68.10 (2013): 2234-2244.
- Bradley, Phelim, et al. "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis." *Nature communications* 6 (2015): 10063.
- Gordon, N. C., et al. "Prediction of Staphylococcus aureus antimicrobial resistance from whole genome sequencing." *Journal of clinical microbiology* (2014): JCM-03117.

# Reference (2)


- McDermott, Patrick F., et al. "The use of whole genome sequencing for detecting antimicrobial resistance in nontyphoidal Salmonella." *Antimicrobial agents and chemotherapy* (2016): AAC-01030.
- Li, Yuan, et al. "Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences." *BMC genomics* 18.1 (2017): 621.
- Eyre, David W., et al. "WGS to predict antibiotic MICs for Neisseria gonorrhoeae." *Journal of Antimicrobial Chemotherapy* 72.7 (2017): 1937-1947.
- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016.
- Nguyen, Marcus, et al. "Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae." *Scientific reports* 8.1 (2018): 421.
- Nguyen, Marcus, et al. "Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal Salmonella." (2018): *JCM* 01260-18.

# Thank you